Problem Statement

- X Education an online course providing organisation wants to increase their lead conversion ratio.

- X education want to correctly predict a lead as a 'Hot lead' which has a higher percentage of getting converted.

- X Education has set a target of increasing their lead conversion ratio from 30 % to 80%.

Process Followed

Data Cleaning
- Data Sourcing
- Removing and imputing null values & Handling Outliers
- Removing biased columns

Exploratory Data Analysis
- Count plots to understand data distribution in both cases
- Boxplot to visualize Outliers.
- Heatmap for visualizing correlation of numerical variables

Data Preparation
- Encoding Binary Variables
- Creating Dummy Variables
- Splitting Data into Train Test Split
- Scaling the data

Model Preparation
- RFE to extract top features
- Evaluating Models and refining models basis p values and VIF.
- Evaluating model F1 score, accuracy, sensitivity, recall, precision scores.
- Comparing the training and testing scores

Exploratory Data Analysis (1/4)
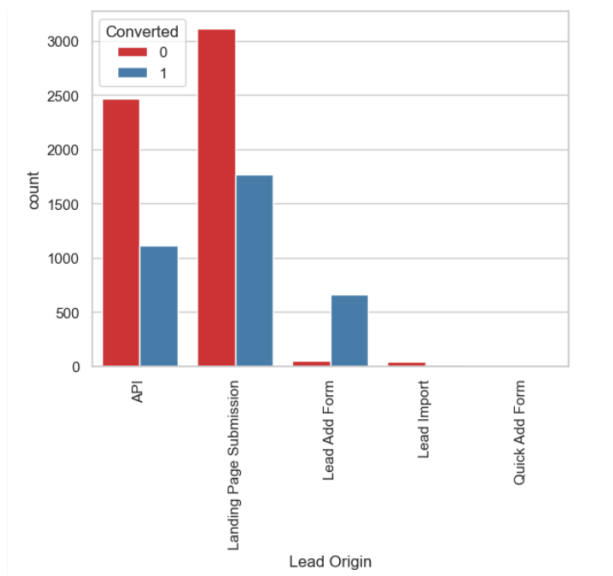
1.API and Landing Page Submission have 30-35% conversion rate but count of lead originated from them are considerable.
2.Lead Add Form has more than 90% conversion rate but count of lead are not very high.
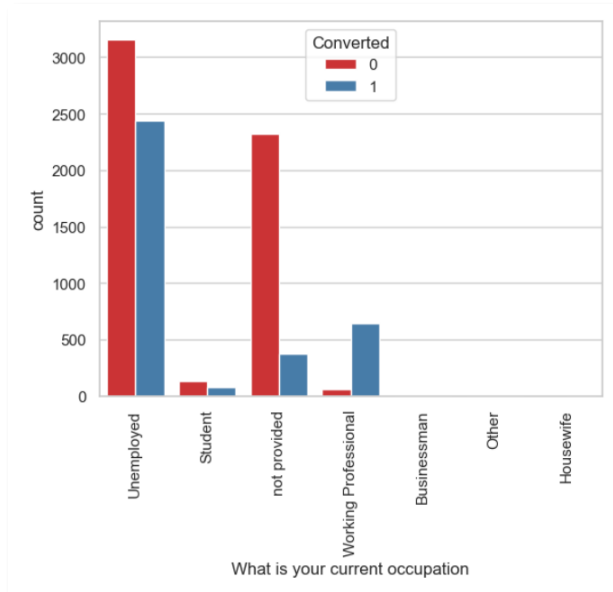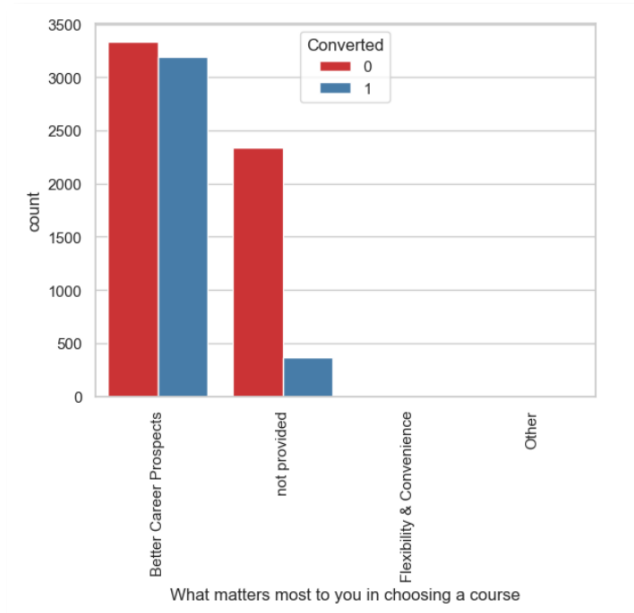

1. Unemployed leads are the most in numbers but has around 30-35% conversion rate.
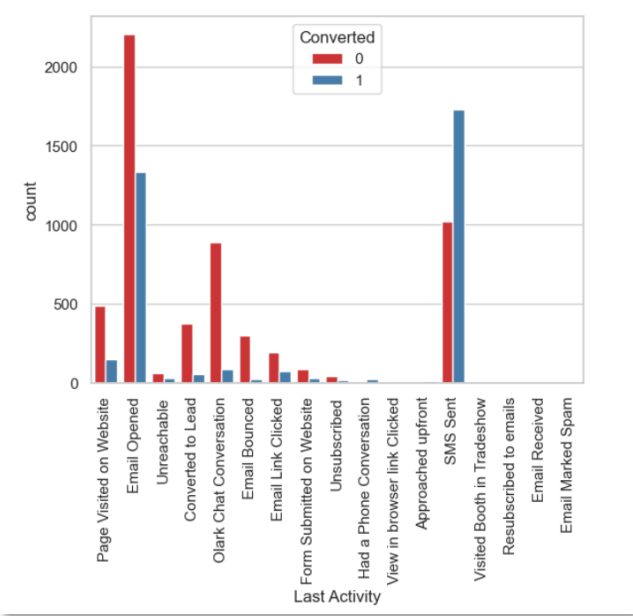
Exploratory Data Analysis (2/4)
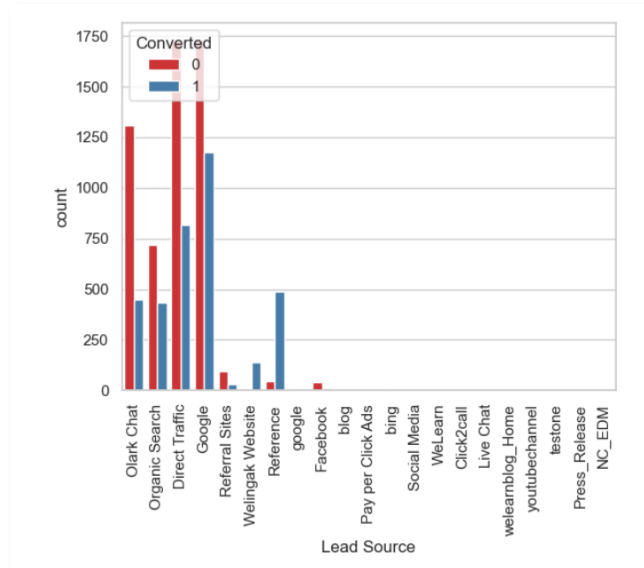
1. Google and Direct traffic generates maximum number of leads.
2. Conversion Rate of reference leads and leads through welingak website is high.

1. Leads joining course for Better Career Prospects have higher conversion rate
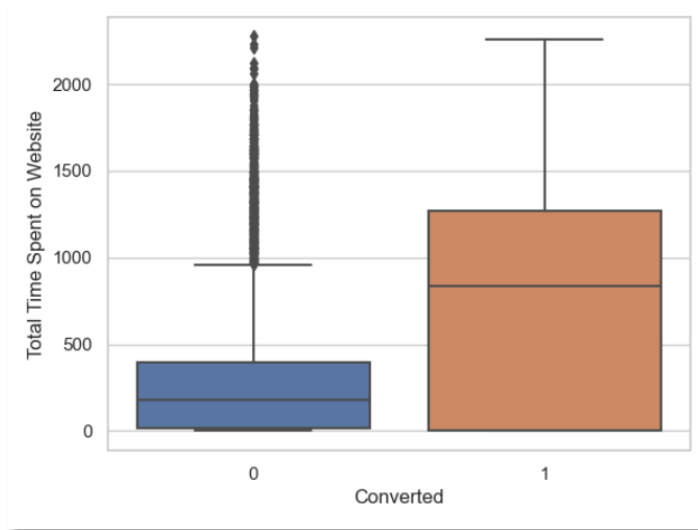
Exploratory Data Analysis (3/4)

1. Most of the lead have their Email opened as their last activity. Conversion rate for leads with last activity as SMS Sent is almost 60%.
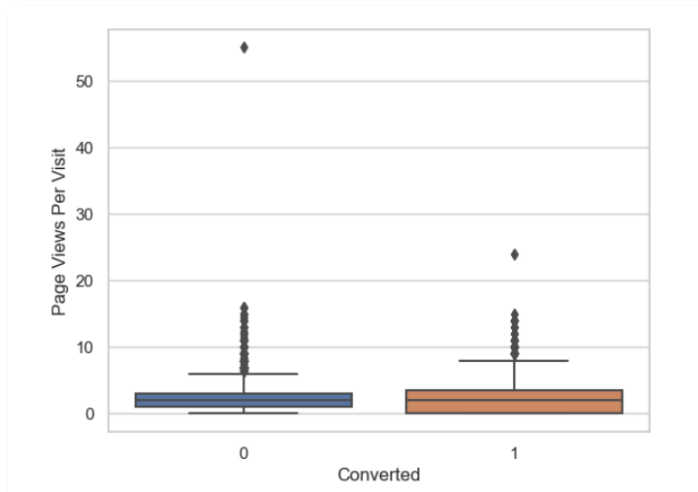

1. 'TotalVisits' and 'Page Views per Visit' are highly correlated
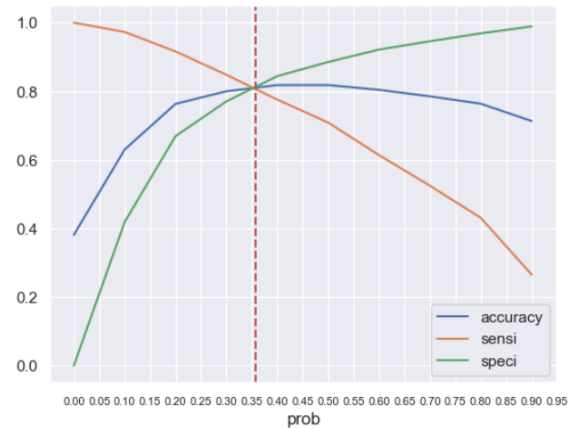
Exploratory Data Analysis (4/4)

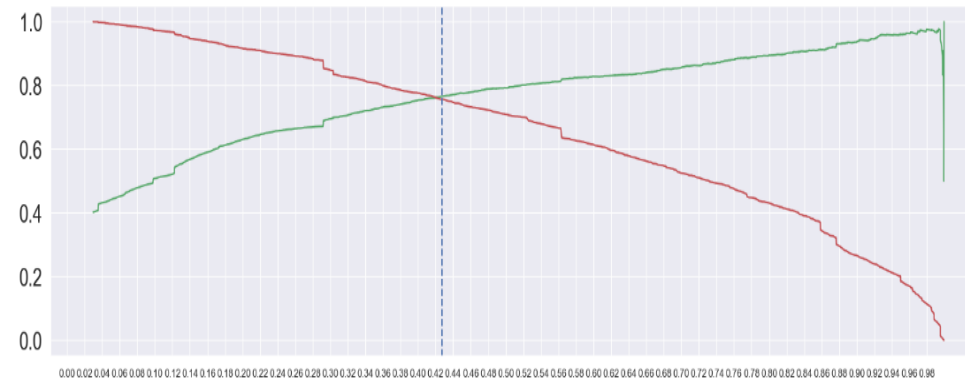1. Leads spending more time on the website are more likely to be converted.

1. Median for converted and unconverted leads is the same for 'Page Views per visit'
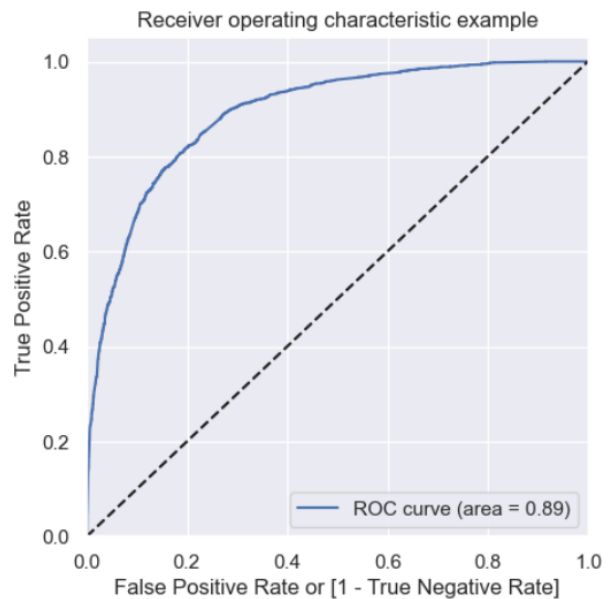
Model Evaluation

Accuracy-Sensitivity-Specificity Curve

Precision Recall Curve

ROC Curve

Model Evaluation Metrics

- Accuracy : 81.1%
- Sensitivity : 80.1%
- Specificity : 81.8%
- Pricision: 73.1%
- Recall: 80.1%

Conclusion

- The top most affecting variable/fields are
    - Lead Origin - Landing Page Submission, Lead Add Form
    - Lead Source - Olark Chat, Welingak Website
    - Last Activity - Email Opened, Olark Chat Conversation, Other_Activity, SMS Sent, Unsubscribed
    - Specialization – Hospitality Management
    - What is your current occupation – Working Professional
    - Do Not Email
    - Total Time Spent on Website
- The accuracy for training and test data are 81.1% and 80.9% respectively.