Lead Score Case Study

by Pritam Chatterjee

Data Understanding and Preprocessing Correlation

First step after receiving data is to clean and analyze the data. Following are the steps.

1

Data Cleaning

The initial step involved understanding the data and performing necessary preprocessing, including dropping irrelevant columns, handling missing values, analyzing the outliers and removing (using boxplot for visualization), and converting categorical features to dummy variables.

2

Exploratory Data Analysis

Correlation analysis, box plots, and bar charts, heat maps, pair plots were used to explore the relationships between features and the target variable (Converted).

Feature Engineering and Selection

Dummy Variables

Dummy variables were created for all categorical columns to prepare the data for the logistic regression model.

Feature Selection

Recursive Feature Elimination (RFE) was used to select the most important features for the logistic regression model. This helped identify the columns with the highest predictive power.

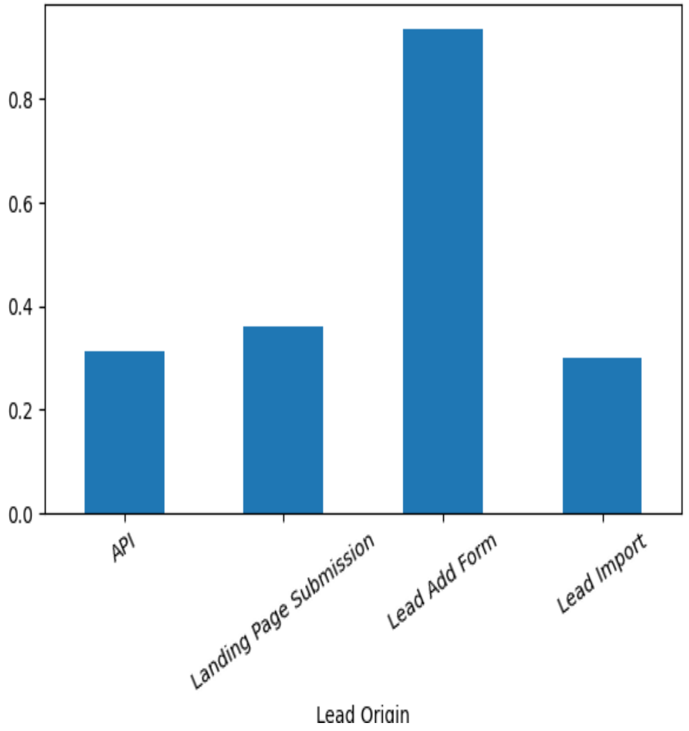
Multicollinearity

The model was further refined by addressing multicollinearity issues, dropping columns with high Variance Inflation Factor (VIF) values and high p-values.

Correlation Between Categorical Features and Convert

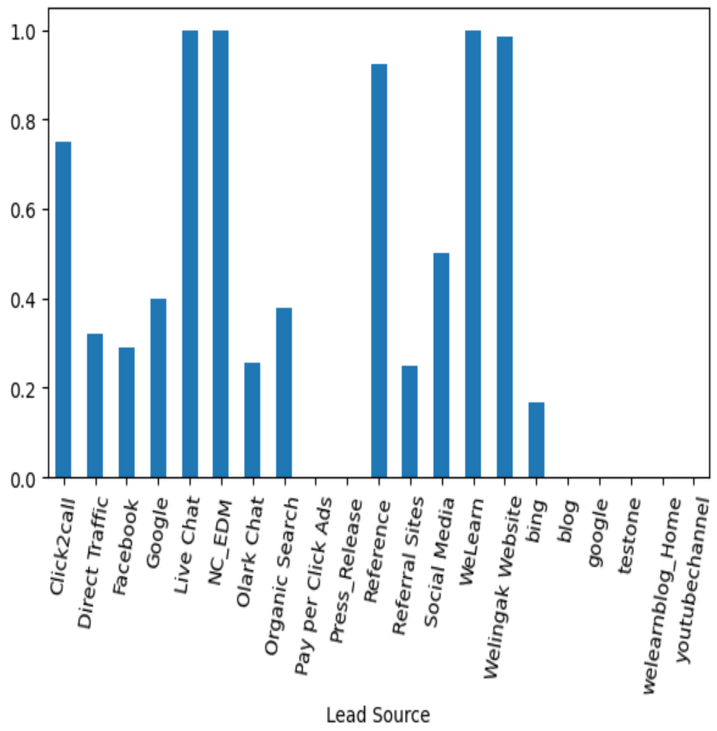
Lead Origin

Leads with different Origin impact the conversion rate as can be seen from chat below that lead added from Form have highest rate of conversion.



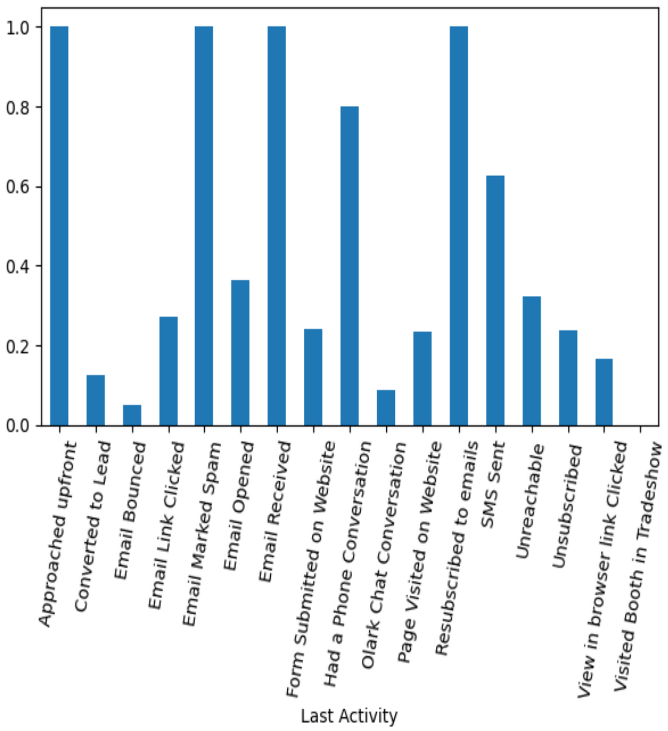
Lead Source

Source of the lead also impact the conversion rates. Leads that were sourced from live chat facility, that have come from references, from we learn and welingak website have high rate of conversions.



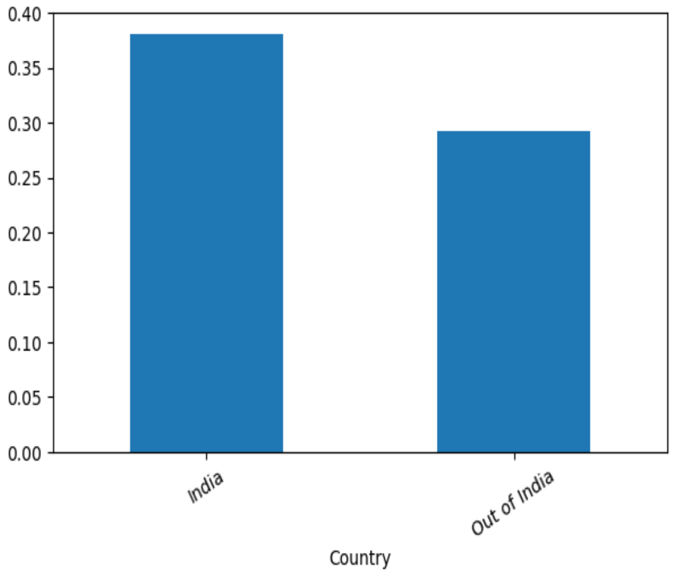
Last Activity

Last activity performed by user also impact the conversion rate. Users that approached upfront, subscribed, received email acknowledgement have high rate of conversion.



Country

The users from India have higher rate of conversion over users from outside India.



Correlation Between Categorical Features and Convert

Occupation

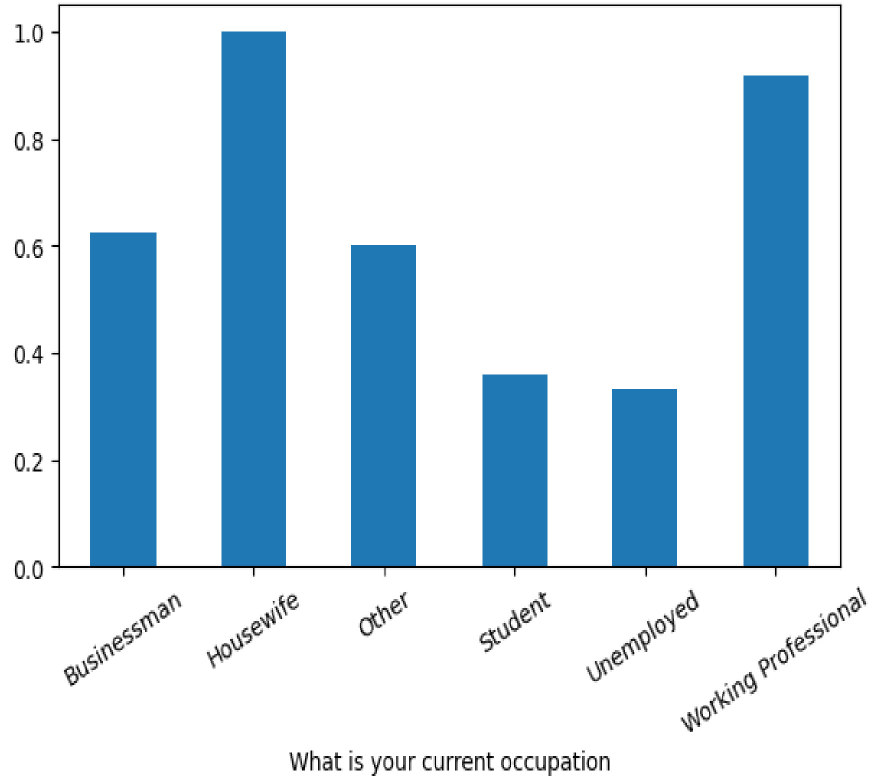
Occupation of user have impact on conversion rate. Working professionals and housewives are highly likely to convert

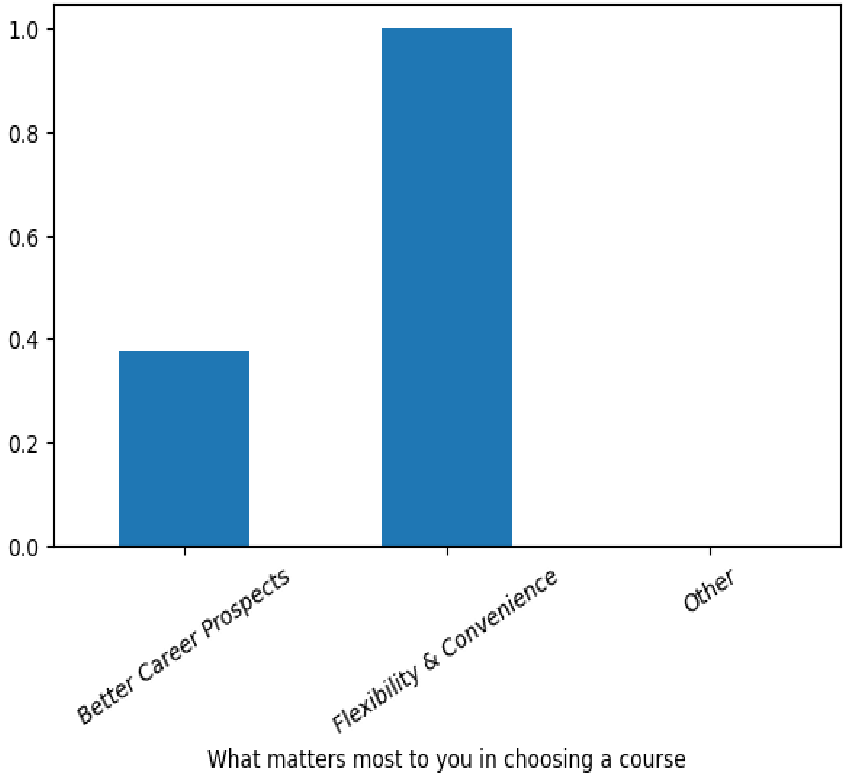
Important factor in Course

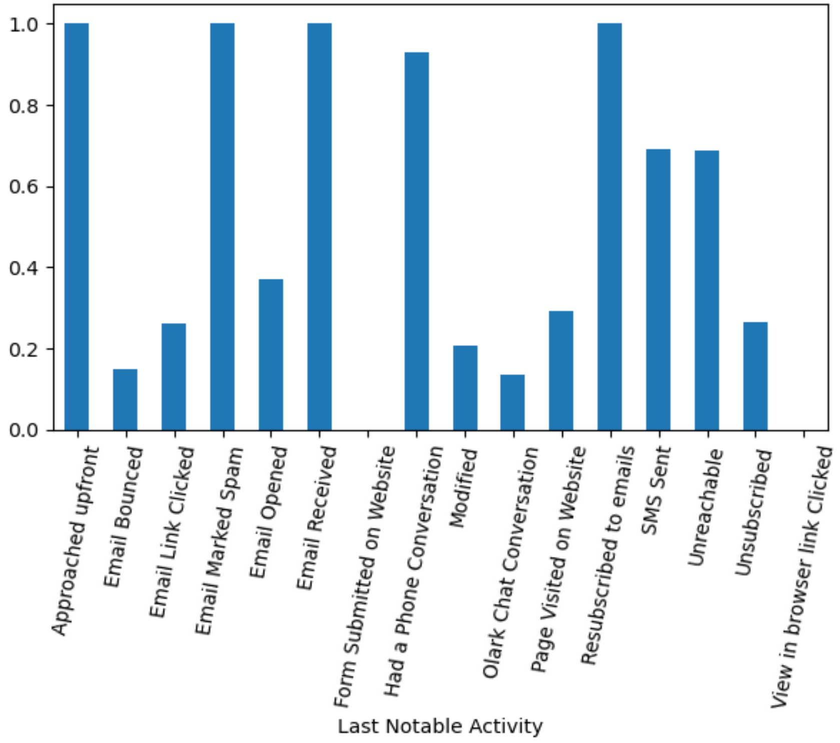
What maters most chosen by user can be a predictor of conversion rate. User who prefer flexibility and comfort have higher rate of conversion.

Specialization

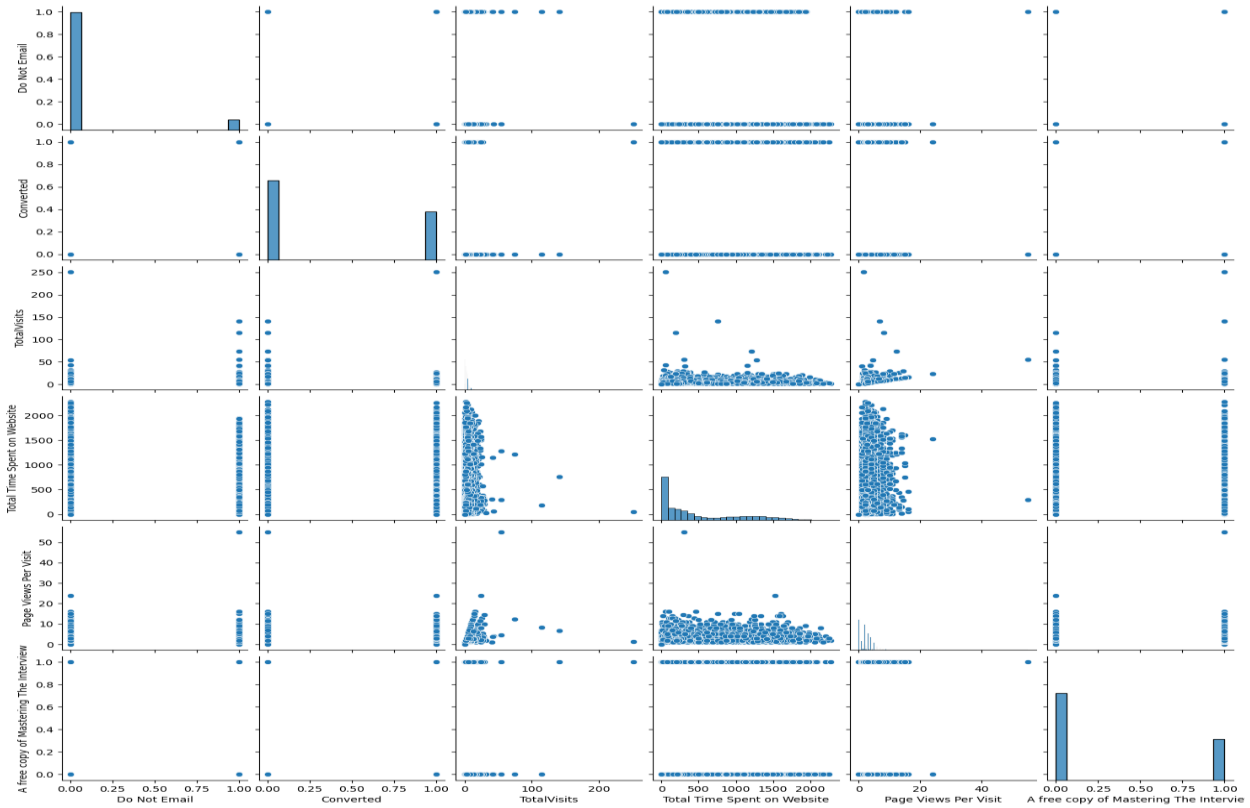
Specialization of the user in professional field also impact the conversion rate. Users that specialize in Banking, healthcare, marketing, operational have high rate of conversion.

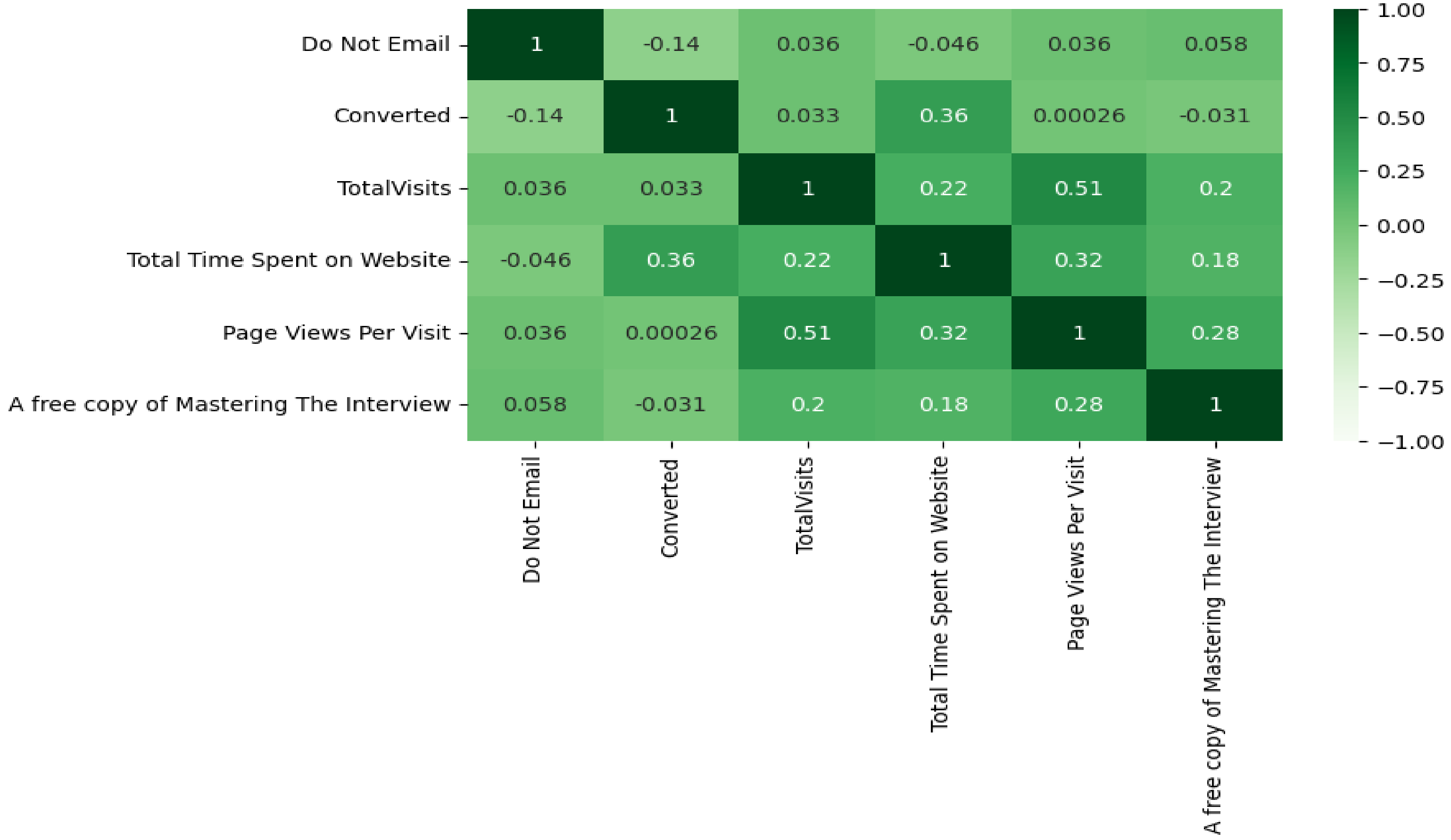






Discovering Numerical Data Patterns





Pair Plots

Pair plots of the numerical columns were generated to discover any relationships and data patterns. Some correlations were observed, which will be further explored in the heatmap analysis.

Correlation Heatmap

The correlation between the numerical columns and the target variable was plotted using a heatmap. This revealed that the target variable is correlated with features such as Do Not Email, Total Visits, Total Time Spent on Website, and A free copy of mastering the interview.

Model Evaluation and Optimization

1

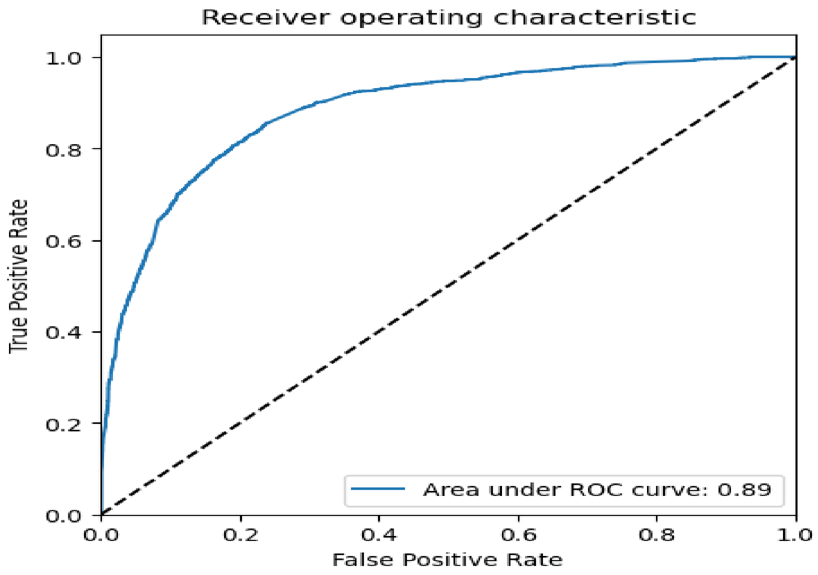
Model Performance

The logistic regression model achieved an overall accuracy of 81% on the training dataset. The sensitivity, specificity, recall, and precision were also evaluated to assess the model's performance.

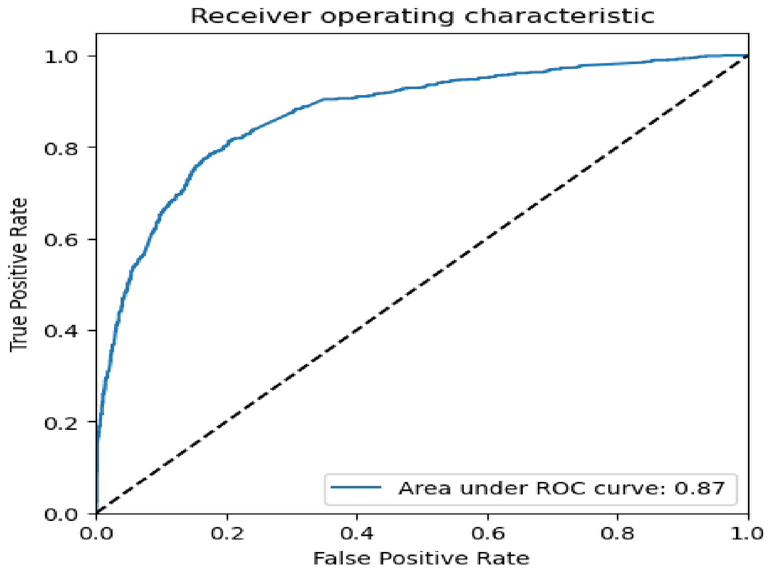
2

ROC Curve

The ROC curve was plotted to visualize the trade-off between sensitivity and specificity. The area under the curve (AUC) was 0.89, indicating a strong predictive power of the model.



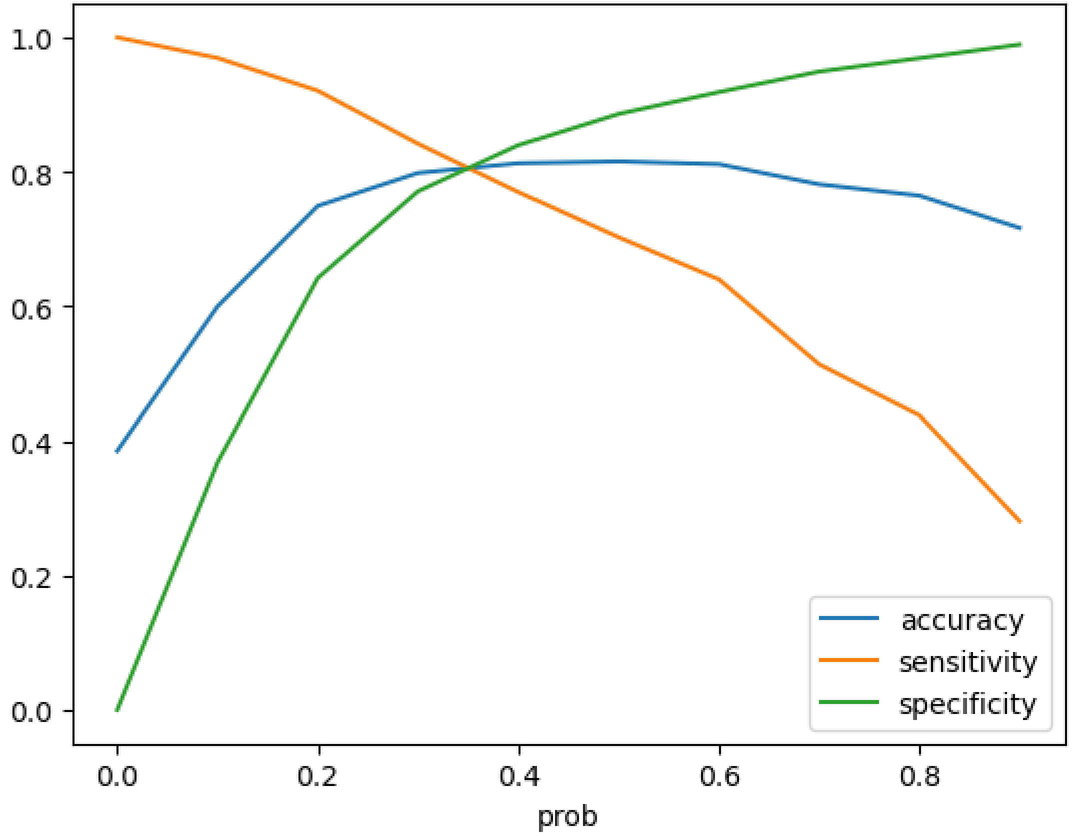
Training ROC



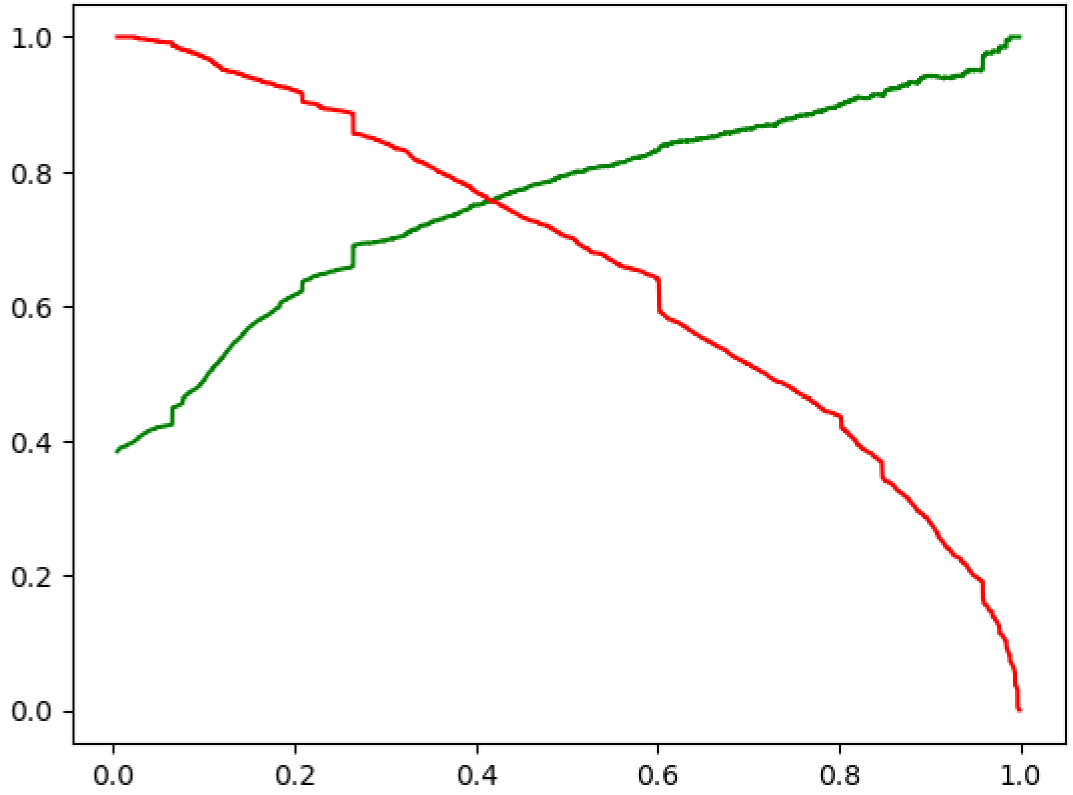
Test ROC

Probability Thresholds

Different probability thresholds were explored to optimize the model's performance, with a 0.4 cutoff identified as the optimal balance between accuracy-sensitivity-specificity and considering the precision-recall curve.



accuracy-sensitivity-specificity



Precision vs Recall

Model Deployment and Scoring

Test Set Evaluation

The model was evaluated on the test dataset, achieving an overall accuracy of 81%, with similar sensitivity (78%), specificity (84%), recall (78%), and precision (75%), AUC ROC (89%) metrics as the training dataset.

Lead Scoring

The model's predicted probabilities were used to calculate a lead score for each lead, ranging from 0 to 100. This score can be used to prioritize and target potential leads effectively and sort the leads based on potential to conversion.

Key Features

The most important features identified by the model include Total Visits, Lead Source, Total Time Spent on Website, Current Occupation, and Last Notable Activity, among others.

Most Important Predictors

Total Visits

The number of total visits to the website is the most important positive factor, indicating that leads with more website engagement are more likely to convert.

Lead Source

The lead source is also a crucial factor, with 'Welingak Website', 'Reference', and 'Live Chat' being the most effective lead sources for conversion.

Time Spent on Website

The total time spent on the website is another important positive factor, suggesting that leads who spend more time exploring the website are more likely to convert.

Last Activity

The last notable activity, such as email link clicks, phone conversations, and website visits, can have a significant impact on the lead's likelihood of conversion.

Key Predictive Features

Positive Influences

Features like TotalVisits, Lead Source - Welingak Website, Total Time Spent on Website, and Lead Source - Reference had strong positive coefficients, indicating their importance in predicting lead conversion.

Negative Influences

Activities and behaviors like Last Notable Activity - Email Link Clicked, Do Not Email, Olark Chat Conversation, and Email Opened had negative coefficients, suggesting they were less likely to lead to conversion.

Adjusting the Model Threshold – Business Requirement

Lower Threshold

Lowering the cutoff probability threshold will classify more leads as "high potential," ensuring that a greater number of leads are flagged for follow-up. This can be used when sales team have higher requirements for more conversions.

Increase Threshold

Increasing the threshold value will reduce the number of conversions but leads will have higher potentials of conversion. This can be used when sales team are busy elsewhere and only high potential leads needs to be followed up.

Leveraging the Lead Score

Lead Score

The lead score (100 to 0) column in the model output is the indicator of how much potential is there for conversion. Higher value indicates higher probability of conversion.

Sales Prioritization

Sales personnel can start from highest lead score and work their way down, targeting leads with the highest potential for conversion.

Key Takeaways

1

Predictive Power

The logistic regression model achieved an accuracy of approximately 81%, demonstrating its ability to effectively identify high-potential leads.

1

Strategic Advantage

The model provides X Education with a clear roadmap to focus their efforts on leads with the highest potential for conversion, while understanding the behaviors that may require intervention.

3

Targeted Outreach

The model's insights on positive and negative influencing factors can guide the sales team's outreach strategies, helping them focus on the most promising leads.

4

Targeted Outreach

The model's insights on positive and negative influencing factors can guide the sales team's outreach strategies, helping them focus on the most promising leads.

5

Threshold Adjustment

Adjusting the cutoff probability threshold can help X Education balance the trade-off between sensitivity and specificity, depending on their sales goals and resource constraints.

Conclusion

Key Takeaways	- The logistic regression model achieved an overall accuracy of 81% on both the training and test datasets. - The most important features identified by the model include Total Visits, Lead Source, Total Time Spent on Website, and Last Notable Activity. - The lead scoring system can be used to prioritize and target potential leads more effectively, with a 0.4 probability cutoff as the optimal balance between sensitivity and specificity.
Next Steps	- Continuously monitor and update the model as new data becomes available to ensure its ongoing effectiveness. - Explore additional features and data sources that may further improve the model's predictive power. - Implement the lead scoring system and track its impact on the organization's lead conversion and sales performance. Adjust Threshold Cut Off value depending on how availability of sales team and aggressiveness of the sales need.

Thank You