

**Capstone Project – Interim Report**

|  |  |
| --- | --- |
| **Batch details** | **PGP-DSE PUNE JUL’20** |
| **Team members** | **Praful Bhoyar**  **Soyeb Kapasi**  **Varun Kukday**  **Vinay Deokar**  **Yagjna Kurra** |
| **Domain of Project** | **Healthcare** |
| **Proposed project title** | **Diabetes Patient Readmission Prediction**  **Analysis of 100,000 Clinical Database Patient Records** |
| **Group Number** | **Group 3** |
| **Mentor Name** | **Srikar Muppidi** |

**Business Understanding:**

It is increasingly recognized that the management of Hyperglycemia (High Blood Sugar) in the hospitalized patient has a significant bearing on the outcome, in terms of both morbidity and mortality. This recognition has led to the development of formalized protocols in the intensive care unit (ICU) setting with rigorous glucose targets in many institutions.

In particular, we examined the use of HbA1c as a marker of attention to diabetes care in a large number of individuals identified as having a diagnosis of diabetes mellitus. We hypothesize that measurement of HbA1c is associated with a reduction in readmission rates in individuals admitted to the hospital.

**Business Objective:**

A hospital readmission is when a patient who is discharged from the hospital, gets re-admitted again within a certain period of time. Hospital readmission rates for certain conditions are now considered an indicator of hospital quality, and also affect the cost of care adversely.

Determining which factors are the strongest predictors of hospital readmission in diabetic patients & the efficiency and accuracy of the model in predicting hospital readmission with limited features.

**Data Description:**

**Encounter ID** - Unique identifier of an encounter

**Patient number** - Unique identifier of a patient

**Race** – Race of patient (Caucasian, Asian, African American, Hispanic, and other)

**Gender** - Male, Female, and unknown/invalid

**Age** - Grouped in 10-year intervals: 0-10, 10-20, etc

**Weight** - Weight in pounds

**Admission type** - Integer identifier corresponding to 9 distinct values, for example, emergency, urgent, elective, new-born and not available

**Discharge disposition** - Integer identifier corresponding to 29 distinct values, for example, discharged to home, expired, and not available

**Admission source** - Integer identifier corresponding to 21 distinct values, for example, physician referral, emergency room, and transfer from a hospital

**Time in hospital** - Integer number of days between admission and discharge

**Payer code** - Integer identifier corresponding to 23 distinct values, for example, Blue Cross\Blue Shield, Medicare, and self-pay

**Medical specialty** - Integer identifier of a specialty of the admitting physician, corresponding to 84 distinct values, for example, cardiology, internal medicine, family\general practice, and surgeon

**Number of lab procedures** - Number of lab tests performed during the encounter

**Number of procedures** - Number of procedures (other than lab tests) performed during the encounter

**Number of medications** - Number of distinct generic names administered during the encounter

**Number of outpatient visits** - Number of outpatient visits of the patient in the year preceding the encounter

**Number of emergency visits** - Number of emergency visits of the patient in the year preceding the encounter

**Number of inpatient visits** - Number of inpatient visits of the patient in the year preceding the encounter

**Diagnosis 1** - The primary diagnosis (coded as first three digits of ICD 9); 848 distinct values

**Diagnosis 2** - Secondary diagnosis (coded as first three digits of ICD 9); 923 distinct values

**Diagnosis 3** - Additional secondary diagnosis (coded as first three digits of ICD 9); 954 distinct values

**Number of diagnoses** - Number of diagnoses entered to the system

**Glucose serum test result** - Indicates the range of the result or if the test was not taken.

**A1c test result** - Indicates the range of the result or if the test was not taken.

**Metformin -** sold under the brand name Glucophage among others, is the first-line medication for the treatment of type 2 diabetes, particularly in people who are overweight.

**Repaglinide -** isused alone or with other medications to control high blood sugar along with a proper diet and exercise program. It is used in people with type 2 diabetes.

**Nateglinide -** is a drug for the treatment of type 2 diabetes

**Chlorpropamide -** isan oral antihyperglycemic agent used for the treatment of non-insulin-dependent diabetes mellitus (NIDDM).

**Glimepiride -** isan oral diabetes medicine that is used together with diet and exercise to improve blood sugar control in adults with type 2 diabetes mellitus.

**Acetohexamide -** is a first-generation sulfonylurea medication used to treat diabetes mellitus type 2, particularly in people whose diabetes cannot be controlled by diet alone.

**Glipizide** - sold under the brand name Glucotrol among others, is an anti-diabetic medication of the sulfonylurea class used to treat type 2 diabetes

**Glyburide -** is a diabetes medicine used to help control blood sugar levels and treat type 2 diabetes.

**Tolbutamide -** is a first-generation potassium channel blocker, sulfonylurea oral hypoglycemic medication. This drug may be used in the management of type 2 diabetes if diet alone is not effective.

**Pioglitazone -** is a diabetes drug (thiazolidinedione-type, also called "glitazones") used along with a proper diet and exercise program to control high blood sugar in patients with type 2 diabetes.

**Rosiglitazone** - is an insulin sensitizing agent and thiazolidinedione that is indicated for the treatment of type 2 diabetes.

**Acarbose -** is an anti-diabetic drug used to treat diabetes mellitus type 2 and, in some countries, prediabetes.

**Miglitol** - is an oral anti-diabetic drug that acts by inhibiting the ability of the patient to break down complex carbohydrates into glucose.

**Troglitazone -** is an antidiabetic and anti-inflammatory drug, and a member of the drug class of the thiazolidinediones. It was prescribed for people with diabetes mellitus type 2.

**Tolazamide -** is an oral blood glucose lowering drug used for people with Type 2 diabetes. It is part of the sulfonylurea family.

**Citoglipton (Sitagliptin)** - Sitagliptin is a diabetes drug that works by increasing levels of natural substances called incretins. Incretins help to control blood sugar by increasing insulin release, especially after a meal. They also decrease the amount of sugar your liver makes.

**Glyburide-Metformin** - The combination of glyburide and metformin is used to treat type 2 diabetes (condition in which the body does not use insulin normally and therefore cannot control the amount of sugar in the blood) in people whose diabetes cannot be controlled by diet and exercise alone.

**Glipizide-Metformin** - Glipizide and Metformin combination is used to treat high blood sugar levels that are caused by a type of diabetes mellitus or sugar diabetes called type 2 diabetes

**Glimepiride-Pioglitazone** - Pioglitazone and glimepiride combination is used with proper diet and exercise to treat high blood sugar levels caused by type 2 diabetes. Pioglitazone works by helping your body use insulin better. Glimepiride stimulates the release of insulin from the pancreas which will help your body turn food into energy

**Metformin-Rosiglitazone** - Rosiglitazone and metformin combination is used to treat a type of diabetes mellitus called type 2 diabetes. It is used together with a proper diet and exercise to help control blood sugar levels.

**Metformin-Pioglitazone** - Metformin/pioglitazone is used to improve blood sugar control in adults with type 2 diabetes. It's used along with diet and exercise. Metformin/pioglitazone isn't used to treat type 1 diabetes.

**24 features for medications** - The feature indicates whether the drug was prescribed or there was a change in the dosage. Values: “up” if the dosage was increased during the encounter, “down” if the dosage was decreased, “steady” if the dosage did not change, and “no” if the drug was not prescribed

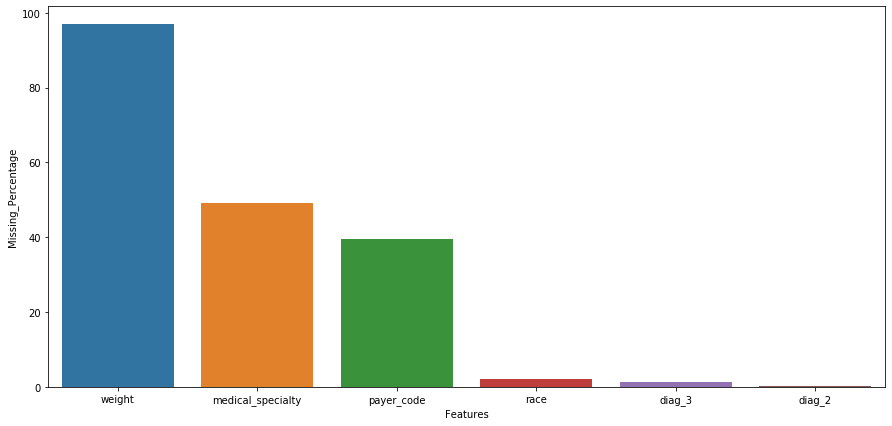
**Change of medications** - Indicates if there was a change in diabetic medications (either dosage or generic name).

**Diabetes medications** - Indicates if there was any diabetic medication prescribed

**Readmitted** - Whether the Patient was Readmitted or not OR whether the Patient was readmitted within 30 days or not.

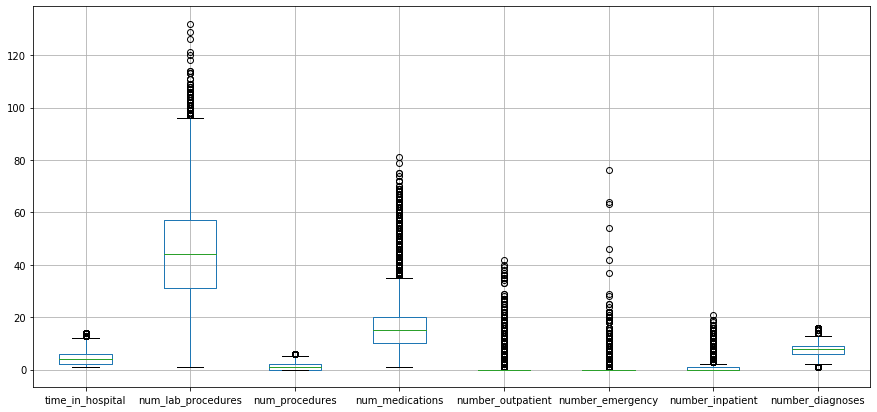
**Data Pre-Processing and Preparation:**

**NULL Handling:**

****

There are no traditional “Null Values” in the data rather there are some values which are missing that have been filled with “?” . We have treated these values in most cases with assigning it a class or dropping the variable altogether in case of weight where 97% values were missing.

**Outlier Handling:**

****

As this is critical patient data, we have chosen not to do outlier treatment so as to preserve the integrity of the dataset which is linked to real world outcomes.

**Exploratory Data Analysis & Business Insights:**

**####################################**

**UNIVARIATE:**

We can see from the data that on average, a patient spends about 4.3 days in the hospital, and around 76% of the patients we have in the data are of the Caucasian race i.e White, with African Americans patients in the data being around 19%. As for the Gender of the patients we have around 53% Female patients and 47% Male Patients.

The majority of the patients in the data are between the ages of 50 and 90. We have around 97% Missing values in the weight column, deeming it unfit for consideration. We have 71518 unique patients with some of them having multiple visits. As for the target variable “Readmitted”, we have 54864 patients who were not readmitted while 35545 patients having to be readmitted after 30 days and 11357 patients having to be readmitted before 30 days from date of discharge.

We can see from the analysis of the Drugs in the data, the majority of the patients are being prescribed one or more of the drugs with a steady dosage.

**BIVARIATE:**

We can see from the bi-variate analysis done with the target variable, we can see that the majority of the variables follow a similar trend to the distribution of the target variable, with majority coming under the ‘NO’ classification in terms of having to be readmitted.

**#####################################**

**Feature Engineering:**

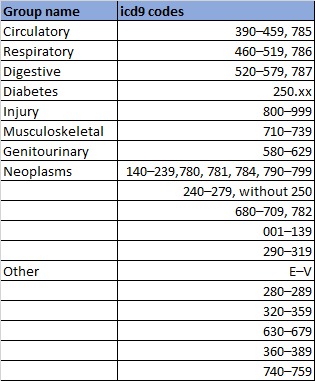
New Features Created from existing Features for making Interpretations:

* Diagnosis Kind
* Diagnosis Complications

These Features are created based on the ICD-9 Codes which are shared under the diagnoses noted down. These ICD-9 Codes are the method of capturing the diagnoses in US Hospitals.

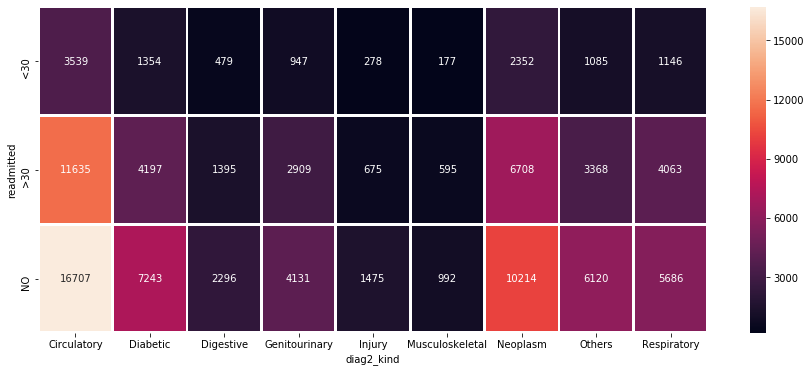
**Diagnosis Kind:**

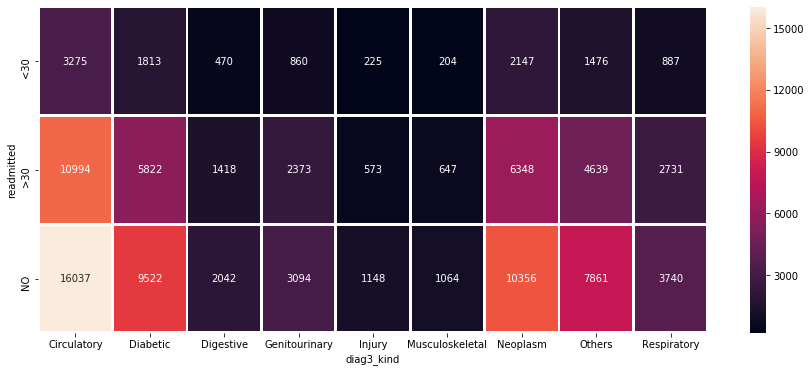
The code-mapping for Diagnosis Kind is described below:



We could see here how the Diagnosis Kind was affecting my Target of Readmission:





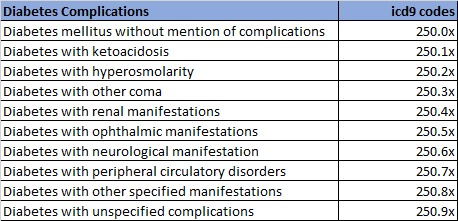


Based on the above graphs we could see that the people getting diagnosed are getting detected by the 3rd diagnosis, as the number of diabetic kind diagnoses are increasing from 1st to 2nd to 3rd Diagnostic Test.

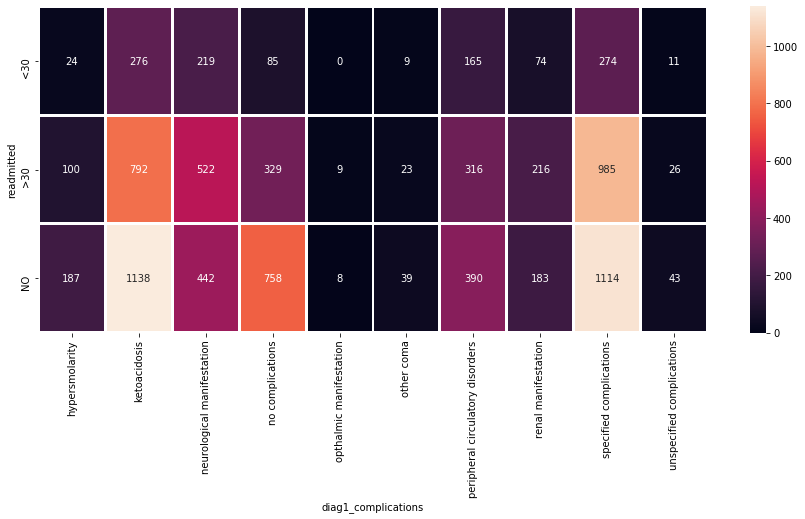
This might be a reason for the increase in readmission percentage of the patients!

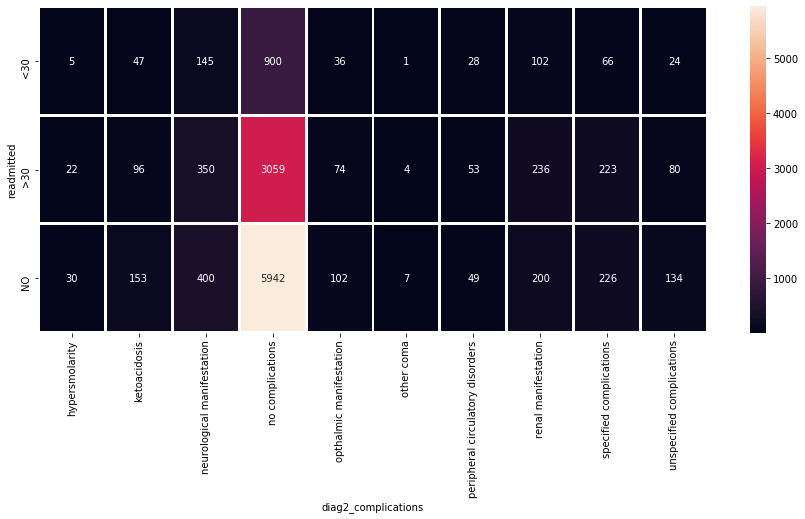
**Diabetes Complications:**

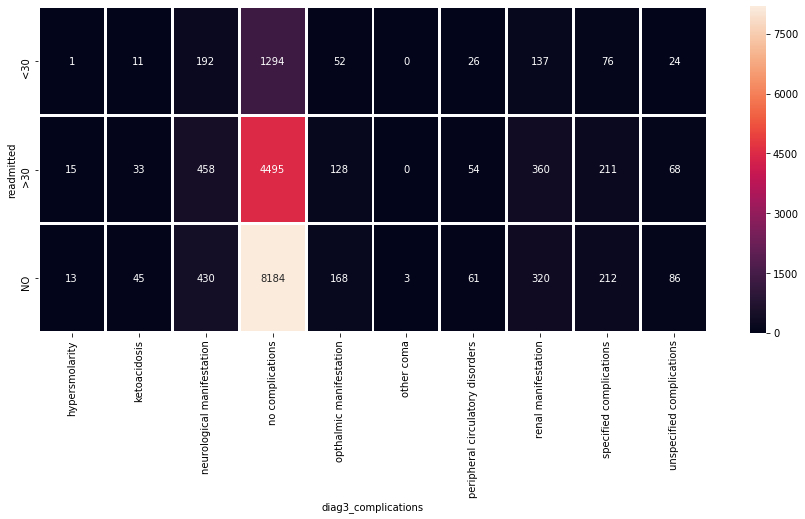
The Codes mapped for Diagnosis Complications are only considered here towards Diabetes Patients and None was assigned for others. The Mapping is as given below:



We could see here how the Diagnosis Kind was affecting my Target of Readmission:







We could see that major diagnoses of Diabetic people are related to Ketoacidosis and Neurological manifestations.

**Ketoacidosis** is a serious complication of diabetes that occurs when your body produces high levels of blood acids called ketones. The condition develops when your body can't produce enough insulin. Insulin plays a major role for Glucose(Sugar) to enter the cells which inturn provides Energy!

**Neurologic** disorders are a common and often disabling aspect of diabetes mellitus. Pain and sensory disturbances, weakness and paralysis and symptoms of autonomic dysfunction may be experienced by the diabetic patient.

These might be the Major Focus Areas for the Hospital for preemptive care Measures to be  
taken which would in turn reduce the Patient Readmission.

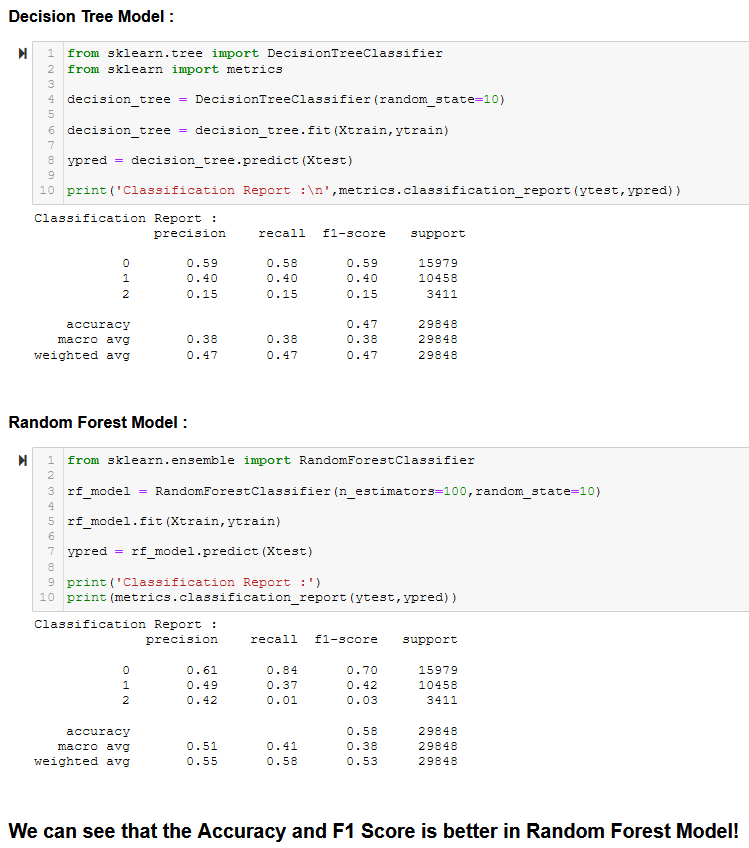
**Patient Visits** : A new feature of Patient Visits was manufactured from the patient number. As this data is around 9 years from 1999-2008, the first occurrence of individual patients is given as count 1 and then the count was given in an incremental fashion for each patient.



Majority of patients have visited less than 6 times, although there are people who have visited multiple times!

Based on this new information gained from the **codes** and **patient\_nbr**, we made new features of Kind and Complications for each diagnosis 1,2 and 3 and patient\_visits.   
Post this, we chose to drop the icd codes and patient\_nbr which were not giving any other information.

**Basic Model:**

****