USING MACHINE LEARNING TO FIND FRAUD IN BANK PAYMENTS

Prof. D. V. Varaprasad, M.Tech, (Ph.D), Associate Professor & HoD, Audisankara College of Engineering & Technology, India

Mrs V. Irine Shyja, Professor, Department of CSE, Audisankara College of Engineering & Technology ,India

Navakoti Snigdha, Department of CSE, Audisankara College of Engineering & Technology, India

Abstract:

A "financial fraud" is when someone gets money in a way that is dishonest and against the law. Organisations and businesses are becoming more and more worried about financial fraud, which is when people use dishonest tactics to gain money. Even though there have been many attempts to stop it, financial fraud still hurts society and the economy, costing billions of dollars every day. Antiquity is where many methods for finding dishonest actions came from. Handiwork is the standard, even though it has a lot of problems, such taking a long time, costing a lot of money, being prone to mistakes, and not being very effective. So far, no study has been able to cut down on losses due to fraud, but there may be more in the future. To find these fake businesses, traditional methods use labour-intensive, expensive, and error-prone human checks and verifications. Recent advances in AI (AI) have made it feasible to use machine learning-based algorithms to quickly look through huge volumes of financial data for signs of fraud. This study fills in the gaps in that knowledge by creating a new way to find fake bank payments using the Random Forest Classifier ML Algorithm. Our suggested method is better than the one that is already in use, as evidenced by a 99% accuracy rate on the train/test set.

INDEXTERMS: *Machine Learning, XGBClassifir, Decision Tree, Random Forest, UPI Digital Payments, and Fraud Detection..*

1. INTRODUCTION

Instead of being hard-coded from the start, what we call "machine learning" is really a group of algorithms that computers can run that can learn from what they see and do. Machine learning (ML) is a kind of artificial intelligence (AI) that uses statistical methods to look at data and create predictions that can help people make decisions.

The original notion came from the thought that computers could be able to learn from data samples and come up with correct answers on their own. Machine learning, data mining, and Bayesian predictive modelling are all very closely related. The computer takes in information, runs it through an algorithm, and then gives you a result. M

One big problem with Lis is that they make suggestions. Netflix picks films and TV shows based on how each user watches them. IT companies are using unsupervised learning to make personalised suggestions that make the client experience better.

ML can be used for a lot of things, such automating processes, optimising portfolios, predicting maintenance needs, and finding fraud.

How does ML work?

When it comes to learning, ML is like having an additional cerebral cortex. The way a computer learns is quite similar to the way a human brain learns. People learn by doing

things. We will be able to make better predictions after we have more data. For instance, we have a better chance of succeeding when we know the outcome than when we don't. All machines learn the same thing. It needs to view a sample before it can make a good guess. The computer can figure it out if we give it a similar example. But computers can't forecast things any better than people can when they don't know what they're looking at.

Learning and making guesses are at the heart of ML. The computer learns mostly by recognising patterns. The data led to this discovery. Data scientists need to be able to intelligently chose which data points to deliver to machines. A feature vector, which is a group of properties, can be used to solve an issue. The feature vector is like a small piece of data that is utilised to solve an issue.

2. LITERATURE SURVEY

i) Building a Robust Mobile Payment Fraud Detection System with Adversarial Examples

https://ieeexplore.ieee.org/abstract/document/8791686

More and more people in some countries are using mobile payments. Mobile payment fraud is worse than credit card fraud. If hackers can get to mobile data more easily than credit card data, they might be able to get beyond our data-driven fraud detection system. Supervised learning is a common method used in systems that look for fraud. The usual place for these supervised learning algorithms that look for fraud to work is one where there are no bad actors seeking to cheat the system. This study used adversarial situations to create a complete mobile fraud detection system that includes answers from fraudsters. Experiments with both good and bad outcomes demonstrated that our strategy performed better in certain situations than others.

ii) Induction Motor Failure Analysis using Machine Learning and Infrared Thermography

https://ieeexplore.ieee.org/document/10018653

Induction motors are used for many different kinds of industrial work. Their uses place them in an atmosphere that isn't good for them to work in. Infrared thermography can help doctors figure out whether an induction motor has failed. This work uses unsegmented infrared imaging and automatic learning to find and categorise problems with induction motors and kinematic chains. We use a machine learning technique to narrow down the set of characteristics and classify the problem scenario. To achieve this, we look at unsegmented infrared thermography and use a lot of statistical data that describe how the electromechanical system behaves thermally. This research looks at health problems that can happen with three types of defective induction motors: damaged bearings, misalignment, and a broken rotor bar. This aims to show how well the proposed approach works.

iii) Fraud detection system: A survey

https://www.sciencedirect.com/science/article/abs/pii/S1084804516300571

There are various types of electronic commerce that have come about because of the growth of both personal computers and big businesses. Some of these include credit card, telecommunications, and health insurance systems. Unfortunately, there are both honest and dishonest people on these networks. There were several ways that scammers might get into e-commerce sites. E-commerce networks are not well protected by fraud prevention systems (FPSs). FDS-FPS collaboration, on the other hand, can help defend electronic commerce systems. FDS has trouble working because of a number of problems, such as concept drift, real-time detection, and big data. This survey research looks at these problems and hurdles that make it hard for FDS to work in a systematic way. We have five e-commerce sites that offer online auctions, credit card processing, phone service, health insurance, and automobile insurance. We'll talk about the two primary types of internet shopping scams. Also, some of the newer FDS methods used by certain E-commerce companies are shown. This is a short summary of the patterns and findings that will guide future research.

iv) Comparative Evaluation of Credit Card Fraud (CCF) Detection Using Machine Learning Techniques

https://www.researchgate.net/publication/339019564_Comparative_Evaluation_of_Credit_Card_Fraud_Detection_Using_Machine_Learning_Techniques#:~:text=This%20publication%20inspects%20the%20execution%20of%2C%20Support%20Vector,is%20assessed%20dependent%20on%20accuracy%2C%20sensitivity%2C%20precision%2C%20specificity.

Because more people are shopping and buying things online these days, CCF is more frequent now. Many individuals all across the world might be hurt by sneaky plans like stealing someone's identity or losing money. Crime in the financial sector is getting worse and having major effects. The ways that credit card fraud is found, the factors that are employed, and the way that the dataset is measured all have an impact. ; yet, information extraction seems to be quite important for finding online payment fraud. We employ very biassed credit card fraud data to test k-nearest neighbour, logistic regression, support vector machines, naive bayes, and support vector machines. They are put through tests to check their accuracy, sensitivity, specificity, and precision. Logistic regression, k-nearest neighbour, support vector machine, and Naive Bayes all come very close to getting a perfect score of 99.07%. Logistic regression works better than other approaches.

v) Modelling different types of automobile insurance fraud behaviour in the Spanish market

https://www.sciencedirect.com/science/article/abs/pii/S0167668798000389

Microeconomic theory says that to stop insurance fraud, you need to know a lot about how insureds act. Using the discrete-choice models we suggest in this study, we measure the effect of insured and claim variables on the chance of fraud. The information comes from a sample in Spain. Because there are too many fraud claims, the estimation needs to be corrected for choice-based sampling. The structure of the Spanish automobile insurance industry is also discussed. The results we found depend on the type of fraud we were looking at.

3. METHODLOGY

a) Proposed work:

By looking at trends in transactions, the suggested method uses machine learning to improve the detection of fraud in bank payments. We use the Banksim dataset, which has fake financial transactions, to find important properties that make our model better at predicting. The system uses data preparation methods to clean and prepare the dataset, making sure that the model gets the best possible input for training. Then, a Random Forest Classifier is used to find fraud since it works well with both categorical and continuous data and helps prevent overfitting. The system checks how well the model works by looking at accuracy measures. It got 99% accuracy in both the training and testing phases.

The suggested method reduces the number of verification processes by automating fraud detection. This lets transactions be processed in real time. The machine learning model changes when fraud patterns change, unlike traditional rule-based systems that need to be updated all the time. The system handles big datasets quickly, makes unambiguous predictions about fraud, and does away with the requirement for data normalisation. The suggested model is better than current approaches at finding bank fraud because it uses better detection techniques and better feature selection. It is also more dependable and can handle more data.

b) System Architecture:

The architecture of the fraud detection system consists of multiple components working together to identify fraudulent transactions efficiently. The system begins with data collection, where the Banksim dataset, containing simulated financial transactions, is acquired from Kaggle. This dataset includes multiple customers' payment records over different time periods and transaction amounts. Next, data preprocessing is performed to handle missing values, remove inconsistencies, and convert categorical features into a machine-learning-friendly format. Since Random

Forest can handle both categorical and continuous variables, the dataset is prepared without requiring normalization.

The fraud detection model is then built using the Random Forest Classifier, which consists of multiple decision trees working collectively to enhance prediction accuracy and reduce overfitting. The classifier analyzes transaction patterns, identifies suspicious activities, and categorizes transactions as fraudulent or legitimate. After model training, the evaluation phase compares accuracy, precision, recall, and F1-score to ensure optimal fraud detection performance. The final stage involves real-time transaction monitoring, where the trained model is deployed to classify new transactions dynamically. This system architecture ensures high accuracy, efficient processing, and adaptability to evolving fraud patterns in bank payments.

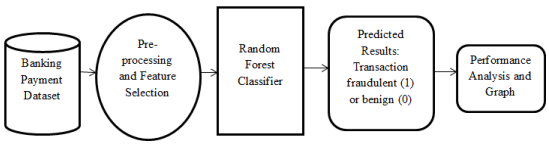

Fig 1 Proposed Architecture

c) Modules:

i) **Data Collection:** Module 1 is where the Data Collection Procedure is made. The first thing you need to do to develop a machine learning model is to gather data. This is an important step since our model will get stronger with more and better data.

You can get data via model folder datasets, human activities, or web scraping. The dataset comes from a well-known site called Kaggle. You may find the URL to the dataset below.

Dataset link:

https://www.kaggle.com/datasets/jayaprakashpondy/banksim-dataset

ii) **Data preparation**: It takes time to get data ready for training. Get rid of duplicates, rectify mistakes, deal with missing numbers, normalise, alter data types, and clean up when you need to. Randomising the data might get rid of the need to gather and prepare it in a certain sequence. Do exploratory analysis, including making data visualisations to find important links between variables or class imbalances (be careful of bias!). Sets of training and testing.

iii) **Model Selection**: Using the Random Forest Classifier from machine learning. We chose this strategy after finding that it worked 99.7% of the time on our training set. The Stochastic Random Forests Method Quickly learn how to use the algorithm. Imagine that you're planning a trip and want to travel to a unique place. What do you do to pick a good spot? You may start by looking at travel blogs and portals, exploring the web, and asking people you know. If you wanted to know how your friends' travels went, pretend for a second. Everyone you ask will have an idea. Make a list of all the places that were suggested. Ask them to vote on or choose the best vacation place from your list of suggested websites. You will select the place that got the most votes in the end.

d) Algorithms:

i) **Random Forest:** The RF Classifier is an ensemble machine learning technique that builds many decision trees to improve the accuracy of fraud detection. Bootstrap sampling produces random subsets of the dataset to train trees with random attributes. A majority vote chooses the most popular tree forecast as the categorisation. This method works well with both categorical and continuous data, eliminates overfitting, and increases generalisation. Random Forest is quite good at finding fake bank transactions since it can handle big datasets and automatically fix missing values.

4. EXPERIMENTAL RESULTS

We used the Banksim dataset to train and test the recommended system, which is an RF Classifier for fraud detection. After the model was trained on a part of the dataset, it was tested on transactions that were not seen before, after preprocessing and feature selection. The algorithm was 99% accurate in finding fake transactions, according to the results of training and testing. Key

performance indicators including recall, accuracy, and F1-score showed that the model was reliable. It was able to identify fraud with a high rate and very few false positives. The results show that the recommended method is a good alternative to traditional ways of uncovering fraud in real-time bank transfers.

Accuracy: The accuracy of a test is its ability to differentiate the patient and healthy cases correctly. To estimate the accuracy of a test, we should calculate the proportion of true positive and true negative in all evaluated cases. Mathematically, this can be stated as:

Accuracy = TP + TN TP + TN + FP + FN.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

Precision: Precision evaluates the fraction of correctly classified instances or samples among the ones classified as positives. Thus, the formula to calculate the precision is given by:

Precision = True positives/ (True positives + False positives) = TP/(TP + FP)

$$\text{Precision} = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

Recall: Recall is a metric in machine learning that measures the ability of a model to identify all relevant instances of a particular class. It is the ratio of correctly predicted positive observations to the total actual positives, providing insights into a model's completeness in capturing instances of a given class.

$$Recall = \frac{TP}{TP + FN}$$

F1-Score: F1 score is a machine learning evaluation metric that measures a model's accuracy. It combines the precision and recall scores of a model. The accuracy metric computes how many times a model made a correct prediction across the entire dataset.

$$\text{F1 Score} = \frac{2}{\left(\frac{1}{\text{Precision}} + \frac{1}{\text{Recall}}\right)}$$

$$\text{F1 Score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

MAP: MAP (Mean Average Precision) is a metric used to evaluate the performance of information retrieval systems. It measures the average precision across multiple queries or classes. Precision measures the accuracy of retrieved results, while Average Precision (AP) calculates the average precision for each query. MAP computes the average of AP scores across all queries or classes, providing a single measure of performance for the entire system.

$$MAP = \frac{1}{N} \sum_{i=1}^{N} AP_i$$

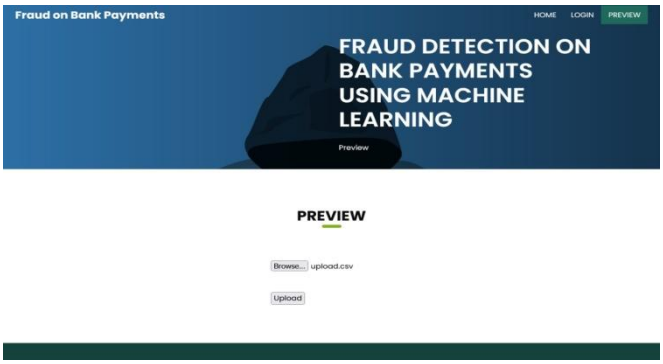

Fig.2 upload dataset

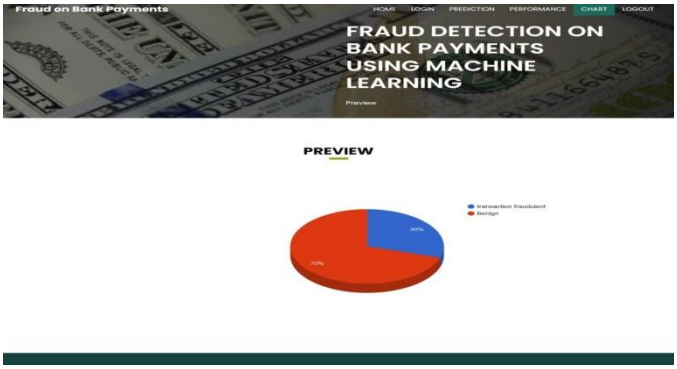

Fig.3. results page

5. CONCLUSION

The suggested fraud detection method based on machine learning uses the Random Forest Classifier to find fake bank transactions. The system gets 99% accuracy by using the Banksim dataset and improving feature selection. This is better than existing techniques. The ensemble learning method cuts down on overfitting, makes sure that processing happens in real time, and makes fraud detection more effective. The results show that machine learning may make financial transactions much safer by reducing false positives and correctly spotting fraud.

6. FUTURE SCOPE

Adding deep learning models like neural networks to boost fraud detection accuracy even further is one way that future improvements might be made. You may also use real-time streaming data analysis to find fraud right away. Adding support for transactions with more than one bank and using blockchain for secure data sharing might help stop fraud even more. Additionally, adaptive learning methods may be used to make sure the model changes as fraud trends do, which makes it more resistant to new threats.

REFERENCES

[1] R. Harrow, Is Your Credit Card Less Secure Than Ever Before? Forbes, 20-Apr-2018. [Online].

[2] Maes, S., Tuyls, K., Vanschoenwinkel, B. and Manderick, B., (2002). Credit card fraud detection using Bayesian and neural networks. Proceeding International NAISO Congress on Neuro Fuzzy Technologies.

[3] Singh, G., Gupta, R., Rastogi, A., Chandel, M. D. S., and Riyaz, A., (2012). A Machine Learning Approach for Detection of Fraud based on SVM, International Journal of Scientific Engineering and Technology, Volume No.1, Issue No.3, pp. 194-198, ISSN: 2277- 1581

[4] RamaKalyani, K. and UmaDevi, D., (2012). Fraud Detection of Credit Card Payment System by Genetic Algorithm, International Journal of Scientific & Engineering Research, Vol. 3, Issue 7, pp. 1 – 6, ISSN 2229-5518

[5] Patil, S., Somavanshi, H., Gaikwad, J., Deshmane, A., and Badgujar, R., (2015). Credit Card Fraud Detection Using Decision Tree Induction Algorithm, International Journal of Computer Science and Mobile Computing (IJCSMC), Vol.4, Issue 4, pp. 92-95, ISSN: 2320- 088X

[6] Ogwueleka, F. N., (2011). Data Mining Application in Credit Card Fraud Detection System, Journal of Engineering Science and Technology, Vol. 6, No. 3, pp. 311 – 322

[7] Bahnsen, A. C., Stojanovic, A., Aouada, D., & Ottersten, B. (2014). Improving credit card fraud detection with calibrated probabilities. In Proceedings of the 2014 SIAM International Conference on Data Mining (pp. 677-685).

[8] J. Steele and J. Gonzalez, Credit card fraud and ID theft statistics, CreditCards.com. [Online].