Introduction (Kaggle Username marianoeliseoamaya)

This report details the methodology and insights derived from analyzing a dataset of credit card transactions to identify fraudulent activities. The primary objective was to leverage machine learning techniques to classify transactions accurately as fraudulent or non-fraudulent. This endeavor involved comprehensive data exploration, intelligent feature engineering, and strategic model selection and tuning.

The dataset comprises 555,719 training observations and 69,465 testing observations, each representing individual credit card transactions with various attributes. Initial exploration revealed a mix of numerical and categorical features, including transaction amount, date and time, merchant details, and customer demographics. A critical step was to assess missing values, ensuring data quality for modeling. The exploration phase included visualizing the distribution of the target variable, which highlighted an imbalanced class distribution favoring non-fraudulent transactions. Additionally, plotting numerical features helped identify potential outliers and trends worth investigating further.

Feature Engineering

During the exploratory data analysis phase, certain patterns emerged that were indicative of potential fraud. It was observed that fraudulent transactions had amounts below $10000. This insight led to the creation of features that could capture extreme values in transaction amounts, aiding the model in identifying transactions that deviate from typical spending patterns. Analysis showed that fraudulent activities tended to peak during certain times of the day (12-11pm) and on specific days of the week (Monday, Tuesday, Sunday). This pattern suggests that fraudsters might prefer times when detection is less likely, either due to lower transaction volume or during non-working hours. Incorporating features like hour_of_day and day_of_week enabled the model to leverage temporal patterns in predicting fraud. Certain merchant categories were more associated with fraud than others. This finding was pivotal in creating categorical features that highlight transactions in high-risk categories. The exploration revealed that certain age groups and jobs might be more targeted for fraud or more prone to making transactions flagged as fraudulent. Though direct causality could not be established, incorporating demographic information as features provided additional context for the model. Looking at a Age by Fraud Status visual, I noticed that people from 40 - 70 are more likely to fall victims of fraud. Another sometimes forgotten variable to look out for is states. Sometimes states or companies do have sufficient resources to fight fraud related transactions, there might be a group of criminals that are residents of these states, and other things like that. By changing the feature extraction, I saw how one feature changes the F1 by a lot. This is why it's important to look at the graphs because they tell a story, and that's why I found it important to see the graphs first and then create my model. Knowing that fraudulent transactions could exhibit extreme values, data normalization techniques were carefully chosen to ensure that these signals were not diminished. Similarly, understanding the transaction frequency across different hours and days informed the decision on how to treat temporal features, whether as continuous or categorical variables. Insights into credit card use also shaped the validation strategy. Specifically, temporal validation (splitting the training and test sets based on time) was considered to mimic real-world deployment scenarios better, ensuring the model's robustness across different times.

Model Selection

The KNN model, before any optimization and with n_neighbors set to 1, achieved a high F1 score of 0.9260204081632653. This initial success can be attributed to the model's ability to capture the local patterns of fraud within the dataset effectively. KNN's performance suggests that fraudulent transactions may cluster closely together in the feature space, allowing the algorithm to accurately identify similar fraudulent instances based on proximity to known fraud cases. However, despite its high F1 score, the KNN model's reliance on distance calculations makes it computationally intensive, especially as the dataset size increases.

The ensemble classifier, combining KNeighborsClassifier, DecisionTreeClassifier, and LogisticRegression through soft voting, aimed to harness the strengths of each individual model. Despite this, it achieved a moderate F1 score of 0.368932. The ensemble method's underperformance compared to the initial KNN model might be due to the dilution of the strong signal captured by KNN when averaged with the other models. This indicates that not all models contributed equally to detecting fraud, possibly because of the varying sensitivity to the patterns and noise within the data.

XGBoost stood out for its superior performance, with an F1 score of 0.833087. Its success can be attributed to its gradient boosting framework, which iteratively corrects the errors of previous trees, allowing it to adaptively focus on challenging cases. XGBoost's ability to handle imbalanced datasets, through mechanisms like weighted sampling of the minority class, made it particularly effective for this task. Moreover, its feature importance utility provided insights into which features were most predictive of fraud, guiding further feature engineering and selection efforts.

Optimization of the KNN model involved tuning parameters such as n_neighbors, weights, and metric, resulting in an improved F1 score of 0.66828087. The optimization process likely addressed overfitting seen in the initial model by considering more neighbors for making predictions, hence capturing broader patterns in the data beyond immediate neighbors. However, the optimized KNN model still lagged behind XGBoost in performance, underscoring the latter's superior ability to model complex non-linear relationships and interactions among features.

Why Some Models Worked Better Than Others

The differential performance of these models underscores several key points about machine learning model selection:

Data Complexity: XGBoost's ability to capture complex patterns through ensemble learning gave it an edge over simpler models like KNN and Logistic Regression, particularly for the intricate task of fraud detection.

- **Imbalance Handling**: Models that offer mechanisms to deal with imbalanced classes (like XGBoost) tend to perform better in scenarios where fraudulent transactions are rare compared to non-fraudulent ones.
- **Computational Efficiency**: While KNN can be highly accurate, its computational cost at prediction time can be a drawback, especially for large datasets and real-time detection systems.

- **Model Interpretability**: Despite their effectiveness, models like XGBoost may sacrifice interpretability for performance, a trade-off that must be considered based on the application's requirements.

In summary, the selection and optimization of models for fraud detection require a careful balance between accuracy, interpretability, and computational efficiency. The superiority of XGBoost and KNN in this context can be attributed to their robust handling of imbalanced data, ability to model complex relationships, and adaptiveness to the training data's nuances.

Challenges and Solutions

Machine learning models, including those mentioned like KNN, XGBoost, and ensemble classifiers, tend to perform better on the majority class while struggling to accurately identify instances of the minority class. In the context of fraud detection, this means a model might predominantly predict transactions as non-fraudulent, achieving high accuracy overall but failing to catch many actual fraud cases. This challenge was evident in the initial phases of model training, where despite achieving high accuracy, the practical utility of the model was questionable due to its inability to effectively detect fraud instances. Standard metrics like accuracy become less informative and potentially misleading in imbalanced datasets. For instance, a model that naively predicts every transaction as legitimate in a dataset with 99% legitimate transactions and 1% fraudulent ones would still achieve 99% accuracy. This scenario necessitated the focus on other metrics like the F1 score, which considers both precision and recall, providing a more balanced view of model performance, especially for the minority class.

As noted, metrics like the F1 score, precision, recall became crucial for assessing model performance more accurately, guiding the optimization and tuning of models beyond mere accuracy. The ensemble classifier, despite its moderate success, was part of an effort to mitigate bias by combining multiple models. The diversity in decision-making strategies was aimed at improving the overall detection of fraudulent transactions, demonstrating a collective intelligence approach. Adjusting the classification threshold based on probabilities, especially in models producing probabilistic outputs, allowed for more nuanced control over the trade-off between precision and recall, tailoring the model's sensitivity to fraud detection. Navigating the challenge of an imbalanced dataset required a multifaceted approach, combining data preprocessing, model selection, and post-modeling adjustments. It highlighted the importance of understanding the underlying data distribution and adjusting the modeling strategy accordingly. This challenge also underscored the necessity of continuous iteration and experimentation in model development to find the optimal balance between identifying fraudulent transactions accurately and maintaining a manageable false positive rate.

Conclusion

This exploration into credit card fraud detection underscored the importance of comprehensive data analysis, innovative feature engineering, and meticulous model selection. The KNN and XGBoost model stood out for its performance, attributed to its ability to handle imbalanced data and leverage engineered features effectively. The insights gained not only facilitated the development of a high-performing classifier but also shed light on behavioral patterns associated with fraudulent transactions.