LEAD SCORE CASE STUDY

PRESENTED BY:

ALDRIN FIGUEREDO

RISHABH GANESH

PRAVEEN KUMAR

DS C45

CONTENTS:

- Problem Statement
- Step by Step process
- Data manipulation
- EDA
- Data Imbalance
- Univariate, Bivariate & Multi-variate
- Visualizing and handling the outliers
- Correlation
- Model Building
- ROC Curve
- Conclusion

PROBLEM STATEMENT

An X Education sells online courses to industry professional. 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.

Although the company gets a lot of leads, its lead conversion rate is very poor. 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 all.

Business Objective:

- A regression model to assign a lead score between 0 to 100 to each lead to target the potential lead.
- Address Business problems presented by company through the model.
- Target conversion rate to be around 80%.

STEP BY STEP PROCESS

- Import Libraries
- Import data file and basic checks
- Data cleaning and pre-processing
- Exploratory Data Analysis
- Creation of Dummy variables
- Splitting the data into test and train
- Scaling the data
- Model Building through feature selection RFE & VIF
- Creating Prediction
- Model Evaluation
- Optimise Cut Odd (ROC Curve)
- Prediction on test data set
- Precision and Recall
- Conclusion

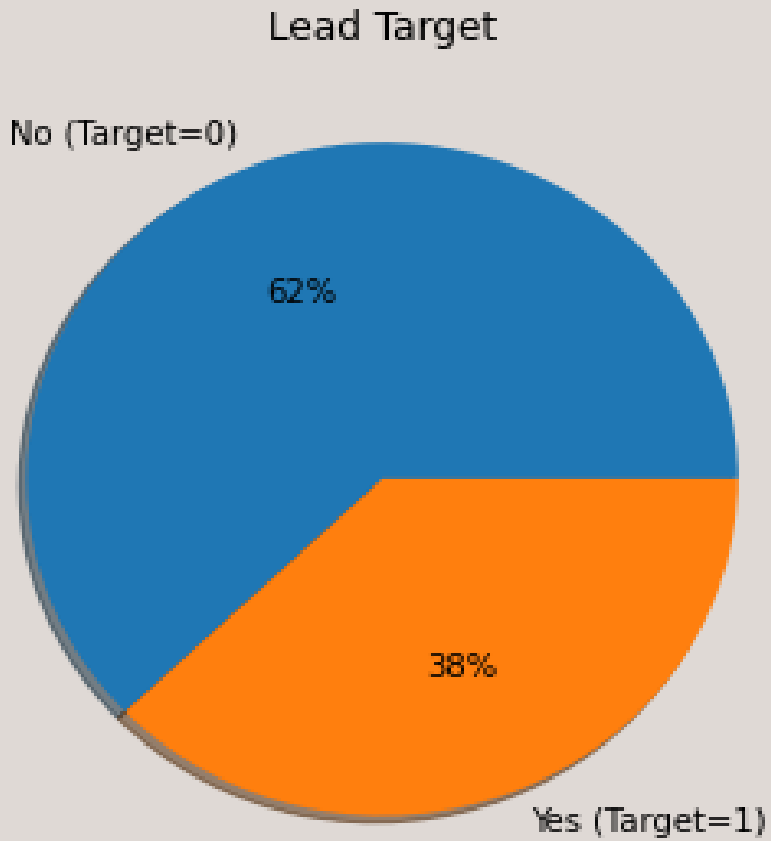
DATA MANIPULATION

- ✓ The Data set contained total number of rows = 37 and columns = 9240.
- ✓ The ‘Select’ option values in the datasets were replaced with NaN.
- ✓ Single value features like ‘Magazine’, ‘Receive More Updates About Our Course’, ‘Update me on supply chain content’. ‘Get updates on DM content’, ‘I agree to pay the amount through cheque’ were dropped along with ‘Prospect ID’ and ‘Lead Number’ which are not necessary for analysis.
- ✓ By taking the threshold at 35%, the columns which have null values more than 35% were dropped and columns with null values less than 2% were imputed.
- ✓ Columns like country, specializations, What matters most to you in choosing a course, What is your current occupation, which had null values were imputed by value ‘not provided’.

EXPLORATORY DATA ANALYSIS

Analysis on Data Imbalance

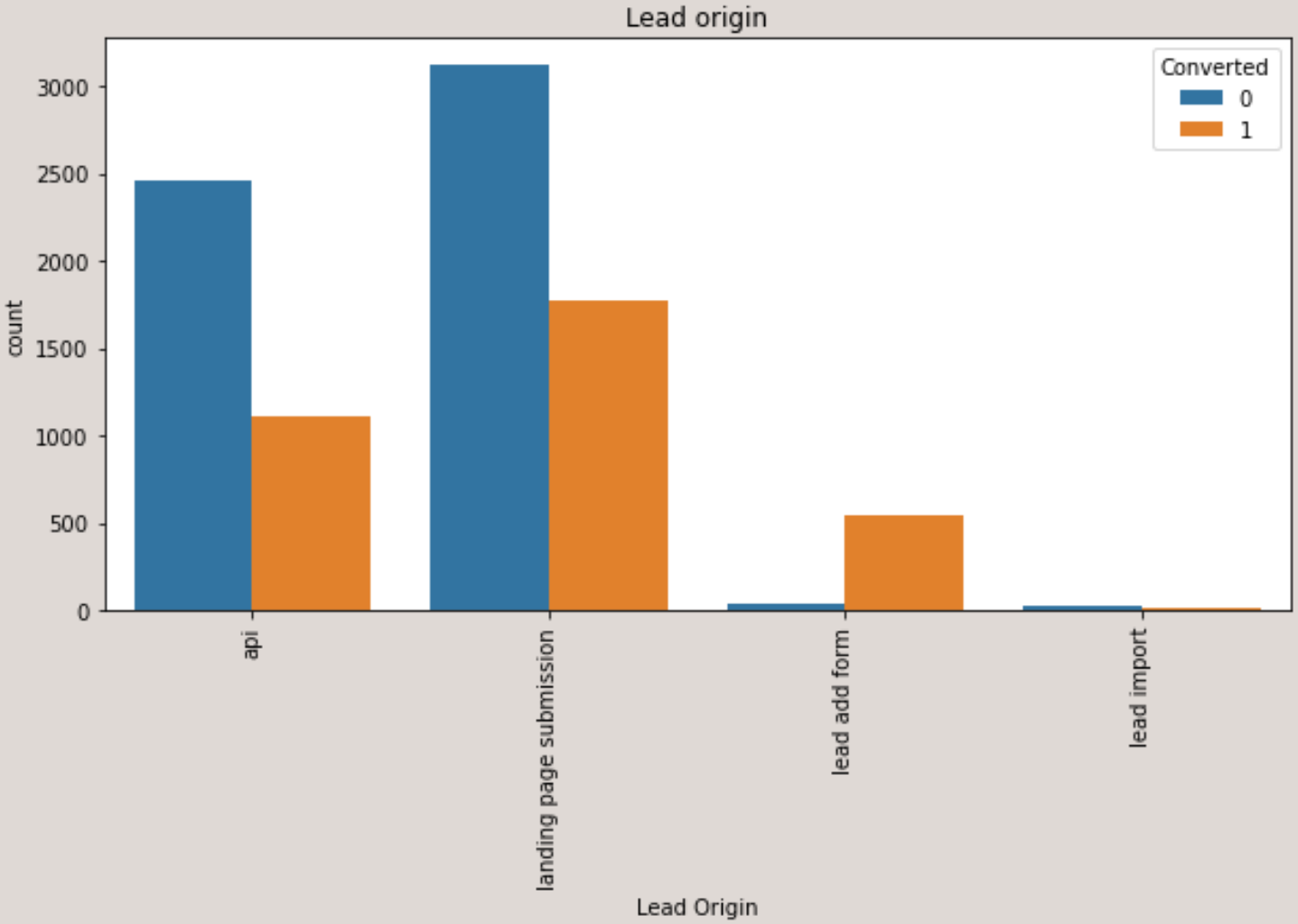
There is a data imbalance in the dataset as only 38% leads are converted and around 62% are not converted.



UNIVARIATE, BI-VARIATE & MULTI-VARIATE ANALYSIS

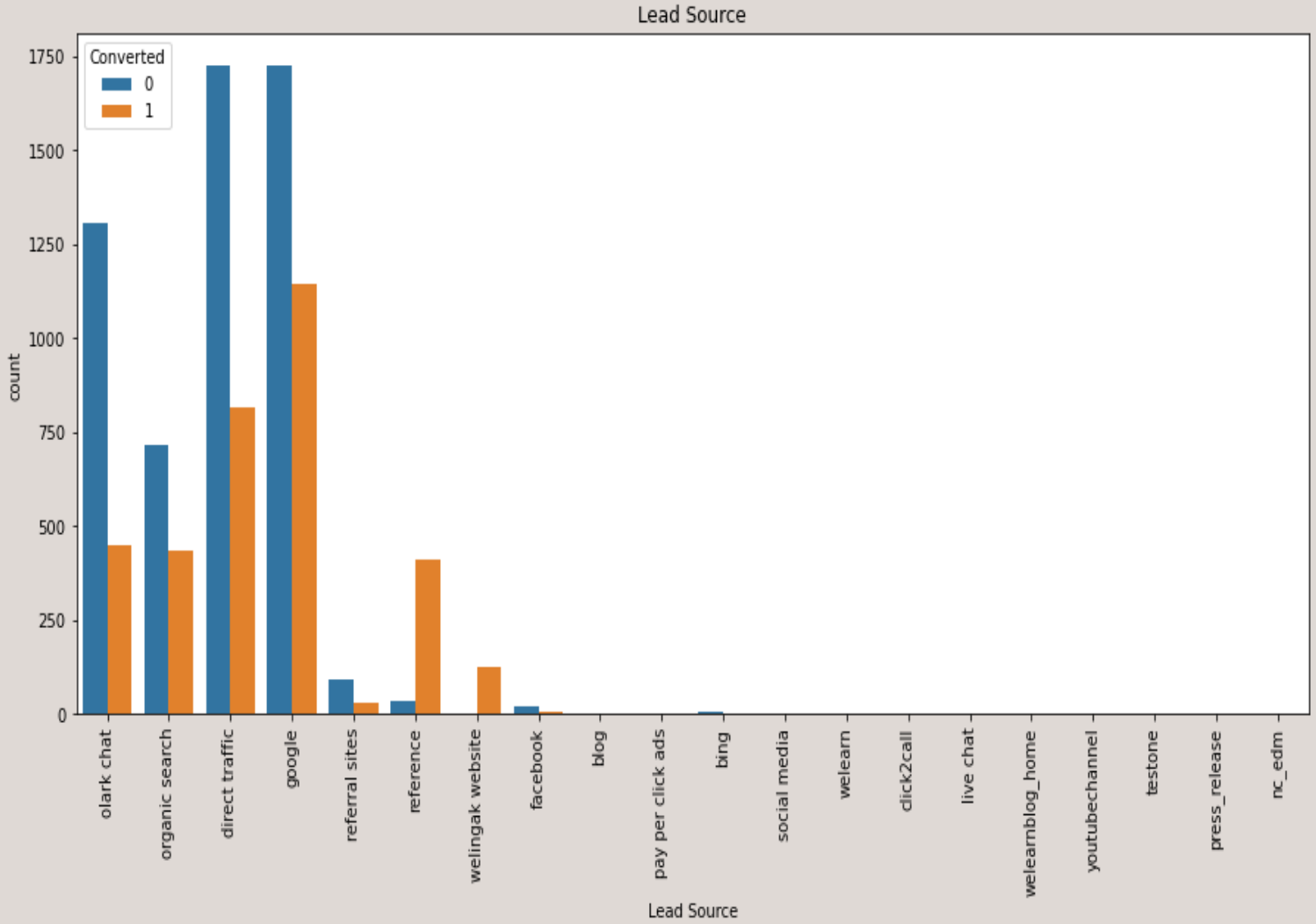
Lead origin

- Customers who were identified as Lead from Landing Page submission, constitute the majority of the leads.
- Customers originating from Lead Add Form have high probability of conversion. These Customers are very few in number.
- Lead origin-API & Lead Import have the least conversion rate. Customers from Lead Import are very few in number.



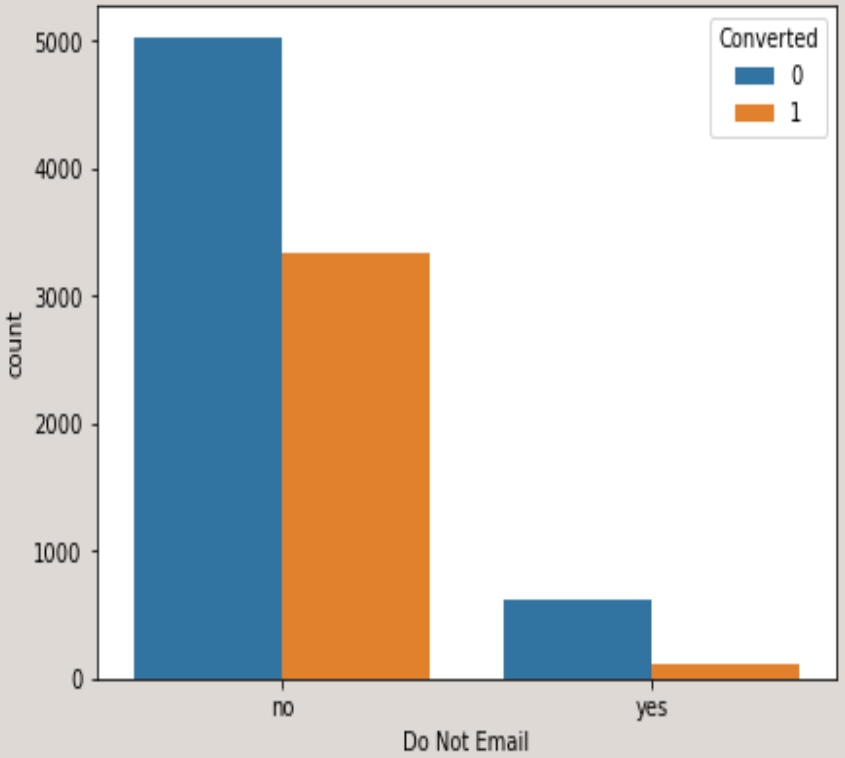
LEAD SOURCE

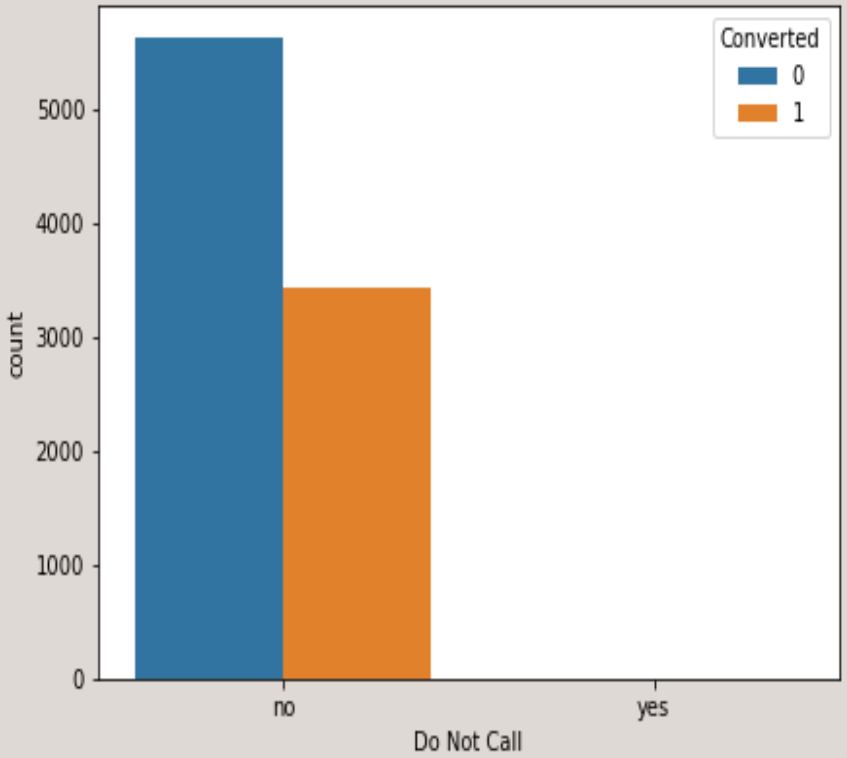
- Majority source of the lead is Google & Direct Traffic.
- Lead source from Google has highest probability of conversion.
- leads with source Reference has maximum probability of conversion.



DO NOT EMAIL AND DO NOT CALL

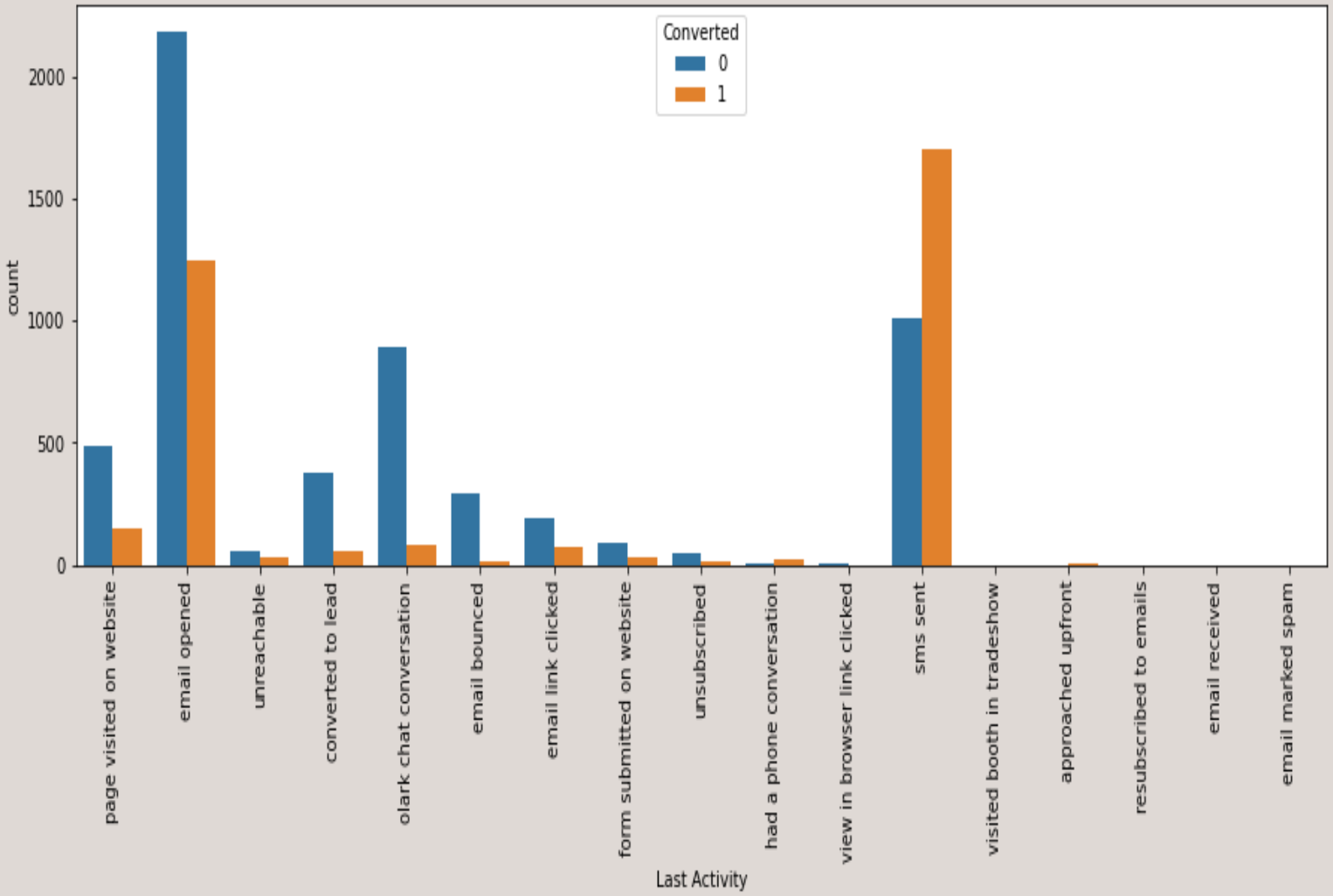
- Customers who opt for Do Not Mail have lower conversion rate.
- Customers who do not opt for Do Not Mail have higher conversion rate which is around 40%. These constitute the majority of the leads.
- Customers who do not opt for Do Not call have Higher conversion rate which is around 38%.These constitute the majority of the leads.





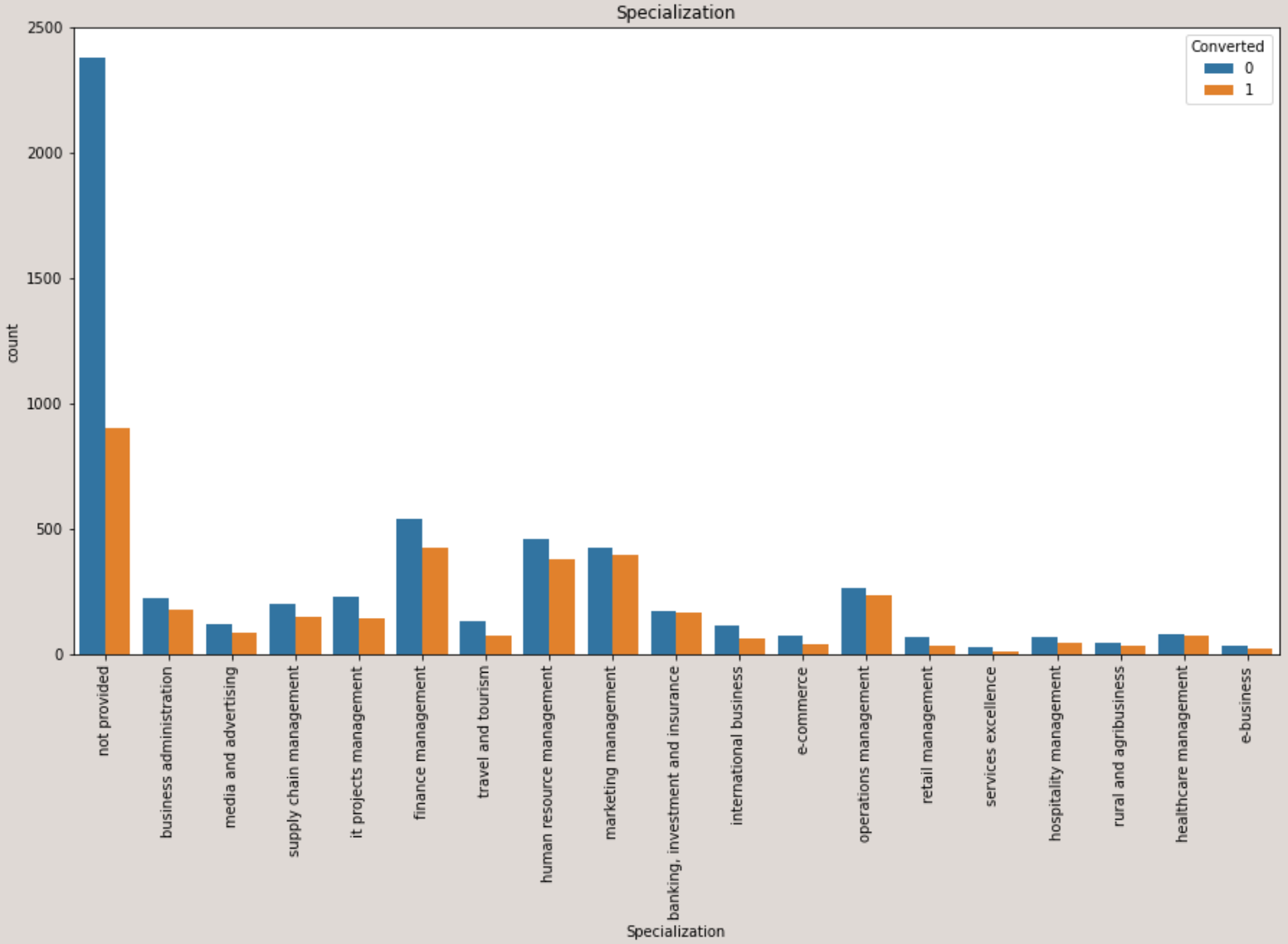
LAST ACTIVITY

- Customers who last activity was SMS Sent have higher conversion rate which is around 63Customers who last activity was Email Opened constitute majority of the customers. They have around 36% of conversion rate.



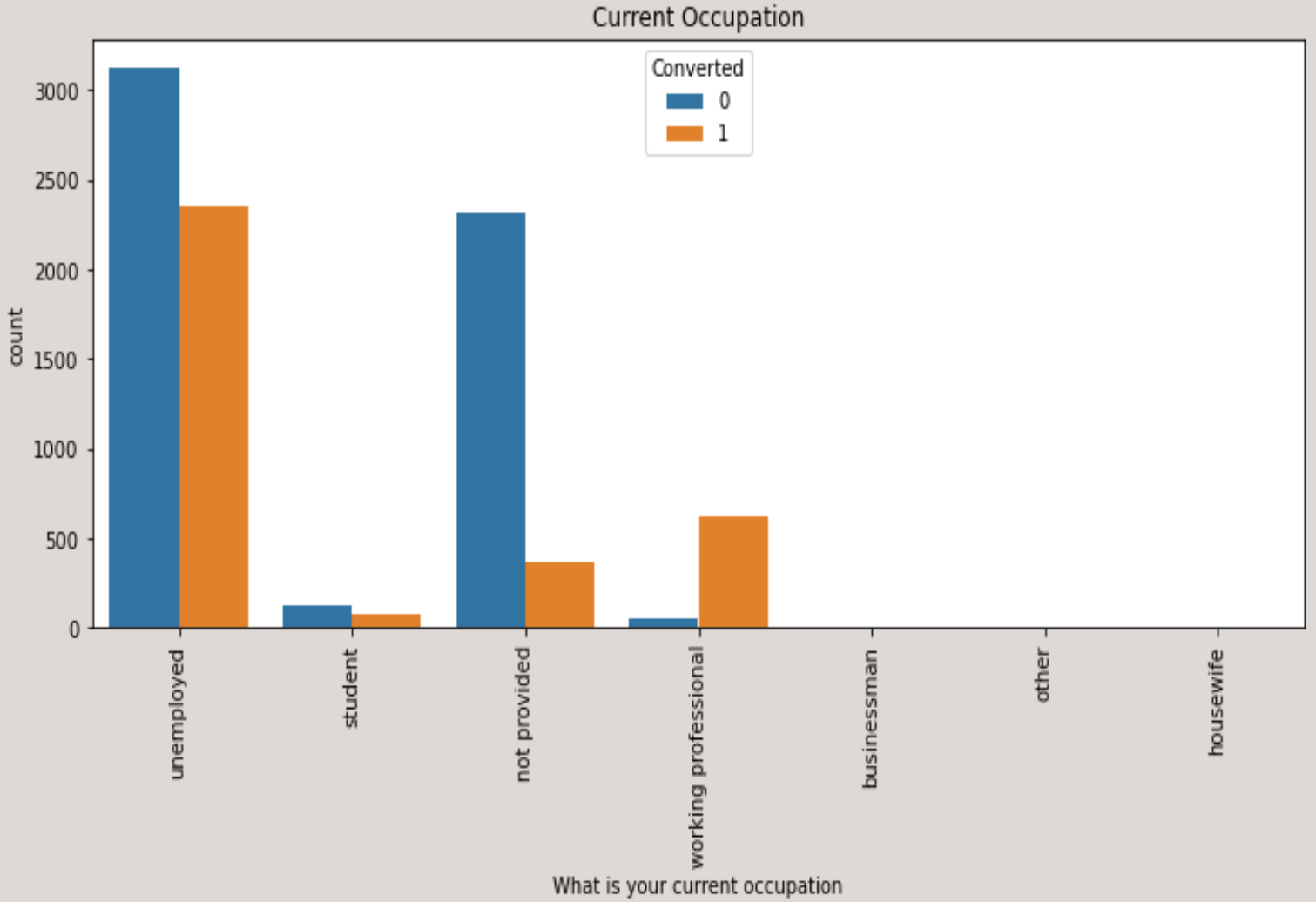
SPECIALIZATION

- Maximum Leads have specialization as Management & Others.
- Leads with specialization as Rural & Agribusiness have least probability of conversion.



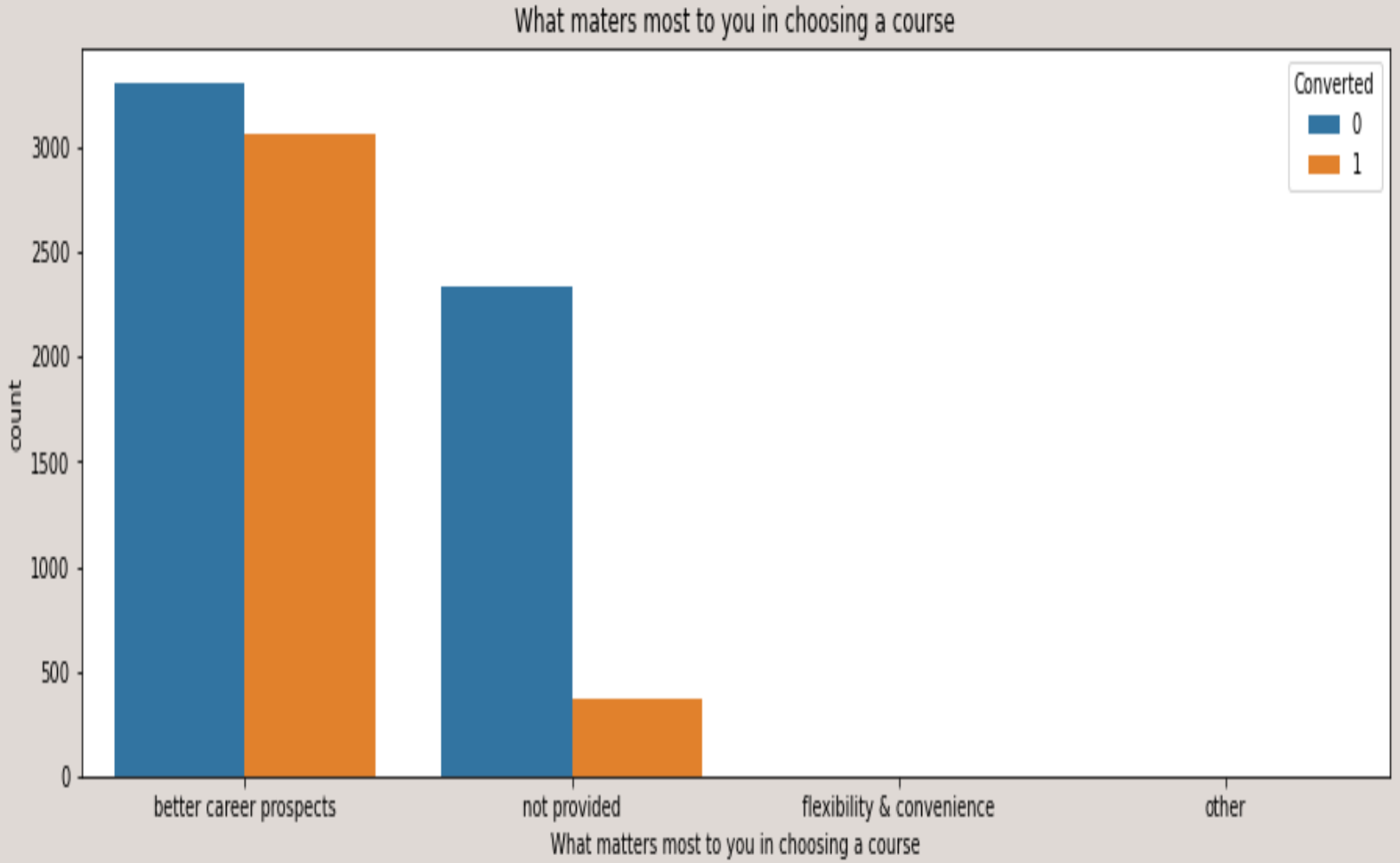
CURRENT OCCUPATION

- Maximum Leads have occupation as Unemployed.
- Very few leads are Housewives.

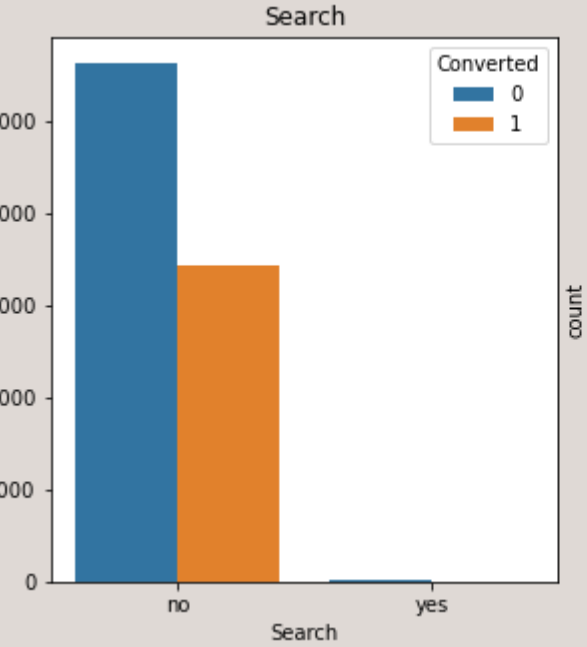


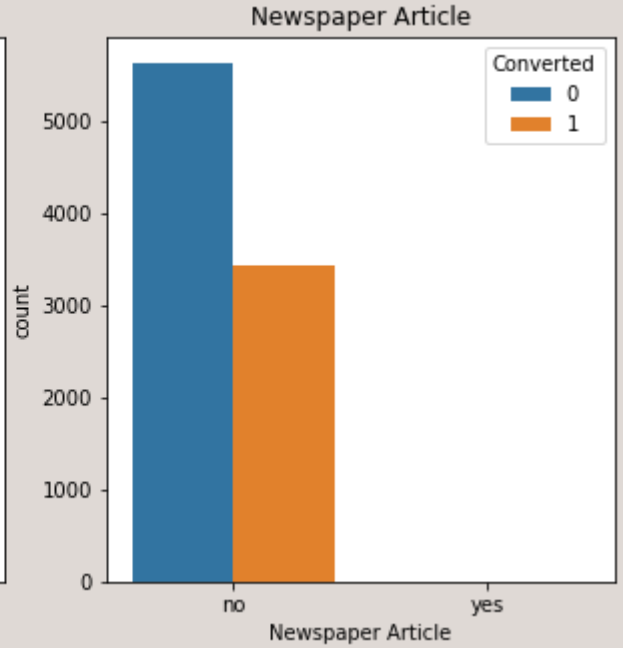
WHAT MATERS MOST TO YOU IN CHOOSING A COURSE

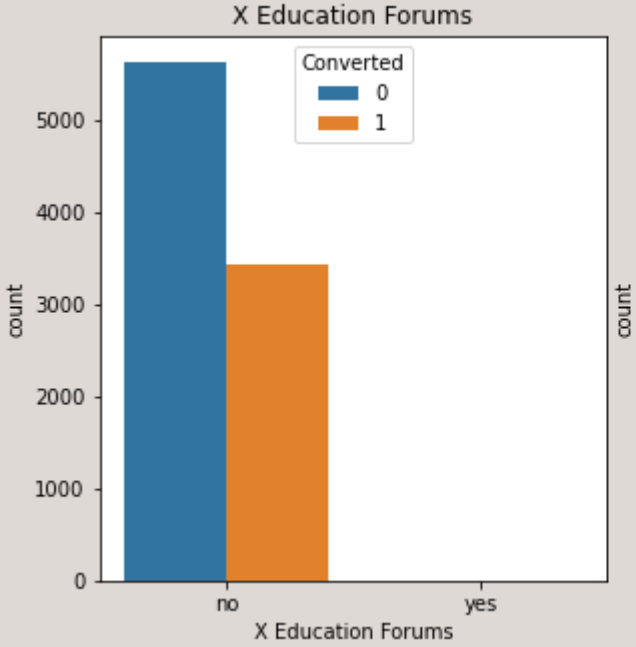
- Number of Leads to whom better career aspects matters most in choosing a career are more & have higher probability of conversion.

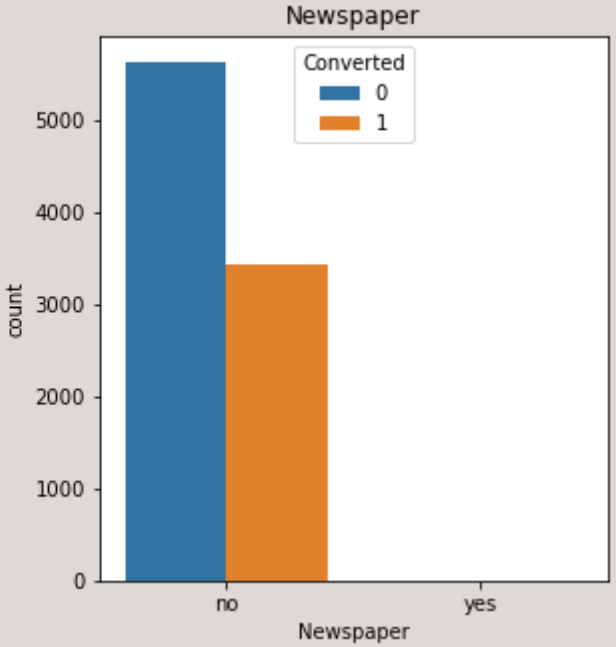


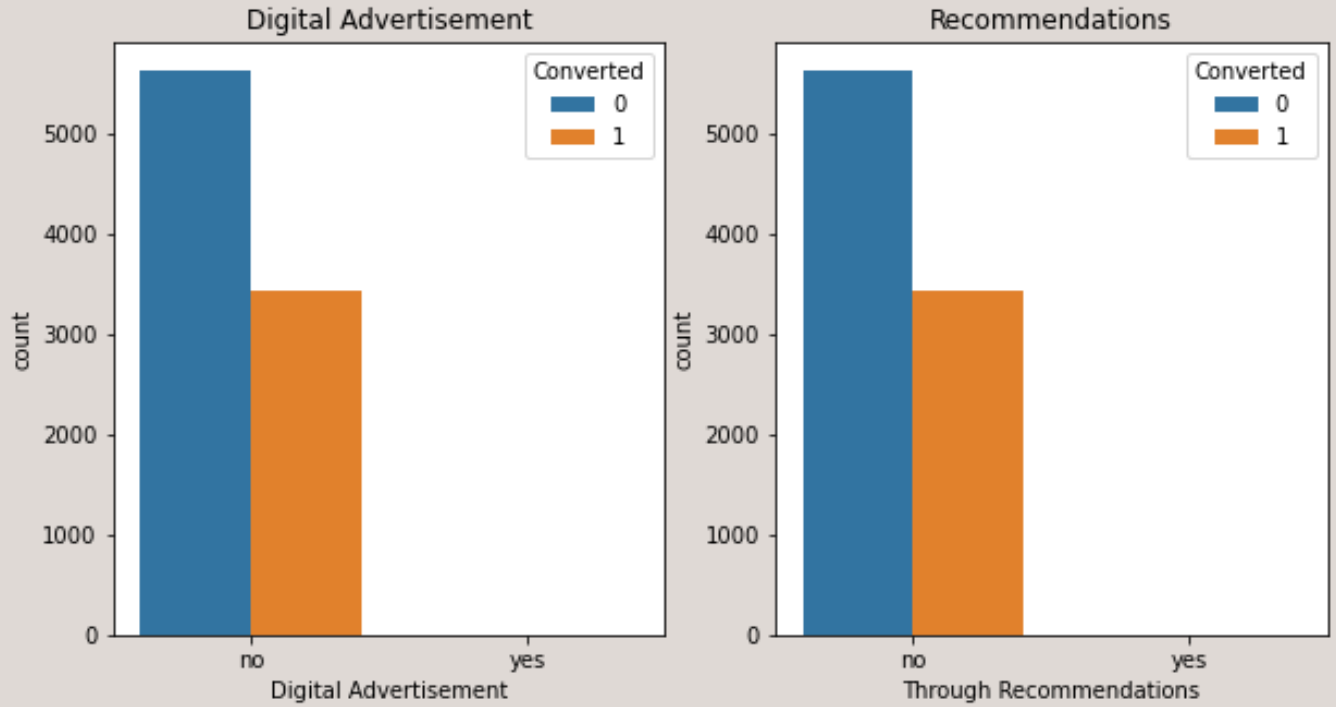
SEARCH, MAGAZINE, NEWSPAPER ARTICLE, X EDUCATION FORUMS, NEWSPAPER, DIGITAL ADVERTISEMENT, RECOMMENDATIONS







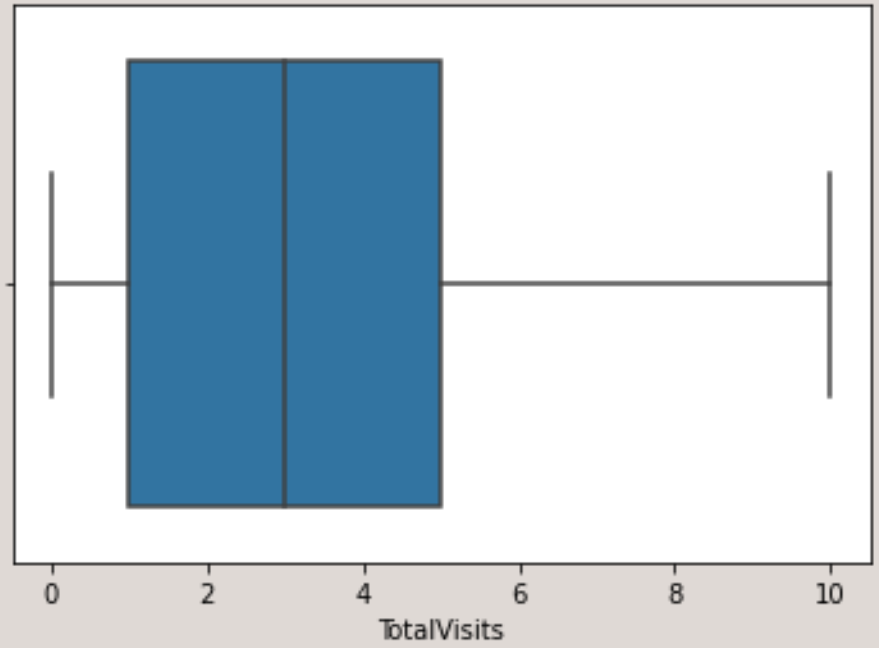


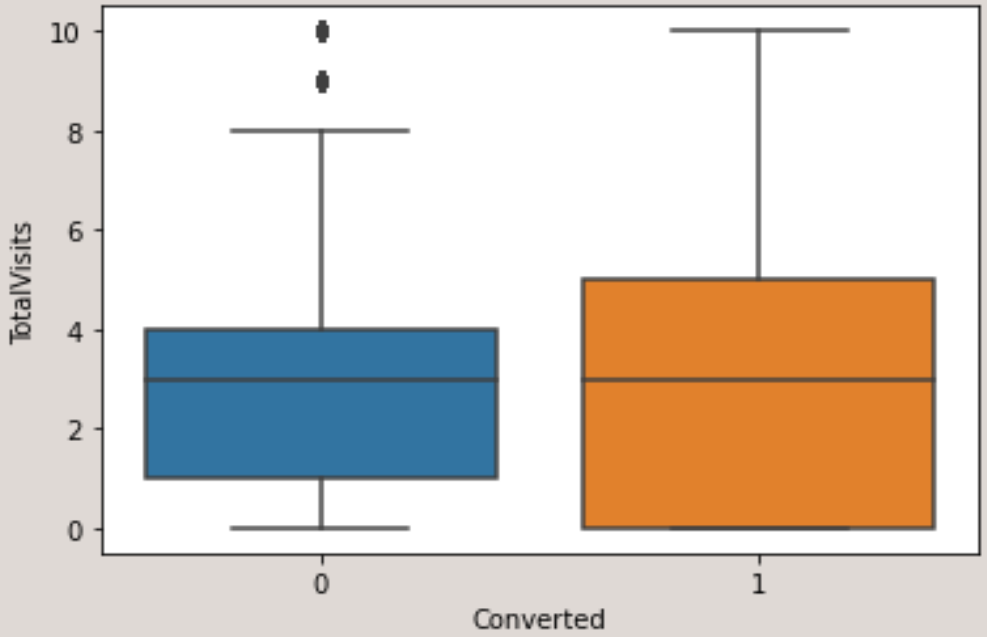


- CUSTOMERS WHO HAVE SEEN THE ADD OF THE EDUCATION COMPANY IN ANY FORM, ARE VERY FEW IN NUMBER. NOTHING MEANINGFUL INSIGHT CAN BE CONCLUDED FROM THE PLOT THAT WILL IMPROVE THE OVERALL LEAD CONVERSION RATE.

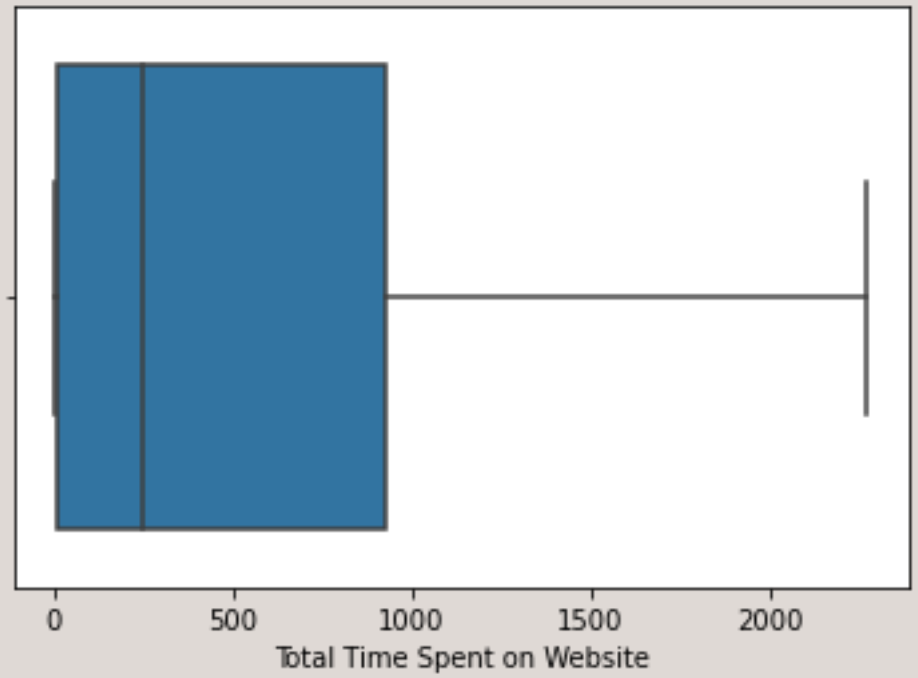
VISUALIZING AND HANDLING THE OUTLIERS

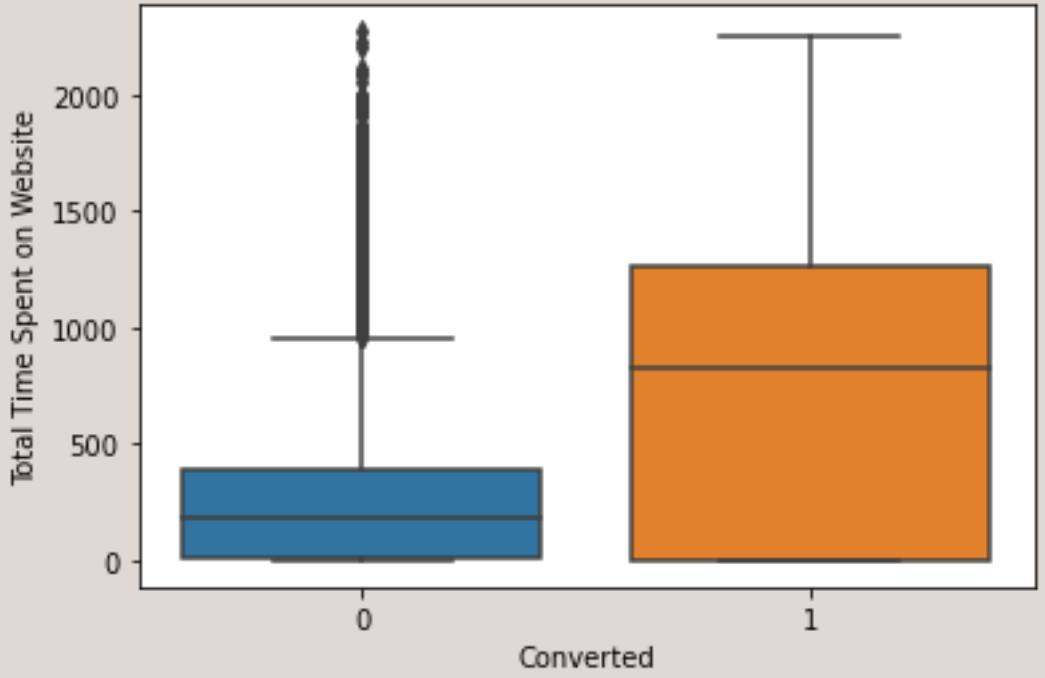
TOTAL VISITS





TOTAL TIME SPENT





CORRELATION

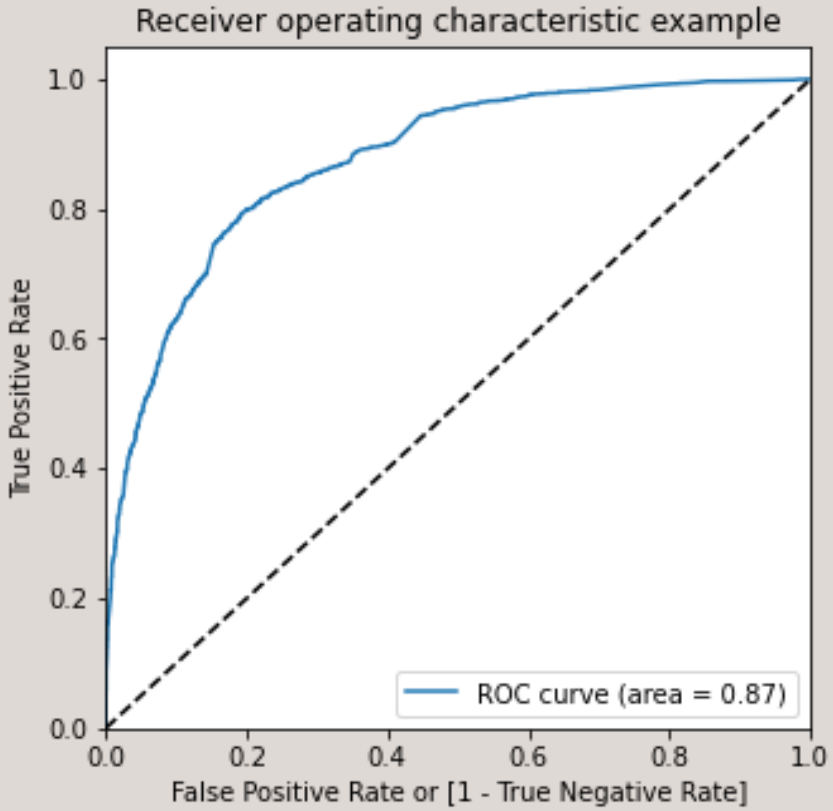
- From the above graph we can conclude that all the variables are multicollinear in nature. The correlation between the Total Visits and Page Views Per visit is high.

MODEL BUILDING

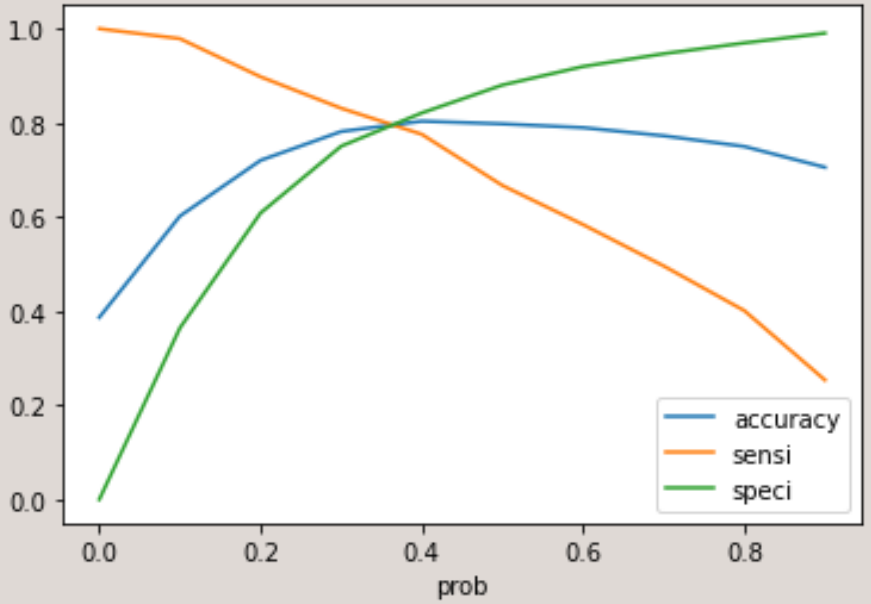
- ✓ Creating Dummies for variables Lead origin, Specialization, Lead Source, Do Not email, Last Activity, What is your current occupation, A free copy of Mastering The Interview and Last Notable Activity.
- ✓ After creating dummies splitting the dataset into 70% Train and 30% Test data sets.
- ✓ Scaling the train data set using Scaler function MinMax.
- ✓ Using feature selection RFE method by selecting 15 variables.
- ✓ Building model by removing the variable whose p-value is greater than 0.5 and VIF value is greater than 5.
- ✓ Model – 4 was the perfect model and was selected for model evaluation on test data set since all the VIF values are good and all the p-values were below 0.5.
- ✓ After predicting on the test data set, the overall accuracy was around 81% which is above the target conversion rate of 80%.

ROC CURVE

- ✓ By using ROC curve method of feature selection, below were the findings:
- ✓ With the cut off 0.5, the accuracy is around 81%, sensitivity is around 70% and specificity is around 88%.
- ✓ By optimising the ROC curve, the area under the curve is 0.88 indicating a very good value.



- ✓ With the cut off of 0.35, the accuracy, sensitivity and specificity is around 80% which can be observed in the below graph.



- ✓ By predicting the model on test data set the accuracy, sensitivity and specificity is around 80%.
- ✓ By performing the precision and recall method, the precision is around 79% and recall is around 70%.
- ✓ With the cut off of 0.41 and performing the trade off of precision and recall, the precision is around 74% and recall is around 76%.
- ✓ The final predication on test data set with the cut off of 0.41, we got the precision around 73% and 76%.

CONCLUSION

- From the analysis it was found that when the time is limited with the company, it needs to approach the hot leads i.e. those customers who are more likely to have conversion chance to achieve maximum conversion.
- When the time is ample with the company along with resources and time it needs to approach all the potential leads and also approach customers who have lesser conversion rate to improve overall conversion rate.
- To improve overall lead conversion rate, we need to focus more on improving lead converion rate of Customers originating from API and Landing Page Submission and generate more leads from Lead Add Form.
- To improve overall lead conversion rate, we need to focus more on improving lead converion rate of Customers whose last activity was Email Opened and generate more leads from the ones whose last activity was SMS Sent.