Lead Scoring Case Study 2023

Kunal Das
Ananth Ram
Rubina D’souza

Problem Statement

- X Education sells online courses to industry professionals. The company markets its courses on several websites and search engines like Google.
- Once these people land on the website, they might browse the courses or fill up a form for the course or watch some videos. When these people fill up a form providing their email address or phone number, they are classified to be a lead. Moreover, the company also gets leads through past referrals.
- Once these leads are acquired, employees from the sales team start making calls, writing emails, etc. Through this process, some of the leads get converted while most do not. The typical lead conversion rate at X education is around 30%.

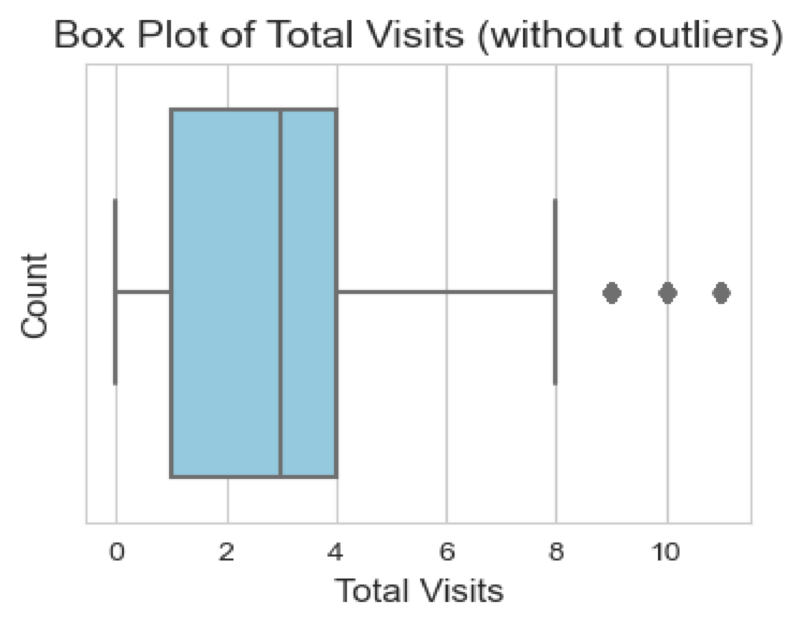
Goal

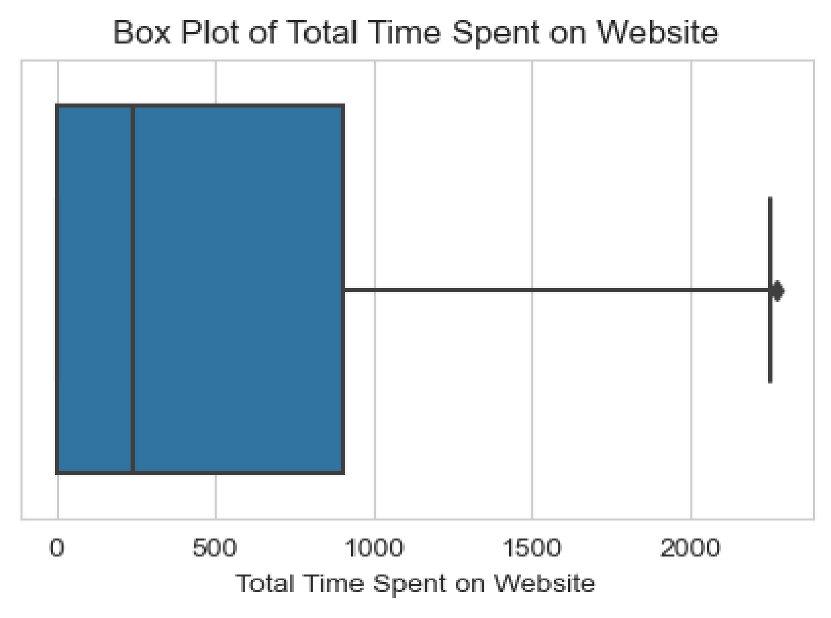
- X Education needs help in selecting the most promising leads, i.e. the leads that are most likely to convert into paying customers.
- The company needs a model wherein you a lead score is assigned to each of the leads such that the customers with higher lead score have a higher conversion chance and the customers with lower lead score have a lower conversion chance.
- The CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80%

Problem Solving Methodology

- ❑ Source Data for analysis
- ❑ Data Pre-processing: Data cleaning, Data Manipulation
- ❑ EDA: Univariate Data Analysis, Bivariate Data Analysis
- ❑ Feature Scaling
- ❑ Model Building: Logistic regression Model
- ❑ Model Training
- ❑ Model Evaluation
- ❑ Model Performance
- ❑ Predictions
- ❑ Conclusion

Exploratory Data Analysis

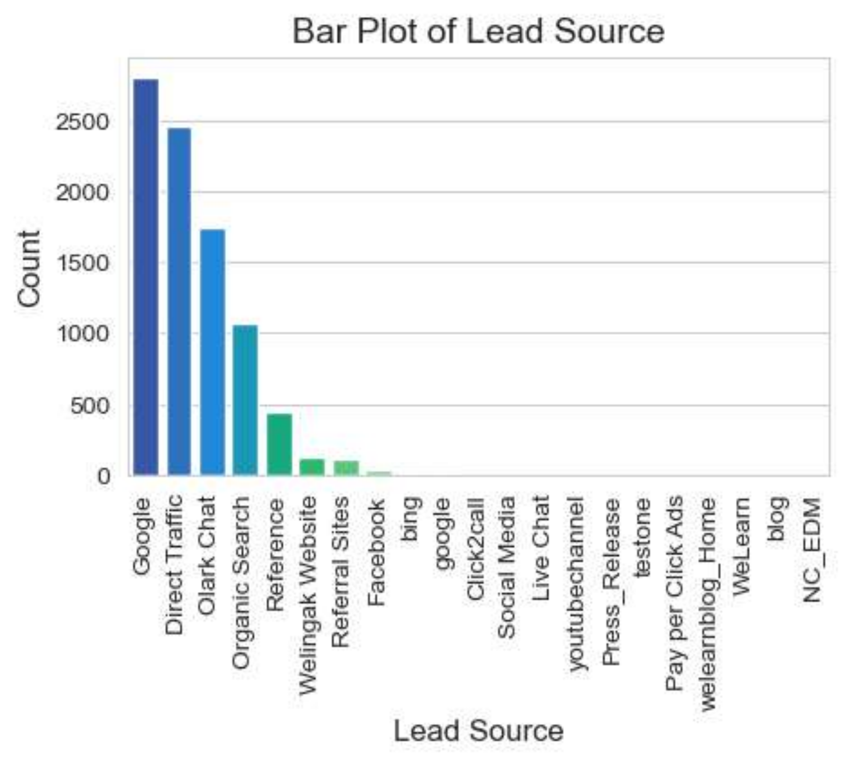




The box ranges from approximately 1 to 4, indicating that the majority of the Total Visits values fall within this range. The median, which is located at around 3, suggests that 50% of the Total Visits values are below this point and 50% are above.

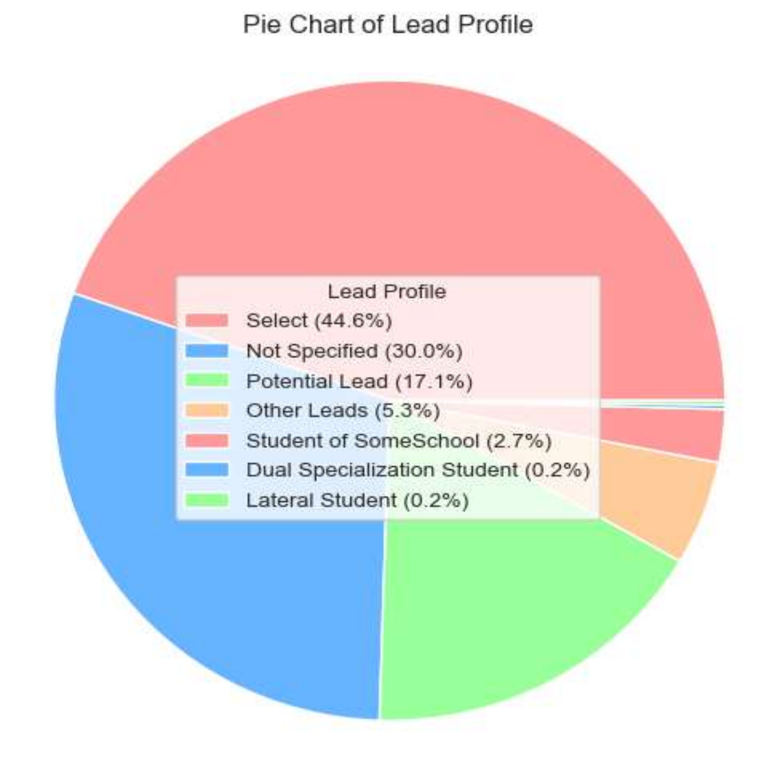
The box plot suggests that the majority of the values for 'Total Time Spent on Website' range from approximately 0 to 800, with the median value being around 250.

Lead Source



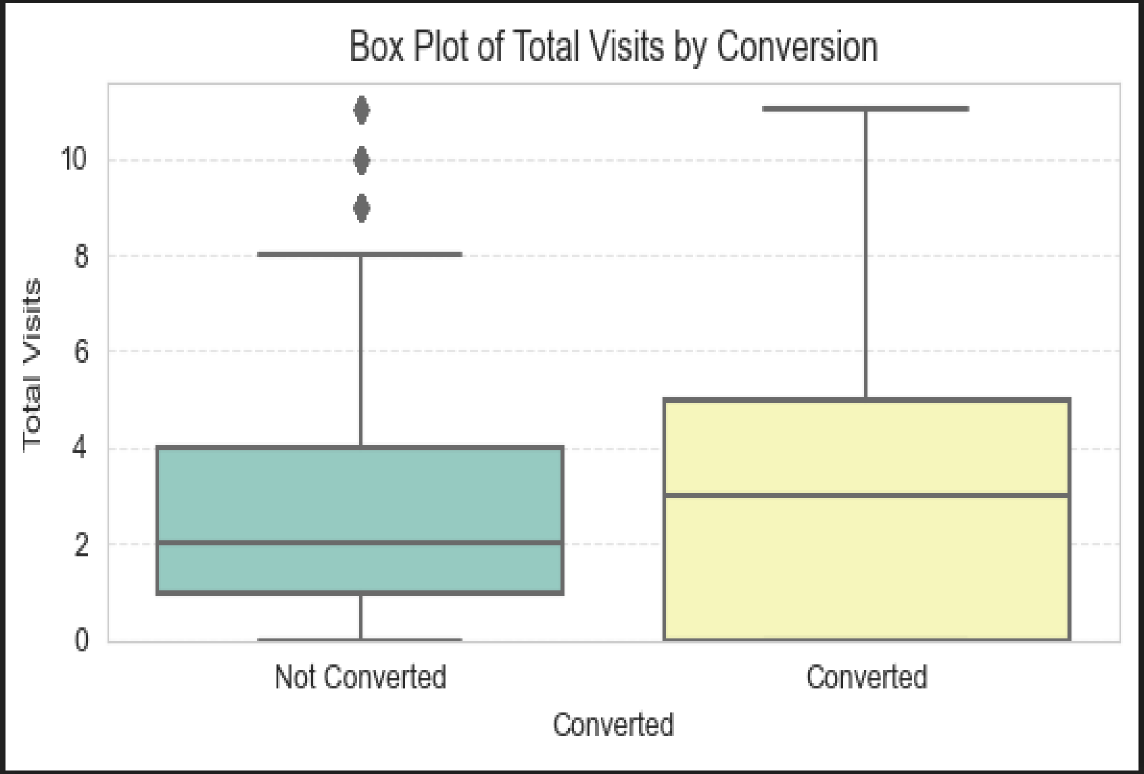
We can observe most of the traffic coming from Google or direct

Lead Profile



The majority of leads have a profile labeled as "Select," indicating incomplete information.

Box Plot - Conversion



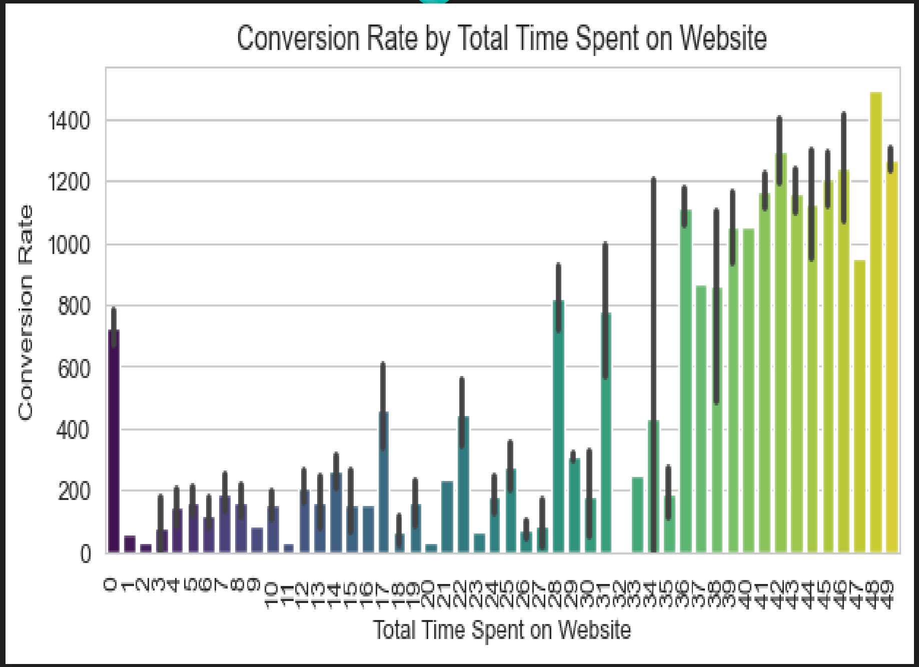
❑ The median of the 'Converted' group is slightly higher than the 'Not Converted' group, it suggests that on average, the 'Total Visits' tends to be slightly higher for the converted leads compared to the non-converted leads.

❑ The vertical extent of the box indicates the range of values where the majority of the data points lie. For the 'Converted' group, the box extends from approximately 0 to 4.5, while for the 'Not Converted' group, the box extends from approximately 1 to 4.

❑ This means that the range of 'Total Visits' values for the converted leads is slightly wider than the range for the non-converted leads.

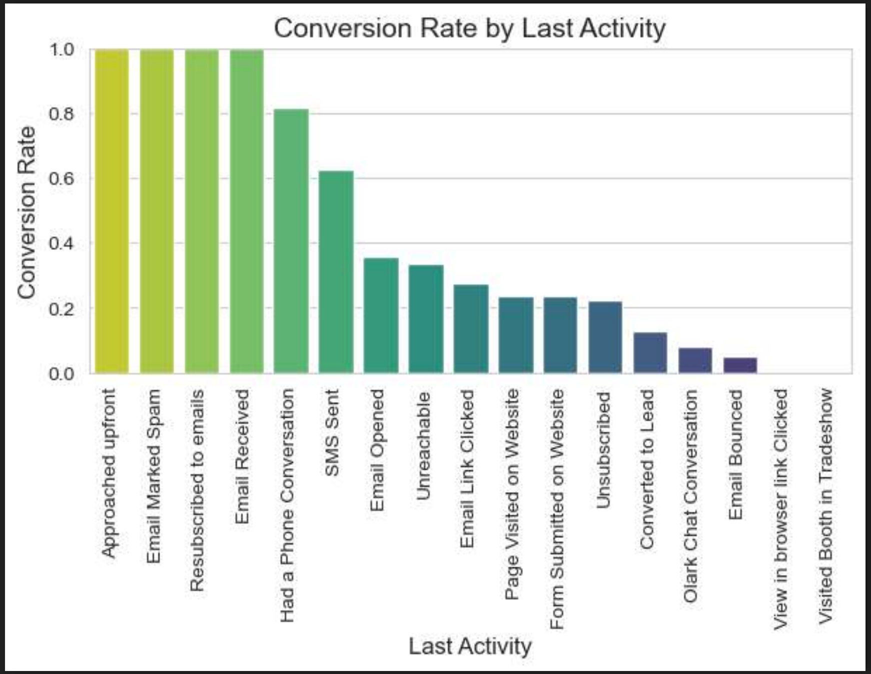
Conversion Rate Graph

Total Time



We can observe the conversion rate increases significantly for the users who spends more time on the website.

Last Activity

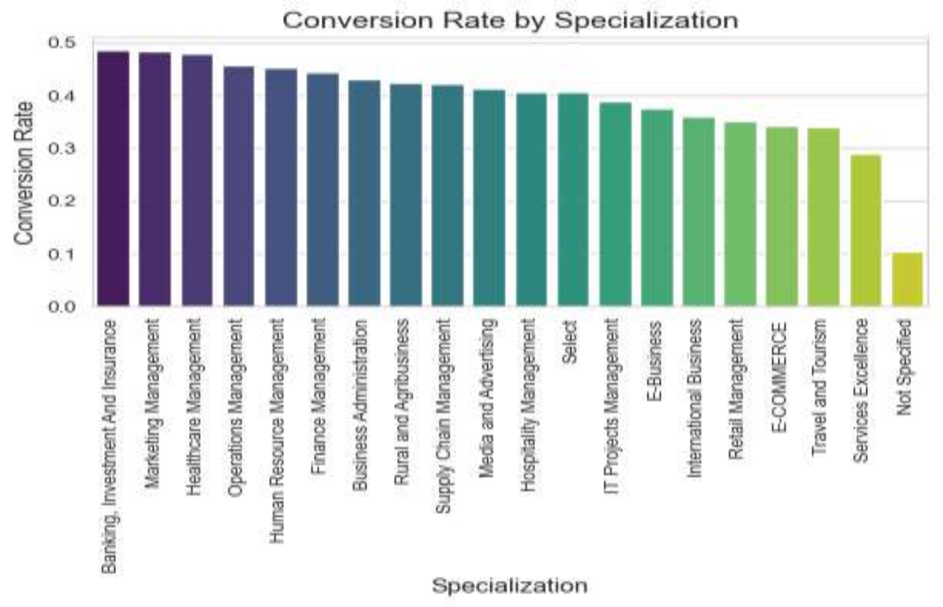


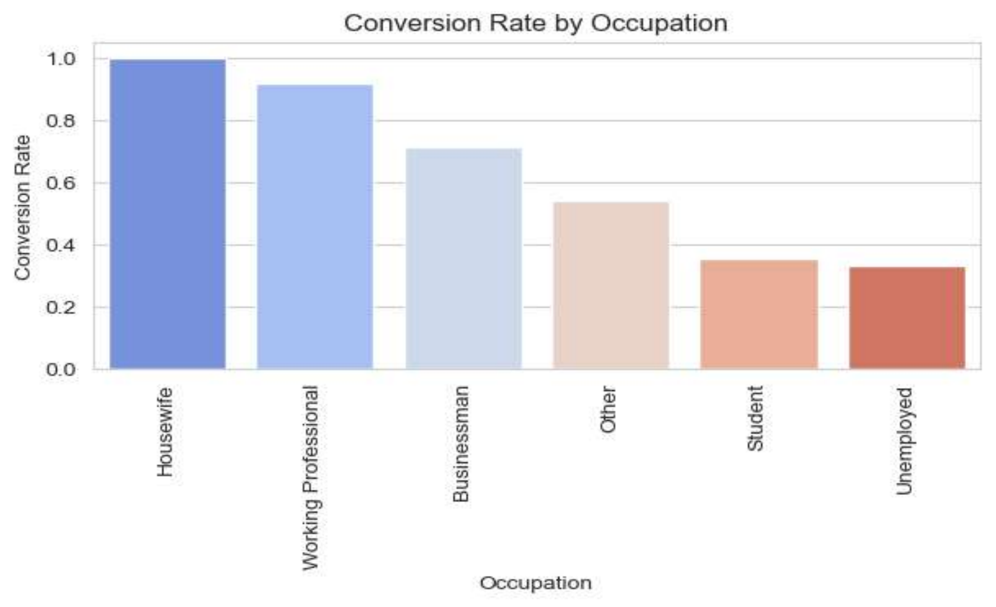
The higher conversion rate for activities like "Approached upfront," "Email Marked Spam," "Resubscribed to emails," and "Emails Received" suggests that these activities have a positive impact on the likelihood of conversion. It indicates that leads who engage in these activities are more likely to be converted into customers.

Conversion Rate Graph

Specialization

Occupation

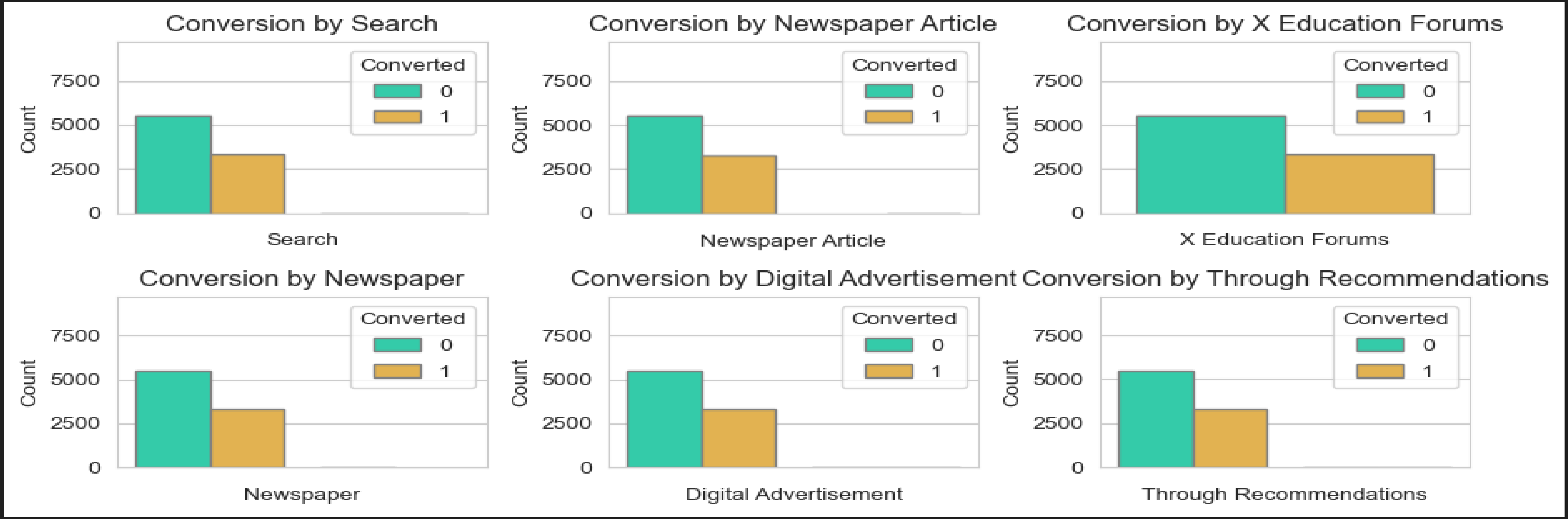




We can see that most of the specialization has >30% conversion rate while some has as high as 50% conversion rate.

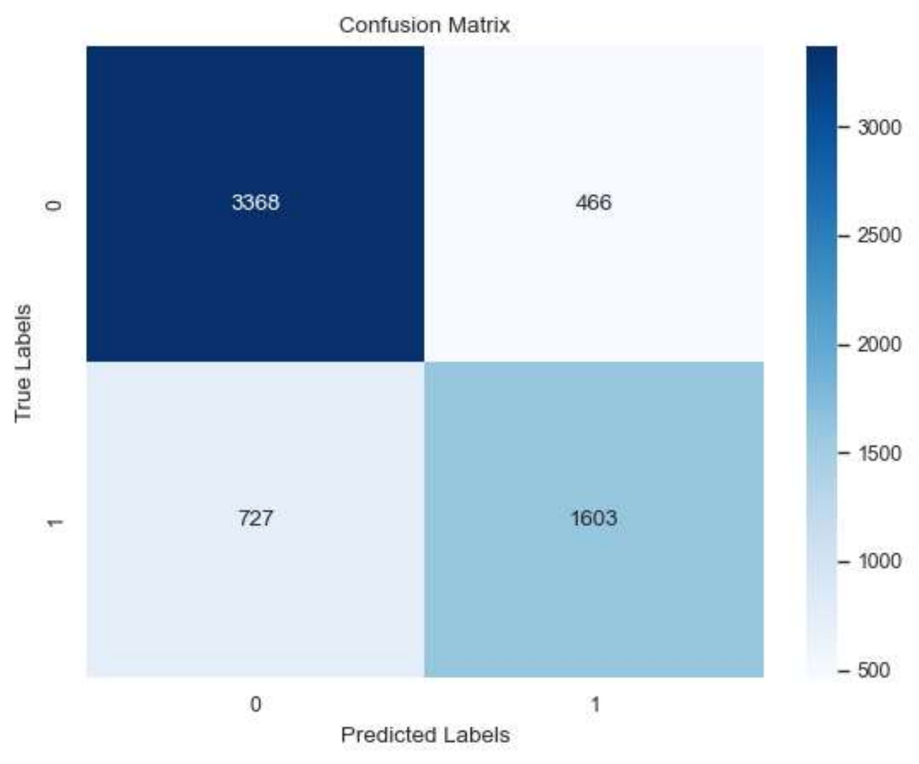
Housewives working professionals has pretty high conversion rates.

Conversion rates - Others



Conversion rates from across all the portals are more or les similar.

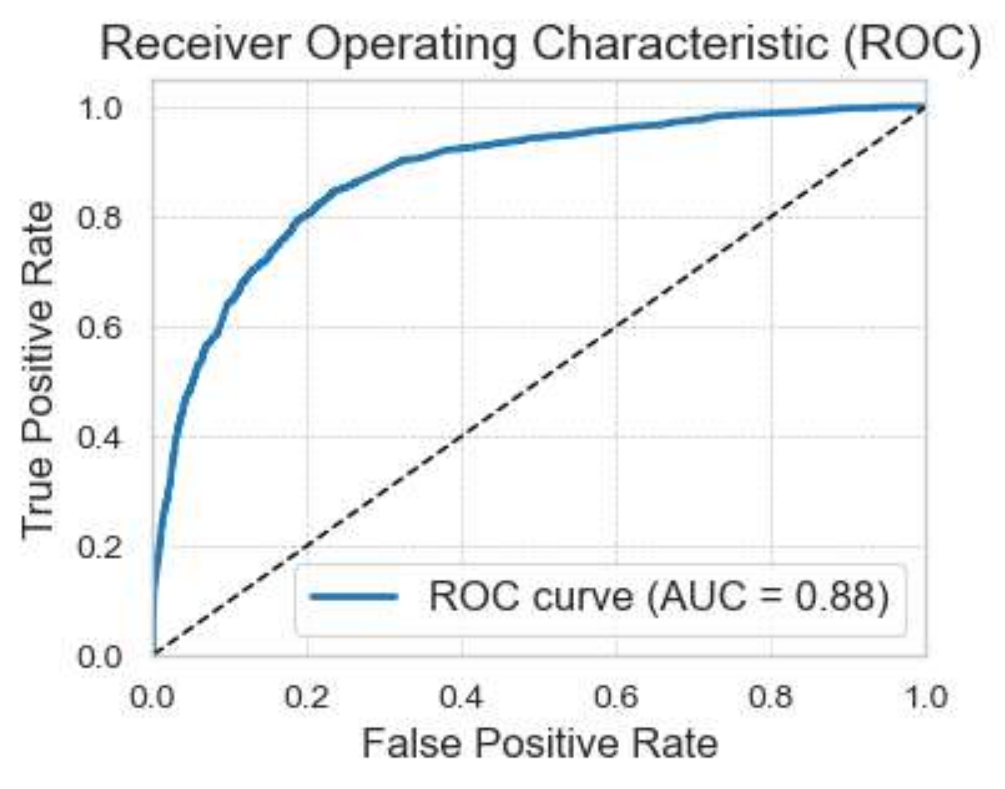
Model Evaluation



Heat Map

- ❑ The confusion matrix provides information about the performance of a binary classification model. In this case, the model has correctly predicted 3368 instances of the positive class (actual converted flag) and 1603 instances of the negative class (actual not converted flag).
- ❑ However, it has made 466 false positive predictions (predicted as converted, but actually not converted) and 727 false negative predictions (predicted as not converted, but actually converted).

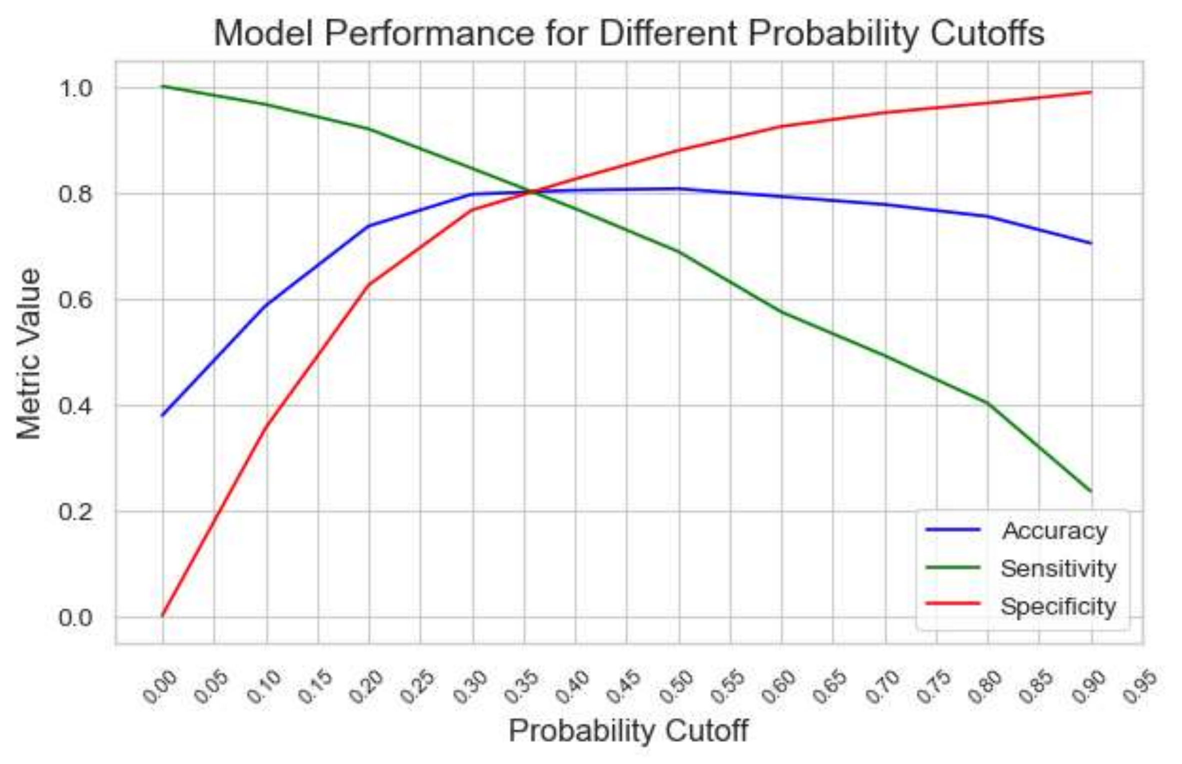
Receiver Operating Characteristic- ROC



- ❑ An ROC AUC (Receiver Operating Characteristic Area Under the Curve) of 0.88 indicates that the model has good discriminatory power in distinguishing between positive and negative instances.

- ❑ The higher the ROC AUC value (ranging from 0 to 1), the better the model's ability to correctly classify instances. In this case, an ROC AUC of 0.88 suggests that the model performs well in predicting the conversion flag.

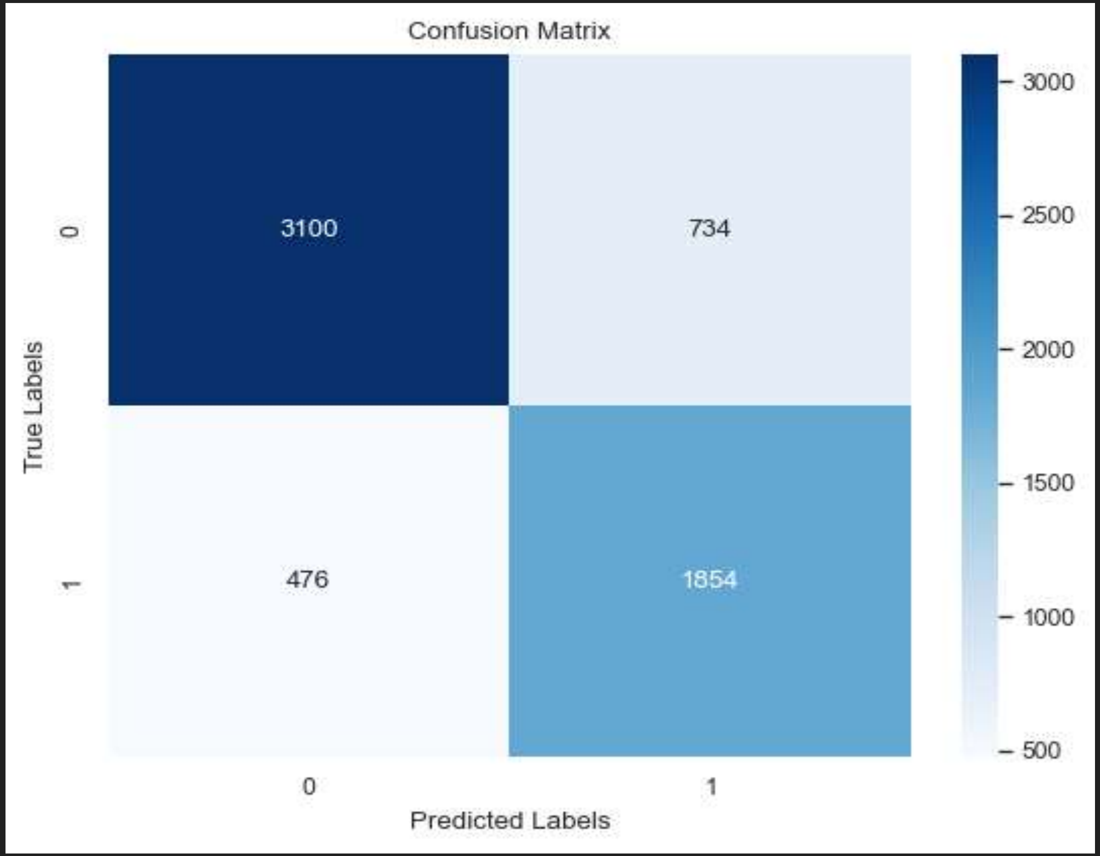
Model Performance



- ❑ We can observe that the optimal cutoff point is approximately at 0.37. At this threshold, the model achieves a balanced trade-off between accuracy, sensitivity, and specificity.

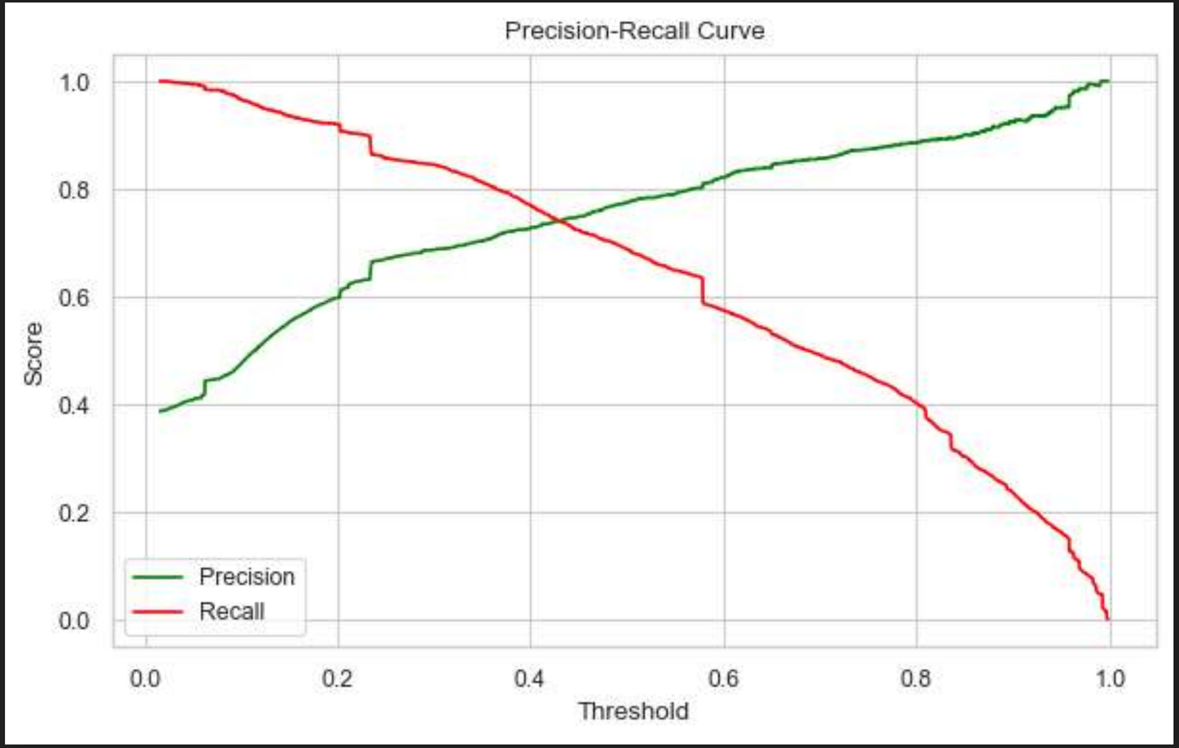
- ❑ This means that the model can effectively classify both positive and negative instances without favoring one over the other.

Binary Classification Model



- ❑ The confusion matrix you provided is a 2x2 matrix representing the performance of a binary classification model. Here's the interpretation of each element in the matrix:
- ❑ True Negative (TN): 3100 - The number of observations that are actually negative and correctly predicted as negative.
- ❑ False Positive (FP): 734 - The number of observations that are actually negative but incorrectly predicted as positive.
- ❑ False Negative (FN): 476 - The number of observations that are actually positive but incorrectly predicted as negative.
- ❑ True Positive (TP): 1854 - The number of observations that are actually positive and correctly predicted as positive.

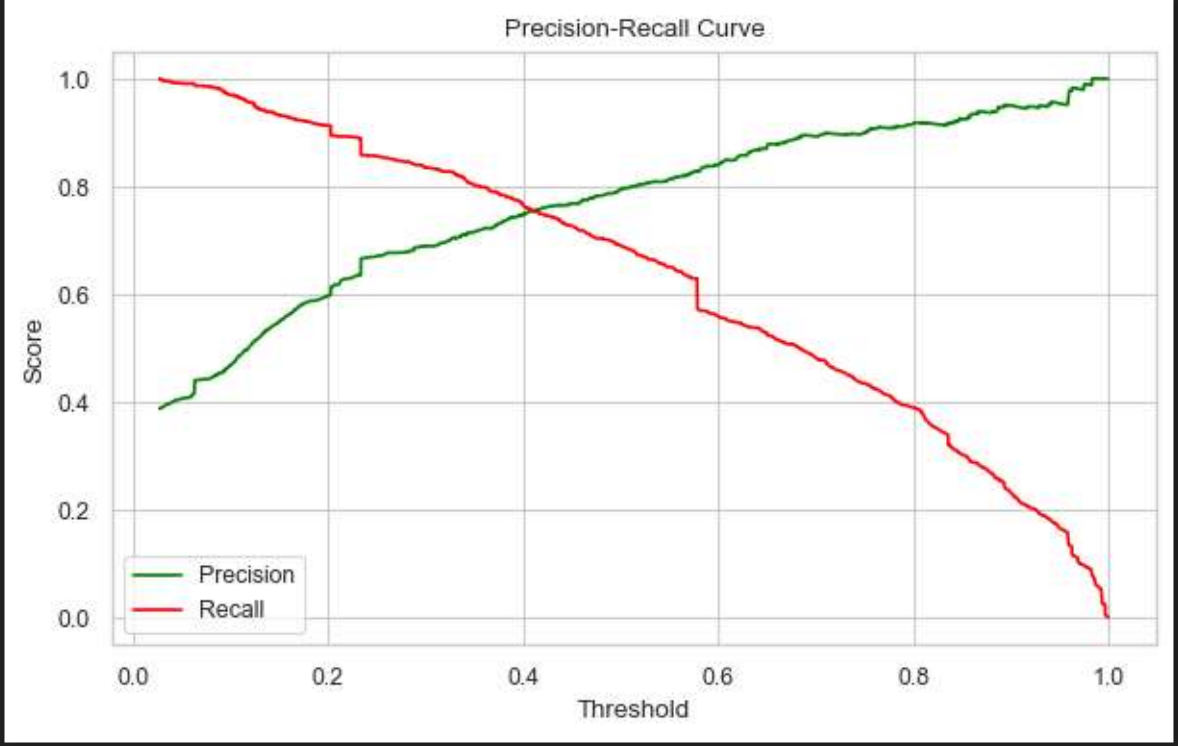
Precision and Recall View



- ❑ At the threshold score of 0.4, the precision and recall values are both relatively high, indicating that the model is able to correctly classify a significant number of positive instances (high recall) while maintaining a relatively low number of false positive predictions.

- ❑ The intersection of the precision and recall curves at a score of 0.7 and threshold of 0.4 suggests a balanced performance of the model in terms of correctly identifying positive instances while minimizing false positives.

Prediction



- ❑ The precision and recall curves intersect at approximately (0.8, 0.4). This means that at a threshold of 0.8, we can achieve a precision of 0.8 and a recall of 0.4.

- ❑ This intersection point represents a threshold value where precision and recall are relatively balanced.

Conclusion

- ❑ **Based on the analysis and results obtained from the logistic regression model, here are some suggestions that can be provided to the customer:**
- ❑ Lead Scoring: Implement the lead scoring system based on the logistic regression model. Assign lead scores between 0 and 100 to prioritize and target potential leads effectively. This will help in identifying hot leads with higher conversion probabilities and allocating appropriate resources for conversion efforts.
- ❑ Focus on Conversion Factors: Identify the key factors that contribute to lead conversion based on the logistic regression model's coefficients. Pay special attention to these factors when designing marketing campaigns, website content, and communication strategies. For example, factors such as total visits, total time spent on the website, and page views per visit have shown significance in predicting lead conversions.
- ❑ Personalized Communication: Utilize the information gathered during lead capture, such as lead origin, lead source, last activity, and lead profile, to personalize communication with leads. Tailor marketing messages and content based on the lead's specific needs and preferences, increasing the likelihood of engagement and conversion.
- ❑ Improve Website Experience: Enhance the user experience on the website to increase engagement and time spent by leads. Optimize page load times, simplify navigation, and provide relevant and valuable content to encourage leads to explore more pages and increase the chances of conversion.
- ❑ Analyze Lead Quality: Use the lead quality information provided by the logistic regression model to identify leads that are more likely to convert. Focus efforts on high-quality leads and allocate resources accordingly. Continuously evaluate and update lead quality metrics to refine targeting strategies.
- ❑ Referral Programs: Leverage the "Through Recommendations" feature to encourage satisfied customers to refer others to X Education. Implement referral programs or incentives to motivate customers to recommend the courses to their networks, expanding the reach and potential conversion opportunities.
- ❑ Continuous Model Monitoring and Improvement: Regularly monitor the performance of the logistic regression model and assess its predictive accuracy. Incorporate new data and retrain the model periodically to ensure it remains up-to-date and maintains its effectiveness in lead scoring.

These suggestions aim to optimize lead conversion strategies by leveraging the insights and predictions provided by the logistic regression model. By implementing these recommendations, X Education can enhance its marketing and sales efforts, increase conversion rates, and improve overall business outcomes.