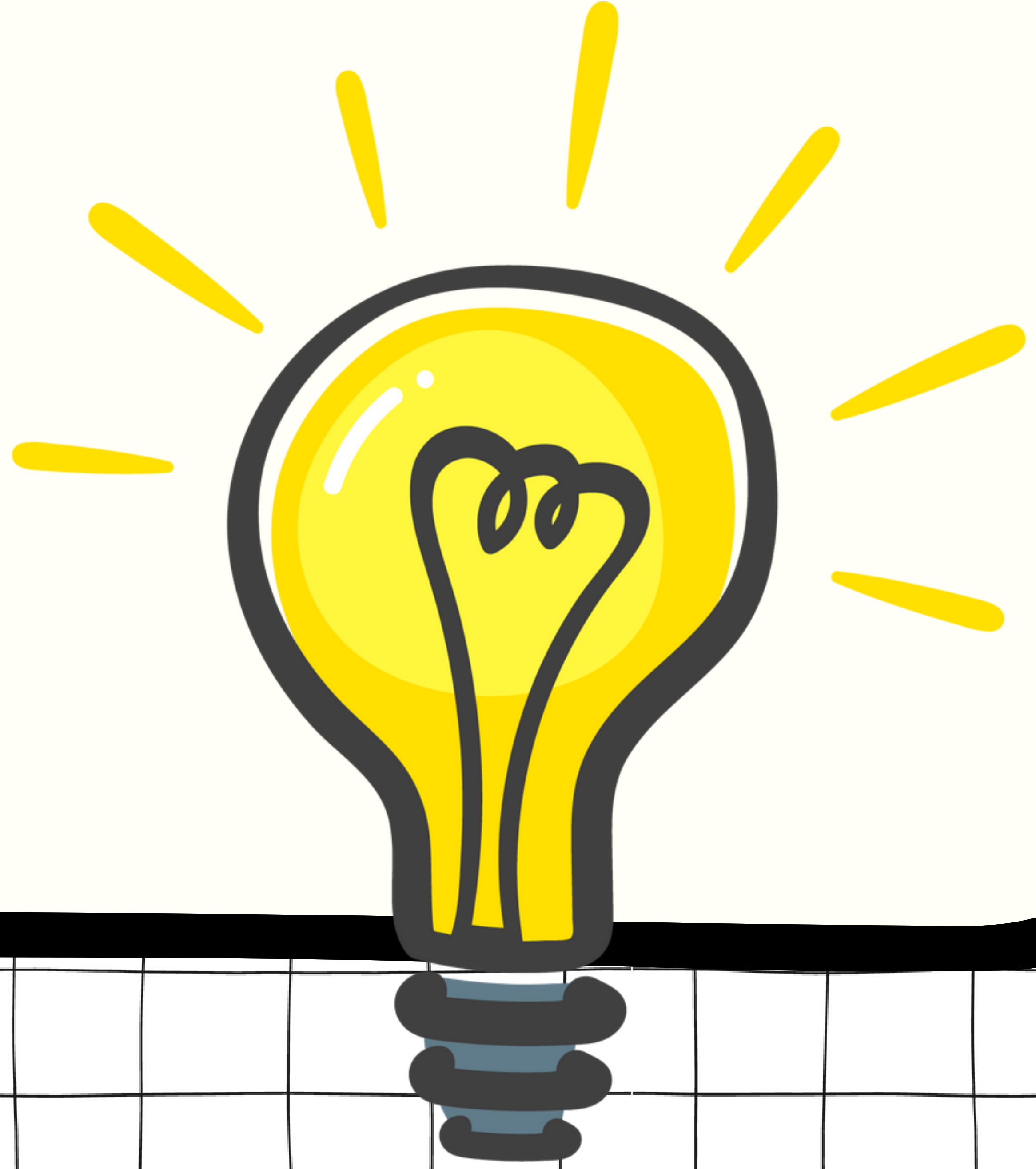
Lead Scoring

Abhishek Singh Dhadwal



Contents

01 Project Description

02 Key Questions

03 Findings and Insights

04 Summary

05 Actions and recommendations

Project Description

X Education, an online education company catering to industry professionals, faces a significant challenge with its low lead conversion rate, currently standing at 30%. They attract potential customers through various online channels, including their website and referrals. To enhance their conversion rate, the company seeks to identify 'Hot Leads', those with the highest potential for conversion into paying customers.

By assigning lead scores to each prospect, the aim is to prioritize interactions with leads who are more likely to convert, thereby increasing the overall conversion rate towards the CEO's ambitious target of 80%. This initiative involves nurturing potential leads and improving the efficiency of the lead conversion process, ultimately resulting in a more effective and profitable customer acquisition strategy.

Key Questions

What is the current ratio of successfully converted leads to the rest of the population?

What is the relationship between the sources and origins of leads when compared to their success?

Do demographic factors like location, and methods of contact play into the likelihood of conversion?

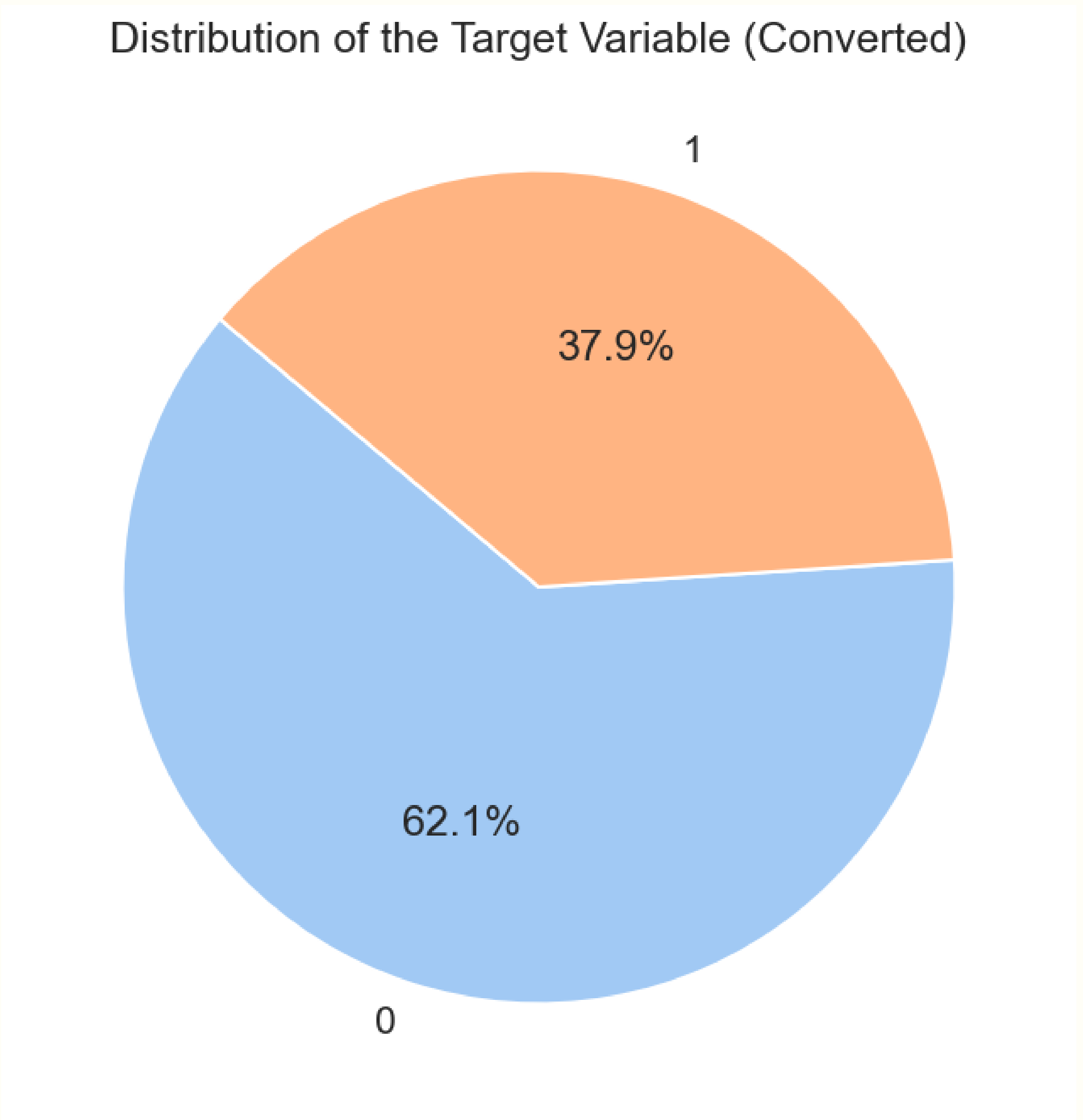
What are the drawbacks of the current methods of storing user data when looking into leads?

How can a model help with the determination of success for any potential customer?

What do stochastic methods of determination state about the deciding factors of “hot leads”?

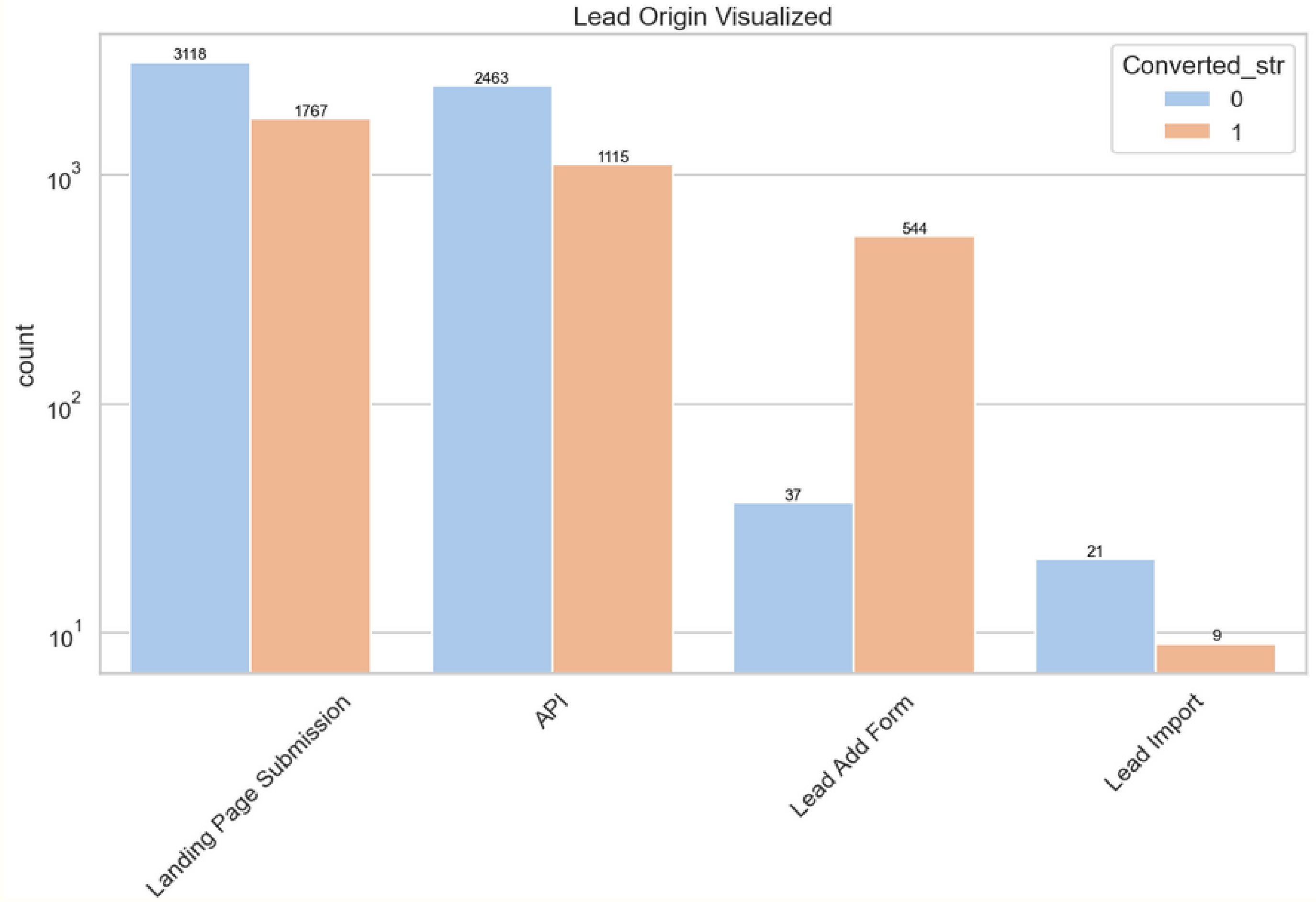
Findings and Insights

How many leads are being converted in the current scenario?

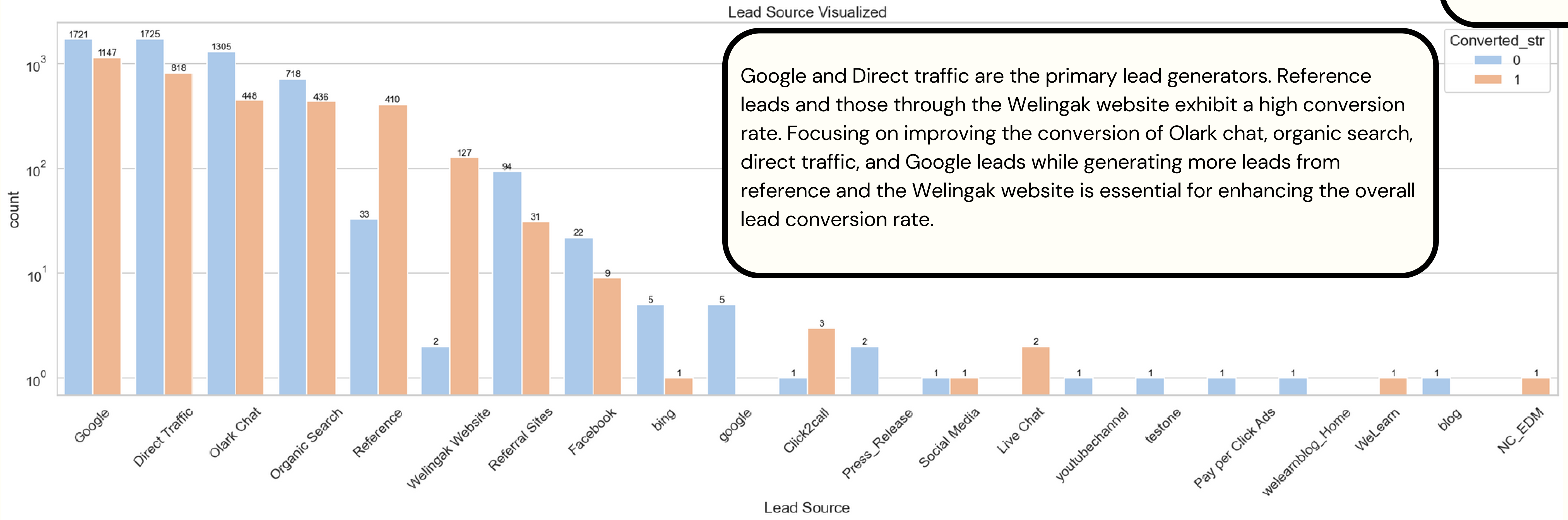


- About 4 in every 10 candidates are being converted in the current methodology, which are lower than our expectations!

What is the relationship between the sources and origins of leads when compared to their success?



API and Landing Page Submission have a 30-35% conversion rate with a substantial lead count. Lead Add Form boasts a conversion rate of over 90%, though the lead count is lower. Lead Import contributes very few leads. To enhance the overall lead conversion rate, efforts should prioritize improving the conversion of leads from API and Landing Page Submission and generating more leads from Lead Add Form.



Google and Direct traffic are the primary lead generators. Reference leads and those through the Welingak website exhibit a high conversion rate. Focusing on improving the conversion of Olark chat, organic search, direct traffic, and Google leads while generating more leads from reference and the Welingak website is essential for enhancing the overall lead conversion rate.

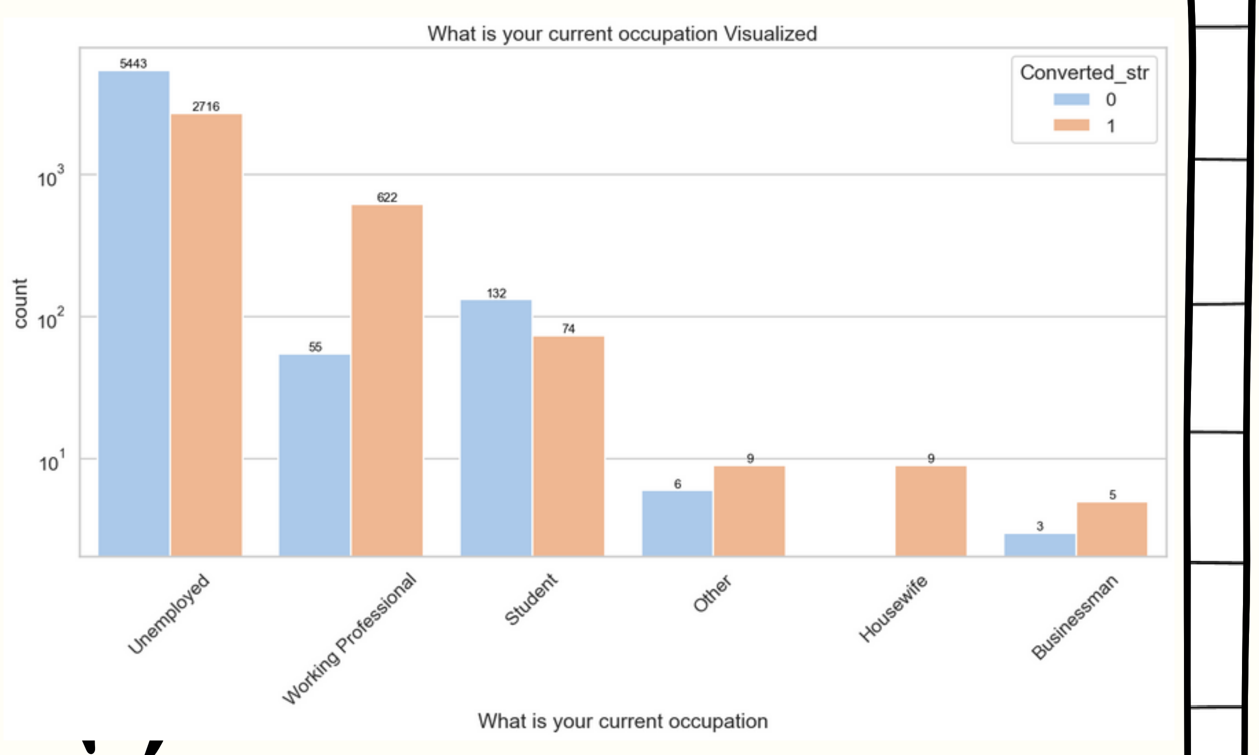
Do demographic factors like location, and methods of contact play into the likelihood of conversion?

Insights

A majority of leads are:

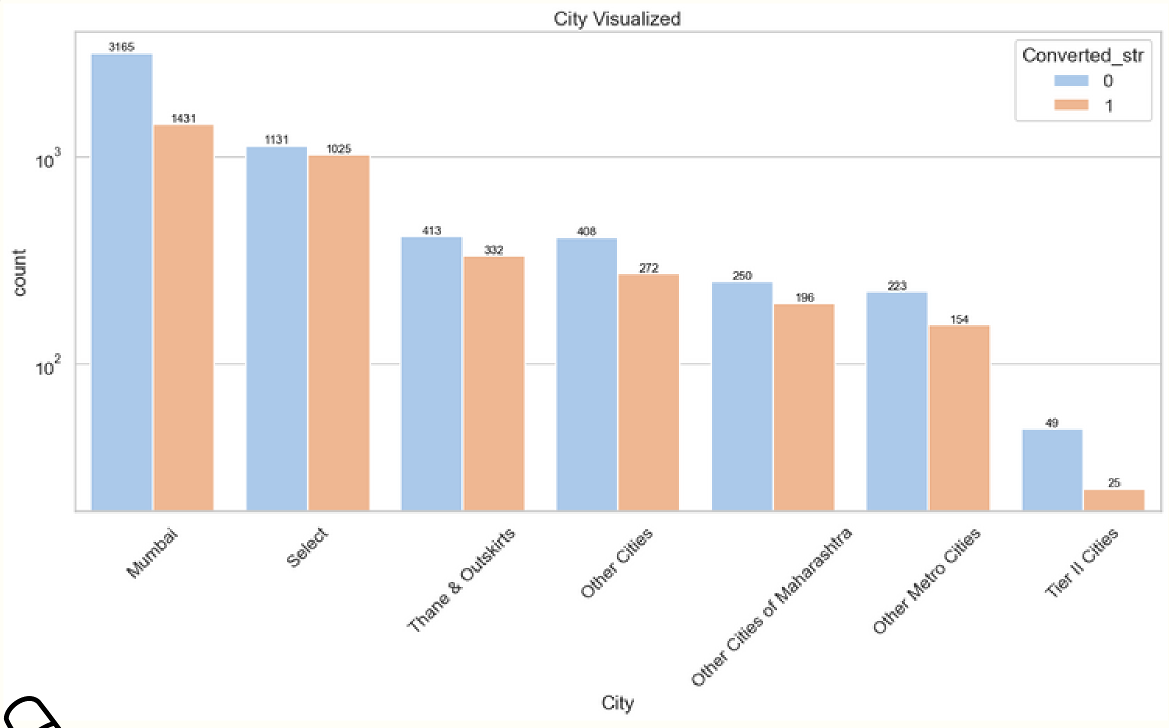
1. From Mumbai
2. Unemployed
3. Not interested in the free textbook

Occupations of leads



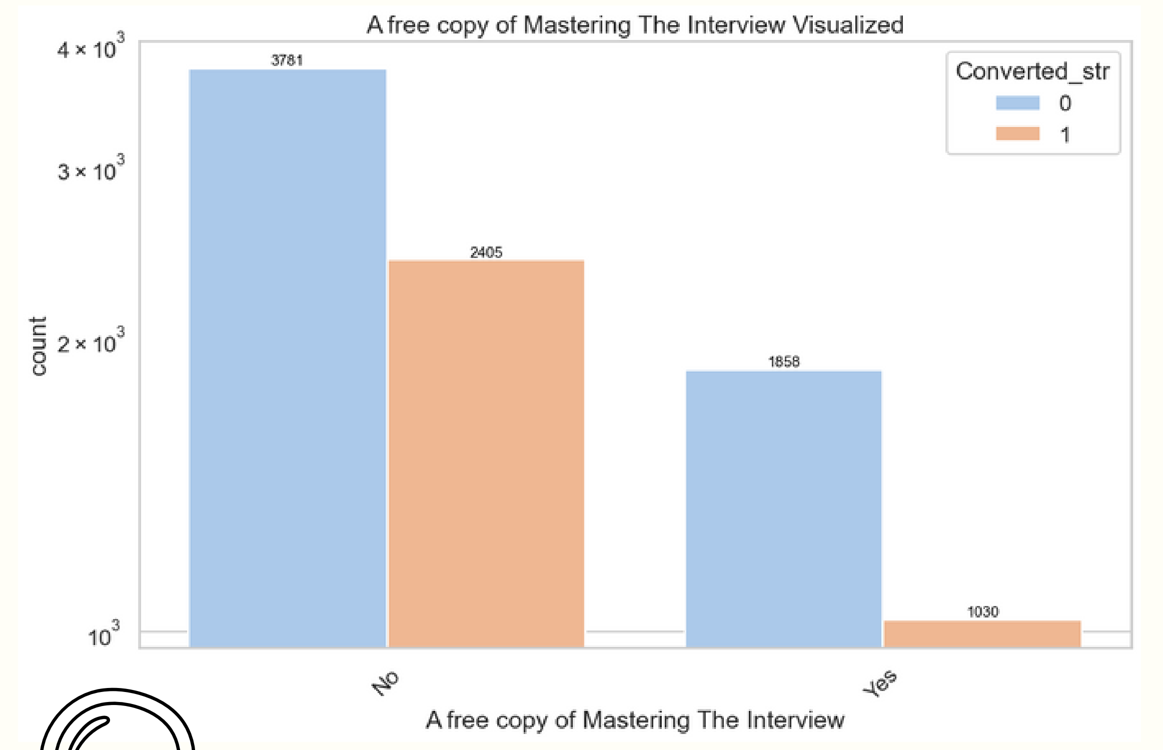
Working professionals opting for the course have a high likelihood of joining it. Unemployed leads are the most numerous but maintain a 30-35% conversion rate.

City Distribution



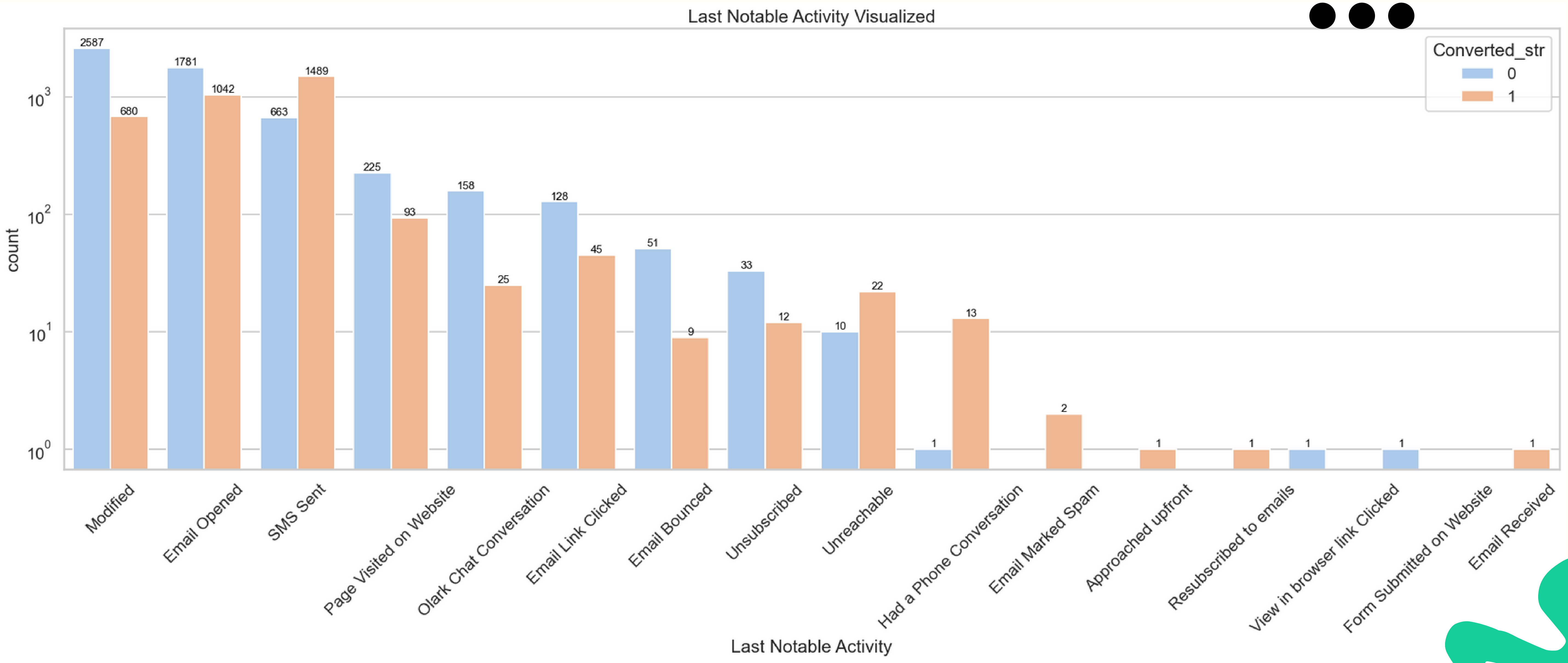
The majority of leads are from Mumbai, boasting a conversion rate of around 50%.

Last activity and interaction



A majority of other factors do not seem to provide an impact into the success of leads.

What last notable activities influence the conversion of a lead?



01 **Last Activity:** Most leads have their last activity as "Email opened." Leads with their last activity as "SMS Sent" achieve an impressive 60% conversion rate.

02 **Total Time Spent on Website:** Leads spending more time on the weblise are more likely to be converted.

Preparing our data for the logistic regression model

Now that we have made variables of note, we can begin making our model for the company

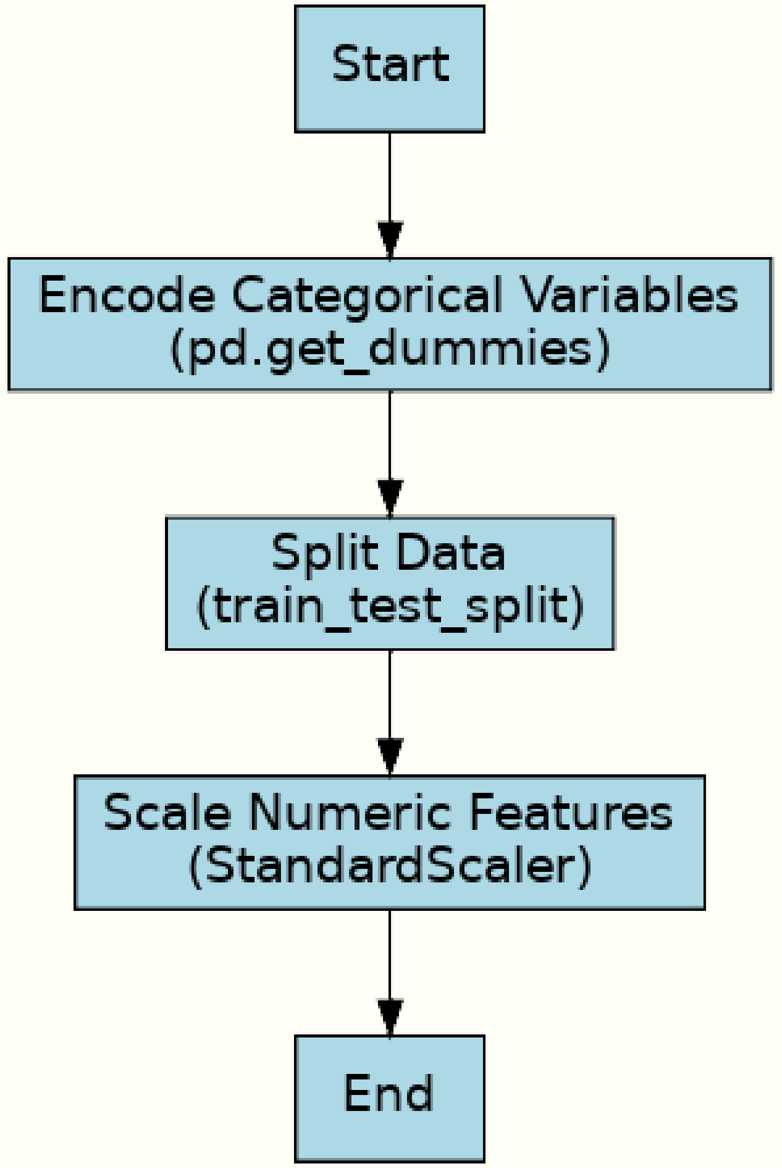
Actions taken for processing columns of the dataset

Removing and aggregating columns: Columns with more than 30% missing values have been removed. Post which many categorical outputs from columns have been combined., for eg. Replacing less frequent Last Activities with 'Other_Activity'.

Feature Engineering: Visits_PageViews has been made as a combination of pre-existing features to provide variation in input data.

Pipeline to process data: A pipeline has been generated to process numerical and categorical columns in preparation for the dataset.

Flowchart for processing columns of the dataset



Start

Encode Categorical Variables
(pd.get_dummies)

Split Data
(train_test_split)

Scale Numeric Features
(StandardScaler)

End

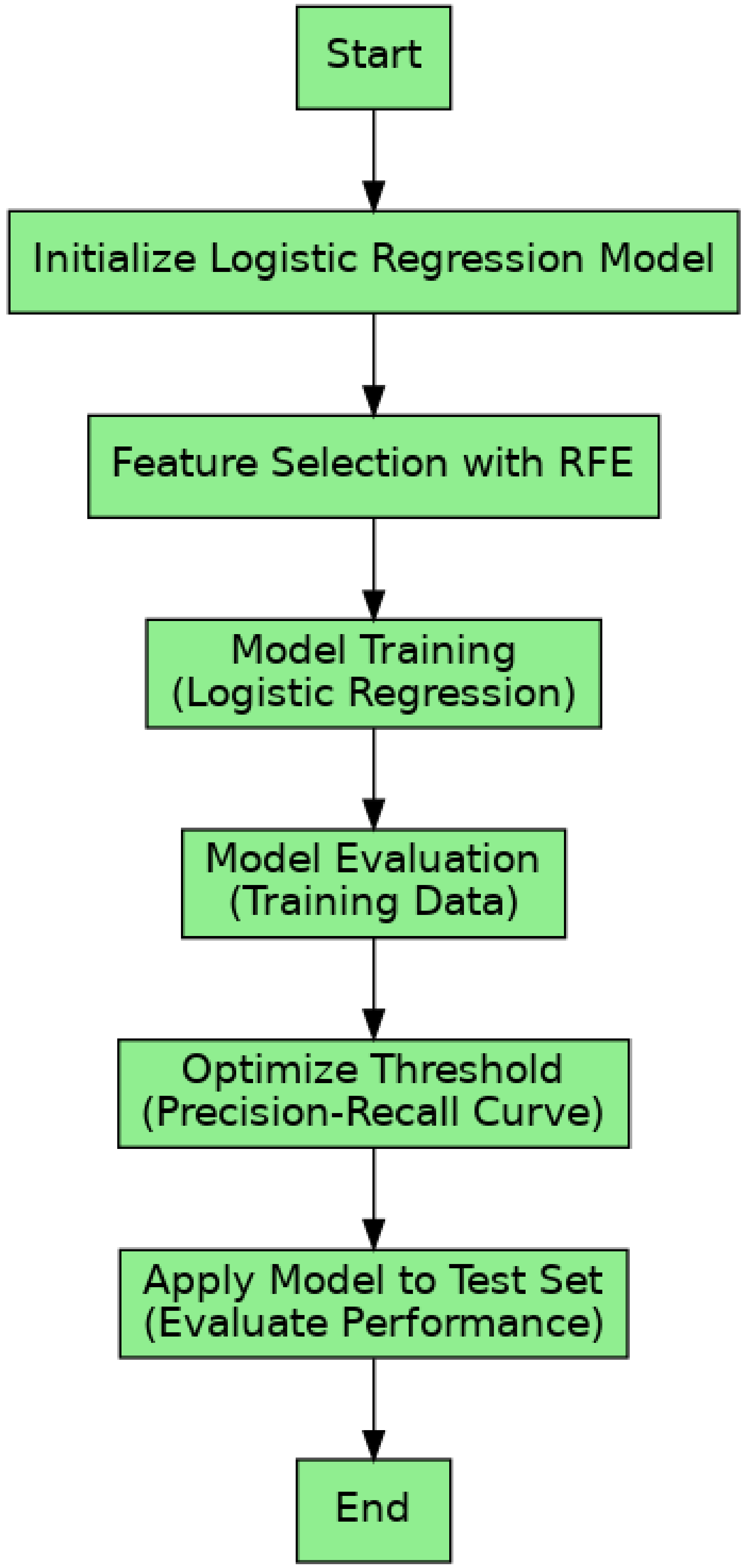
The Model Creation Methodology

Generating LR entities with our data

01 **Systematic Approach to Model Development:**
The methodology begins with initializing a logistic regression model, a choice driven by the binary nature of the lead conversion prediction. This step is followed by a strategic feature selection using Recursive Feature Elimination (RFE), ensuring that only the most relevant predictors are retained, thereby optimizing the model's complexity and performance.

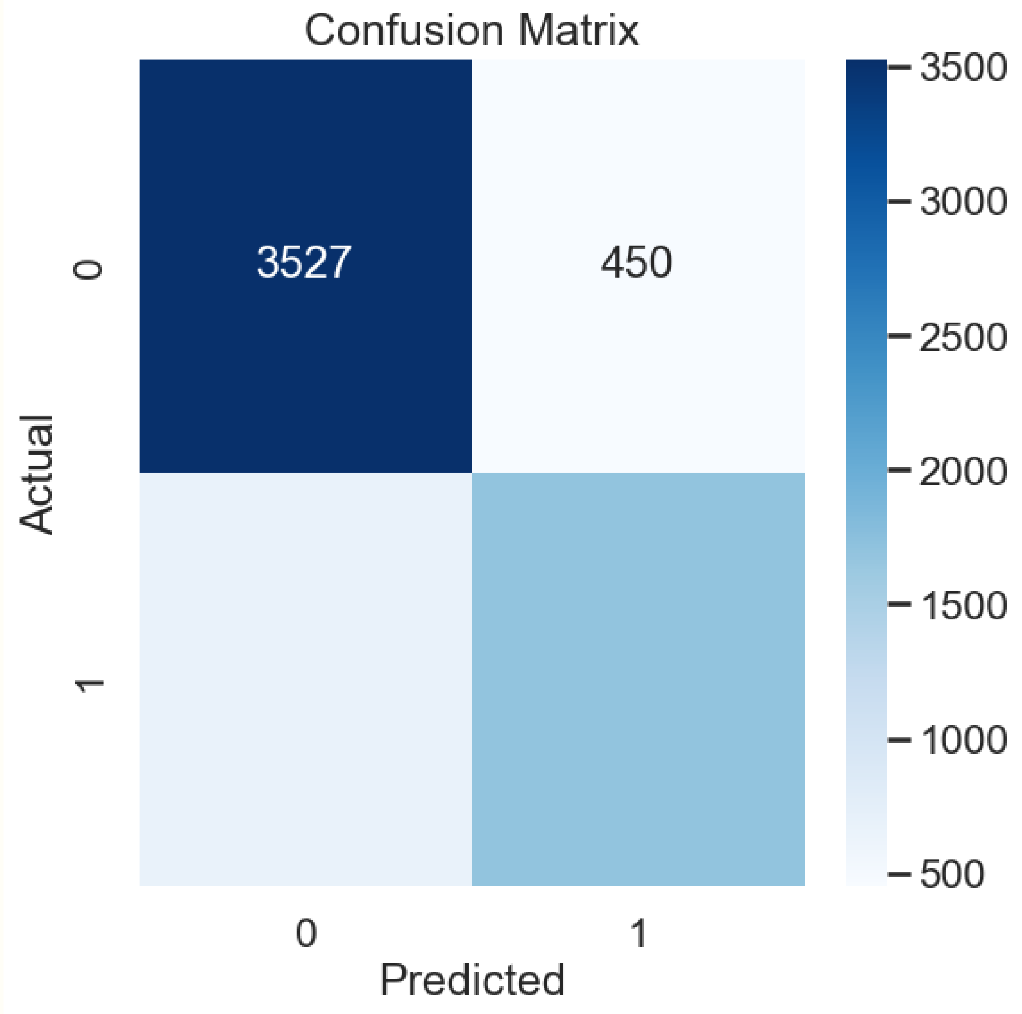
02 **Data-Driven Model Refinement:**
The process emphasizes rigorous model training and evaluation on historical data, allowing for iterative refinement based on performance metrics. Evaluation on the training set offers an initial understanding of the model's predictive power, which is then fine-tuned using a precision-recall curve to find the optimal balance between capturing true positives and minimizing false positives.

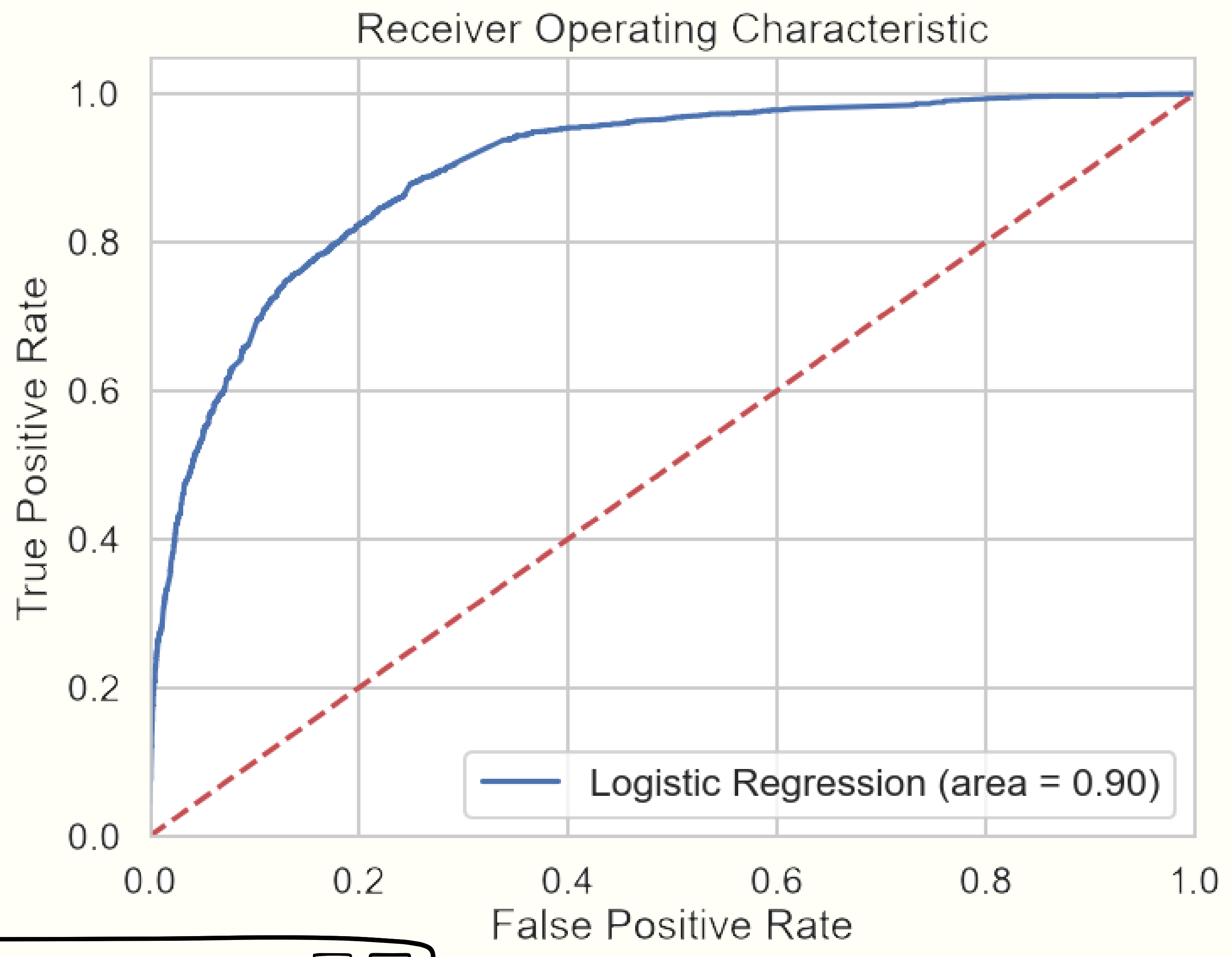
03 **Real-World Application and Validation:**
Post optimization, the model is not just a theoretical construct but is applied to a separate test set, mimicking real-world unpredictability and providing a measure of the model's generalizability. This step is crucial for validating the model's effectiveness in accurately scoring leads for conversion, which is the ultimate goal of the project.



What is the performance of the original model?

The model demonstrates strong predictive power with a ROC AUC of 0.896, indicating excellent capability in distinguishing between converted and not-converted leads. Additionally, a balanced accuracy of 82.27% alongside good precision and recall scores for both classes signifies its effectiveness in accurately classifying leads, making it a reliable tool for prioritizing potential customers.





The ROC curve indicates that the logistic regression model has a high discriminative ability to distinguish between positive and negative classes, with an AUC of 0.90, which is close to 1. This high AUC value suggests that the model has a good measure of separability, meaning it's capable of classifying the leads with a high degree of accuracy.

Analyzing the top coefficients from our logistic regression model

What does the model think about the data?

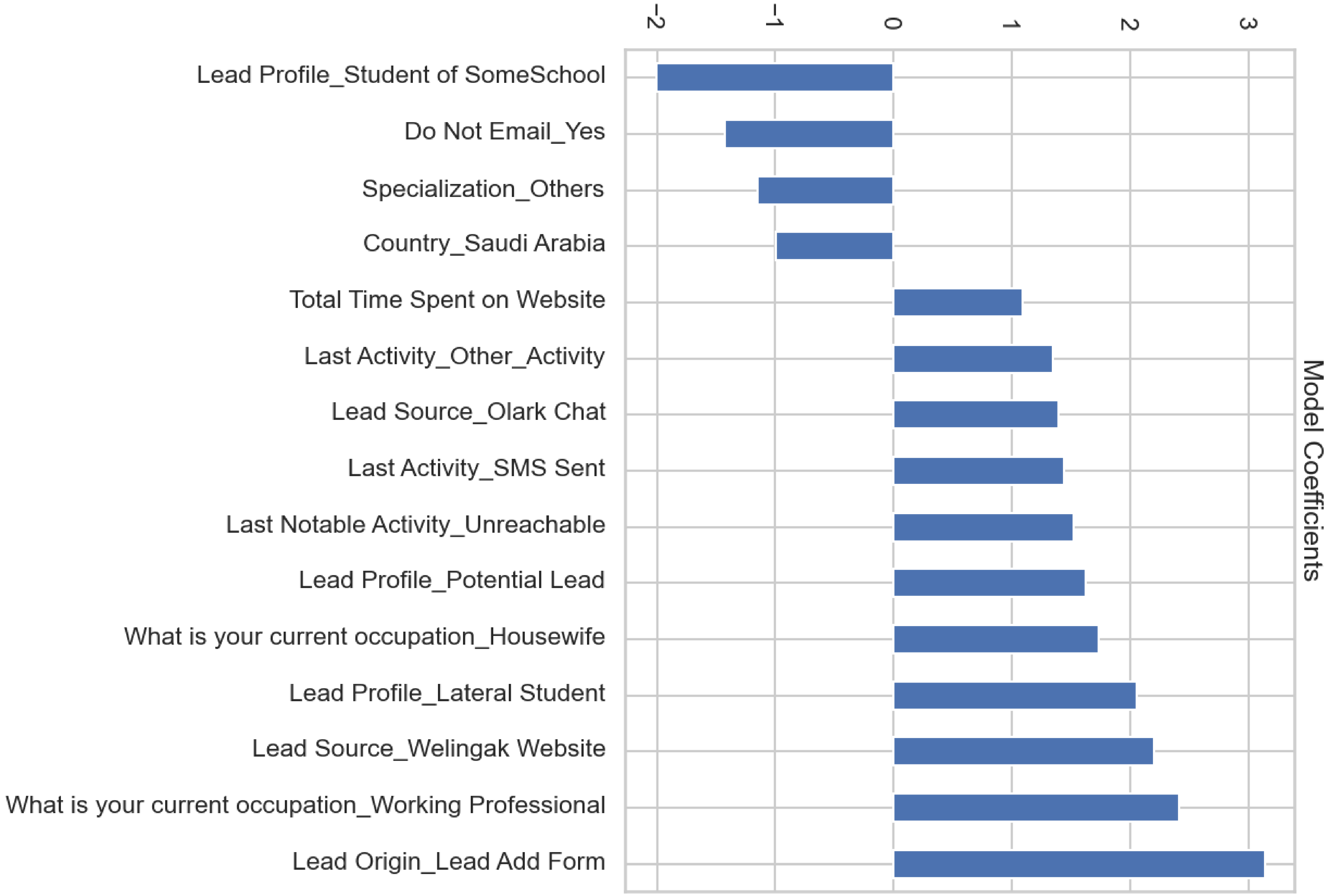
High-Value Actions and Sources: The top positive coefficients like Lead Origin_Lead Add Form, Lead Source_Welingak Website, and actions such as Last Notable Activity_Unreachable suggest that leads from specific sources or those who have been difficult to contact are very likely to convert if engaged correctly.

Occupational Impact: Working professionals are highly likely to convert, indicating the effectiveness of targeting this demographic. Conversely, the negative coefficient for Lead Profile_Student of SomeSchool suggests that this group is less likely to convert, which could indicate a need for different engagement strategies or a reevaluation of the lead scoring for this segment.

Engagement and Content: Total Time Spent on Website and Last Activity_SMS Sent having positive coefficients show that engaged leads and those who respond well to SMS communication are more likely to convert. This implies that enhancing website content and maintaining SMS communication could be beneficial.

Demographic Considerations: The presence of Country_Saudi Arabia suggests geographic location may play a role in lead conversion, possibly due to regional preferences or market fit, which might warrant further investigation.

Communication Preferences: The variable "Do Not Email_Yes" being among the top features suggests that a user's preference regarding email communication is an important predictor, which might reflect on their engagement level or interest in the service/product offered.



Optimizing the threshold of the logistic regression model

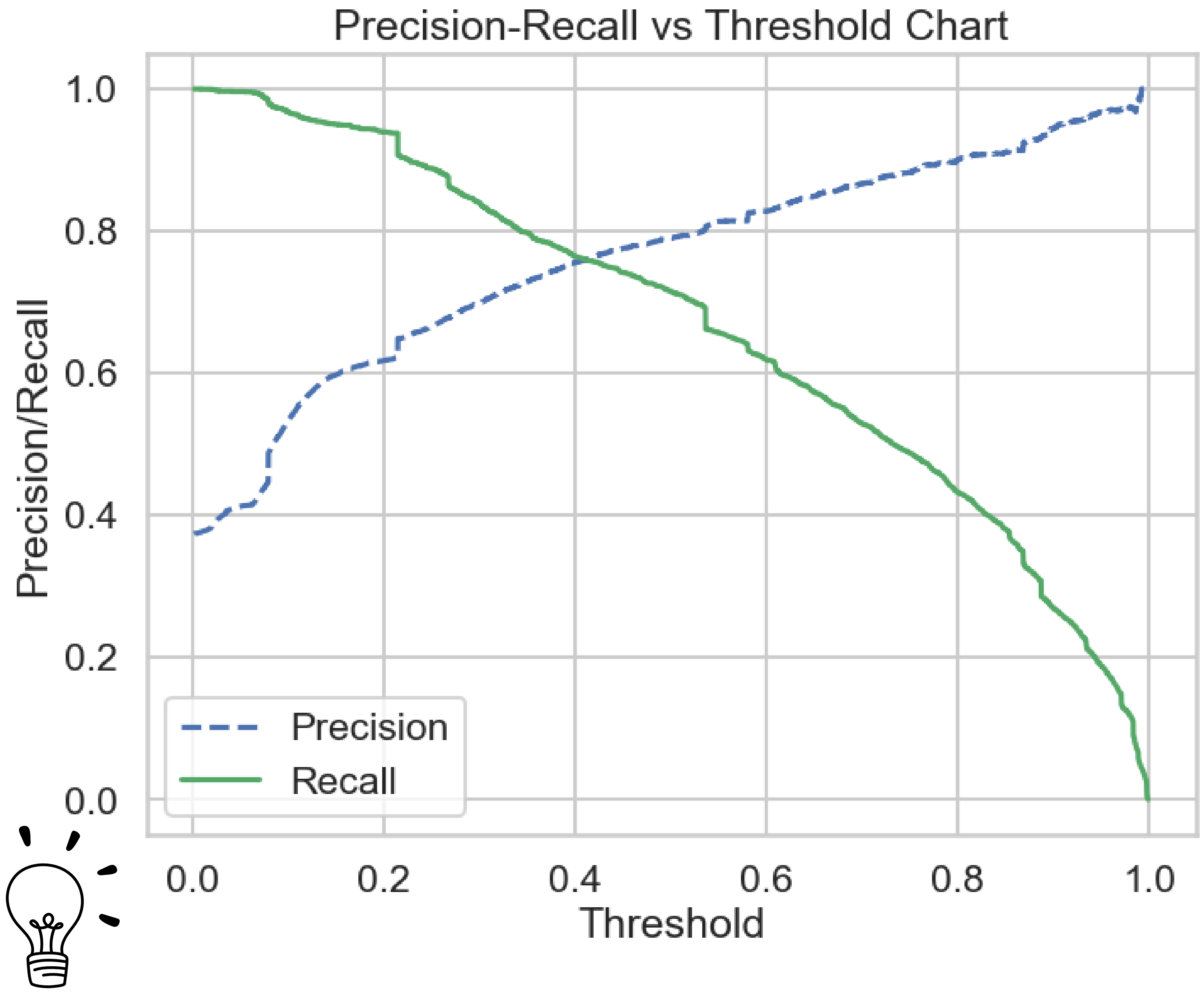
We optimize the threshold of our model via the analysis of a precision recall chart

Insights

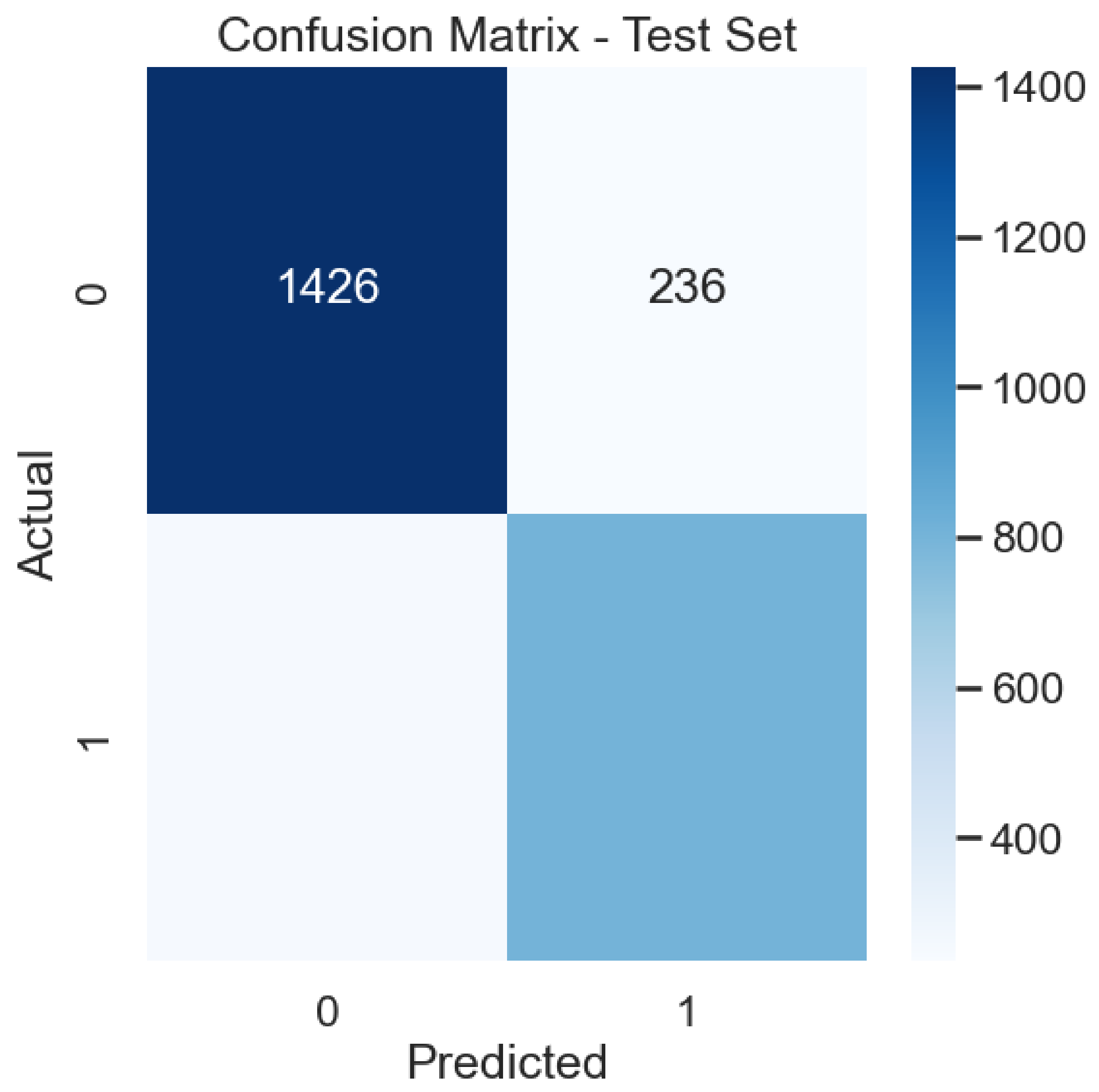
From the curve, we can infer that the best threshold providing a good balance between precision and recall is 0.4.

The optimized model demonstrates enhanced precision and recall balance for both classes on the test set, indicating improved reliability and predictiveness when generalized to unseen data, crucial for practical lead scoring applications.

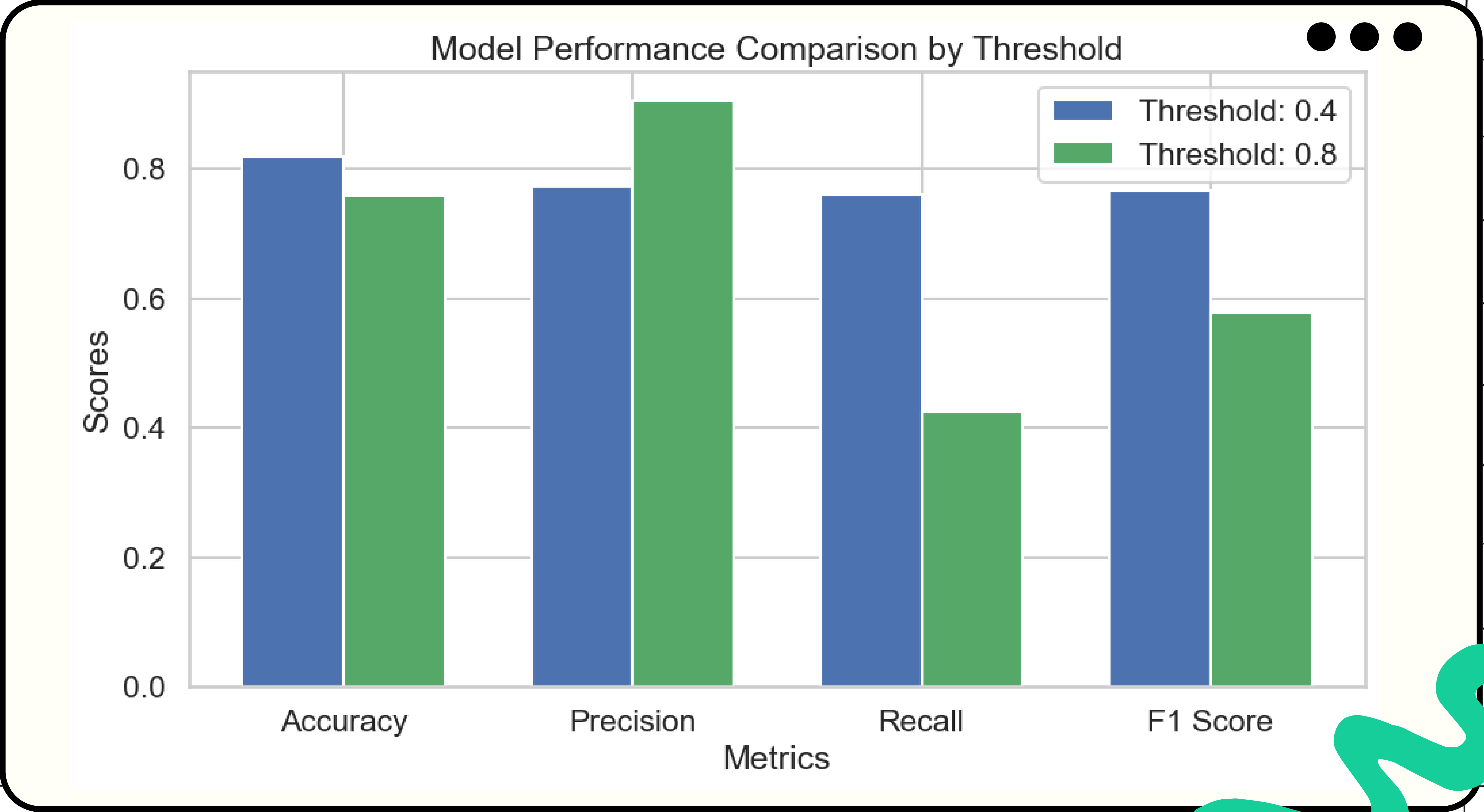
Precision recall chart of the original model



Confusion matrix of the optimized threshold model



Visualizing model performance between the ideal threshold and the threshold for “Hot Leads”requested by the team (80%)



01 The chart shows that at a threshold of 0.4, the model achieves a balance of accuracy, precision, recall, and F1 score, indicating a well-rounded performance.

02 However, when the threshold is raised to 0.8, precision increases significantly at the cost of recall, suggesting that while we are more confident about the leads we classify as 'hot', we also miss out on a higher number of potential leads.

03 This trade-off is important for the business to consider depending on their capacity to pursue leads and the importance of not missing potential opportunities.

Insights

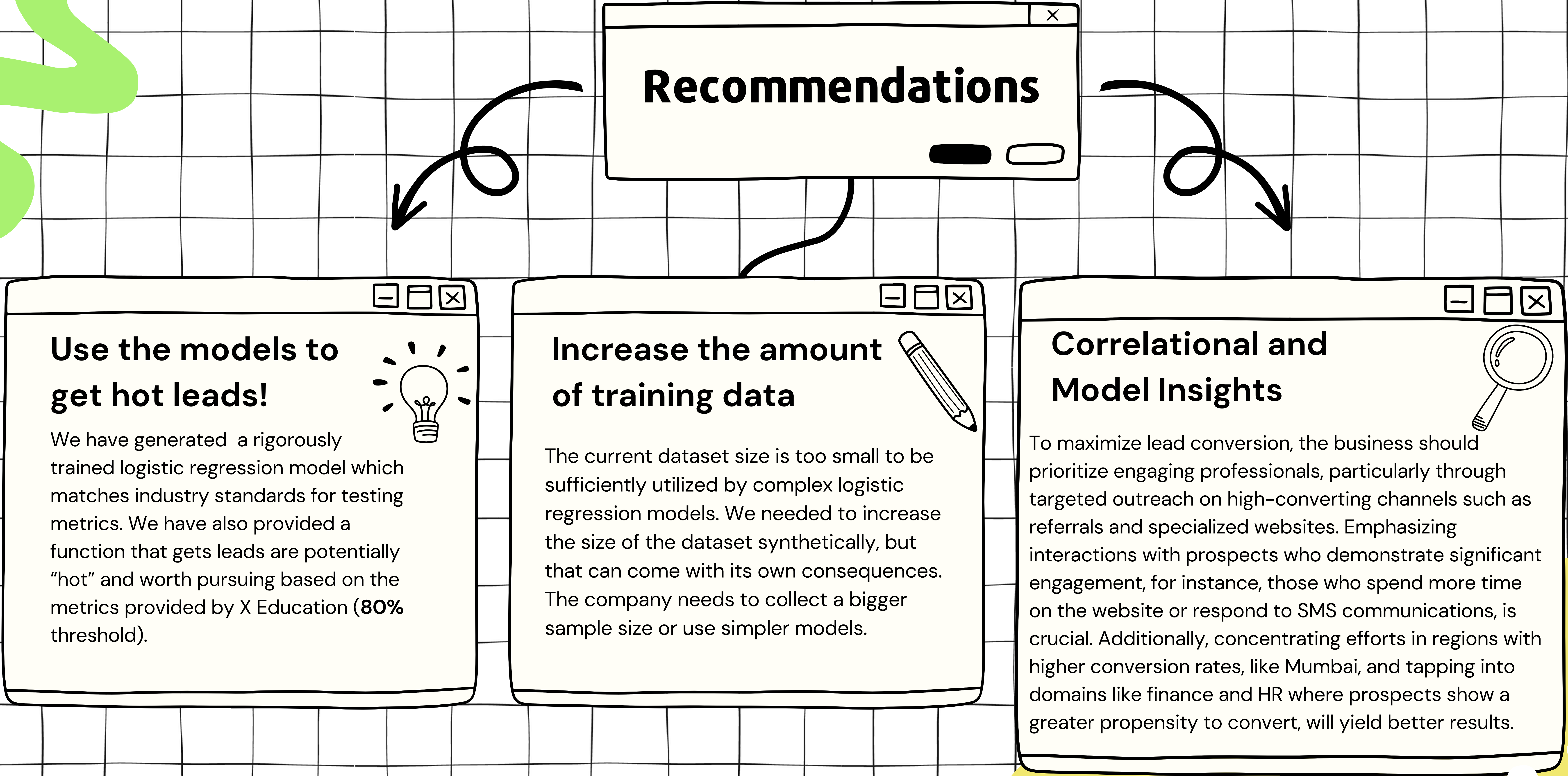
Using a rigorous combination of EDA and model generation, we are able to provide insights into what potentially defines a "Hot lead" which we are providing below :

Exploratory Data Analysis

1. API and Landing Page Submission are major lead sources with 30-35% conversion.
2. Lead Add Form shows over 90% conversion but with fewer leads.
3. Google and Direct traffic are primary lead generators; Reference and Welingak website leads have high conversion.
4. "SMS Sent" as last activity correlates with 60% conversion rate.
5. Working professionals show high likelihood of conversion; Unemployed maintain 30-35% conversion.
6. Most leads are from Mumbai with ~50% conversion rate.
7. Longer time spent on website increases conversion likelihood.
8. Finance and HR specializations have high conversion rates.

Logistic Regression Coefficients

1. "Lead Origin_Lead Add Form" emerges as the strongest predictor, indicating that the origin of the lead significantly influences the likelihood of conversion.
2. "What is your current occupation_Working Professional" and "Lead Source_Welingak Website" are notable predictors, suggesting that a lead's occupational status and specific source channels are critical for predicting conversion.
3. User activities such as "Last Activity_SMS Sent" highlight the importance of engagement metrics in determining lead quality and propensity to convert.



Recommendations

Use the models to get hot leads!

We have generated a rigorously trained logistic regression model which matches industry standards for testing metrics. We have also provided a function that gets leads are potentially “hot” and worth pursuing based on the metrics provided by X Education (**80%** threshold).

Increase the amount of training data

The current dataset size is too small to be sufficiently utilized by complex logistic regression models. We needed to increase the size of the dataset synthetically, but that can come with its own consequences. The company needs to collect a bigger sample size or use simpler models.

Correlational and Model Insights

To maximize lead conversion, the business should prioritize engaging professionals, particularly through targeted outreach on high-converting channels such as referrals and specialized websites. Emphasizing interactions with prospects who demonstrate significant engagement, for instance, those who spend more time on the website or respond to SMS communications, is crucial. Additionally, concentrating efforts in regions with higher conversion rates, like Mumbai, and tapping into domains like finance and HR where prospects show a greater propensity to convert, will yield better results.

Thank you