**Springboard Foundations of Data Science**

**Capstone Project**

**Diabetes**

November 22, 2016

By: Swati Jani Joshi

Project Description:

The Dataset is from Kaggle – Prima Indians Diabetes Data set.

The objective of the capstone project is to provide analysis on the data set so as to derive a model that can predict the onset of diabetes.

Diabetes:

This project would be beneficial to scientists/researchers in the health services field as it would give a better insight of who would be pre disposed to diabetes. This is turn will benefit doctors to monitor people who are determined to be predisposed to diabetes. They can conduct frequent tests on such patients to monitor them and provide them with guidance on how to prevent or delay the onset of diabetes.

Diabetes is the seventh leading cause of deaths in the US. Diabetes affects men and women differently. On an average, men with diabetes live 7.5 years less than men without diabetes. On an average, women with diabetes live 8.2 years less than women without diabetes. Diabetes increases the risk of heart disease 6 times more in women and 3 times more in men

Data:

The Kaggle Data set has nine features in the dataset.

* Pregnancies
* Glucose Reading
* Blood Pressure Reading
* Skin Thickness
* Insulin
* BMI
* Diabetes Pedigree Function
* Age
* Outcome

The dataset contains data on women only.

Data Wrangling:

The Dataset had many observations with values of 0 for features such as Blood Pressure, Skin Thickness, Insulin etc. Since it is not medically possible for an alive person to have such values, those records were deleted. Since it is possible for women to have no children, observations where pregnancies were zero, were unchanged in the dataset.

The dataset was split randomly into 70% training set and 30% test set using the sort function.

Data Analysis:

A dot graph was plotted to visualize the correlations between the nine features. The corrplot function was used for the dot graph.

Some interesting observations were made based on the corrplot.

Age, Glucose, BMI and Insulin have significant impact on Outcome.

Age had a significant impact on pregnancies, which is puzzling, as generally with older women, number of pregnancies should decrease.

Modelling:

Logistic Regression Model

The glm function was used to determine logistic regression model fitting.

Logistic regression was chosen as the Outcome variable is a 1 or 0 i.e. yes or no which is a binary format.

Looking at the summary of the model displays that not all variables are significant to the model. Hence the logistic regression model that was ultimately used had the variables Glucose, BMI and Diabetes Pedigree Function.

The Logistic Regression Model Accuracy achieved was 82.4%

Random Forest Model

The randomForest function was used to determine the Random Forest model fitting.

The Variable Importance Plot was plotted to display the importance of the different variables as measured by the Random Forest model.

The Random Forest Model Accuracy achieved was 80%.

Comparison of the two models

The Random Forest Model performed better than the Logistic Regression Model as the Root Mean Squared Error for the Random Forest Model was smaller that the Root Mean Squared Error for the Logistic Regression Model.