Lead Scoring Case Study

TEAM MEMBERS

DS C70-DEEPTI GUPTA, VAISHNAVY MUTHUKRISHNAN AND DHRUV SINGH

Business Objectives

1. Analyse the conversion data of a education company named X Education that sells online courses to industry professionals.

2. Identify the lead attributes and their actions on landing of the company website which lead to lead conversion.

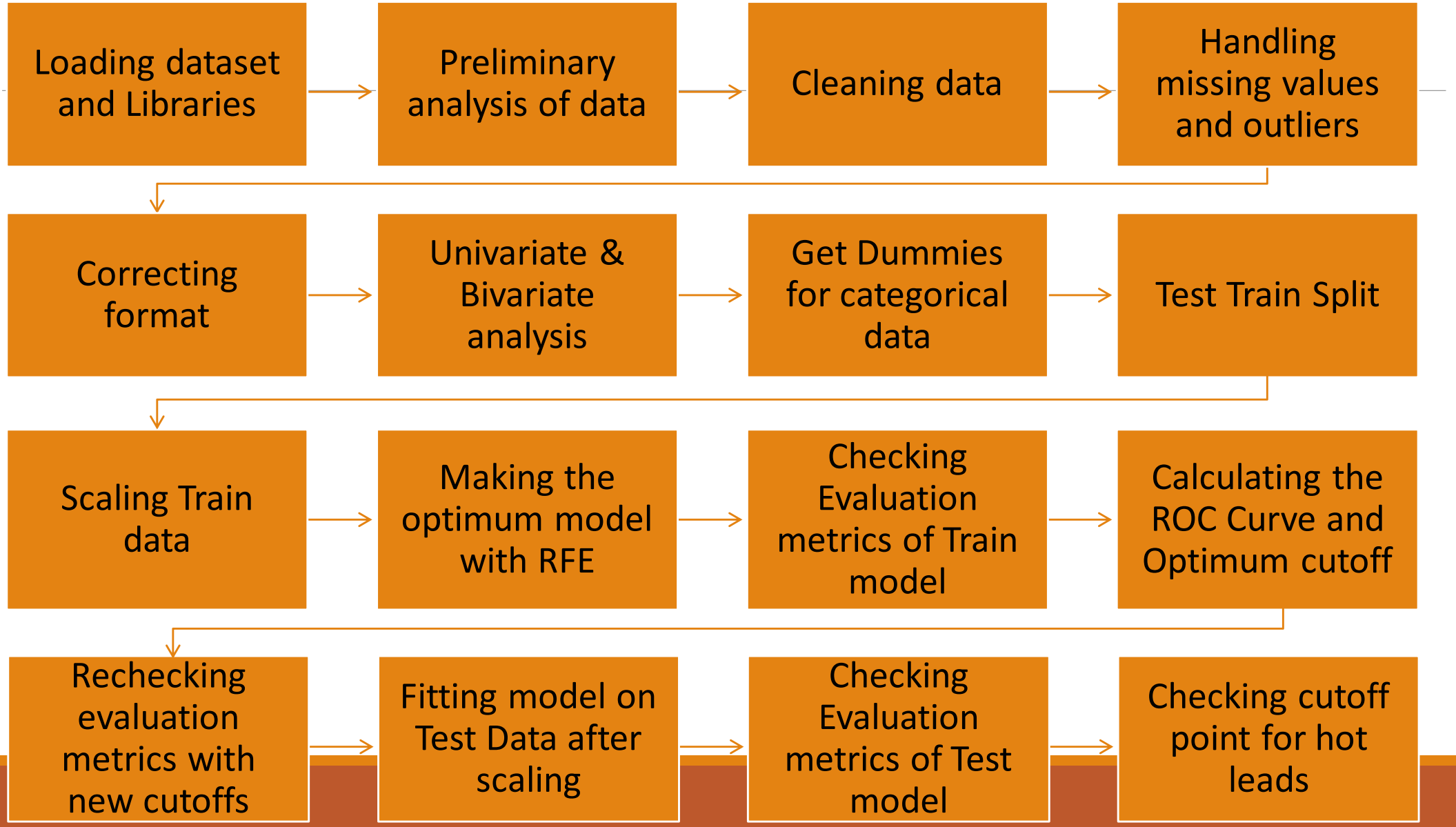
3. The typical lead conversion rate at X education is around 30%.

4. Use these patterns, which are predictive of conversion, for making logistic regression model for increasing conversion of potential leads to have lead conversion at about 80%.

5. Generate lead score between 0 and 100 for each of the leads to identify and focus on most potential leads, also known as '**Hot Leads**'.

6. Ensure that 'hot leads' are 100% converted leading to higher conversion.

Methodology



Loading Dataset and Relevant Libraries

- Loaded the data into Jupyter notebook from the csv file.
- Imported relevant libraries
  - Numpy
  - Pandas
  - Matplotlib
  - Seaborn
  - Warnings (to ignore warnings given by Jupyter notebook)
  - sklearn.model_selection (for train_test_split)
  - sklearn.preprocessing (for StandardScaler OR MinMaxScaler)
  - sklearn.linear_model (for LogisticRegression)
  - sklearn.feature_selection (for RFE)
  - statsmodels.stats.outliers_influence (for variance_inflation_factor)
  - sklearn (for metrics)
  - sklearn.metrics (for confusion_matrix , precision_score, recall_score and precision_recall_curve)

Preliminary Analysis of Conversion Data

Data contains information about approximately 9000 leads who may either have come to explore courses on the company website or have been referred by past alumini.

After contact by the sales team, the lead may:

1. Get **converted** to a paying customer : Denoted by 1

2. Remain **unconverted** : Denoted by 0

Shape of dataframe : (9240, 37)

37 columns = 30 Categorical features + 7 Numerical features

dtypes: float64(4), int64(3), object(30)

Cleaning Conversion Data:
Null Values

1. Identified count and % of null values across columns with df.isnull().sum() function
2. Dropped columns with more than 35% null values after ensuring that they do not include any important columns
3. Dropped Lead Number column as it was not helpful in building model.
4. Replaced ‘Select’ value in certain columns where the user had not chosen any option while filling the website form with NaN value.
5. Recalculated the % of null values
6. Dropped columns with more than 40% null values and rows with 1.5% null values
7. Checked rest of columns one by one for null values and handled them by:
   - Imputing more than 30% null values in column by replacing with ‘Missing’

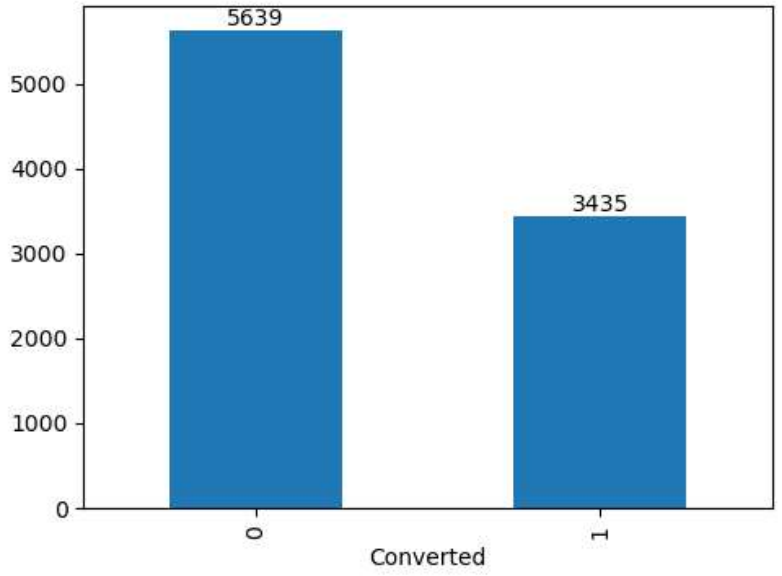
Cleaning Current Application Data:
Outliers, Variance Check and Correcting Format

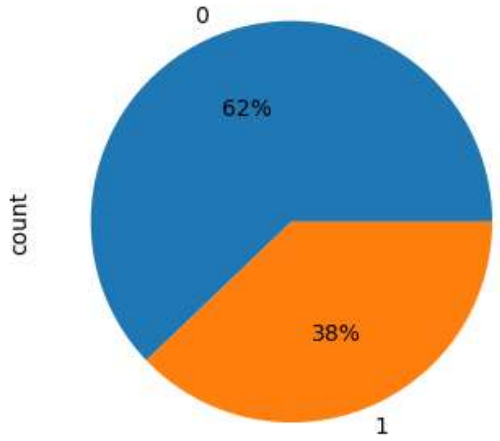
1. Checked the columns with skewed distribution for outliers by plotting histograms, boxplots and using the describe () function.
2. Removed outliers by:
    - Dropping few extreme outliers which were positioned extremely far from the rest of the data.
3. Checked the columns for unique values with barplots and removed columns with only 1 unique value or where more than 95% of the column values consisted of only 1 value as they columns had limited variance.
4. Corrected the case of categories to converge similar data
5. Collapsed categories with less than 1% data to a common consolidated category

Univariate & Bivariate Analysis

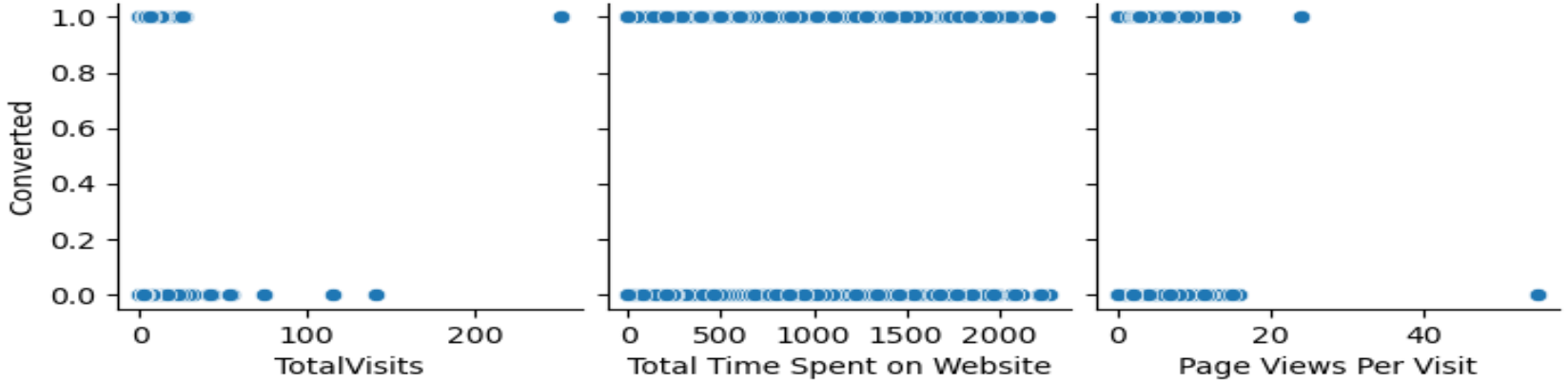
ANALYSED EACH VARIABLE IN TERMS OF DISTRIBUTION, UNIQUE VALUES AND COMPARED WITH THE CONVERTED VARIABLE

Graph 1: Converted (Target Variable) – Barplot and Pie chart



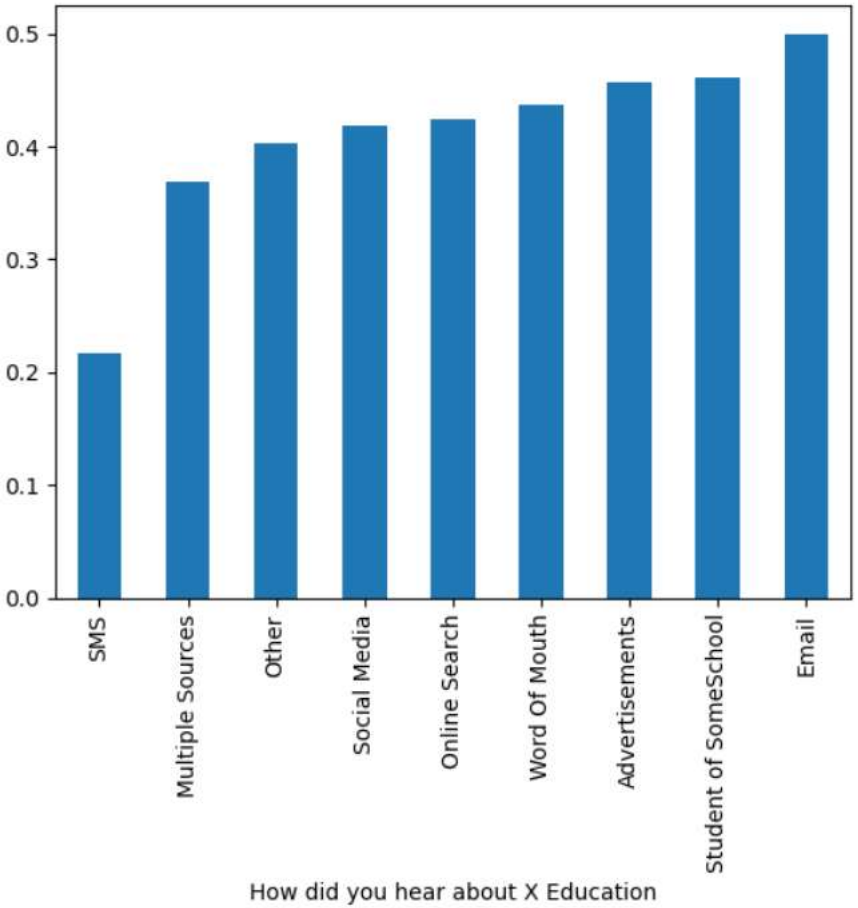


Graph 2: Converted with Numerical Variables - Pairplot

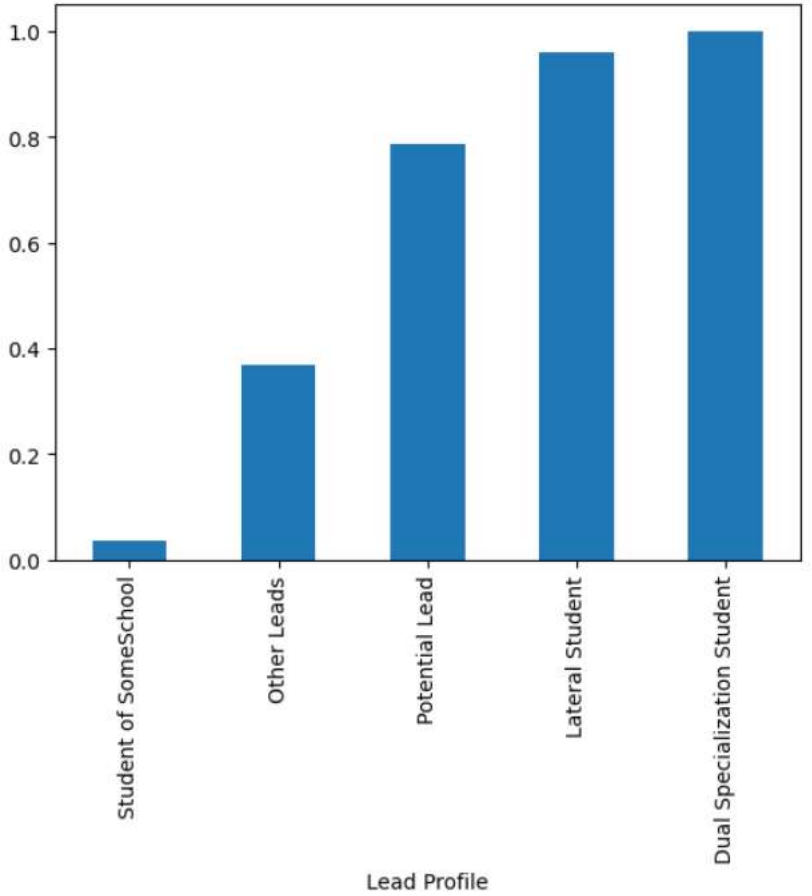


- 'Converted' variable indicates whether a lead has been successfully converted or not.
- It has only 2 unique values: 0 (Not Converted) and 1 (Converted)
- With **imbalance ratio at 0.61**, the dataset is moderately imbalanced with Majority class being unconverted and Minority class being converted.

Graph 1: How did you hear about X Education groupby with Converted



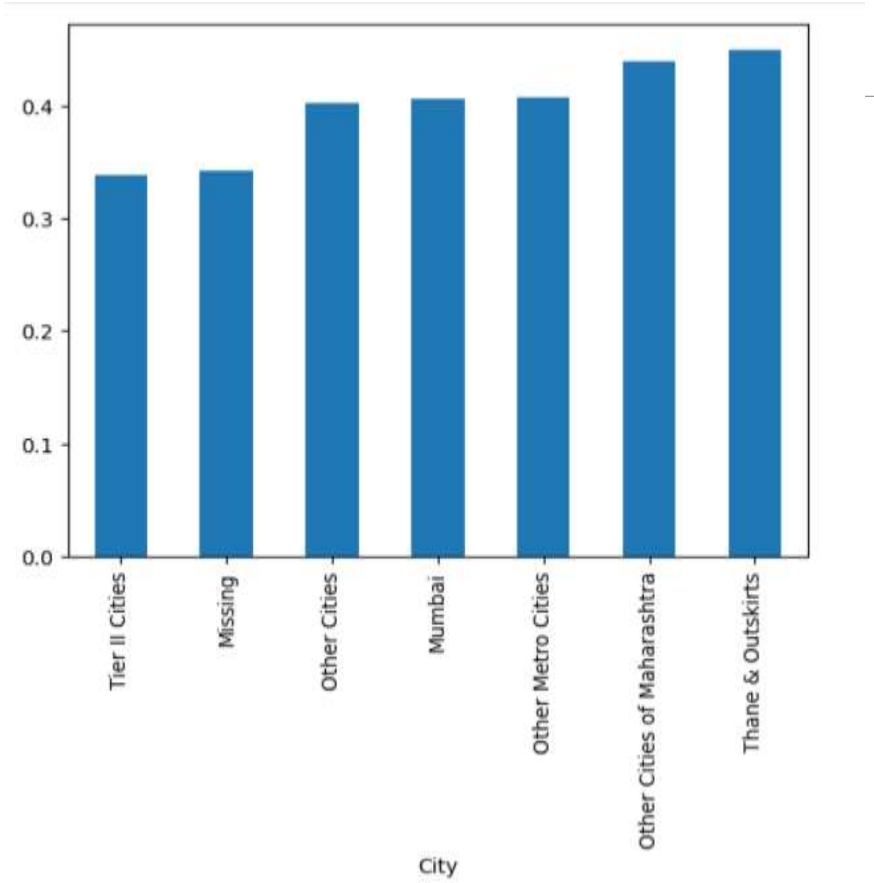
Graph 2: Lead Profile groupby with Converted



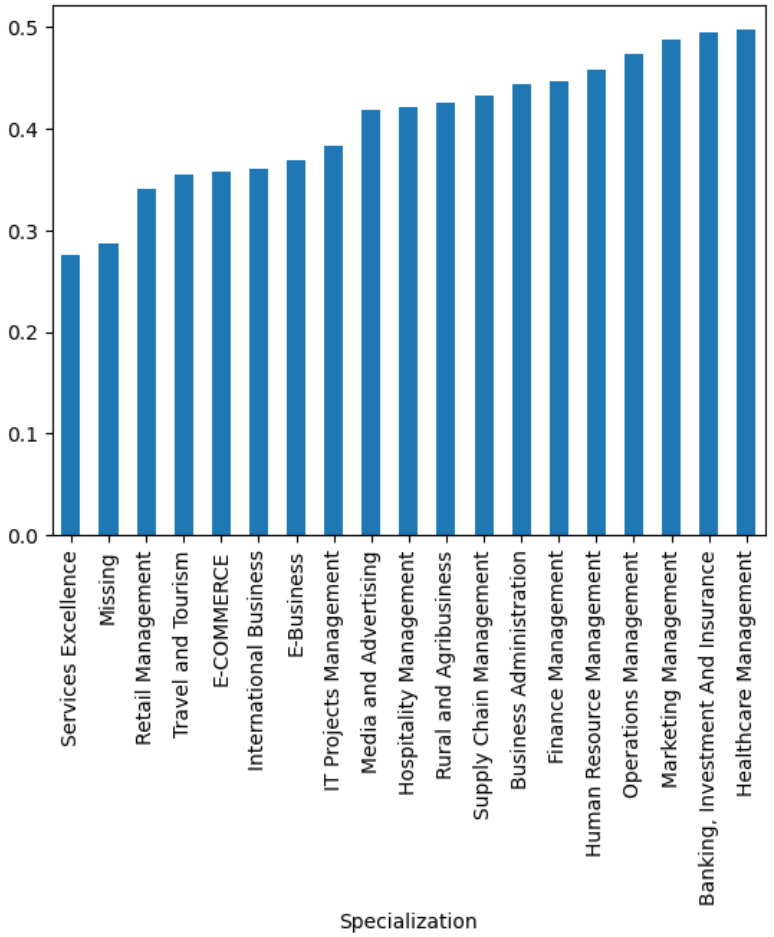
Graph 1 shows than leads showing highest conversion came to hear about X education through email.

Graph 2 shows that highest conversions seem to be among students with Dual specialization and those who were lateral students

Graph 1: City groupby with Converted



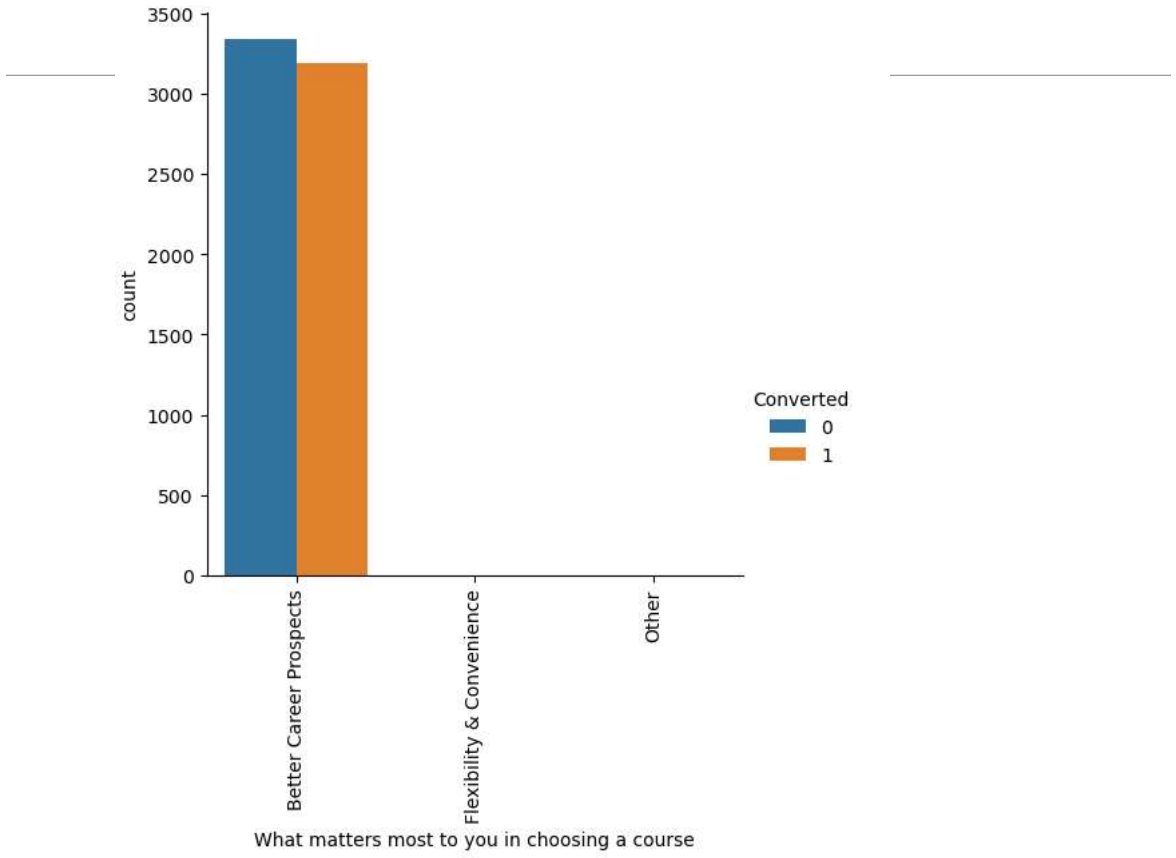
Graph 2: Specialization groupby with Converted



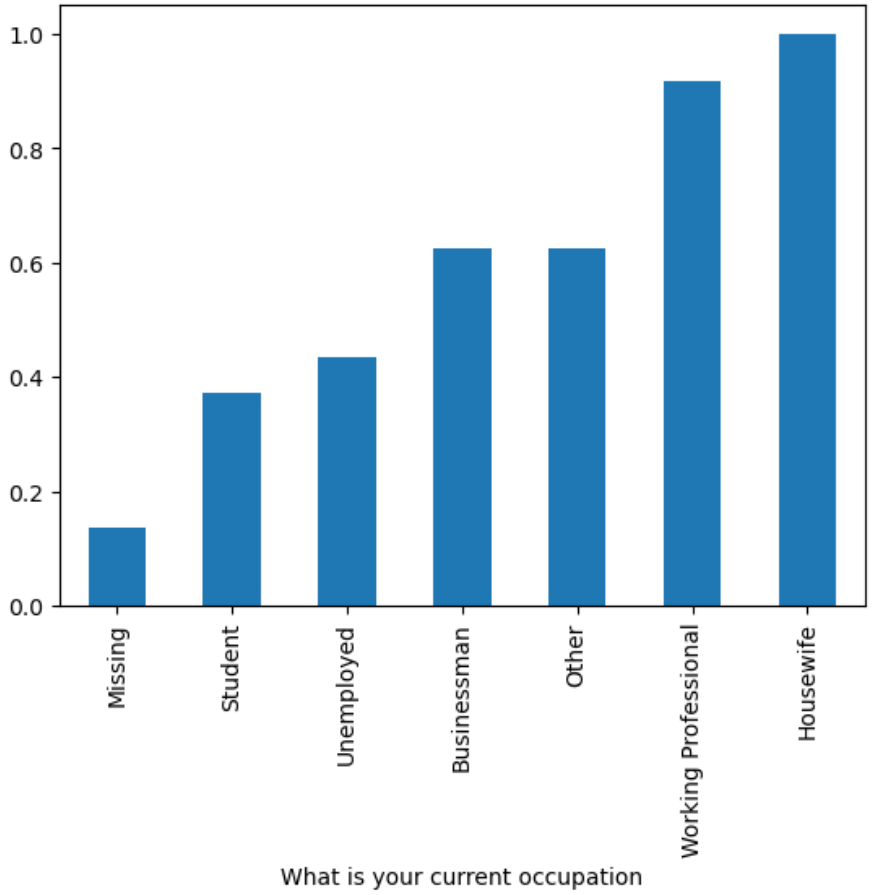
Graph 1 shows than leads showing highest conversion belong to Mumbai and surrounding areas.

Graph 2 shows that highest conversions seem to be among leads with core business specialization (Healthcare mgmt. / Banking, Investment & Insurance / Ops) and those in support functions (HR mgmt. / Finance mgmt.)

Graph 1: What matters most to you in choosing a course catplot with Converted



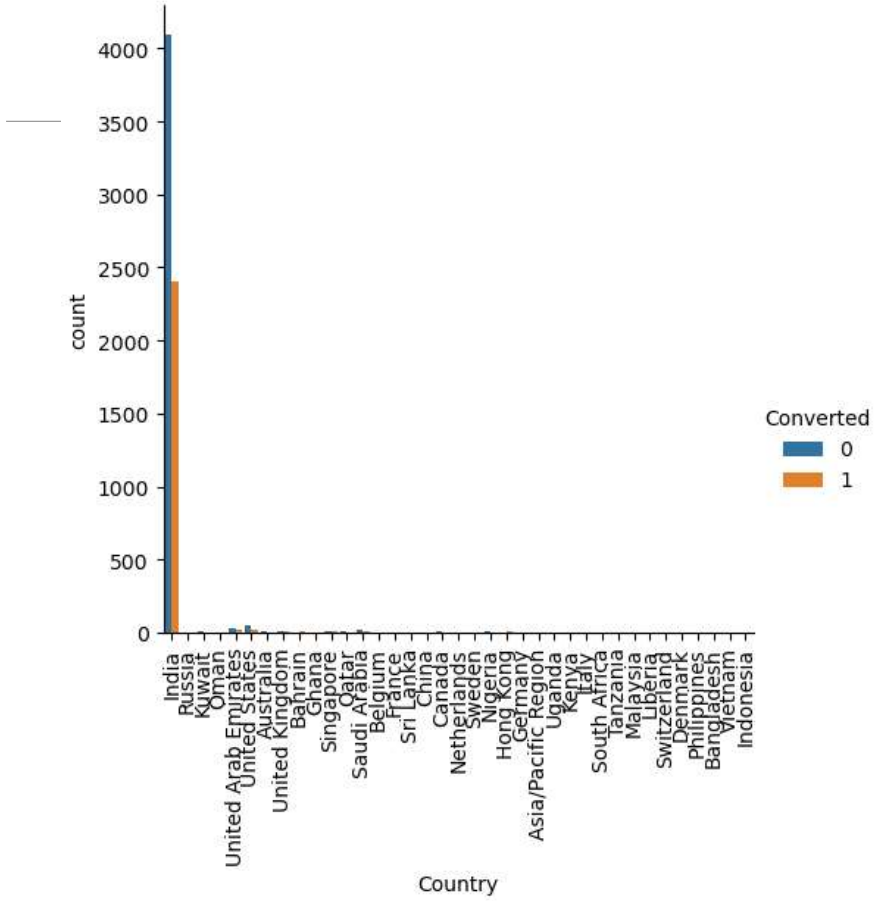
Graph 2: What is your current occupation groupby with Converted



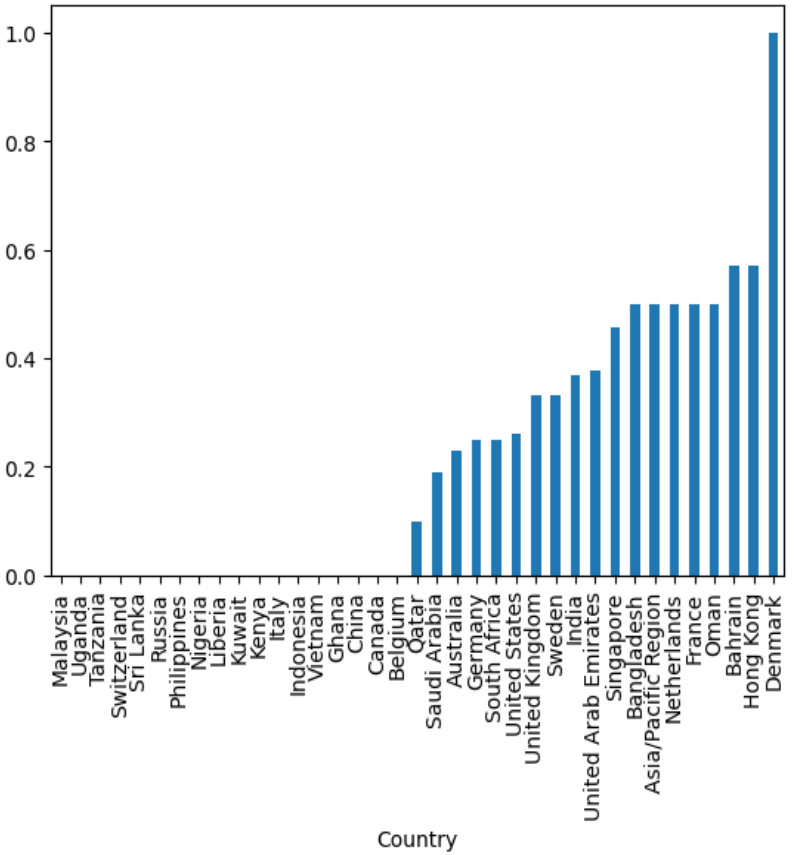
Graph 1 shows than leads with highest conversion chose the option better career prospects with only 0.033 choosing other options.

Graph 2 shows that though unemployed leads were highest in number, the highest conversions seem to be among leads who are housewives or are working professionals.

Graph 1 : Country catplot with Converted



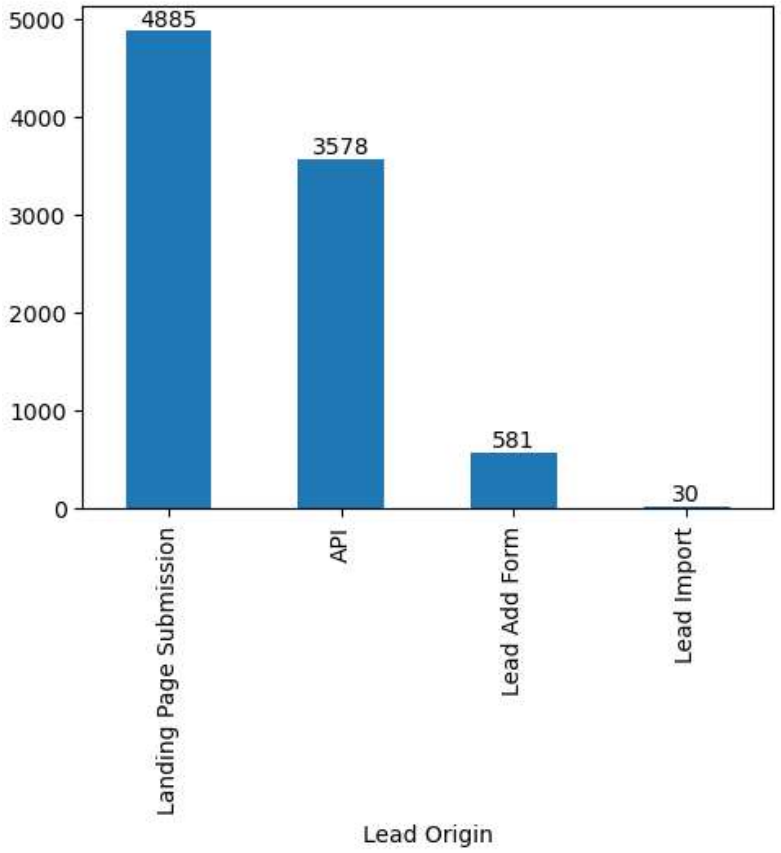
Graph 2: Country groupby with Converted



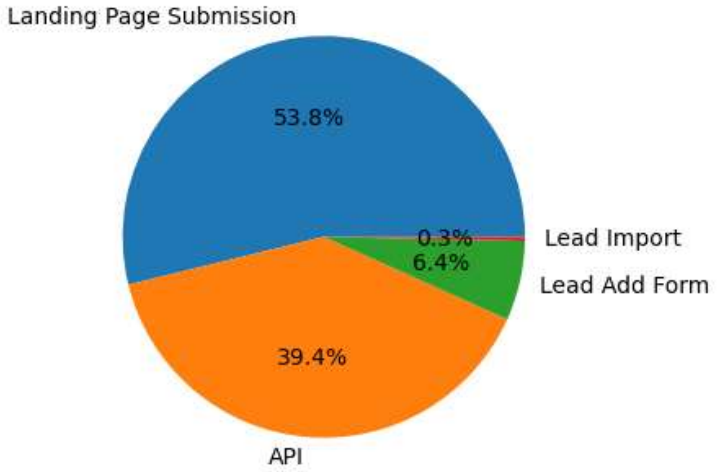
Graph 1 shows than most leads were Indias with very few belonging to other countries.

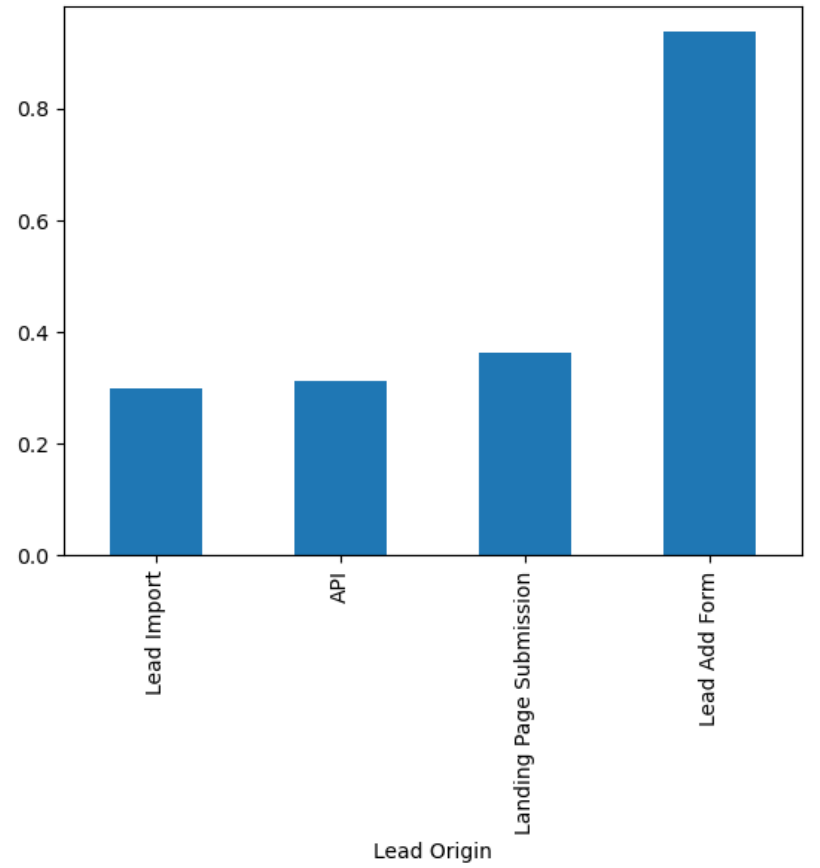
Graph 2 shows that highest conversions seem to be among leads from Denmark and Middle Eastern countries.

Graph 1: Lead Origin distribution



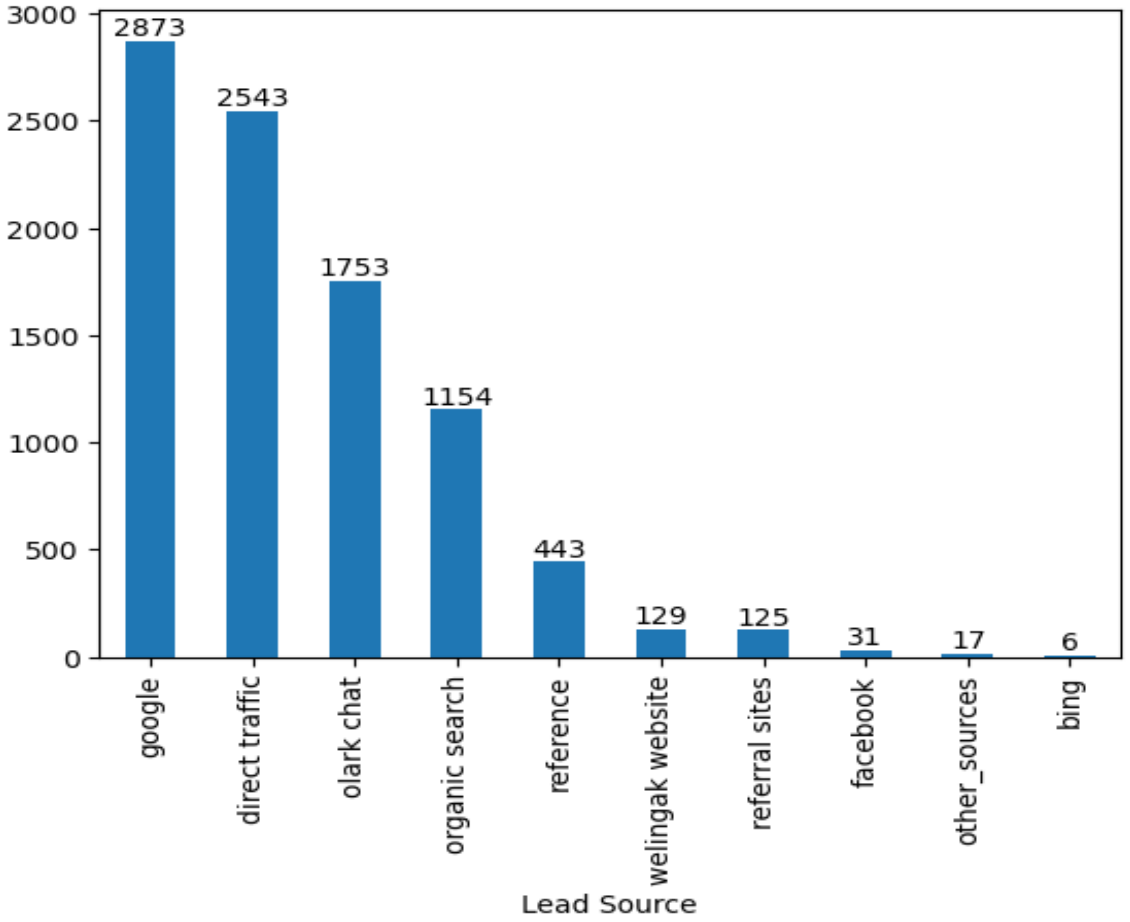
Graph 2: Lead Origin groupby with Converted



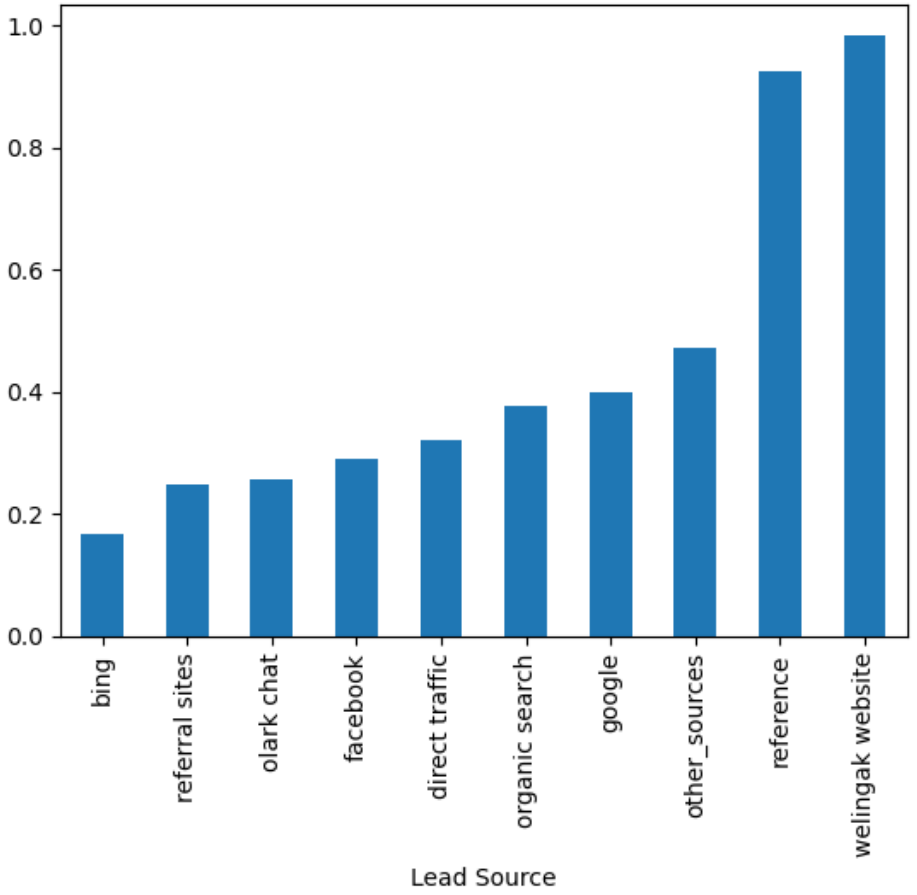


Graph 1 shows than highest leads were identified from the Landing page submission.
Graph 2 shows that highest conversions seem to be among leads who added form.

Graph 1: Lead Source distribution



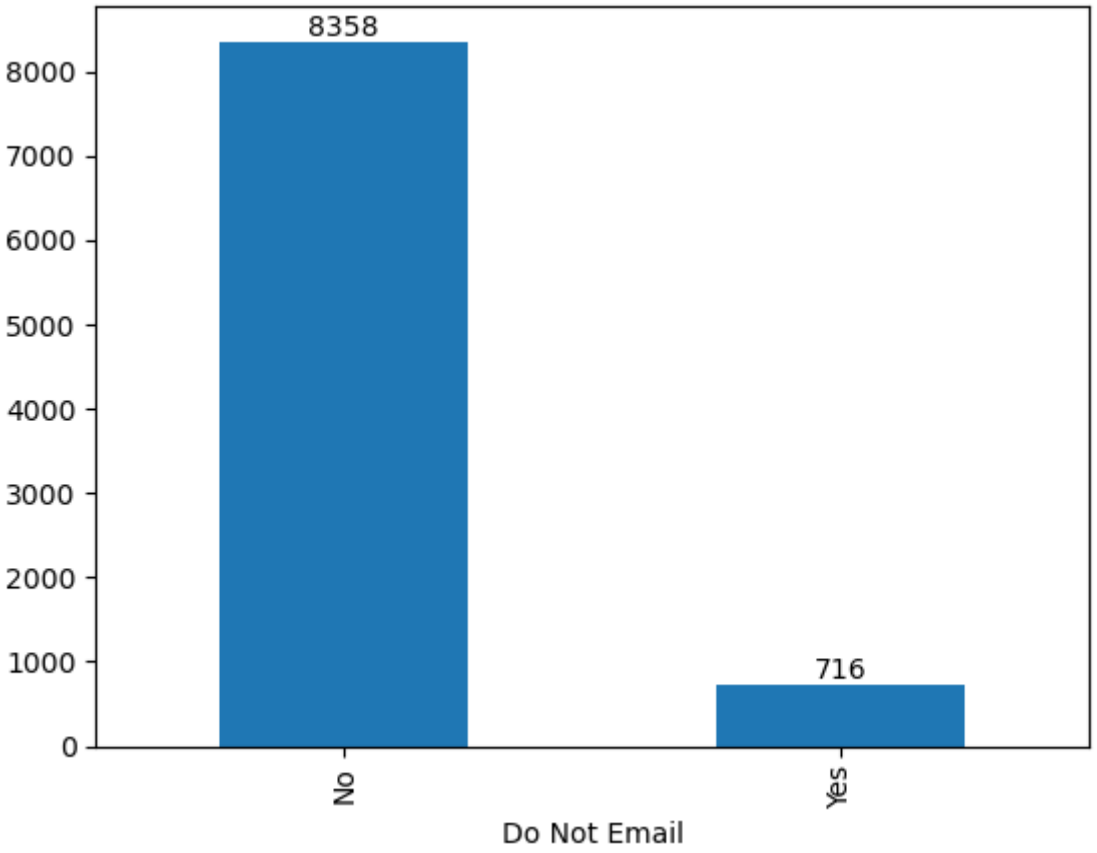
Graph 2: Lead Source groupby with Converted



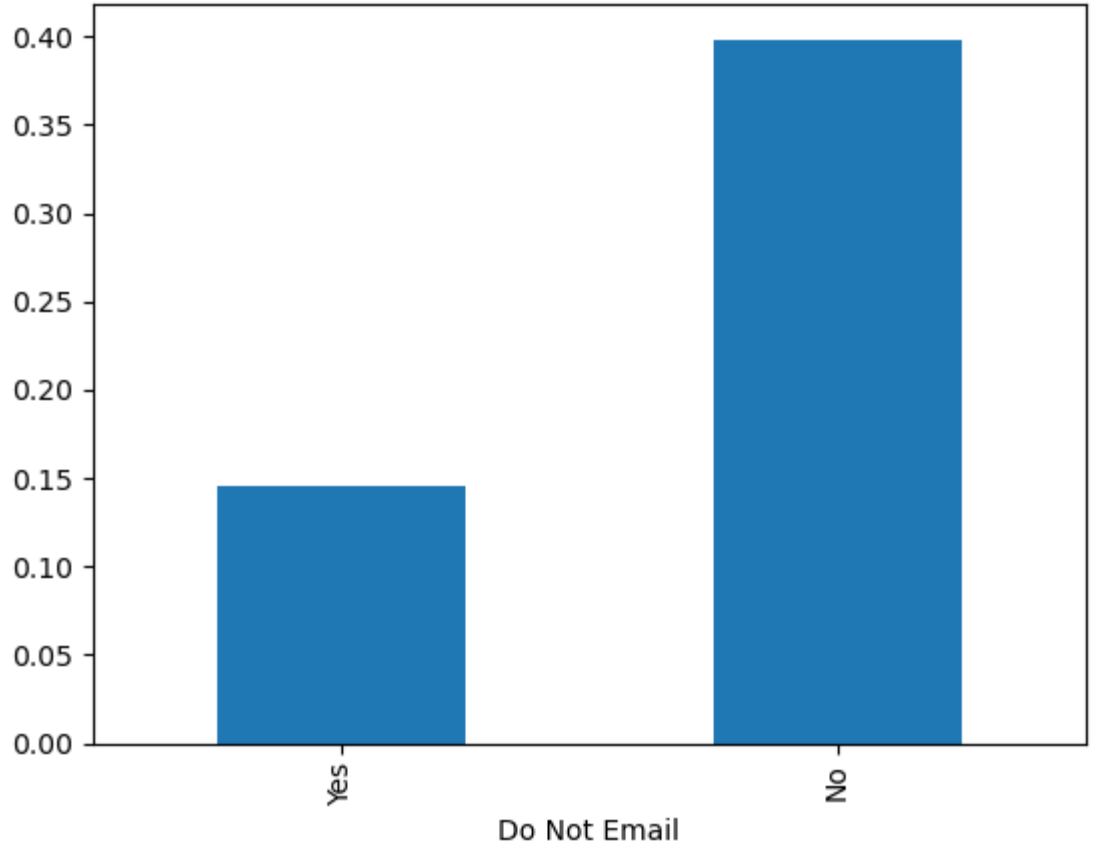
Graph 1 shows than highest leads were sourced from Google.

Graph 2 shows that highest conversions seem to be among leads sourced from welingak website

Graph 1: Do Not Email distribution

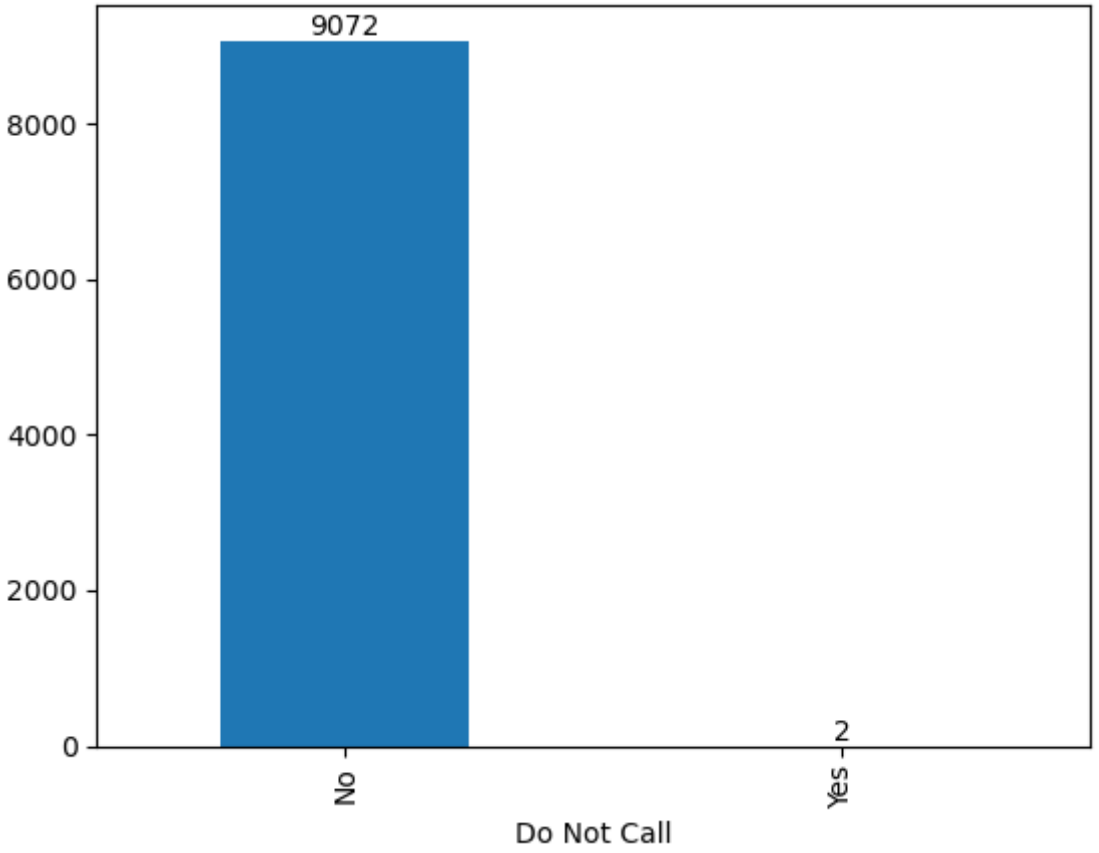


Graph 2: Do Not Email groupby with Converted

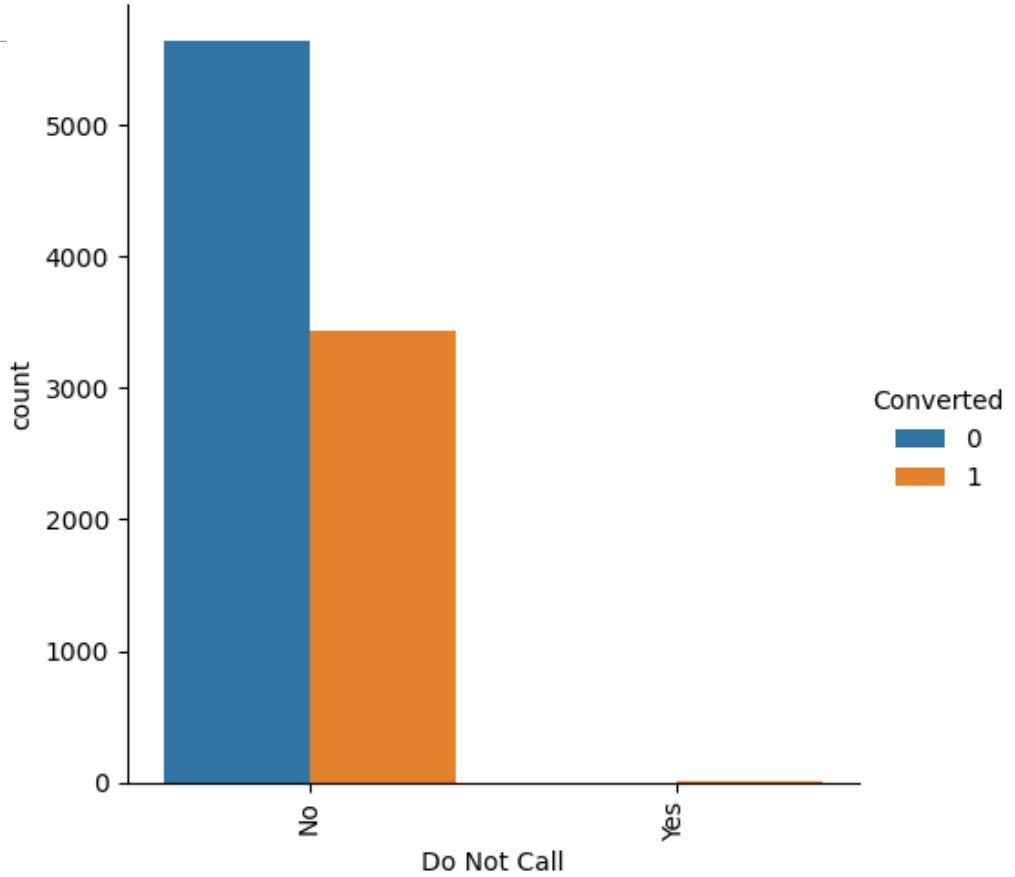


Graph 1 shows than highest leads chose the option ‘No’ for Do not Email.
Graph 2 shows that highest conversions are also among the leads who chose the option ‘No’

Graph 1: Do Not Call distribution

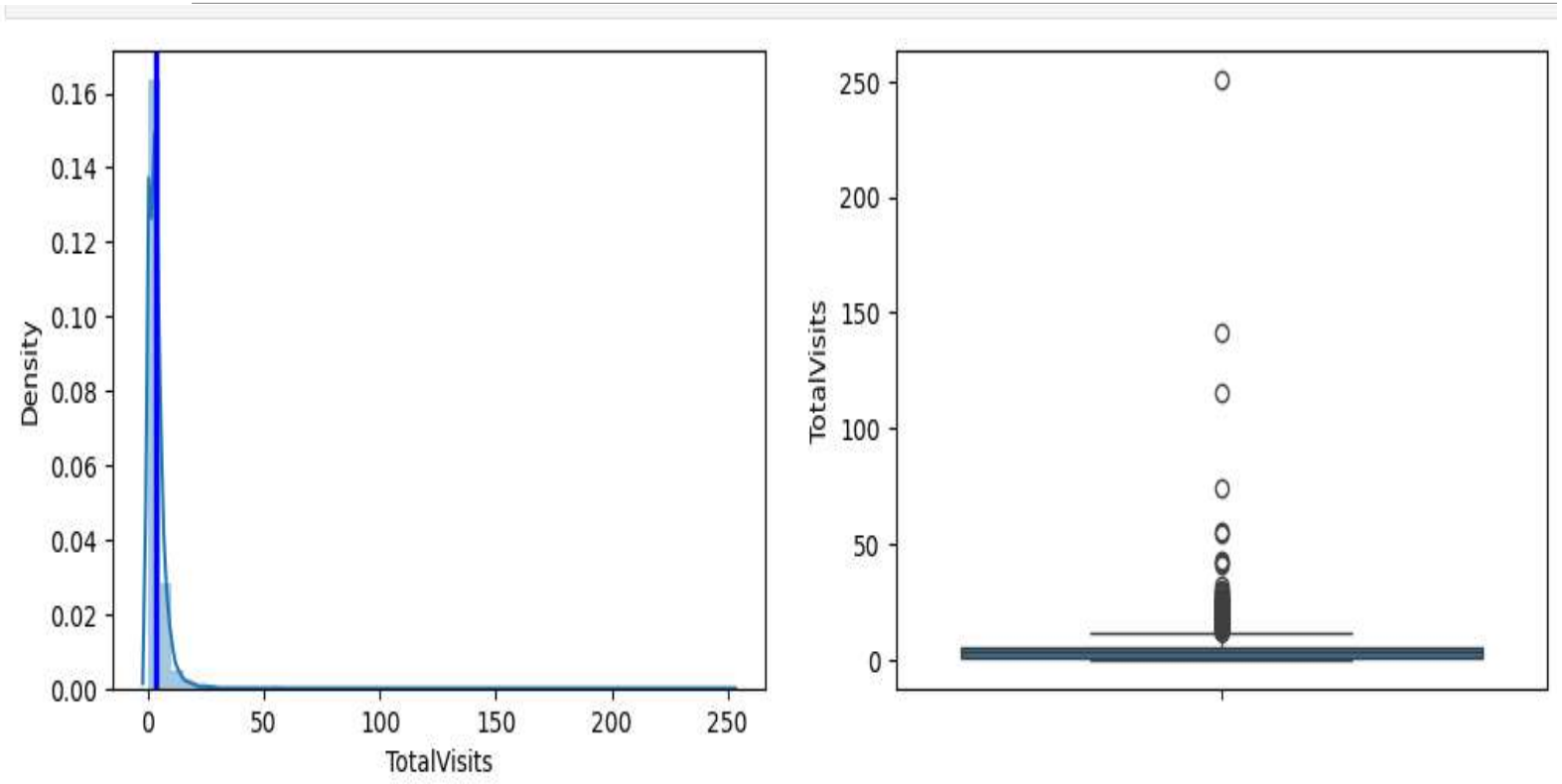


Graph 2: Do Not Call catplot with Converted

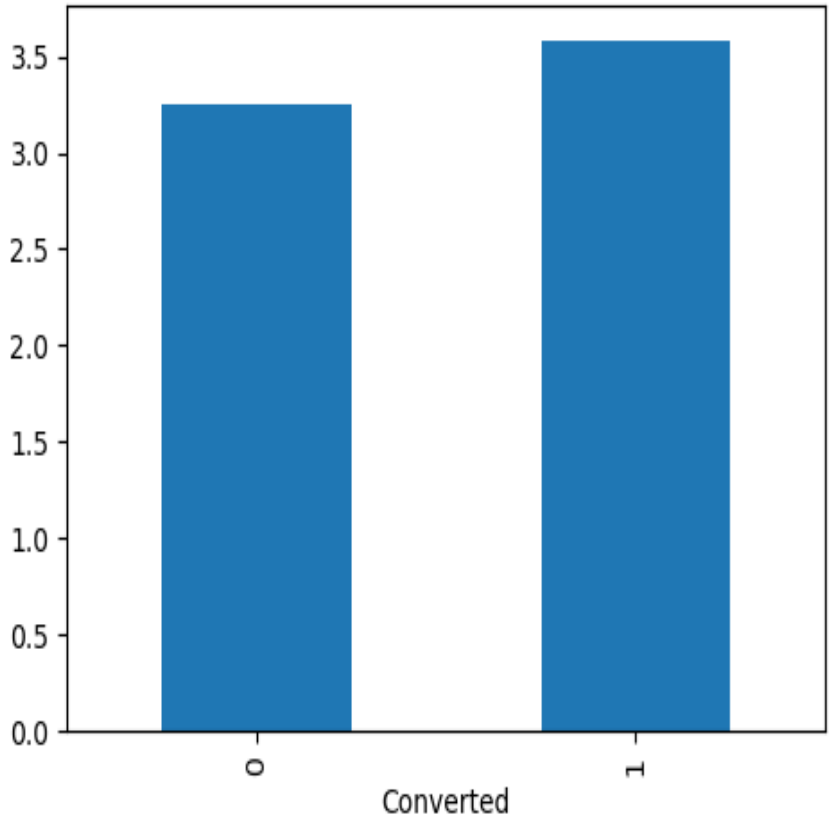


Graph 1 shows that highest leads chose the option ‘No’ for Do not call while only 2 leads chose ‘Yes’.
Graph 2 shows that highest conversions are also among the leads who chose the option ‘No’

Graph 1: Total Visits distribution



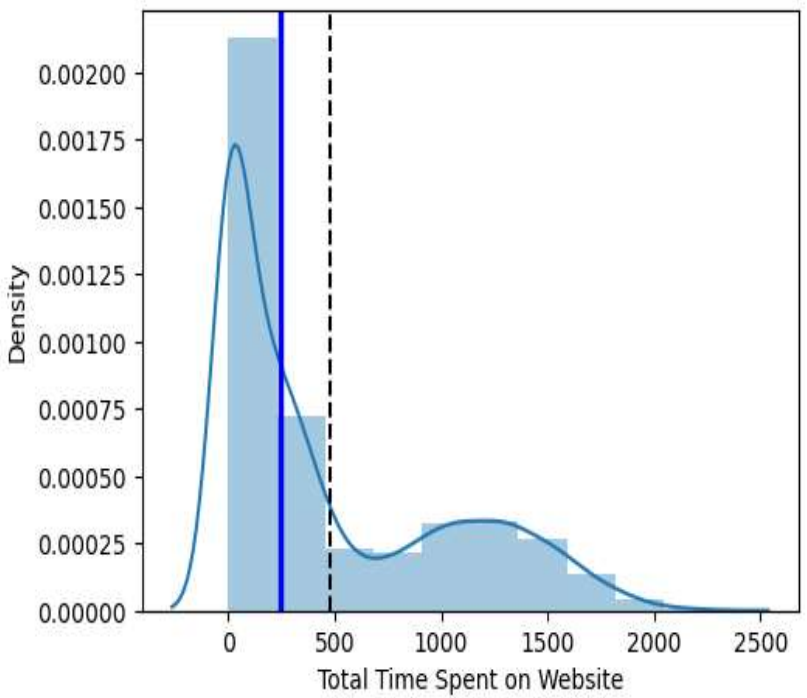
Graph 2: Total Visits groupby with Converted



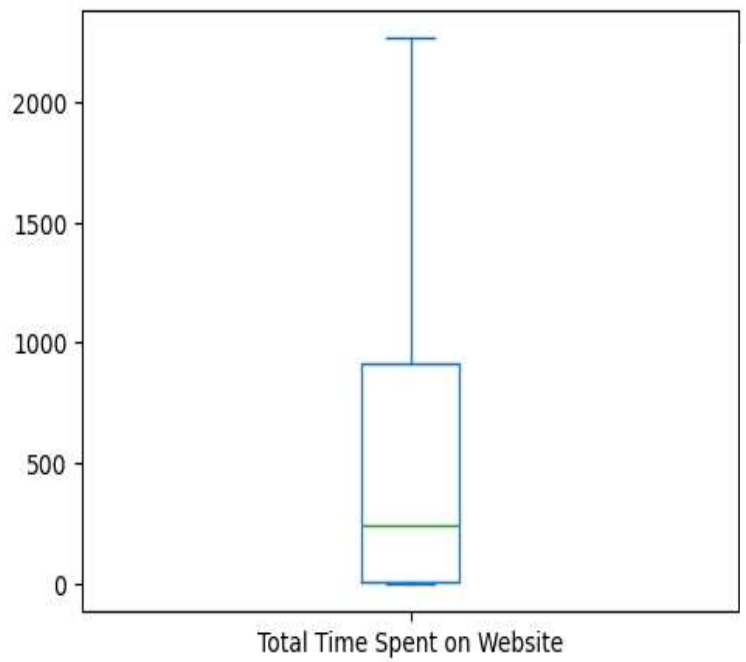
Graph 1 shows than highest leads visited the website 0 number of times.

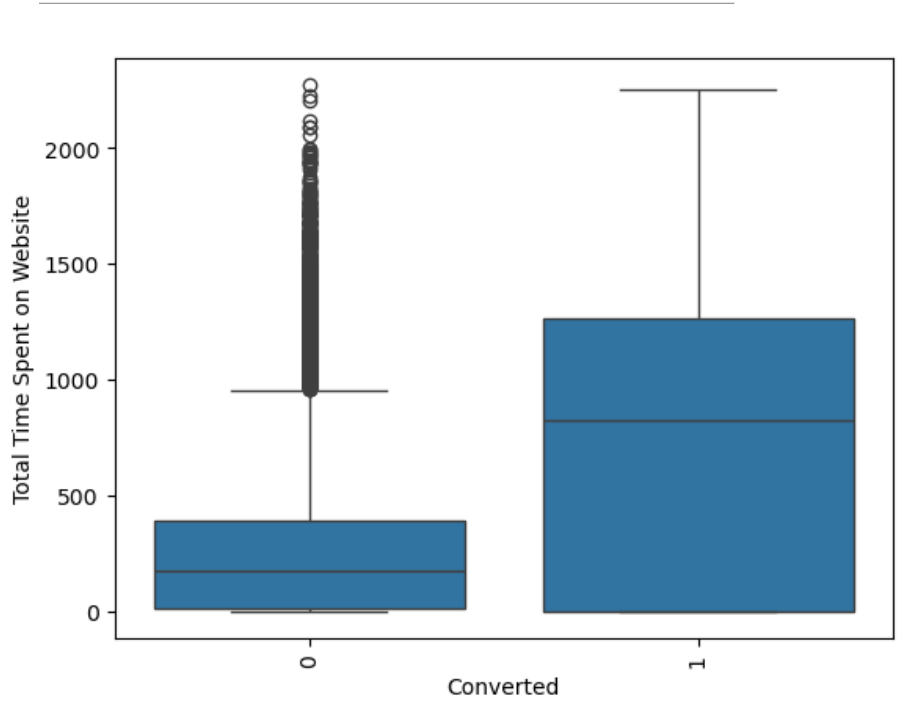
Graph 2 shows that highest conversions are among the leads who visited the website about 3.5 times

Graph 1: Total Time Spent on Website distribution



Graph 2: Total Time Spent on Website boxplot with Converted

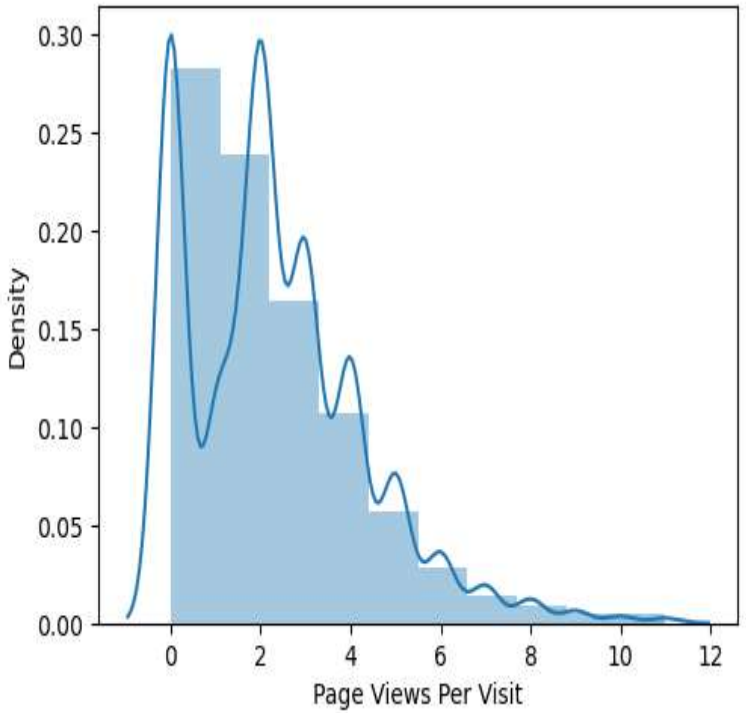


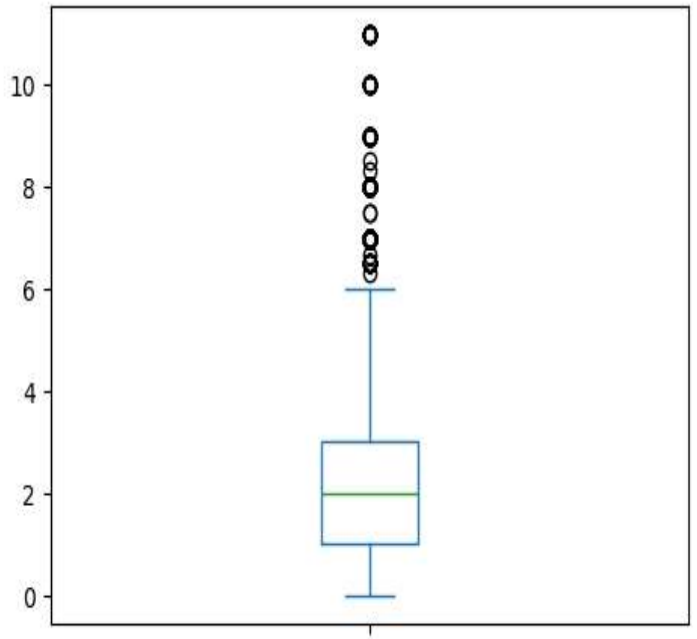


Graph 1 shows than highest leads spent 0 units of time on the website with a small minority having spent more than 1500 units of time.

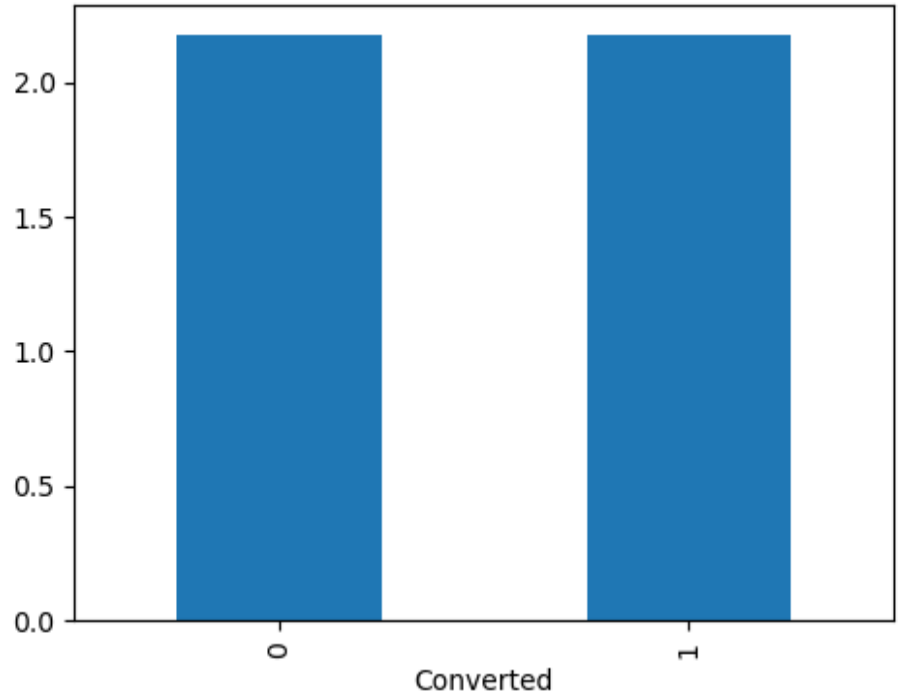
Graph 2 shows that leads who converted to customers spent a lot more time on the website than those who didnt though there are also cases where the conversion happened without the lead having spent any time on the website too

Graph 1: Page Views Per Visit distribution





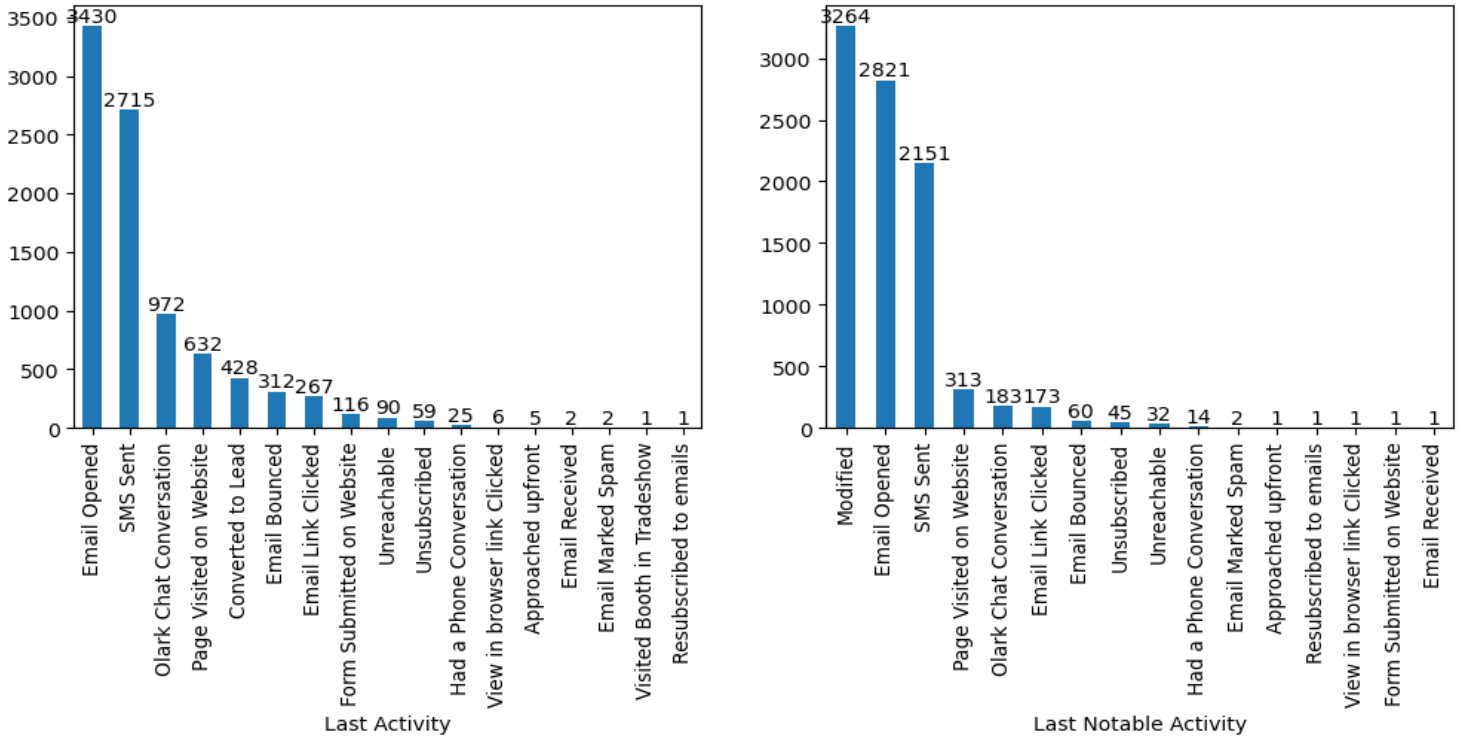
Graph 2: Page Views Per Visit groupby with Converted



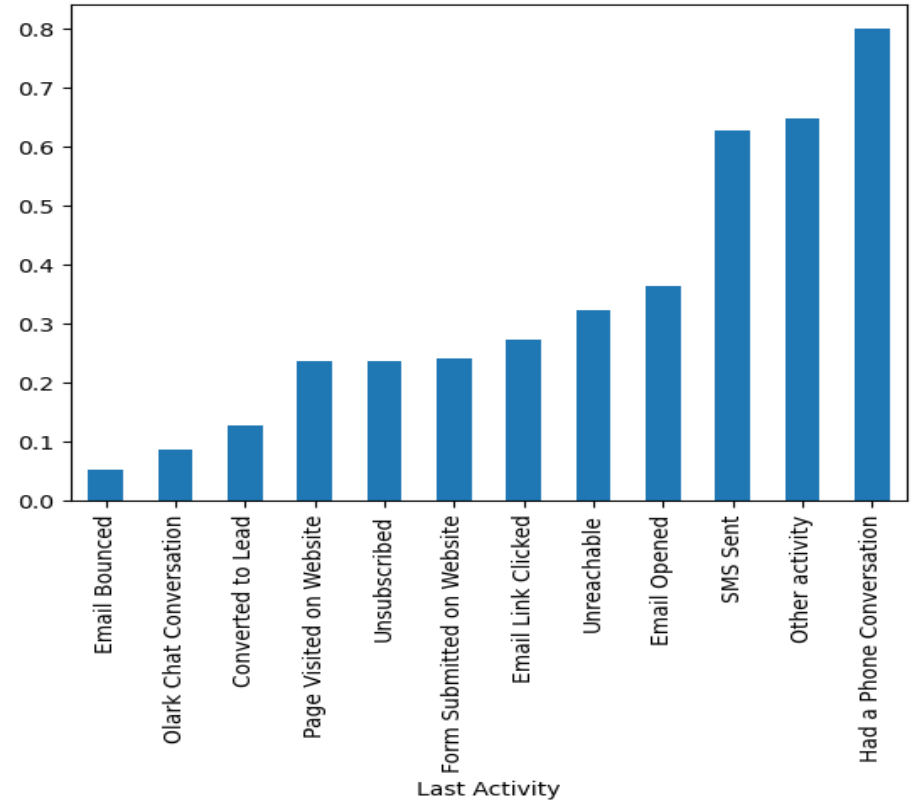
Graph 1 shows that most of the leads viewed between 0 to 3 pages per visit to the website though a minority viewed 6-7 pages too.

Graph 2 shows that though there is a slight variation in the range of pages viewed per visit between the converted and unconverted leads, on the whole there is no significant difference in the mean pages viewed per visit among two groups

Graph 1: Last Activity & Last Notable Activity distribution



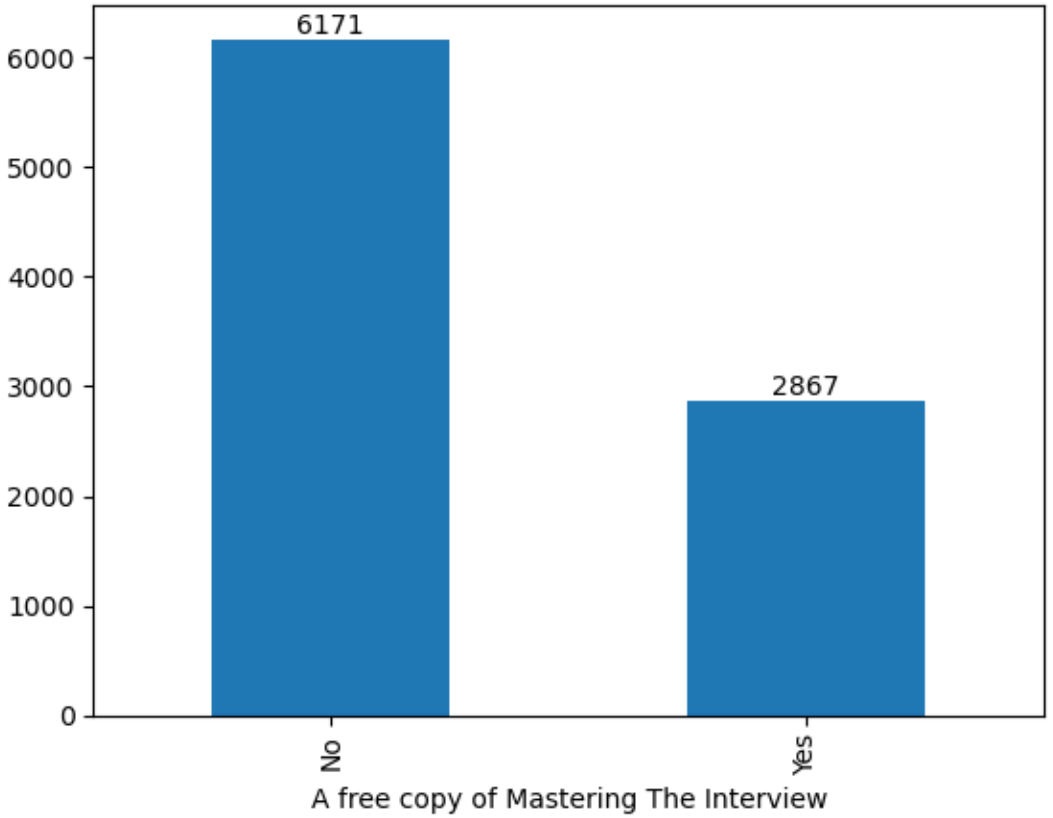
Graph 2: Last Activity groupby with Converted



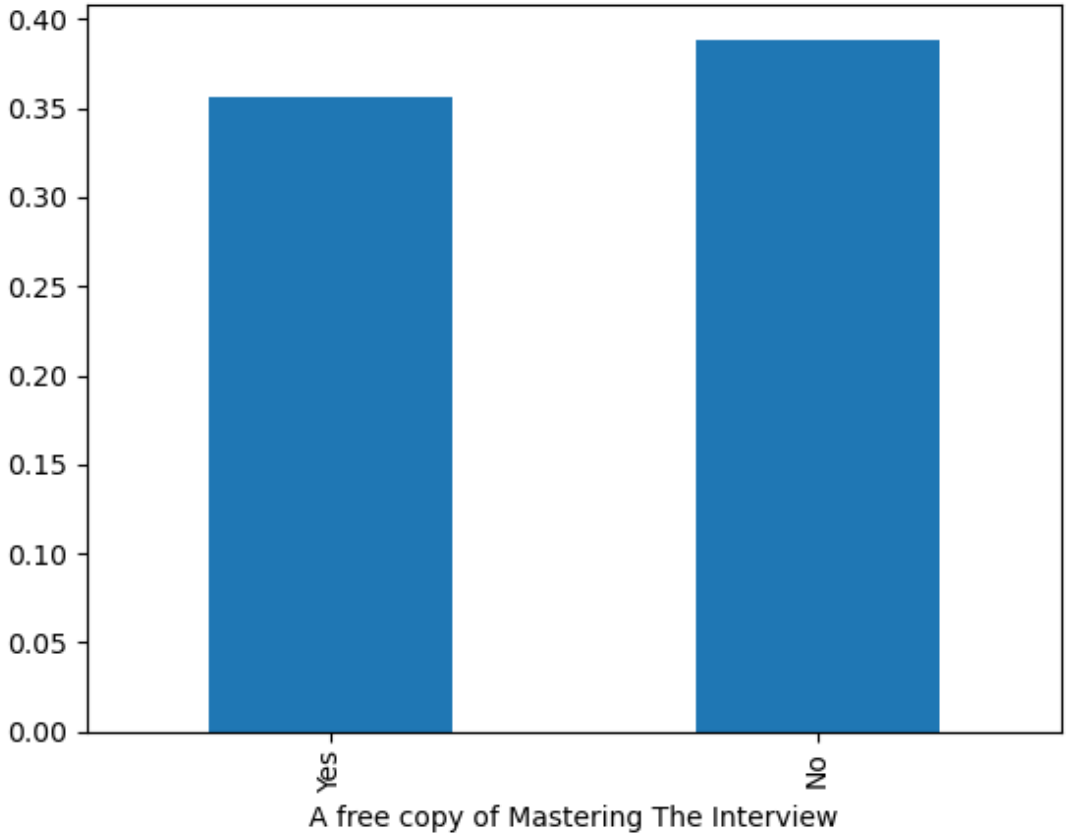
Graph 1 shows 'Last Activity' and 'Last Notable Activity' columns only vary by 3267 cells which only contain the value 'Modified'. So 'Last Notable Activity' column was dropped.

Graph 2 shows that maximum lead conversion was seen among leads whose last activity was 'Had a phone conversation'.

Graph 1: A free copy of Mastering The Interview distribution

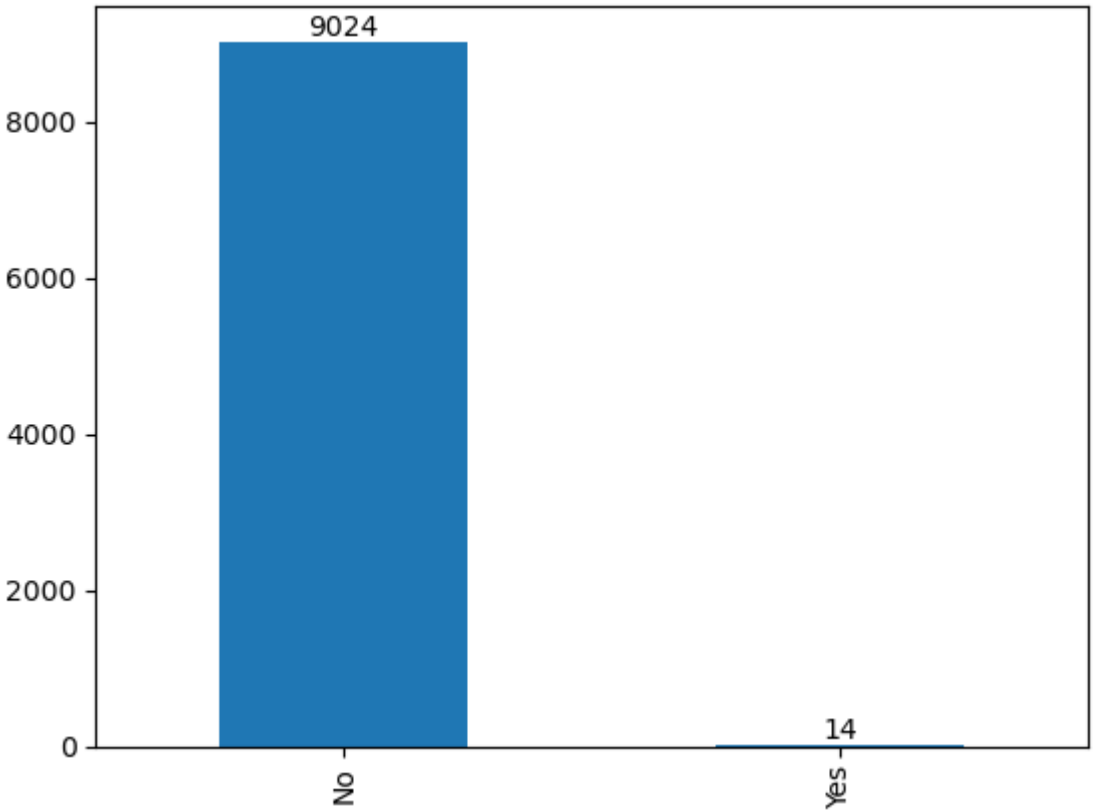


Graph 2: A free copy of Mastering The Interview groupby with Converted

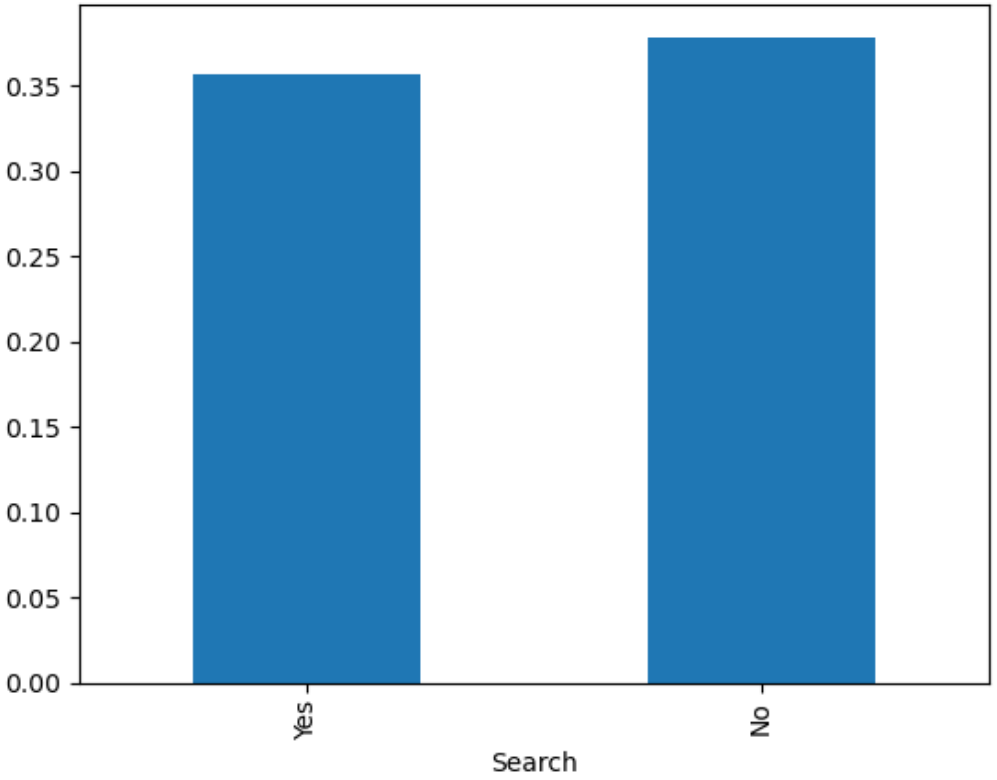


Graph 1 shows most leads chose the option ‘No’ for ‘A free copy of Mastering The Interview’.
Graph 2 shows that maximum lead conversion was seen among leads selected the option ‘No’.

Graph 1: Search distribution



Graph 2: Search groupby with Converted



Graph 1 shows that most leads chose the option ‘No’ for Search while only 14 leads chose ‘Yes’.
Graph 2 shows that maximum lead conversion was seen among leads who chose ‘No’. As more than 99% of leads chose No, this column was deleted

Other columns deleted due to limited variance

Columns having 99% same class

1. Newspaper Article
2. Newspaper
3. X Education Forums
4. Digital Advertisement
5. Through Recommendations

Column with almost 40% missing values

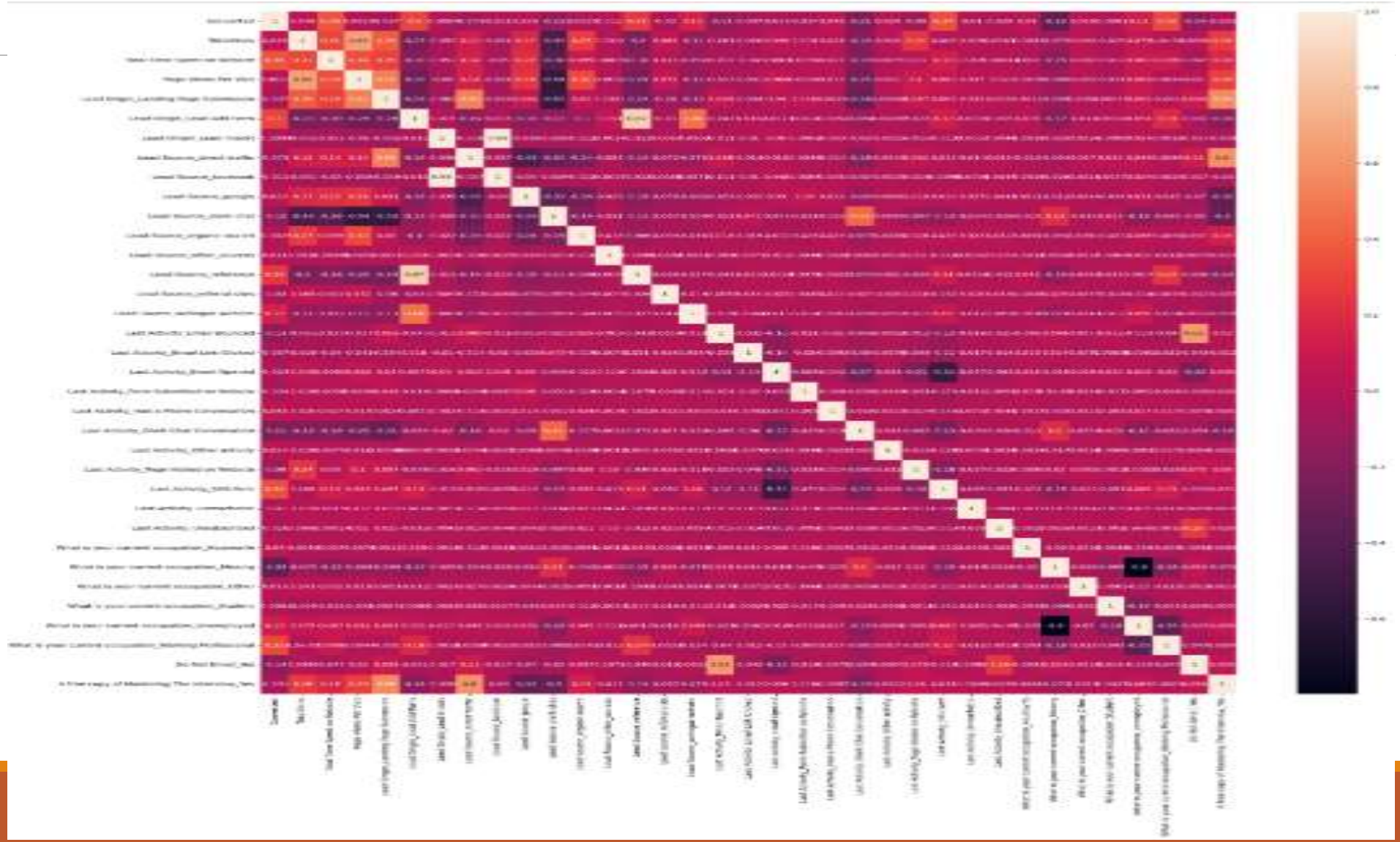
1. City

Data Preparation: Dummies

Created dummies for the categorical features left in the dataset and dropped the original variables

1. Lead Origin
2. Lead Source
3. Last Activity
4. What is your current occupation
5. Do Not Email
6. A free copy of Mastering The Interview'

Correlations Matrix & Heatmap



Model Building

1. Test-Train Split

X feature: All variables except Prospect ID and Converted (target variable)

y feature: Converted

2. Scaling

Applied MinMaxScaler to all numerical features in train data to bring them all in the same range

3. Running first model with Statsmodel library

4. RFE

Automated feature selection using RFE keeping the number of features to be selected at 15:

'TotalVisits', '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', 'Lead Source_welingak website', 'Last Activity_Had a Phone Conversation', 'Last Activity_Olark Chat Conversation', 'Last Activity_SMS Sent', 'What is your current occupation_Housewife', 'What is your current occupation_Missing', 'What is your current occupation_Working Professional', 'Do Not Email_Yes'

Building optimum model

Running successive iterations of model to check p-values and VIFs to remove features which are insignificant or causing collinearity.

	coef	std err	z	P>|z|	[0.025	0.975]
const	-0.8467	0.092	-9.224	0.000	-1.027	-0.667
TotalVisits	2.4860	0.482	5.157	0.000	1.541	3.431
Total Time Spent on Website	4.4655	0.167	26.807	0.000	4.139	4.792
Lead Origin_Lead Add Form	2.4365	0.238	10.236	0.000	1.970	2.903
Lead Source_direct traffic	-1.6876	0.124	-13.606	0.000	-1.931	-1.444
Lead Source_google	-1.3518	0.121	-11.214	0.000	-1.588	-1.116
Lead Source_organic search	-1.5810	0.147	-10.734	0.000	-1.870	-1.292
Lead Source_referral sites	-1.6236	0.331	-4.906	0.000	-2.272	-0.975
Lead Source_welingak website	2.4932	1.034	2.412	0.016	0.467	4.519
Last Activity_Had a Phone Conversation	2.1523	0.698	3.084	0.002	0.785	3.520
Last Activity_Olark Chat Conversation	-1.2571	0.163	-7.724	0.000	-1.576	-0.938
Last Activity_SMS Sent	1.2737	0.074	17.157	0.000	1.128	1.419
What is your current occupation_Housewife	23.0995	1.59e+04	0.001	0.999	-3.11e+04	3.12e+04
What is your current occupation_Missing	-1.1806	0.087	-13.597	0.000	-1.351	-1.010
What is your current occupation_Working Professional	2.5062	0.183	13.660	0.000	2.147	2.866
Do Not Email_Yes	-1.3334	0.166	-8.039	0.000	-1.658	-1.008

	Features	VIF
0	TotalVisits	2.800
4	Lead Source_google	2.430
1	Total Time Spent on Website	2.370
3	Lead Source_direct traffic	2.150
5	Lead Source_organic search	1.790
2	Lead Origin_Lead Add Form	1.500
10	Last Activity_SMS Sent	1.500
12	What is your current occupation_Missing	1.410
7	Lead Source_welingak website	1.330
13	What is your current occupation_Working Profes...	1.200
9	Last Activity_Olark Chat Conversation	1.150
14	Do Not Email_Yes	1.110
6	Lead Source_referral sites	1.080
8	Last Activity_Had a Phone Conversation	1.010
11	What is your current occupation_Housewife	1.000

Model Accuracy: 0.81

Removed ‘What is your current occupation_Housewife’ feature as its p-value is more than 0.05.

Building optimum model

Second iteration.

	coef	std err	z	P>|z|	[0.025	0.975]
const	-0.8417	0.092	-9.177	0.000	-1.022	-0.662
TotalVisits	2.4523	0.482	5.092	0.000	1.508	3.396
Total Time Spent on Website	4.4622	0.166	26.813	0.000	4.136	4.788
Lead Origin_Lead Add Form	2.4496	0.238	10.302	0.000	1.984	2.916
Lead Source_direct traffic	-1.6812	0.124	-13.571	0.000	-1.924	-1.438
Lead Source_google	-1.3439	0.120	-11.163	0.000	-1.580	-1.108
Lead Source_organic search	-1.5651	0.147	-10.648	0.000	-1.853	-1.277
Lead Source_referral sites	-1.6202	0.331	-4.897	0.000	-2.269	-0.972
Lead Source_welingak website	2.4789	1.034	2.398	0.016	0.453	4.505
Last Activity_Had a Phone Conversation	2.1438	0.698	3.072	0.002	0.776	3.511
Last Activity_Olark Chat Conversation	-1.2614	0.163	-7.753	0.000	-1.580	-0.943
Last Activity_SMS Sent	1.2661	0.074	17.073	0.000	1.121	1.411
What is your current occupation_Missing	-1.1849	0.087	-13.655	0.000	-1.355	-1.015
What is your current occupation_Working Professional	2.4996	0.183	13.628	0.000	2.140	2.859
Do Not Email_Yes	-1.3372	0.166	-8.065	0.000	-1.662	-1.012

	Features	VIF
0	TotalVisits	2.800
4	Lead Source_google	2.420
1	Total Time Spent on Website	2.370
3	Lead Source_direct traffic	2.150
5	Lead Source_organic search	1.790
2	Lead Origin_Lead Add Form	1.500
10	Last Activity_SMS Sent	1.500
11	What is your current occupation_Missing	1.410
7	Lead Source_welingak website	1.330
12	What is your current occupation_Working Profes...	1.200
9	Last Activity_Olark Chat Conversation	1.150
13	Do Not Email_Yes	1.110
6	Lead Source_referral sites	1.080
8	Last Activity_Had a Phone Conversation	1.010

Model Accuracy: 0.81

Now all variables have low p-value and VIFs less than 5.

Other Metrics with Confusion Matrix:
Cutoff at 0.5

Actual / Predicted	Unconverted	Converted
Unconverted	3453 (True Negative TN)	457 (False Positive FP)
Converted	747 (False Negative FN)	1669 (True Positive TP)

Sensitivity = TP/TP+FN = 0.69

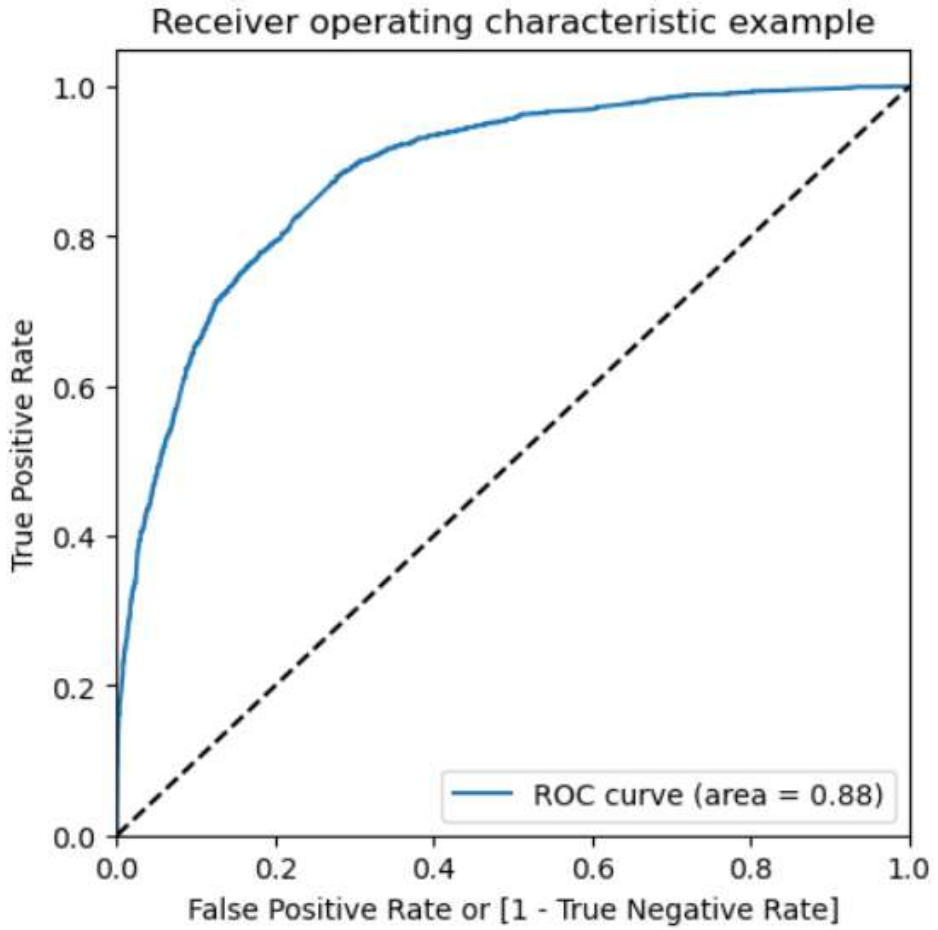
Specificity = TN/TN+FP = 0.88

False positive rate = FP/TN+FP = 0.15

Positive predictive value = TP/TP+FP = 0.79

Negative predictive value = TN/TN+FN = 0.82

Plotting the ROC Curve



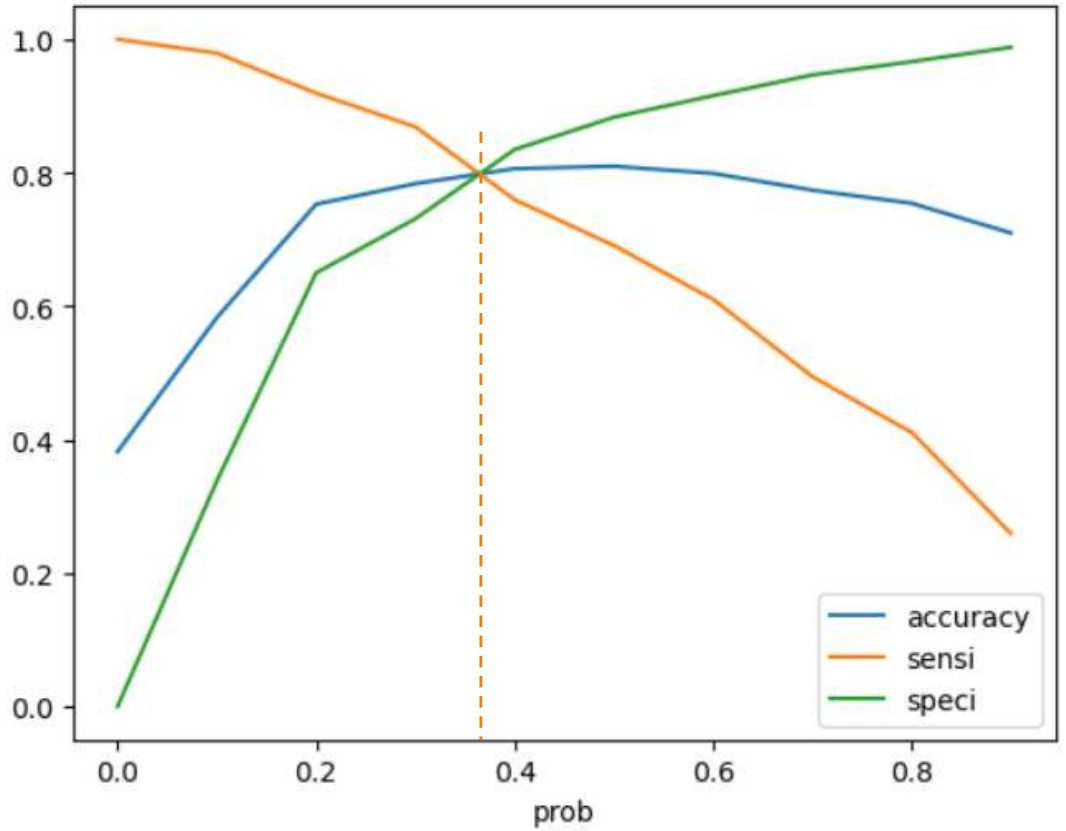
An ROC curve demonstrates several things:

- It shows the tradeoff 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.

ROC curve area: 0.88

Finding Optimal Cutoff Probability Point 1

Point where sensitivity and specificity are balanced



From the curve above, 0.35 is the optimum cutoff point to divide the probabilities of converted and unconverted leads.

Rechecking Metrics with Confusion Matrix: Cutoff at 0.35

Actual / Predicted	Unconverted	Converted
Unconverted	3132 (True Negative TN)	778 (False Positive FP)
Converted	501 (False Negative FN)	1915 (True Positive TP)

Accuracy = TP+TN/TP+FN+FP+FN = 0.80

Sensitivity = TP/TP+FN = 0.79

Specificity = TN/TN+FP = 0.80

False positive rate = FP/TN+FP = 0.2

Positive predictive value = TP/TP+FP = 0.71

Negative predictive value = TN/TN+FN = 0.86

Precision and Recall with Confusion Matrix: Cutoff at 0.5

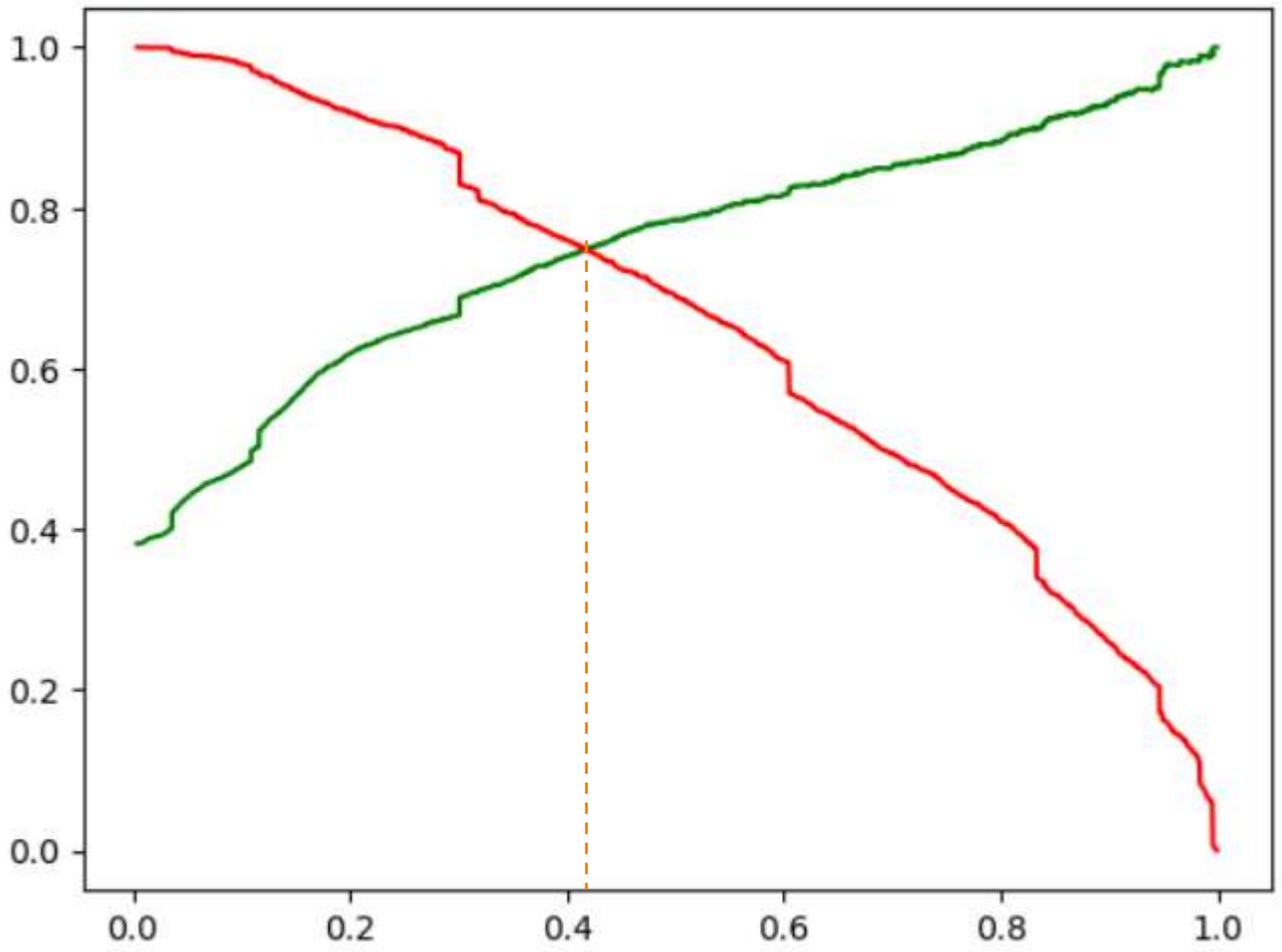
Actual / Predicted	Unconverted	Converted
Unconverted	3453 (True Negative TN)	457 (False Positive FP)
Converted	747 (False Negative FN)	1669 (True Positive TP)

Precision = TP/TP+FP = 0.79

Recall = TP/TP+FN = 0.69

Finding Optimal Cutoff Probability Point 2

Point where precision and recall are balanced



Optimal cutoff point: 0.41

Making predictions on the test set

- Calculating conversion probabilities for test data on basis of training model

	Prospect ID	Converted	Conversion_Prob
0	5150	0	0.313
1	3799	0	0.048
2	3588	0	0.014
3	6447	0	0.043
4	565	1	0.518

Checking Evaluation Metrics for Test Data

Using the probability cutoff of 0.35 based on the Accuracy-Sensitivity-Specificity tradeoff curve

Accuracy = TP+TN/TP+FN+FP+FN = 0.82

Sensitivity = TP/TP+FN = 0.80

Specificity = TN/TN+FP = 0.83

Precision = TP/TP+FP = 0.74

Recall = TP/TP+FN = 0.80

Using the probability cutoff of 0.41 based on the precision-recall tradeoff curve

Accuracy = TP+TN/TP+FN+FP+FN = 0.82

Sensitivity = TP/TP+FN = 0.76

Specificity = TN/TN+FP = 0.86

Precision = TP/TP+FP = 0.76

Recall = TP/TP+FN = 0.76

As there is very little difference in Training accuracy and Test Accuracy, the model is not overfitting or underfitting the data.

Identifying Hot Leads

Leads with Conversion Probability percentage > = 80

	Prospect ID	Converted	Conversion_Prob	final_predicted	Conversion Percentage Probabilities
5	4125	1	0.911	1	91.103
6	4941	1	0.973	1	97.338
8	6499	1	0.918	1	91.785
10	2166	1	0.861	1	86.062
12	392	1	0.966	1	96.639
...	...	...	...	...	...
2689	4005	1	0.972	1	97.158
2690	1618	0	0.995	1	99.529
2702	6127	1	0.833	1	83.311
2705	8330	1	0.922	1	92.192
2708	4081	1	0.954	1	95.435

476 rows × 5 columns