Lead Scoring Case Study

Saurabh Saxena

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 gets a lot of leads, its lead conversion rate is very poor. For example, 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.

Data

Data

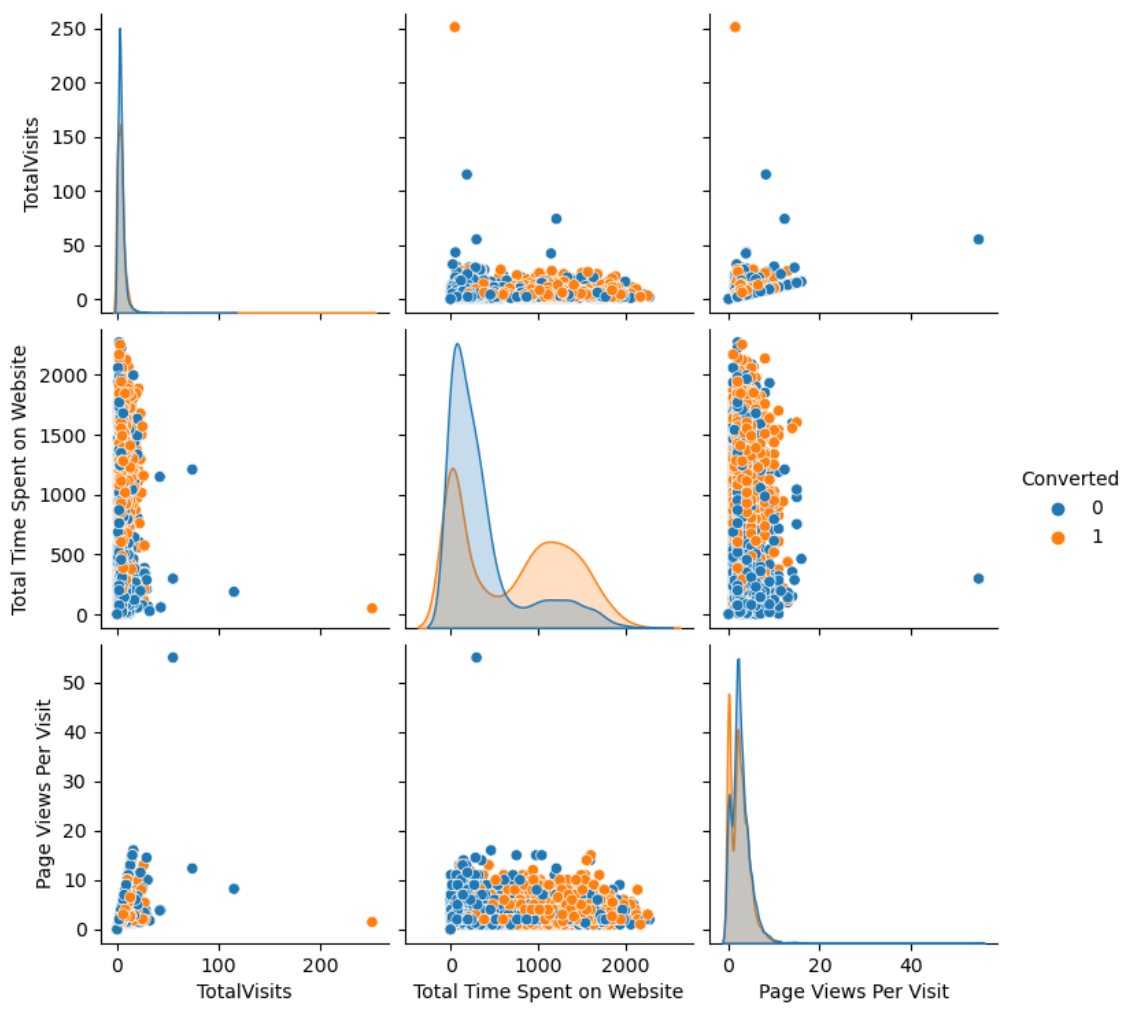
You have been provided with a leads dataset from the past with around 9000 data points. This dataset consists of various attributes such as Lead Source, Total Time Spent on Website, Total Visits, Last Activity, etc. which may or may not be useful in ultimately deciding whether a lead will be converted or not. The target variable, in this case, is the column ‘Converted’ which tells whether a past lead was converted or not wherein 1 means it was converted and 0 means it wasn’t converted. You can learn more about the dataset from the data dictionary provided in the zip folder at the end of the page. Another thing that you also need to check out are the levels present in the categorical variables. Many of the categorical variables have a level called 'Select' which needs to be handled because it is as good as a null value.

Goals of the Case Study

quite a few goals for this case study:

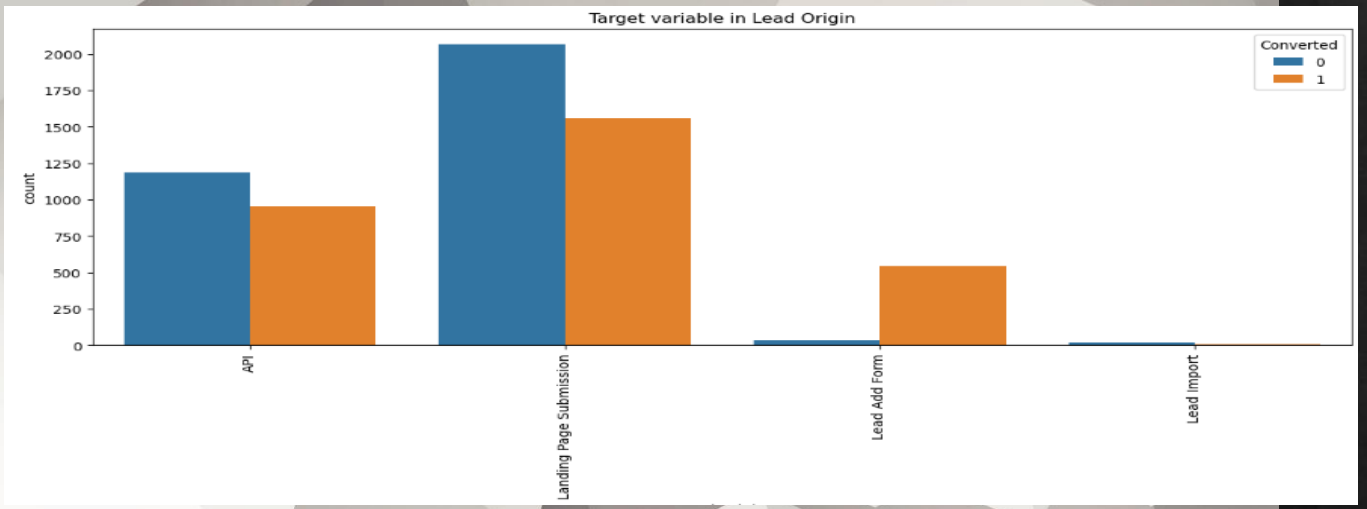
ogistic regression model to assign a lead score between 0 and 100 to each of the leads which can be used by the o target potential leads. A higher score would mean that the lead is hot, i.e. is most likely to convert whereas a lower ld mean that the lead is cold and will mostly not get converted.

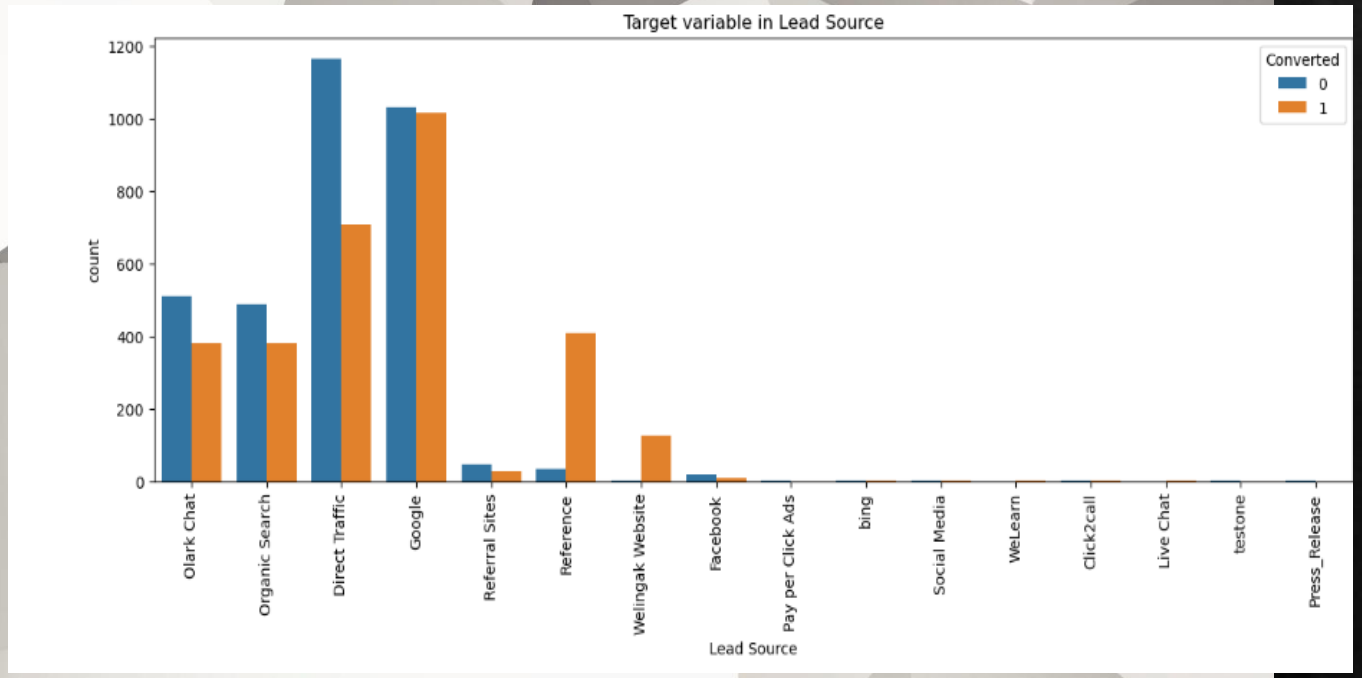
e some more problems presented by the company which your model should be able to adjust to if the company's nt changes in the future so you will need to handle these as well. These problems are provided in a separate doc file. t based on the logistic regression model you got in the first step. Also, make sure you include this in your final PPT 'll make recommendations.



Actions performed

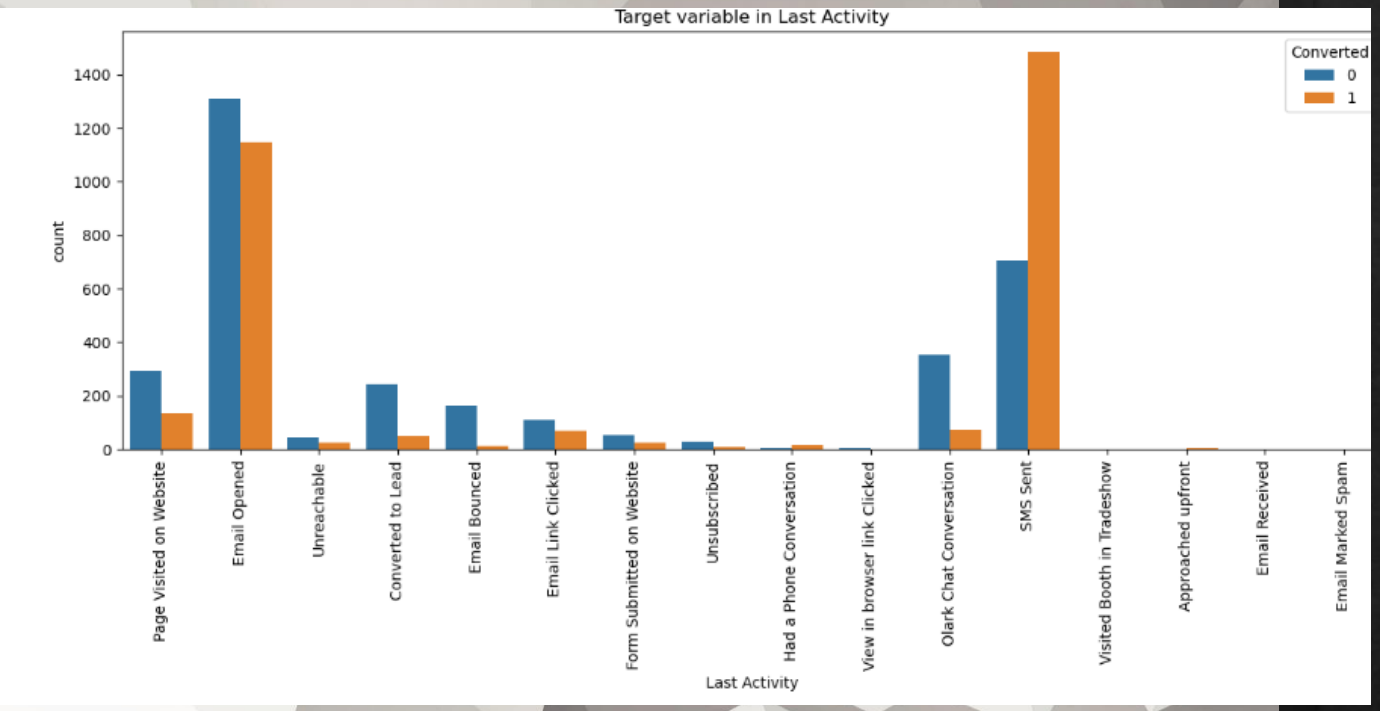
- Importing of data
- Data preparation
- EDA
- Logistic Regresssion
  - Dummy variable creation
  - Split data in test and training data
  - Scaling
  - Checking correlation
  - Model Building
    - RFE
    - Manual feature rejection (Pvalue and VIF)
  - Model evaluation
    - Accuracy, sensitivity, Selctivity
    - ROC curve

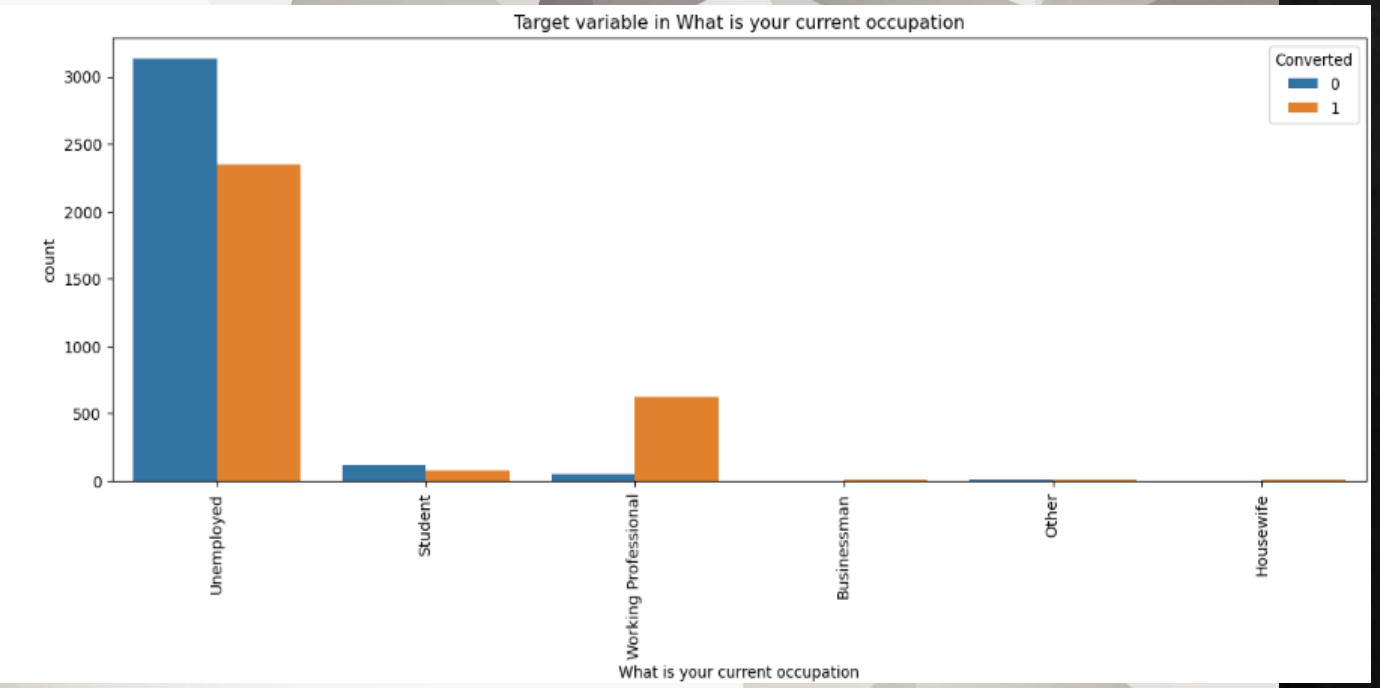




EDA

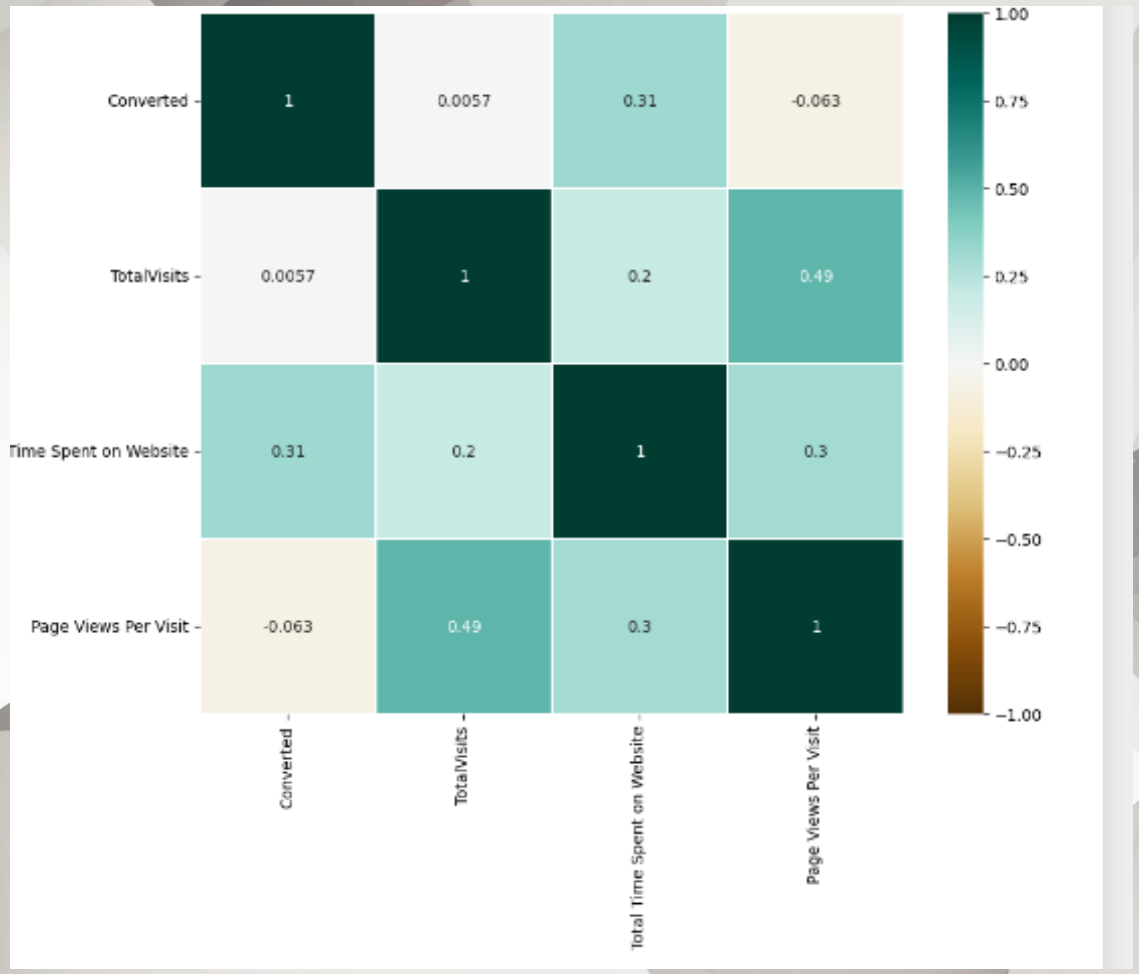
- Those leads from lead add form have higher conversions
- Most traffic is generated from Google and direct website traffic
- Those coming with reference have higher conversion.





EDA

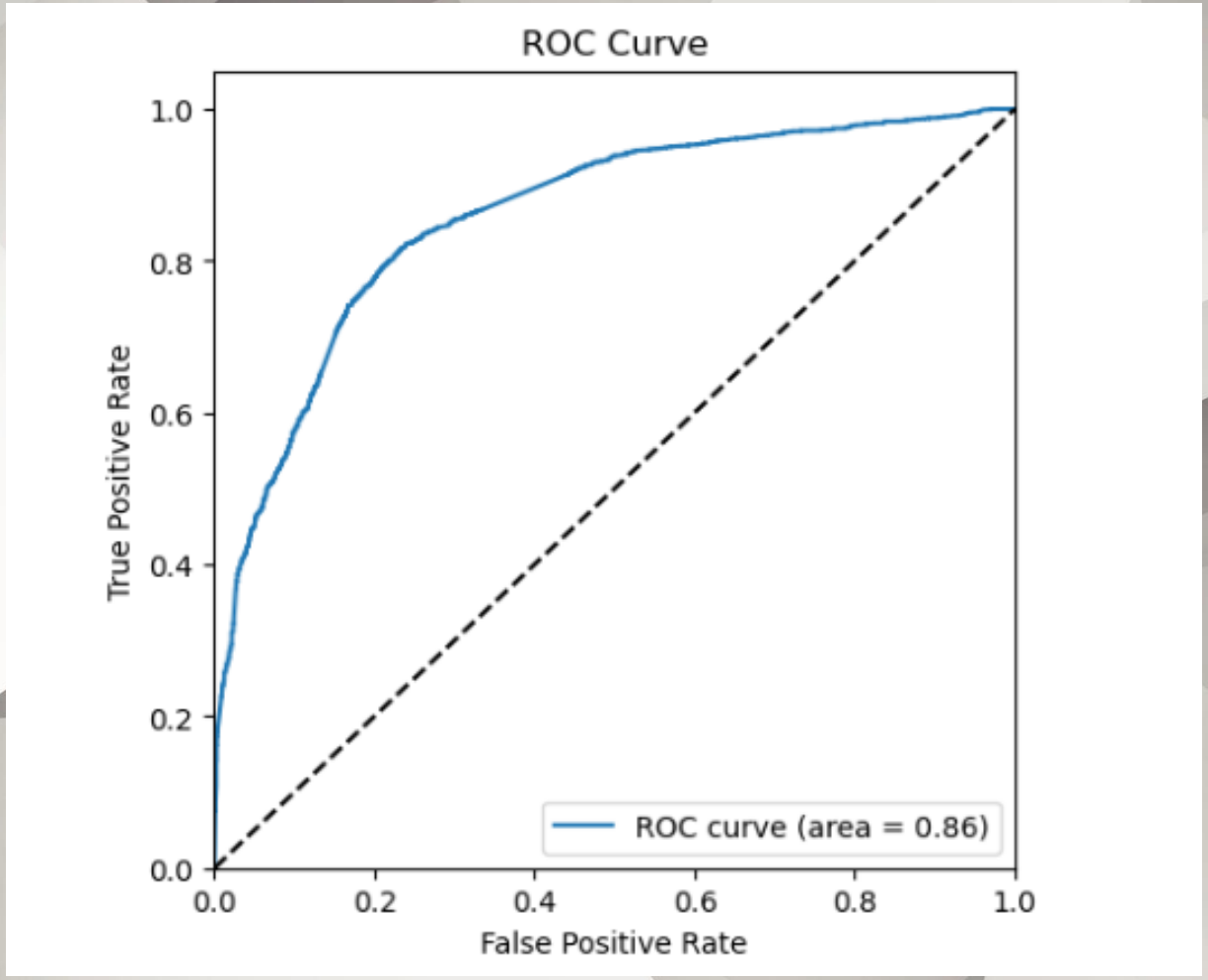
- SMS has high conversion rate
- Most traffic is generated from Google and direct website traffic
- Those coming with reference have higher conversion.
- Working professionals have high conversion



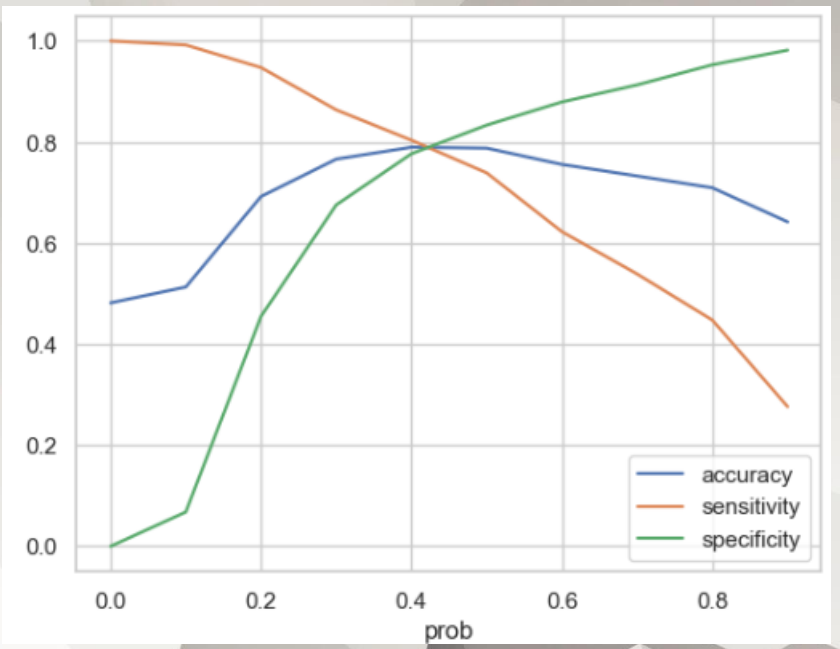
Correlation

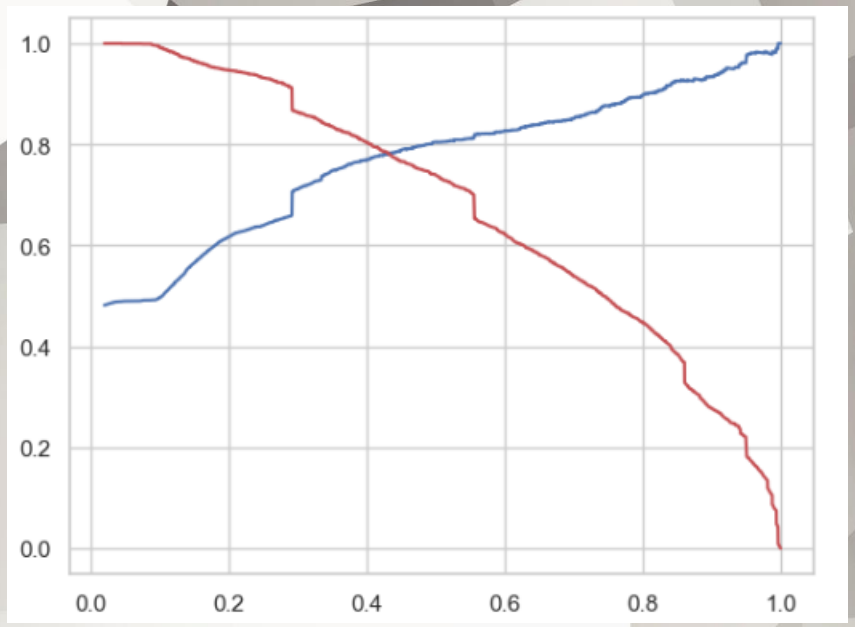
- There is little to no correlation between the variables

Regression Model



- The model is considered good as it has high area under ROC curve.





Model Evaluation

- The optimum value for probability cutoff is 0.43

- Metrics – Train Data
  - Accuracy – 79%
  - Sensitivity – 77%
  - Specificity – 98%
- Metrics – Test Data
  - Accuracy – 79%
  - Sensitivity – 77%
  - Specificity – 80%