Case Study

Leads Scoring

Chinmay Biradar

Problem Statement

X Education offers professional training courses delivered through an online format, catering to industry professionals seeking to enhance their skills and knowledge remotely. As an education company operating in the digital realm, X Education recognizes the critical importance of effectively converting website visitors into enrolled students.

Understanding the multitude of factors that influence student conversion is paramount for X Education's success in a competitive online landscape. Factors such as website engagement, course relevance, pricing, and communication strategies play pivotal roles in shaping enrollment rates.

To address this challenge, the objective of the analysis is to identify and analyze these key factors to inform more effective marketing strategies and ultimately drive improved enrollment rates. By discerning the nuances of what motivates prospective students to enroll in their courses, X Education can tailor its approach to attract and convert leads more efficiently, thereby fostering sustained growth and success in the online education market.

Chinmay Biradar

Analysis Approach

Libraries:

The code begins by importing necessary libraries for data manipulation (pandas, numpy), visualization (matplotlib, seaborn), model building (scikit-learn, statsmodels), and warning suppression.

Data Loading and Exploration:

The leads data is loaded from a CSV file using pandas. Initial exploration involves checking the shape of the dataset, basic statistics of numerical columns, detecting duplicates, and identifying null values in each column.

Data Cleaning:

- **Nomenclature:** Column names are converted to snake case for consistency.
- **Handling "Select":** "Select" entries in certain categorical columns are replaced with null values.
- **Dropping Columns:** Irrelevant columns and those with a high percentage of null values or related to sales generated activities are dropped.
- **Handling Null Values:** Null values in categorical columns are imputed by either combining categories or proportionately imputing based on existing distribution.
- **Data Imbalance:** Binary columns with significant data imbalance are dropped.
- **Outliers:** Outliers in numerical columns are capped at the 99th percentile to mitigate their impact.
- **Exploratory Data Analysis (EDA):** EDA includes visualizations of the distribution of numerical columns, correlations between them, and bar plots for categorical variables to understand data patterns and relationships.

Analysis Approach

Exploratory Data Analysis (EDA):

EDA includes visualizations of the distribution of numerical columns, correlations between them, and bar plots for categorical variables to understand data patterns and relationships.

Data Preparation:

- **Converting Binary Variables:** Binary variables (Yes/No) are mapped to 0/1 for modeling purposes.
- **Creating Dummy Variables:** Dummy variables are created for categorical columns using one-hot encoding.
- **Outliers Treatment:** Remaining outliers in numerical columns are capped at the 99th percentile.
- **Train-Test Split:** The dataset is split into training and testing sets for model evaluation.

Feature Scaling:

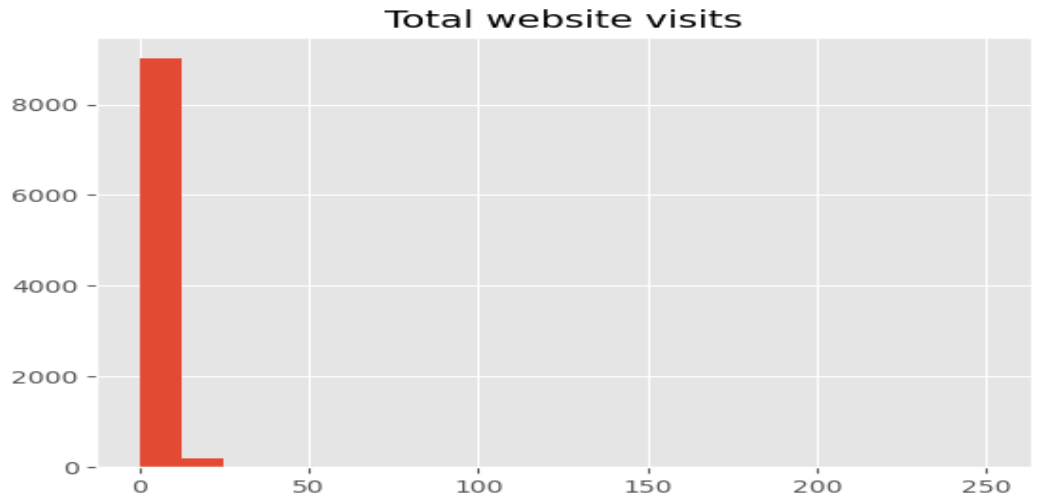
Numerical features are scaled using StandardScaler to ensure uniformity and prevent dominance by features with larger scales.

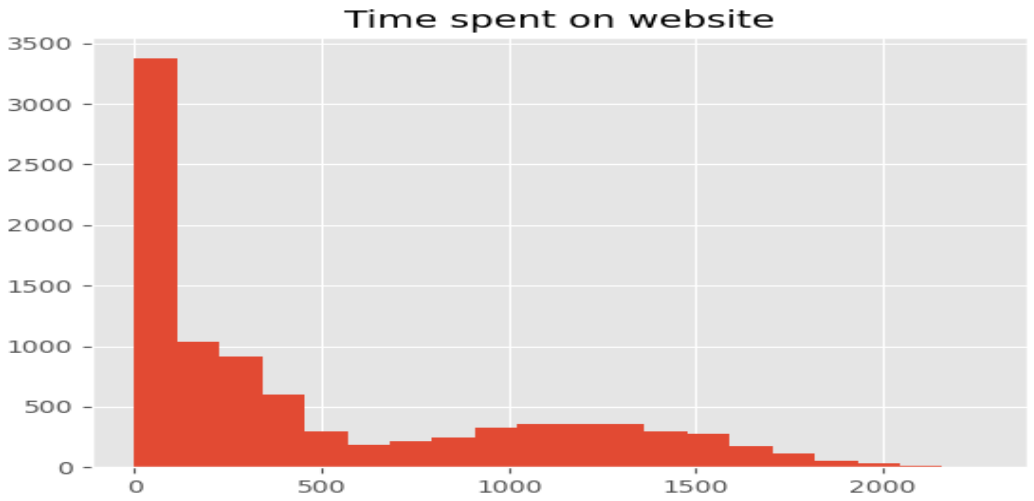
Model Building:

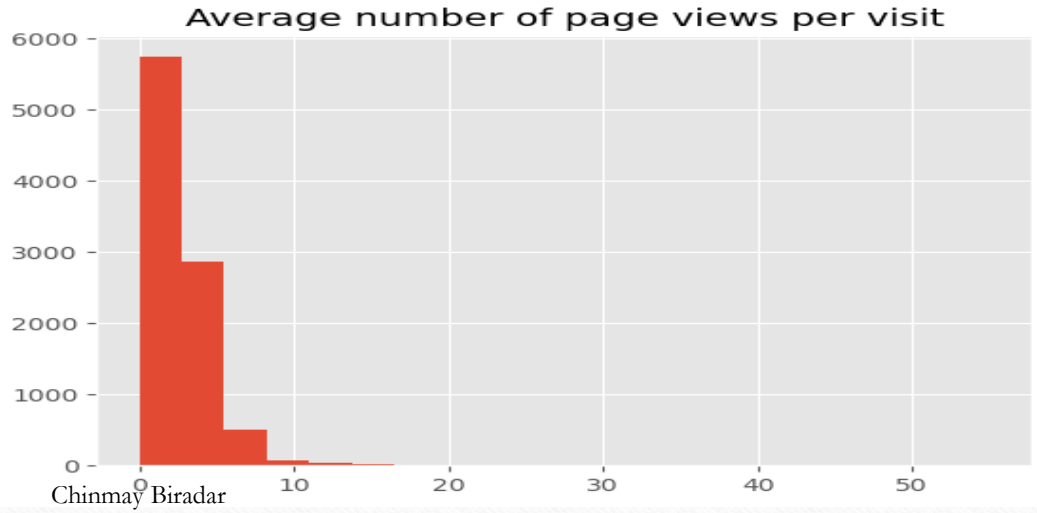
- **Initial Model:** An initial logistic regression model is built using all variables to predict lead conversion, and the model summary is displayed.
- **Feature Selection:** Recursive Feature Elimination (RFE) is employed to select the most important variables for predicting conversion.
- **Final Model:** A second logistic regression model is built using the selected features, and its performance is evaluated.

Results

Website Engagement: Increased total visits, longer time spent on the website, and higher page views per visit correlated with heightened conversion rates.

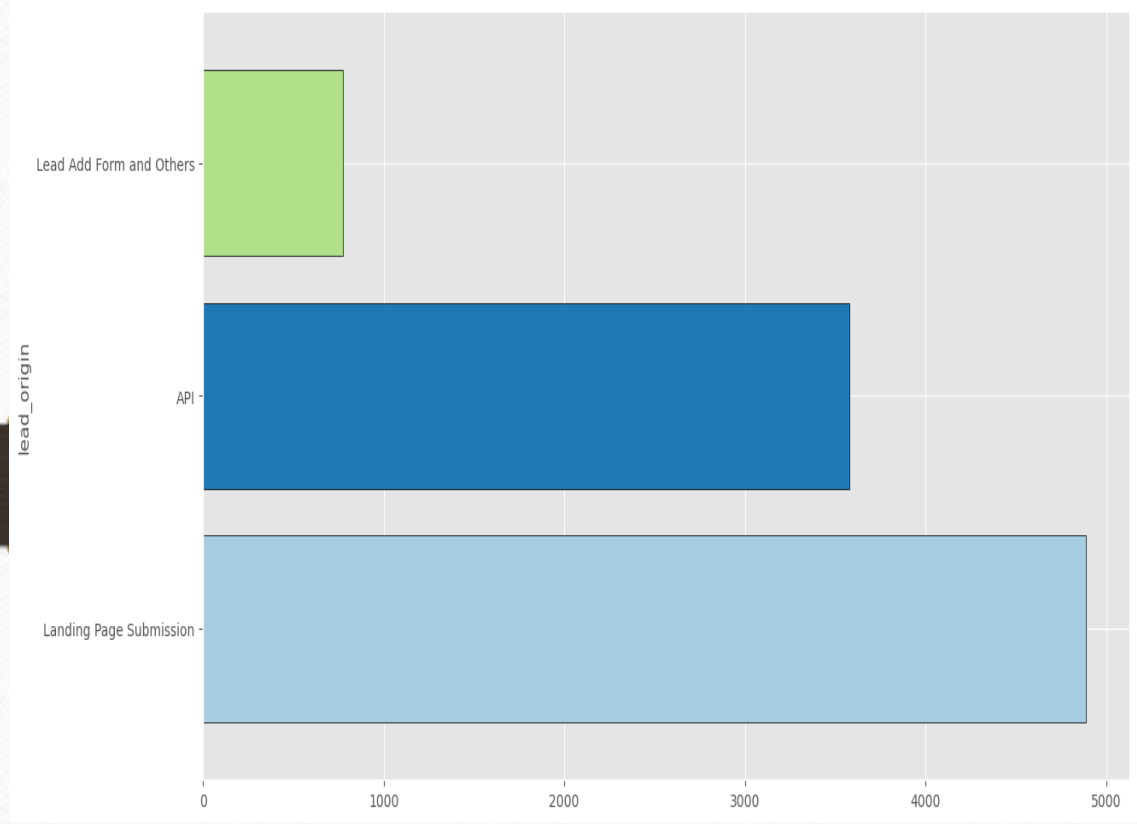


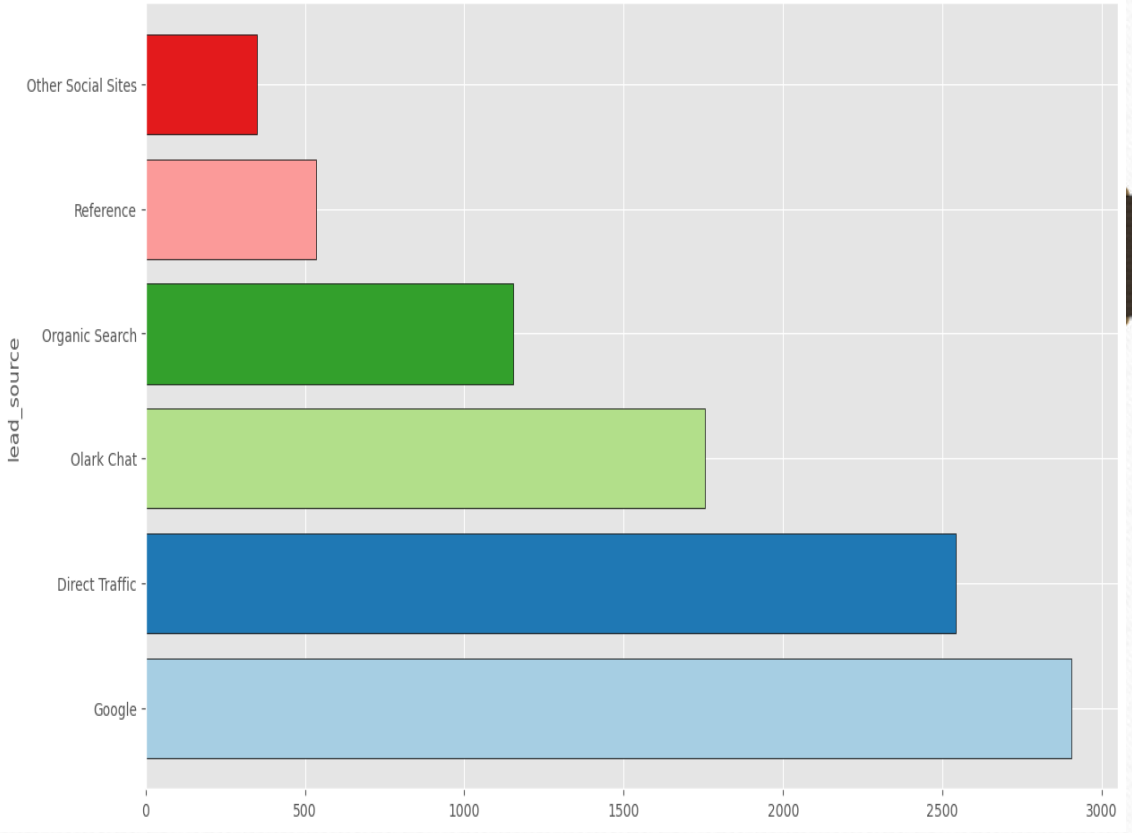




Chinmay Biradar

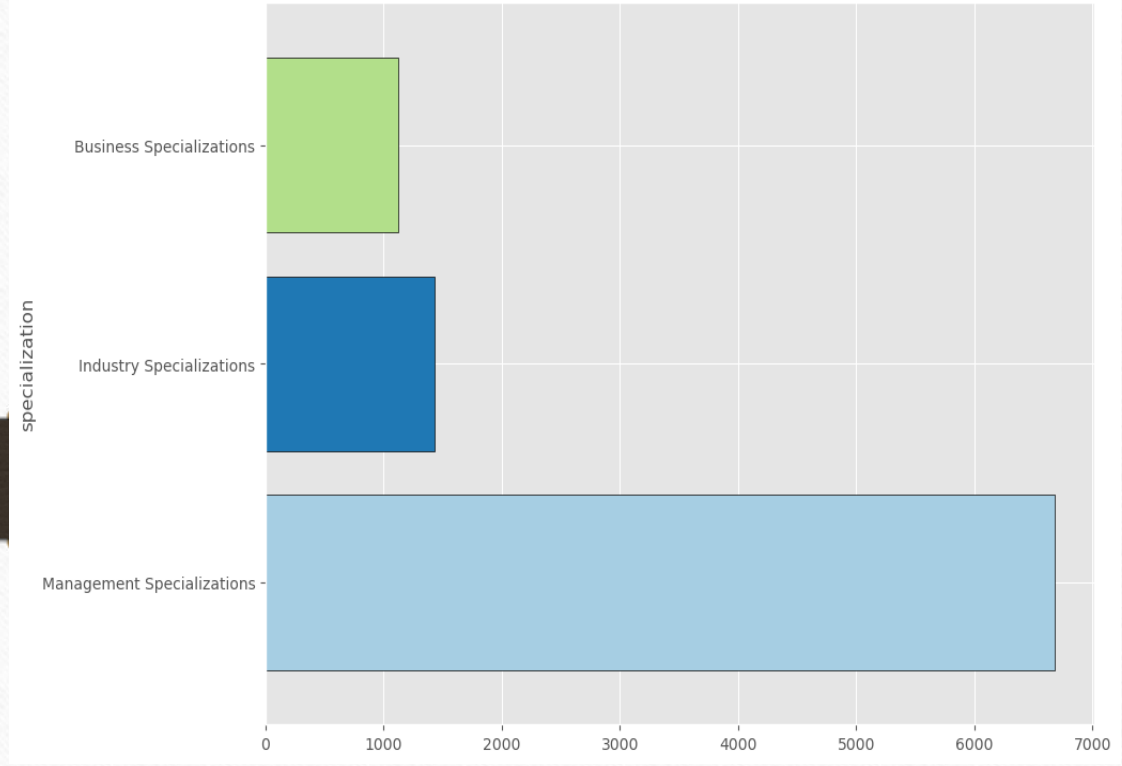
Results



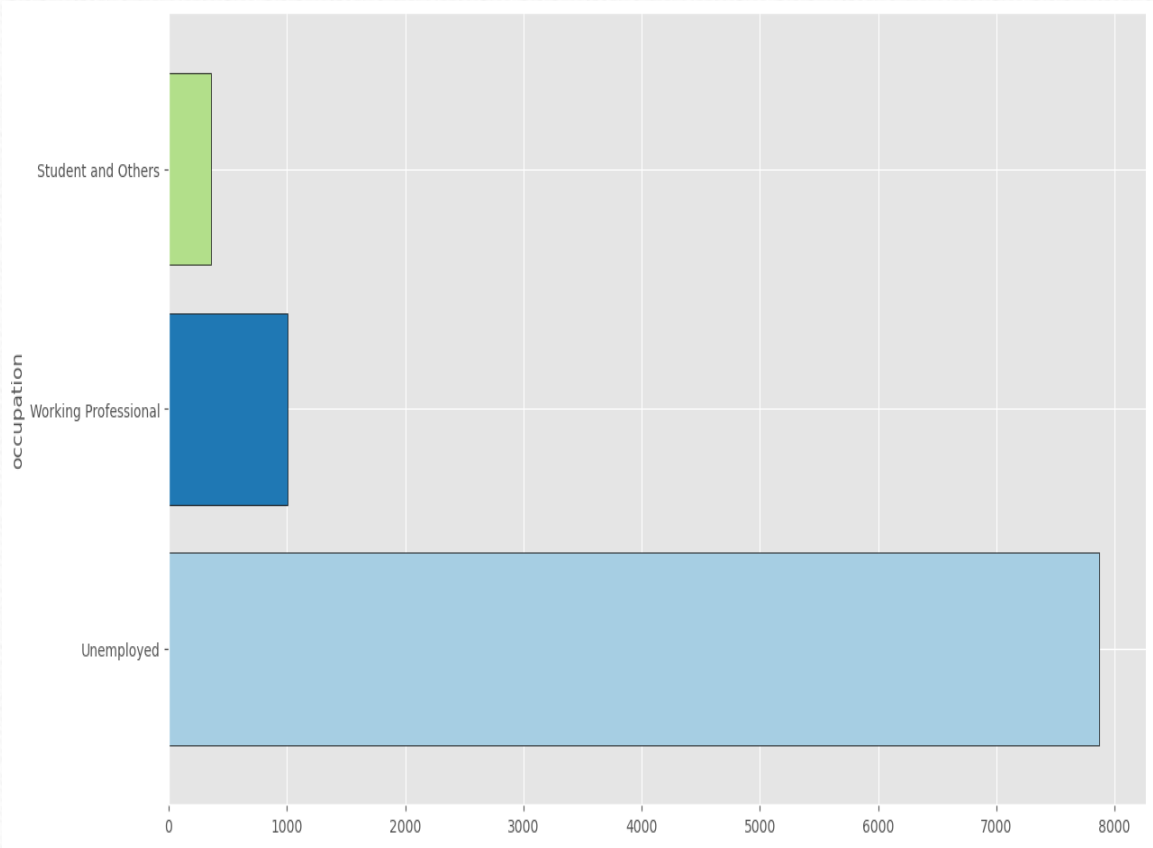


Chinmay Biradar

Results

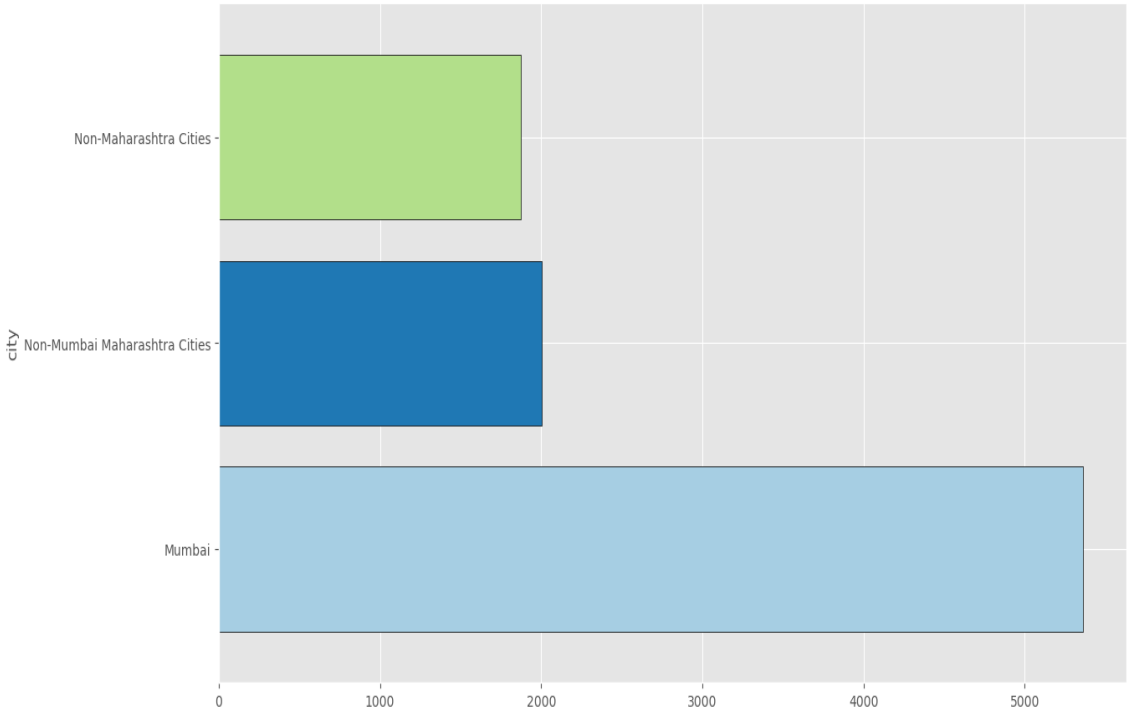


Most of the speciliazation taken are management
Unemployed users are the most significant leads_data

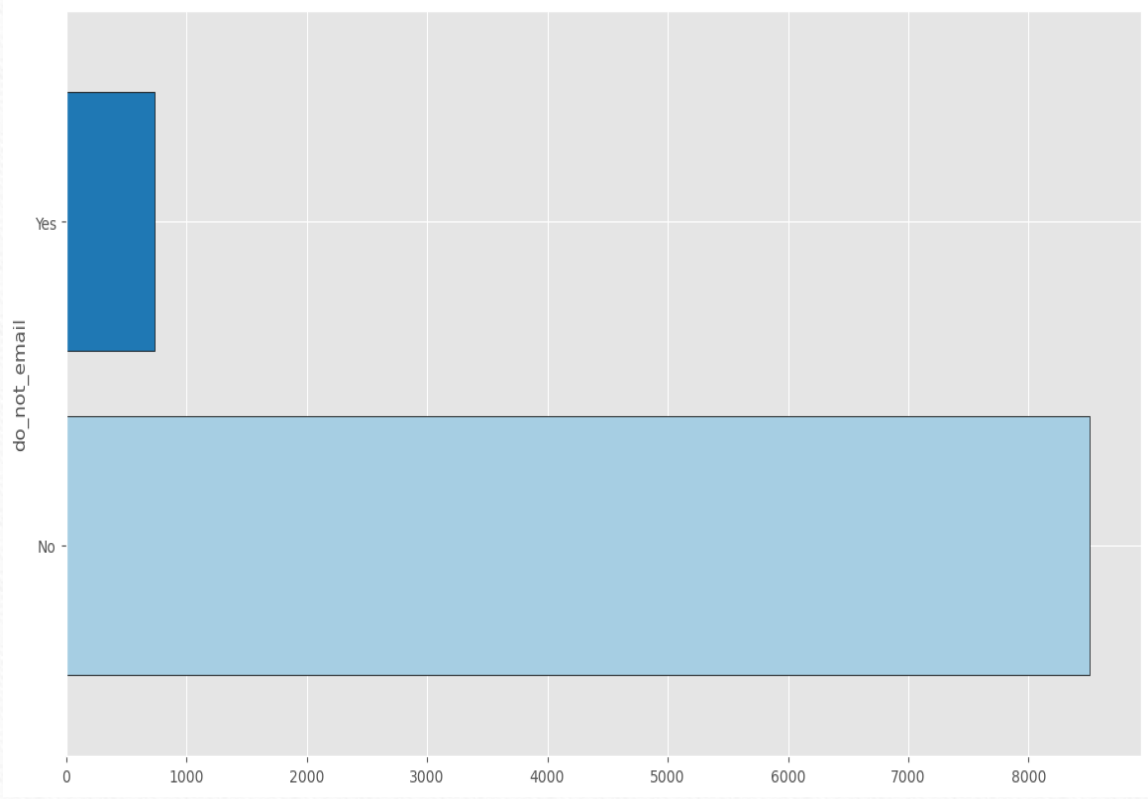


Chinmay Biradar

Results



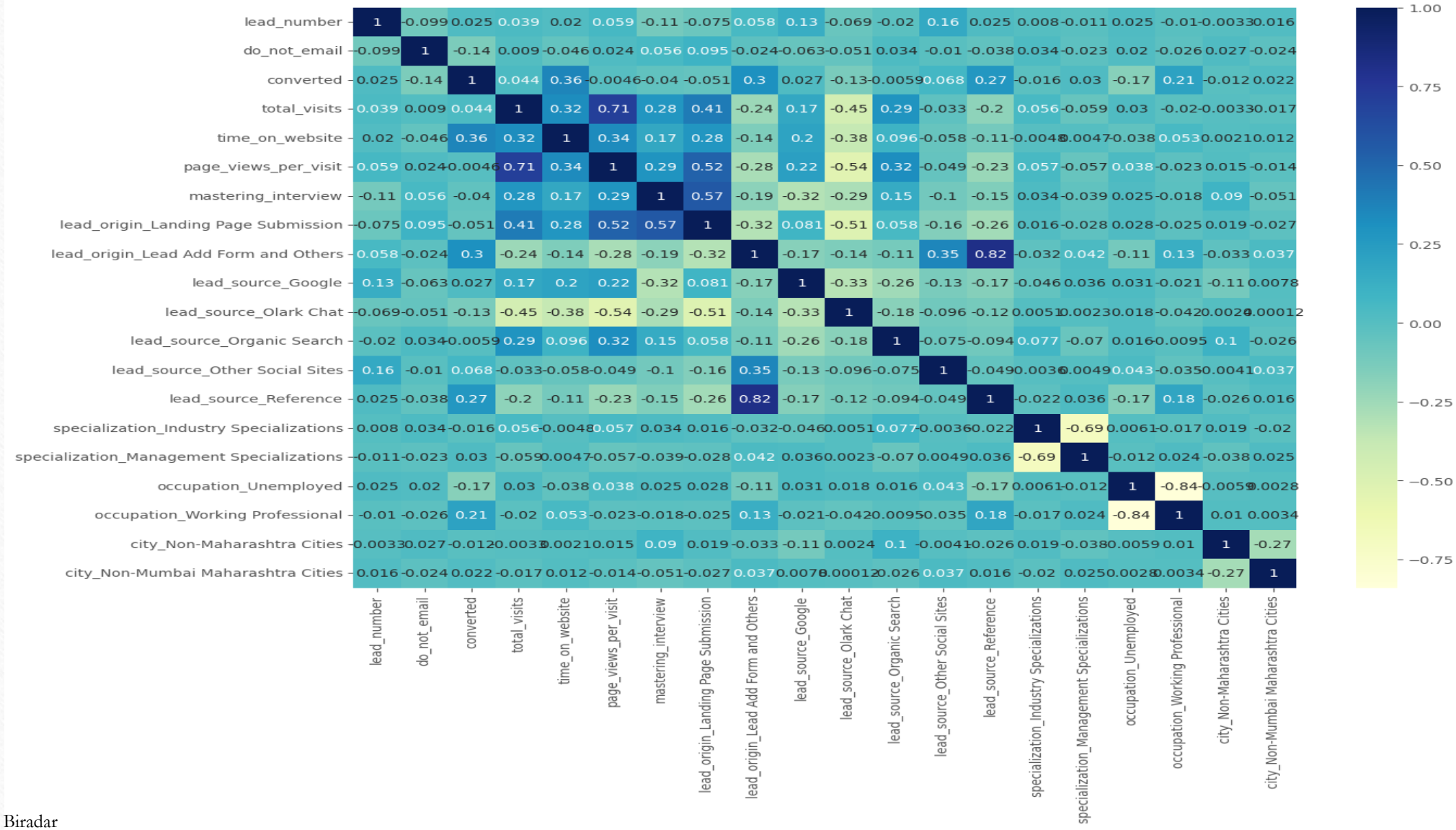
Mumbai in particular and Maharashtra in general dominates the lead. This is likely due to the fact that the courses are based in Mumbai



Model Performance: The model employing RFE-based feature selection showcased marginally superior performance compared to the model with all features. This underscores the potential for enhancing model accuracy and interpretability by excluding irrelevant features.

Chinmay Biradar

Results



Chinmay Biradar

Summary

Identifying High Potential Leads:

Features like 'Total Visits', 'Total Time Spent on Website', and 'Page Views Per Visit' were prioritized to identify leads with higher conversion probabilities.

Personalized Lead Nurturing Strategy:

Leads were segmented based on preferences and interests to provide personalized communication about relevant courses, services, and job opportunities.

Avoiding Non-Converting Leads:

Leads unlikely to convert, were excluded from outreach efforts.
Leads with the specialization labeled as "Others" and those who opted for "Do not Email" were also avoided to optimize resources and focus on more promising leads.

By following these guidelines, X-Education aims to enhance its lead conversion rate by targeting the most promising leads while avoiding outreach efforts that are less likely to result in conversions.

Chinmay Biradar