Lead Scoring Case Study Summary

1. **Data Cleaning:**

   First we checked for missing and duplicate data, there were no duplicate data but there was considerable amount of missing values, we eliminated columns with more than 2500 missing values and rows with any missing values and we were left with around 98% of the original data. And the next step in data cleaning was outliers treatment, we found that there were some significant outliers in 3 columns we treated them and reduced the outliers significantly without losing too much data.

2. **Creating Dummy Variables/mapping of categorical variables:**

   As we know there were lot of categorical variables we used pandas get_dummies to create dummy variables and we used map function to map categorical variables which had only 2 categories yes/no in them.

3. **Splitting and Scaling the data:**

   As we know its best practice to scale the data before creating model, we used standardScaler module to scale the data and we spit the data into training data(70%) and testing data(30%).

4. **Model Creation:**

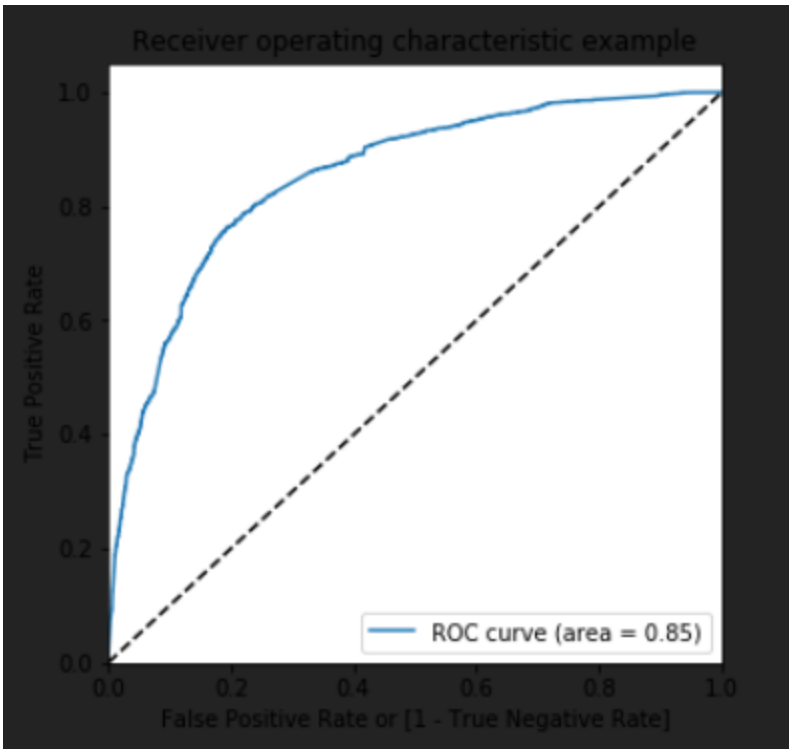
   Before we go ahead with model creation we had to eliminate some features as we checked there were significant correlations between the variables. So we decided to retain around top 25 features using RFE. Later we built the model using statsmodels api .And found that still there are some variables which are still correlated, so we used variance_inflation_factor module to check the vif score and eliminated manually the top variable and rebuilt the model, we repeated this steps until all the variables had vif score below 5.And we also checked the 'p-value to check If the variable is significant or not, we again manually eliminated the variables which had high 'p-values ' and rebuilt the model, continued the process till there were no insignificant variables present in the model. After which we had a stable model .

5. **Model Evaluation:**

   After building the stable model ,we made the prediction and checked the statistics of the mode. We got the following statistics:

   1. Accuracy : 0.7864
   2. Sensitivity : 0.6600
   3. Specificity : 0.8693

   And we drew the ROC curve and found that Area under the curve is around 0.85 which is very good.



6. **Making Prediction on test data:**

   After training and evaluation of the model we went ahead with test data to check the stability of the model y predicting on the test data. We found that it also has very good statistics ,which is very similar to the train data, from which we can conclude our model has not over fitted the data to make prediction. Below are the statistics on the test data.

   1. Accuracy : 0.7929
   2. Sensitivity : 0.6796
   3. Specificity : 0.8594