horizontal line

Data Mining and Advanced Statistical Modelling - Project Report

**Jikku Abraham**

# OVERVIEW

Data mining deals with the extraction of interesting and non-obvious patterns or knowledge from huge amount of data. Data pre-processing, correlation analysis, modelling and evaluation are major steps in data mining. Linear regression, logistic regression, decision trees are some examples of modelling methods. In this project, we looked into two major types of modelling: linear and logistic regression.

# GOALS

1. To perform a multi-linear predictive model to predict the strength of High-Performance Concrete (HPC) with respect to the supplementary materials such as fly ash and blast furnace slag, and chemical admixture, such as superplasticizer
2. To build a logistic regression predictive model to forecast the health of a subject if the city of residence, gender, age and income were provided

# Linear Regression

## The Dataset

The dataset was taken from UCI Machine Learning Repository. The dataset had a total of 1,030 instances. The total number of attributes were 9, out of which 8 were quantitative input variables and 1 quantitative output variable (refer figure). The input attributes were Cement, Blast Furnace Slag, Fly Ash, Water, Superplasticizer, Coarse Aggregate, Fine Aggregate & Age and the target/ dependant variable is the Strength of the High-Performance Concrete. All the dependant variables were numeric and there were no missing values.

## Data Analysis

The heat map shown in the figure shows the target variable, Strength, has close to zero to no correlation with the independent variables. The pair plot of all the variables bears the same witness (figure below).

# 

# 

# 

# 

# 

# 

# 

# 

## 

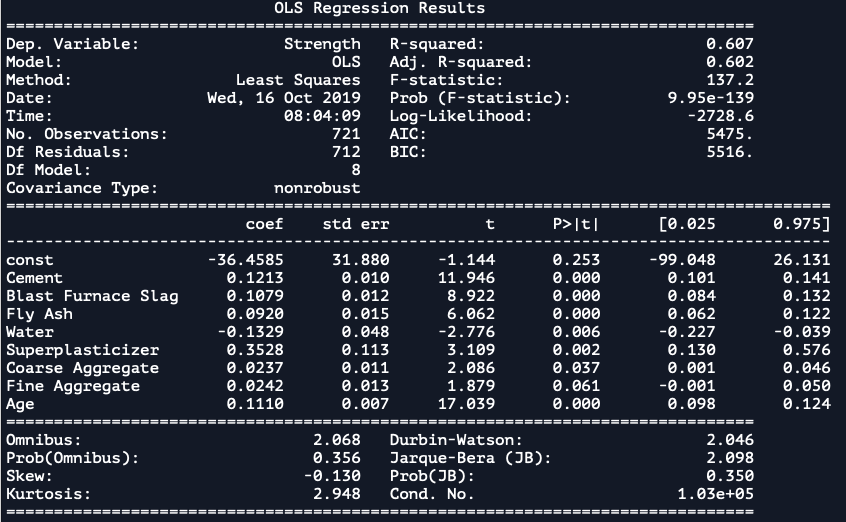
## Data Pre-processing

Since all the variables in the dataset were numeric, no encoding was needed and also since there were no missing values, no filling of missing values addition steps were required. The dataset was scaled during an earlier pre-testing phase and the R-square score of scaled and unscaled was found and observed to be of a negligible difference. So in light of this observation, the dataset was kept as-is.

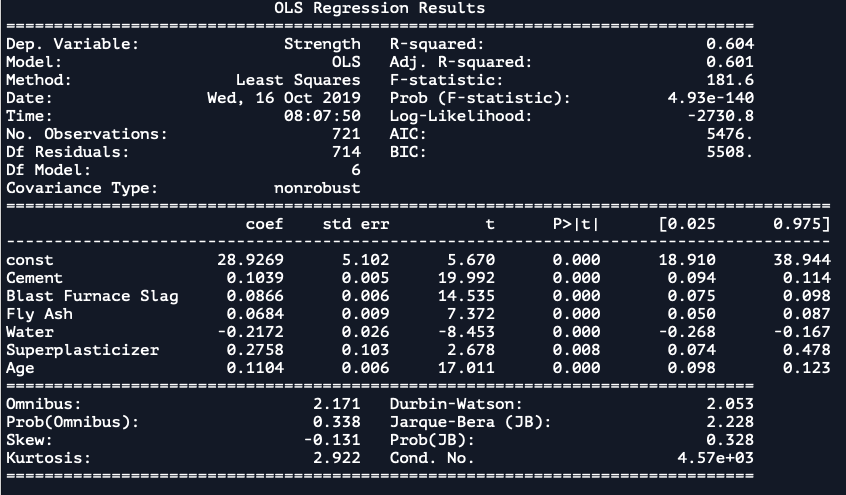
The dataset was split into ‘training’ and ‘testing’ of the ratio 70:30, respectively, using the sklearn model selection, train\_test\_split for the modelling.

## Modelling

### OLS - Ordinary Least Square testing

After the dataset was split into 70:30 ratio, ordinary least squares (OLS) regression model was used to estimate the relationship between one or more independent variables and a target variable. This method estimates the relationship by minimizing the sum of the squared errors. The table below shows a raw summary of OLS regression. 

The independent variables that have a P>|t| value greater than 0.05 were dropped from the analysis starting with the variables with the highest P value. The model is run till all the relevant variables are identified. The table below shows the final OLS Regression Results with only the applicable variables.



# 

### 

The R-squared and Adjusted R-squared output were 60.4% and 60.1%, respectively.

### Linear Regression

Linear Regression model from sklearn, linear model was run to identify the y-intercepts(b0) and the x-coefficients(b1 to b6) of the best fit line and a function for the strength of HPC is formulated. The function is shown below:

# 

# 

# 

# 

# 

The R-square values for the testing and training set were estimated as 63.73% and 60.41%, respectively. The Mean Square Error was predicted as 93.0845.

### k-fold Cross-Validation

A k-fold Cross-Validation was performed on the same data set for 3 folds. The mean cross validation score for the set k-value was observed as 54.848%.

## Analysis

The linear regression model gave an r-square value of about 63% and running a k-fold Cross-Validation gave a mean average of about 54.848%. Linear Regression turned out to be not an accurate prediction model to find the Strength evaluation of the concerte given the Cement, Blast Furnace Slag, Fly Ash, Water , Superplasticizer, Age. A different model should be attempted to see if we can get a better prediction model to find the Concrete Strength

### Decision Tree Regression (RTR)

A decision tree regression was performed on the same data set to see if we could achieve a model with a better accuracy.

The max depth level of DTR was set at 20 to see the performance of the tree. The DTR showed a better performance compared to the rest of the models. The accuracy plot of the training set in shown in the figure.

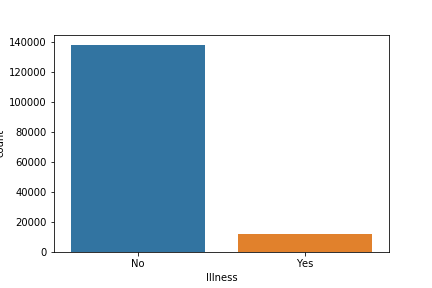
The R-square value of the testing set under DTR was observed as 80.37%.

## Conclusion

The Linear Regression model turned out to be not a better model to find the linear relationship between Strength of HPC against its independent variables. The Decision Tree Regressor proved to be a better contender in predicting the linear relation between the dependent and independent variables in this dataset with the R-squared value of 80.37%.

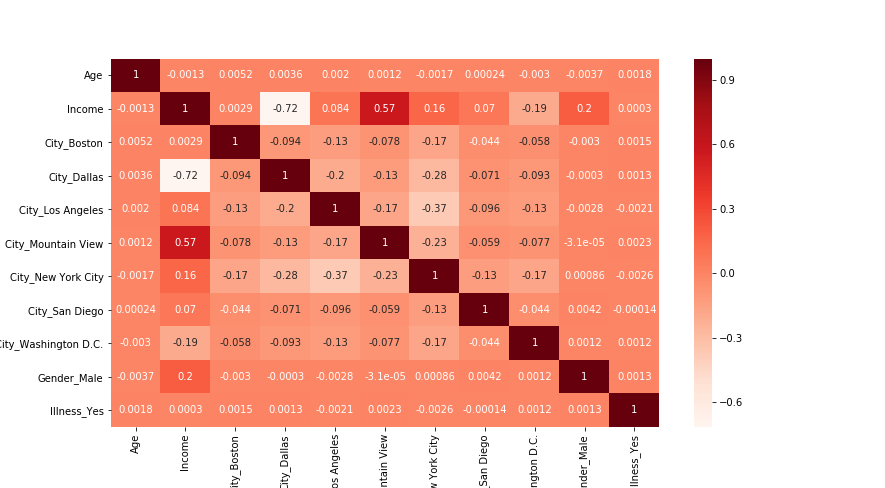
# Logistic Regression

## The Dataset

The dataset was procured from Kaggle.com. The dataset had a total of 150,000 instances. The total number of attributes were 5, out of which 4 were independent variables and 1 target/ dependent variable (refer figure). The input variables were: City: (Dallas, New York City, Los Angeles, Mountain View, Boston, Washington D.C., San Diego and Austin), Gender: Gender of a person (Male or Female), Age: The age of a person (Ranging from 25 to 65 years) and Income: Annual income of a person (Ranging from -674 to 177175).

## Data Analysis & Preparation

Initial analysis of the dataset revealed some key revelations that could not be discounted naming the target variable, Illness showed that 90% of the data comprised of patients who were not sick and only 10% were found to be sick (see figure). This showed data bias and there is a need to address this issue.

This data bias setback was taken care by splitting the dataset by the target variable where the number of nos where 10% of the total Illness No rows and concatenating that with the sort-separated Illness Yes dataset. The original dataset was hence reduced to 25,,926 instances. The total sick was 12139 and not sick was 13,787. The heatmap of the final dataset is shown above.

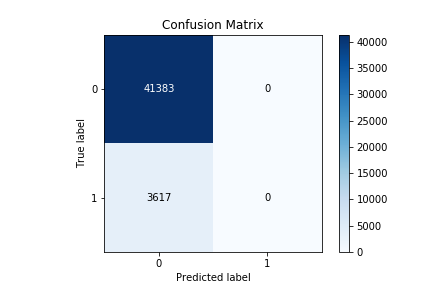
The dataset didn’t have any missing values. Some attributes were categorical and it was encoded using dummies. The dataset was finally split into ‘training’ and ‘testing’ of the ratio 70:30, respectively, using the sklearn model selection, train\_test\_split for the modelling.

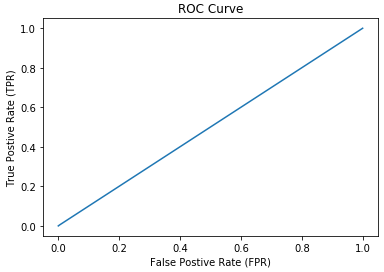
## Modelling

### Logistic Regression

Logistic regression is a statistical method for predicting binary classes. The outcome or target variable is dichotomous in nature. Dichotomous means there are only two possible classes. For example, it can be used for cancer detection problems. It computes the probability of an event occurrence.

Logistic regression from the sklearn package is called and was fitted to the training dependent and independent variables. The accuracies of the training and the testing set were observed to be 53.01% and 53.46%, respectively.

The classification report, confusion matrix and ROC curve of the model are shown in the figures below.



The AUC was observed as 0.498

### k-fold Cross-Validation

A k-fold Cross-Validation was performed on the same data set for 3 folds. The mean cross validation score for the set k-value was observed as 51.35%

## Analysis

The logistic regression model gave an accuracy score of about 53.46% and running a k-fold Cross-Validation gave a mean average of about 51.35%. Logistic Regression turned out to be not an accurate prediction model to predict if the Illness of the subject if and when the city, gender, age and income of the subject were provided. An alternately model should be figured out to measure the relationship between the independent and dependent variables.

### Decision Tree Classifier (DTC)

A decision tree classifier was performed on the same data set to see if we could achieve a model with a better accuracy.

The max depth level of DTC was set at 70 to see the performance of the tree. The DTC showed a better performance compared to the rest of the models. The accuracy plot of the training set in shown in the figure.

The accuracy score was still observed to be approximately 51.67%. The ROC plot remained the same. The DTC still turned out to be not an accurate model for prediction.

## Conclusion

The Logistic Regression model was observed to be not a better model to predict if the subject will be ill if given the above mentioned independent variables. The Decision Tree Classifier also failed to prove to be a superior model. This could maybe because of underfitting, the bias in the dataset or lack of a better model.