Pima Diabetes Data Analytics

Some details about the dataset:

Pregnancies: Number of times pregnant
Glucose: Plasma glucose concentration a 2 hours in an oral glucose tolerance test
BloodPressure: Diastolic blood pressure (mm Hg)
SkinThickness: Triceps skinfold thickness (mm)
Insulin:2-Hour serum insulin (mu U/ml)
BMI: Body mass index (weight in kg/(height in m)^2)

Aim:

Building a machine learning model to accurately classify whether or not the patients in the dataset have diabetes or not

Here are the set the analytics that has been run on this data set

- •Data Cleaning to remove null values and impractical datas
- •Data Exploration for Y and Xs
- •Descriptive Statistics – Numerical Summary and Graphical (Histograms) for all variables
- •Screening of variables by segmenting them by Outcome
- •Study bivariate relationship between variables using pair plots, correlation and heat map
- •Validation of the model

Steps involved in project:

1. Importing neccessary packages
2. Importing the Diabetes CSV data file:
   - Import the data and test if all the columns and respective datas are loaded
   - The Data frame has been assigned a name of ‘data’

3. Using **.info**() function to get a concise summary of the dataframe.

4. Using **.describe**() function to view some basic statistical details like percentile, mean, std etc. of a data frame.

5.Plotting histograms for visual representation of data destribution.

6. Using .drop() to drop impractical datas

7. Split the data frame into two sub sets for convenience of analysis

As we wish to study the influence of each variable on Outcome (Diabetic or not), we can subset the data by Outcome

8. Screening of Association between Variables to study Bivariate relationship

.We are using pairplot to study the association between variables – from individual scatter plots

.Then we are using compute pearson correlation coefficient

. Then we are summarizing the same as heatmap

9. Inference from Pair Plots

•From scatter plots, BMI & SkinThickness and Pregnancies & Age seem to have positive linear relationships. Another likely suspect is Glucose and Insulin.

•There are no non-linear relationships

10. Inference from heat map

- Glusose is having the moderate correlation and maximum among all of them

11.identifying diabetes with positive and negative results

12. splitting into training set and test set

13. displaying number of test and train sets

14. importing model Random Forest

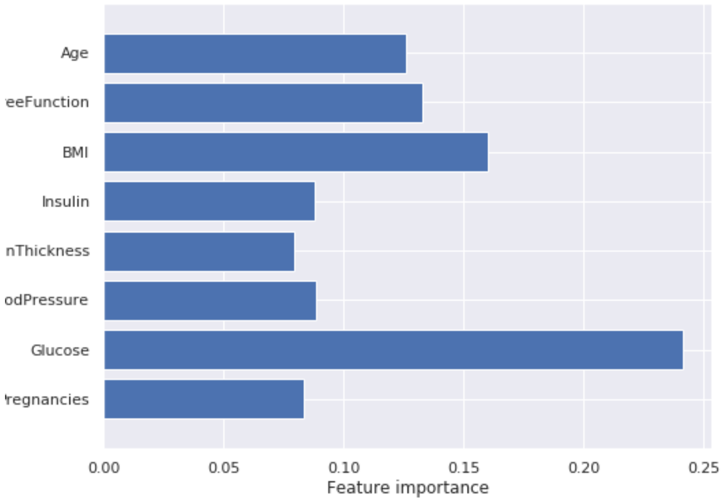
- Precision of the model is 79%

We are using random forest because:

•The random forest combines hundreds or thousands of decision trees, trains each one on a slightly different set of the observations, splitting nodes in each tree considering a limited number of the features. The final predictions of the random forest are made by averaging the predictions of each individual tree.

Conclusion from Feature Importance Plot:

•the random forest also gives a lot of importance to the “Glucose” feature, but it also chooses “BMI” to be the 2nd most informative feature overall. The randomness in building the random forest forces the algorithm to consider many possible explanations, the result being that the random forest captures a much broader picture of the data than a single tree.


Age
eeFunction
BMI
Insulin
nThickness
odPressure
Glucose
'regnancies
Feature importance