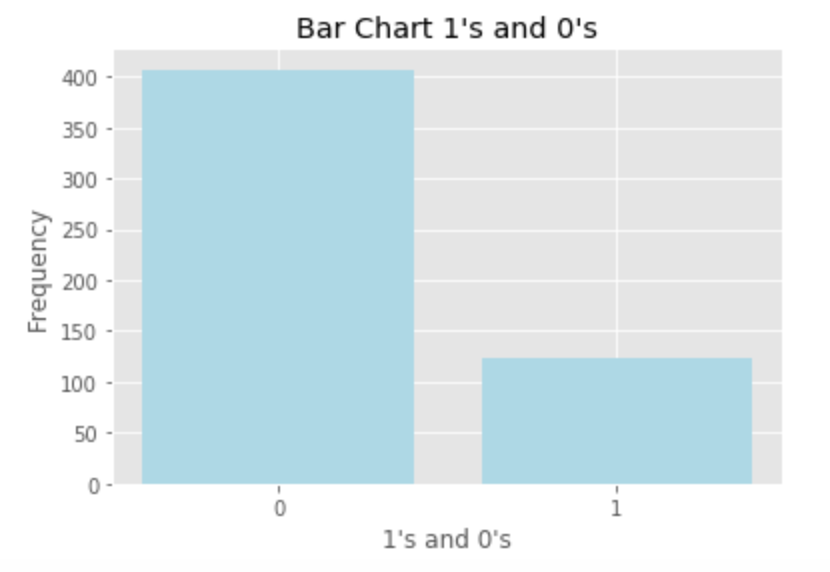
**NLP Take Home Exam Documentation**

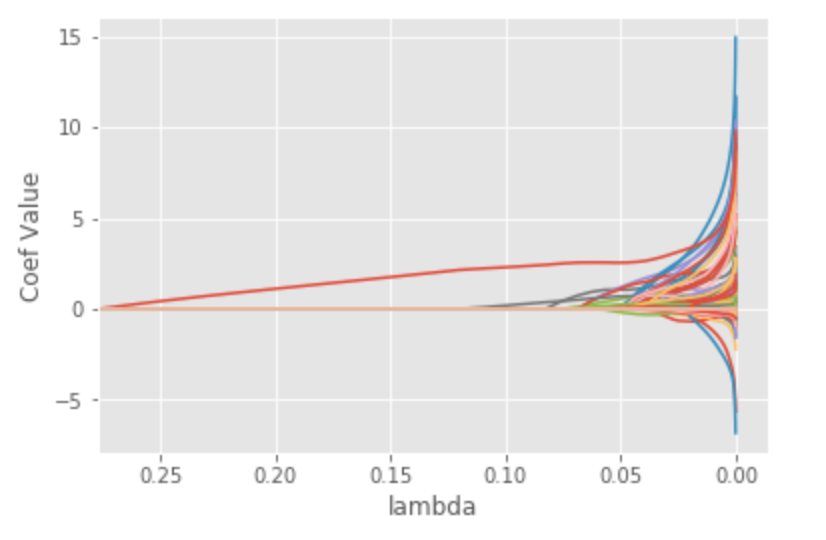
Approach I followed in the Assignment

1. Given Datafile is in the format of text file. I read the file and converted it into a dataframe using pandas builtin function read\_csv with the help of a parameter delimiter = ‘\t’
2. Then I separated the response column(output) and variables column(input)
3. Then I checked the shape, info, summary statistics, Looked at Head and Tail of two Dataframes response and variables
4. Then I checked for NaN values in both the variables and response Dataframes I found there are No NaN values in both Dataframes from this Now, I’m sure that my dataset is clean
5. Then I Looked at the response column to see How Many 0’s and 1’s are present in the response column 0’s are almost 3.3(approx) times more than 1’s this tell that my dataset is unbalanced.

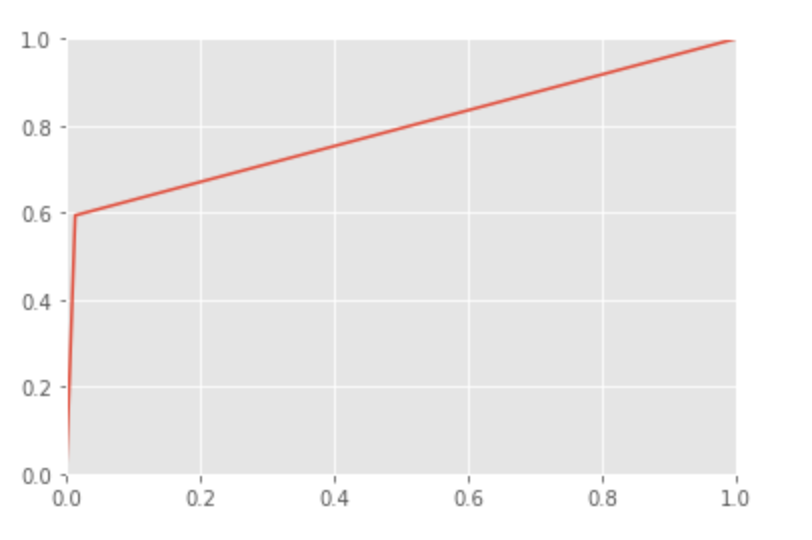


1. Then I also Looked at the response column of response dataframe how the 0’s and 1’s are arranged. I noticed that there are sequence of 1’s and sequence of 0’s. There are there are many (16562 features) (so I also want to do feature selection)
2. I split the data using train\_test\_split method by setting the parameter shuffle = True from sklearn library model\_selection by choosing the different train & test sizes and calculated the training and validation errors to check whether my model is overfit or underfit
3. My dataset is unbalanced, and also my model doesn’t want to overfit, I also want to do feature selection, I also want to do regularization because of unbalanced data set. Considering all these features I selected the LogitNet From glmnet and imported the LogitNet and setting alpha to one gives a Lasso Model LogitNet also gives best lambda value and best parameters of the model. I used Lasso Model to predict the accuracy for this Dataset
4. I also measured mean\_squared\_error, generated confusion matrix, classification report of my model on different train and test sizes
5. I also split my data into 0.6% train, 0.2% test, 0.2% validation data by keeping the test\_size data to 0.2 just wanted to split into three data sets and wanted to see what happens
6. I also did 5 -Fold cross Validation to measure the skill of my model Plotted Roc Curve’s for every iteration and calculated the mean for accuracy\_score, validation errors, training errors and AUC

**Graph of calculating minimum Lambda Value**



ROC CURVE



Conclusion: In my case since I have an unbalanced data, the accuracy is not a good metric and can make it better by using other metrics and this is also not the best model for good performance there are also other models to get good performance