Patient Readmission Prediction Model Documentation

1. Project Overview

Objective:

The goal of this project is to build a predictive model that identifies patients at high risk of readmission. Reducing patient readmission rates is critical for improving healthcare quality and reducing costs, and this model aims to assist healthcare providers in early intervention.

Dataset:

The dataset includes patient demographics, medical history, hospital stay details, and various other clinical features. The target variable indicates whether a patient was readmitted within a certain period after discharge.

2. Data Description

Key Features:

Demographics: Age, gender, race

Clinical Data: Time in hospital, number of inpatient visits, number of emergency visits, medications, discharge disposition.

Target Variable: Binary variable indicating readmission status (1 = readmitted, 0 = not readmitted).

Data Summary: The dataset contains a mix of numerical and categorical data, requiring preprocessing to make it suitable for model training.

3. Data Preprocessing

Steps:

Missing Value Treatment:

Imputed missing values based on feature type (e.g., mean/median for numerical features, mode for categorical features).

Encoding Categorical Variables:

Converted categorical features into numerical format using techniques like one-hot encoding for compatibility with machine learning algorithms.

Feature Scaling:

Applied StandardScaler to standardize numerical features, ensuring that features with different scales do not disproportionately affect the model.

Handling Class Imbalance:

Implemented resampling methods to handle class imbalance, particularly targeting models sensitive to imbalance.

4. Exploratory Data Analysis (EDA)

Key Insights:

Correlation Analysis:

Identified highly correlated features to ensure no redundancy.

Feature Distributions:

Examined distributions of key features such as time_in_hospital and number_inpatient to understand typical patient profiles.

Target Distribution:

Confirmed class imbalance with more non-readmitted cases than readmitted ones.

5. Model Selection

Models Evaluated:

Logistic Regression: Chosen as a baseline model for its simplicity and interpretability.

Decision Tree: Provides an interpretable model structure with decision rules.

Random Forest: An ensemble method that improves accuracy by averaging over many decision trees.

XGBoost: An advanced boosting technique that often performs well on structured data.

Hyperparameter Tuning:

Used GridSearchCV to find the optimal hyperparameters for each model, balancing parameters like max_depth in trees and learning_rate in boosting algorithms.

6. Model Evaluation

Evaluation Metrics:

Accuracy: Provides an overall measure of correct predictions but is less informative for imbalanced classes.

ROC-AUC: Measures the model's ability to distinguish between readmitted and non-readmitted patients, used to compare model performance.

Precision and Recall:

Precision: Indicates the proportion of true readmissions out of all predicted readmissions.

Recall: Emphasized to reduce false negatives and avoid missing high-risk patients.

Confusion Matrix Analysis:

Examined true positives, false negatives, and false positives to understand misclassification patterns and improve recall.

Model Performance:

After comparison, [final chosen model, e.g., Random Forest] was selected for its high ROC-AUC and balanced precision-recall tradeoff.

7. Feature Importance and Interpretability

Feature Importance:

Analyzed feature importance in the final model to identify key predictors, such as time_in_hospital, number_inpatient, and number_emergency.

Visualized feature importance to provide insights into the main factors affecting readmission risk, offering transparency and interpretability.

8. Insights and Recommendations

Insights:

High-Risk Indicators: Patients with multiple inpatient visits, longer hospital stays, and frequent emergency visits are at a higher risk of readmission.

Model Confidence: The model demonstrates a strong ability to identify high-risk patients, aiding in proactive healthcare intervention.

Recommendations:

Real-World Application: Deploy the model in a clinical setting to prioritize high-risk patients for follow-up care.

Continuous Training: Regularly update the model with new data to maintain accuracy over time.

Patient Interventions: Use model predictions to guide targeted interventions, such as post-discharge follow-up and specialized care plans.

9. Future Work

Enhanced Feature Engineering: Explore additional derived features, such as risk scores, to improve prediction.

Addressing Bias: Assess and address any potential biases, ensuring fair predictions across different demographic groups.

Advanced Model Interpretability: Utilize interpretability techniques like SHAP values for deeper insights into individual predictions.

10. Appendix

Codebase:

The code used to preprocess data, build and evaluate models, and visualize results is contained in the project notebook.