Lead Scoring Case Study

Submitted By:
Harsh Deep Jaggi
DSC 57

Objective

The company is required to build a model wherein you need to assign a lead score to each lead such that the customers with a higher lead score have a higher conversion chance and the customers with a 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%.

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.

X Education has appointed you to help them select the most promising leads, i.e. the leads that are most likely to convert into paying customers. 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 a higher lead score have a higher conversion chance and the customers with a 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%.

Business Objective

This case study aims to identify patterns that indicate if a client has difficulty paying their instalments which may be used for taking actions such as denying the loan, reducing the amount of the loan, lending (to risky applicants) at a higher interest rate, etc. This will ensure that the consumers capable of repaying the loan are not rejected. Identification of such applicants using EDA is the aim of this case study.

In other words, the company wants to understand the driving factors (or driver variables) behind loan default, i.e. the variables that are strong indicators of default. The company can utilise this knowledge for its portfolio and risk assessment.
To develop your understanding of the domain, you are advised to independently research a little about risk analytics - understanding the types of variables and their significance should be enough.

Datasets

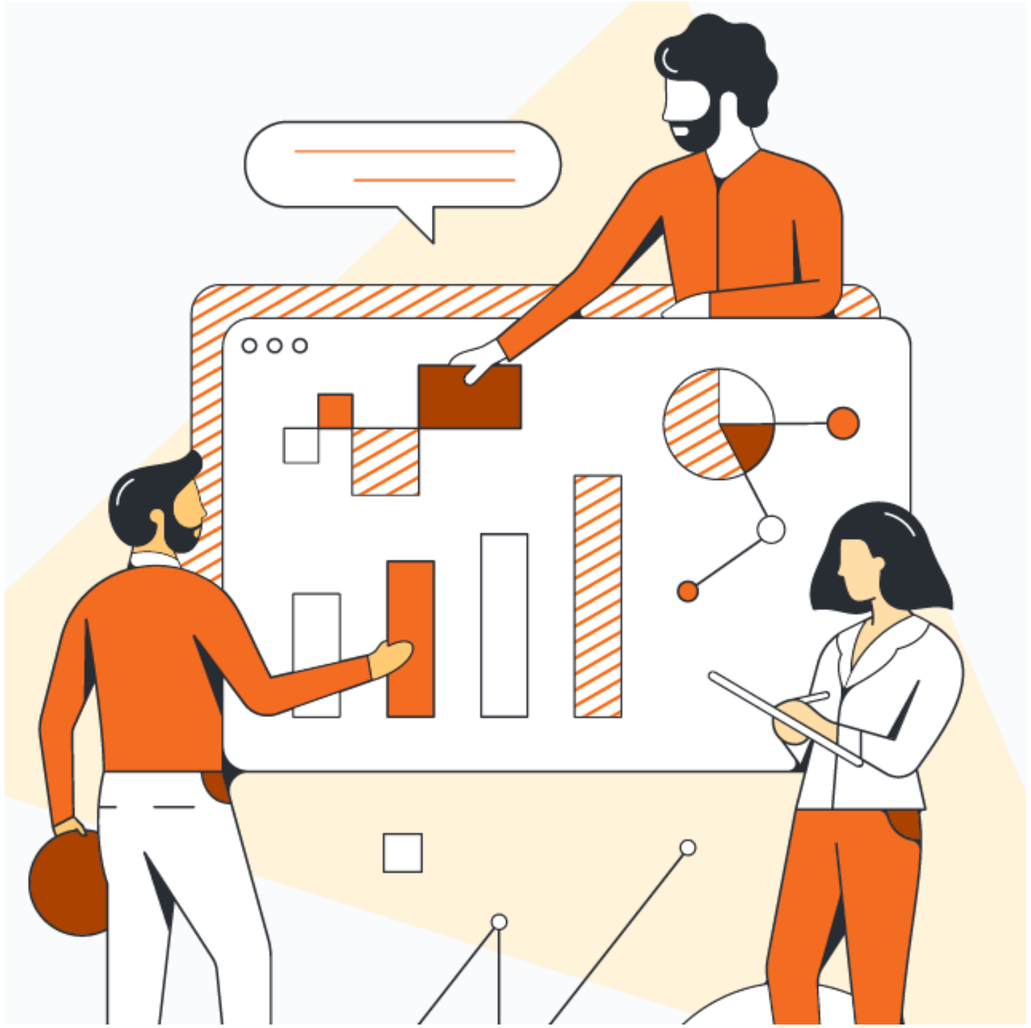
This dataset has 1 file as explained below:

'leads.csv'

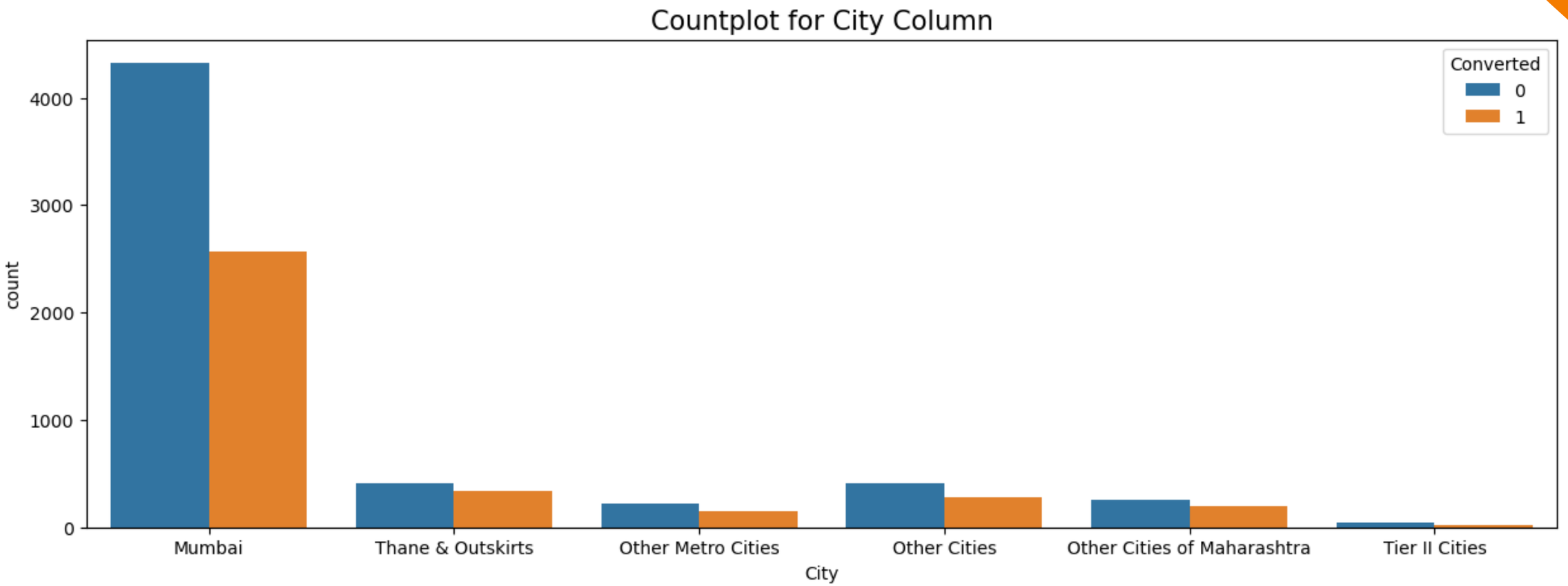
contains information about the leads. All the data of leads are in this file.

Steps:

- Understanding business problem
- Importing the data
- Understanding the data
- Check the structure of the data
- Data Transformation
- EDA:
- Dummy Variable Creation
- Model Building
- Recursive Feature Elimination
- Variance Inflation Factor (VIF)
- ROC Curve
- Confusion Matrix
- Conclusion

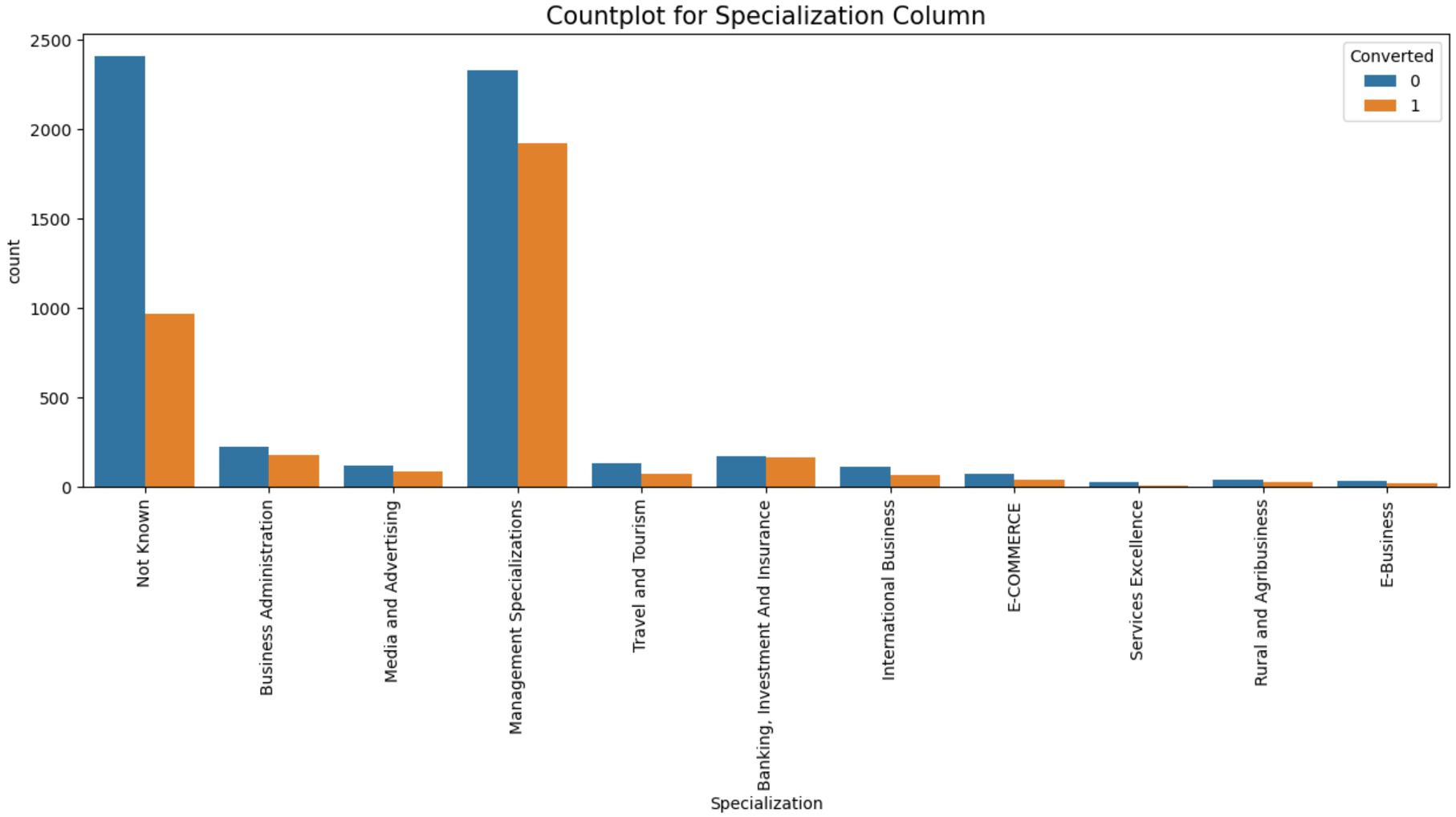


Analysis of City Column



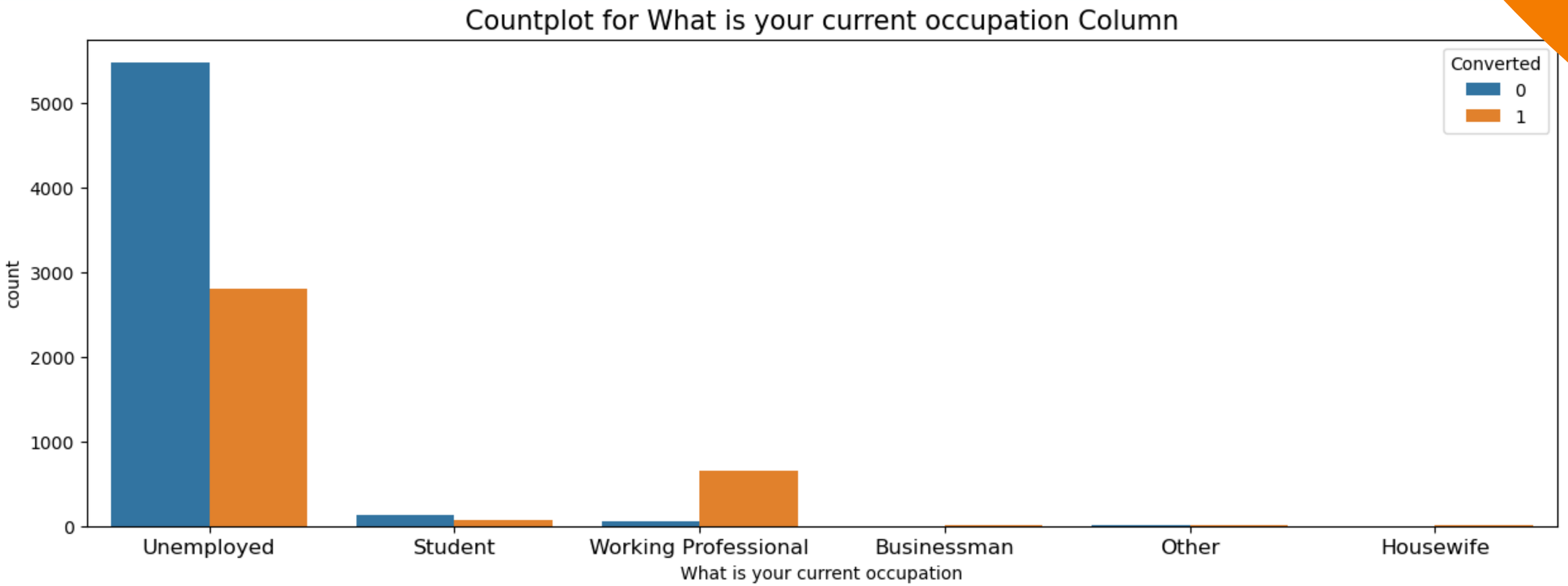
1. Most leads are generated from Mumbai parts.
2. Company should not focus on “Tier II Cities”.

Analysis of Specialization Column



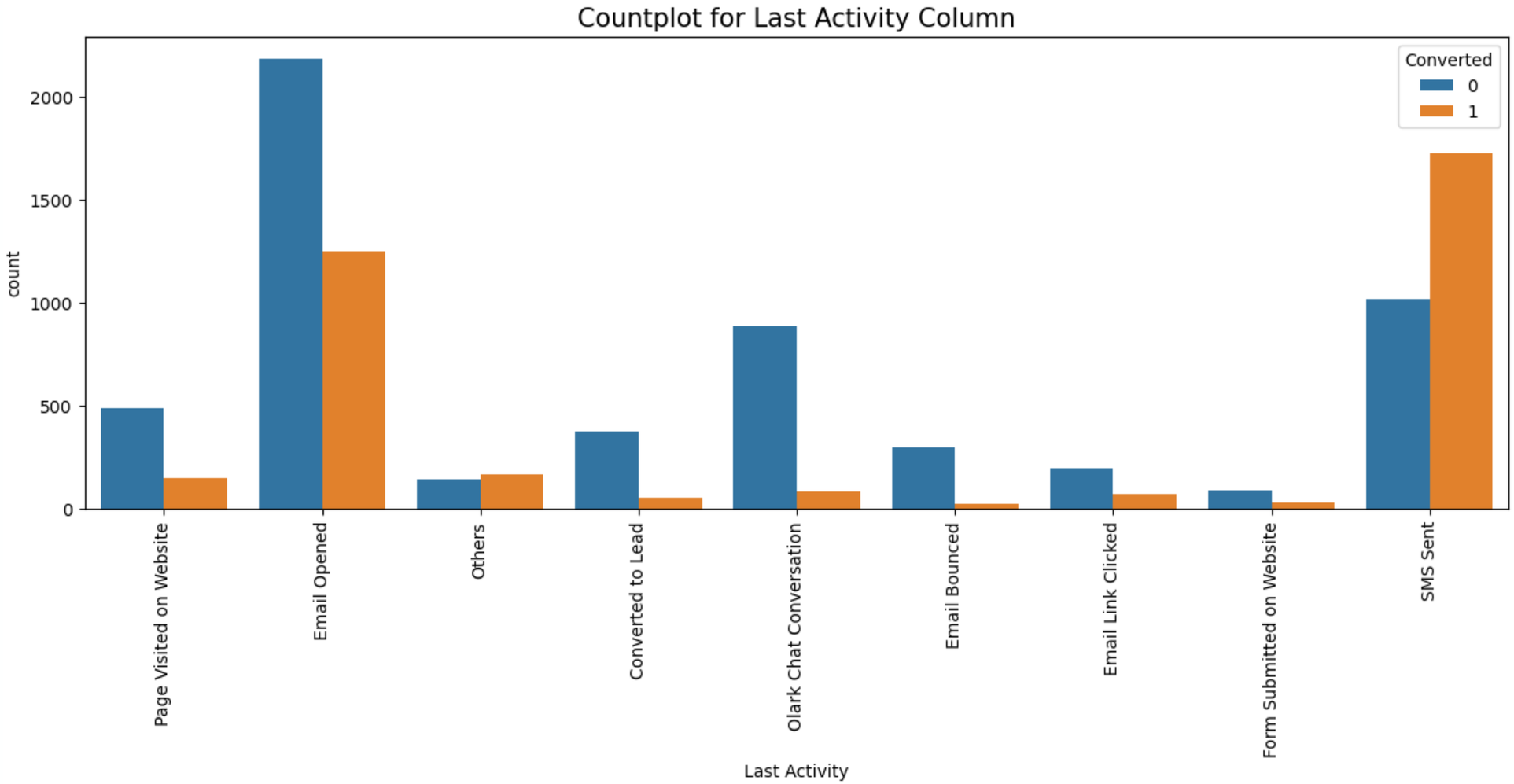
1. I've combined multiple management values into the 1 specialization named "management specialization".
2. "Management Specialization" generates higher traffic (when combined)
3. The company should focus on "Banking, Investment & Insurance" sector.

Analysis of What is your current occupation



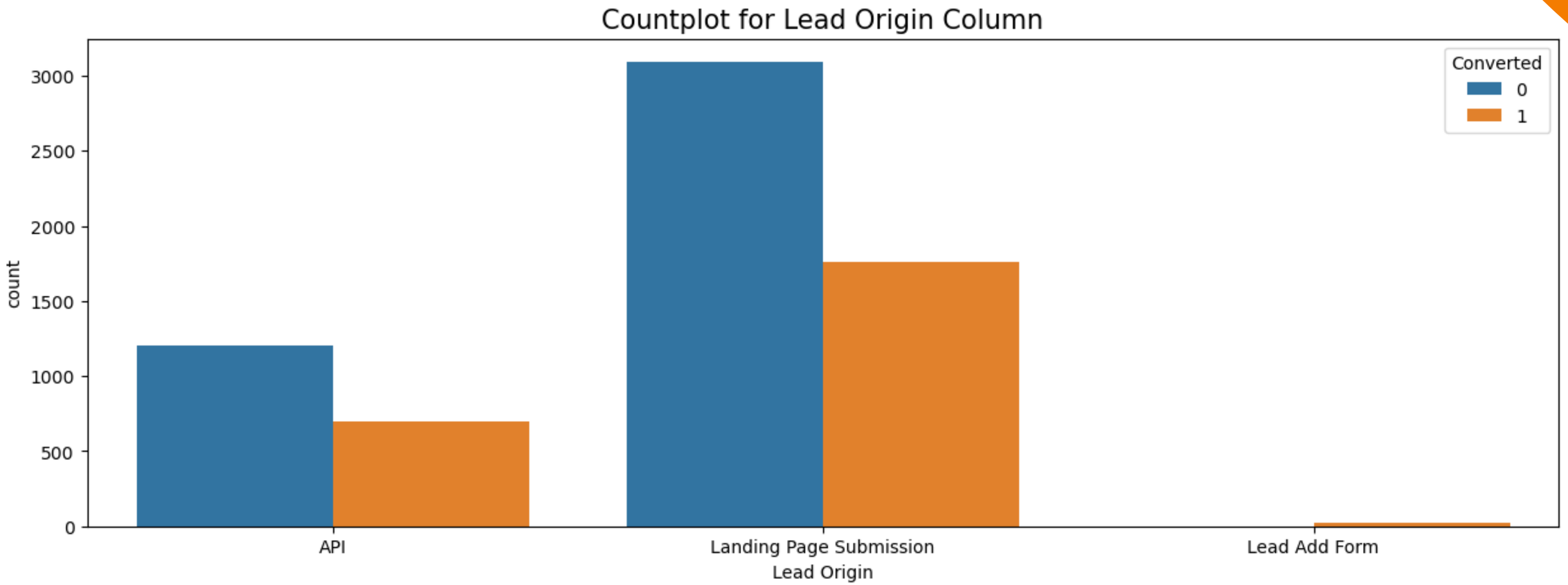
1. “Working Professional” have higher conversion rate.
2. “Unemployed” have higher leads generated but it should be improved.

Analysis of Last Activity Column



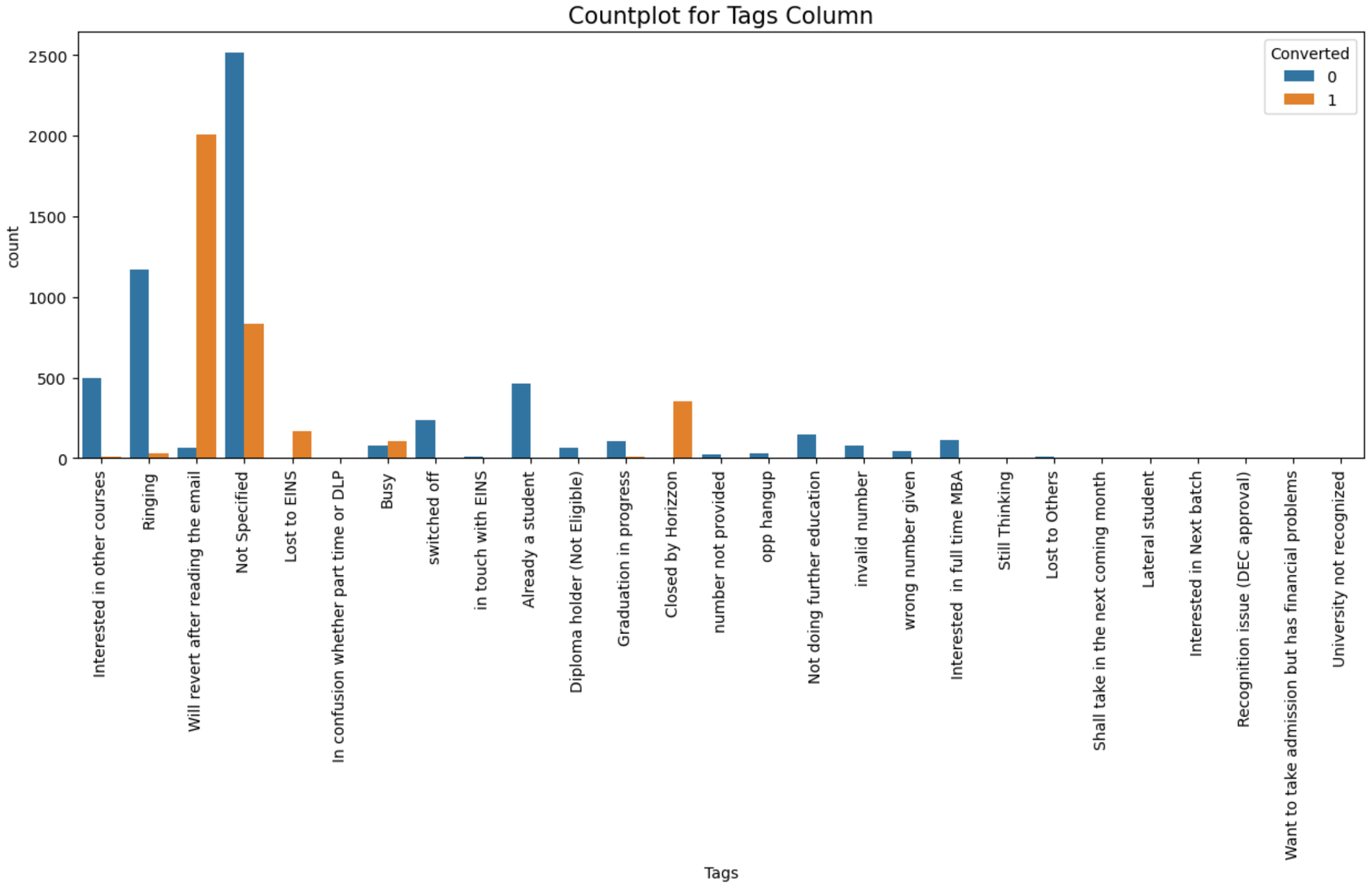
1. “Email Opened” generates higher traffic.
2. Conversion Rate of “SMS Sent” is higher.
3. The Company should work on “Olark Chat Conversation” to improve their conversions as it has great traffic generated.

Analysis of Lead Origin Column



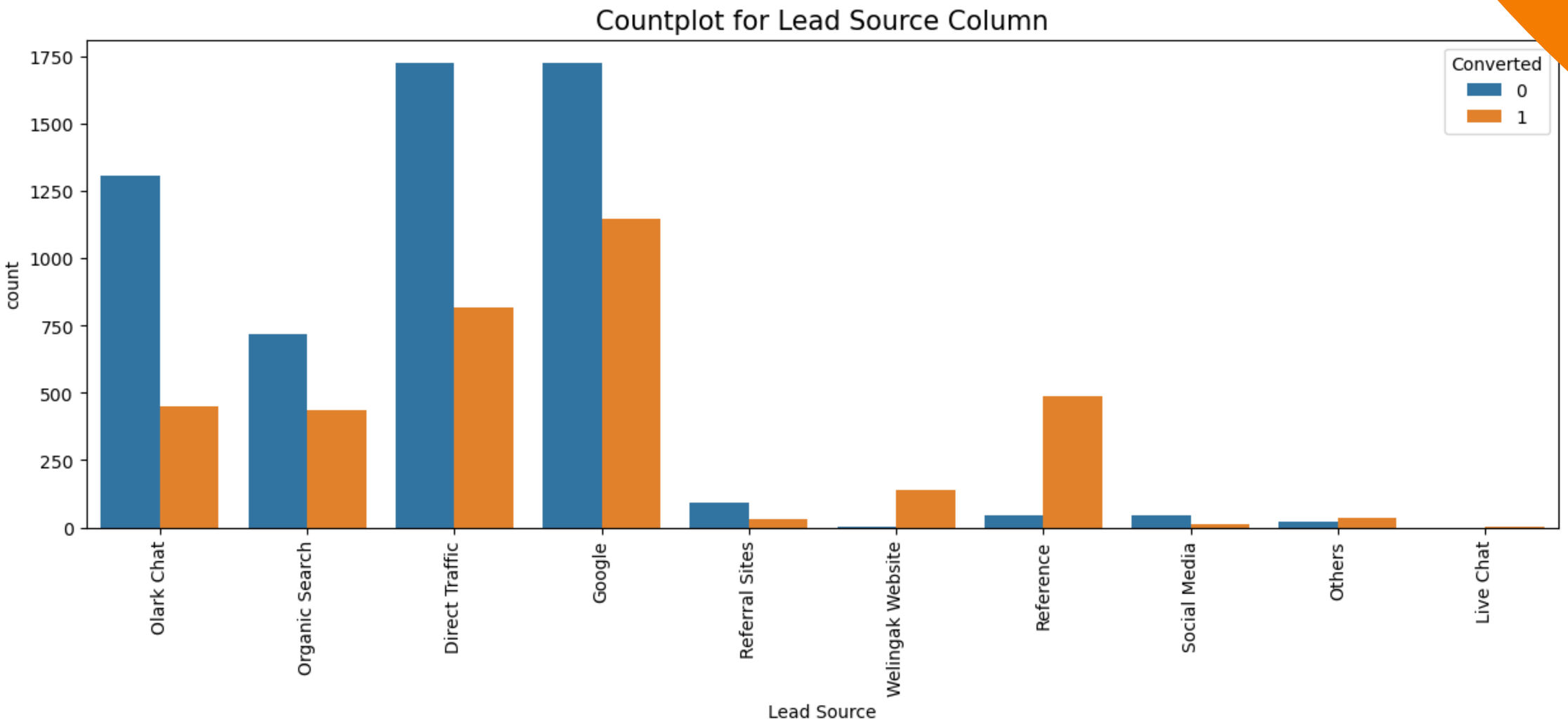
1. Leads generated from “Landing Page Submission” are higher & higher conversions too.
2. Leads generated from “Lead Add Form” are the least but conversions are more.

Analysis of Tags Column



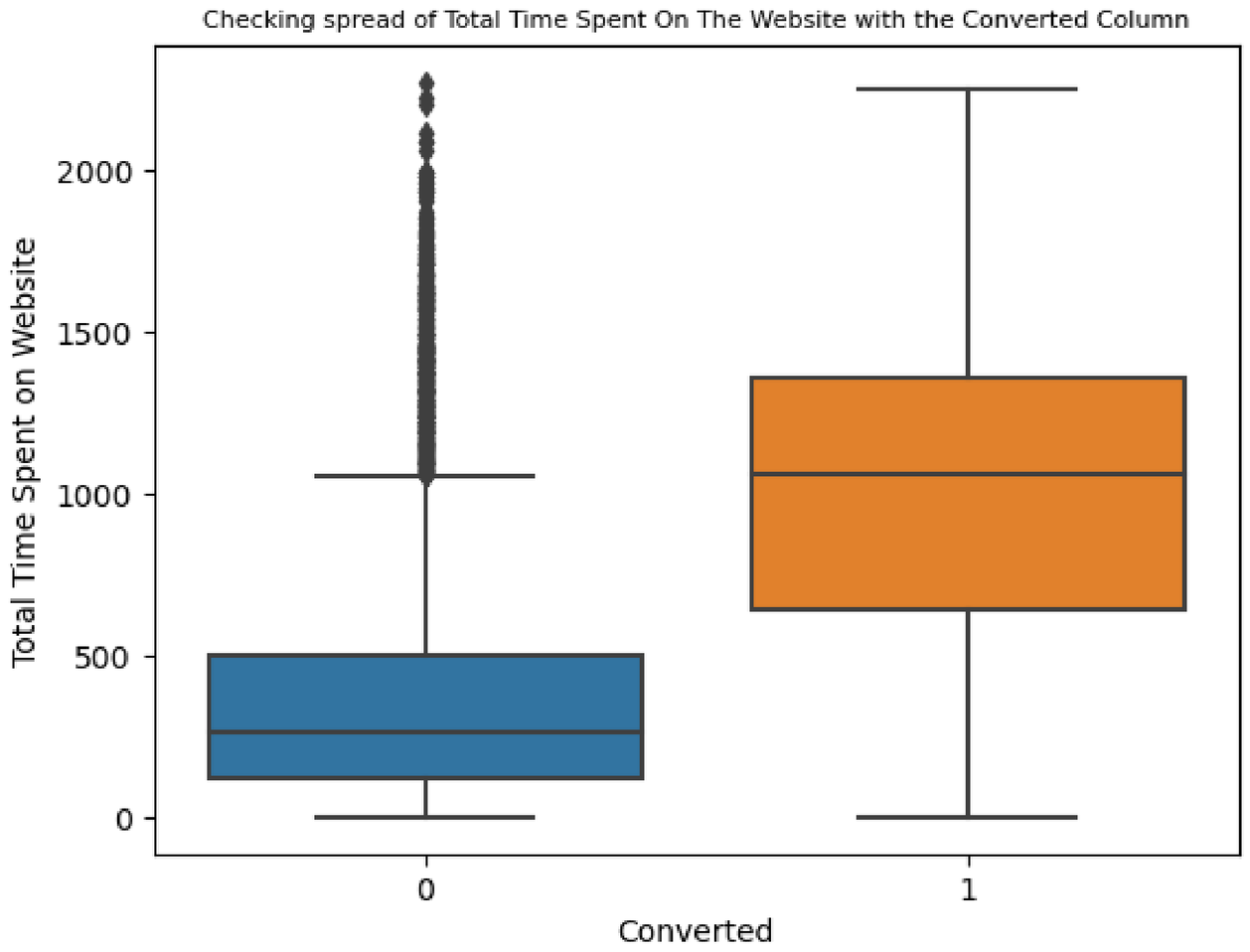
1. Most of the people chose not specified in the tags column.
2. most of the leads are converted from the tag which have “will revert after reading the mail”.

Analysis of Lead Source Column



1. I've combined two entries of google.
2. Direct Traffic & Google generates higher traffic.
3. Conversion Rate of "Reference" & "Welingak Website" are higher

Analysis of Total Time Spent on Website With Converted



1. People who spent more time on website likely to convert more.

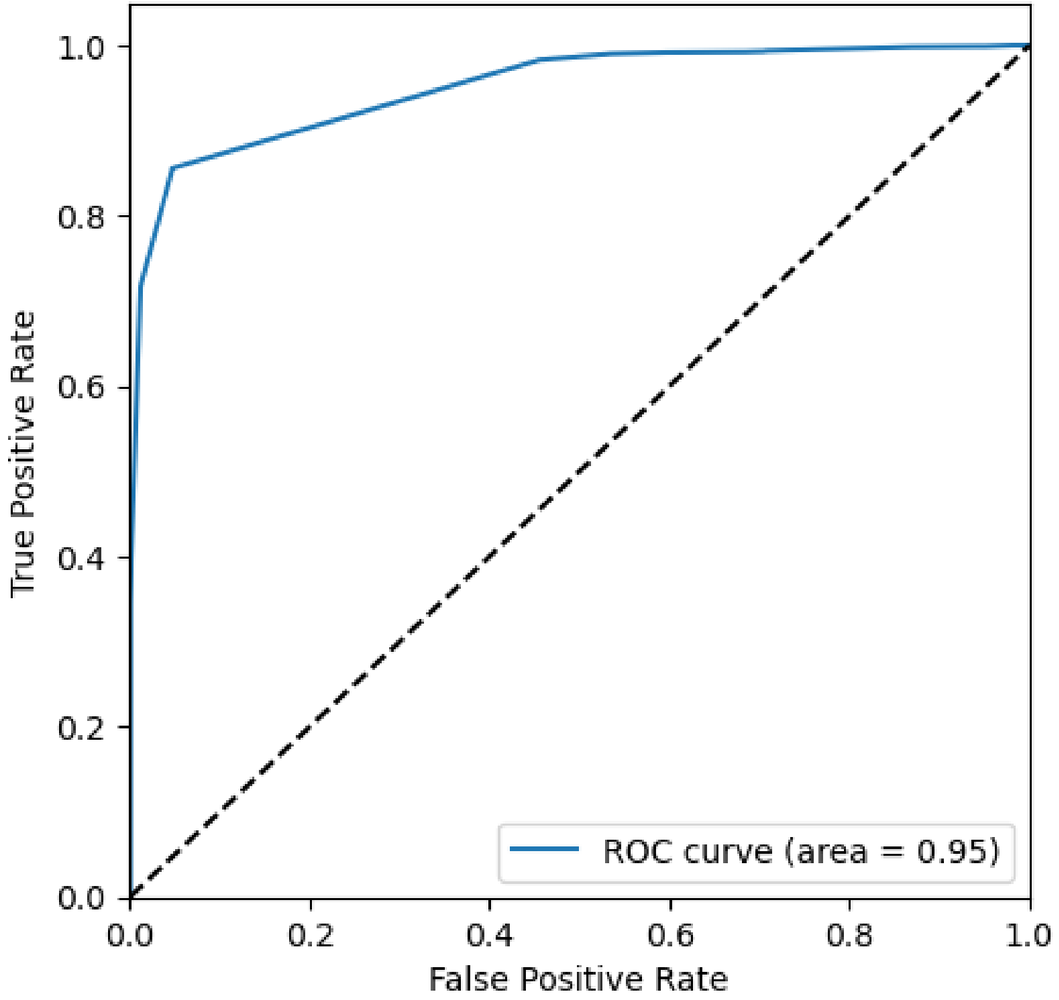
Data Preparation For Model Building

1. After EDA, I'm left with 14 Columns.
2. Then I've created Dummy Variables.
3. I've split this dataset into Train & Test Dataset.
4. Performed the feature scaling using the Min Max Scaler on the numerical variables.

Model Building Part

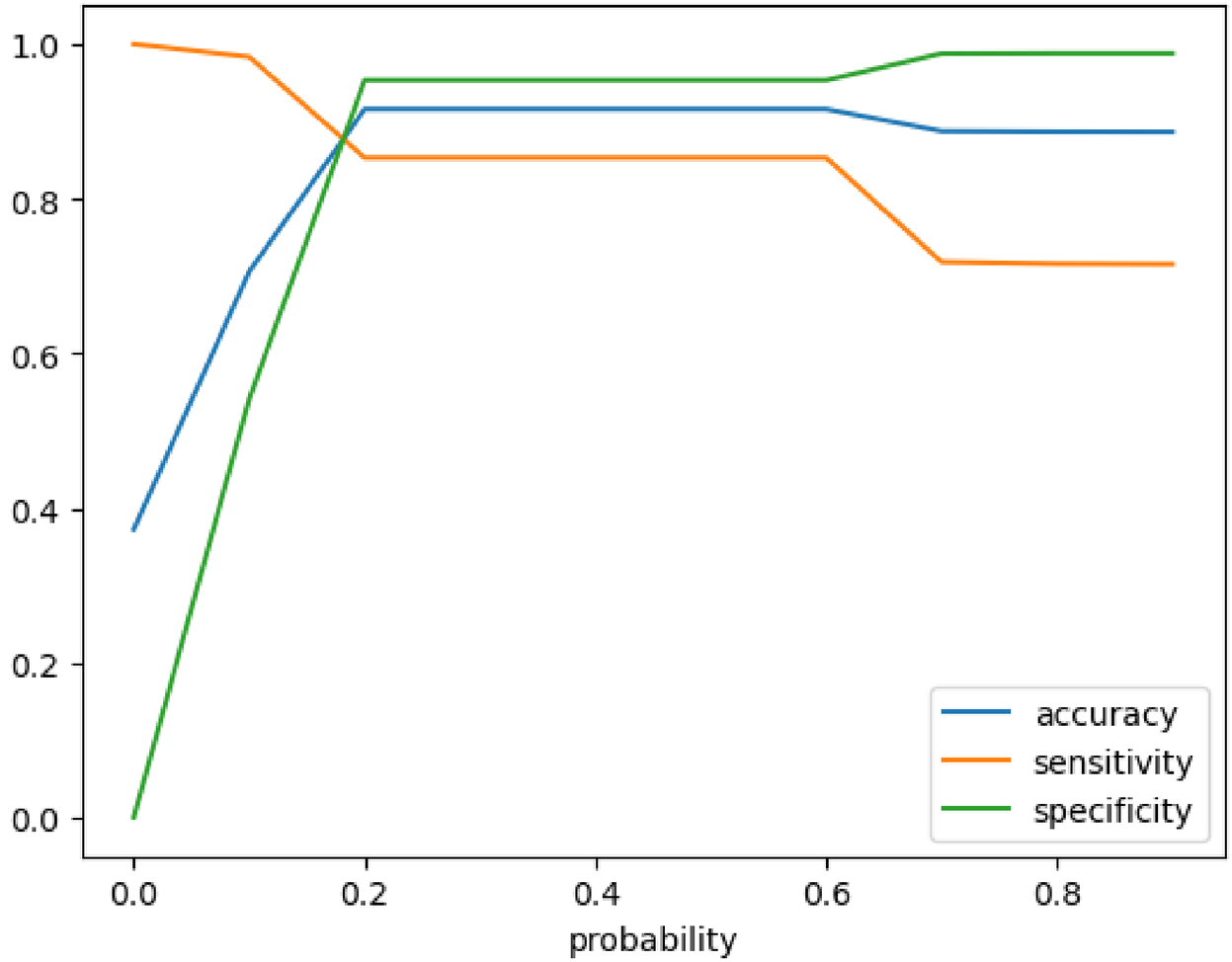
1. With the help of Recursive Feature Elimination (RFE), I've selected the Top 15 important features.
2. After that I've dropped some of the features which have high p-value and high VIF.
3. I've left with final 12 Variables

Analysis of ROC Curve



1. The ROC Curve came with an area coverage of 95% which is too good.

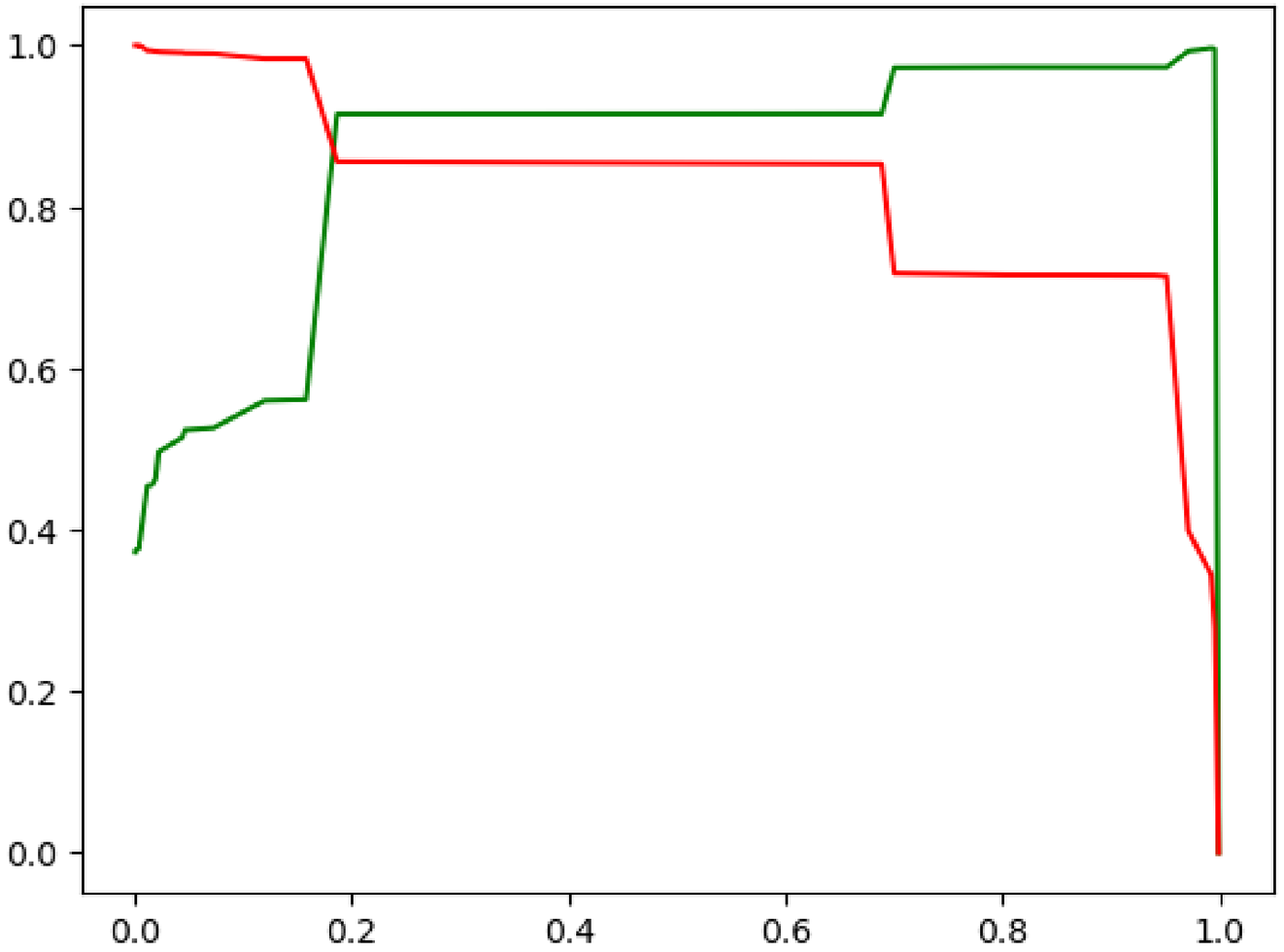
Analysis of Accuracy, Sensitivity, Specificity



Model Evaluation with Accuracy, Sensitivity, Specificity

1. The optimal probability cut-off turns around 0.19 approximately.
2. For Train Dataset:- Accuracy (91.58%); Sensitivity (85.30%), Specificity (95.31%).
3. For Test Dataset:- Accuracy (91.65%); Sensitivity (86.61%), Specificity (94.35%).

Analysis of Precision Recall Curve



- Model Evaluation using Precision and Recall Metrics Technique:
  With Precision and Recall, we got a cutoff value of 0.18 approximately.

Conclusion

- While I’ve checked Sensitivity and Specificity along with Precision and Recall Metrics Techniques. I’ve decided that the Sensitivity and Specificity technique is more efficient.
- Accuracy, Sensitivity and Specificity values of the Test Dataset are around 90% which is closed to the respective values calculated using the Trained Set. Which means the Conversion rate on The Final Predicted Model is around 85%.
- These Features contributed the most to get the lead converted:
  - Total Tme Spent on Website
  - Lead Source
  - Tags (Interested in MBA)

The Company should focus more on the below listed variable/columns to increase their conversions:-

1. **Last Notable Activity:** A person who had a Phone Conversation, Email Opened and SMS Sent is more likely to be converted.
2. **Specialization:** People who have Management in their specializations are more likely to be converted.
3. **Total Time Spent On Website:** People who spend more time on the website are more likely to be converted.
4. **What is your current occupation:** People who are working professionals and Unemployed are more likely to be converted.
5. **Lead Source:** Leads which are from Google and Direct Traffic are more likely to be converted.

THANK YOU

Harsh Deep Jaggi
DSC-57 Batch