ST 541 Final Project: Analysis of Diabetes Risk Factors

Benjamin A. Ajibade and Jacob E. Waggoner

2022-11-17

Introduction

Diabetes is a very prevelent disease that affects the human body's core functions of getting energy from food. In the United States, more than 37 million adults have diabetes, and it is the seventh-leading cause of death. Concerningly, 1 of 5 adults with diabetes are not aware they have the disease. A lack of knowledge and action for diabetes patients leads to an increased liklihood of extreme complications such as hear disease, kidney disease, vision loss, and death.
As a result of the severity of this disease, it is important that individuals are screened and informed of their risk level. This is important because diabetes can be present in a person without any obvious symptoms for a long period of time, and early detection allows for better outcomes. Understanding of the risk levels for those without diabetes is also important, because preventative measures can be taken before the disease is contracted.
The objective of this study is to determine which clinical measurements and records make individuals more or less susceptible to diabetes. Our key research question is as follows:

> How can clinical datapoints be used to determine an individual's risk of having diabetes?

The methodology we used to accomplish this goal was based around performing a logistic regression to attain a ratio, between 0 and 1, that a person would be diagnosed with diabetes (outcome 1). Before constructing and assessing this model, we conducted data visualizations and statistical tests that would help us understand the data and what outcomes we should expect from our model. The data used to perform this analysis and modeling comes from the United States' National Institute of Diabetes and Digestive and Kidney Diseases and was downloaded from Kaggle.

Data Exploration

Variables

The dataset contains 8 numeric predictors and one binary outcome. This aligns with the requrients for multiple logistic regression. Descriptions of each variable are:

- **Pregnancies**: Number of times pregnant
- **Glucose**: Plasma glucose concentration from an oral glucose tolerance test
- **Blood-Pressure**: Diastolic blood pressure (mm Hg)
- **Skin-Thickness**: Triceps skin fold thickness (mm)
- **Insulin**: 2-Hour serum insulin level (mu U/ml)

- **BMI**: Body mass index (weight in kg/ (height in m) ^2)
- **DiabetesPedigreeFunction**: Likelihood of diabetes based on family history
- **Age**: Age (in years) of the individual in observation
- **Outcome**: The target variable, a binary variable of 0 or 1, in which 1 is interpreted as "tested positive for diabetes"

  The first step in our data exploration was to assess the data quality and structure. We viewed he first few rows of the dataset to visually assess the data. We can see that out of our eight numeric variables, only two, BMI and the pedigree function, are continuous, and the rest are discrete.

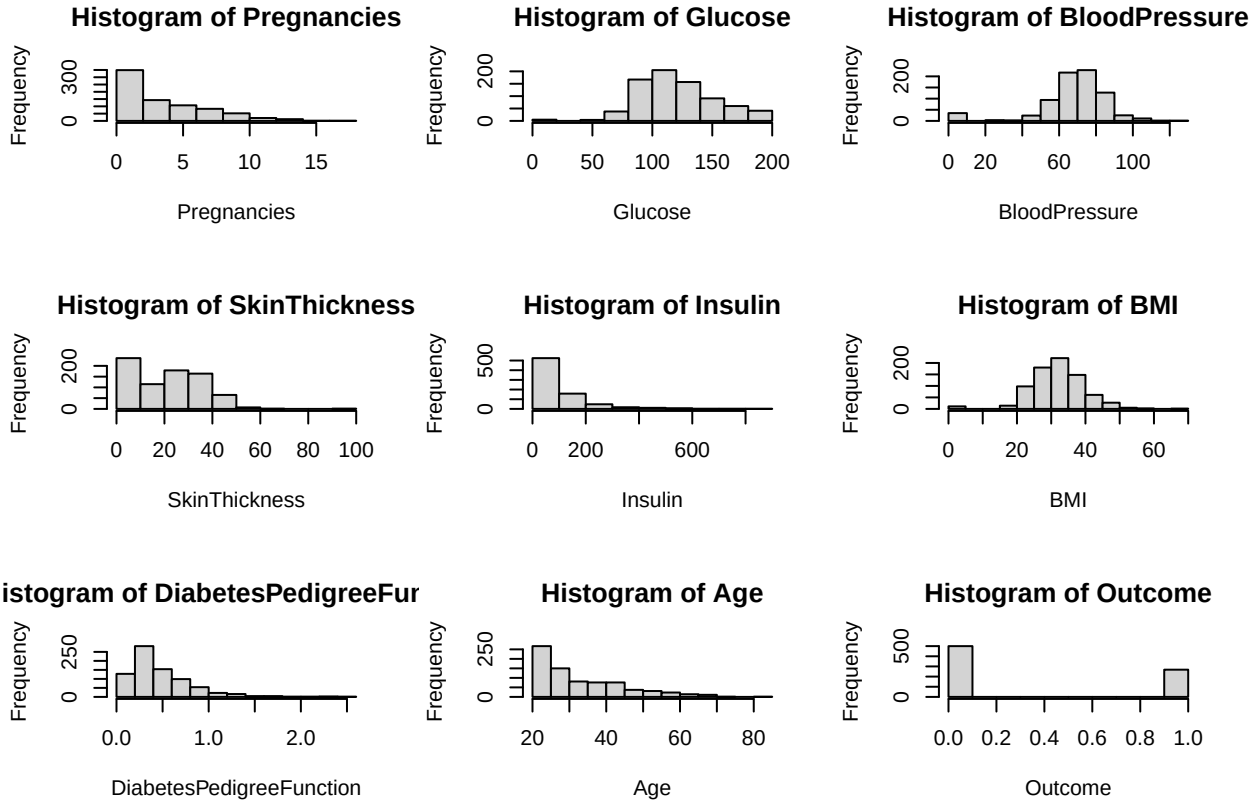
```
##   Pregnancies Glucose BloodPressure SkinThickness Insulin  BMI
## 1           6     148            72            35       0 33.6
## 2           1      85            66            29       0 26.6
## 3           8     183            64             0       0 23.3
## 4           1      89            66            23      94 28.1
## 5           0     137            40            35     168 43.1
## 6           5     116            74             0       0 25.6
##   DiabetesPedigreeFunction Age Outcome
## 1                    0.627  50       1
## 2                    0.351  31       0
## 3                    0.672  32       1
## 4                    0.167  21       0
## 5                    2.288  33       1
## 6                    0.201  30       0
```

We also generated summary statistics of the data, as shown below. In addition to explaining the distributions and central tendancies of each variable, the output also demonstrated that the dataset contained no null values.

```
##   Pregnancies        Glucose      BloodPressure    SkinThickness
##  Min.   : 0.000   Min.   :  0.0   Min.   :  0.00   Min.   : 0.00
##  1st Qu.: 1.000   1st Qu.: 99.0   1st Qu.: 62.00   1st Qu.: 0.00
##  Median : 3.000   Median :117.0   Median : 72.00   Median :23.00
##  Mean   : 3.845   Mean   :120.9   Mean   : 69.11   Mean   :20.54
##  3rd Qu.: 6.000   3rd Qu.:140.2   3rd Qu.: 80.00   3rd Qu.:32.00
##  Max.   :17.000   Max.   :199.0   Max.   :122.00   Max.   :99.00
##     Insulin           BMI        DiabetesPedigreeFunction      Age
##  Min.   :  0.0   Min.   : 0.00   Min.   :0.0780           Min.   :21.00
##  1st Qu.:  0.0   1st Qu.:27.30   1st Qu.:0.2437           1st Qu.:24.00
##  Median : 30.5   Median :32.00   Median :0.3725           Median :29.00
##  Mean   : 79.8   Mean   :31.99   Mean   :0.4719           Mean   :33.24
##  3rd Qu.:127.2   3rd Qu.:36.60   3rd Qu.:0.6262           3rd Qu.:41.00
##  Max.   :846.0   Max.   :67.10   Max.   :2.4200           Max.   :81.00
##     Outcome
##  Min.   :0.000
##  1st Qu.:0.000
##  Median :0.000
##  Mean   :0.349
##  3rd Qu.:1.000
##  Max.   :1.000
```

Variable Distributions

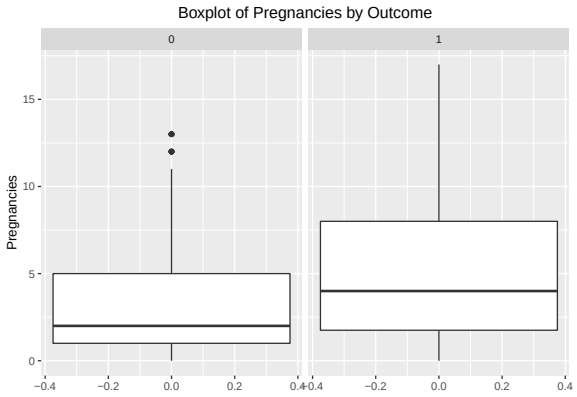
Our next step in data exploration was to assess the distributions of all our variables. We generated the following histograms for this assessment.

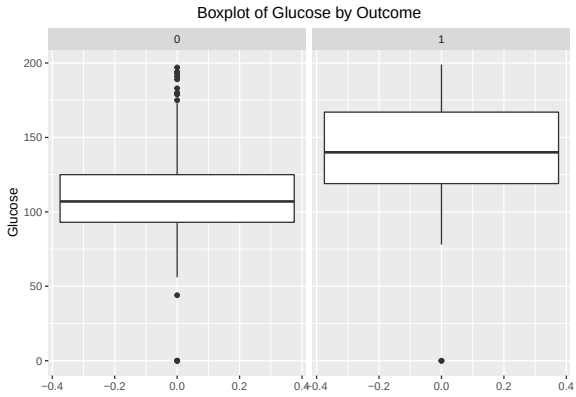


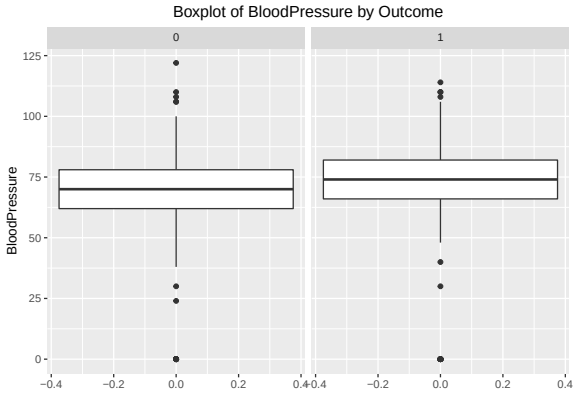
From these visualizations, we determined that Pregnancies, SkinThickness, Insulin, DiabetesPedigree, and Age were right-skewed and that Glucose, Blood Pressure, and BMI were approximately normal. We observed an irregularity inthe glucose distribution in which there were a few values at 0, far from the normal range of that distribution. It would seem counterintuitive that a human could survive with such a low glucose level, and these may be an indicator of a missing measurement. We also observed in the "Pregnancies" histogram that the mode and a large portion of the observations had zero pregnancies. This suggests that the sample, of which we do not have extensive information about, does include both men and women.

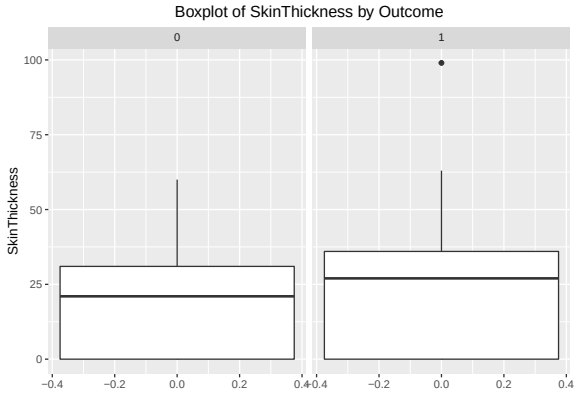
Interactions with Target

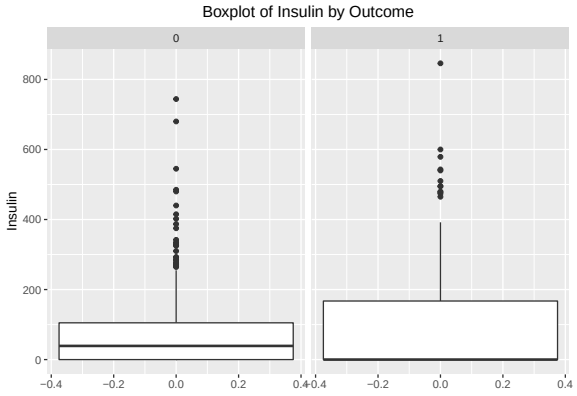
Following our assessment of the variables within the entire sample, we then visualized the variables by target class to see which variables differed greatly in those who weren't diagnosed with diabetes, and those who were. This gave us an initial idea of which variables would provide value as a predictor, based on which have less overlap between the central areas of the boxplot between the two outcomes.

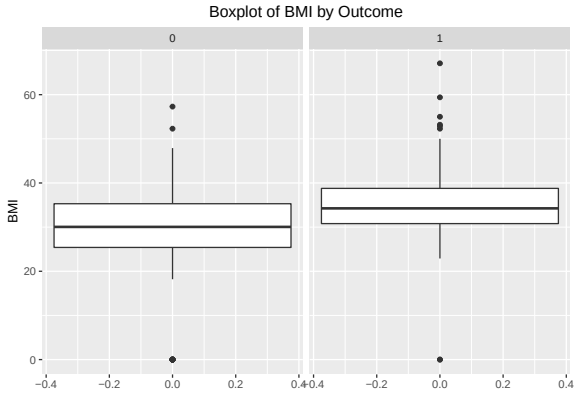

Boxplot of Pregnancies by Outcome
0
1
Pregnancies
15
10
5
0
−0.4
−0.2
0.0
0.2
0.4

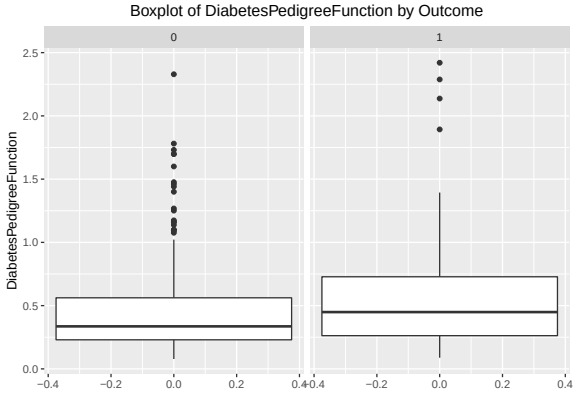

Boxplot of Glucose by Outcome
0
1
Glucose
200
150
100
50
0
−0.4
−0.2
0.0
0.2
0.4

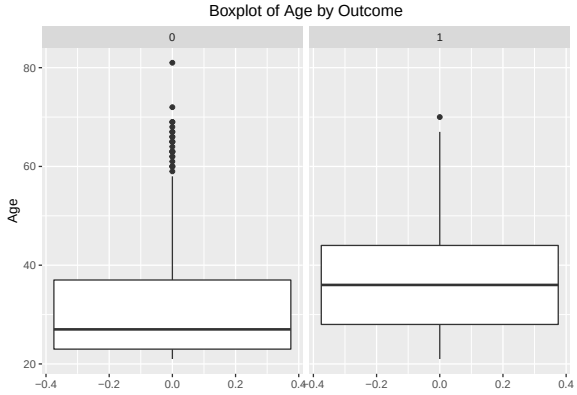

Boxplot of BloodPressure by Outcome
0
1
BloodPressure
125
100
75
50
25
0
−0.4
−0.2
0.0
0.2
0.4


Boxplot of SkinThickness by Outcome
0
1
SkinThickness
100
75
50
25
0
−0.4
−0.2
0.0
0.2
0.4


Boxplot of Insulin by Outcome
0
1
Insulin
800
600
400
200
0
−0.4
−0.2
0.0
0.2
0.4


Boxplot of BMI by Outcome
0
1
BMI
60
40
20
0
−0.4
−0.2
0.0
0.2
0.4


Boxplot of DiabetesPedigreeFunction by Outcome
0
1
DiabetesPedigreeFunction
2.5
2.0
1.5
1.0
0.5
0.0
−0.4
−0.2
0.0
0.2
0.4


Boxplot of Age by Outcome
0
1
Age
80
60
40
20
−0.4
−0.2
0.0
0.2
0.4

Statistical Tests

Differences Between Outcome Groups

Building upon our previous exploratory analysis, we wanted to assess which predictors had statistically significant differences in the means between the two outcome classes. This is an objective approach to glean similar information from the class boxplots in the previous section and will give us another indication of which predictors will be useful. To test this, we utilized Welch Two-Sample T-Tests for each predictor. The results are shown below.

```
##
##  Welch Two Sample t-test
##
## data:  df$Pregnancies by df$Outcome
## t = -5.907, df = 455.96, p-value = 6.822e-09
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
## 95 percent confidence interval:
##  -2.089219 -1.046125
## sample estimates:
## mean in group 0 mean in group 1
##        3.298000        4.865672
##
##
##  Welch Two Sample t-test
##
## data:  df$Glucose by df$Outcome
## t = -13.752, df = 461.33, p-value < 2.2e-16
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
## 95 percent confidence interval:
##  -35.74707 -26.80786
## sample estimates:
## mean in group 0 mean in group 1
##        109.9800        141.2575
##
##
##  Welch Two Sample t-test
##
## data:  df$BloodPressure by df$Outcome
## t = -1.7131, df = 471.31, p-value = 0.08735
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
## 95 percent confidence interval:
##  -5.669580  0.388326
## sample estimates:
## mean in group 0 mean in group 1
##        68.18400        70.82463
##
##
##  Welch Two Sample t-test
##
## data:  df$SkinThickness by df$Outcome
## t = -1.9706, df = 472.1, p-value = 0.04936
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
## 95 percent confidence interval:
```

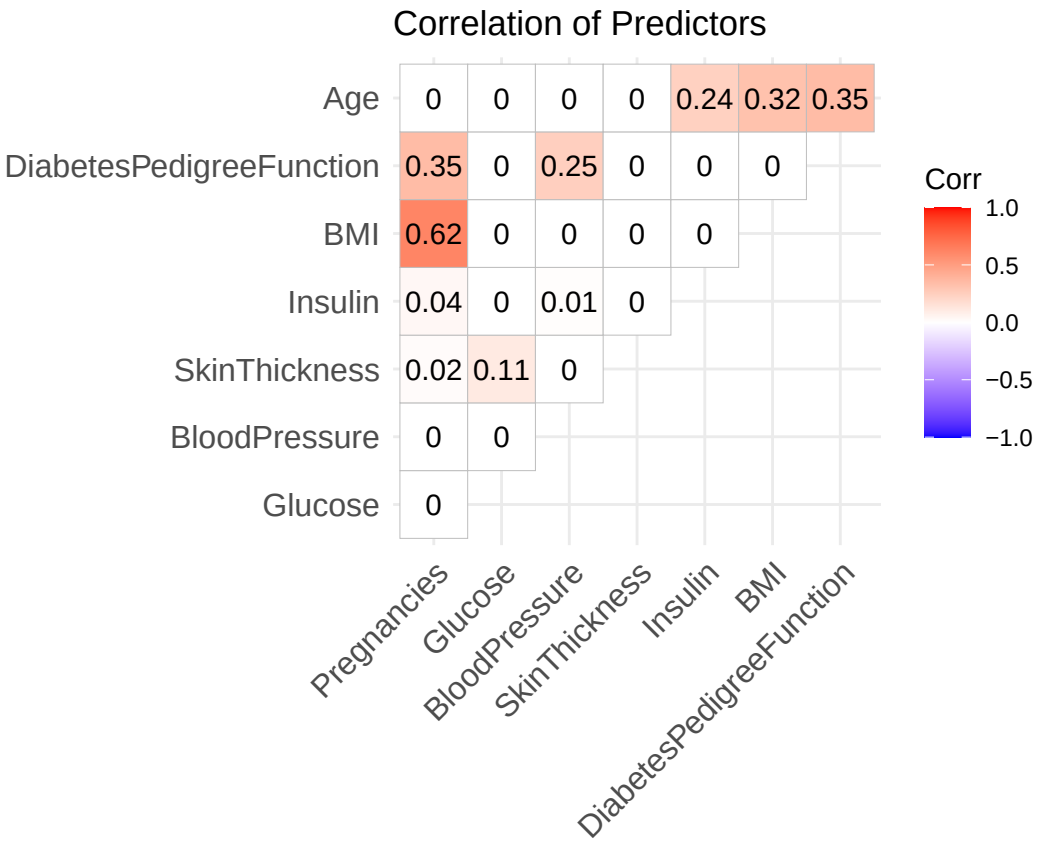
```
##  -4.993281565 -0.007076644
## sample estimates:
## mean in group 0 mean in group 1
##        19.66400        22.16418
##
##
##  Welch Two Sample t-test
##
## data:  df$Insulin by df$Outcome
## t = -3.3009, df = 415.75, p-value = 0.001047
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
## 95 percent confidence interval:
##  -50.32820 -12.75944
## sample estimates:
## mean in group 0 mean in group 1
##         68.7920        100.3358
##
##
##  Welch Two Sample t-test
##
## data:  df$BMI by df$Outcome
## t = -8.6193, df = 573.47, p-value < 2.2e-16
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
## 95 percent confidence interval:
##  -5.940864 -3.735811
## sample estimates:
## mean in group 0 mean in group 1
##        30.30420        35.14254
##
##
##  Welch Two Sample t-test
##
## data:  df$DiabetesPedigreeFunction by df$Outcome
## t = -4.5768, df = 454.51, p-value = 6.1e-06
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
## 95 percent confidence interval:
##  -0.17262065 -0.06891135
## sample estimates:
## mean in group 0 mean in group 1
##        0.429734        0.550500
##
##
##  Welch Two Sample t-test
##
## data:  df$Age by df$Outcome
## t = -6.9207, df = 575.78, p-value = 1.202e-11
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
## 95 percent confidence interval:
##  -7.545092 -4.209236
## sample estimates:
## mean in group 0 mean in group 1
##        31.19000        37.06716
```

Six of the eight variables resulted in p-values below the .05 alpha level in our T-tests. This indicated that

the null hypothesis was rejected, and there is a statistically significant difference between the means for a positive diabetes result (1) and a negative diabetes result (0). For Blood Pressure and Skin Thickness, the p-value was higher than .05 and the null hypothesis was not rejected, indicating that there is no difference between the means for each outcome group. As a result, these two predictors will probably be less valuable for predicting the diabetes outcome than the six with a statistically significant difference in means.

Correlation Between Predictors

We were also interested in whether any of the predictors had cross-correlations that may be of note when building our model. To do this, we calculated correlation values between all predictors. We also plotted these to provide a visual representation. We can see that most of our predictors have a low or non-existant correlation coeffecient; however, the correlation between BMI and Pregnancies stands out as it is the highest at .62.



Statistical Modeling

After concluding our exploratory analysis, we moved on to developing a statistical model to explain how each clinical predictor related to the diabetes outcome, and to calculate risk scores for our sample.

Model Selection

Full Model

We started by plugging in all predictors into the model

```
##
## Call:
## glm(formula = Outcome ~ ., family = binomial(), data = df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.5566  -0.7274  -0.4159   0.7267   2.9297
##
## Coefficients:
##                            Estimate Std. Error z value Pr(>|z|)
## (Intercept)              -8.4046964  0.7166359 -11.728  < 2e-16 ***
## Pregnancies               0.1231823  0.0320776   3.840 0.000123 ***
## Glucose                   0.0351637  0.0037087   9.481  < 2e-16 ***
## BloodPressure            -0.0132955  0.0052336  -2.540 0.011072 *
## SkinThickness             0.0006190  0.0068994   0.090 0.928515
## Insulin                  -0.0011917  0.0009012  -1.322 0.186065
## BMI                       0.0897010  0.0150876   5.945 2.76e-09 ***
## DiabetesPedigreeFunction  0.9451797  0.2991475   3.160 0.001580 **
## Age                       0.0148690  0.0093348   1.593 0.111192
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 993.48  on 767  degrees of freedom
## Residual deviance: 723.45  on 759  degrees of freedom
## AIC: 741.45
##
## Number of Fisher Scoring iterations: 5
```

Despite being significant the DiabetesPedigreeFunction has a high standard error. A quick look at the variance inflation factor for the predictor shows that there is no redundant predictor.

```
## Loading required package: carData
```

```
##              Pregnancies                  Glucose            BloodPressure
##                 1.408434                 1.214367                 1.175283
##            SkinThickness                  Insulin                      BMI
##                 1.522040                 1.467918                 1.220416
## DiabetesPedigreeFunction                      Age
##                 1.034318                 1.502069
```

Backward Selection, BIC Criterion

Backward elimination starts with all potential predictor variables in the regression model.This amounts to deleting the predictor with the largest p -value each time. BIC has a strict penalty and we expect a reduced number of predictors with good significance at the end of the process. Final model: Outcome ~ Pregnancies + Glucose + BMI + DiabetesPedigreeFunction

```
## Start:  AIC=783.24
## Outcome ~ Pregnancies + Glucose + BloodPressure + SkinThickness +
##     Insulin + BMI + DiabetesPedigreeFunction + Age
```

```
##
##                            Df Deviance    AIC
## - SkinThickness             1   723.45 776.60
## - Insulin                   1   725.19 778.34
## - Age                       1   725.97 779.12
## - BloodPressure             1   729.99 783.14
## <none>                          723.45 783.24
## - DiabetesPedigreeFunction  1   733.78 786.94
## - Pregnancies               1   738.68 791.83
## - BMI                       1   764.22 817.38
## - Glucose                   1   838.37 891.52
##
## Step:  AIC=776.6
## Outcome ~ Pregnancies + Glucose + BloodPressure + Insulin + BMI +
##     DiabetesPedigreeFunction + Age
##
##                            Df Deviance    AIC
## - Insulin                   1   725.46 771.97
## - Age                       1   725.97 772.48
## <none>                          723.45 776.60
## - BloodPressure             1   730.13 776.64
## - DiabetesPedigreeFunction  1   733.92 780.42
## - Pregnancies               1   738.69 785.20
## - BMI                       1   768.77 815.27
## - Glucose                   1   840.87 887.38
##
## Step:  AIC=771.97
## Outcome ~ Pregnancies + Glucose + BloodPressure + BMI + DiabetesPedigreeFunction +
##     Age
##
##                            Df Deviance    AIC
## - Age                       1   728.56 768.42
## <none>                          725.46 771.97
## - BloodPressure             1   732.51 772.37
## - DiabetesPedigreeFunction  1   734.99 774.85
## - Pregnancies               1   741.27 781.13
## - BMI                       1   769.24 809.10
## - Glucose                   1   845.76 885.62
##
## Step:  AIC=768.42
## Outcome ~ Pregnancies + Glucose + BloodPressure + BMI + DiabetesPedigreeFunction
##
##                            Df Deviance    AIC
## - BloodPressure             1   734.31 767.52
## <none>                          728.56 768.42
## - DiabetesPedigreeFunction  1   738.43 771.65
## - Pregnancies               1   760.56 793.78
## - BMI                       1   770.21 803.43
## - Glucose                   1   862.96 896.18
##
## Step:  AIC=767.52
## Outcome ~ Pregnancies + Glucose + BMI + DiabetesPedigreeFunction
##
##                            Df Deviance    AIC
```

```
## <none>                           734.31 767.52
## - DiabetesPedigreeFunction  1   744.12 770.70
## - Pregnancies               1   762.87 789.45
## - BMI                       1   771.27 797.85
## - Glucose                   1   864.84 891.41
```

```
##              (Intercept)              Pregnancies                  Glucose
##              -8.41585098               0.14192631               0.03382636
##                      BMI DiabetesPedigreeFunction
##               0.07809694               0.90129355
```

Stepwise Selection, BIC Criterion

Using BIC criteria Stepwise selects 4 variables: Pregnancies, Glucose, BMI and DiabetesPedigreeFunction. Procedure starts with no potential predictor variables in the regression equation. Then, it adds the predictor with the smallest p-value. Next, it adds second predictor with smallest p-value while checking if we can drop any previously added variable. This process is continued until adding an additional predictor does not yield P-value below requirement. Final model: Outcome ~ Pregnancies + Glucose + BMI + DiabetesPedigreeFunction

```
## Start:  AIC=783.24
## Outcome ~ Pregnancies + Glucose + BloodPressure + SkinThickness +
##     Insulin + BMI + DiabetesPedigreeFunction + Age
##
##                            Df Deviance    AIC
## - SkinThickness             1   723.45 776.60
## - Insulin                   1   725.19 778.34
## - Age                       1   725.97 779.12
## - BloodPressure             1   729.99 783.14
## <none>                          723.45 783.24
## - DiabetesPedigreeFunction  1   733.78 786.94
## - Pregnancies               1   738.68 791.83
## - BMI                       1   764.22 817.38
## - Glucose                   1   838.37 891.52
##
## Step:  AIC=776.6
## Outcome ~ Pregnancies + Glucose + BloodPressure + Insulin + BMI +
##     DiabetesPedigreeFunction + Age
##
##                            Df Deviance    AIC
## - Insulin                   1   725.46 771.97
## - Age                       1   725.97 772.48
## <none>                          723.45 776.60
## - BloodPressure             1   730.13 776.64
## - DiabetesPedigreeFunction  1   733.92 780.42
## + SkinThickness             1   723.45 783.24
## - Pregnancies               1   738.69 785.20
## - BMI                       1   768.77 815.27
## - Glucose                   1   840.87 887.38
##
## Step:  AIC=771.97
## Outcome ~ Pregnancies + Glucose + BloodPressure + BMI + DiabetesPedigreeFunction +
##     Age
```

```
##
##                             Df Deviance    AIC
## - Age                        1   728.56 768.42
## <none>                           725.46 771.97
## - BloodPressure              1   732.51 772.37
## - DiabetesPedigreeFunction   1   734.99 774.85
## + Insulin                    1   723.45 776.60
## + SkinThickness              1   725.19 778.34
## - Pregnancies                1   741.27 781.13
## - BMI                        1   769.24 809.10
## - Glucose                    1   845.76 885.62
##
## Step:  AIC=768.42
## Outcome ~ Pregnancies + Glucose + BloodPressure + BMI + DiabetesPedigreeFunction
##
##                             Df Deviance    AIC
## - BloodPressure              1   734.31 767.52
## <none>                           728.56 768.42
## - DiabetesPedigreeFunction   1   738.43 771.65
## + Age                        1   725.46 771.97
## + Insulin                    1   725.97 772.48
## + SkinThickness              1   728.00 774.51
## - Pregnancies                1   760.56 793.78
## - BMI                        1   770.21 803.43
## - Glucose                    1   862.96 896.18
##
## Step:  AIC=767.52
## Outcome ~ Pregnancies + Glucose + BMI + DiabetesPedigreeFunction
##
##                             Df Deviance    AIC
## <none>                           734.31 767.52
## + BloodPressure              1   728.56 768.42
## - DiabetesPedigreeFunction   1   744.12 770.70
## + Insulin                    1   731.51 771.37
## + Age                        1   732.51 772.37
## + SkinThickness              1   733.06 772.92
## - Pregnancies                1   762.87 789.45
## - BMI                        1   771.27 797.85
## - Glucose                    1   864.84 891.41

##              (Intercept)              Pregnancies                  Glucose
##              -8.41585098               0.14192631               0.03382636
##                      BMI DiabetesPedigreeFunction
##               0.07809694               0.90129355
```

The Backward selection BIC and Hybrid BIC arrives at the same predictors.

What is the final model of your recommendation?

Backward selection with BIC selected 4 variables and BIC selected 4 variables.Any of the two BIC is preferred because both models have similarly adjusted R-squared and anyone with less variables is preferred.

Reduced model based on the most significant predictors

```
##
## Call:
## glm(formula = Outcome ~ Pregnancies + Glucose + BloodPressure +
##     BMI + DiabetesPedigreeFunction, family = binomial(), data = df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.7931  -0.7362  -0.4188   0.7251   2.9555
##
## Coefficients:
##                            Estimate Std. Error z value Pr(>|z|)
## (Intercept)              -7.954952   0.675823 -11.771  < 2e-16 ***
## Pregnancies               0.153492   0.027835   5.514  3.5e-08 ***
## Glucose                   0.034658   0.003394  10.213  < 2e-16 ***
## BloodPressure            -0.012007   0.005031  -2.387  0.01700 *
## BMI                       0.084832   0.014125   6.006  1.9e-09 ***
## DiabetesPedigreeFunction  0.910628   0.294027   3.097  0.00195 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 993.48  on 767  degrees of freedom
## Residual deviance: 728.56  on 762  degrees of freedom
## AIC: 740.56
##
## Number of Fisher Scoring iterations: 5
```

```
## Analysis of Deviance Table
##
## Model 1: Outcome ~ Pregnancies + Glucose + BloodPressure + BMI + DiabetesPedigreeFunction
## Model 2: Outcome ~ Pregnancies + Glucose + BloodPressure + SkinThickness +
##     Insulin + BMI + DiabetesPedigreeFunction + Age
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1       762     728.56
## 2       759     723.45  3   5.1142   0.1636
```

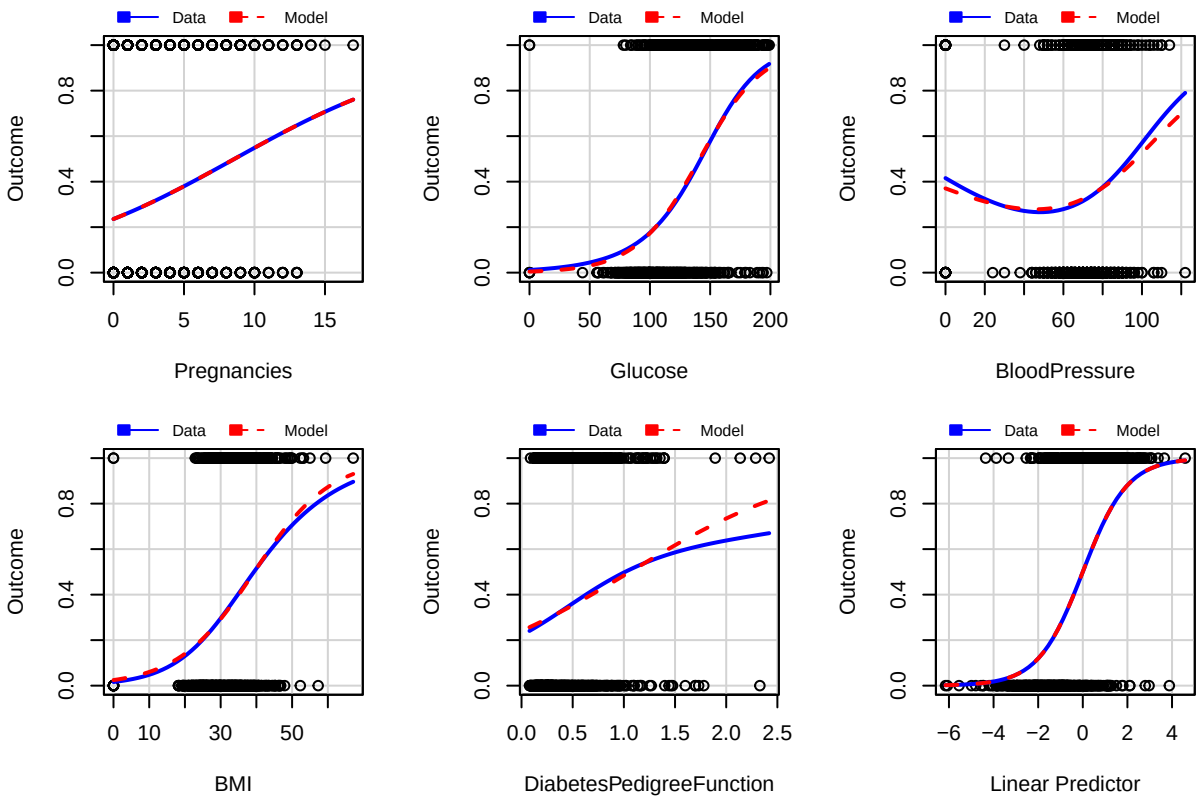
All of the regression coefficients in the model are now highly significant at the 5% level. The coefficients of the predictors Pregnancies, Glucose, BMI and DiabetesPedigreeFunction are positive implying that (all other things equal) higher Pregnancies, Glucose, BMI and DiabetesPedigreeFunction increases the chance of being diabetes patient as one would expect. However the coefficient of BloodPressure is negative.

Model Adequacy

Marginal model plots for model

```
## Loading required package: effects
```

```
## lattice theme set by effectsTheme()
## See ?effectsTheme for details.
```




```
## ResourceSelection 0.3-5   2019-07-22

##
##  Hosmer and Lemeshow goodness of fit (GOF) test
##
## data:  df$Outcome, fitted(m2)
## X-squared = 9.4733, df = 8, p-value = 0.304
```

Model Accuracy

ROC

```
##           pred
## 1 1 0.65750317
## 2 0 0.04428403
## 3 1 0.80775101
## 4 0 0.04863693
## 5 1 0.88621766
## 6 0 0.15434062

##
##   0   1
## 555 213

## [1] 213
```

```
##          response
## predicted   0   1
##         0 441 114
##         1  59 154
```

Out of 768 observations, We see that the model predicted 555 zeros and 213 ones. Out of the 213 actual diabetes patient, The model correctly predicted 154(True positive 72.3) and misclassified 59(False Positive 27.7). Also for the other part, the model correctly predicted 441 patient without diabetes(True Negative 79.5) and wrongly predicted 114 (False Negative 20.5)

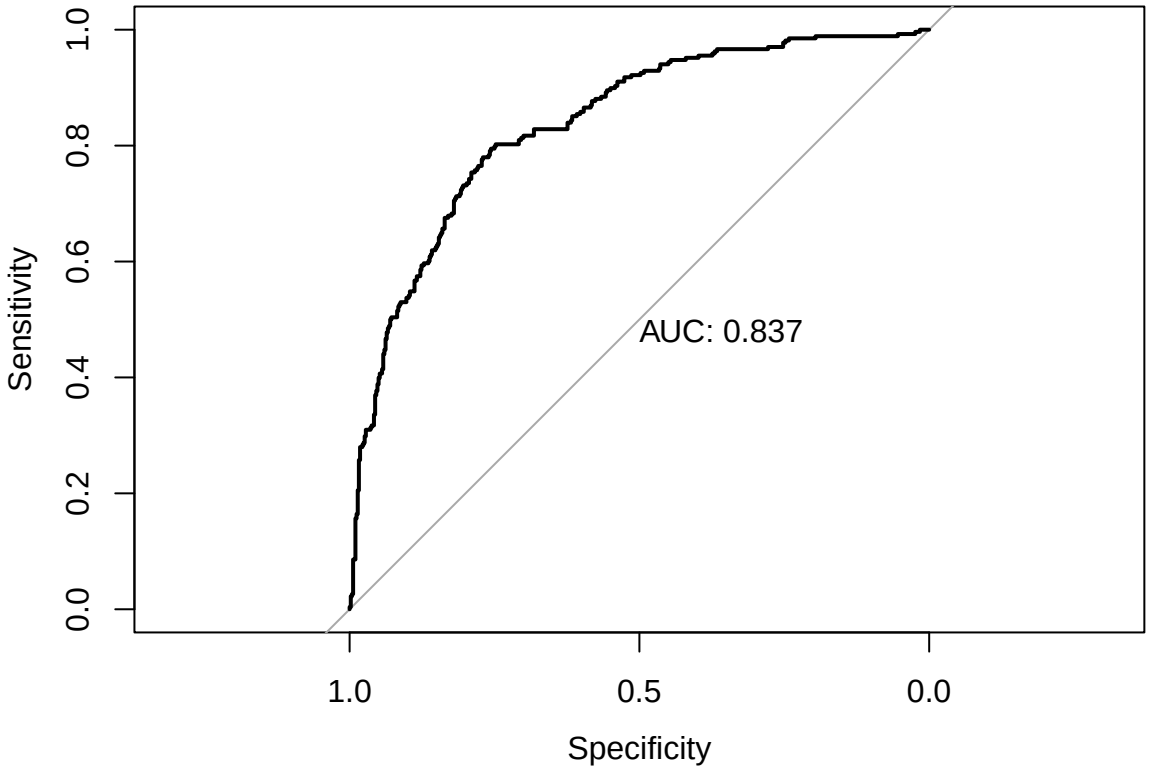
```
## Type 'citation("pROC")' for a citation.

##
## Attaching package: 'pROC'

## The following objects are masked from 'package:stats':
##
##     cov, smooth, var

## Setting levels: control = 0, case = 1

## Setting direction: controls < cases
```



Conclusion

In summary, we see that higher number of Pregnancies, Glucoselevel, BMI and DiabetesPedigreeFunction increases the chance of being diabetes greatly (with diabetePedegree being the most significant factor), while a low low blood pressure increases an individual's chances of being diagnosed with diabetes. This information would be valuable to researchers and medical practitioners looking for early warning signs of diabetes. Individuals who are aware that they have these high-risk traits could take more preventative measures and be screened for diabetes more often. The model as a whole could be used in the future to develop diabetes risk scores for medical patients. If an individual sees their risk score is getting closer to 1, or rising in that direction, they should be taking more preventative measures. Some limitations of this model for application in a clinical setting is that individuals may not have all the measurements that are necessary. Some of the predictors, such as glucose and insulin, are not measured frequently in doctors' visits unless the patient's risk of certain diseases is high. In conclusion, at AUC 83.7, the model is a relatively accurate to predict whether an individual has diabetes and can be used to improve medical outcomes.

References

Data: https://www.kaggle.com/datasets/mathchi/diabetes-data-set

Diabetes Information: https://www.cdc.gov/diabetes/basics/diabetes.html

Appendix