SKCET

Name: Rudhramani K

Regno: 727723eucs195

Boosting for Imbalanced Datasets with XGBoost

Problem Statement

Handling imbalanced datasets is a major challenge in machine learning. Traditional algorithms such as SVM and Random Forest often perform poorly when one class significantly outnumbers the other. Boosting techniques like XGBoost improve classification by focusing more on misclassified and minority class samples. This project applies XGBoost with imbalance handling techniques to enhance predictive performance.

Objectives

• Apply XGBoost for classification tasks.

• Address class imbalance using SMOTE and weighted loss functions.

• Tune hyperparameters such as learning rate, max depth, and n_estimators.

• Evaluate performance using metrics suited for imbalanced data, including Precision-Recall and ROC-AUC.

Methodology

1. Data Preprocessing: Replaced invalid zero values with NaN and filled them using median imputation.

2. Feature Engineering: Created interaction features such as Glucose_BMI, Age_BMI, and Insulin_Glucose.

3. Handling Class Imbalance: Applied SMOTE to oversample the minority class.

4. Model Training: Trained XGBoost with tuned hyperparameters.

5. Evaluation: Used Precision-Recall curves, ROC-AUC, and classification report.

XGBoost Model Implementation

The XGBoost classifier was trained using parameters like n_estimators=600, learning_rate=0.05, max_depth=3, and scale_pos_weight to handle imbalance. The model was trained on SMOTE-resampled data to improve recall and precision for the minority class.

Class Imbalance Handling

• SMOTE (Synthetic Minority Over-sampling Technique) was used to balance the dataset.

• scale_pos_weight in XGBoost adjusted the loss function to emphasize minority class samples.

Performance Evaluation

The model achieved strong results using metrics suited for imbalanced data:

• ROC-AUC: Measures class separation ability.

• Precision-Recall AUC: Focuses on minority class performance.

• Classification Report: Shows precision, recall, and F1-score.

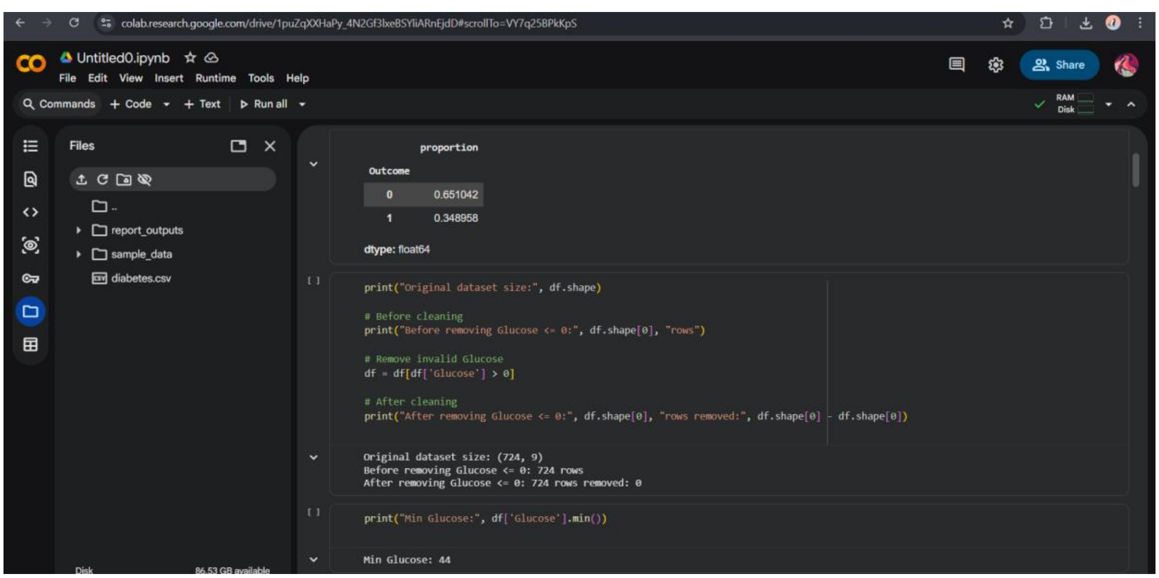
Outcomes

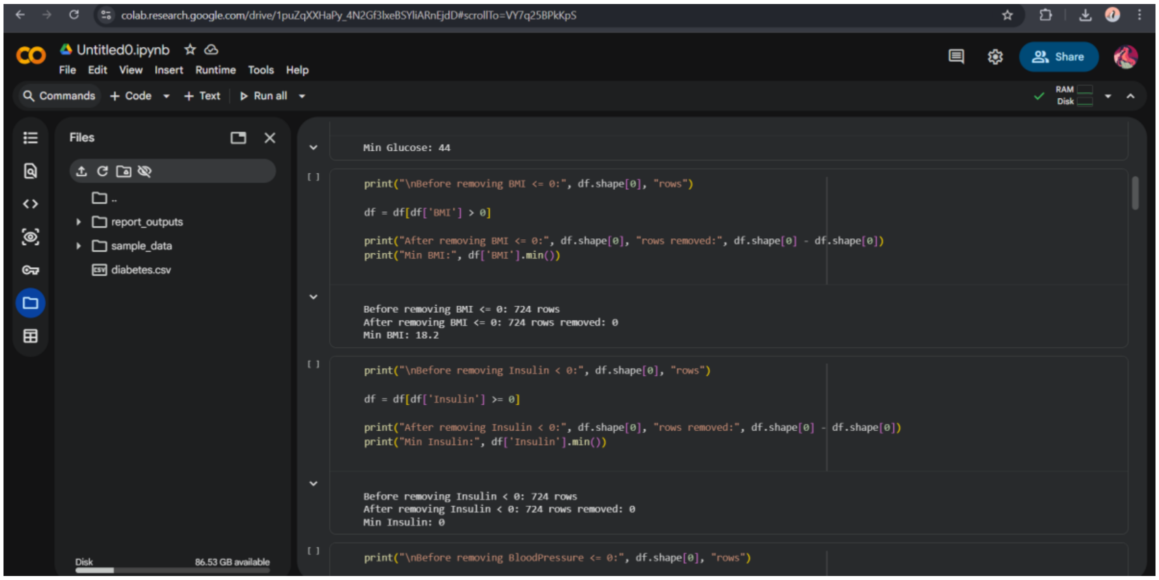
An optimized XGBoost model capable of handling imbalanced datasets effectively. Performance metrics demonstrate improved detection of the minority class with balanced precision and recall.

Conclusion

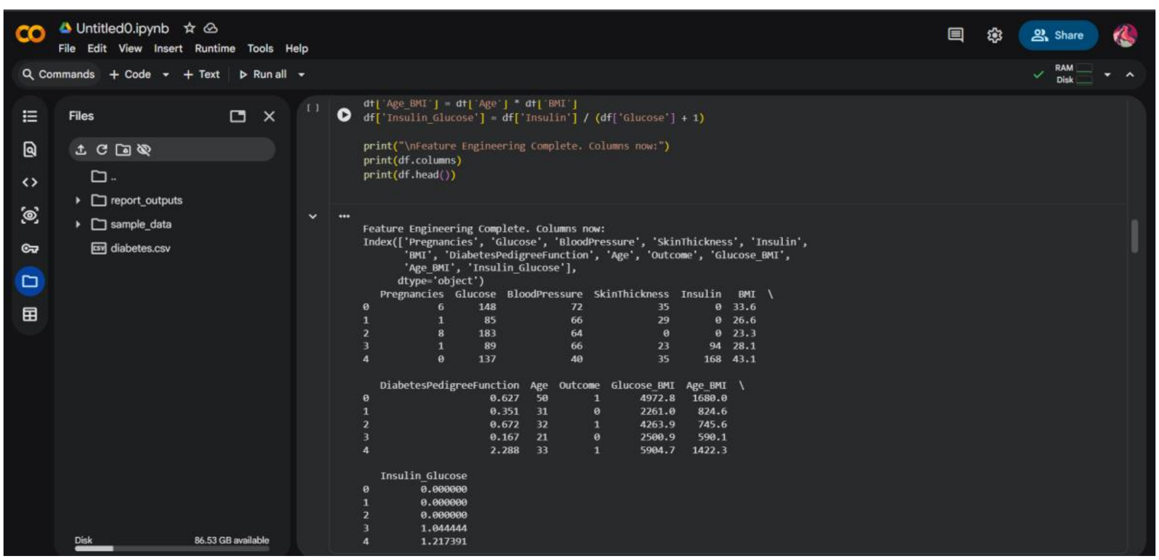
Boosting with XGBoost combined with SMOTE and proper hyperparameter tuning significantly improves classification on imbalanced datasets. Precision-Recall curves and ROC-AUC confirm the model's effectiveness in identifying minority class samples.

Data cleaning:

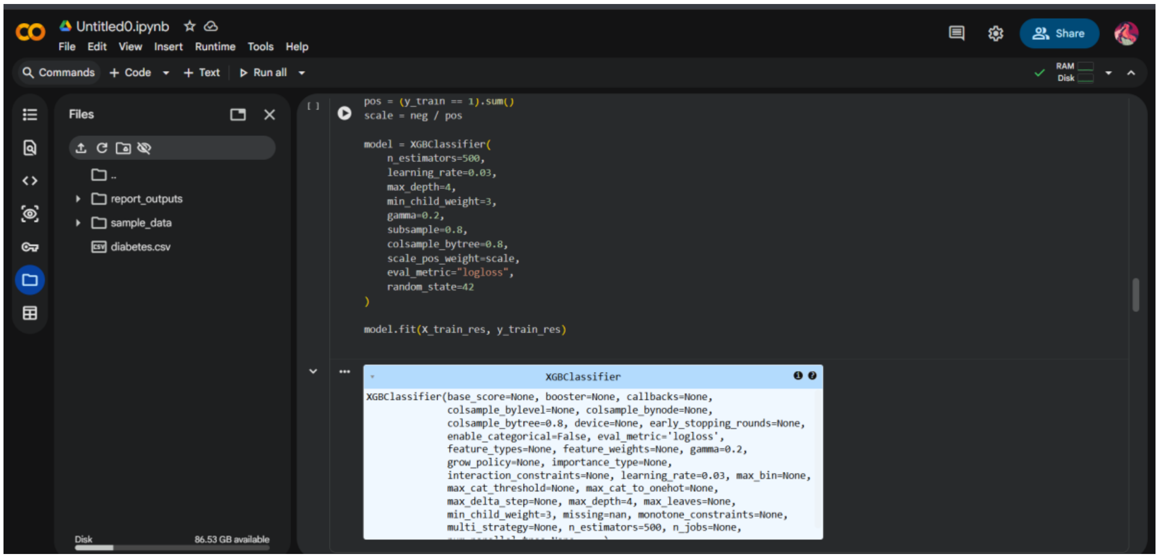




Feature Eng:



XGBoost model:



Final Report: