Insurance Fraud Detection using Machine Learning

Soham Shah[1], Shrutee Phadke[2], Princia Koli[3], Shweta Sharma[4]

[1,2,3]Student, Degree (Computer Engineering), Atharva College of Engineering, Mumbai University
[4]Assistance Professor (Department of Computer Engineering), Atharva College of Engineering, Mumbai University

---***---

Abstract - *Nowadays fraudulent insurance claims is the problem faced by many of the insurance companies which leads to huge financial loss yearly. These frauds have adverse consequences on society as the losses are settled down by increasing the premium cost of policy holders. Also the traditional claim investigation process being time consuming and tedious that generally leads to inaccurate results has been identified as main culprit. Thus, in this paper we develop an automated fraud detection application framework based on machine learning and XGBoost algorithm. The aim is to identify fraud claims accurately within shorter period of time. Throughout the process data analysis is used to validate, clean and extract the relevant data. Hence, by using this framework insurance company can maintain its respectability in outside world and can also share trustworthy relationship with customers.*

Key Words: **Fraud Insurance Claims, Traditional Process, Automated Framework, Machine Learning, XGBoost Algorithm, Data Analysis**

1. INTRODUCTION

Fraud in insurance is an unethical activity performed systematically to get some financial gain. These fraudulent claims present overpriced and large problem for insurance company leading to billions of dollars of needless expenses per year. Also due to some flaws in traditional process most of the companies are in search of some new technique to find fraud claims. So here we propose machine learning based automated framework employed with XGBoost algorithm to classify claims. We also compare the performance of XGBoost algorithm with other algorithms to obtain most accurate results.

2. LITERATURE SURVEY

Rama Devi Burri et all [1] presented several machine learning techniques to analysis insurance claims efficiently. They also mentioned three ways to transform machine learning techniques into insurance industry. Additionally they specified different resistances for adapting machine learning to classify claims and challenges in implementing machine learning. Also they evaluated the performance of different algorithms for claim predictions.

Pinak Patel et all [2] proposed a fraud detection framework for health care industry. They classified the fraudulent behaviour in two categories period based claim anomalies and disease based anomalies. Their framework was evaluated on real world medical data which showed efficient results to determine fraud claims.

Sunita Mall et all [3] proposed a study to identify important triggers of fraud and to predict the fraudulent behavior of customers using those identified triggers. They used statistical techniques to identify and predict the triggers.

Najmeddine Dhieb et all [4] evaluated the performance XGBoost algorithm for detecting and classifying different types of auto insurance fraud claims. They compared the proposed algorithm with other state-of-the-art solutions. Also they evaluated the algorithms for several metrics by applying data analysis and exploration techniques.

Shimin LEI et all [5] presented an XGBoost based system for financial fraud detection. They divided system into two parts as automatic part and manual part. The automatic part used large database to train model and manual part is used to monitor the transactions. Then for making final decision they combined machine scoring and manual feedback.

3. METHODOLOGY

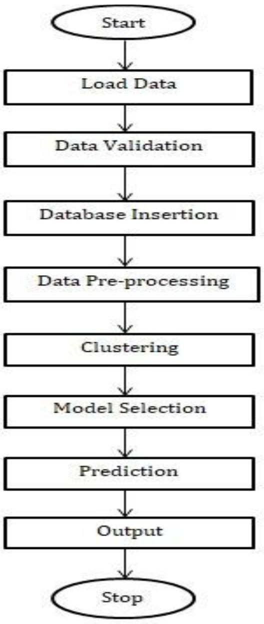


Fig-1: Flow Diagram of the Proposed Model.

Fig.1 shows the flow of the model where dataset is in form of csv (comma separated values) format. Classification is performed using the XGBoost (eXtreme Gradient Boosting) algorithm. K-Means Clusting technique have also been employed for the recognition task and as a result the recognition accuracy has improved significantly.

3.1 Data Validation

1. Number of Column - We validate the number of columns present in the files, and if it doesn't match with the value given in the schema file, then the file is discarded.

2. Name of column - The name of the columns is validated and should be the same as given in the schema file. If not, then the file is discarded.

3. The datatype of columns - The datatype of columns is given in the schema file. It is validated when we insert the files into Database. If the datatype is wrong, then the file is rejected.

4. Null values in columns - If any of the columns in a file have all the values as NULL or missing, we discard such a file.

3.2 Data Insertion in Database

1. Database Creation and connection - A database is created using the given name passed. If the database has already been created, open a connection to the database.

2. Table creation in the database - Table named as - "Good_Data", is being created in the database for inserting all the files accepted based on given column names and datatype in the schema file. If the table is already present, then the new table is not created, and new files are inserted in the already present table as we want training to be done on new along with old training files.

3. Insertion of files in the table - All the files which were accepted are inserted in the above-created table. If any file has invalid data type in any of the columns, the file is not loaded in the table and is discarded.

3.3 Data Preprocessing

The data stored in the database is exported and preprocessed using following methods:

1. Drop the columns not required for prediction.

2. Handling Missing values - For this dataset, the null values were replaced with '?' in the dataset. Those '?' have been replaced with NaN values. Check for null values in the columns. If present, impute the null values using the categorical imputer.

3. Handling Categorical values - Replace and encode the categorical values with numeric values. For this One_Hot_Encoding of Scikit-learn is used.

3.4 Clustering

After data preprocessing, the data is ready to be used for training the model. Before feeding the data to the model for training, the data is divided into number of clusters. To create cluster in the preprocessed data, KMeans algorithm is used. The ideal number of clusters is selected by plotting the elbow plot, and "KneeLocator" function is used for the dynamic selection of the number of clusters. The idea behind clustering is to implement different algorithms. The Kmeans model is trained over preprocessed data, and the model is saved for further use on prediction data.

3.5 Model Selection

After the clusters have been created, we find the best model for each cluster. We are using four algorithms, “Logistic Regression”, “Random Forest”, “SVM” and "XGBoost". For each cluster, all the algorithms are passed with the best parameters derived from GridSearch. We calculate the Accuracy, Precision, Recall, F1-score and ROC scores for all four models and select the model with the best score. Similarly, the model is selected for each cluster. All the models for every cluster are saved in .sav format for use in prediction.

4. RESULT AND ANALYSIS

To evaluate the performance of the machine learning algorithms (Logistic Regression, Random Forest, Support Vector Machine, XGBoost), following metrics were used: Precision, Recall, F1-score, ROC Area.

For each cluster, XGBoost algorithm has given better results when compared with Logistic Regression, Support Vector Machine and Random Forest.

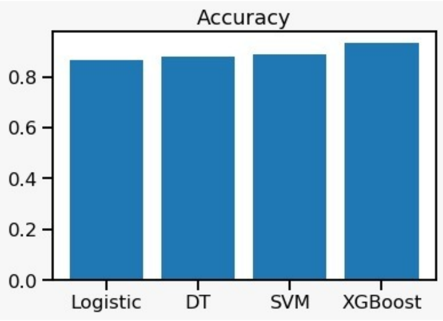


Fig-2: Accuracy Graph

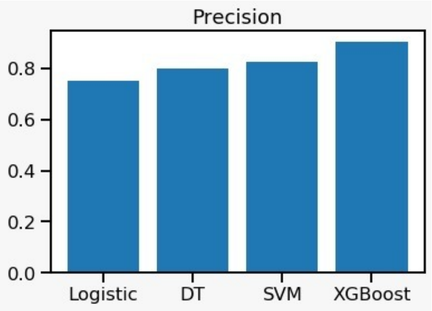


Fig-3: Precision Graph

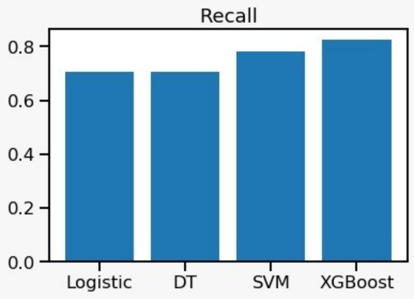


Fig-4: Recall Graph

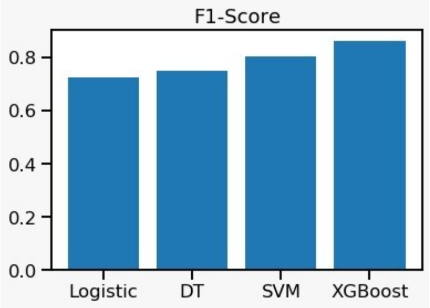
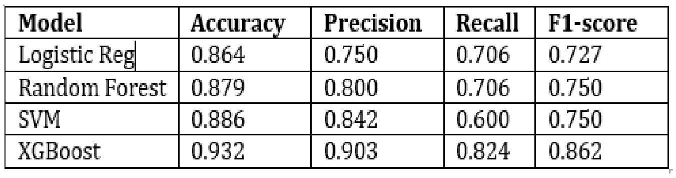


Fig-5: F1 Score Graph

Table 1 compares the performance measures of all the algorithms used for training the model.

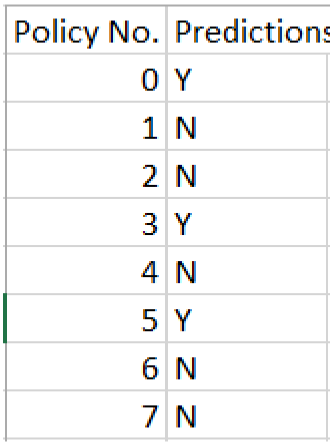
Table – 1: Performance measures of different models



Model	Accuracy	Precision	Recall	F1-score
Logistic Reg	0.864	0.750	0.706	0.727
Random Forest	0.879	0.800	0.706	0.750
SVM	0.886	0.842	0.600	0.750
XGBoost	0.932	0.903	0.824	0.862

From the table we conclude that the XGBoost algorithm based model outperformance other three models (Logistic Regression, Random Forest, SVM). Which means XGBoost can detect fraud more accurately than other three models.

The prediction provided by the model is stored as a csv file in which the first column provides the policy number and the second column gives prediction.



Policy No.	Predictions
0	Y
1	N
2	N
3	Y
4	N
5	Y
6	N
7	N

5. CONCLUSIONS

In this paper, we presented an automated model to identify fraud claims in insurance industry. As explained in section 4 the results for XGBoost algorithm contain 94% which is the highest precision accuracy for fraud detection problem with machine learning data. Hence by implementation of this model the insurance company can get accurate results in short duration of time. Thus, this automated framework can be used by any insurance company to reduce human labor and also to minimize the monetary loss in the insurance industry.

6. ACKNOWLEDGEMENT

No Project is ever completed without the guideline of those experts who have already traded this past before and hence become master of it. We would like to take this opportunity to take all those individuals who helped in visualizing this project. We are grateful to our principal Dr S.P. Kallurkar for providing us with an environment to complete our project successfully. We would like to express our gratitude to our H.O.D. Prof. Suvarna Pansambal for giving us valuable guidance. We would like to thank Prof. Shweta Sharma who was also our project guide for giving us valuable guidance and timely suggestion. Last but not least we would like to thank all the members of Computer Engineering Department who helped us directly or indirectly in completing our project.

7. REFERENCES

[1]" Insurance Claim Analysis Using Machine Learning Algorithms" – Rama Devi Burri et all, IJITEE 2019

https://www.ijitee.org/wpcontent/uploads/papers/v8i6s4/F11180486S419.pdf

[2]" A Survey Paper on Fraud Detection and Frequent Pattern Matching in Insurance claims using Data Mining Techniques" – Pinak Patel et all, IRJET 2019 https://www.irjet.net/archives/V6/i1/IRJET-V6I1104.pdf

[3]" Management of Fraud: Case of an Indian Insurance Company" – Sunita Mall et all, Accounting and Finace Research 2018

http://www.sciedu.ca/journal/index.php/%20afr/article/download/13474/8333

[4]" Extreme Gradient Boosting Machine Learning Algorithm for Safe Auto Insurance operations" – Najmeddine Dhieb, et all, LCVES 2019

https://www.researchgate.net/publication/337508754_Extreme_Gradient_Boosting_Machine_Learning_Algorithm_For_Safe_Auto_Insurance_Operations

[5]" An XGBoost Based System for Financial Fraud Detection" – Shimin Lei, et all, E3S Web of Conferences 2020

https://www.e3sconferences.org/articles/e3sconf/pdf/2020/74/e3sconf_ebldm2020_02042.pdf