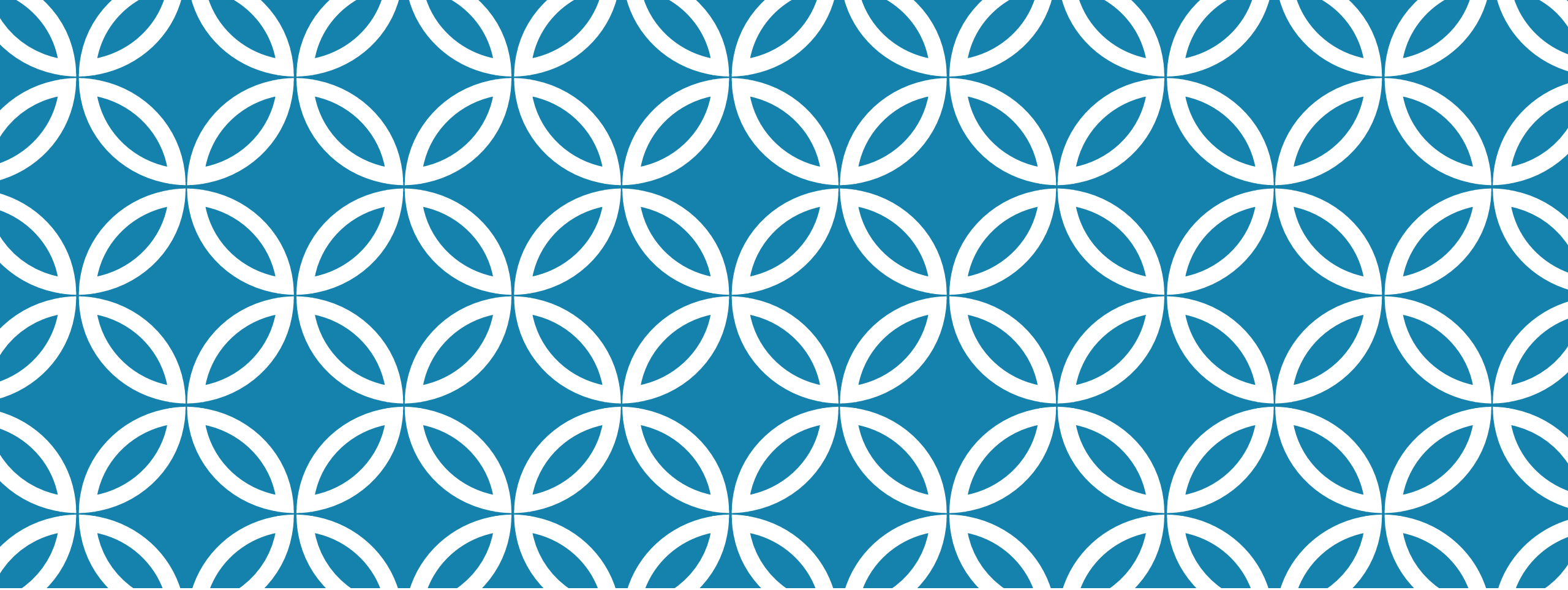


LEADING SCORING CASE STUDY

Shivani

Intoduction

PROBLEM STATEMENT

X Education is an organization which provide online courses for industry professionals. The company marks it courses on several popular website like google.

It want to select most promising leads that can be converted to paying customers.

Although the company generates a lot of leads only a few are converted into paying customers. Mostly leads come through numerous model like email, advertisements on website, googles searches etc.

The company has had 38.5% conversion rate through the whole process of turning leads into customers by approaching those leads which are to be found having interest in taking the course. The implementation process of lead generating attributes are not sufficient in helping conversions.

BUSINESS GOAL

The company requires a model to be built for selecting most promising leads. Lead score to be given to each leads such that it indicates how promising that lead could be. The higher the lead score the more promising the lead to get converted, the lower it is the lesser the chances of conversion. The model to be built in the lead conversion rate maximum.

Approach and Strategy

1. Importing and cleaning the data
2. Exploratory Data Analysis
3. Data Preparation
4. Building the Logistic regression model
5. Prediction and evaluation on Train data
6. Prediction and evaluation on Test data
7. Conclusion
8. Lead Scoring (0-100)

1. Import and Clean the data

DATA INSPECTION

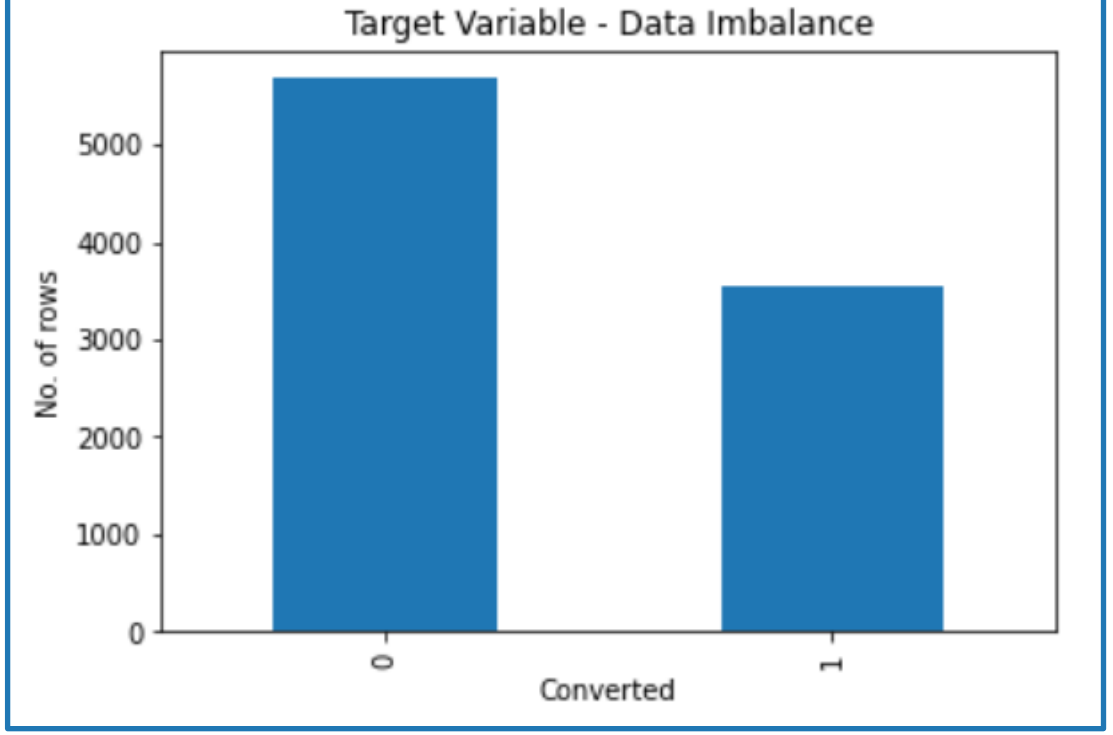
After importing the file ‘Leads.csv’, we did a rough inspection and found that the table contained 9240 columns, 29 Categorical and 5 continuous variables. ‘Converted’ was the target variable and the columns ‘Prospect ID’ and ‘Lead Number’ were identifiers and need not be considered in the case study

DATA IMBALANCE

The Target variable ‘Converted’ had a data imbalance of ratio 62:38, which is similar to the expectations set in the problem statement.

In other words the conversion rate is ~38%

```
Distribution of Target Variable 'Converted' in % is
0    61.461039
1    38.538961
Name: Converted, dtype: float64
================================================
```



Data imbalance in Target Variable

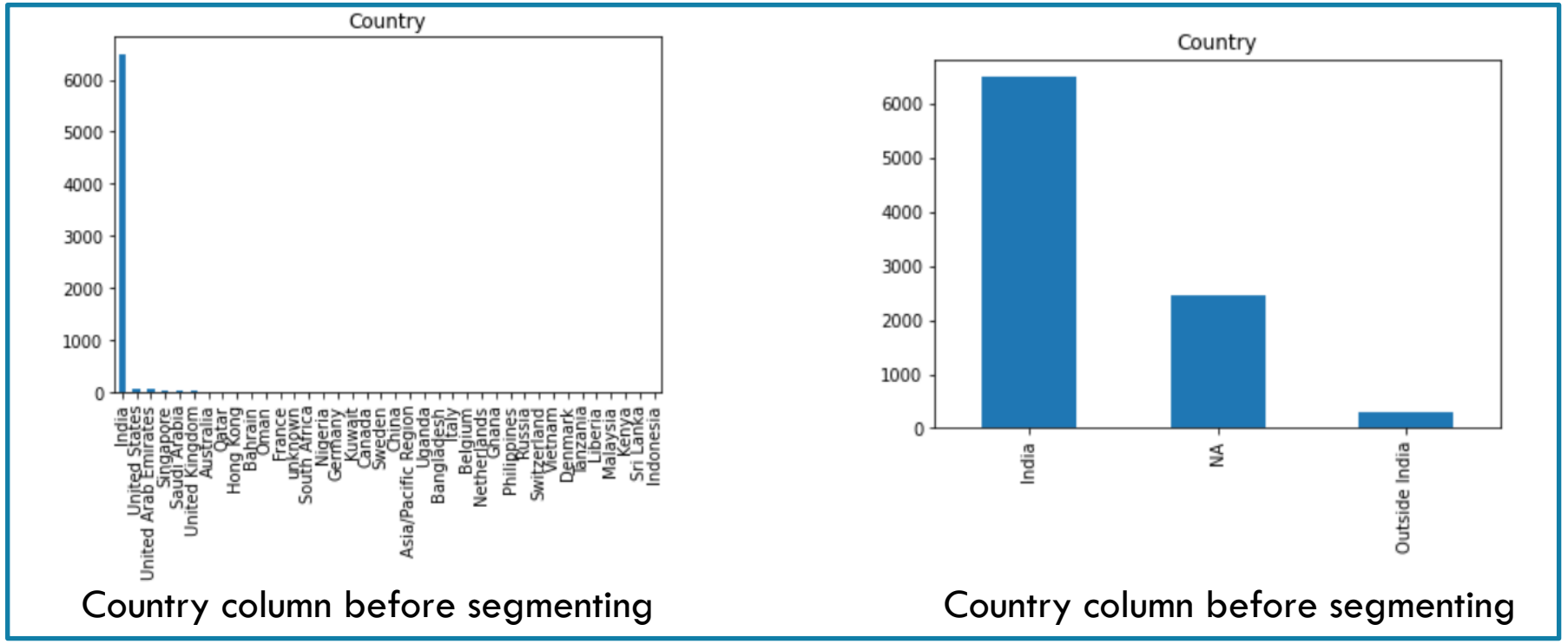
1. Import and Clean the data

REPLACING SELECT

Select' is a recurring value in many columns and as per the problem statement it is to be considered as blank. The reason is that 'Select' is the default option in forms. So we replaced 'Select' with np.nan

COUNTRY

Most of the leads were from India. We segmented this column's values as 'India', ''Outside India' and 'NA'



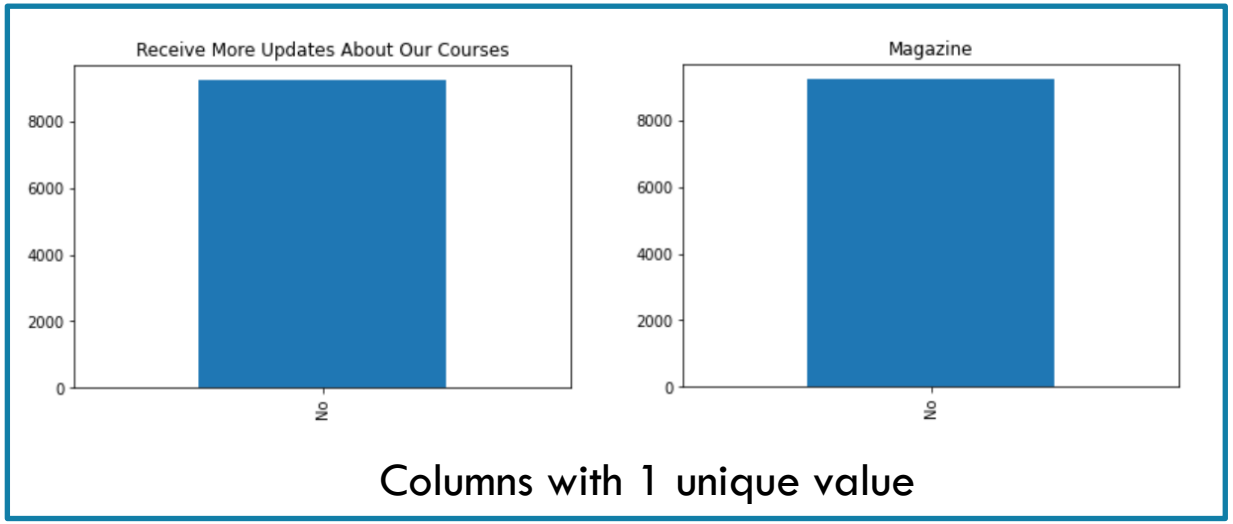
Country column before segmenting

Country column before segmenting

1. Import and Clean the data

DROPPING COLUMNS WITH 1 UNIQUE VALUES

Many columns like ‘Magazine’, ‘Receive More Updates About Our Courses’ etc had extreme data imbalance or one unique value. So we dropped these columns



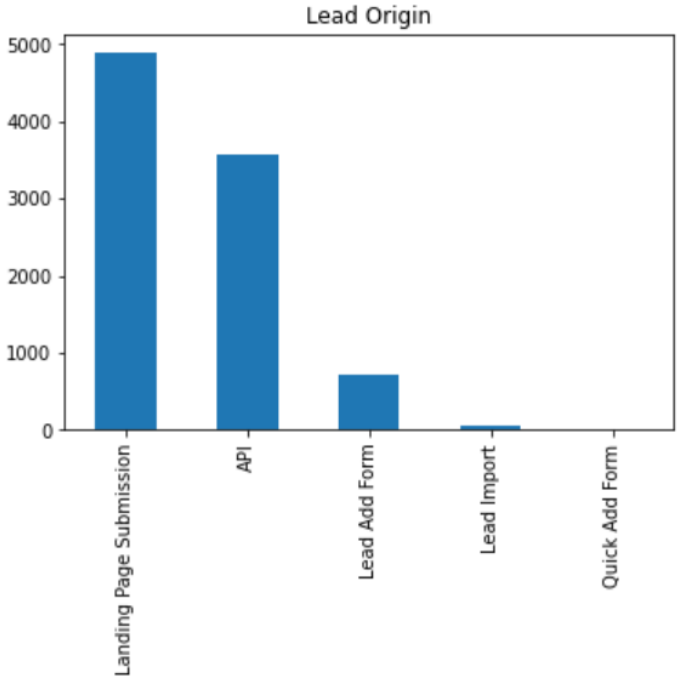
Columns with 1 unique value

MISSING VALUE TREATMENT

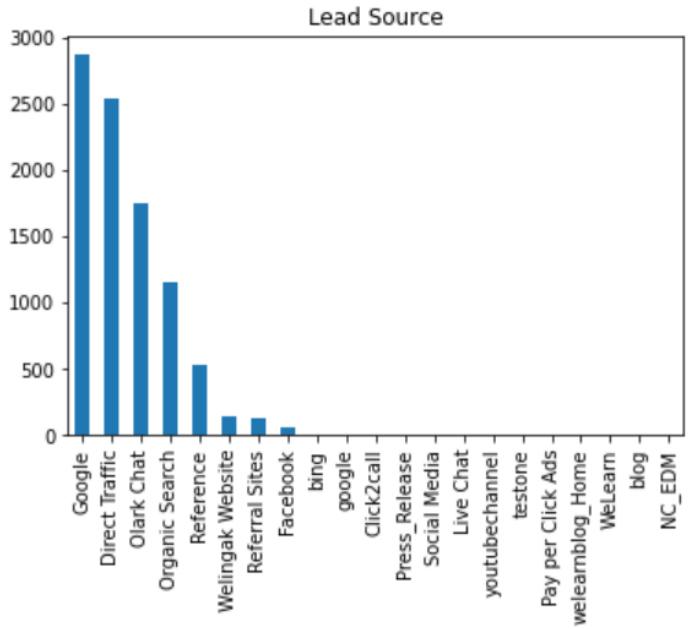
We deleted columns that have >30% missing values, deleted the rows of columns that have <1.5% missing values

imputed missing values with ‘NA’ for columns that had missing value between 1.5% and 30%

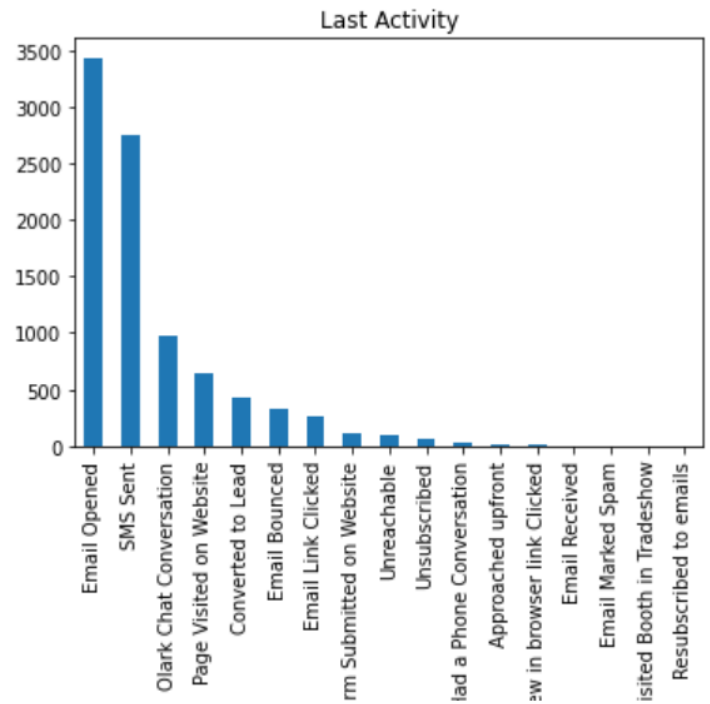
2. Exploratory Data Analysis



Most of the leads had the origin ‘Landing Page Submission’ and ‘API’

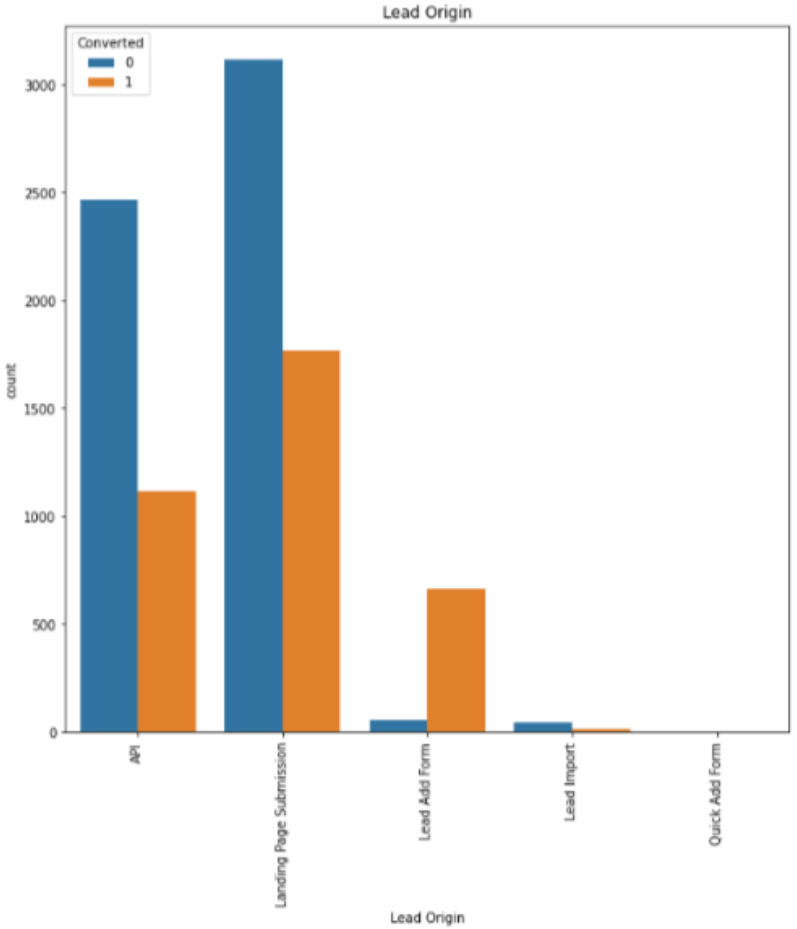


Most of the leads’ source was Google followed by Direct Traffic and Olark Chat

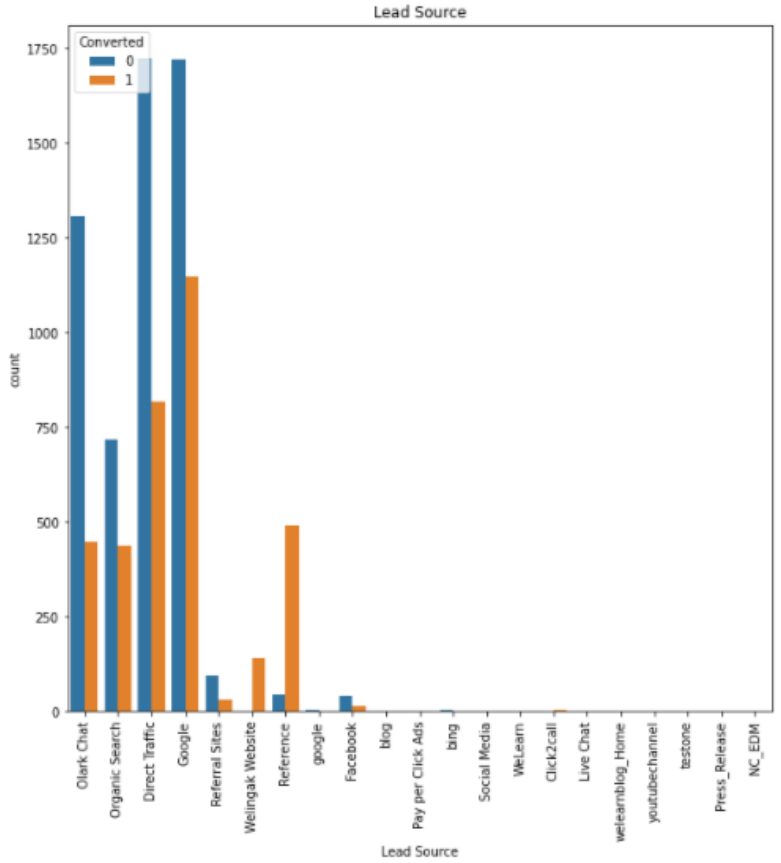


The last activity of most leads was ‘Email opened’ followed by ‘SMS Sent’

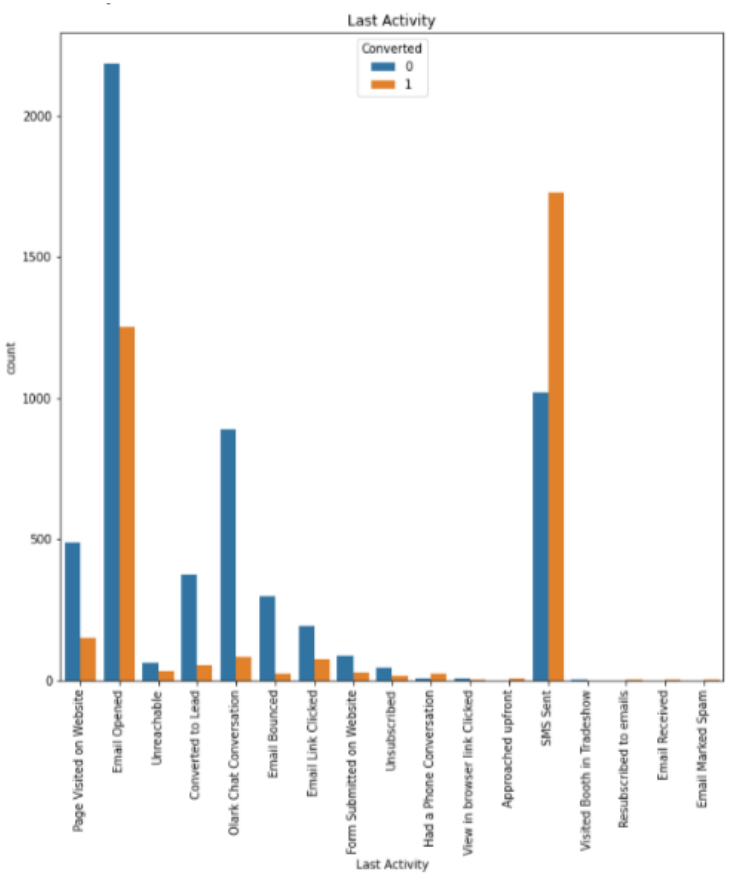
2. Exploratory Data Analysis



Leads originated from ‘Lead Add Form’ have a higher conversion rate

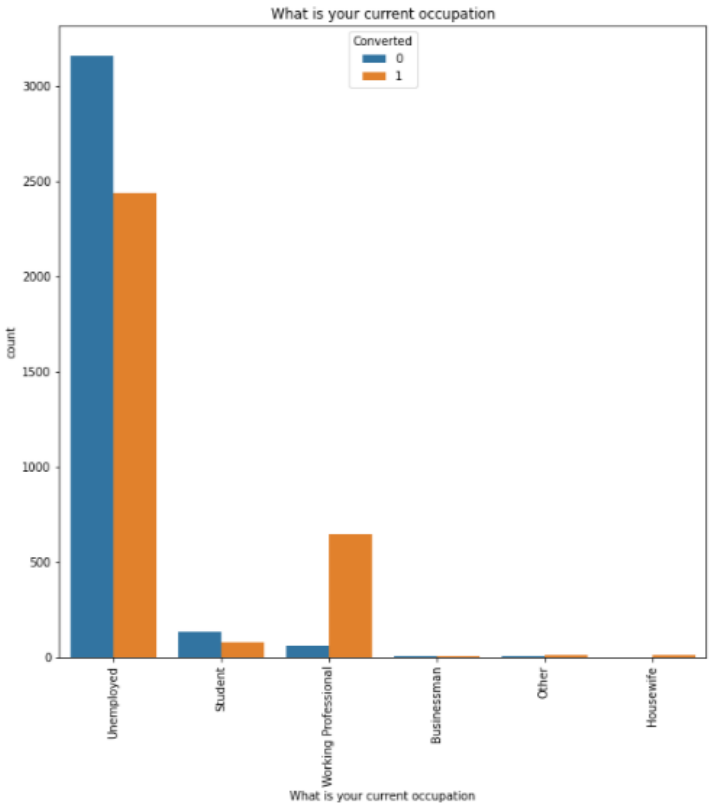


Leads having Source ‘Reference’ and ‘Welingak Website’ have a higher conversion rate

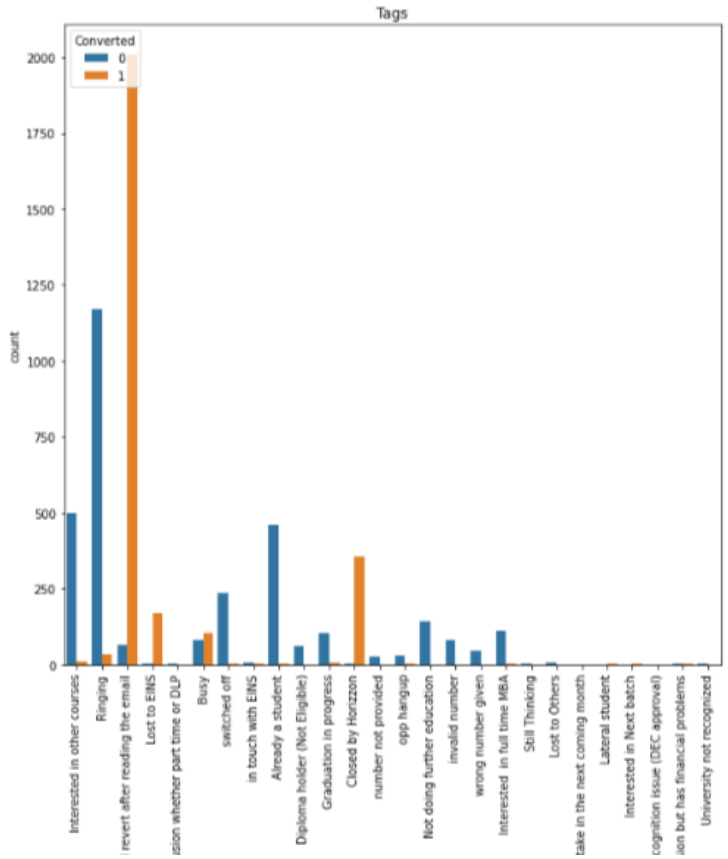


High conversion is seen in leads whose last activity was ‘SMS Sent’

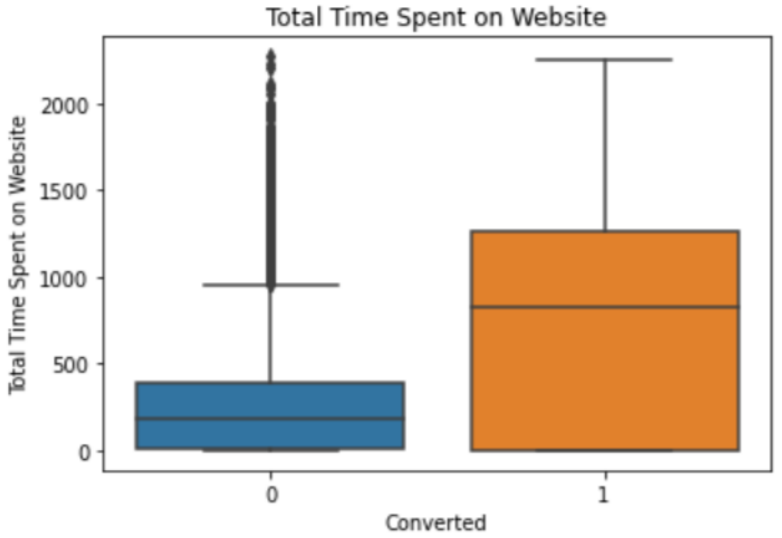
2. Exploratory Data Analysis



Working professionals show high chance of conversion



Leads tagged as ‘Will revert after reading the email’ and ‘Closed by Horizon’ show higher conversion



Converted leads spent on average more time on the website than the others

3. Data Preparation

DUMMY VARIABLES

Binary variables (Yes/No) were converted to 1/0.
Categorical variables were converted to ‘n-1’ dummy variables (n is the unique values of each variable)

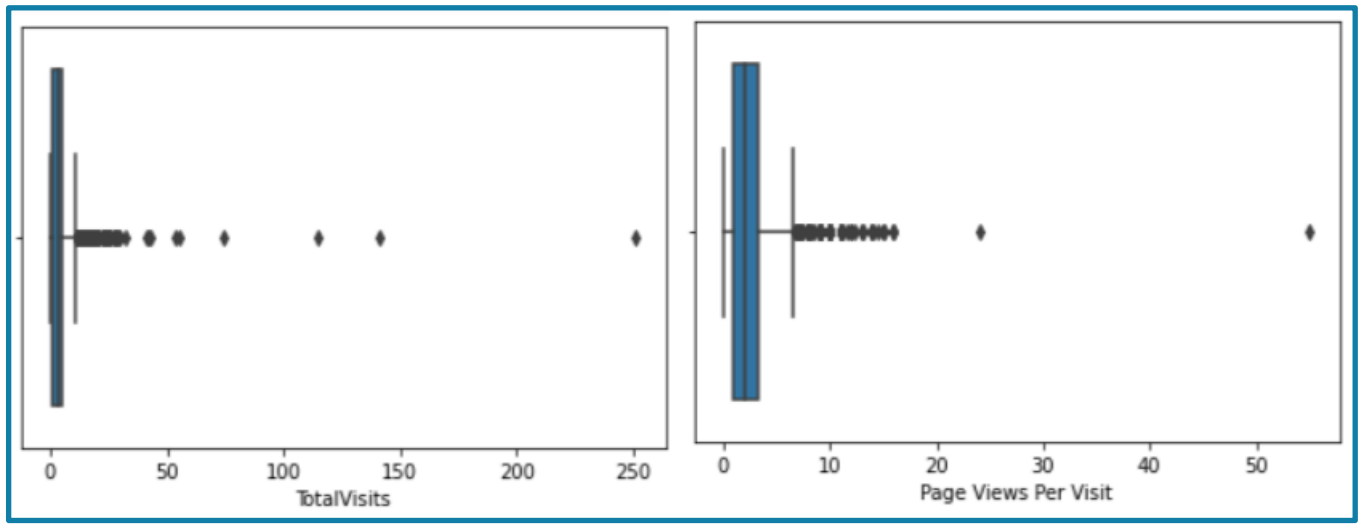
OUTLIER TREATMENT

Outliers in the numeircal column were identified and replaced with 99-percentile value

TRAIN-TEST SPLIT

Train-Test split was done in the ratio 7:3

FEATURE SCALING
We scaled the numerical variables using standard-scaler



Outliers were present in ‘TotalVisits’ and ‘Page Views Per Visit’

4. Model Building

We did RFE on the remaining variables and filtered the top 15 variables.

Then we removed the variables with p-value>0.05 and VIF<5

The Country_NA did not make business sense, so we dropped this variable

For all remaining variables were the p-value< 0.05 and VIF<5, we finalized this model.

	coef	std err	z	P>|z|	[0.025	0.975]
const	-0.8515	0.051	-16.660	0.000	-0.952	-0.751
Do Not Email	-1.3314	0.187	-7.114	0.000	-1.698	-0.965
Total Time Spent on Website	0.9156	0.035	26.066	0.000	0.847	0.984
Lead Origin_Lead Add Form	3.0116	0.208	14.464	0.000	2.604	3.420
Lead Source_Welingak Website	2.8002	1.038	2.698	0.007	0.766	4.835
Last Activity_Converted to Lead	-1.3517	0.214	-6.329	0.000	-1.770	-0.933
Last Activity_Email Bounced	-1.0676	0.405	-2.634	0.008	-1.862	-0.273
Last Activity_Olark Chat Conversation	-0.6862	0.152	-4.503	0.000	-0.985	-0.388
Last Activity_SMS Sent	1.1970	0.075	16.027	0.000	1.051	1.343
What is your current occupation_Working Professional	2.5540	0.198	12.893	0.000	2.166	2.942
What matters most to you in choosing a course_NA	-1.2199	0.087	-14.035	0.000	-1.390	-1.050
Last Notable Activity_Had a Phone Conversation	2.9531	1.144	2.581	0.010	0.711	5.196
Last Notable Activity_Unreachable	2.0178	0.557	3.624	0.000	0.927	3.109

5. Prediction and Evaluation on Train data

ROC CURVE

The Area under ROC curve is 0.88 which is a good value

OPTIMAL CUT-OFF

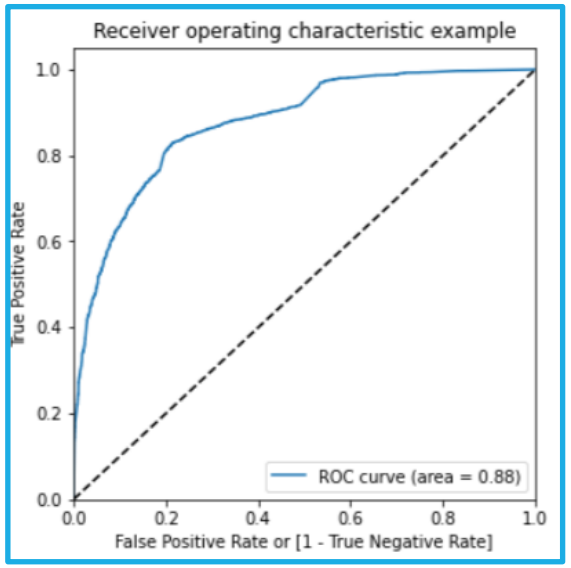
Using the Accuracy, Sensitivity and Specificity curve, we found the optimal cut-off to be 0.35

PRECISION-RECALL CURVE

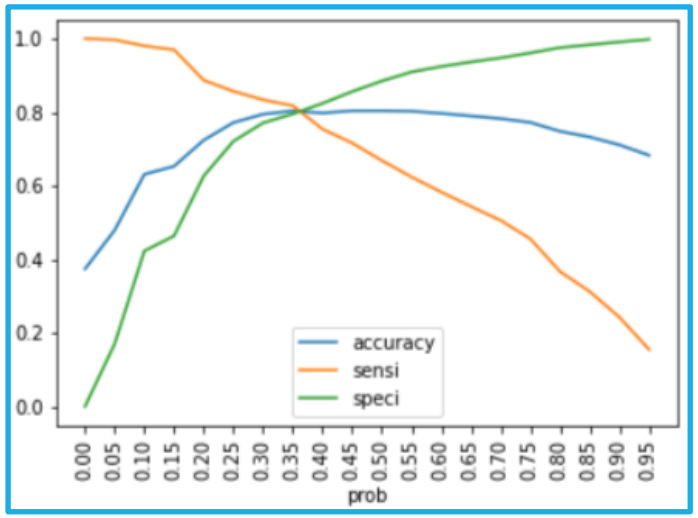
using the precision-recall curve, the optimal cut-off was 0.45. But to keep a higher recall value, we took cut-off=0.35

EVALUATION METRICS

We have a good accuracy of 80%. The recall is also very good at 81%. Precision is 70%



ROC Curve



Precision-Recall Trade-off curve

```
The confusion matrix is :
 [[3162  815]
 [ 433 1941]]
================================================
Accuracy                  : 0.8034955125177138
Sensitivity/Recall        : 0.8176074136478517
Specificity               : 0.7950716620568268
False positive rate       : 0.20492833794317325
False predictive value    : 0.7042815674891146
Negative predictive value : 0.8795549374130737
Precision                 : 0.7042815674891146
```

Evaluation Metrics

6. Prediction and Evaluation on Test data

On predicting the conversion on the test data we got a good accuracy of 79.55% with an 81.24% Recall.

Since we do not want to miss out on conversion opportunities, the trade-off between Recall and Precision was inclined towards Recall (High Recall and low precision) but we still managed to get a good precision of 70.66%

We can see that the accuracy has dropped when compared to the train data, but only by 1% (<5%). This is a good value.

```
The confusion matrix is :
 [[1304  358]
 [ 199  862]]
================================================
Accuracy                    : 0.7954461990451708
Sensitivity/Recall          : 0.8124410933081998
Specificity                 : 0.7845968712394705
False positive rate         : 0.21540312876052947
False predictive value      : 0.7065573770491803
Negative predictive value   : 0.867598137059215
Precision                   : 0.7065573770491803
```

Evaluation Metrics

7. Conclusion

After evaluating the model we can conclude that the Probability conversion can be given by the equation -

```
Conversion Probability =
  - 0.8515
  - 1.3314 * Do Not Email
  + 0.9156 * Total Time Spent on Website
  + 3.0116 * Lead Origin_Lead Add Form
  + 2.8002 * Lead Source_Welingak Website
  - 1.3517 * Last Activity_Converted to Lead
  - 1.0676 * Last Activity_Email Bounced
  - 0.6862 * Last Activity_Olark Chat Conversation
  + 1.1970 * Last Activity_SMS Sent
  + 2.5540 * What is your current occupation_Working Professional
  - 1.2199 * What matters most to you in choosing a course_NA
  + 2.9531 * Last Notable Activity_Had a Phone Conversation
  + 2.0178 * Last Notable Activity_Unreachable
```

The top 3 variables in our model that contribute the most towards the probability of a lead getting converted are –

1. Last Activity
2. Total Time Spent on Website
3. Lead Origin

The top 3 categorical/dummy variables in the model which should be focused the most on in order to increase the probability of lead conversion are -

1. Lead Origin_Lead Add Form (Coeff: 3.0116)
2. Last Notable Activity_Had a Phone Conversation (Coeff: 2.9531)
3. Lead Source_Welingak Website (Coeff: 2.8002)

7. Lead Scoring (0-100)

Based on the conversion probability, we multiply the probability values by 100 to get the lead scores

	Prospect ID	Lead Score
0	7927b2df-8bba-4d29-b9a2-b6e0beafe620	15.86
1	2a272436-5132-4136-86fa-dcc88c88f482	36.84
2	8cc8c611-a219-4f35-ad23-fdfd2656bd8a	71.06
3	0cc2df48-7cf4-4e39-9de9-19797f9b38cc	23.91
4	3256f628-e534-4826-9d63-4a8b88782852	34.81

THANK YOU