Lead Score – Case Study

Lead Score – Case Study

Problem Statement

An 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’.
The company requires you to build a model wherein 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%.**

Lead Score – Case Study

Goals of the Case Study

There are quite a few goals for this case study.

1.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.

2.There are some more problems presented by the company which your model should be able to adjust to if the company's requirement changes in the future so you will need to handle these as well. These problems are provided in a separate doc file. Please fill it based on the logistic regression model you got in the first step. Also, make sure you include this in your final PPT where you'll make recommendations.

Lead Score – Case Study

Strategy

ØSource the data for analysis

ØClean and prepare the data

ØExploratory Data Analysis.

ØFeature Scaling

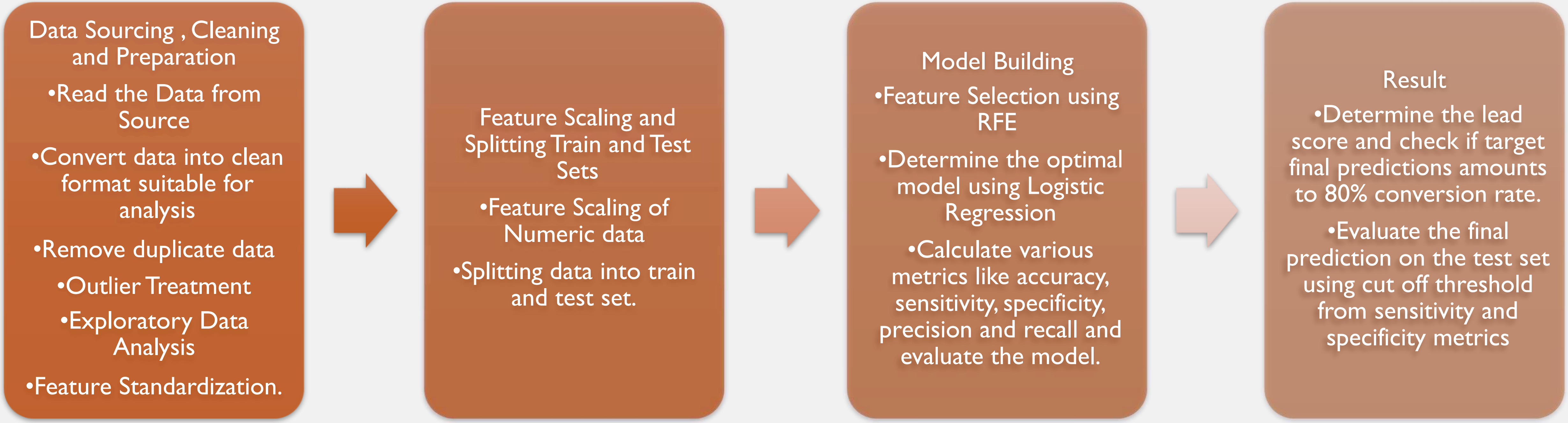
ØSplitting the data into Test and Train dataset.

ØBuilding a logistic Regression model and calculate Lead Score.

ØEvaluating the model by using different metrics - Specificity and Sensitivity or Precision and Recall.

ØApplying the best model in Test data based on the Sensitivity and Specificity Metrics.

Problem solving methodology



Data Sourcing , Cleaning and Preparation

- Read the Data from Source
- Convert data into clean format suitable for analysis
- Remove duplicate data
- Outlier Treatment
- Exploratory Data Analysis
- Feature Standardization.

Feature Scaling and Splitting Train and Test Sets

- Feature Scaling of Numeric data
- Splitting data into train and test set.

Model Building

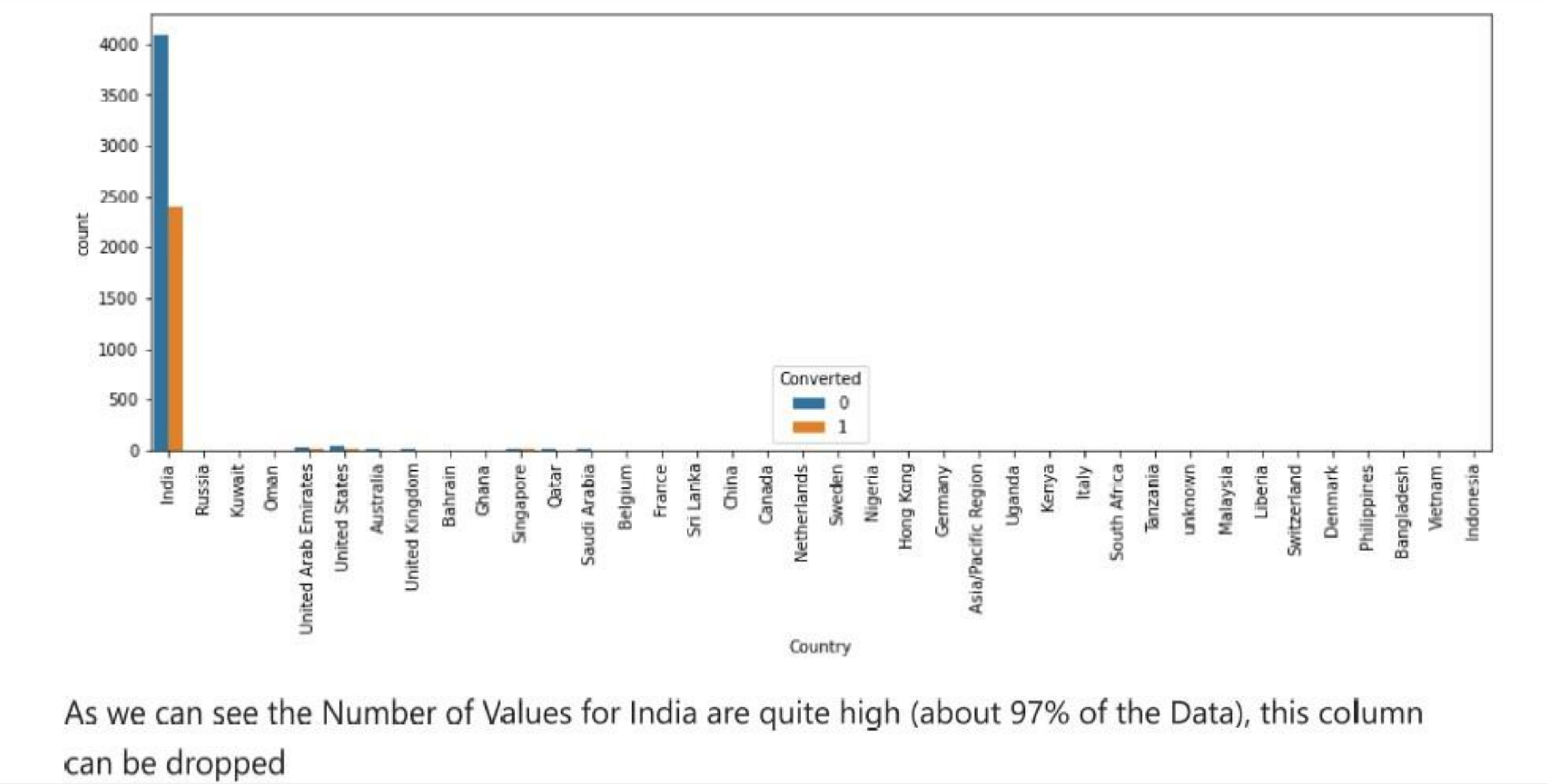
- Feature Selection using RFE
- Determine the optimal model using Logistic Regression
- Calculate various metrics like accuracy, sensitivity, specificity, precision and recall and evaluate the model.

Result

- Determine the lead score and check if target final predictions amounts to 80% conversion rate.
- Evaluate the final prediction on the test set using cut off threshold from sensitivity and specificity metrics

Lead Score – Case Study

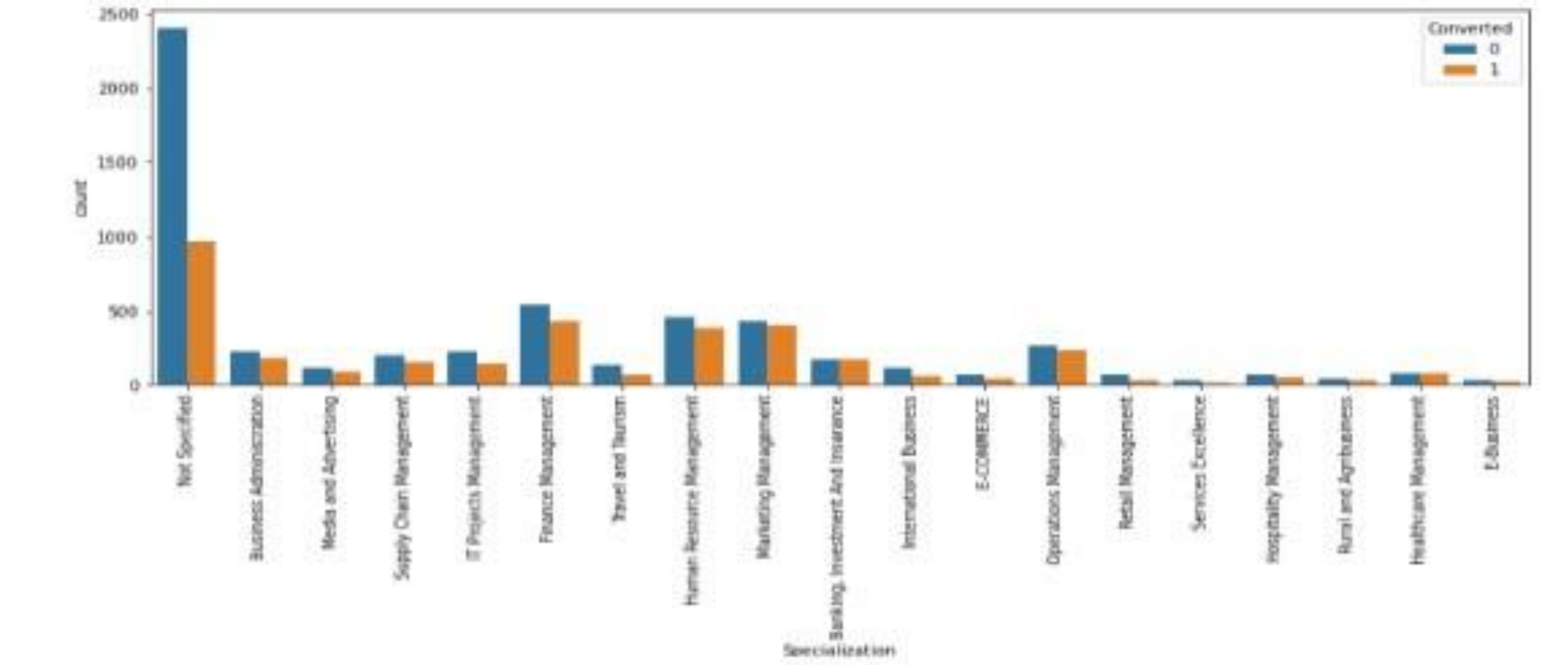
Exploratory Data Analysis



As we can see the Number of Values for India are quite high (about 97% of the Data), this column can be dropped

Lead Score – Case Study

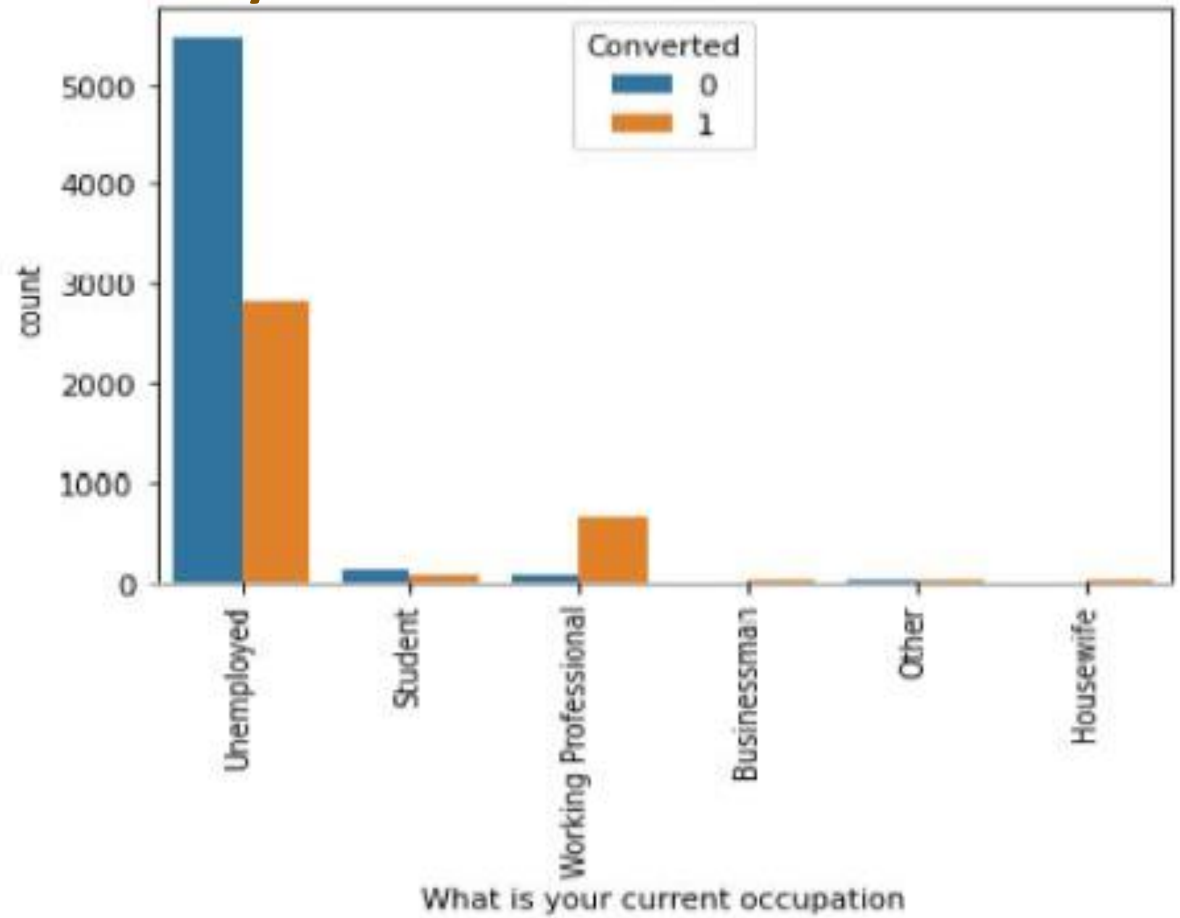
Exploratory Data Analysis



We see that specialization with Management in them have higher number of leads as well as leads converted. So this is definitely a significant variable and should not be dropped.

Lead Score – Case Study

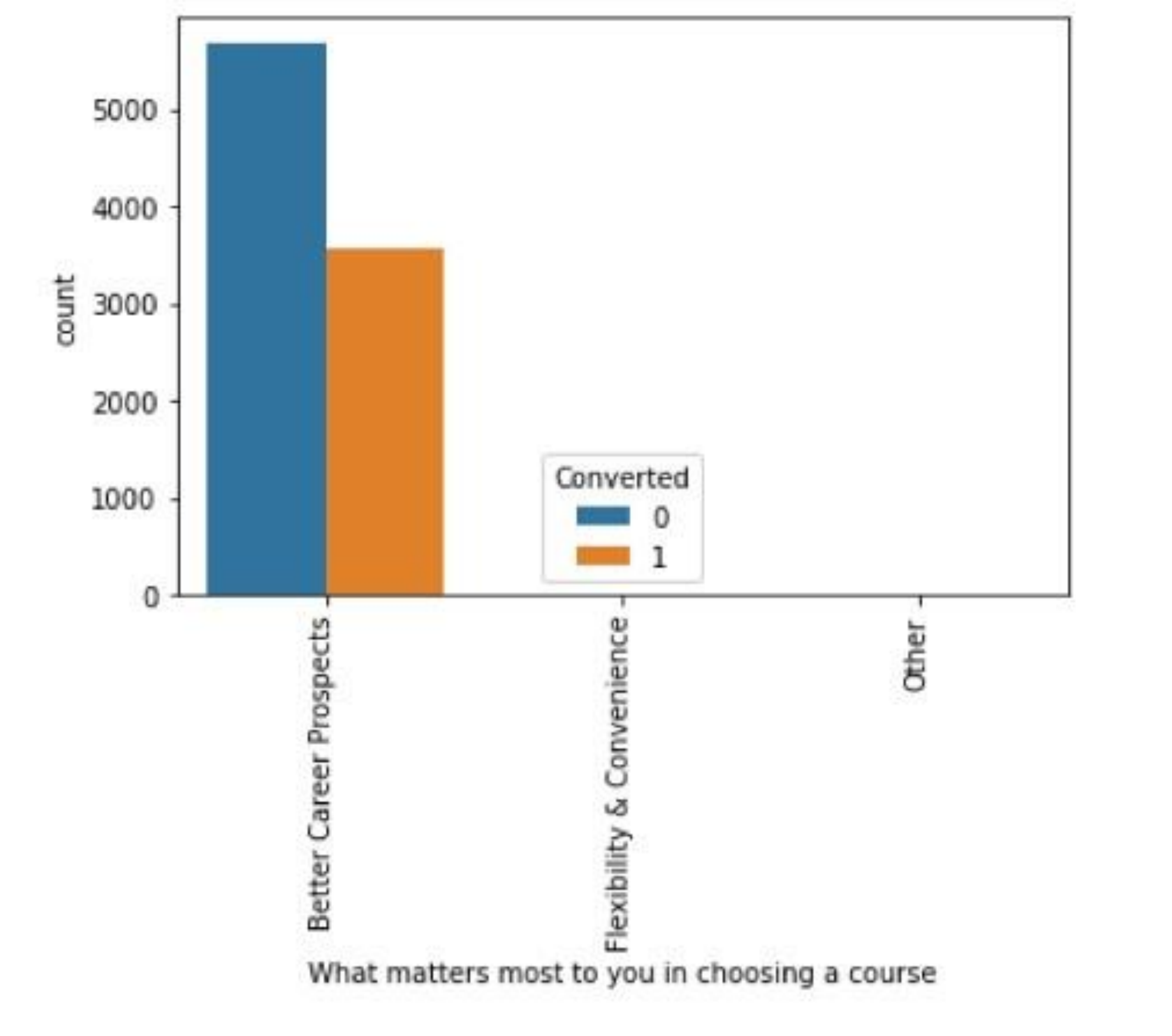
Exploratory Data Analysis



Working Professionals going for the course have high chances of joining it. Unemployed leads are the most in terms of Absolute numbers.

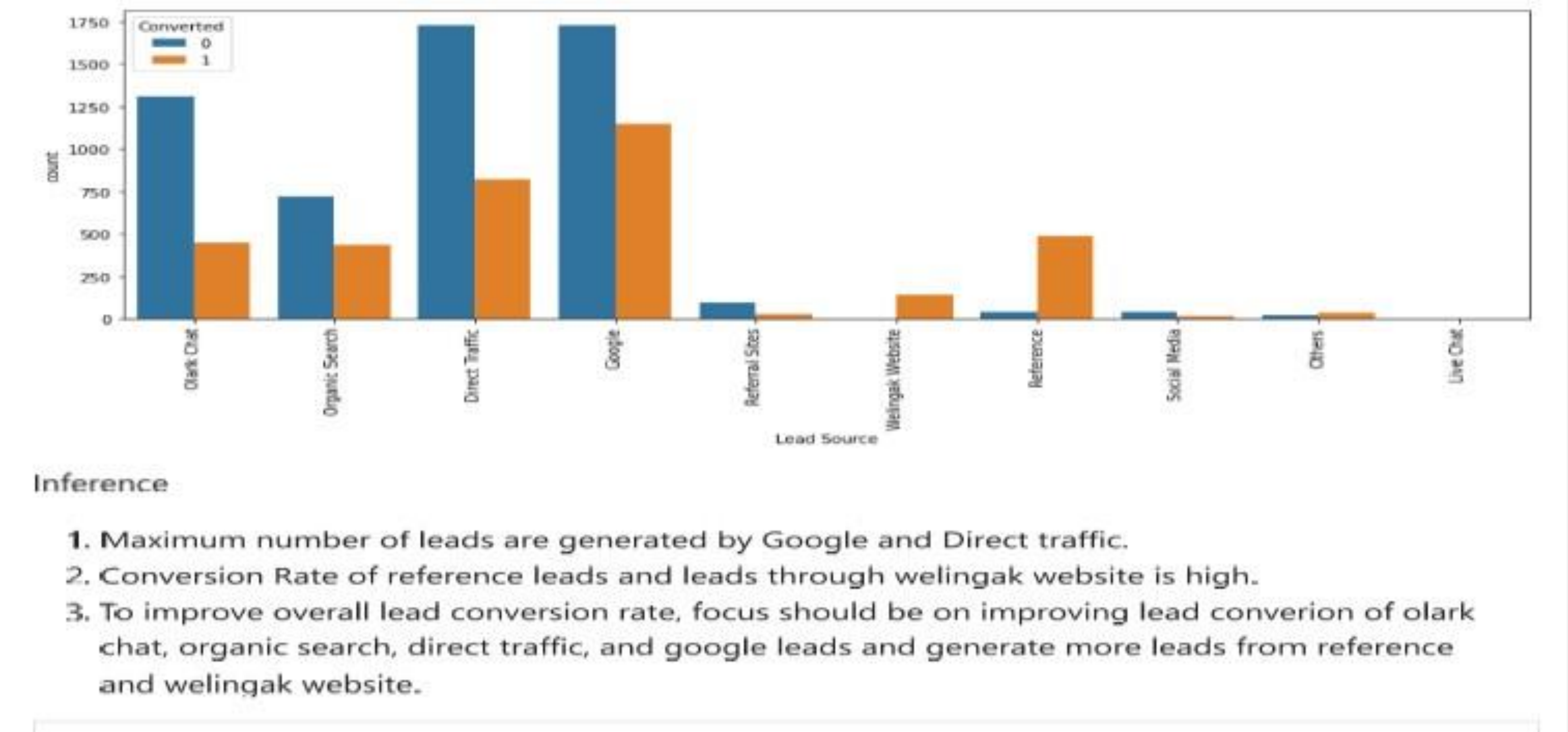
Lead Score – Case Study

Exploratory Data Analysis



Lead Score – Case Study

Exploratory Data Analysis

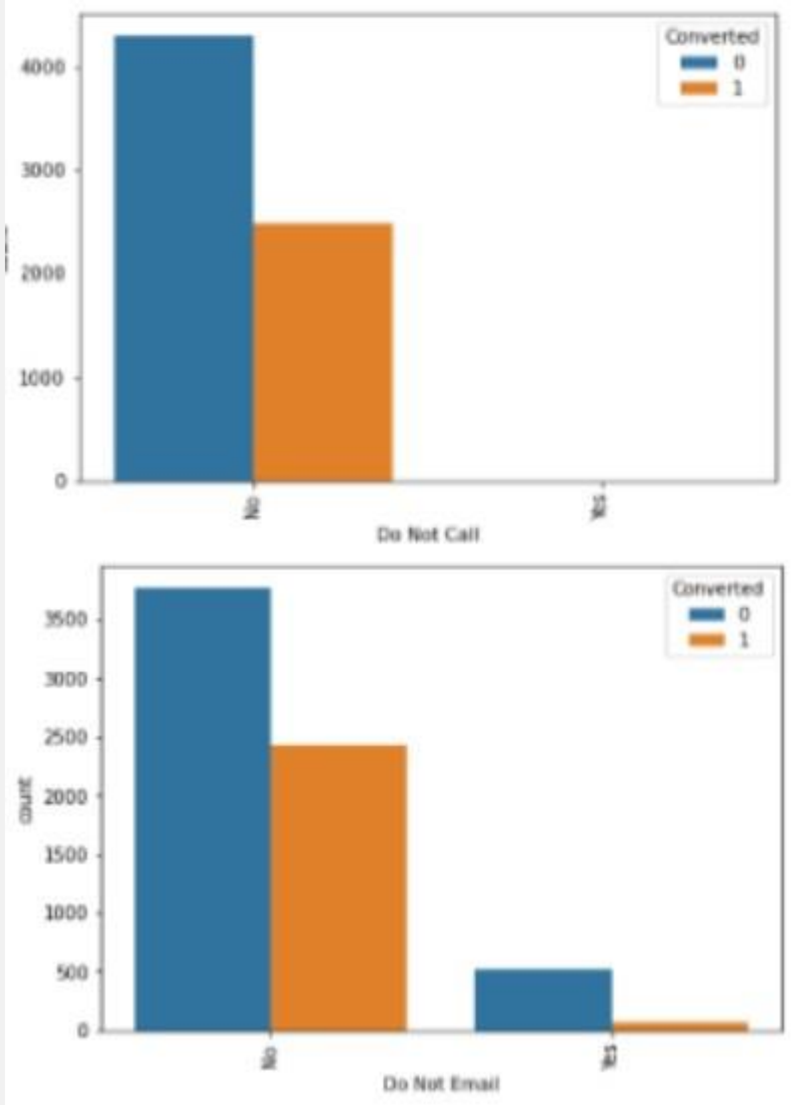


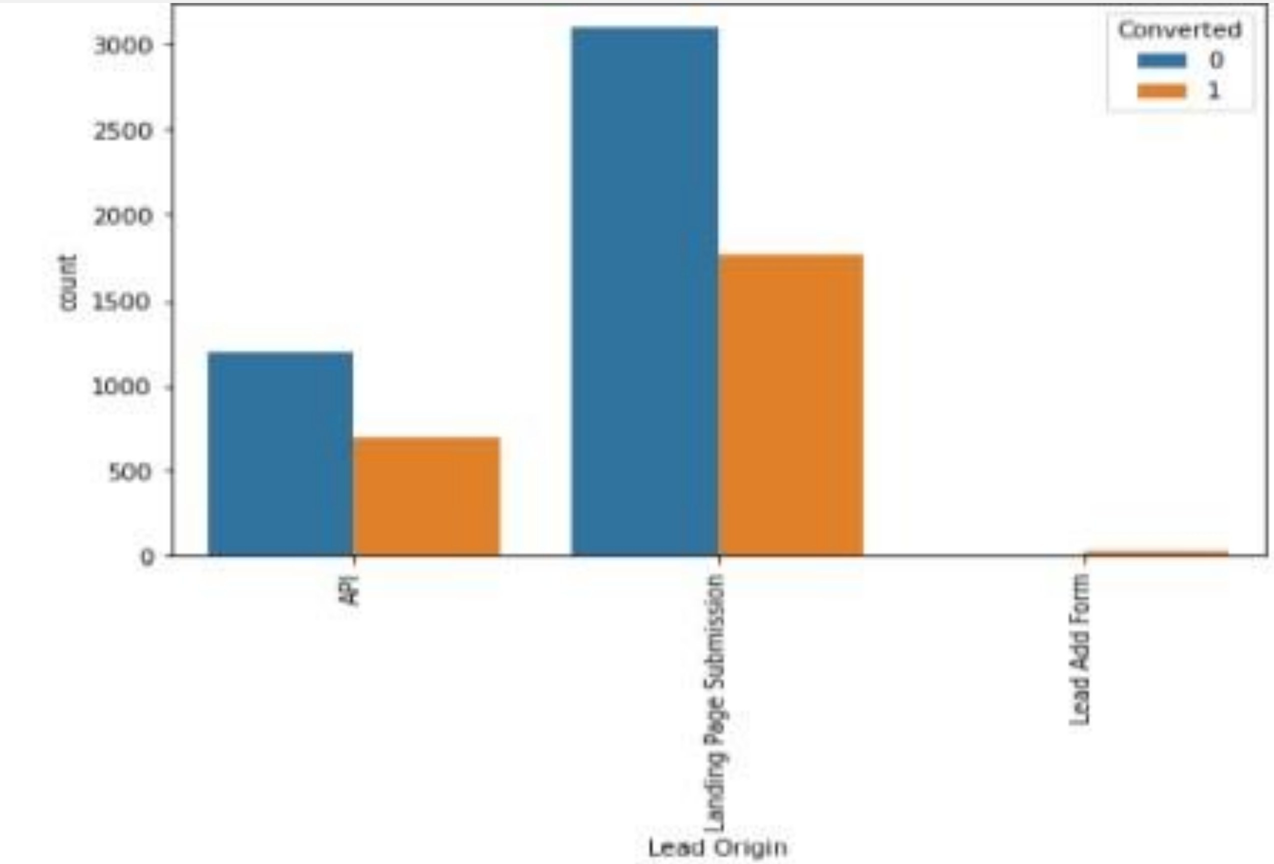
Inference

1. Maximum number of leads are generated by Google and Direct traffic.
2. Conversion Rate of reference leads and leads through welingak website is high.
3. To improve overall lead conversion rate, focus should be on improving lead converion of olark chat, organic search, direct traffic, and google leads and generate more leads from reference and welingak website.

Lead Score – Case Study

Exploratory Data Analysis



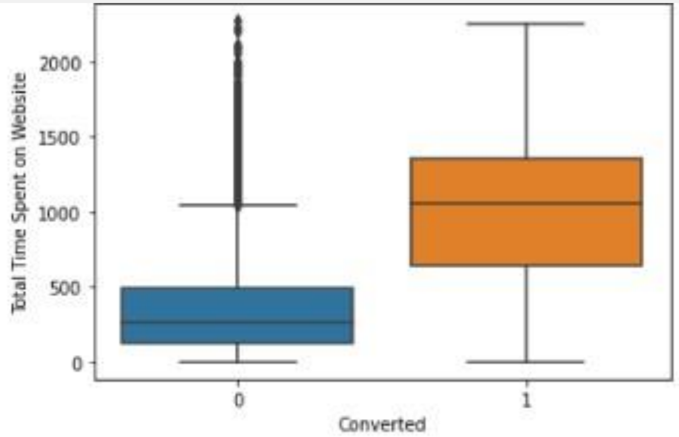


Inference

1. API and Landing Page Submission bring higher number of leads as well as conversion.
2. Lead Add Form has a very high conversion rate but count of leads are not very high.
3. Lead Import and Quick Add Form get very few leads.
4. In order to improve overall lead conversion rate, we have to improve lead converion of API and Landing Page Submission origin and generate more leads from Lead Add Form.

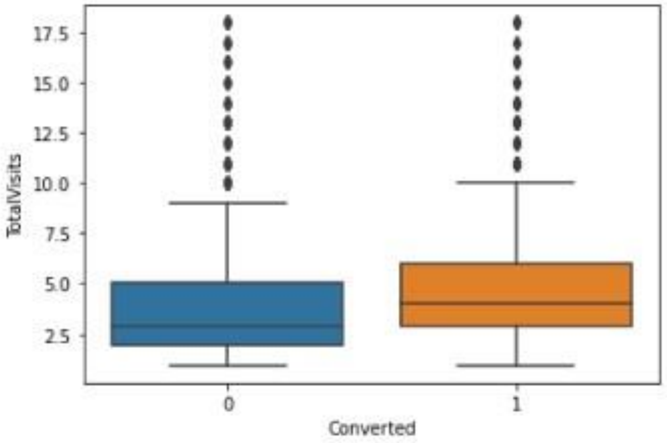
Lead Score – Case Study

Exploratory Data Analysis



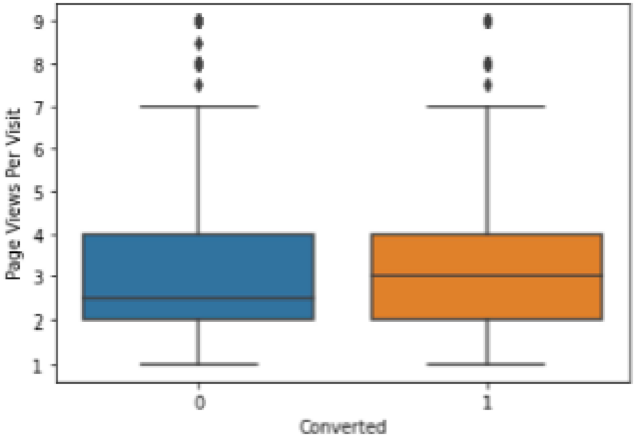
Inference

1. Leads spending more time on the website are more likely to be converted.
2. Website should be made more engaging to make leads spend more time.



Inference

1. Median for converted and not converted leads are the close.
2. Nothng conclusive can be said on the basis of Total Visits

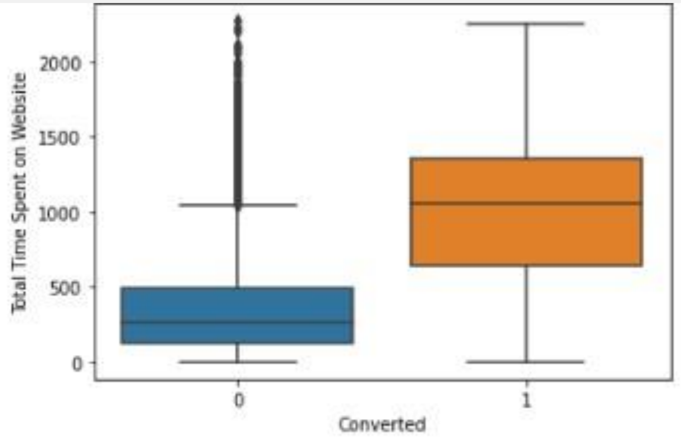


Inference

1. Median for converted and unconverted leads is the same.
2. Nothing can be said specifically for lead conversion from Page Views Per Visit

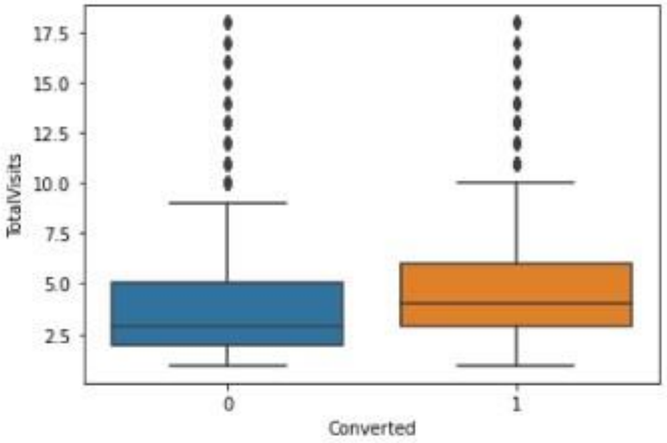
Lead Score – Case Study

Exploratory Data Analysis



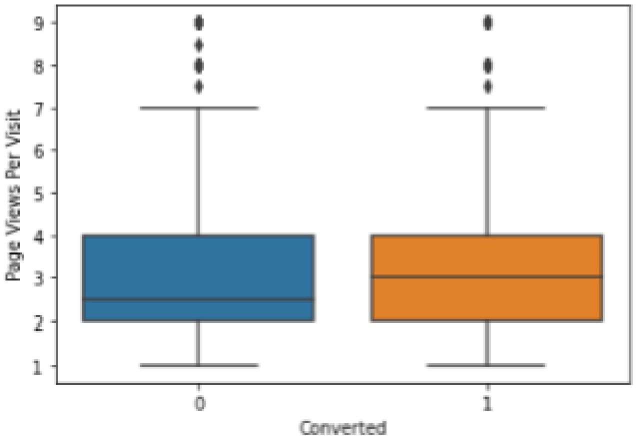
Inference

1. Leads spending more time on the website are more likely to be converted.
2. Website should be made more engaging to make leads spend more time.



Inference

1. Median for converted and not converted leads are the close.
2. Nothng conclusive can be said on the basis of Total Visits



Inference

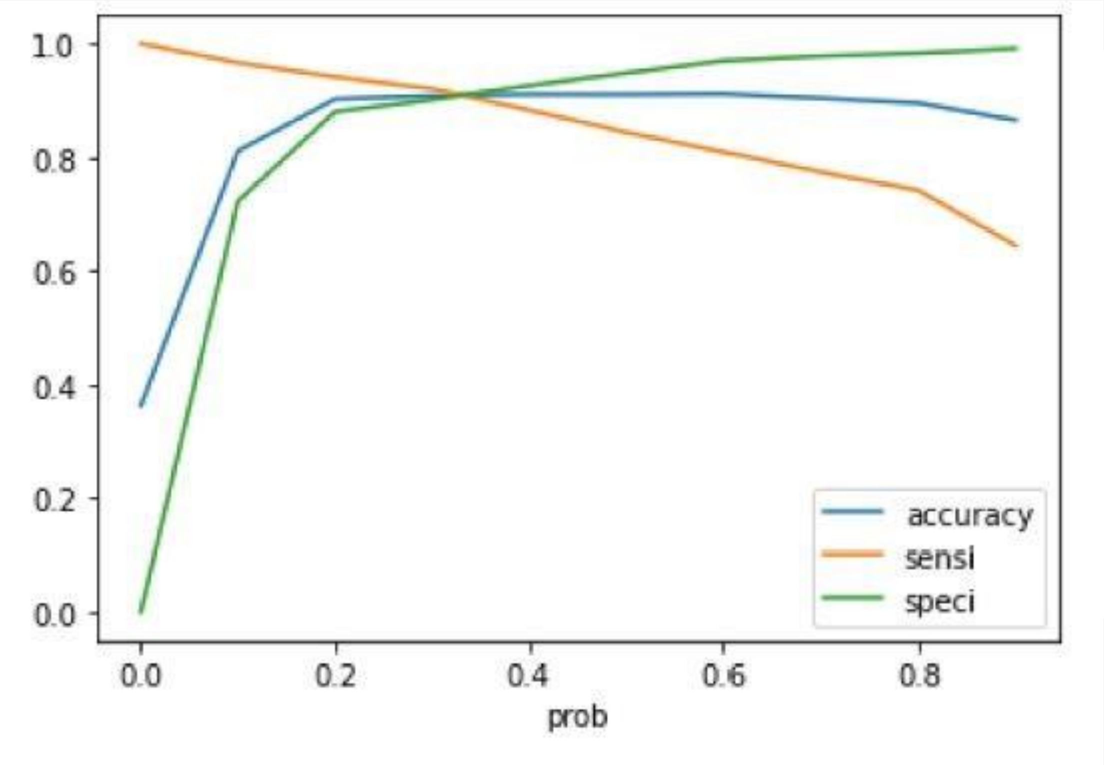
1. Median for converted and unconverted leads is the same.
2. Nothing can be said specifically for lead conversion from Page Views Per Visit

Lead Score – Case Study

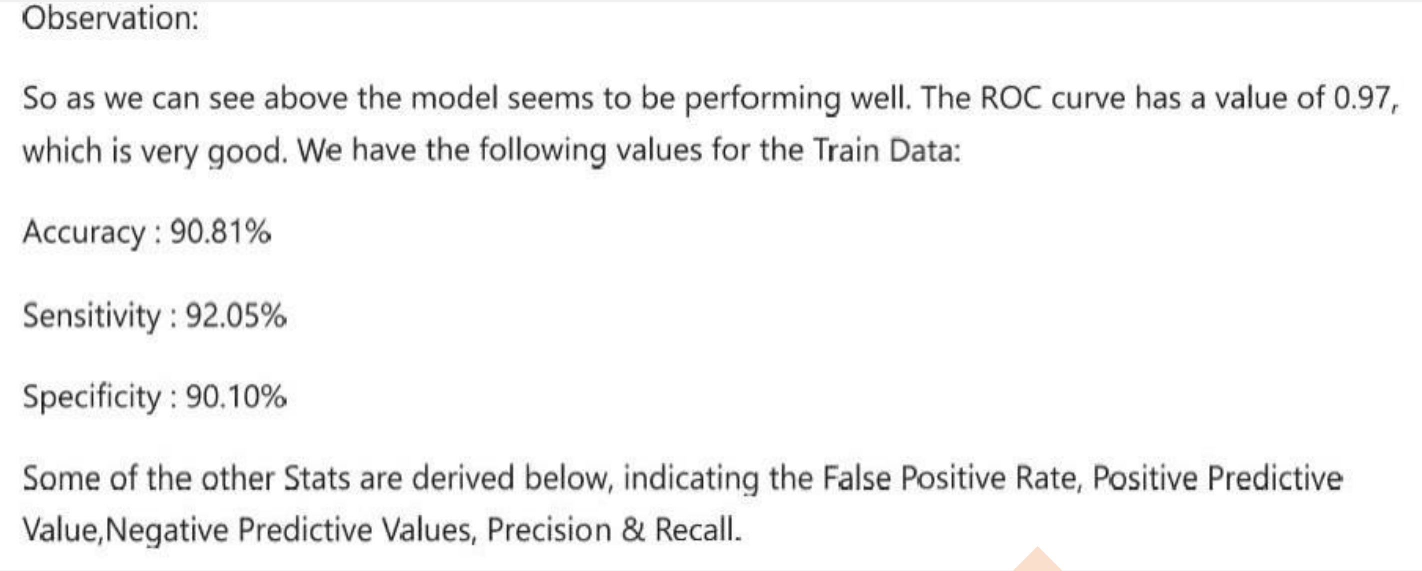
Variables Impacting the Conversion Rate

- Do Not Email
- Total Visits
- Total Time Spent On Website
- Lead Origin – Lead Page Submission
- Lead Origin – Lead Add Form
- Lead Source - Olark Chat
- Last Source – Welingak Website
- Last Activity – Email Bounced
- Last Activity – Not Sure
- Last Activity – Olark Chat Conversation
- Last Activity – SMS Sent
- Current Occupation –No Information
- Current Occupation – Working Professional
- Last Notable Activity –Had a Phone Conversation
- Last Notable Activity - Unreachable

Model Evaluation -Sensitivity and Specificity on Train Data Set



The graph depicts an optimal cut off 0.37 based on Accuracy, Sensitivity and Specificity



Observation:

So as we can see above the model seems to be performing well. The ROC curve has a value of 0.97, which is very good. We have the following values for the Train Data:

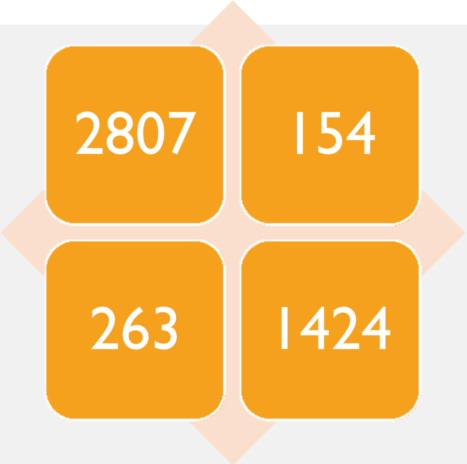
Accuracy : 90.81%

Sensitivity : 92.05%

Specificity : 90.10%

Some of the other Stats are derived below, indicating the False Positive Rate, Positive Predictive Value,Negative Predictive Values, Precision & Recall.

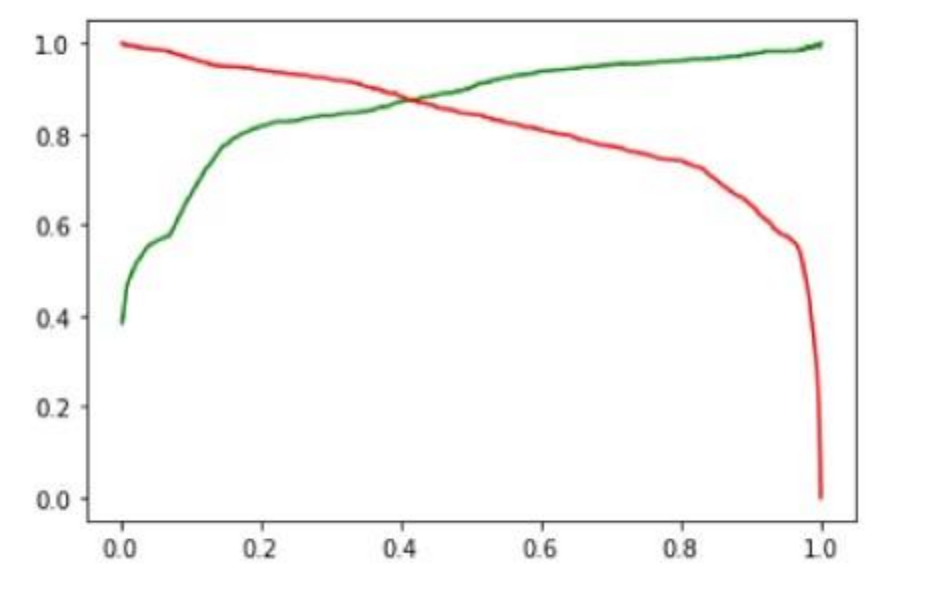
Confusion Matrix



2807	154
263	1424

Lead Score – Case Study

Model Evaluation -Precision and Recall on Train Data Set



- Precision-84.12%
- Recall-92.05 %

The graph depicts an optimal cut off 0.42 based on Precision and Recall

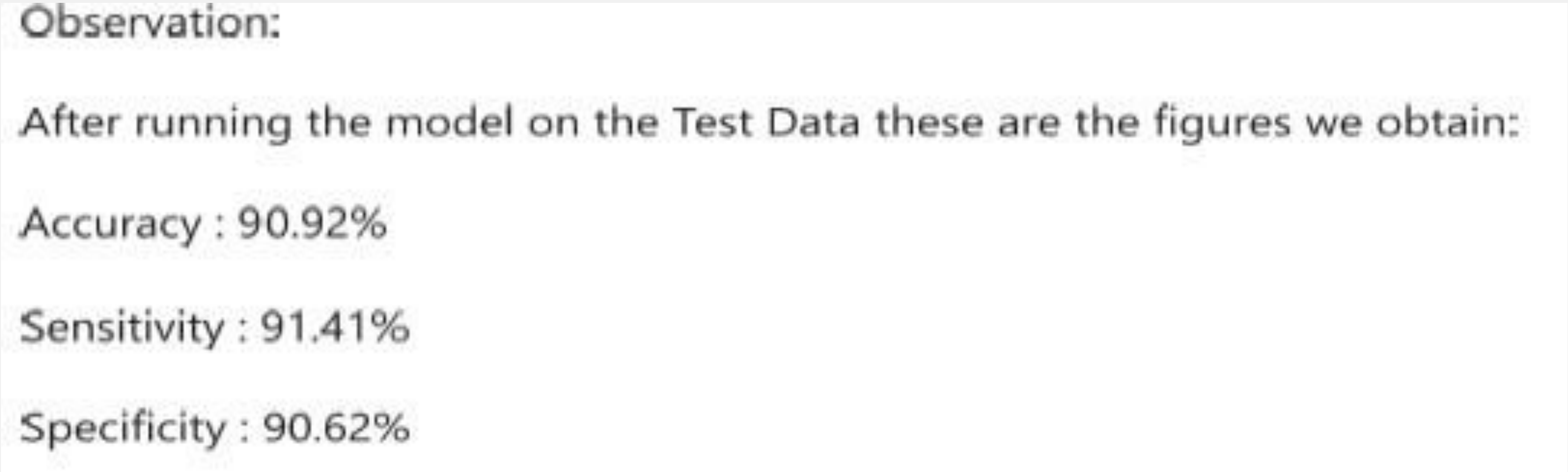
Confusion Matrix

2668	293
134	1553

Model Evaluation -Precision and Recall on Test Data Set

Confusion Matrix

1140	118
63	671



Observation:

After running the model on the Test Data these are the figures we obtain:

Accuracy : 90.92%

Sensitivity : 91.41%

Specificity : 90.62%

Lead Score – Case Study

Conclusion

- 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 91%, 91.41% and 90.62% which are approximately closer to the respective values calculated using trained set.

- lead score calculated shows the conversion rate on the final predicted model is around 92.05% (in train set) and 91.41% in test set

- The top variables that contribute for lead getting converted in the model are

1. Total time spent on website
2. What is your current occupation
3. Lead Add Form from Lead Origin
4. Had a Phone Conversation from Last Notable Activity

- Hence overall this model seems to be good.