Lead Scoring Case Study

Logistic Regression Assignment

By:

Rakesh Verma

Renu Vind

Sabita Rana

Content

- Problem Statement
- Business Objective
- Solution Approach
- Data Cleaning and Visualization
- EDA: Univariant Analysis
- EDA: Bivariant Analysis
- Feature Selection
- Data Preparation & Modelling
- Finding the Cut-off
- Model Evaluation
- Final Equation
- Observations
- Recommendation

Problem Statement:

- X Education sells online courses to industry professionals.
- X Education gets a lot of leads, its lead conversion rate is low. For example, if, say, they acquire 100 leads in a day, only about 30 of them are converted.
- 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

- To identify most promising leads i.e. the leads that are most likely to convert into paying customers.
- To build a wherein a score can be assigned to each of the leads such that the customers with a higher lead score have a higher conversion chance and the customers with a lower lead score have a lower conversion chance.
- This shall help in improving target lead conversion rate to be around 80%.
- The model can be used further in future.

Solution Approach

Here Logistic Regression has been used to create Model.

Logistic regression is a supervised machine learning algorithm used for classification i.e. to predict the probability that an instance belongs to a given class or not. It helps to make predictions for categorical variables.

To validate the model Confusion matrix and ROC curve has been used.

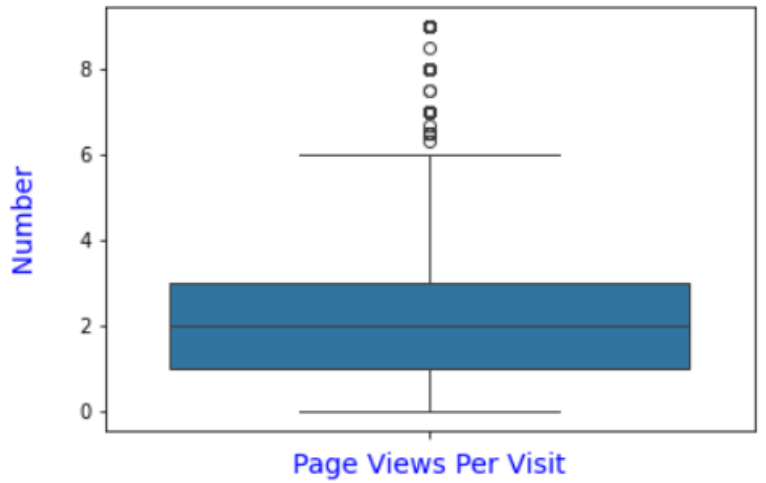
Data Cleaning and Visualization

Given Dataset Info

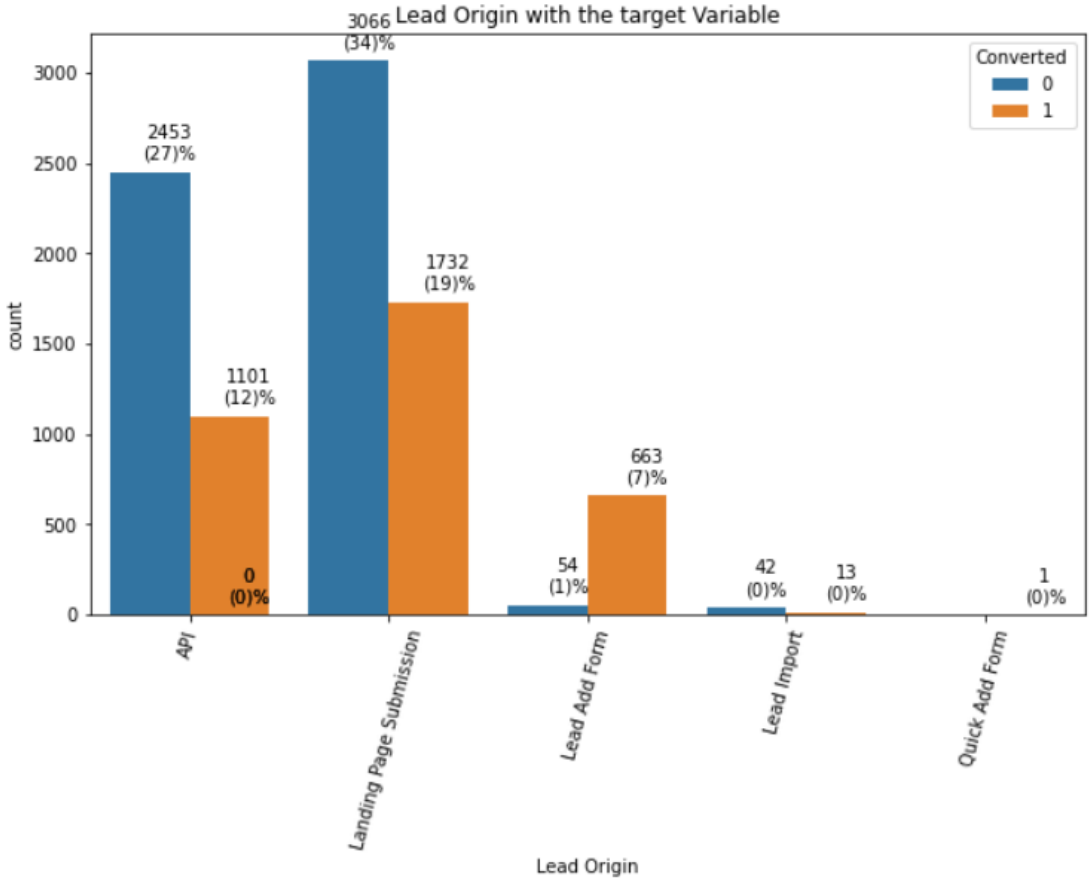
1. Shape of the given dataset: (9240, 37)
2. There are 7 Numerical Columns and 30 categorical columns
3. Around 17 columns have NaN values present in them.
4. Numerical columns such as 'TotalVisits', 'Total Time spent on Website', and 'Page Views Per Visit' have outliers present in them

Below steps take to Clean and Prepare the data.

1.Removed Columns with 40% of missing values
2.Removed columns having no variance in the data(presence of same data)
3.Missing value treatment
4.Data Imputation
5.Outliers Treatment.

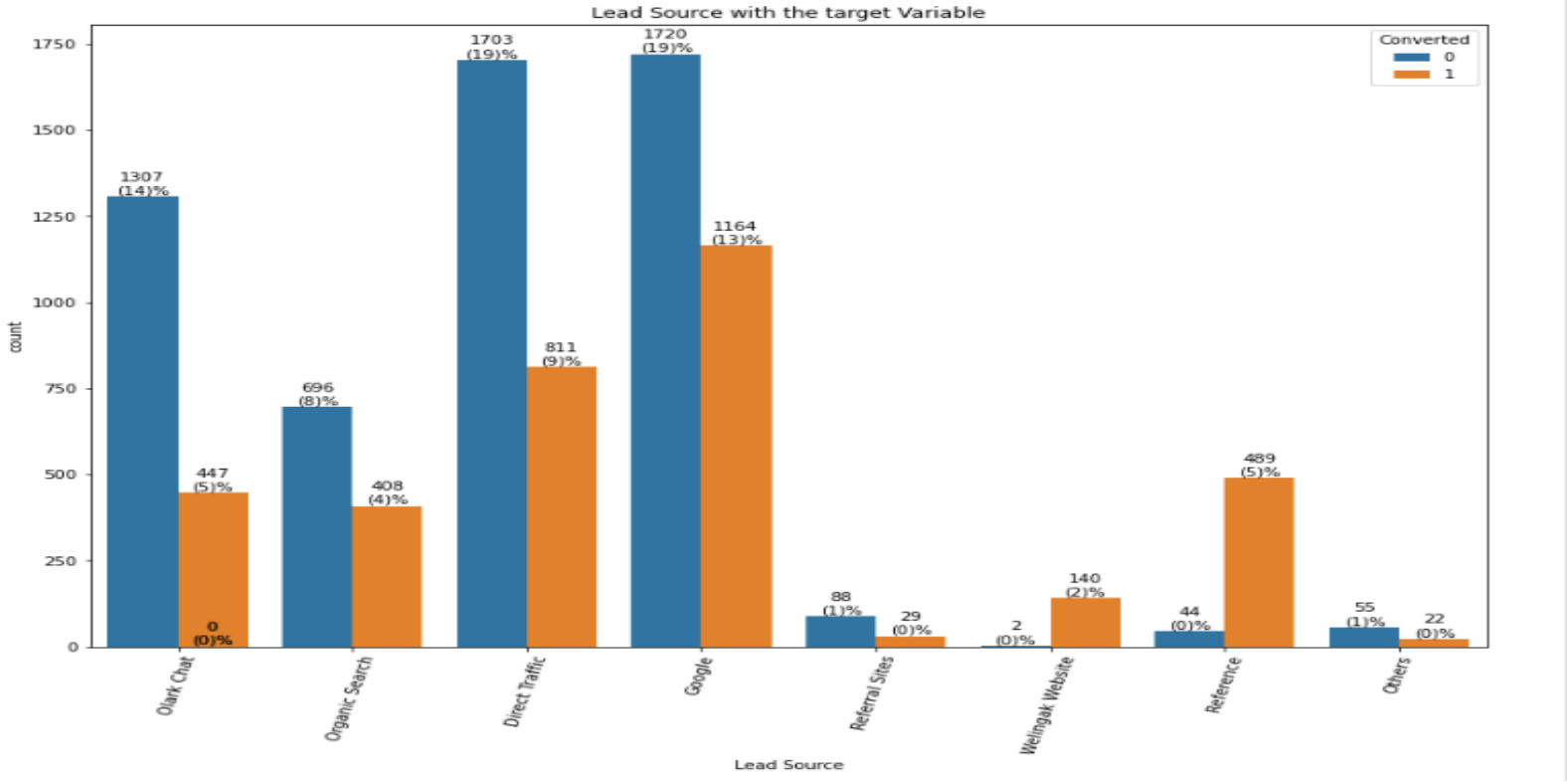


Univariate Data Analysis
(Lead Origin)



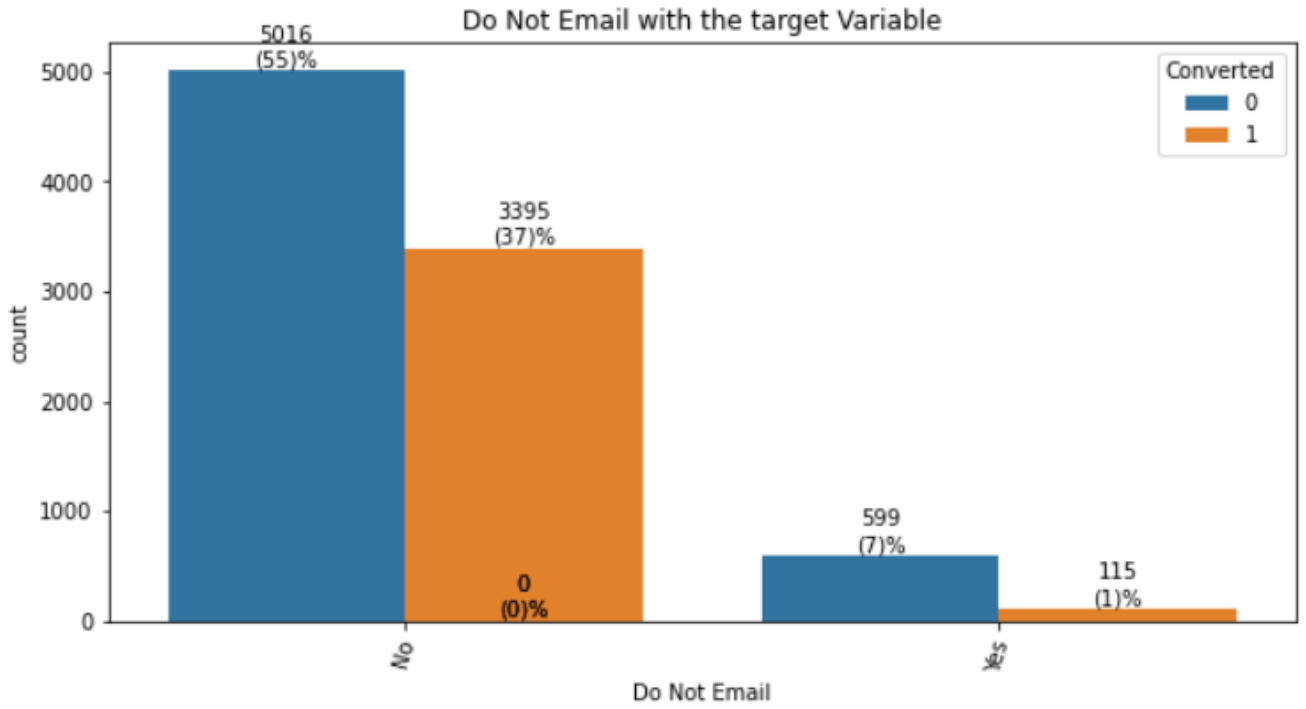
- Landing Page Submission and API have high number of Leads as compared to other Origins
- Conversion of Lead is also higher for landing Page submission (~36%) than API (~31%)
- In order to improve the overall conversion rate, X education should focus on increasing the conversion of API and Landing Page submission.
- They should also increase Leads form Lead Ads Form as they have the highest conversion among all

Univariant Analysis
(Lead Source)



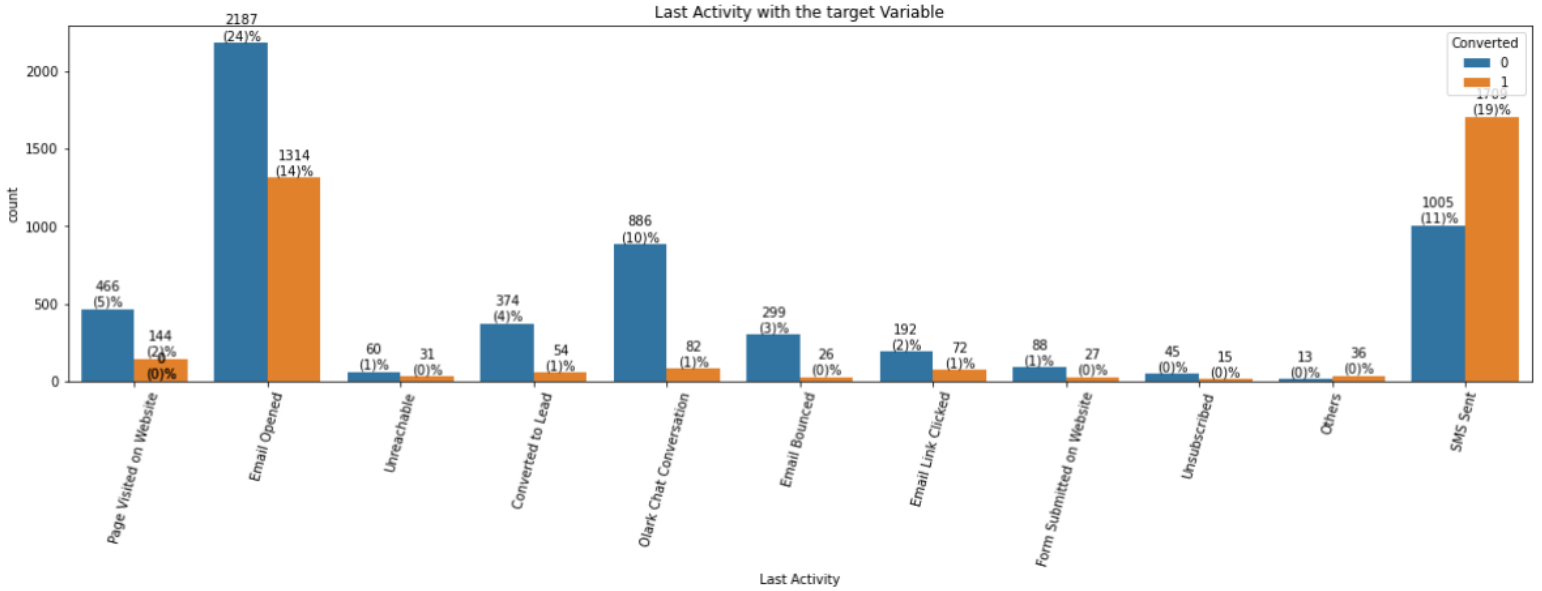
- Google and Direct Traffice are generating high number of leads followed by Olark Chat
- The conversion rate is highest for Reference approx. 92%, than for google (~40%) and Organic Search (~37)
- In order to improve the overall conversion rate, X education should focus on increasing the Lead conversion through Google, Organic Search, Direct Traffic.
- They should increase more Leads from References and Welingak WebSites

Univariant Analysis
(Do Not Email)



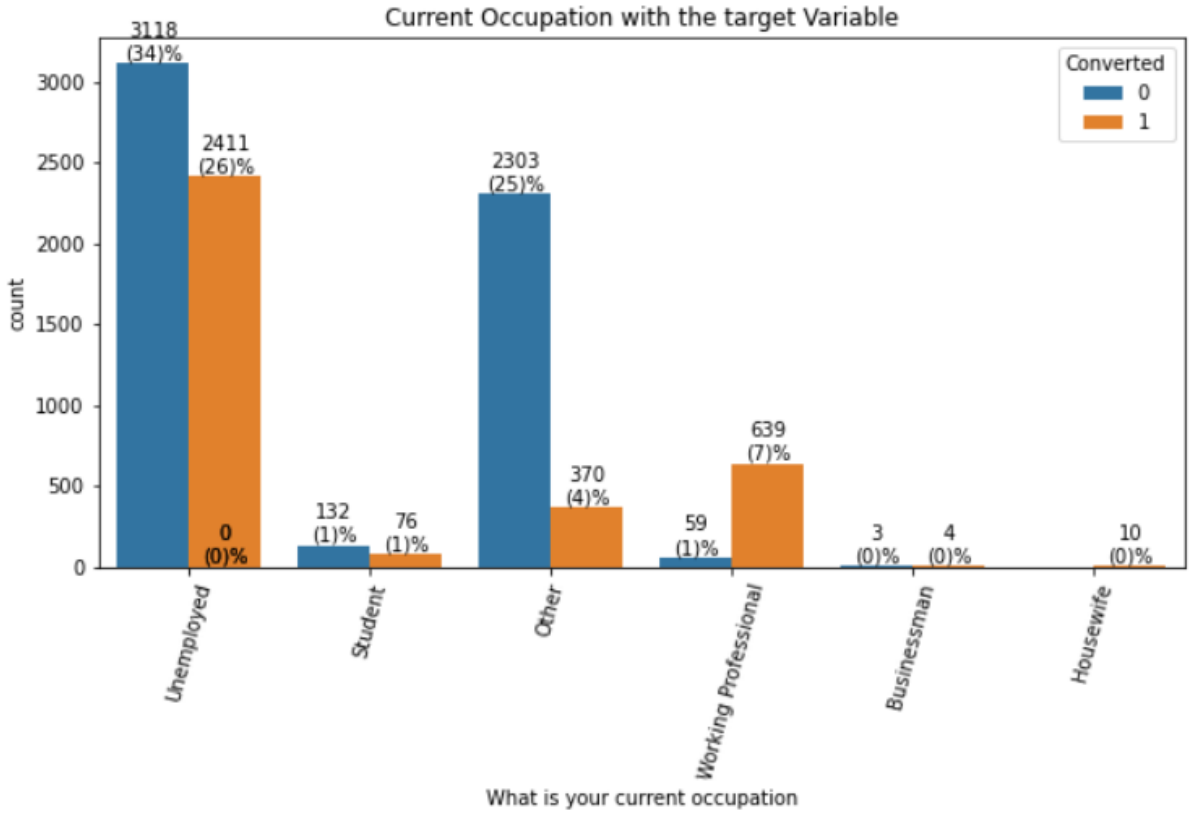
- Conversion Rate of Leads who opted for Email sharing 40
- Majority of Leads have asked to share the course details over email
- Out of these, around 40% of the Leads got converted
- Steps to be taken to share more informative content over email so that conversion rate will increase

Univariant Analysis
(Last Activity)



- Email and SMS are the mail source which Leads uses on daily basis, and these should be targeted to converts the leads

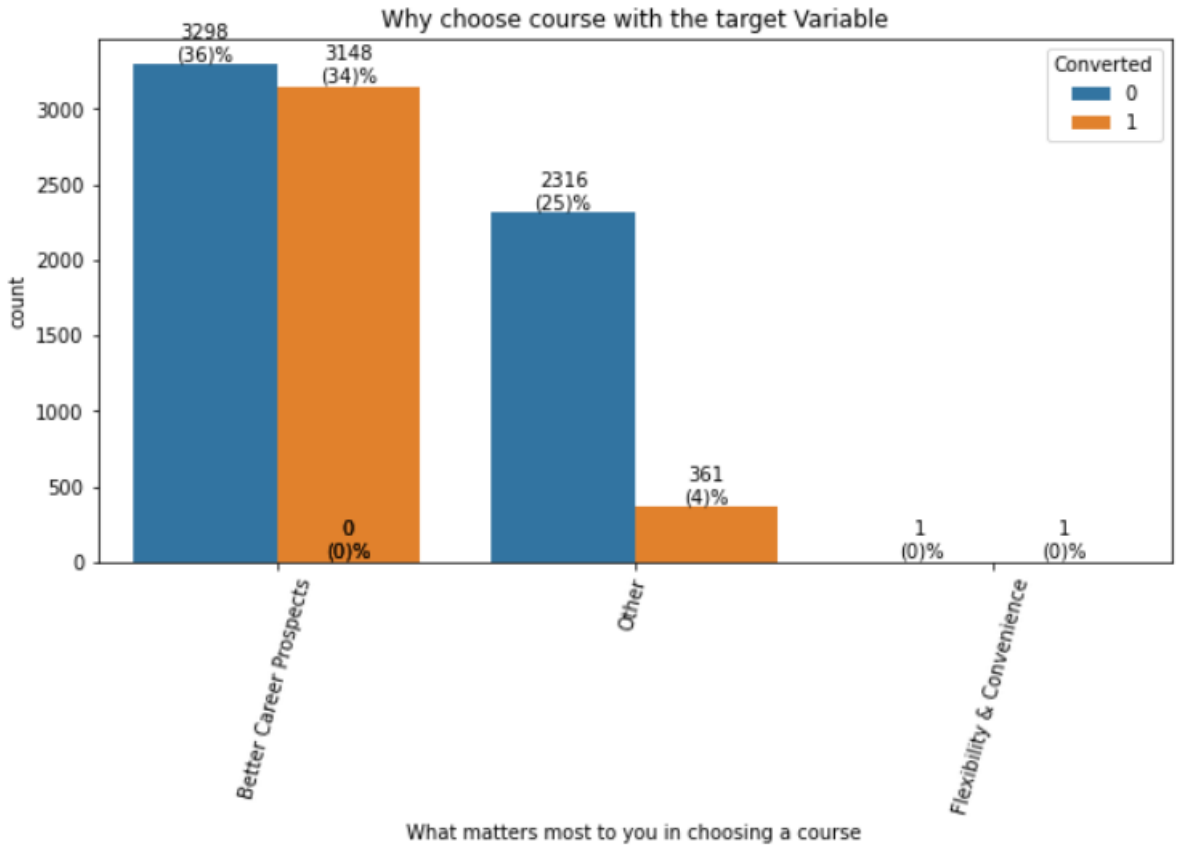
Univariant Analysis
(What is your current occupation)



- Conversion Rate of Leads are unemployed 44
- Observation
- Unemployed users are the majority of the Leads but only 44% are converted.
- We should focus on converting these set of users and increase out conversion percent

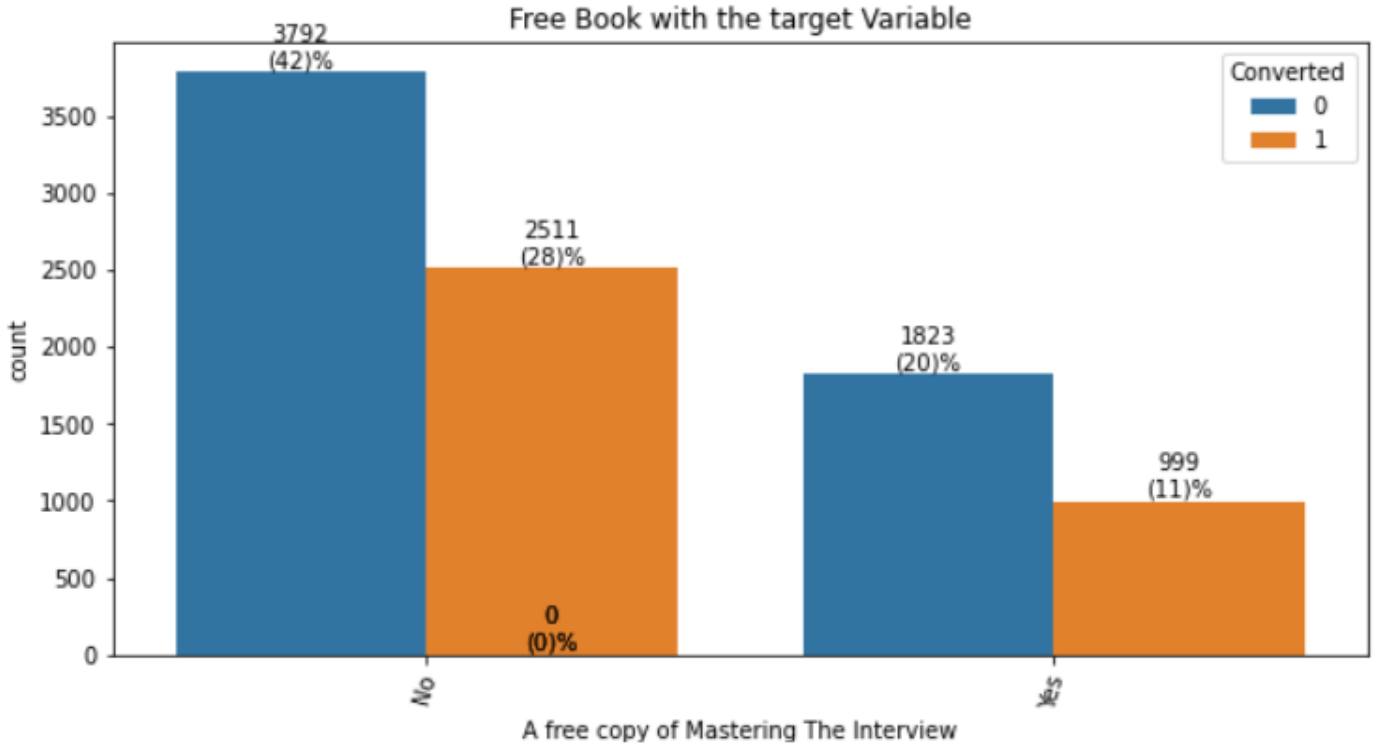
Univariant Analysis

(What matters most to you in the choosing a course)



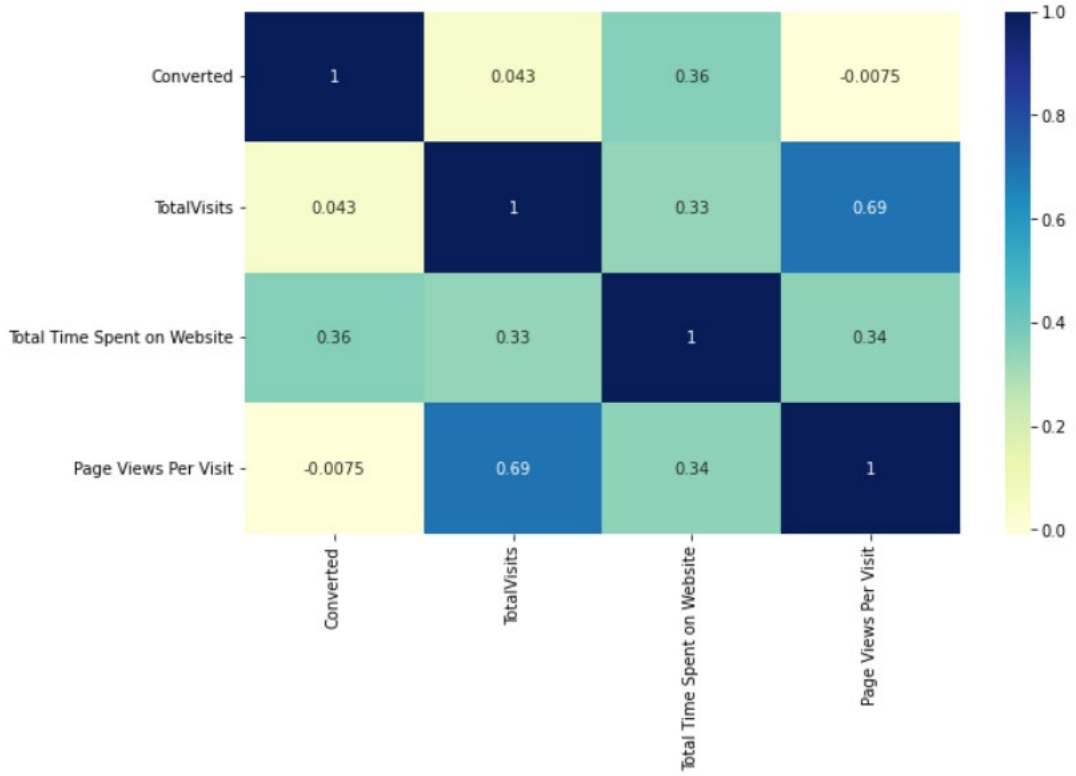
- Conversion Rate of Leads who want Better Career Prospects 49
- More than 70% of the Leads want to opt for course for Better Career Prospects but only 49% are converted

Univariant Analysis
(A free copy of mastering the interview)



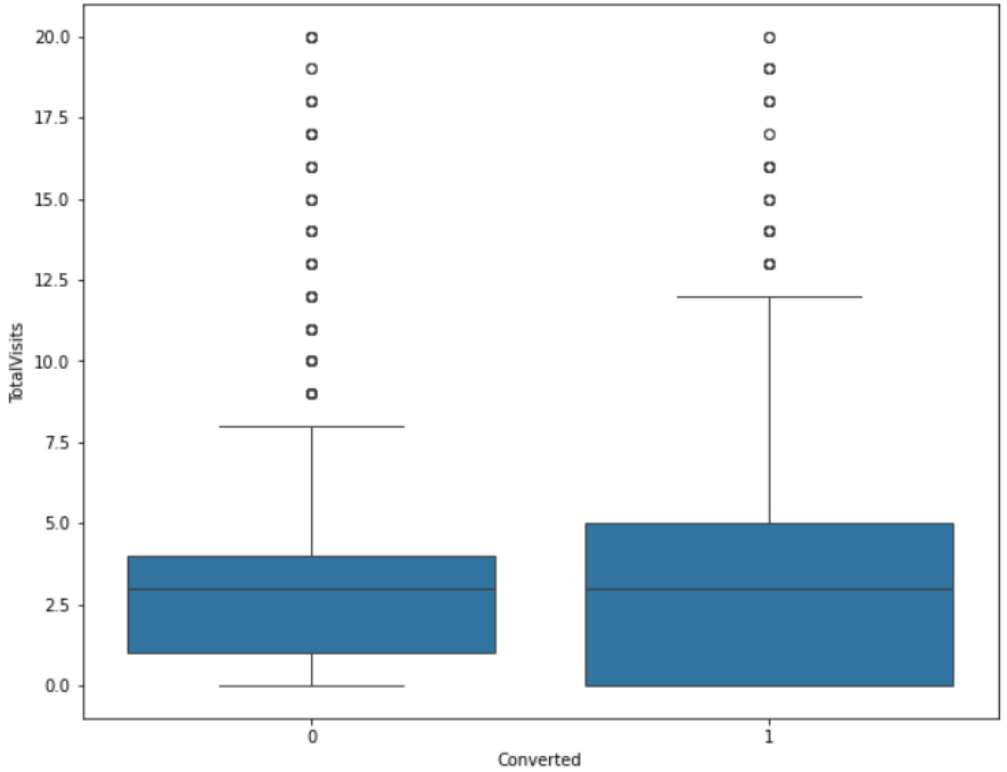
- Looking at the data, most of the Leads do not opt for Free Books.
- And for those who opted for free books, only 50% get converted.

Bivariant Analysis
(Correlation Matrix)



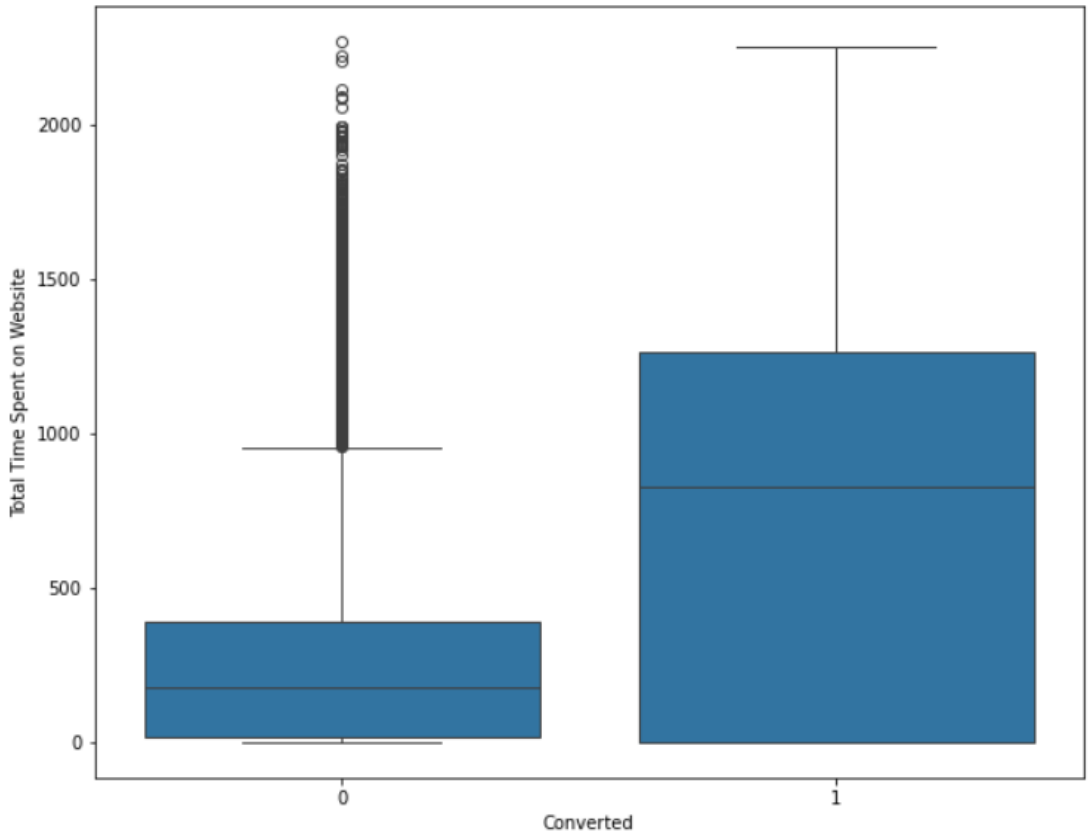
- There is high correlation of around 70% in Total Vists and Pages Views per visit
- Looking at the data, we can see the total visit and converted have less small correlation whereas have moderate correlation with 'Total time spent on the Website'.

Bivariant Analysis
(TotalVisits)



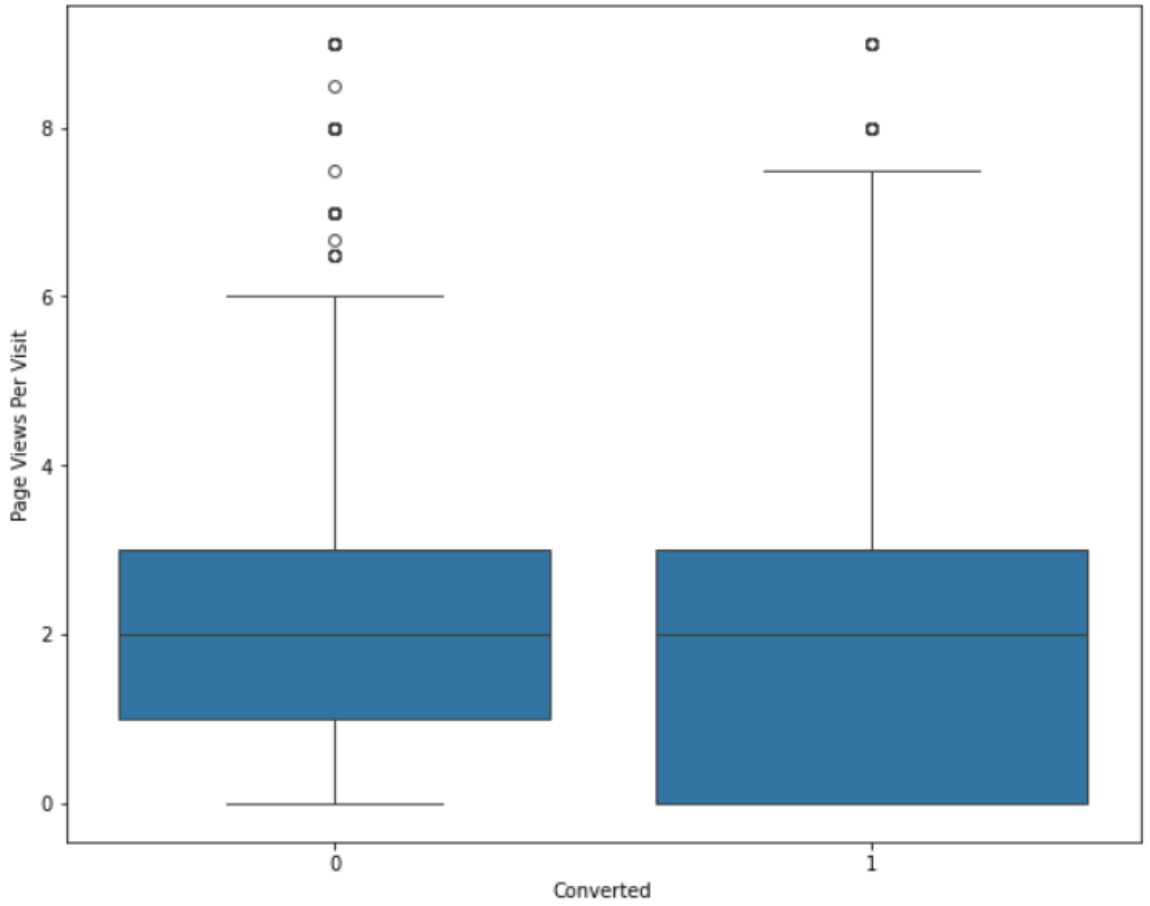
- Leads that are converted have more number of visits, hence company should focus on Leads who are visiting the site more than other

Bivariant Analysis
(Total Time spent on website)



- Leads that are converted have spent more time on the Website than others, hence company should focus on such Leads as they have high chance of converting

Bivariant Analysis
(Page Views Per Visit)



- Pages Viewd per visit does not show any remarkable difference between who are converted to those who are not converted

Feature Selection

After analyzing the correlation between dependent variables we found, there are few variables which are highly correlated.

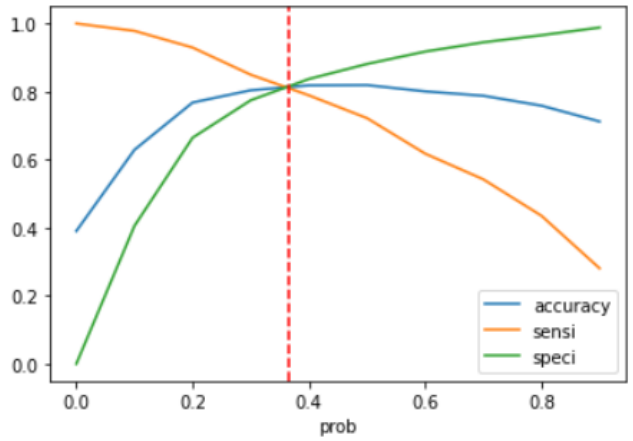
- TotalVisits and Page Views per visits have high correlation of 0.70
- Do Not Email and Last Activity - Email Bounced are highly correlated 0.64
- Lead Origin - Landing Page Submission and Page Views per Visits
- Lead Source - Reference and Lead Origin - Lead Add Form

Hence we prefer to use RFE to select best features for feature selection.

Data Preparation and Modelling

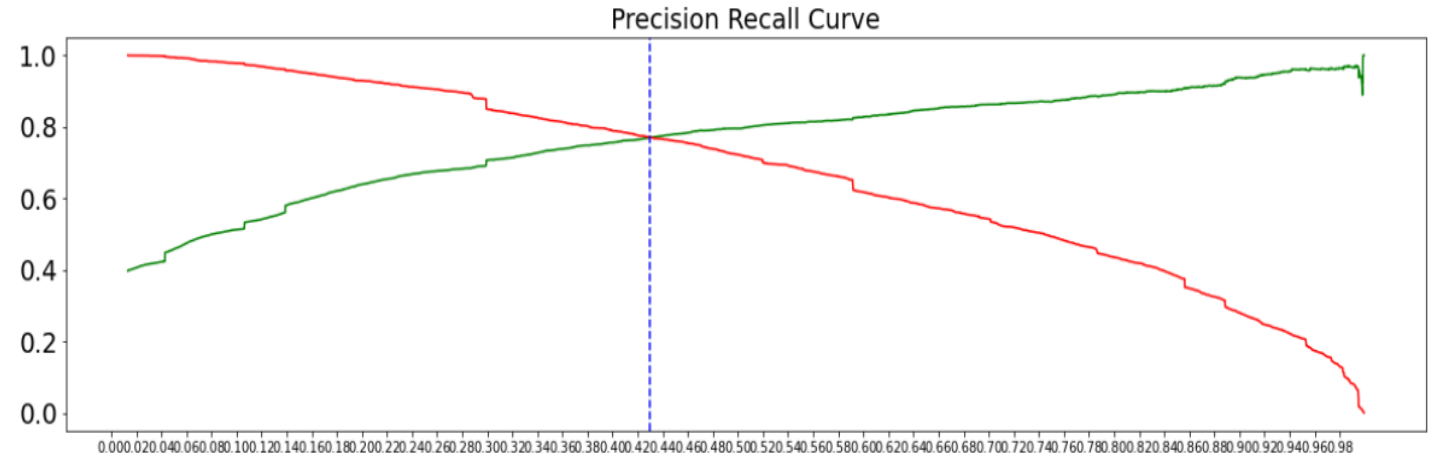
- Creating Dummy Variables
- Splitting the data into Train and Test datasets
- Perform Feature Scaling
- Creating X and y variable for both train and test data
- Feature selection using RFE
- Model Building
- Eliminating features with based on p-value and VIF
- Selecting Final model for evaluation
- Predicting the y-train values
- Finding Option Cut-off

Finding the Cut-off



The cutoff value from the above graph looks like 0.365

- We found True Positive number has decrease and True Negative number has increase using Precision-Recall trade-off method.
- Thus, we cannot use Precision-Recall trade-off method as it reduced True Positive so 'Recall'/'sensitivity' decreased for this point. We have to increase Sensitivity Recall value to increase True Positives. Thus we will use **0.365** as optimal cutoff point.

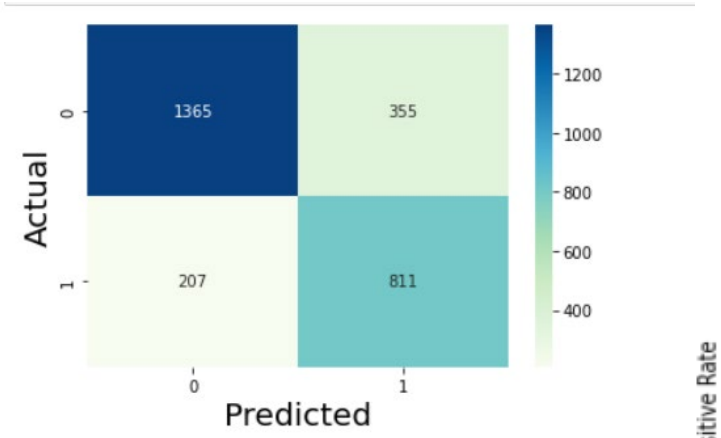


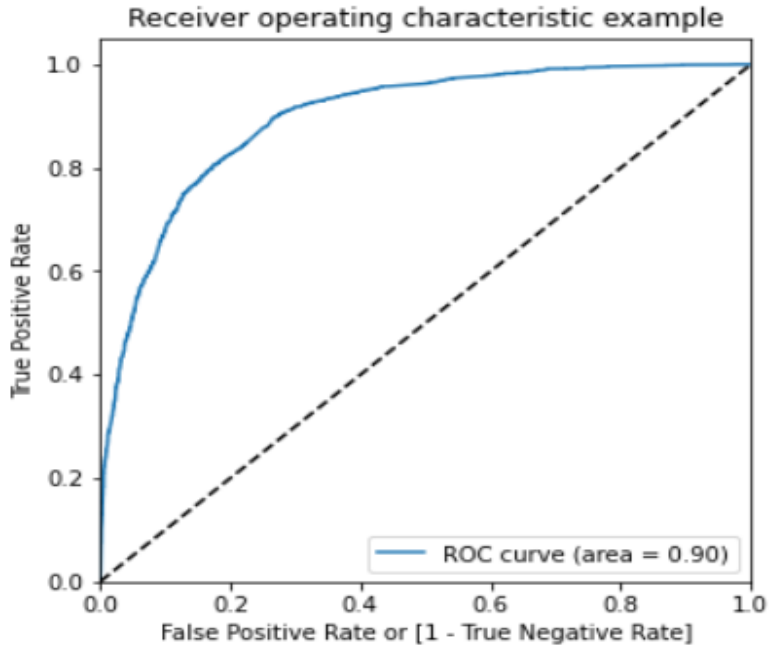
Observation

1. From above above 'precision_recall_curve' we can see that cutoff point is 0.430

Model Evaluation

Model Evaluation using Confusion Matrix and ROC Curve





Observation

1. We got ROC value of 0.90, which is a good value

Final Equation

Converted = -2.175775 - (1.199395 * Do Not Email) + (1.159936 * TotalVisits) + (4.396581 * Total Time Spent on Website) - (1.000800 * Lead Origin_Landing Page Submission) + (2.987703 * Lead Origin_Lead Add Form) + (1.281876 * Lead Source_Olark Chat) + (3.170651 * Lead Source_Welingak Website) + (0.972868 * Last Activity_Email Opened) + (1.788474 * Last Activity_Others) + (2.194420 * Last Activity_SMS Sent)+ (1.060620 * Last Activity_Unreachable) + (1.801875 * Last Activity_Unsubscribed) - (0.930426 * Specialization_Others) + (2.295716 * What is your current occupation_Working Professional) - (1.278123 * What matters most to you in choosing a course_Other)

Observation

The sensitivity value for test data is 80% while for train data is also 81% . The accuracy values is 79%. Which shows that model is performing well for test data set also.

Evaluation Metrics for the train Dataset:-

- Accuracy : 82%
- Sensitivity: 81%
- Specificity: 82%
- Precision: 74%
- Recall: 81%

Evaluation Metrics for the test Dataset:-

- Accuracy : 79%
- Sensitivity: 80%
- Specificity: 79%
- Precision: 70%
- Recall: 80%

Recommendation

X-Education will have to mainly focus below important features responsible for good conversion rate are :

- Total Time Spent on website: If Leads are spending more time on website, they be easy to convert into a hot lead.
- What is your current occupation_Working Professional : Those who are 'Working Professional' have higher lead conversion rate ,company should focus on working professionals and should focus on getting more number of leads.
- Lead Source_Welingak Website : Those leads who got to know about course from 'Welingak Website' have higher conversion rate, so company can focus on this website to get more number of potential leads.
- Lead Origin_Lead Add Form: Leads who have engaged through 'Lead Add Form' having higher conversion rate so company can focus on it to get more number of leads cause have a higher chances of getting converted.
- Last Activity_SMS Sent: Lead whose last activity is sms sent can be potential lead for company.