Lead Score Case Study

Presented by :
AJEET KUMAR
SHRAWANI DAS
SAURAV KUMAR

Problem Statement

- X Education sells online courses to industry professionals.
- X Education gets a lot of leads, its lead conversion rate is very poor. For example, if, say, they acquire 100 leads in a day, only about 30 of them are converted.
- 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

Business Objective

- X education wants to know most promising leads.
- For that they want to build a Model which identifies the hot leads.
- Deployment of the model for the future use.

Solution Methodology

- Data cleaning and data manipulation.

1. Check and handle duplicate data.

2. Check and handle NA values and missing values.

3. Drop columns, if it contains large amount of missing values and not useful for the analysis.

4. Imputation of the values, if necessary.

5. Check and handle outliers in data.

- EDA

1. Univariate data analysis: value count, distribution of variable etc.

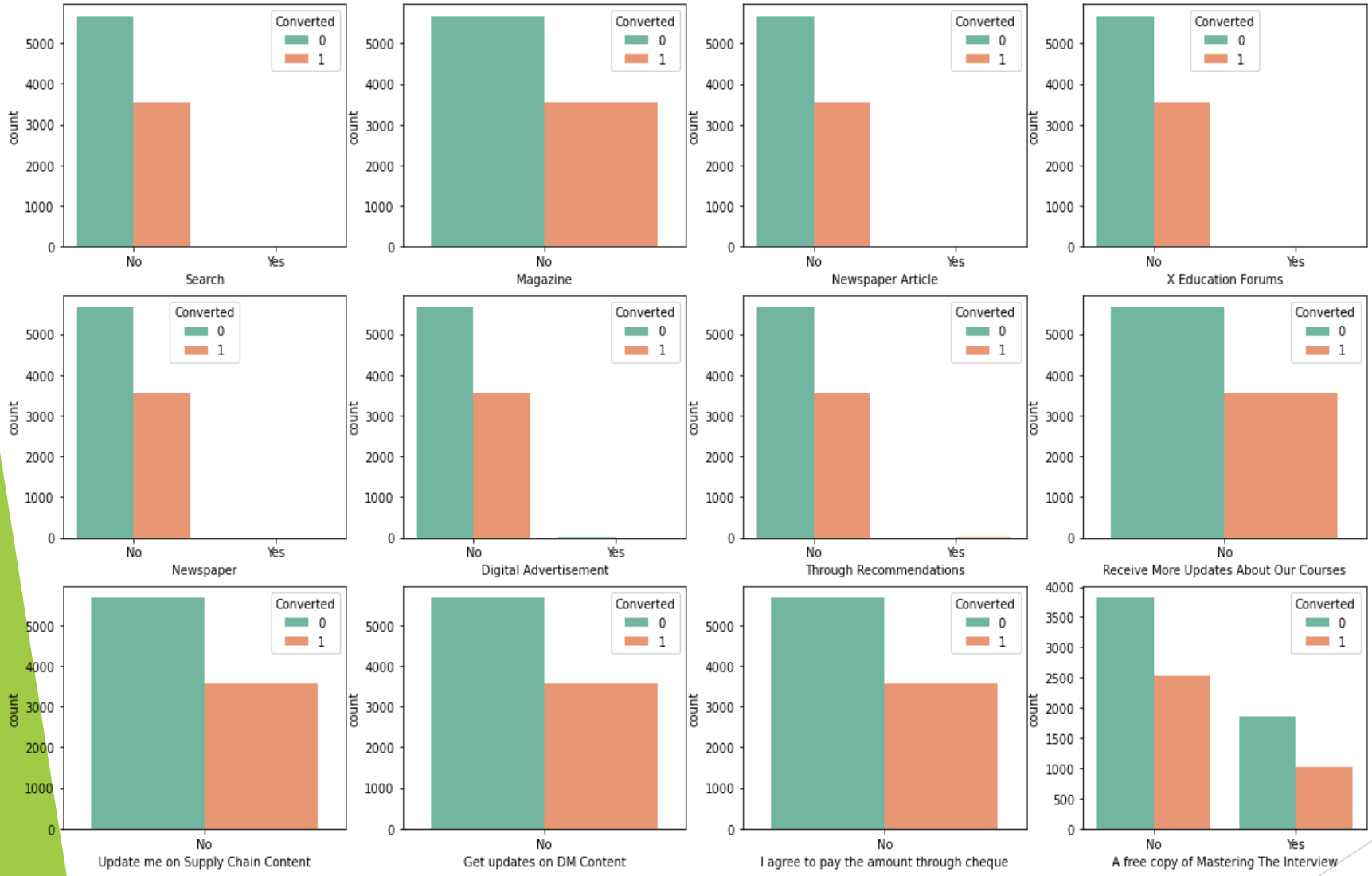
2. Bivariate data analysis: correlation coefficients and pattern between the variables etc.

- Feature Scaling & Dummy Variables and encoding of the data.
- Classification technique: logistic regression used for the model making and prediction.
- Validation of the model.
- Model presentation.
- Conclusions and recommendations

Data Manipulation

- Total Number of Rows =37, Total Number of Columns =9240.
- Single value features like "Magazine", "Receive More Updates About Our Courses", "Update me on Supply".
- "Chain Content", "Get updates on DM Content", "I agree to pay the amount through cheque" etc. have been dropped.
- Removing the "Prospect ID" and "Lead Number" which is not necessary for the analysis.
- After checking for the value counts for some of the object type variables, we find some of the features which has no enough variance, which we have dropped, the features are: "Do Not Call", "What matters most to you in choosing course", "Search", "Newspaper Article", "X Education Forums", "Newspaper", "Digital Advertisement" etc.
- Dropping the columns having more than 35% as missing value such as 'How did you hear about X Education' and 'Lead Profile'.

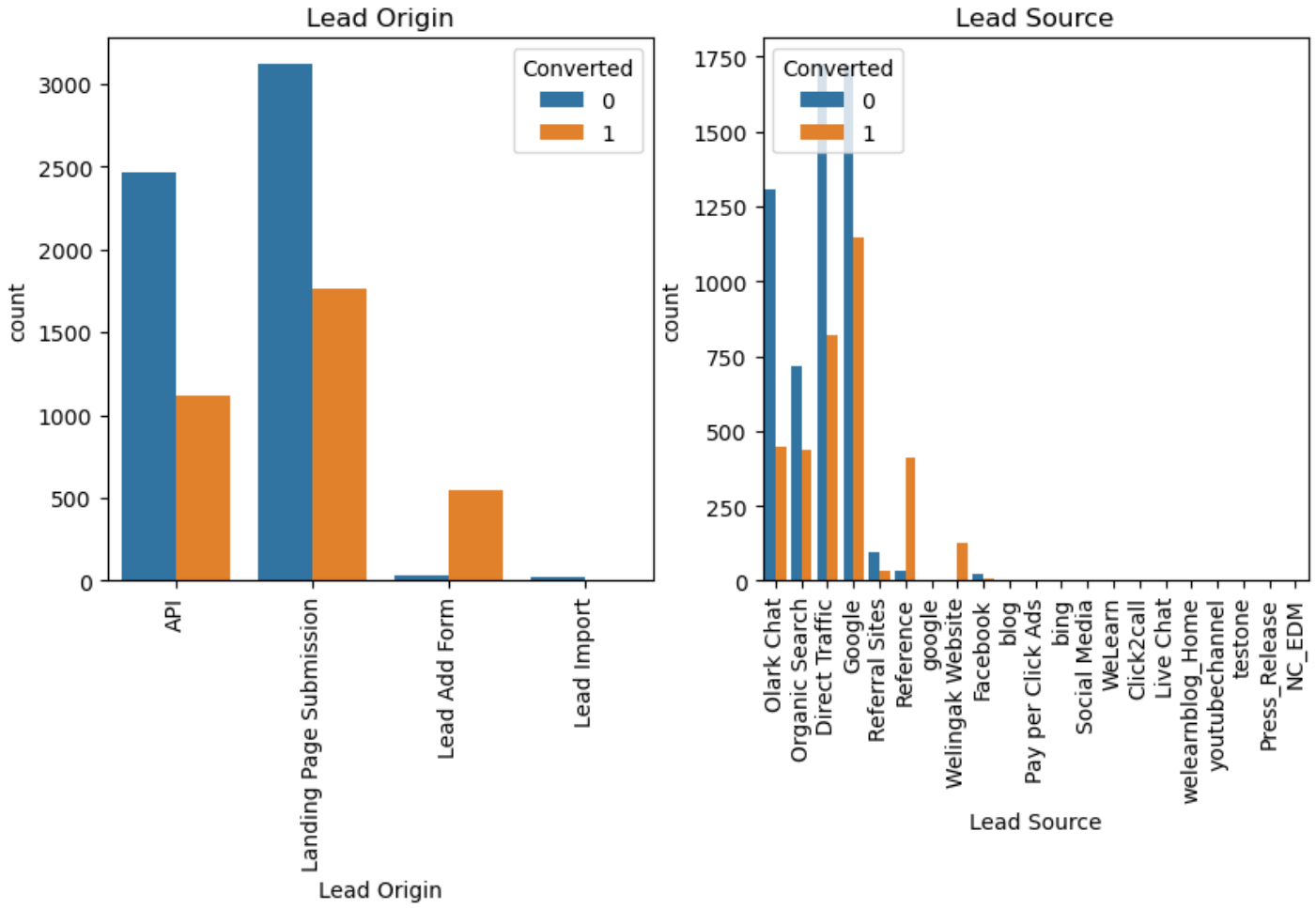
EDA



Interface:

- For all these columns except 'A free copy of Mastering The Interview' data is highly imbalanced, thus we will drop them
- "A free copy of Mastering The Interview" is a redundant variable so we will include this also in list of dropping columns.

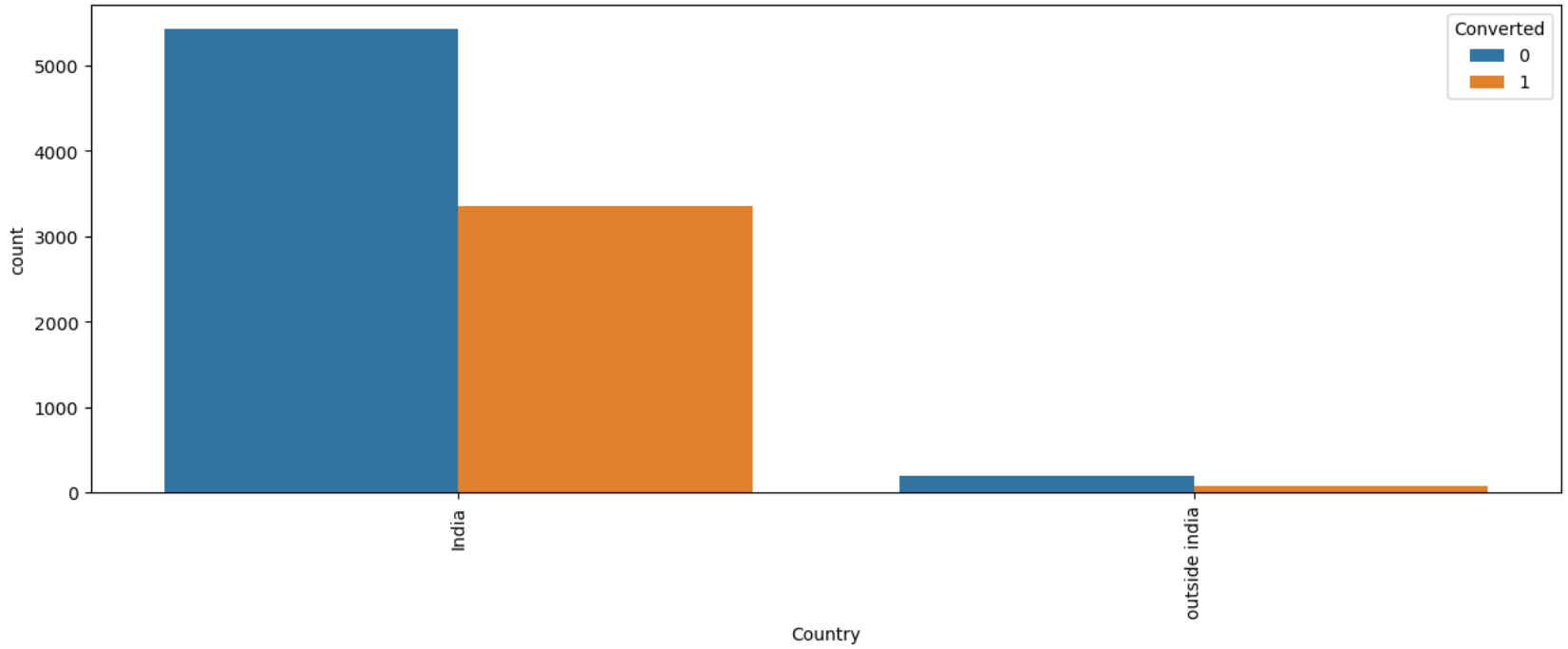
Lead Source Variable based on Converted value



Interface :

- Maximum Leads are generated by Google and Direct Traffic.
- Conversion rate of Reference leads and Welinkgak Website leads is very high.

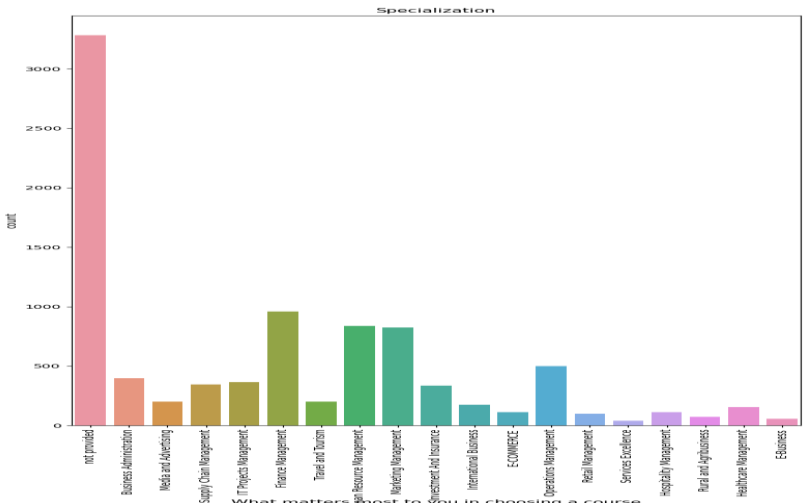
Country variable

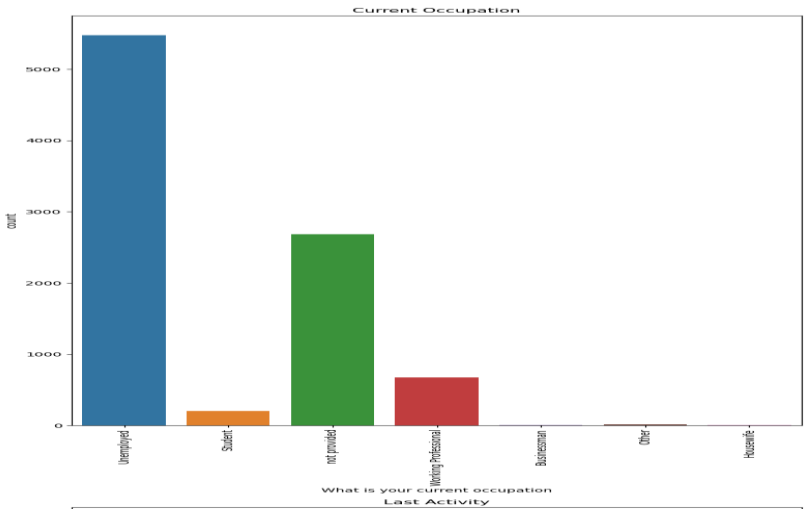


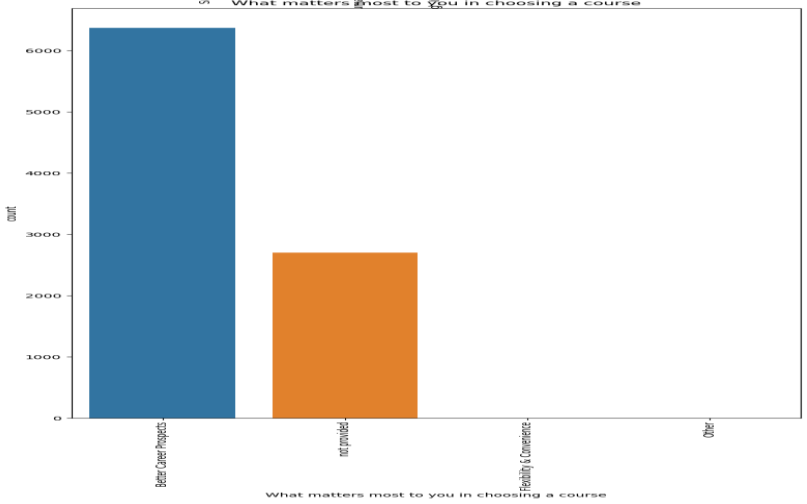
Interface :

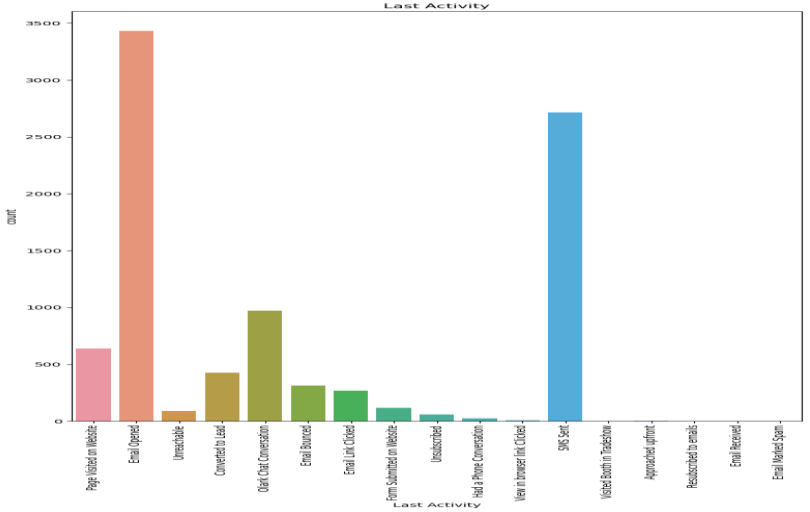
- As we can see that most of the data consists of value 'India', no inference can be drawn from this parameter.Hence, we can drop this column

Count of Variable based on Converted value





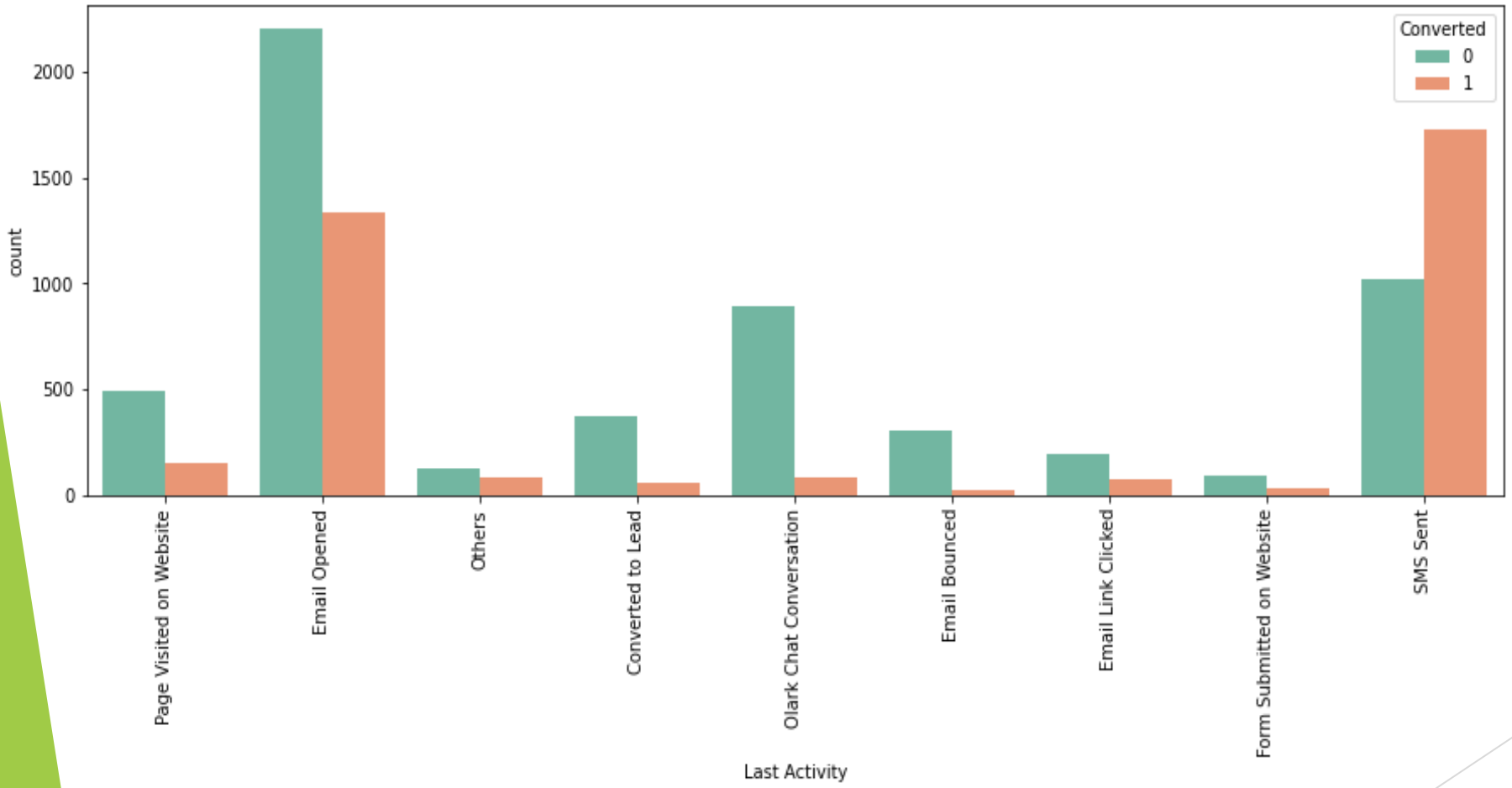




Interface :

- Maximum leads generated are unemployed and their conversion rate is more than 50%.
- Conversion rate of working professionals is very high.

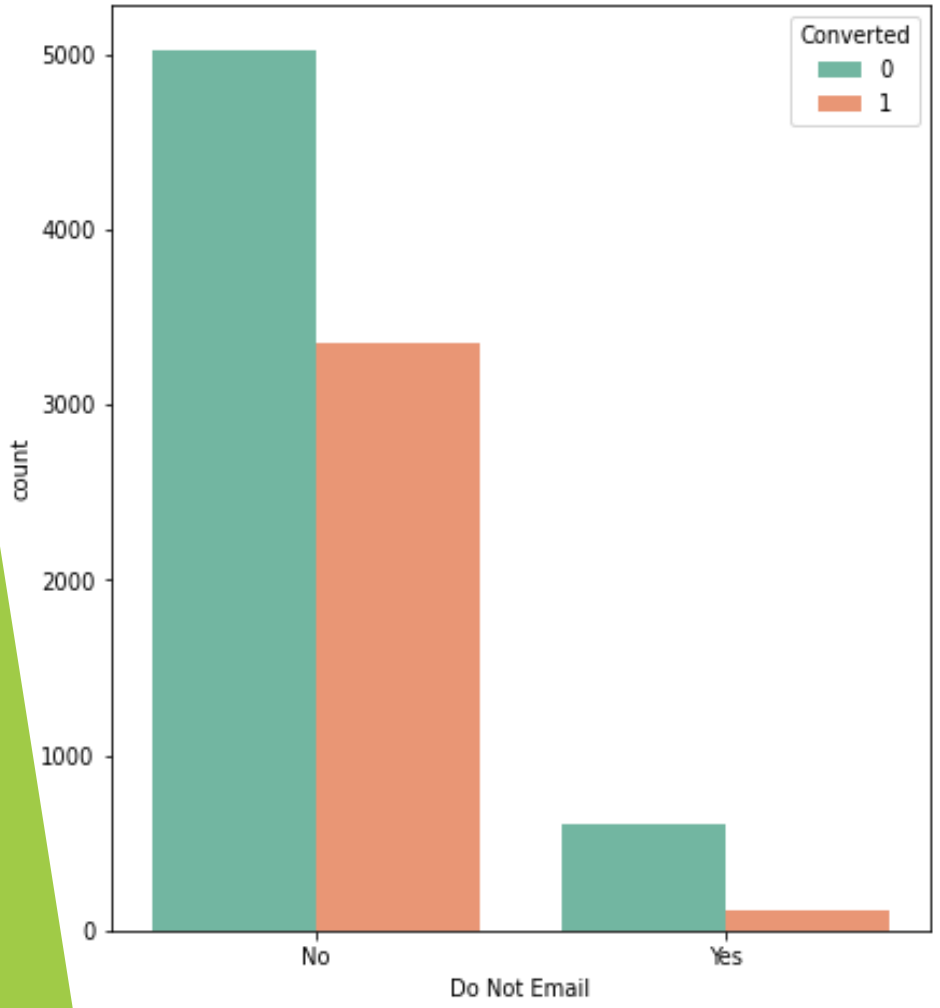
Count of Last Activity Variable

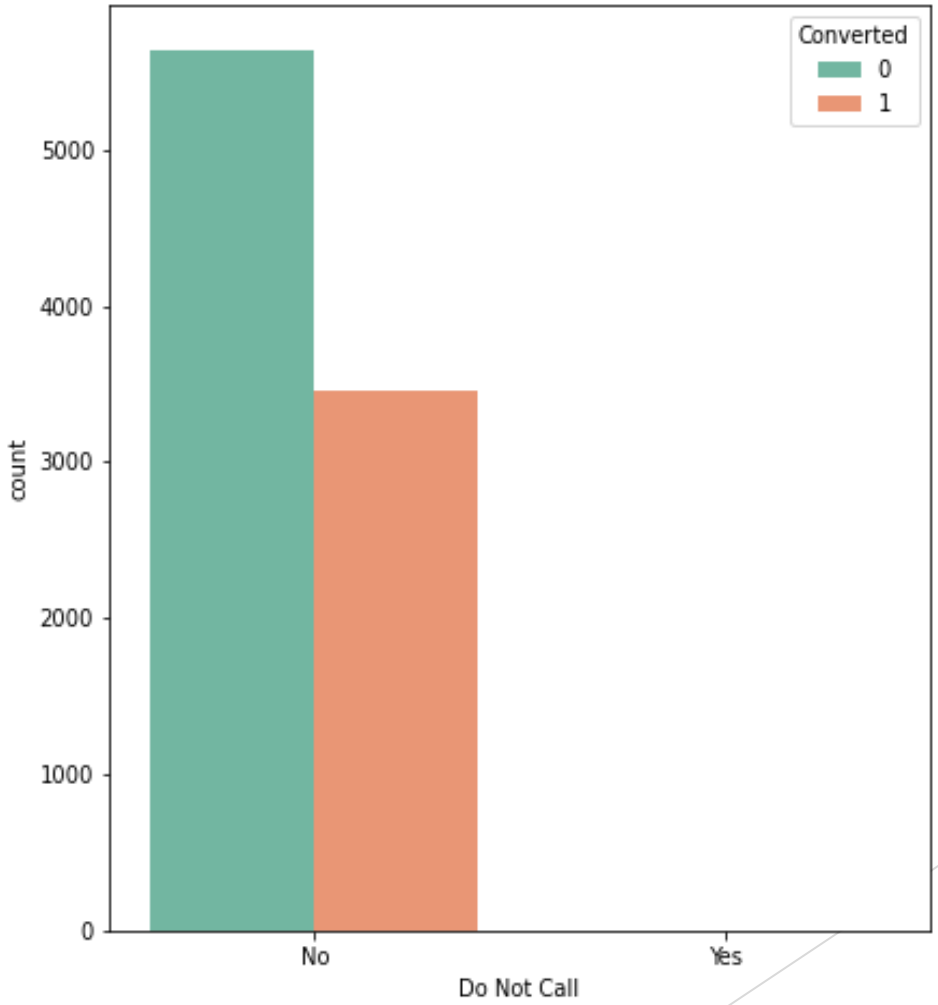


Interface :

- Maximum leads are generated having last activity as Email opened but conversion rate is not too good.
- SMS sent as last acitivity has high conversion rate.

Do Not Email & Do Not Call

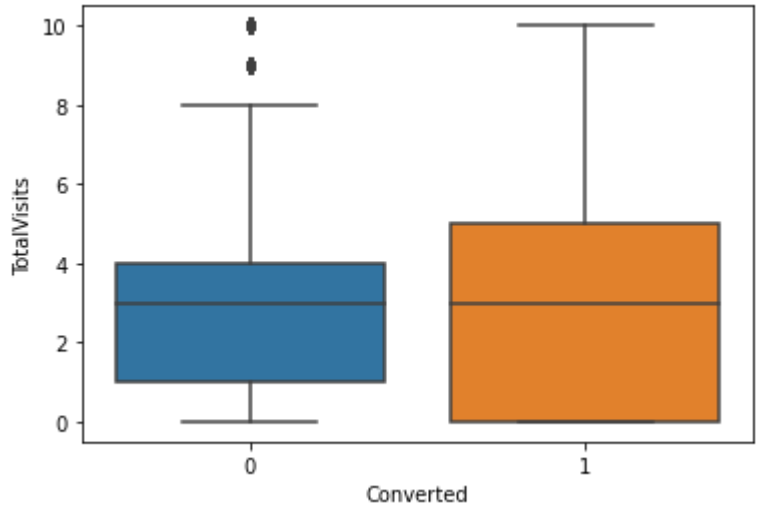




Interface :

- We can append the do not call column to the list of columns to be dropped data is higjly skewed

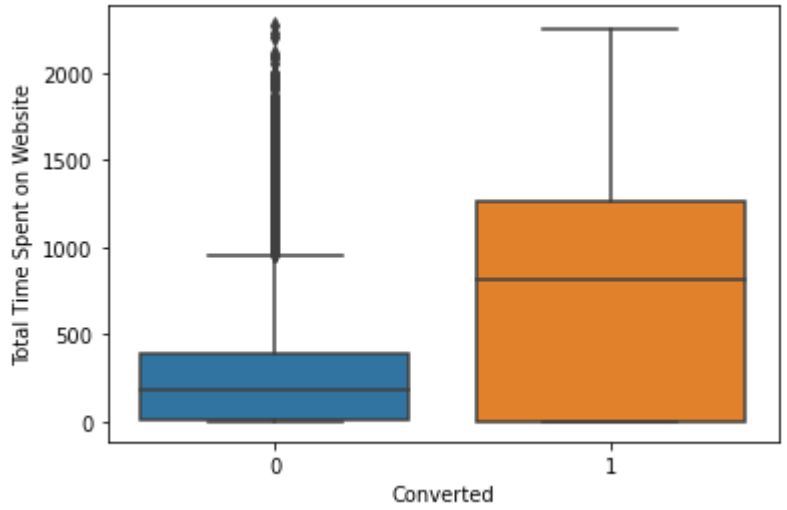
TotalVisits w.r.t Target Variable 'Converted'



Interface :

- As the median for both converted and non-converted leads are same , nothing coclusive can be said on the basis of variable TotalVisits

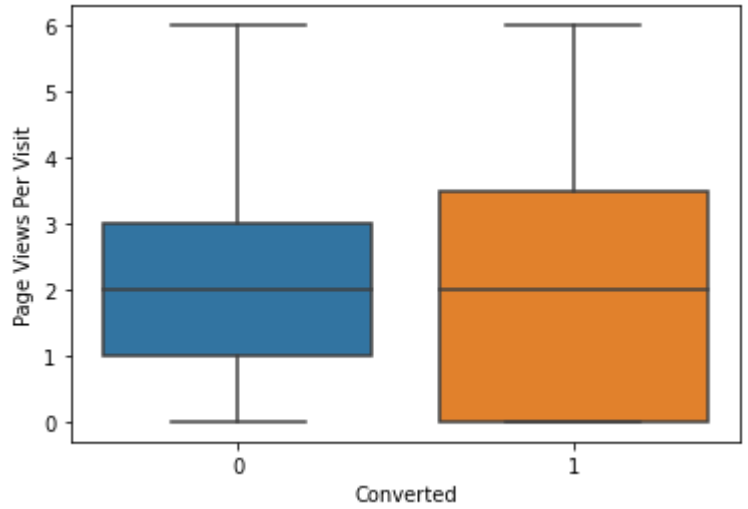
'Total Time Spent on Website' w.r.t Target Variable 'converted'



Interface :

- As can be seen, leads spending more time on website are more likely to convert , thus website should be made more enagaging to increase conversion rate

'Page Views Per Visit' w.r.t Target variable 'Converted'



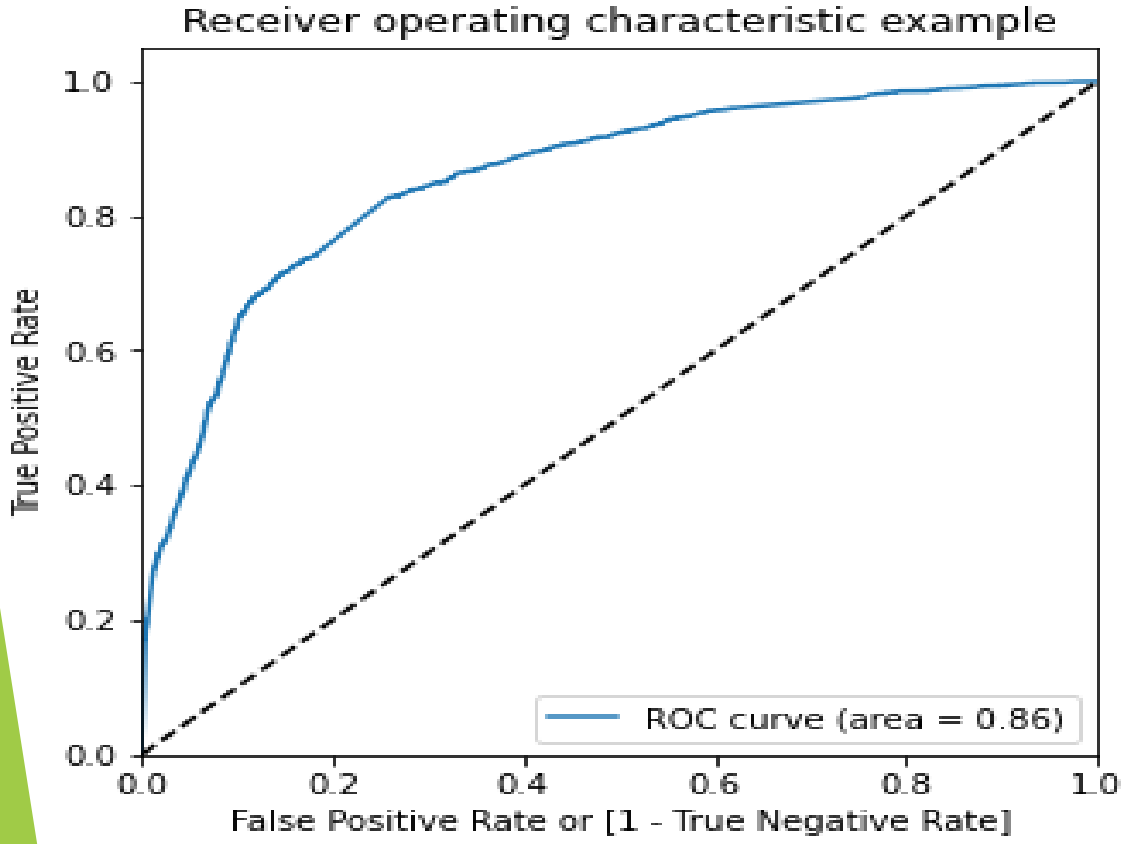
Interface :

- Median for converted and not converted leads is almost same.
- Nothing conclusive can be said on the basis of Page Views Per Visit.

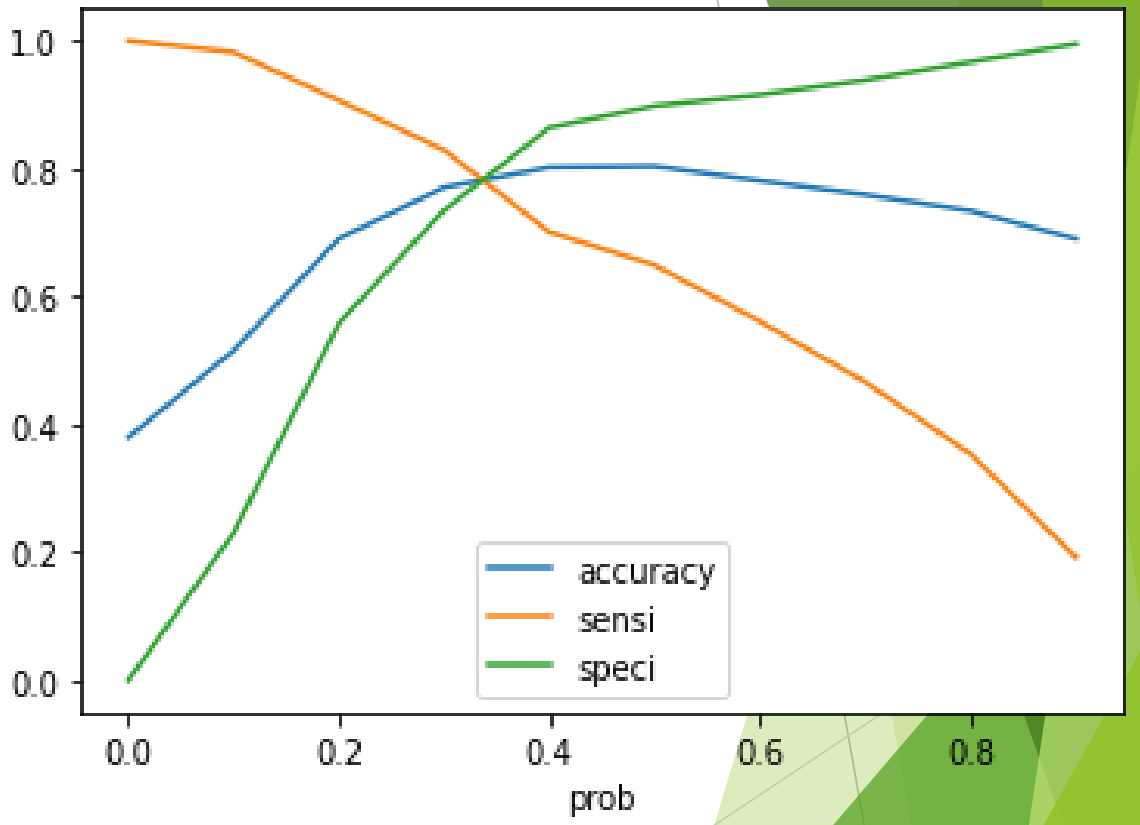
Model Building

- Splitting the Data into Training and Testing Sets.
- The first basic step for regression is performing a train-test split, we have chosen 70:30 ratio.
- Use RFE for Feature Selection.
- Running RFE with 15 variables as output.
- Building Model by removing the variable whose p-value is greater than 0.05 and VIF value is greater than 5.
- Predictions on test data set.
- Overall accuracy 81%.

ROC Curve



The ROC Curve should be a value close to 1. We are getting a good value of 0.86 indicating a good predictive model.



From the curve above, 0.3 is the optimum point to take it as a cutoff probability.

Conclusions

- After running the model on the Test Data these are the figures we obtain:
  - Accuracy : 77.52%
  - Sensitivity :83.01%
  - Specificity : 74.13%
- While 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 77%, 83% and 74% which are approximately closer to the respective values calculated using trained set.
- Also the lead score calculated in the trained set of data shows the conversion rate on the final predicted model is around 80%
- Hence overall this model seems to be good.

Recommendations

- The company should make calls to the leads coming from the lead sources "Welling Websites" and "Reference" as these are more likely to get converted.
- The company should make calls to the leads who are the "working professionals" as they are more likely to get converted.
- The company should make calls to the leads who spent "more time on the websites" as these are more likely to get converted.
- The company should make calls to the leads coming from the lead sources "Olark Chat" as these are more likely to get converted.
- The company should make calls to the leads whose last activity was SMS Sent as they are more likely to get converted.

Recommendations

- The company should not make calls to the leads whose last activity was "Olark Chat Conversation" as they are not likely to get converted.
- The company should not make calls to the leads whose lead origin is "Landing Page Submission" as they are not likely to get converted.
- The company should not make calls to the leads whose Specialization was "Others" as they are not likely to get converted.
- The company should not make calls to the leads who chose the option of "Do not Email" as "yes" as they are not likely to get converted.

THANK YOU