LEAD SCORING CASE STUDY

(LOGISTIC REGRESSION)

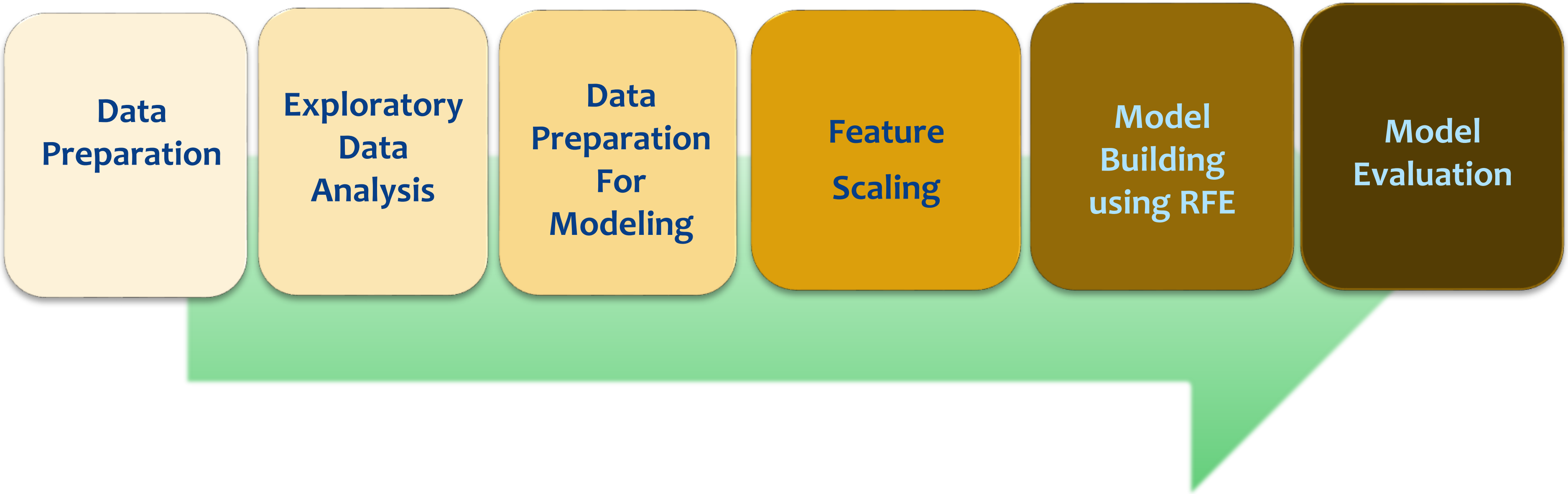
Business Problem Statement :

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%.

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 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%.

Technical Approach :



Data Preparation

Exploratory Data Analysis

Data Preparation For Modeling

Feature Scaling

Model Building using RFE

Model Evaluation

Data Understanding

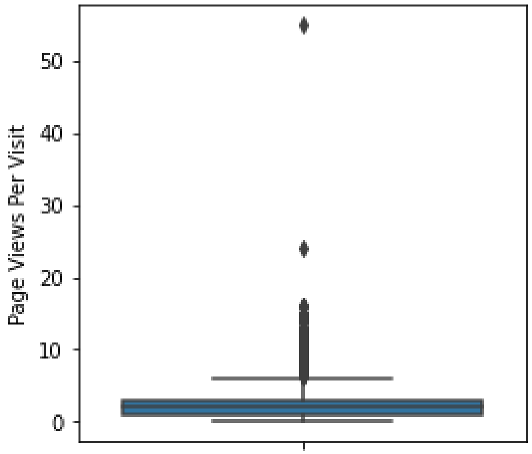
Data Preparation :

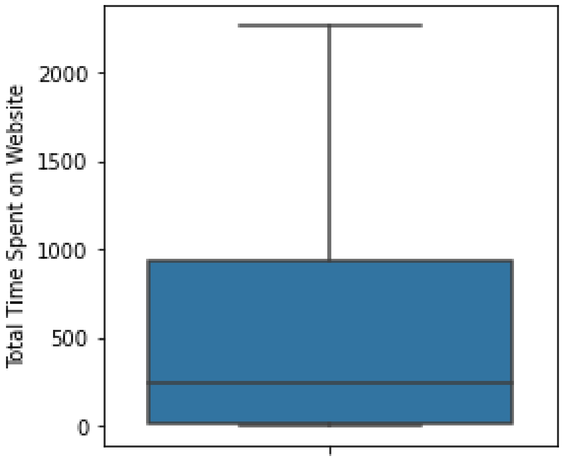
- There are some categorical features having a label as “SELECT”. This means the person might not have selected any value for that field. Hence this is as good as a missing value. So converting SELECT into the NaN.
- After identifying all the missing data, dropped columns having more than 70% null values.
- As the Lead Quality depends upon the intuition of the employee, it will be safer to update the NaN to “Not Sure”.
- There are too many variations in the columns (‘Asymmetrique Activity Index’,’Asymmetrique Activity Score’,’Asymmetrique Profile Index’,’Asymmetrique Profile Score’) and it is not safer to impute any values in the columns and hence we will drop these columns with very high percentage of missing data.
- We can impute the MUMBAI into all the NULLs as most of the values belong to MUMBAI.
- Since there is no significant difference among top 3 specialisation , hence it will be safer to impute NaN with Others.
- For Tags column, more than 30% data is for “Will revert after reading the email” and hence we can impute NULLS with Will revert after reading the email.
- More than 99% data is of “Better Career Prospects” and hence it is safer to impute NULLS with this value.
- More than 85% data is of “Unemployed” and hence it is safer to impute NULLS with this value.
- More than 95% data is of “India” and hence it is safer to impute NULLS with this value.

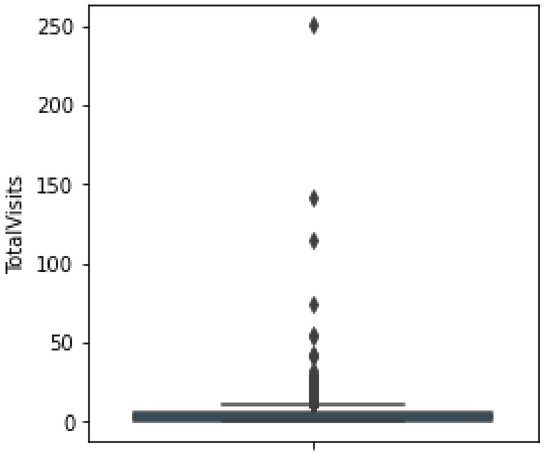
Exploratory Data Analysis

(Univariate Analysis)

Univariate Analysis of Numerical Variable:



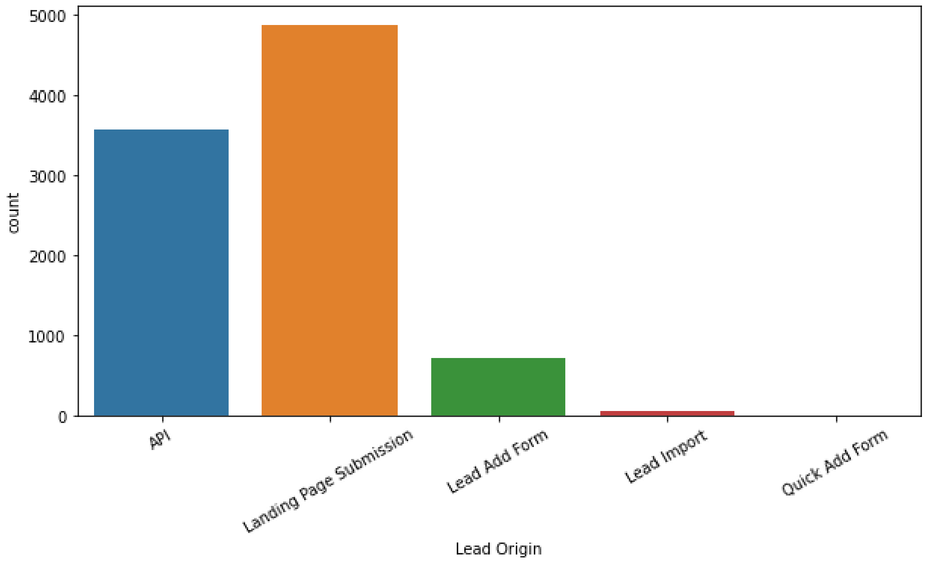


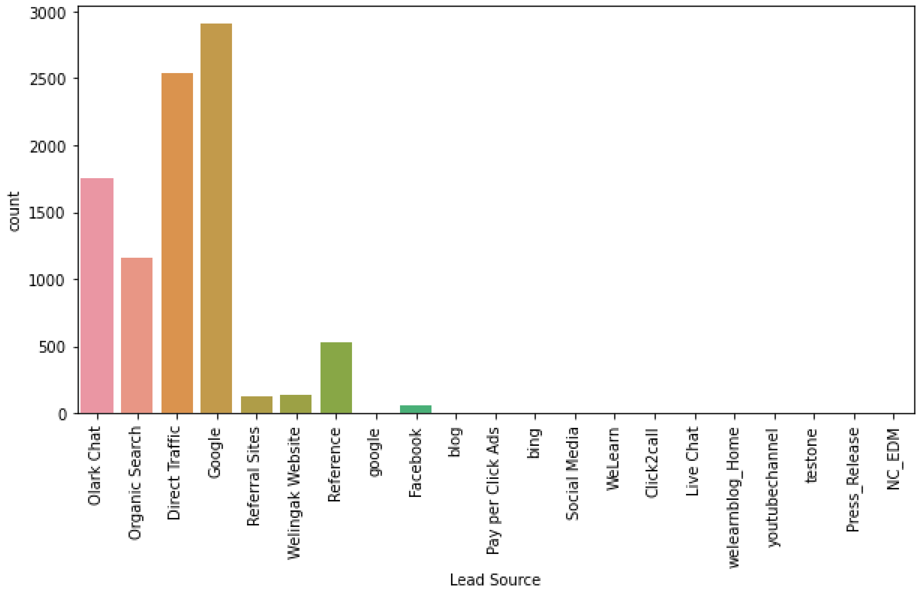


Observation :

- There is higher variation in Total Time Spent on Website.
- There is low variation and lot of outliers in TotalVisits which needs to be treated before modelling.
- There is low variation and lot of outliers in Page Views Per Visit which needs to be treated before modelling.

Univariate Analysis of Categorical Variable:

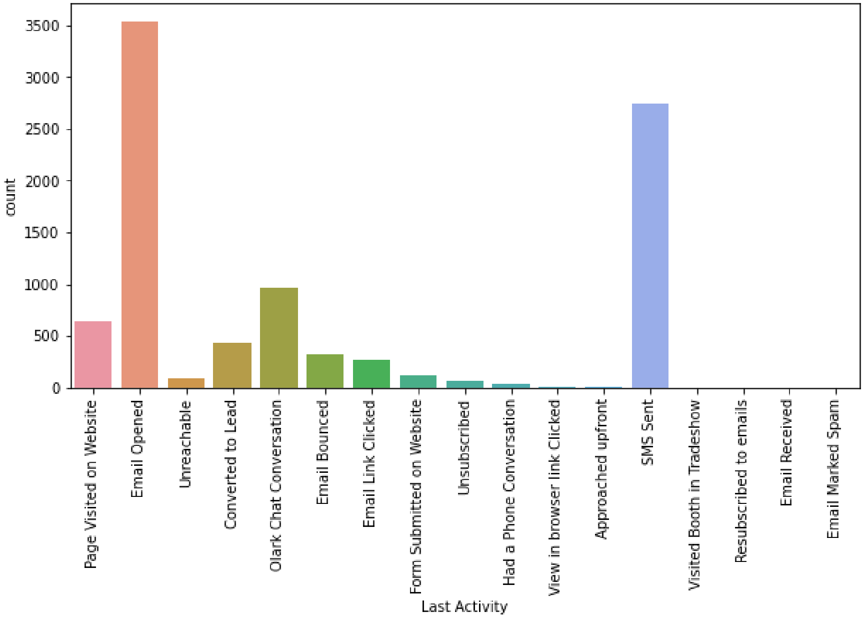


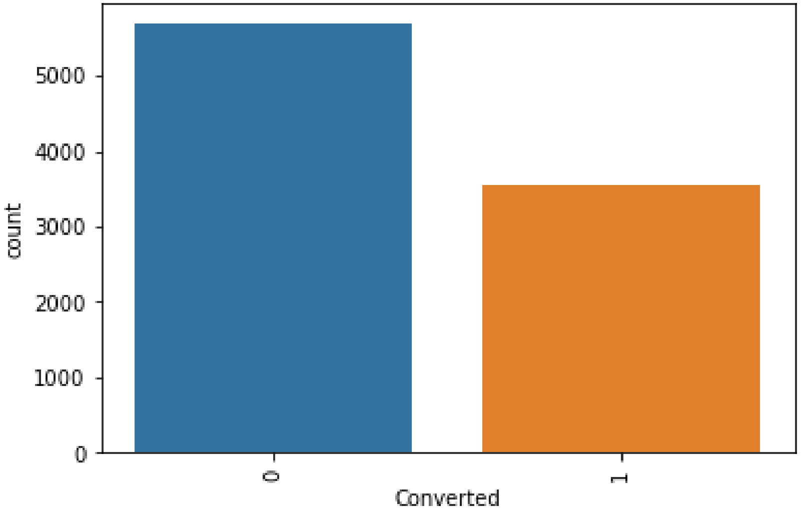


Observation :

- API and Landing Page Submission bring higher number of leads as well as conversion.
- Olark Chat, Organic Search, Direct Taaffic, Google & Reference bring higher number of leads.

Univariate Analysis of Categorical Variable:





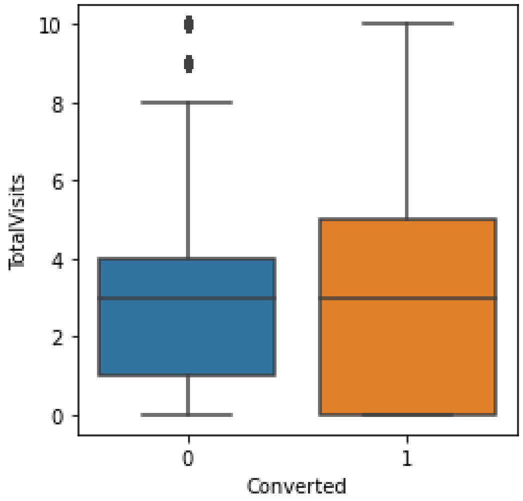
Observation :

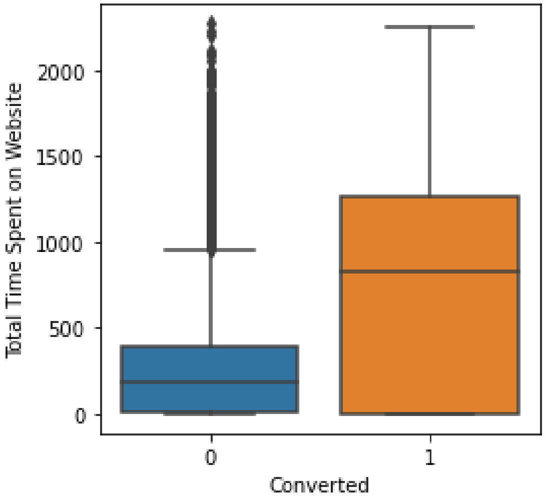
- Email Opened & SMS Sent bring higher number of leads.

Exploratory Data Analysis

(Bivariate Analysis)

Bivariate Analysis Numerical Variable:





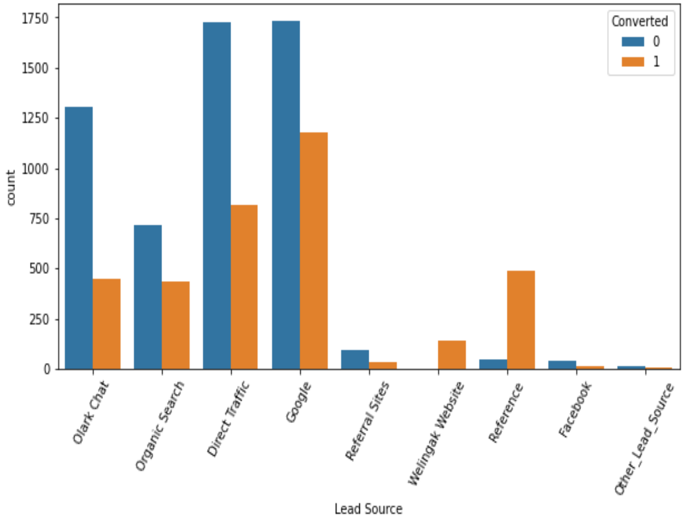
Observation :

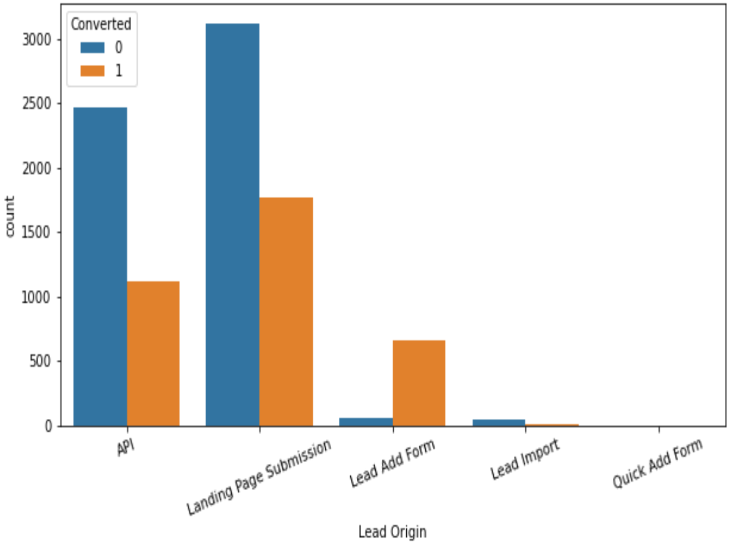
- We didn't get any conclusion from above information since their median of conversion and non-conversion are same.
- Users spending more time on the website are more likely to get converted.

Suggestion:

Websites can be made more appealing so as to increase the time of the Users on websites

Bivariate Analysis Categorical Variable:





Observation:

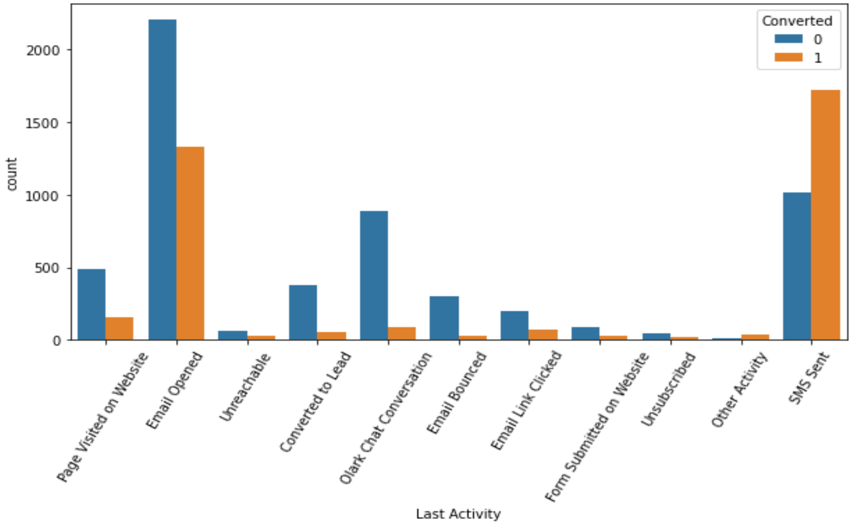
- The count of leads from the Google and Direct Traffic is maximum.
- The conversion rate of the leads from Reference and Welingak Website is maximum.
- API and Landing Page Submission has less conversion rate but counts of the leads from them are considerable.
- The count of leads from the Lead Add Form is pretty low but the conversion rate is very high.
- Lead Import has very less count as well as conversion rate, hence can be ignored.

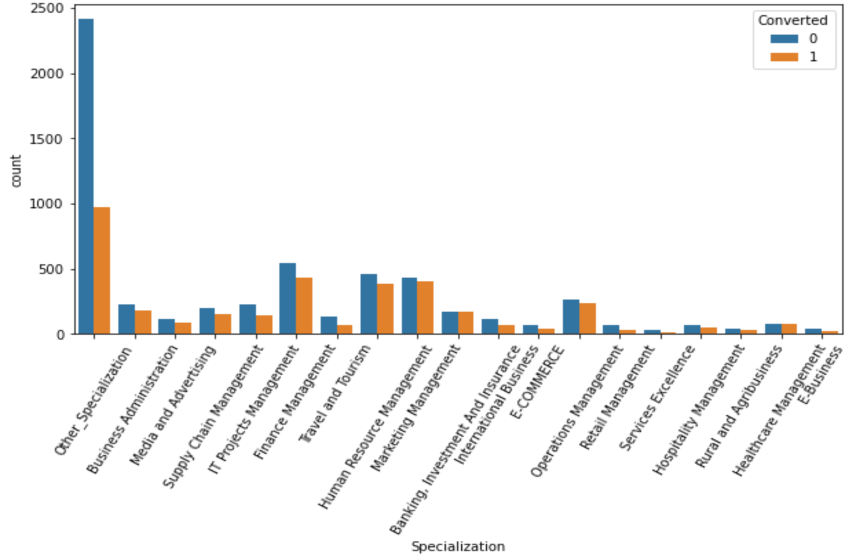
Suggestion:

To improve the overall lead conversion rate, we need to focus on increasing the conversion rate of 'Google', 'Olark Chat', 'Organic Search', 'Direct Traffic' and also increasing the number of leads from 'Reference' and 'Welingak Website‘.

To improve the overall lead conversion rate, we need to focus on increasing the conversion rate of 'API' and 'Landing Page Submission' and also increasing the number of leads from 'Lead Add Form‘.

Bivariate Analysis Categorical Variable:





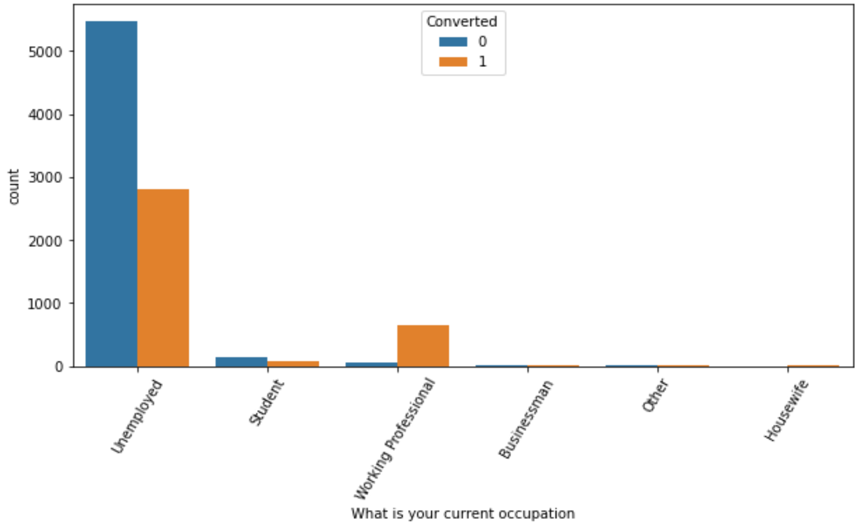
Observation:

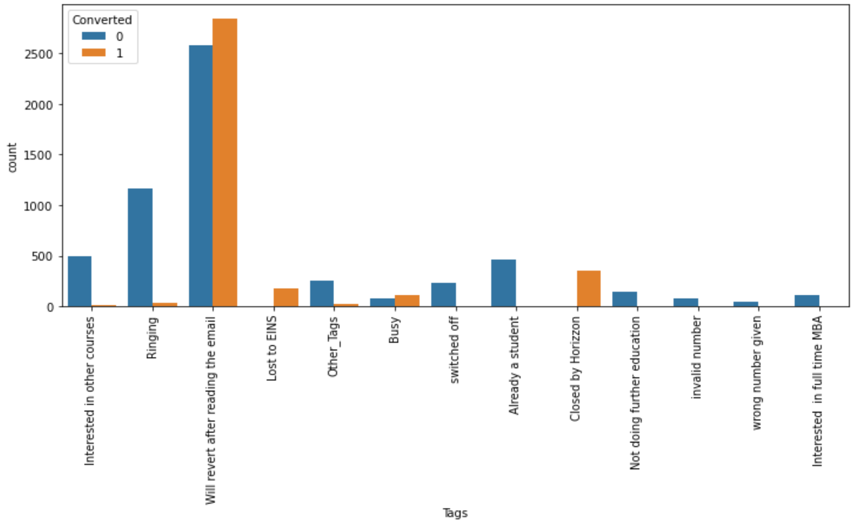
- The count of last activity as "Email Opened" is max
- The conversion rate of SMS sent as last activity is maximum
- Looking at above plot, no particular inference can be made for Specialization

Suggestion

We should focus on increasing the conversion rate of those having last activity as Email Opened by making a call to those leads and also try to increase the count of the ones having last activity as SMS sent.

Bivariate Analysis Categorical Variable:





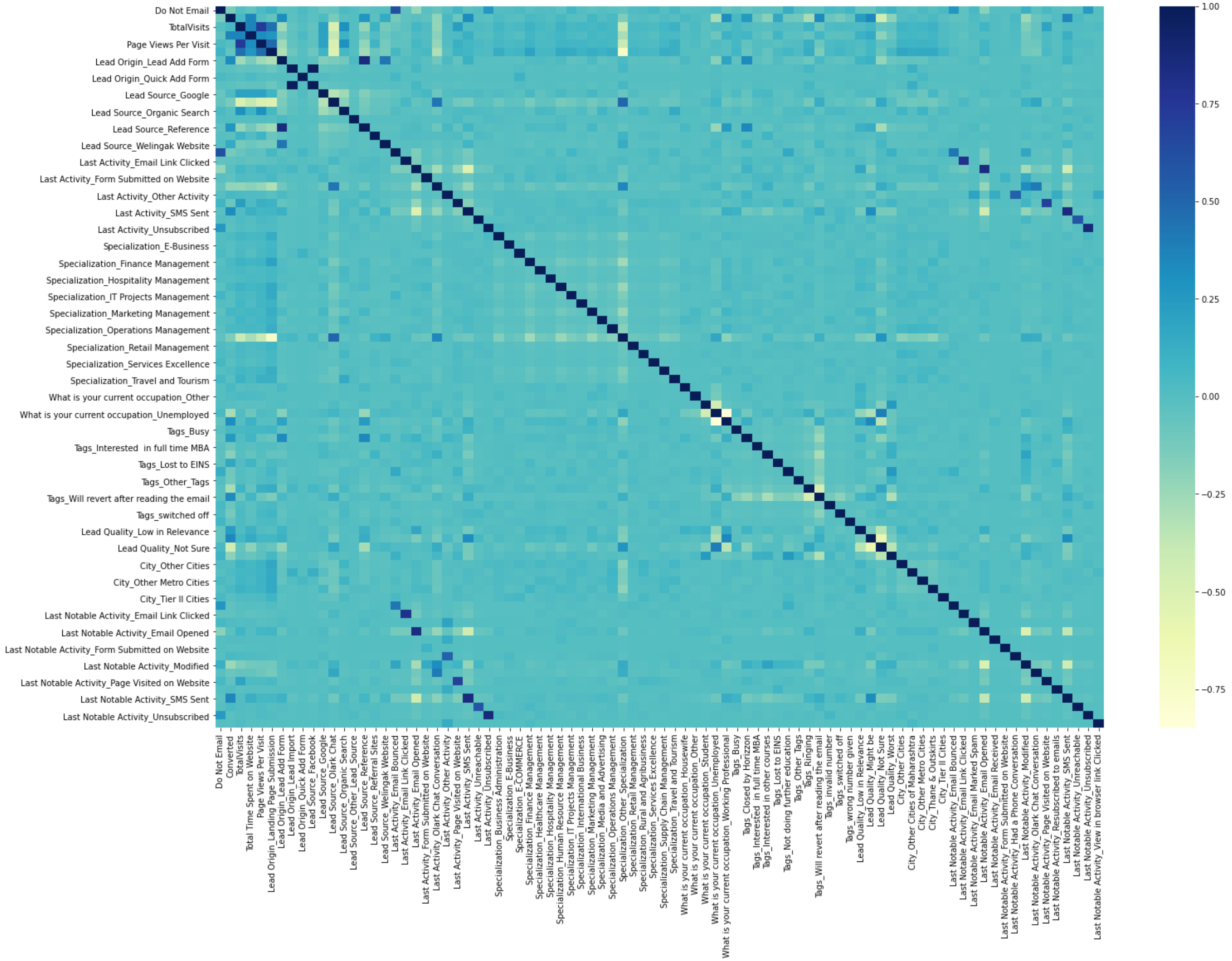
Observation:

- Looking at above plot, we can say that working professionals have high conversion rate also Number of Unemployed leads are more than any other category.
- 'Will revert after reading the email' and 'Closed by Horizzon' have high conversion rate

Suggestion:

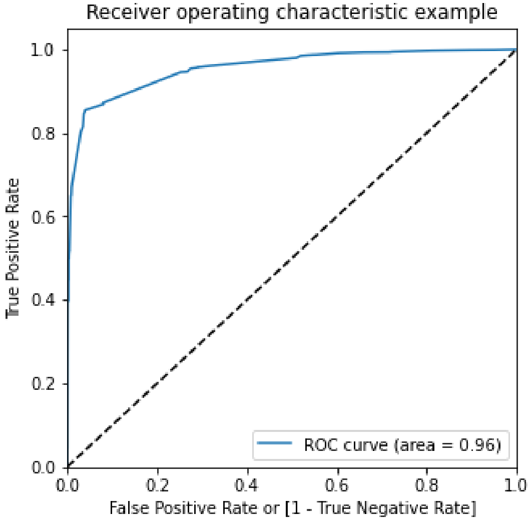
Need to increase the conversion rate of Unemployed leads also we need to increase the number of Working Professional leads by reaching out to them through different social sites such as LinkedIn etc.

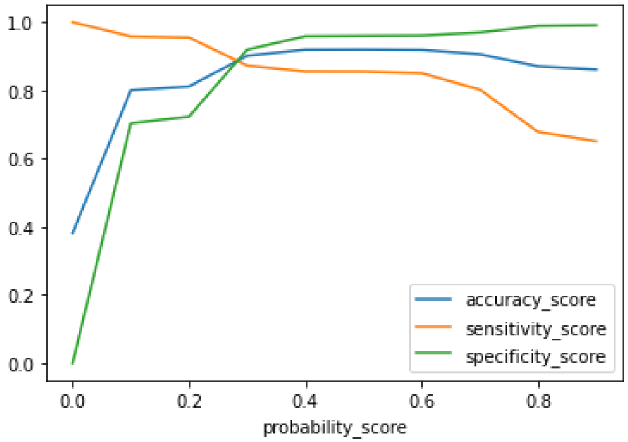
EDA plots depicting correlation (Heat Map) of all selected columns (numerical columns and dummy columns).

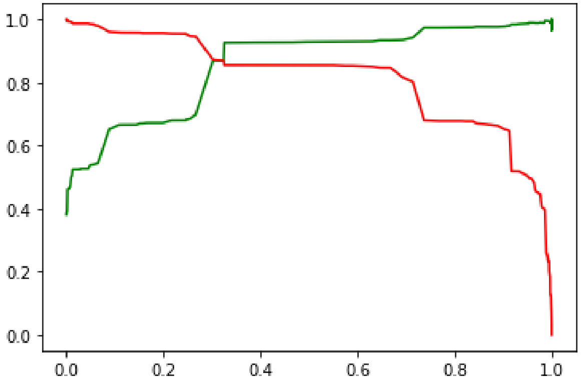


Model Summary

Bivariate Analysis Numerical Variable:







Observation :

- Optimal cutoff probability is that prob where we get balanced sensitivity and specificity
- In Sensitivity-Specificity-Accuracy plot 0.27 probability looks optimal. In Precision-Recall Curve 0.3 looks optimal.
- We are taking 0.27 is the optimum point as a cutoff probability and assigning Lead Score in training data.

Model Analysis

Performance of our Final Model

Final Observation:

Train Data:

Accuracy : 90.15%

Sensitivity : 87.26%

Specificity : 91.92%

Test Data:

Accuracy : 89.93%

Sensitivity : 87.94%

Specificity : 93.23%

The final model has Sensitivity of 0.879, this means the model is able to predict 88% customers out of all the converted customers, (Positive conversion) correctly.

The final model has Precision of 0.86, this means 86% of predicted hot leads are True Hot Leads.

Inferences from Model

Business Insights Derived from our Model

Top 3 variables in model, that contribute towards lead conversion are:

- Total Time Spent on Website
- Tags_Will revert after reading the email
- Lead Origin_Lead Add Form

- Top 3 variables in my model, that should be focused are:

- Last Activity_SMS Sent (positively impacting)
- Last Activity_Olark Chat Conversation (negatively impacting)
- Lead Source_Olark Chat (negatively impacting)

Conclusion

Conclusion:

X Education Company needs to focus on following key aspects to improve the overall conversion rate:

- Increase user engagement on their website since this helps in higher conversion
- Increase on sending SMS notifications since this helps in higher conversion
- Get TotalVisits increased by advertising etc. since this helps in higher conversion
- Improve the Olark Chat service since this is affecting the conversion negatively