# LEAD SCORING CASE STUDY

PRESENTED BY:

RICHA RASHMI (DSC 47)

HARSHADKUMAR CHAVDA(DSC 47)

TEJ VAMSI (DSC 47)

# SCOPE

- X Education is an education company who sells online courses to industry professionals.
- With the help of ML model company wishes to find out the hot leads , who have more likely chances of conversion.
- The target of  lead conversion rate is set to be around 80% by CEO.
- Logistic model will be built based on data provided

# AIM

- To 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 most likely to convert
- Model should be able to adjust to if the company's requirement changes in the future.

# DATA UNDERSTANDING

- STEP FOLLOWED
  - READING THE DATASET
  - UNDERSTANDING THROUGH DATA DICTIONARY
  - FINDING THE MISSING VALUES

- RESULT
  - 9240 ROWS AND 37 COLUMNS
  - DATA CONTAINS ‘SELECT’ AS IMPURITY WHICH NEEDS TREATMENT
  - FEW COLUMNS CONTAINS MORE THAN 40 % NULL VALUES

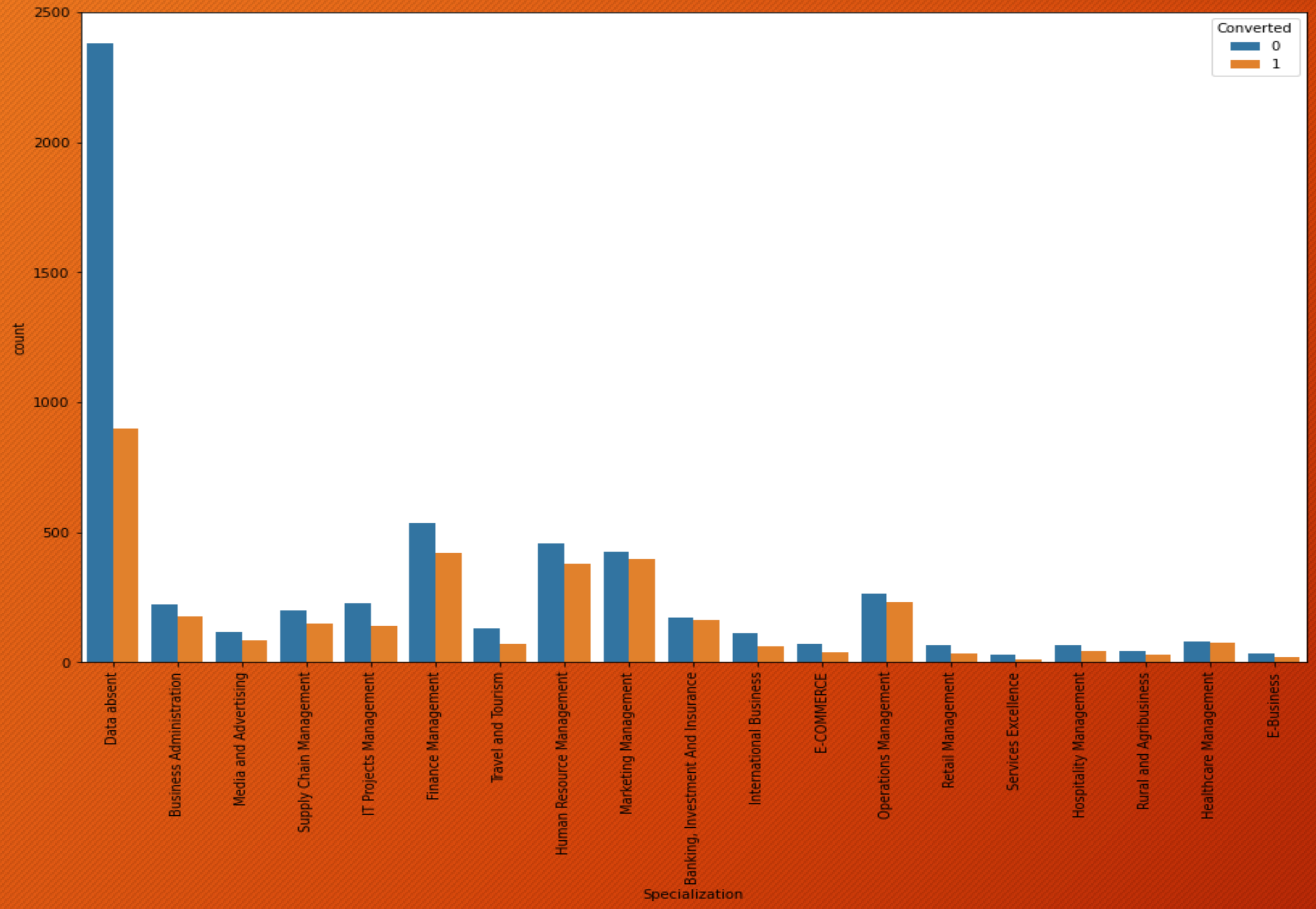
# DATA CLEANING

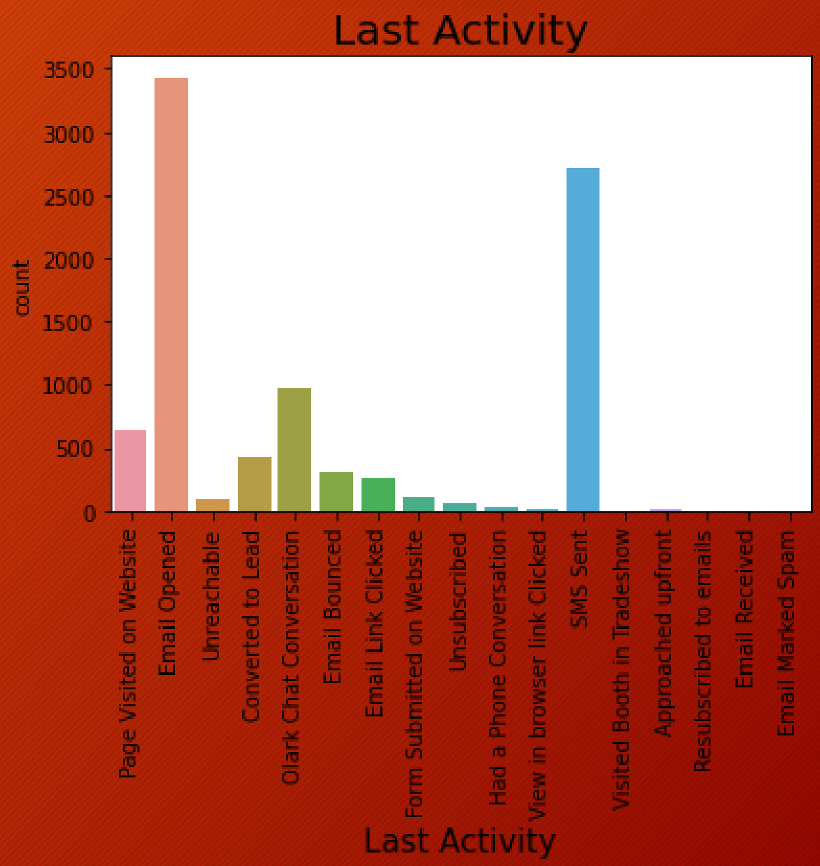
Following steps were undertaken to perform missing data identification and treatment:

- UNDERSTANDING THE DATA THROUGH COLUMNS
- COLUMNS CONTAINING HIGHER MISSING VALUES HAS BEEN DROPPED AND ROWS HAVING CONSIDERABLE MISSING VALUES WERE REMOVED.
- DATA CORRECTION WAS APPLIED BY REPLACING 'SELECT' WITH APPROPRIATE ATTRIBUTE TO MAKE IT MORE EASY FOR ANALYSIS.

# EDA

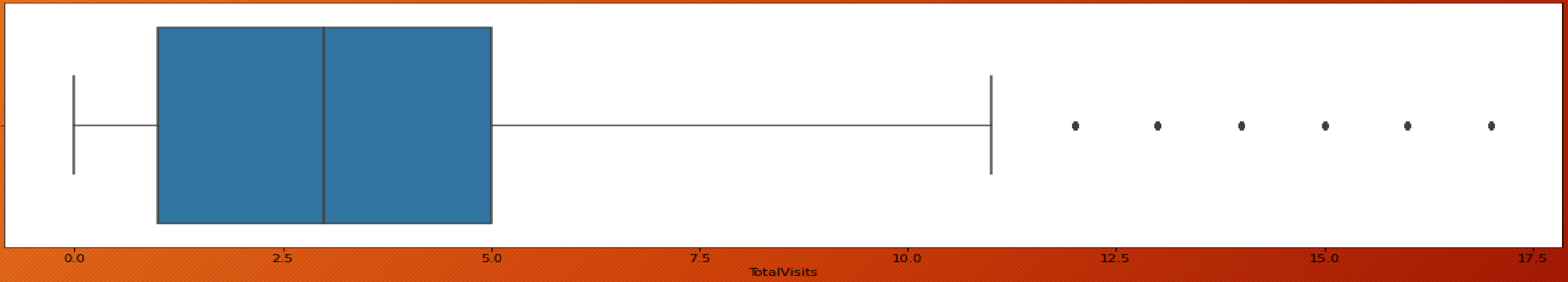
- Performed Univariate analysis and bivariate analysis on categorical variable and numerical variable.

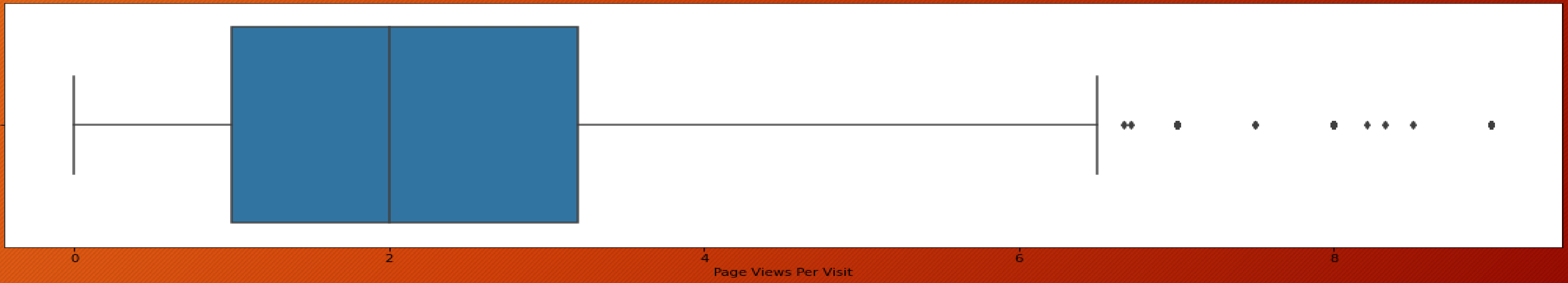




# EDA

- Outlier treatment





# MODEL PREPERATION

- CREATION OF DUMMY VARIBLES ON SELECTED ATTRIBUTES
  - Lead origin
  - Lead source
  - Do not email
  - Specialization
  - What is your current occupation
  - A free copy of mastering the interview

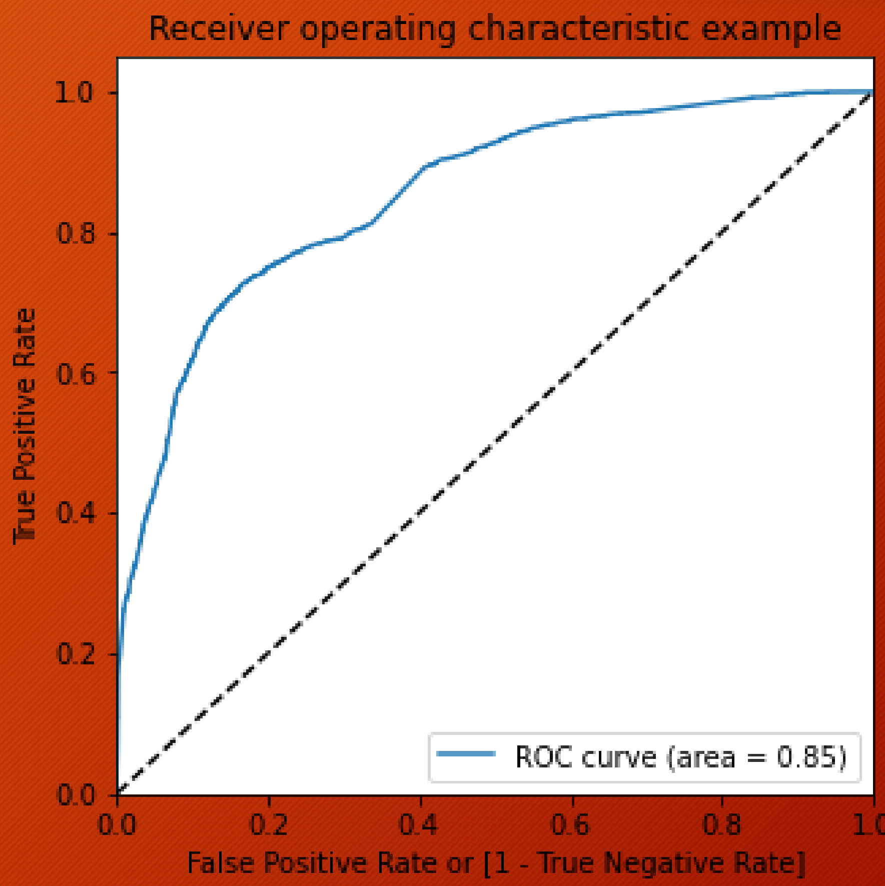
# MODEL BUILDING

- SPLITTING MODEL INTO 70:30 TRAIN TEST
- SCALING WITH MIN MAX SCALER ON THE COLUMNS
- RFE FOR FEATURE SELECTION AND SELECT BEST 15 FEATURES
- SELECTION OF FINAL MODEL AFTER 5 ITERATIONS WHERE VIF <5 AND P VALUE<0.05
- MAKING PREDICTIONS
  - PREDICT PROBABILITIES ON TRAIN SET
  - CREATING DATAFRAME

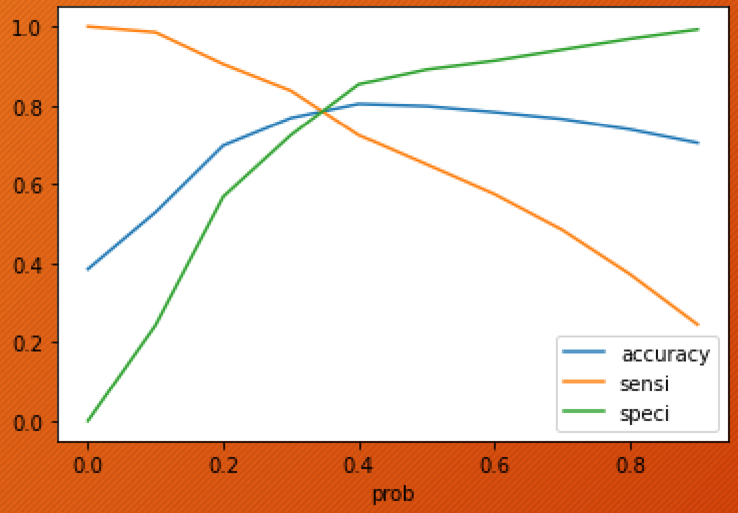
# MODEL EVALUATION

## PLOTTING ROC CURVE

- Area= 0.85



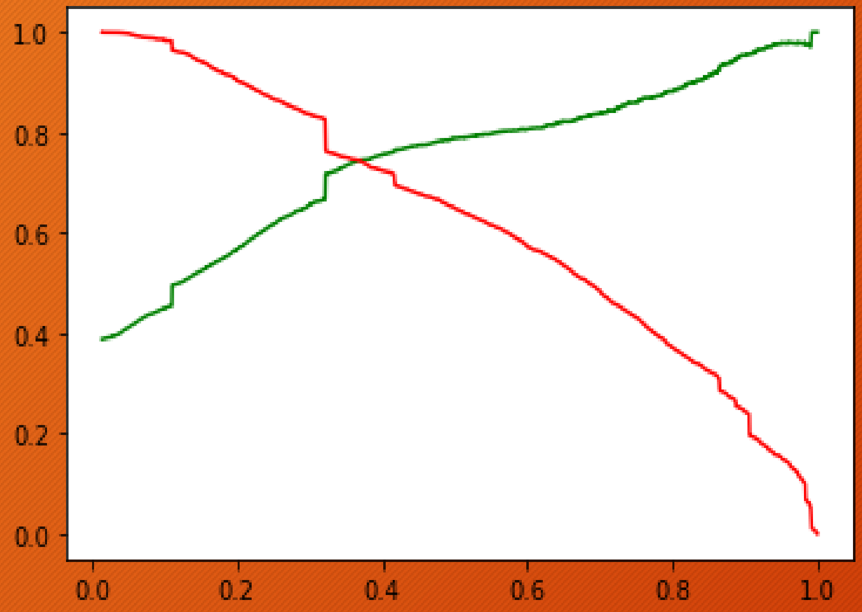
# OPTIMAZTION OF THRESHOLD POINT



| Train Data | | Test Data | |
|---|---|---|---|
| 1. Accuracy | 76.88% | 1. Accuracy | 77.30% |
| 2. Sensitivity | 76.61% | 2. Sensitivity | 75.02% |
| 3. Specificity | 77.05% | 3. Specificity | 78.60% |

0.30 AS NEW CUTOFF POINT

# OPTIMAZTION OF THRESHOLD POINT THROUGH PRECISION AND RECALL



| Train Data | | Test Data | |
|---|---|---|---|
| 1. Accuracy | 78.74% | 1. Accuracy | 78.99% |
| 2. Sensitivity | 73.50% | 2. Sensitivity | 71.28% |
| 3. Specificity | 82.02% | 3. Specificity | 83.39% |

# FINAL MODEL WITH LEAD SCORE

- MOST CONRTIBUTING VARIABLE
- Lead source from Welingak Website.
- People with maximum time spent on website.
- Lead source from Reference.
- Working professional.

Le
Landing Page Submission

# CONCLUSION

- The company should make calls to the leads coming from the lead sources "Welingak Websites" and "Reference" as these are more likely to get converted.
- The company should make calls to the leads who are the "working professionals" as they are more likely to get converted.
- The company should make calls to the leads who spent "more time on the websites" as these are more likely to get converted.
- The company should make calls to the leads whose current occupation was Student, Unemployed, Other as they are more likely to get converted.
- The company should not make calls to the leads whose Lead origin was “landing page submission" as they are not likely to get converted.
- The company should not make calls to the leads whose Lead Source was "Referral Sites" as they are not likely to get converted.
- The company should not make calls to the leads who opted for"Do Not Email" as they are not likely to get converted.