Problem statement:

An education company named X Education sells online courses to industry professionals. On any given day, many professionals who are interested in the courses land on their website and browse for courses.

The company markets its courses on several websites and search engines like Google. Once these people land on the website, they might browse the courses or fill up a form for the course or watch some videos. When these people fill up a form providing their email address or phone number, they are classified to be a lead. Moreover, the company also gets leads through past referrals. Once these leads are acquired, employees from the sales team start making calls, writing emails, etc. Through this process, some of the leads get converted while most do not. The typical lead conversion rate at X education is around 30%.

The company requires you to build a model wherein you need to assign a lead score to each of the leads such that the customers with higher lead score have a higher conversion chance and the customers with lower lead score have a lower conversion chance. The CEO has given a ballpark of the target lead conversion rate to be around 80%.

1.Exploratory Data Analysis:

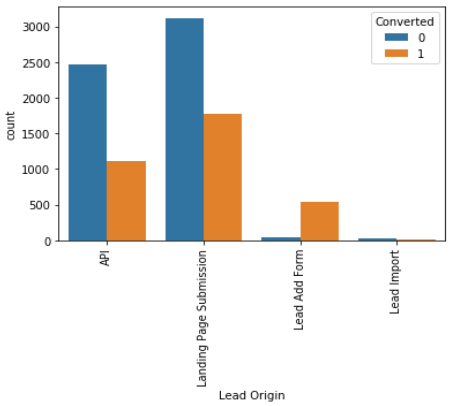
Data understanding and Data cleaning:

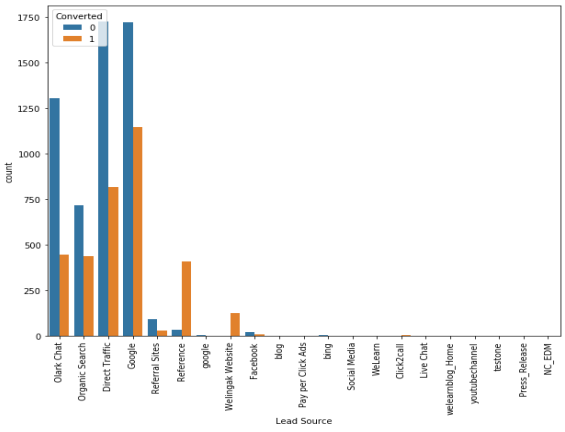
We noticed the shape of the data to be (9240,37) and observed missing values present in the data.

Firstly, we converted all the select values to Nan values and then we dropped the columns with more than 70% missing values. We also dropped the columns 'Asymmetrique Activity Index','Asymmetrique Profile Index','Asymmetrique Activity Score','Asymmetrique Profile Score' as they don't have any pattern and has 45% of missing values. And we have imputed the remaining missing column values with mode of the column. Finally we did not lose significant amount of data and are left 98% of data which is good to go.

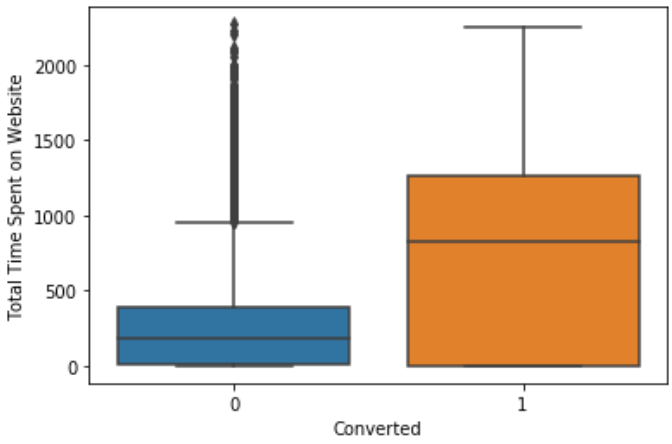
2.Univariant Analysis:

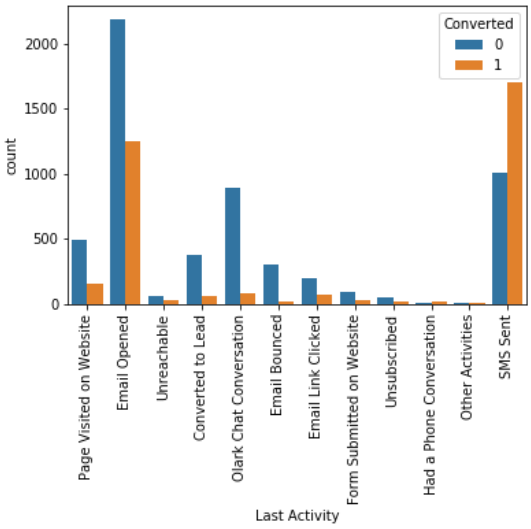
We started plotting count and boxplots for each variable to identify significant points to be noted and found few as below.





From the above two plots it is noticeable that more conversions are from “landing page submission lead origin” and “reference” lead source.





From these above two plots it is observed that most conversion rate happened when the user spent most time on website and if the last activity was of SMS sent.

Similarly, we analyzed all the variable by plotting the count plots and we treated the outlier by observing the boxplots for some variables by taking in to the account of their quantiles.

3.Dealing with Dummy Variable/converting of categorical variables:

We performed label encoding technique on the variables and mapped “Yes/no” columns with 0 and 1.

Splitting and Standardizing the data:

We split the data to train(70%) and test(30%) data, then performed scaling using StandardScaler.

4.Model Building:

As we have so many features to work with, first we will eliminate some of them by using RFE technique and finally left took 30 variables. Later we imported statsmodel api and built a model with these 30 variables. We analyzed the summary taking p values in to consideration. And we also analyzed the VIF values of these columns. We stared eliminating the variables manually with high p and high VIF one by one, then rebuilt the model. This process is carried out until the p value is less than 0.05 and VIF less than 5. Finally we arrived at a model with desired p and VIF values.

5.Model Evaluation:

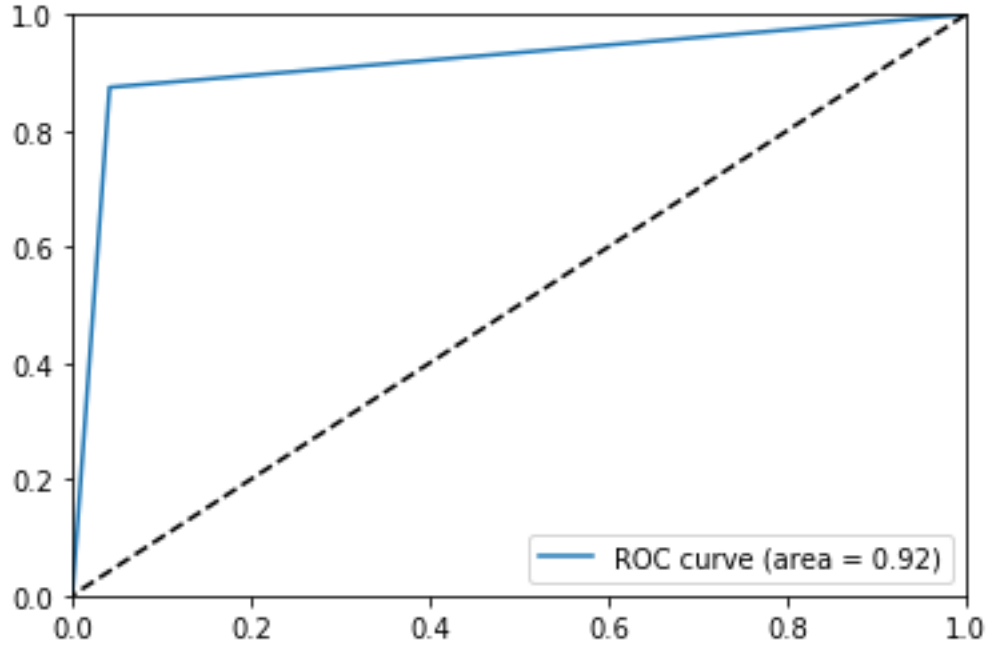
We made prediction after building the final model and calculated the stats parameters such as sensitivity, specificity, accuracy.

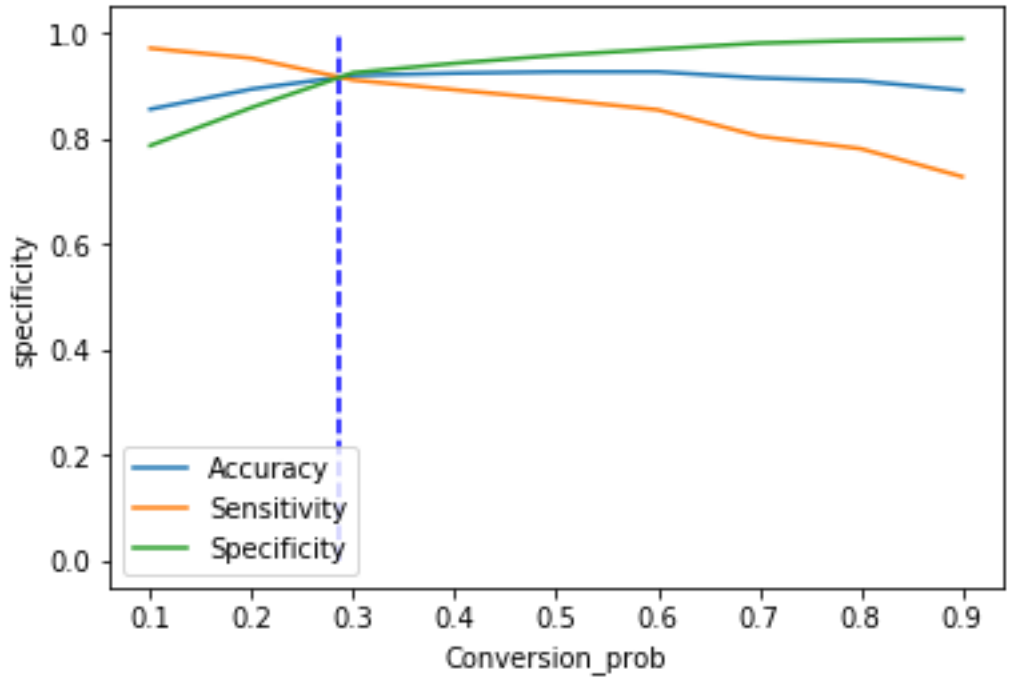
Confusion matrix with cutoff 0.5 comes out to be as

[[3811, 166],[297, 2077]]

Accuracy score = 0.92

ROC curve with area under curve value as 0.92.





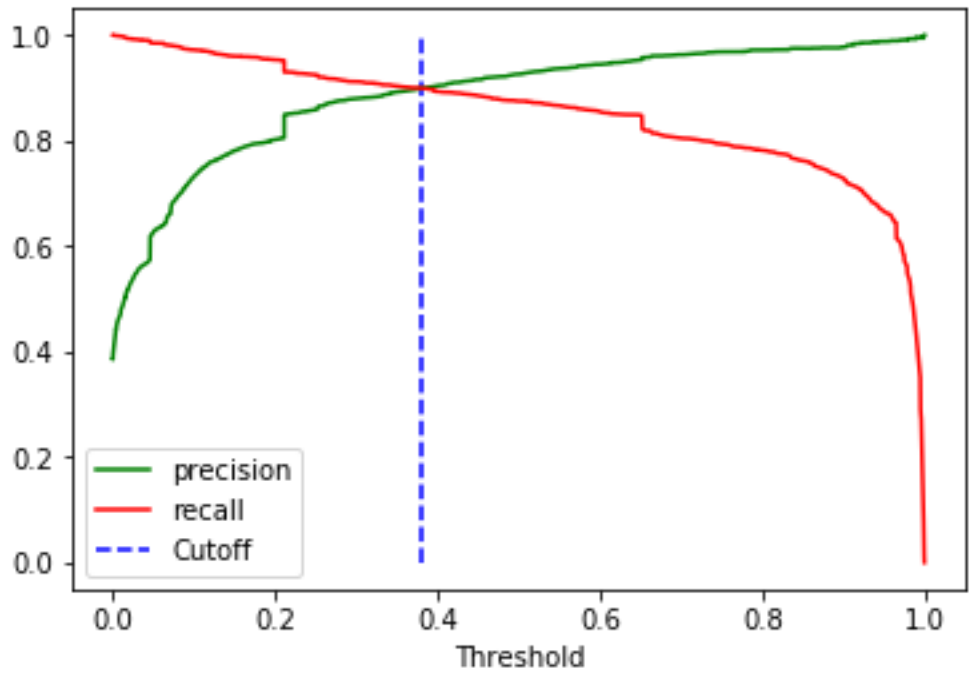
From the above curve 0.285 is the optimal point where we have sensitivity and specificity are high. Hence we chose the point as 0.285 and made predictions again which resulted in 0.92 accuracy.

Sensitivity = 0.91

Specificity = 0.92

precision score = 0.876
recall score = 0.914

Precision-recall plot.



6. Predictions on Test Data set:

Finally, we predicted the test data on the trained model which resulted in the following parameters.

Sensitivity =0.92
Specificity = 0.918
Accuracy score= 0.919

By looking at the above parameters we can conclude that built model is good and stable.