LEAD SCORING CASE STUDY

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. 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.

The company requires you to build a model wherein you need to assign a lead score to 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%.

Goals of the Case Study:

Build a logistic regression model to assign a lead score between 0 and 100 to each of the leads which can be used by the company to target potential leads. A higher score would mean that the lead is hot, i.e. is most likely to convert whereas a lower score would mean that the lead is cold and will mostly not get converted.

Steps Involved In Solving The Case Study:

1. Importing and reading the dataset.
2. Inspecting the dataset.
3. Data preparation.
4. EDA.
5. Dummy variable creation.
6. Train-Test split.
7. Feature scaling.
8. Model building.
9. Plotting ROC curve.
10. Finding optimal cutoff point.
11. Making predictions on the test data.

1. Importing and Reading the Dataset:

- Importing the necessary libraries.
- Importing the “Leads.csv” file into python.

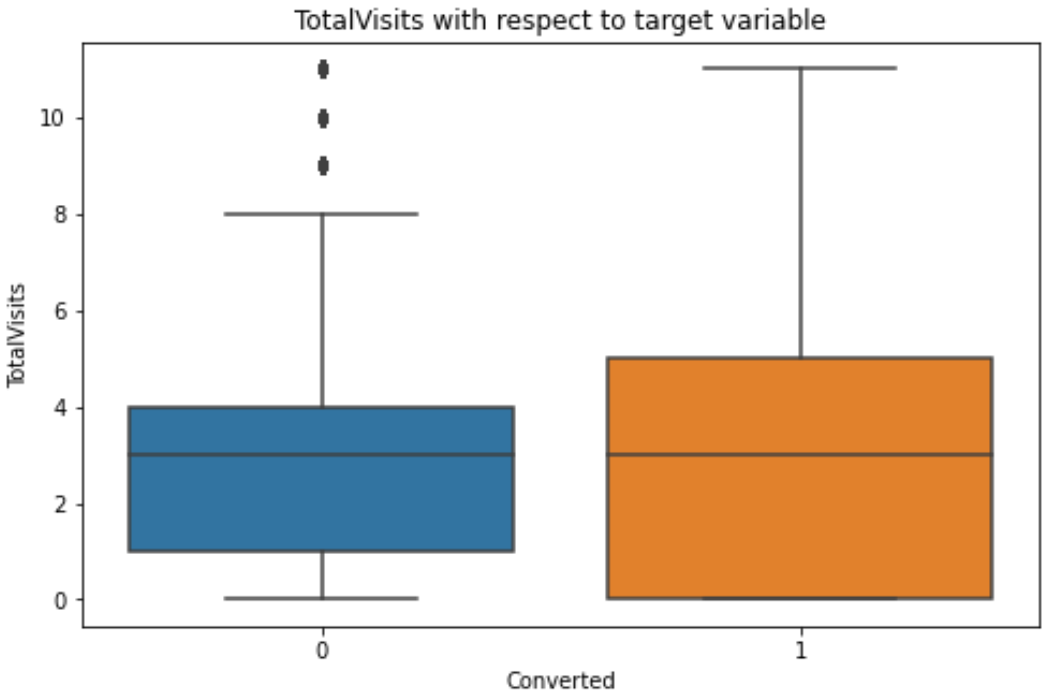
2. Inspecting the Dataset:

- Checking the head of the dataframe.
- Checking the shape of the dataframe.
- Checking the statistical aspect of the dataframe.
- Checking the basic info of each column.

3. Data Preparation:

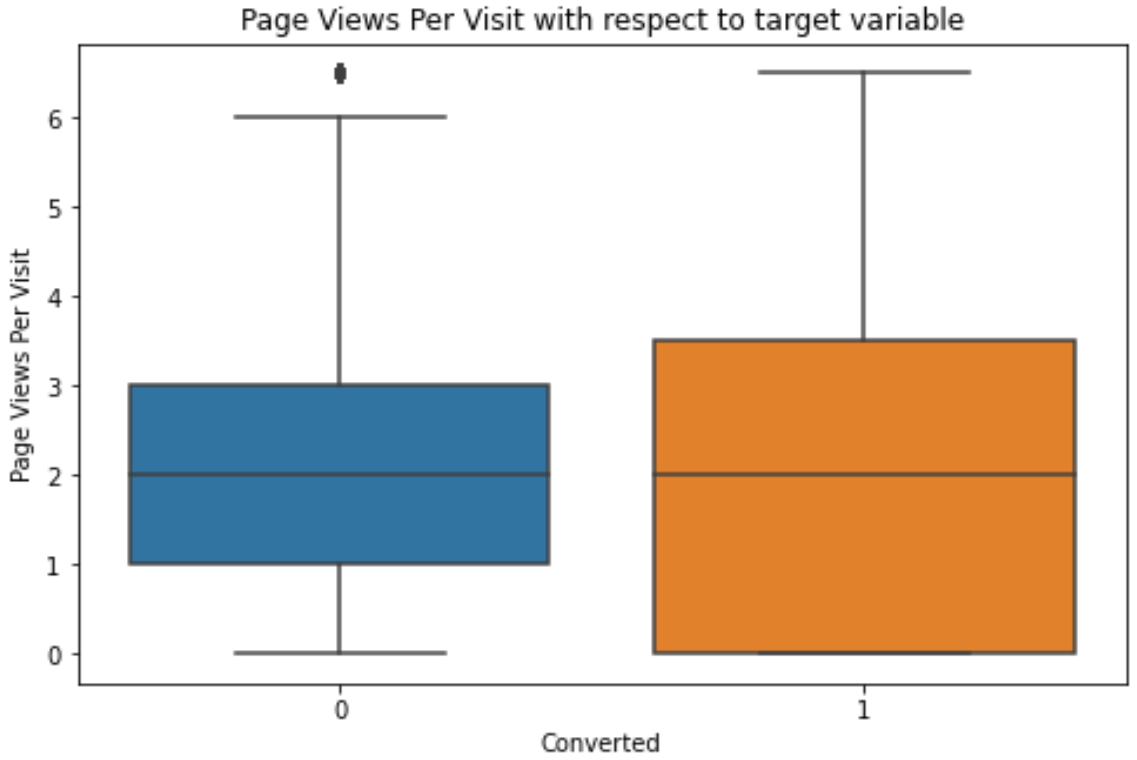
- Replacing the level “Select” in categorical variables with NaN values.
- Checking the percentage of missing values in each column.
- Dropping redundant columns and columns having missing value percentage greater than 40%.
- Replacing NaN values in certain columns with mode/new category values.
- Dropping Rows having Nan values in very small proportion.
- Dropping columns which had a particular level representing the categorical variable majorly.
- Checking for outliers using percentiles.
- Capping and flooring method used to treat outliers.

4. EDA:

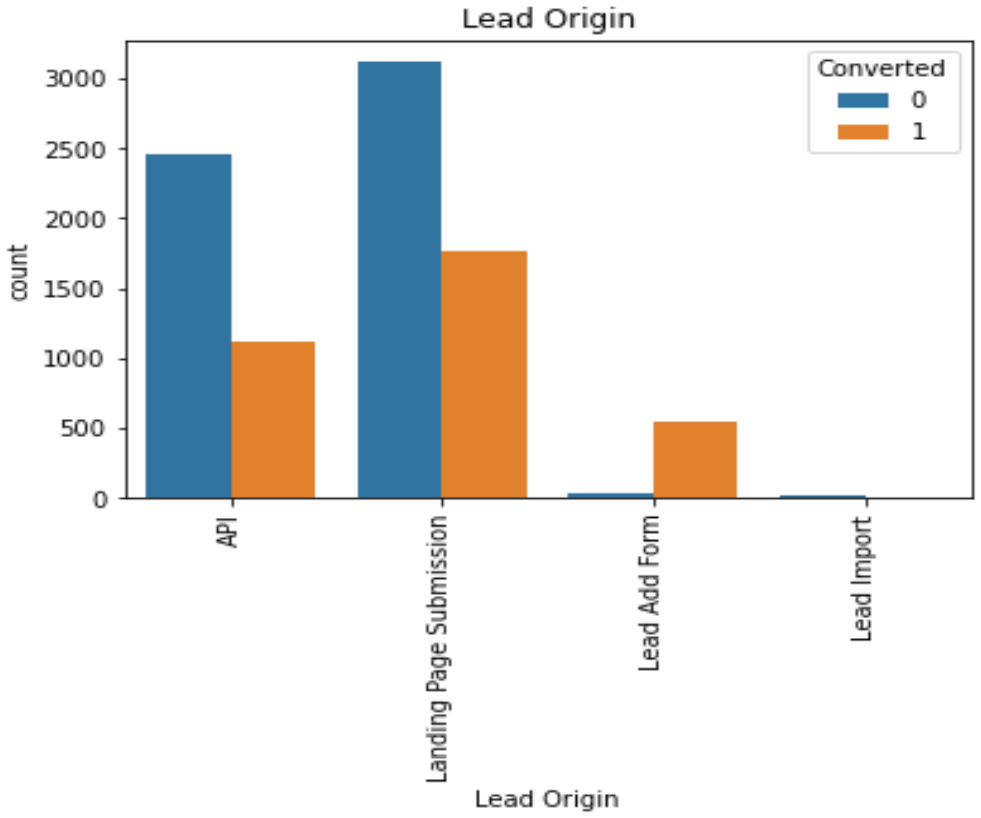


- The Median of “TotalVisits” for both non-converted leads as well as converted leads is same around 3.

- The Maximum value of “TotalVisits” for converted leads is higher than that of non-converted leads.

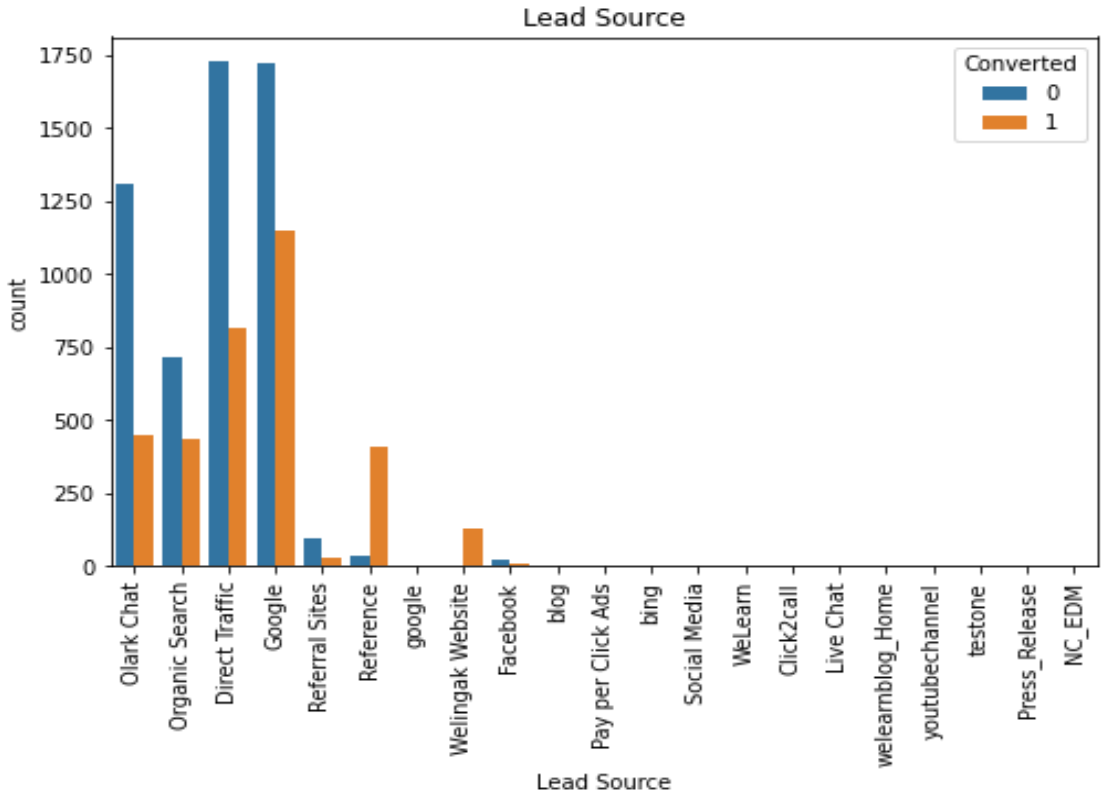


- The Median Number of “Page Views Per Visit” for both non-converted leads as well as converted leads is 2.



- “API” and “Landing Page Submission” are the Lead Origins that generate the most leads as well as converted leads.

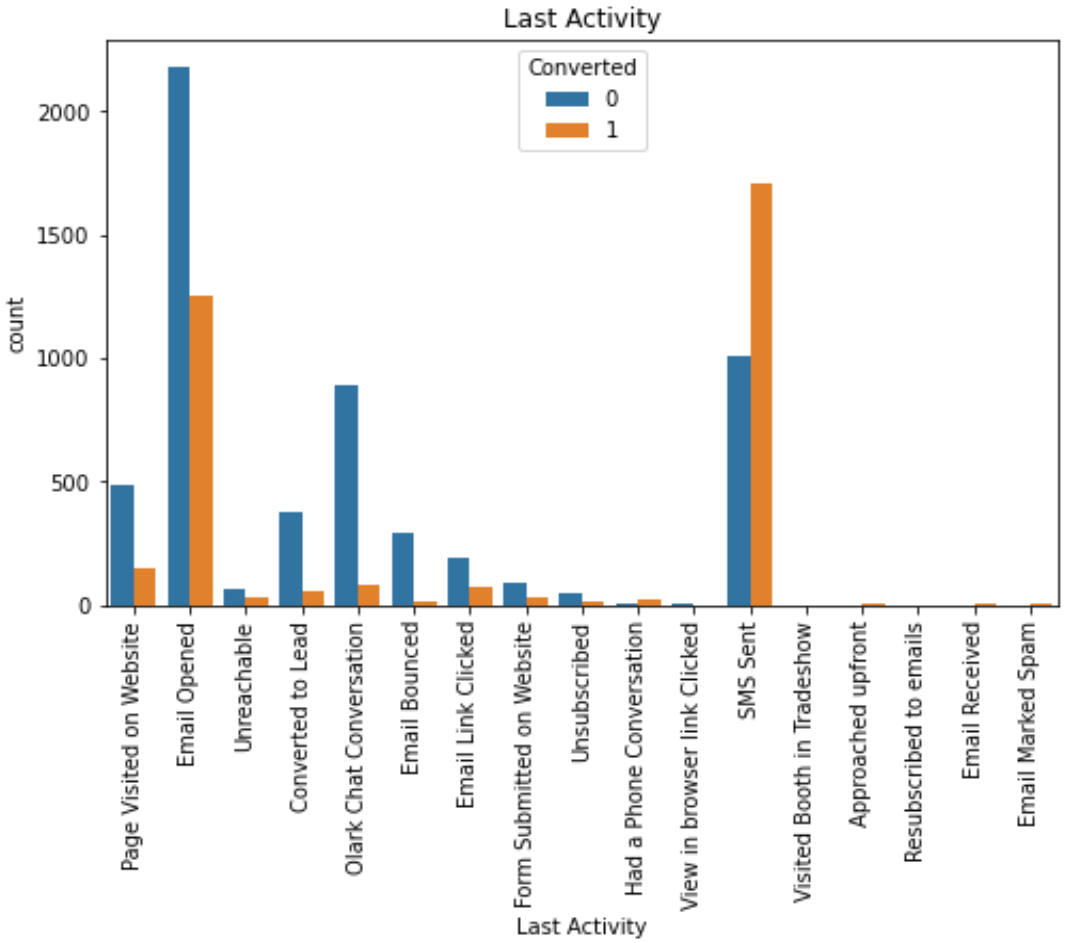
- Total Leads from “Lead Add Form” are very less in number but there is a higher conversion rate there.



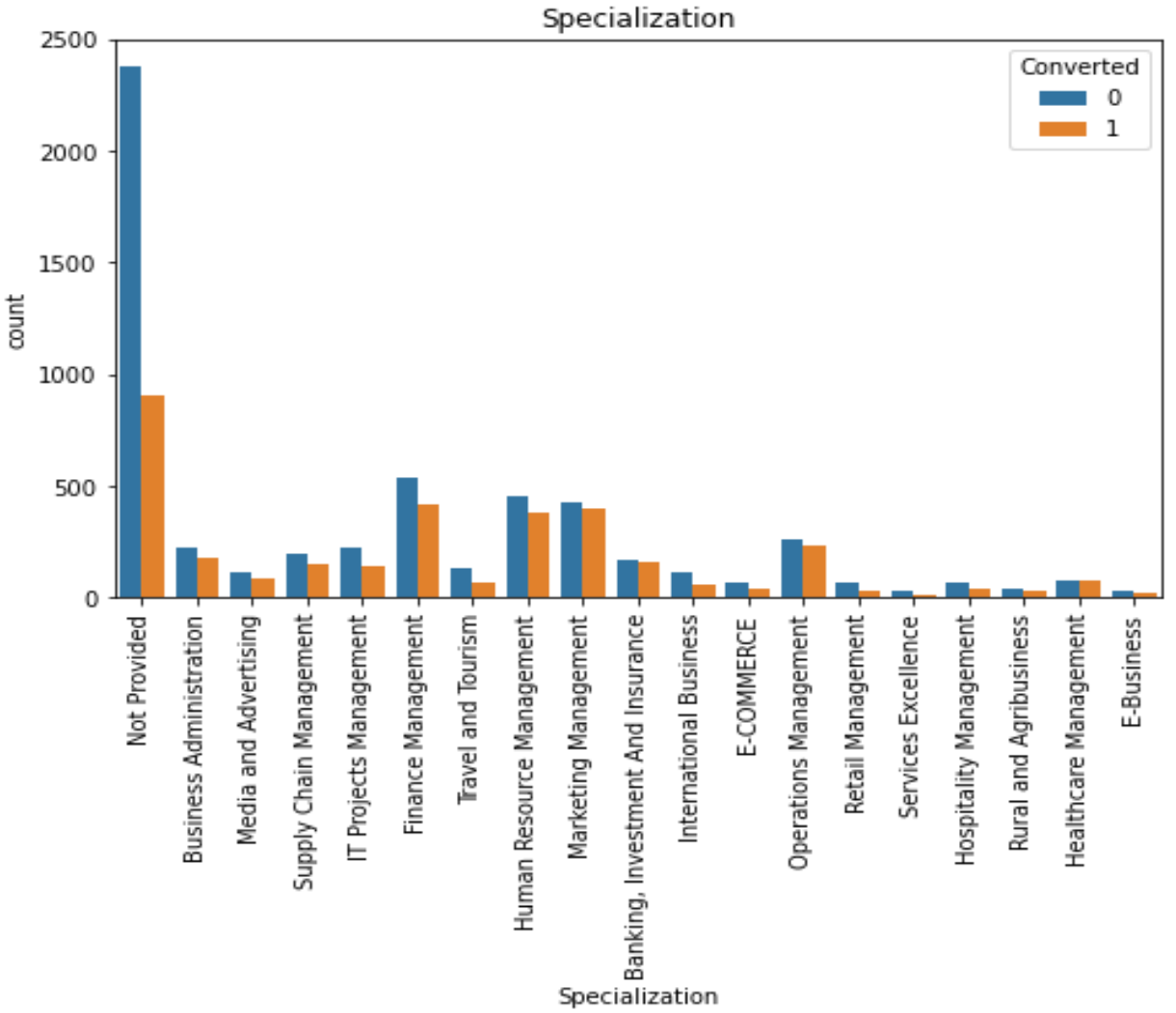
- Total number of leads generated by “Direct Traffic” and “Google” are the highest. They also have the highest converted leads.

- Leads through “Reference” are low in number, but they have a higher conversion rate.

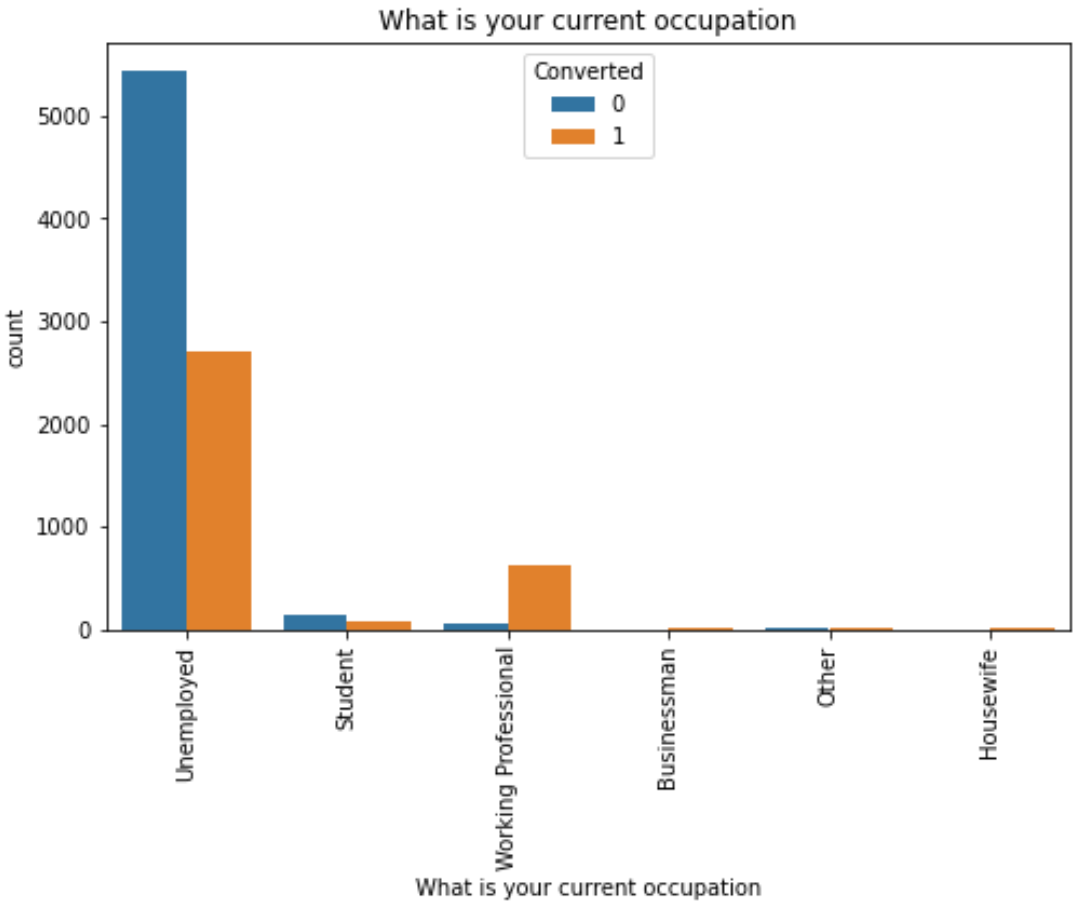
- Numbers of leads from “Olark Chat” and “Organic Search” are high, but their conversion rates are low.



- Customers whose Last Activity was either “Email Opened” or “SMS Sent” have a high chance of being turned to converted leads.

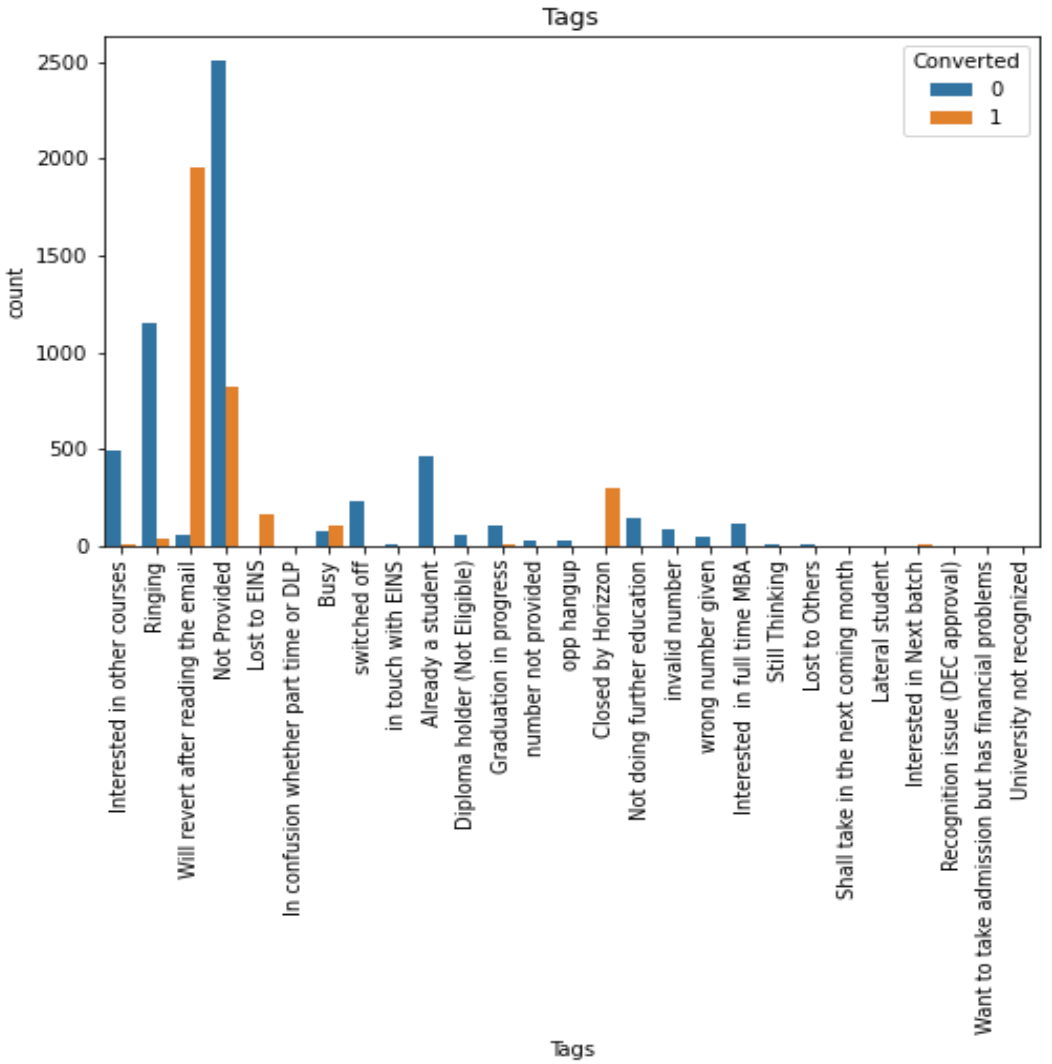


- Specialization having different type of Management in them are the ones having high number of leads as well as converted leads.

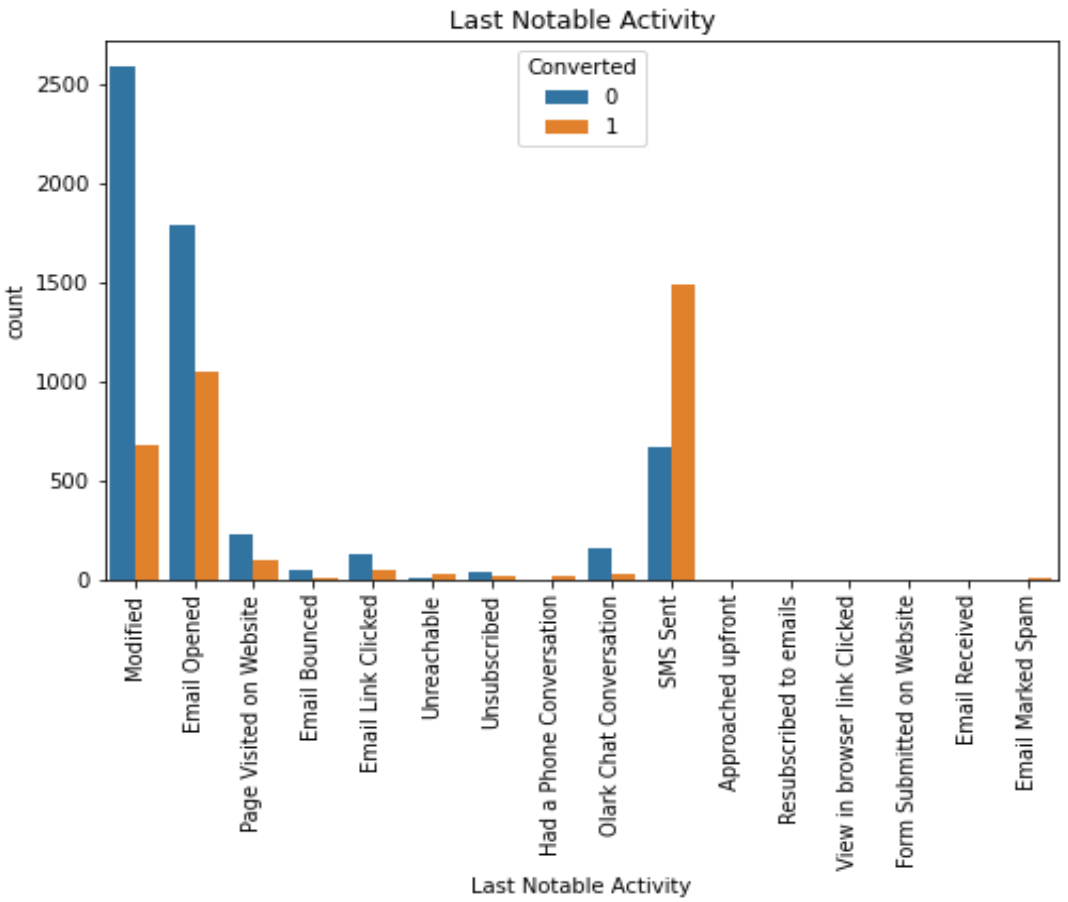


- “Unemployed” leads are highest in number. They also have the highest converted leads as well.

- “Working Professional” are low in number, but they have a high conversion rate.



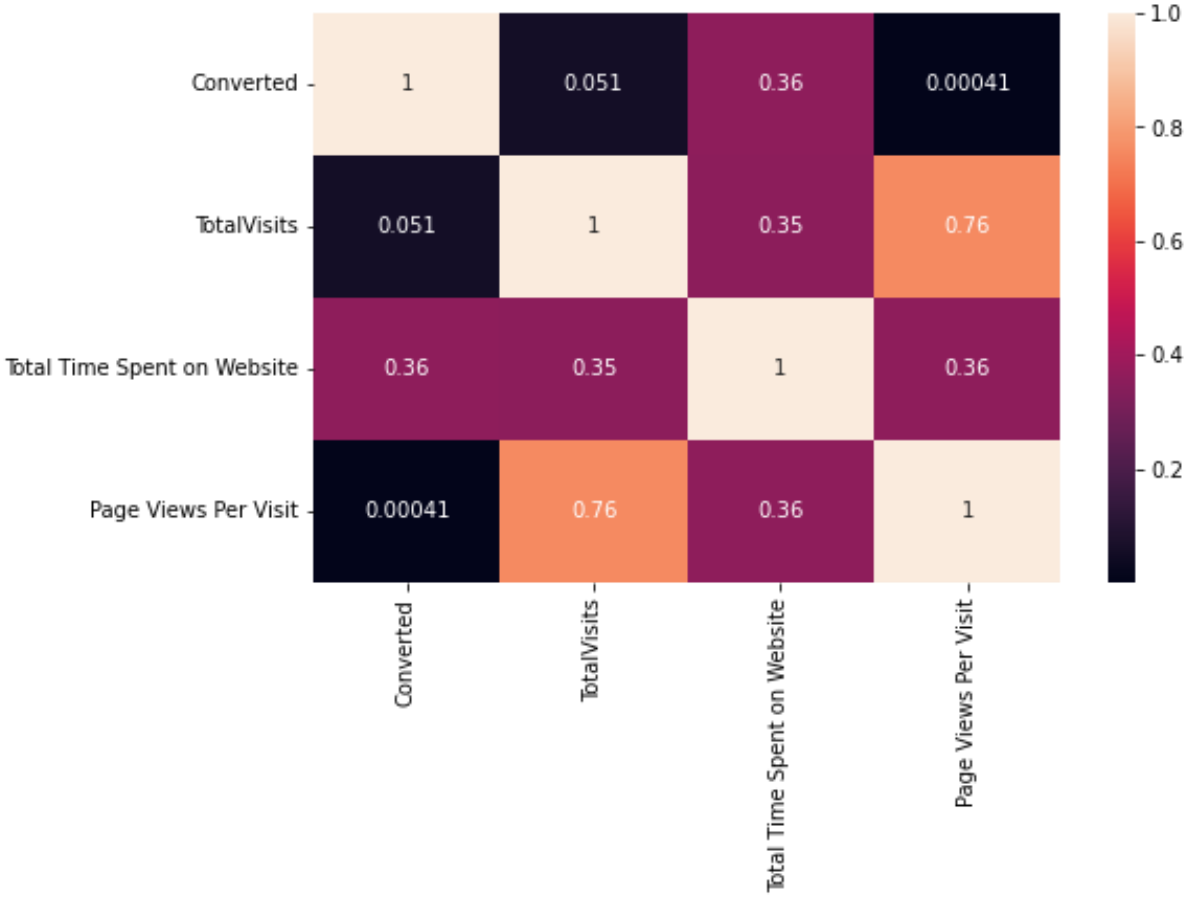
- Customers whose current status is “Will revert after reading the email” are most likely turned to converted leads.



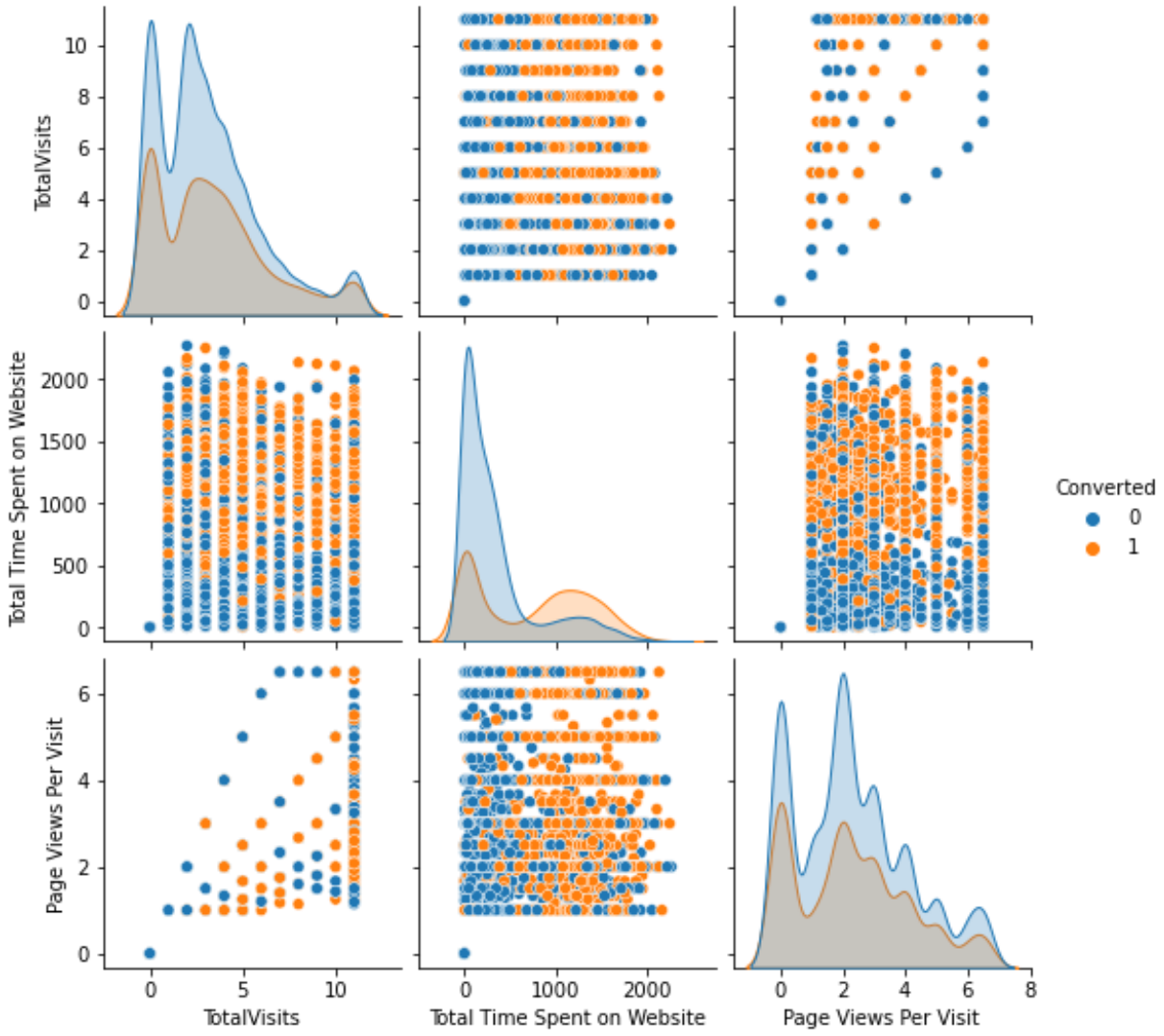
- Customers whose Last Notable Activity was either “Email Opened” or “SMS Sent” are having high number of leads. They also have higher converted leads.

- Customers having “Modified” as their Last Notable Activity have a very low conversion rate.

Correlation Heatmap:



Pairplot:


TotalVisits
Total Time Spent on Website
Page Views Per Visit
Converted
0
1

5. Dummy Variable Creation:

- Mapping Binary variables with two levels(yes & no) to 1’s and 0’s.
- Creating dummy variables for categorical variables using pd.get_dummies.
- Dropping categorical columns for which dummies have been created.

6. Train-Test Split:

- Importing train_test_split from sklearn.model_selection.
- Putting all feature variables in X.
- Putting Target variable in y.
- Splitting the dataset into train and test set in the ratio of 70:30.

7. Feature Scaling:

- Using Standard scaler from sklearn.preprocessing.
- Scaling numerical features using Standard scaler.

8. Model Building:

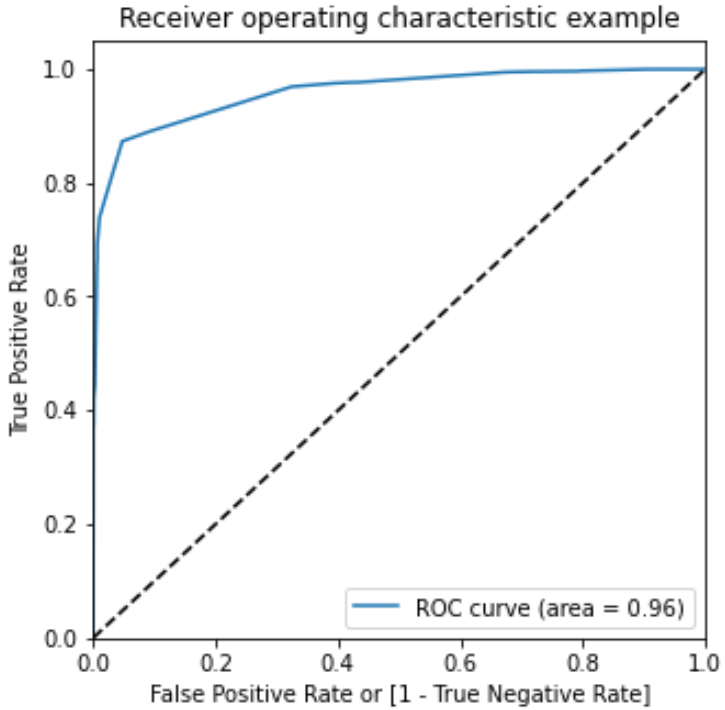
- Using RFE(recursive feature elimination) to select top 15 variables for our model.
- Building a model with the features selected by RFE using StatsModels.
- Dropping feature with a high p-value.
- Checking VIF value for features, and dropping the feature with a high VIF value.
- We repeated the model building process till we arrived to a model with features having normal p-value and VIF values.

- Predicting values on the train set.

- Creating a confusion matrix.

- Checking accuracy, sensitivity and specificity.

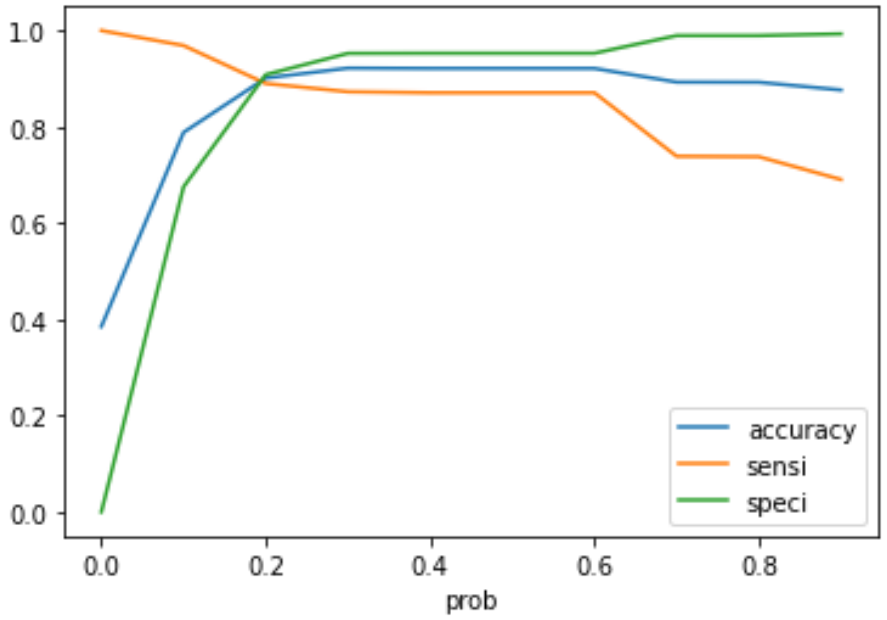
9. Plotting ROC curve:



- A Good ROC curve is the one which touches the upper left corner of the graph. we have a similar curve over here.

- Higher the area under the ROC curve the better is your model. The value of ROC curve should be closer to 1, we have the area under ROC curve = 0.96.

10. Finding Optimal Cutoff Point:



- From the plot we decided to take a cutoff point of 0.2.
- Checking accuracy after taking 0.2 cutoff.
- Creating a confusion matrix.
- Checking sensitivity, specificity and false positive rate.
- Checking precision and recall scores.

11. Making Predictions on the Test Set:

- Scaling the variables in test set.
- Predicting values on the test set.
- Checking the accuracy.
- Creating confusion matrix.
- Checking sensitivity and specificity.
- Checking the precision and recall scores.

Results:

Final Comparison of Observation values between Train and Test dataset.

Train Dataset.

- Accuracy : 90%
- Sensitivity : 89%
- Specificity : 91%

Test Dataset.

- Accuracy : 89%
- Sensitivity : 87%
- Specificity : 90%

- We have very close values for accuracy, sensitivity and specificity when comparing results from train and test sets.
- These values show that the model is performing well.

Summary:

- The top three variables contributing most towards the probability of a lead getting converted are:
  1. Tags_Closed by Horizzon.
  2. Tags_Lost to EINS.
  3. Tags_Will revert after reading the email.

- The top three categorical variables contributing most towards the probability of a lead getting converted are:
  1. Tags_Closed by Horizzon.
  2. Tags_Lost to EINS.
  3. Tags_Will revert after reading the email.

- We have chosen a cutoff point of 0.2.

- We have the are under the ROC curve = 0.96 which shows that our model is performing well.

- Accuracy, sensitivity and specificity for train set are 90%, 89%, 91% respectively.

- Accuracy, sensitivity and specificity for test set are 89%, 87%, 90% respectively.

Conclusions:

- “API” and “Landing Page Submission” are the Lead Origins that generate the most leads as well as converted leads.
- Total number of leads generated by “Direct Traffic” and “Google” are the highest. They also have the highest converted leads.
- Leads through “Reference” are low in number, but they have a higher conversion rate.
- Customers whose Last Activity was either “Email Opened” or “SMS Sent” have a high chance of being turned to converted leads.
- “Working Professional” have a high conversion rate, meaning they are more likely to be converted.
- Customers whose Last Notable Activity was either “Email Opened” or “SMS Sent” are having high number of leads. They also have higher converted leads.
- Keep approaching people who spend more time on the website.
- Emphasize more on leads from lead source – Olark Chat, as the total number leads are high but conversion rate is low.

THANK YOU!

Submitted By:
Shyam Dalsaniya
Iranna Chatti