Machine Learning - Assignment 3

Assignment on handling imbalanced datasets in credit card fraud detection using the "**Credit Card Fraud Detection**" dataset from Kaggle:

Task 1. Data Exploration and Preprocessing

1. Load the dataset and explore its structure.
2. Check for missing values and handle them (e.g., by imputation or removal).
3. Visualize the distribution of classes (fraudulent vs. non-fraudulent transactions).

Task 2. Baseline Model Training

1. Split the dataset into training and testing sets.
2. Train a baseline model using Logistic Regression without addressing the class imbalance.
3. Evaluate the model using metrics like accuracy, precision, recall, F1-score, and ROC-AUC.

Task 3. Handling Class Imbalance

1. Apply oversampling techniques (e.g., SMOTE) to balance the classes in the training set.
2. Train the model again using Logistic Regression on the balanced dataset.
3. Evaluate the model and compare the results with the baseline model.

Task 4. Advanced Techniques and Final Evaluation

1. Apply a combination of oversampling and undersampling techniques to further balance the classes.
2. Train a more complex model (e.g., Random Forest or XGBoost) on the balanced dataset.
3. Evaluate the model using the same metrics as before.

Task 5. Summary of Findings

1. Summarize the results and discuss the impact of different techniques on model performance.

Deliverables

A Jupyter Notebook containing all the above steps.