


The background features abstract green geometric shapes. On the left, a solid green trapezoid points upwards. On the right, a complex arrangement of overlapping translucent green triangles and polygons creates a layered, crystalline effect. The central text is positioned in the white space between these green elements.

Lead Scoring

Educational Lead Conversion


Problem Statement

X Education, an online education company, faces a challenge with low lead conversion rates despite acquiring many leads daily. Leads are obtained through website interactions and referrals, and only 30% of leads typically convert. The company seeks to enhance efficiency by identifying high-potential "Hot Leads" for better focus. A model is needed to assign lead scores, indicating conversion likelihood. The goal is to increase the lead conversion rate from 30% to the CEO's target of around 80%. This involves optimizing lead nurturing efforts to guide potential leads through the conversion process and improve overall sales funnel performance.



A logistic regression model has been developed to assign lead scores between 0 and 100, aiding in the identification of potential leads for a company's target conversion. Higher scores indicate greater likelihood of conversion, while lower scores suggest lower conversion probability. The model incorporates various factors influencing lead conversion and accommodates potential future adjustments to address evolving company requirements.

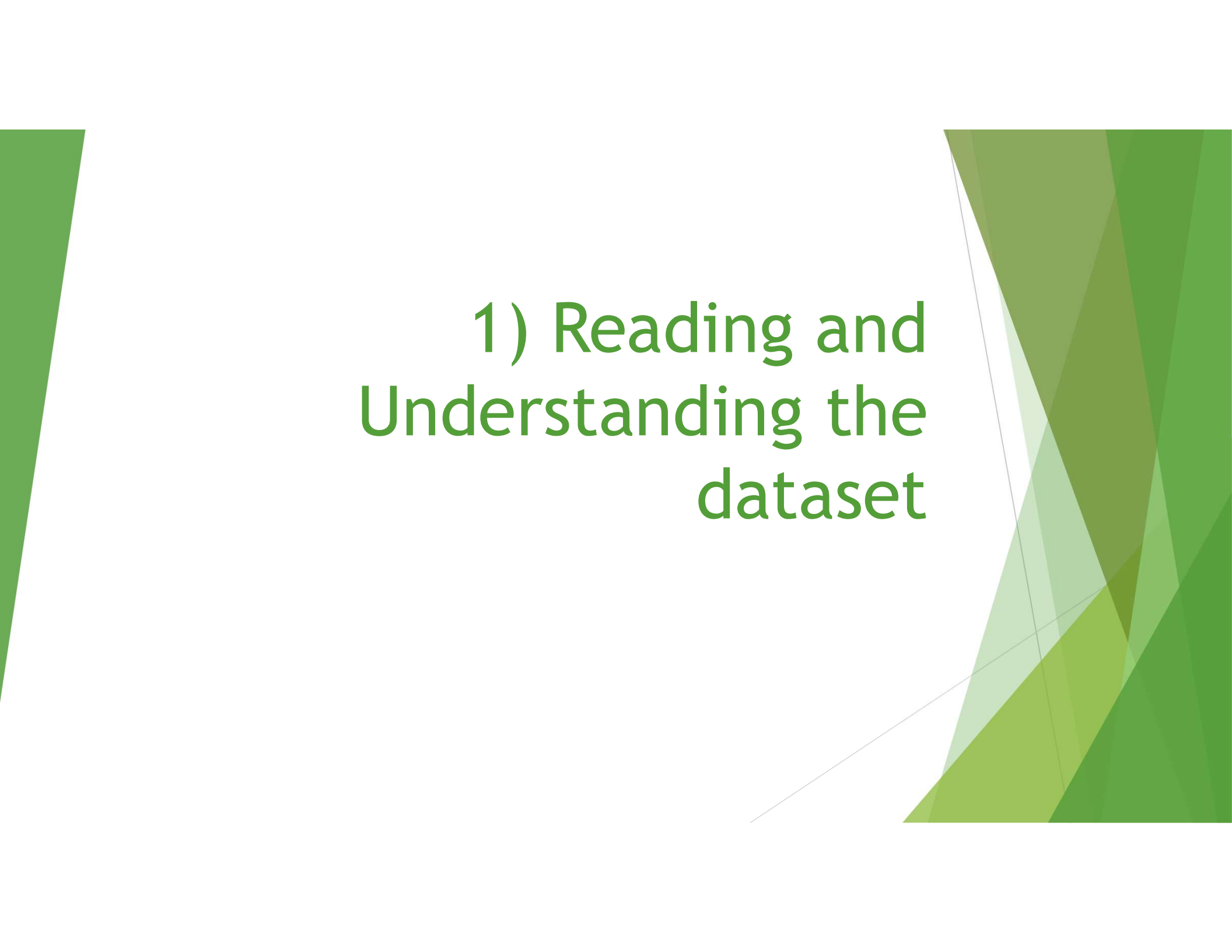
To further enhance the model's capabilities, a separate document has been prepared to address additional challenges presented by the company. This document outlines how the logistic regression model can be adapted to accommodate changing requirements and effectively tackle specific problems. The insights from this model, along with its ability to adapt, will be summarized and presented in the final PowerPoint presentation, offering recommendations to optimize lead conversion strategies.



Steps involved for Solving problem

- 1) Reading and Understanding the dataset
- 2) Exploratory Data Analysis
- 3) Pre-processing the data for Model Building
- 4) Building the Model
- 5) Evaluating the Model
- 6) Making predictions on the test set
- 7) Business aspects of the model



The background features abstract green geometric shapes. On the left, a solid green trapezoid points towards the center. On the right, a complex arrangement of overlapping translucent green triangles and polygons creates a layered, dynamic effect. The text is centered in a clean, green, sans-serif font.

1) Reading and Understanding the dataset

Information of dataset

► Shape:

- 9240 rows and 37 columns
- 30 Columns with data type as Object
- 4 Columns with data type as float64
- 3 Columns with data type as int64



► Target Variable - Converted (Categorical)

- Value 1 means lead is converted successfully
- Value 0 means lead is not converted successfully
- There are around 38% values for which value of converted is 1 and 62% values for which value of converted is 0.
- So, the data is nicely balanced for performing analysis

The background features abstract green geometric shapes. On the left, a solid green trapezoid points upwards. On the right, a complex arrangement of overlapping, semi-transparent green triangles and polygons creates a layered, dynamic effect. The central text is positioned between these two main graphic elements.

2) Exploratory Data Analysis

2) Exploratory Data Analysis - Steps

- 1) Dropping Columns and Handling Outliers for Categorical Columns
- 2) Analysing Numerical Columns and Checking Outliers
- 3) Univariate Analysis
- 4) Bivariate Analysis
- 5) Multivariate Analysis



1) Dropping Columns - Null Values

1	Prospect ID	0.000000
2	Lead Number	0.000000
3	Lead Origin	0.000000
4	Lead Source	0.389610
5	Do Not Email	0.000000
6	Do Not Call	0.000000
7	Converted	0.000000
8	TotalVisits	1.482684
9	Total Time Spent on Website	0.000000
10	Page Views Per Visit	1.482684
11	Last Activity	1.114719
12	Country	26.634199
13	Specialization	15.562771
14	How did you hear about X Education	23.885281
15	What is your current occupation	29.112554
16	What matters most to you in choosing a course	29.318182
17	Search	0.000000
18	Magazine	0.000000
19	Newspaper Article	0.000000

20	X Education Forums	0.000000
21	Newspaper	0.000000
22	Digital Advertisement	0.000000
23	Through Recommendations	0.000000
24	Receive More Updates About Our Courses	0.000000
25	Tags	36.287879
26	Lead Quality	51.590909
27	Update me on Supply Chain Content	0.000000
28	Get updates on DM Content	0.000000
29	Lead Profile	29.318182
30	City	15.367965
31	Asymmetrique Activity Index	45.649351
32	Asymmetrique Profile Index	45.649351
33	Asymmetrique Activity Score	45.649351
34	Asymmetrique Profile Score	45.649351
35	I agree to pay the amount through cheque	0.000000
36	A free copy of Mastering The Interview	0.000000
37	Last Notable Activity	0.000000

1) Dropping Columns - Null Values

- ▶ There are total 6 columns from total of 37 columns for which missing value percentage is more than 35%. All those columns are dropped.
- ▶ Dropped 6 columns are:
 - 1) Asymmetrique Profile Score
 - 2) Asymmetrique Activity Index
 - 3) Asymmetrique Profile Index
 - 4) Asymmetrique Activity Score
 - 5) Tags
 - 6) Lead Quality

Country Column - Dropped

- ▶ There are around 26% null values in this column and of the remaining 74% values, around 96% country value is India. So, there is high imbalance in the data.
- ▶ And column named City is present in the data which gives idea about the geographical location.
- ▶ So, the column country is removed from the data as it doesn't provide any significant information.

Prospect ID and Lead Number - Dropped

- ▶ Both columns Prospect ID and Lead number are like unique Identifier for each row. Thus, it has different value for every record in the data.
- ▶ There are not any feature which can change the value of target variable. So, both columns are dropped from the analysis.

Do Not Call Column - Dropped

- ▶ There are only 2 values for which value of “Do Not Call” is Yes and all the other values are No.
- ▶ So, this column has very high imbalance and as there are only 2 Yes values compared to huge number of records. It doesn't provide any significant insights for the analysis.
- ▶ So, the column “Do Not Call” is dropped from the analysis.

How did you head about X Education - Dropped

- ▶ There are around 4800+ values as “Select” for this column, which means user hasn’t provide any information for this.
- ▶ When all the null values are replaced with “Select” for this column and then total number of “Select” value is counted, it is turned out that there are around 78% Select values in the column.
- ▶ 78% null value in this column is very high percentage. So, this column is dropped from analysis.

What matters most to you in choosing a course - Dropped

- ▶ There are around 6300+ points for which value is “Better Career Prospects” and there are only 3 types of values. Other value like ‘Flexibility & Convenience’ and ‘Other’ have occurred only 1 times each. And other records has null values.
- ▶ So, there is high imbalance in this column and it doesn’t hold much importance for the analysis.
- ▶ So, this column is dropped.

Search, Magazine, Newspaper Article, X Education Forums, Newspaper, Digital Advertisement - Dropped

- ▶ There are mostly 'No' values in all these sources and very few Yes values with maximum as 14 for Search column, which is very low compared to the original size of the dataset.
- ▶ So, all these columns can be dropped as combined Yes from all these columns is very less compared to size of the data.



Through Recommendations - Dropped

- ▶ There are only 7 'Yes' values compared to more than 9000+ 'No' values for this column.
- ▶ As there is high imbalance in this column and it doesn't provide much important information, it is dropped from the analysis.

Receive More Updates About Out Courses

- Dropped

- ▶ There are all 'No' values in this column and there is 0 'Yes' value for this column.
- ▶ This column doesn't provide any useful information for the analysis and hence it is dropped from the analysis.

Update me on Supply Chain Content and Get updated on DM Content - Dropped

- ▶ There are all 'No' values in both these columns and there is 0 'Yes' value in both these columns.
- ▶ These columns don't provide any useful information for the analysis and hence they are dropped from the analysis.

I agree to pay through the amount through cheque - Dropped

- ▶ There are all 'No' values in this column and there is 0 'Yes' value for this column.
- ▶ This column doesn't provide any useful information for the analysis and hence it is dropped from the analysis.

Handling Outliers for Categorical Columns

- ▶ Do Not Email column
 - ▶ For this column, there are 8400+ 'No' values and around 700+ 'Yes' values.
 - ▶ There is imbalance in Yes and No values but more than 700+ Yes values may provide significant information for analysis of the data.
 - ▶ So, there are no outliers in this column

Lead Source

- There are very few number of observations (1 or 2) compared to original datasize for many categories in Lead Source which will lead to more number of variables for analysis, when dummy variables will be created from it. So, records with those categories are dropped whose count is less than 10.

Value	Counts
Google	2873
Direct Traffic	2543
Olark Chat	1755
Organic Search	1154
Reference	534
Welingak Website	142
Referral Sites	125
Facebook	55

Last Activity

- There are very few number of observations (<10) compared to original datasize for many categories in Last Activity which will lead to more number of variables for analysis, when dummy variables will be created from it. So, records with those categories are dropped whose count is less than 10.

Value	Count
Email Opened	3422
SMS Sent	2720
Olark Chat Conversation	972
Page Visited on Website	634
Converted to Lead	426
Email Bounced	321
Email Link Clicked	267
Form Submitted on Website	116
Unreachable	93
Unsubscribed	58
Had a Phone Conversation	30

Specialization

- ▶ There are 1942 datapoints which has value 'Select' which is equivalent to null. So, all null values are replaced with 'Select'.
- ▶ There are around 36% null values but this column is kept for the bivariate analysis and the decision on keeping it or not is taken during that because other type of value maybe significant for the analysis.

Lead Profile

- ▶ There are around 4000+ 'Select' values in this column which is equivalent to null value. Value 'Select' in this column doesn't provide any useful information.
- ▶ But if user has any other value than 'Select' it maybe proved significant for conversion to the lead. So, this column is further investigated while doing bivariate analysis.



2) Analysing Numerical columns and checking Outliers

► Total Time Spent on Website

- There seems gradual increase in this column after 75 percentile, so there is no outlier in this column.

Parameter	Value
Count	9019.000000
Mean	482.692981
Std	545.425814
Min	0.000000
25%	10.000000
50%	246.000000
75%	921.000000
Max	2272.000000

Percentile	Value
90	1375
95	1557.1
99	1838.64

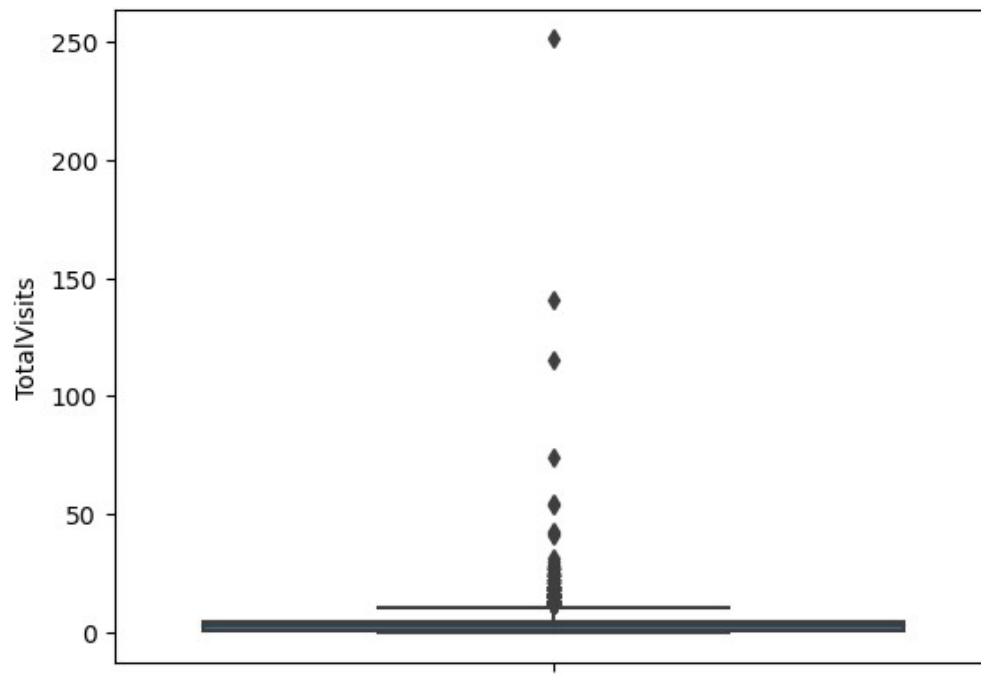
TotalVisits Column

- Maximum number of visit is 251 and 75% quantile value is 5 only. So, there is need to check if progress is quite sharp or gradual by getting different quantiles and using boxplots.

Parameter	Value
Count	9019.000000
Mean	3.452320
Std	4.861811
Min	0.000000
25%	1.000000
50%	3.000000
75%	5.000000
Max	251.000000

Quantile	Value
75	5
85	6
90	7
95	10
99	17
99.5	20.86
99.9	31.944

- There are only 10 observations for which value is more than 30. So, those records are considered as an outliers and can change the predictions significantly in the model. So, those records are dropped.



Page Views Per Visit

- ▶ Maximum number of Pages is 24 and 75% quantile value is 3 only. So, there is need to check if progress is quite sharp or gradual by getting different quantiles and using boxplots.

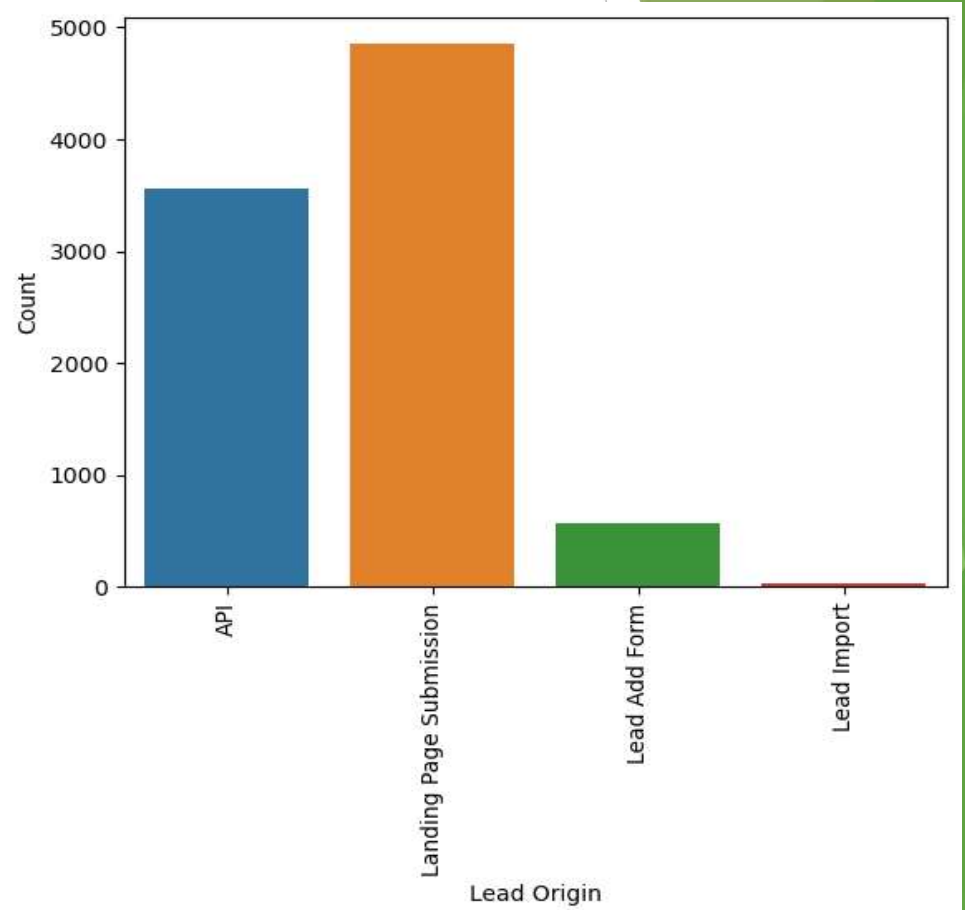
Parameter	Value
Count	9019.000000
Mean	2.359522
Std	2.083752
Min	0.000000
25%	1.000000
50%	2.000000
75%	3.000000
Max	24.000000

Quantile	Value
90	5
95	6
99	9
99.5	11
99.9	14
99.95	15

Univariate Analysis - Lead Origin

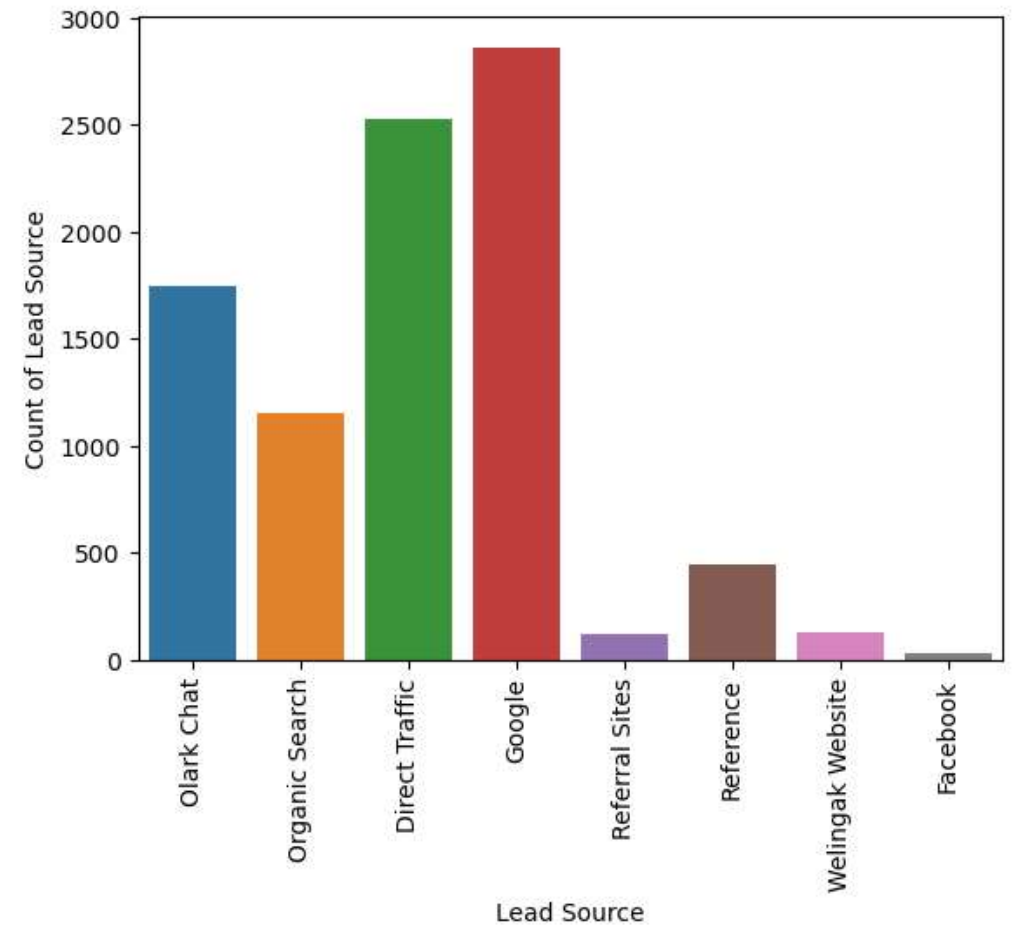
- ▶ For more number of persons, origin of lead is through Landing page submission.
- ▶ Different values and their counts for Lead Origin row is as below:

Value	Count
Landing Page Submission	4886
API	3580
Lead Add Form	718
Lead Import	55



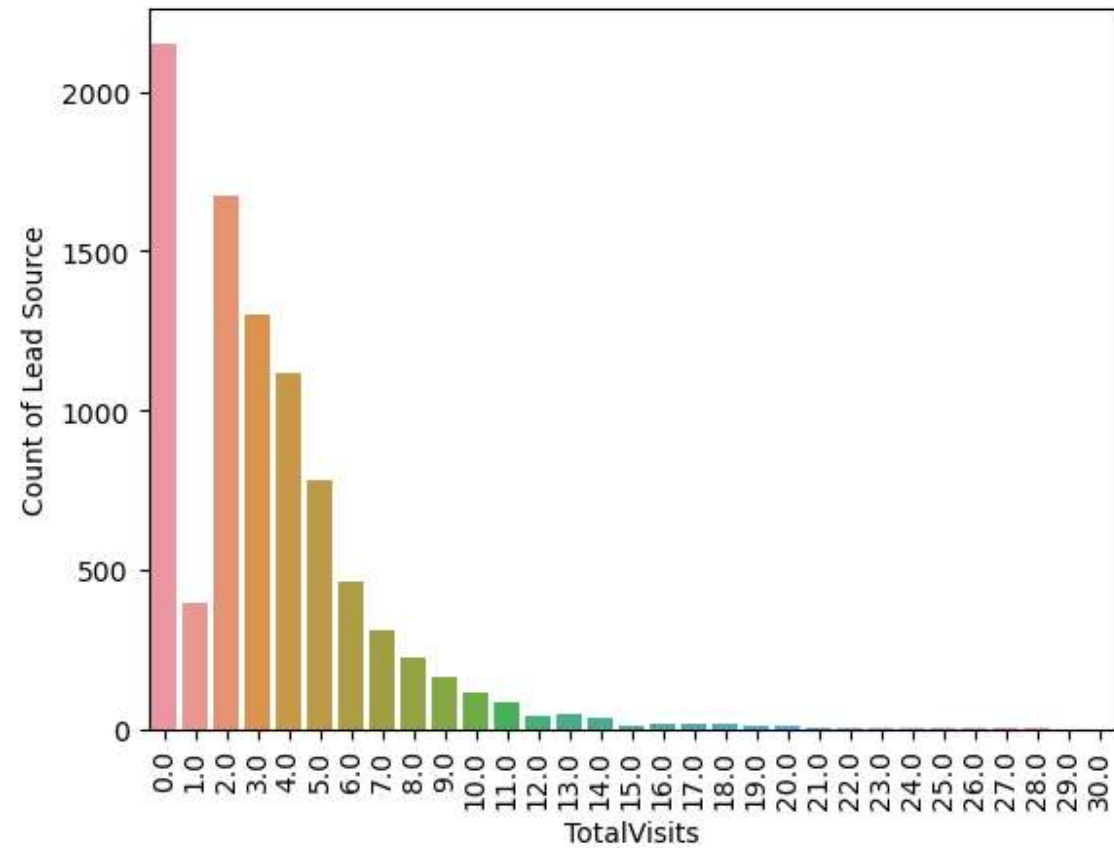
Lead Source

- ▶ Google is the Lead Source for more number of users.



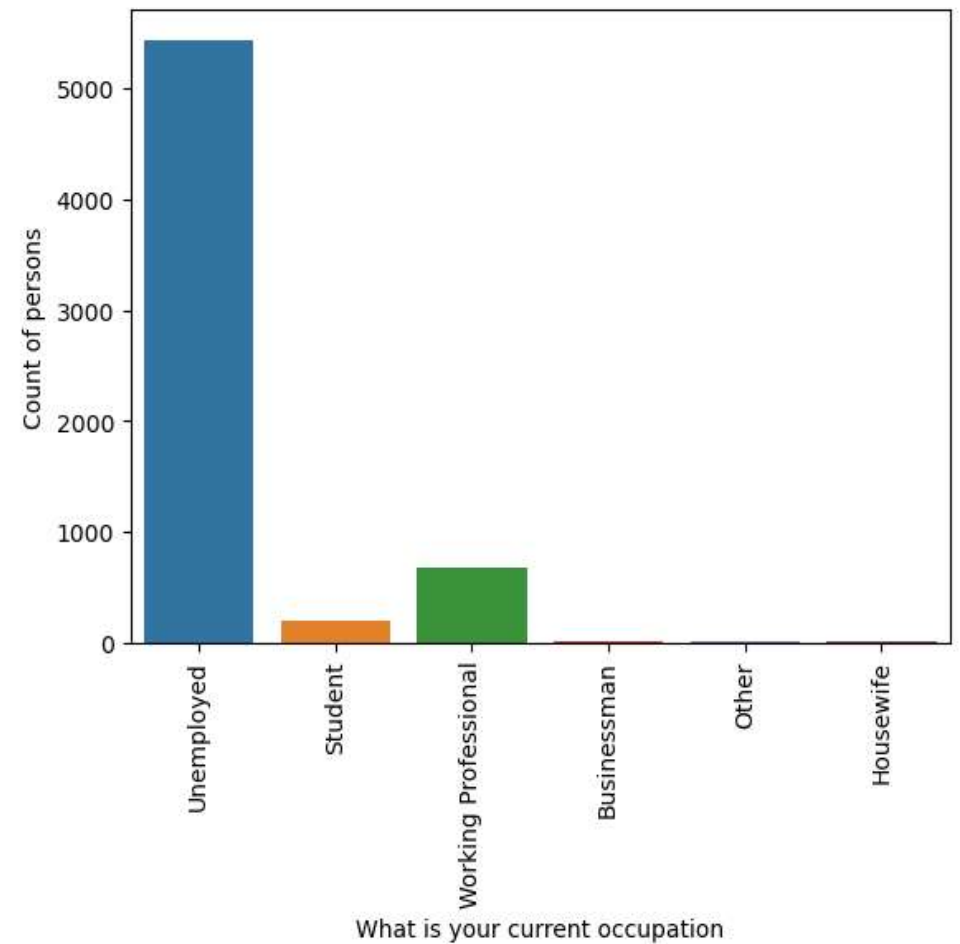
TotalVisits

- For most of the users, number of total visits is between 0 to 9.



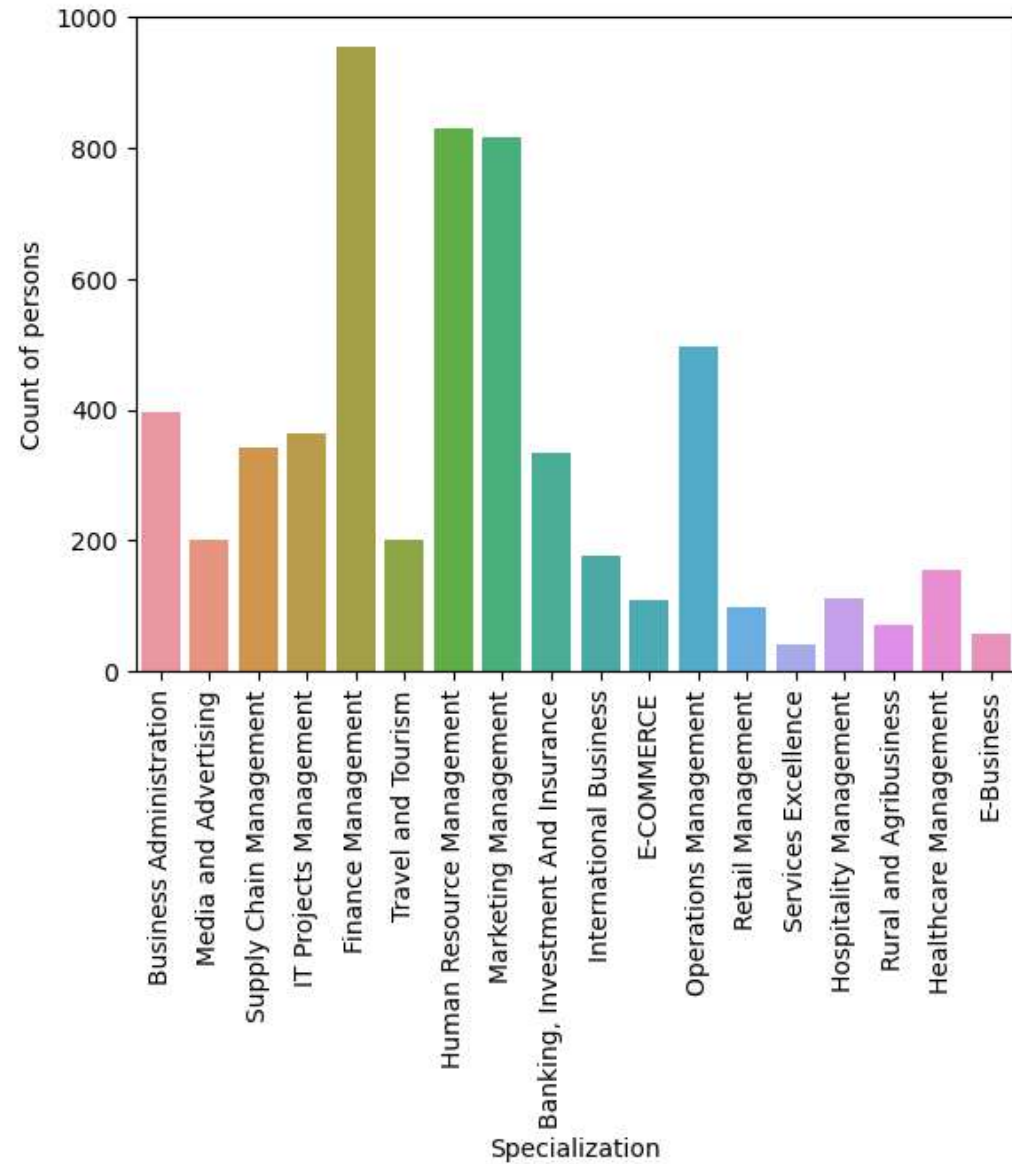
What is your current occupation

- More number of people are unemployed and there are very few businessman and housewife in the given dataset.



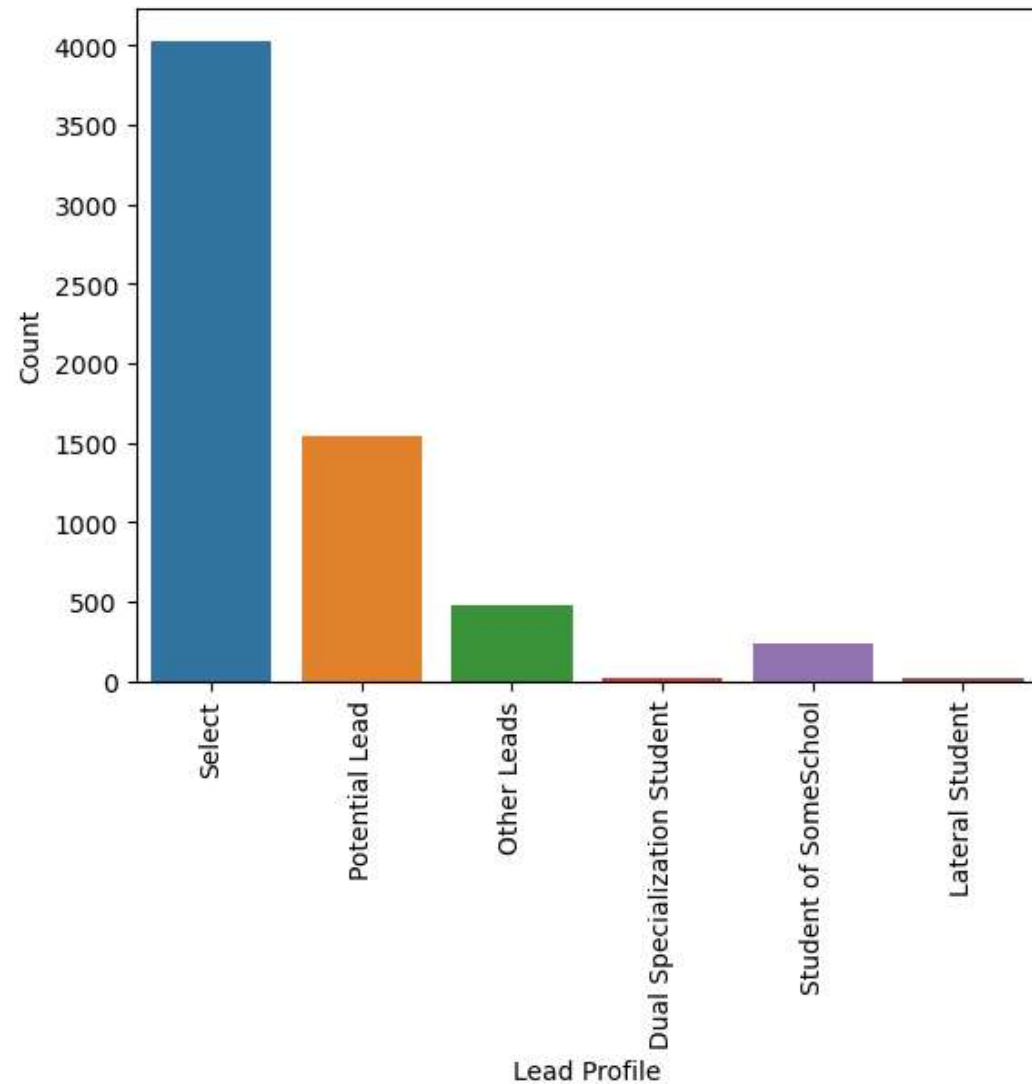
Specialization

- More number of people are from Finance or Human Resource or Marketing management.



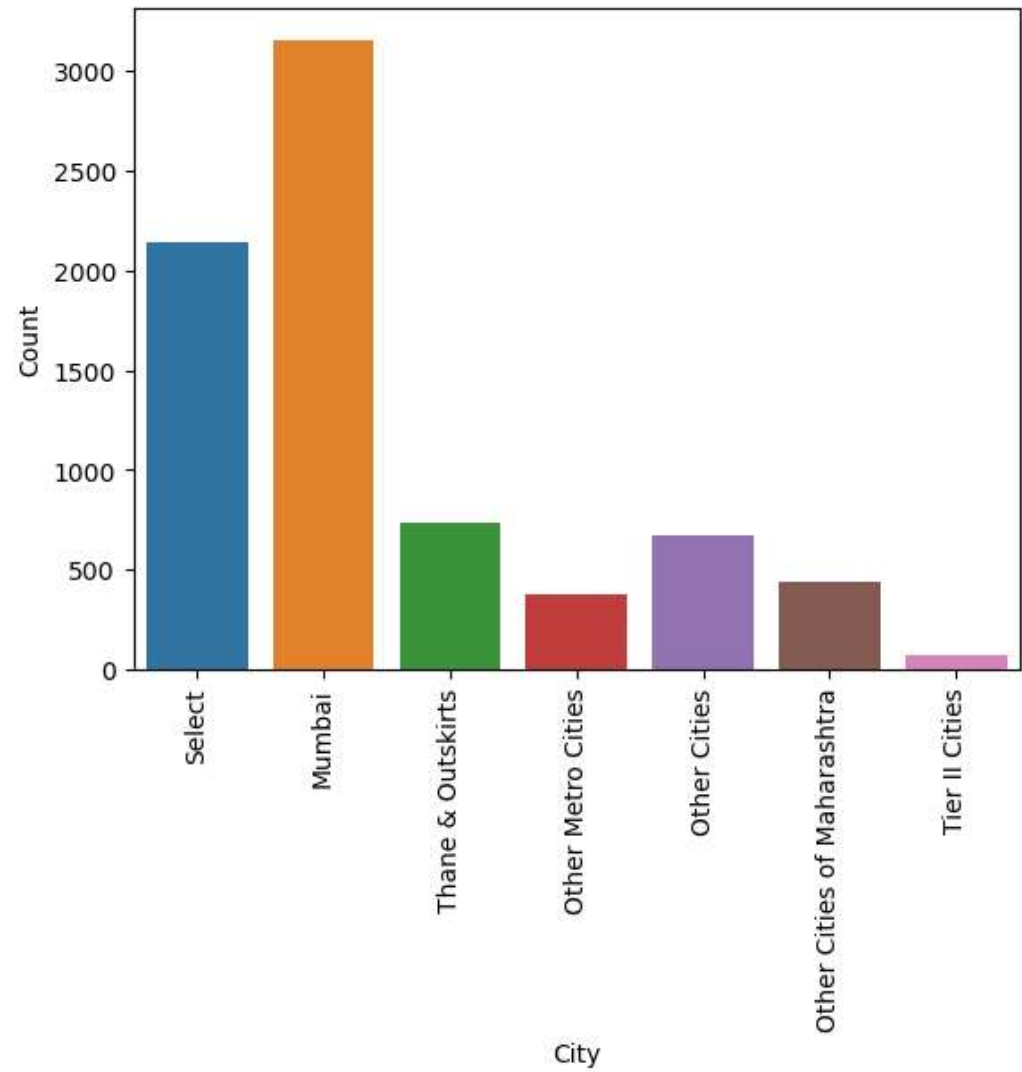
Lead Profile

- There are many users for whom Lead Profile is not determined. There are very few lateral students and students with dual specialization.



City

- More number of users belongs to Mumbai city. And there are least number of users for Tier II cities.



A free copy of Mastering The Interview

- This column has good balance of 'Yes' and 'No' values. So, this column could be very significant for the analysis.

Value	Count
No	6180
Yes	2879

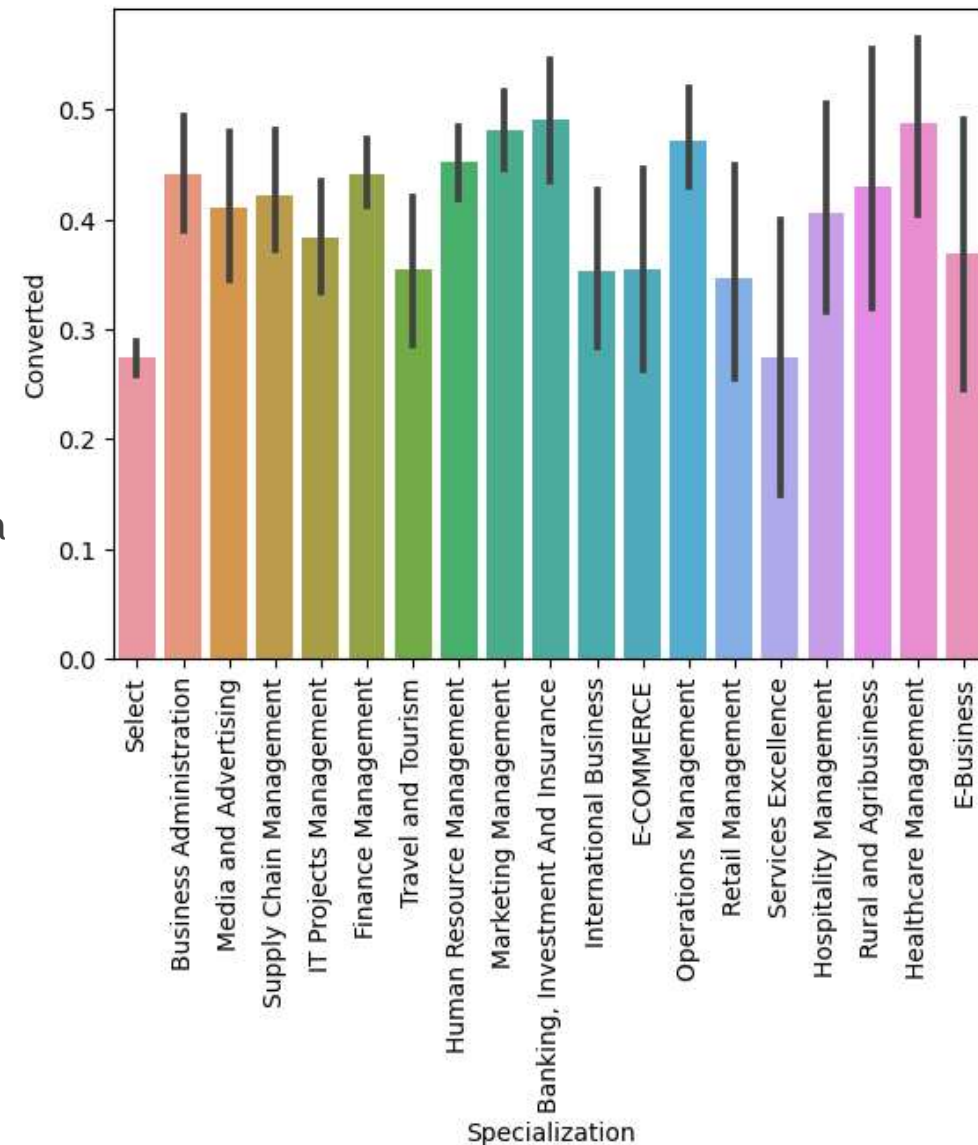
Bivariate Analysis

- ▶ It is divided into 2 parts:
 - 1) Categorical - Target (Categorical) variable
 - 2) Numerical - Target (Categorical) variable



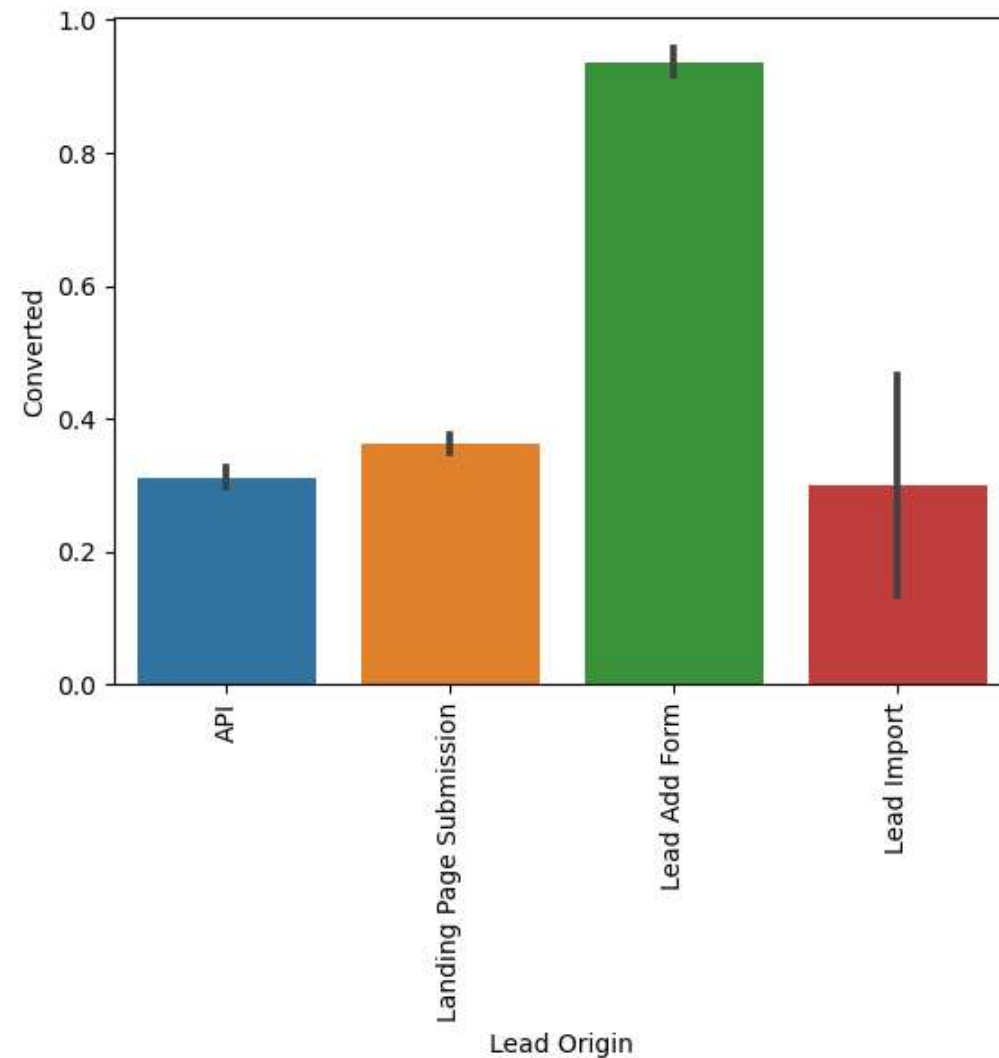
Specialization v/s Converted

- ▶ After replacing the null value with 'Select', it is clear that person who hasn't specified his/her Specialization has low chance of conversion. So, it may not be the case of missing value.
- ▶ So, despite having around 36% missing value, this column is not deleted.



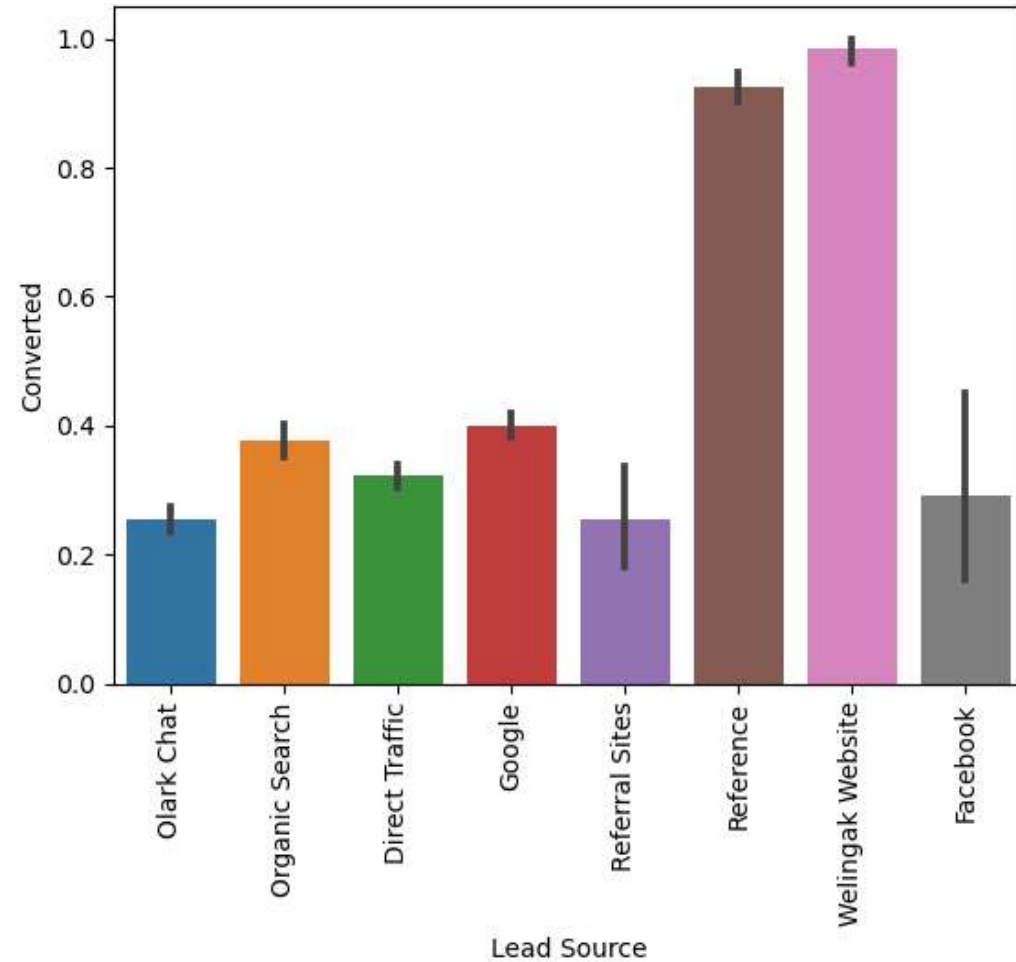
Lead Origin v/s Converted

- ▶ When Lead origin is “Lead Add Form” there is very high chance of conversion, which is close to 93%. So, this variable maybe given more importance while building the model.



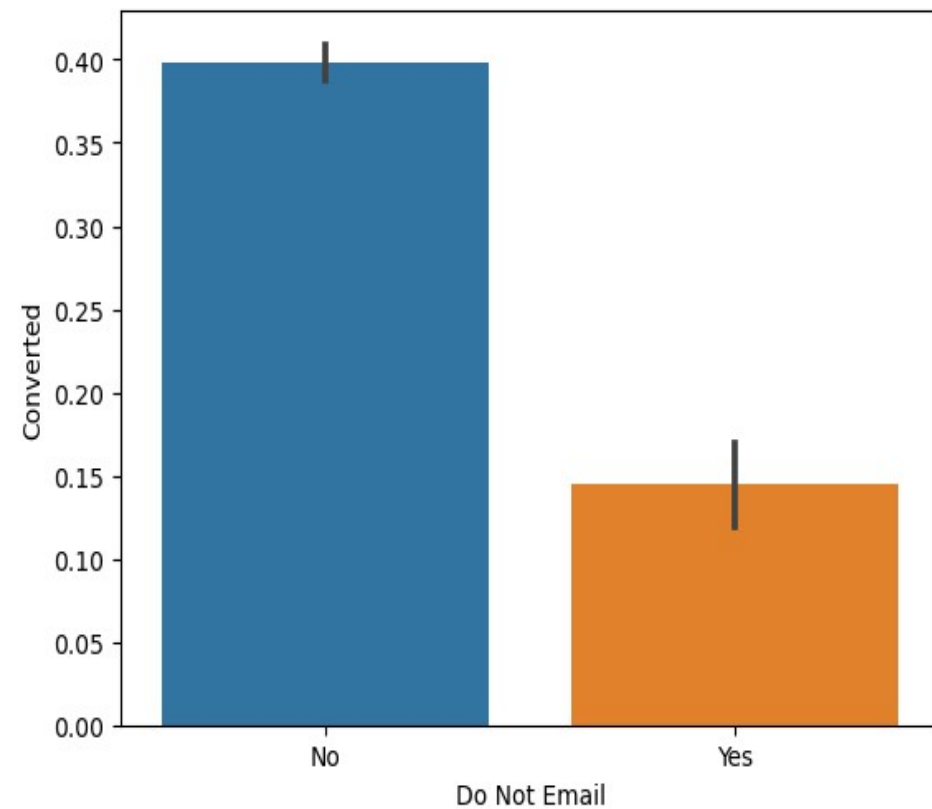
Lead Source v/s Converted

- There is very high percentage of conversion rate when source of lead is either through reference or through Welingak website.



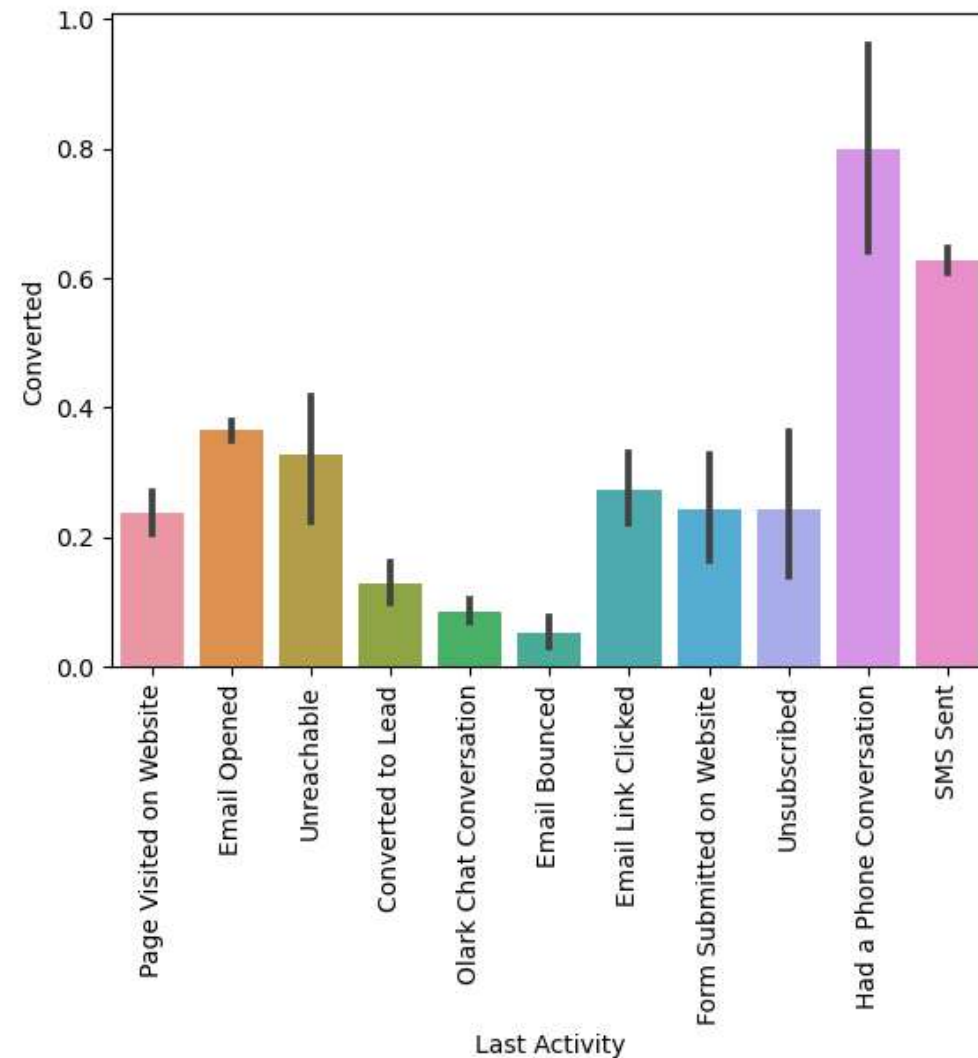
Do Not Email v/s Converted

- ▶ Person who doesn't opt for do not email has higher conversion rate compared to person who choose not to receive email which seems logical.



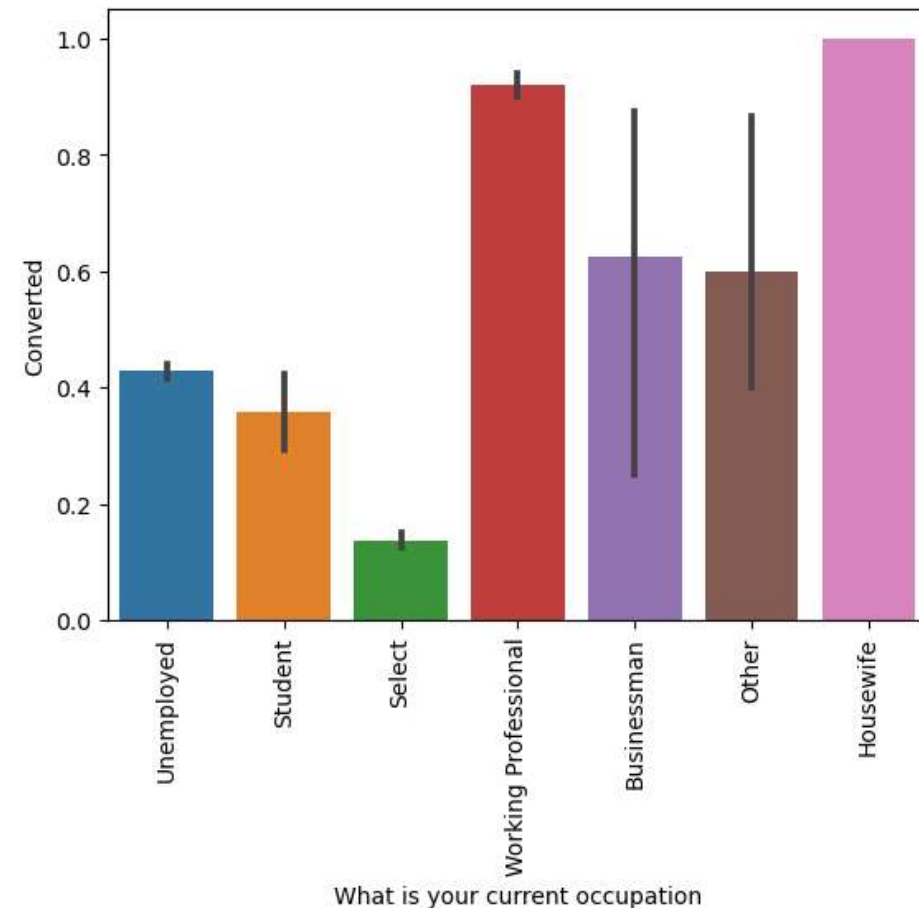
Last Activity v/s Converted

- ▶ When user had a phone conversation as last activity, then there is very high chance that lead is converted and on the other hand, when email is bounced, then there is very low chance of conversion.



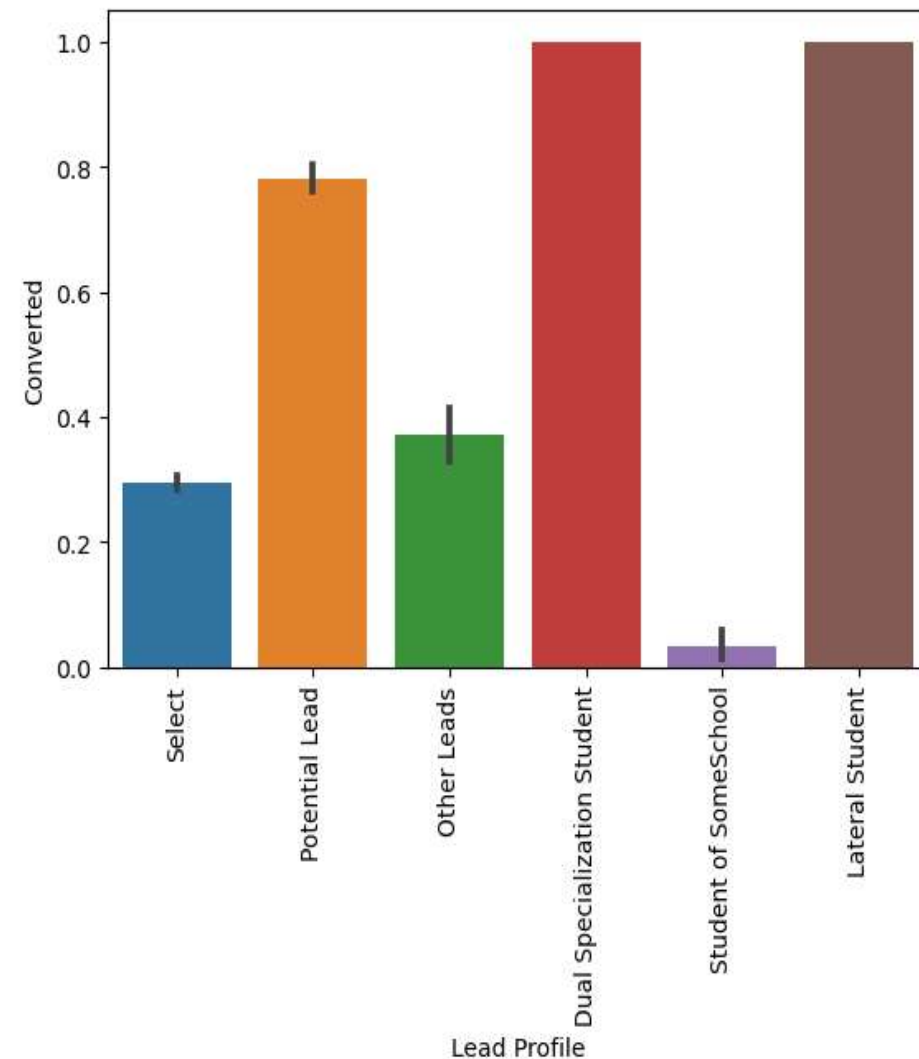
What is your current occupation v/s Converted

- There are around 29% missing values, which is replaced with 'Select' to check conversion rate for missing values.
- It is clear from the mean values, that when detail of current occupation is missing, there is very low chance of conversion to lead. For housewife and working professional, conversion chance is very high.



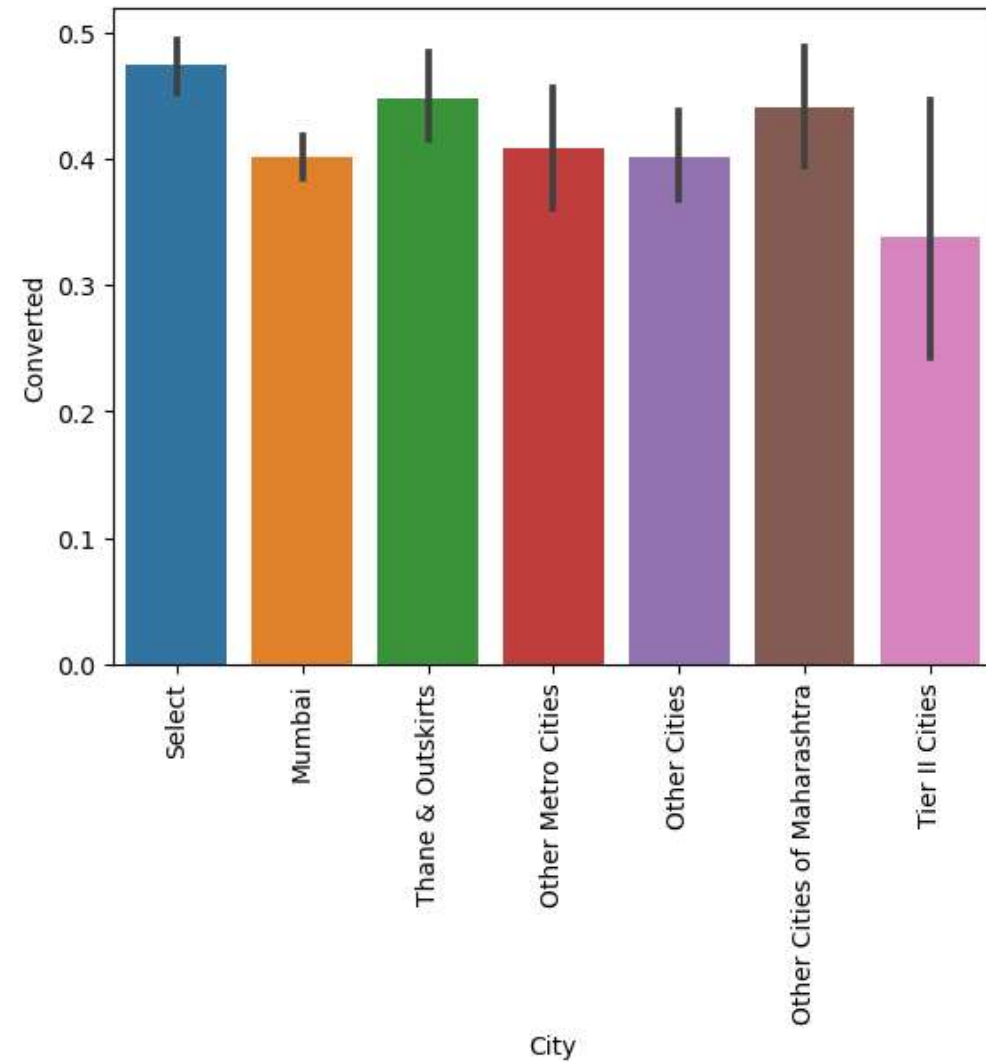
Lead Profile v/s Converted

- ▶ There are around 74% missing values, when all null values are replaced with 'Select'.
- ▶ But other 26% value can provide more useful information because for dual specialization and lateral student, conversion rate is 100%. And for student of some school. Conversion rate is very low.



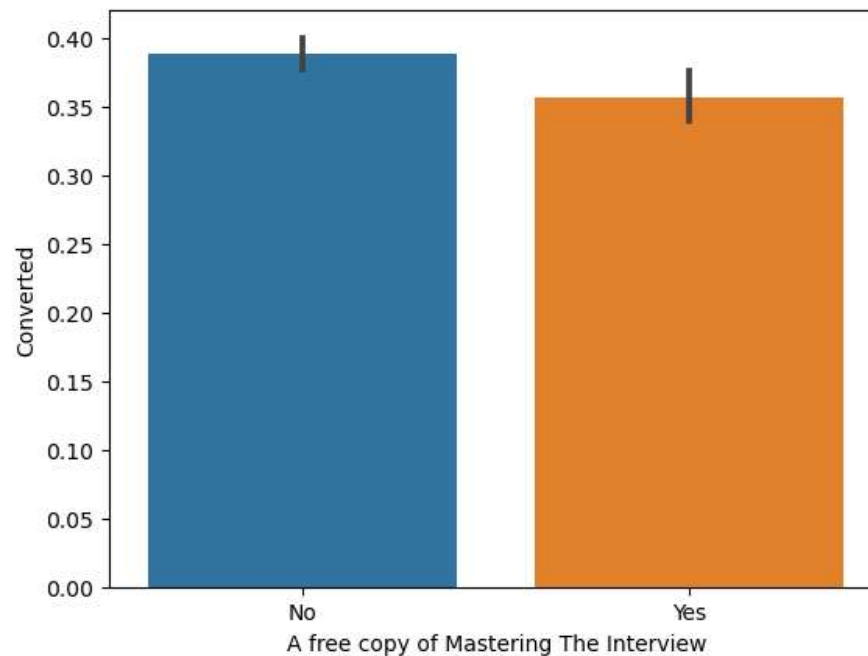
City v/s Converted

- It seems that value of city doesn't make that much difference in terms of conversion rate, but it is still kept in the dataset.



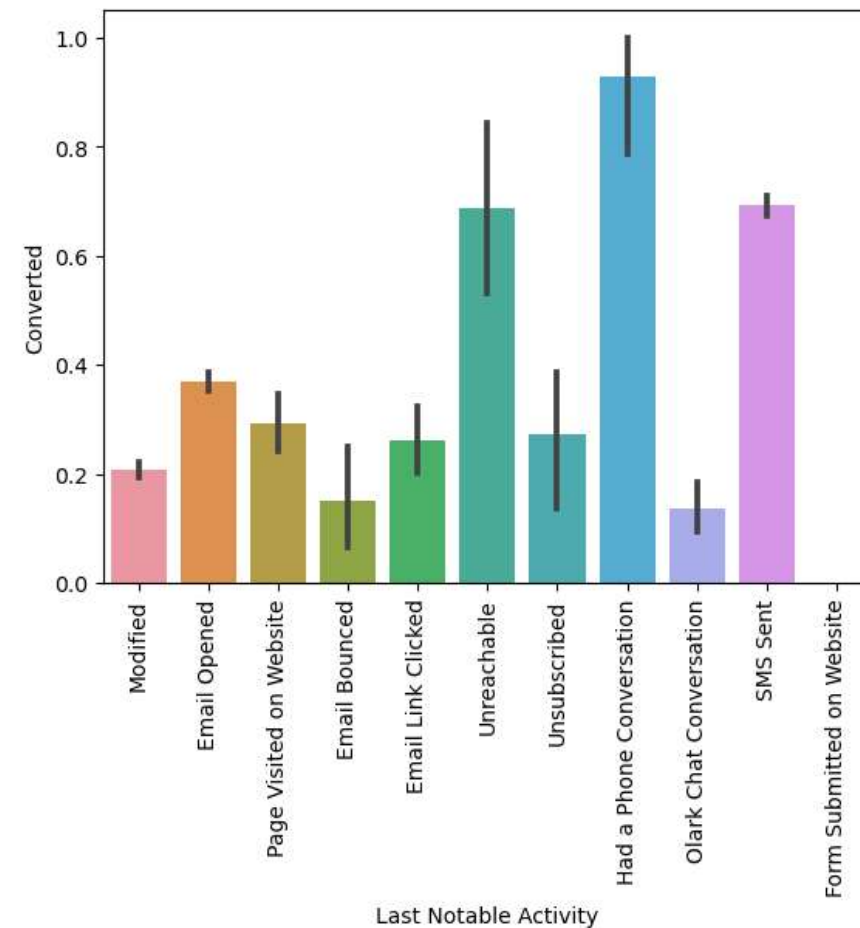
A free copy of mastering The Interview v/s Converted

- There is not much difference in conversion rate based on whether free copy of mastering interview value is 'Yes' or 'No'



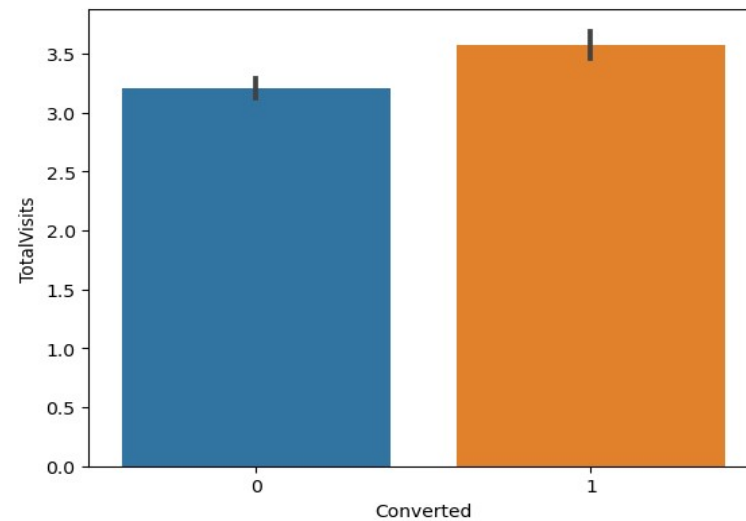
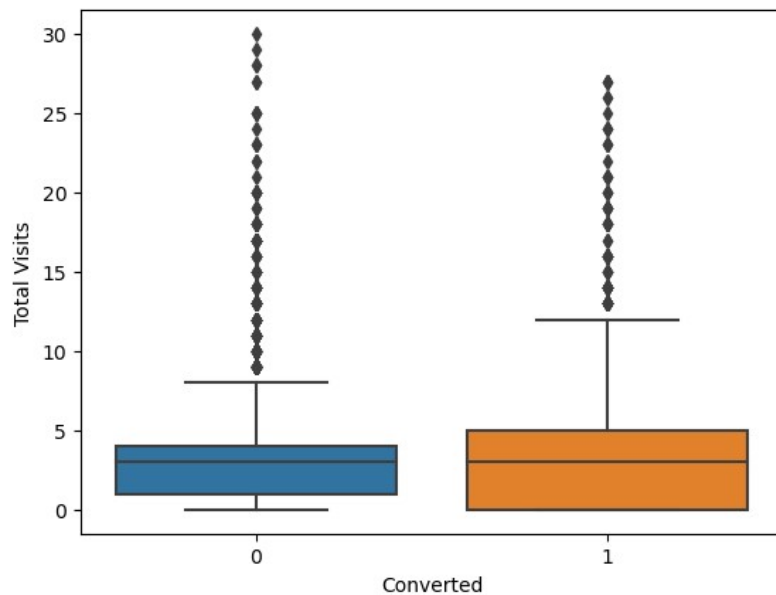
Last Notable Activity v/s Converted

- In case of having last notable activity as having a phone conversation or sending SMS, conversion rate is high.



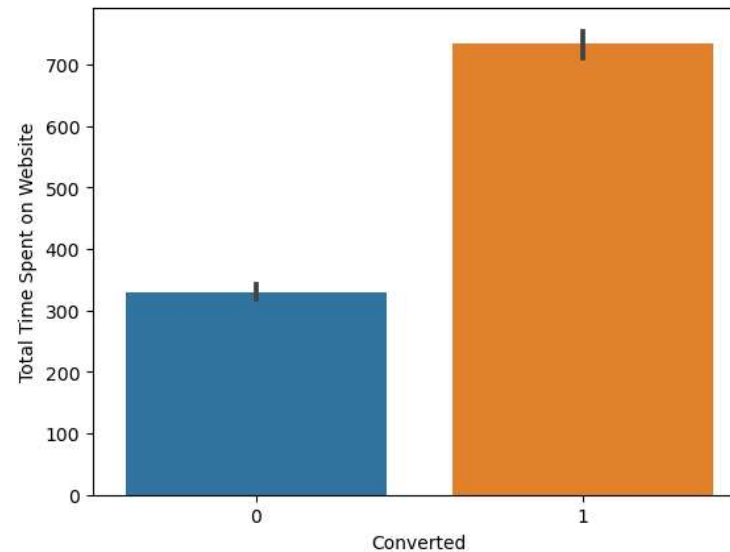
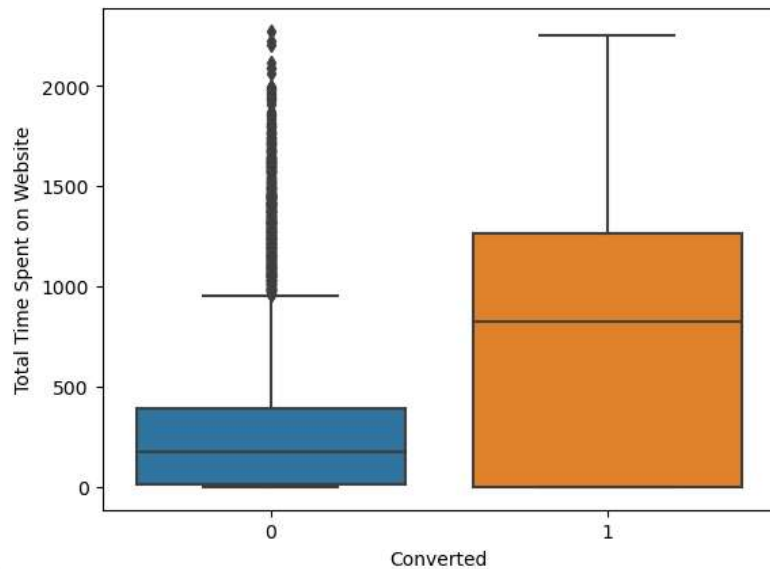
Numerical - Target (TotalVisits v/s Converted)

- There is not huge difference in average value of total visits whether lead is converted or not. Whereas median is same in both case with slight difference in box plot.



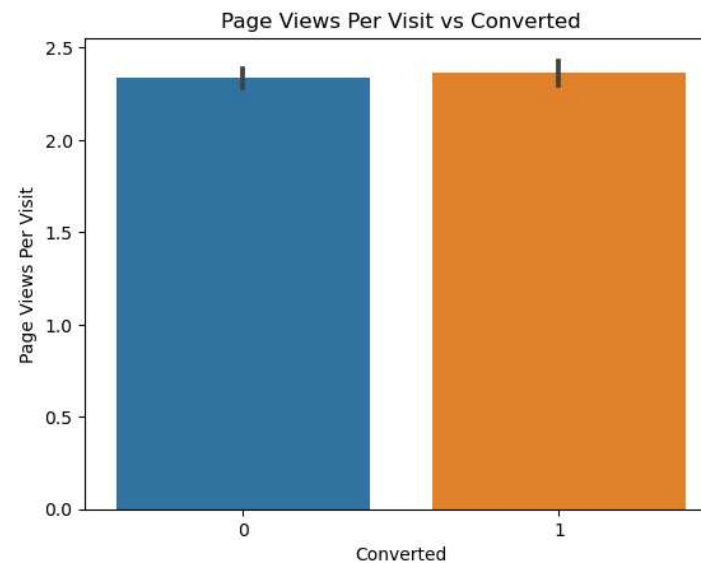
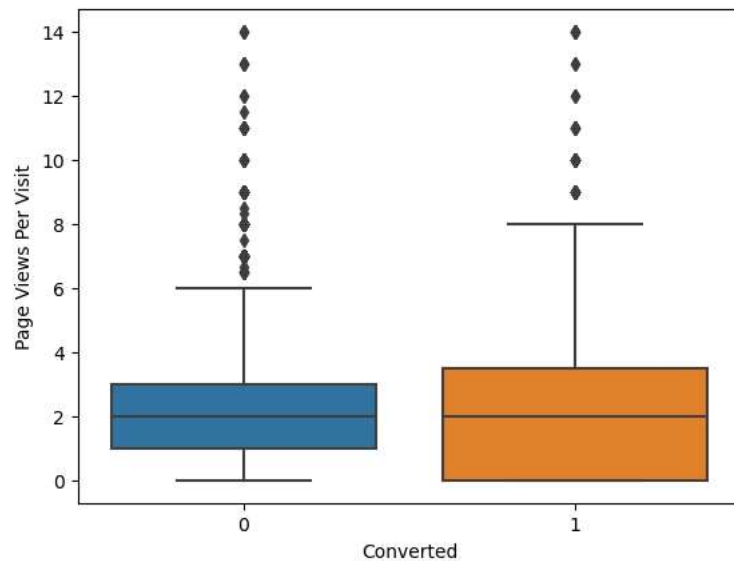
Total Time Spent on Website v/s Converted

- Average amount of total time spent on website is very high when lead is converted successfully compared to non-conversion of the lead.



Page Views Per Visit v/s Converted

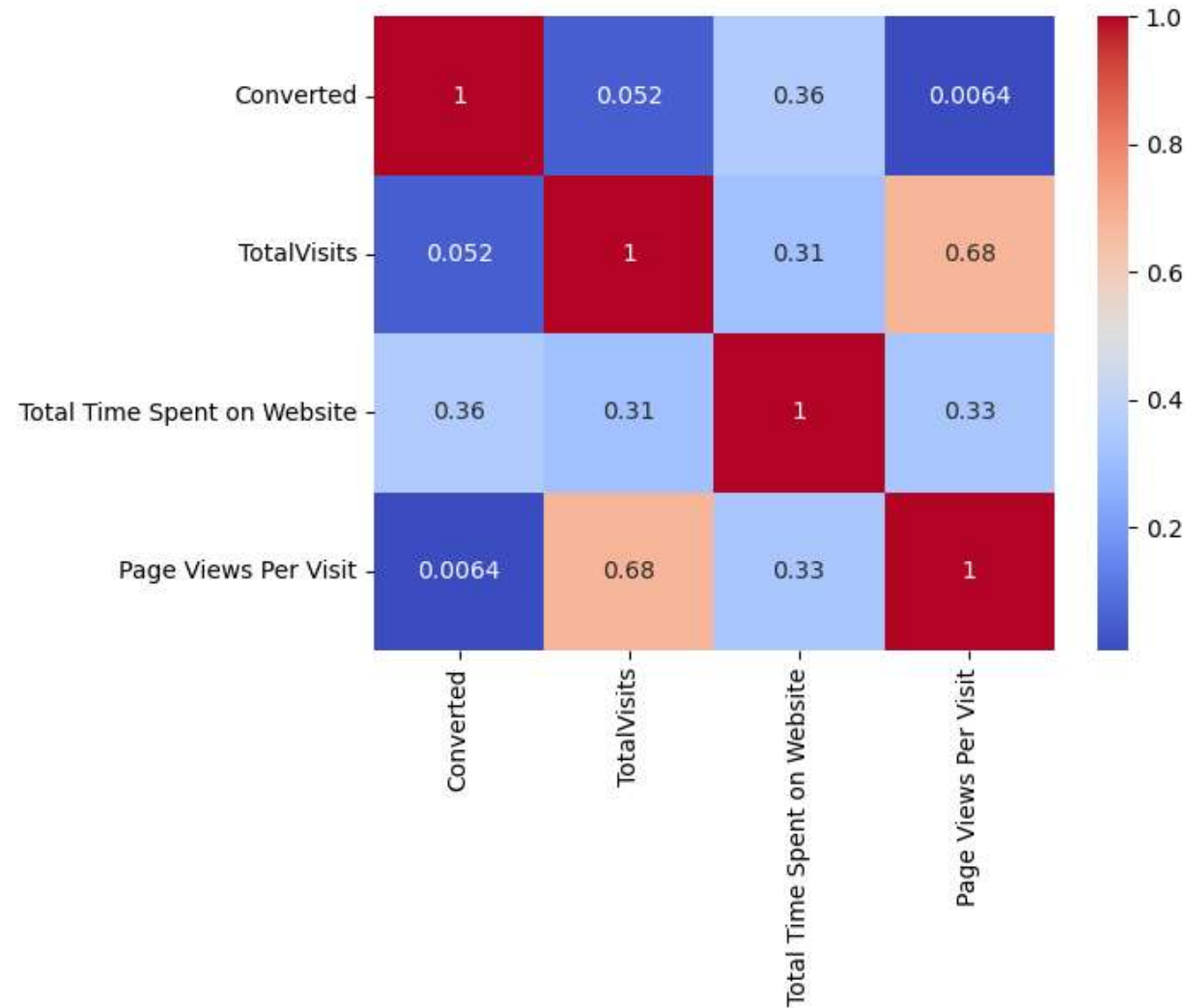
- ▶ Median is same and average of Page Views Per Visit is also very similar for both converted and non-converted. There isn't much difference in boxplot of both as well.



Multivariate Analysis

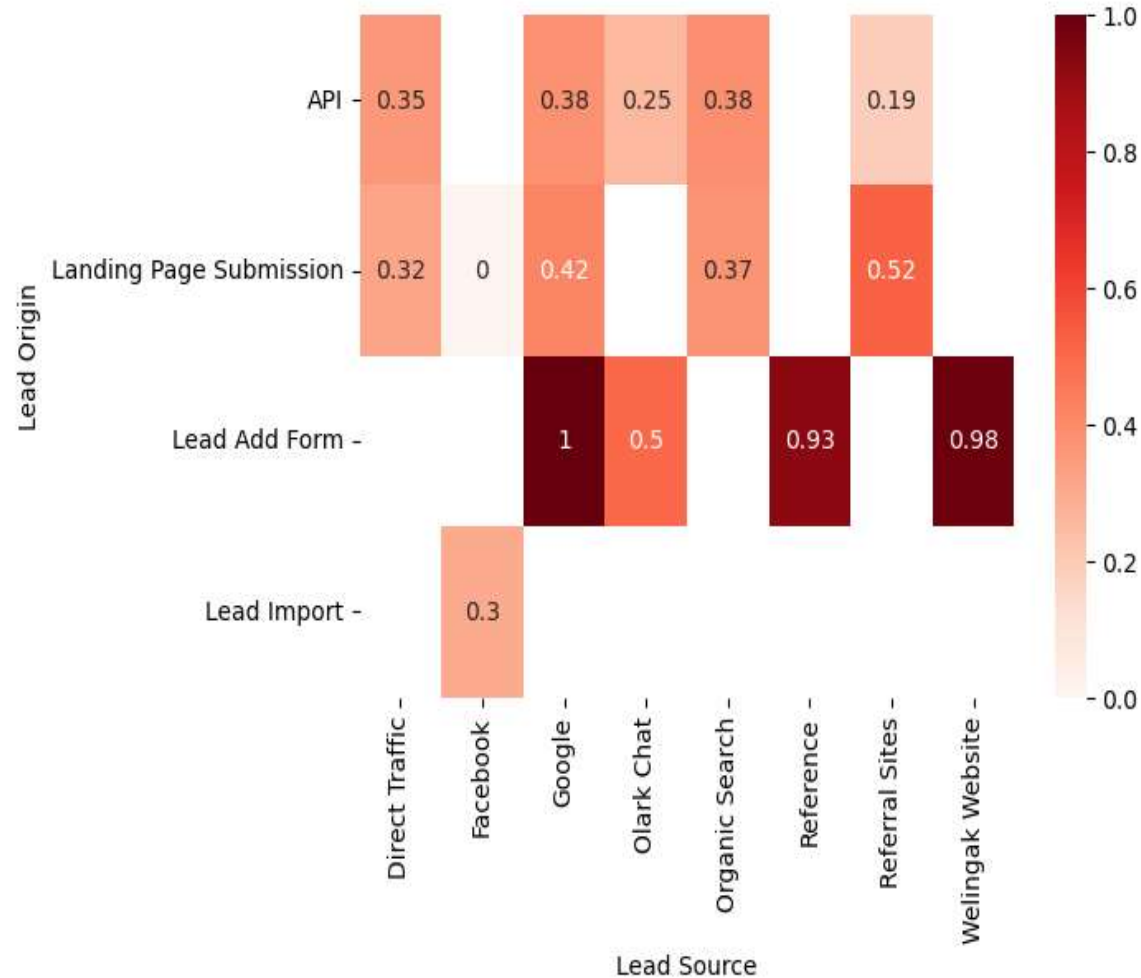
- Correlation matrix

- There is high correlation between Page Views Per Visit and TotalVisits.



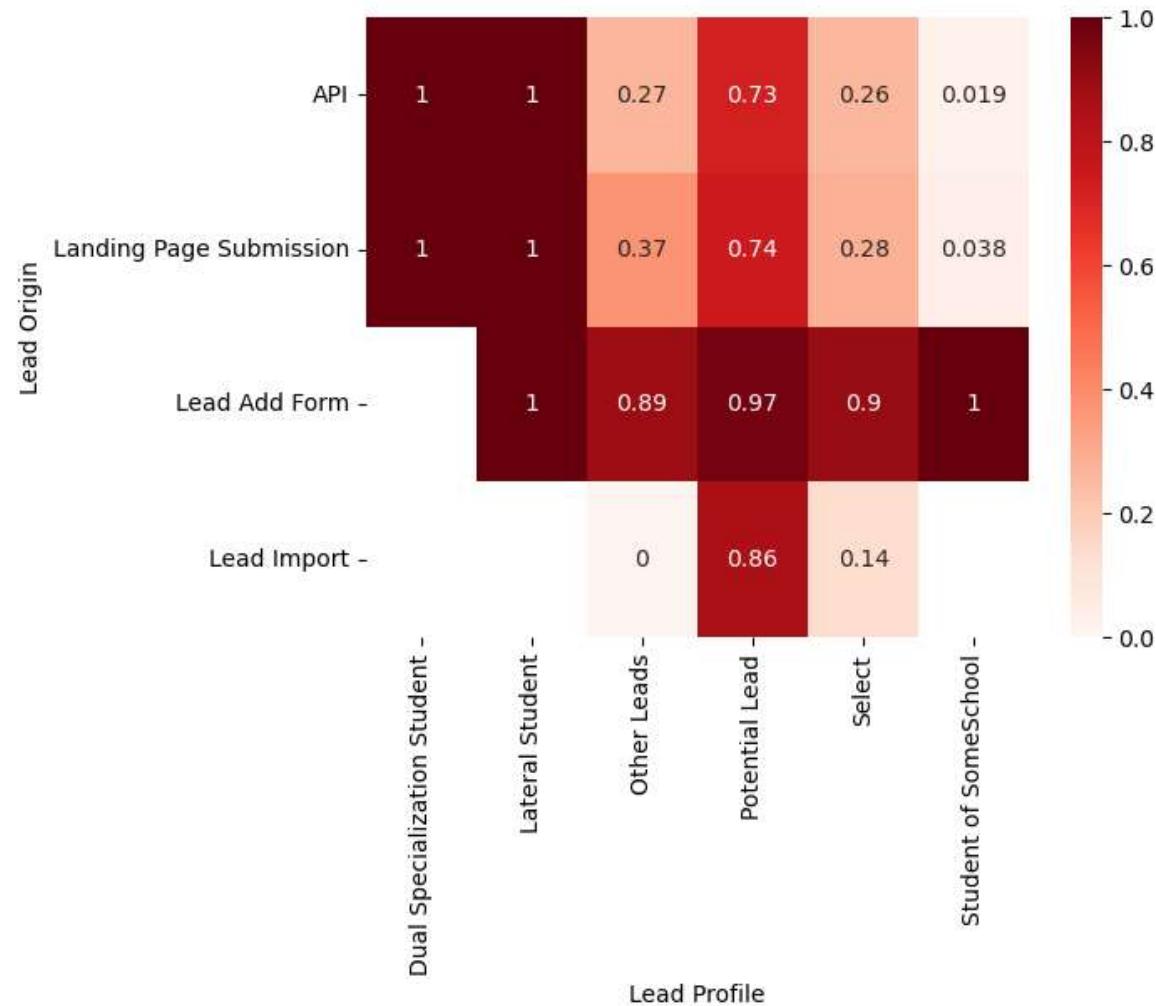
3 Variable correlations - Lead Origin and Lead Source v/s Converted

- There is 100% conversion, when Origin of lead is through Lead Add Form and Source of Lead is Google.



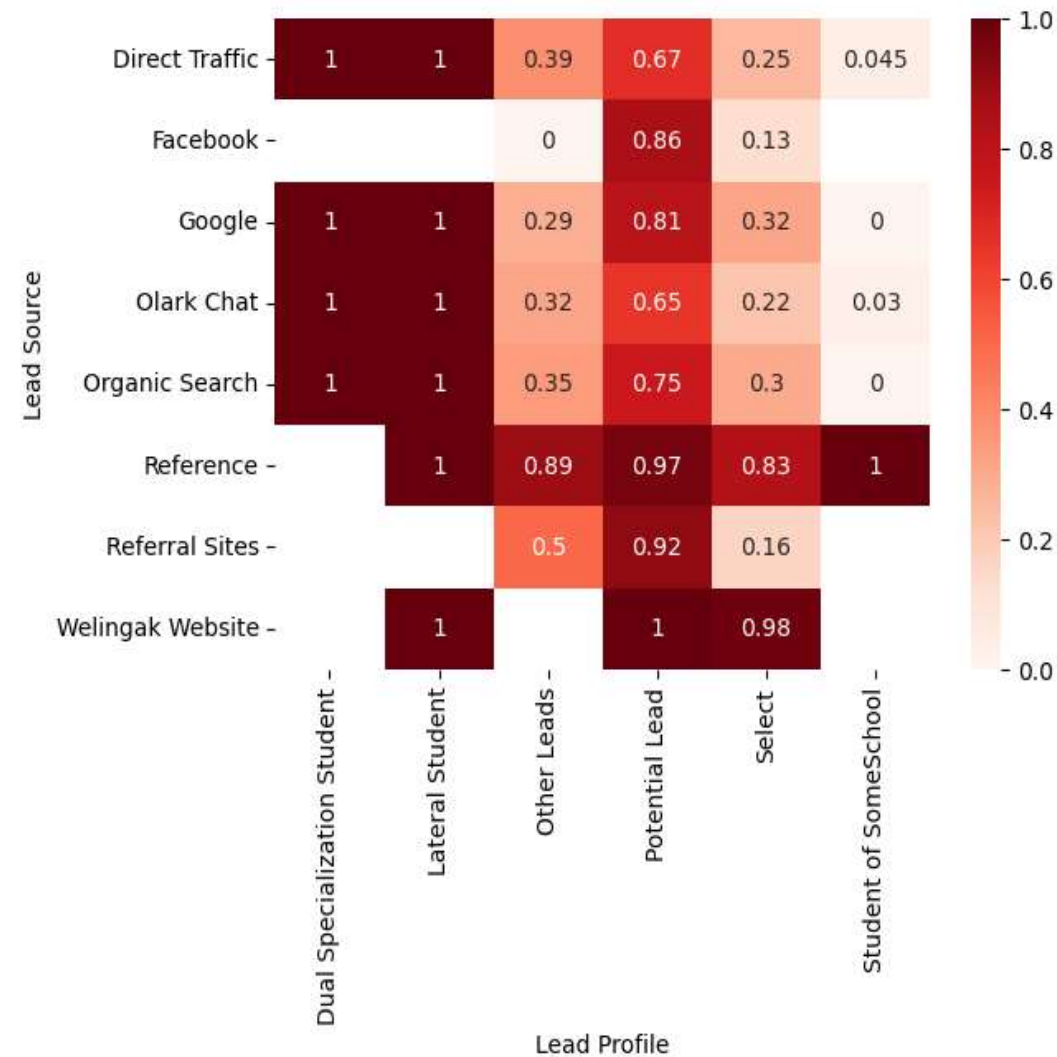
Lead Origin and Lead Profile v/s Converted

- For Dual Specialization student and lateral students, conversion rate is 100%. Conversion rate is also high when lead origin is lead add form.



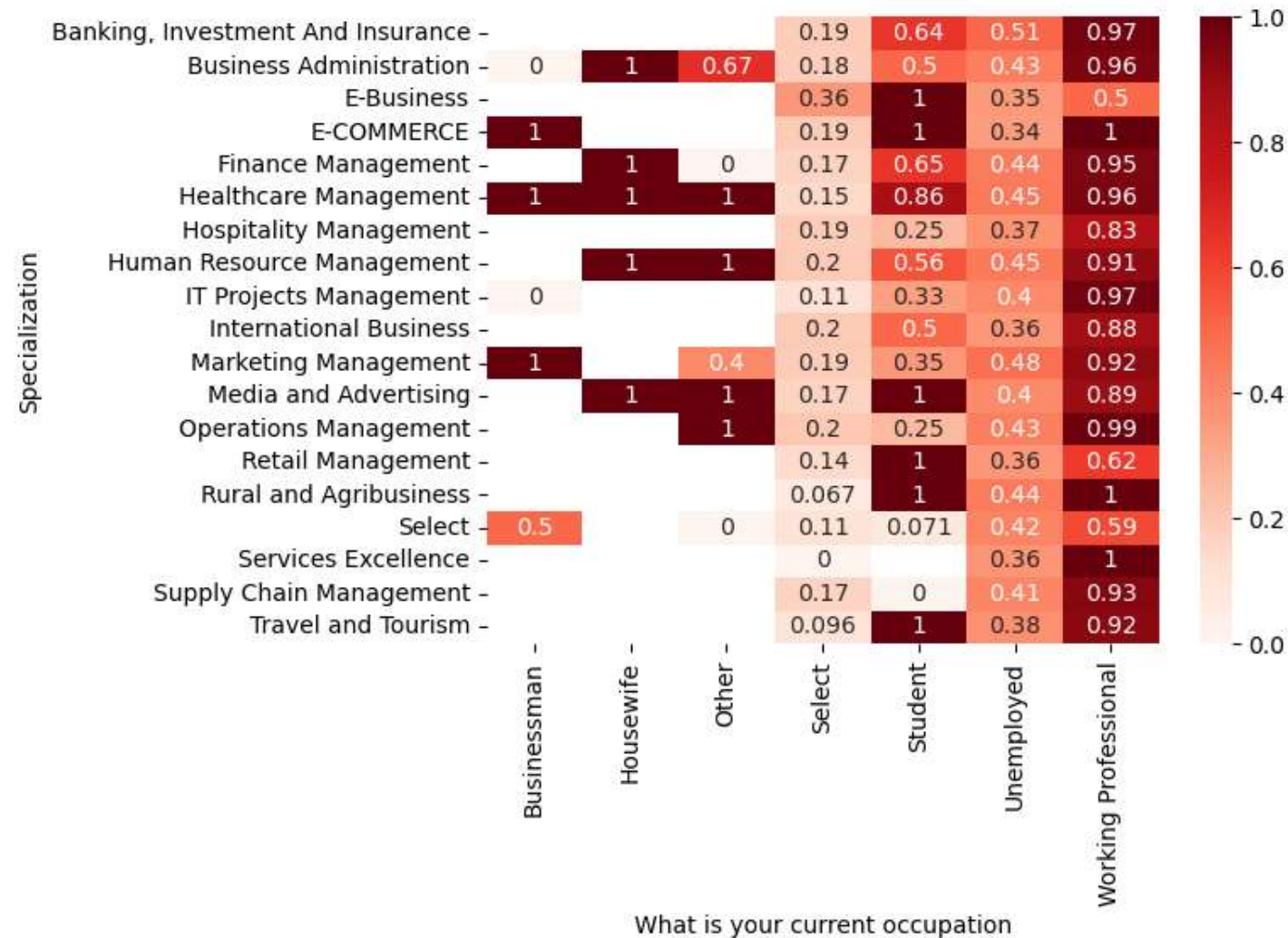
Lead Source and Lead Profile v/s Converted

- Conversion rate is 100% for Dual Specialization students and lateral students. When source of lead is through reference or Welingak website, conversion rate is very high. There is 0% conversion rate for student of some school when source of lead is through organic search.



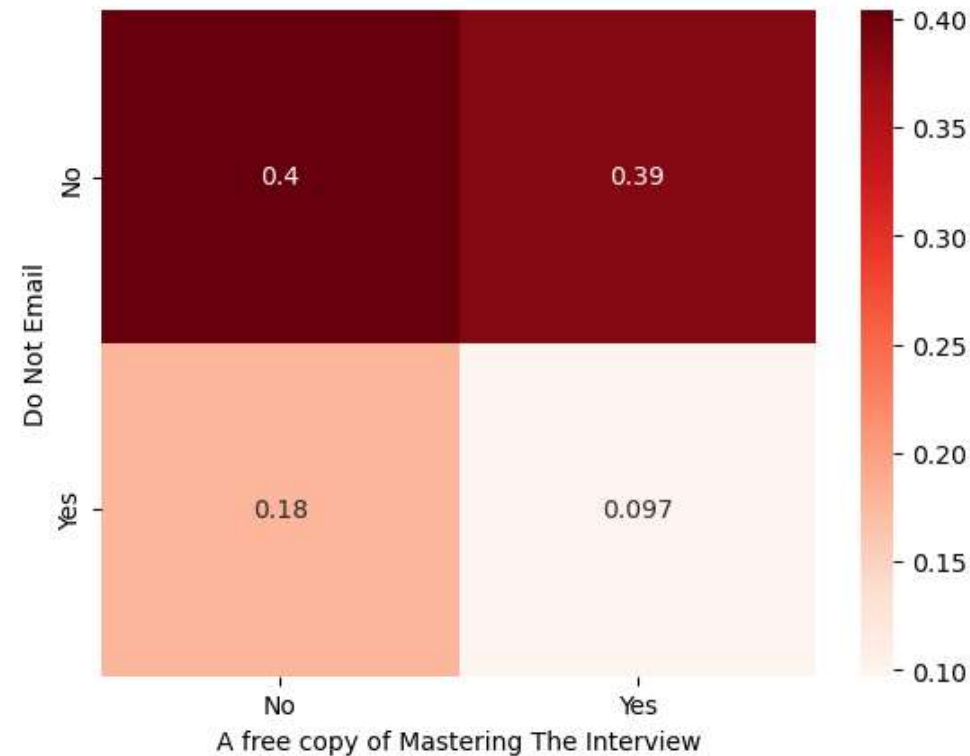
Specialization and occupation v/s Converted

- Conversion rate is very low when occupation is 'Select' means occupation is not declared by user. And for working professional, generally conversion rate is high.



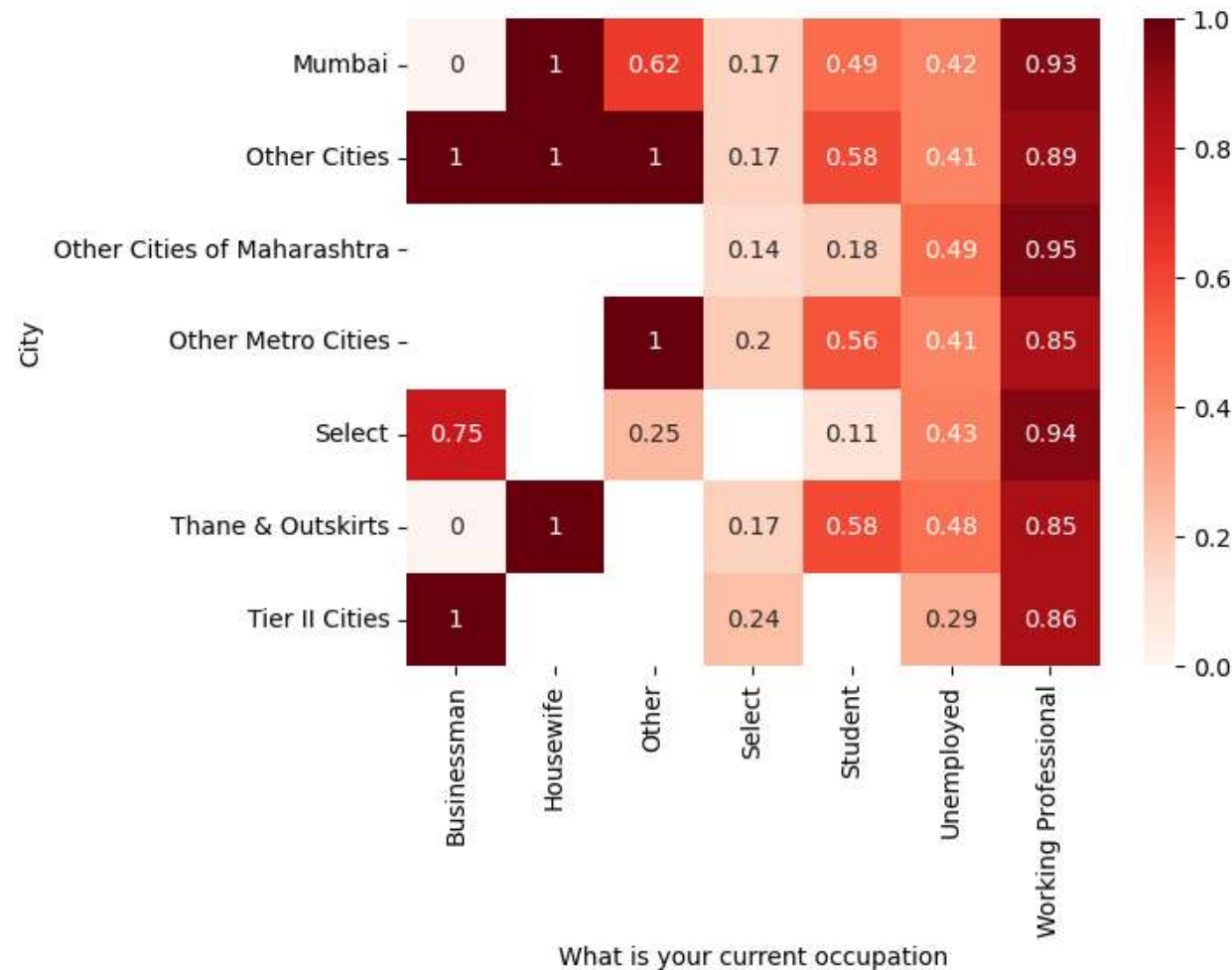
Do Not Email and A free copy of Mastering The Interview v/s Converted

- When user preferred not receive Email and also has free copy of mastering the interview, then conversion rate is low.



City and What is your current occupation v/s Converted

- For working professional, generally the conversion rate is higher. And when occupation is 'Select' which means occupation is not declared, then conversion rate is very low.



The background features abstract green geometric shapes. On the left, a solid green trapezoid points towards the center. On the right, a complex arrangement of overlapping, semi-transparent green triangles and polygons creates a layered, dynamic effect. The text is centered in the white space between these elements.

3) Pre-processing the data for Model Building

3) Pre-processing the data for Model Building

- ▶ It is divided into 3 parts:
 - ▶ 1) Creating Dummy Variables for Categorical Columns
 - ▶ 2) Splitting data into train and test set
 - ▶ 3) Scaling of Data



Creating Dummy Variables for Categorical Columns

- ▶ For binary columns like 'Do Not Email' and 'A free of Mastering The Interview' value 'Yes' is converted to 1 and value 'No' is converted to 0.
- ▶ For columns 'Last Activity', 'Lead Source', 'Lead Origin', 'Last Notable Activity' dummy variables are created with `drop_first = True`
- ▶ For other categorical columns 'Specialization', 'What is your current occupation', 'Lead Profile', 'City' after creating dummy variables, column with value 'Select' is deleted

Splitting data into train and test set

- ▶ Data is divided into train and test set with 70 %- 30% ratio, which means 70% data is used for training and 30% data is used for testing.
- ▶ In variable `y_train` and `y_test`, Converted column from the dataset is used. Whereas, in `X_train` and `X_test`, all the other columns are used.

Scaling of Data

- ▶ MinMaxScaler is used for scaling the data from sklearn.
- ▶ On training set, `fit_transform()` method is used whereas on test set, `transform()` method is used.



The background features abstract green geometric shapes. On the left, a solid green trapezoid points upwards. On the right, a complex arrangement of overlapping, semi-transparent green triangles and polygons creates a layered, crystalline effect. The central text is positioned between these two main graphic elements.

4) Building the Model

- ▶ LogisticRegression() class is used with automated RFE for building the model
- ▶ Earlier, the model is build with high number of features around 20.
- ▶ Functions are created to fit the model and get the summary of the model. So, we don't need to repeatedly write the same code.

Model - 1: With 20 features

- ▶ With this model, there are many features for which p-value is very high (Close to 1)
- ▶ Log-likelihood value is -2386.2 for this model
- ▶ So, the next model is tried with less number of features (15)

Model - 2: With 15 features

- ▶ There still few features for which p-value is high. So, those features are dropped one-by-one to check the further models.
- ▶ Log-likelihood value is -2478.0 for this model, which is high compared to the model with high number of features
- ▶ So, the next model is tried after dropping `occupation_Housewife` column

Model - 3: With dropping occupation_Housewife column

- ▶ There still few features for which p-value is high. So, those features are dropped one-by-one to check the further models.
- ▶ Log-likelihood value is -2481.9 for this model, which is high compared to the model with high number of features
- ▶ So, the next model is tried after dropping Last Activity_Had a Phone Conversation column. Because there is also one column Last Notable Activity_Had a Phone Conversation, which looks similar to this one.

Model - 4: With dropping Last Notable Activity_Had a Phone Conversation column

- ▶ There still 2 features (Lead Profile_Lateral Student and Lead Profile_Dual Specialization Student) for which p-value is high. So, those features are dropped one-by-one to check the further models.
- ▶ Log-likelihood value is -2485.6 for this model, which is higher compared to the model with high number of features
- ▶ So, the next model is tried after dropping Lead Profile_Lateral Student

Model - 5: After dropping Lead Profile_Lateral Student column

- ▶ In this model, there is only 1 feature Lead Profile_Dual Specialization Student for which p-value is high.
- ▶ In the next model, Lead Profile_Dual Specialization Student is dropped.
- ▶ Log-likelihood value is -2493.0 for this model, which is high compared to the model with higher number of features

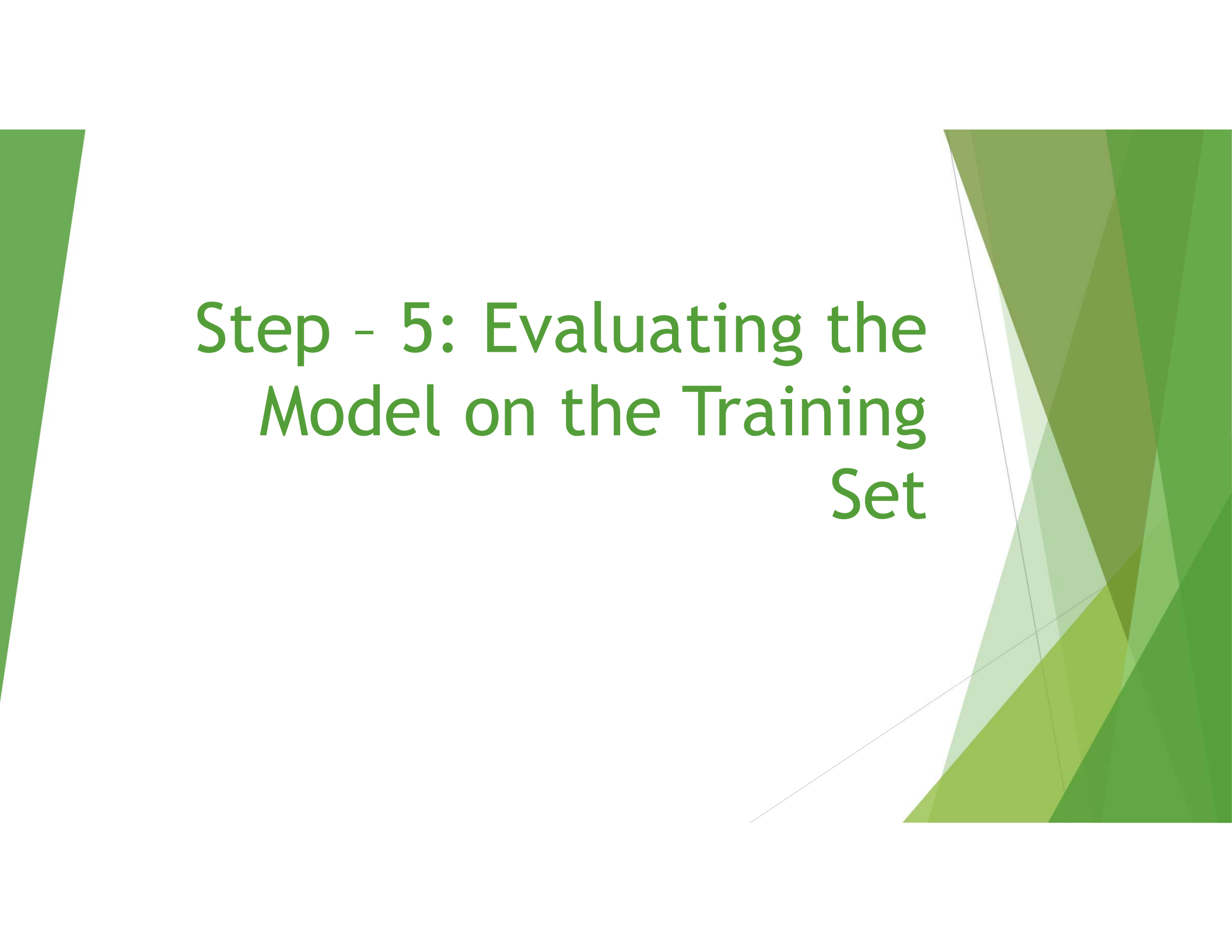
Model - 6: Final Model (After dropping Lead Profile_Dual Specilization Student)

- ▶ P-values for all the features are below 0.05 in this model.
- ▶ Log-likelihood value is -2505.9 for this model
- ▶ Next, Variance Inflation Factor (VIF) is used for checking multi-collinearity present in the model

VIF of Final Model and Selected Features

Features	VIF
Total Time Spent on Website	1.91
TotalVisits	1.78
Lead Origin_Lead Add Form	1.64
Last Activity_SMS Sent	1.44
Lead Profile_Potential Lead	1.40
Lead Source_Welingak Website	1.39
occupation_Working Professional	1.25
Do Not Email	1.06
Lead Source_Olark Chat	1.05
Lead Profile_Student of SomeSchool	1.03
Last Activity_Had a Phone Conversation	1.01

- ▶ VIF for all the features should be less than 5. And for our model, VIF value for all the features is less than 2.
- ▶ As for all the feature in this model, p-value is less than 0.5 and VIF is less than 5, it is selected as final model.



Step - 5: Evaluating the Model on the Training Set

Different evaluation Measures used

- ▶ Accuracy
- ▶ Specificity
- ▶ Sensitivity
- ▶ Precision
- ▶ Recall
- ▶ F1_score



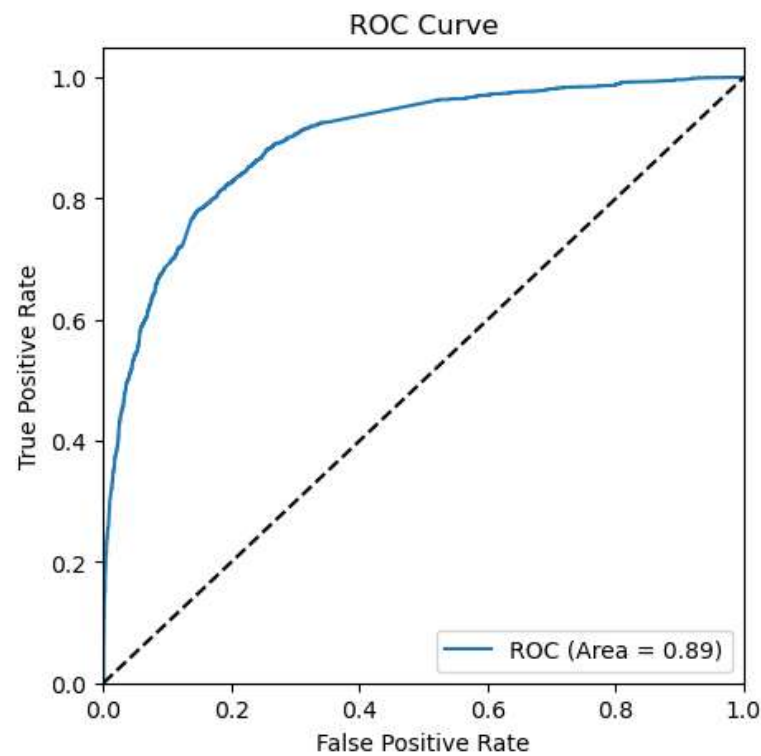
Checking the model with cut-off value 0.5 for prediction

- Various evaluating metrics are as below with this cut-off:

Measure	Value
Accuracy	0.8201267828843106
Specificity	0.9061621068780363
Sensitivity	0.6798666110879533
Precision	0.8163163163163163
Recall	0.6798666110879533
F1_score	0.741869456447578

ROC Curve

- Area under the ROC curve is 0.89 which is really good.

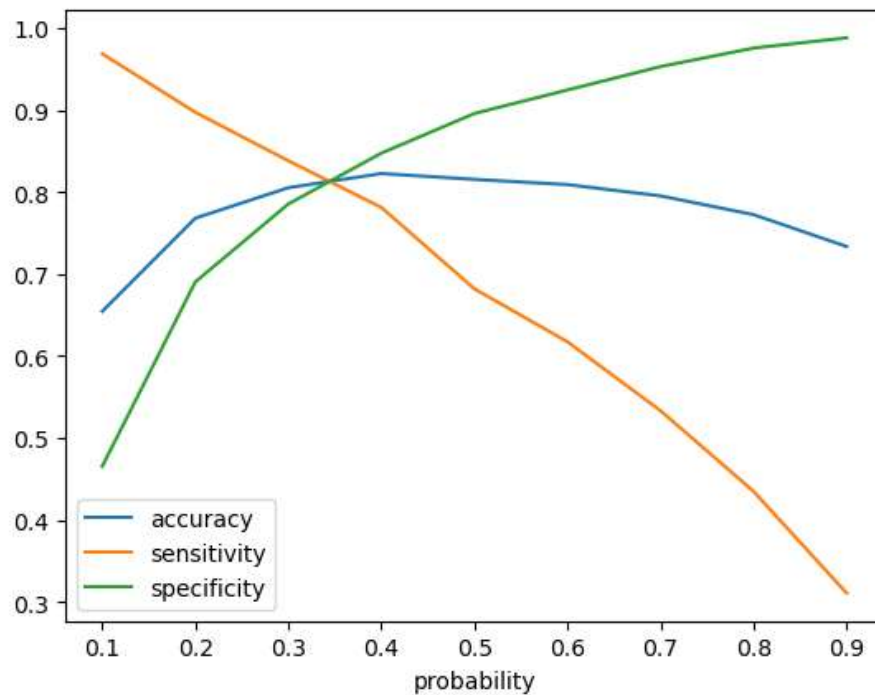


Evaluation metrics with different cut-offs

Cut-off	Accuracy	Sensitivity	Specificity	Precision	Recall
0.1	0.654511	0.968354	0.465583	0.521709	0.968354
0.2	0.768036	0.897046	0.690373	0.635575	0.897046
0.3	0.805137	0.837975	0.785370	0.701519	0.837975
0.4	0.822420	0.781013	0.847346	0.754894	0.781013
0.5	0.815285	0.681857	0.895606	0.797237	0.681857
0.6	0.808942	0.617300	0.924308	0.830778	0.617300
0.7	0.795148	0.533333	0.952756	0.871724	0.533333
0.8	0.772316	0.435021	0.975362	0.914007	0.435021
0.9	0.733471	0.310970	0.987808	0.938854	0.310970

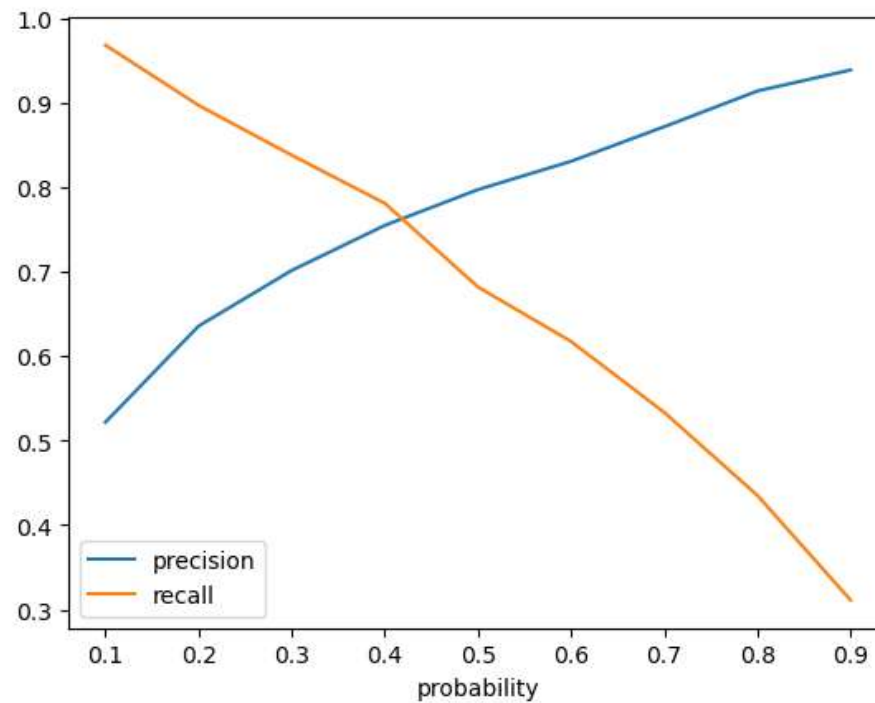
Finding optimal value of cut-off

- With accuracy, sensitivity and specificity graph, optimal value of threshold seems close to 0.35.



Finding optimal value of cut-off

- With precision and recall graph, optimal value of threshold seems close to 0.42



Making prediction with cut-off (0.35)

- With cut-off of 0.35, accuracy is close to 81%, whereas value of precision is close to 73% and recall is close to 81%. Value of specificity is close to 82%.

Measure	Value
Accuracy	0.8256735340729001
Specificity	0.8578368703656354
Sensitivity	0.7732388495206336
Precision	0.7693902944836167
Recall	0.7732388495206336
F1_score	0.7713097713097713

Making prediction with cut-off (0.42)

- ▶ With cut-off of 0.42, accuracy is close to 82%, whereas value of precision and recall is almost same (Close to 77%). And specificity is also very high which is close to 86%. So, it is chosen as final cut-off.
- ▶ Various measures are as below:

Measure	Value
Accuracy	0.8256735340729001
Specificity	0.8578368703656354
Sensitivity	0.7732388495206336
Precision	0.7693902944836167
Recall	0.7732388495206336
F1_score	0.7713097713097713

The background features abstract green geometric shapes. On the left, a solid green trapezoid points towards the center. On the right, a complex arrangement of overlapping translucent green triangles and polygons creates a layered effect. A thin, light gray line extends from the bottom left towards the right side of the composition.

Step - 6: Making predictions on the test set

Scaling and transforming features

- Numerical features TotalVisits, Total Time Spent on Website and page Views Per Visit is transformed with same `MinMaxScaler()` that is applied on training set

Prediction on the Test Set (0.42)


- On the test set as well, model is performer quite well.

Measure	Value
Accuracy	0.8125693160813309
Specificity	0.8421985815602837
Sensitivity	0.7630799605133267
Precision	0.7432692307692308
Recall	0.7630799605133267
F1_score	0.7530443253774963

Comparison of results on Train and Test Set

Train Set	
Measure	Value
Accuracy	0.8256735340729001
Specificity	0.8578368703656354
Sensitivity	0.7732388495206336
Precision	0.7693902944836167
Recall	0.7732388495206336
F1_score	0.7713097713097713

Test Set	
Measure	Value
Accuracy	0.8125693160813309
Specificity	0.8421985815602837
Sensitivity	0.7630799605133267
Precision	0.7432692307692308
Recall	0.7630799605133267
F1_score	0.7530443253774963

The background features abstract green geometric shapes. On the left, a solid green trapezoid points upwards. On the right, a complex arrangement of overlapping translucent green triangles and polygons creates a layered, crystalline effect. The central text is positioned between these two main graphic elements.

7) Business aspects of the model

Co-efficient of all the variables in Final Model

Variable_Name	Coefficient
Const	-3.1351
Do Not Email	-1.4010
TotalVisits	1.7680
Total Time Spent on Website	4.6546
Last Activity_Had a Phone Conversation	2.1145
Last Activity_SMS Sent	1.5764
Lead Origin_Lead Add Form	3.8776
Lead Source_Olark Chat	1.2890
Lead Source_Welingak Website	2.4821
Occupation_Working Professional	2.7065
Lead Profile_Potential Lead	1.8078
Leaf Profile_Student of SomeSchool	-2.1685

Top 3 Contributing variables

- ▶ 1) Total time spent on website
- ▶ 2) Lead add from (From lead origin)
- ▶ 3) Working professional (From occupation)



Business aspects of the model

- For different cut-offs in the test set, value of evaluation metrics are as below in the table:

Cut-off	Accuracy	Sensitivity	Specificity	Precision	Recall
0.1	0.605915	0.967423	0.389480	0.486836	0.967423
0.2	0.757116	0.889437	0.677896	0.623098	0.889437
0.3	0.790388	0.836130	0.763002	0.678686	0.836130
0.4	0.809242	0.769990	0.832742	0.733772	0.769990
0.5	0.803697	0.663376	0.887707	0.779582	0.663376
0.6	0.800000	0.615005	0.910757	0.804910	0.615005
0.7	0.786322	0.536032	0.936170	0.834101	0.536032
0.8	0.771904	0.449161	0.965130	0.885214	0.449161
0.9	0.734935	0.314906	0.986407	0.932749	0.314906

Lead conversion with more aggressiveness

- ▶ For making the sales more aggressive, so that almost all the potential leads are contacted via phone calls, we can use low threshold value.
- ▶ On making the threshold value low, there is very less chance of missing any potential leads.
- ▶ For lower cut-off value, sensitivity is very high which is given by the formula:
Sensitivity = True Positive / (True Positive + False Negative)
- ▶ For lower cut-off value, False Negative records will be very less as most of the records will be categorized as potential leads.

- So, in case company wants to be more aggressive and had a more phone conversation, they should opt for the lower cut-off value like 0.2. For the test set, with cut-off 0.2, sensitivity value is close to 89% which means there is only around 11% chance of missing successfully converted lead.

Cut-off	Accuracy	Sensitivity	Specificity	Precision	Recall
0.1	0.605915	0.967423	0.389480	0.486836	0.967423
0.2	0.757116	0.889437	0.677896	0.623098	0.889437
0.3	0.790388	0.836130	0.763002	0.678686	0.836130
0.4	0.809242	0.769990	0.832742	0.733772	0.769990
0.5	0.803697	0.663376	0.887707	0.779582	0.663376
0.6	0.800000	0.615005	0.910757	0.804910	0.615005
0.7	0.786322	0.536032	0.936170	0.834101	0.536032
0.8	0.771904	0.449161	0.965130	0.885214	0.449161
0.9	0.734935	0.314906	0.986407	0.932749	0.314906

Reducing rate of useless phone calls

- ▶ When company reaches its target and wants to only focus on minimizing the rate of useless phone calls, company could increase the cut-off value.
- ▶ On making the threshold value high, there is very less chance of selecting lead which is not converted.
- ▶ For higher cut-off value, specificity is very high which is given by the formula:
Specificity = True Negative / (True Negative + False Positive)
- ▶ For higher cut-off value, False Positive records will be very less as only most promising lead for conversion is chosen.

- So, in case company want to have minimal useless phone calls, they should opt for the higher cut-off value like 0.8. For the test set, specificity value is close to 96.5% which means there are only 3.5% chances of selected lead is not converted and company makes useless phone calls.

Cut-off	Accuracy	Sensitivity	Specificity	Precision	Recall
0.1	0.605915	0.967423	0.389480	0.486836	0.967423
0.2	0.757116	0.889437	0.677896	0.623098	0.889437
0.3	0.790388	0.836130	0.763002	0.678686	0.836130
0.4	0.809242	0.769990	0.832742	0.733772	0.769990
0.5	0.803697	0.663376	0.887707	0.779582	0.663376
0.6	0.800000	0.615005	0.910757	0.804910	0.615005
0.7	0.786322	0.536032	0.936170	0.834101	0.536032
0.8	0.771904	0.449161	0.965130	0.885214	0.449161
0.9	0.734935	0.314906	0.986407	0.932749	0.314906