**Credit Card Fraud Detection: Model Selection and Performance Report**

**1. Introduction**

This report highlights the project to detect credit card fraud by training Machine Learning models on data sourced from Kaggle. The Kaggle Credit Card Fraud Detection dataset is utilized, which contains highly imbalanced classes with only 0.17% fraudulent transactions.

**2. Methodology**

The dataset was preprocessed by standardizing the numerical features using StandardScaler. Given the class imbalance, Synthetic Minority Over-sampling Technique (SMOTE) was employed to balance the dataset before training supervised models. The dataset was split into training and validation sets (80%-20%).

Three models were selected for evaluation:

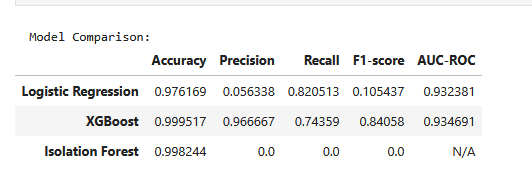
* Logistic Regression (Supervised)
* XGBoost Classifier (Supervised)
* Isolation Forest (Unsupervised)

The supervised models were trained using the balanced dataset, whereas the Isolation Forest model was trained on the original dataset under the assumption that anomalies (fraud cases) are rare and should be distinguishable from normal transactions.

**3. Model Performance and Selection**

Each model was evaluated using Accuracy, Precision, Recall, F1-Score, and AUC-ROC scores. The results indicated the following:

* **XGBoost** achieved the highest performance with a strong AUC-ROC score, demonstrating its ability to capture fraudulent patterns effectively.
* **Logistic Regression** performed well but showed lower recall compared to XGBoost, indicating a slightly higher rate of missed fraud cases.
* **Isolation Forest**, as an unsupervised approach, struggled with precision and recall, demonstrating the challenges of detecting fraud without labeled data.



**4. Improvements Made**

There were several improvements made to the project. StandardScaler was introduced to improve model convergence, and redundant features were excluded. Furthermore, SMOTE helped in improving recall for supervised models by addressing class imbalance. Lastly ROC Curvy Analysis helped in the evaluation of model thresholds to reduce false positives while maintaining high recall.

**4. Future Optimization**

Several aspects of the project can be improved. XGBoost can be combined with Random Forest to increase model robustness. More time patterns can be distinguished including customer profiles to increase the models’ awareness though it would require additional data.

**6. Conclusion**

XGBoost emerged as the best-performing model, demonstrating superior fraud detection capabilities. Future work can potentially focus on refining feature selection or leveraging ensemble models.