DETECTION AND INVESTIGATION OF SUSPICIOUS INSURANCE CLAIMS USING DATA-DRIVEN APPROACHES

#1PRABIN KUMAR BISOI, *Assistant Professor,*
#2TRILOCHAN DAS, *Assistant Professor,*
Department of CSE,
ADARSHA COLLEGE OF ENGINEERING, ANGUL, ODISHA, INDIA.

ABSTRACT: The deliberate commission of an illegal act in order to profit is known as insurance fraud. This is currently causing problems for many insurance companies all over the globe. In most cases, the main problem is that there is a defect in the evaluation of false accusations. Therefore, measures to prevent fraud using computers had to be implemented. In addition to providing a risk-free setting for customers, this significantly reduced the number of fraud claims. To prove our point, we used Data Analytics and Machine Learning to automatically detect fraudulent claims and automated the process of examining insurance claims using several data methodologies. Heuristics for fraud indications may also be generated by the system. All throughout the insurance industry, this tactic boosts consumer satisfaction and company reputation.

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

1. INTRODUCTION

An act of insurance fraud happens when a policyholder fraudulently obtains financial recompense by misleading their insurance agent or firm. The problem is becoming worse quickly because insurance premiums are going up due to false applications. New studies show that the old ways of identifying fraud are inaccurate and unreliable. In order to find answers, the data analytics and machine learning groups are concentrating on these problems. By using our suggested process, we are able to successfully differentiate between genuine and fraudulent claims. Because of this, we can quickly discover potentially fraudulent claims and process valid ones.

2. LITERATURE REVIEW

The use of statistical analysis and machine learning in evaluating insurance claims was investigated by Rama Devi Burri and colleagues. They spoke about the problems that need fixing and the benefits that could come from using machine learning in the insurance industry.

Shivani Waghade looked into healthcare system operations and fraud in the medical field. They went a step further by compiling cutting-edge data mining and machine learning techniques that can be used to identify signs of counterfeit goods.

The mobile payment system has been the target of financial fraud investigations led by Dahee Choi and colleagues. This study didn't limit itself to just one data mining technique; instead, it included supervised and unsupervised approaches.

Sunita Mall et al. intend to detect fakes in the automotive industry with their planned study. Logistic regression is one of the statistical methods used in this investigation of statement dependability and fraud origins.

An approach to healthcare fraud detection and punishment that Pinak Patel and colleagues are planning to use is rule-based pattern mining. Anomalies within the normal distribution of statistical decision rules, k-means clustering, and association rule mining all point to insurance claim fraud.

The authors of the study, Najmeddine Dhieb et al., use methods that are based on the XGBoost technology to identify car insurance claims that are false. Data analysis makes use of a wide variety of approaches for cleaning,

examining, and extracting insights from data, including data exploration, privacy protection, and data purification.

The goal of the framework for an automated fraud detection software that Soham Shah and colleagues created was to improve the quality and efficiency of claims processing via the use of XGBoost techniques and machine learning. Data analytic methods such as clustering, variable selection, data insertion, and extraction are used for cleaning, validating, and transforming the data.

The hybrid approach developed by Vipula Rawate combines supervised and unsupervised machine learning methods for the purpose of detecting medical claim fraud. Similar insurance claims are grouped using a support vector machine. To group statements about the same situation, we use dynamic clustering.

3. EXISTING SYSTEM

Fraud can take many different forms, and each has its own criminal penalties. Theft without remuneration or intentional damage to someone else's property are prevalent types of fraud. The insurance sector has unfortunately always had to deal with the problem of theft. Because it is impossible to do exhaustive background checks on all applicants, detecting insurance fraud is difficult. Investigations into insurance fraud are costly and time-consuming. An example with a computer is the best way to demonstrate this point. Preemptive coding was once necessary due to technological constraints; as a result, a reliable framework for detecting fraudulent applications was established. If a claim follows this pattern, it would be considered fraudulent or rejected in some other way.

AI has multiple tools at its disposal to detect dishonesty. Here are a couple of approaches:

The use of data mining techniques for data segmentation, aggregation, and categorization in order to discover rules and patterns that were previously undiscovered, potentially pertaining to fraud.

Legislative corruption detection program that was custom-developed. Using ML to detect the root causes of claim fraud automatically.

4. PROPOSED METHODOLOGY.

BLOCK DIAGRAM OF PROPOSED MODEL

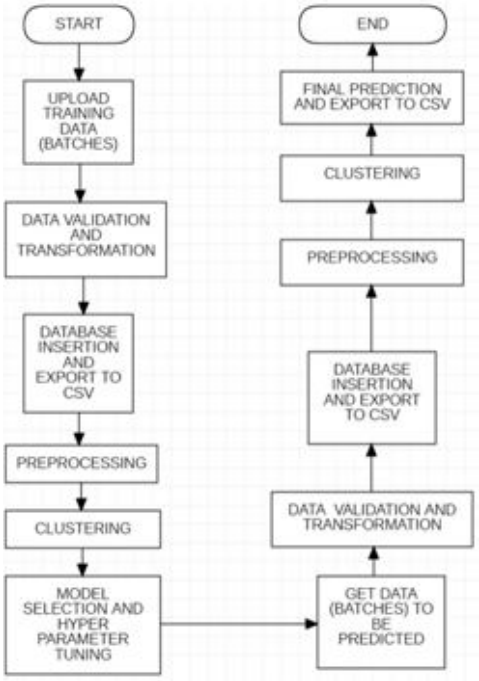

Fig. 1

TRAINING STAGE

The first step is to make sure the client's data is formatted correctly. Another option is to put the item in storage and then discard it. Next, we'll make some changes to the data so it can be easily entered into a database. After that, the data is saved in a CSV file, where it is first processed and then assembled. We apply any improvements or hyperparameter tweaks during training and choose the best models for each cluster. You must save the models right away.

PREDICTION STAGE

The training has allowed the system to now offer predictions. The client's prediction data is carefully reviewed and any necessary revisions are made before entering it into the database. The clustering process is applied once the data is cleaned and standardized. Every group receives a model in the end. With this, accurate predictions are made. An Excel spreadsheet contains the data.

DEVELOPMENT STAGE

The model is ready to be deployed to Heroku's

cloud platform as soon as the necessary push files are received. Anyone can now access the contribution. At the start of the program, we collect data for training purposes and give you an Excel file with all the expected results.

5. MODULES

Some of the most important parts of this program are these:

Data Validation-

The following criteria are used to classify the data as good or negative.

Name Validation:

The customer gave us permission to make the file, so we make sure that the name of the file matches the name in the schema file. By comparing the name to the file name, a Regular Expression Function verifies the name. The next step is to move the file to the designated subdirectory for successful data. Incorrect data is stored in a dedicated directory.

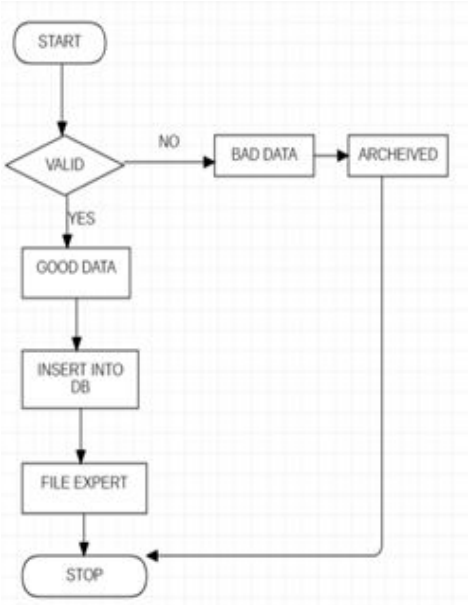

Fig. 2

Number of Columns:

The number of columns is determined when the filename has been authenticated. The agreed upon quantity of characteristics must match the quantity that was previously discussed and agreed upon with the client. If necessary, you can reduce the number of columns. The "bad data" subdirectory is where rejected files are saved when they do not have enough columns.

Column Name:

The schema file is used to verify that the column names are equivalent. Column names are changed from one backtick to two backticks to ensure the database recognizes them as varchar data types.

Nan values in Columns:

The database will work best if all Nan numbers are empty. The file is directed to the invalid data folder if a whole column is either blank or contains NULL values.

Data insertion in a database:

Databases are essential and should not be left out. No individual models will be developed in the event that the client supplies data in multiple file formats. A consolidated table contains all the information.

Database Creation and Communication

Initiating a connection is the first step in creating or checking the availability of a SQLite database. If it does, the database is connected; otherwise, a new database is established using the same name.

Table creation in a data base:

The processing of files is assigned to a dedicated table. An existing table is updated with new data. If it doesn't work, we'll create a new table and add files to it.

File Insertion:

The software iteratively processes all CSV files, adding each element to the table as it goes. The original carton can be discarded once data entry is finished. No action was taken on a file that included incorrect data.

Exporting Data:

The CSV file format allows for the export of data from databases. This data fixes the shortcomings of the model's original sources.

Pre-processing:

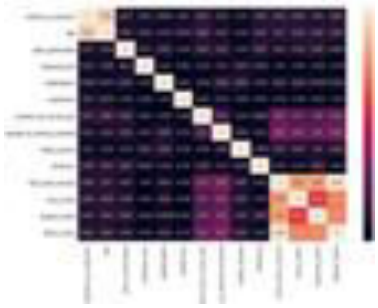

Fig.3

Dropping columns: First, we examine the data and eliminate any unnecessary columns.

Handling missing values : Discovering the missing values in each column and substituting them with the appropriate imputation technique is our process.

Encoding:

The data is classified once it has been captured. Specific variables are automatically wrapped and ordinal variables are uniquely translated by the Pandas system.

Correlation:

An examination of the relationships between the numerical values inside the columns is used to exclude those that show a high degree of similarity.

Preparing Data:

Following the same preprocessing steps as the training data, the prediction data is validated, uploaded to the database, and ready for use.Signifying "Yes," a "Y" is used, while a "N" implies "No."

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

N=No, Y=Yes

Final Output:

The KMeans model classifies information by assigning a post-training label to each input. Then, depending on the number of clusters, a forecast is made using the matching model. Next, the prediction is written to a file with values separated by commas.

Deployment:

Among the many variables that are strongly related to one another are "age" and "number of months," "property claim," "vehicle claim," "injury claim," and "total claim amount." For the moment, you can delete the "age" and "total claim amount" options.

Training:

- **Separating columns:** The first step in training is to identify which columns in the final table will serve as targets and which will serve as features.
- **Clustering:** According to our research, training multiple models on subsets of data improves accuracy more than training a single model on the entire dataset. As a result, the K-Means clustering algorithm is employed to determine the ideal cluster size.
- **Grouping:** Every node in the dataset is given a unique cluster ID, which allows it to be categorized into different groups.
  - **Model Selection:** The goal is to find the best model and tweak the hyperparameters for every cluster. Using the data, we determine which model is best for the given cluster.

Prediction:

- **Host:** The Heroku Cloud platform was utilized for the hosting and deployment of our application.
- **Working:** Improved teamwork on forecasting documents is a result of the web app's intuitive design. The files that are generated are then saved in the CSV format.

6. CONCLUSION

The main goal of this study is to help the insurance industry increase its profitability by decreasing the number of false claims and improving the settlement of valid claims. However, unfounded assertions are investigated promptly. A method for

automatically detecting fraud using policy knowledge to evaluate the validity of a claim without human intervention is proposed in the paper. By combining the XGBClassifier with SVMClassifier for prediction, the model demonstrated improved accuracy and precision. The customer can use a default file that has already been uploaded to create a prediction and find out what will happen. Clients can also specify the location of their batch files using our web service, making it easier to use them as input. The finalized document will be returned to the specified spot. Upon completion, you will possess a complete inventory of insurance numbers together with a deep understanding of the authenticity of each one. With this design, businesses may evaluate multiple policy claims at once, which boosts operational efficiency. Insurers stand to gain in more ways than one from this initiative's potential boost to their image and bottom line.

REFERENCES

1. Insurance Claim Analysis Using Machine Learning Algorithms – Rama Devi Burri et al., IJITEE, 2019
2. A Comprehensive Study of Healthcare Fraud Detection based on Machine Learning – Shivani S. Waghade, Int. J. Appl. Eng. Res., 2018
3. Machine Learning based Approach to Financial Fraud Detection Process in Mobile Payment System – Dahee Choi and Kyungho Lee, IT Convergence Practice (INPRA), Vol. 5, No. 4, pp. 12–24, December 2017
4. Management of Fraud: Case of an Indian Insurance Company – Sunita Mallet et al., Accounting and Finance Research, 2018
5. Prodduturi, S. M. K. (2024). Investigating the challenges and opportunities of cybersecurity in the era of remote work. European Journal of Advances in Engineering and Technology, 11(10), 80-84.
6. A Survey Paper on Fraud Detection and Frequent Pattern Matching in Insurance Claims using Data Mining Techniques – Pinak Patel et al., IRJET, 2019
7. Nandigama, N. C. (2016). Scalable Suspicious Activity Detection Using Teradata Parallel Analytics And Tableau Visual Exploration
8. Mallick, P. (2022). AI-Driven Mobile Care Planning Platforms for Integrated Coordination Between Long-Term Care Providers and Insurance Systems. Available at SSRN 6066586.
9. The Detection of Professional Fraud in Automobile Insurance using Social Network Analysis – Arezo Bodaghi et al., 2018
10. Extreme Gradient Boosting Machine Learning Algorithm for Safe Auto Insurance Operations – Najmeddine Dhieb et al., LCVES, 2019
11. Insurance Fraud Detection using Machine Learning – Soham Shah et al., IRJET, 2021
12. An Effective Algorithm for Hyperparameter Optimization of Neural Networks – Diaz, Gonzalo; Fokoue, Achille; Nannicini, Giacomo; Samulowitz, Horst, IBM Journal of Research and Development, 61, 2017
13. Todupunuri, A. (2024). Develop Machine Learning Models to Predict Customer Lifetime Value for Banking Customers, Helping Banks Optimize Services. International Journal of All Research Education & Scientific Methods, 12(10), 1254–1259. https://doi.org/10.56025/ijaresm.2024.1210241254
14. A Comprehensive Survey of Data Mining-based Fraud Detection Research (Bibliography) – Phua, Clifton; Lee, Vincent; Smith-Miles, Kate; Gayler, Ross, 2013
15. Fraud Detection in Health Insurance using Data Mining Techniques – Vipula Rawte et al., IEEE, 2015
16. An XGBoost Based System for Financial Fraud Detection – Shimin Lei et al., E3S Web of Conferences, 2020