Credit Card Transactions Fraud Detection using Big Data and Machine Learning Techniques

Taksh Rastogi
Business Analytics and Big Data
Thapar University

Abstract — As the discipline of big data and data science develops throughout the world, it has become increasingly popular to predict the effect before the cause. The growth of data volume and the discovery of new uses for machine learning, which give researchers a means to identify vulnerability, are both results of science and technological advancement. Now, turning our attention to the financial sector, the number of fraud cases has sharply grown as a result of the growing number of clients and businesses that complete financial transactions using credit cards. In this paper, several models are tested on a big data set, using various combinations of parameters/features, and the best model is suggested as per the performance metrics of high R square and low RMSE.

***Index Terms* — Credit Card fraud, Fraud Transactions, Fraud Detection, Machine Learning, Big Data.**

I. INTRODUCTION

Today's generation relies heavily on the finance and banking industries, as almost every person interacts with banks either in person or online. Because of the banking information system, both the public and private sectors' productivity and profitability have greatly increased. Credit cards and online net banking are now used for the majority of transactions on e-commerce application systems. These systems are weak to increasingly sophisticated attacks and methods. Due to the importance of finance in our lives today, fraud detection in the banking industry is one of the key factors.

There is not a single company that is completely immune from fraud, whether it is in the conventional or developing industries. According to some research, frauds of all types might cost companies between 1% and 1.75 percent of their yearly revenue, or over $200 billion annually.

Credit card transaction fraud affects over 127 million people and results in about $8 billion in attempted fraudulent charges on credit and debit cards used by Americans each year, making it one of the most prevalent kinds of fraud. The data shows that 3,432 incidents of credit and debit card fraud were reported from all over India in 2021, an increase of about 20% from the previous year. Such scams surged by more than 70% in 2020. It demonstrated an almost two-fold increase in credit and debit card-related fraud in only two years. With 336 million cardholders and over 40 million shops accepting it, Visa is the largest major payment network. Mastercard has 231 million cardholders, Citibank has 95 million cardholders, and American Express has 63 million cardholders; followed by HDFC, SBI, and ICICI in India.

The features of a fraudulent transaction must be understood by credit card firms in order to create prediction algorithms that can identify potentially dangerous behaviour and prevent fraud. As the amount of data increases in terms of Peta Bytes (PB), they have integrated an analytical framework with Hadoop that can efficiently read data and give it to an analytical server for fraud prediction.

The recent rise in credit card fraud has been striking. It is actually one of the most pervasive threats to the BFSI sector. Credit card companies must be able to identify fraudulent credit card transactions in order to prevent customers from being charged for goods they did not buy. When a card is lost, skimmed, or accessed by fraudsters, the transactions that result have an unusual spending pattern and are therefore referred to as fraudulent transactions. However, these kinds of transactions are relatively uncommon and few in comparison to legitimate transactions that take place in large volumes. As a result, locating such fraudulent transactions is a challenging task that falls under the purview of fraud analytics. Identification of fraudulent transactions has historically been a fascinating research topic for the banking and financial industries, the research community, and academia due to the complexity of fraud analytics. The creation of a classifier that is proficient at detecting credit card fraud is the aim of this project. The project's dataset will be a collection of credit card transaction datasets that include both legitimate and fraudulent transactions. The project can make use of a variety of machine learning (ML) techniques, including Gradient Boosting Classifiers, Artificial Neural Networks, Decision Trees, and Logistic Regression. With the aid of these ML algorithms, the system will be able to differentiate between fake and real calls. This project enables us to perform practical classification using ML algorithms.

II. LITERATURE REVIEW

A variety of models are proposed for Credit Card Fraud Detection through different approaches and using different machine learning, AI and big data techniques.

Namrata et al. [1] claims that big data technologies like Spark, Kafka, and Zookeeper are utilised to store and handle huge amounts of user history transactions. The end result demonstrates the accuracy and efficiency of the Hidden Markov Model over a wide range of data. Patil et al. [2] proposes a strong framework that can handle a lot of data, with features that can be expanded to retrieve real-time data from several unreliable sources. A robust analytical model is subsequently constructed using the retrieved data. Three distinct analytical methodologies have been used to increase the analytical precision of fraud prediction. Decision Trees, Logistic Regression, and Random Forest Decision Trees. Kamaruddin et al. [3] employs Particle swarm optimization

(PSO) and auto-associative neural network (AANN) which are used in a hybrid architecture to solve the one-class classification (OCC) problem in the SAPRK cluster. Emmanuel et al. [4], states a Genetic Algorithm based feature selection method in conjunction with the Random Forest, Decision Tree, Artificial Neural Network, Naïve Bias, and Logistic Regression to improve the accuracy of the current credit card fraud detection model. Puninder et al. [5], employs data mining to recognise and catch credit card fraud. All algorithms for supervised and unsupervised learning are used. Machine learning techniques are used in conjunction with data mining to manage complicated and vast amounts of data. Preprocessing is used in data mining to eliminate extraneous information and retrieve relevant information from large datasets. You Dai et. al [6] discusses the Random Forest method and how it may be used to find fraud in this article. There are two types of random forests: CART-based random forests and random forests based on random trees. Second kind is superior to first type, according to this essay's summary. Sathyapriya et al. [7], proposes that the approaches that were previously being utilised in a different context, such as processing speed, latency, fault tolerance, performance, and scalability, have been examined to help the forensic investigator do big data forensic. Kulatilleke et al. [8], specifies a method and system for choosing classifiers for highly unbalanced sensitive, encoded data, which is characteristic in fraud detection jobs. We demonstrate that the best approach is to choose the best classifier based on the performance of all combinations of PCA feature selection, under sampling, and multiple classifiers, then construct a voting classifier to see if it can improve the score even more. Kulatilleke et al. [9], With an emphasis on the nature, difficulties, and consequences of obfuscation as a technique to preserve secrecy, this study offers descriptive and experimental insights into the features of credit card fraud data. Shakya et al. [10], devised a data-level strategy that combines tomek link removal, random under sampling, random oversampling, the Synthetic Minority Over-sampling Technique (SMOTE), and a hybrid approach combining SMOTE and tomek links removal. A logistic regression-based classifier was proposed by Alenzi et al. [11] using their techniques. The mean-based approach and the clustering-based method are the first two strategies used to clean the data. The suggested classifier performs better in terms of accuracy, sensitivity, and error rate when compared to the support vector machine classifier and the voting classifier, two well-known classifiers. According to Shao et al. [12], a model based on frequency domain features is integrated with a generative adversarial network (GAN) to improve minority class. This strategy employs the frequency domain amplitude properties of the data to produce different training data that reflects the trend of data changes, in contrast to the typical adversarial network, which only generates adversarial samples from the training data itself. Using numerous original datasets reduces over-fitting, according to experimental results, and classification performance is significantly enhanced. superior than other state-of-the-art strategies that are already in use. Naik et al. [13] proposed employing machine learning methods such as logistic regression, J48, Nave Bayes, and AdaBoost to eliminate fraudulent transactions. An online dataset is used to build and test the same set of algorithms. Through comparison, it is possible to determine that the Logistic Regression and AdaBoost algorithms perform better in terms of fraud detection. Arora et al. [14] used a variety of supervised machine learning methods on a dataset that was 70% training and 30% testing. Random forest, stacking classifier, XGB classifier, Decision tree, SVM, naive Bayes, and KNN algorithms are compared. In summary, SVM has the greatest FPR with 0.5360, and stacking classifier has the lowest FPR with 0.0335. Malini et al. [15] explored credit card fraud kinds, precautionary measures to avoid credit card fraud, and big data tools to identify card theft. However, the use of big data for this purpose is still in its early stages, thus a significant amount of effort is required to integrate big data techniques into real-time fraud detection. Because it can operate with massive and real-time transaction data sets, it reduces risk and reaction time to milliseconds. Benchaji et al. [16] created a model by combining the strengths of three sub-methods: uniform manifold approximation and projection (UMAP) networks for selecting the most useful predictive features, Long Short Term Memory (LSTM) networks for incorporating transaction sequences, and the attention mechanism for improving LSTM performance. Sulaiman et al. [17] developed a hybrid method in which neural networks (ANN) are used in a federated learning framework. It has been reported to be an excellent method for improving CCFD accuracy while maintaining anonymity. To prevent the problem of data imbalance, we require huge datasets to train the model. The utilisation of real-time information can offer us with a wider range of data, yet privacy remains a concern. We may use real-time datasets to train the model while maintaining anonymity, according to the suggested strategy. As a result, a Federated learning architecture with ANN can improve the ML model's capacity to detect fraudulent transactions. Kavipriya et al. [18] contrast data mining strategies including Simple K-means, Hidden Markov Model, Bayesian Network, KNN algorithm, and Outlier detection. They investigate various data mining approaches in order to identify and forecast credit card fraud. Different researchers' analyses demonstrate that different data mining strategies work. In addition to these strategies, the "Hidden Markov Model" optimises the optimum option for fraud detection. Acosta et al. [19] discuss a Balanced Random Forest, which may be employed in both supervised and semi-supervised circumstances using a co-training strategy. In order to solve the class imbalance problem, two distinct strategies for the co-training approach are evaluated. Furthermore, a Spark platform and the Hadoop file system complement our approach, allowing it to be scalable. In terms of geometric mean, the suggested technique provides an absolute improvement of roughly 24% when compared to a normal random forest learning strategy. Han et al. [20] advocated building and fitting several models, such as logistic regression and decision trees, using the training data. Finally, we discovered that the random forest model had the highest fraud detection rate, hitting 54% in an out-of-time test. The model developed can be used in anti-fraud monitoring systems, or a similar model creation approach can be used in related business domains to detect and minimise the recurrence of such behaviours.

Table 1: Summary of related work

References	Technique Used	Dataset Used	Details	Performance measure used
Namrata et al. [1]	Hidden Markov model (HMM)	Synthetic dataset multiple transaction details taken from different sources	A credit card fraud detection workflow is proposed which can fuse different detection models to improve accuracy, this workflow works with HMM model	-
Patil et al. [2]	Decision Trees, Logistic Regression, and Random Forest Decision Trees.	German Credit Card Fraud Dataset	These analytical models are applied to the credit card dataset, and the confusion matrix is used to assess the model's accuracy.	Confusion Matrix
Kamaruddin et al. [3]	Particle swarm optimization (PSO), auto-associative neural network (AANN)	ccFraud dataset	The suggested parallel method parallelizes datasets in a distributed clustered computation using Apache Spark. The AANN method has also been implemented in parallel.	MSE and Mean MSE convergence plot
Emmanuel et al. [4]	Random Forest, Decision Tree, Artificial Neural Network, Naïve Bias, and Logistic Regression	European Credit Card Fraud Dataset	The Random Forest was used to implement the Genetic Algorithm's (GA) fitness function. Five ideal feature vectors were produced when the GA was further applied to the dataset of credit card transactions made by European cardholders.	Accuracy, Recall, Precision and F1 Score
Puninder et al. [5]	Logistic and linear regression and hierarchal clustering	Multiple weak entity dataset	Data mining techniques might be used to find solutions to the complex issues facing the banking industry. It publishes precise findings that assist the cardholder in locating instances of fraud.	Accuracy, Recall, Precision and F1 Score
You Dai et. al [6]	Genetic Programming and Fuzzy control system, Hadoop and spark.	Synthetic dataset	Created a hybrid architecture to address the issue of the surge in trading activity that occurs every day. Our system uses a four-layer design to manage data storage, model training, data sharing, and online detection with the goal of merging several detection methods to increase accuracy.	Throughput rate
Sathyapriya et al. [7]	Apache Hadoop, Apache FLINK, Apache SPARK and MapReduce	Dataset from Digital Forensic by Alessandro Guarino.	Spark is recommended as the best performing approach among the four after all four have been compared based on variables including processing speed, latency, fault tolerance, performance, and scalability.	Latency, Fault Tolerance, Performance and scalability
Kulatilleke et al. [8]	25 Machine Learning Models	IEEE Symposium dataset	Choosing a classifier based on data is basically what this is. While employing the entire enormously imbalanced real data distribution, the model that was developed with the technique surpasses even powerful generative models,	G-Mean Score and F1 Score

			proving the usefulness of the strategy.	
Kulatilleke et al. [9]	15 Machine Learning Classifiers	Pise, N.N. Kulkarni Dataset	Tested 15 machine learning classifiers' algorithmic performance on sensitive transaction data that had been PCA encoded. The results demonstrate that, while PCA does not significantly worsen performance, care should be taken to use the right principle component size (dimensions) to prevent overfitting.	G-Mean and F1-Score
Shakya et al. [10]	Logistic Regression . Random Forest and XGBoost	Dataset by ULB (Universit Libre de Bruxelles)	Utilized algorithmic strategies like bagging and boosting to address the issue of class imbalance. As a bagging approach for this, we chose the random forest model, and as a boosting method, we chose XGBoost. In addition to these models, we selected the logistic regression model to contrast with others.	Precision, Recall, F1 Score, PR and ROC
Alenzi et al. [11]	Logistic Regression Classifier	Kaggle Machine learning group – ULB Credit Card Fraud Detection Dataset	The proposed system uses logistic regression to build the classifier to prevent frauds in credit card transactions.	Accuracy, Sensitivity, and Error Rate
Shao et al. [12]	XGBoost, SVM, and LOR	Real World Credit (RWC), Kaggle Credit card (KC) and UCI Statlog German Credit (USGC)	To achieve high quality data augmentation , they suggest a novel imbalanced data augmentation technique based on frequency domain transformation and generative modelling.	AUC, F1-score and Fit Ratio
Naik et al. [13]	Naïve Bayes, Logistic regression , J48 and AdaBoost	Online Sample Credit card Dataset	Implemented 4 models and calculated their accuracy and time for training. Concluded that Logistic regression and AdaBoost algorithms perform better	Accuracy and Time duration.
Arora et al. [14]	16 models tested	--	This method is used to rank fraud detection models in order to discover the best model from among the available fraud detection models. Various model selection criteria, as well as a collection of FDMs, are required for rating the FDMs. The Coefficient sum approach is shown using real-world data sets.	Variance, RSS, Bias, MSE, MAE, PRR, RMSPE, ,TS and Accuracy
Malini et al. [15]	Hadoop, Spark, Regression	Data mined	Detected many forms of credit card theft involving actual or virtual cards	--
Benchaji et al. [16]	LSTM	Kaggle credit card dataset	The goal was to increase prediction efficiency during the detection of fraudulent transactions by integrating the strengths of several Machine Learning approaches.	Accuracy, recall and precision
Sulaiman et al. [17]	Federated Learning and ANN	Real-time transaction datasets	A hybrid strategy using ANN for	--

			effective detection and federated learning to create a framework for data privacy will give a fresh contribution.	
Kavipriya et al. [18]	Hidden Markov Model, Neural Network, Bayesian Network, Genetic Algorith, K- nearest neighbor algorithm, Support Vector Machine, Decision Tree, Fuzzy Logic Based System	--	They look at various data mining techniques for detecting and forecasting credit card fraud. Various academics' findings show that various data mining methodologies succeed.	--
Acosta et al.[19]	BRF	Colombian payment gateway company dataset	The concept is built on a Balanced Random Forest, which may be employed in supervised and semi-supervised settings using a co-training approach.	Sensivity, Specificity, AUC, G-mean and Weighted-Accuracy
Han et al. [20]	9 models	Credit card applications dataset	On the training set, many models, such as logistic regression and decision trees, are created and fit. Finally, we discovered that the random forest model had the highest fraud detection rate, hitting 54% in an out-of-time test.	OOT score

III. PROBLEM FORMULATIONS AND RESEARCH METHODOLOGY

3.1 Problem Formulation

In the last few years, the technology industry has changed quite a lot and people and companies are shifting from the physical place to the online websites and ecommerce in each sector like BFSI, healthcare, education, retail, media, hospitality etc. As companies are becoming digitalized, a lot of the customer information is collected by the company through various sources like transactions, registration forms and login credits. The COVID-19 pandemic has become the catalyst for the digitalization of each sector and increased the rate of company becoming digitalized. In the Banking industry, as the data security has become the major concern for the company about their customers, many new and established companies in this industry are increasing at very fast pace and are coming with new innovative products.

But there are many problems that the company is facing. As a company exists to serve the customers best, they are required to secure their transactions and deal with/defend against any fraud act that may happen. The company is needed to analyze information and conduct analytics and strategic research based on the data and changing trends/ practices in fraud. In accordance with the project timeline, they offer real-time detection solutions to the complex problems at hand. The problem is fraud detection in daily transactions of credit card users' purchases at various POS. They need to analyse them further and validate with the customer to minimize the damage and do needful procedures to safeguard the card and customer details as soon as possible.

3.1.1 Trend and Sentiment Analysis

We want to collect information and spot the trend of people's concern in credit card fraud in various regions and platforms like Twitter and Google searches to further know about the problem and its occurrences.

3.1.1.1 Twitter Analysis

A set of tweets were scrapped from twitter on various posts matching keywords like credit card, credit card fraud, fraud transactions etc. The scrapped tweets/posts were further cleaned and analysed to produce WordCloud and Sentiment analysis using the R program and JMP tool.

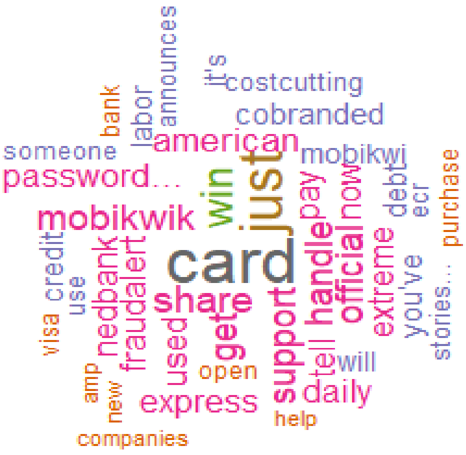

Figure 1: Word Cloud of Tweets

The top keywords from Word Cloud generated are Card, Share, Fraudalert, Support, Password, Pay, Daily, Used and Handle, which clearly express the concern of the posts that the customers are sharing their fraud experiences and demand support for the same.

The Sentiment Analysis conveys their mood and feelings as Negative (towards the fraudsters) – 1000 count, Trust (in the company for support) – 2500 count, Anger (on the act) – 490 count, and Sadness (with the act) – 500 counts.

3.1.1.2 Google Trend Analysis

Google Trends show a fluctuation and some seasonality in the searches for ‘credit card fraud’ for the past 12 months. For the worldwide, it is found that the searches are consistent with most searches from Singapore, United States, Japan, New Zealand, Canada and so on. People are also searching for ‘credit card scamming methods’, ‘how to spot a credit card skimmer’, ‘phishing’, and ‘atm skimming’.

For India region, it indicates that the frauds happen mostly on festivals times. These searches were mostly conducted in Telangana, Andhra Pradesh, Karnataka, Tamil Nadu, and Chandigarh. A recent fraud was in news in the month of April which showed spike in searches for the same.

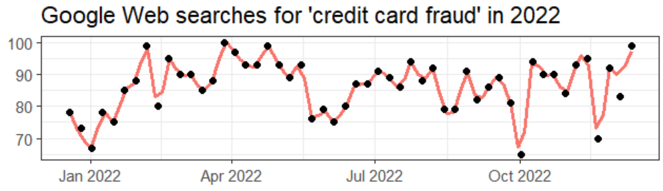

Figure 2: Google Trend in India
(Plot using R)

3.2 Gap Analysis

From the literature review, we have knowledge and clarity of various fraud detection techniques, big data techniques and machine learning algorithms to detect them. But credit card fraud detection using big data and tools is a field that is not explored much. Also, the trend and sentiment analysis from Twitter and Google Trends clearly provide evidence of increasing fraudulent activities and concerns in the market and among consumers.

Hence, we would like to utilize big data and the available predictive tools and technology to find the best model for fraud detection.

3.3 Objectives

Detecting credit card fraud is a significant issue for all banking companies. Banks must be able to recognise and stop credit card fraud.

Therefore, the objective of this project is:

- To create a machine learning model that identifies fraudulent transactions using historical transactional data from a pool of merchants' customers.
- We'll also give stakeholders a cost-benefit analysis of the model and the appropriate recommendations they should follow to reduce the risk of fraud.

This project is divided into three parts which cover all the problems that companies and customers are facing associated with the project. The first part is the Data collection and Pre-processing which will help the company to know about their actual figures of specific researched company or person. Using tools: Kaggle, Excel, SPSS and Data type: Structured.

The second part of the project i.e., Data manipulation and modelling will help to deal with the expansion and finding of the potential fraud and will help them to improve their security and fraud detection analysis. Using tools: Excel, JMP Pro.

The third part of the project i.e., Analysis and Interpretation, which will help to understand and deliver the right results.

3.4 Research Methodology

In the current context, it has been noticed that knee fraud detection for companies and consumers is a difficult and complex job, and their dominance in the banking space is expanding rapidly. The database of various credit card transactions is taken into account in this job.

3.4.1 Procedures and Methods

The methodology began with searching the dataset or gathering data for the identified problem. This step consisted of three steps: identifying various data sources, such as the Kaggle, UCI, and UCSD repositories; collecting data; and finally, integrating the data obtained from various sources. Afterwards, data preprocessing is carried out to transform the raw data into a clean data set and carry out operations to fix the problems of size, duplication or redundancy, invalid data, and noise in the data set by first importing all the necessary libraries from loading the dataset to performing manipulation, and also performing feature scaling when and where needed to carry out the Explanation. In EDA, the analysis of the dataset is done, and then plots and others, are used.

The steps to design the proposed system are as follows

1. Importing Data
2. Preprocessing
3. Choose the Model(s)
4. Run for prediction
5. Analyze Predictions or Detection
6. End

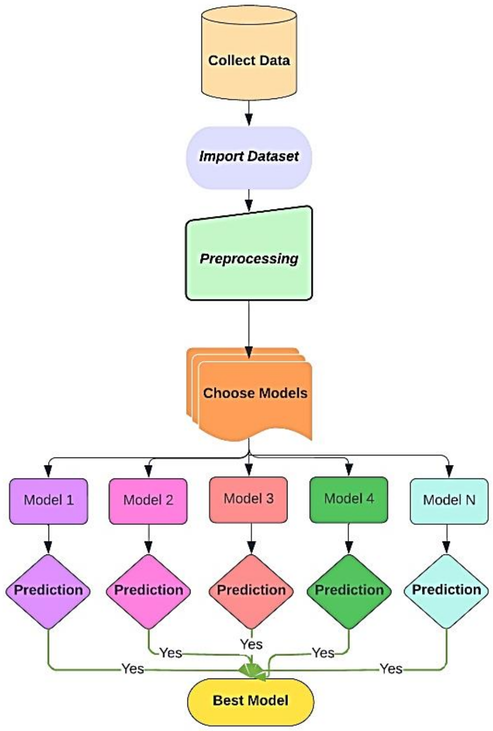

Figure 3: Proposed System Framework

3.4.2 Proposed model and implementation

The proposed framework is implemented on Credit Card Transactions Fraud Detection Dataset from Kaggle. The dataset has independent and dependent variable, so supervised learning techniques had to be used, in which the machine learns with supervision or when the learning involves training with labeled data and letting the model act on that information. In order for the model to find information and patterns, it must be allowed to operate independently. Here are the models used for the analysis of the data and to make a detection system:

i. EDA and Data Preprocessing

The process of Exploratory Data Analysis and preprocessing can be initially done using R/Python/Excel tools like PowerQuery and Pivots. Check for correlation and collinearity between the variables to reduce the features and increase the accuracy of models. Perform encoding on categorical variables and remove the features that do not contribute to the model. Then, standardize the columns to make better predictions using the Z-Score method i.e. $Z = (X - \mu)/\sigma$.

ii. Model Selection

The model selection process is to choose the best available models and their knowledge and support to test your hypothesis and decide which one performs better than the other. We will use JMP to Fit the Models and analyse which features of the data are contributing well to detecting fraud. We will be testing our data with the following models:

a. Boosted Tree

Strong decision trees are created by boosting weak decision trees, often known as weak learners. The methods employ a collection (ensemble) of decision trees for bagging and boosting. Ensemble meta-algorithms are what are used for boosting and bagging. Iterative processes are used in boosting. Every tree is reliant on the one before it. As a result, it is challenging to parallelize the training of boosting algorithms. The trees that have undergone the boosting procedure are known as boosted trees.

b. Bootstrap Forest

A method called Bootstrap Forest produces numerous decision trees and, in essence, averages them to produce a final predicted value. From a unique random sample with replacement, each tree is built. Additionally, the technique only allows for a random split.

c. Generalized Regression

The generalised linear model (GLM), which generalises linear regression, allows the linear model to be connected to the response variable via a link function and allows the variance of each measurement to be a function of the projected value. Other statistical models like Poisson regression, logistic regression, and linear regression are all brought together by this one.

d. Neural Boosted

Boosting is a technique used in machine learning to cut down on mistakes in the processing of predicted data. Data scientists use labelled data to train machine learning software, sometimes known as machine learning models, to infer information from unlabeled data. Depending on how accurately the training dataset was generated, a single machine learning model may produce incorrect predictions. Boosting makes an effort to address the problem by progressively training many models to increase the overall system's accuracy.

e. K Nearest Neighbors

The k-nearest neighbours algorithm, sometimes referred to as KNN or k-NN, is a supervised learning classifier that employs proximity to produce classifications or predictions about the grouping of a single data point. Although it may be applied to classification or regression issues, it is commonly employed as a classification method since it relies on the idea that comparable points can be discovered close to one another.

iii. Metrics

The quality of a statistical or machine learning model is measured using evaluation metrics. Any project requires the evaluation of machine learning models or algorithms. There are several assessment measures available for testing a model. R/AdjR square, p-value/significance, T-test, VIF, Accuracy, confusion matrix, and others are examples.

IV. EXPERIMENTAL RESULTS AND ANALYSIS

The experimentation was conducted on a Windows 11 OS with 24 GB RAM, 4GB NVIDIA RTX 3050 GPU and a 512 GB SSD disk. Used tools include Excel and JMP. In order to assess the effectiveness of the suggested methodology, we worked on the dataset: Credit Card Transactions Fraud Detection Dataset from Kaggle. This section describes the dataset and details the analysis process and findings.

A. Datasets Description

In order to test the proposed framework, a standard benchmark has been used. The details of this dataset are given below

i. Credit Card Transactions Fraud Detection Data Set

This dataset of simulated credit card transactions includes both valid and fraudulent purchases made between January 1 and December 31, 2019. It includes transactions made with a pool of 800 businesses using the credit cards of 1000 consumers. The Credit Card Transactions Fraud Detection Dataset, which consists of a training dataset (12.5 lac rows X 22 columns) and a testing dataset (5.5 lac rows X 22 columns), is the subject of this project's analysis.

Data Dictionary: cc num - Credit card number, merchant - merchant name, category - transaction category, trans date trans time - Transaction time stamp, trans amt - Transaction amount, first - Cardholder's first name, last - Cardholder's last name, gender - Street address of cardholder's sex, city of the transaction, state of the transaction, zip code of the transaction, and longitude of the transaction lengthy transaction, latitude, Population of the city, job - Job of Cardholder, dob - Date of Cardholder's Birth, trans num - Transaction Number, unix time - Time in Unix Format, is fraud - type of transaction (fraud or not fraud). Our target variable in this case is the 'is fraud' variable.

Table 2: State-wise Fraud Transactions in the dataset

State	Fraudulent Transactions
AK	50
AL	278
AR	195
AZ	64
CA	402

B. Performance Evaluation

There are various methods for assessing machine learning models. We have used R Squared and Root Average/Mean Square Error (RASE/RMSE) to evaluate the fraud detection module. The performance metrics are given in the following equations (7,8):

$$R^2 = 1 - \frac{RSS}{TSS} \qquad (7)$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N} (Predicted_i - Actual_i)^2}{N}}$$

(8)

Where R^2 = coefficient of determination, RSS = sum of squares of residuals, TSS = total sum of squares, N = number of observations.

Table 3: Comparison of performance metrics

Method	R Square	RMSE
Boosted Tree	0.91	0.036
Neural Boosted	0.90	0.045
Bootstrap Forest	0.85	0.046
Generalized Regression	0.77	0.060
K Nearest Neighbors	0.77	0.060
Naive Bias	0.74	0.062

C. Discussion

Data preprocessing has been done before, and the model is deployed on the dataset so that it doesn't face any issues while running the model. In order to identify the features that could be associated with fraudulent behaviours, we first did an exploratory data analysis on the data, including a correlation and collinearity check between the variables (accepted below or equal to 5). Most transactions are made after noon, and during the Christmas season, both regular and fraud ones rise. Older adults over 75 are especially vulnerable to scams. This is because scammers could try to take advantage of their ignorance of the continuously evolving processes used to conduct transactions. Gas, grocery, house, shopping, and pets are the top 5 categories with the highest fraud rates in the categories. For model testing, the variables "lat," "long," "zip," "city pop," "unix time," "merch lat," and "merch long" have been assumed to be insignificant. They were therefore encoded or removed from the dataset together with the original variables. After that, we standardize the dataset (18 lac rows X 67 columns) and run models using those features and evaluate how well they predict fraud with the output metrics where we wanted to look for the one with high R Square and low RMSE/RASE.

V. CONCLUSION AND FUTURE WORK

A number of algorithm-based strategies were used and built into the methodology. The Boosted Tree approach outperforms the other in terms of higher R Square and lowest RMSE. Finally, the future focus of this study will be on increasing the functionality and efficacy of the present models utilised by several banking institutions and researchers. For data of this complexity, it is advised to test a sample with multiple models before deciding to perform the entire thing with the one that produces the best results. The accuracy of the current model can also be improved by increasing the data size and fitting the complete model which would require a much more sophisticated hardware and performance system. In addition to this, a big data ecosystem (on Hadoop/Spark) can help in fast processing, application and analysis at a organization level.

VI. REFERENCES

[1] Namrata Pandey, Rajeshwari S,Shobha Rani BN, Mounica B,"CREDIT CARD FRAUD DETECTION USING BIG DATA FRAMEWORK", 2018 IJCRT | Volume 6, Issue 2 April 2018 | ISSN: 2320-2882

[2] Patil, S., Nemade, V., & Soni, P. K. (2018). Predictive Modelling For Credit Card Fraud Detection Using Data Analytics. Procedia Computer Science, 132, 385-395. https://doi.org/10.1016/j.procs.2018.05.199

[3] Sk. Kamaruddin Vadlamani Ravi, " Credit Card Fraud Detection using Big Data Analytics: Use of PSOAANN based One-Class Classification"

[4] Emmanuel Ileberi, Yanxia Sun and Zenghui Wang, "A machine learning based credit card fraud detection using the GA algorithm for feature selection", Ileberi et al. Journal of Big Data (2023) 9:24

[5] P. Kaur, A. Sharma, J. K. Chahal, T. Sharma and V. K. Sharma, "Analysis on Credit Card Fraud Detection and Prevention using Data Mining and Machine Learning Techniques," 2021 International Conference on Computational Intelligence and Computing Applications (ICCICA), 2021, pp. 1-4, doi: 10.1109/ICCICA52458.2021.96`97172

[6] You Dai, Jin Yan, Xiaoxin Tang, Han Zhao and Minyi Guo, "Online Credit CardFraud Detection: A Hybrid Framework with Big Data Technologies", IEEE TrustCom/BigDataSE/ISPA , pp 1644 -1651, 2016

[7] Sathyapriya, M., & Thiagarasu, D.V. (2017). Big Data Analytics Techniques for Credit Card Fraud Detection: A Review.

[8] Kulatilleke, G.K. (2023). Credit card fraud detection - Classifier selection strategy. ArXiv, abs/2208.11900.

[9] Kulatilleke, G.K. (2023). Challenges and Complexities in Machine Learning based Credit Card Fraud Detection. ArXiv, abs/2208.10943.

[10] Ronish Shakya, "Application of Machine Learning Techniques in Cr techniques in Credit Car edit Card Fraud Detection", http://dx.doi.org/10.34917/14279175

[11] Alenzi, H.Z., & O, N. (2020). Fraud Detection in Credit Cards using Logistic Regression. International Journal of Advanced Computer Science and Applications.

[12] M. Shao, N. Gu and X. Zhang, "Credit Card Transactions Data Adversarial Augmentation in the Frequency Domain," 2020 5th IEEE International Conference on Big Data Analytics (ICBDA), 2020, pp. 238-245, doi: 10.1109/ICBDA49040.2020.9101344.

[13] Heta Naik and Prashasti Kanikar. Credit card Fraud Detection based on Machine Learning Algorithms. International Journal of Computer Applications 182(44):8-12, March 2019.

[14] Suman Arora , "Selection of Optimal Credit Card Fraud Detection Models Using a Coefficient Sum Approach" , International Conference on Computing, Communication and Automation (ICCCA2017), pp 482 - 487, 2017

[15] N.Malini, Dr.M.Pushpa, Analysis on credit card fraud detection techniques by data mining and big data approach, pp. 38-45 ISSN 2320-7345 vol 5.

[16] Benchaji, I., Douzi, S., El Ouahidi, B. et al. Enhanced credit card fraud detection based on attention mechanism and LSTM deep model. J Big Data 8, 151 (2021). https://doi.org/10.1186/s40537-021-00541-8

[17] Bin Sulaiman, R., Schetinin, V. & Sant, P. Review of Machine Learning Approach on Credit Card Fraud Detection. Hum-Cent Intell Syst 2, 55–68 (2022). https://doi.org/10.1007/s44230-022-00004-0

[18] Kavipriya, T & Natarajan, Geetha. (2017). STUDY ON CREDIT CARD FRAUD DETECTION USING DATA MINING TECHNIQUES. ISSN: 2395-5325 3. https://www.researchgate.net/publication/344489468_STUDY_ON_CREDIT_CARD_FRAUD_DETECTION_USING_DATA_MINING_TECHNIQUES

[19] G. E. Melo-Acosta, F. Duitama-Muñoz and J. D. Arias-Londoño, "Fraud detection in big data using supervised and semi-supervised learning techniques," 2017 IEEE Colombian Conference on Communications and Computing (COLCOM), 2017, pp. 1-6, doi: 10.1109/ColComCon.2017.8088206.

[20] Yaodong Han et al Detection and Analysis of Credit Card Application Fraud Using Machine Learning Algorithms 2020 J. Phys.: Conf. Ser. 1693 012064. doi:10.1088/1742-6596/1693/1/012064