X Education - Lead Scoring Case Study

Detection of Hot Leads to concentrate more of marketing efforts on them, improving conversion rates for X Education

Teammates : Yash Mehta , Vinay Saraf , Sagar Varshney

Table of Contents

- Background of X Education Company
- Problem Statement & Objective of the Study
- Analysis Approach
- Data Cleaning
- EDA
- Data Preparation
- Model Building (RFE & Manual fine tuning)
- Model Evaluation
- Recommendations

Background of X Education Company

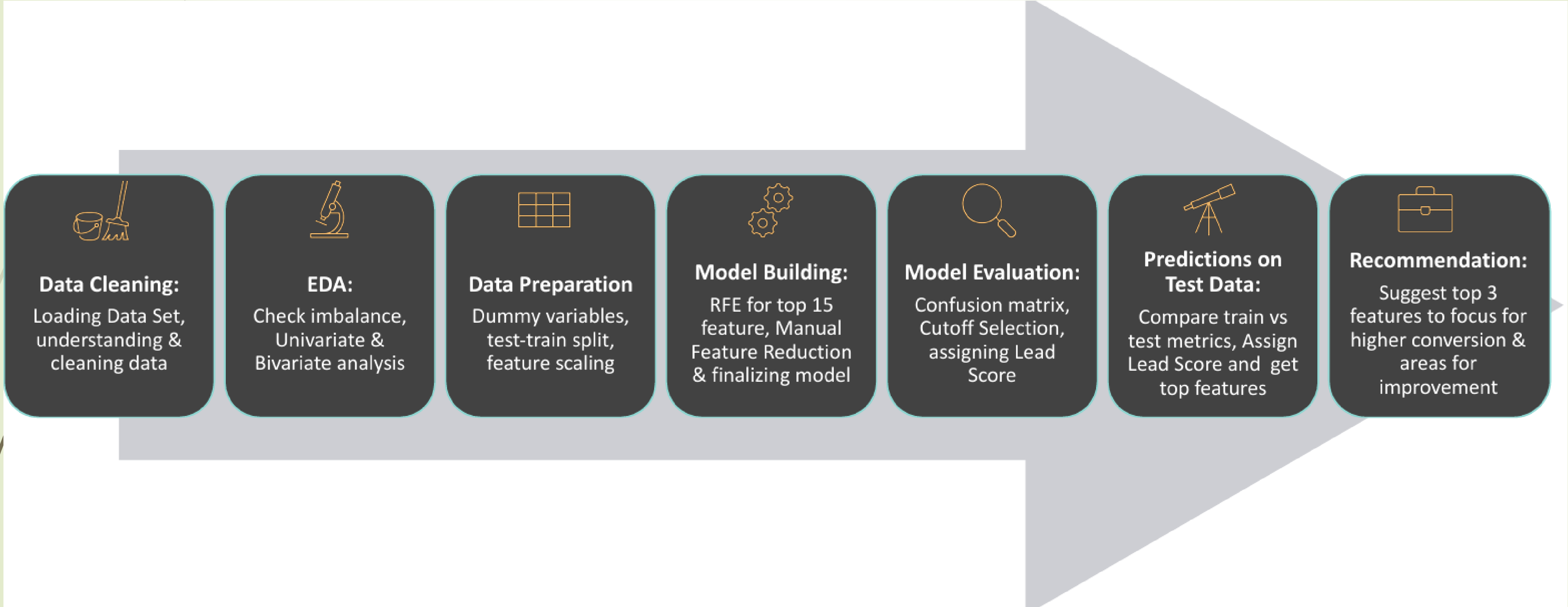
- An education company named X Education sells online courses to industry professionals.
- On any given day, many professionals who are interested in the courses land on their website and browse for courses.
- The company markets its courses on several websites and search engines like Google.
- Once 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.
- 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.

Problem Statement & Objective of the Study

- Problem Statement:
- X Education gets a lot of leads, its lead conversion rate is very poor at around 30%
- X Education wants to make lead conversion process more efficient by identifying the most potential leads, also known as Hot Leads
- Their sales team want to know these potential set of leads, which they will be focusing more on communicating rather than making calls to everyone.

- Objective of the Study:
- To help X Education select the most promising leads, i.e., the leads that are most likely to convert into paying customers.
- The company requires us to build a model wherein we need to assign a lead score to each of the leads such that the customers with a higher lead score have a higher conversion chance and the customers with a lower lead score have a lower conversion chance.
- The CEO has given a ballpark of the target lead conversion rate to be around 80%.

Analysis Approach



Data Cleaning:
Loading Data Set, understanding & cleaning data

EDA:
Check imbalance, Univariate & Bivariate analysis

Data Preparation
Dummy variables, test-train split, feature scaling

Model Building:
RFE for top 15 feature, Manual Feature Reduction & finalizing model

Model Evaluation:
Confusion matrix, Cutoff Selection, assigning Lead Score

Predictions on Test Data:
Compare train vs test metrics, Assign Lead Score and get top features

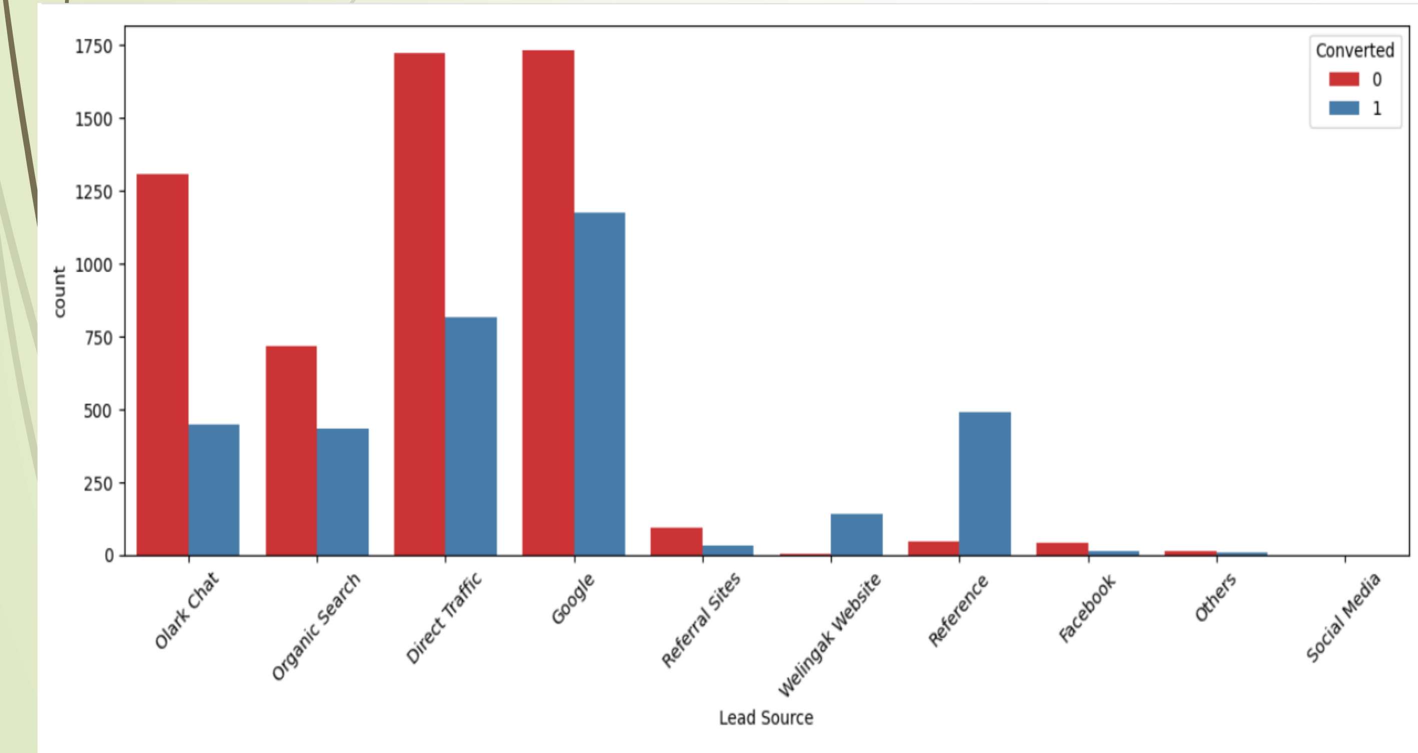
Recommendation:
Suggest top 3 features to focus for higher conversion & areas for improvement

Data Cleaning

- "Select" level represents null values for some categorical variables, as customers did not choose any option from the list.
- Columns with over 35% null values were dropped.
- Missing values in categorical columns were handled based on value counts and certain considerations.
- Drop columns that don't add any insight or value to the study objective (tags, country)
- Imputation was used for some categorical variables.
- Additional categories were created for some variables.

EDA

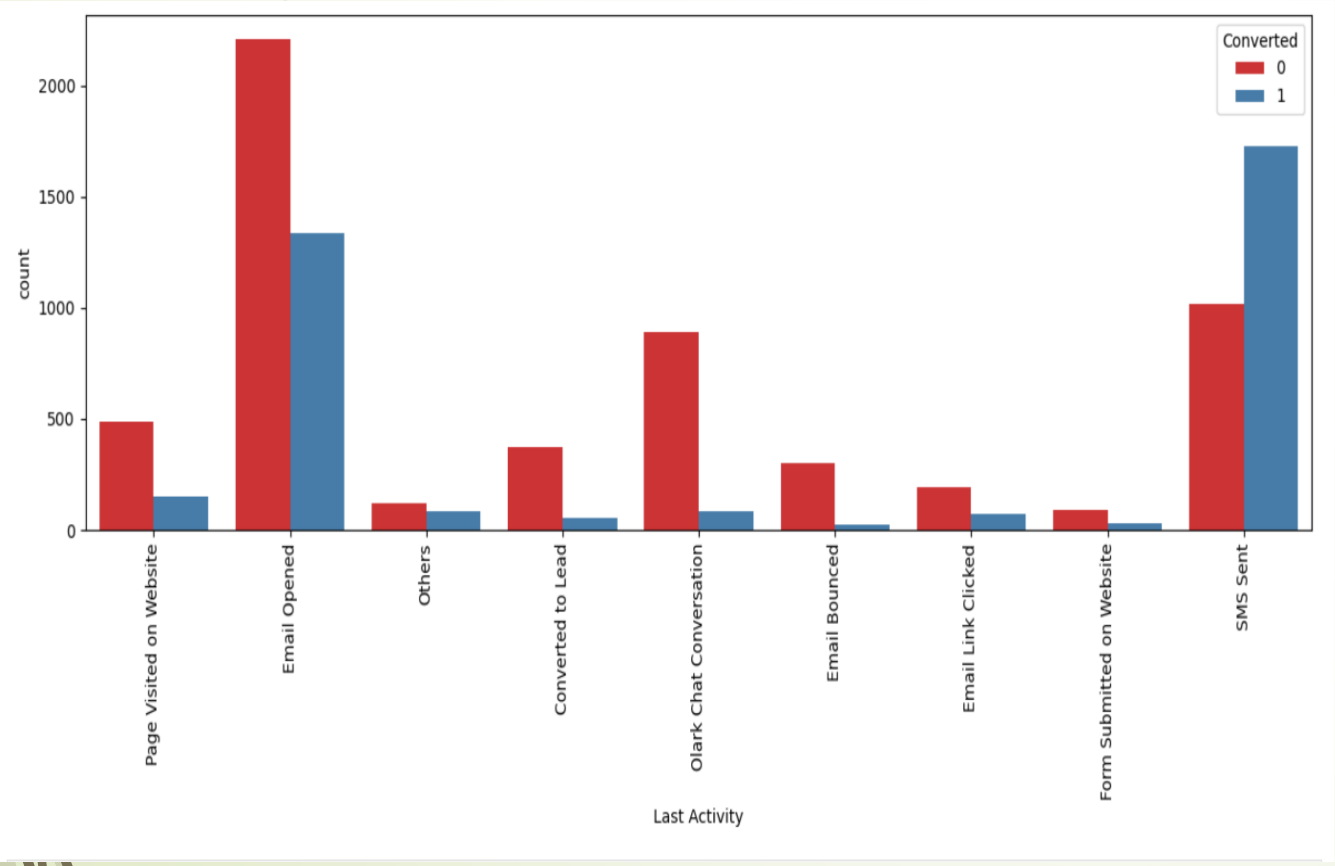
Univariate Analysis – Categorical Variables



- The Lead source is from Google & Direct Traffic combined as compared with other sources.

EDA

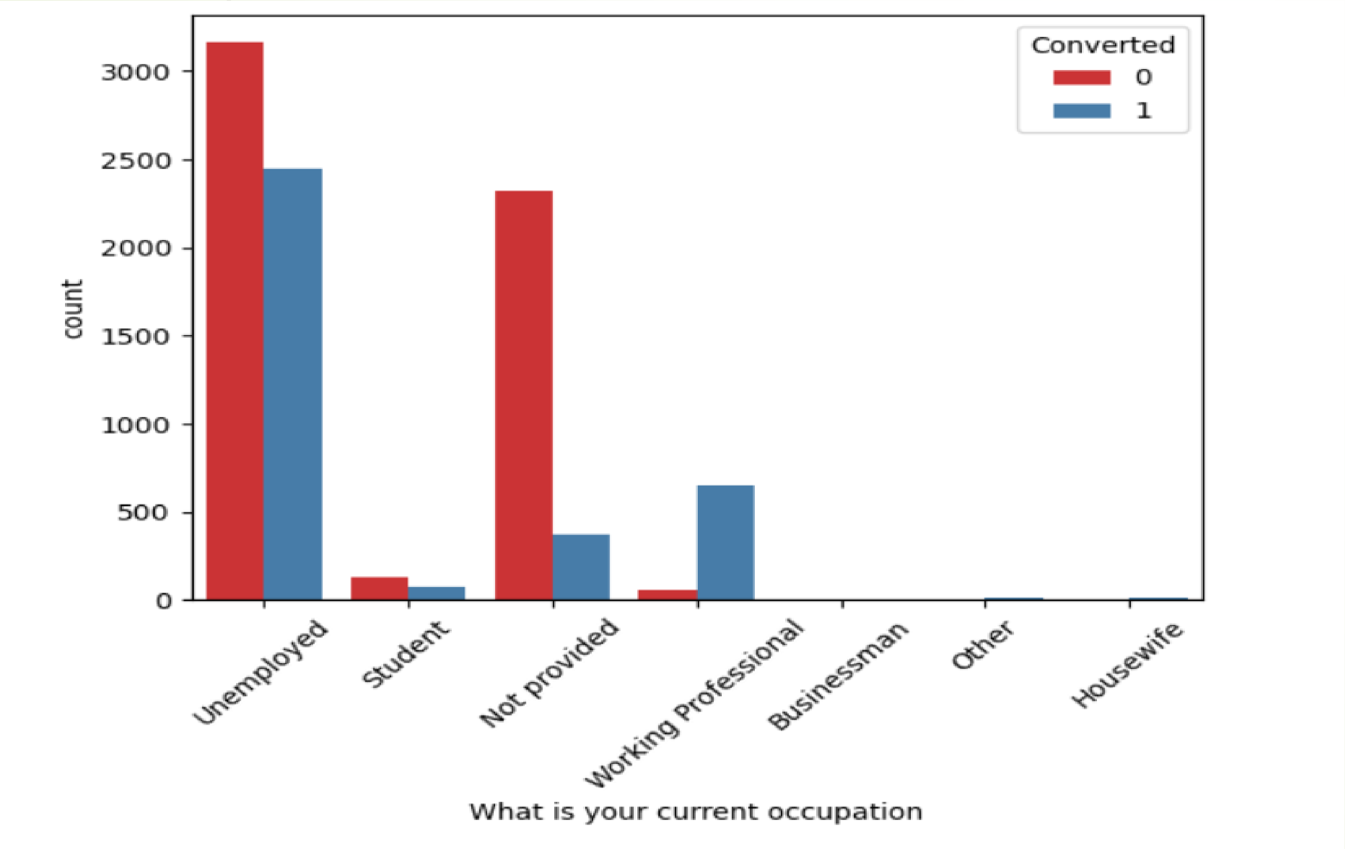
Univariate Analysis – Categorical Variables



- The highest customers contribution comes from SMS Sent & Email Opened activities.

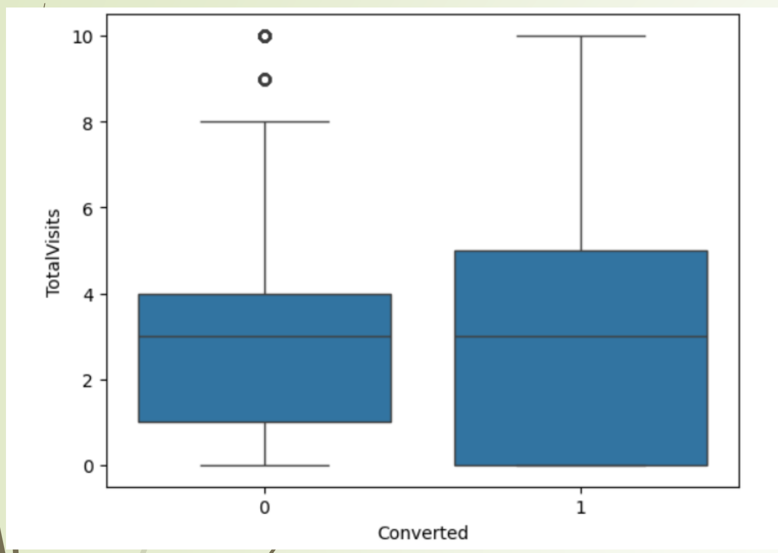
EDA

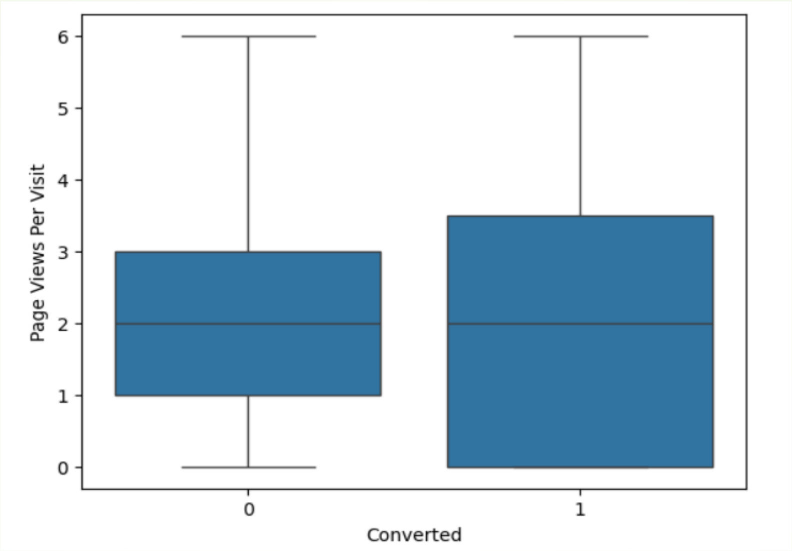
Univariate Analysis – Categorical Variables

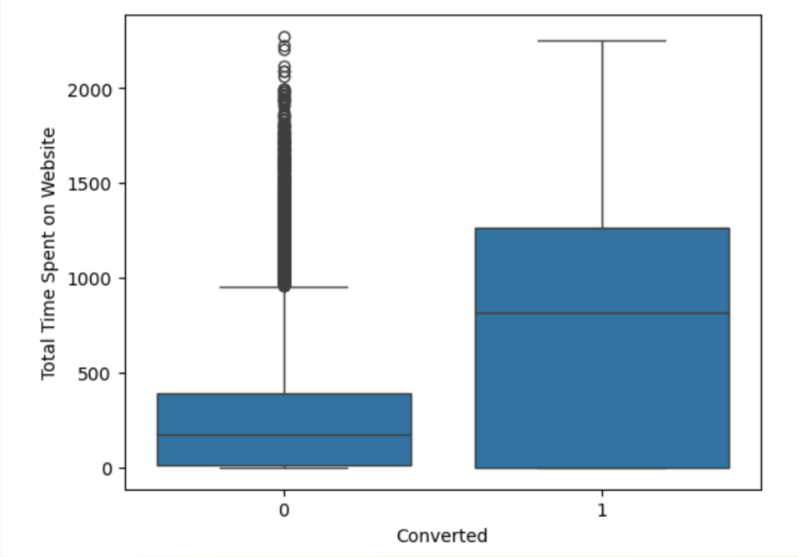


- As per the chart the highest number of customers are unemployed.

EDA – Bivariate Analysis for Numerical Variables







- Past Leads who spends more time on the Website have a higher chance of getting successfully converted than those who spends less time as seen in the box-plot

Data Preparation before Model building

- Binary level categorical columns were already mapped to 1 / 0 in previous steps
- Created dummy features (one-hot encoded) for categorical variables – Lead Origin,Lead Source_Others, What is your current occupation_Not provided
- Splitting Train & Test Sets ○ 70:30 % ratio was chosen for the split
- Feature scaling ○ Standardization method was used to scale the features

Model Building

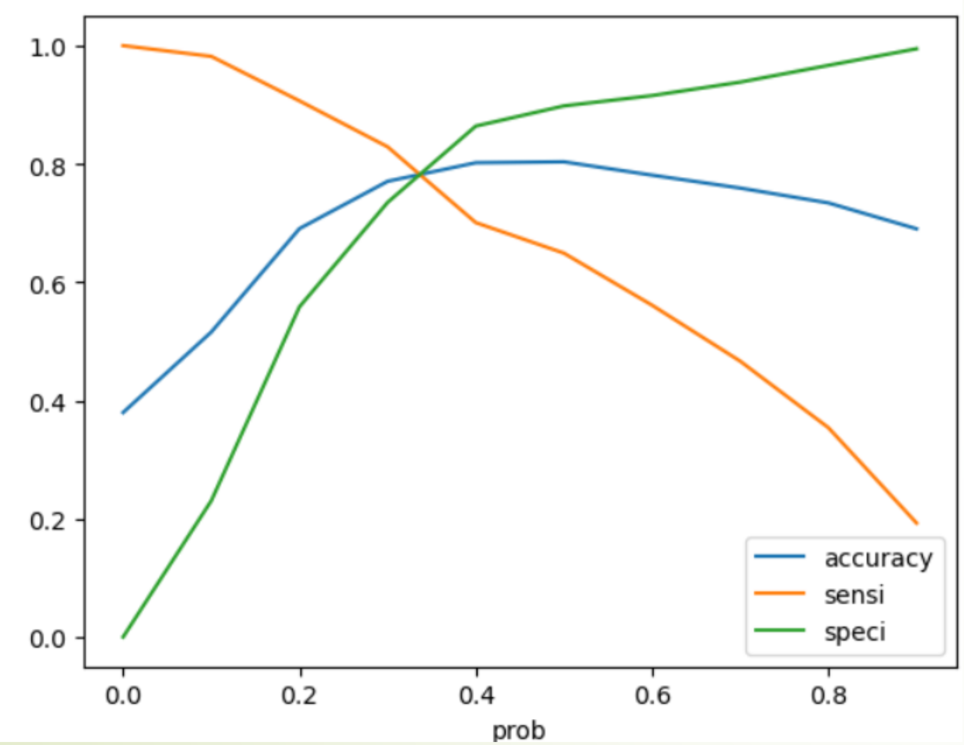
- Feature Selection
- The data set has lots of dimension and large number of features.
- This will reduce model performance and might take high computation time.
- Hence it is important to perform important columns. Recursive Feature Elimination (RFE) and to select only the

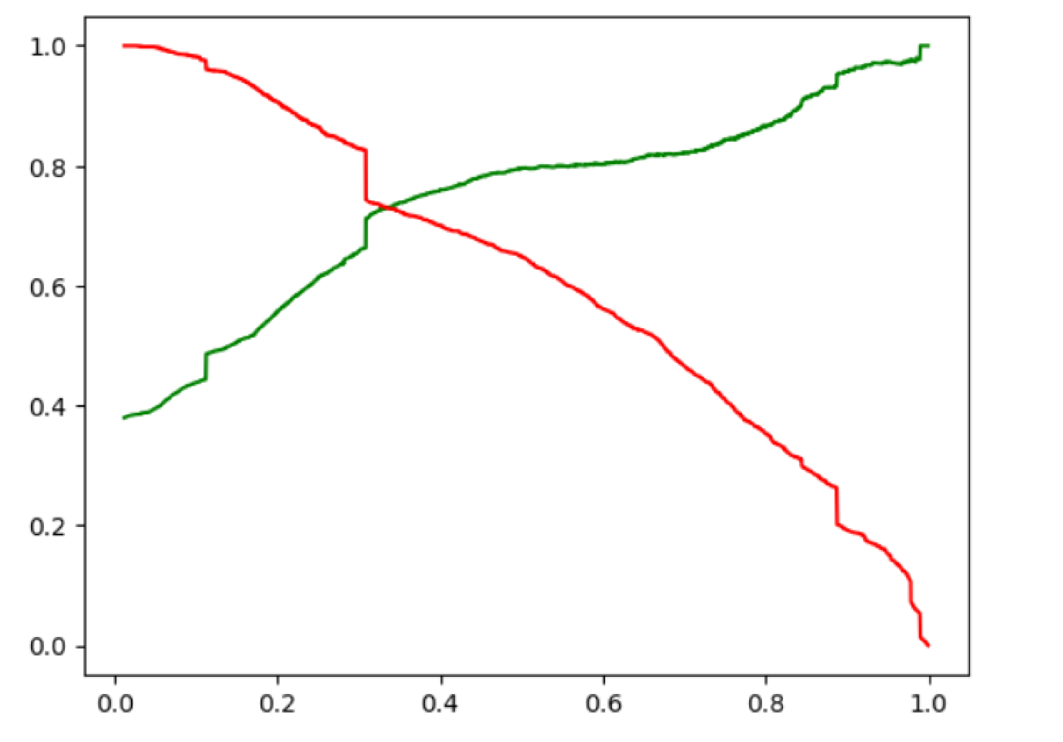
- Then we can manually fine tune the model.
- RFE outcome
- Pre RFE – 22 columns & Post RFE – 15 columns

Model Building

- Manual Feature Reduction process was used to build models by dropping variables with p – value greater than 0.05.
- Model 4 looks stable after four iteration with:
- Significant p-values within the threshold (p-values < 0.05) and No sign of multicollinearity with VIFs less than 5
- Hence, logm4 will be our final model, and we will use it for Model Evaluation which further will be used to make predictions

Model Evaluation - Train Dataset





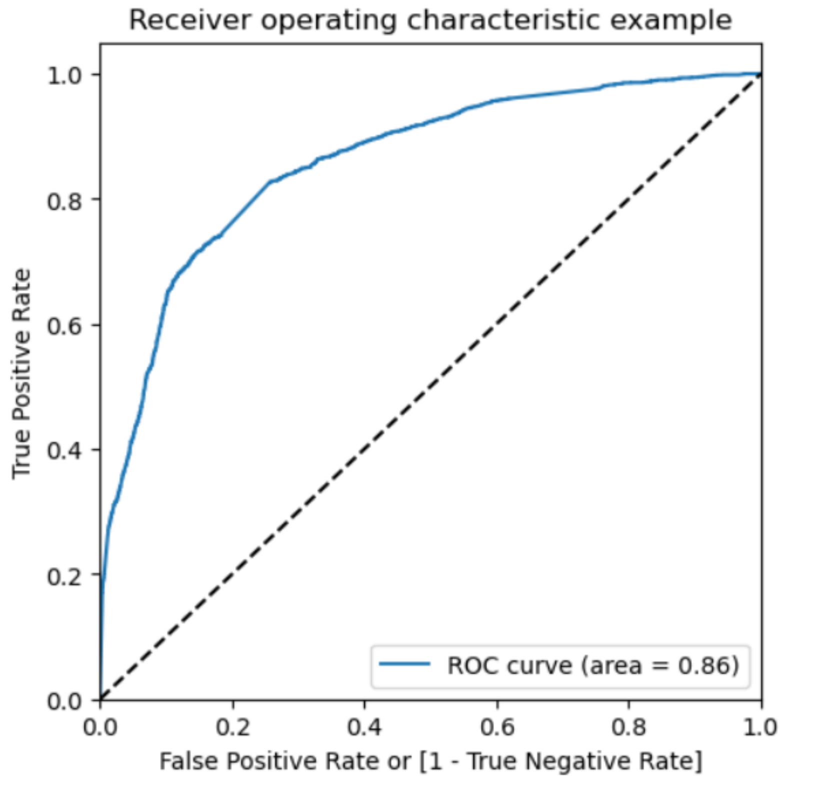
The graph shows the optimal cut off of 0.3 based on Accuracy, Sensitivity and Specificity.

The graph shows an optimal cut off of 0.4 based on Precision and Recall

Model Evaluation

ROC Curve – Train Data Set

- Area under ROC curve is 0.86 out of 1 which indicates a good predictive model.
- The curve is as close to the top left corner of the plot, which represents a model that has a high true positive rate and a low false positive rate at all threshold values.



Model Evaluation - Confusion Matrix & Metrics

- Using a cut-off value of 0.345, the model achieved a sensitivity of 80.35% in the train set and 77.05% in the test set.
- Sensitivity in this case indicates how many leads the model identify correctly out of all potential leads which are converting
- The CEO of X Education had set a target sensitivity of around 82%.
- The model also achieved an accuracy of 80.35%, which is in line with the study's objectives.

Recommendations

- The leads having high ‘Lead Score’ can be focused on more for better conversion rate.
- Marketing on Google, since the conversion rate from the traffic from Google is high.
- Encouraging existing converted leads for referrals by providing some incentives for the referrals.
- Since the number of leads is high in Mumbai as compared to other major cities, the company can increase marketing in the other cities as well to achieve more leads.
- The unemployed category can be focused on more and also individuals having Finance Management as specialization.
- Focus on the students can be minimized since the conversion rate is significantly low.