Pima-Indian Women Diabetes: Analysis & Prediction

4/30/19
Alexander Yang, Che-An Lin

Description of Project & Dataset:

The project is about what causes diabetes in women. The purpose is to see whether variables like number of pregnancies or glucose levels have a large factor in having diabetes. The main learning goal to compare what different models suggest are the biggest factors and find the common culprit. From there, we would advocate what can be done to prevent diabetes in women based on our results.

Research Question: What are the biggest factors contributing to the diagnosis of diabetes in women?

We performed logistic regression and decision tree models (unpruned & pruned) to compare results and collectively come to our conclusion. We also trained these models so they can potential be useful in future diagnosis of diabetes. The partition was 60% was the training data and 40% was the test data.

Resources:

For this project, we used the Pima Indians Diabetes Database provided by UCI Machine Learning on Kaggle. The samples in this dataset are all female patients.

Link to the dataset: https://www.kaggle.com/uciml/pima-indians-diabetes-database
Link to the Plotly library: https://plot.ly/r/

Software/Language:

- R & RStudio

Libraries:

- plotly
- rpart
- rpart.plot

- corrplot
- tree
- caret
- Hmisc
- randomForest

Summary:

For our logistic regression the most significant variables was Glucose, BMI, pregnancies, and DiabetesPedigreeFunction. With each increase of 1 unit of each of these variables, the likeness of the woman having diabetes increases. We performed the step function so that the RStudio can decide which variables are not helpful to the model rather than us hand-picking the variables. The model's accuracy ended up being at 79.6%.

For our unpruned tree, the variables used include Glucose, BMI, Age, DiabetesPedigreeFunction, Pregnancies, and Blood Pressure. Glucose was the primary factor of the split. For the 2nd split, it was Age and BMI. The rpart library was used to generate the tree. An example of a prediction would be when a woman as less than a glucose level of 128 and less than the age of 29, there is a 8% chance the woman does not have diabetes. The model's accuracy ended up being at 85.7%.

For our pruned tree, the variables used include Glucose and BMI. Like our last tree, Glucose was the primary factor of the split. The prune function is used from the tree library. An example of a prediction would be when a woman has greater than a glucose level of 128 and greater than a BMI of 30, there is a 72% chance the woman has diabetes. The model's accuracy ended up being at 79.2%.

The logistic regression model and the pruned tree model has about 80% accuracy which isn't bad. While the unpruned tree had a higher accuracy at 85.7%, it could be due to overfitting.

Conclusion:

The best predictors for diabetes based on the models are glucose and BMI. These predictors were mainly used or had the highest significance towards the response in the models. In the future, glucose levels and BMI should be used to predict diabetes in women that are Pima-Indian. What worked really well was the logistic regression and trees as the the response variable is binary. However, random forest algorithm was not a good model for this dataset and the output of the algorithm was difficult to interpret in the context of binary variable. We were surprised that blood pressure was not one of the top variables. Usually when we think of diabetes, we would think that blood pressure should be a major contributor.