Case Study – Lead Scoring

By:
Supriya Ayinampudi
Surinder Pal Kaur
Suranjan Banerjee

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. They have a process of form filling out on their website, after which the company uses that individual as a lead.
- Once these leads are acquired, employees from the sales team start making calls, writing emails, etc. Through this process, the leads are converted to paying customers who are students of the courses.
- The typical lead conversion rate at **X Education** is around **30%**. To make this process more efficient, the company wishes to identify the most potential leads, also known as hot leads.
- If they successfully identify this set of leads, the lead conversion rate should go up as the sales team will now be focusing more on communicating with the potential leads rather than making calls to everyone
- **X Education** has appointed us to help them select the most promising leads.

Project Goal

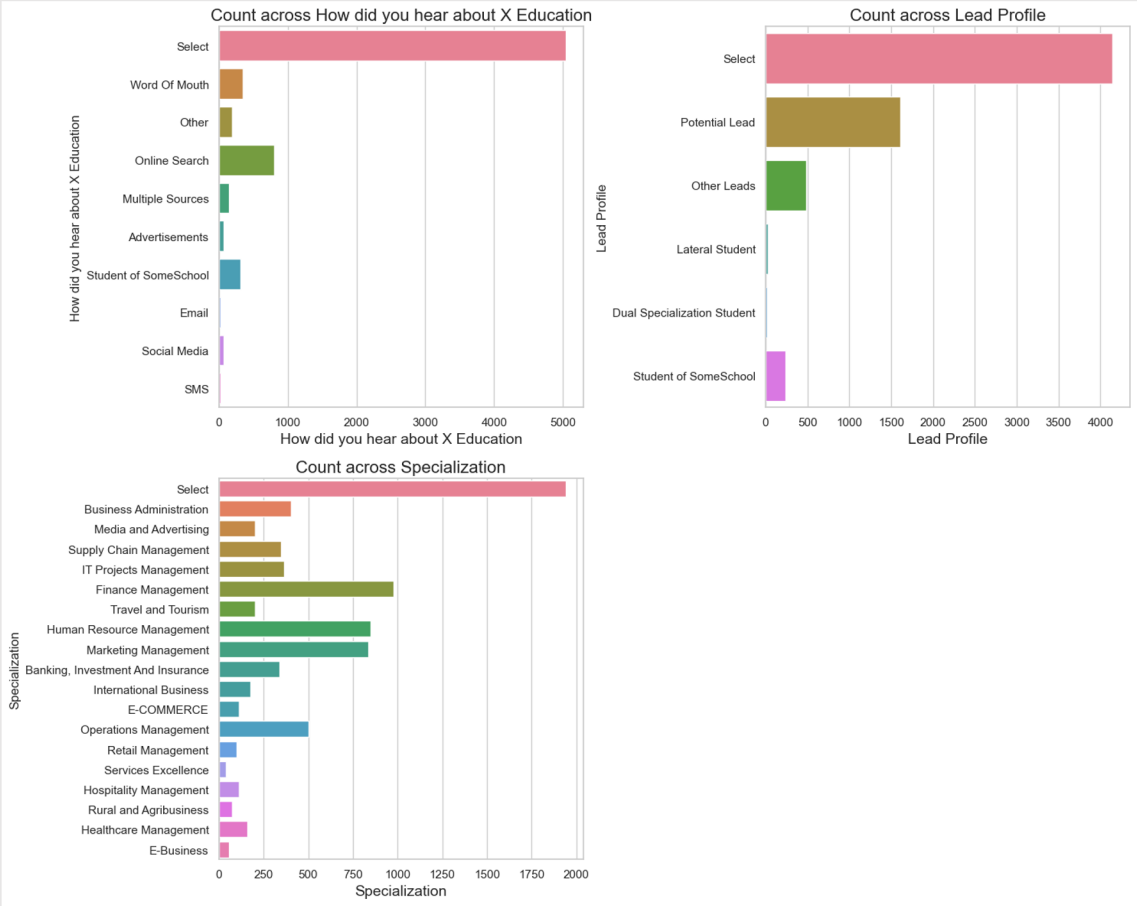
1. Build a model wherein we assign a lead score to each of the leads
2. Customers with a higher lead score have a higher conversion chance, and customers with a lower lead score have a lower conversion chance.
3. We are given a ballpark of the target lead conversion rate to be around **80%**.

Analysis Approach

- **Data Understanding & Preparation:**
  - Import, clean, and prepare the dataset.
- **Exploratory Data Analysis (EDA):**
  - Univariate and bivariate analysis to find key insights.
- **Feature Engineering:**
  - Scaling, creating dummy variables.
- **Model Building:**
  - RFE for feature selection, multicollinearity check with VIF.
- **Model Evaluation:**
  - Evaluation using accuracy, precision, recall, etc.
- **Final Model & Results:**
  - Present final predictions and outcomes.

EDA: Select

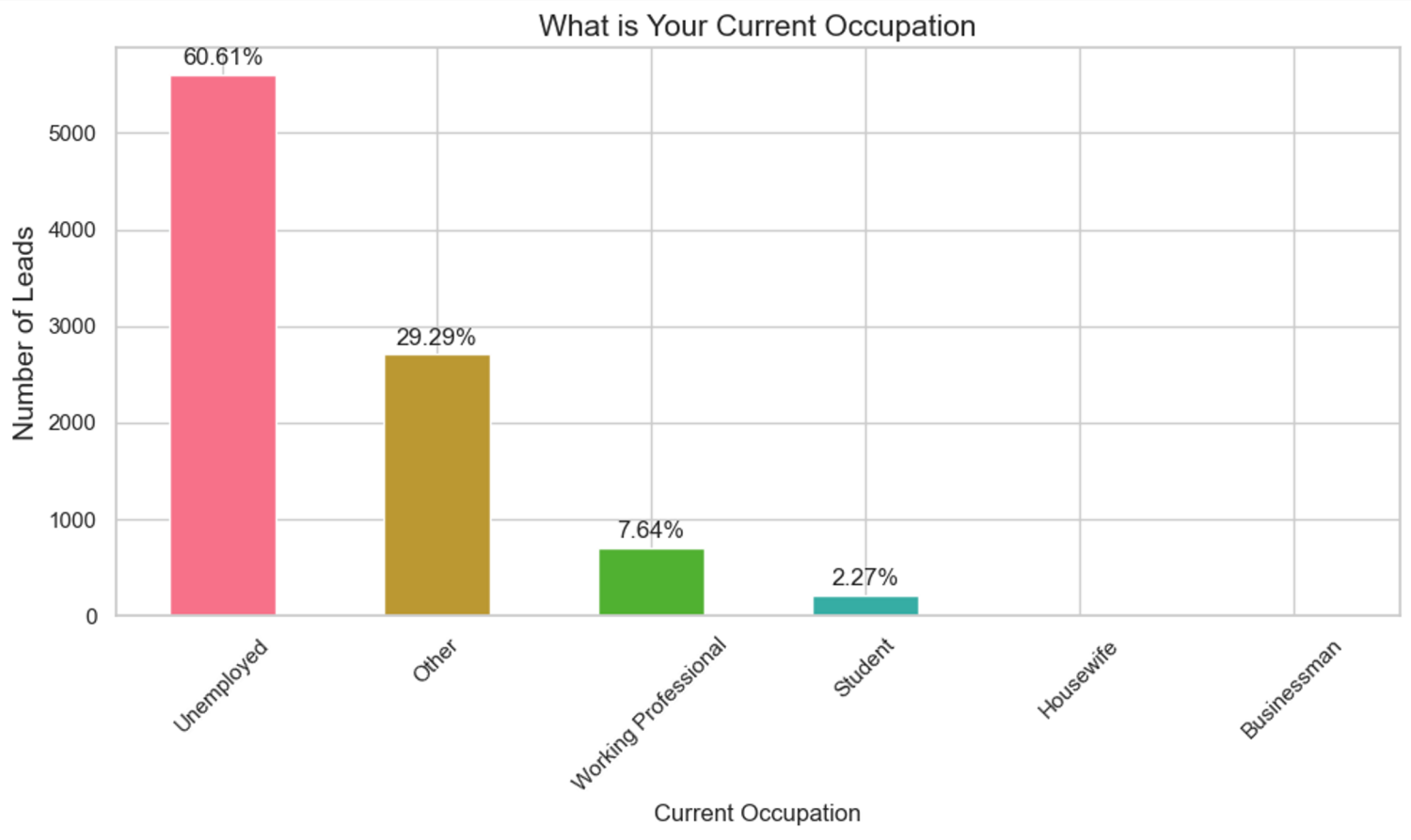
- The data originally contained 37 columns across 9240 records.
  - 17 columns have null values, with 5 containing more than 25% nulls.
  - 4 columns have ‘Select’ as a value, which is most likely the default value on the form where the customer has not made an explicit choice.



EDA: Missing Values

- Treating select as a missing / Null value and imputing accordingly
  - Dropping columns with more than 40% missing values
  - Replacing nulls with 0, unknown or other as appropriate

For example, in Current Occupation, the null values are imputed with other

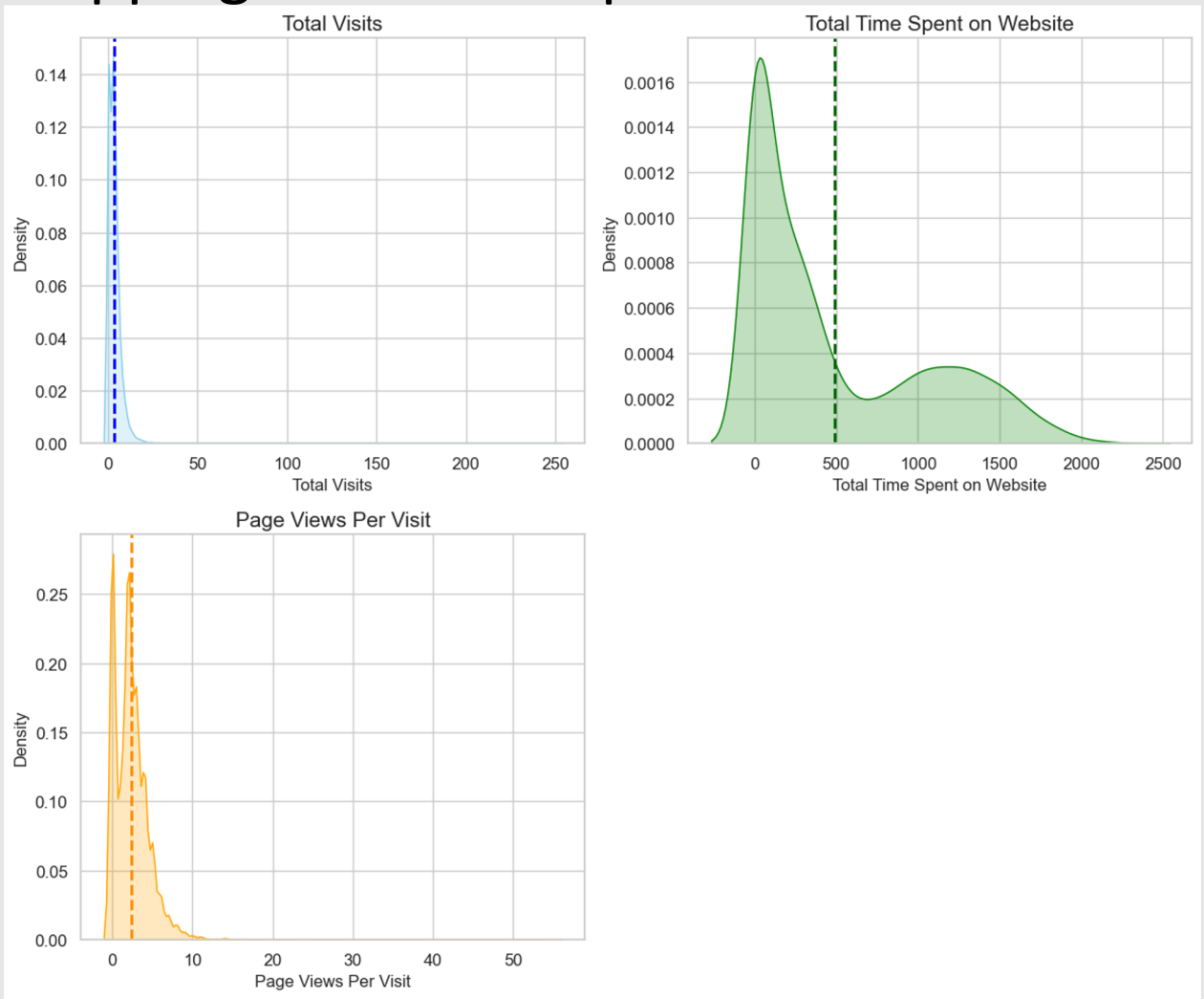


Data Understanding

- Exploring all columns to understand data
  - Converted is the Target variable (y)—with *38.5% leads converted*
  - Product ID & Lead number are unique ID values which do not contribute to the analysis
  - Columns which have heavily skewed data are dropped (95%+ values are the same)
  - Since this is an online education, it is assumed that city and country do not play a part in the targeting of customers.

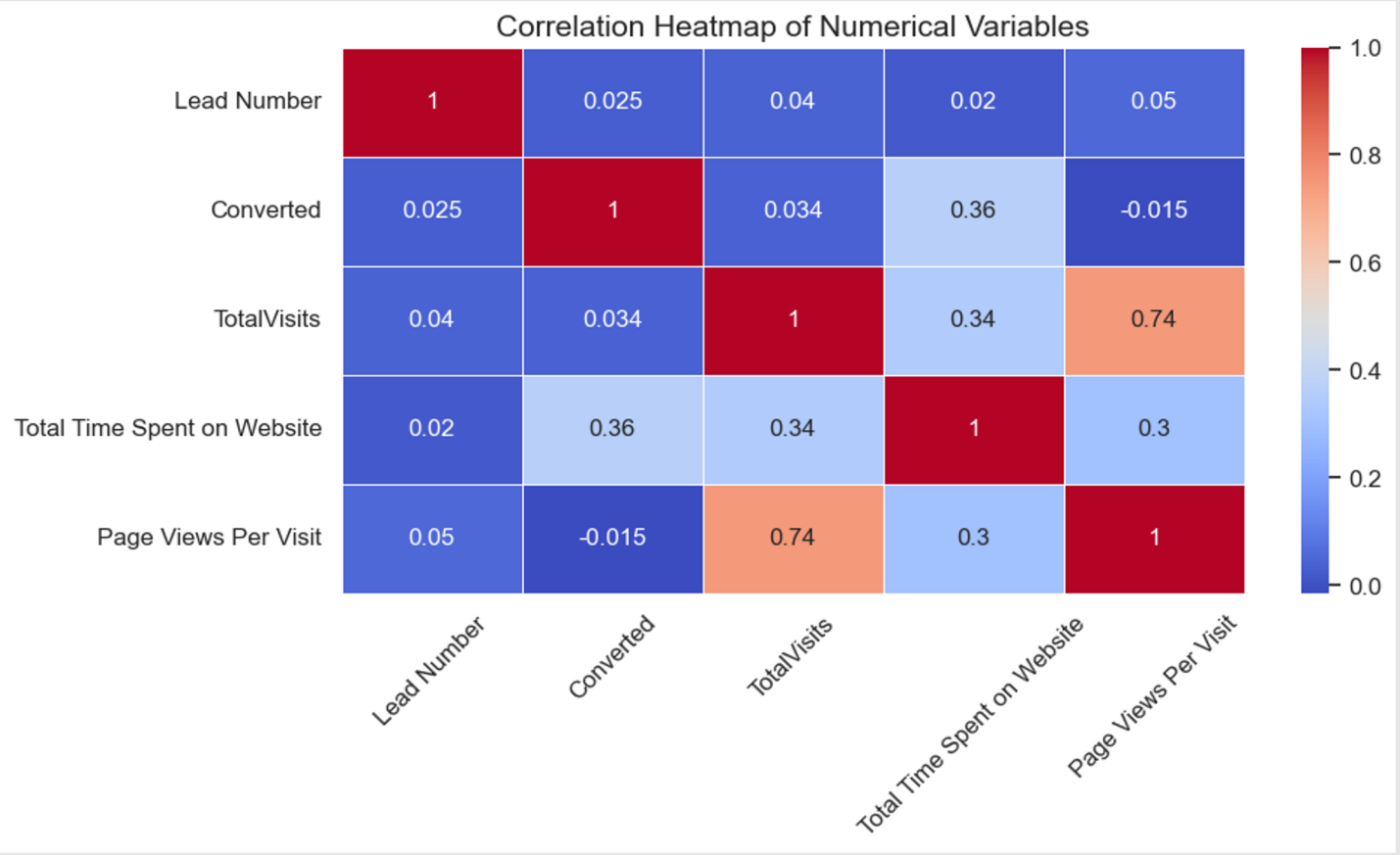
EDA: Handing outliers

- Handling Outliers- The outliers were found in the below features and were handled by capping them at 95 percentile.



EDA: Numerical correlation

- Correlation Heatmap of numerical features



- **Total Visits** and **Page Views Per Visit** have a high positive correlation (0.74), suggesting potential multicollinearity, as these variables might provide redundant information.
- **Converted** has a moderate positive correlation (0.36) with **Total Time Spent on Website**, indicating that users who spend more time on the website are more likely to convert.

Feature Engineering

- Feature Engineering
  - Numeric variables were scaled
  - Categorical variables we split to dummy variables to allow for logistic regression
- **There is no clear or high correlation between the numeric variables.**

Modelling Methodology

Test / train split:

- Converted is used as the ‘y’ variable, with the remaining 127 columns being focused on as the ‘X’ features.
- Data is split, with 70% being used to train the model and 30% used to test the final model.

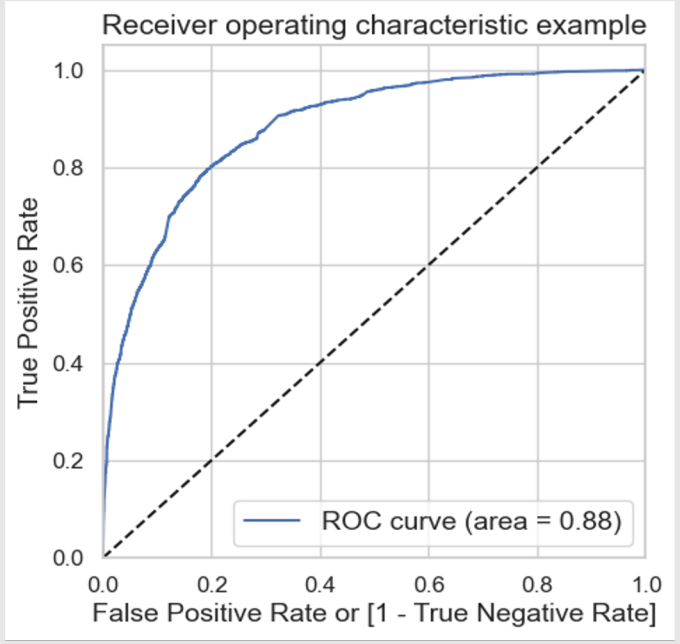
Model building:

- Recursive feature elimination (RFE) is used to select the 15 most relevant features.
- Dropped features with high p-value as they do not suggest a significant relationship with the target variable
- Using VIF to identify features to drop to reduce multicollinearity

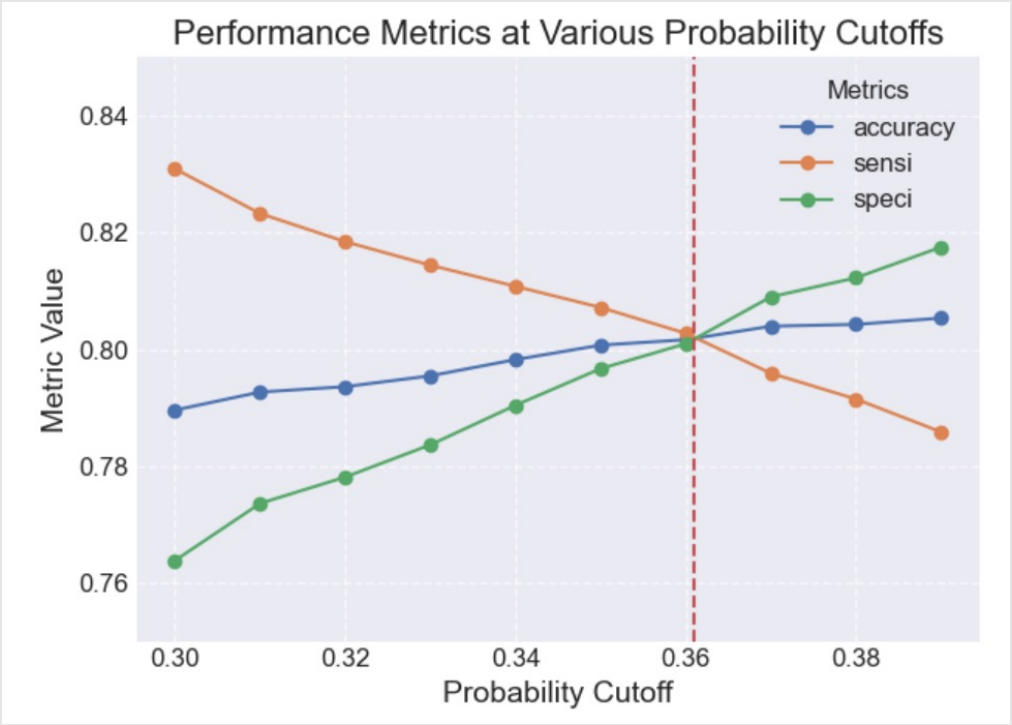
Modelling Methodology

Predicting conversion:

- Using ROC curve to identify the optimum cut-off point.



- Evaluating the model using Accuracy, Confusion Matrix, Specificity, Sensitivity, etc to arrive at the optimal cut-off.



Model Evaluation

- Confusion matrix*:

Training Data	Actual Positive	Actual Negative
Predicted Positive	**3195**	**789**
Predicted Negative	**491**	**1993**

Test Data	Actual Positive	Actual Negative
Predicted Positive	**1348**	**347**
Predicted Negative	**168**	**909**

- Comparison of evaluation metrics for final model*:

Model evaluation	Accuracy	Sensitivity	Specificity	Precision	Recall	F1 score
Train Data	**80.21**	**80.23**	**80.2**	**71.64**	**80.23**	**75.69**
Test Data	**81.42**	**84.4**	**79.53**	**72.37**	**84.4**	**77.93**

* The above are at a prob cut-off of 0.361

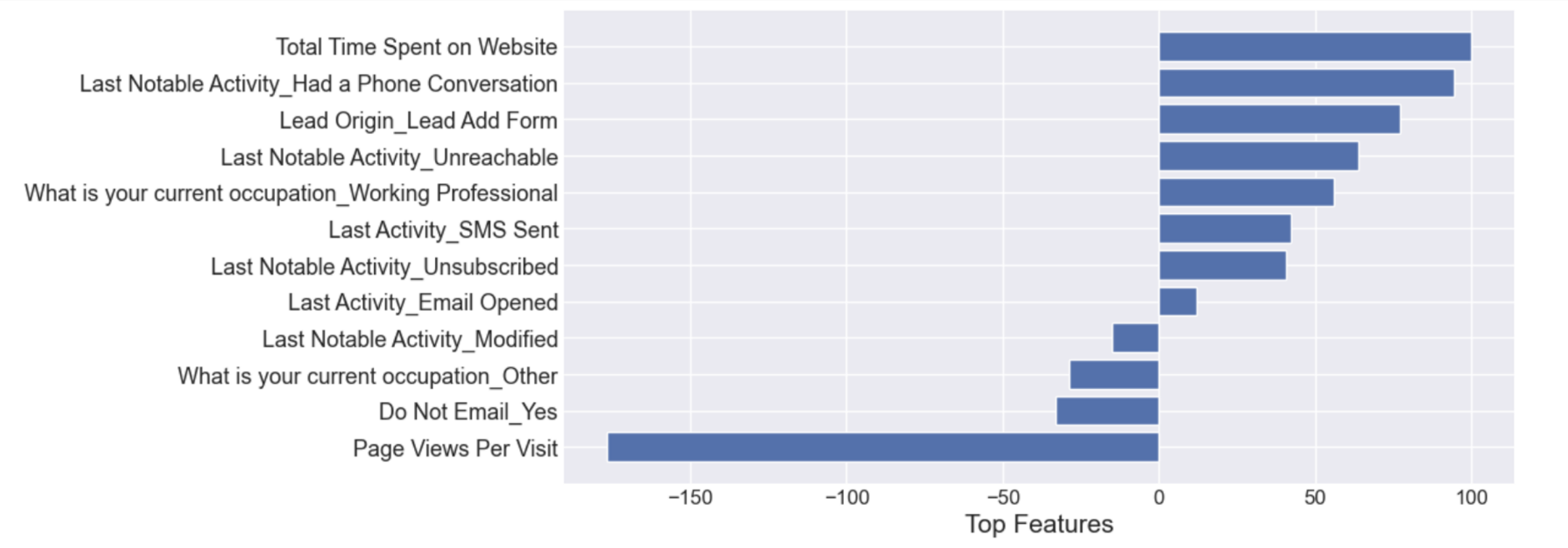
Final Model Stats

Generalized Linear Model Regression Results

Dep. Variable:	Converted	No. Observations:	6468
Model:	GLM	Df Residuals:	6455
Model Family:	Binomial	Df Model:	12
Link Function:	Logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-2712.7
Date:	Sun, 20 Oct 2024	Deviance:	5425.3
Time:	16:59:28	Pearson chi2:	6.49e+03
No. Iterations:	6	Pseudo R-squ. (CS):	0.3893
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-1.6412	0.106	-15.442	0.000	-1.850	-1.433
Total Time Spent on Website	4.0079	0.151	26.482	0.000	3.711	4.304
Page Views Per Visit	-7.0742	1.039	-6.809	0.000	-9.111	-5.038
Lead Origin_Lead Add Form	3.0971	0.193	16.011	0.000	2.718	3.476
Do Not Email_Yes	-1.3191	0.171	-7.709	0.000	-1.654	-0.984
Last Activity_Email Opened	0.4766	0.101	4.710	0.000	0.278	0.675
Last Activity_SMS Sent	1.6939	0.102	16.617	0.000	1.494	1.894
What is your current occupation_Other	-1.1521	0.086	-13.420	0.000	-1.320	-0.984
What is your current occupation_Working Professional	2.2526	0.180	12.505	0.000	1.900	2.606
Last Notable Activity_Had a Phone Conversation	3.7927	1.090	3.479	0.001	1.656	5.930
Last Notable Activity_Modified	-0.5996	0.085	-7.016	0.000	-0.767	-0.432
Last Notable Activity_Unreachable	2.5542	0.561	4.551	0.000	1.454	3.654
Last Notable Activity_Unsubscribed	1.6306	0.531	3.071	0.002	0.590	2.671

Interpreting Results



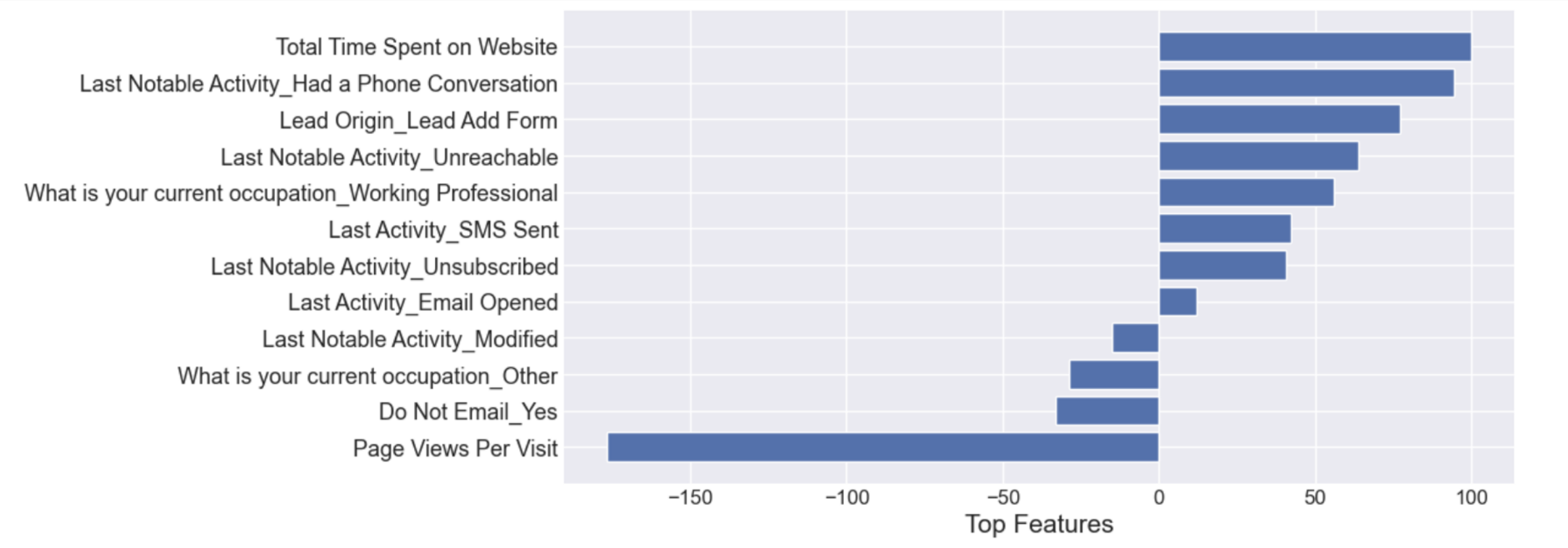
Positive Features:

- **Total Time Spent on Website**: Leads spending more time on the website are more likely to convert.
- **Phone Conversations**: Having a phone conversation increases conversion likelihood.
- **Lead Add Form**: Leads from direct form submissions show a higher chance of conversion.
- **SMS sent**: Leads who receive an SMS are more likely to convert.
- **Working professionals**: Leads who are working professionals have a higher conversion rate.

Negative Features:

- **Do Not Email - Yes**: Leads who opt out of emails are less likely to convert.

Interpreting Results



Key Takeaways:

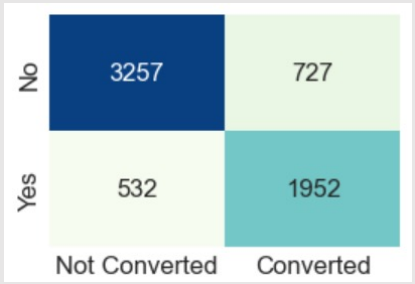
- **Engagement through Phone Conversations and Website Interaction**: The most important factors for conversion are related to lead engagement through time spent on the website and direct communication through phone calls.
- **Effective Channels**: Features like SMS and form submissions play a key role in converting leads, and these should be prioritized by the sales team.
- **Caution with Email Opt-Outs**: Leads who choose not to receive emails are less likely to convert, indicating the importance of maintaining an open line of communication.

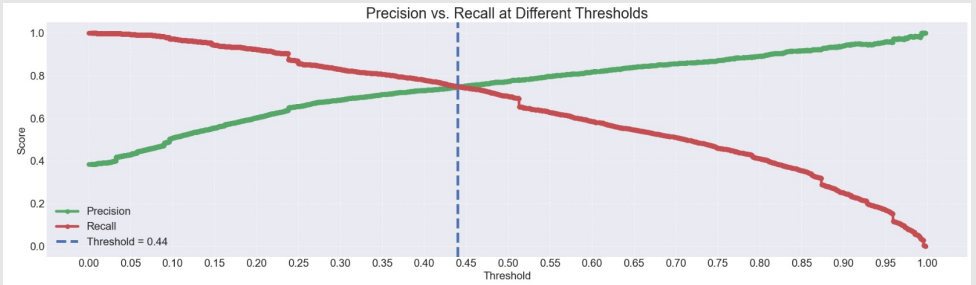
Recommendations and Learnings

- **Target High-Engagement Leads**: Focus on leads that spend more time on the website, originated from the Lead Add Form, and opened emails from X Education.
- **Prioritize Key Variables**: Key factors such as Total Time Spent on Website, Last Notable Activity, and Lead Origin should be prioritized.
- **Iterative Model Improvement**: Regularly update and refine the model based on new data and feedback to enhance accuracy.
- **Strategic Resource Allocation**: Utilize interns and sales teams effectively during peak and off-peak times to optimize lead conversion and minimize unnecessary calls.

Learnings from the data:

- **Importance of Data Cleaning:** Highlighting how meticulously handling categorical variables with missing or placeholder values drastically improved model performance.
- **Precision-Recall Trade-Off**: Would like to emphasizing the need to balance precision and recall, and how this trade-off influences lead prioritization. Illustrating the impact of different thresholds on false positives and false negatives.





Potential Next Steps

- **Advanced Model Testing**:
  - Explore more complex models like decision trees or random forests to potentially improve predictive accuracy.
  - These models can complement the existing logistic regression model.

- **Enhanced Data Collection**:
  - We suggest collecting additional data points that could further refine the model, such as detailed customer interaction histories or demographic information.
  - Recommend integrating data from different touchpoints including lead source, website visit and reactions on calls / sms / emails to create a more holistic view of each lead.

Thank You!