International Journal of Future Generation Communication and Networking
Vol. 13, No. 3, (2020), pp. 1855–1860

Motor Insurance Claim Processing and Detection of Fraudulent Claims Using Machine Learning

Mariya Mathew[1], Nimitha M Kunjumon[1], Ria Maria Lalji[1], Kency Susan Skariah[1], Dr Jeyakrishnan V[2]

[1]UG Students, Department of Computer Science and Engineering, Saintgits College of Engineering, Kottayam, Kerala- 686536, India, APJ Abdul Kalam Technological University
mariyamathew014@gmail.com, nimithanimi1998@gmail.com, riamaria584@gmail.com, kencysusans@gmail.com

[2]Associate Professor, Department of Computer Science and Engineering, Saintgits College of Engineering, Kottayam, Kerala- 686536, India, APJ Abdul Kalam Technological University
jeyakrishnan.v@saintgits.org

Abstract

A claim that is poorly handled will bring criticism through social media and hence detection of fraudulent claims are highly alarmed in the society. The paper focuses on estimating the insurance amount provided to the customer based on the severity of vehicle damage during an accident and detecting fraudulent claims. Random Forest is used to build a regression model that might be applied in forecasting the insurance claim. The algorithm involves identification of associations between claims, high dimensionality application to cover all levels, identification of missed observations, etc. By that way, the portfolio is made for the particular customer. An automobile insurance company's actual data was chosen to create the random forest model centered on the automotive insurance fraud removal theory. The data's were analyzed, and the dependency of each input variable to the corresponding output variable was obtained. The error of each model was examined. Finally, the model was tested by empirical study. The empirical findings indicate that: the automotive insurance fraud mining paradigm that incorporates Random Forest is ideal for wide data sets and unstable data relative to the traditional model. It can be well used to predict the claim amount to be given to the insured person and to detect the fraudulent claim of insurance.

Keywords: *Machine Learning, Random forest, Multiple Linear Regression, Overfitting, Data Mining.*

1. Introduction

Time and money are essential factors for the compilation of claims. The standard approach is to collect insurance manually. Insurers are liable on their balance sheet to compensate for potential settlements of claims on policies previously issued and this responsibility is known as the Unpaid Claims Fund [1]. There is a huge quantity of data to store in the database, and the size and complexity of the database makes it difficult to determine manually, so it is relevant to automate the system to help the process. The goal of the project is therefore to replace the current manual system with a system that automatically calculates the value of the insurance payment based on vehicle damage and calculates false insurance claims. For the insurance industry, it is very difficult to detect the insurance fraud. The conventional strategy for detecting fraud focuses on growing heuristics around the sign of fraud. It is the most communal form of insurance fraud that can be committed based on a false claim to an accident [1].

The ability to estimate the right amount of claims has a huge effect on the management decisions and financial declarations of the insurers. The two main categorizations for fraud are, hard insurance and soft insurance fraud. Hard insurance fraud is termed as the deliberate forgery or false making of a human accident; whereas soft insurance fraud means that a person is having a reasonable claim, but forges a part of it [2]. Training is conducted using some part of the data collection, and then the remaining information is used to check the result. Auto insurance fraud is slowly increasing across the

ISSN: 2233-7857 IJFGCN

world, and the industry is growing. Auto insurance fraud is slowly increasing across the globe and the company is deeply concerned about mining automobile insurance fraud.

2. Literature Review

Works relating insurance claim prediction focus on a technique which accurately predict the claim amount and those relating fraud detection aims at developing a model which is capable of detecting the fraudulent claims more accurately.

[1] Dal Pozzolo, "Comparison of Data Mining Techniques for Insurance Claim Prediction"

In this thesis, techniques of data mining are applied for prediction of claim amount. In this article, the prediction will be fully dependent on the customer's vehicle. Insurance claim amount was predicted by using various data mining techniques and different methods were compared. The techniques used were Decision Tree, Random Forest, Naive Bayes, K Nearest Neighbors (KNN), Neural Networks (NN), Support Vector Machine (SVM), Linear Discriminant Analysis (LDA) and Unsupervised technique Principal Components Analysis (PCA).

[2] Jessica Pesantez-Narvaez, Montserrat Guillen and Manuela Alcaniz, "Predicting Motor Insurance Claims Using Telematics Data- XGBoost"

In this, Logistic Regression and another regression method XGBoost is compared on the basis of their predictability, that is how well or how accurate these algorithms predict the claim amount. In this paper, the data used for prediction is telematics. Hence, contains more vehicle information. Without model- tuning XGBoost showed more predictive performance than Logistic Regression for training sample but comparatively poorer for testing sample. By regularizing overfitting, the predictive performance of XGBoost and Logistic Regression became same. This shows that XGBoost needs additional tuning for achieving a higher prediction.

[3] Roel Verbelen, Katrien Antonio and Gerda Claeskens, "Unravelling the Predictive Power of Telematics Data in Car Insurance Pricing"

The predictive capability of telematics data is investigated in this research paper based on a Belgian telematics dataset. Earlier the data used for insurance premium is the data provided by the policyholder. But in this model, a black box is installed in the vehicle of policyholder and it keep track of the driving habits of policyholder like his speed and other characteristics. This work was mainly focused on young people and it aims at selecting variables that contributes more for prediction.

[4] G. Kowshalya and M. Nandhini, "Predicting Fraudulent Claims in Automobile Insurance"

A synthetic dataset is used in this case study for predicting whether a claim is fraudulent or not. Since the insurance dataset is not easily available, a mock dataset for insurance claim is created based on the case studies on insurance fraud. The estimation of the premium percentage and the prediction of fraudulent claims is based on a data mining process. In the dataset developed, there are two categories one for vehicle theft claim and the other for accident claim. Naïve Byes, J48 and Random Forest classification algorithms are used for prediction of fraudulent claims. For enhancing the accuracy of this model, data preprocessing is performed.

[5] T. Badriyah, L. Rahmaniah, and I. Syarif, "Nearest Neighbor and Statistics Method Based For Detecting Fraud in Auto-Insurance"

The study proposes a model for detection of fraud using nearest neighbor and statistics method. Here, two approaches distance-based and density-based in nearest neighbor method is used for detection of fraudulent claims. The resultant model is also compared with other models developed using the same dataset. For increasing accuracy, they also performed feature selection and concluded that the dataset which undergoes feature selection gives higher accuracy. This model cannot be used if the dataset is large enough, since selecting the attributes from it is a cumbersome process.

3. Proposed Model

Using Machine Learning, we propose a model which can help to predict the insurance charge and to detect fraudulent claims. Definitely, Random Forest can aid in improved claim handling by utilizing faster and more effective claim management processes. In the risk mitigation side, managing fraud and making sure that you are paying the right clients with right insurance for the right reasons can be accomplished with AI [3]. One delicate problem to be avoided when implementing predictive learning algorithms is overfitting. The basic steps that are commonly taken to resolve these issues are not only meant to prevent overfitting, but are the simplest to understand and use [4].

A. Block Diagram

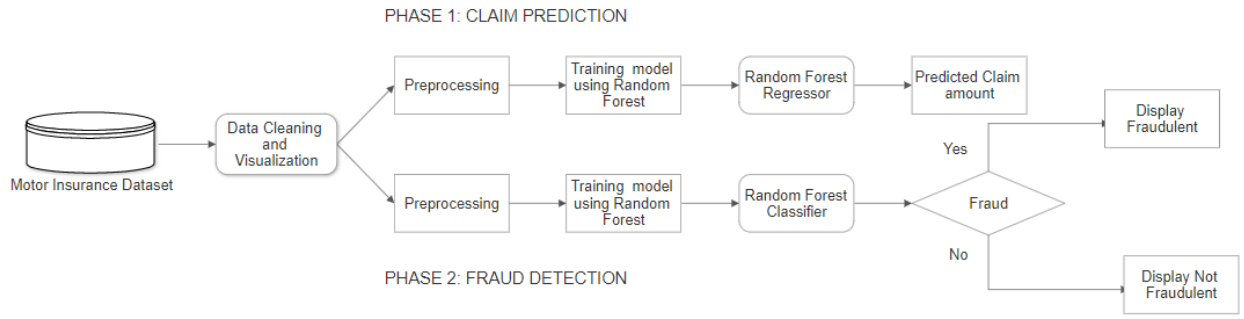

Fig. 1 Block Diagram of suggested model

The steps regarding model building process are:

a. Data pre-processing: This process comprises of cleaning up the data and choosing ones that will be included in the modeling step. In this paper, which deals with auto insurance fraud cases, the raw data consists of following attributes: Age, Policy_number, Policy_csl, Policy_deductible, Policy_annual_ premium, Insured_sex, Insured_relationship, Incident_type, Collision_type, Incident_severity, Number_of_vehicles_involved, Property_damage, Bodily_injuries, Auto_year, Auto_make, Total_claim_amount, Fraudulent_or_Not_Fraudulent.

b. Data splitting: The total dataset comprises of 1000 samples and is separated into a learning, testing and test sample, mitigating the overfitting that happens when a model tends to be incredibly effective on the same data used to determine the basic structure while displaying considerably less efficiency on undisclosed data[5].

c. Adjusting the chosen models to the training set: Most model families tend to set one or more tuning parameters in advance to establish a model individually.

d. Model selection: It is a valuation of which method among the experiments on a test set is better performed, rendering the findings generalizable to unused data[6]. Various models were tried for claim prediction and fraud detection, among them Random Forest model was chosen to be the one with higher accuracy for both the phases.

B. Random Forest Model

The algorithm randomly develop the forest, and an increase in the amount of trees in the forest creates a combined estimating process [7].The Model is trained using a Python library called Scikit-learn (sklearn) which is probably most useful library for machine learning in Python. It improves the model's forecast precision by summing up a huge number of classification trees.

a. Random Forest Model for Claim Prediction

We used the motor insurance dataset as the test model to create a model which predicts the claim amount using Random Forest. The dataset consists of 1000 observations, the attributes are described as dataset description in data preprocessing step. Among them, Total_Claim_Amount is the predicted output, and the other attributes except Fraudulent or_Not_Fraudulent are used for prediction. The tuning parameters used for claim prediction are n_estimators : 95, max_depth : 4, n_jobs : -1.

b. Random Forest Model for Fraud Detection

The same dataset exploited for claim prediction was used for detecting frauds. Fraudulent_or_Not Fraudulent is the predicted output in this model. The output variable for claim prediction, Total_Claim_Amount along with all other attributes are the input for this model. The tuning parameters for fraud detection are n_estimators : 500, max_depth : 3, n_jobs : -1, verbosity : -1.

4. Experimental Results

Various machine learning algorithms are used to predict charges and fraud detection. We have to choose the most accurate algorithm which contributes more accuracy. This step is important to compare how well different algorithms perform on a particular dataset. The project consists of two phases:

A. Claim Processing

This is to estimate the sum of insurance provided to the customer based on the severity of vehicle damage during an accident and to detect fraudulent claims. For calculating the insurance claim, the data was trained using various models.

TABLE. 1 Comparison of Random Forest with other models for Claim Processing

SI NO	ALGORITHM	ACCURACY (%)
1.	Multiple Linear Regression	59.20
2.	Decision Tree	41.05
3.	Adaboost Classifier	69.61
4.	Random Forest	76.25

Here, the highest accuracy was obtained by using Random Forest model. Root Mean Squared Error (RMSE) is actually the standard divergence of the prediction errors, which is also called residuals[8]. Hence RMSE is an estimate of how extensive these residuals are. n_estimator values controls the trees to be used in the process, since Random Forest itself is an ensemble method containing numerous decision trees.

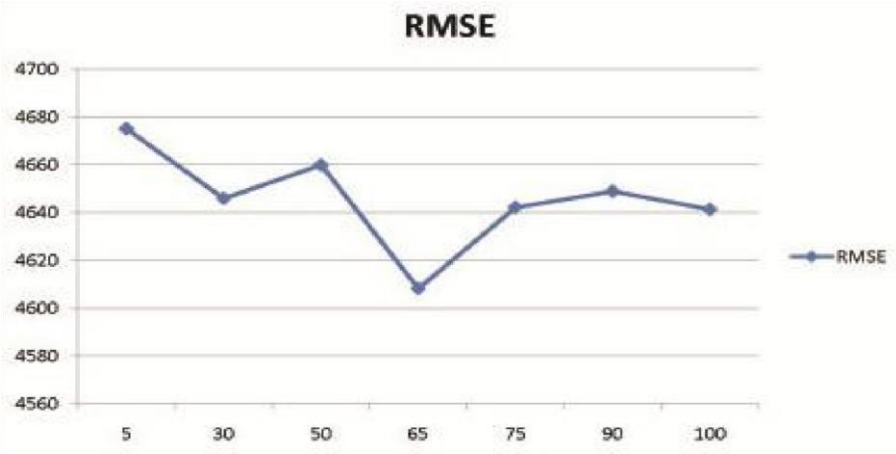

Fig. 2 RMSE v/s n_ estimator curve.

B. Detection of Fraudulent Claims

The dataset is trained using Random Forest classifier and AdaBoost classifier to make sure that a person does not make a fraud insurance claim in order to receive its benefits.

TABLE. 2 Comparison of Random Forest with other models for Fraud Detection

SI NO	ALGORITHM	ACCURACY (%)
1.	Adaboost Classifier	74.62
2.	Random Forest Classifier	91.01

For fraud detection, we used Random Forest Classifier which obtained the highest accuracy, 93.01%. The following ROC curve was obtained by plotting true positive (TP) rate and false positive (FP) rate at an assortment of thresholds. The true-positive rate and false positive rate are known as sensitivity and probability of false alarm respectively.

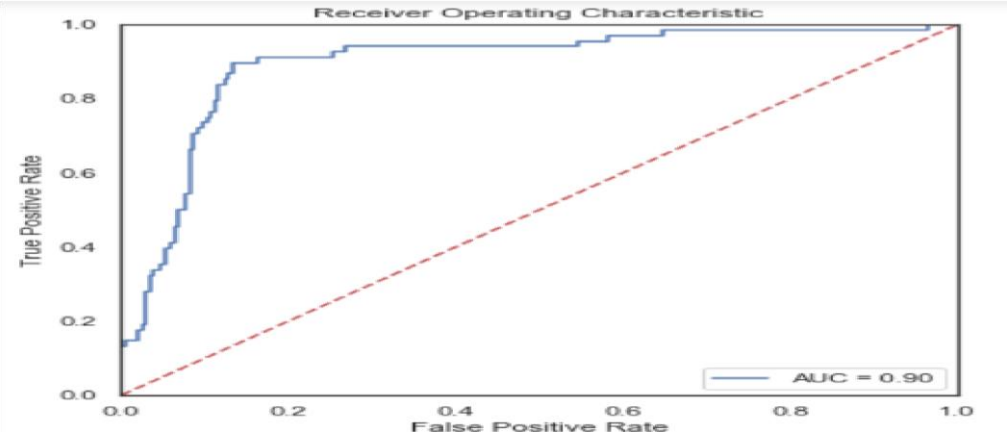

Fig3 Receiver Operating Characteristics

5. Conclusion

This paper suggests an approach to predict the motor insurance amount based on the concept of Random Forest. Our goal was accomplished by utilizing data mining procedures to generate a learning model that was created from past customer data. When insurance claim patterns are created it is easy to predict new insurance claim. Past consumer information is also significant. The Random Forest model used in this paper proposes a different approach for vehicle insurance fraud mining and has certain reference value. In this paper, we selected an insurance company's real data to create the random forest fraud mining model centered on the automotive mining insurance fraud theory. Therefore, the value of each input variable was measured and got the important input variables as the outputs. Hence the importance of each input variable on the fraud detection was obtained. Finally, the accuracy of different models were analyzed and found Random Forest better. Compared to the old model, this model of car insurance presenting Random Forest can be well used f or the classification and forecasting of motor insurance claims data and mining fraud rules.

Reference

1. Dal Pozzolo, ―Comparison Of Data Mining Techniques For Insurance Claim Prediction‖, (MSc thesis). Universit_adegliStudi di Bologna, 2010.
2. Jessica Pesantez-Narvaez, Montserrat Guillen and Manuela Alcaniz ―Predicting Motor Insurance Claims Using Telematics Data—Xgboost―Department of Econometrics, Riskcenter-IREA,Universitat de Barcelona, 08034 Barcelona, Spain, June 20, 2019

3. Rama Devi Burri, Ram Burri, Ramesh Reddy Bojja, SrinivasaRaoBuruga ―Insurance Claim Analysis Using Machine Learning Algorithms‖, International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-8,Issue- 6S4, April 2019.
4. RoelVerbelen, Katrien Antonio, GerdaClaeskens ―Unravelling The Predictive Power Of Telematics Data In Car Insurance Pricing‖, Appl. Statist. (2018) 67, Part 5, pp. 1275–1304, March 2018.
5. HanwuLuo, Xiubao Pan , Qingshun Wang , Shasha Ye , Ying Qian, ‖Logistic Regression And Random Forest For Effective Imbalanced Classification,‖ in IEEE 43rd Annual Computer Software and Applications Conference (COMPSAC),Milwaukee, WI, USA, USA, July 2019.
6. Prof. Gianluca Bontempi Dott. Yann Ael Le Borgne "Comparison Of Data Mining Techniques For Insurance Claim Prediction", Facolt`a di ScienzeStatisticheCorso di LaureaMagistrale in SistemiInformativi per l'Aziendae la FinanzaTesi di Laurea in Data Mining e SupportoalleDecisioni, 2010/2011 - Sessione II
7. Riya Roy, Thomas George K, ‖DETECTING INSURANCE CLAIMS FRAUD USING MACHINESofia Benbelkacem, Baghdad Atmani, ―RANDOM FOREST FOR DIABETICS DIAGNOSIS,‖ in IEEE International Conference on Computer and Information Sciences (ICCIS), Sakaka, Saudi Arabia, April 2019.
8. Alexandra Khalyasmaa ; Stanislav A. Eroshenko ; TejaPiepurChakravarthy ; VenuGopalGasi ;Sandeep Kumar YadavBollu, RaphaëlCaire, Sai Kumar Reddy Atluri, Suresh Karrolla, ―PREDICTION OF SOALR POWER GENERATION BASED ON RANDOM FOREST REGRESSOR MODEL,‖ ,in IEEE International Multi-Conference on Engineering, Computer and Information Sciences (SIBIRCON), October 2019.