CMSE 492 Project Report - HW2:
Alice Shirley

Project Overview

Detecting Credit Fraud with Machine Learning
GitHub Repository

This project is based on a dataset containing transactions made by credit cards in September 2013 by European cardholders. This dataset presents transactions that occurred over two days, with 492 frauds out of 284,807 transactions. The motivation for this project is to improve upon my previous work in CMSE 202, which predicted credit fraud using methods such as Logistic Regression, PCA, Naive Bayes Classifier, and Random Forest. Using the knowledge gained in CMSE 492, I will construct a machine learning model that creates better predictions than those produced in CMSE 202.

Project Setup

My directory is set up as follows:

- **Folders**: Data, Features, Models, Notebooks, Reports, Tests, Visualizations
- **Key Files**: README.md and .gitignore
- **Data Folders**: Raw and Preprocessing
- **Models Folder**: Train SVM and Tuning & Evaluation
- **Notebooks Folder**: Exploratory and Final

Explanation of Key Files and Directories

The **Preprocessing** directory contains data with additional feature-engineered columns and is necessary for creating preprocessed.csv.

List of Dependencies and Setup Instructions

Setup: The initial dataset, creditcard.csv, is too large to upload to GitHub, so you need to download it from Kaggle and upload it into the Raw and Preprocessing directory. Additionally, preprocessed.csv, which is produced from running the preprocessing notebook, should be uploaded into any directory where the data is manipulated, by copying it from the Preprocessing directory.

Completed Tasks

List of Tasks Completed from the Homework

Preprocessing

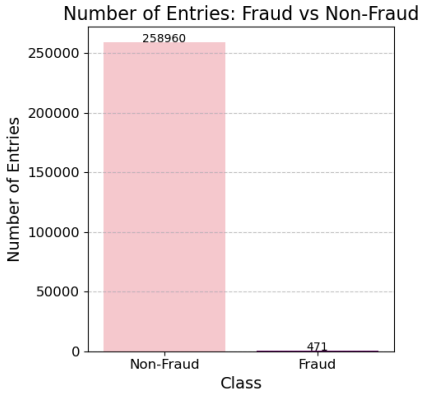
- **Check for Missing Values**: Identifying and addressing missing values is crucial for ensuring data quality and integrity, as incomplete data can lead to inaccurate model predictions and analyses.
- **Correlation Analysis**: Assessing high correlation between columns helps eliminate multicollinearity, enhancing model performance by reducing redundancy and simplifying the feature set.
- **Feature Engineering**: Transforming TransactionTime into HourOfDay and MinuteOfHour enables the model to capture temporal patterns, providing valuable categorical variables that can enhance the detection of fraudulent transactions.

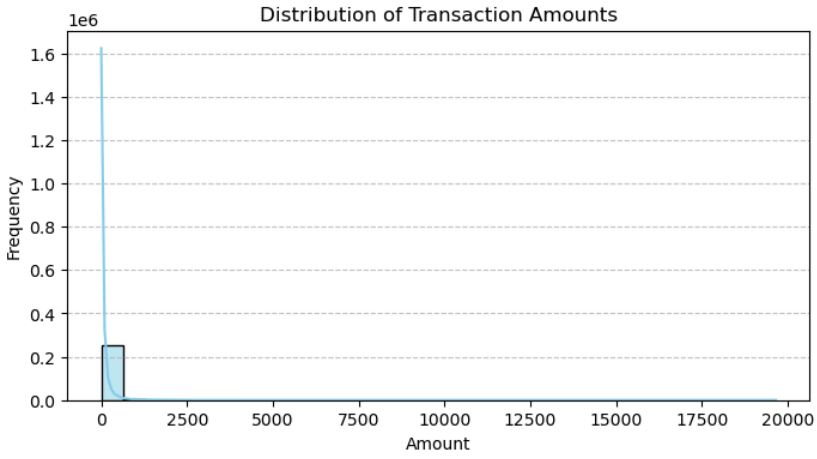
Exploration

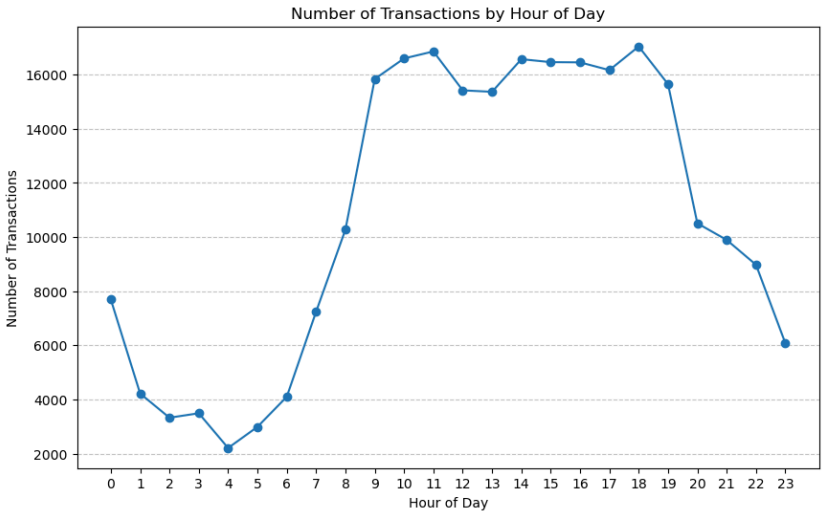
- **Data Visualization**: Creating visualizations to analyze class concentration and transaction distributions provides insights into the underlying data structure, helping identify trends, anomalies, and potential biases.
- **Understanding Transaction Patterns**: Examining the number of transactions by hour reveals behavioral patterns critical for detecting fraud, as certain times may correlate with higher fraud rates, allowing for more targeted monitoring and prevention strategies.

Initial Analysis and Findings

Key Findings: I found that 0.18% of entries are fraudulent while 99% of entries are non-fraudulent. This significant imbalance (99% vs. 0.18%) presents a challenge for machine learning models.







Proposed Approach

The proposed approach is to use Support Vector Machines (SVM). SVMs are effective in high-dimensional spaces and can utilize different kernels to handle non-linearity. I will use this model because it excels in handling imbalanced datasets, allowing the adjustment of the class_weight parameter to emphasize minority classes like fraudulent transactions. Despite being sensitive to outliers, careful tuning of parameters and preprocessing can mitigate these effects. By implementing SVMs with these strategies, organizations can enhance their fraud detection capabilities.

Preliminary Results

At this point, I have no preliminary results from my work in CMSE 492 beyond exploratory work. However, my work on this dataset in CMSE 202 shows that Logistic Regression and Naive Bayes Classifier are not sufficient for predicting credit fraud.

Challenges and Solutions

- **Challenge**: I am encountering the challenge of a widely unbalanced dataset. My plan to combat this is to use an SVM, which allows for the adjustment of class weight to address the imbalance, as well as hyperparameter tuning.
- **Challenge**: The original dataset is too large to push to GitHub. I am addressing this by providing direct instructions for users to download the dataset from Kaggle and upload it to their local directories.

Next Steps

1. **Complete Preprocessing**: Finalize feature engineering and ensure preprocessed.csv is correctly formatted.
2. **Conduct Exploratory Data Analysis (EDA)**: Further analyze and visualize transaction patterns, focusing on fraud detection features.
3. **Model Development**: Implement SVM with hyperparameter tuning.
4. **Evaluate Model Performance**: Use metrics suited for imbalanced datasets.
5. **Documentation and Reporting**: Document findings, challenges, and methodologies in the project report and prepare visuals and summaries for the final presentation.

Timeline

- **Week 1**: Complete preprocessing and EDA; ensure all visualizations are created and insights gathered.
- **Week 2**: Implement SVM model, including parameter tuning; conduct preliminary evaluations.
- **Week 3**: Analyze results; refine the model based on evaluation metrics.
- **Week 4**: Finalize the report, prepare presentation materials, and conduct a final review of findings.

Conclusion

Summary of Current Project Status

The project is currently focused on preprocessing and exploratory analysis, with key tasks completed, including missing value checks and feature engineering. Visualizations have revealed a significant class imbalance, with only 0.18% of entries classified as fraudulent.

Reflection on Progress and Lessons Learned

This project has deepened my understanding of the importance of data preprocessing and its impact on model performance. Addressing the imbalanced dataset has underscored the need for careful algorithm selection and parameter tuning.

Outlook for Project Completion

The next steps involve implementing the Support Vector Machine model and optimizing its parameters for better fraud detection. I am confident that with a structured approach, I will successfully complete the project and gain valuable insights into credit fraud detection.

References

The dataset is sourced from Kaggle.