

A Study on Credit card Fraud in Online Transactions

A PROJECT REPORT

Submitted by

MOHANAPRIYA E (2116210701164)

in partial fulfillment for the award of

the degree of

BACHELOR OF ENGINEERING

in

COMPUTER SCIENCE AND ENGINEERING



RAJALAKSHMI ENGINEERING

COLLEGE ANNA UNIVERSITY,

CHENNAI

MAY 2024

**RAJALAKSHMI ENGINEERING COLLEGE,
CHENNAI**

BONAFIDE CERTIFICATE

Certified that this Thesis titled “**A Study on Credit card Fraud in Online Transactions**” is the bonafide work of “**MOHANAPRIYA E (2116210701164)**” who carried out the work under my supervision. Certified further that to the best of my knowledge the work reported herein does not form part of any other thesis or dissertation on the basis of which a degree or award was conferred on an earlier occasion on this or any other candidate.

SIGNATURE

Dr . S Senthil Pandi M.E.,Ph.D.,

PROJECT COORDINATOR

Professor

Department of Computer Science and Engineering

Rajalakshmi Engineering College

Chennai - 602 105

Submitted to Project Viva-Voce Examination held on_____

Internal Examiner

External Examiner

ABSTRACT:

Emerging techniques rooted in data science and machine learning have emerged to combat the rising tide of credit card fraud. To ascertain the most effective model for detecting fraudulent transactions, this investigation scrutinizes three distinct approaches: decision trees, random forests, and logistic regression. Notably, the random forest model exhibits exceptional performance, boasting a 99.9% area under the curve and a leading accuracy rate of 99%. Consequently, we recommend the adoption of the random forest algorithm for credit card fraud detection. Furthermore, our analysis unveils a prevalent targeting of individuals aged 60 and above by these fraudulent schemes.

INTRODUCTION:

Before 1996, banking services were only available in person. But everything changed when Citibank and Wells Fargo Bank introduced the first internet banking application in the United States. This marked the beginning of a major shift towards online financial transactions, including using credit cards for internet purchases. Over the past decade, there has been a significant increase in online activities such as shopping, working remotely, banking online, and socializing. Unfortunately, along with this increase, there has been a rise in fraudulent activities targeting online transactions and payment systems.

The Recent advancements in digital technology have transformed how people manage their finances, especially in day-to-day transactions. Many payment systems have moved from physical to digital platforms to improve efficiency and stay competitive. This has made internet banking and other online transactions more convenient for customers, with credit cards playing a central role.

Credit cards, often seen as plastic cards with personal information, are issued by banks to allow customers to make purchases worldwide. However, using someone else's credit card without permission to get money or goods, whether physically or digitally is considered credit card fraud. These fraudulent incidents often lead to significant financial losses. The digital transaction environment has made it easier for fraud to occur, as transactions can be completed using just the card information without needing the physical card.

The Bank of Ghana reported a substantial increase in credit card fraud losses, from GH 1.26 million (\$250,000) in 2019 to GH 8.20 million (\$1.46 million) in 2020. Fraudulent activities have been on the rise across all payment channels, with digital transactions experiencing the most significant surge. Perpetrators of such fraud often use tactics like VPN connections through Anchor Free software or engage in physical robberies, making it difficult for authorities to apprehend them.

2. LITERATURE REVIEW:

A regular distribution or relationships among explanatory variables aren't important for the binary final results variable prediction method referred to as logistic regression. While the explanatory variables in logistic regression fashions may be both numerical or categorical, the final results variable in those fashions is qualitative. Logistic regression has been extensively utilized by lecturers to come across economic bankruptcy.

The Decision trees, a non-linear classification technique, segment a sample into progressively smaller subgroups based on various explanatory variables. At each branch, the algorithm selects the explanatory variable with the strongest correlation to the outcome variable according to a predetermined criterion. Being nonparametric, decision trees accommodate both quantitative and qualitative data structures. However, when applied to the entire dataset, decision trees tend to overfit, potentially yielding poor results. They find application in filtering spam emails and predicting susceptibility to certain viruses in medical contexts.

The Random forests, introduced by, add an extra layer of randomness to bagging. Unlike conventional trees, random forests use different bootstrap samples for each tree's construction and randomly select subsets of variables at each node for splitting. The average prediction of all trees constitutes the model's output [19]. Despite their capability to measure the significance of each feature, random forests exhibit bias towards attributes with numerous levels, including qualitative variables. They find application in diverse domains such as bioinformatics, video segmentation, and image classification. Identifies various categories of credit card fraud, including bankruptcy fraud, counterfeit fraud, application fraud, and behavioural fraud. Depending on the type of fraud, different preventive measures can be devised. Machine learning methods like logistic regression, naive Bayes, random forest, k-nearest neighbours, gradient boosting, support vector machines, and neural networks have been employed for identifying fraudulent transactions in various jurisdictions. Feature importance approaches are often used to select the most relevant features for the model developed a machine learning-based technique for fraud detection using hybrid models with AdaBoost and majority voting strategies. They introduced noise to enhance the robustness of their models and achieved a high accuracy score. Other studies have explored the use of random forests to identify credit card fraud, although challenges like imbalanced data hinder their effectiveness investigated practical methods for detecting credit card fraud in financial institutions. They employed various machine learning algorithms and found that random forest, Boost, and decision tree performed best, with high AUC values . By identifying and categorizing fraudulent transactions, machine learning algorithms may be able to stop financial losses. Credit card fraud detection involves modelling past transactions, distinguishing between genuine and fraudulent transactions in real-time .

3. Data and methods:

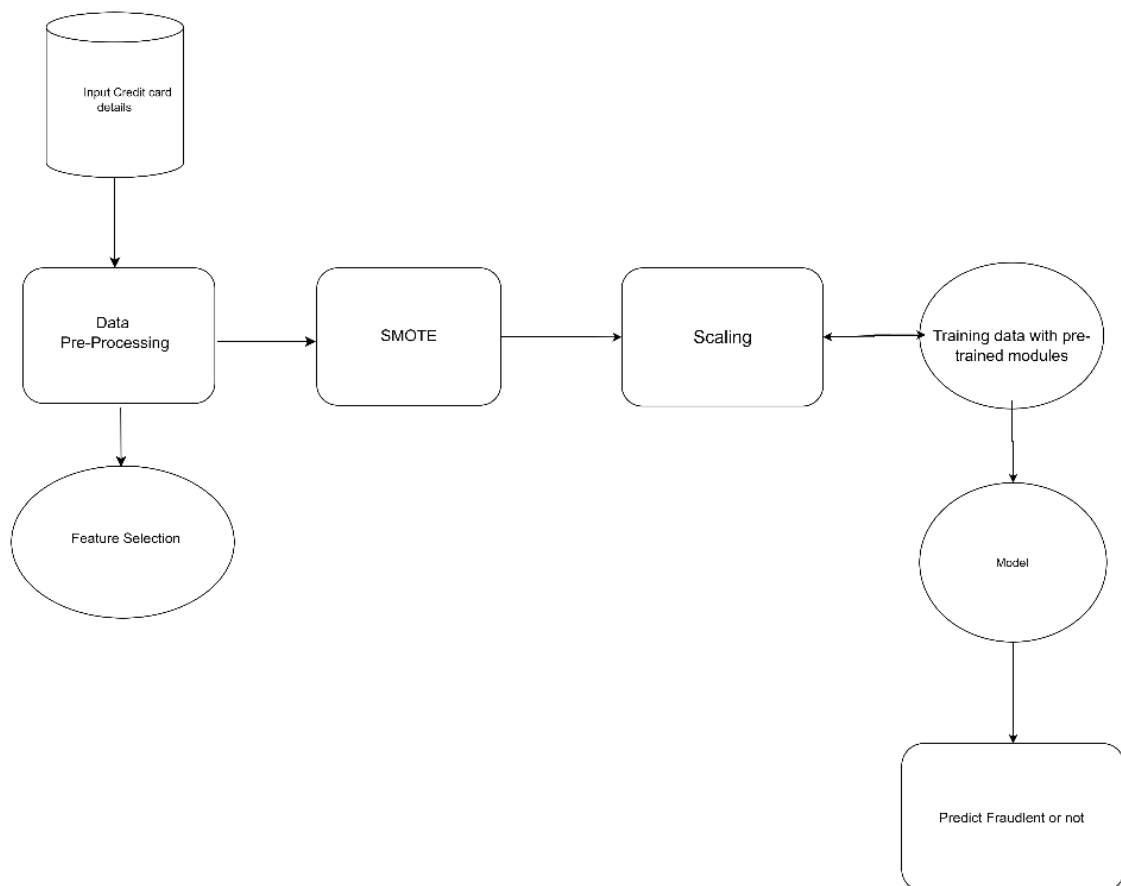
3.1 Data:

The research utilized a dataset containing simulated credit card transactions spanning on September of 2013, originating from the western region of the United States. These legitimate and fraudulent transactions were generated made by the transactions by European cardholders over data generation methodology, miming purchases made by 1000 customers across 800 businesses. Each entry in the dataset includes details such as the customer's

identity, the merchant involved, the nature of the purchase, and a flag indicating whether the transaction is fraudulent. With a total of 555,719 observations and 23 variables, including 12 qualitative ones, this dataset offers a comprehensive foundation for analysis.

During the pre-processing phase, rigorous cleaning procedures were applied to eliminate missing values, ensuring the integrity of the dataset for subsequent analysis. To standardize numeric features and mitigate the impact of varying scales, feature scaling techniques were employed, transforming all numerical variables into a uniform range between 0 and 1. Moreover, to address the imbalance in class distribution, undersampling was implemented, removing instances where transaction frequencies fell below five or exceeded 1250. While the dataset exhibits significant skewness, rather than employing the Synthetic Minority Oversampling Technique (SMOTE) as suggested by previous studies, this research opted for under-sampling to rectify class imbalances. Specifically, under-sampling was conducted in the majority class to align its frequency with or slightly surpass that of the minority class, as depicted.


Architecture Diagram :





3.2 Methods:

In order to categorise fraudulent transactions, we check out supervised device gaining knowledge of on this section, searching at famous fashions which include logistic regression, Random Forest, and Decision Tree.

Summary of numerical variables' basic statistics:



| | count | mean | std | min | 25% | 50% | 75% | max |
|------|--------|------------|------------|------------|------------|------------|-------------|-------------|
| Time | 1986.0 | 761.035750 | 451.034025 | 0.000000 | 366.000000 | 750.000000 | 1161.000000 | 1526.000000 |
| V1 | 1986.0 | -0.284195 | 1.353508 | -11.140706 | -1.045512 | -0.437621 | 1.095047 | 1.685314 |
| V2 | 1986.0 | 0.266886 | 1.142026 | -12.114213 | -0.204111 | 0.314294 | 0.926126 | 6.118940 |
| V3 | 1986.0 | 0.848005 | 1.012645 | -12.389545 | 0.280517 | 0.864505 | 1.486942 | 4.017561 |
| V4 | 1986.0 | 0.151216 | 1.264932 | -4.657545 | -0.670513 | 0.190698 | 1.002546 | 6.013346 |
| V5 | 1986.0 | -0.077457 | 1.272512 | -32.092129 | -0.576269 | -0.154843 | 0.376901 | 7.672544 |
| V6 | 1986.0 | 0.050205 | 1.274204 | -3.498447 | -0.691393 | -0.198063 | 0.389714 | 21.393069 |
| V7 | 1986.0 | 0.138347 | 1.140750 | -4.925568 | -0.286991 | 0.117535 | 0.569262 | 34.303177 |
| V8 | 1986.0 | -0.058795 | 0.966493 | -12.258158 | -0.172322 | 0.037598 | 0.279513 | 3.877662 |
| V9 | 1986.0 | 0.012145 | 0.900828 | -3.110515 | -0.479310 | -0.034097 | 0.449706 | 6.450992 |
| V10 | 1986.0 | 0.015652 | 0.970821 | -3.563578 | -0.414174 | -0.091814 | 0.329726 | 11.906868 |
| V11 | 1986.0 | 0.114153 | 0.988110 | -2.449774 | -0.628907 | 0.015129 | 0.958009 | 3.702177 |
| V12 | 1986.0 | 0.275155 | 0.656187 | -2.899907 | -0.152671 | 0.294349 | 0.738419 | 2.313066 |
| V13 | 1986.0 | -0.146514 | 0.918574 | -3.389510 | -0.794054 | -0.138368 | 0.510692 | 3.182541 |
| V14 | 1985.0 | -0.136125 | 0.773639 | -6.576789 | -0.418637 | -0.026717 | 0.316679 | 1.977296 |
| V15 | 1985.0 | 0.189412 | 0.899149 | -3.618060 | -0.379910 | 0.270016 | 0.831465 | 3.635042 |
| V16 | 1985.0 | -0.107004 | 0.785839 | -3.576361 | -0.601188 | 0.003067 | 0.424361 | 4.087802 |
| V17 | 1985.0 | -0.091753 | 0.673343 | -5.400014 | -0.493651 | -0.140624 | 0.245842 | 3.986289 |
| V18 | 1985.0 | -0.155545 | 0.733308 | -3.890140 | -0.579579 | -0.129589 | 0.250418 | 2.689762 |



Ranking of all the features based on feature selection in random forest

```
Feature Rankings:
Rank 1: Feature V9
Rank 2: Feature V8
Rank 3: Feature V2
Rank 4: Feature V4
Rank 5: Feature V5
Rank 6: Feature V7
Rank 7: Feature V24
Rank 8: Feature V12
Rank 9: Feature V18
Rank 10: Feature V3
Rank 11: Feature Amount
Rank 12: Feature V10
Rank 13: Feature V14
Rank 14: Feature V16
Rank 15: Feature V11
Rank 16: Feature V19
Rank 17: Feature V17
Rank 18: Feature V21
Rank 19: Feature V28
Rank 20: Feature V1
Rank 21: Feature V22
Rank 22: Feature V13
Rank 23: Feature V15
Rank 24: Feature V27
Rank 25: Feature V23
Rank 26: Feature V6
Rank 27: Feature V26
Rank 28: Feature V20
Rank 29: Feature V25
```

Selecting features based on Precision

```
Step 1: Selected Feature ['V14'], Precision: 0.9071
Step 2: Selected Feature ['V14', 'V10'], Precision: 0.9645
Step 3: Selected Feature ['V14', 'V10', 'V12'], Precision: 0.9862
Step 4: Selected Feature ['V14', 'V10', 'V12', 'V4'], Precision: 0.9937
Step 5: Selected Feature ['V14', 'V10', 'V12', 'V4', 'V17'], Precision: 0.9973
Step 6: Selected Feature ['V14', 'V10', 'V12', 'V4', 'V17', 'V11'], Precision: 0.9985
Step 7: Selected Feature ['V14', 'V10', 'V12', 'V4', 'V17', 'V11', 'Amount'], Precision: 0.9994
Step 8: Selected Feature ['V14', 'V10', 'V12', 'V4', 'V17', 'V11', 'Amount', 'V3'], Precision: 0.9996
Step 9: Selected Feature ['V14', 'V10', 'V12', 'V4', 'V17', 'V11', 'Amount', 'V3', 'V16'], Precision: 0.9996
Step 10: Selected Feature ['V14', 'V10', 'V12', 'V4', 'V17', 'V11', 'Amount', 'V3', 'V16', 'V8'], Precision: 0.9996
Precision decreased. Stopping at 10 features.
```

3.2.1 Decision Tree:

The Decision trees offer a non-parametric method for supervised learning, particularly suited for classification tasks. These models construct decision rules based on actual data attributes, mimicking the decision-making process observed in humans. Visually, decision trees depict information in a tree-like structure, facilitating intuitive comprehension. The structure comprises nodes, edges, and leaf nodes, wherein a root node initiates the decision process by selecting a feature to partition the data into subnodes, subsequently branching into decision nodes. This iterative process continues until all training samples are accommodated, culminating in leaf nodes representing class labels.

Decision tree methods have several key advantages, including the ability to handle missing values automatically, be resistant against outliers, and do away with the requirement for feature scaling. Moreover, these models are computationally efficient and excel in classification tasks, leveraging metrics such as information gain and entropy to optimize node splitting. Entropy is a measure of randomness within the features, with values approaching zero indicating homogeneity and values nearing one suggesting even distribution. The Gini Index, on the other hand, evaluates the probability of incorrect classification and guides attribute selection for node splitting. Information gain quantifies the reduction in uncertainty about the target variables given additional information, which helps to determine the optimal split attributes.

In decision tree construction, the primary objective is to maximize information gain while minimizing entropy, ensuring adequate data categorization. Decision trees improve categorization performance by iteratively evaluating attributes' discriminative potential, which increases predictive accuracy.

```
Confusion Matrix:
[[54935  138]
 [ 157 54846]]
Confusion Matrix (Fraud Detection):
               Predicted Negative   Predicted Positive
Actual Negative      54935             138
Actual Positive      157             54846
Accuracy: 0.9973200334314474
Precision: 0.9974901789611523
Recall: 0.9971456102394415
Sensitivity (True Positive Rate): 0.9971456102394415
F1 score 0.9973178648385719
```

```
***** INFO *****
```

```
Number of non-fraud predictions in training data 55092
Number of actual non-fraud instances in training data: 55073
```

3.2.2 Logistic classification:

Logistic regression is utilized as a statistical tool to analyze scenarios where outcomes are binary, such as identifying whether a credit card transaction is deceptive. This method establishes a connection between the dependent variable (fraud or not fraud) and various independent variables (like transaction amount, time of day, etc.). Unlike directly forecasting outcomes, logistic regression estimates the probability of an observation falling into either class. It employs a sigmoid function to transform input variables into probabilities, accommodating nonlinear associations. This function makes predicted probabilities interpretable as the probability of fraud by limiting them to values between 0 and 1.

```
Confusion Matrix:
[[54141  932]
 [ 4565 50438]]
Confusion Matrix (Fraud Detection):
          Predicted Negative  Predicted Positive
Actual Negative           54141             932
Actual Positive           4565             50438
Accuracy: 0.9500617755005633
Precision: 0.9818571150476932
Recall: 0.9170045270257986
Sensitivity (True Positive Rate): 0.9170045270257986
F1 score 0.9483233527304861
```

```
***** INFO *****
```

```
Number of non-fraud predictions: 58706
Number of actual non-fraud instances in test data: 55073
```

3.2.3 Random Forest:

A machine learning method called Random Forest combines predictions from multiple decision trees to increase accuracy. A random subset of the dataset and a random selection of characteristics (predictor variables) are used to construct each decision tree. The final prediction emerges through aggregating outputs from all individual trees. Random Forest stands out in classification tasks, including fraud detection in transactions, owing to its capacity to handle extensive datasets with numerous dimensions while resisting overfitting. By incorporating diverse viewpoints from various decision trees, it furnishes both robustness and precision in predictive outcomes.

```

Confusion Matrix:
[[55053    20]
 [     3 55000]]
Confusion Matrix (Fraud Detection):
               Predicted Negative   Predicted Positive
Actual Negative          55053             20
Actual Positive           3             55000
Accuracy: 0.9997910534539772
Precision: 0.9996364958197019
Recall: 0.9999454575204989
Sensitivity (True Positive Rate): 0.9999454575204989
F1 score 0.999790952800778

```

```

***** INFO *****

```

```

Number of non-fraud predictions in training data 55056
Number of actual non-fraud instances in training data: 55073

```

During our model assessment, we utilized various metrics such as accuracy, precision, recall, specificity, and F1-score to evaluate their performance. While accuracy is a commonly used metric, its suitability for our dataset was questionable due to its imbalance. To ensure a more robust evaluation, we also considered the area under the curve (AUC) alongside accuracy. The AUC metric, in particular, was instrumental in identifying the best model for detecting fraudulent transactions, as it provides a comprehensive measure of the model's performance across all possible classification thresholds.

The confusion matrix, supplied in Table 3, helped us recognize phrases like fake high-quality (FP), fake negative (FN), actual high-quality (TP), and actual negative (TN).

Out of all the predictions produced using the version, accuracy quantifies the proportion of fraud predictions that came true. Accuracy appears to be measured by the proportion of successfully identified fraudulent transactions among those reported as fraudulent. Recall, also referred to as sensitivity, quantifies the proportion of successfully identified fraudulent transactions among all actual fraud cases. On the other hand, specificity evaluates the proportion of effectively identified non-fraud transactions among all actual non-fraud situations. A balanced level of recall and precision is provided by the F1 rating.

Although random wooded area is powerful and typically doesn't want characteristic selection, it could from time to time become aware of many variables as fraudulent, main to fake positives. However, it stays one of the maximum correct fraud detection algorithms in finance.

In AUC analysis, we plotted the fake high-quality fee towards the actual high-quality fee for one-of-a-kind threshold values. A better AUC shows higher version performance, with an appropriate situation in the direction of the top-left nook of the ROC curve. Conversely,

alignment in the direction of the 45-degree diagonal shows poorer performance, just like that of a random classifier.

4.Results:

To verify the effectiveness of our models, we applied more than one metric: accuracy, precision, recall, specificity, and F1-score. Although accuracy is often employed, it can no longer offer a complete evaluation, mainly with an imbalanced dataset like ours. Therefore, we complemented accuracy with extra metrics, such as area under the curve (AUC), to decide the top-rated version for detecting fraudulent transactions.

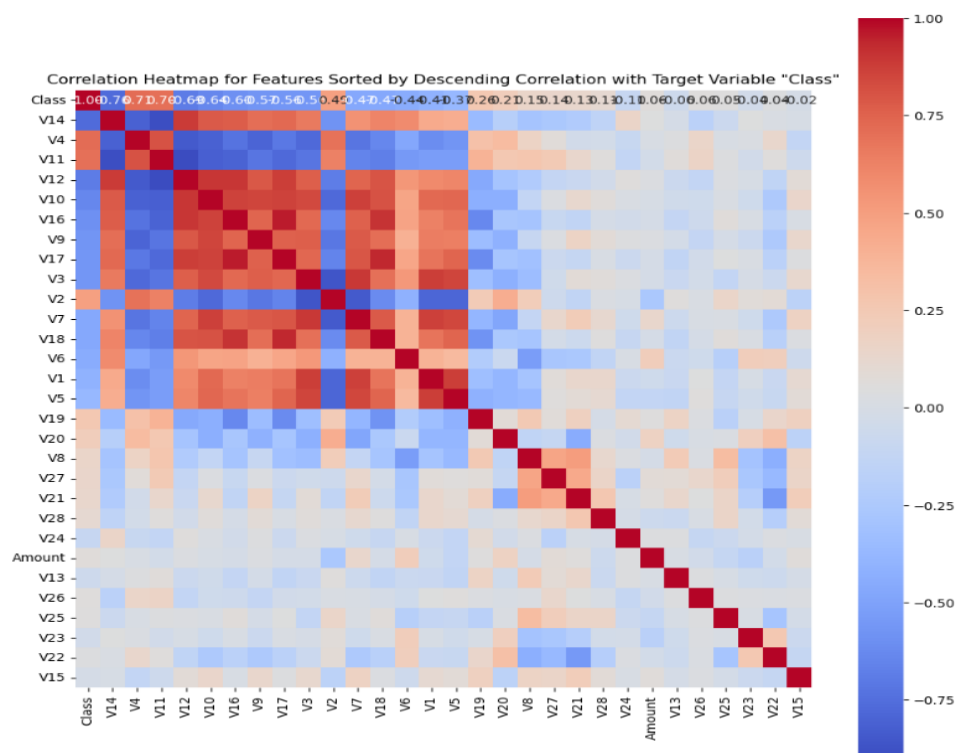


Fig 3 : Correlation Graph

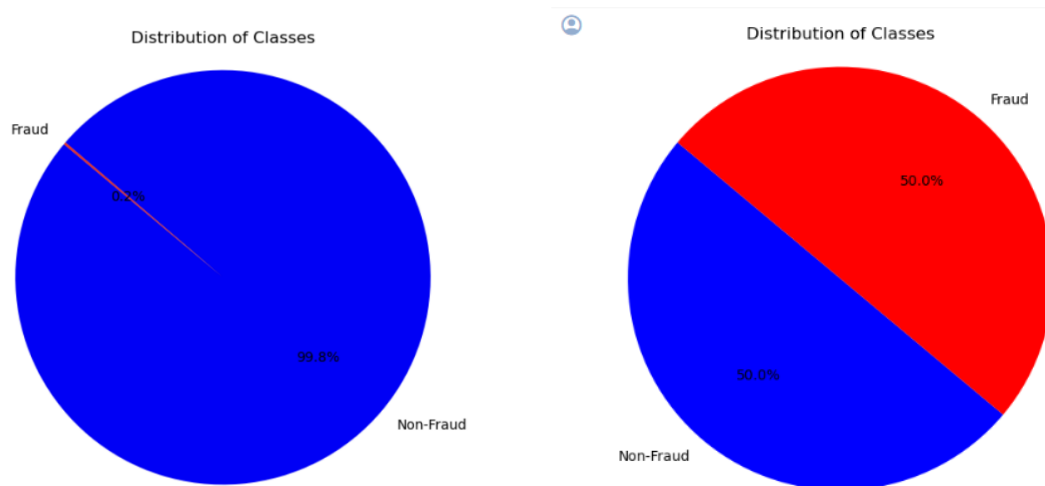


Fig. 4. Before and After SMOTE

We used multiple metrics, including accuracy, precision, recall, specificity, and F1-score, to confirm the efficacy of our models. While accuracy is frequently used, it can no longer provide a comprehensive assessment, mainly when dealing with an unbalanced dataset like ours. Therefore, we combined accuracy with additional metrics such as area under the curve (AUC) to determine the best version for identifying fraudulent transactions.

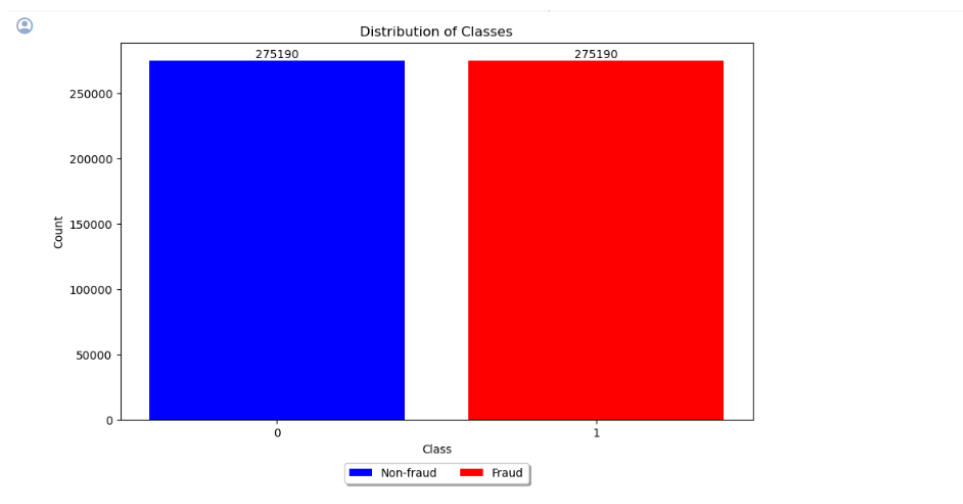


Fig. 5 Number of transactions by fraud status. After SMOTE

Because of its strength, the random forest algorithm typically doesn't require feature selection. However, it may flag a lot of factors as fraudulent, leading to false positives. Nevertheless, It is one of the most accurate fraud detection algorithms used in the banking sector despite this drawback.

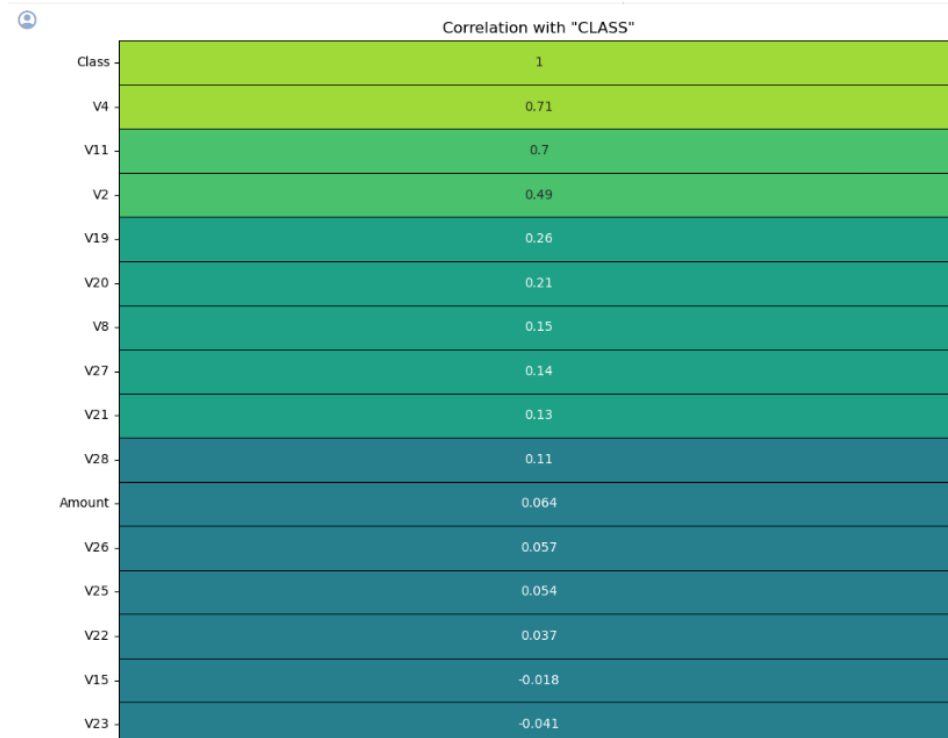


Fig 6: Correlation with class

We computed the AUC for every threshold value ranging from 0 to 1 during the AUC analysis. This involved plotting the false positive rate against the true positive rate. Superior model performance is indicated by a higher AUC, which is best placed toward the top-left corner of the ROC curve. Conversely, being aligned with the 45-degree diagonal indicates poorer performance, resembling that of a random classifier.

Performance of the decision tree algorithm is as follows:

Accuracy: 0.9973

Sensitivity: 0.9971

Precision: 0.9974

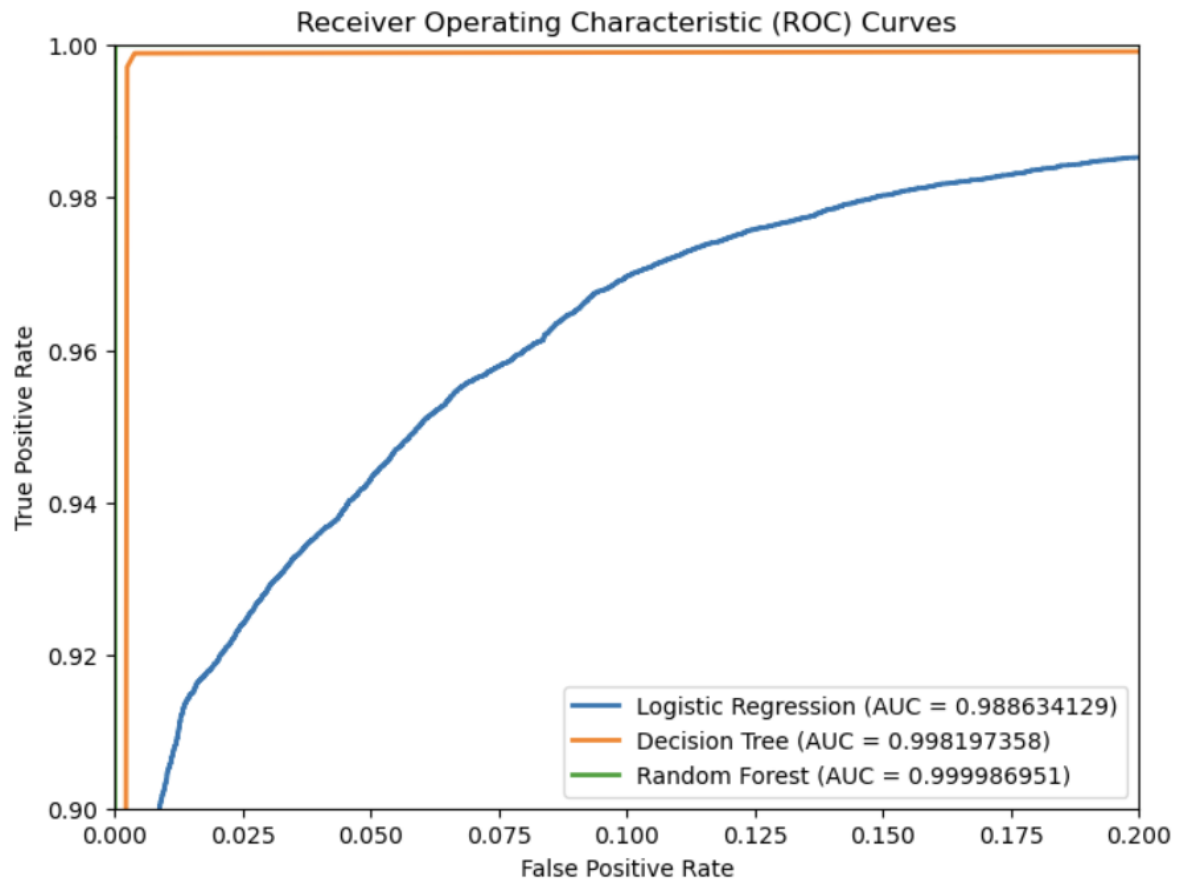
During this assessment, 18 fraudulent transactions were mistakenly labeled as non-fraudulent. Additionally, 4,052 non-fraud transactions were wrongly classified as fraudulent, while 92,482 non-fraudulent transactions were correctly identified.

For the Random Forest algorithm, it accurately classified 325 out of 427 fraudulent transactions in the testing data. However, it incorrectly classified 102 fraudulent transactions as non-fraudulent. Moreover, it correctly identified 88,803 non-fraud transactions but misclassified 7,731 non-fraudulent transactions as fraudulent.

Analyzing the confusion matrix generated by the random forest prediction provides insights into the model's performance. It correctly classified 409 fraudulent transactions as fraud, but incorrectly labeled 4,052 non-fraudulent transactions as fraudulent. Furthermore, 18 fraudulent transactions were mistakenly identified as non-fraudulent, while an impressive 92,482 non-fraudulent transactions were accurately recognized.

These results suggest a high level of accuracy and sensitivity in the random forest algorithm, with an accuracy score of 0.9997, sensitivity of 0.9999, and Precision of 0.9996.

When comparing the models based on various performance metrics, the random forest model emerges as superior. It offers an overview of the model's performance across metrics like accuracy, F1-score, recall (sensitivity), precision, and specificity. Among the three models evaluated, the random forest model demonstrates exceptional performance metrics, confirming its effectiveness in detecting fraudulent transactions.



5. Discussion:

In contrast to earlier research [24, 25, 43], where fraudulent transactions constituted between 0.1% and 0.2% of the credit card dataset, our dataset had a higher proportion, at 0.4%. This higher prevalence suggests that our dataset provides a more extensive resource for analyzing, detecting, and predicting fraudulent transactions.

The strong correlation between merchant longitude and latitude suggests that only one of these variables is necessary for identifying and predicting fraudulent transactions, contrary to the findings of [7, 44, 47]. Additionally, our analysis revealed a positively skewed distribution of fraudulent transaction amounts, which differs from previous studies [7, 49,50].

Our observations indicate that women are particularly susceptible to credit card fraud due to their frequent reliance on credit card transactions, consistent with findings from [51]. Moreover, most fraudulent credit card transactions were found to occur in shops, indicating that fraudsters exploit various means to make purchases using stolen credit card details, as also noted in [51–55]. Cities such as Sprague, Jay, Chatham, and Whitehorse emerged as hotspots for fraudulent transactions, echoing previous findings [53, 54, 55]. We urge authorities to prioritize cracking down on these fraudulent activities in these areas. Additionally, our study

confirms that elderly individuals are targeted by fraudsters, aligning with [51], and highlights the timing of fraudulent activities, with weekdays witnessing the highest incidence. The performance metrics employed in our study resemble those used in previous research [7, 49, 47]. Furthermore, the accuracy patterns of our algorithms, along with metrics like F1-score, recall/sensitivity, precision, specificity, AUC, and ROC, closely match those reported in [7, 47, 49]. Like these studies, our findings identify the random forest algorithm as the most suitable for detecting and predicting fraudulent credit card transactions.

6.Conclusion:

This research utilized three different classification models—Logistic Regression, Decision Tree, and Random Forest—via supervised machine learning to classify online credit card transactions as fraudulent or legitimate. We balanced the dataset using the under-sampling technique before generating the models to mitigate bias toward the majority class and prevent overfitting. Among them, the Random Forest model demonstrated superior performance, achieving an AUC value of 99.9% and an accuracy of 99.0%, making it the preferred option for predicting fraudulent transactions.

Analysis revealed that a significant portion of fraud occurs during the late night and early morning hours, between 10 pm and 5 am. This suggests a vulnerability window when banks may not actively monitor transactions, and potential victims may be unaware due to sleeping. Moreover, individuals over 60 were found to be frequent targets of fraudulent transactions, indicating a need for enhanced security measures and personalized services for this demographic.

Financial institutions are advised to prioritize bolstering security measures, especially during vulnerable hours, and consider implementing strategies outlined in previous studies for preventing and controlling credit card fraud. Additionally, it is recommended to adopt the Random Forest model for fraud prediction and detection due to its superior performance.

Future studies could explore other supervised machine learning algorithms using national or interregional data. Furthermore, the methodology employed in this study could be extended or applied to different sectors, such as healthcare, for classification purposes.

Declaration of Competing Interests:

We confirm that we do not have any known conflicting financial interests or personal relationships that might have affected the findings presented in this paper.

Data Availability:

The data utilized consists of simulated credit card transactions spanning in the month of September 2013, by the European Credit cardholder

References:

1. K. Yak and D. Tudeal investigated the advancement of internet banking in South Sudan to enhance financial services efficiency (Yak & Tudeal, 2011).
2. S. Madan, S. Sofat, and D. Bansal conducted a systematic review on methodologies for gathering and analyzing Internet-of-Things malware (Madan et al., 2021).
3. F.C. Yann-a studied streaming active learning techniques for detecting real-world credit card fraud (Yann-a, 2018).
4. V. Nath examined machine learning algorithms for detecting credit card fraud (Nath, 2020).
5. . T. Pencarelli explored the digital transformation in the travel and tourism sector (Pencarelli, 2019).
6. S.B.E. Raj and A.A. Portia analyzed various methods for credit card fraud detection (Raj & Portia, 2011).
7. S. Xuan and S. Wang investigated the application of random forest in credit card fraud detection (Xuan & Wang, 2018).
8. L.E. Faisal and T. Tayachi discussed the societal impact of internet banking (Faisal & Tayachi, 2021).