Lead Scoring Case Study using Machine Learning

-By

Sahil Gera
Ch. Chaitanya

Problem Statement

- An education company named X Education sells online courses to industry professionals
- X Education gets a lot of leads, its lead conversion rate is very poor. For example, if, say, they acquire 100 leads in a day, only about 30 of them are converted. 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. Business

Objective:

1. X education wants to know most promising leads.
2. To attain there Targets they want to build a logistic Regression Model which identifies the “HOT LEADS”.

Approach to the Problem with given Data Set

❖ Data cleaning and data manipulation.

1. Check and handle duplicate data.
2. Check and handle NA values and missing values.
3. Drop columns, if it contains large amount of missing values and not useful for the analysis.
4. Imputation of the values, if necessary.
5. Check and handle outliers in data.

❖ EDA

1. Univariate data analysis: value count, distribution of variable etc.
2. Multi- variate data analysis: correlation coefficients and pattern between the variables etc.
3. Feature Scaling & Dummy Variables and encoding of the data.

❖ Classification technique

- Logistic regression used for the model making and prediction.
- Validation of the model.
- Model presentation.
- Conclusions and recommendations

Data Manipulations

lead_score.shape gives us Rows = 9240 and columns = 37

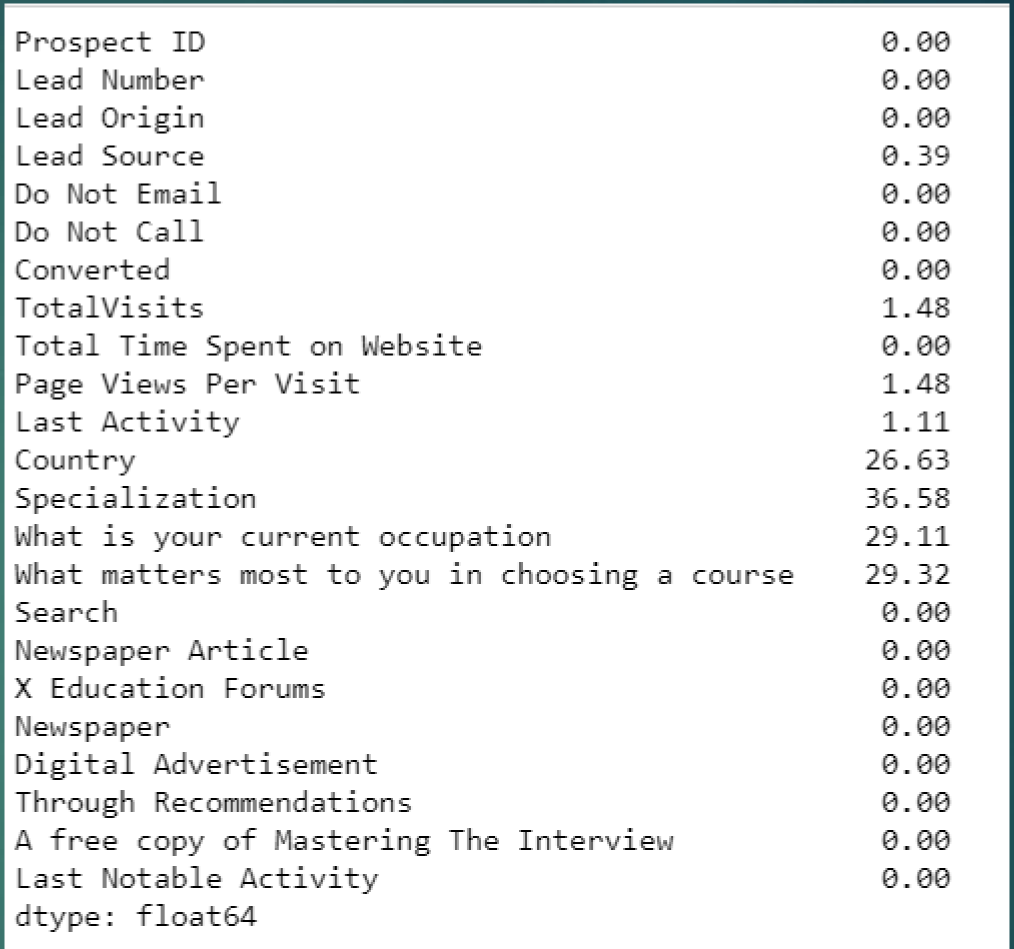
- **Removing the unique values columns**:
- After inspection of Data set we found these columns are not necessary for our analysis
- **Dropping these Columns:**

['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’]

- **Checking for Null values**
- According to the Industry standards, If there are more than 30 % null - values in any column , then we have to drop that column.
- **Dropping these columns:**

['How did you hear about X Education’, 'Tags’, 'Lead Quality','Lead Profile','City’, 'Asymmetrique Activity Index', 'Asymmetrique Profile Index’, 'Asymmetrique Activity Score’, 'Asymmetrique Profile Score']

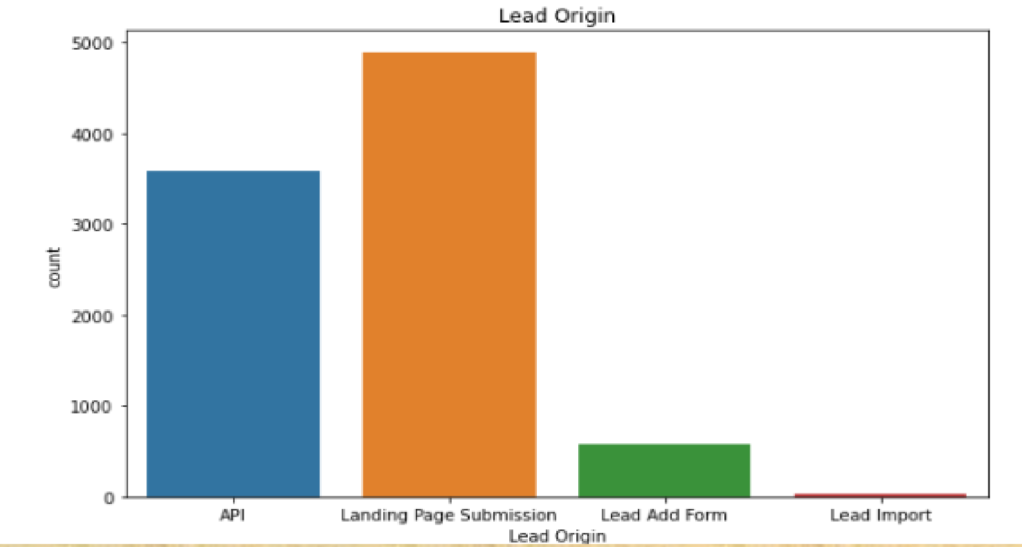
❖ **Manipulating some Null Value Columns as there are very important for further Analysis**

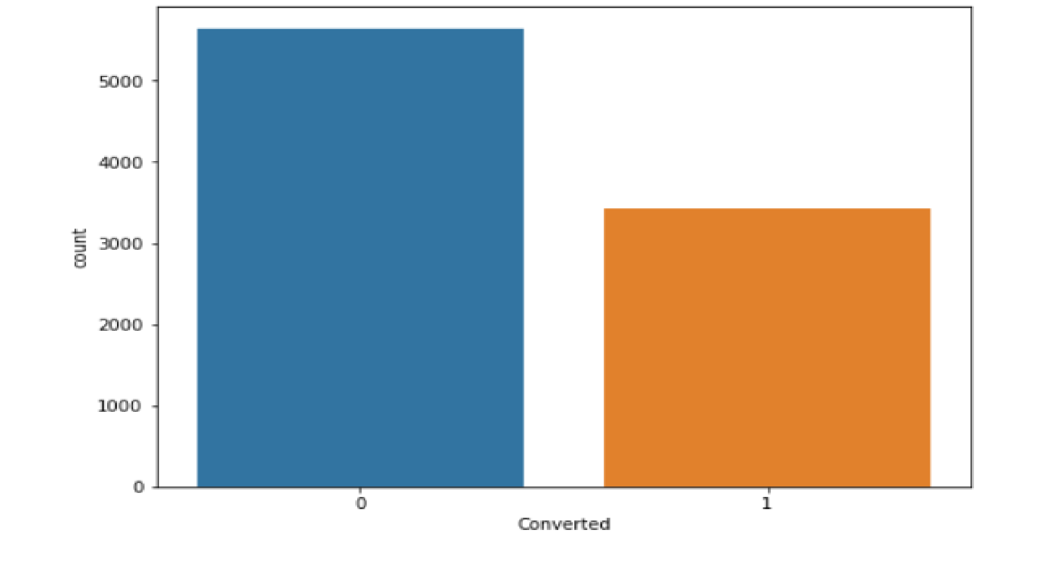


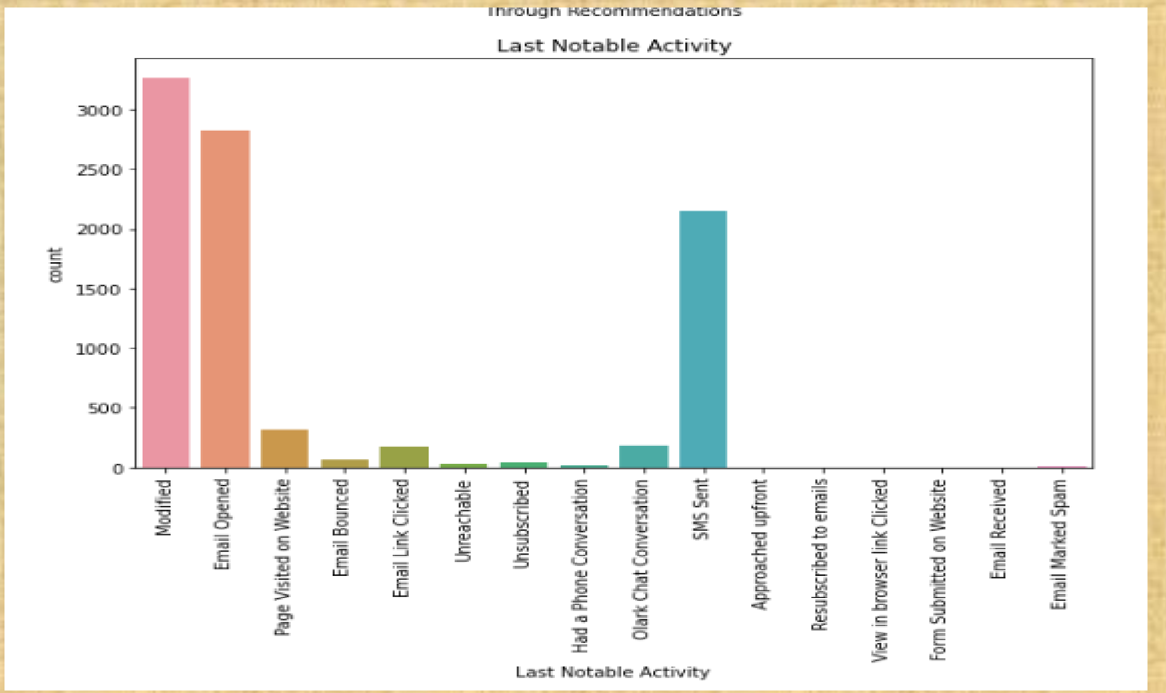
```
Prospect ID                                       0.00
Lead Number                                       0.00
Lead Origin                                       0.00
Lead Source                                       0.39
Do Not Email                                      0.00
Do Not Call                                       0.00
Converted                                         0.00
TotalVisits                                       1.48
Total Time Spent on Website                       0.00
Page Views Per Visit                              1.48
Last Activity                                     1.11
Country                                          26.63
Specialization                                   36.58
What is your current occupation                  29.11
What matters most to you in choosing a course    29.32
Search                                            0.00
Newspaper Article                                 0.00
X Education Forums                                0.00
Newspaper                                         0.00
Digital Advertisement                             0.00
Through Recommendations                           0.00
A free copy of Mastering The Interview            0.00
Last Notable Activity                             0.00
dtype: float64
```

- Here we see that some of the null values are still there
- As these columns are so important , we cannot Drop these values in the columns,
- so we have to impute the null values with some logical values. We are replacing the null values with “Unknown” as mode doesn’t make any sense for these columns

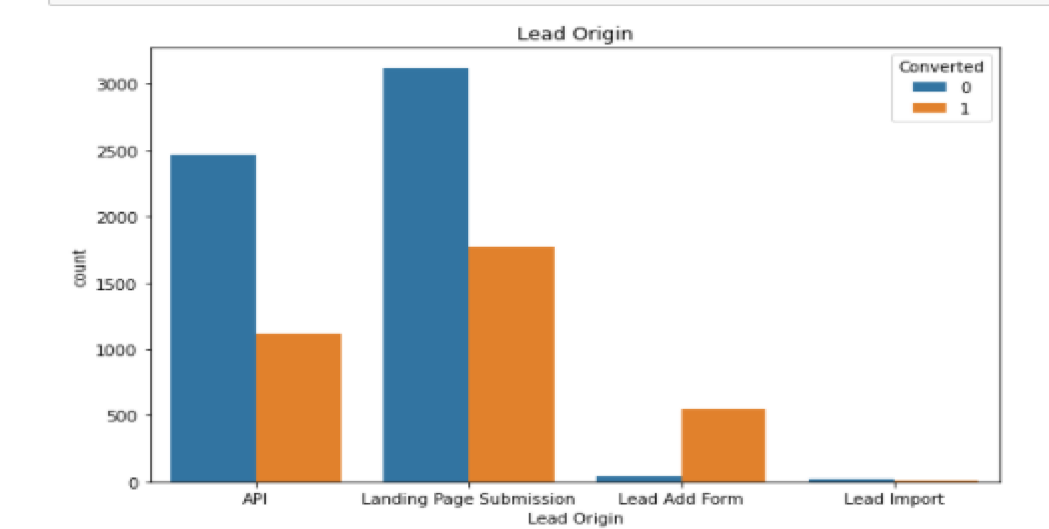
Data visualization: Uni-Variant Analysis

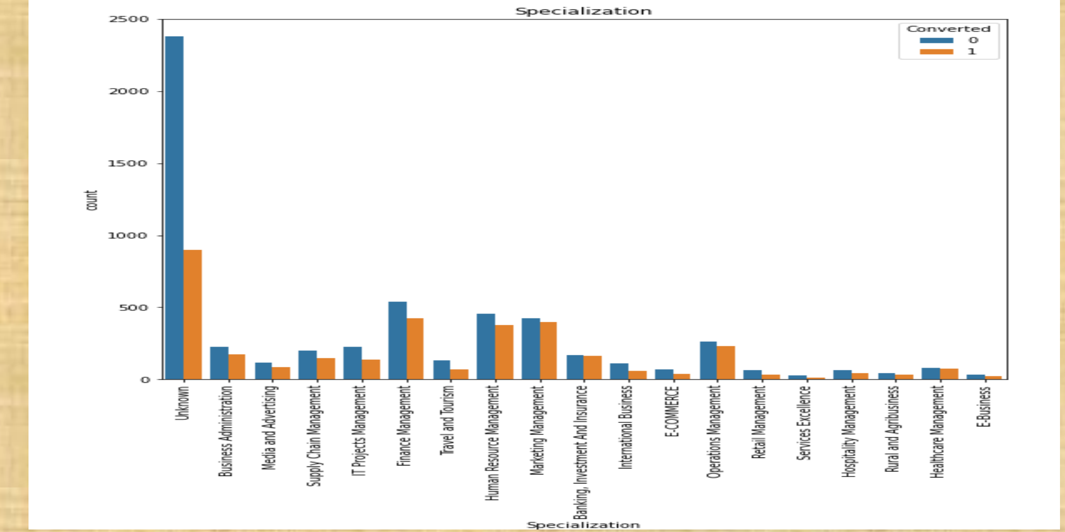


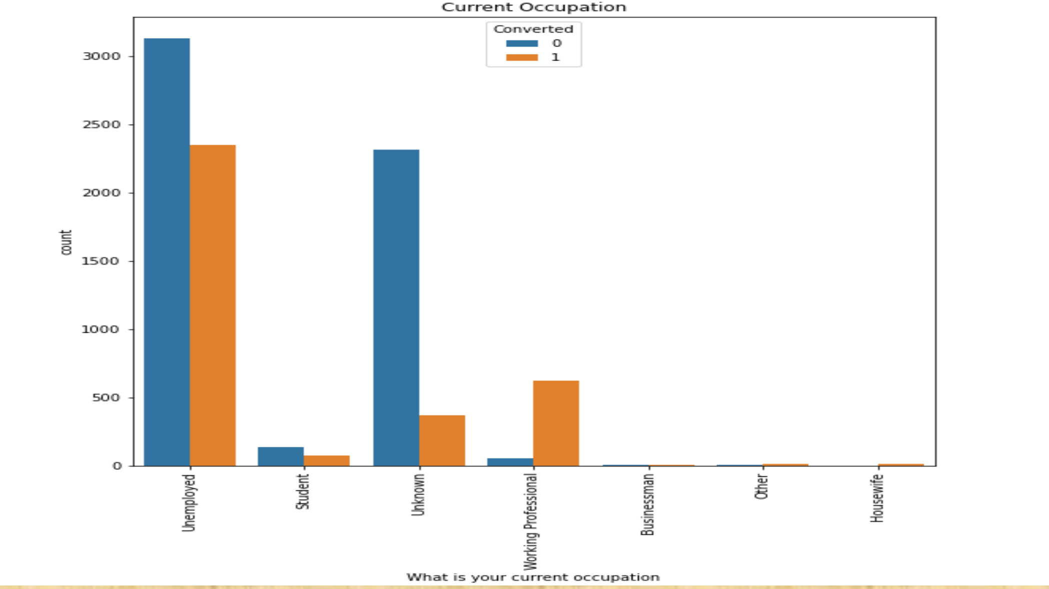


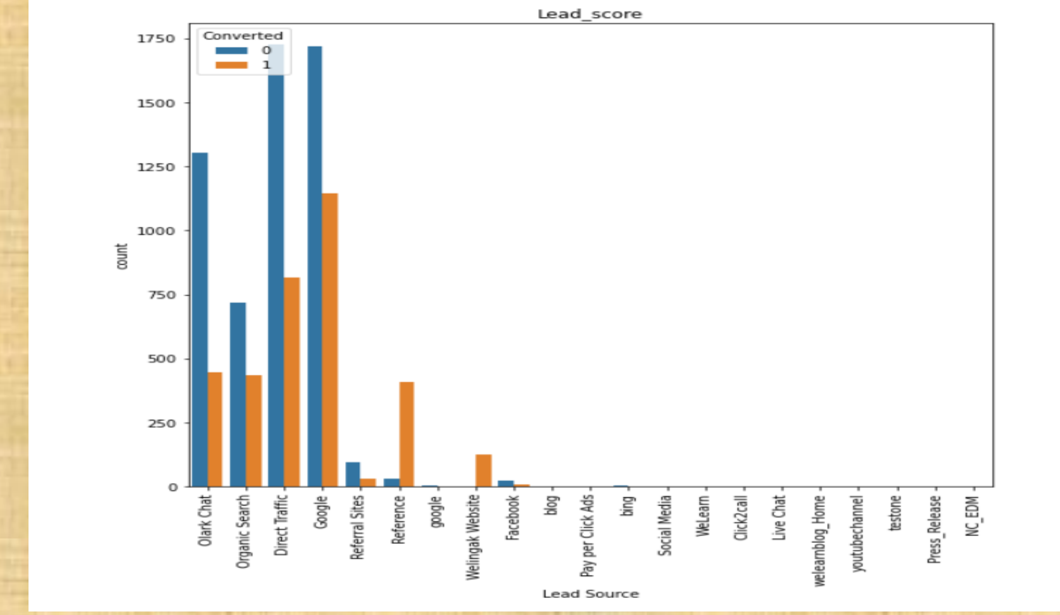


Multi-Variant Analysis



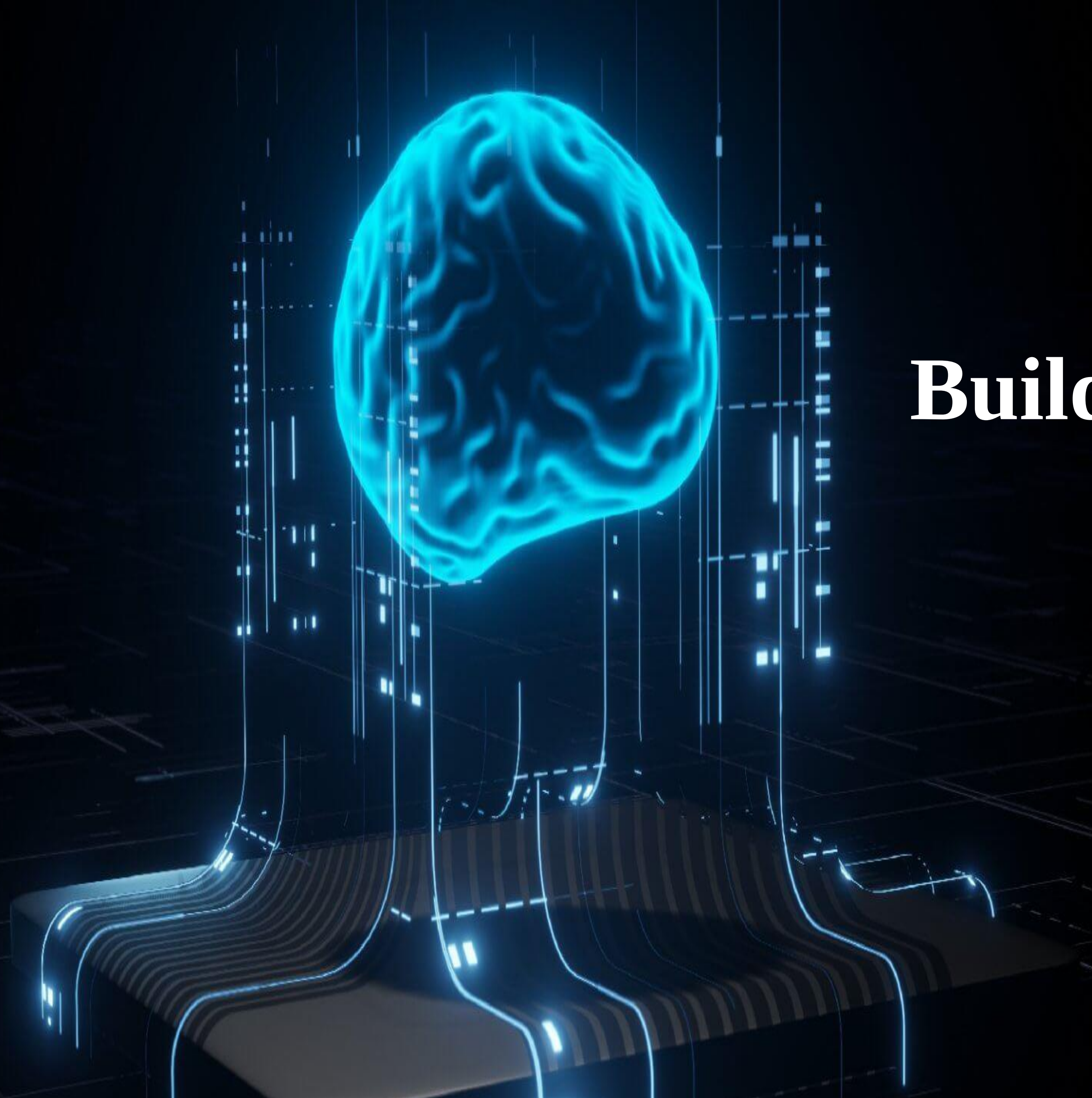






Data Preparation For Logistic Regression Analysis

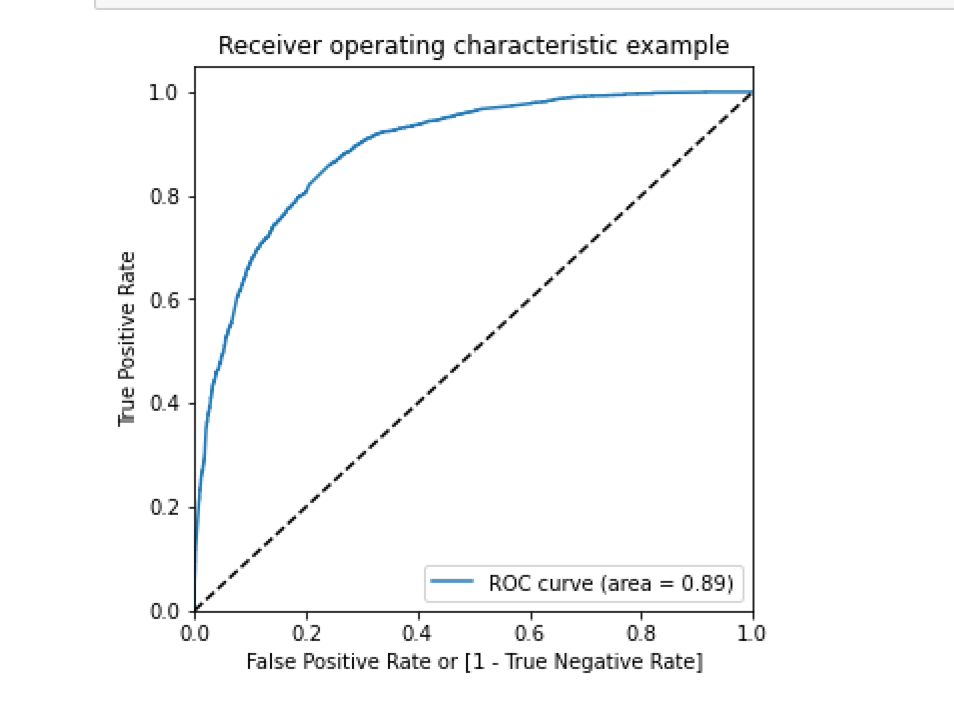
- Numerical Variables are Normalized
- Dummy Variables are created for object type variables
- Total Rows for Analysis: 9072
- Total Columns for Analysis: 132

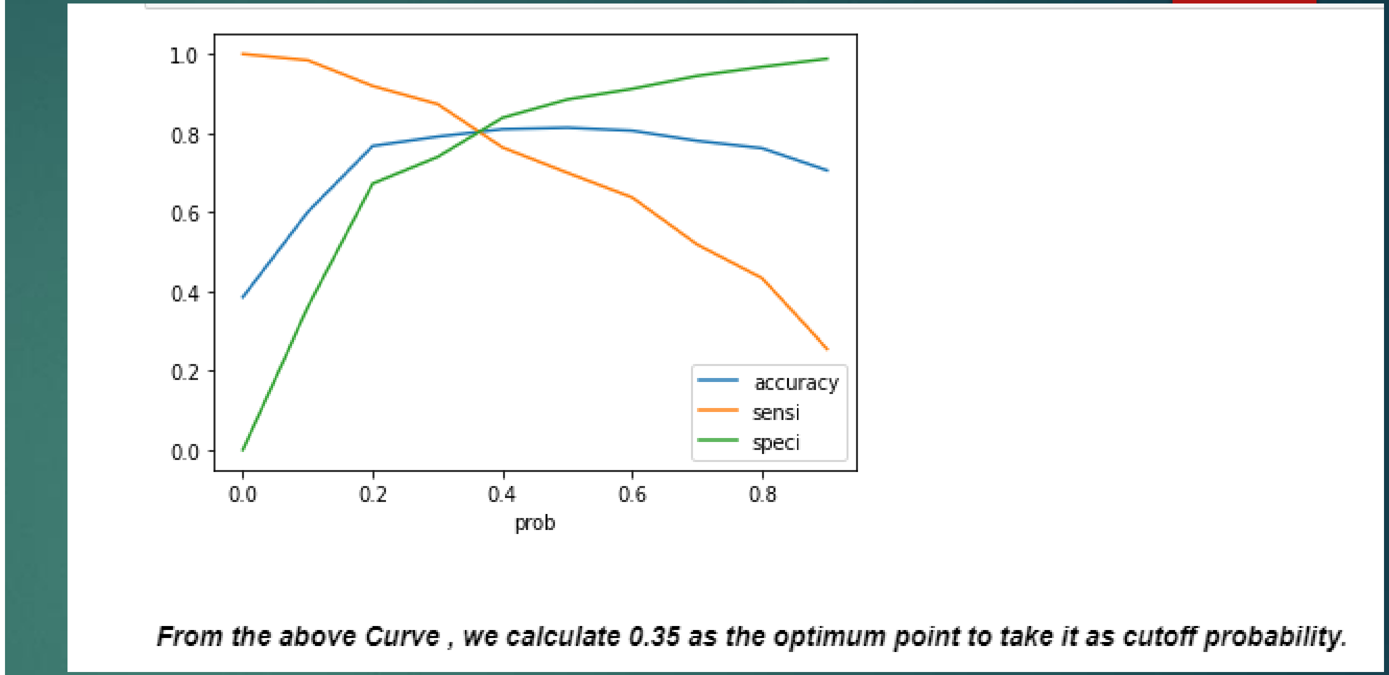


Building ML Model

❖ Splitting the Data into Training and Testing Sets

- The first basic step for regression is performing a train-test split, we have chosen 70:30 ratio.
- **Use RFE for Feature Selection**
- **Running RFE with 15 variables as output**
- Building Model by removing the variable whose p-value is greater than 0.05 and VIF value is greater than 5
- Predictions on test data set
- Overall accuracy 82%





From the above Curve , we calculate 0.35 as the optimum point to take it as cutoff probability.

Finding Optimal Cut off Point:

- Optimal cut off probability is that
- probability where we get balanced sensitivity and specificity.
- From the second graph it is visible that the optimal cut off is at 0.35.

Conclusion And recommendations

The variables that affects our lead mostly are the following:

1.The TotalVistits
2.The Total Time Spend on the Website
3. Last Notable Activity_Had a Phone Conversation
4. When the last activity was:
  a). SMS sent
  b). Last Activity_Olark Chat Conversation
5.What is their current occupation_Working Professional
6.When Lead Source_Welingak Websit