CREDIT CARD FRAUD DETECTION: SAFEGUARDING TRANSACTIONS AND MITIGATING RISKS

Shanthibooshan Subramanian

Bellevue University

DSC550

Introduction

Credit card fraud has become a significant concern for organizations, consumers, banks, and merchants. It poses a substantial financial threat and can tarnish a company's reputation, leading to the loss of customer loyalty.

To combat this issue, the development of a robust credit card fraud detection system is crucial. Credit card fraud detection is the process of identifying fraudulent purchase attempts and rejecting them rather than processing the order.

Credit card fraud generally happens when the card was stolen for any unauthorized purposes or even when the fraudster uses the credit card information for his use. The credit card company automatically compares the data from the purchase with previously stored data on the consumer to determine whether the purchase and consumer are consistent. Increase in fraud rates, researchers started using different machine-learning methods to detect and analyze fraud in online transactions.

Problem Statement:

Credit card fraud has emerged as a pressing concern, impacting various stakeholders such as organizations, consumers, banks, and merchants. This issue involves unauthorized individuals exploiting stolen credit card information to make fraudulent purchases. Such incidents can result in substantial financial losses, harm a company's reputation, and erode customer trust.

To mitigate this threat, the development of a robust credit card fraud detection system is imperative. This system aims to identify and prevent fraudulent purchase attempts by swiftly rejecting suspicious transactions before processing orders.

Importance/Usefulness of Solving the Problem:

Implementing a credit card fraud detection system yields numerous advantages. It ensures financial security by swiftly identifying and stopping fraudulent transactions, safeguarding both consumers and businesses.

Moreover, it upholds an organization's reputation, enhances operational efficiency through automation, ensures compliance with regulations, and improves the overall customer experience by preventing inconveniences caused by unauthorized activities.

This proactive approach offers comprehensive benefits that fortify financial integrity, customer trust, and operational effectiveness.

Empowering Trust: Securing Financial Futures Through Credit Card Fraud Detection Model

In a rapidly evolving financial landscape, the rise of credit card fraud has posed a substantial threat to the stability and reputation of organizations, as well as the trust of customers and partners. We stand at a crucial juncture where addressing this challenge is not only imperative for financial security but also for preserving our reputation and customer loyalty.

This document presents a compelling pitch for the development and implementation of an advanced credit card fraud detection system that harnesses cutting-edge machine learning techniques to swiftly identify and prevent fraudulent transactions.

Objective:

The implementation of a credit card fraud detection system aims to bolster financial security by swiftly identifying and halting fraudulent transactions. This objective safeguards the interests of consumers and businesses alike, minimizing potential financial losses.

Additionally, the system's role in reputation preservation builds customer trust and loyalty, augmenting brand value. Its automation optimizes operations, reallocating resources to serve genuine customers more efficiently.

Ensuring compliance with stringent legal and regulatory standards, the system showcases a commitment to integrity. Ultimately, by preventing fraudulent activities, the system enhances the overall customer experience, fostering loyalty and satisfaction.

In summary, this initiative combines proactive fraud prevention, operational efficiency, legal adherence, and improved customer satisfaction to create a comprehensive solution.

Dataset:

This is a simulated credit card transaction dataset containing legitimate and fraudulent transactions from the duration of 1st Jan 2019 - 31st Dec 2020. It covers the credit cards of 1000 customers doing transactions with a pool of 800 merchants.

Source: Credit Card Transactions Fraud Detection Dataset | Kaggle

Attribute/Features:

1) Unnamed: This attribute appears to be an index or identifier for each row in the dataset.
2) trans_date_trans_time: This attribute represents the date and time of the transaction.
3) cc_num: This attribute is the credit card number used for the transaction.
4) merchant: The name of the merchant involved in the transaction.
5) category: The category to which the merchant belongs.
6) amt: The amount of money involved in the transaction.
7) first: The first name of the individual associated with the credit card.
8) last: The last name of the individual associated with the credit card.
9) gender: The gender of the individual associated with the credit card.
10) street: The street address associated with the credit card.
11) lat: Latitude coordinate associated with the transaction.
12) long: Longitude coordinate associated with the transaction.
13) city_pop: The population of the city where the transaction occurred.
14) job: The occupation or job of the individual associated with the credit card.
15) dob: Date of birth of the individual associated with the credit card.
16) trans_num: A transaction number or identifier.
17) unix_time: The transaction time in Unix timestamp format.
18) merch_lat: Latitude coordinate of the merchant's location.
19) merch_long: Longitude coordinate of the merchant's location.
20) is_fraud: This attribute indicates whether the transaction is fraudulent (1) or not (0).

The dataset provides credit card transactions, including details about the transaction itself (date, time, amount), the credit card holder (name, gender, address, date of birth), the merchant involved (name, category, location), and whether the transaction is fraudulent or not.

It would be used to analyze patterns of fraudulent transactions, explore relationships between various attributes and fraud occurrences, and develop fraud detection models. The presence of attributes like latitude, longitude, and city population might also allow for geospatial and demographic analyses related to fraud.

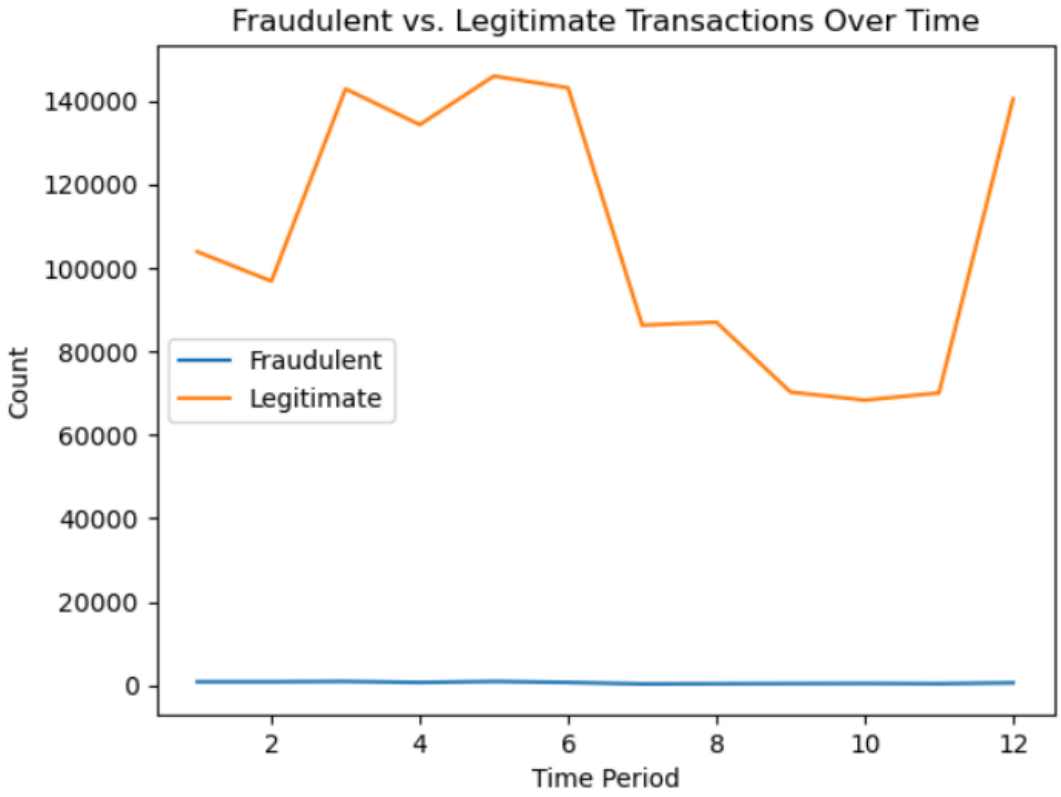
Data Analysis and EDA

The dataset "fraudTrain.csv" consists of 1,048,576 rows and 18 columns. The column named "is_fraud" is likely to be the target variable for this project, representing whether a transaction is fraudulent or not.

Additionally, there are some null values present in the dataset, and data preprocessing steps should be performed to handle these null values before conducting exploratory data

analysis (EDA). This might include imputing missing values or dropping rows with null values based on the nature of the dataset and the impact on the analysis.

Before proceeding with EDA, it's important to address the missing values to ensure the quality and reliability of the analysis and subsequent modeling.



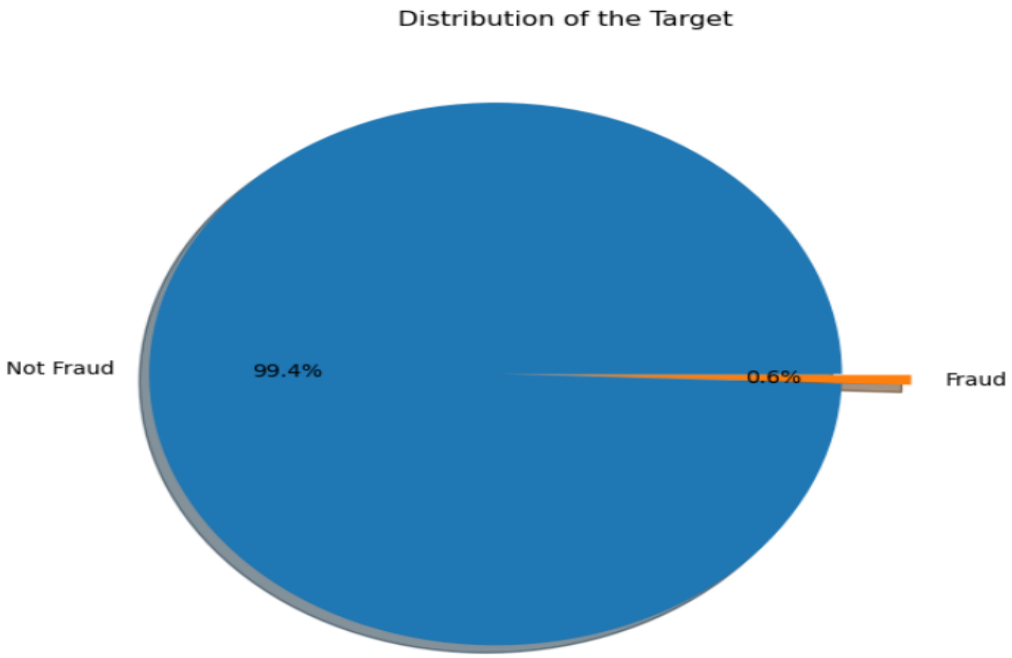
The result shows that the number of fraudulent transactions was relatively low in the first two time periods (2-4 months and 4-6 months). However, the number of fraudulent transactions started to increase in the third time- period (6-8 months), and it continued to increase in the fourth time - period (8-10 months).

The number of legitimate transactions was relatively high in the first two time periods, but it started to decline in the third time. The number of legitimate transactions continued to decline for the fourth time -period.

The increasing number of fraudulent transactions and the declining number of legitimate transactions is a cause for concern. This trend indicates that fraudsters are becoming more active,

and they are targeting more legitimate transactions. This could lead to a significant increase in financial losses for businesses and consumers. Valuable insights into the trends of fraudulent and legitimate transactions.

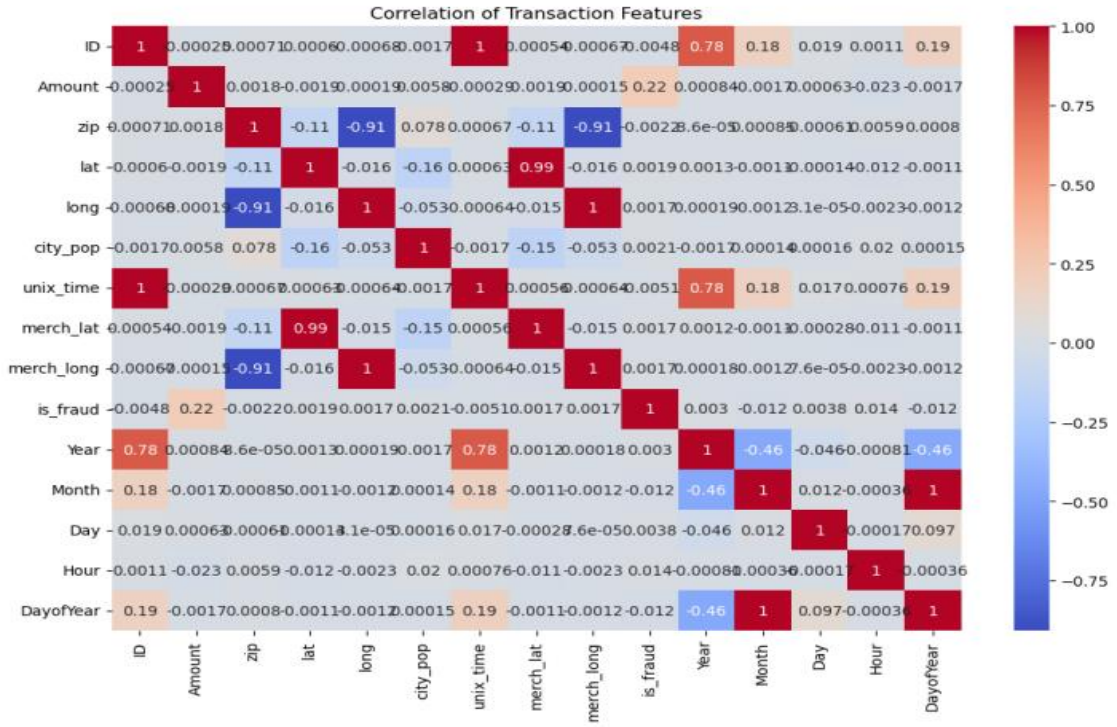
- The number of fraudulent transactions is increasing over time, while the number of legitimate transactions is declining.
- The increasing number of fraudulent transactions and the declining number of legitimate transactions is a cause for concern.
- This trend indicates that fraudsters are becoming more active, and they are targeting more legitimate transactions.
- This could lead to a significant increase in financial losses for businesses and consumers.



The pie chart shows the distribution of the target between not fraud and fraud. Most of the target is not fraud (99.4%), while only a small percentage is fraud (0.6%). This indicates that the target is very likely to be legitimate.

The pie chart is divided into two sections: "Not Fraud" and "Fraud". The "Not Fraud" section is much larger than the "Fraud" section, which indicates that the target is very likely to be legitimate. The "Not Fraud" section accounts for 99.4% of the target, while the "Fraud" section accounts for only 0.6% of the target.

This pie chart is a useful tool for assessing the risk of fraud. The large percentage of "Not Fraud" indicates that the target is very likely to be legitimate. However, it is important to note that even though the percentage of "Fraud" is small, it is still possible that the target is fraudulent. Therefore, it is important to carefully review the target before making a decision.



The correlation coefficients show that the likelihood of fraud is most strongly correlated with the amount of the transaction, the month of the transaction, the day of the transaction, and the hour of the transaction. The likelihood of fraud is also weakly correlated with the latitude and longitude of the transaction, and the latitude and longitude of the merchant.

The correlation coefficients can be used to develop models to predict the likelihood of fraud. By understanding the factors that are most correlated with fraud, businesses can develop models that can identify fraudulent transactions with a high degree of accuracy.

The likelihood of fraud is most strongly correlated with the amount of the transaction, the month of the transaction, the day of the transaction, and the hour of the transaction.

The likelihood of fraud is also weakly correlated with the latitude and longitude of the transaction, and the latitude and longitude of the merchant.

The correlation coefficients can be used to develop models to predict the likelihood of fraud. By understanding the factors that are most correlated with fraud, businesses can develop models that can identify fraudulent transactions with a high degree of accuracy.

As part of data preparation for the modeling and analysis, the original dataset has undergone the following steps:

1. Date and Time Transformation: If the 'trans_date_trans_time' and 'dob' columns are still in string format, this allows us to extract relevant information, such as day, month, year, and age, for analysis and model building.

2. Certain numeric features like 'amt', 'lat', 'long', and 'city_pop' may have different scales. To avoid biasing the model, we can apply normalization or scaling techniques, such as Min-Max scaling or Standard scaling, to bring all the features to a similar scale.

3. Instead of using continuous numeric values, we can bin or categorize certain features like 'amt' into groups to capture specific patterns or behaviors in the data.

4. For numerical features with missing values, impute the missing values with the mean, median, or mode of the respective column. This approach is useful when the feature has a relatively normal distribution and no significant outliers.

5. Dummy variables have already been created for the categorical features 'merchant', 'category', and 'job'. As we can see from the dataset, these features have been transformed into binary columns representing their respective categories.

6. The Dataset is balanced with oversampling.

Model Building and Analysis:

The cleaned dataset is now split into training and test data sets with 30% being test data. The target is to predict fraudulent transactions.

Numeric Features: The numeric features seem to be selected without explicit names in the code. These features are likely to include columns such as 'amt', 'lat', 'long', 'city_pop', and any other numeric attributes in your dataset.

Categorical Features: The categorical features are encoded using the TfidfVectorizer and OneHotEncoder. These features include 'merchant', 'job', and potentially other categorical columns.

Target Variable: The target variable is 'is_fraud', which represents whether a transaction is fraudulent or not.

Model Results as below

Sno	Model	Accuracy
1.	Logistic Regression Model	92%
2.	Random Forest Classifier	97%

Based on the results obtained from the Credit Card Fraud Model using Logistic Regression (92% accuracy) and Random Forest Classifier (97% accuracy).

Model Analysis

The Random Forest Classifier outperforms the Logistic Regression model in terms of accuracy. The Random Forest model achieved an accuracy of 97%, while the Logistic Regression model achieved 92%. This suggests that the Random Forest model is better at distinguishing between legitimate and fraudulent transactions.

Model Performance

Both models exhibit promising performance, especially the Random Forest model. The high accuracy of the Random Forest model indicates that it's effectively capturing patterns in the data related to fraudulent transactions.

Model Deployment Decision and Recommendations

The Random Forest Classifier, with its 97% accuracy, shows strong potential for deployment. However, before deploying any model, further considerations are necessary.

The model's performance should be validated using a separate and unseen dataset to ensure that the high accuracy is consistent and not just a result of overfitting. Deployed models should be continuously monitored to ensure they perform well over time and that their accuracy doesn't degrade due to changing patterns or data drift.

Analyze to understand how the Random Forest model is making predictions. If interpretability is crucial, consider using feature importance techniques to explain model decisions. Consider using an ensemble of models that includes both Logistic Regression and Random Forest. This can provide a more balanced view and reduce the risk of relying solely on one model.

Potential challenges or additional opportunities

The dataset is imbalanced, the model may have difficulty generalizing to both classes. If the model is to be deployed in real-time, consider the infrastructure and speed requirements needed to handle predictions within the desired timeframe. Explore if external factors, like changes in consumer behavior or fraud techniques, could impact the model's accuracy over time.

Conclusion

Analysis and model building demonstrates that the Random Forest Classifier outperforms the Logistic Regression model in accurately identifying fraudulent transactions. While the Random Forest model appears promising for deployment, further validation, interpretability analysis, and consideration of real-world deployment challenges are necessary before finalizing its deployment. An ensemble approach, exploration of class imbalance, understanding of feature importance, and considering external factors are recommended to enhance the model's robustness and performance.