Lead Scoring Model Development

Date: December 16, 2024

Present By: Shivam Thacker
Shobha K A
Shubham Sarvesh

Overview

- Objective
- Data Overview
- Key Insights from EDA
- Model Development & Results
- Recommendations
- Conclusion

Why Build a Lead Scoring Model?:

- Assign a lead score between 0-100 to rank potential leads.
- Prioritize leads for better targeting and higher conversions.

Data Overview & Cleaning

- Dataset Details: 9,240 rows, 37 columns

Check data size

```
[9]: user.shape

[9]: (9240, 37)
```

- Cleaning Process: Dropped columns with >40% missing values

```
[21]: #Dropping columns having more than 40% null values

user_data = user_data.drop(columns=user_data.columns[(round(user_data.isnull().sum() / user_data.shape[0], 2)) > 0.40])
```

Data Overview & Cleaning

- Cleaning Process: Imputed missing data logically

Since the majority of users are from Mumbai, we can impute the missing values in the "City" field with "Mumbai."

```
user_data['City']=user_data['City'].replace(np.nan,'Mumbai')
```

The "Specialization" column has 37% missing values. This could be because the lead is a student, lacks a specialization, or their specialization is not listed among the available options. To address this, we can create a new category labeled "Others."

```
user_data['Specialization'] = user_data['Specialization'].replace(np.nan, 'Others')
```

Majority of values in this column are "Will revert after reading the email," we can use this value to impute the missing entries.

```
user_data['Tags'] = user_data['Tags'].replace(np.nan,'Will revert after reading the email')
```

Data Overview & Cleaning

- Cleaning Process: Retained 98% of rows

We can calculate the percentage of rows retained.

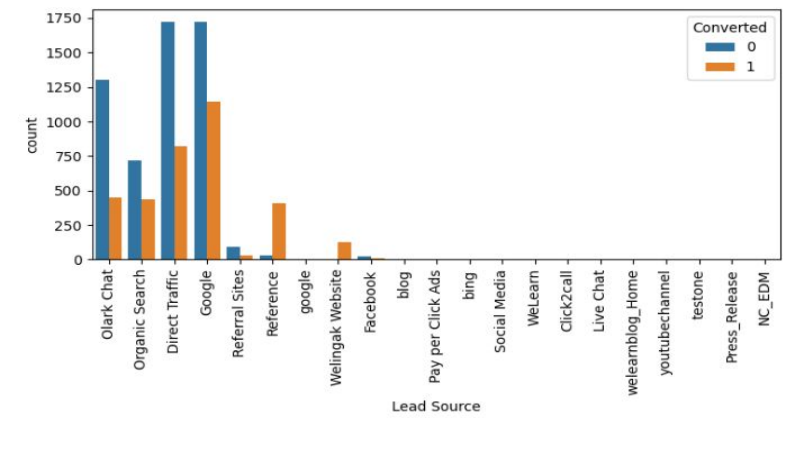
```
(len(user_data.index)/9240)*100
```

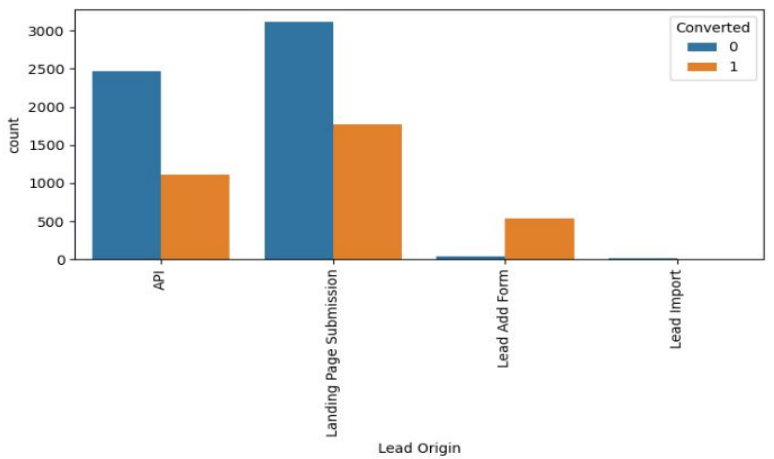
```
98.2034632034632
```

After cleaning the data, we have retained 98% of the rows.

Insights from Data Analysis

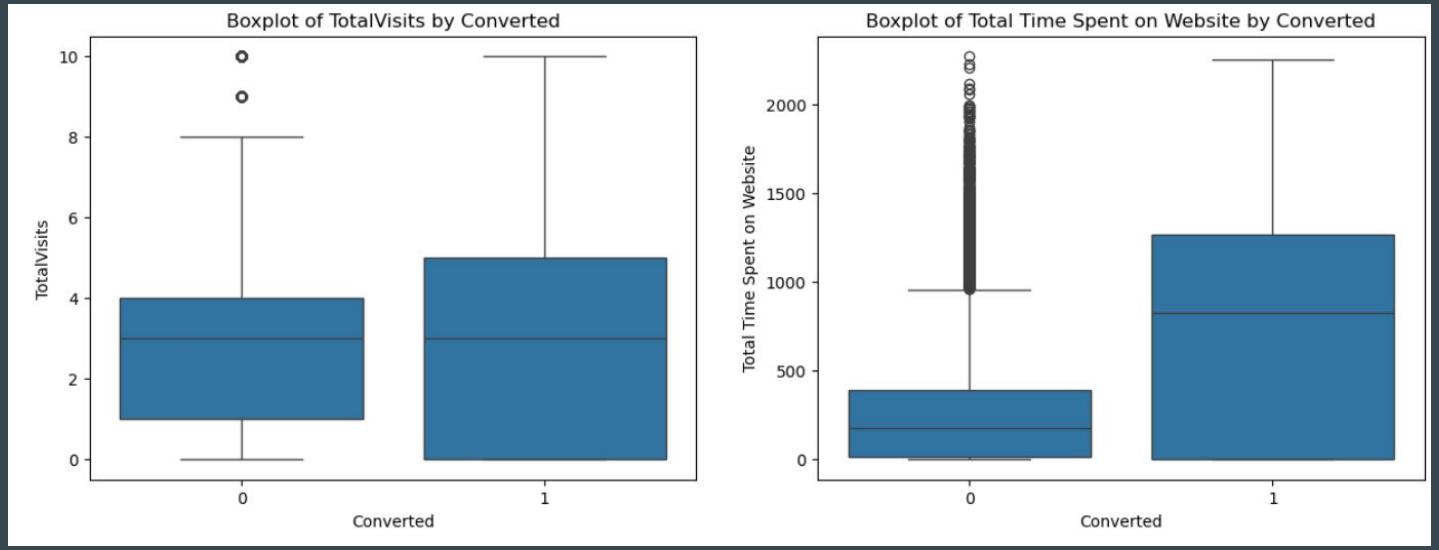
- Lead Sources & Origins:
  - API & Landing Page Submission need improvement (high volume, low conversion).
  - Google, Direct Traffic, and Olark Chat require focused campaigns.
  - Reference & Welingkar Website: Increase leads.





Insights from Data Analysis

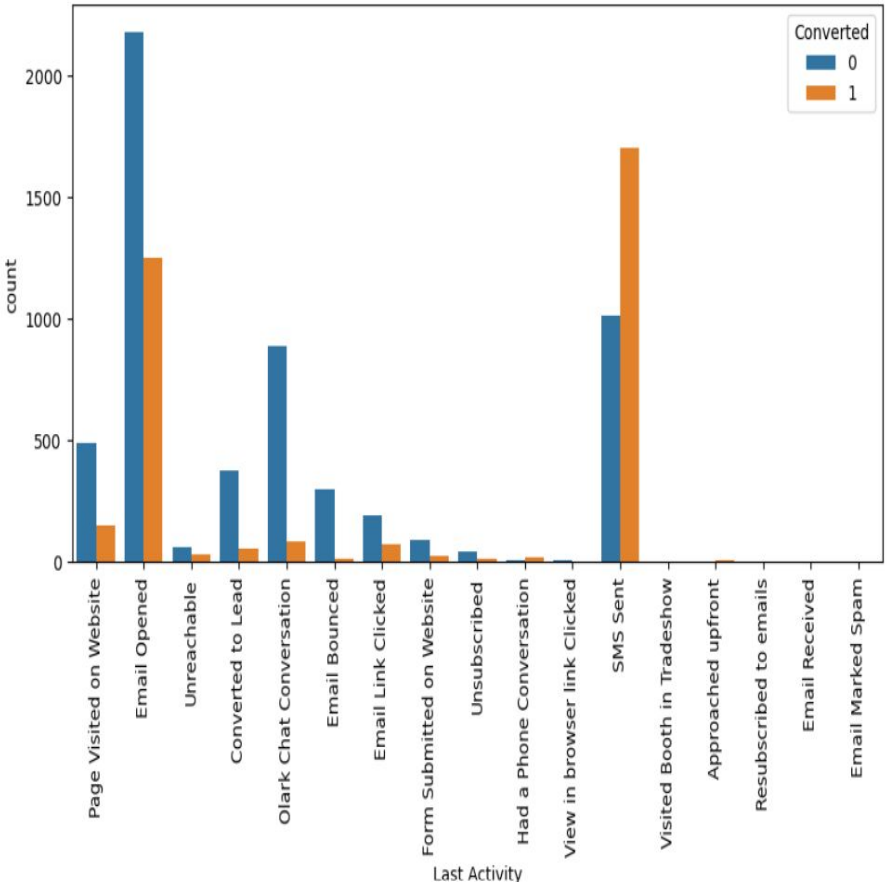
- User Engagement:
  - Leads spending more time on the website convert better.
  - Website engagement improvements are recommended.



Insights from Data Analysis

Last Activity:

- Follow-up on leads with "Email Opened."
- Increase leads with "SMS Sent" for higher conversion.



Model Development & Key Metrics

Methodology

- Feature Selection (RFE)

Feature Selection Using RFE

```
#Starting with 15 features selected by RFE
#We will then optimize the model further by inspecting VIF and p-value of the features

logreg = LogisticRegression()
rfe = RFE(estimator=logreg, n_features_to_select=15)
rfe = rfe.fit(X_train, y_train)

list(zip(X_train.columns, rfe.support_, rfe.ranking_))
```

Model Development & Key Metrics

Methodology

- Feature Selection (P-Values)

	coef	std err	z	P>|z|	[0.025	0.975]
const	-2.4546	0.193	-12.745	0.000	-2.832	-2.077
Do Not Email	-1.5558	0.203	-7.672	0.000	-1.953	-1.158
Lead Origin_Landing Page Submission	-1.6117	0.151	-10.683	0.000	-1.907	-1.316
Lead Origin_Lead Add Form	1.5833	0.315	5.034	0.000	0.967	2.200
Lead Source_Welingak Website	2.1487	0.795	2.704	0.007	0.591	3.706
Last Activity_Other_Last_Activity	2.4096	0.602	4.001	0.000	1.229	3.590
Last Activity_SMS Sent	1.9109	0.089	21.376	0.000	1.736	2.086
Last Activity_Unsubscribed	2.0761	0.520	3.989	0.000	1.056	3.096
Specialization_Others	-2.1929	0.154	-14.257	0.000	-2.494	-1.891
Last Notable Activity_Modified	-1.6376	0.093	-17.560	0.000	-1.820	-1.455
Last Notable Activity_Olark Chat Conversation	-1.5503	0.326	-4.755	0.000	-2.189	-0.911
Tags_Busy	3.3535	0.255	13.176	0.000	2.855	3.852
Tags_Closed by Horizzon	9.3350	0.735	12.696	0.000	7.894	10.776
Tags_Lost to EINS	9.2079	0.742	12.406	0.000	7.753	10.663
Tags_Will revert after reading the email	4.1573	0.151	27.549	0.000	3.862	4.453

Model Development & Key Metrics

Methodology

- Feature Selection (VIF)

	Features	Variance Inflation Factor
13	Tags_Will revert after reading the email	2.40
1	Lead Origin_Landing Page Submission	2.19
7	Specialization_Others	2.01
8	Last Notable Activity_Modified	1.67
2	Lead Origin_Lead Add Form	1.64
5	Last Activity_SMS Sent	1.58
3	Lead Source_Welingak Website	1.34
11	Tags_Closed by Horizzon	1.21
0	Do Not Email	1.20
6	Last Activity_Unsubscribed	1.08
9	Last Notable Activity_Olark Chat Conversation	1.07
10	Tags_Busy	1.07
12	Tags_Lost to EINS	1.06
4	Last Activity_Other_Last_Activity	1.01

Model Development & Key Metrics

Model Accuracy	• Training - 85%, Test - 83%

```
print("Accuracy score", metrics.accuracy_score(y_train_pred_final.Convert, y_train_pred_final.predicted))
```

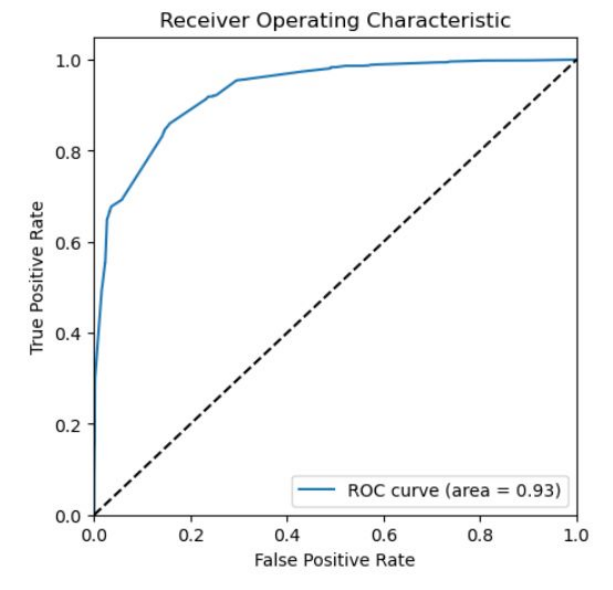
```
Accuracy score 0.8501023460872303
```

```
---Model Evaluation Metrics---
Confusion Matrix :
 [[1453  281]
 [ 162  827]]
Accuracy : 0.8373117884686008
Sensitivity : 0.8361981799797775
Specificity : 0.8379469434832757
Precision : 0.7463898916967509
```

Model Development & Key Metrics

ROC-AUC

- 0.93



Model Development & Key Metrics

Key Features

- "Closed by Horizzon" (Coefficient: 9.33)
- "Lost to EINS" (Coefficient: 9.21)
- "Will revert after reading the email" (Coefficient: 4.16)

```
==========================================================
Tags_Closed by Horizzon                                9.334976
Tags_Lost to EINS                                      9.207855
Tags_Will revert after reading the email               4.157329
Tags_Busy                                              3.353539
Last Activity_Other_Last_Activity                      2.409595
Lead Source_Welingak Website                           2.148682
Last Activity_Unsubscribed                             2.076098
Last Activity_SMS Sent                                 1.910934
Lead Origin_Lead Add Form                              1.583328
Last Notable Activity_Olark Chat Conversation         -1.550329
Do Not Email                                          -1.555791
Lead Origin_Landing Page Submission                   -1.611658
Last Notable Activity_Modified                        -1.637641
Specialization_Others                                 -2.192860
const                                                 -2.454629
dtype: float64
```

Recommendations for Improvement

Lead Conversion Strategy

- Improve conversion from API & Landing Page Submission.
- Increase leads from Reference & Welingkar Website.
- Use LinkedIn for working professionals.

Follow-Up Process

- Call leads with "Email Opened" as last activity.
- Send more SMS to potential leads.

Website Engagement

- Make the website more engaging to increase time spent.

Conclusion

- The model predicts lead scores effectively with high accuracy.
- Following recommendations can improve conversion rates significantly.

print(“Thank You!”)