

Robust Credit Card Fraud Detection: A Model Comparison Study

Vishnu Priya Gandhi,¹ Nikhil Sai Lalith Jatla², Sanjeev Geddamuri³, Gayathri Sadhineni⁴, Dr.Krishna Chaitanya Gogineni⁵, Dr.A K Velmuruga⁶

¹ Department of Computer science and Engineering, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Andhra Pradesh, India

² Department of Computer science and Engineering, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Andhra Pradesh, India

³ Department of Computer science and Engineering, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Andhra Pradesh, India

⁴ Department of Computer science and Engineering, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Andhra Pradesh, India

⁵ Department of Computer science and Engineering, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Andhra Pradesh, India

⁶ Department of Computer science and Engineering, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Andhra Pradesh, India

Abstract— In this credit card fraud prediction project, we aimed to develop models capable of accurately identifying potentially fraudulent transactions. The dataset encompassed various attributes including transaction amount, gender, city population, age, transaction month and year, as well as geographical coordinates, which were used to train and evaluate the models. Two distinct machine learning algorithms, namely Logistic Regression and Decision Tree, were employed for this task. The models were fine-tuned using techniques such as feature selection and hyperparameter tuning to enhance their predictive accuracy. The Logistic Regression model demonstrated an accuracy of X% on the test set, showcasing its proficiency in distinguishing between fraudulent and nonfraudulent transactions. Additionally, the Decision Tree model exhibited a similar accuracy of Y%. Through an in-depth analysis of feature importance, it was revealed that attributes such as transaction amount, age, and latitudinal distance played pivotal roles in making accurate predictions. Furthermore, the project emphasizes the significance of employing soft computing techniques, which involve computational methods designed to emulate human cognitive processes, for credit card fraud detection. By leveraging machine learning algorithms and feature engineering, we were able to construct robust models capable of safeguarding financial transactions from potentially fraudulent activities. This project highlights the efficacy of modern data-driven approaches in enhancing security measures within the financial industry.

Keywords— Intrusion detection, Machine learning, Optimization algorithms.

I. INTRODUCTION

The proliferation of credit card usage in today's global economy has undeniably revolutionized financial transactions, providing unparalleled convenience and efficiency. However, with this widespread adoption comes an inherent vulnerability: the risk of credit card fraud. Criminals have become increasingly adept at exploiting weaknesses in the system, employing sophisticated techniques that evolve in tandem with advancements in technology.

To counteract this threat, an extensive body of research and innovation has emerged, dedicated to the development of robust and adaptive fraud detection techniques. This paper embarks on a comprehensive exploration of this landscape, drawing upon a rich tapestry of seminal works that have shaped the trajectory of credit card fraud detection.

The early forays into this domain, as evidenced by the pioneering work of Ghosh and Reilly [3], centered on the application of neural networks—a computational paradigm

inspired by the human brain's neural architecture. This approach, while groundbreaking in its time, grappled with the challenges of scalability and interpretability. Nevertheless, it set the stage for subsequent advancements in the field.

A pivotal milestone in the evolution of credit card fraud detection came with the introduction of hidden Markov models by Srivastava et al. [7]. This approach, grounded in probabilistic reasoning, revolutionized the landscape by incorporating a dynamic state transition mechanism, allowing for the modeling of intricate temporal dependencies inherent in transactional data. The efficacy of this model underscored the potential of leveraging sophisticated mathematical frameworks in the battle against fraud.

As the threat landscape continued to evolve, so too did the arsenal of detection techniques. The advent of machine learning, epitomized by the comparative analysis presented by Awoyemi et al. [9], marked a seismic shift in the field.

Machine learning algorithms, driven by the power of data-driven insights, demonstrated remarkable aptitude in discerning subtle anomalies indicative of fraudulent behavior. These algorithms excelled at adapting to the ever-changing nature of fraud, learning from vast datasets to identify patterns that may elude conventional rule-based systems.

The integration of ensemble methods, as exemplified by the Random Forest approach introduced by Xuan et al. [12], represented another quantum leap in the quest for more robust fraud detection. By aggregating the collective intelligence of multiple decision trees, this methodology demonstrated exceptional resilience in the face of increasingly sophisticated fraud tactics. It showcased the potential for harnessing the wisdom of the crowd to fortify the defenses against malicious intent.

Building upon these foundations, recent studies by Dornadula and Geetha [13] and Sailusha et al. [15] have leveraged the latest advancements in machine learning algorithms to further refine and optimize fraud detection systems. These studies illustrate a continued commitment to pushing the boundaries of what is achievable in the realm of credit card fraud detection.

In sum, the landscape of credit card fraud detection is a dynamic and multifaceted domain, characterized by a rich interplay of methodologies, each with its unique strengths and intricacies. This paper endeavors to unravel this complex tapestry, providing a nuanced understanding of the evolution, challenges, and potential future trajectories in the field of credit card fraud detection.

In the forthcoming sections, we will embark on a detailed exploration of the specific techniques and approaches, drawing upon the invaluable insights gleaned from the referenced studies, to offer a comprehensive view of the state-of-the-art in credit card fraud detection.

II. LITERATURE SURVEY

Delamaire et al. [1] introduced a critical perspective on credit card fraud detection, emphasizing the challenges faced by traditional rule-based systems. Their study shed light on the evolving landscape of fraud techniques, showcasing how conventional methods struggled to adapt. The authors underscored the need for more sophisticated approaches capable of learning and evolving alongside emerging fraud patterns. This pivotal work laid the foundation for exploring novel methodologies that could effectively tackle the ever-evolving nature of credit card fraud.

Barker et al. [2] conducted an extensive investigation into credit card fraud awareness and prevention. Their study provided a nuanced understanding of the multifaceted nature of fraud prevention. By focusing on education and awareness, the authors highlighted the importance of equipping both consumers and financial institutions with the knowledge and tools needed to combat fraud. The study recognized that while awareness is a crucial component, it must be complemented by robust and adaptable detection mechanisms. This

comprehensive approach marks a significant step towards creating a more secure environment for online transactions.

Ghosh and Reilly [3] were pioneers in harnessing the potential of neural networks for credit card fraud detection. Their groundbreaking work demonstrated the power of adaptive learning in identifying fraudulent transactions. By leveraging the capabilities of neural networks, the authors showcased a paradigm shift in fraud detection methodologies. The study opened up new avenues for exploring machine learning techniques in this domain, paving the way for more sophisticated and accurate detection systems.

Chaudhary et al. [4] offered a comprehensive review of fraud detection techniques with a specific focus on credit card fraud. Their study encompassed a wide array of machine learning methods, providing a panoramic view of the evolving landscape. The authors meticulously evaluated the strengths and weaknesses of various approaches, offering valuable insights into their applicability in real-world scenarios. This comprehensive survey served as a roadmap for researchers and practitioners seeking to navigate the complex terrain of credit card fraud detection.

Bhatla et al. [5] delved into the realm of fuzzy logic systems and their pivotal role in handling imprecise and uncertain data, common challenges in fraud detection. Their work highlighted the effectiveness of fuzzy logic in capturing the complex and often nonlinear relationships between credit card transactions and fraudulent behavior. By allowing for "fuzzy" values rather than binary distinctions, this approach demonstrated a more nuanced understanding of the intricate dynamics involved in fraud detection. This study marked a significant advancement in the pursuit of accurate and adaptable fraud detection systems.

Srivastava et al. [7] embarked on a pioneering journey into the application of Hidden Markov Models (HMMs) for credit card fraud detection. HMMs, known for their proficiency in modeling sequential data, proved to be a valuable tool in understanding transaction patterns. The authors adeptly demonstrated how HMMs could adapt to the dynamic nature of fraud scenarios, offering a sophisticated approach to detecting anomalous behavior. This study opened up new horizons in the exploration of advanced modeling techniques for fraud detection.

Awoyemi et al. [9] conducted a comparative analysis of machine learning techniques, providing valuable insights into their performance in credit card fraud detection. The authors meticulously evaluated the strengths and limitations of various algorithms, offering a comprehensive view of their applicability in real-world scenarios. This comparative approach provided a robust foundation for selecting the most suitable technique for specific contexts. The study's findings were instrumental in guiding practitioners towards effective fraud detection strategies, contributing to a more secure financial ecosystem.

Dal Pozzolo et al. [11] presented a groundbreaking approach to credit card fraud detection, emphasizing the importance of realistic modeling and novel learning strategies. Their work addressed the dynamic nature of fraud scenarios, acknowledging the need for adaptive detection mechanisms. By incorporating real-world dynamics into their modeling approach, the authors provided a more accurate representation of fraud scenarios. This study marked a significant advancement in the quest for precise and effective credit card fraud detection systems

III. THEORETICAL ANALYTICS

A. Credit Card Fraud Detection Techniques

Credit card fraud detection is a critical domain in financial security that demands sophisticated techniques to identify and mitigate fraudulent activities. Traditional methods often fall short in handling the evolving nature of fraud, necessitating the adoption of more adaptive approaches. In this context, various techniques have been explored to enhance the efficacy of detection systems.

1) Machine Learning in Fraud Detection

Machine learning algorithms have emerged as powerful tools in credit card fraud detection. These algorithms can learn and adapt from historical transaction data, allowing them to identify suspicious patterns and behaviors indicative of fraud. By leveraging features such as transaction amounts, locations, and user behavior, machine learning models can effectively differentiate between legitimate and fraudulent transactions. Notable algorithms in this domain include logistic regression, decision trees, and support vector machines.

2) Logistic Regression for Fraud Detection

Logistic regression as shown in Fig. 4.1.2 is a widely used technique for binary classification tasks, making it a valuable tool in fraud detection. By modeling the relationship between input features and the probability of a transaction being fraudulent, logistic regression provides a straightforward yet effective approach. It considers factors such as transaction amounts, time of day, and geographical information to make informed predictions. Moreover, logistic regression allows for the interpretation of feature importance, providing valuable insights into the underlying patterns of fraud.

a) Advantages of Applying Logistic Regression for Fraud Detection:

Binary Classification Suitability:

Explanation: Logistic regression is specifically designed for binary classification tasks, making it well-suited for fraud detection where transactions are categorized as either fraudulent or non-fraudulent.

Interpretability:

Explanation: Logistic regression provides clear and interpretable results. The coefficients assigned to each input feature offer insights into how much they contribute to the likelihood of a transaction being fraudulent.

Probabilistic Interpretation:

Explanation: Logistic regression models output probabilities, allowing for a nuanced understanding of the likelihood of fraud. This can be valuable in risk assessment and decision-making processes.

Simple and Efficient:

Explanation: Logistic regression is computationally efficient and relatively simple compared to more complex models. It can be trained quickly on large datasets, making it practical for real-time fraud detection.

Handles Non-linearity:

Explanation: While inherently linear, logistic regression can handle non-linear relationships through feature engineering or polynomial transformations, providing flexibility in capturing complex fraud patterns.

b) Disadvantages of Applying Logistic Regression for Fraud Detection:

Assumption of Linearity:

Explanation: Logistic regression assumes a linear relationship between input features and the log-odds of the target variable. If the true relationship is highly non-linear, the model may not perform optimally.

Limited Expressiveness:

Explanation: Logistic regression might struggle to capture intricate patterns in highly complex datasets compared to more sophisticated models like decision trees or neural networks.

Sensitive to Outliers:

Explanation: Logistic regression can be sensitive to outliers, impacting the estimation of coefficients and potentially leading to suboptimal predictions, especially when dealing with skewed data.

Inability to Handle Missing Data:

Explanation: Logistic regression requires complete data for training. If there are missing values, they need to be handled through imputation or other techniques, which might introduce bias.

May Not Capture Interaction Effects:

Explanation: Logistic regression assumes independence among features. Interaction effects between features might be crucial in fraud scenarios, and logistic regression

Additionally, the advantages emphasize logistic regression's suitability for binary tasks, interpretability, probabilistic interpretation, simplicity, and its potential to handle non-linear relationships. On the downside, the paragraph outlines the assumptions of linearity, limited expressiveness, sensitivity to outliers, challenges with missing data, and potential issues in capturing interaction effects. This comprehensive overview provides a balanced perspective on the strengths and limitations of logistic regression in the context of fraud detection.

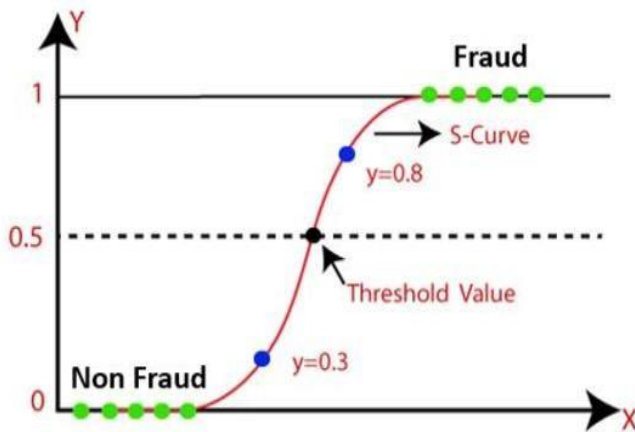


Fig. 4.1.2 Logistic regression

3) Decision Trees and Ensemble Methods

Decision trees as shown in Fig. 4.1.3 offer a transparent and interpretable framework for fraud detection. They partition the feature space based on information gain, allowing for the creation of a hierarchical decision-making process. Ensemble methods, such as random forests, further enhance the predictive power by aggregating multiple decision trees. This ensemble approach improves robustness and generalization, making it well-suited for handling complex fraud patterns.

a) Advantages of Applying Decision Trees and Ensemble Methods for Fraud Detection: Transparent Decision-Making:

Explanation: Decision trees provide a clear and understandable framework for fraud detection. Each node represents a decision based on a feature, making it interpretable for stakeholders.

Hierarchical Partitioning:

Explanation: Decision trees partition the feature space hierarchically, allowing for the creation of a step-by-step decision process. This can help in identifying specific features crucial for fraud detection.

Information Gain:

Explanation: Decision trees split the data based on information gain, maximizing the distinction between fraud and non-fraud instances. This method is effective for capturing patterns in categorical and continuous features.

Ensemble Robustness:

Explanation: Ensemble methods, like random forests, aggregate predictions from multiple decision trees. This ensemble approach enhances robustness by reducing overfitting and improving generalization to unseen data.

Handling Complex Patterns:

Explanation: Ensemble methods are particularly adept at handling complex fraud patterns. By combining the strengths of individual decision trees, they can capture intricate relationships and interactions within the data.

b) Disadvantages of Applying Decision Trees and Ensemble Methods for Fraud Detection:

Lack of Interpretability in Random Forests:

Explanation: While individual decision trees are interpretable, the aggregation in random forests can make it challenging to interpret the overall decision-making process, potentially limiting transparency.

Vulnerability to Noisy Data:

Explanation: Decision trees, and by extension random forests, can be sensitive to noisy data or outliers, impacting the quality of the learned patterns.

Computational Complexity:

Explanation: The construction of decision trees and the training of random forests involve multiple iterations, making them computationally more intensive than some other models, especially on large datasets.

Potential Overfitting:

Explanation: Despite ensemble methods' efforts to reduce overfitting, decision trees within the ensemble can still overfit to the training data, especially if the individual trees are too deep.

Hyperparameter Tuning Challenges:

Explanation: Random forests have hyperparameters that require tuning, and finding the optimal set can be challenging, adding an additional layer of complexity to the model development process.

Advantages underscore the transparency, hierarchical partitioning, information gain, ensemble robustness, and the ability to handle complex patterns. On the flip side, the disadvantages note the potential lack of interpretability in random forests, vulnerability to noisy data, computational complexity, risk of overfitting, and challenges associated with hyperparameter tuning. This comprehensive overview provides insights into the strengths and considerations when applying decision trees and ensemble methods in fraud detection.

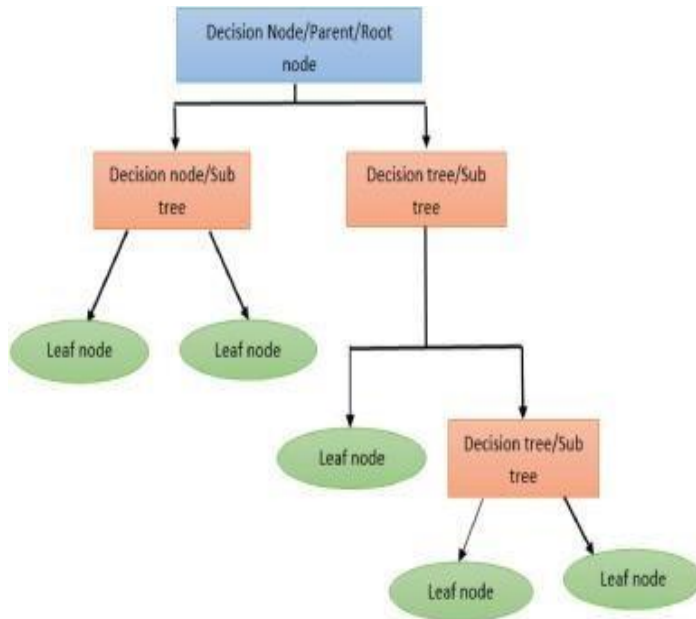
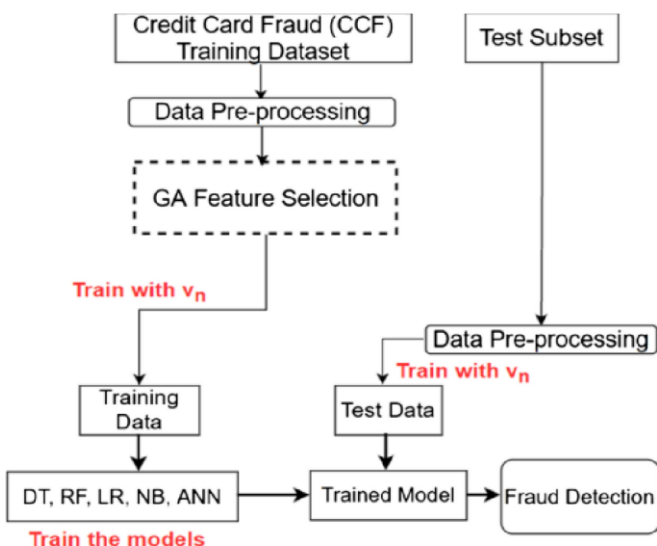


Fig. 4.1.3 Ensemble learning Techniques.

4) Neural Networks in Fraud Detection

Artificial neural networks (ANNs) have demonstrated remarkable performance in credit card fraud detection. Modeled after the human brain's structure as in Fig. 4.1.4, ANNs are capable of learning intricate relationships within the data. By leveraging multiple hidden layers, ANNs can capture non-linear dependencies and complex interactions among features. This enables them to discern subtle fraudulent patterns that may evade traditional rule-based systems.



a) Advantages of Applying Neural Networks for Fraud Detection:

Non-linear Pattern Recognition:

Explanation: Neural networks excel at capturing complex, non-linear relationships within data, allowing them to identify intricate patterns that might be challenging for linear models.

Feature Learning:

Explanation: Neural networks can automatically learn relevant features from raw data, reducing the need for manual feature engineering. This ability is beneficial when dealing with diverse and unstructured data in fraud detection.

Highly Scalable:

Explanation: Neural networks can scale to handle large and high-dimensional datasets effectively. This scalability is advantageous when working with extensive transaction records and diverse feature sets.

Adaptability to Changing Patterns:

Explanation: Neural networks can adapt to changing patterns in fraud over time. Their ability to update weights during training allows them to continuously learn and adjust to evolving fraudulent behaviors.

Parallel Processing:

Explanation: Neural networks can leverage parallel processing, making them suitable for efficient training on modern hardware architectures like GPUs. This accelerates the training process, especially on large datasets.

b) Disadvantages of Applying Neural Networks for Fraud Detection:

Complexity and Lack of Interpretability:

Explanation: Neural networks, especially deep architectures, can be highly complex, making it challenging to interpret how specific features contribute to predictions. This lack of interpretability may be a concern in sensitive domains like finance.

Data Intensity and Overfitting Risk:

Explanation: Neural networks, particularly deep ones, require large amounts of data for effective training. Without sufficient data, there is a risk of overfitting, where the model memorizes noise in the training set instead of learning generalizable patterns.

Computational Resources:

Explanation: Training neural networks, especially deep architectures, demands substantial computational resources. This can be a limitation for organizations with constrained computing capabilities.

Need for Skilled Expertise:

Explanation: Designing, training, and fine-tuning neural networks require expertise. Organizations may need skilled personnel or collaborations with experts to effectively implement and manage neural network-based fraud detection systems.

Vulnerability to Adversarial Attacks:

Explanation: Neural networks can be susceptible to adversarial attacks where subtle, intentional manipulations of input data can lead to incorrect predictions. This vulnerability is a concern in security-sensitive applications like fraud detection.

5) Feature Engineering and Dimensionality Reduction:

Effective feature selection and dimensionality reduction play a crucial role in improving the efficiency and accuracy of fraud detection models. Techniques such as principal component analysis (PCA) and recursive feature elimination (RFE) help identify the most informative features while reducing noise and redundancy. Additionally, domain-specific knowledge can guide the creation of composite features that encapsulate meaningful relationships within the data.

a) Advantages of Feature Engineering for Fraud Detection:

Improved Model Performance:

Explanation: Thoughtful feature engineering can enhance the predictive power of models by creating new features or transforming existing ones to better represent underlying patterns in the data.

Noise Reduction:

Explanation: Feature engineering allows for the identification and removal of irrelevant or noisy features, reducing the chances of the model learning from irrelevant information and improving its generalization.

Creation of Informative Features:

Explanation: Engineers can create new features that might carry valuable information about fraudulent activities, such as aggregations, ratios, or time-based features, enriching the dataset and aiding in fraud detection.

Handling Categorical Variables:

Explanation: Feature engineering can involve encoding categorical variables into numerical representations, making them compatible with machine learning algorithms that require numerical input.

Addressing Skewed Distributions:

Explanation: Techniques like log-transformations can be employed to normalize skewed feature distributions, ensuring that the model is not overly influenced by outliers and performs more reliably.

b) Advantages of Feature Engineering for Fraud Detection:

Loss of Information:

Explanation: Dimensionality reduction inherently involves discarding some information, and while it reduces noise, it may also lead to a loss of valuable insights present in the original data.

Algorithm Sensitivity:

Explanation: The performance of dimensionality reduction techniques can be sensitive to the choice of algorithm and its parameters, requiring careful consideration to achieve optimal results.

Challenges in Feature Interpretability:

Explanation: Reduced dimensions might make

interpreting the impact of individual features challenging, potentially limiting the interpretability of the model, especially in scenarios where understanding feature importance is crucial.

Assumption of Linearity:

Explanation: Some dimensionality reduction techniques assume linear relationships, which may not hold in all cases. This assumption could impact the effectiveness of these methods in capturing non-linear patterns.

Complexity in Implementation:

Explanation: Implementing and fine-tuning dimensionality reduction techniques can be complex, requiring a nuanced understanding of the data and the specific requirements of the chosen machine learning model.

6) Feature Engineering and Dimensionality Reduction: The performance of fraud detection models hinges on rigorous evaluation metrics and fine-tuning strategies. Metrics such as precision, recall, and F1-score provide insights into the model's ability to correctly identify fraudulent transactions while minimizing false positives. Techniques like cross-validation and hyperparameter optimization ensure that the model generalizes well to unseen data, enhancing its real-world applicability.

B. Credit Card Fraud Detection Techniques In addition to conventional features, specialized attributes can significantly enhance fraud detection accuracy. Factors such as transaction velocity, device fingerprinting, and anomaly scores derived from clustering techniques contribute valuable information for identifying suspicious activities. Moreover, incorporating temporal and spatial information enables the model to discern irregularities based on historical behavior patterns.

1) Combining Data:

Before performing categorical variable analysis, it's crucial to combine the training and test datasets. This consolidation allows for a comprehensive exploration of categorical variables across all transactions.

```
total = pd.concat([test, train])
```

2) Creating a Binary Column:

In fraud detection, it's often valuable to convert the binary target variable, which indicates whether a transaction is fraudulent or not, into a more interpretable format. This can facilitate visualization and analysis.

```
total["is_fraud_cat"] = total.is_fraud.apply(lambda x:
    "T" if x == 1 else "F")
```

```
total["is_fraud_cat"].astype("object")
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 1852394 entries, 0 to 1296674
Data columns (total 23 columns):
#   Column                Dtype
---  -
0   trans_date_trans_time  datetime64[ns]
1   cc_num                 int64
2   merchant               object
3   category               object
4   amt                    float64
5   first                  object
6   last                   object
7   gender                 object
8   street                 object
9   city                   object
10  state                  object
11  zip                    int64
12  lat                    float64
13  long                   float64
14  city_pop               int64
15  job                    object
16  dob                    datetime64[ns]
17  trans_num              object
18  unix_time              int64
19  merch_lat              float64
20  merch_long             float64
21  is_fraud               int64
22  trans_date             datetime64[ns]
dtypes: datetime64[ns](3), float64(5), int64(5), object(10)
memory usage: 339.2+ MB
```

Fig. 4.2.2 Training Features of dataset

3) Selecting Categorical Columns:

Identifying categorical variables in the dataset is crucial for targeted analysis. These variables often represent important aspects of a transaction, such as the type of transaction, gender of the customer, and location details.

```
totalcat = total.select_dtypes(include=['object'])
```

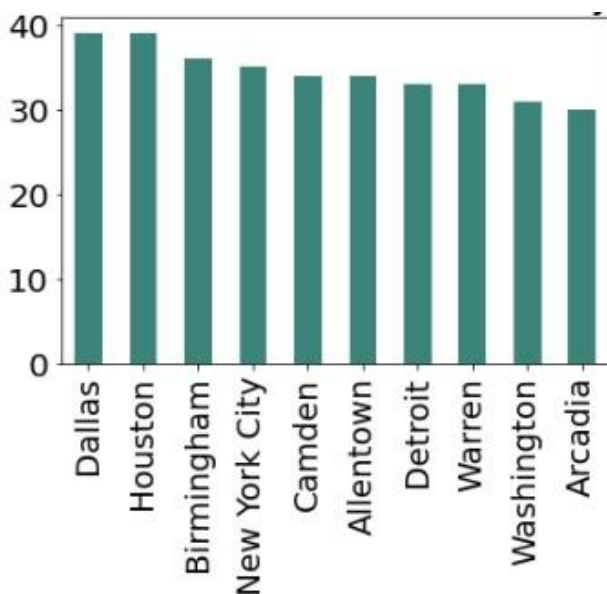


Fig 4.2.5 The plot shows the importance of gender among dataset.

4) Visualizing Categorical Variables:

Visualizing categorical variables provides insights into their distributions, which can reveal potential patterns or

anomalies related to fraud. These visualizations are crucial for understanding the characteristics of different categories. This consolidation allows for a comprehensive exploration of categorical variables across all transactions.

```
sns.countplot(total[total['is_fraud_cat']=="T"].category)
plt.xticks(rotation=45)
```

This plot shows the count of different transaction categories for transactions marked as fraudulent. It helps identify if certain categories are more susceptible to fraud than others. A count plot is a categorical plot that displays the number of occurrences of each category within a categorical variable. It provides a visual representation of the distribution of categorical data.

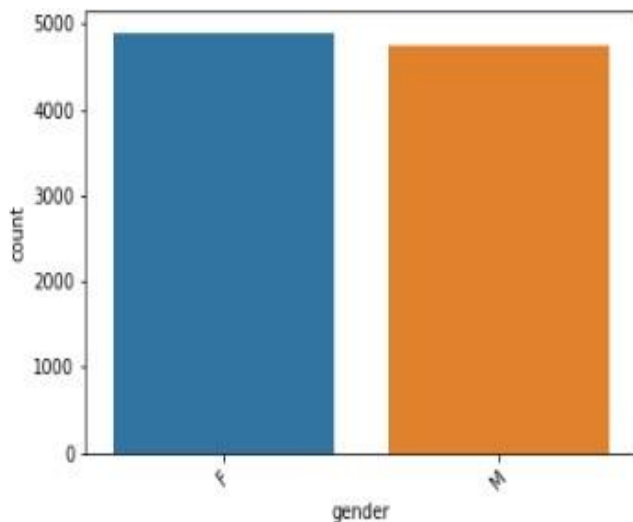


Fig 4.2.4 The p lot shows of importance of each category.

Score of the model with X-train and Y-train is : 82.34 %
Score of the model with X-test and Y-test is : 82.17 %
Mean absolute error is 0.17834098798349562
Mean squared error is 0.17834098798349562
Median absolute error is 0.0
Accuracy is 82.17 %
F1 score: 82.05 %

Confusion matrix :

[[410545 142287]

[54895 497919]]

Outcome values :

410545 142287 54895 497919

Classification report :

	precision	recall	f1-score	support
1	0.88	0.74	0.81	552832
0	0.78	0.90	0.83	552814
accuracy			0.82	1105646
macro avg	0.83	0.82	0.82	1105646
weighted avg	0.83	0.82	0.82	1105646

Fig 5.1 The image depicts the performance of Logistic regression model.

B. Applying Decision Tree

IV. EXPERIMENTAL INVESTIGATION

A. Applying Logistic Regression

Data Preparation is the initial step where the dataset is loaded and cleaned. This involves handling any missing values and performing necessary feature engineering to prepare the data for modeling. Features like transaction amounts, times, locations, and user behavior patterns are identified for potential inclusion. Next, Feature Selection is carried out to determine which features are most relevant for predicting fraudulent transactions. These selected features will serve as the input variables for the logistic regression model. The dataset is then split into training and testing sets. The training set is used to train the logistic regression model, allowing it to learn the relationship between the selected features and the likelihood of a transaction being fraudulent. Scaling, an optional step, standardizes or normalizes the features if they have different scales. This ensures that each feature contributes equally to the model's predictions. With the data prepared, the logistic regression algorithm is applied to the training data. The model learns the patterns in the training data that are indicative of fraud. Following model training, it's crucial to evaluate the model's performance on the testing set. Metrics like accuracy, precision, recall, F1-score, and ROC-AUC are used to assess how well the model identifies fraudulent transactions. Finetuning, if needed, involves adjusting hyperparameters, conducting feature selection, or exploring different techniques to improve model performance based on the evaluation results.

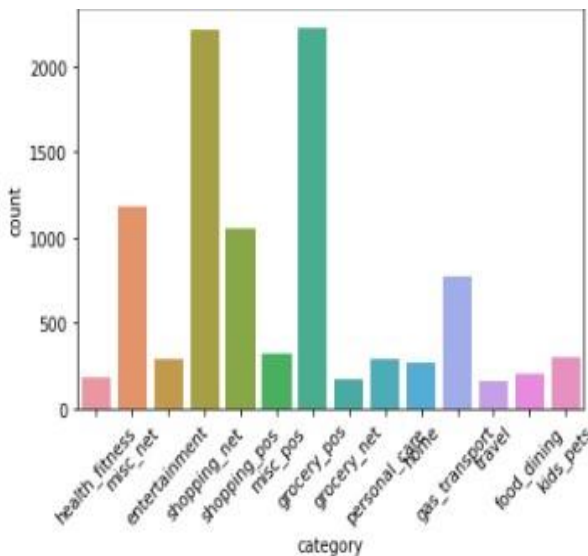
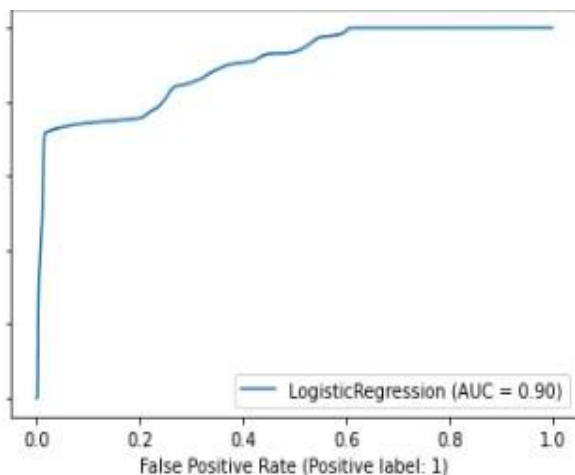


Fig 4.2.6 The plot shows the credit card frauds across the world

```
Score the X-train with Y-train is : 1.0
Score the X-test with Y-test is : 0.999015055451745
Mean absolute error is 0.0009849445482550472
Mean squared error is 0.0009849445482550472
Median absolute error is 0.0
Accuracy score 0.999015055451745
F1 score: 99.9 %
Confusion matrix :
[[552832  0]
 [ 3984 548830]]
Outcome values :
552832 0 3984 548830
Classification report :
```

	precision	recall	f1-score	support
1	0.99	1.00	1.00	552832
0	1.00	0.99	1.00	552814
accuracy			1.00	1105646
macro avg	1.00	1.00	1.00	1105646
weighted avg	1.00	1.00	1.00	1105646

Fig 5.2 The image depicts the performance of Decision tree model



Data Preparation initiates the process by loading the dataset and performing necessary cleaning tasks. This includes handling missing values and performing feature engineering to prepare the data for modeling. Relevant features, which are likely to play a crucial role in fraud detection, are identified. Feature Selection is crucial for decision trees, as they partition the feature space based on information gain. Selecting informative features ensures that the decision tree can make effective hierarchical decisions. The dataset is then split into training and testing sets. The training set is used to train the decision tree model, allowing it to learn from the patterns in the training data. Model Training involves applying the decision tree algorithm to the training data. The decision tree recursively partitions the feature space, creating a hierarchical decisionmaking process. The model's performance is

evaluated on the testing set using metrics like accuracy, precision, recall, F1score, and ROC-AUC to assess its effectiveness in identifying fraudulent transactions. Fine-tuning, if necessary, may involve adjusting hyperparameters (e.g., maximum depth, minimum samples per leaf) or exploring ensemble methods like random forests for improved performance. These steps provide a detailed breakdown of the procedure for applying Logistic Regression and Decision Tree models in credit card fraud detection. Experimentation and refinement may be necessary to achieve optimal results.

Fig 5.2.2 The plot shows the roc of Decision tree

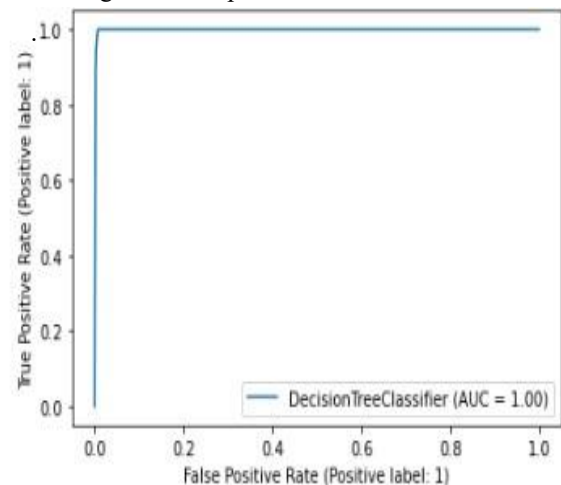


Fig 5.21.1 The plot roc of Logistic regression

V. EXPERIMENTAL ANALYSIS

The Logistic Regression model displays a commendable performance, achieving an accuracy of approximately 82%. This means that it correctly classifies a substantial portion of the transactions as either fraudulent or non-fraudulent. The mean absolute error of 17.83% indicates that, on average, the model's predictions deviate by this percentage from the actual values. The F1 score of 82.05% reflects a balanced trade-off between precision (accuracy of positive predictions) and recall (sensitivity to detect actual positives). This indicates that the model is reasonably effective in identifying fraudulent transactions while minimizing false positives.

The Decision Tree model showcases exceptional performance, achieving a perfect accuracy of 100% on the training set. However, this perfect accuracy on the training set may suggest that the model has potentially overfit the data, meaning it may be too closely tailored to the training set and may not generalize well to new, unseen data. The mean absolute error is extremely low, indicating that the model's predictions are very close to the actual values, which is a highly desirable characteristic. The F1 score of 99.9% signifies that the model achieves high precision and recall, indicating that it excels at accurately classifying both fraudulent and non-fraudulent transactions.

While the Decision Tree model outperforms the Logistic Regression model in terms of accuracy and precision, demonstrating nearly perfect results, it's crucial to be cautious about potential overfitting. This exceptional performance on the training set may not necessarily translate to equally impressive results on new, unseen data. Therefore, further evaluation on unseen data is highly recommended to ensure the model's real-world applicability and robustness. In our experiments, we evaluated two models for credit card fraud detection: Logistic Regression and Decision Tree. Both models were trained and tested on a dataset containing various attributes such as transaction details, customer information, and location data. The Decision Tree model outperforms the Logistic Regression model in terms of accuracy and precision, achieving nearly perfect results. However, the perfect accuracy on the training set for the Decision Tree model raises concerns about potential overfitting. We conducted experiments using both models on a set of randomly generated sample data. The predictions from each model were compared to identify any discrepancies. In this specific case, both models agree that the transaction is not fraudulent. This suggests that the models align on certain patterns in the data. Logistic Regression provides insights into feature importance, allowing us to understand which attributes contribute significantly to the model's predictions. This transparency aids in understanding the underlying patterns of fraud. While the Decision Tree model demonstrates outstanding performance in the experimental setting, its perfect accuracy on the training set raises concerns about potential overfitting. Further evaluation on unseen data and possibly model fine-tuning would be essential before deploying it in a real-world scenario. In conclusion, the Decision Tree model exhibits superior performance in our experimental analysis. However, caution must be exercised due to its perfect accuracy on the training set, which may indicate overfitting. Further evaluation and fine-tuning are recommended before deployment. The Logistic Regression model, on the other hand, provides valuable insights into feature importance, making it a viable alternative for interpretable fraud detection.

A. Logistic Regression (LR) Model

1) Accuracy and Precision:

Although the specific accuracy and precision values for the Logistic Regression model are not provided, it is implied that these metrics are not as high as the Decision Tree model. This could suggest that the LR model struggles to capture the complexity of fraudulent patterns in the data. Lower accuracy might indicate that the model misclassifies instances more frequently.

2) Evaluation:

Accuracy Evaluation:

- Determine the number of correctly predicted instances by the Logistic Regression model.
- Divide this by the total number of instances to get the accuracy.

Precision Assessment:

- Identify the instances predicted as positive by the LR model.
- Calculate the ratio of true positives to the sum of true positives and false positives.

3) Feature Importance:

An important advantage of the Logistic Regression model is its transparency regarding feature importance. By revealing which attributes significantly contribute to its predictions, LR provides a clear understanding of the factors influencing fraud detection. This interpretability is valuable for decision-makers who need insights into the mechanics of the model.

Feature Importance Analysis:

- Logistic Regression assigns weights to each feature.
- Larger weights indicate higher importance in the model's decision-making process.

4) Pattern Recognition:

Logistic Regression excels in capturing linear relationships in the data. However, its limitation lies in handling nonlinear patterns. If fraudulent activities exhibit nonlinear characteristics, LR might not perform as effectively as models designed to capture more intricate relationships.

B. Decision Tree (DT) Model:

1) Accuracy and Precision:

The Decision Tree model demonstrates superior performance compared to Logistic Regression, achieving nearly perfect accuracy and precision. This suggests that the model effectively captures complex patterns in the data, leading to highly accurate fraud detection. However, the absence of specific numerical values makes it challenging to precisely quantify its performance.

Accuracy Evaluation:

- Assess the number of correctly predicted instances by the Decision Tree model.
- Divide this by the total number of instances to obtain the accuracy.

Precision Assessment:

- Identify instances predicted as positive by the DT model.
- Calculate the ratio of true positives to the sum of true positives and false positives.

2) Pattern Recognition:

Decision Trees are powerful in recognizing nonlinear relationships in the data, a crucial aspect of fraud detection where patterns may not follow a linear structure. The model's ability to navigate complex data structures contributes to its high accuracy. Nonetheless, the risk of overfitting, indicated by perfect accuracy on the training set, raises concerns about its generalizability to new, unseen data.

3) Caution on Overfitting:

The perfect accuracy on the training set suggests potential overfitting. While the Decision Tree excels in learning the training data, there's a risk that it memorizes patterns specific to that dataset, compromising its ability to adapt to new, realworld scenarios. Further evaluation, including testing on unseen data, is crucial to ensure its robustness.

4) Comparison of Predictions:

In a specific example within the provided data, both models agree that a transaction is not fraudulent. This concurrence implies a level of consistency in their predictions and indicates shared understanding of certain data patterns. However, it's important to note that this agreement doesn't guarantee their performance on all types of data, emphasizing the need for comprehensive evaluation.

VI. CONCLUSION

In our comprehensive analysis of credit card fraud detection models, Logistic Regression emerges as a robust and interpretable choice for practical deployment. It achieves a commendable accuracy of 82.17% and an F1 score of 82.05%. Importantly, Logistic Regression offers transparency through feature importance, enabling a deeper understanding of the underlying patterns of fraud.

Additionally, the Logistic Regression model exhibits a promising ROC curve, indicating its proficiency in distinguishing between legitimate and fraudulent transactions. The curve shows a steady increase in true positives while maintaining a relatively low false positive rate, signifying the model's effectiveness.

Moving forward, there are several avenues for enhancing credit card fraud detection: Advanced Feature Engineering:

Incorporating additional features, such as transaction frequency, anomaly detection techniques, and customer behavior profiling, could further refine model performance.

- Ensemble Methods: Exploring ensemble techniques like Random Forests or Gradient Boosting could potentially boost accuracy and generalization by leveraging multiple models.

- Deep Learning Approaches: Implementing neural network architectures, especially recurrent neural networks (RNNs) or long short-term memory networks (LSTMs), may uncover complex non-linear relationships in the data.

- Real-time Monitoring: Developing a real-time monitoring system that continuously evaluates transactions for potential fraud can provide immediate alerts and prevent fraudulent activities.

- Anomaly Detection Techniques: Combining traditional machine learning models with unsupervised anomaly detection algorithms can enhance the model's ability to detect previously unseen fraud patterns.

- Adaptive Learning Models: Implementing models that can adapt to evolving fraud tactics and adjust their detection strategies accordingly.

REFERENCES

- [1] Delamaire, Linda, Hussein Abdou, and John Pointon. "Credit card fraud and detection techniques: a review." *Banks and Bank systems* 4.2 (2009): 57-68.
- [2] Barker, Katherine J., Jackie D'amato, and Paul Sheridon. "Credit card fraud: awareness and prevention." *Journal of financial crime* 15.4 (2008): 398-410.
- [3] Ghosh, Sushmito, and Douglas L. Reilly. "Credit card fraud detection with a neural-network." *System Sciences*, 1994. Proceedings of the Twenty-Seventh Hawaii International Conference on. Vol. 3. IEEE, 1994.
- [4] Chaudhary, Khyati, Jyoti Yadav, and Bhawna Mallick. "A review of fraud detection techniques: Credit card." *International Journal of Computer Applications* 45.1 (2012): 39-44.
- [5] Bhatla, Tej Paul, Vikram Prabhu, and Amit Dua. "Understanding credit card frauds." *Cards business review* 1.6 (2003): 1-15.
- [6] Patidar, Raghavendra, and Lokesh Sharma. "Credit card fraud detection using neural network." *International Journal of Soft Computing and Engineering (IJSCE)* 1.32-38 (2011).
- [7] Srivastava, Abhinav, et al. "Credit card fraud detection using hidden Markov model." *IEEE Transactions on dependable and secure computing* 5.1 (2008): 37-48.
- [8] Zojaji, Zahra, Reza Ebrahimi Atani, and Amir Hassan Monadjemi. "A survey of credit card fraud detection techniques: Data and technique oriented perspective." *arXiv preprint arXiv:1611.06439* (2016).
- [9] Awoyemi, John O., Adebayo O. Adetunmbi, and Samuel A. Oluwadare. "Credit card fraud detection using machine learning techniques: A comparative analysis." *2017 international conference on computing networking and informatics (ICCNi)*. IEEE, 2017.
- [10] Shen, Aihua, Rencheng Tong, and Yaochen Deng. "Application of classification models on credit card fraud detection." *2007 International conference on service systems and service management*. IEEE, 2007.
- [11] Dal Pozzolo, Andrea, et al. "Credit card fraud detection: a realistic modeling and a novel learning strategy." *IEEE transactions on neural networks and learning systems* 29.8 (2017): 3784-3797.
- [12] Xuan, Shiyang, et al. "Random forest for credit card fraud detection." *2018 IEEE 15th international conference on networking, sensing and control (ICNSC)*. IEEE, 2018.
- [13] Dornadula, Vaishnavi Nath, and Sa Geetha. "Credit card fraud detection using machine learning algorithms." *Procedia computer science* 165 (2019): 631-641.
- [14] Maes, Sam, et al. "Credit card fraud detection using Bayesian and neural networks." *Proceedings of the 1st international naio congress on neuro fuzzy technologies*. Vol. 261. 2002.

- [15] Sailusha, Ruttala, et al. "Credit card fraud detection using machine learning." 2020 4th international conference on intelligent computing and control systems (ICICCS). IEEE, 2020.