**Diabetes Readmission Prediction**

Balaji Venktesh (11910041) | ISB CBA Practicum 2 | 22-Sep-2018

# Table of contents:

|  |  |  |
| --- | --- | --- |
| **Topic #** | **Topic** | **Page #** |
| 1 | Executive summary | 3 |
| 2 | Introduction | 3 |
| 3 | Motivation for the study | 3 |
| 4 | Questions & Hypothesis | 4 |
| 5 | Data collection | 4 |
| 6 | Tools used | 4 |
| 7 | Data cleaning and preparation | 4 |
| 8 | Analysis | 6 |
| 9 | Conclusion and business recommendations | 8 |
| 10 | Appendix | 8 |
| 11 | Reference | 8 |

# Executive Summary:

# Diabetes is a medical condition that is caused due to insufficient production. More than 100 million U.S. adults are now living with diabetes or prediabetes as per report from CDC. Hospital readmission for diabetic patients is a major concern in the United States. Over $250 million dollars was spent on treatment of readmitted diabetic inpatients in 2011 alone. Diabetes is chronic and does not have any specific cure. Hence Hospital readmission is considered an effective measurement of care provided within healthcare. Being able to risk identify patients facing a high likelihood of unplanned hospital readmission in the next 30-days could allow for further investigation and possibly prevent the readmission.

# Introduction:

# A survey conducted by the Agency for Healthcare Research and Quality (AHRQ) found that in the year 2011 more than 3.3 million patients were readmitted in the United States within 30-days of being discharged. Over $250 million was spent on treatment of readmitted diabetic patients in 2011 (Hines et al., 2014). Current practice to identify at-risk diabetic patients are subjective: a clinician will assess the patient and decide what the appropriate care plan is for that individual. Research has shown that these subjective methods for determining readmission are slightly better than random guessing. However, there are tools to objectively score readmission risk, such as LACE. These objective tools are seen to be useful because end-users can make these calculations manually and offer improved accuracy over subjective techniques.

# Machine learning models can be used to create objective models which then can be used to measure risk (Mingle, 2015). These models are more complex, but may be able to create more accurate risk predictions that should lead to improved diabetic patient outcomes.

# Motivation for this study:

# Hospital readmissions of diabetic patients are expensive as hospitals face penalties if their readmission rate is higher than expected and reflects the inadequacies in health care system. Most hospitals can agree that their main goals are centered on improving outcomes, creating more satisfied patients, and better value. For these reasons, it is important for the hospitals to improve focus on reducing readmission rates. Identify the key factors that influence readmission for diabetes and to predict the probability of patient readmission. Hence a Machine Learning Model to understand the diabetics patients and their treatments and to predict potential patients under Age 30 with higher chances. This machine learning framework solves a general problem for diabetic patients who discharge from the hospital and as a single comprehensive solution can be easily implemented. Also, to identify potential modifiable risk factors leading to readmission rates.

# Questions & Hypothesis:

The questions that I would like to test using the data collected are:

1. Does this patient will be readmitted for Diabetics treatment again?
2. What are all the potential modifiable risk factors leading to readmission rates?
3. How accurate our model for predicting readmission?

# Data collection:

Dataset is taken from publicly available UCI Repository.

1. https://archive.ics.uci.edu/ml/datasets/Diabetes+130-US+hospitals+for+years+1999-2008#

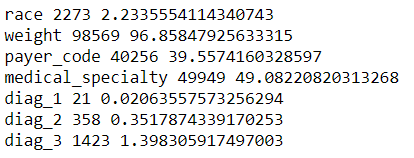
# The dataset represents 10 years (1999-2008) of clinical care at 130 US hospitals and integrated delivery networks. It includes over 50 features representing patient and hospital outcomes. There are around 1,00,000 records for you to analyze.

# Tools used:

1. Python and its libraries
   1. NumPy, Pandas for data processing
   2. Matplotlib in python for visualization
   3. SciKit learn for ML algorithms

# Data cleaning and preparation:

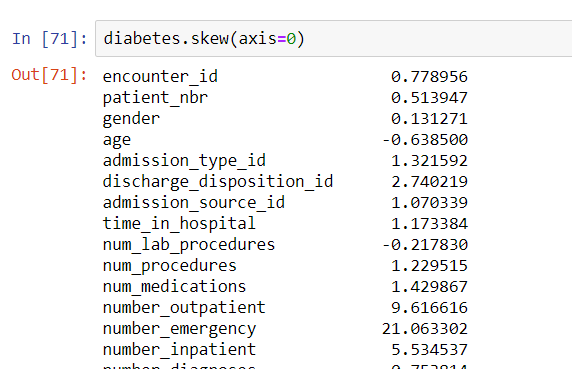
1. From the given data, found we have totally 101766 records with 50 features.
2. Below highlighted 7 features had wrong value “**?**” in the dataset



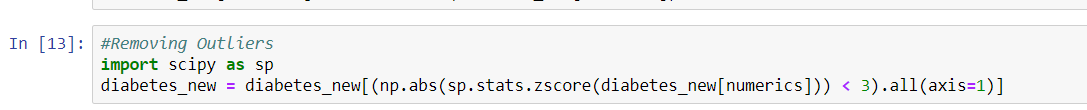
1. Also, Column “**Gender**” has 3 records with value “Unknown”



1. Columns which has more than 40% of data with improper data was dropped from the analysis and it is not suitable for imputation from the remaining data. For rest of the data, we dropped only those records which had wrong data.
2. New Feature “Service Utilization” has been created to identify services availed by patient relations
3. Based on ICD Values, Categorizing diagnoses level into smaller categories like **Circulatory, Respiratory, Digestive, Diabetes, Injury, Musculoskeletal, Genitourinary, Neoplasms, and Others**
4. Conversion of Categorical Variables into equivalent Dummy Variables
5. Removal of Duplicates records and duplicate features. Ex: **Examide** & **Citoglipton** had same value of each rows.
6. Check for skewness and normalization of the data.



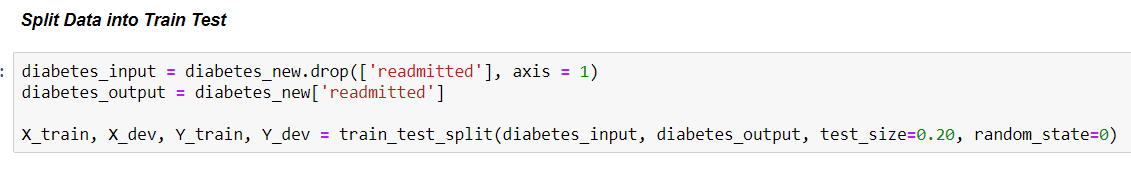
1. Removing Outliers in the data



# Analysis:

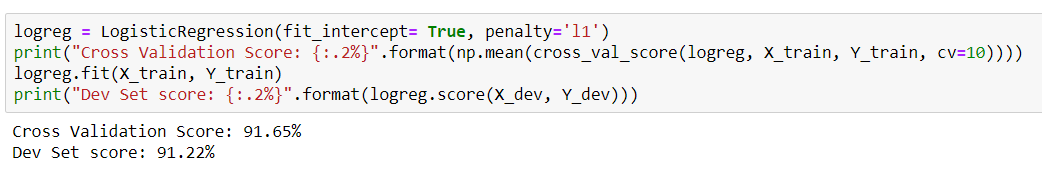
Model Building

**Splitting Data into Train & Test first**

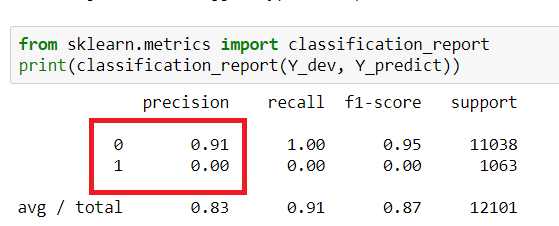


1) **Logistic Regression**

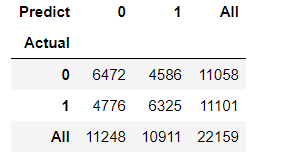
Using K-Fold Cross Validation with N=10, we achieved nearly 91% of accuracy



**But when check CONFUSION MATRIX, model performed poorly in identifying those diabetics readmitted patients due to unbalanced data**

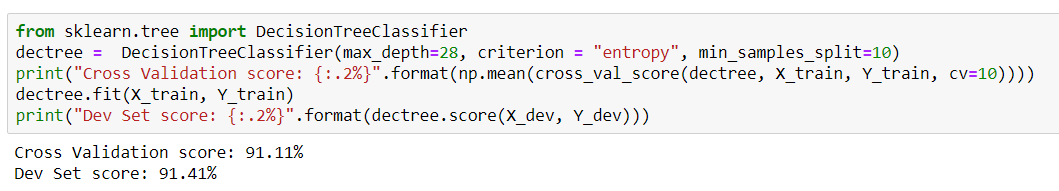


After applying SMOTE for Data Balancing, we saw **57% accuracy**

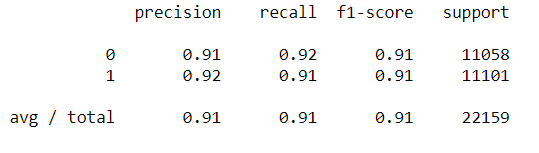


2) Decision Tree Classifier

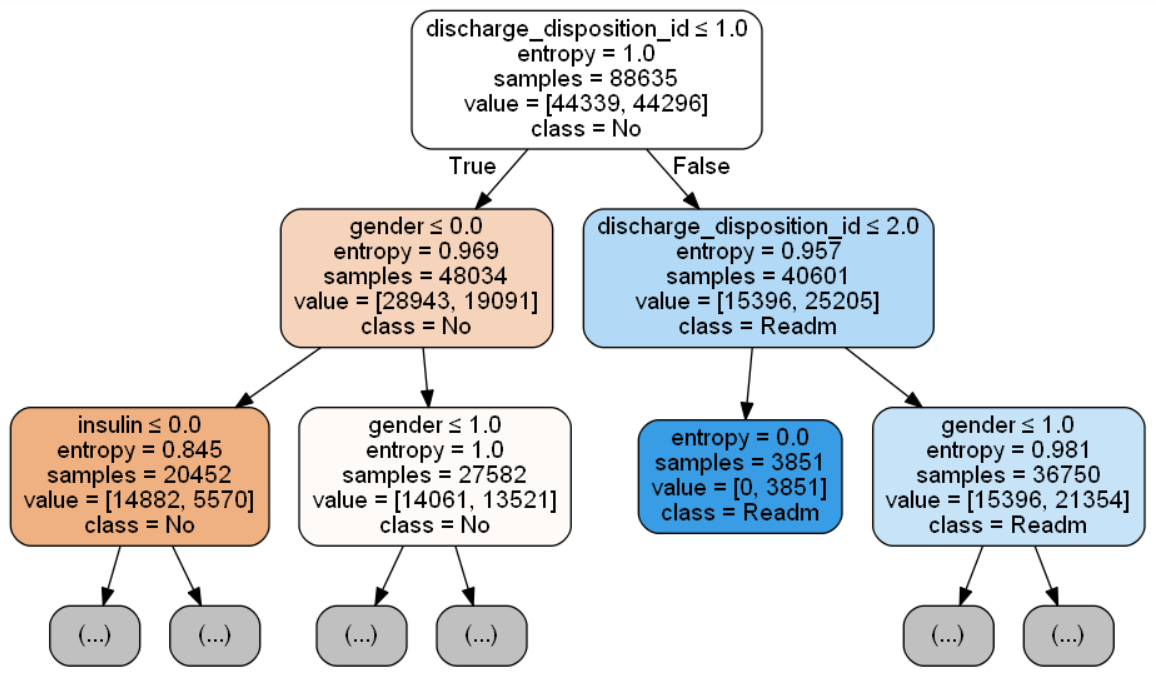
Using K-Fold Cross Validation with N=10, we achieved nearly 91% of accuracy



Model did good performance while predicting Re-Admission



Decision Tree Graph:

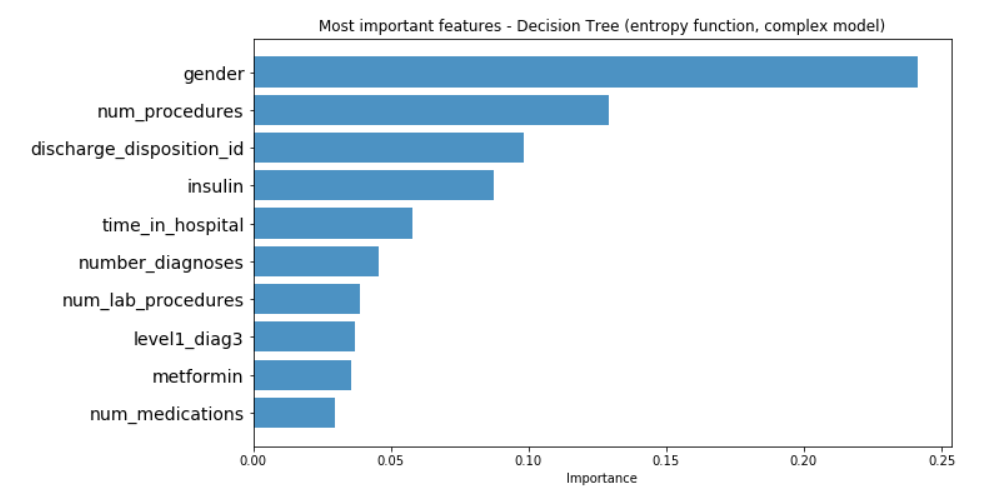


# Conclusion and Business Recommendations:

Based on Decision Tree Fitted Model, below are list of high important features affecting Re-Admission of Diabetics Patient.

**1) Number of Procedures a Patient undergoes while Treatment along with Gender import highly for Re-Admission.**

**2) Similarly, Time in Hospital also play major role.**



# References and Credits:

1. **“Impact of HbA1c Measurement on Hospital Readmission Rates: Analysis of 70,000 Clinical Database Patient Records,” BioMed Research International, vol. 2014, Article ID 781670, 11 pages, 2014. https://doi.org/10.1155/2014/781670.**