Lead Score Case Study

TEAM MEMBER

1. PALASH SAMAR
2. ANJALI SHARMA

Problem Statement

- An education company named X Education sells online courses to industry professionals.
- 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.

Business Objective

- X Education want to know the most promising leads.
- For they want to build a Model to identify the potential leads.
- Deployment of the model and EDA for the future use/

Solutions Steps

- Reading and Understanding of the data
  - Null columns, datatypes of the columns found.
  - Understanding of different columns.
- Data Cleaning
  - Handle missing values.
  - Unnecessary columns are dropped (containing large missing values)
  - Imputation of the values
  - Checking and handling of Outliers.
- EDA
  - Univariate data analysis
  - Bivariate data analysis
- Data Preparation
  - Feature Scaling
  - Dummy Variable
  - Splitting of the data – Train and Test data
- Model Building - Logistic Regression Model
- Model Evaluation
- Conclusion

Data Cleaning

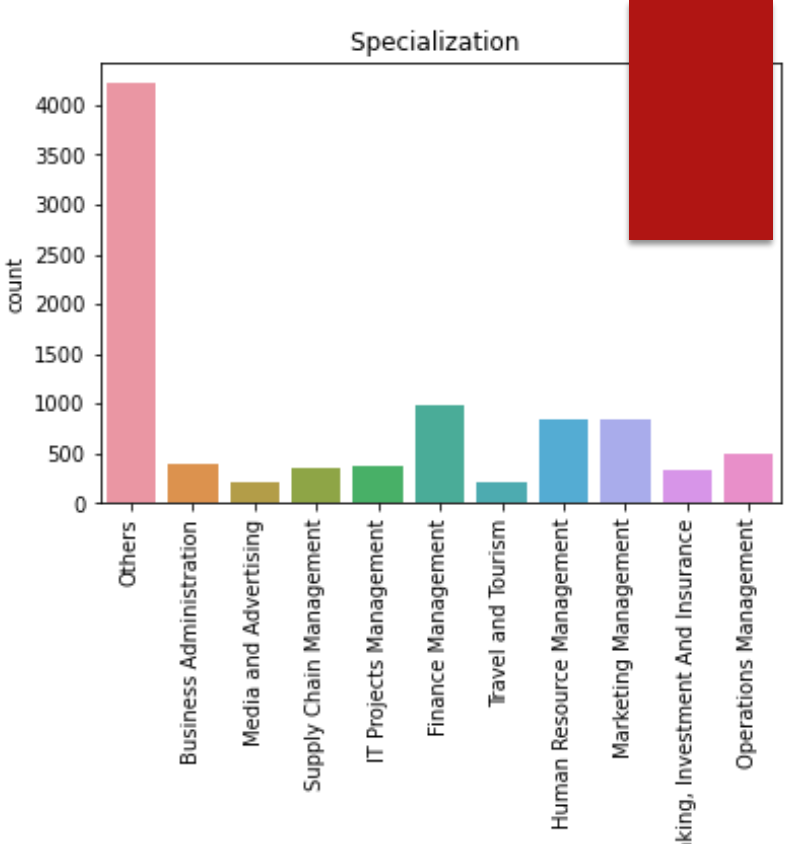
- Dropping the columns having null values more than 45%.
- Dropping the columns having highly skewed data or only single categorical values, as they are helpful for analysis.
- Imputing the null values -
  - Categorical Columns – with most no. Of times of occurrence of value I.e. taking Mode
  - Numerical Columns – taking mean/median depending on the outliers.

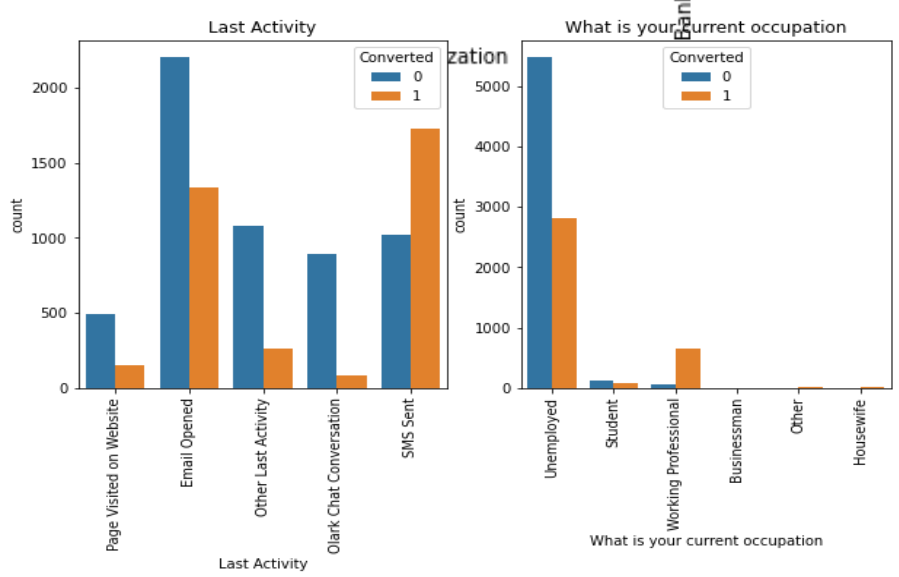
Outliers have been removed using MinMaxScaler.

EDA

Observations:

- It can be seen that many of the learners are looking for some different kind of specialization, which can be added into the curriculum.
- Conversion rate coming from other leads is very much higher which means they have approach to the website via some sort of referral, which means have some bit of early makeup of mind of pursuing the course.
- Similarly is the case for the one who is landing on the website from some other source





Data Preparation

- Numerical Variables are Normalized.
- Dummy Variables are created.
- Unnecessary columns like prospectId , which is not useful for analysis are removed.
- Data split into two parts - Train (70%) and Test (30%)

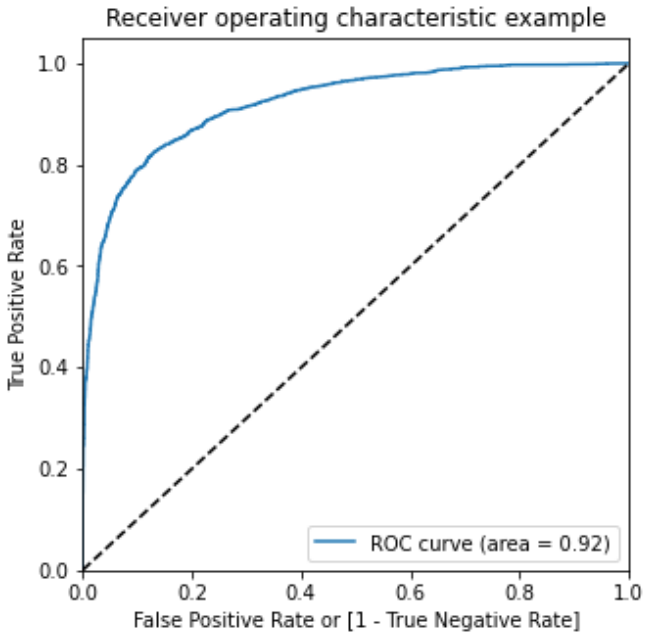
- Total rows – 9240
- Total Columns - 36

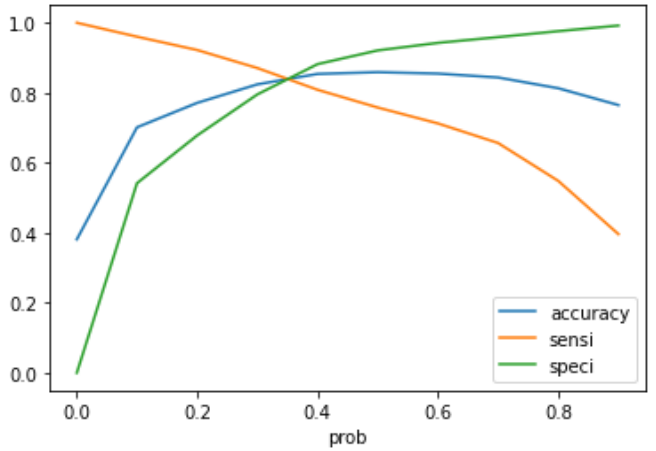
Model Building

- Running RFE with 20 variables.
- Removing other variables manually whose p-value > 0.05 or VIF > 5.
- Prediction on data set.
- Overall Accuracy - 86%.

ROC Curve

- Finding Optimal Cut off point.
- Need to improve sensitivity, from graph 0.35 is seems to be optimal.





Conclusion

- It was found that the variables that mattered the most in potential buyers are :
  - Total Visits
  - Current Occupation (professional)
  - Total time spent on the website
  - Tags - Revert after reading the email
  - Specialization.