Insurance Fraud Detection

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

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

---***---

Abstract *- Insurance fraud is one of the major problems facing many insurance companies of the world and some loop holes during traditional manual fraud investigation process have been identified as a major culprit. This is one of the motivations for this research, to deploy computing techniques in creating a barrier to fraud claims in order to not only provide trustworthy environment to the customers, but also to reduce the percentage of such illegal fraud activities to a greater extent. We presented our research by automating whole insurance claiming process through the use of different technologies in its design, development, and implementation. The system used machine learning and data analytics to automate the process of identifying fraudulent claims and has the ability to develop heuristics around fraud indicators. Thus, implementation of this model has good impact on insurance company's reputation in market and also on the customer's satisfaction.*

Key Words: **Fraud detection, Machine Learning, Data Analytics, Insurance Company's Reputation, Customer Satisfaction**

1. INTRODUCTION

Insurance fraud is deliberate illegal activity done to get financially benefitted. It is a serious and growing problem as fraudulent insurance claims increase the burden on society and as that demands the attention of communities such as machine learning to find solution to this problem. Also now there is a wide recognition that traditional approach for identifying the frauds is inadequate. At the time of building classify models; the savings from loss prevention needs to be balanced. The study on same reported that total cost of insurance fraud estimated is more than billion dollars worldwide.

1.1 Need and motivation

Insurance fraud covers the range of improper and illegal activities for achieving the favorable outcomes. Hence, there is a definite need to build an automated model capable for identifying potential frauds with high degree of accuracy. Also the model is required for improving the process efficiency and innovation.

Motivation for developing this model is to avoid the frauds in insurance claims by automatically acknowledging whether customer's claim is fraud or not. Also the other motivation is to maintain the customer's satisfaction by rapidly clearing their honest claims and scrutinizing the fraud identified cases in detail.

2. LITERATURE SURVRY

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 challenges in implementing machine learning.

Shivani Waghade [2] reviewed frauds in healthcare insurance industry including types of healthcare frauds and types as well as sources of healthcare data. She also reviewed techniques for detecting frauds such as machine learning.

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.

3. PRESENT SYSTEM

General Working of Existing System:

Insurance fraud exists since the beginning of insurance organizations. Different types of frauds lead to various crimes, however, in most of the cases in include deliberate damage to the insured item or the purpose to obtain goods without paying. Detection of insurance fraud turns out to be a tedious job as not every claim can be investigated toughly. Also this process is not only costly but also time consuming.

The efficient strategy till now is a computerized system. However, the technologies available in the past were pre-programmed, which means a particular fixed template was designed for detecting fraud claims; and if a particular claim fits in that template then only it is identified as flatulent or else it would not be recognized.

AI Techniques Used For Fraud Detection are as follows:

• To classify, cluster and segment data, data mining is used which can find rules in data and also is able to highlight some patterns, including those ones as well that are related to fraud.

• Expert systems for detecting fraud in the form of rules.

• Machine learning techniques are also used to automatically determine the characteristics of fraud claims.

4. AIM AND OBJECTIVES

4.1 Aim

The aim of ours is to determine whether the customer is claiming fraudulent insurance claim. We automate the process instead of using traditional approach of developing heuristics around fraud indicators.

4.2 Objectives

- To ease the task of finding insurance fraud claims.
- To provide quickness & high accuracy for claiming process.
- To minimize number of fraud claim cases.
- To reduce the amount of financial loss of company due to such illegals frauds.
- To maintain insurance companies reputation in the market.
- To improve customers trust & satisfaction about an insurance organization.

5. PROPOSED SYSTEM

5.1 System Design

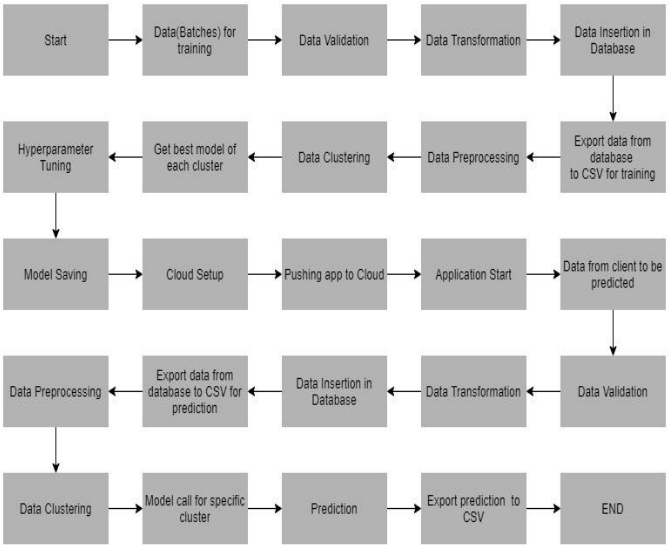
Start
Data(Batches) for training
Data Validation
Data Transformation
Data Insertion in Database
Export data from database to CSV for training
Data Preprocessing
Data Clustering
Get best model of each cluster
Hyperparameter Tuning
Model Saving
Cloud Setup
Pushing app to Cloud
Application Start
Data from client to be predicted
Data Validation
Data Transformation
Data Insertion in Database
Export data from database to CSV for prediction
Data Preprocessing
Data Clustering
Model call for specific cluster
Prediction
Export prediction to CSV
END

FIG 1: Block Diagram

The figure 1 shows the block diagram which explains how we implement this model. The whole diagram can be broadly divided into 3 main steps as follows:-

1. Data Ingestion Stage:

First the data from all the customers is collected. Then data validation is performed which means it is checked if the customer has sent data in correct format or not. After passing through all the validations some transformations on that validated data are done such as replacing missing values by null, etc. Then all data is aggregated in form of database and it is being exported as a csv file which acts as an input for training model.

2. Model Selection and Tuning:

In this stage data preprocessing is performed in which the data is further divided into several substeps. After cleaning all the needed data clustering is implemented before feeding it to the machine learning algorithm. After that hyper parameter tuning is performed to enhance efficiency of the model and at last the best fitted model is being selected. Then lastly best fitted model predicts inputted claims as fraud or legit.

3. Deployment:

Here an application model is pushed on cloud platform but before doing it some files that are required to run the model correctly on cloud platform are included in model. Then the predicted results can be seen in form excel file.

6. SCOPE AND FEASIBILITY

6.1 Scope

The model classifies claims as fraud or genuine therefore it can be used for an insurance company so that it can higher less employee for investigation process and also to get financially benefitted. The company will be capable to manage its reputation by knowing fraud claims in short time.

6.2 Feasibility:

6.2.1 Technical Feasibility:

Technical Feasibility includes the development of working prototype of the final product. Python is a powerful object-oriented, general purpose and simple to use programming language which is used by us in this project. Also we have used eXtreme Gradient Boosting popularly known as XGBoost which is one of the powerful machine learning algorithms to predict fraudulent claims accurately.

6.2.2 Economical Feasibility:

Economic feasibility can be referred as the cost and logistical outlook guideline for a business project. Cost incurring will be at time of deployment i.e. Heroku, GCP or AWS, Maintenance, which will be required if the user pool is higher than expected.

6.2.3 Operational Feasibility:

Operational Feasibility is the ability to utilize, support and perform the necessary tasks of a system. It includes everyone who uses the application. The solution uses python libraries and also machine learning techniques. It can be used by any type of insurance company. Also the solution can be deployed on any cloud platform depending on company's requirements.

7. MODULE

7.1 EXPLANATION

This tool comprises of following major modules:

1. Data Validation:

- Name Validation: We validate the name of the files based on the given name in schema file.
- Number of Columns: We validate the number of columns present in the files.
- Name of Columns: The name of columns is validated and should be same as given in schema file.
- Null values in Columns: Suppose any of the columns in a file contains all the NULL values or missing values, we discard such a file.

2. Data Insertion in Database:

- Database Creation and Connection: We create a database with given name and if it is already has been created we open connection to that database.
- Table Creation in database: Table with a particular name is created for inserting the files. If that table is already present then we insert new files in that table only instead of creating new table.
- Insertion of files in the Table: All the files are inserted in the above created table.

3. Prediction:

- Data Export from Db: The data in the stored database is exported as a CSV file for prediction.
- Data Pre-processing: Dropping the columns which are not required for prediction.
- Clustering: KMeans models are created during training is loaded, and clusters for pre-processed prediction data is predicted.

- Prediction: Based on the cluster number, the respective model is loaded and used o predict the data for that cluster.

4. Deployment:

- We deploy the model to the Heroku cloud Platform.

8. FUTURE SCOPE

From a future perspective, this project allows for multiple algorithms to be combined together in a single module and their results can be merged to increase the accuracy of the final result. This model can further be improved with the addition of more algorithms like addition Facebook Prophet library into it. More improvement can be found in the dataset. Thus, large amount of data will make the model more accurate in detecting frauds and reduce the number of false positives.

9. CONCLUSION

On implementing this technology it becomes helpful to the insurance company to identify genuine and fraud claims. As the model automatically acknowledges the insurance company owner about claim status instantly. Another advantage of this project is that it decreases the human labor. This technology can be implemented in any type of insurance organization .So with the implementation of this technology it is possible to optimize marketing strategies, to improve the business, to enhance the income of company, and to reduce costs due to fraud claims.

10. 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.

11. 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 Comprehensive Study of Healthcare Fraud Detection based on Machine Learning" - Shivani S.Waghade, Int. J. Appl. Eng. Res. 2018
https://www.ripublication.com/ijaer18/ijaerv13n6_140.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