Lead Score Case Study Assignment

Group Members: Jyotishmoi Phukan, Sakthi Sudarsan S

Batch: Upgrad DSC57 | Data Science – June 2023

Lead Score Case Study

Let us understand the problem statement

- There is a company named X Education who provides online courses to the working professionals. They use various strategies to gather customers. They do marketing on various search engines, online streaming ads, etc. by posting their website details and course curriculum.

- When people visits their website or fill-up a form then they get classified as leads. The company even collects leads from past referrals. Once the list of leads is prepared, then the company connects with the sales team to make phone calls or SMS to get the leads converted into paying customers.

- Although in the initial stage, the company is getting many leads but at the end only a few gets converted. It is found that the company's rate of lead conversion is very poor which is somewhat around 30% only. Now the company wants to improve their lead conversion rate and identify the potential leads.

- We will now look at the upcoming slides for the methodologies and techniques applied to solve this problem.

Goals and Objectives

The objective of doing this assignment is that:

- The X Education company needs help in selecting the most potential leads, i.e. the leads that are most likely to get converted into paying customers.

- The company needs a model which determines a lead score for each of the leads such that the customers with higher lead score have a higher conversion chance and the customers with lower lead score have a lower conversion chance.

- The CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80%.

Methodology

✓ At first we have loaded our dataset to the jupyter notebook.

✓ Then after reading the dataset we have cleaned our data using various techniques of Exploratory Data Analysis.

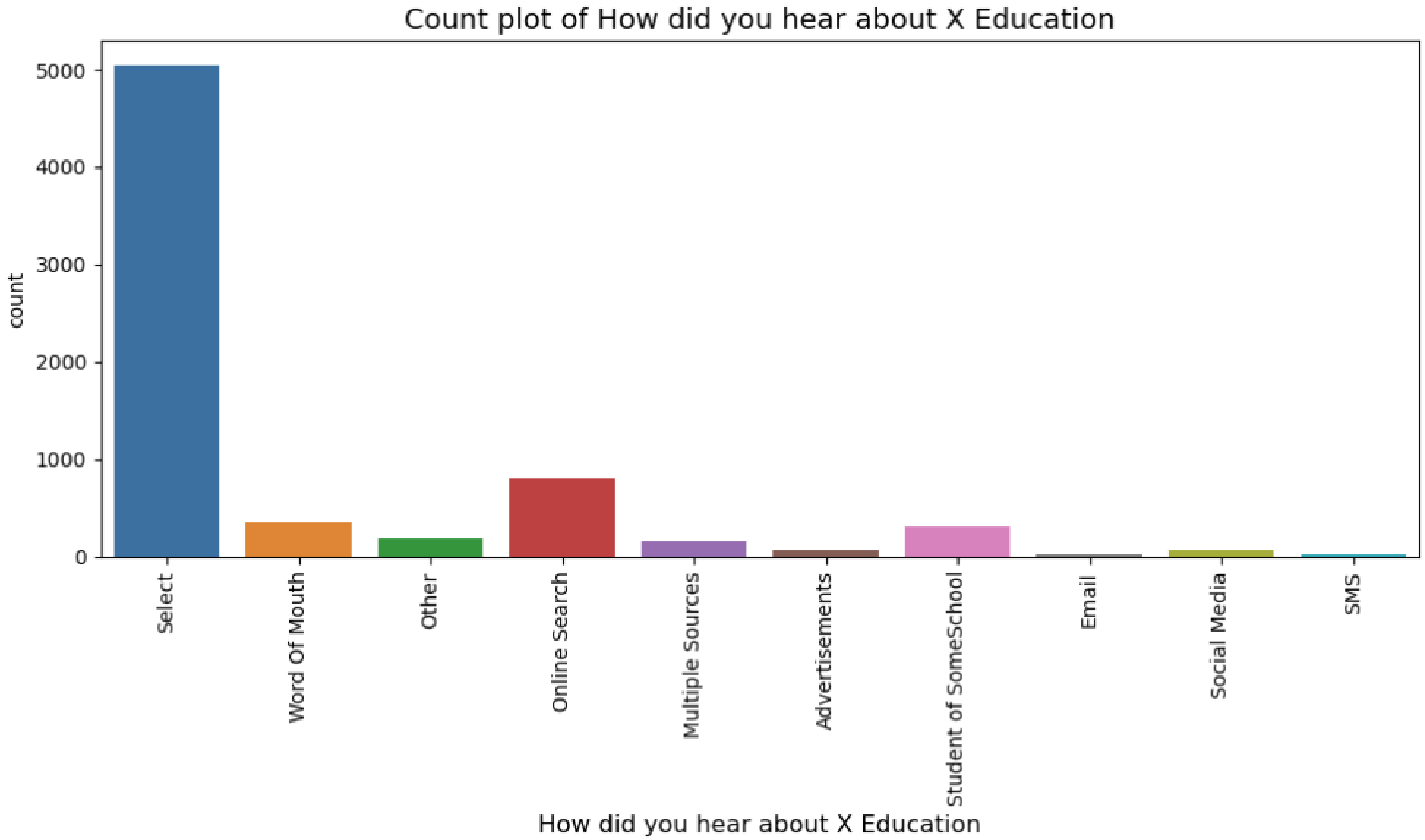
✓ Important variables were selected for drawing meaningful insights and redundant variables were eliminated from our data.

✓ Logistic Regression model was used to make predictions on train and test sets of data.

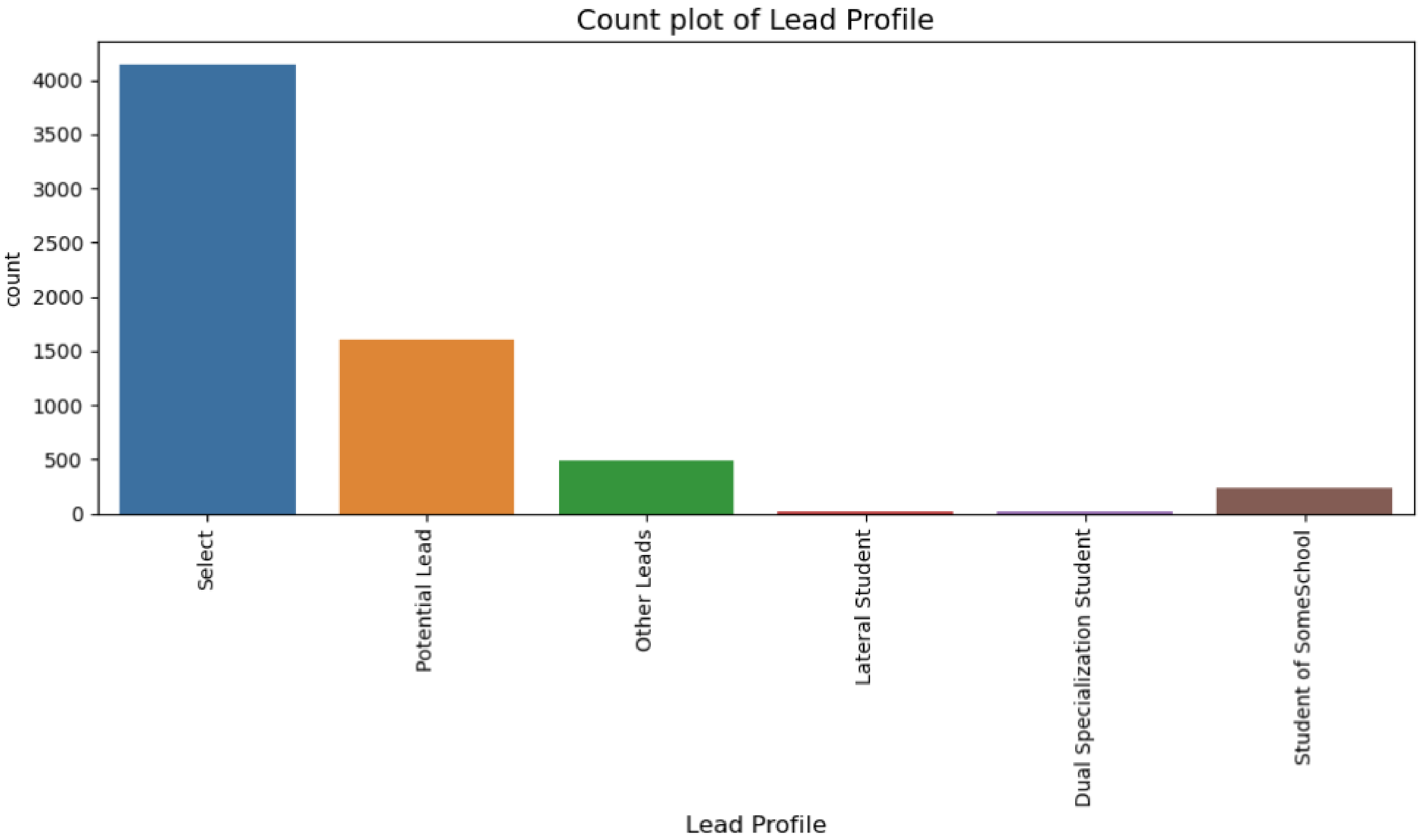
✓ Precision and Recall was used to correctly identify positive instances and measure the accuracy of the positive predictions.

✓ Confusion matrix was used to measure the overall accuracy, sensitivity and specificity of the model.

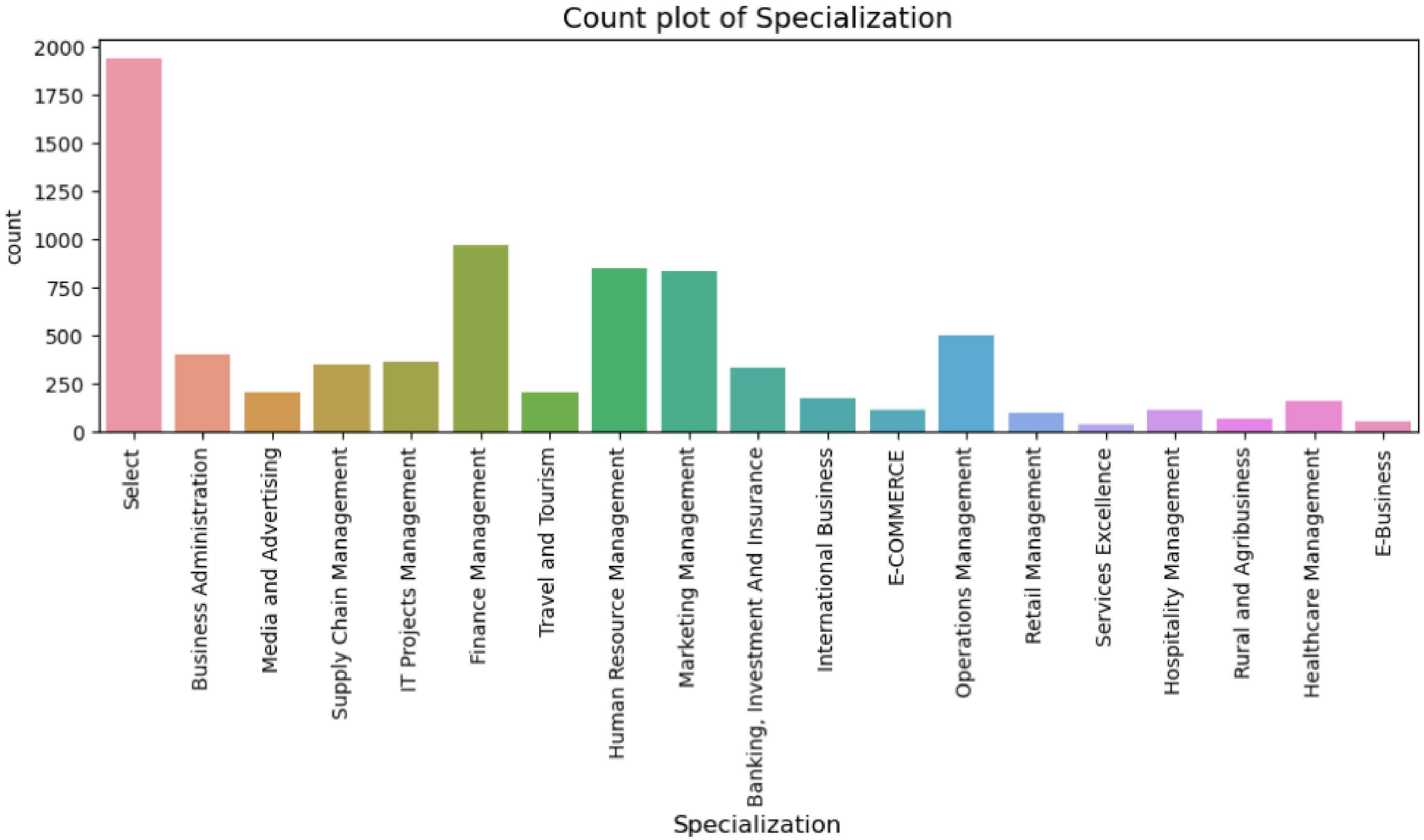
Exploratory Data Analysis



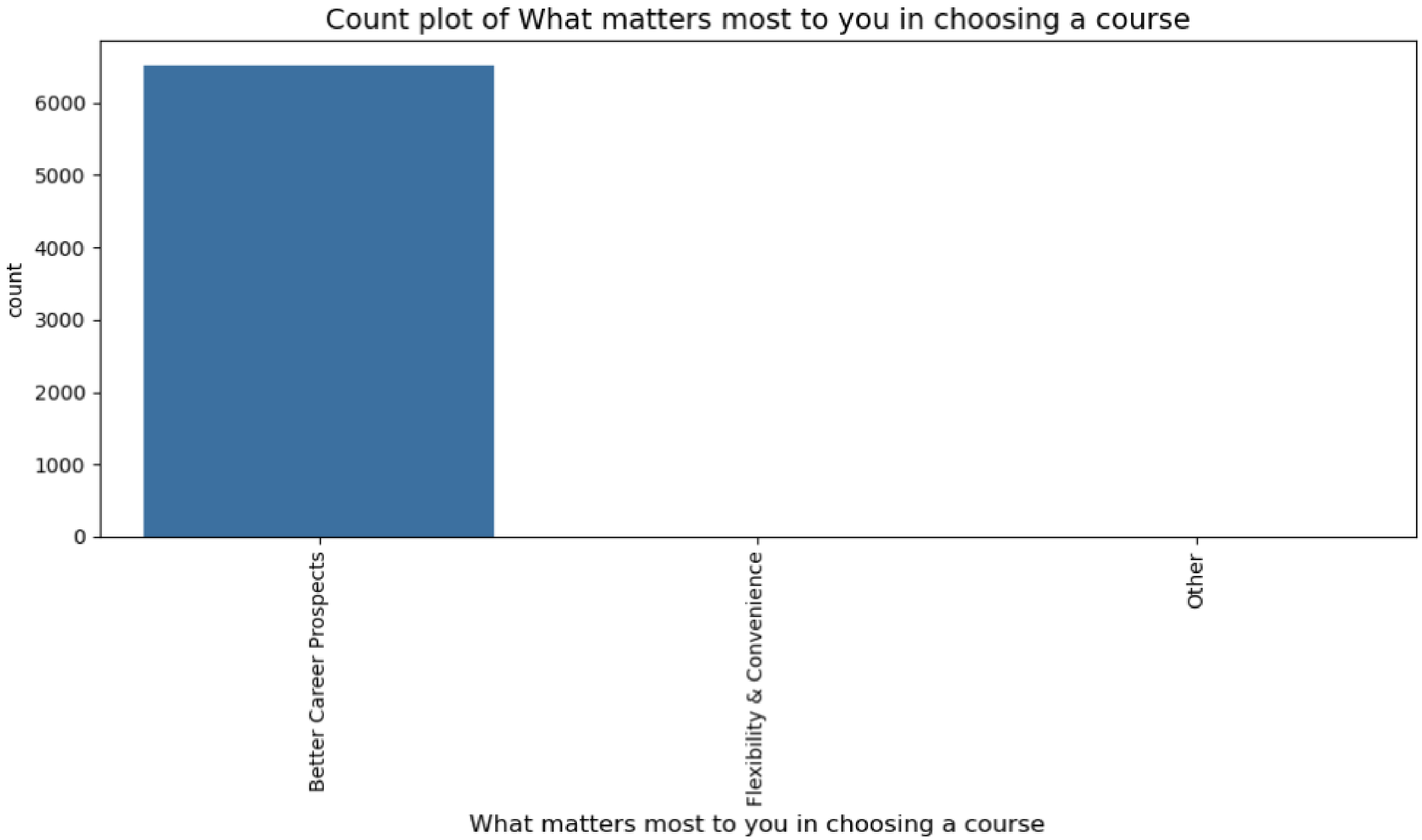
In the above plot we can see that most of the leads got to know about X Education from “Online Search”. Here we are not considering the level ‘Select’ because this is a default value which means that the customer have not selected any option from this column which resulted in default level ‘Select’.



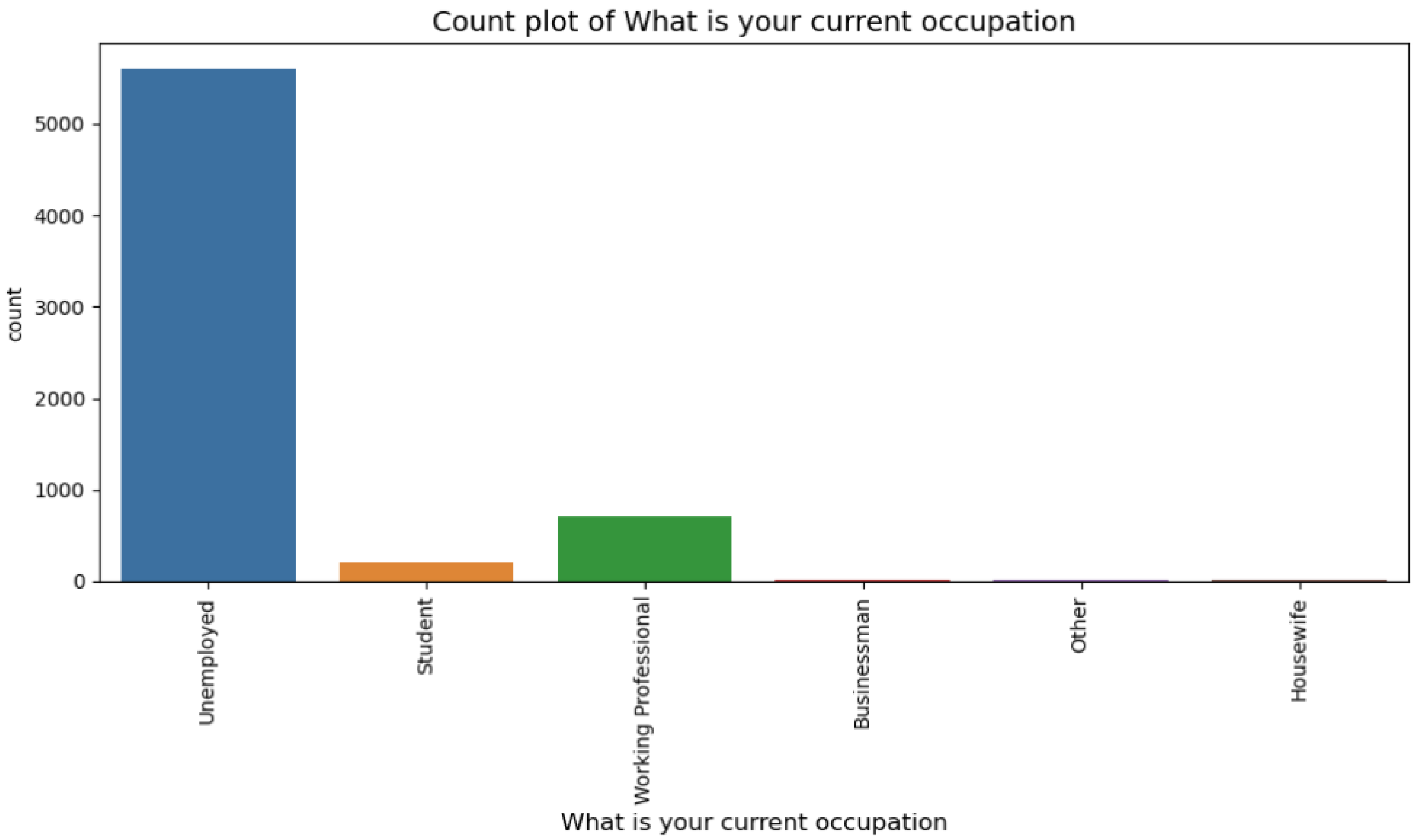
In the above plot we are getting a significant number of Lead Profile from “Potential Lead” Category.



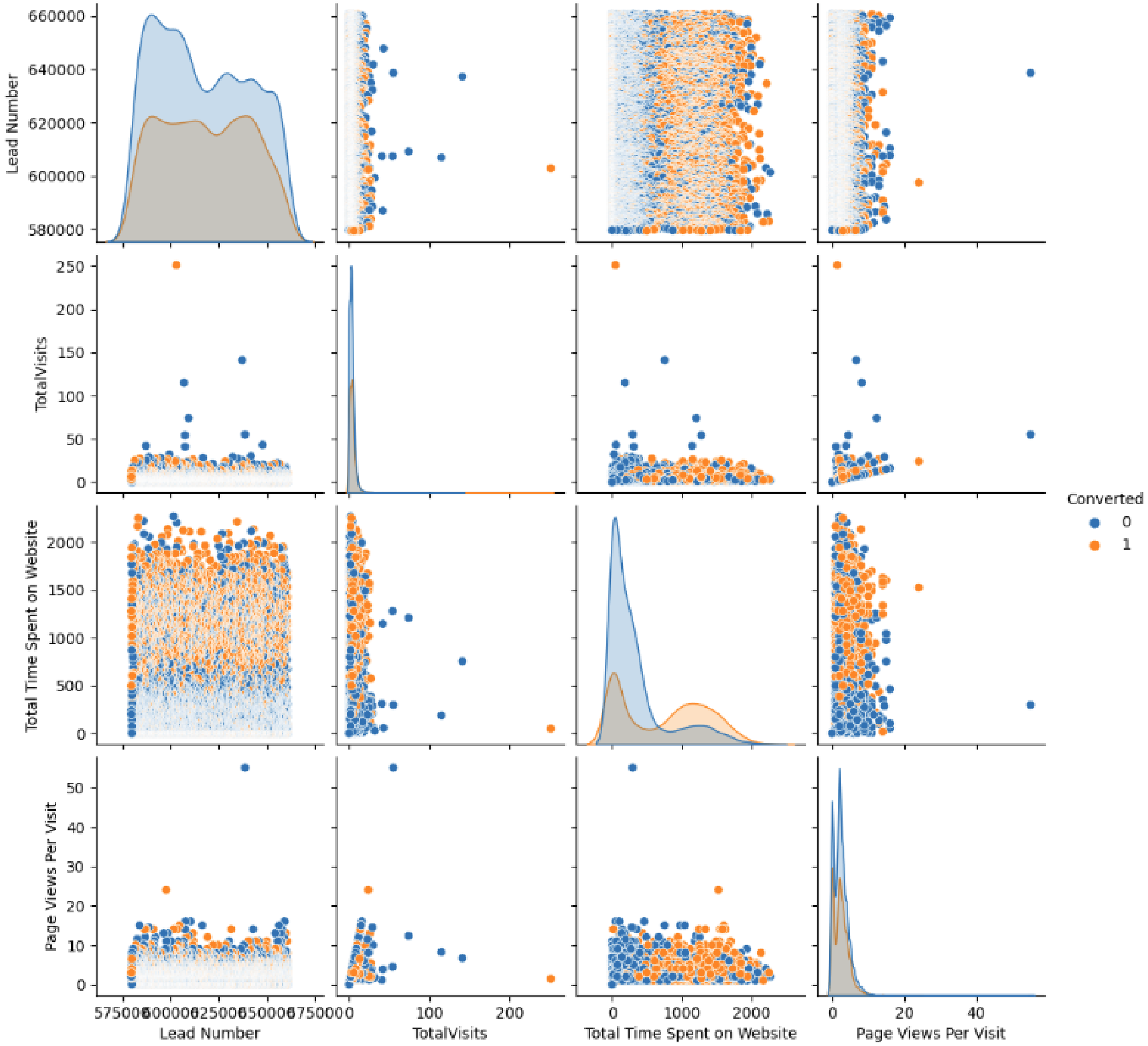
In the above plot we can observe that most the leads are from “Finance Management” specialization.



In the above plot we can observe that most of the leads are from “Better Career Prospects” category which means that leads keep a vision for better career option while selecting a course.

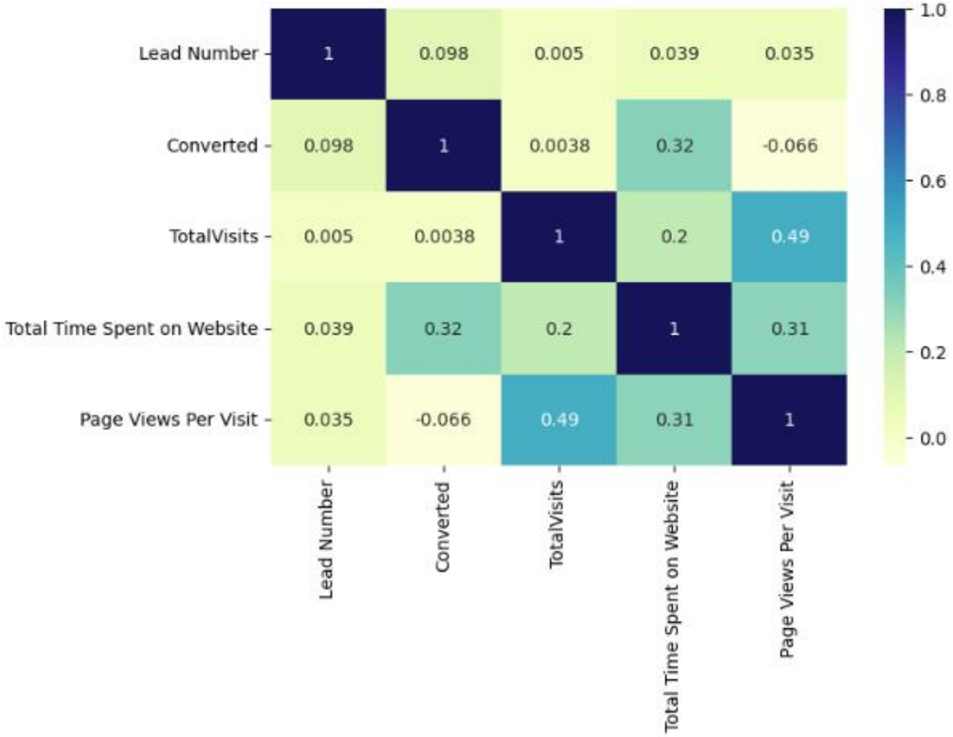


In the above plot we can observe that there are a high number of “Unemployed” leads and they have a low chance of getting converted compared to the “Working Professional” which have a high chance of getting converted.



This is a pair plot which we have prepared during our analysis which gives us a comparison between the Converted and Not Converted leads from the past data.

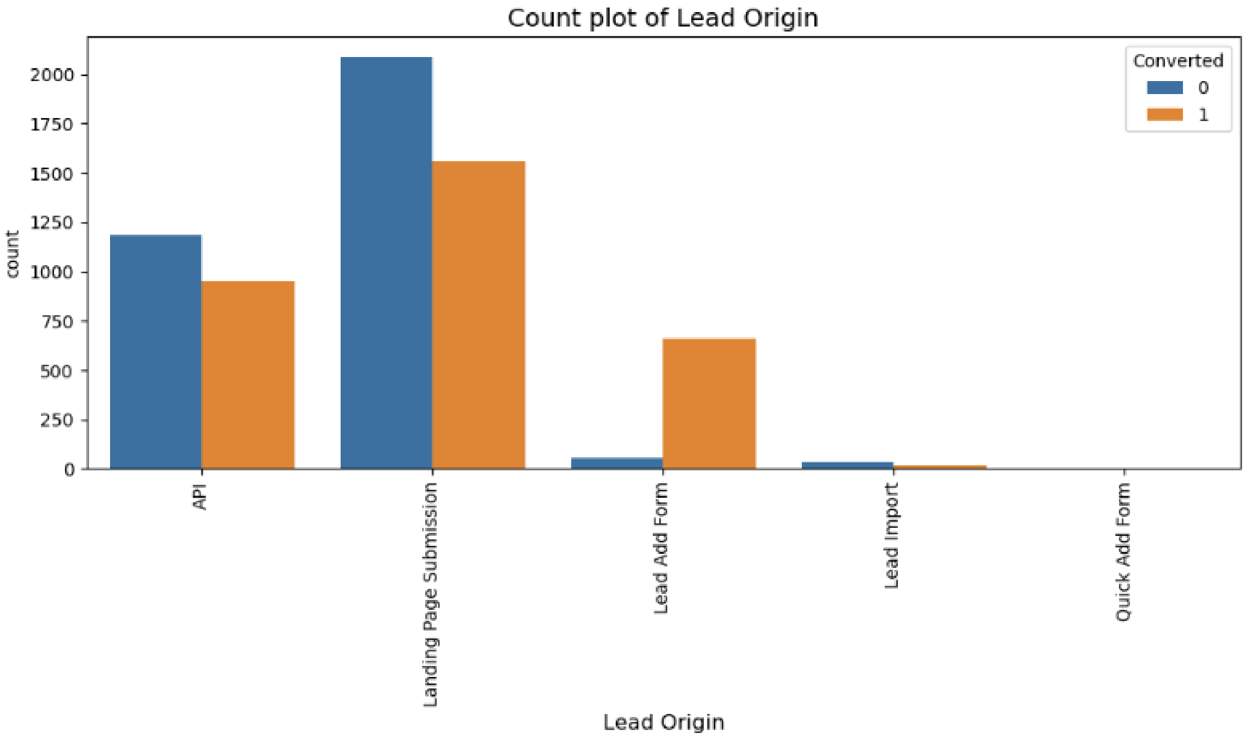
Here 1 means it was converted and 0 means it wasn't converted.



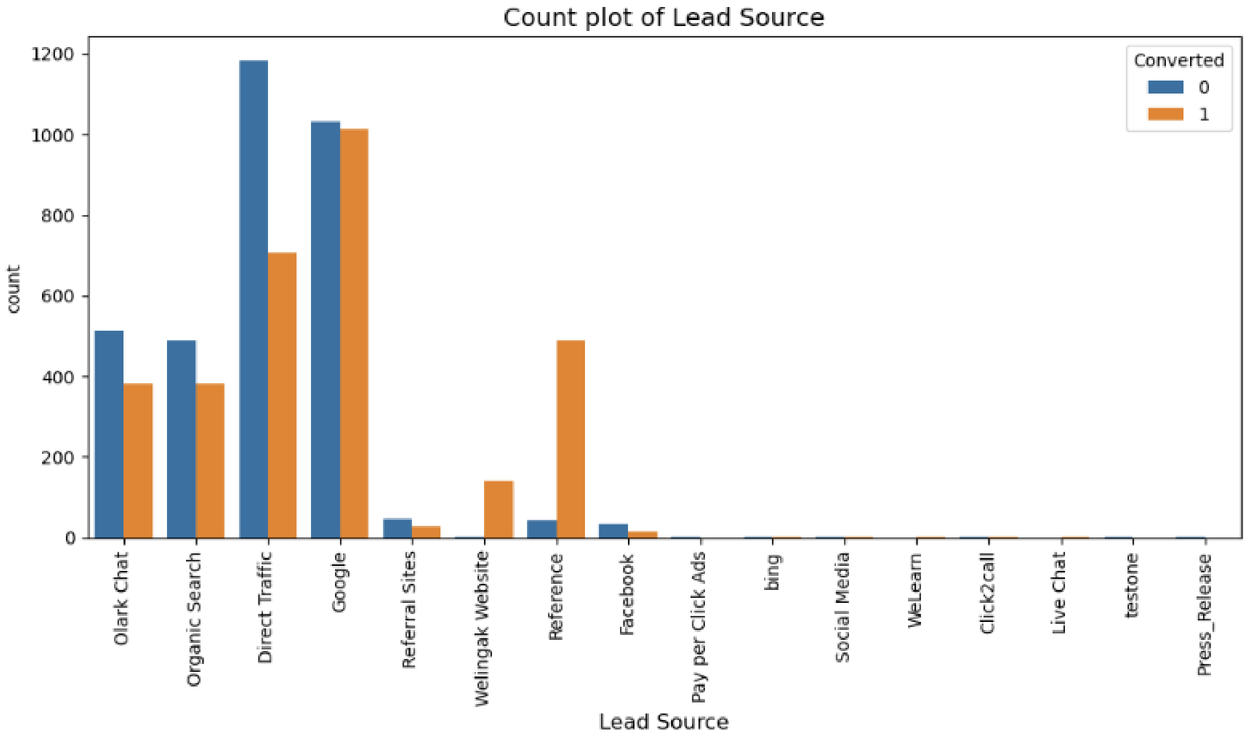
This is a correlation heatmap which is prepared to understand the linear relationship between the dependent and independent numerical variables.

Here we can observe that there is a positive correlation between “Total Time Spent on Website” with the leads getting “Converted”.

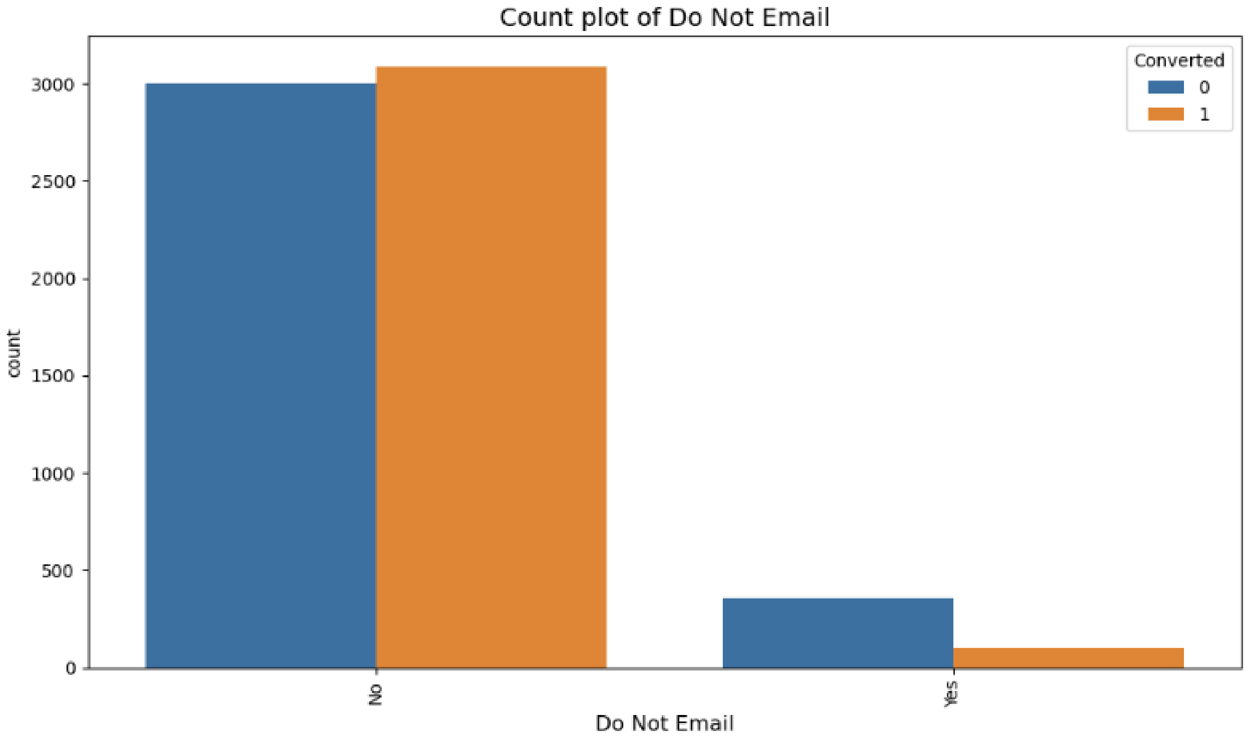
Analysis based on Target variable – “Converted”



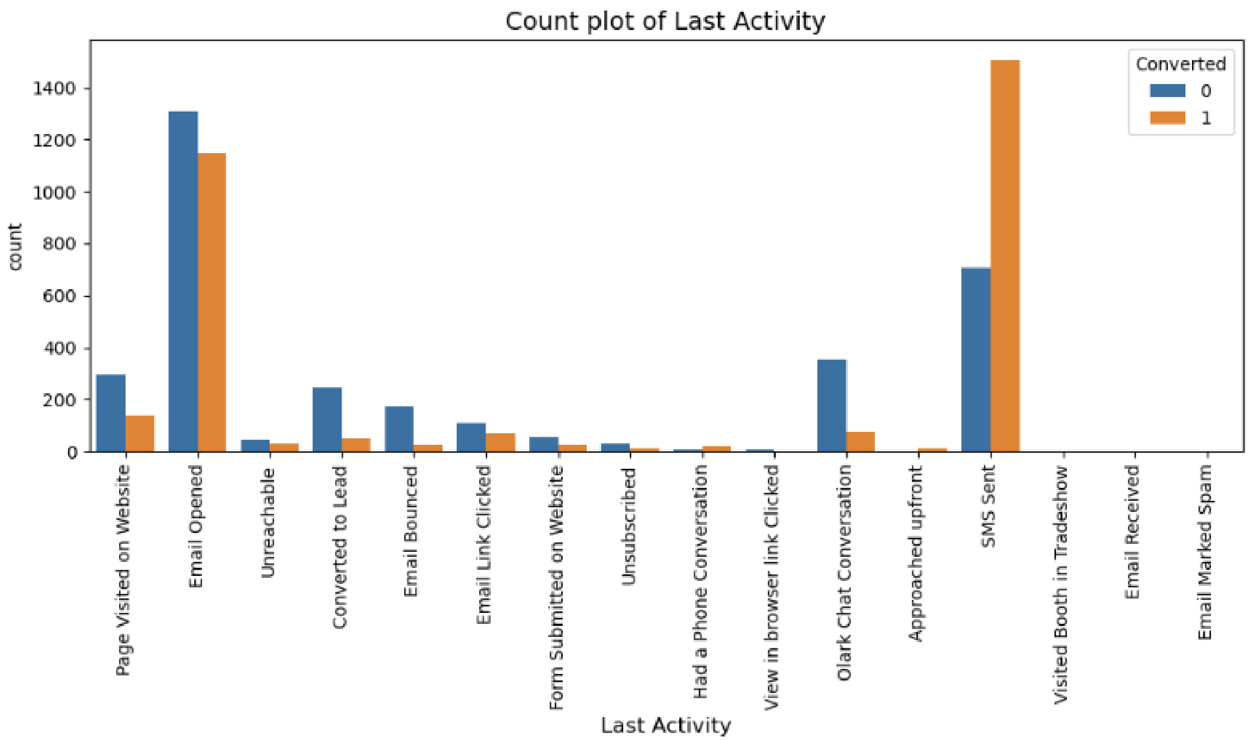
This is a comparison plot between the target variables where we can observe that high number of leads are converted in the “Lead Add Form” category.



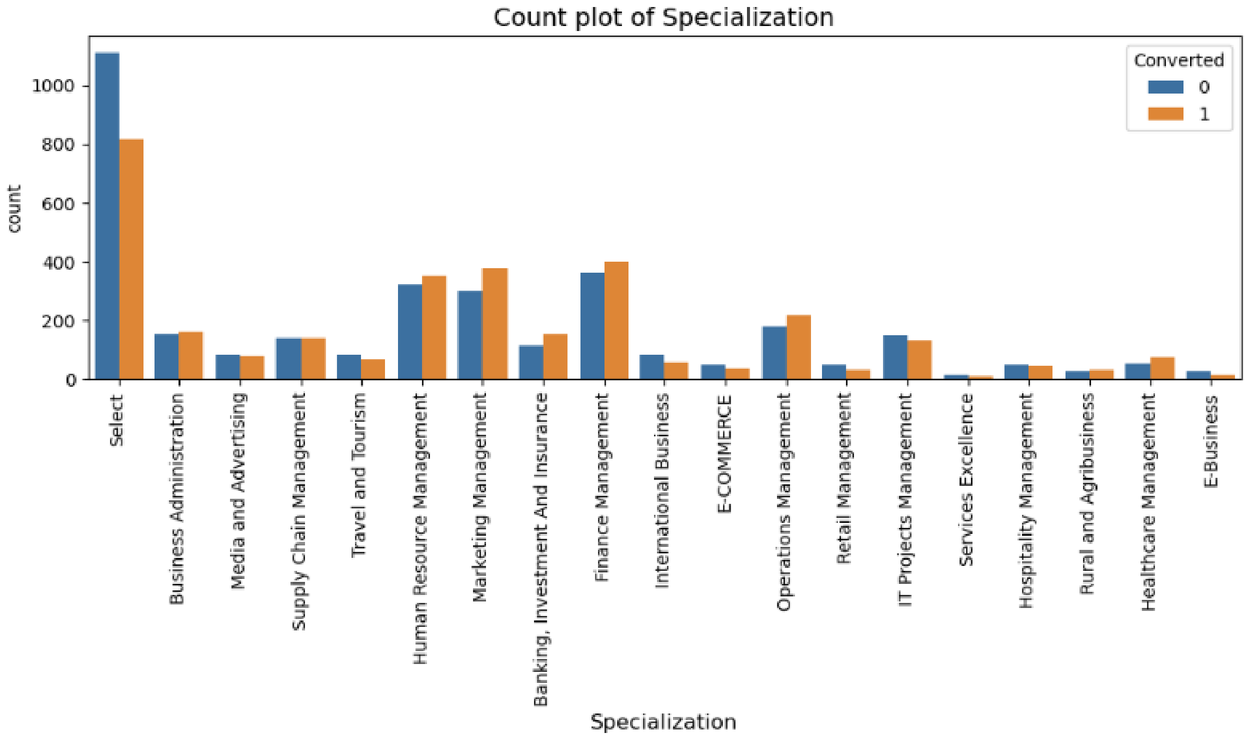
This is a comparison plot between the target variables where we can observe that high number of leads are converted in the "Reference" category.



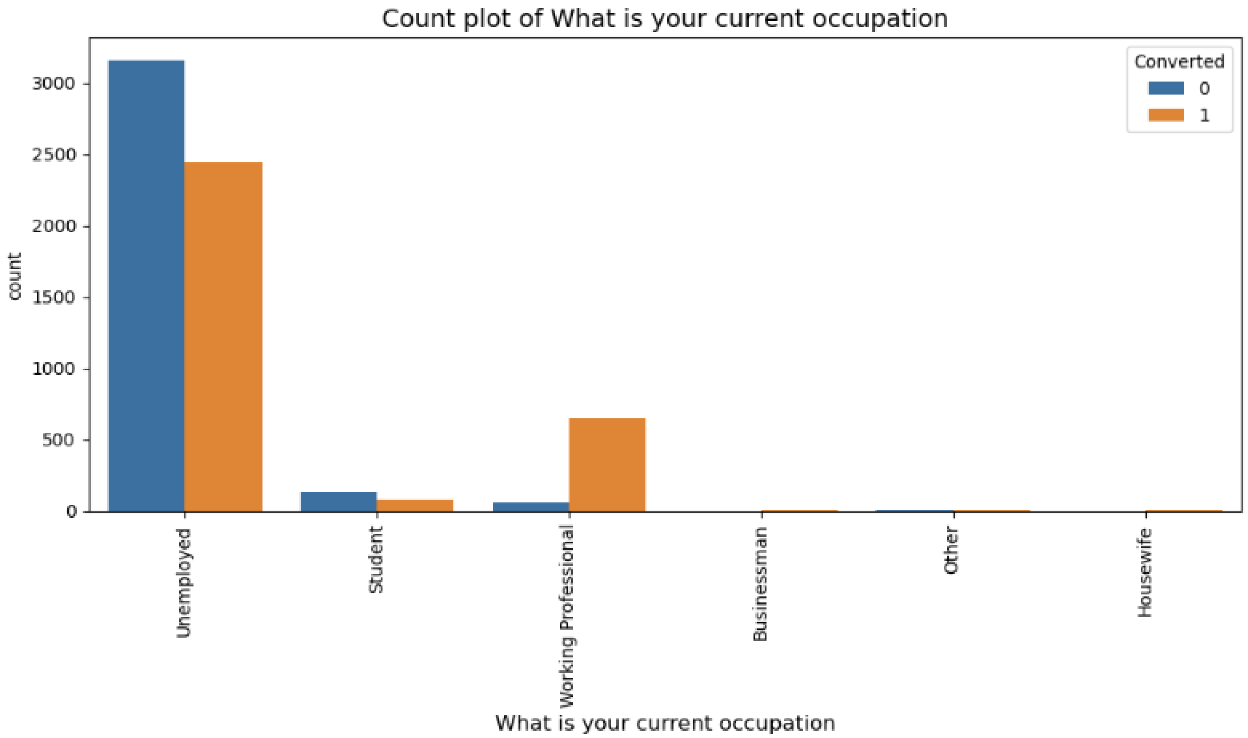
This is a comparison plot between the target variables where we can observe that high number of leads are converted in the “No” category for “Do Not Email”.



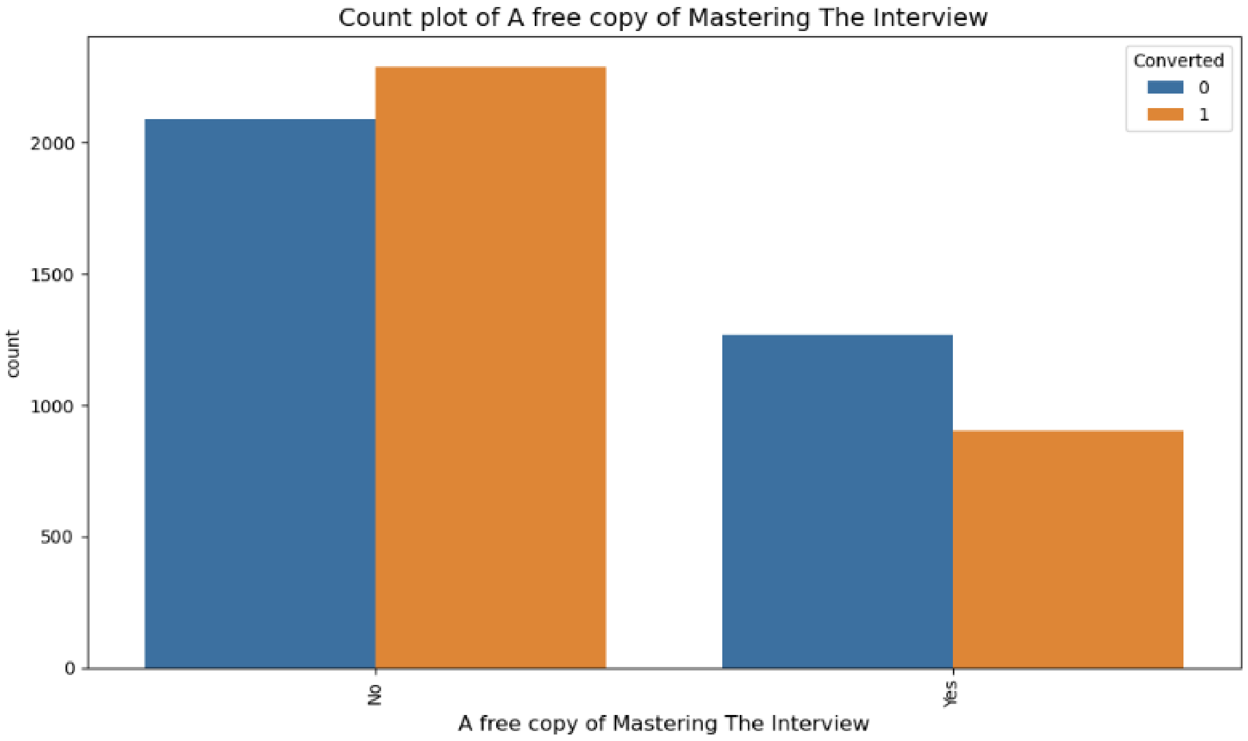
This is a comparison plot between the target variables where we can observe that high number of leads are converted in the “SMS Sent” category and in second position we have the “Email Opened” category.



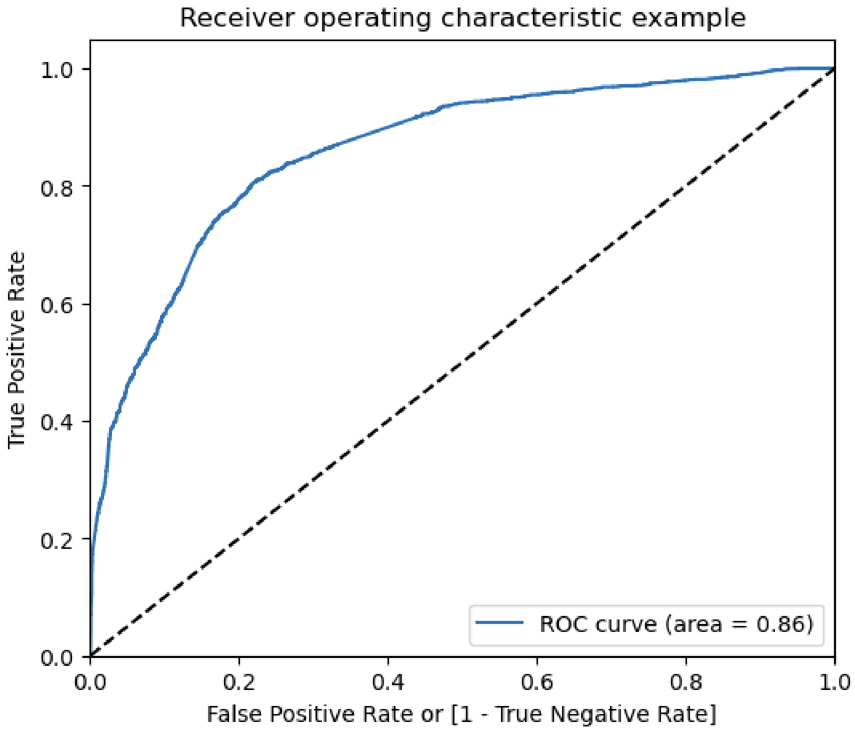
This is a comparison plot between the target variables where we can observe that high number of leads are converted in the “Finance Management” category which means that potential leads are from finance background.



This is a comparison plot between the target variables where we can observe that high number of leads are converted in the “” category.



This is a comparison plot between the target variables where we can observe that high number of leads are converted in the "Lead Add Form" category.

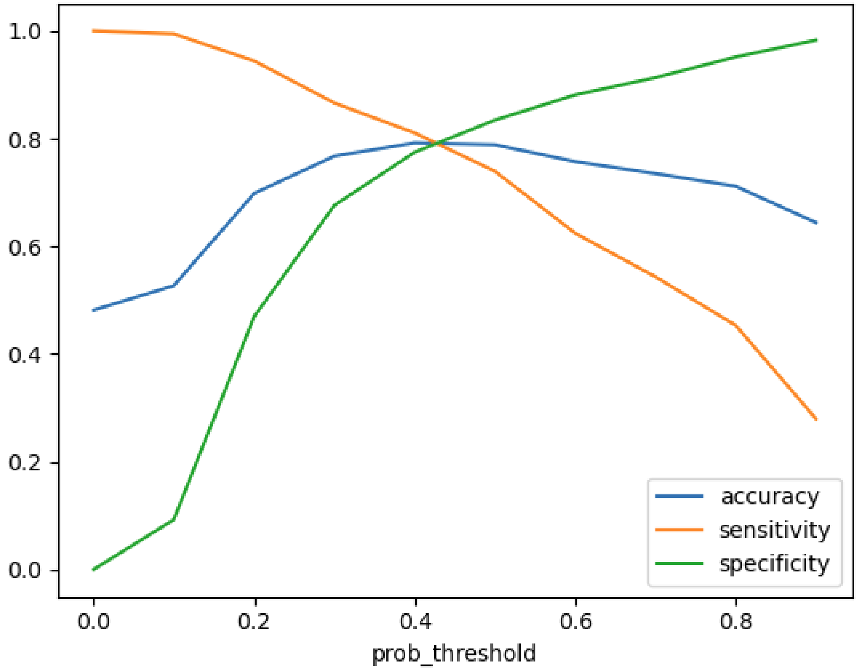


This is the ROC curve that we have prepared and the AUC is clearly visible in the ROC curve. The ROC (Receiver Operating Characteristic) is a graphical representation of the performance of a classification model at various threshold settings.

Whereas, AUC (Area Under the Curve) quantifies the overall performance of a binary classification model based on the ROC curve.

Model Evaluation -Sensitivity and Specificity on Train Data Set

The graph depicts an optimal cut off of 0.42 based on Accuracy, Sensitivity and Specificity



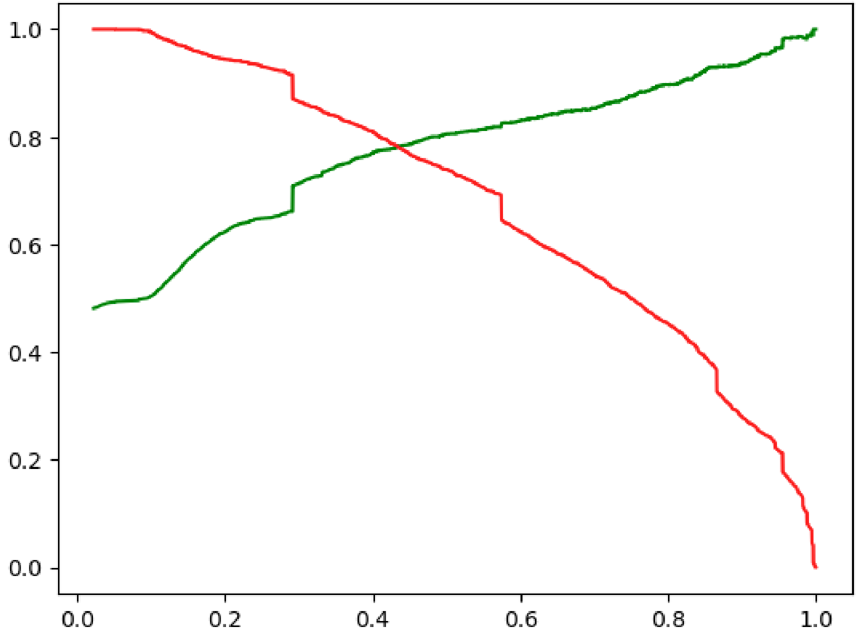
Confusion Matrix

Predicted	Not_converted	Converted
Actual		
Not_converted	1823	489
converted	444	1705

- Accuracy - 79%
- Sensitivity – 79.3%
- Specificity – 78.8%
- False Positive Rate - 21%
- Positive Predictive Value – 77.7%
- Negative Predictive Value – 80.4%

Model Evaluation -Sensitivity and Specificity on Train Data Set

The graph depicts an optimal cut off of 0.42 based on Precision and Recall



Here we can see the trade-off between Precision and Recall

- Precision – 80.5%
- Recall – 73.9%

Model Evaluation –Sensitivity and Specificity on Test Dataset

Confusion Matrix

Predicted	Not_converted	Converted
Actual		
Not_converted	786	210
converted	202	714

- Accuracy – 78.4%
- Sensitivity – 77.9%
- Specificity – 78.9%

Conclusion

- In the logistic regression model we have checked both Sensitivity, Specificity as well as Precision and Recall metrics.

- We have considered the optimal cut-off based on Sensitivity and Specificity for calculating the final prediction.

- Accuracy, Sensitivity and Specificity values of test set are around 78.4%, 77.9% and 78.9% which are approximately closer to the respective values calculated using trained set.

- The top three variables that contribute for successful lead conversion in the model are:

a. Total time spent on website
b. Lead Add Form from Lead Origin
c. Had a Phone Conversation from Last Notable Activity.

- Hence overall this model seems to be good.