LEAD SCORE CASE STUDY

By:

Tejas Pathak

Gayathry Sadananda Kaimal

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%.

X Education has appointed us to help them select the most promising leads, i.e. the leads that are most likely to convert into paying customers. The company requires us to build a model wherein we need to assign a lead score to each of the leads such that the customers with a higher lead score have a higher conversion chance and the customers with a lower lead score have a lower conversion chance. The CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80%.

Goals of the Case Study

Build a logistic regression model to assign a lead score between 0 and 100 to each of the leads which can be used by the company to target potential leads. A higher score would mean that the lead is hot, i.e. is most likely to convert whereas a lower score would mean that the lead is cold and will mostly not get converted.

There are some more problems presented by the company which your model should be able to adjust to if the company's requirement changes in the future so you will need to handle these as well. These problems are provided in a separate doc file. Please fill it based on the logistic regression model you got in the first step. Also, make sure you include this in your final PPT where you'll make recommendations.

Solution Overview and Techniques

1) Importing Necessary Libraries like pandas, numpy, seaborn, matplotlib and scikitlearn

2)Data Cleaning:

Identify and Remove Duplicates: Scan the dataset for any duplicated rows and eliminate them to avoid skewing the results.

Assess Missing Values: Investigate missing or null entries in the dataset and handle them appropriately based on the context.

Drop Irrelevant Columns: If certain columns have a significant number of missing values and don’t contribute meaningfully to the analysis, consider dropping them to improve model performance.

Impute Missing Data: Where needed, apply imputation techniques to fill in missing values, ensuring data consistency and completeness.

Solution Overview and Techniques

3) Performed EDA and Various Analysis such as Bivariate Analysis

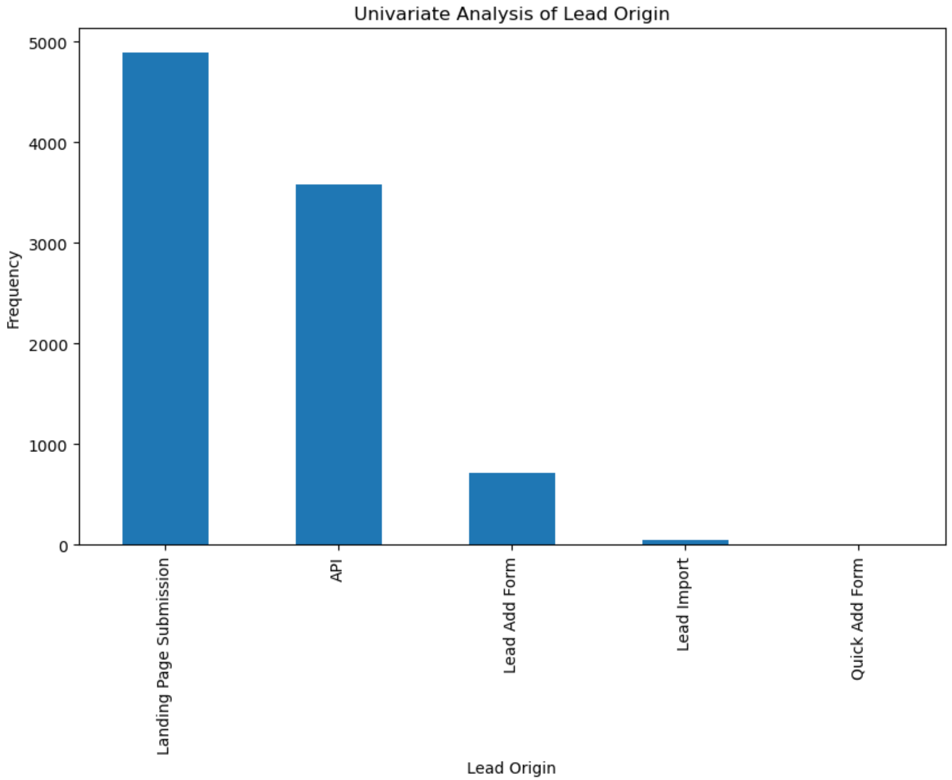
4) Feature Scaling & Dummy Variables plus One hot encoding of the data.

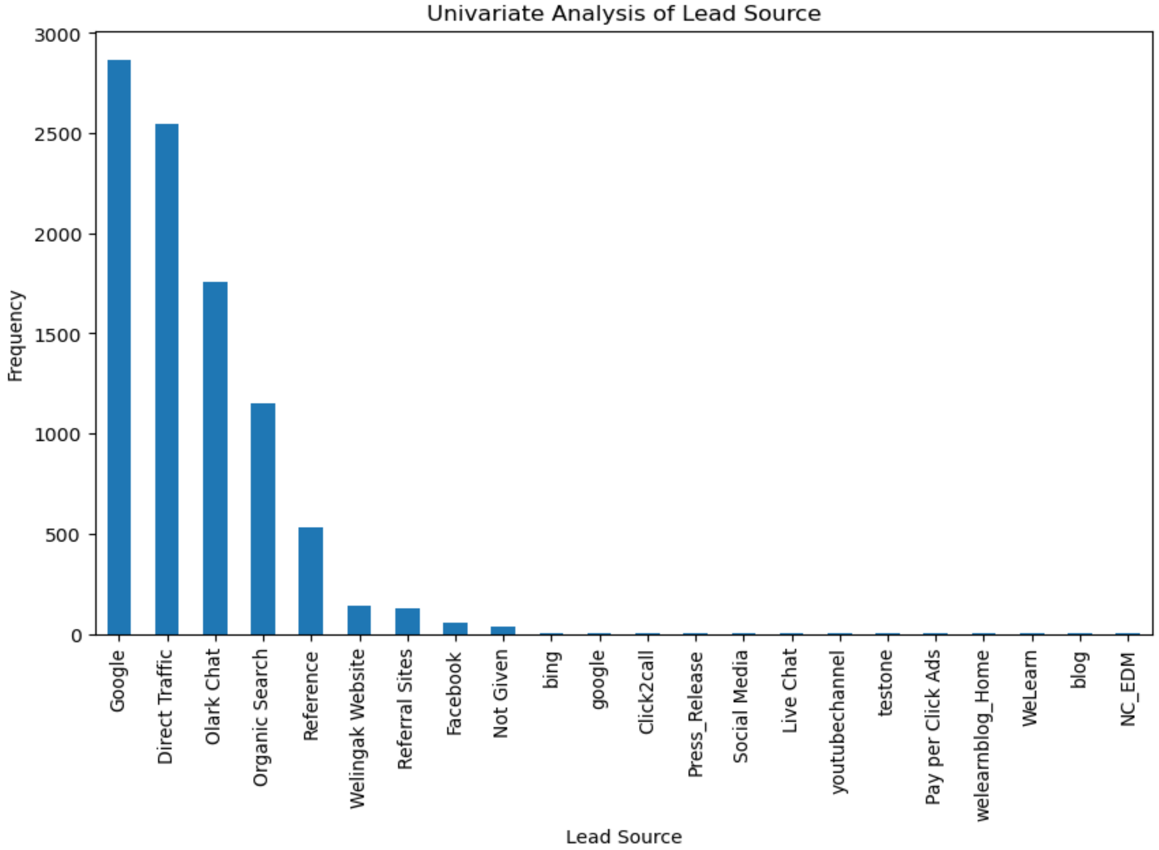
5) Performed Feature Selection

6) Created the model and configured it based on iterative building

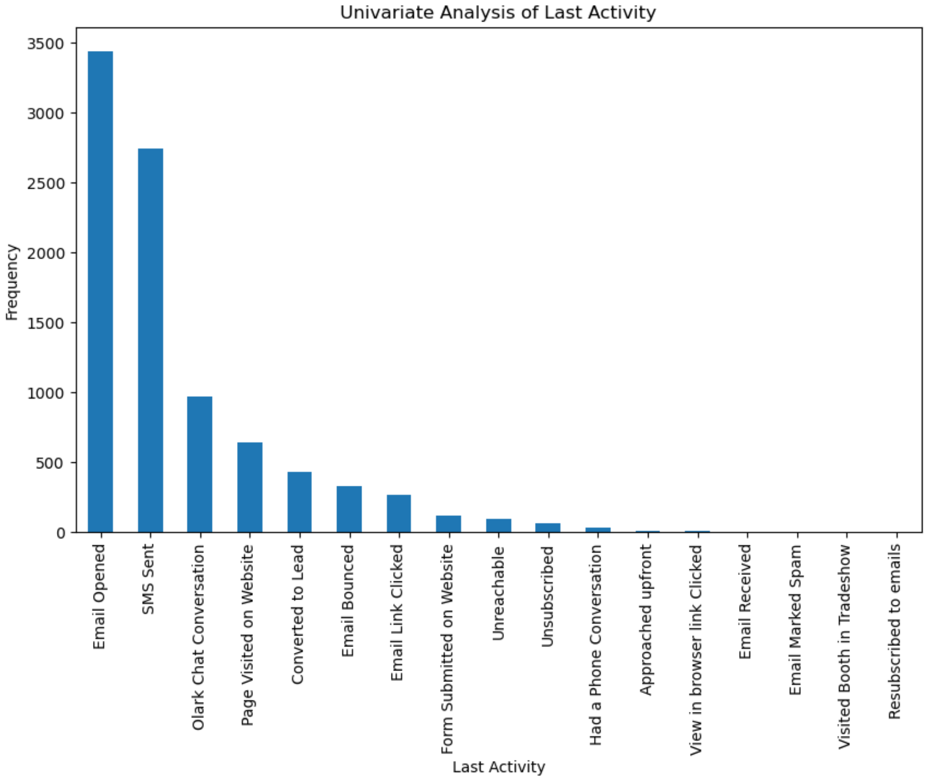
7) Evaluated and Predicted the model

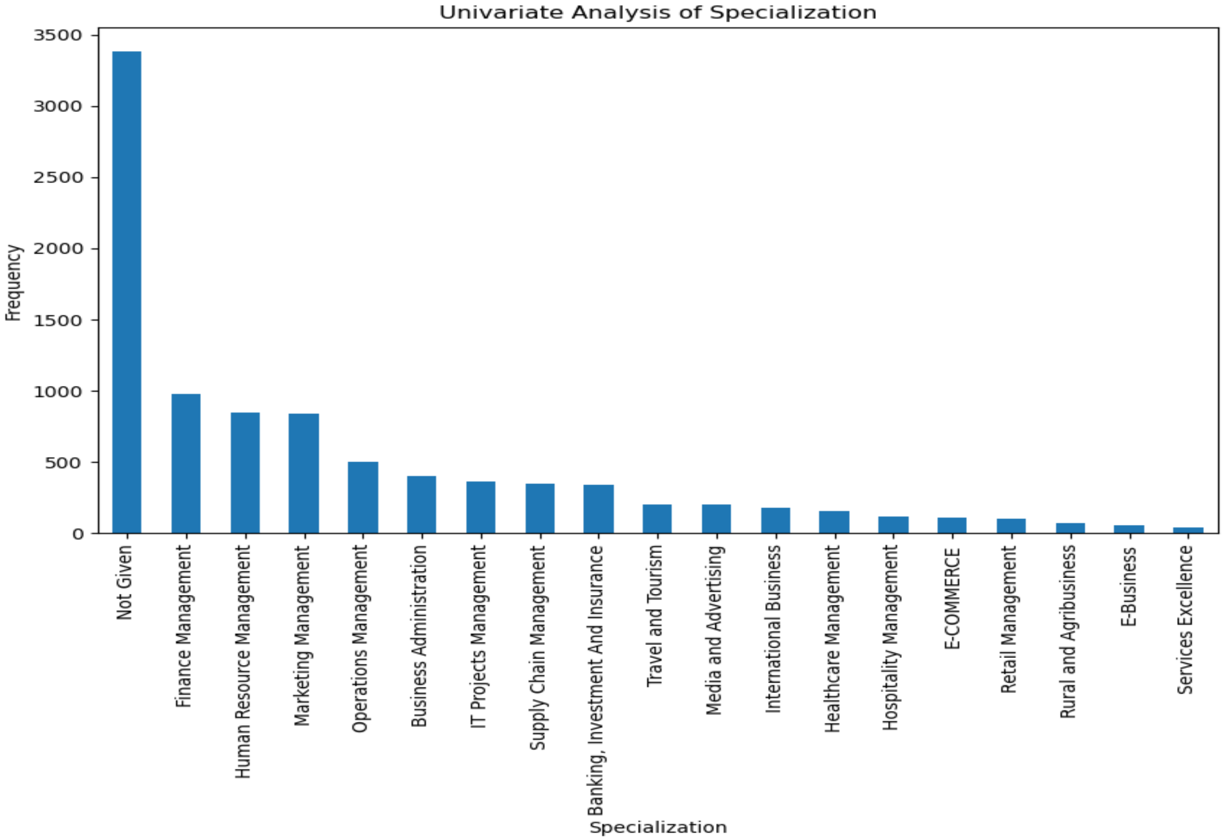
Uni Variate Analysis



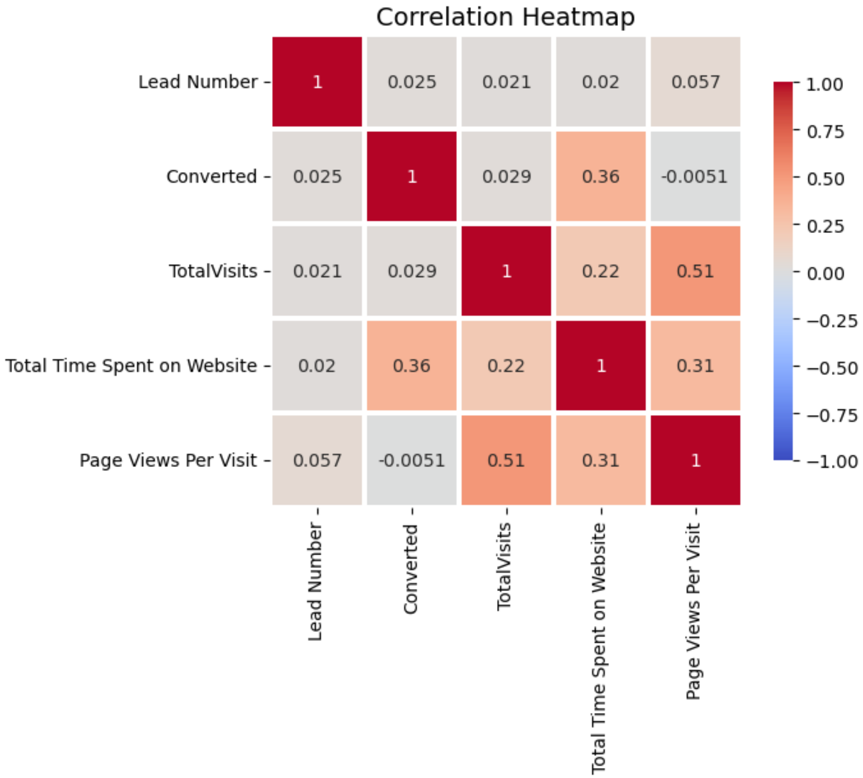


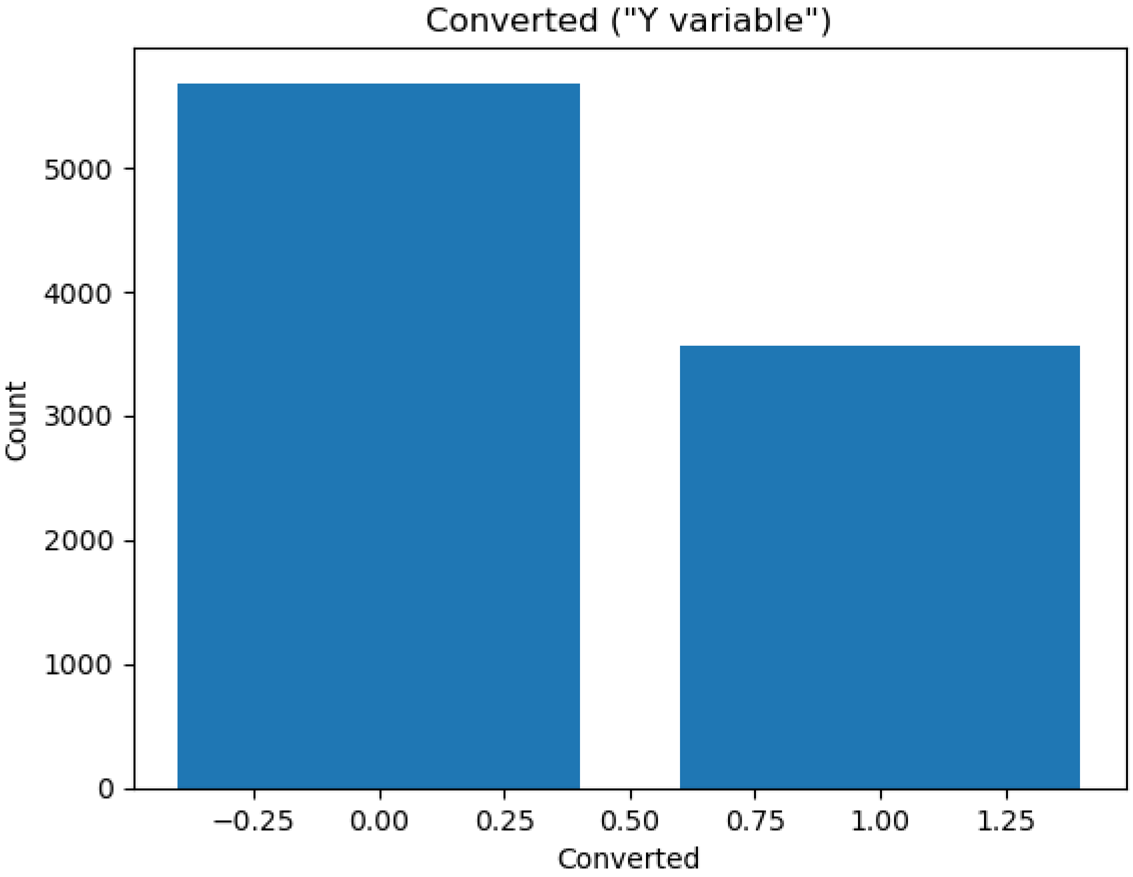
Uni Variate Analysis





Few More Analysis





Feature Selection

- **Recursive Feature Elimination (RFE)**: RFE was employed to select the top 15 important features for lead conversion prediction.
- **Multicollinearity Check**: Variance Inflation Factor (VIF) analysis was used to remove features with high multicollinearity (VIF > 5), ensuring the model was not biased by redundant information.
- **Statistical Significance**: Variables with p-values > 0.05 were removed, leaving only statistically significant features for model building.

Model Building & Evaluation

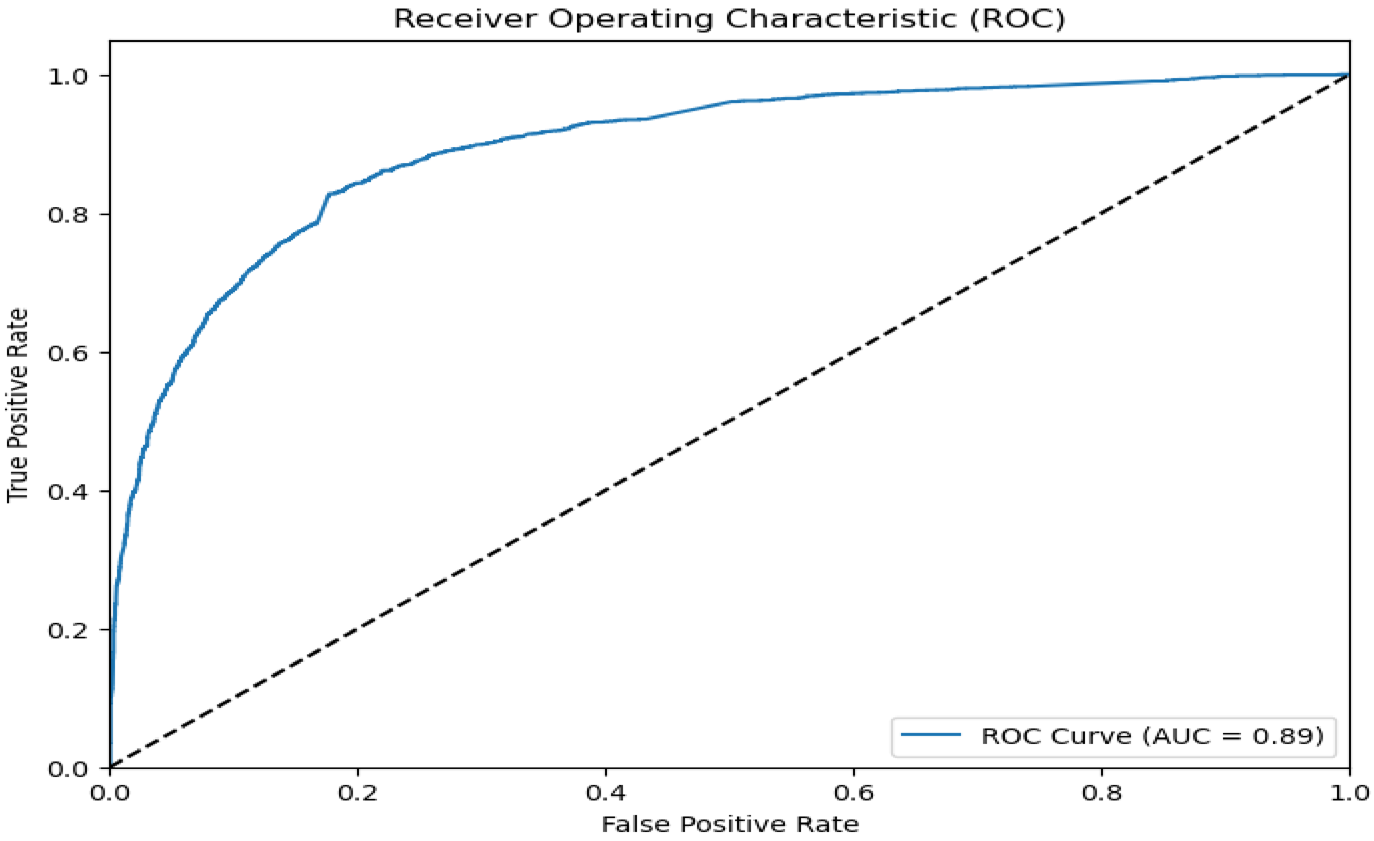
- **Train-Test Split**: The data was split into a 70:30 ratio for training and testing the model, ensuring the model’s ability to generalize.
- **Logistic Regression**: A logistic regression model was built using the selected features. The model was evaluated based on accuracy, sensitivity, and specificity.
- **Confusion Matrix**: To evaluate the model's performance, a confusion matrix was used to calculate accuracy, sensitivity, and specificity, which all achieved around 80%.
- **ROC Curve**: The ROC curve was plotted to determine the optimal probability cutoff. A cutoff value of 0.35 was selected for maximizing the model’s performance in lead conversion prediction.
- **Precision-Recall Curve**: A precision-recall analysis was also conducted, with an optimal cutoff value of 0.41, yielding a precision of 73% and a recall of 75% on the test dataset.

Model Metrics

Model metrics included:

- **Accuracy: 82.03%**
- **Precision: 80.38%**
- **Recall: 69.95%**
- **F1 Score: 74.80%**
- **ROC-AUC Score: 89.46%**

ROC Curve


Receiver Operating Characteristic (ROC)
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.2
0.4
0.6
0.8
1.0
ROC Curve (AUC = 0.89)

Feature Importance

1) Total Spent on Website - For every unit increase in the total time spent on the website, the log-odds of conversion increase by 4.06, meaning that users who spend more time on the website are significantly more likely to convert. This feature is the most impactful because of its high coefficient and extremely low p-value, indicating strong predictive power.

2) Page Views Per Visit - A negative coefficient of -7.15 suggests that higher page views per visit are associated with a lower likelihood of conversion. This could mean that users who browse more pages without converting might be less engaged. The significance and large magnitude of this coefficient make it one of the most important factors.

3) Lead Origin_Lead Add Form - Leads originating from the "Lead Add Form" are 2.7 times more likely to convert, showing a strong association with conversions. This is another key feature due to its high coefficient and significance.

4) Lead Source_Welingak Website - Leads coming from the "Welingak Website" are also more likely to convert, with a coefficient of 2.54, making it an important feature, though less impactful compared to "Lead Add Form."

5) Do Not Email_Yes - Leads who opted out of receiving emails are less likely to convert, with a negative coefficient of -1.15. This shows that email communication plays an important role in conversions.

6) Last Activity_SMS Sent - If the last activity was an SMS sent, the likelihood of conversion increases. This feature is quite important as well, with a moderately high positive coefficient.

7) Lead Profile_Not Given - If the lead profile is not provided, the likelihood of conversion significantly decreases. This suggests that having more detailed information about the lead increases the chances of conversion.

Recommendations for X Education

- **Increase Time Spent on Website**: Focus on content that increases visitor engagement, such as course previews, testimonials, and interactive tools.
- **Optimize Lead Sources**: Continue focusing on high-performing lead sources like Google and Welingak Website, and invest in direct traffic campaigns.
- **Leverage Communication Channels**: Focus on SMS and chat as effective follow-up methods. Be cautious with leads marked as "Do Not Email" as they are less likely to convert.
- **Target Working Professionals**: Marketing strategies should focus on working professionals seeking to upskill, as they have the highest conversion rates.

By focusing on these key factors, X Education can improve its lead conversion rate, making better use of marketing efforts and outreach to potential customers.

Conclusion

Here are the key factors influencing lead conversion, ranked by importance:

1. **Total Time Spent on Website**: More time spent on the site indicates higher engagement and conversion likelihood.
2. **Page Views per Visit**: Higher page views suggest interest in courses, increasing the chances of conversion.
3. **Lead Origin**: Leads from the "Lead Add Form" demonstrate strong intent to convert.
4. **Lead Source**: Conversion rates are highest from:
    1. Welingak Website
    2. Google Search
    3. Direct Traffic
5. **Email Communication**: Lack of email communication negatively impacts conversions.
6. **Last Activity**: SMS or Olark chat as the last activity boosts conversion, along with phone conversations.
7. **Current Occupation**: Working professionals are more likely to convert, indicating career advancement interest.