**CAPSTONE PROJECT REPORT**

***Detection of Type II Diabetes Mellitus and Impact of HbA1c Measurement on Hospital Readmission Rates***

*Submitted towards partial fulfillment of the criteria for award of* ***PGP-BABI*** *by Great Lakes*

***Submitted By:***

*Girish Velivala - BAHMAR17018*

*Aravind Rao Bhaskar - BAHMAR17003*

*Ashwanth Prathapani - BAHMAR17007*

*Sumit Chander - BAHMAR17067*

***Course and Batch: PGP-BABI MARCH 2017-18***

**Mentor: Mr. Y. L. Prasad**



B-Block, 1st Floor, Kavuri Incor9,

Kavuri Hills Road, Madhapur, Telangana

Tel: +91-040-485 2700

[www.greatlakes.edu.in](file:///C:\GirFtp\zcap\www.greatlakes.edu.in)

[www.greatlearning.in](file:///C:\GirFtp\zcap\www.greatlearning.in)

**Acknowledgements**

We wish to place on record our deep appreciation for the guidance and help provided to us by our ***Mentor Y. L. Prasad*** for completing this project on time.

Mr. Prasad reviewed the design and model proposed by us, the R scripts written and the analytics generated, providing us valuable feedback at every stage to embellish the process and the outputs.

We would also like to place on record our appreciation and special thanks to ***Dr. Srabashi Basu*** for giving us valuable feedback and being a source of inspiration in helping us to work on this project.

We certify that the work done by us for conceptualizing and completing this project is original and authentic.

Date : January 28, 2018 Girish Velivala

Place : Hyderabad Aravind Rao Bhaskar

Ashwanth Prathapani

Sumit Chander

**Certificate of Completion**

I hereby certify that the project titled “**Detection of Type II Diabetes Mellitus and Impact of HbA1c Measurement on Hospital Readmission Rates**” was undertaken and completed under my supervision by Girish Velivala, Aravind Rao Bhaskar, Ashwanth Prathapani and Sumit Chander, students of the Postgraduate Program in Business Analytics and Business Intelligence (PGP-BABI MARCH 2017-18).

Date : January 28, 2018 (Y. L. Prasad)

Place : Hyderabad Mentor & Faculty Guide

Table of Contents

[2 Abstract 6](#_Toc508958677)

[3 Literature Overview 7](#_Toc508958678)

[4 Executive Summary 8](#_Toc508958679)

[4.1 Introduction 8](#_Toc508958680)

[4.2 Need of Study 8](#_Toc508958681)

[4.3 Objectives 8](#_Toc508958682)

[4.4 Data Sources 8](#_Toc508958683)

[4.5 Statistical Tools and Techniques 10](#_Toc508958684)

[5 Data Management 11](#_Toc508958685)

[5.1 Check for Missing Values 12](#_Toc508958686)

[5.2 Feature Selection 13](#_Toc508958687)

[5.3 Missing value treatment/coding 14](#_Toc508958688)

[5.4 Removal of unwanted rows 14](#_Toc508958689)

[5.5 No. of unique patients 14](#_Toc508958690)

[5.6 Factor levels handling 14](#_Toc508958691)

[5.7 Feature aggregation 16](#_Toc508958692)

[5.8 Final dataset 16](#_Toc508958693)

[6 Exploratory Data Analysis 17](#_Toc508958694)

[6.1 Gender vs Readmission 17](#_Toc508958695)

[6.2 Race vs Readmission 18](#_Toc508958696)

[6.3 Age vs Readmission 19](#_Toc508958697)

[6.4 Medical Specialty vs Readmission 20](#_Toc508958698)

[6.5 Time in Hospital vs Readmission 21](#_Toc508958699)

[7 Chi-Square Test 21](#_Toc508958700)

[7.1 Chi-Square Summary 21](#_Toc508958701)

[8 Information Gain 22](#_Toc508958702)

[8.1 Summary Plot 22](#_Toc508958703)

[9 Training and Test data 23](#_Toc508958704)

[10 Model Development 23](#_Toc508958705)

[10.1 Logistic regression model summary 23](#_Toc508958706)

[10.2 Model Odds Ratios 25](#_Toc508958707)

[10.3 Predicted probabilities for readmitted 25](#_Toc508958708)

[11 Model Validation 26](#_Toc508958709)

[11.1 AUC Curve 26](#_Toc508958710)

[11.2 Accuracy 27](#_Toc508958711)

[11.3 HbA1c vs Readmission 27](#_Toc508958712)

[12 New Model Development on segregated data 28](#_Toc508958713)

[12.1 Removal of encounters where HbA1c test was not performed 28](#_Toc508958714)

[12.2 Training and test split 28](#_Toc508958715)

[12.3 New Logistic regression model summary 29](#_Toc508958716)

[12.4 New Model Odds Ratios 31](#_Toc508958717)

[12.5 New Model Predicted probabilities for readmitted 31](#_Toc508958718)

[13 New Model Validation 32](#_Toc508958719)

[13.1 AUC Curve 32](#_Toc508958720)

[13.2 Accuracy 33](#_Toc508958721)

[14 Recommendations and Conclusions 33](#_Toc508958722)

[14.1 Best Model Accuracy 33](#_Toc508958723)

[14.2 Best Model Recommendation 34](#_Toc508958724)

[14.3 Conclusion 37](#_Toc508958725)

[15 Bibliography 37](#_Toc508958726)

[16 Annexure 38](#_Toc508958727)

[16.1 Contacts 38](#_Toc508958728)

[16.2 LDA 38](#_Toc508958729)

[16.2.1 Summary 38](#_Toc508958730)

[16.2.2 Accuracy 39](#_Toc508958731)

[16.3 MDA 39](#_Toc508958732)

[16.3.1 Summary 39](#_Toc508958733)

[16.3.2 Accuracy 40](#_Toc508958734)

[16.4 Neural Net 40](#_Toc508958735)

[16.4.1 Summary 40](#_Toc508958736)

[16.4.2 Accuracy 41](#_Toc508958737)

[16.5 SVM 41](#_Toc508958738)

[16.5.1 Summary 41](#_Toc508958739)

[16.5.2 Accuracy 42](#_Toc508958740)

[16.6 KNN 42](#_Toc508958741)

[16.6.1 Summary 42](#_Toc508958742)

[16.6.2 Accuracy 43](#_Toc508958743)

[16.7 Naïve Bayes 43](#_Toc508958744)

[16.7.1 Summary 43](#_Toc508958745)

[16.7.2 Accuracy 44](#_Toc508958746)

[16.8 CART 44](#_Toc508958747)

[16.8.1 Summary 44](#_Toc508958748)

[16.8.2 Accuracy 45](#_Toc508958749)

[16.9 C4.5 45](#_Toc508958750)

[16.9.1 Summary 45](#_Toc508958751)

[16.9.2 Accuracy 46](#_Toc508958752)

[16.10 PART 46](#_Toc508958753)

[16.10.1 Summary 46](#_Toc508958754)

[16.10.2 Accuracy 47](#_Toc508958755)

[16.11 Bagging CART 47](#_Toc508958756)

[16.11.1 Summary 47](#_Toc508958757)

[16.11.2 Accuracy 48](#_Toc508958758)

[16.12 Random Forest 48](#_Toc508958759)

[16.12.1 Summary 48](#_Toc508958760)

[16.12.2 Accuracy 49](#_Toc508958761)

[16.13 C5.0 49](#_Toc508958762)

[16.13.1 Summary 49](#_Toc508958763)

[16.13.2 Accuracy 50](#_Toc508958764)

# Abstract

Management of hyperglycemia in hospitalized patients has a significant bearing on outcome, in terms of both morbidity and mortality. However, there are few assessments of diabetes care during hospitalization which could serve as a baseline for change. This analysis of a large clinical database was undertaken to provide such an assessment and to find future directions which might lead to improvements in patient safety. Almost 70,000 inpatient diabetes encounters were identified with sufficient detail for analysis. Multivariable logistic regression was used to fit the relationship between the measurement of HbA1c and early readmission while controlling for covariates such as demographics, severity and type of the disease, and type of admission. Results show that the measurement of HbA1c was performed very infrequently. The statistical model suggests that the relationship between the probability of readmission depends on the HbA1c measurement and based on the results of the measurement whether any change of medications has been administered or not. We hypothesize that measurement of HbA1c is associated with a reduction in readmission rates in individuals admitted to the hospital. The data suggest further that the greater attention to diabetes reflected in HbA1c determination may improve patient outcomes and lower cost of inpatient care.

**Techniques : Predictive modeling and data mining**

**Tools : R, Tableau**

**Domain : Health Care**

# Literature Overview

Hyperglycemia in hospitalized patients is associated with increased morbidity, mortality, and length of hospital stay. Insulin counteracts the damaging processes caused by hyperglycemia and is therefore a logical choice in treating inpatient hyperglycemia. Hyperglycemia in hospitalized patients is a common, serious, and costly health care problem with profound medical consequences.

Increasing evidence indicates that the development of hyperglycemia during acute medical or surgical illness is not a physiological or benign condition, but rather is a marker of poor clinical outcome and mortality. Evidence from observational studies indicates that development of hyperglycemia in critical illness is associated with an increased risk of complications and mortality, a longer hospital stay, a higher admission rate to the intensive care unit (ICU), and a higher likelihood that transitional or nursing home care after hospital discharge will be required.

The importance of hyperglycemia also applies to adult patients admitted to general surgical and medical wards. In such patients, the presence of hyperglycemia is associated with prolonged hospital stays, infection, and disability after hospital discharge, and death. In a retrospective study of 1,886 patients admitted to a community hospital, mortality in the general floors was significantly higher in patients with newly diagnosed hyperglycemia and those with known diabetes than in those who were normoglycemic (10, 1.7, and 0.8%, respectively; P < 0.01).

In a prospective cohort multicenter study of 2,471 patients, those with admission glucose levels of > 198 mg/dl had a greater risk of mortality and complications than those with lower glucose levels. The risk of in-hospital complications increased 3% for each 18 mg/dl increase in admission glucose. In a retrospective study of 348 patients with chronic obstructive pulmonary disease and respiratory tract infection, the relative risk of death was 2.10 in those with a blood glucose of 126–160 mg/dl and 3.42 for those with a blood glucose > 160 mg/dl compared to patients with a blood glucose < 110 mg/dl.15 The median length of hospital stay was 7 days in those with a blood glucose < 110 mg/dl, 10 days in those with a blood glucose of 126–160 mg/dl, and 12 days in those with a blood glucose > 160 mg/dl. Furthermore, each 18 mg/dl increase in blood glucose was associated with a 15% increase in the risk of an adverse clinical outcome, which was defined as death or length of stay of > 9 days.

Inpatient hyperglycemia is increasingly recognized as a contributor to in-hospital complications and prolonged hospital stays. Protocols to assist in management of hyperglycemia are becoming more widely used and have been shown to improve outcomes for hyperglycemic patients. Clinicians should consider implementing protocols for hyperglycemia management in the inpatient setting.

# Executive Summary

## Introduction

The primary goal that this project is to establish that measurement of HbA1c and based on the test results change in medication is associated with a reduction in readmission rates in individuals admitted to the hospital due to diabetes related complications. This may improve patient outcomes, lower cost of inpatient care and increase in the number of providers who order the HbA1c test.

## Need of Study

Inpatient hyperglycemia is increasingly recognized as a contributor to in-hospital complications and prolonged hospital stays. Protocols to assist in management of hyperglycemia are becoming more widely used and have been shown to improve outcomes for hyperglycemic patients. Clinicians should consider implementing protocols for hyperglycemia management in the inpatient setting.

## Objectives

Applying machine learning and data mining methods in DM research is a key approach to utilizing large volumes of available diabetes-related data for extracting knowledge. The severe social impact of the specific disease renders DM one of the main priorities in medical science research, which inevitably generates huge amounts of data. Undoubtedly, therefore, machine learning and data mining approaches in DM are of great concern when it comes to diagnosis, management and other related clinical administration aspects. Hence, in this project, efforts were made use analytics techniques and design a Data Mining pipeline to derive a set of predictive model/s for DM complications based on electronic health record data.

## Data Sources

This project used the Health Facts database (Cerner Corporation, Kansas City, MO), a national data warehouse that collects comprehensive clinical records across hospitals throughout the United States. The database contains data systematically collected from participating institutions electronic medical records and includes encounter data (emergency, outpatient, and inpatient), provider specialty, demographics (age, sex, and race), diagnoses and in-hospital procedures documented by ICD-9-CM codes, laboratory data, pharmacy data, in-hospital mortality, and hospital characteristics.

The Health Facts data we used was an extract representing 10 years (2007–2016) of clinical care at 130 hospitals and integrated delivery networks throughout the United States.

The database consists of 41 tables in a fact-dimension schema and a total of 117 features. The database includes 74,036,643 unique encounters (visits) that correspond to 17,880,231 unique patients and 2,889,571 providers. Because this data represents integrated delivery network health systems in addition to stand-alone hospitals, the data contains both inpatient and outpatient data, including emergency department, for the same group of patients. However, data from out-of-network providers is not captured.

The dataset was created in two steps. First, encounters of interest were extracted from the database with 50 attributes. Second, preliminary analysis and preprocessing of the data were performed resulting in retaining only these features (attributes) and encounters that could be used in further analysis, that is, contain sufficient information.

Information was extracted from the database for encounters that satisfied the following criteria.

* It is an inpatient encounter (a hospital admission).
* It is a “diabetic” encounter, that is, one during which any kind of diabetes was entered to the system as a diagnosis.
* The length of stay was at least 1 day and at most 14 days.
* Laboratory tests were performed during the encounter.
* Medications were administered during the encounter.

Criteria were applied to remove admissions for procedures and so forth, which were of less than 23 hours of duration and in which changes in diabetes management were less likely to have occurred.

101,766 encounters were identified to fulfill all of the above five inclusion criteria and were used in further analysis. Attribute/feature selection was performed and only attributes that were potentially associated with the diabetic condition or management were retained. From the information available in the database, we extracted 50 features describing the diabetic encounters, including demographics, diagnoses, diabetic medications, number of visits in the year preceding the encounter, and payer information.

## Statistical Tools and Techniques

Statistical analyses will be performed using R statistical software, version 3.3.1 for Windows in conjunction with RStudio, version 1.0.143 for windows; and data mining prediction models will be constructed using R.

Descriptive statistical analyses will be carried out for all variables, using appropriate statistical tests to examine differences between proportions with significance level determined. The primary analyses were stratified by demographic characteristics and lifestyle risk factors.

Logistic regression and decision tree (5.0) models will then be constructed using the training dataset and tested by the testing dataset. The original dataset will be randomly divided into two parts, with the training dataset containing about 80% training of the encounters, and the testing dataset containing 20% of the encounters by the partition methods of R statistical software.

A decision tree is a form for expressing mappings, and it consists of tests or attribute nodes linked to two or more subtrees and leaves or decision nodes labeled with a class that reflects the decision. Popular decision-tree algorithms include Quinlan’s ID3, C4.5, C5.0, and CART.

Graphics will be used to help in the interpretation of interaction terms in the final models.

We will use confusion matrix to appraise the performance of the models for readmission and three evaluated indices for accuracy, sensitivity and specificity. The classification accuracy measures the proportion of cases correctly classified.

Sensitivity measures the fraction of positive cases that are classified as positive.

Specificity measures the fraction of negative cases that are classified as negative.

* **Accuracy = (TP + TN)/(TP + FP + TN + FN)**
* **Sensitivity = TP/(TP + FN)**
* **Specificity = TN/(FP + TN)**

where TP, TN, FP and FN denote true positives, true negatives, false positives and false negatives, respectively. The model with highest the sensitivity, specificity, and accuracy is the best predictive model.

# Data Management

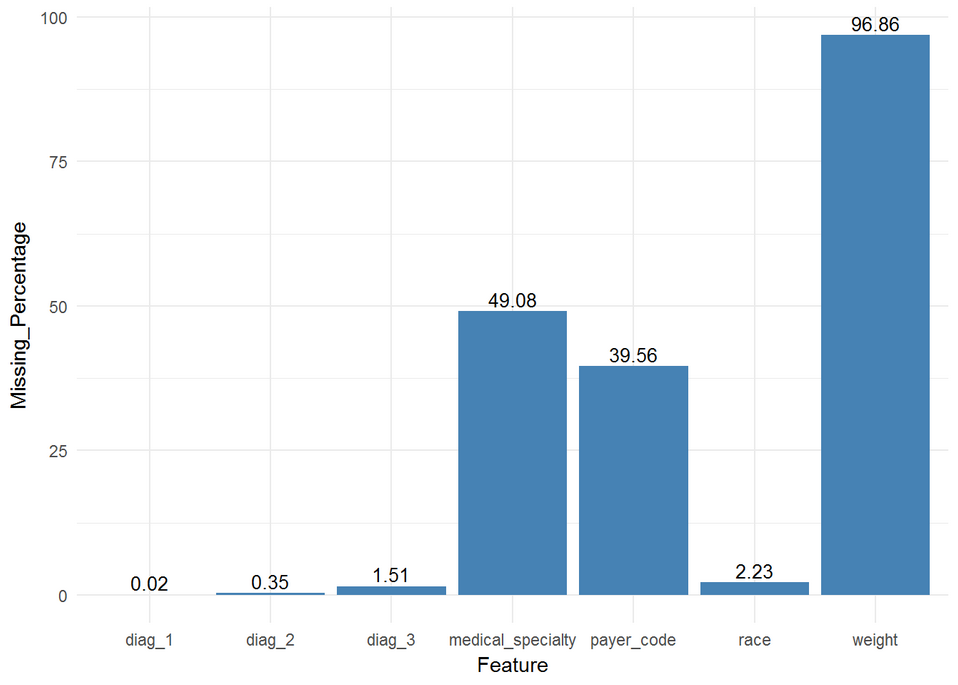
We have 50 features in the final dataset and 101766 patient encounters. Following are the features and their data type, description and values:

|  |  |  |
| --- | --- | --- |
| **Feature name** | **Type** | **Description and values** |
| Encounter ID | Numeric | Unique identifier of an encounter |
| Patient number | Numeric | Unique identifier of a patient |
| Race | Nominal | Values: Caucasian, Asian, African American, Hispanic, and other |
| Gender | Nominal | Values: male, female, and unknown/invalid |
| Age | Nominal | Grouped in 10-year intervals: 0, 10), 10, 20), …, 90, 100) |
| Weight | Numeric | Weight in pounds. |
| Admission type | Nominal | Integer identifier corresponding to 9 distinct values, for example, emergency, urgent, elective, newborn, and not available |
| Discharge disposition | Nominal | Integer identifier corresponding to 29 distinct values, for example, discharged to home, expired, and not available |
| Admission source | Nominal | Integer identifier corresponding to 21 distinct values, for example, physician referral, emergency room, and transfer from a hospital |
| Time in hospital | Numeric | Integer number of days between admission and discharge |
| Payer code | Nominal | Integer identifier corresponding to 23 distinct values, for example, Blue Cross/Blue Shield, Medicare, and self-pay |
| Medical specialty | Nominal | 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 | Numeric | Number of lab tests performed during the encounter |
| Number of procedures | Numeric | Number of procedures (other than lab tests) performed during the encounter |
| Number of medications | Numeric | Number of distinct generic names administered during the encounter |
| Number of outpatient visits | Numeric | Number of outpatient visits of the patient in the year preceding the encounter |
| Number of emergency visits | Numeric | Number of emergency visits of the patient in the year preceding the encounter |
| Number of inpatient visits | Numeric | Number of inpatient visits of the patient in the year preceding the encounter |
| Diagnosis 1 | Nominal | The primary diagnosis (coded as first three digits of ICD9); 848 distinct values |
| Diagnosis 2 | Nominal | Secondary diagnosis (coded as first three digits of ICD9); 923 distinct values |
| Diagnosis 3 | Nominal | Additional secondary diagnosis (coded as first three digits of ICD9); 954 distinct values |
| Number of diagnoses | Numeric | Number of diagnoses entered to the system |
| Glucose serum test result | Nominal | Indicates the range of the result or if the test was not taken. Values: “>200,” “>300,” “normal,” and “none” if not measured |
| A1c test result | Nominal | Indicates the range of the result or if the test was not taken. Values: “>8” if the result was greater than 8%, “>7” if the result was greater than 7% but less than 8%, “normal” if the result was less than 7%, and “none” if not measured. |
| Change of medications | Nominal | Indicates if there was a change in diabetic medications (either dosage or generic name). Values: “change” and “no change” |
| Diabetes medications | Nominal | Indicates if there was any diabetic medication prescribed. Values: “yes” and “no” |
| 24 features for medications | Nominal | For the generic names: metformin, repaglinide, nateglinide, chlorpropamide, glimepiride, acetohexamide, glipizide, glyburide, tolbutamide, pioglitazone, rosiglitazone, acarbose, miglitol, troglitazone, tolazamide, examide, sitagliptin, insulin, glyburide-metformin, glipizide-metformin, glimepiride-pioglitazone, metformin-rosiglitazone, and metformin-pioglitazone, 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 |
| Readmitted | Nominal | Days to inpatient readmission. Values: “<30” if the patient was readmitted in less than 30 days, “>30” if the patient was readmitted in more than 30 days, and “No” for no record of readmission. |

## Check for Missing Values

As a first step we have checked for missing values in each of the features.

Below are the features that have the missing values.



We have observed that Weight though is an important feature has 96.86% of missing values.

Payer code has 39.56% and medical specialty has 49.08% missing values.

## Feature Selection

Below are features that are unique to patient/encounter and are not relevant in our model

* encounter\_id
* patient\_nbr

Below are features that have high missing values and so are dropped

* weight
* payer\_code

Below are redundant/not relevant features that are dropped

* admission\_type\_id

Below are medications used by patients and they are also removed.

* metformin
* repaglinide
* nateglinide
* chlorpropamide
* glimepiride
* acetohexamide
* glipizide
* glyburide
* tolbutamide
* pioglitazone
* rosiglitazone
* acarbose
* miglitol
* troglitazone
* tolazamide
* examide
* citoglipton
* glyburide.metformin
* glipizide.metformin
* glimepiride.pioglitazone
* metformin.rosiglitazone
* metformin.pioglitazone

## Missing value treatment/coding

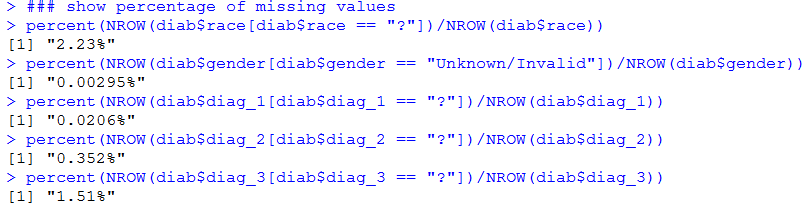
Below mentioned is the way we have handled missing value treatment for features that are identified are needed.

* Medical specialty is coded with value "Unknown" where ever it was missing.

## Removal of unwanted rows

We have removed unwanted rows from the dataset as described below.

* Removed rows that had gender as "Unknown/Invalid".
* Removed rows that had missing values in race.
* Removed rows that had missing diagnosis codes.
* Removed rows where discharge disposition code resulted in death of the patient.
* Only retained the first occurrence of each unique patient and discarded the rest of encounters.



## No. of unique patients

The data set has 67571 unique patients.



## Factor levels handling

Features that have factor levels are handled in below way.

* Three diagnosis codes are split into below factors based on their ICD9 diagnosis values.

|  |  |  |
| --- | --- | --- |
| **Group name** | **ICD9 codes** | **Description** |
| Circulatory | 390–459, 785 | Diseases of the circulatory system |
| Respiratory | 460–519, 786 | Diseases of the respiratory system |
| Digestive | 520–579, 787 | Diseases of the digestive system |
| Diabetes | 250.xx | Diabetes mellitus |
| Injury | 800–999 | Injury and poisoning |
| Musculoskeletal | 710–739 | Diseases of the musculoskeletal system and connective tissue |
| Genitourinary | 580–629, 788 | Diseases of the genitourinary system |
| Neoplasms | 140–239 | Neoplasms |
| 780, 781, 784, 790–799 | Other symptoms, signs, and ill-defined conditions |
| 240–279, without 250 | Endocrine, nutritional, and metabolic diseases and immunity disorders, without diabetes |
| 680–709, 782 | Diseases of the skin and subcutaneous tissue |
| 001–139 | Infectious and parasitic diseases |
| Other | 290–319 | Mental disorders |
| E–V | External causes of injury and supplemental classification |
| 280–289 | Diseases of the blood and blood-forming organs |
| 320–359 | Diseases of the nervous system |
| 630–679 | Complications of pregnancy, childbirth, and the puerperium |
| 360–389 | Diseases of the sense organs |
| 740–759 | Congenital anomalies |

* Age is treated as three factor levels: “<30”, “30 to 60” and “60 to 100”.
* Medical specialty is divided into factor levels as "Unknown", "Cardiology", "General practice", "Internal medicine", "Surgery" and "Other".
* No readmission and readmission after 30 days (>30) are treated as "NoReadmission" and if readmitted within 30 days (<30) is treated as "Readmission".

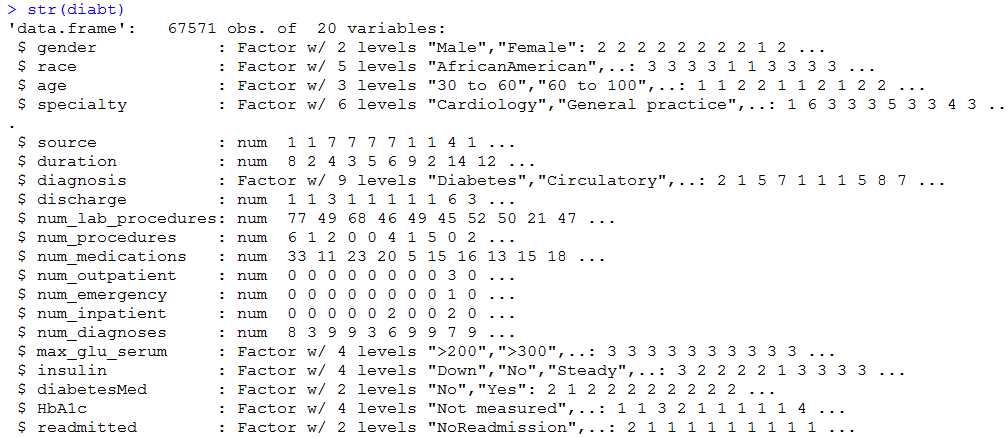
## Feature aggregation

Feature that are related with each other are aggregated in below fashion.

* Single diagnosis code is derived from the three available diagnosis codes. If one of the three diagnosis code is diabetes related then it is kept otherwise the primary diagnosis code is considered.
* Hb1Ac is arrived using below conditions using the actual Hb1Ac measure and Change in medications based on the test.
* If value is "None" then it is imputed as "Not measured".
* If value is "Norm" then it is imputed as "Normal".
* If value is ">8" or ">7" and there is no change in medication prescribed then value of "High and not changed" is imputed.
* If value is ">8" or ">7" and there is change in medication prescribed then value of "High and changed" is imputed.

## Final dataset

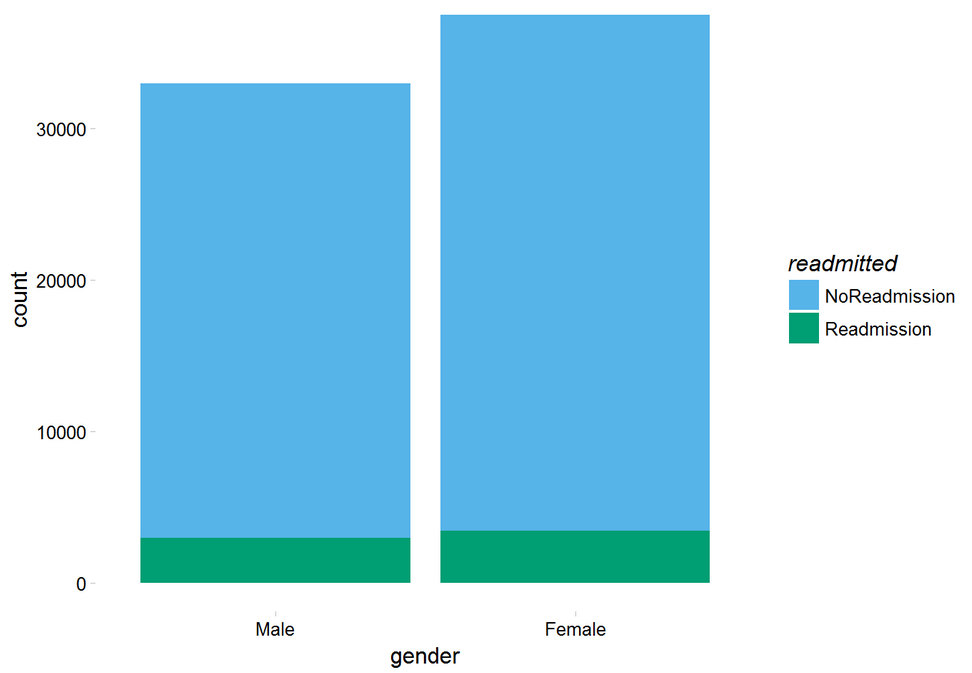
Final dataset that was derived consists of 20 features and 67571 rows (unique patients). Dependent feature is “readmitted” and rest 19 features are independent variables.



# Exploratory Data Analysis

Following were the insights generated from the final dataset:

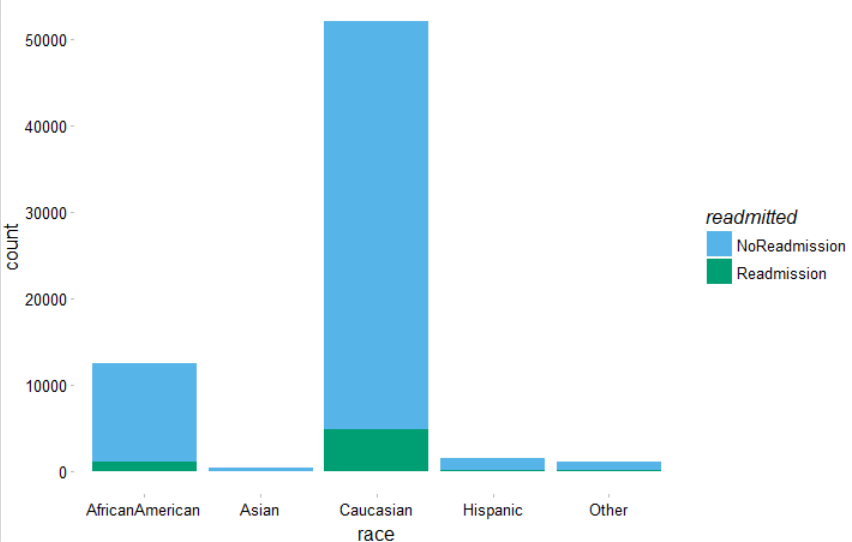
## Gender vs Readmission

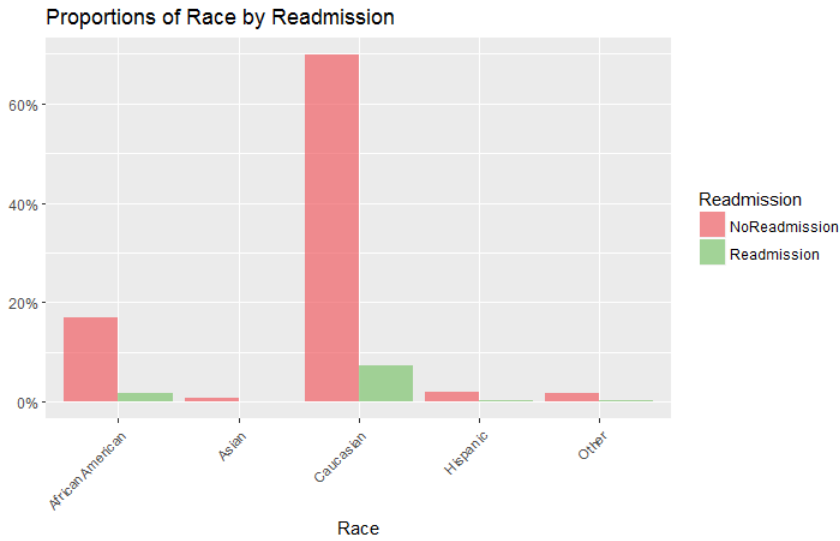




From the above charts, we can gather that the readmission percent is similar in both the genders. We see that female patients are more than male patients.

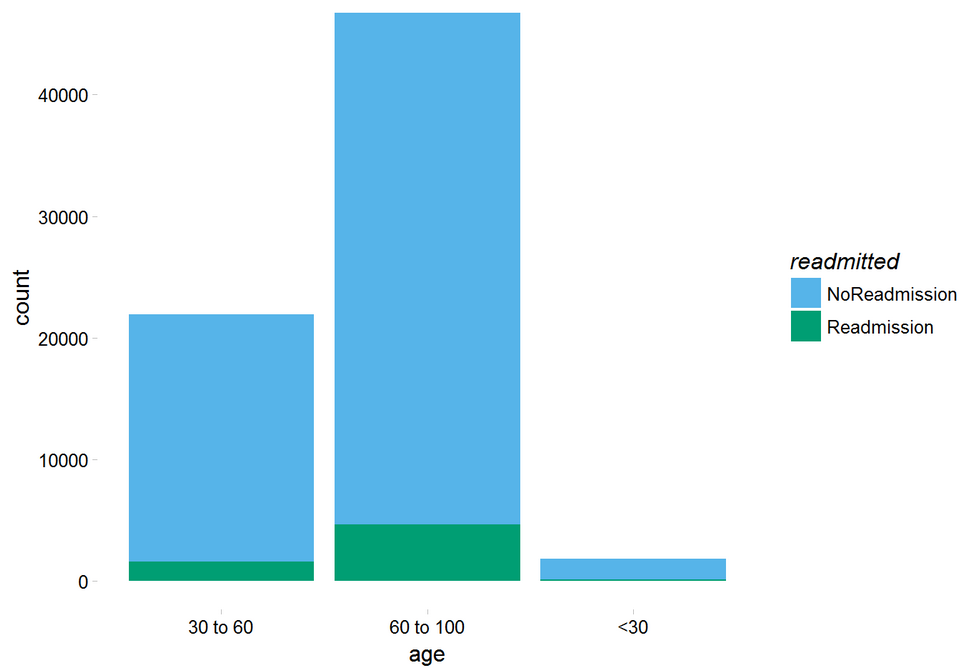
## Race vs Readmission

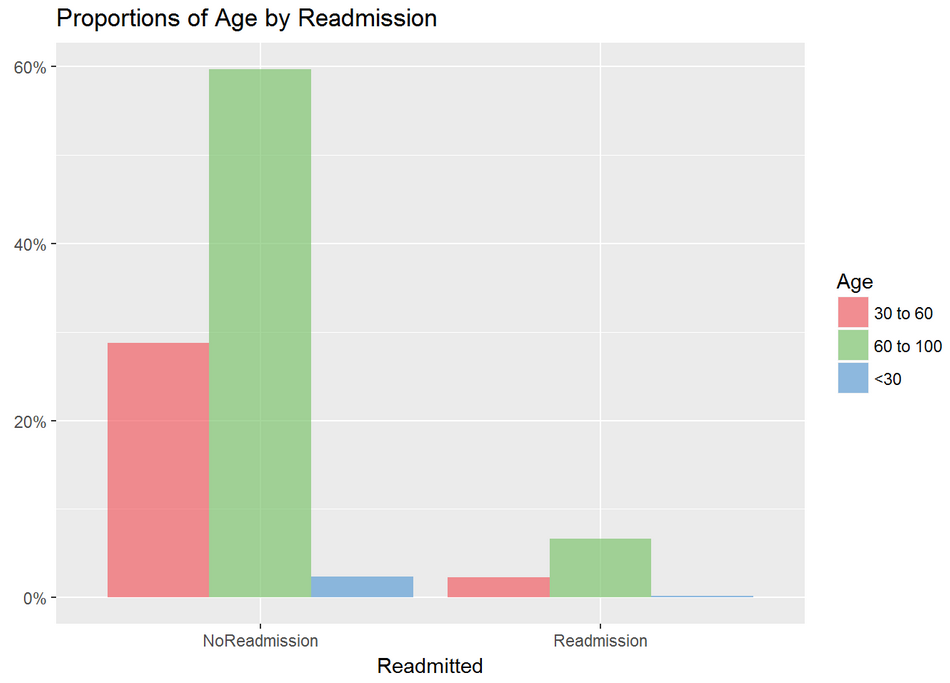




From the above charts, we can gather that the readmission percent is similar across different type of races. We observe that Caucasian patients are more in numbers followed by African American.

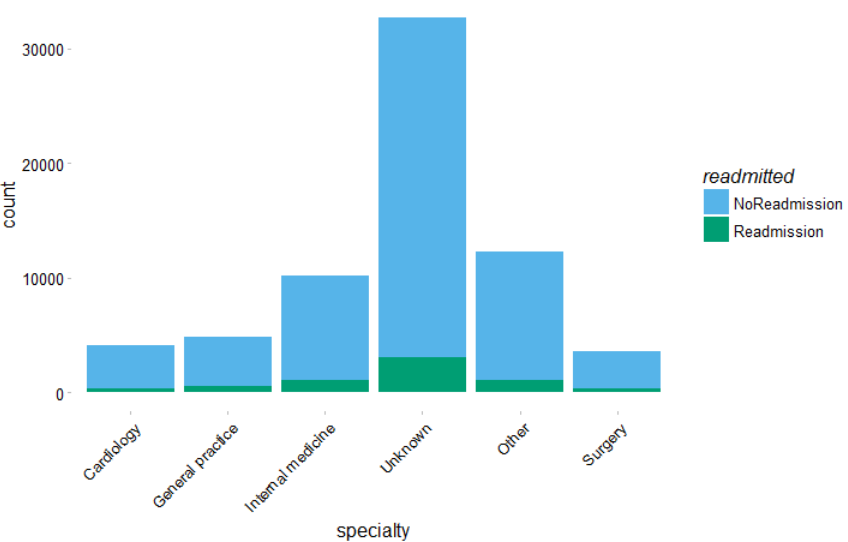
## Age vs Readmission

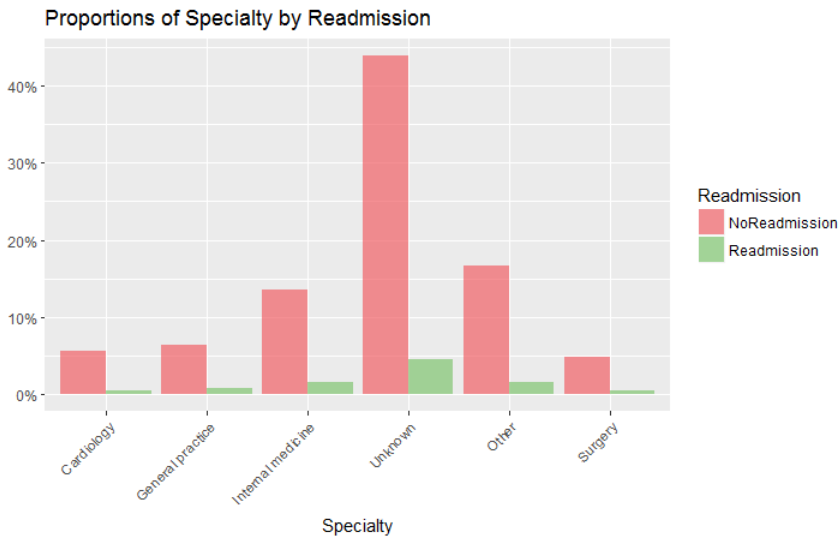




From the above charts, we can gather that the readmission percent is similar across different age groups. We observe that patients in age group 60 to 100 are huge in numbers.

## Medical Specialty vs Readmission





From the above charts, we can gather that the readmission percent is similar across different medical specialties.

## Time in Hospital vs Readmission

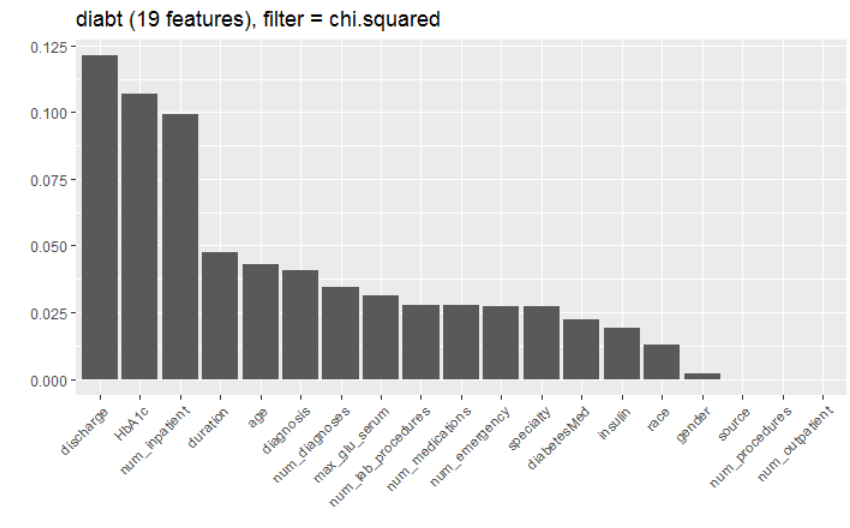


From the above chart, we can gather that the readmission percent is similar across different length of stay in hospital.

# Chi-Square Test

Post our data exploration and preparation; we went about doing chi-square test for all the features. Below is the graphic output of the results.

## Chi-Square Summary

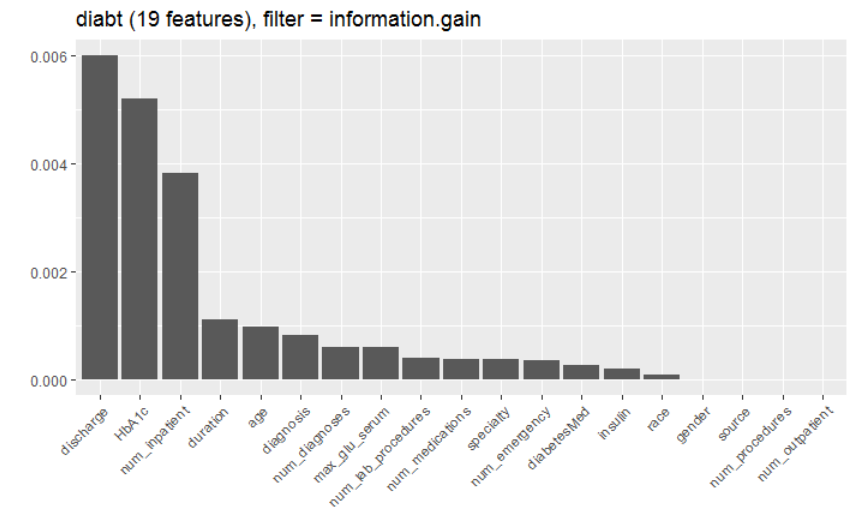


We can observe that highest dependence is on discharge, followed by HbA1c test measure.

# Information Gain

Post this; we have gone about doing Information Gain for the purpose of checking significant variables.

## Summary Plot



We can observe that discharge is the most significant feature followed by HbA1c.

Based on the above graph we have dropped variables “race”, “gender”, “source”, “num\_procedures” and “num\_outpatient” .

# Training and Test data

We divided the data into training and test with a split of 70:30 and developed the model on the training data set.

# Model Development

On the training data set first we have built base logistic regression model.

## Logistic regression model summary

Below is the summary of the base model.

Call:

glm(formula = readmitted ~ ., family = binomial, data = train)

Deviance Residuals:

Min 1Q Median 3Q Max

-2.0236 -0.4562 -0.3927 -0.3283 2.8160

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) -3.4112974 0.1793916 -19.016 < 2e-16 \*\*\*

age60 to 100 0.2530220 0.0386998 6.538 6.23e-11 \*\*\*

age<30 0.1154554 0.1270686 0.909 0.363558

specialtyGeneral practice 0.3252368 0.0950711 3.421 0.000624 \*\*\*

specialtyInternal medicine 0.2388585 0.0855965 2.791 0.005262 \*\*

specialtyUnknown 0.2088953 0.0783799 2.665 0.007695 \*\*

specialtyOther 0.2111094 0.0855856 2.467 0.013639 \*

specialtySurgery 0.1805421 0.1050135 1.719 0.085573 .

duration 0.0228241 0.0063641 3.586 0.000335 \*\*\*

diagnosisCirculatory 0.1534930 0.0474347 3.236 0.001213 \*\*

diagnosisDigestive 0.0389079 0.0734872 0.529 0.596492

diagnosisGenitourinary 0.0177188 0.0920661 0.192 0.847384

diagnosisInjury 0.2449036 0.0741492 3.303 0.000957 \*\*\*

diagnosisMusculoskeletal -0.0430022 0.1012270 -0.425 0.670976

diagnosisNeoplasms 0.2041384 0.0633850 3.221 0.001279 \*\*

diagnosisOther 0.0441704 0.0718192 0.615 0.538540

diagnosisRespiratory -0.0093296 0.0645797 -0.144 0.885133

discharge 0.0330053 0.0026995 12.226 < 2e-16 \*\*\*

num\_lab\_procedures 0.0027803 0.0009451 2.942 0.003263 \*\*

num\_medications 0.0003060 0.0023687 0.129 0.897222

num\_emergency 0.0786125 0.0269126 2.921 0.003489 \*\*

num\_inpatient 0.3643774 0.0195979 18.593 < 2e-16 \*\*\*

num\_diagnoses 0.0385217 0.0103049 3.738 0.000185 \*\*\*

max\_glu\_serum>300 -0.8792816 0.2064893 -4.258 2.06e-05 \*\*\*

max\_glu\_serumNone -0.1306328 0.1311985 -0.996 0.319402

max\_glu\_serumNorm 0.0247823 0.1619284 0.153 0.878363

insulinNo -0.1391967 0.0601960 -2.312 0.020756 \*

insulinSteady -0.1563706 0.0583763 -2.679 0.007392 \*\*

insulinUp -0.0815810 0.0731453 -1.115 0.264710

diabetesMedYes 0.2137961 0.0471793 4.532 5.85e-06 \*\*\*

HbA1cHigh and changed -0.7533224 0.0904944 -8.325 < 2e-16 \*\*\*

HbA1cHigh and not changed 1.0395852 0.0580358 17.913 < 2e-16 \*\*\*

HbA1cNormal -0.7764978 0.0988797 -7.853 4.06e-15 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 28815 on 47303 degrees of freedom

Residual deviance: 27486 on 47271 degrees of freedom

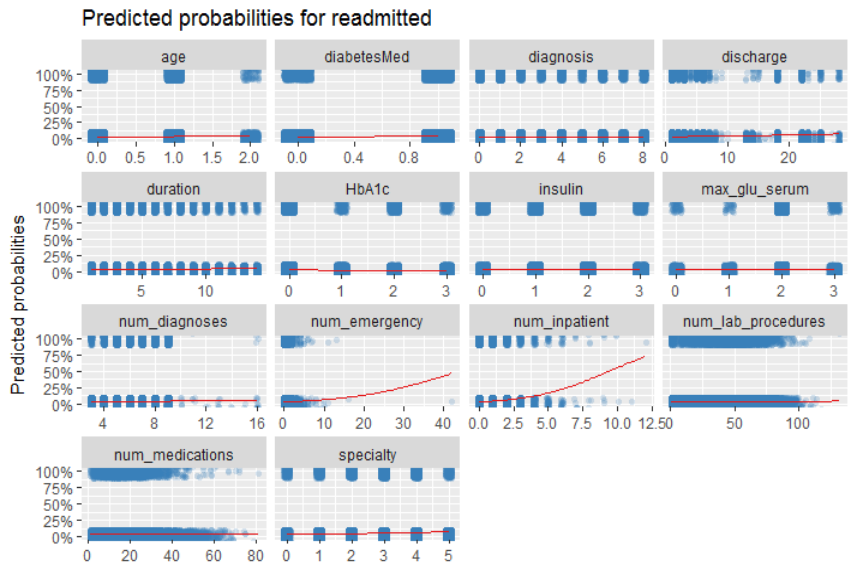
AIC: 27552

Number of Fisher Scoring iterations: 6

## Model Odds Ratios



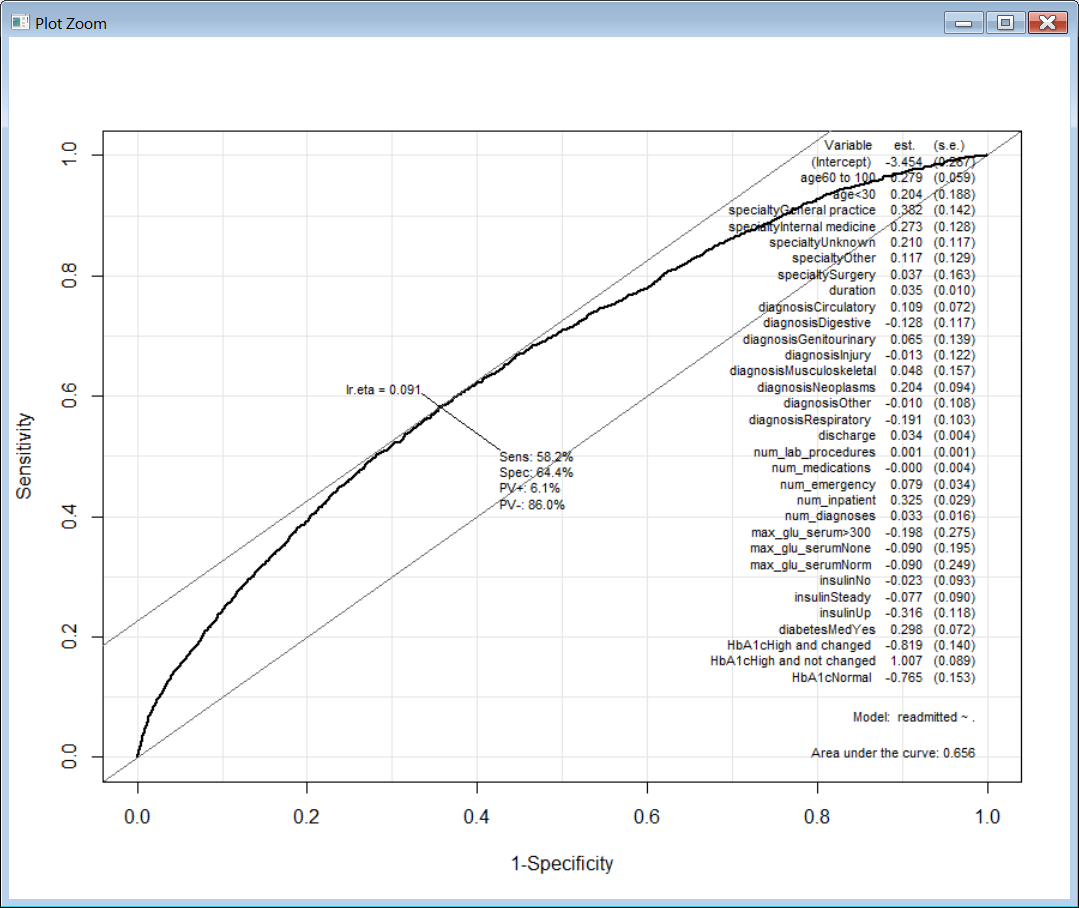
## Predicted probabilities for readmitted



# Model Validation

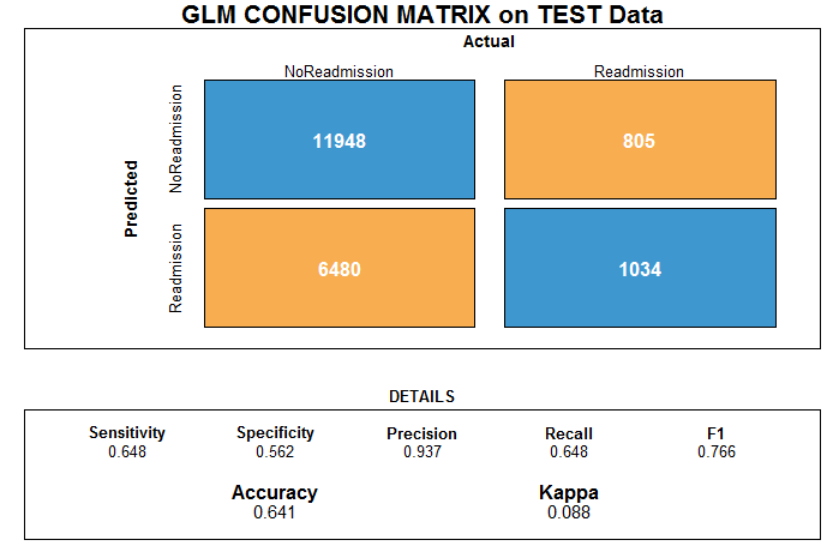
## AUC Curve

Below is the AUC curve for the base model on test data. Area under curve is 65.6%



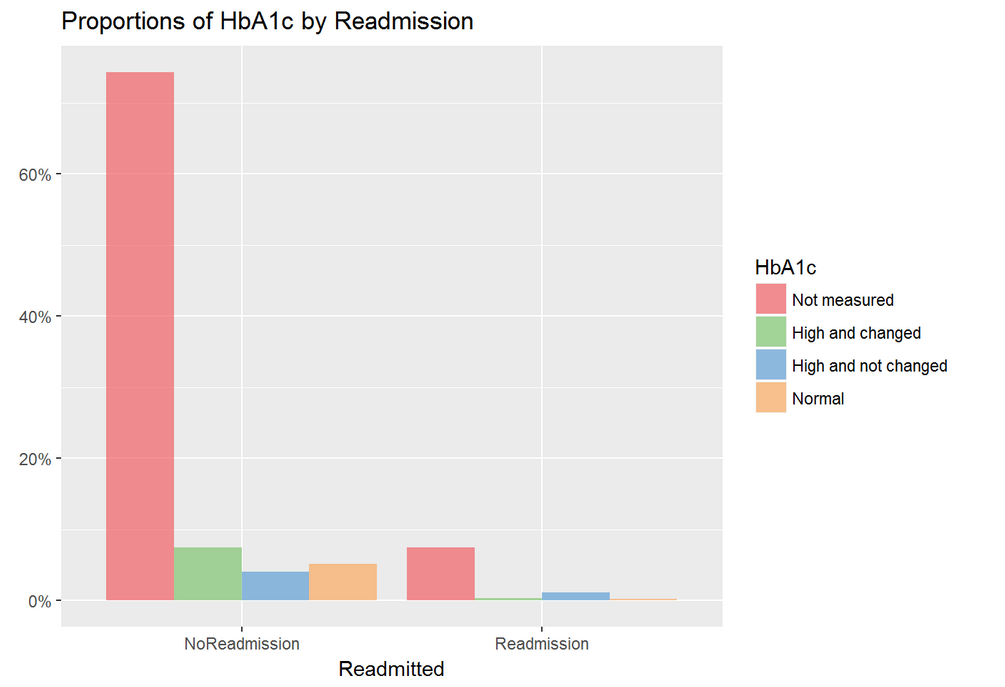
## Accuracy

We have run our base model on the test data set and we got an accuracy of 64%.



## HbA1c vs Readmission

From below graph we can observe that on 80% of the cases HbA1c test was not performed.

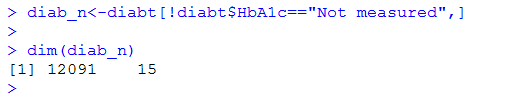


So we have decided to remove the encounters where HbA1c test was not performed and verify how the model will perform.

# New Model Development on segregated data

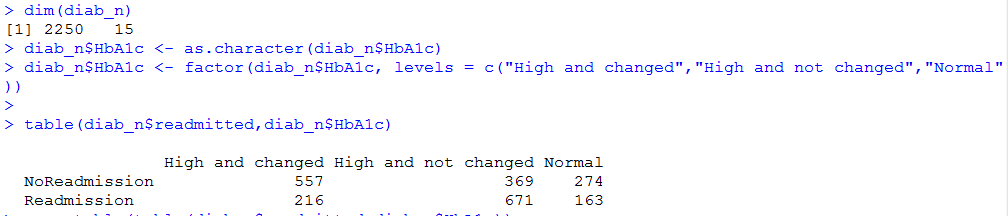
## Removal of encounters where HbA1c test was not performed

We have removed encounters where HbA1c test was not performed and the resultant dataset has 12091 encounters.

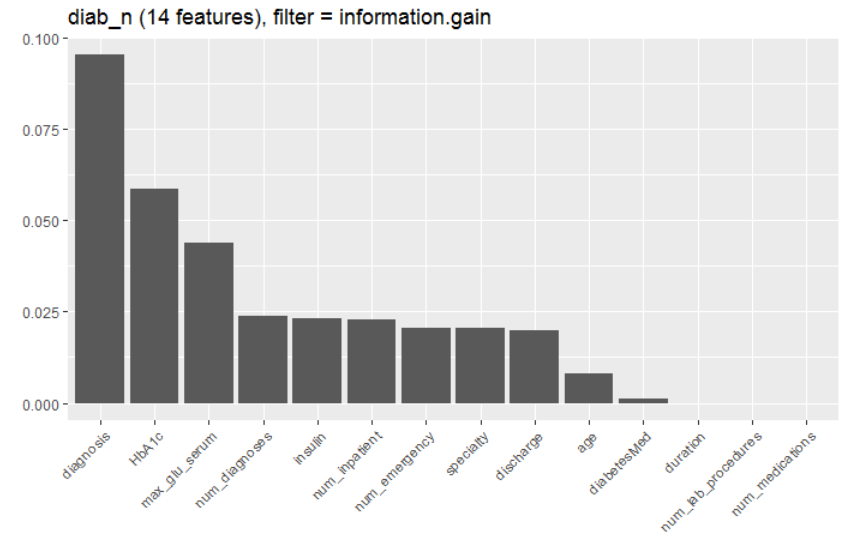


## Training and test split

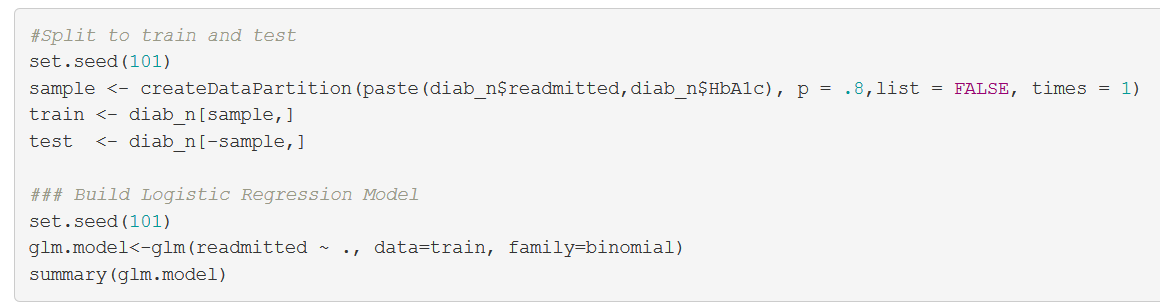
We have decided to split this new dataset that has 12091 encounters into training and test data sets with split ration of 70:30 by maintain equal proportions of HbA1c feature and build new logistic regression model on the same to verify how it performs on this new data that has encounters where HbA1c test was performed.



Information gain chart below for the data set.



Based on above result drop features “duration”, “num\_lab\_procedures” and “num\_medications”.



## New Logistic regression model summary

Below is the summary of the new logistic regression model on train data set.

Call:

glm(formula = readmitted ~ ., family = binomial, data = train)

Deviance Residuals:

Min 1Q Median 3Q Max

-2.7499 -0.8040 -0.2433 0.7928 2.4212

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) 12.317659 430.449496 0.029 0.977171

age60 to 100 0.200807 0.129531 1.550 0.121078

age<30 -0.148389 0.332334 -0.447 0.655232

specialtyGeneral practice 0.474481 0.306893 1.546 0.122085

specialtyInternal medicine 0.218878 0.275162 0.795 0.426351

specialtyUnknown 0.889505 0.268188 3.317 0.000911 \*\*\*

specialtyOther 0.768341 0.294740 2.607 0.009138 \*\*

specialtySurgery 0.401597 0.379072 1.059 0.289407

diagnosisCirculatory 1.294457 0.158931 8.145 3.80e-16 \*\*\*

diagnosisDigestive 1.253622 0.303105 4.136 3.54e-05 \*\*\*

diagnosisGenitourinary 1.371082 0.401431 3.415 0.000637 \*\*\*

diagnosisInjury 0.861176 0.317786 2.710 0.006730 \*\*

diagnosisMusculoskeletal 0.801364 0.501315 1.599 0.109926

diagnosisNeoplasms 2.233156 0.225186 9.917 < 2e-16 \*\*\*

diagnosisOther 0.988502 0.293431 3.369 0.000755 \*\*\*

diagnosisRespiratory 1.189842 0.228943 5.197 2.02e-07 \*\*\*

discharge -0.015923 0.009122 -1.746 0.080876 .

num\_emergency 1.622735 0.434170 3.738 0.000186 \*\*\*

num\_inpatient 0.821402 0.156989 5.232 1.67e-07 \*\*\*

num\_diagnoses 0.173854 0.034226 5.080 3.78e-07 \*\*\*

max\_glu\_serum>300 -18.495114 430.449504 -0.043 0.965728

max\_glu\_serumNone -16.507710 430.449325 -0.038 0.969409

max\_glu\_serumNorm -15.449155 430.449814 -0.036 0.971369

insulinNo 0.167940 0.240426 0.699 0.484857

insulinSteady -0.529029 0.244495 -2.164 0.030483 \*

insulinUp 0.460733 0.323325 1.425 0.154162

diabetesMedYes 0.909510 0.171556 5.302 1.15e-07 \*\*\*

HbA1cHigh and not changed 1.390286 0.170007 8.178 2.89e-16 \*\*\*

HbA1cNormal 0.136124 0.185668 0.733 0.463462

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

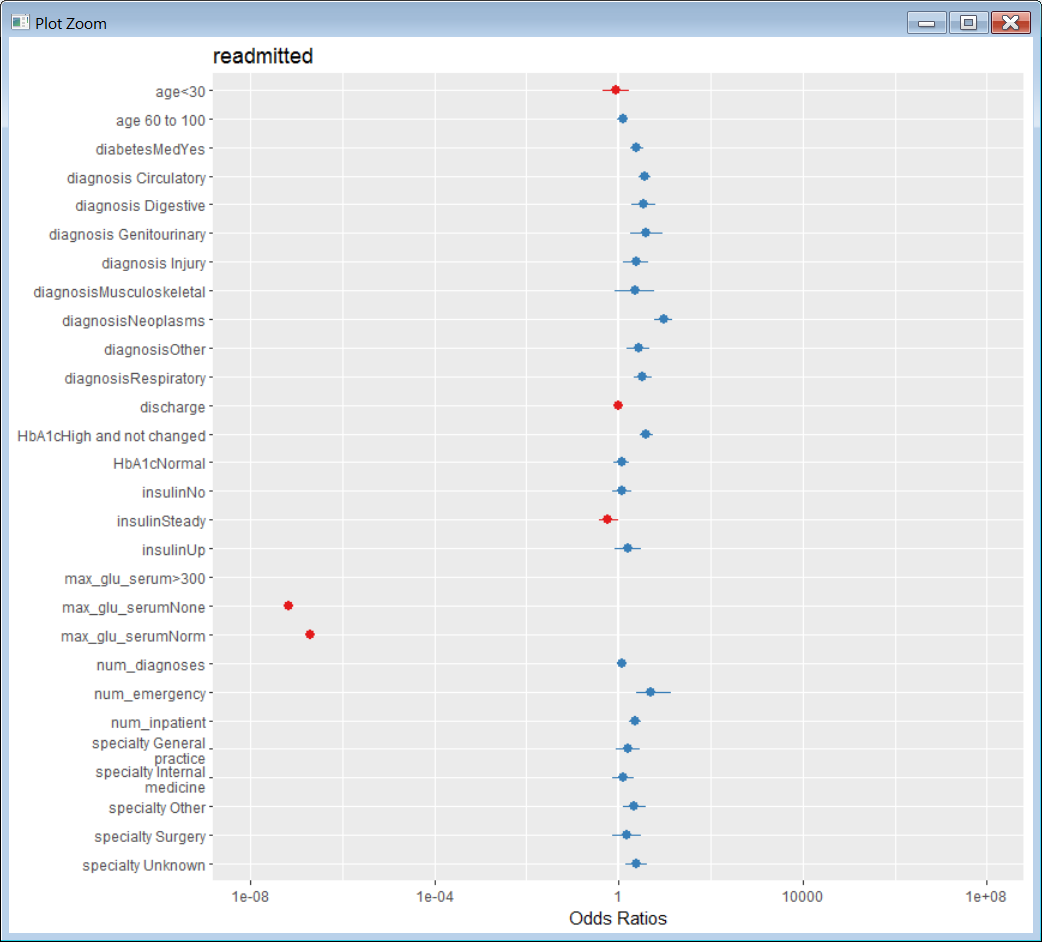
Null deviance: 2491.4 on 1802 degrees of freedom

Residual deviance: 1760.9 on 1774 degrees of freedom

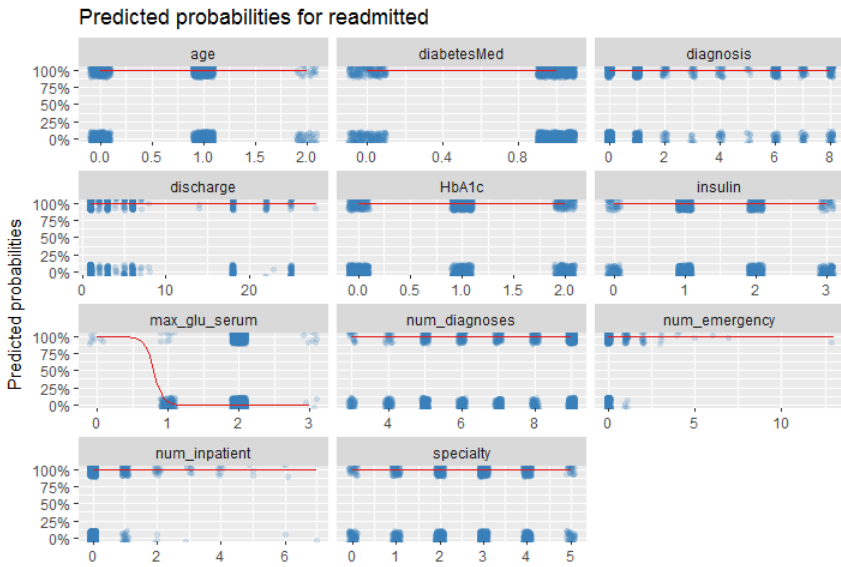
AIC: 1818.9

Number of Fisher Scoring iterations: 14

## New Model Odds Ratios



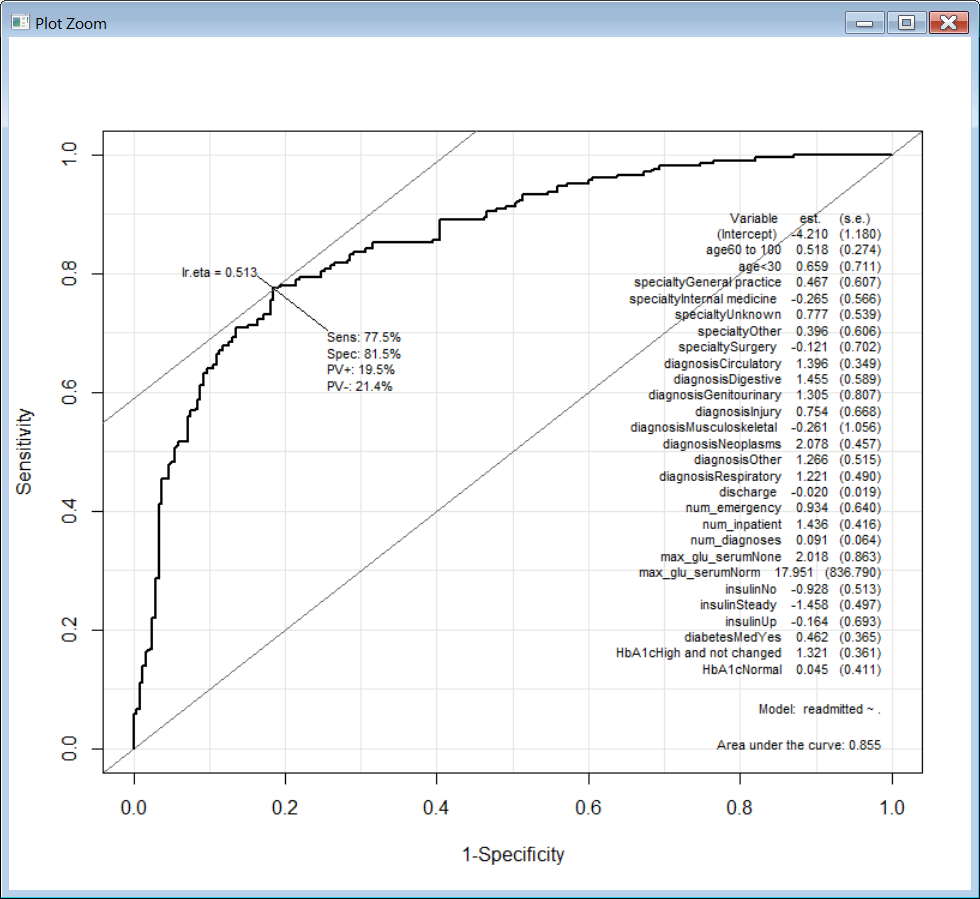
## New Model Predicted probabilities for readmitted



# New Model Validation

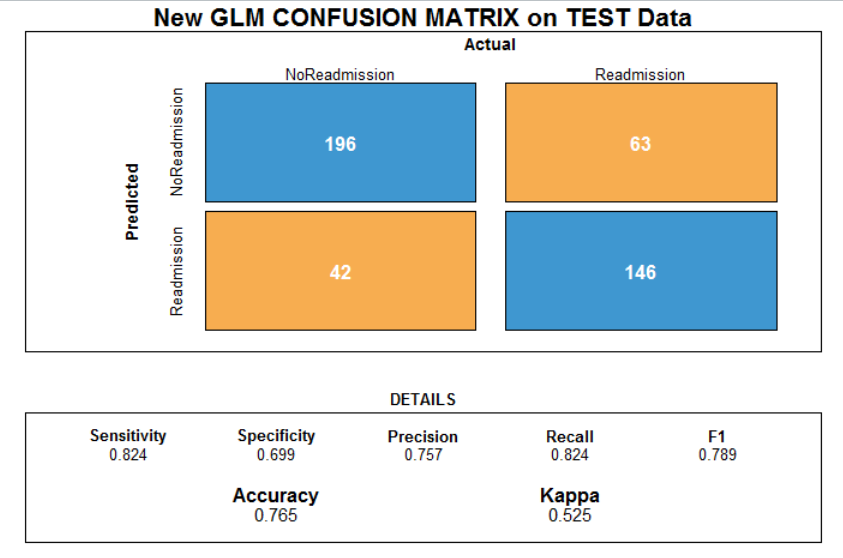
## AUC Curve

Area under curve has improved to 85.5% on the segregated data that had only encounters where HbA1c test was performed.



## Accuracy

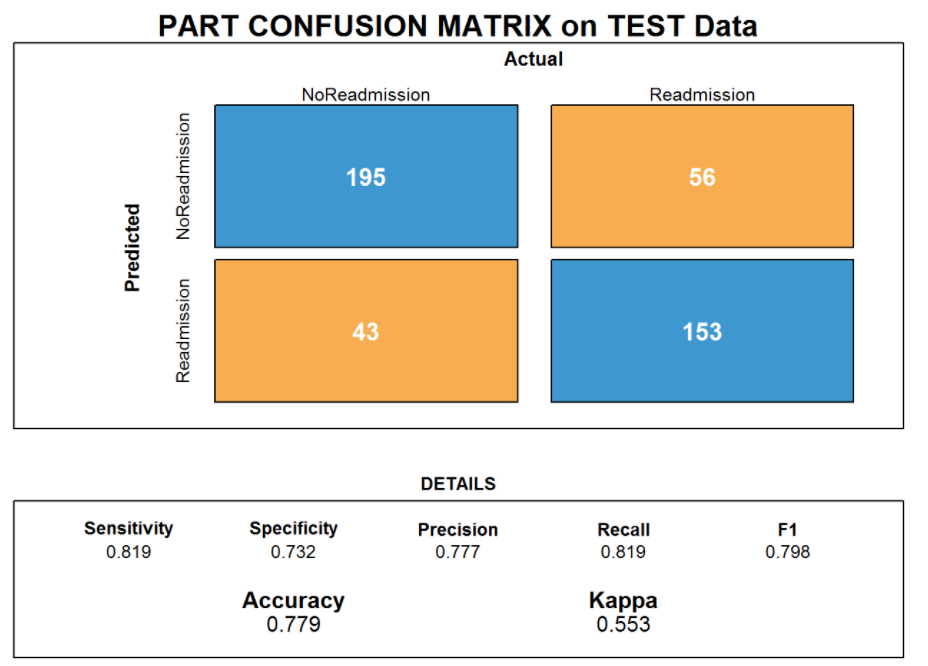
On the new data set we have observed that the accuracy has improved to 76.5% on the test data using the logistic regression model.



# Recommendations and Conclusions

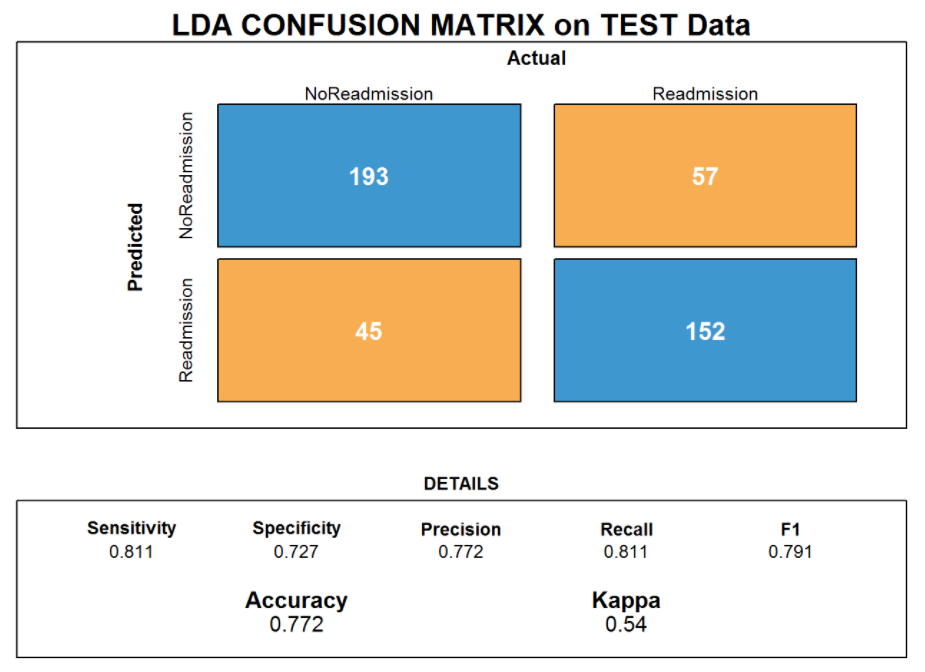
## Best Model Accuracy

Apart from our initial logistic regression model we have run different other models and have observed that **PART** model has given us the highest accuracy of **77.9%.**

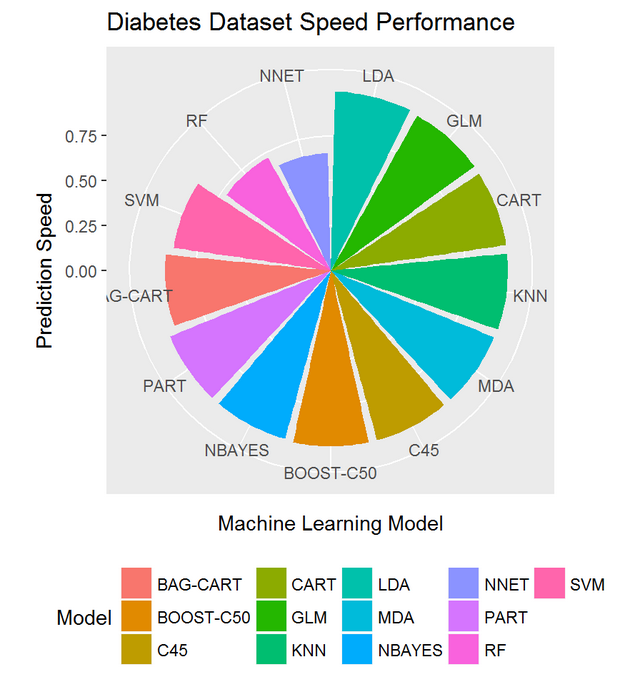


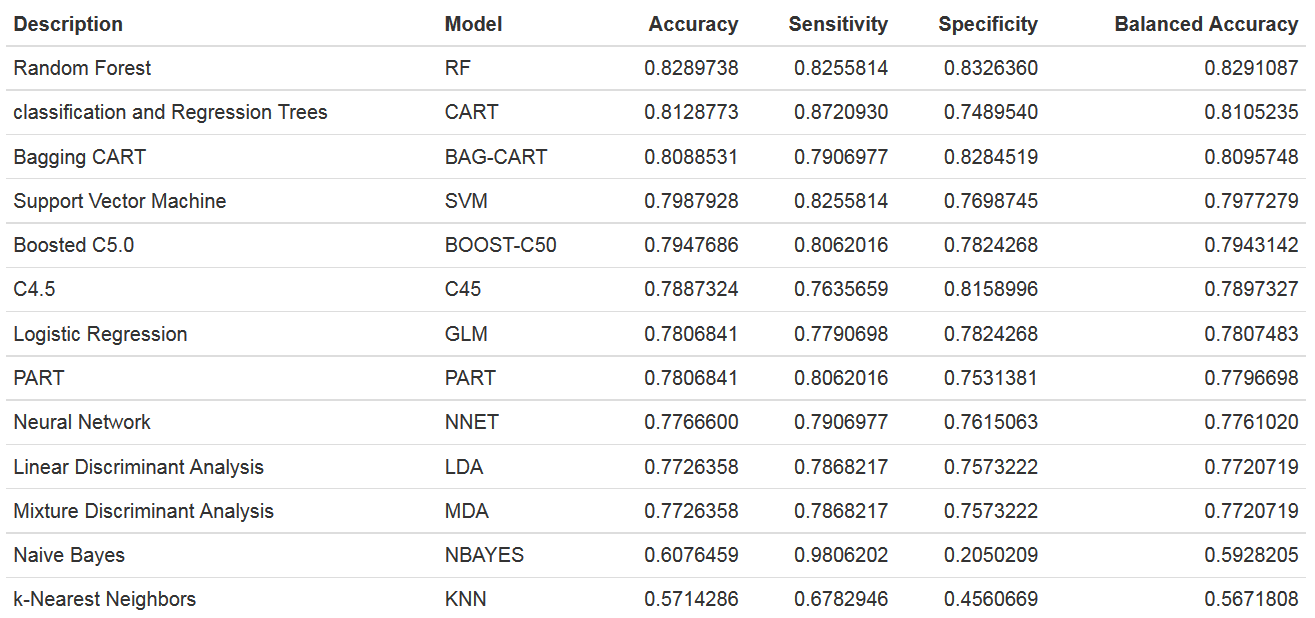
## Best Model Recommendation

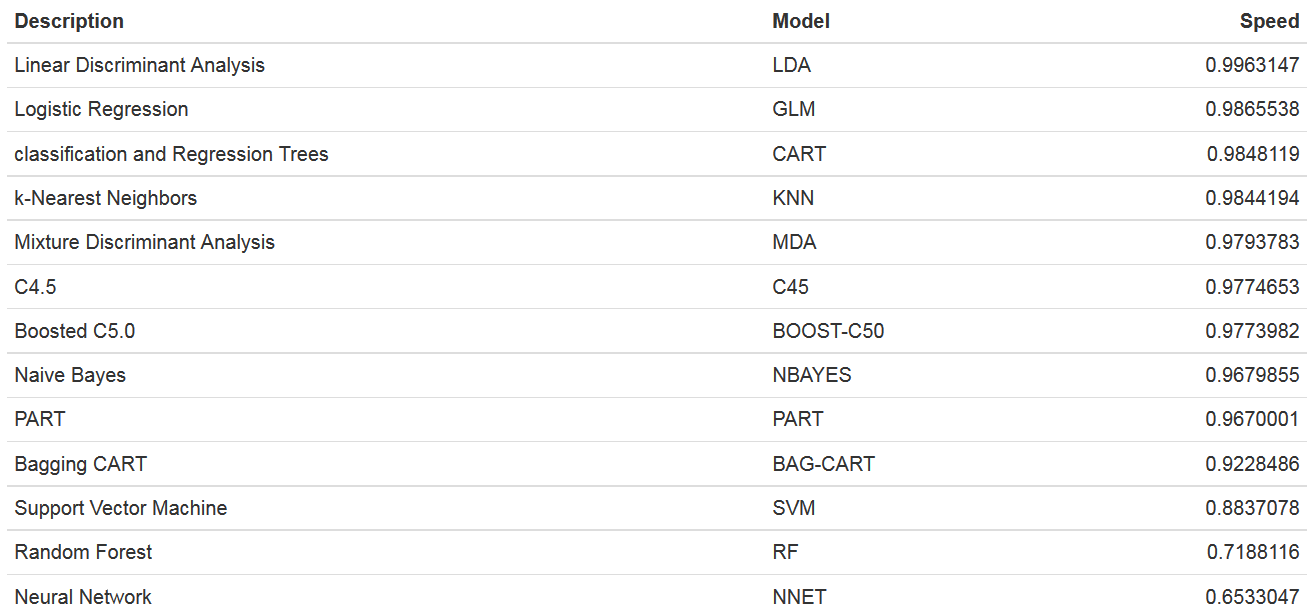
After running various models and verifying their accuracy and run time we would recommend **LDA** model which looks best in terms of both and accuracy and run time.











## Conclusion

In conclusion, the decision to obtain a measurement of HbA1c for patients with diabetes mellitus is a useful predictor of readmission rates which may prove valuable in the development of strategies to reduce readmission rates and costs for the care of individuals with diabetes mellitus. For instance, our model showed that the profile of readmission differed significantly in patients where HbA1c was checked in the setting of a diabetes diagnosis, when compared to those with other diagnosis.

# Bibliography

<https://archive.ics.uci.edu/ml/machine-learning-databases/00296/>

<http://spectrum.diabetesjournals.org/content/21/4/248>

<https://www.scopus.com/record/display.uri?eid=2-s2.0-79961200771&origin=inward&txGid=fe39fc366b8dd9713a6282de99fa6cd2>

<https://www.r-project.org/>

<http://rmarkdown.rstudio.com/>

<http://r-statistics.co/Top50-Ggplot2-Visualizations-MasterList-R-Code.html>

# Annexure

## Contacts

Contact details of the students who worked on this capstone project.

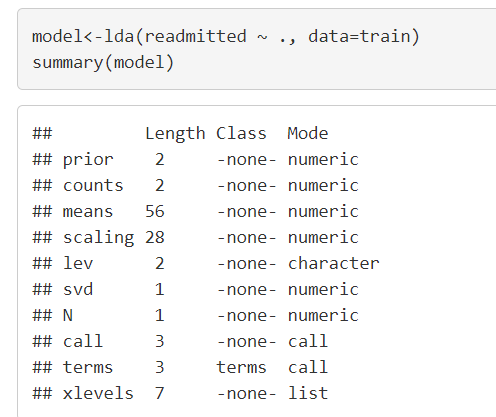
|  |  |  |  |
| --- | --- | --- | --- |
| **NAME** | **ENROLLMENT#** | **MOBILE#** | **E-MAIL** |
| Girish Velivala | BAHMAR17018 | 98497 20131 | [girish.velivala@gmail.com](mailto:girish.velivala@gmail.com) |
| Aravind Rao Bhaskar | BAHMAR17003 | 90031 22543 | <raoaravindbhaskar91@gmail.com> |
| Ashwanth Prathapani | BAHMAR17007 | 97004 82282 | <ashwanthprathapani@gmail.com> |
| Sumit Chander | BAHMAR17067 | 99854 40964 | <sumit.rhl@gmail.com> |

Below are various other models run and the details.

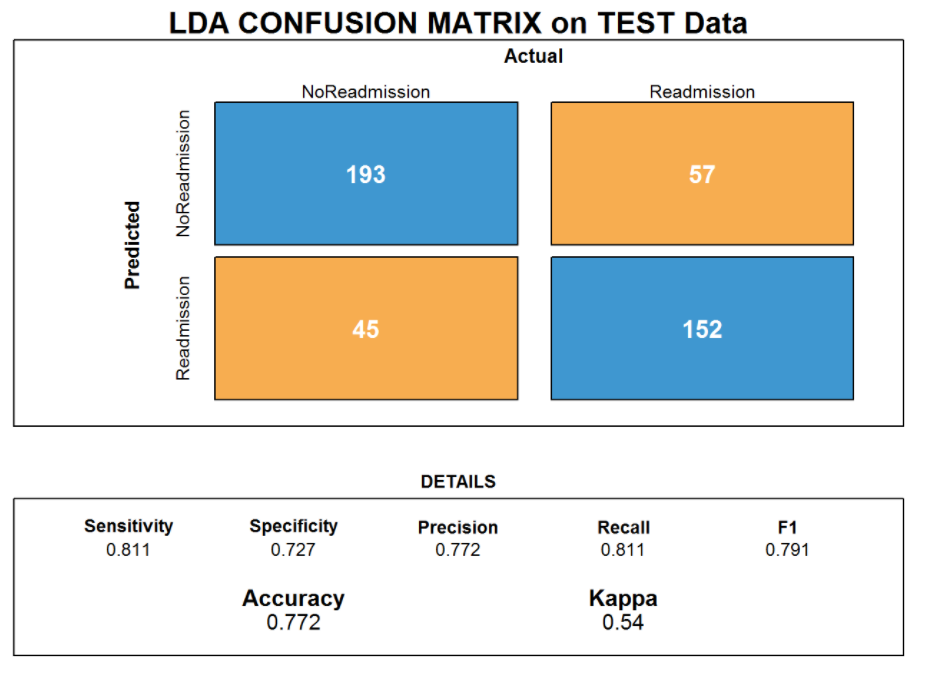
## LDA

Linear Discriminant Analysis

### Summary



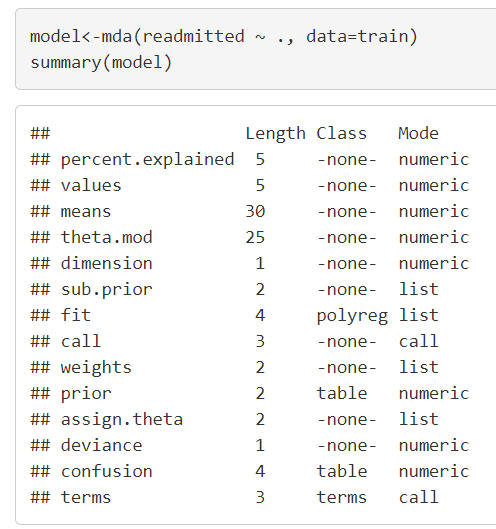
### Accuracy



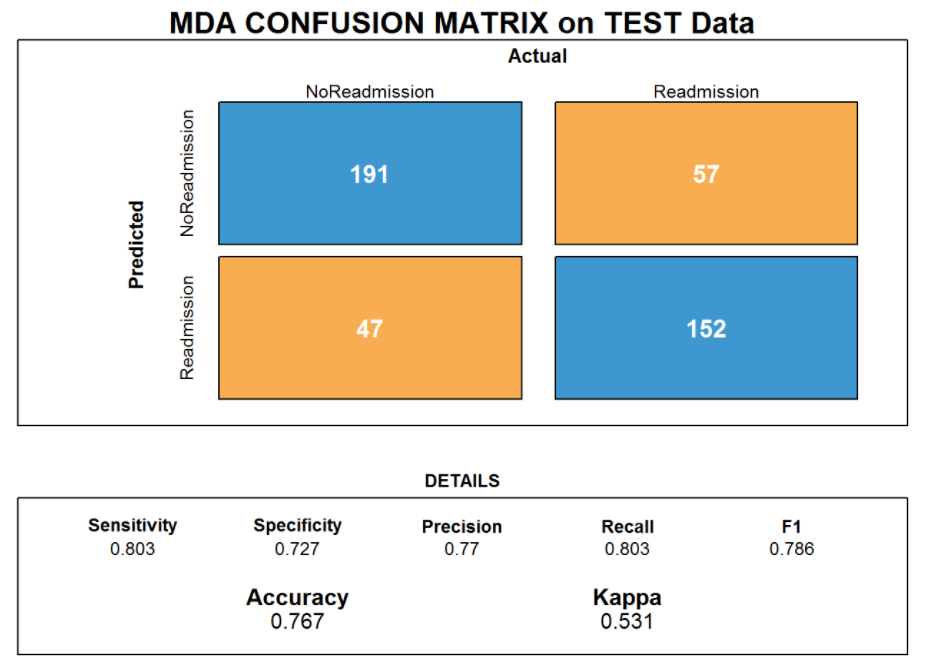
## MDA

Mixture Discriminant Analysis

### Summary



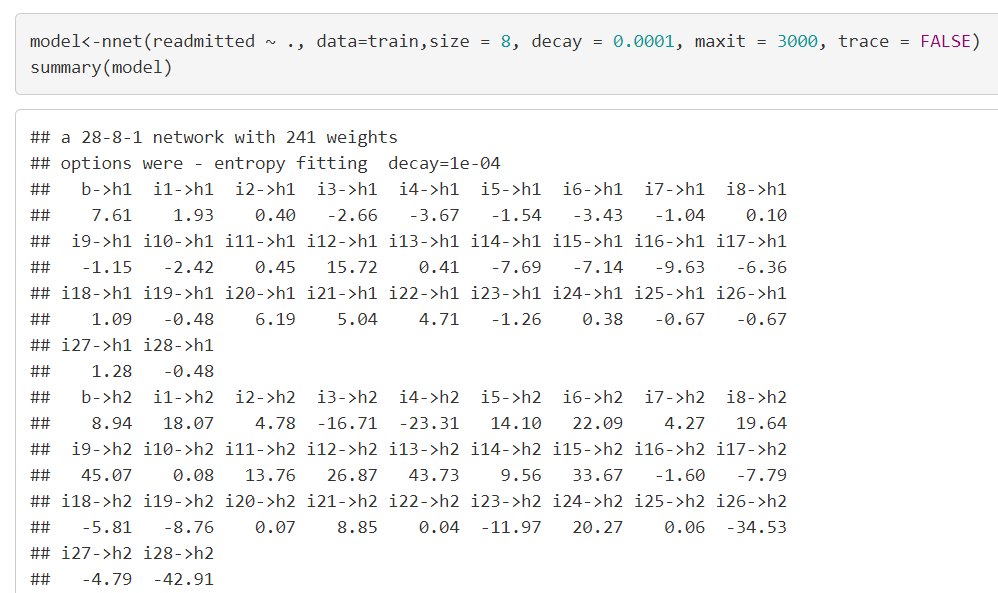
### Accuracy



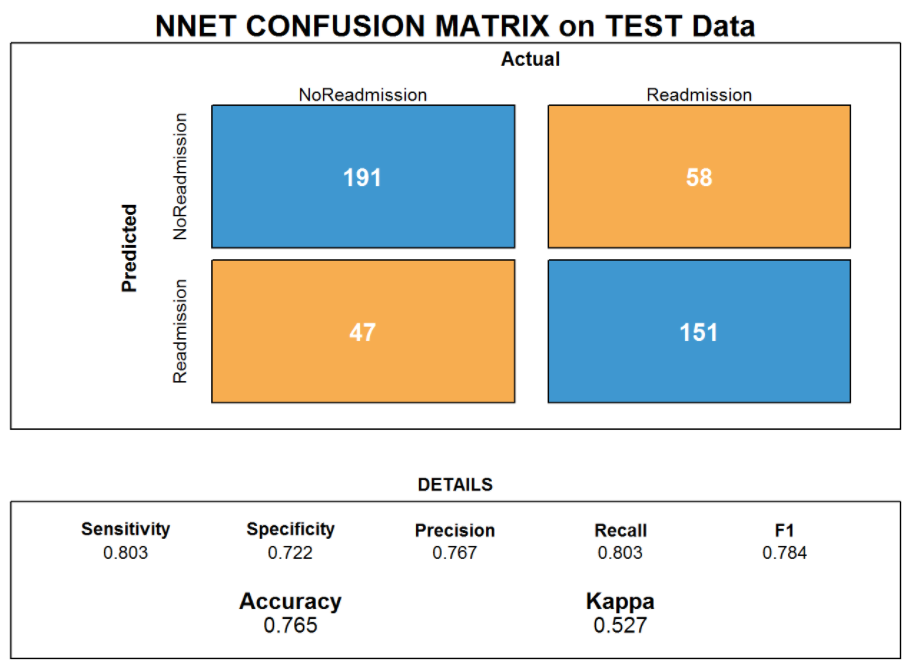
## Neural Net

Neural Net

### Summary



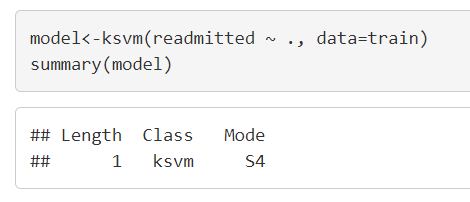
### Accuracy



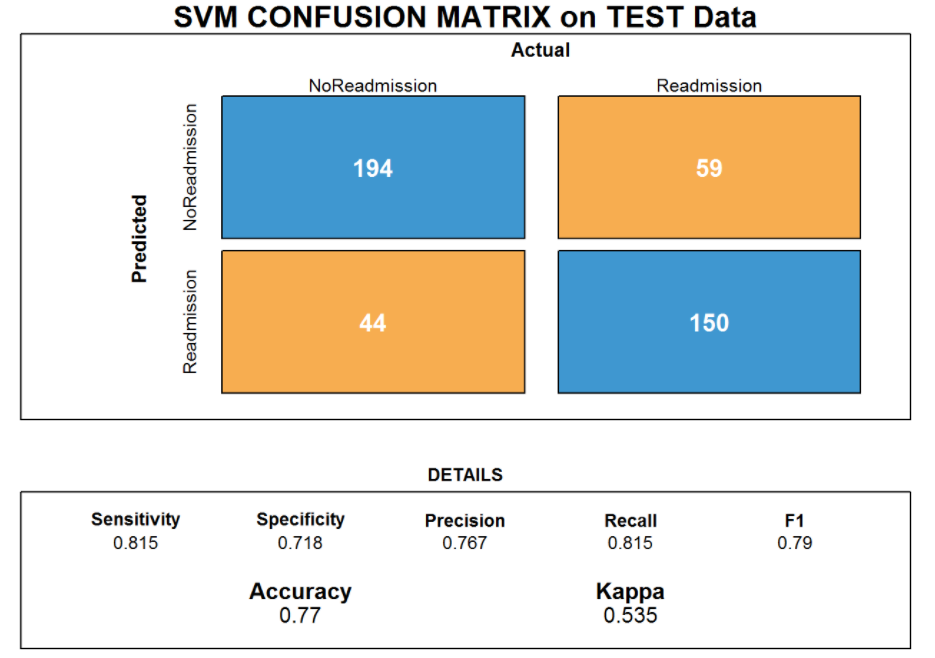
## SVM

Support Vector Machine

### Summary



### Accuracy



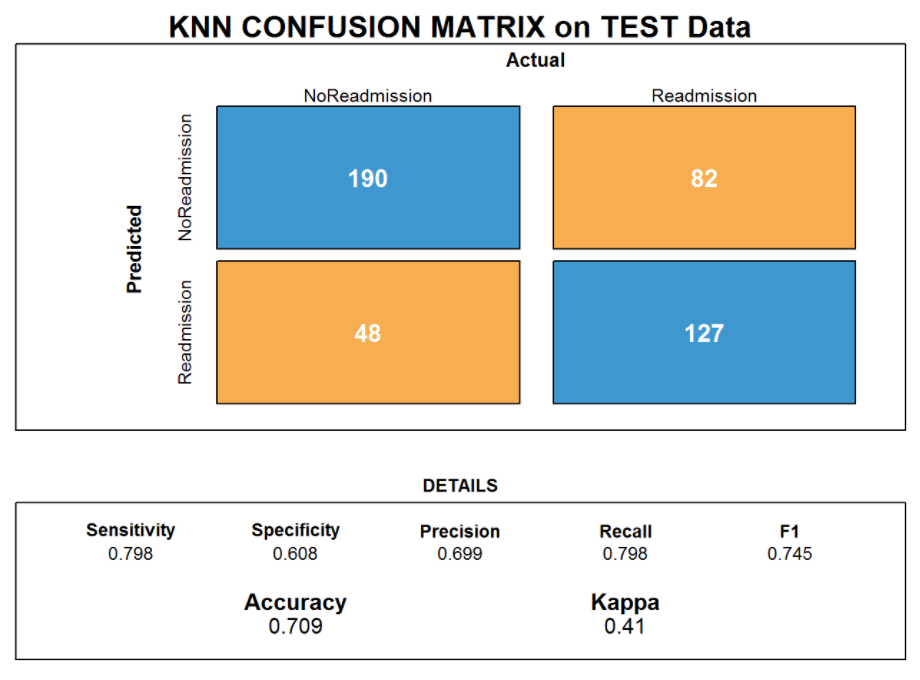
## KNN

k-Nearest Neighbors

### Summary



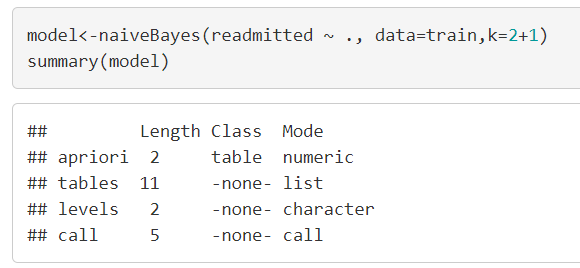
### Accuracy



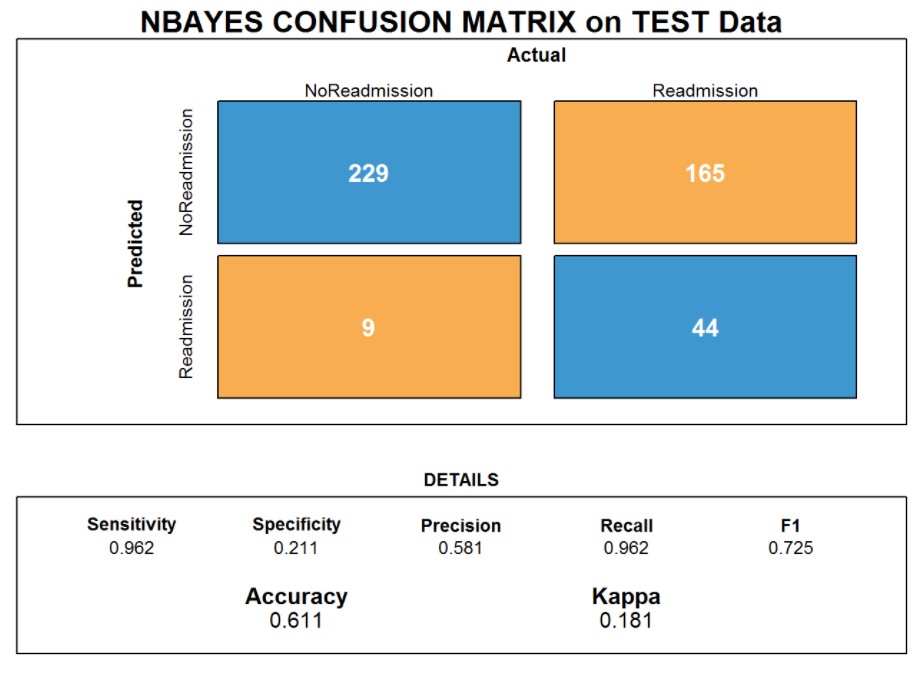
## Naïve Bayes

Naïve Bayes

### Summary



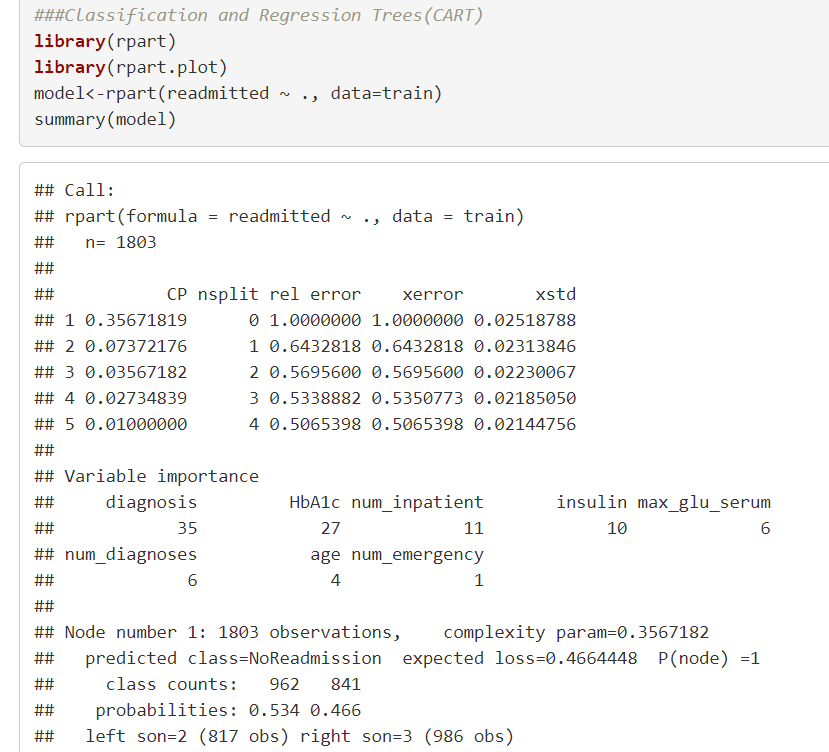
### Accuracy



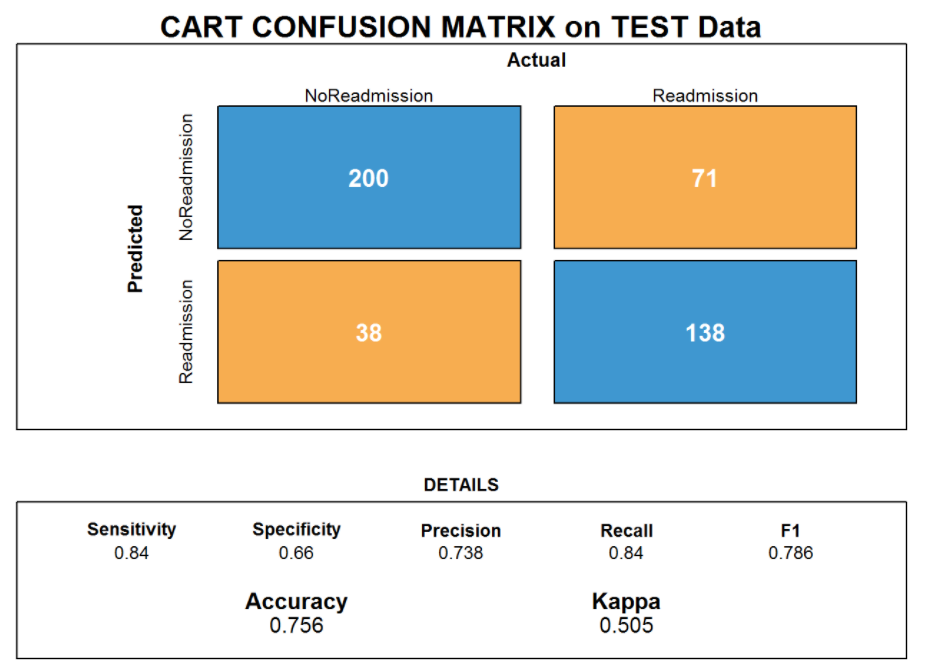
## CART

Classification and Regression Tree

### Summary



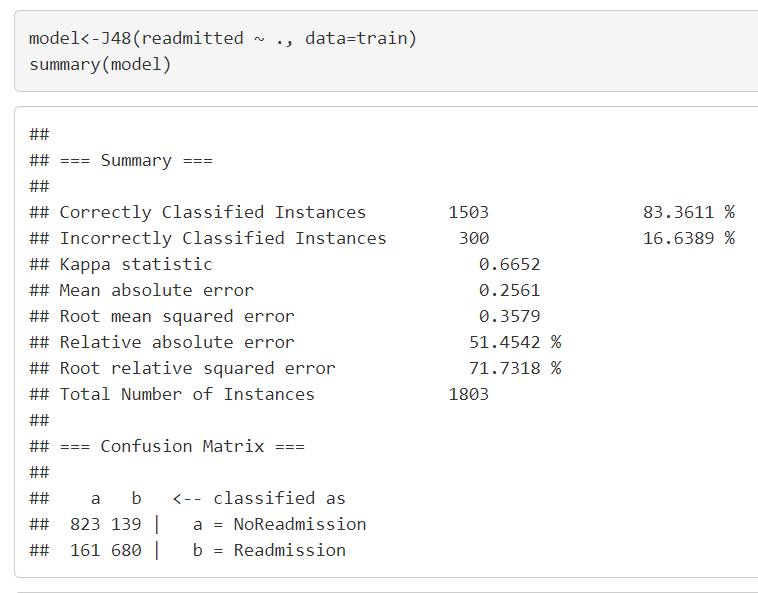
### Accuracy



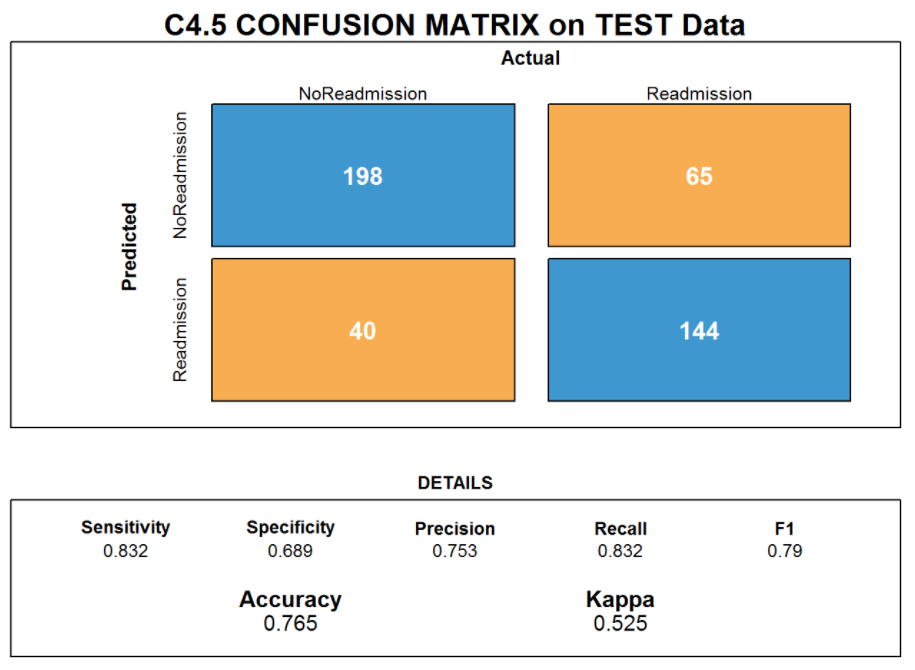
## C4.5

C4.5 algorithm

### Summary



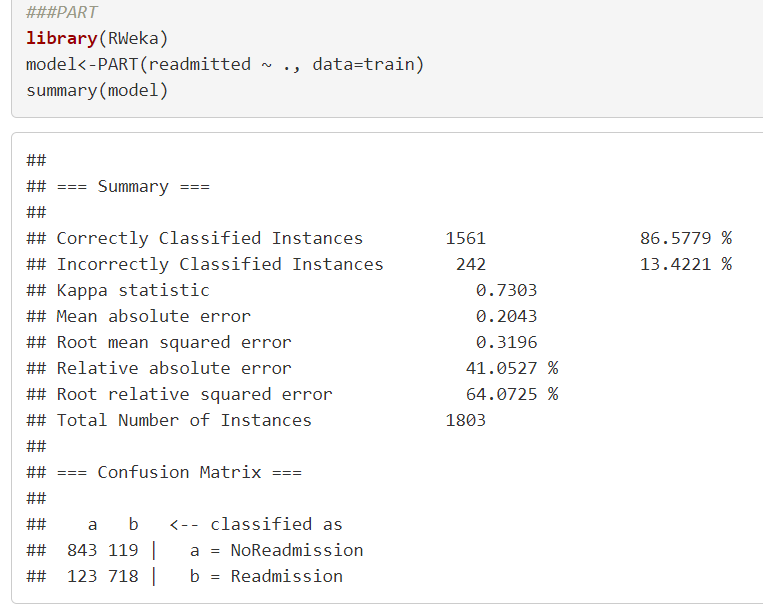
### Accuracy



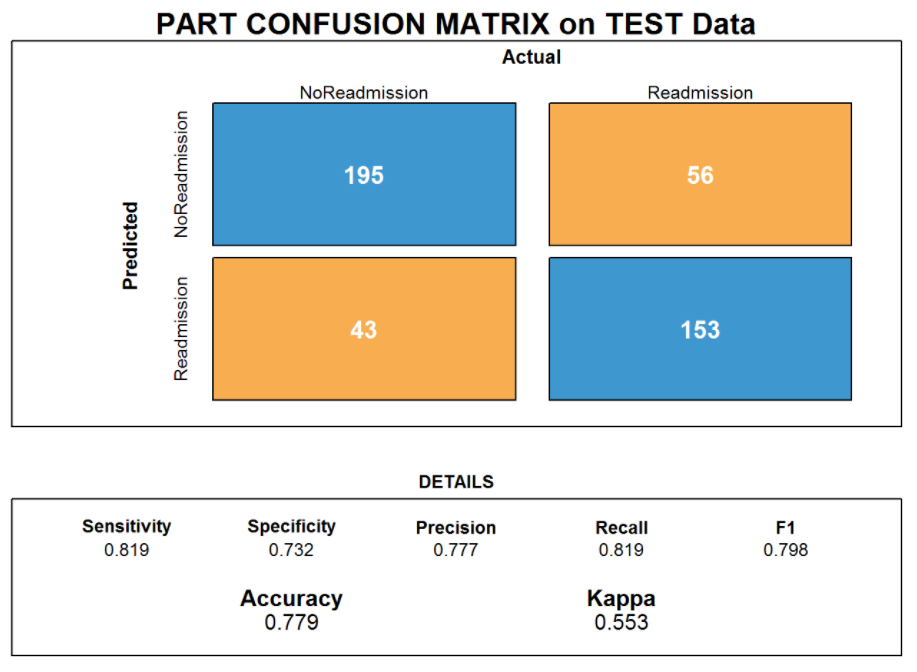
## PART

PART decision lists.

### Summary



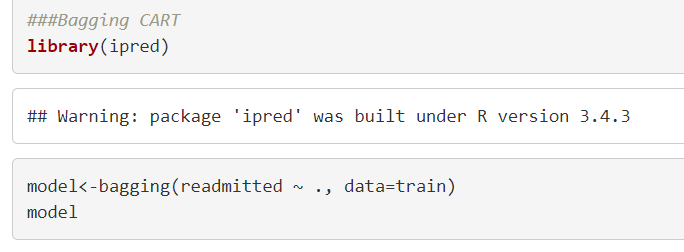
### Accuracy



## Bagging CART

Bagging CART

### Summary



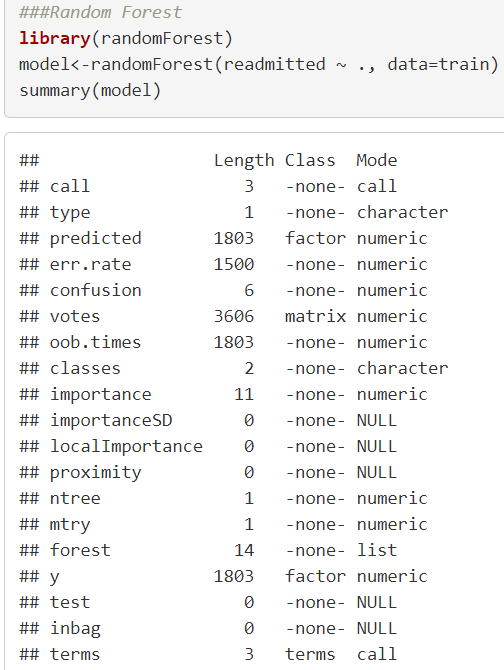
### Accuracy



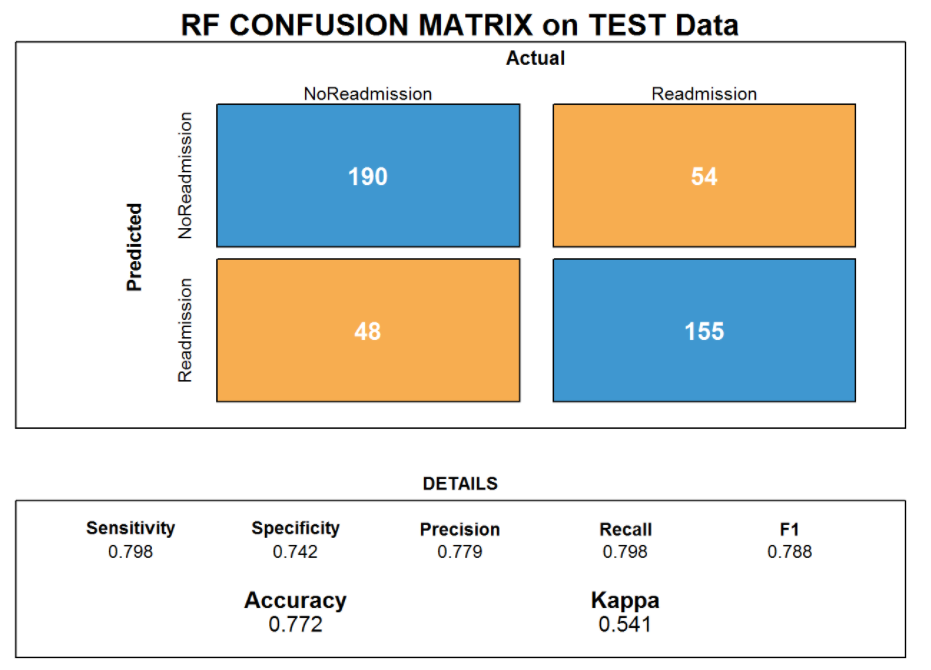
## Random Forest

Random Forest

### Summary



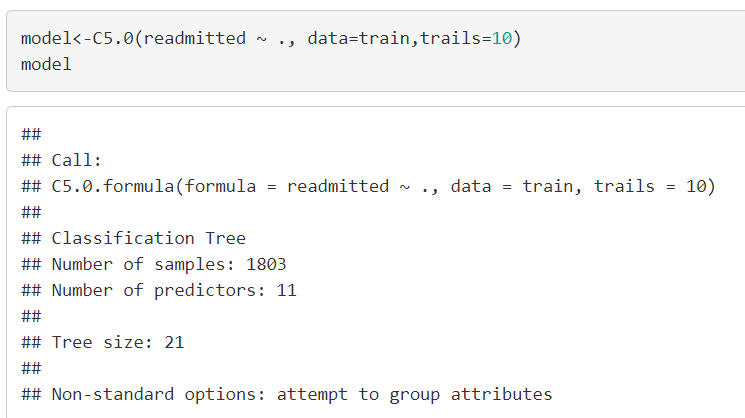
### Accuracy



## C5.0

C5.0 Algorithm

### Summary



### Accuracy

