LEAD SCORING CASE STUDY

Siddhartha Krishna J

M H Keerthana

Deepthi Mangal

Problem Statement :

X Education sells online courses to industry professionals. 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%.

Business Goal:

X Education needs help in selecting the most promising leads, i.e. the leads that are most likely to convert into paying customers.
The company needs a model wherein you a lead score is assigned 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%.

Solution methodology:

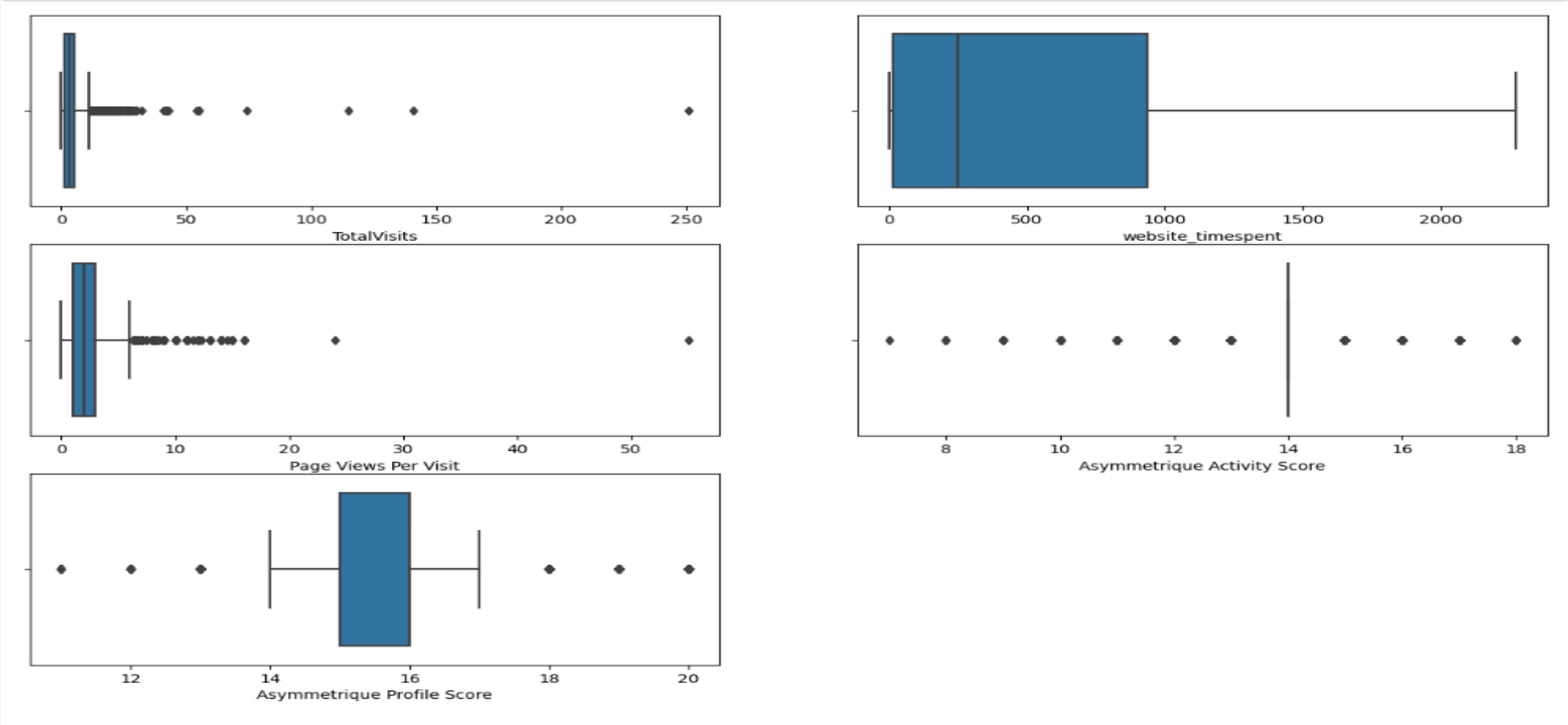
- Data loading and understanding data
- Data cleaning
- EDA
- Data pre-processing and feature scaling
- Model Building
- Model Evaluation
- Conclusion

Data Cleaning:

- Total Number of Rows =37, Total Number of Columns =9240.
- Single value features like “Magazine”, “Receive More Updates About Our Courses”, “Update me on Supply” ,“Chain Content”, “Get updates on DM Content”, “I agree to pay the amount through cheque” etc. have been dropped.
- Dropping “Last Notable Activity” because values in “Last Activity” is similar.
- Removing the “Prospect ID” and “Lead Number” which is not necessary for the analysis.
- The columns "XEducation_hear", "Lead Quality" and "Lead Profile" have null percent which greater than 50 which is very high so we dropped these columns.
- The missing value in the columns 'Asymmetrique Activity Index', 'Asymmetrique Profile Index','Asymmetrique Activity Score' and 'Asymmetrique Profile Score' are of type missing not at random, so after performing EDA on it we dropped ‘Asymmetrique Activity Score’ column.
- We replaced the null values with mode for the columns with type object and with median for numerical data types.

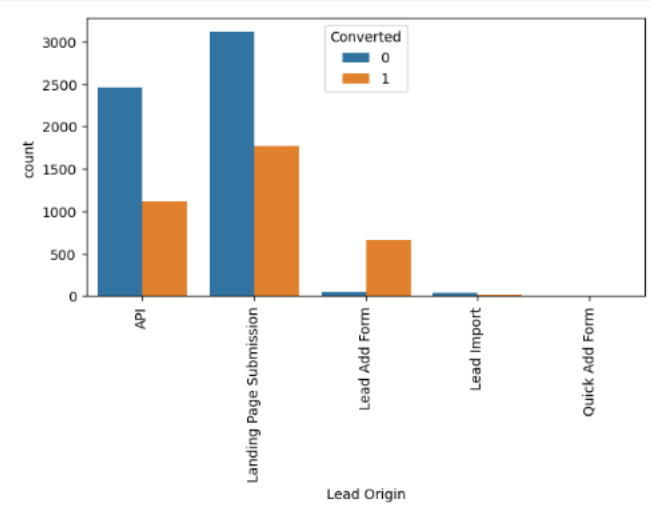
EDA

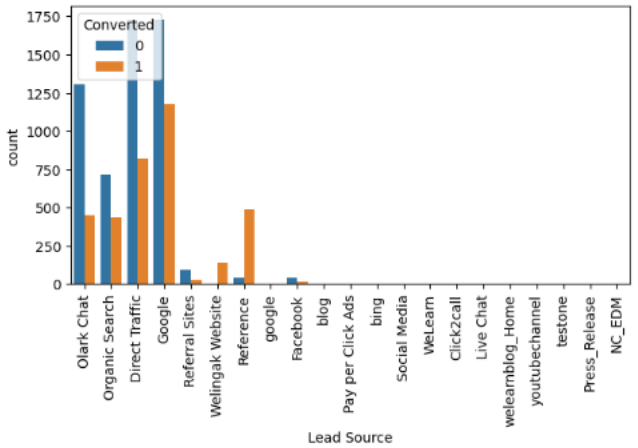
Univariate Analysis of Numerical variables



Outliers are present in the columns TotalVists, page views per visit, asymmetric activity and profile score, so we capped values according to the distribution of data in its respective columns, as part of outlier treatment.

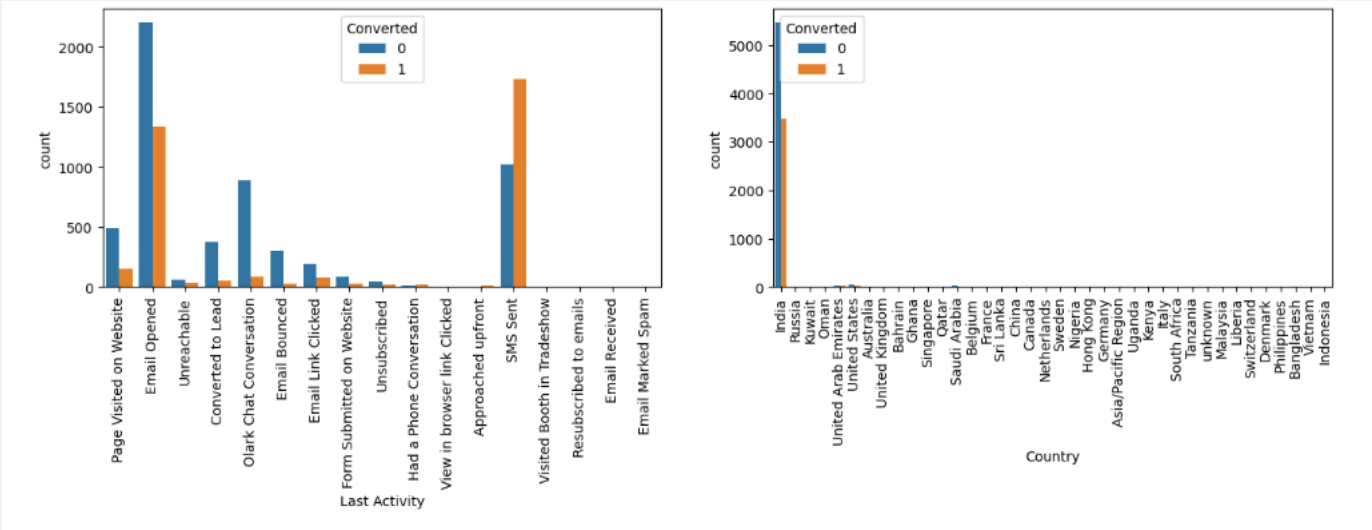
Bivariate Analysis for Categorical Variables

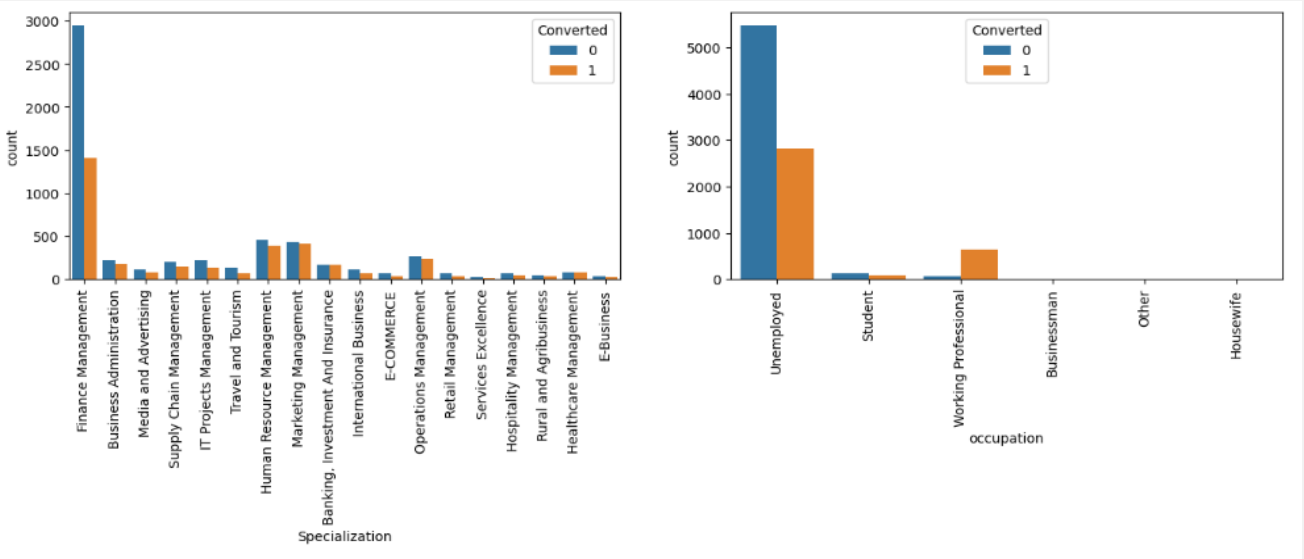




Most number of converted members has lead origin as landing page submission, and google as a lead source

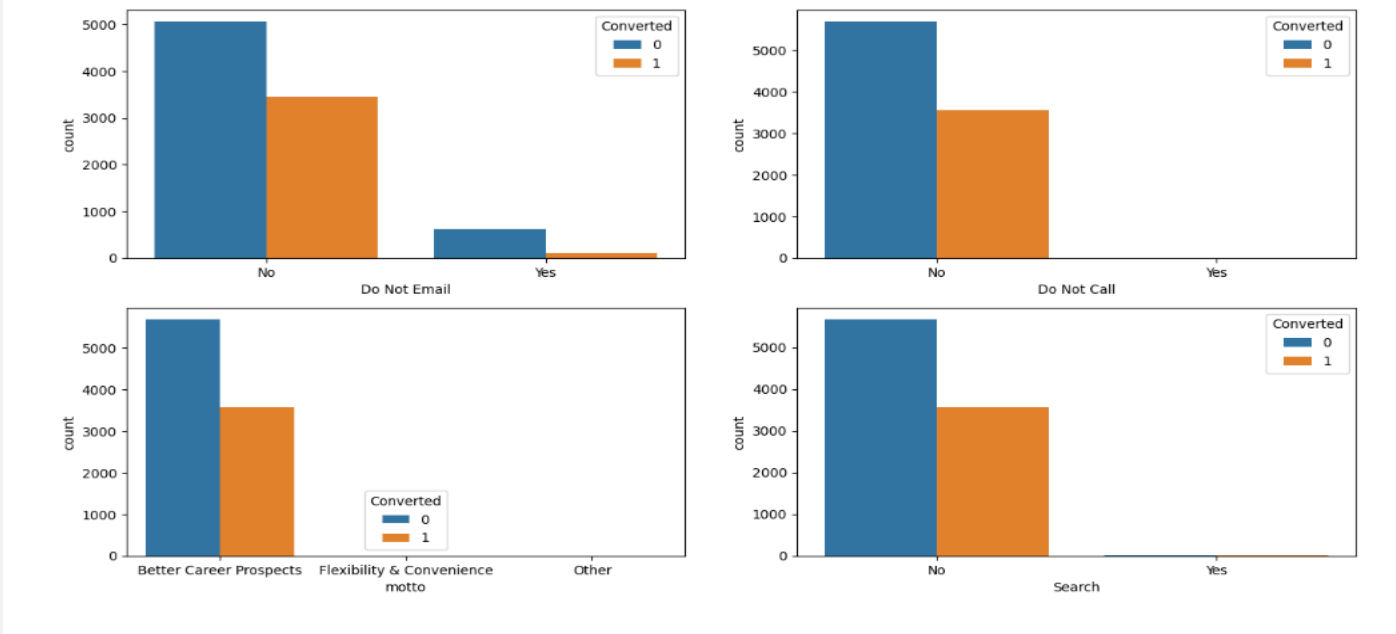
Lead who converted most have sms set as last activity and are mostly from country india

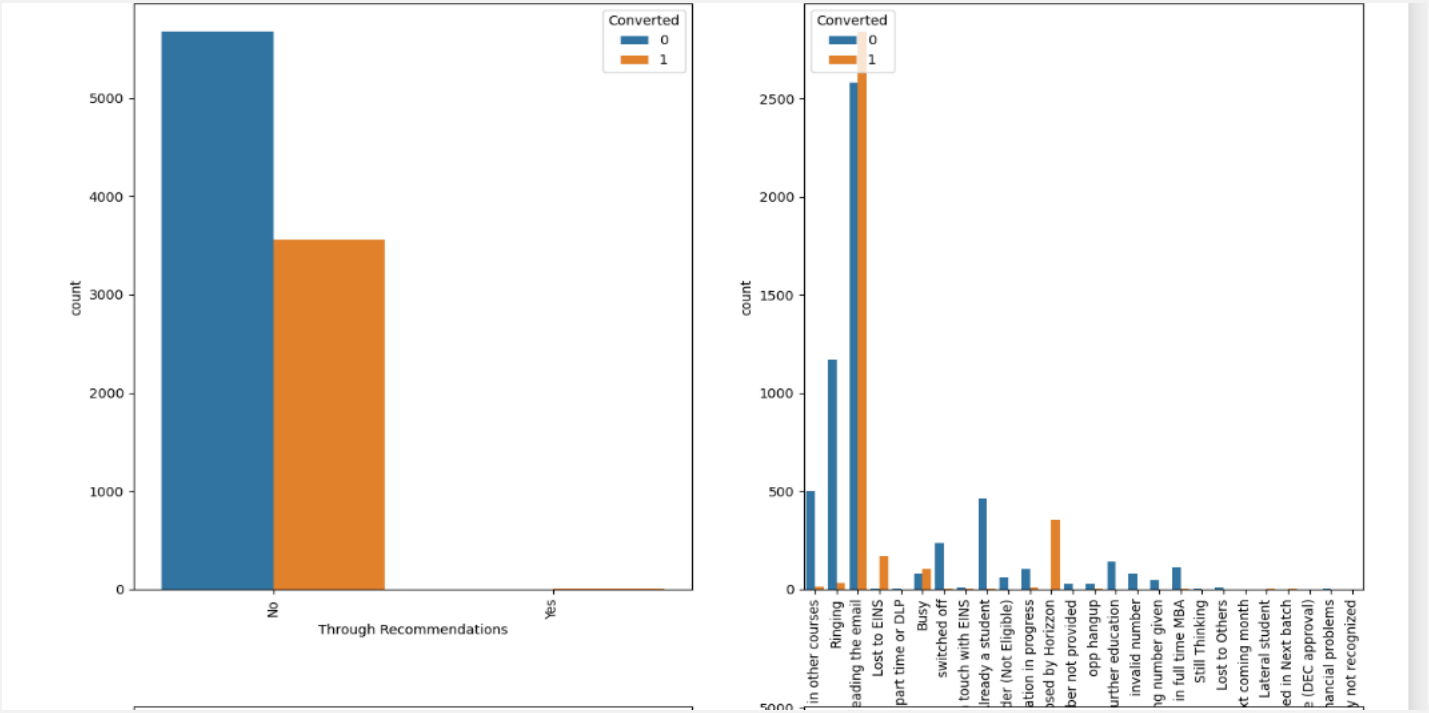


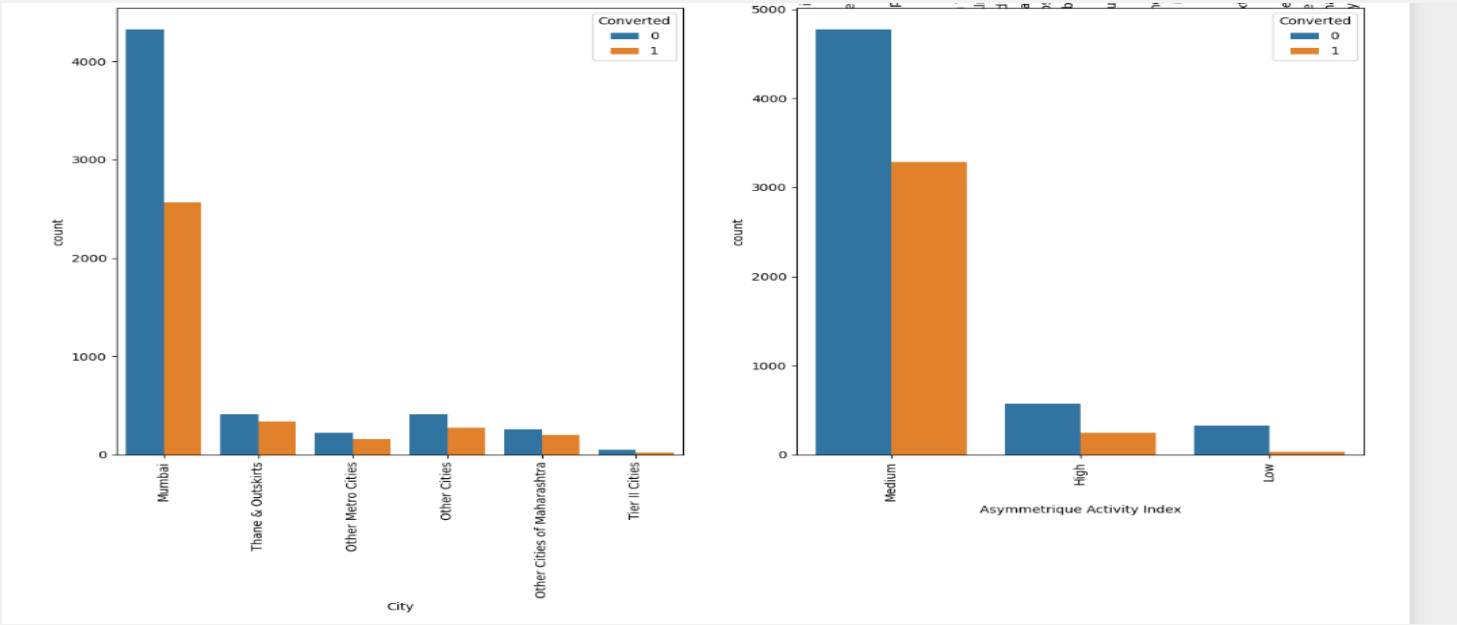


Most number of converted memebrs are from finance management as specialisation, unemployed as occupation.
We should mainly focus on the unemployed people because thier conversion rate is higher when compared to working professionals

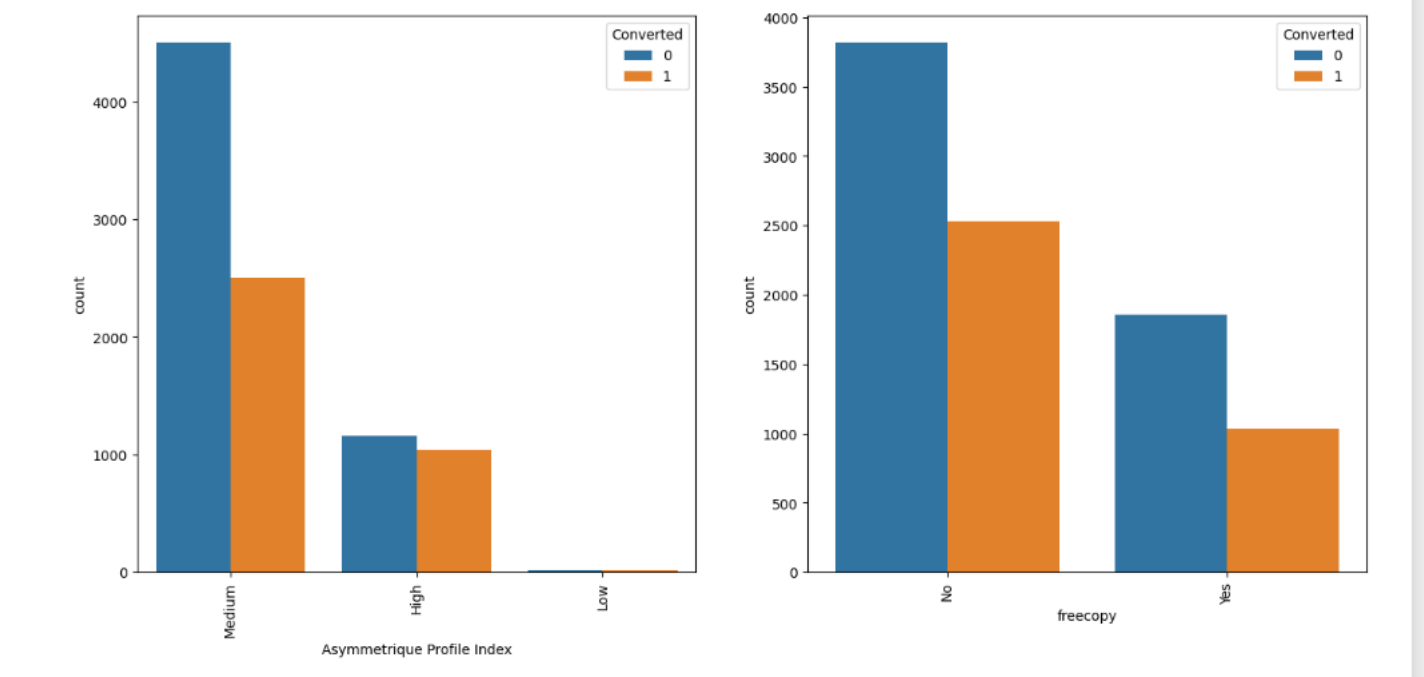
Leads who are okay to get the course details through emails and calls have high chance of saying okay to the course and whose motto is for 'better career prospect’ have high chance of being converted





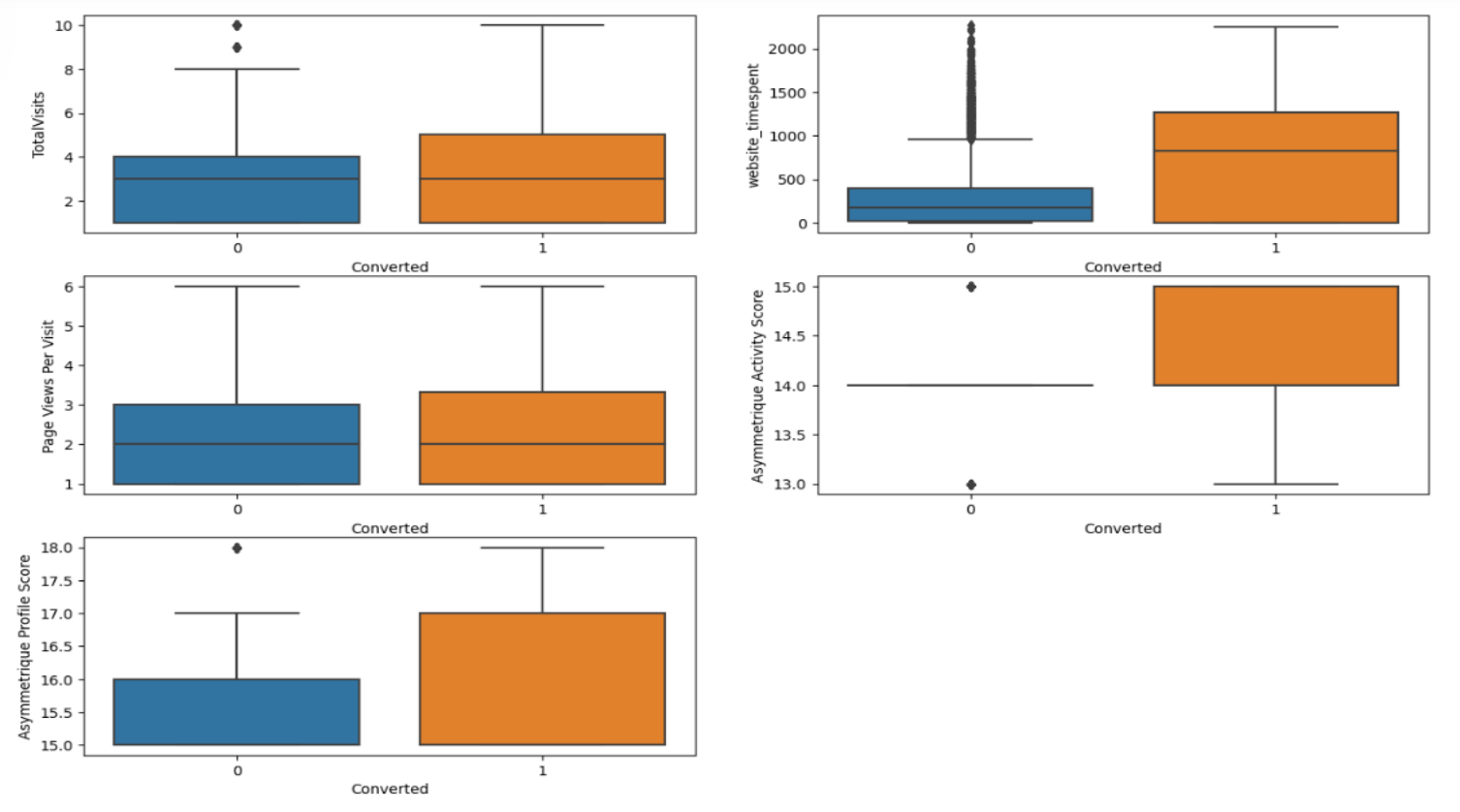


Leads who are reading and responding to mails, whose activity index is medium and are not recommended by other people are the ones who joined.
Most of the people who are in Maharashtra(assumption) are more likely to enroll



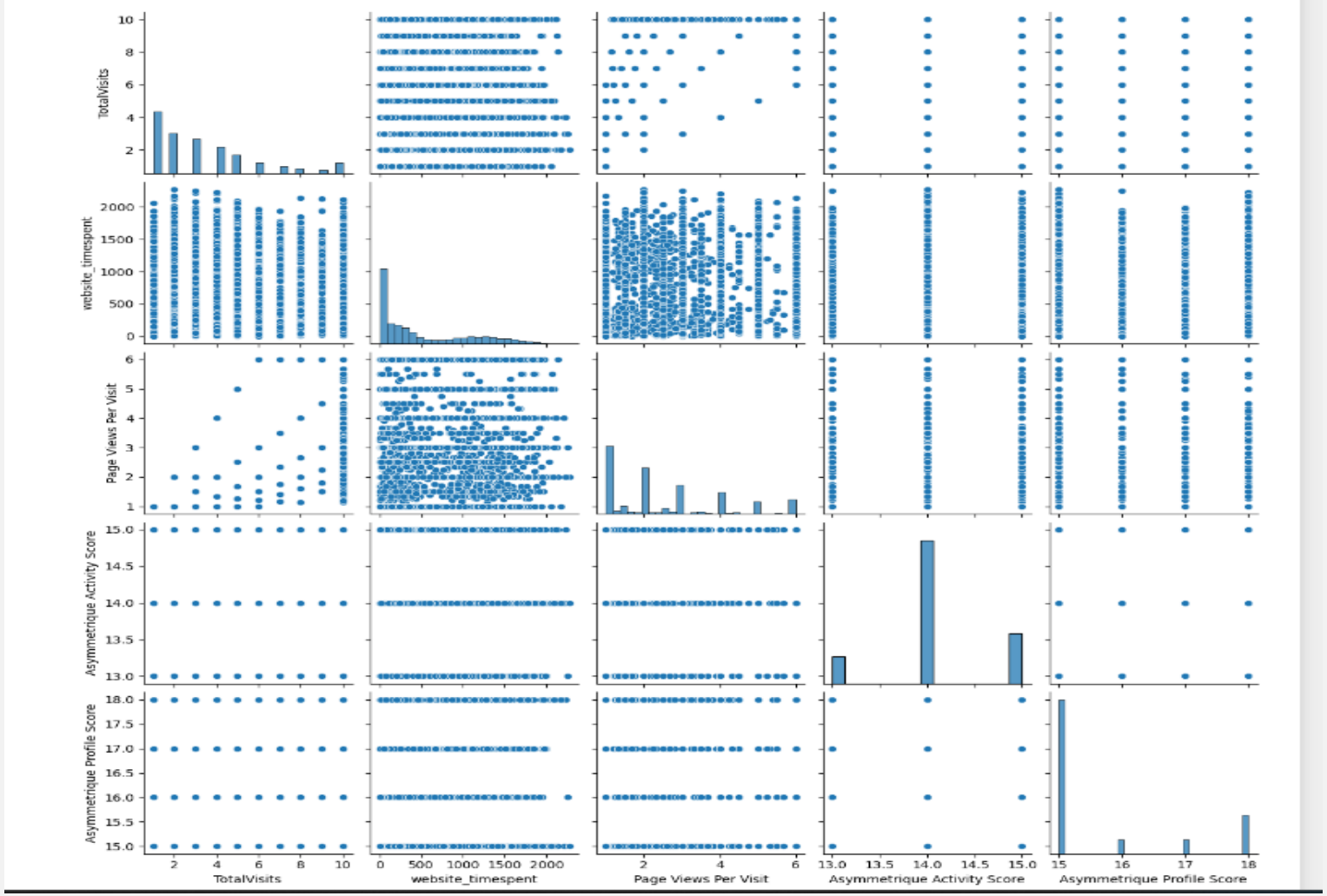
Most number of members who are interested in free copy, whose profile index is of medium and contacted through sms have high chance of getting enrolled in the course

Bi-variate Analysis for Numerical variables



Most number of converted members total visits and website_timespent are high. Median of both converted and not converted is almost same in the case of page views per visit. If asymmetric activity score more than 14 and profile score' is greater than 16, chances of conversion is high

Multi-variate Analysis



- The distribution of total amount of time spent on the website is same across the people with n number of visits
- There is a linear relationship between page views per visit and total visits

Data Pre-processing and Feature Scaling

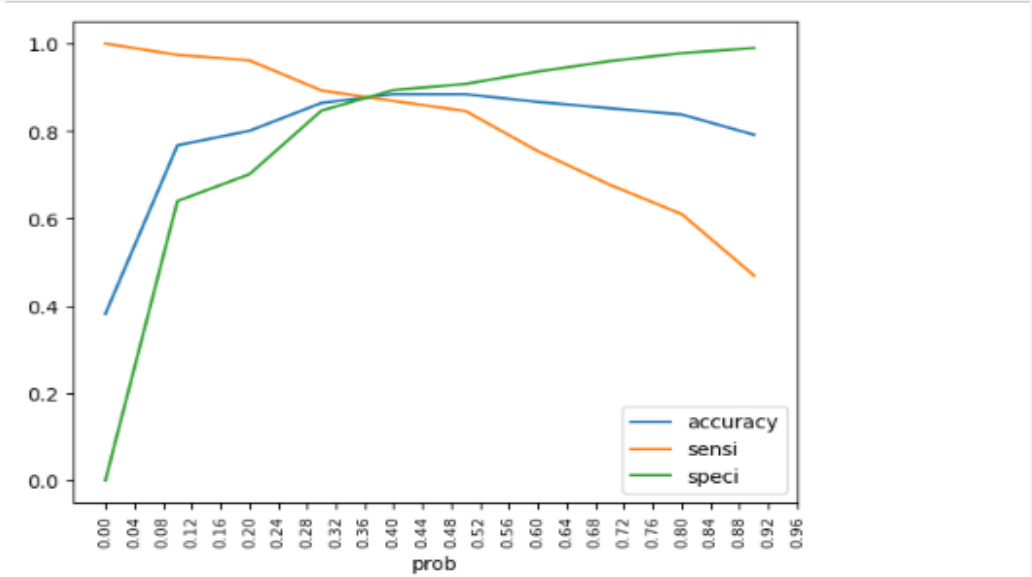
- Converted Nominal categorical variables to Numerical variables using one hot encoding and ordinal variables with label encoding.
- Feature scaling was done on Numerical variables using standardization method.

Model Building

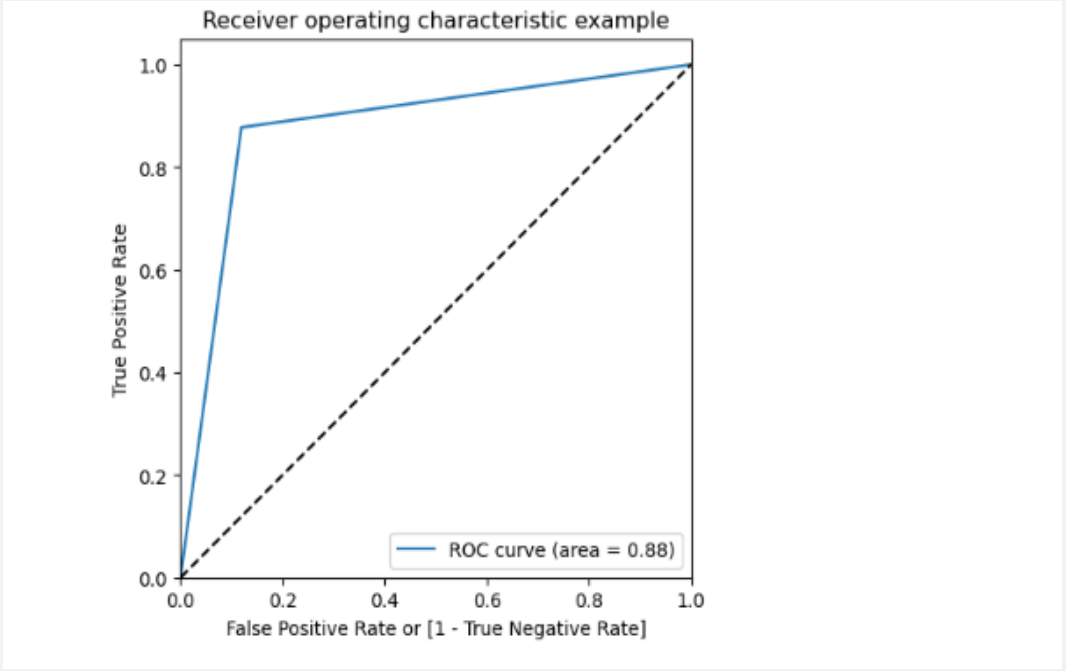
- Data is split into Train and Test Sets with 70:30 ratio respectively as first step of regression
- Used RFE for Feature Selection with 30 variables as output
- Model building was done by removing the variable whose p-value is greater than 0.05 and VIF value is greater than 5.
- Variables that impact conversion rate is seen in the figure →

Features
Do Not Email
Tags_Will revert after reading the email
Last Activity_Email Bounced
Last Activity_SMS Sent
Lead Origin_Lead Add Form
Tags_Closed by Horizzon
Last Activity_Olark Chat Conversation
occupation_Working Professional
website_timespent
Last Activity_Page Visited on Website
Tags_Ringing
Last Activity_Converted to Lead
Tags_Busy
Tags_Graduation in progress
Tags_Lost to EINS
Last Activity_Had a Phone Conversation
Country_Australia
Tags_in touch with EINS

Optimal cut-off probability and ROC Curve



By Observing the above graph we can say that the optimal prob value is 0.36



Model Evaluation

```
Accuracy of the model with cutoff probability of 0.36: 87.96
Sensitivity: 0.88
Specificity: 0.12
FalsePositive rate: 0.12
Positive Predictive rate: 0.82
Negative predictive: 0.92
```

➡ Train dataset

```
Acuuracy of final model on test data set: 82.97
Confusion Matrix: [[1641   36]
 [ 436  659]]
Sensitivity: 0.6
Specificity: 0.02
FalsePositive rate: 0.02
Positive Predictive rate: 0.95
Negative predictive: 0.79
```

➡ Test dataset

CONCLUSION

- Company should focus on Working professionals as they are more likely to get converted.
- Leads whose current status is Ringing are more likely to not get converted, so company should focus on leads whose status is Lost to EINS,Closed by Horizzon,Graduation in progress, Will revert after reading the email.
- Leads who spend more time on website are more likely to get converted
- Leads who are willing to receive emails have higher chances of conversion
- Leads whose last activity is "Converted to Lead", "Olark Chat Conversation", "Email Bounced" are not likely to get converted, so focus on leads whose last activity is "Had a Phone Conversation" and "SMS Sent"Leads whose location is out of India are not likely to convert