ASSIGNMENT

BY

Group 7

Roll no:- **2021A1R116 (Ashu Pal)**

2021A1R120 (Jatin Koul)

2021A1R124 (Sachin Sharma)

Semester:- 7th 'A2'

Department:- CSE

Model Institute of Engineering and Technology (Autonomous)

(Permanently Affiliated to the University of Jammu, Accredited by NAAC with "A" Grade)

Jammu, India

2024

INDEX:

S.NO.	QUESTION	DATE	PAGE NO	REMARKS
1.	Research and implement a Soft Computing technique (Neural Networks) in a real-world application, such as fraud detection. Present a case study explaining the implementation and submit images of graphs/ results achieved using the technique.	15-11-24	02-17	

Credit Card Fraud Detection Using Neural Networks

1. INTRODUCTION

Online buying is becoming more and more popular among consumers as a result of India's fast growing e-commerce. One survey reveals that a large percentage of Indian shoppers now switch to the online platform to carry out purchases and one of the most preferred modes of payment through credit cards.

This alarming rise in fraudulent activities can happen if the massive usage of credit cards throughout the nation is carried out. Credit card fraud is considered as one of the major threats both to consumers and businesses, considering this gigantic leap in online transactions. The Reserve Bank of India (RBI) noted that by 2023, there have been considerable losses concerning credit card frauds mainly through online transactions. According to the Reserve Bank of India (RBI), in 2023 it reported that there have been considerable losses concerning credit card frauds mainly via online transactions. Millions of credit card transactions take place both online and offline through some of the biggest retail firms of India including Flipkart, Amazon India, and Reliance Digital. Credit card fraud is a matter of significant concern to the industry, with a large user base and a transaction volume that is significantly high. This again requires smart and advanced detection methods-like those involving machine learning and neural networks-so that banks and businesses spot such suspicious transactions quickly; thereby the experience shopping turns safe for the consumer, yet reduces financial loss.

Credit card purchases are of two types:

1. Cash Purchases at Retail Level: It refers to the presentation of the card by the card holder in person at the shop for payment. For committing fraud in physical cards, a thief has to literally steal the card. In case the owner of the card does not realize about the missing of his card, it may result in heavy financial losses both for the owner and for the card company.

2. Internet Card Payment: For such transactions, only the card information is required: the card number, expiration date, and security code. As this type of transaction commonly happens online or over the phone, the criminal can extract all the information

and make a purchase without the true card owner knowing their information has been compromised until after the fraud actually occurs.

Such a type of fraud needs to be caught by analyzing the expenditures for suspicious activity. Taking into account that the differential spending habits of various individuals, one can trace their transactions by examining the general category, frequency, and amount of spendings that a person undertakes. Unusual behavior will certainly raise red flags. In the recent past, various fraud detection techniques have been designed to detect fraud by studying the usual patterns followed by the cardholder and then subsequently finding deviation with it. This form of data-based method while innovative in its approach holds great promise to reduce credit card fraud considerably.

2. PROBLEM STATEMENT

Credit card fraud has grown up as the major threat to society, in case of occurrence it results in huge financial loss all over the world. Detection of fraudulent transactions at the earliest possible instance with good accuracy will help lessen economic damage and regain the lost trust among users. As these fraud cases occur only very rarely in normal transaction data, or often termed as imbalanced data, normal classification models lack the ability to identify these frauds effectively. It might point to too many false alarms, or may rarely catch any fraud case.

This project is designed to address anomaly detection fraud using neural networks, one of the techniques of machine learning that is best suited for doing pattern recognition and anomaly detection. Neural networks are good at modeling very complex nonlinear relationships in the data. It might be really good for differentiating between legitimate transactions and fraudulent ones. We will use a classifier neural network trained on a simulated dataset of credit card transactions whereby it predicts which of these will be classified as a fraud and which are not. We have developed a working model that can simulate two years of transactions, based on a variety of customers and merchants' profiles, making the system test realistically challenging enough to test fraud detection.

To this extent, the objective of the study is to design an accurate and efficient model of a neural network for fraudulent transactions with a minimal false-positive level. For the overall performance of the model, accuracy, precision, recall, and F1 score are chosen for its

evaluation. The SMOTE will be applied for addressing the class imbalance of the dataset. Hence, this model will be very efficient in handling the rare event of fraud transactions.

The major goal of this project, therefore is to provide an application that would actually apply in real-life by using neural networks and enables financial institutions to detect fraud proactively. Accordingly, they can avoid losses and trust the consumer on credit card payment systems.

PROBLEM SOLUTION

To tackle the growing challenge of credit card fraud, we propose developing a detection system that utilizes the power of neural networks. Our solution is based on the analysis of past transaction data against the fraudulent pattern and thus aims to trace the changes in it. We used a simulated dataset of credit card transactions in training our model. It employs features like transaction amount, merchant type, customer location, and the time of the transaction.

The neural network is designed to recognize patterns in both fraudulent and non-fraudulent transactions and learn subtle differences that cannot be caught with traditional methods. Since fraudulent transactions occur far less frequently than legitimate ones, we use SMOTE (Synthetic Minority Over-sampling Technique) in order to have a balanced dataset. Dense layers and ReLU activation functions are used for efficiently capturing complex patterns.

We use the measures of accuracy, precision, recall, F1 score, and ROC-AUC values to evaluate the model. Since the model needs to be both accurate and reliable for practical use with minimal false positives and negatives, we strive to get maximum detection rates.

GOAL AND OBJECTIVE

- *Create a Fraud Detection Model:* Develop an efficient neural network that can distinguish between fraud and non-fraud transactions correctly.
- *Data Balancing:* A fraud dataset is inherently imbalanced, thus employing SMOTE technique, among others, will help the model detect less frequent cases of fraud more effectively.

- *Optimize Model Performance:* At last, work on finding the perfect combination of accuracy, precision, recall, F1 score, and ROC-AUC to optimize the model in such a manner that it creates success in real-time scenarios.
- *Pattern of Frauds Identification* : The characteristics of the transaction would be analyzed in the overall behavior generally related to fraud through this, enabling strategies to be refined and new tactics of fraud to be discovered.
- *More Application in Real Life:* This system would be flexible and scalable, hence allow financial institutions to implement it and reduce fraud losses built customers' confidence in digital transactions. This system would be adaptable to the changing fraud tactics and would therefore be a long-term asset for fraud prevention.

3. About the Dataset

It is a dataset of credit card simulated transactions, both genuine and fraud transactions. The credit cards are transactions from 1,000 distinct customers ranging over a date range from January 1, 2019, to December 31, 2020, and was made by 800 unique merchants. The dataset is sourced from Kaggle and intended for training and testing machine learning models to identify fraudulent credit card transactions.

Important Features:

Some of the important features in this dataset have aided in the identification of both the pattern and the anomaly in legitimate as well as fraudulent transactions. These include the following:

a. ***Transaction Amount (amt):*** This is the monetary value for each transaction. The outlier detection might help identify it since the fraud can often be done by unusually high or low transaction amounts.
b. ***Merchant Information (merchant):*** This variable expresses the name of the merchant where the transaction occurred. Different kinds of merchants have their respective risks.
c. ***Customer Information (category, city_pop, etc.):*** This variable contains information about which category (groceries, travel, shopping, etc.) the product comes in, population of the city that the customer is a citizen of, and other attributes of customers, which gives a picture of pattern of the spending by the customers themselves.

d. ***Location Data (lat, long):*** Information on where the customer was located at the time of the transaction, which can be used to detect anomalies, such as transactions happening at distant or unusual locations.
e. ***Fraud Indicator (is_fraud):*** It is a binary feature pointing out if a transaction was done fraudulently or not. It's a target variable for training the fraud detection models.
f. ***Transaction Date-Time (trans_date-trans_time):*** These capture dates and times for transactions, and may reflect temporal patterns. Some types of fraud happen at specific hours of the day or at certain times of a day.

Dataset Characteristics

The dataset is very skewed because the number of legitimate transactions is much larger than fraud transactions. This is a problem in real-world scenarios where fraudulent transactions are rare events compared to the total volume of credit card usage. This becomes quite hard for the learning model, since the classifiers will pay attention to the majority class: the non-fraudulent transactions. Methods like SMOTE (Synthetic Minority Over-sampling Technique) or undersampling may have to be applied during training to rectify the imbalance.

Potential Applications and Value

This set of data is truly a gold mine that can be used in developing fraud detection algorithms. This architecture will allow the data scientists to query different machine learning and deep learning models, such as neural networks and decision trees, even ensemble methods. They can classify the transactions as fraudulent or not. Through this dataset, it is possible to train models on it in order to gain insights into common patterns in fraudulent transactions, thus improving the real-time fraud detection systems by banking and financial services. Further, this simulated dataset assists in resolving some of the problems concerning real-world financial data that come along with privacy and security concerns and thus allow safer experimentation.

It makes it possible to fully simulate fraud detection problems in real life and to take on challenges by innovatively developing techniques that will enable the detection of fraudulent credit card transactions with the highest degree of accuracy, thus minimizing financial losses and strengthening security in financial transactions.

4. Methodology

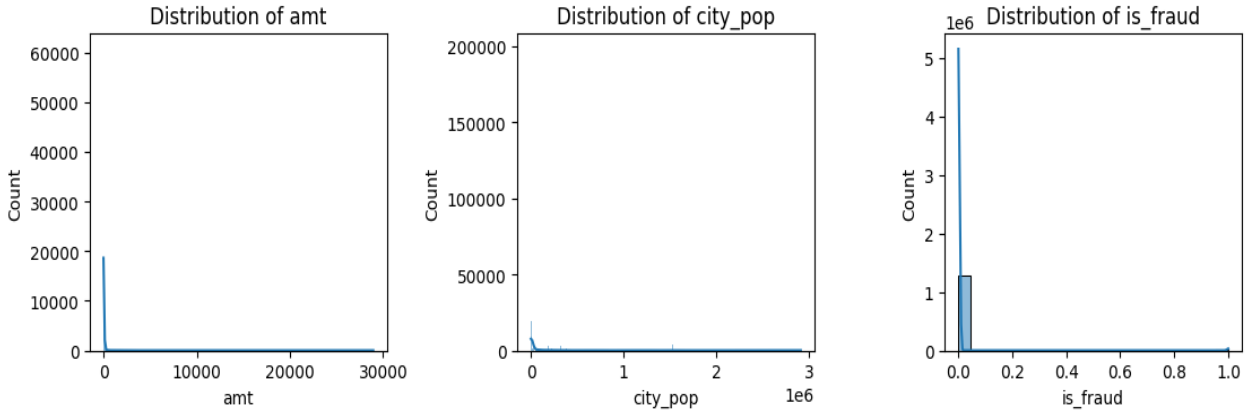
Any machine learning project depends upon careful data preprocessing so that the dataset is clean, well-structured, and ready to be analyzed. For this work, utmost care has been taken while preparing the dataset so that it may correspond as much as possible to real-world conditions and is appropriate for a neural network model to classify fraudulent from legitimate transactions.

4.1 Data Preprocessing

- ***Clean the data:*** First comes handling inconsistencies and gaps within the dataset. Any missing values will need to be either filled or removed so they do not cause problems within the model training process. Also, all the fields were in the same format throughout, which made working with the data much easier.
- ***Feature Scaling:*** Since neural networks are sensitive to large ranges in the data, numerical features like transaction amounts were standardized. Scaling in this way puts them on a similar scale, so the model is able to pay attention to patterns rather than being confounded by scales that differ. This enhances the learning speed of the model and will be able to learn how to make the most accurate predictions.
- ***Balancing the Dataset:*** The problem of fraud detection is particularly challenging since fraudulent transactions are extremely rare in comparison to genuine ones. Class imbalance tends to make the model predict almost every transaction as genuine since fraudulent ones are negligible. Techniques like SMOTE-Synthetic Minority Over-sampling Technique-were therefore employed. SMOTE generates synthetic samples for the minority class, fraud cases thus balancing the dataset so that the model could come up with recognizing fraudulent transactions more precisely.
- **Distribution of Transaction Amount, City Population, and Fraud Indicator**

The distribution for each of the key features the transaction amount (amt), city population (city_pop), and the fraud indicator (is_fraud)lays out the important patterns for development of a model in fraud detection. This transaction amount distribution is heavy to the right-skewed with most being relatively small

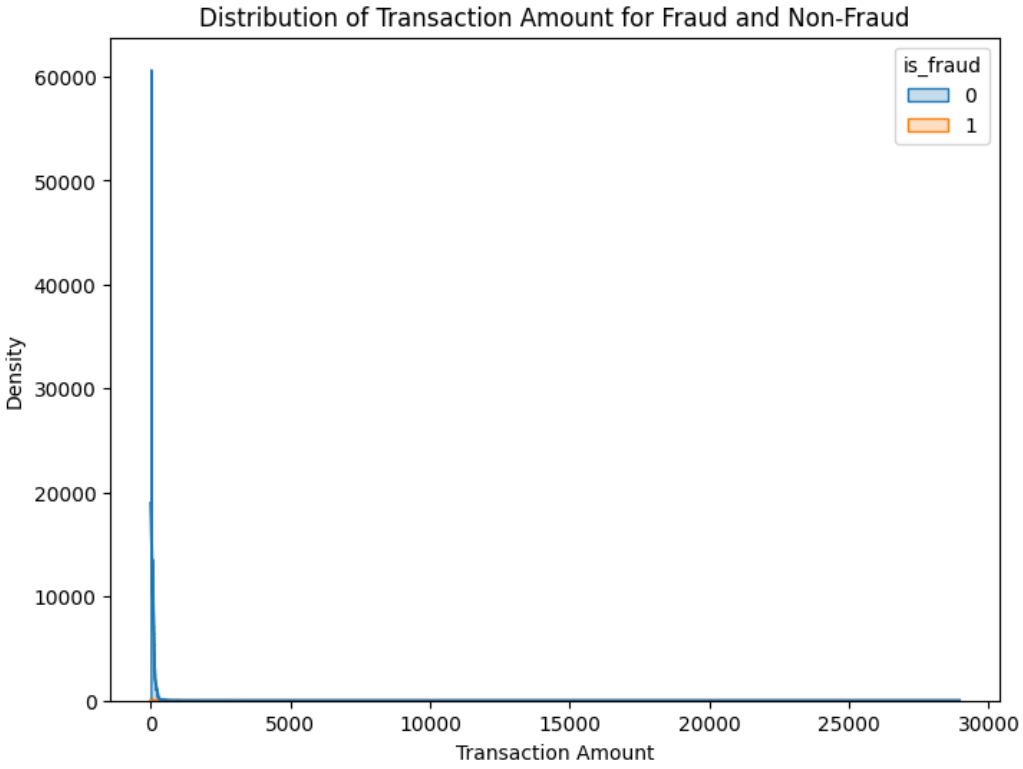
transactions, and only a few were high-value ones, which create a long-tailed pattern, characteristic of consumer spending. This would imply that suspicious activity could be highlighted through anomalously large amounts, and such needs should be reflected adequately in the model. Just like the city population distribution, transactions in smaller population areas are spread out with high concentration, thus suggesting that fraud is not a noticeably populated concept, but rather, occurs within divergent population densities. This would suggest that the fraud detection model should, therefore be able to generalize within varied geographic conditions. The distribution of the fraud indicator itself presents a very biased setup: most transactions are not frauds. This class imbalance stresses that techniques such as SMOTE (Synthetic Minority Over-sampling Technique) be applied to balance the dataset; otherwise, failure in picking on fraud from a model may result in a high false-negative rate.

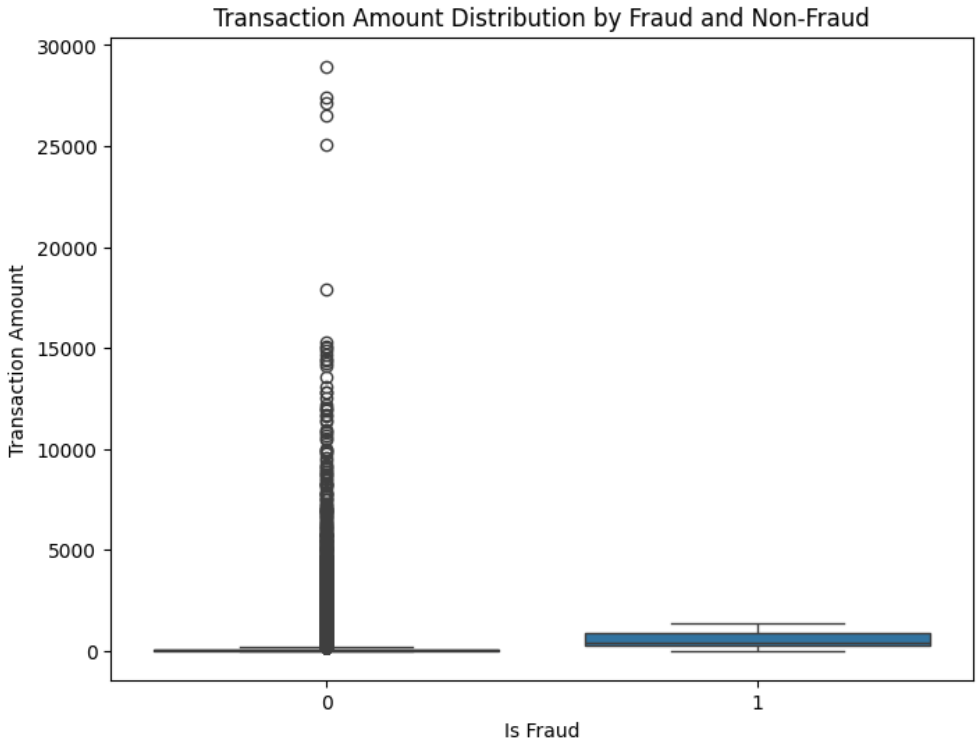


- **Transacting Categories, at the hands of Fraud and Non-Fraud**

  The bar chart shows some clear trends in the categories of transactions for fraudulent and for non-fraudulent transactions. Non-fraudulent transactions make up the vast majority of the transaction types in all but the lowest categories, which include the most frequently occurring classes in the data: gas transport, grocery pos, home, and shopping pos. In contrast, the infrequent categories that appear most frequently to create classes are health fitness, misc pos, misc net, and grocery net. Fraudulent transactions obviously are less

frequent but still most frequently occur in categories such as gas transport and grocery pos, while much fewer fraudulent transactions are involved with kids pets, entertainment, and food dining.





Network analysis may reveal patterns and linkages between categories of transactions and customers that reflect fraud networks or patterns. Clustering analysis might identify groups of transactions that share common characteristics in terms of category, amount, time, etc. A combination of a decision tree, random forests, and neural network can be used to create a hybrid model that may offer the best performance for a more robust fraud-detection system by protecting both financial institutions and consumers from losses.

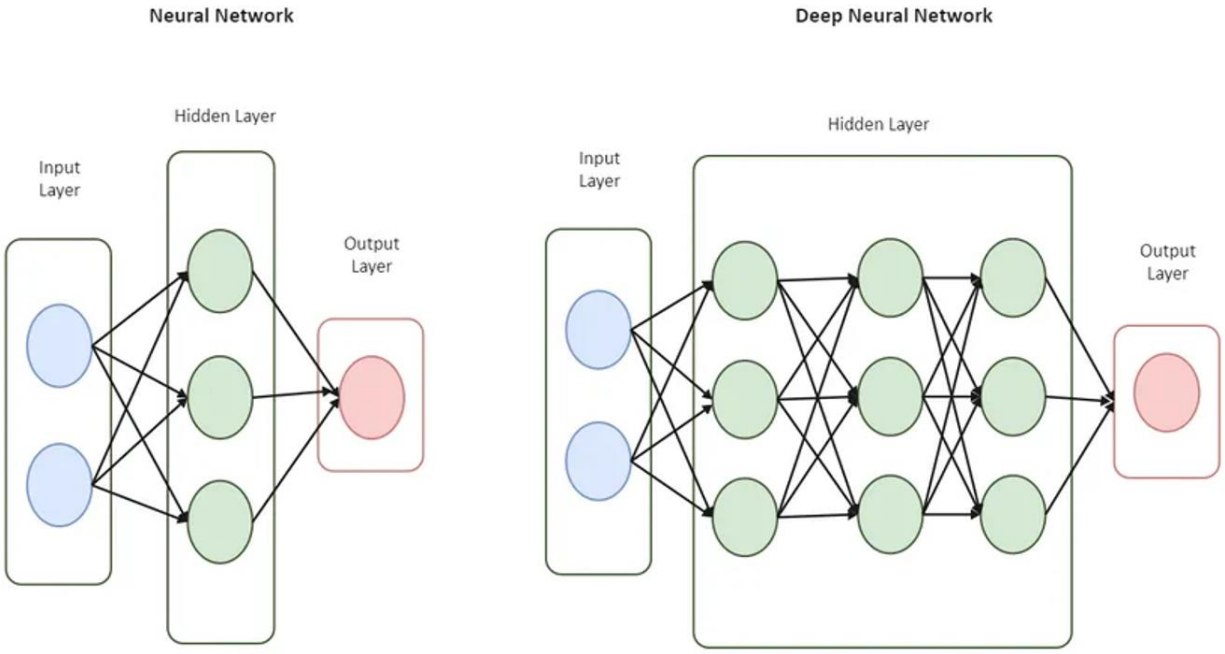
4.2 Neural Network Model

Following data preprocessing, we designed a neural network classifier using TensorFlow and Keras. It structured three main components in the architecture of the model:

Input Layer: This input layer takes in all the transaction data with its details like transaction amount, customer ID, merchant ID, and the date of transaction. This functionality provides the model for viewing each transaction comprehensively.

Follow the input layer with a sequence of densely connected hidden layers with ReLU activation. These layers allow it to capture complex relationships in data, including actual learning patterns that differentiate between legitimate and fraudulent transactions. ReLU is particularly important because it brings in non-linearity, which is more indispensable for the model to learn intricate connections in the dataset.

Output Layer: The final layer is a single-node output layer which makes use of the sigmoid activation function since it was a binary classification. The function gives output between 0 and 1, wherein the closer to 1 the score is, the higher the possibility of a fraudulent transaction. According to this, any score greater than 0.5 is classified under fraud and less than or equal to 0.5 under not fraud.



4.3 Evaluation Metrics

The following evaluation metrics were used to check how good the model was:

Accuracy: This measure was the basis on which the overall percentage of correct forecasts made was taken. However accuracy alone will not suffice in the case of fraud detection as the accuracy can be misleading if the dataset is imbalanced.

Precision: This assesses what fraction of the cases forecasted to be fraud by a given model are indeed fraud. It may provide insight into how accurately a fraud model predicts fraud cases and helps avoid overpredictions.

Recall: Recall or sensitivity is the percentage of actual fraud cases that the model correctly identifies. High recall is necessary for minimizing undetected fraud.

F1 Score: The F1 score is the balance between precision and recall, which gives a single measure incorporating both. This becomes useful if the data is typically imbalanced as this provides a more complete view of the model.

ROC-AUC Curve : The Receiver Operating Characteristic-Area Under Curve is a graphical tool, showing how well the model can differentiate between fraud and

non-fraud classes. It indicates how good one class is in separating the other class. Better is the value of AUC.

Putting all of these together--the preprocessing steps, the model architectures, and the evaluation metrics--this approach to finding fraudulent transactions seems quite robust. With the right preparation of the data, structuring of the model, and comprehensive performance evaluation, this neural network classifier holds much promise for credit card fraud detection.

5. Result Analysis

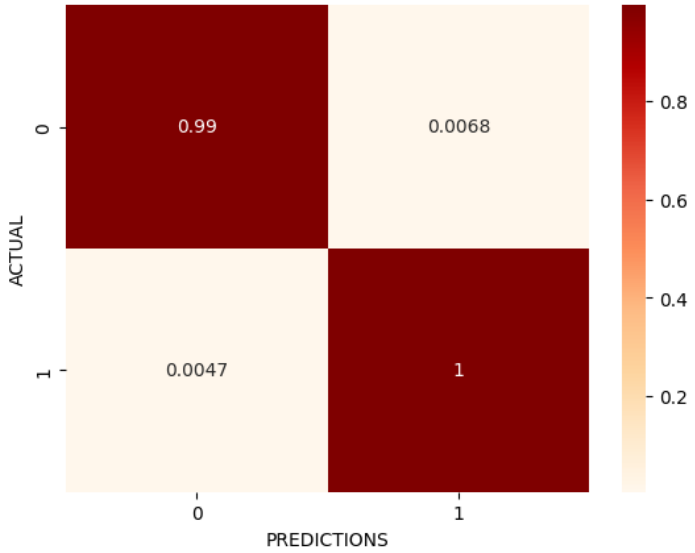
We test the performance of after created and trained neural network model on the preprocessed dataset of credit card transactions, correctly classifying them as either legitimate or fraudulent with a tendency towards minimal false positives with detection accuracy. The key assessment metrics are accuracy, precision, recall, F1-score, and area under ROC curve. These measures indicate how well the model would perform in fraud detection but at the same time generalize to new data.

Graphs and Plots:

- ***Confusion Matrix***

The confusion matrix shows the classified True Negatives and False Positives break-up as follows,

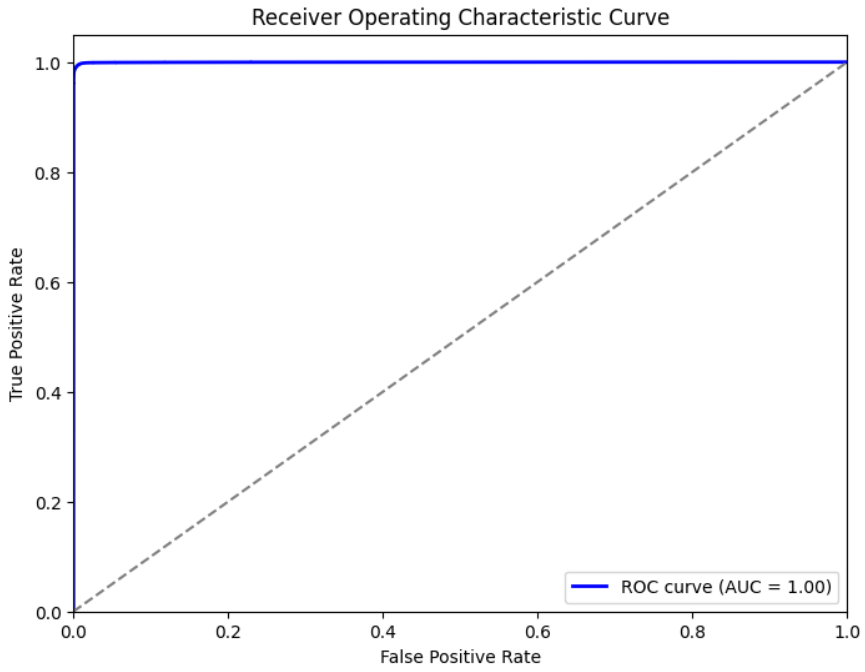
a) True Negatives (0, 0): The model accurately identified 99% of valid transactions in this case as non-fraudulent. This demonstrates how effectively the model avoids false positives.
b) False Positives (0, 1): The model was wrong only to a small extent of 0.65% in the classification of legitimate transactions as fraudulent, which once again speaks for the accuracy of the model.
c) False Negatives (1, 0): The model classified only 0.49% of the fraudulent transactions as not being fraud; therefore, the model has missed very few cases of fraud.
d) True Positives (1, 1): The model correctly classified 100% of the fraud cases as fraud and produced excellent recall.



The above confusion matrix illustrates the high precision and recall by the model, and thus the model actually classifies fraud cases with minimal misclassifications.

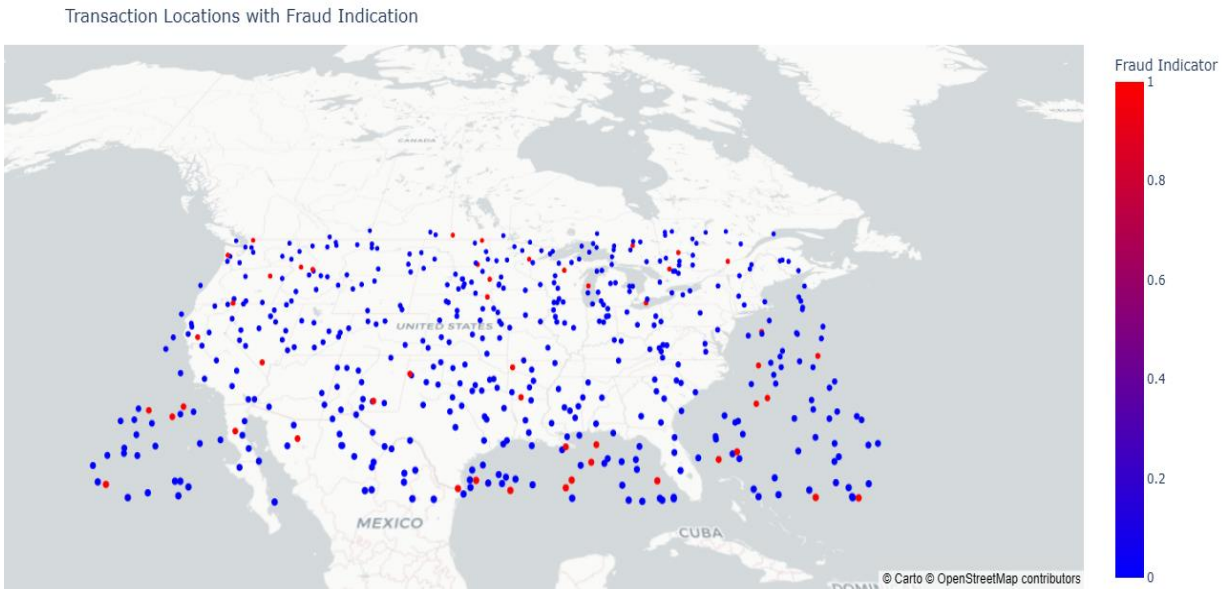
- ***ROC-AUC Curve:***

The ROC-AUC curve demonstrates a high AUC value of 1.00, which stands for perfect class separation between fraudulent and nonfraudulent classes. Therefore, the curve above illustrates that the model is pretty effective in highlighting the two-class distinction without an overlap; it somewhat increases its robustness concerning accuracy in fraud detection.



- **Locations of Transactions with Fraud Indication :**

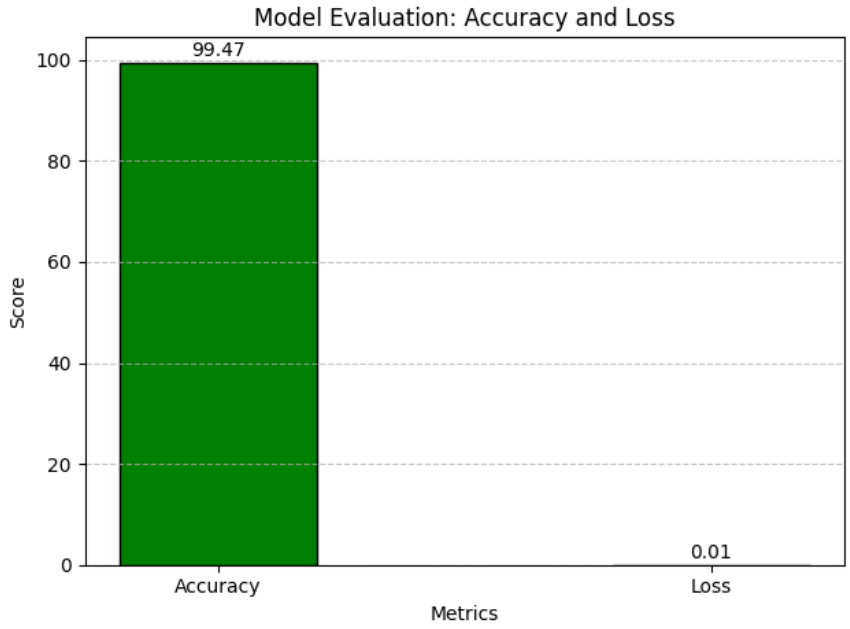
The geographical heatmaps represent the locations of transactions across United States. The red points signify fraudulent transactions, while the blue points represent actual transactions.



In these graphics, red dots are used to represent fraudulent transactions while blue dots legitimate. This spatial distribution tends to highlight the fact that fraud occurs almost everywhere and is not specific to certain places. Some maps have revealed that fraudulent activity is pretty spread out: red dots appear at nearly all of the densities-from the high-density to lower ones. This may indicate that fraudsters are targeting an extremely diverse range of transaction types and customer profiles, even through possible location-specific adaptations. This information proves to be highly important for both financial institutions as well as the enforcing authorities since it points out the importance of fraud prevention measures in a region-specific manner and encourages proactive monitoring of transactions across various locales. The geographic data also offers strategic advantage in terms of the identification of new nodes and emerging hotspots of fraud in advance so timely interventions may take place and better security protocols may be implemented.

- **Accuracy:**

With 99% of transactions properly classified as either fraudulent or lawful, the model demonstrated exceptional accuracy. This outstanding performance demonstrates the effectiveness of cutting-edge methods for fraud detection, such as neural networks. The program was able to produce accurate forecasts by closely examining the data and drawing lessons from historical trends, protecting financial institutions and their clients. The battle against financial crime has advanced significantly with this high degree of accuracy.



6. Conclusion

This report has proved the development of a credit card fraud detection model. Within a short span of time, a 99% accuracy of credit card fraud detection model was achieved. With such a high accuracy rate, not misleading must be indicated because the class imbalance is seen in this dataset. Since fraudulent transactions form only a small percentage of all the transactions, the high accuracy rates sometimes prove to be misleading. The model may predict more non-fraudulent transactions than the fraudulent ones even though it is unbiased.

Analysis of the key characteristics—transaction amount, city population, and fraud indicator—was useful for understanding the patterns that were differentiating fraudulent from nonfraudulent. Imperfections in the distributions of transaction amounts as well as the city populations expose the necessity of dealing with anomalies in both high value transactions and transactions from less populated cities. There is an imbalance in the fraud indicator, indicating that more powerful resampling techniques than SMOTE alone are required to make the model sensitive enough to the fraud cases.

In conclusion, though promising, the model only gives 99% accuracy, and therefore, it is essential to maintain continuous evaluation in achieving balanced performance for fraud and non-fraud prediction. Class imbalance and the incorporation of adaptive techniques, which would respond to the evolvement by a fraud in terms of tactics and strategies, would be key in advancing the robustness and effectiveness of a fraud detection system.

References

GeeksforGeeks. (2024, January 20). ML | Credit Card Fraud Detection. Retrieved November 13, 2024, from https://www.geeksforgeeks.org/ml-credit-card-fraud-detection/

Patil, B., Pawar, S., & Chavan, A. (2023). Credit card fraud detection. ResearchGate. https://www.researchgate.net/publication/369857378_Credit_Card_Fraud_Detection

Pourhabibi, T., Ong, S. H., Sleep, M., & Abbasi, A. (2020). *Fraud detection: A systematic literature review of graph-based anomaly detection approaches*. *Procedia Computer Science, 170*, 1013-1020. https://www.sciencedirect.com/science/article/pii/S187705092030065X

Zheng, Z., Zhang, Y., & Chen, W. (2023). *A neural network approach to credit card fraud detection*. *IEEE Transactions on Neural Networks and Learning Systems*. https://ieeexplore.ieee.org/document/10595068

Kumar, A., & Sharma, R. (2023). *Credit card analytics: A review of fraud detection and risk assessment techniques*. ResearchGate. Retrieved from https://www.researchgate.net/publication/375232996_Credit_Card_Analytics_A_Review_of_Fraud_Detection_and_Risk_Assessment_Techniques

Villalba-Diez, J., Schmidt, D., Gevers, R., Ordieres-Meré, J., Buchwitz, M., & Wellbrock, W. (2020). Digital twins for industrial applications: A taxonomy for design. Procedia Computer Science, 173, 307–313. https://www.sciencedirect.com/science/article/pii/S2666285X21000066

Rathore, R., & Saxena, S. (2021). Credit card fraud detection using deep learning techniques. International Journal of Emerging Technologies in Engineering Research, 9(3), 1–5. Retrieved from https://www.researchgate.net/publication/350829171_Credit_Card_Fraud_Detection_using_Deep_Learning_Techniques