Diabetes 130-US hospitals for years 1999-2008

Group 4

Goal

To use logistic regression to predict the risk of readmission for diabetic patients in US hospitals based on various patient and hospital characteristics.

Dataset

UCI Machine Learning Repository: Diabetes 130-US hospitals for years 1999-2008 Data Set

Dataset Description

The diabetes dataset from the UCI Machine Learning Repository contains data on over 100,000 hospital admissions of patients with diabetes to 130 hospitals in the United States from 1999 to 2008. The goal of this analysis is to build a logistic regression model to predict whether a patient will be readmitted within 30 days of their initial hospitalization based on their demographic information, medical history, and medications. There are 37 categorical and 13 numerical variables.

- Age: age of patient in years
- Gender: gender of patient (male or female)
- Admission_type_id: identifier corresponding to the type of admission (emergency, urgent, elective, etc.)
- Time_in_hospital: number of days between admission and discharge
- Medical_specialty: specialty of the admitting physician
- Num_lab_procedures: number of lab procedures performed during the stay
- Num_procedures: number of non-lab procedures performed during the stay
- Num_medications: number of distinct medications administered during the stay
- Change: whether there was a change in medication during the stay (yes or no)

- DiabetesMed: whether any diabetes medication was prescribed (yes or no)
- Readmitted: whether the patient was readmitted within 30 days (yes, no, or >30)
- The dependent variable for the logistic regression model is the binary variable "Readmitted", which can be coded as 1 for patients who were readmitted within 30 days and 0 for those who were not.

Plan

Data Preprocessing:

1. Remove instances with missing values or incomplete data
2. Encode categorical variables using one-hot encoding or label encoding
3. Standardize numerical variables using Z-score normalization

Feature Selection:

1. Use domain knowledge and exploratory data analysis to identify relevant features
2. Apply statistical techniques like correlation analysis, chi-square test, or feature importance score to select important features.

Model Training:

1. Split the dataset into training and validation sets
2. Fit a logistic regression model to the training data
3. Evaluate the model's performance on the validation set using metrics such as accuracy, precision, recall, and F1 score.
4. Tune the model hyperparameters using techniques to improve the model's performance.

Model Interpretation:

1. Interpret the model coefficients to understand the impact of each feature on the target variable.
2. Analyze the model's predictions using tools like confusion matrix, ROC curve, and precision-recall curve to gain insights into the model's behavior.
3. Use the model to generate predictions on new data and evaluate its performance on an independent test set.