CLASSIFICATION USING LOGISTIC REGRESSION (For PIMA Indians Diabetes Dataset)

31/10/2018

BY
Priyanka Kalena 1511020
Aromal Nair 1511034
Saurabh Parkar 1511037

1. Problem Definition:

The objective of the dataset is to diagnostically predict whether or not a patient has diabetes, based on certain diagnostic measurements included in the dataset. Several constraints were placed on the selection of these instances from a larger database. In particular, all patients here are females at least 21 years old of Pima Indian heritage. The datasets consists of several medical predictor variables and one target variable, Outcome. Predictor variables includes the number of pregnancies the patient has had, their BMI, insulin level, age, Glucose level, Insulin, Skin Thickness, and Blood Pressure. According to the values, outcome is displayed as 0 or 1, which depicts whether a person has diabetes or not based on features. We would be training the model using logistic regression and random forest selection and compare the accuracy of the two models.

2. Preparing Data:

First we checked for unique values in the attributes to check if any of the attributes could be converted into factor type for easy of analysis. Since Outcome was only such attribute we converted it into factor. Rest remained either numerical or integers. There were no missing values in the dataset but through some medical sources found out that it was impossible for 5 of these attributes to have a zero values. These were – Glucose concentration, Blood Pressure, Skin Thickness, 2 hour serum insulin and Body Mass Index.

We plotted histograms for each of them to decide on the method for dealing with missing values. Glucose and Blood Pressure had very few missing values with a symmetric histogram and no outliers, so we decided to go with mean. For BMI, the histogram was symmetric but had a few outliers. So we replaced the missing values with mean of all the values by excluding the outliers to avoid any anomaly. Insulin had more than 300 missing values in a dataset of 700 values. The histogram itself was distributed unevenly with no visible trend or correlation with other attributes. As in such case it won’t be wise to assume as much as half the values we decided to drop this attribute altogether.

Lastly, we checked for correlation between SkinThickness and other attributes using scatterplots and corrplots and found that only BMI was related. We built a linear model to check the same and found that the model worked the best when only BMI was used. We therefore used this model to predict the missing values from the SkinThickness column and replaced the NA values.

3. Algorithms:

Mainly 2 algorithms have been used in the whole process. First being the linear regression which was used to predict the missing values of the SkinThickness column and Logistic Regression which was used to predict the Outcome i.e. whether the person has diabetes or not. The SkinThickness column contained unusually high number of missing values and hence it was not suitable to replace them simply with mean or median.

Using splom we generated scartterplot matrices to check for correlation between SkinThickness and other attributes and look for trends. We also used ‘corrplot’ to better visualize the relationship between the different features. Using the above plots we saw that BMI was the only attribute that was related enough so as to generate a pattern and use it for linear regression.

LINEAR REGRESSION:

We generated the linear model using the lm() method which takes 2 main arguments – formula and the data. The data is typically a data. Frame object and the formula is an object of class formula. By building the linear regression model, we established the relationship between the predictor and response in the form of a mathematical formula.
lm(Formula = SkinThickness~., data = new_d1)

Doing this gives us the values for intercepts and beta coefficients as in

SkinThickness = Intercept + (β ∗ Glucose) + (β ∗ BMI) + ………

Finally we predicted the values using the generated values and replaced them with the NA values in the SkinThickness column.

LOGISTIC REGRESSION:

We used Logistic Regression for classification – for predicting whether the person has diabetes or not. Logistic regression achieves this by taking the log odds of the event ln(P/1−P), where, P is the probability of event. So P always lies between 0 and 1. Taking exponent on both sides of the equation gives:

$$P_i = E(y = 1|x_i) = \frac{e^z}{1 + e^z} = \frac{e^{\alpha + \beta_i x_i}}{1 + e^{\alpha + \beta_i x_i}}$$

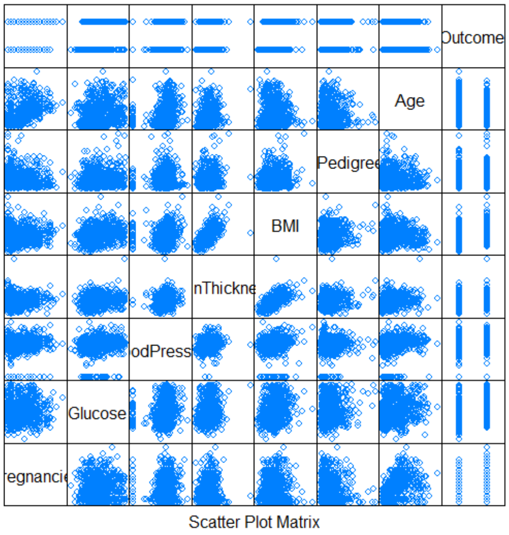
This can be easily implemented in R by using the glm() function and by setting the family argument to "binomial" which means that it can take only 2 values i.e. 1 or 0.
Similar to linear regression we will be dividing our dataset into training and testing datasets first.
glm() method also takes 2 main arguments – formula and the data

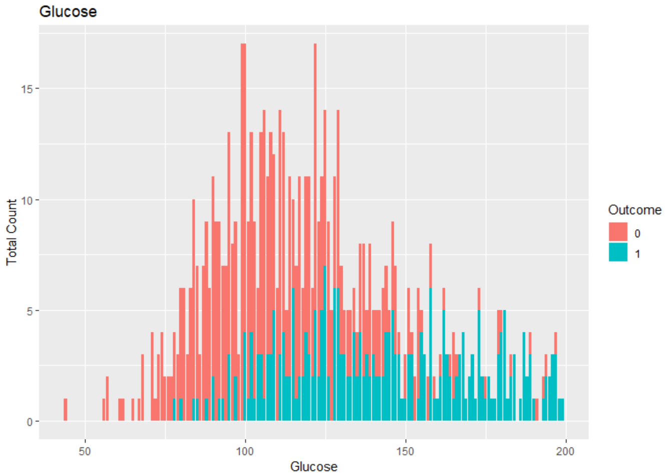
modelF <- glm(Outcome~., trainingF, family = "binomial")

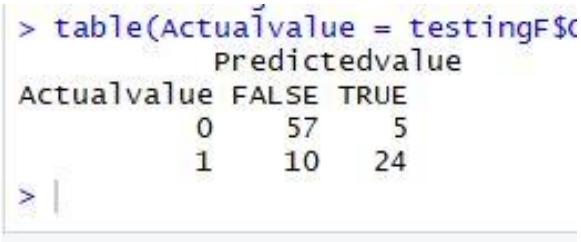
STATISTIC	**CRITERION**
AIC	Model with the minimum AIC is preferred
Null Deviance	Lower the better
Residual deviance	Lower the better
p- value	less than 0.05 or as small as possible

Present Result:

Our current model predicted the Outcome with a sizeable accuracy of 85.4% when the threshold was set at 0.4. Also the number of risk cases (when the model predicts that a person doesn’t have diabetes when he actually does) is also pretty low: 7. The AIC for the model is 617.15 and Residual Deviance= 661.15 which are also significantly lower than those from the previous models.







```
> table(Actualvalue = testingF$C
             Predictedvalue
Actualvalue FALSE TRUE
          0    57    5
          1    10   24
> |
```

Conclusion:

Thus we successfully predicted the Outcomes of the Diabetes dataset with a sizeable accuracy of 85.4% using logistic regression. Building the model and classifying the Outcome is only half work done because, the scope of evaluation metrics to judge the efficacy of the model is vast and requires careful judgment to choose the right model. Data analysis and exploratory analysis were the most important part of the whole process as it helped us to decide on choosing the right method to deal with missing values and for choosing proper variables while building the model. Visualizing the data using different techniques and verifying it using the statistical values from the model summary greatly helped to generate the best possible model. Also, the test error rate was not significantly different from the estimated test error and so we were also confident that the logistic regression model was not over fitting.