X-Education - Leads Scoring Case Study Assignment

upGrad & IIITB | Data Science Program - April 2023

Name	Mohd Sami Uzzaman / Agney Ravi O / Aditi Mishra
Batch	DS C55 IITB EPGDS

Lead Scoring Case Study

16-10-2023

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%.
Now, 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 Objectives

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 a higher lead score have a higher conversion chance and the customers with a 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%.

Goals of the Case Study

Build a logistic regression model to assign a lead score between 0 and 100 to each of the leads which can be used by the company to target potential leads. A higher score would mean that the lead is hot, i.e. is most likely to convert whereas a lower score would mean that the lead is cold and will mostly not get converted.
There are some more problems presented by the company which your model should be able to adjust to if the company's requirement changes in the future so you will need to handle these as well. These problems are provided in a separate doc file. Please fill it based on the logistic regression model you got in the first step. Also, make sure you include this in your final PPT where you'll make recommendations.

Summary of the step used for Assignment

Data Cleansing

1.Data set have total 9240 Rows and 37 Columns after dropped 35% null value columns data set have 9240 Rows and 31 Columns left And after removing unused column 12 Columns and 9240 rows left.

2.Select – There was 4 columns in which customer was not selected any option, we have consider as null and imputed accordingly

3.Outlier – 'TotalVisit' and 'Page Views Per Visit' both columns was having outlier hence treated as upper outliers at u_bound and the lower outliers at l_bound while removing the top and bottom 1% of the both columns values

EDA

1. **Data Imbalance** – Checked Data Imbalance on Conversion Columns

2.Univariate Analysis – Plotted count plot and Pair plot each Numerical and Categorical Chart for Analysis

3. **BiVariate Analysis** - Plotted each 2 Features Chart for Analysis

Data Preparation

1.Dummy Variable – Created Dummy for all Text Variable which was having more than 2 categories

2.Train, Test Split – Splited the Data into Train and Test data set by using Train Test Split Method

3.Feature Scaling – Scale the Numerical Data set using MinMax Scaler

Model Building

1.Using RFE Selected Random 15 Feature for Model Building

2.Fitted the Logestice Regression Model into Train and Test Data

3.Checked P Value

4.Plotted Confusion Metrics

Model Evaluation

1.Plotted ROC Curve to to find ROC Area

2.Plotted Accuracy, sensitivity and specificity for various probabilities

3.Plotted Confusion metrics using Cuttoff

Prediction on the Test data

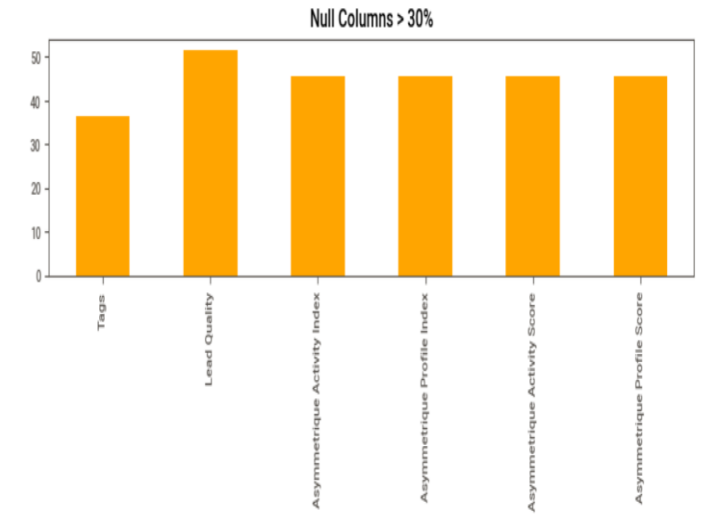
1.Used the Model into Test Data set

2.Checked the Accuracy, Sensitivity and Specificity for bot Train and Test Data

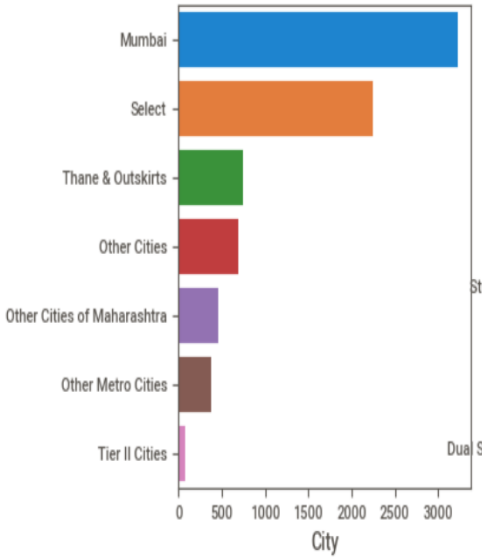
3.Get the prediction

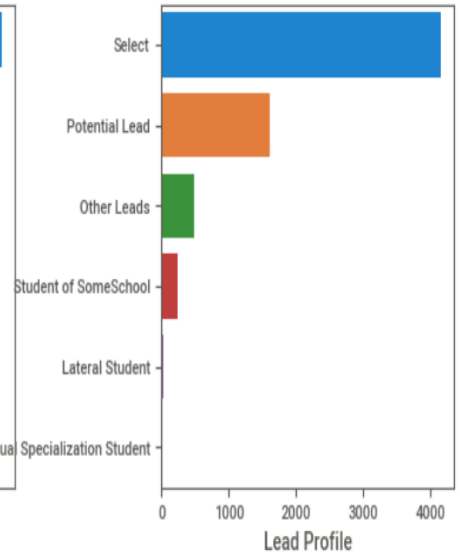
Data Cleansing Summary

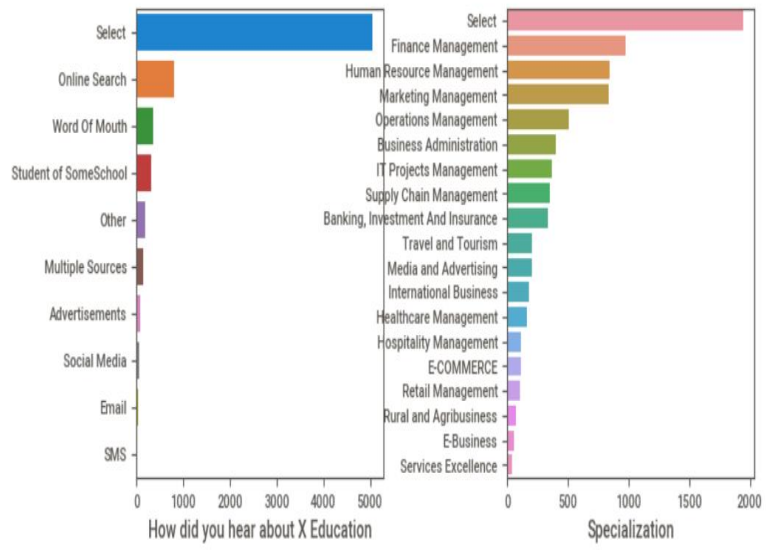
Missing Value Above 30% Columns column



Select – Client has not Selected any Category hence considered as Nan and imputed with EDA Method



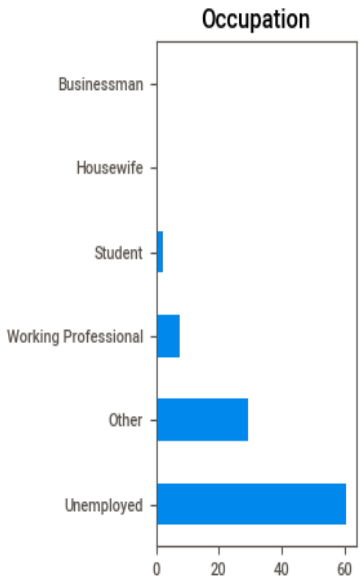


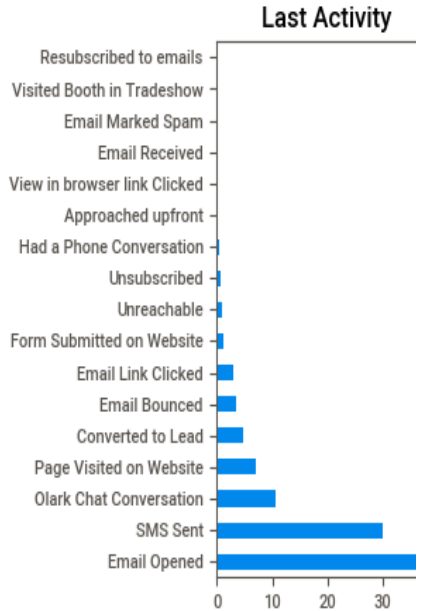


Summary:-

- Leads Data set was having 37 Col and 9240 Rows
- Removed 30% above 6 Missing Value columns
- In 4 Columns Client was not selected category hence considered as Nan
- Removed Unused Columns
- Total 12 important Columns and 924 Rows Left after droping all unused ar missing values columns
- Imputed all Missing values which was Important features with New Creatin New category or imputed using EDA Methods

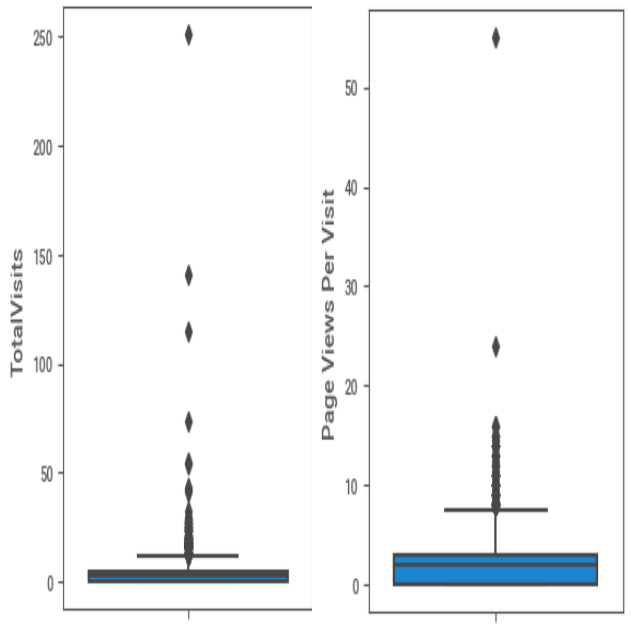
Missing Values– Imputed Missing Values into the given columns using EDA methods



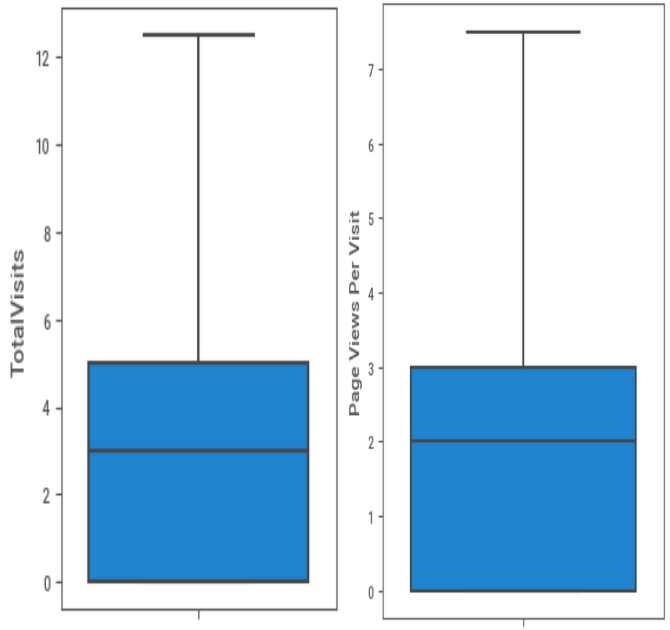


Detect Outlier

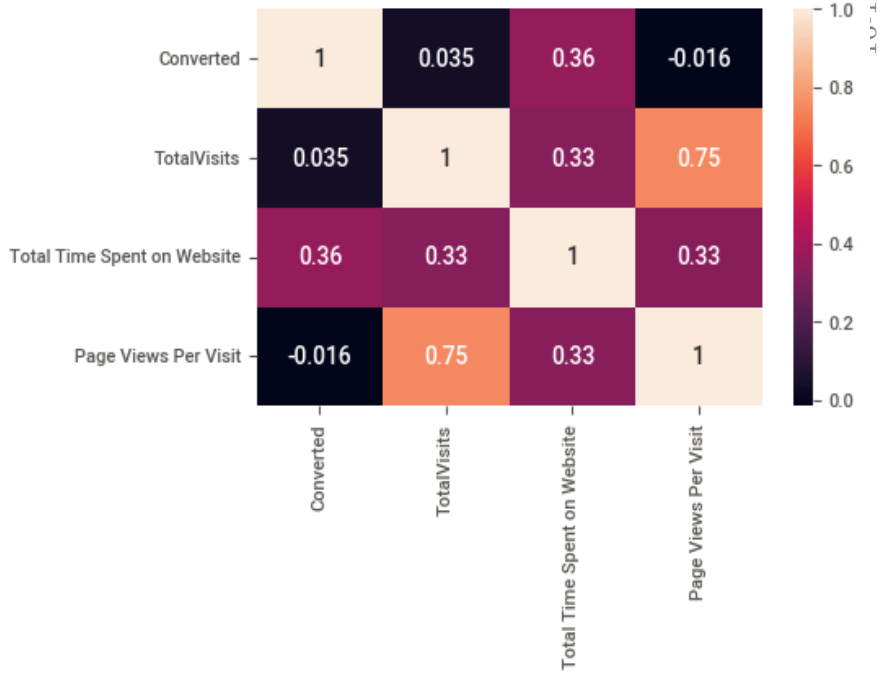
Before Imputation



After Imputation



Checked Correlation Metrics



Summary

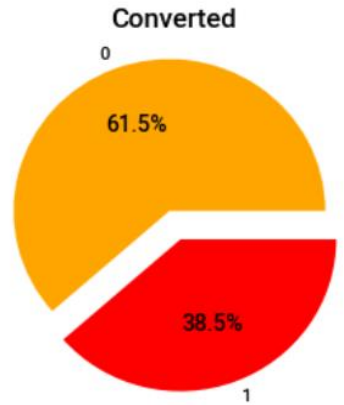
Outlier - We have found outlier in given 2 columns, both columns hence treated as upper outliers at u_bound and the lower outliers at l_bound while removed the top and bottom 1% of the both columns values.

Data Imbalance – We have found 61.6 % Data Imbalance while plotting pie chart on conversion column.

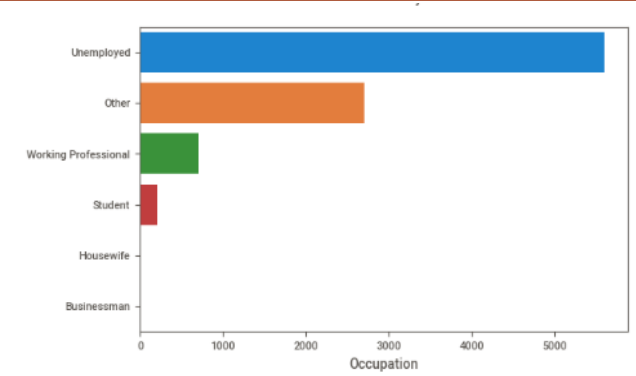
Correlation – Checked Correlation on numeric columns, we have find highest correlation between TotalVisits and Converted column

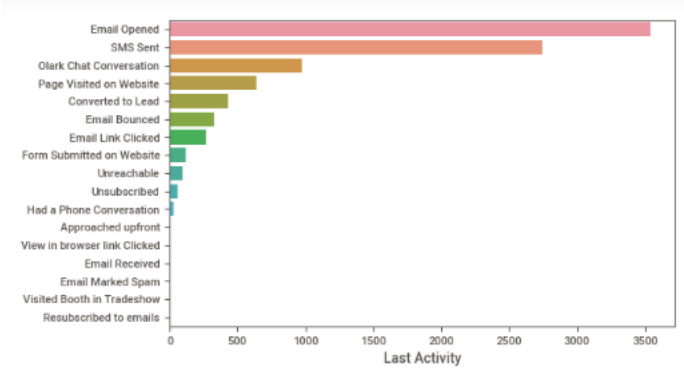
Data Imbalance check on Converted Variable

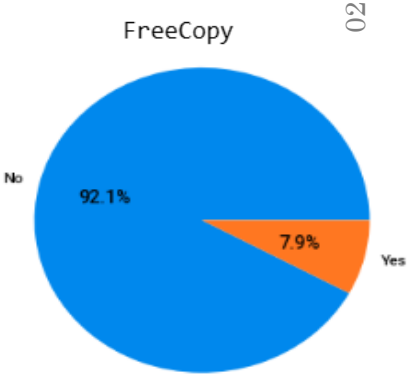
Conveted : 38.5%
Not Conveted : 61.5%

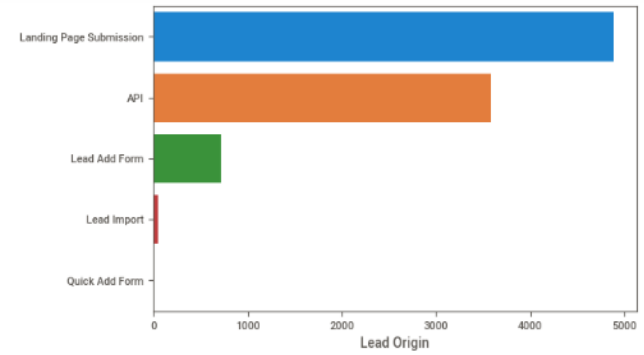


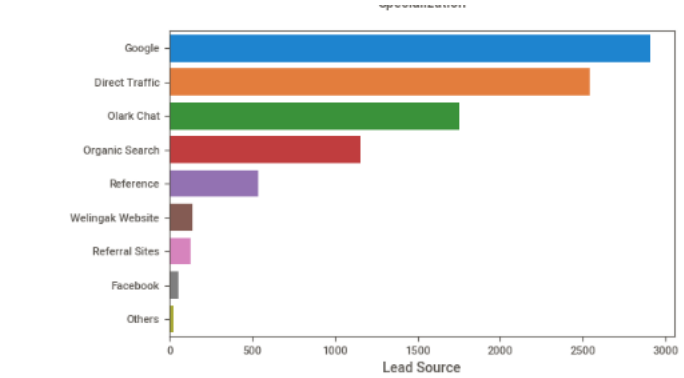
Univariate Analysis

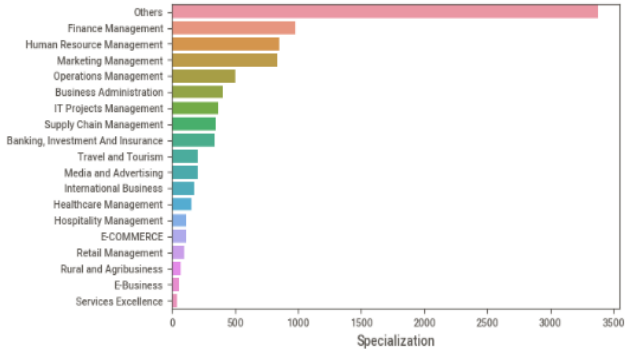


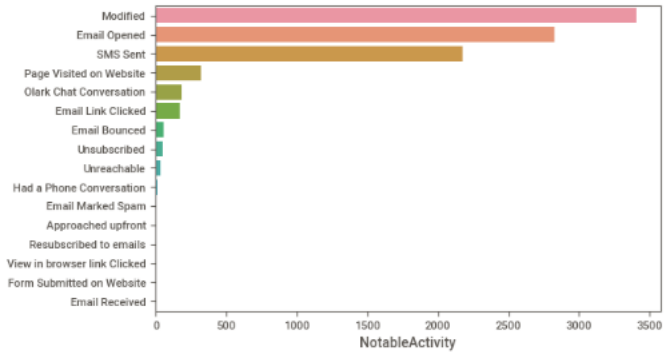












Univariate Analysis Summary

- **Do Not Email** – 68.7% client has asked for not to email wherein 31.3% client is interested to send email
- **Free Copy** – Approx. 8% client is interested to share Free copy
- **Occupation** – Compare to Student Unemployed
- **Last Activity** – Email Opened and SMS received % client are very high compare to Others
- **Lead Origin** – Landing Page Submission clients are high
- **Lead Source** – Most Numbers of client visited through Google and the very less from others source
- **Specialization** – Mostly client visited for Finance, HR, Marketing, however large customer has not selected any category

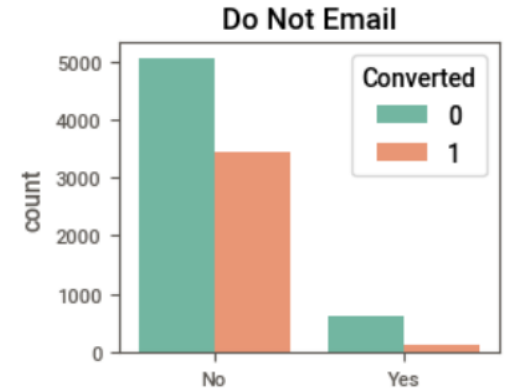
Observation :-

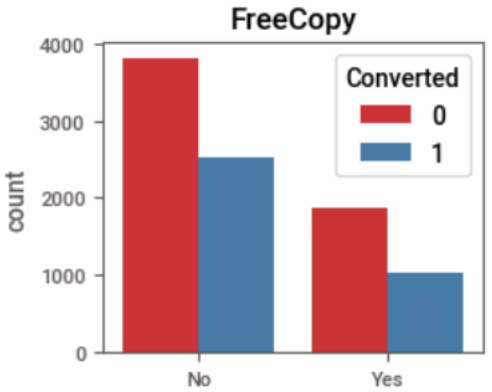
- **As We can See most of the clients visited through the advertising using google and Direct Traffic hence they should increase Online Advertising**
- **Unemployed client is high hence we should focus on Unemployed clients**
- **Finance, HR and Marketing specialization clients are very high, we should firs focus on these clients**

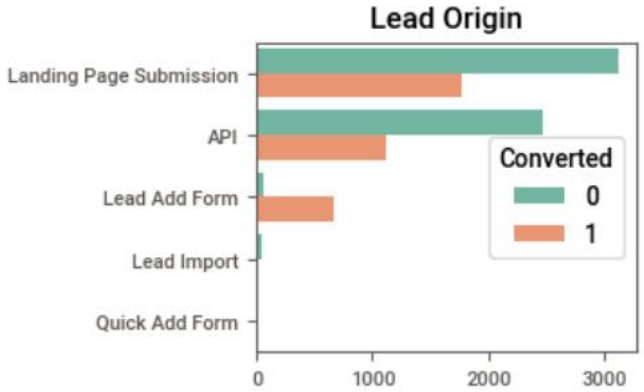
Bivariate Analysis

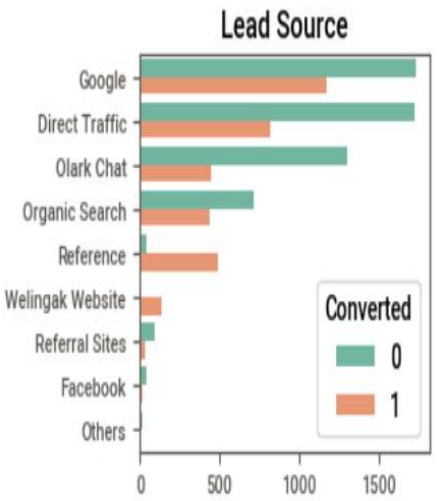
Bivariate Analysis Summary

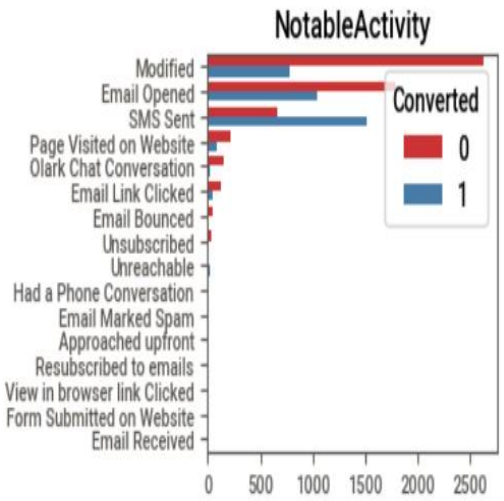
- **Do Not Email** – Lead conversion for the Client who has asked for mail is highest conversion rate wherein not to email has very less conversion
- **Lead Source**-Mostly client wo has visited through online source such as google has heighest conversion rate. Company should focus on advertising in online portal
- **Specialization** –Finance, HR and Marketing Specialization conversion is high chances of lead conversion hence we should focus on this area
- **Occupation** –Unemployed client has heighest conversion rate where in Business occupation no conversion rate. Company should also target on Student and Working professional as working professional have very high conversion chances companion to others











16-10-2023

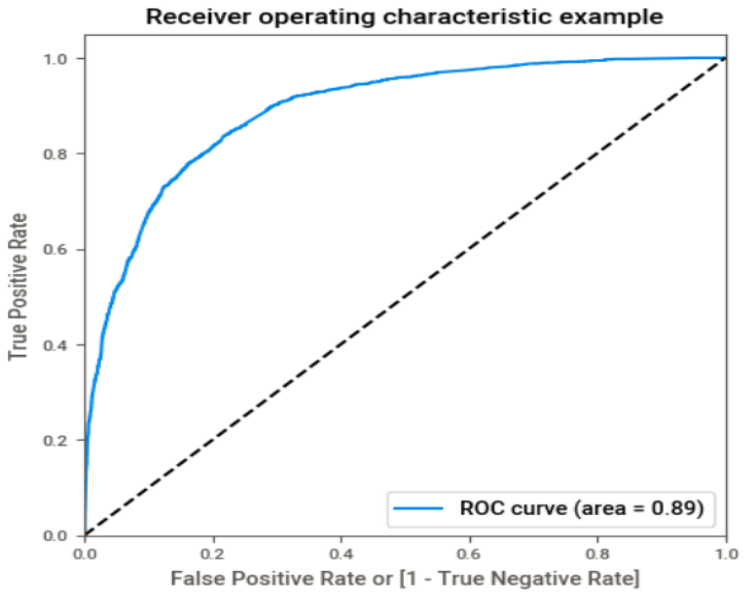
Data Preparation for Modeling

- Data Preparation for Modeling - Created dummy variable for more than 2 category columns
- Dropped Columns - After Creating Dummy variable dropped the dummy variable columns
- Train Test Split - Splitted Data into Train and Test by takin 70 % and 30% size
- Feature Scaling - Scale the numeric variable using MinMax scaler methods

Model Building

- REF - Used RFE method to get the top 15 important feature selection as the data was having more that 70 columns it was very difficult to choose important feature
- Fitting Logestic Regression - After added a constant into train data fitted a logestic regression model
- Checked P Value - Checked the P value in which features was having more that 0.05 P Value
- Dropped Columns - Dropped Features which was having > 0.05 P Value and build the model till all the features P Value converted to less then 0.05.

Model Evaluation



- Calculated accuracy sensitivity and specificity for both plot and get the final cut-off as 0.35
- The area under the ROC curve is 0.89 which is indicating that we have a good model.
- Sensitivity of Test set is 81.00% and Train set is 80.17% using cutoff 0.35
- X Education company set target of lead conversion rate to be around 80% and the model is also giving 81.38% which is meeting the objective

Test Data:

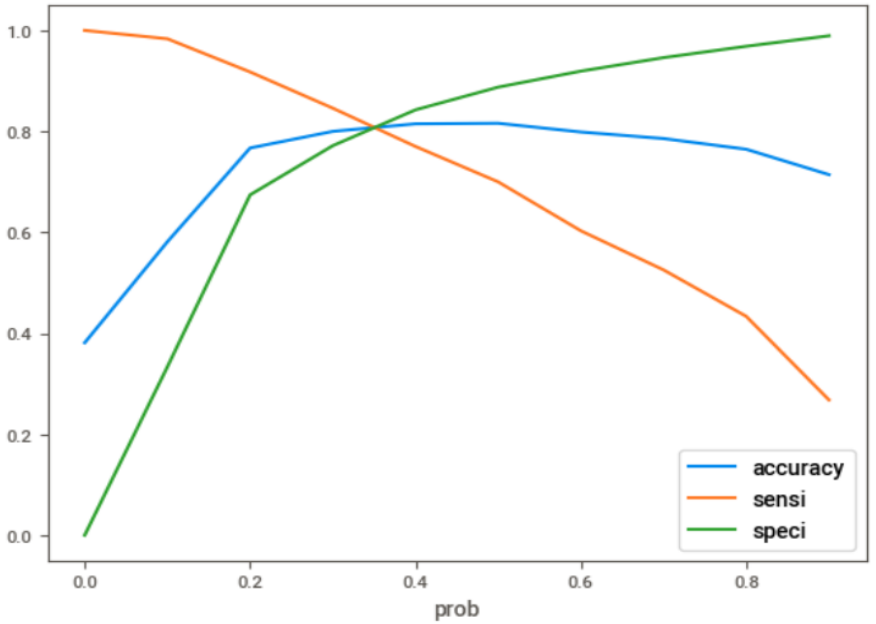
```
Accuracy    : 81.38%
Sensitivity : 81.00%
Specificity : 81.63%
```

Train Data:

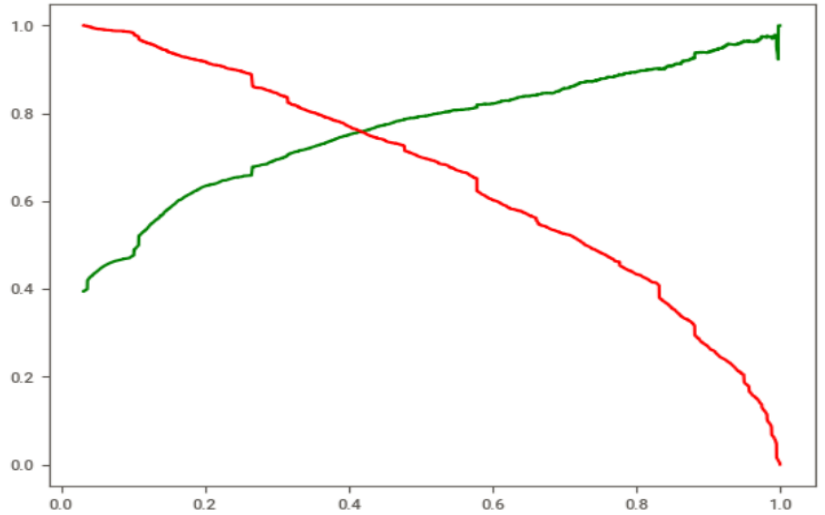
```
Accuracy    : 80.76%
Sensitivity : 80.17%
Specificity : 81.13%
```

Note: This model is reflecting very close to each others. This indicates that the model is performing consistently across different evaluation metrics in both test and train dataset.

Train Data Accuracy, sensi and Speci Prob



Test Data Accuracy, sensi and Speci Prob



X-Education - Leads Scoring Analytical Insights

- As per problem statement we have build the logistic regression model that can help in identify the most significant factor and help in increasing the lead conversion

- Company should focus on the given features as these feature is having highest coefficient and that is indicate the hot lead:
  - > TimeSpent
  - > NotableActivity_Had a Phone Conversation
  - > Lead Origin_Lead Add Form
  - > Lead Source_Welingak Website

- There are also negate coefficient wich company need to analyze:
  - > Do Not Email
  - > Last Activity_Olark Chat Conversation
  - > Occupation_Other
  - > Lead Origin_Landing Page Submission

- Company should more focus on advertising through various source

- Focus on positive coefficient, making more calls to convert.

Thanks & Regards,
Mohd Sami Uzzaman
Agney Ravi O
Aditi Mishra

16-10-2023