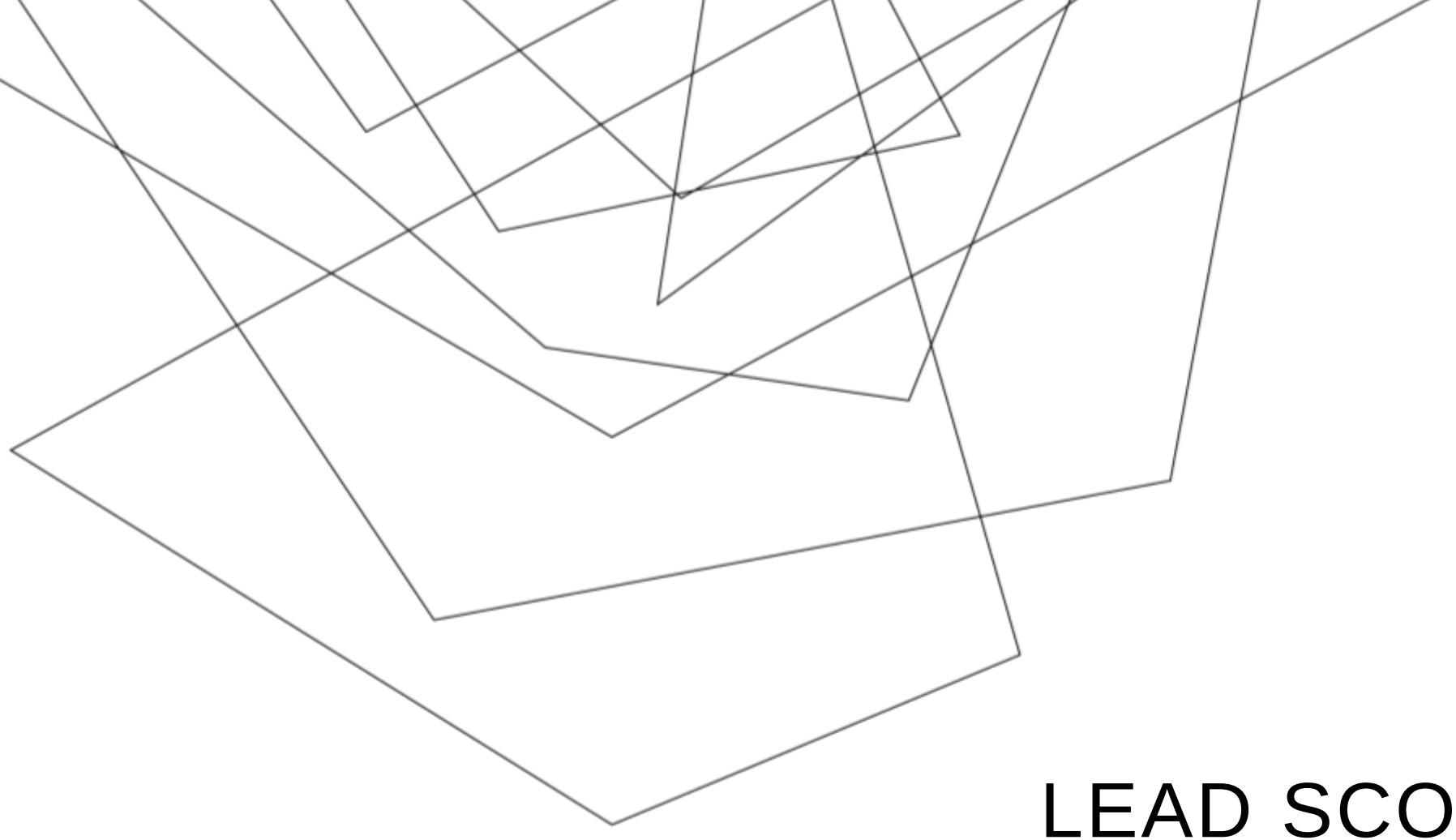


LEAD SCORE CASE STUDY

By Maria J Peter

PROBLEM STATEMENT

Problem Statement : X Education sells online courses to industry professionals. 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%.

Business Goal:

X Education needs help in selecting the most promising leads, i.e., the leads that are most likely to convert into paying customers.

The company needs a model wherein you a lead score is assigned 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%.

OVERALL APPROACH

Load and Clean Data	Data Analysis	Data Modelling	Prediction & Evaluation	Recommendation
* Size * Shape * Datatype * Identifying Null percentage *Standardizing *Handling Outliers	*Data Imbalance *Univariate & Bivariate analysis of categorical columns *Univariate & Bivariate analysis of numerical columns *Correlation between numerical columns	* Creating dummy variable * Splitting Train & Test * Rescaling * Checking Correlation *RFE *P-Value &VIF	*Prediction & Probability * Confusion Matrix * ROC *Optimal Cut off * Precession & Recall * Precession & Trade off	*Concluding who are the best targeted group.

DATA UNDERSTANDING

	CURRENT
SHAPE	(9240 , 37)
DATA TYPE	float64(4), int64(3), object(30)
DUPLICATES	NIL
OBSERVATION	Many place data is mentioned as select

DATA CLEANING

*Replaces Select with NaN

* Removing columns with null % above 35%

* Handling null between 2% and 20%we use impute

1) numerical variables impute with median

2) categorical with mode

*Handling null less than 2%

Drop the rows

*Converted Columns with Yes/No to 1/0

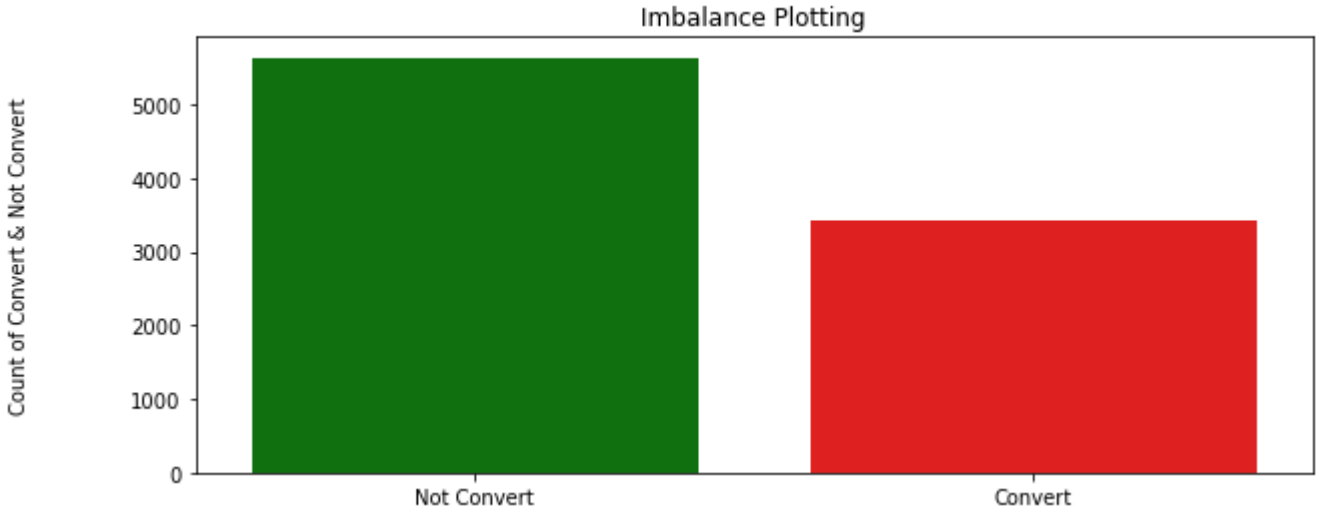
* Handling Outliers

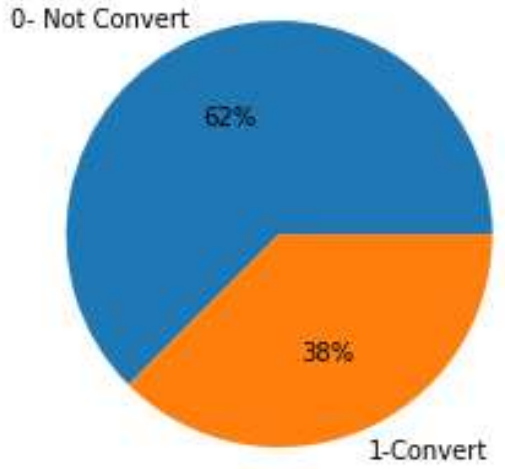
Total Visits

Total Time Spent on Website

Page Views Per Visit have large difference in Max and 75%

EXPLORATORY DATA ANALYSIS

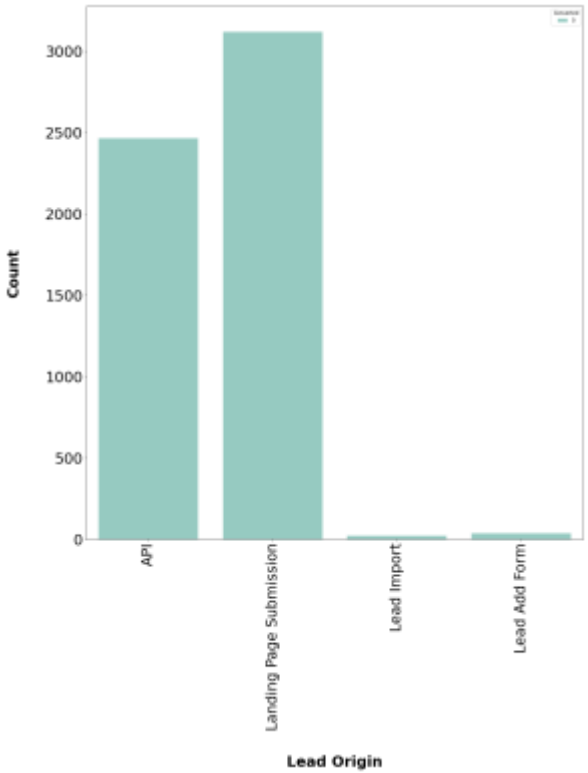


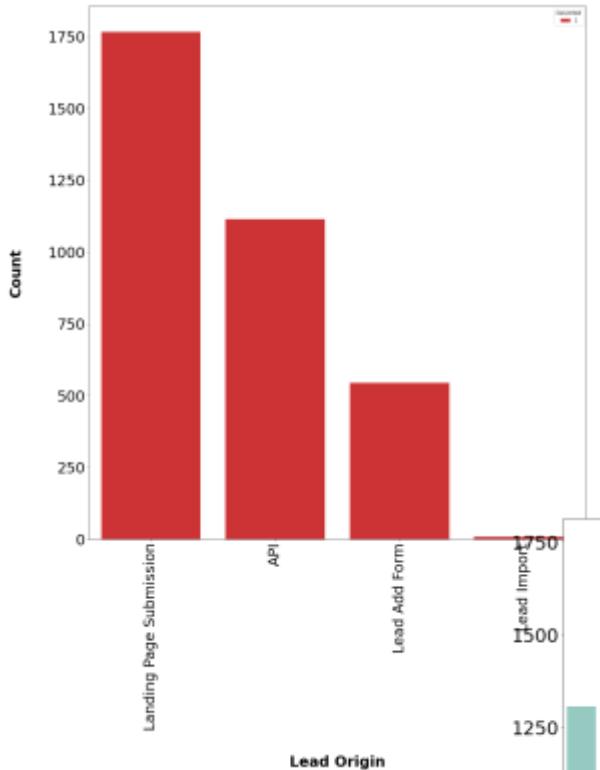


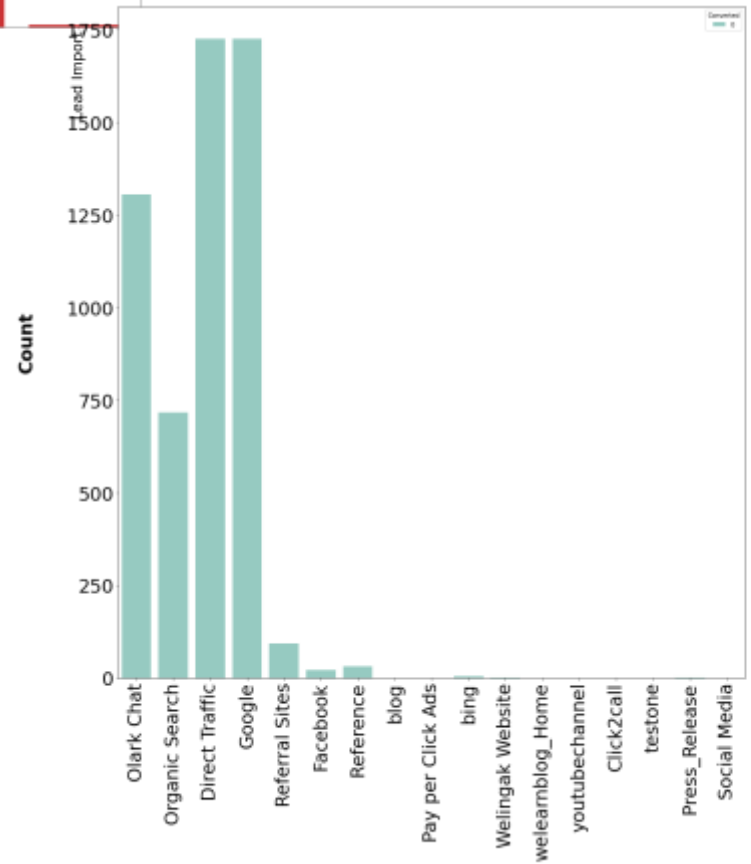
Observation

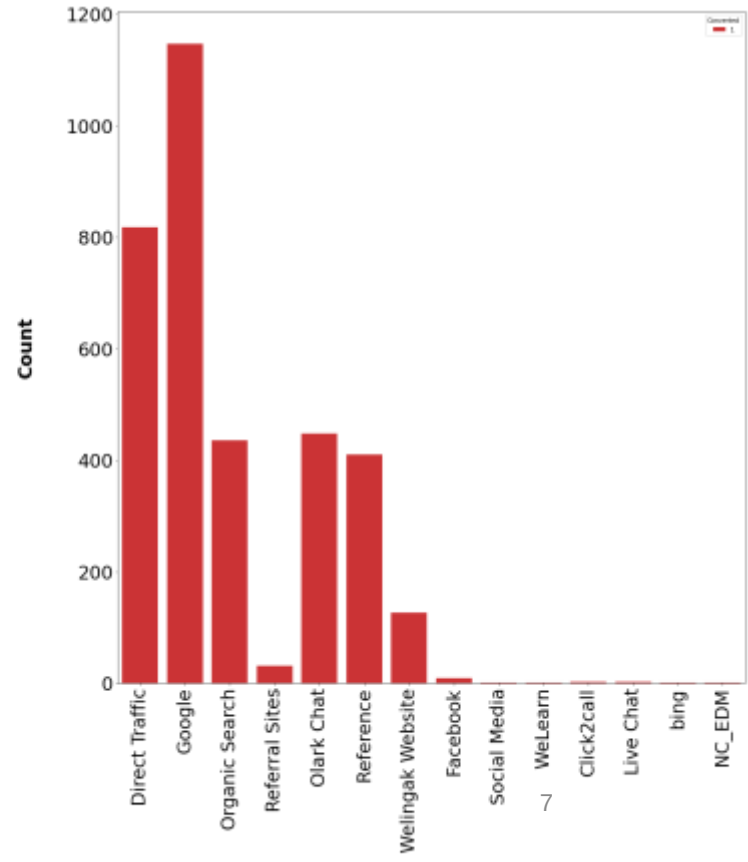
•Highly imbalanced with 62 % not convert and 38% convert.

ANALYSIS OF CATEGORICAL COLUMNS

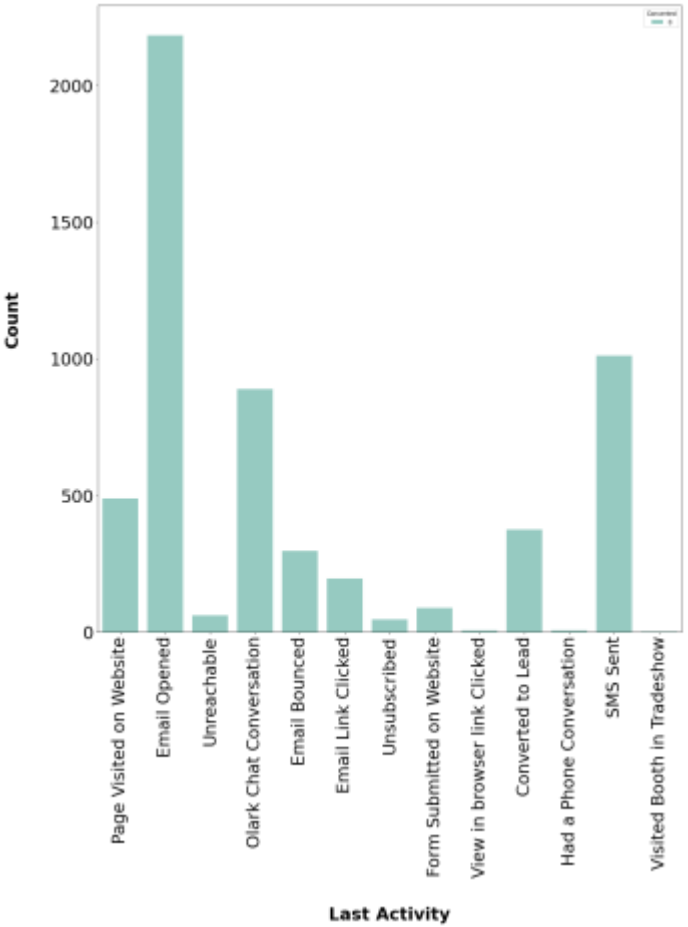


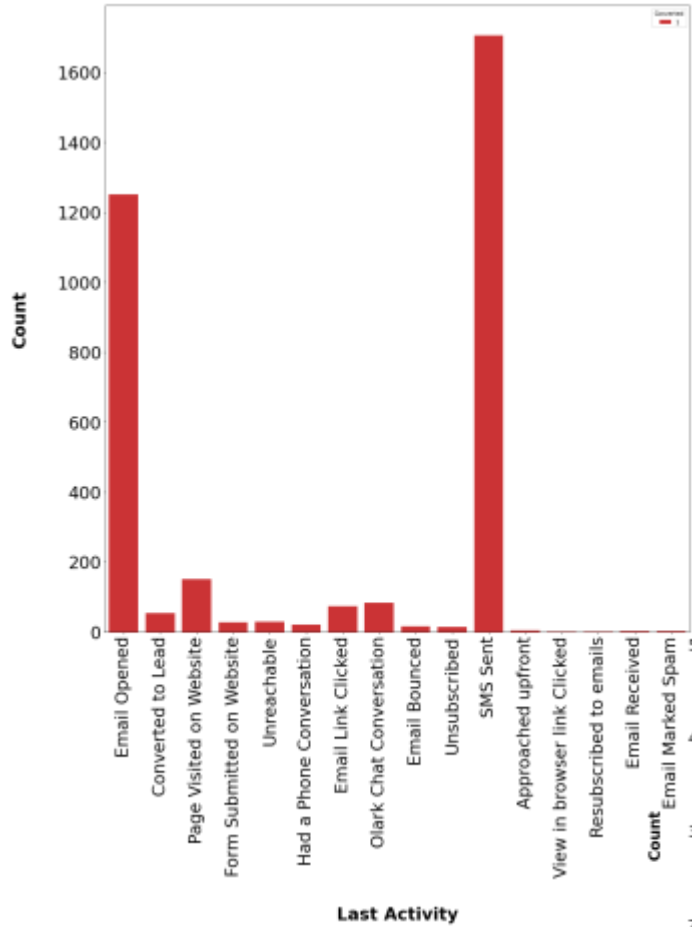


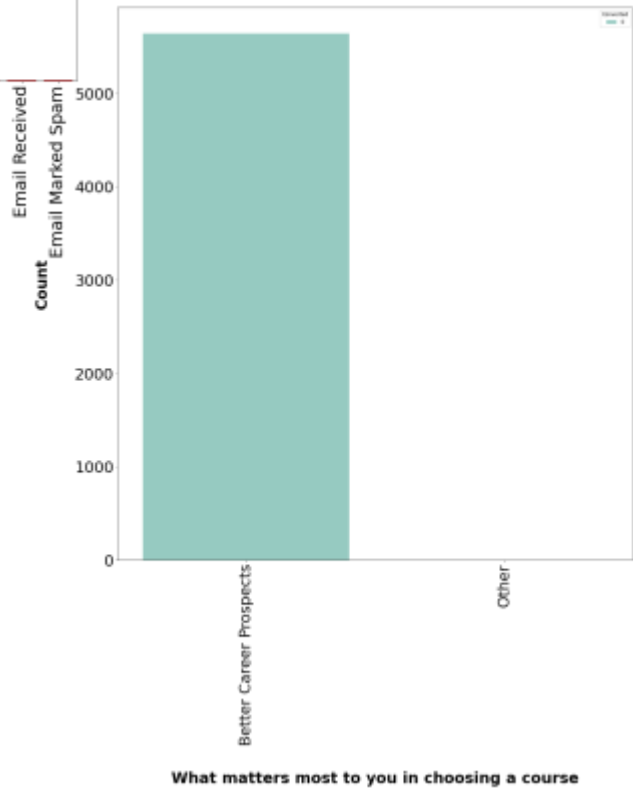


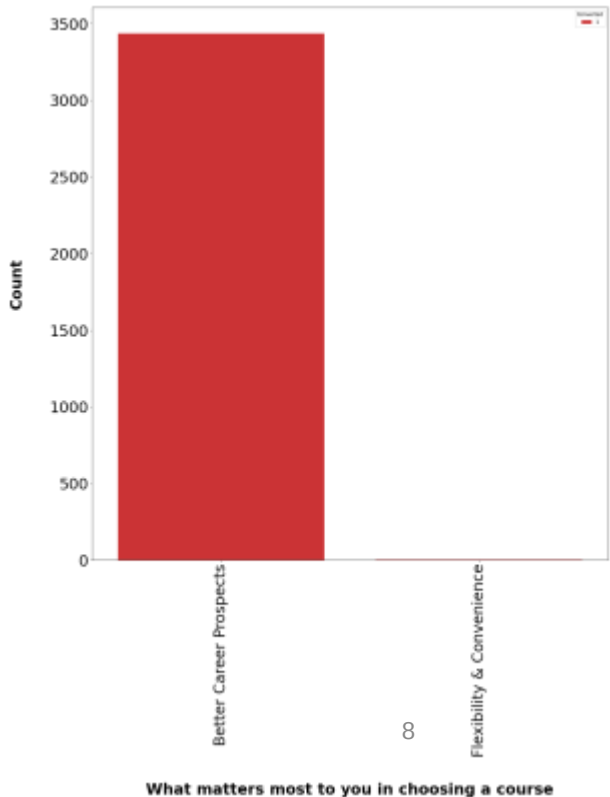


ANALYSIS OF CATEGORICAL COLUMNS

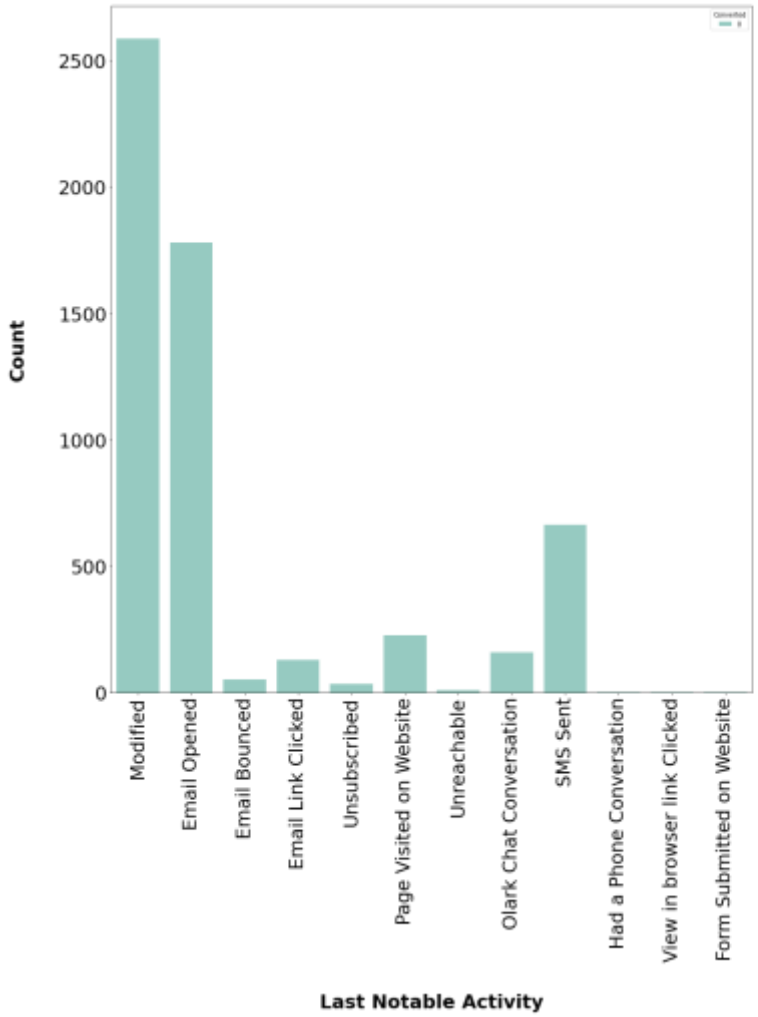


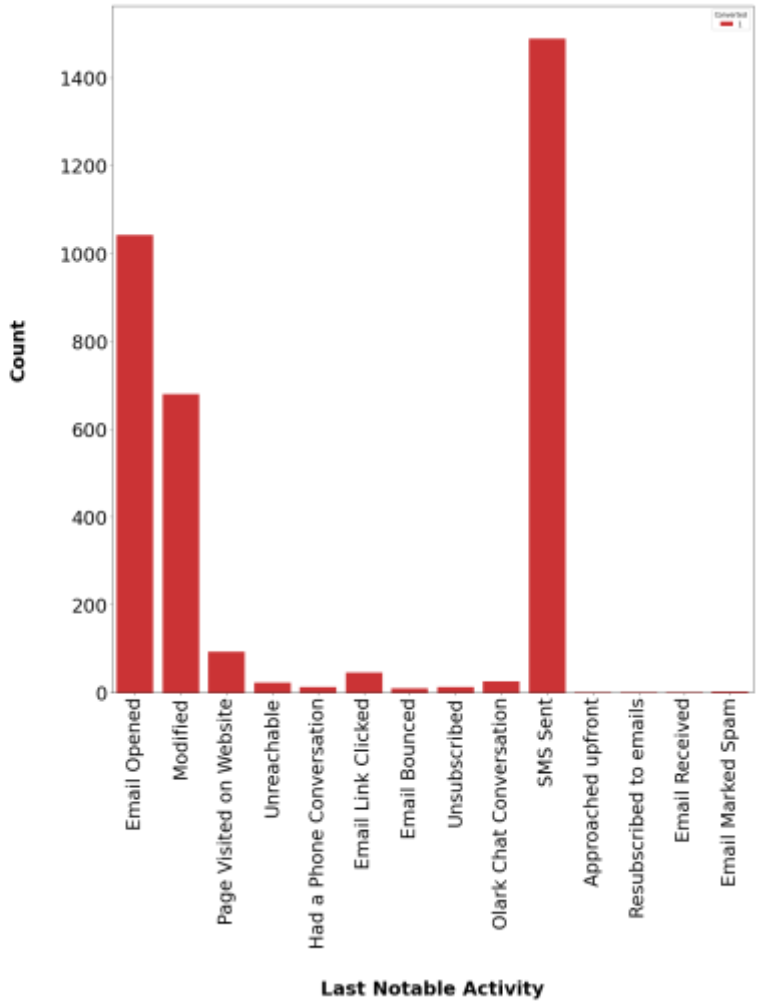






ANALYSIS OF CATEGORICAL COLUMNS

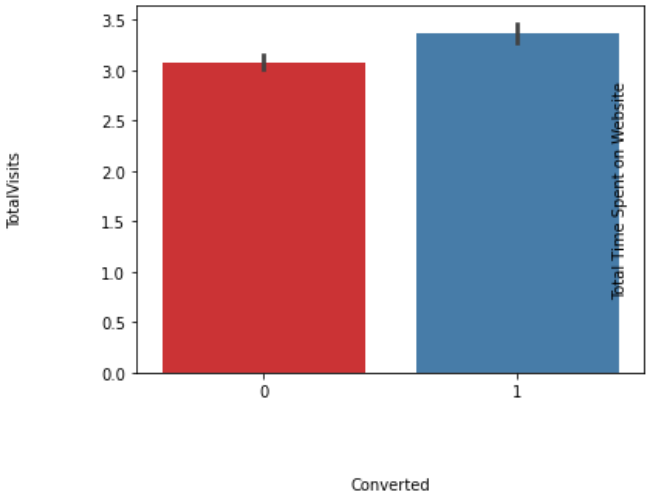


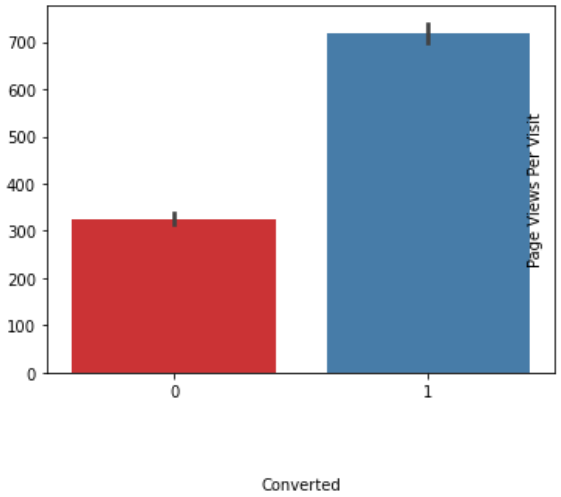


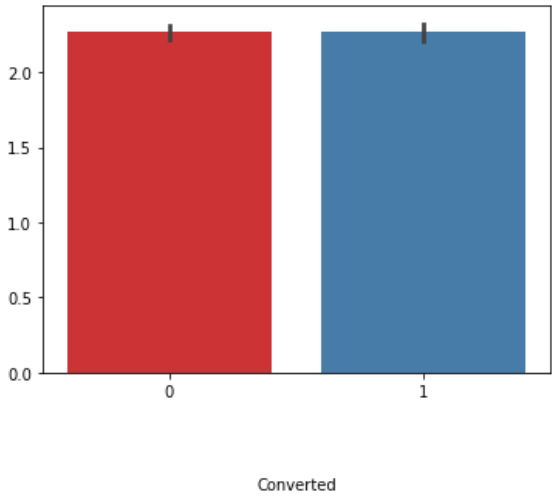
OBSERVATION

- Maximum conversion happened from Landing Page Submission Also there was only one request from quick add form which got converted.
- From the above graph, major conversion in the lead source is from google
- last activity value of 'SMS Sent' had more conversion
- India has the highest number od people
- In current occupation majority are unemployed followed by working professionals
- Reason for choosing course is better career
- Last notable activity is SMS sent

ANALYSIS OF NUMERICAL COLUMNS







OBSERVATIONS

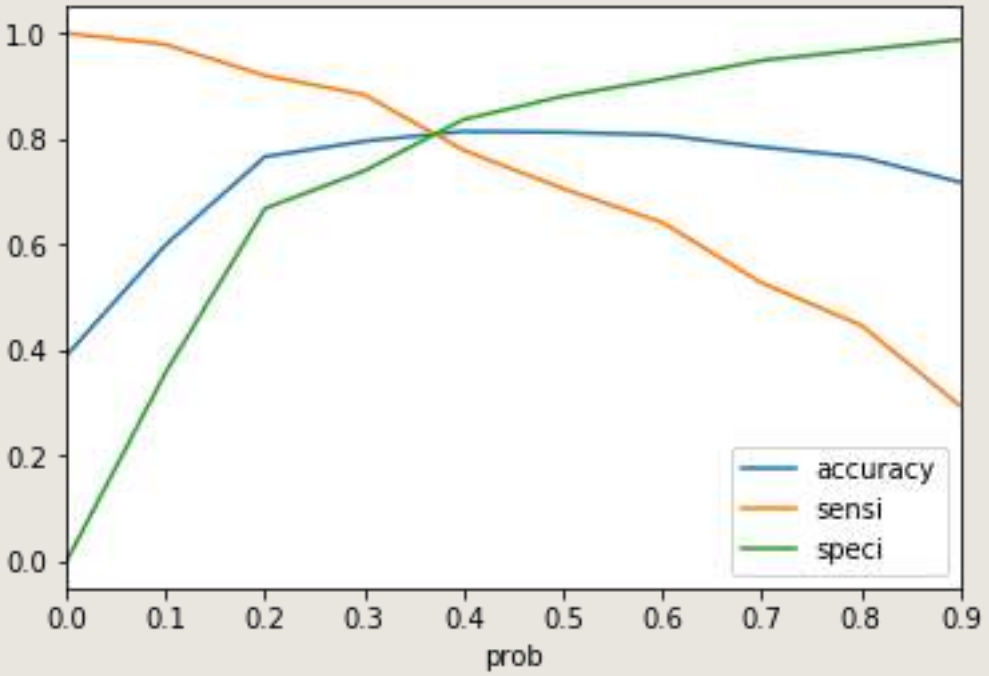
The conversion rated were high for Total Visits, Total Time Spent on Website and Page Views Per Visit

VARIABLES IMPACTING THE CONVERSION RATE

- Do Not Email
- Total Time Spent on Website
- Lead Origin _ Lead Add Form
- Lead Source _ Direct Traffic
- Lead Source _ Google
- Lead Source _ Organic Search
- Lead Source _ Referral Sites
- Last Activity _ Had a Phone Conversation
- Last Activity _ Olark Chat Conversation
- Last Activity _ SMS Sent
- Last Activity _ Unsubscribed
- Last Notable Activity _ Modified
- Last Notable Activity _ Unreachable

MODEL PREDICTION AND EVALUATION

Train Dataset-Model Evaluation



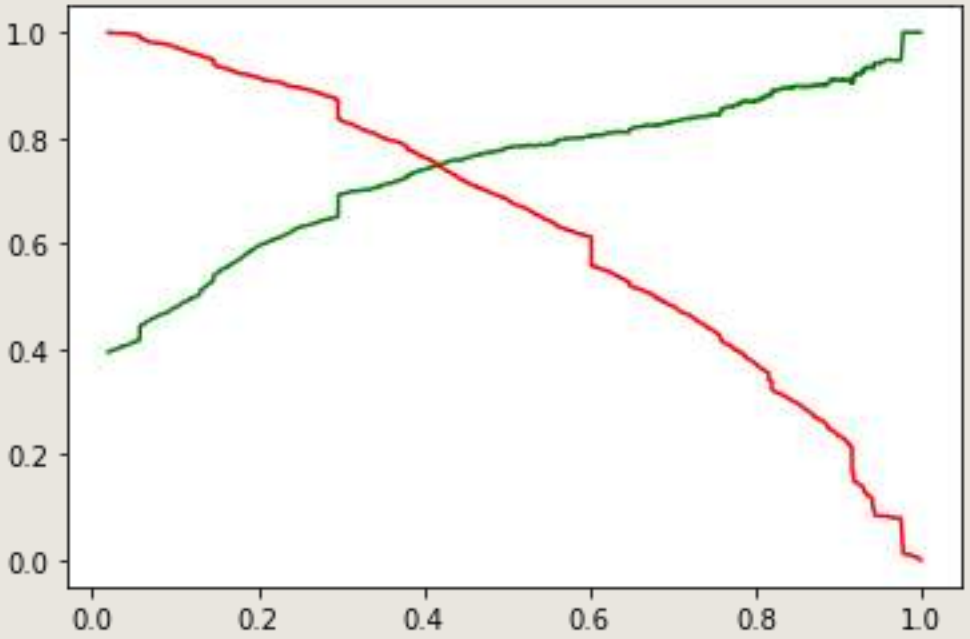
From the curve above, 0.37 is the optimum point to take it as a cutoff probability.

- Accuracy :80%,
- Sensitivity :68%,
- Specificity :88%
- False Positive Rate : 12%
- Positive Predictive Value :78%
- Negative predictive value : 81

Confusion Matrix

3151	**754**
511	**1935**

Train Dataset-Precision and Recall



Confusion Matrix

3436	**469**
774	**1672**

Precesion :78%

Recall: 68%

From the curve above, Based on the Precision and Recall tradeoff, we got a cut off value of approximately 0.42

PREDICTION ON TEST DATASET

- Accuracy :80.6%,
- Sensitivity :68.35%,
- Specificity :87.9%

Confusion Matrix

1409	**325**
202	**787**

CONCLUSION

- Accuracy, Sensitivity and Specificity values of test set are around 81.6%, 68% and 88% which are approximately closer to the respective values calculated using trained set.
- Also the lead score calculated shows the conversion rate on the final predicted model is around 80.9% (in train set) and 79.5% in test set
- The top 3 variables that contribute for lead getting converted in the model are

  Total time spent on website

  Lead Add Form from Lead Origin

  Had a Phone Conversation from Last Notable Activity
- Hence overall this model seems to be good.

RECOMENDATIONS

- The company should make calls to the leads who are the "Housewife" as they are more likely to get converted.
- The company should make calls to the leads who are the "Last Notable Activity_Had a Phone Conversation " as they are more likely to get converted.
- The company should make calls to the leads coming from the lead sources "Welingak Websites" as these are more likely to get converted.
- The company should make calls to the leads who spent "Lead Origin_Lead Add Form" as these are more likely to get converted.
- The company should not make calls to the leads whose last activity was "Olark Chat Conversation" as they are not likely to get converted.
- The company should not make calls to the leads whose lead origin is "Referral Sites" as they are not likely to get converted.