PREDICT THE CANCER MORTALITY RATE PER CAPITA

This Documentation is made to brief about the work done to solve the Machine Learning Problem.

Data Analysis:

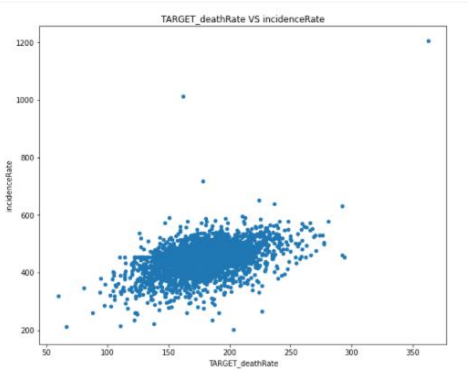
- Given dataset has lot of columns in which few are understandable and few are not.
- So we cannot ignore any of the columns just like that.
- Dataset has total 34 columns in which 2 are of data type objects and remaining are of data type int and float.

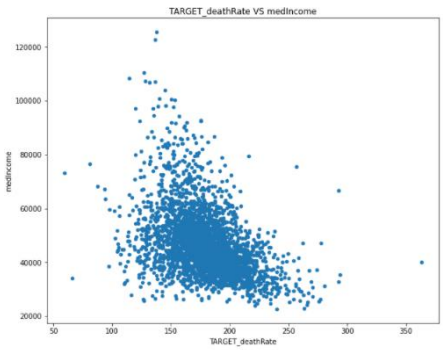
Data Cleaning:

- Checked for the Null values. The 3 columns of data set has null values which are replaced by the mean values of that particular columns.
- Dropped all the duplicate values.
- As 2 columns has object data type hence did One Hot Encoding to the columns i.e for the columns binnedInc and Geography.
- Now the Total columns are 93.

Data Visualization:

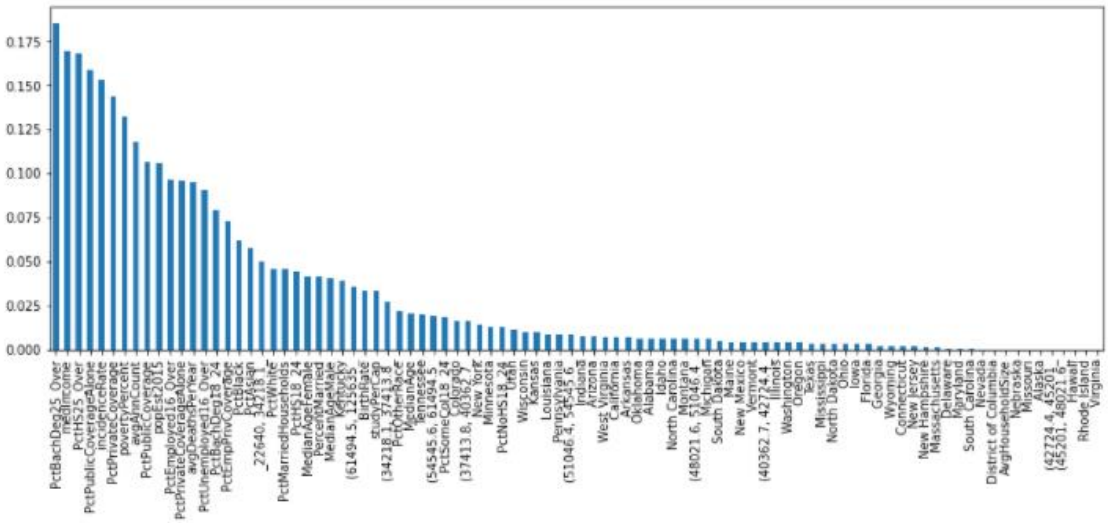
- Calculated correlation between all the variables and plotted heat map. But as the columns are huge it is difficult to find the relation.
- Hence visualized few features vs Target variable. Could find that the Target variable is not so much positively dependent on these features.





Data Featurization:

- There are many techniques for feature selection. One using PCA through which we can reduce the dimension but in this it might be possible that we may lose some data which may affect the models.
- Hence in this mutual_info_regression is used. This is for feature selection with information gain.
- **Mutual information** between two random variables is a non-negative value, which measures the dependency between the the variables with the Target variable.
- It is equal to zero if and only if two random variables are independent, and higher values mean higher dependency.
- Finally only 20 features are selected because these have 99% of the valuable information.
- Feature selection graph in the Descending order of dependency values.



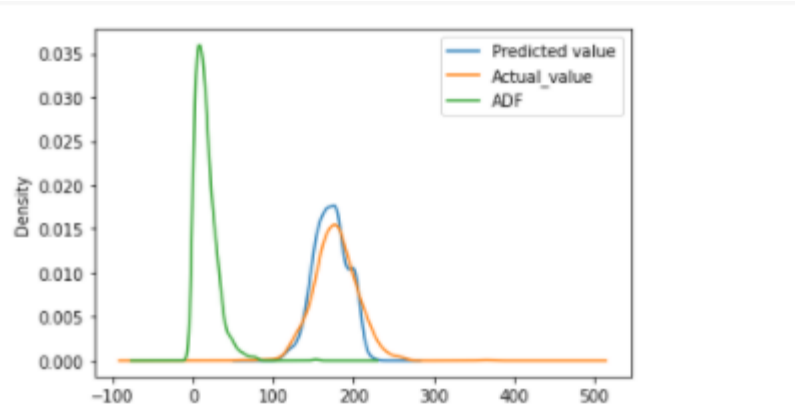
Performance Metric:

- In any machine learning problem it is very important to select right performance metric.
- For this problem statement Mean Absolute Error is used. This is because because as the columns are more it is very difficult to find the outliers in each column and deal with them individually. Hence in order to ignore the outliers and check the performance MAE is used.

Modelling:

- Initially data is trained and tested with 7 Regression models by giving default parameters.
- Models are:
  1. Linear Regression
  2. XGBRegressor
  3. DecisionTreeRegressor
  4. ExtraTreeRegressor
  5. SVR
  6. RandomForestRegressor
  7. LGBMRegressor
- For each model the absolute difference is calculated and plotted them to check the spread of their differences.

- After finding MAE values for all the models XGBRegressor is chosen as best choice as the spread of the difference is less compared to other models.
- Density plot of XGBRegressor



- From above graph we can see that the spread of ADF is less.

Hyper Parameter Tuning

- Once best model is selected we can improve the MAE values by doing a hyper parameter tuning.
- The result of Hyper Parameter Tuning was not so much effective. Because the MAE of test value is decreased but the model is overfiting. Tried to tune the model with more parameters but model is still getting overfitt.