LEAD SCORING CASE STUDY

SUMEDHA CHAVAN

SIDDHARTH JHA

THE PROBLEM

What is the problem?	• To identify the most potential lead (Hot leads)
Who has this problem?	• X Education
Why should this problem be solved?	• To improve the lead conversion rate
How will I know this problem has been solved?	• If we can successfully identify the hot leads

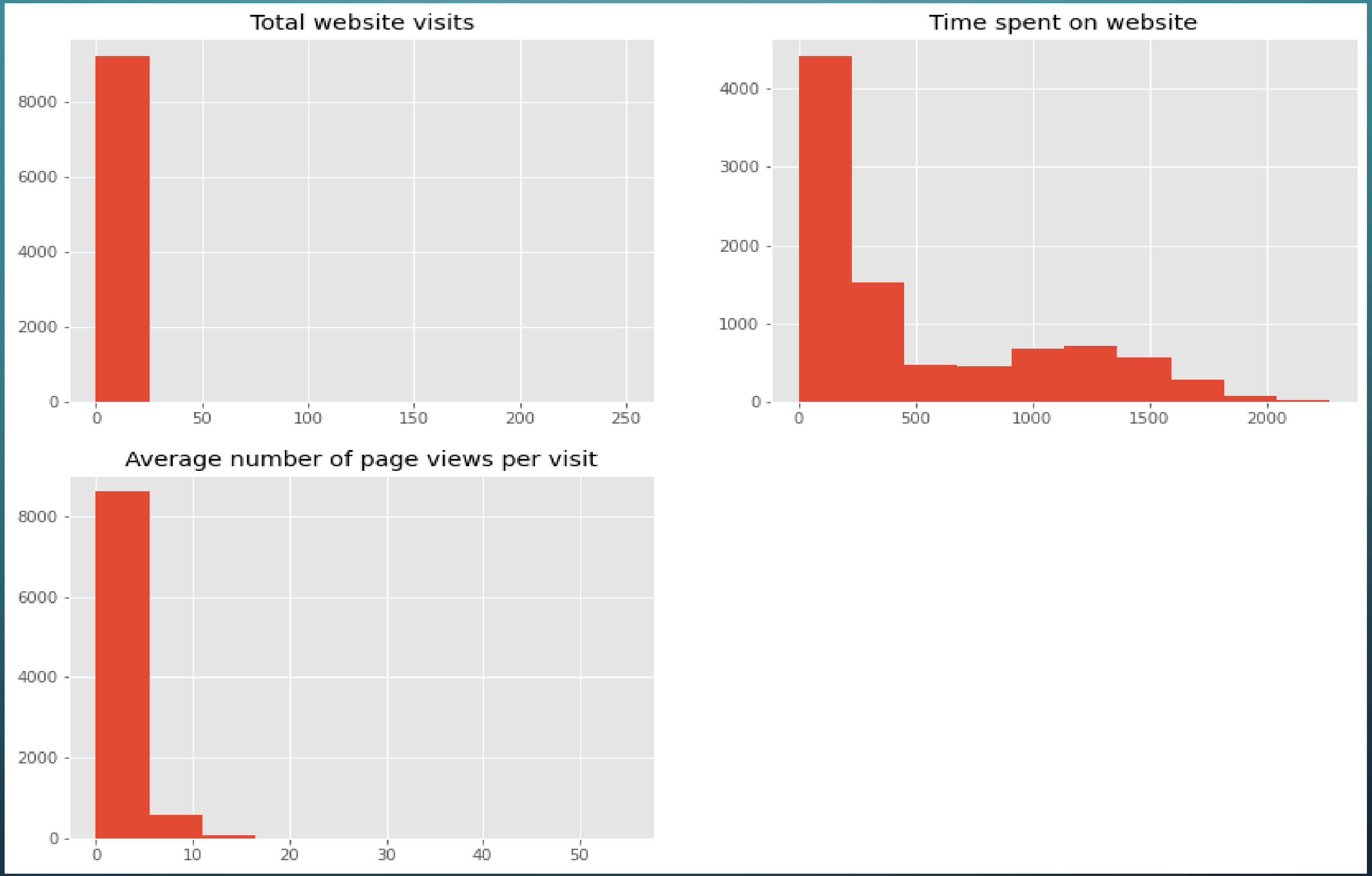
SOLUTION OUTLINE

- Data Cleaning
- EDA
- Data Preparation
- Model Selection – Logistic Regression
- Model Validation
- Conclusion

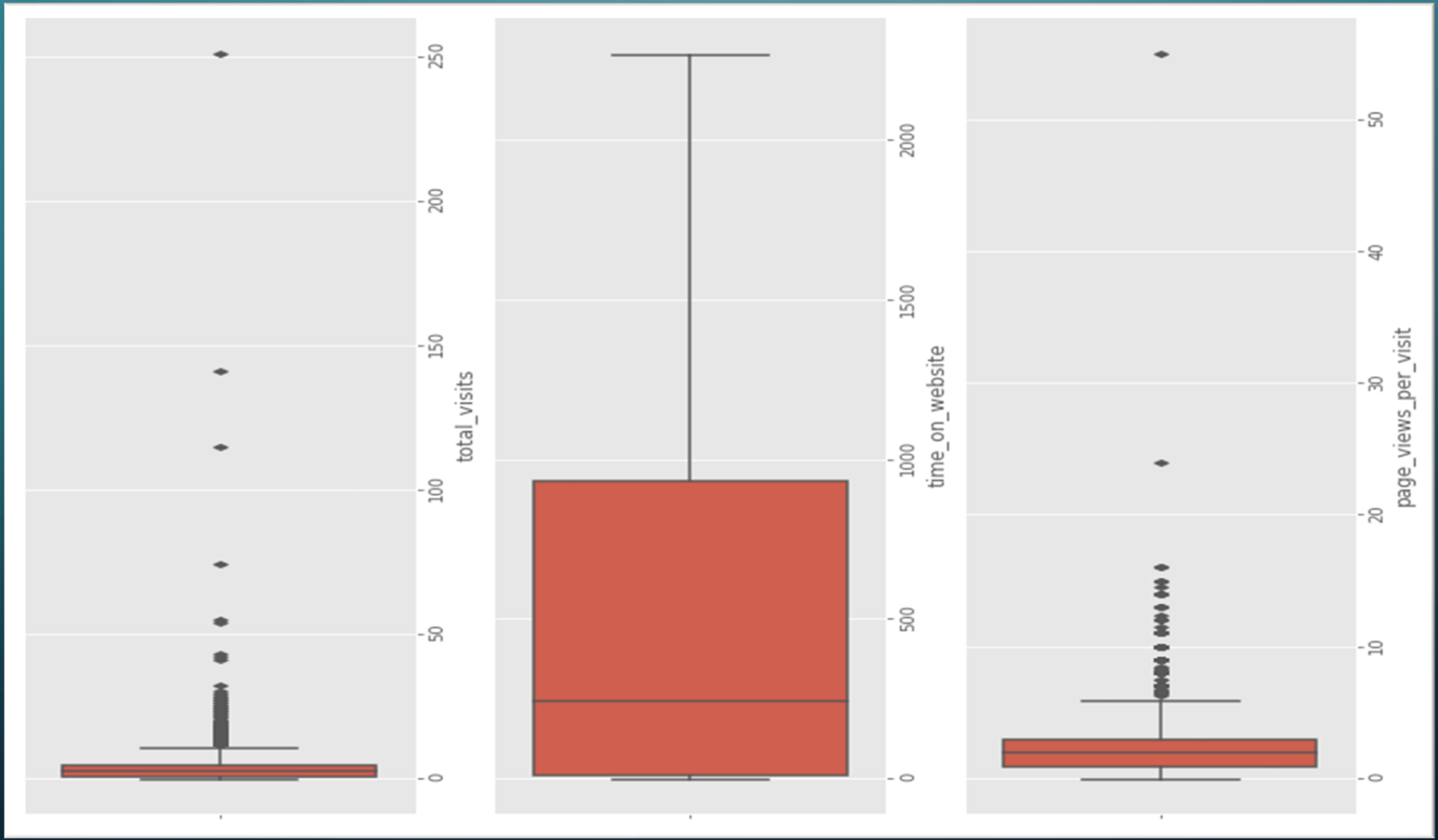
DATA CLEANING

- Shape of the data – (9240, 37)
- Dropping unnecessary columns like prospect_id
- Dropping columns having null values > 40%
- Categorize all the courses into Management, Business & Industry
- Grouping of other similar categorical variables such as city
- Typecasting columns into their respective type such as lead_number

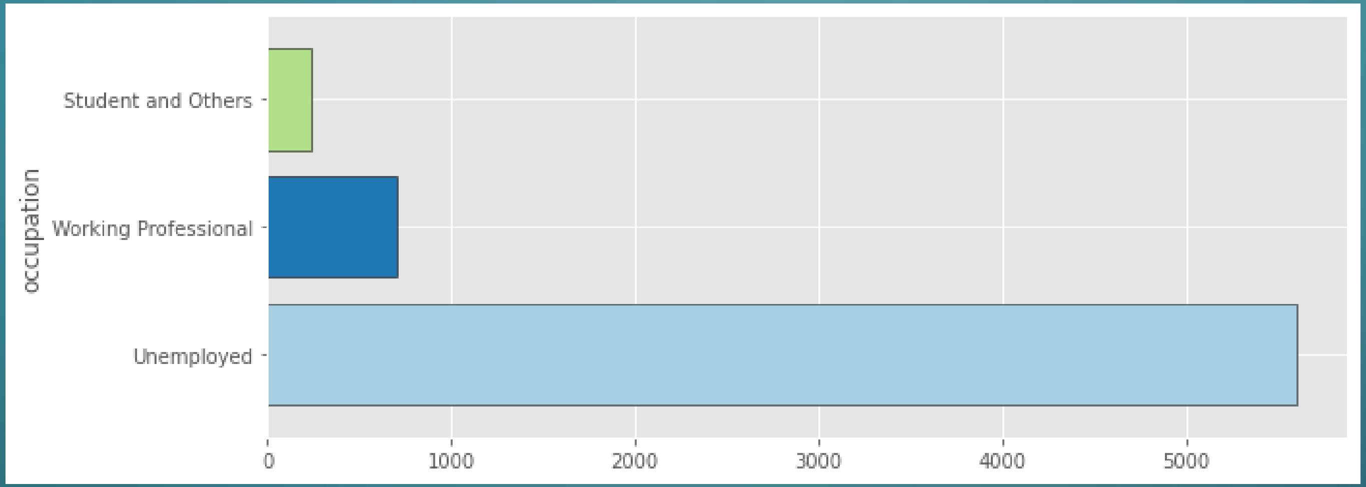
EDA

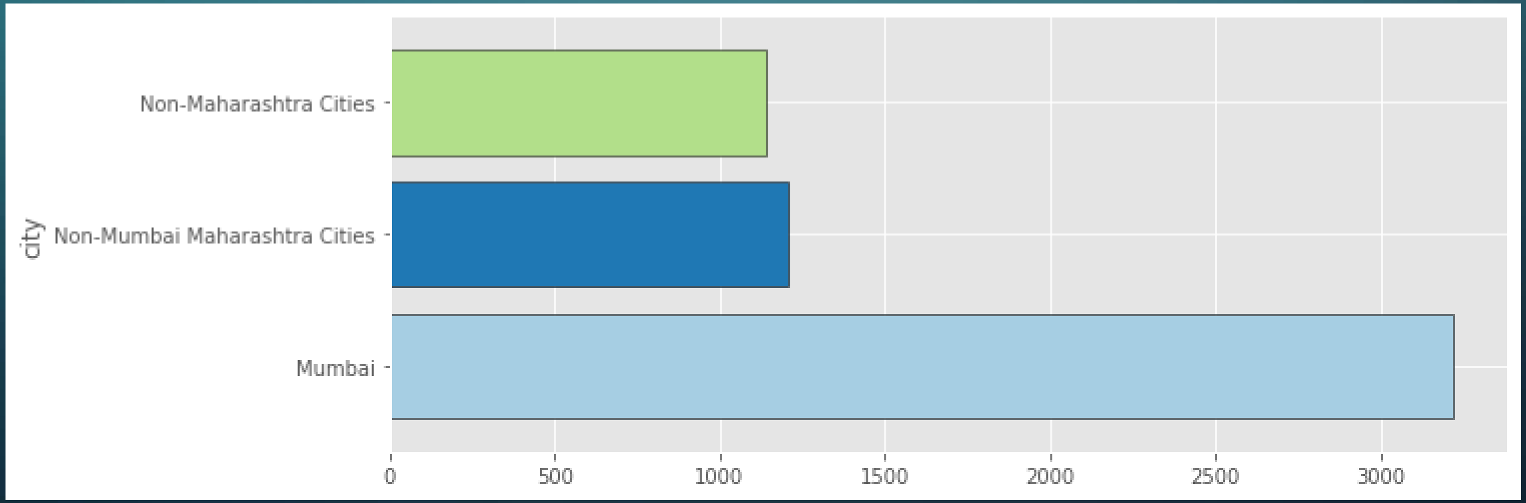

Total website visits
0
2000
4000
6000
8000
0
50
100
150
200
250
Time spent on website
0
1000
2000
3000
4000
0
500
1000
1500
2000
Average number of page views per visit
0
2000
4000
6000
8000
0
10
20
30
40
50

EDA


total_visits
0
50
100
150
200
250
time_on_website
0
500
1000
1500
2000
page_views_per_visit
0
10
20
30
40
50

EDA


occupation
Student and Others
Working Professional
Unemployed
0
1000
2000
3000
4000
5000


city
Non-Maharashtra Cities
Non-Mumbai Maharashtra Cities
Mumbai
0
500
1000
1500
2000
2500
3000

DATA PREPARATION

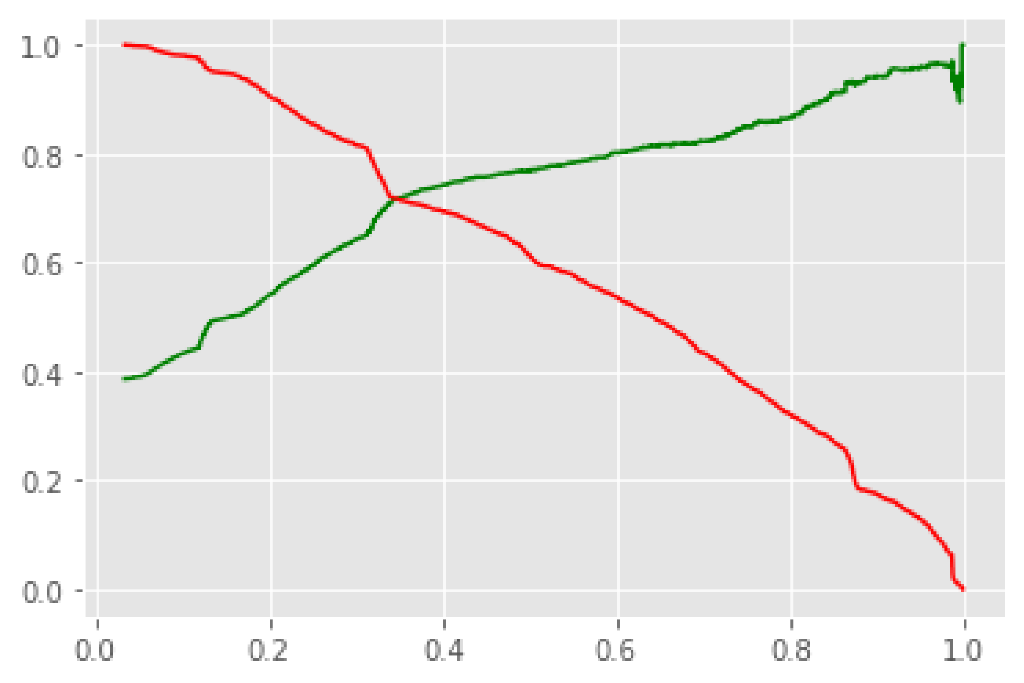
- Converting binary categorical variables to numeric
- Splitting the data into train & test set with a ratio of 70% and 30% respectively
- Dropping the columns causing multicollinearity

MODEL BUILDING

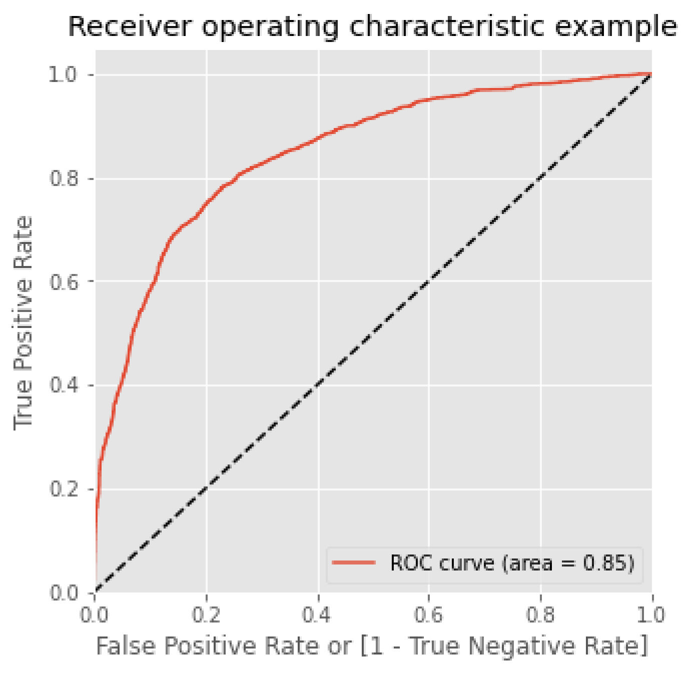
- Selecting the 12 most important features using RFE
- Fitting the model on those parameters
- Dropping the columns having greater P value than our threshold (0.05)

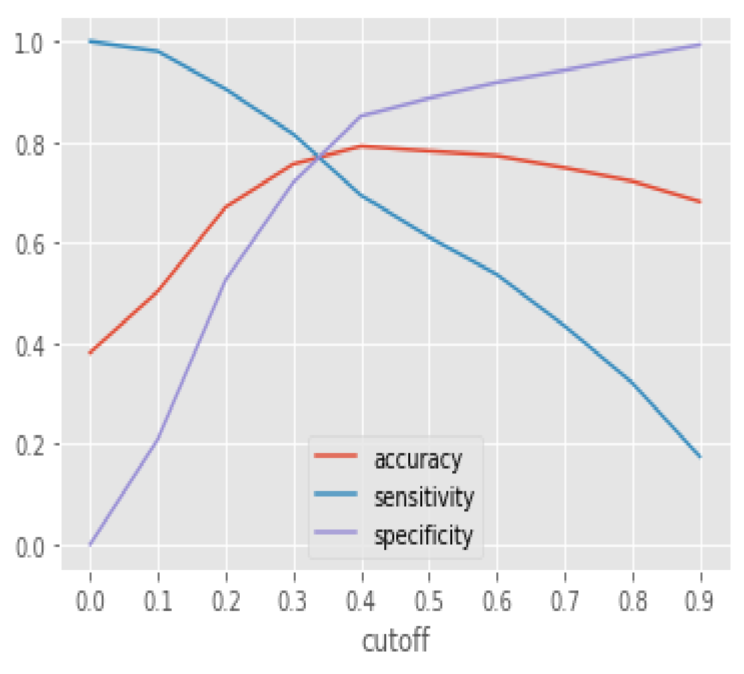
MODEL VALIDATION

- Precision – 0.59
- Recall – 0.85



MODEL VALIDATION





CONCLUSION

This analysis is done for X education to target potential leads.

Input data gives the information about customer activities like how they reached to the website, how many times they have visited and how much time they spent on it, occupation and lead conversion rate etc.

The following steps were followed:

Reading Data

Data is read from CSV file and stored in pandas' data frame.

Cleaning Data

Data was cleaned except for some null values. 'Select' is changed to null value in some columns. For some columns null values are imputed with median. Grouping is also followed for some columns

EDA

From EDA, it is observed that there is no specific relation in categorical variables. Also some outliers are found in case of numeric variables, Such outliers are capped to 90th percentile.

Creation of Dummy Variables

Dummy variables are created for Categorical Variables and MinMaxSaler is used for numeric ones.

Creation of Train and Test Dataset

Entire dataset is divided in train (70%) and Test (30%) datasets.

CONCLUSION

Model Building

RFE was done to attain top 15 relevant variables. Later variable having VIF >5 and p-value > 0.05 were removed manually.

Model Evaluation

A confusion matrix is created, optimum cut off value (using ROC curve) was used to find the accuracy (72.18), sensitivity (85.44) and specificity (64.01) .

Prediction

Prediction was done on test data set and with optimum cut off as with accuracy (72.47), sensitivity (86.48) and specificity (63.32).

Precision: Recall

Used to recheck and a cut off of 0.4 was found.

CONCLUSION

It is also observed that variables mattered the most in finding potential buyers are:
Total time spent on website
Total number of visits
When a lead source was,
Google
Direct Traffic
Organic Search
When the last activity was SMS or Olark chat conversation
When current occupation is working professional