LEAD SCORE CASE STUDY

Powering X Education's Sales Transformation

Submitted by:
- Sachin Kanchan
- Smita Gaikwad
- Megha Jain

Business Challenge - Lead Conversion

X Education - Online Professional Training Platform

- Current Situation:
  - Generates leads through websites, search engines, and referrals
  - Existing lead conversion rate: Only 30%
  - Significant resource wastage on low-potential leads
  - Inefficient sales process

- Key Pain Points:
  - Time and effort spent on unproductive leads
  - Missed opportunities with high-potential prospects
  - Lack of systematic lead prioritization

Our Strategic Solution - Lead Scoring Model

- **Objective: Develop a Predictive Lead Score**
- Model Highlights:
  - Scoring Range: 0-100
  - Purpose: Identify "Hot Leads" with high conversion potential
  - Goal: Improve conversion rate from 30% to 80%
- Key Deliverables:
  1. Logistic Regression Predictive Model
  2. Data-Driven Insights - Questionnaire
  3. Performance Visualization - PPT
  4. Actionable Recommendations - Summary
- Expected Outcomes:
  - Optimize sales team's efforts
  - Increase conversion efficiency
  - Reduce wasted resources
  - Systematic lead qualification process

Methodology

- Importing Libraries & Setting up Analytics Environment
- Dataset Inspection
- Data Pre-Processing
- Exploratory Data Analysis
- Model Building – Logistic Regression
- Model Evaluation
- Predictions on Test Set
- Lead Score Generation
- Findings & Recommendations

Dataset Inspection

- We start with 37 columns and over 9240 rows.
- Most of these columns are string, with only a handful of numerical features

	Lead Number	Converted	TotalVisits	Total Time Spent on Website	Page Views Per Visit	Asymmetrique Activity Score	Asymmetrique Profile Score
count	9240.000	9240.000	9103.000	9240.000	9103.000	5022.000	5022.000
mean	617188.436	0.385	3.445	487.698	2.363	14.306	16.345
std	23405.996	0.487	4.855	548.021	2.161	1.387	1.811
min	579533.000	0.000	0.000	0.000	0.000	7.000	11.000
25%	596484.500	0.000	1.000	12.000	1.000	14.000	15.000
50%	615479.000	0.000	3.000	248.000	2.000	14.000	16.000
75%	637387.250	1.000	5.000	936.000	3.000	15.000	18.000
max	660737.000	1.000	251.000	2272.000	55.000	18.000	20.000

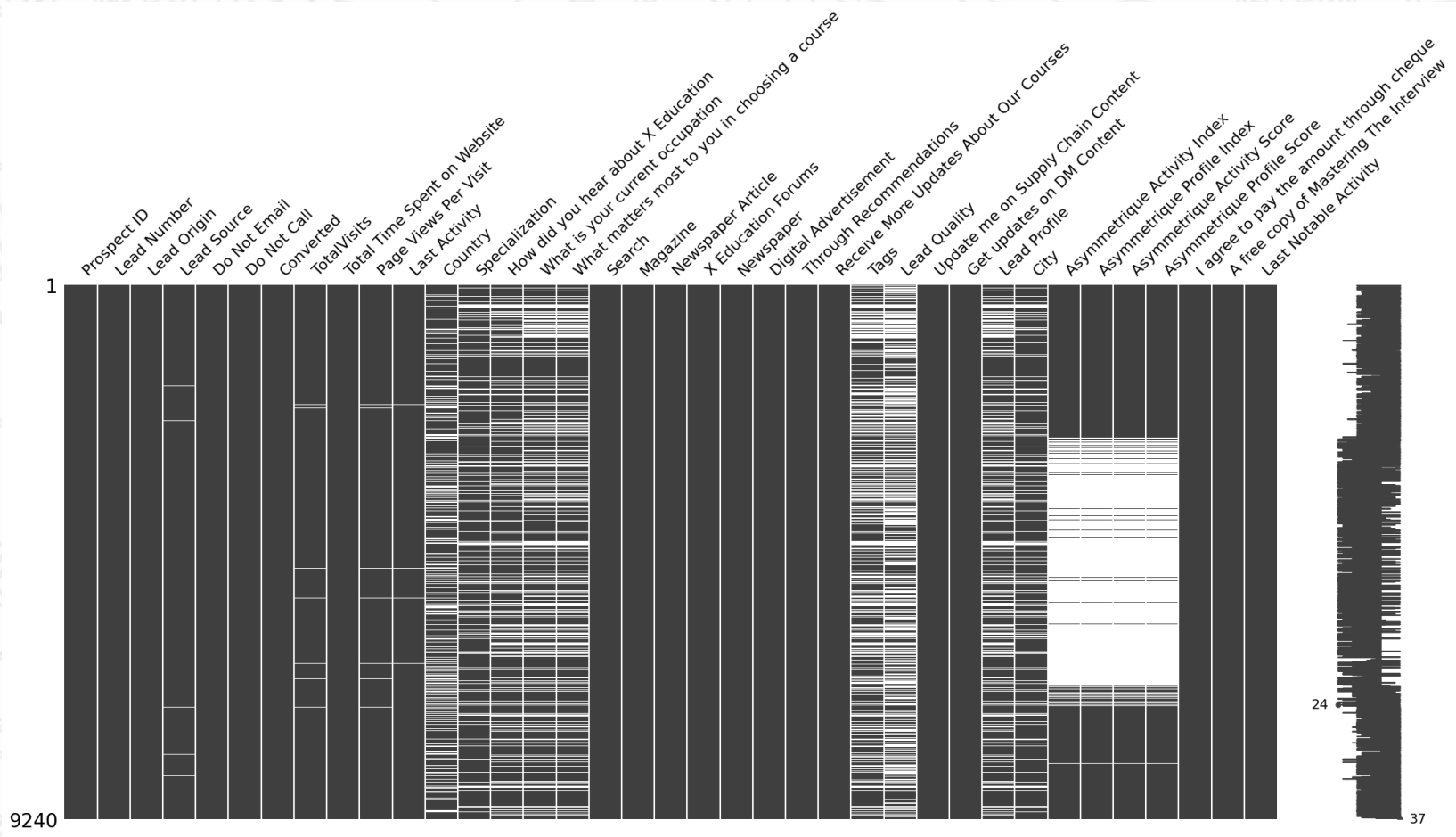
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 9240 entries, 0 to 9239
Data columns (total 37 columns):
 #   Column                                         Non-Null Count  Dtype
---  ------                                         --------------  -----
 0   Prospect ID                                    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
```

Data Pre-Processing

- `Select` seems to be erroneously captured in the data collection process despite not being a valid data point.
- We replaced this with 'Unknown'

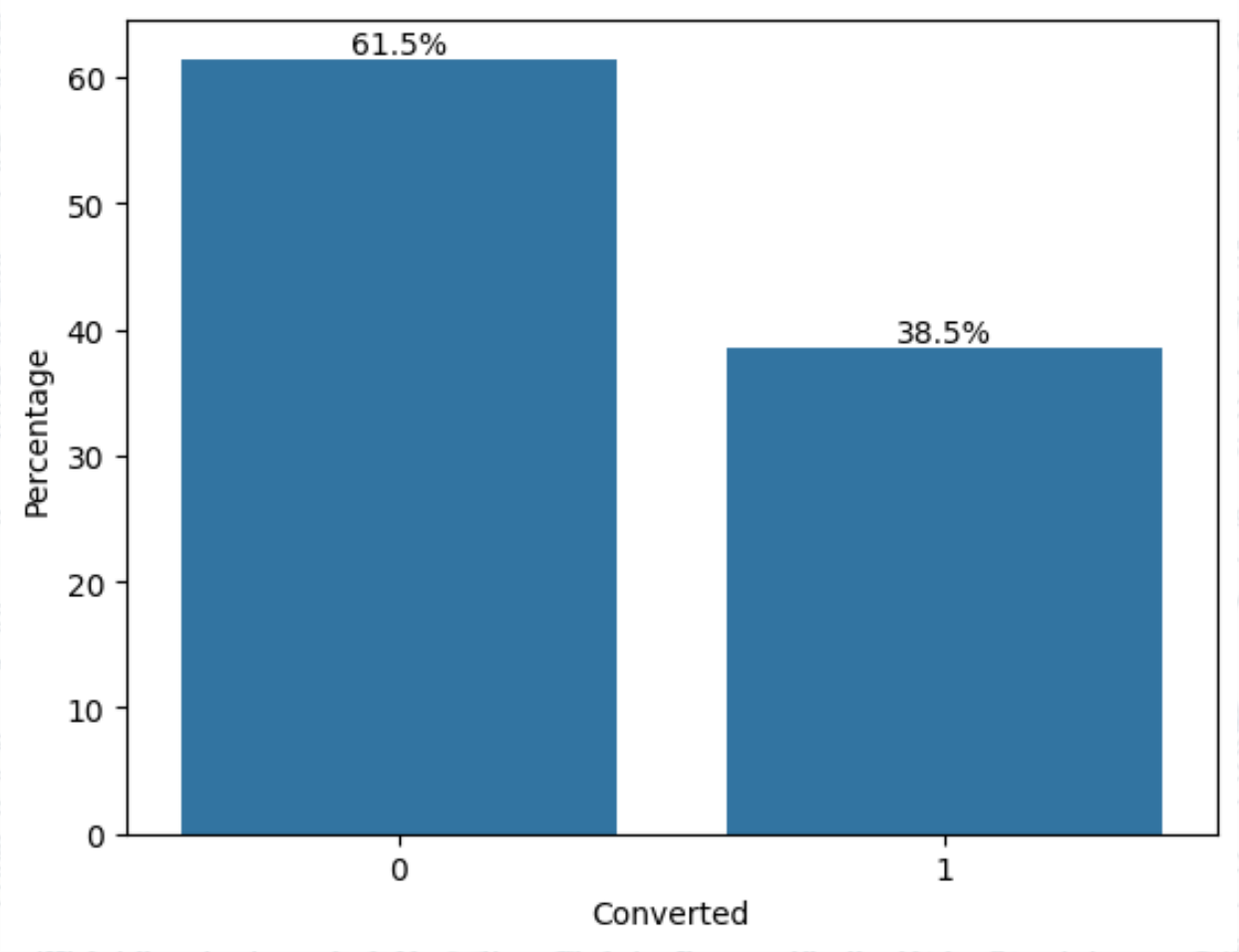
NULLS

- We dropped features with Null % over 30%
- In low null columns
  - for Numerical Features – Imputed nulls with median
  - For Categorical Features – imputed nulls with mode
- Capped Outliers in Numerical features
- Reduced sub-categories in 'Lead Source'



Exploratory Data Analysis

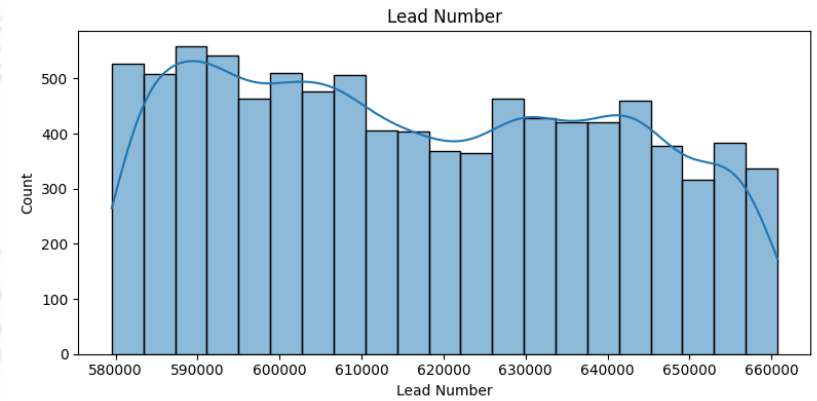
Target Imbalance

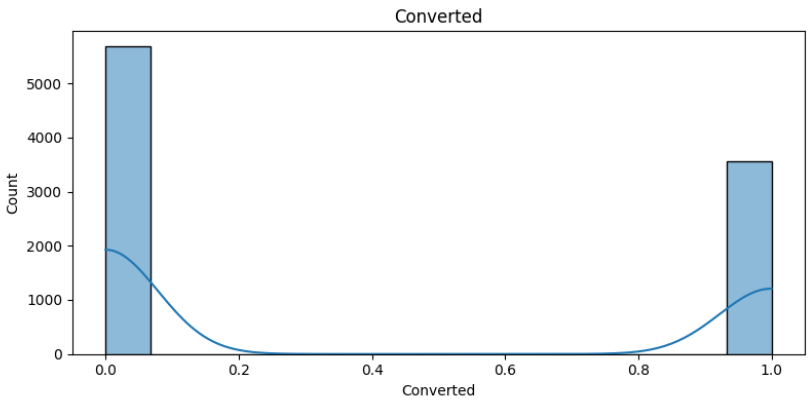


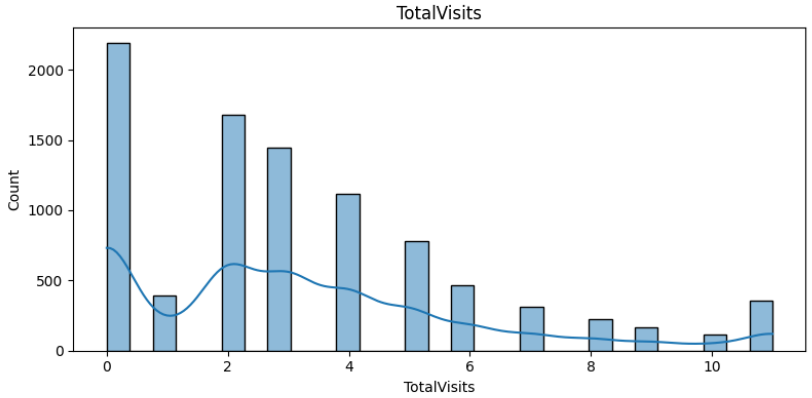
There is a slight imbalance in the Target variable in the given dataset.

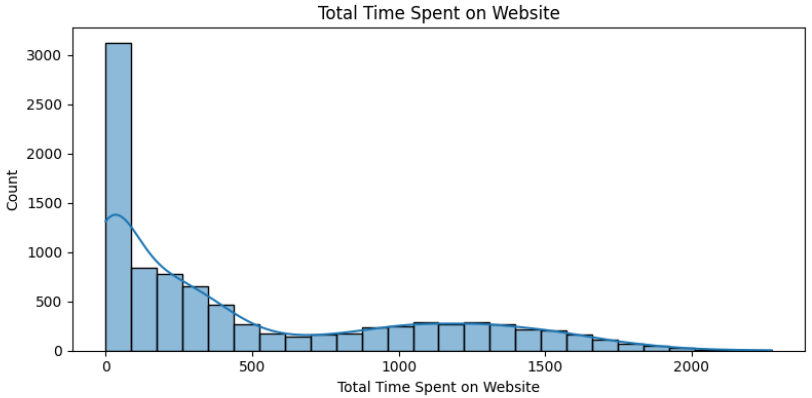
Exploratory Data Analysis

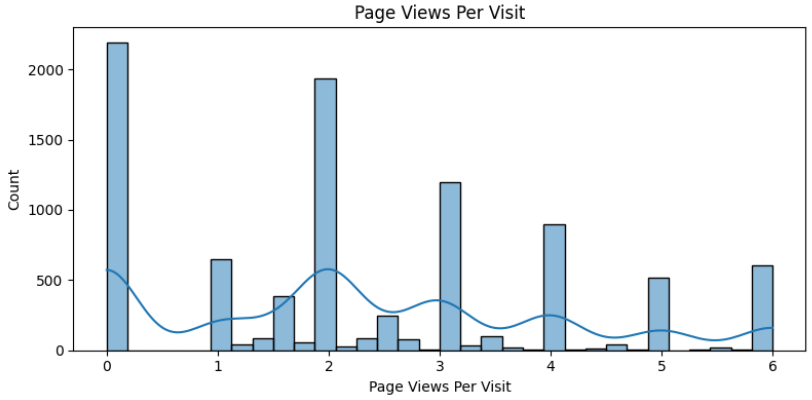
Univariate Analysis - Numerical









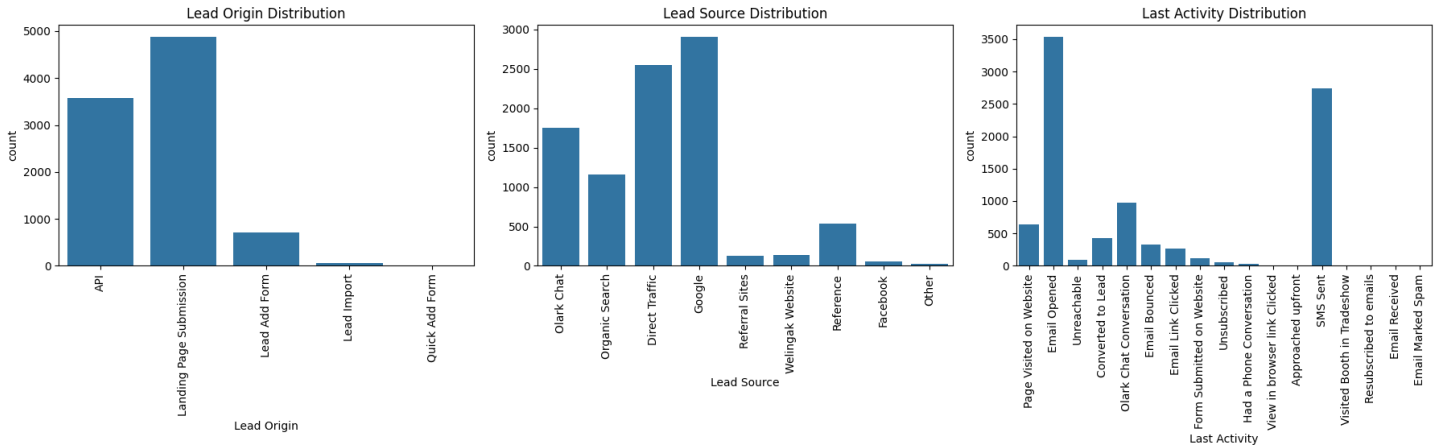


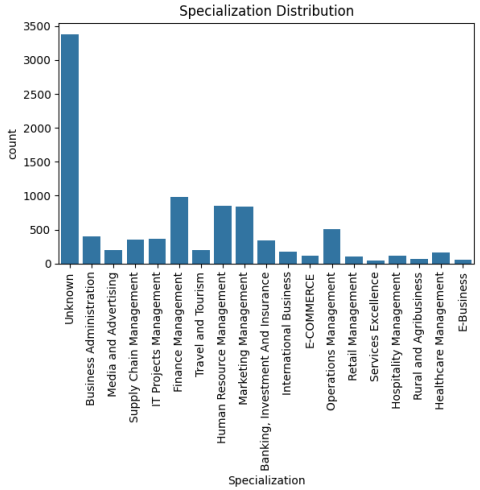
- We can see a slight skewness in the dataset
- It’s a right tailed distribution for most of the numerical features.

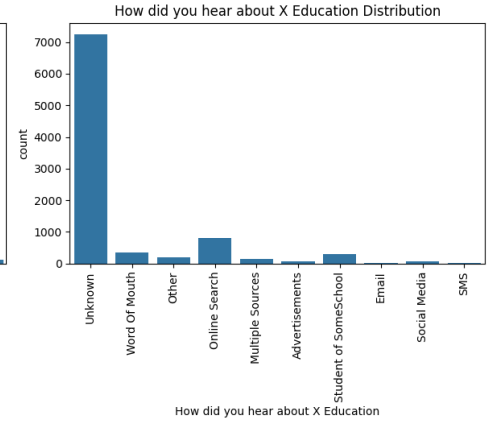
Exploratory Data Analysis

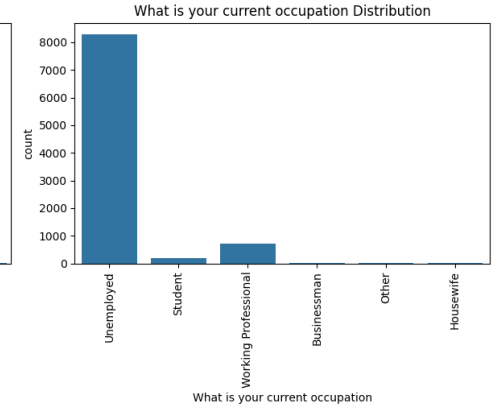
Univariate Analysis - Categorical

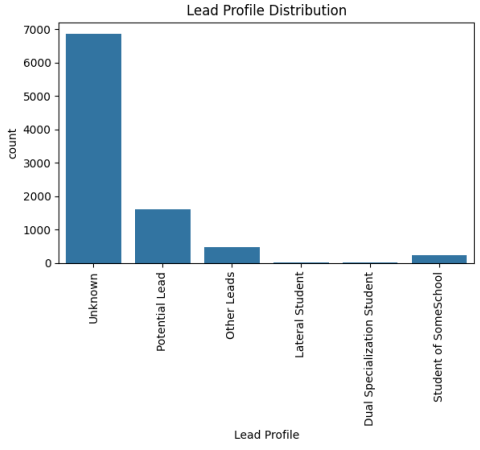
- We can see a huge imbalance in most of the categorical features
- Some of these seem moderately balanced

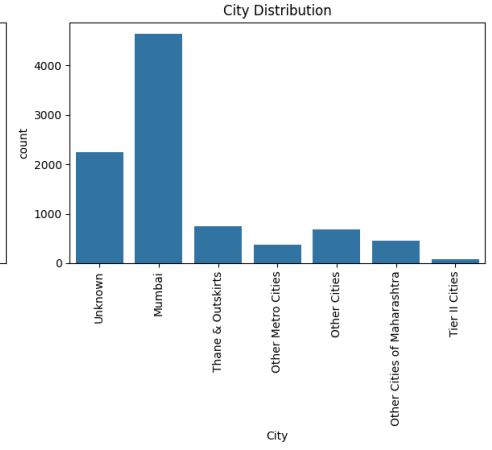


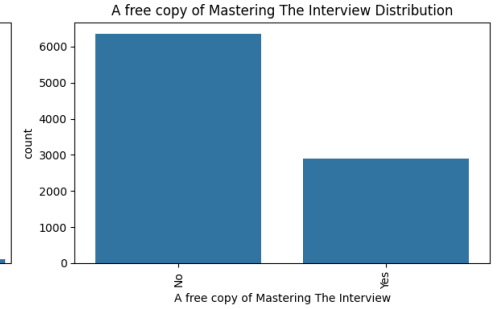








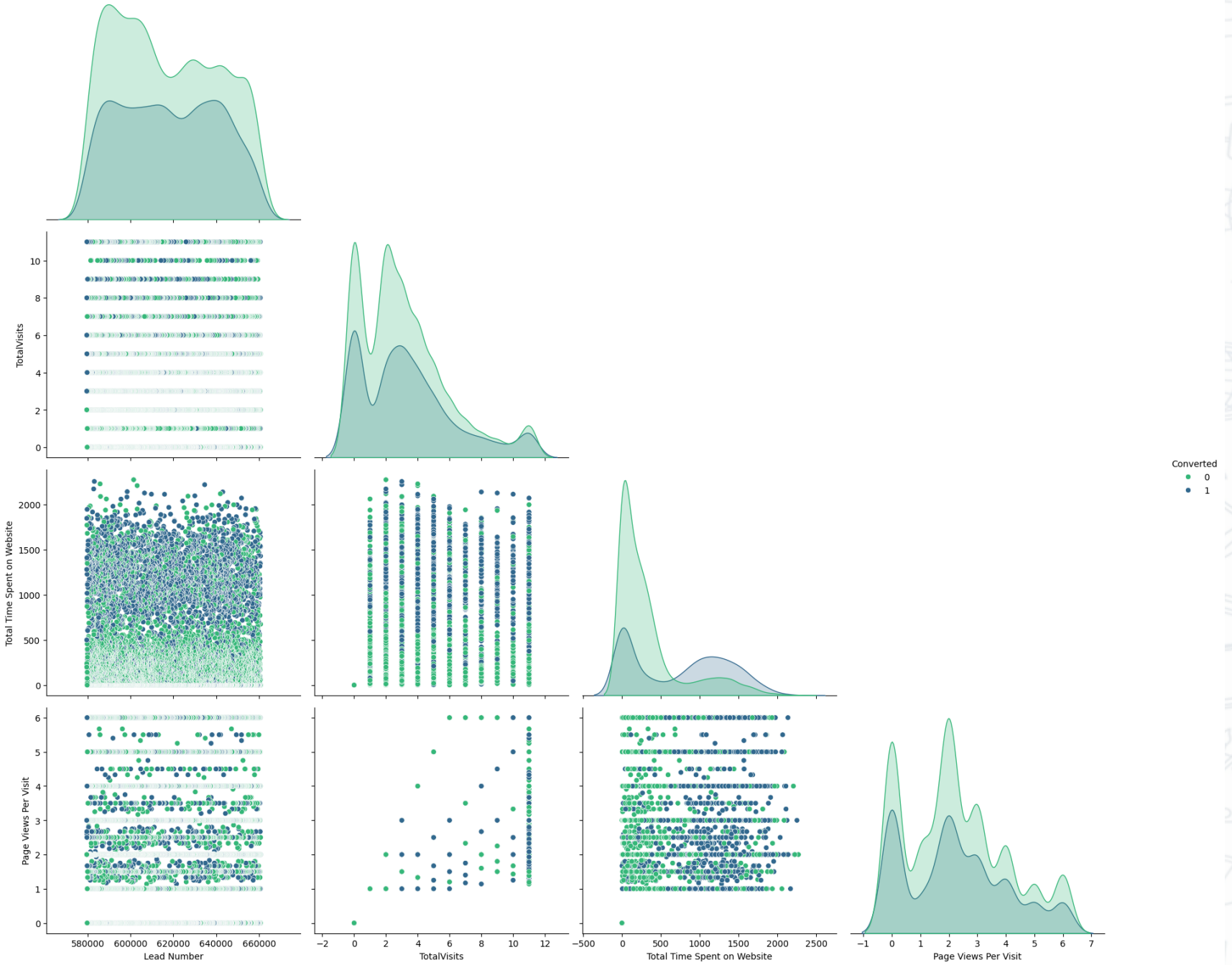




Exploratory Data Analysis

Bivariate Analysis - Numerical

- The only pair showing somewhat linear relationship is between - `TotalVisits` & `Page Views Per Visit`



Exploratory Data Analysis

Multivariate Analysis - Numerical

- A high correlation can be seen between `Page Views Per Visit` & `Total Time Spent on Website`
- A good Correlation can also be seen between `Total Time Spent on Website` & `Converted`
- This could imply that those who are highly interested to buy an education program visit the website often, or spend more time exploring the programs during their visits.

Exploratory Data Analysis

Target Variable Segregated

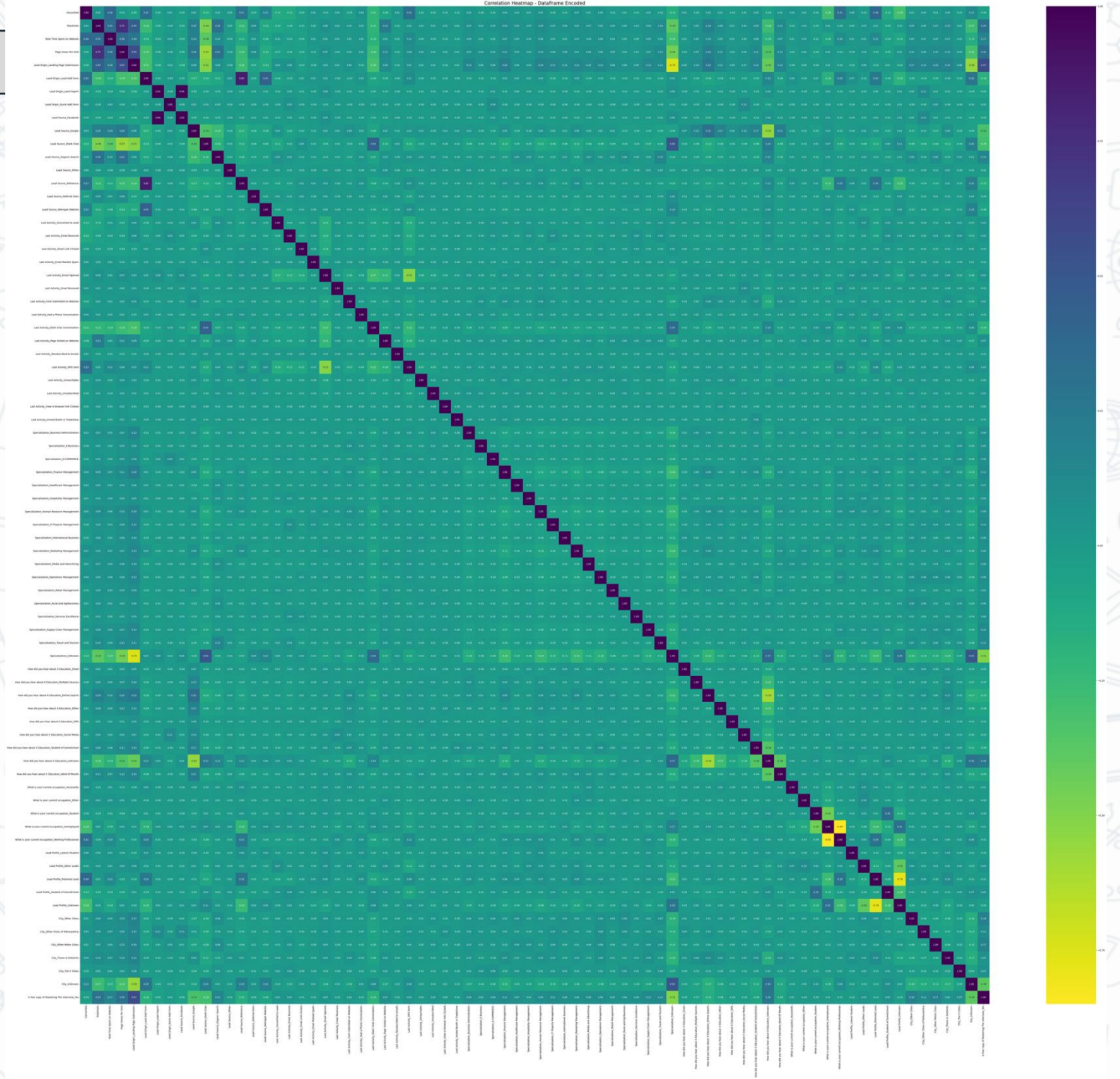
- The correlation between Total Time Spent on Website and both TotalVisits (0.47) and Page Views Per Visit (0.52) is significantly stronger for leads who converted (Target = 1).
- This indicates that for high-potential leads, more visits directly translate into more time spent learning about the offerings.

What it means -

A lead's journey isn't just about the number of visits; it's about the depth and quality of their engagement, which is a powerful signal of their intent to buy.

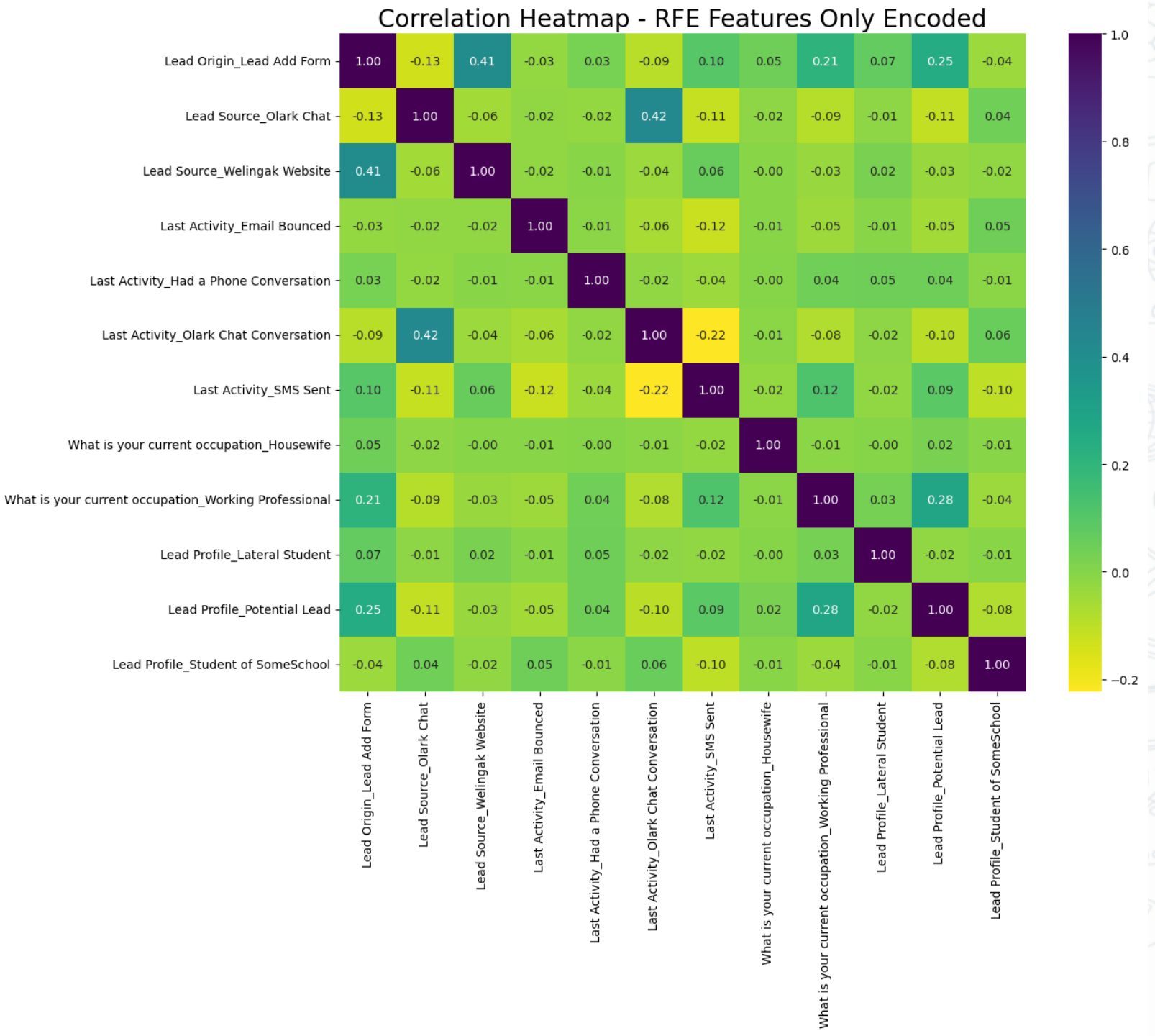
Model Building – Logistic Regression

- We start with one-hot encoding the categorical columns
- As a result, we end up with 76 features as we start our model building process.



Train-Test Split, Scaling & RFE

- We split the data into train & test sets
- Scale the Numerical features using StandardScaler
- Using RFE to quickly filter down 12 features for analysis
- We don’t see extremely high correlation between features here, but we’ll manually check using statsmodels



Final Model

- At the end of the 3rd model, we have no longer any feature with high p-values or high VIFs
- We stop dropping any more features and are left with 10 features

	feature	VIF
0	const	2.096
1	Lead Origin_Lead Add Form	1.360
2	Lead Source_Welingak Website	1.243
9	Lead Profile_Potential Lead	1.164
7	What is your current occupation_Working Profes...	1.134
6	Last Activity_SMS Sent	1.102
5	Last Activity_Olark Chat Conversation	1.076
3	Last Activity_Email Bounced	1.030
10	Lead Profile_Student of SomeSchool	1.018
8	Lead Profile_Lateral Student	1.011
4	Last Activity_Had a Phone Conversation	1.008

Generalized Linear Model Regression Results

Dep. Variable:	Converted	No. Observations:	7392
Model:	GLM	Df Residuals:	7381
Model Family:	Binomial	Df Model:	10
Link Function:	Logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-3513.8
Date:	Tue, 05 Aug 2025	Deviance:	7027.7
Time:	14:41:46	Pearson chi2:	8.68e+03
No. Iterations:	7	Pseudo R-squ. (CS):	0.3154
Covariance Type:	nonrobust		

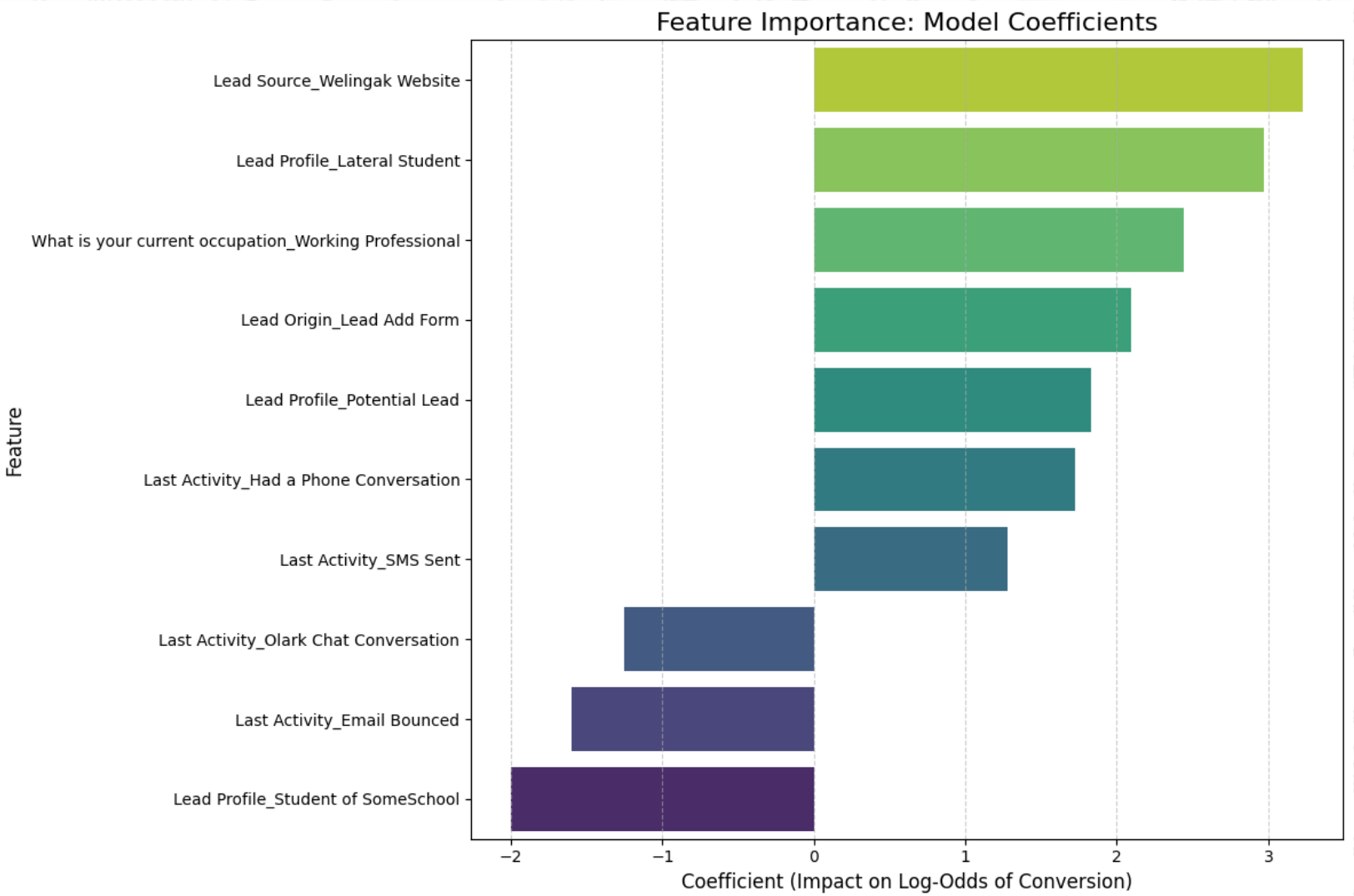
	coef	std err	z	P>|z|	[0.025	0.975]
const	-1.3573	0.043	-31.719	0.000	-1.441	-1.273
Lead Origin_Lead Add Form	2.0947	0.178	11.801	0.000	1.747	2.443
Lead Source_Welingak Website	3.2306	1.023	3.159	0.002	1.226	5.235
Last Activity_Email Bounced	-1.5951	0.264	-6.045	0.000	-2.112	-1.078
Last Activity_Had a Phone Conversation	1.7219	0.557	3.092	0.002	0.630	2.813
Last Activity_Olark Chat Conversation	-1.2498	0.139	-8.961	0.000	-1.523	-0.976
Last Activity_SMS Sent	1.2804	0.063	20.262	0.000	1.157	1.404
What is your current occupation_Working Professional	2.4405	0.164	14.854	0.000	2.119	2.763
Lead Profile_Lateral Student	2.9690	1.081	2.746	0.006	0.850	5.088
Lead Profile_Potential Lead	1.8338	0.083	22.216	0.000	1.672	1.996
Lead Profile_Student of SomeSchool	-2.0008	0.426	-4.701	0.000	-2.835	-1.167

Final Model

- Top Positive Drivers: The strongest signals of a "Hot Lead" are their ‘Lead Source Welingak Website’, or ‘Lead Profile Lateral Student’. A lead’s ‘Occupation Working Professional’ is also a key indicator.
- Top Negative Drivers: Leads profiled as a ‘Student of SomeSchool’ or whose last activity was an ‘Olark Chat’ or a ‘Bounced email’ are highly unlikely to convert.

Strategic Action –

Focus marketing spend on top sources and sales efforts on leads with positive indicators. Automate or deprioritize leads with strong negative indicators.



Model Evaluation – Metrics – Train Set

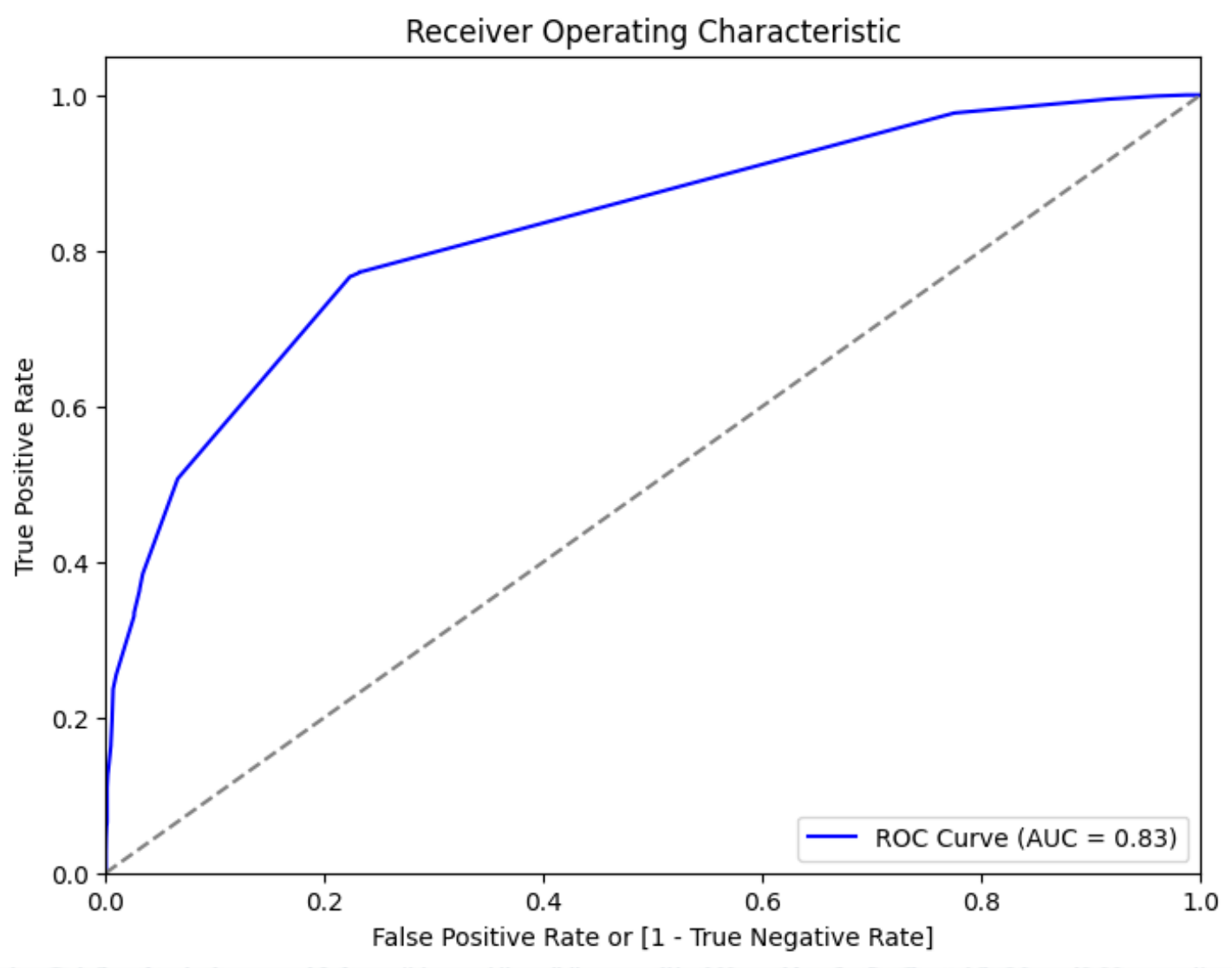
```
Training Performance:
              precision    recall  f1-score   support

           0       0.75      0.93      0.83      4572
           1       0.83      0.51      0.63      2820

    accuracy                           0.77      7392
   macro avg       0.79      0.72      0.73      7392
weighted avg       0.78      0.77      0.76      7392

Confusion Matrix (Training):
[[4271  301]
 [1391 1429]]
```

Accuracy	0.7711
Sensitivity (Recall)	0.5067
Specificity	0.9342

- We take a look at the Classification Report & Confusion Matrix of the Train Set
- Cross-Validation Scores: [0.96178344 0.96496815 0.95329087 0.96815287 0.95855473]
- Mean CV Accuracy: 96.14% (+/- 1.03%)

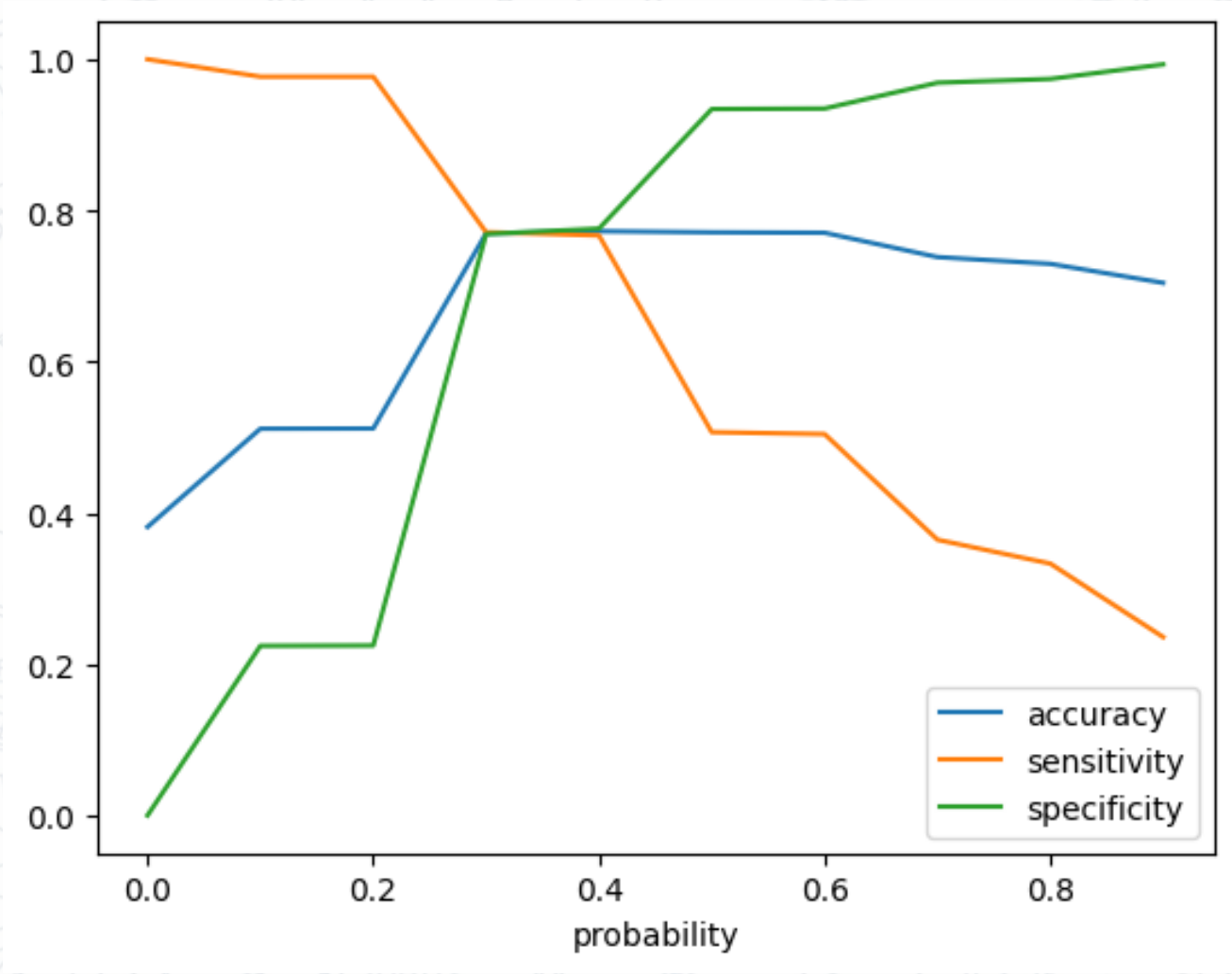
Model Evaluation – ROC AUC – Train Set

- The ROC curve with an AUC of 0.83 indicates that the logistic regression model is performing well.
- This means the model is highly accurate in distinguishing between positive and negative classes. It has a strong ability to correctly classify instances into their respective categories.



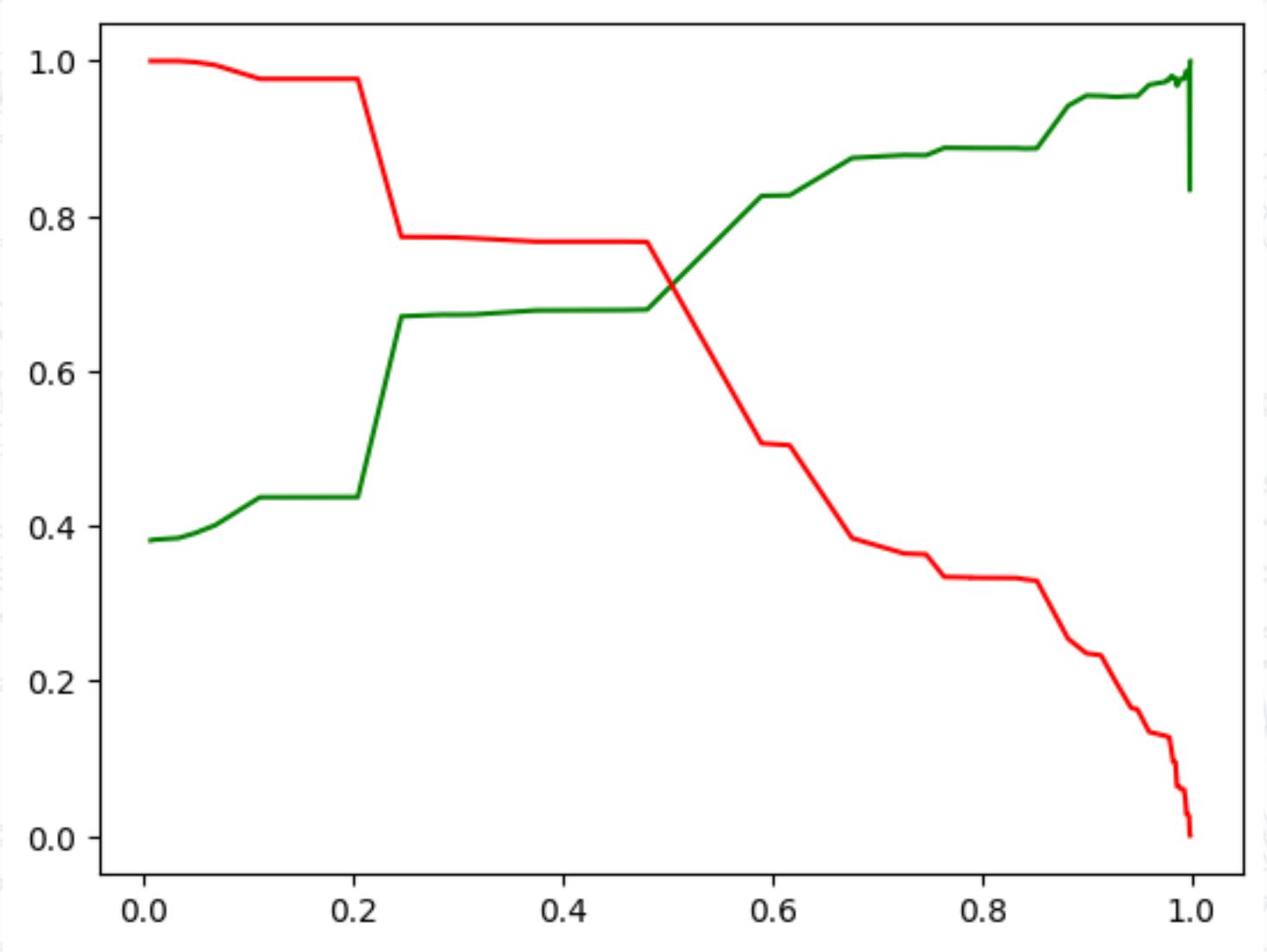
Optimal Cutoff – Accuracy-Sensitivity-Specificity

- We can see that all 3 curves intersect at about 0.35
- The accuracy at this threshold is 0.7723



Optimal Cutoff – Precision-Recall

- We can see that all Precision & Recall intersect at about 0.5
- The accuracy at this threshold is 0.7711



Predictions on Test Set – Evaluation Metrics

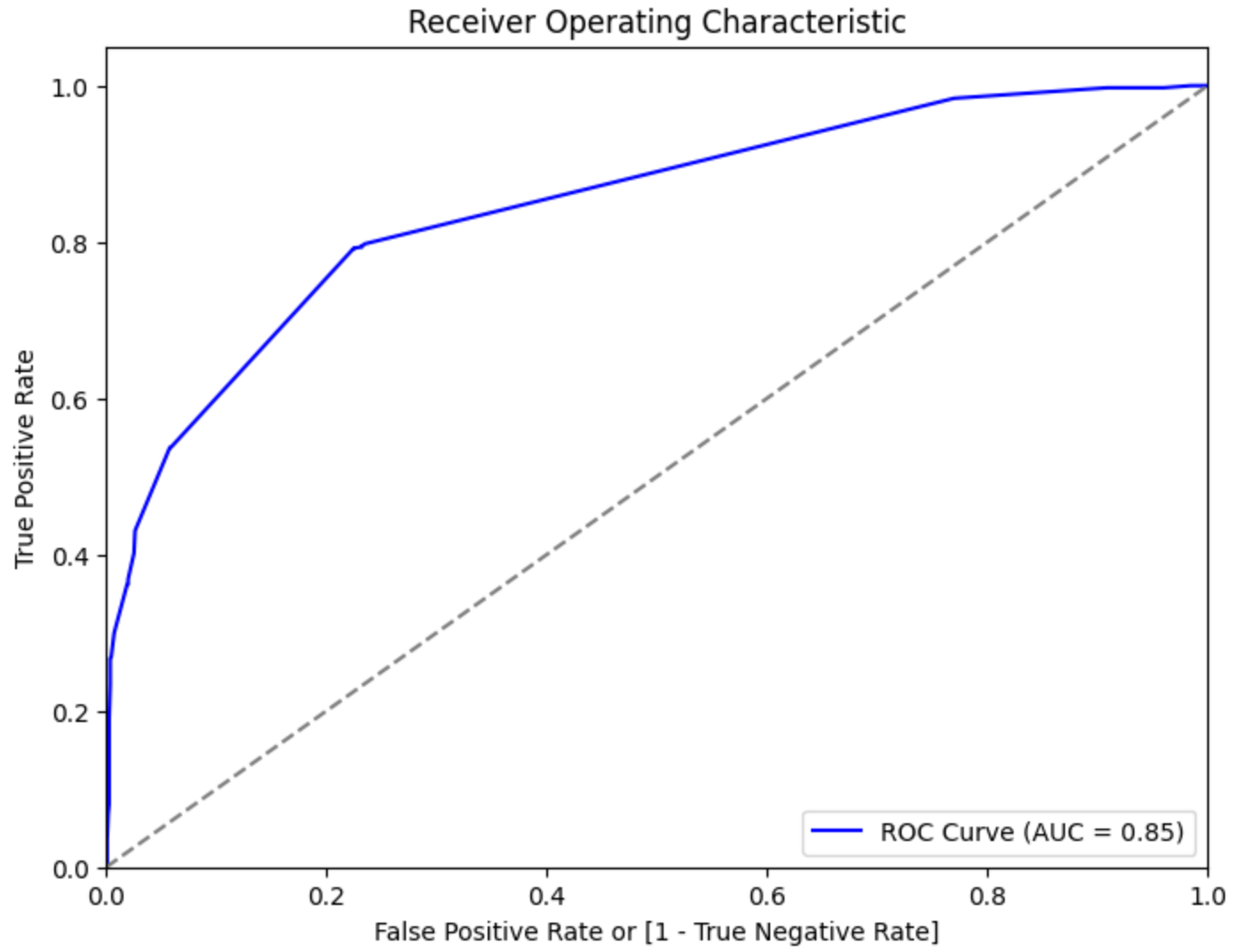
- We check for Accuracy on Test Set using both thresholds we found in the earlier sliders
- The ‘Accuracy-Sensitivity-Specificity’ threshold of 0.4 gives slightly higher accuracy in Test set, so we’ll proceed with this value.

```
Testing Performance:
              precision    recall  f1-score   support

           0       0.85      0.78      0.81      1107
           1       0.70      0.79      0.74       741

    accuracy                           0.78      1848
   macro avg       0.77      0.78      0.78      1848
weighted avg       0.79      0.78      0.78      1848

Confusion Matrix (Testing):
[[858 249]
 [154 587]]
```

Accuracy	0.7819
Sensitivity (Recall)	0.7922
Specificity	0.7751

Model Evaluation – ROC AUC – Test Set

- In the Test Set we see an ROC curve with an AUC of 0.85
- This means the model is highly accurate in distinguishing between positive and negative classes and can correctly classify instances into their respective categories.



Lead Score & Priority Labels

- Finally, we assign Lead Scores to each Lead
- Lead Score is basically the probability of the Lead to Convert multiplied by 100
- We also categorized the Leads as – Very High, High, Medium & Low Priority – based on their Lead Scores
- priority level based on a lead score:
  - Score > 80: Very High
  - Score > 60: High
  - Score > 40: Medium
  - Score ≤ 40: Low
- Higher scores indicate higher priority levels.

Key Findings

- **Overall Accuracy**: 77.11 % on train set, with consistent performance of 78.19% on the test set
- **ROC AUC Score**: 0.85, indicating excellent discrimination between converted and non-converted leads
- **Good Sensitivity (Recall):** 79.22%, demonstrating strong ability to identify actual conversions
- **Good Specificity:** 77.51%, showing robust performance in correctly identifying non-converting leads
- **Optimal Probability Threshold:** Identified at 0.35 using Accuracy-Sensitivity-Specificity curve analysis
- **Feature Significance:** Successfully reduced feature set while maintaining high predictive performance

Recommendations

1. **Deploy the Lead Scoring Model**
   - Use the model's score (0-100) to prioritize all incoming leads. This allows the sales team to instantly focus on "Hot Leads" and stop wasting time on those with low conversion probability.
2. **Focus Marketing on High-Performing Segments**
   - Allocate more budget and strengthen partnerships with the Welingak Website source.
   - Create targeted campaigns to attract Lateral Student and Working Professional profiles, as the model proves these segments have the highest ROI.
3. **Differentiate Outreach Based on Lead Quality**
   - Use high-touch engagement (personal calls, SMS) for top-scoring leads.
   - Automate communication for low-potential segments (e.g., those from Olark Chat or profiled as Student of SomeSchool) to save valuable sales time.
4. **Enhance Data Collection for Future Success**
   - Make key fields like Specialization and Occupation mandatory on all inquiry forms. This will improve future model accuracy and provide richer business insights.
5. **Adopt a Dynamic Strategy for Sales Efforts**
   - Aggressive Phase (e.g., Intern Period): Target leads with scores > 35 to maximize reach and capture ~79% of all potential conversions.
   - Efficiency Phase: Target leads with scores > 80 to minimize effort and ensure every call has the highest chance of success.