Lead Score Case Study

Group Members –

1. Prasad Sunil Mahurkar
2. Srikaran Goud Chagapuram

1 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%.
- As you can see, there are a lot of leads generated in the initial stage (top) but only a few of them come out as paying customers from the bottom. In the middle stage, you need to nurture the potential leads well (i.e., educating the leads about the product, constantly communicating etc.) in order to get a higher lead conversion.
- X Education has appointed you to help them select the most promising leads, i.e., the leads that are most likely to convert into paying customers. The company requires you to build a model wherein you need to assign a lead score to each of the leads such that the customers with higher lead score have a higher conversion chance and the customers with lower lead score have a lower conversion chance. The CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80%.

2 Business Objective

- X education wants to know most promising leads.
- Build a Model which identifies the hot leads.
- Deployment of the model for the future use.

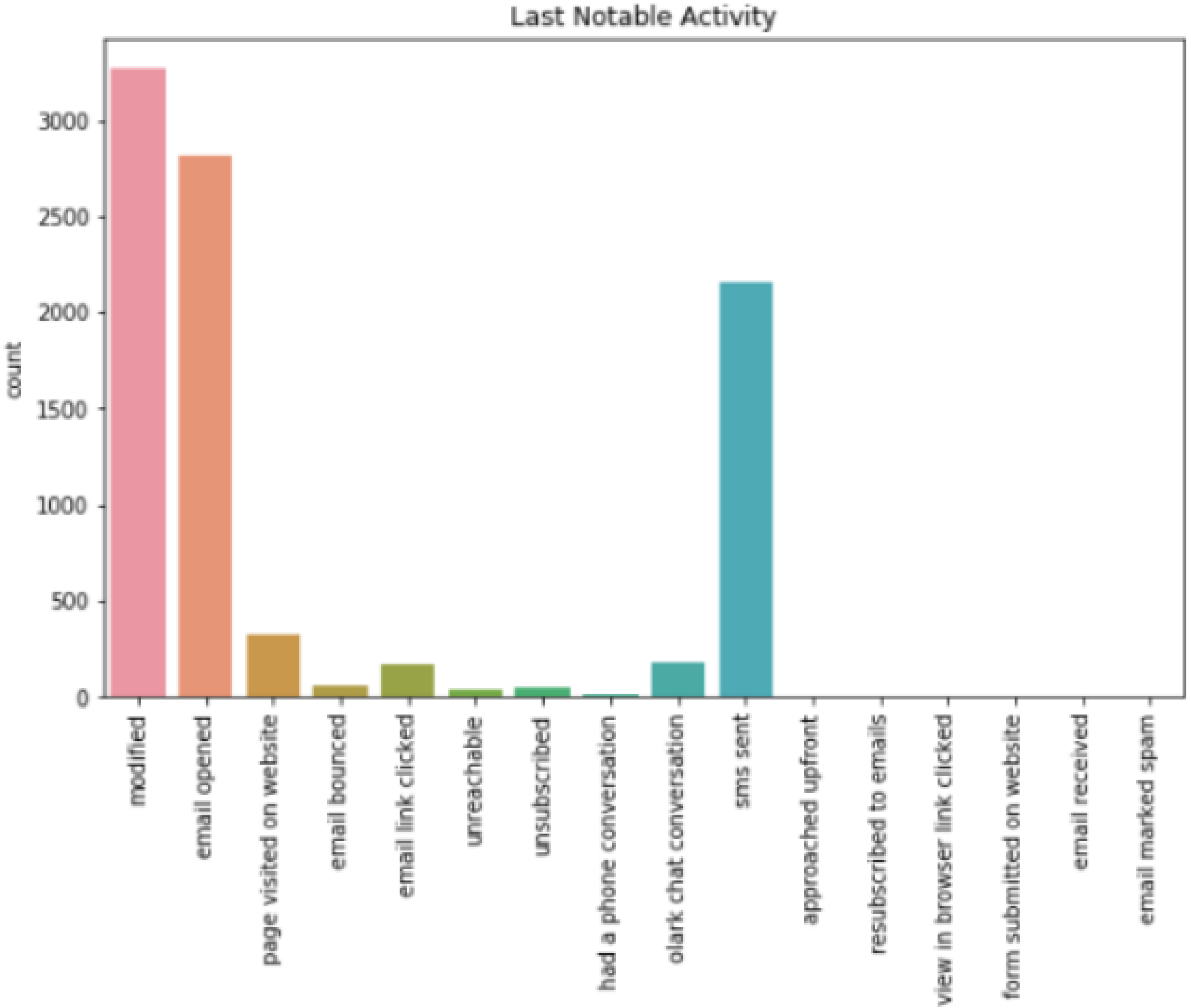
3 Data Cleaning

- Check and handle duplicate data.
- Check and handle NA values and missing values.
- Drop columns, if it contains large number of missing values and not useful for the analysis.
- Imputation of the values, if necessary.
- Check and handle outliers in data.

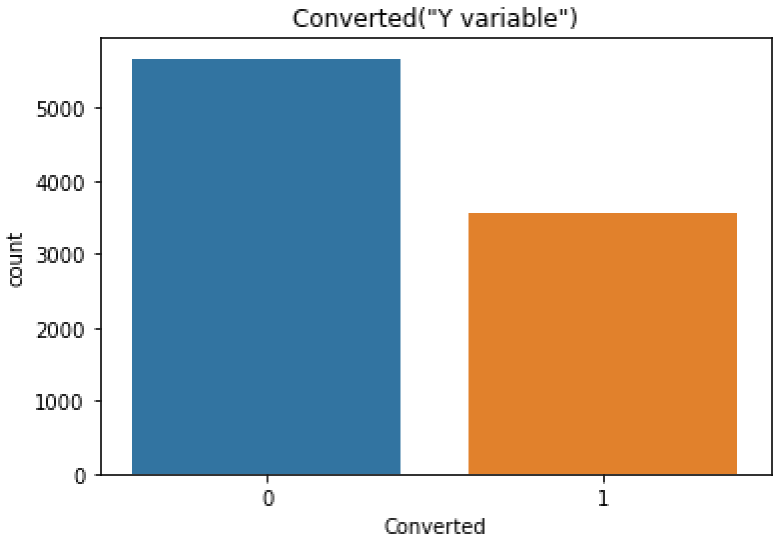
4 Exploratory Data Analysis

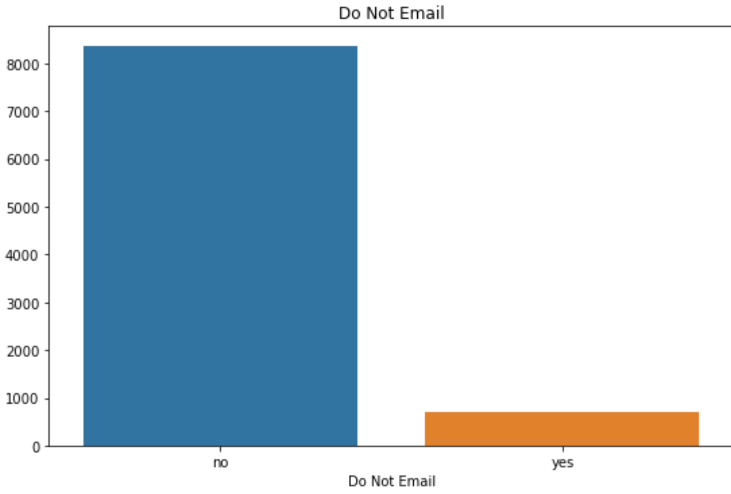
- Univariate data analysis: value count, distribution of variable etc.
- Bivariate data analysis: correlation coefficients and pattern between the variables etc.

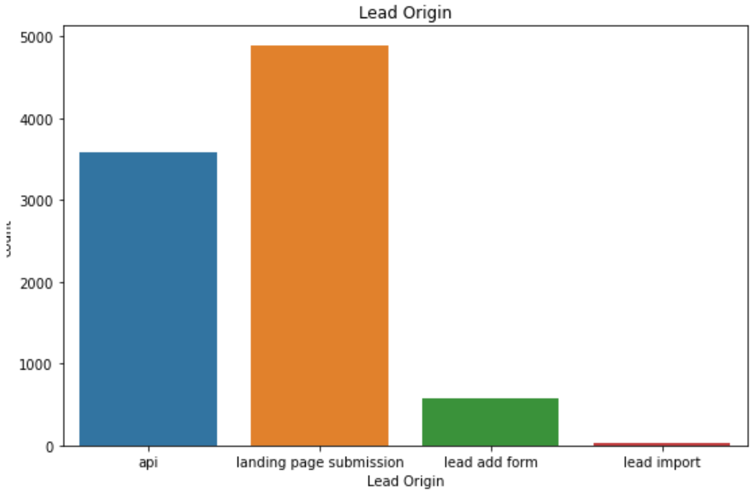
4 Exploratory Data Analysis



4 Exploratory Data Analysis

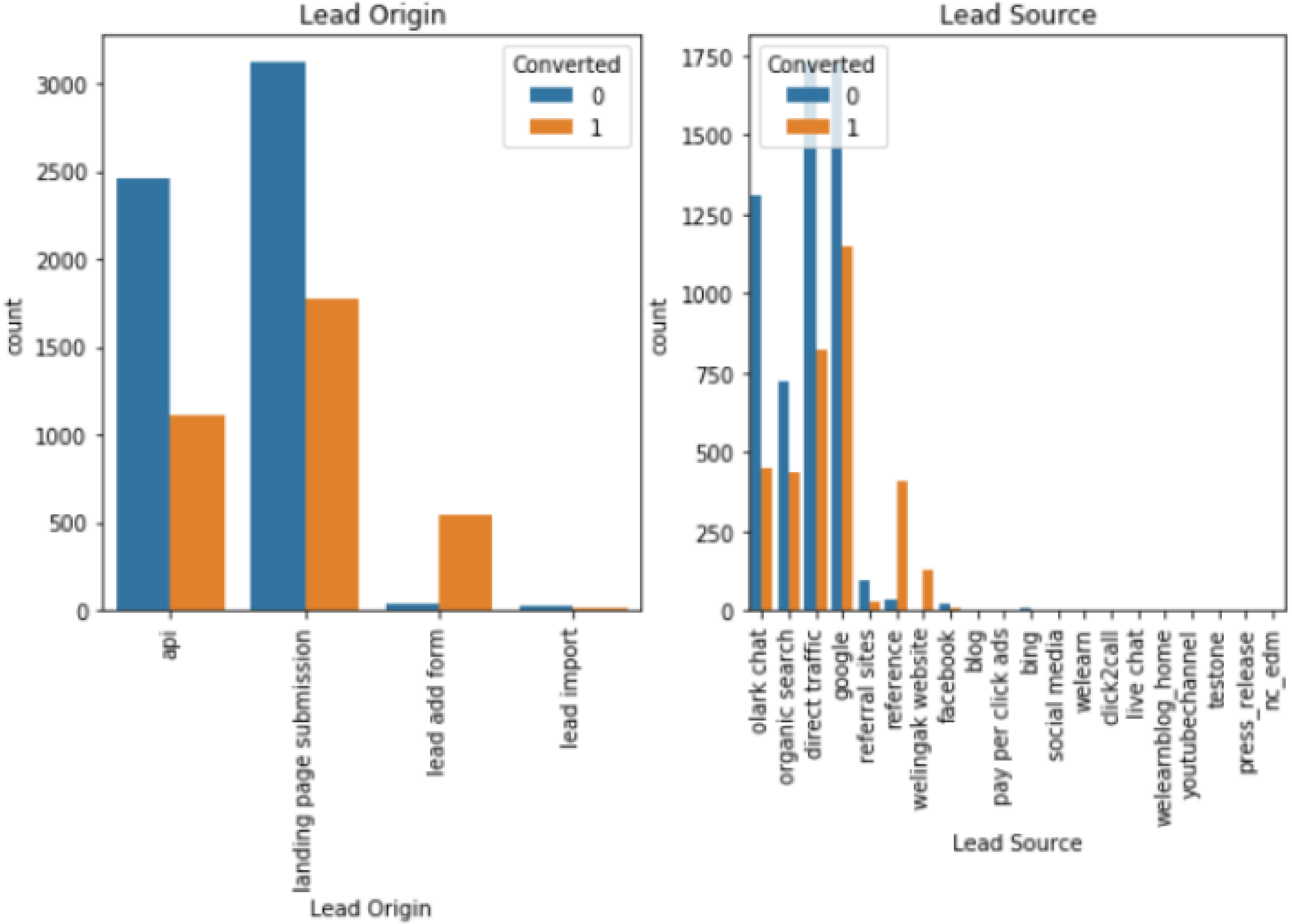






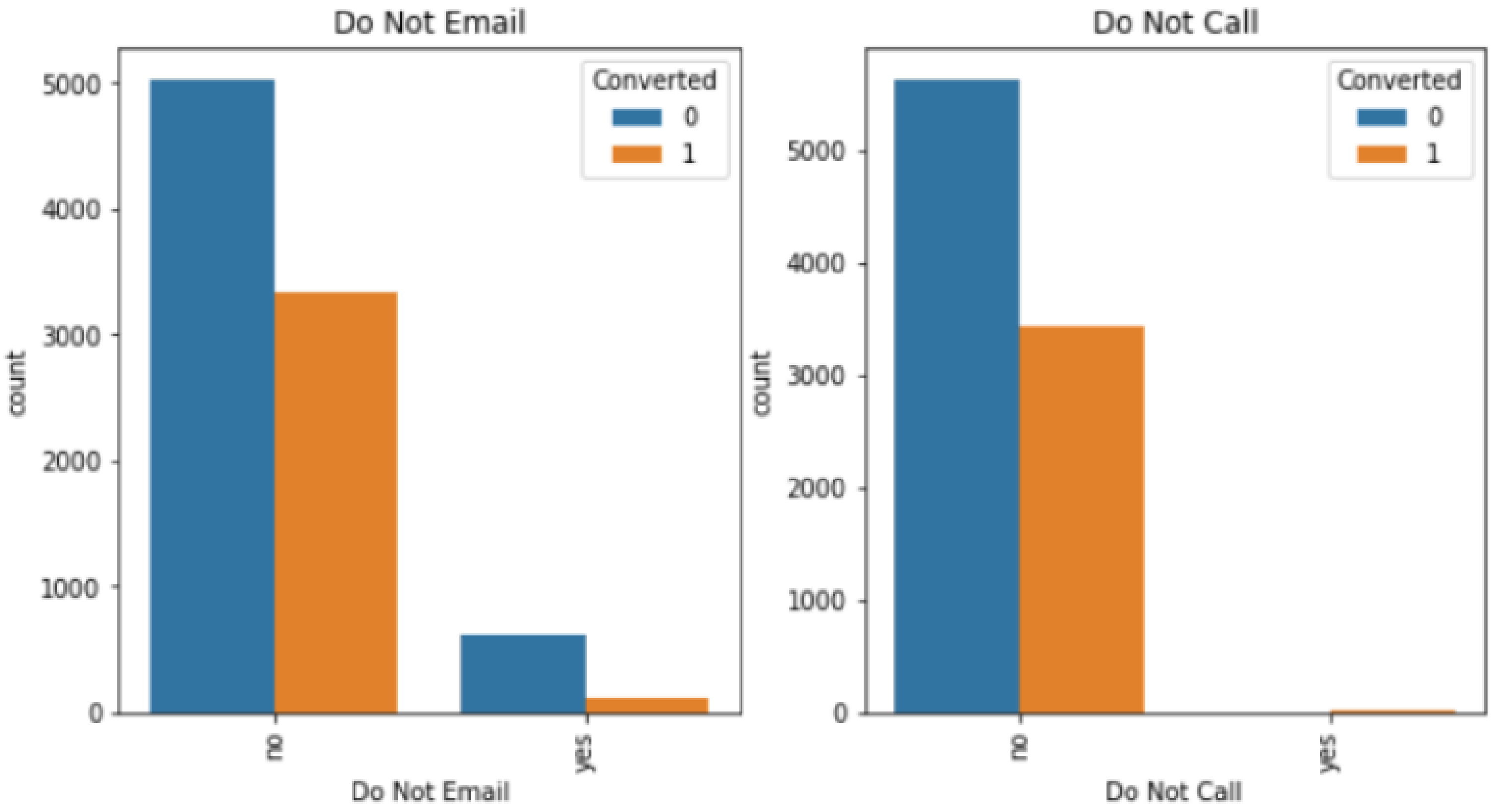
4 Exploratory Data Analysis

Categorical Variable Relation



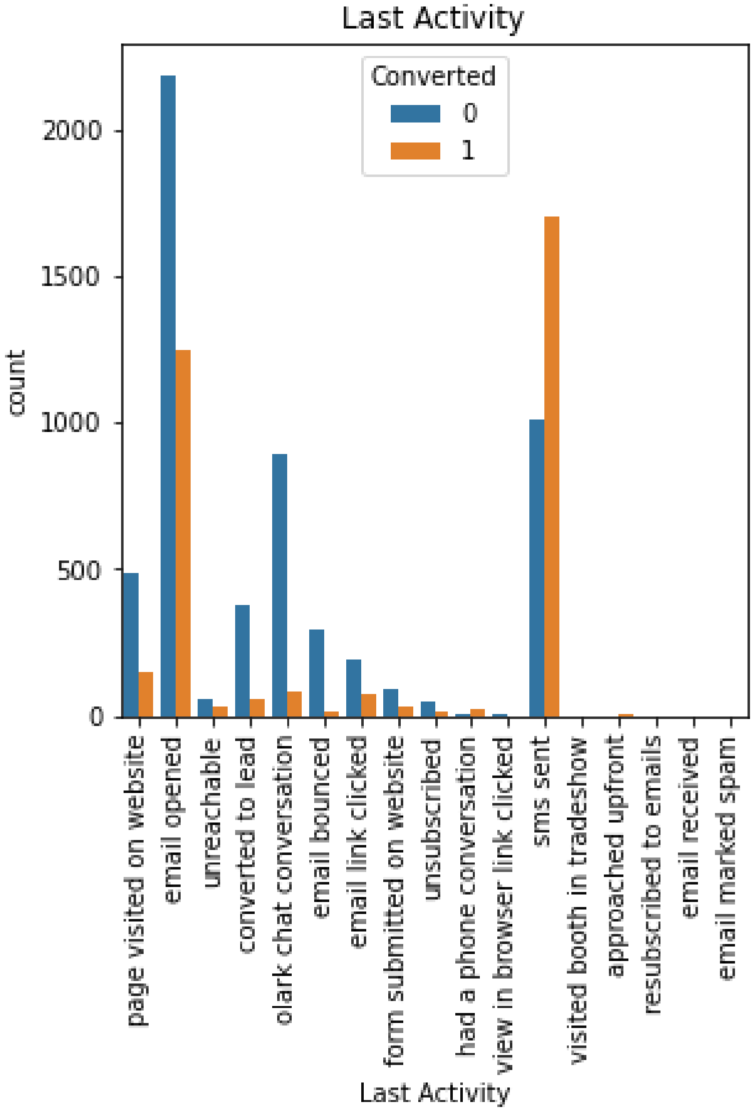
4 Exploratory Data Analysis

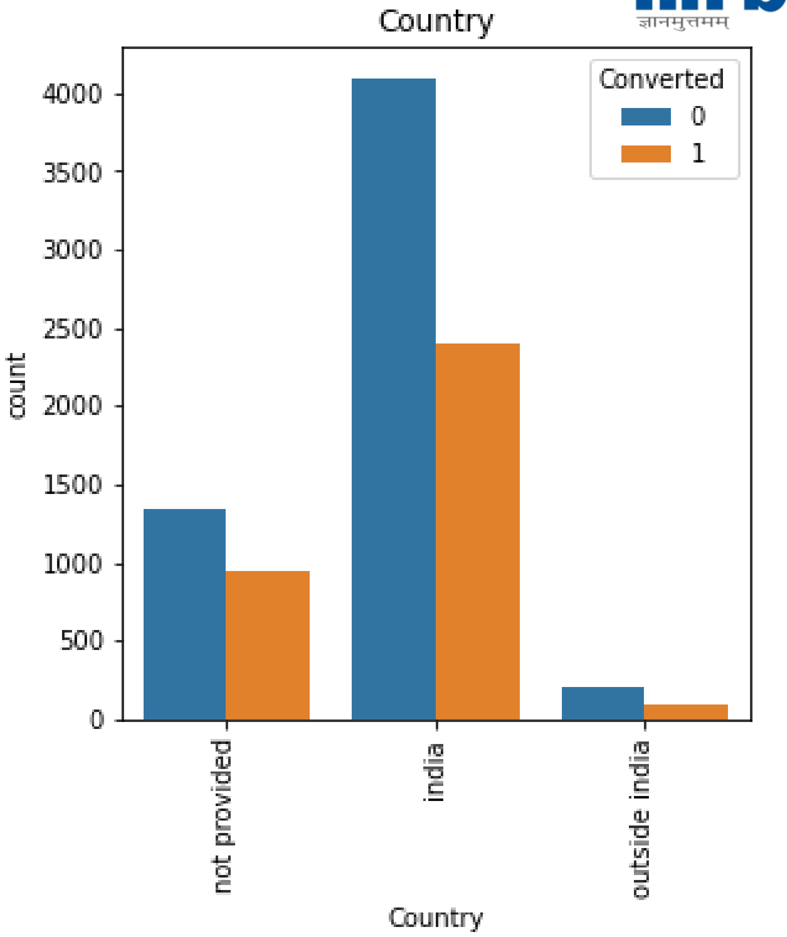
Categorical Variable Relation



4 Exploratory Data Analysis

Categorical Variable Relation





5 Data Manipulation

- Total Number of Rows =37, Total Number of Columns =9240.
- Single value features like "Magazine", "Receive More Updates About Our Courses", "Update me on Supply"
- Chain Content", "Get updates on DM Content", "I agree to pay the amount through cheque" etc. have been dropped.
- Removing the "Prospect ID" and "Lead Number" which is not necessary for the analysis.
- After checking for the value counts for some of the object type variables, we find some of the features which has no enough variance, which we have dropped, the features are: "Do Not Call", "What matters most to you in choosing course", "Search", "Newspaper Article", "X Education Forums", "Newspaper", "Digital Advertisement" etc.
- Dropping the columns having more than 35% as missing value such as 'How did you hear about X Education' and 'Lead Profile'.

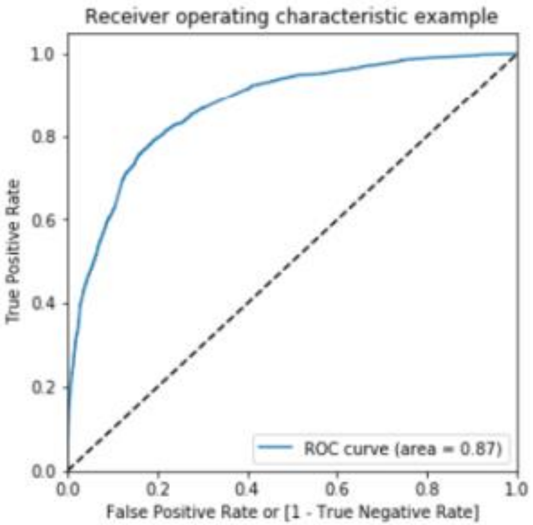
6 Model Building

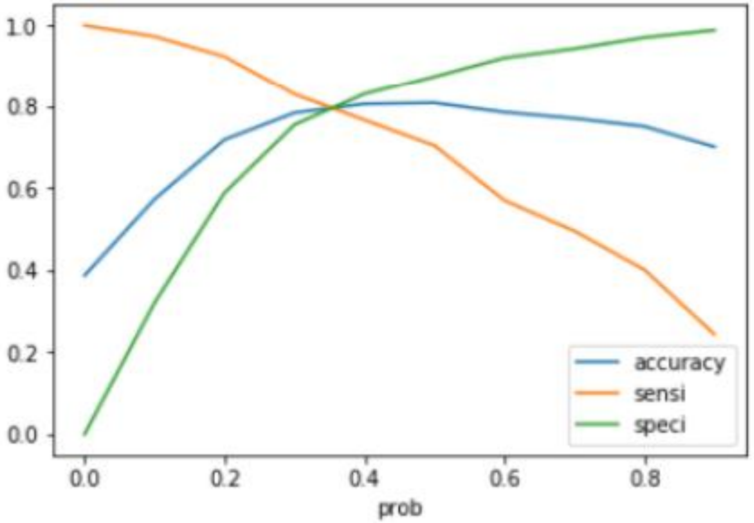
- Splitting the Data into Training and Testing Sets
- The first basic step for regression is performing a train-test split, we have chosen 70:30 ratio.
- Use RFE for Feature Selection
- Running RFE with **12 variables** as output
- Building Model by removing the variable whose p- value is greater than 0.05 and VIF value is greater than 5
- Predictions on test data set
- Overall accuracy 77%

7

ROC Curve

- Finding Optimal Cut off Point
- Optimal cut off probability is that probability where we get balanced sensitivity and specificity
- From the second graph it is visible that the optimal cut off is at 0.3.





8 Conclusion

- Increasing interactivity on the site and improving the time spent would lead to high conversion rate
- Lead source like Google, Direct Traffic and Organic search are the top sources to get potential buyers