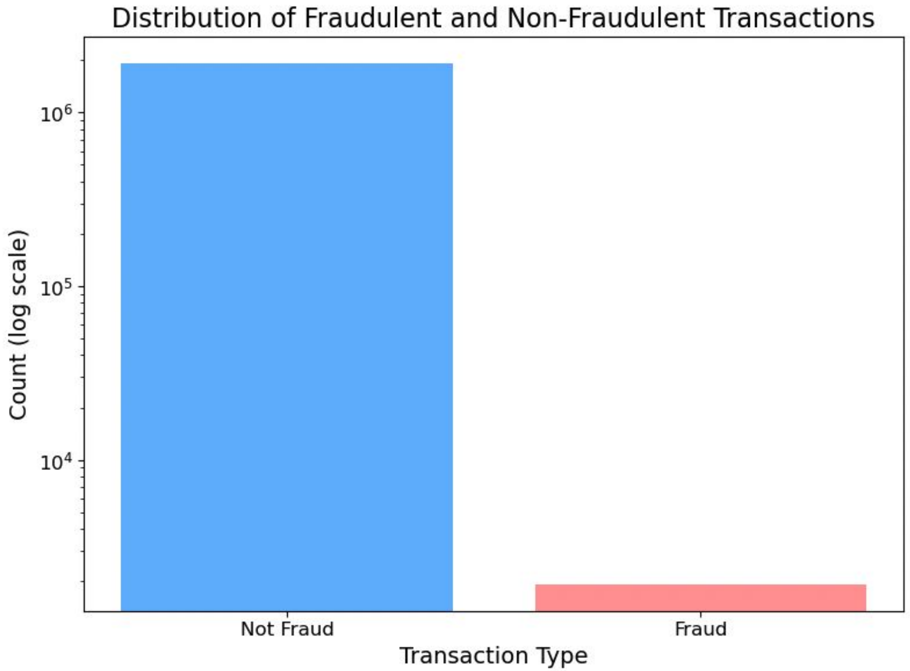
Mobile Phone Transactions Fraud Detection

Chujun Chen
Brown Data Science Institute
Dec, 4th 2023
Github:https://github.com/Christina-Chen01/DATA1030-FinalProject-Fraud-Detection

Classify a Transaction Based on Historical Transaction Patterns and Features

- Importance:
  - Ensure Financial Security
  - Prevent Financial Loss
  - Maintain trust in mobile platforms among users.
- Characteristics:
  - 100K+ records
  - Non-iid (time series)
- Data Source: PaySim synthetic dataset on Kaggle
- Data Collection:
  - Simulates real transactions from a global mobile financial service provider.

EDA



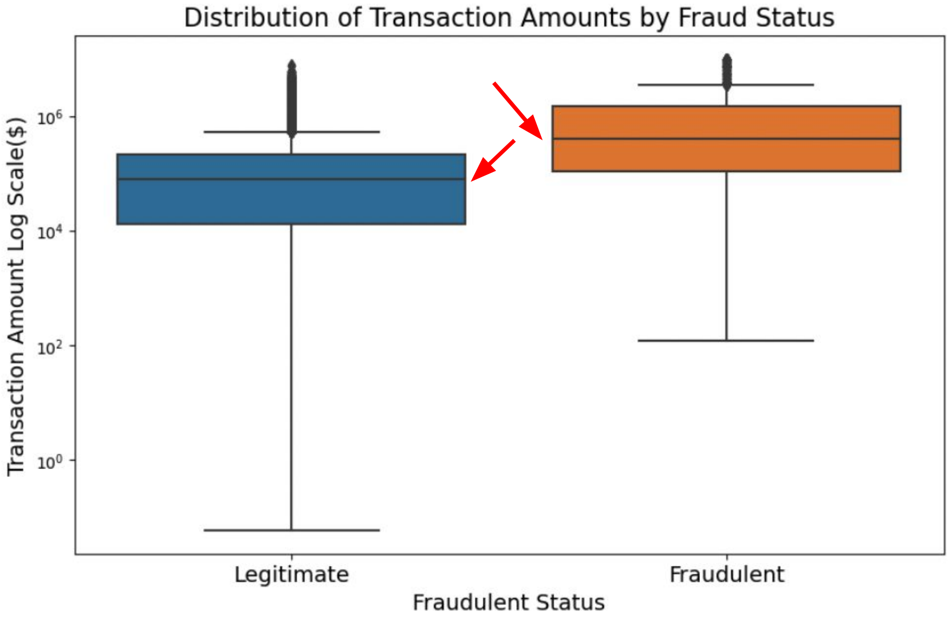


Figure 1. Highlights the imbalance in the 'isFraud' target variable, where 0.018% of transaction is labeled as fraudulent

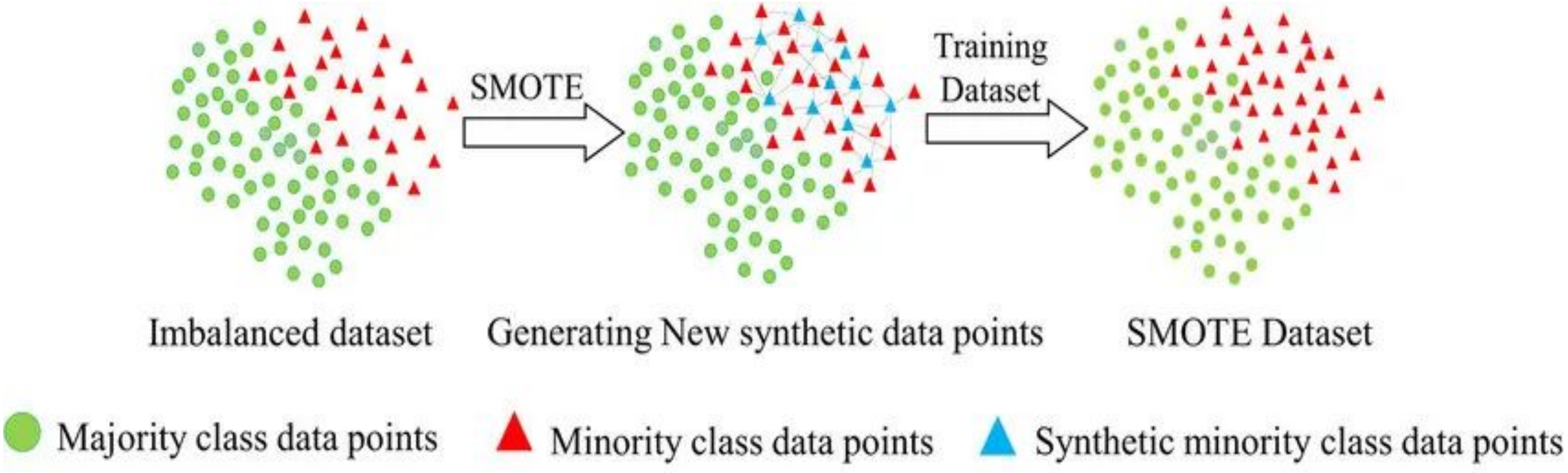
Figure 2. highlights a higher median of transaction amount for fraudulent transactions

Data Splitting & Preprocessor

- TimeSeriesSplit (*n_split = 4*):
  - Preserves chronological order
  - Prevents future data leakage
- OneHotEncoder & StandardScaler

- Synthetic Minority Over-sampling Technique (SMOTE)
  - Balance the class distribution by creating synthetic examples of the minority class

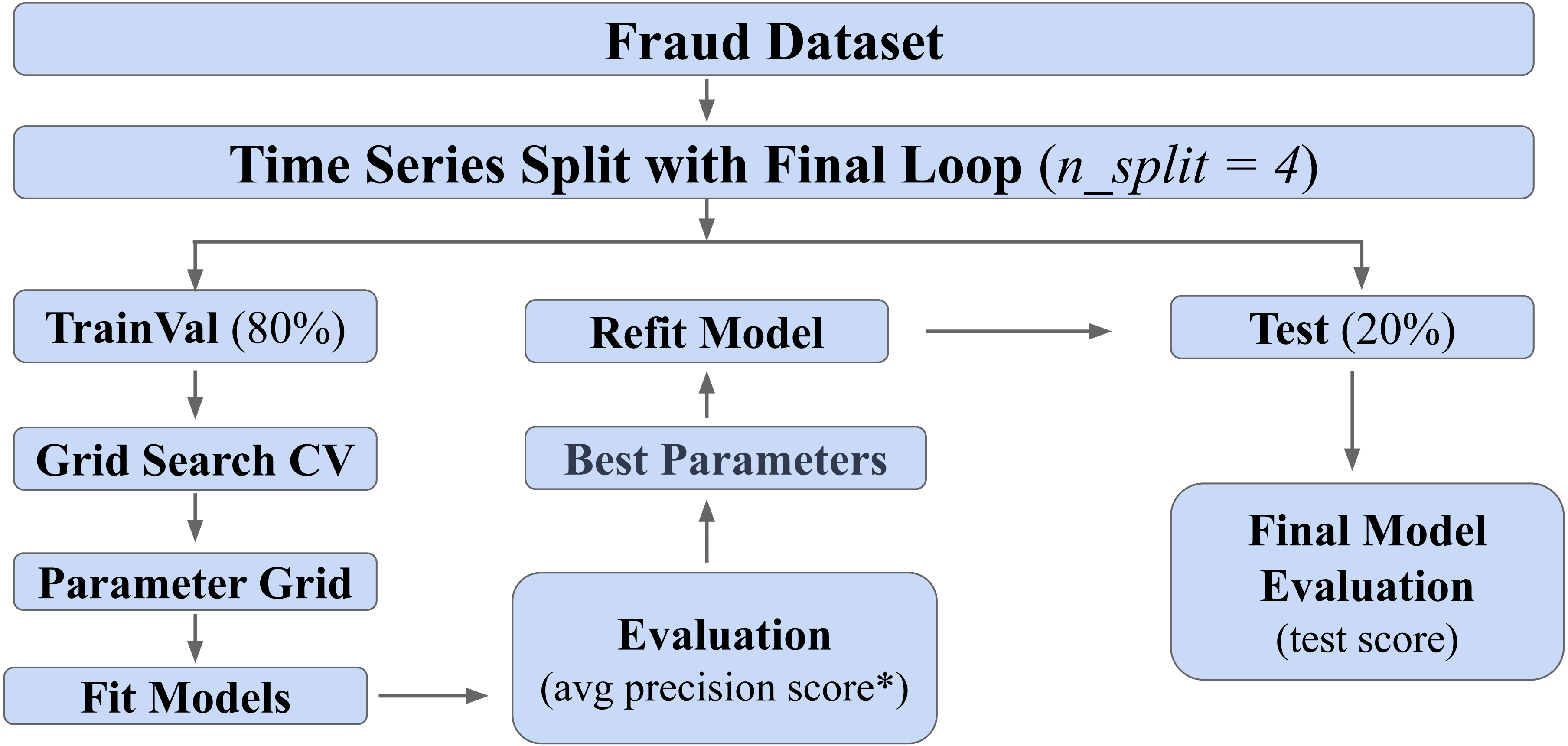
Data Splitting & Preprocessor



- Synthetic Minority Over-sampling Technique (SMOTE)
    - Balance the class distribution by creating synthetic examples of the minority class

ML Model	Hyperparameter	Values
Logistic Regression	C penalty solver **class_weight**	0.001, 0.01, 0.1, 1, 10, 100, 1000 l1, l2 saga balanced
Random Forest Classifier	n_estimators max_depth max_features **class_weight**	25, 50, 100, 200 10, 20, 30, None None, sqrt balanced, balanced_subsample
KNeighbors Classifier	n_neighbors **weights**	5, 10, 15, 20 distance
XGBoost Classifier	**scale_pos_weight** learning_rate reg_alpha	weight = (y == 0).sum() / (1.0 * (y == 1).sum()) 0.01, 0.03 0.01, 1, 100

ML Models and their Corresponding Hyperparameters



ML Pipelines

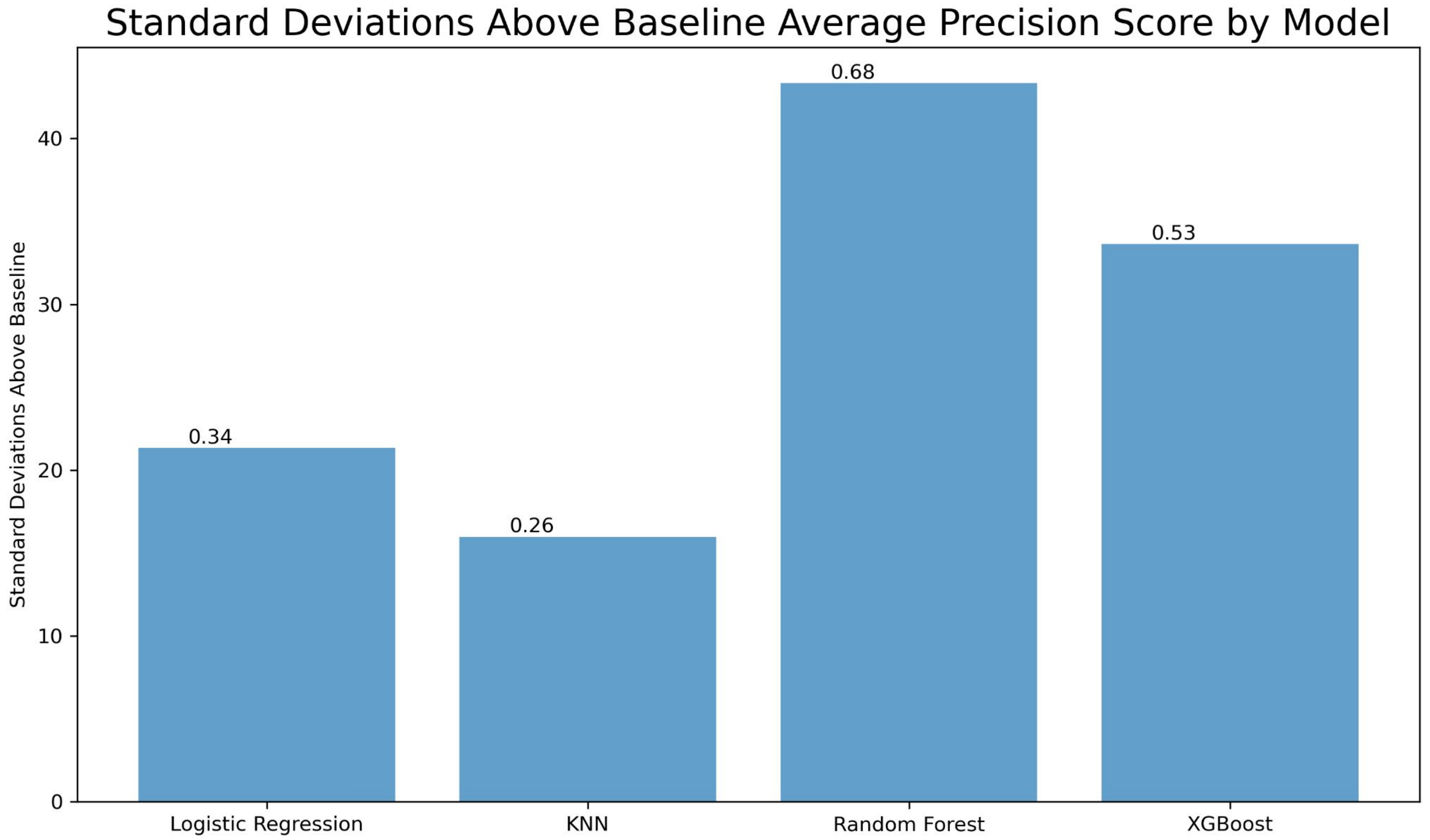
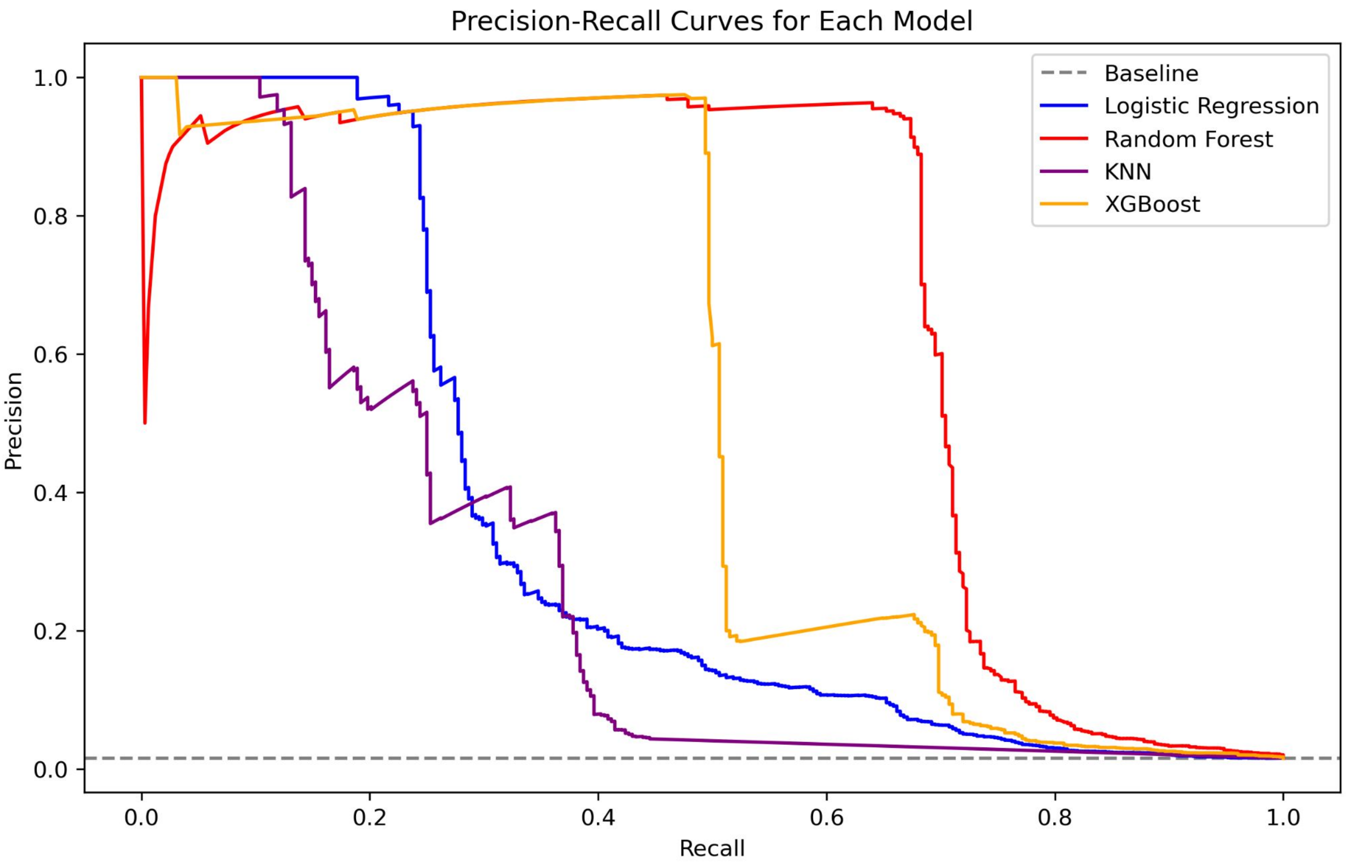


Figure 3. displays the standard deviations above the baseline performance and the actual test score labeled above each bar for four predictive models.

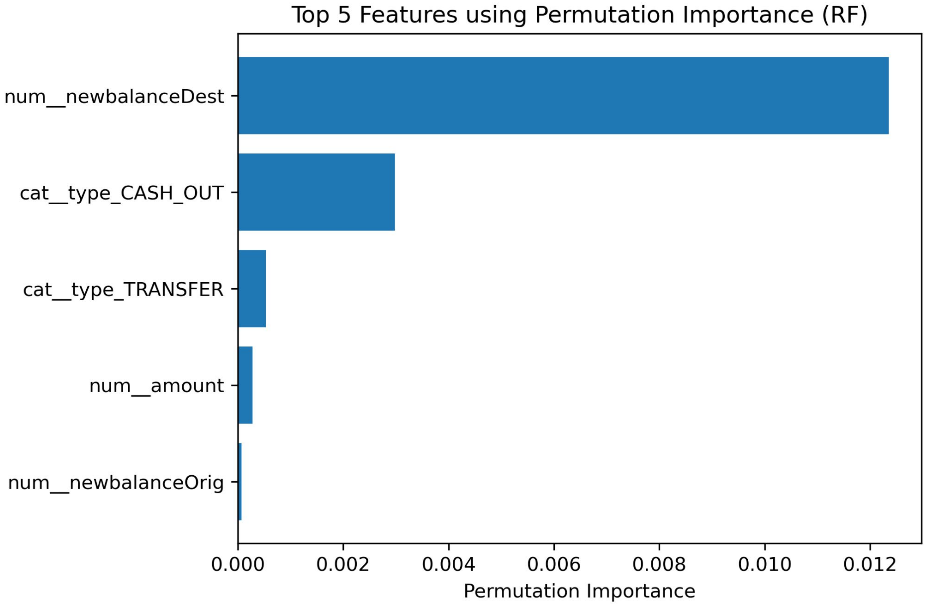


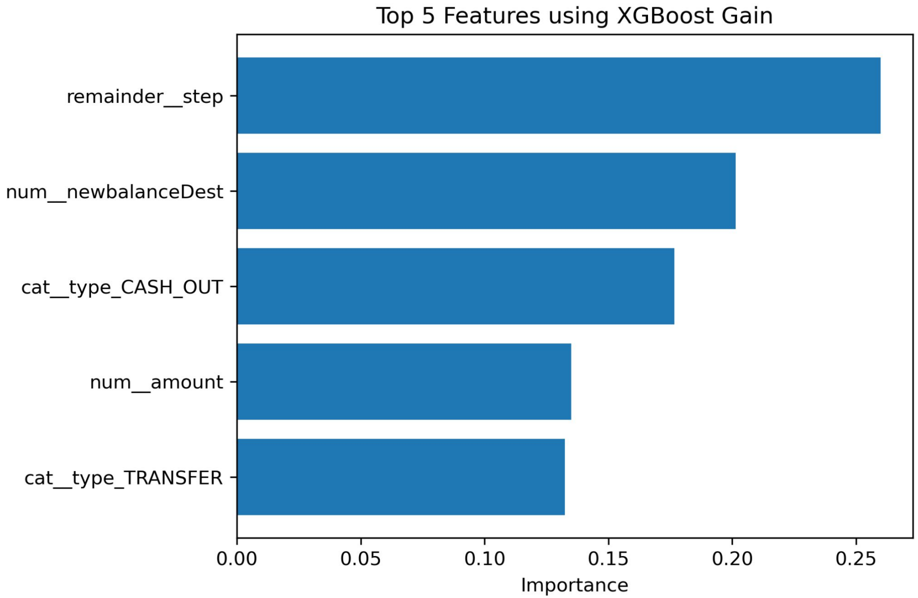
- Higher area under the curve represents both high recall and high precision.

Baseline in precision-recall Curve

Figure 4. presents precision-recall curves for four different models, demonstrating their trade-offs between precision and recall.

Interpretability (Global & Local Features Importances)





Outlook

- Alternative Techniques for Imbalanced Data
  - SMOTE might be misleading, i.e. high false positive
  - Adaptive Synthetic Sampling (ADASYN) or Tomek Links to refine the way synthetic samples are generated

- Feature Engineering and Selection
  - Create new features that might capture fraud patterns
  - Eg: ‘is_weekend’, ‘transaction_location_frequency’, etc.

Thanks!