LEAD SCORE CASE STUDY

BY: SAI KISHAN K

PROBLEM STATEMENT

- A X Education company sells on courses online in their website. Many users use to access the website to browse and learn some courses. The website includes a form to be filled by the user after which the company makes that user as lead
- The lead conversion rate of the company for this is 30%. o make this process more efficient, the company wishes to identify the most potential leads, also known as ‘Hot Leads’.
- Upon successfully identifying the 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

- 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.
- They also want the model to be able to handle future constraints and some other necessary strategies to get from the model like how to utilize full manpower and how to approach the user

SOLUTION APPROACH

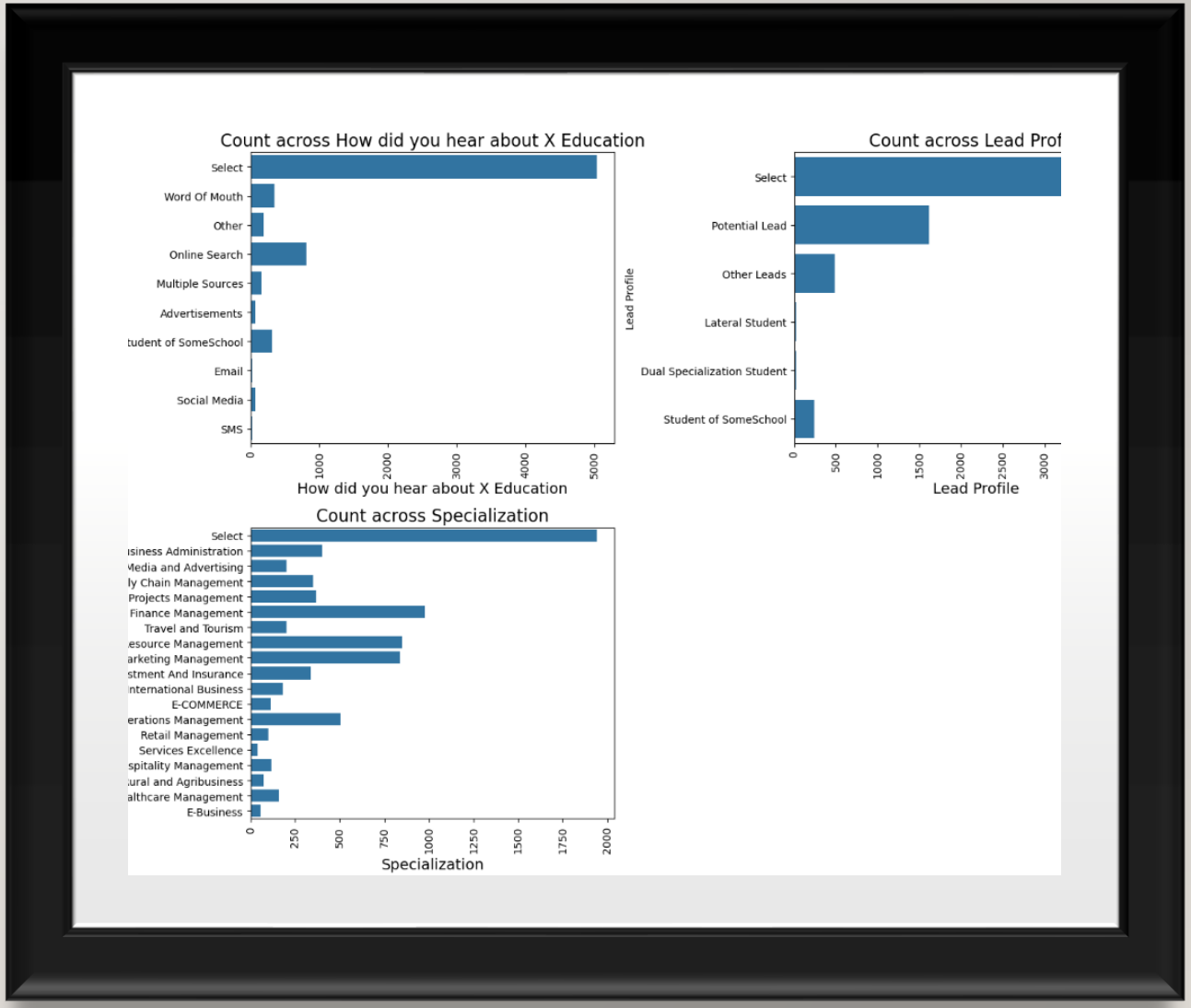
- Importing the Dataset
- Data Cleansing
- Visualising Necessary features
- Creating Dummy Variables
- Test Train Splitting
- Model Building
- Model Evaluation
- Cut off Optimisation
- Test Set Predictions

DATA CLEANSING

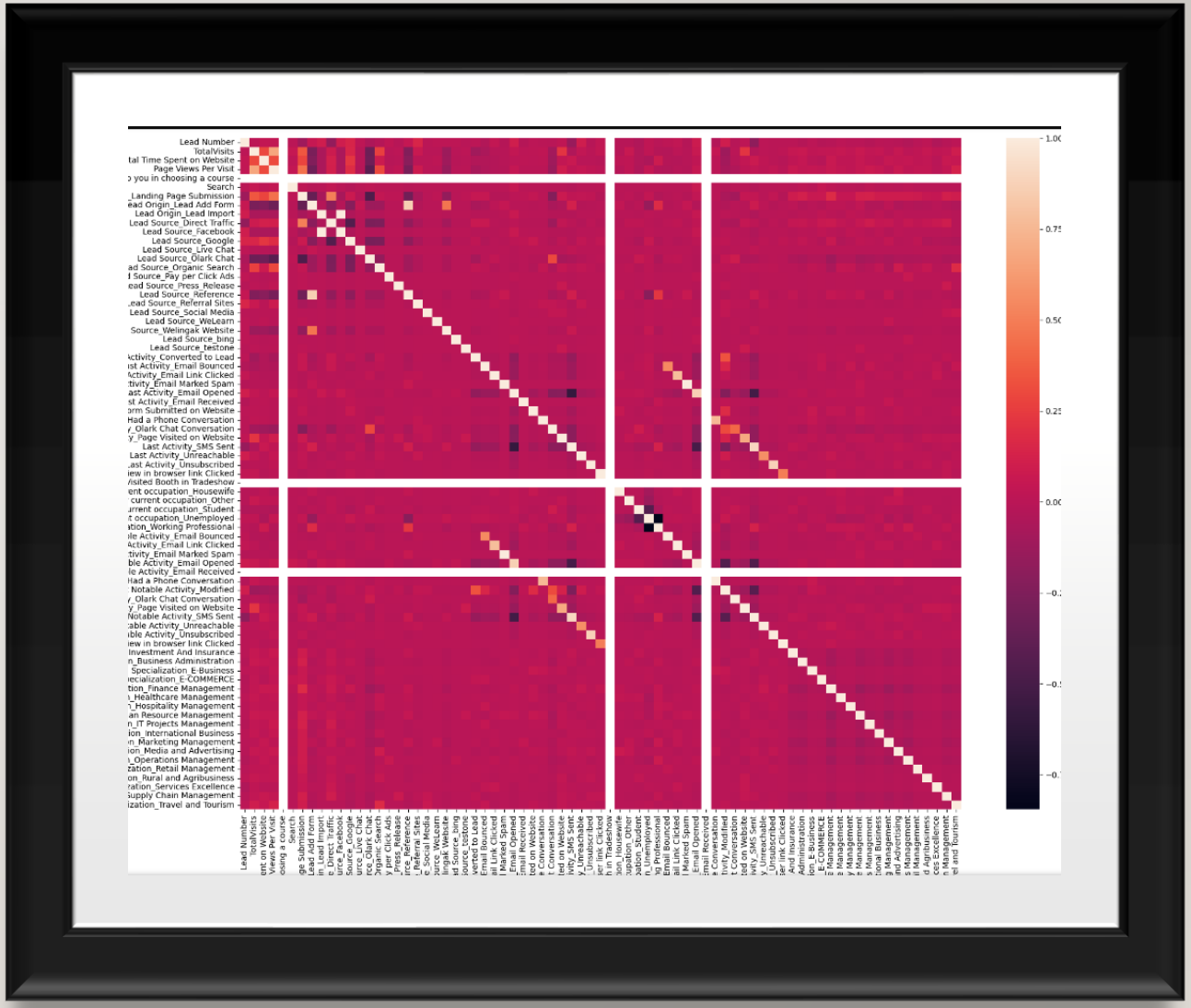
- Total number of columns = 9000. Therefore, Dropped columns that are of no use for our analysis and having more than 3000 null values.
- Dropped columns that have single dominant value.(e.g. India for Country column etc.).
- Dropped columns that have more values as ‘Select’ which means the user haven’t selected any option.
- Dropped columns that directly means ‘NO’ (‘Do Not Call’, ‘Search’,’ Magazine’ etc)
- After cleaning the data, we got a 69% of retention rate.

VISUALIZING THE DATA

- Visualized dataset columns that have large amount of ‘Select values



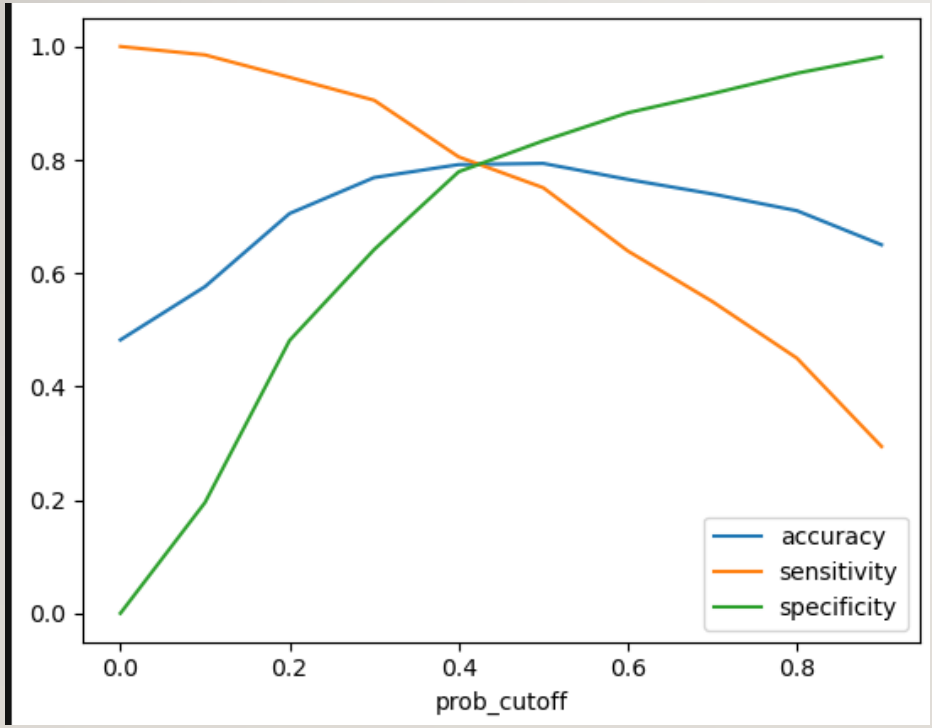
- Viewing correlation using heatmap between the columns in the dataset.

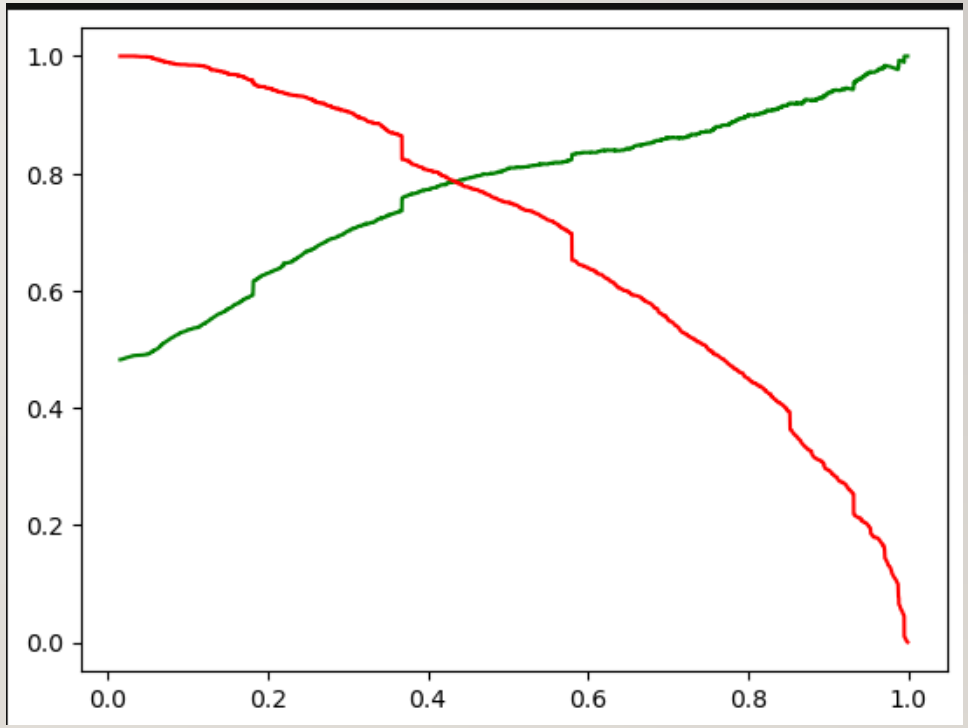


MODEL BUILDING

- Split the data into Test and Train sets to build the model.
- Chose 70-30 ratio for performing regression in the train test split.
- Used and ran RFE for Feature selection with 15 variable as output.
- Removed variables that have p-value greater than 0.05 and VIF value greater than 5.
- Found out optimal cut off to be as 0.42.
- With optimal cut off as 0.42 we have accuracy of 79%.

MODEL EVALUATION (ROC CURVE)





OBSERVATIONS

- Train Data
  - Accuracy – 79.39%
  - Sensitivity – 79.59%
  - Specificity – 79.2%
- Test Data
  - Accuracy – 76.52%
  - Sensitivity – 85.34%
  - Specificity – 68.44%

CONCLUSION

- From the above we can conclude that priority of the dataset is as follows:
  1. The total time spend on the website
  2. Total number of visits
  3. When the lead source is Google
  4. When the last activity is SMS Olark chat conversation
  5. When the lead origin is lead add format
- Based on the above findings the X education can conclude that they have a very chance to get all the potential buyers to change their mind and buy thier courses