Lead Score Case Study

Group Members

1) Siddhant Singh
2) Dakshin JV

Problem Statement

- X Education sells online courses to industry professionals.
- 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 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.

Solution Methodology

□ **Data cleaning and data manipulation.**

1. **Check and handle duplicate data.**
2. **Check and handle NA values and missing values.**
3. **Drop columns, if it contains large amount of missing values and not useful for the analysis.**
4. **Imputation of the values, if necessary.**
5. **Check outliers in data.**

□ **EDA**

□ **Univariate Data analysis: value count, distribution of variable etc.**

□ **Feature Scaling & Dummy Variables and encoding of the data.**

□ **Classification technique: logistic regression used for the model making and prediction.**

□ **Validation of the model.**

□ **Model presentation.**

□ **Conclusion.**

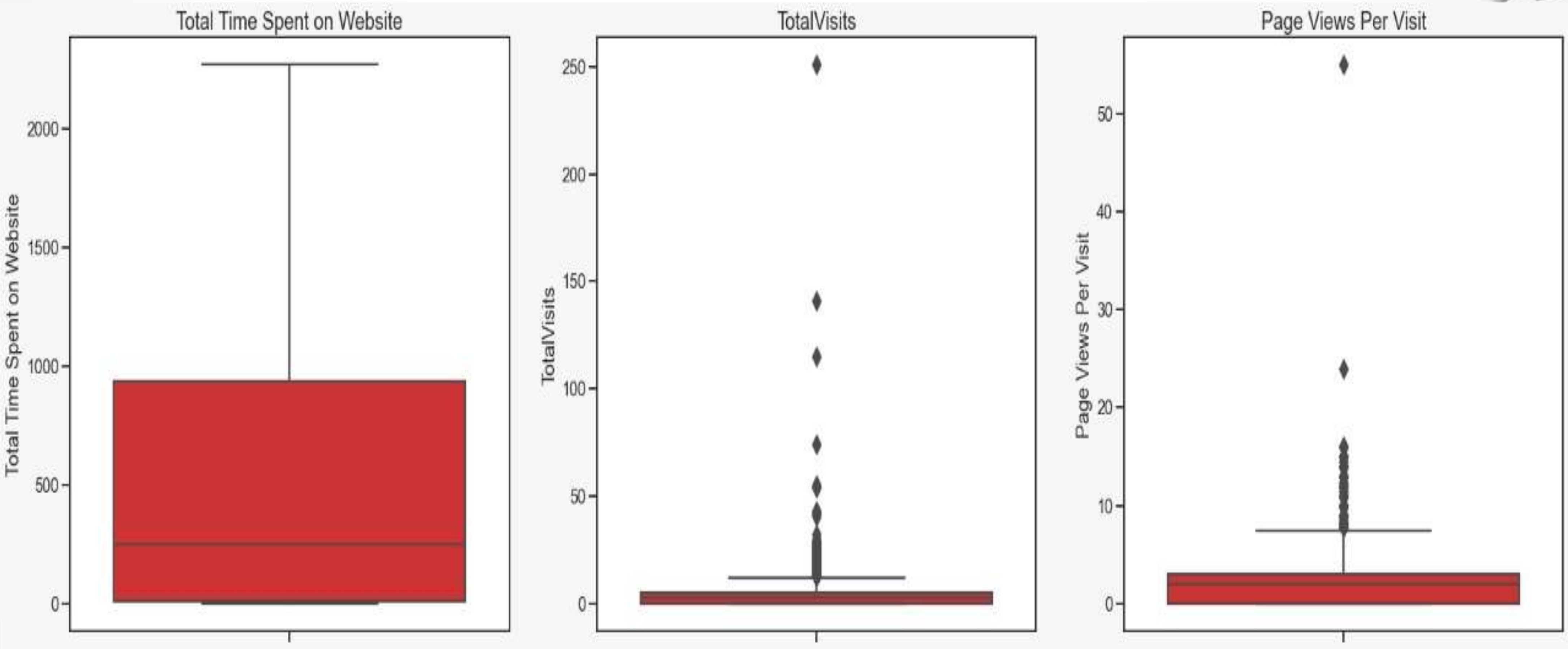
Data Manipulation

- Total Number of Rows = 37, Total Number of Columns = 9240.
- There are some columns/categorical variables having label as 'Select' which means the customer was not selected any option hence it is better to put it as null value - because there was no suitable option present to select for the customer’s searching.
- We have some sales genrated data in our dataset which are not required, so we will drop them also. Sales generated columns are: 'Tags','Last Notable Activity','Last Activity','Prospect ID','Lead Number'.
- There are some columns which are having higher frequncy for a single variable (greater than 99%) which will not help our model.Thus needs to be deleted. The columns: 'Do Not Call', 'Search', 'Newspaper Article', 'X Education Forums', 'Newspaper’, 'Through Recommendations', 'Digital Advertisement, 'What matters most to you in choosing a course', 'Country'.

Data Manipulation

- There are 3 categorical columns : 'Specialization','What is your current occupation' and 'City' where there are a lot of missing data and we will replace them with mode.
- Dropping the columns having more than 40% as missing value.

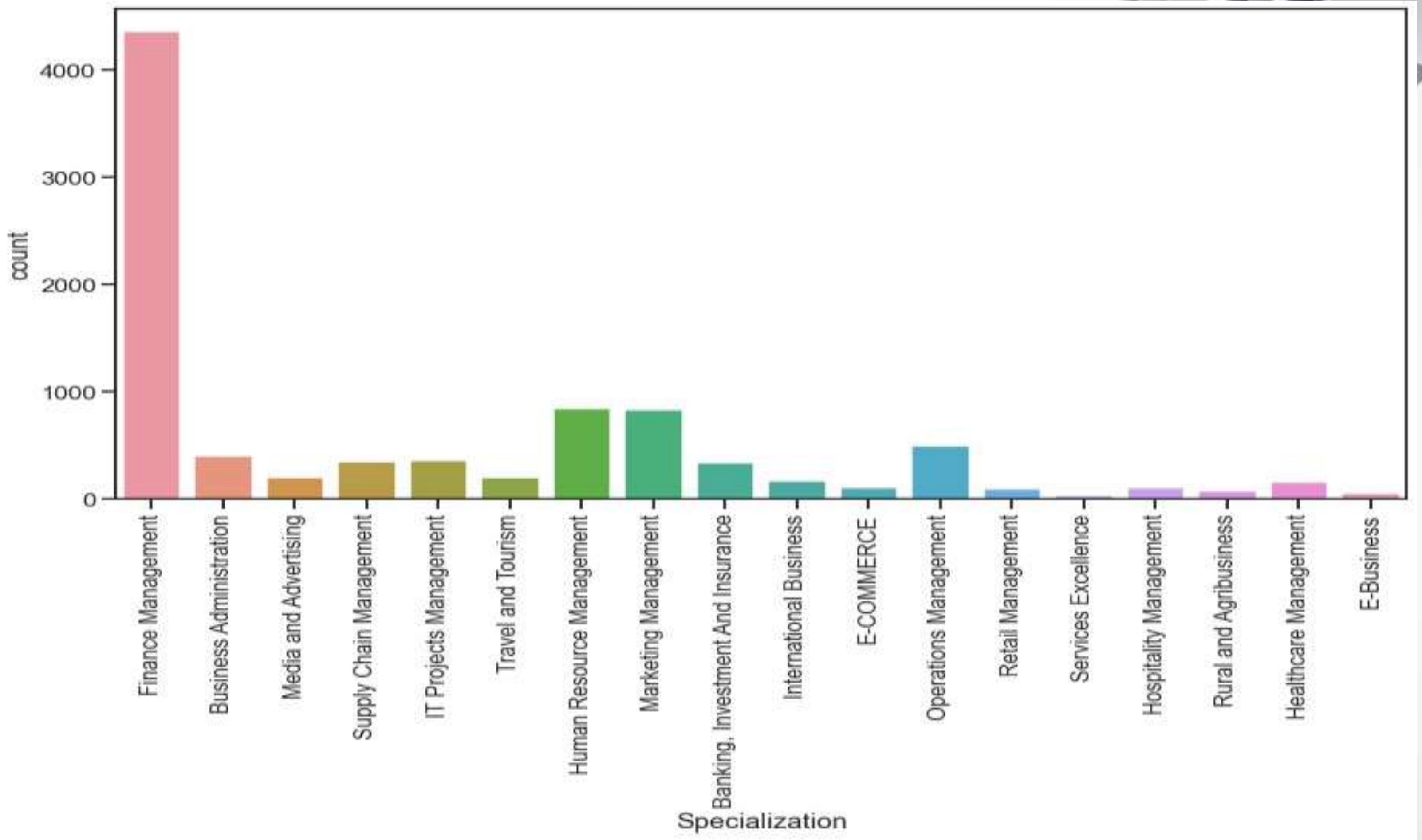
Outliers



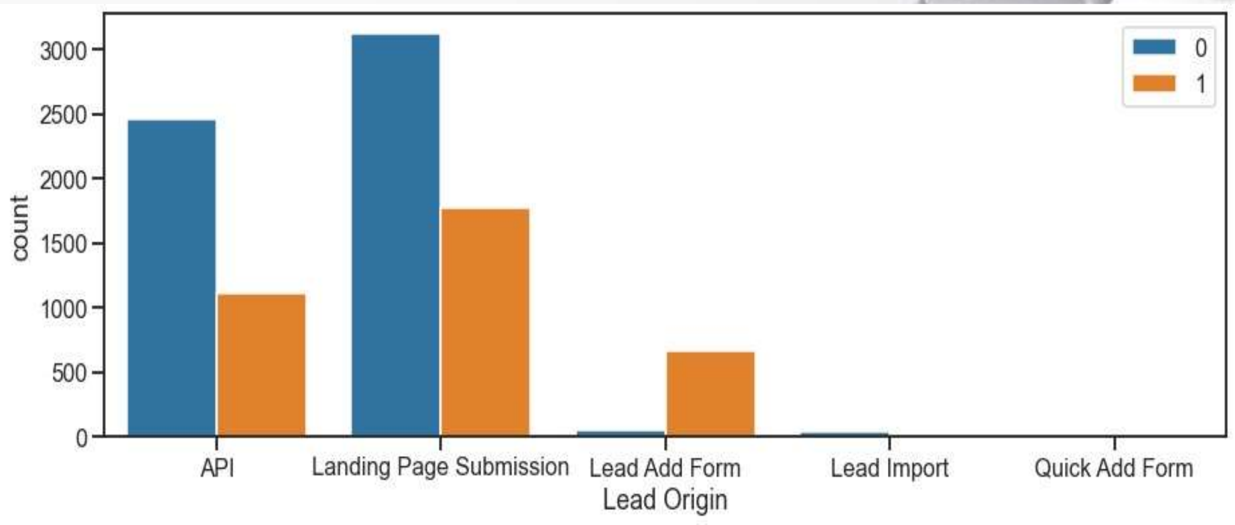
❑ **As we can see there are outliers in 2 variables 'TotalVisits' and 'Page Views Per Visit'.**

EDA

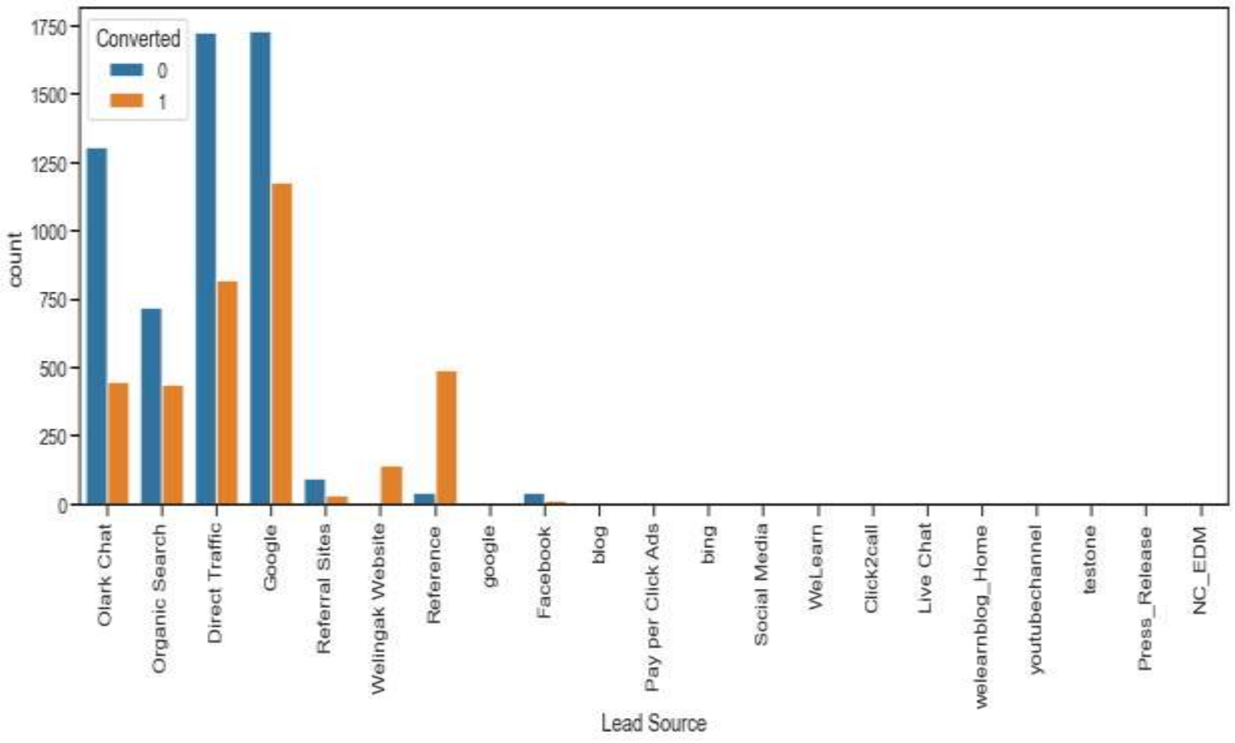
- "Finance Management" is mostly choosen by the people.
- There are equal number of the people opted for HR Management and Marketing Management.



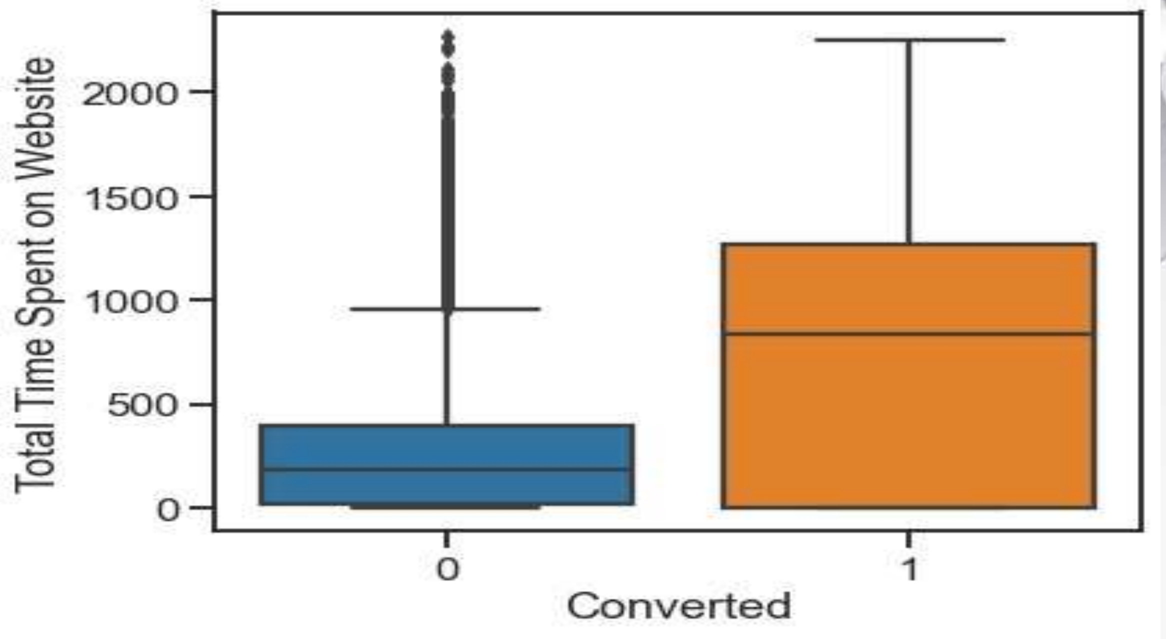
- **API and Landing Page Submission have 30-35% conversion rate but count of lead originated from them are considerable.**
- **Lead Add Form has more than 90% conversion rate but count of lead are not very high.**
- **Lead Import are very less in count.**



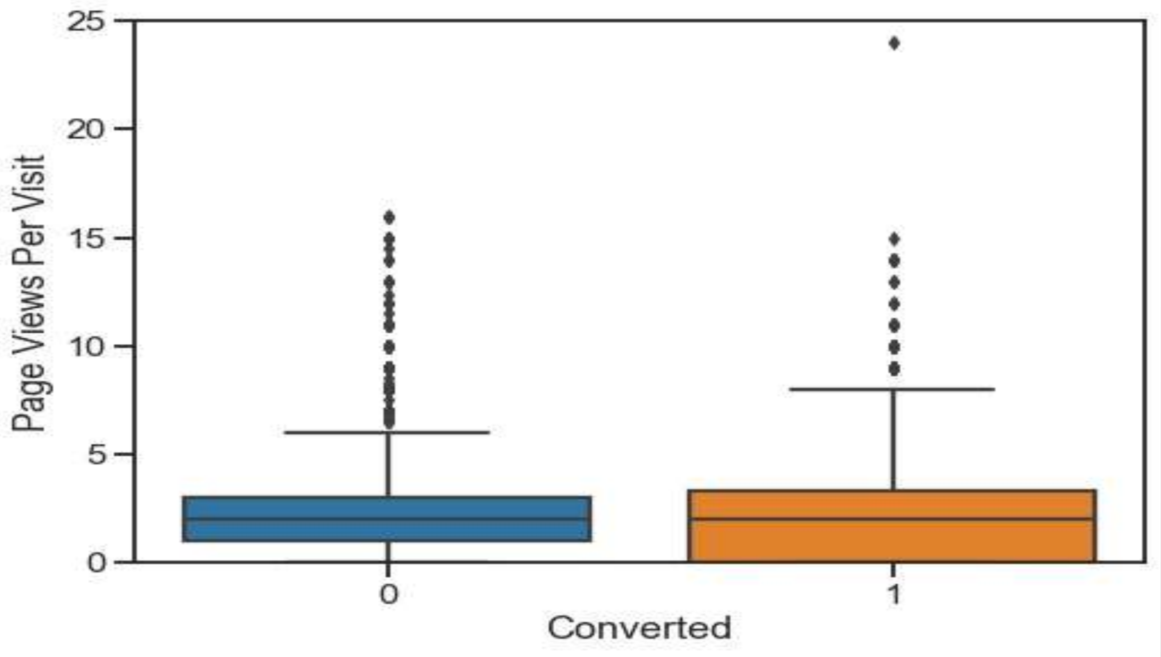
- **Google and Direct traffic generates maximum number of leads.**
- **Conversion Rate of reference leads and leads through welingak website is high.**



- ❑ **Leads spending more time on the website are more likely to be converted.**



- ❑ **Median value for converted and unconverted leads is same.**



Data Conversion

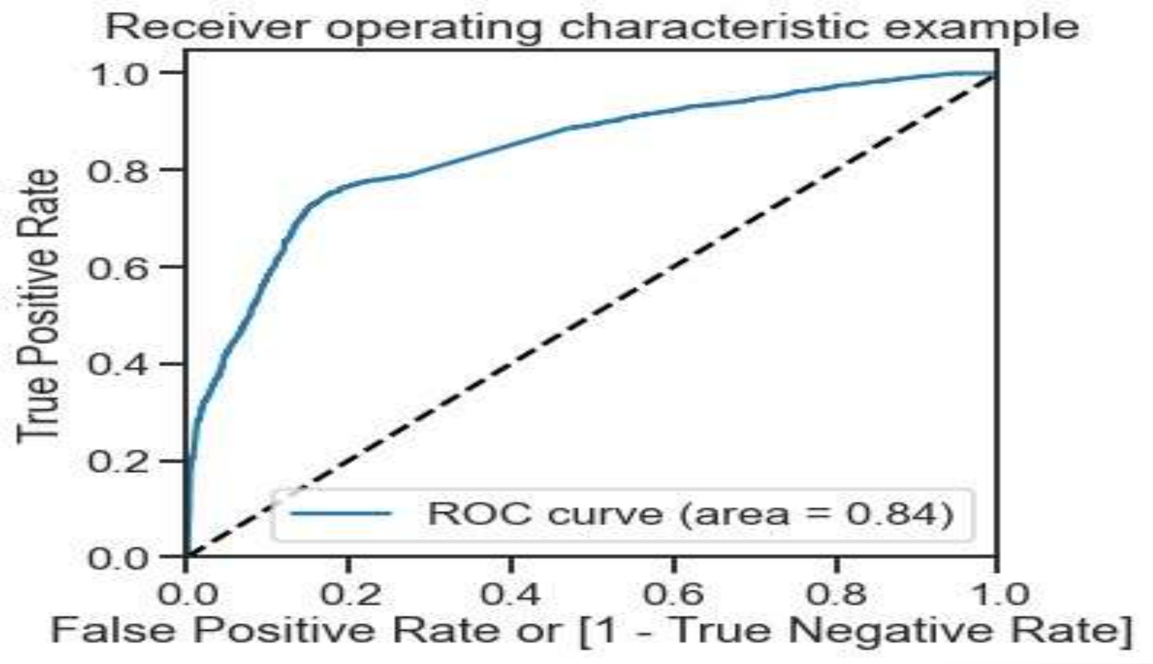
- Numerical Variables are Normalised
- Dummy Variables are created for object type variables
- Total Rows for Analysis: 9240
- Total Columns for Analysis: 57

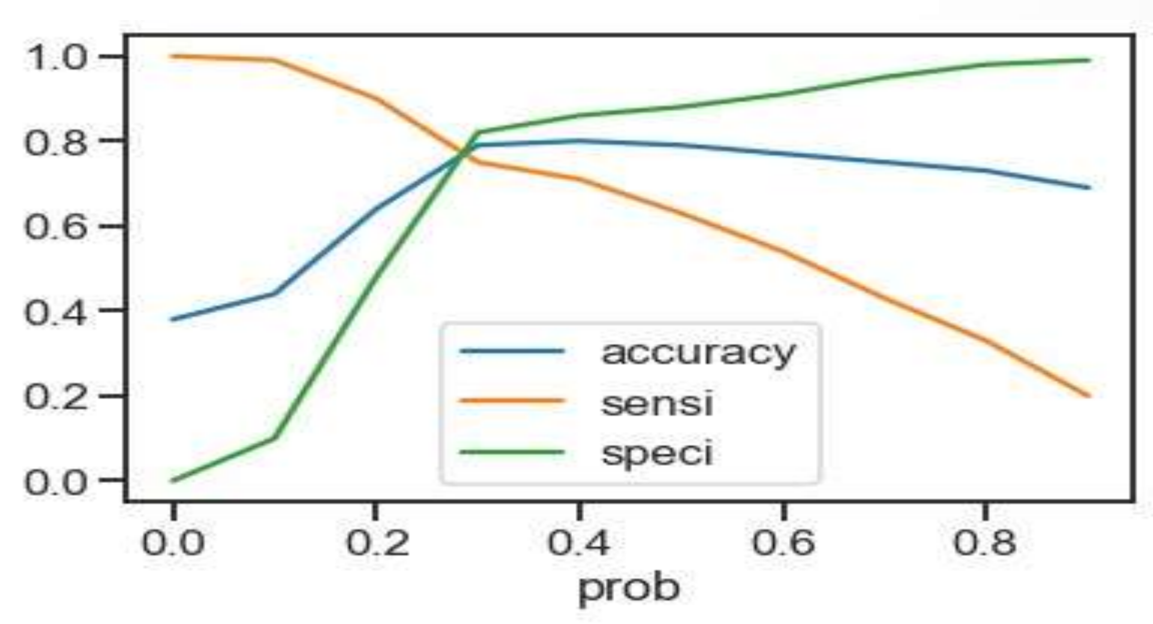
Model Building

- Splitting the Data into Training and Testing Sets
- The first basic step for regression is performing a train-test split, we have chosen 80:20 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 of the model is 79%

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 0.25





Conclusion

- The Accuracy, Precision and Recall score we got from test set are present in acceptable range.
- We have high recall score than precision score as per the requirement.
- Important features responsible for good conversion rate or the ones' which contributes more towards the probability of a lead getting converted are :
  1. Total Time Spent on Website
  2. Total visits to the website and
  3. Lead_origin_Lead Add Form
- When their current occupation is as a working professional, keeping these in mind the X Education can flourish as they have a very high chance to get almost all the potential buyers to change their mind and buy their courses.
- For the business terms, this model has an ability to adjust with the company’s requirements in coming future.