Fraud Claims Detection in Insurance Using Machine Learning

Hritik Kalra[1], Ranvir Singh[2], Dr.T. Senthil Kumar[3]

[1]Dept. of Computing Science Eng., School of Computing, SRM Institute of Science and Technology, Kattankulathur, Chennai, India.
E-mail: hk3657@srmist.edu.in

[2]Dept. of Computer Science Eng., School of Computing, SRM Institute of Science and Technology, Kattankulathur, Chennai, India.
E-mail: rs8679@srmist.edu.in

[3]Dept. of Computing Technologies, School of Computing, SRM Institute of Science and Technology, Kattankulathur, Chennai, India.
E-mail: senthilt2@srmist.edu.in

Abstract

Insurance fraud is an illegal activity that is deliberately done for financial gain. In the present era, this is the prime severe problem encountered by various insurance companies all over the globe. In most cases, some shallow or deep holes in process of fraudulent claims investigation have been identified as a major cause. Due to this reason, the need of using computer techniques to curb fraud activities grew as a motivation, providing not only a reliable and robust environment for customers but also reducing fraud claims to a large scale. We presented our research by automating the assessment process of insurance claims through various data techniques, where identification of false claims will be automatized by utilizing Data Analytics and Machine Learning methods. Moreover, the system has the ability to potentially develop heuristics around fraud indicators. Therefore, this model creates a positive impact on the whole insurance industry with increased company credibility as well as customer satisfaction.

Keywords: Machine Learning, Data Analytics, Fraud Detection, Insurance Company's Reputation, Customer Satisfaction.

DOI: 10.47750/pnr.2022.13.S03.053

Introduction

Insurance Fraud is basically intentional deceit that can be done by or against an insurance company or an agent with the motive of monetary gain. It is a serious and acute growing menace as fraudulent insurance applications increase the burden on the community in the form of high premiums rates. Recent studies propose widespread recognition that classical methods of fraud identification are quite inaccurate and non-reliable. As a result, these concerns bring the attention of machine learning and data analytics communities to find a solution to this problem. Similarly, our proposed work differentiates fraudulent and non-fraudulent claims with high accuracy, so that only fraud cases need to be scrutinized and legit cases get claimed swiftly without wasting time and resources.

Need and Motivation

Insurance fraud involves multifarious illegitimate and illicit activities either by the claimant or by the insurer in order to achieve favorable benefits. According to current reports, insurance fraud costs several billion dollars to consumers every year. Therefore, there exists an exigency to seek a suitable way that can ascertain possible frauds with high precision and accuracy. As a result, this brings out the motivation to build an automated model which helps to increase the robustness, efficiency, and effectiveness of the insurance claim process. With the automation of process and reduction of human intervention, non-fraudulent claims are assessed and passed in far lesser time, maintaining customer satisfaction and the company's credibility

Literature Review

Rama Devi Burri et all [1] reviewed several machine learning and statistical techniques which are used to analyze insurance claims efficiently. They also mentioned various ways to use machine learning techniques in the insurance industry along with the challenges faced while implementing these techniques.

Shivani Waghade [2] gave a review on frauds that are committed in the health care and medical industry. They also mentioned various advanced machine learning and data mining techniques which can be used to detect these frauds by analyzing unusual patterns.

Dahee Choi et all [3] aims to recognize frauds of finances in the mobile payment system. The research while analyzing methods of data mining, rather than focussing on a single aspect, used a blend of both supervised as well as unsupervised techniques.

Sunita Mall et all [4] proposed work that recognizes frauds done in the industry of automobile. Multiple statistical techniques are used in this research like Logistic Regression to recognize fraud triggers and to calculate the probability of claims to check whether they are accepted or rejected.

Pinak Patel et all [5] aim is to identify and gauge the frauds in the health care industry using rule-based pattern mining. Outliers from statistical decision rules, k means clustering and association rule-based mining using gaussian distribution depict the fraud insurance claims in the given data.

Najmeddine Dhieb et all [7] identifies fraudulent claims in the auto industry by using approaches based on an extreme gradient boosting algorithm known as XGBoost. Various data analysis techniques like data cleansing, data exploration, and privacy-preserving are also used to clean, explore, and extract relevant features from the given data.

Soham Shah et all [8] developed an automated fraud detection application framework based on machine learning and XGBoost algorithms to get fraud claims accurately within a short period. Various data analysis techniques like - data validation, data Insertion, data preprocessing, and clustering are used to clean, validate and extract the relevant features from given data.

Vipula Rawate [11] aims to identify fraudulent claims in the healthcare industry by using a hybrid approach based on the advantages of both supervised and unsupervised machine learning techniques. Evolving clustering method is used to cluster insurance claims according to diseases and a support vector machine is used for the classification of duplicate claims.

Existing System

General Performance of Existing System: Different forms of fraud lead to various crimes, however, many cases involve intentional injury to an insured object or the purpose of obtaining assets without payment. It is a well cognizant fact that fraud cases had been evident even from the start of the insurance industry. The discovery of insurance fraud is already a daunting task as not all applications can be severely investigated. The process of detecting fraud in the insurance industry is not only expensive but also time-consuming. The method that is working so far is the computer machine instance. However, the existing technology in the past was pre-programmed, which means that a consistent template was designed to detect fraudulent applications; and if a particular claim fits that figure it will only be identified as illegal, or else it will not be recognized.

There are various AI methods with which frauds can be detected. Some techniques are as follows:

- To classify, combine cluster data, and segment, using data mining that can find rules in data and be able to highlight specific patterns, including those related to fraud.
- Professional programs to detect fraud in the form of laws.
- To automatize determining factors of false claims, ML techniques are employed.

Aim and Objectives

AIM:

The aim of this research paper is to find out whether the insurance claim filed by the customer is genuine or illegal by automatization of the process instead of applying the classical approach by evolving heuristics around triggers related to fraud.

OBJECTIVES

The present paper concentrates on the following objectives:

- To provide the procedure for detecting illegal claims of insurance.
- To ensure the processes of application have high reliability as well as accuracy.
- To lessen the existence of false claim cases.
- To minimize monetary losses of companies as a result of illegal fraud.
- For supporting the credibility of insurance organizations.
- For elevating client trustworthiness and satisfaction with the insurance organizations.

Proposed Methodology

- BLOCK DIAGRAM OF PROPOSED MODEL

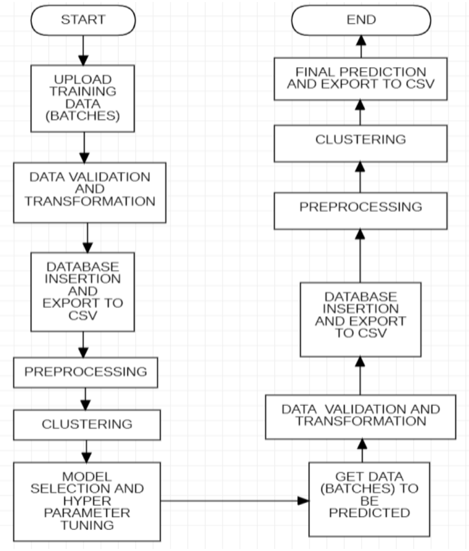

START
UPLOAD TRAINING DATA (BATCHES)
DATA VALIDATION AND TRANSFORMATION
DATABASE INSERTION AND EXPORT TO CSV
PREPROCESSING
CLUSTERING
MODEL SELECTION AND HYPER PARAMETER TUNING
GET DATA (BATCHES) TO BE PREDICTED
DATA VALIDATION AND TRANSFORMATION
DATABASE INSERTION AND EXPORT TO CSV
PREPROCESSING
CLUSTERING
FINAL PREDICTION AND EXPORT TO CSV
END

Fig. 1

- TRAINING STAGE

Firstly, data received in batches from the client is validated to check whether data is in the same format as agreed with the client, if not, gets discarded and sent to the archived folder. Then, data transformation is done, in which data format is amended to make it suitable for insertion in the database. In the next stage, this data is exported in the form of CSV and before doing data clustering, data is preprocessed. In the training stage, different models suitable for each cluster are chosen and, hyperparameter tuning or optimization is done to choose a set of parameters for optimized model learning. Then, the models are saved.

- PREDICTION STAGE

Once the training is completed, now, the model is fit for prediction. Now, data taken from the client for prediction is again checked, and appropriate changes are performed to make it suitable for insertion in the database. Subsequently, after pre-processing of data, it is sent for the process of clustering. Finally, to each cluster, a specific model is allocated. Then, prediction is performed accordingly. The output is exported to a CSV output file

- DEVELOPMENT STAGE

After setting up the Heroku cloud platform, required files needed for pushing the model are added to the model, and the application is pushed on the cloud. Now, the application is ready to get launched. After the application starts, input is taken for training, and predicted output is received in the form of an excel file.

MODULES

This tool contains the following main modules:

1. Data Validation - The data is divided into good and bad data based on the following parameters.

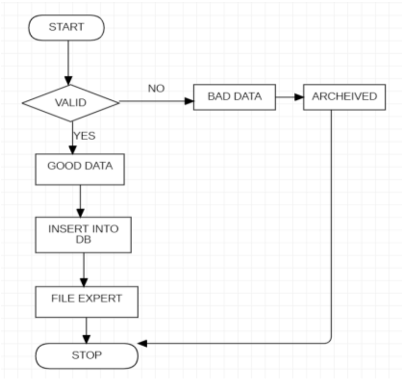


Fig. 2

   - Name Validation: We verify the file name based on the name provided in the schema file which is initially designed as per the agreement of the client. The name is verified with the help of a Regular Expression Function which checks for the name as well as the file type. Then, the valid file is sent to the good data folder, otherwise, it is sent to the bad data folder.
   - Number of Columns: After checking the file name, we confirm the number of columns available in the files. Here, the number of features must be the same as the client initially agreed to. If the number of columns is higher than the defined, some columns are dropped off, else if they are lesser, the complete file is sent to the bad data folder.
   - Column Name: The name of the columns is verified and must match the one provided in the schema file. If the name of columns is in single inverted commas, we convert them into double so that the database can read it as a varchar data type.
   - Nan values in Columns: We convert all Nan values to NULL which is readable to our database. In a file, suppose any of the columns is entirely containing missing values or NULL values, we discard such a file into a folder containing bad data.
2. Data insertion in a database: The requirement of a database is a necessity. As the client sends data in the form of various files, we would not make separate models for each file because it decreases efficiency. So, we combine all the data and make a single table out of it.
   - Database Creation and Communication: We build a connection to create an SQLite database with a given name and check if the database is already available. If yes, we open a link to that database, else we create a new database with that name.
   - Table creation in a database: A table with a specific name is created to insert files. If that table already exists, we add new files to that table, else we create a new table and insert files into it.
   - file Insertion: We write each CSV file in the table row by row until all files are inserted in the table created above. After inserting all the data, the good raw data folder is deleted as it is of no use now. Also, a bad raw data folder is archived.
   - Exporting Data: Data on a database is exported as a CSV file. Finally, this data serves the need for input for the purpose of model creation.

3. Pre-processing:

- Dropping columns: Initially, after scrutinizing the data, we rectify the table by dropping unnecessary columns which are not needed.
- Handling missing values: We find missing values in each column and impute them using the corresponding imputation strategy.
- Encoding: We extract the categorical columns and perform encoding. While custom mapping is done on ordinal variables, autoencoding is done on remaining using the Pandas framework.
- Correlation: By analyzing the correlation between each numerical column, columns that have a high degree of correlation are finally dropped.

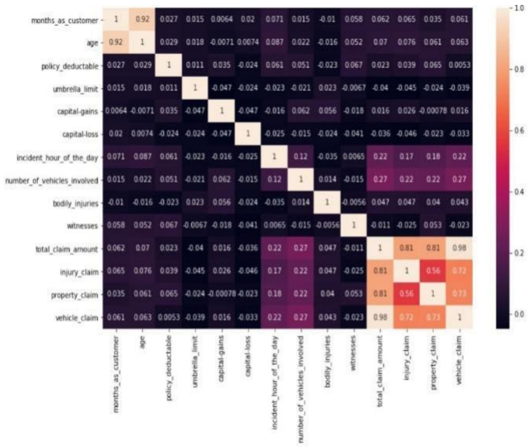


Fig. 3

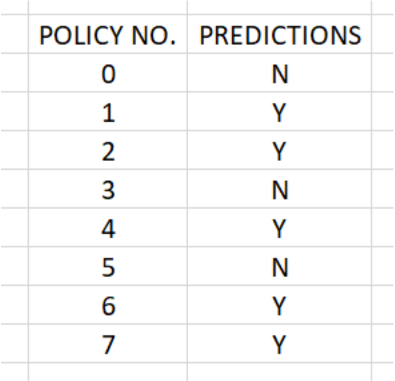
Here, we can see "age" with "number of months", as well as "property claim", "vehicle claim", and "injury claim", with "total claim amount" are highly correlated. Therefore, we can drop columns "age" and "total claim amount".

4. Training:

- Separating columns: Firstly, feature columns and target columns from the final table are separated for training.
- Clustering: Through various researches, we found that if the data is first divided into clusters and each cluster is trained using separate suitable models, then, it provides better accuracy than the data entirely trained using one model. Wherefore, to find out and create the optimal number of clusters, we use the K-Means clustering algorithm.
- Grouping: We group the whole data under different clusters created and each row is assigned with the corresponding cluster number.
- Model Selection: For every cluster, we try to choose the best model and also perform hyperparameter tuning. Finally, we compare the accuracy and select the best model for the particular cluster.

5. Prediction:

- Preparing Data: We prepare the data for predicting output by performing validation, insertion into the database, and pre-processing on prediction data similar to what is done on training data.
- Final Output: Now, clustering is done using the KMeans model developed during training and for each row appropriate cluster is predicted. Based on the cluster number, the corresponding model is loaded and output is predicted. Finally, the prediction is saved into a CSV file.



POLICY NO.	PREDICTIONS
0	N
1	Y
2	Y
3	N
4	Y
5	N
6	Y
7	Y

*N=No, Y=Yes

Fig. 4

6. Deployment:

- Host: We have performed the hosting and deployment of our project on Heroku Cloud.
- Working: It provides a simple user interface, where users can upload files to be predicted and after processing, our web application provides output files in the form of CSV.

Future Scope

The deceits in the contemporary world are getting complex and advanced with various different patterns. Therefore, it becomes equally important to analyze suspicious actions more carefully by doing intensive research in this area and consider those factors in detecting frauds using machine learning techniques. As an outcome, the machines will be able to develop better insights into data and correspondingly provide more efficient output while detecting fraud claims. Scrutinizing more real-world data and performing better data pre-processing using advanced methods may also help in increasing the efficiency of models. Although, it is always difficult to eradicate every false claim completely, bifurcating claims into fraudulent and non-fraudulent with high accuracy can be achieved.

Conclusions

In this research, the prime objective is to increase the revenue of the insurance industry by avoiding money wastage on false claims and increasing customer satisfaction by processing legit cases in very less time. However, at the same time subjecting non-legit cases to immediate inspection. The proposed work provides an automatic fraud detection application with no human intervention, which takes policy information as input to perform prediction as to whether the claim is legit or illegal within a fraction of time. We have used integration of XGBClassifier and SVMClassifier models while predicting which helped in increasing the accuracy and precision of the model to a large extent. The application provides the functionality to perform prediction with a default uploaded file, where the client can get an overview of the predicted output. Further, our web app allows the client to provide the absolute path of the custom input batch files, and the output file will be downloaded in a specified folder. The result consists of all the policy numbers along with the prediction as to whether the particular policy is verified as fraudulent or legit. Also, this framework allows companies to check any number of policies together at a given time which increases the overall throughput while accessing multiple policy claims. Therefore, present work can provide various monetary and credibility benefits to insurance organizations.

Acknowledgment

We would like to thank our guide Dr. T. Senthil Kumar for helping us in every step of the project, guiding us with every aspect of the project, motivating us at each stage, and aiding us with writing this research paper. We are also grateful to SRM Institute of Science and Technology (formerly known as SRM University) for believing in us and providing us with this great opportunity.

References

”Insurance Claim Analysis Using Machine Learning Algorithms” – Rama Devi Burri et all, IJITEE 2019

“A Comprehensive Study of Healthcare Fraud Detection based on Machine Learning” - Shivani S. Waghade, Int. J. Appl. Eng. Res. 2018

"Machine Learning based Approach to Financial Fraud Detection Process in Mobile Payment System" - Dahee Choi and Kyungho Lee, IT CoNvergence PRActice (INPRA), volume: 5, number: 4 (December 2017), pp. 12-24.

”Management of Fraud: Case of an Indian Insurance Company” – Sunita Mall et all, Accounting and Finance Research 2018.

“A Survey Paper on Fraud Detection and Frequent Pattern Matching in Insurance claims using Data Mining Techniques” –Pinak Patel et all, IRJET 2019

"The detection of professional fraud in automobile insurance using social network analysis" - Arezo Bodaghi et all.(2018)

“Extreme Gradient Boosting Machine Learning Algorithm for Safe Auto Insurance operations” – Najmeddine Dhieb, et all, LCVES, 2019.

“Insurance Fraud Detection using Machine Learning” – Soham Shah et all, IRJET 2021.

Diaz, Gonzalo & Fokoue, Achille & Nannicini, Giacomo & Samulowitz, Horst. (2017). An effective algorithm for hyperparameter optimization of neural networks. IBM Journal of Research and Development. 61.10.1147/JRD.2017.2709578.

Phua, Clifton & Lee, Vincent & Smith-Miles, Kate & Gayler, Ross. (2013). A Comprehensive Survey of Data Mining-based Fraud Detection Research (Bibliography).

“Fraud Detection in health insurance using data mining techniques“ – Vipula Rawte et all, IEEE 2015.

“An XGBoost Based System for Financial Fraud Detection” – Shimin Lei, et all, E3S Web of Conferences 2020.

“Analytics of Insurance fraud detection: An Empirical Study “ – Carol Anne Hargreaves et all, American Journal of Mobile Systems, Applications and Services.

"Predicting medical provider specialties to detect anomalous insurance claims." – Bauder, Richard A., Taghi M. Khoshgoftaar, Aaron Richter, and Matthew Herland, IEEE 2016.

Raghavan, Pradheepan & Gayar, Neamat. (2019). Fraud Detection using Machine Learning and Deep Learning. 334-339. 10.1109/ICCIKE47802.2019.9004231.

Rodriguez MZ, Comin CH, Casanova D, Bruno OM, Amancio DR, Costa LdF, et al. (2019) Clustering algorithms: A comparative approach. PLoS ONE 14(1): e0210236.

Piernik, M., Morzy, T. A study on using data clustering for feature extraction to improve the quality of classification. Knowl Inf Syst 63, 1771–1805 (2021).

Hämäläinen, Wilhelmiina & Kumpulainen, Ville & Mozgovoy, Maxim. (2014). Evaluation of Clustering Methods for Adaptive Learning Systems. 10.4018/978-1-4666-6276-6.ch014.

Sarker, I.H. Machine Learning: Algorithms, Real-World Applications and Research Directions. SN COMPUT. SCI. 2, 160 (2021).

"Tunability: Importance of Hyperparameters of Machine Learning Algorithms" - Philipp Probst et all, Journal of Machine Learning Research 20 (2019).

Ibrahim, S., & Koksal, M. E. (2021). Realization of a fourth-order linear time-varying differential system with nonzero initial conditions by cascaded two second-order commutative pairs. Circuits, Systems, and Signal Processing, 40(6), 3107-3123.