Predicting Hospital Readmission for Diabetic Patients

Repository:
[https://github.com/subu53/Predicting-Hospital-Readmission-for-Diabetic-Patients]

Project Description:

This project focuses on predicting hospital readmission for diabetic patients within 30 days of discharge using a dataset from 130 US hospitals. The goal is to develop machine learning models to identify patients at high risk of early readmission, which is a significant challenge in healthcare impacting patient outcomes and costs.

The dataset, sourced from the "Diabetes 130-US hospitals for years 1999–2008," contains over 100,000 patient records with diverse features including demographics, encounter details, diagnoses, procedures, medications, lab results, and prior visit history.

Key Challenges Addressed:

- **Severe Class Imbalance:** The target variable (readmission within 30 days) is a minority class.
- **High Feature Cardinality:** Many categorical features have a large number of unique values.
- **Missing and Inconsistent Data:** Handling '?' values and other data inconsistencies.

Project Objectives:

- Implement a robust data preprocessing and cleaning pipeline.
- Perform feature engineering and dimensionality reduction to create a suitable feature set.
- Develop and evaluate classification models for predicting 30-day readmission.
- Apply techniques to address the class imbalance issue.
- Evaluate models using appropriate metrics (F1-score, Recall, Precision, ROC AUC) for imbalanced data.
- Discuss model interpretability and identify areas for future improvement.

Key Steps Taken:

- Loaded and performed initial data inspection.
- Handled missing values (dropping columns, replacing '?').
- Dropped identifier columns (encounter_id, patient_nbr).
- Grouped rare categories in high-cardinality nominal features (medical_specialty,

payer_code, discharge_disposition_id).
- Grouped rare individual diagnosis codes (diag_1, diag_2, diag_3).
- Engineered aggregate medication features (num_active_meds, num_med_changes) and selectively kept/dropped individual medication columns based on frequency.
- Engineered prior visit features (total_prior_visits, has_outpatient_prior, etc.) and dropped original count columns.
- Mapped binary and ordinal features to numerical values (diabetesMed, change, age).
- Performed one-hot encoding on remaining nominal features.
- Defined the binary target variable (readmitted_within_30_days).
- Split the data into training and testing sets using stratified sampling.
- Trained and evaluated baseline models (Logistic Regression, LightGBM, Random Forest) with techniques like class_weight and SMOTE to handle imbalance.

Models Explored:

- Logistic Regression
- LightGBM (LGBMClassifier)
- Random Forest (RandomForestClassifier)

Evaluation Metrics:

- Precision, Recall, F1-score (for the minority class)
- ROC AUC Score
- Confusion Matrix

Key Findings & Conclusion:

The project successfully demonstrates a practical approach to tackling a real-world imbalanced classification problem. While the initial models achieved modest performance (minority class F1-scores around 0.27-0.28), the refined data preparation and use of imbalance handling techniques showed improvements over a naive approach. The results highlight the inherent difficulty of this prediction task. Further optimization through hyperparameter tuning, exploring other advanced models, and potentially more sophisticated feature engineering are necessary steps to build a more highly predictive model.

Future Work:

- Complete hyperparameter tuning for LightGBM and other models.
- Implement and evaluate XGBoost and CatBoost classifiers.
- Explore advanced resampling and ensemble methods from imblearn.

- Investigate classification threshold tuning.
- Perform additional feature engineering.
- Consider feature selection techniques.