APPLIED DATA SCIENCE

elangi

gokul

VVCET

PROBLEM STATEMENT

The problem is to develop a machine learning-based system for real-time credit card fraud detection.

The goal is to create a solution that can accurately identify fraudulent transactions while minimizing false positives.

This project involves data preprocessing, feature engineering, model selection, training, and evaluation to create a robust fraud detection systems

INTRODUCTION

Machine learning algorithms play a crucial role i dentifying fraudulent transactions by analyzing patterns a nd anomalies in credit card data.

Commonly used machine learning algorithms for credit card fraud detection include logistic regression, deci sion trees, random forests, and neural networks.

DESIGN THINKING PROCESS

Data Collection :

Obtain a dataset of credit card transactions, which should include both legitimate and fraudulent transactions.

Data Preprocessing :

Clean and preprocess the data. This involves handling missing values, outlier detection, and potentially scaling or normalizing features

Feature Engineering :

Create meaningful features from the transaction data. Feature engineering is crucial for improving the model's ability to detect fraud.

Model Selection:

Choose appropriate machine learning algorithms for fraud detection. Common choices include logistic regression, decision trees, random forests, or more advanced methods like deep learning.

Training:

Split your data into training and testing sets. Train your chosen models on the training data.

Hyper parameter Tuning:

Fine-tune your model's hyper parameters to optimize its performance.

Monitoring and Updates:

Continuously monitor the model's performance in the real-time system and update it as needed to adapt to changing fraud patterns.

Real-Time Implementation:

Deploy your model in a real-time environment where it can process credit card transactions as they occur. Consider using technologies like streaming data processing.

PHASE OF DEVELOPMENT

Exploratory Data Analysis (EDA):

Perform data visualization and statistical analysis to understand the characteristics of the data. Identify patterns and trends that could be indicative of fraud

Threshold Selection:

Determine an appropriate threshold for fraud detection to balance false positives and false negatives based on business requirements.

Monitoring and Alerts:

Continuously monitor the model's performance in real-time.

Implement alerting systems to notify relevant personnel of suspicious transactions.

Documentation and Reporting:

- Document the entire development process, including data sources, preprocessing, model selection, and deployment procedures.
- Provide regular reports on model performance, false positive rates, and fraud detection rates.

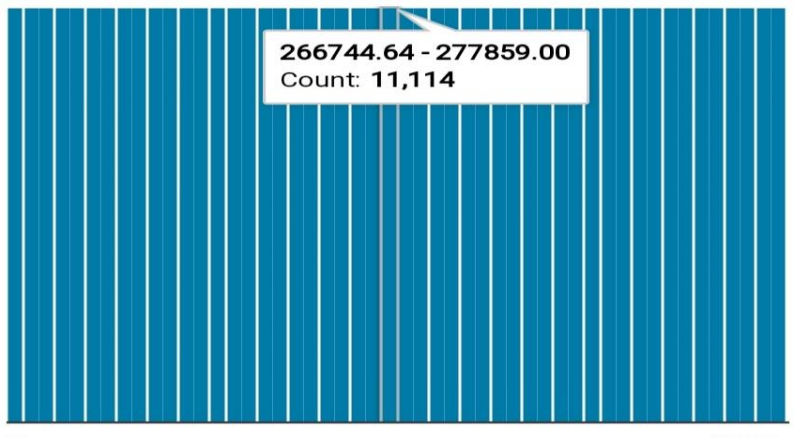
Compliance and Security:

- Ensure the system complies with data privacy and security regulations and standards, such as PCI DSS.
- Implement robust security measures to safeguard the model and sensitive data.

Communication and Stakeholder Involvement:

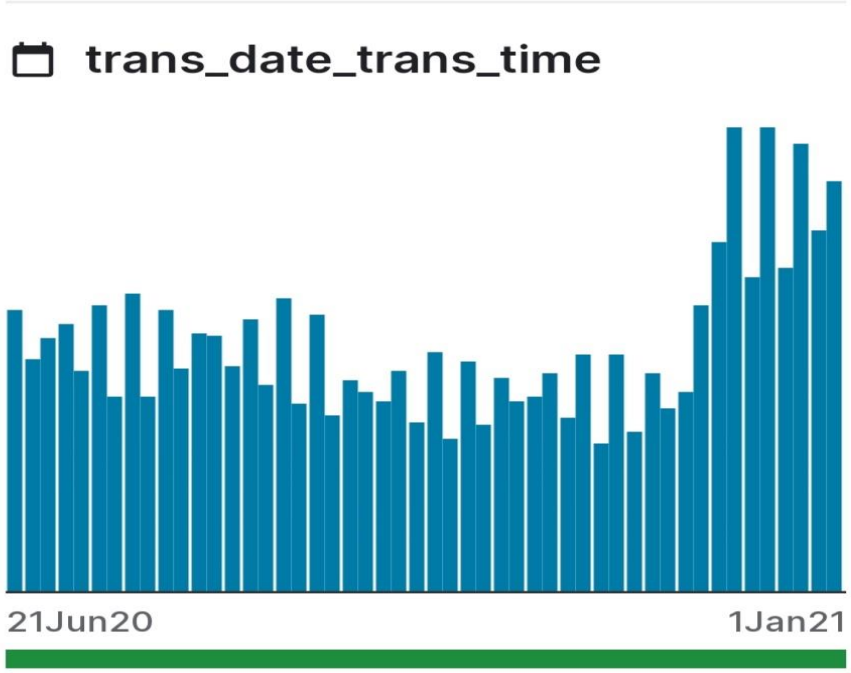
- Maintain open communication with stakeholders, including fraud analysts and business experts.
- Gather feedback to enhance the system's performance and address emerging fraud patterns.

COLUMNS



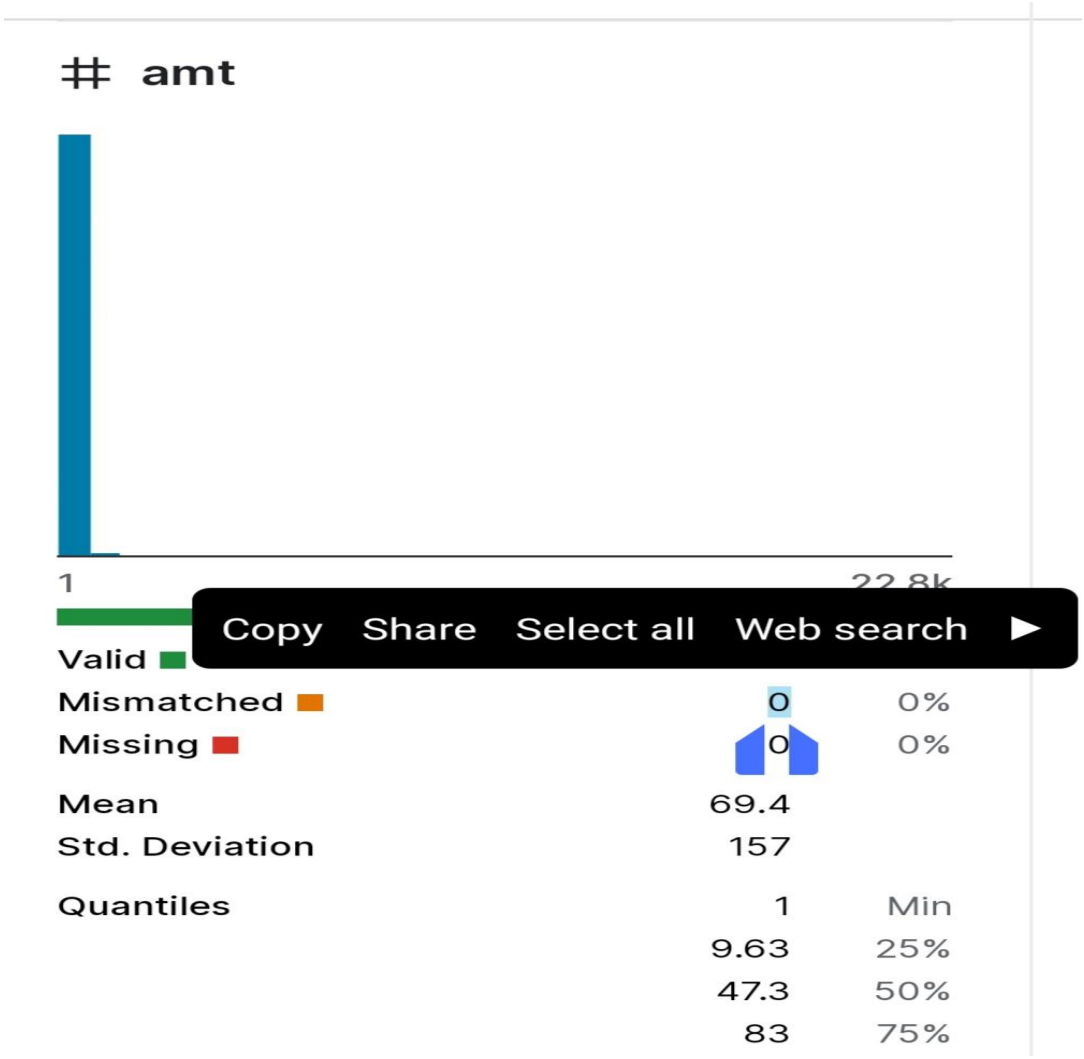
Valid	556k	100%
Mismatched	0	0%
Missing	0	0%
Mean	278k	
Std. Deviation	160k	
Quantiles	0	Min
	139k	25%
	278k	50%
	417k	75%
	556k	Max

TRANS_DATE_TRANS_TIME

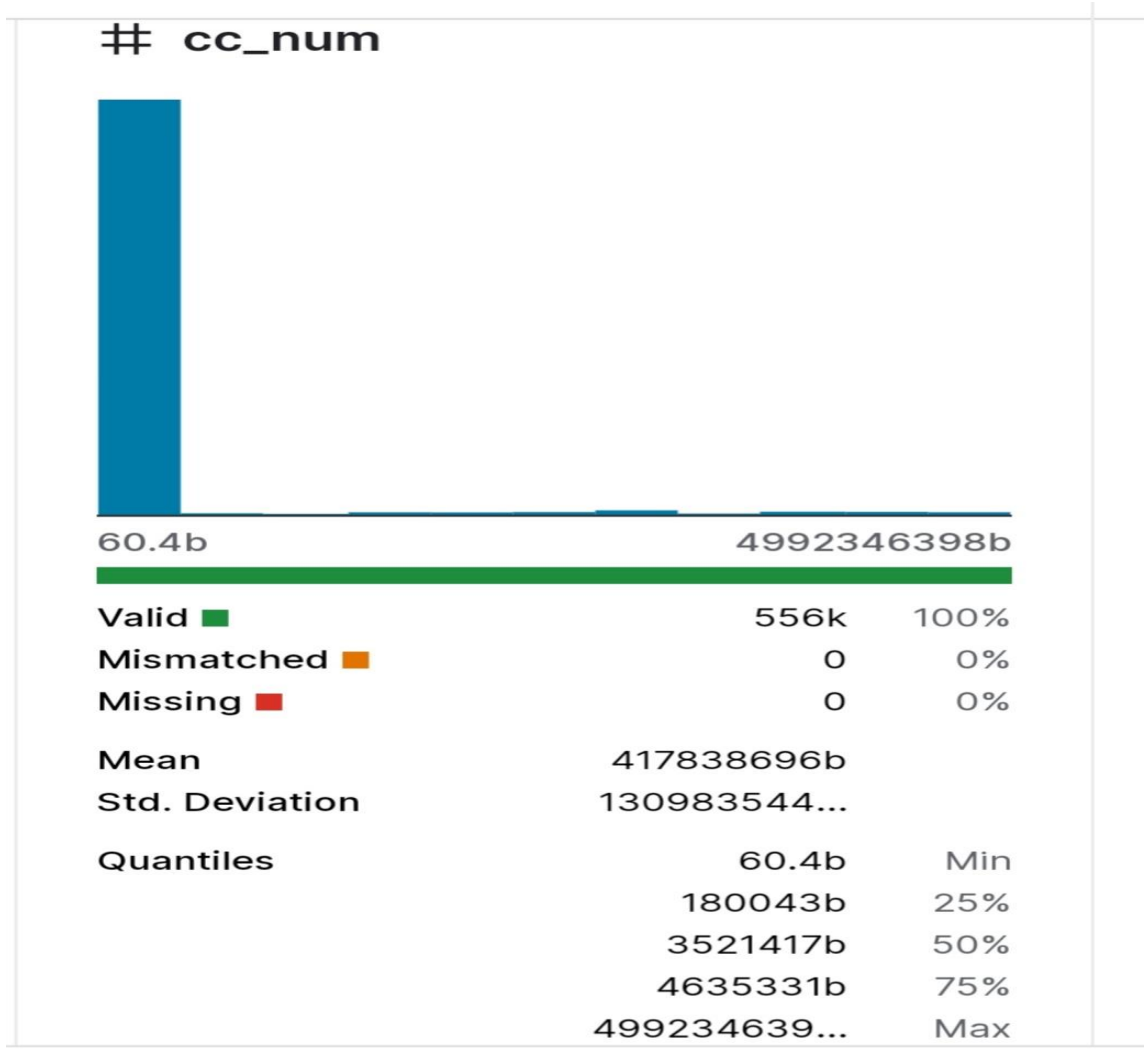


Valid	556k	100%
Mismatched	0	0%
Missing	0	0%
Minimum	21Jun20	
Mean	2Oct20	
Maximum	1Jan21	

#AMT


amt
1
22.8k
Copy Share Select all Web search
Valid
Mismatched 0 0%
Missing 0 0%
Mean 69.4
Std. Deviation 157
Quantiles
1 Min
9.63 25%
47.3 50%
83 75%

#CC_NUM


cc_num
60.4b
4992346398b
Valid 556k 100%
Mismatched 0 0%
Missing 0 0%
Mean 417838696b
Std. Deviation 130983544...
Quantiles
60.4b Min
180043b 25%
3521417b 50%
4635331b 75%
499234639... Max

gender

F	55%
M	45%

Valid	556k	100%
Mismatched	0	0%
Missing	0	0%
Unique	2	
Most Common	F	55%

street

924
unique values

Valid	556k	100%
Mismatched	0	0%
Missing	0	0%
Unique	924	
Most Common	444 Robert ...	0%

first

Christopher	2%
Robert	2%
Other (535200)	96%

Valid	556k	100%
Mismatched	0	0%
Missing	0	0%
Unique	341	
Most Common	Christopher	2%

last

Smith	2%
Williams	2%
Other (533517)	96%

Valid	556k	100%
Mismatched	0	0%
Missing	0	0%
Unique	471	

DATA PREPROESSING SETPS

Handling Imbalanced Data:

Credit card fraud datasets are typically highly imbalanced, with a majority of legitimate transactions and a small fraction of fraudulent ones.

Outlier Detection:

Identify and handle outliers in the data, which could represent potentially fraudulent transactions or data entry errors.

Techniques like the Z-score, IQR, or robust statistical methods can help identify outliers

Data Scaling for Anomaly Detection Models:

If using anomaly detection techniques (e.g., isolation forests or one-class SVMs), scaling may not be necessary. These models can handle data with varying scales effectively.

Data Privacy and Security:

Implement data privacy measures, ensuring that sensitive customer information is anonymized or encrypted to comply with data protection regulations like GDPR.

Time-Based Data Handling:

Credit card transactions often involve time-related features. Ensure that time-based features are appropriately processed and transformed. Consider creating time-based aggregations or features to capture temporal patterns.

Dimensionality Reduction:

If the dataset has high dimensionality, consider dimensionality reduction techniques like Principal Component Analysis (PCA) or feature selection to reduce computation time and improve model performance.

Handling Skewed Features:

Some features may be highly skewed. Apply transformations like log or Box-Cox to make the distributions closer to normal, which can help certain models perform better.

Data Splitting:

Split the preprocessed data into training, validation, and test sets to evaluate the model's performance.

Documentation:

Keep detailed records of all data preprocessing steps and transformations for transparency and reproducibility.

FEATURE EXTRACTION TECHNIQUES TRANSACTION

Transaction Amount Statistics:

Extract statistics such as mean, standard deviation, and percentiles (e.g., 25th, 75th) of transaction amounts for each user or card. Anomalies in transaction amounts can be indicative of fraud.

Transaction Frequency:

Create features related to the frequency of transactions for each user or card, such as the number of transactions in a given time window (e.g., daily, weekly).

Time-Based Features:

Extract temporal information, including hour of the day, day of the week, and month, as well as time between transactions (inter-transaction time).

Geographic Information:

Incorporate location-related features, such as the geographic distance between consecutive transactions or the number of unique locations where transactions occur.

Merchant Category Codes (MCC):

Utilize the MCC of the merchant to create features related to the type of business where the transaction occurred. Unusual MCCs or changes in spending patterns can signal fraud.

Transaction Aggregations:

Aggregate transactions over time windows (e.g., daily, weekly) to create features like total transaction amount, total transaction count, and average transaction amount for each time period.

Velocity Checking:

Calculate features related to the velocity of transactions, such as the number of transactions in a short time interval or the total transaction amount in a given time period.

Address Verification System (AVS) Features:

If available, include features that relate to the AVS, such as the match between the billing address and the shipping address.

Card-Related Features:

Extract features specific to the card, such as the age of the card, credit limit, and the number of previous fraudulent transactions associated with the card.

Fraud Flags:

Use external fraud flags or indicators if available, such as indicators from third-party fraud detection systems or card associations (e.g., Visa or MasterCard).

Behavioral Biometrics:

Incorporate behavioral biometric features, such as typing speed, click patterns, or device attributes if available for online transactions.

Text Analysis (NLP):

Analyze text descriptions associated with transactions, such as transaction comments or merchant descriptions, to extract keywords or sentiments that may signal fraud.

Graph Features:

Create graph-based features to represent transaction networks, such as features based on graph centrality, clustering, or connectivity.

Customer Behavior Analysis:

Create features that capture customer behavior, like spending habits, transaction patterns, and deviations from historical behavior.

MACHINE LEARINING ALGORITHM

Logistic Regression:

Logistic regression is a simple and interpretable algorithm that is often used for binary classification tasks like fraud detection.

Decision Trees:

Decision trees are used to create a flowchart-like model of decisions and their possible consequences.

Random Forests:

Random forests are an ensemble learning method that combines multiple decision trees to improve accuracy and reduce overfitting

Gradient Boosting:

Gradient boosting algorithms, such as XGBoost, LightGBM, and CatBoost, are known for their strong predictive power and the ability to handle imbalanced datasets.

Support Vector Machines (SVM):

SVMs are effective at separating data points into different classes using a hyperplane.

Neural Networks:

Deep learning techniques, such as feedforward neural networks, can be used for credit card fraud detection

Anomaly Detection:

Anomaly detection techniques, such as Isolation Forest, One-Class SVM, and Autoencoders, are particularly suitable for detecting rare and unusual patterns in credit card transactions, making them well-suited for fraud detection.

MODEL TRAINING

Selecting the Model:

Choose an appropriate machine learning or deep learning model for credit card fraud detection. Common choices include:

- Logistic Regression
- Decision Trees
- Random Forests
- Gradient Boosting (e.g., XGBoost, LightGBM)
- Support Vector Machines (SVM)
- Neural Networks (e.g., deep learning models)

Deployment:

Deploy the trained model into a production environment where it can make real-time predictions on incoming credit card transactions.

Hyperparameter Tuning:

This may involve adjusting parameters like learning rates, regularization strength, tree depth, and more. Consider using techniques like grid search or random search for hyperparameter tuning.

Cross-Validation:

Implement cross-validation techniques (e.g., k-fold cross-validation) to assess the model's generalization performance.

Threshold Selection:

Determine an appropriate decision threshold for classifying transactions as legitimate or fraudulent. This threshold may need to be adjusted to balance the trade-off between false positives and false negatives, depending on business requirements.

Alerting and Response:

Implement alerting mechanisms to notify relevant personnel of potentially fraudulent transactions. Develop response procedures to address flagged transactions.

EVALUATION OF METRICS

Accuracy:

Accuracy is the ratio of correctly classified transactions (both legitimate and fraudulent) to the total number of transactions. It is a basic metric but can be misleading when dealing with imbalanced datasets.

Precision (Positive Predictive Value):

Precision measures the ratio of correctly identified fraudulent transactions to the total number of transactions predicted as fraudulent. It is a critical metric as it helps in minimizing false positives, which can inconvenience legitimate customers.

Recall (Sensitivity or True Positive Rate):

Recall is the ratio of correctly identified fraudulent transactions to the total number of actual fraudulent transactions.

F1-Score:

The F1-Score is the harmonic mean of precision and recall. It provides a balance between precision and recall and is useful when there's a trade-off between these two metrics.

Area Under the Receiver Operating Characteristic (ROC AUC):

The ROC AUC measures the model's ability to distinguish between legitimate and fraudulent transactions.

Area Under the Precision-Recall Curve (AUC-PR):

The AUC-PR is particularly important for imbalanced datasets. It quantifies the area under the precision-recall curve, where precision is plotted against recall.

Specificity (True Negative Rate):

Specificity measures the ratio of correctly identified legitimate transactions to the total number of actual legitimate transactions.

False Positive Rate (FPR):

FPR is the ratio of false positives to the total number of actual legitimate transactions. Minimizing the FPR is crucial to reduce the number of false alarms.

Conclusion

In conclusion, credit card fraud detection is a vital defense against financial crime. Leveraging data and machine learning, it identifies and prevents fraudulent transactions, safeguarding both individuals and financial institutions.

The ongoing battle against fraud requires continual monitoring, data security, and transparency in model decisions to adapt to emerging threats and maintain trust in financial transactions.