# Lead Score Case Study

ABHISHEK, SAPNA & SANJEEV

# Problem Statement

- X Education, an online education company, is facing a challenge with its lead conversion process. Despite acquiring a substantial number of leads daily through website visits, form submissions, and referrals, their current lead conversion rate is only around 30%. This poor conversion rate indicates inefficiency in identifying and targeting potential leads effectively.
- In an effort to improve lead conversion rates, X Education seeks a solution to identify and prioritize "Hot Leads" – those most likely to convert into paying customers. The company aims to enhance its lead management system to increase the conversion rate, optimize the sales team's efforts, and ultimately improve overall business profitability.
- How can X Education efficiently identify and prioritize potential "Hot Leads" to enhance its lead conversion rate

# Case Study Goals

- Develop a logistic regression model to assign lead scores from 0 to 100, enabling efficient targeting of potential leads. Higher scores signify hotter leads with a greater likelihood of conversion, while lower scores represent colder leads less likely to convert.
- Ensure model flexibility to address future changes in the company's requirements by adapting to additional problems and challenges presented by the company

# Model Building Steps

- Import all libraries, read and clean the input data, and carrying out EDA
- Train-Test Split
- Feature Scaling
- Checking Conversion Rate
- Feature Selection using RFE, looking at correlations
- Model Building
- Plotting ROC Curve
- Finding Optimal Cut-off to measure Accuracy metrics of the model
- Making predictions and calculating score on Test dat

# Detailed Steps

1. Reading the dataset from csv
2. Data Cleaning: Replacing "Select" as NULL values , Removing/Dropping of columns whose percentage of null values greater than 25%, Handling Missing values , Checking for any other null values in the dataframe, Removing unwanted columns that do not give relevant information and have only one unique value
3. Exploratory Data Analysis: Checking Data Imbalance percentage and Lead Converted Ratio and Univariate and Bivariate Analysis
4. Data Preparation
5. Splitting of Train and Test set
6. Scaling of features
7. Model Building Using Statsmodel and RFE
8. Model Evaluation for Train set : Creating Confusion matrix, Calculate : Accuracy, Sensitivity, Specificity, False positive rate, Positive predictive value, Negative predictive value , Precision and Recall

# Detailed Steps Contd.

- Plotting the RoC curve
- Determining optimal cut-off point or probability
- Model evaluation after obtaining optimal cut-off point or probability method
- Model Evaluation using Precision-Recall Trade off method
- Comparing the metrics values from Optimal cut-off point method and Precision-Recall Trade off method

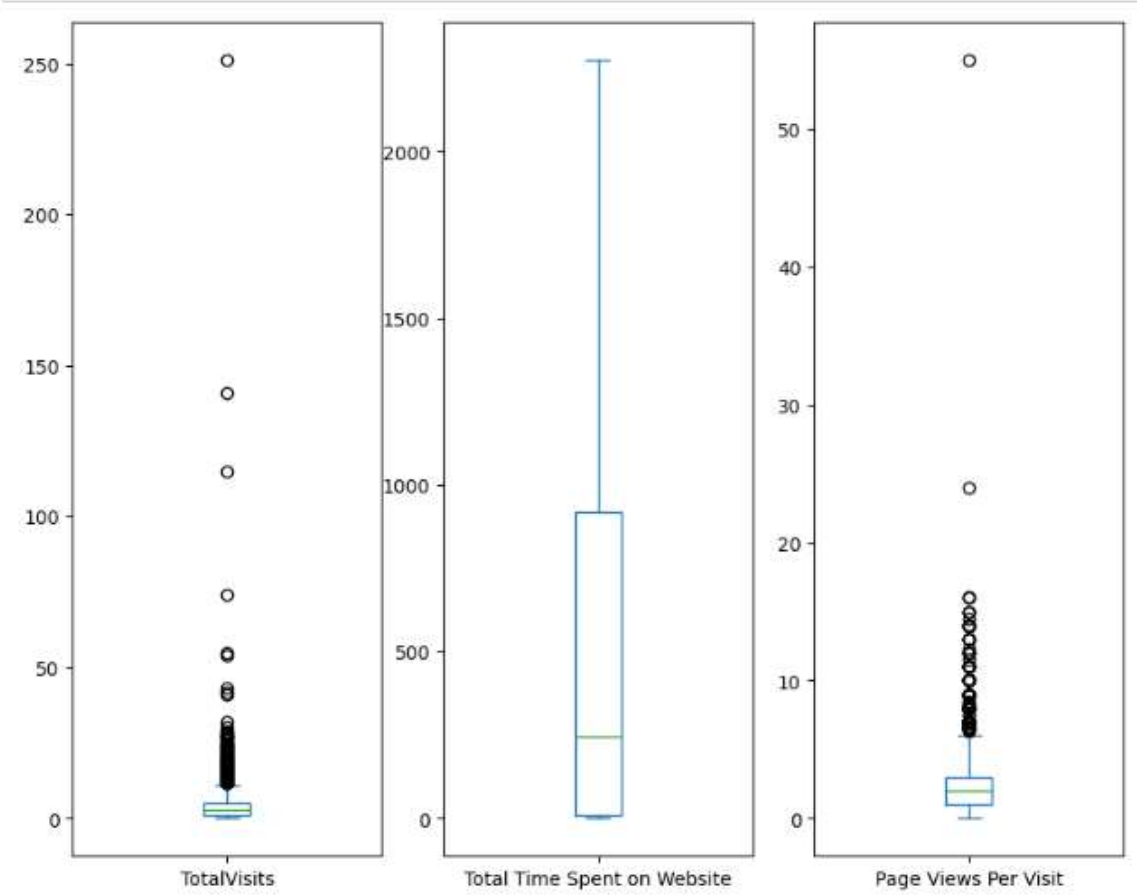
9.Model Evaluation for Test set

10. Results:

- Metrics Value
- Hot leads
- Prospect ID of the customers
- Features of the model

# Exploratory Data Analysis (EDA)

It is observed that both "TotalVisits" and "Page Views Per Visit" contain outliers, as evident from the box plots. We shall drop these outliers as these may disturb our analysis and model building



# Exploratory Data Analysis (EDA)

- Columns which shows no variance with the Target variable Converted can be dropped :

  "Magazine“ ,

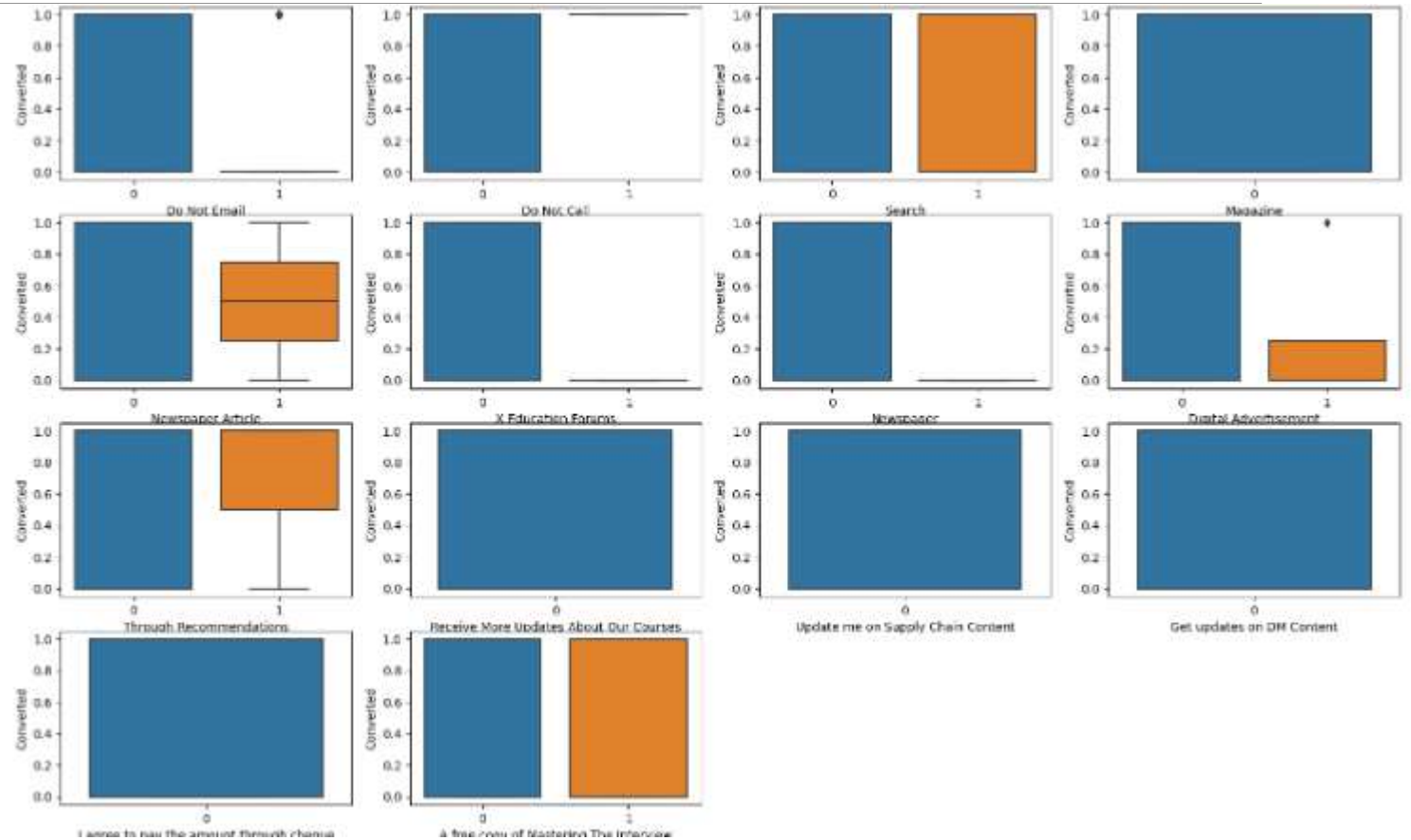
  "Receive More Updates About Our Courses",

  "Update me on Supply Chain Content",

  "Get updates on DM Content",

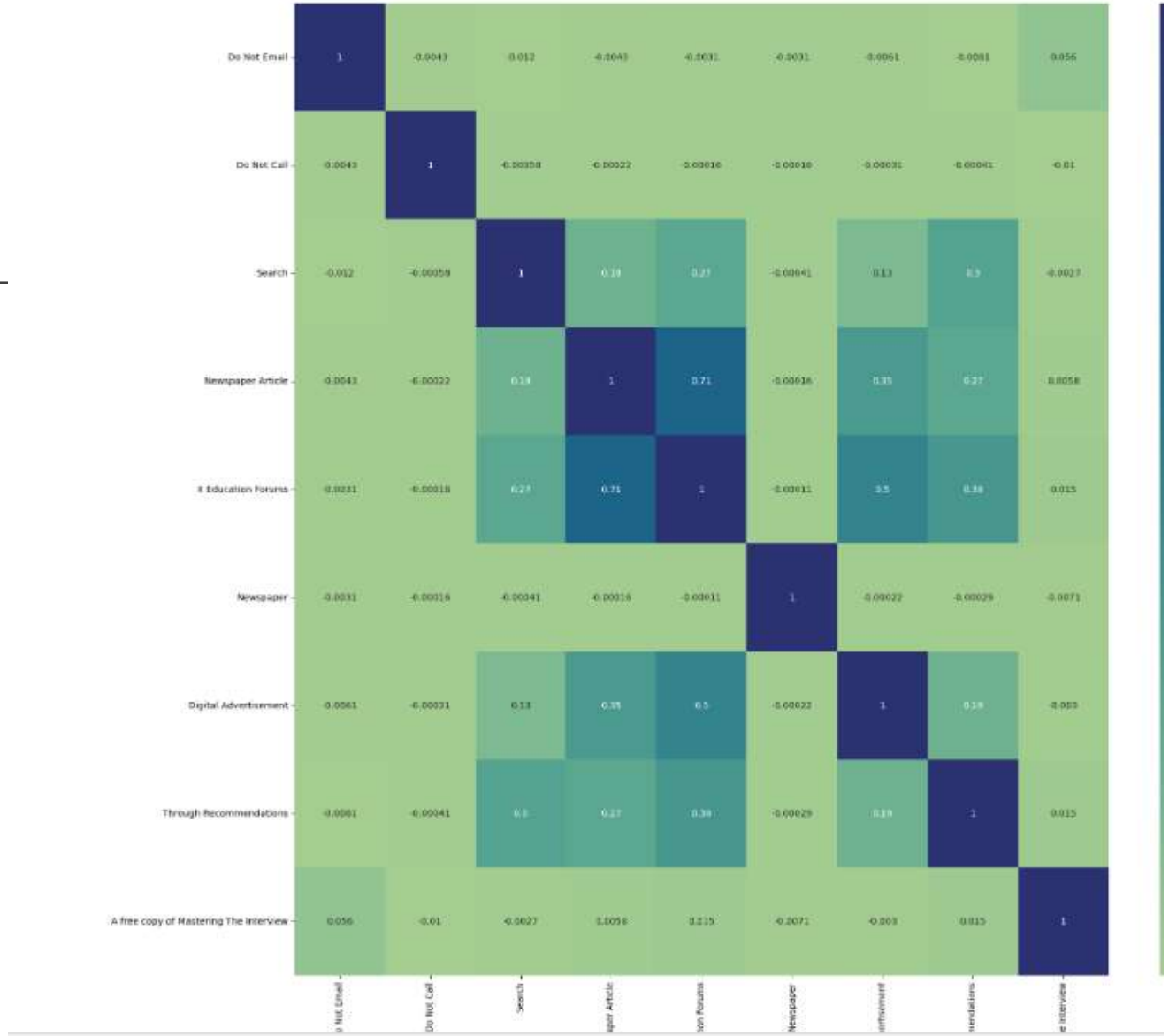
  "I agree to pay the amount through cheque“

- Other variables shows some significant variation with the target variable (i.e. Converted) and are retained



# Correlations

Some of the variables show significance correlations



# Model Evaluation

RFE and Logistic regression was used

8 iterations were used to obtain final model

P value is <0.05 and VIF <2 in the final model

Generalized Linear Model Regression Results

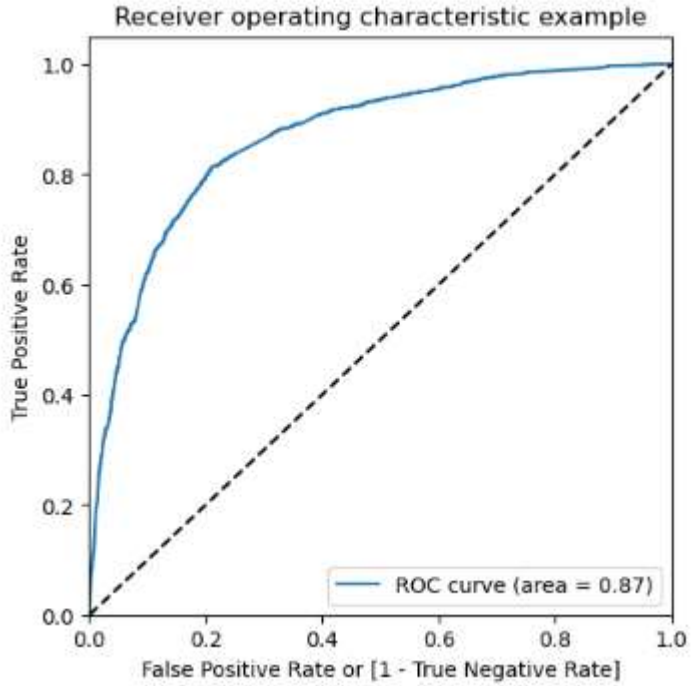
Generalized Linear Model Regression Results

| | | | |
|---|---|---|---|
| **Dep. Variable:** | Converted | **No. Observations:** | 6367 |
| **Model:** | GLM | **Df Residuals:** | 6355 |
| **Model Family:** | Binomial | **Df Model:** | 11 |
| **Link Function:** | Logit | **Scale:** | 1.0000 |
| **Method:** | IRLS | **Log-Likelihood:** | -2814.4 |
| **Date:** | Tue, 28 May 2024 | **Deviance:** | 5628.7 |
| **Time:** | 07:55:18 | **Pearson chi2:** | 6.66e+03 |
| **No. Iterations:** | 6 | **Pseudo R-squ. (CS):** | 0.3552 |
| **Covariance Type:** | nonrobust | | |

| | coef | std err | z | P>\|z\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| **const** | -0.9685 | 0.066 | -14.778 | 0.000 | -1.097 | -0.840 |
| **Do Not Email** | -1.3539 | 0.158 | -8.547 | 0.000 | -1.664 | -1.043 |
| **TotalVisits** | 0.1984 | 0.039 | 5.092 | 0.000 | 0.122 | 0.275 |
| **Total Time Spent on Website** | 1.0709 | 0.038 | 27.948 | 0.000 | 0.996 | 1.146 |
| **A free copy of Mastering The Interview** | -0.2910 | 0.076 | -3.837 | 0.000 | -0.440 | -0.142 |
| **Lead Origin_Lead Add Form** | 4.4919 | 0.214 | 21.009 | 0.000 | 4.073 | 4.911 |
| **Lead Source_Olark Chat** | 1.0581 | 0.110 | 9.581 | 0.000 | 0.842 | 1.275 |
| **Last Activity_Converted to Lead** | -0.9704 | 0.210 | -4.625 | 0.000 | -1.382 | -0.559 |
| **Last Activity_Olark Chat Conversation** | -1.4460 | 0.163 | -8.880 | 0.000 | -1.765 | -1.127 |
| **Last Activity_Page Visited on Website** | -0.6325 | 0.145 | -4.348 | 0.000 | -0.918 | -0.347 |
| **Last Notable Activity_Modified** | -0.3035 | 0.086 | -3.514 | 0.000 | -0.473 | -0.134 |
| **Last Notable Activity_SMS Sent** | 1.4083 | 0.083 | 16.962 | 0.000 | 1.246 | 1.571 |

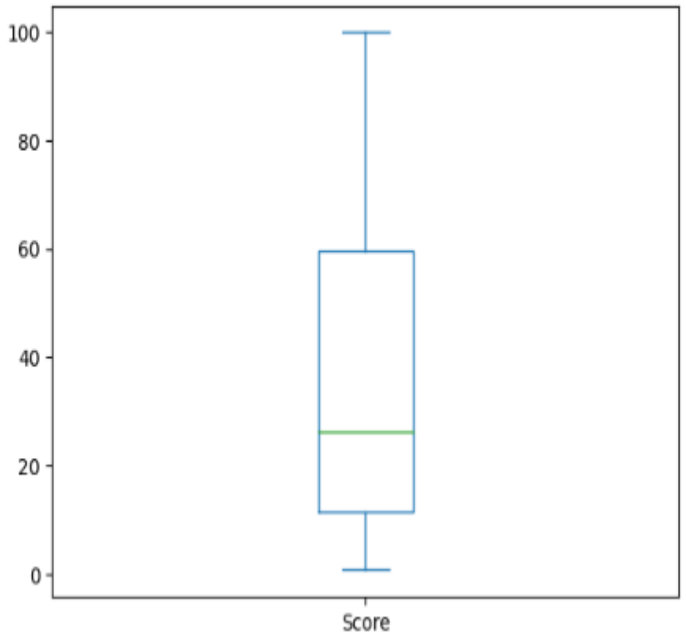
# VIF & ROC Curve

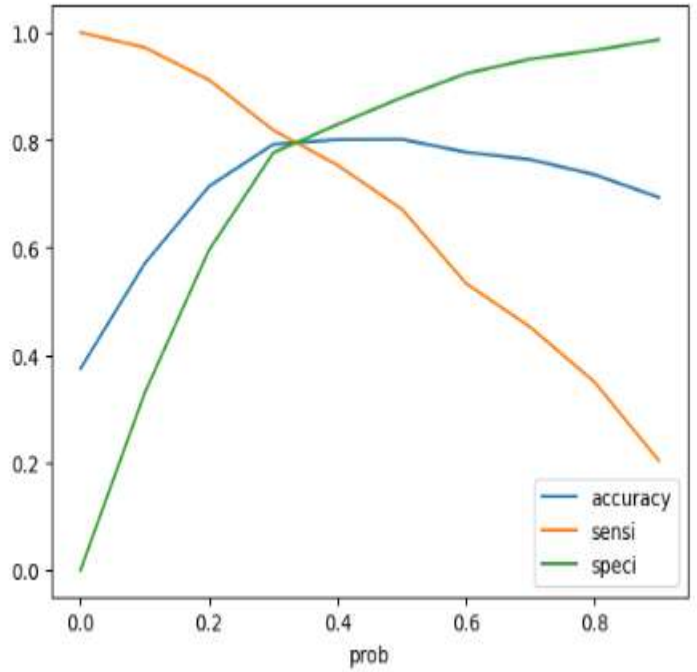
| | Features | VIF |
|---|---|---|
| 9 | Last Notable Activity_Modified | 1.93 |
| 5 | Lead Source_Olark Chat | 1.80 |
| 7 | Last Activity_Olark Chat Conversation | 1.61 |
| 1 | TotalVisits | 1.47 |
| 3 | A free copy of Mastering The Interview | 1.36 |
| 10 | Last Notable Activity_SMS Sent | 1.28 |
| 2 | Total Time Spent on Website | 1.27 |
| 6 | Last Activity_Converted to Lead | 1.26 |
| 4 | Lead Origin_Lead Add Form | 1.23 |
| 8 | Last Activity_Page Visited on Website | 1.17 |
| 0 | Do Not Email | 1.14 |



# Optimal cut-off and Metrics

- From ROC curve we could see 0.3 is the optimal cutoff
- Accuracy : 79%
- Specificity : 78%
- Sensitivity : 82%
- Precision Score – 69%
- Recall Score – 82%



# Test Data Metrics

- Accuracy - 78%
- Sensitivity - 83%
- Specificity - 75%
- Precision – 69%
- Recall - 83%

# Summary

- Score assigned to the Leads fall under the range i.e. 0 to 100, in both train and test dataset
- Critical parameter in model

| | |
|---|---|
| Lead Origin_Lead Add Form | 4.491904 |
| Last Notable Activity_SMS Sent | 1.408319 |
| Total Time Spent on Website | 1.070866 |
| Lead Source_Olark Chat | 1.058071 |
| TotalVisits | 0.198407 |
| A free copy of Mastering The Interview | -0.291003 |
| Last Notable Activity_Modified | -0.303451 |
| Last Activity_Page Visited on Website | -0.632515 |
| const | -0.968522 |
| Last Activity_Converted to Lead | -0.970432 |
| Do Not Email | -1.353942 |
| Last Activity_Olark Chat Conversation | -1.446049 |

- A total of 11 features are able to define the maximum variance in the conversion rate. In these 11 features, dummy variables from “Lead Source” and “Lead Notable Activity” have greater weightage. Other features like “Do Not Email” and “Total Time Spent on Website” are significant enough in scoring the leads