Lead Scoring Case Study

Identifying Potential Leads for X Education

X Education is an online education company that sells courses to industry professionals. However, their lead conversion rate is only around 30%, which means that most of the leads they acquire do not end up converting into paying customers. In order to make their sales process more efficient, the company wants to identify the most potential leads, also known as Hot Leads.

By identifying these hot leads, X Education can focus its sales efforts on communicating with the leads that are most likely to convert into paying customers. This will increase the efficiency of their sales process and ultimately lead to higher revenue for the company.

Business Objective

Lead X has approached us to build a lead scoring model that can assign a score between 0-100 to every lead. The purpose of this model is to help them identify the hot leads and increase their conversion rate. The CEO has set a target of achieving an 80% lead conversion rate.

The model should be able to handle future constraints such as peak time actions required, utilizing full manpower, and determining the approach after achieving the target. By implementing this model, Lead X hopes to streamline their sales process and achieve higher revenue.
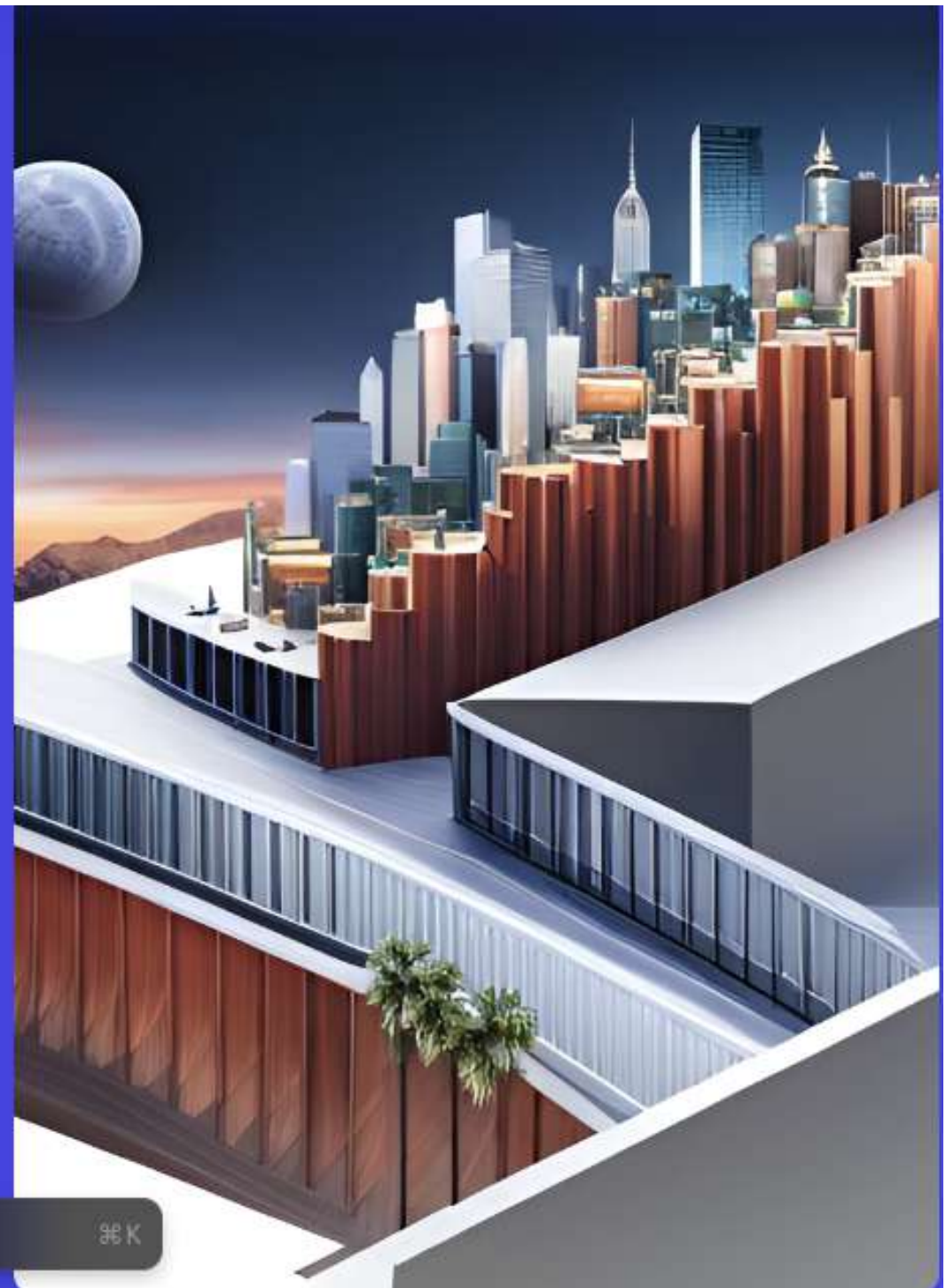
Problem Approach

The Lead Scoring Case Study presented a number of challenges that needed to be addressed in order to build an effective model. The first step was to import the data and inspect the data frame, identifying any issues with the data that needed to be corrected. Once the data was prepared, exploratory data analysis was conducted to identify patterns and relationships between different variables. Dummy variables were created to enable categorical variables to be included in the model, and a test-train split was performed to evaluate the model's performance on new data. Feature scaling was applied to ensure that all variables had equal weight in the model, and correlations between different variables were examined to identify any potential issues.

After completing these initial steps, the model building process began. Recursive feature elimination, R-squared values, variance inflation factors, and p-values were used to identify the most important variables for inclusion in the model. Once the model was built, it was evaluated using a range of metrics to assess its performance. Finally, predictions were made on the test set to determine how well the model would perform in real-world scenarios.

EDA – Data Cleaning

During the exploratory data analysis phase of the Lead Scoring Case Study, it was discovered that there were some columns in the data frame that contained a level called 'Select', which required special attention. This level was found to be taking care of some missing values in the data set. To avoid any bias in the model, these missing values were imputed with the median value of the respective column. Additionally, outliers were identified and removed from the data set to ensure accurate results.

The process of cleaning the data was crucial for ensuring the accuracy and effectiveness of the lead scoring model. By identifying and addressing any issues with the data, the model was able to make more accurate predictions and provide valuable insights into the lead generation process.

Leads from HR, Finance & Marketing management specializations are high probability to convert

In the Lead Scoring Case Study, it was found that leads from HR, Finance, and Marketing management specializations had a higher probability of converting. This insight was gained through analyzing the data and identifying patterns in the behavior of these leads. By focusing on these high-probability leads, businesses can optimize their lead generation process and increase their chances of success.

It is important for businesses to understand the characteristics of their most valuable leads. By identifying which types of leads are most likely to convert, businesses can tailor their marketing efforts and allocate resources more effectively. The Lead Scoring Case Study provides valuable insights into this process and highlights the importance of data analysis in optimizing lead generation strategies.

Count across Specialization
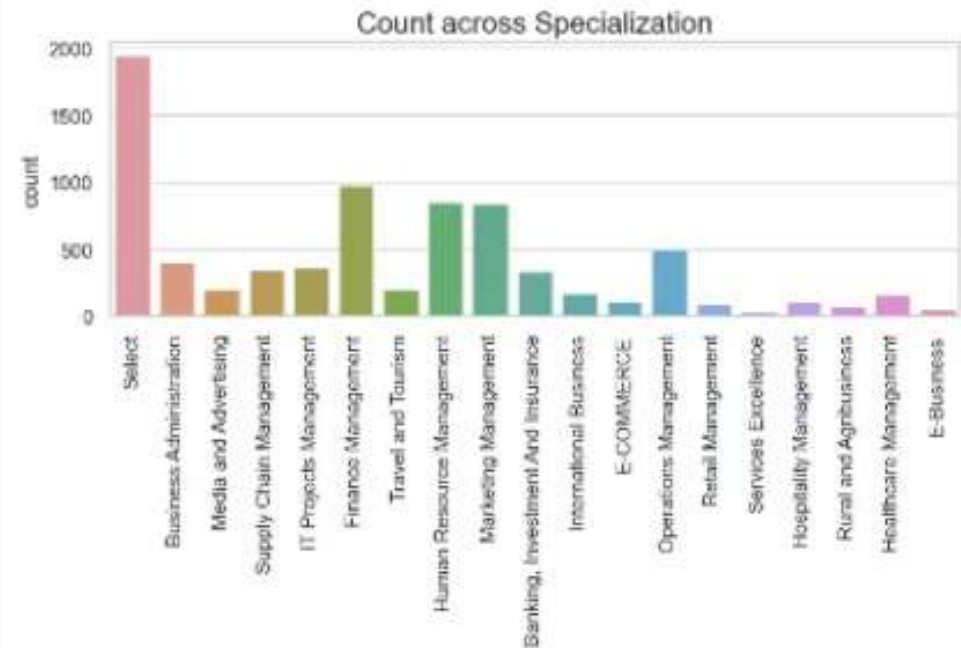
Lead Source & Lead Origin

In the Lead Scoring Case Study, it was observed that leads obtained through Google and direct traffic tend to have a higher probability of converting. This finding is based on an analysis of lead behavior patterns and can be used by businesses to optimize their lead generation process.

By understanding the source and origin of their leads, businesses can tailor their marketing efforts to better target high-probability leads. This can help them allocate resources more effectively and ultimately increase their chances of success.

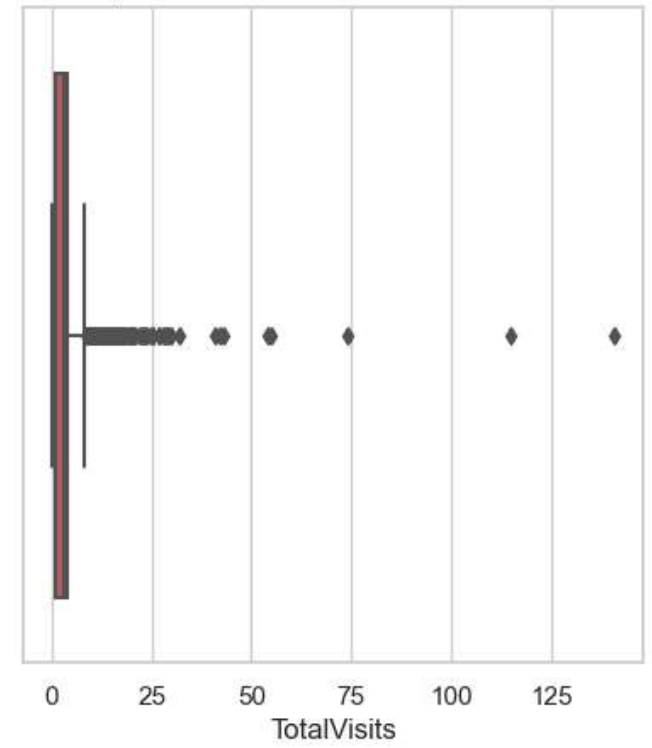Whereas in Lead origin most number of leads are landing on submission
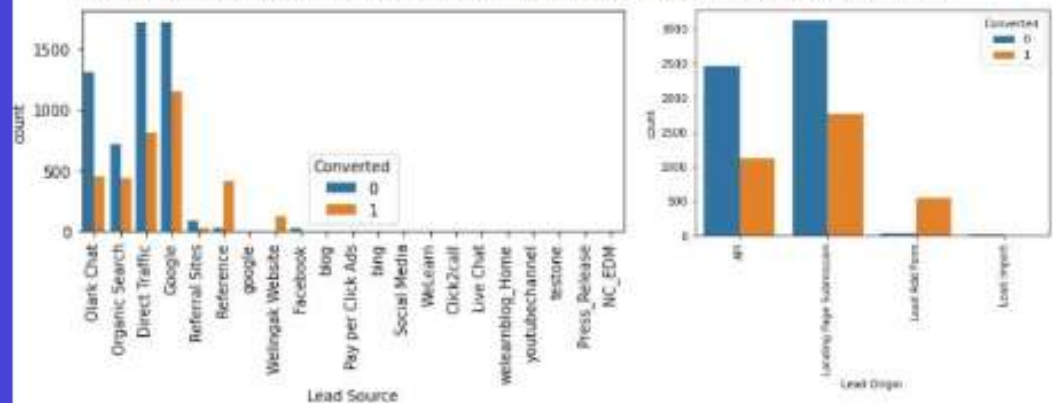
Boxplot for Total Visits
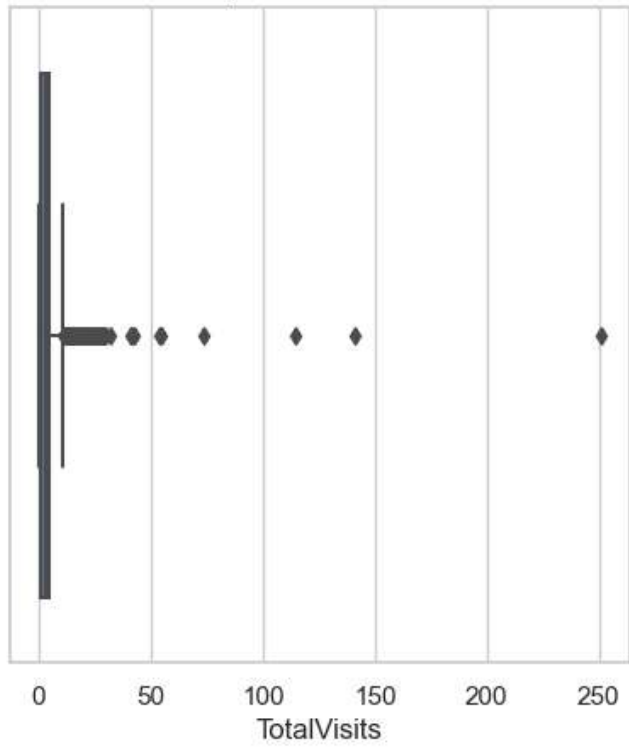
Boxplot for Total Visits with Conversion
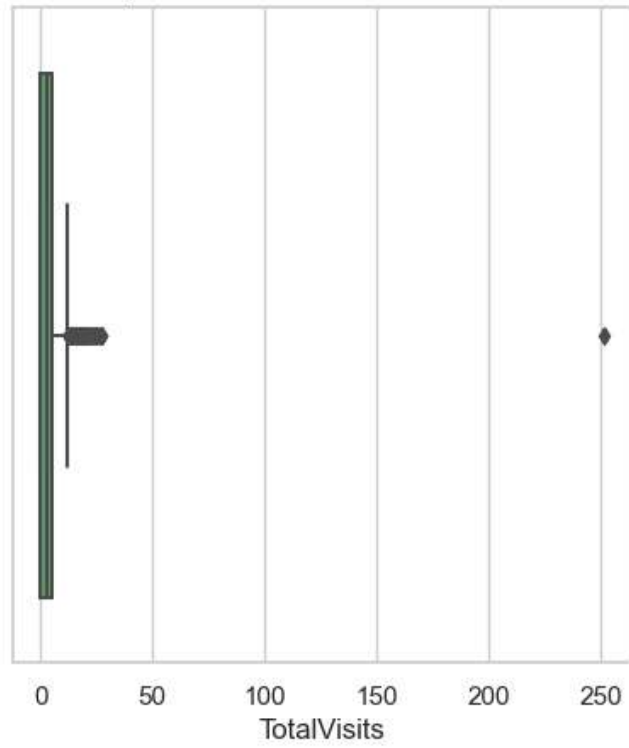
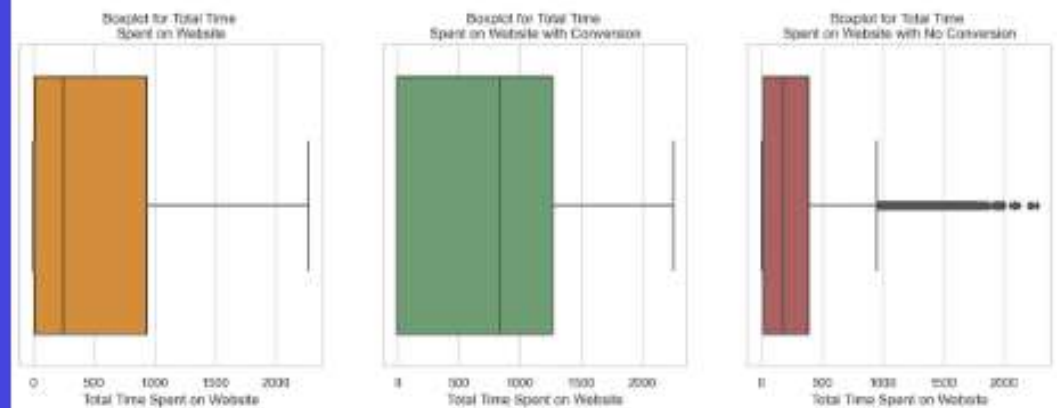Boxplot for Total Visits with No Conversion

Potential Outliers in No Conversion Class

In analyzing the lead scoring case study, we have identified some potential outliers within the 'No Conversion' class. However, upon further examination of the data at an aggregate level, it appears that the effect of this class is neutralized by the fact that the 'Conversion' class is far less skewed in comparison.

This suggests that while there may be some individual instances where the 'No Conversion' class has a significant impact, overall it does not have as much influence on the final outcome as initially anticipated.

Boxplot for Total Time Spent on Website

Boxplot for Total Time Spent on Website with Conversion

Boxplot for Total Time Spent on Website with No Conversion

Fixing Outliers in the Lead Scoring Case Study

After analyzing the data, we have identified some extreme values that need to be fixed before proceeding with our analysis. These outliers can significantly impact the final outcome and may lead to inaccurate results if left unaddressed.

To fix these outliers, we will need to examine each instance individually and determine whether it is a legitimate data point or an error. Once we have identified the outliers, we can then decide how to handle them - either by removing them from the dataset or by adjusting their values to better fit within the overall distribution.
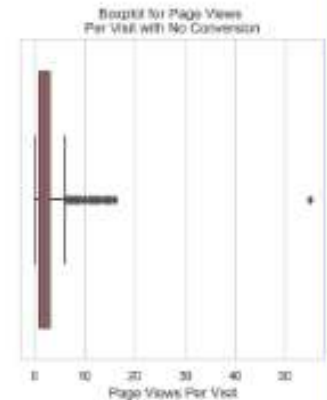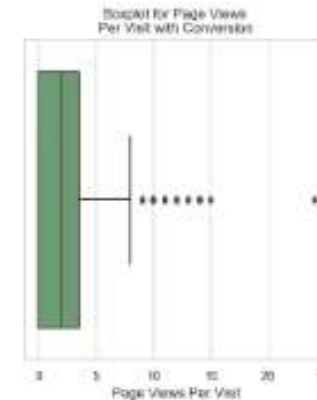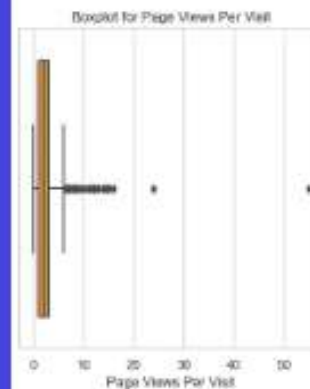
ROC curve

0.42 is the tradeoff between
Precision and Recall -
Thus we can safely choose to
consider any Prospect Lead with
Conversion Probability higher
than 42 % to be a hot Lead
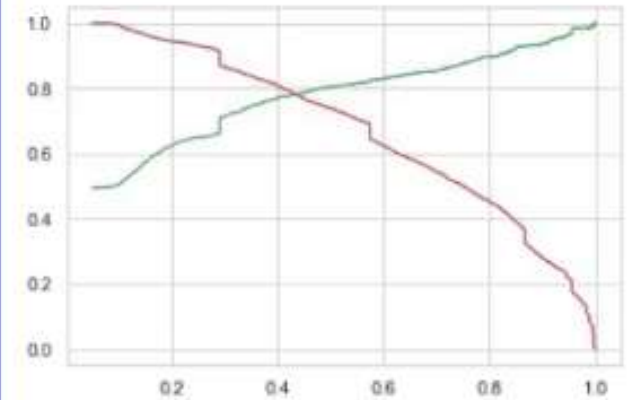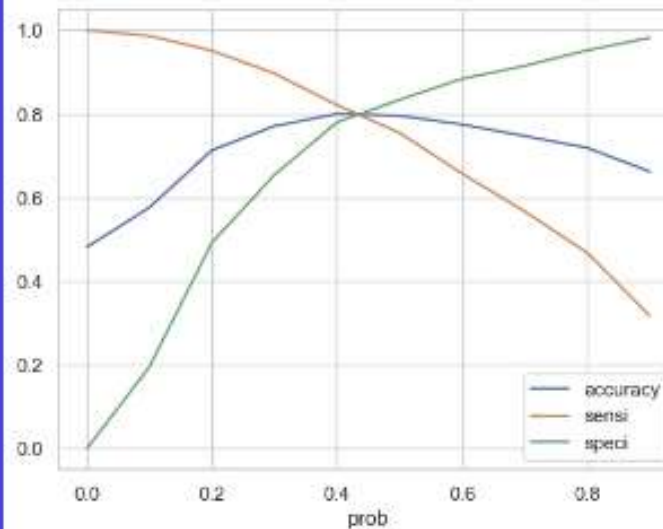
Last Lead Activity

The behavior of leads can be analyzed to determine their likelihood of converting into customers. One important factor to consider is the last activity of the lead. If a lead has recently opened an email, they are more likely to convert than leads who have not engaged with any content. This information can be used to prioritize follow-up efforts and focus on leads who are most likely to convert.

Sending SMS messages can also be an effective way to engage with leads and increase the likelihood of conversion. Similar to email opens, leads who respond to SMS messages are more likely to become customers. By tracking these activities and incorporating them into lead scoring models, businesses can optimize their sales efforts and improve overall conversion rates.
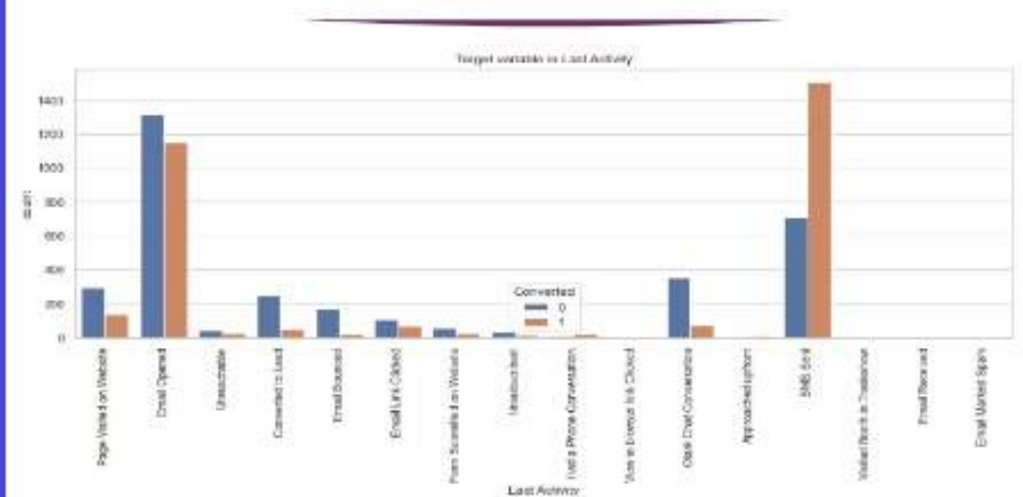
Train and Test Data Accuracy, Sensitivity, and Specificity

The train data accuracy, sensitivity, and specificity for the lead scoring case study are at 78%, 77%, and 80% respectively. The test data also has the same values for accuracy, sensitivity, and specificity. These numbers indicate that the lead scoring model is performing well, with a high level of accuracy and precision in predicting which leads are most likely to convert into customers.

Based on these results, the final features list for the lead scoring model includes Lead Source_Olark Chat, Specialization_Others, Lead Origin_Lead Add Form, Lead Source_Welingak Website, Total Time Spent on Website, Lead Origin_Landing Page Submission, What is your current occupation_Working Professionals, and Do Not Email. By incorporating these features into the model, businesses can better prioritize their follow-up efforts and focus on leads who are most likely to convert.