Progress Report: Credit Card Fraud Detection

Team Members (Data Minds): Hilda Ogamba, Joyce Malicha, Lynn Obadha

1. Project Progress Overview

The goal of this project is to develop a model that accurately identifies fraudulent transactions in a highly imbalanced credit card dataset. Below is a summary of the progress made thus far:

1.1 Completed Tasks

1.1.1 Data Preprocessing

- **Data Scaling**: Applied scaling techniques to the "Amount" feature to align it with the PCA-transformed features (V1-V28).
- **Class Imbalance Handling**: Successfully implemented Synthetic Minority Over-sampling Technique (SMOTE) to address class imbalance. Fraud cases were undersampled and oversampled for balanced training, ensuring a robust comparison of sampling methods.
- **Train-Test Split**: Split the dataset into training and test sets using stratified sampling to maintain fraud-to-non-fraud proportions.

1.1.2 Model Development:

- **Model Training**: Trained machine learning models - Logistic Regression, Random Forest, XGBoost and Neural Networks.
- **Hyperparameter Tuning**: Employed GridSearchCV for hyperparameter optimization on each model.
- **Ensemble Learning**: Began preliminary tests with ensemble methods, specifically Random Forest and XGBoost, for potentially enhanced performance.

1.1.3 Evaluation Setup:

- **Evaluation Metrics**: Established AUPRC as the primary evaluation metric, supported by precision, recall, confusion matrix, ROC-AUC score and F1-score due to dataset imbalance.
- **Comparative Analysis**: Created a structured evaluation pipeline to compare models, focusing on fraud detection performance under varying resampling techniques.

1.2 Data

The dataset used is sourced from the European credit card transactions dataset published by Worldline and the Machine Learning Group at ULB. This data, containing 492 fraudulent transactions out of 284,807, is hosted on Kaggle.

2. Challenges

2.1 Extended Training Time:

Model development for the Neural Network, Random Forest, Logistic Regression and XGBoost models required substantial computation time. The training session took approximately 7 hours to complete.

2.2 Interpretability

Interpretability is challenging with Random Forests and XGBoost, as they are effective for fraud detection but difficult to interpret and understand in terms of insights provided.

2.3 Hyperparameter Tuning Complexity:

Selecting optimal parameters proved challenging due to model sensitivity to data imbalance and variance in AUPRC scores.

Solution: Limited the hyperparameter search space initially to shorten tuning time..

3. Collaboration

Group Meetings: The team has been meeting twice weekly to discuss progress and challenges, with additional sessions scheduled as needed.

Contribution: All members are actively contributing to data preprocessing, model development, and evaluation setup. Any gaps in understanding are addressed promptly to ensure equal participation.

4. Next Steps

4.1 Further Model Tuning:

- Conduct final parameter tuning to refine model precision and recall balance.

4.2 Evaluation and Comparative Analysis:

- Document performance insights for each model, including ensemble learning results.

4.3 Potential Challenges:

- **Compute Resources**: Extended training times may continue to pose an issue.
- **Model Overfitting**: Adjustments to prevent overfitting may be required, particularly with ensemble methods like XGBoost.

Conclusion

The project is progressing well with completed preprocessing and model training. Key challenges around training time and model tuning have been addressed. We are on track to complete the comparative analysis and finalize model evaluation.