Assignment - IV

Siddharth Chandrasekaran
690V

October 7, 2017

Food frequency questionnaire with additional demographics and computed food content for anonymized data for 54 anonymized individuals and with over 1000 variables was provided. This dataset consisted of three very crucial information about each individual:

- Disease information (cancer, diabetes etc)
- Characteristics (Smoker, non smoker etc)
- Food Habits (How much sandwiches does he eat etc.)

Since food habits and characteristics are crucial information required to predict disease possibility, classification models were trained on the dataset comprising characteristics and food habits with the target variable as each disease (cancer, diabetes and heart disease). Classification models used:

- Linear SVC
- Decision Trees
- Random forest

Each of the models give out feature importance (a number between 0 & 1 that informs how important the feature was in that particular model).

This information gives us a pretty good intuition on what all patterns exist within the dataset with respect to diseases.

Example : We can see that GROUP_YOGURT_PLAIN_NON_FAT_TOTAL_GRAMS is very important feature to predict Diabetes.

So to visualize this, we plot a bar chart which contains all the features whose importance score ¿ 0.01. (This plot is enabled with hover which tells us the feature and the value since some classification techniques yield a large number of features and it looks cluttered owing to screen resolution).

But importance alone wont help solve the problem statement of being able to put it in phrases such as ”having a cat and eating bagels often decreases your chances of having heart disease”.

Hence, to know whether some the important features are positively correlated or not, we plot the heat map of the correlation matrix of all the features with respect to the diseases.

Blue implies negative correlation and Red implies positive. Black implies extremely high positive/negative correlation since black catches attention upon first glance in a predominantly white based background and these features are extremely important to look at.