MACHINE LEARNING FOR DETECTING AND INVESTIGATING POTENTIALLY FRAUDULENT INSURANCE CLAIMS

[#1]KANAPARTHY KHYATEESHWAR, Dept of *MCA*,

[#2]Dr.V.BAPUJI, *Associate Professor& HOD*,

Department of Master of Computer Application,

VAAGESWARI COLLEGE OF ENGINEERING, KARIMANGAR, TELANAGANA

ABSTRACT: Insurance fraud is a deliberate criminal act committed with the intent of profiting financially. This is currently the most critical issue confronting many insurance firms throughout the world. The primary issue in the majority of cases has been found as one or more gaps in the investigation of false claims. As a result, the desire to deploy computer solutions to combat fraud activities arose, providing clients with not only a dependable and stable environment, but also significantly reduced fraud claims. We demonstrated our findings by automating the examination of insurance claims utilizing a range of data methodologies, with the detection of erroneous claims performed automatically using Data Analytics and Machine Learning techniques. Furthermore, the system may be able to generate heuristics for fraud indicators. As a result, this technique benefits the whole insurance industry by improving both firm reputation and customer satisfaction.

Keywords: MachineLearning,DataAnalytics,FraudDetection,InsuranceCompany'sReputation,CustomerSatisfaction.

I. INTRODUCTION

If you lie to your insurance company or agent for financial gain, you are committing insurance theft. Because of the ripple effect that fraudulent applications have on insurance premiums, this problem is rapidly expanding in severity. According to recent research, conventional methods of detecting fraud are both inaccurate and unreliable. These concerns have piqued the interest of the machine learning and data analytics groups, who are eager to find a solution. Similarly, our proposed work accurately distinguishes between fraudulent and non-fraudulent claims, allowing for efficient processing of valid claims and efficient investigation of just potentially fraudulent situations.

NEED AND MOTIVATION

It is possible for either the claimant or the insurance company to engage in illegal or unethical behavior in order to increase their financial gain. Recent estimates place the annual cost of insurance fraud at several billion dollars. Therefore, it is crucial to develop a method for accurately and precisely detecting such forgeries. That's why an automated model is being developed to streamline and improve the entire insurance claims procedure. Legitimate claims are assessed and authorized in a lot less time thanks to automation and the reduced need for human interaction. This preserves patron satisfaction and shields the firm's credibility

LITERATURE REVIEW

Rama Devi Burri and coworkers [1] explored applications of machine learning and statistical analysis in the context of evaluating insurance claims. They discussed not only the potential benefits of using machine learning in the insurance industry, but also the challenges that must be overcome.

Shivani Waghade investigated the practice of health care and medical fraud. They also provided a list of state-of-the-art machine learning and data mining techniques that can be utilized to identify telltale

signs of forgery.

Finding instances of financial fraud in the mobile payment system is the focus of Dahee Choi et al. [3]. This research did not focus on just one form of data mining but rather made use of both supervised and unsupervised techniques.

Finding counterfeits in the automotive industry is the goal of the study proposed by Sunita Mall et al. [4]. Logistic Regression is just one of several statistical tools used in this investigation of the reasons of fraud and the acceptability of claims.

The mission of Pinak Patel et al.'s rule-based pattern mining for healthcare fraud detection and adjudication. In the provided data, anomalies in the Gaussian distribution of statistical decision rules, k-means clustering, and association rule-based mining stand in for fraudulent insurance claims.

To identify fraudulent insurance claims in the automotive industry, Najmeddine Dhieb et al. [7] employ methods based on the extreme gradient boosting technique (XGBoost). The data is cleaned, explored, and the relevant bits extracted using a variety of data analysis techniques, such as data cleansing, data exploration, and privacy protection.

In order to expedite and improve the quality of claims processing, Soham Shah et al. [8] developed an autonomous fraud detection application framework based on machine learning and XGBoost algorithms. Data cleaning, validation, and extraction using clustering, variable selection, data insertion, and other data analytic techniques.

The hybrid approach proposed by Vipula Rawate [11] combines supervised and unsupervised machine learning approaches in an effort to detect fraudulent healthcare claims. Similar insurance claims are sorted using a support vector machine, and claims associated with the same ailment are clustered using dynamic clustering.

II. EXISTING SYSTEM

There is a wide range of fraudulent activities, each with its own set of criminal outcomes. However, many instances of fraud entail wilfully damaging a valuable or obtaining valuables without paying for them. Theft has unfortunately always been a problem in the insurance industry. Insurance fraud is already difficult to detect due to the fact that not all applications can be thoroughly investigated. Investigating insurance fraud is a time-consuming and costly process. Using a computer as an illustration has proven to be the most effective method thus far. However, previously, technology had to be pre-programmed, therefore a trustworthy template was built to identify fraudulent applications. A claim would be flagged as fraudulent or otherwise not accepted if it conformed to such pattern.

Artificial intelligence can detect dishonesty in a number of ways. Here are a few illustrations of possible approaches: • Data mining-based classification, clustering, and segmentation, which can unearth previously unsuspected rules and patterns in data (including those pertaining to fraud). • Specially developed software to detect corruption in the legislative branch. Automatically identifying the root causes of claim fraud using machine learning approaches.

III. PROPOSED METHODOLOGY

AIM AND OBJECTIVES

AIM:

Instead of manually constructing heuristics around fraud-related events, the purpose of this research is to automate the process of determining whether or not a customer's insurance claim is legitimate.

OBJECTIVES

- The primary foci of this paper are the following: A method of identifying fraudulent insurance claims needs to be developed.
- So that the application procedure is foolproof.
- In an effort to reduce the number of fraudulent insurance claims.
- So that less money is lost by companies as a result of fraud.
- To verify the legitimacy of insurance providers.
- In an effort to boost satisfaction rates and boost

confidence in insurance firms, naturally.

BLOCK DIAGRAM OF PROPOSED MODEL

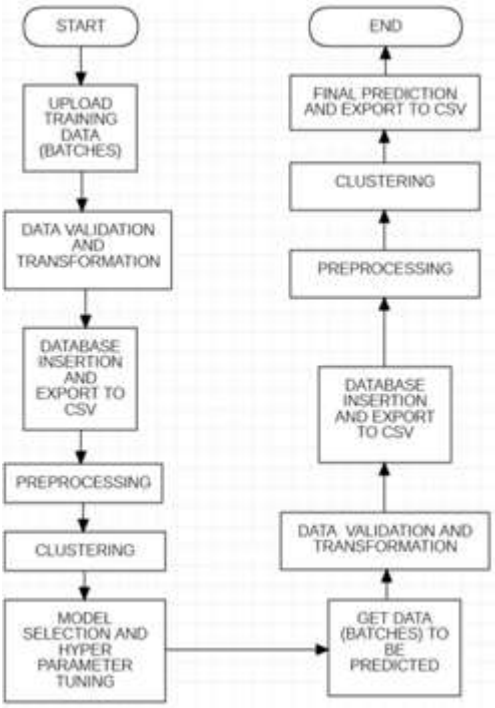

Fig. 1 block diagram

TRAINING STAGE

The first step is verifying that the client's data is provided in the expected format. If not, the information is rejected and placed in the archive. To prepare the information for entry into a database, the next step is data transformation. This information is then written to a comma-separated values (CSV) file, where it undergoes preliminary processing before being aggregated. During training, we select the most appropriate models for each cluster and perform any necessary hyperparameter adjustment or optimization. Models must be saved as soon as possible.

PREDICTION STAGE

Now that it has been trained, the machine can make forecasts. The client's prediction data is reviewed once more, and any necessary adjustments are made, before being entered to the database. Once the data has been cleaned and normalized, it is sent on to the clustering technique. Finally, a model is assigned to each classification. After that, reliable forecasts can be produced. The information is written out to a CSV file.

DEVELOPMENT STAGE

The model is prepared for deployment to Heroku's cloud platform by providing the necessary push files. The submission can now be made available to the general audience. After the program has begun, training data is collected and a file in Excel format detailing the expected outcomes is given.

IV. MODULES

The following components make up the backbone of this program:

Data Validation-

The data is either excellent or poor depending on the following criteria.

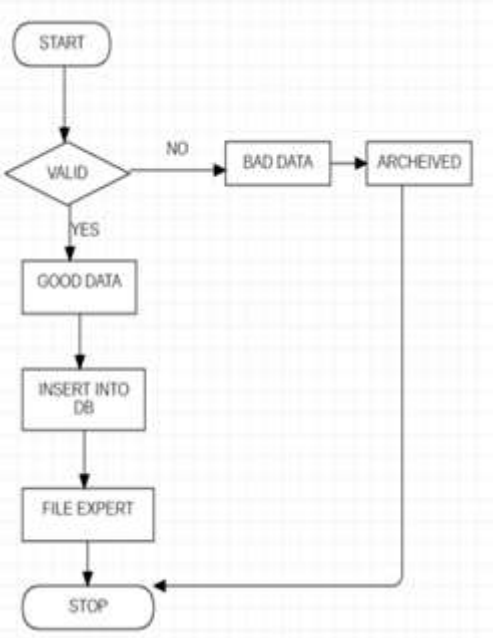

Fig. 2 model digagram

Name Validation:

The schema file was created with the client's blessing, therefore we check that the file name is the same as it is in there. The name is checked

using a Regular Expression Function that analyzes the name and the file's name. The successful file is subsequently transferred to the successful data folder. Bad information is stored in a separate directory..

Number of Columns:

First, we make sure the filename is correct, and then we count how many columns we have. In this scenario, the quantity of features must be consistent with what was first discussed and agreed upon with the client. The number of columns is reduced if necessary. If a file doesn't have enough columns, it gets discarded and placed in the "bad data" subdirectory.

Column Name:

Column names are validated against the schema file to ensure consistency. To ensure that the database treats a column as a varchar data type, a single reversed comma in its name is converted to two commas.

Nan values in Columns:

For our database to understand Nan numbers, they must all be blank. If a file's whole column contains missing or NULL values, the file is transferred to the invalid data folder.

Data insertion in a database:

A database is essential and cannot be avoided. If the client were to send data in more than one file format, we wouldn't create separate models for each file. A consolidated table has all the data.

Database Creation and Communication

In order to create and verify the availability of a certain SQLite database, a connection is established. If so, a connection to that database is established; otherwise, a new database is created and given the same name.

Table creation in a data base:

A specific table is created for reading in files. New data is appended to an existing table. If not, then a new table will be created and files will be added to it.

File Insertion:

One row at a time from each CSV file is read in and added to the table until all files have been processed. After everything has been entered, the original box containing the data can be discarded. In addition, there is a file containing corrupt information that has not been processed.

Exporting Data:

Information can be transferred from a database to a CSV file. Last but not least, these data finish up the sources missing from the model.

Pre-processing:

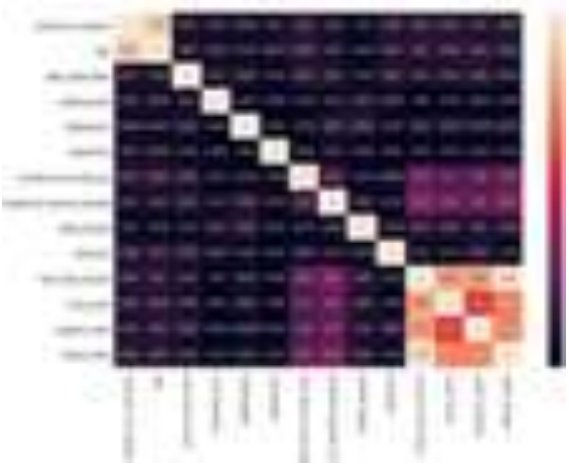

Fig.3 pre processing steps

Droppingcolumns:First, we examine the data and get rid of any unnecessary columns.

Handlingmissingvalues:Each column's missing values are determined and filled up using the appropriate imputation technique.

Encoding:

The category data is encoded after collection. Some variables are automatically encoded by the Pandas system, while ordinal variables are mapped in a special fashion.

Correlation:

Columns of numbers that are highly correlated are eliminated after being analyzed for their relationships to one another.

Preparing Data:

In the same way that the training data was checked, uploaded to the database, and preprocessed, so too is the prediction data.With a Y, you're saying "Yes," while a N indicates "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:

After being trained, the KMeans model classifies data by predicting which category each entry belongs in. The matching model is then loaded, and a prediction is made based on the total number of clusters. The forecast is then written to a comma-separated values file.

Deployment:

A substantial correlation exists between "age" and "number of months," as well as "property claim," "vehicle claim," "injury claim," and "total claim amount." The "age" and "total claim amount" boxes can thus be deleted.

Training:

- Separating columns: The first step in training is splitting the feature columns from the objective columns in the final table.
- Clustering: Our research shows that training several models on subsets of data improves accuracy over training a single model on the entire dataset. Therefore, we employ the K-Means clustering technique to establish an appropriate number of clusters.
- Grouping: After assigning a unique cluster identifier to each row, the dataset is partitioned into those groups.
- Model Selection: Finding the optimal model and tuning the hyper parameters for each cluster is our objective. Finally, we use the data to determine which model is more accurate for the specified cluster.

Prediction:

- Host: Our application's hosting and deployment were handled by Heroku Cloud.
- Working: The user-friendly layout of the online software makes it easy for users to collaborate on forecasting files. The output files are then stored as CSV documents.

V. CONCLUSION

The primary objective of this research is to help the insurance business increase profits by decreasing the number of fraudulent claims and increasing the number of successfully resolved claims. However, phony reports are investigated immediately. By using policy information as input to forecast whether a claim is true or not, the proposed study proposes a means to discover fraud automatically, without the assistance of a human. When creating predictions, we combined the XGBClassifier and SVMClassifier models, dramatically improving the model's accuracy and precision. The client can utilize a default file that has already been uploaded to the server in order to execute prediction and get an overview of the anticipated results. In addition, our web service allows users to specify the location of their own batch files to be used as input, with the resulting file being downloaded to that location. The end result is a comprehensive list of all the policy numbers and an educated guess as to whether or not each policy is genuine. This architecture not only improves efficiency but also allows businesses to verify numerous policy claims simultaneously. Therefore, the present work can provide insurance firms with a variety of monetary and reputational gains.

REFERENCES

1. ”Insurance Claim Analysis Using Machine Learning Algorithms” –

RamaDeviBurrietall,IJITEE2019
2. “AComprehensiveStudyofHealthcareFraudDetectionbasedonMachineLearning” -ShivaniS. Waghade, Int.J.Appl. Eng.Res.2018
3. "Machine Learning based Approach to Financial Fraud Detection Processin Mobile Payment System" - Dahee Choi and KyunghoLee,ITCoNvergence PRActice (INPRA), volume: 5, number: 4 (December2017),pp.12-24.
4. ”Management of Fraud: Case of an Indian Insurance Company” – SunitaMalletall,Accounting andFinanceResearch2018.
5. “A Survey Paper on Fraud Detection and Frequent Pattern Matching inInsuranceclaimsusingDataMiningTechniques ”–PinakPateletall,
6. IRJET2019
7. "Thedetectionofprofessionalfraudinautomobileinsuranceusingsocialnetwork analysis"-ArezoBodaghietall.(2018)
8. “ExtremeGradientBoostingMachineLearningAlgorithmforSafeAutoInsuranceoperations”–NajmeddineDhieb,etall, LCVES,2019.
9. “InsuranceFraudDetectionusingMachineLearning”–SohamShahetall,IRJET2021.
10. Diaz, Gonzalo & Fokoue, Achille & Nannicini, Giacomo & Samulowitz,Horst.(2017).Aneffectivealgorithmforhyperparameteroptimizationofneuralnetworks.IBMJournalofResearchandDevelopment.61.10.1147/JRD.2017.2709578.
11. Phua,Clifton&Lee,Vincent&Smith-Miles,Kate&Gayler,Ross.(2013).AComprehensiveSurveyofDataMining-basedFraudDetectionResearch (Bibliography).
12. “Fraud Detection in health insurance using data mining techniques“ –Vipula Rawteetall,IEEE2015.
13. “An XGBoost Based System for Financial Fraud Detection” – Shimin Lei,etall,E3SWebofConferences2020.
14. “Analytics of Insurance fraud detection: An Empirical Study “ – CarolAnneHargreavesetall,AmericanJournalofMobileSystems,ApplicationsandServices.
15. "Predicting medical provider specialties to detect anomalous insuranceclaims."–Bauder,RichardA.,TaghiM.Khoshgoftaar,AaronRichter,andMatthewHerland,IEEE2016.
16. Raghavan, Pradheepan & Gayar, Neamat. (2019). Fraud Detection usingMachine LearningandDeepLearning.334-339.
17. 10.1109/ICCIKE47802.2019.9004231.
18. Rodriguez MZ, Comin CH, Casanova D, Bruno OM, Amancio DR, CostaLdF, et al. (2019) Clustering algorithms: A comparative approach.PLoSONE14(1):e0210236.
19. Piernik,M.,Morzy,T.Astudyonusingdataclusteringforfeatureextraction to improve the quality of classification. Knowl Inf Syst 63,1771–1805 (2021).
20. Hämäläinen, Wilhelmiina & Kumpulainen, Ville & Mozgovoy, Maxim.(2014).EvaluationofClusteringMethodsforAdaptiveLearningSystems.10.4018/978-1-4666-6276-6.ch014.
21. Sarker, I.H. Machine Learning: Algorithms, Real-World Applications andResearch Directions. SNCOMPUT. SCI.2,160(2021).
22. "Tunability:ImportanceofHyperparametersofMachineLearningAlgorithms"-PhilippProbst et all, Journal ofMachineLearningResearch20 (2019).
23. Ibrahim, S., & Koksal, M. E. (2021). Realization of a fourth-order lineartime-varying differential system with nonzero initial conditions bycascaded two second-order commutative pairs. Circuits, Systems, andSignal Processing,40(6),3107-3123.