Logistic Regression- Lead Scoring

Bhagabati Prasad Mishra
Swamy Bore
Krishna Prasanth

Content

- **Introduction**
- **Current Lead Conversion Rate**
- **Objective**
- **Data Overview**
- **Key Features Impacting Conversion**
- **Machine Learning Approach**
- **Model Performance**
- **Recommendations and Implementation**

Optimizing Lead Conversion at X Education

Introduction-

- X Education, a leading online course provider for industry professionals, faces challenges with its current lead conversion process.
- Despite acquiring a significant number of leads daily, the lead conversion rate stands at 30%, highlighting room for improvement.

Current Scenario-

- On any given day, professionals land on the website, browse courses, and may fill out forms or watch videos.
- The company acquires leads through various channels, including website interactions, form submissions, and past referrals.
- The sales team engages in extensive outreach, but only about 30% of acquired leads convert.

Objective:

- The primary goal is to identify 'Hot Leads,' focusing the sales team's efforts on potential customers with a higher likelihood of conversion.
- The aim is to increase lead conversion efficiency, making the outreach more targeted and effective.

Data Overview

There are 9240 rows or records provided with 37 features
Feature “Converted” is the target feature
Multiple features are identified with null values or missing values

Handling Null Values:

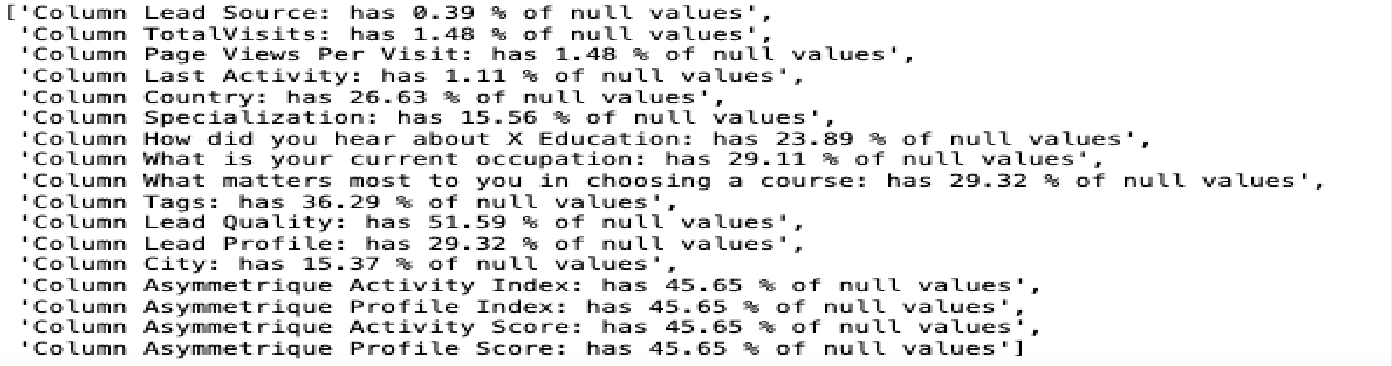
- Removing columns with significant null values (>45%): 'Lead Quality', 'Lead Profile', 'Asymmetrique Activity Index', 'Asymmetrique Profile Index', 'Asymmetrique Activity Score', 'Asymmetrique Profile Score'.
- Avoiding imputation to prevent biases in data.

Removing Imbalanced or Irrelevant Columns:

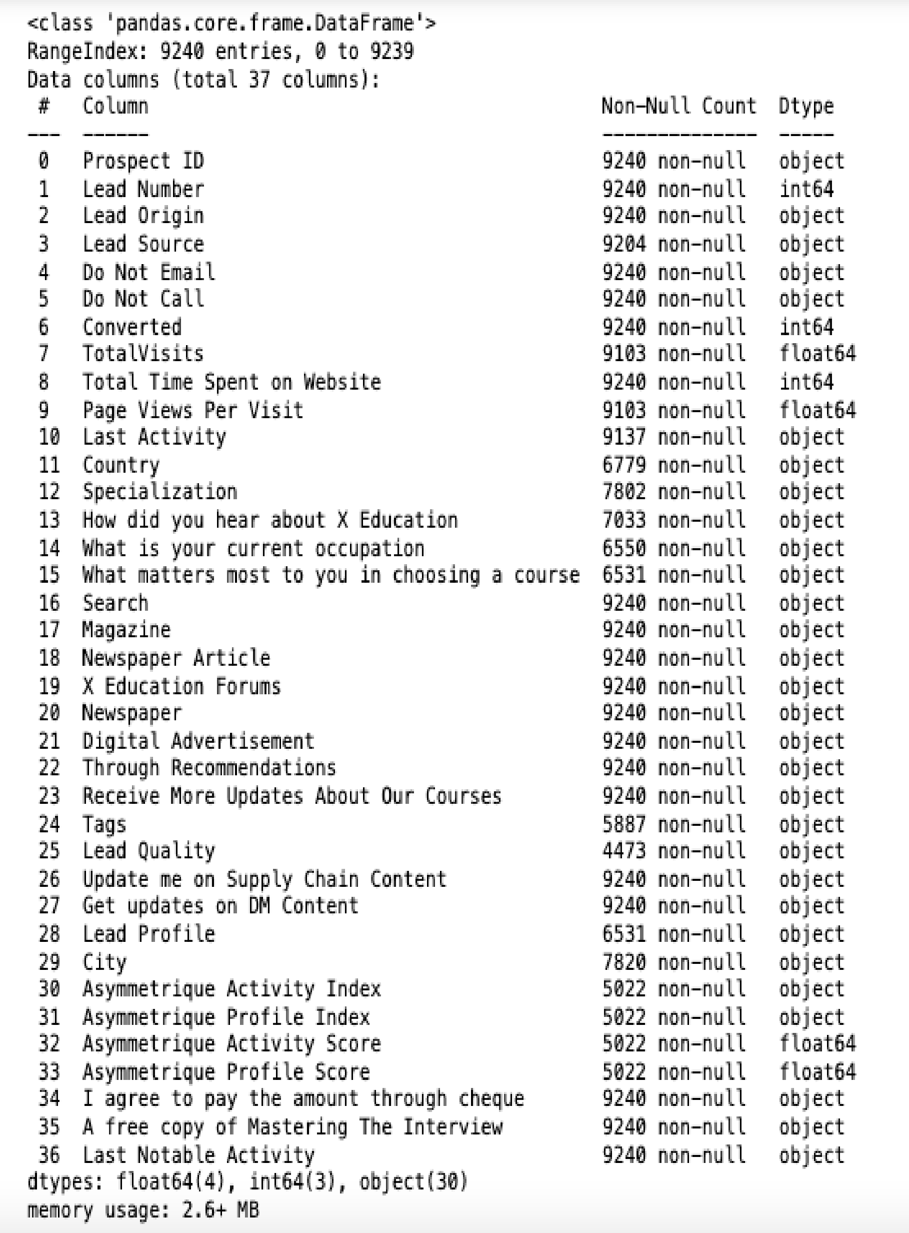
- Dropping columns with high imbalance: 'Country', 'What matters most to you in choosing a course', 'What is your current occupation'.
- Removing 'Tags' due to its impact on subsequent lead action processes.

Additional Column Removals:

- Dropping less relevant columns for model building: 'Search', 'Magazine', 'Newspaper Article', 'X Education Forums', 'Newspaper', 'Digital Advertisement', 'Through Recommendations', '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', 'A free copy of Mastering The Interview', 'Lead Number', 'Last Activity', 'Prospect ID', 'Last Notable Activity', 'Lead Profile', 'City', 'How did you hear about X Education'.



```
['Column Lead Source: has 0.39 % of null values',
 'Column TotalVisits: has 1.48 % of null values',
 'Column Page Views Per Visit: has 1.48 % of null values',
 'Column Last Activity: has 1.11 % of null values',
 'Column Country: has 26.63 % of null values',
 'Column Specialization: has 15.56 % of null values',
 'Column How did you hear about X Education: has 23.89 % of null values',
 'Column What is your current occupation: has 29.11 % of null values',
 'Column What matters most to you in choosing a course: has 29.32 % of null values',
 'Column Tags: has 36.29 % of null values',
 'Column Lead Quality: has 51.59 % of null values',
 'Column Lead Profile: has 29.32 % of null values',
 'Column City: has 15.37 % of null values',
 'Column Asymmetrique Activity Index: has 45.65 % of null values',
 'Column Asymmetrique Profile Index: has 45.65 % of null values',
 'Column Asymmetrique Activity Score: has 45.65 % of null values',
 'Column Asymmetrique Profile Score: has 45.65 % of null values']
```

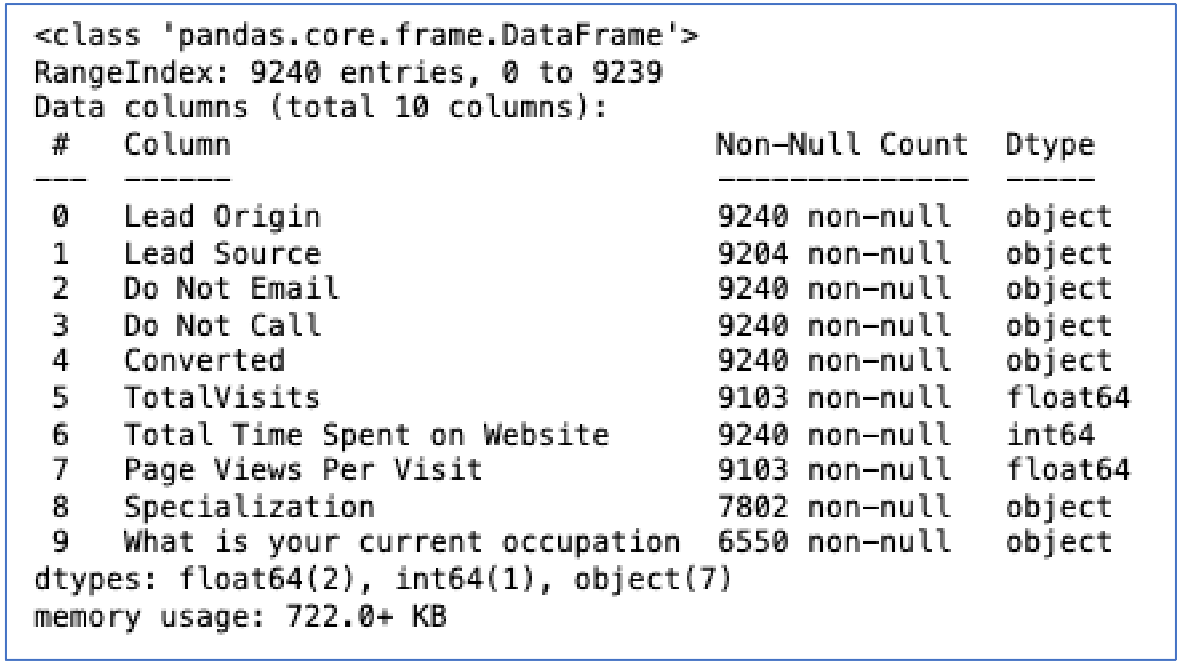


```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 9240 entries, 0 to 9239
Data columns (total 37 columns):
 #   Column                                         Non-Null Count  Dtype
---  ------                                         --------------  -----
 0   Prospect ID                                    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
 20  Newspaper                                      9240 non-null   object
 21  Digital Advertisement                          9240 non-null   object
 22  Through Recommendations                        9240 non-null   object
 23  Receive More Updates About Our Courses         9240 non-null   object
 24  Tags                                           5887 non-null   object
 25  Lead Quality                                   4473 non-null   object
 26  Update me on Supply Chain Content              9240 non-null   object
 27  Get updates on DM Content                      9240 non-null   object
 28  Lead Profile                                   6531 non-null   object
 29  City                                           7820 non-null   object
 30  Asymmetrique Activity Index                    5022 non-null   object
 31  Asymmetrique Profile Index                     5022 non-null   object
 32  Asymmetrique Activity Score                    5022 non-null   float64
 33  Asymmetrique Profile Score                     5022 non-null   float64
 34  I agree to pay the amount through cheque       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
```

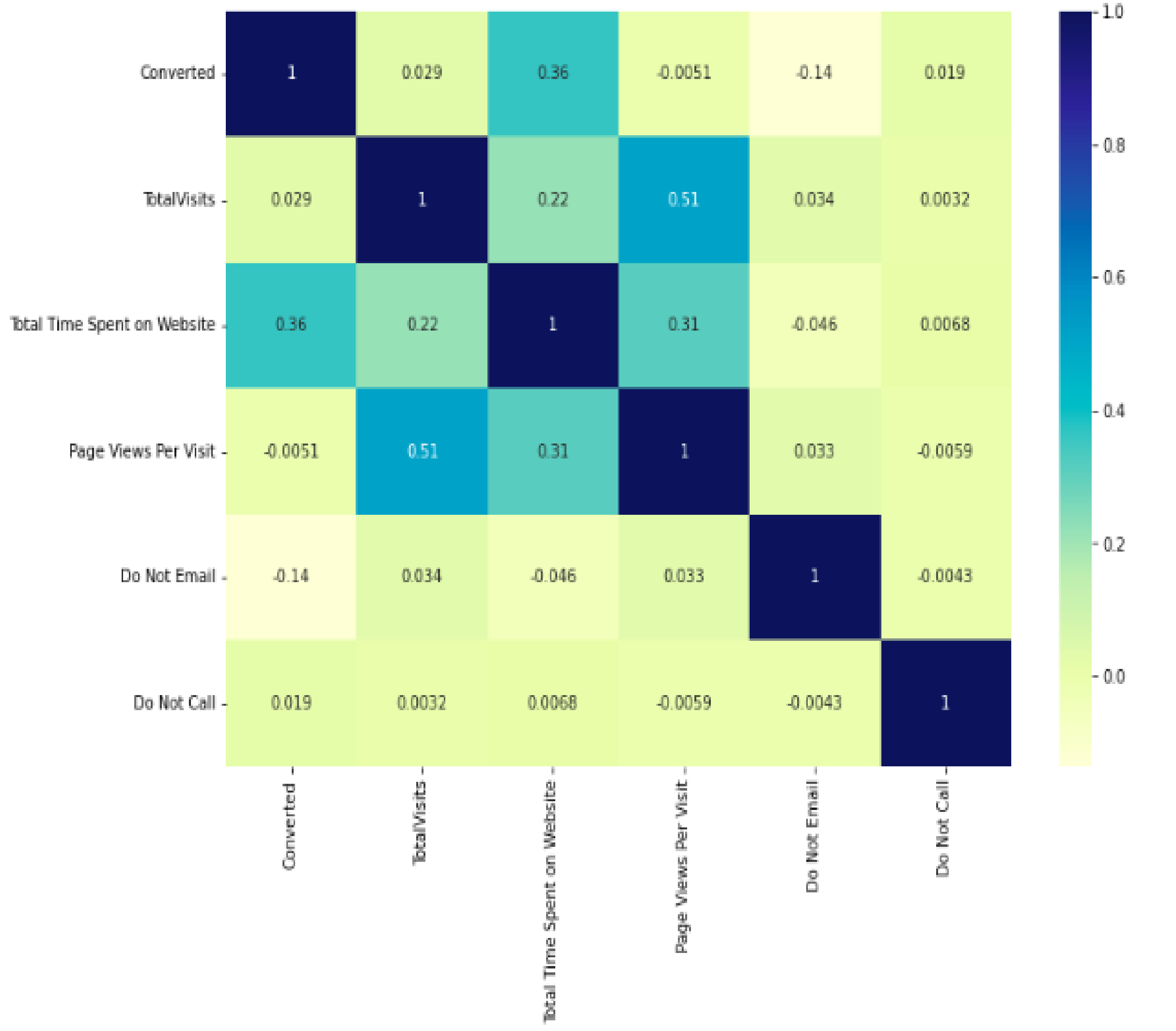
Key Features Impacting Conversion

Post thorough analysis of the features, following features are retained and rest of the features were not identified as critical for model performance stand point and hence, were removed from the data frame.

Post identifying the important feature, data imputation and replacing attributes where ever necessary are conducted.

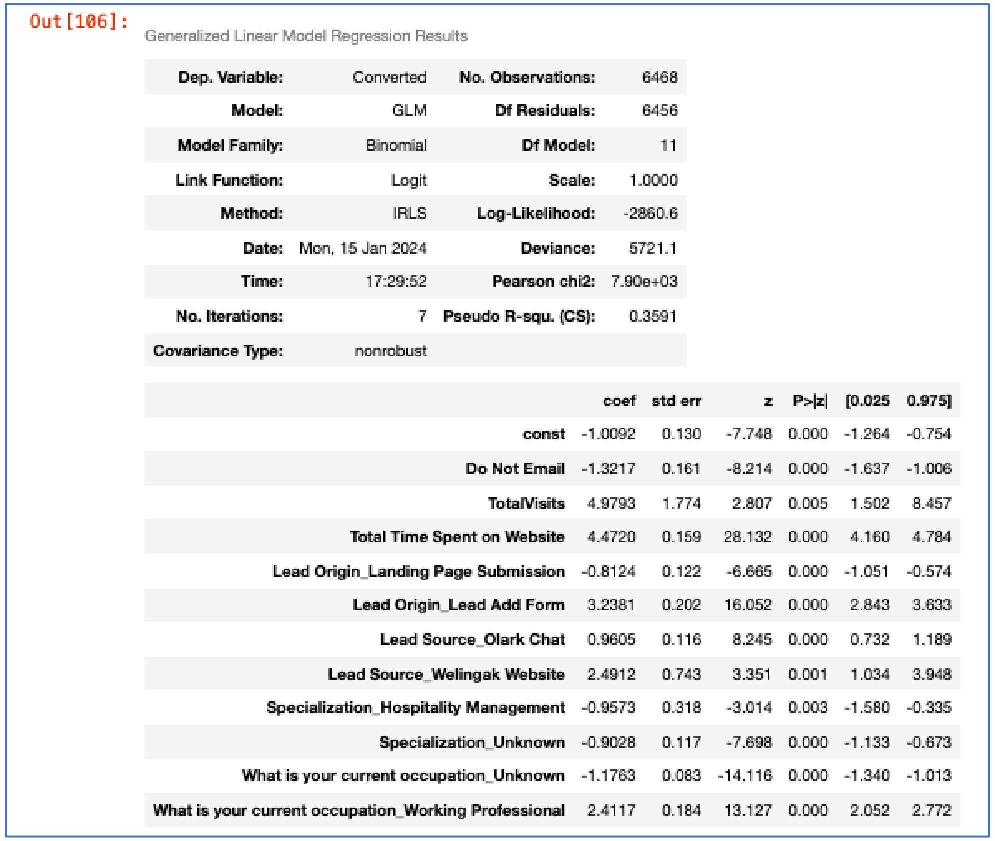


```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 9240 entries, 0 to 9239
Data columns (total 10 columns):
 #   Column                            Non-Null Count  Dtype
---  ------                            --------------  -----
 0   Lead Origin                       9240 non-null   object
 1   Lead Source                       9204 non-null   object
 2   Do Not Email                      9240 non-null   object
 3   Do Not Call                       9240 non-null   object
 4   Converted                         9240 non-null   object
 5   TotalVisits                       9103 non-null   float64
 6   Total Time Spent on Website       9240 non-null   int64
 7   Page Views Per Visit              9103 non-null   float64
 8   Specialization                    7802 non-null   object
 9   What is your current occupation   6550 non-null   object
dtypes: float64(2), int64(1), object(7)
memory usage: 722.0+ KB
```



Machine Learning Approach

- Data Preparation:
  - Null Values: Columns with >45% null values removed to prevent biases.
  - Imbalanced Data: Columns with imbalances and irrelevant data removed for model clarity.
- Feature Selection:
  - Relevance: Removed less relevant columns for streamlined model building.
  - Tags: Omitted due to potential impact on real-world data for subsequent lead actions.
- Target Variable Definition:
  - Define 'Hot Leads' as the target variable for the model.
- Feature Engineering:
  - Logistic Regression: Chosen for its simplicity and effectiveness in binary classification tasks.
  - Recursive Feature Elimination (RFE): Used to select the most relevant features for the model.
- Model Training:
  - Train-Test Split: Split the data into training and testing sets for model evaluation.
  - Logistic Regression Training: Utilized logistic regression to predict 'Hot Leads.'
- Evaluation Metrics:
  - Precision, Recall, F1 Score: Key metrics to assess model performance.
  - ROC Curve: Visualized to analyze the trade-off between true positive rate and false positive rate.
- Optimization:
  - Fine-Tuning: Adjusted model parameters for optimal performance.
  - Iterative Process: Refinement based on evaluation metrics.

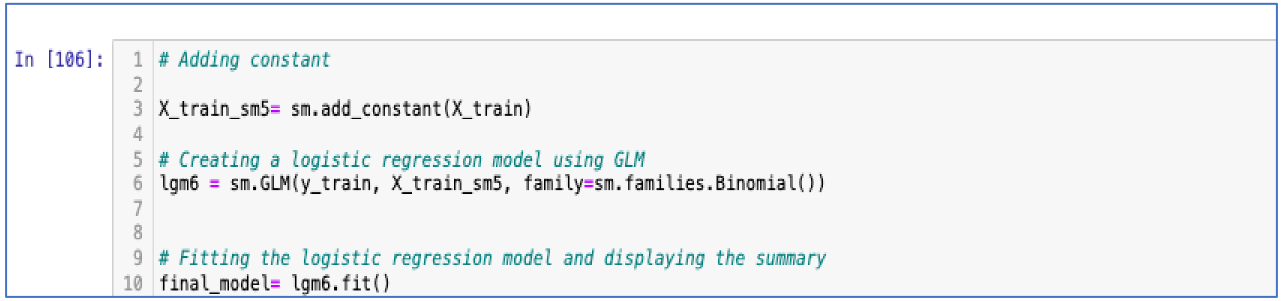


Out[106]:

Generalized Linear Model Regression Results

Dep. Variable:	Converted	**No. Observations:**	6468
Model:	GLM	**Df Residuals:**	6456
Model Family:	Binomial	**Df Model:**	11
Link Function:	Logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-2860.6
Date:	Mon, 15 Jan 2024	**Deviance:**	5721.1
Time:	17:29:52	**Pearson chi2:**	7.90e+03
No. Iterations:	7	**Pseudo R-squ. (CS):**	0.3591
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-1.0092	0.130	-7.748	0.000	-1.264	-0.754
Do Not Email	-1.3217	0.161	-8.214	0.000	-1.637	-1.006
TotalVisits	4.9793	1.774	2.807	0.005	1.502	8.457
Total Time Spent on Website	4.4720	0.159	28.132	0.000	4.160	4.784
Lead Origin_Landing Page Submission	-0.8124	0.122	-6.665	0.000	-1.051	-0.574
Lead Origin_Lead Add Form	3.2381	0.202	16.052	0.000	2.843	3.633
Lead Source_Olark Chat	0.9605	0.116	8.245	0.000	0.732	1.189
Lead Source_Welingak Website	2.4912	0.743	3.351	0.001	1.034	3.948
Specialization_Hospitality Management	-0.9573	0.318	-3.014	0.003	-1.580	-0.335
Specialization_Unknown	-0.9028	0.117	-7.698	0.000	-1.133	-0.673
What is your current occupation_Unknown	-1.1763	0.083	-14.116	0.000	-1.340	-1.013
What is your current occupation_Working Professional	2.4117	0.184	13.127	0.000	2.052	2.772



In [106]:

```python
# Adding constant

X_train_sm5= sm.add_constant(X_train)

# Creating a logistic regression model using GLM
lgm6 = sm.GLM(y_train, X_train_sm5, family=sm.families.Binomial())


# Fitting the logistic regression model and displaying the summary
final_model= lgm6.fit()
```

Model Performance

Model Performance Metrics:

- Accuracy: 0.71
- Precision: 0.94
- Recall: 0.27
- F1-Score: 0.42

Threshold Selection:

- Optimal Threshold: 0.35
- Intersection Point: Precision, Recall, and F1-Score intersect at this threshold.

Evaluation Analysis:

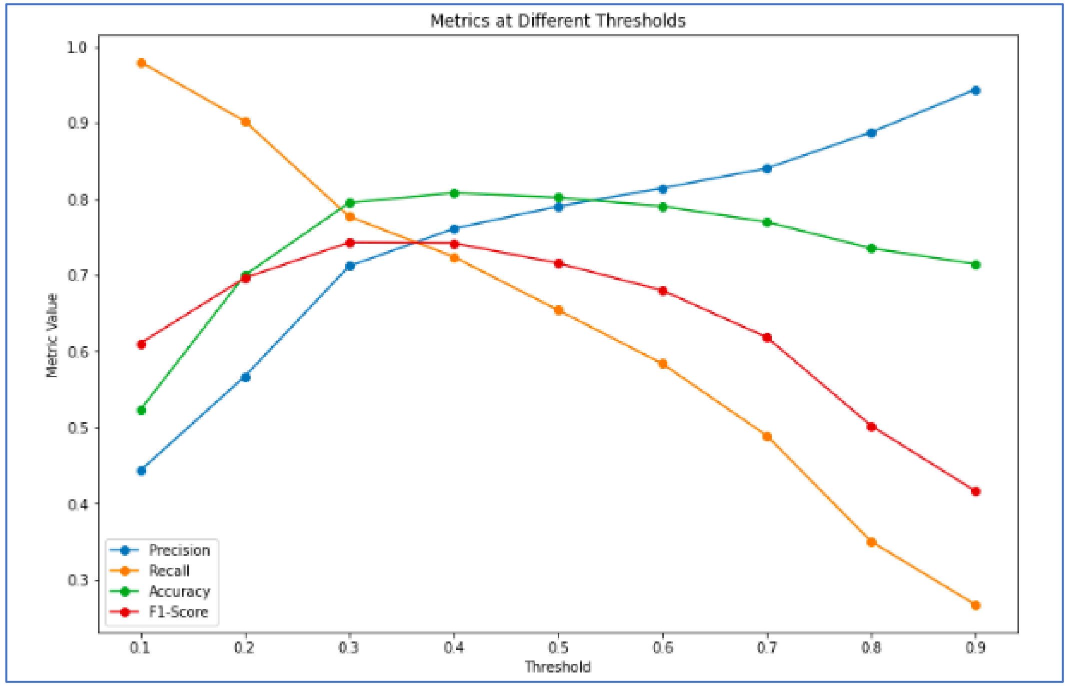
- Accuracy: Represents overall correctness, currently at 71%.
- Precision: High precision at 94% indicates low false positives.
- Recall: Recall at 27% signifies the model's ability to capture actual positive instances.
- F1-Score: Balanced measure considering precision and recall.

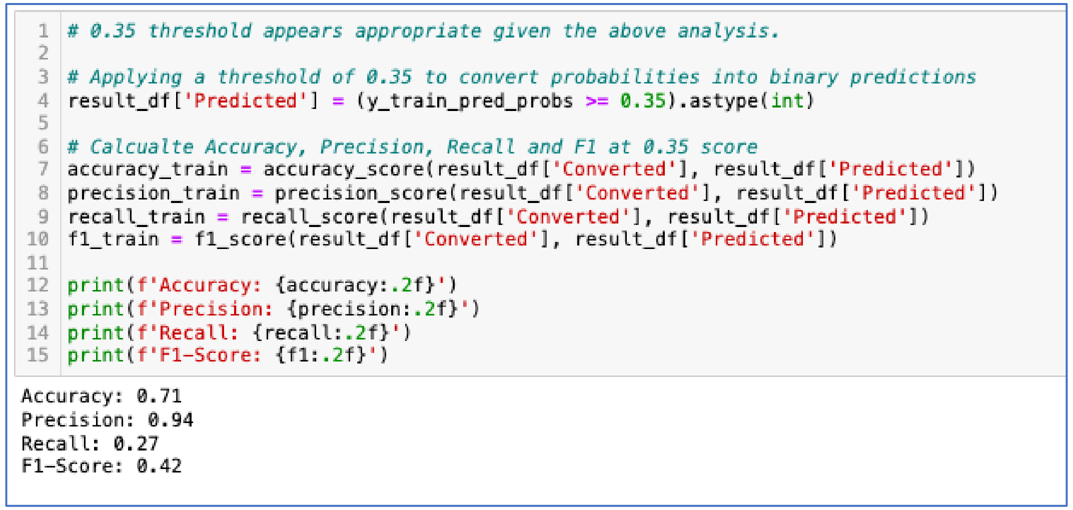
Threshold Rationale:

- Precision-Recall Trade-off: Threshold selected at the intersection point for a balanced approach.
- Optimizing for Precision: Lower threshold reduces false positives, crucial for targeted outreach.

Recommendations:

- Consideration for Business Goals: Precision-oriented threshold aligns with the objective of focusing on 'Hot Leads.'
- Trade-offs: Acknowledge the trade-off between precision and recall.





```
# 0.35 threshold appears appropriate given the above analysis.

# Applying a threshold of 0.35 to convert probabilities into binary predictions
result_df['Predicted'] = (y_train_pred_probs >= 0.35).astype(int)

# Calcualte Accuracy, Precision, Recall and F1 at 0.35 score
accuracy_train = accuracy_score(result_df['Converted'], result_df['Predicted'])
precision_train = precision_score(result_df['Converted'], result_df['Predicted'])
recall_train = recall_score(result_df['Converted'], result_df['Predicted'])
f1_train = f1_score(result_df['Converted'], result_df['Predicted'])

print(f'Accuracy: {accuracy:.2f}')
print(f'Precision: {precision:.2f}')
print(f'Recall: {recall:.2f}')
print(f'F1-Score: {f1:.2f}')
```

```
Accuracy: 0.71
Precision: 0.94
Recall: 0.27
F1-Score: 0.42
```

Recommendations and Implementation

Key Recommendations:

Focus on 'Hot Leads':

1. **Leverage the model's precision-oriented threshold to target potential leads more effectively.**
2. **Prioritize resources on leads with higher conversion probability.**

Continuous Monitoring:

1. **Regularly evaluate model performance and consider feedback from the sales team.**
2. **Adjust thresholds if needed to align with evolving business goals.**

Feedback Loop:

1. **Establish a feedback loop between the model predictions and the sales team.**
2. **Refine the model based on real-world outcomes.**

Implementation Plan:

Training:

1. **Conduct training sessions for the sales team on leveraging model predictions.**
2. **Emphasize the importance of targeting 'Hot Leads' identified by the model.**

Communication:

1. **Communicate the model's capabilities and limitations to the sales team.**
2. **Foster collaboration between data science and sales teams.**

Documentation:

1. **Document the model implementation process and guidelines.**
2. **Create a playbook for the sales team on utilizing model predictions.**