Predictive Modeling for Hospital Readmission

Name: Richa Sharma
Student ID: 23037468
Project Link

1 Introduction

1.1 Background

Diabetes requires careful monitoring during hospitalization to reduce readmission risks. This study predicts whether diabetic patients may return within 30 days post-discharge. The dataset covers 10 years (1999-2008) from 130 US hospitals, focusing on diabetic diagnoses. Accurate readmission predictions promise significant improvements for these patients.

1.2 Dataset Overview

The dataset used spans a decade (1999-2008), gathering data from 130 US hospitals. It comprises 50 features and 101,766 samples, encompassing patient demographics, medical history, medications, and diagnostic codes. Categorical features include race, gender, and medical procedures, while continuous features cover attributes like hospital stay duration, lab procedures, and medications. Each entry relates to diabetic patients' hospital records, aiming to predict their likelihood of returning to the hospital within 30 days post-discharge.

Table 1. Small snippet of actual data

encounter_id	patient_nbr	race	gender	age	medical_specialty	diag_1	readmitted
2278392	8222157	Caucasian	Female	[0-10)	Pediatrics-Endocrinology	250.83	NO
149190	55629189	Caucasian	Female	[10-20)	?	276	>30

1.2.1 Readmission categorization

In the dataset, there are three distinct classes used for readmission categorization:

- '<30' indicates patients who were readmitted in less than 30 days.
- '>30' represents patients who were readmitted in more than 30 days.
- 'No' signifies no record of readmission.

2 Exploratory Data Analysis

Initial analysis identified categorical and continuous features, revealing a high incidence of missing 'weight' values. Further examination through violin plots illuminated distributions across significant continuous variables, uncovering insights into patient stay durations, lab procedures, and medication frequencies.

2.1 Medication Counts Across Age Groups by Readmission

The grouped bar chart shows medication distributions by age groups for readmitted and non-readmitted patients. It consistently reveals higher medication counts among readmitted patients across all age groups, suggesting a possible link between increased medications and higher readmission rates.

2.2 Imbalanced Classes and Data Insights:

Preliminary exploration involved assessing imbalanced classes. Moreover, violin plots elucidated distributions of major continuous variables like 'admission_type_id', 'time_in_hospital', 'num_lab_procedures', and others, revealing their respective distributions and potential outliers.

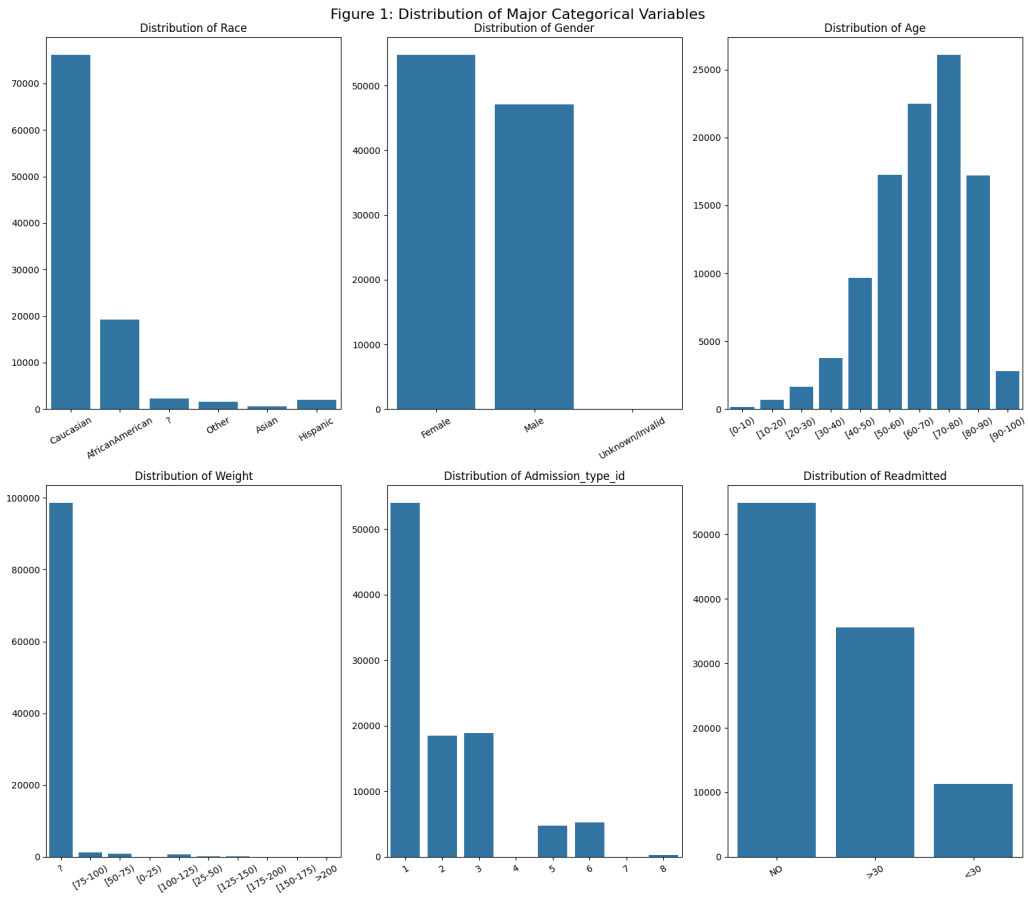

Figure 1. Distribution of Major Categorical Variables

3 Data Cleaning and Pre-processing

Data preparation involved dropping columns with substantial missing values `'weight'`, `'payer_code'`. Encoding categorical variables e.g. `'diag_1'`, `'diag_2'`, `'diag_3'` and scaling continuous variables were executed using pipelines and column transformers. We use StandardScaler to make continuous variables similar in range. It's a usual step before using machine learning. With categorical variables, we fill in missing values with the most common option, then turn them into numbers (one-hot encoding) or order them if it makes sense (ordinal encoding).

4 Machine Learning Model Evaluation and Hyperparameter Tuning

4.1 Model Selection and Performance Metric Justification

The F1 score we use checks for things we missed in our guesses, especially the things we got right. It puts together two things: how exact we are and how much we find of the actual correct stuff.
We run our analysis comparing three models:

- K Nearest Neighbors Classifier
- Logistic Regression Classifier
- Random Forest Classifier

Hyperparameter tuning via GridSearchCV aimed to optimize model parameters for improved predictive performance. Cross-validation strategies (Stratified K-Fold) provided robust estimations across different k-fold values.

4.2 Model Performance Analysis

In our performance summary, the second table displays 3-fold cross-validation results. The random forest classifier scored the highest F1 at 0.623, followed closely by logistic regression at 0.603 and KNN at 0.588. Expanding to 7-fold cross-validation in the third table, logistic regression slightly outperformed random forest with an F1 of 0.615 versus 0.607 and 0.575 for random forest and KNN, respectively. Precision, recall, and accuracy follow similar trends. Overall, random forest and logistic regression show strong performance, with random forest excelling under 3-fold CV and logistic regression under 7-fold. KNN consistently lags across metrics and cross-validation strategies. These insights will aid in selecting and evaluating the final model using hold-out test data.

Table 2. Results for 3-fold CV

Fold	Model	precision	recall	f1	accuracy
3	RandomForestClassifier	0.637933	0.617065	0.622894	0.617065
3	LogisticRegression	0.610920	0.599443	0.602998	0.599443
3	KNeighborsClassifier	0.596609	0.584442	0.588279	0.584442

Table 3. Results for 7-fold CV

Fold	Model	precision	recall	f1	accuracy
7	LogisticRegression	0.628441	0.609499	0.614927	0.609499
7	RandomForestClassifier	0.614939	0.603472	0.606999	0.603472
7	KNeighborsClassifier	0.579792	0.572028	0.574666	0.572028