LEAD SCORING CASE STUDY (LOGISTIC REGRESSION)

BY

KOMAL NALAWADE AND PRABHJOT SINGH

LEAD SCORING CASE STUDY

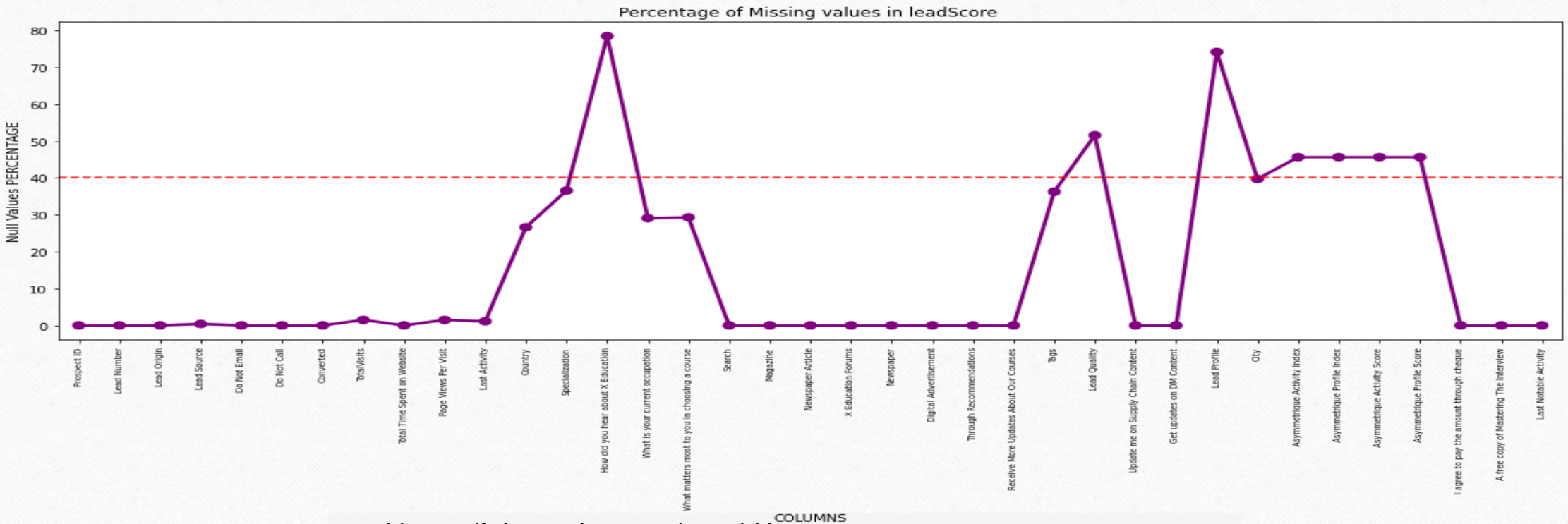
Table of contents

Business Objective

Problem Statement

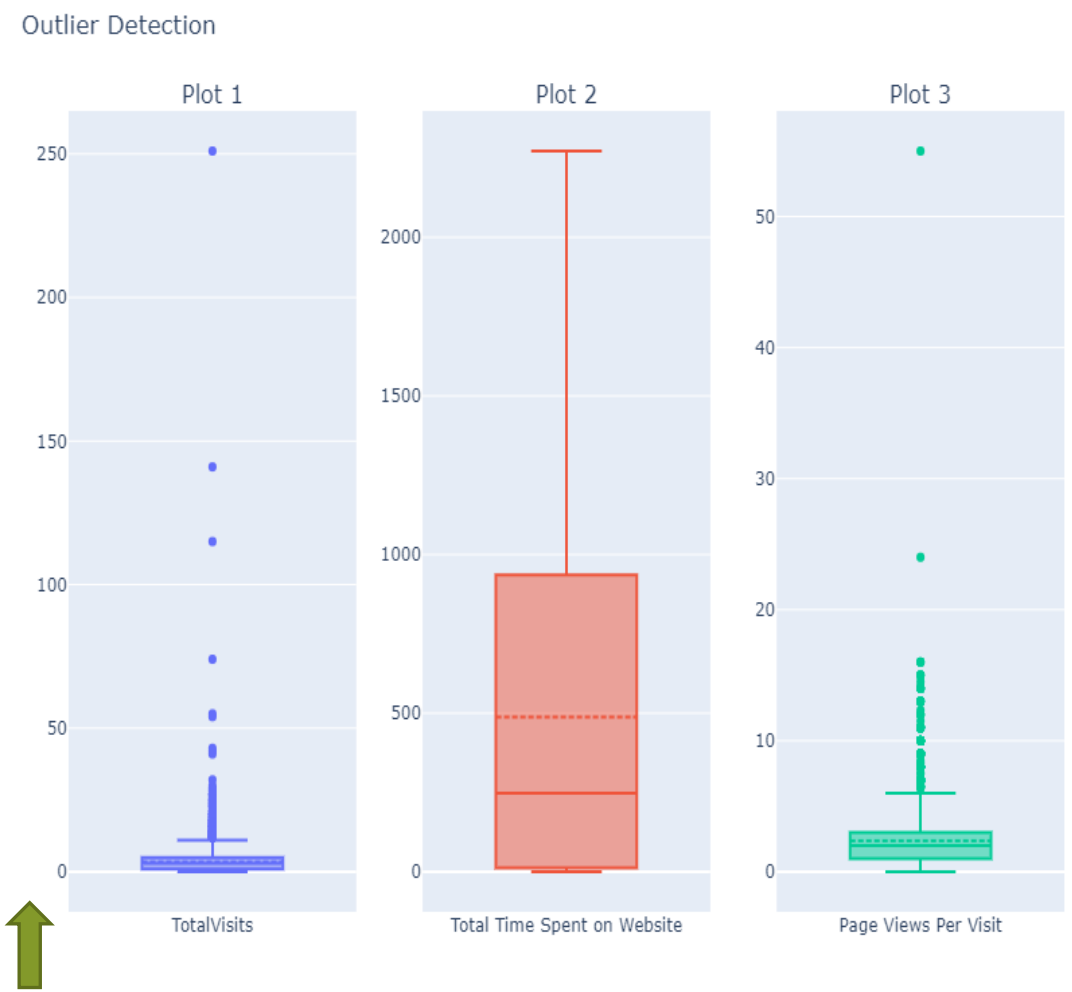
An education company named X Education sells online courses to industry professionals. On any given day, many professionals who are interested in the courses land on their website and browse for courses.
The company markets its courses on several websites and search engines like Google. Once these people land on the website, they might browse the courses or fill up a form for the course or watch some videos. When these people fill up a form providing their email address or phone number, they are classified to be a lead. Moreover, the company also gets leads through past referrals. Once these leads are acquired, employees from the sales team start making calls, writing emails, etc. Through this process, some of the leads get converted while most do not. The typical **lead conversion rate at X education is around 30%.**
Now, although X Education gets a lot of leads, its lead conversion rate is very poor. To make this process more efficient, the **company wishes to identify the most potential leads, also known as 'Hot Leads'**. If they successfully identify this set of leads, the lead conversion rate should go up as the sales team will now be focusing more on communicating with the potential leads rather than making calls to everyone.

Null Value Calculation



COLUMNS	
How did you hear about X Education	78.46
Lead Quality	51.59
Lead Profile	74.19
Asymmetrique Activity Index	45.65
Asymmetrique Profile Index	45.65
Asymmetrique Activity Score	45.65
Asymmetrique Profile Score	45.65

Outlier Detection



Insight:

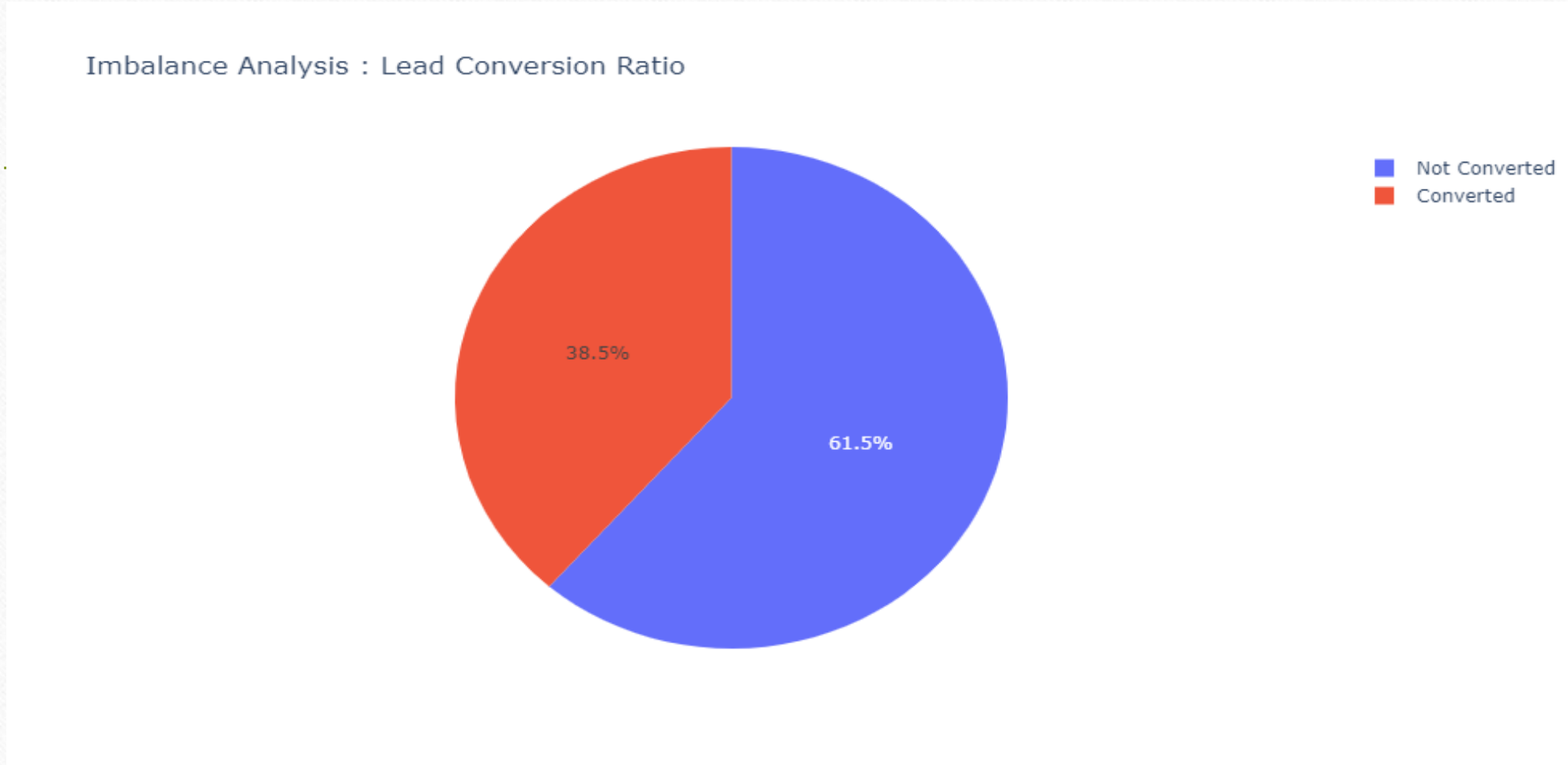
•From the BoxPlot, TotalVisits and Page Views Per Visit show anomalies.

1. But there is single record for TotalVisits>250 and can not be considered as outlier as the record shows **converted as true** so its important and can not be dropped. Hence capping them is suitable.

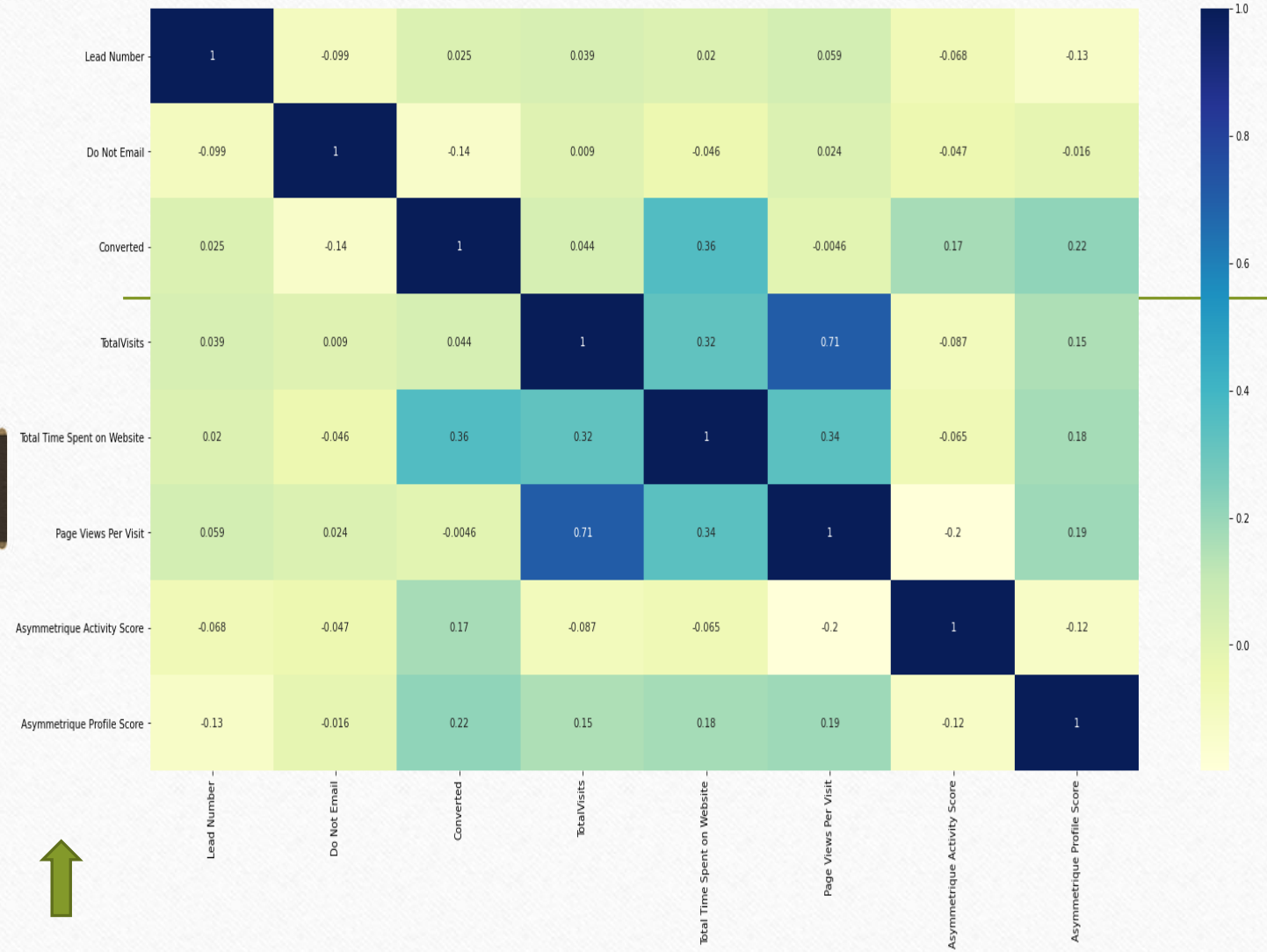
2. There is a single record for Page Views Per Visit >50 and capping at 0.99 percentile need to be done

Data Analysis Imbalance Analysis

Ratios of imbalance in percentage with respect to Not converted and Converted(Hot Leads) data are: 61.5 and 38.5



Numerical Variables' Analysis



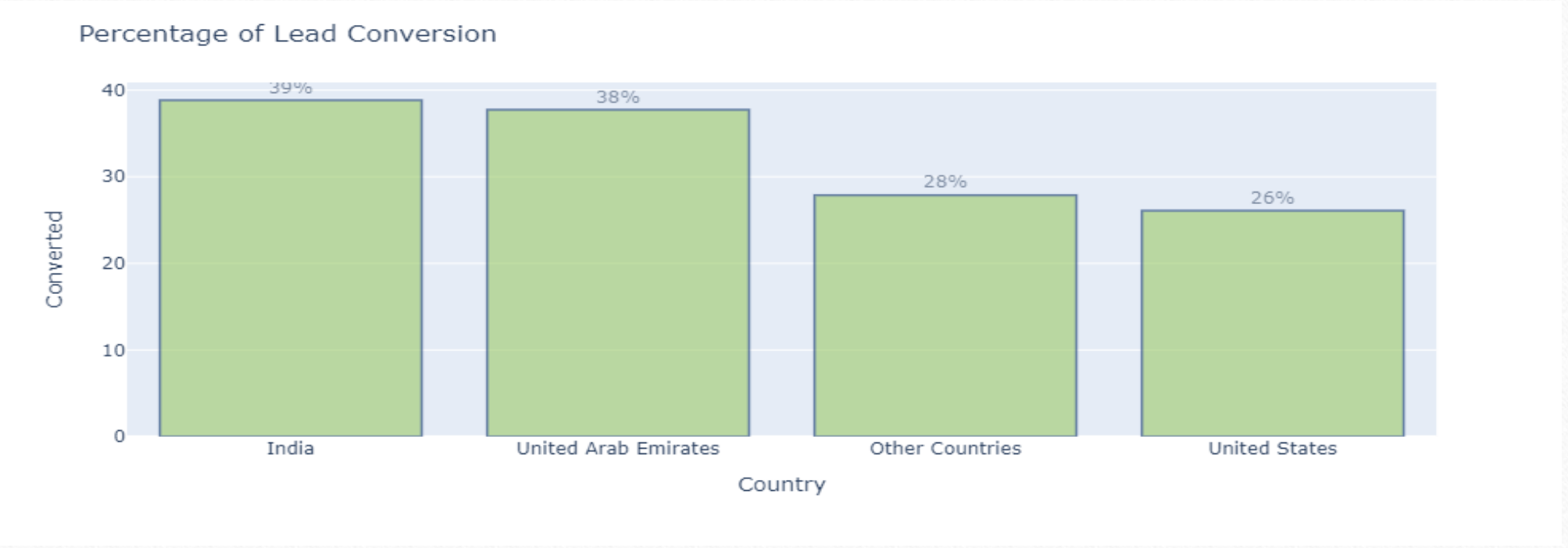
Inferences:

- Page Views Per Visit and total Visits have high positive correlation
- Total time spent on website shows positive collinearity with converted

Categorical Variables' Analysis

Inferences:

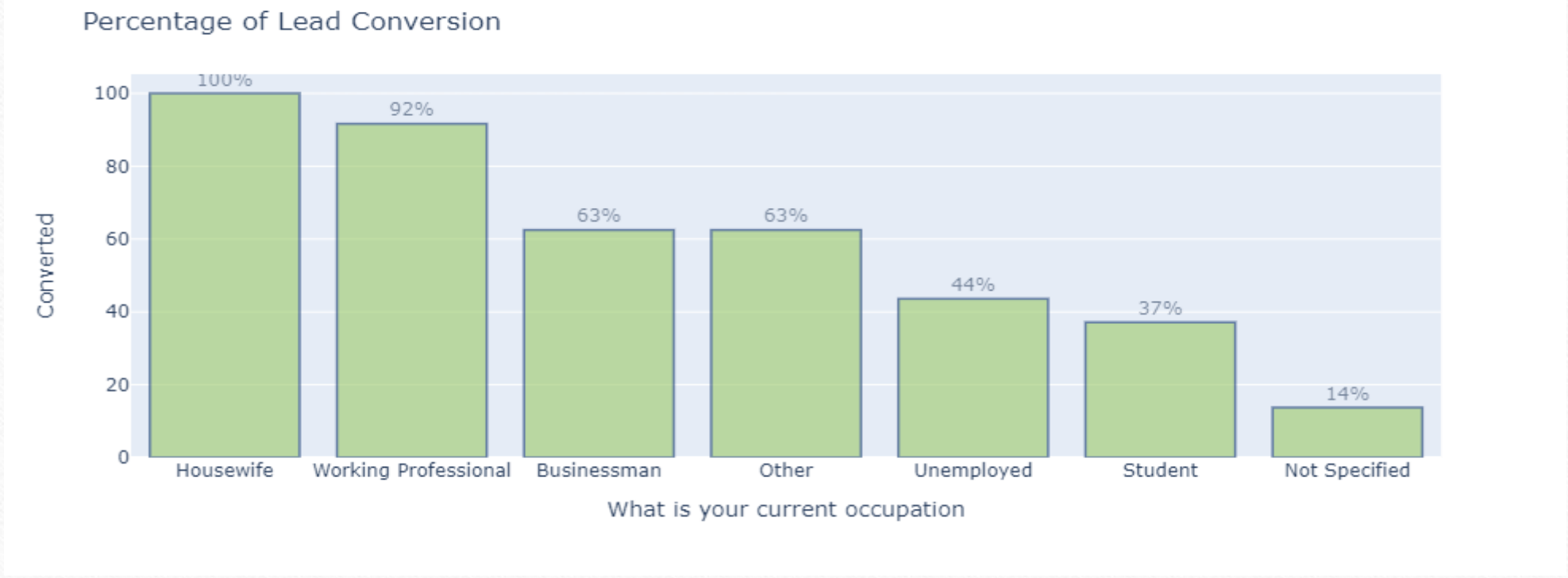
From graph, it can be said that Country have no much influence on lead conversion



Categorical Variables' Analysis

Inferences:

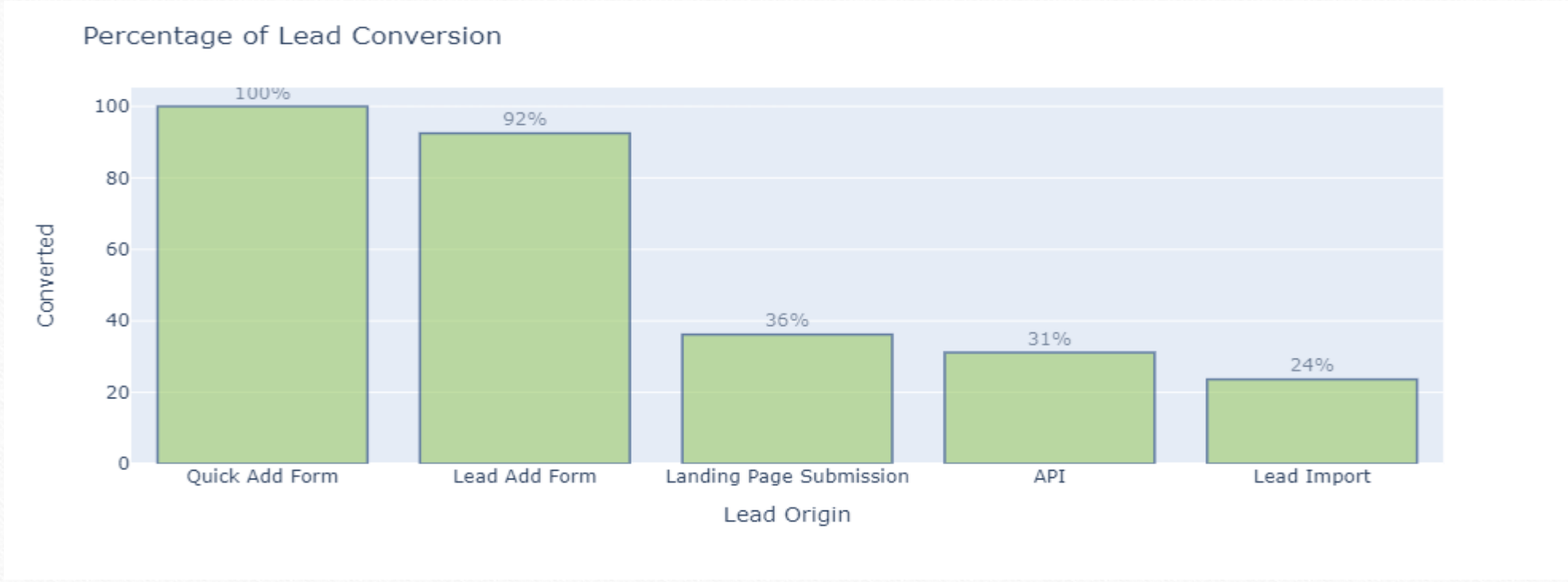
From graph, Working Professionals are likely to become 'Hot Leads'



Categorical Variables' Analysis

Inferences:

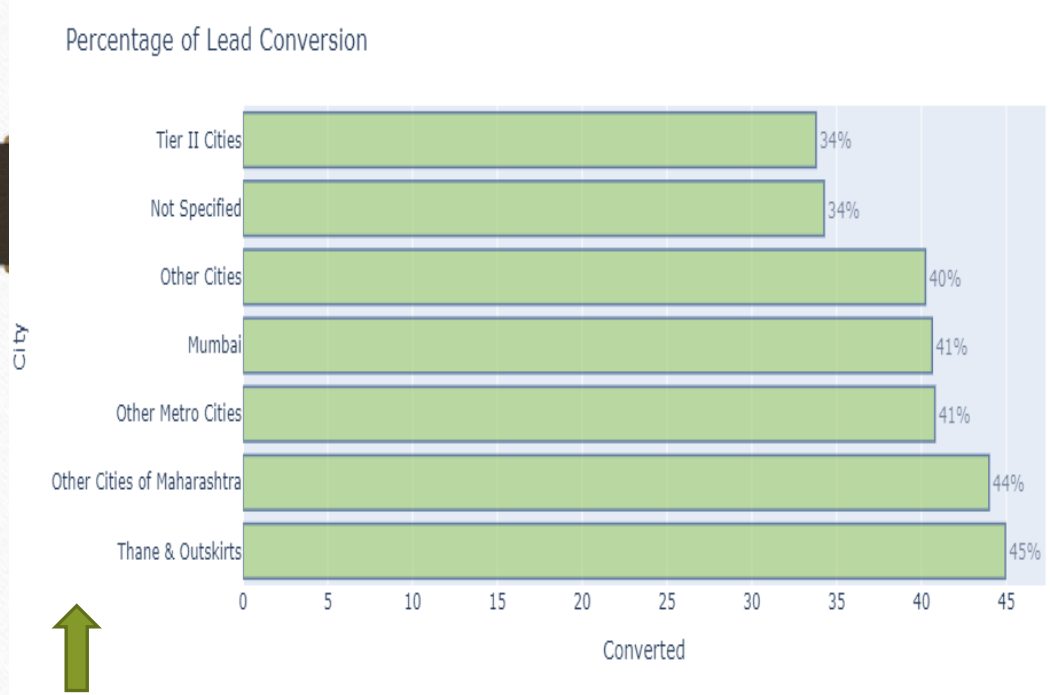
From graph, Quick Add Form followed by lead Add Form are most converted in lead Origin

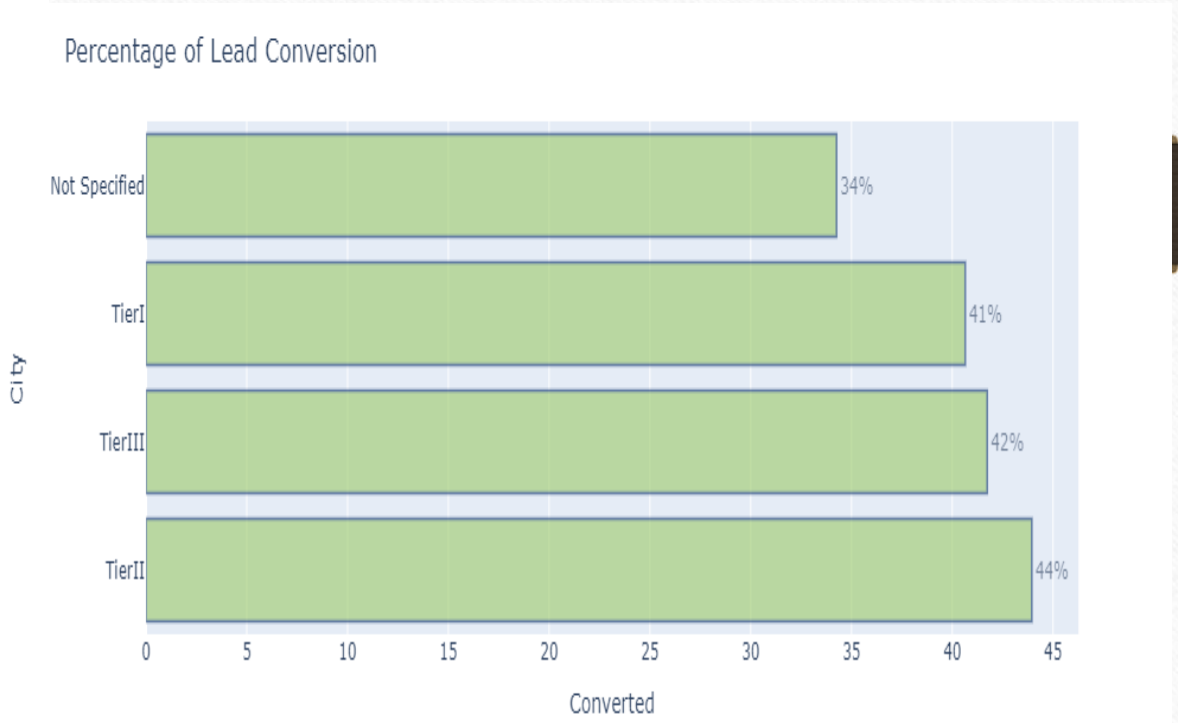


Categorical Variables' Analysis

Inferences:

From graph, Most levels in the City are equally likely to get converted ,has no much influence on Lead Conversion

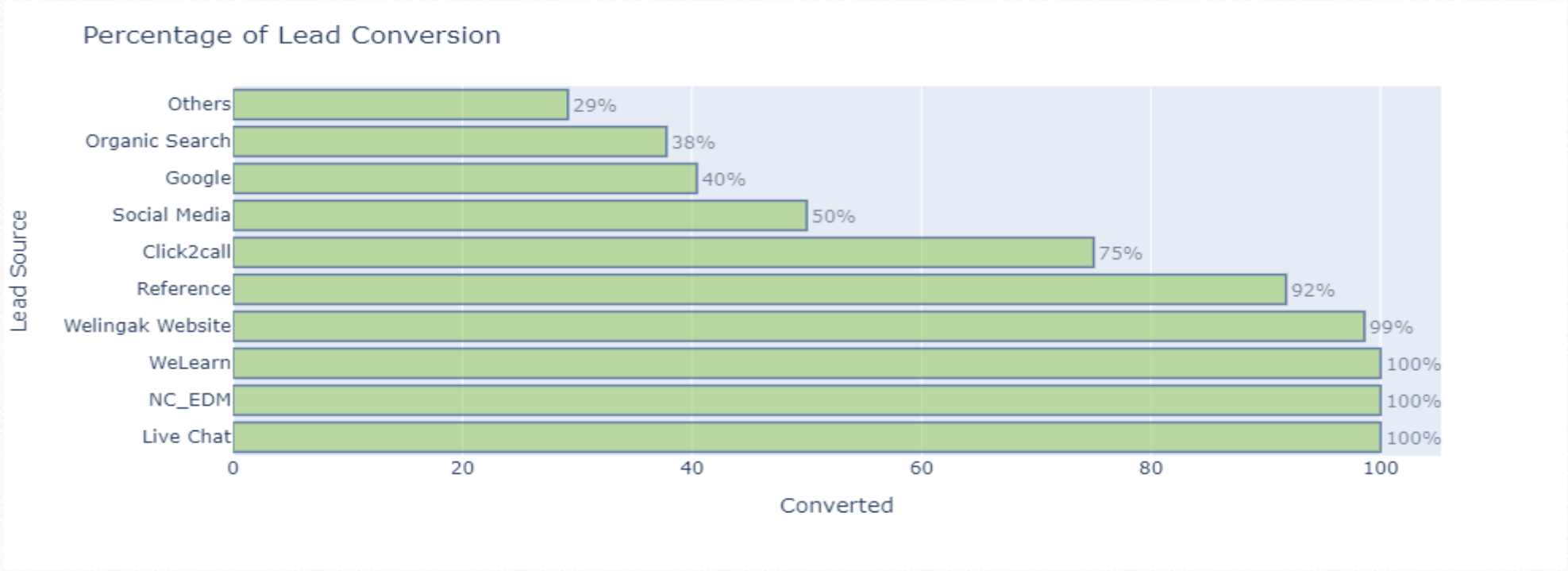




Categorical Variables' Analysis

Inferences:

From graph, Live Chat, NC_EDM ,WeLearn ,Welingak Website and Reference seems to influence the lead conversion

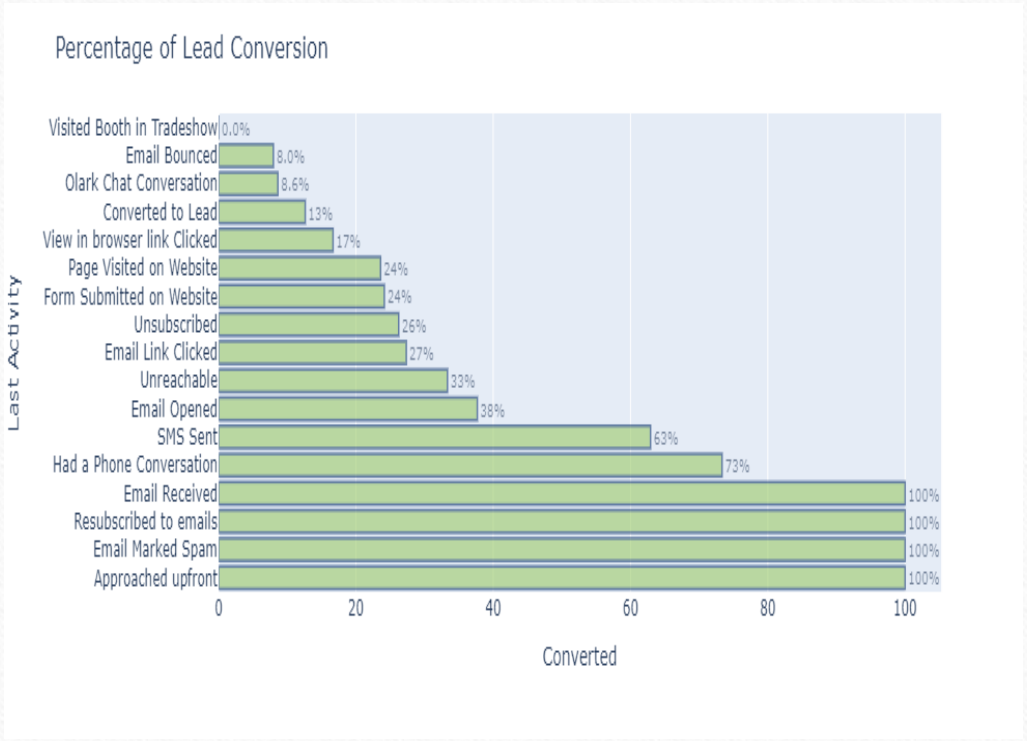


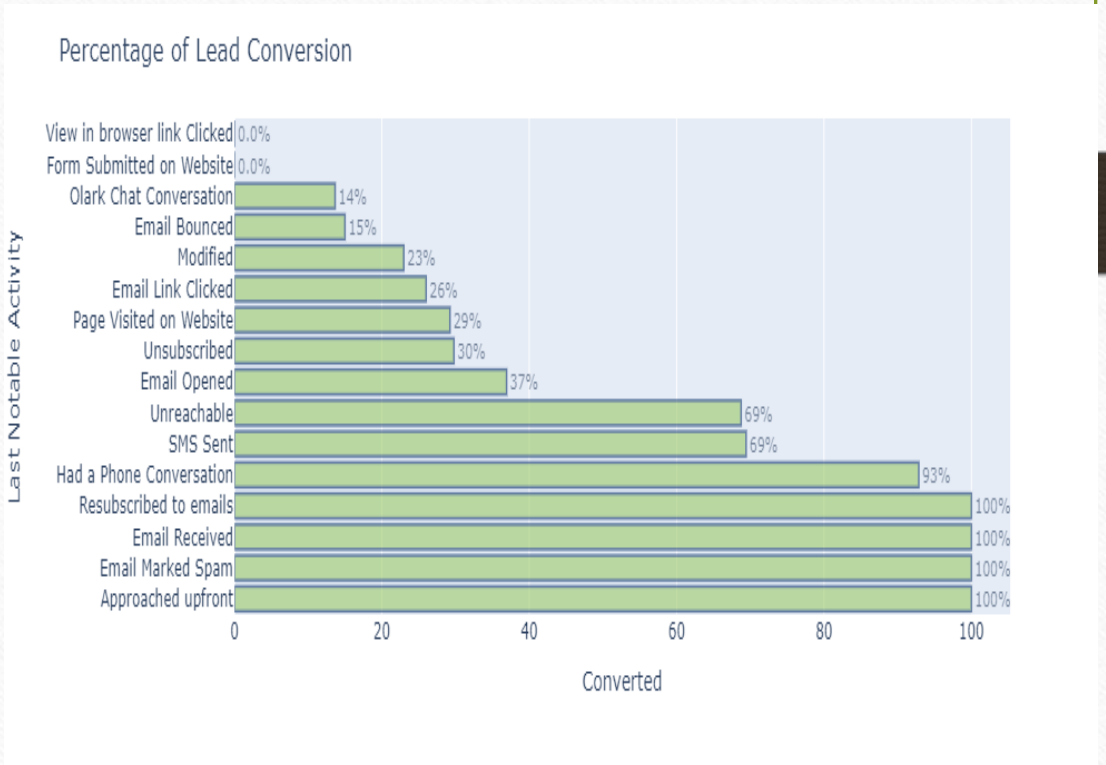
Categorical Variables' Analysis

Inferences:

From graphs, Last Notable Activity and last activity seems important variable in potential Lead conversion as four different activities show nearly 100% lead conversion.

- But Had a phone conversation shows good lead conversion in both

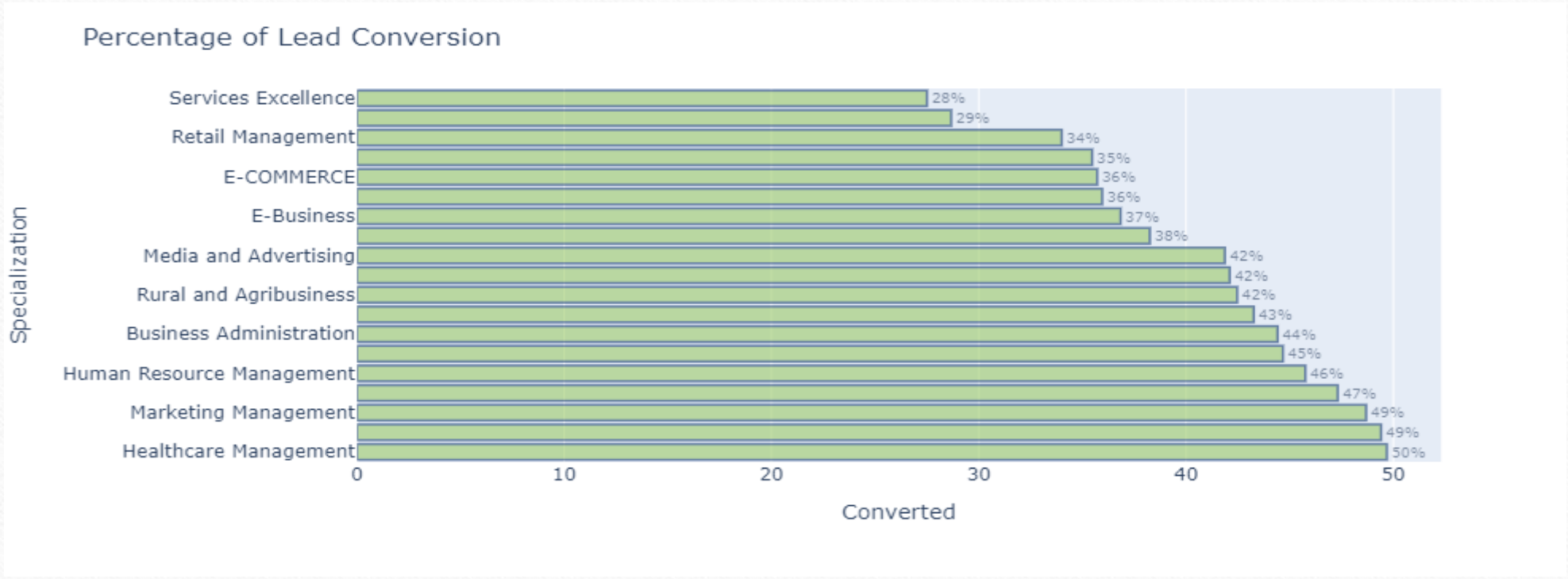




Categorical Variables' Analysis

Inferences:

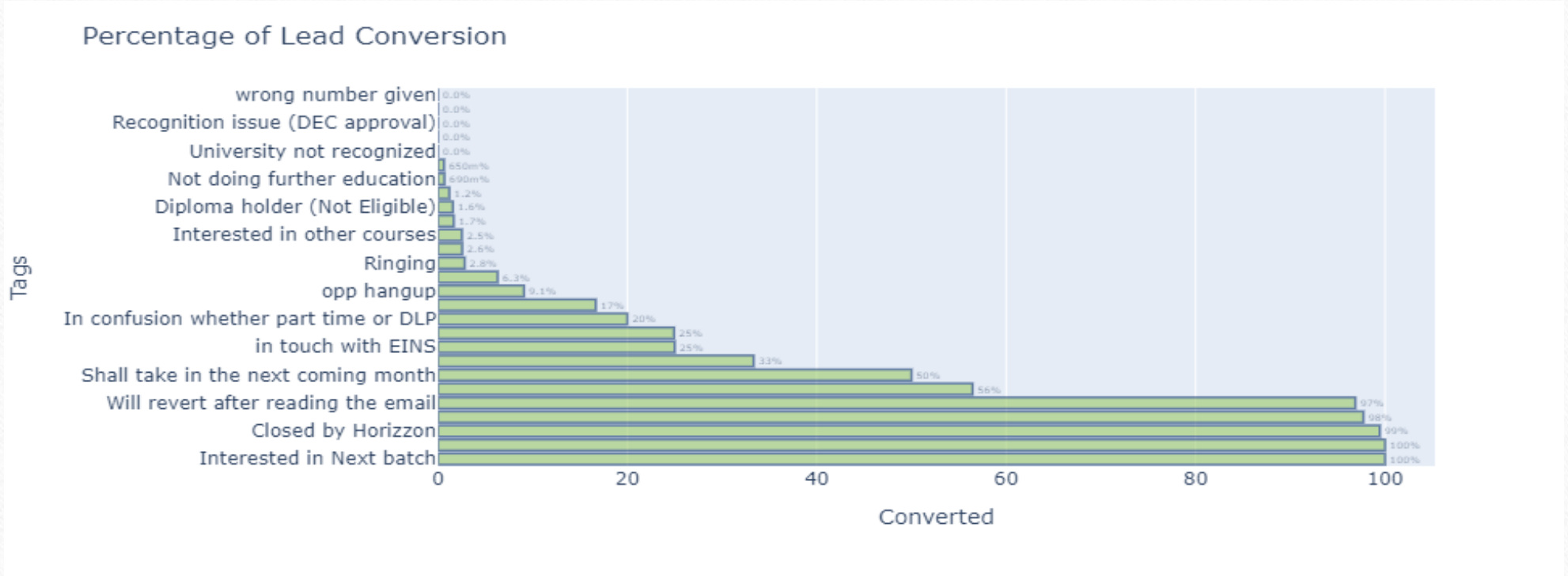
From graph, it can be said that Specialization have no much influence on lead conversion.
- It Can be seen that Customer belongs to Management Profession like Banking ,Investment ,Insurance, Healthcare, Marketing



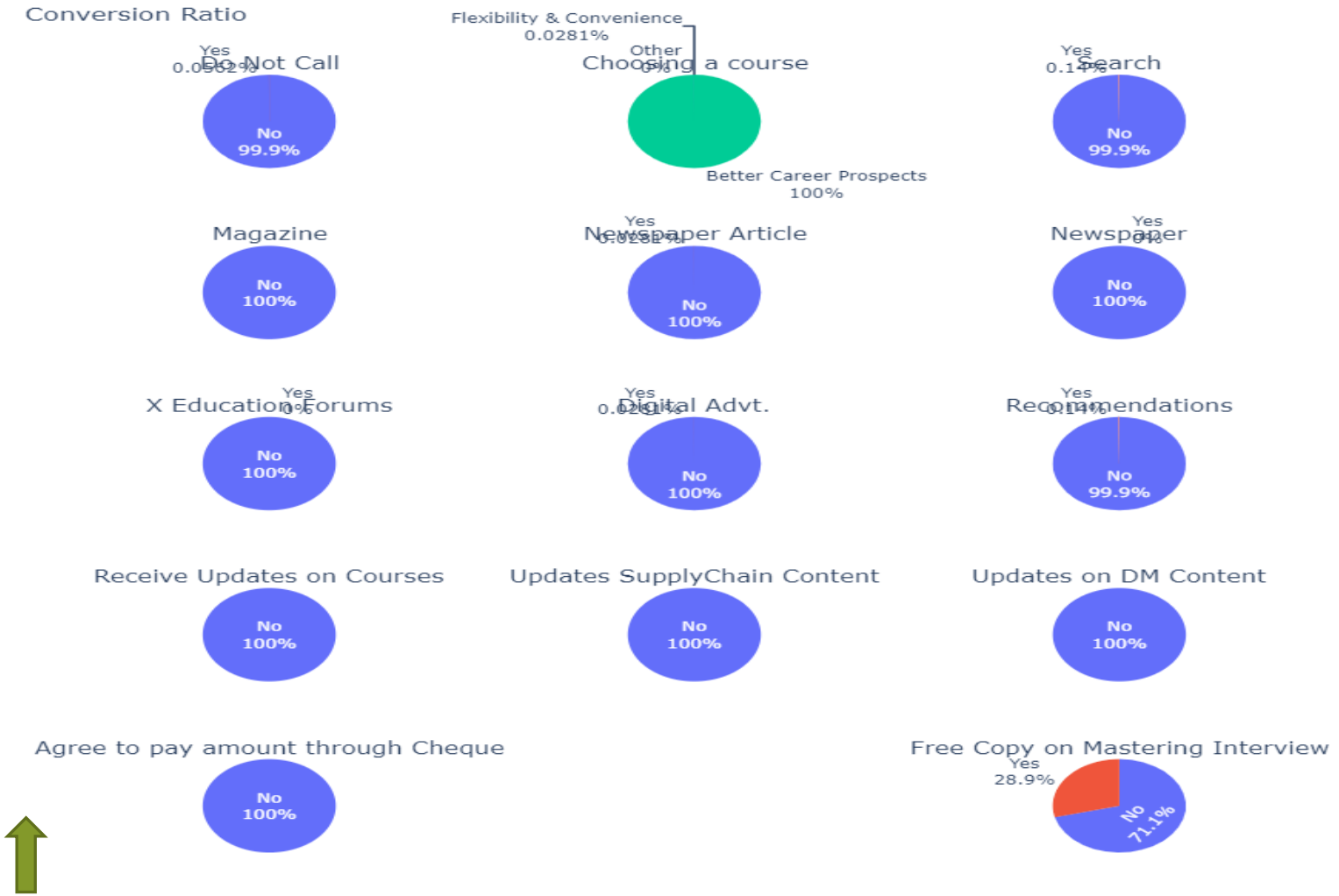
Categorical Variables' Analysis

Inferences:

•From graph, in Tags variable four levels such as 'Interested in next batch', 'Closed by Horrizon', 'Lost to EINS' and 'will revert back after reading email' have almost 100% conversion rate to be potential Leads
•But these tags are not known primarily ,they are given after lead identification



Categorical Variables' Analysis



Inferences:

From graph, given Variables has no much influence on Lead Conversion .

Evaluation Definitions

	Predicted Negative(0)	Predicted Positive(1)
Actual Negative(0)	**True Negative (TN)**	**False Postive (FP)**
Actual Positive(1)	**False Negative (FN)**	**True Positive (TP)**

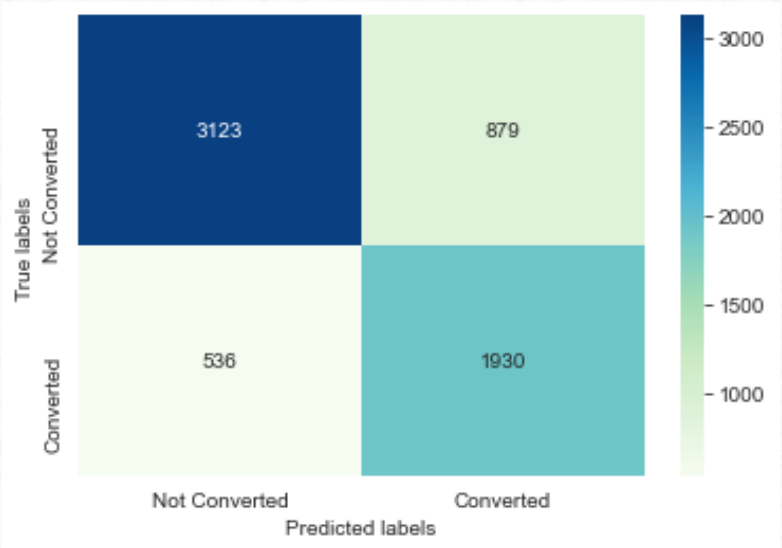
Accuracy = (TP + TN)/(TP + TN + FP + FN)
Sensitivity = TP/(TP + FN)
Specificity = TN/(TN + FP)
Precision =TP/(TP + FP)
Recall = TP/(TP + FN)
F Measure (F1) = 2 * (Precision * Recall)/(Precision + Recall)
TPR (True Positive Rate) = TP/(TP + FN)
TNR (True Negative Rate) = TN/(TN + FP)
FPR (False Positive Rate) = FP/(TN + FP)
FNR (False Negative Rate) = FN/(TP + FN)

Confusion Matrix for train data

Inferences:

•From graph, 1980 have been converted correctly with cutoff value 0.35 for probability calculation

Model Accuracy value is	: 78.12 %
Model Sensitivity value is	: 78.26 %
Model Specificity value is	: 78.04 %
Model Precision value is	: 68.71 %
Model Recall value is	: 78.26 %
Model True Positive Rate (TPR)	: 78.26 %
Model False Positive Rate (FPR)	: 21.96 %
Model Positive Prediction Value is	: 68.71 %
Model Negative Prediction value is	: 85.35 %

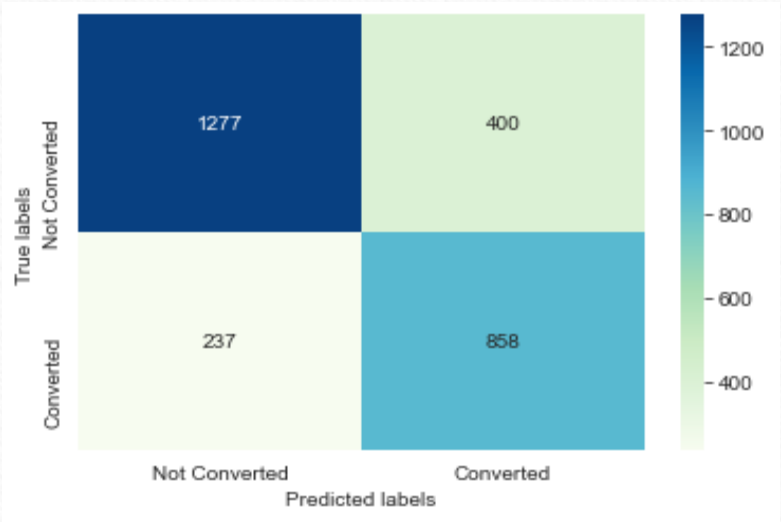


Confusion Matrix for test data

Inferences:

•From graph, 1980 have been converted correctly with cutoff value 0.35 for probability calculation

Model Accuracy value is	: 77.02 %
Model Sensitivity value is	: 78.36 %
Model Specificity value is	: 76.15 %
Model Precision value is	: 68.2 %
Model Recall value is	: 78.36 %
Model True Positive Rate (TPR)	: 78.36 %
Model False Positive Rate (FPR)	: 23.85 %
Model Positive Prediction Value is	: 68.2 %
Model Negative Prediction value is	: 84.35 %

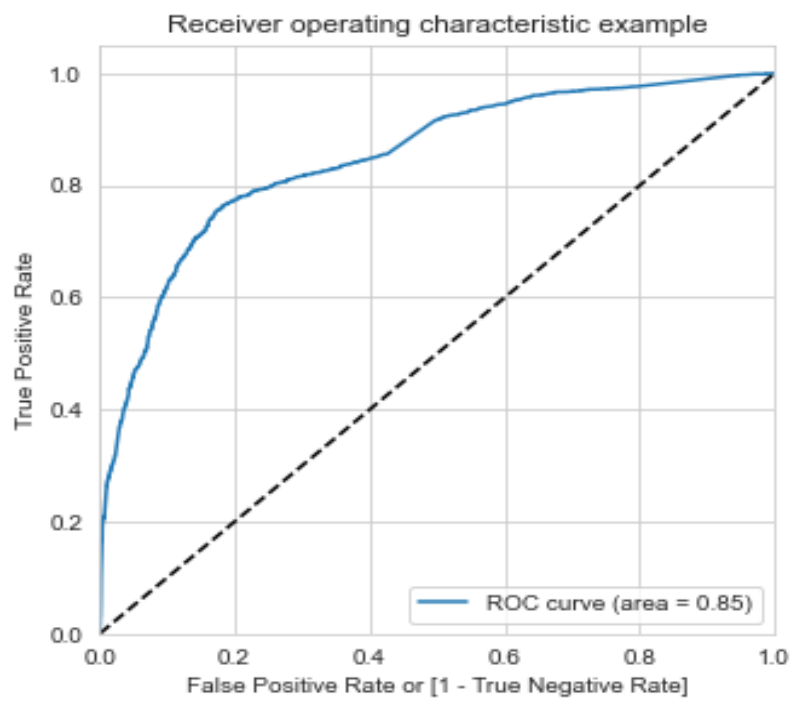


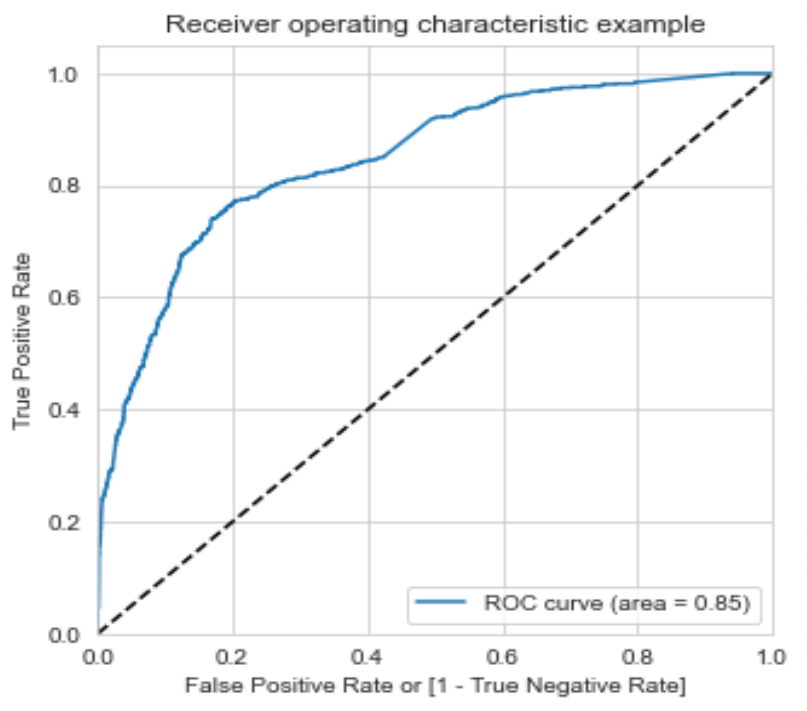
ROC Curve for train data

ROC Curve for test data

Inferences:

- From graph, **ROC Curve area is 0.85 for both train and test, which indicates that the model is good and not overfitting.**
- Gini = 0.85





Conclusions

After analyzing the dataset and model building using logistic regression , there are few attributes of a customers(leads) with which the X Education would be able to identify the Hot Leads depending upon the lead score calculated between 0 to 100.

•We can use the lead_score column to identify which potential leads to prioritize first. The higher the score, the higher chances are there for the lead to convert. If there are limited sales representatives, then score cut-off should be higher to ensure a higher conversion probability people are contacted further to turn them into a potential customer.

•In case there are interns, then the score cut-off can be lowered. As there are more human resources, the company can afford a higher rate of False positives as it will increase the customer outreach and, in turn, increase the potential customer who will take the online courses.

•Tags were not included as predictor variable. Another model including Tags as a predictor variable can be built and segmented accordingly

Conclusions

After analyzing the dataset and model building using logistic regression , there are few attributes of a customers(leads) with which the X Education would be able to identify the Hot Leads. The factors with relative importance are as below

•Lead Origin_Lead Add Form	100.00
•Occupation_Working Professional	94.41
•Lead Source_Welingak Website	56.31
•Lead Origin_API	30.47
•Occupation_Unemployed	28.89
•Occupation_Student	26.01
•Total Time Spent on Website	24.97
•City_TierIII	22.98
•City_TierI	22.17
•City_TierII	20.59
•Specialization_Hospitality Management	-23.80