Lead scoring Case study using Logistic Regression

Submitted by: Pavithra Sri S

Date: 18.02.2025

Contents

1. Problem statement and Objective
2. Solution approach
3. Dataset overview and Manipulations
4. Exploratory data analysis
5. Model Building
6. ROC curve and optimal cutoff.
7. Evaluation metrices and observation
8. Business recommendations and conclusion

Problem statement

X Education faces a low lead conversion rate of around 30%, despite acquiring a high volume of leads through online marketing and referrals. The sales team currently targets all leads uniformly, leading to inefficiencies. To improve this, the company needs to identify high-potential leads (Hot Leads) to focus sales efforts more effectively and boost conversion rates.

Objective:

- Build a logistic regression model to predict lead conversion probability.
- Assign lead scores (0 to 100) to prioritize high-potential leads.
- Increase overall lead conversion rate to the target of 80%.

Solution approach

Data understanding and Cleaning:

- Imported dataset and checked data types and null values.
- Replaced 'Select' values with nulls and removed high-null columns.
- Grouped low-frequency categories and Dropped unique-valued columns.

Data Preparation and Visualization:

- Analyzed categorical columns using graphs and removed outliers for numerical columns.
- Converted binary categories to 0 and 1. And created dummy variables for better interpretation.

Model Building:

- Split data into train and test sets and standardized values.
- Built logistic regression model and optimized using p-values and VIF with top 15 features selected through RFE.

Model Evaluation and Optimization:

- Predicted on train data with initial cutoff of 0.5.
- Calculated accuracy, sensitivity, and other metrics.
- Used ROC curve to find optimal cutoff and recalculated metrics.

Final Predictions:

- Made predictions on test data and analyzed precision-recall tradeoff.
- Assigned lead scores to prioritize high-potential leads.

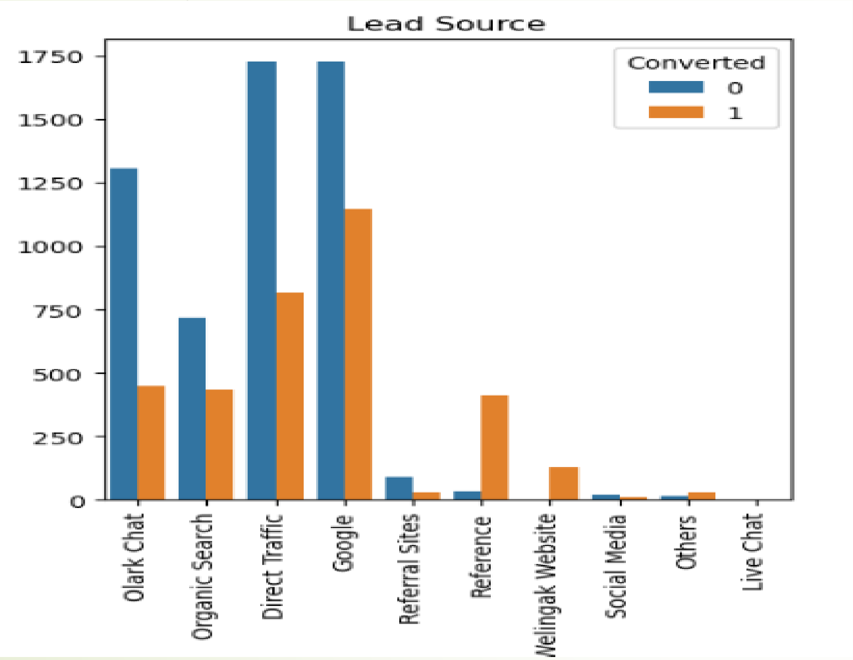
Dataset overview and Manipulation

- Data set has 37 columns and 9240 rows initially with categorical and numeric data.
- Replaced 'Select' values with NULL, as they represent missing data.
- Dropped columns with more than 40% missing values like 'How did you hear about X Education’, 'Lead Quality’, 'Lead Profile’ etc.
- Replaced NaN values with 'Not Provided' or 'Others' to preserve information for the columns like 'Specialization’, 'What is your current occupation’, 'City’, ‘Tags’etc.
- Grouped low-frequency categories under Single category for the columns like ‘Lead source’etc.
- Dropped Prospect ID as it serves the same information as Lead Number.
- Removed unique-valued columns that didn’t add value to the analysis like 'Magazine’, 'Receive More Updates About Our Courses’, 'I agree to pay the amount through cheque’ etc.
- Numerical columns are standardized and outliers are removed.
- Dummy variables are created for easy interpretation and analysis.

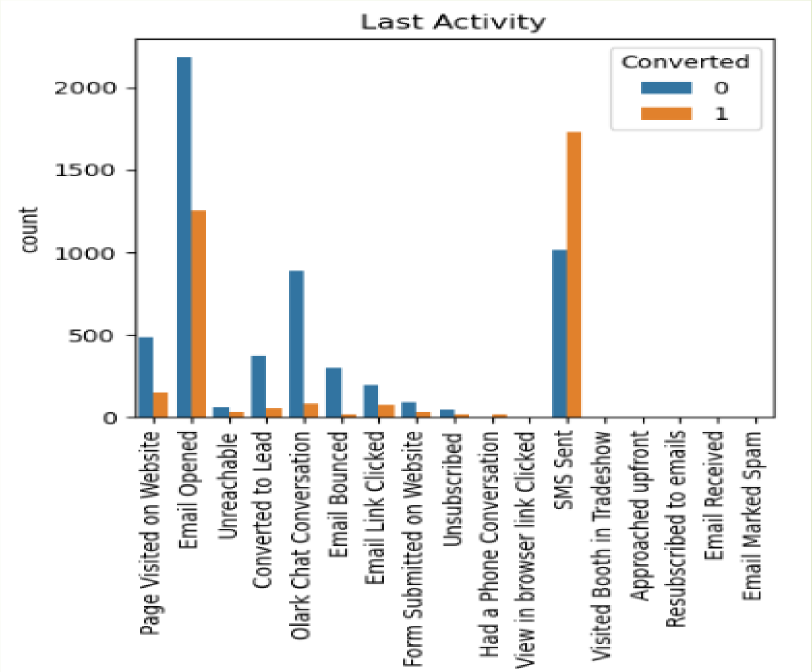
Outcome:

After all the manipulation being done, we retained with 79 columns and 9103 rows for the model building.

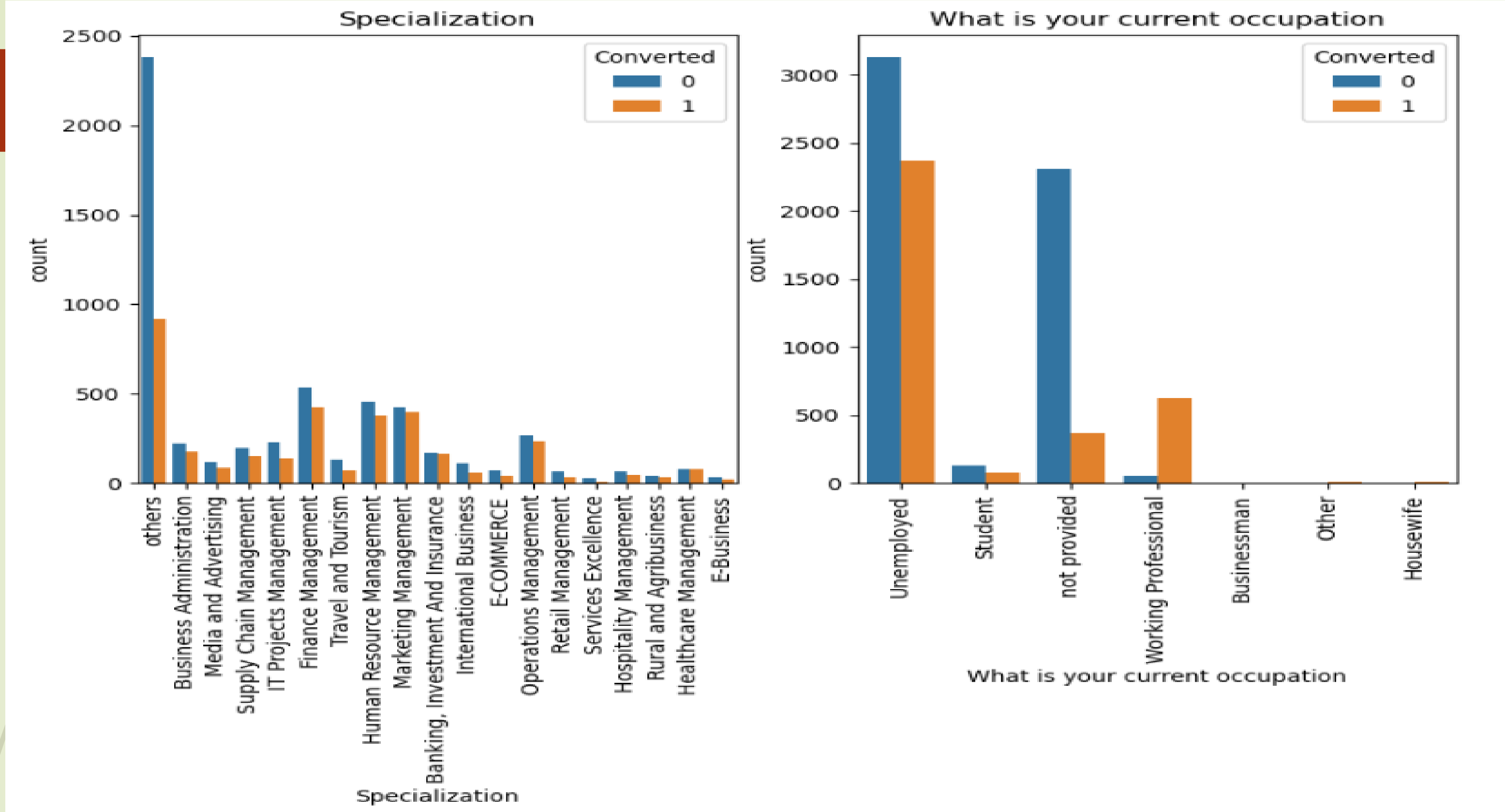
Exploratory Data analysis

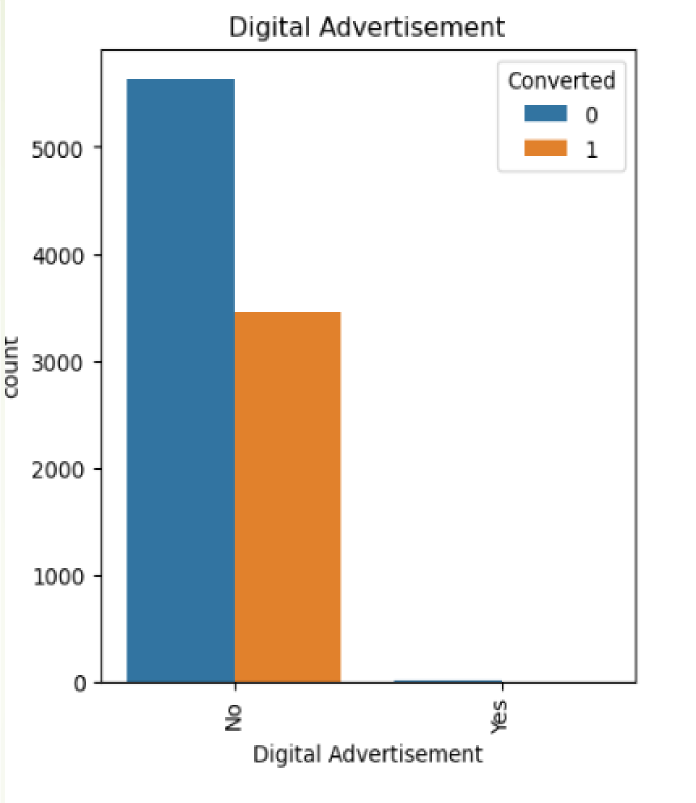


From the graph, Google searches generated more leads, but referrals had the highest conversion rate.



The graph clearly shows that SMS is the most effective method for higher conversion, followed closely by email.





- From a specialization perspective, most conversions lack specific information. However, focusing on Marketing and Finance Management can lead to higher conversions.
- The graph indicates that unemployed individuals showed more interest and a higher conversion rate, while working professionals exhibited a higher overall conversion rate.
- Digital advertisement dosen’t show any promising lead conversion.

Model Building

•Split the dataset into training (70%) and testing (30%) sets for model evaluation.

•Scaled features using MinMax Scaler to normalize the values and ensure consistent scaling across variables.

•Applied Recursive Feature Elimination (RFE) to select the top 15 most relevant features, reducing overfitting.

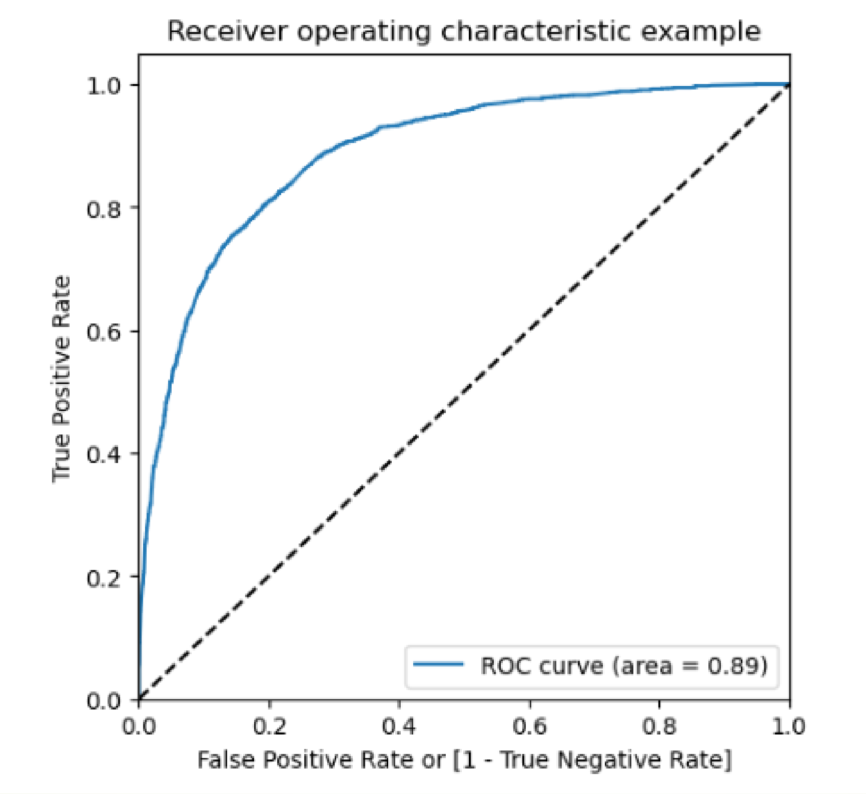
•Built the logistic regression model by removing features with a p-value > 0.05 and VIF > 5 to ensure stability and avoid multicollinearity.

•Made predictions on both the training and testing datasets to evaluate model performance.

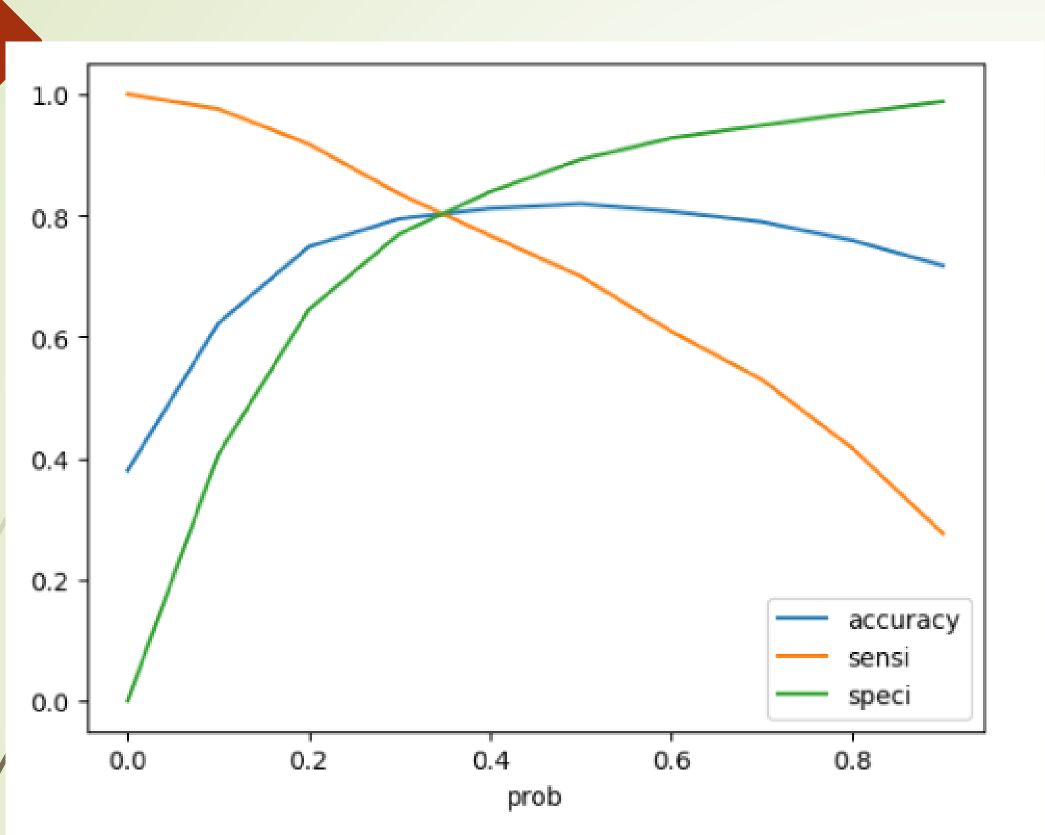
•Used the ROC curve to determine the optimal cutoff probability and adjusted the threshold for better conversion predictions.

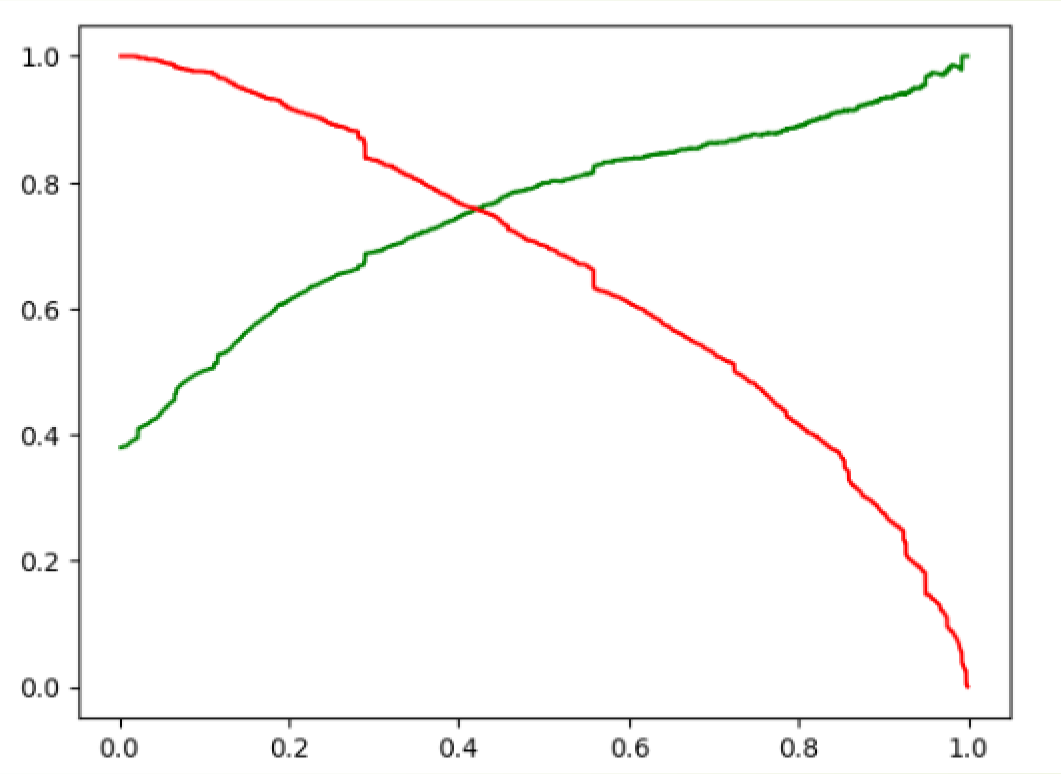
•Evaluated the model using accuracy, sensitivity, specificity, and other relevant metrics to assess its effectiveness in predicting lead conversions.

ROC curve and Optimal cutoff



The ROC curve has an area of 0.89, indicating a strong model performance with good classification ability.





From the above graphs, a tradeoff of 0.38 between Precision and Recall is observed, which we have identified as the optimal cutoff. This means that any Prospect Lead with a Conversion Probability greater than 38% can be confidently classified as a "Hot Lead." This cutoff allows us to effectively prioritize leads that have a higher chance of conversion, improving the efficiency of the sales team.

Evaluation metrices and observation

Evaluation metrices	Train data	Test data
Accuracy	80.8 %	81.6 %
Sensitivity	78.2 %	80 %
Specificity	82.4 %	82.7 %
Precision	73.2 %	74 %
Recall	78.2 %	80 %

Thus we have achieved our goal of getting a ballpark of the target lead conversion rate to be around 80% . The Model seems to predict the Conversion Rate very well and we should be able to give the CEO confidence in making good calls based on this model to get a higher lead conversion rate of 80%.

With the provided data set, there are 435 leads which can be contacted and have a high chance of getting converted.

Business recommendations and conclusion

- **Prioritize High-Scoring Leads:** Leads with a score above 70 should be given higher priority for sales follow-ups to improve conversion rates.

- **Enhance Website Engagement:** Optimize user experience and interaction to increase conversions, especially focusing on increasing time spent on the website.

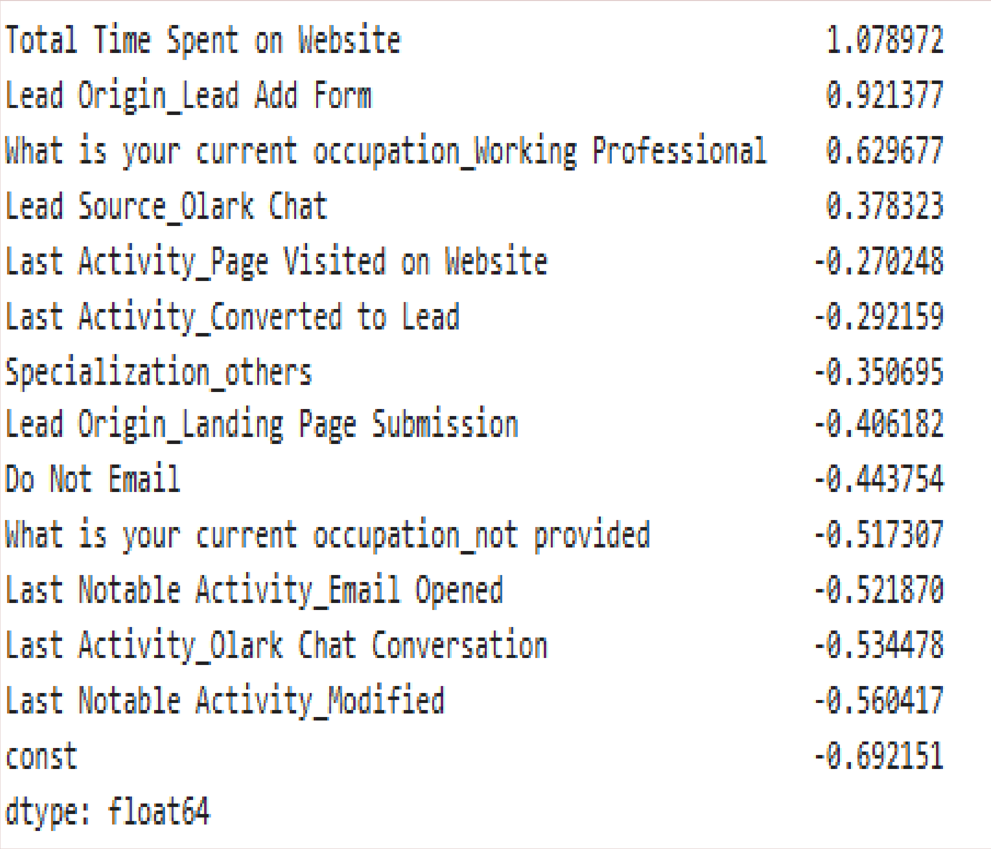
- **Improve Lead Tracking:** Utilize insights from last activity trends to refine follow-up strategies and ensure timely engagement.

- **Focus on High-Performing Acquisition Channels:** Channels such as the Welingak Website and Olark Chat have shown strong correlations with conversions. Invest in these channels to maximize lead quality.

- **Ensure Model Scalability:** Adapt the model to evolving business requirements and lead acquisition strategies, ensuring long-term effectiveness.

Finding out the Important Features from our final model

13 important features to be considered



```
Total Time Spent on Website                            1.078972
Lead Origin_Lead Add Form                              0.921377
What is your current occupation_Working Professional   0.629677
Lead Source_Olark Chat                                 0.378323
Last Activity_Page Visited on Website                 -0.270248
Last Activity_Converted to Lead                       -0.292159
Specialization_others                                 -0.350695
Lead Origin_Landing Page Submission                   -0.406182
Do Not Email                                          -0.443754
What is your current occupation_not provided          -0.517307
Last Notable Activity_Email Opened                    -0.521870
Last Activity_Olark Chat Conversation                 -0.534478
Last Notable Activity_Modified                        -0.560417
const                                                 -0.692151
dtype: float64
```

- Top three variables in model which contribute most towards the probability of a lead getting converted are:-

Total Time Spent on Website
Lead Origin
What is your current occupation

- Leads came from “Reference “and “Welingak website “ should be approached through phone call.
- Leads who had “Add form” should be approached through phone call.
- Leads whose Last activity and Last Notable Activity is “SMS sent” should be approached.
- Leads whose current occupation is “Working Professional” should be approached.
- Leads with total high “Time spent on Website” should be approached.

Thank you !!!