Lead Scoring Logistic Regression

Team:-

Elumalai C

Hariharan J

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.
- X Education wants to find out the leads that are most likely to convert into paying customers. The company requires to build a model and 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, in particular, has given a ballpark of the target lead conversion rate to be around 80%.

Steps involved in the Analysis

1)Reading and Understanding the dataset.
2)EDA Building
    Removing Missing/Null values
    Outlier Treatment
    Rescaling
    Converting some binary variables
3)Data Preparation
    create dummy features
    Perform train-test split
    Perform Scaling
4)Model Building
5)Model Evaluation
    ROC Curve
    Optimal Cut off Point
    Predictions on the test set
6)Conclusion

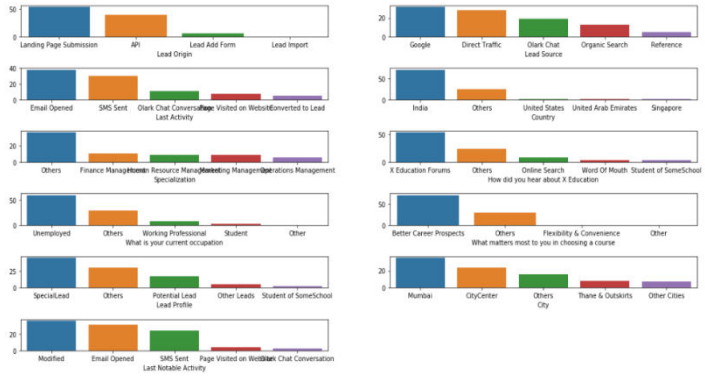
Reading and Understanding the dataset

- Importing the data into pandas data frame.
- Inspecting the data frame using inspections methods like Info, shape, describe etc.
- Handling Select Level in Categorical Columns

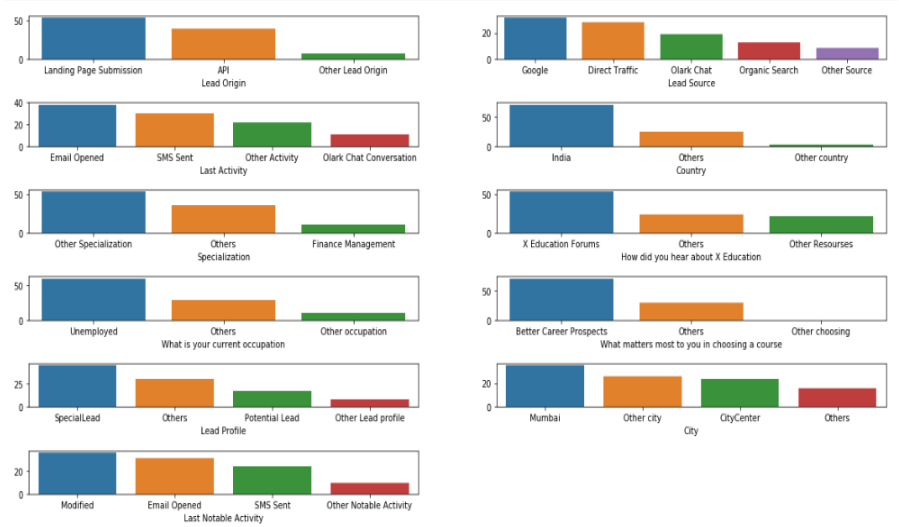
EDA Building

- There was no duplicate records in the given data set.
- Unique Categories in the Categorical Columns

Before

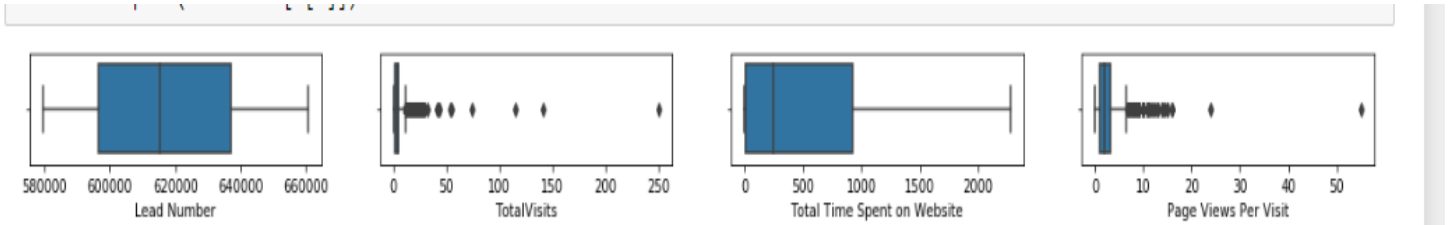


After

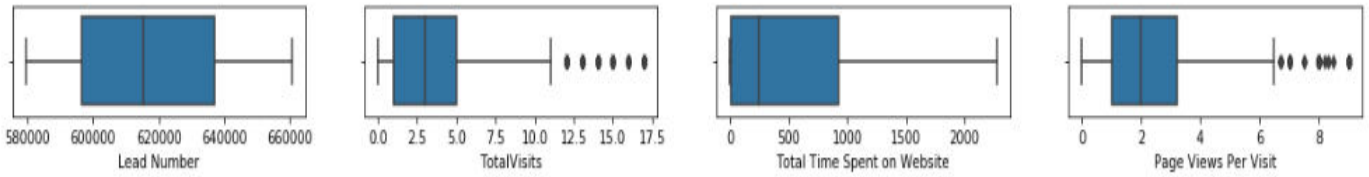


EDA Building

- Outliers
  - Box plot is used to find out the outliers of the variables



- Capping method is used to fix the outliers.

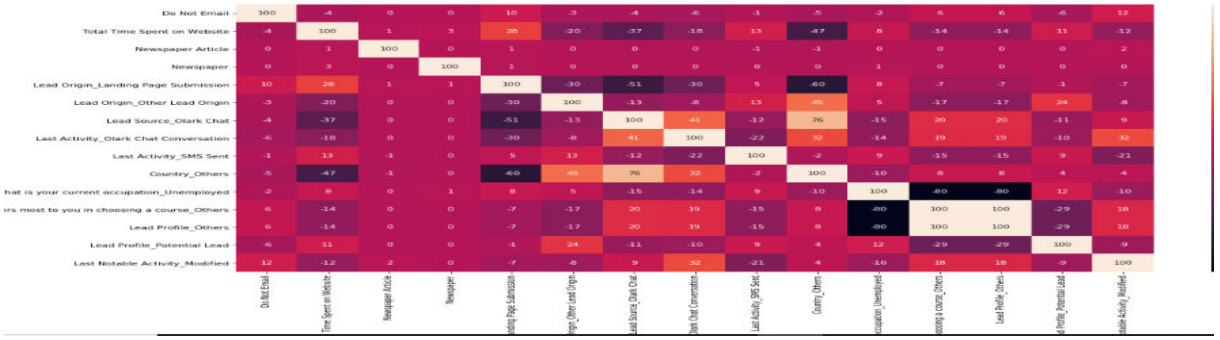


Data Preparation

- For all categorical variables with multiple levels, created the dummy features (one-hot encoded)using the pandas get dummies functions.
- Data is split into train and test set where 70 percentage of data in train set and 30 percentage in test data
- Using Standard Scalar the variables are scaled.

Model Building

- Using Stats model build the first model and analyzed the coefficient of each variable.
- Logistic Regression model is used to find out the top 15 variables to build model.
- Heat Map is build to show the correlation of the variables.

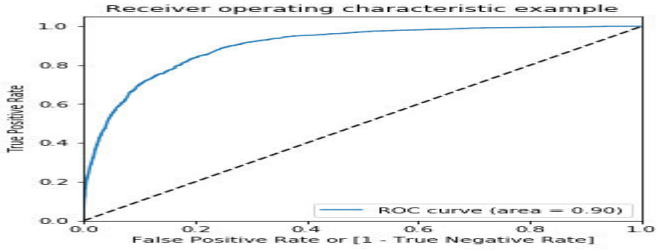


Model Building

- Using stat model the model is build for the top 15 variables.
- Confusion matrix is build for the model and analyzed the Accuracy, Sensitivity, Precision, Recall.
- VIF score is calculated for the variables and eliminated the variables whose value is more than 5.

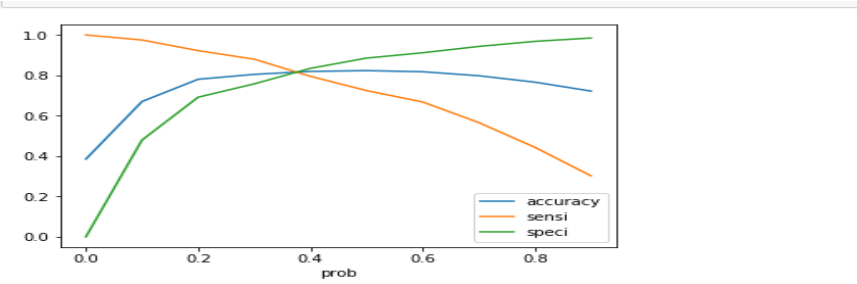
Model Evaluation

- ROC Curve.
  - It shows the trade off between sensitivity and specificity (any increase in sensitivity will be accompanied by a decrease in specificity).
  - The closer the curve follows the left-hand border and then the top border of the ROC space, the more accurate the test.
  - The closer the curve comes to the 45-degree diagonal of the ROC space, the less accurate the test.



Model Evaluation

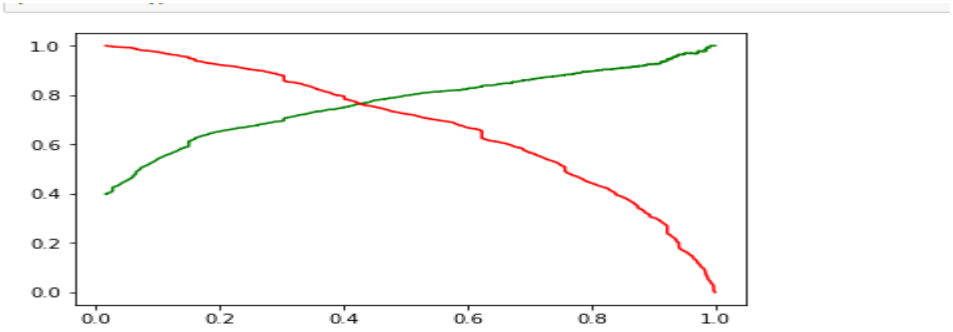
- Optimal Cut off Point:-
  - For the different probability from 0.0 to 0.9 calculated the Accuracy, Sensitivity, Precision, Recall.
  - Line chart is plotted for Accuracy, Sensitivity, Precision, Recall and found the actual cut of point.



From the curve above, 0.4 is the optimum point to take it as a cut off probability.

Model Evaluation

- Precision and Recall curve
  - Precision and Recall curve is calculated for thresholds using the precision_recall_curve() function that takes the true output values and the probabilities for the positive class as output and returns the precision, recall and threshold values



Model Evaluation

- Predictions on the test set
  - Using the train model the values are predicated for the test data set.
  - Accuracy, Sensitivity, Precision, Recall is calculated for the test modal.
  - Summary:-

Summary of the train modal

Evaluation Result

Accuracy :-82 %

Sensitivity :- 79%

Specificity :-83 %

Precision :-75 %

Recall :-79 %

Summary of the Test modal

Evaluation Result

Accuracy :-81 %

Sensitivity :- 78 %

Specificity :-83 %

Precision :-72 %

Recall :-78 %

Conclusion

- From the train and test model evaluation - we can infer that all Accuracy,Sensitivity & specificity are all close to each other- which determines the ROC cut off - 0.4 is good.

- And also we can see all the variables P values ate <0.5 which are more significant.
- And also VIF values of all the variables are < 5 , which is also more significant.

- _ve coefficient values variables - has less percentage of chance of probability of lead getting converted.
- +ve coefficient values variables - has high percentage of chance of probability of lead getting converted.
- High Chance of getting lead to converted variables are listed below with their co-efficient value.

- Lead Origin_Other Lead Origin : Coefficient values = 3.1409 times of Lead origin probability of lead getting converted.
- Newspaper Article : Coefficient values = 1.7885 times of Newspaper Article probability of lead getting converted.
- Lead Profile_Potential Lead : Coefficient values = 1.6258 times of Lead origin probability of lead getting converted .

- Overall, the Company should concentrate on the above variables to increase the number of leads percentage.

End of the Analysis-Thanks.