Lead Scoring Case Study

Created by

Tanmay Kakde

Satish Reddy

Vivek Ghadiyaram

INDEX

Sl. No.	Topic
1	Problem Statement
2	Business Goals
3	Strategy Used
4	Data Preparation
5	Data Standardization and Outlier Check
6	Exploratory Data Analysis (Univariate analysis)
7	Exploratory Data Analysis (Bivariate analysis)
8	Correlation Test
9	Model Building
10	Model Evaluation
11	Insights
12	Conclusions
13	Recommendations

Problem Statement

- An education company named X Education sells online courses to industry professionals. The company markets its courses on several websites and search engines like Google.
- X Education wants to select the most promising leads, i.e. the leads that are most likely to convert into paying customers.
- As interested leads may browse the courses or fill up a form for the course or watch some videos. Once the leads are acquired, the sales team will start interacting with the leads to convert them into customers.
- Through the above-explained process, so far the company is able to reach a 30% lead conversion rate.
- To make the lead scoring process more efficient, the company has now decided to identify the most potential leads, which are known as ‘Hot Leads’, and focus on communicating more with these potential leads rather than focusing on nonpotential leads.

Business Goals

- **Logistic Regression Modelling:**
  The company wants to build a logistic regression model that assigns a lead score between 0 and 100 to each lead which can later be used by X Education to target potential leads. The higher the lead score higher is the chance of the lead getting converted.

- **Recommendations:**
  The problems presented by the X Education company need to be resolved by the model to fit the X Education company's requirements which may even change in the future. Hence the model needs to be adjusted based on the dynamic needs of X Education, which will be discussed further in this presentation.

Strategy Used

- Import required libraries.
- Import data.
- Cleaning data (removing null values, removing unwanted columns, etc).
- Exploratory Data Analysis on data to get a further understanding of data.
- Data preparation for model building.
- Build a Logistic Regression model.
- Assigning a lead score to each lead.
- Test the model on the training dataset.
- Evaluation of the model based on different measures and metrics.
- Test the model on the test dataset.
- Evaluation of the accuracy of the model based on different measures and metrics.

Data Preparation

- In the dataset the categorical variables that are having value as “Select” shows that no option was chosen and is hence it is treated as a null value. The value “Select” has been replaced with NaN for the same.
- There are 7 columns with null values greater than 40% in the dataset and hence it is dropped.
- Missing values in ‘What is your current occupation’, ‘Last Activity’, and ‘Lead Source’ columns are imputed with the modal values for each column respectively.
- ‘TotalVisits’ and ‘Page Views per Visit’ columns are continuous columns with 1.48% null values in each of them. These are imputed with the modal values of the respective columns.
- Columns that are irrelevant in building the regression model or that are having only one unique value have been dropped.
- Highly skewed columns can affect the result of logistic regression models, as they can lead to biased results or inaccurate parameter estimation. Hence these column needs to be dropped as they will not add any value to the model.

Data Standardization and Outlier Check

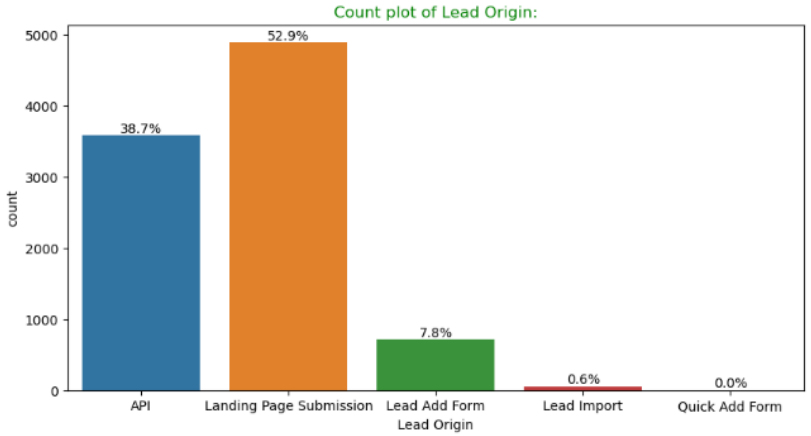
- After analyzing the data we found that the "Lead Score" and "Last Activity“ columns have a number of labels whose value count is negligible and hence can be grouped together under "Others" to remove columns that are irrelevant for regression analysis.

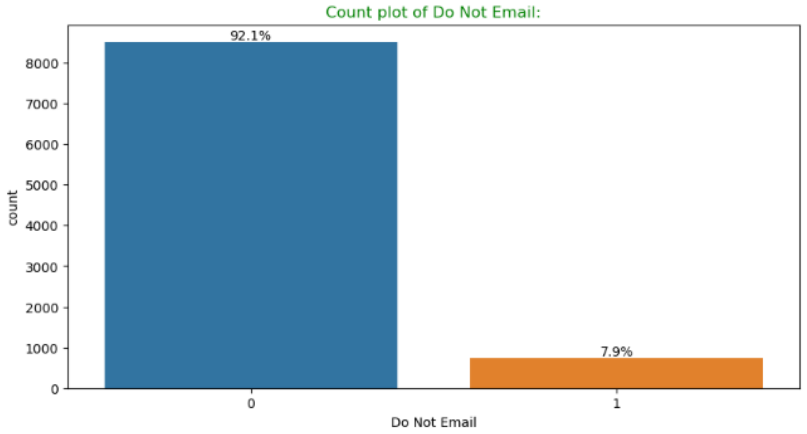
- "Free_copy" and "Do Not Email" columns are both binary categorical columns. To standardize these columns, the values ‘yes’ and ‘no’ needs to be converted to 0 & 1.

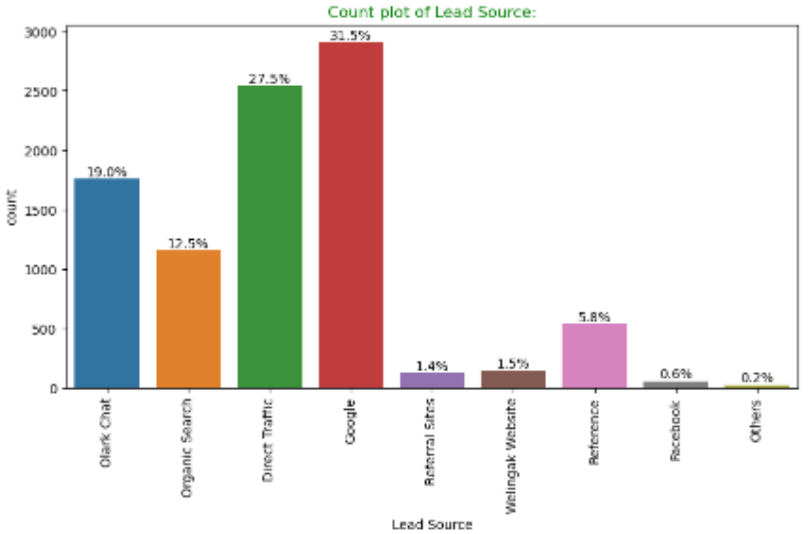
- Columns that are having same value but are present in different cases for example "Google" & "google" are the same in "Lead Source", and are sanitized by replacing lower case values with upper case, resulting in standardizing the data.

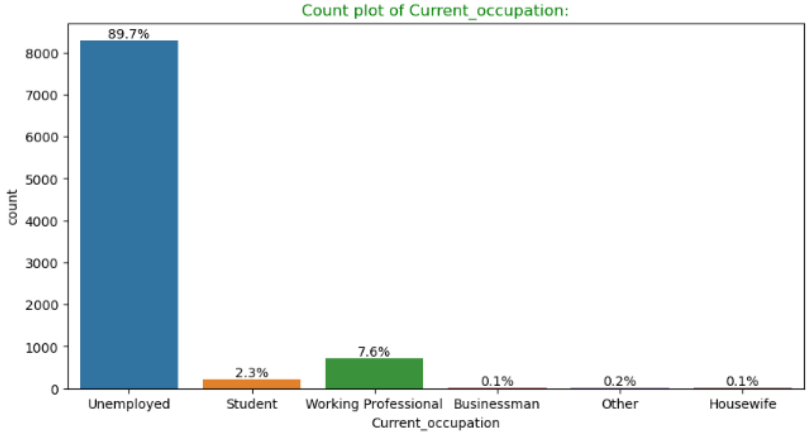
- An Outliers in the data are checked using boxplots. They are treated by defining the upper and lower limits and replacing the values that lie outside the defined range with the correct value to remove an outlier.

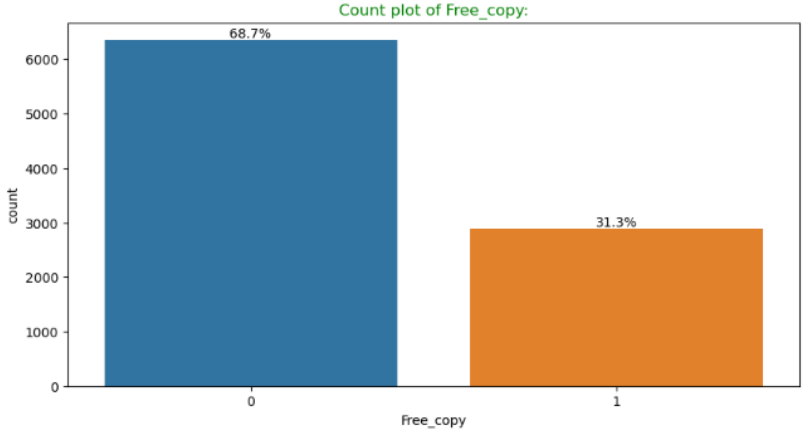
Exploratory Data Analysis (Univariate analysis)

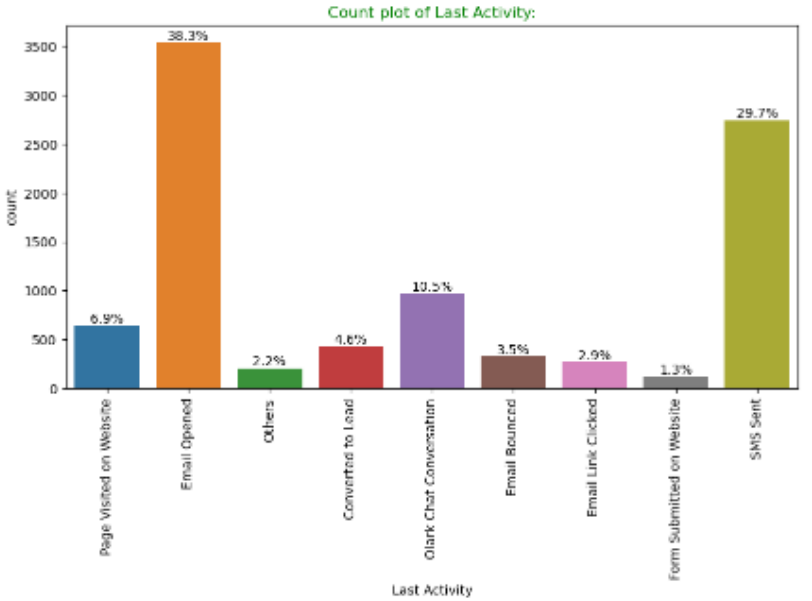




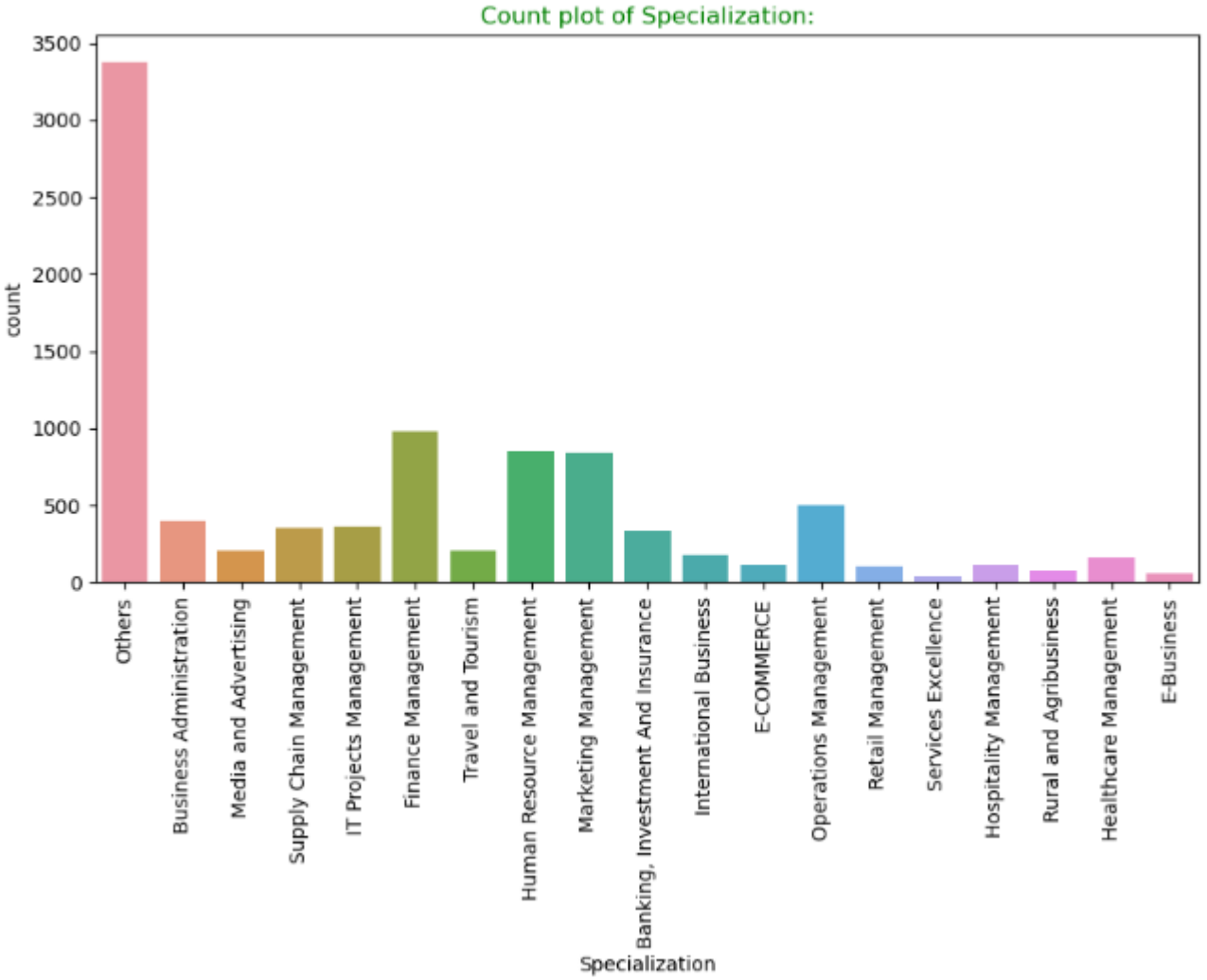








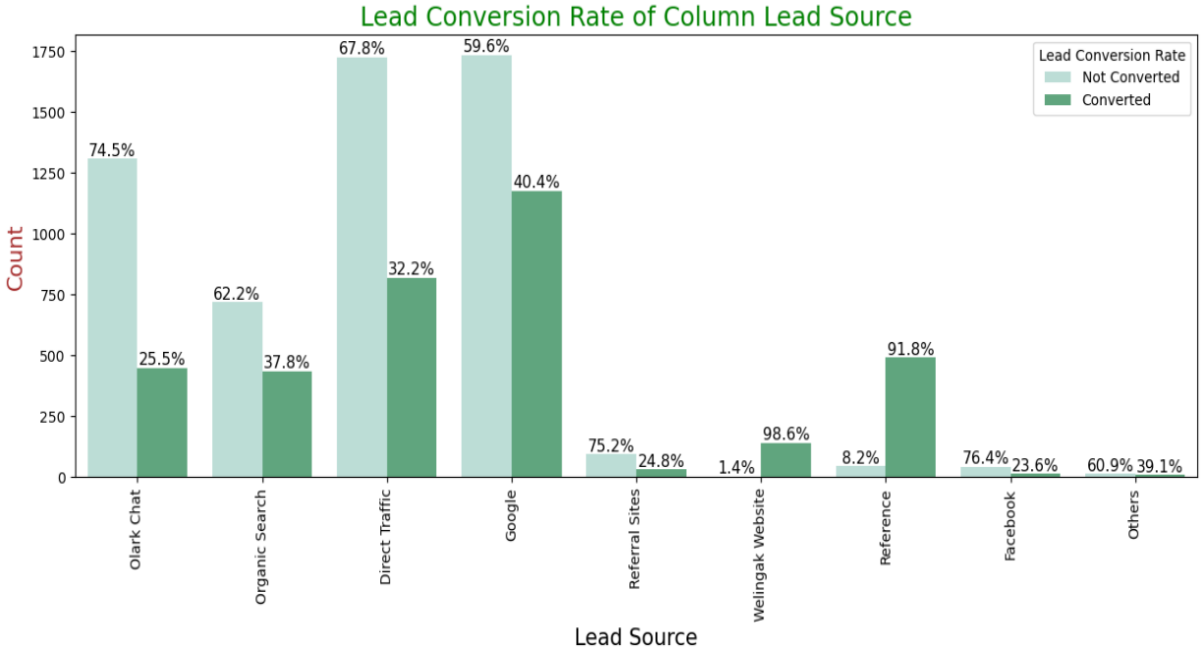
Exploratory Data Analysis (Univariate analysis)



Insights From Univariate Analysis:

- **Do Not Email**: 92% of the people has opted that they dont want to be emailed about the course.
- **Lead Source**: 58% Lead source is from Google & Direct Traffic combined
- **Current_occupation**: It has 90% of the customers as Unemployed
- **Lead Origin**: "Landing Page Submission" identified 53% customers, "API" identified 39%.
- **Last Activity**: 68% of customers contribution in SMS Sent & Email Opened activities

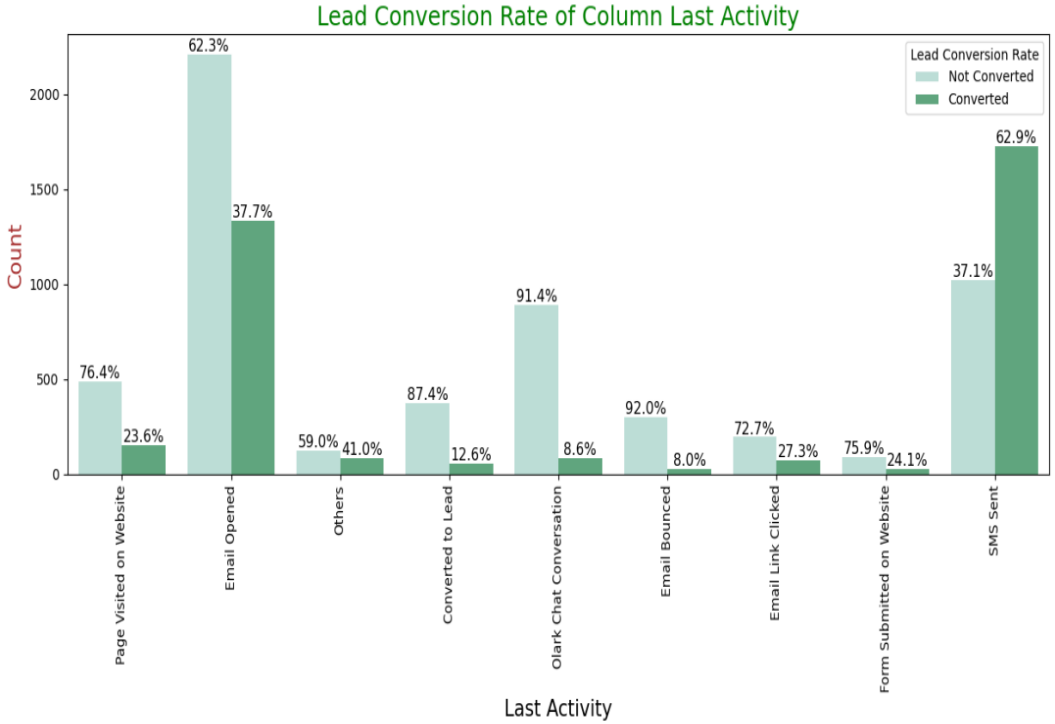
Exploratory Data Analysis (Bivariate analysis)



The conclusion from the above Graph:-

(Lead Source vs Conversion Rate)

- **Google** has a **High** Lead Conversion Rate than other modes.
- **Direct Traffic** comes next to Google search that have the **High** Lead Conversion Rate.
- **Organic Search** has a **Moderate** Lead Conversion Rate but the contribution is by only 12.5% of customers.
- **Reference** has a **High** Lead Conversion Rate of 91% but there are only around 6% of customers through this Lead Source.

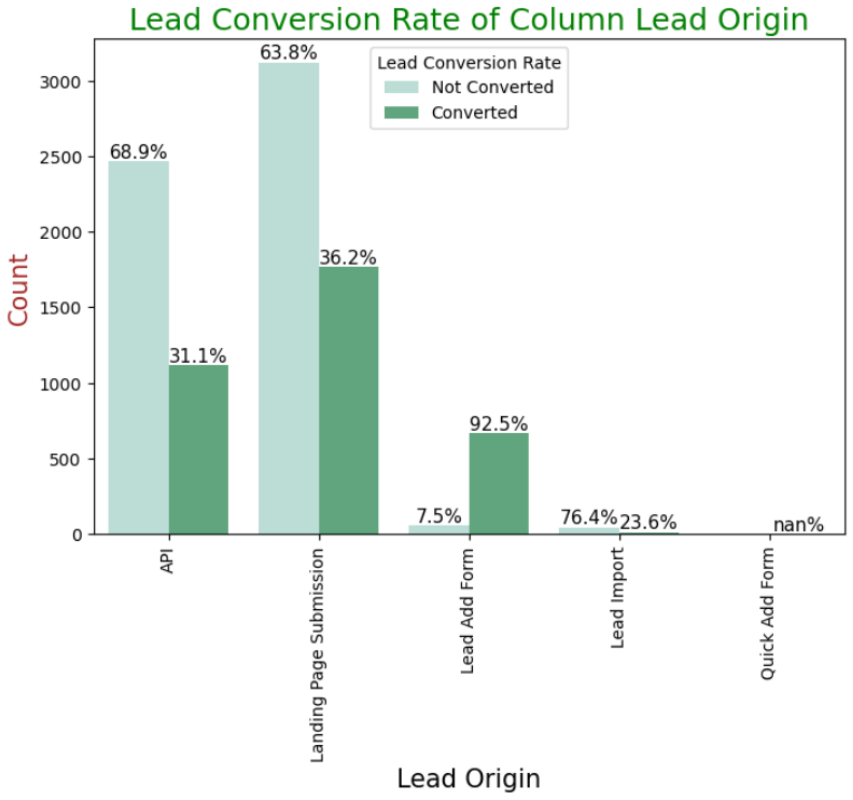


The conclusion from the Graph:-

(Last Activity vs Conversion Rate)

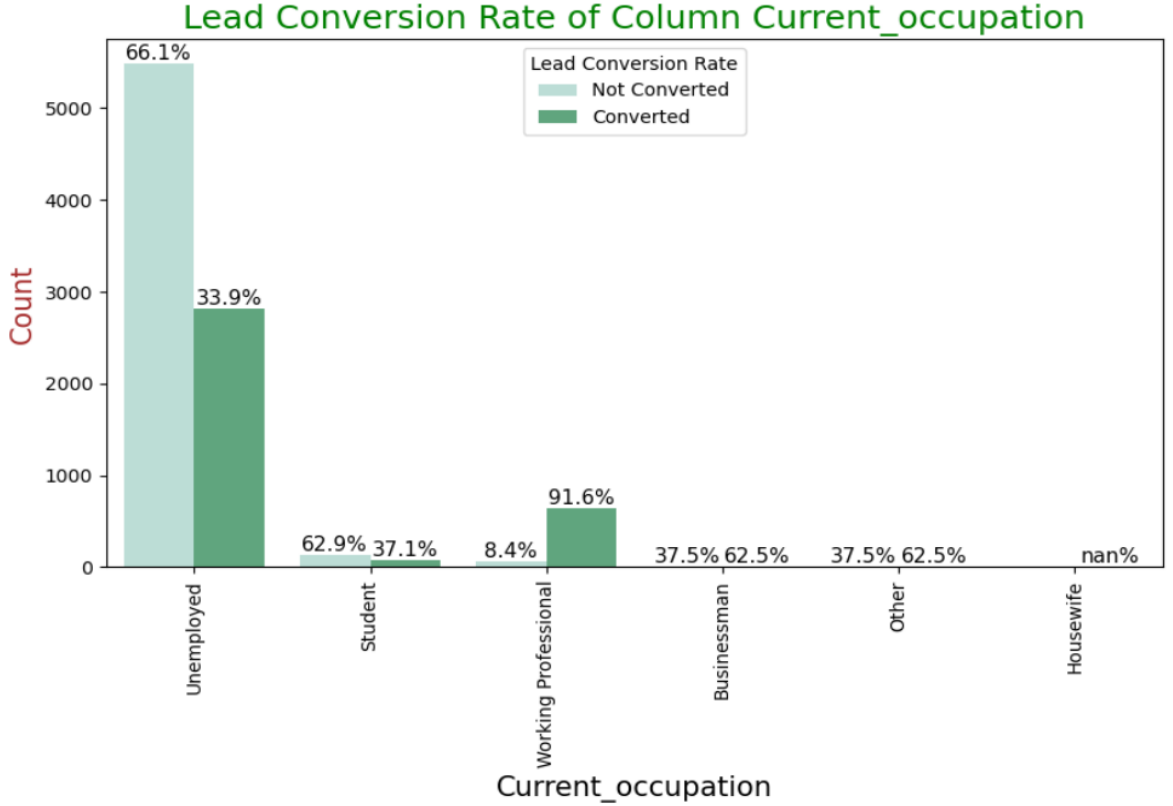
- **SMS sent** has the **highest** lead conversion rate among the other last activity mode.
- **Email Opened** is the next to SMS sent that has the **highest** conversion rate.

Exploratory Data Analysis (Bivariate analysis)



The conclusion from the above Graph:-
(Lead Origin vs Conversion Rate)

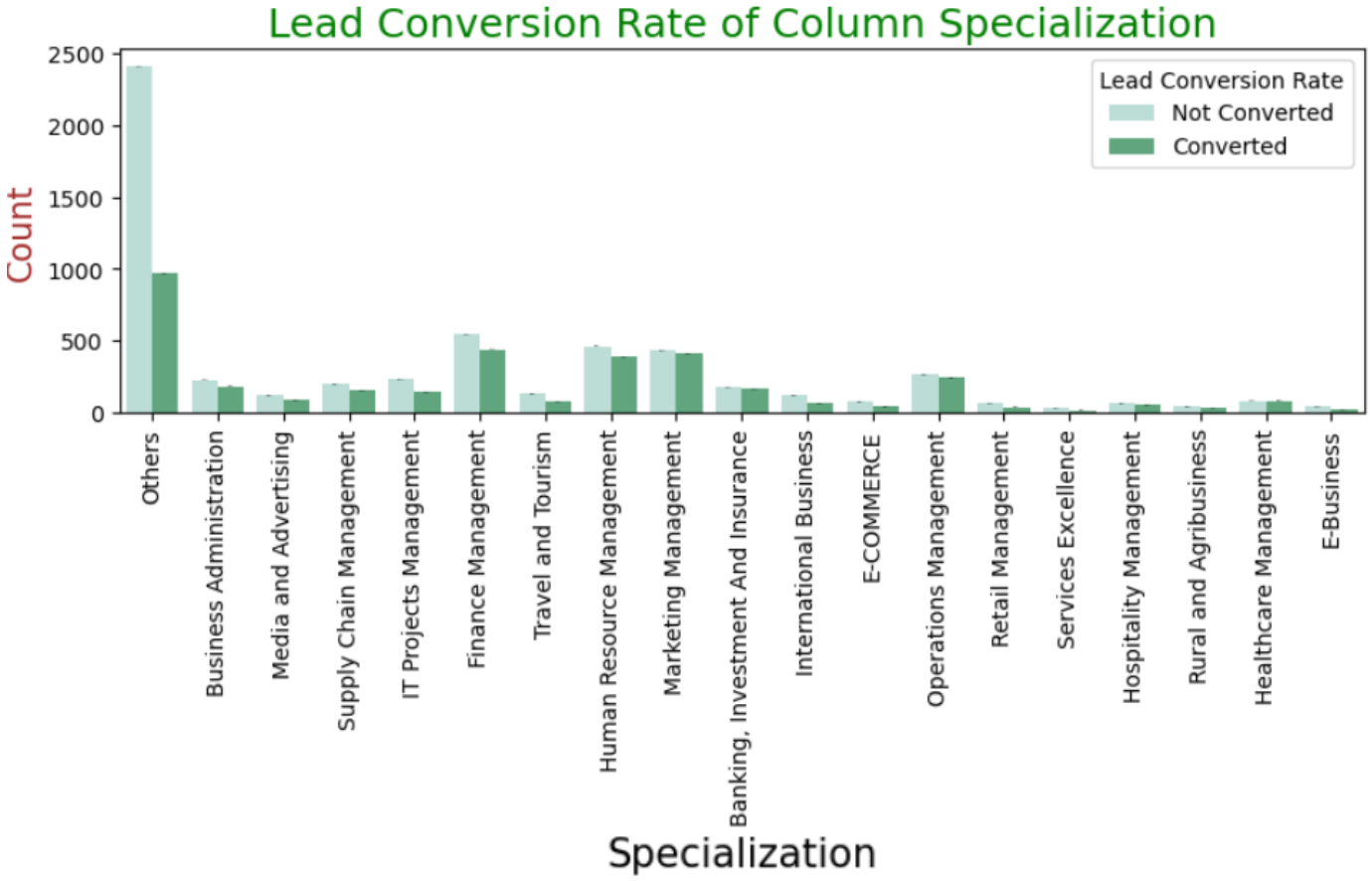
- **Landing Page Submission** has the **highest** Lead conversion rate among the other lead origins.



The conclusion from the above Graph:-
(Lead Current_Occupation vs Conversion Rate)

- Customers who are **unemployed** have the **highest** Lead conversion rate than that of the customers who are employed.

Exploratory Data Analysis (Bivariate analysis)

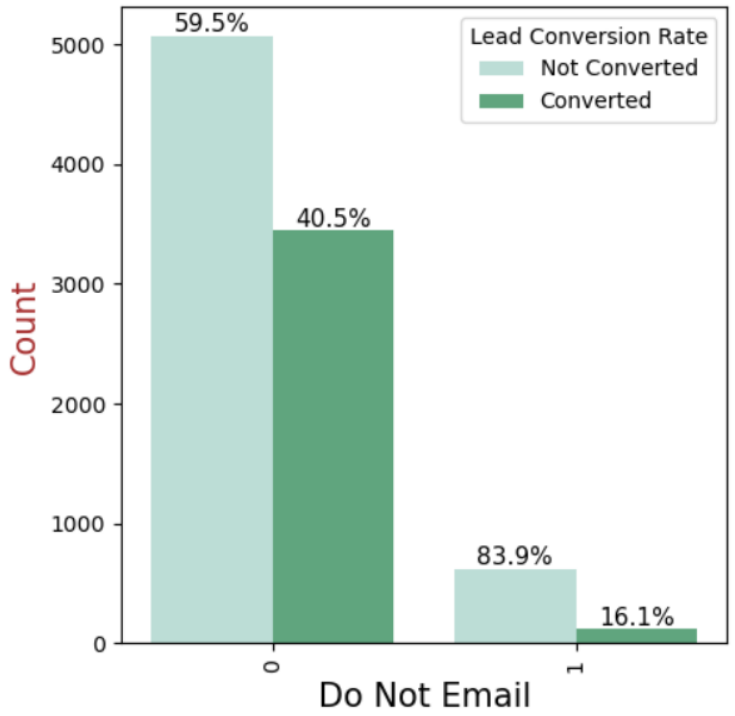


The conclusion from the Graph:-
(Specialization vs Conversion Rate)

- **Marketing Management**, **HR Management**, and **Finance Management** show **good** contributions in converting the leads among the other specialization.

Exploratory Data Analysis (Bivariate analysis)

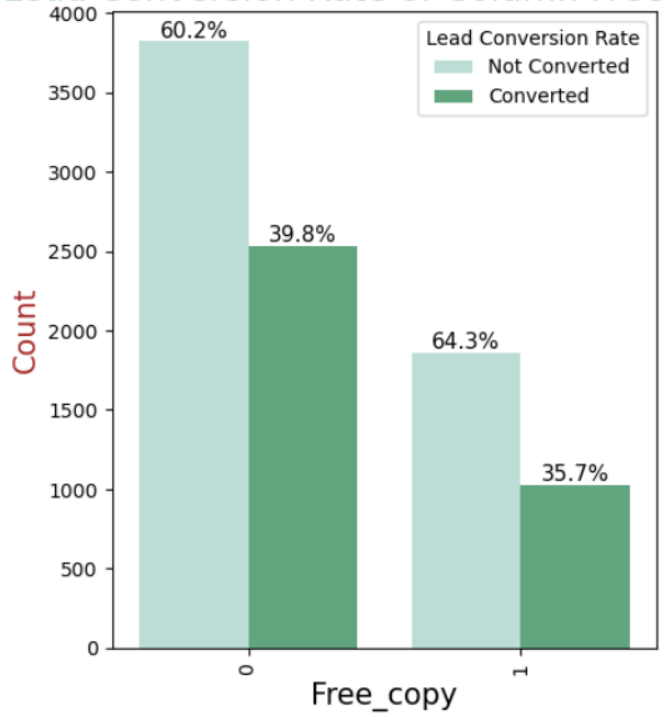
Lead Conversion Rate of Column Do Not Email



The conclusion from the above Graph:-
(Do not Email vs Conversion Rate)

- **Majority** of the leads got converted, those who are **not contacted by E-mail**.

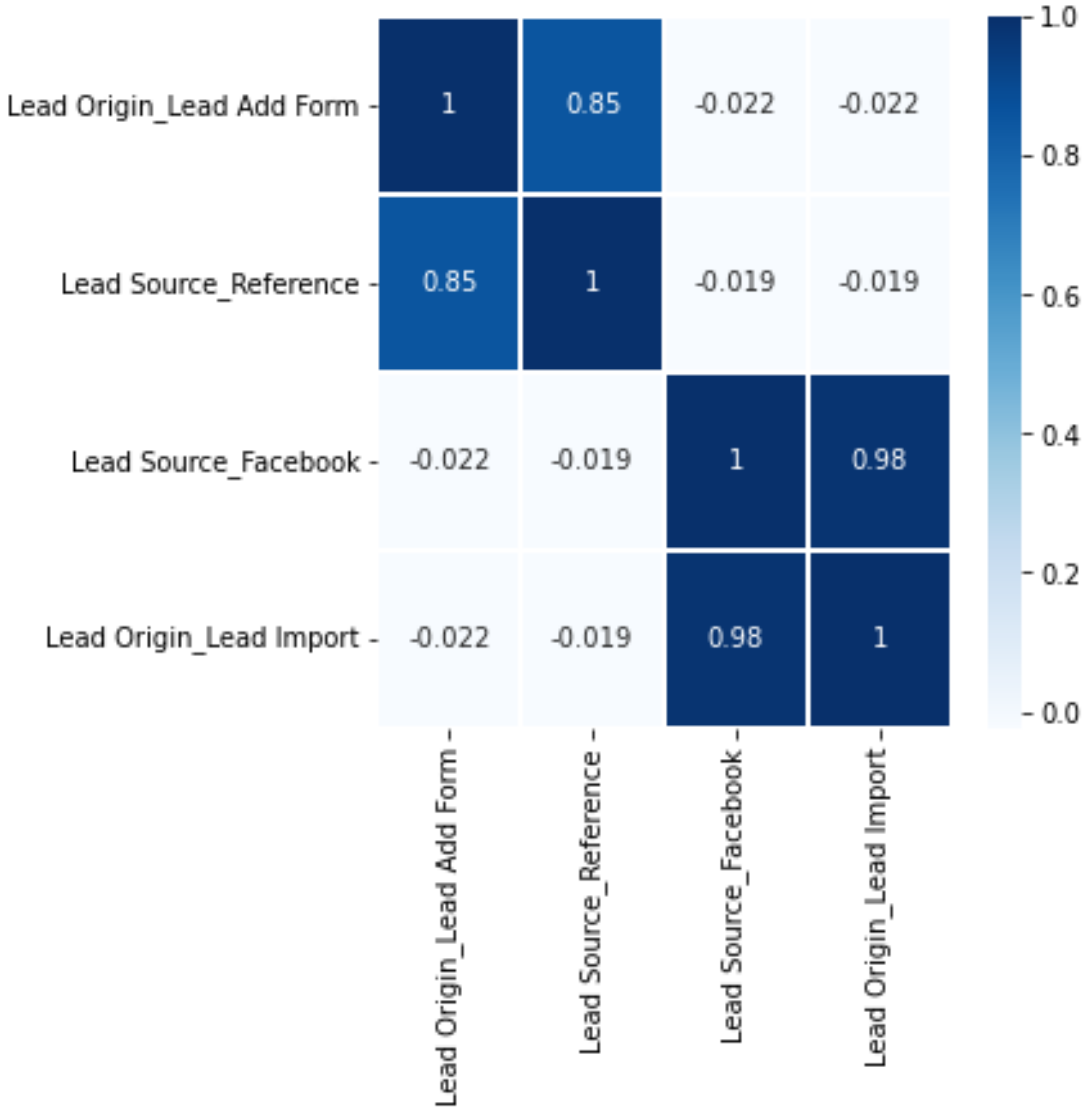
Lead Conversion Rate of Column Free_copy



The conclusion from the Graph:-
(Lead Source vs Conversion Rate)

- Customers who didn't opt for a free copy show **good** contributions in converting the leads than the ones who opted for a free copy.

Correlation Test



The conclusion from the HeatMap:-

- **Lead Source_Facebook** has a **high correlation** among the others.
- **Lead Origin_Lead** Add Form has the next **high correlation** with Leas Source_Reference.

Model Building

- Here we have used a logistic Regression Model which has been used for predicting categorical variables in this case study. The procedure involves two stages:
  1. RFE (Recursive Feature Elimination) with coarse tuning.
  2. Manual fine-tuning using p-values and VIFs (Variance Inflation Factors).

- The steps to build the model are as follows:
  - Creating Dummy Variables.
  - Splitting the Dataset into train set and test set.
  - Scaling of Features
  - Feature Elimination based on Correlation Check.
  - Feature Selection Using RFE (Recursive Feature Elimination)
  - Model building
  - Removing the less relevant variables based on RFE, p-values, and VIFs value.
  - Evaluating the accuracy and other metrics of the model.

- Using the stats model, a detailed model is built.
  - This procedure is repeated 4 times it consists of the following steps:-
  - Columns with a high p-value above the accepted threshold of 0.05 p-value are dropped.
  - Model 4 is coming out to be the final model. As this is stable and has significant p-values within the threshold (p-values less than 0.05) with VIF <5 for all columns, and thus can be used for further analysis.

Model Building

Analysis of Model 4

Generalized Linear Model Regression Results

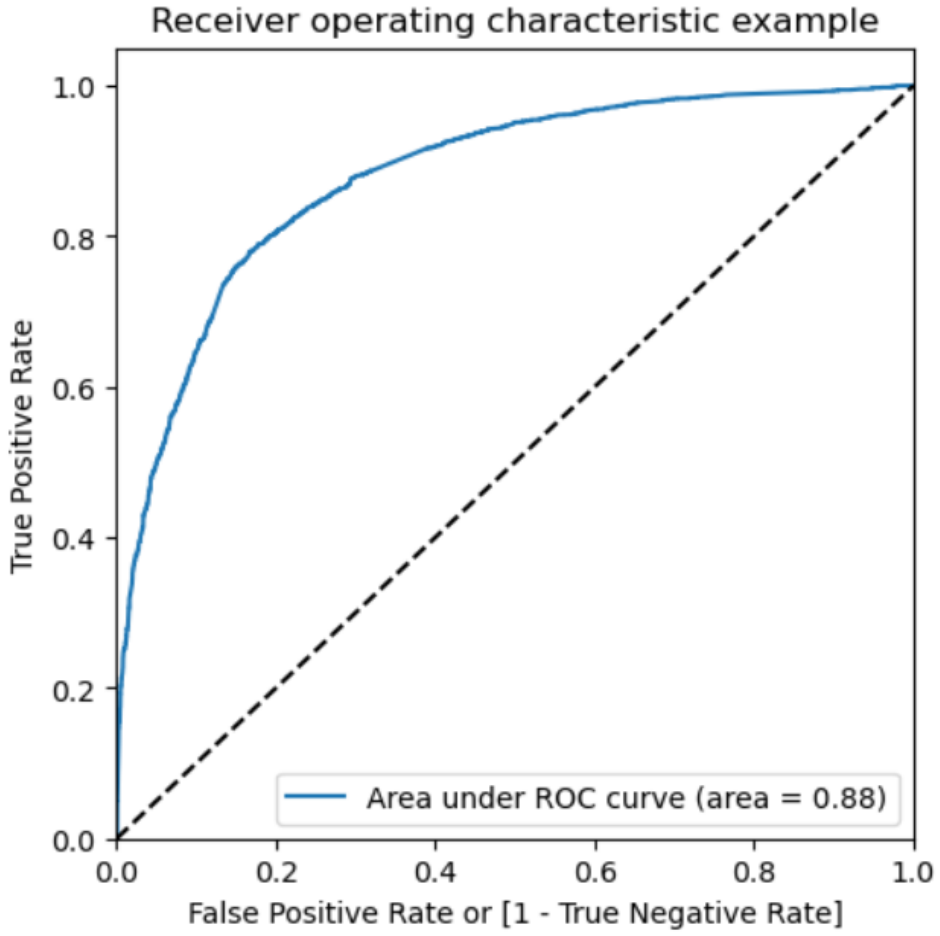
Dep. Variable:	Converted	No. Observations:	6468
Model:	GLM	Df Residuals:	6455
Model Family:	Binomial	Df Model:	12
Link Function:	Logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-2743.1
Date:	Sun, 21 May 2023	Deviance:	5486.1
Time:	18:47:33	Pearson chi2:	8.11e+03
No. Iterations:	7	Pseudo R-squ. (CS):	0.3819
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-1.0236	0.143	-7.145	0.000	-1.304	-0.743
Total Time Spent on Website	1.0498	0.039	27.234	0.000	0.974	1.125
Lead Origin_Landing Page Submission	-1.2590	0.125	-10.037	0.000	-1.505	-1.013
Lead Source_Olark Chat	0.9072	0.118	7.701	0.000	0.676	1.138
Lead Source_Reference	2.9253	0.215	13.615	0.000	2.504	3.346
Lead Source_Welingak Website	5.3887	0.728	7.399	0.000	3.961	6.816
Last Activity_Email Opened	0.9421	0.104	9.022	0.000	0.737	1.147
Last Activity_Olark Chat Conversation	-0.5556	0.187	-2.974	0.003	-0.922	-0.189
Last Activity_Others	1.2531	0.238	5.259	0.000	0.786	1.720
Last Activity_SMS Sent	2.0519	0.107	19.106	0.000	1.841	2.262
Specialization_Hospitality Management	-1.0944	0.323	-3.391	0.001	-1.727	-0.462
Specialization_Others	-1.2033	0.121	-9.950	0.000	-1.440	-0.966
Current_occupation_Working Professional	2.6697	0.190	14.034	0.000	2.297	3.042

	Features	VIF
0	Specialization_Others	2.47
1	Lead Origin_Landing Page Submission	2.45
2	Last Activity_Email Opened	2.36
3	Last Activity_SMS Sent	2.20
4	Lead Source_Olark Chat	2.14
5	Last Activity_Olark Chat Conversation	1.72
6	Lead Source_Reference	1.31
7	Total Time Spent on Website	1.24
8	Current_occupation_Working Professional	1.21
9	Lead Source_Welingak Website	1.08
10	Last Activity_Others	1.08
11	Specialization_Hospitality Management	1.02

- **From the analysis of model 4 we can conclude:-**
- **Model 4 is stable** and has **significant p-values within the threshold** (p-values less than 0.05).
- Hence we will use this model 4 for further analysis.

Model Evaluation

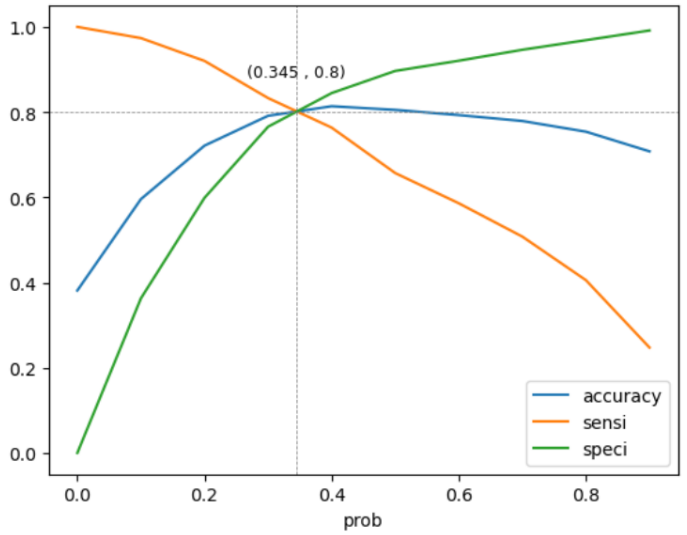


ROC Curve

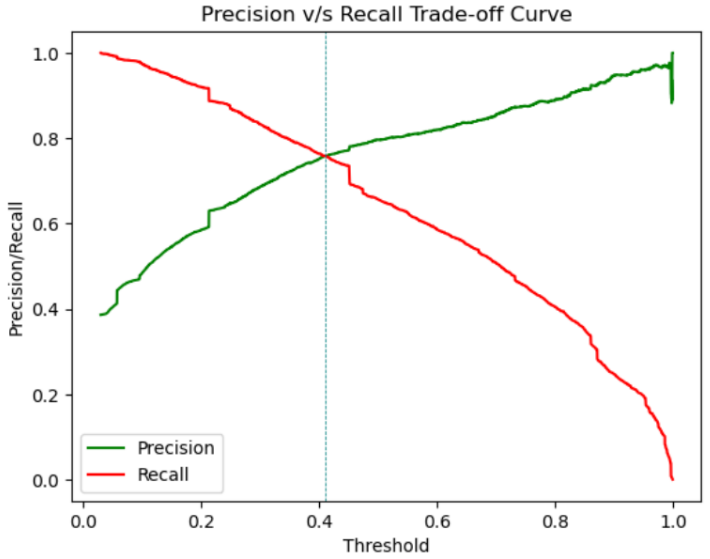
- **Tools used to evaluate the model are:**
  1. Confusion Matrix
  2. Accuracy
  3. Sensitivity and Specificity
  4. Threshold determination using ROC & Finding Optimal cutoff point.
  5. Precision and Recall

- **Analysis made from the graph:**
  1. Increase in sensitivity will be accompanied by a decrease in specificity.
  2. The more accurate the test shows the the curve following the top-left hand border and then the top border of the ROC space.
  3. the less accurate the test shows the curve comes closer to the 45-degree diagonal of the ROC space.
     (Area under the ROC curve is 0.88 out of 1 which indicates a good predictive model)

Model Evaluation



- **Optimal Cut-Off Point:**
  - The graph shows the optimal cutoff point or probability.
  - This analysis is necessary as it identifies the threshold that maintains a balance between sensitivity and specificity.
  - The given graph shows that point **0.345 (approx.)** is the optimal cut-off value.



- **Precision vs Recall Trade-Off:**
  - Comparing the Precision-Recall view with the Specificity - Sensitivity view, we can conclude that the threshold value provides the best balance between these metrics.
  - From the graph, we can conclude that Precision v/s Recall curve achieves balance at an intersection that is the location where we find the optimal threshold value which in this case is **0.41 approx**.

Insights

- **Based on the model evaluation:**
  - As mentioned in the Business Requirements the metrics need to be close nearby 80% and when compared the metrics obtained from both Precision-Recall and sensitivity-specificity the values of those metrics were found to be 75% which is less than that of committed value.

  - So to obtain 80% for the metrics with the sensitivity-specificity cut-off threshold value of 0.345. We will consider a sensitivity-specificity view for our Optimal cut-off for final predictions.

- **Adding Lead Score Feature to Training Dataframe:**
  - As we found that the higher the score the greater the chance that the lead will get converted. (i.e HOT)
  - Whereas a lower score means that the lead is cold and lesser chance to get converted.

- **The test dataset is scaled to make predictions:**
  - Below given is the evaluation of the test dataset :
  - **Accuracy**: 80.34%
  - **Sensitivity**: 79.82% ≈ 80% (approx)
  - **Specificity**: 80.68%

Note: Since the matrices obtained from the test set is very close to the train set, it shows that the model logsm4 is performing consistently.

Conclusions

- **The final Logistic Regression Model has 12 features:**
  Features that are contributing positively to predicting hot leads(leads that can be converted) in the model are:
  1. Lead Source_Welingak Website
  2. Lead Source_Reference
  3. Current_occupation_Working Professional

- From the analysis, we got that the Optimal threshold/cut-off probability point was found to be 0.345.

- Converted probability predicted having a value **greater than 0.345** will be predicted as Converted lead (i.e. Hot lead) while those **smaller than 0.345** will be predicted as not Converted lead (Cold lead).

- The Final model achieved an accuracy of 80.34% on the test data set, which is in line with the study's objectives as mentioned in the problem statement.

Parameters from the Final model

Lead Source_Welingak Website	5.388662
Lead Source_Reference	2.925326
Current_occupation_Working Professional	2.669665
Last Activity_SMS Sent	2.051879
Last Activity_Others	1.253061
Total Time Spent on Website	1.049789
Last Activity_Email Opened	0.942099
Lead Source_Olark Chat	0.907184
Last Activity_Olark Chat Conversation	-0.555605
const	-1.023594
Specialization_Hospitality Management	-1.094445
Specialization_Others	-1.203333
Lead Origin_Landing Page Submission	-1.258954

dtype: float64

Recommendations

In-Order to increase our Lead Conversion Rates following steps are to be taken:-

- We need to focus on the features that are having positive as well as high coefficients to target our marketing strategies.
- Need to engage working professionals with tailored messaging.
- We should optimize communication channels based on lead engagement impact.
- We can spend more budget on Welingak Website as it is the top performer, etc.
- Incentives/discounts for providing a reference that converts to lead, encourage providing more references.
- Develop strategies to attract high-quality leads from top-performing lead sources.
- We can target working professionals as well as they have high conversion rates as well as they are ready to pay as they have a better financial situation to pay higher fees too.

Areas of improvement:-

- Analyse negative coefficients in specialization offerings.
- Review the **Landing page submission** process for areas of improvement.

Thank you