Diabetic Patients Data Of 130-US hospitals for years 1999-2008

Introduction & Aim of the Project

The main intention to choose Healthcare data is to understand the characteristics and importance of each feature in the dataset and how better models can be built using the right set of data. Around 7% of the population worldwide are suffering with Diabetes. It is a chronic disease characterized by elevated levels of blood glucose which increases the risk of stroke by 1.8 times, increases the mortality rate by 1.8 times when compared to undiagnosed diabetic patients. We will be determining the readmission of patients to the hospitals after their medication during the first visit.

Our aim is to understand the data and focus on EDA which is an important step that consumes most of the time and effort in any Data Science Project. Having the right set of data will help to develop robust algorithms. As we learned **'No Free Lunch Theorem'**, which states No single Machine Learning Algorithm is the best solution for all problems, we wish to explore the data and try different types of classification algorithms to find out the best fit algorithm.

Data Sources:

UCI Machine Learning Repository:
This data has been prepared to analyze factors related to readmission as well as other outcomes pertaining to patients with diabetes. This data set was used in a research work https://www.hindawi.com/journals/bmri/2014/781670/

Problem Definition:

The goal is to predict the chances of readmission of patients to the hospital considering certain changes in medication procedures, length of stay at the hospital, HbA1c measurement, age group, and multiple factors. We will classify the patients into multiple classes. Whether they are readmitted within 30 days, readmitted after 30 days or not at all readmitted.
Hence, this project is a Classification project.

Data Description:

The dataset represents 10 years (1999-2008) data of clinical care at 130 US hospitals. It includes 55 features representing patient and hospital outcomes. Information was extracted from the database for encounters that satisfied the following criteria.

(1) It is an inpatient encounter (a hospital admission).
(2) It is a diabetic encounter, that is, one during which any kind of diabetes was entered into the system as a diagnosis.
(3) The length of stay was at least 1 day and at most 14 days.
(4) Laboratory tests were performed during the encounter.
(5) Medications were administered during the encounter.
The data contains such attributes as patient number, race, gender, age, admission type, time in hospital, medical specialty of admitting physician, number of lab tests performed, HbA1c test

result, diagnosis, number of medications, diabetic medications, number of outpatient, inpatient, and emergency visits in the year before the hospitalization, etc.

Understanding of Data:

As the weight column contains 97% missing values, it is not considered in the further analysis. Payer code contains 40% missing values and doesn't impact the outcome of the project so it is not considered for further analysis. Also, the medical specialty column has 47% missing values, as it is important for the analysis. There are multiple types of admission types, admission sources and discharge types which are unique. Patient's data consists of the age group, gender, race, length of their stay, number of lab procedures, number of medications, number of diagnoses, HbA1C result, label data related to multiple drugs given to patients, change of medication.

Exploratory Data Analysis & Model Development

Our idea is to experiment with different techniques in each technique of EDA starting from Data Visualization, Feature Selection, Feature Engineering, Data Cleaning, Null Value Handling, Model Development, Model Evaluation, tuning the model to finalize the techniques that are best fit to this nature and characteristics of data. We focus on understanding the nature of data and try to use techniques that do not alter the characteristics of data(Example: We Prefer to use those imputation techniques for features that result in less variance difference from the previous values)

1. Dimensionality Reduction techniques will be implemented to find out the best influencing features.
2. Plot multiple histograms for skewness, boxplots for outliers, scatter plots for relationship type, correlation heat map
3. Null Values will be handled using different techniques like : Mean, Median Imputation, Other Imputation methods, Prediction of missing values
4. Outliers will be handled using techniques that are suitable to the dataset.
5. No duplicates as each record portrays a unique patient history
6. For the medical specialty column, we will populate the missing values with any keyword and consider for further analysis.
7. Training data, testing data percentages to be decided
8. Since it is a classification problem, we wish to code Logistic Regression initially and later use advanced classification algorithms like Decision Trees, Random Forest, SVM etc.

We believe deploying ML models is the final and most underrated step in any Data Science Project. We would also like to deploy this model as the final step in the project.