Lead Scoring Case Study

By

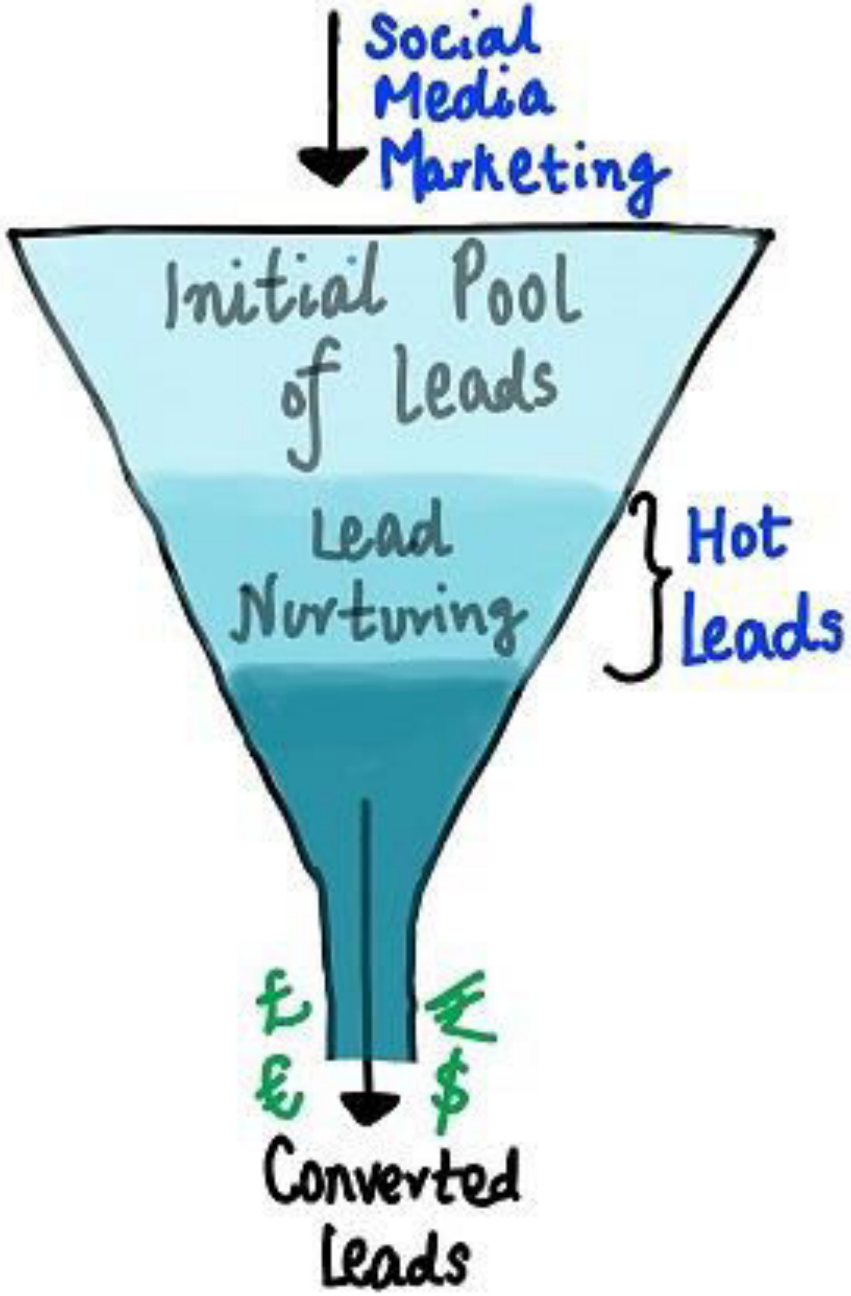
Malini S

Ricky Samuel

Sumit Tiwari

Introduction

- An education company named X Education sells online courses to industry professionals.
- The company markets its courses on several websites and search engines like Google.
- Once these people land on the website, they might browse the courses or fill up a form for the course, or watch some videos.
- When these people fill up a form providing their email address or phone number, they are classified to be a lead.
- The company also gets leads through past referrals.
- Once these leads are acquired, employees from the sales team start making calls, writing emails, etc
- Through this process, some of the leads get converted while most do not.
- The typical lead conversion rate at X education is around 30%.
- To make this process more efficient, the company wishes to identify the most potential leads, also known as ‘Hot Leads’.
- If they successfully identify this set of leads, the lead conversion rate should go up as the sales team will now be focusing more on communicating with the potential leads rather than making calls to everyone.

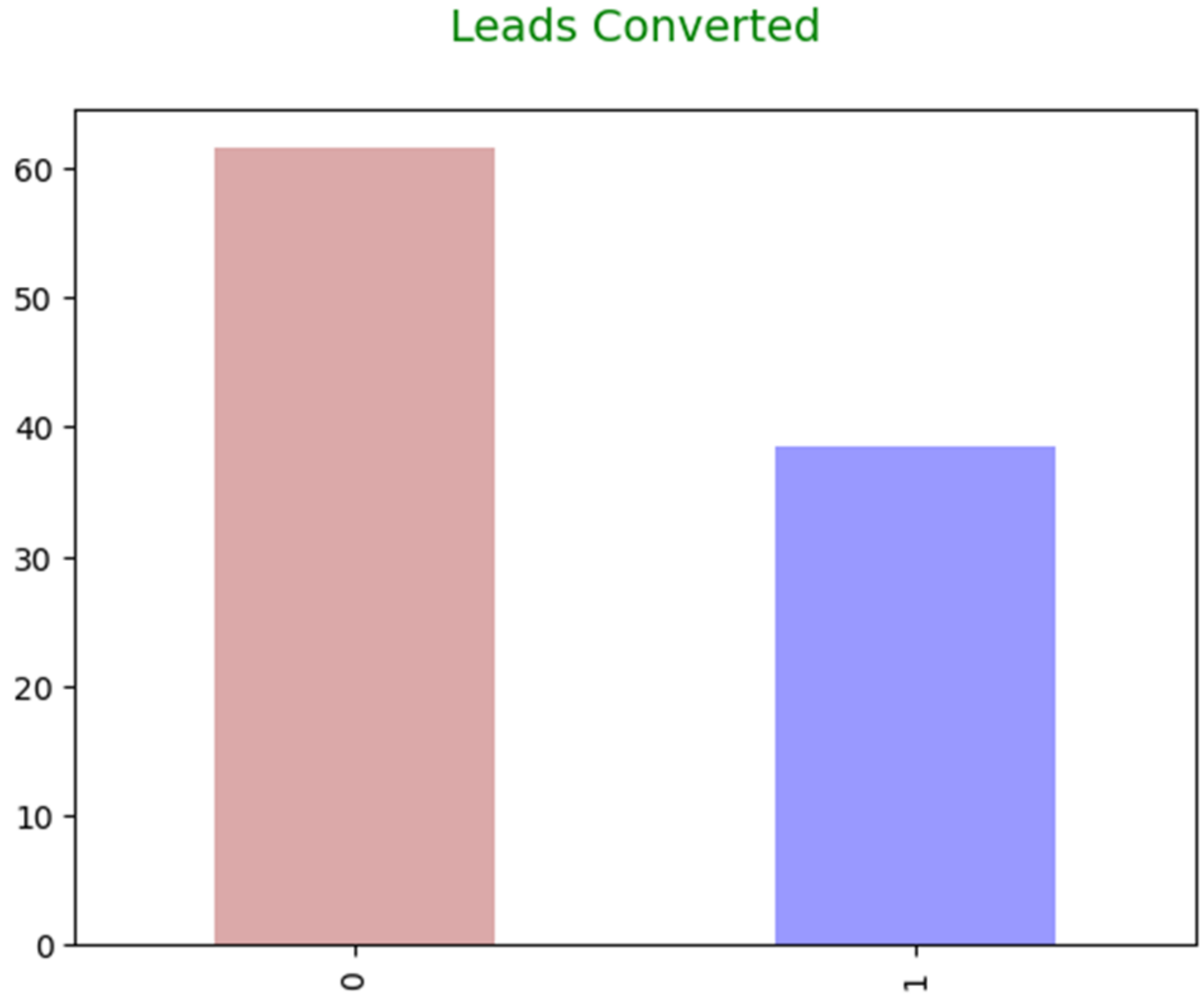

Social Media Marketing
Initial Pool of Leads
Lead Nurturing
Hot Leads
£
€
₹
$
Converted Leads

Business Objectives

- 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 hot, i.e. is most likely to convert whereas a lower score would mean that the lead is cold and will mostly not get converted.

Only 38.54% are converted into leads
Other 61.46% are not converted into leads

Lead Converted Ratio



ALGORITHM

The following steps are performed:

1. Reading the dataset
2. Data Cleaning
3. Exploratory Data 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 the Train set
9. Model Evaluation for Test set
10. Results

1. Reading the dataset

1. Reading the dataset
    - Conversion of data into dataframe
    - Examining the dataframe

2. Data Cleaning

- Replacing "Select" as NULL values
- Calculating the percentage of null values in each column
- Removing/Dropping of columns whose percentage of null values is greater than 40%
- Checking the dimensions of the dataframe after removing 40% null values columns
- Checking columns whose null values percentages are less than 40%
- Handling Missing values:
    - A. Segregation of numerical and categorical column
    - B. Imputation on the numerical and categorical column
- Checking for any other null values in the dataframe.
- Removing unwanted columns
    - A. Removing columns that do not give relevant information
    - B. Removing columns which have only one unique value

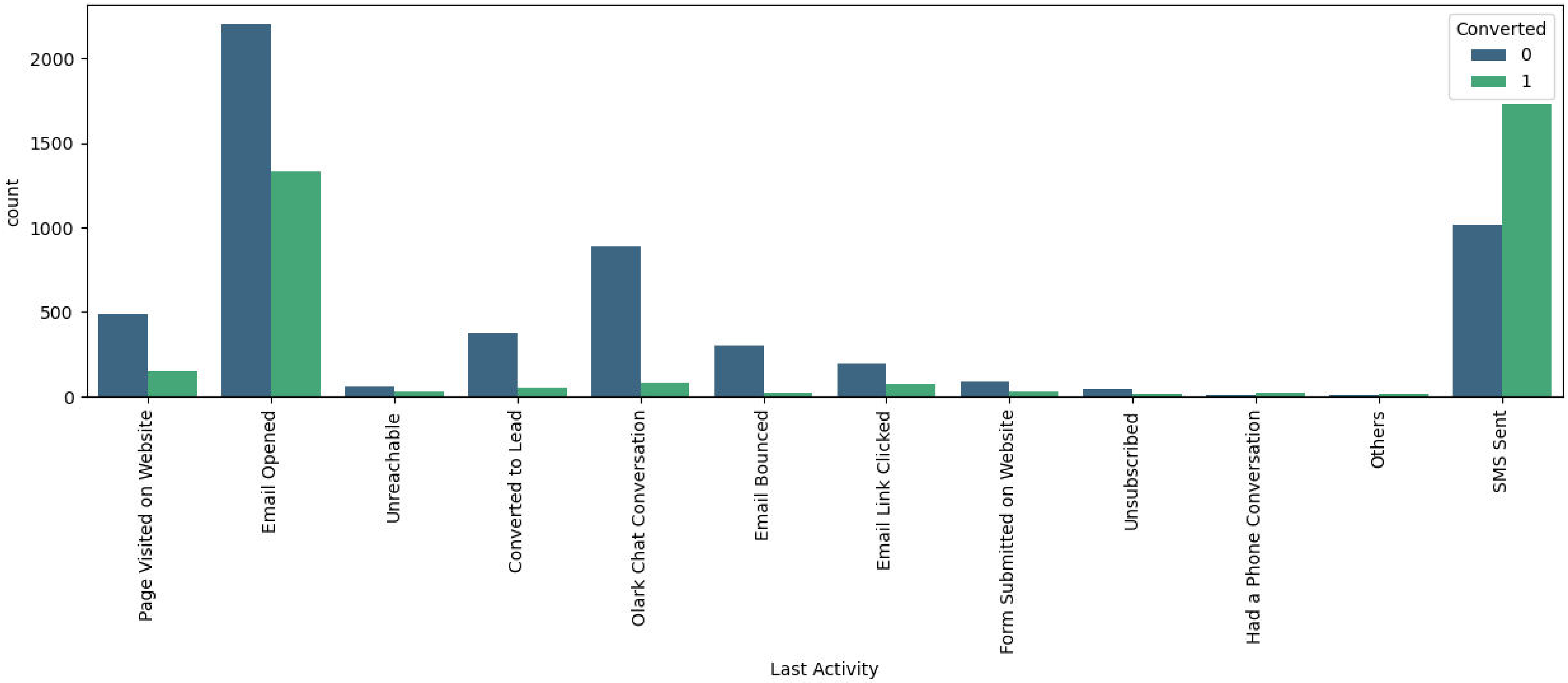
3.Exploratory Data Analysis

- Checking Data Imbalance percentage and Lead Converted Ratio
- Univariate Analysis

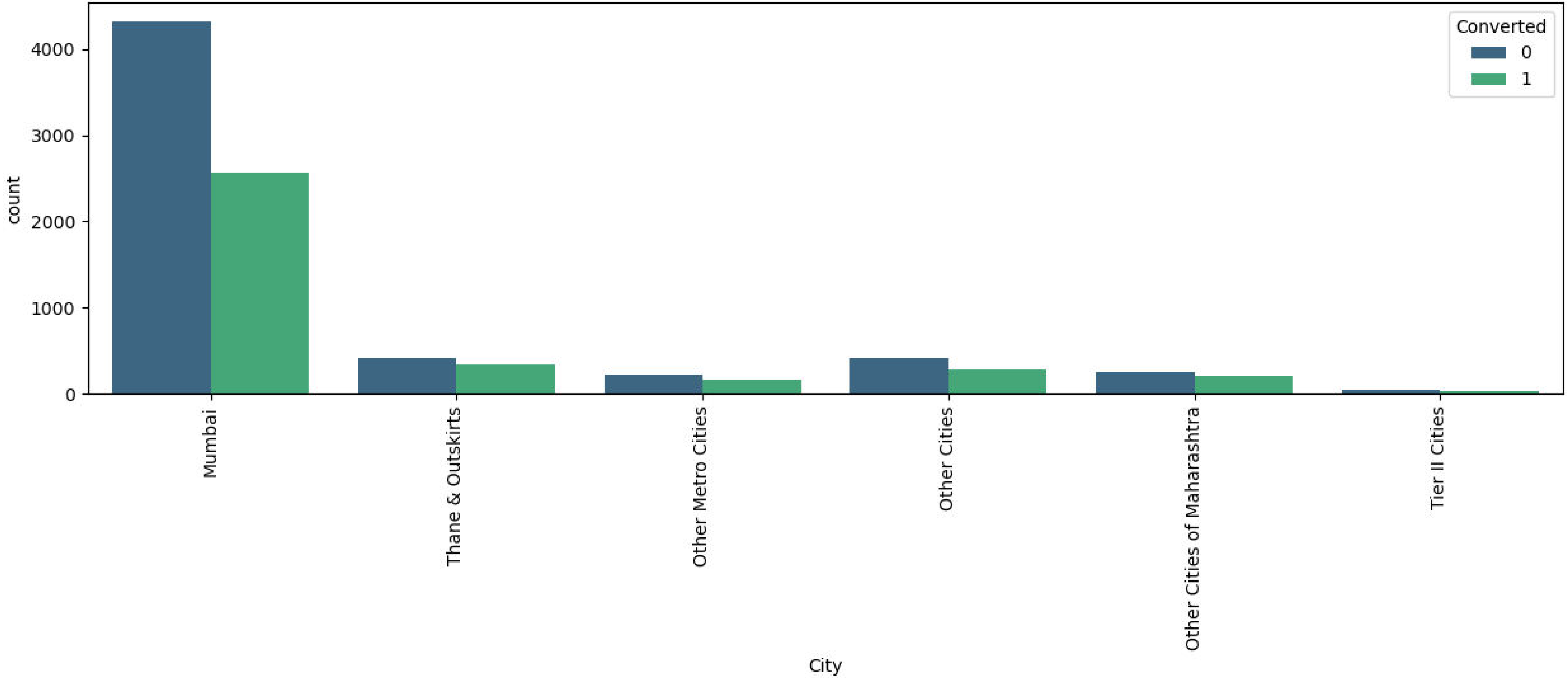
  A. Categorical columns

  B. Numerical columns
- Bivariate Analysis for numerical variables.

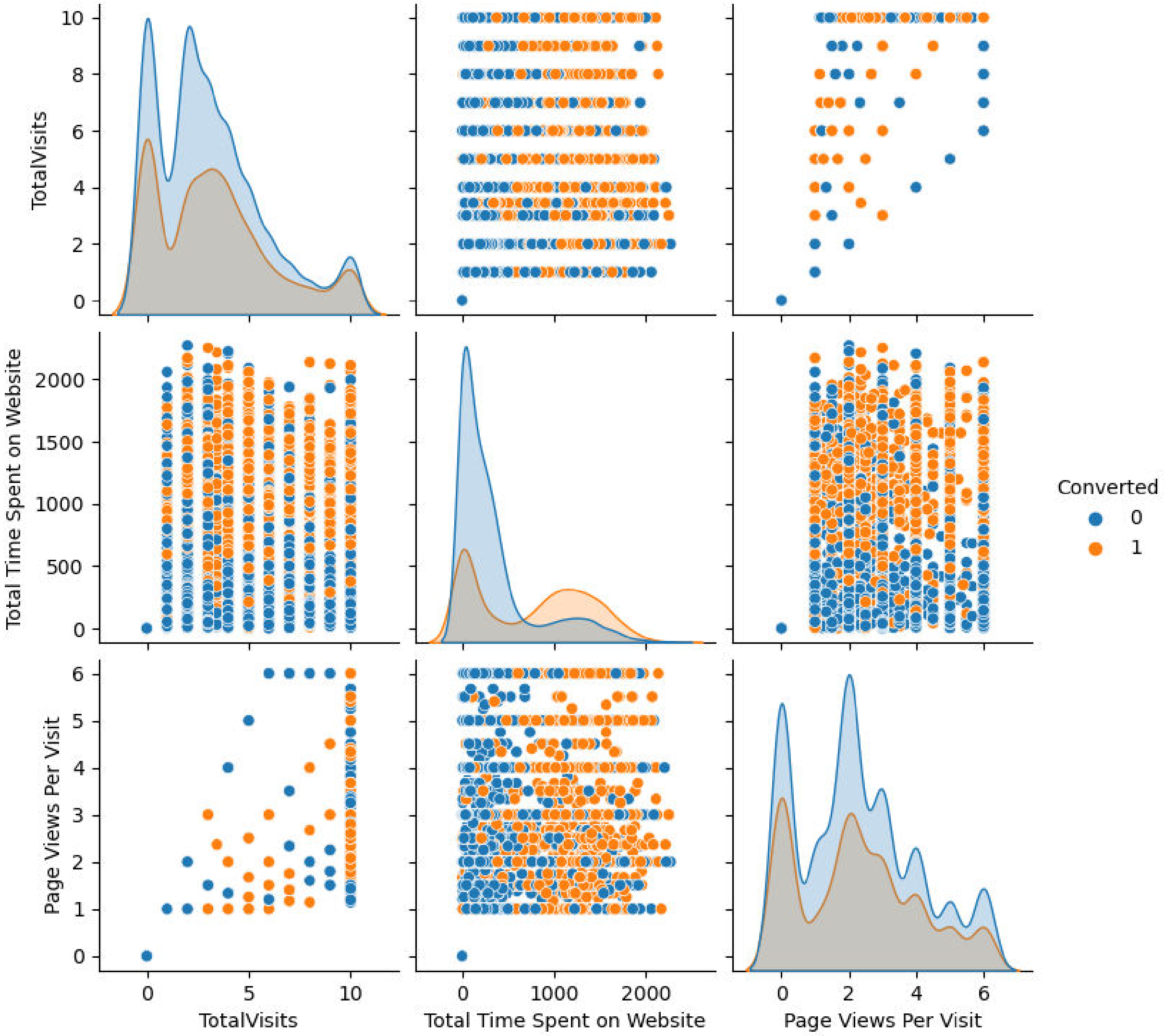
Univariate Analysis – Last Activity



Univariate Analysis – City



Bivariate Analysis – Numerical columns



4. Data Preparation

- Mapping of binary categorical variables
- Checking the datatypes of columns
- Creation of dummy variables

5. Splitting of Train and Test set
6. Scaling of features
7. Model Building Using Statsmodel and RFE

8.Model Evaluation for Train set

- Obtaining predicted value on y_train set
- Creating a dataframe with the actual converted flag and predicted probabilities
- Creating a threshold value for the new column "predicted"
- Creating a Confusion matrix

    A. Accuracy

    B. Sensitivity

    C. Specificity

    D. False positive rate

    E. Positive predictive value

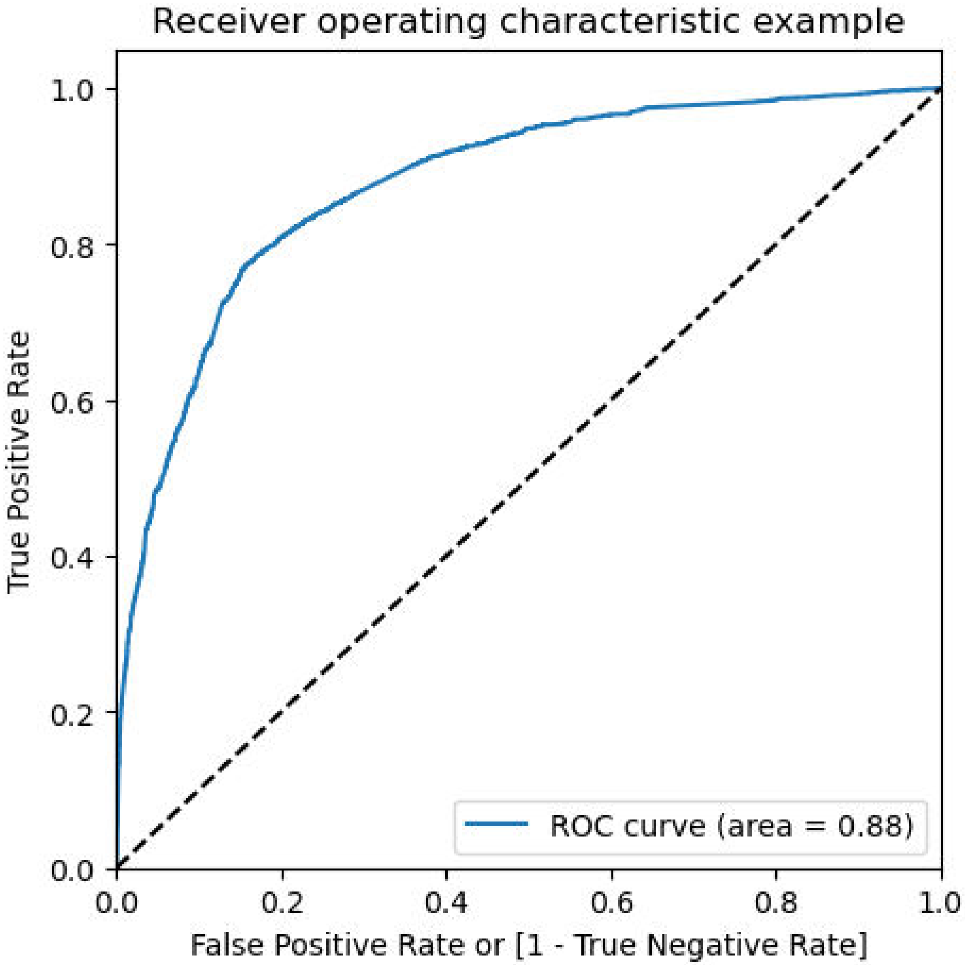
    F. Negative predictive value

    G. Precision

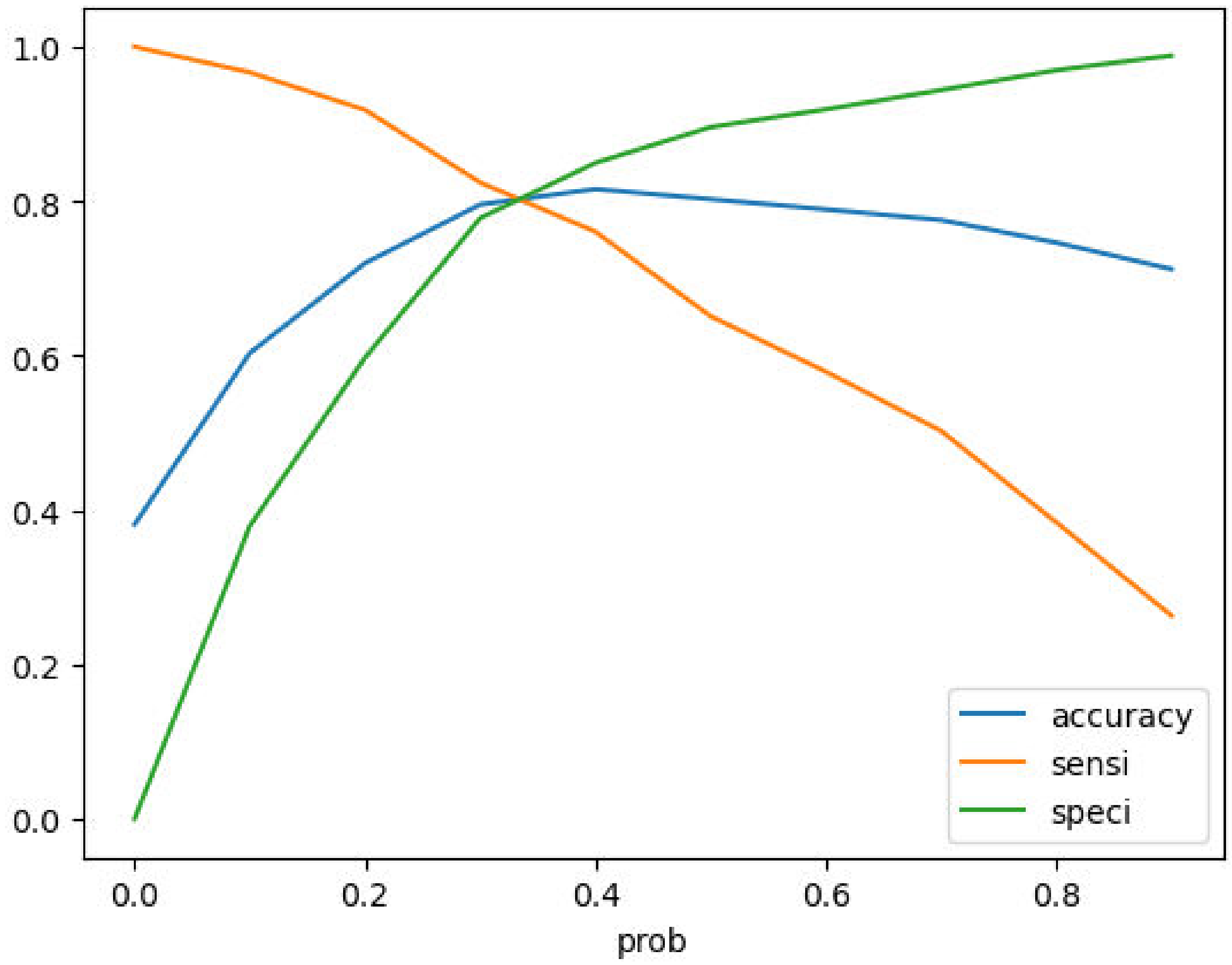
    H. Recall

- 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 the Precision-Recall Trade-off method
- Comparing the metrics values from the Optimal cut-off point method and the Precision-Recall Trade-off method

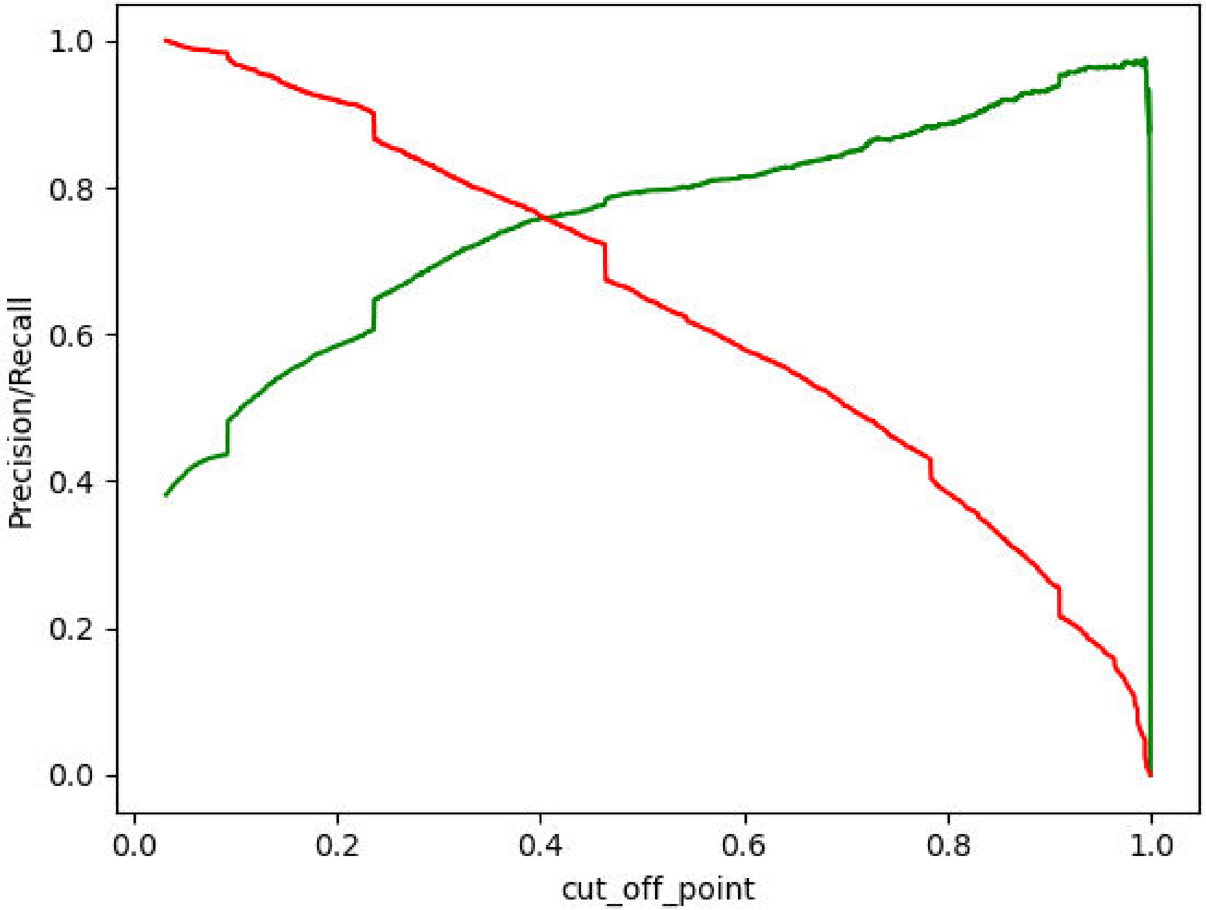
Area under ROC curve – 0.88



Determining optimal cut-off point probability – 0.35



Cut-off point probability using Precision–Recall method



9. Model Evaluation for Test set

- Prediction on Test Dataset
- Test model Evaluation
- Comparision of train set and Test set values
- Adding Lead_Score to test data

10. Results

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

10. Results

1. The lead conversion rate is approximately 80%

Metrics		Train Set	Test Set
Accuracy	:	80.94	80.16
Sensitivity	:	79.24	78.02
Specificity	:	81.98	81.51

2. Hot leads:

The customers who have lead_Score greater than or equal to 85 are termed as Hot leads.

From this model, 357 customers could be converted into leads

10. Results

3. Prospect ID of the customers:

2376, 2935, 2907, 8429, 1200, 5638, 6666, 5448, 1287, 8103, 3444, 2392, 5363, 8499, 4830, 7306, 3192, 2451,
1365, 6687, 5793, 8099, 4868, 8120, 2844, 7396, 88, 7683, 6860, 4058, 269, 5666, 8113, 4645, 2481, 3518, 1965,
3845, 2946, 4869, 7627, 954, 4902, 5818, 2804, 446, 4786, 8348, 1026, 76, 5825, 8055, 2266, 2926, 2524, 1867,
7503, 1467, 5697, 6725, 6156, 2055, 2549, 2653, 3478, 5687, 5832, 3190, 77, 9026, 6243, 4038, 7187, 5812,
4646, 7033, 3188, 8556, 7818, 1675, 3321, 2515, 1973, 7053, 1350, 8576, 6632, 7877, 7334, 6375, 7222, 5586,
7482, 6383, 1425, 8904, 3172, 2158, 3919, 5784, 3455, 5942, 6046, 472, 833, 4612, 2670, 9087, 3456, 2688,
2914, 5263, 4613, 3945, 8098, 2662, 4281, 6010, 7636, 507, 2631, 7448, 2578, 8920, 8412, 3339, 8054, 8082,
3113, 4607, 8087, 3488, 8888, 2764, 6760, 3120, 7963, 5671, 8901, 7570, 8110, 4803, 5571, 8641, 918, 3244,
8451, 818, 7453, 6987, 4285, 8495, 7433, 1283, 6778, 7055, 7150, 220, 2122, 7191, 8687, 819, 8042, 6158, 4971,
5979, 7467, 2489, 6457, 8107, 6784, 6092, 8117, 4180, 1254, 9162,

4812, 2600, 6736, 4116, 6884, 64, 1470, 6932, 7691, 7216, 7672, 6008, 1573, 7977, 2410, 5300, 5700, 7814,
7941, 6171, 8413, 1534, 1995, 7653, 7753, 2727, 8591, 5365, 1436, 7039, 2131, 1248, 4660, 3679, 8959,
3424, 939, 2391, 2976, 5802, 4707, 5562, 2538, 2014, 6822, 4860, 1575, 4005, 4879, 4592, 3019, 4050,
8097, 5307, 1267, 5459, 4461, 7094, 449, 3001, 2852, 5026, 2358, 8314, 3267, 8106, 7537, 270, 1379, 8650,
2397, 4778, 318, 3034, 829, 1899, 5353, 4775, 8509, 459, 5915, 2426, 6913, 5576, 7268, 3248, 8434, 8268,
9062, 6422, 1263, 5159, 7947, 8187, 133, 1490, 3736, 2085, 8146, 6729, 6362, 373, 4771, 3180, 4573, 5084,
3168, 4438, 6953, 8075, 3932, 5443, 3660, 8745, 7789, 1614, 2651, 7082, 7652, 5346, 9022, 7552, 3723,
5807, 6126, 6157, 4112, 3851, 4913, 3926, 134, 5921, 4480, 2614, 5183, 5960, 4442, 7905, 7469, 1860,
8088, 4360, 5439, 4935, 5206, 7666, 4157, 8282, 4955, 1625, 3291, 5210, 785, 6011, 2675, 4675, 1190,
4466, 7327, 8204, 6332, 7719, 5390, 3310, 2765, 6947, 260, 6663, 1588, 8897, 4013, 4407, 4941, 5418,
8092, 8761, 5362, 5741, 6944, 2152, 2960

10. Results

4. Features of the model

- Lead Origin_Lead Add Form
- Last Activity_Had a Phone Conversation
- current_occupation_Working Professional
- Last Activity_SMS Sent
- Lead Source_Welingak Website
- Last Activity_Others
- Lead Source_Olark Chat
- Last Activity_Unreachable
- Last Activity_Email Opened
- Total Time Spent on Website
- Specialization_Rural and Agribusiness
- Specialization_Finance Management

Suggestions:

- High lead conversation rates are possible for the "Lead Origin_Lead Add Form".
- The company should focus on the customers who had a last activity of phone conversation
- Working professional has a higher rate of lead conversation rate
- The company should concentrate on sending SMS to customers to convert them into leads
- Lead source like "Welingak Website" and "Olark Chat" to be focused
- The company should send email to convert into leads
- The customers who spent time on the website have a higher lead conversion rate. The company should know how to convert customers into leads when they enter into website
- The customers whose specializations are "Rural and Agri business" and "Financial management" are likely to be converted into leads.