Presentation

PROBLEM STATEMENT:

X Education is an organization which provides online courses for industry professional. The company markets its courses on several popular websites like google.
X Education wants to select most promising leads that can be converted to paying customers. Although the company generates a lot of leads only a few are converted into paying customers, wherein the company wants a higher lead conversion. Leads come through numerous modes like email, advertisements on websites, google searches etc.
The company has had 30% conversion rate through the whole process of turning leads into customers by approaching those leads which are to be found having interest in taking the course.

BUSINESS GOAL:

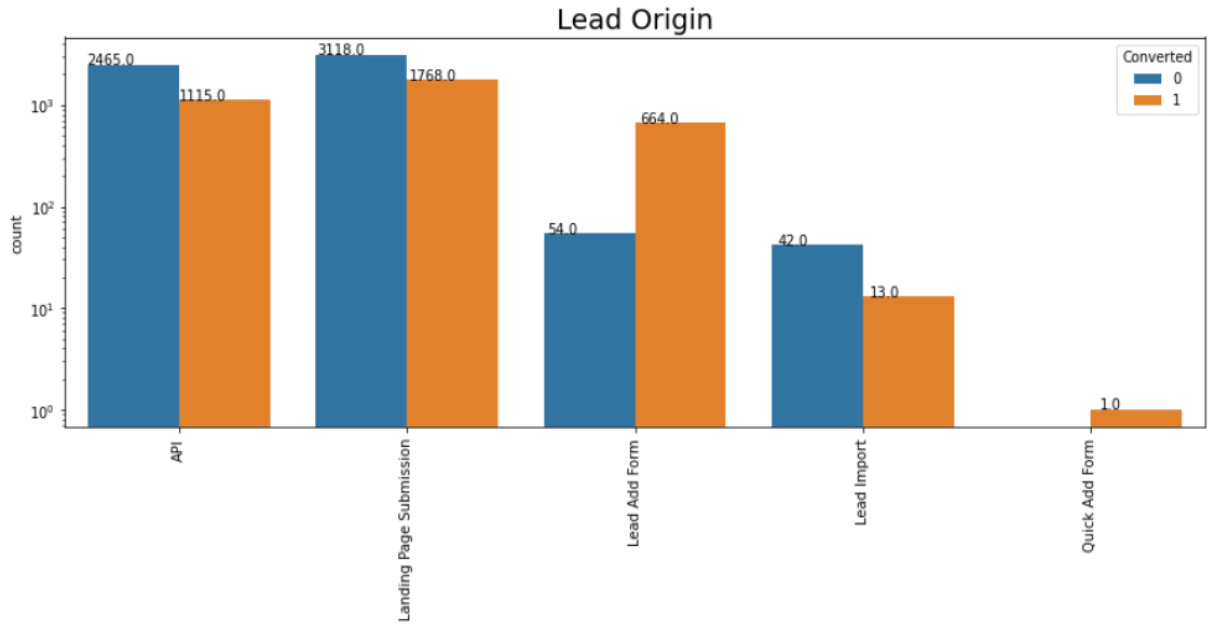
The company requires a model to be built for selecting most promising leads. Lead score to be given to each leads such that it indicates how promising the lead could be. The higher the lead score the more promising the lead to get converted, the lower it is the lesser the chances of conversion. The model to be built in lead conversion rate around 80% or more.

STRATEGY:

- Import data
- Clean and prepare the data for further analysis
- Exploratory data analysis for figuring out most helpful attributes for conversion
- Scaling features
- Prepare the data for model building
- Build a logistic regression model
- Test the model on train set
- Evaluate model by different measures and metrics
- Test the model on test set
- Measure the accuracy of the model and other metrics for evaluation

Plots and Visualization:

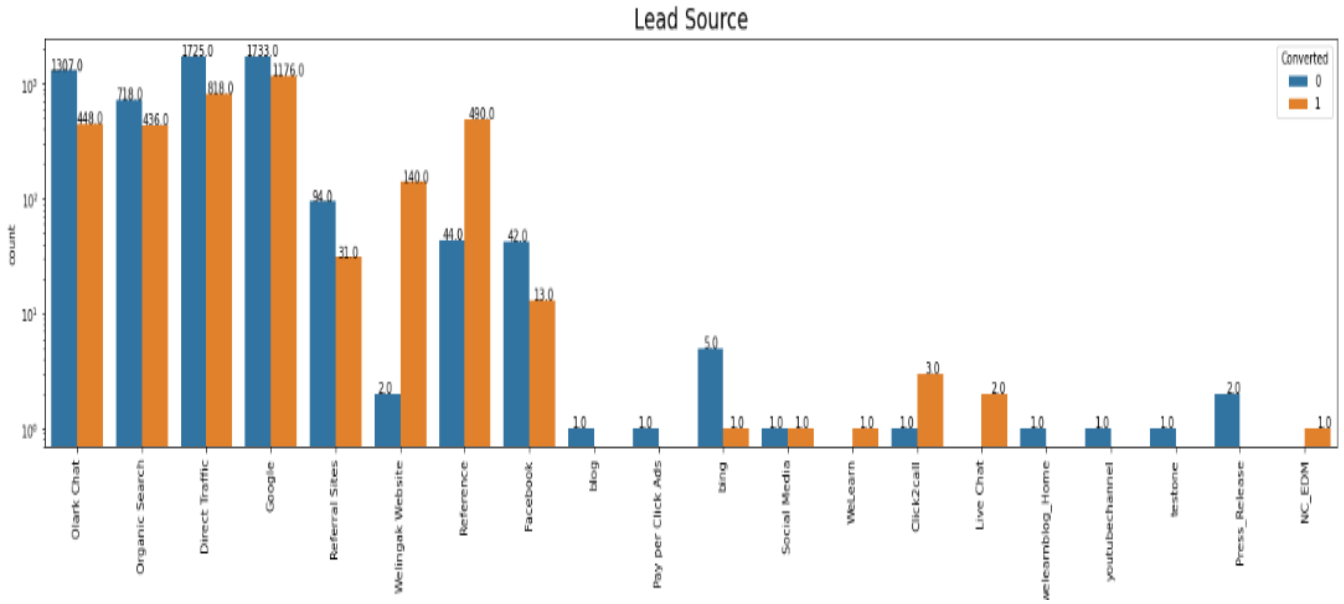
Lead Origin Vs Converted:



From above plot we can observed that:

- Conversion rate for 'API' is ~ 31% and for 'Landing Page Submission' is ~36%.
- For 'Lead Add Form' number of conversion is more than unsuccessful conversion.
- Count of 'Lead Import' is lesser.

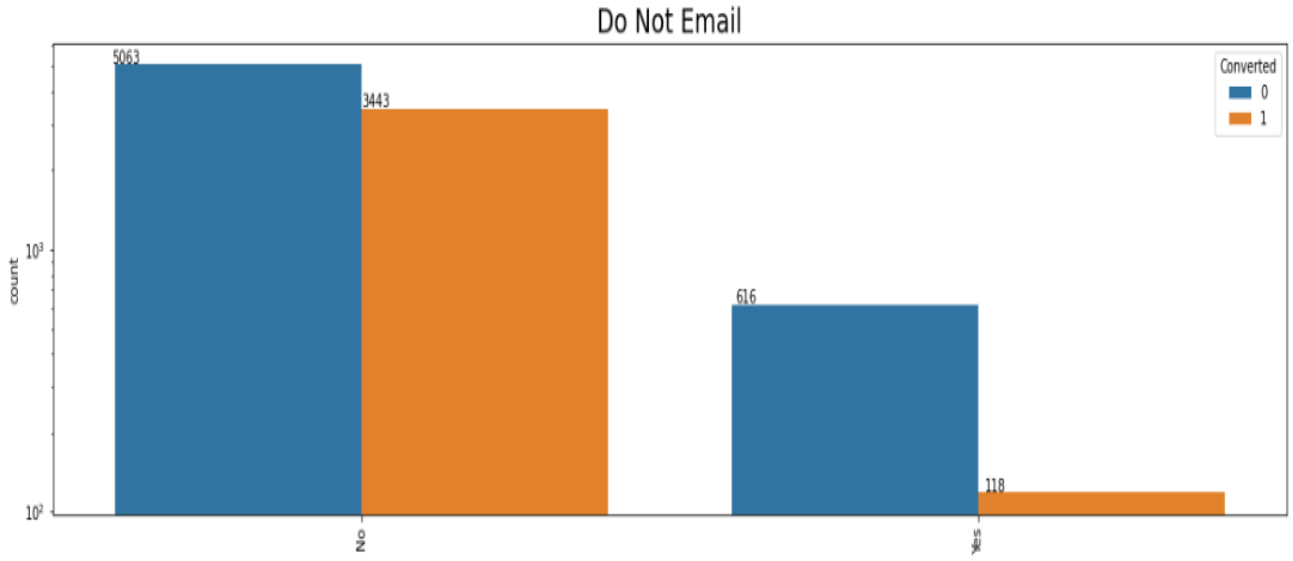
Lead Source Vs Converted:



From above plot we can observed that:

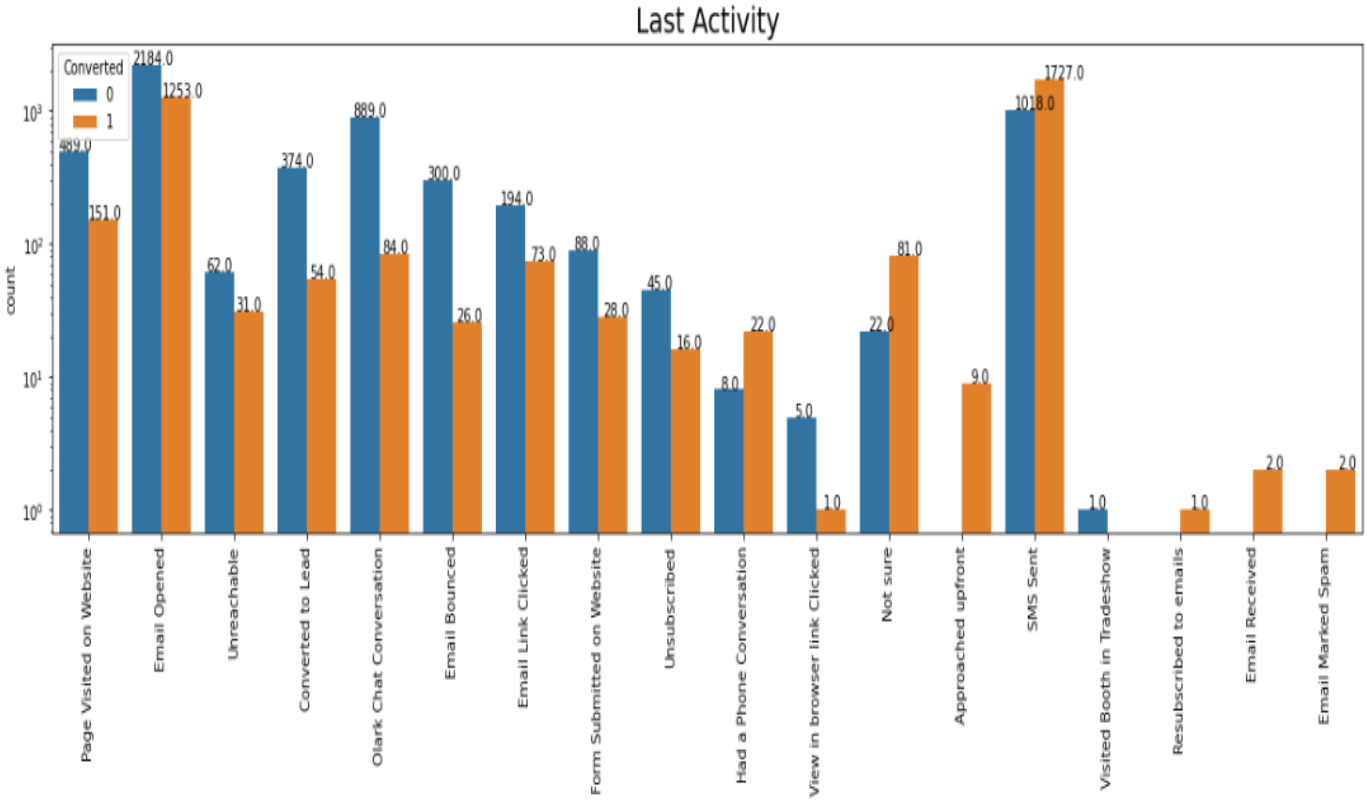
- Google and Direct traffic generates maximum number of leads.
- Conversion rate of 'Reference' and 'Welingak Website' leads is high.

Do not Email Vs Converted:



People who opted for mail option are becoming more leads.

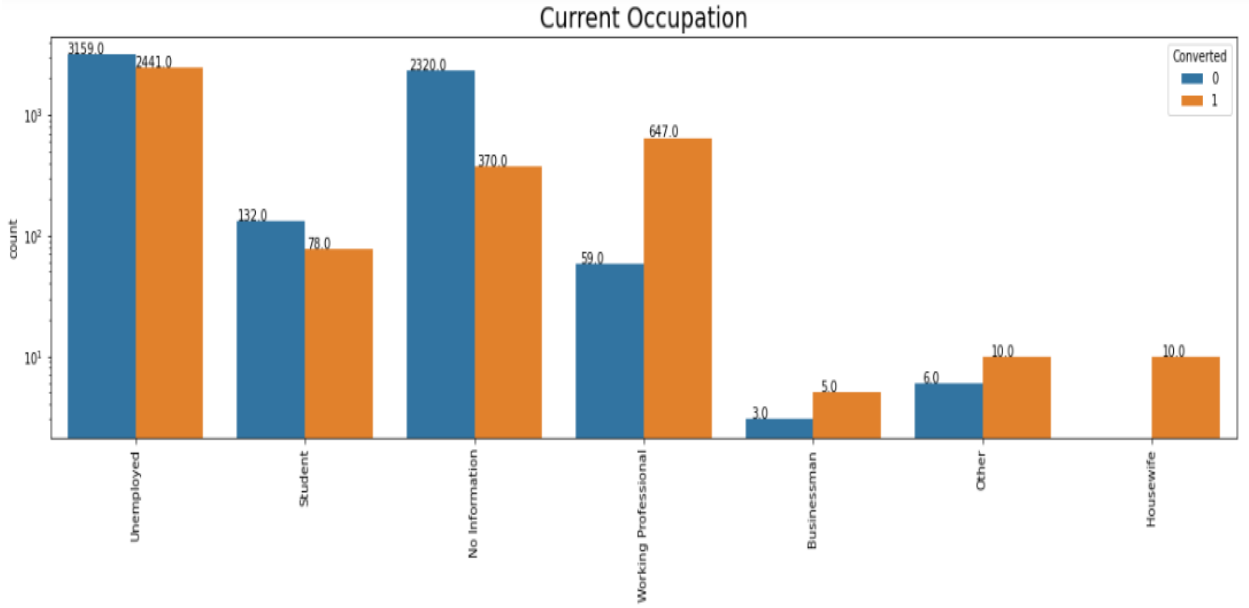
Last Activity Vs Converted:



From above plot we can observed that:

- Conversion rate for last activity of 'SMS Sent' is ~63%.
- Highest last activity of leads is 'Email Opened'.

Current Occupation Vs Converted:

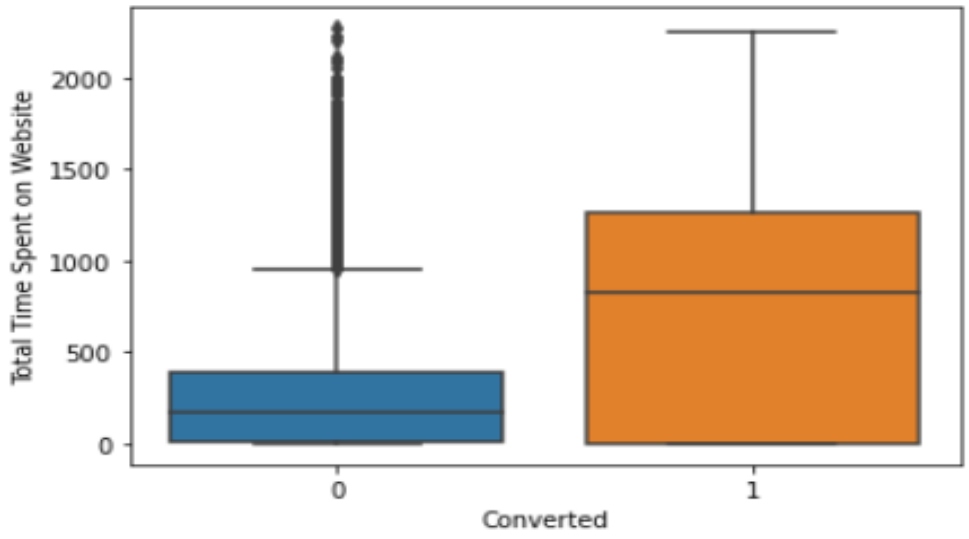


From above graph we conclude that:

- 'Unemployed' leads are generating more number of leads and having ~45% conversion rate.
- Conversion rate is higher for 'Working Professionals'.

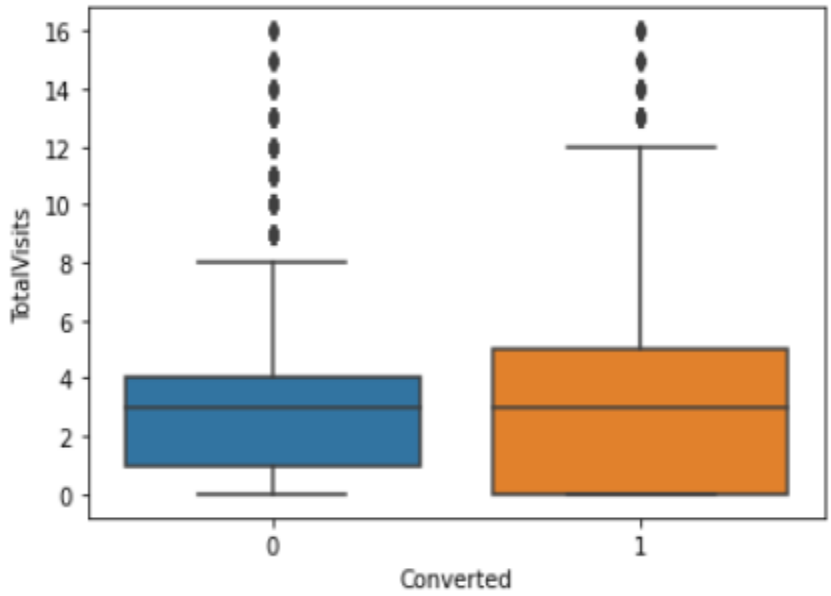
Outliers:

Converted Vs Total Time Spent on Website:



Leads spending more time on website are more likely to opt for courses or converted.

Converted VS TotalVisits:

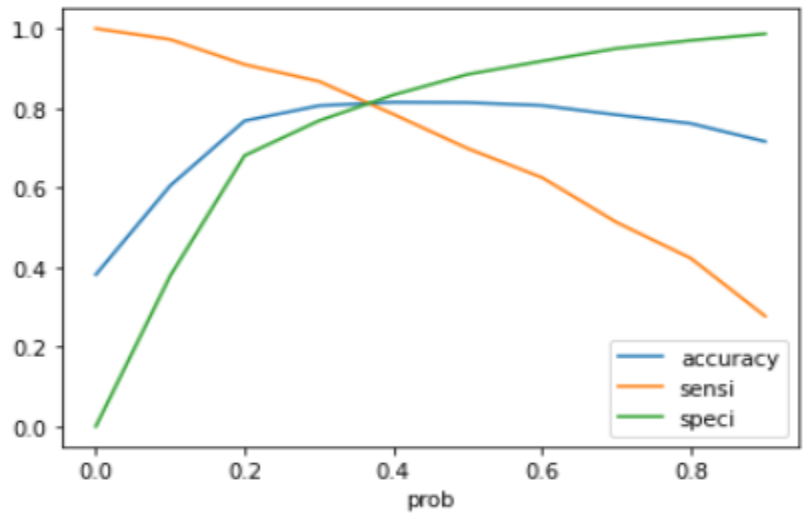


From above plot we can see that the median for Converted and Non-Converted is approximately same.

Correlation Matrix:

- TotalVisits' and 'Page Views per Visit' are highly correlated with correlation of 0.72.
- 'Total Time Spent on Website' has correlation of 0.36 with target variable 'Converted'.

Model Evaluation - Sensitivity and Specificity for Train Dataset:



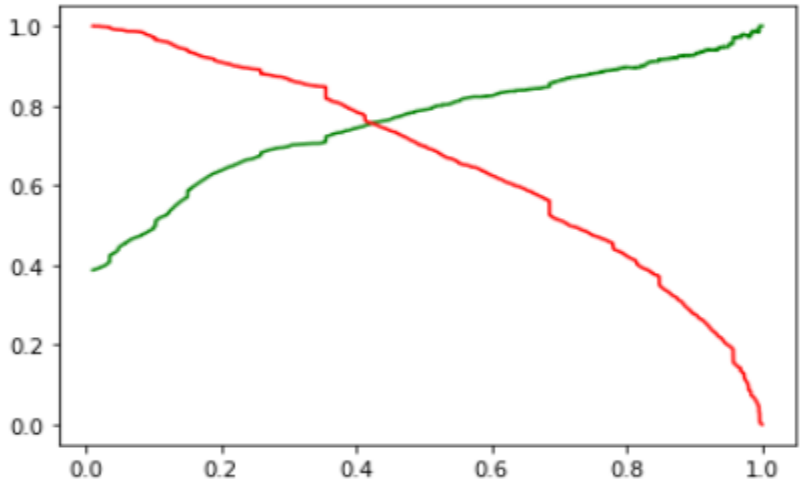
The graph described an optimal cut off of 0.37 based on Accuracy, Sensitivity and Specificity.

Confusion Matrix:

3459	450
727	1684

Accuracy: 81%
Sensitivity: 70%
Specificity: 88%

Model Evaluation - Precision and Recall for Train Dataset:



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

Confusion Matrix:

3459	450
727	1684

Precision: 79%
Recall: 70%

Model Evaluation - Sensitivity and Specificity for Test Dataset:

Confusion Matrix:

1330	313
200	866

Accuracy: 81%
Sensitivity: 81%
Specificity: 80%