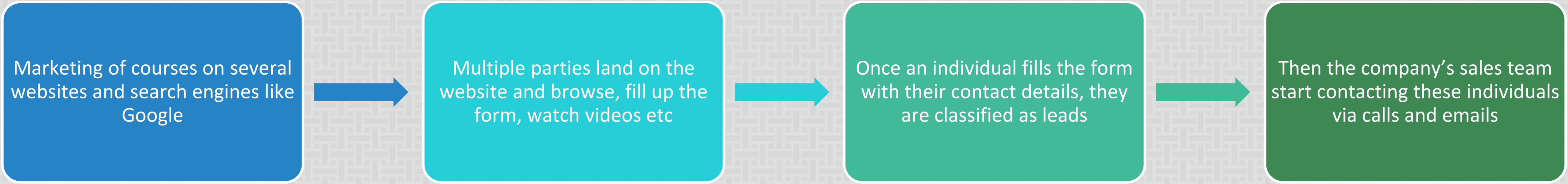
LEAD SCORING CASE STUDY

Presented by:
Shubhangi Prakash
Deepika Chinnala

BUSINESS PROBLEM STATEMENT

Our Client – X Education is an online course provider to industry professionals. It’s process of Lead Conversion is given in the below flowchart.



Marketing of courses on several websites and search engines like Google

Multiple parties land on the website and browse, fill up the form, watch videos etc

Once an individual fills the form with their contact details, they are classified as leads

Then the company’s sales team start contacting these individuals via calls and emails

PROBLEM STATEMENT:

The client identifies that the typical lead conversion rate is 30% only. To make this process more efficient, the company wishes to identify the most potential leads, also known as ‘Hot Leads’.

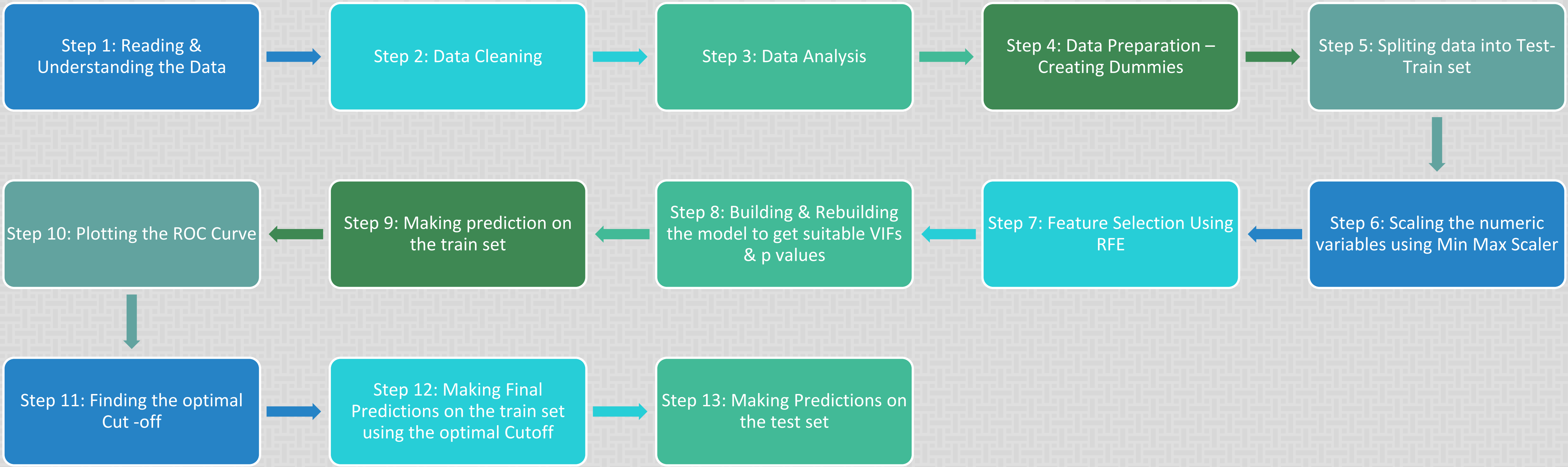
If they successfully identify this set of leads, the lead conversion rate should go up as the sales team will now be focusing more on communicating with the potential leads rather than making calls to everyone.

The company needs a model wherein 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, in particular, has given a ballpark of the target lead conversion rate to be around 80%.

MODEL BUILDING PROCESS

The following steps are used to build the model .



Step 1: Reading & Understanding the Data

Step 2: Data Cleaning

Step 3: Data Analysis

Step 4: Data Preparation – Creating Dummies

Step 5: Spliting data into Test-Train set

Step 6: Scaling the numeric variables using Min Max Scaler

Step 7: Feature Selection Using RFE

Step 8: Building & Rebuilding the model to get suitable VIFs & p values

Step 9: Making prediction on the train set

Step 10: Plotting the ROC Curve

Step 11: Finding the optimal Cut -off

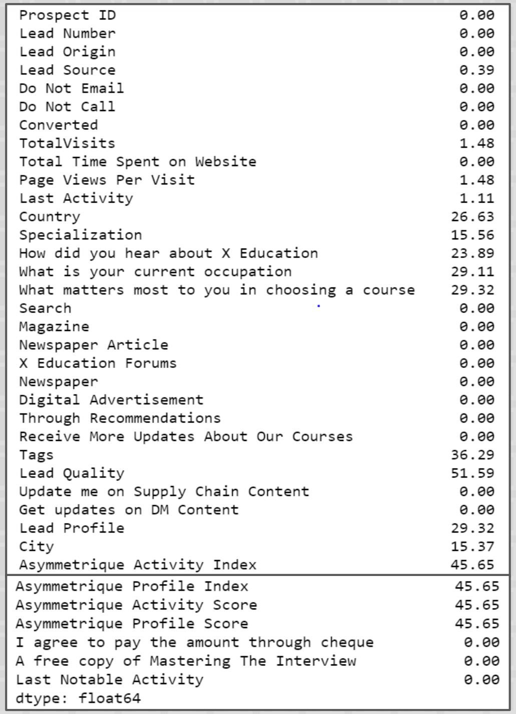
Step 12: Making Final Predictions on the train set using the optimal Cutoff

Step 13: Making Predictions on the test set

UNDERSTANDING & CLEANING THE DATA

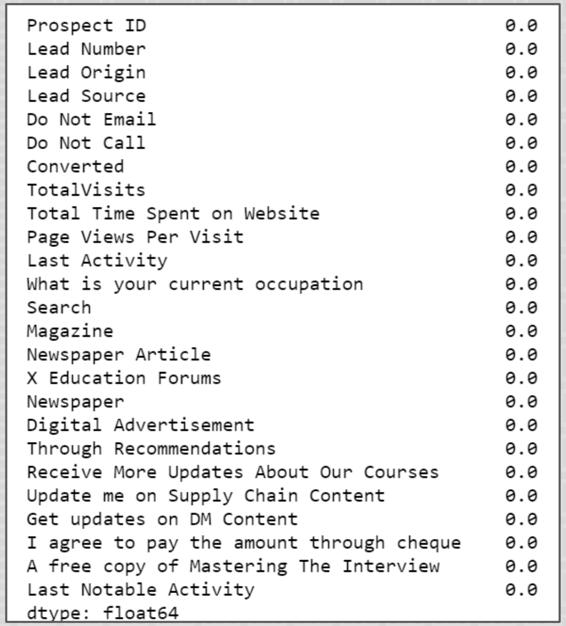
The major step performed here is:

1. Identification of the null values in the data set and dropping columns and rows or imputing values accordingly.
2. Identifying the outliers in all numeric columns and removing values beyond 99%.



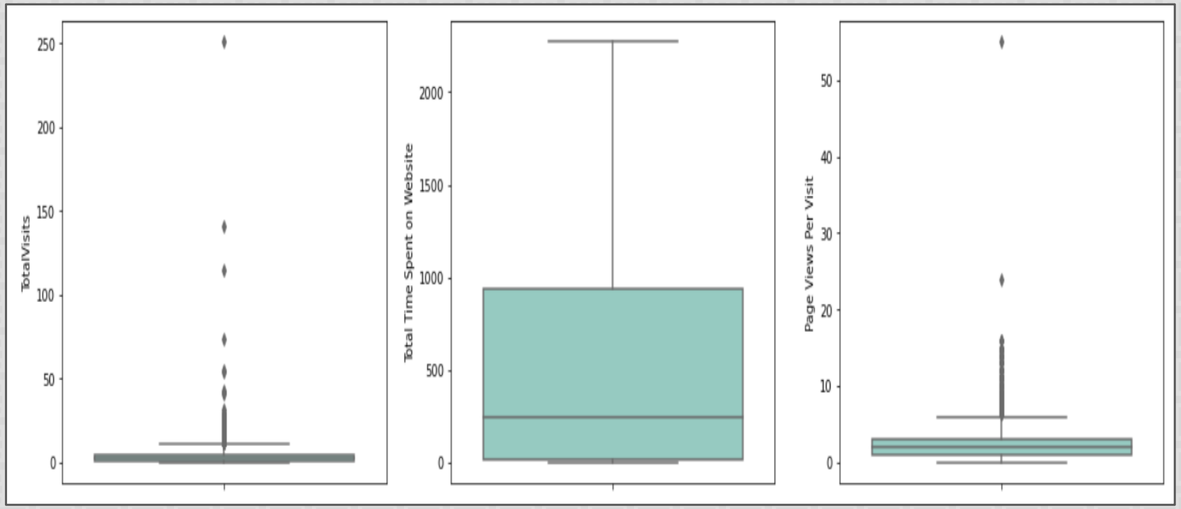
```
Prospect ID                                      0.00
Lead Number                                      0.00
Lead Origin                                      0.00
Lead Source                                      0.39
Do Not Email                                     0.00
Do Not Call                                      0.00
Converted                                        0.00
TotalVisits                                      1.48
Total Time Spent on Website                      0.00
Page Views Per Visit                             1.48
Last Activity                                    1.11
Country                                         26.63
Specialization                                  15.56
How did you hear about X Education              23.89
What is your current occupation                 29.11
What matters most to you in choosing a course   29.32
Search                                           0.00
Magazine                                         0.00
Newspaper Article                                0.00
X Education Forums                               0.00
Newspaper                                        0.00
Digital Advertisement                            0.00
Through Recommendations                          0.00
Receive More Updates About Our Courses           0.00
Tags                                            36.29
Lead Quality                                    51.59
Update me on Supply Chain Content                0.00
Get updates on DM Content                        0.00
Lead Profile                                    29.32
City                                            15.37
Asymmetrique Activity Index                     45.65
Asymmetrique Profile Index                      45.65
Asymmetrique Activity Score                     45.65
Asymmetrique Profile Score                      45.65
I agree to pay the amount through cheque         0.00
A free copy of Mastering The Interview           0.00
Last Notable Activity                            0.00
dtype: float64
```

The initial Dataframe had the following % missing values



```
Prospect ID                                 0.0
Lead Number                                 0.0
Lead Origin                                 0.0
Lead Source                                 0.0
Do Not Email                                0.0
Do Not Call                                 0.0
Converted                                   0.0
TotalVisits                                 0.0
Total Time Spent on Website                 0.0
Page Views Per Visit                        0.0
Last Activity                               0.0
What is your current occupation             0.0
Search                                      0.0
Magazine                                    0.0
Newspaper Article                           0.0
X Education Forums                          0.0
Newspaper                                   0.0
Digital Advertisement                       0.0
Through Recommendations                     0.0
Receive More Updates About Our Courses      0.0
Update me on Supply Chain Content           0.0
Get updates on DM Content                   0.0
I agree to pay the amount through cheque    0.0
A free copy of Mastering The Interview      0.0
Last Notable Activity                       0.0
dtype: float64
```

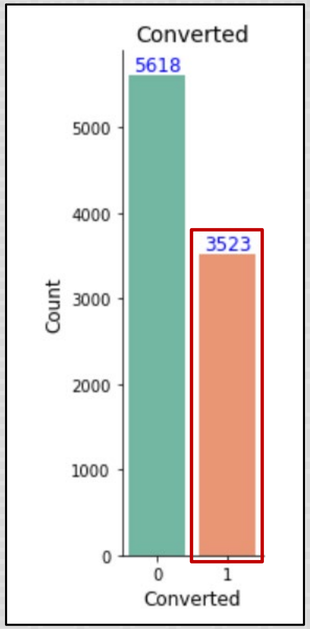
The cleaned dataset had the following columns retained with zero missing values



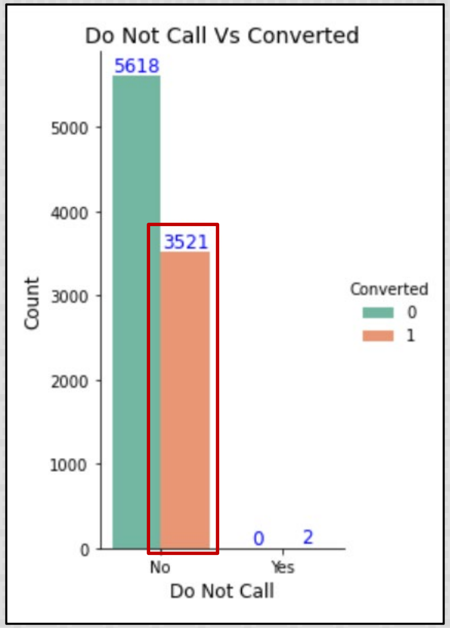
The numeric variable ‘Total Visits’ and ‘Page Views Per Visits’ had outliers hence values beyond 99% were removed.

DATA ANALYSIS

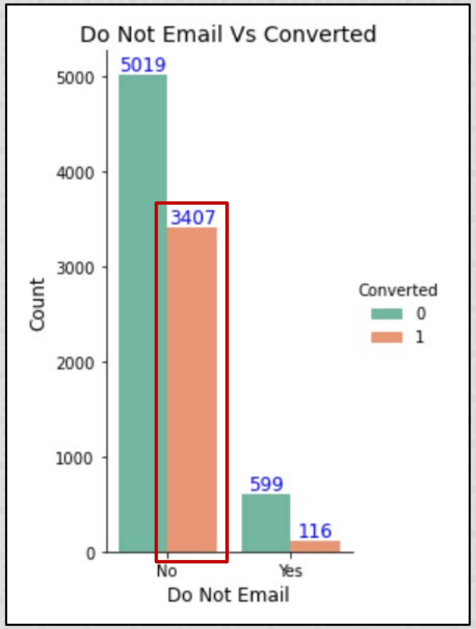
Understanding Each Columns Effect On The Conversion Rate



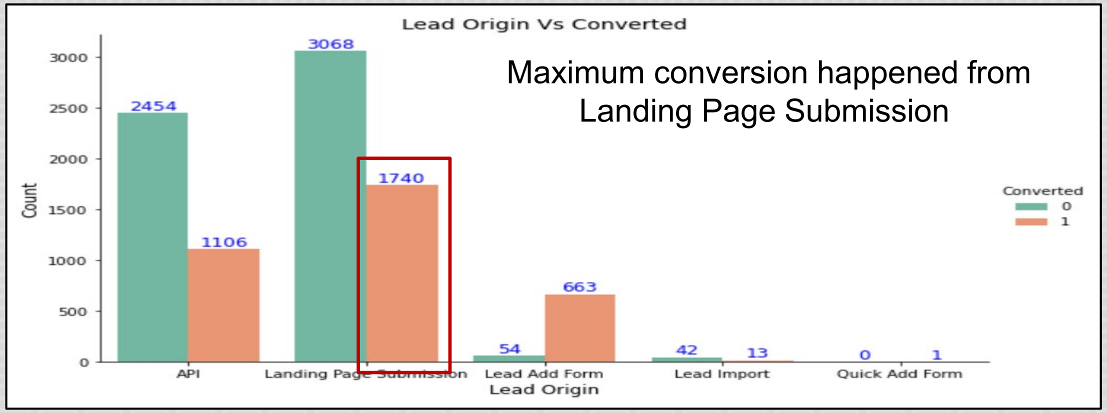
Shows overall conversion rate of 39%



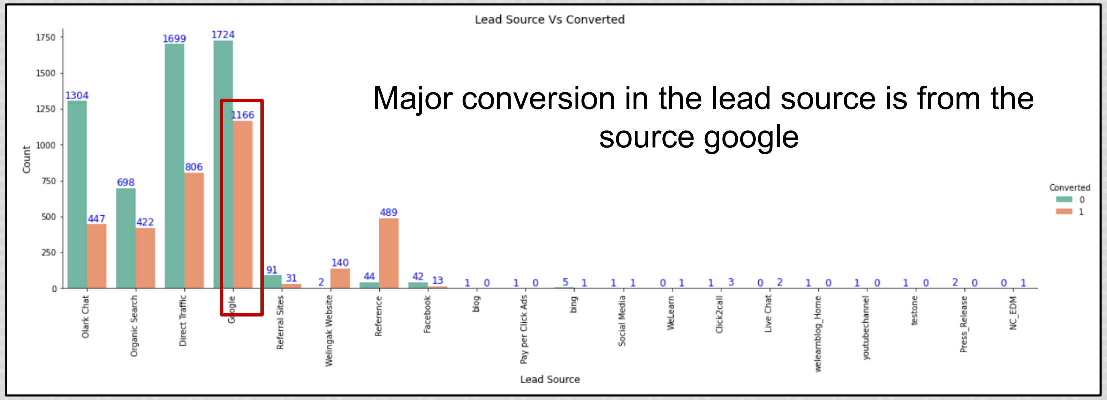
Major conversions happened when calls were made



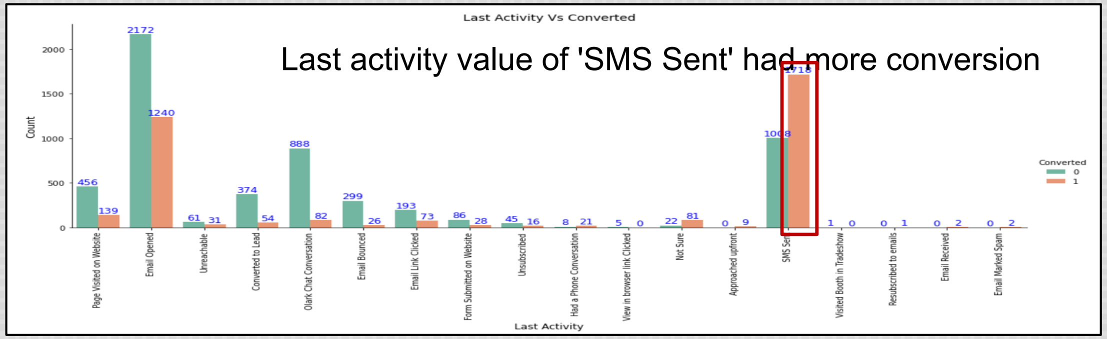
Major conversion has happened from the emails that have been sent



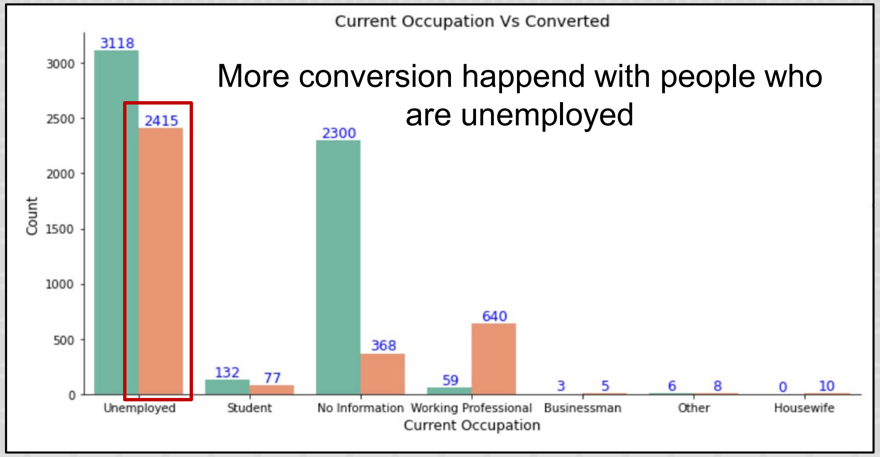
Maximum conversion happened from Landing Page Submission



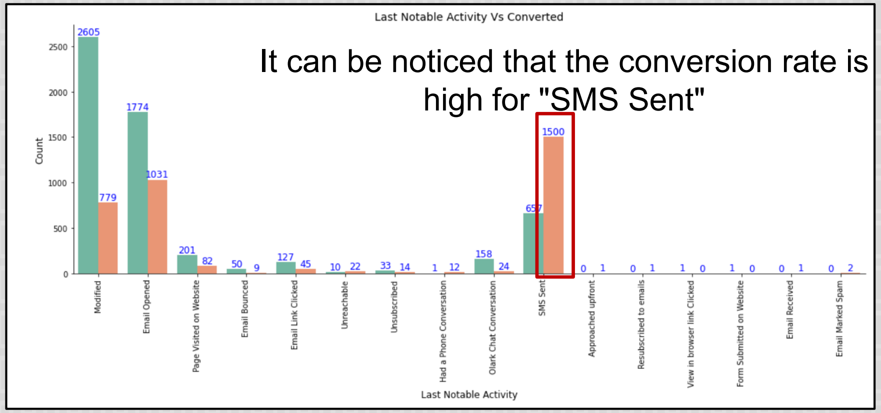
Major conversion in the lead source is from the source google



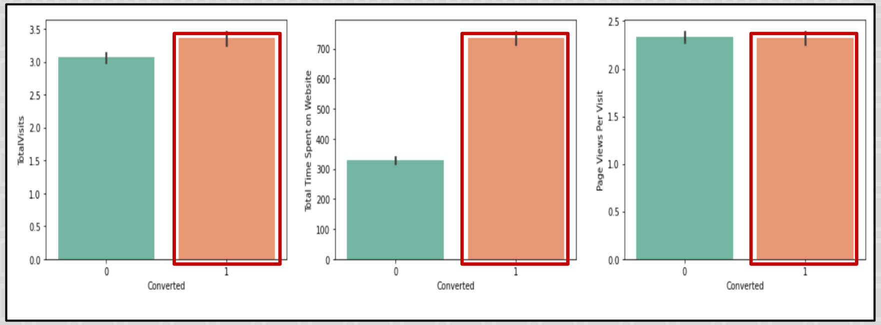
Last activity value of 'SMS Sent' had more conversion



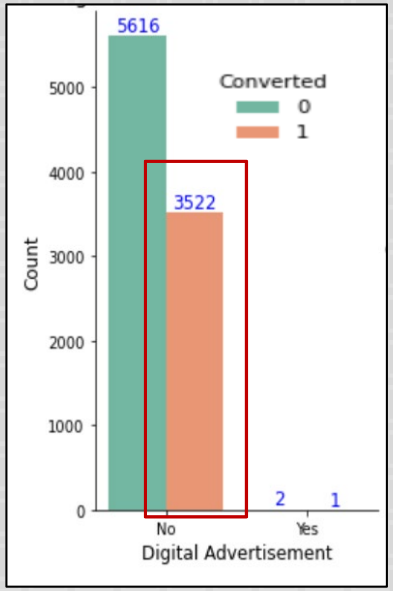
More conversion happend with people who are unemployed

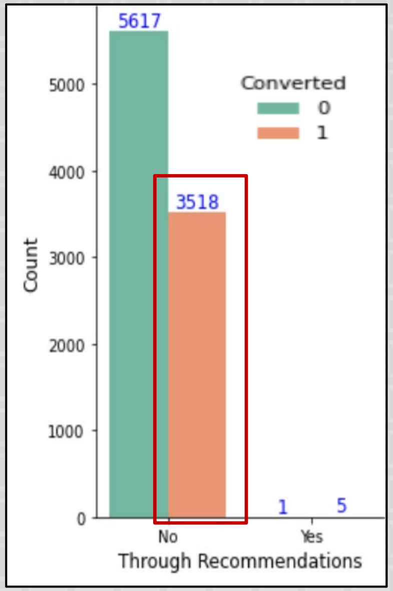


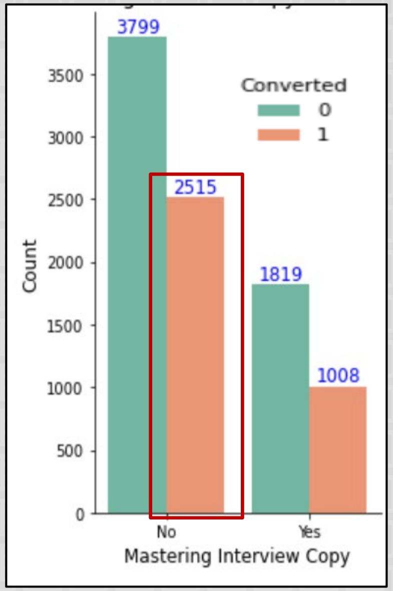
It can be noticed that the conversion rate is high for "SMS Sent"

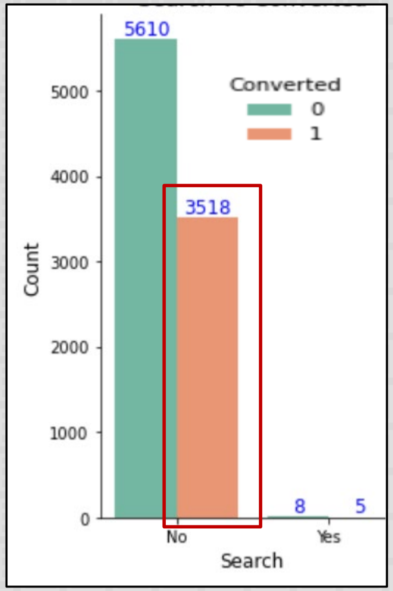


The conversion rates were high for Total Visits, Total Time Spent on Website and Page Views Per Visit

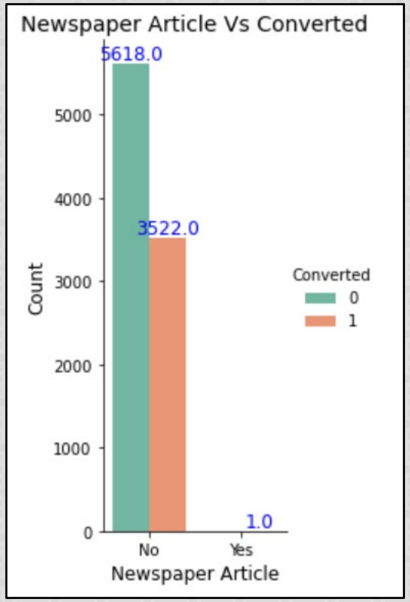


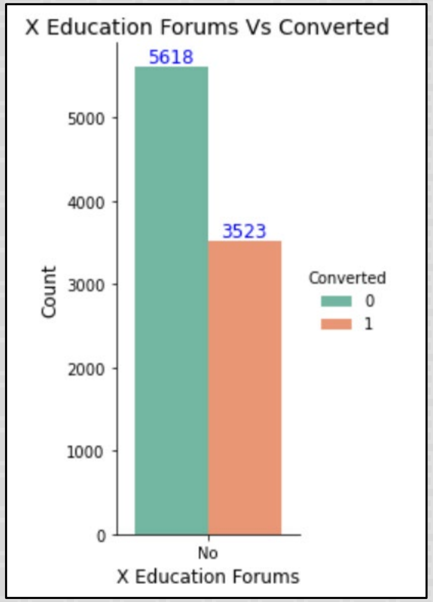


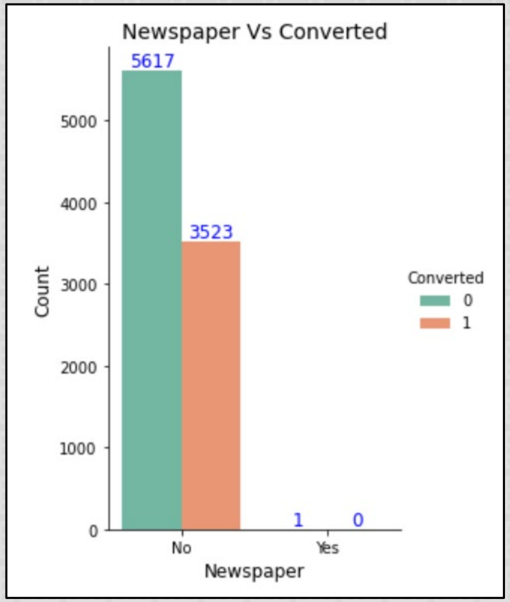




COLUMNS DROPPED DUE TO MAJORITY ROWS HAVING VALUE AS 'NO'

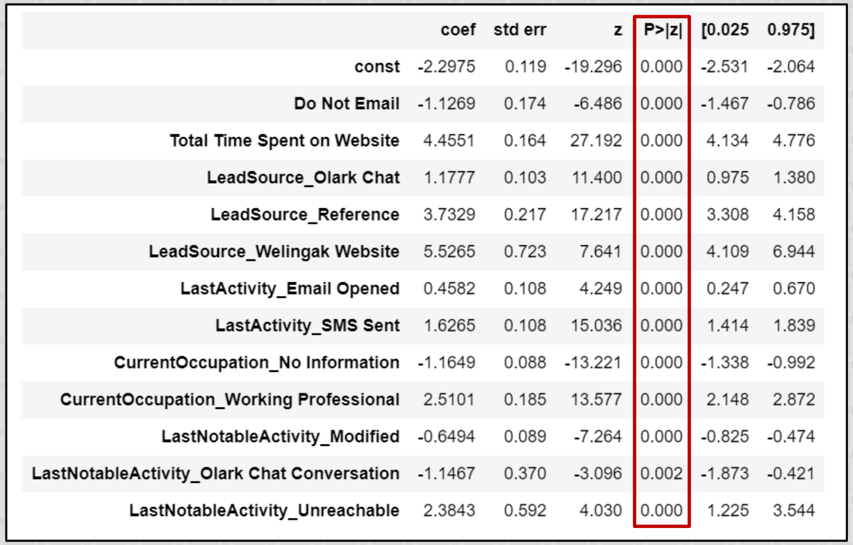




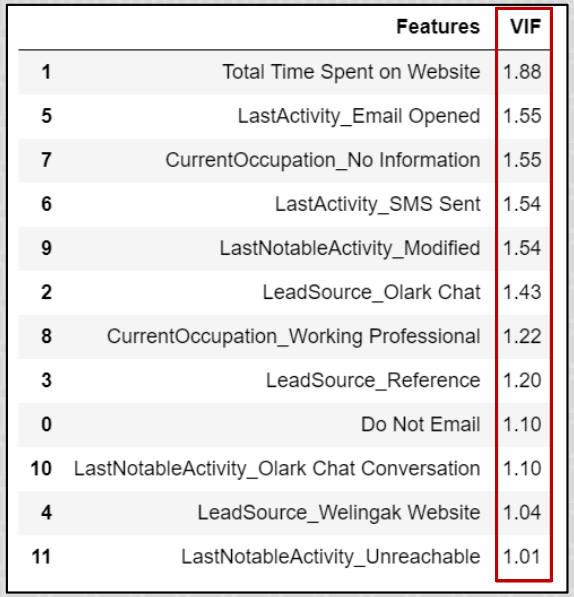


DATA PREPARATION AND MODEL BUILDING

1. Converting **binary variables to 0/1**.
2. Creating **dummy variables** for non binary categorical variables and dropping the original columns.
3. Splitting the data frame into **test and train sets** for training our model and testing.
4. Scaling the numeric variables using **Min Max Scaler** for normalization of data.
5. Selecting columns using RFE.
6. Building the model using statsmodel.api.
7. Checking the **p values and VIFs** and dropping variables one by one on the basis of the same.
8. Re building the model post dropping the column with high p value or high VIF.
9. The final model results are shown below:



	coef	std err	z	P>|z|	[0.025	0.975]
const	-2.2975	0.119	-19.296	0.000	-2.531	-2.064
Do Not Email	-1.1269	0.174	-6.486	0.000	-1.467	-0.786
Total Time Spent on Website	4.4551	0.164	27.192	0.000	4.134	4.776
LeadSource_Olark Chat	1.1777	0.103	11.400	0.000	0.975	1.380
LeadSource_Reference	3.7329	0.217	17.217	0.000	3.308	4.158
LeadSource_Welingak Website	5.5265	0.723	7.641	0.000	4.109	6.944
LastActivity_Email Opened	0.4582	0.108	4.249	0.000	0.247	0.670
LastActivity_SMS Sent	1.6265	0.108	15.036	0.000	1.414	1.839
CurrentOccupation_No Information	-1.1649	0.088	-13.221	0.000	-1.338	-0.992
CurrentOccupation_Working Professional	2.5101	0.185	13.577	0.000	2.148	2.872
LastNotableActivity_Modified	-0.6494	0.089	-7.264	0.000	-0.825	-0.474
LastNotableActivity_Olark Chat Conversation	-1.1467	0.370	-3.096	0.002	-1.873	-0.421
LastNotableActivity_Unreachable	2.3843	0.592	4.030	0.000	1.225	3.544



	Features	VIF
1	Total Time Spent on Website	1.88
5	LastActivity_Email Opened	1.55
7	CurrentOccupation_No Information	1.55
6	LastActivity_SMS Sent	1.54
9	LastNotableActivity_Modified	1.54
2	LeadSource_Olark Chat	1.43
8	CurrentOccupation_Working Professional	1.22
3	LeadSource_Reference	1.20
0	Do Not Email	1.10
10	LastNotableActivity_Olark Chat Conversation	1.10
4	LeadSource_Welingak Website	1.04
11	LastNotableActivity_Unreachable	1.01

All variables have a good value of VIF and p-value. So we need not drop any more variables.

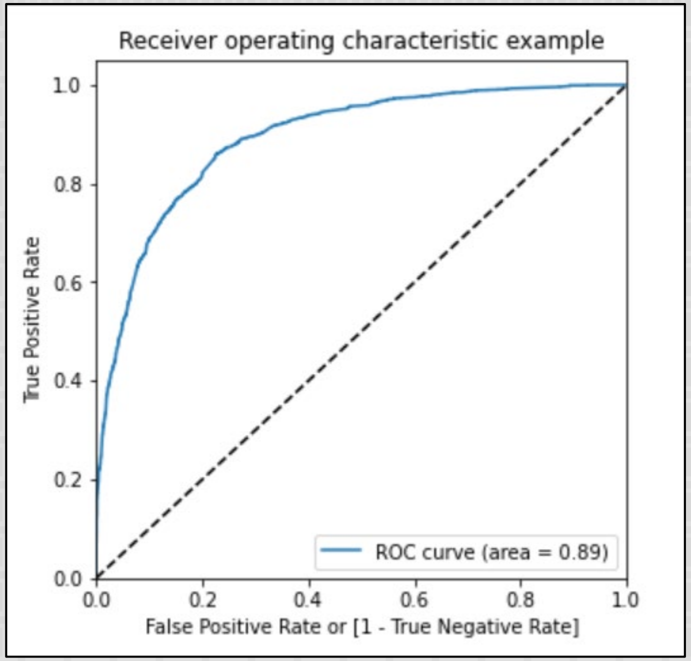
PREDICTIONS ON TRAIN SET & PLOTTING THE ROC

1. The below confusion matrix has been derived to draw conclusions.

	Predicted	
Actual	Not Converted	Converted
Not Converted	3481	442
Converted	728	1747

From the matrix the following metrics can be calculated

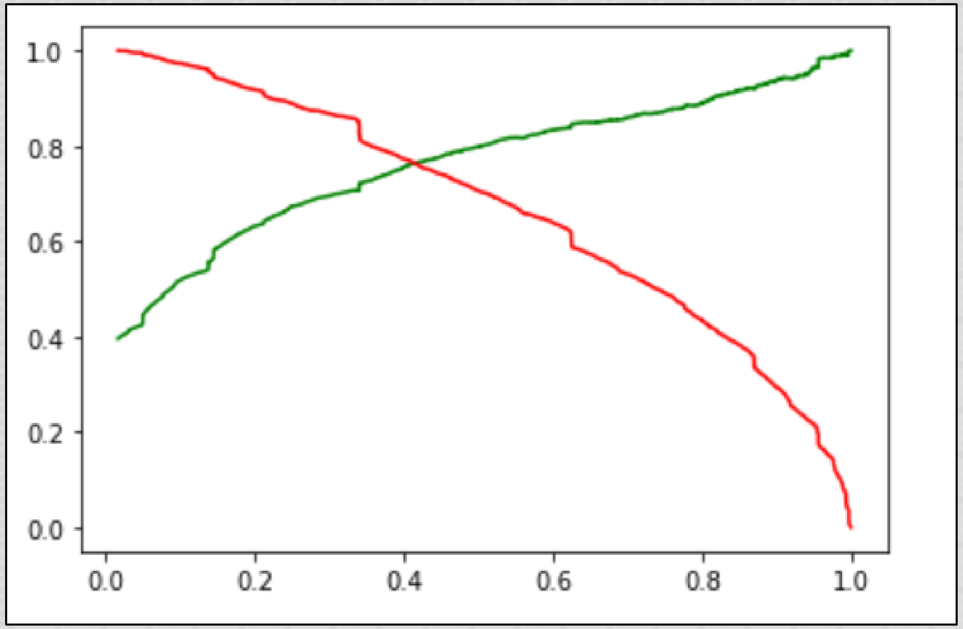
- Accuracy – 82%
- Sensitivity – 71%
- Specificity – 89%
- False Positive Rate – 11.2%
- Positive Predictive Value – 80%
- Negative Predictive Value – 83%



- An ROC curve shows the trade – off between sensitivity and 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.

FINDING THE OPTIMUM POINT

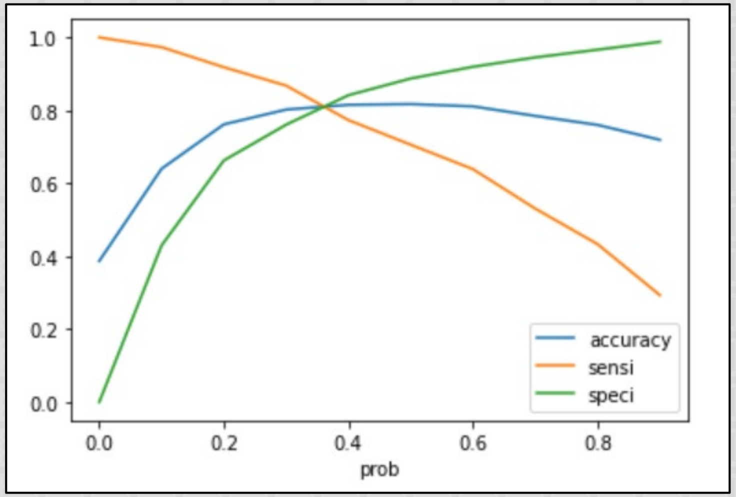
Making Final Predictions using the optimum cut off



Precision – Recall Trade off for the train set :

The precision and recall for the test set is as follows

1. Precision – 80%
2. Recall – 71%



Accuracy, sensitivity and specificity are used to find the optimum cut off which is 0.38 in the following model

<u>THE PREDICTIONS ARE MADE WITH 0.38 CUT-OFF</u>

1. Overall Accuracy – 81%

2. The new confusion matrix is as below:

	Predicted	
Actual	Not Converted	Converted
Not Converted	3247	676
Converted	529	1946

- Sensitivity – 79%
- Specificity – 83%

- False Positive Rate – 17%
- Positive Predictive Value – 74%
- Negative Predictive Value – 86%

PREDICTIONS ON THE TEST SET

Using a cut-off of 0.38 on the test set the following confusion matrix is derived:

	Predicted	
Actual	Not Converted	Converted
Not Converted	1385	310
Converted	240	808

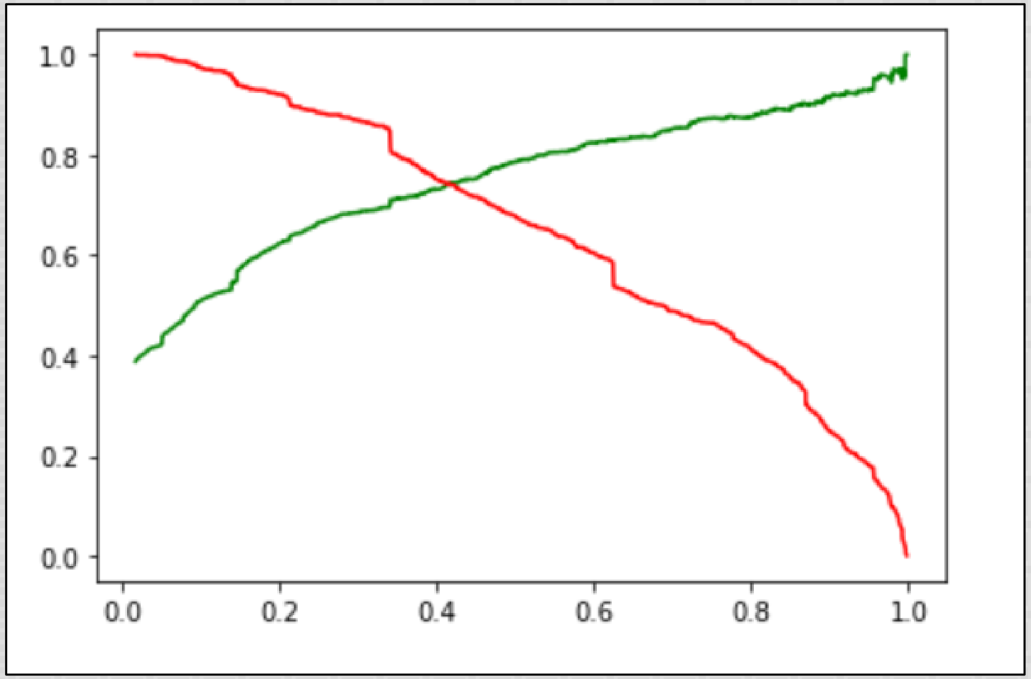
From the matrix the following metrics can be calculated

- Accuracy – 80%
- Sensitivity – 77%
- Specificity – 82%

Precision – Recall Trade off :

The precision and recall for the test set is as follows

1. Precision – 72%
2. Recall – 77%



CONCLUSION

- While we have checked both **Sensitivity-Specificity as well as Precision and Recall Metrics**, we have considered the **optimal cut off** based on Sensitivity and Specificity for calculating the final prediction.
- **Accuracy, Sensitivity and Specificity values of test set are around 80%, 77% and 82%** which are approximately closer to the respective values calculated using trained set.
- Also the lead score calculated in the trained set of data shows the **conversion rate on the final predicted model is around 79%.**
- Hence overall this model seems to be good.

It was found that the variables that mattered the most in the potential buyers are:

- **The total time spend on the Website.**
- **Do not Email**
- **When the lead source was: a. Olark Chat b. Reference c. Welingak Website**
- **When the last activity was: a. SMS b. Email opened**
- **When the last Notable activity was: a. Modified b. Unreachable c. Olark Chat Conversation**
- **When their current occupation is as a working professional and also has no information**

Keeping these in mind the X Education can flourish as they have a very high chance to get almost all the potential buyers to change their mind and buy their courses.

THANK YOU

PRESENTATION BY:
DEEPIKA CHINNALA
SHUBHANGI PRAKASH