A Case Study of U.S. Hospitals

Hospital Readmission Prediction for Diabetes Patients

Group 2

ROSSILLON Timothée, FAOUZI Walid, SHAN Jiameng,

NDAW Abdoulaye, THEODORE Mary Sheeba

Abstract

1. Introduction

The dataset used in this analysis (*Diabetes 130-US Hospitals for Years 1999-2008,* from UCI Machine Learning Repository) comes from the field of health and medicine and focuses on inpatients diagnosed with diabetes. Diabetes readmissions present a critical challenge for healthcare systems. This study develops a machine learning model to predict 30-day readmissions using patient data from 130 U.S. hospitals between 1999 and 2008. It contains 101,766 instances (patient records) and 48 features. The dataset is multivariate, with both categorical and numerical variables representing various medical and demographic information about the patients.

Attributes include information such as race, gender, age, type of admission, time spent in hospital, attending physician's medical specialty, number of laboratory tests performed, HbA1c test results, diagnosis, number of prescribed medications, diabetes medications, as well as information about outpatient, inpatient and emergency visits prior to hospitalization.

The main purpose of this dataset is to predict diabetes-related categories, which makes it a classification problem. However, the dataset also contains missing values, which requires appropriate management during pre-processing. The data is divided into 11 numerical variables (int64 type) and 37 categorical variables (object type).

2. Methods

The first process was to load the data directly from the UCI Machine Learning Repository using the dedicated library. In the initial step of our analysis, we performed an Exploratory Data Analysis. This allowed us to understand the structure of the dataset and gain insights into the distribution of the variables, the presence of missing values and other essential characteristics of the data. Then we conducted univariate analysis to visualize the distribution of each feature. This step was crucial for understanding the nature of the variables, whether categorical or continuous, and allowed us to spot any skewness or outliers. We found that some features had only one modality, or that some modalities were present for only very few individuals. Following that, we performed a bivariate analysis to investigate the relationships between pairs of variables. This helped us identify potential correlations that could inform the model-building process.

After the exploratory data analysis, the data cleaning process began.

As part of our project to analyze data on diabetic patients, a thorough cleaning of the dataset was undertaken. The aim of this process was to reduce the number of modalities and improve feature consistency, while ensuring that the data was usable for our machine learning models.

A copy of the original dataset was created to preserve the raw data. Next, we identified and removed several columns considered irrelevant or unusable for our analysis, such as ‘weight’ (97% empty), ‘payer_code’, and ‘medical_specialty’, as well as various drugs that did not provide significant information. This reduction was guided by observations during the exploratory analysis of the data and was confirmed by the subsequent use of feature importance.

To ensure consistency within the dataset, manual encoding of the categorical variables was used. For example, for the race variable, we grouped the minority modalities under a single ‘Other’ category and assigned numerical values to the different modalities. Similarly, the gender variable was recoded in binary (0 for female and 1 for male), while age was transformed using the midpoints for each age bracket. We also dealt with missing values in critical columns such as ‘max_glu_serum’ and ‘A1Cresult’, by

replacing these values with meaningful codes. Diagnoses (diag_1, diag_2, diag_3) were cleaned up by replacing invalid values. The other categories were encoded by converting the character string into a float and then reducing the number of modalities in accordance with ICD-9 codes. The variables 'admission_type_id', 'discharge_disposition_id' and 'admission_source_id' were grouped together using mapping files provided by the UCI site. This made it possible to simplify these characteristics while retaining their relevance for the analysis. After these steps, the dataset was reduced to 29 features that could be used with the target. The modifications improved the quality and relevance of the data while reducing the potential noise from infrequent or uninformative modalities.

Initially, the data was standardized using StandardScaler. After applying principal component analysis (PCA) and the Kaiser criterion to a dataset of 28 columns. PCA, while retaining 95% of the variance, reduced the columns to 25, a reduction deemed insignificant. Furthermore, the Kaiser criterion led to an excessive loss of variance, retaining only 12 components that explained only 60% of the variance. The following conclusions were drawn: The reduction of only 3 columns does not bring a substantial simplification, the PCA complicates the interpretation of the results, and the principal components make interpretation more difficult. Not applying to the PCA was chosen in favor of explainability. A feature permutation method was also used to assess the relevance of variables.

We decided to build five models: DecisionTreeClassifier, RandomForestClassifier, GradientBoostingClassifier, GaussianNB and XGBClassifier. These choices were motivated by the fact that these models assume no ordering between non-ordinal variables, which is particularly suited to our situation. A LogisticRegression model was also built. Although this model may introduce a fictitious order (and therefore bias), it is interesting to compare its performance. Moreover, it's a model that can be easily interpreted and that can perform well.

3. Results

Overall, all models show high accuracy due to the predominance of the majority class (class 0), but struggle to correctly identify the minority class (class 1). The low recall and F1 Score values for the positive class underline the challenge posed by the imbalance of classes in the dataset. A use of the SMOTE (Synthetic Minority Over-sampling Technique) to balance our training set, to manage the significant class imbalance to oversample the minority class, resulting in resampled data sets. Then we chose to proceed with hyperparameter tuning for the RandomForestClassifier, GradientBoostingClassifier and XGBClassifier models.

To this end, we first applied RandomizedSearchCV for faster exploration of the hyperparameter space, followed by GridSearchCV for more precise optimization. This strategy enables us to efficiently identify the best configurations to improve the performance of our models. The shap library has also been used to understand and explain how the models work, thereby increasing their explainability.

4. Discussion

The integration of predictive models in healthcare, particularly for the prediction of readmissions, can transform the management of patient care. The ability to anticipate readmissions within 30 days of discharge enables hospitals to classify patients according to their risk (high, medium, low) and implement targeted interventions, such as frequent follow-ups and personalized education. This improves follow-up and reduces readmission rates. Economically, this approach optimizes resource management by proactively identifying high-risk patients, enabling early intervention programs and reducing the costs associated with prolonged care. Ensuring adequate follow-up also improves patient satisfaction. In conclusion, the use of predictive models not only strengthens clinical decision-making, but also fosters innovation in medical organizations. This leads to more efficient resource allocation and better care coordination, which is essential in a sector where improving care and reducing costs are crucial.