Lead Score Analysis Project

Machine learning project for educational courses sales enhancement

Client : X Education

Prepared by : Vinay Chawla

No of slides : 10

Contents

Problem Statement

•**Challenge:** The average lead conversion rate is only **30%**, meaning a significant portion of leads do not convert.

•**Current Scenario:** Many professionals visit the website and fill out forms, becoming leads.
•The sales team then engages with these leads through calls and emails.

•**Objective:** Identify **Hot Leads**—the most promising leads—so the sales team can focus their efforts efficiently.

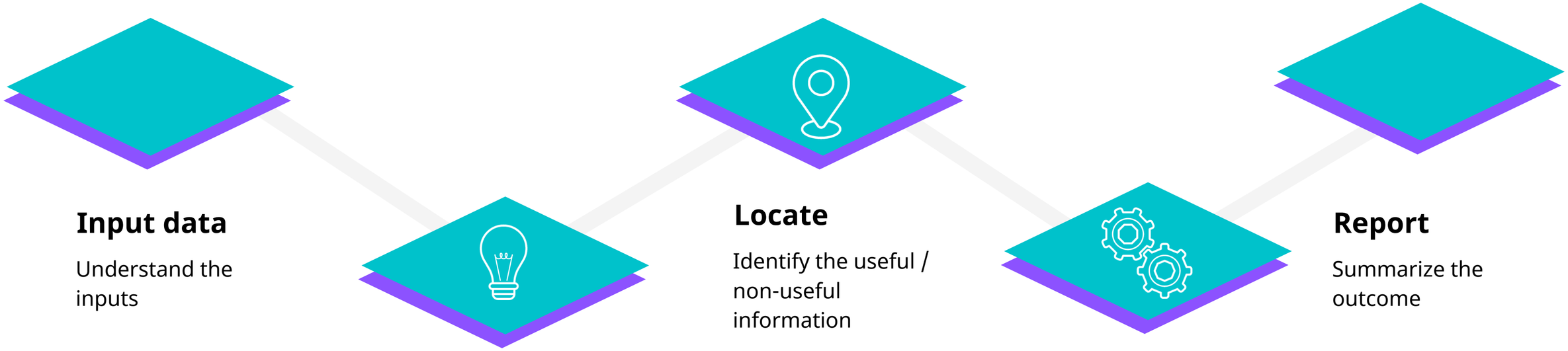
•**Expected Impact:** Improved lead conversion rates and better resource utilization, leading to higher sales efficiency.

Objective of Business

- Calculate Lead score
- Optimize the resources
- Increase Lead conversion to 80% from 30%
- Define strategies for peak and lull period

Establish the methodology

Overall Project Strategy



Input data

Understand the inputs

Refinement

Process data to make it Fit for ML Model

Locate

Identify the useful / non-useful information

Apply in Model

Fit on model and evaluate the outcome

Report

Summarize the outcome

Methodology - Detailing

Understanding & cleaning data

Understanding data

- Importing the data
- Inspecting data features, their data types etc

Cleaning the dataset

- Removing Nulls
- Removing outliers
- Grouping unwanted categories

Exploratory data analysis

- Understanding data through count plot, boxplot, scatterplots
- Identifying columns usability and removing if not useful

Data Modelling and Evaluation

Data Preparation

- Converting to Dummies
- Test-Train Split
- Feature Scaling
- Clearing Collinearity using VIF

Modelling

- Applying Logistic Regression
- Identify right Lambda
- Re-fitting the Model

Evaluation and Testing

- Finding Predictions from Model
- ROC curve , Confusion Matrix ,
- Optimizing threshold tradeoff
- Calculating Lead score

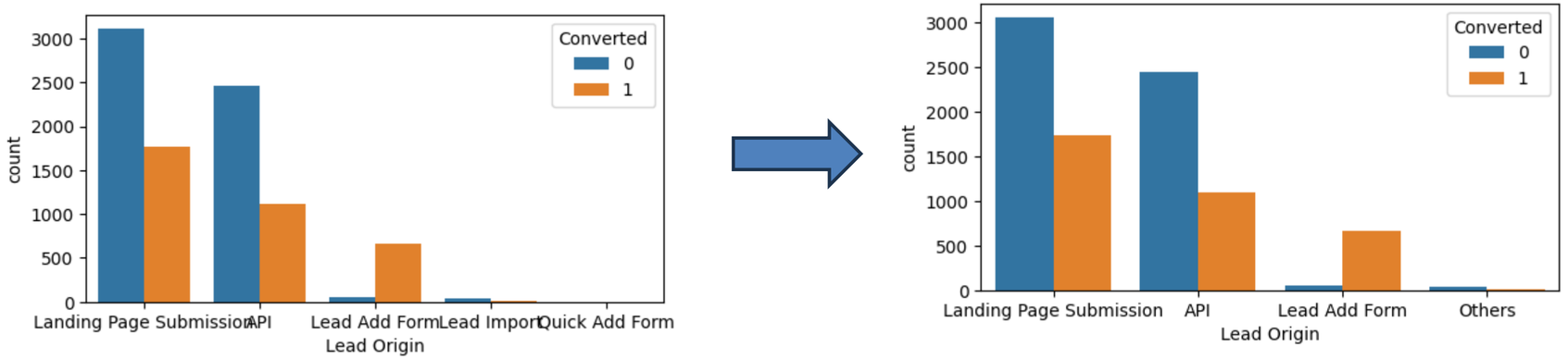
Report and Result

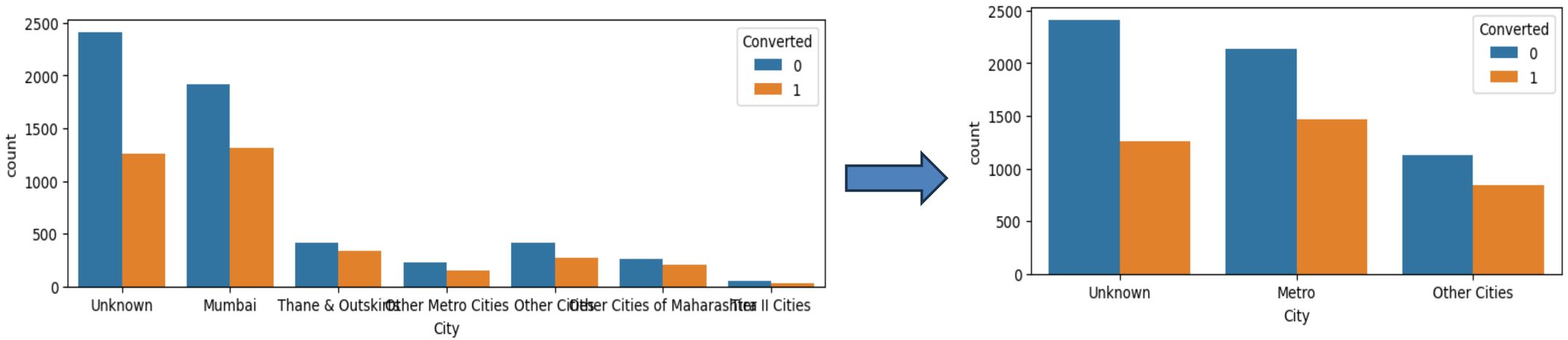
- Calculate Success rate

Observations : EDA

- Preliminary Data Cleaning
  - Removed Features with Nulls >45%
  - Removed Index columns
- Exploratory Data Analysis
  - Individual Column wise approach
  - Checking Categorical data
  - Removing Outliers
  - Grouping categorical data if needed

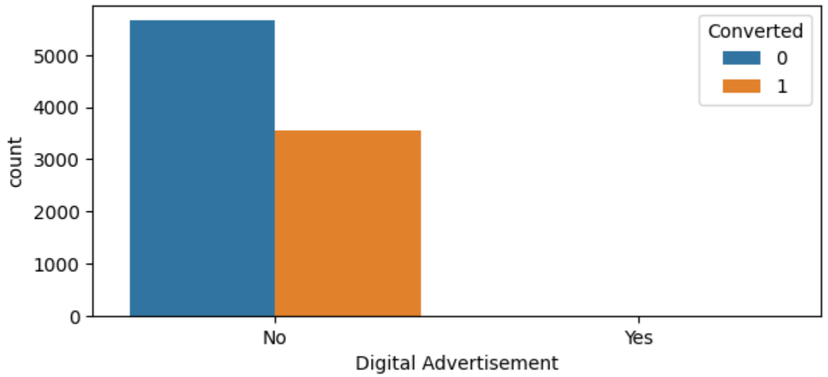
Examples of Category Grouping

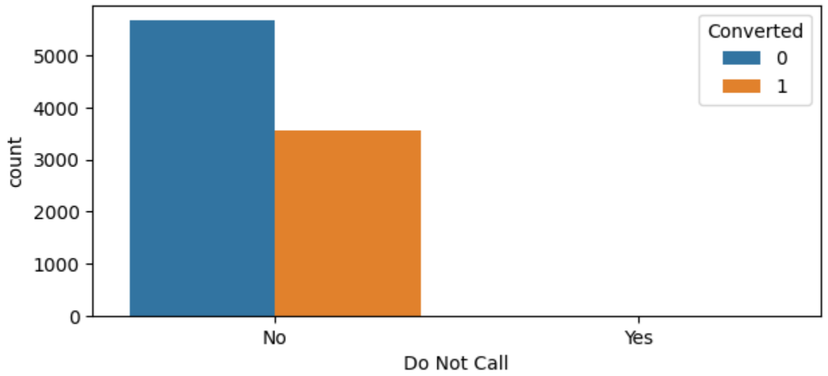


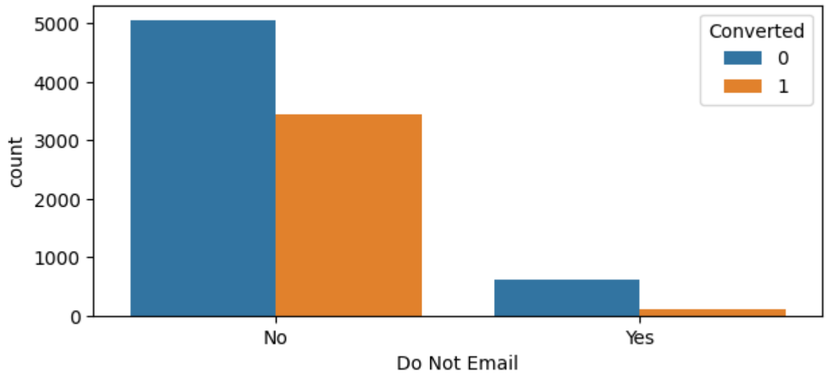


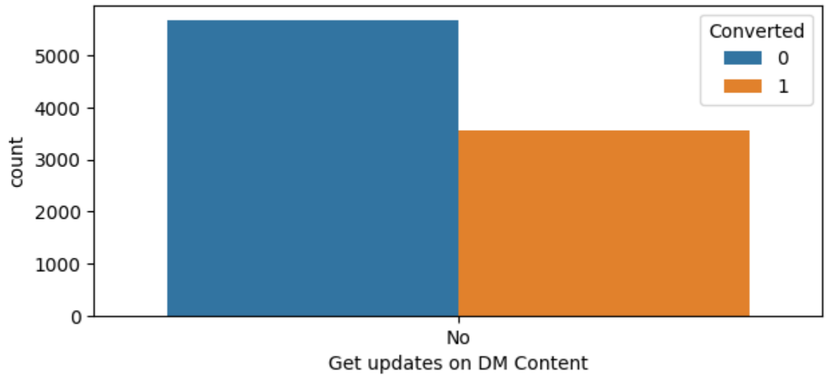
Observations : EDA

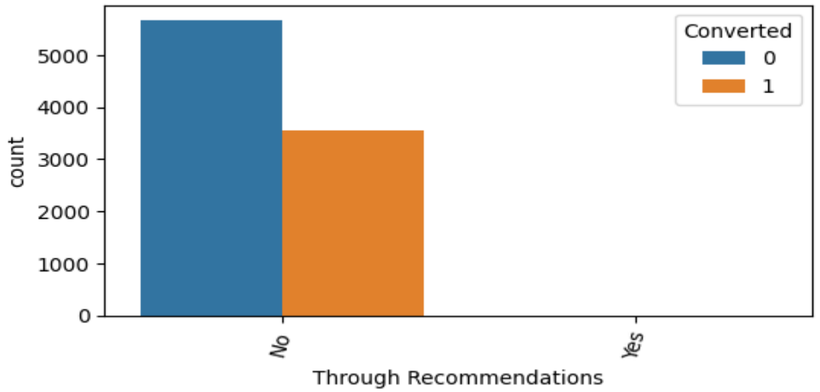
Feature dropped based on their data set nature







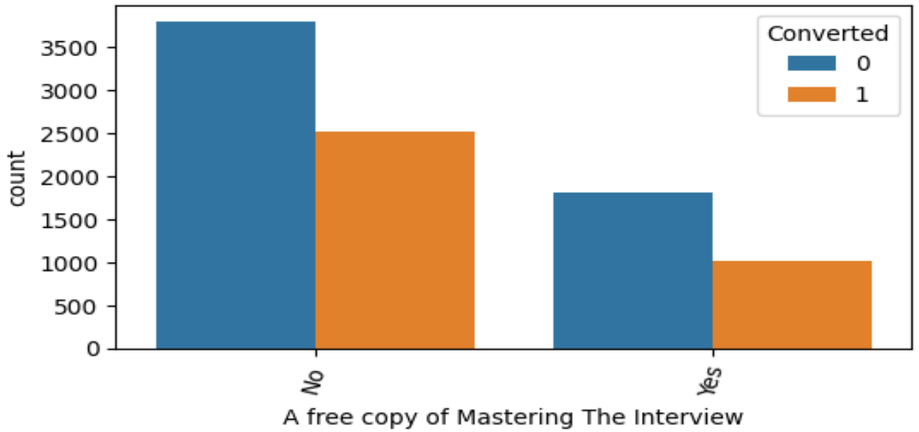


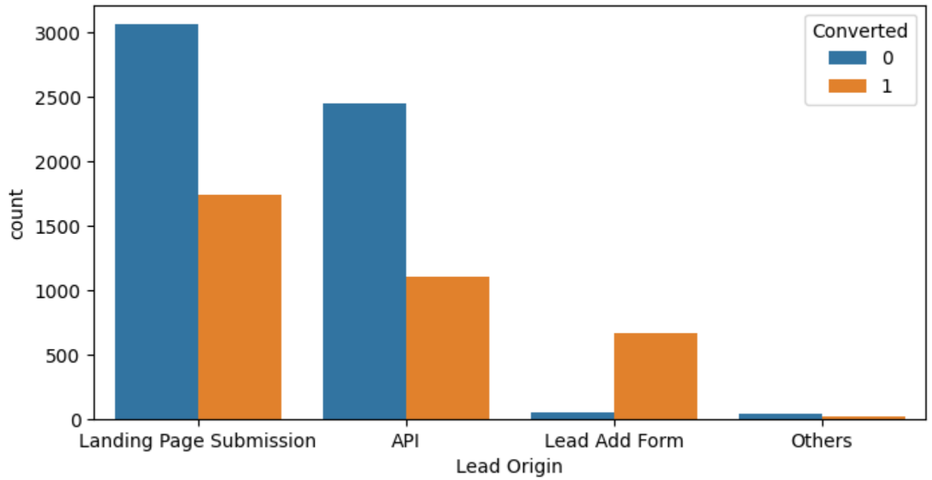


Such Features are dropped due to their Biased (highly imbalanced data) nature. Such data can cause wrong predictions in model.

Observations : EDA

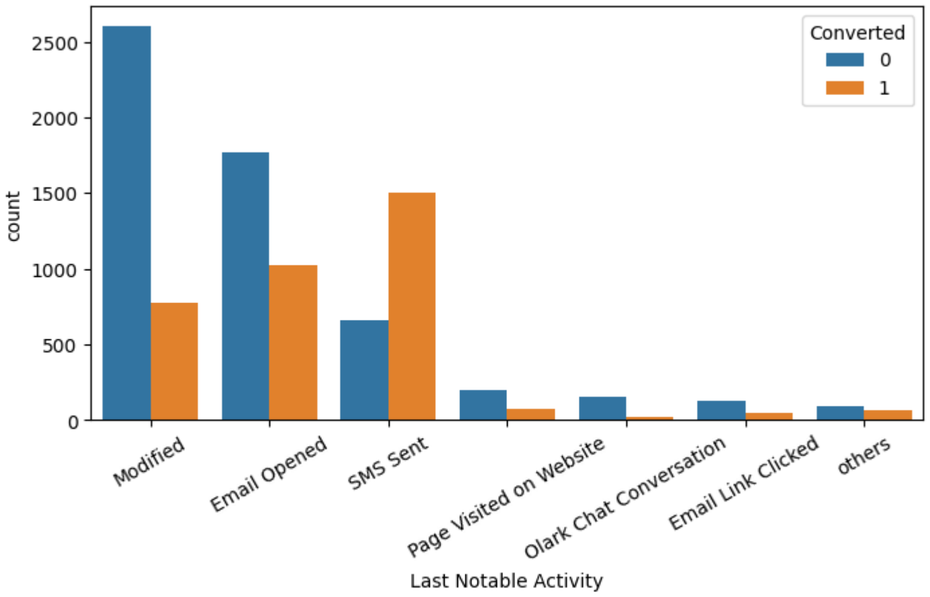
Features that were retained

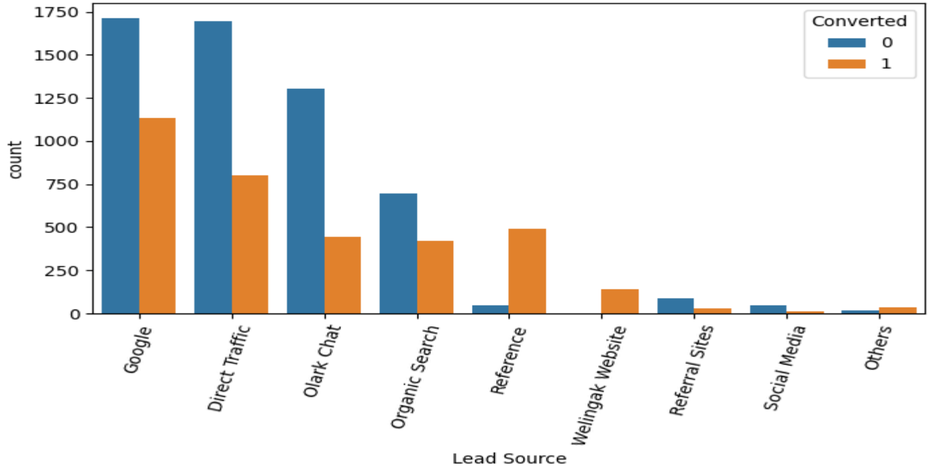




These features are retained with a bit of grouping as needed because the data isn't imbalanced, and the categories are minimal.

Moreover, data visualizations indicate these maybe some of useful features that may decide the outcome of Model.





Outcome of EDA

37 features got reduced to 13 !

```
1  #  2.2 Displaying data info
2  lead.info()
```

4] ✓ 0.0s

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 9240 entries, 0 to 9239
Data columns (total 37 columns):
 #   Column                                         Non-Null Count  Dtype
---  ------                                         --------------  -----
 0   Prospect ID                                    9240 non-null   object
 1   Lead Number                                    9240 non-null   int64
 2   Lead Origin                                    9240 non-null   object
 3   Lead Source                                    9204 non-null   object
 4   Do Not Email                                   9240 non-null   object
 5   Do Not Call                                    9240 non-null   object
 6   Converted                                      9240 non-null   int64
 7   TotalVisits                                    9103 non-null   float64
 8   Total Time Spent on Website                    9240 non-null   int64
 9   Page Views Per Visit                           9103 non-null   float64
 10  Last Activity                                  9137 non-null   object
 11  Country                                        6779 non-null   object
 12  Specialization                                 7802 non-null   object
 13  How did you hear about X Education             7033 non-null   object
 14  What is your current occupation                6550 non-null   object
 15  What matters most to you in choosing a course  6531 non-null   object
 16  Search                                         9240 non-null   object
 17  Magazine                                       9240 non-null   object
 18  Newspaper Article                              9240 non-null   object
 19  X Education Forums                             9240 non-null   object
...
 35  A free copy of Mastering The Interview         9240 non-null   object
 36  Last Notable Activity                          9240 non-null   object
dtypes: float64(4), int64(3), object(30)
memory usage: 2.6+ MB
```

Output is truncated. View as a scrollable element or open in a text editor. Adjust cell output settings...

```
1  lead.info()
```

✓ 0.0s

```
<class 'pandas.core.frame.DataFrame'>
Index: 9120 entries, 0 to 9239
Data columns (total 14 columns):
 #   Column                                 Non-Null Count  Dtype
---  ------                                 --------------  -----
 0   Lead Origin                            9120 non-null   object
 1   Lead Source                            9120 non-null   object
 2   Do Not Email                           9120 non-null   object
 3   Converted                              9120 non-null   int64
 4   TotalVisits                            9120 non-null   float64
 5   Total Time Spent on Website            9120 non-null   int64
 6   Page Views Per Visit                   9120 non-null   float64
 7   Last Activity                          9120 non-null   object
 8   Specialization                         9120 non-null   object
 9   What is your current occupation        9120 non-null   object
 10  Tags                                   9120 non-null   object
 11  City                                   9120 non-null   object
 12  A free copy of Mastering The Interview 9120 non-null   object
 13  Last Notable Activity                  9120 non-null   object
dtypes: float64(2), int64(2), object(10)
memory usage: 1.0+ MB
```

Data Preparation : Converting to Dummies

13 Features converted to 48 Dummies !

```
1  lead.info()
```

✓ 0.0s

```
<class 'pandas.core.frame.DataFrame'>
Index: 9120 entries, 0 to 9239
Data columns (total 14 columns):
 #   Column                                  Non-Null Count  Dtype
---  ------                                  --------------  -----
 0   Lead Origin                             9120 non-null   object
 1   Lead Source                             9120 non-null   object
 2   Do Not Email                            9120 non-null   object
 3   Converted                               9120 non-null   int64
 4   TotalVisits                             9120 non-null   float64
 5   Total Time Spent on Website             9120 non-null   int64
 6   Page Views Per Visit                    9120 non-null   float64
 7   Last Activity                           9120 non-null   object
 8   Specialization                          9120 non-null   object
 9   What is your current occupation         9120 non-null   object
 10  Tags                                    9120 non-null   object
 11  City                                    9120 non-null   object
 12  A free copy of Mastering The Interview  9120 non-null   object
 13  Last Notable Activity                   9120 non-null   object
dtypes: float64(2), int64(2), object(10)
memory usage: 1.0+ MB
```

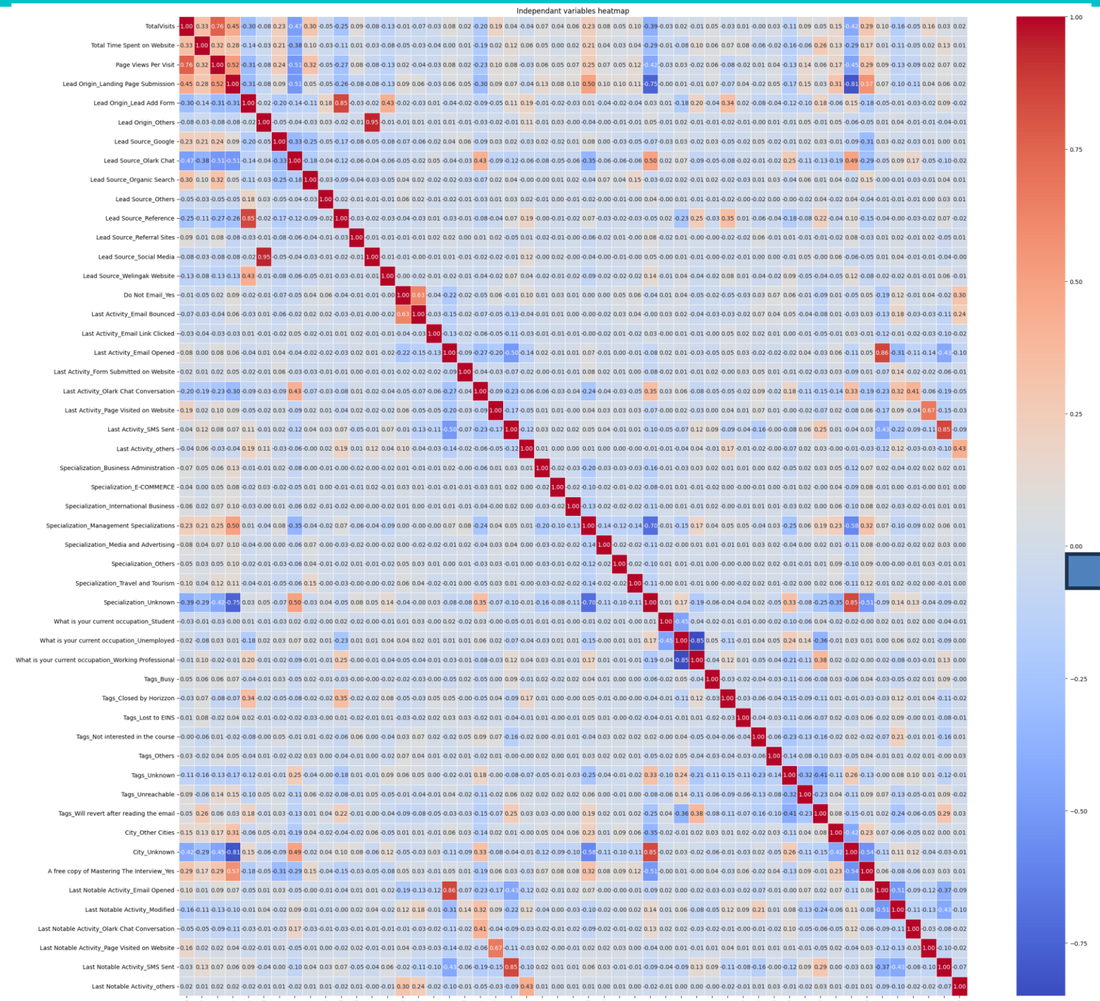
```
1  # Apply One-hot Encoding on Object features .
2
3  lead_dummy = pd.get_dummies(lead_obj, drop_first=True).astype(int)
4
5  lead_dummy.info()
```

✓ 0.0s

```
<class 'pandas.core.frame.DataFrame'>
Index: 9120 entries, 0 to 9239
Data columns (total 48 columns):
 #   Column                                  Non-Null Count  Dtype
---  ------                                  --------------  -----
 0   Lead Origin_Landing Page Submission     9120 non-null   int64
 1   Lead Origin_Lead Add Form               9120 non-null   int64
 2   Lead Origin_Others                      9120 non-null   int64
 3   Lead Source_Google                      9120 non-null   int64
 4   Lead Source_Olark Chat                  9120 non-null   int64
 5   Lead Source_Organic Search              9120 non-null   int64
 6   Lead Source_Others                      9120 non-null   int64
 7   Lead Source_Reference                   9120 non-null   int64
 8   Lead Source_Referral Sites              9120 non-null   int64
 9   Lead Source_Social Media                9120 non-null   int64
 10  Lead Source_Welingak Website            9120 non-null   int64
 11  Do Not Email_Yes                        9120 non-null   int64
 12  Last Activity_Email Bounced             9120 non-null   int64
 13  Last Activity_Email Link Clicked        9120 non-null   int64
 14  Last Activity_Email Opened              9120 non-null   int64
 15  Last Activity_Form Submitted on Website 9120 non-null   int64
 16  Last Activity_Olark Chat Conversation   9120 non-null   int64
 17  Last Activity_Page Visited on Website   9120 non-null   int64
 18  Last Activity_SMS Sent                  9120 non-null   int64
 19  Last Activity_others                    9120 non-null   int64
...
 46  Last Notable Activity_SMS Sent          9120 non-null   int64
 47  Last Notable Activity_others            9120 non-null   int64
dtypes: int64(48)
memory usage: 3.4 MB
```

Output is truncated. View as a scrollable element or open in a text editor. Adjust cell output settings...

Data Preparation : Collinearity



Removing Multi-Collinearity through VIF method

40 Out of 48 features are left – 8 Removed.

04] ✓ 3m 30.5s

```
Removing 'What is your current occupation_Unemployed' with VIF: 121.5
Removing 'Lead Origin_Lead Add Form' with VIF: 50.2
Removing 'Last Notable Activity_Email Opened' with VIF: 23.6
Removing 'Specialization_Unknown' with VIF: 13.7
Removing 'Last Activity_SMS Sent' with VIF: 10.9
Removing 'Lead Origin_Others' with VIF: 10.1
Removing 'Lead Origin_Landing Page Submission' with VIF: 8.2
Removing 'Page Views Per Visit' with VIF: 6.6
Removing 'Last Activity_Email Opened' with VIF: 5.9
Removing 'Tags_Unknown' with VIF: 5.0

Final VIF DataFrame:
```

	Feature	VIF
0	TotalVisits	3.776272
1	Total Time Spent on Website	2.498008
2	Lead Source_Google	2.840219
3	Lead Source_Olark Chat	2.769970
4	Lead Source_Organic Search	1.703093
5	Lead Source_Others	1.053308
6	Lead Source_Reference	1.957285
7	Lead Source_Referral Sites	1.124110
8	Lead Source_Social Media	1.050599
9	Lead Source_Welingak Website	1.151211
10	Do Not Email_Yes	1.999297
11	Last Activity_Email Bounced	1.982836
...		
37	Last Notable Activity_Olark Chat Conversation	1.404365
38	Last Notable Activity_Page Visited on Website	2.143333
39	Last Notable Activity_SMS Sent	1.902990
40	Last Notable Activity_others	1.585538

Output is truncated. View as a scrollable element or open in a text editor. Adjust cell output settings...

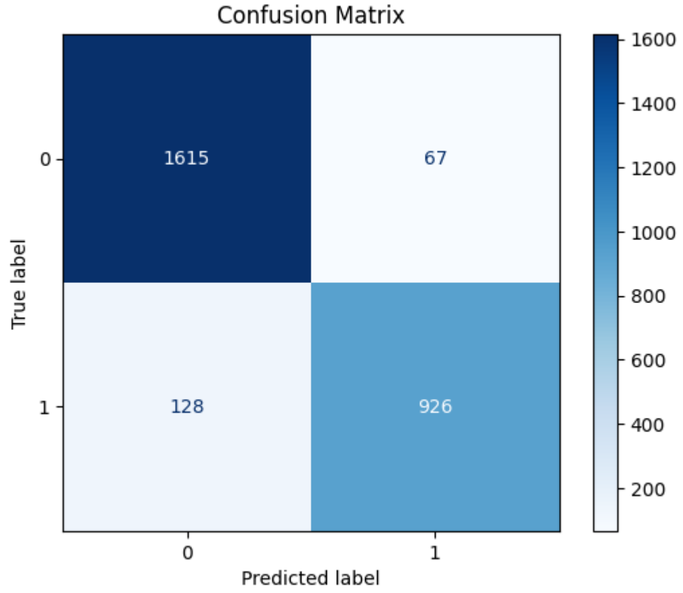
Data Modelling - OUTCOME

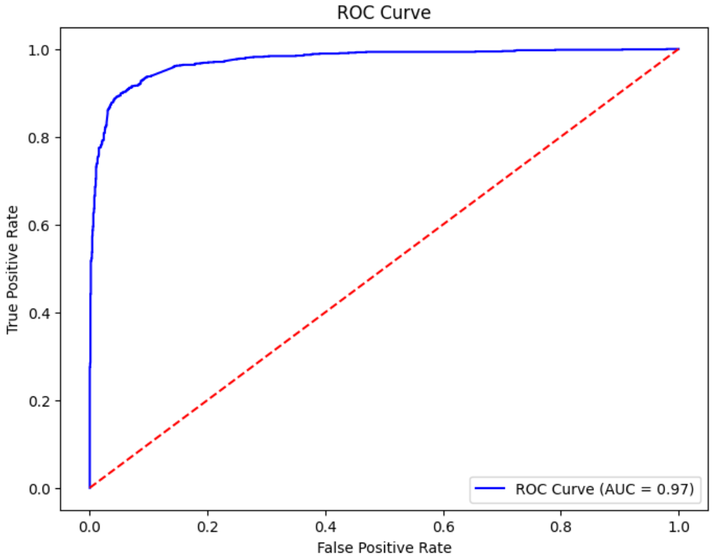
Following are the results found :

- Accuracy: 0.929 (92.9%)
- Precision: 0.933 (93.3%)
- Recall: 0.879 (87.9%)
- F1 Score: 0.905 (90.5%)
- AUC-ROC: 0.974 (97.4%)

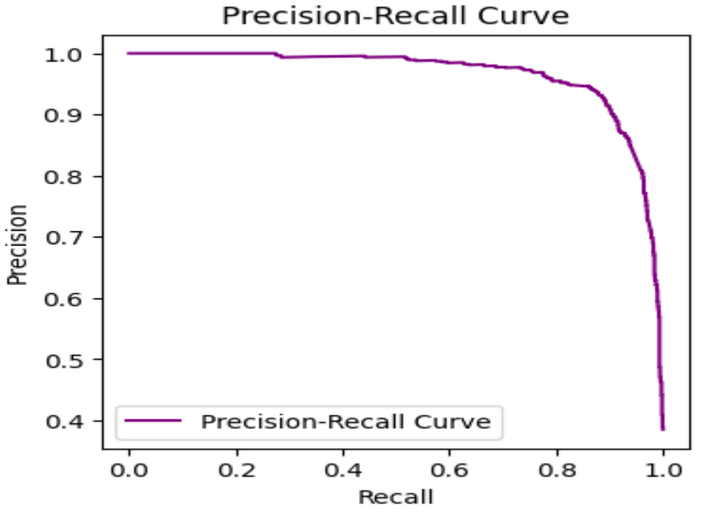
- Confusion Matrix:
- [[1615 67]
- [128 926]]

Indicates that this model is performing excellently to predict the right leads.

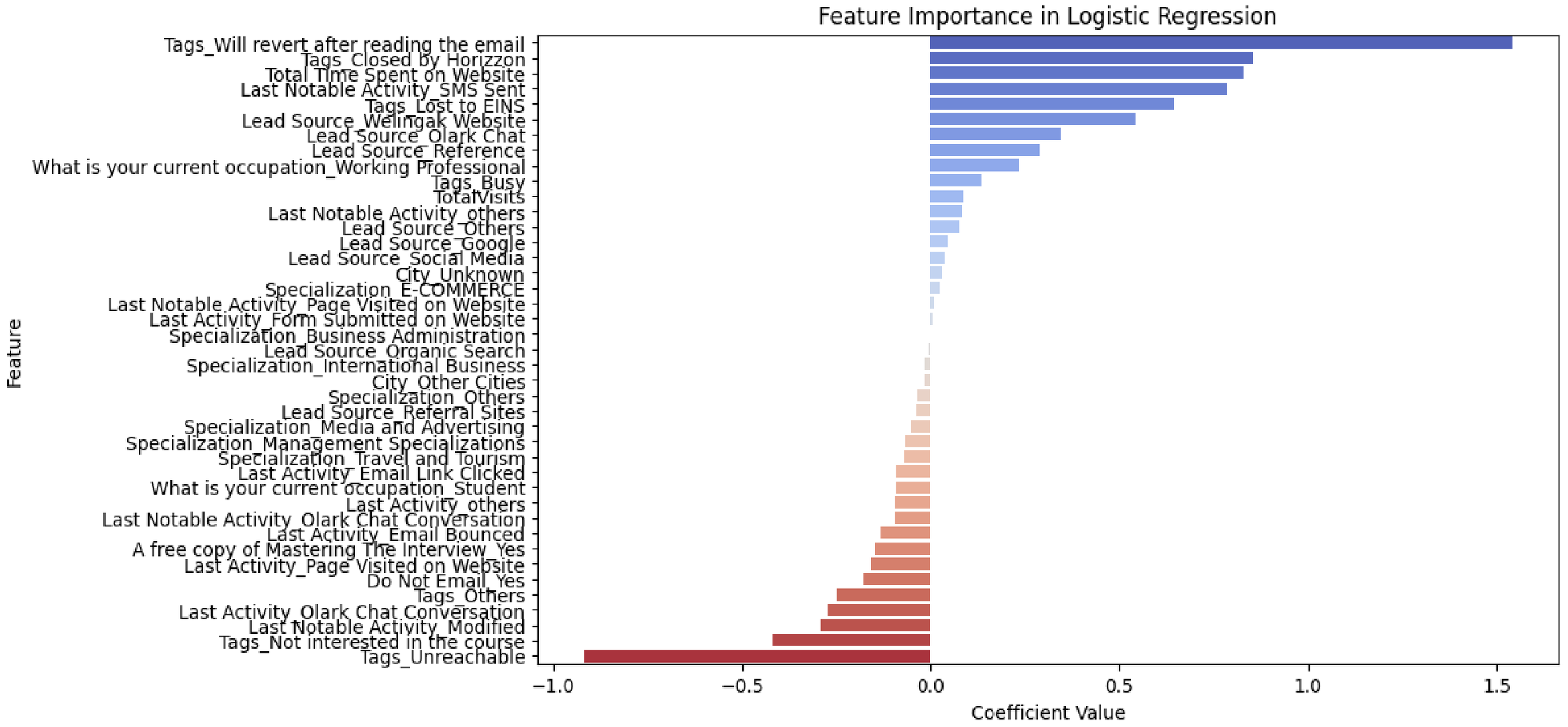




Classification Report:	precision	recall	f1-score	support
0	0.93	0.96	0.94	1682
1	0.93	0.88	0.90	1054
accuracy			0.93	2736
macro avg	0.93	0.92	0.92	2736
weighted avg	0.93	0.93	0.93	2736



Data Modelling - Outcome



RESULT - LEAD CONVERSION

```
df_lead_scores['lead_status'].value_counts()
```

[26] ✓ 0.0s

```
lead_status
Good drop         1622
Good selection     920
Missed Lead        134
Bad selection       60
Name: count, dtype: int64
```

```
# Client wants to achieve the lead conversion above 80%. lets see if we get there.
# Dividing Good selection by all selected leads

total_leads_selected = df_lead_scores[df_lead_scores['lead_status']=='Good selection'].shape[0]+df_lead_scores[df_lead_scores['lead_status']=='Missed Lead'].shape[0]
total_successful_leads = df_lead_scores[df_lead_scores['lead_status']=='Good selection'].shape[0]

success_rate = round(total_successful_leads*100/total_leads_selected,1)

print(f'success rate of hot leads as per this model :  {success_rate} %')

```

[27] ✓ 0.0s

```
success rate of hot leads as per this model :  87.3 %
```

Following this Model, the Company X Lead Success can straight away increase from 30% to Above 80% !!

Recommended Strategy for Client

Peak time

Reduce the Threshold to increase efforts

Comfort zone

Preserve resources by keeping high cut-off

Focus point

Focus on key features that are suggested by Model.

Thank You

"Without data, you're just another person with an opinion."

– W. Edwards Deming