LEAD SCORING CASE STUDY

PRESENTED BY:

KRISHNA K P

HRISHIKESH BASHABOINA

SAROJ SHARMA

AGENDA

- The aim of this case study is to optimize the lead scoring mechanism based on their fit, demographics, behaviours, buying tendency etc. By implementing explicit and implicit lead scoring modelling with lead point system.

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.
- 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%.
- Now although X Education get a leads, its lead conversion rate is low. For example out of 100 only 30 are converted. To increase this conversion rate, the company wishes to identify the most potential leads aka 'Hot Leads". If they successfully identify this set of leads then the conversion rate increases and sales team will now focusing more on communicating with the potential leads rather than making calls to everyone.

GOALS OF THE CASE STUDY

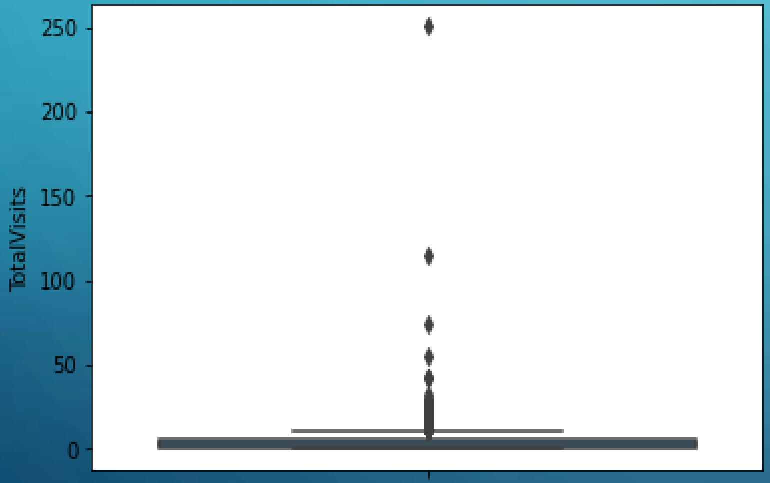
- Build a logistic regression model to assign a lead score between 0 and 100 to each of the leads which can be used by the company to target potential leads. A higher score would mean that the lead is hot, i.e. is most likely to convert whereas a lower score would mean that the lead is cold and will mostly not get converted.
- There are some more problems presented by the company which your model should be able to adjust to if the company's requirement changes in the future so you will need to handle these as well. These problems are provided in a separate doc file. Please fill it based on the logistic regression model you got in the first step. Also, make sure you include this in your final PPT where you'll make recommendations.

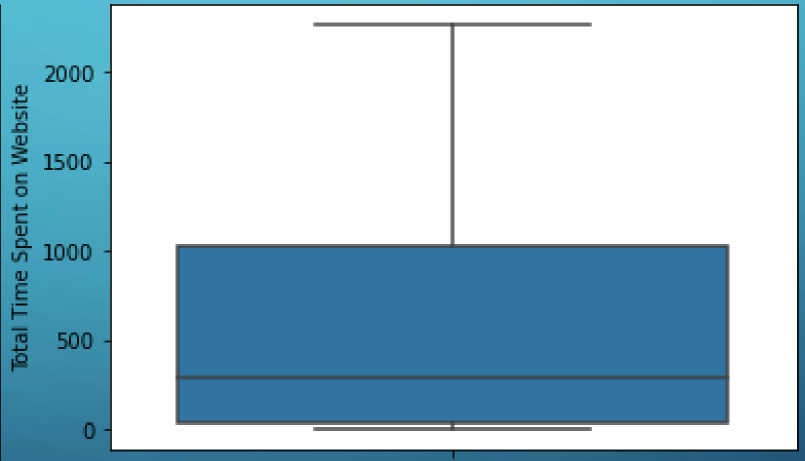
APPROACH / STRATEGY

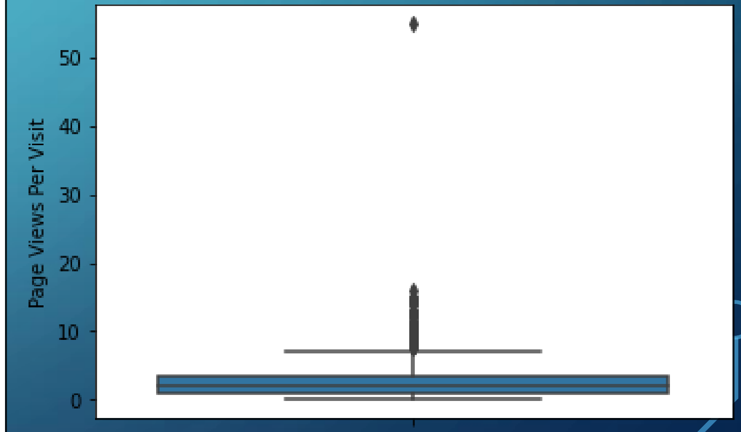
- Importing necessary libraries
- Reading & Understanding the data
- EDA – Data Cleaning and Data Preparation
- Feature scaling
- Splitting data into train and test set
- Model building
- Model evaluation – accuracy, sensitivity & specificity or Precision & Recall
- Finding final optimal cut-off
- Making predictions on the test dataset – accuracy, Precision & Recall
- Summary and Conclusion

READING, CLEANING & PREPARATION OF DATA

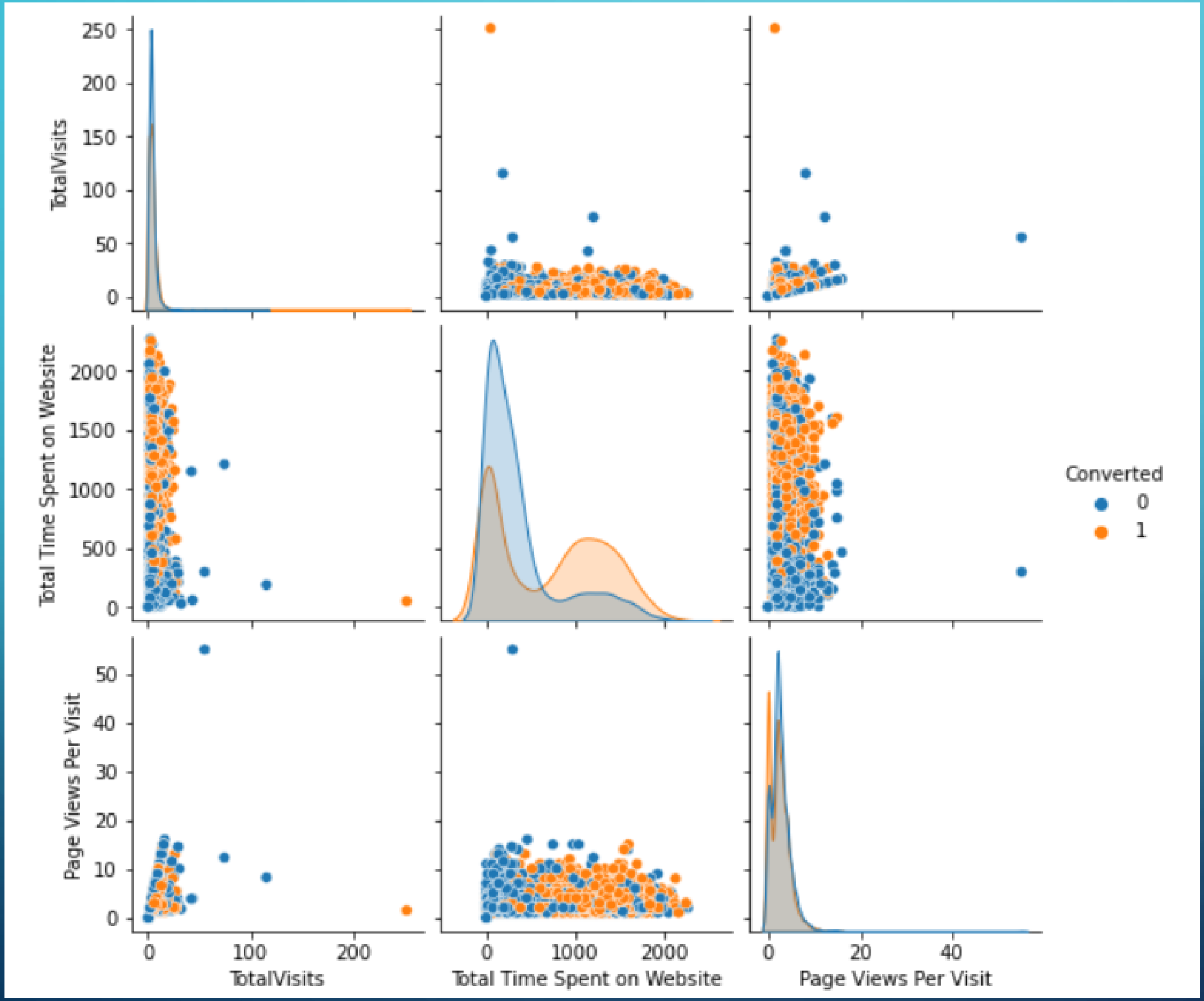
- Reading the data from csv file
- Data cleaning – Inspecting the columns and handling of null values and dropping the columns which have high null values in it.
- Checking the outliers in numerical columns



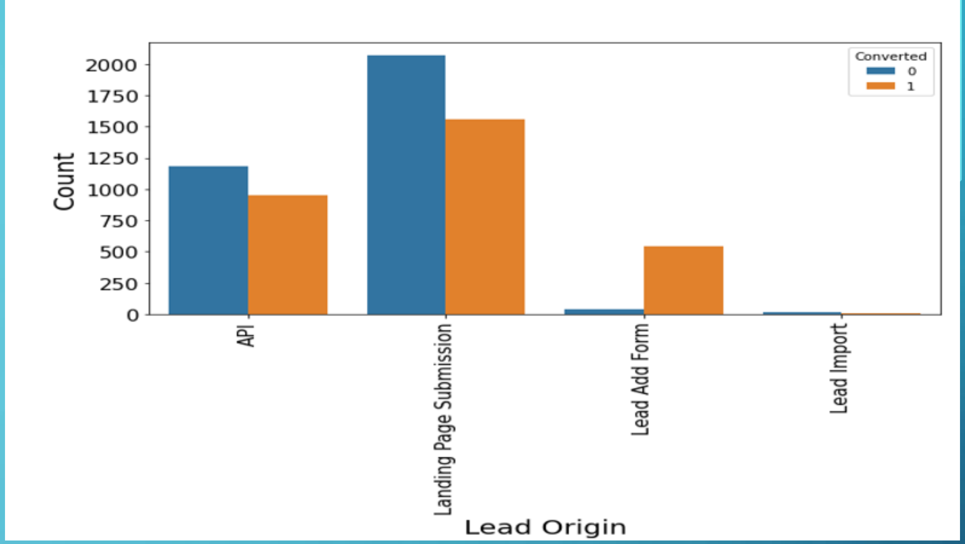




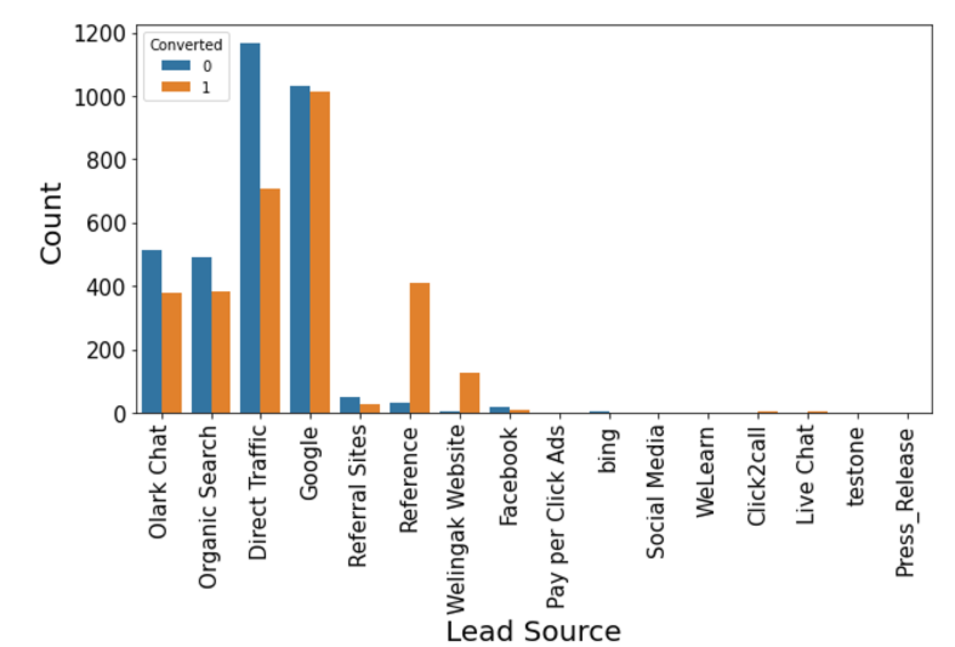
- Pair plot for numerical columns:



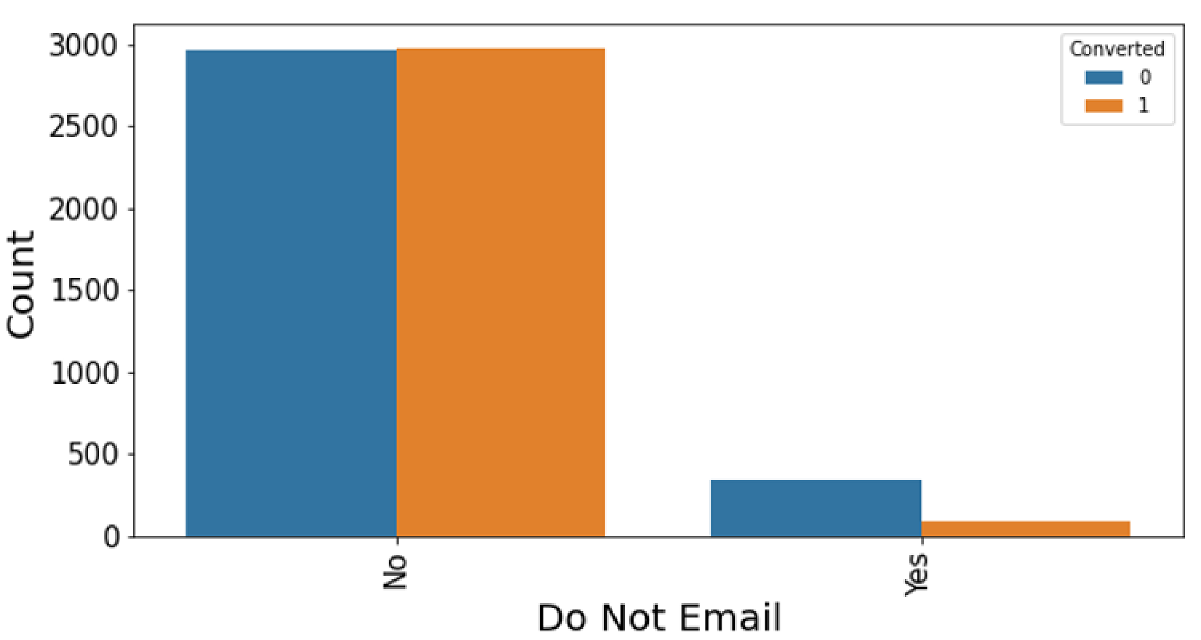
- EDA on categorical columns with Target column:
- Lead Origin: Most of the leads converted are Landing Page Submission followed by API & Lead Add form.



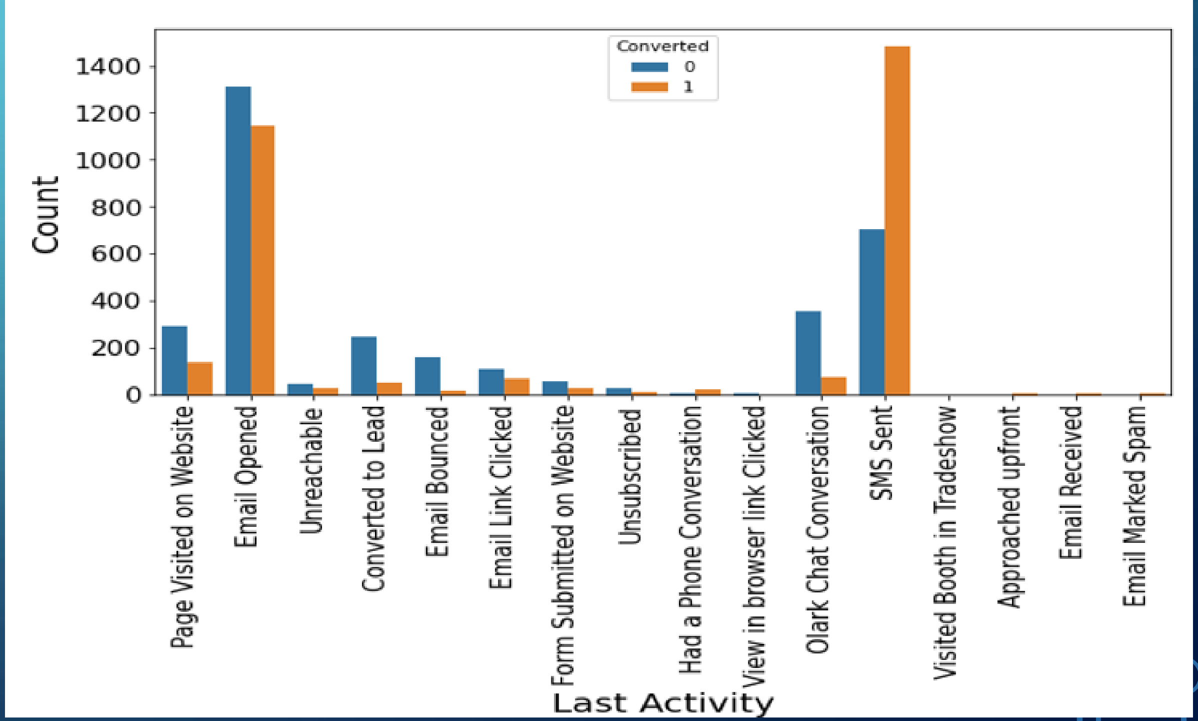
- Lead Source: The leads which are converted are majority from Google source and followed by Direct Traffic, Olark Chat, Reference & Organic Search.



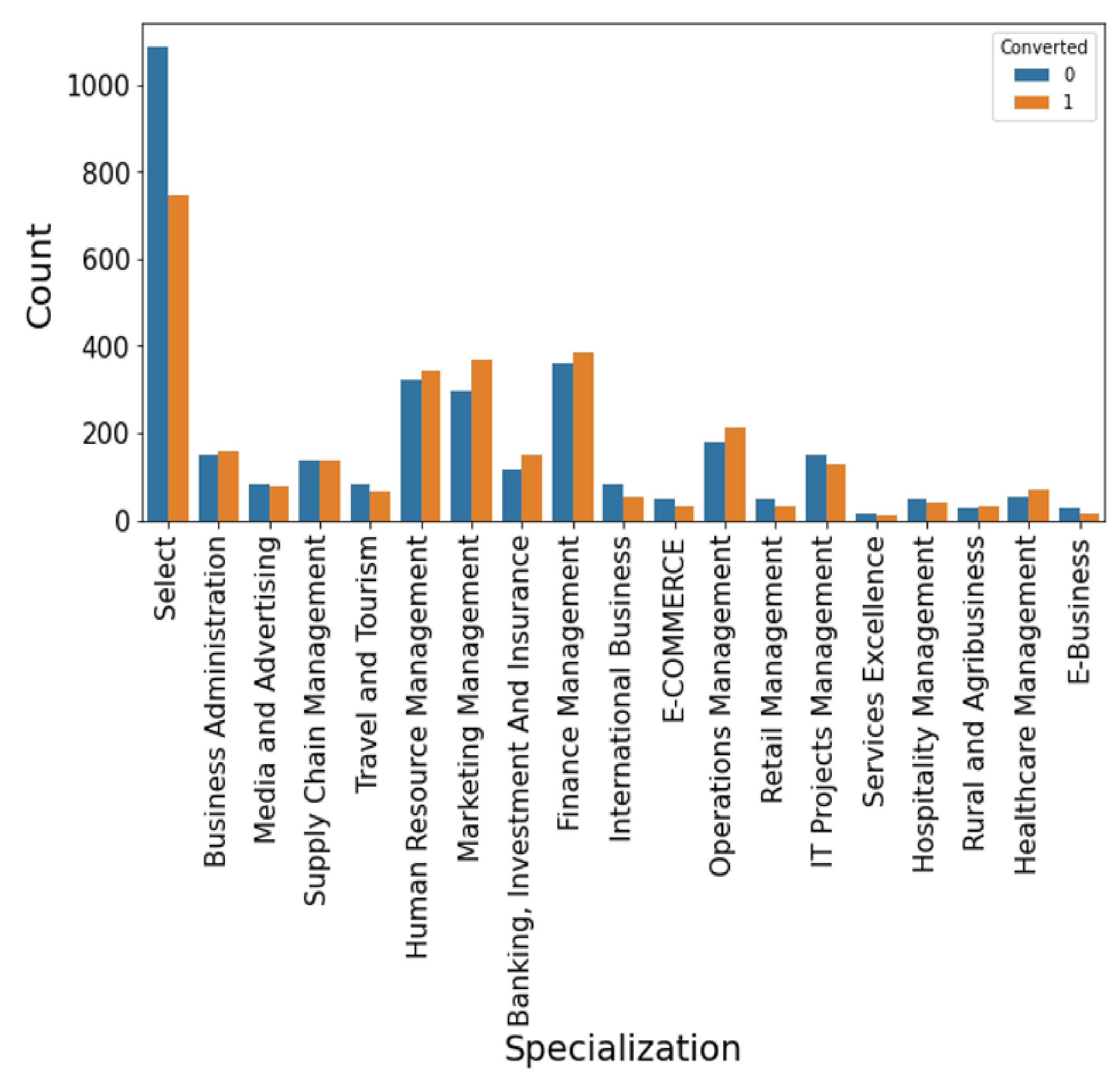
- Do Not Email: Most of the Converted leads are opted not to email about the course.



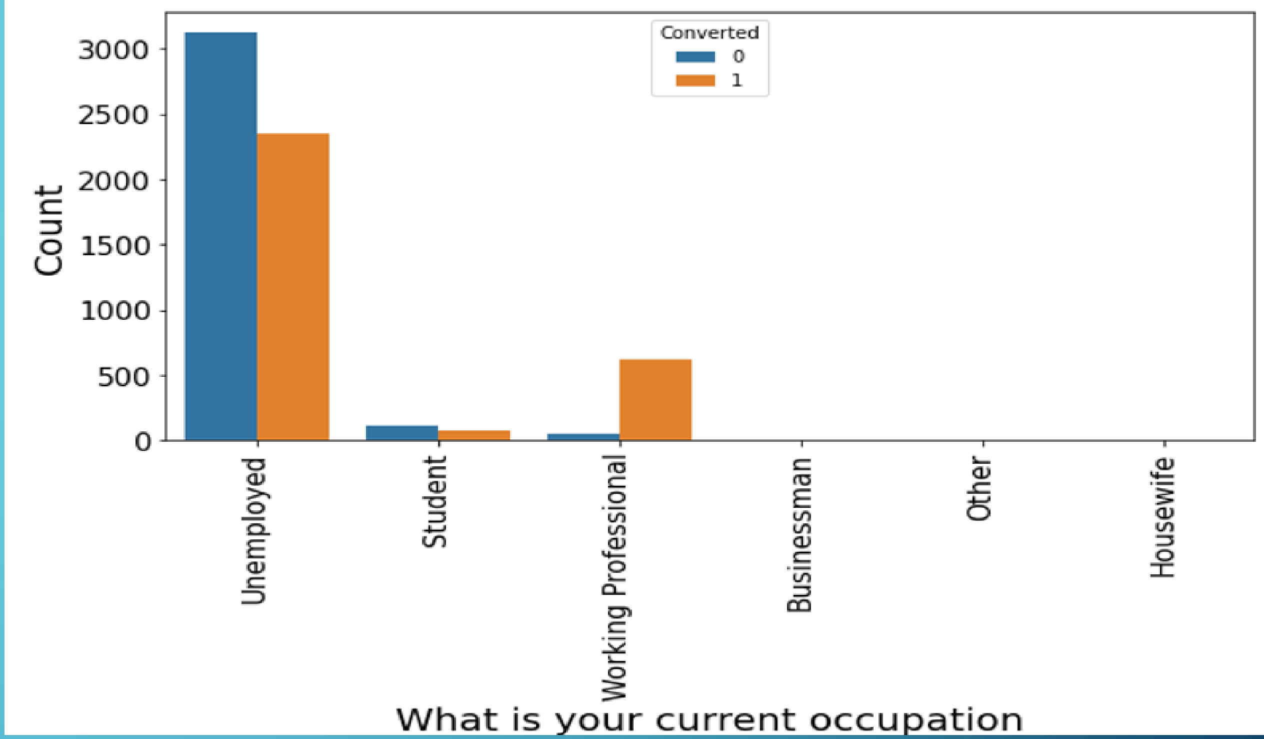
- Last Activity: SMS has shown to be a promising method for getting higher confirmed leads, emails also has high conversions.



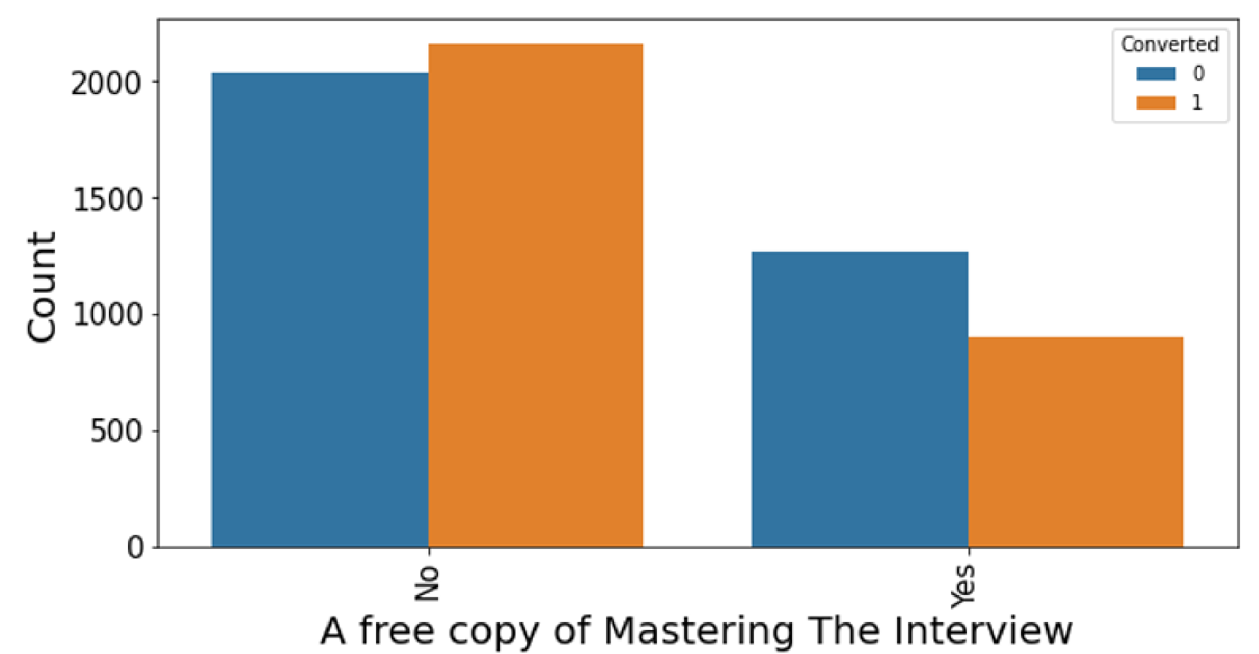
- Specialization: Most of the leads Have no information about specialization. Apart from that customers with marketing, human and finance management has high conversion rates. People with these specialization have shown promising leads.



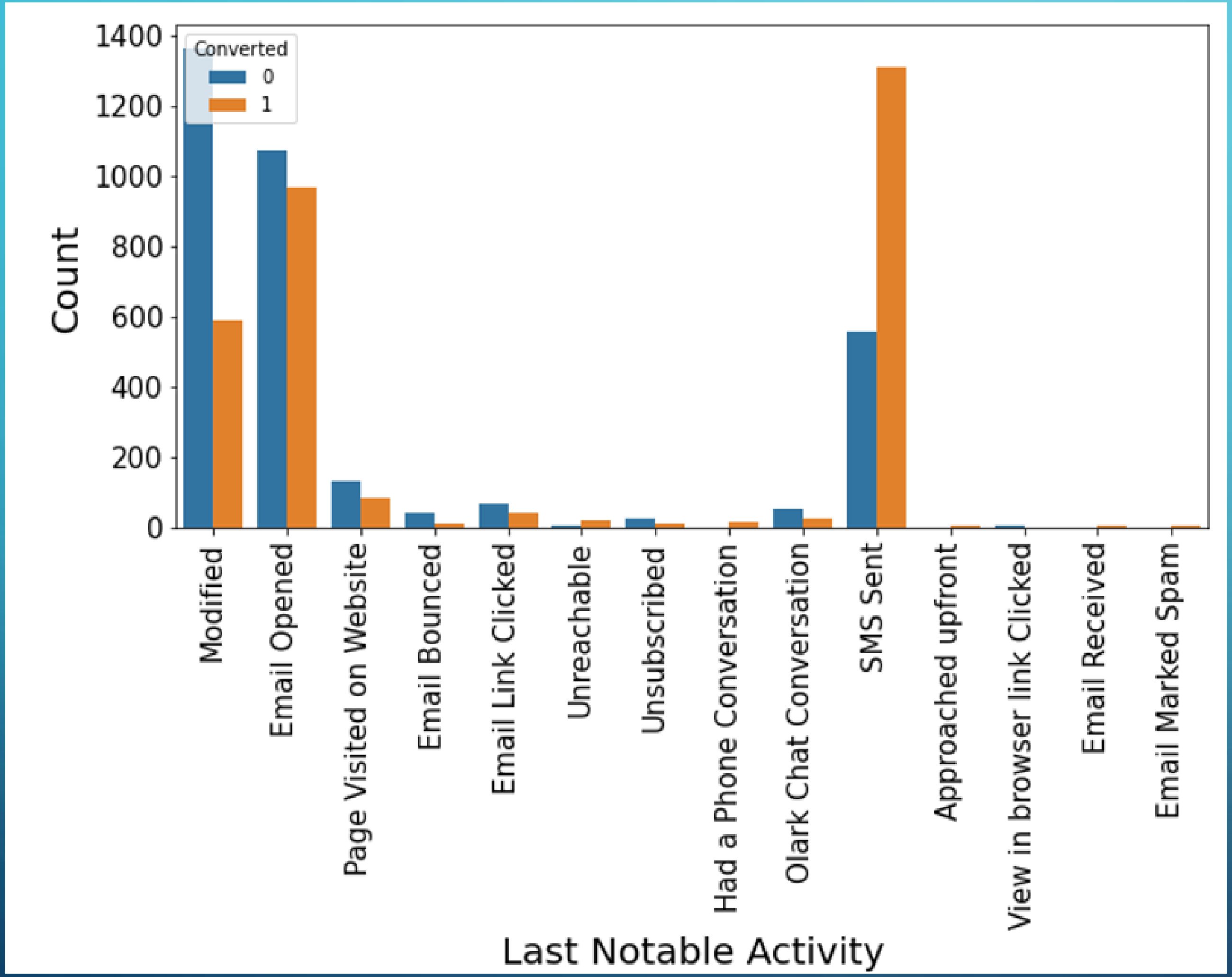
- What is your Current occupation: Customers who are Unemployed and working professionals has shown Promising leads.



- A free copy of Mastering the Interview: Customer who prefers less copies of interviews are having high chance of becoming leads.



- Last Notable Activity: Most leads are converted with messages followed by email and modified as their last notable activity.



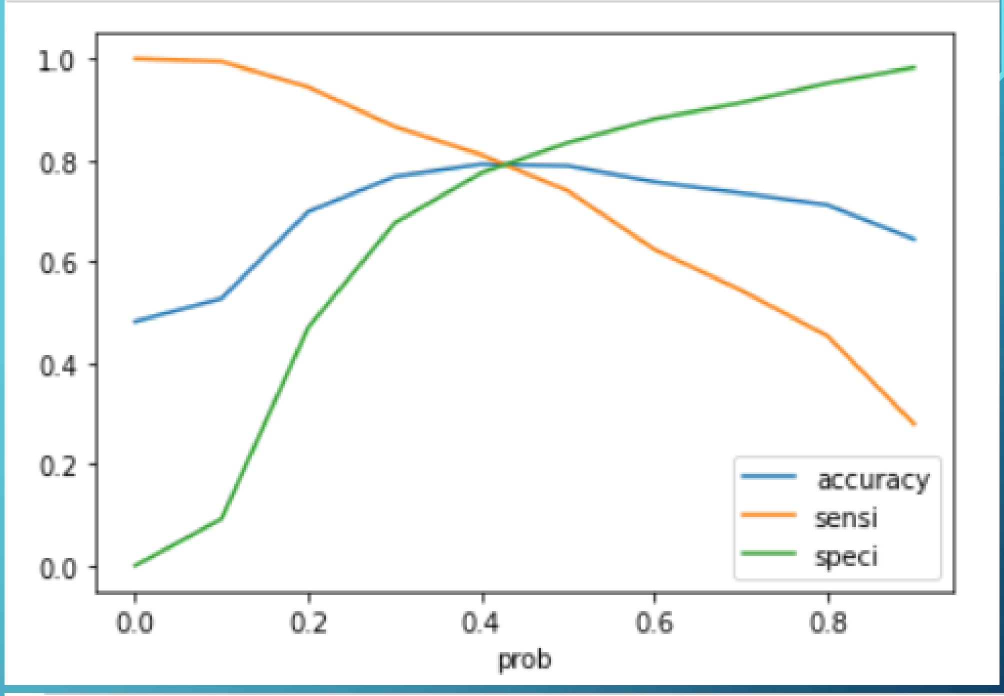
MODEL BUILDING

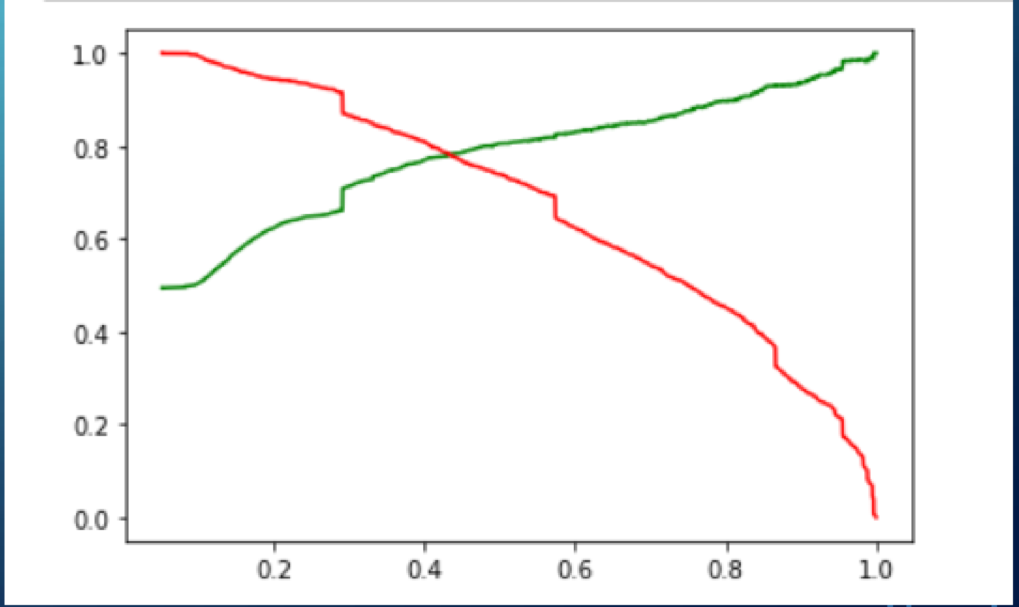
- Splitting the dataset into train and test dataset
- Scaling the numerical variables in train dataset
- Use RFE to eliminate the less relevant variables
- Build the first Logistic regression model
- Eliminate the variables based on high p-value and high VIF values
- Repeat the processing of building the model and eliminating the less relevant variable till all the variables in model are having p-value and VIF value less than threshold frequency
- Model Evaluation – confusion matrix, accuracy, sensitivity & specificity
- Predict on test data – confusion matrix, accuracy, sensitivity & specificity

MODEL EVALUATION ON TRAIN DATASET

- Accuracy: 79.1%
- Sensitivity: 79.3%
- Specificity: 78.8%
- Precision: 77.7%
- Recall: 79.3%
- Confusion Matrix:

1823	489
444	1705





MODEL EVALUATION ON TEST DATASET

- Accuracy: 78.7%
- Sensitivity: 76.7%
- Specificity: 80.4%
- Precision: 78.3%
- Recall: 76.7%
- Confusion Matrix:

801	**195**
213	703

SUMMARY

- There are a lot of leads generated in the initial stage (top) but only a few of them come out as paying customers from the bottom. In the middle stage, you need to nurture the potential leads well (i.e. educating the leads about the product, constantly communicating etc.) in order to get a higher lead conversion. First, sort out the best prospects from the leads you have generated. 'TotalVisits' , 'Total Time Spent on Website' which contribute most towards the probability of a lead getting converted. Then, You must keep a list of leads handy so that you can inform them about new courses, services, job offers and future higher studies. Monitor each lead carefully so that you can tailor the information you send to them. Carefully provide job offerings, information or courses that suits best according to the interest of the leads. A proper plan to chart the needs of each lead will go a long way to capture the leads as prospects. Focus on converted leads. Hold question-answer sessions with leads to extract the right information you need about them. Make further inquiries and appointments with the leads to determine their intention and mentality to join online courses.

CONCLUSION

- The logistic regression model shows accuracy of 79% which close to 80%.
- The threshold has been selected from accuracy, sensitivity, specificity measures and precision, recall curves.
- The model shows 77.5% sensitivity and 79% specificity.
- Overall this model is good and below are the variables that mattered most in converting the leads:

a. TotalVistis

b. Total Time Spent on Website

c. Lead Origin_Lead Add Form

d. Lead Source_Welingak Website & Lead Source_Olark Chat

e. Last Activity_Had a Phone Conversation & Last Activity_SMS Sent

f. Last Notable Activity_Unreachable

g. Do Not Email_Yes

h. What is Your Current Occupation_Student &

What is Your Current Occupation_Unemployed