LEAD SCORE CASE STUDY

GROUP MEMBERS:

SHYAM NARAYAN TRIPATHY

SATYAPRIYA MAHAPATRA

SHILPA SHARMA

Problem Statement & Business Objective

X Education sells online courses to industry professionals.

- X Education gets a lot of leads, its lead conversion rate is very poor which is around 38%
- 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 to know most promising leads.
- For that they want to build a Model which identifies the hot leads.
- Deployment of the model for the future use.

****** Note: To build the model X education has provided the dataset and we have used the same data for our analysis and building the model.***

Steps we followed

- Read and understand the data
- EDA & Clean the data
- Prepare data for Model Building
- Model Building
- Model Evaluation
- Making Predictions on the Test Set
- Conclusion

Read & Understand the Data

- X education has been 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.
- To learn more about the dataset a data dictionary has been provided

EDA & Data Cleaning

- There were lot of columns which have high number of missing values. Since, there were 9000+ datapoints in our data set and those columns were not useful. So, we eliminated the columns having greater than 3000 missing values .

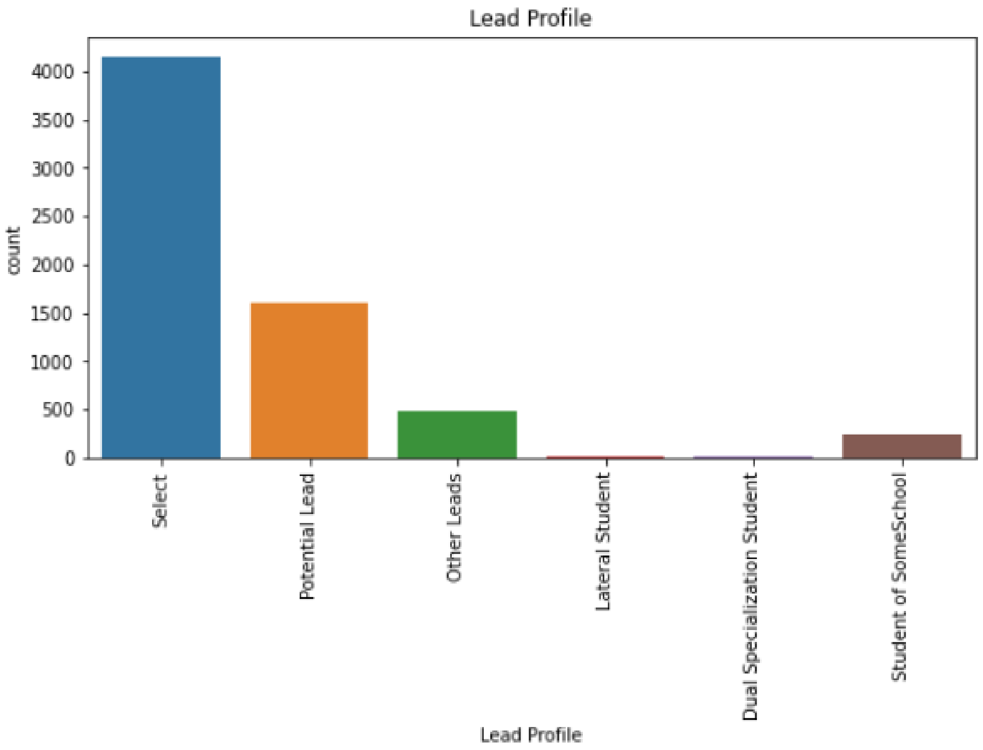
- There were few columns in which there is a level called **'Select'** which basically means that the student had not selected the option for that particular column which is why it shows 'Select'. These values are as good as missing values and hence we need to identify the value counts of the level 'Select' in all the columns that it is present.

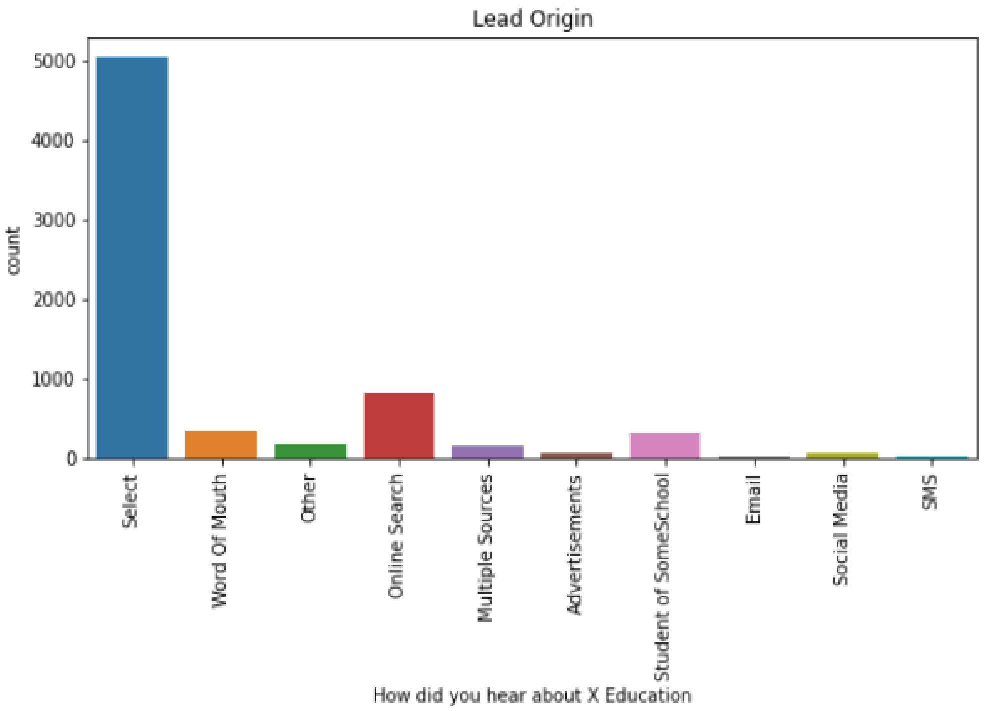
- The levels "Lead Profile" and "How did you hear about X Education" have a lot of rows which have the value “Select” which is of no use to the analysis so we dropped them.

- Also notice that when we got the value counts of all the columns, there were a few columns in which only one value was majorly present for all the data points . This is short of data imbalance. So we dropped these columns as they won't help with our analysis.

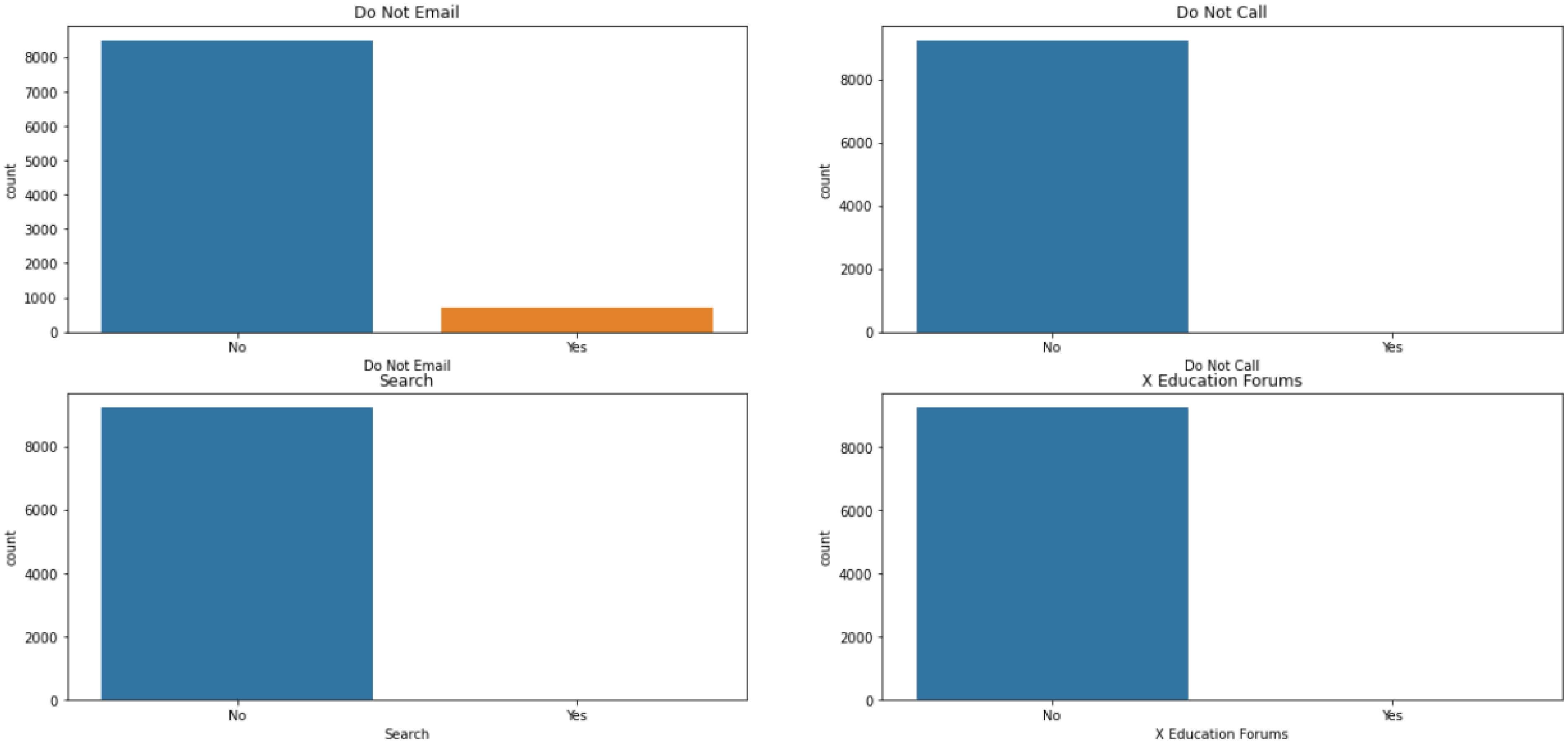
- So, after all cleaning operation we got around **69%** of the rows which seems good number for our analysis.

EDA & Data Cleaning



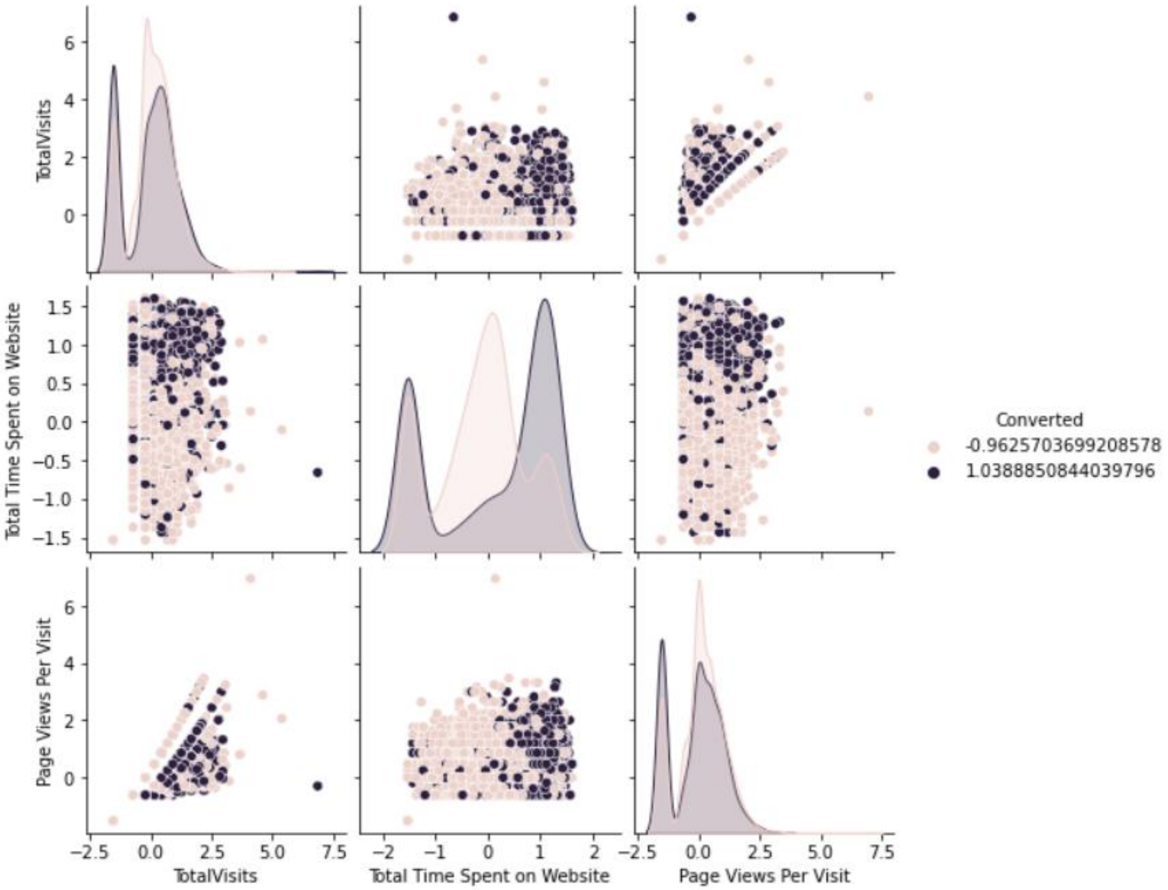


EDA & Data Cleaning



Prepare Data for Model Building

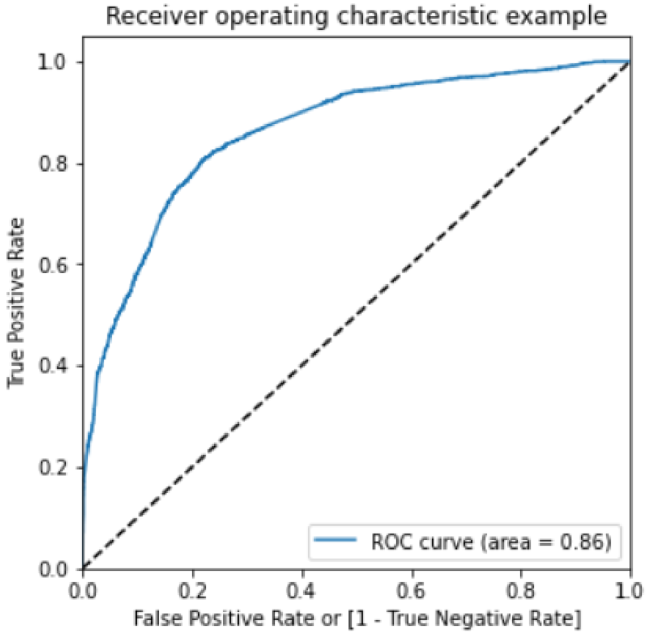
- Numerical Variables are Normalised
- Dummy Variables are created for object type variables
- Total Rows for Analysis: 6373
- Total Columns for Analysis: 75

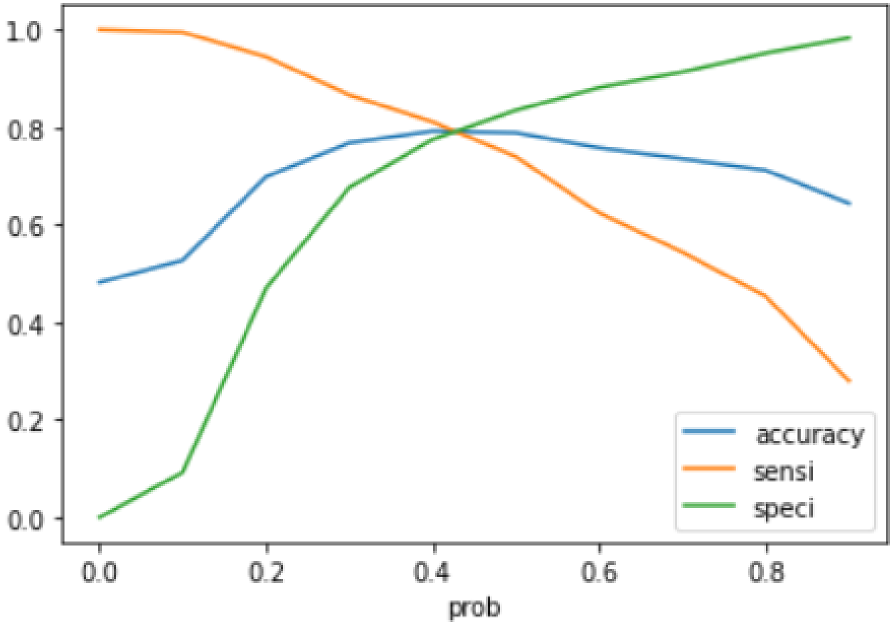


Model Building

- Splitting the Data into Training and Testing Sets
- The first basic step for regression is performing a train-test split, we have chosen 70:30 ratio.
- Use RFE for Feature Selection
- Running RFE with 15 variables as output
- Building Model by removing the variable whose p-value is greater than 0.05 and VIF-value is greater than 5
- Predictions on test data set
- Overall accuracy around 80%

Model Evaluation





Finding Optimal Cut off Point

- Optimal cut off probability is that probability where we get balanced model accuracy, sensitivity and specificity.

- From the second graph it is visible that the optimal cut off is at 0.42.

Conclusion

- It was found that the variables that mattered the most in the potential buyers are
  - Total Visits
  - Total Time Spent on Website
  - Lead Origin_Lead Add Form
  - Lead Source_Olark Chat
  - Lead Source_Welingak Website
  - Do Not Email_Yes
  - Last Activity_Had a Phone Conversation
  - Last Activity_SMS Sent
  - What is your current occupation_Student
  - What is your current occupation_Unemployed
  - Last Notable Activity_Unreachable

- Recommendations:
  - 'Total Visits', 'Total Time Spent on Website', 'Page Views Per Visit' which contribute most towards the probability of a lead getting converted.
  - The Sales team needs to target the people to below leads.
    - The people who all are visiting the site and spending maximum time.
    - Connect the people where earlier a phone conversation happened but didn’t enroll for any course yet
    - People belongs to category Unemployed/ Student category as they might be looking for any courses to upskill their skill set and looking for Job opportunities out of that.