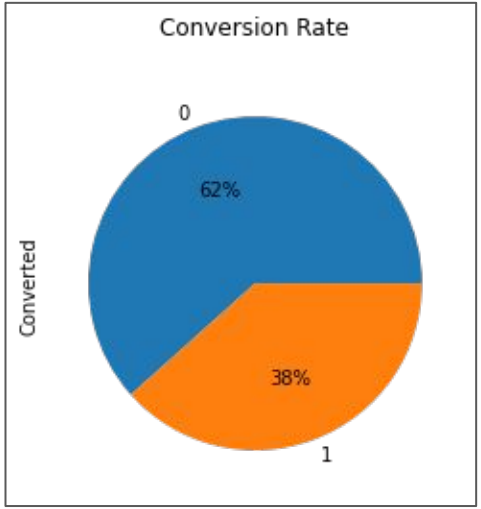
LEAD SCORING CASE STUDY

Submitted by :-
Priyanka Tanpure
Biswakesh Naik

PROBLEM STATEMENT

- An Education company named X Education sells online courses to industry professionals.
- It markets its courses on several websites and search engines like Google.
- People interested in the courses land on their website are called the Leads who further might get converted to take up the course.
- The typical lead conversion rate at X Education is around 30% which is very poor.



SOLUTION APPROACH

- Build a logistic regression model
- Assign a lead score between 0 to 100 to each of the leads which can be used by the company to target Potential leads.
- A Lead score higher than a particular threshold would mean that the lead is most likely to convert.
- A lower Lead score would mean that the lead is cold and will mostly not get converted.
- The Model should have a sensitivity value more than 80%.

OBJECTIVE OF THE COMPANY

- The company wishes to identify the most potential leads, also known as 'Hot Leads'.
- Thus, the sales Team of the company can only focus on the Hot Leads.
- By this, they would be able to increase the Conversion rate by approximately 80%.

ROADMAP TO CASE STUDY

1. BUSINESS UNDERSTANDING
2. SOLUTION APPROACH
3. DATA CLEANING AND VALIDATION
4. EXPLORATORY DATA ANALYSIS
5. MODEL BUILDING
6. MODEL EVALUATION
7. ANSWERING BUSINESS QUESTIONS

- 1. Handling Duplicate rows and Missing Values
- 2. Handling Outliers
- 3. Checking row wise null percentage
- 4. Datatype correction

- 1. Univariate Analysis
- 2. Segmented Univariate Analysis
- 3. Bivariate and Multivariate Analysis

- 1. Train Test Split
- 2. Feature Scaling
- 3. Feature Selection
- 4. ML Model Building

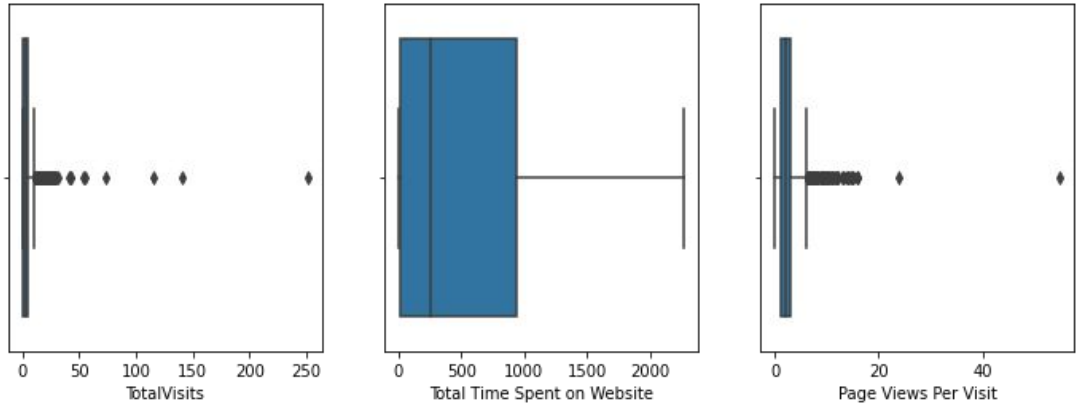
- 1. Predicting the Lead Probability Score
- 2. Accuracy, Sensitivity and Specificity
- 3. Threshold Detection using ROC Curve
- 4. Predicting for Test Data

- 1. Technical Solutions
- 2. Business Solutions

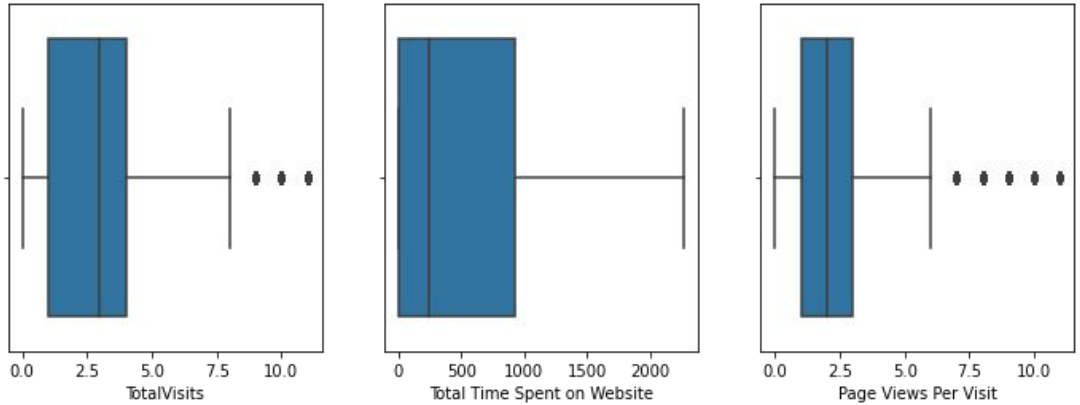
DATA CLEANING AND VALIDATION

1. Handled the “Select” value and converted to NaN.
2. Handled Duplicate rows.
3. Dropped columns with >40% Missing Values.
4. Imputed Missing Values
5. Handled Outliers
6. Checked row wise null percentage and dropped Rows with >70% Missing Values.
7. Corrected Data Types for required Columns.
8. Removed highly skewed and Redundant Columns.
9. Standardized the features.

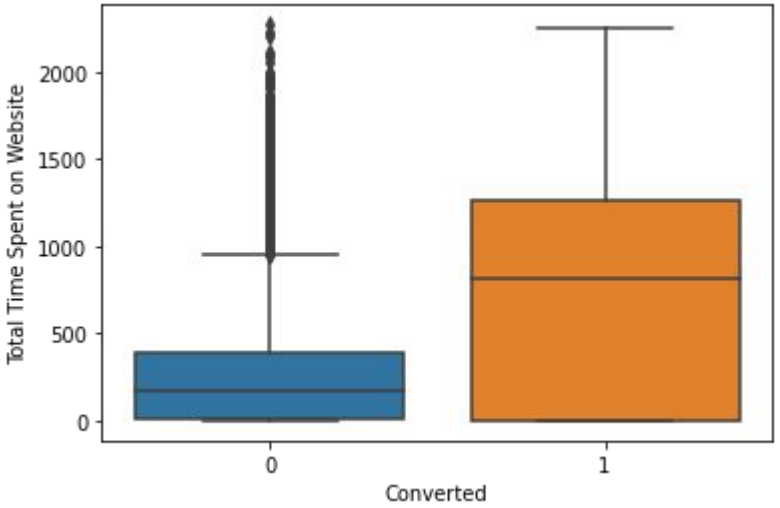
OUTLIERS

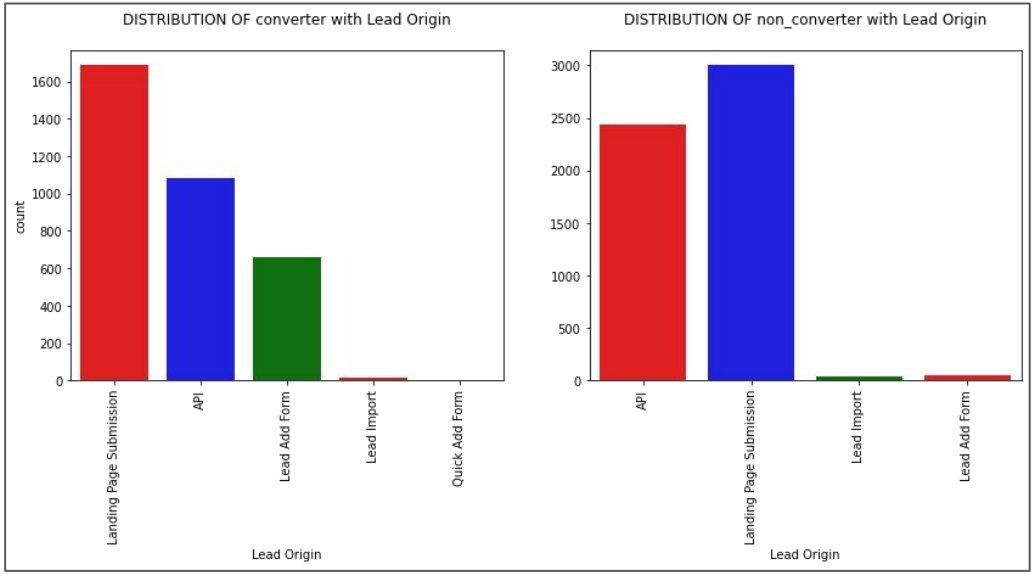


AFTER REMOVING OUTLIERS



EXPLORATORY DATA ANALYSIS





Conclusion:

1. If Total Time Spent on Website is more than Leads Converted are more
2. Add forms Lead have more chances of Conversion

DATA PREPARATION

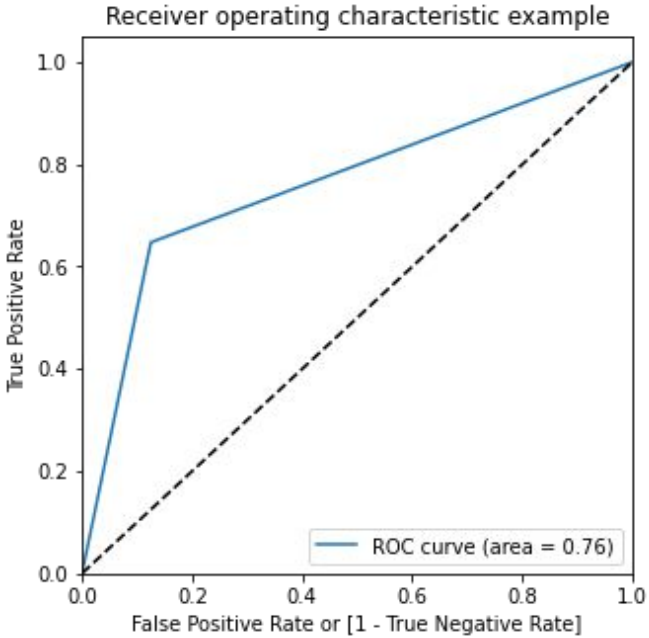
- Data Standardisation :
  - Used custom function to convert Binary variables into 0 or 1.
- Feature Scaling :
  - Using StandardScaler to scale the numerical variables
- Data Encoding:
  - Used One-hot Encoding Technique to create dummy variables for categorical columns.
- Splitting into Train and Test Data:
  - Used sklearn module to split 70 % of the data for Training purpose and 30% for Testing purpose.
- Removing the Sales data columns which aren’t fit for modelling.
  - These columns are Tags, Last Activity, Last Notable Activity etc.

MODEL BUILDING

1. Feature Selection using RFE
2. Used statsmodel to filter the insignificant features using p-value
3. Used Variance Inflation factor with a threshold value of 5.0(Maximum) to check for Multicollinearity between the features.
4. Used sklearn to create the optimal Logistic Regression Model

DRIVING VARIABLES:

1.	**Do Not Email**	**-1.2726**
2.	**TotalVisits**	**0.6336**
3.	**Total Time Spent on Website**	**4.5348**
4.	**LeadOrigin_Landing Page Submission**	**-0.9451**
5.	**LeadOrigin_Lead Add Form**	**3.6990**
6.	**LeadSource_Olark Chat**	**0.9773**
7.	**LeadSource_Referral Sites**	**-0.8589**
8.	**LeadSource_Welingak Website**	**3.3476**
9.	**Specialization_Not Specified**	**-1.0744**
10.	**CurrentOccupation_Other**	**-2.5817**
11.	**CurrentOccupation_Student**	**-2.4031**
12.	**CurrentOccupation_Unemployed**	**-2.4595**



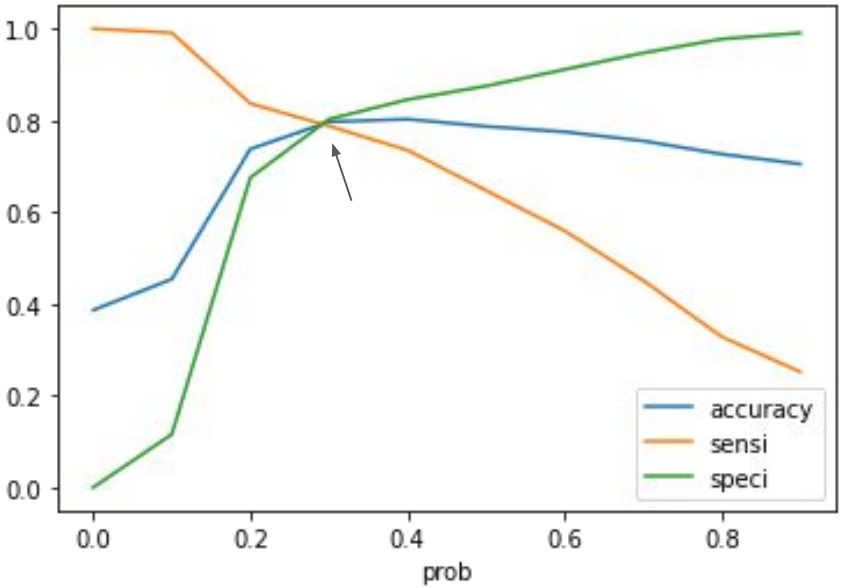
MODEL EVALUATION

Training data:

- Accuracy : ~80%
- Sensitivity : ~78.4%
- Specificity : ~81%

Test data:

- Accuracy : ~80%
- Sensitivity : ~77%
- Specificity : ~82%



CONCLUSION:
So, based on the above graph we have taken the threshold value for lead score to be 0.3.

TECHNICAL SOLUTIONS

- We have checked both sensitivity- specificity graph as well as the Precision-Recall graph.
- Based on the observations, we have decided 0.3 as the optimal threshold for deciding whether the lead has the potential or not.
- Hence, the Lead with a Lead score >= 0.3 can be converted and hence is considered as Potential Lead.
- Accuracy, Sensitivity and Specificity for the Test Data comes out to be 80%, 77% and 82% respectively which is almost similar to the corresponding values of Trained data.

BUSINESS SOLUTIONS

1. For Converting Maximum Number of Potential Leads, the company needs to focus on the following aspects:
    a. Total time spent on the website by the Person
    b. Leads who were identified by lead Add Form Origin
    c. Leads whose source are Welingak Chat and Olark Chat.
    d. Current Occupation of the lead whether they are Unemployed or working Professionals.
2. In quarters when the conversion rate is high and Making a phone call is not necessary, The team can focus on Email and SMS mode of communication.
3. If the Lead is a student then they need to be communicated more frequently.