Detection of Diabetes using Machine Learning

STAT 835: Fall Project, 2020

Bibekananda Mishra

Department of Biostatistics and Data Science
University of Kansas, USA

December 16, 2020

Contents

Abstract

In order to predict diabetes disease from a set of 17 indicators containing data of 520 patients in Sylhet Diabetes Hopsital, we devlope a model having only four predictors which play very siginificant role in predicting the existence of the disease in a person. The four important inidcators are as follows: gender (being female increases the probability of having disease), polyurea, genital-thrush and Itching.

Introduction

Diabetes is one of the most dangerous noncummunicative diesease prevalent currently in the world. Around 8.8% of the adult population (and more than 10% of US population) are affected with this disease. With change in our lifestyle, this percntage is bound increase significantly in coming years. One of the key factors in diabetes is early detection i.e. before onset of any severe symptoms. The earlier one detects the disease the easier it becomes to treat and take effective care. We are analysing the data here collected from the patients of Sylhet Diabetes Hopsital, Bangladesh(https://archive.ics.uci.edu/ml/datasets/Early+stage+diabetes+risk+prediction+dataset.#).

The data contains 17 parameters including a parameter called 'class' which takes two possible values: positive and negative, where positive indicates presence of disease. Some of the parameters here like 'Age' and 'Gender' are self-explainatory. I have given a brief explanation of the other parameters involved below:

[1] Polyuria: Large production of urine for a person (more than 3 litre which is normal for an adult).

[2] Polydipsia: Feeling of excessive thirst.

[3] Polyphagia: Feeling of excessive hunger.

[4] Genital thrush: Fungal infection caused by Candida yeast.

[5] Partial paresis: Weakening of muscle.

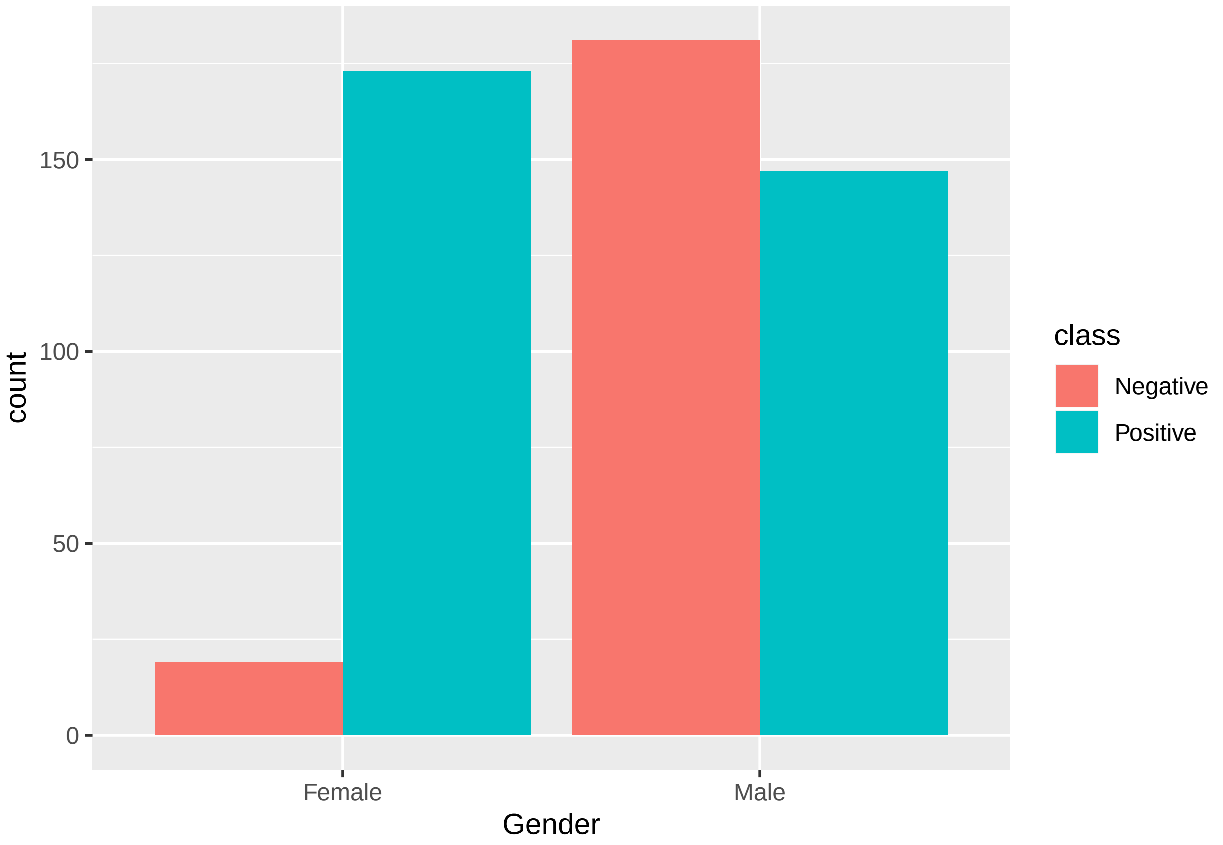
[6] Alopecia: Sudden hairloss in one or more circular patches.

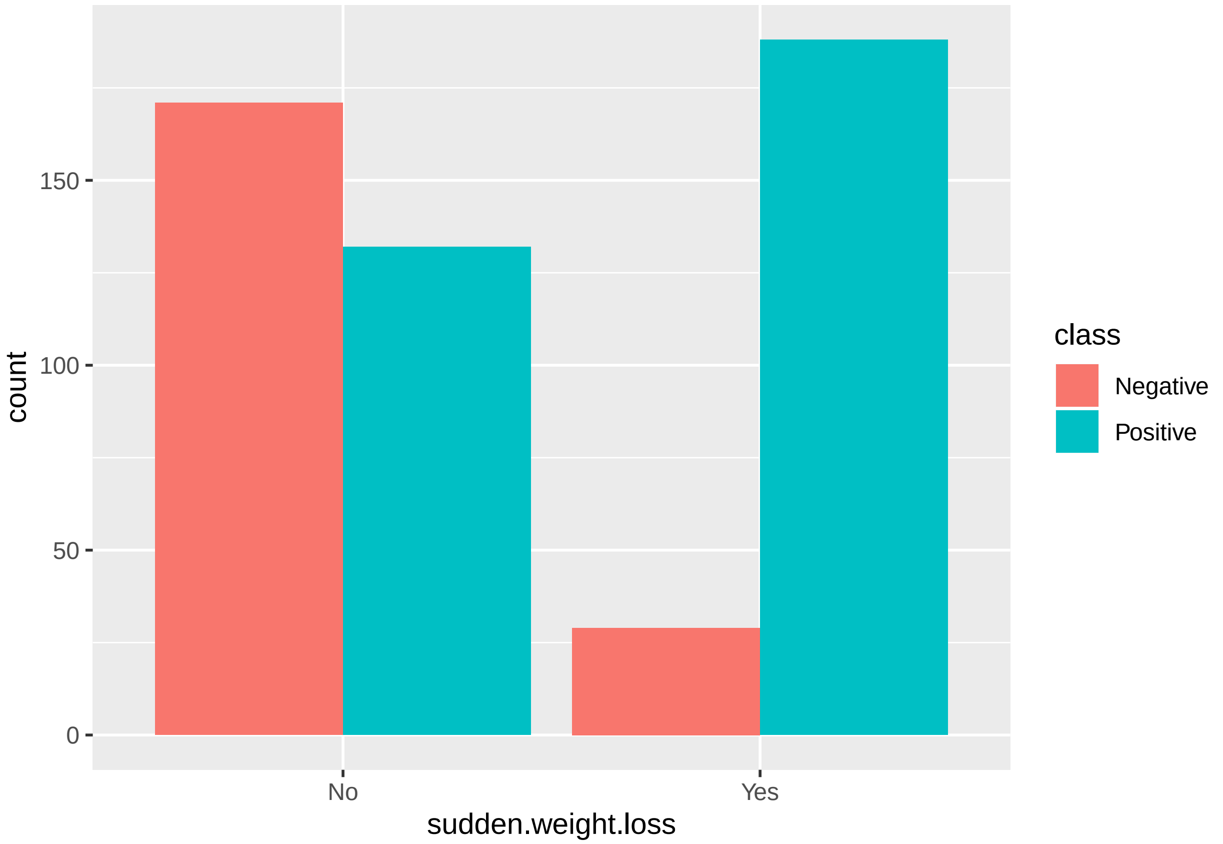
We have referred the book(3rd edition): Categorical data analysis by Alen Agresti and the lecture notes by Dr. Robert Montgomery for this work.

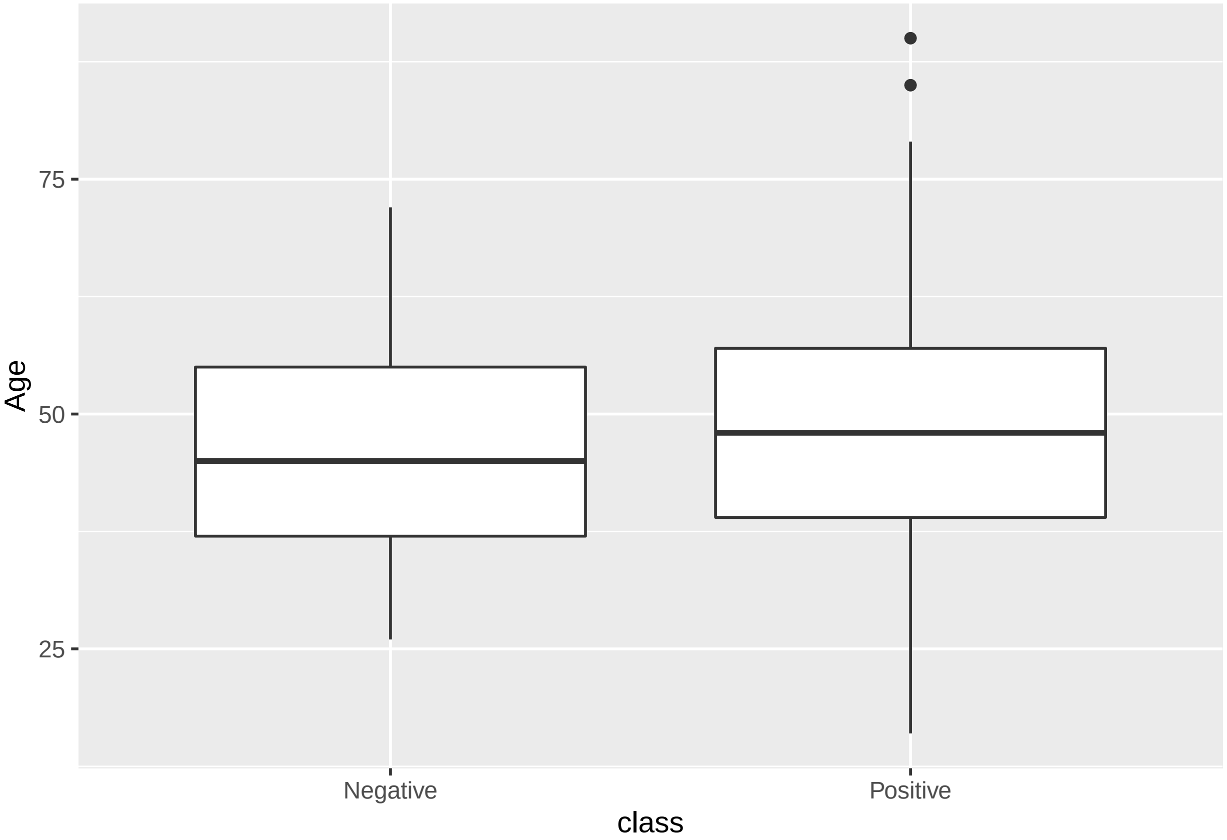
Primary Analysis Objectives

Our goal is to find a simple model, possibly having relatively fewer predictors and good accuracy in predicting the disease. First we explore the relations between the predictors and the respond variable. We will plot the graphs as follows.

From the graph below, it is clear that females have higher incidence of having diabetes compared to their male counterparts.


count
150
100
50
0
Female
Male
Gender
class
Negative
Positive


count
150
100
50
0
No
Yes
sudden.weight.loss
class
Negative
Positive



Similarly if we look at the persons experiencing sudden weight-loss we fing that they have higher incidence of diabetes compared to people not experiencing so among whom it is relatively evenly distributed. So, weight-loss is certianly a good indicator od having diabetes.

However, if we look at the boxplots of age distribution with incidence of diabetes, we notice that both have similar distributions. Even though the ‘positive’ boxplot is slightly above the ‘negative’ boxplot, there is not significant difference between the mean age of the people having diabetes and people not having so. So, it is clear that age is playing limited role in predicting the occurance of the disease. However, it may happen sometimes that with the combination of other predictors, it may play a decisive role which we will explore later.

From our preliminary analysis, it is intutive to suspect that some of the parameters must be related with each other. For example, a person having polydipsia also having polyuria is not very uncommmon. So, we will investigate the pairwise relation between the parameters. Since all the predictor variables except ‘Age’ are categorical, we look at the Chi-square test to figure out any pairwise relation. See the following table having the pairs and their corresponding value of Chi-squared statistics and p-values.

Polyuria	Polydipsia	186.3	0
Polyuria	sudden.weight.loss	104	0
Polydipsia	sudden.weight.loss	85.7	0
Polyuria	weakness	36	0
Polydipsia	weakness	57.5	0

sudden.weight.loss	weakness	41.6	0
Polyuria	Polyphagia	72.7	0
Polydipsia	Polyphagia	52.2	0
sudden.weight.loss	Polyphagia	30.8	0
weakness	Polyphagia	16.9	0
Polyuria	Genital.thrush	4	0.0466
Polydipsia	Genital.thrush	0.4	0.5219
sudden.weight.loss	Genital.thrush	4.2	0.0405
weakness	Genital.thrush	0.4	0.5264
Polyphagia	Genital.thrush	2.1	0.1463
Polyuria	visual.blurring	28.7	0
Polydipsia	visual.blurring	57.1	0
sudden.weight.loss	visual.blurring	2.5	0.1169
weakness	visual.blurring	47.1	0
Polyphagia	visual.blurring	44.8	0
Genital.thrush	visual.blurring	11.5	7e-04
Polyuria	Itching	4.1	0.0441
Polydipsia	Itching	8.6	0.0033
sudden.weight.loss	Itching	0	0.918
weakness	Itching	49.8	0
Polyphagia	Itching	10.8	0.001
Genital.thrush	Itching	8.2	0.0043
visual.blurring	Itching	44.1	0
Polyuria	Irritability	29.4	0
Polydipsia	Irritability	21.5	0
sudden.weight.loss	Irritability	10.2	0.0014
weakness	Irritability	11.2	8e-04
Polyphagia	Irritability	29.8	0
Genital.thrush	Irritability	13.4	3e-04
visual.blurring	Irritability	3.1	0.0787
Itching	Irritability	6.8	0.0093
Polyuria	delayed.healing	11.7	6e-04
Polydipsia	delayed.healing	7	0.0083
sudden.weight.loss	delayed.healing	4	0.0444
weakness	delayed.healing	58.5	0
Polyphagia	delayed.healing	36.2	0
Genital.thrush	delayed.healing	9.6	0.0019
visual.blurring	delayed.healing	16.4	1e-04
Itching	delayed.healing	106.9	0
Irritability	delayed.healing	8.4	0.0038
Polyuria	partial.paresis	101.4	0
Polydipsia	partial.paresis	101.7	0
sudden.weight.loss	partial.paresis	36.2	0
weakness	partial.paresis	38.7	0

Polyphagia	partial.paresis	72.6	0
Genital.thrush	partial.paresis	19.9	0
visual.blurring	partial.paresis	69	0
Itching	partial.paresis	7.1	0.0078
Irritability	partial.paresis	11.9	5e-04
delayed.healing	partial.paresis	18.3	0
Polyuria	muscle.stiffness	12.2	5e-04
Polydipsia	muscle.stiffness	17	0
sudden.weight.loss	muscle.stiffness	6.3	0.0123
weakness	muscle.stiffness	36	0
Polyphagia	muscle.stiffness	53.3	0
Genital.thrush	muscle.stiffness	5.2	0.0223
visual.blurring	muscle.stiffness	88.4	0
Itching	muscle.stiffness	24.2	0
Irritability	muscle.stiffness	21.1	0
delayed.healing	muscle.stiffness	32.5	0
partial.paresis	muscle.stiffness	28.1	0
Polyuria	Alopecia	10.8	0.001
Polydipsia	Alopecia	50.3	0
sudden.weight.loss	Alopecia	21.4	0
weakness	Alopecia	4.3	0.0391
Polyphagia	Alopecia	1.5	0.2225
Genital.thrush	Alopecia	21.8	0
visual.blurring	Alopecia	0.1	0.7391
Itching	Alopecia	36.9	0
Irritability	Alopecia	1	0.3189
delayed.healing	Alopecia	43.8	0
partial.paresis	Alopecia	25.5	0
muscle.stiffness	Alopecia	0.9	0.3527
Polyuria	Obesity	8.3	0.0039
Polydipsia	Obesity	5.1	0.0244
sudden.weight.loss	Obesity	14.9	1e-04
weakness	Obesity	1.1	0.2977
Polyphagia	Obesity	0.5	0.497
Genital.thrush	Obesity	1.5	0.2196
visual.blurring	Obesity	6.2	0.0129
Itching	Obesity	0	0.9655
Irritability	Obesity	8.5	0.0036
delayed.healing	Obesity	2.3	0.1303
partial.paresis	Obesity	0	0.8302
muscle.stiffness	Obesity	13.1	3e-04
Alopecia	Obesity	0.4	0.5051

Pairs having Chi-squared statistics value significantly above 2 indicate that there is not enough evidence for them being independednt of each other. For example as we suspected above, polyuria and polydipsia are related with each Chi-square statistics value greater than 186, significantly above 1.4 (with df=1) and p-value very close to 0. We will incorporate some of these dependency factors in finalising our best model for this data below.

Secondary Analysis Objectives

First we start with the full glm model having all the variables in them. Since the response variable is categorical with two levels, we will apply the binomial logistic regression model. See the summary below at Table-2.

Table 2: Full Model

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	2.7466	1.0755	2.55	0.0107
Age	-0.0512	0.0254	-2.02	0.0437
GenderMale	-4.3512	0.5982	-7.27	0.0000
PolyuriaYes	4.4395	0.7053	6.29	0.0000
PolydipsiaYes	5.0704	0.8289	6.12	0.0000
sudden.weight.lossYes	0.1903	0.5477	0.35	0.7282
weaknessYes	0.8171	0.5368	1.52	0.1280
PolyphagiaYes	1.1938	0.5335	2.24	0.0253
Genital.thrushYes	1.8637	0.5533	3.37	0.0008
visual.blurringYes	0.9159	0.6512	1.41	0.1596
ItchingYes	-2.8029	0.6727	-4.17	0.0000
IrritabilityYes	2.3407	0.5905	3.96	0.0001
delayed.healingYes	-0.3916	0.5500	-0.71	0.4765
partial.paresisYes	1.1593	0.5248	2.21	0.0272
muscle.stiffnessYes	-0.7288	0.5802	-1.26	0.2091
AlopeciaYes	0.1504	0.6201	0.24	0.8084
ObesityYes	-0.2890	0.5443	-0.53	0.5954

Looking at the p-values of the coefficients of each of the 16 predictor variables, we find that there are number of predictors which have limited impact on the probability, $P(Y = 1)$. Note that p-value for a predictor in this case tests the null-hypothesis that the respective coefficient in the logistic regression model is 0. If it is very close to 0 (let's say less than a threshold value of 0.05), we conclude that there is not enough evidence to support the null-hypothesis. So factors like 'delayed-healing' have less impact on the final prediction. So, we consider the following model, denoted as 'model1' removing such 'unimportant' predictors. We will also remove the predictor 'age' as we see above that it has limited role in the final prediction. Interestingly, sudden-weight-loss, though found to be an important indicator of the disease, turns out be insignificant in the logistic regression model. See the table-3 below.

Table 3: Reduced Model

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	1.0772	0.4468	2.41	0.0159
GenderMale	-4.4616	0.5801	-7.69	0.0000
PolyuriaYes	3.9117	0.5589	7.00	0.0000
PolydipsiaYes	5.1549	0.7375	6.99	0.0000
PolyphagiaYes	0.8298	0.4261	1.95	0.0515
Genital.thrushYes	1.7632	0.5021	3.51	0.0004
ItchingYes	-2.6816	0.4931	-5.44	0.0000
IrritabilityYes	2.1240	0.5087	4.18	0.0000
partial.paresisYes	1.0459	0.4371	2.39	0.0167

Next, to find the best model having the most appropriate predictors, we will consider the AIC score which find the suitable balance between fitness and the number of predictors (punishing models with greater number of predictors).

```
library(MASS)
model1_2 <- stepAIC(model_1, direction = "both", trace = FALSE)
model1_2
```

```
Call: glm(formula = class ~ Gender + Polyuria + Polydipsia + Polyphagia +
Genital.thrush + Itching + Irritability + partial.paresis, family = binomial, data =
diabetes)

Coefficients: (Intercept) GenderMale PolyuriaYes
1.08 -4.46 3.91
PolydipsiaYes PolyphagiaYes Genital.thrushYes
5.15 0.83 1.76
ItchingYes IrritabilityYes partial.paresisYes
-2.68 2.12 1.05

Degrees of Freedom: 519 Total (i.e. Null); 511 Residual Null Deviance: 693 Residual
Deviance: 186 AIC: 204
```

```
AIC(model_full)
```

```
[1] 205.65
```

Interestingly the resultant model having the least AIC score, called 'model1_2' turns out to be the same model as the 'model-1' This is probably because we have already removed the non-essential variables from the full model thus reducing the AIC value already in the process.

Lastly, we will consider the following model, called 'model-3', where we have removed some of the predictors that are strongly correlated with each other (which we analysed above).

This is pretty much based on heuristic sense and explored for experimental purpose. This is also along our main goal of finding a sufficiently simple model for prediction. We will remove polydipsia as it is strongly related to polydipsia. Similarly, we will remove 'Irritability' as it is significantly correlated to many of the predictors considered already. So, our final model will be as follows. See table-4.

Table 4: Final Model

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	1.3567	0.3287	4.13	0.0000
GenderMale	-2.9112	0.3497	-8.33	0.0000
PolyuriaYes	3.9439	0.3528	11.18	0.0000
Genital.thrushYes	1.0372	0.3334	3.11	0.0019
ItchingYes	-1.0621	0.3105	-3.42	0.0006

[1] 341.1

Now, we will consider the performances of the three models: full-model, model-1 and model-3. We have considered the AIC score above and it is least for model1. However, as we have reamrked above, AIC takes into account the number of predictors and punishes for higher such number. We will consider other parameters for evaluating the performances. One such criteria is the deviance. Usually lower the deviance better the model.

```r
cbind(deviance(model_full), deviance(model_1), deviance(model_3))
```

```
##        [,1]   [,2]  [,3]
## [1,] 171.65 185.69 331.1
```

```r
K <- matrix(nrow = 3, ncol = 4)
K[1, ] <- cbind(mean(rstandard(model_full, type = "pearson")),
    mean(residuals(model_full, type = "pearson")), mean(residuals(model_full,
        type = "deviance")), mean(rstandard(model_full, type = "deviance")))
K[2, ] <- cbind(mean(rstandard(model_1, type = "pearson")), mean(residuals(model_1,
    type = "pearson")), mean(residuals(model_1, type = "deviance")),
    mean(rstandard(model_1, type = "deviance")))
K[3, ] <- cbind(mean(rstandard(model_3, type = "pearson")), mean(residuals(model_3,
    type = "pearson")), mean(residuals(model_3, type = "deviance")),
    mean(rstandard(model_3, type = "deviance")))
print(K)
```

```
##           [,1]      [,2]      [,3]      [,4]
## [1,] 0.0039193 0.0020288 -0.013985 -0.012270
## [2,] 0.0152287 0.0155010 -0.010423 -0.010873
## [3,] 0.0363353 0.0363822  0.024090  0.024101
```

Clearly, the deviance of the full model is the least indicating the best fitness of the model. However, it is possible that it may be over-fitting. we need to check this using on a test dataset and compare the performance of the three models. At this moment due to lack of time, we will go with the simple model i.e. model_3 as the performance is not being compromised too much in this model and we are getting a relatively simpler model.

Now looking at the summary of model_3(table-4), we will interpret the coefficients. The model actually looks like this:

$$P(Y=1) = \frac{e^{1.3567-2.911.Gender+3.944.Polyurea+1.037.thrush-1.062Itching}}{1+e^{1.3567-2.911.Gender+3.944.Polyurea+1.037.thrush-1.062Itching}}$$

The coefficient of Polyuria being 3.94 implies with each unit increase (in this case, simply patient having polyurea) contribute 3.94 times the $logit(P(Y=1))$ or the logit value of the probability of the patient having diabetes. Similarly, Gender has negative effect as we saw in our explorative analysis above.

Discussion and Conclusion

We finnally conclude that gender, polyurea, genital thrush and itching play important role in diagonising whether a person has diabetes or not. We must note that we have found a simple model by compromising the deviance. As mentioned above, we can apply this model on a test dataset to see whether the other models available are over fitting. One direction we have not explored is the higher degree interaction with numerical avriable Age or models having interaction terms of predictors. It is also worthwhile to mention here that when I checked this model with one of my friends who is in medical profession, I found that the above three factors (polyurea, thrush and ithcing) may be common symptoms of Urinary Track Infection (UTI). So, clearly we need a better model to improve our identifier. For the time being this is our final model.

Appendix: R-code

```
#https://archive.ics.uci.edu/ml/datasets/Early+stage+diabetes+risk+prediction+dataset.
library(data.table)
library(ggplot2)
library(plyr)
diabetes<-read.csv("~/Downloads/diabetes_data_upload.csv", quote="\"", comment.char="")
head(diabetes)
#data exploration
ggplot(diabetes, aes(x= Gender, y= ..count.., fill= class)) + geom_bar(aes(fill = class), posi
ggplot(diabetes, aes(x= sudden.weight.loss, y= ..count.., fill= class)) + geom_bar(aes(fill = 
ggplot(diabetes, aes(x=class, y= Age)) + geom_boxplot()
dbts<-diabetes[,c('Age', 'Polyphagia','delayed.healing', 'class')]
model_full<- glm(class~., diabetes, family = binomial)
summary(model_full)
library(MASS)
model1<-glm(class~ Gender+ Polyuria + Polydipsia + Polyphagia + Genital.thrush+ Itching +Irrit
step.model <- stepAIC(model1, direction = "both", trace = FALSE)
summary(step.model)
AIC(step.model)
deviance(model1)
deviance(step.model)
B<-count(diabetes, c('Polyuria', 'Polydipsia'))
tbl = matrix(data=c(222, 40, 65, 193), nrow=2, ncol=2, byrow=T)
chi2 = chisq.test(tbl, correct=F)
chi2
fisher.test(tbl)
x=3
J= matrix(nrow=91, ncol=4)
i=1
while (x<17)
{
   y=3
   while(y<x)
   {
      B= count(diabetes, c(A[x], A[y]))
      tbl= matrix(data=c(B[1,3], B[2,3], B[3,3], B[4,3]), nrow=2, ncol=2, byrow=T)
      J[i,3]= chisq.test(tbl, correct=F)$statistic
      J[i,2]= C[x]
      J[i,1]= C[y]
      J[i,4]= chisq.test(tbl, correct=F)$p.value
      i<-i+1
      y<-y+1
   }

   x<-x+1
}
```