Lead Scoring Case Study

Logistic Regression

Group Members –
Bhargav Ram
Sandhyarani Sahoo

Problem Statement

- X Education sells online courses to industry professionals.
- Although X Education gets a lot of leads, its lead conversion rate is very poor. The typical lead conversion rate at X education is around 30%.
- To make this process more efficient, the company wishes to identify the most potential leads, also known as ‘Hot Leads’.
- If they successfully identify this set of leads, the lead conversion rate should go up as the sales team will now be focusing more on communicating with the potential leads rather than making calls to everyone.

Business Objective:

- X Education wants to identify the most promising leads.
- The company wants to build a model to identify the hot leads.
- Improve the lead conversion rate from 30% to 80%

Solution Methodology

- Read and understand Data
- Data Cleaning
  - Handle Missing values
  - Handle Skewed Categorical Columns
  - Handle Outliers

EDA
- Univariate Analysis of Categorical Variables with respect to Target Variable
- Bivariate Analysis on Numerical Columns

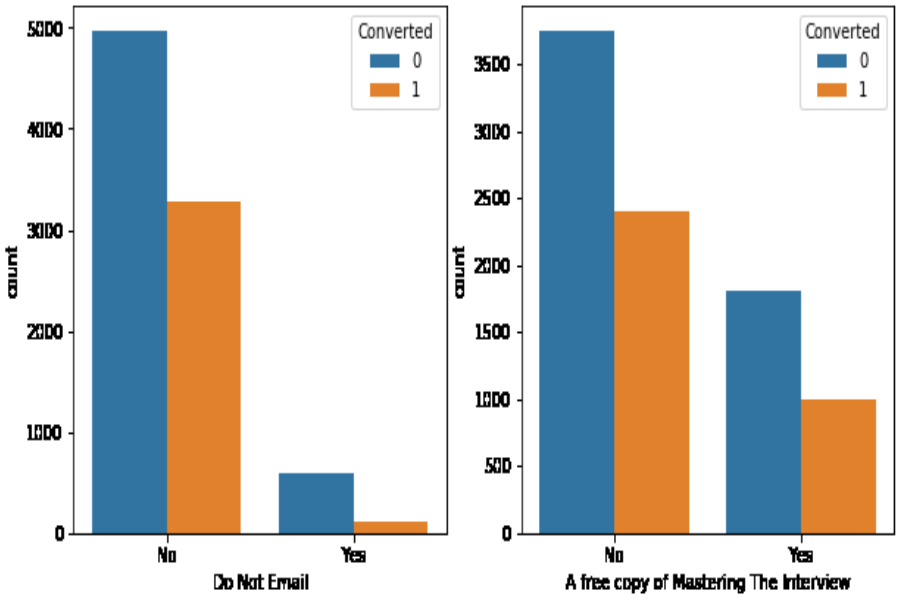
Data Preparation
- Dummy Variable Creation
- Splitting the Data Set into Train and Test Set
- Feature Scaling

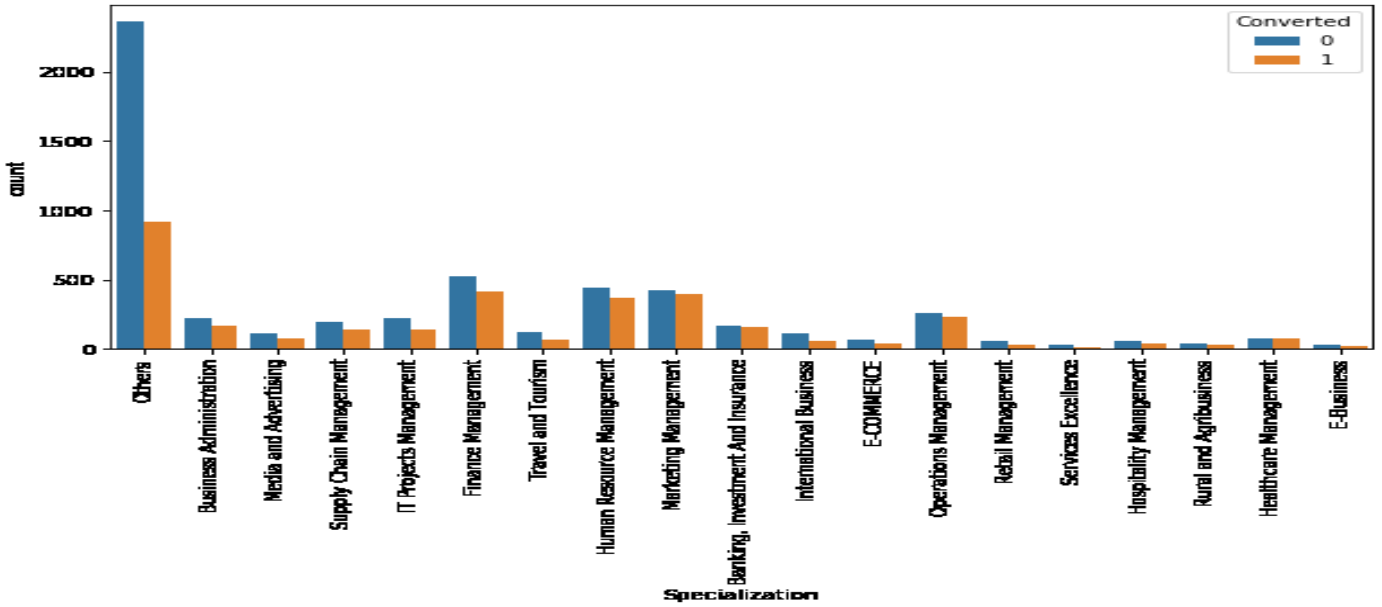
Modelling – Logistic Regression for Model Building and Prediction

Model Validation and Analysis of Different Metrics

Conclusions and Recommendation

EDA(1/5)

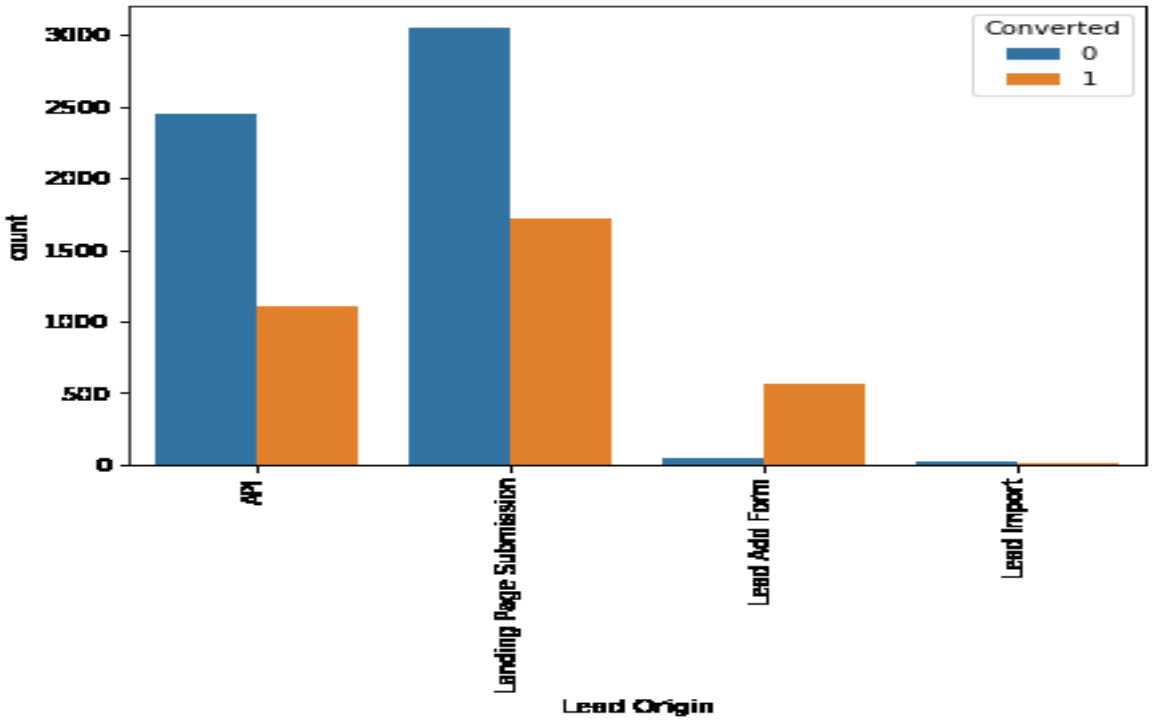


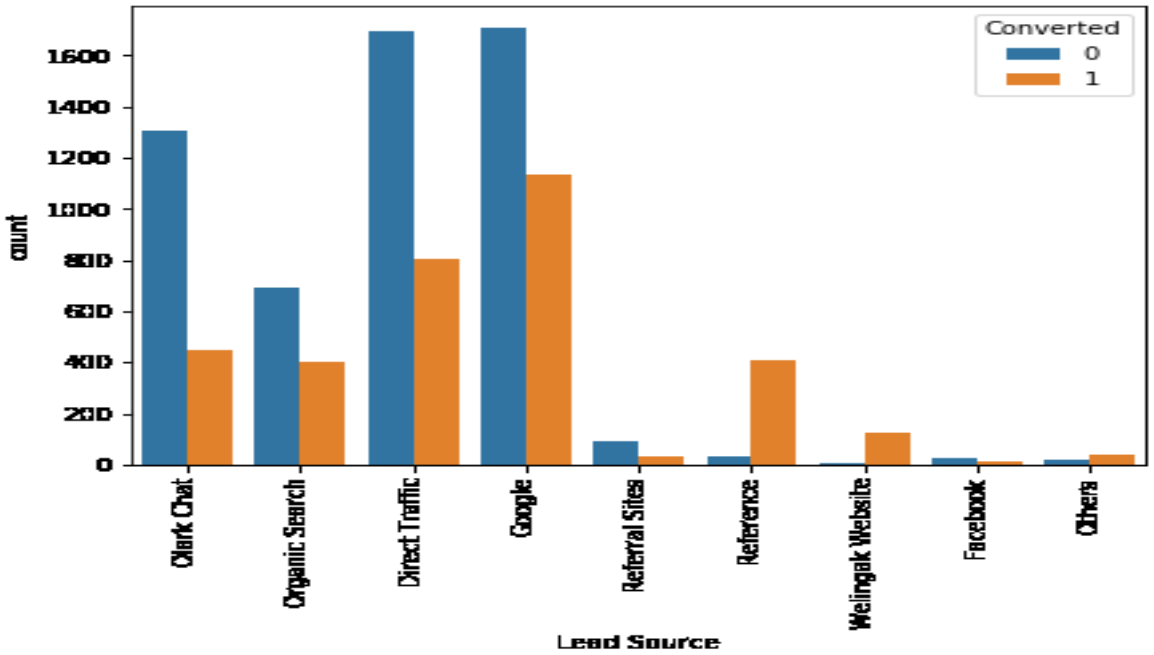


1.Most of the leads do not want to be emailed about the course and also do not want the free copy of mastering the interview.
2.Those leads who do not want to be emailed have high chances of getting converted

Inferences from 'Specialization' -
1.Highest number of leads as well as conversion fall in 'Others' category. However, this category is the bucket of 'Not Specified' Values.
2.Finance Management, Human Resource Management, Marketing Management, Operations Management are showing reasonably good results in terms of count of leads as well as conversion.

EDA(2/5)

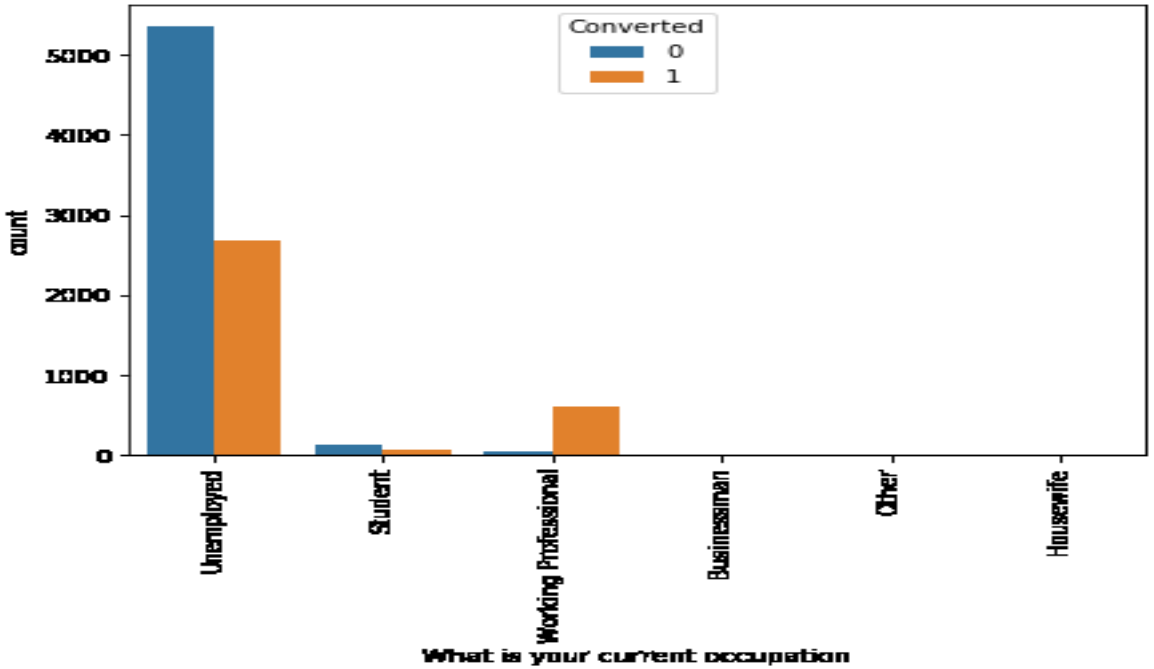


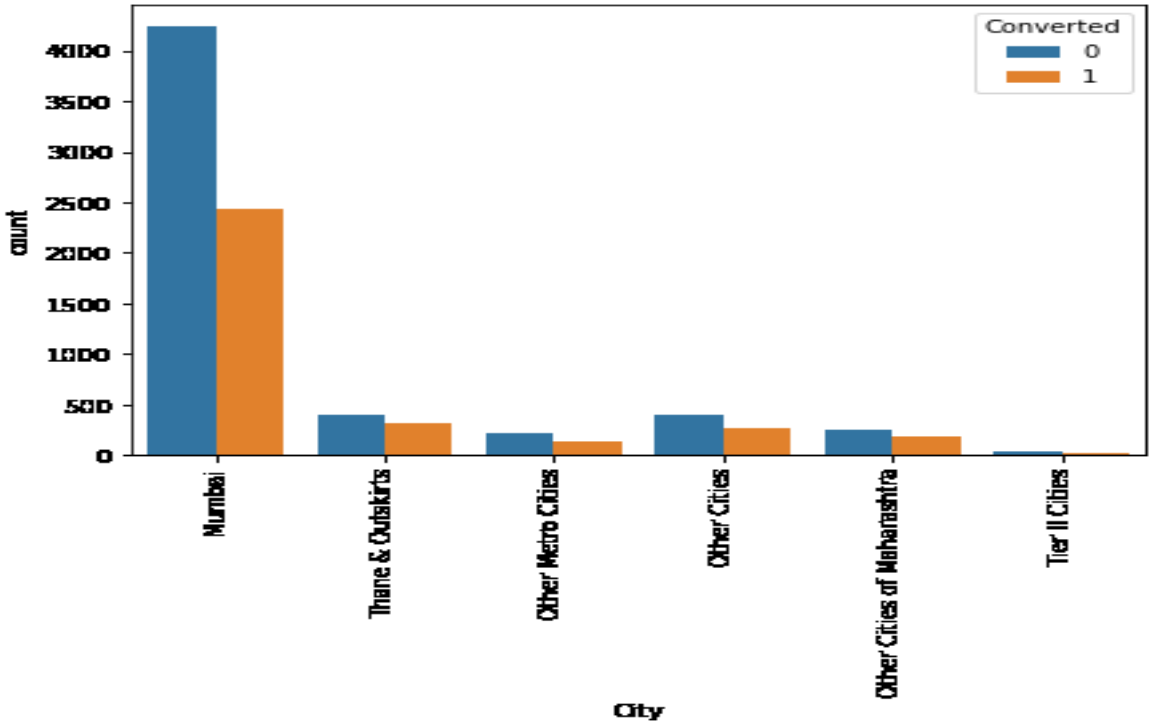


Inferences from Lead Origin -
1.Lead Add Form has a very high conversion rate but count of leads are not very high.
2.API and Landing Page Submission bring higher number of leads as well as conversion.
3.In order to improve overall lead conversion rate, we need to improve lead conversion of API and Landing Page Submission origin and generate more leads from Lead Add Form.

Inferences from Lead Source -
1.Reference and Welingak Website have very high conversion rate but count of leads are not very high.
2.Google, Direct Traffic and Olark Chat bring higher number of leads as well as conversion.

EDA(3/5)

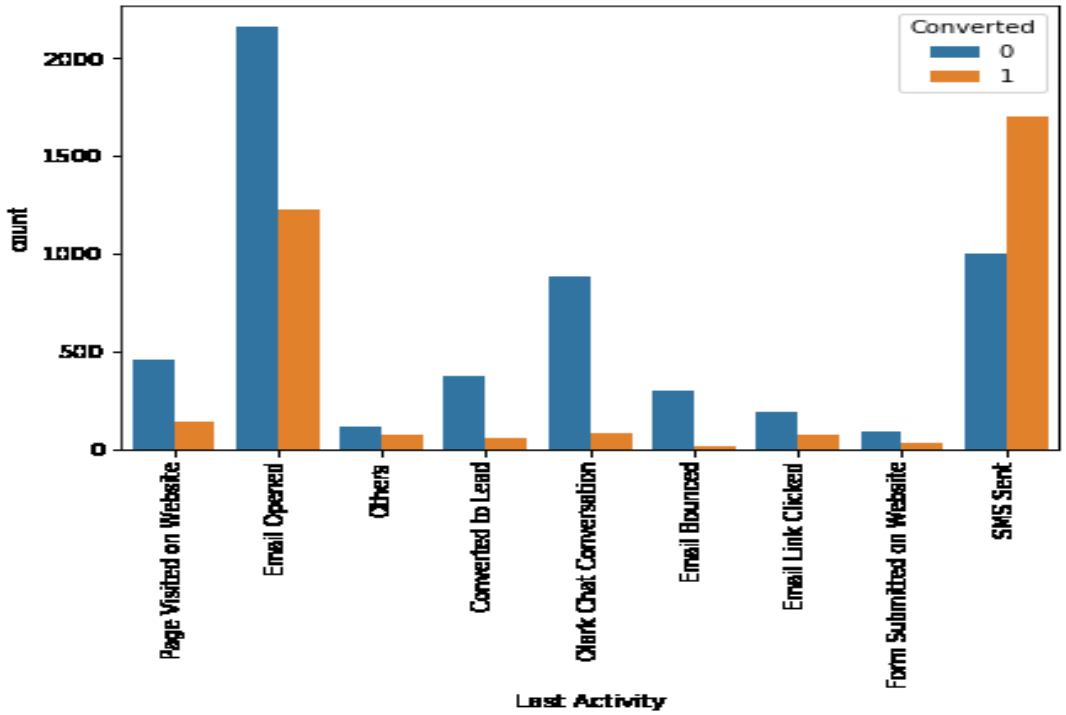


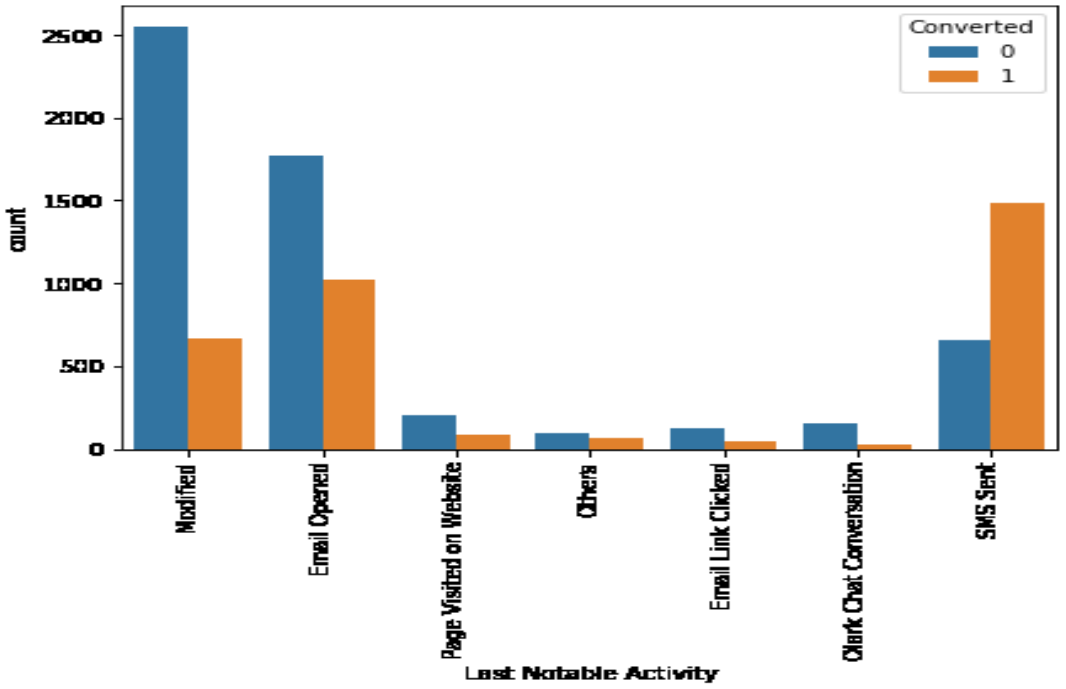


Inferences from 'What is your current occupation' -
1.Working professional checking about the course have high chances of joining the course.
2.Higher number of leads as well as conversion from Unemployed category.

Inferences from 'City' -
1.Mumbai has the highest number of leads as well as conversion.

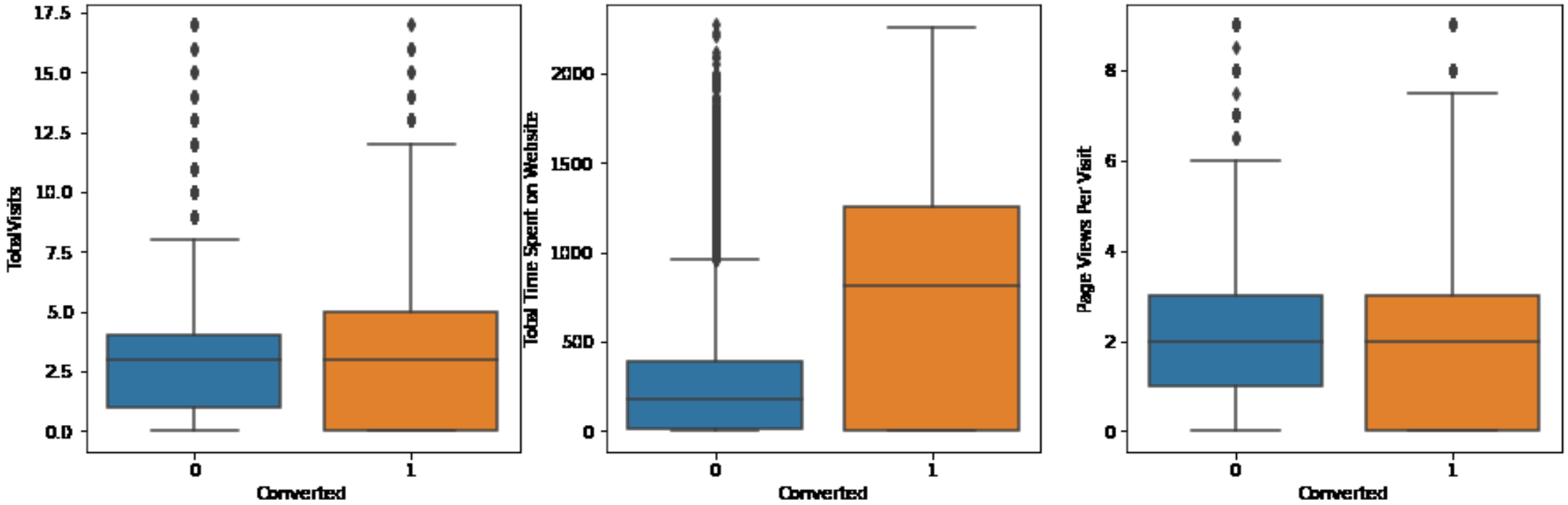
EDA(4/5)





Inferences from 'Last Activity' -
1.Although the count is high for 'Email Opened', but the highest conversion rate from 'SMS Sent' Category.
Inferences from 'Last Activity' -
1.High Conversion rate is for 'Email Opened' and 'SMS Sent' Category.
2.Lead count is highest for 'Modified' and 'Email Opened' category.

EDA(5/5)



1.Nothing conclusive can be said on the basis of Total Visits and Page Views Per Visit.
2.Leads spending more time on the website are more likely to be converted.

Data Preparation

Dummy variables are created for categorical variables.

For performing train-test split, we have chosen 70:30 ratio.

After cleaning and dummy variable creation, total number of rows and columns are 8953 and 59 respectively.

Modelling

Used Logistic Regression Technique

First used RFE method to come up with 15 important variables

Then used manual method for dropping variables to make the model significant by looking at p-value and VIF.

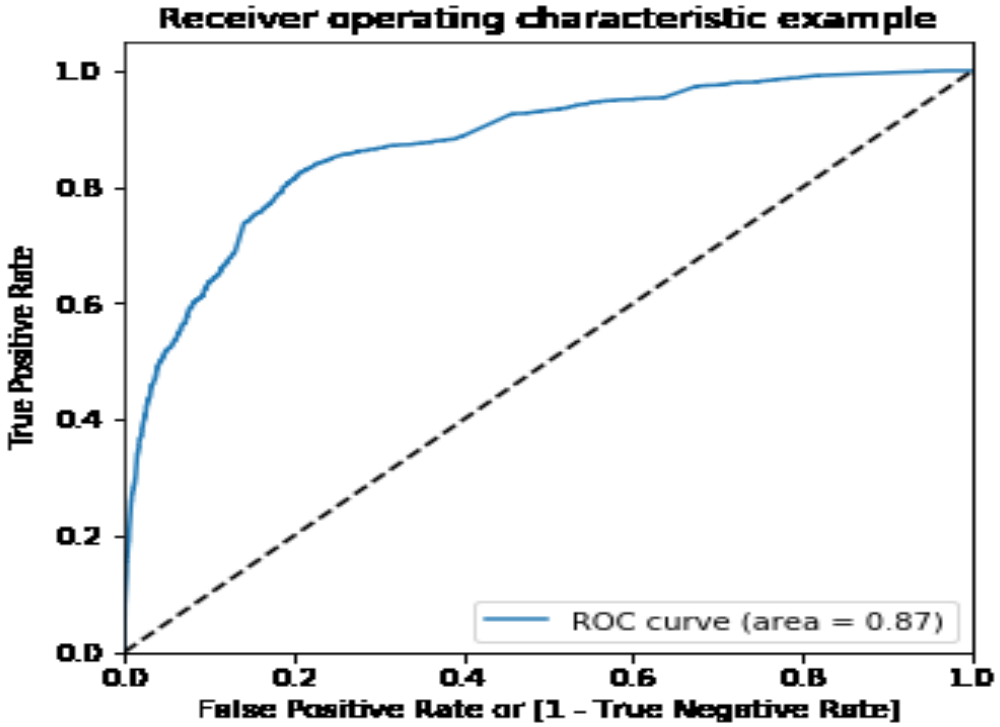
In our final model, all VIFs are low and p-values are below 0.05.

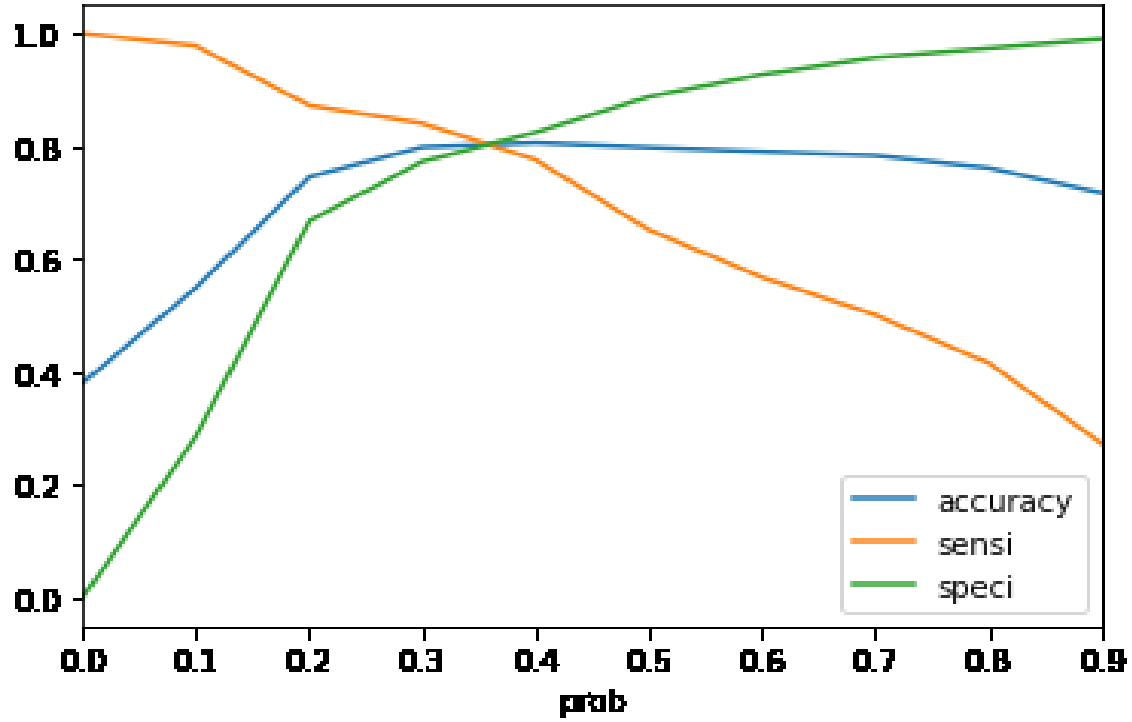
Model Evaluation is done using accuracy, sensitivity and specificity.

Used ROC to check the trade off between True Positive Rate and False Positive Rate.

Found the Optimal Cut-off point to improve sensitivity.

ROC and Optimal Cut-off Point





- Area under ROC is 0.87, which is a very good value for a model.
- From the curve above, 0.35 is the optimum point to take it as a cutoff probability.

Conclusion

	Train Set	Test Set
Accuracy	80.53%	80.15%
Sensitivity	81.38%	81.09%
Specificity	80.01%	79.59%

The variables that mattered most in the lead conversion are -

- •Lead Origin_Lead Add Form
- •What is your current occupation_Working Professional
- •Total Time Spent on Website
- •Lead Source_Organic Search
- •Lead Source_Direct Traffic
- •Last Activity_Olark Chat Conversation
- •Last Notable Activity_Email Opened
- •Do Not Email
- •Last Notable Activity_Page Visited on Website
- •Last Notable Activity_Email Link Clicked
- •Last Notable Activity_Modified
- •Last Notable Activity_Olark Chat Conversation

The ability of accurately predicting the conversion rate of any lead from the model is around 80%.

The probability of predicting a promising lead from the model is around 81%.

So we can go ahead with the deployment of this model

Recommendation

X Education must focus on the people if –

- The lead origin is through Lead Add Form
- The prospect is a working professional
- The prospect spends a lot of time on the website

As these parameters have positive coefficients, this will improve the lead score and in turn, these people will be ‘hot leads’ for X Education

THANK YOU