Analysis of Heart Disease Mortality

Deepesh Nair, July 2018

Summary

This chapter describes the dataset used to develop a Machine Learning model to predict Heart Disease Mortality in the US region .

After exploring the data by calculating initial summary and descriptive statistics, and by creating visualisations of the data, several potential relationships between key characteristic variables and mortality rate were identified. After exploring the data, a regression model to predict the heart mortality rate from its features were created.

- **Area**: variables that contain information about the county.
- **Economy**: they are categories of economic dependence, labour force, unemployment and insurance.
- **Health**: indicators about obesity, smoking, diabetes and other characteristics of the county's population.
- **Demographics**: County's characteristics such as age percentage distribution, education, and others.

The report uses the techniques and analysis of the rate of heart disease (per 100,000 individuals) across the United States at the county-level from socioeconomic indicators. The data is taken from the United States Department of Agriculture Economic Research Service (USDA ERS) and the University of Wisconsin Population Health Institute. County Health Rankings & Roadmaps.

There are 33 variables in this dataset. Each row in the dataset represents a United States county that is unidentifiable in the dataset. While all the variables were studied, the discoveries lead to detect outliers, missing data and skewed values. And, after using categories for some variables, both area information about the county and the economic typology play a major role in the prediction of heart disease.

The target variable for prediction is: *heart_disease_mortality_per_100k*. It is defined as the rate of heart disease (per 100,000 individuals).

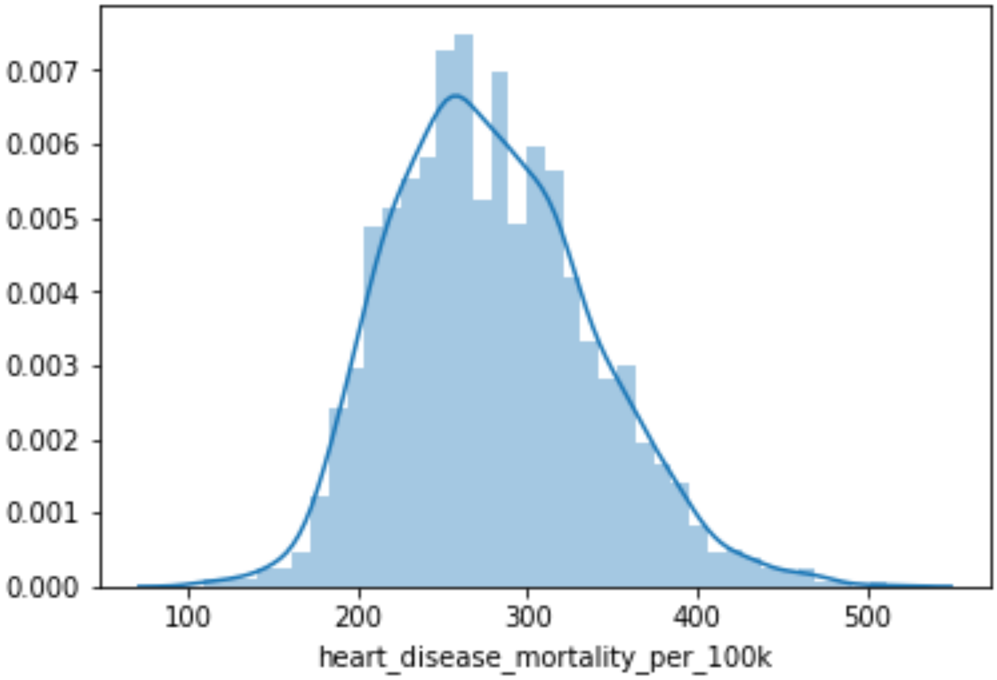
When developing the Machine Learning model, only one model couldn't get an effective prediction. Therefore Stacking Regressions were used to improve accuracy.

Initial Data Exploration

A quick overview on the Summary Statistics from original Training & Label set

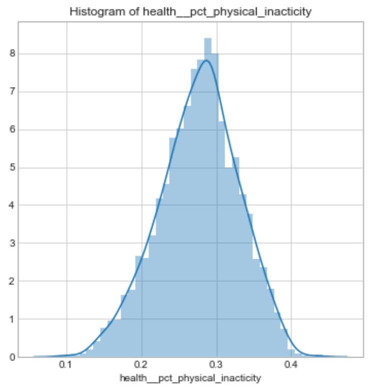
count	3198.000000
mean	279.369293
std	58.953338
min	109.000000
median	275.000000
max	512.000000

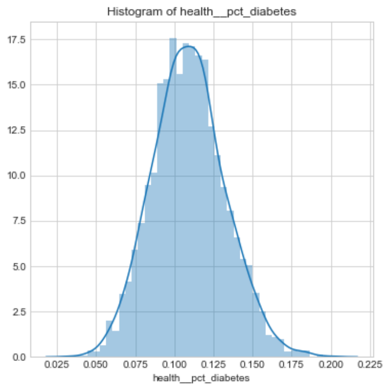
A Histogram showing the distribution of Mortality Rate (Per 100K) from the Label set

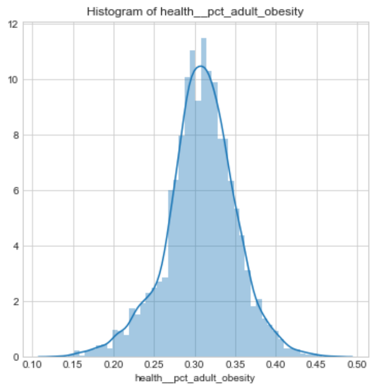


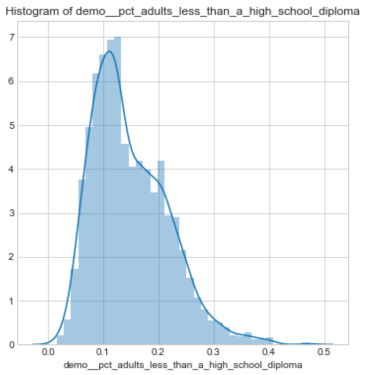
Histogram

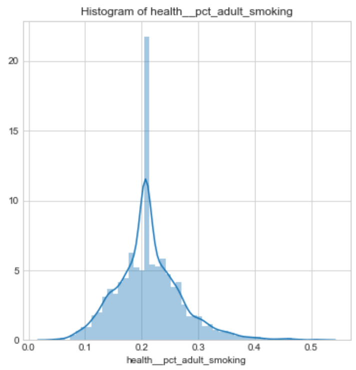
Histogram plots helped in understanding the distribution of features in the dataset . Below are some of the examples taken from training values.

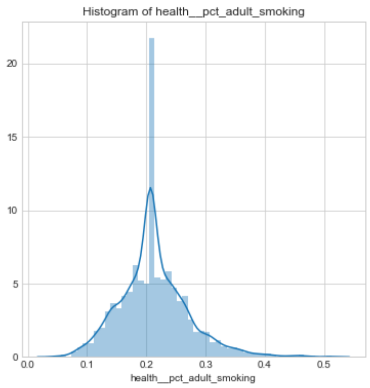


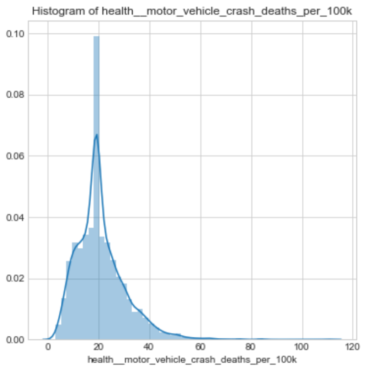


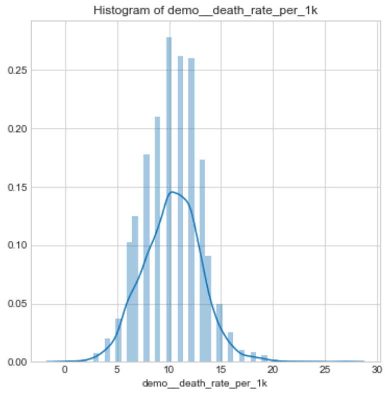


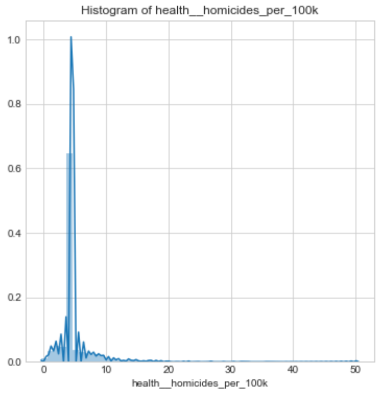


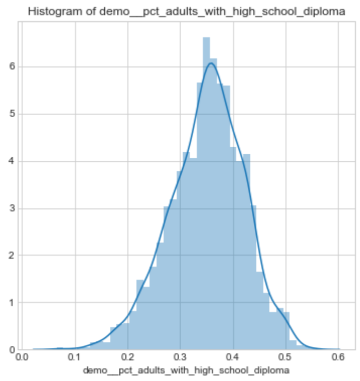


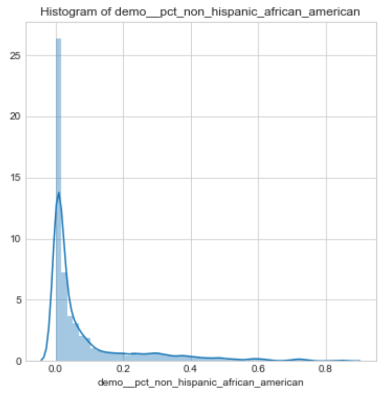


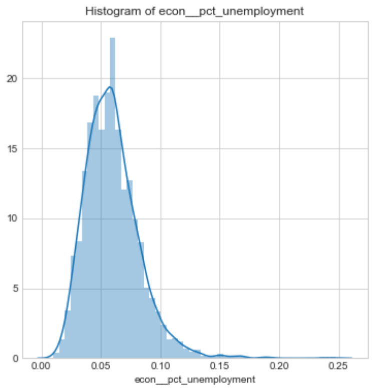






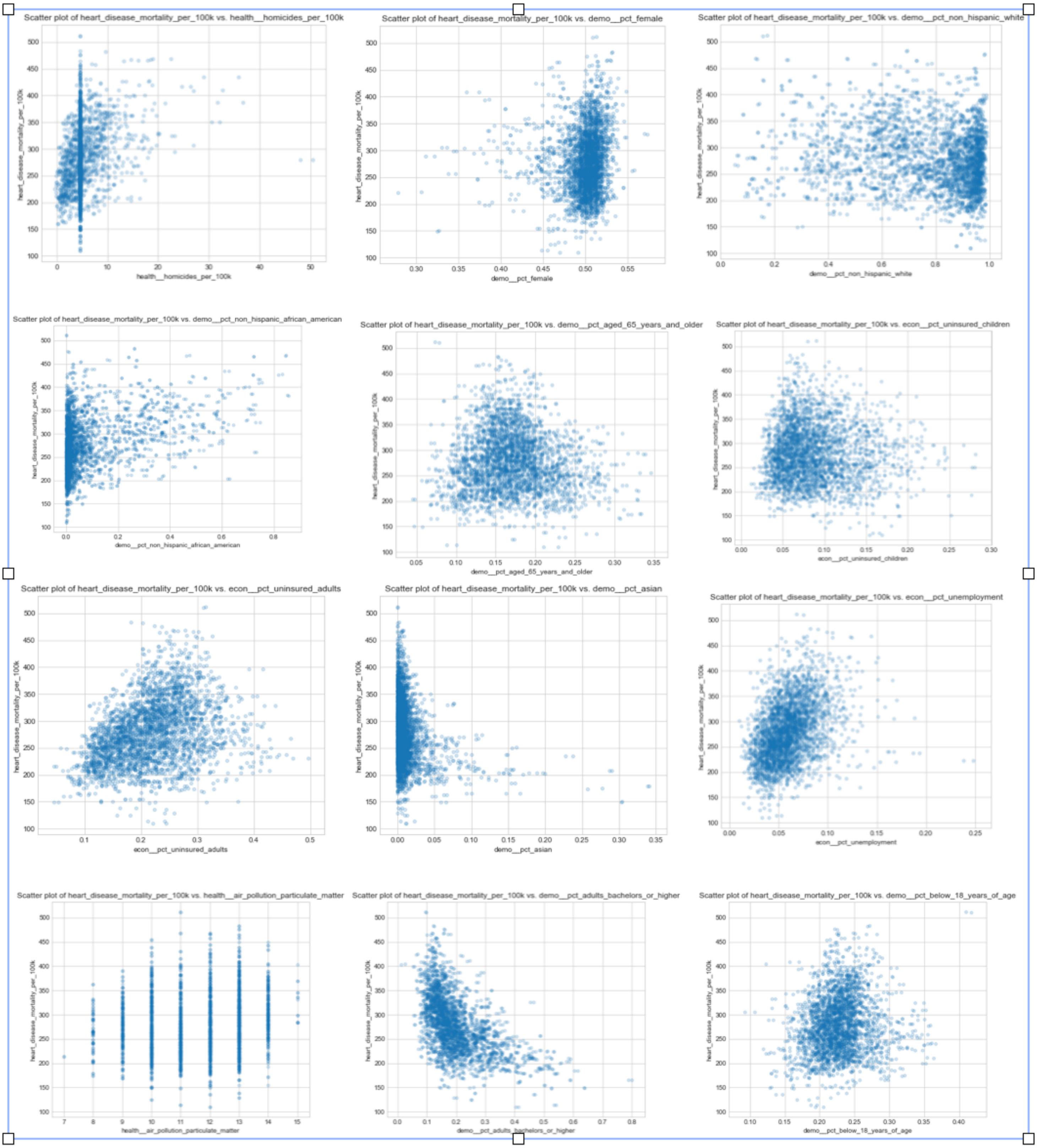






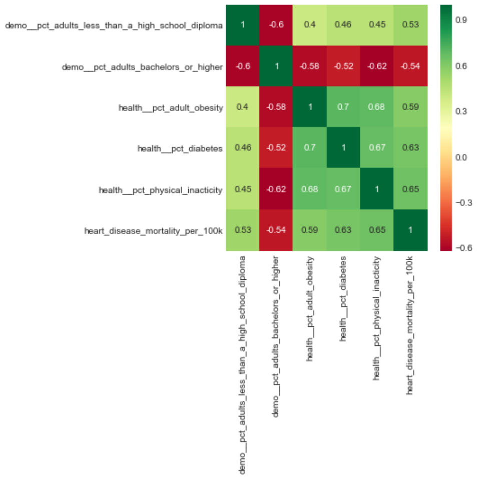
Scatter Plot

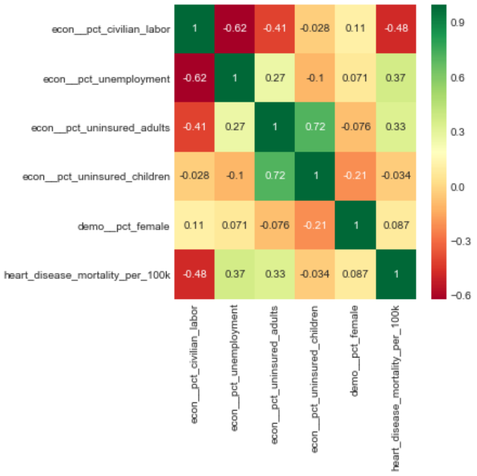
Moving ahead with further Data exploration , Scatter plots were used to carry out bi-variate analysis. A relationship between each numerical features with heart disease mortality . Below is a snapshot from some variables.

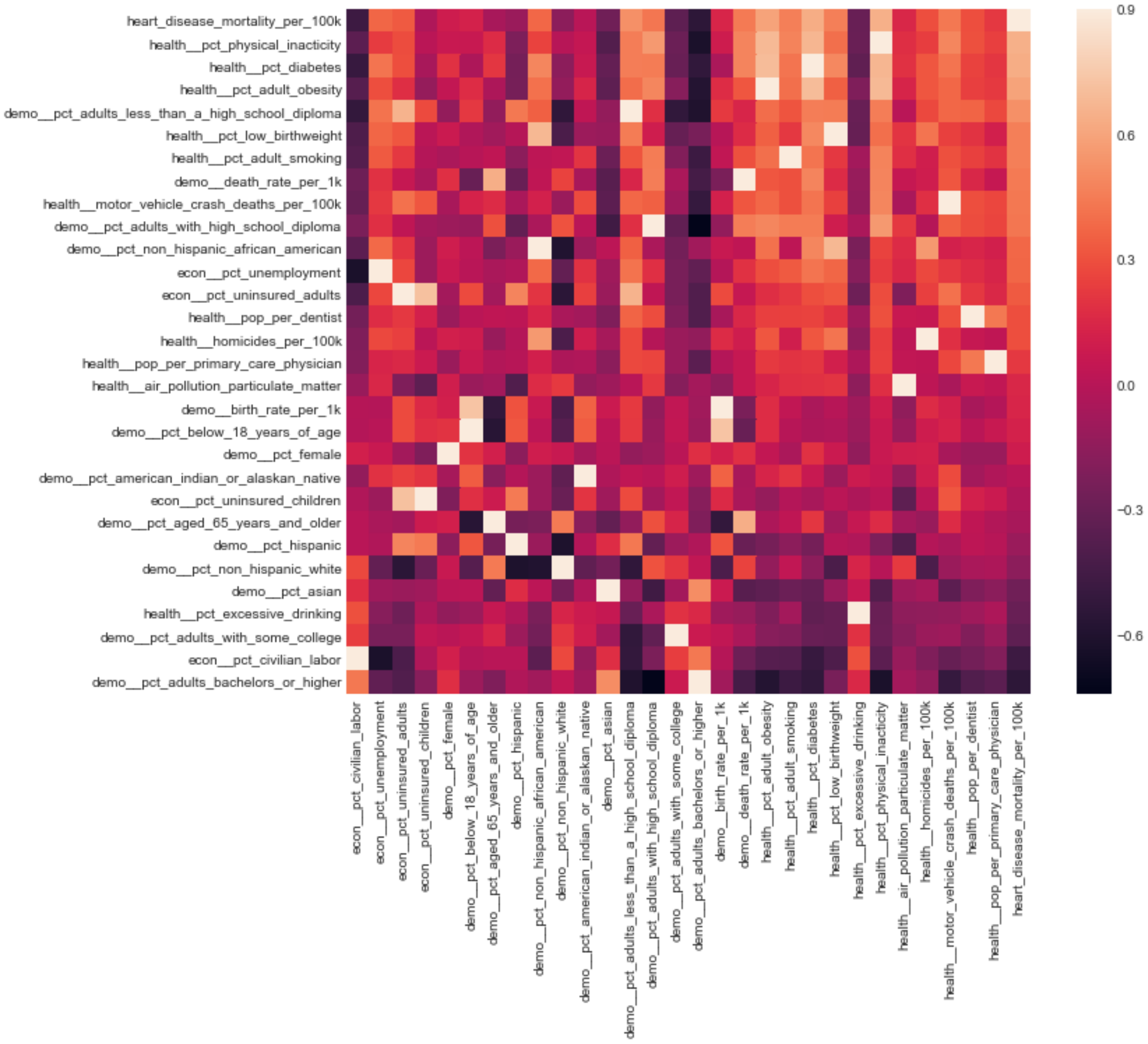


Correlation and All Apparent Relationships

Scatter plots and Histogram gave us some brief visual assistance . However , to gather more conclusive information , Correlation matrices were generated to provide concrete numerical metrics between mortality against all variables in the training dataset



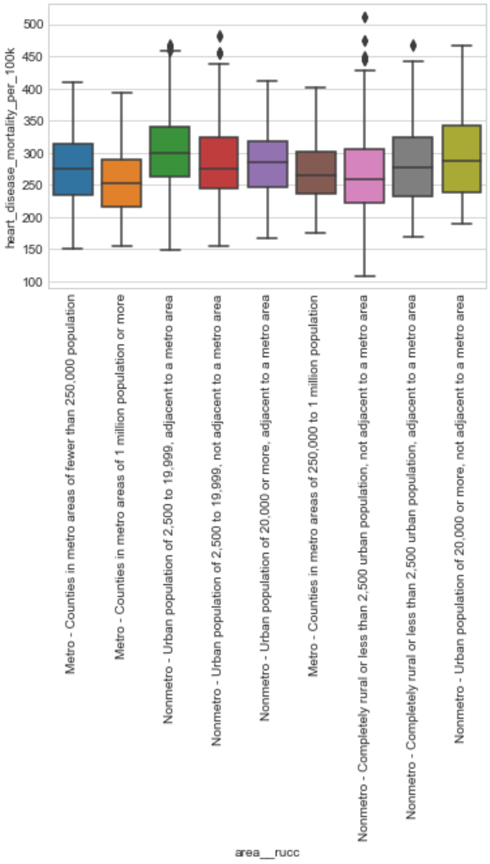


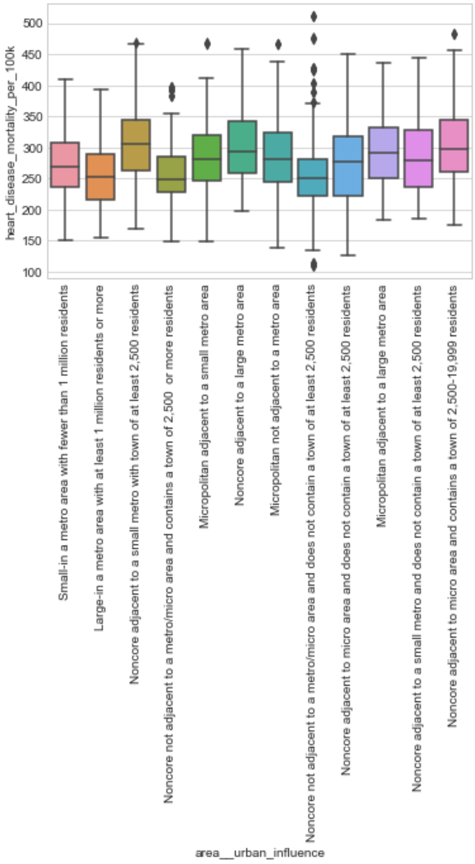


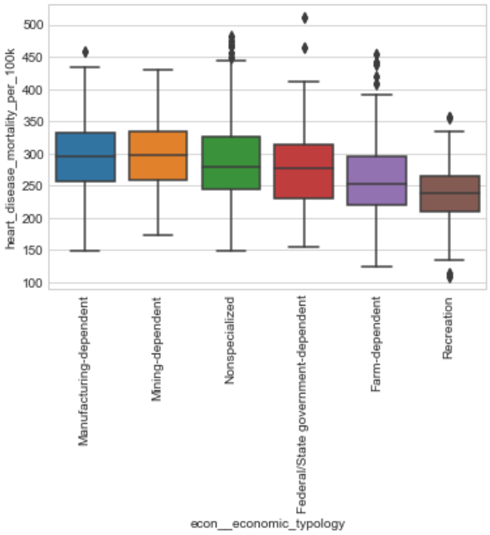
Categorical Relationships

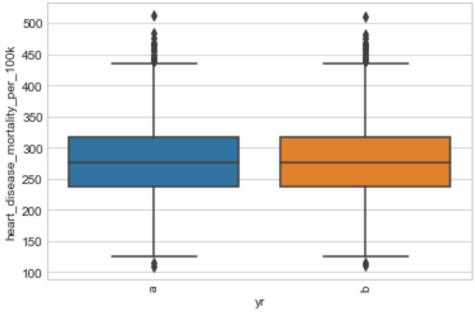
These findings suggest that the lifestyle factors that come with income can reduce or increase the risk for heart disease. With this in mind its logic to think that factors like education, income and wealth play an important role in overall health. Social position can influence a person’s behaviour, impacting decisions related to diet, exercise and smoking.

Exploring the data in more detail, the analysis proceeds to the data set distribution.









Regression

Based on the above training dataset and subsequent data exploration , we now proceed by selecting the best predictive model using the train and test values.

Below is the list of model(s) that were tested on the training dataset.

Model	RMSE	R^2 (%)
Lasso (L1)	33.81538189006959	66.13
Ridge (L2)	33.78989886513291	66.18
Gradient Boost	27.569181843597576	77.49
Light GBM	30.724454201370765	72.04
XGboost	27.656712439795168	77.37

As you can observe from above , **Boosting** mechanisms yields in a lower RMSE and an even better R^2 value (*>70%*) as compared to the other models . The residuals also look normally distributed with a constant variance.

Boosting is a machine-learning, distribution-free approach, for regression and classification problems, based on the idea of creating a highly accurate predictive model by combining many weak models, performance slightly better than chance. The combined model produces strong predictive accuracy.

Boosting's performance lies in the dynamics of its algorithm: the minimisation of the loss function is done sequentially using residuals, not the target variable.

Why use boosting? They are slow to overfit. And, it accounts for most wins in data science competitions.

The best model (Gradient Boosting) was implemented on the test dataset.

However , upon multiple submissions and learning from the aspects of using a stacked generalisation & ensemble concepts , it was observed the predictions improved drastically by stacking an average of different best model scores.

Additionally , an Ensemble score can be weighted (80/20 rule) by tweaking the result of the stacked predictions with individual best model.

Key Points

Below are some of the considerations , observations and conclusions made during the course of this regression problem.

- Missing Data were replaced with the median of the column values in order to impute the dataset towards a normal distribution
- Categorical Fields were mapped to binary values.
- Data Analysis proved health, economic and education features attributed towards strong correlation.
- Mortality rate was more in areas where Adult Smoking were prevalent.
- A single effective prediction model wasn’t able to demonstrate predictive values .
- Multiple regression models were ensembles in order to improve the accuracy of the Target Variable.

Conclusion

A Stacked Prediction Model was constructed by collectively identifying individual decent (based on RMSE) base models. To leverage the benefits of improved accuracy , the final prediction for the Heart Disease Mortality was generated by assigning weights to the stacked model along with best Boosting result using Ensemble techniques.