Lead Score Case Study

By -

- Rohan Sharma
- Roy Mathew Benjamin
- Sahil Sharma

PROBLEM STATEMENT

An education company named X Education sells online courses to industry professionals. On any given day, many professionals who are interested in the courses land on their website and browse for courses. Although X Education gets a lot of leads, its lead conversion rate is very poor. 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.

GOAL

Our goal is to analyze the data and build a logistic regression which will help us identify the ‘Hot Leads’ which will eventually help the company and its sales team to focus on these particular leads which will increase the revenue and help business grow

Current conversion rate – 30%

Target conversion rate – 80%

STEPS TAKEN

1. Data cleaning
2. Exploratory data analysis (Univariate & Bivariate analysis)
3. Data preparation
4. Train-Test split
5. Feature scaling
6. Model Building (Recursive feature elimination, Variance inflation factor, Accuracy, Plotting, receiver operating characteristic curve)
7. Finding optimal curve-off (Precision & Recall)
8. Model prediction on the Test Set

DATA CLEANING

Total number of rows – 9240

Total number of columns – 37

Number of columns with null values – 17

Number of columns with null values (>39%) – 8

Columns with all Unique values – 2

Columns with 1 Unique value – 5

Skewed Columns (>90%) – 8

DATA CLEANING

- STEP 1 – Few rows have a value ‘Select’ which means the option in the given column was left blank hence it can replaced with NaN

- STEP 2 – Check for columns with null values & eliminate those which have 40% or more.

(Columns with >39% null values – ‘How did you hear about X Education’, ‘Lead Profile’, ‘Lead Quality’, ‘Asymmetrique Profile Score’, ‘Asymmetrique Activity Score’, ‘Asymmetrique Activity Index’, ‘Asymmetrique Profile Index’ , ‘City’)

- STEP 3 – Dropping columns which do not add much value to the data analysis

(Useless Columns - 'Country’, 'What matters most to you in choosing a course’)

- STEP 4 – Checking the number of unique values in each column and eliminating those columns which have either 1 unique value or all unique values

(Columns with all Unique values – ‘Prospect ID’, ‘Lead Number’)

(Columns with 1 Unique value – '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’

DATA CLEANING

- STEP 5 – Separating categorical value columns and numerical value columns

(Categorical value columns – ‘Lead Origin’, ‘Lead Source’, ‘Do Not Email’, ‘Do Not Call’, ‘Last Activity’, ‘Specialization’, ‘What is your current occupation’, ‘Search’, ‘Newspaper Article’, ‘X Education Forums’, ‘Newspaper’, ‘Digital Advertisement’, ‘Through Recommendations’, ‘Tags’, ‘A free copy of Mastering The Interview’, ‘Last Notable Activity’)

(Numerical value columns – ‘Converted’, ‘TotalVisits’, ‘Total Time Spent on Website’, ‘Page Views Per Visit’)

- STEP 6 – Dropping categorical value columns with skewness more than 90%

(Skewed columns – ‘Do Not Call’, ‘Search’, ‘Newspaper Article’, ‘X Education Forums’, ‘Newspaper’, ‘Digital Advertisement’, ‘Through Recommendations’, ‘Do Not Email’)

EXPLORATORY DATA ANALYSIS

After the data cleaning process we are left with 12 columns for EDA

➢ UNIVARIATE ANALYSIS

- STEP 1 – Clubbing variables of the columns which show similar attributes

❖ Specialization

(Variables clubbed – ‘Finance Management’, ‘Human Resource Management’, ‘Marketing Management’, ‘Operations Management’, ‘IT Projects Management’, ’Retail Management’, ‘Supply Chain Management’, ‘Healthcare Management’, ’Hospitality Management’)

❖ Lead Source

(Variables clubbed – ‘bing’, ‘Click2call’, ‘Press_Release’, ‘youtubechannel’, ‘welearnblog_Home’, ‘WeLearn’, ‘blog’, ‘Pay per Click Ads’, ‘testone’, ‘NC_EDM’

❖ Last Activity

(Variables clubbed - (‘Unreachable’, ‘Unsubscribed’, ‘Had a Phone Conversation’, ‘View in browser link Clicked’, ‘Approached upfront’, ‘Email Received’, ‘Email Marked Spam’, ‘Visited Booth in Tradeshow’, ‘Resubscribed to emails’)

EXPLORATORY DATA ANALYSIS

- Tags

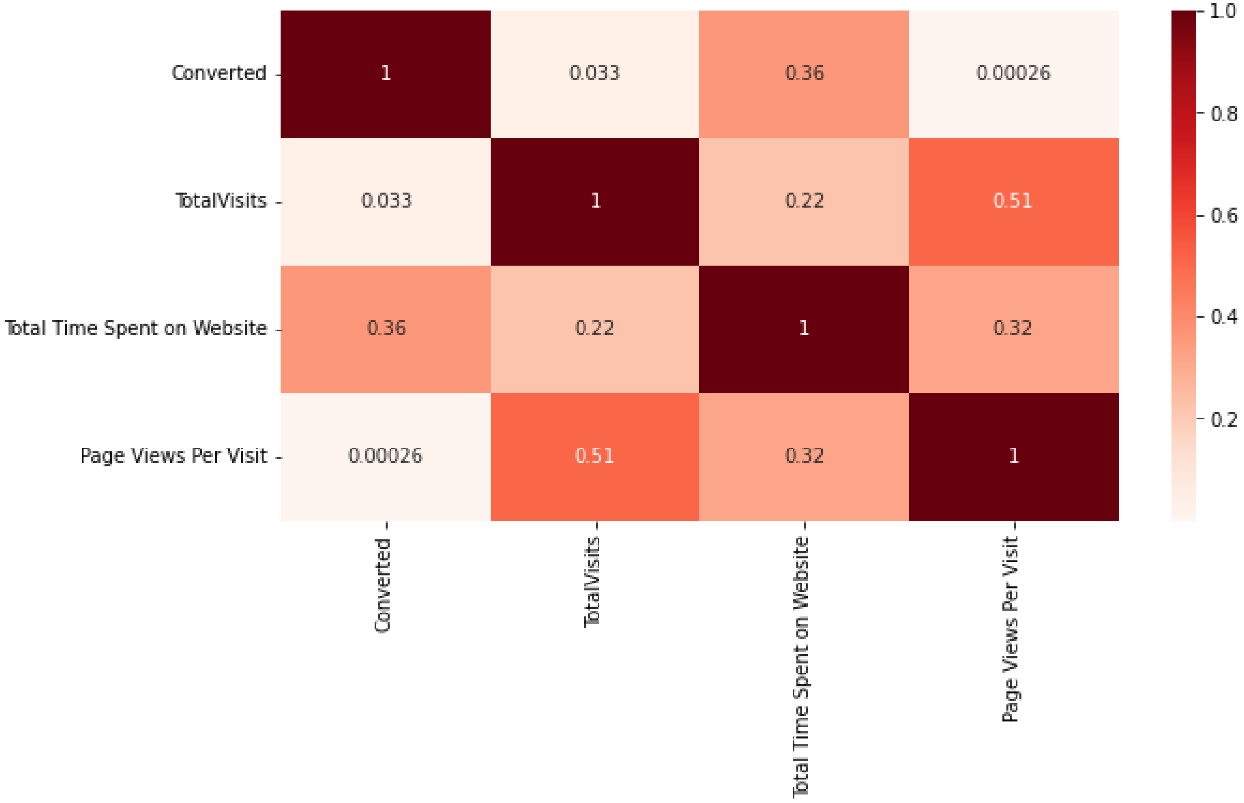
(Variables clubbed – ‘invalid number’, ‘Diploma holder (Not Eligible)’, ‘wrong number given’, ‘opp hangup’, ‘number not provided’, ‘in touch with EINS’, ‘Lost to Others’, ‘Still Thinking’, ‘Want to take admission but has financial problems’, ‘In confusion whether part time or DLP’, ‘Interested in Next batch’, ‘Lateral student’, ‘Shall take in the next coming month’, ‘University not recognized’, ‘Recognition issue (DEC approval)’, ‘switched off’, ‘Graduation in progress’, ‘Interested in full time MBA’

- Last Notable Activity

(Variables clubbed – ‘Email Bounced’, ‘Unsubscribed’, ‘Unreachable’, ‘Had a Phone Conversation’, ‘Email Marked Spam’, ‘Approached upfront’, ‘Resubscribed to emails’, ‘View in browser link Clicked’, ‘Form Submitted on Website’, ‘Email Received’)

EXPLORATORY DATA ANALYSIS

- STEP 2 – Check for correlations between numerical values



- STEP 3 – Check for Outliers

(Columns with outliers – ‘TotalVisits’, ‘Page Views Per Visit’)

EXPLORATORY DATA ANALYSIS

INFERENCE (after proceeding with the following steps) –

- Those who have a Management specialization have a high chance of conversion
- business administration and & BFSI sector also have a high chance of conversion
- API and Landing Page Submission brings higher number of leads as well as conversion.
- Lead Add Form has a very high conversion rate but count of leads are not very high.
- Maximum number of leads are generated by Google and Direct traffic.
- Conversion Rate of reference leads and leads through welingak website is high.
- We should focus to improve leads conversion on Olark chat, organic search, direct traffic and google because their count of leads are quite high
- SMS Sent activity has a very high conversion rate
- Focus should be on Email opened & those who are visiting website pages conversion rates
- Working Professionals going for the course have high chances of joining it.
- Unemployed leads are the most in terms of Absolute numbers.
- SMS Spent activity has a pretty high conversion rate
- To increase conversion rates, focus should be on Modified & email opened levels
- Quite a good linear relationship b/w Total visits & Total time spent on website
- Leads spending more time on the website are more likely to be converted.
- Website should be made more engaging to make leads spend more time.

EXPLORATORY DATA ANALYSIS

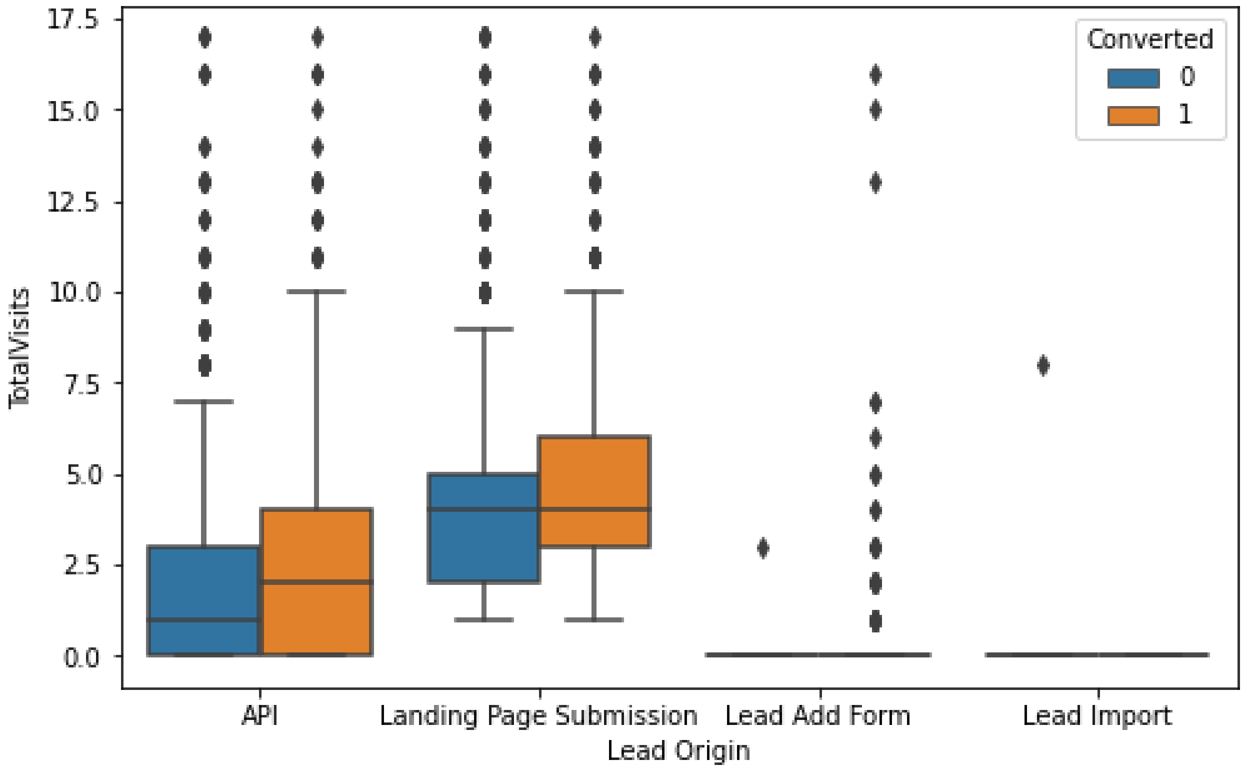
➢ BIVARIATE ANALYSIS

INFERENCE -

- All graphs look good, there are not as such outliers.
- All outliers are within a range, so we will not remove any, as these can be handy for further analysis

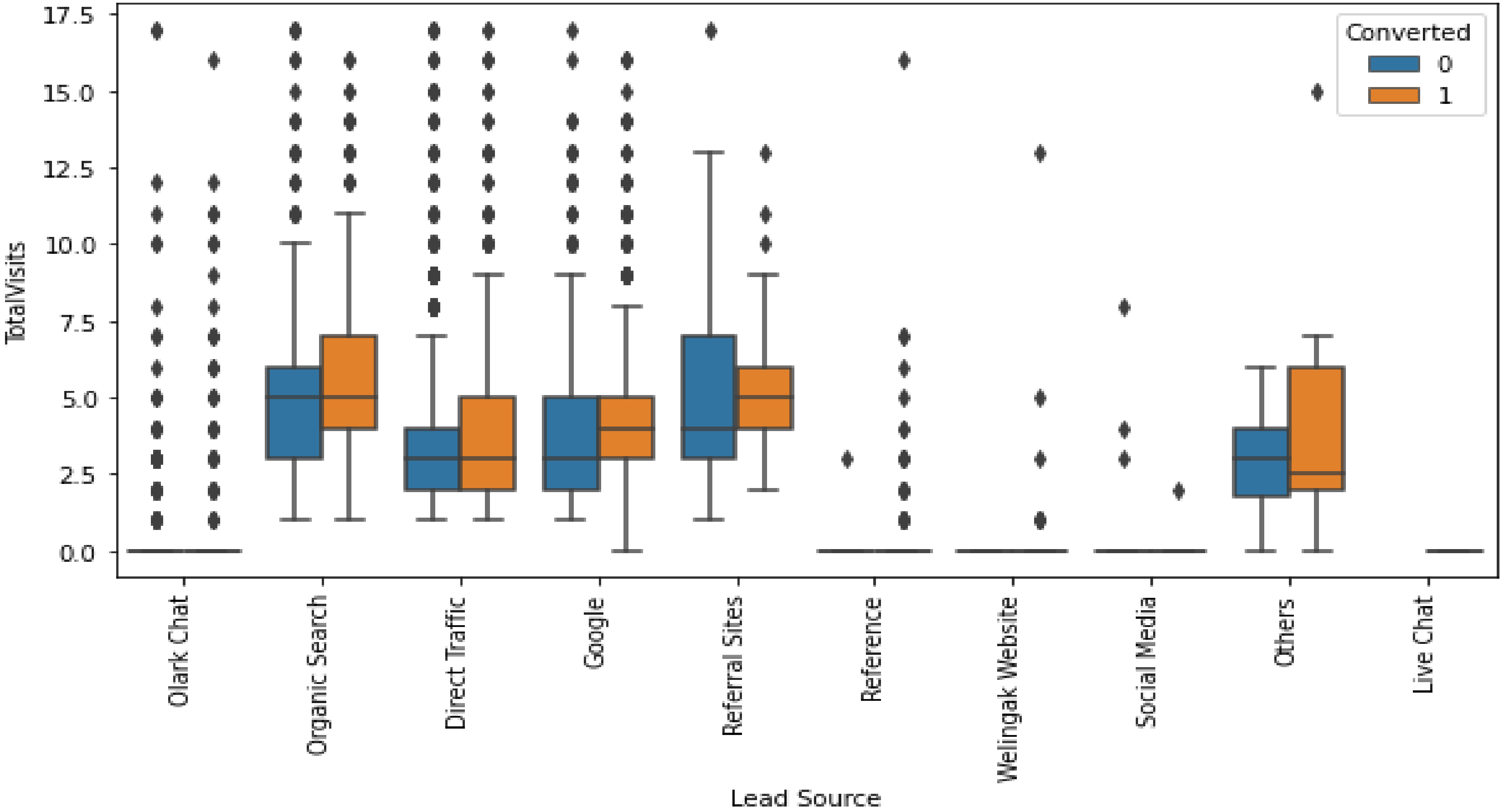
EXPLORATORY DATA ANALYSIS

- ‘TotalVisits’ vs ‘Lead Origin’



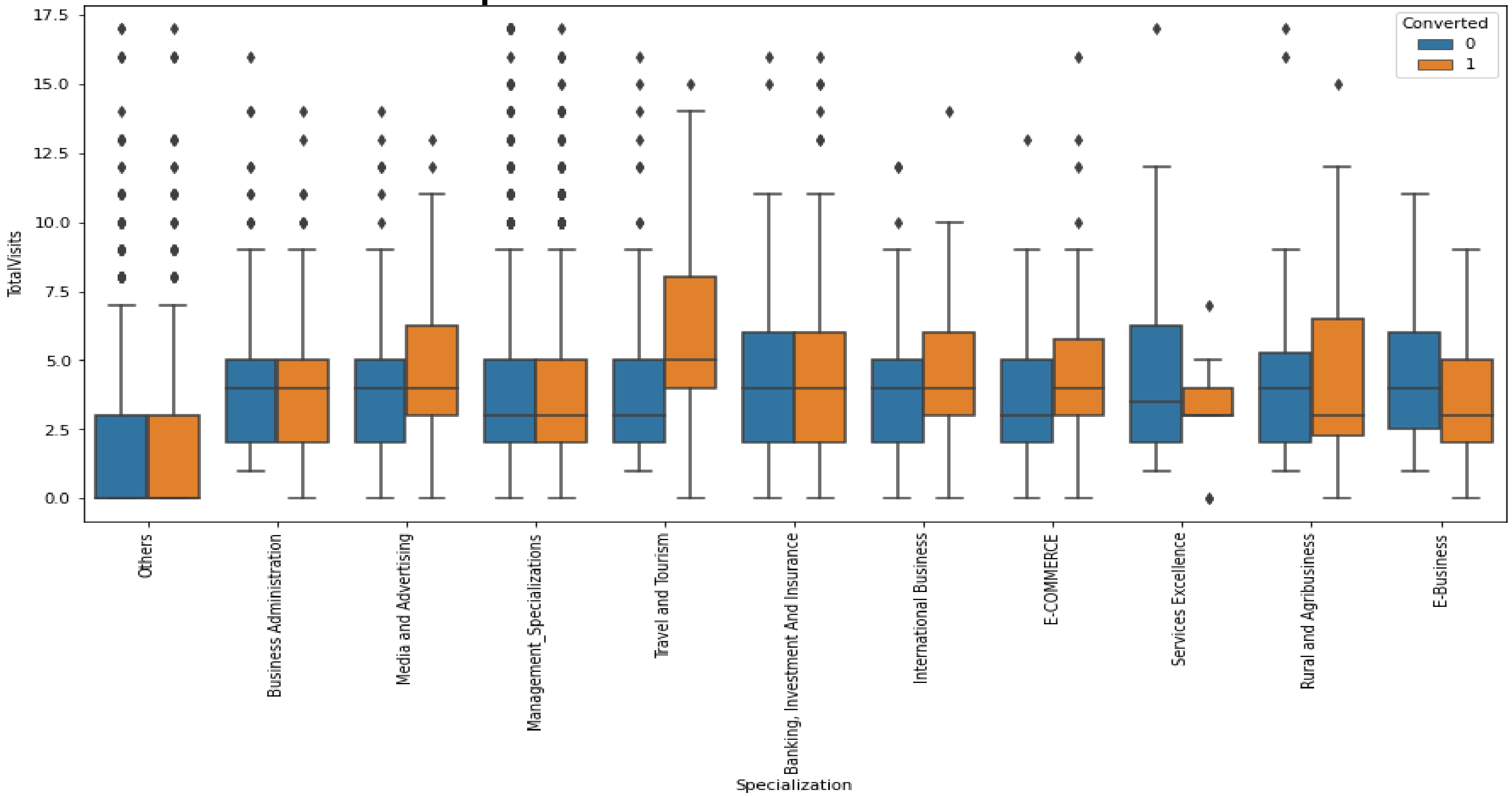
EXPLORATORY DATA ANALYSIS

- ‘TotalVisits’ vs ‘Lead Source’



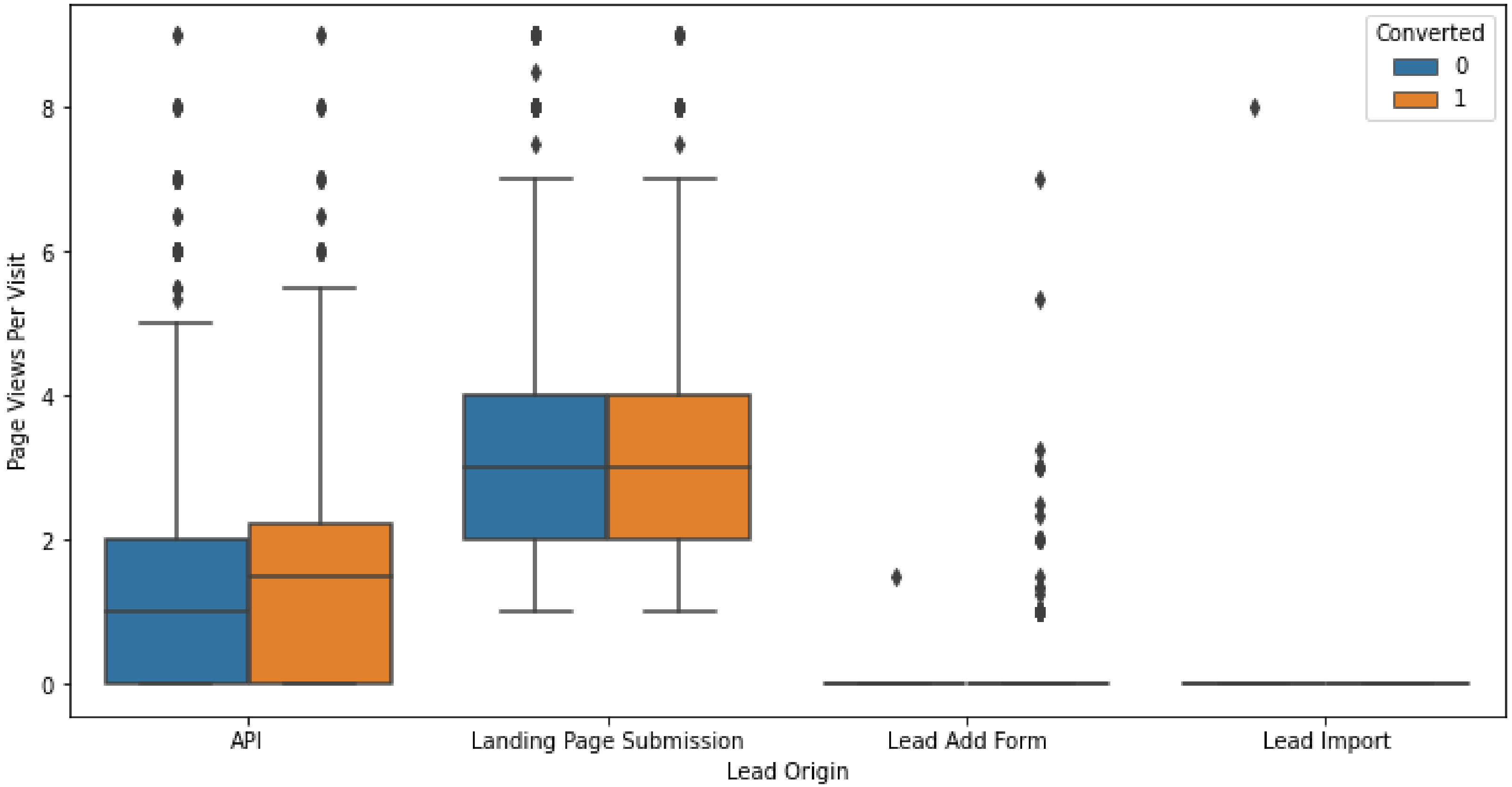
EXPLORATORY DATA ANALYSIS

- ‘TotalVists’ vs ‘Specialization’



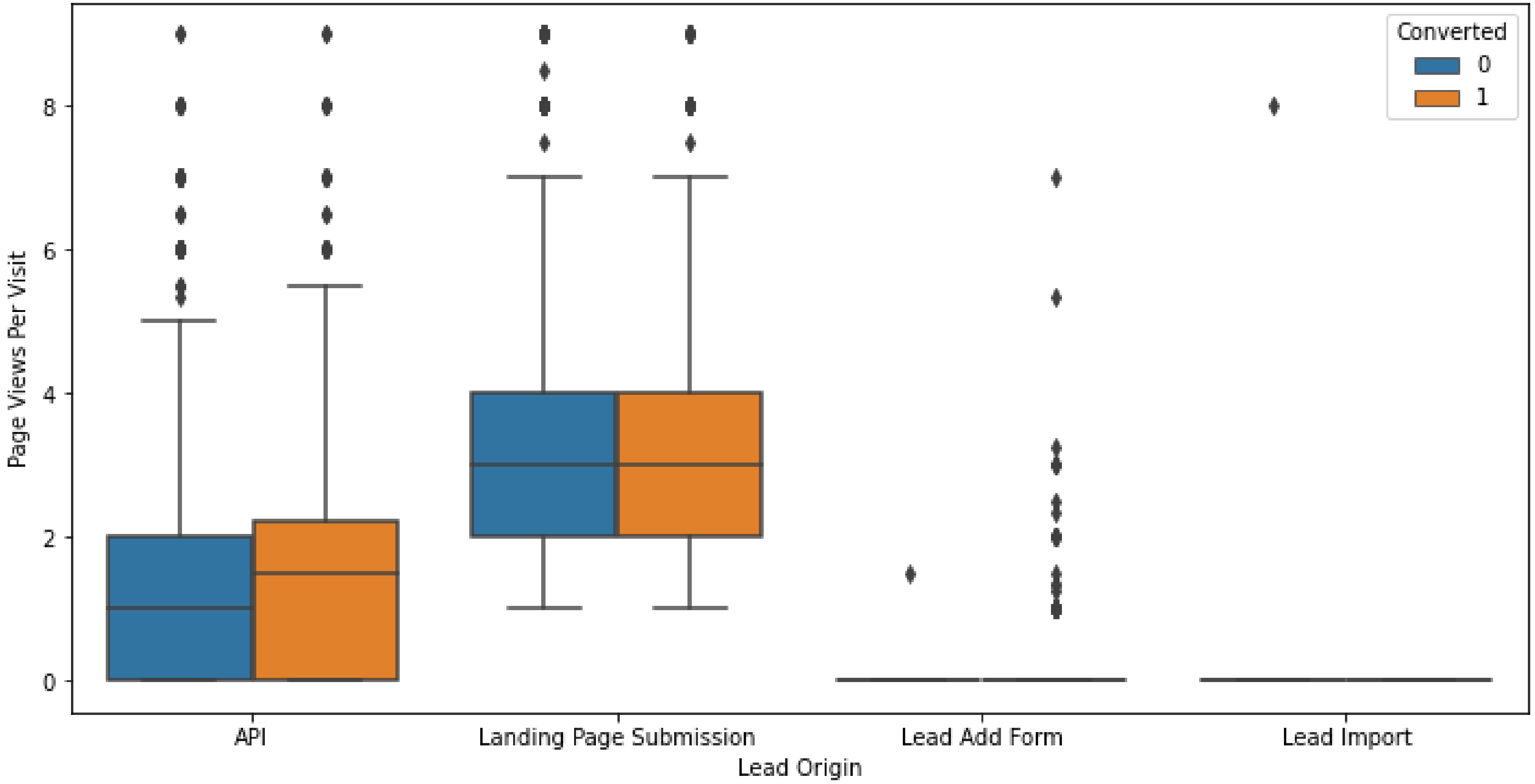
EXPLORATORY DATA ANALYSIS

- ‘Page views Per Visit’ vs ‘Lead Origin’



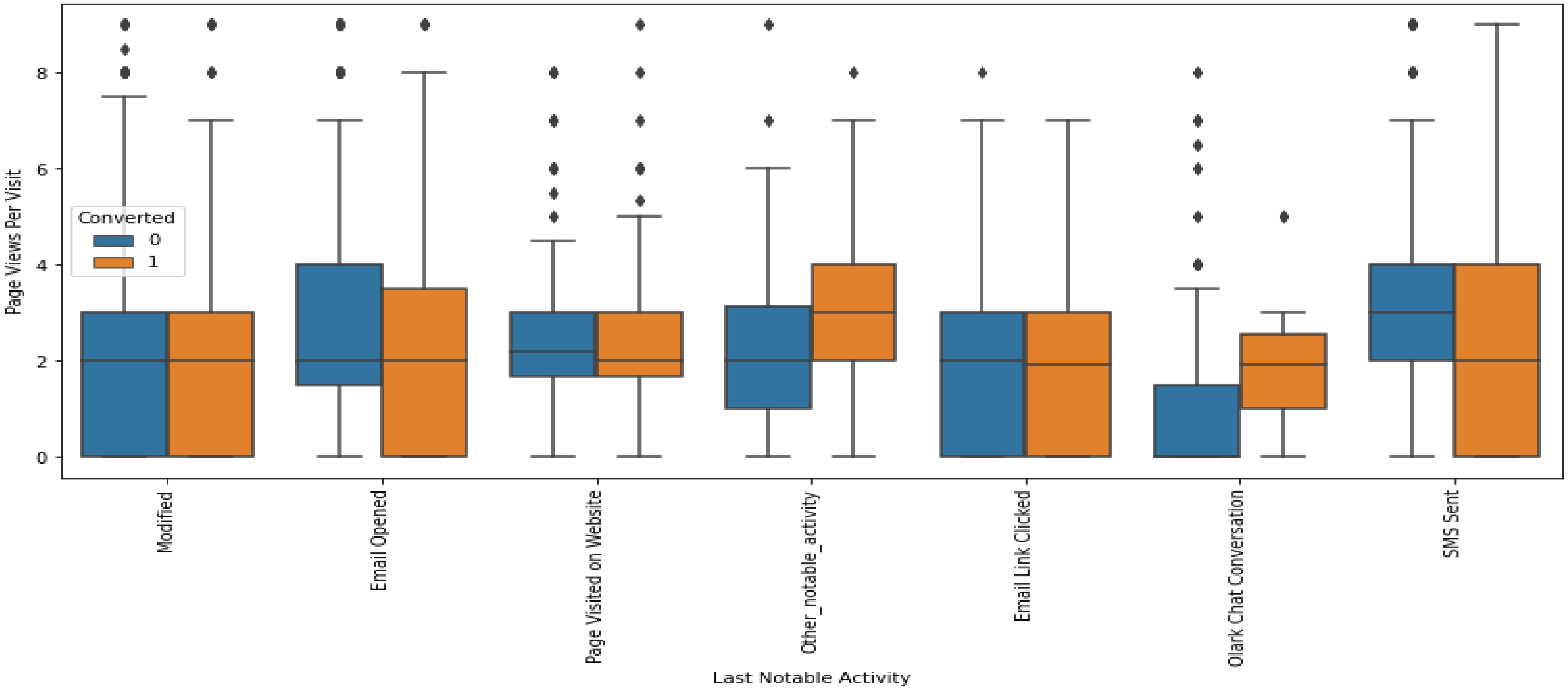
EXPLORATORY DATA ANALYSIS

- ‘Page Views Per Visit’ vs ‘Lead Source’



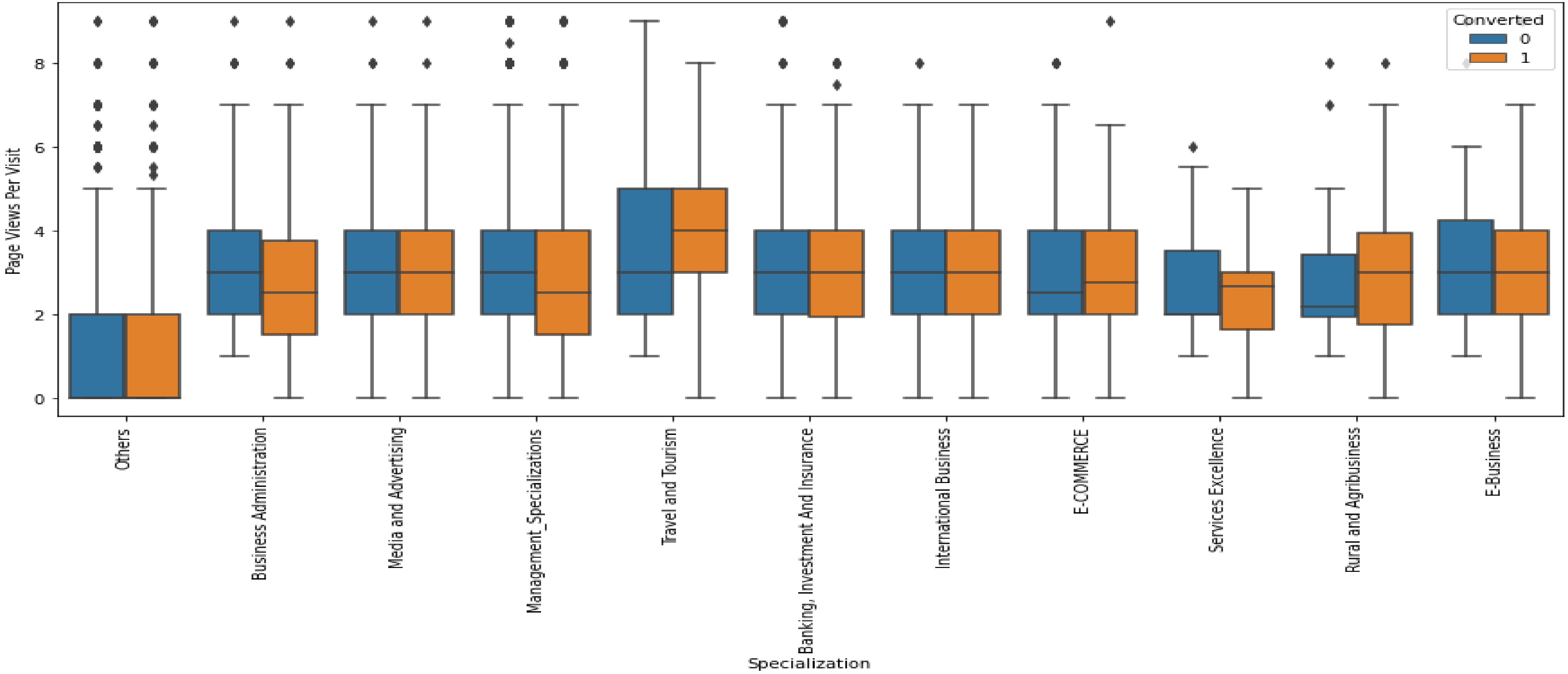
EXPLORATORY DATA ANALYSIS

- ‘Page Views Per Visit’ vs ‘Last Notable Activity’



EXPLORATORY DATA ANALYSIS

- ‘Page Views Per Visit’ vs ‘Specialization’



DATA PREPARATION

- STEP 1 – Converting binary variables (Yes/No) to (1/0)

(Binary variable column – ‘A free copy of Mastering the Interview’)

- STEP 2 - For categorical variables with multiple levels, we will create dummy features

(Columns with Categorical variable with multiple columns – ‘Lead Origin’, ‘What is your current occupation’)

- STEP 3 - Creating dummy variables for other categorical variables and dropping the level with others level

(Columns – ‘Lead Source’, ‘Last Activity’, ‘Specialization’, ‘Tags’, ‘Last Notable Activity’)

- STEP 4 – Dropping the original columns after creating dummy variables

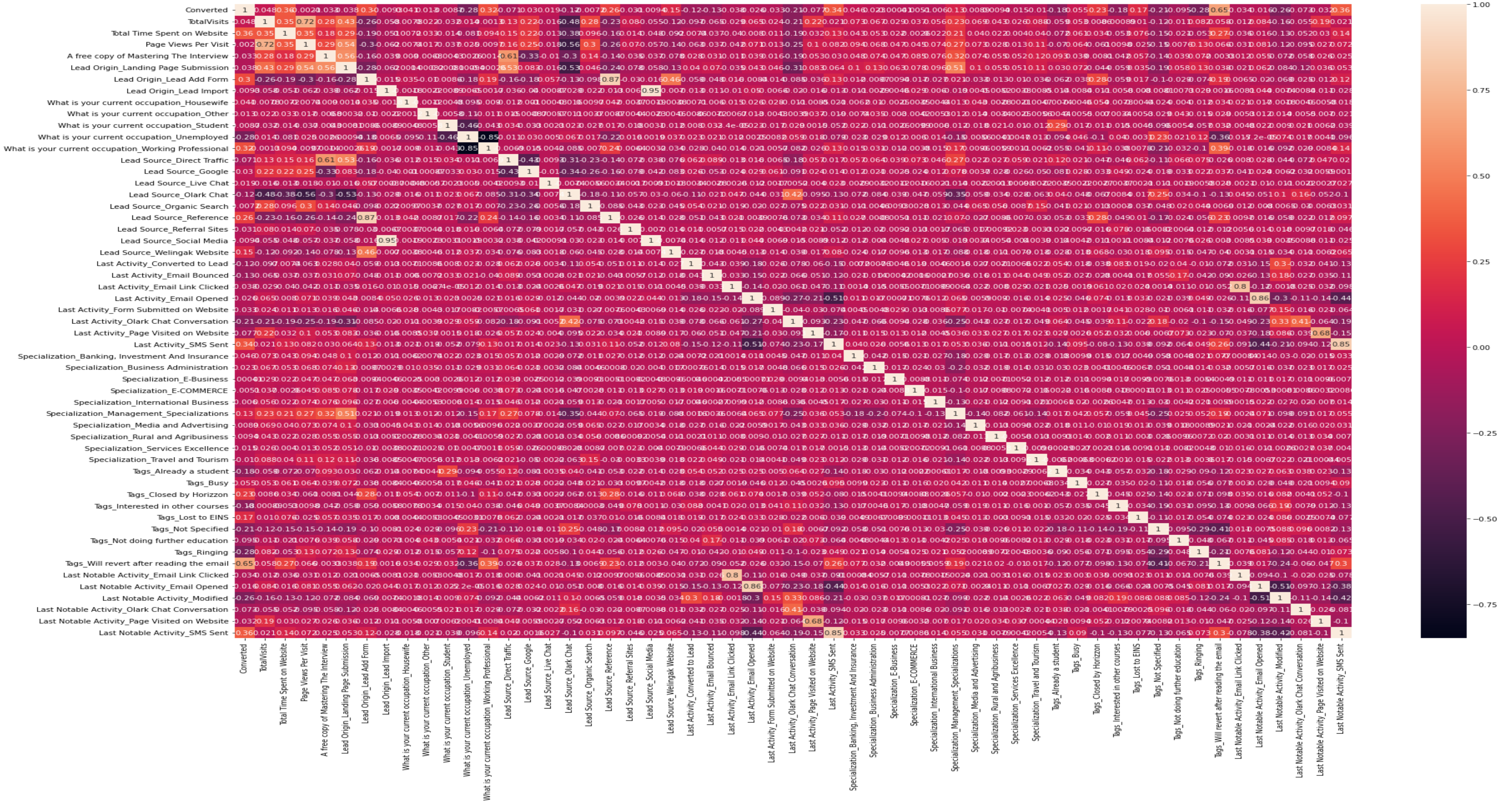
TRAIN-TEST SPLIT

- STEP 1 – Putting feature variable to X
- STEP 2 – Putting response variable to y

(Response variables – ‘Converted’)

- STEP 3 – Splitting the data into train-test

FEATURE SCALING



MODEL BUILDING

- STEP 1 – Assessing the model with Statsmodel
- STEP 2 – Building model 1
- STEP 3 – Dropping column with high P-value ('Tags_Already a student’)
- STEP 4 – Building model 2 (after this our model is fit)

MODEL BUILDING

- STEP 5 – Checking VIF

Features	VIF	
1	Lead Origin_Lead Add Form	1.77
9	Tags_Not Specified	1.65
11	Tags_Will revert after reading the email	1.65
2	Lead Source_Olark Chat	1.62
5	Last Activity_SMS Sent	1.62
12	Last Notable Activity_Modified	1.48
0	Total Time Spent on Website	1.46
3	Lead Source_Welingak Website	1.32
7	Tags_Closed by Horizzon	1.21
10	Tags_Ringing	1.12
4	Last Activity_Email Bounced	1.09
8	Tags_Lost to EINS	1.07
13	Last Notable Activity_Olark Chat Conversation	1.07
6	Tags_Busy	1.05

MODEL BUILDING

- STEP 6 – Creating a dataframe with the actual churn flag and the predicted probabilities
- STEP 7 – Creating new column 'Predicted' with 1 if Converted_prob > 0.5 else 0
- STEP 8 – Creating a confusion matrix and checking the overall accuracy (92.7%)
- STEP 9 – Metrics beyond Accuracy

➢ Sensitivity – 87.8%

➢ Specificity – 95.7%

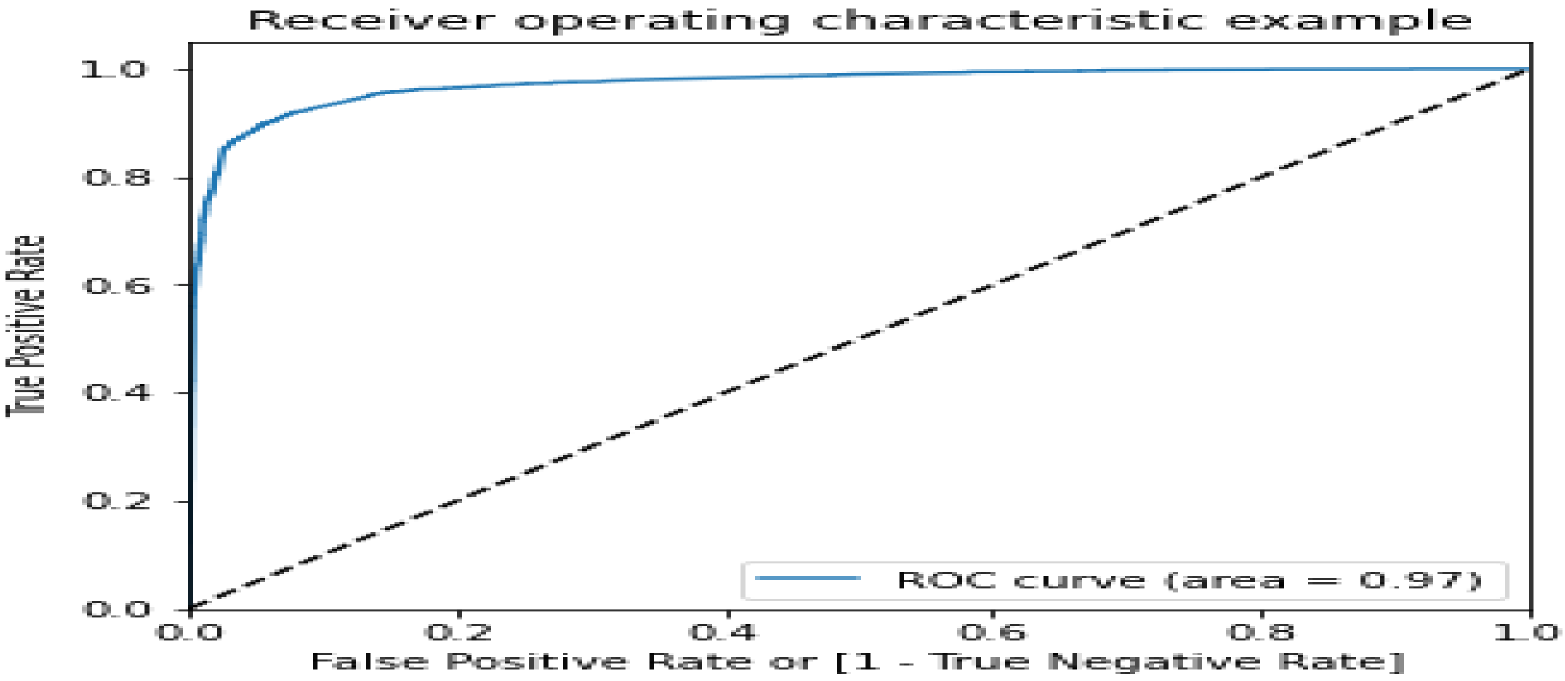
➢ False positive rate – 4.3%

➢ Positive predicted value – 92.6%

➢ Negative predicted value – 92.7%

MODEL BUILDING

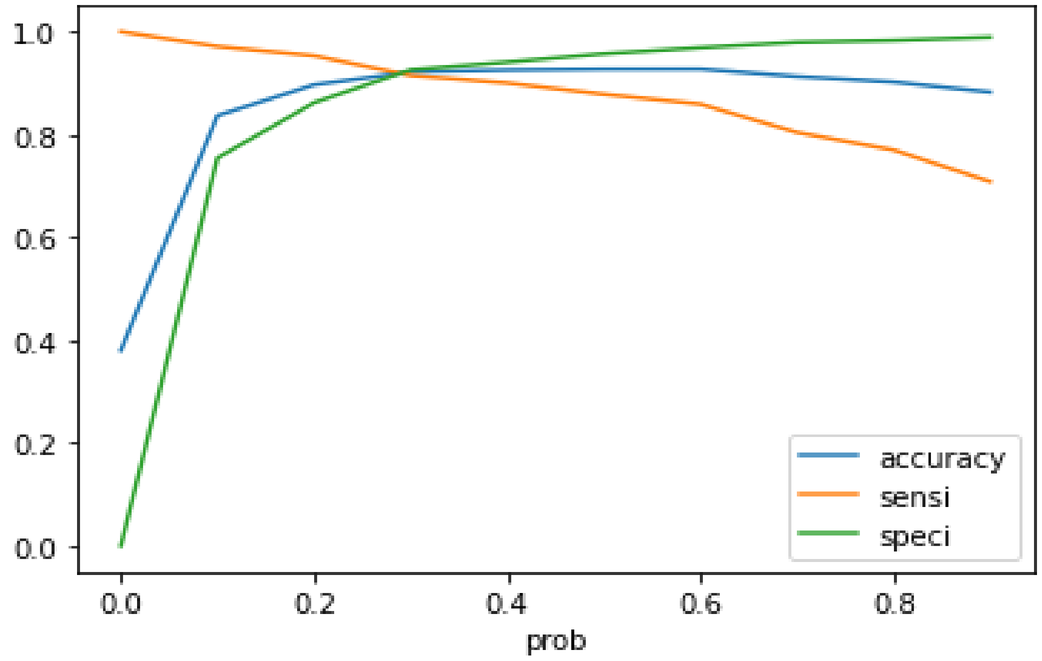
- STEP 10 – Plotting ROC curve



➢ ROC value should be close to 1. We are getting around 0.92 indicates a good predictive model

FINDING OPTIMAL CUT-OFF

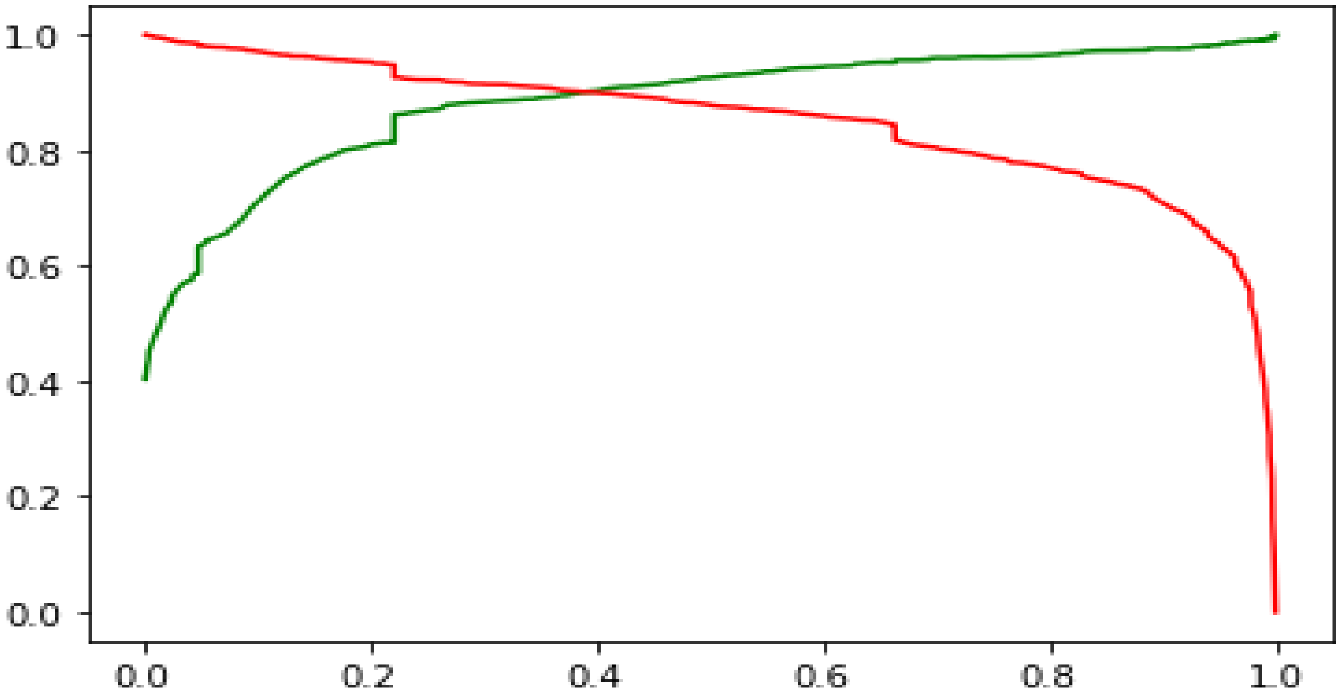
- STEP 1 – Plotting accuracy sensitivity and specificity for various probabilities



➢From the above curve Optimal cut-off is 0.3, intersection of accuracy,sensitivity & specificity

FINDING OPTIMAL CUT-OFF

- STEP 2 – Precision and Recall tradeoff



MODEL PREDICTION ON TEST SET

- STEP 1 – Converting y_test to dataframe
- STEP 2 – Removing index for both dataframes to append them side by side
- STEP 3 – Appending ‘y_test_df’ and ‘y_pred_1’
- STEP 4 – Renaming & rearranging the column
- STEP 5 – Check the sensitivity & specificity for the final model

FINAL CONCLUSION

- **Train Data :**

➢ Accuracy : 92.17%

➢ Sensitivity : 91.49%

➢ Specificity : 92.58%

- **Test Data :**

➢ Accuracy : 92.30%

➢ Sensitivity : 91.85%

➢ Specificity : 92.57%

- **The Model seems to be predicted very well by looking at the score mentioned above**