Lead Score Case Study

Submitted by :

Sumit Bhatia

Bimla Bisht

Jyoti Sharma

Lead Score Case Study for X Education

Problem Statement :

X Education sells online courses to industry professionals. 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%.

Business Goal:

X Education needs help in selecting the most promising leads, i.e. the leads that are most likely to convert into paying customers.

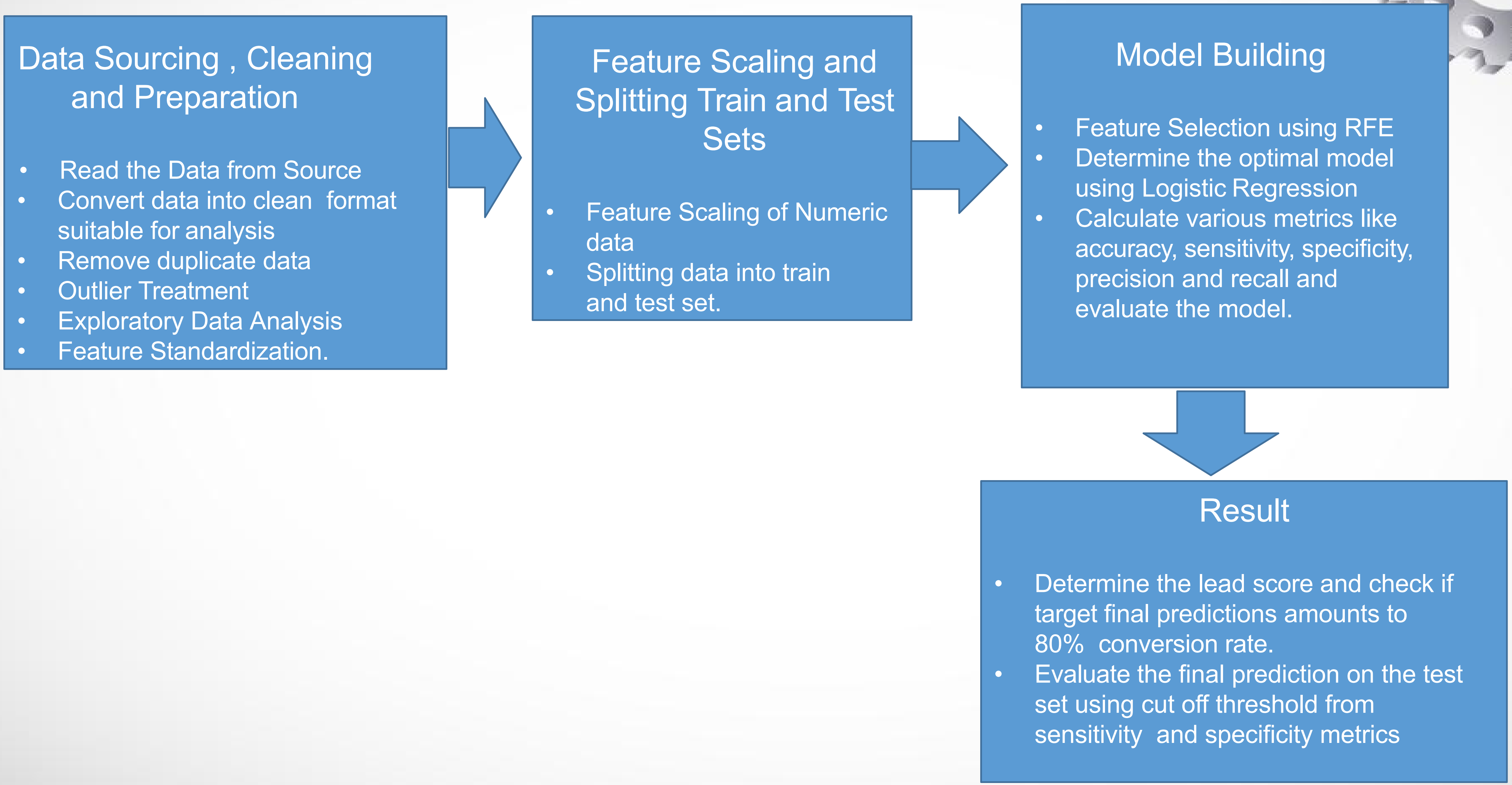
The company needs a model wherein your lead score is assigned to each of the leads such that the customers with higher lead score have a higher conversion chance and the customers with lower lead score have a lower conversion chance.

The CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80%

Strategy :-

- Source the data for analysis
- Clean and prepare the data
- Exploratory Data Analysis.
- Feature Scaling
- Splitting the data into Test and Train dataset.
- Building a logistic Regression model and calculate Lead Score.
- Evaluating the model by using different metrics - Specificity and Sensitivity or Precision and Recall.
- Applying 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.

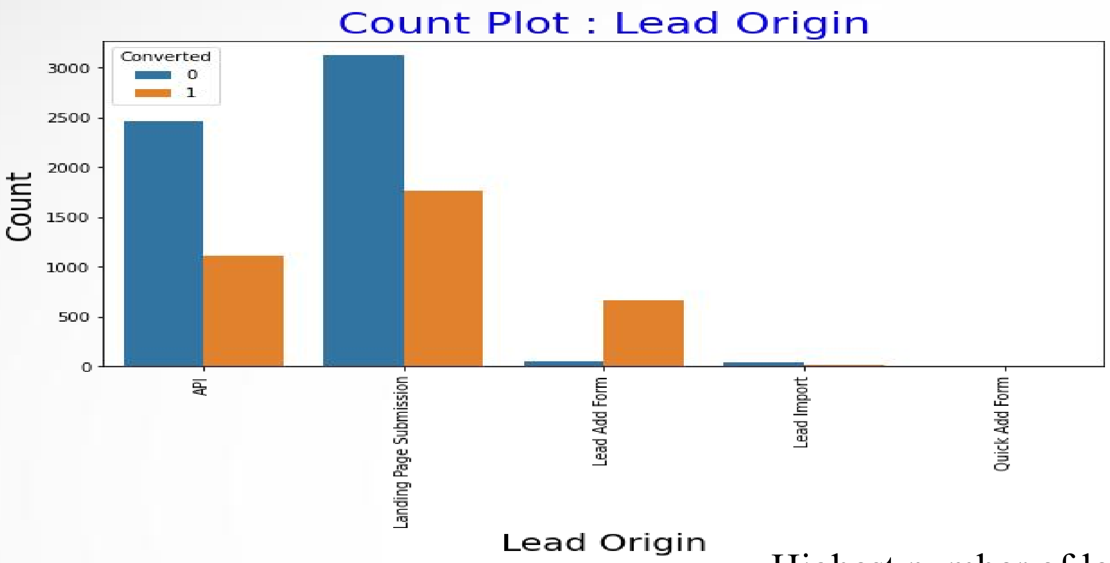
Result

- 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

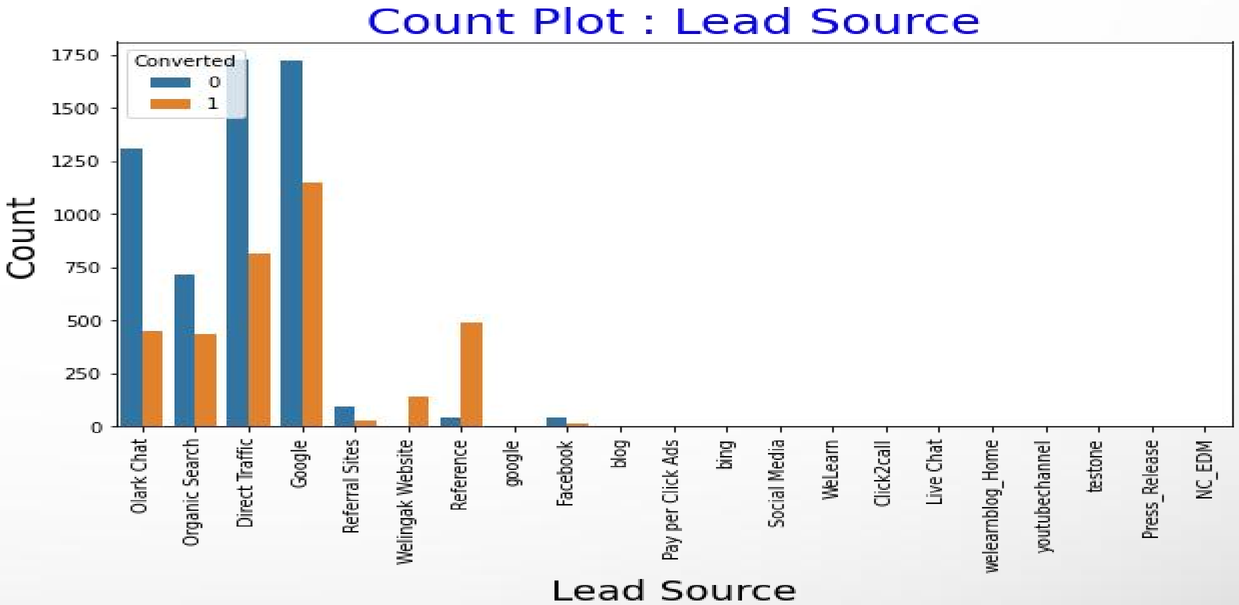
Exploratory data analysis and data visualization

Exploratory data analysis and data visualization

In Lead Origin, maximum conversion happened from Landing Page Submission

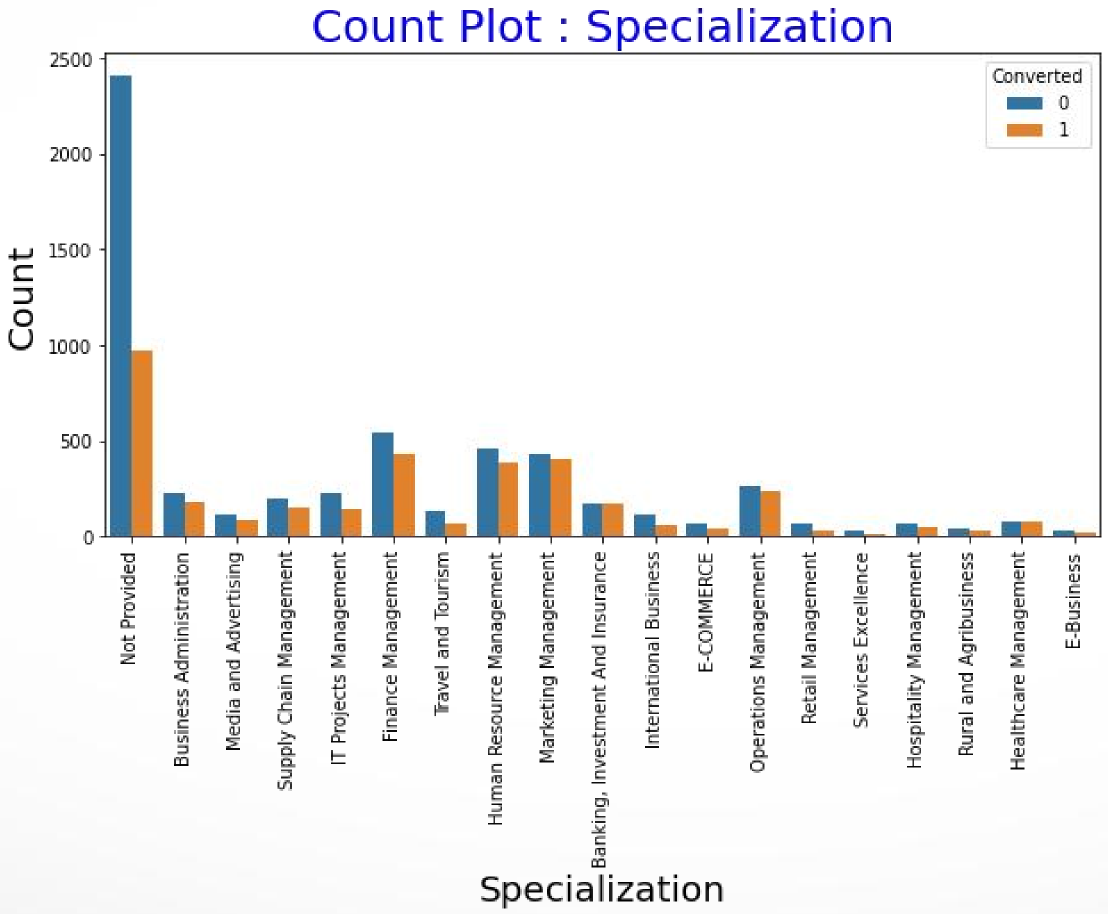


Highest number of leads are coming from "Google" and "Direct Traffic" followed by "Olark Chat".



Exploratory data analysis and data visualization

Highest number of lead generators are the people from management specialization.



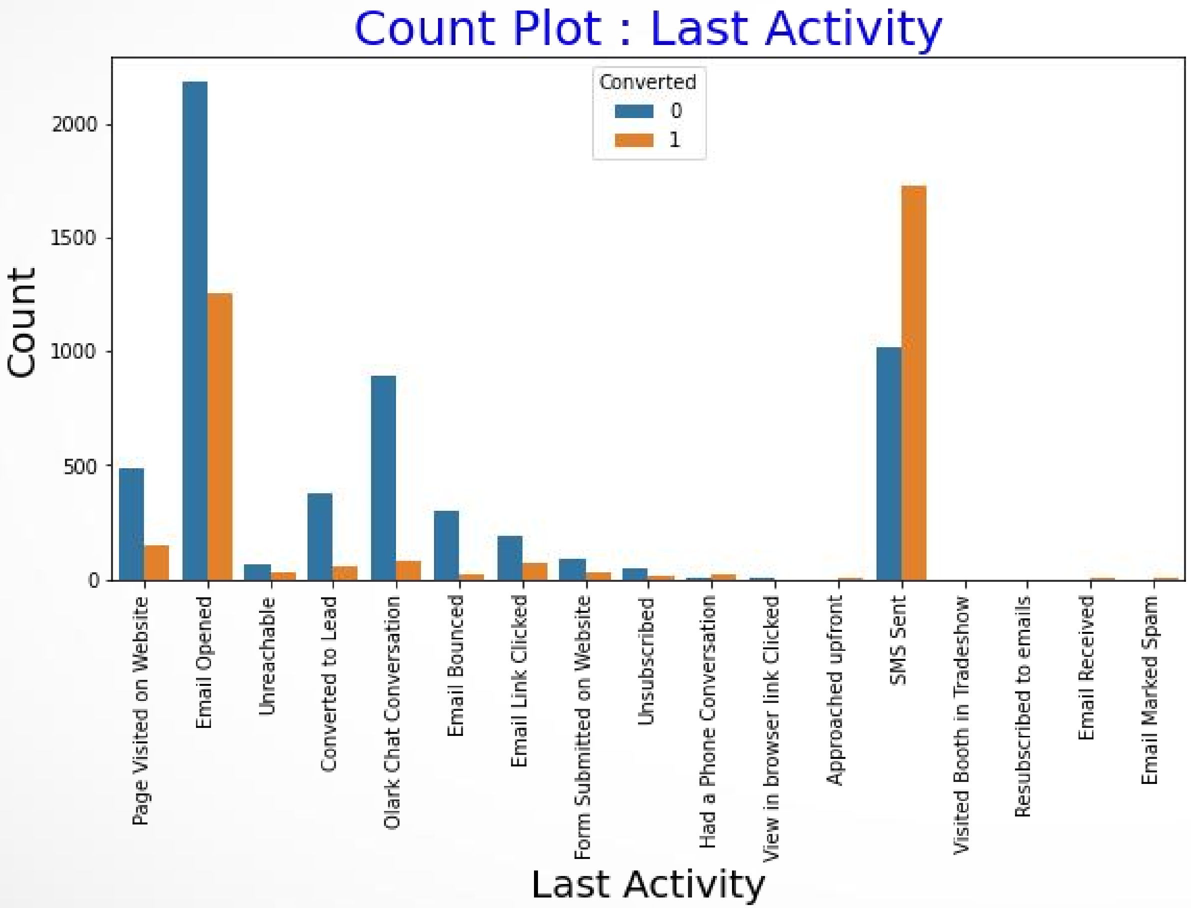
INSIGHTS

1. Highest number of leads are coming from "Google" and "Direct Traffic" followed by "Olark Chat".

2. Highest lead conversion is recorded by "Reference" and "Welingak Website".

3. We should focus on lead conversion rate of "Google" , "Direct Traffic" , "Olark Chat" and "Organic Search".

INSIGHTS



1. Most leads are coming from "Email Opened" followed by "SMS Sent"

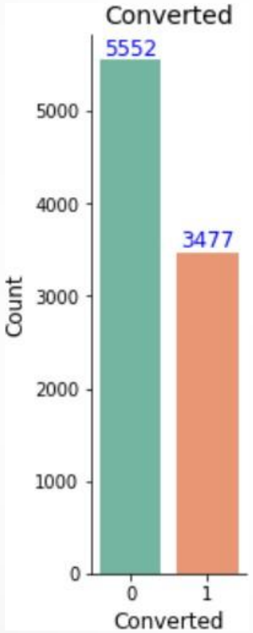
2. "Olark Chat Conversation" is generating high humber of leads but its conversion rate is very poor.

3. The conversion rate is good in case of "SMS Sent".

4. We should focus on improving the conversion rate of "Email Opened" as well as "Olark Chat Conversation".

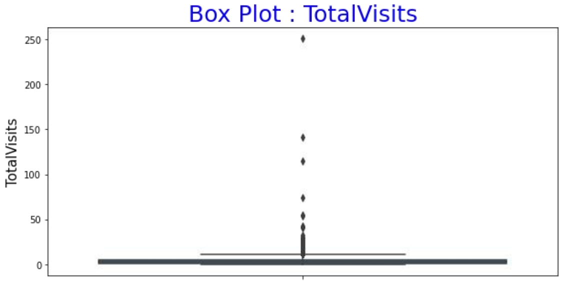
Exploratory Data Analysis

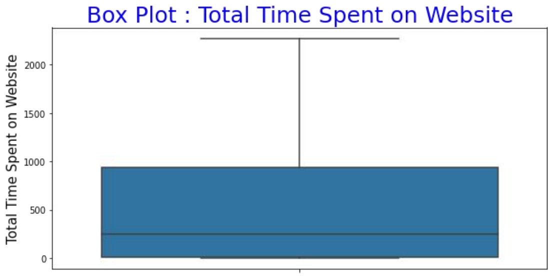
We have around 39% Conversion rate in Total

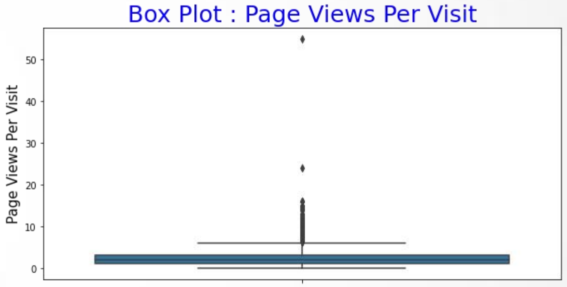


.

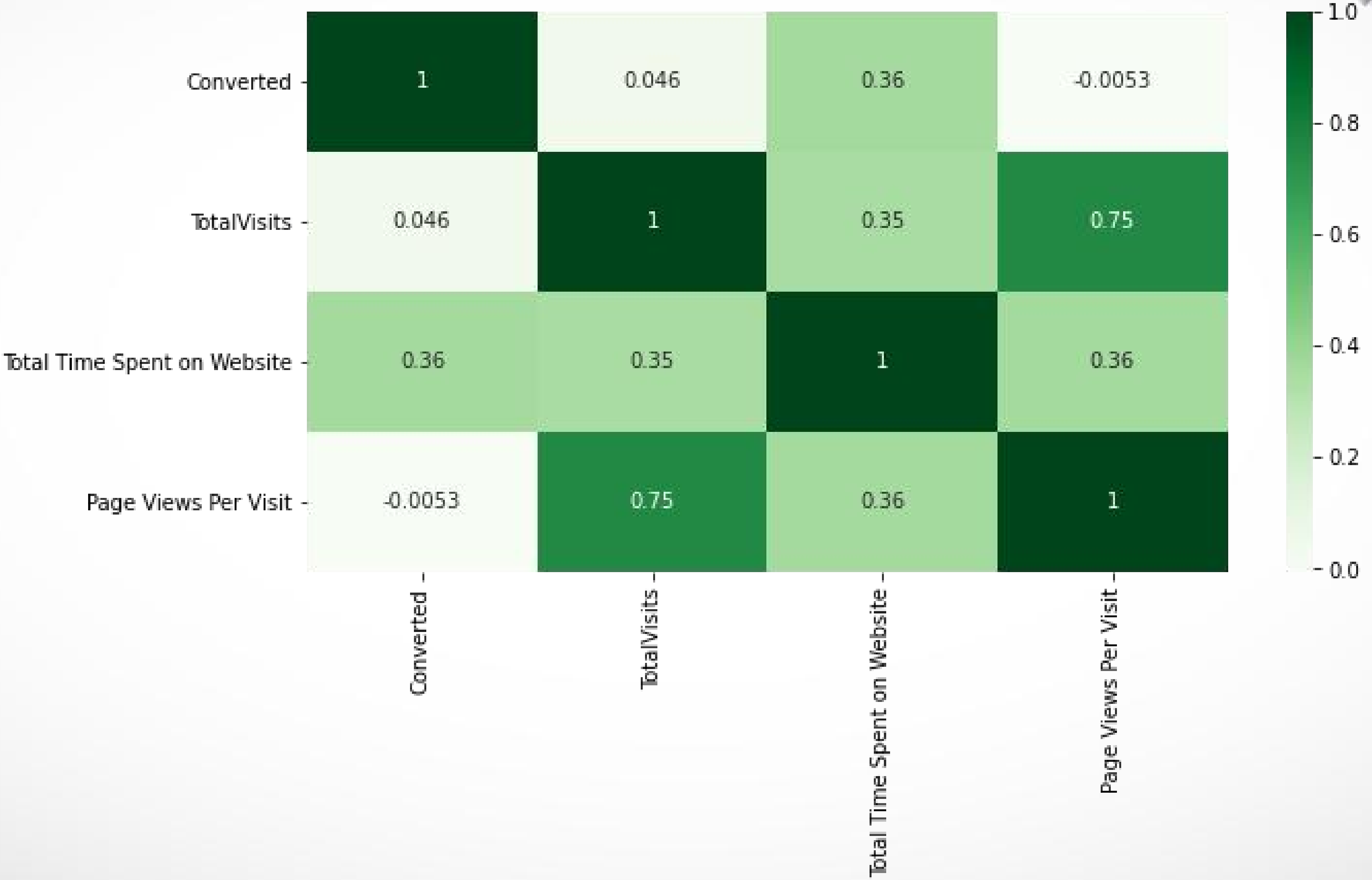
The conversion rates were high for Total Visits, Total Time Spent on Website and Page Views Per Visit





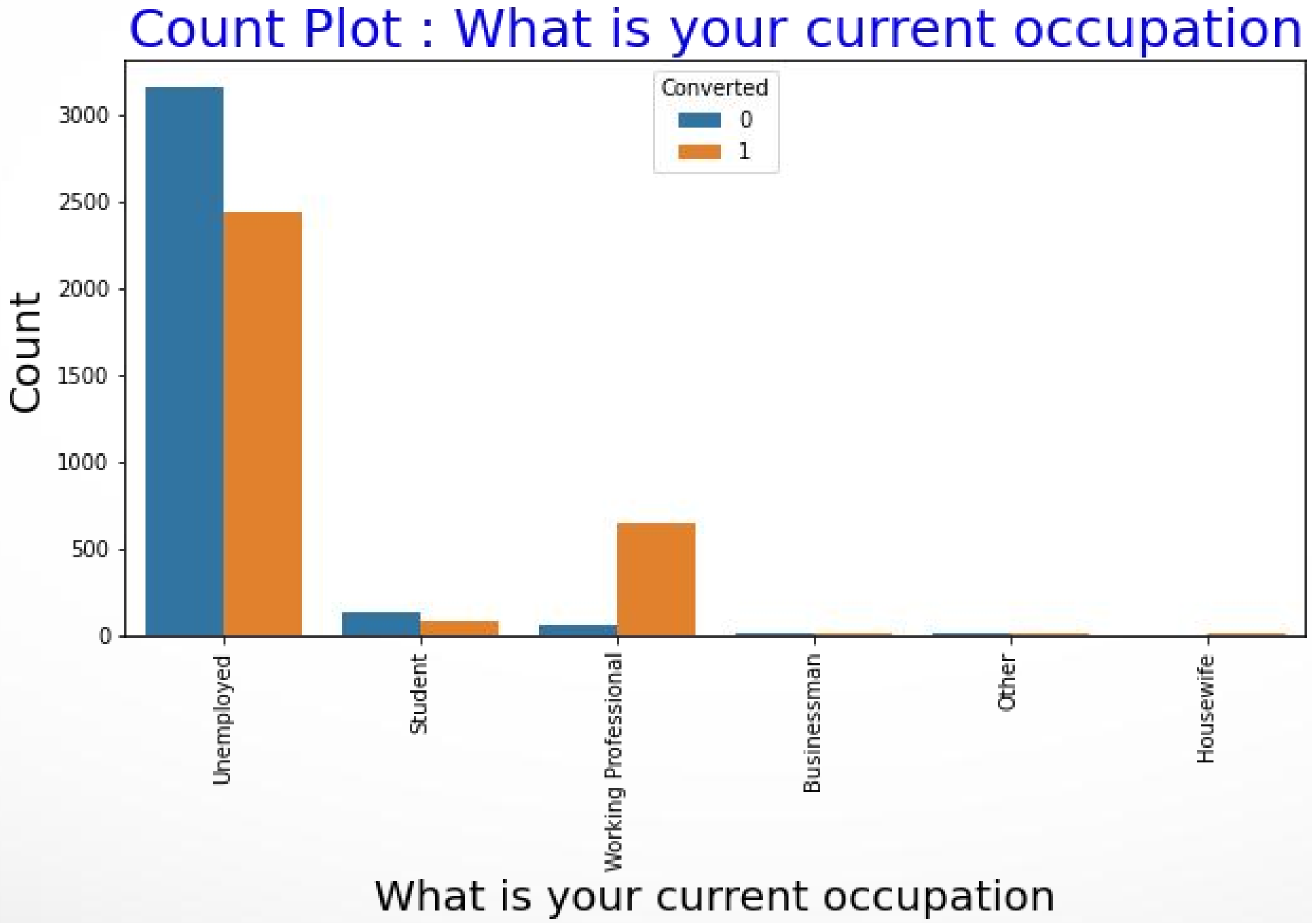


Variables "TotalVisits" and "Page Views Per Visit" have a high correlation.



INSIGHTS

1. Highest number of leads are coming from the unemployed category.
2. Conversion rate of "Working Professional" category is quite high as compared to other categories.

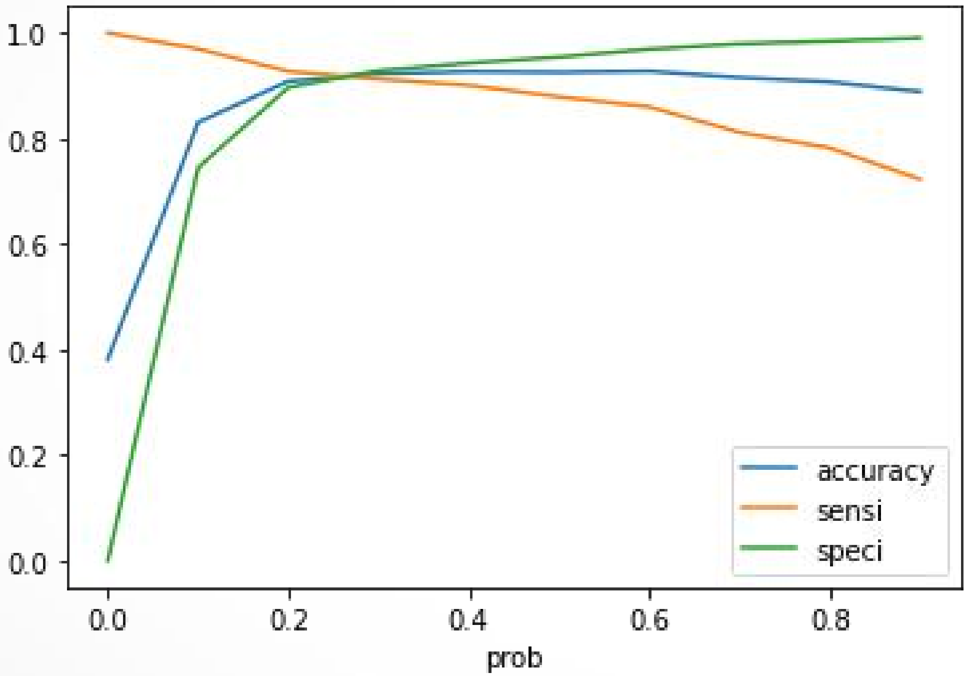


Variables Impacting the Conversion Rate

- 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 of 0.3 based on Accuracy, Sensitivity and Specificity



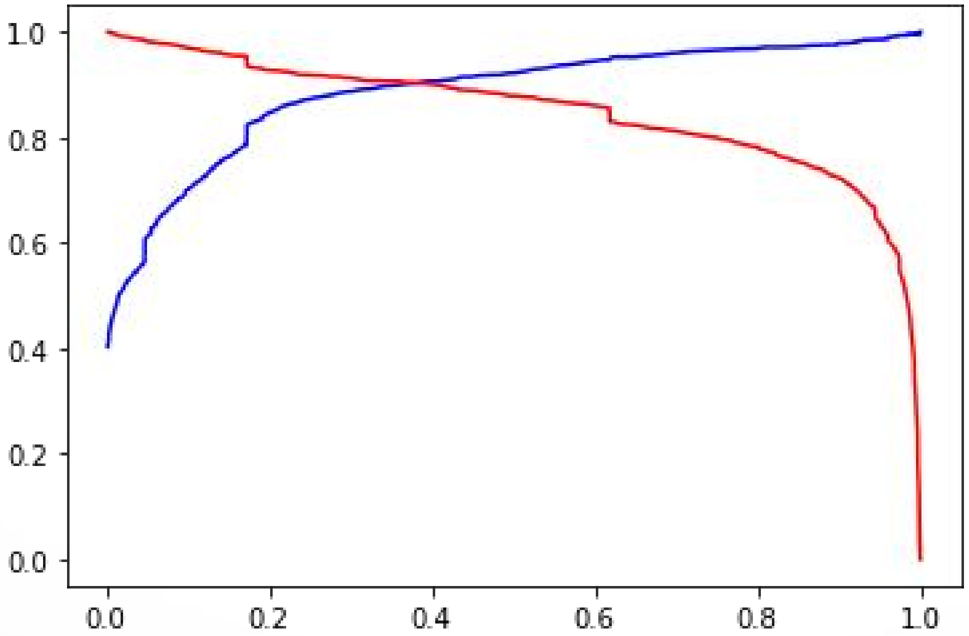
Confusion Matrix

3717	285
213	2253

- Accuracy - 92%
- Sensitivity - 91 %
- Specificity - 92 %
- False Positive Rate - 7 %
- Positive Predictive Value - 88 %
- Negative Predictive Value – 94%

Model Evaluation- Precision and Recall on Train Dataset

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



Conclusion

Top three variables in our model which contribute most towards the probability of a lead getting converted :

1. Tags_Closed by Horizzon

2. Tags_Lost to EINS

3. Tags_Will revert after reading the email

- The above features are the dummy features created from original categorical variables. These contribute positively in increasing the conversion probability.

- The management should focus more on leads with above mentioned tags to improve their conversion rate.

- The management should also focus on features with negative coefficients also so that they know which features decrease the conversion probability.

THANK YOU