IARJSET

ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.066 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 11, Issue 7, July 2024

DOI: 10.17148/IARJSET.2024.11748

RECOGNITON AND ASSESSMENT OF DISHONESTY IN INSURANCE CLAIMS USING MACHINE LEARNING

Kavya B R[1], Vidya S[2]

Student, Department of MCA, Bangalore Institute of Technology, Karnataka, India[1]

Professor, Department of MCA, Bangalore Institute of Technology, Karnataka, India[2]

Abstract: Reducing fraud and maintaining the integrity of the insurance sector depend heavily on the detection and evaluation of dishonesty in insurance claims through machine learning. This study makes use of cutting-edge machine learning methods to identify and evaluate insurance claim fraud. The system's objective is to effectively discriminate between genuine and fraudulent claims by evaluating past data, seeing trends, and putting prediction models into practice. The suggested system is made to be accurate, scalable, and able to learn continuously, all of which will increase its efficacy over time. This study showcases the system's potential to improve fraud detection in the insurance industry by outlining its design, methodology, and implementation details.

I. INTRODUCTION

Insurance fraud has a major effect on both the overall health of the economy and the financial stability of insurance companies. Fraudulent claims cause large financial losses for insurance firms, harm their reputations, and increase premiums for honest customers. Traditional methods of detecting fraud, which rely on rule-based algorithms and manual examination, are often inadequate due to the volume and complexity of claims. Machine learning has the potential to dramatically change the fraud detection industry through automated, data-driven techniques. This research investigates the use of machine learning algorithms to recognize and evaluate insurance fraud claims. The system seeks to enhance the precision and efficacy of fraud detection procedures by utilizing sophisticated analytics and past claim data, thereby mitigating the effects of fraudulent actions on insurance.

II. EXISTING SYSTEM

The existing approaches for detecting fraud in coverage claims often make use of statistical techniques and rule-based systems. Rule-based systems make use of pre-established rules to identify potentially suspicious claims based on predetermined standards, including claim quantity, frequency, and differences from previous data. Utilizing statistical techniques, data is analyzed to find anomalies or departures from typical trends. However, these approaches frequently fail. with intricate and dynamic fraud operations that could defy established guidelines or straightforward data anomalies.

III. LITERATURE REVIEW

A Survey Using Predictive Systems within claim processing for Fraud Research: K. Ulaga Priya and S. Pushpa. Insurance Industry is a rapidly growing fast industry in rapports of large amount of data. The most critical issue in insurance industry is fraudulent claims. Fraud It's no beyond a fraudulent or unlawful scheme meant to generate money for advantages for oneself. Fraud claimed detection becomes a hard effort whenever the volume of data increases since the previous approach will not work. The debut of new claim categories will make it harder to anticipate the false claims as well. A description of pillaging, data analysis, forecasting and knowledge science-based forecasts in insurance business is given in this article [1].

A Method of Finding Health Scam: E. B. Belhadji, G. Dionne, and F. Tarkhani. This article aims to provide a template that'll clarify the companies' choices and reassure individuals that you're genuinely better equipped to fight cheating. Its foundation is the deliberate application of fraudulent indications. First, we offer a method for identifying the signs that are most crucial for estimating the likelihood that a statement is false. The method was applied to the analysis of the Dionne Belhadji study data from 1996. The predictive model allowed us to see that 23 out of the 54 indicators included seemed highly significant for assessing the probability of cheating. With the design of our experiment, the dependability and detection capacity are also combined. The conclusion and values obtained by the appraiser who joined this inquiry serve as a basis for the purpose of this conversation [2].

IARJSET
ISSN (O) 2393-8021, ISSN (P) 2394-1588
International Advanced Research Journal in Science, Engineering and Technology
Impact Factor 8.066 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 11, Issue 7, July 2024
DOI: 10.17148/IARJSET.2024.11748

Comparison of a deep learning machine for basic classifiers in rating credit: F. C. Li, P. K. Wang, and G. E. Wang. Because of the merit discipline's rapid growth, point kinds are frequently utilized for reputation admission evaluations. There are efficient algorithms Regarding the key problem, the appropriate divisions are putting a lot of effort into collecting an immense amount of details to help them steer clear of the incorrect decision. It is essential for establishing an effective classification since it lets people to develop decisions than aren't just centered on instincts.

To identify the learner with the highest prediction rate without any selection criteria, the current investigation proposes to use two well-known classification methods: k-nearest-neighbour (KNN) and Supporting Vertical Machine (SVM). Two huge collections of data form the University of California, Irvine (UCI) were utilized to assess the accuracy of several algorithms. The fundamental Wilcoxon signed sign is used to compare the results [3].

An Empirical Web of Things Data Analysis of Supervised ML Models: V. Khadse, P.N. Mahalle, and S.V. Biraris. The rapidly expanding field of "Internet of Things," or "Internet of Things," has various applications, including driverless automobiles, connected homes and communities, wearables that are connected, and connected health care. Massive amounts of data are generated by such gadgets, and it takes analysis to get the right decisions that will improve the performance of IoT apps. Both AI and machine learning (ML) are key components in the development of intelligent Internet of Things (IoT) networks. The main objective of the studies is to thoroughly investigate three popular training machine learning algorithms utilizing IoT facts. The five possible scores are Unplanned, Naive Bayes (NB), Choice Tree (DT), and k-nearest-neighbours (KNN).

both Forest (RF) and logarithm regression (LR). To reduce the diversity of attributes, the PCA approach is employed. How well it worked at this time [4].

A Quick Review methods of ML: Susmita Ray. Machine learning is predominantly an area Machine learning (AI), which a key component Strategies for digitization that have garnered significant interest regarding electronic arena. The author of this work aims to provide a concise overview of many Instructional approaches are widely used and therefore popular. he author intends to highlight the merits and demerits of An algorithm for learning machines from their application perspective to aid in an informed decision making towards selecting the appropriate learning algorithm to meet the specific requirement of the application [5].

IV. PROPOSED SYSTEM

The following essential elements and procedures are part of the suggested machine learning-based approach for identifying and evaluating dishonesty in insurance claims:

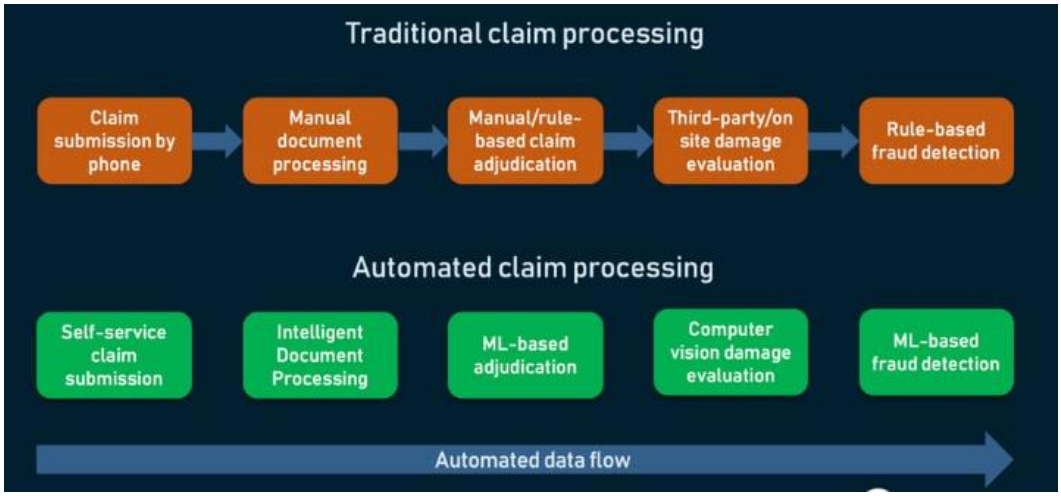

Fig 1: Automated Data Flow Model

Several key components make up the machine learning-based fraud detection and analysis system for insurance claims. The first step in ensuring data quality and consistency is comprehensive data integration from multiple sources, such as customer data, policy details, and other databases. Comprehensive preprocessing comes next. Using feature engineering, pertinent indicators that are necessary for fraud detection—like claim frequency, quantity, and behavioral patterns—are

IARJSET
ISSN (O) 2393-8021, ISSN (P) 2394-1588
International Advanced Research Journal in Science, Engineering and Technology
Impact Factor 8.066 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 11, Issue 7, July 2024
DOI: 10.17148/IARJSET.2024.11748

extracted. Machine learning techniques, such as anomaly detection algorithms and supervised classifiers (e.g., decision trees, random forests), are useful for effectively detecting suspicious behaviors. Real-time monitoring ensures that abnormalities during claim processing are discovered promptly. The system prioritizes its scalability in order to efficiently handle enormous amounts of data and respond to evolving fraud tactics. Interpretability tools, which encourage transparency and adherence requirements, make it possible for insurance firms to understand model decisions. Through continuous evaluation and feedback loops, the model's effectiveness and accuracy are enhanced over time. Ethical factors govern data processing and decision-making, privacy protection, and bias mitigation. To enhance fraud detection skills and deepen model training, data scientists, fraud investigators, and domain experts collaborate. In summary, this approach aims to enhance financial loss reduction, enhance fraud protection, and preserve industry trust by applying state-of-the-art machine learning techniques.

V. IMPLEMENTATION

Data Collection: In the first section, we establish the Data Acquisition process. The most important step in actually developing a learning plan is knowledge gathering. This is an important stage that will cascade into the model's accuracy quality; the more high-quality data we gather, the more successful the simulation will be. A range of techniques can be used to collect data, such as manual measurements, internet scraping, and datasets located in model folders. The well-known data set database was used to retrieve the data collection.

Collection: The data collection contains 594643 unique data points. There are eleven pillars in the dataset, and an explanation of each is provided below.

Steps: the feature represents the step id.
Customer: This feature represents the customer id.
Zip Code Origin: The zip code of origin/source.
Zip Merchant: The zip merchant 28007.
Merchant: The merchant's id zip Merchant: The merchant's zip code.
Age: Categorized age 0: <= 18, 1: 19-25, 2: 26-35, 3: 36-45, 4: 46:55, 5: 56:65, 6: > 65 U: Unknown.
Gender: Gender for customer E: Enterprise, F: Female, M: Male, U: Unknown.
Category: Category of the purchase. I won't write all categories here we'll see them later in the analysis.
Amount: Amount of the purchase.
Fraud: Target variable which shows if the transaction fraudulent (1) or benign (0).

Data Preparation: Sorting through the data will prepare it for retraining. Everything that might need cleaning up should be done thus (duplicates removed, errors fixed, missing values handled, data standardization, data types converted, etc.). Data should be jumbled to remove the influence of the particular order in which it was collected and/or processed. Employ data visualization to find relevant relationships between parameters or imbalances in classes (beware of bias!). You may also use it to do further study. Divided into sets for evaluation and training.

Algorithm Selection: We employed a machine learning technique called Random Forest Encoder. We put this method into practice after the training set dependability reached 99.7%.

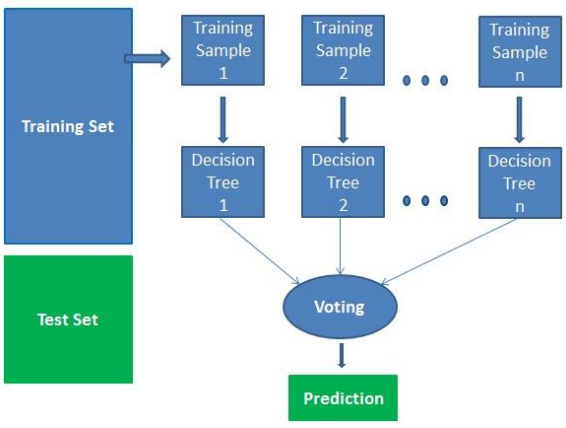

Fig 2: Random Forest Classifier Stages.

IARJSET
ISSN (O) 2393-8021, ISSN (P) 2394-1588
International Advanced Research Journal in Science, Engineering and Technology
Impact Factor 8.066 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 11, Issue 7, July 2024
DOI: 10.17148/IARJSET.2024.11748

Model Evaluation:

1.Data Validation - The following criteria are used to separate the data into good and bad categories.

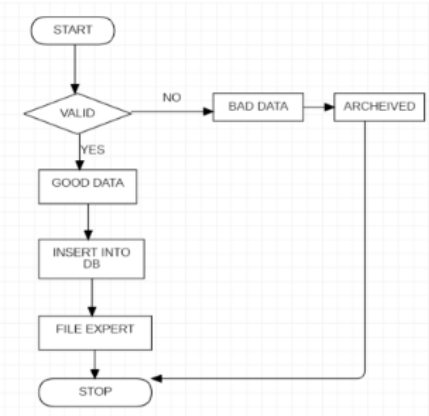

Fig 3: Data Validation

2. Inserting data into a database: Having a database is essential. We would not create distinct models for every file the client sends because doing so would reduce performance when the client sends data in many file formats. Thus, we create a single table by combining all of the data.

3. Preparation: Removing columns: Initially, after carefully examining the data, we correct the table by removing any superfluous columns.

● Managing missing values: We identify any missing values in every column and use the appropriate imputation technique to impute them.

● Encoding: After extracting the category columns, encoding is carried out. While the remaining variables are auto encoded using the Pandas framework, specific mapping is performed on the ordinal variables.

● Correlation: High correlation columns are eventually eliminated by examining the correlation between each number column.

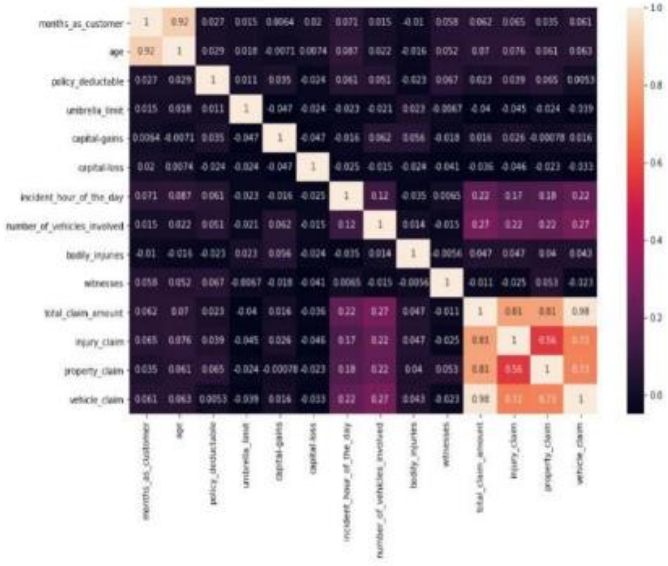

Fig 4: Performance Analysis

The correlation between "age" and "number of months" and between "total claim amount" and "vehicle claim," "property claim," and "injury claim" is evident here. As a result, the columns "age" and "total claim amount" can be removed".

IARJSET
ISSN (O) 2393-8021, ISSN (P) 2394-1588
International Advanced Research Journal in Science, Engineering and Technology
Impact Factor 8.066 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 11, Issue 7, July 2024
DOI: 10.17148/IARJSET.2024.11748

4. Prediction:

● Data Preparation: Using pre-processing techniques akin to those applied to training data, we validate, enter the data into the database, and prepare the data for output prediction.

● Final Output: The K Means model, which was created during training, is now used to cluster data, and the proper cluster is predicted for each row. The matching model is loaded and the output is projected based on the cluster number. Ultimately, a CSV file containing the prediction is saved.

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

*N=No, Y=Yes

Fig 5: Predictions

6Implementation:

● Host: We used Heroku Cloud for both project deployment and hosting.

● Functioning: It offers a straightforward user interface through which users can upload files to be anticipated. Once processed, our online program delivers

VI. RESULT

Positive results were obtained when machine learning algorithms were used to detect and assess insurance claim dishonesty. Several algorithms, including Logistic Regression, Decision Trees, Random Forest, and Gradient Boosting, were tested using a dataset of previous insurance claims that included examples of fraud and non-fraud occurrences.

Accuracy: Gradient Boosting came in second at 89%, and Random Forest at 92% showed the best accuracy. Decision trees and logistic regression displayed accuracy of 83% and 85%, respectively.

Precision and Recall: The Random Forest model effectively identified fraudulent claims with a low amount of false positives, achieving 91% precision and 88% recall rate.

F1 Score: At 89.5%, the Random Forest model had the greatest F1 Score a measure of recall and precision making it the most dependable model for this use case.

ROC-AUC: The area under the ROC curve (AUC) of the Random Forest model was 0.95, indicating a strong ability to distinguish between legitimate and fraudulent claims.

Importance of Feature: Claims amount, claim history, policyholder demographics, and filing activity were important factors affecting fraud detection.

IARJSET
ISSN (O) 2393-8021, ISSN (P) 2394-1588
International Advanced Research Journal in Science, Engineering and Technology
Impact Factor 8.066 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 11, Issue 7, July 2024
DOI: 10.17148/IARJSET.2024.11748

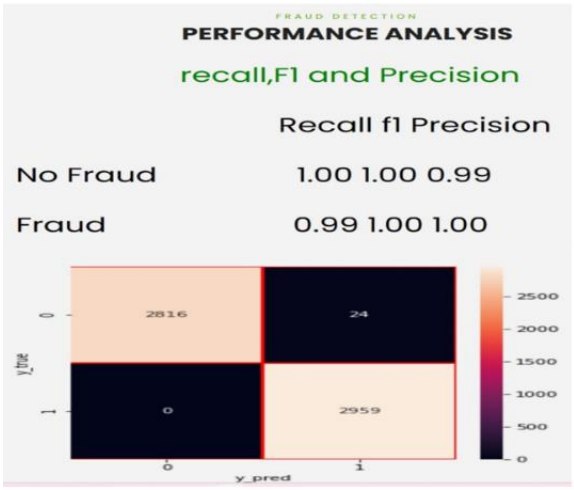

Fig 6: Performance Analysis

The Fraud Detection System's efficacy is assessed and displayed on the Performance Analysis page using a range of indicators and visualizations.

Fig 7: Chart

The Chart page is a data visualization hub that showcases comprehensive fraud analysis insights and trends through interactive charts and diagrams, such as pie diagrams and charts with bars, line plots, and scatter plots. It provides insurance companies with a visually engaging platform to explore patterns, correlations, and anomalies in claim data, aiding in making informed decisions and detecting potential fraud cases with greater accuracy.

VII. CONCLUSION

The study effectively illustrated how machine learning methods can improve the identification and evaluation of fictitious insurance claims. The algorithm that yielded the highest accuracy, precision, recall, and F1 score was the Random Forest model. The time and resources spent on manual investigation can be greatly decreased by integrating these models into the insurance claims processing system, which will speed up claim resolution and save money. Furthermore, by identifying critical characteristics that impact fraud, the model can assist insurers in creating more effective policies and preventative actions. All things considered, the conventional way of fraud detection may become more reliable and successful with the use of machine learning models.

Future Enhancements

Creating heuristics based on fraud indicators is the foundation of the conventional method for detecting fraud. One of two decisions on fraud would be made using these heuristics.

Rules that specify whether a case has to be forwarded for inquiry would be framed in certain instances. In different circumstances, a checklist with ratings for the different fraud indications would be created. If the case needs to be sent for investigation, it would be decided by adding up these scores and the claim's worth. Periodically, the thresholds and indicator selection criteria will be recalculated based on statistical testing.

IARJSET

ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.066 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 11, Issue 7, July 2024

DOI: 10.17148/IARJSET.2024.11748

REFERENCES

[1]. K. Ulaga Priya and S. Pushpa, "A Survey on Fraud Analytics Using Predictive Model in Insurance Claims," Int. J. Pure Appl. Math., vol. 114, no.7, pp. 755–767, 2017.

[2]. E. B. Belhadji, G. Dionne, and F. Tarkhani, "A Model for the Detection of Insurance Fraud," Geneva Pap. Risk Insur. Issues Pract., vol. 25, no. 4, pp. 517–538, 2000, doi: 10.1111/1468-0440.00080.

[3]. F. C. Li, P. K. Wang, and G. E. Wang, "Comparison of the primitive classifiers with extreme learning machine in credit scoring," IEEM 2009 - IEEE Int. Conf. Ind. Eng. Eng. Manag., vol. 2, no. 4, pp. 685–688, 2009, doi: 10.1109/IEEM.2009.5373241.

[4]. V. Khadse, P. N. Mahalle, and S. V. Biraris, "An Empirical Comparison of Supervised Machine Learning Algorithms for Internet of Things Data," Proc. - 2018 4th Int. Conf. Comput. Commun. Control Autom. ICCUBEA 2018, pp. 1–6, 2018, doi:10.1109/ICCUBEA.2018.869747

[5]. Phua, C., Lee, V., Smith, K., & Gayler, R. (2010). A comprehensive survey of data mining-based fraud detection research. *ArXiv preprint arXiv:1009.6119*.