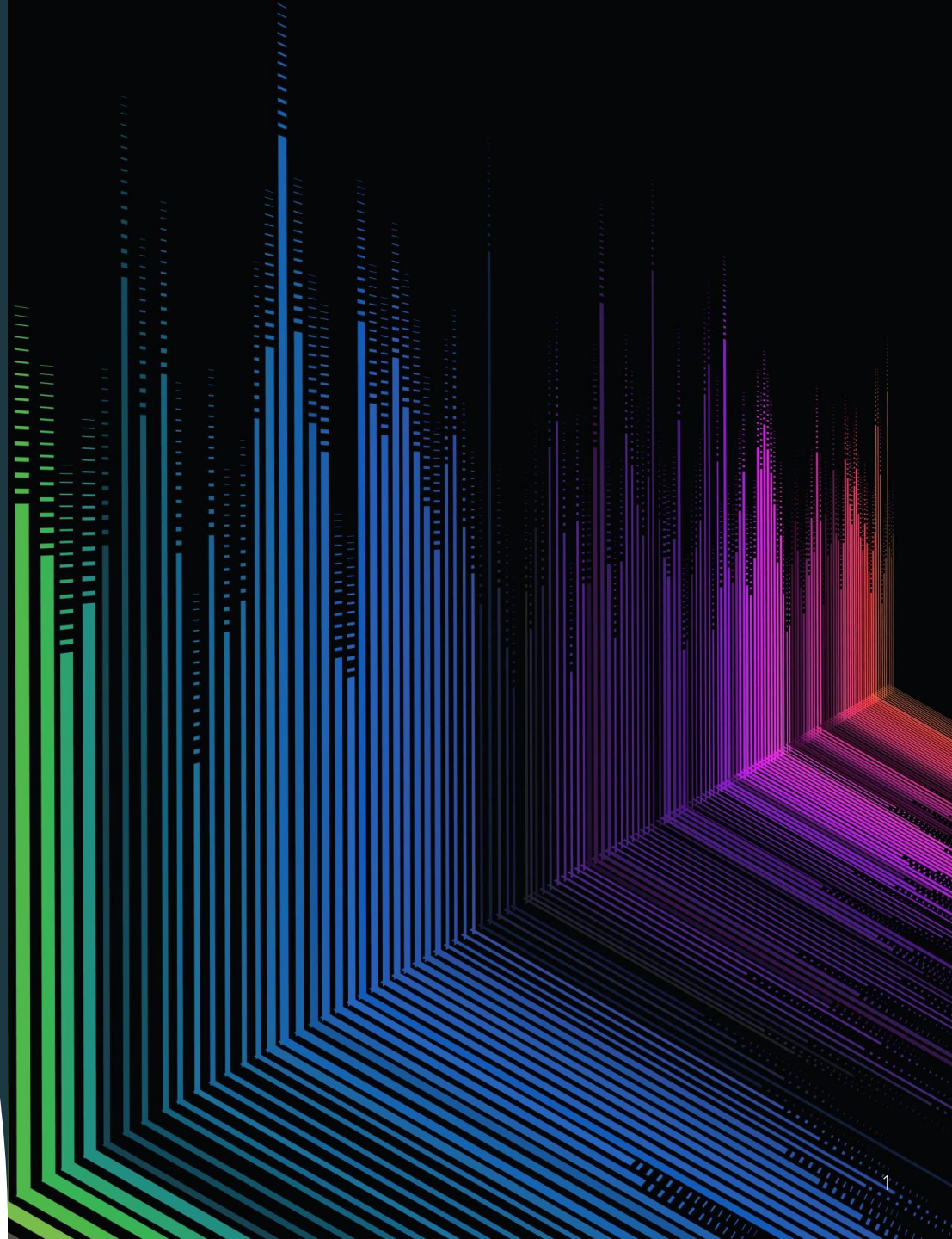
Lead_Scoring_Case_Study

By

Thejus VS and PremKumar



Problem Statement

X Education is an organization which provides online course to Industry professionals. The company markets its Courses on several websites and search engines like Google.

X Education wants to select most promising leads that can be converted to paying customers.

Although X education gets a lot of leads, its lead conversion rate is very poor, where in the company wants a higher lead conversion. Leads come through numerous mode like email, advertisements on websites, google search etc.

The company has had 30% conversion rate through the whole process of turning lead into customers by approaching those leads which are to be found having interest in taking the course. The implementation process of lead generating are not efficient in helping conversion.

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.

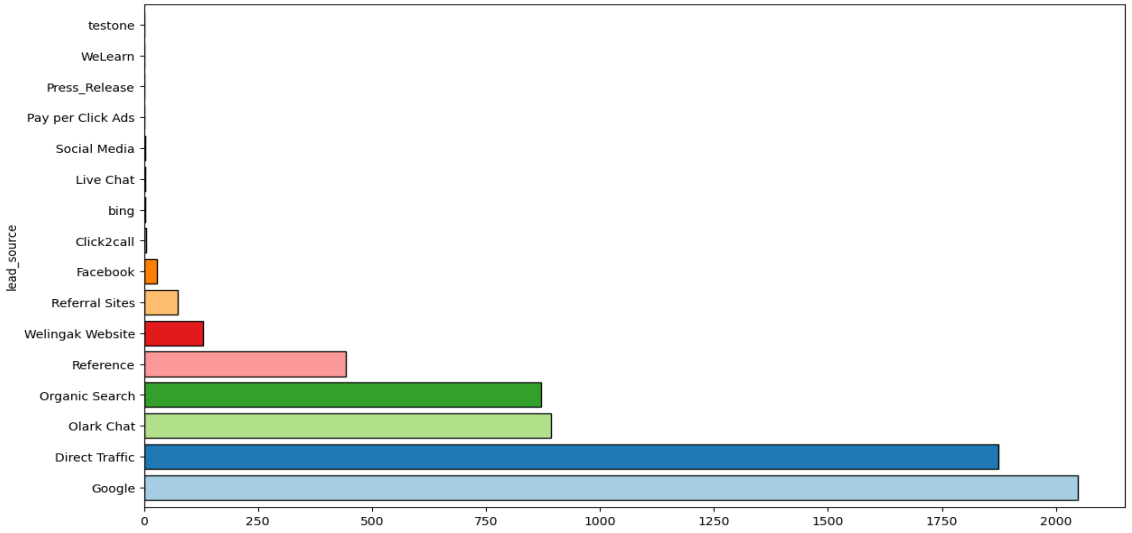
Strategy

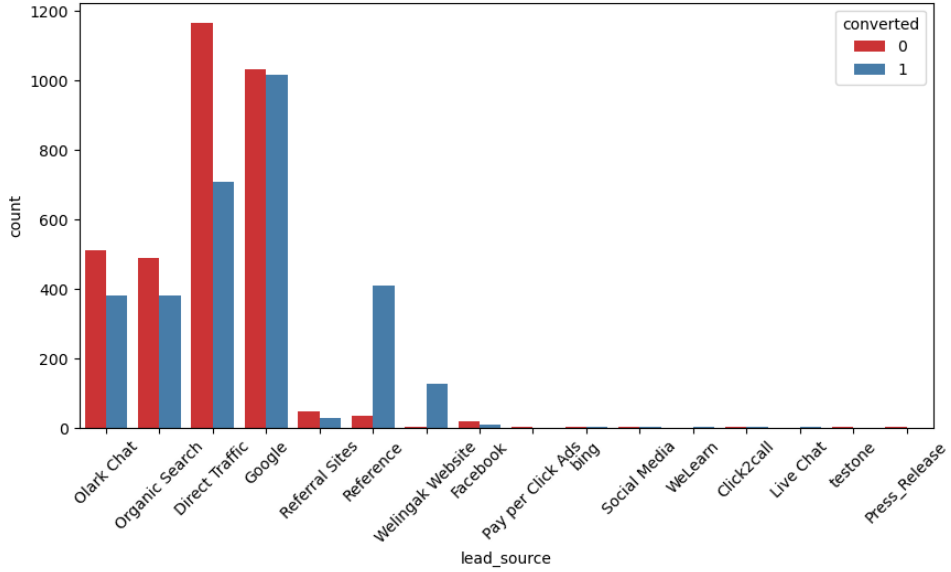
- Import data
- Data Cleaning and Preparation for further analysis
- Exploratory data analysis for figuring out most helpful attributes for conversion
- Scaling features Preparing Categorical and Numerical variable analysis
- Prepare the data for model building
- KDE Mean square transformation
- Finding Outliers
- Dummy variable Creation
- Test the model on train set
- Looking at the Correlations
- Model Building
- Model Evaluation
- Precision-Recall View
- Precision and and Recall Trade Off
- Making Prediction on the Test set
- Summary

Exploratory Data Analysis

Lead Source VS Converted

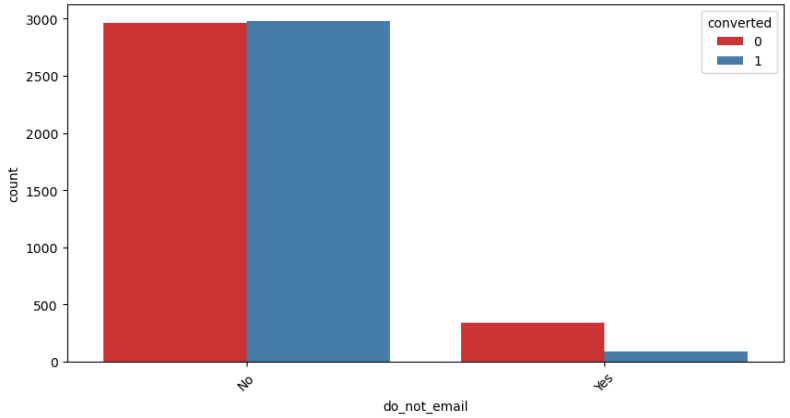
Google search has had high conversion compared to other modes. Behind that Welligak website Direct Traffic and Olark chat modes are also in contribution.

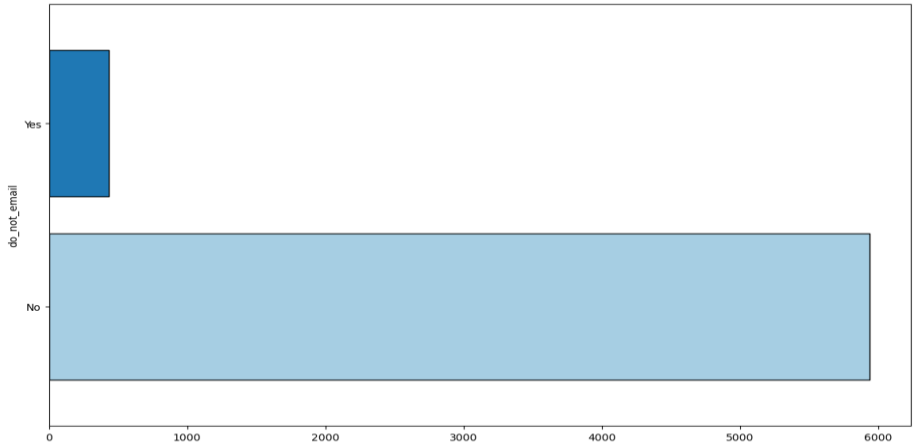




Do Not Email VS Converted

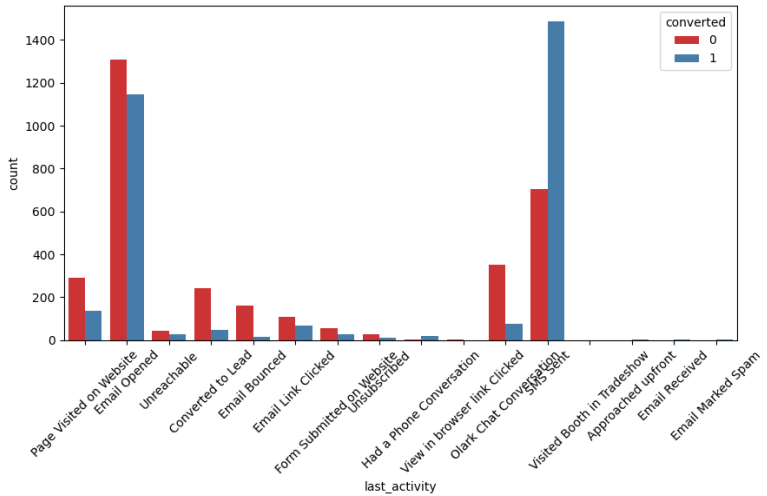
Most leads do not prefer to be informed through Email modes.

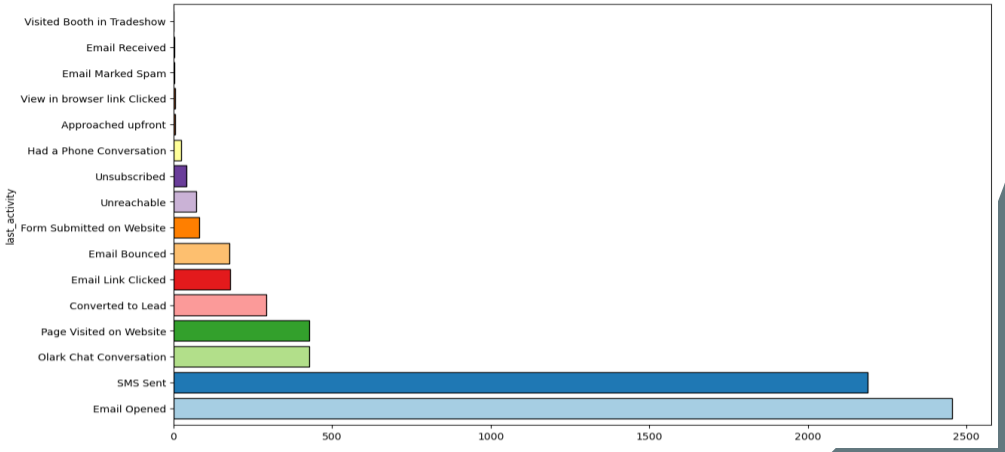




Last activity VS Converted

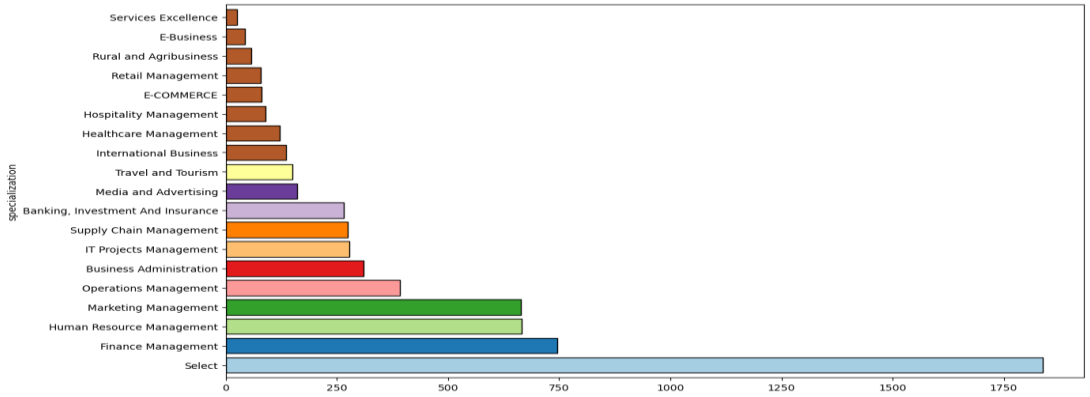
The chance are in higher for conversion to those Email opened and SMS sent.

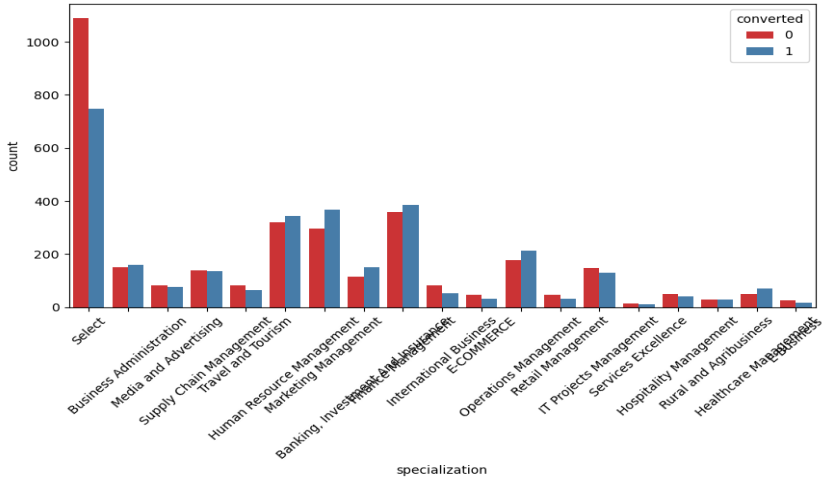




Specialization VS Converted

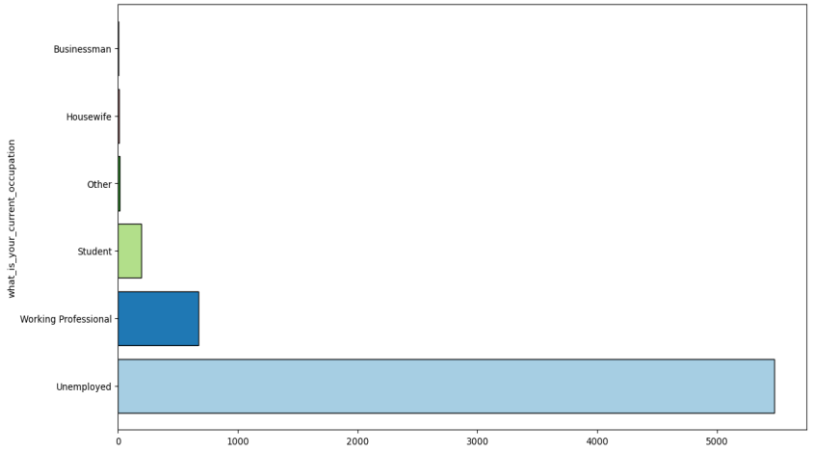
Most leads prefers Finance Management , Human Resource Management and Marketing Management

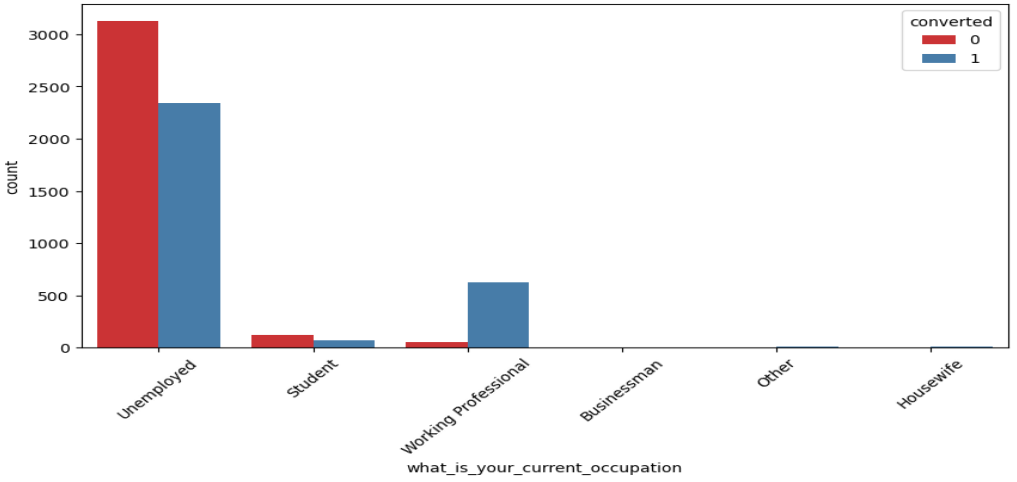




What is your current Occupation VS Converted

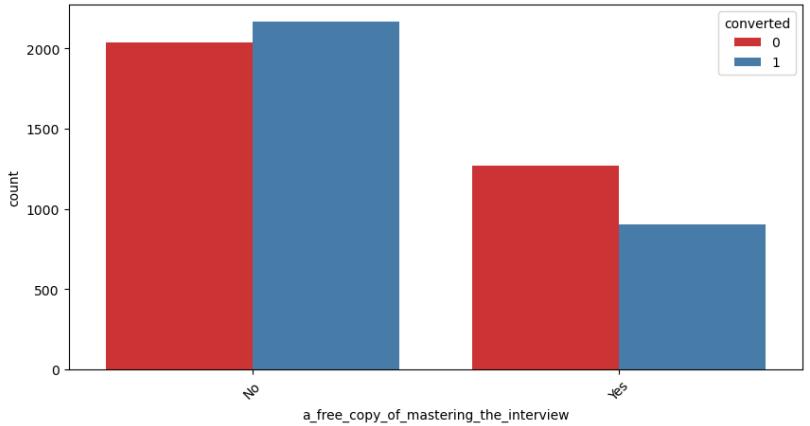
Those are Working professional having higher conversion that rest of the categories. Unemployed having comparatively more Conversion and leads that off Students.

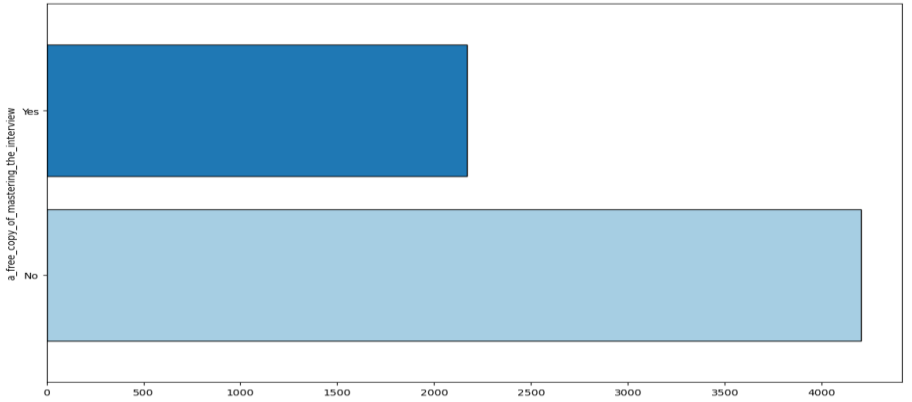




A Free Copy Of Mastering The Interview VS Converted

A free copy of mastering the interview had lesser preference and not shows as important to a lead conversion.

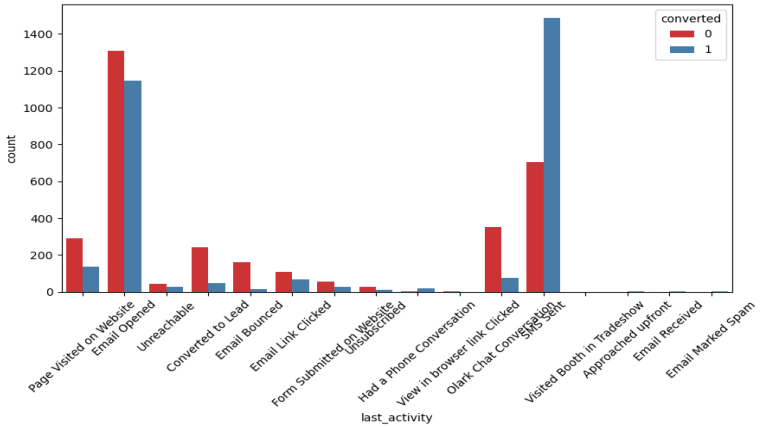


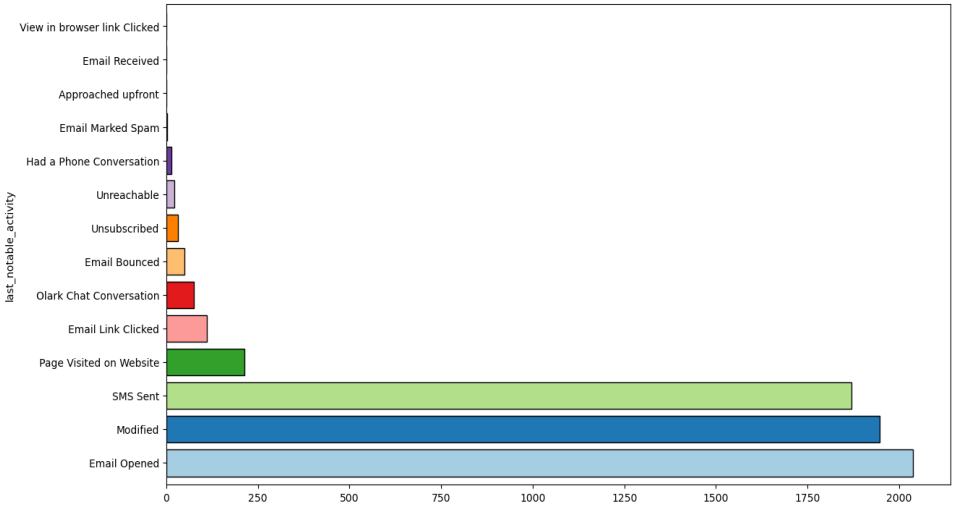


The Last Notable activity VS Converted

The people who opened Email , modified and SMS sent having

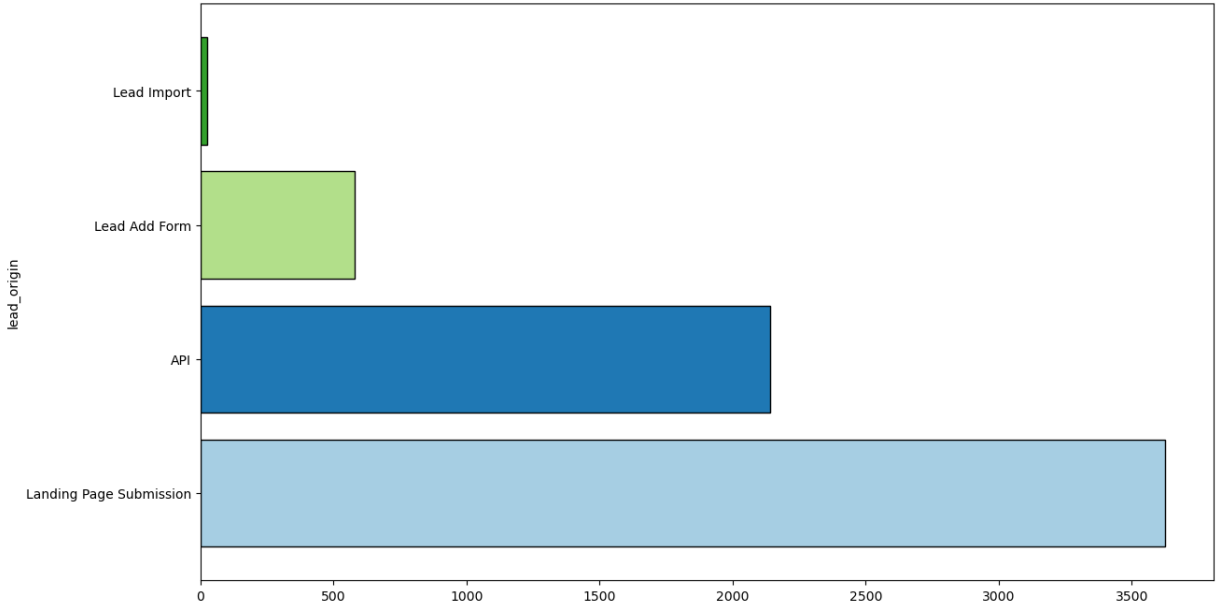
Higher chance of being a promising lead.

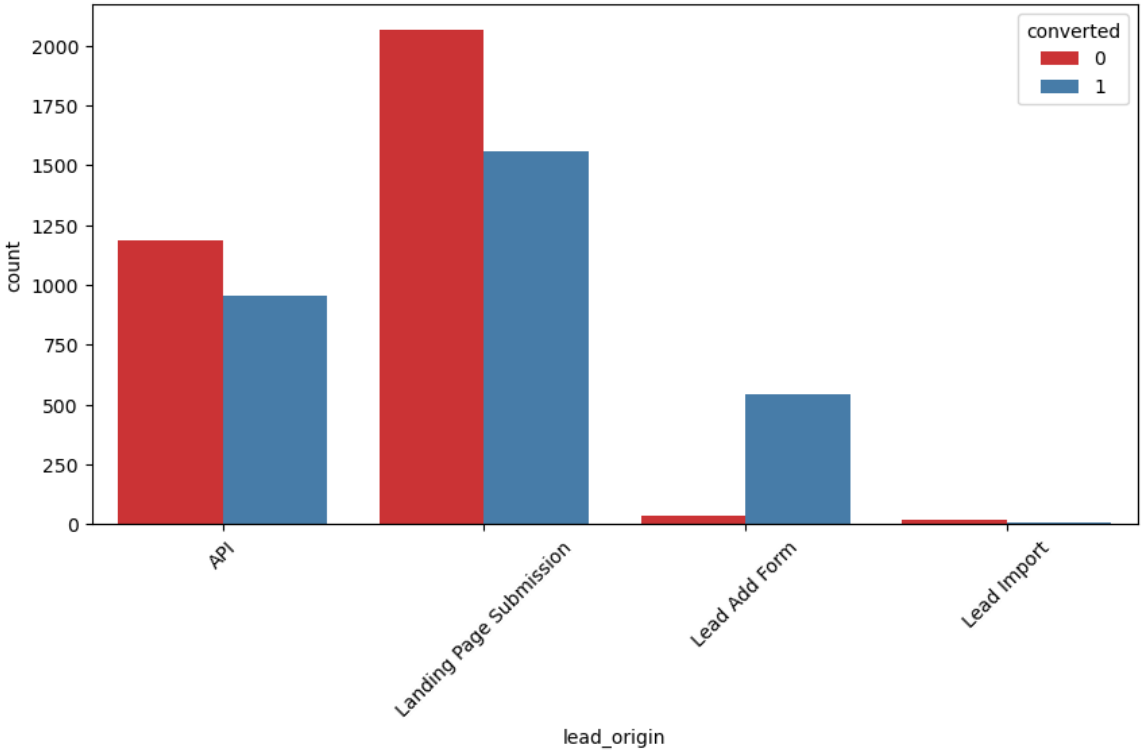




Lead Origin VS Converted

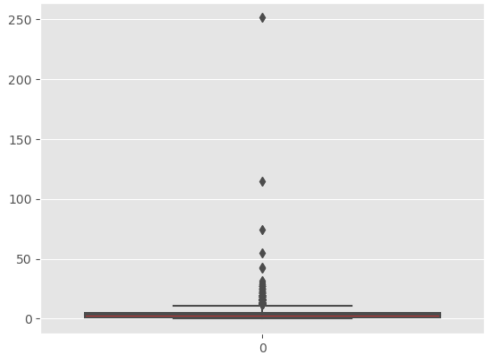
The conversion rate is high in Lead Add form, whereas other likes, submission page and API have more numbers but also, they have less attractive say ‘NO’ more number of times

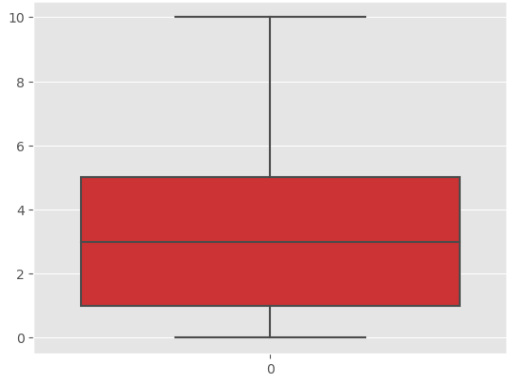


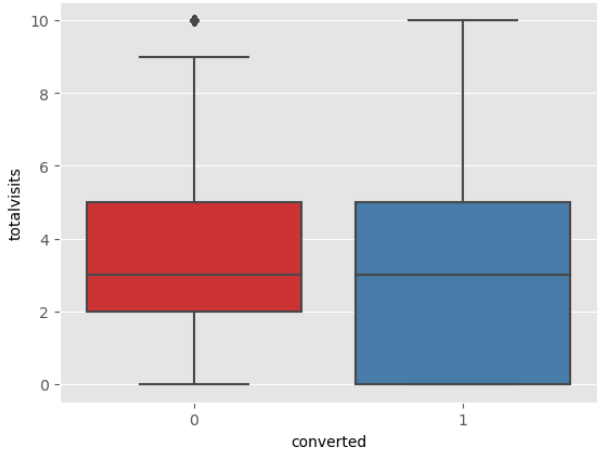


Outlier Treatment

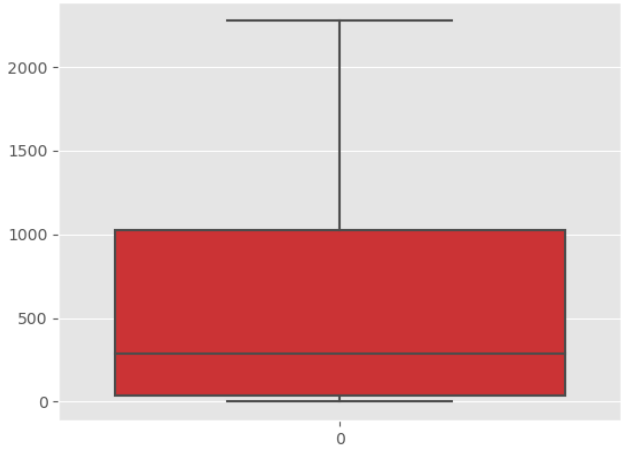
Total Visit

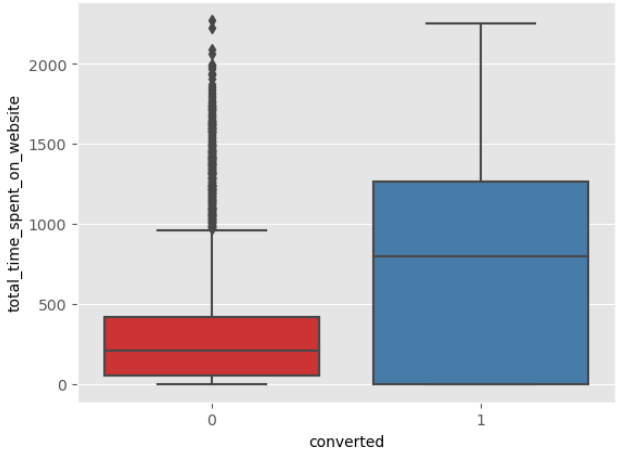




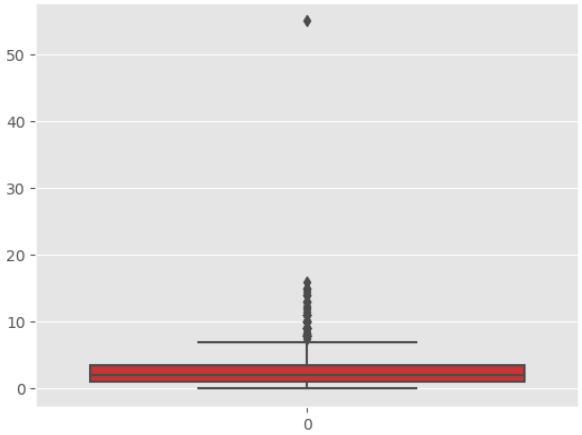


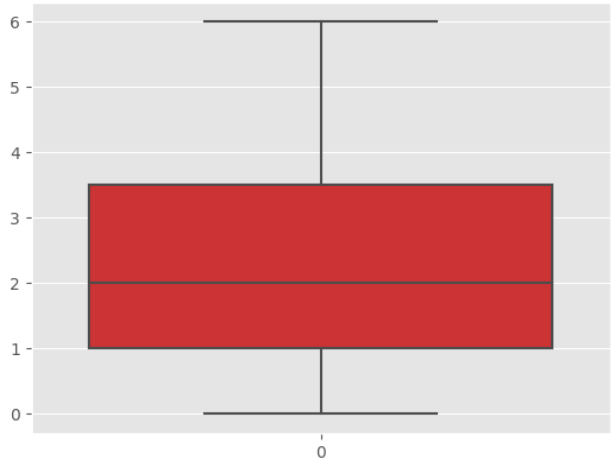
Total time spent on website

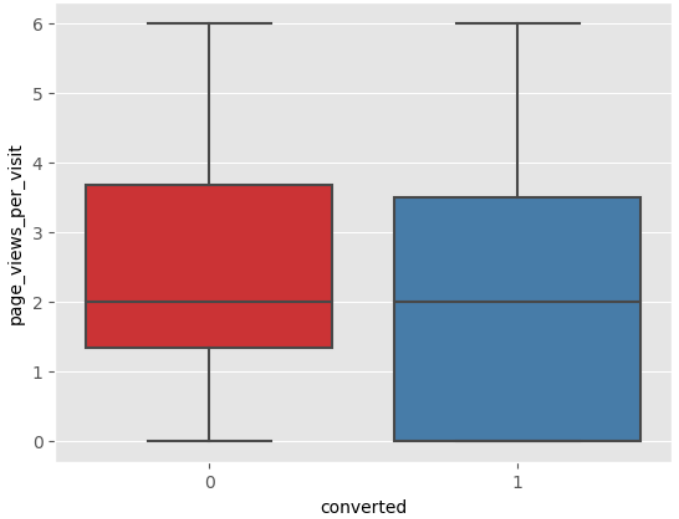




The plots represent before and after outliers' treatment for each of the variables totalvisit, total time spend on website and page views per visit against convert variable .
Though outliers in Total time spend on website shows valid values, this will misclassify the outcomes consequently create problems when making inference with wrong model. Logestic regression heavily influenced by outliers. Those data having heavily with outliers cap it to 95% value of analysis





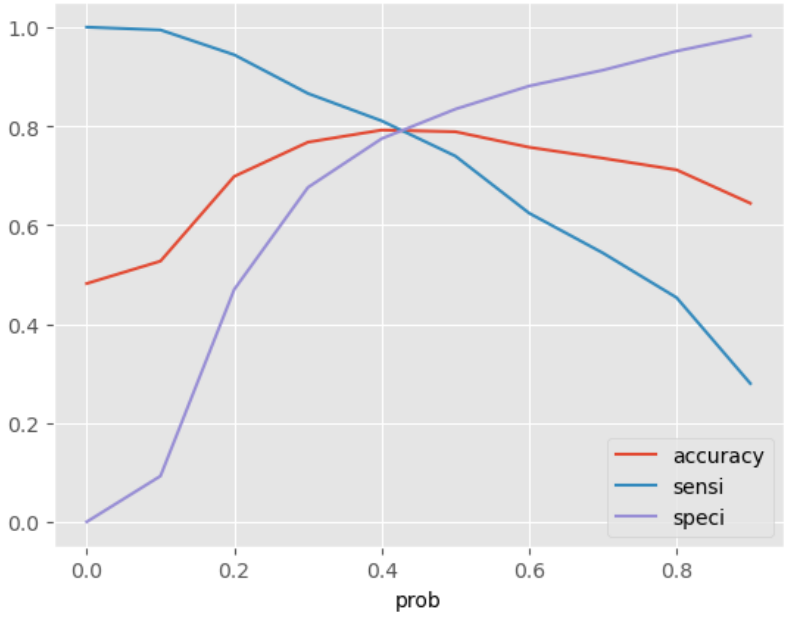


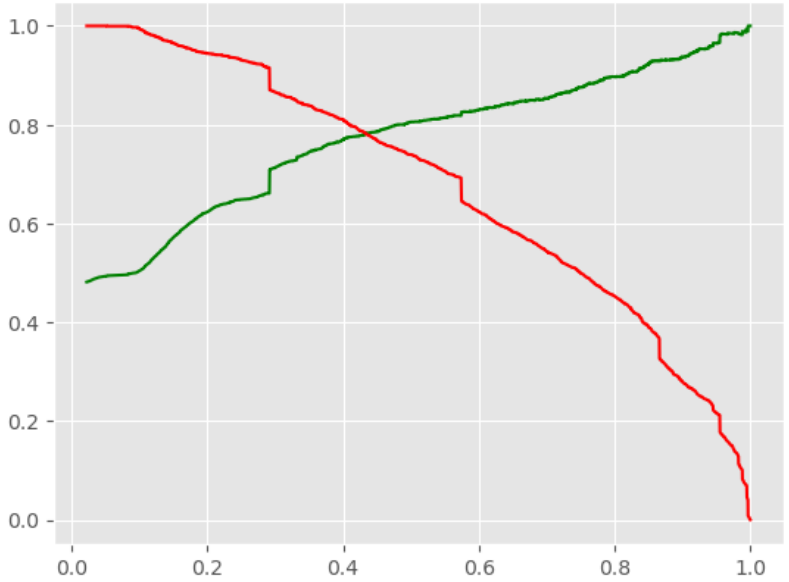
Page views per visit

Model Building

- Splitting Into Train and Test Set Data.
- Scale variables in train set.
- Build the first model.
- USE RFE to eliminate less relevant variables.
- Build next model.
- Eliminate variables for all the existing columns.
- Predict using train set.
- Create Confusion metrics
- Precision and recall analysis on test prediction.

Model Evaluation





Accuracy Sensitivity and Specificity

- 79.00 % Accuracy
- 79.43% Sensitivity
- 78.67% Specificity
- can see that around 0.42, you get the optimal values of the three metrics. So , choose 0.42 as our cutoff now.

1819 493

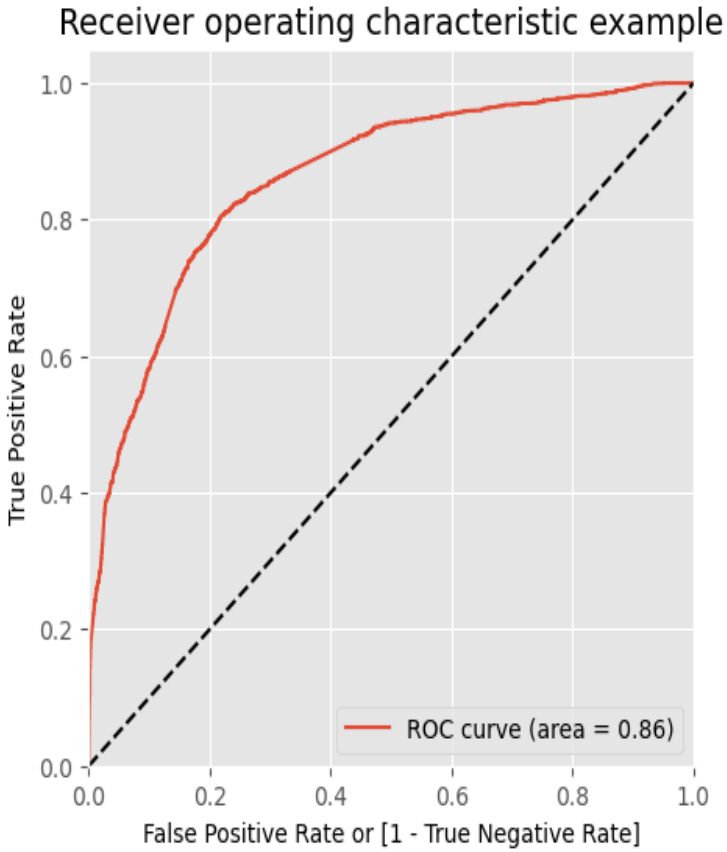
442 1707

Precision and Recall

- 78.40% Precision
- 78.36% Recall

Finding Optimal Cut-Off

- **ROC CURVE**



The area under the curve of the ROC is 0.86 which is quite good. So we seem to have a good model.

Conclusion

EDA:

- The people spending higher time(totalvisit) than average time are promising leads ‘(Total Time Spent on Website' , 'Page Views Per Visit’), so targeting them and approaching them can be helpful in conversions.
- Target those from google ,Direct traffic and Olark chat, Organic search are great internet tools to get promising lead contact.
- SMS messages have a high impact on lead conversion.
- Working professional conversion is significantly higher as comparing to students. Whereas unemployed people as they prefer to take course but less chances and slighter chance to become a prospective lead.
- Need to follow up those prospective leads who come through Lead Add form and consider those landed on the submission page and API.

Logistic Regression Model:

- The model shows high close to 79% accuracy.
- The Threshold has been selected from Accuracy, Sensitivity, Specificity measures and precision, recall curves.
- The model shows 78.40% Sensitivity and 78.80% Specificity.
- The model finds promising leads and leads that have less chance of getting converted.
- Overall, this model proves to be accurate.