9/20/2016

Noura Abu Khamiss

**Predicting Type 2 Diabetes**

Contents

[1. Abstract: 2](#_Toc462135932)

[2. The goal of this project: 2](#_Toc462135933)

[3. Problem statement: 2](#_Toc462135934)

[4. Background: 3](#_Toc462135935)

[4.1. What is type 2 diabetes? 3](#_Toc462135936)

[4.2. What is insulin? 3](#_Toc462135937)

[4.3. How insulin works? 3](#_Toc462135938)

[4.4. The role of glucose 3](#_Toc462135939)

[4.5. Diabetes risk factors: 4](#_Toc462135940)

[5. Data Set: 6](#_Toc462135941)

[6. Methodology: 6](#_Toc462135942)

[7. Variable Selection: 7](#_Toc462135943)

[7.1. Dependent Variable: 7](#_Toc462135944)

[7.2. Independent Variables: 7](#_Toc462135945)

[8. Exploratory Data Analysis: 10](#_Toc462135946)

[8.1. Univariate: 10](#_Toc462135947)

[8.2. Multivariate: 15](#_Toc462135948)

[9. Modeling Approach: 21](#_Toc462135949)

[10. Results: 21](#_Toc462135950)

[10.1. Logistic Regression: 21](#_Toc462135951)

[10.2. Lasso: 22](#_Toc462135952)

[11. Conclusion and Recommendation: 26](#_Toc462135953)

[12. Future Work: 26](#_Toc462135954)

# Abstract:

This project was done to predict type 2 diabetes diagnosis in a data set of 9948 patients. Logistic regression and Lasso algorithms were preformed to predict the binary outcome of 1 if the patient is diabetic and 0 if the patients is not. The two methods performed reasonably well with accuracy of 80% based on the area under the ROC curve.

# The goal of this project:

The goal of this project is classify patients who have type 2 diabetes and to develop predictive model(s) to predict the risk of developing the disease. This could help medical practitioners in assessing and recognizing the risk of developing this disease in their patients and by that they could take some preventive measures to reduce those risks.

# Problem statement:

According to World Health Organization WHO Type 2 diabetes affect around 8.5% of the adult population above 18 years of age. It is a major cause for blindness, kidney disease and lower limb amputation. WHO projects that diabetes will be the 7th leading cause of death in 2030.

Measures like exercise and healthy diet could delay or prevent the onset of the disease therefore it is important to identify individual who are at risk of developing it in order to implement an intervention plan.

# Background:

## What is type 2 diabetes?

Type 2 diabetes develops when the insulin-producing cells in the body are unable to produce enough insulin, or when the insulin that is produced does not work properly (known as insulin resistance)

## What is insulin?

Insulin is a hormone. It works as a chemical messenger that helps your body use the glucose in your blood to give you energy.

## How insulin works?

Insulin comes from the gland situated behind and below the stomach (pancreas).

* The pancreas secretes insulin into the bloodstream.
* The insulin circulates, enabling sugar to enter your cells.
* Insulin lowers the amount of sugar in your bloodstream.
* As your blood sugar level drops, so does the secretion of insulin from your pancreas.

## The role of glucose

Glucose — a sugar — is a main source of energy for the cells that make up muscles and other tissues.

* Glucose comes from two major sources: food and your liver.
* Sugar is absorbed into the bloodstream, where it enters cells with the help of insulin.
* Your liver stores and makes glucose.
* When your glucose levels are low, such as when you haven't eaten in a while, the liver breaks down stored glycogen into glucose to keep your glucose level within a normal range.

In type 2 diabetes, this process doesn't work well. Instead of moving into your cells, sugar builds up in your bloodstream. As blood sugar levels increase, the insulin-producing beta cells in the pancreas release more insulin, but eventually these cells become impaired and can't make enough insulin to meet the body's demands.

In the much less common type 1 diabetes, the immune system destroys the beta cells, leaving the body with little to no insulin.

## Diabetes risk factors:

### **Diet:**

What you eat has a profound effect on your blood glucose levels and your diabetes risk factors. A diet high in saturated fat decreases insulin sensitivity. Also [starches, sugars](http://www.drwhitaker.com/diabetic-diet-starches-and-sugars) and other [high-glycemic carbohydrates](http://www.drwhitaker.com/how-to-plan-diabetic-diet) that rapidly break down into glucose and drive up blood sugar levels significantly increases the risk of type 2 diabetes.

### **Weight:**

One of the most significant type 2 diabetes risk factors is obesity. Statistics show that 90 percent of all people with type 2 diabetes are overweight. Where you store those extra pounds is also an issue.

### Fat distribution:

If you carry them in the abdominal area, you are at an even greater risk of [insulin resistance](http://www.drwhitaker.com/insulin-resistance-pre-diabetes#insulin_resistance) and type 2 diabetes. That’s because abdominal fat is more metabolically active than fat stored in the hips or buttocks. It is more easily broken down into free fatty acids that enter the bloodstream, interfere with the action of insulin and raise triglyceride and glucose levels. Fortunately, losing weight is often all it takes to lower risk of type 2 diabetes.

### Race:

Although it's unclear why, people of certain races — including blacks, Hispanics, American Indians and Asian-Americans — are more likely to develop type 2 diabetes than whites are.

### **Activity level:**

People who lead a sedentary lifestyle are at higher risk of developing type 2 diabetes. Lack of [regular exercise](http://www.drwhitaker.com/how-to-stick-with-an-exercise-program/) lowers insulin sensitivity and increases the possibility of [metabolic syndrome](http://www.drwhitaker.com/insulin-resistance-pre-diabetes#metabolic_syndrome), another type 2 diabetes risk factor.

### **Smoking:**

Everyone knows smoking increases your risk of cancer, but many probably don’t know that people who smoke are at increased risk of developing type 2 diabetes. People who smoke are more likely to have high blood pressure, elevated cholesterol, and other lipid abnormalities, all of which are other type 2 diabetes risk factors.

### **Prescription drugs:**

Drugs that increase your risk of type 2 diabetes include corticosteroids (typically prescribed to treat asthma and arthritis); thiazide diuretics (often used to treat heart failure and high blood pressure); and beta blockers (drugs used to treat hypertension).

### Age:

The risk of type 2 diabetes increases as you get older, especially after age 45. That's probably because people tend to exercise less, lose muscle mass and gain weight as they age. But type 2 diabetes is also increasing dramatically among children, adolescents and younger adults.

### Family history:

The risk of type 2 diabetes increases if your parent or sibling has type 2 diabetes.

### Prediabetes:

Prediabetes is a condition in which your blood sugar level is higher than normal, but not high enough to be classified as diabetes. Left untreated, prediabetes often progresses to type 2 diabetes.

### Gestational diabetes:

If you developed gestational diabetes when you were pregnant, your risk of developing type 2 diabetes increases. If you gave birth to a baby weighing more than 9 pounds (4 kilograms), you're also at risk of type 2 diabetes.

### Polycystic ovarian syndrome:

For women, having polycystic ovarian syndrome — a common condition characterized by irregular menstrual periods, excess hair growth and obesity — increases the risk of diabetes.

### Other conditions:

History of hypertension, high cholesterol levels or heart disease increases the risk of developing type 2 diabetes.

# Data Set:

The data used for this project is from kaggle competition Identify patients diagnosed with Type 2 Diabetes

<https://www.kaggle.com/c/pf2012-diabetes>

The data consists of the following CSV files

|  |
| --- |
| SyncCondition |
| SyncSmokingStatus |
| training\_SyncAllergy |
| training\_SyncDiagnosis |
| training\_SyncImmunization |
| training\_SyncLabObservation |
| training\_SyncLabPanel |
| training\_SyncLabResult |
| training\_SyncMedication |
| training\_SyncPatient |
| training\_SyncPatientCondition |
| training\_SyncPatientSmokingStatus |
| training\_SyncPrescription |
| training\_SyncTranscript |
| training\_SyncTranscriptAllergy |
| training\_SyncTranscriptDiagnosis |
| training\_SyncTranscriptMedication |

# Methodology:

The solution for this problem started with identifying the factors that contribute to the diagnosis of type 2 diabetes and finding the correspond data within the data set. After that the data was transformed into features for the main table that contains information about 9948 patients so that each patient will have no more than one record in the analysis. The data is then split into training and testing sets. The split was done where the first 2/3 of the data set was assigned to the training set and the last 1/3 was assigned to the testing set.

# Variable Selection:

Based on the Type 2 diabetes risk factors the following variables were selected from the data set:

## Dependent Variable:

The dependent variable to be predicted is the DMIndicator which has the value of 1 if the patient has been diagnosed with type 2 diabetes and 0 otherwise. The variable is found in training\_SyncPatient.csv file.

## Independent Variables:

Merging the data tables together would result in multiple records per patient which would cause difficulty in the analysis process so instead the following variables will be added to the patient table as features

### Age:

Age is to be calculated by subtracting the year of birth from the training\_SyncPatient.csv file from the year 2012

### Age Grouping (AgeDesc):

The variable AgeDesc is used to group the Age variable into 5 groups as per the below table

|  |  |
| --- | --- |
| Age Range | Description |
| Between 13 and 17 | Teenager |
| Between 18 and 29 | Young Adult |
| Between 30 and 39 | Thirties |
| Between 40 and 64 | Middle Aged |
| Above 65 | Elder |

### BMI:

BMI is to be extracted from the BMI field in the training\_SyncTranscript.csv file by filtering the last visit year where the BMI is greater than 0

### BMI Groups (BMIDesc):

The variable BMIDesc is used to group the BMI variable into 4 groups as per the below table

|  |  |
| --- | --- |
| BMI Range | Description |
| Less than 18.5 | Underweight |
| Between 18.5 and 25 | Healthy Weight |
| Between 25 and 30 | Overweight |
| Greater than 30 | Obese |

### Blood Pressure(BP):

Blood pressure is to be extracted from the SystolicBP field in the training\_SyncTranscript.csv filtered by the last visit and maximum SystolicBP value

### Blood Pressure Groups (BPDesc):

The variable BPDesc is used to group the BP variable into 4 groups as per the below table

|  |  |
| --- | --- |
| BMI Range | Description |
| Less than 120 | Normal |
| Between 120 and 139 | Prehypertension |
| Between 140 and 159 | Stage I Hypertension |
| Greater than 160 | Stage II Hypertension |

### Cholesterol:

The variable will be given a binary value of 0 or 1 according to the following condition:

If the patient has and ICD9Code between 272 and 273 in the training\_SyncDiagnosis.csv table then Cholesterol = 1 otherwise Cholesterol = 0

### Hypertension:

The variable will be given a binary value of 0 or 1 according to the following condition:

If the patient has and ICD9Code between 401 and 402 in the training\_SyncDiagnosis.csv table then Hypertension = 1 otherwise Hypertension = 0

### Allergy:

This variable will be assigned a value of the number of patient records in the training\_SyncAllergy.csv table

### Medication:

This variable will be assigned a value of the number of patient records in the training\_SyncMedication.csv table

### Labs:

This variable will be assigned a value of the number of patient records in the training\_SyncLabResult.csv table

### PreviousSmoker:

The variable will be assigned a binary value 0 or 1 based on the following criteria:

In the SyncSmokingStatus.csv table

|  |  |
| --- | --- |
| Current status unknown | PreviousSmoker |
| Not a current tobacco user | 0 |
| 0 cigarettes per day (previous smoker) | 1 |
| Few (1-3) cigarettes per day | 1 |
| Current status unknown | 0 |
| 2 or more packs per day | 1 |
| Up to 1 pack per day | 1 |
| 1-2 packs per day | 1 |
| 0 cigarettes per day (non-smoker or less than 100 in lifetime) | 0 |
| Current Tobacco user | 1 |

Then the status will be assigned to each patient in the training\_SyncPatientSmokingStatus.csv then the data will be assigned to matching patients in the training\_SyncPatient.csv table

### CigarettesNum:

The value of this variable is assigned based on the following:

|  |  |
| --- | --- |
| Current status unknown | PreviousSmoker |
| Not a current tobacco user | 0 |
| 0 cigarettes per day (previous smoker) | 0 |
| Few (1-3) cigarettes per day | 3 |
| Current status unknown | 0 |
| 2 or more packs per day | 40 |
| Up to 1 pack per day | 20 |
| 1-2 packs per day | 30 |
| 0 cigarettes per day (non-smoker or less than 100 in lifetime) | 0 |
| Current Tobacco user | 10 |

Then the number of cigarettes is be assigned to each patient in the training\_SyncPatientSmokingStatus.csv then the data is assigned to matching patients in the training\_SyncPatient.csv table

After populating the Patient table with the variables all NA values have been replaced with 0

# Exploratory Data Analysis:

Exploratory Data analysis was done on the full data set including training and testing data sets. After transforming the data into one data frame the data was plotted as bar charts to see the relationship between diabetes and the different variables

## Univariate:

### Diabetes diagnosis

About 20% of the patients are diagnosed with type 2 diabetes in the data set

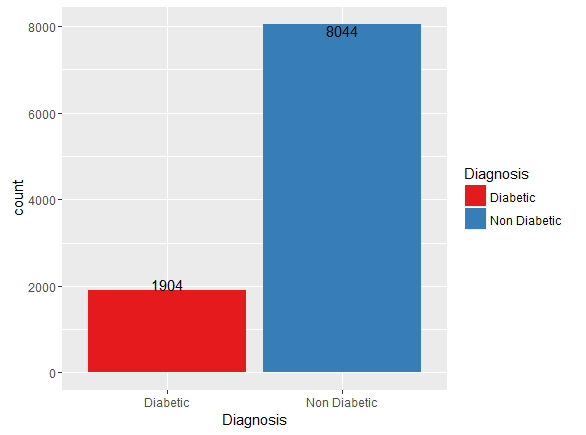


Figure 1:Diabetic vs No Diabetic Patients

### Gender

More than 50% of the patients are female

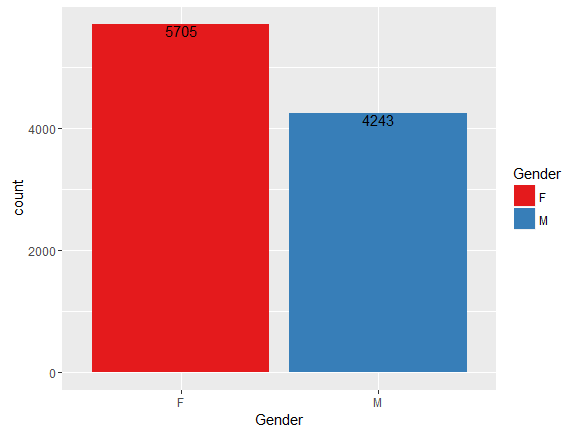


Figure 2: Female vs Male patients

### Age Group

Most of the patients are older than 40 years of age

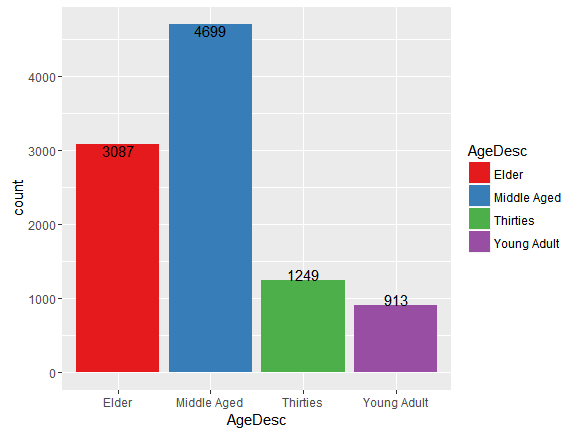


Figure 3: Patient Age Groups

### Blood Pressure

A large number of patients fall within the prehypertension group followed by stage I hypertension

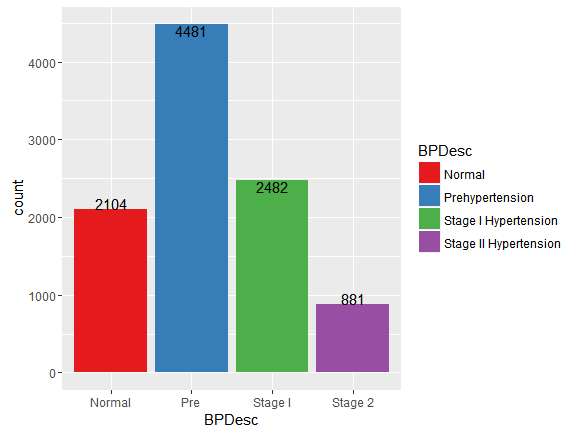


Figure 4: Blood Pressure Groups Patient Distribution

### Body Mass Index

Most of the patients in the data set are above the healthy weight

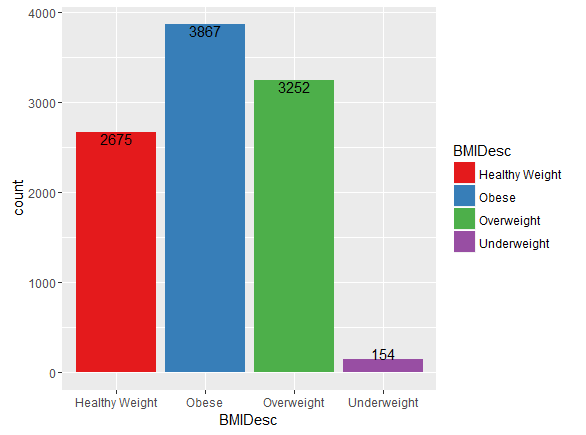


Figure 5: BMI Groups patient Distribution

### Smoking Status

Most of the patients in the data set are non smokers

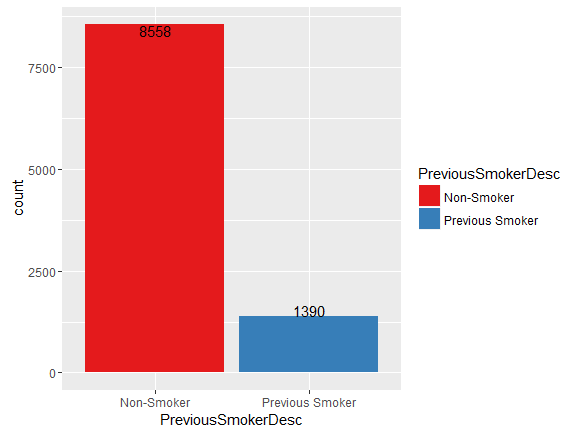


Figure 6: Patient Smoking Status

### Number of Cigarettes

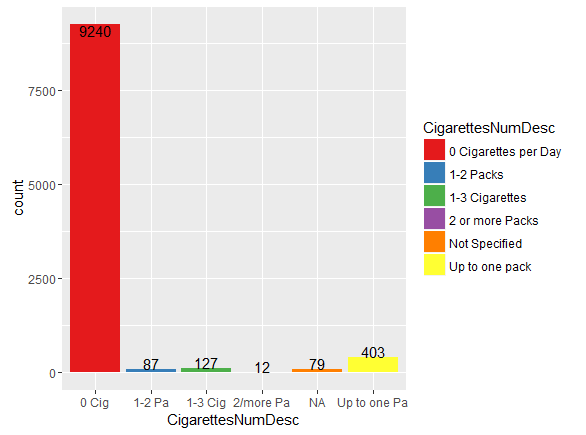


Figure 7: Number of Cigarettes per day

### Cholesterol

Most of the patients do not have high cholesterol levels

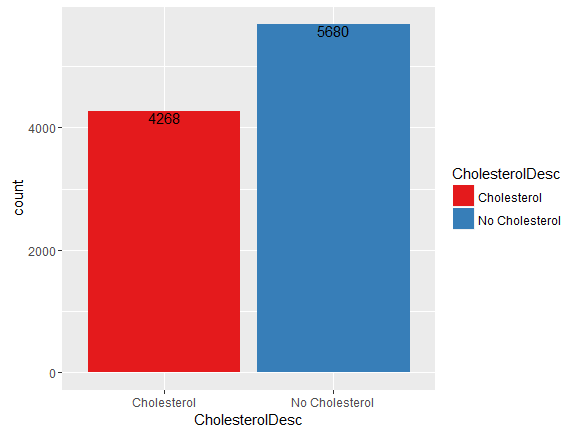


Figure 8: Cholesterol among patients

### Labs

Only a small percentage of patients have lab results

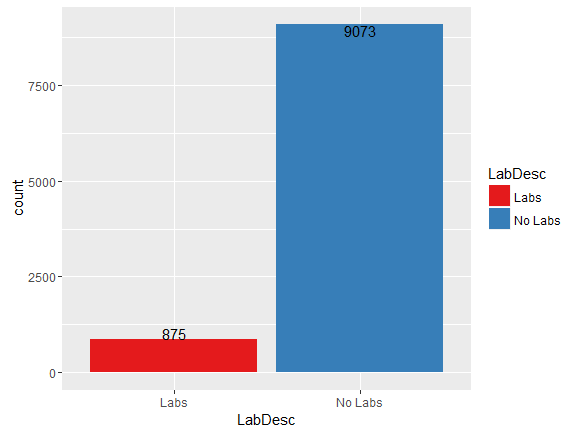


Figure 9: Labs among Patients

### Allergy

Most of the patients have no allergy

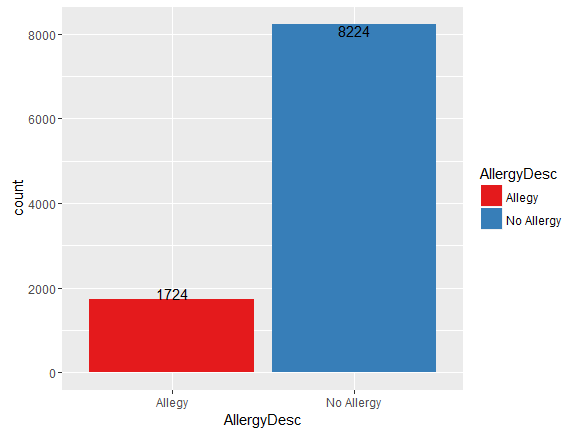


Figure 10: Allergy among Patients

### Medication

Most patients are taking prescribed medications

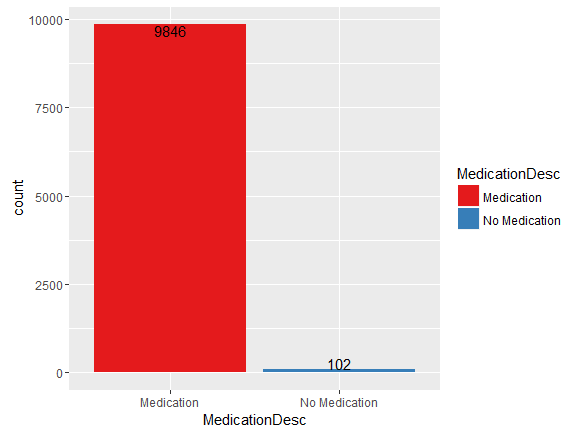


Figure 11: Medication among patient

## Multivariate:

### Diabetes diagnosis by gender

From the chart we can see that the percentage of male patients diagnosed with diabetes is higher than the female patients so we can infer that gender is an important predictor for diabetes

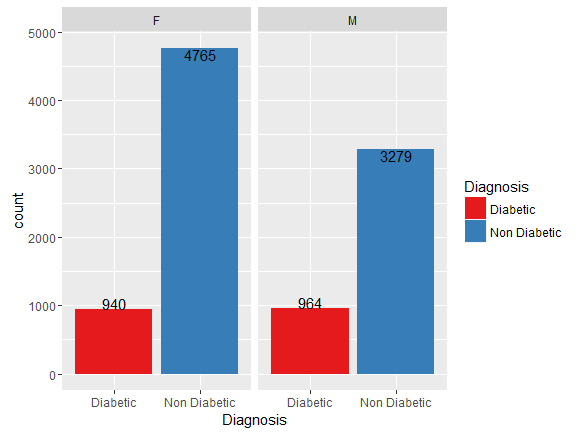


Figure 12: Diabetes Diagnosis by Gender

### Diagnosis by Age group:

From the below chart it is visible that the diagnosis of diabetes is higher among people above 40 years of age

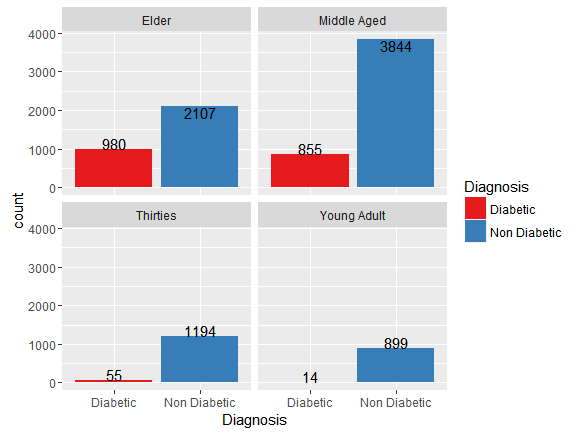


Figure 13: Diabetes Diagnosis by Age Group

### Diabetes and Blood pressure:

From the chart below we can see that the percentage of patients with diabetes diagnosis increases with the level of their blood pressure

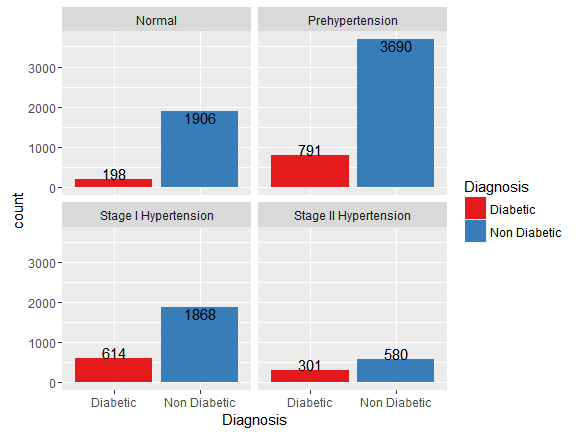


Figure 14: Diabetes Diagnosis by Blood Pressure Group

### Diabetes and BMI

From the chart below we can see that most patients with diabetes diagnosis are over the healthy weight

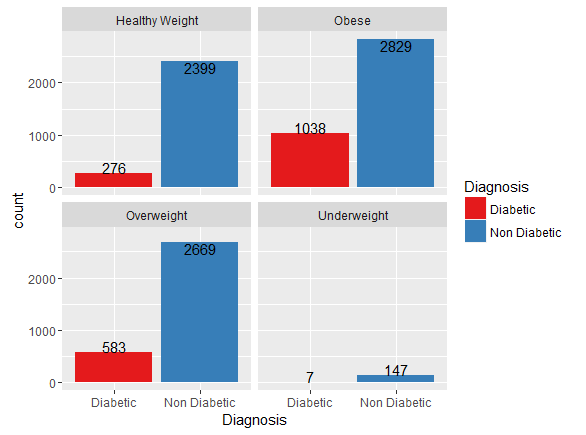


Figure 15: Diabetes Diagnosis by Weight Group

### Cholesterol

We can see that proportion of diabetic people is higher within people who have cholesterol so we can infer that it will be an important variable in our model

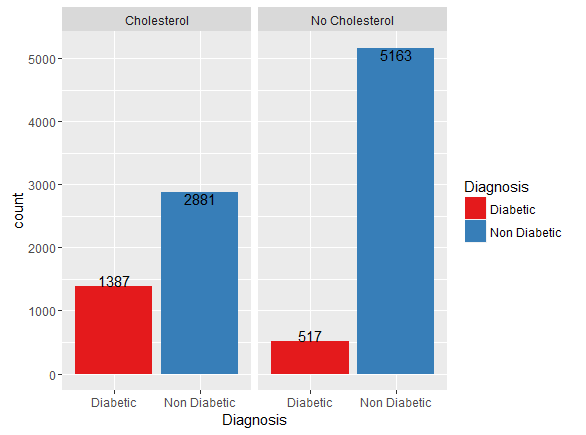


Figure 16: Diabetes Diagnosis by Cholesterol

### Allergy

We can see from the chart below that about 20% of patients with allergy are diagnosed with diabetes

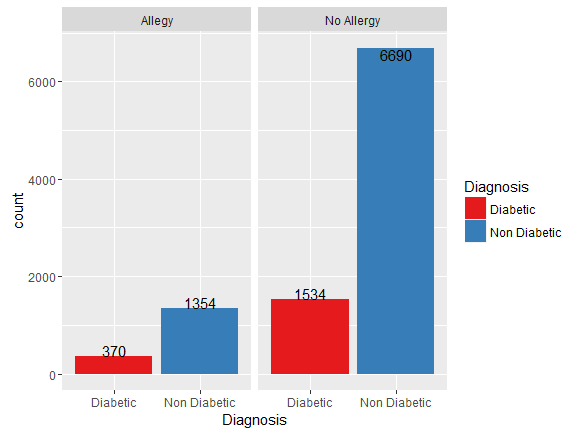


Figure 17: Diabetes Diagnosis by Allergy Status

### Medication

Most of the diabetic patients are on prescribed medications and since most of the patients are taking medications it is difficult to see a relationship between the two variables

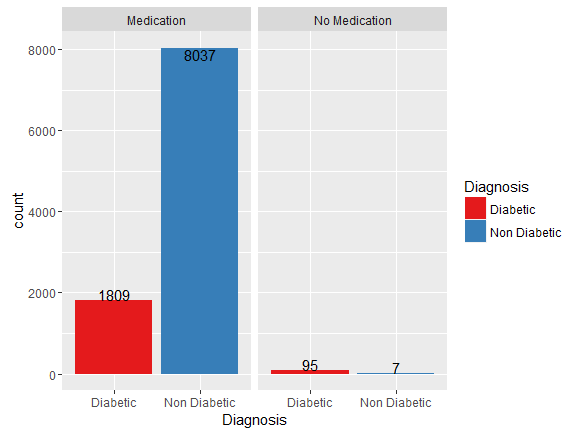


Figure 18: Diabetes Diagnosis by Medication status

### Labs

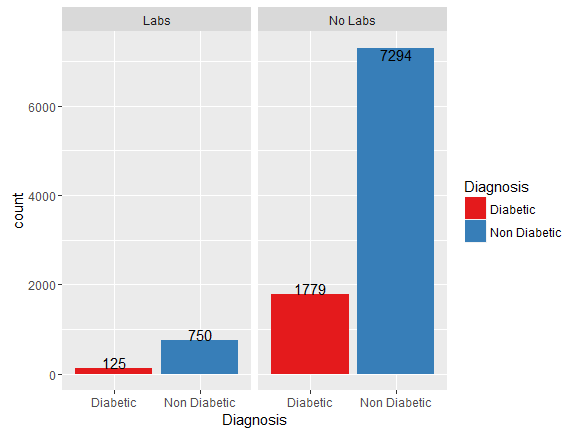


Figure 19: Diabetes Diagnosis by Lab Status

### Smoking Status

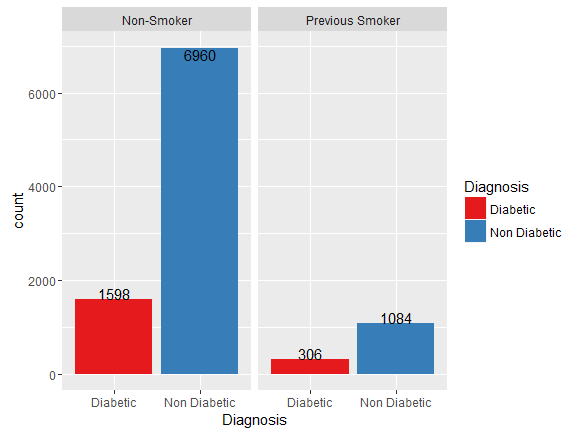


Figure 20: Diabetes Diagnosis by Smoking Status

### Number of Cigarettes

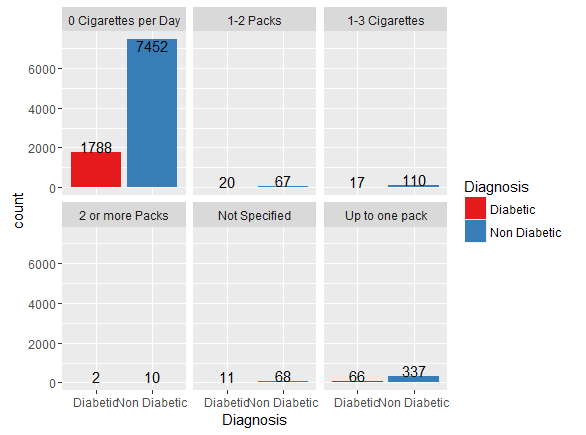


Figure 21: Diabetes Diagnosis by Number of Cigarettes per Day

# Modeling Approach:

Based on the problem definition and the binary outcome variable a logistic regression model was fitted to the data and then to better enhance the results a Lasso model was used for better feature selection.

# Results:

## Logistic Regression:

### Coefficient Table:

Below is the results of the logistic regression using all the variables. We can see that the Allergy, PreviousSmoker and the CigarettesNum are not significant variables for the model

|  |
| --- |
| Coefficients: |
| Estimate Std. Error z value Pr(>|z|) |
| (Intercept) -7.334486 0.337100 -21.758 < 2e-16 \*\*\* |
| Age 0.029763 0.002552 11.663 < 2e-16 \*\*\* |
| BMI 0.072314 0.005709 12.667 < 2e-16 \*\*\* |
| Cholesterol 1.021415 0.076067 13.428 < 2e-16 \*\*\* |
| Allergy 0.038627 0.042852 0.901 0.367370 |
| Medication 0.019835 0.007023 2.824 0.004740 \*\* |
| GenderM 0.272960 0.070402 3.877 0.000106 \*\*\* |
| Hypertension 0.773865 0.078516 9.856 < 2e-16 \*\*\* |
| BP 0.005948 0.001923 3.093 0.001979 \*\* |
| Labs -0.171605 0.050493 -3.399 0.000677 \*\*\* |
| PreviousSmoker -0.105840 0.117869 -0.898 0.369213 |
| CigarettesNum -0.001424 0.008428 -0.169 0.865852 |

After removing the two variables we have the following coefficient table

|  |
| --- |
| Coefficients: |
| Estimate Std. Error z value Pr(>|z|) |
| (Intercept) -7.351850 0.336287 -21.862 < 2e-16 \*\*\* |
| Age 0.029936 0.002534 11.814 < 2e-16 \*\*\* |
| BMI 0.072573 0.005688 12.759 < 2e-16 \*\*\* |
| Cholesterol 1.017249 0.075979 13.389 < 2e-16 \*\*\* |
| Medication 0.019540 0.006970 2.803 0.005058 \*\* |
| GenderM 0.260677 0.069649 3.743 0.000182 \*\*\* |
| Hypertension 0.770457 0.078417 9.825 < 2e-16 \*\*\* |
| BP 0.005984 0.001921 3.115 0.001842 \*\* |
| Labs -0.170515 0.050159 -3.400 0.000675 \*\*\* |

### Confusion Matrix:

|  |
| --- |
| Predicted\_Non\_Diabetic Predicted\_Diabetic |
| Actual\_Non\_Diabetic 2067 615 |
| Actual\_Diabetic 198 436 |

True Positive (TN) = 436

False Positive (FN)= 615

True Negative (TN)= 2067

False Negative (FN)= 198

**True Positive** refers to diabetic patients who got classified as diabetics.

**False Positive** refers to non diabetic patients classified as diabetic.

**True Negative** refers to non diabetic patients classified as non diabetic

**False Negative** refers to diabetic patients classified as non diabetic

Sensitivity = TP/(TP + FN) = 0.69

Specificity = TN/(TN + FP) = 0.77

Accuracy = (TN + TP)/(TN + TP + FN + FP) = 0.75

### ROC Curve

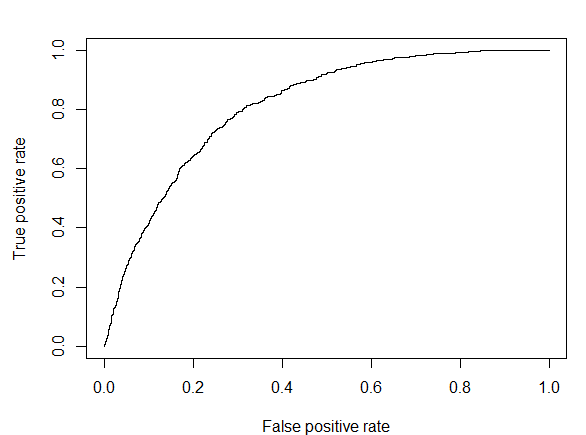


Figure 22: Logistic Regression ROC Curve

### Area under the curve

AUC = 0.81

## Lasso:

### Coefficients

We can see from the coefficients table that having cholesterol and hypertension are of the strong variables related to being diagnosed with type 2 diabetes

|  |
| --- |
| coef |
| (Intercept) -2.884461104 |
| patientTrain.Age 0.008847891 |
| patientTrain.BMI 0.014254451 |
| patientTrain.BP 0.000000000 |
| patientTrain.Medication 0.000000000 |
| patientTrain.Labs 0.000000000 |
| patientTrain.GenderM 0.000000000 |
| patientTrain.Cholesterol1 0.558880189 |
| patientTrain.Hypertension1 0.519539110 |
| patientTrain.PreviousSmoker1 0.000000000 |
| patientTrain.CigarettesNum3 0.000000000 |
| patientTrain.CigarettesNum10 0.000000000 |
| patientTrain.CigarettesNum20 0.000000000 |
| patientTrain.CigarettesNum30 0.000000000 |
| patientTrain.CigarettesNum40 0.000000000 |

From the chart below we can see that a lot of our variables are approaching zero which signifies high correlation levels among independent variables

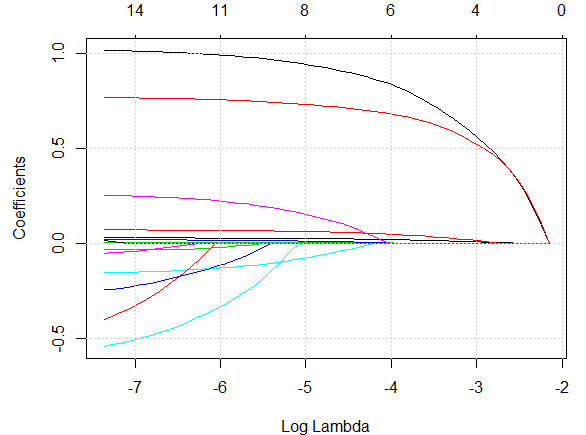


Figure 23: LASSO Coefficients

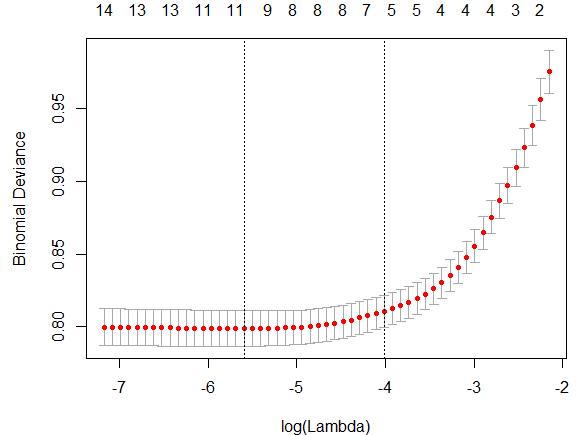


Figure 24:LASSO Best Lambda

### Confusion matrix

|  |  |
| --- | --- |
| |  | | --- | |  |   Predicted\_Non\_Diabetic Predicted\_Diabetic |
| Actual\_Non\_Diabetic 2073 609 |
| Actual\_Diabetic 196 438 |

True Positive = 438

False Positive = 609

True Negative = 2073

False Negative = 196

Sensitivity = TP/(TP + FN) = 0.69

Specificity = TN/(TN + FP) = 0.77

Accuracy = (TN + TP)/(TN + TP + FN + FP) = 0.76

### ROC Curve

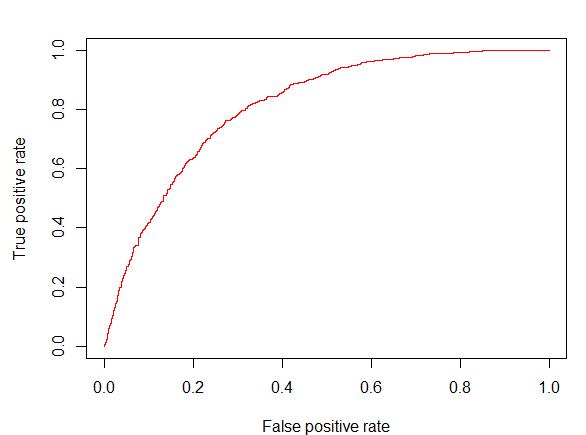


Figure 25: LASSO ROC Curve

### Area under the curve

AUC = 0.8115371

The two algorithms chosen performed at similar accuracy level of 80%. The classification of patients diagnosed with type 2 diabetes appears to be related to features other than the variables selected for this study. Overall the algorithms performed reasonably well

# Conclusion and Recommendation:

From the above results we can see than some factors increase the risk of patients getting type 2 diabetes. These factors include age of over 40 years, overweight, hypertension and cholesterol. With proper health management and dietary choices the risk could be reduced so following is our recommendations:

1. For physicians: be aware of the patient overall health and risk of getting the disease and take preventive measures
2. For the patients if you notice any of the risk factors like overweight manage it with proper exercise regimen and healthy diet
3. And finally for hospitals and clinics we recommend installing proper analytics to give insights to the physicians about risk scores for different diseases for their patients so they can take proper precautions

# Future Work:

Due to time constraint we could not cover the following areas and we recommend them for future work:

1. Assigning a risk score to patients based on the probability of developing type 2 diabetes
2. Exploring options of more independent variables to enhance the predictive ability of the models
3. A system for preventive measures recommendations personalized for each patient based on their risk factors.