Lead Scoring Case Study Assignment

Problem Statement

- To help X education to select the most promising leads known as ‘hot leads’ who are most likely to convert into paid customers.
- Build a logistic regression model to assign a lead score between 0 and 100 to each of the leads where the leads with higher lead score have a higher conversion chance and the leads with lower lead score have a lower conversion chance.
- Identify the driver variables and understand their significance which are strong indicators of lead conversion.
- Identify the outliers, if any, in the dataset and justify the same.
- Consider both technical and business aspects while building the model.
- Summarize the conversion predictions by using evaluation metrics like accuracy, sensitivity, specificity and precision

Data Understanding

- ‘Leads.csv’ contains all the information about the leads generated through various sources and their activities.
- This file contains 9240 rows and 37 columns.
- Out of 37 columns, 7 are numeric columns and 30 are non-numeric or categorical columns.
- Current conversion rate of the leads is 39%.
- ‘Leads Data Dictionary.csv’ is data dictionary which describes the meaning of the variables present in the “Leads” dataset

Date Cleaning

- Leads.csv :

Below columns contain more than 30% null values initially:

1. What is your current occupation
2. What matters most to you in choosing a course
3. Tags
4. Lead Quality
5. Lead Profile
6. Asymmetrique Activity Index
7. Asymmetrique Profile Index
8. Asymmetrique Activity Score
9. Asymmetrique Profile Score

Below columns have default value of ‘select’ as a dominating value which is same as null value. So, we have converted ‘select’ to ‘NA’.

1. Specialization
2. How did you hear about X Education
3. Lead Profile
4. City

All the missing values of categorical columns have been imputed with ‘NA’

Date Cleaning

- All the missing values of quantitative columns have been imputed with median as the difference between mean and median is insignificant.
- Following columns have been dropped which contain single value as their contribution is insignificant:

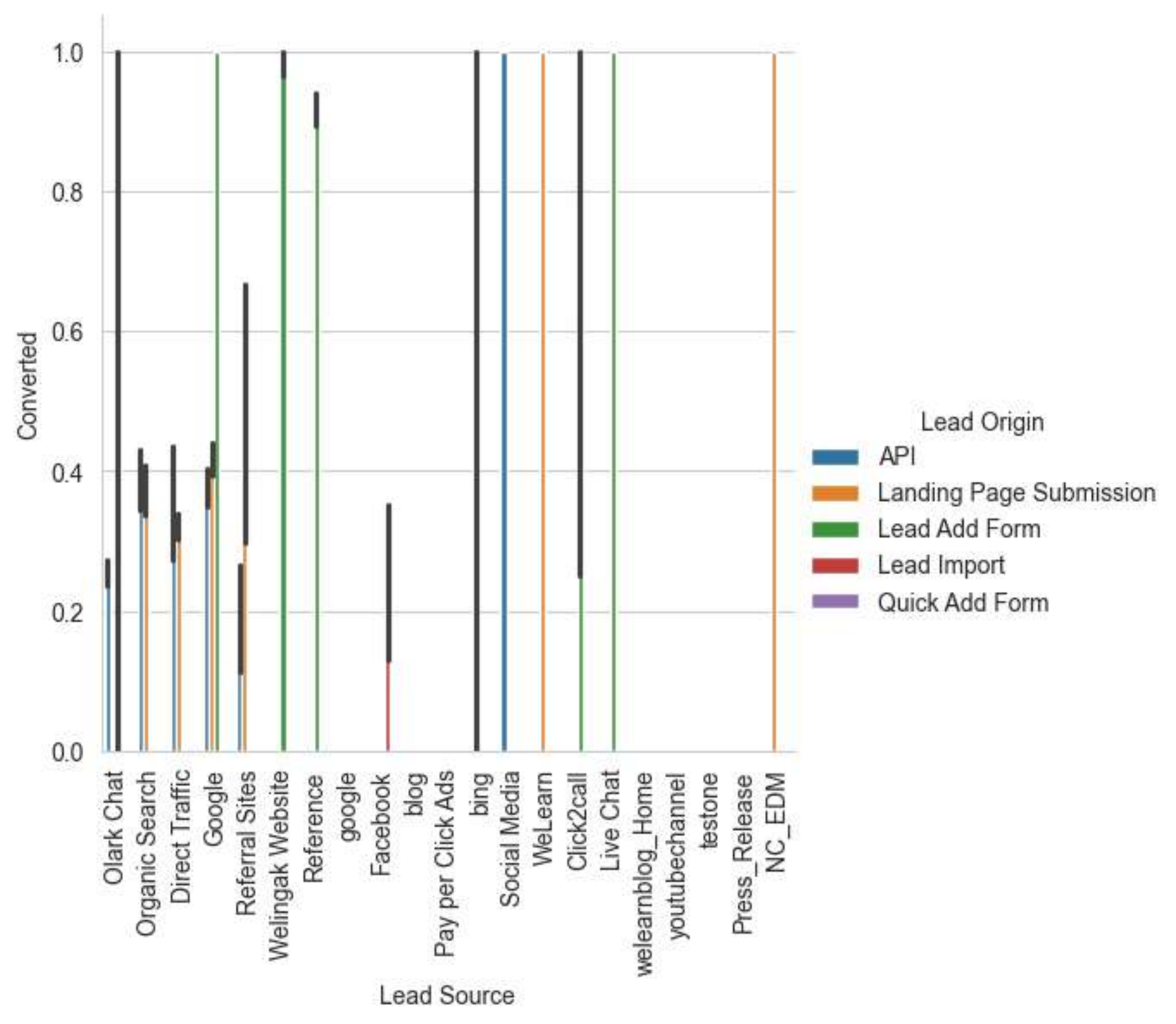
1. Magazine
2. Receive More Updates About Our Courses
3. Update me on Supply Chain Content
4. Get updates on DM Content
5. I agree to pay the amount through cheque

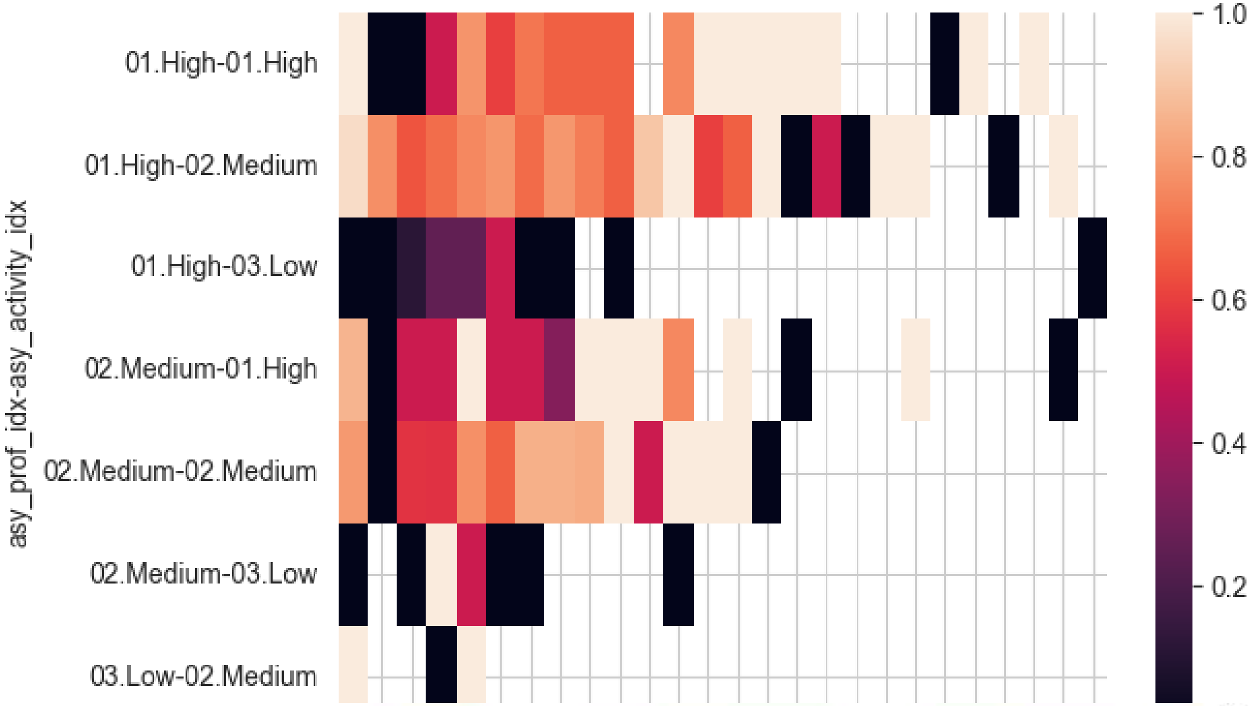
- Following columns have been dropped since percentage of missing value is more than 70%:
- How did you hear about X Education
- Lead Profile

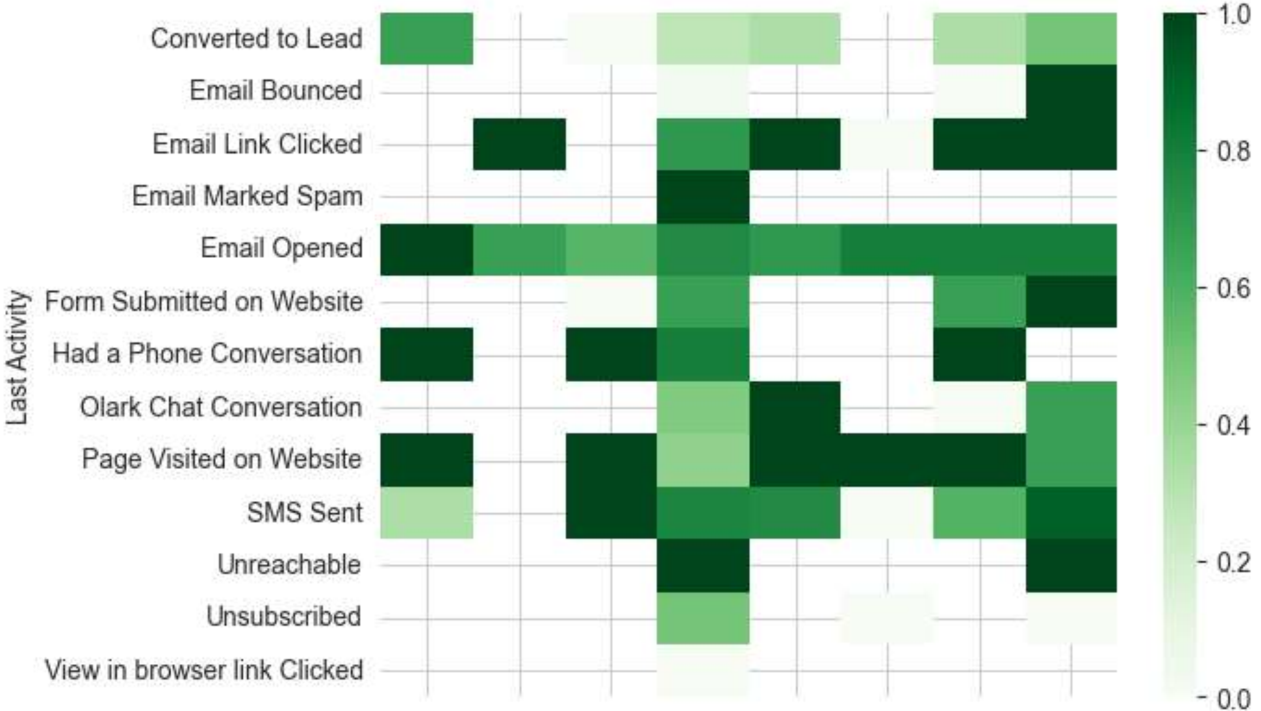
- Following columns have been imputed with mode since the percentage of missing value is low.
- Lead Source
- Lead activity

UNIVARIATE ANALYSIS – OUTLIERS

Lead Source: Google is the most effective Lead Source with an LCR of 40.4%, followed by Direct Traffic at 32.2% and Organic Search at 37.8% (contributing to only 12.5% of customers). Reference has the highest LCR at 91.8%, but there are only 5.8% of customers through this Lead Source

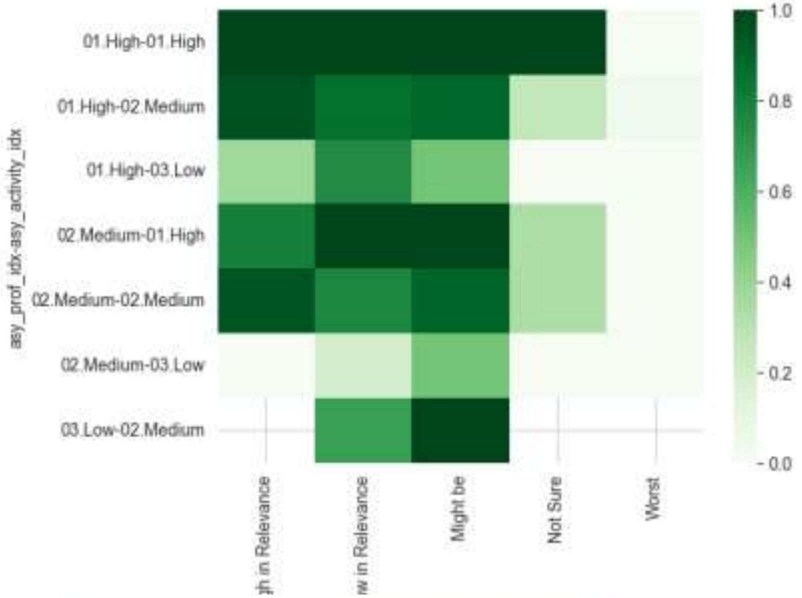


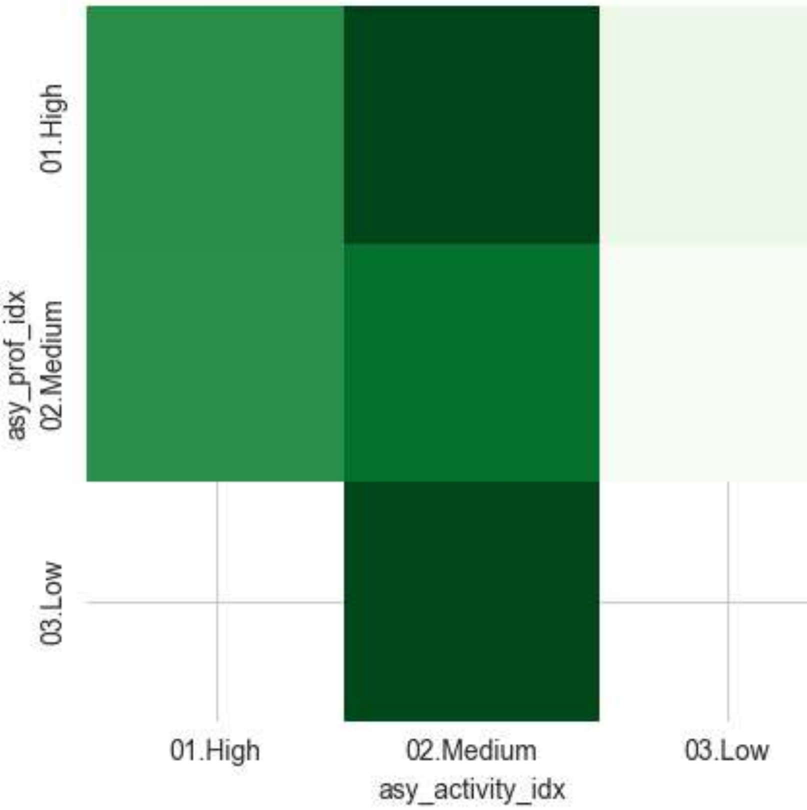




BIVARIATE ANALYSIS: CATEGORICAL VARIABLE

- Lead originated through “Lead Add Form” and “Quick Add Form” has high possibility of getting converted.
- Email is the most common last activity, with 38.3% of customers having opened an email, and 29.7% having sent an SMS





BIVARIATE ANALYSIS: CATEGORICAL VARIABLE

- Lead quality tagged with “High in Relevance” has high conversion rate history
- Low assymetric profile index and high activity index tends to be hot leads while assymetric activity index itself is another great indicator Lets check our assumptions from model itself

Data Preparation for Modeling

- Dummy variables were created for categorical columns and scaling was done. In simpler words, we can say that the scaling is used for making data points generalized so that the distance between them will be lower. Scaling was carried out in order to bring all the features into a comparable range. Then we achieved the splitting of train and test dataset with 70% and 30%.

- Feature selection was applied using RFE technique and then the elimination was done according the steps followed for fetching the column having high Pvalue & VIF, a final model was obtained after occurrence of 4 times until both VIF and p-values reached under acceptable range.

BIVARIATE ANALYSIS: CHECKING CORRELATION

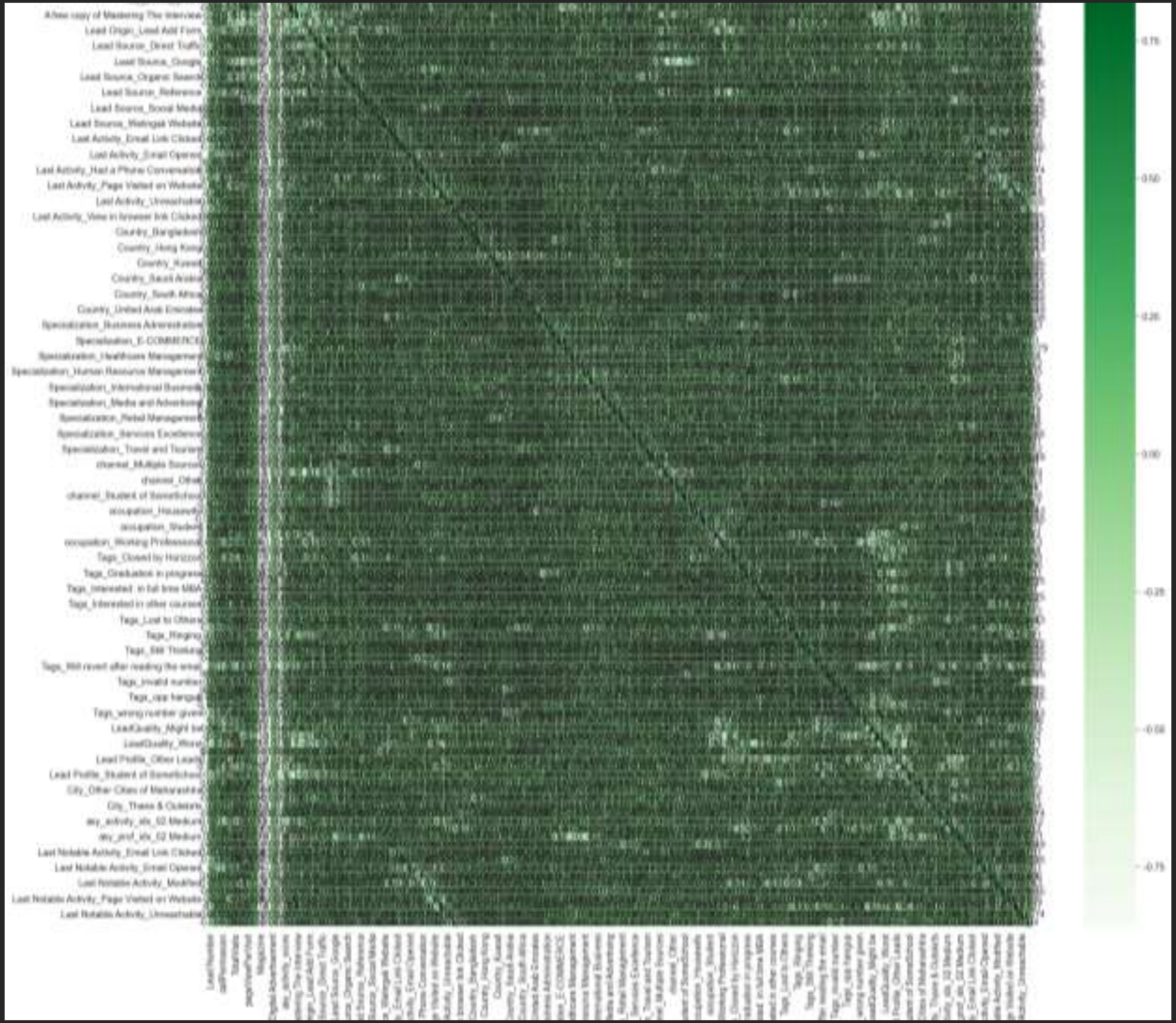
Following group of columns are positively highly correlated with each other:

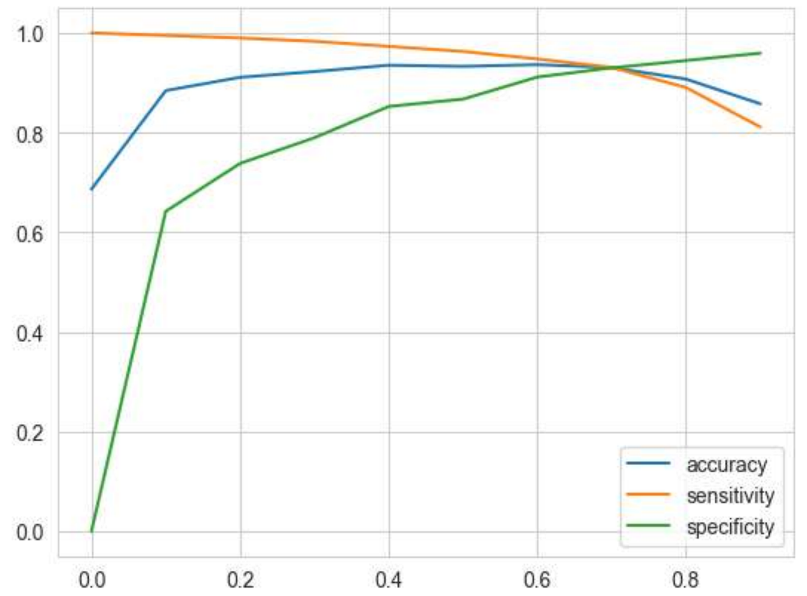
1. Search
2. Newspaper Article
3. X Education
4. Digital Advertisement
5. Through Recommendations

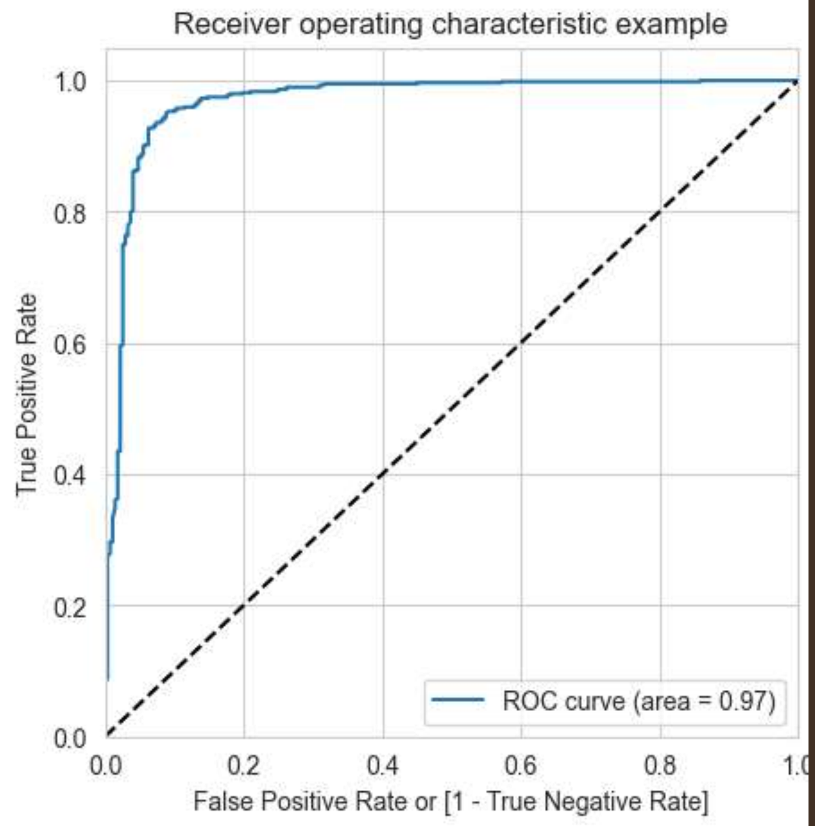
Another set of columns are also positively highly correlated with each other:

1. Total Visits
2. Total Time Spent on Website
3. Page Views Per Visit

There is a strong positive correlation between Asymmetrique Activity Index and Asymmetrique Profile Index.







MODEL EVAULATION

Train Data:

- Accuracy : 0.91
- Sensitivity : 0.97
- Specificity : 0.94

Test Data:

Accuracy : 0.90

Sensitivity : 0.87

Specificity : 1

- Based on the curve analysis, a cutoff probability of 0.35(approx.) is suggested as the optimal point for classification.
- The customers with a high lead score have a higher chance of conversion and low lead score have a lower chance of conversion

Conclusion and Recommendation

- It was found that the variables that mattered the most in the potential buyers are (In descending order) :

- The total time spend on the Website.
- Total number of visits.

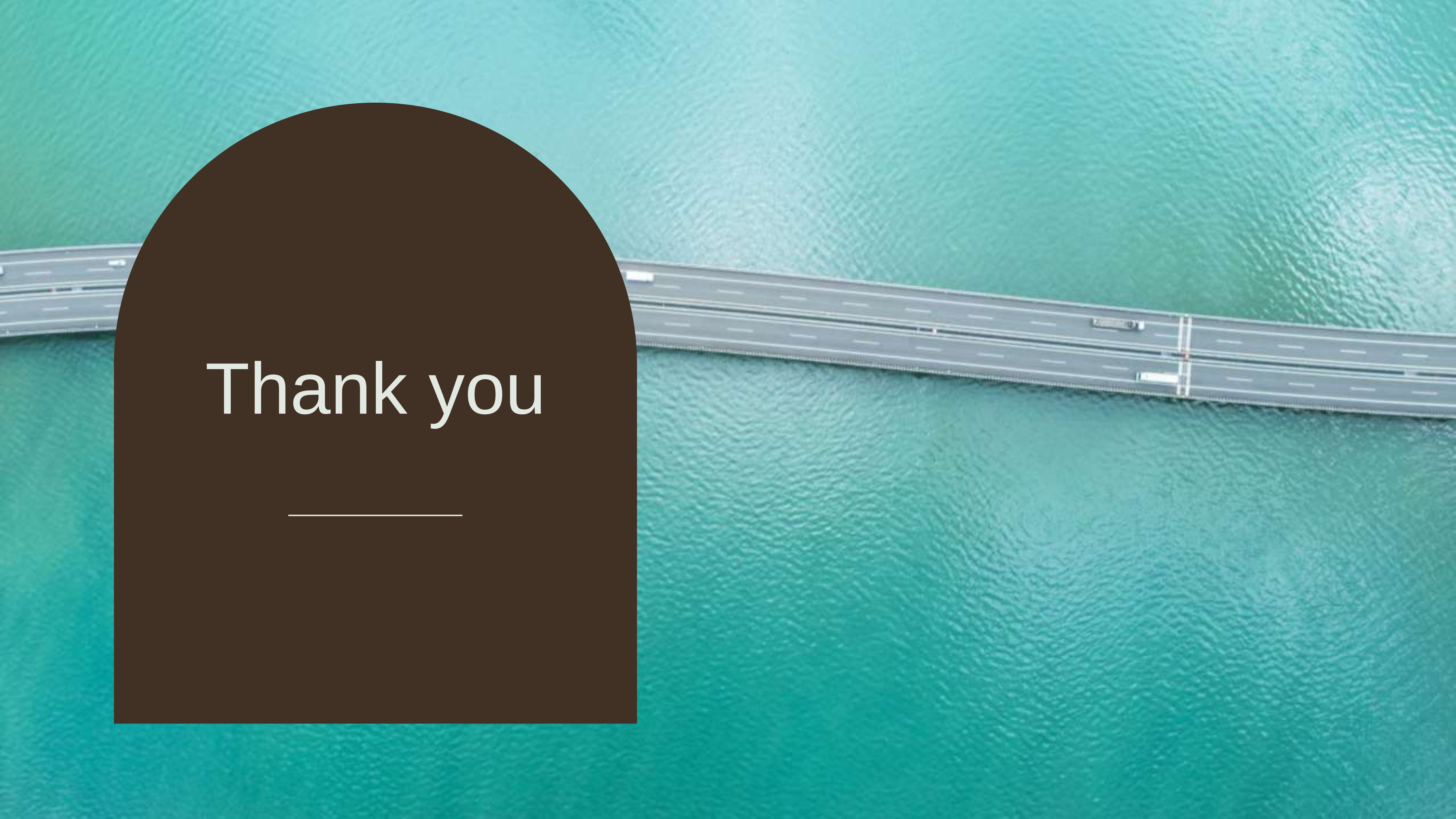
- When the lead source was:

a. Google
b. Direct traffic
c. Organic search

- When the last activity was:

a. SMS
b. Olark chat conversation

- When the lead origin is Lead add format.
- When their current occupation is as a working professional.
- Keeping these in mind the X Education can flourish as they have a very high chance to get almost all the potential buyers to change their mind and buy their courses



Thank you