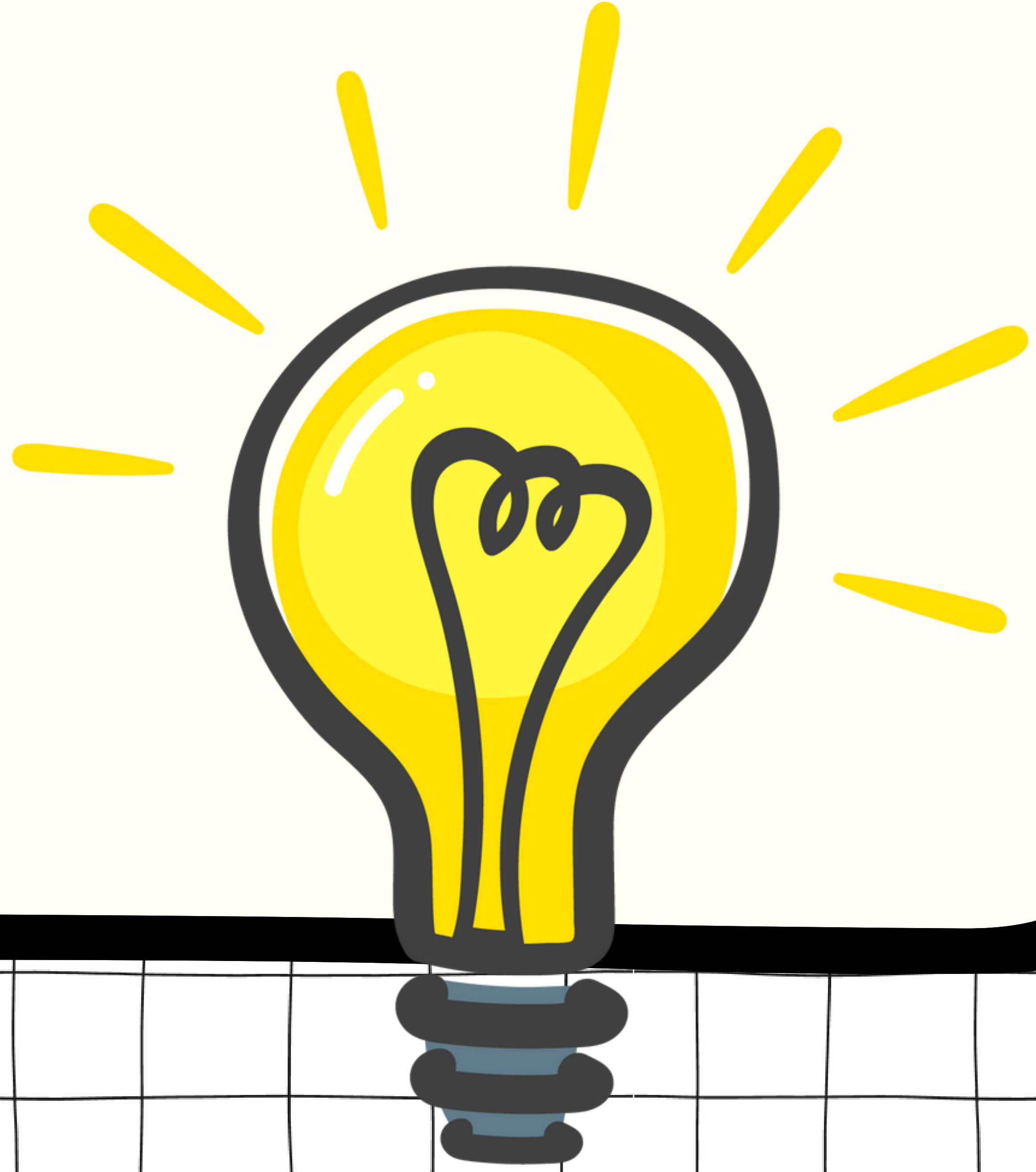
Lead Scoring

Abhishek Singh Dhadwal



Contents

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?

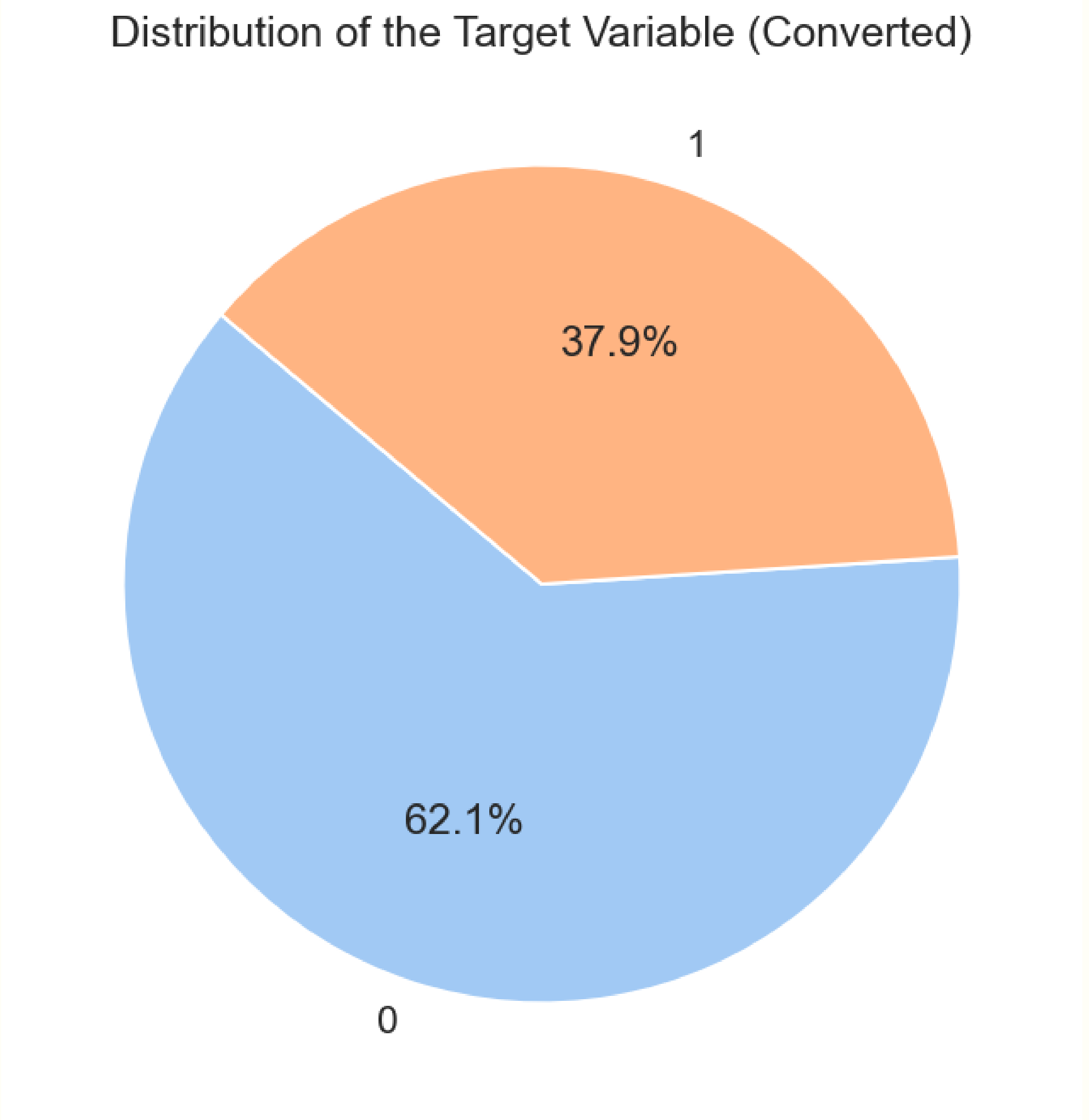
How can a model help with the determination of success for any potential customer?

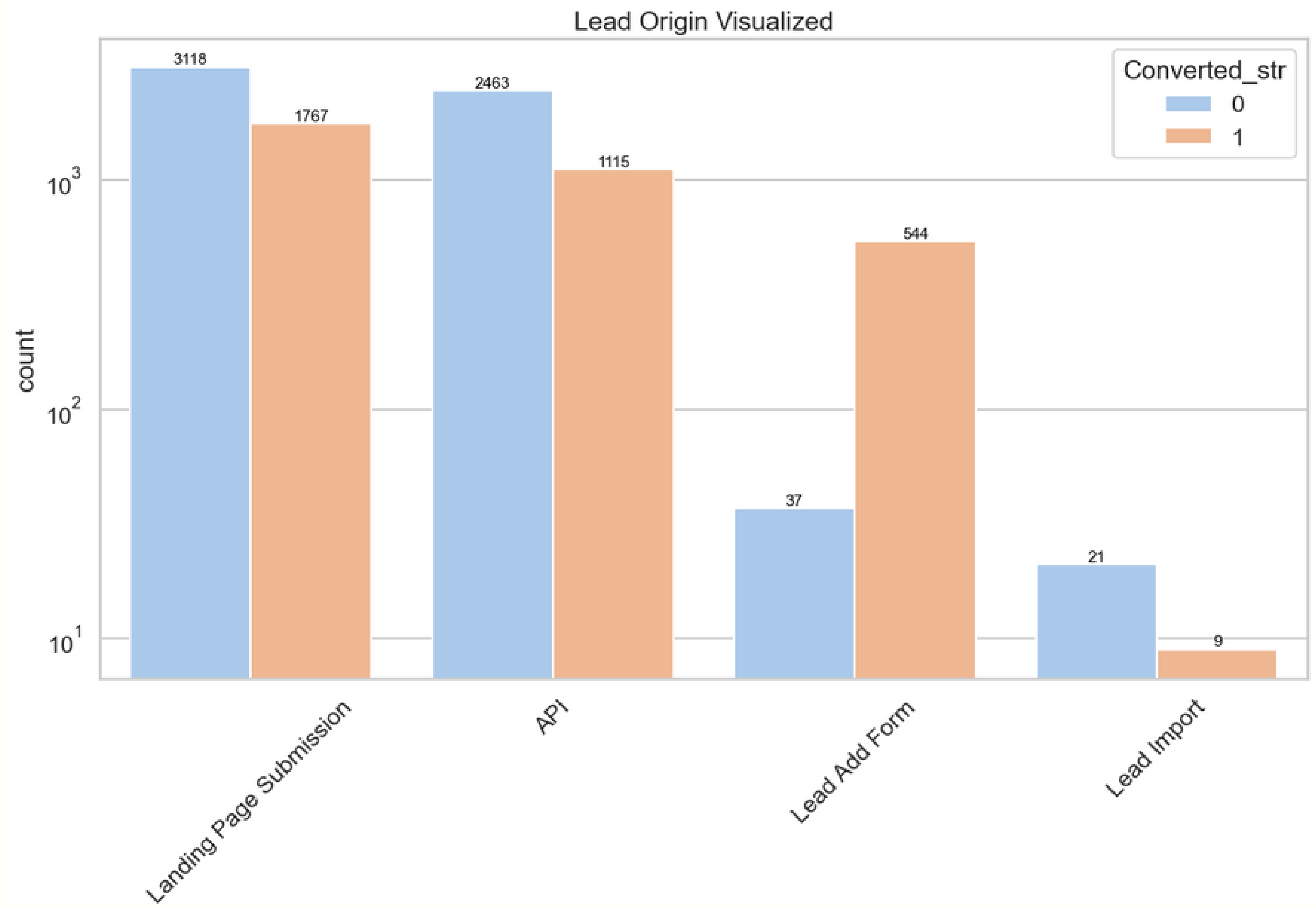
What do stochastic methods of determination state about the deciding factors of "hot leads"?

Findings and Insights

How many leads are being converted in the current scenario?

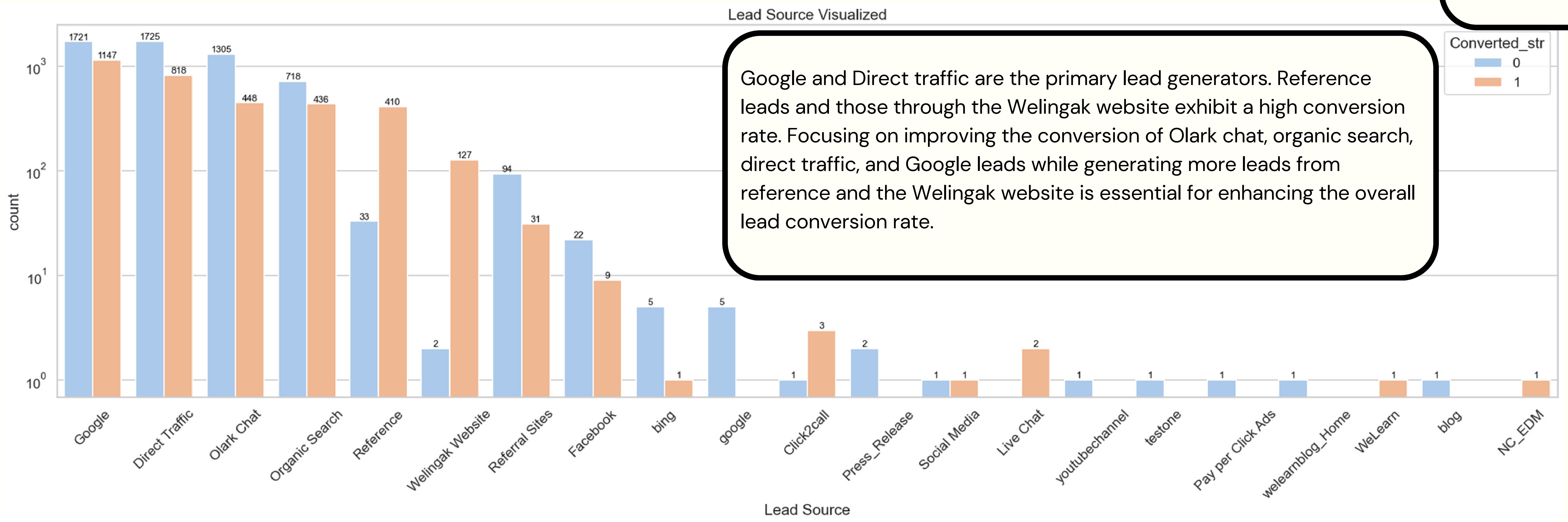
- About 4 in every 10 candidates are being converted in the current methodology, which are lower than our expectations!





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.

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



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.

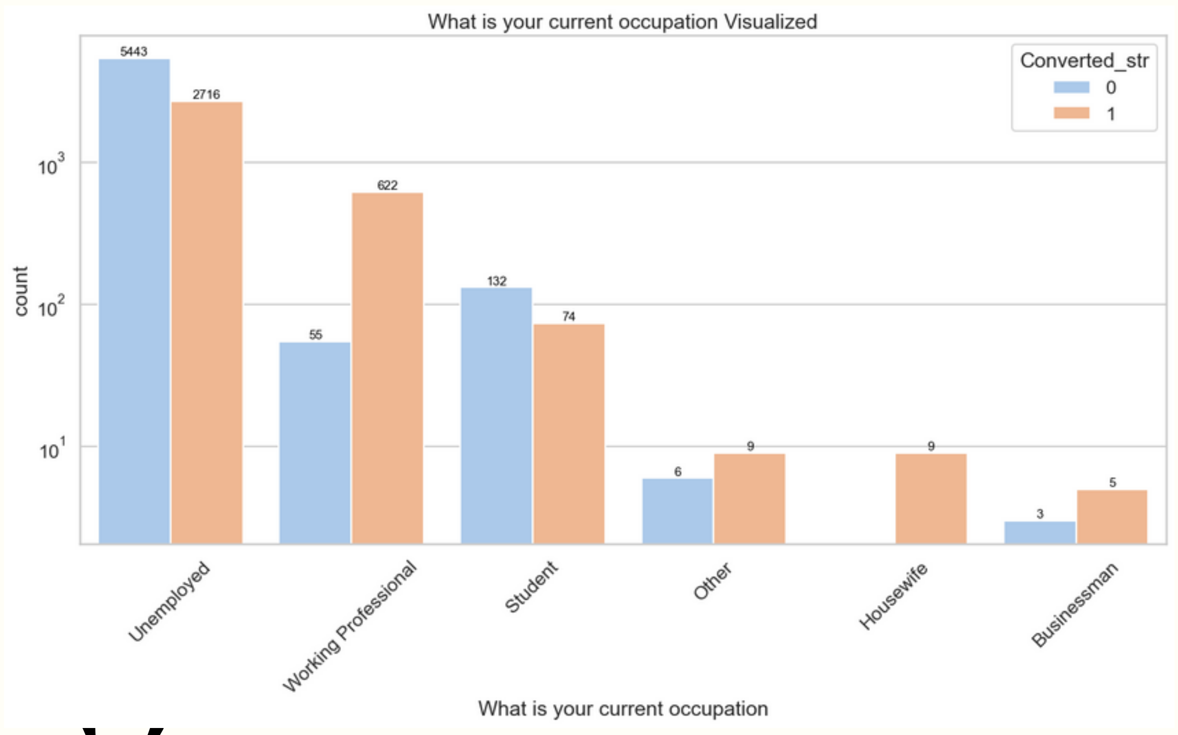
Do demographic factors like location, and methods of contact play into the likelihood of conversion?

Insights

A majority of leads are:

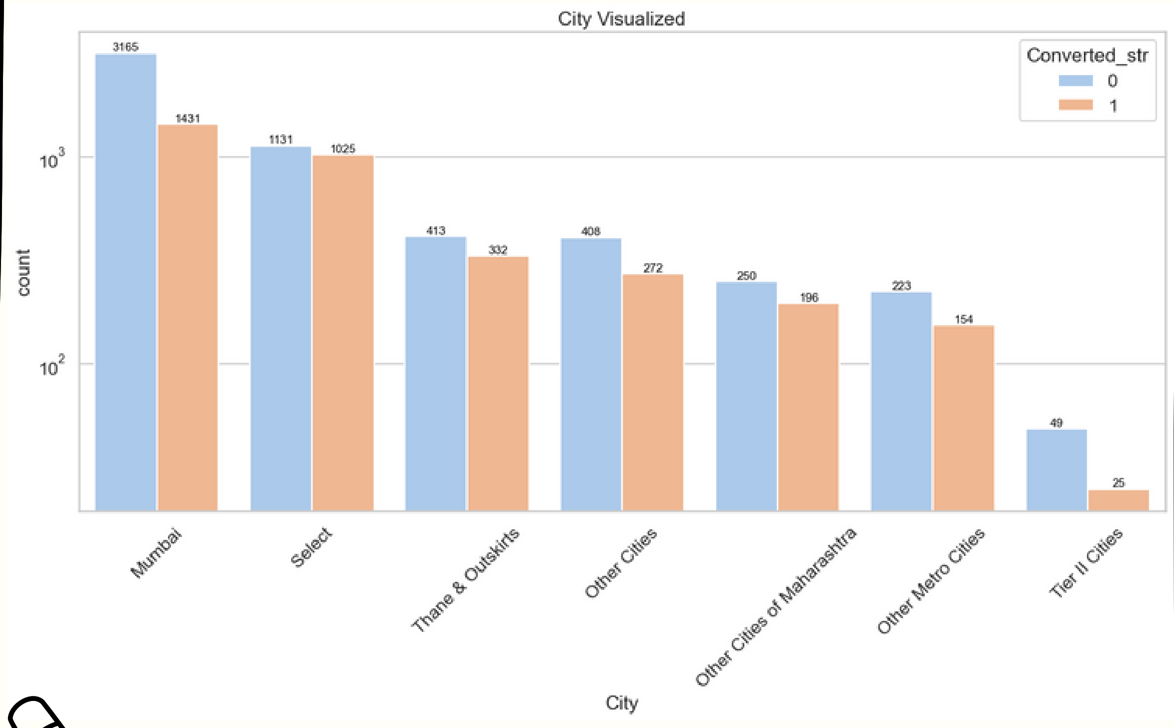
1. From Mumbai
2. Unemployed
3. Not interested in the free textbook

Occupations of leads



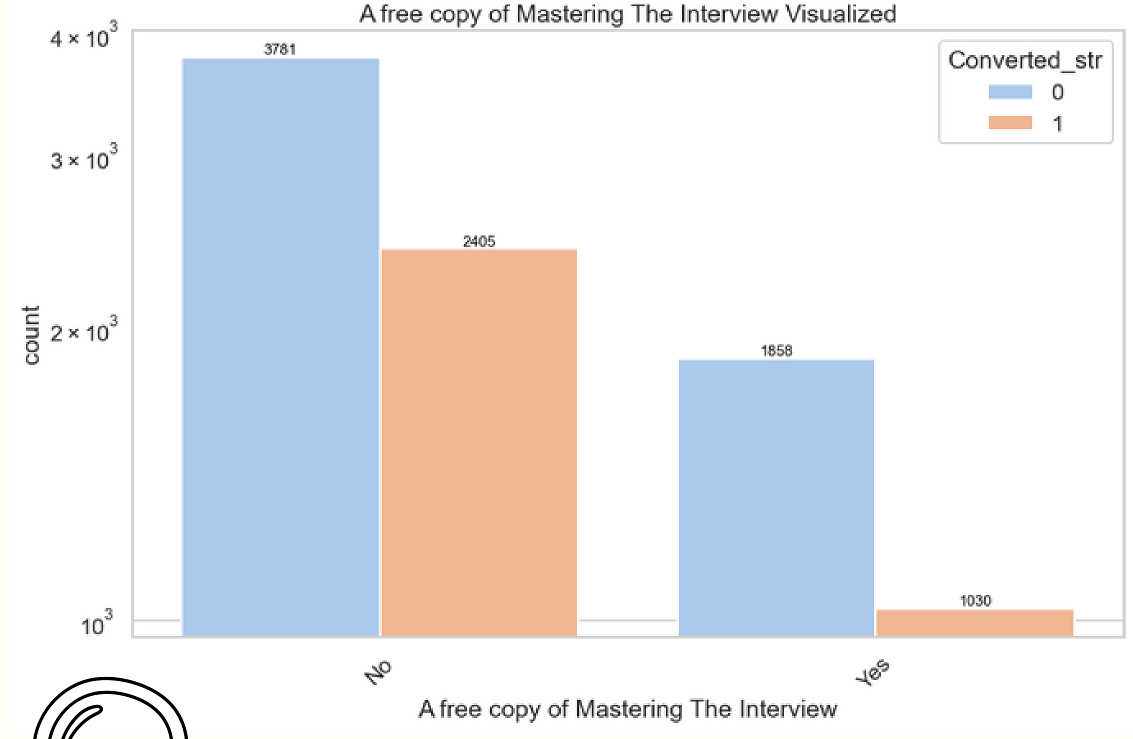
Working professionals opting for the course have a high likelihood of joining it. Unemployed leads are the most numerous but maintain a 30-35% conversion rate.

City Distribution



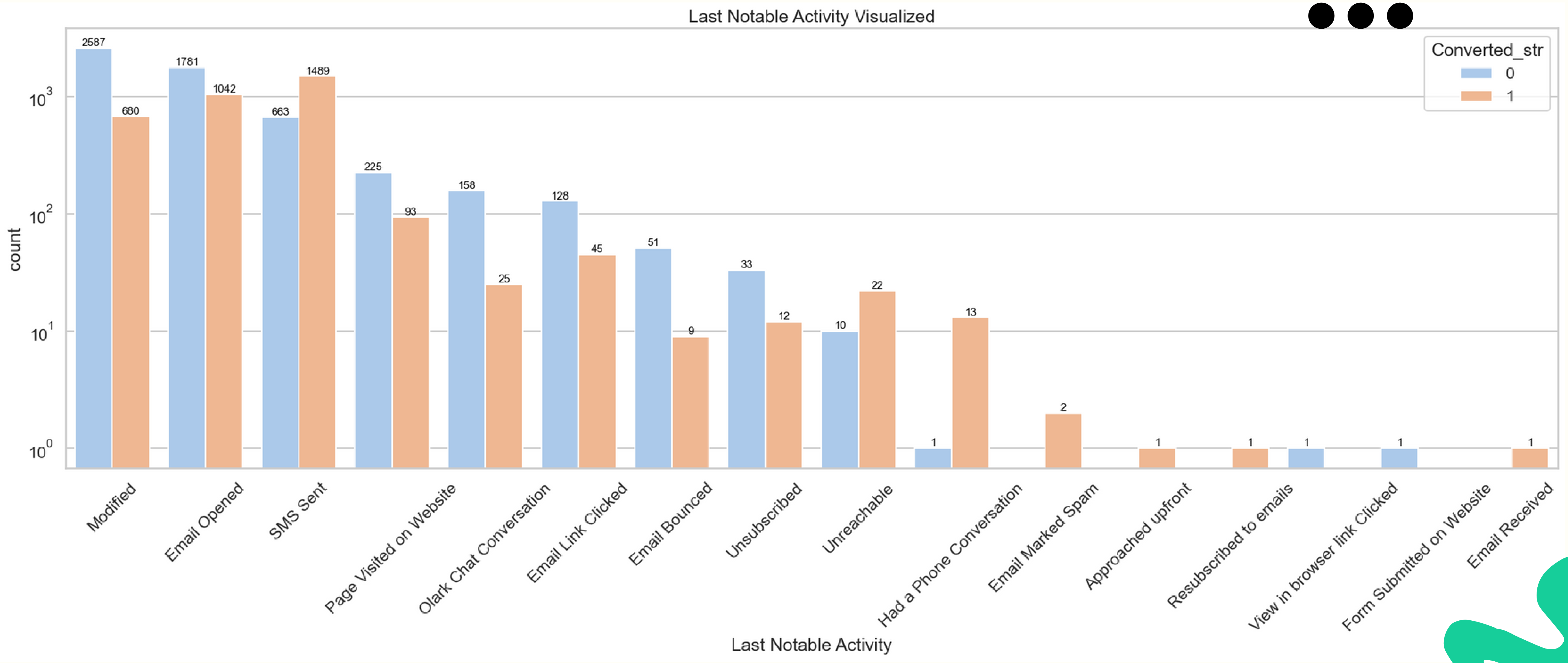
The majority of leads are from Mumbai, boasting a conversion rate of around 50%.

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.

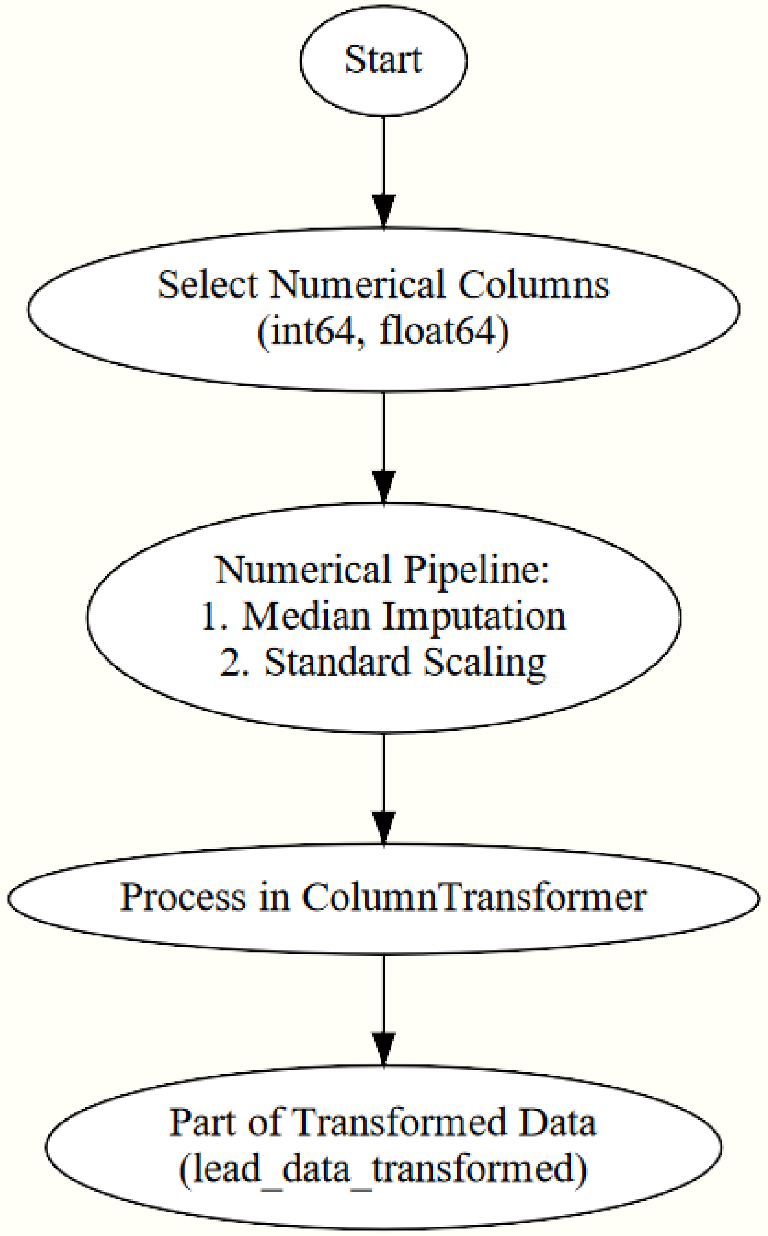
02 **Total Time Spent on Website:** Leads spending more time on the weblise are more likely to be converted.

Preparing our data for the logistic regression model

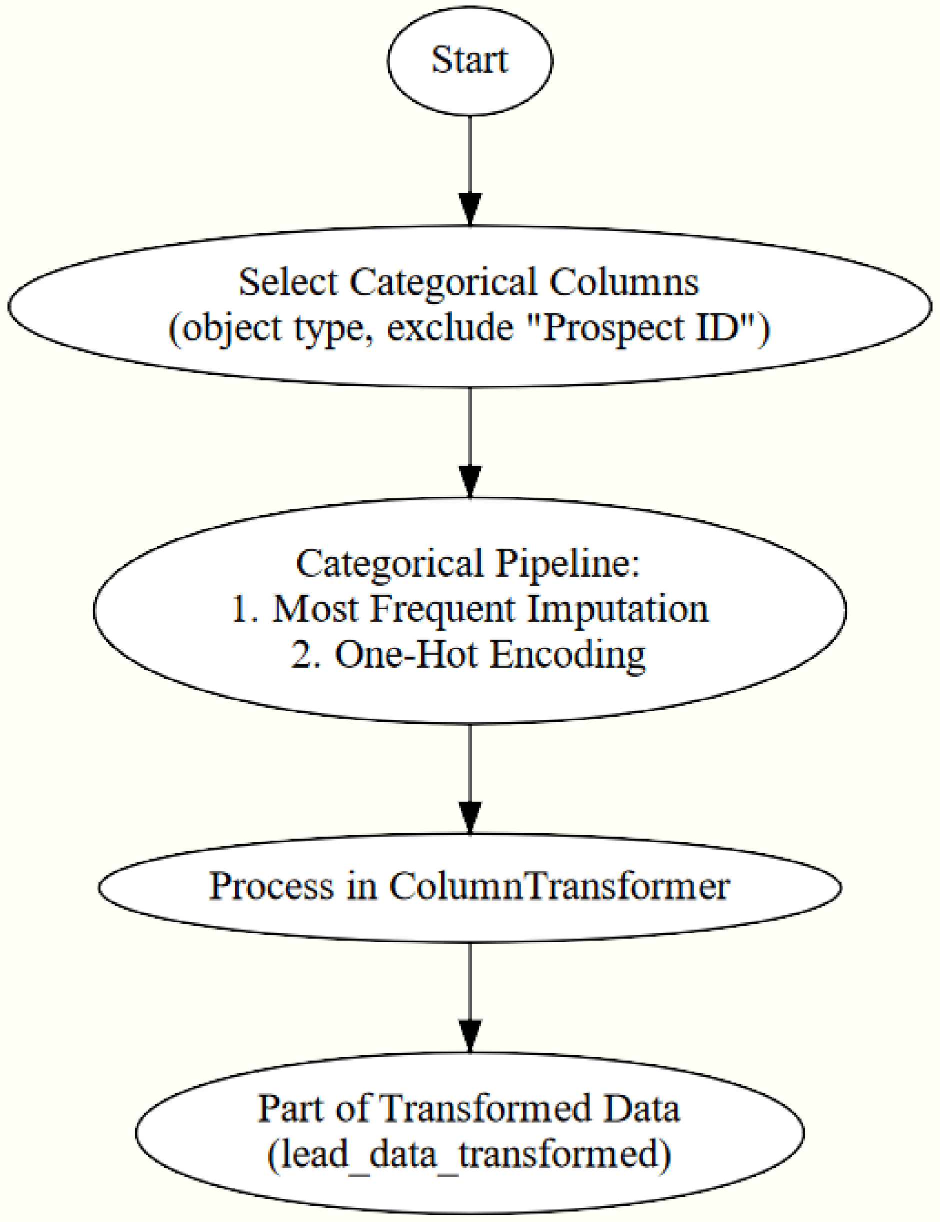
Now that we have made variables of note, we can begin making our model for the company

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 numerical columns of the dataset



Flowchart for processing categorical columns of the dataset



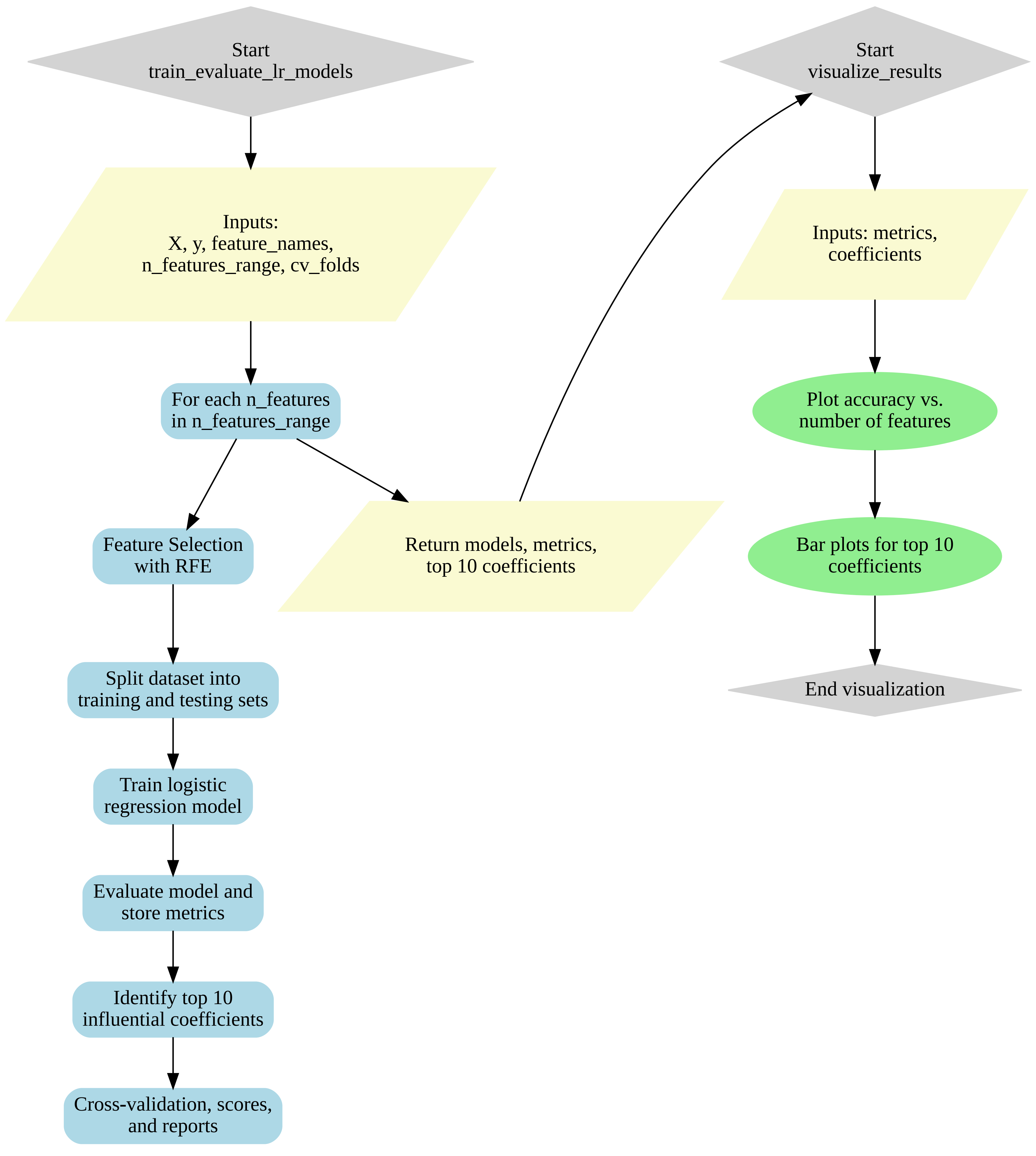
The Original Model Creation Methodology

Generating LR entities with our data

01 **Feature Selection and Model Training:** The train_evaluate_lr_models function employs Recursive Feature Elimination (RFE) to select varying numbers of features from the dataset for training multiple logistic regression models, aiming to identify the optimal feature subset for model performance.

02 **Model Evaluation and Analysis:** Each trained model is evaluated using classification metrics, with the process including the identification of the top 10 most influential features. The evaluation encompasses cross-validation scores, classification reports, and confusion matrices to comprehensively assess model quality.

03 **Results Visualization:** The visualize_results function showcases the efficacy of each model by plotting accuracy against the number of features used and visually represents the significance of the top 10 features through bar plots, facilitating an intuitive understanding of feature importance and model performance.



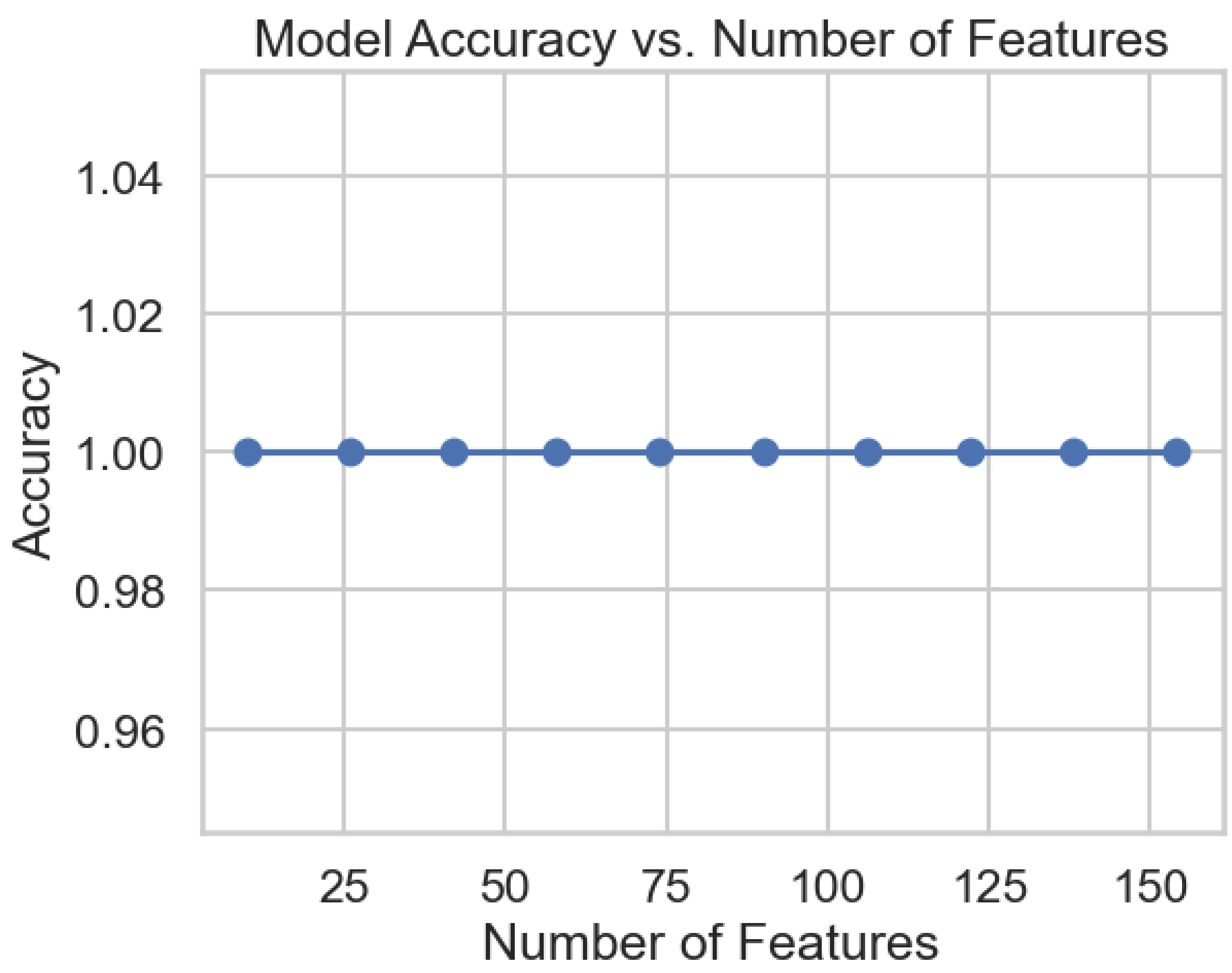
Too perfect? The models reveal a major source of worry for future implementations

Overfitting is a clear consequence

01
Increased Data Dependency:
Logistic regression models are particularly reliant on having access to large volumes of data. Their performance, in terms of accuracy and predictive reliability, significantly improves with the availability of more extensive datasets for training purposes.

02
Heightened Overfitting Risks with Complex Models:
When faced with a constrained dataset size, the application of more sophisticated models introduces a heightened risk of overfitting. Such models might learn to replicate the training data's noise rather than capturing the underlying patterns, thereby failing to predict accurately on new, unseen data.

03
The solution? Expanding the Dataset through Synthesis:
To mitigate the issues arising from a limited dataset, synthesizing additional data can be an effective strategy. Dataset synthesis helps in augmenting the training data, enhancing model robustness by providing a broader variety of training examples.



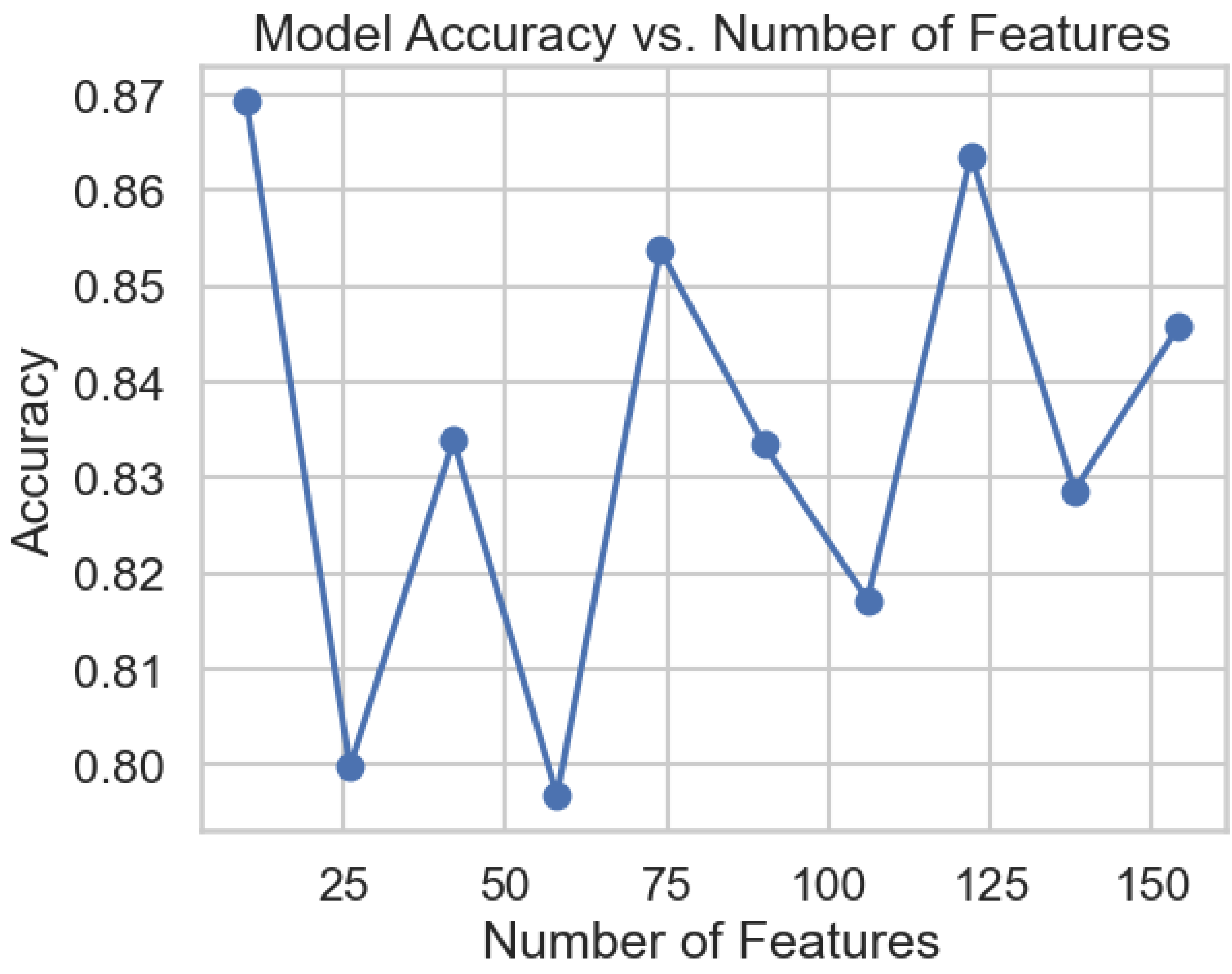
The solution?
Increase the training data via dataset synthesis!

Dataset synthesis helps with accuracy

01 **Integration of Synthetic Data:** To combat the challenge of overfitting and enhance model robustness, the process now includes the generation of synthetic datasets via make_classification. By progressively increasing the size of these datasets, the approach aims to examine how logistic regression models fare across a spectrum of data volumes and complexities, ensuring a more comprehensive evaluation of model performance.

02 **Ensuring Reproducibility:** The deliberate setting of a fixed random_state ensures that both the synthetic data generation and the subsequent model training and evaluation phases produce consistent and reproducible outcomes. This methodological rigor facilitates accurate comparisons and analyses over repeated experiments, contributing significantly to the reliability of the findings.

03 **Adapting to Synthetic Data for Robust Evaluation:** By applying the iterative training and evaluation to synthetically expanded datasets, this method tests logistic regression models more thoroughly. It ensures models meet industry standards across different dataset sizes, effectively addressing overfitting concerns and demonstrating the value of synthetic data in improving model generalization.



Analyzing the top 10 coefficients across all of our models

What do the combined models think about the data?

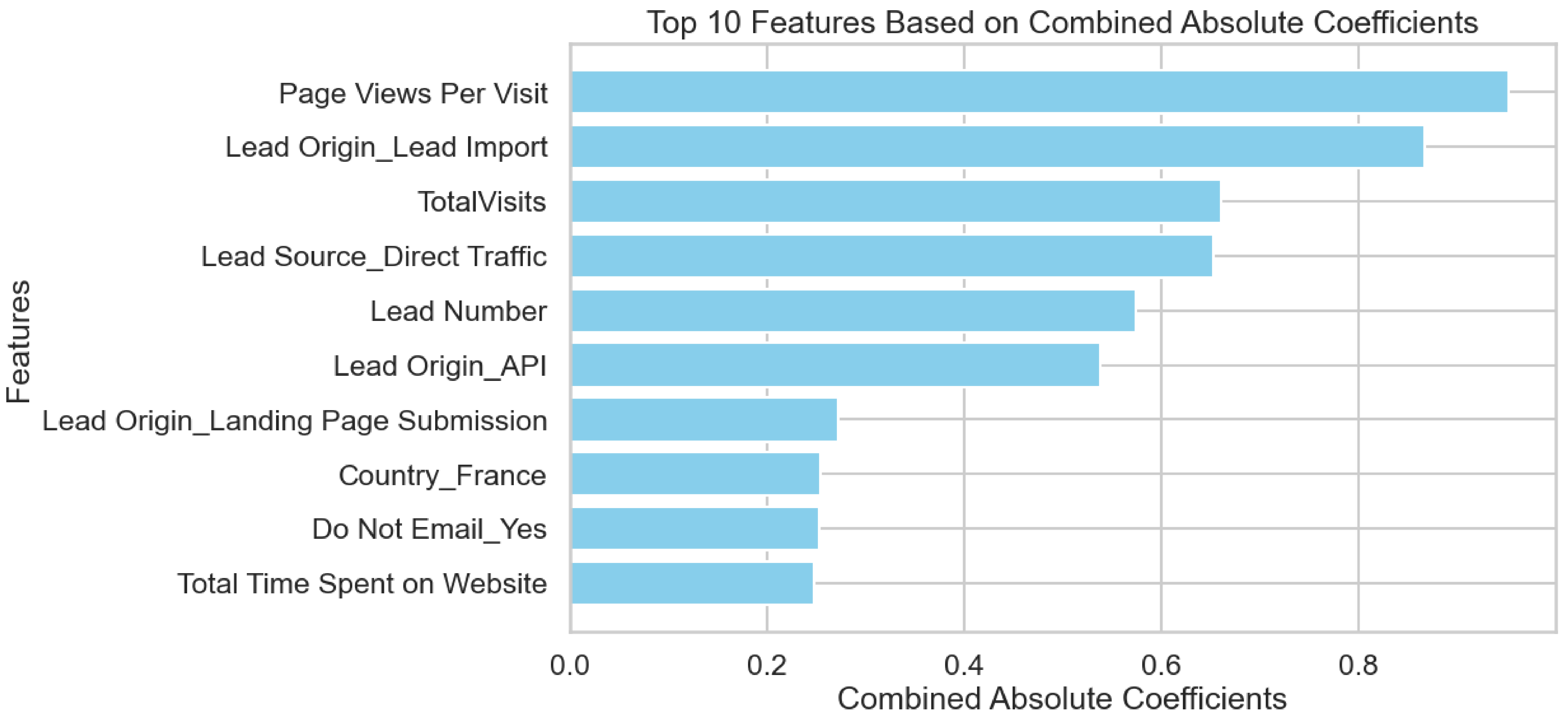
Most Influential Features: The feature "Total Time Spent on Website" has the highest combined absolute coefficient value, indicating it is the most influential factor in the model's outcomes. It is closely followed by "Do Not Email_Yes" and "Country_France", which also appear to be important predictors.

Variety of Features: The features range from user engagement metrics like "Page Views Per Visit" and "TotalVisits" to categorical variables such as "Lead Origin_Lead Import", "Lead Source_Direct Traffic", and "Lead Origin_API". This variety suggests that both quantitative user behavior data and categorical source information are valuable for the model's predictions.

Lead Origin Significance: Several 'Lead Origin' type features are within the top 10, implying that the origin of the lead is a key determinant in the model's decision process.

Geographical Influence: The presence of "Country_France" within the top features indicates that geographic location, or at least being associated with France, is a significant factor for the model, potentially affecting the outcome of the prediction.

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.



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. "Total Time Spent on Website" is the most influential predictor.
2. "Do Not Email_Yes" and "Country_France" are significant predictors.
3. Lead Origin features, especially API, indicate lead source importance.
4. Geographic location, particularly France, significantly affects predictions.
5. User engagement metrics and source information are key to model predictions.
6. Communication preferences, e.g., email opt-out, influence conversion likelihood.

Recommendations

Use the models to get hot leads!

We have generated an ensemble of rigorously trained logistic regression models which match industry standards for testing metrics. These are trained across a wide variety of feature sets and can be used to understand which leads are potentially “hot” and worth pursuing.

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

Our insights underscore the critical impact of lead origin and engagement on conversion rates, with "Total Time Spent on Website" and "Lead Add Form" being key drivers. Notably, "SMS Sent" activity, preferences like "Do Not Email_Yes," and specific geographies (e.g., "Country_France") emerge as significant predictors. This highlights the importance of targeted lead sourcing and personalized engagement strategies in enhancing conversion effectiveness.

Thank you