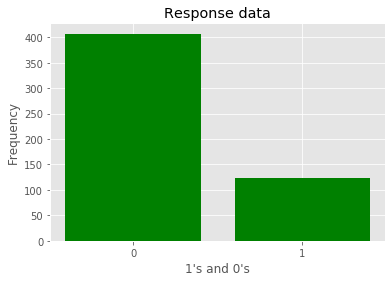
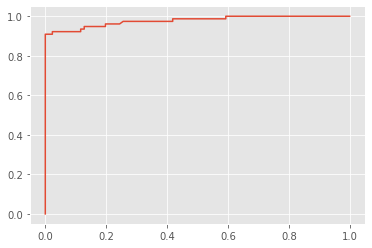
DS casestudy Take Home Documentation

1. The given dataset is in the form of a text file. As the data was tab-separated, It was loaded into a dataframe using pandas read\_csv.
2. Firstly; shape,head, statistics of the data was checked.
3. Secondly, searched for the missing values and could not find any.
4. Although they were no missing values, expected some redundancy between the columns and checked for them.
5. Removed the found redundant columns.
6. By using describe () on the dataframe found the summary of the data.
7. From the statistics found columns with all 0’s and all 1’s and removed those columns.
8. Then separated the response variable from the predictor variables.
9. Statistics of response variable implies that the data is unbalanced. They are high number of zero’s.

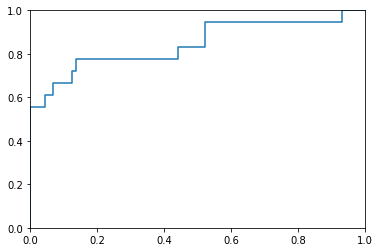


1. From Correlation plot, it was evident that they are many features correlated to one another both positively and negatively. It also seems that the variables are interdependent.
2. From the above analysis, we see that the data we are dealing with is wide and not balanced.
3. Then split the prepared 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 the model is overfitting or underfitting.
4. After looking at the above two problems, initially implemented RandomForestClassifier by using sklearn with RandomizedSearchCV. As the random forest would handle large data with higher dimensions greatly. Class imbalance was controlled by Smote(over\_sampling)
5. Mean squared error, confusion matrix, accuracy, and classification report for different test sizes has been calculated. Obtained results were good.
6. But in real-world scenarios where we work with real-time health data, we would not be using any of the oversampling techniques. As the data is unbalanced and model should not overfit the data. Logistic regression with lasso regularization was other model which was implemented.
7. Again, mean squared error, confusion matrix, accuracy, and classification report for different test sizes has been calculated.
8. Validated both the models by dividing the dataset into 60% train, 20% test and 20% validation and calculated the errors.
9. Also performed 5-fold cross-validation, so as to measure the performance of the models plotted roc curve for each iteration.
10. Finally mean accuracy score, mean validation errors, mean training errors and mean AUC of both the models has been calculated.

ROC for Random Forest Classifier



ROC for Logistic Regression



Conclusion:

As the given data is unbalanced, accuracy is not a good metric. Based on the training error and the test error the model is neither overfitting nor underfitting. Model 1 performs better than Model 2 as the data considered is balanced.

Both the models have low root mean squared error values and low false negatives values.

Future Work:

In model 1 instead of oversampling, under sampling can be used in the health industry when there is large amount of data.