Lead Scoring Assignment

Using Logistic Regression

- Subramanyeswara and Suresh

Problem Statement

An X Education need help to select the most promising leads, i.e. the leads that are most likely to convert into paying customers. The company requires us 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, in particular, has given a ballpark of the target lead conversion rate to be around 80%.

Steps Followed

Data Cleaning and Data Manipulation

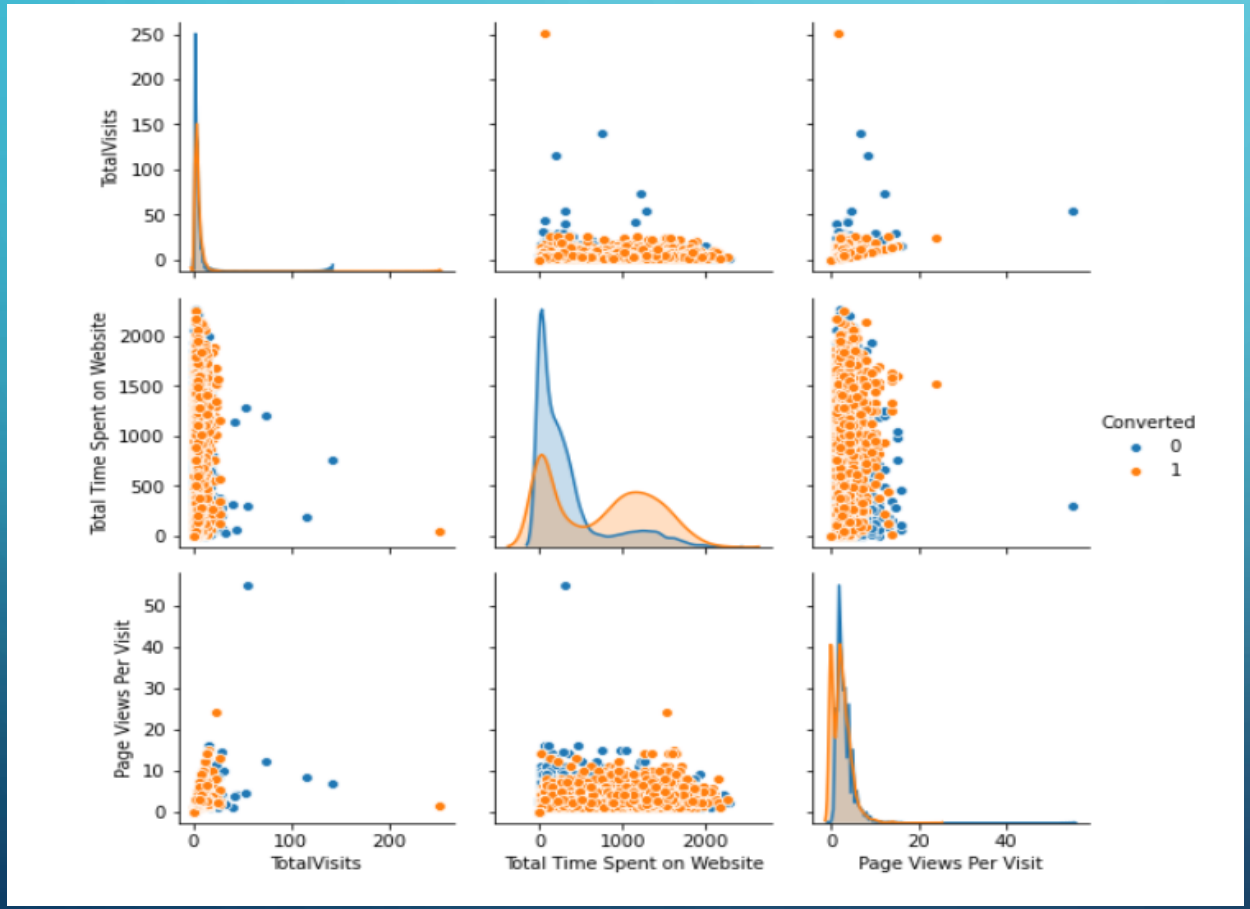
- Checking for Number of missing values and replacing ‘select’ with NULL values
- Removing columns which has 70% of Missing Values
- Imputation of Missing Values if missing values are less than 70%

Exploratory data analysis

- Univariate Analysis
- Bi variate analysis

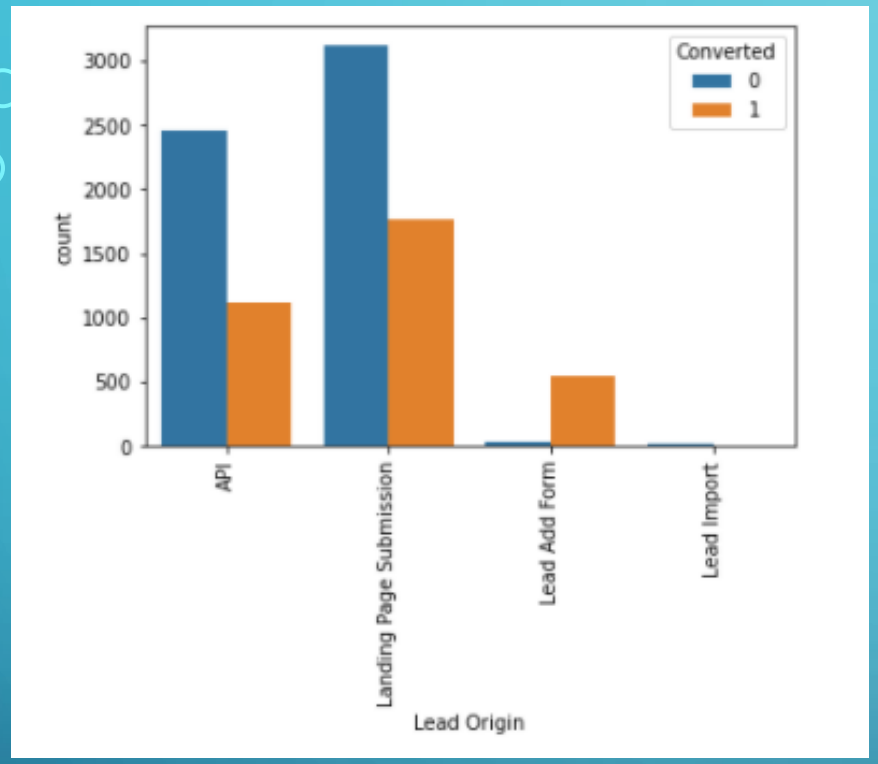
Scaling the Numerical Variable
Creation of Dummy Variables for Categorical columns
Splitting the data like Test and Train data set
Model Building : Logistic Regression Method
RFE for Feature selection
Validating and Evaluating the model

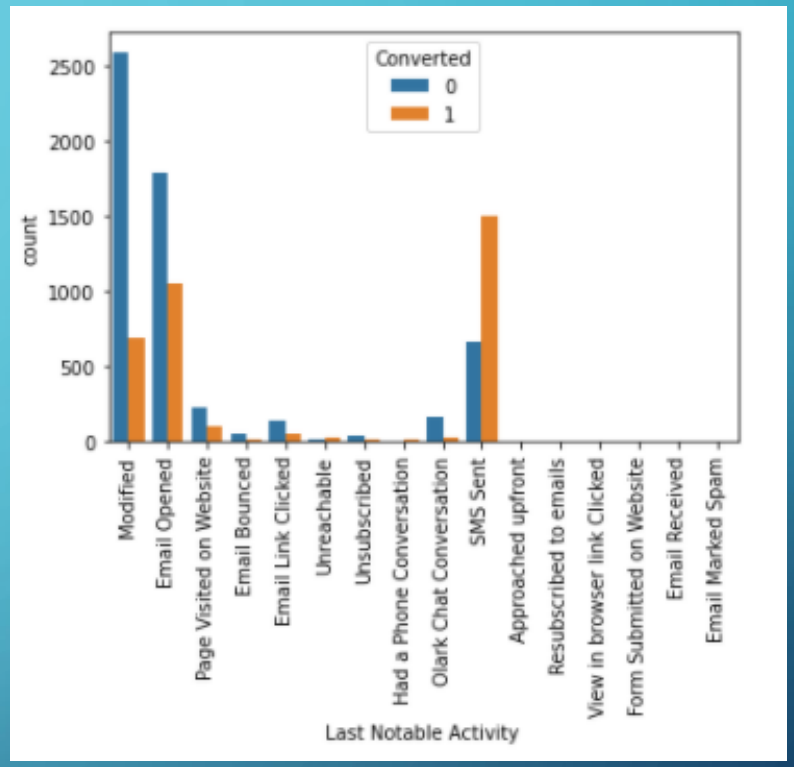
PAIR PLOT BETWEEN NUMERICAL VARIABLES



- Total time spent on website as well as page views per visit has higher lead rate

Results of Univariate Analysis



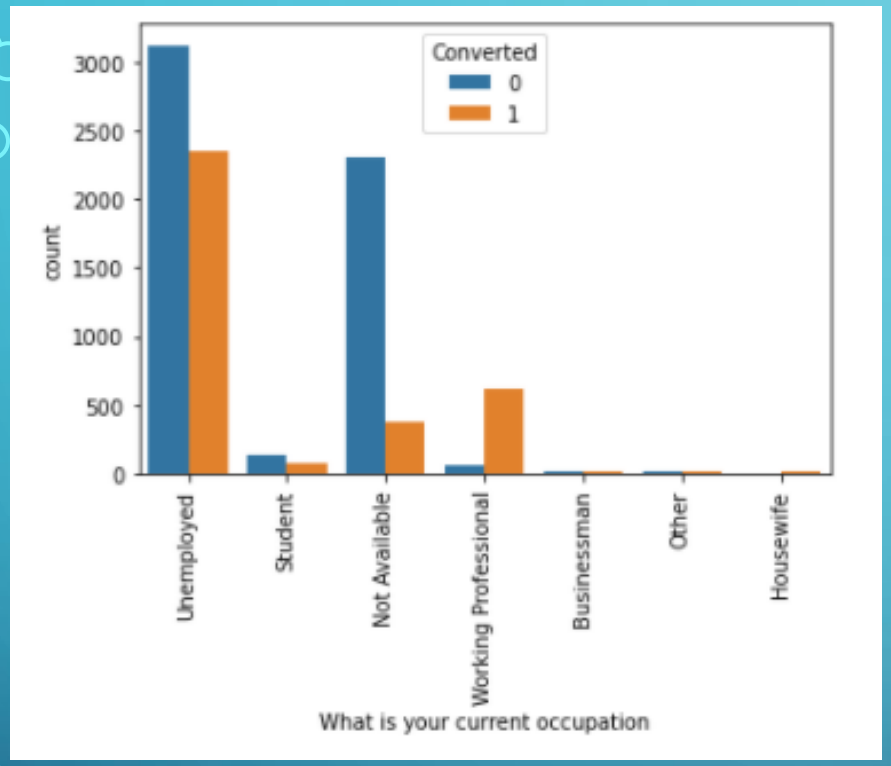


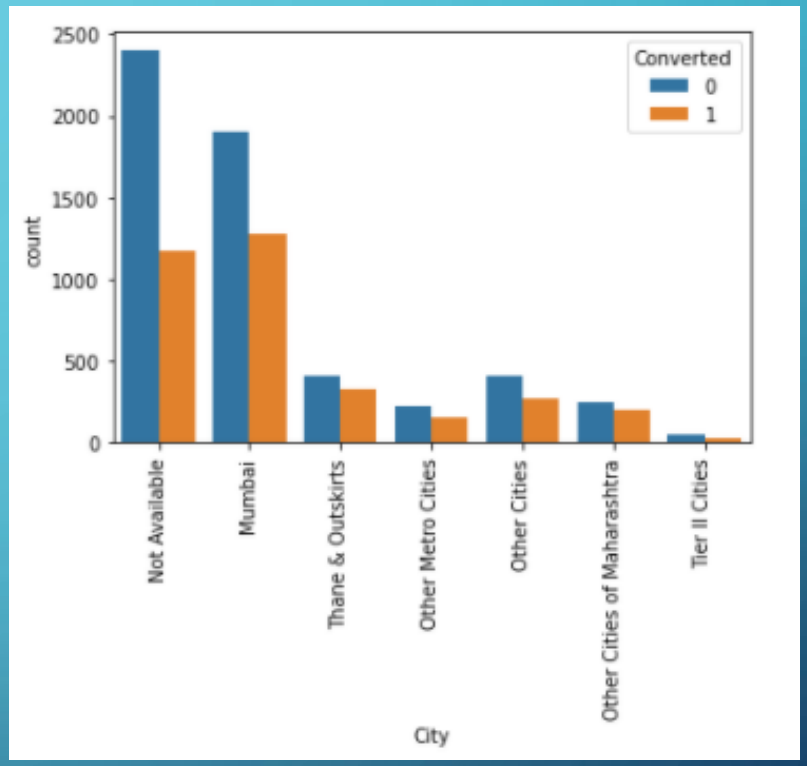
Inferences:

Success rate is high if the lead origin is Lead Add form.

If last notable activity is SMS Sent, then those leads are getting converted

Results of Univariate Analysis





Inferences:

People with current occupation is unemployed are approaching more but Success rate is little less and working professional success rate is high so better to reach out to working professionals.

People with Mumbai as City are approaching more but success rate is less , overall when compared with other cities Mumbai is higher

Bi Variate Analysis

Time spent on website is highly correlated with Converted, that means people who spent more time and high likely to be converted

VIF CALCULATION

- Top 15 features are selected
- What is your current occupation_unemployed as

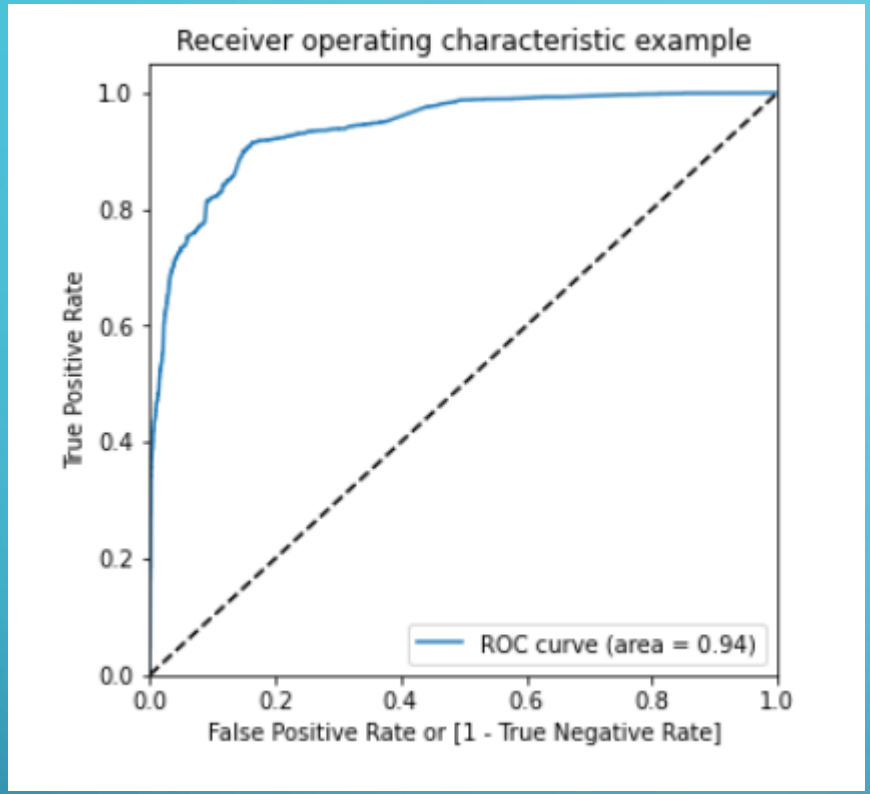
High VIF , this will be highly correlated with other

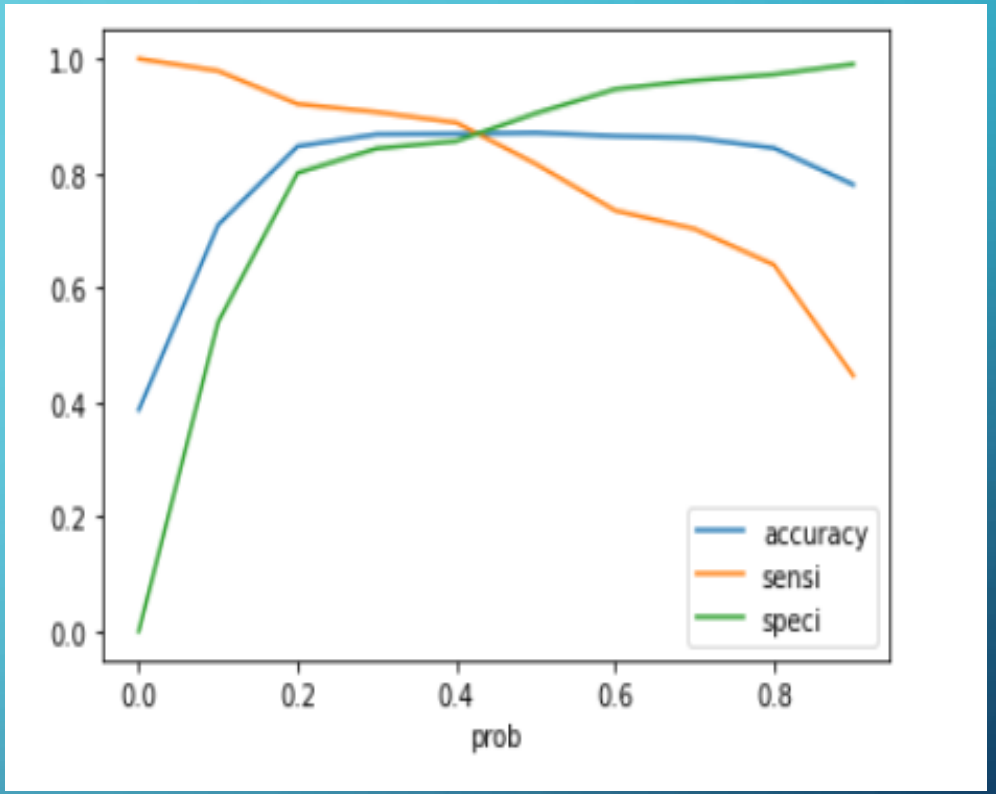
Variables hence dropped this feature

	Features	VIF
12	What is your current occupation_Unemployed	3.96
6	Tags_Will revert after reading the email	2.78
0	Total Time Spent on Website	1.99
5	Tags_Ringing	1.80
10	Last Activity_SMS Sent	1.68
13	What is your current occupation_Working Profes...	1.68
4	Tags_Not Available	1.56
14	Last Notable Activity_Modified	1.56
8	Lead Quality_Worst	1.42
2	Tags_Closed by Horizzon	1.23
11	What is your current occupation_Student	1.18
1	Tags_Busy	1.16
7	Tags_switched off	1.16
3	Tags_Lost to EINS	1.10
9	Lead Source_Welingak Website	1.10

	Features	VIF
0	Total Time Spent on Website	1.53
9	Last Activity_SMS Sent	1.45
12	Last Notable Activity_Modified	1.45
4	Tags_Not Available	1.42
7	Lead Quality_Worst	1.15
11	What is your current occupation_Working Profes...	1.14
5	Tags_Ringing	1.12
2	Tags_Closed by Horizzon	1.09
10	What is your current occupation_Student	1.08
3	Tags_Lost to EINS	1.06
1	Tags_Busy	1.05
8	Lead Source_Welingak Website	1.05
6	Tags_switched off	1.03

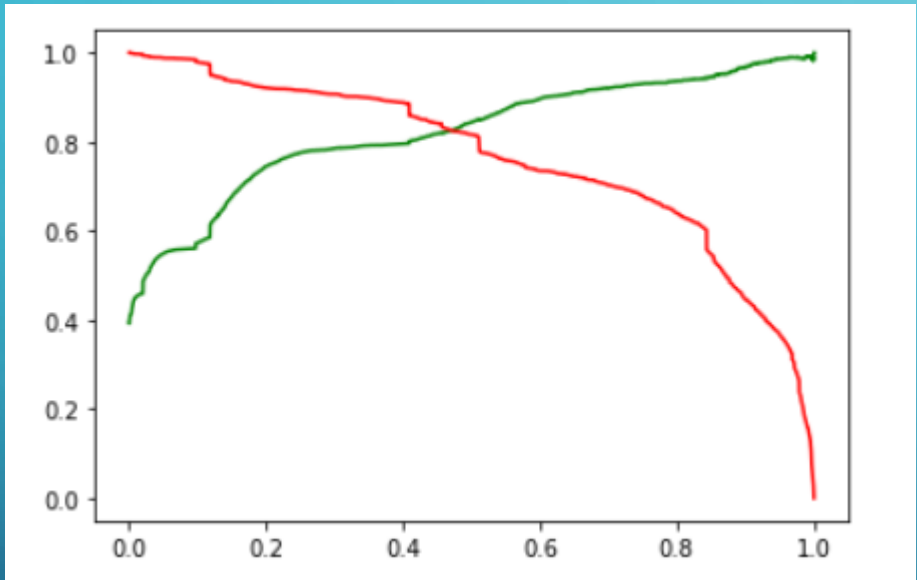
After dropping features with high correlation, finally model is built with 13 features and they are listed here





ROC curve denotes the area under the curve has more hence it interprets a useful test
We can notice sensitivity, specificity and accuracy meets at 0.45 so we can consider optimal cut off is 0.45

PRECISION RECALL CURVE



A high area under the curve represents both high recall and high precision, where high precision relates to a low false positive rate, and high recall relates to a low false negative rate. High scores for both show that the classifier is returning accurate results (high precision), as well as returning a majority of all positive results (high recall).

Conclusion

Based on model built, overall accurary came out to be 86.15% and so this model is very good and variables considered are effective

Below points are derived:

- ***Person who spent more time on website will be converted to positive lead***

- ***Based on VIF - These 3 variables are not highly correlated with remaining independent variables hence these variables are good***
- **Tags_Closed by Horizzon - has higher coefficent based on the model, hence we can target these people**

- **We can concentrate on the People enquired from Mumbai, with specialization as Finance management and lead source as Google**