DESIGN AND DEVELOPMENT OF A NOVEL APPROACH TOWARDS HUMAN INTERPRETABILITY IN MACHINE LEARNING BASED SOLUTIONS

A Thesis
Submitted in partial fulfillment of the requirements for the award of the degree of

DOCTOR OF PHILOSOPHY

in

Computer Applications

by

Pawan Kumar

Registration No. 41500197

Supervised by

Dr. Manmohan Sharma

LOVELY PROFESSIONAL UNIVERSITY
PUNJAB - 144 411, India
November 2021

Dedicated to my parents, family and friends

DECLARATION

I declare that

(a) The work contained in the thesis is original and has been done by myself under the general supervision of my supervisor;

(b) The work has not been submitted to any other institute for any degree or diploma;

(c) I have followed the guidelines provided by the University in writing the thesis;

(d) I have conformed to the norms and guidelines given in the Ethical Code of Conduct of the University;

(e) Whenever I have used materials (data, theoretical analysis, and text) from other sources, I have given due credit to them by citing them in the text of the thesis, giving their details in the references, and taking permission from the copyright owners of the sources, if necessary;

(f) Whenever I have quoted written materials from other sources, I have put them under quotation marks and given due credit to the sources by citing them and giving required details in the references.

Place: Phagwara

Date: 19 / 11 / 2021

Pawan Kumar

CERTIFICATE

Date: 19 / 11 / 2021

This is to certify that the thesis entitled **Design and Development of a Novel Approach Towards Human Interpretability in Machine Learning Based Solutions**, submitted by **Mr. Pawan Kumar** to Lovely Professional University Phagwara, is a record of bona fide research work carried out by him under my supervision and is worthy of consideration for the award of the degree of **Doctor of Philosophy** of the University.

Dr. Manmohan Sharma
Associate Professor
School of Computer Applications
Lovely Professional University Phagwara
Punjab - 144 411, India

Acknowledgements

I am thankful to my supervisor **Dr. Manmohan Sharma**, who has been a consistent motivator and mentor to me during nurturing of the ideas presented in this thesis. Whenever I was finding it tough to keep moving, he was always there to hold my hand and take me through. Without your support, it was just not possible to come this far, Sir.

I am thankful to **Dr. M Syamala Devi**, Professor, Department of Computer Science and Applications, Panjab University Chandigarh. She has been my first mentor when I started my journey as a research student. I admire her for her core ethical values and expertise in the field of Artificial Intelligence. The fundamental lessons you taught me have been a guiding force always, Madam.

I am thankful to **Dr. Rajesh Kumar Gupta**, Professor, Department of Mathematics, Lovely Professional University, who always gave me time for the discussions despite his busy schedule. Sir, your strong mathematical concepts and research experience made our discussions fruitful. I always came out from your office full of new ideas.

I am thankful to **Ashwani Tewari Sir**, Head, School of Computer Applications, Lovely Professional University, for his motivation and feedback related to my work. Sir, your inputs always provided me with new insights and helped me improve.

I am thankful to **Dr. Santosh Kumar Henge**, Associate Professor, School of Computer Applications, for his valuable inputs regarding how to write a quality scientific manuscript and present your work to journals. Sir, you have been my true support.

I am thankful to **Dr. Amar Singh**, **Dr. Anil Sharma**, **Dr. Ramandeep Singh**, **Dr. Devendran** and **Dr. Mithilesh Dubey** from Lovely Professional University, for sparing time to review my work and provide valuable feedback. Sirs, your valuable inputs helped me improve my work.

I am thankful to the **Support Staff** at the School of Computer Applications, Lovely Professional University for the facilities provided. You all have been a helping hand always.

I am thankful to my friend **Rajesh Gupta**, research scholar at Nirma University, Gujarat, India. Your suggestions regarding how to write scientific papers and use tools helped me improve my productivity as a research scholar.

This list can not be complete without acknowledging my friend **Sanjay**, who has always been after me to stay focused and complete my dissertation work.

Pawan Kumar

Abstract

Machine learning (ML) refers to training a machine to learn from experience. This experience is provided as input to a machine in the form of a dataset. Intuitively, the behaviour of the learned ML model is expected to agree with the provided dataset. Moreover, to win the trust of its intended users, the ML model must agree with the perception of human domain experts regarding the problem domain. So, agreement of an ML model with the underlying dataset and perception of human domain experts are the keys to accurate and trustable learning.

ML is moving out from the hands of scientists and research labs to the masses. It is becoming important that the human users of ML-based solutions should be able to interpret ML model behaviour as well as to provide feedback based on their experience-based perception regarding the problem domain. A majority of the existing work towards conferring interpretability to ML models has been focusing on explaining the behaviour of ML models. However, interpretability is a means to an end and not the end itself. There can be other ways to achieve the broader goals of interpretability. An important motivation behind conferring interpretability is to facilitate trust in an ML-based solution. There is a need for a framework that can help explore the solution space of an ML problem to find out an ML-based solution that is accurate as well as agrees with its intended human users.

The goal of this thesis is to propose a framework that enables verification of the learning acquired by an ML model against the provided dataset and the perception of the human domain experts regarding the problem domain. This two-way verification ensures that the learning process has been reliable and the acquired learning is in sync with the prevailing domain knowledge. Additionally, this framework should be capable of making the ML model explore the possibility of adapting itself to incorporate human expert feedback, in case it is lacking in terms of the user agreement. The underlying dataset, the ML model learned using that dataset and feedback of human experts are the three pillars of this framework. The proposed framework listens to each of these pillars individually and analyzes the agreement between these three pillars.

To listen to the dataset, this research used entropy and Gini index as information gain measures to identify important characteristics of the dataset. To listen to the ML model, this research used variable importance measures to identify features that were most affecting the decision-making behaviour of the model. To listen to the human domain experts, this research collected feedback from human experts in terms of feature impor-

tance as per their experience-based perception of the problem domain. To get unbiased feedback from human experts, outcomes of listening to the dataset and the ML model were not shared with human domain experts. The problem of measuring agreement between these three pillars (data, model and human) has been modelled as a '2-Judges n-participants' Spearman's rank correlation problem, taking a pair of judges at a time. As an analogy, dataset, ML model and human experts were considered as equivalent to judges. Each feature used in the learning ML model was considered equivalent to a participant. Hypothesis testing using rank correlation coefficient has been used to verify the statistical significance of the agreement between dataset, ML model and human domain experts.

First, this research demonstrates the above framework to verify the learning acquired by an ML model against the evaluated dataset and the prevailing domain knowledge (extracted through feedback from human domain experts). Second, this research expands the proposed framework to enable human-feedback adaptive learning to improve user agreement. The underlying motivation is that humans and machines have unique capabilities and they can collaborate to complement each other. Human-feedback adaptive learning aims to make an ML model adapt itself to incorporate feedback from human domain experts in case it is lacking in terms of the agreement. The proposed framework has been demonstrated using a novel problem domain of predicting the joining behaviour of freshmen students using a primary dataset. The proposed framework has been also demonstrated for predicting the onset of diabetes using a publicly available standard dataset. We demonstrate human-feedback adaptive learning by enabling the ML system to move from a state of 'insignificant agreement' to a state of 'significant agreement' without losing much in terms of accuracy.

Humans are inherently good at interpreting visualizations. They have been using visualizations to explain different phenomena under study. It is no surprise that most of the approaches towards conferring human interpretability to ML models lead to visualization as an outcome. This research proposes a novel visualization FIFI graph that can enable interpreting ML models by analyzing feature importance and feature interactions using a single visualization. As this visualization is modelled as a graph, the existing research and software tools for network analysis can be used to analyze this visualization. The proposed visualization has been demonstrated using the problem domain of predicting the joining behaviour of freshmen students using a primary dataset and predicting customer churn using the telco customer-churn dataset available on Kaggle.

The potential applications of this research work are in designing ML-based solutions for problem domains involving critical decision-making. In such contexts, accuracy, as well as user agreement, are both equally important for an ML-based solution. For example, in the medical domain, any ML-based solution for assisting human medical experts is likely to succeed only if its decision-making is in sync with the experience-based perception of the human medical experts. The limitations of this research work include a requirement of human domain experts and dependency on the selection of quality human experts. It requires features to concepts mapping for complex domains where it is not feasible to provide feedback in terms of feature importance.

The future scope of this research includes exploring alternatives of entropy and Gini index as information gain measures, alternative human interpretability techniques for listening to the ML model and alternatives other than the ranking of features to collect human expert feedback. Instead of Spearman's rank correlation, weighted correlation can be explored as the degree of agreement between the three pillars in terms of top-ranked features is more crucial. Innovative variants of existing algorithms can be explored to implement human-feedback adaptive learning aiming at better user agreement. Enabling human users with varying levels of domain expertise interact with ML systems and making use of UI/UX practices is an opportunity to take ML to the masses. Due to their analogy being a graph, exploring the interaction between a FIFI graph and a knowledge graphs is another interesting research direction with potential impact.

Contents

List of Figures

List of Tables

Nomenclature

n	Number of features used for learning
$IG_{e,f}$	Information gain using entropy for feature ‘f’
$IG_{g,f}$	Information gain using gini index for feature ‘f’
$R_{e,f}$	Rank of a feature ‘f’ using $\mathrm{IG}_{e,f}$
$R_{g,f}$	Rank of a feature ‘f’ using $\mathrm{IG}_{g,f}$
$R_{dataset,f}$	Average of $\mathrm{R}_{e,f}$ and $\mathrm{R}_{g,f}$
Imp_f	Importance measure of feature ‘f’ as per model
acc_d_f	Mean decrease in accuracy for feature ‘f’ as per model
$gini\_d_f$	Mean decrease in the Gini index of node impurity for feature ‘f’ as per model
mmd_f	Mean minimal depth for feature ‘f’ as per model
$times\_root_f$	Number of times a feature ’f’ is used a root node
$R_{Imp,f}$	Rank of a feature ‘f’ as per Model using Imp_f
$R_{acc\_d,f}$	Rank of a feature ‘f’ as per Model using $\mathrm{acc\_d}_f$
$R_{gini\_d,f}$	Rank of a feature ‘f’ as per Model using $\mathrm{gini\_d}_f$
$R_{mmd,f}$	Rank of a feature ‘f’ as per Model using mmd_f
$R_{times\_root,f}$	Rank of a feature ‘f’ as per Model using $\mathrm{times\_root}_f$
$R_{Model,f}$	Average rank for a feature ’f’ as per Model
K	Number of human domain experts shortlisted
$R_{h_i,f}$	Rank assigned to a feature ‘f’ by human expert h_i
$R_{Human,f}$	Average rank assigned to a feature by human experts
ρ_s	Spearman’s rank correlation coefficient

Chapter 1

Introduction

Humans and machines have unique strengths. Humans are good at decision-making in unseen situations owing to their experience. Machines are good at processing data with the desired level of detail and accuracy.

Machine Learning (ML) refers to training a machine to learn from experience (Mitchell, 2006). This experience is usually provided in the form of a dataset. Also, it is imperative that this acquired learning is in sync with the prevailing domain knowledge as it facilitates the trust of users in an ML model. Therefore, in any ML-based solution, the underlying dataset, ML model and domain-specific human experts are expected to agree with each other in terms of their perception regarding the problem domain. The agreement between the dataset and the learned ML model indicates that the model has been able to capture important characteristics from the dataset. The agreement between the decision-making behaviour of the model and the perception of human domain experts facilitates trust in the learning acquired.

There are several challenges to such a collaboration. These include the availability of human domain experts, they may or may not have any knowledge about ML, and even there can be variations in the level of their expertise regarding the problem domain. Moreover, these human experts must be able to interpret the ML model so that they can provide feedback based on their domain knowledge. There is a need for ML algorithms that can incorporate this feedback from human domain experts. Another constraint is to ensure that while attempting to incorporate this feedback, the ML model remains accurate enough.

The field of ML has been witnessing tremendous growth during recent years owing to its wide scope of applications. Some of the important applications of ML include speech recognition, bio-surveillance and robot control (Guyon and Elisseeff, 2003). Some of the most recent applications of ML include advertisement classification (Jain et al., 2021), life insurance (Rani et al., 2021), secure data analytics (Gupta et al., 2020) and cryptocurrency price prediction (Patel et al., 2020).

Software engineers are coming up with tools with ML capabilities integrated into them. Using these tools, even novice users can build ML models for their respective problem domains. However, creativity, experience and domain knowledge are critical to the success of ML models (Wagstaff, 2012)(Storcheus et al., 2015)(Domingos, 2012).

As ML is moving out of research labs to the hands of common users, these users must be able to interpret the outcome of these models. However, a majority of the advanced ML models behave like a black box in the sense that their outcome is not interpretable to human users, particularly if they are not ML experts. This is termed as lack of human interpretability in ML and is a major hurdle in the further growth of ML (Ribeiro et al., 2016*b*). Also, human users must be able to interact with these ML-based systems. It provides an opportunity to learn from the experience of human users (Kim, 2015).

During recent years, a renewed interest has been observed among the research community towards conferring human interpretability to ML models. Human interpretability has several advantages like facilitating trust, debugging ML models, potential to discover new knowledge, ensuring fairness and addressing right-to-explanation situations (Ribeiro et al., 2016*a*)(Lipton, 2016)(Doshi-Velez and Kim, 2017). However, this field is still evolving and there are several research gaps. These research gaps include discovering important interactions among variables to provide new insights and facilitate verification by human experts (Strumbelj and Kononenko, 2010), exploring alternatives as explainers (Ribeiro et al., 2016*b*), evaluation metrics, providing global explanations (Ribeiro et al., 2016*a*), and the need of incorporating domain knowledge into the ML process (Kim, 2015)(Yang et al., 2019)(Holzinger et al., 2019). However, it is important to understand that interpretability is a means to an end and not the end itself. There may be alternate ways to achieve the broader goals of interpretability.

The goal of this thesis is to propose and demonstrate a framework for collaboration between data, ML model and human domain experts. The objective of this collaboration framework is to verify the learning acquired by an ML model against the provided dataset and prevailing domain knowledge. For this verification, the idea is to listen to the dataset, ML model and human experts individually to know their perception regarding the problem domain. The framework then measures and analyze the degree of agreement between these three. A significant agreement between these three indicates that learning acquired by the model is accurate as well as in agreement with the prevailing domain knowledge. Moreover, in case of a lack of agreement with human experts, the framework makes the ML model attempt to adapt itself by re-exploring the solution space, to align with the human user's perception regarding the problem domain. Feature importance is a commonly used measure to interpret the decision-making behaviour of an ML model. Studying important interactions between features is important to understand which features interact more. This thesis also proposes a graph-based visualization that can be used to analyze feature importance and feature interactions using a single visualization. Figure 1.1 gives a brief outline of this thesis.

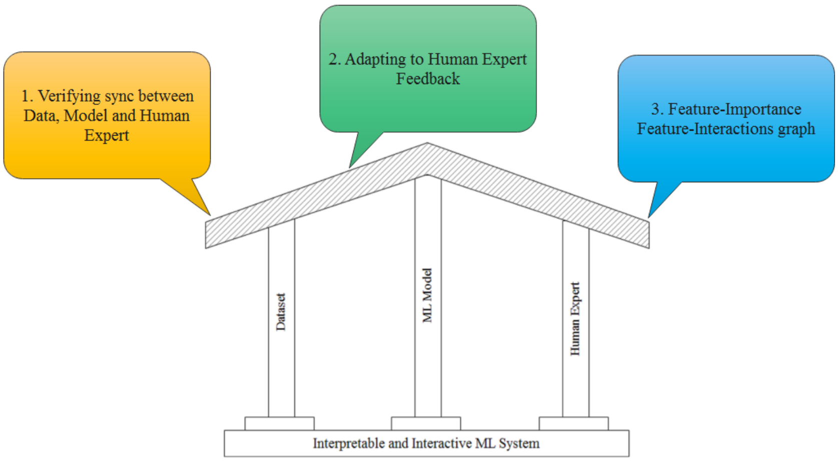

1. Verifying sync between Data, Model and Human Expert
2. Adapting to Human Expert Feedback
3. Feature-Importance Feature-Interactions graph
Dataset
ML Model
Human Expert
Interpretable and Interactive ML System

Figure 1.1: Thesis outline diagram

Organization of the chapter: The goal of this chapter is to summarize the key innovations presented in this thesis. The section structure follows the chapter structure of the thesis. Section 1.1 gives a summary of the existing literature in the field of human interpretability. It talks about the motivation behind studying interpretability, the taxonomy of the existing work towards conferring interpretability, key ideas and explainers used for explaining outcomes, desired characteristics of an explanation, evaluation metrics, and problem domains attempted. Section 1.2 proposes and demonstrates a framework that considers dataset, ML model and human experts as the three pillars of an ML-based solution. The framework is capable of verifying learning acquired by the ML model against the underlying dataset and the prevailing domain knowledge. Section 1.3 discusses the extension of the proposed framework to enable an ML model to incorporate feedback from human domain experts. The proposed framework listens to human feedback and aims to adapt itself to improve user agreement. In case of a conflict between human feedback and data, the ML system is capable of reporting it as a conflict. Section 1.4 presents Feature-importance Feature-Interactions (FIFI) graph, a novel graph-based visualization for analyzing important features and important inter-feature interactions using a graph object. The fact that it is a graph brings the potential advantage of applying existing research and software tools for network analysis.

1.1 Human interpretability in ML-based solutions

The goal of this chapter is to investigate the approaches proposed towards conferring human interpretability to ML-based solutions. The focus has been to identify the key ideas and motivations. The sources of papers selected for review included prominent ML journals and leading conferences. The research studies were shortlisted using predefined search strategies and inclusion criteria. The conclusions from the investigation are presented in the form of answers to the important research questions like underlying motivations, characteristics of a good explanation, the taxonomy of approaches proposed, key ideas used for conferring interpretability, alternatives used as explainers, ML algorithms attempted for interpretability, problem domains attempted, evaluation metrics, opportunities in human-computer interaction and possible future research directions.

Human interpretability in machine learning (ML) based solutions refer ability to explain its outcome in a manner that is understandable to a human expert. Human interpretability of ML-based solutions has potential advantages like facilitating trust, model debugging and ensuring fairness (Ribeiro et al., 2016*b*)(Doshi-Velez and Kim, 2017). A renewed interest in making ML-based solutions human-interpretable has been observed among the ML community during recent years.

Goal: A critical evaluation of the existing work done in the field of conferring human interpretability

The following research questions were formulated to keep the investigation guided:

- What are the underlying motivations for human interpretability?
- What are the characteristics of a good explanation?
- What is the taxonomy of the approaches proposed?
- What are the key ideas and explainers used for conferring interpretability?
- What are the metrics used for evaluating interpretability?
- Which ML models have been attempted for interpretability?
- What are the types of problems attempted?
- Is Interpretability compulsory always?
- Can human-machine collaboration be useful in incorporating domain expertise?
- What are the open research directions?

Approach

Search terms were constructed using common terms related to human interpretability, alternate spellings, synonyms and keywords from relevant papers. Some of the most relevant search terms included: human interpretability in ML; explainable ML;

transparency in ML; understanding the outcome of classification models; understanding black-box models and accuracy-interpretability trade-off. Sources of the search included the reputed ML journals and conferences. A few of the most prominent ones include Journal of Machine Learning Research (JMLR), Pattern Recognition, IEEE transactions, Machine Learning – Springer, International Conference on Machine Learning (ICML), and Neural Information Processing Systems (NIPS).

The papers covering one or more of the following criteria were shortlisted: (i) proposing an approach for conferring interpretability (ii) accuracy-interpretability trade-off (iii) implementing human interpretability approach on an ML problem and (iv) position papers on human interpretability in ML, and (v) human-machine collaboration focusing incorporating prevailing domain knowledge.

Findings from the literature survey

The potential advantages of human interpretability include facilitating trust of the end-user in the ML model, providing new insights about the phenomena under study, providing scrutinizing ability to a human expert, and ensuring fairness or unbiasedness. Accuracy-interpretability trade-offs also make it important. Starting May 2018, as per new Global Data Protection Regulations (GDPR), human subjects that are going to be affected by the outcomes of an ML-based solution have the right to get an explanation for that outcome.

A good explanation should be interpretable and easy to understand. The number of pieces of information should not exceed the comprehension capabilities of the user. An explanation should be in terms of concepts known to the target audience. An explanation must at least be locally true. It must agree to the behaviour of the model at least in the neighbourhood of the example for which model outcome is being explained. Explanations should be model agnostic i.e. these are applicable irrespective of which underlying ML model is used.

The existing work towards conferring interpretability can be broadly classified into model-agnosticism(Ribeiro et al., 2016*b*)(Ribeiro et al., 2016*a*)(Baehrens et al., 2010), approaches for interpreting tree ensembles (Hara and Hayashi, 2016)(Deng, 2019)(Vandewiele et al., 2016), approaches for interpreting neural networks (Hechtlinger, 2016)(Samek et al., 2016)(Montavon et al., 2017) and addressing lack of consensus in this evolving field (Bibal and Frénay, 2016)(Weller, 2017)(Doshi-Velez et al., 2017). Model-agnostic approaches can be further classified into (a) work done towards approximating a complex model with a simple interpretable model and (b) modelling input-output relationship without attempting transparency in the underlying model, termed as visual analytics (Krause et al., 2016)(Villagrá-Arnedo et al., 2017).

Key ideas employed for conferring interpretability to ML models included: approximating a complex model using a simpler but interpretable model (Ribeiro et al., 2016*b*); identifying important features using local explanation vectors (Baehrens et al., 2010); explaining ML model outcome in terms of feature-wise contribution; listing important features using explanation curves; interpretability using partial derivatives with respect to input; using programs snippets as explanations (Singh et al., 2016). Approaches for interpreting tree ensembles include: extracting an interpretable decision tree from an ensemble (Hara and Hayashi, 2016); using Genetic Algorithms to extract a single interpretable tree from a population of trees (Vandewiele et al., 2016); and finding prototypes in tree space (Tan et al., 2016). Other ideas included: extracting interpretable features and summarizing results in the form of rule sets; Bayesian framework for learning a falling rule list (Wang and Rudin, 2015); learning a small set of rules in disjunctive normal form; heat maps to show most contributing pixels using deep-Taylor decomposition. Interpretability has also been attempted through modelling input-output relationships and making outcomes of a black-box model more expressive using heatmaps and progression charts (Krause et al., 2016)(Villagrá-Arnedo et al., 2017).

Metrics used for evaluating interpretability include: volume of the information in the explanation to be comprehended; ease of identifying how to move a data point to get its prediction outcome label changed; quality of explanations evaluated by human experts; how many outcomes are explained using only top-ranked features?; using explainability score to compute MEP(Mean Explainability Precision) and (MER) Mean Explainability Recall; quantifying interpretability using information entropy instead of the count of features.

Interpretability is not compulsory always and in fact can be harmful in certain contexts. It is particularly useful in situations that involve critical decision making and a say of human domain experts. Possible dangers of transparency include divergence between intended audience and the actual beneficiary; transparency in government use of algorithms; lack of motivation of intellectual property if all algorithms are open source; discrimination of sub-groups based on sensitive features.

Humans and machines have unique strengths and can complement each other. Humans-machine collaboration has potential advantages like improved user agreement, accelerated exploration of solution space, providing a bigger role to end-users in the design of interactive and interpretable ML systems, and incorporating multi-disciplinary expertise. The key idea is to present the learned model to human domain experts in an interpretable manner. The human experts in turn provide feedback back to the ML system based on their experience-based perception regarding the problem domain. The ML systems attempt to adapt themselves to incorporate the feedback provided by human experts.

There are several interesting opportunities for carrying out future research. There is a need to discover important interactions of variables to provide new insights; facilitate verification by human experts; explore alternatives as explainers; provide global explanations; address inconsistencies in local explanations; evaluate alternatives for proximity measures between a complex model and its interpretable approximation; evaluating alternatives as metrics; finding new graphical tools to improve expressiveness. As this field is still evolving and is quite subjective, there is a need of addressing the lack of consensus regarding definitions, types and metrics to measure interpretability. There is a need to define criteria for the faithfulness of an explanation. As interpretability is not always compulsory, there is a need for a framework to understand when transparency is useful and when it is harmful. As ML is moving from research labs to the hands of the masses, it is important to utilize human domain expertise for incorporating domain knowledge into the ML process. There is a need for clearly defined requirements for input and goals of the output; preparing a front-end for the intended audience in terms of concepts known to them to improve interpretability. Some generic research directions include reducing computational complexity for interpretability in real-time and extending approaches implemented for classification problems to regression tasks.

1.2 Data, Machine Learning and Human Experts: A collaboration

The goal in this chapter is to propose a three-pillar framework with the provided dataset, learned ML model and human expert feedback as to its three pillars. The objective of the framework is to verify the learning acquired by the model against the provided dataset and prevailing domain knowledge. This chapter describes the proposed framework, methods used to listen to each of these three pillars and computing degree of agreement between these. The intuitive idea is that in an ideal ML-based solution, the provided dataset, ML model learned using the provided dataset, and human domain experts must agree in terms of their perception regarding the problem domain. The proposed approach has been evaluated using two problem domains: (i) a novel problem domain of predicting the joining behaviour of freshmen students using a primary dataset and (ii) ’diabetes’, a publicly available standard dataset. A positive degree of agreement was observed between the dataset, ML model and human experts in terms of importance assigned to features. The agreement between the provided dataset learned model and human domain experts helps verify learning acquired by the ML model and facilitates trust in the ML model. This work can potentially form the basis for developing formal quantitative metrics for evaluating ML models in terms of reliable learning and capability to facilitate trust of human users.

Goal: A framework that verifies the agreement between the dataset, ML model and human expert after listening to each individually

The goal is to propose and demonstrate a framework that verifies the learning acquired by an ML model against the provided dataset and human expert feedback. The intuitive idea is that in any ML-based solution, the provided dataset learned ML model and human expert feedback should agree. These three can be regarded as the three pillars of an ML-based solution. The proposed framework aims to listen to each of these pillars individually and attempts to answer the following research question:

Q. Is the provided dataset, the learned ML model and human expert feedback in agreement regarding their perception of the problem domain?

This chapter describes the proposed framework along with methods used to listen to the dataset, ML model and human experts. The proposed approach has been demonstrated using a primary dataset from a novel problem domain.

Approach used

The input to the framework is the underlying dataset and the ML model learned using this dataset. The output of the framework is to verify whether the learning acquired by the model is accurate and in sync with the prevailing domain knowledge. The framework aims to listen to the provided dataset, learned ML model and human expert feedback. Listening to the dataset refers to identifying important characteristics by computing information gain using entropy and the Gini index. Listening to the ML model refers to interpreting its decision-making behaviour by computing feature importance as per the model. Listening to human experts refers to taking feedback in terms of important features as per their perception of the problem domain. After listening to each of these three pillars, the framework verifies the degree of agreement between them regarding the perception of the problem domain.

Results and contributions

For the problem domain of predicting the joining behaviour of freshmen students, the input dataset consisted of records of freshmen students enrolled in a higher education institute. The objective was to classify which of these enrolled students are likely to join at the start of the session. For the ’diabetes’ dataset, the objective was to predict the onset of diabetes based on different diagnostic measures. We demonstrated that the proposed framework was capable of verifying the learning acquired by the ML model against the provided dataset and prevailing domain knowledge as per feedback from human domain experts. A statistically significant degree of agreement was observed between the dataset, ML model and human expert feedback. This chapter has the following contributions:

(i) A framework that listens to the important pillars of an ML-based solution: the dataset, ML model and human domain experts.

(ii) Enable verification of the learning acquired and facilitating trust of human users.

(iii) Due to its quantitative nature, this approach is having the potential of forming a basis for developing formal metrics for facilitating trust in an ML model.

1.3 Human-feedback Adaptive Learning

Humans and machines have their unique strengths and collaboration between these two have the potential to improve ML systems further. For any such collaboration, human users must be able to interpret the behaviour of ML systems. Moreover, human users should have a provision to give feedback to ML systems based on their domain knowledge. Consequently, apart from accuracy, human interpretability and the ability to interact with human experts are becoming crucial parameters for ML systems (Yang et al., 2019)(Kim, 2015)(Holzinger et al., 2019).

The solution space of an ML problem may have multiple solutions that are equally good in terms of internal optimization metrics. However, these solutions may differ in terms of their alignment with the user's perspective of the problem domain. Including human experts in this exploration of the solution space of a problem has the potential to help search for a solution that not only satisfies the threshold for internal metrics but has improved agreement with the user's perspective regarding the problem domain under study.

The goal in this chapter is to extend the framework proposed in chapter 3 by incorporating human-feedback adaptive learning capability. In case there is a lack of agreement between the ML model and human domain experts, the proposed algorithm makes the ML model search for a solution that is improved in terms of agreement with the user. Moreover, it takes care that the accuracy of the model does not go below pre-decided threshold limits in the context of the problem domain.

Goal: Human-feedback adaptive learning

The objective of the proposed framework is to incorporate prevailing domain knowledge collected in the form of feedback from human experts regarding their perception of the problem domain.

This chapter also provides an algorithmic and graphical description of the workflow involved in the design of ML systems with human-feedback adaptive capability. As the field of conferring the interactive ability to ML systems is new, there is a lack of established guidelines and evaluation metrics. This chapter lists a set of guidelines for the design of interpretable and interactive ML systems. Novel evaluation metrics that can be used for evaluating these interpretable and interactive ML systems have also been proposed. These metrics have been thought of keeping in mind that human users who are interacting with ML systems may not be ML experts at all.

Approach used

This chapter extends the capability of the framework proposed in chapter 3. The proposed approach is based on an intuitive idea that the provided dataset, ML model learned using that dataset and feedback from human domain experts are important pillars of an ML-based solution. The proposed approach listens to each of these three pillars and measures the degree of agreement between them. The problem of measuring the degree of agreement between these three is modelled as a rank correlation problem with three judges (dataset, model and domain knowledge) and 'n' participants (features used for learning).

If the agreement between these three pillars is lacking, the model is rebuilt to incorporate feedback from human experts. The revised model is acceptable only if it remains accurate enough as per the requirements of the problem domain. If the revised model can achieve agreement between dataset, model and human experts without a significant decrease in classification accuracy, it is considered as a gainful trade-off. An intuitive idea to align the ML model with the perception of human experts regarding the problem domain is to relearn while giving lesser importance to features ranked least important by human experts. Hypothesis testing using Spearman's rank correlation has been used to verify the statistical significance of the agreement between the ML model and human domain experts.

Results and contributions

The proposed human-feedback adaptive algorithm has been demonstrated using our running example of prediction joining behaviour of freshmen students. The capability of the framework has been demonstrated by reaching out from a 'lack of agreement situation' to 'Presence of agreement situation' by making the ML model adapt itself to incorporate human expert feedback. The major contributions of this chapter are:

(i) An algorithmic and graphical description of the working of interactive ML systems

(ii) Proposes human-feedback adaptive learning algorithm to enable an ML model to adapt to human expert feedback

(iii) Proposing novel opportunities in human-feedback adaptive learning

(iv) Proposing metrics for the evaluation of interpretable and interactive ML systems.

1.4 Feature-importance and Feature-interactions graph

Humans are inherently good at interpreting visualizations and have been using these visualizations to explain a phenomenon under investigation. As a result, most of the approaches for conferring human interpretability to ML models have resulted in a visualization. These visualizations enable a human user to interpret the decision-making behaviour of an ML model. Most of the existing visualizations focus on identifying features that are relatively more important in terms of affecting the decision-making behaviour of the ML model. Apart from studying the importance of features, it is also important to identify important interactions between features. The identification of important interactions brings an opportunity to uncover important interactions that are still unknown to the community concerned with the problem domain.

Goal: To propose a novel visualization for human interpretability

The goal of this chapter is to propose a novel visualization FIFI(Feature-importance Feature-interactions) graph, for interpreting ML models by analyzing important features as well as important interactions. The objective is that a human user should be able to identify relatively important features and relatively strong interactions. Moreover, the visualization should provide control in the hands of the user to filter out important interactions and plot the visualization corresponding to the subset shortlisted. The proposed visualization has been demonstrated using two problem domains, one using a primary dataset and the other using a publicly available standard dataset. As the proposed visualization has been modelled as a graph, the existing research on network analysis can be applied to analyze this visualization. Also, due to its natural analogy, a comparison has been made with a knowledge graph.

Approach used

The proposed visualization has been modelled as a graph. Each node of this graph represents a feature used by the ML model for learning. Each edge of this graph represents the interaction between a pair of features. This visualization depicts the relative importance of features and the relative magnitude of interactions between features used in learning an ML model. The size of a node is made proportional to the importance of the corresponding feature. The width of an edge is proportional to the magnitude of the interaction between the corresponding pair of features.

Results and contributions

The proposed approach has been demonstrated using two problem domains. The first problem domain was our running example of predicting the joining behaviour of freshmen students. The second problem was of predicting customer churn using a standard dataset from Kaggle. In both the demonstrations, the proposed visualization has been able to identify important features and important interactions between features. This chapter has the following contributions:

(i) To propose a novel visualization for interpretability that can enable a human user to interpret an ML model behaviour.

(ii) The proposed visualization is a compact alternative to existing visualizations as it enables the identification of important features and important interactions using a single plot.

(iii) The proposed visualization provides control to the user in terms of filtering out the most important interactions.

(iv) Existing tools and software for network analysis can be applied to analyze this visualization as it is modelled as a graph object.

1.5 Problem domains identified for experimentation

For experimentation, the following problem domains were used:

1.5.1 Predicting joining behaviour of freshmen students

As per the University Grant Commission (UGC) report, there are almost 900 universities across India including more than 300 private universities (UGC, 2019). Most of these private universities spend hugely in competing to reach out to students who are exploring different options to take admission into. Every admission year, from the students, enrolled in an educational institution, some students do not join the same institution. The reasons include getting a higher scholarship in some other institution, preferred college or discipline of interest in some other institution. Each student, who takes admission but do not join is a loss to the concerned educational institute in terms of resources invested. Ability to foresee such students is important for an educational institute. Institutes can use this information to plan activities towards improving the retention of enrolled students.

The objective is to explore the applicability of ML in helping educational institutions predict the joining behaviour of their freshmen students. The idea is to formulate this research question as a binary classification ML problem. The target variable is the actual joining status of a student with ‘Joined’ or ‘Lost’ as possible outcomes. The first objective is to learn an ML-based model using admission details and the actual joining status of freshmen students of previous batches for learning. Such a model can be used for predicting the joining behaviour of freshmen students of the new batch. The second objective is to analyze important factors that contribute towards the joining behaviour the most. This problem domain is dynamic as student joining behaviour is affected by multiple factors that include changes in admission policies and trends in the education sector every year. The outcomes of this study can be used by an educational institution to improve its admission processes.

1.5.2 Predicting onset of diabetes using diagnostic measures

’Diabetes’ dataset is a standard dataset contributed originally by the National Institute of Diabetes and Digestive and Kidney diseases (Kaggle, 2016). The objective was to use ML to learn a model that can be used to predict ‘diabetes’ taking different diagnostic measures as input. The diagnostic measures included age of the person, diastolic blood pressure, body mass index, diabetes pedigree function, glucose level, insulin level, number of pregnancies and skin thickness. The target variable ’outcome’ is a binary variable with values as 0 (not diabetes) or 1 (diabetes). In this research work, a framework for collaboration between dataset, machine and human experts is proposed. The use of this dataset is justified as in the medical domain, say of human medical experts is important for the acceptance of an ML-based solution. The dataset consisted of 768 records of female patients of Pima Indian origin. Out of 768 records, 268 are 1(diabetic) and 500 are 0 (not diabetic).

1.5.3 Predicting customer churn

Telco-customer churn dataset referred to as ’tcc’ here onwards, consisted of customer records and is available publicly on Kaggle (Kaggle, 2019). The target variable ‘Churn’ contained a ’Yes’ for the customers who left within the last month and a ’No’ for the retained customers. This data set includes information about the customers like services availed, customer account information(period of association, payment method etc.) and demographic information about customers. The dataset consisted of 7032 observations and 21 variables.

1.6 Organization of the thesis

The rest of the thesis is organized as per the outline given in figure 1.1.

The chapter 2 is dedicated to the review of existing literature in the field of human interpretability. It talks about the strategy used for doing the literature review and the prior formulated research questions aiming to be answered through the literature review. These questions included identifying the motivation behind studying interpretability, desired characteristics of an explanation, broad categorization of the existing work towards conferring interpretability, key ideas employed for explaining outcomes, and evaluation metrics used.

The chapter 3 describes the proposed framework for collaboration between dataset, ML model and human experts. The objective of this framework is to verify the learning acquired by an ML model against the provided dataset and prevailing domain knowledge. The chapter includes motivation for this collaboration, the detailed algorithm, and the framework diagram. The framework has been demonstrated using two datasets, one primary dataset and one standard dataset.

The chapter 4 discusses how the framework proposed in the chapter 3 can be extended to address the situation where the learned ML model lacks in agreement with the prevailing domain knowledge. The proposed approach has been demonstrated using a primary dataset. The algorithm demonstrated that an ML model can be made to adapt itself to incorporate feedback from human domain experts. The chapter details the enhanced algorithm and a framework with the capability of human-feedback adaptive learning. The chapter also identifies desired characteristics of interpretable and interactive ML systems, opportunities in human-feedback adaptive learning and metrics for evaluation of ML systems with human-feedback adaptive learning capability. The chapter also coins the idea of human-capability adaptive interfaces as all human domain experts are not equally competent.

The chapter 5 proposes Feature-importance Feature-Interactions (FIFI) graph, a visualization for interpreting ML models. The chapter also discusses the experimental work to demonstrate the idea of a FIFI graph. The highlight of this visualization is that it is capable of presenting feature importance as well as interactions between features using one visualization, thereby offering a compact alternative. Also, being a graph, it has the advantage that the existing theory of network analysis can be applied. It offers flexibility in terms of filtering out sparse versions in case the user is interested in top features or top interactions only. The chapter also provides a comparative analysis of a FIFI graph and a knowledge graph due to their natural analogy being graph objects.

The chapter 6 presents a summary of the complete research work including conclusions and future research opportunities related to this work. A justification of each future research direction has also been provided to motivate the readers.

Chapter 2

Human Interpretability in Machine Learning Based Solutions – Literature Survey

2.1 Introduction

The goal of this chapter is to review the state-of-the-art in the field of conferring human interpretability to ML-based solutions. The chapter starts with an introduction to the concepts of ML and human interpretability. This introduction is followed by discussing investigation strategies adopted. To keep the review of the literature focused, a set of fundamental research questions were formulated. These included investigating the underlying motivations behind human interpretability, key ideas employed for conferring human interpretability, evaluation metrics, opportunities in human-computer interaction, and identifying future research directions.

ML attempts to train a machine to learn from experience like human beings do. Here, the experience input to a machine is provided in the form of a set of known examples. The objective is that after training the machine should generalize well to predict the outcome for unseen examples. "A machine learns with respect to a particular task T, performance metric P, and type of experience E, if the system reliably improves its performance P at task T, following experience E" (Mitchell, 2006).

ML problems where target variable labels are available along with examples are known as supervised ML problems. Supervised ML problems are broadly categorized into classification and regression problems. In a classification problem, the aim is to predict the class of a categorical variable for a given unseen example. For example, predicting whether a student will pass or fail, whether a tumour is malignant or benign or as simple as whether there will be rain tomorrow or not? In a regression problem, the target variable is of continuous type, for example, predicting the price of a house.

ML is a multi-disciplinary field with computer science and statistics having made the most significant contributions. Speech recognition, computer vision, bio-surveillance, robot control and accelerating empirical sciences are among important applications of ML (Guyon and Elisseeff, 2003). A few recent explorations of ML include advertisement classification of online newspapers (Jain et al., 2021), recommender systems for life insurance (Rani et al., 2021), commercial mobile adhoc networks (Taneja et al., 2021), ML in secure data analytics (Gupta et al., 2020) and cryptocurrency price prediction (Patel et al., 2020).

However, building ML models is not that straightforward as factors like creativity, experience and domain knowledge are critical to their success (Wagstaff, 2012)(Storcheus et al., 2015)(Domingos, 2012).

Human interpretability in ML refers to the ability to explain model outcomes for a particular instance in a way that is understandable to a human user who may not know ML. Models like regression and decision trees are interpretable inherently. For example, looking at the coefficients in a regression equation, one can interpret the sign and magnitude of the contribution of features towards prediction outcome. Similarly, by looking at the level of a node (feature) in a decision tree, one can interpret the relative importance of each feature in the decision making of the ML model. Advanced ML models like random forest and neural networks are lacking in human interpretability. For these models, it is not easy to understand why a particular label outcome is predicted for a given instance. Human interpretability offers advantages like facilitating trust in end users, model debugging and ensuring that the ML solution is unbiased. It is useful in ML-based solutions that affect human subjects and are liable for an explanation to the affected human subjects regarding the decision outcome. Examples include the recidivism problem where an ML-based solution is used for accepting or rejecting a parole application, early detection of diseases in medicine and fraud detection in banking. Human interpretability is crucial in application domains that have the potential to impact a large audience and involve human subjects.

In recent years, conferring human interpretability to ML models has been given importance during top ML conferences like ICML (International Conference on Machine Learning) and NIPS (Neural Information Processing Systems). A few of these events have been listed below:

- ICML 2020 Workshop on Human Interpretability in Machine Learning (WHI), Virtual

- ICML 2019 Workshop on Human In the Loop Learning (HILL), California, USA

- ICML 2018 Workshop on Human Interpretability in Machine Learning (WHI), Stockholm, Sweden
- ICML 2017 Workshop on Human Interpretability in Machine Learning (WHI), Sydney, Australia
- NIPS 2017 Interpretable ML Symposium
- FAT-SG 2017 Workshop on Fairness, Accountability and Transparency in AI and Big Data
- ICML 2016 Workshop on Human Interpretability in Machine Learning (WHI), New York, USA
- NIPS 2016 Workshop on Interpretable Machine Learning in Complex Systems

The lack of human interpretability in ML-based solutions remains a major obstacle in the acceptance of ML-based solutions in these problem domains. Due to its potential advantages, a renewed interest has been observed among the research community in conferring human interpretability to ML-based solutions. The objective of this chapter is to review the work done towards making ML models interpretable to their human users. The following research questions were formulated before the investigation of the state-of-the-art:

Q.1 What are the underlying motivations for human interpretability?

Q.2 What are the characteristics of a good explanation?

Q.3 What is the taxonomy of the approaches proposed?

Q.4 What are the key ideas and explainers used for conferring interpretability?

Q.5 What are the metrics used for evaluating interpretability?

Q.6 Which ML models have been attempted for interpretability?

Q.7 What are the types of problems attempted?

Q.8 Is Interpretability compulsory always?

Q.9 Can human-machine collaboration be useful in incorporating domain expertise?

Q.10 What are the open research directions?

The contribution of this chapter is to summarize the findings of the literature review in the field of human interpretability as answers to the research questions formulated above.

Methodology - Search terms, literature resources and selection criteria

Search terms were constructed using common terms related to human interpretability, alternate spellings, synonyms and keywords from relevant papers. Some of the most relevant search terms included: human interpretability in ML; explainable ML; transparency in ML; understanding the outcome of classification models; understanding black-box models and accuracy-interpretability trade-off. Sources of the search included the reputed ML journals and conferences. A few of the most prominent ones include Journal of Machine Learning Research (JMLR), Pattern Recognition, IEEE transactions, Machine Learning – Springer, ICML, and NIPS.

The research studies covering one or more of the following criteria were shortlisted: (i) proposing an approach for conferring interpretability (ii) accuracy-interpretability trade-off (iii) implementing human interpretability approach on an ML problem (iv) position papers on human interpretability in ML, and (v) human-machine collaboration focusing incorporating prevailing domain knowledge.

2.2 Literature review

The presentation of this section has been organized into model-agnostic approaches, interpreting tree ensembles, interpreting neural networks, human-machine interaction and lack of consensus existing in this evolving field.

Model-agnostic approaches: Model-agnostic approaches are applicable irrespective of the underlying ML algorithm. This independence of the underlying ML algorithm brings advantages like flexibility in the choice of the underlying ML algorithm.

Using fundamental concepts of coalitional game theory (Strumbelj and Kononenko, 2010), a general method for explaining outcomes of any classification model has been proposed. Using the Titanic dataset, the outcomes of different classifiers like naïve Bayes (NB), artificial neural networks (ANN), logistic regression (LR), and support vector machine (SVM) have been explained in terms of the contribution from each feature.

Local explanation vectors have been used to understand an instance-specific prediction outcome of a classifier (Baehrens et al., 2010). These vectors are an estimation of local gradients and characterize how a data point should be moved to change its prediction outcome. The proposed approach was validated using three problem domains: (i) Iris flowers classification in Fisher's data set (ii) Distinguishing 2 from 8 in USPS data set and (iii) a drug discovery problem.

Local Interpretable Model-agnostic Explanations (LIME) algorithm has an underlying assumption that every complex classifier can be approximated by a simpler model at a local scale (Ribeiro et al., 2016*b*). The key idea has been to learn a local interpretable model to explain the prediction outcome of any classifier. The SP-LIME algorithm that uses a set of representing instances, has also been proposed to address the "Trusting a model" problem. This work has already been implemented in Python and R as a package.

Model-agnostic approaches have advantages like model flexibility, explanation flexibility, representation flexibility, low switching cost and ability to compare models (Ribeiro et al., 2016*a*). The associated challenges include: (i) getting a global understanding of the model (ii) inconsistency in local explanations (iii) some domains require exact explanations, and (iv) incorporating more powerful forms of user feedback.

Program snippets have been used as local explanations of any black-box model (Singh et al., 2016). The use of program snippets offers several advantages: (i) capability to capture complex behaviour of black-box systems (ii) each existing interpretable representation can be translated into a program and in fact, can represent their arbitrary combinations (iii) level of detail can be controlled and (iv) research done on software or programs analysis can be used to analyse complex systems being represented by programs.

Local explanations methods like LIME are capable of identifying when a model is right for the wrong reason. However, these models neither scale to the whole dataset nor correct these problems. Input gradients-based explanations are consistent with local explanation methods like LIME. Examining and selectively penalizing input gradients can help regularize differentiable models (Ross et al., 2017). Given annotations about incorrect explanations, alternate explanations were explored to find better reasons to be right. In the absence of annotations, classifiers with different decision boundaries were explored and inspected by human experts. The right-for-the-right reasons approach helps bring robustness to explainable systems despite the presence of adversarial examples.

A model-agnostic explanation system based on if-then rules called 'anchors' has been proposed (Ribeiro et al., 2018). An anchor explanation highlights part of the input that is sufficient to make a prediction. A human user study was conducted to demonstrate how anchors help humans predict the outcome for unseen instances with fewer efforts and higher precision. The limitations of the approach include (i) overly specific anchors for rare classes resulting in increased complexity and lesser coverage (ii) potentially conflicting anchors where anchors with the different outcomes apply to the same instance and (iii) the possibility of a variety of explanations for a given instance.

Influence functions have been used to explain prediction outcomes of black-box models (Koh and Liang, 2017). Influence functions are based on the core idea of studying models through the lens of their training data. The idea is to trace back a prediction outcome through the underlying ML algorithm to identify the most responsible training points behind that prediction. The implementation of this approach required only oracle access to gradients and Hessian-vector products. Even in the case of non-differentiable and non-convex models, an approximation to influence functions using second-order optimization techniques can still provide valuable information. This work demonstrated a variety of applications including training-set attacks, debugging models and fixing datasets.

High-level functional programs, which capture the invariant structure in the observed data have been proposed to represent abstract models (Penkov and Ramamoorthy, 2017). The authors propose a program-induction machine, an architecture capable of inducing interpretable explanations in the form of LISP-like programs from observed transition system data traces. The optimization procedure for program learning is based on backpropagation, gradient descent and A* search. This work allows the users to specify the properties and context in which data is to be explained by accepting a set of predicates of interest from the user.

A crucial property that interpretability methods should satisfy is the robustness of explanations to local perturbations. Such a requirement states that similar inputs should give rise to similar explanations (Alvarez-Melis and Jaakkola, 2018). The authors have formalized the intuitive notion of robustness and investigated popular gradient and perturbation-based interpretability methods in terms of robustness of explanations. Experimental results showed that model-agnostic perturbation methods are more prone to instability in comparison to their gradient-based counterparts. The authors also proposed ways to enforce robustness in interpretability methods.

A novel method, ASTRID (Automatic STRucture IDentification), has been proposed for investigating which feature interactions are exploited by the classifier for making predictions (Henelius et al., 2017). Attributes are said to interact whenever they are conditionally dependent given the class. Knowledge of such interactions has applications in the field of pharmacovigilance and bioinformatics. This paper studies two problems: (i) To determine if a particular grouping of attributes represents attribute interactions structure in a given dataset and (ii) to automatically find a grouping of attributes in the given dataset with maximum cardinality. ASTRID found this grouping in polynomial time and did not make any assumption on the data distribution or classifier thereby increasing its applicability.

Explaining outcomes of complex models in terms of feature importance becomes infeasible in the case of a high dimensional dataset without restraining the number of features. A model-agnostic approach to limit the length of explanations using contrastive explanations has been proposed (van der Waa et al., 2018). This approach utilizes the human tendency of asking questions like “Why this output (the fact) instead of that output (the foil)?” to shortlist features playing an important role. An arbitrary model was trained to distinguish between the fact and the foil. From this model, two sets of rules, one for identifying data point as a fact and another for identifying data point as a foil, were distilled. The factual rule set was subtracted from the foil rule set to obtain a relative complement of the fact rules. This relative complement was used to construct contrastive explanations. The proposed approach constructed explanations that were shorter than the full feature list, provided more information on the contribution of features and this contribution matched the underlying model more closely.

Assessment of algorithmic fairness is among the several advantages of interpretability. This assessment is affected by explanation styles as certain explanations are inherently less fair while some can enhance confidence in the algorithmic fairness (Dodge et al., 2019). Assessment of fairness is also affected by whether it is model-wide or instance-specific. There is a need for personalized and adaptive explanations to support the assessment of fairness.

Interpreting tree ensembles: Tree ensembles utilize a large number of decision trees to improve predictive performance. The outcome is computed using the outcomes of individual trees. This increased predictive performance comes at the cost of interpretability as compared to that using a single decision tree.

To avoid compromise on accuracy or interpretability in the case of tree ensembles, using a pair of models, P and I, respectively for prediction and interpretation (Hara and Hayashi, 2016), has been proposed. The tree ensemble has been approximated by an interpretable model using the Expectation-Maximization (EM) algorithm that attempts to minimize Kullback-Leibler (KL) divergence from the complex tree.

A framework named ‘inTrees’ that can extract, process, prune, summarize rules and discover frequent variable interactions from a tree ensemble has been proposed (Deng, 2019). One limitation is that its current version in ‘R’ handles trees with binary splits only, whereas, the algorithms apply to trees with multiple splits also.

GENESIM a genetic algorithm that learns a single interpretable decision tree after starting with an initial population of decision trees has been proposed (Vandewiele et al., 2016).

Finding prototypes in tree space has been proposed as a new approach for interpreting tree ensembles (Tan et al., 2016). Prototypes are “representative” observations. This approach benefits from leveraging tree structure and naturally learned similarity measures. The idea has been to focus on sparsity-in-observations instead of sparsity-in-features.

Explaining tree ensemble methods usually rely on feature attribution. Existing feature attribution methods can assign higher importance to features with a lower impact on the model's behaviour (Lundberg and Lee, 2017). This problem has been addressed by developing fast exact solutions for SHAP (Shapley Additive explanations) values using conditional expectations which are consistent as well as locally accurate. XGBoost has been used to demonstrate the inconsistencies of current methods and improved clustering performance using SHAP values. The complexity of computing SHAP values has also been reduced from exponential to $O(TLD2)$, where, T, L and D respectively represents the number of trees, maximum number of leaves in any tree and the maximum depth of any tree.

The importance of defining the right locality, while constructing local surrogate models has been demonstrated (Laugel et al., 2018). Defining the locality right has a major impact on the relevance and quality of the approximation of the local decision boundary of a black-box model and thus on the quality of the explanation for an individual instance. This work proposed a sampling strategy centred on a particular place of the decision boundary, relevant for a prediction to be explained, rather than on the prediction itself. The intuition behind the approach has been that as a local surrogate aim at approximating local black-box decision boundary, this boundary should be sought first to sample instances in the neighbourhood of the instance. The experimental results demonstrated improvements in the local fidelity of the surrogate both for synthetic and UCI repository datasets.

In contexts, where multiple distinct but accurate models for some datasets exist for classification, current ML methods are likely to recover a complex model that combines them (Ross et al., 2018). While individual models may be more or less interpretable, their combinations are likely to be harder for humans to understand. The authors introduced a way to identify a maximal distinct set but accurate models for a dataset through optimizing orthogonal gradients during training. The proposed approach has been demonstrated empirically to recover simpler more interpretable classifiers instead of complex ones. The novelty of the work lay in the use of independence-based proxy, focus on maximal sets, and training methods.

Interpreting neural networks: Artificial Neural Networks (ANNs) are capable of providing accurate solutions in complex problem domains. However, they do not fare well on the interpretability front.

Partial derivative or input gradient of the model with respect to input has been used for interpreting any model, both for classification as well as regression problems (Hechtlinger, 2016). If a partial derivative with respect to input was zero, it indicated that this input did not affect the prediction function. If the same is non-zero, that input was considered as important to the model prediction.

Layer-wise relevance propagation (LRP) framework has been used to explain predictions of a deep neural network by the decomposing outcome in terms of the input variables (Samek et al., 2016). In the context of an image classification problem, LRP operates by building a local redistribution rule for each neuron. These rules are applied in a backward pass to obtain pixel-wise decomposition. LRP is useful in analysing differences in DNN architectures.

A novel technique called deep-Taylor decomposition that decomposes decisions of a neural network as a relative contribution from features has been proposed (Montavon et al., 2017). Existing approaches for explaining decisions of complex non-linear models have been categorized as (i) functional approaches involving a local investigation of the prediction function or (ii) message passing approaches viewing approximation as an outcome of a computational graph and explanation is produced using a backward pass on this graph. This work has reconciled these two approaches in the context of deep neural networks.

Deep neural networks are opaque and there is a need to explain neural computation. The explanation problem itself has been modelled as a learning problem (Rosenbaum et al., 2017). The authors present a new dataset and user simulator, e-QRAQ (explainable – Query, Reason and Answer Question), to test an agent's capability to read a short ambiguous story and answer a challenge questions associated with the given story. The story is made ambiguous by replacing some of the entities with variables. The authors demonstrated e-QRAQ by training a new neural architecture based on end-to-end memory networks to generate predictions as well as partial explanations. A strong correlation was observed between the quality of the prediction and the explanation.

Sensitivity Analysis (SA) and Layer-wise Relevance Propagation (LRP) have been evaluated for explaining prediction outcomes of a deep neural network (Samek et al., 2017). Dataset used included image classification (ILSVRC2012 dataset), classification of text documents (20Newsgroup dataset), and recognition of human actions in videos (HMDB51 dataset).

Deep neural networks have shown success in processing images and audio data whereas tree-based models have been popular for processing tabular data. Deep Neural Decision Tree (DNDT), an intersection of these two approaches, (Yang et al., 2018) is a neural network, where each set of its weights corresponds to a specific decision tree and is interpretable. In a DNDT, parameters are simultaneously optimized using stochastic gradient descent instead of a greedy splitting procedure. A DNDT self-prunes at split as well as feature-level. The experimental work demonstrated better results than neural networks for several tabular datasets while providing interpretability.

The automated discovery of new or unusual information in large datasets is a key to scientific progress. Many image data analysis systems are making use of CNNs (Convolutional Neural Networks). However, most of the existing anomaly and novelty detection methods are difficult to interpret. An approach to generate human-comprehensible explanations for novel discoveries in large image datasets has been proposed (Wagstaff and Lee, 2018). Novelty detection is combined with CNN image features to achieve rapid discovery and interpretable explanations. The proposed approach has been demonstrated using the ImageNet images dataset. Experimental results implied that in the case of novel image detection, explainable results in tandem with best discovery performance are obtainable.

Textual explanations using natural language has been used to explain a decision made by a deep neural network. Current textual explanations discuss class discriminative features in an image. Counterfactual explanations which reason about information that is not present in an image, but might impact the decision if it were available, has been proposed (Hendricks et al., 2018). The intuitive idea behind counterfactual explanations is to identify the evidence which is discriminative for one class, but not present in the other class. Experimental results demonstrated good counterfactual explanations both qualitatively and quantitatively.

Image classifiers operate on low-level features rather than high-level concepts. Concept Activation Vectors(CAVs) which represents the internal state of a neural network in terms of high-level human-friendly concepts has been proposed (Kim et al., 2018). The idea has been to quantify the importance of a user-defined concept to a classification outcome. Image classification has been used as a testing ground to demonstrate the approach. Contributions of this work include (i) human-friendly linear interpretation and (ii) the use of natural high-level concepts in explaining model decision-making.

Visual Analytics: This approach is based on using visualizations for modelling input-output relationships of an ML classifier, to achieve interpretability (Krause et al., 2016). This work has talked about the role that visual analytics can play in interpretability. Making the outcomes of the black-box model more expressive instead of trying to make the underlying complex model transparent has been attempted (Villagrá-Arnedo et al., 2017). The idea has been to provide more information than mere classification like interpreting progress, trends in learning and identifying causes for learning problems, in the context of student performance. Expressive representation tools like progression charts and heat maps have been used.

Addressing lack of consensus in this evolving field: As the field of human interpretability is still evolving, there is a lack of consensus among the ML community regarding the formal definitions and evaluation metrics for interpretability.

An attempt has been made to give more specific definitions of interpretability (Lipton, 2016). This study aims to answer questions like: What are the underlying motivations for human interpretability in ML? What model properties and techniques confer interpretability to ML models? The common notion that linear classifiers are interpretable whereas deep neural networks are not, has also been questioned.

The issue of lack of consensus related to definition and measurability of interpretability has been addressed (Bibal and Frénay, 2016). Different terms being used for interpretability has been categorized into (i) synonyms for interpretability and (ii) the ones that depend on interpretability but are related to distinct problems e.g. acceptability, justifiability. Approaches to measuring interpretability have been categorized as heuristics and user-based surveys.

The need to distinguish between types of transparency and identify settings, where transparency may be harmful, has been addressed (Weller, 2017). Motivations and benefits are different for different contexts as well as different stakeholders i.e. developer, deployer, user and society. There is a need to (i) Define criteria and tests for faithfulness of an explanation (ii) Metrics for comparing the performance of explainers (iii) Develop outlines to understand what kinds of transparency are supportive and contexts in which it can be harmful.

Data-driven ways have been suggested to derive operational descriptions and assessments of explanations to address the issue of lack of consensus in defining and evaluating interpretability (Doshi-Velez and Kim, 2017). The questions raised include: Are all so-called interpretable models like decision trees, falling rule lists etc. equally interpretable? Do all applications have the same interpretability needs? The authors have suggested application-grounded, human-grounded and functionally-grounded as approaches for interpretability evaluation.

Lack of clarity in existing definitions of interpretability and its cognates has been addressed (Krishnan, 2019). The authors have advocated that the notion of interpretability is over-hyped and may not be the only solution to the broader goals of interpretability. The research community should talk more about end goals rather than interpretability which might be one tool among a possible set of solutions. Strategies like reviewing the construction of training set and strategies for testing of classifiers have been talked about to achieve many of the goals of interpretability without actually knowing the inner working of ML algorithms.

Transparency is often advocated to facilitate trust in ML-based solutions and their successive deployment in real-life situations. Motivations and benefits of transparency vary for different types of transparency and in different contexts (Weller, 2019). The author highlights and review settings where transparency may be harmful. Also, sometimes, transparency is a means to an end and not the end in itself. Also, transparency can be used by actors with misaligned interests as a channel for manipulation or inappropriately use information gained due to the transparency of an ML model.

Explanations produced by existing tools are neither standardized nor systematically assessed. There is a need for defining explainability to create best practices and identify existing open challenges. To trust black-box models, there is a need for explainability in addition to interpretability(Gilpin et al., 2018). Explainability implies interpretability but the reverse may not be always true. Most of the work focussing on explaining deep networks involves evaluating explanations for (i) completeness as compared to the original model (ii) Completeness as measured on a substitute task (iii) ability to detect biases in models and (iv) Human evaluation for reasonableness.

Human-Machine Interaction: As ML is moving out from research labs to the hands of the end-users, it is becoming important that future ML systems should be capable of incorporating feedback from human domain experts. For this to happen, end-users should be able to interpret the ML model as well as revert back with their experience-based feedback regarding the problem domain.

Well-designed explanation interfaces have the potential to facilitate the trust-building of end-users. A roadmap identifying important design issues worth investigation and a set of general principles for the construction of explanation interfaces in the context of trust-inducing interfaces has been provided (Pu and Chen, 2006).

The democratization of ML by providing a bigger role to end-users in the design of ML systems has been advocated (Amershi et al., 2014). Involving users in the learning process results in rapid and specific incremental updates in the ML model. This involvement brings the challenge of understanding the capabilities, behaviour and needs of the end-user. There are several open research challenges like common language different fields, evaluation of interactive ML systems, establishing principles and guidelines, leveraging the masses and addressing algorithmic problems (time-space trade-off).

Human-in-the-loop ML that enables interaction of human experts with ML systems to incorporate domain knowledge, has been attempted (Kim, 2015). An interpretable model iBCM to support human decision-making has been developed. This research developed a model that allows transparent interaction with humans who may not be ML experts. The model is demonstrated in the education domain for streamlining the grading process.

A novel notion of interpretability has been provided looking beyond human understanding of model outcome (Dhurandhar et al., 2017). Interpretability is not an absolute concept and must be considered relative to a target model that may be a human in the most obvious setting. This work proposes a framework to compare interpretable procedures on practical aspects like robustness and accuracy. As a future direction, all humans may not be equal relative to a task as expertise in a domain may increase the level of detail consumable by that human.

An approach to efficiently optimize human interpretable models by directly involving human subjects in the optimization loop has been proposed (Lage et al., 2018). The idea is to develop a diverse collection of models that can explain data well. From this collection, pick models that are highly human interpretable. As the evaluation of an interpretable model involves human user studies, it helps to reduce the number of user studies and make it cost-effective. The algorithms used included decision trees and neural networks. The Proxies used included average path length, the average number of distinct features in a path, number of nodes, and number of non-zero features.

Overall decision-making in high-stakes settings is an ML model communicating with a human who makes the final decision. Therefore, the relevant performance criteria should be for the entire system and not only for the ML component. The performance of such two-node tandem data fusion systems architecture has been characterized using the theory of distributed detection (Varshney et al., 2018). The ML model node transmits its observation to the human node. The final decision is produced after fusing independent local observation of the human node with information received from the model node. The experimental results have shown the overall system of a human combined with an interpretable classifier outperformed the systems with a black-box classifier only. This approach is termed a hybrid of human and ML models.

A hybrid human-machine intelligence approach has been proposed where initially a white-box ML model is learned (Yang et al., 2019). The model is presented to human experts to get their feedback on the learned model. The model is refined to incorporate the feedback given by human experts. This iterative cycle goes on until an equivalent ML model is reached that is as competitive as a black-box model in terms of internal optimization metrics. The internal metrics used in this case were Precision, Recall and F1-score. A user interface called 'Ruleslearner' has been used to present the learned model as a set of rules. These rules are in disjunctive normal form. Using this interface, feedback from users is collected in the form of addition, deletion, modification, ranking or filtering of rules.

The involvement of humans in the exploration of solution space can positively affect computationally hard problems(Holzinger et al., 2019). This involvement can address the disadvantages of auto-ML by reducing resource consumption, engineering efforts required, and the amount of training data required. For demonstration, the ACO (Ant colony optimization) algorithm was used for solving a TSP (Travelling Salesman Problem). The TSP problem was modelled as of finding the optimal path on a weighted connected graph. The key idea has been to increase the probability of selection of a human-traversed path, by artificial agents also.

Interdisciplinary expertise has been advocated to develop explanatory learning models that can provide interpretable and achievable insights in addition to accurate predictions (Rosé et al., 2019). The interdisciplinary expertise spans AI/ML engineers, cognitive, education and UI/UX designers. The proposed approach demonstrated (i) improvement in student learning (ii) better pedagogical practices and (iii) advancement of learning science.

Setting up appropriate expectations from AI systems is important as they vary among end-users. The impact of setting up expectations has been demonstrated using two versions of an AI-based scheduling assistant, with similar accuracy but different focus (avoiding False Positives Vs False Negatives). Expectation adjustment techniques prepare users for imperfections in AI systems and improve user acceptance of such systems (Kocielnik et al., 2019).

The impact of combinations of explanations and feedback on users' perception of ML systems has been investigated (Smith-Renner et al., 2020). For the low-quality model, explanations without the feedback opportunity lead to a negative user experience. For the high-quality model, requesting feature-level feedback without explanation reduced trust.

Results showed that relational style leads to favourable perceptions as compared to functional style. Technologies that offer high levels of automation and human control have the potential to improve human performance and thus leading to wider adoption of these technologies. The human-Centered Artificial Intelligence (HCAI) framework aims to explain how to design for a high level of automation and human control, deciding when to provide full control to humans and when to computer, and how to avoid pitfalls associated with giving full control to humans or computer (Shneiderman, 2020).

Research on AI ethics should also focus on everyday interactions with personal technology as AI can commit moral wrongs (Shank and Gott, 2020). The findings reported that two types of exposure lead to such wrongs most frequently: (i) exposing personal information (31%) due to information sharing between devices and (ii) exposing people (especially children) to undesirable content (20%) due to their proximity to audio devices.

With the increase in demand for online education, the use of ML or AI-based teaching assistants is an upcoming area for improving the learning experience of students. Using an online survey, the perception of students regarding using machine teachers or AI assistants in higher education was investigated (Kim et al., 2020). The findings of the survey indicated that usefulness and ease of communication with these AI assistants were key to the adoption of such assistants. The authors examined the impact of communication style (relational vs functional) of machine teachers on students' perception about AI-based education (Kim et al., 2021).

Other related work: Falling rule lists (FRL) are classification models consisting of a set of rules where (i) the order of rules decides which example to be classified by each rule and (ii) the probability of success decreases monotonically downward. A Bayesian framework to learn an FRL has been proposed (Wang and Rudin, 2015). FRLs are particularly useful in healthcare applications where the patients are stratified into risk-sets with a different priority.

Supersparse Linear Integer Model (SLIM) has been used to produce an accurate and interpretable model for recidivism prediction (Zeng et al., 2015). Running a SLIM was modelled as an integer programming problem. It has been found that traditional methods like ridge regression perform equally well as modern methods like Stochastic Gradient Boosting.

An interpretable predictive model by integrating data-driven model and domain knowledge has been proposed (Jovanovic et al., 2016). Hierarchy of diseases ICD-9-CM (International Classification of Diseases 9th - revision clinical modifications) has been taken as the domain knowledge. In this work, a way to quantify interpretability using information entropy and not just the count of features has also been proposed. Tree-Lasso regularized regression model was found competitive in accuracy while being more interpretable also.

Making interpretable models using the human mind's construction of concepts and meaning has been proposed (Condry, 2016). Using a relational model for meaning, the concepts can be classified as form (symbol) or functions (meaning), given a context.

The problem of interpreting ML models built from high dimensional and sparse data has also been attempted (Moeyersoms et al., 2016). The basic idea has been to list important features using "Explanation curves". It evaluates in terms of how many examples have been explained for a given number of features listed. Online browsing records for predicting product interest was used for empirical evaluation.

Conferring interpretability to Restricted Boltzmann Machines (RBM), a commonly used algorithm for collaborative filtering in recommender systems, has been attempted using explainable RBM technique (Abdollahi and Nasraoui, 2016). This technique picks top-n recommendations using an explainability score for each item and each user. MovieLens rating data has been used for testing the approach.

A two-stage approach to increase the interpretability of a Computer-aided Diagnosis (CAD) system has been proposed (Gallego-Ortiz and Martel, 2016). The first stage extracted features that reflect simple and interpretable characteristics of lesions. The second stage summarized CAD results in the form of rules that can be explained in terms of lesions characteristics. The problem of classification of breast MRIs as cancerous or non-cancerous was taken for testing the proposed approach.

The Bayesian framework has been used to learn a classifier that consists of a small set of rules in the disjunctive normal form to achieve interpretability (Wang et al., 2017). The basic idea has been that in the space of good predictive models, a very sparse but accurate rule set may exist. The problem was to find that rule set in a computationally effective manner so that it can be used in practical settings.

Using the technique called "cloaking", social-networking site users can inhibit inferences being drawn from their activities as a matter of privacy (Chen et al., 2017). This work demonstrated that users have to cloak a small portion of their likes to inhibit any inferences being drawn about them. Also, the targeted users for posting advertisements can seek an explanation of why are they being targeted for this particular advertisement.

There exist legitimate concerns about the intentional and unintentional consequences of AI systems. The authors advocated the use of explanations to make AI systems accountable (Doshi-Velez et al., 2017). The authors reviewed current societal, moral and legal norms around explanations and different contexts that required explanations under the law. This work listed the technical considerations if AI systems are to produce kinds of explanations currently required of humans under the law. Also, there is a need to think about why and when explanations are useful enough to outweigh their cost.

For a given input, output and explanation, a human is expected to verify is the output as per the input and expected rationale. Variation in factors like size of explanation, creating new types of cognitive chunks, and repeated terms in an explanation have been analysed to identify what factors of an explanation, affect its human-interpretability in the context of verification and what factors are relatively insensitive (Narayanan et al., 2018). Variation in the form of the domain was also evaluated. Any increase in explanation complexity increased the response time of humans for verification tasks and decreases subjective satisfaction with the explanation. Embedding a new concept improved response time as compared to creating a new definition. New concepts and the number of lines increased response time more as compared to the repetition of concepts or longer lines.

The interest in interpretability has been extended to active learning, where the explanation is of interest added to the labeller apart from the receiver of the decision and creator of the model (Phillips et al., 2018). This work was an extension of the work on LIME to interpret what specific trends and patterns an active learning strategy may be exploring and how LIME can be used to generate locally faithful explanations for an active learning strategy. These explanations help understand how different models and datasets explore a problem space over time. The notion of uncertainty bias has been introduced to quantify per-subgroup differences in terms of how active learning queries spatial regions.

Arguing that a majority of the existing feature contribution-based explanation methods lack a formal mathematical definition, a novel definition of the feature score has been proposed (Hara et al., 2018). The idea has been to find the maximally invariant data perturbation that does not change the model's output. The features that allow small data perturbations are considered as the important ones. The intuition behind this is that if a feature is relevant to the output, it cannot change a lot to keep the output changed. The problem of finding maximal data perturbation has been formulated as semi-infinite programming and approximated as a linear programming problem using the first-order

Taylor expansion. The proposed approach has been demonstrated on image classification with VGG16 and was able to identify important parts of the images effectively.

It has been argued that the interpretability of an ML system should be defined concerning a specific task or agent. To answer this question, a model identifying different roles that agents can fulfil concerning the ML system has been described (Tomsett et al., 2018). The model has been illustrated in different scenarios like how the role of an agent influences its goals and the implications for defining interpretability. This model can be useful for the research community working on interpretability, developers involved in such systems, and people involved in the auditing of ML systems. This approach can help system creators-owners identify interpretability needs of different agents and specify the level of access to explanations based on the role of an agent.

Open Learner Modelling (OLM) provide tools for supporting human learning and teaching by developing models of learner's cognition and emotions (Conati et al., 2018). The authors summarized how Artificial Intelligence can be used in education through the OLM research and what are the necessary considerations for making it interpretable and explainable. This work provided a starting point for an interpretable AI in education. OLM are student models that allow users to access their content with varying levels of interactivity. Key considerations for OLMs include identifying: (i) What is the need of building the OLM (ii) Which aspects of the model are to be made available to the user? (iii) How is the model accessed and the degree to which it can be manipulated? and (iv) Who all have access to the model?

Knowledge bases are now a reality owing to their success in natural language processing and semantic web search. Knowledge bases suffer from incompleteness and rely on mapping entities and relations into a low-dimensional vector space via embedded models. Embedded models are accurate but hard to interpret. Two pedagogical methods have been proposed to interpret embedding models associated with large knowledge bases(Gusmao et al., 2018). The idea has been to extract human interpretable weighted horn rules from embedding models.

Sharing understanding of an ML model's decision-making behaviour can be risky from the privacy aspect of the training data (Shokri et al., 2019). This work initiated a new research direction termed as 'Privacy – Quality Explanation trade-off'. The objective was to explore whether an adversary inferred private information regarding the training data by leveraging explanations. Privacy risks of feature-based model explanations have been investigated using membership-inference attacks. The idea was to find out how much information about the presence of a training data point gets leaked through model predictions plus explanations. Results showed that back-propagation based explanations are prone to such information leak due to revealing statistical information regarding the model's decision boundaries about an input.

2.3 Discussion on literature review

This section attempts to summarize the findings from the literature review. The existing work can be broadly categorized as model-agnostic approaches, interpreting tree ensembles, interpreting neural networks, visual analytics and human-machine interaction approaches. A few studies are addressing the lack of consensus regarding the definition of interpretability and its evaluation. The findings have been presented as per the above categorization. For each category, key ideas have been summarized and existing research gaps have been identified. Figure 2.1 presents a broad categorization of the state-of-the-art. Table 2.1 summarizes the key ideas, explainers and metrics used.

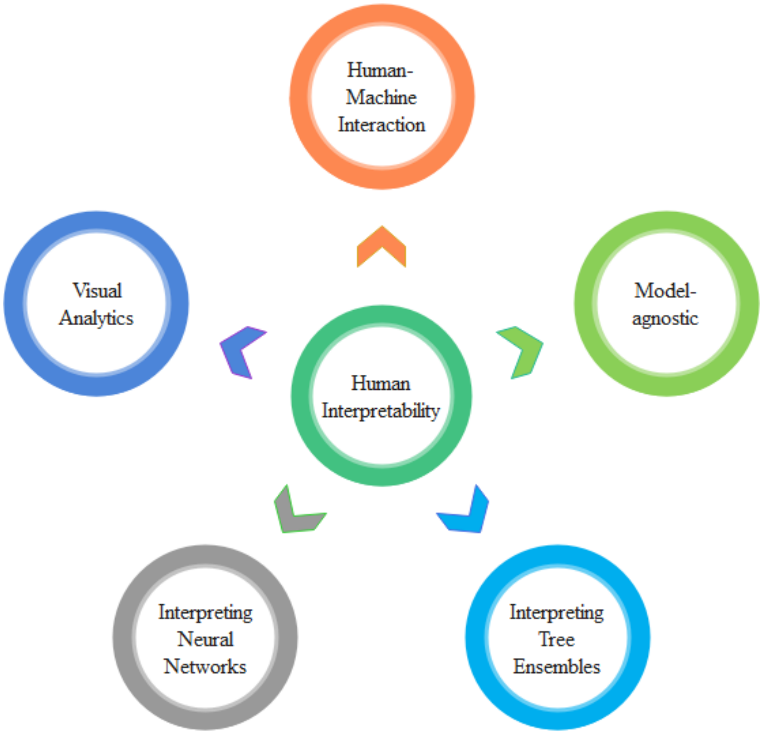

Human-Machine Interaction
Visual Analytics
Model-agnostic
Human Interpretability
Interpreting Neural Networks
Interpreting Tree Ensembles

Figure 2.1: A broad categorization of the existing work

Model-agnostic approaches: Concepts of coalitional game theory have been used for explaining an ML model outcome in terms of the relative contribution of input features (Strumbelj and Kononenko, 2010). Local explanation vector has been used to explain an individual outcome (Baehrens et al., 2010). The algorithm LIME learns an interpretable model locally around the instance to explain its prediction outcome (Ribeiro et al., 2016*b*). Model agnostic approaches have potential advantages of model flexibility and low switching cost (Ribeiro et al., 2016*a*). Program-snippets have also been used as model-agnostic local explanations (Singh et al., 2016). Their advantages include the ability to capture complex behaviour, the existence of a corresponding program for each existing interpretable model, control on the level of detail in the program and using existing research in program or software analysis. Research gaps associated with these approaches include: discovering important individual interactions (Strumbelj and Kononenko, 2010); computational optimization(Strumbelj and Kononenko, 2010)(Ribeiro et al., 2016*b*); extension to regression tasks (Strumbelj and Kononenko, 2010)(Baehrens et al., 2010); exploring alternatives of a sparse linear model as explainer(Ribeiro et al., 2016*b*); addressing challenges associated with model-agnosticism like global explanations, inconsistency among local explanations, more powerful forms of user feedback (Ribeiro et al., 2016*a*); investigating methods for inducing programs, including more expressive syntax in programs, exploring recently introduced differentiable program induction technique, and evaluating the approach in other real-world problems (Singh et al., 2016).

Interpreting tree ensembles: Using a pair of models, ‘P’ and ‘I’, respectively for prediction and interpretation, has been demonstrated, where ‘I’ is an interpretable approximation of prediction model ‘P’ (Hara and Hayashi, 2016). A framework ‘inTrees’ demonstrated extracting, processing, pruning, summarizing and deciding on rules from an ensemble of trees (Deng, 2019). Genetic algorithms have been used for extracting an interpretable model starting with an initial population of decision trees (Vandewiele et al., 2016). Prototypes in tree space have been found using the two unique aspects of tree ensembles, tree structure and naturally-learned similarity measure (Tan et al., 2016). Associated research gaps include: evaluating alternatives as proximity measure between ‘P’ and ‘I’, and automatic calculation of the number of regions ‘K’ (Hara and Hayashi, 2016); improving ‘inTrees’ package to handle trees with more than two splits (Deng, 2019); reducing computational complexity (Vandewiele et al., 2016); using human experts to check the quality of prototypes, investigating other prototype-finding procedures, evaluating on datasets having confusing objects or higher-dimensions and optimization (Tan et al., 2016).

Interpreting neural networks: Approaches used for interpreting neural networks model include input gradient, sensitivity analysis, layer-wise relevance propagation (LRP) and deep Taylor decomposition. Input gradient was obtained using back-propagation and chain rule, in case of a neural network (Hechtlinger, 2016). LRP is useful in explaining predictions of a deep neural network (DNN) and analyzing differences in DNN architectures (Samek et al., 2016). Sensitivity Analysis (SA) and LRP are useful in explaining the outcomes of a deep neural network (Samek et al., 2017). Deep Taylor decomposition (Montavon et al., 2017) reconciles functional and message passing approaches, to attribute the outcome of a neural network in terms of influence from input features. There is a need for evaluating these approaches in problem domains other than image classification(Montavon et al., 2017).

Visual Analytics: This approach has aimed to keep the underlying model as black-box so that we do not gain on interpretability at the cost of accuracy due to the interpretability-accuracy trade-off. The idea is representing input-output relationships using visual analytics to bring interpretability (Krause et al., 2016). Another idea has been to focus on making the outcome more expressive rather than just classifying using tools like heat maps, progression charts, and confidence level in the prediction (Villagrá-Arnedo et al., 2017). Research gaps associated with visual analytics include: identifying metrics to compare interpretability through visual analytics (Krause et al., 2016); finding new graphical tools to improve expressiveness, using this expressive information to guide interventions, exploring in domains other than education (Villagrá-Arnedo et al., 2017).

Lack of consensus regarding interpretability concepts: A lack of consensus has been reported in defining interpretability, its types and ways of measuring. Many terms being used are related to interpretability but are distinct from interpretability (Bibal and Frénay, 2016). Transparency is of different types and may not be useful in every setting. Different stakeholders (developer, deployer and user) have different benefits or motivations (Weller, 2017). Approaches for evaluation include (i) real humans, real tasks (ii) real humans, simplified tasks (iii) No humans, proxy tasks (Doshi-Velez and Kim, 2017). The notion of interpretability is over-hyped and may not be the only solution to the broader goals of interpretability like public trust, non-discrimination, and causal explanations. Strategies like the construction of training sets and testing of classifiers are also promising towards understanding the outcomes of an ML classifier (Krishnan, 2019).

Associated research gaps and open research questions include: distinguishing interpretability measures of models and representations, linking results of user-based surveys and heuristics by translating former to later (Bibal and Frénay, 2016); criteria for the faithfulness of an explanation, framework to understand what transparency is useful and what is harmful; metrics to compare two explainers (Weller, 2017), creating links between evaluation approaches, what proxies for what applications, important factors while designing simplified tasks, factors while characterizing proxies for explanation quality (Doshi-Velez and Kim, 2017).

Human-Machine interaction: To incorporate feedback from human domain experts regarding the behaviour of ML systems, the behaviour of the ML system must be interpretable. Moreover, human experts must be enabled to provide feedback to the ML system based on their domain knowledge. Bi-directional communication between machine and human in terms of prototypes and subspaces has been demonstrated using an interactive ML model 'iBCM' for clustering (Kim, 2015). A hybrid human-machine intelligence approach where the model is refined to incorporate the feedback given by human experts in the form of modifications in the rules explaining the ML model has been demonstrated (Yang et al., 2019). The involvement of humans in the exploration of solution space has been demonstrated to positively affect computationally hard problems (Holzinger et al., 2019). Through the democratization of ML, end users can be provided with a bigger role in the design of ML systems (Amershi et al., 2014). Interdisciplinary expertise spanning AI/ML engineers, cognitive, education and UI/UX designers have been useful in developing explanatory learning models that can provide interpretable and achievable insights in addition to accurate predictions (Rosé et al., 2019). Human interpretability can be made cost-effective by reducing the number of human user studies by involving human experts in the learning loop (Lage et al., 2018).

Other related studies: There are several other approaches proposed for conferring interpretability. A falling rule list (FRL), an ordered list of rules, has been extracted using a Bayesian framework (Wang and Rudin, 2015). Integration of data-driven model and domain knowledge (hierarchy of diseases ICD-9-CM) using Tree-Lasso regularized regression has been done for predicting pediatric hospital readmission (Jovanovic et al., 2016). Interpretability can be improved by utilizing the conceptual structure of meaning, focusing on the psychology of human learning of concepts (Condry, 2016). A classifier comprising a small set of rules in disjunctive normal form can be learned using a Bayesian framework. Also provides user-adjustable priors for desired shape and size and domain-specific interpretability (Wang et al., 2017). A two-stage process has been proposed for the diagnosis of breast MRI (Gallego-Ortiz and Martel, 2016), first extracting simple and interpretable features and then summarizing results in the form of rules. Research gaps include: exploring the integration of other sources of domain knowledge hierarchies, incorporating additional patient follow-up data after discharge to improve the model (Jovanovic et al., 2016); evaluating proposals to improve interpretability like clearly defined requirements for input and goals of the output, minimize the number of attributes, models that complement and expedite existing processes, preparing a front-end for the intended audience in terms of concepts known to them (Condry, 2016).

Table 2.1: Summary of key ideas, explainers and evaluation metrics used.

Key Idea	Explainer	Evaluation Metrics
Coalitional game theory (Strumbelj and Kononenko, 2010)	Outcome explained in terms of feature-wise contribution	Captivating both the degree as well as the sign of the influence.
Local explanation vector to explain an instance-specific outcome (Baehrens et al., 2010)	Scatter plots of the explanation vectors (local gradients)	Ease of characterizing how a data point is to be moved to change its label.
Using two models, P and I, Where I is an interpretable approximation of P (Hara and Hayashi, 2016)	Simplified decision Tree	Number of regions(nodes) in the simplified tree
Every complex model can be approximated at a local scale (Ribeiro et al., 2016*b*)	Sparse linear model	Human subjects on Amazon Mechanical Turk provided with explanations to evaluate the following: (i) Can users choose the best classifier among the two (ii) improving an untrustworthy classifier and (iii) Can explanations lead to insights?
Listing important features in case of high-dimensional and sparse data (Moeyersoms et al., 2016)	Explanation curves	Computing feature-coverage (how many outcomes are explained using only top-ranked features)

Continued on next page

Table 2.1 – continued from previous page

Key Idea	Explainer	Evaluation Metrics
Visual analytics (Krause et al., 2016)	Graphical tools	Uses the power of visual perception in humans. Visualizing model behaviour by looking exclusively at the relationship between input and output.
Pick top 'n' recommendations that are interpretable (Abdollahi and Nasraoui, 2016)	User-based neighbor-style explanations	Explainability score for user-based neighbour-style explanations, MEP (Mean Explainability Precision) and MER(Mean Explainability Recall)
A two-stage approach: 1. Extract interpretable features 2. Summarize results in form of rules (Gallego-Ortiz and Martel, 2016)	Only the most relevant nodes of random forest are highlighted. The size of the node is also proportionate to their weight.	Degree of agreement with the radiologist
Learning an FRL using bayesian framework (Wang and Rudin, 2015)	FRL (Set of Rules with ordering)	Falling rule lists with "Probability" and "Support" for each rule
Integrating data-driven model and domain knowledge (Jovanovic et al., 2016)	Plotting information loss	Quantifying interpretability using information entropy instead of just a count of features

Continued on next page

Table 2.1 – continued from previous page

Key Idea	Explainer	Evaluation Metrics
Extracting a single interpretable tree from a population of trees using Genetic Algorithm (Vandewiele et al., 2016)	Interpretable tree	Number of nodes
Making outcomes of a black-box model more expressive (Villagrá-Arnedo et al., 2017)	Representation tools like Progression charts and heat maps	Ability to identify factors affecting the outcome
SLIM (using integer programming) (Zeng et al., 2015)	Rule Set, Decision tree	Number of nodes
Interpretability using input gradient (Hechtlinger, 2016)	Gradient vector	Ability to identify important features
LRP framework (Samek et al., 2016)	decomposing classification decision in terms of input variables	Highlighting most contributing pixels in image classification
The relational model for meaning of a concept (Condry, 2016)	Features ->Concepts ->meaning	1. Clearly defined input and output requirements 2. Front-end for user experience
Using program snippets as explanations (Singh et al., 2016)	Program snippets	The complexity of program snippets

Continued on next page

Table 2.1 – continued from previous page

Key Idea	Explainer	Evaluation Metrics
Interpreting tree ensembles by finding prototypes in tree space (Tan et al., 2016)	Prototypes in tree space	Human experts to evaluate the quality of found prototypes
Learning a small set of rules in the disjunctive normal form (Wang et al., 2017)	Set of rules in disjunctive normal form	Number of rules can be one straightforward measure
Deep-Taylor decomposition (Montavon et al., 2017)	Decomposing classification decision of a DNN into the contribution from of its input elements. Heat Maps to show the most contributing pixels.	Highlighting the most contributing pixels in image classification. The correctness of highlighted pixels as per domain knowledge experts.
Interpretability prior (Lage et al., 2018)	Decision tree	Optimizing interpretable models
Bayesian case model (Kim, 2015)	Prototypes and subspaces	User agreement regarding clustering performed by the model
Allowing human users to modify the rules (Yang et al., 2019)	Rules set	Improvement in the user agreement
Human guided exploration of solution space (Holzinger et al., 2019)	Modelled as a TSP problem	Savings in terms of computational complexity

Continued on next page

Table 2.1 – continued from previous page

Key Idea	Explainer	Evaluation Metrics
Learning alternate explanations (to be right for the better reasons) (Ross et al., 2017)	Input-gradient based explanations	Robustness of explanations
Highlighting part of the input that is sufficient to make a prediction (Ribeiro et al., 2018)	If-then rules called ‘anchors’	How humans can predict with fewer efforts and higher precision for unseen instances.
Quantifying the importance of a user-defined concept towards a classification result (Kim et al., 2018)	Representing the internal state of a neural network in terms of human-friendly concepts called Concept activation vectors (CAVs).	Answering questions about model decision-making in terms of natural high-level concepts.
Trace a model’s prediction through its learning algorithm and back to training points (Koh and Liang, 2017)	Identify training points most responsible for a given prediction	Debugging models and fixing datasets
Inducing interpretable programs from observed transition system data traces (Penkov and Ramamoorthy, 2017)	LISP-like programs	Inducing interpretable programs

Continued on next page

Table 2.1 – continued from previous page

Key Idea	Explainer	Evaluation Metrics
Investigating which feature interactions are exploited by the classifier (Henelius et al., 2017)	Identification of most contributing feature interactions	Whether a particular grouping of attributes represents attribute interactions structure in a given dataset
Utilizing human tendency of asking "Why this output (the fact) instead of that output (the foil)?" (van der Waa et al., 2018)	Contrastive explanations using a relative complement of the fact rules	Shorter explanations as compared to the full feature list.

2.4 Answers to the research questions formulated prior to literature review

This section concludes the findings of the literature review in the form of answers to the research questions formulated in section 2.1.

Q1. What are the underlying motivations for interpretability?

The key motivations for conferring human interpretability to ML models are listed below:

(i) Trust: ML solutions with interpretability are more probable of winning a trust vote from the end-user.

(ii) New insights: Interpretability has the potential to provide useful insights about the underlying processes and discover new knowledge.

(iii) Scrutinizing ability: A human expert can scrutinize the decision-making process of the classifier.

(iv) Fairness or un-biasedness: Interpretability helps to ensure that solution is not biased towards a race or gender.

(v) Accuracy-interpretability trade-off: It is important to critically evaluate existing interpretability approaches so that new ideas can be generated to handle this trade-off.

(vi) Right to explanation: Starting May 2018, as per new GDPR, human subjects that are going to be affected by outcomes of an ML-based solution have the right to get an explanation for that outcome.

Figure 2.2 presents these motivations graphically.

Q2. What are the desired characteristics of an explanation?

The desired characteristics of an explanation are listed below:

(i) Interpretable: The first and foremost essential characteristic of an explanation is its ability to provide an understanding of the relation between input variables and the model outcome.

(ii) Easy to understand: The number of pieces of information should not be beyond the comprehension capabilities of the user.

(iii) Target audience: An explanation should be in terms of concepts known to the target audience. It may require mapping of features to concepts known to the user and clearly defined requirements for input and goals of the output.

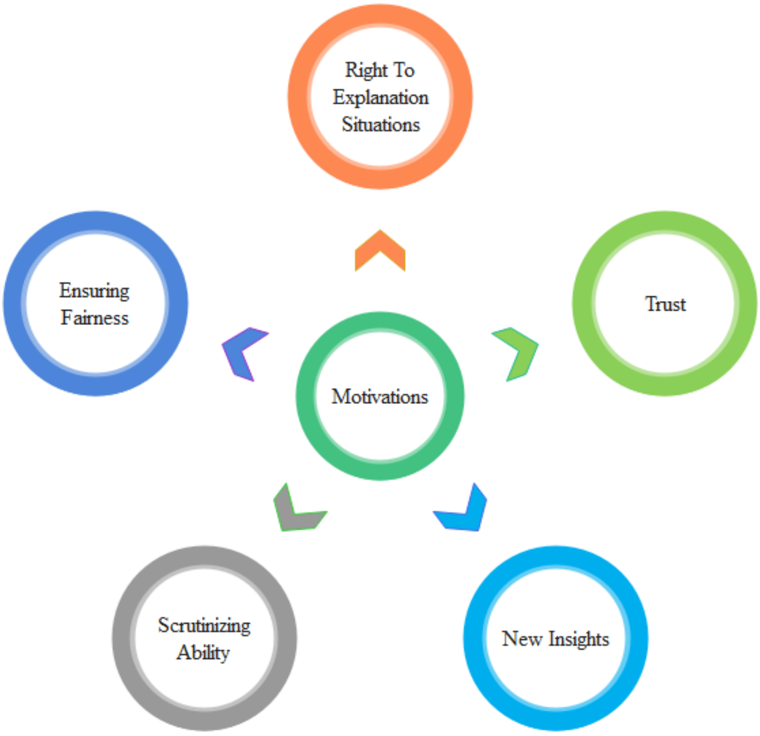


Figure 2.2: Key motivations behind conferring human interpretability

(iv) Local fidelity: an explanation must at least be locally true i.e. must agree to the behaviour of the model in the neighbourhood of the example that is being explained.

(v) Model agnosticism refers to interpretability approaches that are applicable irrespective of which underlying ML model is used.

Figure 2.3 presents these desired characteristics using a visualization.

Q3. What is the taxonomy of the approaches proposed?

The existing work towards conferring interpretability can be broadly classified into model-agnosticism, approaches for interpreting tree ensembles, approaches for interpreting neural networks, human-machine collaboration and addressing lack of consensus regarding terms related to interpretability. Model-agnostic approaches can be further classified into (a) work done towards approximating a complex model with a simple interpretable model and (b) Visual analytics where the input-output relationship is modelled without attempting transparency in the underlying model. This taxonomy can be represented as below:

- Model-agnostic
  - Making underlying model transparent
  - Visual analytics

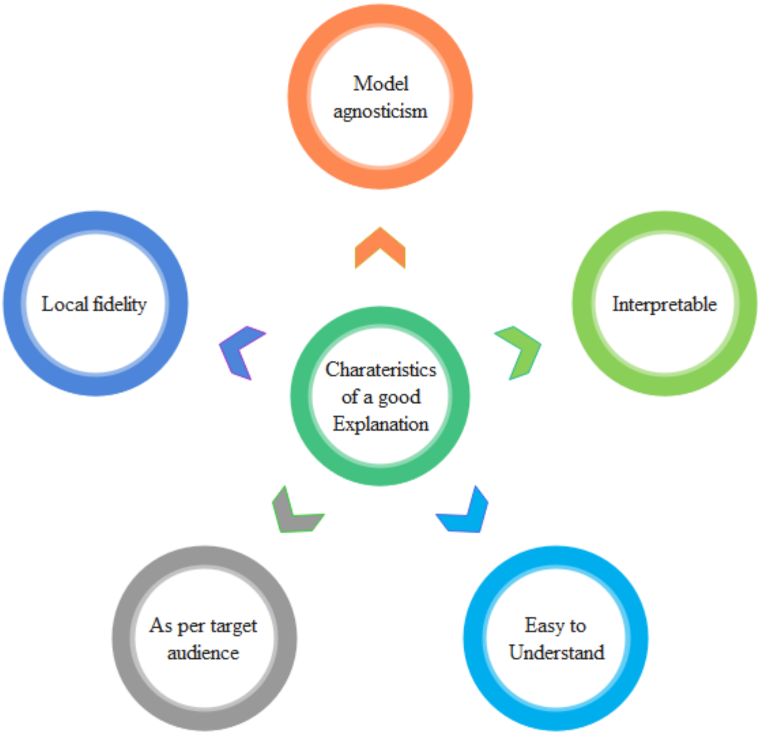


Figure 2.3: Desired characteristics of an explanation

- Model-specific
    - Tree ensembles
    - Neural Networks
- Human-machine interaction
- Addressing lack of consensus

Q4. What are the key ideas and explainers used for conferring interpretability?

The key ideas and explainers used for conferring interpretability are listed below:

(i) approximating a complex model at a local scale using a linear model

(ii) local explanation vector to explain an instance-specific outcome by identifying more influential features

(iii) using fundamental concepts of coalitional game theory to explain an outcome in terms of feature wise contribution

(iv) listing important features using explanation curves

(v) interpretability using partial derivative with respect to input

(vi) using program snippets as explanations

(vii) extracting an interpretable decision tree from a tree ensemble using Expectation-Maximization algorithm

(viii) extracting a single interpretable tree from a population of trees using Genetic Algorithms

(ix) interpreting tree ensembles by finding prototypes in tree space

(x) extracting interpretable features and summarizing results in the form of rule sets

(xi) learning a falling rule list using Bayesian framework

(xii) learning a small set of rules in disjunctive normal form

(xiii) heat maps to show most contributing pixels using deep-Taylor decomposition

(xiv) modeling input-output relationships using visual analytics

(xv) making outcomes of black-box model more expressive using heatmaps and progression charts.

Figure 2.4 presents these key ideas using a visualization.

Q5. What are the metrics used for evaluating interpretability?

(i) decomposing classification outcomes into feature-wise contribution, both the magnitude and direction of contribution

(ii) complexity of the explanation measured as the volume of the information in the explanation to be comprehended. These include the number of nodes in the decision tree, number of rules in the ruleset, and complexity of program snippets

(iii) ease of identifying how a data point is to be moved to change its label

(iv) evaluation of the quality of explanations by human experts. Can these explanations enable human subjects in choosing the best classifier? Can explanations help improve an untrustworthy classifier? Can explanations lead to insights?

(v) how many outcomes are explained using only top-ranked features

(vi) utilizing the power of visual perception in humans

(vii) MEP (Mean Explainability Precision) and MER (Mean Explainability Recall) using explainability score

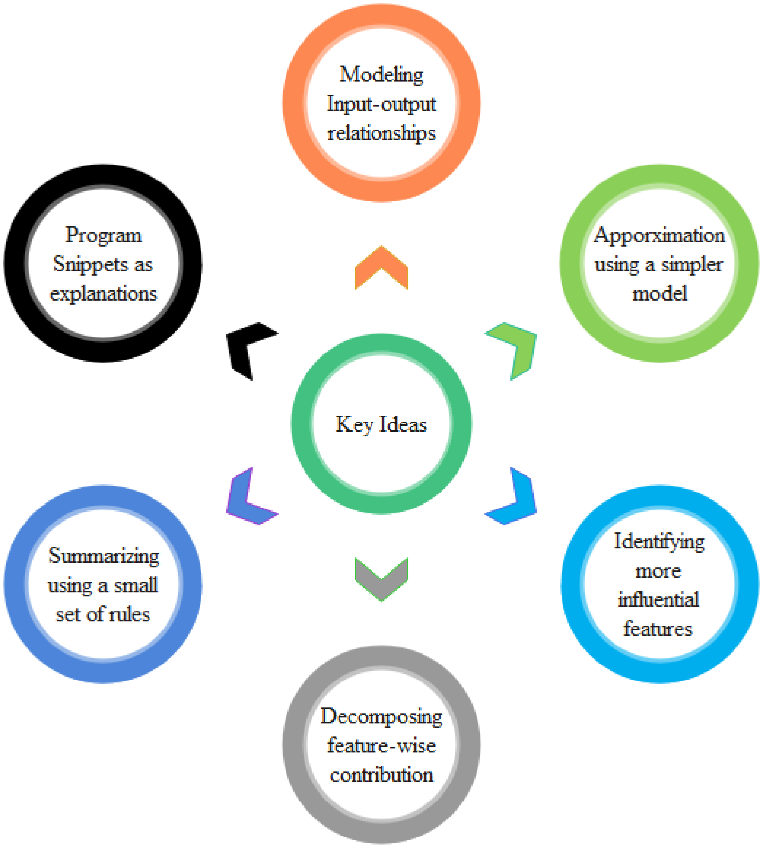

Modeling Input-output relationships
Program Snippets as explanations
Apporximation using a simpler model
Key Ideas
Summarizing using a small set of rules
Identifying more influential features
Decomposing feature-wise contribution

Figure 2.4: Key Ideas for conferring human interpretability

(viii) quantifying interpretability using information entropy instead of just count of features.

Figure 2.5 presents these metrics using a visualization.

Q6. Which ML models have been attempted for interpretability?

The ML models that have been attempted for interpretability included decision trees, Naïve Bayes, tree ensembles, Artificial Neural Networks, logistic regression, support vector machine, AdaBoost, k-nearest Neighbors, Xgboost, Random forests, Classification and regression trees, and simulated annealing. Almost all common ML algorithms have been attempted for conferring interpretability.

Q7. What are the types of problems attempted?

Classification using Titanic dataset, Fisher's IRIS dataset, UCI datasets; Text classification; Sentiment analysis; Image classification; Predicting product interest using online browsing data; Explaining movie recommendations; Classification of breast Magnetic Resonance Images (MRIs); Recidivism prediction; Facebook data; Predicting heating load of a building using synthetic and energy efficiency data.

Q8. Is Interpretability compulsory always?

Interpretability is particularly useful in situations that involve critical decision mak-

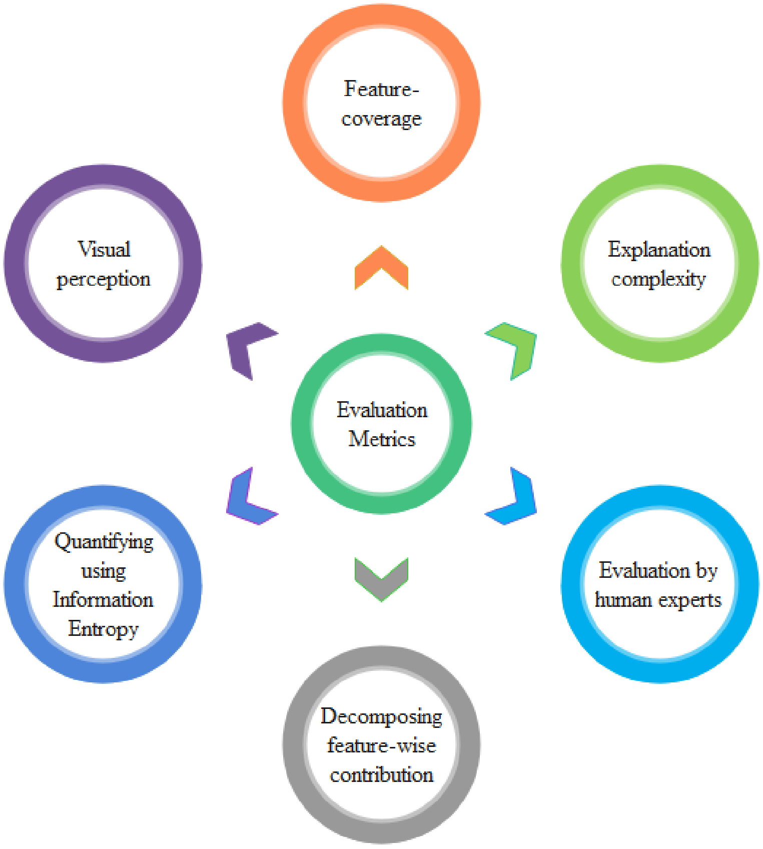

Feature-coverage
Visual perception
Explanation complexity
Evaluation Metrics
Quantifying using Information Entropy
Evaluation by human experts
Decomposing feature-wise contribution

Figure 2.5: Metrics for evaluating human interpretability

ing and a say of human domain experts. It is not compulsory always and can be harmful rather in certain contexts. Possible dangers of transparency include

(i) divergence between intended audience and actual beneficiary

(ii) transparency in government use of algorithms

(iii) gaming of rules and lack of motivation of intellectual property if all algorithms are open source

(iv) discrimination of sub-groups based on sensitive features.

Q9. Can human-machine collaboration be useful in incorporating domain expertise?

Humans and machine have their unique strengths and a collaboration between the two can help to incorporate domain expertise into the learning of ML systems. Potential advantages of a human-machine collaboration include improved user agreement, accelerated exploration of solution space, providing a bigger role to end-users in the design of interactive and interpretable ML systems, and incorporating multi-disciplinary expertise. The key idea is to present the learned model to human domain experts in an interpretable manner. The human experts in turn provide feedback to the ML system based on their experience-based perception regarding the problem domain. The ML system attempts to adapt itself to incorporate the feedback provided by human experts. In case there is a contradiction between what a human expert says and what data says, the situation is reported as a conflict.

Q10. What are the open research directions?

The field of human interpretability is still evolving and offers the following opportunities for future research:

(i) Discovering important individual interactions of variables to provide new insights and facilitate verification by human experts.

(ii) Need of optimization to reduce the computational complexity for interpretability in real-time.

(iii) Extending proposed approaches from classification to regression tasks.

(iv) Exploring alternatives as explainers and metrics for their performance comparison.

(v) Addressing challenges like providing global explanations, inconsistency in local explanation, and powerful forms of user feedback.

(vi) Explore alternatives for proximity measures between predictive model and its interpretable approximation, varying number of regions in the interpretable mode and automatic computation of the number of regions, while interpreting tree ensembles.

(vii) Using human experts to check the quality of prototypes and investigating prototype-finding procedures while using tree space prototypes for interpretability.

(viii) Identifying metrics to compare interpretability through visual analytics, finding new graphical tools to improve expressiveness, and using this expressive information to guide interventions.

(ix) Addressing lack of consensus in definition, types and metrics to measure interpretability. There is a need to distinguish interpretability measures of models and representations, linking results of user-based surveys and heuristics by translating former to later.

(x) There is a need to define criteria for the faithfulness of an explanation.

(xi) There is a need for a framework to understand when transparency is useful and when it is harmful.

(xii) The need of incorporating domain knowledge into the ML process; clearly defined requirements for input and goals of the output; models that complement and expedite existing processes; preparing a front-end for the intended audience in terms of concepts known to them to improve interpretability.

Chapter 3

Data, ML Model and Human domain experts: A collaboration for reliable and trustable learning

3.1 Introduction

Every ML solution acquires its learning from the existing experience fed to the machine, usually, in the form of a dataset. Any ML algorithm employed aims to capture important characteristics from this dataset provided for learning. Moreover, to facilitate the trust of human users, the behaviour of the learned ML model must be in sync with the prevailing domain knowledge. So, an ML-based solution is expected to be in sync with the provided dataset and human domain experts. Intuitively, the provided dataset, the learned ML model and human domain experts can be considered as the three important pillars for constructing an ML-based solution.

The goal of this chapter is to propose a framework for verifying the learning acquired by an ML model. The underlying dataset, ML model and human domain experts are the three pillars of this framework. The proposed approach aims to listen to each of these three pillars regarding their perception of the problem domain under study. By 'listen to' we refer to understanding what features are important in terms of affecting the outcome of the target variable. The dataset is listened to by computing information gain measures using entropy and Gini index. The ML model is listened to in terms of variable importance measures. The human experts are listened to by collecting the importance of features as per their perception of the problem domain based on their experience. After listening to each of the three pillars, the framework measures the degree of agreement between them in terms of their perception of the problem domain. The problem of measuring the agreement between these three pillars has been modelled as a Spearman's rank correlation problem.

The proposed approach has been validated using two problem domains: (i) predicting the joining behaviour of freshmen students using a primary dataset and (ii) predicting the onset of diabetes using a publicly available standard dataset. We demonstrated that the proposed framework was capable of verifying the learning acquired by the ML model against the provided dataset and the prevailing domain knowledge as per feedback taken from human domain experts. This chapter has the following contributions:

(i) A framework that listens to the important pillars of an ML-based solution: the dataset, ML model and human domain experts

(ii) Enable verification of the learning acquired and facilitating trust of human users

(iii) Due to its quantitative nature, this approach is having the potential of forming a basis for developing formal metrics for facilitating trust in an ML model

(iv) Validation of the proposed approach using a primary dataset and a standard dataset

The rest of the chapter is organized as follows. Section 3.2 summarizes the related work done in the field of conferring human interpretability. Section 3.3 describes the problem formulation and hypothesis framed. Section 3.4 describes the proposed framework using a framework diagram and an algorithmic description. Section 3.6 discusses the datasets used for validation of the proposed approach. Section 3.7 describes the experimental setup for validating the proposed approach along with evaluation methods. Section 3.8 Summarizes and discusses the results obtained during the experiments. Section 3.9 summarizes conclusions of the chapter.

3.2 Related Work

Advanced ML models have demonstrated their supremacy in terms of prediction accuracy. However, their prediction outcomes are not easily interpretable to non-ML experts. This lack of human interpretability is of concern in problem domains involving critical decision making. Human interpretability in ML offers advantages like facilitating trust, new knowledge discovery, model debugging and ensuring fairness. As a result, a renewed interest has been observed during recent years, towards conferring human interpretability to ML-based solutions.

The existing work can be categorized into model-specific and model-agnostic methods. Model-specific methods apply to a particular underlying ML algorithm whereas model-agnostic refer to techniques that apply to any underlying model. The ‘inTrees’ framework that consists of algorithms to extract, process, prune and summarize rules from a tree ensemble has been proposed (Deng, 2019). The use of two models, ’P’ for prediction and ’I’ for interpreting has been proposed to approximate the learned complex tree ensemble by a simple interpretable model using KL-divergence as proximity measure (Hara and Hayashi, 2016). An approach benefitting from two unique aspects of tree ensembles i.e. leveraging tree structure and naturally-learned similarity measure has been proposed for Interpreting tree ensembles by finding prototypes in tree space (Tan et al., 2016). GENESIM algorithm for extraction of a single interpretable model from an initial population of decision trees using a genetic algorithm has been proposed (Vandewiele et al., 2016). The Deep Taylor decomposition technique has been used to decompose decisions of neural networks in terms of contribution from input elements (Montavon et al., 2017). Layer-wise relevance propagation (LRP) framework has been proposed to explain predictions of a deep neural network (Samek et al., 2016).A general method to explain an outcome in terms of individual contributions of features using concepts of coalitional game theory has been proposed (Strumbelj and Kononenko, 2010). Local explanation vector that is an estimation of local gradients has been used to understand instance-specific outcome (Baehrens et al., 2010). Algorithm LIME has been proposed to explain the prediction outcome of any classifier by learning an interpretable model locally around the prediction (Ribeiro et al., 2016*b*). Use of input gradient i.e. partial derivative of the model with respect to the input is proposed for interpreting any model (Hechtlinger, 2016). The paper (Ribeiro et al., 2016*a*) advocates the use of a model-agnostic approach covering advantages and associated challenges. Using programs as model-agnostic local explanations have also been proposed (Singh et al., 2016).

Due to its subjective nature, the field of interpretable ML is still evolving. There is a lack of consensus related to the definition and evaluation of interpretability. There is a lack of established formal metrics to compare ML models in terms of interpretability. Moreover, a collaboration between machines and humans has been advocated due to their unique strengths.

A majority of the existing work to confer human interpretability to ML models focuses on identifying features that are important in terms of their contribution towards the predicted outcome. As discussed in the introduction section, the dataset, ML model and human domain experts can be considered as the three pillars of an ML-based solution. Any ML-based solution is expected to agree with the provided dataset and prevailing domain knowledge. Agreement with the dataset implies that the learning process has been accurate. Agreement with prevailing domain knowledge implies that the ML solution is in sync with human domain experts. Ensuring this two-way agreement, with dataset and human experts, is indeed an intuitive idea, to ensure accuracy as well as facilitate acceptability of the developed solution. If any ML-based solution is in agreement with the human domain experts regarding their perception of the problem domain, it helps in facilitating trust in the ML-based solution. In this chapter, we demonstrate this idea of collaboration between dataset, ML model and human domain experts to verify the learning acquired by an ML model.

3.3 Problem formulation

This research work is based on an intuitive idea that the underlying dataset, ML model learned using that dataset and feedback from human domain experts are the pillars of an ML-based solution. The desired characteristics of any ML-based solution include (i) the learning process must result in an ML model that incorporates important characteristics of the underlying dataset and (ii) the behaviour of an ML model is expected to be in sync with the perception of the human domain experts regarding the problem domain. For example, let us consider an ML-based model for the classification of mammograms as malignant or benign. Medical experts are likely to trust this model only if its decision-making behaviour is as per the perception of the medical experts based on years of medical practice.

The idea is to listen to each of these three pillars (dataset, model and human expert) and measure the degree of agreement between these. The dataset is listened to in terms of identifying important characteristics using entropy and the Gini index. The ML model is listened to in terms of interpreting its decision-making behaviour using feature importance measures. The human experts are listened to by taking feedback in terms of the importance of a feature as per their experience-based perception regarding the problem domain.

The proposed collaboration aims to answer the following questions:

Q1. Did the learned ML model capture important characteristics of the provided dataset?

Q2. Is the learning acquired by the ML-based model in sync with the prevailing domain knowledge?

If the answer to both the questions is a 'Yes', the model can be recommended for reliability and trust vote. If the answer to any of the questions is a 'No', it is an indicator that there is a need to investigate the input dataset, the learning process or the process of capturing human expert feedback. Outcomes of this investigation may lead to any of the following:

(i) A problem in the dataset preparation

(ii) A problem in the machine learning process adopted

(iii) A problem in the process of capturing feedback of human experts

(iv) A case of discovering a new knowledge

The problem of measuring the degree of agreement between dataset, ML model and human domain experts has been modelled as a Spearman's rank correlation problem. Dataset, ML model and human domain experts are considered as Judges. The features used for learning the ML model are considered as participants to be ranked by the judges. As Spearman's rank correlation test applies to 2-Judges at a time, the problem of measuring agreement between three judges has been addressed by taking a pair of judges at a time and repeating the test for each of the three pairs of judges. Spearman's rank correlation is denoted by the symbol (ρ_s).

To verify the statistical significance of the agreement observed between these three judges, hypothesis testing has been used. The following hypothesis was formulated:

Null hypothesis (H_0): Ranking of features by two judges are independent ($\rho_s = 0$)

The alternate hypothesis (H_1): Ranking of features by two judges have a positive association ($\rho_s > 0$)

Acceptance of the null hypothesis for any of the pair of judges is an indicator that the concerned judges are not in sync with each other or do not have an agreement in terms of the importance of features. On the other hand, acceptance of alternate hypothesis is an indicator that the concerned pair of judges have a positive association or they are in agreement in terms of their perception regarding the problem domain (rankings given to the features).

3.4 Proposed framework

This section describes the proposed framework using a framework diagram and an algorithmic description of the proposed framework.

Figure 3.1, shows the diagrammatic representation of the proposed framework. Entropy and Gini index have been computed as dataset measures to listen to the dataset. Feature importance measures have been used for interpreting the decision-making behaviour of the model. After listening to the dataset and model, the next step is to measure the degree of agreement between the dataset and the model regarding the ranking of features in terms of importance. Spearman's rank correlation has been used to measure this degree of agreement. If the dataset and the model are in agreement, the agreement with the human expert feedback is measured. If the dataset, model and human expert are in

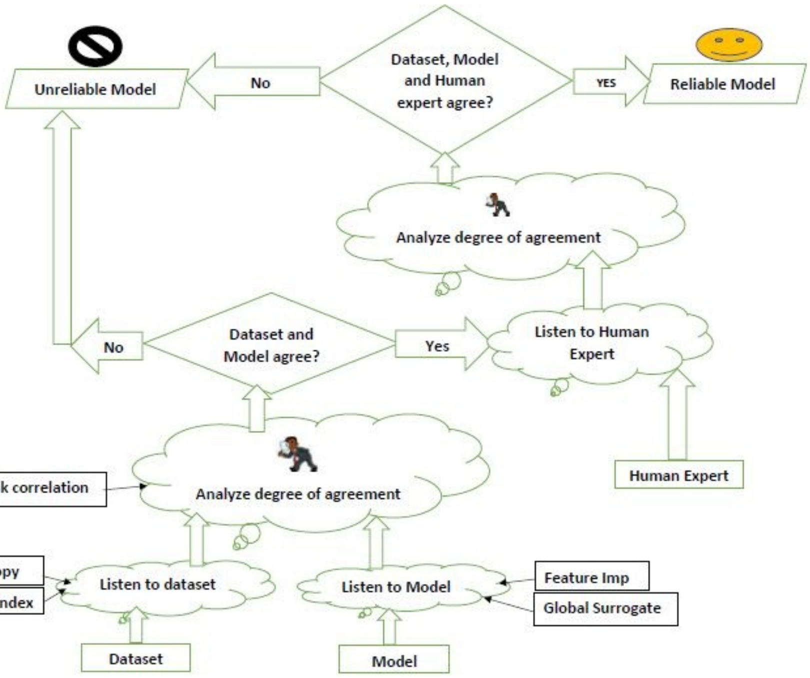


Figure 3.1: Framework for data, model and human expert collaboration

agreement, the model is considered trustworthy. If the degree of agreement is not found at any of the two stages, the model is considered unreliable.

The input to the system was the provided dataset and a black-box ML-based model learned using that dataset. For each feature used, information gain using entropy and Gini index was computed. Using these values features were ranked, with rank 1 given to feature with maximum information gain. To find a single rank for each feature, the average of the two ranks, one using entropy and one using Gini index, was taken. To listen to the model, multiple variable importance measures were computed for each feature. Using these measures, an average rank was assigned to each feature as per its importance to the model. To strengthen our understanding of model behaviour, a global surrogate model was also used. To measure the degree of agreement, spearman's rank correlation was used. To verify the significance of the degree of agreement, hypothesis testing using one-tail Spearman's rank correlation was used. If the dataset and model do not agree, the model is considered unreliable, else stage two processing starts. To listen to the human experts, each expert was asked to assign a unique rank to each feature based on his or her perception regarding the problem domain. For each feature, an average rank was computed by taking an average of ranks assigned to that feature by all human experts. The degree of agreement between dataset, model and human experts was computed and tested using spearman's rank correlation. If the dataset, the ML model, and human experts do not have a degree of agreement, the model was not considered trustable. The basic approach was quite intuitive. It aimed to listen to the dataset, the model and the

```
Data: Dataset, a black-box ML model
Result: Interpreting model behaviour, verifying learning, facilitating trust
FEATURES = Set of features used for learning;
HUMANS = Set of human domain experts;
for each f in FEATURES do
    Compute $IG_{e,f}$ and $IG_{g,f}$;
end
for each f in FEATURES do
    Assign $R_{e,f}$ using $IG_{e,f}$;
    Assign $R_{g,f}$ using $IG_{g,f}$;
end
Compute $R_{dataset,f}$ as average of $R_{e,f}$ and $R_{g,f}$;
for each f in FEATURES do
    Compute $mmd_f$, $acc\_d_f$, $gini\_d_f$, $times\_root_f$ and $Imp_f$ ;
end
for each f in FEATURES do
    Assign $R_{mmd,f}$ using $mmd_f$;
    Assign $R_{acc\_d,f}$ using $acc\_d_f$;
    Assign $R_{gini\_d,f}$ using $gini\_d_f$;
    Assign $R_{times\_root,f}$ using $times\_root_f$;
    Assign $R_{Imp_f}$ using $Imp_f$;
end
Compute $R_{Model,f}$ as average of $R_{mmd,f}$, $R_{acc\_d,f}$, $R_{gini\_d,f}$, $R_{times\_root,f}$ and
  $R_{Imp_f}$;
Analyze degree of agreement between $R_{dataset,f}$ and $R_{Model,f}$;
if Dataset and Model do not agree then
    Model is not reliable. Exit.
end
for each h in HUMANS do
    for each f in FEATURES do
        Collect $R_{h,f}$ based on human expert perception
    end
end
for each f in FEATURES do
    Compute $R_{Human,f}$ as average of ranks assigned to that feature $R_{h,f}$;
end
Analyze degree of agreement between $R_{dataset,f}$ , $R_{Model,f}$ and $R_{Human,f}$;
if Dataset, Model and Human agree then
    Model is trustworthy.
else
    Model is not reliable.
end
```

Algorithm 1: Dataset, model, and human expert collaboration

human domain experts and then analyze the agreement between them. A higher degree of agreement between these three was taken as an indicator that the model had listened to the dataset properly and is also in agreement with the prevailing domain knowledge. Algorithm 1 describes the steps involved and related computations.

Computational complexity analysis: The algorithm 1 comprises of 5 for loops, 1 nested for loop and 2 conditional statements. Each of the for loops iterates over the number of features 'n' and has a complexity of O(n). The nested for loop represents the collection of feedback from human experts and does not require any computational efforts. The conditional statement is executed only once giving a complexity of O(1). So, the computational complexity of the algorithm is O(4n)+O(1)+O(n)+O(1) = O(5n+2) = O(n).

3.5 Experiments designed

The experiments designed and the underlying objectives have been compiled in table 3.1.

Table 3.1: Description and objective of experiments designed

Experiment #	Experiment description	Experiment Objective
1	Exploring multiple ML algorithms to learn an accurate model	Learning an accurate ML model
2	Identifying important features from the dataset using entropy and Gini index measures	Listening to what data says?
3	Understanding model behaviour using interpretability techniques	Listening to what the model says?
4	Capturing feedback from human domain experts based on their perception regarding the problem domain	Listening to what human domain experts say?
5	Measuring and analyzing agreement between dataset, model, and human experts	Determining whether dataset, model and human experts are in agreement?

3.6 Material (Dataset)

3.6.1 Problem domain 1: Predicting Joining behavior of freshmen students

As per the University Grant Commission (UGC) report, there are almost 900 universities across India including more than 300 private universities UGC (2019). Most of these private universities spend hugely in competing to reach out to students who are exploring different educational institutes to take admission into. Due to this competition, educational institutions start their admission process for the next session quite early, usually at the start of the year. Every admission year, from the students, enrolled in an educational institution, some students do not join the same institution. The reasons include getting a higher scholarship in some other institution, preferred college or discipline of interest in some other institution. Each student, who takes admission but do not join is a loss to the concerned educational institute in terms of resources invested. Ability to foresee such students is important for an educational institute. Institutes can use this information to plan activities towards improving the retention of enrolled students.

Justification of choosing this problem: The objective is to explore the applicability of ML in helping educational institutions predict the joining behaviour of their freshmen students. The related work in the field of educational data mining focusing the use of ML solutions to predict student performance, predict student dropouts and predict the employability of students. Exploring ML to foresee the joining behaviour of freshmen students is a novel problem domain. This problem domain is dynamic in nature as student joining behaviour is affected by multiple factors like changes in admission policies and trends in the education sector every year. As the proposed approach requires feedback from human domain experts, the availability of experienced admission counsellors justifies it as a suitable choice for validating this approach. The outcomes of this study can be used by an educational institution to improve its admission processes and retention of enrolled students.

The idea is to formulate the problem of predicting the joining behaviour of freshmen students as a binary classification ML problem. The target variable is the actual joining status of a student with ‘Joined’ or ‘Lost’ as possible outcomes. The first objective is to learn an ML-based model using admission details and the actual joining status of freshmen students of previous batches for learning. Such a model can be used for predicting the joining behaviour of freshmen students of the upcoming batches. The second objective is to analyze important factors that contribute towards the joining behaviour the most.

The dataset consisted of students enrolled in an educational institute in North India,

Table 3.2: Demography of the subjects used

Attribute	Type	Description
RegistrationNumber	Int	Unique ID for each student
AdmissionMonth	Int	Month of admission e.g 5 = May, 6 = June
Gender	Factor	F - Female, M – Male
State	Factor	Home state of student
HomeTownType	Factor	Rural, Urban (Metropolitan), Urban (Town)
ProgrammeName	Factor	Programme enrolled e.g. MBA, B. Arch. etc.
Discipline	Factor	Agriculture, Management etc.
QualifyingExam	Factor	Eligibility qualification e.g. 10+2, Graduation
MarksPercent	Factor	Marks in qualifying exam
CategoryCode	Factor	General, Scheduled Caste (SC), Scheduled Tribe (ST) etc.
TransportAvailed	Factor	Transport facility of university availed? [Yes/No]
LoanLetter	Factor	Education loan availed? [Yes/No]
PreviouslyStudied	Factor	Whether student studied earlier in the same institution? [Yes/No]
HostelAvailed	Factor	Hostel availed or not? [Yes/No]
MessAvailed	Factor	Mess availed or not? [Yes/No]
ScholarshipPercentage	Numeric	Scholarship amount offered as percentage of tuition fee
ScholarshipBracket	Factor	High, Low, Medium
EconomicCondition	Factor	AboveAverage, Average. BelowAverage, Good
FeePaidPercentage	Numeric	Percentage of fee paid so far by student
MediumOfStudy	Factor	English, NonEnglish
FeePaidCategorized	Factor	High, Low, Medium
MarksCategory	Factor	Excellent, Fail, FirstDivision, SecondDivison, ThirdDivision
HostelOrTransport	Factor	Hostel or transport availed? [Yes/No]
StudentStatus	Factor	Joined, Lost (Did not join)

during 2018 (Kumar et al., 2020). The target variable was ‘JoiningStatus’ with values as ‘Joined’ or ‘Lost’. The dataset consisted of 13125 observations and 25 variables. Out of 13125 students, 8374 joined while the remaining 4751 did not join, giving a baseline accuracy of 63.8%. Table 3.2 shows the structure of our dataset with the type and description of each attribute.

Feature selection: To shortlist the relevant features for learning a ML model, correlation of each predictor was analyzed with the target variable ‘JoiningStatus’. As majority of the predictors are categorical in nature, chi-square test statistic was used to identify significant correlation. Also, a few features were dropped on the basis of commonsense.

(i) Redundant features:

- RegistrationNumber: Registration number of a student is not expected to affect joining behavior of a freshmen student.

- BatchYear: Redundant as data is of 2018 batch only

- MessAvailed: Highly correlated with ‘HostelAvailed’

(ii) Transformed features:

- MarksPercent: Converted to categorical ‘MarksCategory’

- TransportAvailed: Utilized for deriving ‘HostelOrTransport’

- HostelAvailed: Utilized for deriving ‘HostelOrTransport’

- ScholarshipPercentage: Converted to categorical ‘ScholarshipBracket’

- FeePaidPercentage: Converted to categorical ‘FeePaidCategorized’

(iii) Features not correlated with the target variable:

- Gender: p-value = 0.1724, Not correlated with the target variable

- MediumOfStudy: p-value = 0.2655, Not correlated with the target variable

Features correlated with the target variable:

- AdmissionMonth: p-value = 3.695e-06

- HomeTownType: p-value < 2.2e-16

- QualifyingExam: p-value = 3.598e-08

- LoanLetter: p-value < 2.2e-16

- PreviouslyStudied: p-value < 2.2e-16

- ScholarshipBracket: p-value < 2.2e-16

- FeePaidCategorized: p-value < 2.2e-16

- MarksCategory: p-value < 2.2e-16

- HostelOrTransport: p-value < 2.2e-16

Based on the results of using feature selection techniques the following 9 features were shortlisted for learning ML model:

(i) AdmissionMonth

Table 3.3: Demography of the subjects for ’diabetes’ dataset

Attribute	DataType	Description
Age	Int	Age (years)
BloodPressure	Int	Diastolic blood pressure (mm Hg)
BMI	Numeric	Body mass index (weight in kg/(height in m)^2)
DiabetesPedigreeFunction	Numeric	Diabetes pedigree function
Glucose	Int	Plasma glucose concentration 2 hours in an oral glucose tolerance test
Insulin	Int	2-Hour serum insulin (mu U/ml)
Pregnancies	Int	Number of times pregnant
SkinThickness	Int	Triceps skin fold thickness (mm)
Outcome	Int	0 (not diabetes), 1(diabetes)

(ii) MarksCategory

(iii) LoanLetter

(iv) HomeTownType

(v) QualifyingExam

(vi) PreviouslyStudied

(vii) ScholarshipBracket

(viii) FeePaidCategorized

(ix) HostelOrTransport

3.6.2 Problem domain 2: Predicting onset of diabetes using diagnostic measures

As a second problem, a standard dataset ‘diabetes’ contributed by the National Institute of Diabetes and Digestive and Kidney diseases was selected. This dataset is available on Kaggle (Kaggle, 2016). The objective was to use ML to learn a model that can be used to predict ‘diabetes’ taking different diagnostic measures as input. As our proposed approach is useful in situations involving feedback as well as a say from human domain experts, this dataset was found a fit for authenticating our approach.

The dataset consisted of 768 records of female patients of Pima Indian origin. Out of 768 records, 268 are 1(diabetic) and 500 are 0 (not diabetic), giving a baseline accuracy of 65.1%. Table 3.3 shows the structure of our dataset with the type and description of each attribute.

Feature selection: To identify the features relevant for learning, the correlation of each predictor with the target variable was computed as compiled in table 3.4. To identify any correlation between predictors, a correlation matrix was computed as detailed in table 3.5.

Table 3.4: Correlation of features with the target variable

Feature(Diagnostic measure)	Correlation with Outcome (diabetes)
Age	0.238
BloodPressure	0.065
BMI	0.293
DiabetesPedigreeFunction	0.174
Glucose	0.467
Insulin	0.131
Pregnancies	0.222
SkinThickness	0.075

Table 3.5: Correlation among predictors

	Preg nan- cies	Glucose	Blood Pres- sure	Skin Thick ness	Insulin	BMI	Diabetes Pedigree Function	Age
Pregnancies	1.00	0.13	0.14	-0.08	-0.07	0.02	-0.03	0.54
Glucose	0.13	1.00	0.15	0.06	0.33	0.22	0.14	0.26
BloodPressure	0.14	0.15	1.00	0.21	0.09	0.28	0.04	0.24
SkinThickness	-0.08	0.06	0.21	1.00	0.44	0.39	0.18	-0.11
Insulin	-0.07	0.33	0.09	0.44	1.00	0.20	0.19	-0.04
BMI	0.02	0.22	0.28	0.39	0.20	1.00	0.14	0.04
Diabetes Pedigree Function	-0.03	0.14	0.04	0.18	0.19	0.14	1.00	0.03
Age	0.54	0.26	0.24	-0.11	-0.04	0.04	0.03	1.00

Outcome: All the predictors were having a positive correlation with the target variable. 'BloodPressure' and 'SkinThickness' were having a very weak correlation. No two features were strongly correlated as correlation was not very high (>.7). So, all features were selected for participating in the learning process.

3.7 Methods

This section talks about the methods used for learning and evaluating an accurate ML-based model. It also discusses methods used for identifying dataset characteristics, interpreting the behaviour of the learned model, taking feedback from human experts and analyzing agreement between the dataset, the model and human experts.

3.7.1 Learning an accurate ML model

Learning algorithms: Logistic Regression (LR), Naive Bayes (NB), Classification and Regression Tree (CART) and Random Forest(RF) algorithms were explored to learn an ML-based model to predict the joining behaviour of freshmen students. LR uses the logistic function to estimate the probability of each class of target variable. NB uses the famous Bayes theorem to calculate the probability that a new instance belongs to a particular category. CART is a decision tree algorithm that has the advantage of being easily human interpretable (Wittkowski, 1986). RF is an ensemble approach to improve the prediction accuracy of CARTs. In RF, for every instance, the classification outcome is the most frequent outcome when each decision tree in the forest is made to give its classification outcome (Breiman, 2001).

Performance metrics for model evaluation Classification accuracy (Acc), sensitivity (Se), specificity (Sp), Precision (P), F1-score and AUC values (area under the ROC curve) were used as performance evaluation metrics. Sensitivity and specificity are not equally critical in all ML problems. In the problem domain of predicting the joining behaviour of freshmen students, it is important that no student who is likely to be 'Lost' should be missed out. These performance metrics were computed using equations 3.1, 3.2, 3.3 and 3.4 respectively.

$$Acc = \frac{TP+TN}{N} \tag{3.1}$$

$$Se = \frac{TP}{TP+FN} \tag{3.2}$$

$$Sp = \frac{TN}{TN+FP} \tag{3.3}$$

$$P = \frac{TP}{TP+FP} \tag{3.4}$$

TP, TN, FP, FN represent true positives, true negatives, false positives and false negatives as computed from the confusion matrix. N is the total number of predictions made.

3.7.2 Listening to the dataset by identifying important characteristics

Information gain measures using entropy and Gini index were computed for each feature to identify important features as per the provided dataset.

Entropy is a measure of disorder or impurity. It is used to measure information gain if a particular feature is selected for splitting while constructing trees (Hausser and Strimmer, 2014).Entropy for a feature 'f' is computed using equation 3.5 (Shannon, 1948). Gini index is a measure of inequality in distribution (Handcock and Morris,

2006) and is always in the range of 0 to 1. It is computed for a feature using equation 3.6.

$$Entropy = -\sum(p_j log_2(p_j) \tag{3.5}$$

where p$_j$ is probability of a class 'j'.

$$Gini\,index = 1 - \sum p_j^2 \tag{3.6}$$

where p$_j$ is probability of a class 'j'.

In our case, there only two possible classes for the target variable as specified below:

- freshmen: 'lost' or 'joined' as possible outcome for the target variable, so j $\in \{lost, joined\}$
- diabetes: 'Yes' or 'No' as possible outcome for the target variable, so j $\in \{Yes, No\}$

Information gain using entropy(IG$_{e,f}$) = entropy(parent)- Weighted average entropy of children

Both entropy and Gini index have been used for information gain to have more matured inferences regarding the importance of individual features as they use different splitting criteria. Also, in this work, multiple ML algorithms were explored before measuring agreement between data, ML model and human experts. However, the proposed approach will work using one of these two also.

3.7.3 Listening to the ML model

Listening to the ML model referred to identifying features that are affecting its decision-making the most. Feature importance measures and global surrogate model techniques were used to interpret ML model behaviour.

(a) Feature importance measures: Feature importance provides a global insight into a model's behaviour. It is a measure of the increase in the model's error rate when the values of a feature are permuted. A higher increase indicates higher importance for the feature. The error rate was measured using 'iml' package in R(Molnar et al., 2018). The importance of a feature 'f' as per model is computed using equation 3.7.

$$Imp_f = 1 - AUC \tag{3.7}$$

where AUC represents area under the ROC curve.

Apart from feature importance, the following additional variable importance measures were used:

- mmd: mean minimal depth
- acc_d: mean decrease in accuracy after a feature is permuted
- gini_d: mean decrease in the Gini index of node impurity
- times_root: total number of trees in which feature is used for splitting the root node

These importance measures were computed using the 'randomForestExplainer' package in R (Liaw et al., 2002). The average rank assigned to a feature 'f' as per model was computed by finding the average of ranks assigned to a feature using mmd, acc_d, gini_d, times_root and Imp_f.

(b) Surrogate model: It is an approximation of a complex black box model using a simpler model that may not be that accurate but is easy to interpret. R squared value is taken as a measure of how well our surrogate model replicates the original black-box model. It measures the amount of variance captured by our surrogate model.

3.7.4 Listening to human domain experts

Listening to human domain experts referred to taking feedback from them as per their perspective of the problem domain. While taking feedback from human experts, outcomes of listening to data and interpreting ML model behaviour were not shared with them to get unbiased feedback. No knowledge of ML was assumed on the part of human experts.

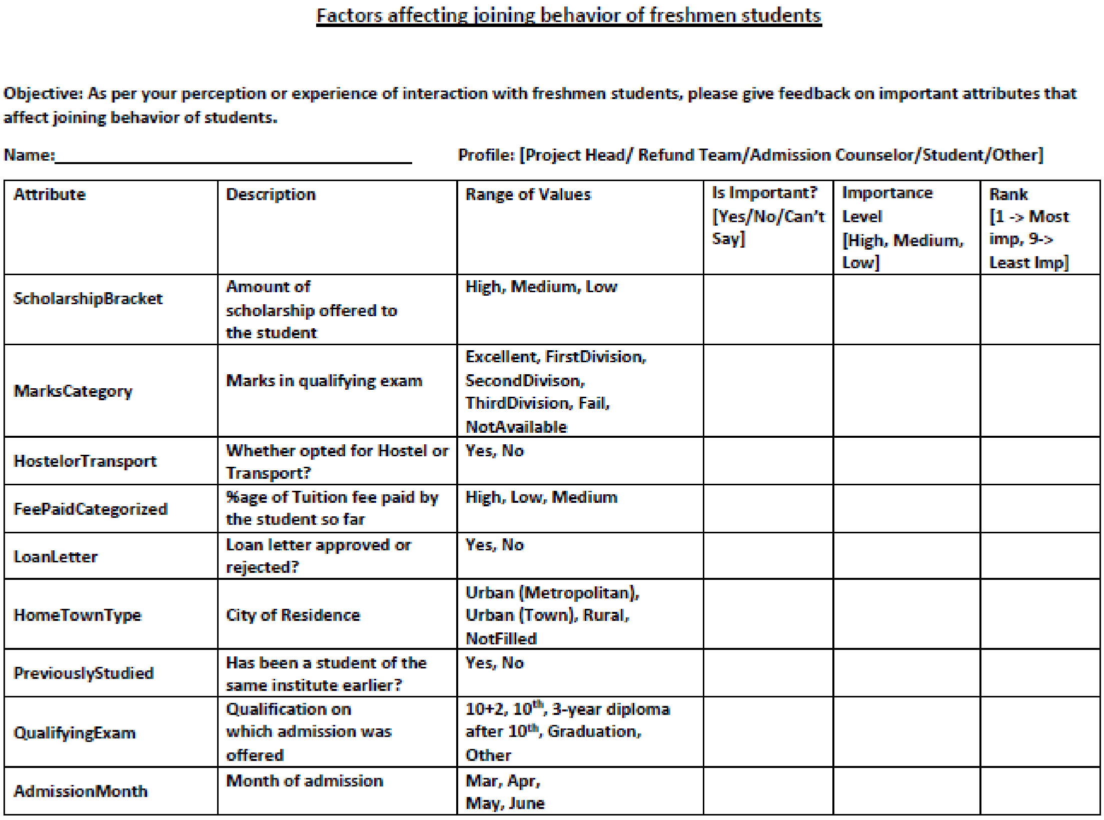
Problem domain 1: Predicting Joining behaviour of freshmen students

Human domain experts in this problem domain: Admission counsellors working in the admissions department were picked as human domain experts. Their job profile was handling queries of aspiring students in-person or telephonically. Based on their conversation with prospective students over the years, they were having a perception developed in their minds regarding the factors that affect the joining behaviour of freshmen students the most. The objective of taking feedback from these human experts was to extract inputs from their experience-based perception.

Eligibility condition: Minimum five years of work experience and recommendation from their project head was set as eligibility criteria for selection as a human expert. The project head was the reporting manager of these admission counsellors.

The process followed for taking feedback: To ensure that the feedback of human experts was based on their experience-based perception only, the outcomes from dataset measures and understanding of the model behaviour was not shared with the identified human experts. Each expert was asked to give feedback on each of the features used for developing an ML-based model. For each feature, an expert had to answer whether that feature is important, level of its importance and assign a relative rank to each feature.

The orientation of human experts: To ensure that the feedback from a human expert is based on their experience only, a training session was organized before taking feedback. Before taking feedback, it was ensured that all experts understood the feedback format and questions. Figure 3.2 shows the format used.



Factors affecting joining behavior of freshmen students

Objective: As per your perception or experience of interaction with freshmen students, please give feedback on important attributes that affect joining behavior of students.

Name:____________________________ **Profile:** [Project Head/ Refund Team/Admission Counselor/Student/Other]

Attribute	Description	Range of Values	Is Important? [Yes/No/Can't Say]	Importance Level [High, Medium, Low]	Rank [1 -> Most imp, 9-> Least Imp]
ScholarshipBracket	Amount of scholarship offered to the student	High, Medium, Low			
MarksCategory	Marks in qualifying exam	Excellent, FirstDivision, SecondDivison, ThirdDivision, Fail, NotAvailable			
HostelorTransport	Whether opted for Hostel or Transport?	Yes, No			
FeePaidCategorized	%age of Tuition fee paid by the student so far	High, Low, Medium			
LoanLetter	Loan letter approved or rejected?	Yes, No			
HomeTownType	City of Residence	Urban (Metropolitan), Urban (Town), Rural, NotFilled			
PreviouslyStudied	Has been a student of the same institute earlier?	Yes, No			
QualifyingExam	Qualification on which admission was offered	10+2, 10th, 3-year diploma after 10th, Graduation, Other			
AdmissionMonth	Month of admission	Mar, Apr, May, June			

Figure 3.2: Feedback form for human domain experts - problem domain 1

The average rank assigned to a feature ‘f’ was computed using equation 3.8.

$$R_{Human,f} = \frac{\sum_{i=1}^{K} R_{h_i,f}}{K} \tag{3.8}$$

where ’K’ denotes number of human domain experts shortlisted for taking feedback. In our case, K = 10.

The following is the list of human domain experts from whom the feedback regarding the problem domain was taken:

(i) h_1: Maninder Kaur, Admission counselor

(ii) h_2: Vandana Sharma, Admission counselor

(iii) h_3: Iesha Saroya, Admission counselor

(iv) h_4: Rajwinder Kaur, Admission counselor

(v) h_5: Pooja, Admission counselor

(vi) h_6: Neha Vaish, Admission counselor

(vii) h_7: Anchal Rai, Admission counselor

(viii) h_8: Prachi Sharma, Admission counselor

(ix) h_9: Anjali Chakma, Admission counselor

(x) h_{10}: Ritu Singla, Admission counselor

Problem domain 2: Predicting the onset of diabetes

A similar approach as in problem domain 1, was followed for listening to human domain experts in predicting the onset of diabetes. Instead of admission counsellors, experienced human medical experts were picked as domain experts.

Eligibility condition: The selection criteria was having a professional medical degree and a minimum of 5 years of medical practice. Also, before taking feedback, it was ensured that the medical expert feels confident in terms of understanding regarding the clinical measures used in the feedback form.

The orientation of human experts: To ensure that the feedback from human medical experts is based on their experience during medical practice only, an interaction before taking feedback was conducted individually with each medical expert. The objective was to ensure that the medical expert understood the feedback format and questions.

Figure 3.3 shows the format used. In our case, K (number of human medical experts) = 5.

The feedback was taken from the following human medical experts:

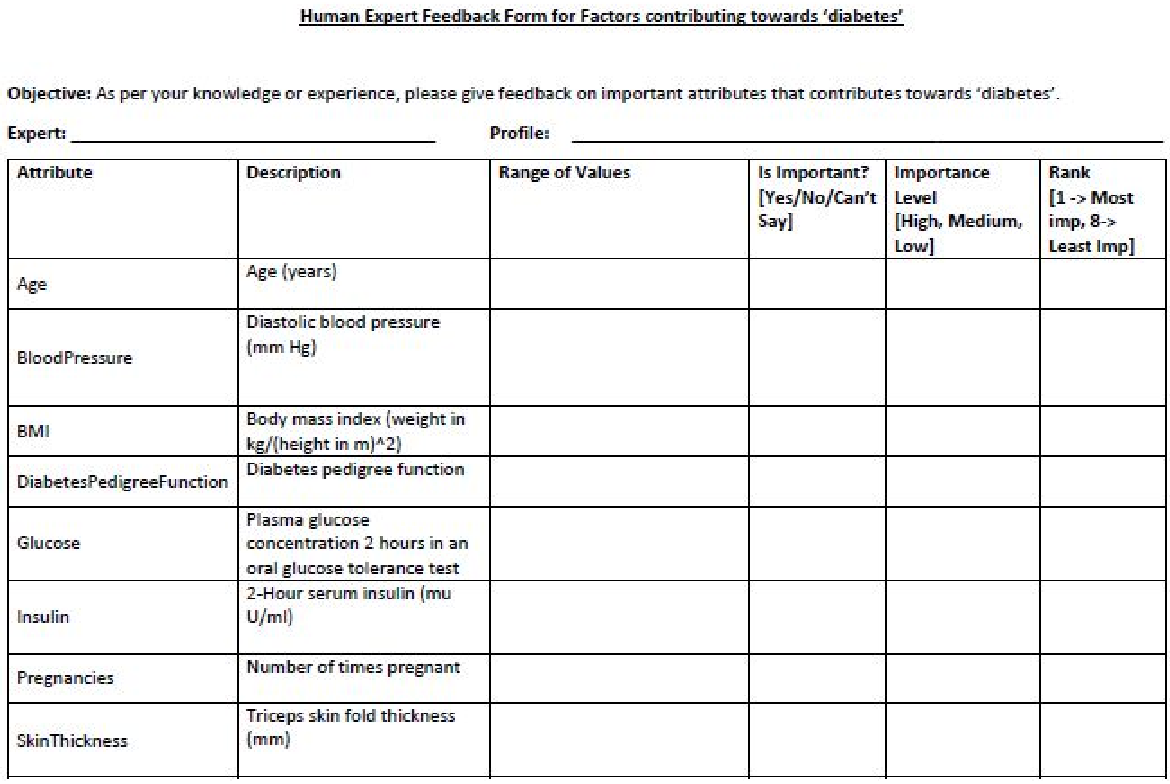
(i) h_1: Dr Shruti Sood, M.B.B.S., DCH

(ii) h_2: Dr Dr Yogesh Kalra, M.B.B.S., M.D.

(iii) h_3: Dr Manjot Kaur, M.B.B.S., M.D.

(iv) h_4: Dr Dhawal Kaushal, M.B.B.S., M.D.

(v) h_5: Dr Vipin Rai, M.B.B.S., M.D.



Human Expert Feedback Form for Factors contributing towards 'diabetes'

Objective: As per your knowledge or experience, please give feedback on important attributes that contributes towards 'diabetes'.

Expert: ______________________ **Profile:** ______________________

Attribute	Description	Range of Values	Is Important? [Yes/No/Can't Say]	Importance Level [High, Medium, Low]	Rank [1 -> Most imp, 8-> Least Imp]
Age	Age (years)				
BloodPressure	Diastolic blood pressure (mm Hg)				
BMI	Body mass index (weight in kg/(height in m)^2)				
DiabetesPedigreeFunction	Diabetes pedigree function				
Glucose	Plasma glucose concentration 2 hours in an oral glucose tolerance test				
Insulin	2-Hour serum insulin (mu U/ml)				
Pregnancies	Number of times pregnant				
SkinThickness	Triceps skin fold thickness (mm)				

Figure 3.3: Feedback form for human domain experts - problem domain 2

3.7.5 Statistical tests and hypothesis testing for agreement analysis

The problem of measuring agreement between the dataset, ML model and human expert was modelled as a '2-Judges and n-participants' rank correlation problem. In a rank correlation problem, there are two judges and 'n' participants. Each of the judges is asked to rank each participant. The objective of the test is to check whether the two judges are in sync in terms of their rating of the participants. The value of the rank correlation coefficient(ρ_s) varies from -1 to +1. The Spearman's rank correlation coefficient (Spearman, 1987)is computed using equation 3.9.

$$\rho_s = 1 - \frac{6\sum_{i=1}^{n} {d_i}^2}{n(n^2-1)} \tag{3.9}$$

where d$_i$ is the difference between the ith pair of ranks and 'n' is number of features used for learning.

The use of Spearman's rank correlation is justified as it has an analogy with the problem of measuring agreement between data, ML model and human experts. In a '2-Judges n-participants' rank correlation problem, judges are asked to rank participants on the same agreed-upon parameters. The objective of using Spearman's rank correlation is to measure the agreement between the two judges in terms of ranks assigned to participants. On the same lines, the problem of measuring agreement between dataset, ML model and human experts can be modelled as a Spearman's rank correlation problem. As an analogy, each feature used for learning was considered equivalent to a participant. Dataset, ML model and human domain expert were considered as judges. As we were to analyze the agreement between three judges instead of two, the problem was addressed by taking a pair of two judges at a time and repeating for each of the three possible pairs.

Figure 3.4 presents an analogy of the problem of measuring agreement between data, model and human experts with the Spearman's rank correlation problem. Each feature used for learning was considered equivalent to a participant. Dataset, ML model and human domain expert were considered as judges.

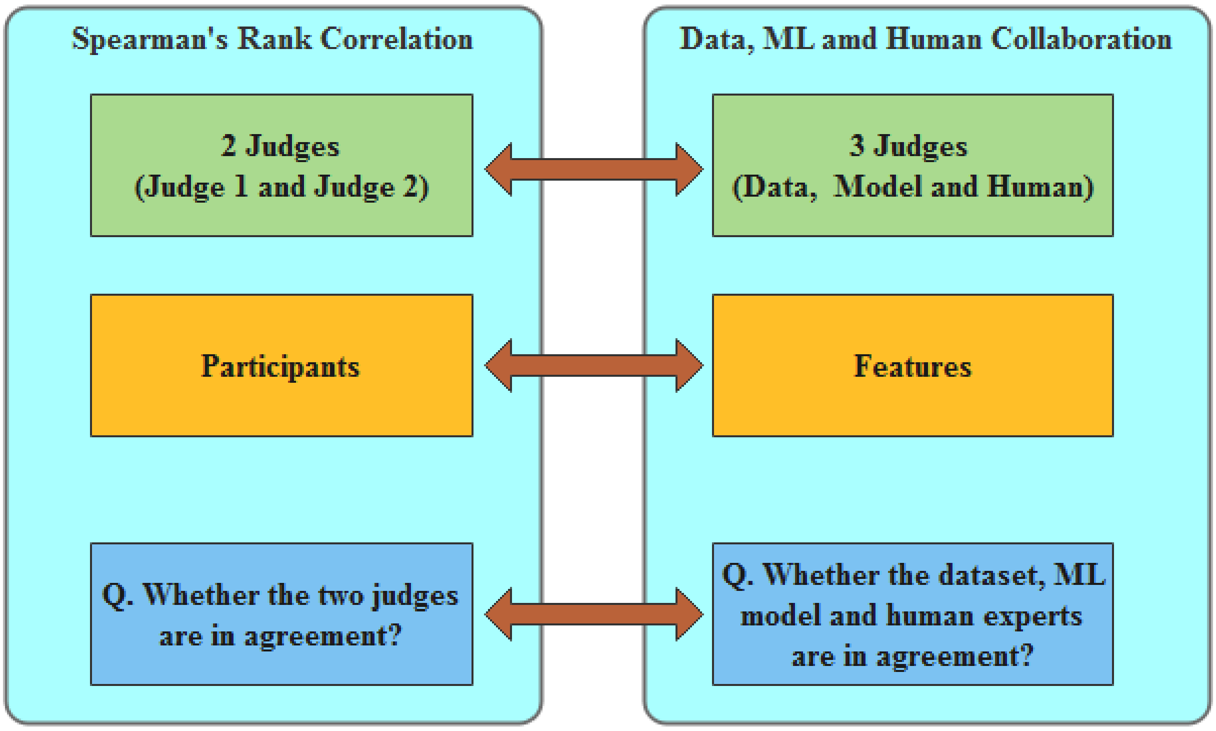

Spearman's Rank Correlation
2 Judges
(Judge 1 and Judge 2)
Participants
Q. Whether the two judges are in agreement?
Data, ML amd Human Collaboration
3 Judges
(Data, Model and Human)
Features
Q. Whether the dataset, ML model and human experts are in agreement?

Figure 3.4: Analogy with Spearman's rank correlation problem

To verify the statistical significance of the degree of agreement among judges, hypothesis testing using 1-tail Spearman's rank correlation test was used. The following three hypotheses were framed:

Hypothesis Test-1: Agreement between Dataset and Model

Null Hypothesis (H_0): Ranking of features as per dataset and ML model is independent ($\rho_s = 0$)

Alternate Hypothesis (H_1): Ranking of features as per dataset and ML model has a positive association ($\rho_s > 0$)

Hypothesis Test-2: Agreement between Model and Human

Null Hypothesis (H_0): Ranking of features as per ML model and human domain experts is independent ($\rho_s = 0$)

Alternate Hypothesis (H_1): Ranking of features as per ML model and human domain experts has a positive association ($\rho_s > 0$)

Hypothesis Test-3: Agreement between Dataset and Human

Null Hypothesis (H_0): Ranking of features as per dataset and human domain experts is independent ($\rho_s = 0$)

Alternate Hypothesis (H_1): Ranking of features as per dataset and human domain experts has a positive association ($\rho_s > 0$)

3.8 Results and Discussion

This section summarizes the outcomes and observations of the experimental work conducted.

3.8.1 Problem domain 1: Predicting joining behavior of freshmen students

Experiment 1: Learning an accurate model

Table 3.6 mentions the performance evaluation metrics computed from the confusion matrix.

Table 3.6: Performance comparison of ML models

Model	Acc		Se		Sp		Precision		F1 Score		AUC	
	Trng	Test	Trng	Test	Trng	Test	Trng	Test	Trng	Test	Trng	Test
LR	0.818	0.81	0.654	0.633	0.911	0.91	0.806	0.8	0.722	0.707	0.862	0.861
NB	0.796	0.79	0.672	0.657	0.867	0.866	0.742	0.735	0.705	0.694	0.849	0.845
CART	0.802	0.789	0.66	0.627	0.883	0.88	0.762	0.748	0.707	0.682	0.793	0.78
RF	0.832	0.816	0.653	0.625	0.933	0.925	0.848	0.825	0.738	0.711	0.834	0.835

Outcome of the experiment: It was observed that all the four classification algorithms gave accuracy in the range of around 78%-80%, a significant improvement over baseline accuracy of 63.8%. All four algorithms competed well in terms of sensitivity, specificity, precision, F1 score and AUC values. RF model has given the best classification accuracy. A good generalization of classification accuracy from training to test data was observed. RF model was taken as input for the next stages of the experimental work.

Experiment 2: Listening to the dataset Table 3.7 mentions the information gain computed using entropy and Gini index for each feature in ’freshmen’ dataset.

Table 3.7: Ranking of features as per dataset

Feature(f)	$IG_{e,f}$	$R_{e,f}$	$IG_{g,f}$	$R_{g,f}$	$R_{Dataset,f}$
ScholarshipBracket	0.197	1	0.125	1	1
MarksCategory	0.116	2	0.067	2	2
HostelorTransport	0.1	3	0.062	3	3
FeePaidCategorized	0.091	4	0.052	4	4
LoanLetter	0.027	5	0.015	5	5
HomeTownType	0.022	6	0.014	6	6
PreviouslyStudied	0.005	7	0.003	7	7
QualifyingExam	0.002	8	0.001	8	8
AdmissionMonth	0.001	9	0.001	9	9

Outcome of the experiment: Each feature has been ranked in descending order of $IG_{e,f}$ and $IG_{g,f}$. The column $R_{Dataset,f}$ is the average of the above two ranks ($R_{e,f}$ and $R_{g,f}$). It was observed that the features giving maximum information gain included 'ScholarshipBracket', 'MarksCategory', 'HostelorTransport', 'FeePaidCategorized' and 'LoanLetter'. The rank assigned to each feature was the same using entropy and Gini index.

Experiment 3: Listening to the model

The RF model was picked for subsequent phases of the experimentation work owing to its high accuracy and black-box nature. Table 3.8 mentions the quantitative measure of importance given to a feature by the model and corresponding ranks are compiled in Table 3.9 for ‘freshmen’ dataset.

Table 3.8: Feature Importance measures

Feature	mmd	acc_d	gini_d	times_root	Imp_f
ScholarshipBracket	1.028	0.125	1147.8	169	1.694
MarksCategory	1.348	0.043	467.8	105	1.198
HostelorTransport	1.758	0.028	356.6	98	1.14
FeePaidCategorized	1.786	0.03	322.5	70	1.123
LoanLetter	2.528	0.008	106.6	38	1.12
HomeTownType	2.634	0.007	130.2	14	1.078
PreviouslyStudied	3.25	0.004	42.4	6	1.066
QualifyingExam	2.598	0.006	109.5	0	1.061
AdmissionMonth	2.596	0.014	247.4	0	1.032

Global surrogate model: Figure 3.5 shows the global surrogate model with depth fixed to two levels for our datasets. Globally top significant features selected for tree construction included 'ScholarshipBracket', 'MarksCategory' and 'FeePaidCategorized'.

Outcome of the experiment: Top features to which model outcome is sensitive includes ‘ScholarshipBracket’, ‘FeePaidCategorized’, ‘HostelorTransport’, and ‘MarksCategory’. The features selected for the construction of the global surrogate model were

Table 3.9: The rank of a feature as per ML model

Feature	$R_{mmd,f}$	$R_{acc\_d,f}$	$R_{gini\_d,f}$	$R_{times\_root,f}$	R_{Imp_f}	$R_{Model,f}$
ScholarshipBracket	1	1	1	1	1	1.0
MarksCategory	2	2	2	2	5	2.6
HostelorTransport	3	4	3	3	3	3.2
FeePaidCategorized	4	3	4	4	2	3.4
LoanLetter	5	6	8	5	8	6.4
HomeTownType	8	7	6	6	6	6.6
PreviouslyStudied	9	9	9	7	9	8.6
QualifyingExam	7	8	7	8.5	7	7.5
AdmissionMonth	6	5	5	8.5	4	5.7

matching closely with the most significant variables found using feature importance.

Experiment 4: Feedback from human domain experts

Each human expert assigned a unique rank from 1 to 9 to the nine features that were used in learning the ML-based model. Rank 1 was assigned to the feature perceived as affecting joining behaviour the most as per that human expert. Successive ranks are assigned in the same order. Figure 3.6 shows a snapshot of a feedback form filled by a human expert.

Table 3.10 compiles the feedback from each human expert against each feature. $R_{Human,f}$ is the average rank assigned to each feature and is computed using equation 3.8.

Table 3.10: Feedback from human experts

Attribute	$R_{h_1,f}$	$R_{h_2,f}$	$R_{h_3,f}$	$R_{h_4,f}$	$R_{h_5,f}$	$R_{h_6,f}$	$R_{h_7,f}$	$R_{h_8,f}$	$R_{h_9,f}$	$R_{h_{10},f}$	$R_{Human,f}$
ScholarshipBracket	1	1	1	1	1	1	1	1	1	1	1.0
MarksCategory	2	3	2	2	2	4	2	6	2	2	2.7
HostelorTransport	6	6	7	8	6	9	3	7	4	4	6.0
FeePaidCategorized	7	5	9	4	9	2	9	5	5	8	6.3
LoanLetter	8	7	5	3	4	5	7	9	3	3	5.4
HomeTownType	9	9	8	6	7	3	8	3	9	6	6.8
PreviouslyStudied	5	8	6	5	5	8	4	4	6	7	5.8
QualifyingExam	3	2	3	7	8	7	5	8	7	9	5.9
AdmissionMonth	4	4	4	9	3	6	6	2	8	5	5.1

Outcome of the experiment: All human experts were having consensus that 'ScholarshipBracket' affects the joining behaviour of freshmen students the most. 'ScholarshipBracket' and 'MarksCategory' were the top two features as per the majority of human experts.

Analysis of Agreement between dataset, model and human experts After collecting rankings of features as per the three judges (dataset, model and human experts), the degree of agreement between these three in terms of ranking of features was computed

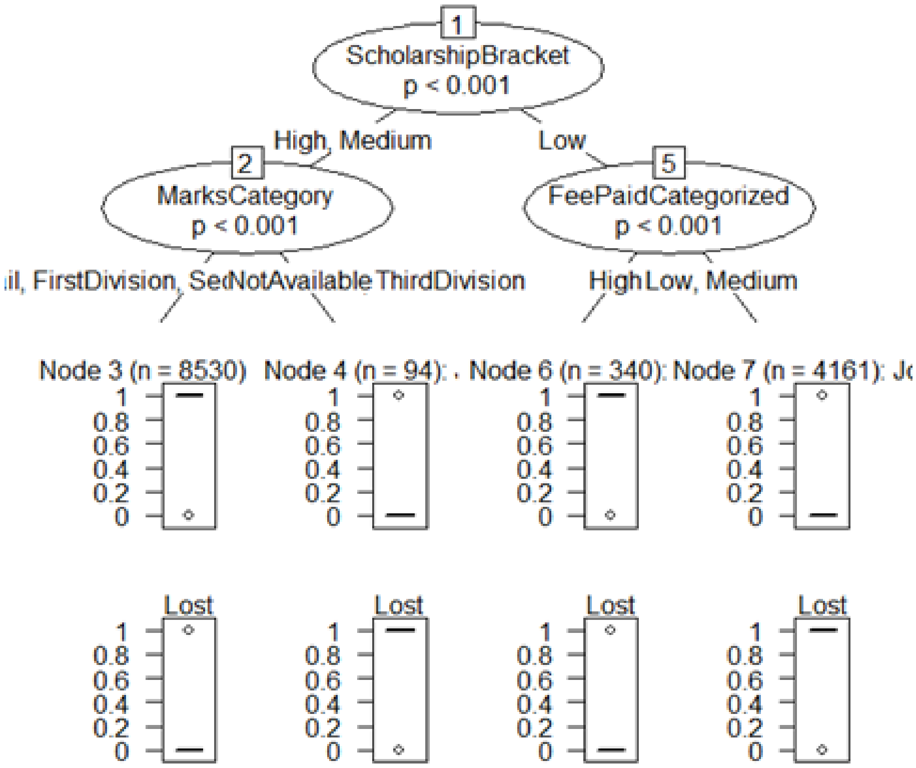


Figure 3.5: Global surrogate model

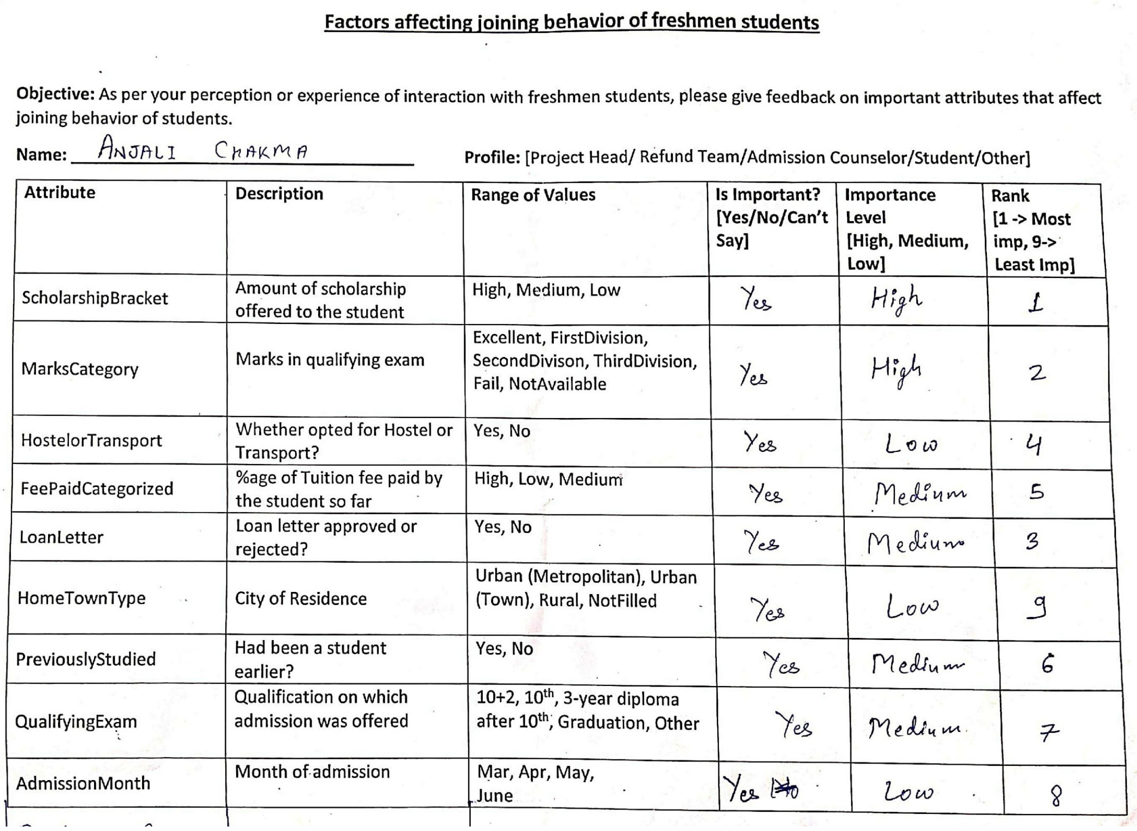
and verified for statistical significance. Table 3.11 gives a comparison of ranks assigned to features based on dataset measures, model behaviour, and human expert feedback.

Table 3.12 compiles the value of Spearman's rank correlation coefficient to measure agreement between data, model and human feedback, taking two at a time. Also, it compiles the outcomes of the three hypothesis tests formulated to verify the statistical significance of the agreement observed. The value of 'n' in our case was 9 as a total of nine features were used for learning the ML model. Referring to the table of critical values for one-tail Spearman's ranked correlation coefficient, the critical value for n=9 was 0.6 at a 5% level of significance.

Outcomes of the experiment:

(i) A positive value of correlation coefficient (test statistic) was observed for each pair of judges. The positive values indicate agreement between dataset measures, model behaviour and human perception.

(ii) As the value of the test statistic was greater than the critical value in each of the three tests, the null hypothesis was rejected for each pair of judges. It indicates that there is a statistically significant positive association between rankings of features by dataset, model and human experts.



Factors affecting joining behavior of freshmen students

Objective: As per your perception or experience of interaction with freshmen students, please give feedback on important attributes that affect joining behavior of students.

Name: ANJALI CHAKMA **Profile:** [Project Head/ Refund Team/Admission Counselor/Student/Other]

Attribute	Description	Range of Values	Is Important? [Yes/No/Can't Say]	Importance Level [High, Medium, Low]	Rank [1 -> Most imp, 9-> Least Imp]
ScholarshipBracket	Amount of scholarship offered to the student	High, Medium, Low	Yes	High	1
MarksCategory	Marks in qualifying exam	Excellent, FirstDivision, SecondDivison, ThirdDivision, Fail, NotAvailable	Yes	High	2
HostelorTransport	Whether opted for Hostel or Transport?	Yes, No	Yes	Low	4
FeePaidCategorized	%age of Tuition fee paid by the student so far	High, Low, Medium	Yes	Medium	5
LoanLetter	Loan letter approved or rejected?	Yes, No	Yes	Medium	3
HomeTownType	City of Residence	Urban (Metropolitan), Urban (Town), Rural, NotFilled	Yes	Low	9
PreviouslyStudied	Had been a student earlier?	Yes, No	Yes	Medium	6
QualifyingExam	Qualification on which admission was offered	10+2, 10th, 3-year diploma after 10th, Graduation, Other	Yes	Medium	7
AdmissionMonth	Month of admission	Mar, Apr, May, June	Yes ~~No~~	Low	8

Figure 3.6: Feedback form for human experts

(iii) The model has been able to extract important features of the dataset into its learning and this learning was also in sync with the prevailing domain knowledge as per human domain experts.

3.8.2 Problem domain 2: Predicting onset of diabetes using diagnostic measures

Experiment 1: Learning an accurate model

Table 3.13 mentions the performance evaluation metrics computed from the confusion matrix for LR, NB, CART and RF algorithms.

Outcome of the experiment: It was observed that all the four classification algorithms gave accuracy in the range of around 75%-79%, a significant improvement over baseline accuracy of 65.1%. All four algorithms competed well in terms of sensitivity, specificity, precision, and AUC values. RF model has given the best classification accuracy for training data and an almost equally good performance using test data. RF model was taken as input for the next stages of the experimental work.

Experiment 2: Listening to the dataset

Table 3.14 mentions the information gain using entropy and Gini index for each

Table 3.11: Ranks assigned by dataset, model and human experts

Feature	$R_{Dataset,f}$	$R_{Model,f}$	$R_{Human,f}$
ScholarshipBracket	1	1	1.0
MarksCategory	2	2.6	2.7
HostelorTransport	3	3.2	6.0
FeePaidCategorized	4	3.4	6.3
LoanLetter	5	6.4	5.4
HomeTownType	6	6.6	6.8
PreviouslyStudied	7	8.6	5.8
QualifyingExam	8	7.5	5.9
AdmissionMonth	9	5.7	5.1

Table 3.12: Degree of agreement and hypothesis testing

Pair of Judges	Hypothesis Test	Test Statistic (TS)	Critical Value (CV)	Statistical Test Outcome
Dataset & Model	Hypothesis Test-1	0.860	0.6	H0 is Rejected
Model & Human	Hypothesis Test-2	0.766	0.6	H0 is Rejected
Dataset & Human	Hypothesis Test-3	0.695	0.6	H0 is Rejected

Table 3.13: Performance comparison of ML models

Model	Acc		Se		Sp		Precision		AUC	
	Trng	Test	Trng	Test	Trng	Test	Trng	Test	Trng	Test
LR	0.773	0.786	0.557	0.567	0.888	0.904	0.727	0.76	0.84	0.831
NB	0.762	0.771	0.607	0.582	0.845	0.872	0.678	0.709	0.824	0.83
CART	0.793	0.745	0.552	0.448	0.923	0.904	0.793	0.714	0.767	0.704
RF	1.000	0.766	1.000	0.552	1.000	0.88	1.000	0.712	0.814	0.814

feature in 'diabetes' dataset and corresponding ranks assigned.

Table 3.14: Ranking of features as per dataset

Feature(f)	$IG_{e,f}$	$R_{e,f}$	$IG_{g,f}$	$R_{g,f}$	$R_{Dataset,f}$
Age	0.038	2.5	0.025	2	2.3
BloodPressure	0.02	5	0.012	5.5	5.3
BMI	0.038	2.5	0.024	3	2.8
DiabetesPedigreeFunction	0.019	6	0.012	5.5	5.8
Glucose	0.162	1	0.095	1	1.0
Insulin	0.016	7	0.01	7	7.0
Pregnancies	0.033	4	0.021	4	4.0
SkinThickness	0.013	8	0.008	8	8.0

Outcome of the experiment: Each feature has been ranked using $IG_{e,f}$ and $IG_{g,f}$ and the column $R_{Dataset,f}$ is the average of these two ranks. It was observed that the features giving maximum information gain included 'Glucose', 'BMI', and 'Age'. The ranks assigned to each feature using entropy and Gini index were very close.

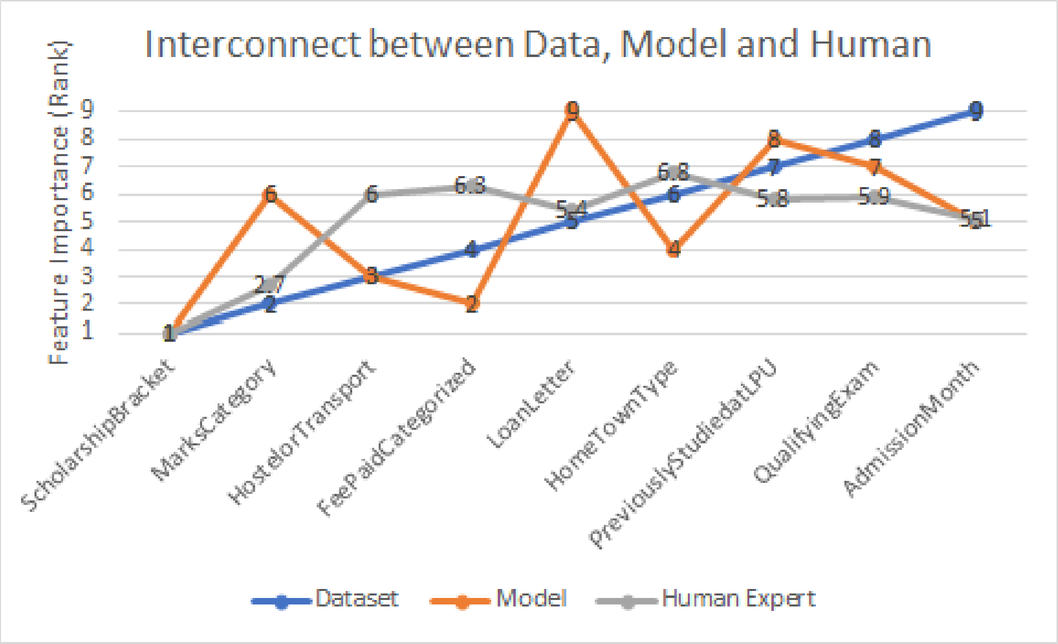


Figure 3.7: Data, Model and Human opinion - problem domain 1

Experiment 3: Listening to the model

The RF model was picked for subsequent phases of the experimentation work owing to its high accuracy and black-box nature. Table 3.15 mentions the quantitative measure of importance given to a feature by the model and corresponding ranks are compiled in Table 3.16 for 'diabetes' dataset.

Table 3.15: Feature Importance measures

Feature	mmd	acc_d	gini_d	times_root	Imp$_f$
Age	0.038	2.5	0.025	2	2.3
BloodPressure	0.02	5	0.012	5.5	5.3
BMI	0.038	2.5	0.024	3	2.8
DiabetesPedigreeFunction	0.019	6	0.012	5.5	5.8
Glucose	0.162	1	0.095	1	1.0
Insulin	0.016	7	0.01	7	7.0
Pregnancies	0.033	4	0.021	4	4.0
SkinThickness	0.013	8	0.008	8	8.0

Outcome of the experiment: Top features to which model outcome is sensitive to included 'Glucose', 'BMI', and 'Age'.

Experiment 4: Feedback from medical domain experts

Each human expert assigned a unique rank from 1 to 8 to the eight features used in learning the ML-based model. Rank 1 was assigned to the feature perceived as affecting the onset of diabetes the most as per that human medical expert. Successive ranks

Table 3.16: The rank of a feature as per ML model

Feature	R$_{mmd,f}$	R$_{acc\_d,f}$	R$_{gini\_d,f}$	R$_{times\_root,f}$	R$_{Model,f}$
Age	3	3	3	2	2.75
BloodPressure	7	8	5	8	7
BMI	2	2	2	3	2.25
DiabetesPedigreeFunction	5	5	4	6	5
Glucose	1	1	1	1	1
Insulin	6	6	7	5	6
Pregnancies	4	4	6	4	4.5
SkinThickness	8	7	8	7	7.5

are assigned in the same order. Table 3.17 compiles the feedback from each of the 5 human medical experts against each feature. Each of these human experts was having experience of more than 5 years. R$_{Human,f}$ is the average rank assigned to each feature and is computed using equation 3.8.

Table 3.17: Feedback from human experts

Attribute	R$_{h_1,f}$	R$_{h_2,f}$	R$_{h_3,f}$	R$_{h_4,f}$	R$_{h_5,f}$	R$_{Human,f}$
Age	2	2	5	5	3	3.4
BloodPressure	4	4	6	6.5	5	5.1
BMI	1	1	2	2.5	2	1.7
DiabetesPedigreeFunction	5	5	1	2.5	6	3.9
Glucose	3	3	3	2.5	1	2.5
Insulin	7	7	4	2.5	4	4.9
Pregnancies	8	8	7	6.5	7	7.3
SkinThickness	6	6	8	8	8	7.2

Outcome of the experiment: The set of top three features were the same as per the model and human experts. However, the internal ranking was not the same. All human experts were having consensus that ‘Glucose’, ‘BMI’ , and ‘Age’ affect the onset of diabetes the most.

Analysis of Agreement between dataset, model and human experts After collecting rankings of features as per the three judges (dataset, model and human experts), the degree of agreement between these three in terms of ranking of features was computed and verified for statistical significance. Table 3.18 gives a comparison of ranks assigned to features based on dataset measures, model behaviour, and human expert feedback.

Table 3.19 compiles the value of Spearman’s rank correlation coefficient to measure agreement between data, model and human feedback, taking two at a time. Also, it compiles the outcomes of three hypothesis tests formulated to verify the statistical significance of the agreement observed. The value of ‘n’ in our case was 8 as a total of eight features were used for learning the ML model. Referring to the table of critical

Table 3.18: Ranks assigned by dataset, model and human experts

Feature	$R_{Dataset,f}$	$R_{Model,f}$	$R_{Human,f}$
Age	2.3	2.75	3.4
BloodPressure	5.3	7	5.1
BMI	2.8	2.25	1.7
DiabetesPedigreeFunction	5.8	5	3.9
Glucose	1.0	1	2.5
Insulin	7.0	6	4.9
Pregnancies	4.0	4.5	7.3
SkinThickness	8.0	7.5	7.2

values for one-tail Spearman's ranked correlation coefficient, the critical value for n=8 was 0.643 at 5% level of significance.

Table 3.19: Degree of agreement and hypothesis testing

Pair of Judges	Hypothesis Test	Test Statistic (TS)	Critical Value (CV)	Statistical Test Outcome
Dataset & Model	Hypothesis Test-1	0.933	0.643	H0 is Rejected
Model & Human	Hypothesis Test-2	0.798	0.643	H0 is Rejected
Dataset & Human	Hypothesis Test-3	0.714	0.643	H0 is Rejected

Outcomes of the experiment:

(i) A positive value of correlation coefficient (test statistic) was observed for each pair of judges. The positive values indicate agreement between dataset measures, model behaviour and human perception.

(ii) As the value of the test statistic was greater than the critical value in each of the three tests, the null hypothesis was rejected for each pair of judges. It indicates that there is a statistically significant positive association between rankings of features by dataset, model and human experts.

(iii) It is concluded that the model has been able to extract important features of the dataset into its learning and this learning is also in sync with the prevailing domain knowledge as per human domain experts.

3.9 Conclusions and Future work

Dataset, model and human domain experts can be considered as the three pillars for constructing an ML-based solution. Listening to each of these pillars helps verify the acquired learning against the provided dataset and prevailing domain knowledge. Agreement between these three pillars regarding their perception of the problem domain facilitates confidence in the ML model in terms of accuracy and user agreement.

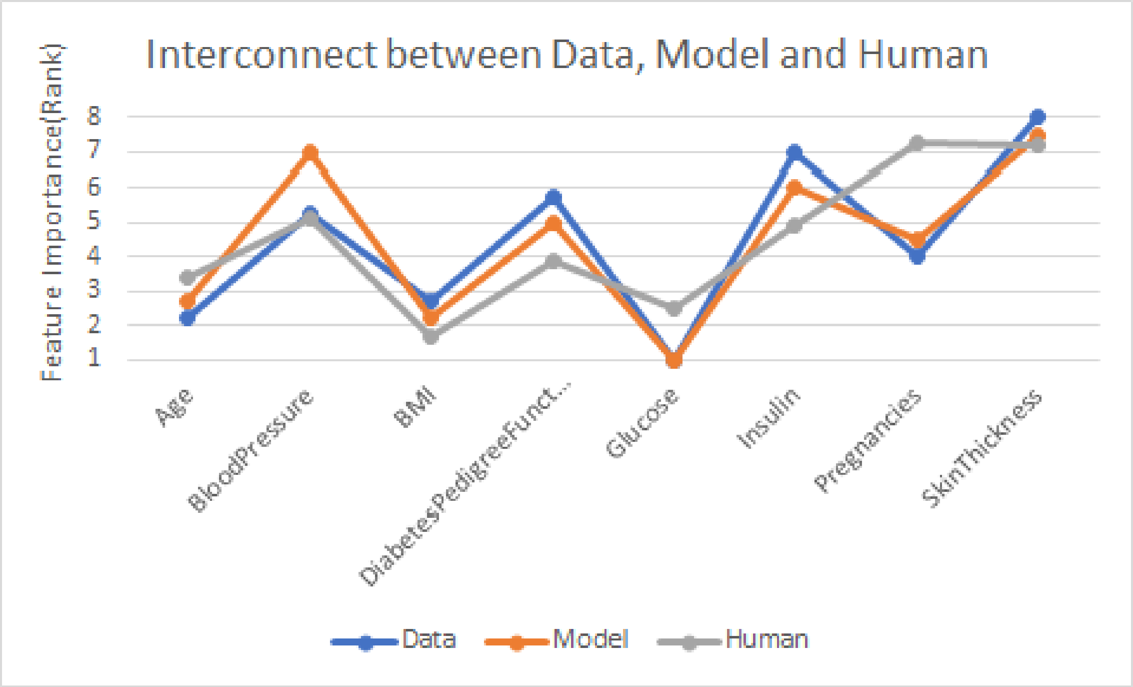


Figure 3.8: Data, Model and Human opinion - problem domain 2

The problem of measuring the degree of agreement between data, model and human can be modelled as a rank correlation problem. A higher degree of agreement between dataset and model indicates that the model has listened well to the dataset. A higher degree of agreement between model and human indicates that the behaviour of the model is in sync with the prevailing domain knowledge. Statistical significance of the agreement can be verified by hypothesis testing using one-tail Spearman's rank correlation coefficient.

The proposed approach has the potential of forming the basis for developing formal quantitative metrics to facilitate trust in ML-based solutions. Information gain computed using entropy and Gini index are useful in identifying features having higher predictive power. Variable importance measures and model-agnostic techniques help diagnose the behaviour of any complex ML model. Human expert feedback is important to validate the behaviour of an ML model. RF algorithm outperformed LR, NB and CART algorithms in terms of classification accuracy.

Future research directions include exploring additional dataset measures and model-agnostic interpretability techniques to listen to the provided dataset and learn the model more comprehensively. For measuring agreement between stakeholders, weighted rank correlation can be explored as an improvement over spearman's rank correlation. Also, the proposed approach can be extended to regression problems and other problem domains.

Chapter 4

Human-feedback Adaptive Learning

4.1 Introduction

Humans and machines have their unique strengths and collaboration between these two have the potential to improve ML systems. For any such collaboration, human users must be able to interpret the behaviour of the ML system. Moreover, human users should have a provision to give feedback to the ML system based on their domain knowledge. Consequently, apart from accuracy, human interpretability and the ability to interact with human experts are the crucial parameters for the success of ML systems.

The solution of an ML problem involves the optimization of one or more internal metrics. The most popular ones among these internal metrics include accuracy, precision, recall, AUC values, and F1-score. Solution space of an ML problem may have, in general, multiple solutions that are equally good in terms of these internal optimization metrics. However, these solutions may differ significantly in terms of their alignment with the user's perspective of the problem domain. Including human domain experts in this exploration of the solution space of a problem has the potential to help search for a solution that not only satisfies the threshold for internal metrics but has improved agreement with the user. Moreover, the involvement of human experts has the potential to accelerate this exploration of solution space, therefore, helping in terms of algorithmic complexity.

During recent years, a renewed interest in conferring interpretability and interact ability (with human users) to ML systems have been observed. However, a lack of consensus regarding definitions, standard practices and evaluation metrics still exists.

The goal in this chapter is to extend our framework to enable human-feedback adaptive learning. The objective is to make the ML model adapt to human expert feedback in case there is a lack of agreement with the human domain experts regarding their perception of the problem domain. The chapter provides a base algorithm for the design of interpretable and interactive ML systems that gives a common starting point for further improvements. The chapter also identifies potential metrics for the evaluation of interpretable and interactive ML systems. Establishing principles and guidelines for the design of interactive ML systems will help in standardization and building a consensus. The chapter drafts a set of desired characteristics for interpretable and interactive ML systems. To take ML to the masses, human users with different expertise levels of the domain must be enabled to interact with and provide feedback to ML systems. The chapter proposes an interface that can capture feedback from human experts with different levels of expertise. This will help make experiments involving human experts more economical and enable leveraging of the masses in improving or verifying ML systems.

The major contributions of this chapter are:

(i) An algorithmic and graphical description of the working of interactive ML systems

(ii) Proposes human-feedback adaptive learning algorithm to enable an ML model to adapt to human expert feedback

(iii) A set of desired characteristics for interpretable and interactive ML systems

(iv) A set of potential evaluation metrics for interpretable and interactive ML systems

These contributions have the potential to positively impact the progress in the field of interpretable and interactive ML systems.

The rest of the chapter is organized as follows: Section 4.2 summarizes the related work and existing research gaps. Section 4.3 describes the problem formulation. Section 4.4 describes the proposed framework using an algorithm and a framework diagram. Section 4.5 describes the experimental work to demonstrate human-feedback adaptive learning. Section 4.6 summarizes the outcomes of the experimental work. Section 4.7 lists desired characteristics of an interactive and interpretable ML system. Section 4.8 proposes metrics that can be used for the evaluation of interactive ML systems. Section 4.9 describes potential improvements to improve the working of interactive ML systems in form of human-feedback adaptive learning. Section 4.10 summarizes conclusions and possible lines for future work.

4.2 Related work

To take ML to the masses, end users must trust ML-based solutions. To facilitate this trust-building, these users should be enabled to interpret the decision-making behavior of an ML model. Interpretable ML is gaining attention due to its potential advantages like facilitating trust, debugging model, ensuring fairness and RTE (Right-to-explanation) obligation (Ribeiro et al., 2016*b*)(Lipton, 2016)(Doshi-Velez and Kim, 2017). A few of the most prominent approaches in conferring interpretability to ML models include Partial dependence plots (Friedman, 2001), Individual conditional expectation plots (Goldstein et al., 2015), feature interaction plots (Friedman et al., 2008), feature importance plot (Breiman, 2001), global surrogate models (Molnar, 2020), local interpretable model-agnostic explanations (LIME) (Ribeiro et al., 2016*b*), and Shapley explanations (Shapley, 1953).

Humans and machines have a distinct set of capabilities that can complement each other via collaboration between the two. Humans are good at making decisions in never-seen-before situations based on their prior experiences. Machines are good at processing a large volume of data without getting tired. A collaboration of humans and machines has the potential to affect the learning process positively in terms of accelerating the exploration of solution space. The idea is to have this search space exploration process guided via human interaction. Such collaboration has other potential advantages like reducing engineering efforts, learning with a lesser number of human experts or training data. Moreover, a collaboration of humans and machines is useful in reaching out to ML solutions that have better user agreement. It involves iterative interaction cycles between the two sides. As human users who are not ML experts are getting access to ML tools, it becomes imperative to enable these users to interpret the decision-making behaviour of the model. Also, they should be enabled to provide feedback to the ML model based on their domain expertise. So, the ML community need to continue improving ML systems in terms of interpretability as well as the ability to interact with human users. Figure 4.1 shows idea of bi-directional interaction between human and machine.

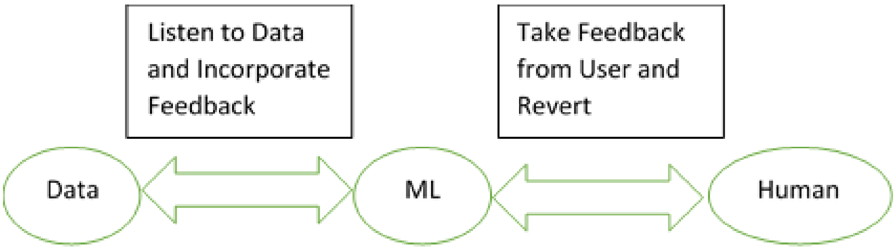

Listen to Data and Incorporate Feedback
Take Feedback from User and Revert
Data
ML
Human

Figure 4.1: Human-Machine Interaction diagram

A hybrid human-machine intelligence approach has been proposed (Yang et al., 2019) where the model is refined to incorporate the feedback given by human experts. A user interface called 'Ruleslearner' has been used to present the learned model as a set of rules in disjunctive normal form and collect feedback from users in the form of addition, deletion, modification, ranking or filtering of rules.

The involvement of humans in the exploration of solution space can positively affect computationally hard problems (Holzinger et al., 2019). The ACO (Ant colony optimization) algorithm was used for solving a TSP (Travelling Salesman Problem). The key idea has been to increase the probability of selection of a path that is traversed by a human, by artificial agents also.

An interactive ML model 'iBCM' that enables a two-way communication (Kim, 2015) with human experts has been proposed for clustering. Human to ML model communication is in terms of prototypes and subspaces. ML model reverts through explanations after incorporating user feedback.

The democratization of ML by providing a bigger role to end-users in the design of ML systems has been advocated (Amershi et al., 2014). Involving users in the learning process results in rapid and specific incremental updates in the ML model. This involvement brings the challenge of understanding the capabilities, behaviour and needs of the end-user.

Interdisciplinary expertise has been advocated to develop explanatory learning models that can provide interpretable and achievable insights in addition to accurate predictions(Rosé et al., 2019). The interdisciplinary expertise spans AI/ML engineers, cognitive, education and UI/UX (User Interface/User Experience) designers.

Research gaps: The associated future lines of work in the design of interactive and interpretable ML-based systems involving human-machine collaboration include:

(i) understanding capabilities and needs of each user

(ii) developing a common language across different domains

(iii) establishing principles and guidelines for the design of interactive ML systems

(iv) developing metrics for the evaluation of interactive ML systems

(v) leveraging human expertise in reducing computation cost of problems that involve the exploration of a large solution space

(vi) utilizing established practices in UI/UX design

(vii) developing innovative and intuitive ideas to accommodate human expert feedback.

4.3 Problem Formulation

This section describes the workflow involved in an interpretable and interactive ML system in the form of an algorithm and flowchart. The objective is to develop a framework that can present an ML model to a human expert in an interpretable manner. Additionally, it can collect feedback from human experts regarding their perception of the problem domain. The proposed framework must be capable of measuring agreement with the human user and validating the statistical significance of this agreement. Moreover, in case the agreement between the ML model and human experts is lacking, it should be able to make the ML model adapt itself to incorporate human feedback. While attempting to incorporate human feedback, the ML model should remain within acceptable limits of accuracy. Whenever there is a conflict between what data and humans say, the same should be reported as a conflict.

First, an accurate ML model is developed using the provided dataset and an appropriate ML algorithm. The behaviour of this ML model is explained to the human domain expert using interpretability techniques like feature importance plot. After presenting the ML model to the human, the system collects the level of agreement between the ML model and the user's perspective of the problem domain. The Likert scale is one commonly used technique to collect user agreement. The choice of the scale can be kept flexible and be decided as per the requirement of the problem domain. A commonly used scale is 1-5 where 1 represents strongly disagree and 5 represents strongly agree. After collecting user agreement, the system prompts the human user to provide feedback, if any, to the ML model in an intuitive and human-friendly manner. A dedicated user interface is used to collect this feedback. This interface provides users with an option to play with existing model behaviour like modifying feature importance or modifying rules. As a next step, the system goes back to incorporate the feedback provided and revert with a refined model to the user. In case, the user feedback is conflicting with what data is saying, the system reverts with a conflict. Again, the system prompts the human user to provide feedback on the revised version of the model or resubmit feedback in case a conflict is reported. This iterative cycle of collecting and incorporating feedback from a human user goes on until the human user is done with providing feedback. After the human user has no more feedback for the ML model, the system collects agreement between the human user and the ML model again. The difference in user agreement before and after human interaction is measured and checked for statistical significance using statistical tests. If the interaction has resulted in an improvement in the user agreement, it is termed as a fruitful interaction else a non-fruitful interaction.

The workflow in an interactive ML system has been presented using algorithm 2 and figure 4.2 shows its graphical representation using a flowchart.

Novelty of the proposed human-feedback adaptive algorithm: The algorithm takes motivation from the intuitive idea that in an ML-based solution, the ML model should agree with the provided dataset as well as the perception of human domain experts regarding the problem domain under study. The agreement with the dataset is

```
   Data: Dataset for learning
   Result: Interpreting model behaviour, Improved user agreement
 1 Learn an accurate ML model using the provided dataset;
 2 Present the ML model to the human expert for interpretation;
 3 Collect the agreement between user's perspective and ML model;
 4 Prompt the human user to provide feedback to the ML model ;
 5 if User feedback conflict with data then
 6     Report as a conflict
 7 else
 8     Incorporate feedback and revert with a revised model
 9 end
10 if User is done with giving feedback then
11     goto 15
12 else
13     goto Step 4
14 end
15 Collect agreement between user and model;
16 Measure the statistical significance of the change in user agreement before and
     after interaction with human;
17 if Change in the agreement is significant then
18     Human interaction is fruitful.
19 else
20     Human interaction is not fruitful.
21 end
```

Algorithm 2: ML system with Human-feedback adaptive learning capability

crucial for verifying the correctness of the learning acquired. The agreement with the perception of human domain experts is crucial for facilitating trust in ML-based solutions. This algorithm also addresses the situation where the learned model lacks agreement with human experts. The algorithm makes the ML model adapt by exploring the solution space again to reach out to its accurate version as well as in agreement with prevailing domain knowledge. The highlight of this algorithm is that it ensures agreement with the dataset as well as human perception. Due to the two-way verification (against the dataset and the human perception), the algorithm is capable of identifying problems in the preparation of the dataset, problems in the learning process, problems in capturing or analyzing human feedback. Also, it is capable of addressing the case of discovering new knowledge when the ML model lacks agreement with human experts despite being correct. This may be useful in problem domains where the phenomenon under investigation is complex and not well understood yet.

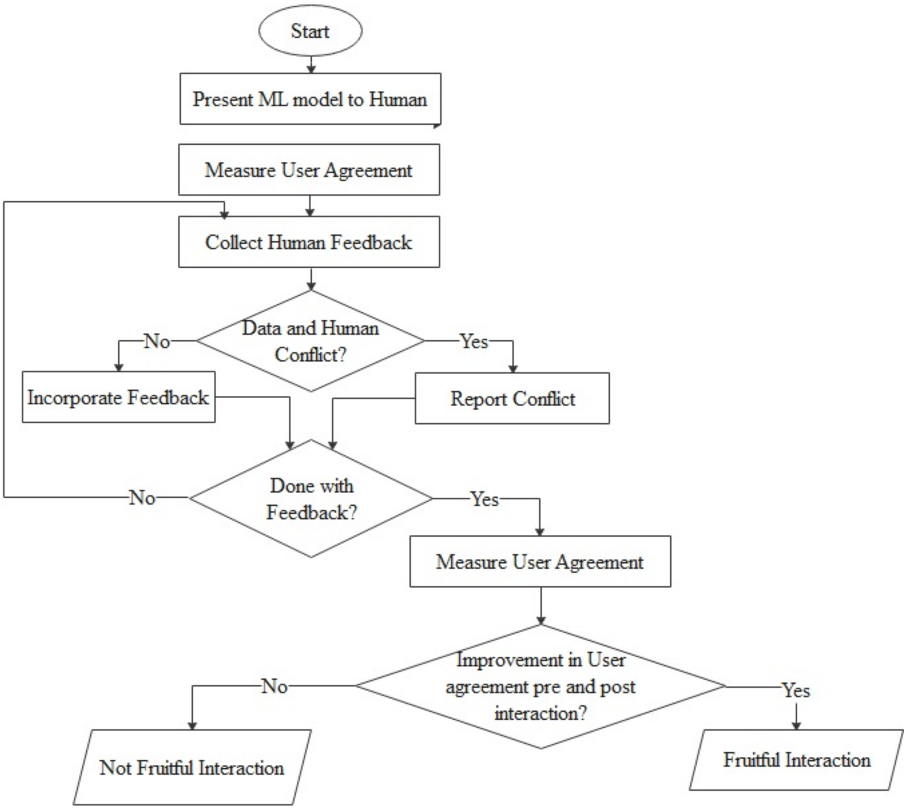


Figure 4.2: Human-feedback adaptive interactive ML system

4.4 Extended framework with human-feedback adaptive learning capability

This section discusses extending the framework to enable human-feedback adaptive learning. The basic idea remains to listen to the dataset, the model and the human domain experts separately in terms of ranking of features. A higher degree of agreement between these three is an indicator that the model had captured important features as per the dataset and its decision-making is in agreement with the prevailing domain knowledge. The degree of agreement between dataset, model and human experts was computed using spearman's rank correlation. If the dataset, the ML model, and human experts do not have a degree of agreement, the model was not considered trustable. In case this agreement is lacking, the ML model attempts to explore the solution space again with a revised strategy to find its version that is better aligned with human users and is still accurate. To measure the degree of agreement, spearman's rank correlation was used. To verify the significance of the degree of agreement, hypothesis testing using one-tail Spearman's rank correlation was used.

The algorithm 3 depicts the flow of the work and the figure 4.3 presents the extended framework graphically with changes highlighted.

These highlighted changes represent the scenario when a lack of agreement between dataset, ML model and human domain experts is observed in terms of their perception regarding the problem domain. In such a case, the ML model attempts to adapt itself to human expert feedback by exploring the solution space again. The revised model is again checked for agreement with the perception of human domain experts. Also, it is ensured that in search of agreement with human perception, the accuracy of the model remains within acceptable limits. These limits are bound to vary from one problem domain to another depending upon the criticality of the decision making involved. In case, the algorithm is not able to find a version of the ML model that is in agreement with the dataset as well as the perception of human experts, it is reported as a conflict.

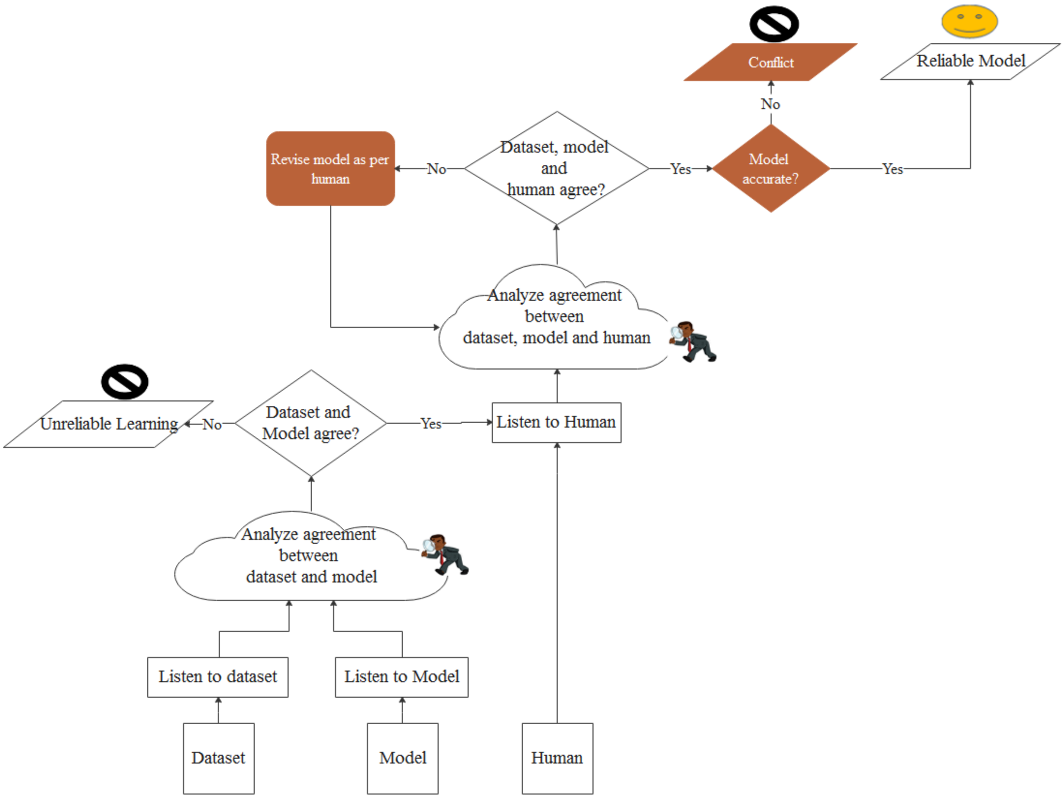


Figure 4.3: Human-feedback adaptive learning framework

4.5 Experimentation

This section describes the experimental work conducted for demonstrating how an ML system can adapt itself to incorporate human expert feedback. The objective is to propose and demonstrate a framework that verifies agreement between the dataset, the ML model and human experts in terms of their perception regarding the problem domain. The framework aims to listen to the dataset, model and human expert and analyze the degree of agreement between these three. In case of a lack of agreement, the framework

```
Data: Dataset, a black-box ML model
Result: Interpreting model behaviour, verifying learning, facilitating trust
FEATURES = Set of features used for learning;
HUMANS = Set of human domain experts;
for each f in FEATURES do
    Compute $IG_{e,f}$ and $IG_{g,f}$;
end
for each f in FEATURES do
    Assign $R_{e,f}$ using $IG_{e,f}$;
    Assign $R_{g,f}$ using $IG_{g,f}$;
end
Compute $R_{dataset,f}$ as average of $R_{e,f}$ and $R_{g,f}$;
for each f in FEATURES do
    Compute $Imp_{f}$;
end
for each f in FEATURES do
    Compute $R_{Model,f}$ using $Imp_{f}$;
end
Analyze the degree of agreement between $R_{dataset,f}$ and $R_{Model,f}$;
if Dataset and Model do not agree then
    Model is unreliable. Exit.
end
for each h in HUMANS do
    for each f in FEATURES do
        Collect $R_{h,f}$ based on human expert perception
    end
end
for each f in FEATURES do
    Compute $R_{Human,f}$ as average of ranks assigned to that feature $R_{h,f}$;
end
Analyze degree of agreement between $R_{dataset,f}$ , $R_{Model,f}$ and $R_{Human,f}$;
if Dataset, Model and Human agree then
    Model is trustworthy
else
    Rebuild Model incorporating human expert feedback.
end
if Dataset, Model and Human agree then
    if Model is still accurate enough then
        Model is trustworthy.
    else
        Report as a conflict.
    end
end
```

Algorithm 3: Human-Feedback Adaptive Learning

Table 4.1: Experiments designed and the underlying objective

Experiment #	Experiment description	Experiment Objective
1	Listening to the provided dataset, learned ML model and human domain experts	Extracting their perception regarding the problem domain in terms of feature importance
2	Measuring and analyzing the degree of agreement between dataset, model, and human expert	To check whether the dataset, model and human experts are in sync in terms of their perception regarding the problem domain
3	Revising ML model to align with user's feedback (if feasible and required)	To achieve improved agreement of ML model with the user's perception regarding the problem domain

makes the ML model revise itself to align with the user's perspective by exploring the solution space again while giving minimum importance to least ranked features as per human domain experts. The experiments designed and the underlying motivations have been compiled in Table 4.1.

4.6 Results and Discussion

This section compiles the experimental outcomes and discusses the observations from these experiments.

As detailed in 3.6, all the four classification algorithms gave accuracy in the range of around 78%-80%. All four algorithms competed well in terms of sensitivity, specificity, precision, F1 score and AUC values. RF model has given the best classification accuracy. Moreover, a good generalization of classification accuracy from training to test data was observed. RF model was taken as input for the next stages of the experimental work.

4.6.1 Listening to Dataset, Model and Human

Table 3.7 mentions the importance of each feature as per the dataset. Each feature has been ranked in descending order of $\mathrm{IG}_{e,f}$ and $\mathrm{IG}_{g,f}$. The column $\mathrm{R}_{Dataset,f}$ is the average of the above two ranks. It was observed that the features giving maximum information gain included 'ScholarshipBracket', 'MarksCategory', 'HostelorTransport', 'FeePaidCategorized' and 'LoanLetter'. The rank assigned to each feature was the same using entropy and Gini index.

Table 4.2 mentions the quantitative measure of importance given to a feature by

Table 4.2: Feature Importance

Feature(f)	Imp$_f$	R$_{Model,f}$
ScholarshipBracket	1.694	1
FeePaidCategorized	1.198	2
HostelorTransport	1.14	3
AdmissionMonth	1.123	4
MarksCategory	1.12	5
HomeTownType	1.078	6
QualifyingExam	1.066	7
LoanLetter	1.061	8
PreviouslyStudied	1.032	9

Table 4.3: Ranks as per dataset, model and human experts

Feature	R$_{Dataset,f}$	R$_{Model,f}$	R$_{Human,f}$
ScholarshipBracket	1	1	1.0
MarksCategory	2	6	2.7
HostelorTransport	3	3	6.0
FeePaidCategorized	4	2	6.3
LoanLetter	5	9	5.4
HomeTownType	6	4	6.8
PreviouslyStudied	7	8	5.8
QualifyingExam	8	7	5.9
AdmissionMonth	9	5	5.1

the model. Top features to which model outcome is sensitive to includes 'ScholarshipBracket', 'FeePaidCategorized' ,'HostelorTransport', and 'MarksCategory'.

Table 3.10 compiles the feedback from each human expert against each feature. Each human expert assigned a unique rank from 1 to 9 to the nine features that are used in learning the ML-based model. Rank 1 is assigned to the feature that is perceived as affecting joining behaviour the most as per that human expert. Successive ranks are assigned in the same order. R$_{Human,f}$ is the average rank assigned to each feature is computed.

4.6.2 Verifying agreement between dataset, model and human experts

After collecting rankings of features as per the three judges (dataset, model and human experts), the degree of agreement between these three in terms of ranking of features was computed and verified for statistical significance. Table 4.3 gives a comparison of ranks assigned to features based on dataset measures, model behaviour, and human expert feedback.

Table 4.4 compiles the value of Spearman's rank correlation coefficient to measure agreement between data, model and human feedback, taking two at a time. Also, it compiles the outcomes of three hypothesis tests formulated to verify the statistical significance of the agreement observed. The value of 'n' in our case was 9 as a total of nine features were used for learning the ML model. Referring to the table of critical values for one-tail Spearman's ranked correlation coefficient, the critical value for n=9 was 0.6 at 5% level of significance.

Table 4.4: Degree of agreement and hypothesis testing

Pair of Judges	Hypothesis Test	Test Statistic (TS)	Critical Value (CV)	Statistical Test Outcome
Dataset & Model	Hypothesis Test-1	0.517	0.6	H0 is Accepted
Model & Human	Hypothesis Test-2	0.456	0.6	H0 is Accepted
Dataset & Human	Hypothesis Test-3	0.695	0.6	H0 is Rejected

Observations:

(i) Dataset and Model: A positive value of correlation coefficient (test statistic) was observed between the ranking of features as per dataset and model, however, the magnitude of correlation was 'moderate' only. As the value of the test statistic was smaller than the critical value, the null hypothesis was accepted. It indicates that rank orders of features as per dataset and model are independent or not in agreement.

(ii) Model and Human Expert: A positive value of correlation coefficient (test statistic) was observed between the ranking of features as per ML model and human experts. However, the magnitude of the correlation was 'moderate' only. As the value of the test statistic was smaller than the critical value, the null hypothesis was accepted. It indicates that rank orders of features as per the ML model and human experts are independent.

(iii) Dataset and Human Expert: A strong positive value of correlation coefficient (test statistic) was observed between the ranking of features as per dataset and human experts. As the value of the test statistic was greater than the critical value, the null hypothesis was rejected. It indicates that there was a statistically significant positive association between rankings of features by Dataset and Human experts.

(iv) As two out of three hypothesis tests resulted in the rejection of the alternate hypothesis, it was concluded that the dataset, model and human experts were not in sync in terms of rank orders assigned to features.

Table 4.5: Revised ranks as per dataset, model and human experts

Feature	$R_{dataset,f}$	$R_{Model,f}$	$R_{Human,f}$
ScholarshipBracket	1	1	1
MarksCategory	2	4	2.7
HostelorTransport	3	3	6
FeePaidCategorized	4	2	6.3
LoanLetter	5	8	5.4
PreviouslyStudied	6	7	5.8
QualifyingExam	7	6	5.9
AdmissionMonth	8	5	5.1

4.6.3 Making ML model adapt to human expert feedback

To achieve a positive degree of association between rank orders by dataset, model and human experts, 'HomeTownType', the least average ranked feature by human experts was removed as a predictor. The ML model was rebuilt using all earlier features except 'HomeTownType'. Table 4.5 compiles the revised ranks assigned to features as per dataset, model and human experts.

Table 4.6 compiles the value of Spearman's rank correlation coefficient taking two judges at a time. Also, it compiles the outcomes of three hypothesis tests to verify the statistical significance of the agreement observed. The value of 'n' was revised to 8 as now a total of eight features were used for learning the revised ML model. Referring to the table of critical values for one-tail Spearman's ranked correlation coefficient, the critical value for n=8 was 0.643 at 5% level of significance.

Table 4.6: Degree of agreement and hypothesis testing after adapting to human feedback

Pair of Judges	Hypothesis Test	Test Statistic (TS)	Critical Value (CV)	Statistical Test Outcome
Dataset & Model	Hypothesis Test-1	0.767	0.643	H0 is Rejected
Model & Human	Hypothesis Test-2	0.688	0.643	H0 is Rejected
Dataset & Human	Hypothesis Test-3	0.795	0.643	H0 is Rejected

Observation(s):

(i) A positive value of the correlation coefficient (test statistic) was observed for each pair of judges. The positive values indicate agreement between dataset measures, model behaviour and human perception.

(ii) As the value of the test statistic was greater than the critical value in each of the three tests, the null hypothesis was rejected for each pair of judges. It indicates that there is a statistically significant positive association between rankings of features by dataset, model and human experts.

(iii)The model has been able to extract important features of the dataset into its learning and this learning is also in sync with the prevailing domain knowledge as per human domain experts.

(iv) Accuracy of the revised model dropped from 0.832 to 0.82 which is a marginal decrease only. Any gain in facilitating trust in the model at the cost of a minimal drop in accuracy might be a worthy trade-off in many problem domains.

4.7 Desired characteristics of an interactive and interpretable ML system

Based on the state-of-the-art and results of the experimental work, the following are the desired characteristics of interpretable and interactive ML systems:

(i) No ML knowledge should be assumed on the part of human users

(ii) Human users are expected to be conversant with data points

(iii) The system should have the ability to present the learned ML model to a human domain expert in an interpretable manner

(iv) The system should support an intuitive medium of interaction with a human user to collect feedback and revert after incorporating the feedback

(v) The system should provide a user interface for collecting feedback from human experts. This interface should be easy to use and support 'what-if' analysis

(vi) The system should have provision to assess the degree to which the ML model agrees with the user's perspective

(vii) The system should have a provision to test and report the significance of the difference in user agreement before and after the interaction cycle.

(viii) The system should be able to report conflicts with explanation wherever user feedback is contradicting what the data says

4.8 Metrics for evaluation of interactive ML systems

Future ML systems shall be expected to be accurate, interpretable and interactive. The field of evaluating ML models in terms of prediction accuracy is quite established. The most popular metrics that are in use include accuracy, precision, recall, F1-score and

AUC values. The evaluation of ML models in terms of interpretability is still evolving. Evaluation metrics for human interpretability that has been proposed include (i) the Size of the explanation (For example, the number of nodes in the tree) (ii) the Rating of explanations by human users. The evaluation of ML models in terms of interact ability is still evolving and there is a lack of established metrics. The following metrics are proposed for evaluating interactive ML systems:

(a) Improvement in user agreement: A natural expectation from an interactive ML system is its capability to align itself with the user's perspective of the problem domain. So, an ML system should be capable of collecting agreement between the ML model and human users in terms of their perception regarding the problem domain. Also, it should be capable of verifying whether the interaction with the human expert has been fruitful or not in terms of improvement in the user agreement.

(i) Collecting user agreement: Agreement of the ML model with the user's perspective of the domain knowledge can be collected using the following techniques:

Likert Scale: Each human expert is asked to rate the ML model in terms of its alignment with the perspective of that human expert regarding the problem domain. The most commonly used range of the Likert scale is 1-5 where 1 represents "Strongly disagree" and 5 represents "Strongly agree". Agreement of the ML model with the user's perspective is collected before and after human interaction.

Spearman's rank correlation coefficient: In this approach, the problem of measuring agreement between the ML model and human experts is modelled as a '2-Judge and n-participants' Spearman's rank correlation problem. The 'n' features used for learning are ranked in terms of importance both by the ML model and human expert. The Spearman's rank correlation is computed between these two vectors of feature ranks. The value of the correlation coefficient indicates the magnitude of agreement between the ML model and human experts.

(ii) Measuring improvement in user agreement: The improvement in user's agreement can be measured using the following alternatives:

Hypothesis Testing: To verify the statistical significance of the change in user agreement before and after interaction of the ML model with human users, hypothesis testing is used. Wilcoxon Signed rank test or Spearman's rank correlation can be used as the statistical significance test. An example of the hypothesis has been framed below:

Null hypothesis (H0): There is no change in the user agreement level before and after interaction

The alternate hypothesis (H1): There is a positive change in user agreement level after interaction

Human Support (HS): It refers to the number of human users feeling that the ML model agrees with their perspective. The number is counted before and after the interaction. A positive change can be taken as an improvement in user agreement after the ML model refined itself to accommodate feedback from human experts. The happiness threshold for a human expert can be kept at >4 (using a 1-5 Likert scale) or can be made available for finetuning depending on the requirement of the problem domain.

Happiness Average (HA): To keep HS independent of the number of human experts used, change in the average user agreement before and after the interaction, is computed. The average is taken of Likert scale values collected.

(b) Accelerated Exploration: Gain in terms of decrease in the computational cost of exploring solution space can be measured to check how much acceleration has been possible as a result of involving a human expert. Comparison is made with the computation cost without involving a human expert.

(c) Economical Learning: In problem domains, having less training data or expensive availability of human experts for evaluating ML systems, iterative human interaction has the potential to make the learning process economical. The economic gain can be measured in terms of the decrease in the number of human experts required or the associated cost.

(d) UI/UX: An interactive ML system must provide an interface through which human users can provide feedback to the system. The desired characteristics of this interface include (i) User-friendliness (ii) Provision of 'Playground' or 'What-if' analysis to foresee impact before submitting feedback to the system, and (iii) Intuitiveness of the medium for taking feedback from human experts.

(e) Conflict Reporting: Whenever human user feedback is different from what the underlying dataset says, the ML system must be capable of detecting and reporting this conflict. An intuitive rule to detect a conflict can be fixing a threshold in terms of prediction accuracy. If incorporating feedback from human user result in a decrease in accuracy below the agreed-upon pre-decided threshold value, it is reported as a conflict. This threshold can be decided after consultation with human domain experts. Additionally, the system should be able to explain this conflict to the human expert so that the human expert can interpret the conflict and resubmit the feedback.

4.9 Other Novel opportunities

This section describes the novel opportunities with the potential to improve the state-of-the-art in the field of interpretable and interactive ML systems. The idea is to improve the agreement of the ML system with the user's perspective by making the ML model adapt to human feedback without losing much on accuracy.

(a) Human-feedback adaptive random forest (HARF) – Weighted Selection of features

The selection of feature subspaces for growing decision trees is a key step in building random forest (RF) models. The most commonly used approach for the selection of feature subspaces has been to randomly select a subset of features. However, this random selection can be replaced by weighted subspace selection where features are selected as per feature weightage provided as an input vector to the RF algorithm. An application of this approach has been to improve the classification performance of RF models in case of high-dimensional data (Xu et al., 2012)(Zhao et al., 2017). In such a scenario, random selection may not be the appropriate choice as many features may not have even decent predictive power.

On the same lines, this approach can be applied to the design of interpretable and interactive ML systems where feature importance is being used as an intuitive medium of collecting feedback from human experts. Using a dedicated user interface, the feature importance as per the user's perspective of the domain knowledge is collected as feedback to the ML model. The ML model then adapts itself to incorporate the feedback provided by the human expert. The key idea is to use feature importance as per user feedback while selecting a feature subspace instead of going for random selection. As a result, the refined version of the ML model will aim to align itself more with the user in terms of importance assigned to the feature.

Mathematical specification of the bootstrapping process for the three variants shown in Figure 4.4 is provided as below:

Random Selection:

$$p(f_i) = p(f_j) \quad \forall\ i, j \in \{1, 2, ..., n\}\ \ and\ \ f \in F \tag{4.1}$$

Weighted Selection:

$$p(f_i) \propto Rank_{\{Human,f\}} \quad \forall\ i \in \{1, 2, ..., n\}\ \ and\ \ f \in F \tag{4.2}$$

Feature Elimination:

$$p(f_i) = p(f_j) \quad \forall i, j \in \{1, 2, ..., n\}\ \ and\ \ f \in F - \{f_l\} \tag{4.3}$$

where,

n = number of features used by the original model

F = Set of all features

f$_i$ = ith feature from set of features 'F'

p(f$_i$) = Probability of an ith feature getting selected during bootstrapping

f_l = least ranked feature from set ‘F’

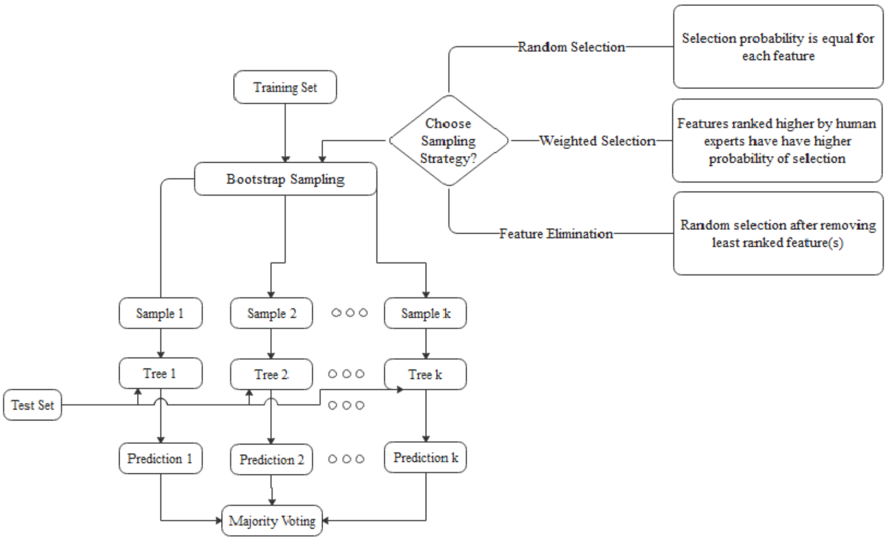

Training Set
Bootstrap Sampling
Choose Sampling Strategy?
Random Selection
Selection probability is equal for each feature
Weighted Selection
Features ranked higher by human experts have have higher probability of selection
Feature Elimination
Random selection after removing least ranked feature(s)
Sample 1
Sample 2
Sample k
Tree 1
Tree 2
Tree k
Test Set
Prediction 1
Prediction 2
Prediction k
Majority Voting

Figure 4.4: Human-feedback adaptive Random forest

(b) User-interface that can adapt to human capability

As different users from the same domain may not have the same level of expertise, there is a need to have a flexible interface for collecting feedback from the human user. For example, in the case of feature importance being used as a medium for collecting user feedback, the following flexibility can help address the issue of different users possessing a different level of expertise in terms of domain knowledge:

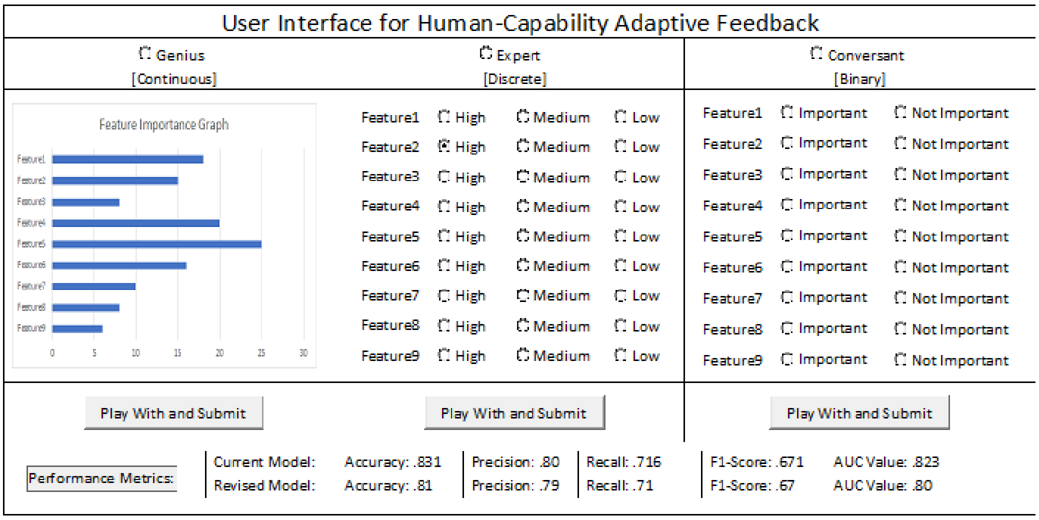
(i) Allowing the human user to play with (increasing or decreasing) the size(width) of the bars showing the importance of each feature.

(ii) Showing feature importance as High, Medium, or Low and allowing the user to play with this labelling in terms of changing labels.

(iii) Showing feature importance as binary values like {Useful, Not Useful} and allowing the user to change this labelling as per his or her perception of the problem domain.

This approach attempts to match experts with varying domain expertise. Users can choose between giving feedback as a continuous value interval, discrete {High, Medium, Low} or even binary {Useful, Not Useful}.

Comparison to existing related work: A collaboration between humans and machines has exciting opportunities to explore due to the unique strengths of each. ’RulesLearner’ an interface to learn a model in the form of rules in disjunctive normal form has been developed (Yang et al., 2019). The interface allows human experts to provide feedback in



User Interface for Human-Capability Adaptive Feedback

○ Genius [Continuous]

○ Expert [Discrete]

○ Conversant [Binary]

Feature1	○ High	○ Medium	○ Low
Feature2	● High	○ Medium	○ Low
Feature3	○ High	○ Medium	○ Low
Feature4	○ High	○ Medium	○ Low
Feature5	○ High	○ Medium	○ Low
Feature6	○ High	○ Medium	○ Low
Feature7	○ High	○ Medium	○ Low
Feature8	○ High	○ Medium	○ Low
Feature9	○ High	○ Medium	○ Low

Feature1	○ Important	○ Not Important
Feature2	○ Important	○ Not Important
Feature3	○ Important	○ Not Important
Feature4	○ Important	○ Not Important
Feature5	○ Important	○ Not Important
Feature6	○ Important	○ Not Important
Feature7	○ Important	○ Not Important
Feature8	○ Important	○ Not Important
Feature9	○ Important	○ Not Important

Performance Metrics:	Current Model:	Accuracy: .831	Precision: .80	Recall: .716	F1-Score: .671	AUC Value: .823
	Revised Model:	Accuracy: .81	Precision: .79	Recall: .71	F1-Score: .67	AUC Value: .80

Figure 4.5: Human-capability adaptive feedback form

the form of modifications to the learned rules. Another related work has been 'iBCM', a model that helps clustering of students assignments in an interactive manner(Kim, 2015). The feedback from human experts has been taken in the form of 'prototypes' and 'subspaces'. Another related work has been to accelerate searching of solution space in computationally hard problems using feedback from human experts in the form of must-be-travelled paths (Holzinger et al., 2019). This input helps the machine reduce the size of the solution space to be explored. In this research work, the feedback from human domain experts has been taken in the form of ranks assigned to features used for learning the ML model. The ranks are assigned based on the experience-based perception of human experts regarding the problem domain. The objective of the collaboration between humans and machines in this work has been to reach out to a version of the ML model that is in agreement with the provided dataset as well as the perception of the human domain experts.

4.10 Conclusions and Future Work

The solution space of an ML problem generally has multiple solutions that are almost equally competitive in terms of accuracy but differ significantly in terms of the user agreement. Interaction of an ML system with human domain experts has the potential to help an ML system reach for a solution with improved user agreement and still have accuracy within acceptable limits. Any gain in terms of the user agreement without resulting in an inaccurate model might be a good trade-off in many problem domains involving critical decision-making and a say of human users. Improved user agreement facilitates developing trust of human users in the ML system.

Accuracy, interpretability and ability to interact with human users are going to be important characteristics and evaluation criteria for future ML systems. ML community need to come up with variants of existing ML algorithms which have established supremacy in terms of accuracy. These variants should have the ability to explain their decision-making process to their users and adapt themselves as per feedback from domain experts aiming for improved user agreement. There is a need to bring a consensus on principles, guidelines and formal evaluation metrics regarding interpretable and interactive ML systems. These systems must take care of the fact that the human users whom these are going to interact with are not ML engineers and may not have the same expertise level regarding the problem domain.

Future lines of this work include: (i) evaluating the proposed framework for human-feedback adaptive learning in other problem domains. (ii) experimental work to evaluate the other variants of the human-feedback adaptive algorithms, and (iii) designing a human-capability adaptive user interface to enable interaction of users with different levels of domain expertise.

Chapter 5

Feature-Importance and Feature-Interactions (FIFI) Graph: A visualization for human interpretability in machine learning

5.1 Introduction

Humans have been inherently good at interpreting visualizations and have been using these visualizations as a medium for explaining a phenomenon or process. Therefore, it is no surprise that the outcome of most of the approaches towards conferring human interpretability to ML models has been a visualization in one or the other form. Examples include showing feature importance as a horizontal bar chart where the width of a bar against a feature indicates its relative importance and highlighting the most contributing pixels in problem domains involving image processing. These visualizations aim to explain ML model behaviour to human users by highlighting the most contributing features.

Most of the existing work has focused on identifying important features to interpret the decision-making behaviour of a model. However, studying interactions between features is also important. It facilitates knowledge discovery and provides new insights into the underlying physical phenomena. Analyzing interactions between features is particularly important in problem domains where the underlying physical phenomena are complex and not yet well understood.

The goal of this chapter is to propose a visualization that can enable interpreting an ML model in terms of important features as well as important interactions among features. The idea is to plot the relative importance of features and relative strength of interactions between features into a single plot. The proposed visualization has been modelled as a network graph. In this graph, each node represents a feature and each edge represents an interaction between a pair of corresponding features. As the proposed visualization is a graph, existing research and software tools for network analysis can be applied for interpreting ML model behaviour, especially, in the case of high-dimensional data sets. The proposed approach has been demonstrated using two problem domains. The first problem is our running example of predicting the joining behaviour of freshmen students using a primary dataset. The second problem is of predicting customer churn using a secondary dataset. Due to its natural analogy, a comparison of the proposed FIFI graph with a knowledge graph has also been made.

The rest of the chapter is organized as follows: Section 5.2 summarizes the related work focusing on existing visualizations for interpreting the behaviour of ML models. Section 5.3 describes the proposed visualization and the steps involved in its construction. Section 5.4 describes the datasets and methods used to implement the FIFI graph. Section 5.5 demonstrates the implementation of the FIFI graph and observations from the results. Section 5.6 draws an analogy of the FIFI graph with the concept of a knowledge graph. Section 5.7 gives conclusions and possible future work along with the potential applications of the FIFI graph.

5.2 Related Work

Due to the advantages associated with the interpretability of ML models, a renewed interest in conferring human interpretability to ML models has been observed among the research community over the past a few years (Ribeiro et al., 2016*a*)(Lipton, 2016)(Doshi-Velez and Kim, 2017). The existing work on conferring human interpretability can be classified into model-specific and model-agnostic approaches. Model-specific methods are approaches that apply to a specific ML model say random forest or artificial neural network. Model-agnostic interpretability methods are independent of the underlying ML model. Model-agnostic approaches offer the advantage of flexibility in terms of choice of which ML model to use. This makes the corresponding visualizations for human interpretability also independent of the underlying ML algorithm (Molnar, 2020). Further, the approaches for conferring human interpretability to ML models can be classified as local or global. The approaches that explain the prediction behaviour for a single instance are classified as local. The approaches that explain the behaviour of the ML model overall, are classified as global. The use of local approaches is important when we are specifically interested in investigating the decision making of the ML model for a particular instance. Global approaches are useful in winning the trust vote of a user by explaining the overall decision-making process of the ML model.

In Table 5.1, a comparison of popular model-agnostic approaches has been made focusing on the objective of each explanation along-with their pros and cons. Most of the visualizations make use of a measure of feature importance. Feature importance is indeed an intuitive way of communicating the behaviour of an ML model to its human users. However, studying the type and magnitude of the interaction between features has the potential to provide new insights into the underlying physical phenomena. It may lead to knowledge discovery especially in problem domains that are not well understood yet.

As studying the individual importance of features and interactions between features has potential advantages, a visualization that supports analysis of them both is worth using. In this chapter, the aim is to propose and demonstrate a visualization in the form of a graph that can represent feature importance as well as feature interactions in a single visualization. A human interpretable explanation to understand ML model behaviour in the form of a graph is an intuitive idea. It has an advantage that research and software tools available on network analysis can be used for interpreting ML models. To the best of our knowledge, there is no such alternative visualization for human interpretability.

5.3 Proposed approach

This section describes how to create a visualization for human interpretability of ML models with the following capabilities:

- Should enable identification of features that are contributing most towards the decision-making process of the model.
- Should enable identification of important interactions between features.
- Should provide control to the user in terms of filtering out important interactions.

Feature importance: The importance of a feature is measured as its sensitivity (in terms of increase in the prediction error) to permuting its feature values (Breiman, 2001). If any such permuting increases the model error, the concerned feature is termed as "Important" as per the model's decision-making behaviour. Otherwise, it is termed as "Unimportant" or not contributing towards the decision-making behaviour of the model.

Feature interaction: A pair of input features in an ML model are termed as "Interacting" if their combined influence on the prediction outcome is beyond their influence. For an interacting pair of features, their combined effect is not simply additive but usually, more complex (Molnar, 2020), "The interaction between two features is the change in the prediction that occurs by varying the features, after having accounted for the individual feature effects" .

Table 5.1: Pros and cons of existing visualizations for human interpretability

Visualization	Scope	Objective	Pros	Cons
Partial Dependence Plot (PDP) (Friedman, 2001)	Local	Shows the marginal effect of a feature on the predicted outcome	Intuitive enough to be understood by a layman. Measures the causal relationship between a feature and an outcome.	Can plot a maximum of two features. Assumes features are uncorrelated.
Individual Conditional Expectation (ICE) Plot (Goldstein et al., 2015)	Global	Plots one line per instance showing how prediction changes with change in feature values	More intuitive compared to PDPs. Has the potential to uncover heterogeneous relationships.	Can plot only one feature meaningfully. Requires features to be uncorrelated.
Feature Interaction plot (Friedman et al., 2008)	Global	To plot the strength of interactions between features. Uses H-Statistic to compute interactions.	H-Statistic has a meaningful interpretation. Detects all kinds of interactions.	Computationally expensive. Estimates will vary from run to run due to the sampling involved.
Feature Importance plot (Breiman, 2001)	Global	To plot the relative importance of features in terms of increase in prediction error when feature values are permuted.	Nice interpretation. Global insight. The use of error ratio instead of error difference makes feature importance comparable.	Needs access to the actual outcome target.
Global Surrogate Model (Molnar, 2020)	Global	Approximating a complex black-box model using a simple interpretable model	Easy to explain to non-ML experts. Using the R-square measure, it is easy to measure how well is our approximation.	Conclusions are about the model, not the data. Difficult to decide cut-off for R-square as a good approximation.
LIME Explanation Ribeiro et al. (2016*b*)	Local	Fitting an interpretable model that is locally faithful, at least.	A better choice where sparse explanations are required. Highly popular due to its sparse nature.	Shows only top contributing features. Does not guarantee perfection in the distribution of contribution.
Local Surrogate Model – Shapley Values (Shapley, 1953)	Local	Plots distribution of contribution of features in a comprehensive manner.	Delivers a full explanation. Suitable in situations that demand explainability as a right. A method with solid theory.	Computationally expensive. Not suitable for Sparse explanations.

FIFI graph is a visualization that depicts the relative importance of features and relative magnitude of interactions between features used in learning an ML model. The visualization is modelled as a graph where each node represents a feature and edge represent the interaction between the corresponding pair of features. The size of a node is made proportional to the importance of the corresponding feature. The width of an edge is made proportional to the magnitude of the interaction between the corresponding pair of features.

Figure 5.1 shows a model FIFI graph consisting of 4 nodes and 6 edges. From this model diagram, it is easy to interpret that ’X1’ and ’X3’ are the features that are relatively more important in decision-making by the model (represented by node size). Also, the interactions between the pairs (X1, X2) and (X2, X3) are relatively strong (represented by edge width). A limitation of this visualization is that increase in computation cost is combinatorial. For an ML model that learns using ‘n’ features, the number of interactions to be computed is given by the equation 5.1.

$$\binom{n}{2} = \frac{n!}{2!(n-2)!} \tag{5.1}$$

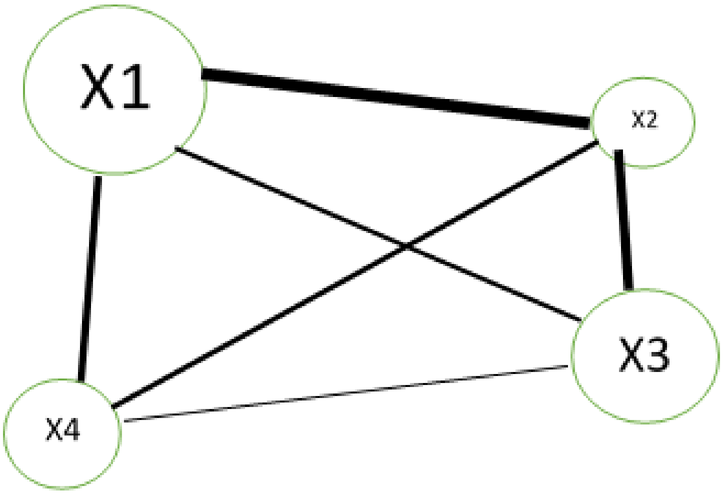


Figure 5.1: A model FIFI graph

5.4 Experimental Setup

This section describes the datasets and methods used for conducting experimental work. The proposed visualization was demonstrated using the following two problem domains:

(i) Predicting joining behaviour of freshmen students (our running example)

(ii) Predicting customer churn using Telco customer-churn(tcc) dataset. The 'tcc' dataset consisted of customer records and is available publicly on Kaggle (Kaggle, 2019). The target variable was 'Churn' with values as 'Yes' or 'No'. The dataset consisted of 7032 observations and 21variables. There were 1869 customers with churn status 'Yes' and 5163 customers with churn status 'No'.

For learning an ML model, Random Forest (RF) algorithm was used. RF algorithm was selected due to two reasons. First, it is a black-box algorithm and lacks human interpretability. Second, software tools are available to compute feature importance and feature interaction for an RF model.

To construct a FIFI graph for our RF model, the following steps were performed for features used in learning the RF model:

(a) Feature importance was computed for each feature

(b) Feature interaction was computed for each pair of features

(c) A visualization in the form of a graph was plotted where nodes represent features and edges represent interactions

(d) Node size was made proportionate to feature-importance

(e) Edge width was made proportionate to strength of feature-interaction

(f) To improve interpretability, a sparse version of the graph was created by leaving out interactions with strength lesser than the mean interaction strength

The implementation was done using 'iml' and 'igraph' packages in R, an open-source statistical software tool (Molnar et al., 2018)(Csardi et al., 2006).

5.5 Results and Discussion

This section compiles the results of the experimental work and associated observations.

Performance of the RF model: Table 5.2 compiles baseline accuracy, prediction accuracy and Area under the ROC curve value for 'tcc' and 'freshmen' datasets. Baseline accuracy is prediction accuracy if the majority class is always predicted by an ML model. It is useful as a benchmark to evaluate the worthiness of employing ML in a problem domain.

As observed from Table 5.2, the RF algorithm achieved a significant improvement in accuracy for both the datasets. ML was found worth applying in both the problem domains. Also, AUC values above 0.8 were reported for both datasets.

Table 5.2: Performance measures for RF Model

Dataset	Baseline accuracy	Accuracy	AUC Value
freshmen	0.638	0.816	0.832
tcc	0.734	0.791	0.822

FIFI graph of RF model for 'freshmen' dataset:

Table 5.3 represents feature importance as per our RF model for 'freshmen' dataset. Table 5.4 mentions the interaction strength of each feature with each of the other features except itself. Figure 5.2 shows the FIFI graph for the RF model developed for 'freshmen' dataset and Figure 5.3 shows a sparse variant of the FIFI graph. It was obtained after deleting edges whose interaction strength is lesser than the mean interaction strength overall

Table 5.3: Feature-Importance for 'freshmen' dataset

id	featureName	featureImportance
f01	ScholarshipBracket	1.5608
f02	MarksCategory	1.1672
f03	HostelorTransport	1.1033
f04	FeePaidCategorized	1.1234
f05	LoanLetter	1.0949
f06	HomeTownType	1.0464
f07	PreviouslyStudied	1.0334
f08	QualifyingExam	1.0259
f09	AdmissionMonth	1.0145

Table 5.4: Feature-Interactions for 'freshmen' dataset

	f01	f02	f03	f04	f05	f06	f07	f08	f09
f01	0.0000	0.2674	0.3555	0.3999	0.3503	0.2972	0.1436	0.2027	0.0928
f02	0.4105	0.0000	0.4567	0.2115	0.2169	0.3134	0.3205	0.3076	0.2877
f03	0.2622	0.5093	0.0000	0.4320	0.5212	0.1787	0.0799	0.3476	0.0497
f04	0.3360	0.2506	0.4773	0.0000	0.3585	0.1671	0.1771	0.2230	0.0635
f05	0.2958	0.3099	0.4939	0.3414	0.0000	0.2048	0.1195	0.0915	0.0838
f06	0.1963	0.3121	0.1915	0.1927	0.2386	0.0000	0.2787	0.1902	0.2881
f07	0.1381	0.3892	0.0847	0.2102	0.1364	0.2553	0.0000	0.1537	0.3363
f08	0.3243	0.3030	0.2568	0.1710	0.0934	0.1885	0.1339	0.0000	0.0701
f09	0.0601	0.2144	0.0496	0.0534	0.0810	0.3413	0.2586	0.0843	0.0000

Observation: It was observed that 'ScholarshipBracket', 'MarksCategory', and 'HostelorTransport' features contribute more in the decision-making process of the model. Talking about interactions between features, the pairs ('HostelorTransport', 'LoanLetter'), and ('HostelorTransport', 'MarksCategory') were observed to have rela-

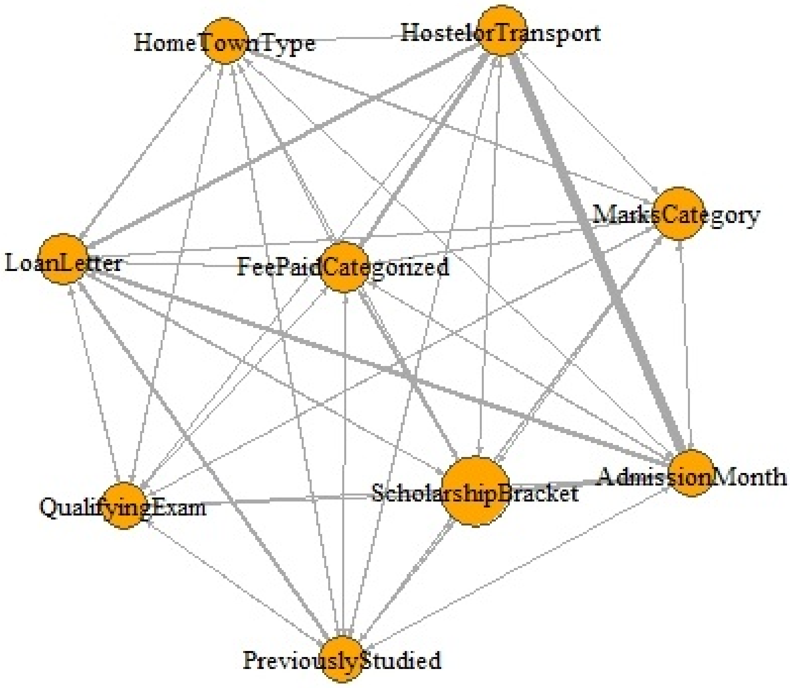


Figure 5.2: FIFI graph for 'freshmen' dataset

tively strong interactions.

FIFI graph of RF model for 'tcc' dataset:

Table 5.5 compiles feature importance as per our RF model for 'tcc' dataset. The features are listed in descending order of importance. As each of these features is going to be a node in the FIFI graph, a unique ID was assigned to each feature. Table 5.6 mentions the interaction strength of each feature with each of the other features except itself. Figure 5.4 shows the FIFI graph for the RF model developed for the 'tcc' dataset. Figure 5.5 shows a sparse variant of the FIFI graph. It was obtained after deleting edges with interaction strength lesser than the mean interaction strength overall.

Table 5.5: Feature-Importance for 'tcc' dataset

id	featureName	featureImportance
f01	tenure	1.9826
f02	MonthlyCharges	1.6380
f03	Contract	1.4963
f04	OnlineSecurity	1.4427
f05	TechSupport	1.4164
f06	PaymentMethod	1.3955
f07	OnlineBackup	1.2471
f08	InternetService	1.2864
f09	DeviceProtection	1.2447

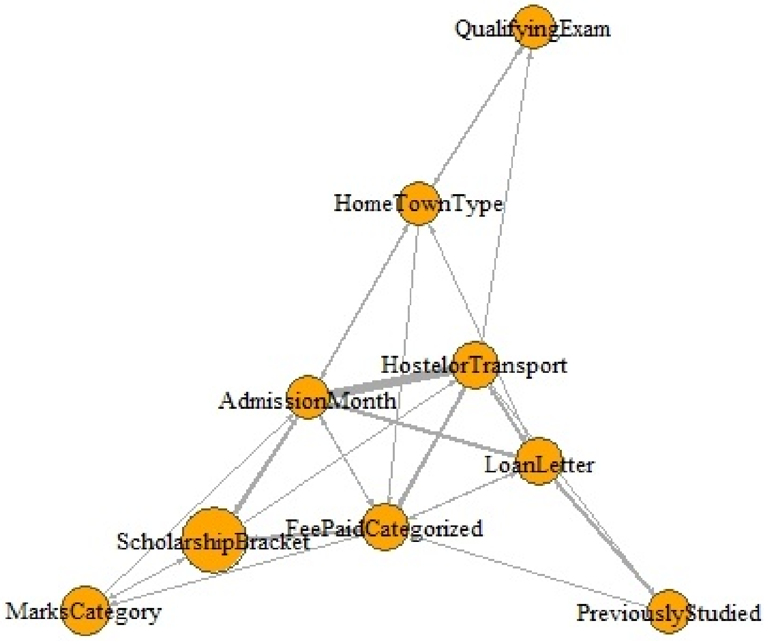


Figure 5.3: Sparse FIFI graph for ’freshmen’ dataset

Observation: It was observed that ‘tenure’, ‘MonthlyCharges’, and ‘Contract’ features contribute more to the decision-making process of the model. Talking about interactions between features, the pairs (‘Contract’, ‘TechSupport’), and (‘Contract’, ‘MonthlyCharges’) were observed to have relatively strong interactions.

5.5.1 A comparative analysis

Table 5.1 in section 5.2 gives a comparative analysis of the existing visualizations for interpreting ML models. In this section, a comparative analysis on the same lines is made for a FIFI graph. A FIFI graph has a global scope, unlike PDP, LIME and Shapley

Table 5.6: Feature-Interactions for ‘tcc’ dataset

	f01	f02	f03	f04	f05	f06	f07	f08	f09
f01	0.0000	0.2674	0.3555	0.3999	0.3503	0.2972	0.1436	0.2027	0.0928
f02	0.4105	0.0000	0.4567	0.2115	0.2169	0.3134	0.3205	0.3076	0.2877
f03	0.2622	0.5093	0.0000	0.4320	0.5212	0.1787	0.0799	0.3476	0.0497
f04	0.3360	0.2506	0.4773	0.0000	0.3585	0.1671	0.1771	0.2230	0.0635
f05	0.2958	0.3099	0.4939	0.3414	0.0000	0.2048	0.1195	0.0915	0.0838
f06	0.1963	0.3121	0.1915	0.1927	0.2386	0.0000	0.2787	0.1902	0.2881
f07	0.1381	0.3892	0.0847	0.2102	0.1364	0.2553	0.0000	0.1537	0.3363
f08	0.3243	0.3030	0.2568	0.1710	0.0934	0.1885	0.1339	0.0000	0.0701
f09	0.0601	0.2144	0.0496	0.0534	0.0810	0.3413	0.2586	0.0843	0.0000

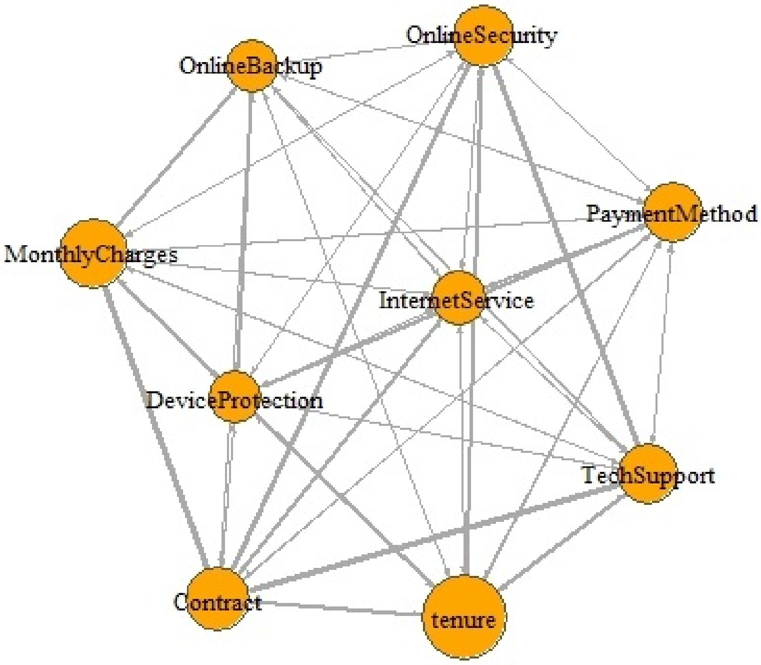


Figure 5.4: FIFI graph for 'tcc' dataset

explanations as it presents the overall behaviour of a model. Also, a FIFI graph can present the importance of all features used for learning instead of one feature at a time like in the case of PDPs and ICEs. A FIFI graph is a compact alternative as it is capable of presenting feature importance as well as interactions between features using a single visualization. Also, being a graph object, it offers an option to derive a sparse version highlighting important features or important interactions only. In that sense, it provides advantages of both LIME as well as Shapley-valued based explanations. However, it is also computationally expensive due to the computing of feature interactions involved.

5.6 FIFI graph Versus a Knowledge-graph

A knowledge graph is a way to organize and retrieve information in the form of a graph. It is used by search engines to improve their search results in response to user queries. Social networking sites and e-commerce sites are also using a knowledge graph to store and retrieve useful information. In a knowledge graph, nodes represent real-world entities and edges represent the existence of a relationship between corresponding nodes(entities). Edge labels specify the type of relationship. Figure 5.6 shows a sample knowledge graph (Nickel et al., 2015). As both FIFI graph and a Knowledge-graph are a representation of knowledge, it is worth making a comparison between the two. Table 5.7 makes a tabular comparison between these two graphical structures.

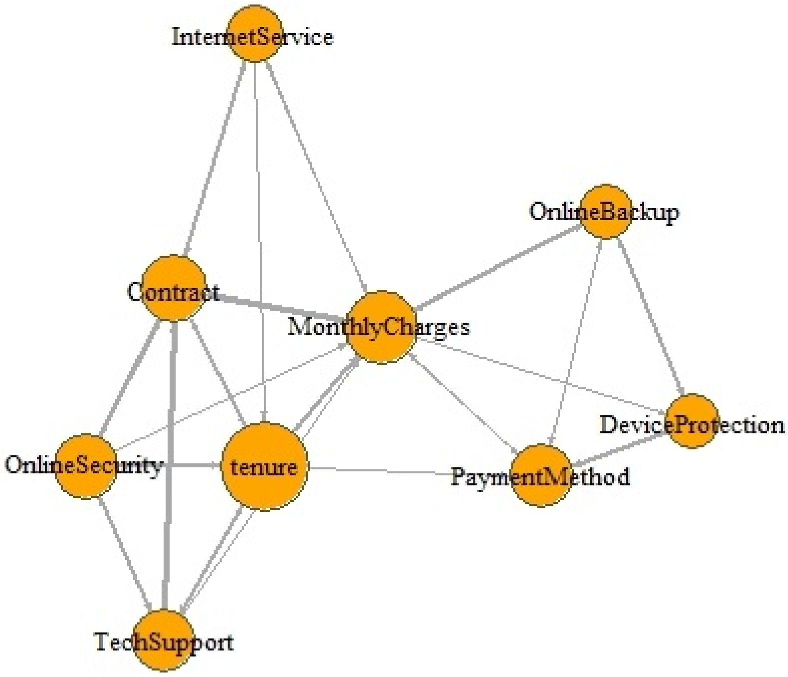


Figure 5.5: Sparse FIFI graph for 'tcc' dataset

5.7 Conclusions and Future work

Feature-importance and Feature-interactions can be plotted into a single visualization in the form of a FIFI graph, making it a compact alternative to existing visualizations. It is easy to interpret as the size of a node is proportional to the importance of a feature and the width of an edge is proportional to the magnitude of the interaction between the pair of corresponding features. Modelling this visualization as a graph brings an advantage that existing theory and tools for network analysis can be employed for analyzing it. A sparse version of a FIFI graph is useful in improving interpretability as most of the time, the user is interested in identifying top features and top interactions only. Being a graphical structure, a FIFI graph has a natural analogy with the concept of a knowledge graph.

A FIFI graph has several potential applications. It is useful in interpreting which features are affecting the decision-making process of the underlying ML model the most and which are the prominent interactions between features. Using an interactive version of a FIFI graph or a customized user interface, feedback from human experts can be taken in terms of feature importance as well as feature interaction by allowing the users to play with the size of a node and width of an edge.

Future directions of this work include additional controls that can be given in the hands of end-user like interactively filtering important features. Another variant of a

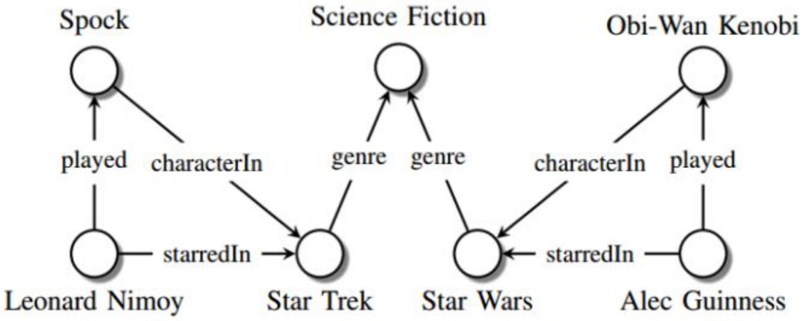


Figure 5.6: A sample knowledge graph

FIFI graph can be created where feature-importance is replaced by 1-way feature interaction. Such a visualization will be useful in identifying the most interacting features and then zooming out which of its interactions are relatively stronger. Another possible exploration is using a FIFI graph and a knowledge graph in a complementary manner.

Table 5.7: FIFI graph Versus Knowledge graph

S. No.	FIFI graph	Knowledge graph
1	Nodes represent features used for learning an ML model.	Nodes represent real-world entities.
2	Scope of nodes(features) is limited to features expected to contribute towards a target variable.	Scope of nodes(entities) is global as any association may be found useful in some search query
3	An edge represents the presence of interaction between features used in learning an ML model.	An edge represents an association between two real-world entities in the form of relationships.
4	Node size and edge width are proportional to feature importance and feature interaction strength respectively.	Node size and edge width do not vary. However, making these two proportionate has the potential to enable comparison of the strength of associations and the importance of entities associated with the entity of interest.
5	Representation of interactions (maybe known as well as unknown) as per ML model.	Representation of existing knowledge in a form that machine can store and retrieve.
6	Reliability of a FIFI graph depends on the quality of the ML process	Makes use of existing published or commercially available information
7	The objective is to enable human users to interpret ML model behaviour.	The objective is to improve the quality and justify search results of a search engine
8	Construction of a FIFI graph requires an ML model. Domain experts may be or may not be involved in the learning process.	Construction of a Knowledge graph requires domain experts, data interlinking and ML algorithms

Chapter 6

Summary and Conclusions

In this thesis, the goal has been to develop a framework that can enable verification of the learning acquired by an ML model against the provided dataset and prevailing domain knowledge. The framework has been demonstrated to be capable of listening to the dataset, the ML model and human domain experts. After listening to these three, the framework measures the degree of agreement between them regarding their perception of the problem domain in terms of feature importance and reports the statistical significance of this agreement. In case, the agreement of the ML model with the human user is lacking, the framework attempts to make the ML model adapt itself to incorporate human expert feedback. FIFI graph, a graph-based visualization, has also been demonstrated for interpreting ML models by analyzing important features and important interactions between features using a single visualization. In chapter 2, the state-of-the-art in the field of conferring human interpretability to ML models has been reviewed. A set of ten research questions were formulated to keep the review guided. The findings from the review have been presented in the form of answers to these formulated research questions. In chapter 3, a framework that has the provided dataset, the learned ML model and human expert feedback as to its three pillars, has been demonstrated. This framework aimed to listen to each of these three pillars in terms of the importance of features regarding the problem domain. After listening to each of these pillars, the framework verifies whether these three are in sync with each other in terms of their perception of the problem domain. In chapter 4, the framework demonstrated in chapter 3 was enhanced to enable human-feedback adaptive learning in case the ML model is not in agreement with human experts. The proposed framework attempts to make the ML model adapt itself to incorporate human expert feedback without losing much on the accuracy front. The idea has been to relearn a solution by eliminating features that were considered least important by human domain experts. This chapter also proposed a set of desired characteristics for the design of interpretable and interactive ML systems. Novel opportunities in human-feedback adaptive learning and evaluation metrics for interactive ML systems were also proposed. An interface for human-capability adaptive feedback has also been discussed that can allow human users with varying levels of domain expertise to interact with an ML system. In chapter 5, FIFI graph, a novel visualization has been proposed and demonstrated. This graph is capable of presenting feature-importance and feature interactions using a single plot. As this visualization has been modelled as a graph object, the existing tools and software for network analysis can be applied to analyze this visualization.

6.1 Conclusions

Motivations for conferring human interpretability include facilitating trust, providing new insights, scrutinizing ability, ensuring fairness and right-to-explanation situations. As per the new Global Data Protection Regulation (GDPR), with effect from May 2018, human subjects that are going to be affected by outcomes of an ML-based solution have the right to ask for an explanation leading to that particular outcome.

A good explanation of an ML model should be interpretable, easy to understand and should be in terms of concepts known to the target audience. It may require mapping of features to concepts known to the user and clearly defined requirements for input and goals of the output.

Model-agnostic approaches, approaches for interpreting tree ensembles, approaches for interpreting neural networks and human-machine interaction are the broad categories into which the existing work towards conferring interpretability can be broadly classified.

The key ideas used for conferring interpretability has been quite intuitive. These included interpreting a complex model using its interpretable approximation, explaining an instance-specific outcome by identifying influential features, explaining an outcome in terms of feature-wise contribution, computing input gradient and using programs snippets as explanations. Tree ensembles have been interpreted by extracting an interpretable decision tree from a population of trees and finding prototypes in tree space. Rulesets have been used for extracting interpretable features and summarizing results. Visual analytics has been used for modelling input-output relationships and making outcomes of a black-box model more expressive.

Human interpretability in ML is an evolving field and there is still a lack of formal metrics. Key ideas used for evaluating interpretability include measuring the complexity of the explanation in terms of the volume of information to be comprehended and ease of identifying how a data point is to be moved to change its label. Evaluation of the quality of explanations by human experts is another measure of interpretability. Can these explanations enable human subjects in choosing the best classifier, improving an untrustworthy classifier, and find new insights? The other ideas include feature-coverage (how many outcomes are explained using only top-ranked features), MEP (Mean Explainability Precision) and MER (Mean Explainability Recall) using explainability score, and quantifying interpretability using information entropy instead of just count of features.

Interpretability is useful in situations that involve critical decision making and a say of human domain experts. It is not compulsory always and rather can be harmful in certain contexts. Possible dangers of transparency include (i) divergence between intended audience and actual beneficiary (ii) transparency in government use of algorithms (iii) gaming of rules and lack of motivation of intellectual property if all algorithms are open source (iv) discrimination of sub-groups based on sensitive features.

Humans and machine have their unique strengths and a collaboration between the two can help incorporate domain expertise into the learning of ML systems. A human-machine collaboration brings potential advantages of an improved user agreement, accelerated exploration of solution space, providing a bigger role to end-users in the design of interactive and interpretable ML systems, and incorporating multi-disciplinary expertise. The key idea is to present the learned model to human domain experts in an interpretable manner. The human experts in turn provide feedback back to the ML system based on their experience-based perception of the problem domain. The ML system attempts to adapt itself to incorporate the feedback provided by human experts. In case there is a contradiction between what a human expert says and what data says, the situation is reported as a conflict.

The provided dataset, the ML model learned using the provided dataset and feedback from human domain experts are the three pillars for constructing an ML-based solution. An ML-based solution is expected to have incorporated important characteristics of the provided dataset and should also be in sync with human domain experts in terms of their perception regarding the problem domain. Listening to each of these three pillars and measuring the degree of agreement between them is a way to verify the acquired learning against the provided dataset and prevailing domain knowledge. A collaboration between these three helps to verify the learning acquired by the ML model and facilitate the trust of human users in an ML-based solution.

Information gain computed using entropy and Gini index are useful in identifying features having higher predictive power. Variable importance measures and model-agnostic techniques help diagnose the behaviour of a complex ML model. Human expert feedback is important to validate the behaviour of an ML model.

The problem of measuring the degree of agreement can be modelled as a rank correlation problem where features are considered as participants and the above three pillars (data, model and human) are considered as Judges. A higher degree of the agreement indicates that the model has listened to what the dataset is saying and is in sync with the human expert's perspective of the prevailing domain knowledge. The statistical significance of this agreement can be verified by hypothesis testing using one-tail Spearman's rank correlation coefficient.

The approach of developing a framework with the dataset, ML model and human expert feedback as to its pillars and making these three collaborate is quantitative and has the potential of forming the basis for developing formal quantitative metrics to facilitate trust in ML-based solutions.

Solution space of an ML problem generally has multiple solutions that have accuracy within acceptable limits but differ significantly in terms of agreement with the human user's perception regarding the problem domain under study. Interaction of an ML system with human domain experts has the potential to help an ML system reach out solutions with improved user agreement and acceptable accuracy. Any gain in terms of agreement with a human user, if feasible without resulting in an inaccurate model, is a good trade-off in many of the problem domains. Improved user agreement facilitates developing trust of human users in the ML system.

Feature-importance and Feature-interactions can be plotted into a single visualization in the form of a FIFI graph. It is a compact alternative to visualizing n+1 graphs (a feature interaction graph for each of the 'n' features and a graph for feature importance). Modelling a FIFI graph as a graph object provides the advantage that existing theory and tools for analysis of graphs can be employed for analyzing it. Sparse versions are useful in improving interpreting of the FIFI graph as most of the time, the user is interested in identifying top features and top interactions only instead of scanning all interactions exhaustively. Being a graphical structure, a FIFI graph has an analogy with the concept of a knowledge graph.

Using a FIFI graph, stakeholders of an ML solution can interpret which features are affecting the decision-making process of the underlying ML model the most. Also, it is useful in identifying prominent interactions between features. A FIFI graph can be used in the design of interactive and interpretable ML systems where feedback from human users is part of the ML learning process. Using an interactive version of a FIFI graph or a customized user interface, feedback from human experts can be taken in terms of feature importance as well as feature interaction by allowing the users to play with the size of a node and width of an edge.

Accuracy, interpretability and ability to interact with human users are going to be important characteristics and evaluation criteria for future ML systems. ML community need to come up with variants of existing ML algorithms which have established supremacy in terms of accuracy. These variants should have the ability to explain their decision-making process to their users and adapt themselves as per feedback from domain experts. These systems must take care of the fact that the human users whom these are going to interact with are not ML engineers and do not have the same expertise level about the problem domain.

The field of interpretable and interactive ML systems is an evolving field and there is a need to bring a consensus on principles, guidelines and formal evaluation metrics.

6.1.1 Scope and limitations of this research work

Scope

- Easy to understand and intuitive
- Quantitative approach
- Relies on well-established concept of hypothesis testing for measuring user agreement
- Useful in situations where user agreement is also critical in addition to accuracy
- Existing research on graph theory can be applied to the FIFI graph to identify important features as well as interactions.

Limitations

- Requirement of human domain experts
- Selection and orientation of the human experts is crucial to the success of this approach
- In problem domains like Image processing, feedback of human experts can not be taken in terms of the importance of pixels. Need to take feedback in terms of human-friendly concepts using techniques like concept activation vectors(CAVs).

6.2 Future research

This research work has several opportunities as the possible lines for future work:

Listening to the three pillars

To extract important characteristics of the provided dataset, alternatives of entropy and Gini index as information gain measures can be explored. One alternative measure that can be explored is the Gain ratio. In this work, feature importance measures have been used to interpret ML model behaviour. There is a scope of exploring other human interpretability methods for listening to the ML model better. Examples of other interpretability methods include interpretable decision trees and surrogate models. Moreover, there are multiple ways of computing feature importance. It will be interesting to evaluate these alternatives in terms of their efficacy and ease of use. Feature importance is quite an intuitive idea of collecting the perception of human experts regarding the problem domain. Alternatives other than the ranking of features in terms of their importance can be used to collect human expert feedback regarding the problem domain. These alternatives may be evaluated in terms of ease of use by human experts and their applicability in the proposed approach.

Weighted rank correlation

In the case of spearman's rank correlation, all ranks are given the same importance. However, as in the proposed approach, the degree of agreement between the three pillars in terms of top-ranked features is more important than the agreement of features that are least-ranked by the ML model. It is because the agreement of a feature that is ranked high by dataset in terms of information gain or by ML model in terms of feature importance is more crucial as compared to a low-ranked feature. So, instead of spearman's rank correlation, weighted correlation can be explored.

Evaluating human-feedback adaptive learning in other problem domain

The proposed human-feedback adaptive learning approach has potential applications in problem domains where there is a lack of training data and in problems where the agreement of the ML model with human users is crucial. For example, an ML-based solution for disease diagnosis requires to have a significantly high user agreement to win a trust vote from medical domain experts.

Variants of the human-feedback adaptive algorithms

As future ML systems are expected to be interpretable and interactive, there is a need to come up with variants of learning algorithms that can enable an ML model to adapt itself to incorporate human expert feedback to result in the improved user agreement. These variants should be able to address the issue of variety in feedback collected from a diverse set of human users who may not have the same level of domain expertise. Another opportunity is to evaluate alternative ideas to adapt to human expert feedback collected that can be used as an alternative to the 'eliminating least-ranked feature' approach demonstrated in this work.

Metrics for evaluation of interactive and interpretable systems

Human interpretability and the ability to interact with human users are going to be among the desired characteristics of future ML systems, in addition to being accurate. The task of evaluating ML systems in terms of accuracy is quite established using metrics like accuracy, precision and recall. However, there is still a lack of formal and established metrics for evaluating ML systems in terms of interpretability and interaction ability.

Novel opportunities in UI/UX design

As human users are poised to play a bigger role in future ML systems, these systems must be equipped with a user interface that can enable human users to interact with ML systems irrespective of their ML expertise. The medium of communicating to the user and collecting feedback from the user should be in terms known to these users. Moreover, all human users interacting with the ML system may not have the same level of expertise in the problem domain. So, the ML system should have a user interface that can accommodate human users with varying levels of expertise regarding the problem domain. There is a need to bring established UI/UX practices into the design of this user interfaces.

Extension to regression problems

The proposed framework has been demonstrated using a binary classification problem. A future line of work can be to evaluate this framework onto regression problems.

Controls in the hands of the FIFI graph user

As future work, a few controls can be given in the hands of the end-user. A user can be enabled to create a subset of the graph that interactively filter top important features. A user can be enabled to filter out important interactions only as per user criteria.

Variants of the FIFI graph

Another variant of a FIFI graph can be created where feature-importance is replaced by 1-way feature interaction. Such a visualization will be useful in identifying the most interacting features and then zooming out which of its interactions are relatively stronger.

Communicating with a Knowledge-graph

Possible exploration of using a FIFI graph and a knowledge graph in a complementary manner is another future line of work.

Performance comparison of FIFI graph in human interpretability

As future work, experimental work can be conducted involving human subjects to evaluate the efficacy of a FIFI graph. The experimental work can be focused on measuring how a FIFI graph performs against other visualization techniques for interpretability.

Plotting higher-order interactions

The proposed FIFI graph plots visualization between a pair of features. However, in complex problem domains like computational biology, there may be interactions of order higher than two. Identification of these higher-order interactions has the potential to discover important interactions that are still unknown to the experts of the problem domain under study.

Addressing computational complexity

As the number of features used for learning by an ML model increases, the number of feature interactions to be computed also increases combinatorically. There is a need to find computationally efficient ways for computing feature interactions apart from the H-Statistic based computations.

References

Abdollahi, B. and Nasraoui, O. (2016), ‘Explainable restricted boltzmann machines for collaborative filtering’, *arXiv preprint arXiv:1606.07129* .

Alvarez-Melis, D. and Jaakkola, T. S. (2018), ‘On the robustness of interpretability methods’, *arXiv preprint arXiv:1806.08049* .

Amershi, S., Cakmak, M., Knox, W. B. and Kulesza, T. (2014), ‘Power to the people: The role of humans in interactive machine learning’, *Ai Magazine* **35**(4), 105–120.

Baehrens, D., Schroeter, T., Harmeling, S., Kawanabe, M., Hansen, K. and MÃžller, K.-R. (2010), ‘How to explain individual classification decisions’, *Journal of Machine Learning Research* **11**(Jun), 1803–1831.

Bibal, A. and Frénay, B. (2016), Interpretability of machine learning models and representations: an introduction., *in* ‘ESANN’.

Breiman, L. (2001), ‘Random forests’, *Machine learning* **45**(1), 5–32.

Chen, D., Fraiberger, S. P., Moakler, R. and Provost, F. (2017), ‘Enhancing transparency and control when drawing data-driven inferences about individuals’, *Big data* **5**(3), 197–212.

Conati, C., Porayska-Pomsta, K. and Mavrikis, M. (2018), ‘Ai in education needs interpretable machine learning: Lessons from open learner modelling’, *arXiv preprint arXiv:1807.00154* .

Condry, N. (2016), ‘Meaningful models: Utilizing conceptual structure to improve machine learning interpretability’, *arXiv preprint arXiv:1607.00279* .

Csardi, G., Nepusz, T. et al. (2006), ‘The igraph software package for complex network research’, *InterJournal, complex systems* **1695**(5), 1–9.

Deng, H. (2019), ‘Interpreting tree ensembles with intrees’, *International Journal of Data Science and Analytics* **7**(4), 277–287.

Dhurandhar, A., Iyengar, V., Luss, R. and Shanmugam, K. (2017), ‘A formal framework to characterize interpretability of procedures’, *arXiv preprint arXiv:1707.03886* .

Dodge, J., Liao, Q. V., Zhang, Y., Bellamy, R. K. and Dugan, C. (2019), Explaining models: an empirical study of how explanations impact fairness judgment, *in* ‘Proceedings of the 24th international conference on intelligent user interfaces’, pp. 275–285.

Domingos, P. (2012), ‘A few useful things to know about machine learning’, *Communications of the ACM* **55**(10), 78–87.

Doshi-Velez, F. and Kim, B. (2017), ‘Towards a rigorous science of interpretable machine learning’, *arXiv preprint arXiv:1702.08608* .

Doshi-Velez, F., Kortz, M., Budish, R., Bavitz, C., Gershman, S., O’Brien, D., Scott, K., Schieber, S., Waldo, J., Weinberger, D. et al. (2017), ‘Accountability of ai under the law: The role of explanation’, *arXiv preprint arXiv:1711.01134* .

Friedman, J. H. (2001), ‘Greedy function approximation: a gradient boosting machine’, *Annals of statistics* pp. 1189–1232.

Friedman, J. H., Popescu, B. E. et al. (2008), ‘Predictive learning via rule ensembles’, *The Annals of Applied Statistics* **2**(3), 916–954.

Gallego-Ortiz, C. and Martel, A. L. (2016), ‘Interpreting extracted rules from ensemble of trees: Application to computer-aided diagnosis of breast mri’, *arXiv preprint arXiv:1606.08288* .

Gilpin, L. H., Bau, D., Yuan, B. Z., Bajwa, A., Specter, M. and Kagal, L. (2018), Explaining explanations: An overview of interpretability of machine learning, *in* ‘2018 IEEE 5th International Conference on data science and advanced analytics (DSAA)’, IEEE, pp. 80–89.

Goldstein, A., Kapelner, A., Bleich, J. and Pitkin, E. (2015), ‘Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation’, *Journal of Computational and Graphical Statistics* **24**(1), 44–65.

Gupta, R., Tanwar, S., Tyagi, S. and Kumar, N. (2020), ‘Machine learning models for secure data analytics: A taxonomy and threat model’, *Computer Communications* **153**, 406–440.

Gusmao, A. C., Correia, A. H. C., De Bona, G. and Cozman, F. G. (2018), ‘Interpreting embedding models of knowledge bases: a pedagogical approach’, *arXiv preprint arXiv:1806.09504* .

Guyon, I. and Elisseeff, A. (2003), ‘An introduction to variable and feature selection’, *Journal of machine learning research* **3**(Mar), 1157–1182.

Handcock, M. S. and Morris, M. (2006), *Relative distribution methods in the social sciences*, Springer Science & Business Media.

Hara, S. and Hayashi, K. (2016), ‘Making tree ensembles interpretable’, *arXiv preprint arXiv:1606.05390* .

Hara, S., Ikeno, K., Soma, T. and Maehara, T. (2018), ‘Maximally invariant data perturbation as explanation’, *arXiv preprint arXiv:1806.07004* .

Hausser, J. and Strimmer, K. (2014), ‘Entropy: estimation of entropy, mutual information and related quantities’, *Cran R* .

Hechtlinger, Y. (2016), ‘Interpretation of prediction models using the input gradient’, *arXiv preprint arXiv:1611.07634* .

Hendricks, L. A., Hu, R., Darrell, T. and Akata, Z. (2018), ‘Generating counterfactual explanations with natural language’, *arXiv preprint arXiv:1806.09809* .

Henelius, A., Puolamäki, K. and Ukkonen, A. (2017), ‘Interpreting classifiers through attribute interactions in datasets’, *arXiv preprint arXiv:1707.07576* .

Holzinger, A., Plass, M., Kickmeier-Rust, M., Holzinger, K., Crişan, G. C., Pintea, C.-M. and Palade, V. (2019), ‘Interactive machine learning: experimental evidence for the human in the algorithmic loop’, *Applied Intelligence* **49**(7), 2401–2414.

Jain, P. et al. (2021), ‘Convolutional neural network based advertisement classification models for online english newspapers’, *Turkish Journal of Computer and Mathematics Education (TURCOMAT)* **12**(2), 1687–1698.

Jovanovic, M., Radovanovic, S., Vukicevic, M., Van Poucke, S. and Delibasic, B. (2016), ‘Building interpretable predictive models for pediatric hospital readmission using tree-lasso logistic regression’, *Artificial intelligence in medicine* **72**, 12–21.

Kaggle (2016), `https://www.kaggle.com/uciml/pima-indians-diabetes-database`.

Kaggle (2019), `https://www.https://www.kaggle.com/blastchar/telco-customer-churn`.

Kim, B. (2015), Interactive and interpretable machine learning models for human machine collaboration, PhD thesis, Massachusetts Institute of Technology.

Kim, B., Wattenberg, M., Gilmer, J., Cai, C., Wexler, J., Viegas, F. et al. (2018), Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav), *in* 'International conference on machine learning', PMLR, pp. 2668–2677.

Kim, J., Merrill Jr, K., Xu, K. and Sellnow, D. D. (2021), 'I like my relational machine teacher: An ai instructor's communication styles and social presence in online education', *International Journal of Human–Computer Interaction* pp. 1–11.

Kim, J., Merrill, K., Xu, K. and Sellnow, D. D. (2020), 'My teacher is a machine: Understanding students' perceptions of ai teaching assistants in online education', *International Journal of Human–Computer Interaction* **36**(20), 1902–1911.

Kocielnik, R., Amershi, S. and Bennett, P. N. (2019), Will you accept an imperfect ai? exploring designs for adjusting end-user expectations of ai systems, *in* 'Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems', pp. 1–14.

Koh, P. W. and Liang, P. (2017), Understanding black-box predictions via influence functions, *in* 'International Conference on Machine Learning', PMLR, pp. 1885–1894.

Krause, J., Perer, A. and Bertini, E. (2016), 'Using visual analytics to interpret predictive machine learning models', *arXiv preprint arXiv:1606.05685* .

Krishnan, M. (2019), 'Against interpretability: a critical examination of the interpretability problem in machine learning', *Philosophy & Technology* pp. 1–16.

Kumar, P., Kumar, V. and Sobti, R. (2020), Predicting joining behavior of freshmen students using machine learning–a case study, *in* '2020 International Conference on Computational Performance Evaluation (ComPE)', IEEE, pp. 141–145.

Lage, I., Ross, A., Gershman, S. J., Kim, B. and Doshi-Velez, F. (2018), Human-in-the-loop interpretability prior, *in* 'Advances in neural information processing systems', pp. 10159–10168.

Laugel, T., Renard, X., Lesot, M.-J., Marsala, C. and Detyniecki, M. (2018), 'Defining locality for surrogates in post-hoc interpretablity', *arXiv preprint arXiv:1806.07498* .

Liaw, A., Wiener, M. et al. (2002), 'Classification and regression by randomforest', *R news* **2**(3), 18–22.

Lipton, Z. C. (2016), 'The mythos of model interpretability', *arXiv preprint arXiv:1606.03490* .

Lundberg, S. M. and Lee, S.-I. (2017), ‘Consistent feature attribution for tree ensembles’, *arXiv preprint arXiv:1706.06060* .

Mitchell, T. M. (2006), *The discipline of machine learning*, Vol. 9, Carnegie Mellon University, School of Computer Science, Machine Learning ....

Moeyersoms, J., d’Alessandro, B., Provost, F. and Martens, D. (2016), ‘Explaining classification models built on high-dimensional sparse data’, *arXiv preprint arXiv:1607.06280* .

Molnar, C. (2020), *Interpretable Machine Learning*, Lulu. com.

Molnar, C., Casalicchio, G. and Bischl, B. (2018), ‘iml: An r package for interpretable machine learning’, *Journal of Open Source Software* **3**(26), 786.

Montavon, G., Lapuschkin, S., Binder, A., Samek, W. and Müller, K.-R. (2017), ‘Explaining nonlinear classification decisions with deep taylor decomposition’, *Pattern Recognition* **65**, 211–222.

Narayanan, M., Chen, E., He, J., Kim, B., Gershman, S. and Doshi-Velez, F. (2018), ‘How do humans understand explanations from machine learning systems? an evaluation of the human-interpretability of explanation’, *arXiv preprint arXiv:1802.00682* .

Nickel, M., Murphy, K., Tresp, V. and Gabrilovich, E. (2015), ‘A review of relational machine learning for knowledge graphs’, *Proceedings of the IEEE* **104**(1), 11–33.

Patel, M. M., Tanwar, S., Gupta, R. and Kumar, N. (2020), ‘A deep learning-based cryptocurrency price prediction scheme for financial institutions’, *Journal of Information Security and Applications* **55**, 102583.

Penkov, S. and Ramamoorthy, S. (2017), ‘Using program induction to interpret transition system dynamics’, *arXiv preprint arXiv:1708.00376* .

Phillips, R., Chang, K. H. and Friedler, S. A. (2018), Interpretable active learning, *in* ‘Conference on fairness, accountability and transparency’, PMLR, pp. 49–61.

Pu, P. and Chen, L. (2006), Trust building with explanation interfaces, *in* ‘Proceedings of the 11th international conference on Intelligent user interfaces’, pp. 93–100.

Rani, A., Taneja, K. and Taneja, H. (2021), ‘Life insurance-based recommendation system for effective information computing’, *International Journal of Information Retrieval Research (IJIRR)* **11**(2), 1–14.

Ribeiro, M. T., Singh, S. and Guestrin, C. (2016*a*), ‘Model-agnostic interpretability of machine learning’, *arXiv preprint arXiv:1606.05386* .

Ribeiro, M. T., Singh, S. and Guestrin, C. (2016*b*), Why should i trust you?: Explaining the predictions of any classifier, *in* ‘Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining’, ACM, pp. 1135–1144.

Ribeiro, M. T., Singh, S. and Guestrin, C. (2018), Anchors: High-precision model-agnostic explanations, *in* ‘Proceedings of the AAAI Conference on Artificial Intelligence’, Vol. 32.

Rosé, C. P., McLaughlin, E. A., Liu, R. and Koedinger, K. R. (2019), ‘Explanatory learner models: Why machine learning (alone) is not the answer’, *British Journal of Educational Technology* **50**(6), 2943–2958.

Rosenbaum, C., Gao, T. and Klinger, T. (2017), ‘e-qraq: A multi-turn reasoning dataset and simulator with explanations’, *arXiv preprint arXiv:1708.01776* .

Ross, A. S., Hughes, M. C. and Doshi-Velez, F. (2017), ‘Right for the right reasons: Training differentiable models by constraining their explanations’, *arXiv preprint arXiv:1703.03717* .

Ross, A. S., Pan, W. and Doshi-Velez, F. (2018), ‘Learning qualitatively diverse and interpretable rules for classification’, *arXiv preprint arXiv:1806.08716* .

Samek, W., Montavon, G., Binder, A., Lapuschkin, S. and Müller, K.-R. (2016), ‘Interpreting the predictions of complex ml models by layer-wise relevance propagation’, *arXiv preprint arXiv:1611.08191* .

Samek, W., Wiegand, T. and Müller, K.-R. (2017), ‘Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models’, *arXiv preprint arXiv:1708.08296* .

Shank, D. B. and Gott, A. (2020), ‘Exposed by ais! people personally witness artificial intelligence exposing personal information and exposing people to undesirable content’, *International Journal of Human–Computer Interaction* **36**(17), 1636–1645.

Shannon, C. E. (1948), ‘A mathematical theory of communication’, *The Bell system technical journal* **27**(3), 379–423.

Shapley, L. S. (1953), ‘A value for n-person games’, *Contributions to the Theory of Games* **2**(28), 307–317.

Shneiderman, B. (2020), ‘Human-centered artificial intelligence: Reliable, safe & trustworthy’, *International Journal of Human–Computer Interaction* **36**(6), 495–504.

Shokri, R., Strobel, M. and Zick, Y. (2019), ‘On the privacy risks of model explanations’, *arXiv preprint arXiv:1907.00164* .

Singh, S., Ribeiro, M. T. and Guestrin, C. (2016), ‘Programs as black-box explanations’, *arXiv preprint arXiv:1611.07579* .

Smith-Renner, A., Fan, R., Birchfield, M., Wu, T., Boyd-Graber, J., Weld, D. S. and Findlater, L. (2020), No explainability without accountability: An empirical study of explanations and feedback in interactive ml, *in* ‘Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems’, pp. 1–13.

Spearman, C. (1987), ‘The proof and measurement of association between two things’, *The American Journal of Psychology* **100**(3/4), 441–471.

Storcheus, D., Rostamizadeh, A. and Kumar, S. (2015), A survey of modern questions and challenges in feature extraction, *in* ‘Feature Extraction: Modern Questions and Challenges’, pp. 1–18.

Strumbelj, E. and Kononenko, I. (2010), ‘An efficient explanation of individual classifications using game theory’, *The Journal of Machine Learning Research* **11**, 1–18.

Tan, H. F., Hooker, G. and Wells, M. T. (2016), ‘Tree space prototypes: Another look at making tree ensembles interpretable’, *arXiv preprint arXiv:1611.07115* .

Taneja, K., Taneja, H. and Kaur, R. (2021), ‘Evolutionary computation techniques for intelligent computing in commercial mobile adhoc networks.’, *International Journal of Next-Generation Computing* **12**(2).

Tomsett, R., Braines, D., Harborne, D., Preece, A. and Chakraborty, S. (2018), ‘Interpretable to whom? a role-based model for analyzing interpretable machine learning systems’, *arXiv preprint arXiv:1806.07552* .

UGC (2019), `https://www.ugc.ac.in/oldpdf/Private%20University/Consolidated_List_Private_Universities.pdf`.

van der Waa, J., Robeer, M., van Diggelen, J., Brinkhuis, M. and Neerincx, M. (2018), ‘Contrastive explanations with local foil trees’, *arXiv preprint arXiv:1806.07470* .

Vandewiele, G., Janssens, O., Ongenae, F., De Turck, F. and Van Hoecke, S. (2016), ‘Genesim: genetic extraction of a single, interpretable model’, *arXiv preprint arXiv:1611.05722* .

Varshney, K. R., Khanduri, P., Sharma, P., Zhang, S. and Varshney, P. K. (2018), ‘Why interpretability in machine learning? an answer using distributed detection and data fusion theory’, *arXiv preprint arXiv:1806.09710* .

Villagrá-Arnedo, C. J., Gallego-Durán, F. J., Llorens-Largo, F., Compañ-Rosique, P., Satorre-Cuerda, R. and Molina-Carmona, R. (2017), ‘Improving the expressiveness of black-box models for predicting student performance’, *Computers in Human Behavior* **72**, 621–631.

Wagstaff, K. (2012), ‘Machine learning that matters’, *arXiv preprint arXiv:1206.4656* .

Wagstaff, K. L. and Lee, J. (2018), ‘Interpretable discovery in large image data sets’, *arXiv preprint arXiv:1806.08340* .

Wang, F. and Rudin, C. (2015), Falling rule lists, *in* ‘Artificial Intelligence and Statistics’, pp. 1013–1022.

Wang, T., Rudin, C., Doshi-Velez, F., Liu, Y., Klampfl, E. and MacNeille, P. (2017), ‘A bayesian framework for learning rule sets for interpretable classification’, *The Journal of Machine Learning Research* **18**(1), 2357–2393.

Weller, A. (2017), ‘Challenges for transparency’, *arXiv preprint arXiv:1708.01870* .

Weller, A. (2019), Transparency: motivations and challenges, *in* ‘Explainable AI: Interpreting, Explaining and Visualizing Deep Learning’, Springer, pp. 23–40.

Wittkowski, K. (1986), ‘Classification and regression trees-l. breiman, jh friedman, ra olshen and cj stone.’, *Metrika* **33**, 128–128.

Xu, B., Huang, J. Z., Williams, G., Wang, Q. and Ye, Y. (2012), ‘Classifying very high-dimensional data with random forests built from small subspaces’, *International Journal of Data Warehousing and Mining (IJDWM)* **8**(2), 44–63.

Yang, Y., Kandogan, E., Li, Y., Sen, P. and Lasecki, W. S. (2019), A study on interaction in human-in-the-loop machine learning for text analytics., *in* ‘IUI Workshops’.

Yang, Y., Morillo, I. G. and Hospedales, T. M. (2018), ‘Deep neural decision trees’, *arXiv preprint arXiv:1806.06988* .

Zeng, J., Ustun, B. and Rudin, C. (2015), ‘Interpretable classification models for recidivism prediction’, *arXiv preprint arXiv:1503.07810* .

Zhao, H., Williams, G. J., Huang, J. Z. et al. (2017), ‘Wsrf: an r package for classification with scalable weighted subspace random forests’, *J Stat Softw* **77**(i03), 1.

Publications out of this work

Journal

- (2021). Human interpretability in machine learning – A review aiming to answer fundamental questions. Pattern Recognition. **Status: Communicated**

- (2021). Data, Machine Learning, and Human Domain Experts: None is better than their Collaboration. International Journal of Human-Computer Interaction. **Status: Accepted**

- (2020). Unboxing the Classification for Visualization of the Outcomes with Naïve User Perspective. International Journal of Control and Automation, 13(4), 1312-1325.

- (2019). INTERNATIONAL STUDENTS'ACADEMIC PERFORMANCE PREDICTION WITH DESCRIPTIVE MACHINE LEARNING EXPLANATIONS. Journal of the Gujarat Research Society, 21(6), 748-758.

- (2019). ANTICIPATING PLACEMENT STATUS OF STUDENTS USING MACHINE LEARNING. Journal of the Gujarat Research Society, 21(6), 738-747.

- (2018). Identifying factors affecting placement status of engineering students using explainable machine learning. Journal of Emerging Technologies and Innovative Research, Vol.5, Issue 12, page no.950-957.

Conference

- (2021) Feature-Importance Feature-Interactions (FIFI) graph: A Novel Visualization for Interpretable Machine Learning, International Conference on Intelligent Technologies, Karnataka, India.

- (2020) Predicting Academic Performance of International Students Using Machine Learning Techniques and Human Interpretable Explanations Using LIME—Case Study of an Indian University, International Conference on Innovative Computing and Communications. Advances in Intelligent Systems and Computing, vol 1087. Springer, Singapore. https://doi.org/10.1007/978-981-15-1286-5_25

- (2020) Predicting Joining Behavior of Freshmen Students using Machine Learning – A Case Study, International Conference on Computational Performance

Evaluation (ComPE), Shillong, India, 2020, pp. 141-145,
doi: 10.1109/ComPE49325.2020.9200167.

Book Chapters

- (2021) Interpretable and interactive machine learning systems using human-machine interaction: state-of-the-art and future opportunities using human-feedback adaptive, Evolutionary Computation with Intelligent Systems: A Multidisciplinary Approach to Society 5.0 **Status: Accepted**

- (2021) Evaluating Machine Learning Capabilities for Predicting Joining Behavior of Freshmen Students enrolled at Institutes of Higher Education: Case study from a novel problem domain, Data Science and Innovations for Intelligent Systems: Computational Excellence and Society 5.0

- (2019) Human Interpretable Machine Learning, Laxmi Publications Pvt. Ltd., 978-93-5274-657-6

Magazine Articles

- (2018). Human interpretability in machine learning based solutions. CSI Communications, Special Research Issue on Pattern Recognition, vol. 41, Issue 11, p.no. 13-14

- (2018). A framework for monitoring healthcare of university students using machine learning and data analytics. CSI Communications, Special Research Issue on Pattern Recognition, vol. 41, Issue 11, p.no. 29-30

Curriculam Vitae of the Scholar

1. **Bio-data**

    - *Name*: Pawan Kumar
    - *Registration No.*:41500197
    - *Father's Name*: Sh Ram Kishan
    - *Date of Birth*: 24th Dec, 1976
    - *Permanent Address*: 56/5 Bhagwati Niwas, Joginder Nagar, Rama Mandi, Jalandhar - 144007, Punjab

2. **Present Status**:

    - Assistant Professor, School of Computer Applications, Lovely Professional University Phagwara

3. **Academic Qualification**:

    - Masters in Computer Applications, 2001
    - Bachelor in Science, 1998

4. **Research Experience**:

    - (2016 - Present) Research Scholar, School of Computer Applications, Lovely Professional University, Phagwara

5. **Journal Publications**:

    - (2020). Unboxing the Classification for Visualization of the Outcomes with Naïve User Perspective. International Journal of Control and Automation, 13(4), 1312-1325.
    - (2019). INTERNATIONAL STUDENTS'ACADEMIC PERFORMANCE PREDICTION WITH DESCRIPTIVE MACHINE LEARNING EXPLANATIONS. Journal of the Gujarat Research Society, 21(6), 748-758.

- (2019). ANTICIPATING PLACEMENT STATUS OF STUDENTS USING MACHINE LEARNING. Journal of the Gujarat Research Society, 21(6), 738-747.
- (2018). Identifying factors affecting placement status of engineering students using explainable machine learning. International Journal of Emerging Technologies and Innovative Research, Vol.5, Issue 12, page no.950-957.

6. **Conference Publications**:

   - (2021) Feature-Importance Feature-Interactions (FIFI) graph: A Novel Visualization for Interpretable Machine Learning, International Conference on Intelligent Technologies, Karnataka, India.
   - (2020) Predicting Joining Behavior of Freshmen Students using Machine Learning – A Case Study, International Conference on Computational Performance Evaluation (ComPE), Shillong, India, 2020, pp. 141-145, doi: 10.1109/ComPE49325.2020.9200167.
   - (2020) Predicting Academic Performance of International Students Using Machine Learning Techniques and Human Interpretable Explanations Using LIME—Case Study of an Indian University, International Conference on Innovative Computing and Communications. Advances in Intelligent Systems and Computing, vol 1087. Springer, Singapore. https://doi.org/10.1007/978-981-15-1286-5_25

7. **Awards**:

   - (2018) Best paper award, Feynman100: International Conference on Computing Sciences
   - (2013) Best paper award, Wilkes100:International Conference on Computing Sciences
   - (2012) Qualified UGC-NET in Computer Science and Applications
   - (2012) Qualified GATE in Computer Science and Information Technology

8. **Professional Affiliations**

   - Life member, Computer Society of India
   - Life member, Indian Science Congress Association