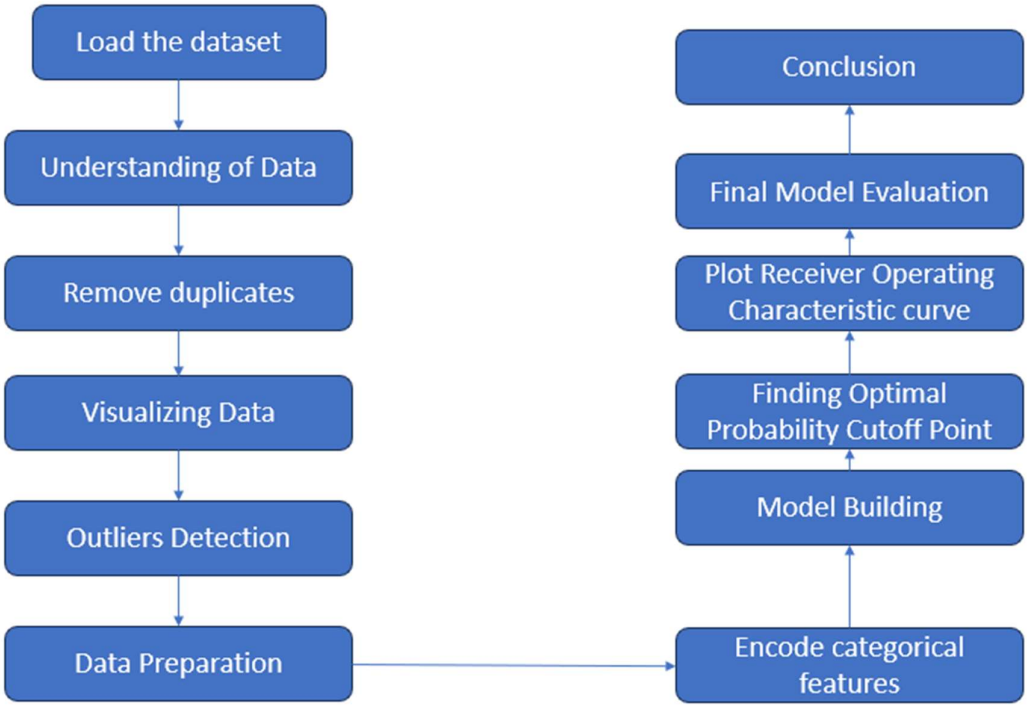
Lead Scoring Case Study Summary Report

Problem Statement:

An education company named X Education sells online courses to industry professionals. On any given day, many professionals who are interested in the courses land on their website and browse for courses.

The company markets its courses on several websites and search engines like Google. Once these people land on the website, they might browse the courses or fill up a form for the course or watch some videos. When these people fill up a form providing their email address or phone number, they are classified to be a lead. Moreover, the company also gets leads through past referrals. Once these leads are acquired, employees from the sales team start making calls, writing emails, etc. Through this process, some of the leads get converted while most do not. The typical lead conversion rate at X education is around 30%.

Goal: Building a logistic regression model to identify the Hot Leads also generating a column that contains Lead Score (Probablity value * 100). It'll help the business in achieving Higher Lead Conversion Rate.



Reading and Understanding the Data

1. The dataset Leads.csv was loaded and a copy was created for further processing.
2. The dataset contains various features related to leads, such as `Lead Origin`, `Lead Source`, `Last Activity`, `Country`, `Specialization`, `TotalVisits`, `Total Time Spent on Website`, etc.
3. Initial exploration included checking the shape, data types, and descriptive statistics of the dataset.

	Prospect ID	Lead Number	Lead Origin	Lead Source	Do Not Email	Do Not Call	Converted	TotalVisits	Time Spent on Website	Page Views Per Visit	...	Get updates on DM Content	Lead Profile	City	Asymmetrique Activity Index	Asymmetrique Profile Index	A
0	7927b2df-8bba-4d29-b9a2-b6e0beafe620	660737	API	Olark Chat	No	No	0	0.0	0	0.0	...	No	Select	Select	02.Medium	02.Medium	
1	2a272436-5132-4136-86fa-dcc88c88f482	660728	API	Organic Search	No	No	0	5.0	674	2.5	...	No	Select	Select	02.Medium	02.Medium	
2	8cc8c611-a219-4f35-ad23-fdfd2656bd8a	660727	Landing Page Submission	Direct Traffic	No	No	1	2.0	1532	2.0	...	No	Potential Lead	Mumbai	02.Medium	01.High	
3	0cc2df48-7cf4-4e39-9de9-19797f9b38cc	660719	Landing Page Submission	Direct Traffic	No	No	0	1.0	305	1.0	...	No	Select	Mumbai	02.Medium	01.High	
4	3256f628-e534-4826-9d63-4a8b88782852	660681	Landing Page Submission	Google	No	No	1	2.0	1428	1.0	...	No	Select	Mumbai	02.Medium	01.High	
5	2058ef08-2858-443e-a01f-a9237db2f5ce	660680	API	Olark Chat	No	No	0	0.0	0	0.0	...	No	NaN	NaN	01.High	02.Medium	
6	9fae7df4-169d-489b-afe4-0f3d752542ed	660673	Landing Page Submission	Google	No	No	1	2.0	1640	2.0	...	No	Potential Lead	Mumbai	02.Medium	01.High	
7	20ef72a2-fb3b-45e0-924e-551c5fa59095	660664	API	Olark Chat	No	No	0	0.0	0	0.0	...	No	NaN	NaN	02.Medium	02.Medium	
8	cfa0128c-a0da-4656-9d47-0aa4e67bf690	660624	Landing Page Submission	Direct Traffic	No	No	0	2.0	71	2.0	...	No	NaN	Thane & Outskirts	02.Medium	02.Medium	
9	af465dfc-7204-4130-9e05-33231863c4b5	660616	API	Google	No	No	0	4.0	58	4.0	...	No	NaN	Mumbai	02.Medium	02.Medium	

Data Cleanup

- Duplicate Rows: No duplicate entries were found in the dataset.

- Handling 'Select' Values: Replaced 'Select' with NaN in columns where it was used as a default value.

- Columns with No Variance: Dropped columns with only one unique value.

- High Missing Values: Dropped columns with more than 40% missing values.

- Categorical Columns: Grouped categories with very few observations into a single group and imputed missing values using business knowledge.

```
<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                                 5860 non-null   object 
 13  How did you hear about X Education             1990 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                                   2385 non-null   object 
 29  City                                           5571 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
```

Visualizing Data

Numerical Variables: Visualized using pair plots and box plots.

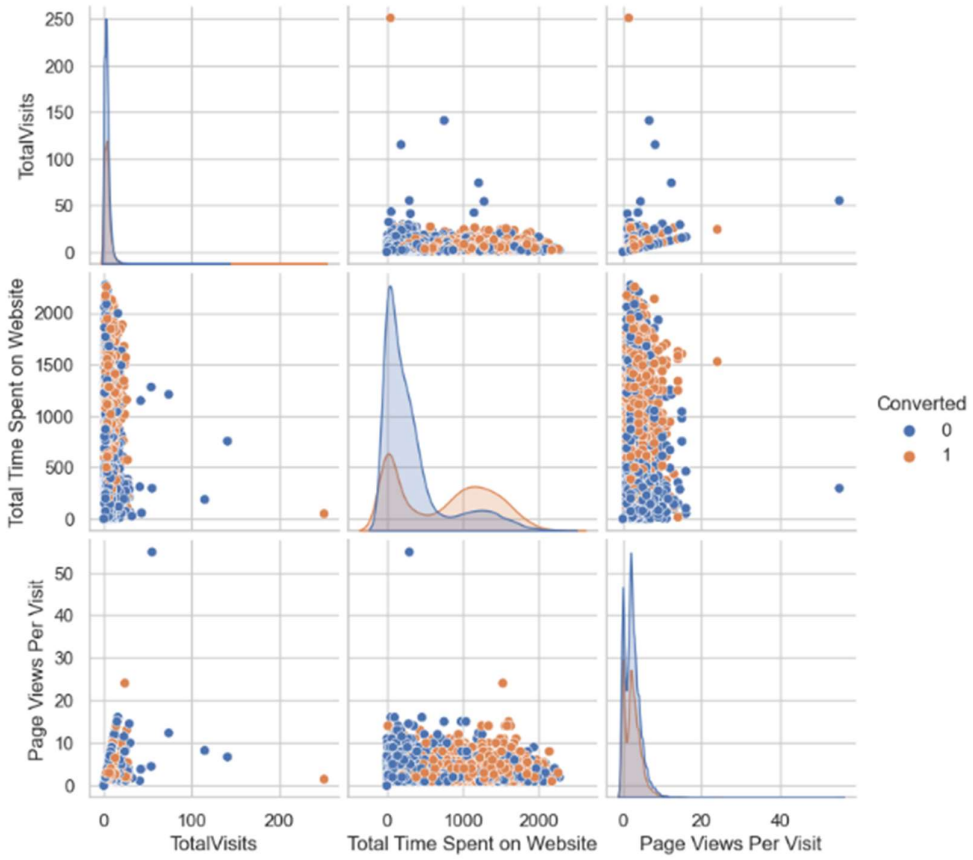
Observed that `Total Time Spent on Website` is significantly higher for converted leads.

Identified outliers in `Total Visits` and `Total Time Spent on Website`.

Categorical Variables: Visualized using count plots.

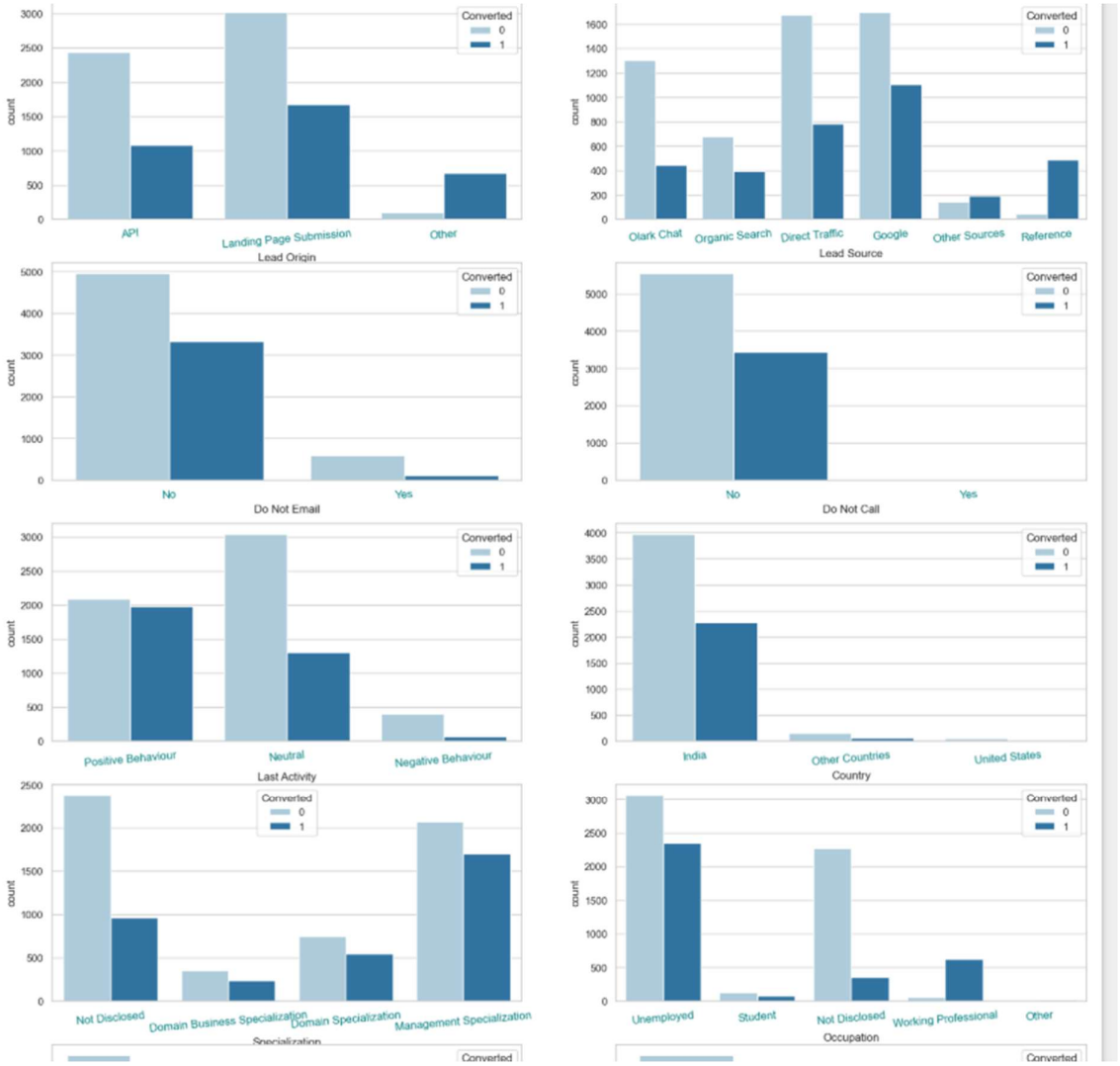
Identified that certain lead origins and sources have higher conversion rates.

Positive behavior in `Last Activity` is associated with higher conversion rates.



Outliers Detection & Visualizing Categorical Values

- Removed outliers in `Page Views Per Visit`, `TotalVisits`, and `Total Time Spent on Website` based on percentile values.



Data Preparation

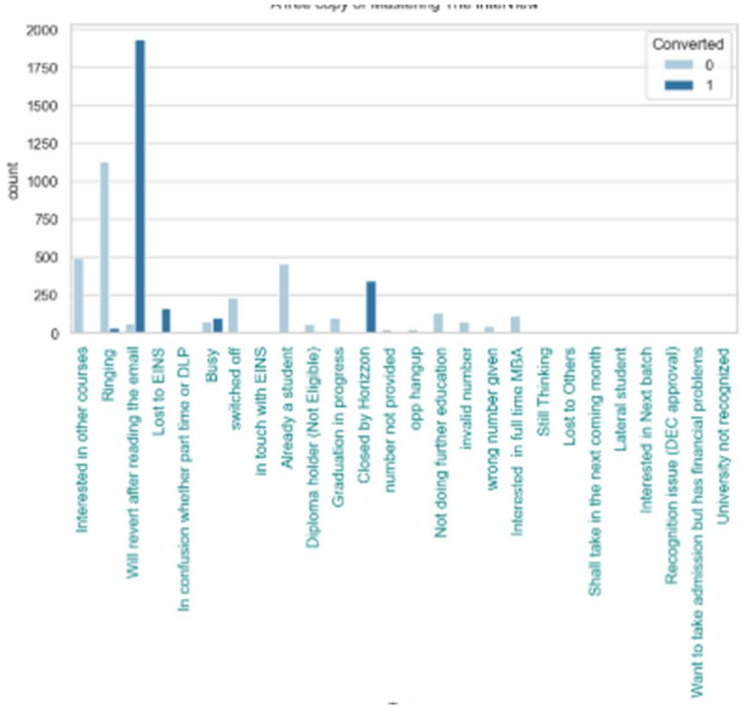
- Train-Test Split: Split the data into training and testing sets.

- Missing Value Imputation: Imputed missing values in categorical features with mode and continuous features with median.

- Encoding Categorical Variables: Converted binary categorical columns to 0 and 1, and used `pd.get_dummies()` for nominal categorical columns.

- Scaling: Applied MinMax scaling to numeric predictors.

- Variance Thresholding: Removed columns with very low variance.



Model Building

- Initial Model: Built a logistic regression model using Recursive Feature Elimination (RFE) to select the top 16 features.

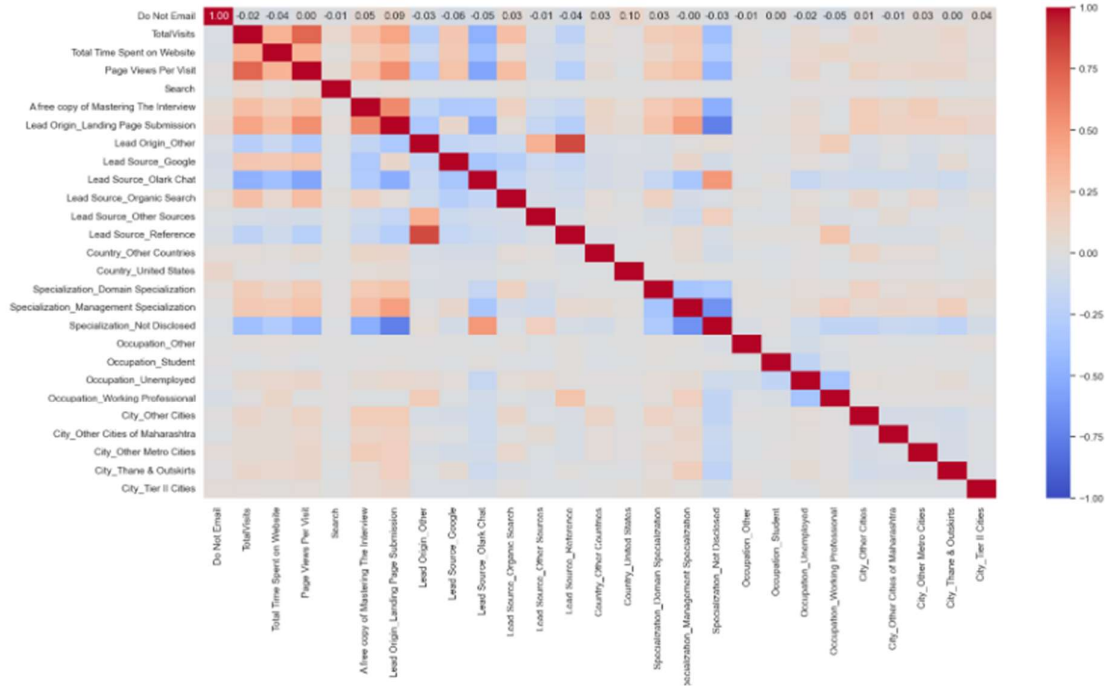
- Model Iterations: Iteratively refined the model by removing features with high p-values and checking for multicollinearity using VIF.

- Final Model: The final logistic regression model (`lreg_model_7`) was built with statistically significant predictors and no multicollinearity.

	Prospect ID	Lead Number	Lead Origin	Lead Source	Do Not Email	Do Not Call	Converted	TotalVisits	Time Spent on Website	Page Views Per Visit	...	Search	Newspaper Article	X Education Forums	Newspaper	Digital Advertisement	R
0	7927b2df-8bba-4d29-b9a2-b6e0beafe620	660737	API	Olark Chat	No	No	0	0.0	0	0.0	...	No	No	No	No	No	
1	2a272436-5132-4136-86fa-dcc88c88f482	660728	API	Organic Search	No	No	0	5.0	674	2.5	...	No	No	No	No	No	
2	8cc8c611-a219-4f35-ad23-fdfd2656bd8a	660727	Landing Page Submission	Direct Traffic	No	No	1	2.0	1532	2.0	...	No	No	No	No	No	
3	0cc2df48-7cf4-4e39-9de9-19797f9b38cc	660719	Landing Page Submission	Direct Traffic	No	No	0	1.0	305	1.0	...	No	No	No	No	No	
4	3256f628-e534-4826-9d63-4a8b88782852	660681	Landing Page Submission	Google	No	No	1	2.0	1428	1.0	...	No	No	No	No	No	

Checking pairwise correlation

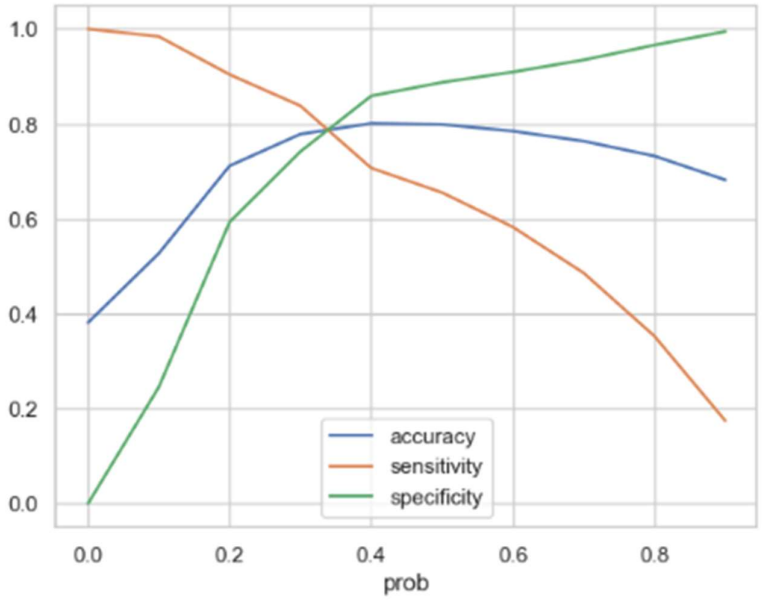
```
[2]: # Creating heatmap

plt.figure(figsize = (20,10))
sns.heatmap(X_train.corr(),annot = True, cmap= 'coolwarm', fmt= '.2f', vmin= -1, vmax= 1)
plt.show()
```



Finding Optimal Probability Cutoff Point

- ROC Curve: Plotted the ROC curve to evaluate model performance.

- Sensitivity-Specificity: Determined the optimal probability cutoff point (0.32) using sensitivity and specificity analysis.



Final Model Evaluation

	Variable	Coefficient	P-Value	Type
3	Total Time Spent on Website	3.979086	1.903845e-182	numerical
10	Occupation_Working Professional	3.814033	3.159241e-88	categorical
4	Lead Origin_Other	3.782189	5.201395e-93	categorical
0	const	-3.411581	5.144206e-188	categorical
7	Occupation_Other	1.586532	1.889389e-03	categorical
5	Lead Source_Olark Chat	1.388053	3.158120e-35	categorical
9	Occupation_Unemployed	1.339241	8.120841e-56	categorical
1	Do Not Email	-1.212806	2.781637e-14	categorical
8	Occupation_Student	1.144419	6.437921e-08	categorical
2	TotalVisits	0.954377	4.089916e-05	numerical
6	Lead Source_Other Sources	-0.585311	7.302927e-03	categorical

- Confusion Matrix: Evaluated the final model using the confusion matrix.

- Accuracy: Overall model accuracy was calculated.

- Sensitivity and Specificity: Sensitivity (Recall) and specificity were calculated.

- Positive and Negative Predictive Values: Calculated to understand the model's precision.

Conclusion

- The final logistic regression model effectively identifies hot leads with a good balance of sensitivity and specificity.

The top contributing variables include:

1. Total Time Spent on Website
2. Occupation_Working Professional
3. Lead Origin_Other

- The model can be used to prioritize leads and improve the lead conversion rate for X Education.

- The optimal probability cutoff point (0.32) ensures a good trade-off between sensitivity and specificity, making the model suitable for aggressive lead conversion strategies during peak periods and conservative strategies during off-peak periods.

- **Insights for Decision Making**:
    - **High-Value Leads**: Leads with high Total Time Spent on Website and those coming from Google and Direct Traffic sources should be prioritized as they have a higher likelihood of conversion.
    - **Lead Origin**: Focus on leads originating from Landing Page Submissions as they show a higher conversion rate.
    - **Behavioural Indicators**: Leads with positive behaviors such as Email Opened in their last activity should be given more attention.
    - **Specialization**: Leads who have disclosed their specialization, especially in Management Specialization, are more likely to convert.

- **Country**: Leads from India show a higher conversion rate and should be prioritized.
- **Communication Preferences**: Avoid leads who have opted for Do Not Email as they are less likely to convert.
- **Intern Hiring Period Strategy**: Lower the probability threshold to include more potential leads and prioritize calling leads with the highest predicted probabilities.
- **Off-Peak Period Strategy**: Raise the probability threshold to focus only on the most likely leads to convert, minimizing unnecessary phone calls.