**Analysis: Team 6**

**Variables Used:**

'ACCESS2\_CrudePrev', 'BINGE\_CrudePrev', 'BPMED\_CrudePrev', 'CHECKUP\_CrudePrev', 'CHOLSCREEN\_CrudePrev', 'COLON\_SCREEN\_CrudePrev', 'COREM\_CrudePrev', 'COREW\_CrudePrev', 'CSMOKING\_CrudePrev', 'DENTAL\_CrudePrev', 'LPA\_CrudePrev', 'MAMMOUSE\_CrudePrev', 'OBESITY\_CrudePrev', 'PAPTEST\_CrudePrev', 'SLEEP\_CrudePrev','LPAandOBESITY','Core','CSMOKING\_Double'

Meaning of variables: percent residents who answered yes to a given question.

**Target Variable:**

'CANCER\_CrudePrev'

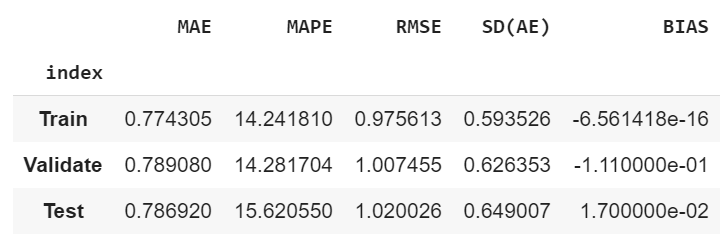
**Data Partition:**

Training - 80%

Validation - 10%

Test - 10%

**Model0: Baseline Model**



Baseline model predicts the mean of the target and results in high errors.

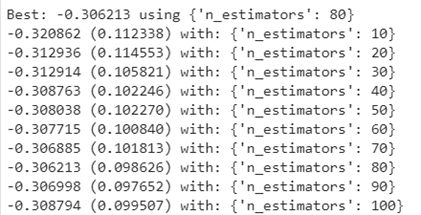
**Model1: Random Forest**

**Algorithm Used:**

K Fold Cross Validation (number of splits as 10)

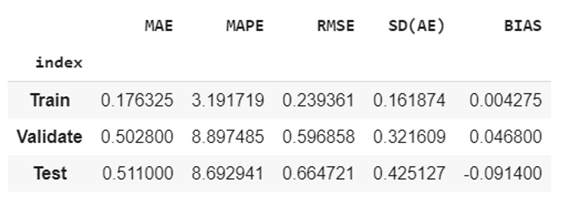
Number of trees built - 10,20,30,40,50,60,70,80,90,100

**Results:**

****

Result shows that number of estimators equals to 80 gives the best result as a negative mean square error of -0.306213.

**Error Metrics:**

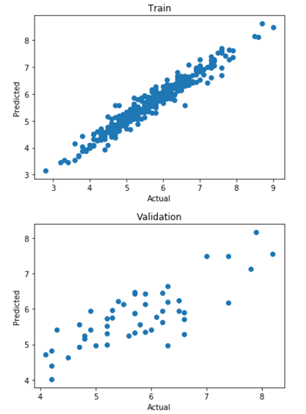


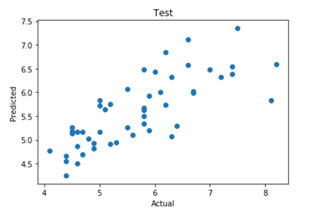
Here we can observe that RMSE in validation has not improved compared to training model. Thus, the model does not give good accurate predictions in validation model.

If we compare the RMSE values for validation and test, there is again a drop in accuracy(High RMSE value).

However, random Forest may address the data overfit, but the model accuracy is a trade off in this case.

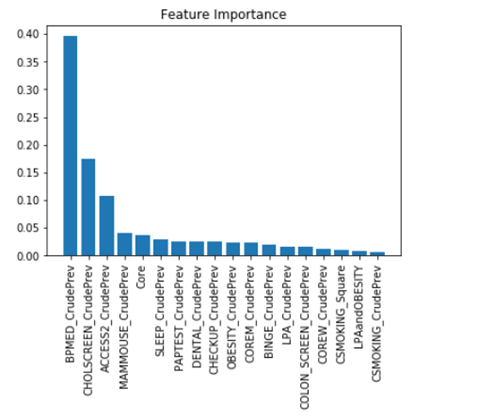
**Comparing Scatter Plot Actual vs Predicted:**





Here the training model gives a tight distribution of points, which is an ideal scenario. However, we check for the validation model results. Scatterplot for validation model and test model shows that the points are comparatively wider. This means that there is no relationship between the model’s predicted and actual values. The nature of the distribution shows that the model is steadily underestimating the actual values.

**Variable Importance:**

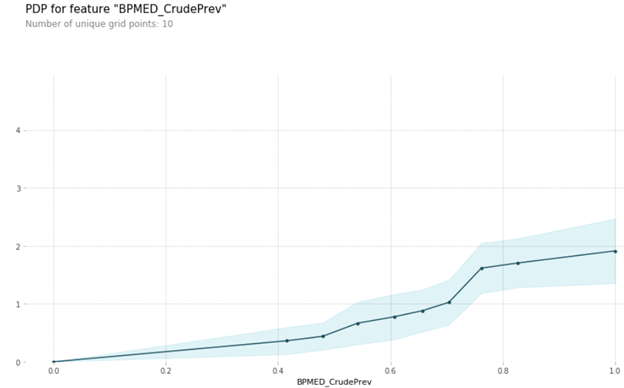
****

This plot suggests that the 3 features - **BPMED\_CrudePrev** , **CHOLSCREEN\_CrudePrev**, and **ACCESS2\_CrudePrev** are informative, while the remaining are not.

**Partial Dependence Plots:**

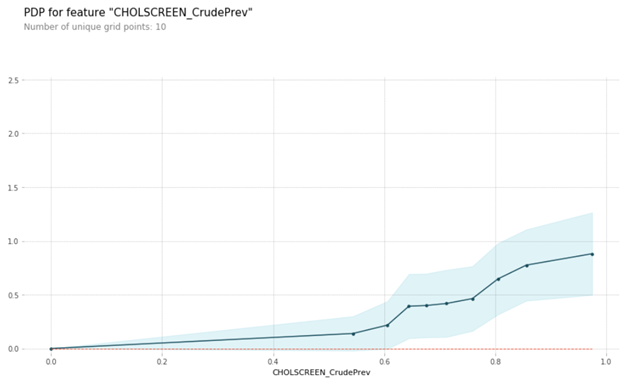
The 3 most important features have been used for partial Dependence Plot”

1. **BPMED\_CrudePrev**



The graph shows that, as the population of adults aged greater than or equal to 18 years who consumes medicine for high blood pressure control increases, the impact of being cancerous is more likely. The model results could have been more effective if we could identify the specific age group that are more likely to be cancerous.

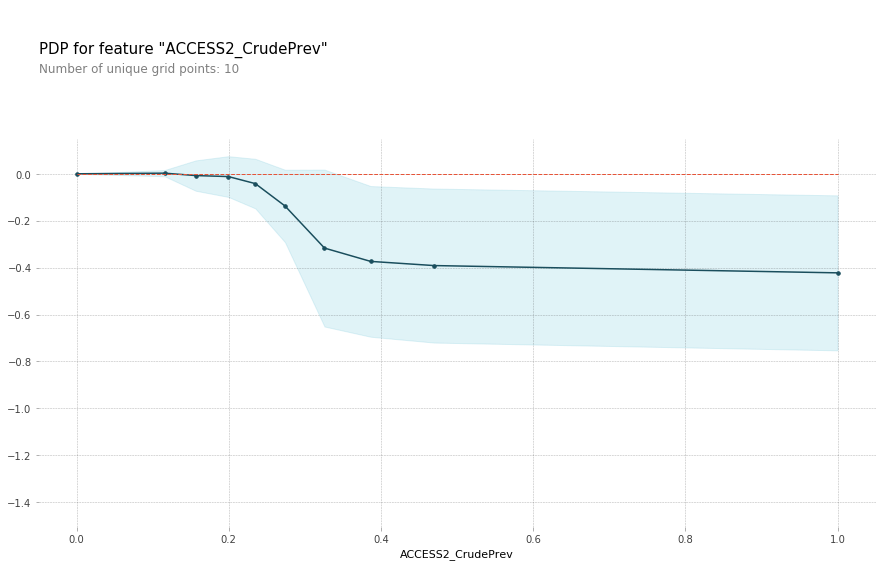
2. **CHOLSCREEN\_CrudePrev**

****

The graph shows that as the population of adults aged greater than or equal to 18 who goes for cholesterol screening increases, the impact of being cancerous is more likely.

As mentioned in the above analysis, the model could have been more effective if it identifies the age group of the particular set of population being cancerous.

3. **ACCESS2\_CrudePrev**

****

The graph shows that as the population of adults who lack insurance protection increased after a certain point, the percentage of cancer tend to stay the same. Which means there is not enough evidence that lacking insurance has impact on having cancer.

**Model 2: GBM**

We use the gradient boosting machine for two reasons. One reason is the ability to determine the relative influence of one or more variables to the target, the percentage of cancer. Another reason is that the GBM has a great prediction performance.

In the GBM model, there are two interaction terms and one risk score: LPA and Obesity, CoreM and CoreW and CSmoking. Firstly, LPA and Obesity might have some interaction with each other since LPA is one of reasons lead to obesity especially when other lifestyles do not lead to obesity such as healthy diet. Secondly, CoreM and CoreW might have some interaction with each other. Suppose each variable has a similar relationship with the target, the interactive term can have more information gathered based on gender. Finally, each of these variables may have different scale of impact on the outcome variable. Since our target is cancer, we are going to give a higher risk score for CSMOKING\_CrudePrev by taking the square of it.

According to the error metrics, the model has 0.177 MAE, 3.185 MAPE, 0.223 RMSE and 0,136 SD in training set, 0.433 MAE, 8.108 MAPE, 0.585 RMSE and 0.393 SD in validation set, and 0.379 MAE, 6.421 MAPE, 0.532 RMSE and 0.373 SD in testing set. Compared with the performance of the random forest model, the overall performance of the GBM was improved significantly.

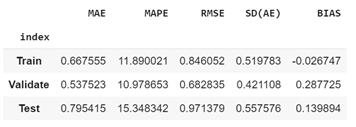
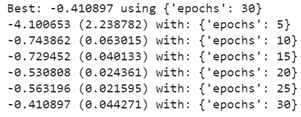
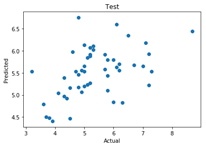
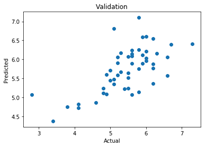
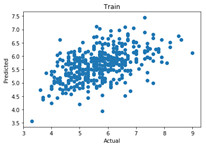
According to the feature importance plot, the combination of LPA and Obesity is not good predictive interaction term since it is the least important feature. In other words, people who never do physical activity in leisure time and have obesity do not have any impact on predicting cancer in the population. Besides, the combination of CoreM and CoreW and the risk core of CSmoking are ranked 8th and 9th, which means that the clinical preventive services and smoking have influence on preventing and causing cancer in both the male and the female aged more than 65. But compared to that of the top four important features, the influence of clinical preventive services and smoking are still light. Therefore, we still recommend people to receive clinical preventive services after 65 years old or even before.

In the top four important features, there are taking medicine for high blood pressure control, taking cholesterol screening, lack of health insurance and using mammography. It can be interpreted as that controlling high blood pressure, taking cholesterol screening and using mammography strongly contribute to increasing a person's risk of cancer. But people do not own health insurance does not influence predicting cancer.

Accordingly, we recommend people to pay more attention to their blood pressure and cholesterol, taking mammography and buy insurance for themselves.

In order to improve the model, we will remove the least three important features, DENTAL\_CrudePrev, LPA\_CrudePrev and LPAandOBESITY. And then there are 13 variables, one interaction item and one risk score variable for the new model to predict the percentage of cancer.

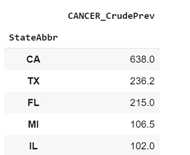
**Model3: Neural Network**



Neural network model gave a best result of negative mean squared error of -0.41 with 30 epochs. It may get better results as the epoch sizes increased. The model is not overfitting to the train and the validation showed that the model fitted well (lower error metrics), while the test partition got a bit worse (higher error metrics). Comparing the scatter plots, training and validation have similar shapes while test partition has a more spread shape.

Overall, neural network did not perform as well as random forest or GBM for this dataset with our trial of the settings, increasing epochs or batch size, change learning rate, and increasing layers may help with improving the performance of the model.

A binary dataset would be very helpful for modeling the health outcomes, which contains the same health outcome as yes or no and same kinds of variables as a yes or no for a certain group of people. With this kind of data, we have more accurate information about how prevention and unhealthy behaviors influence health outcomes. Our data is about the percentage of residents who answered yes to a given question, which means they do not have to be the same group of people; thus, it makes out model less meaningful.



The top five states of people having cancer are shown above. CA is a hotspot for cancer compared with other states.