**Project Overview: Fraud Detection System**

1. **Introduction**

A *fraud transaction* is an unauthorized or deceptive financial activity, such as using stolen credit card details, identity theft, or falsified payments, intended to gain money or goods illegally. These transactions often appear as outliers compared to legitimate ones, exhibiting unusual patterns like high amounts, odd locations, or rapid frequency.

*Fraud detection* is the process of identifying and preventing such transactions using data analysis and machine learning. It involves analyzing transactional data for anomalies, patterns, or behaviors indicative of fraud (e.g., unusual spending, mismatched locations), often employing techniques like classification models (e.g., RandomForestClassifier), resampling for imbalanced data (e.g., SMOTETomek), and real-time monitoring to flag or block suspicious activities, minimizing financial losses while ensuring legitimate transactions proceed smoothly.

**What are we trying to find out?**  
The project aims to develop an effective fraud detection system by determining the best machine learning model to identify fraudulent transactions in a highly imbalanced dataset (0.56% fraud), focusing on maximizing recall and precision for the fraud class.

**What are we aiming to achieve?**  
Success is defined as deploying a model that accurately detects fraud (high recall to catch most fraud cases, high precision to minimize false positives) in real-time, with an F1-score above 0.90 on an imbalanced test set, ensuring practical usability for financial institutions with minimal operational disruption.

1. **Fraud detection**

The dataset was reduced by selecting data from 2019, focusing on the last quarter (November) due to potential fraud events. It was further narrowed by region, selecting the smallest region (250,000 rows), and irrelevant columns like CC number and date of birth were explored for useful data before being omitted. Over all 17 features were selected for the final modelling stage. The final dataset for modelling contains 347,746 after SMOTETomek model.

1. **What is the Likelihood of a Transaction Being Fraudulent?**

Criteria for Classifying a Transaction as "Fraudulent" in This Fraud Detection System:

* High-risk transaction category
* Unusually high transaction amount
* Abnormal transaction frequency
* Occurs during a high-risk time period

Only 0.56% of Transactions Are Identified as Fraudulent in the Dataset.

This Project Investigates the Factors Predicting Fraudulent Transactions:

Beyond the listed criteria, we will analyse the dataset to uncover patterns, such as location, issuer, or user behaviour, to understand and predict the likelihood of identifying fraudulent transactions.

1. **Data from the CSV**

The dataset consisted of 27 columns from two CSV files: credit\_card\_fraud.CSV and customers.CSV. The customers.CSV contained overlapping information with the main transaction’s dataset. They were merged using key identifiers such as trans\_date, trans\_num, and cc\_num, integrating transactional data with customer demographics. This structured approach enhances fraud detection by enabling a more comprehensive analysis of fraud patterns and key indicators.

No external data was used.

1. **Dataset Overview**

**Time Period:**

The original dataset spans transactions from January 1, 2019, to December 31, 2020. After reduction, the dataset focuses on transactions from November 1 to November 30, 2019, narrowing down to the 4th quarter of 2019 for analysis.

**Fraud Detection Target:**

The analysis focuses on detecting fraudulent transactions, with the target variable is\_fraud (0 = non-fraudulent, 1 = fraudulent).

**Key Influencing Factors**

Features like transaction category, amount (amt), transaction time, city population (city\_pop), and location show a strong correlation with fraud occurrence.

**Imbalance Challenge**

Fraudulent transactions make up only 0.56% of the dataset, requiring specialized techniques to balance the data during model training.

**Exploratory Data Analysis (EDA)**

Initial analysis included summary statistics for numerical variables (amt, city\_pop) and frequency distributions for categorical features (category).

Visualizations like histograms and bar plots helped identify trends and anomalies in the data.

Next, feature engineering and encoding will prepare the data, while SMOTE or Tomek Links will address class imbalance. Correlation analysis and advanced models will then enhance fraud detection.

**Data summery**

1. **Data Cleaning and Handling Missing Values**

**Raw Data Processing**

The dataset was filtered to focus on November 2019 (Q4), with new features like day and is\_weekend extracted from trans\_date to capture temporal fraud patterns.

**Outlier Handling**

EDA revealed extreme values in amt (max $16,498.31, mean $67.93) and city\_pop (max 2,504,700). While amt outliers were retained due to potential fraud relevance, city\_pop and amt were log-transformed to reduce skewness.

**Missing Values**

Key variables (amt, category, is\_fraud) had no significant missing data. Minor gaps in non-critical fields (job, dob) were negligible and left untreated.

**Feature Adjustments**

Constant features (year, month, quarter) were removed. trans\_time was categorized into time\_category (Morning, Night) to highlight fraud patterns. The highly imbalanced target variable (is\_fraud at 0.56%) will be addressed using SMOTETemoke during modeling.

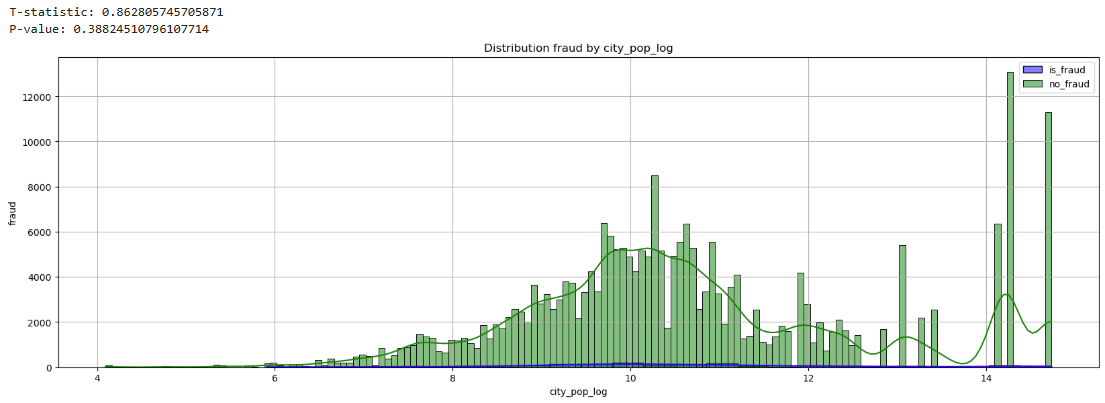
**T-test** - A t-test compares the means of two groups to determine if their differences are statistically significant. It was applied to is\_fraud against features like age, log-transformed city population, time category, day, and gender to identify meaningful patterns.

1. Fraudulent transactions are more common among younger individuals (30-40), while non-fraudulent transactions peak at 40-50. The overall age distribution is right-skewed (20-60). A low P-value (1.95e-11) confirms a significant difference in age, with fraud cases skewing younger. This suggests age may be a valuable feature for fraud detection.

A graph showing a line of green and blue

AI-generated content may be incorrect.

1. City population size does not significantly impact fraud occurrence (p-value = 0.388), as both fraud and non-fraud transactions peak at similar city\_pop\_log values (~10). Fraud remains rare overall. A possible issue with the city\_pop\_log scale should be clarified. More relevant fraud indicators may include transaction time, issuer, or category.



1. The time of day distribution reveals distinct peaks, with non-fraudulent transactions following the general pattern, while fraudulent transactions show higher density around midnight and late evening. The extremely low p-value (4.44e-93) confirms a significant difference in timing. Fraudulent transactions are more likely to occur during off-peak hours, suggesting that time of day is an important factor for fraud detection, as these periods may have less oversight.

A graph with green and blue lines

AI-generated content may be incorrect.

0 = Afternoon 1 = Evening 2 = Morning 3 = Night

1. The transaction data shows a consistent daily pattern, with a significant spike on day 31, driven by non-fraudulent transactions. The is\_fraud group has a flatter distribution with a slight increase at month-end. A low p-value (1.69e-11) and positive T-statistic indicate that fraudulent transactions tend to occur more towards the end of the month, possibly due to financial pressures or behavior changes. This suggests that the day of the month is an important factor for fraud detection, with fraud being more likely at the end.

A graph with green and blue lines

AI-generated content may be incorrect.

1. The P-value of 0.0458 indicates a statistically significant difference in gender representation between fraudulent and non-fraudulent transactions, but the result is borderline and may not be considered highly significant. The negative T-statistic suggests a slight association between fraudulent transactions and one gender (encoded as 0), but the small magnitude and the similar plots indicate that the difference is not substantial.

A graph with numbers and a bar

AI-generated content may be incorrect.

1. **Model Development and Imbalance Handling**

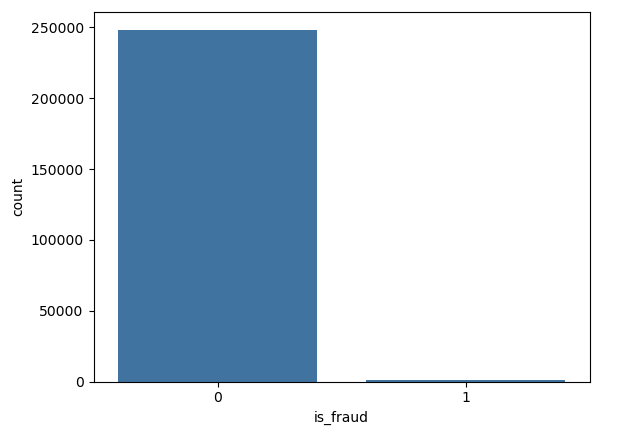
**Model Selection**

After completing the data preparation in earlier stages, we umade a split to train, test, and validation:

* Split into train+val and test sets (80% train+val, 20% test)
* Split train+val into train and val sets (75% train, 25% val from the train+val set)

**Imbalanced Data**

The target variable is is\_fraud, with only 0.56% of transactions being fraudulent. This severe class imbalance requires careful handling to maximize model accuracy and reduce errors in fraud detection.



**Imbalance Handling**

We trained a RandomForestClassifier on the imbalanced dataset. However, this approach failed to detect most fraudulent transactions, as shown by a low recall for the minority class.

Confusion Matrix: TP: 379, FP: 0, TN: 74518, FN: 38 (is\_fraud = 0 - 248391, 1 - 1391)

A diagram of different colored squares

AI-generated content may be incorrect.To address this, we applied a combination of Undersampling and Oversampling:

* **ROS (Random Over Sampling)**: Randomly duplicates instances from the minority class to balance the dataset.
* **RUS (Random Under Sampling)**: Randomly removes instances from the majority class to balance the dataset.
* **SMOTE (Synthetic Minority Over-sampling Technique)**: Generates synthetic samples for the minority class by interpolating between existing instances.
* **SMOTETomek**: Combines SMOTE with Tomek Links to generate synthetic samples and remove ambiguous instances, improving decision boundaries.

This approach aims to improve the model’s ability to detect fraudulent transactions while maintaining overall performance. Eventually after testing all, SMOTETomek performs the best overall, balancing precision and recall for both classes, especially the 'is fraud' (class 1), with high F1-score and accuracy. RUS struggles with low precision for class 1, despite having high recall. It’s less reliable due to many false positives. ROS and SMOTE both perform well, with strong precision and recall for class 1, but SMOTETomek still has a slight edge in balancing both metrics. Overall, SMOTETomek appears to be the most balanced and effective model.

**Final Model Implementation and Evaluation on the Fraud Detection Dataset**

**Dataset Overview**

The dataset used for this fraud detection project has been preprocessed and balanced using SMOTETomek, resulting in a dataset (SMOTETomek\_df.pkl) with 347,746 rows and 17 columns.

**Model Training**

Model training involved using multiple classifiers, including SVM, Logistic Regression, ADABoost, GBM, Decision Tree, and Random Forest. After performing hyperparameter tuning, the optimal parameters for the Random Forest classifier were identified as:

* bootstrap=False
* max\_depth=30
* min\_samples\_split=5
* n\_estimators=300
* random\_state=1

The model was then evaluated on the test set (X\_test, y\_test).

**Conclusion**

The system includes validation steps like cross-validation and evaluation on a holdout test set to ensure robustness. However, the near-perfect validation performance (accuracy 0.9998) on a balanced set suggests potential overfitting or data leakage, however it can be explained due to the imbalanced process the dataset had to go through.

1. **Deployment of the Fraud Detection Model**

**End User and Deployment Strategy**

The end user of this fraud detection machine learning (ML) model will be financial institutions, such as banks and credit card companies, that process large volumes of transactions daily. The model will be deployed on cloud servers, specifically leveraging platforms like AWS or Google Cloud, to ensure scalability and real-time processing capabilities. To protect their customers and reduce financial losses. The model reflects the business success of these institutions by accurately detecting fraud while minimizing false positives, thereby maintaining customer trust and operational efficiency.

**Business Impact**

A robust fraud detection system will enhance customer trust by reducing the incidence of undetected fraudulent transactions, thereby lowering financial losses for both the institution and its customers. By accurately identifying fraud, the model can also reduce the number of false positives, which minimizes customer inconvenience (unnecessary transaction declines) and reduces operational costs associated with manual fraud investigations.

**Nurturing and Motivating Fraud Prevention**

By focusing on high-risk areas, such as transactions in specific categories or during certain times, institutions can allocate resources more effectively. Additionally, the model can help identify "hot spots" for fraud, such as regions with high fraud rates allowing for targeted interventions. For instance, institutions can implement stricter verification processes for transactions in these high-risk areas or times, reducing the likelihood of fraud going undetected.

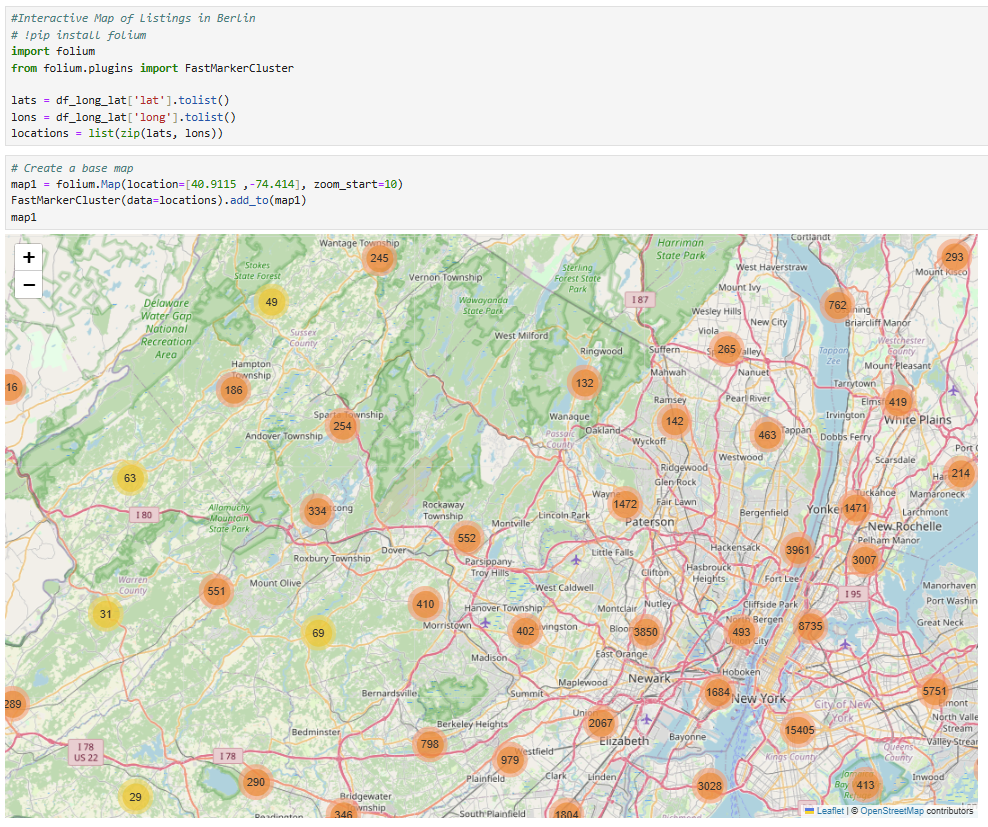
**Deployment Infrastructure**

The fraud detection model will be deployed on cloud servers (AWS, Google Cloud) interfacing with the main transaction processing platforms of financial institutions. This setup ensures scalability to handle large transaction volumes and provides real-time fraud detection capabilities. The model will be integrated into the transaction pipeline, where it will process incoming transactions and flag potential fraud in real time. Continuous updates to the model will be implemented to adapt to evolving fraud patterns, ensuring long-term effectiveness.

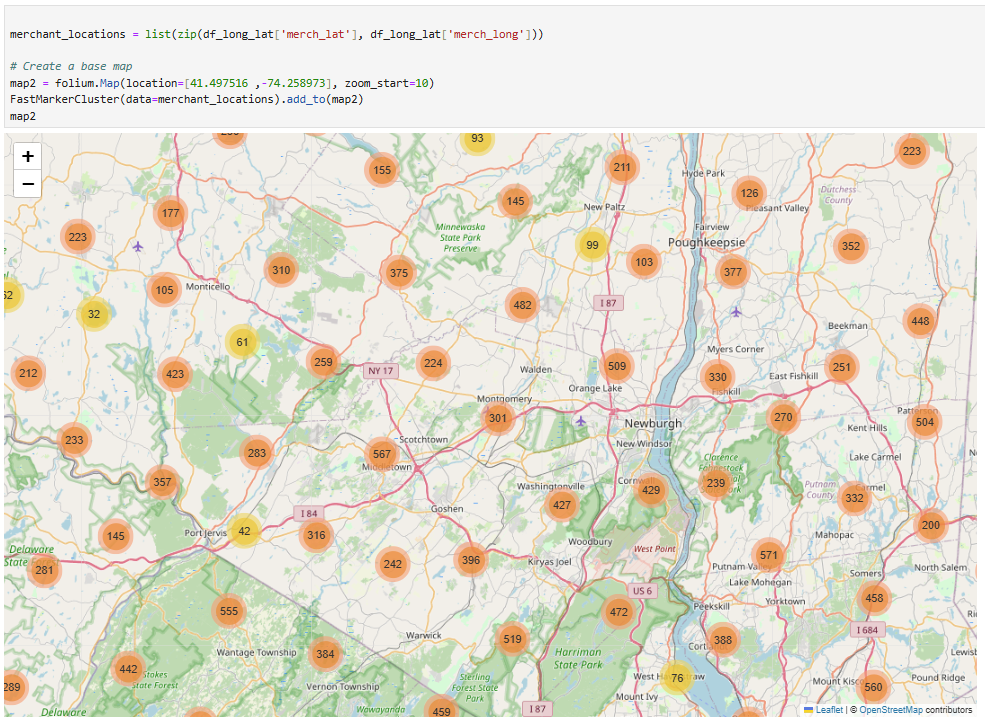
**Conclusion**

The deployment of this fraud detection model on cloud infrastructure will enable financial institutions to identify and prevent fraudulent transactions in real time, focusing on high-risk categories, times, and regions. By reducing false positives and improving detection accuracy, the model will enhance customer trust, reduce financial losses, and optimize fraud prevention efforts. Continuous monitoring and updates will ensure the model remains effective against evolving fraud patterns, making it a valuable tool for the financial industry.

Interactive Map – customers



Interactive Map – Merchant



Word Cloud –

A black background with green text

AI-generated content may be incorrect.