Predicting Early Readmissions of Diabetic Patients: A CRISP-DM Approach

October 1, 2023

AVI AJMERA
SAN JOSE STATE UNIVERSITY

1 Abstract

This study employs the CRISP-DM methodology to address a critical healthcare challenge: predicting early readmissions of diabetic patients within 30 days of discharge. Utilizing a comprehensive dataset, we developed predictive models to identify associated risk factors. While our models achieved promising accuracy, challenges in the precision-recall trade-off were evident, underscoring the intricacies of medical data and the importance of continuous model refinement.

2 Introduction

2.1 Background

The healthcare sector continuously strives to improve patient outcomes, reduce costs, and enhance overall efficiency. A particularly pressing concern is the prevalent rate of diabetes in the U.S. and the consequent hospital readmissions. By identifying factors leading to early readmission, hospitals can optimize care procedures, ensuring better patient outcomes and reducing the financial implications of readmissions.

2.2 CRISP-DM Methodology

The Cross-Industry Standard Process for Data Mining (CRISP-DM) offers a structured approach to tackle such data-driven challenges. This study leverages the CRISP-DM framework, encompassing Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment.

3 Methodology

3.1 Phase 1: Business Understanding

3.1.1 Problem Definition

The arbitrary nature of diabetes management in hospitals often leads to increased readmission rates and potentially worsened patient outcomes. This project aims to predict early readmissions and comprehend the major contributing factors.

3.1.2 Stakeholder Identification

The primary beneficiaries include hospital administrators, healthcare policymakers, and medical researchers.

3.2 Phase 2: Data Understanding

3.2.1 Data Collection

The data, sourced from the hospital's database, provides comprehensive patient information, including demographic details, medical history, and hospital stay specifics.

3.2.2 Data Description

Preliminary data exploration revealed various features, with some missing values and potential outliers.

3.3 Phase 3: Data Preparation

3.3.1 Data Cleaning

We addressed missing values using imputation techniques and filtered out outliers based on domain knowledge.

3.3.2 Feature Engineering

New features were derived from existing ones to enhance the model's predictive power.

3.4 Phase 4: Modeling

3.4.1 Model Selection Criteria

We employed a variety of models, emphasizing interpretability, handling of imbalanced data, and overall performance.

3.4.2 Model Training and Evaluation

Gradient Boosting, Random Forest, and Logistic Regression were the primary contenders. The models were trained on a training dataset and subsequently evaluated on a test dataset.

3.5 Phase 5: Evaluation

3.5.1 Model Performance Metrics

While all models showcased satisfactory accuracy, the precision-recall trade-off, especially recall, presented challenges.

3.5.2 Test Case Demonstrations

A selection of test cases further illustrated the model's predictions, reinforcing the challenges in recall.

4 Discussion

Gradient Boosting emerged as the best-performing model, excelling in handling imbalanced data and capturing intricate data patterns. However, the model's recall remains an area of concern, emphasizing the need for hyperparameter tuning, advanced resampling, or exploration of complex models like neural networks.

5 Conclusion

Predicting early readmissions of diabetic patients within 30 days remains a formidable challenge. Despite achieving decent accuracy, the models grapple with the precision-recall trade-off. Addressing this is paramount, especially in medical contexts where the consequences of false negatives are profound.