PREDICTION OF FILLING INSURANCE CLAIMS BY DRIVERS USING AN EFFECTIVE MACHINE LEARNING MODEL

Dissertation submitted in partial fulfilment of the requirements for the Degree of

MASTER OF TECHNOLOGY

in

Information Technology

By

CHUENFFO TAGNE ARNOLD

11617612

Supervisor

ASEEM KUMAR

School of Computer Science and Engineering

Lovely Professional University

Phagwara, Punjab (India)

November 2017

TOPIC APPROVAL PERFORMA

School of Computer Science and Engineering

Program : P173::M.Tech. (Information Technology) [Full Time]

COURSE CODE : INT548 **REGULAR/BACKLOG :** Regular **GROUP NUMBER :** CSERGD0365

Supervisor Name : Aseem Kumar **UID :** 16839 **Designation :** Assistant Professor

Qualification : _____ **Research Experience :** _____

SR.NO.	NAME OF STUDENT	REGISTRATION NO	BATCH	SECTION	CONTACT NUMBER
1	Chuenffo Tagne Arnold	11617612	2016	K1638	9115514082

SPECIALIZATION AREA : Software Engineering **Supervisor Signature:** _____

PROPOSED TOPIC : Prediction of filling insurance claims by drivers using an effective machine learning model

Qualitative Assessment of Proposed Topic by PAC		
Sr.No.	**Parameter**	**Rating (out of 10)**
1	Project Novelty: Potential of the project to create new knowledge	6.00
2	Project Feasibility: Project can be timely carried out in-house with low-cost and available resources in the University by the students.	7.00
3	Project Academic Inputs: Project topic is relevant and makes extensive use of academic inputs in UG program and serves as a culminating effort for core study area of the degree program.	7.00
4	Project Supervision: Project supervisor's is technically competent to guide students, resolve any issues, and impart necessary skills.	7.00
5	Social Applicability: Project work intends to solve a practical problem.	6.50
6	Future Scope: Project has potential to become basis of future research work, publication or patent.	7.00

PAC Committee Members		
PAC Member 1 Name: Gaurav Pushkarna	UID: 11057	Recommended (Y/N): NA
PAC Member 2 Name: Er.Dalwinder Singh	UID: 11265	Recommended (Y/N): Yes
PAC Member 3 Name: Harwant Singh Arri	UID: 12975	Recommended (Y/N): NA
PAC Member 4 Name: Balraj Singh	UID: 13075	Recommended (Y/N): Yes
PAC Member 5 Name: Raj Karan Singh	UID: 14307	Recommended (Y/N): NA
PAC Member 6 Name: Harleen Kaur	UID: 14508	Recommended (Y/N): NA

PAC Member 7 Name: Sawal Tandon	UID: 14770	Recommended (Y/N): NA
PAC Member 8 Name: Tejinder Thind	UID: 15312	Recommended (Y/N): NA
DAA Nominee Name: Kuldeep Kumar Kushwaha	UID: 17118	Recommended (Y/N): NA

Final Topic Approved by PAC: **Prediction of filling insurance claims by drivers using an effective machine learning model**

Overall Remarks: Approved

PAC CHAIRPERSON Name: 11024::Amandeep Nagpal **Approval Date:** 10 Nov 2017

DECLARATION BY STUDENT

I hereby declare that research work reported in this dissertation proposal entitled "PREDICTION OF FILLING INSURANCE CLAIMS BY DRIVERS USING AN EFFECTIVE MACHINE LEARNING MODEL" in partial fulfilment of the award of the Degree of Master of Technology in Information Technology at Lovely Professional University, Phagwara Punjab is an authentic work carried out under the supervision of Mr Aseem Kumar. I have not submitted this work elsewhere for any degree.

I understand that the work presented herewith is in direct compliance with Lovely Professional University's Policy on plagiarism, intellectual property rights, and highest standard of moral and ethical conduct. Therefore, to the best of my knowledge, the content of this dissertation represents authentic and honest research effort conducted, in its entirety, by me. I am fully responsible for the content of this dissertation work.

Signature of student

Chuenffo T. Arnold

11617612

CERTIFICATE BY RESEARCH SUPERVISOR

This is to certify that the work reported in the M.Tech Dissertation/dissertation proposal entitled "**PREDICTION OF FILLING INSURANCE CLAIMS BY DRIVERS USING AN EFFECTIVE MACHINE LEARNING MODEL** " submitted by **CHUENFFO T. ARNOLD** at **Lovely Professional University, Phagwara, India** is a bonafide record of his / her original work carried out under my supervision. This work has not been submitted elsewhere for any other degree.

Signature of Supervisor

Aseem Kumar

Date:

ACKNOWLEDGEMENTS

I would like to express my deepest appreciation to all those who provided me with comments and suggestions on this research work. I am highly indebted to my supervisor Mr. Aseem Kumar and Associate Dean Amandeep Nagpal for their guidance and constant supervision regarding this research work.

TABLE OF CONTENTS

LIST OF FIGURES, TABLES OR ILLUSTRATIONS

ABSTRACT

Machine learning concept is a paradox for most industries nowadays and automobile insurance industry in particular. This cutting-edge technology is making a lot of remarkable progress and impact in the field of predictive analytics. Business turn to grow faster with machine learning. The objective of this research work is to demonstrate the use and importance of machine learning in the insurance industry and how a model can be used to automate the pattern findings actual data like predicting the probability that a driver will file an insurance claim.

CHAPTER 1 INTRODUCTION

Companies of the insurance industry strive for market growth and profitability which has become highly competitive. In order to increase their market share, policy holders with elevated risk might be underwritten and the benefits or margins of the company might suffer. For the companies to keep and maintain the market share and attain the profit target level, the decision processes are made by setting the profit targets, estimate the risk of each policy holder and setting competitive instalments.

A challenge faced by most insurance companies is to charge each driver an appropriate price for the risk they represent. Risks usually vary from driver (policy holder) to driver, and a deep understanding of these risks parameters is needed in order to make predictions that will allow companies to tailor their prices, and hopefully make auto insurance coverage more accessible to more drivers. We also need to know the effect of price change on customer for customer retention patterns, as well as providing the prospective for market growth, considering the high competitive nature of the business.

Over the years, actuaries and statisticians have used historical data to find patterns in claims and predict future losses for coming years. They've been pretty creative in doing so, using tools in line with the technology of their time from minimum bias all the way up to decision trees [1]. Techniques like data mining are proving to be of enormous benefit to the business world, in terms of identifying hidden patterns in data, as well as predicting future behaviours of customers. The level of sophistication and tools has changed over time and we look at Machine Learning and Artificial Intelligence as transformative way to solve the same problems while also gaining insights from places where traditional methods fail.

The objective of this research work is to build a machine learning model that will predict a driver filling a claim to the insurance company. We have always tried to find patterns in data. What we can do now is automate that pattern finding with machine learning. The data was found at Kaggle (http://kaggle.com), a website that specialises in running statistical analysis and predictive modelling competitions. The Brazilian company Porto Seguro hosted a competition called "Porto Seguro's Safe Driver Prediction (Predict if a driver will file an insurance claim next year).", which

was run from October 2017 to 30th November 2017. The challenge is to build a model that predicts the probability that a driver will initiate an auto insurance claim in the next year.

The data consist of three files:

- training data in train.csv, each row corresponds to a driver, and the target columns indicates that a claim was filed.
- test data in the test.csv file.
- sample_submission.csv is submission file showing the correct format.

For each submission, a Normalized Gini Coefficient is used for evaluation and this coefficient ranges from approximately 0 for random guessing, to approximately 0.5 for a perfect score. The theoretical maximum for the discrete calculation is (1 - frac_pos) / 2.

This stage of the project, we are designing a clear methodology and making some exploratory and data analysis.

CHAPTER 2 REVIEW OF LITERATURE

Most of the work done in insurance claims focuses on fraud detection which is very similar to the work we are carrying out here. Data mining techniques have been widely used in general for claims processing.

The material read in preparation of this research work is based on four different areas. First, the insurance claims in the insurance industry which gave us a clearer understanding of the domain since we are from a different background. Second, data mining used for detecting insurance fraud, insurance claim prediction, customer behavior, retention and so on. Third, how machine learning is making an entrance in to the insurance industry and how it impacts the industry now and in future. Fourth and last is machine learning algorithm and techniques for predictions.

2.1 Insurance claim

Insurance claim could be defined as an application to an insurance company asking for a payment according to the terms of the insurance policy. The company goes through the claim for its validity and then pays out to the requesting party or insured once approved [2]. According to the policies in the insurance claims, then company will cover everything from death benefits to routine health exams at your local hospital.

For a risk to be insured, some assessments need to be done. Insurance companies generally keeps a comprehensive record of the claim history and personal information of their customers. The frequency of claim counts to a great extent reveals the riskiness of the insureds [3].

A generalized insurance claim process pipeline described by [4] as shown on Figure 2:1 below, shows the workflow is as follows:

- Claims submitted to the insurance company.
- The company does automatic validation checks followed by pricing using benefit rules and contracts. Manual intervention is sometimes required when automatic validation fails.
- Once the pricing is done, claims gets finalized and payment is sent to policy holder or issuer.

- The box placed after the pricing is used to detect potential issues related to the claim before it is finalized so it can be corrected before payment is made.

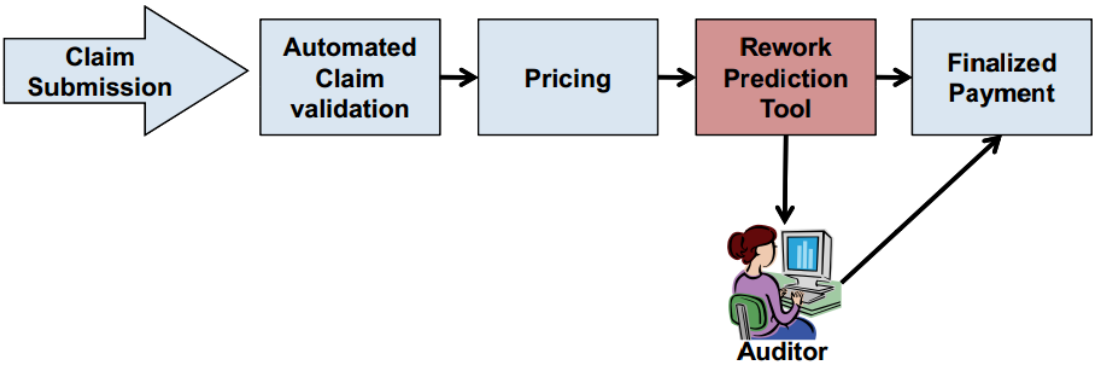

Figure 2:1:Claim Processing Pipeline

Insurance Fraud and mining

Insurance fraud has spread globally, and the society is becoming more concerned by this issue. [5] defines insurance fraud as "the act committed with the intent to obtain a fraudulent outcome from an insurance process".

[6] applies a Boosting Naïve Bayes technique for insurance claim fraud analysis on a case study where the findings of the study turn to be a good method for a valuable contribution to the design of brilliant, liable, and effective fraud diagnosis support.

Mining insurance fraud data has become a major concern in insurance industry and a lot of data mining techniques have been used to so far to find patterns in data.[7] makes use of a random forest model in mining the insurance data. This research evaluates random forest technique as a suitable, accurate and robust technique for large data set and unbalanced data which can be used for prediction of insurance claims, mining fraud rules and classification of claims.

For a risk to be insured, some assessments need to be done then the insurance claim cost can be predicted. [8] uses a case study where the policy holders are classified according to the perceived

risk and a model claim cost is built within each group. The classification is designed to achieve maximum similarities within groups and maximum dissimilarities between groups. This is usually achieved using clustering algorithms. The objective of this clustering is to find a small collection of nuggets that can be further investigated using human resources.

Another case study is described by [9] where an analysis of customer retention and insurance claim patterns findings is done using data mining. A classification algorithm is used for customer retention to group the policy holders as likely to renew and terminate their policies. A holistic framework utilizing hypothesis testing, decision tree, clustering and neural network is used for to find patterns in the data.

Statistical techniques has also been used for a long time in claim mining and interpretations.[10] demonstrate the use of modern statistical techniques in solving actuaries problems. A hierarchical model of three is components is proposed here. A negative binomial regression model for accessing claim frequency, a multinomial logit model to predict the type of insurance claim and severity component.

2.1 Machine learning in insurance

Machine learning is defined as "a field of study (artificial intelligence) that gives computers the ability to learn without being explicitly programmed" [11]. There are three major types of machine learning algorithms: supervised learning and unsupervised learning and reinforcement learning algorithms.

- Labeled training data are analyzed by supervised learning algorithms to produce a model used in predicting new data.
- Finding hidden structures in unlabeled data is done using unsupervised learning algorithms.
- Reinforcement learning algorithms is based on each data point and in turns provide information on the decision taken (either good or not). From time to time, this algorithm changes its learning strategy to achieve a better reward.

The insurance industry sees machine learning as turning point where most companies focus on refining compliance, upgrade cost structures and boost competitiveness. Machine learning appear to make change as can answer to these objectives [12]. Machine learning can be applied across

many business functions in insurance, including: claims forecasting, customer retention, direct marketing, conversion, targeting inspections and audits, predicting litigation and optimal pricing.

Customers keeps demanding about customizing their insurance purchases to their unique needs, leading insurers are evaluate how machine learning can improve customer satisfaction and business operations [13]. Insurance companies are able to use their data and machine learning algorithms to determine customer purchase patterns and manage risk due to the advances in statistical modeling techniques that are available to companies [14].

Wipro [15] released a white paper on detecting insurance claims fraud. A comparative analysis of machine learning techniques such as Logistic Regression, Modified Multi-variant Gaussian (MVG) and Bagging using Random Forest is done based on a data set.

Machine learning algorithms in prediction

There exist several machine learning algorithms of which the frequently used to in data analysis /data mining are [11]:

- **Segmentation Analysis Algorithms**: Data are divided into small groups (clusters) of items having similar properties.
- **Classification Analysis Algorithms**: In these algorithms the attributes in the dataset are used to predict a value.
- **Association Analysis Algorithms**: Find correlations and/or relations between unique features in a dataset. "Application of these algorithms are in market basket analysis.
- **Regression Analysis Algorithms**: In these algorithms data attributes in dataset are used to predict values for one or more variables that take continuous values. It is also used in the process of finding the correlation between variables.
- **Sequence Analysis Algorithms:** Outlines the occurrences in data. It discovers the patterns over time.

The commonly used machine learning algorithms are listed below [16]:

1. **Linear Regression Algorithm**: These algorithms predict values (values for one or more variables) based on the attributes that take continuous values from the data set. Linear Regression assists evaluate risk involved in insurance or financial domain. It is used in health insurance company to analyze the number of claims. This analysis helps insurance

companies find, frequencies in insurance claims filling. Such analysis results play a significant role risk management as well as in business decisions.

2. **Naïve Bayes Classifier Algorithm**: A is a straight forward and powerful algorithm that assigns data's element value from one of the available categories. An example is Spam Filter used in email filtering. The spam filter classifies emails as either "spam or not spam".
3. **K Means Clustering Algorithm**: A simple algorithm that operates on a given data set, partitioning it into a small number of clusters by minimizing the distance between each data point. The result output of this algorithm is "k" clusters with input data distributed among the clusters. Search engines like Bing, Google use K Means to group web pages by similarity and establish the 'relevance rate' of search results.
4. **Random Forest Algorithm**: This algorithm makes a small tweak to bagging approach to create a group of decision trees (powerful classifier) with random subset of the data. To obtain a good prediction performance, the model is trained on several random sample data.
5. **Logistic Regression**: Commonly used to predict class probabilities. Predictions are made based on logistic function applied to linear combination of attributes.

CHAPTER 3 PROBLEM DEFINITION

Insurance bills these days influences drivers on buying a brand-new car. Inaccuracies in car insurance company's claim predictions raise the cost of insurance for good drivers and reduce the price for bad ones. In order to tailor prices with respect to each customer and predict claims costs, data mining techniques/tools have always been at the center of the analytics.

The concept of Machine Learning or Artificial Intelligence (AI) is a still not accepted yet by many insurance companies but It is making a progression impact in predictive analytics. This cutting-edge technology has the power to blow apart old-style thinking and catapult a company into generating business ten, a hundred or even a thousand times faster than anything it has done before[1].

Machine learning has a significant impact in the insurance industry and in this project, we will use it to build a model that predicts the probability that a driver will file an insurance claim.

CHAPTER 4 SCOPE OF STUDY

This study focuses on machine learning in the insurance industry. How machine learning is used to automate predictions and how we can predict if a driver will file a claim. The data collected from Brazilian Company on the Kaggle platform help us get some insights into insurance industry.

The scope of this work is to build an effective the machine learning model to predict if a driver will file an insurance claim.

CHAPTER 5 OBJECTIVE OF THE STUDY

A clear objective of this work is to demonstrate the use and importance of machine learning in the insurance industry that is gradually replacing classic data mining used to find patterns in data. The machine learning model built can be used to automate the pattern findings.

We explore the use of machine learning in the insurance industry, build an automated model the will not only rely on data samples but on actual data of the company for predictive analytics. A model that can work on both structured and unstructured data.

A model that will be trained on a partial set of data and parameters tweaked on a testing set. Several models may be used and the performance of each tested to find the most effective.

CHAPTER 6 PROPOSED RESEARCH METHODOLOGY

We will adopt the following the methodology to carry out this research work.

First, relevant journal articles, publications, white papers, and studies were reviewed in order to get comprehensive information on insurance claims, insurance fraud and machine learning.

Second, we follow the flowchart for supervised learning by [17] as depicted on Figure 6:1. The common steps in a machine learning model building which are define the problem statement, data collection, data cleaning, variable selection, exploratory data analysis, model development, model validation and documentation.

Finally, discussion, conclusion and future work of the study are written down.

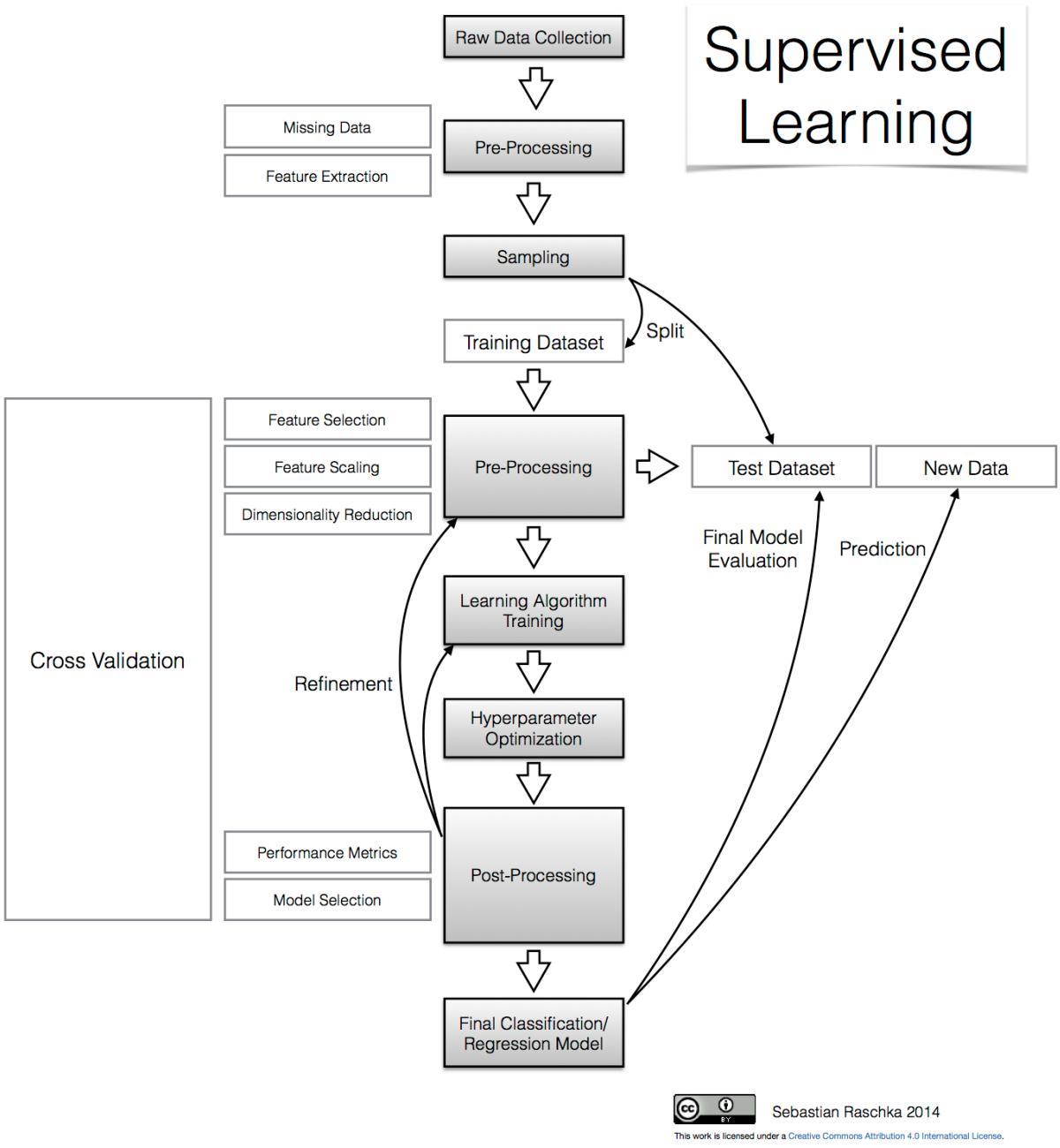

Figure 6:1 Supervised Learning Flowchart

6.1 Data collection

The Brazilian company Porto Seguro hosted a competition called "Porto Seguro's Safe Driver Prediction (Predict if a driver will file an insurance claim next year)", which was run from October 2017 to 30th November 2017. The challenge is to build a model that predicts the probability that a driver will initiate an auto insurance claim in the next year.

The data consist of three files:

- training data in train.csv, where each row corresponds to a driver, and the target columns indicates that a claim was filed.
- test data in the test.csv file.
- sample_submission.csv is submission file showing the correct format.

6.2 Exploratory Data Analysis

Since the sampling has already been done, the main tasks here will be to descript the data, visualize all the different data features, give a statistical summary, find a relation to the target variable, explore interactions between multi-parameter, and perform feature engineering (feature selection and dimensionality reduction). The target variable here, is conveniently named target because it indicates whether this policy holder made an insurance claim in the past.

6.3 Model development and validation

At this stage, cross-validation is used since several machine learning algorithms are tested and their performance evaluated. The most effective model will be selected based on the performance metric. A confusion matrix is generally used in the process of selecting a model

Figure 6:2 depicts the learning process once the model selected

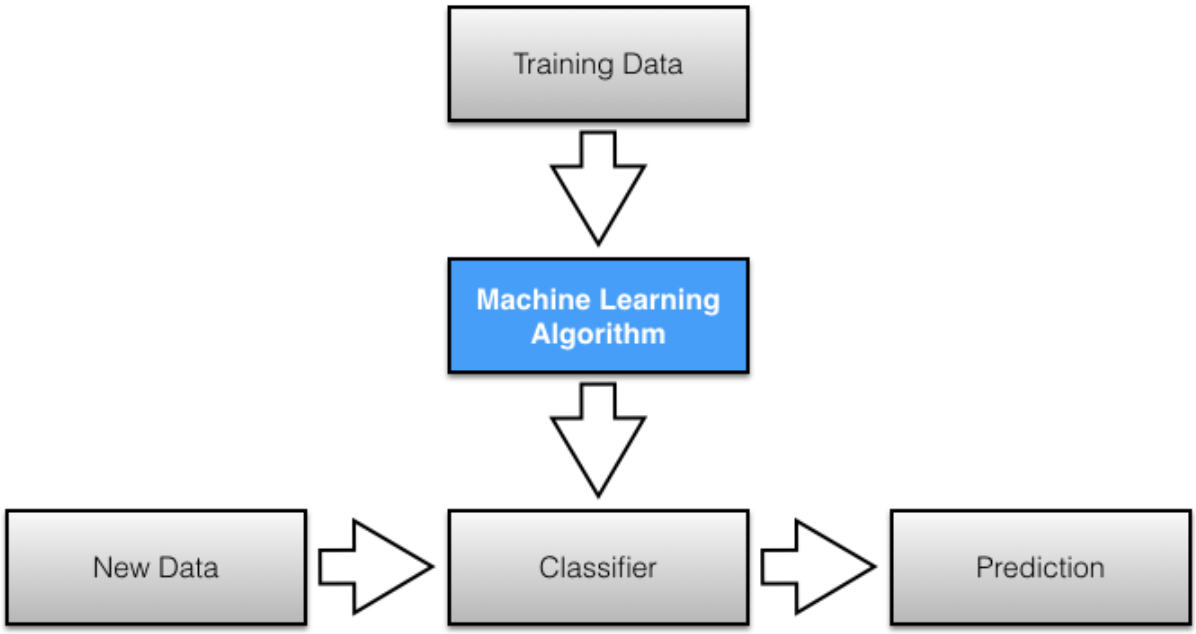

Figure 6:2 Learning algorithm

Machine learning algorithms are trained on the training data set and a model (Classifier as on the Figure 6:2) is built. The test (new) data set is used on this model to produce a new file (prediction) as since on Figure 6:1.

6.4 Model documentation

A proper documentation of the model is written down.

CHAPTER 7 EXPECTED OUTCOMES

The expected outcome of this study is to predict the probability that a driver will file an insurance claim using a machine learning model, with the purpose of providing a fairer insurance cost based on individual driving habits.

CONCLUSIONS

A review of machine learning and data mining in the insurance industry, insurance claims and fraud has been discussed. A challenging part was to find research papers, journal articles on machine learning in the insurance industry since the it is an emerging field. We collected the data for the research work on Kaggle.com.

In future, the proposed methodology will be used to explore the collected data, selected features and build the model in order to meet the expected outcomes of the research work. Data will be trained on the model and testing done to predict the probability that a policy holder (driver) will file an insurance claim.

REFERENCES

[1] M. Learning and A. Intelligence, "Anything You Can Do , AI Can Do Better : Machine Learning and Artificial Intelligence in Insurance Anything You Can Do , AI Can Do Better : Machine Learning and Artificial Intelligence in Insurance Editor."

[2] "Insurance Claim." [Online]. Available: https://www.investopedia.com/terms/i/insurance_claim.asp. [Accessed: 27-Nov-2017].

[3] P. Shi and E. A. Valdez, "Insurance : Mathematics and Economics Multivariate negative binomial models for insurance claim counts," *Insur. Math. Econ.*, vol. 55, pp. 18–29, 2014.

[4] M. Kumar, R. Ghani, and Z. Mei, "Data Mining to Predict and Prevent Errors in Health Insurance Claims Processing Categories and Subject Descriptors," pp. 65–73.

[5] *FBI — Insurance Fraud*. Fbi.gov, 2005.

[6] S. Viaene, R. A. Derrig, and G. Dedene, "A case study of applying boosting naive bayes to claim fraud diagnosis," *IEEE Trans. Knowl. Data Eng.*, vol. 16, no. 5, pp. 612–620, 2004.

[7] Y. Li, C. Yan, W. Liu, and M. Li, "Research and application of random forest model in mining automobile insurance fraud," *2016 12th Int. Conf. Nat. Comput. Fuzzy Syst. Knowl. Discov. ICNC-FSKD 2016*, no. 61502280, pp. 1756–1761, 2016.

[8] A. C. Yeo, K. A. Smith, R. J. Willis, and M. Brooks, "Clustering Technique for Risk Classification and Prediction of Claim Costs in the Automobile Insurance Industry," no. October 2000, pp. 39–50, 2001.

[9] K. A. Smith, R. J. Willis, and M. Brooks, "An analysis of customer retention and insurance claim patterns using data mining : a case study," vol. 51, no. 5, pp. 532–541, 2015.

[10] P. Taylor, E. W. Frees, and E. A. Valdez, "Journal of the American Statistical Association Hierarchical Insurance Claims Modeling," no. December 2013, pp. 37–41.

[11] L. Mcclendon and N. Meghanathan, "U SING M ACHINE L EARNING A LGORITHMS TO A NALYZE C RIME D ATA," vol. 2, no. 1, pp. 1–12, 2015.

[12] Satadru Sengupta, "The Power of Machine Learning in Insurance - Cloudera VISION," 2017. [Online]. Available: https://vision.cloudera.com/the-power-of-machine-learning-in-insurance/. [Accessed: 27-Nov-2017].

[13] Kumba Sennaar, "How America's Top 4 Insurance Companies are Using Machine Learning -," 2017. [Online]. Available: https://www.techemergence.com/machine-learning-at-insurance-companies/. [Accessed: 27-Nov-2017].

[14] Jason Gran, “How Machine Learning is transforming the Insurance industry - Tellius Inc,” 2016. [Online]. Available: http://www.tellius.com/machine-learning-transforming-insurance-industry/. [Accessed: 27-Nov-2017].

[15] “COMPARATIVE ANALYSIS OF MACHINE LEARNING TECHNIQUES FOR DETECTING INSURANCE.”

[16] “Top 10 Machine Learning Algorithms,” 2017. [Online]. Available: https://www.dezyre.com/article/top-10-machine-learning-algorithms/202. [Accessed: 27-Nov-2017].

[17] Sebastian Raschka, “Predictive modeling, supervised machine learning, and pattern classification — the big picture,” 2014. [Online]. Available: http://sebastianraschka.com/Articles/2014_intro_supervised_learning.html.

APPENDIX