## Homework3-INF552-Rahul Ethiraj

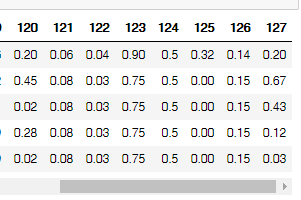
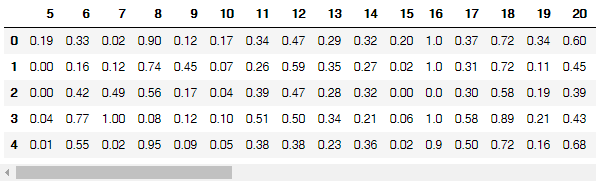
# **1.) The LASSO and Boosting for Regression**

**a)** **Importing data**

**b) Train\_test split data, imputation :**

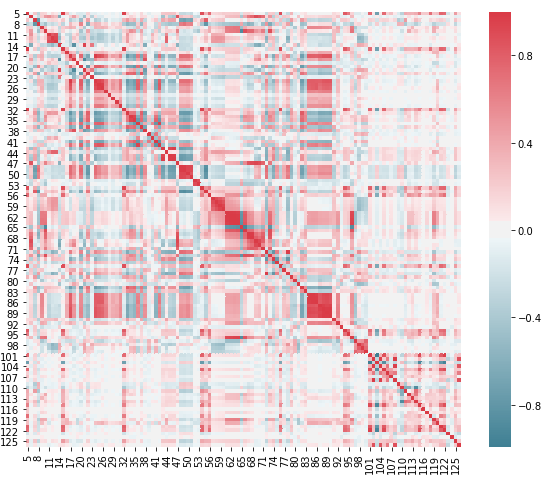
x\_train\_df

x\_test\_df

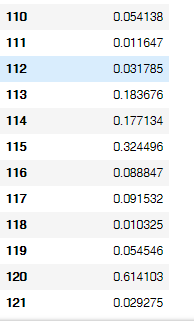
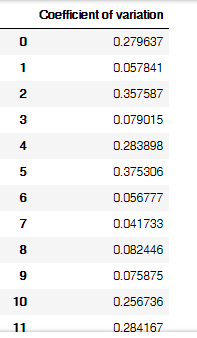


**c)**

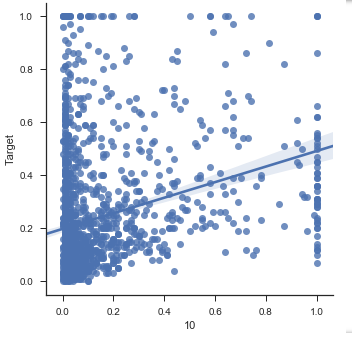
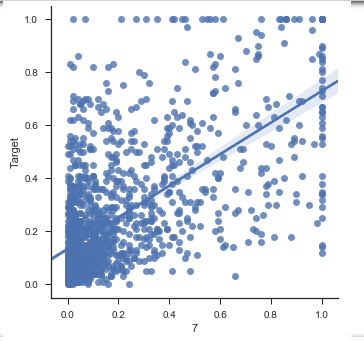
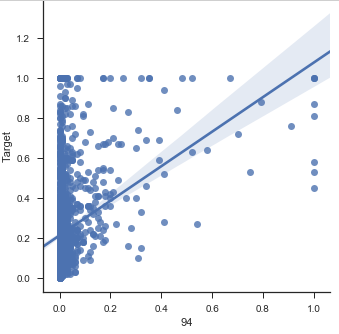
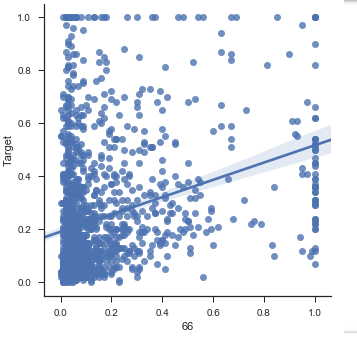
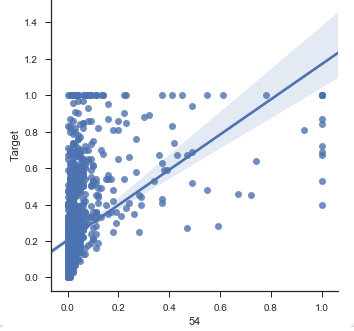
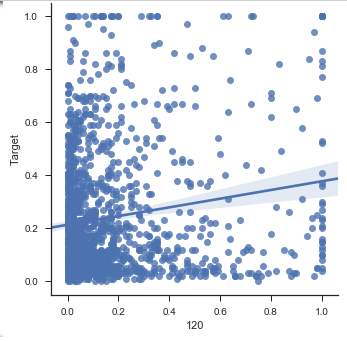
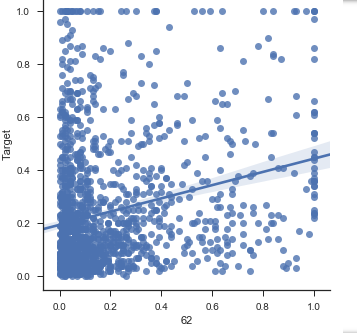
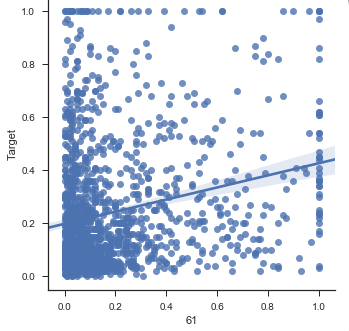
**(i) Correlation matrix:**

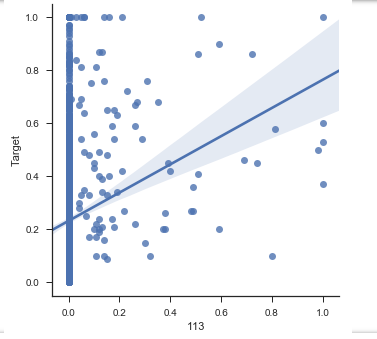
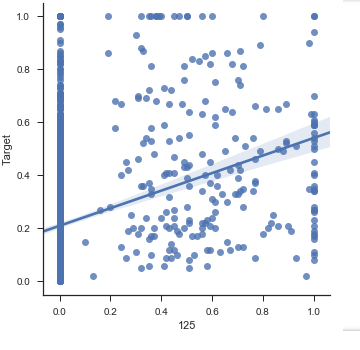
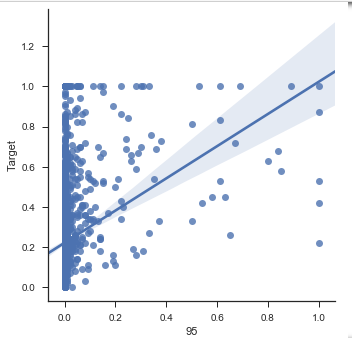


**(d) Coefficient of variation CV**



**(e) Scatter plots**



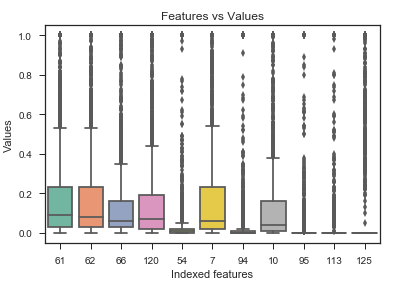


## Linear regression line has been fitted on all the scatter plots, which gives us the correlation.

## Scatter plots are similar to line graphs in that they use horizontal and vertical axes to plot data points. Scatter plots show how much one variable is affected by another. The relationship between two variables is called their correlation.

## Here, we see strong positive correlation between many of the features.

## (e) Box plots



## (f) Linear model using least squares

Mean squared test error : 0.01797697257164362

## (g) Ridge regression model[¶](http://localhost:8888/notebooks/Documents/INF%20ML%20Assignments/Homework3inf552/Homework3inf552.ipynb#(g)-Ridge-regression-model)

Mean squared test error : 0.017630963071701356

## (h) LASSO model

Unnormalized LassoCV :

====================

Mean squared test error : 0.017562937954615554

Total number of selected features : 45

Feature list by column numbers : [2, 7, 11, 13, 14, 15, 17, 18, 22, 23, 24, 25, 26, 28, 33, 38, 44, 45, 46, 48, 50, 59, 67, 68, 69, 71, 72, 74, 75, 76, 78, 82, 85, 86, 88, 90, 91, 94, 101, 102, 114, 115, 119, 120, 121]

Normalized LassoCV :

====================

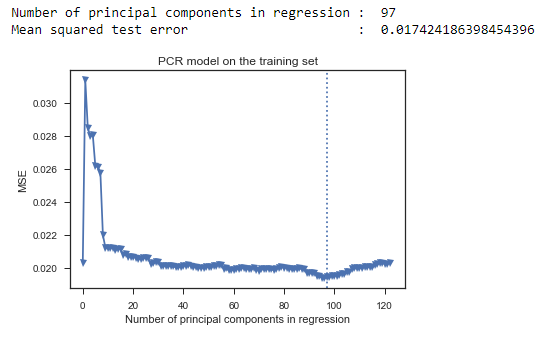
Mean squared test error : 0.017424186398454396

Total number of selected features : 32

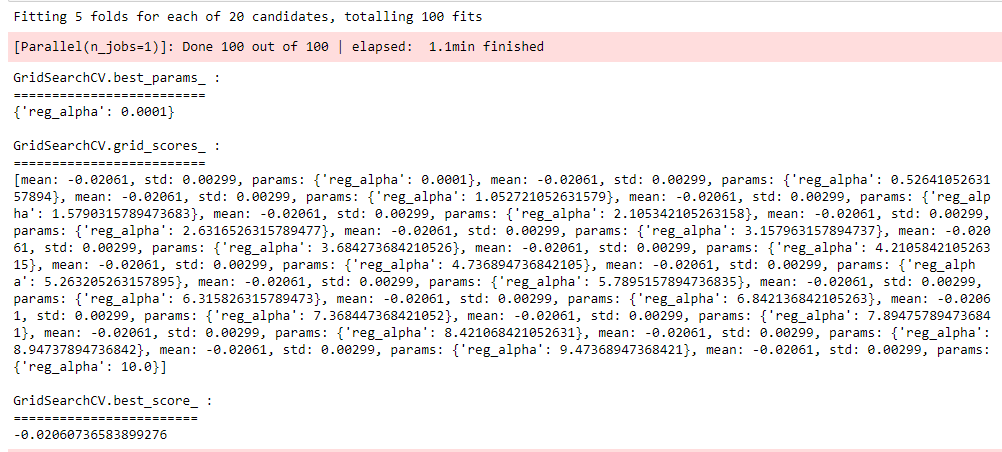
Feature list by column numbers : [2, 3, 7, 11, 15, 17, 18, 23, 24, 25, 38, 44, 45, 48, 50, 68, 69, 71, 72, 74, 75, 86, 88, 90, 91, 94, 99, 102, 104, 108, 119, 121]

The test error for both unnormalized and normalized is almost same, but the number of features is greatly reduced in the nomalized version of LassoCV.

## (i) PCR model



## (j) L1 penalized gradient boosting tree



# **2. Tree based methods**

## (b) Data preparation

### **i.) Train\_test split data, imputation**

The techniques used for dealing with missing values are known as imputation techniques. Examples include mean, median, mode, ffill, bfill, pad etc.

## 

## ii.) Coefficient of variation CV

## 

## (iii) Correlation matrix

## 

## (iv) Scatter plots

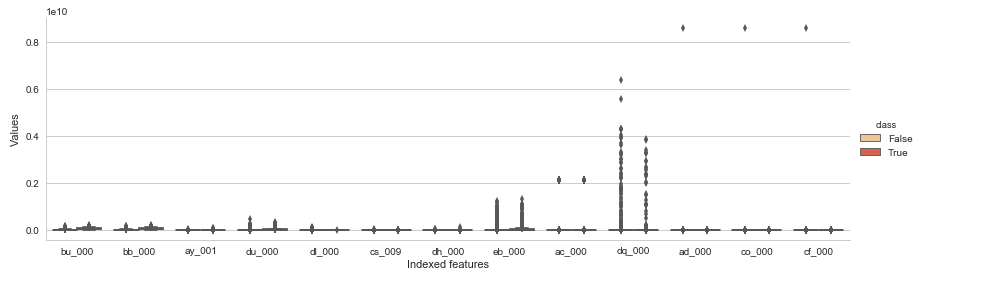
## 

Linear regression line can be fitted on all the scatter plots, which gives us the correlation.

Scatter plots are similar to line graphs in that they use horizontal and vertical axes to plot data points. Scatter plots show how much one variable is affected by another. The relationship between two variables is called their correlation .

Here, we see strong positive correlation between few of the features.

## (iv) Box plots



## (v) Imbalance

## 

## (c) Random forest for training data with imbalance

Out-of-bag error estimate : 0.008633333333333382

Training error : 0.00048333333333333334

## 

## Random forest for test data with imbalance:

Out-of-bag error estimate : 0.009383333333333299

Test error : 0.009

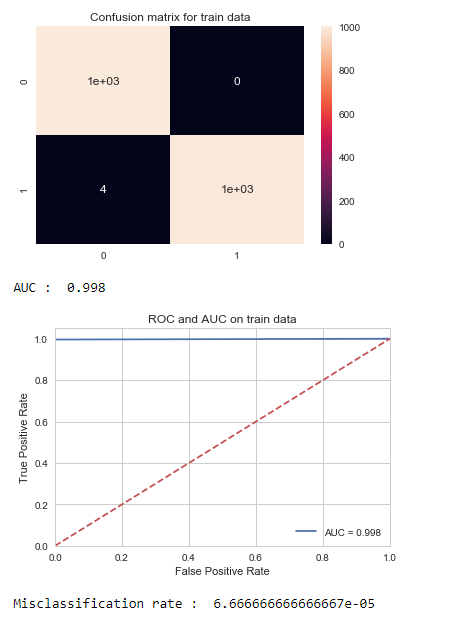
## 

## (d) Random forest for train data with no imbalance

Class imbalance can be addressed in random forests by balancing the classes, resampling, downsampling and upsampling.

Out-of-bag error estimate : 0.06499999999999995

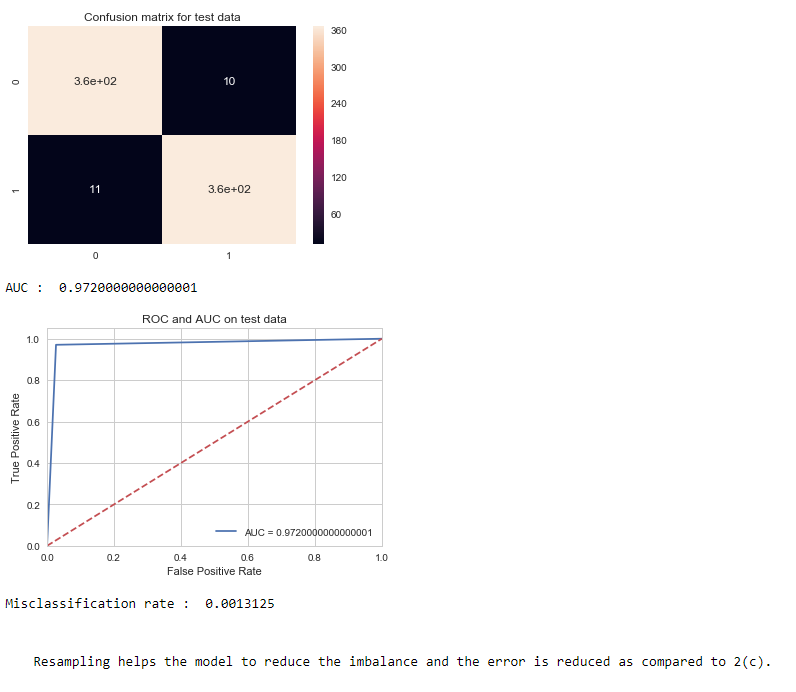
Training error : 0.002



## (d) Random forest for test data with no imbalance

Out-of-bag error estimate : 0.06699999999999995

Training error : 0.028



## (e) Model Trees for training set

## 

## (e) Model Trees for test data

## 

## (f) SMOTE Model Trees for train data

## 

## 

## (f) SMOTE Model Trees for test data

## 

Clearly, using SMOTE has proven beneficial as the results are surprising. AUC has increased when compared to uncompensated case.

## 3.) ISLR 6.8.3, 4.) ISLR 6.8.5

## 

## 5.) ISLR 8.4.5

## 6.) ISLR 9.7.3

## 