Diabetes Patient Readmission Prediction Using Machine Learning

Executive Summary

Hospital readmissions within 30 days are a critical healthcare quality and cost-control challenge. This project focuses on building a robust, industry-aligned machine learning pipeline to predict early readmission risk among diabetic patients. The solution emphasizes interpretability, scalability, and performance under class imbalance, making it suitable for real-world healthcare deployment.

Business Problem Statement

Unplanned readmissions significantly increase operational costs and indicate suboptimal patient outcomes. Healthcare providers require data-driven tools to proactively identify high-risk patients at discharge. The objective of this project is to predict whether a patient will be readmitted within 30 days, enabling targeted intervention, optimized resource allocation, and improved care quality.

Dataset Overview

The dataset consists of historical diabetic patient records, including demographics, clinical test results, treatment details, admission characteristics, and prior utilization metrics. The target variable is binary, indicating readmission within 30 days versus no early readmission.

Data Preprocessing & Feature Engineering

Data preprocessing included systematic handling of missing values, removal of high-null or low-business-value features, and domain-driven transformations. Age categories were converted into numerical midpoints for model compatibility. Categorical variables were encoded using one-hot encoding to preserve information without introducing ordinal bias.

Handling Class Imbalance

Early readmission events are relatively rare, resulting in a highly imbalanced target distribution. Synthetic Minority Oversampling Technique (SMOTE) was applied within a modeling pipeline to ensure balanced learning while preventing data leakage. This approach significantly improved recall and ROC-AUC stability.

Modeling Approach

Multiple classification algorithms were evaluated, including Logistic Regression, Random Forest, Gradient Boosting, Naive Bayes, and K-Nearest Neighbors. Model performance was assessed using Stratified K-Fold cross-validation with ROC-AUC as the primary metric to ensure robustness across class distributions.

Model Selection & Hyperparameter Tuning

Random Forest emerged as the top-performing and most stable model. Hyperparameters such as tree depth, minimum samples per leaf, and number of estimators were tuned to balance bias, variance, and interpretability. Class weighting was incorporated to further address imbalance at the algorithm level.

Threshold Optimization

Rather than relying on the default probability threshold, a custom decision threshold was applied to prioritize recall. This aligns with healthcare business objectives where missing a high-risk patient is costlier than a false alert. The final threshold achieved a strong balance between sensitivity and overall discrimination power.

Model Evaluation

Model performance was evaluated using ROC-AUC, precision, recall, F1-score, and confusion matrix analysis. ROC curves demonstrated consistent generalization across folds, while feature importance analysis provided transparency into the key drivers of readmission risk.

Key Insights & Feature Importance

Feature importance analysis highlighted clinically meaningful predictors such as number of prior admissions, length of stay, medication changes, and lab test outcomes. These insights can support clinical decision-making and guide targeted care management programs.

Business Impact & Industry Relevance

This solution can be integrated into hospital discharge workflows to flag high-risk patients in real time. Potential benefits include reduced readmission penalties, improved patient outcomes, optimized staffing, and data-backed care coordination strategies. The pipeline is scalable and adaptable to other chronic disease domains.

Conclusion

This project demonstrates an end-to-end, production-oriented machine learning workflow tailored for healthcare analytics. By combining strong preprocessing, imbalance handling, model tuning, and business-aligned evaluation, the solution reflects real-world industry expectations rather than academic experimentation.