LEAD SCORING CASE STUDY

BY-
SUVARTA BHAVESH MARU
SHASHANT RATHORE
GAURI AGARWAL

TABLE OF CONTENTS:

BACK GROUND- 'X EDUCATION'

- **X Education'** is an education company which sells online courses to industry professionals.
- The company advertises its courses on several websites like Google.
- When the interested professionals lands on their website they browse the courses or watch the videos related to their interest.
- Once these people fill up a form providing their e-mail address or phone number are classified as a '**lead**'.
- When leads are acquired, employees from the sales team reaches out to them through various methods i.e. e-mails, personal calls, etc.
- Through this process, some of the leads gets converted and mostly do not.

PROBLEM STATEMENT AND GOAL

PROBLEM STATEMENT:

- X education gets a lot of leads but its lead conversion rate is very poor.
- To make this process more efficacious, the company wants to identify the most potential leads, also known as '**Hot Leads**'.
- Our tasks is to give them most promising leads. We are required to build a model where in we need to assign a lead score to each of the leads such that the customers with high conversion lead score have a high conversion chance and vice-versa.
- The CEO has given a ballpark of the target lead conversion rate to be 80%

BUSINESS GOAL:

- Building a logistic regression model to assign a lead score between 0 and 100 to each of the leads.
- A lead with higher score would mean Hot Lead is most likely to convert and lower score lead would be the cold and will mostly not get converted.
- There are some more problems presented by the company with our model they should be able to adjust to if the company's requirement changes in the future

APPROACH

01. Data cleaning and imputing missing values

02. EDA: Bivariate and Univariate analysis.

03. Feature Scaling and creating dummy variables.

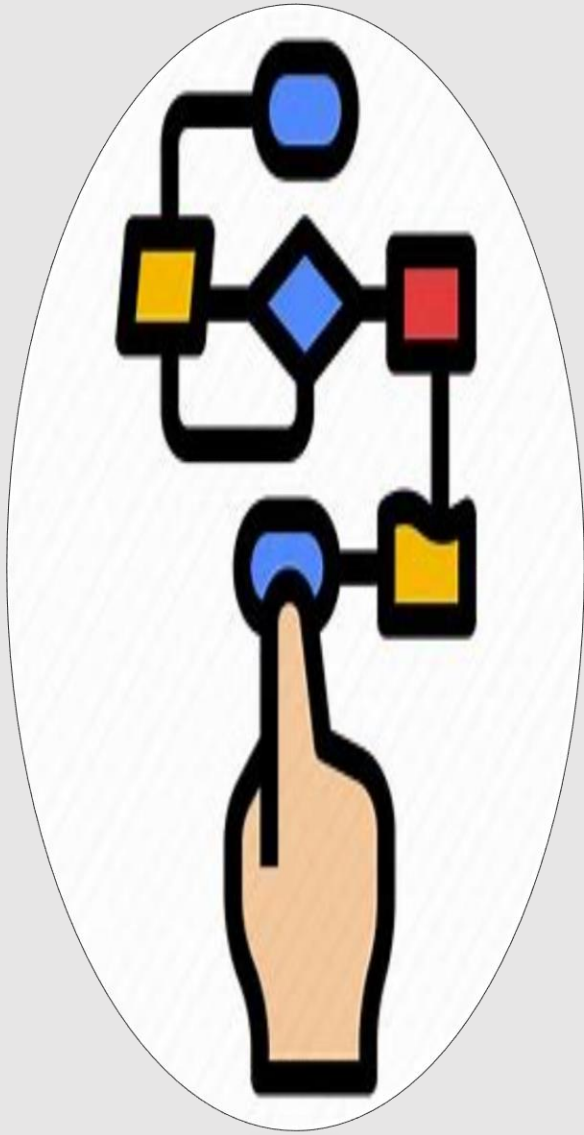
04. Model Building: Logistic Regression

05. Model Evaluation:
Accuracy,
Specificity,
Sensitivity,
Precision and recall.

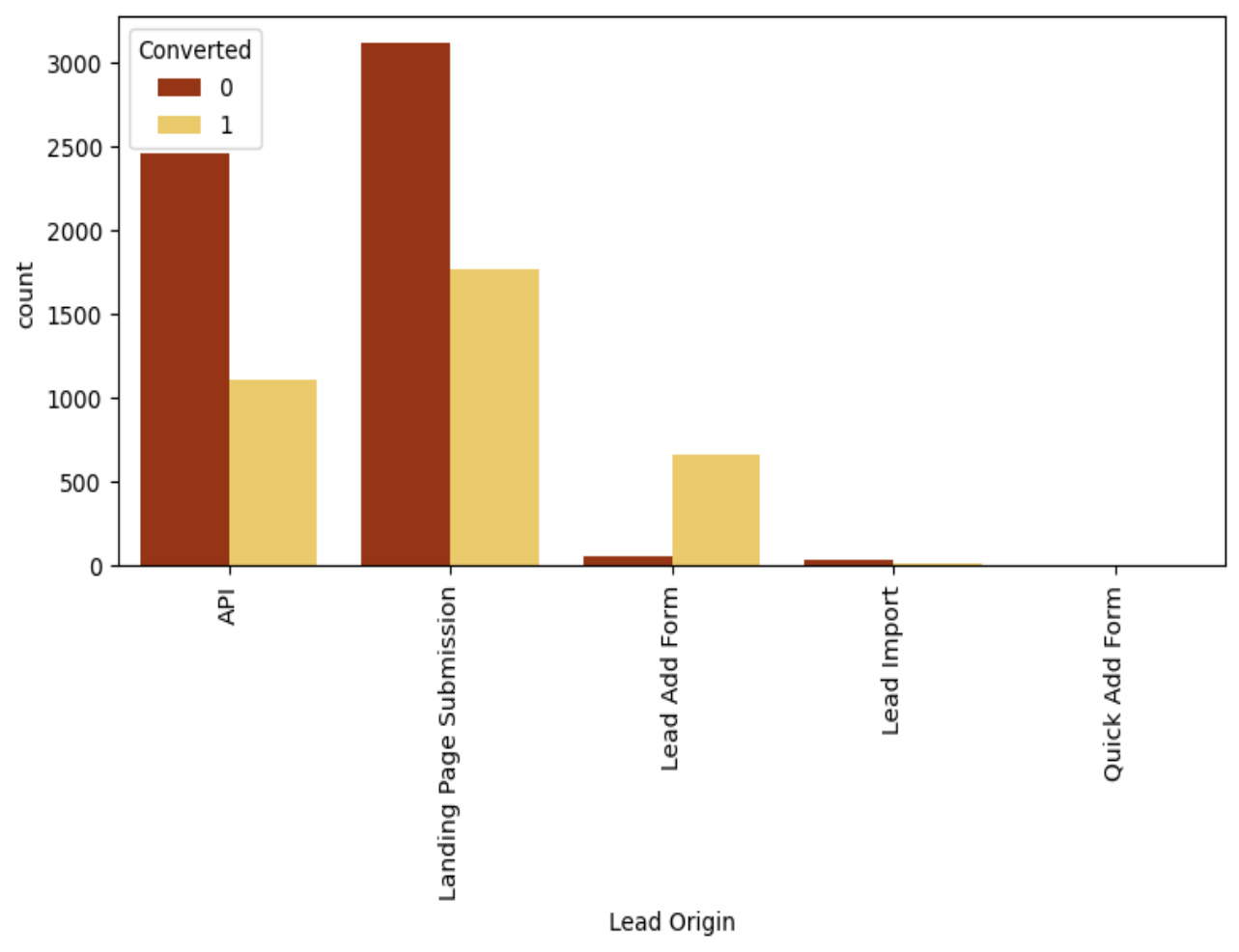
06. Conclusion and recommendation.

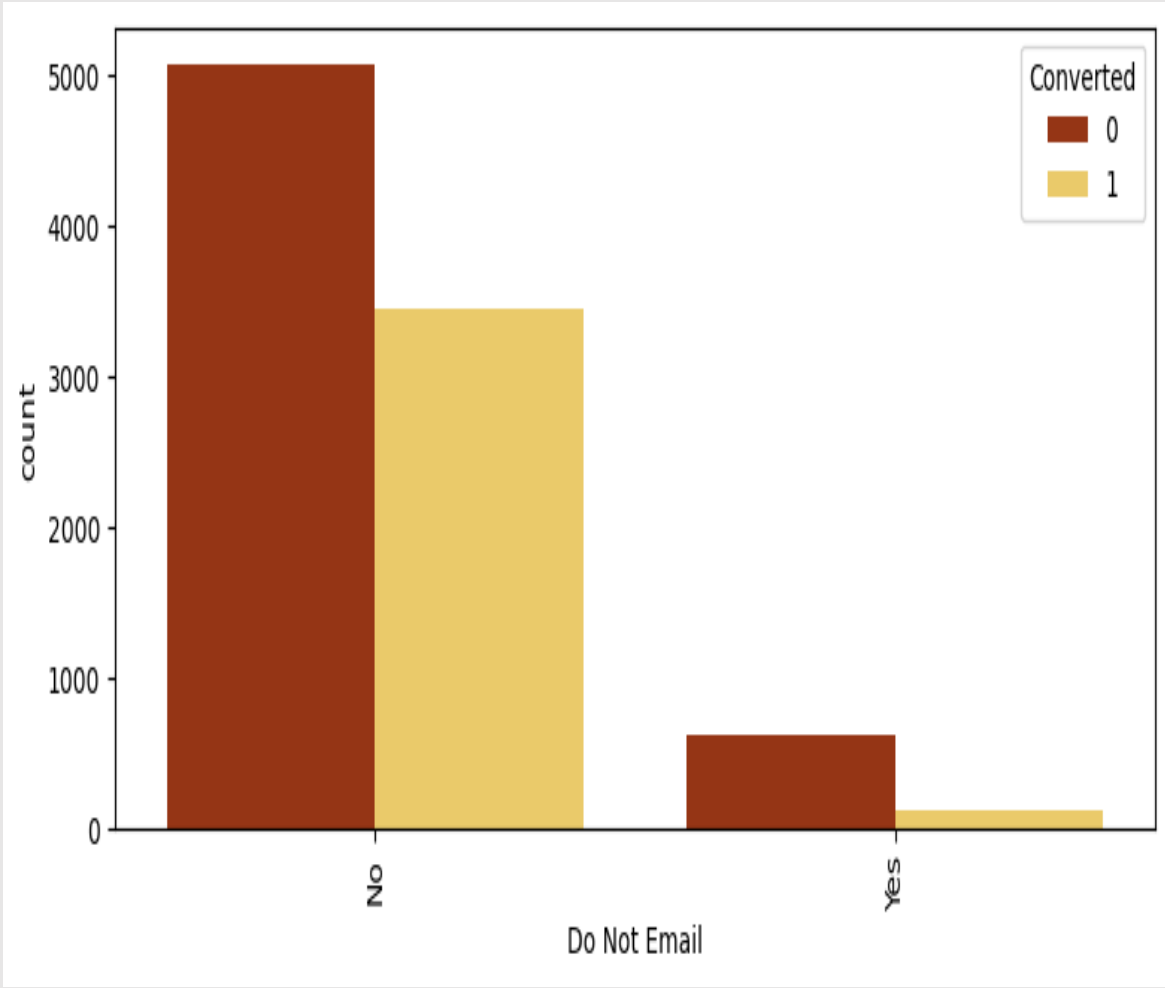
FLOW OF PROCESS

- **Data Cleaning and preparation**
  - Cleaning and removing redundant data.
  - Duplicate data removal.
  - Outlier treatment
  - Exploratory Data Analysis
- **Feature Scaling and Diving the data**
  - Dividing the data into train and test data set.
  - Feature scaling of numerical variables.
- **Model Buiding**
  - Feature selection using RFE, VIF and p-value
  - Determining optimal model using Logistic regression
  - Calculating various evaluation metric
- **Result**
  - Determining the lead score and check if target final prediction greater than 80%
  - Evaluating the final prediction on test set.



EXPLORATORY DATA ANALYSIS

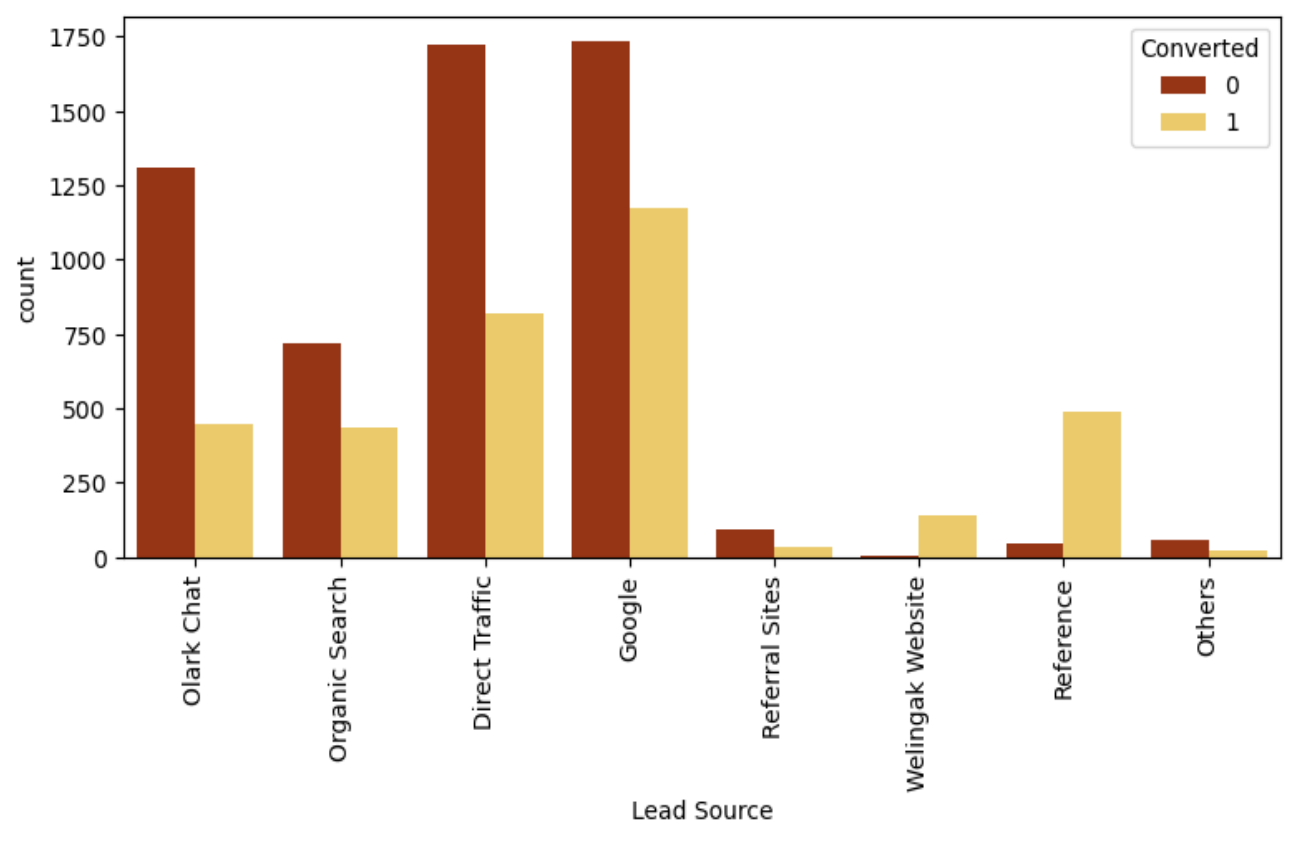


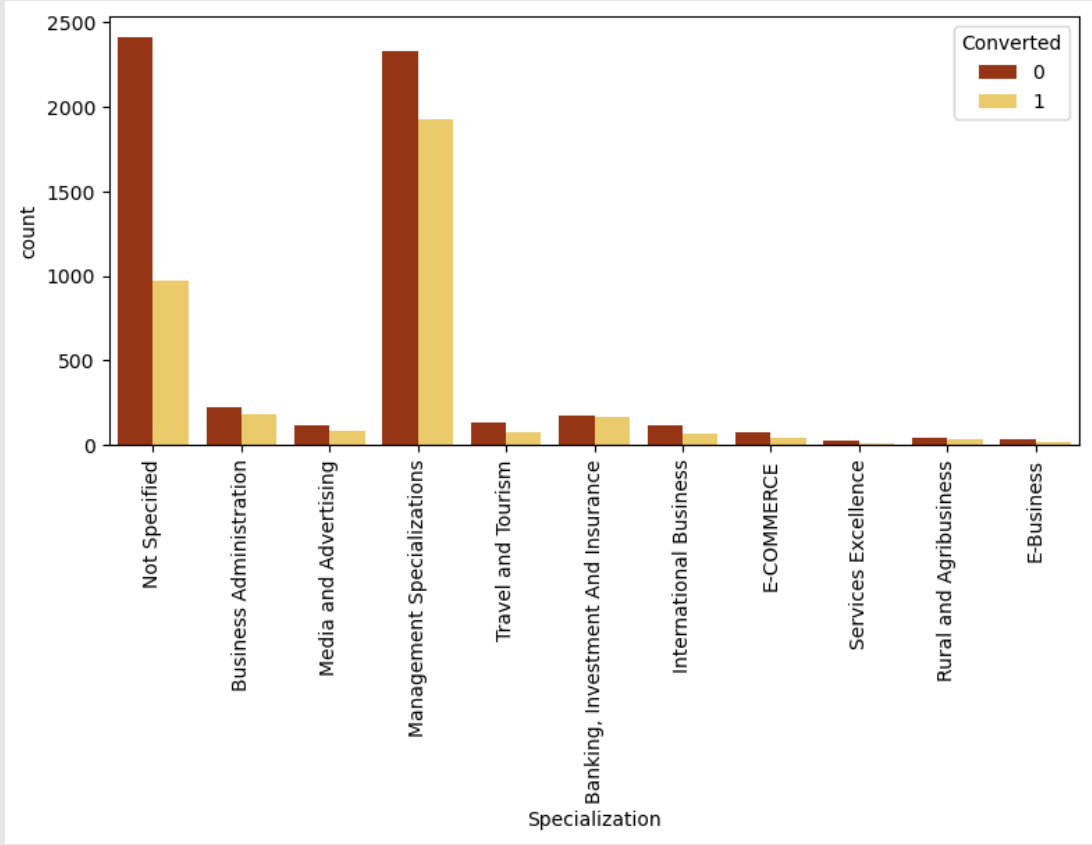


Inferences drawn from "Lead Origin" and "Do not Email"

- The highest conversion rate is from customers of Lead Add Form as compared to other Lead origins.
- People who do not want to receive emails and people who want to receive emails both have negative conversion rates.

EXPLORATORY DATA ANALYSIS

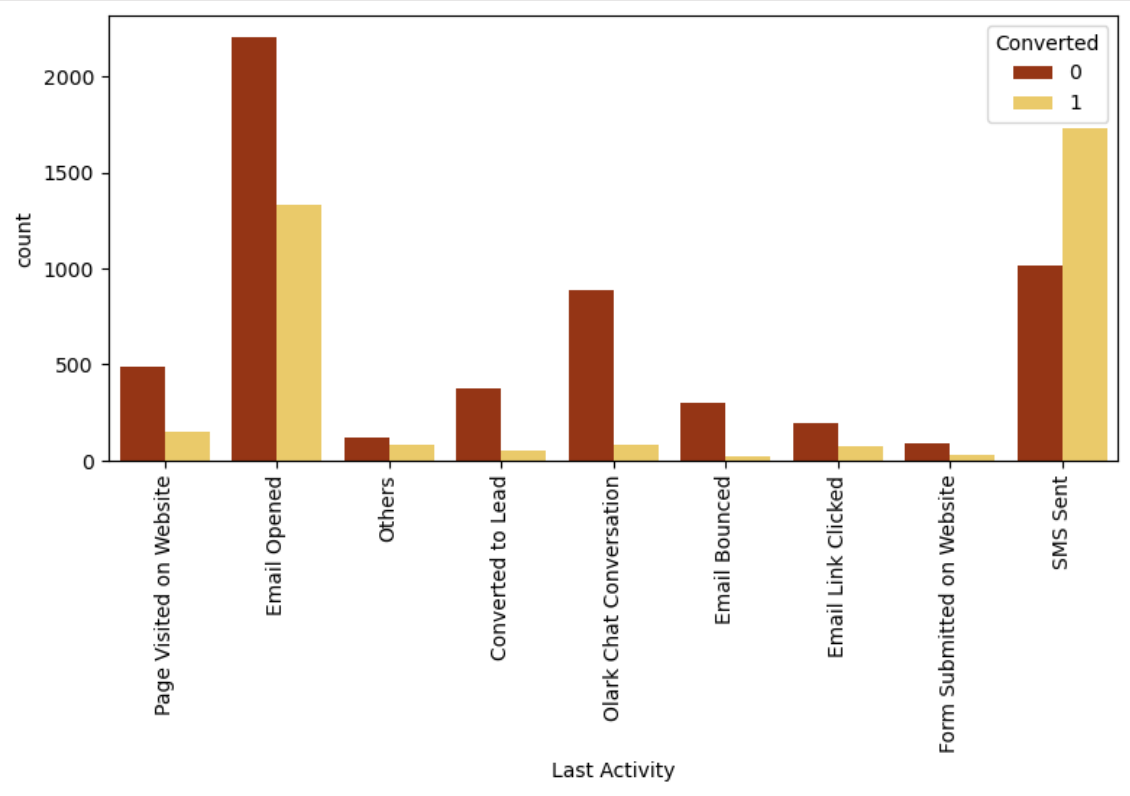


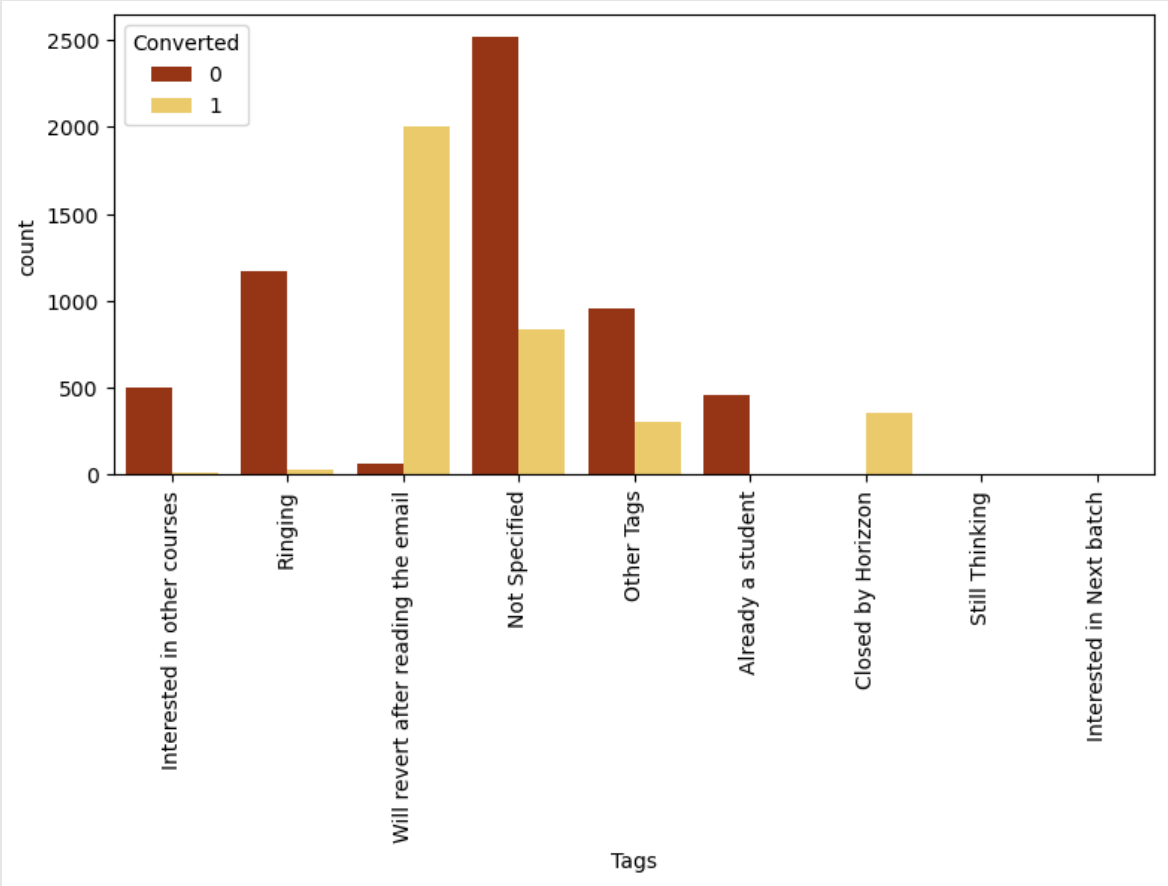


Inferences drawn from Lead Source and Specialization

- Lowest positive conversion rate is of customers who specializes in Services Excellence.
- Highest positive conversion rate is of customers who specializes in Management
- Customer whose specialization is not specified also have lower positive rate
- Direct traffic and google have negative conversion rate.
- Reference shows good positive conversion rate

EXPLORATORY DATA ANALYSIS





Inferences drawn from Last Activity and Tags:

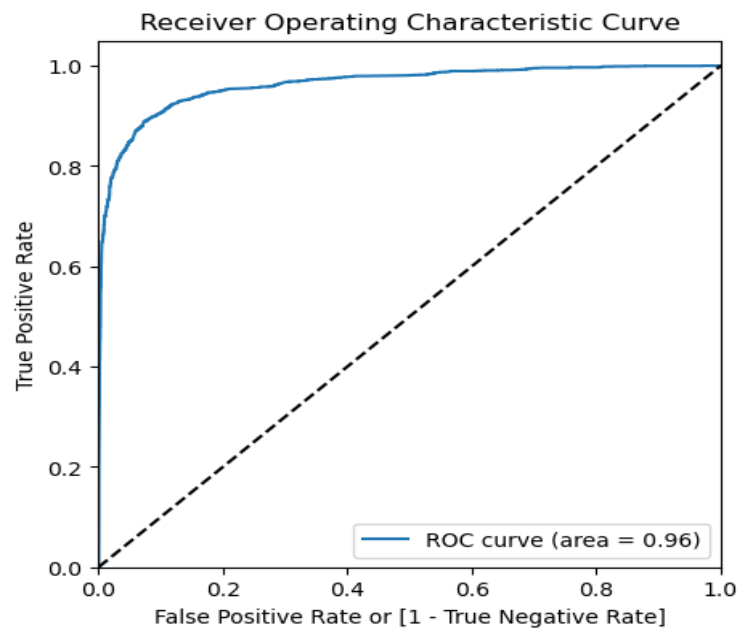
- Last activity performed by customers is SMS sent , since it has the highest positive conversion rate
- Negative conversion rate is of tags 'Ringing', 'Not specified' and 'Other tags'.
- Positive conversion rate is of tags 'Will revert after reading the email' and 'Closed by Horizon'

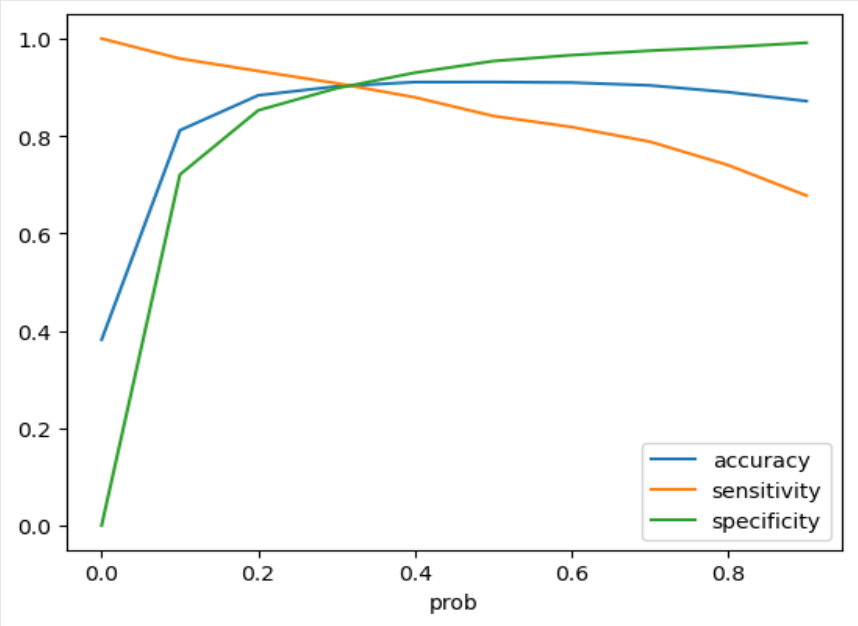
NUMERICAL VARIABLE ANALYSIS

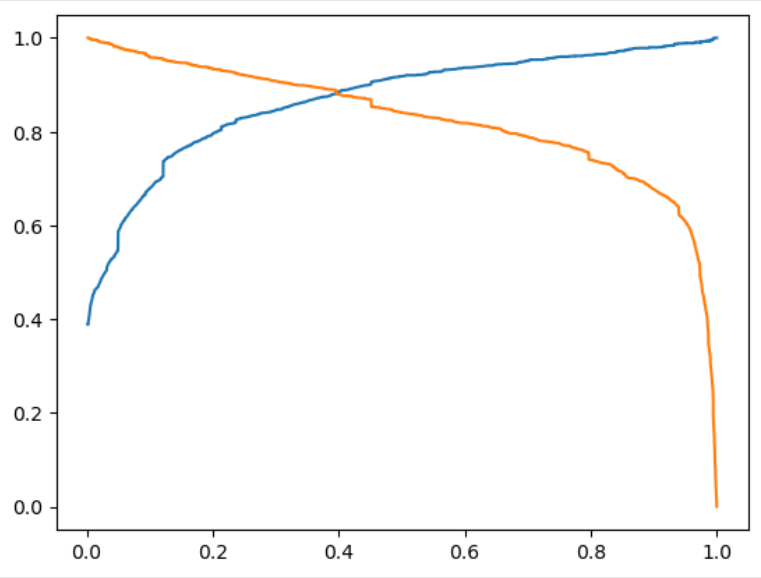
INFERENCE:

The target variable that is Converted and Page Views Per Visit have negative correlation.

MODEL BUILDING

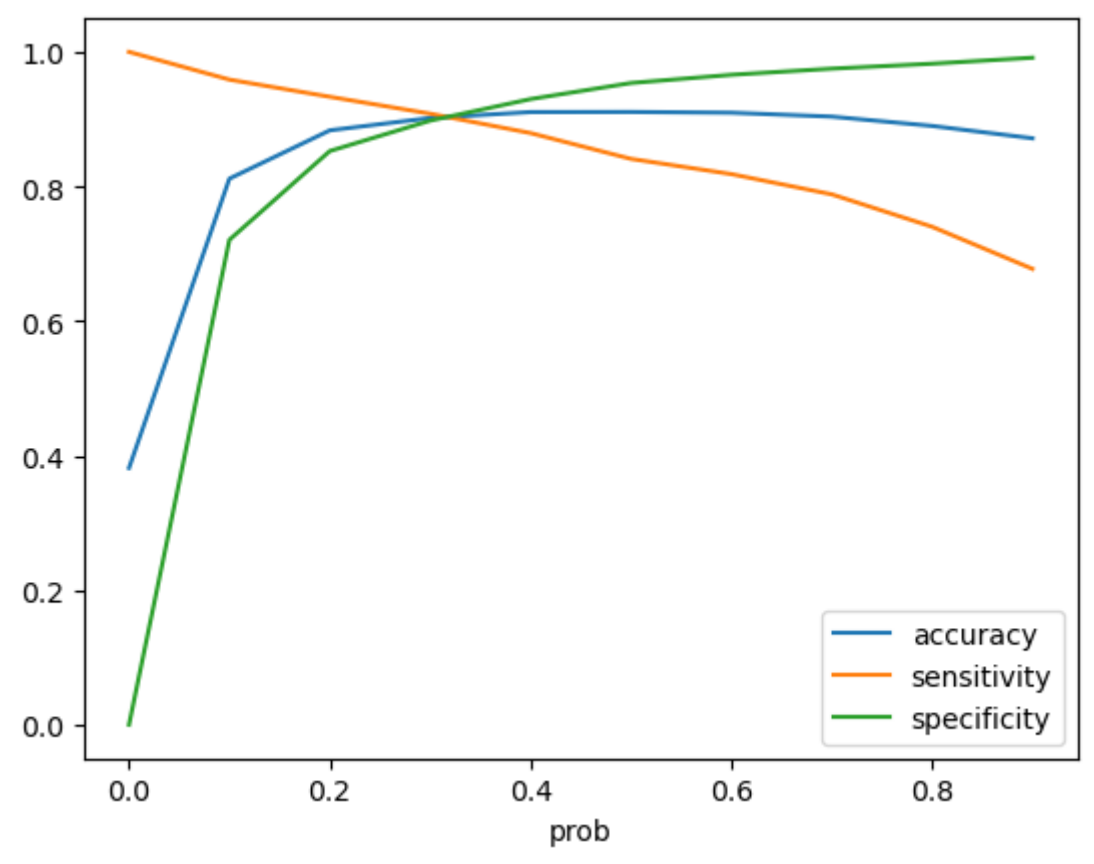






- Test split ratio as 70:30
- Using RFE to choose top 15 variables
- Model building by dropping the variables with P-value > 0.05 and VIF >5
- Overall accuracy is 90%
- ROC curve has a value of 0.96 is a good one

ACCURACY, SENSITIVITY AND SPECIFICITY



	prob	accuracy	sensitivity	specificity
0.0	0.0	0.381581	1.000000	0.000000
0.1	0.1	0.811567	0.958814	0.720712
0.2	0.2	0.883388	0.933278	0.852605
0.3	0.3	0.901776	0.908155	0.897840
0.4	0.4	0.910577	0.879325	0.929860
0.5	0.5	0.910734	0.841021	0.953748
0.6	0.6	0.909634	0.818369	0.965947
0.7	0.7	0.903819	0.788303	0.975095
0.8	0.8	0.889989	0.740115	0.982465
0.9	0.9	0.871601	0.677512	0.991360

- 0.3 is our optimal cut off point, which can be seen in above given graph.

MODEL EVALUATION

TEST DATA	
ACCURACY	0.909
SENSTIVITY	0.913
SPECIFICITY	0.906

TRAINING DATA	
ACCURACY	0.901
SENSTIVITY	0.908
SPECIFICITY	0.906

CONCLUSION

- Logistic regression model is used to predict the probability of conversion of a customer .
- Lead Score and conversion rate of final predicted model is 90% in training and test data both.
- Top variables that contributes for leads conversion in the model are:
    - —Tags will revert after reading the mail
    - — Tags closed by horizon
    - — Last activity SMS sent

RECOMMENDATION

X Education Company needs to focus on the following main aspects to increase the conversion rate :

- Increase on sending SMS notification since this helps in higher conversion.
- Get total visits increased by advertising etc, since this tends to give high conversion.