Predicting the Likelihood of Diabetes Using Common Signs and Symptoms

Project Phase II | MATH1298 Analysis of Categorical Data | RMIT University

Udeshika Dissanayake | s3400652 | Project Groups 60

October 31, 2020

List of Contents

1. Introduction

About one third of patients with diabetes do not know that they have diabetes according to the findings published by many diabetes institutes around the world (ASPE, 2017). Detecting and treating diabetes patients at early stages is critical in order to keep them healthy and to ensure their quality of life is not compromised. Early detection will also help to mitigate the risk of serious complications like heart disease & stroke, blindness, limb amputations, and kidney failures as a result of diabetes (ASPE, 2017).

This study intends to build a logistic regression model to predict the likelihood of having diabetes using common signs and symptoms presented by patients. A successful model will enable early detection of diabetes through signs and symptoms shown by possible patients.

This study consists with two phases: 1) Phase I - preprocess and explore the data set in order to make it ready to consume for model development. 2) Phase II - build a logistic regression model to predict the likelihood of having diabetes based on signs and symptoms. The Phase I part has already been completed under previous work/submission and this report intends to cover the work carried out for Phase II.

All the activities have been performed in R package and the report has been compiled using R-Markdown. This report covers both narratives and R pseudocode for Statistical modelling activities that have been performed under the phase II.

Data Set

The data set consists of signs and symptoms of 516 newly diabetic or would be diabetic patients, who presented at Sylhet Diabetes Hospital in Sylhet, Bangladesh. The data had been collected using direct questionnaires method at the hospital under the supervisor of Doctors. The Source for the data set is the UCI Machine Learning Repository (Dua, D. & Graff, C., 2017) at, archive.ics.uci.edu (https://archive.ics.uci.edu/ml/datasets/Early+stage+diabetes+risk+prediction+dataset.) (Faniqul, M. M. I. et al., 2019). The data set has 16 descriptive features and one target feature.

Descriptive Features

Below table explains the descriptive features in the data set that will be used in the model.

Table 1: Descriptive features

Name	Data Type	Units	Description
Age	numerical	Years	Age of the patient
Gender	binary	Female, Male	Gender of the patient
Polyuria	binary	Yes, No	Body urinates more than usual and passes excessive or abnormally large amounts of urine each time
Polydipsia	binary	Yes, No	Dry mouth and excessive thirst
Sudden weight loss	binary	Yes, No	Unexplained sudden weight loss
Weakness	binary	Yes, No	Fatigue/ Feelings of exhaustion and lethargy

Name	Data Type	Units	Description
Polyphagia	binary	Yes, No	Excessive or extreme hunger
Genital thrush	binary	Yes, No	Itching, irritation and swelling around the Genital organs
Visual blurring	binary	Yes, No	Lack of sharpness of vision and inability to see fine details
Itching	binary	Yes, No	Irritating sensation that makes you want to scratch your skin
Irritability	binary	Yes, No	Feeling frustrated or getting upset easily
Delayed healing	binary	Yes, No	Delayed wound healing, recurrent or severe infections
Partial paresis	binary	Yes, No	Weakening of a muscle or group of muscles
Muscle stiffness	binary	Yes, No	Muscles feel tight and more difficult to move than you usually do
Alopecia	binary	Yes, No	Patches of hair loss on the head and on other parts of the body
Obesity	binary	Yes, No	Excessive amount of body fat

Target Feature

The name of the target feature is “Class” and it’s labels are as follows,

$$\text{Class} = \begin{cases} Positive & \text{if the patient is diagnosed as a diabetic patient} \\ Negative & \text{if the patient is not diagnosed as a diabetic patient} \end{cases}$$

The target feature has two levels. Hence this can be classified as binomial target feature.

Methodology

In order to predict the likelihood of having diabetes using common signs and symptoms, a logistic regression model is formulated. The data set has been pre-processed and explored in the previous Phase (Phase I). For all categorical attributes, the requirement for the Dummy encoding is investigated. The “Class” variable, which is a binary variable is used as the Target feature. The model with main-effects is improved and optimized using the feature selection (forward selection with AIC) method. The model is further improved using incorporation of 2-way interactions of selected features. The Standardized Pearson residual analysis is performed against the “Age” variable in order to check the validity of the model. The Goodness of Fit study is undertaken to evaluate how well the model fits for all the observations at once. The Response Analysis, Confident Interval, Hypothesis Test, and Sensitivity Analysis are performed to observe how the model behaves and responds for different circumstances.

2. Statistical Modelling

Data Preprocessing - Dummy Encoding

The bulk of the data preprocessing has been done under Phase I. However, the dummy encoding of categorical attributes has been kept to preform under Phase II.

Let’s recall the data set by observing 5 random rows of the pre-process data.

```
#Random 5 rows of the data set
kbl(sample_n(df,5), caption = "Table 2: Random 5 rows from data set") %>%
  kable_styling(bootstrap_options = c("striped", "hover", "condensed"), full_width = F, position = "left",font_size = 1
0)
```

Table 2: Random 5 rows from data set

Age	Gender	Polyuria	Polydipsia	sudden weight loss	weakness	Polyphagia	Genital thrush	visual blurring	Itching	Irritability	delayed healing	partial paresis	muscle stiffness	Alopecia	Obesity	class
43	Male	No	No	No	No	No	No	No	No	No	No	No	No	Yes	No	Negative
43	Female	Yes	Yes	Yes	Yes	Yes	No	Yes	No	No	No	Yes	Yes	No	Yes	Positive
54	Male	No	No	Yes	Yes	No	Yes	No	No	No	Yes	No	No	Yes	No	Negative
65	Female	Yes	Yes	No	Yes	Yes	No	No	Yes	No	No	Yes	Yes	No	No	Positive
50	Male	No	No	No	Yes	No	No	No	Yes	No	Yes	Yes	Yes	Yes	No	Negative

As can be seen, all the categorical attributes in the data sets have two levels (Yes or No), therefore each attribute could be encoded to have binary variable (i.e. 1 or 0) using dummy encoding. It is essential to make sure the categorical attributes that are to be dummy encoded are in “Factor” type. As shown in the R-output below, the data type of the ‘Age’ feature is “numeric”, whereas the data type for all the other descriptive features including target is “Factor”.

```
sapply(df, class)#checking variable types in the data frame
```

```
##               Age            Gender           Polyuria         Polydipsia
##          "numeric"          "factor"           "factor"           "factor"
## sudden weight loss          weakness         Polyphagia     Genital thrush
##           "factor"          "factor"           "factor"           "factor"
##    visual blurring           Itching       Irritability    delayed healing
##           "factor"          "factor"           "factor"           "factor"
##    partial paresis   muscle stiffness           Alopecia            Obesity
##           "factor"          "factor"           "factor"           "factor"
##              class
##           "factor"
```

Using the *contrasts()* function, the default indicator values that *R* has sets are observed below:

```
#Checking Levels
for (n in names(df))
  if (is.factor(df[[n]])) {
    print(n)
    print(contrasts(df[[n]]))
  }
```

```
## [1] "Gender"
##        Male
## Female    0
## Male      1
## [1] "Polyuria"
##     Yes
## No    0
## Yes   1
## [1] "Polydipsia"
##     Yes
## No    0
## Yes   1
## [1] "sudden weight loss"
##     Yes
## No    0
## Yes   1
## [1] "weakness"
##     Yes
## No    0
## Yes   1
## [1] "Polyphagia"
##     Yes
## No    0
## Yes   1
## [1] "Genital thrush"
##     Yes
## No    0
## Yes   1
## [1] "visual blurring"
##     Yes
## No    0
## Yes   1
## [1] "Itching"
##     Yes
## No    0
## Yes   1
## [1] "Irritability"
##     Yes
## No    0
## Yes   1
## [1] "delayed healing"
##     Yes
## No    0
## Yes   1
## [1] "partial paresis"
##     Yes
## No    0
## Yes   1
## [1] "muscle stiffness"
##     Yes
## No    0
## Yes   1
## [1] "Alopecia"
##     Yes
## No    0
## Yes   1
## [1] "Obesity"
##     Yes
## No    0
## Yes   1
## [1] "class"
##          Positive
## Negative        0
## Positive        1
```

As shown above, by default R has given “No”=0 and “Yes”=1 for the attributes with “Yes”" and “No” levels. For the attribute “Gender”, R has set “Female”=0 and “Male”=1. Finally for the target attribute, R has set “Negative”=0, “Positive”=1.
Since the encoded values are in meaningful order, no further tweaking in the encoding is required.

2.1. Model Fitting

This section covers the development, attribute selection, and validation of a logistic regression model to predict the likelihood of having diabetes using common signs and symptoms presented by patients. The model fitting task starts with the full-model considering all the main effects and then improving the model by selecting and dropping attributes as needed.

Full-Model with Main Effects

As the starting point of the model development, the full-model is considered with all the main effects. The model coefficients, Z values and P values for each feature are obtained as shown below:

```
full.mod <- glm(formula = class ~., family = binomial(link = logit), data = df)
summary(object = full.mod)
```

```
##
## Call:
## glm(formula = class ~ ., family = binomial(link = logit), data = df)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.86214  -0.23183   0.00322   0.05202   2.80804
##
## Coefficients:
##                         Estimate Std. Error z value Pr(>|z|)
## (Intercept)              2.83038    1.09508   2.585  0.00975 **
## Age                     -0.05369    0.02610  -2.057  0.03968 *
## GenderMale              -4.31790    0.60081  -7.187 6.64e-13 ***
## PolyuriaYes              4.45682    0.70850   6.291 3.16e-10 ***
## PolydipsiaYes            5.02729    0.83049   6.053 1.42e-09 ***
## `sudden weight loss`Yes  0.15843    0.55202   0.287  0.77411
## weaknessYes              0.84479    0.54181   1.559  0.11895
## PolyphagiaYes            1.22159    0.53843   2.269  0.02328 *
## `Genital thrush`Yes      1.82468    0.55831   3.268  0.00108 **
## `visual blurring`Yes     0.90277    0.65182   1.385  0.16605
## ItchingYes              -2.79552    0.67074  -4.168 3.08e-05 ***
## IrritabilityYes          2.33686    0.58750   3.978 6.96e-05 ***
## `delayed healing`Yes    -0.36079    0.55378  -0.651  0.51472
## `partial paresis`Yes     1.18538    0.52982   2.237  0.02527 *
## `muscle stiffness`Yes   -0.76033    0.58474  -1.300  0.19350
## AlopeciaYes              0.13938    0.62271   0.224  0.82289
## ObesityYes              -0.28118    0.54322  -0.518  0.60473
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 689.03  on 515  degrees of freedom
## Residual deviance: 171.37  on 499  degrees of freedom
## AIC: 205.37
##
## Number of Fisher Scoring iterations: 8
```

Further, the LRT test has been performed to identify the important main effects to the model,

```
Anova(full.mod)
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: class
##                       LR Chisq Df Pr(>Chisq)
## Age                      4.524  1  0.0334139 *
## Gender                  75.904  1  < 2.2e-16 ***
## Polyuria                68.088  1  < 2.2e-16 ***
## Polydipsia              73.081  1  < 2.2e-16 ***
## `sudden weight loss`     0.082  1  0.7746134
## weakness                 2.460  1  0.1167578
## Polyphagia               5.422  1  0.0198897 *
## `Genital thrush`        11.593  1  0.0006619 ***
## `visual blurring`        1.979  1  0.1595190
## Itching                 21.643  1  3.285e-06 ***
## Irritability            18.005  1  2.203e-05 ***
## `delayed healing`        0.423  1  0.5154455
## `partial paresis`        5.260  1  0.0218168 *
## `muscle stiffness`       1.695  1  0.1929830
## Alopecia                 0.050  1  0.8228360
## Obesity                  0.269  1  0.6039657
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

It is evident based on the smaller P-values (<0.05) from *Anova()* output, the important main effects to the model are: Age, Gender, Polyuria, Polydipsia, Polyphagia, Genital thrush, Itching, Irritability, Partial Paresis. Also, relatively higher P-values (>0.05) suggest the non-important main effects to the model as: Sudden Weight Loss, Weakness, Visual Blurring, Delayed Healing, Muscle Stiffness, Alopecia, and Obesity.

It is worth noting from the *summary()* output of the full-model, the AIC value is 205.37. This figure will be used to compare the performance of the improved models in future sections.

Feature Selection for Main Effects

In this section, the Forward Selection is used as the model selection criteria, while the Akaike's Information Criterion (AIC) is used as the information criteria in order to compare the performance of different model by changing the attribute combinations. The intention of this task is to optimize the model with main effects by selecting the most impactful features. The most optimized model is derived when the AIC value is the lowest. Below code shows the AIC values for each iteration of the Forward Selection criteria for attribute selection:

```
empty.mod <- glm(formula = class ~ 1, family = binomial(link = logit), data = df)
forw.sel_AIC <- step(object = empty.mod,
                     scope = list(upper=full.mod),
                     direction = "forward",
                     k = 2,
                     trace = TRUE)
```

```
## Start:  AIC=691.03
## class ~ 1
##
##                        Df Deviance    AIC
## + Polyuria              1   426.62 430.62
## + Polydipsia            1   433.81 437.81
## + Gender                1   571.48 575.48
## + `sudden weight loss`  1   584.52 588.52
## + `partial paresis`     1   584.66 588.66
## + Polyphagia            1   625.60 629.60
## + Irritability          1   635.33 639.33
## + Alopecia              1   651.32 655.32
## + `visual blurring`     1   657.02 661.02
## + weakness              1   657.82 661.82
## + `muscle stiffness`    1   682.04 686.04
## + `Genital thrush`      1   683.67 687.67
## + Age                   1   684.55 688.55
## + Obesity               1   686.04 690.04
## <none>                      689.03 691.03
## + `delayed healing`     1   687.90 691.90
## + Itching               1   688.82 692.82
##
## Step:  AIC=430.62
## class ~ Polyuria
##
##                        Df Deviance    AIC
## + Polydipsia            1   338.66 344.66
## + Gender                1   349.43 355.43
## + Alopecia              1   396.47 402.47
## + Irritability          1   401.97 407.97
## + `partial paresis`     1   403.02 409.02
## + `sudden weight loss`  1   405.59 411.59
## + Polyphagia            1   416.43 422.43
## + `visual blurring`     1   418.95 424.95
## + Itching               1   420.54 426.54
## + weakness              1   421.54 427.54
## + `delayed healing`     1   424.24 430.24
## + Age                   1   424.25 430.25
## <none>                      426.62 430.62
## + `Genital thrush`      1   424.92 430.92
## + `muscle stiffness`    1   426.47 432.47
## + Obesity               1   426.51 432.51
##
## Step:  AIC=344.66
## class ~ Polyuria + Polydipsia
##
##                        Df Deviance    AIC
## + Gender                1   256.07 264.07
## + Itching               1   316.63 324.63
## + Alopecia              1   319.74 327.74
## + Irritability          1   329.15 337.15
## + `sudden weight loss`  1   329.74 337.74
## + `partial paresis`     1   329.94 337.94
## + `delayed healing`     1   331.37 339.37
## + Age                   1   331.85 339.85
## + Polyphagia            1   334.51 342.51
## <none>                      338.66 344.66
## + Obesity               1   336.81 344.81
## + `muscle stiffness`    1   336.86 344.86
## + `Genital thrush`      1   337.52 345.52
## + weakness              1   338.40 346.40
## + `visual blurring`     1   338.53 346.53
##
## Step:  AIC=264.07
## class ~ Polyuria + Polydipsia + Gender
##
##                        Df Deviance    AIC
## + Itching               1   227.14 237.14
## + Irritability          1   235.61 245.61
## + `delayed healing`     1   248.32 258.32
## + Alopecia              1   248.50 258.50
## + Age                   1   250.44 260.44
```

```
## + `Genital thrush`      1   250.64 260.64
## + `sudden weight loss`  1   251.52 261.52
## + `partial paresis`     1   252.80 262.80
## + `visual blurring`     1   253.19 263.19
## + `muscle stiffness`    1   253.36 263.36
## + Polyphagia            1   253.57 263.57
## <none>                      256.07 264.07
## + Obesity               1   254.89 264.89
## + weakness              1   255.81 265.81
##
## Step:  AIC=237.14
## class ~ Polyuria + Polydipsia + Gender + Itching
##
##                       Df Deviance    AIC
## + Irritability         1   203.80 215.80
## + `Genital thrush`     1   216.29 228.29
## + weakness             1   220.80 232.80
## + Polyphagia           1   223.24 235.24
## + `partial paresis`    1   223.34 235.34
## + `sudden weight loss` 1   224.39 236.39
## <none>                     227.14 237.14
## + `muscle stiffness`   1   225.32 237.32
## + Alopecia             1   226.79 238.79
## + `delayed healing`    1   226.82 238.82
## + Age                  1   226.87 238.87
## + `visual blurring`    1   226.95 238.95
## + Obesity              1   227.07 239.07
##
## Step:  AIC=215.8
## class ~ Polyuria + Polydipsia + Gender + Itching + Irritability
##
##                       Df Deviance    AIC
## + `Genital thrush`     1   196.81 210.81
## + `partial paresis`    1   200.24 214.24
## + `sudden weight loss` 1   200.78 214.78
## + Polyphagia           1   201.03 215.03
## + Age                  1   201.44 215.44
## + weakness             1   201.69 215.69
## + `muscle stiffness`   1   201.71 215.71
## <none>                     203.80 215.80
## + Obesity              1   202.78 216.78
## + `visual blurring`    1   203.34 217.34
## + `delayed healing`    1   203.49 217.49
## + Alopecia             1   203.50 217.50
##
## Step:  AIC=210.81
## class ~ Polyuria + Polydipsia + Gender + Itching + Irritability +
##     `Genital thrush`
##
##                       Df Deviance    AIC
## + `partial paresis`    1   189.49 205.49
## + Polyphagia           1   191.56 207.56
## + `visual blurring`    1   193.19 209.19
## <none>                     196.81 210.81
## + weakness             1   194.90 210.90
## + `sudden weight loss` 1   195.00 211.00
## + Obesity              1   195.22 211.22
## + Age                  1   195.42 211.42
## + `delayed healing`    1   195.81 211.81
## + Alopecia             1   196.17 212.17
## + `muscle stiffness`   1   196.19 212.19
##
## Step:  AIC=205.49
## class ~ Polyuria + Polydipsia + Gender + Itching + Irritability +
##     `Genital thrush` + `partial paresis`
##
##                       Df Deviance    AIC
## + Polyphagia           1   185.66 203.66
## + weakness             1   186.13 204.13
## + Age                  1   186.32 204.32
## <none>                     189.49 205.49
## + `sudden weight loss` 1   187.81 205.81
## + `visual blurring`    1   188.14 206.14
```

```
## + Obesity               1   188.16 206.16
## + `muscle stiffness`    1   188.79 206.79
## + `delayed healing`     1   188.84 206.84
## + Alopecia              1   189.14 207.14
##
## Step:  AIC=203.66
## class ~ Polyuria + Polydipsia + Gender + Itching + Irritability +
##     `Genital thrush` + `partial paresis` + Polyphagia
##
##                        Df Deviance    AIC
## + Age                   1   179.55 199.55
## + weakness              1   182.39 202.39
## + `muscle stiffness`    1   183.44 203.44
## + `sudden weight loss`  1   183.45 203.45
## <none>                      185.66 203.66
## + Obesity               1   184.98 204.98
## + `delayed healing`     1   185.16 205.16
## + `visual blurring`     1   185.19 205.19
## + Alopecia              1   185.26 205.26
##
## Step:  AIC=199.55
## class ~ Polyuria + Polydipsia + Gender + Itching + Irritability +
##     `Genital thrush` + `partial paresis` + Polyphagia + Age
##
##                        Df Deviance    AIC
## + weakness              1   176.03 198.03
## <none>                      179.55 199.55
## + `visual blurring`     1   177.59 199.59
## + `sudden weight loss`  1   177.92 199.92
## + `muscle stiffness`    1   178.47 200.47
## + Obesity               1   178.81 200.81
## + Alopecia              1   179.33 201.33
## + `delayed healing`     1   179.55 201.55
##
## Step:  AIC=198.03
## class ~ Polyuria + Polydipsia + Gender + Itching + Irritability +
##     `Genital thrush` + `partial paresis` + Polyphagia + Age +
##     weakness
##
##                        Df Deviance    AIC
## <none>                      176.03 198.03
## + `muscle stiffness`    1   174.20 198.20
## + `visual blurring`     1   174.43 198.43
## + Obesity               1   175.32 199.32
## + `delayed healing`     1   175.75 199.75
## + `sudden weight loss`  1   175.78 199.78
## + Alopecia              1   176.02 200.02
```

As can be seen from above outputs, the AIC has been improved to 198.03 (from 205.37 for full model) with the selected attributes as: Polyuria, Polydipsia, Gender, Itching, Irritability, Genital Thrush, Partial Paresis, Polyphagia, Age, and Weakness.

LRT Testing for Reduced Model

In order to test the statistical significance of the above selected features, the LRT test is performed as below:

```
mod.fit_AIC<-glm(formula = class ~ Age + Gender + Polyuria + Polydipsia + weakness + Polyphagia +
                   `Genital thrush` + Itching + Irritability + `partial paresis`,
                 family = binomial(link = logit), data = df)
Anova(mod.fit_AIC)
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: class
##                    LR Chisq Df Pr(>Chisq)
## Age                   6.360  1  0.0116729 *
## Gender               91.314  1  < 2.2e-16 ***
## Polyuria             80.348  1  < 2.2e-16 ***
## Polydipsia           88.864  1  < 2.2e-16 ***
## weakness              3.515  1  0.0608011 .
## Polyphagia            6.272  1  0.0122641 *
## `Genital thrush`     12.594  1  0.0003871 ***
## Itching              28.370  1  1.002e-07 ***
## Irritability         18.624  1  1.592e-05 ***
## `partial paresis`     9.847  1  0.0017014 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

It is evident from the P-values above, almost all the features have P-values less than 0.05 except for the attribute “Weakness”, which has the p-value of 0.0608. This suggested the attribute “Weakness” is marginally significant and it can be ignored from the model.

Adding 2-way Interactions to the model

Based on the observations in the Phase I data exploration task, below two-way interactions have been chosen to be considered to the model:

Polyuria:Age
Polyuria:sudden weight loss
Age:visual blurring
Polyuria:Polydipsia
Polyuria:Gender
partial paresis:muscle stiffness
gender:sudden weight loss
gender:Genital thrush

Using below steps, each two-way interaction is added (one at a time) to the reduced model and respective AIC values have been checked:

```
#Polyuria:Age
mod.fit1<-glm(formula = class ~ Age + Gender + Polyuria + Polydipsia + Polyphagia + `Genital thrush` +
                  Itching + Irritability + `partial paresis` + Polyuria:Age,
              family = binomial(link = logit), data = df)

# Polyuria:`sudden weight loss`
mod.fit2<-glm(formula = class ~ Age + Gender + Polyuria + Polydipsia + `sudden weight loss` + Polyphagia +
                `Genital thrush` + Itching + Irritability + `partial paresis` +
                Polyuria:`sudden weight loss`,
              family = binomial(link = logit), data = df)

# Age:`visual blurring`
mod.fit3<-glm(formula = class ~ Age + Gender + Polyuria + Polydipsia + Polyphagia + `Genital thrush` +
                `visual blurring` + Itching + Irritability + `partial paresis` +
                Age:`visual blurring`,
              family = binomial(link = logit), data = df)

#Polyuria:Polydipsia
mod.fit4<-glm(formula = class ~ Age + Gender + Polyuria + Polydipsia + Polyphagia + `Genital thrush` +
                  Itching + Irritability + `partial paresis` + Polyuria:Polydipsia,
              family = binomial(link = logit), data = df)

#Polyuria:Gender
mod.fit5<-glm(formula = class ~ Age + Gender + Polyuria + Polydipsia + Polyphagia + `Genital thrush` +
                  Itching + Irritability + `partial paresis` + Polyuria:Gender,
              family = binomial(link = logit), data = df)


# `partial paresis`:`muscle stiffness`
mod.fit6<-glm(formula = class ~ Age + Gender + Polyuria + Polydipsia + Polyphagia + `Genital thrush` +
                  Itching + Irritability + `partial paresis` + `muscle stiffness` +
                `partial paresis`:`muscle stiffness`,
              family = binomial(link = logit), data = df)

# gender:`sudden weight loss`
mod.fit7<-glm(formula = class ~ Age + Gender + Polyuria + Polydipsia + `sudden weight loss` +
                Polyphagia + `Genital thrush` + Itching + Irritability + `partial paresis` +
                Gender:`sudden weight loss`,
              family = binomial(link = logit), data = df)

# gender:`Genital thrush`
mod.fit8<-glm(formula = class ~ Age + Gender + Polyuria + Polydipsia + `sudden weight loss` +
                Polyphagia + `Genital thrush` + Itching + Irritability + `partial paresis` +
                Gender:`sudden weight loss` + Gender:`Genital thrush`,
              family = binomial(link = logit), data = df)

inter <- c("Polyuria:Age", "Polyuria:sudden weight loss", "Age:visual blurring",
           "Polyuria:Polydipsia", "Polyuria:Gender", "partial paresis:muscle stiffness",
           "gender:sudden weight loss", "Gender:Genital thrush")

AIC.vec <- c(AIC(mod.fit1), AIC(mod.fit2), AIC(mod.fit3), AIC(mod.fit4), AIC(mod.fit5), AIC(mod.fit6),
             AIC(mod.fit7), AIC(mod.fit8))
all.AIC1 <- data.frame(inter = inter, AIC.vec)
all.AIC1[order(all.AIC1[,2]), ]
```

```
##                             inter  AIC.vec
## 1                     Polyuria:Age 187.7286
## 3              Age:visual blurring 194.6154
## 8            Gender:Genital thrush 196.0349
## 7        gender:sudden weight loss 196.9603
## 6 partial paresis:muscle stiffness 197.1745
## 4              Polyuria:Polydipsia 200.6595
## 5                  Polyuria:Gender 201.4249
## 2      Polyuria:sudden weight loss 201.7522
```

It is evident from the AIC values that only the “Polyuria:Age” interaction improves the model (AIC = 187.7). Therefore, “Polyuria:Age” interaction is considered into the model. Next, other two-way interactions have been added to the improved model sequentially and respective AIC values have been checked.

Adding “Age:Visual Blurring” interaction to the model:

```
#with Polyuria:Age + Age:visual blurring
mod.fit_reduced<-glm(formula = class ~ Age + Gender + Polyuria + Polydipsia +
                       Polyphagia + `Genital thrush`+ Itching + Irritability +
                       `visual blurring` + `partial paresis` + Age:`visual blurring`+
                       Polyuria:Age,
                     family = binomial(link = logit), data = df)
AIC(mod.fit_reduced)
```

```
## [1] 185.29
```

It is worth noting that the AIC value has been reduced to 185.29. This indicates that the model is improved by adding the “Age:Visual Blurring” interaction to the model. Next the AIC value is tested for the improved model (with “Age:Polyuria” and “Age:Visual Blurring”) by adding the next two-way interaction “Gender:Genital Thrush”.

```
#with Polyuria:Age + Age:visual blurring + Gender:Genital thrush
mod.fit_reduced<-glm(formula = class ~ Age + Gender + Polyuria + Polydipsia +
                       Polyphagia + `Genital thrush`+ Itching + Irritability +
                       `partial paresis` + `visual blurring` +
                       Gender:`Genital thrush`+ Polyuria:Age + Age:`visual blurring`,
                     family = binomial(link = logit), data = df)
AIC(mod.fit_reduced)
```

```
## [1] 185.4705
```

Since there is no improvement in the AIC value (AIC = 185.47), it can be concluded that the two-way interaction of “Gender:Genital Thrush” is not important to the model.

Subsequently, the two-way interaction of “Gender:Sudden Weight Loss” is considered next as can be seen below:

```
#with Polyuria:Age + Age:visual blurring + gender:sudden weight loss
mod.fit_reduced<-glm(formula = class ~ Age + Gender + Polyuria + Polydipsia +
                       Polyphagia + `Genital thrush`+ Itching + Irritability +
                       `partial paresis` + `visual blurring` + `sudden weight loss`+
                       Polyuria:Age + Age:`visual blurring`+ Gender:`sudden weight loss`,
                     family = binomial(link = logit), data = df)

AIC(mod.fit_reduced)
```

```
## [1] 182.3148
```

It is evident that the AIC value has been reduced to 182.31 with the incorporation of “Gender:Sudden Weight Loss” to the model. Similarly, all other two -way interactions have been sequentially added to the improved model and the AIC values have been checked. However, there were not significant improvements to the AIC values, hence non of other two-way interactions have been included into the model.

LRT Testing for Improved Model

In order to test the statistical significance of the above selected features (main effects and two-way interactions), the LRT test is performed as below:

```
#LRT testing
Anova(mod.fit_reduced)
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: class
##                              LR Chisq Df Pr(>Chisq)
## Age                             8.582  1  0.0033953 **
## Gender                         74.161  1  < 2.2e-16 ***
## Polyuria                       76.516  1  < 2.2e-16 ***
## Polydipsia                     72.840  1  < 2.2e-16 ***
## Polyphagia                      6.487  1  0.0108647 *
## `Genital thrush`                7.120  1  0.0076247 **
## Itching                        33.473  1  7.227e-09 ***
## Irritability                   27.410  1  1.646e-07 ***
## `partial paresis`               3.229  1  0.0723494 .
## `visual blurring`               3.206  1  0.0733732 .
## `sudden weight loss`            0.324  1  0.5689960
## Age:Polyuria                   11.270  1  0.0007879 ***
## Age:`visual blurring`           4.543  1  0.0330443 *
## Gender:`sudden weight loss`     6.651  1  0.0099104 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

It is evident that the P-values are greater than 0.05 for three attributes: “Partial Paresis”, “Visual Blurring”, and “Sudden Weight Loss”. Out of these three attributes, “Partial Paresis” can be dropped from the model as it is not statistically significant as the P-value (>0.05) is concerned. However, other two attributes can not be dropped due the the significance of their associated two-way interactions.

Finally, the AIC value and LRT test is obtained for final estimated model with selected main effects & two-way interaction as below:

```
#after removing partial paresis
mod.fit_reduced<-glm(formula = class ~ Age + Gender + Polyuria + Polydipsia +
                       `sudden weight loss` + Polyphagia + `Genital thrush`+ `visual blurring` +
                       Itching + Irritability + Polyuria:Age + Age:`visual blurring`+ Gender:`sudden weight loss`,
                     family = binomial(link = logit), data = df)

AIC(mod.fit_reduced)
```

```
## [1] 183.5436
```

```
#LRT testing
Anova(mod.fit_reduced)
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: class
##                              LR Chisq Df Pr(>Chisq)
## Age                             7.176  1  0.0073889 **
## Gender                         76.672  1  < 2.2e-16 ***
## Polyuria                       86.789  1  < 2.2e-16 ***
## Polydipsia                     77.663  1  < 2.2e-16 ***
## `sudden weight loss`            0.331  1  0.5651726
## Polyphagia                      7.120  1  0.0076242 **
## `Genital thrush`                6.077  1  0.0136937 *
## `visual blurring`               6.697  1  0.0096558 **
## Itching                        39.067  1  4.095e-10 ***
## Irritability                   27.181  1  1.853e-07 ***
## Age:Polyuria                   11.100  1  0.0008632 ***
## Age:`visual blurring`           8.262  1  0.0040473 **
## Gender:`sudden weight loss`     6.771  1  0.0092660 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Estimated Model

The coefficients of the estimated logistic regression model is obtained from,

```
round(mod.fit_reduced$coefficients,3)
```

```
##                       (Intercept)                               Age
##                            -1.586                             0.047
##                        GenderMale                       PolyuriaYes
##                            -3.930                            13.858
##                     PolydipsiaYes           `sudden weight loss`Yes
##                             5.296                             2.448
##                     PolyphagiaYes              `Genital thrush`Yes
##                             1.284                             1.410
##             `visual blurring`Yes                         ItchingYes
##                             8.837                            -3.562
##                 IrritabilityYes                  Age:PolyuriaYes
##                             2.891                            -0.167
##          Age:`visual blurring`Yes GenderMale:`sudden weight loss`Yes
##                            -0.147                            -3.156
```

And the estimated logistic regression model to predict the likelihood of having diabetes using common signs and symptoms presented by patients is given by,

$$logit(\hat{\pi}) = -1.586 + 0.047 * Age - 3.93 * Male + 13.858 * Polyuria + 5.296 * Polydipsia + 2.448 * SuddenWeightLoss + 1.284 * Polyphagia + 1.410 * GenitalThrush + 8.837 * VisualBlurring - 3.562 * Itching + 2.891 * Irritability - 0.167 * Polyuria * Age - 0.147 * Age * VisualBlurring - 3.156 * Male * SuddenWeightLoss$$

It is worth noting that the sign “Polyuria” has comparatively large positive coefficient, which suggest it has a dominance contribution to the prediction model. Also, the negative coefficient for “Age:Polyuria” suggests the positive contributions of “Polyuria” sign for diabetes prediction declines with the age.

2.2. Residual Analysis

The Standardized Pearson residual analysis has been performed for the independent numerical variable of “Age”. Firstly, the data set is transformed in to explanatory variable pattern (EVP) form with respect to the “Age” attribute as shown below:

```
# dummy encoding
df1 <- df
df1[2] <- ifelse(df1[2] == "Female", 1,0)
df1[3:16] <- ifelse(df1[3:16] == "Yes", 1,0)
df1[17] <- ifelse(df1[17] == "Positive", 1,0)


# Convert data to EVP form
w <- aggregate(formula = class ~ Age + Gender + Polyuria + Polydipsia  + `sudden weight loss` +
                 Polyphagia + `Genital thrush` + `visual blurring` + Itching + Irritability ,
               data = df1, FUN = sum)

n <- aggregate(formula = class ~ Age + Gender + Polyuria + Polydipsia  + `sudden weight loss` +
                 Polyphagia + `Genital thrush` + `visual blurring` + Itching + Irritability,
               data = df1, FUN = length)

w.n <- data.frame(w, trials = n$class, prop = round(w$class/n$class, 4))

#First 5 rows of the data set -EVP form
kbl(head(w.n,5), caption = "Table 3: Random 5 rows from data set-EVP form") %>%
  kable_styling(bootstrap_options = c("striped", "hover", "condensed"),
                full_width = F, position = "left",font_size = 10)
```

Table 3: Random 5 rows from data set-EVP form

Age	Gender	Polyuria	Polydipsia	sudden.weight.loss	Polyphagia	Genital.thrush	visual.blurring	Itching	Irritability	class	trials	prop
26	0	0	0	0	0	0	0	0	0	0	1	0
27	0	0	0	0	0	0	0	0	0	0	6	0
29	0	0	0	0	0	0	0	0	0	0	1	0
30	0	0	0	0	0	0	0	0	0	0	17	0
32	0	0	0	0	0	0	0	0	0	0	1	0

Five random rows of the EVP form data is shown in above table:

In order to verify that the EVP transformation has happened correctly, the total number of EVP entries and their respective number of observations have been cross checked against the original data set. Also, the logistic regression model is derived again for the EVP data as shown below:

```
nrow(w.n) # Number of EVPs (M)
```

```
## [1] 237
```

```
sum(w.n$trials) # Number of observations
```

```
## [1] 516
```

```
mod.fit.bin<-glm(formula = class/trials ~ Age + Gender + Polyuria + Polydipsia +
                      sudden.weight.loss + Polyphagia + Genital.thrush+ visual.blurring +
                      Itching + Irritability + Polyuria:Age + Age:visual.blurring+ Gender:sudden.weight.loss,
                 family = binomial(link = logit), data = w.n, weights = trials)
round(summary(mod.fit.bin)$coefficients, digits = 4)
```

```
##                            Estimate Std. Error z value Pr(>|z|)
## (Intercept)                 -5.5167     1.6386 -3.3666   0.0008
## Age                          0.0472     0.0359  1.3169   0.1879
## Gender                       3.9303     0.7203  5.4562   0.0000
## Polyuria                    13.8581     3.2366  4.2817   0.0000
## Polydipsia                   5.2959     0.8606  6.1539   0.0000
## sudden.weight.loss          -0.7081     0.6713 -1.0548   0.2915
## Polyphagia                   1.2841     0.4953  2.5924   0.0095
## Genital.thrush               1.4105     0.5863  2.4059   0.0161
## visual.blurring              8.8366     2.8403  3.1112   0.0019
## Itching                     -3.5617     0.7036 -5.0618   0.0000
## Irritability                 2.8908     0.6153  4.6980   0.0000
## Age:Polyuria                -0.1674     0.0562 -2.9780   0.0029
## Age:visual.blurring         -0.1468     0.0542 -2.7082   0.0068
## Gender:sudden.weight.loss    3.1559     1.3686  2.3059   0.0211
```

It is worth noting that the model coefficients obtained from EVP data is almost equal to the coefficients obtained from the original data set. This suggests that the EVP transformation is statistically equivalent to the original data set.

As shown below in the left figure, the Standardized Pearson residuals are plotted against the independent variable of “Age”. It is evident that the points are mostly randomly scattered with a very few points (only 7 observations) outside +/-3 lines. This indicates that the developed logistic regression model is performing well. The middle plot shows the Standardized Pearson residuals against the estimated probability of success. It is worth noticing that the large number of points are gathered close to 0 and 1 in the X axis and their respective residual values are very close to 0. This indicates that for real data, the developed model mostly (except for 7 observations) outputs ‘True’ (close to 1) or ‘False’ (close to 0) with fairly good accuracy. The in between points are also possess Residual values within +/-3 for real data. Finally, the third plot (right most) shows the Standardized Pearson residual against the Linear Predictor. It is evident that the residual values are almost zero for very-high and very-low Linear Predictor values. This indicates that the if symptoms are existing, the model would accurately predicts the likelihood of patience to have diabetes. However, for Linear Predictor values around zero, the Residual values tends to deviates from zero. Among them only 7 points are outsize +/-3 lines.

```
pi.hat <- predict(mod.fit.bin, type = "response")
p.res <- residuals(mod.fit.bin, type = "pearson")
s.res <- rstandard(mod.fit.bin, type = "pearson")
lin.pred <- mod.fit.bin$linear.predictors
w.n <- data.frame(w.n, pi.hat, p.res, s.res, lin.pred)


kbl(round(head(w.n), digits = 3), caption = "Table 4: First 5 rows of new data set with Pi.hat, p.res, s.res and lin pr
ed") %>%
  kable_styling(bootstrap_options = c("striped", "hover", "condensed"),
                full_width = F, position = "left",font_size = 10)
```

Table 4: First 5 rows of new data set with Pi.hat, p.res, s.res and lin pred

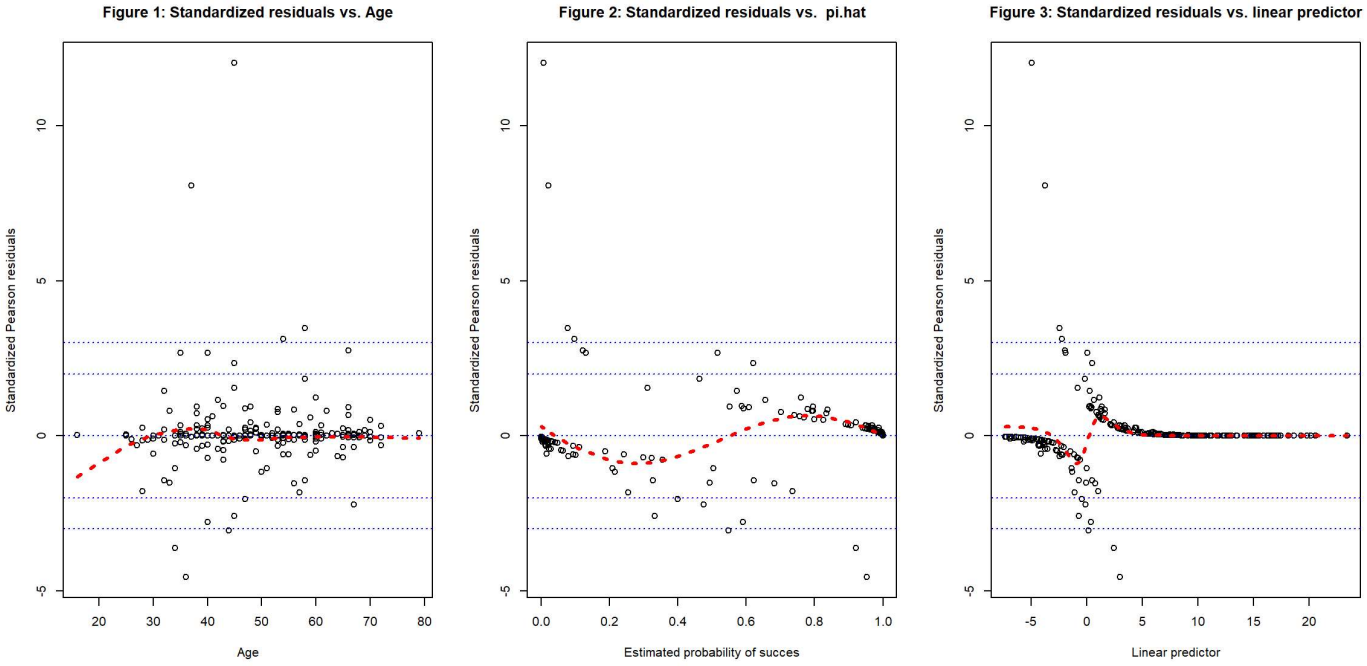
Age	Gender	Polyuria	Polydipsia	sudden.weight.loss	Polyphagia	Genital.thrush	visual.blurring	Itching	Irritability	class	trials	prop	pi.hat	p.res	s.res	lin.pred
26	0	0	0	0	0	0	0	0	0	0	1	0	0.014	-0.117	-0.118	-4.289
27	0	0	0	0	0	0	0	0	0	0	6	0	0.014	-0.294	-0.302	-4.241
29	0	0	0	0	0	0	0	0	0	0	1	0	0.016	-0.126	-0.126	-4.147

Age	Gender	Polyuria	Polydipsia	sudden.weight.loss	Polyphagia	Genital.thrush	visual.blurring	Itching	Irritability	class	trials	prop	pi.hat	p.res	s.res	lin.pred
30	0	0	0	0	0	0	0	0	0	0	17	0	0.016	-0.531	-0.574	-4.100
32	0	0	0	0	0	0	0	0	0	0	1	0	0.018	-0.135	-0.136	-4.005
34	0	0	0	0	0	0	0	0	0	0	3	0	0.020	-0.245	-0.248	-3.911

```
par(mfrow = c(1,3))
# Standardized Pearson residual vs Age plot
plot(x = w.n$Age, y = w.n$s.res, xlab = "Age",ylab = "Standardized Pearson residuals",
     main = "Figure 1: Standardized residuals vs. Age")
abline(h = c(3, 2, 0, -2, -3), lty = 3, col = "blue")
# Add loess model to help visualize tren
smooth.stand <- loess(formula = s.res ~ Age, data = w.n, weights = trials)
order.age <- order(w.n$Age)
lines(x = w.n$Age[order.age], y = predict(smooth.stand)[order.age], lty = 3, col = "red", lwd = 3)

# Standardized Pearson residual vs pi plot
plot(x = w.n$pi.hat, y = w.n$s.res, xlab = "Estimated probability of succes",ylab = "Standardized Pearson residuals",
     main = "Figure 2: Standardized residuals vs.  pi.hat")
abline(h = c(3, 2, 0, -2, -3), lty = 3, col = "blue")
smooth.stand <- loess(formula = s.res ~ pi.hat, data = w.n, weights = trials)
order.pi.hat <- order(w.n$pi.hat)
lines(x = w.n$pi.hat[order.pi.hat], y = predict(smooth.stand)[order.pi.hat], lty = 3, col = "red", lwd = 3)

# Standardized Pearson residual vs linear predictor plot
plot(x = w.n$lin.pred, y = w.n$s.res, xlab = "Linear predictor",ylab = "Standardized Pearson residuals",
     main = "Figure 3: Standardized residuals vs. linear predictor")
abline(h = c(3, 2, 0, -2, -3), lty = 3, col = "blue")
smooth.stand <- loess(formula = s.res ~ pi.hat, data = w.n, weights = trials)
order.lin.pred <- order(w.n$lin.pred)
lines(x = w.n$lin.pred[order.lin.pred], y = predict(smooth.stand)[order.lin.pred], lty = 3, col = "red", lwd = 3)
```



Below table shows all 7 EVP groups that contribute for the Standardized Pearson Residual points outsize +/-3 lines. It is worth noting that each of these groups have very small number of observations, therefore, the model results would not be necessarily aligned with the real observation. These points could be safely ignored despite they are outside +/-3 lines.

Table 5: The residual outside of the ±3 lines.

Age	Gender	Polyuria	Polydipsia	sudden.weight.loss	Polyphagia	Genital.thrush	visual.blurring	Itching	Irritability	class	trials	prop	pi.hat	p.res	s.res	lin.pred
37	0	0	0	0	0	0	0	0	0	3	6	0.5	0.0226	7.8764	8.0678	-3.7690
34	1	0	0	1	0	0	0	0	0	0	1	0.0	0.9218	-3.4339	-3.6263	2.4674
58	0	0	0	0	0	0	1	0	0	1	1	1.0	0.0791	3.4125	3.4710	-2.4549
44	0	1	0	1	0	1	0	1	0	0	4	0.0	0.5481	-2.2026	-3.0481	0.1930
45	0	0	0	1	1	1	0	1	0	1	1	1.0	0.0069	11.9790	12.0245	-4.9663

Age	Gender	Polyuria	Polydipsia	sudden.weight.loss	Polyphagia	Genital.thrush	visual.blurring	Itching	Irritability	class	trials	prop	pi.hat	p.res	s.res	lin.pred
36	1	0	0	0	0	0	0	0	1	0	1	0.0	0.9528	-4.4926	-4.5614	3.0049
54	0	0	0	0	0	1	0	1	1	1	1	1.0	0.0974	3.0441	3.1318	-2.2264

2.3. Response Analysis

Next, lets look how the response variable (i.e. likelihood of having diabetes) behaves against a independent variable. The symptom “Polyuria” has been selected as the independent variable in this analysis due to the fact that it has the largest model coefficient. Using below script, the real observation and model output has been visually represented,

```
pi.hat1 <- predict(mod.fit_reduced, type = "response")
df1 <- data.frame(df, pi.hat1)

p1<- ggplot(df1, aes(x= class,  group=Polyuria)) +
    geom_bar(aes(y = ..prop.., fill = factor(..x..)), stat="count", alpha=0.5,color="dark blue", width = 0.5, show.lege
nd = TRUE) +
    geom_text(aes( label = scales::percent(..prop..),
                   y= ..prop.. ), stat= "count", vjust = 1.5, colour="black",size=3) +
    labs(y = "Percent", fill="Class",title="Figure 4: Polyuria by Class") +
    facet_grid(~Polyuria) +
    scale_y_continuous(labels = scales::percent)+
  scale_fill_discrete(name="Class",labels=c("Negative", "Positive"))+
  theme(plot.title = element_text(size = 14),axis.title.x = element_blank())

p2<-ggplot(df1, aes(x=Polyuria, y=pi.hat1,color=Polyuria)) +
  geom_boxplot() +
  stat_summary(fun.y=mean, geom="point", shape=20, size=6, color="red", fill="red")+
  labs(y = "Estimated Probability", title="Figure 5: Estimated probability by Polyuria")

grid.arrange(p1, p2, ncol=2, widths=c(2.6, 2.6))
```

Figure 4: Polyuria by Class

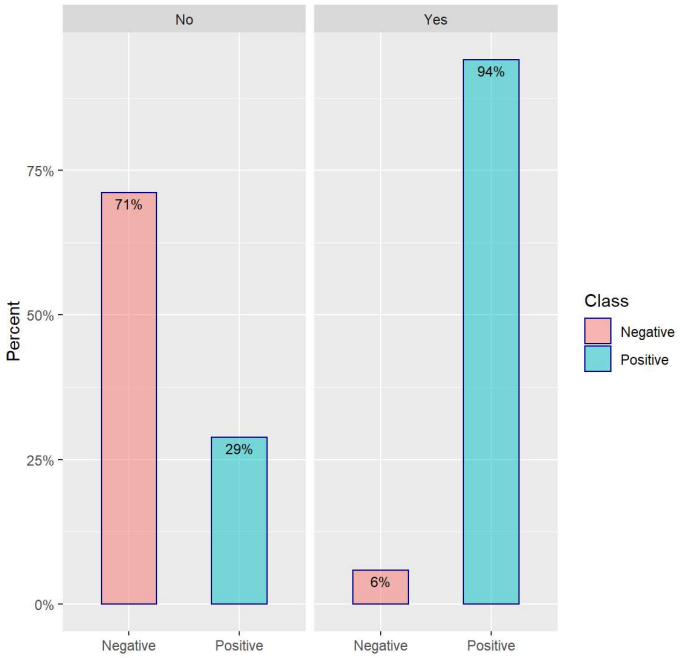
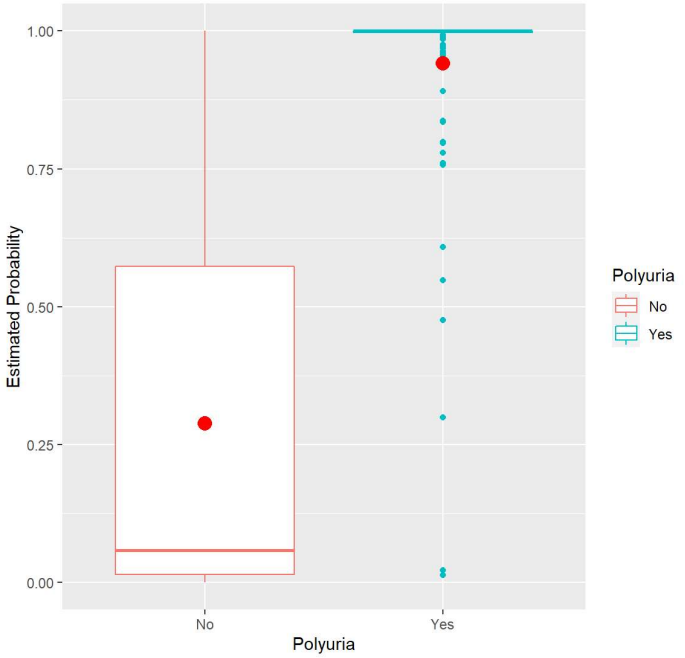


Figure 5: Estimated probability by Polyuria



As can be seen from the above left figure, according to the real observations, 94% of population with “Polyuria” symptom are diagnosed with diabetes, while only 29% of population without “Polyuria” symptom have been diagnosed with diabetes. The plot in the right figure shows model output as the box plots of probability of having diabetes against the “Polyuria” symptom. It is evident from the right figure that the model output of the probability of having diabetes when “Polyuria=Yes”, shows a mean value (the Red dot) of ~90%, while probability of having diabetes when “Polyuria=No” shows a mean value (Red dot) ~29%. This indicates that the model output for probability of having diabetes as a function of “Polyuria” closely aligns with real observation from original data set.

2.4. Goodness of Fit

In order to evaluate how well the model fits for all the observations at once, the Goodness of Fit (GOF) analysis is performed. The “Residual Deviance/df” value is calculated for the model using the below script:

```
# deviance/DF
rdev <- mod.fit.bin$deviance
dfr <- mod.fit.bin$df.residual
ddf <- rdev/dfr
thresh2 <- 1 + 2*sqrt(2/dfr) # potential problem
thresh3 <- 1 + 3*sqrt(2/dfr) # poor fit
round(c(rdev, dfr, ddf, thresh2, thresh3),3)
```

```
## [1] 147.226 223.000   0.660   1.189   1.284
```

As can be seen above, the “Residual Deviance/df” value for the model is 0.673 and it is smaller than the potential fit ratio (1.192) as well as the poor fit ratio (1.287). This indicates that the GOF statistics of this particular model is within the satisfactory limit, hence the developed model is a good fit.

2.5. Confidence Intervals

Primarily, the data set comprises of two populations: Male and Female. It is worth checking whether there is a significant different between males and females in the data set to be a diabetics patient. If there is no significant difference, the attribute “Gender” can be dropped from the model development. The Wald’s confidence interval has been checked to determine this as shown in below steps:

First, the Two-Way Contingency and respective Probability tables have been obtained for female and male populations,

```
c.table <- table(df$Gender,df$class)
c.table
```

```
##
##          Negative Positive
##   Female       19      171
##   Male        181      145
```

```
pi.hat.table<-c.table/rowSums(c.table)
pi.hat.table
```

```
##
##           Negative  Positive
##   Female 0.1000000 0.9000000
##   Male   0.5552147 0.4447853
```

Then, the Wald’s confidence intervals for the probability difference of two populations (male and female) has been calculated as below:

```
#probability of having diabetes for each group:
alpha<-0.05
pi.hat1<-pi.hat.table[1,2]
pi.hat2<-pi.hat.table[2,2]
#Wald
var.wald<-pi.hat1*(1-pi.hat1) / sum(c.table[1,]) + pi.hat2*(1-pi.hat2) / sum(c.table[2,])
round(pi.hat1 - pi.hat2 + qnorm(p = c(alpha/2, 1-alpha/2)) * sqrt(var.wald),3)
```

```
## [1] 0.386 0.524
```

According to the results, the 95% Wald confidence interval is:

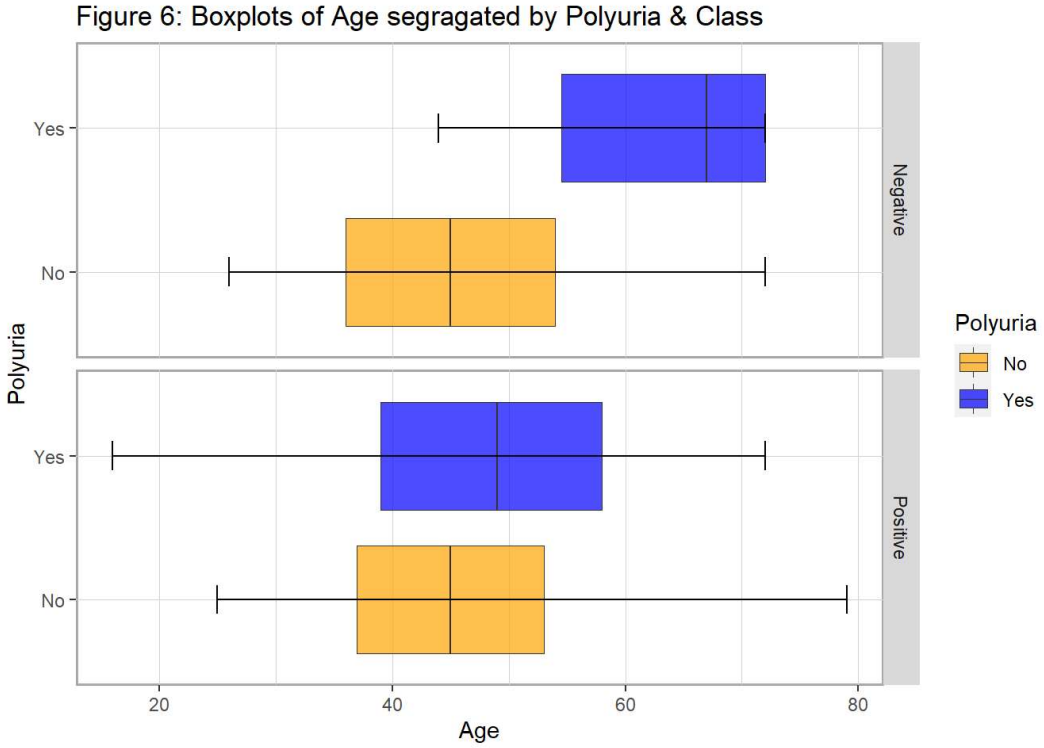
$$0.386 < (\hat{\pi}_1 - \hat{\pi}_2) < 0.524$$

Since this interval does not contain zero, there is a sufficient evidence to indicate a significant difference between male and female populations to the target attribute. Hence “Gender” attribute need to be considered in the model.

2.6. Hypothesis Tests

Under this section the importance Age & Polyuria interaction is analysed using the hypothesis testing method. Below the age distribution of Polyuria symptom segregated by ‘Class’ (i.e. diabetes positive or negative) is shown.

```
bp <- ggplot(data=df, aes(x=Age, y=Polyuria, group=Polyuria)) +
  geom_boxplot(aes(fill=Polyuria), alpha=0.7,outlier.shape=NA,lwd=0.2)
bp + facet_grid(df$class ~.)+ stat_boxplot(geom = 'errorbar', width = 0.2,coef = 3)+
  theme(
  panel.background = element_rect(fill = "white",colour = "dark gray",
                                  size = 1, linetype = "solid"),
  panel.grid.major = element_line(size = 0.2, linetype = 'solid',
                                  colour = "light gray"),
  panel.grid.minor = element_line(size = 0.1, linetype = 'solid',
                                  colour = "light gray"))+
  scale_fill_manual(name = "Polyuria", values = c("orange", "blue"))+
  labs(title="Figure 6: Boxplots of Age segragated by Polyuria & Class") +
  theme(plot.title = element_text(size = 13,colour = "black"))
```



It is obvious from the diabetes negative plot (top) that Polyuria symptoms are presented in older population; the age distributions of Polyuria “yes’ and “no” show a clear separation of age (mean age of Polyuria ‘no’ is 45 years, while mean age of Polyuria “yes” is about 78 years). This suggests that Polyuria is an age-related sign in general community. However, this age separation between Polyuria “yes’ and “no” populations are not prominent within diabetes positive population as shown in second plot. This supports someone to believe Polyuria is a diabetes related symptom at the first sight.

In order to evaluate the importance of the interaction of Polyuria and Age for the model, the LRT test is carried out for the parameters that correspond to Age:polyuria - β_{11}.

The corresponding hypotheses are:

$$H_0 : \beta_{11} = 0$$
$$H_a : \beta_{11} \neq 0$$

```
Anova(mod.fit_reduced)
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: class
##                              LR Chisq Df Pr(>Chisq)
## Age                             7.176  1  0.0073889 **
## Gender                         76.672  1  < 2.2e-16 ***
## Polyuria                       86.789  1  < 2.2e-16 ***
## Polydipsia                     77.663  1  < 2.2e-16 ***
## `sudden weight loss`            0.331  1  0.5651726
## Polyphagia                      7.120  1  0.0076242 **
## `Genital thrush`                6.077  1  0.0136937 *
## `visual blurring`               6.697  1  0.0096558 **
## Itching                        39.067  1  4.095e-10 ***
## Irritability                   27.181  1  1.853e-07 ***
## Age:Polyuria                   11.100  1  0.0008632 ***
## Age:`visual blurring`           8.262  1  0.0040473 **
## Gender:`sudden weight loss`     6.771  1  0.0092660 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

It is shows from the results that the test statistic is -2log(Λ) = 11.10, and the p-value is 0.0008 using a X_2^2 approximation. This indicates that the “Polyuria:Age” interaction is statistically significant for the model, thus the null hypothesis can be rejected safely.

2.7. Sensitivity Analysis

The sensitivity analysis has been performed to check the effect of polyuria symptom on the probability of diagnosed with diabetes depends on the age. The Odds Ratio and Confidence Intervals for Polyuria and its interactions are being considered here. The odds ratio for polyuria comparing polyuria (1) vs non-polyuria (0) holding Age constant is,

$$\hat{Odds Ratio}_{polyuria} = e^{\beta_3 + \beta_{11} * Age}$$

where β_3 is coefficient for polyuria and β_{11} coefficient for polyuria:age interaction. Using below code, the Odd Ratios are obtained for ages from 10 to 80 at 5 year intervals.

```
beta.hat<-mod.fit_reduced$coefficients[2:13]
age<-seq(from = 10, to = 80, by = 5)
OR.polyuria<- exp((beta.hat[3] + beta.hat[11]*age))

cov.mat<-vcov(mod.fit_reduced)[2:13,2:13]
#Var(beta^_4 + age*beta^_11)
var.log.OR<-cov.mat[3,3] + age^2*cov.mat[11,11] + 2*age*cov.mat[3,11]

ci.log.OR.low<-(beta.hat[3] + beta.hat[11]*age) - qnorm(p = 0.975)*sqrt(var.log.OR)
ci.log.OR.up<-(beta.hat[3] + beta.hat[11]*age) + qnorm(p = 0.975)*sqrt(var.log.OR)
OR.low <- exp(ci.log.OR.low)
OR.up <- exp(ci.log.OR.up)
round(data.frame(age = age, OR.hat = OR.polyuria , OR.low , OR.up ),2)
```

```
##    age    OR.hat  OR.low       OR.up
## 1   10 195572.91 1000.45 38231523.34
## 2   15  84666.61  734.34  9761680.81
## 3   20  36653.52  535.86  2507146.11
## 4   25  15867.89  387.77   649327.89
## 5   30   6869.46  277.16   170263.44
## 6   35   2973.90  194.33    45511.14
## 7   40   1287.45  132.04    12553.48
## 8   45    557.36   84.98     3655.40
## 9   50    241.29   49.72     1171.07
## 10  55    104.46   24.86      438.91
## 11  60     45.22   10.15      201.53
## 12  65     19.58    3.45      111.06
## 13  70      8.48    1.04       69.04
## 14  75      3.67    0.29       45.99
## 15  80      1.59    0.08       31.87
```

As can be seen from the OR.hat values in above table, the odds of having diabetics change by 195573 times for the people having polyuria when the age is fixed at a value of 10. However, the odds of having diabetics dramatically changes to 1.59 times for the people having polyuria when the age is fixed at a value of 80. This suggests, the “Ployuria” is very sensitive indicator within the young population to determine whether they have diabetes or not. However, among the older population, the “Polyuria” symptom is not comparatively so sensitive indicator. In other words, according to the analysis, diabetes is more susceptible for “Polyuria” symptoms in younger population compared to

that of older population. Also, through the confident interval shows above, it can be expressed with 95% confidence, the odds of detecting diabetes for 50 years old patients is between 49.72 to 1171.07 times as large for patients with polyuria symptoms (polyuria = 1) than for non-polyuria patients (polyuria = 0). Similarly, for the other age groups the odd of detecting diabetes can be estimated using the values in the table.

3. Critique & Limitations

It is observed that the female population withing the data set is surprisingly low (37%), while their diabetes positive rates are significantly high (90%). This shows a gender biasness of the data set. This could be due to a anomaly in the data set during the recording or could be due to social & cultural issue within the respective community. It will be interesting to re-build the model for similar data sets captured for different socio-economic group.

4. Summary & Conclusions

The objective of this work is to build a logistic regression model to predict the likelihood of having diabetes using common signs and symptoms presented by patients. The initial data exploration indicates that there are number of noticeable relationships between sign/symptoms and having diabetes. A logistic regression model is formulated and was further improved using feature selection techniques and incorporating 2-way interactions of the attributes. The Standardized Pearson residual analysis and Goodness of Fit analysis confirmed that the improved model performs and fits reasonably well. The Response Analysis done for “Polyuria” confirms that the model output is reasonably close to the actual observation. It was found from Confident Interval analysis that the male and female populations in the data set are statistically different, hence kept the Gender as a feature in the model. Using the Hypothesis Testing, the statistical importance of Age:Polyuria interaction was confirmed. The odd ratio of having diabetes for “Polyuria” presented patients is investigate across the “Age” groups. It is evident that the “Ployuria” is very sensitive indicator within the younger population to determine whether they have diabetes. However, among the older population, the “Polyuria” symptom is not comparatively so sensitive indicator.

Surprisingly, the proportion of diabetes positive females in the data set is significant high (90%) compared to that of male (44%), despite the fact the female patients in the data set is noticeably low (37%) compared to male (67%). It will be interesting to conduct a study to investigate the reason behind this. Could this be due to females in Bangladesh are less likely to visit hospitals compared to males or could females be tolerating illnesses more compared to males. Such analysis is out of the scope of this study, therefore did not carry out further analysis on those lines in this study.

In this model, only a few selected 2-way interactions have been considered. It is quite possible that the model could be further improved by considering all the 2-way and higher-order interactions.

5. References

Aksakalli, V., Yenice, Z., Wong, Y. K., Ture, I., & Malekipirbazari, M. (2020). *Www.featureranking.com*. https://www.featureranking.com/ (https://www.featureranking.com/)

ASPE. (2017). *The importance of early diabetes detection*. https://aspe.hhs.gov/report/diabetes-national-plan-action/importance-early-diabetes-detection#:%7E:text=Early%20detection%20and%20treatment%20of,limb%20amputations%2C%20and%20kidney%20failure (https://aspe.hhs.gov/report/diabetes-national-plan-action/importance-early-diabetes-detection#:%7E:text=Early%20detection%20and%20treatment%20of,limb%20amputations%2C%20and%20kidney%20failure)

Dua, D., & Graff, C. (2017). *UCI machine learning repository*. University of California, Irvine, School of Information; Computer Sciences. http://archive.ics.uci.edu/ml (http://archive.ics.uci.edu/ml)

Faniqul, M. M. I., Ferdousi, R., Rahman, S., & Bushra, H. Y. (2019). Likelihood prediction of diabetes at early stage using data mining techniques. *Computer Vision and Machine Intelligence in Medical Image Analysis*, 113–125. https://doi.org/10.1007/978-981-13-8798-2_12 (https://doi.org/10.1007/978-981-13-8798-2_12)

kassambara. (2017). *Plot one variable: Frequency graph, density distribution and more*. http://www.sthda.com/english/articles/32-r-graphics-essentials/133-plot-one-variable-frequency-graph-density-distribution-and-more/#one-categorical-variable (http://www.sthda.com/english/articles/32-r-graphics-essentials/133-plot-one-variable-frequency-graph-density-distribution-and-more/#one-categorical-variable)

Sauer, S. (2016). *How to plot a 'percentage plot' with ggplot2*. https://sebastiansauer.github.io/percentage_plot_ggplot2_V2/ (https://sebastiansauer.github.io/percentage_plot_ggplot2_V2/)

Schneider, W. J. (2017). *Introduction to rmarkdown*. https://my.ilstu.edu/~wjschne/442/IntroductiontoRMarkdown.html (https://my.ilstu.edu/~wjschne/442/IntroductiontoRMarkdown.html)

Xie, Y. (2020). *R markdown cookbook*. https://bookdown.org/yihui/rmarkdown-cookbook/bibliography.html (https://bookdown.org/yihui/rmarkdown-cookbook/bibliography.html)

Zhu, H. (2020). *Create awesome html table with knitr::Kable and kableExtra*. https://cran.r-project.org/web/packages/kableExtra/vignettes/awesome_table_in_html.html (https://cran.r-project.org/web/packages/kableExtra/vignettes/awesome_table_in_html.html)