Project Report

Topic 1

Background

Diabetes mellitus (DM), many times recognized as diabetes, is a group of metabolic disorders, whose most common symptom is blood sugar level stays abnormally high for a long period of time. Such kind of disease has been increasingly noticeable in the past few decades, and it is usually evidently related to unhealthy diets, like smoking or amount of drinking, and some body measures such as BMI, blood pressure, etc.

Objective

The goal of this project is to investigate which features are related to diabetes and how significant the relation is. In addition, a logistic regression is built to predict the probability of having diabetes given a set of data. Finally, the model will be tested to see how well it fits the data.

The selected features are weight, height, BMI, having high blood pressure, having high cholesterol level, and alcohol taking.

Methods

1. Data Processing

The features we needed are in different datasets, so the first step is to extract those features. However, problems may occur if only the feature columns are exacted, because once we do that, the information, or identity, of the owner of that particular value is lost.

Therefore, what is done is that tuples that contain both the feature columns and the identity column, which is the sequence number in this scenario, should be extracted first. After that, all the dataframes are merged to form a matrix using nature join, a concept in relational algebra, which uses the common column as a key and join 2 tables together.

One remark is that what are respected in the features that are discrete are zeros and ones. However, referring to the encoding methods, in some columns, numbers 1 to 6 are used to indicate different statuses, only part of which are of interest. Therefore, extra work is done to convert multi-status data into binary ones, during which missing values (NAs) are excluded as well.

2. Multiple logistic regression

Multiple logistic regression is applied first to identify which characteristics are highly related or are not related to diabetes. After the significant features are found, apply multiple logistic regression again to fit another model.

3. BMI and Diabetes

Attention is paid to seek the relation between BMI and diabetes. Therefore, single logistic regression is applied.

4. Relative Ratios

Relative Ratios is calculated for every noticeable binomial feature in the above tests.

5. ROC - AUC

ROC curve is plotted for original multiple regression model, improved multiple regression model and “BMI Diabetes” model. As the higher the AUC is, the better it is when working as a predictor. If AUC is close to 0.5, it will be considered as a bad predictor.

6. Observations from Figures

By plotting histogram and distribution of BMI and weights of those who had diabetes, some observations can be made.

Results

1. Data Processing

After the processing mentioned, a merged matrix is obtained. The first column is the identity number of that particular tuple. And having high blood pressure and having high cholesterol level are encoded as zero and one correctly by observing the output of the "summary(M)".

2. Multiple logistic regression

Hypothesis-1:

H0: All features are not related to diabetes.

H1: At least one feature is related to diabetes.

The result is as follows.

```
Call:
glm(formula = DIQ010 ~ BMXHT + Overweight + BPQ02
0 + BPQ080 +
    ALQ101 + ALQ130, family = "binomial", data =
M)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.0991  -0.4743  -0.3127  -0.2403   2.7588

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

(Intercept) -3.1484868  1.0592668  -2.972  0.0029
6 **
BMXHT       -0.0007778  0.0060955  -0.128  0.89846

Overweight   0.7222811  0.1189397   6.073 1.26e-0
9 ***
BPQ020       1.3506402  0.1262465  10.698  < 2e-16
 ***
BPQ080       1.1338416  0.1233931   9.189  < 2e-16
***
ALQ101      -0.0688969  0.1391210  -0.495  0.62044

ALQ130      -0.0607901  0.0283760  -2.142  0.03217
 *
```

Therefore, we reject H0 for there are features show statistical significance.

We can see that only Overweight, BPQ020(high blood pressure), BPQ080(high cholesterol level) and ALQ130(average number of alcohol taken in the past year) are statistically significantly.

Therefore, the model is improved by excluding the features of low statistical significance, another model is obtained. However, AIC that is obtained in the original model is 1983.2 while it is 1979.4 this time, where AIC is an indicator of the wellness of the model and the less it is, the better the model is. The difference is marginal, meaning that by excluding the unrelated features, the performance is only improved slightly. This can be further verified when plotting the ROC-AUC.

3. BMI and Diabetes

Hypothesis-2:

H0: BMI is not related to diabetes.

H1: BMI is related to diabetes.

The result obtained from a single logistic regression model is as follows.

```
Call:
glm(formula = M$DIQ010 ~ M$BMXBMI, family = "bino
mial", data = M)
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.3266  -0.5131  -0.4366  -0.3782   2.4849
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -4.154895   0.226233  -18.37   <2e-16
***
M$BMXBMI     0.069645   0.006856   10.16   <2e-16
***
```

We can see that p-value that we obtained is very small, less than 0.05, therefore, H0 is rejected. Therefore, the conclusion is that BMI is highly related to diabetes.

The plot of the model is shown below, where the shadow area stands for the confidence level.

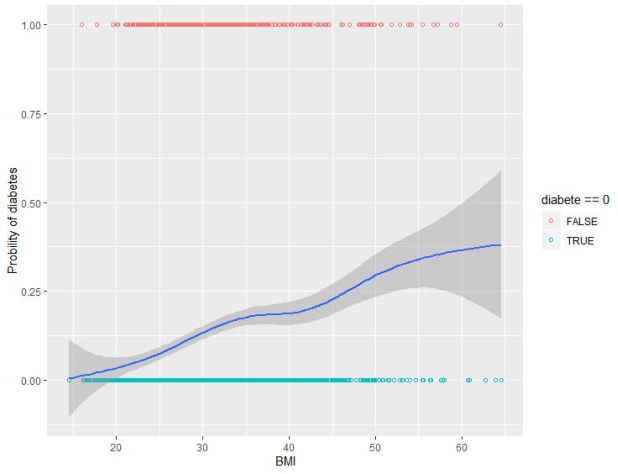

Figure 1. Predicted probability and diabetes situation (red: true positive; blue: true negative) using gam fitting

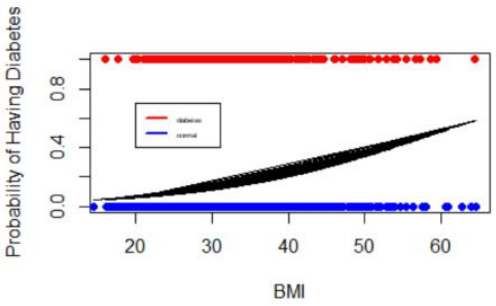

Figure 2. Predicted probability and diabetes situation (red: true positive; blue: true negative) using logistic regression model

We can see that there is a large part of data share the same BMI while they have different diabetes situations. Besides, both of the fitted cures, using gam fitting in R and logistic regression, do not give a good model, which means that BMI alone may not be a good predictor even though it has high statistical significance.

4. Relative Ratios

The relative ratios for diabetes in favor of overweight, high blood pressure and high cholesterol level.

Table 1. Relative ratios of diabetes in favor of overweight, high blood pressure and high cholesterol level respectively

Item	Relative Ratio
overweight	1.54
high blood pressure	2.37
high cholesterol level	2.05

Values obtains is greater than 1 quite obviously, verifying that they are relatively highly related to diabetes.

5. ROC - AUC

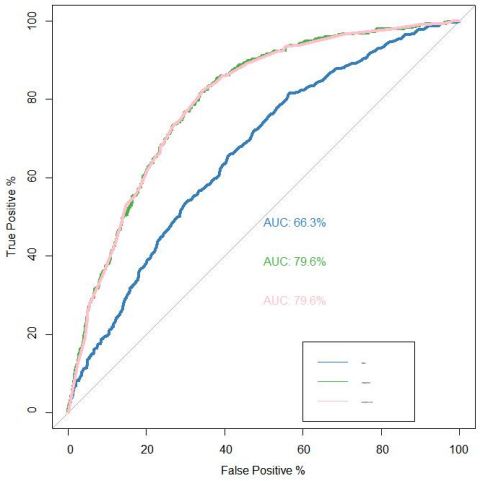

Figure 3. ROC-AUC curve for logistic regression of BMI model(blue), original model (green) and improved model(pink)

It is noticeable that the curve for the original model and the improved model are highly overlapped and they share the same AUC, verifying that the outcome of the improved model is almost the same as the original one.

In addition, the AUC of BMI logistic model is 66.3% while an AUC value near 50% can be considered as a bad predictor. On the other hand, the models using several features performs much better, having an AUC value of 79.6%.

Choosing proper threshold

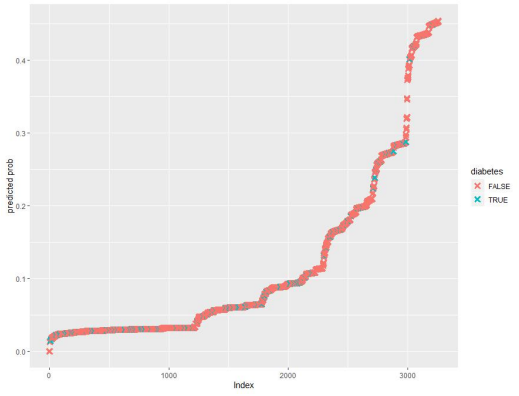

Figure 4. Predicted probability obtained from the model versus the true value

From the above figure, there is not a obvious threshold for being diabetes or not. However, by looking at the true positive percentage (tpp) and false

positive percentage (fpp), a proper threshold can be obtained.

By calling the following line of code, we can get

```
> str(roc.df[roc.df$tpp > 80 & roc.df$fpp < 40,])
'data.frame':       202 obs. of  3 variables:
 $ tpp       : num  85.9 85.9 85.9 85.9 85.9 ...
 $ fpp       : num  40 40 39.9 39.9 39.9 ...
 $ thresholds: num  0.0675 0.0684 0.0689 0.0693 0.
0696 ...
```

The meaning is that there are 202 tuples having tpp > 80% and fpp < 40%, which is a relatively large value. Therefore, by fine tuning the threshold, a relatively precise model can be obtained, which may be the further study goal of this project.

6. Observations from Figures

Weight distribution

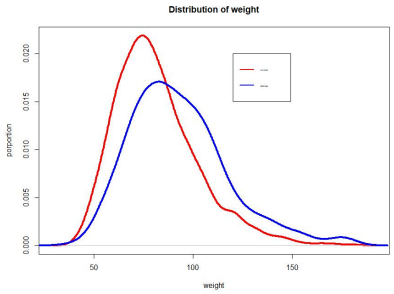

Figure 5. Weight distribution for people who had and did not have diabetes.

From Figure 1, we can see that weights of people who did not have diabetes has a lower mean that that of those who had diabetes, and the density at the center is larger as well. Thus, it can be concluded intuitively that people with a higher weight are more likely to have diabetes.

Histogram of BMI

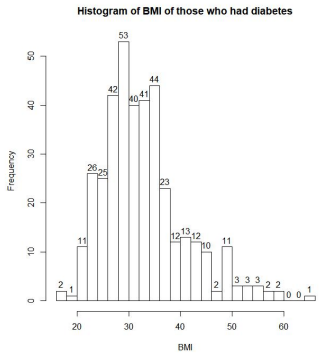

Figure6. Histogram of BMI of those who had diabetes

Hypothesis-3:

H0: There is no relation between BMI and height.

H1: There is relation between BMI and height. (using Pearson correlation test)

From the Pearson correlation test, the estimated correlation is -0.03122215, which is quite small, verified by a p-value = 0.07468. Therefore, H0 cannot be rejected at any significance level less than 0.07. For another thing, the correlation, which is a negative number, indicating a negative relation, conforms to the formula of BMI = weight/(height^2), where height is in meter. It also indicates that if two variable are correlated, their reciprocals may not be correlated.

Conclusion

There is a significant relationship between diabetes and height, overweight, blood pressure, cholesterol level, and alcohol taking, and a relatively good classification model can be got from these features.

There is a significant relationship between diabetes and BMI, however, the model obtained from this does not have a good performance.

The argument that there is a correlation between BMI and height cannot be disproved at any significance level below 0.07(i.e. confidence level higher than 93%).

The logistic model obtained can have true positive percentage > 80 and false positive percentage < 40 at the same time. Further research can be concentrate on selecting a proper threshold.

The model that we improved by excluding features that do not show statistical significance does not show significantly better performance.

Remarks:

Task 1.1 - 2.2-1: Result - 1. Data Processing

Task 2.2-2: Result - 6. Observations from Figures - Hypothesis-3

Task 3.1: Result - 2. Multiple Logistic Regression, 3. BMI and Diabetes

Task 3.2: Result - 6. Observations from Figures - Weight distribution, Histogram of BMI

Task 3.3: Conclusion