Lead Score Case Study

SUBMISSION DATE: 8TH DEC 2021

PREPARED BY RUPAL ACHARYYA & SINDURARUPA KARI

Problem Statement

Background

- **About The 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.

- **What This X Education Company Do**

  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.

- **What We Have To Do For This X Education Company OR What The X Education Company Is Expecting Form Us**

  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%.

Business Objective

To categorize or cluster the leads as hot leads and cold leads. This is to be done by analyzing past data provided by X Education Company. The Company's employee will then focus on communicating effectively with the hot leads so that most of them actually convert. We have been given a target of 80% conversion rate, and thus, in order to achieve that, we must accurately categorize the leads.

The Solution is divided into the following steps:

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

  And Train Set

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.
- □ **EDA**
  1. Univariate data analysis.
  2. Bivariate data analysis
  3. Check and handle outliers in data.
- □ **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.**

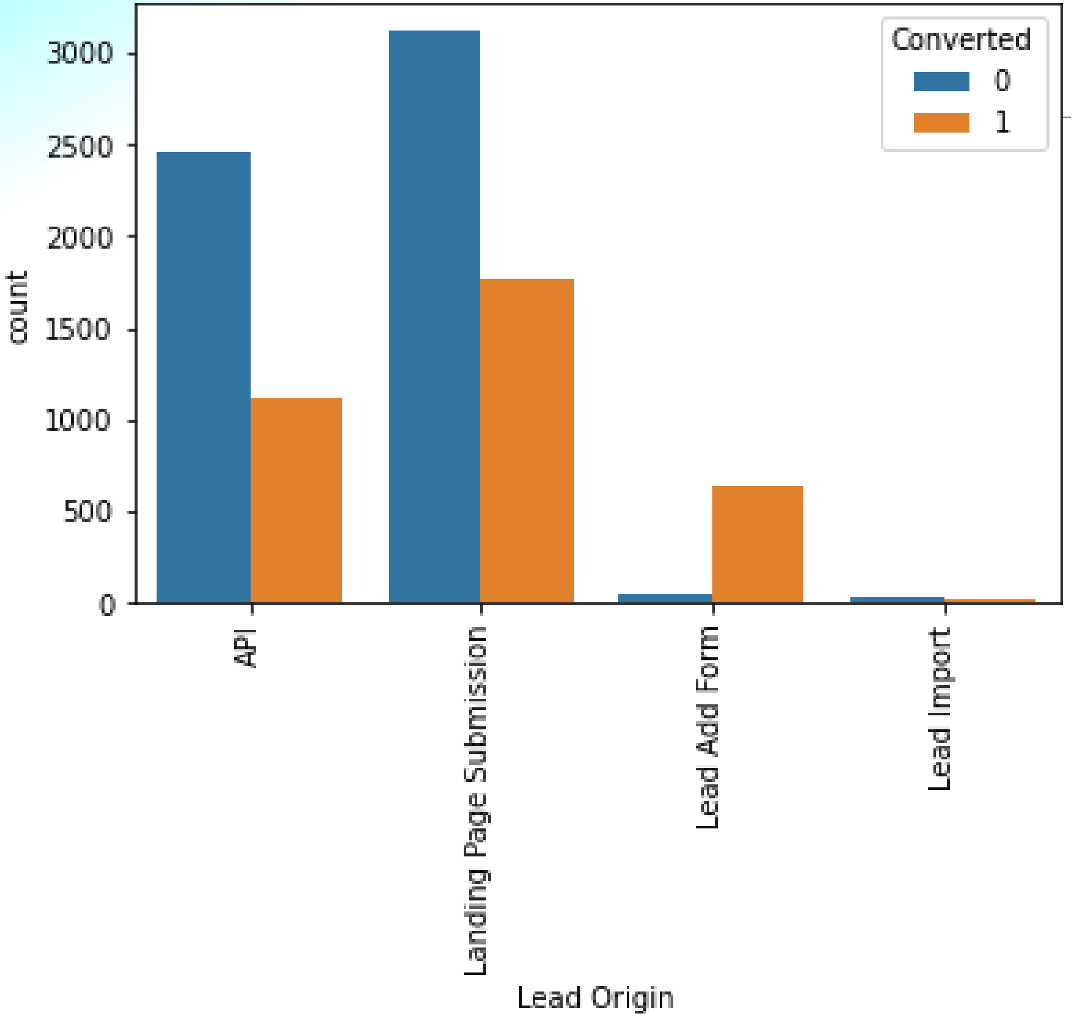
Data Insights

- Total Number of Rows =37, Total Number of Columns =9240.
- Single value features like 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.
- In some variance we replaced with mean and median based on the type of object
- We Dropped the columns having more than 40% as missing value such variables has droped

EDA
Univariate data analysis

upGrad

upGrad – IIITB & LJMU
Executive PG program in ML, AI
&
M.Sc Data Science – June 2021

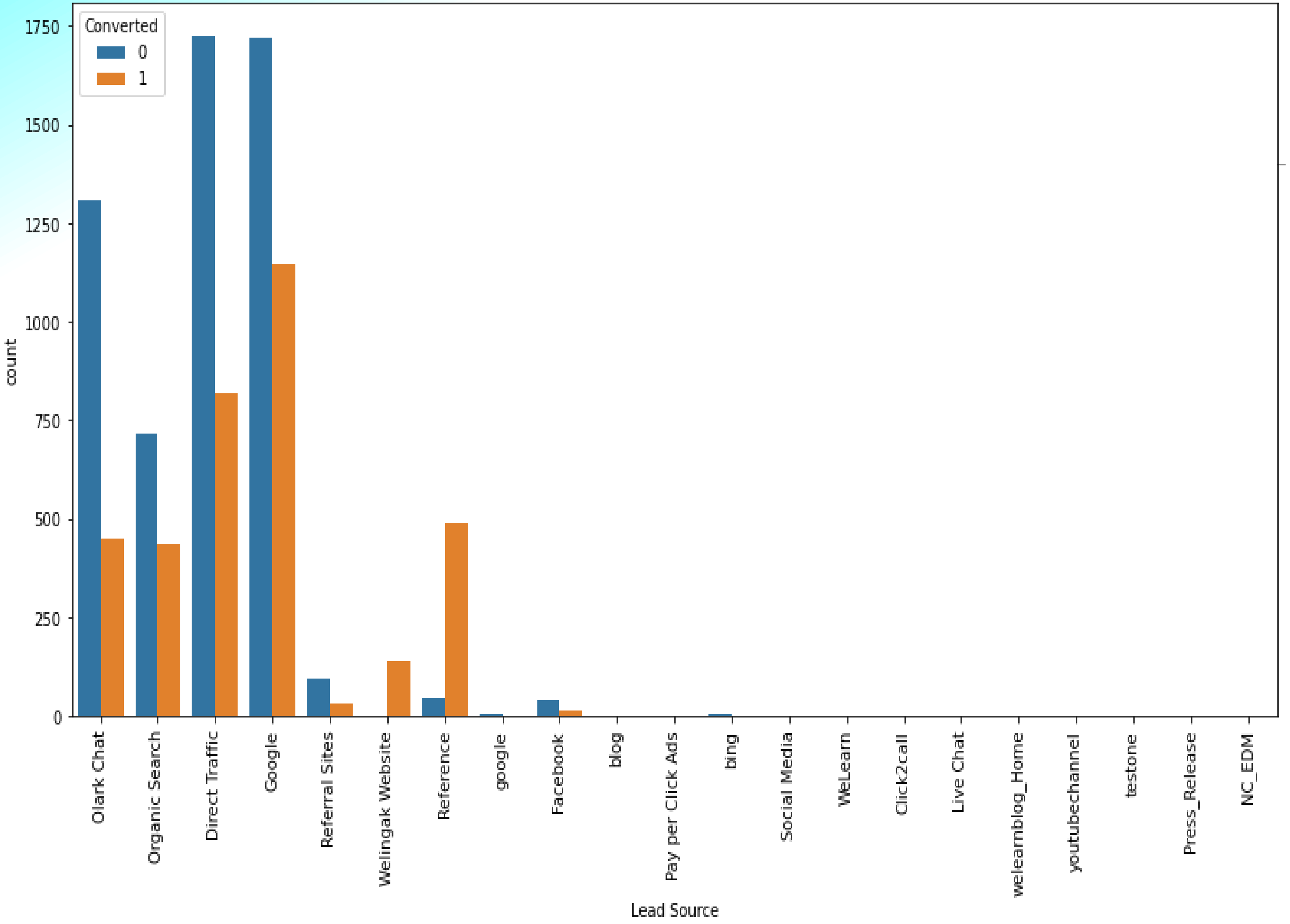


Inference:

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

Note: To improve overall lead conversion rate, we need to focus more on improving lead conversion of API and Landing Page Submission origin and generate more leads from Lead Add Form.

Univariate data analysis



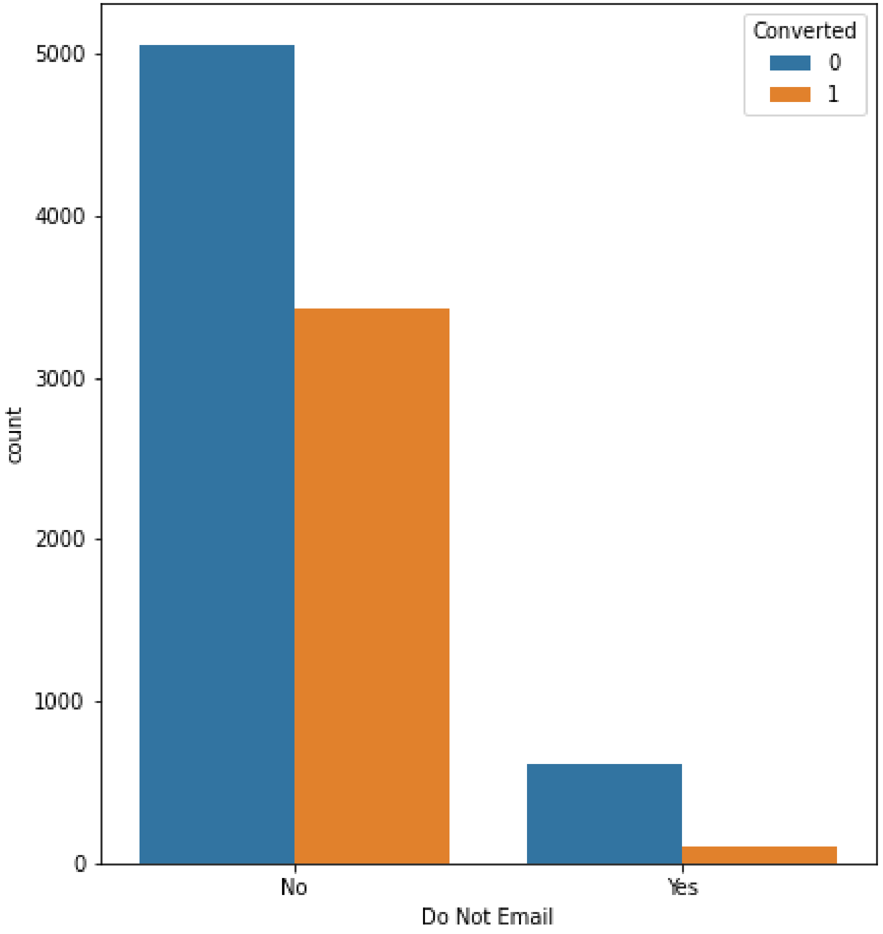
Inference:

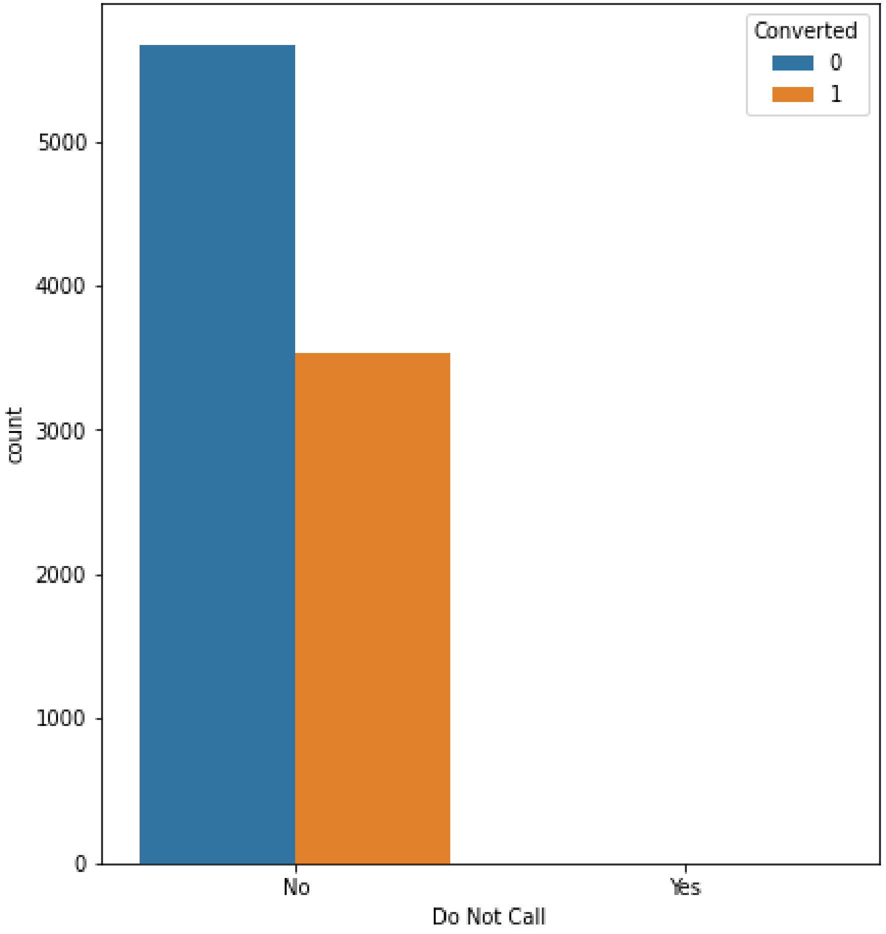
1. Google and Direct traffic generates maximum number of leads.

2. Conversion Rate of reference leads and leads through welingak website is high.

Note: To improve overall lead conversion rate, focus should be on improving lead conversion of olark chat, organic search, direct traffic, and google leads and generate more leads from reference and welingak website.

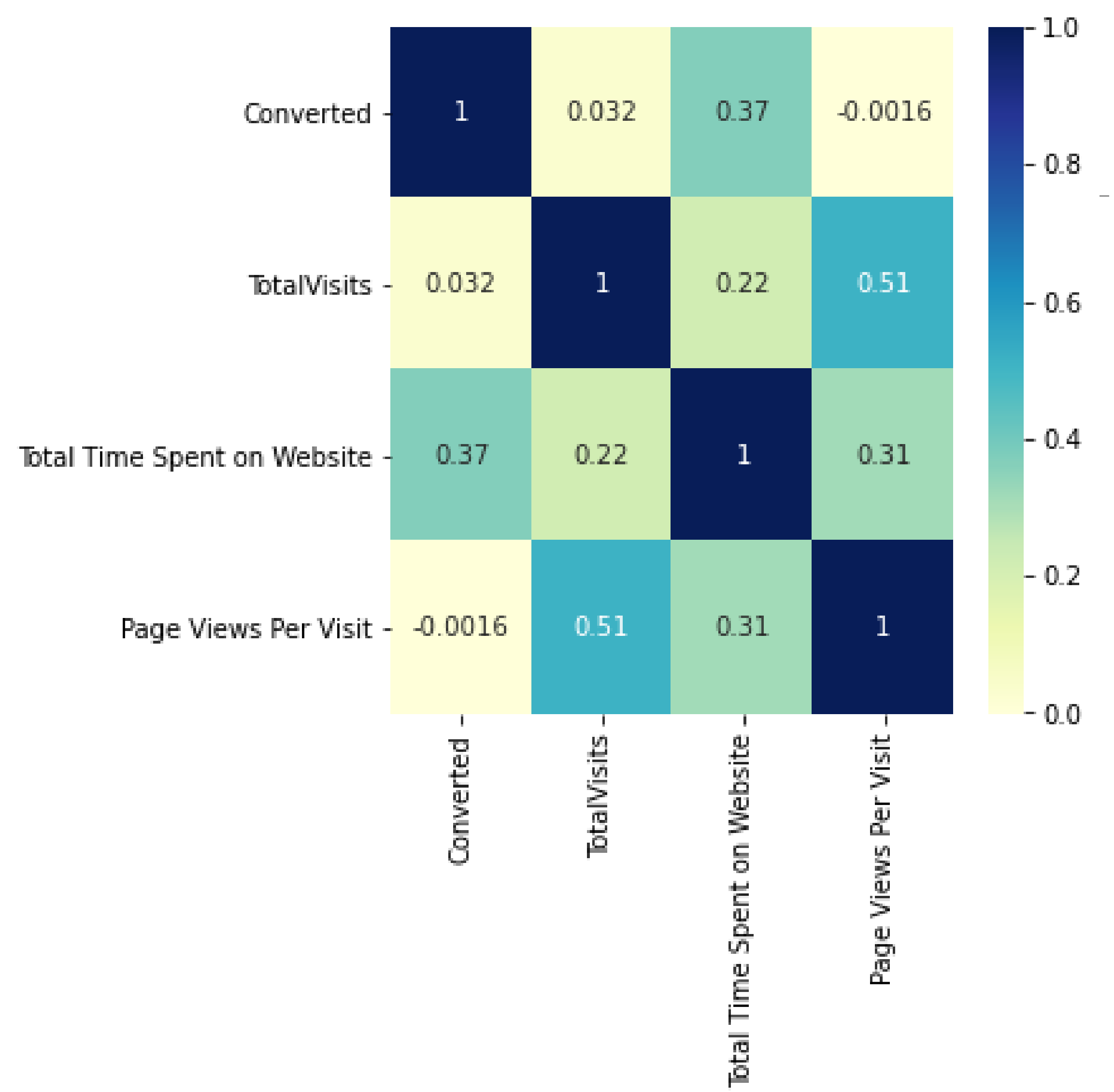
Bivariate Analysis





Inference: As we can see so many customers choose the option for Do Not Emails and Do Not Call

Correlation between all set of usable columns



Inference:

1. There is positive correlation between Total Time Spent on Website and Conversion

2. There is almost no correlation in Page Views Per Visit and Total Visits with Conversion

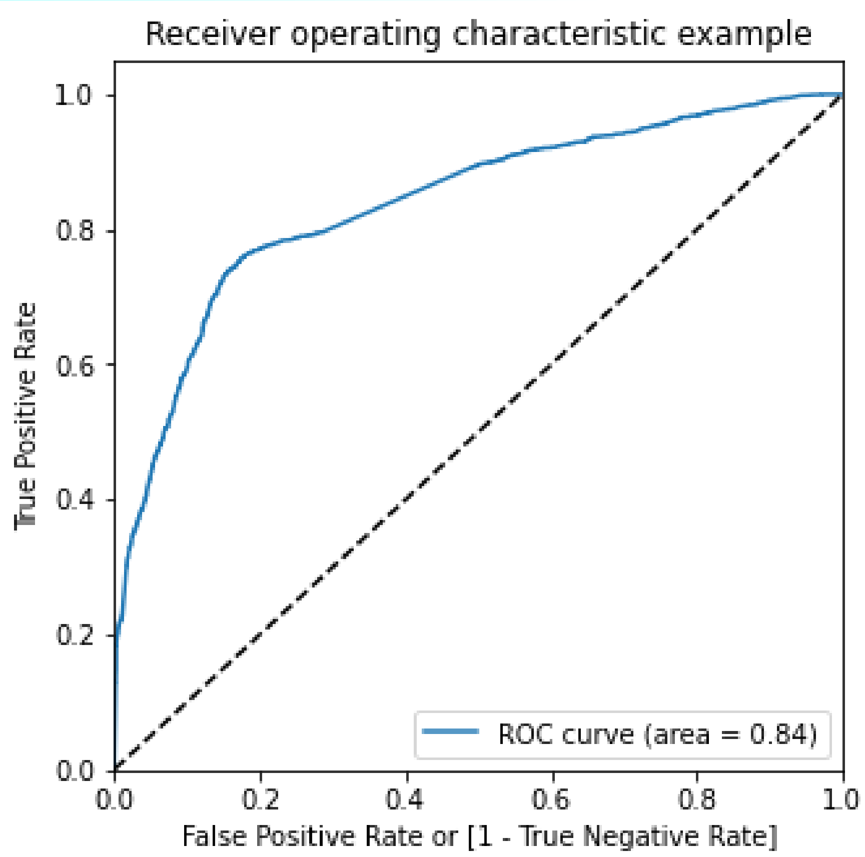
Data Conversion

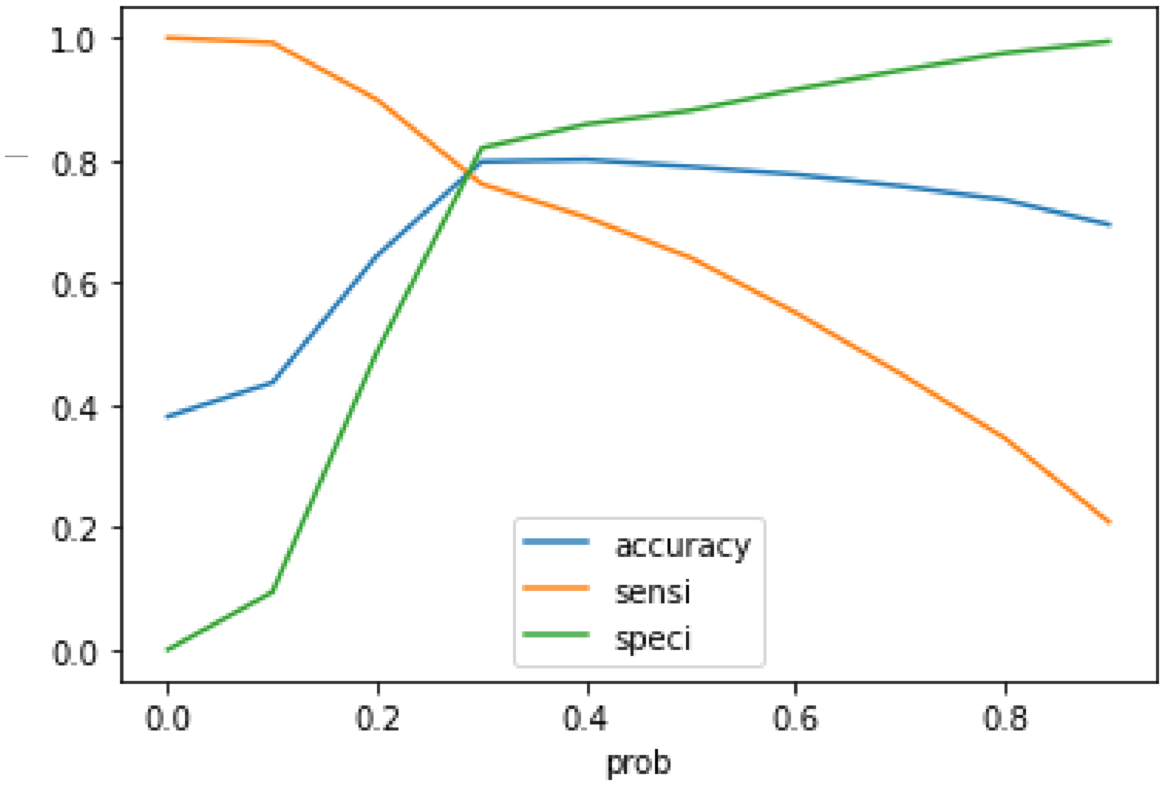
- Numerical Variables are Normalised
- Dummy Variables are created for object type variables
- Total Rows for Analysis: 9204
- Total Columns for Analysis: 56

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
- Approximate accuracy 80%

ROC Curve





- Finding Optimal Cut off Point
- Optimal cut off probability is that 0.84
- Probability where we get balanced sensitivity and specificity.
- From the second graph it is visible that the optimal cut off is at 0.3.

Conclusion

The variables that mattered the most in the potential buyers are

- The total time spend on the Website.
- Total number of visits.
- When the lead source was:
  - ✓ Google
  - ✓ Welingak website
- When the last activity was:
  - ✓ SMS
  - ✓ Olark chat conversation
- When the lead origin is Lead add format.
- 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.

Thank you