Lead Score Case Study Using Logistic Regression

SUBMITTED BY :

- MONITHA MS
- ANJALI SAHU
- MRITYUNJAY SINGH

Content

Problem statement

Business Objective

Problem approach

EDA- Data Cleaning and Visualization

Model Building

Model Evaluation

Final Summary of Train and Test data

Conclusion

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. They have **process of form filling on their website after which the company that individual as a lead**.
- 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, this means 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

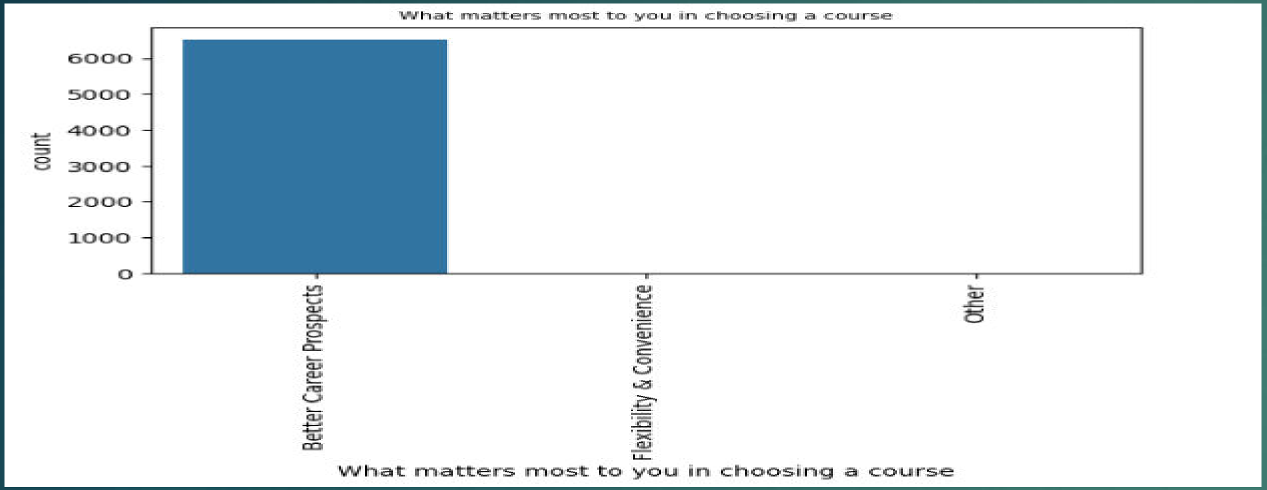
- X Education wants us to build a model to give every lead a lead score between 0 -100 . So that they can identify the Hot leads and increase their conversion rate as well.
- The CEO want to achieve a lead conversion rate of 80%.
- They want the model to be able to handle future constraints as well like Peak time actions required, how to utilize full man power and after achieving target what should be the approaches

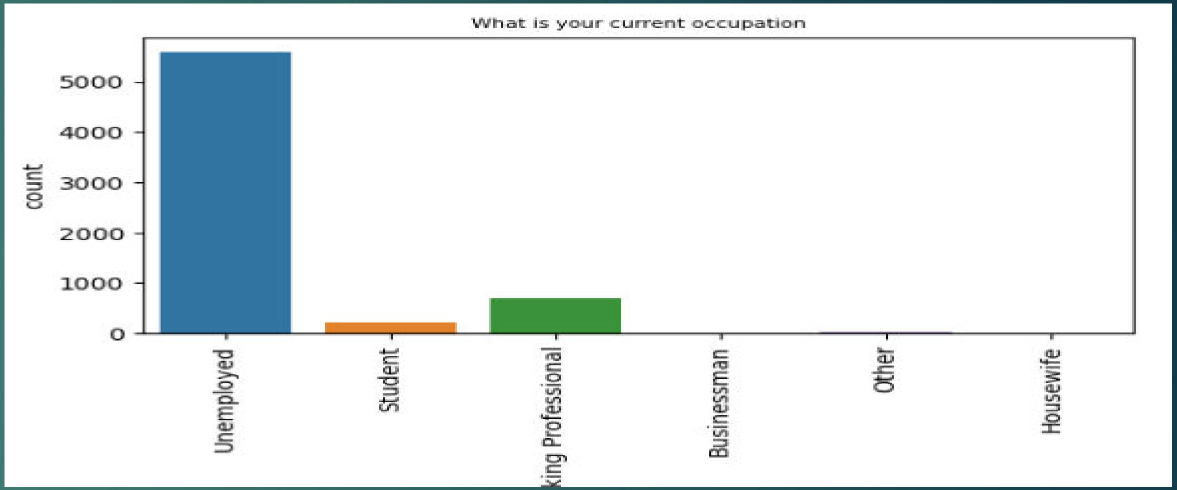
Problem Approach

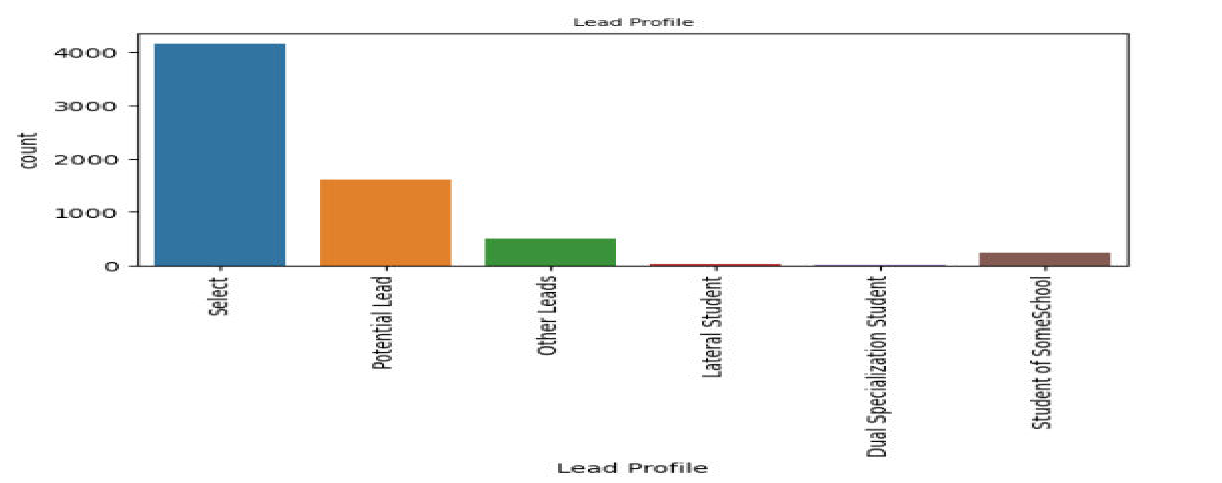
- **Read and Understand the Data**
- **Data Cleaning and Visualization-** EDA
- **Data Preparation** – create dummy variable, Test and Train Split, Scaling
- **Model Building** – RFE Selecting for 20 Variable, Manual Selection of features
- **Model Evaluation** – Lead Score prediction for train data, Matric trade off and cut off selection
- **Model Evaluation on test data** : Run the model on Test data and calculate the metrics.

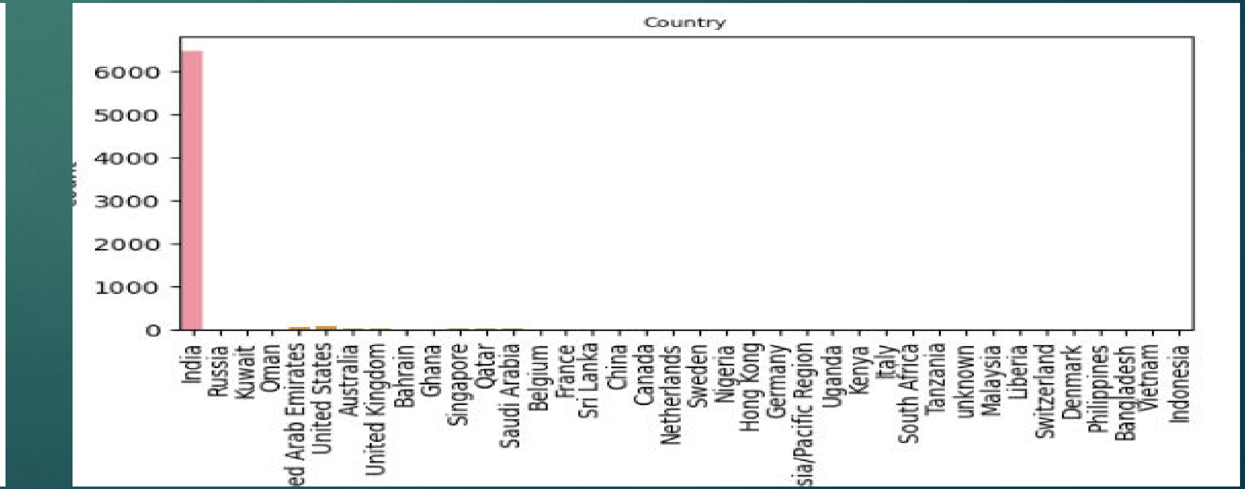
1. EDA – Data Cleaning and Visualization

❑ Below features has more than ~ 29% null values. They are dropped as they are imbalance features. One category count is very high.

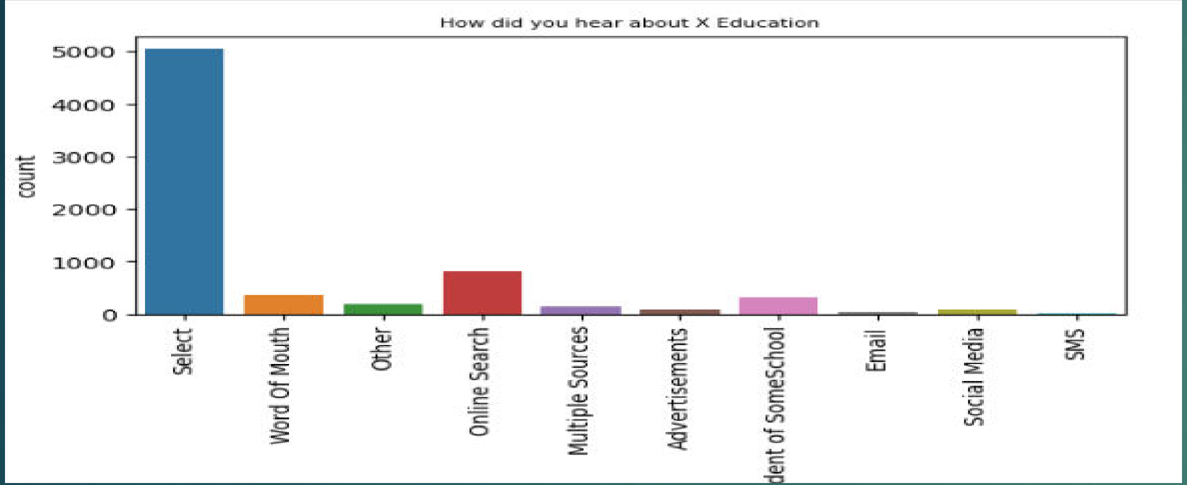


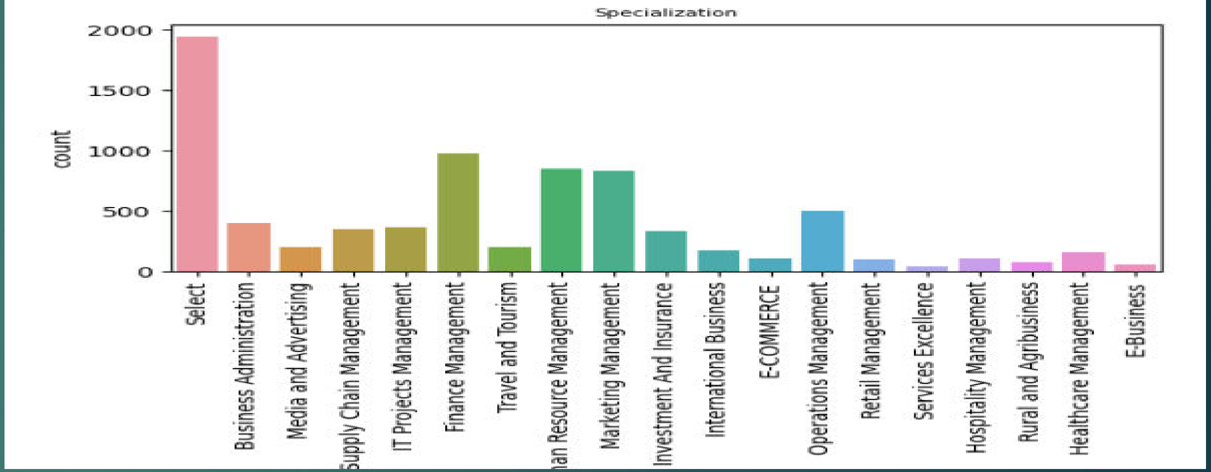


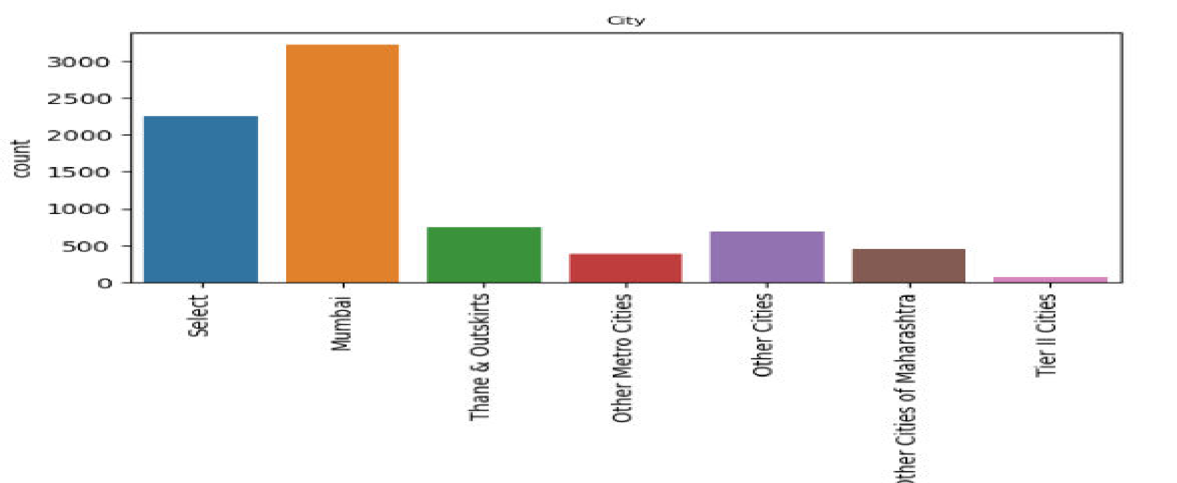




- ❑ Below features has more than ~ 15% null values. They are dropped as they are imbalance features.
- ❑ Specialization column did not seems to be a value add also the Select value is very hign. Together Null and Select is around 36% hence dropping

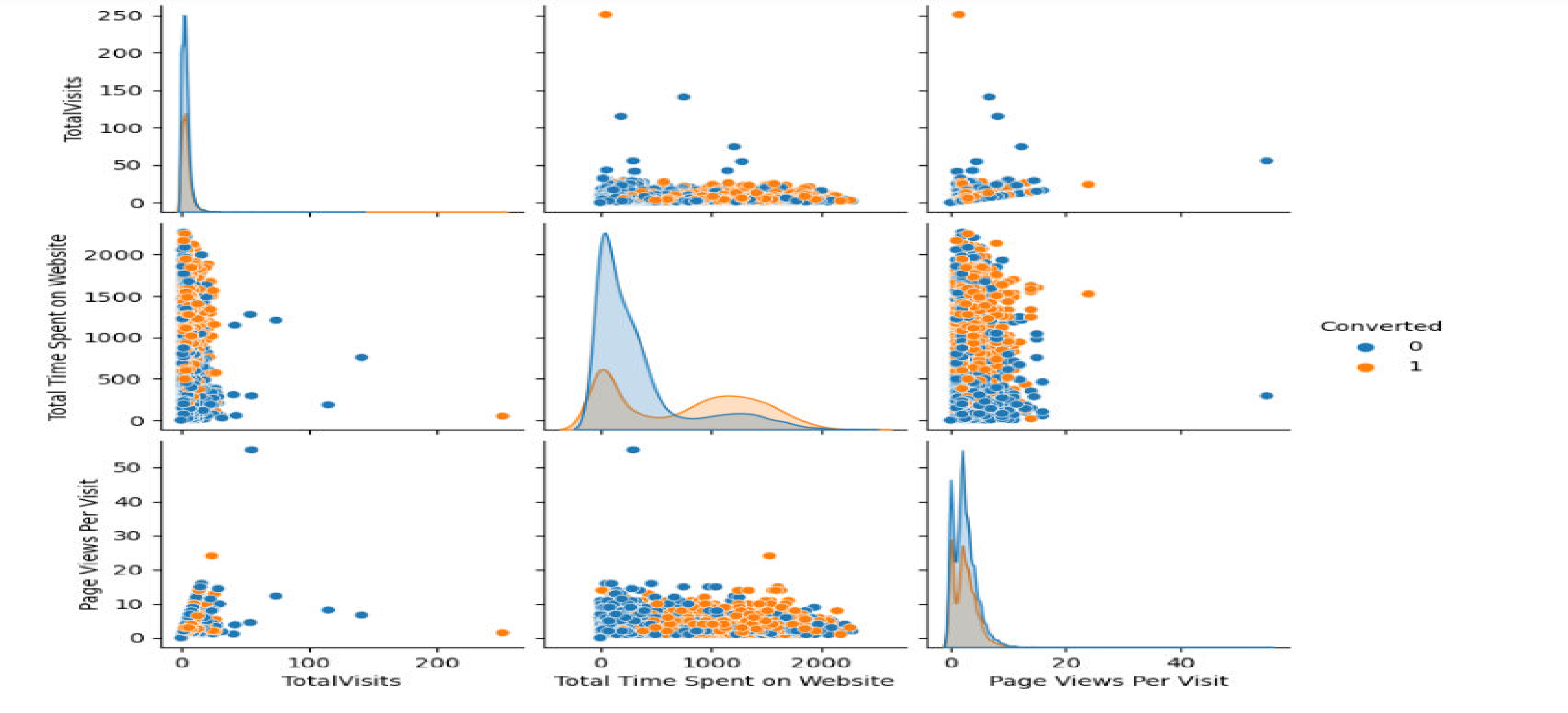




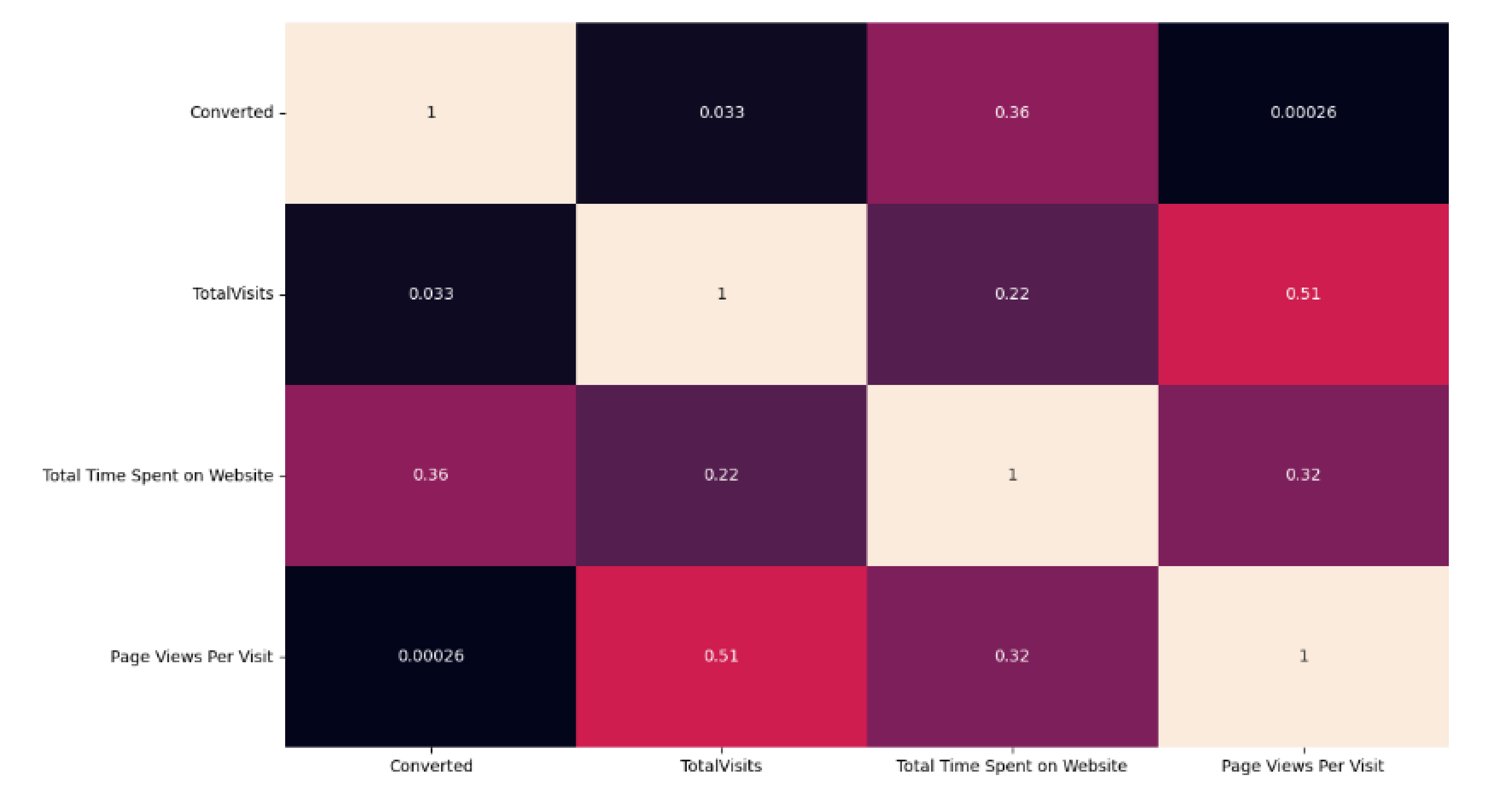


Data Visualization – Numerical Bi-Variate

- Observed that more the users spend time on the site and more the users visit the site has high chances of conversion.



- Observed that more the users spend time on the site and more the users visit the site has high chances of conversion.



Model Building:

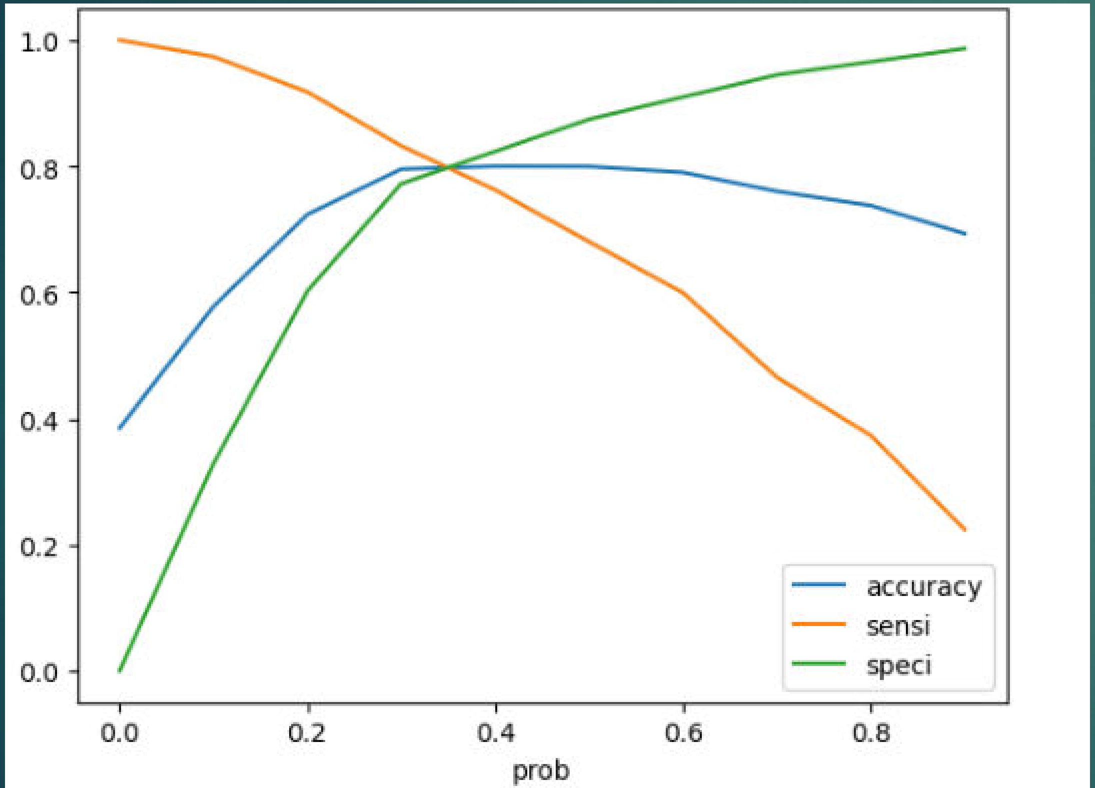
- ❑ 20 Features selected initially with RFE
- ❑ Build models one by one and finally reached model with 16 features where all features have pValue less that 0.05 and VIF less than 5.

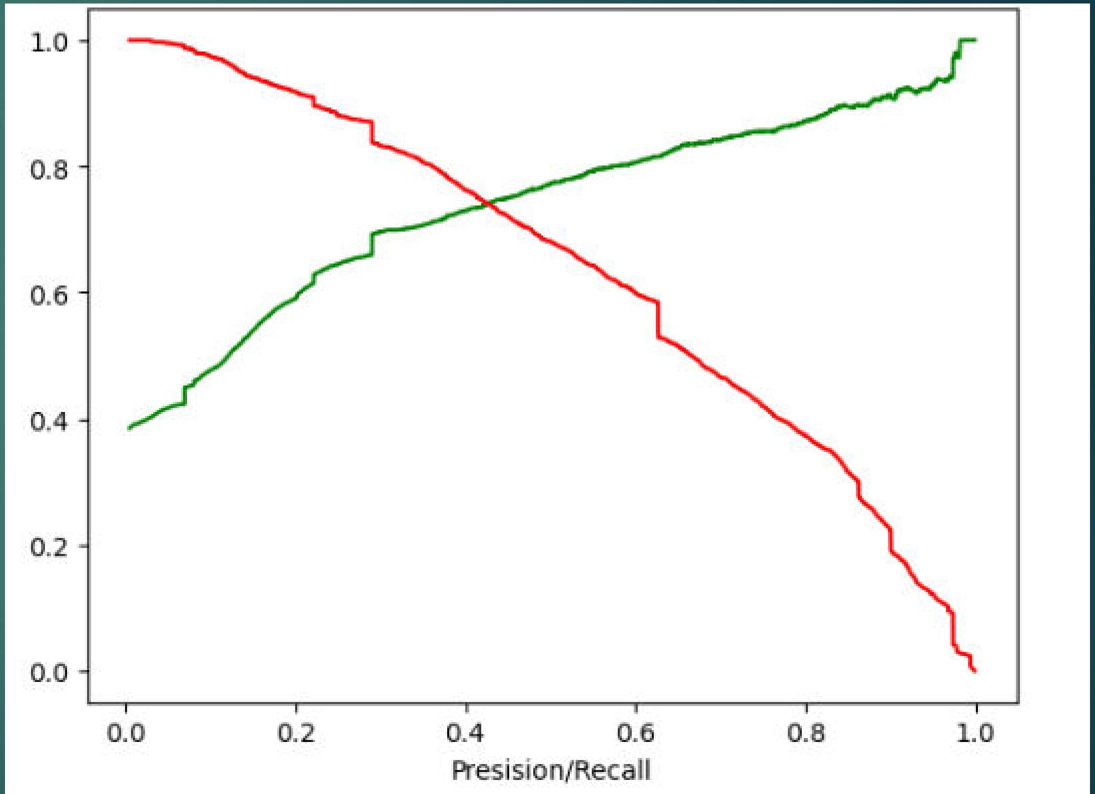
- ❑ Final Features selected:

- ❑ 'Do Not Email'
- ❑ 'TotalVisits'
- ❑ 'Total Time Spent on Website'
- ❑ 'Lead Origin_Lead Import'
- ❑ 'Lead Source_Olark Chat'
- ❑ 'Lead Source_Reference'
- ❑ 'Lead Source_Welingak Website'
- ❑ 'Last Activity_Converted to Lead'
- ❑ 'Last Activity_Email Bounced'
- ❑ 'Last Activity_Had a Phone Conversation'
- ❑ 'Last Activity_Olark Chat Conversation'
- ❑ 'Last Notable Activity_Email Link Clicked'
- ❑ 'Last Notable Activity_Email Opened'
- ❑ 'Last Notable Activity_Modified'
- ❑ 'Last Notable Activity_Olark Chat Conversation',
- ❑ 'Last Notable Activity_Page Visited on Website'

Model Evaluation:

- ❑ Precision Recall trade off is 0.42 but it gets Sensitivity is only ~74%
- ❑ Hence taking 0.33 from the Accuracy and sensitivity trade off. Recall of 82% and Accuracy ~80%





Final Summary of Test and Train

- Final Cut off probability value 3.3 .

Train data Summary:

- Accuracy   : 0.7995591245473154
- Sensitivity: 0.8205233033524121
- Specificity: 0.7810499359795134
- Precision: 0.7012578616352201
- Recall : 0.8205233033524121

Test Data Summary:

- Accuracy   : 0.7818582445831803
- Sensitivity: 0.7906976744186046
- Specificity: 0.7768166089965398
- Precision: 0.6689478186484175
- Recall: 0.7906976744186046

Conclusion

- ❑ Lead score conversion is very high for Lead Source **Welingak Website**, followed by **Reference** and **Olark Chat**.
- ❑ Leads who **spent more time on website**, more likely to convert.
- ❑ Leads who have high **TotalVisits** are more likely to convert.
- ❑ Leads who have selected **Do Not Email** are less likely to be converted.