

LEAD SCORING CASE STUDY USING LOGISTIC REGRESSION

HIEU HAN HAN

PROBLEM STATEMENT



X Education, an online education company, attracts professionals to its website through various marketing channels. Upon landing on the site, visitors may browse courses, fill out forms, or watch videos, thus becoming leads.



Leads are also obtained through referrals. The sales team engages with these leads through calls and emails, aiming for conversion, with a typical rate of around 30%.



Despite a high volume of leads, the conversion rate remains low. To improve efficiency, the company seeks to identify 'Hot Leads' with the highest conversion potential. They aim to increase the conversion rate by focusing efforts on these promising leads.



The task is to build a model assigning lead scores, and prioritizing leads with higher scores for increased conversion chances. The CEO's target conversion rate is around 80%.

BUSINESS OBJECTIVE

Create a logistic regression model assigning lead scores from 0 to 100 (indicating conversion likelihood)

Higher scores represent 'hot' leads with higher conversion potential

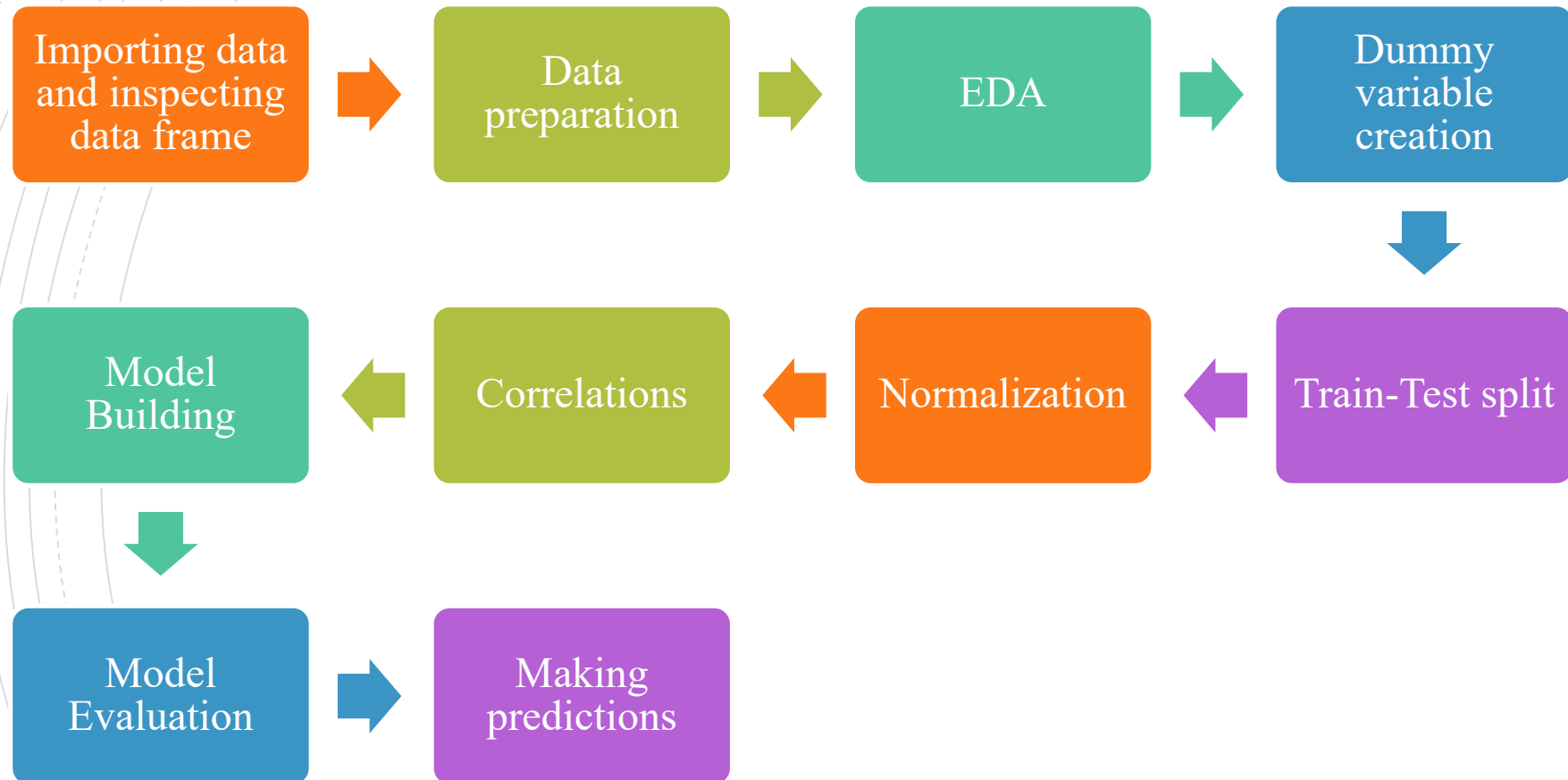
Optimize lead targeting efforts, maximizing sales team productivity

Implementing this model aligns with X Education's goal of improving lead conversion efficiency

Supports sustaining growth in the online education market

Achieve of lead conversion rate of 80%

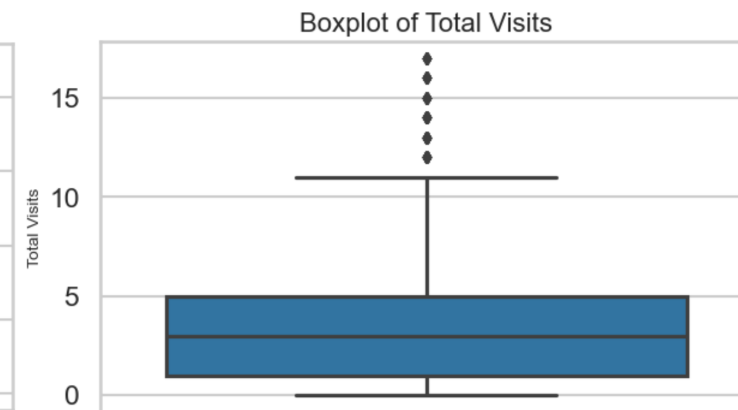
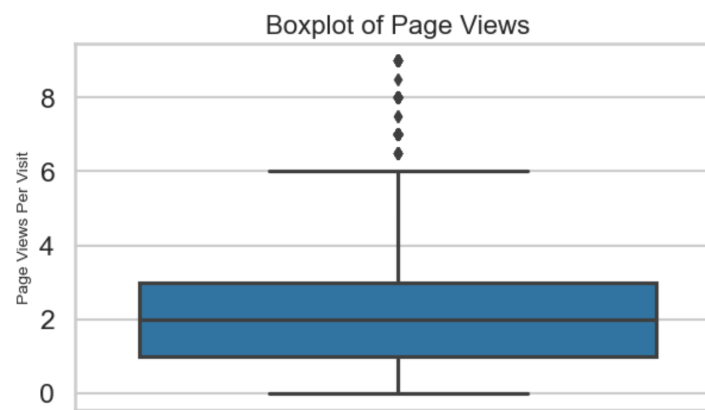
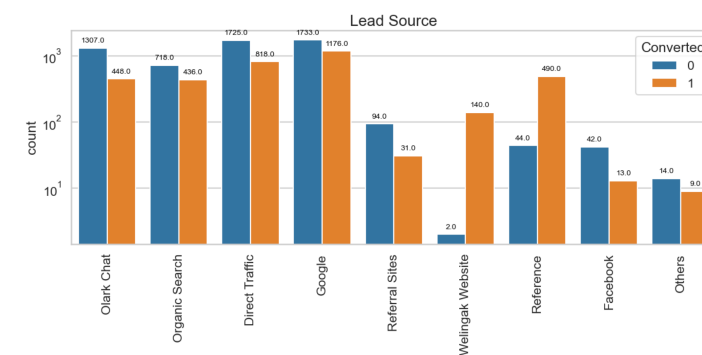
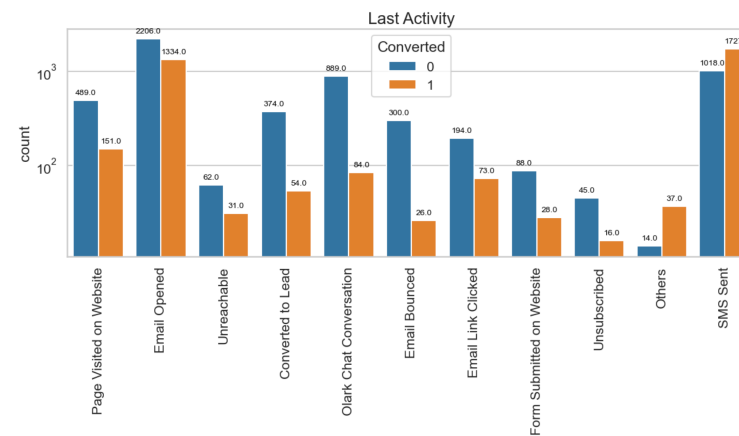
PROBLEM APPROACH



EDA

A lot of elements in the categorical variables were irrelevant

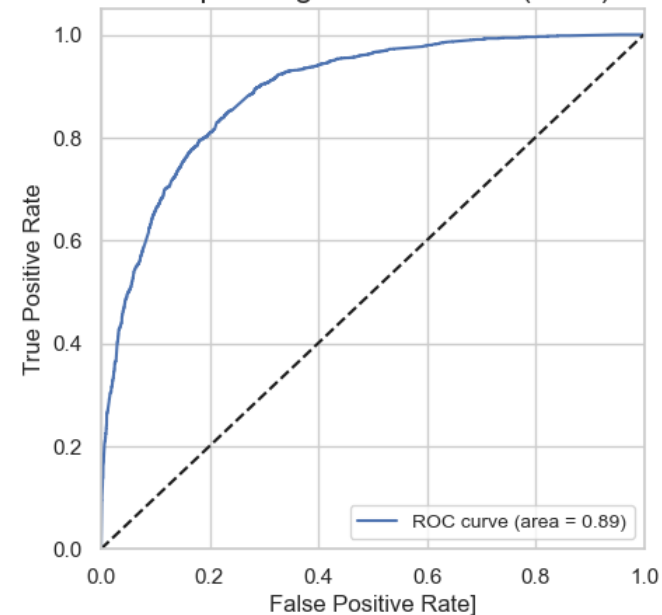
The numeric values seem good and no outliers were found



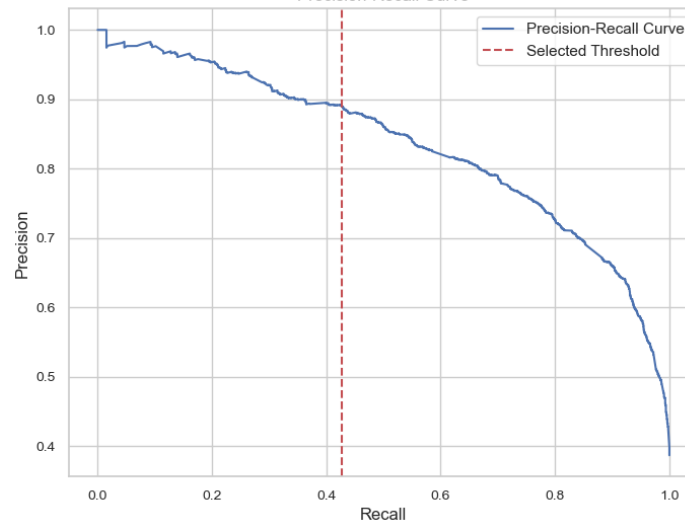
MODEL BUILDING

Recursive Feature Elimination (RFE) was used to identify the top 15 relevant variables. The remaining variables were manually eliminated based on their Variance Inflation Factor (VIF) values and p-values, ensuring the model's focus on the most influential predictors.

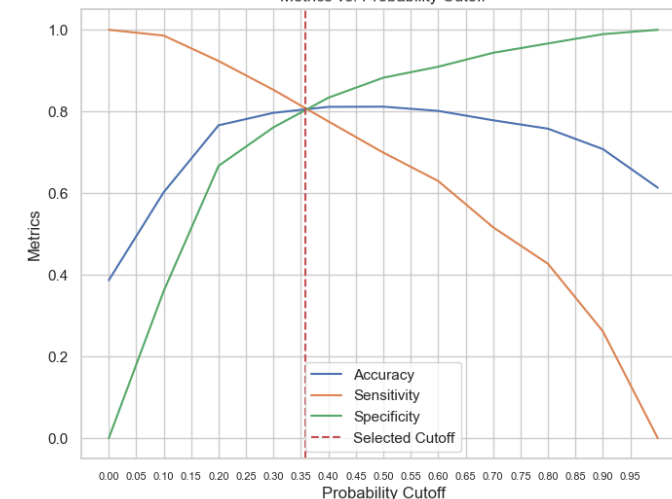
Receiver Operating Characteristic (ROC) Curve



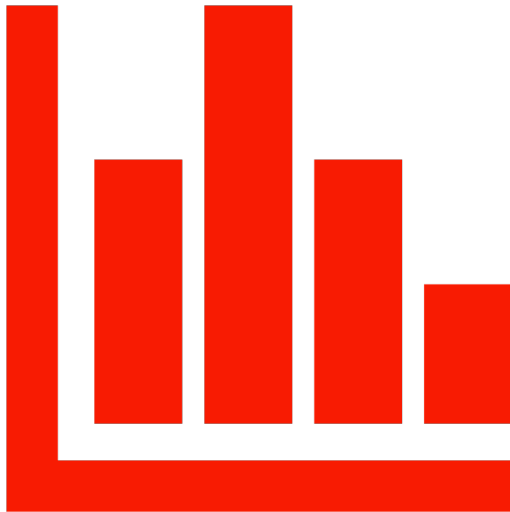
Precision-Recall Curve



Metrics vs. Probability Cutoff




PREDICTION




Using an optimal cutoff of 0.35, our model made predictions on the test dataset, achieving notable performance metrics. With an accuracy, precision, and recall rate all standing at 80%, ensuring reliable lead classification and targeted outreach.

EVALUATION

The model's performance was evaluated using a confusion matrix, which helped analyze its predictive accuracy. Additionally, the optimal cutoff point was determined using the ROC curve.



Accuracy, recall, and precision metrics were computed, yielding an overall performance of approximately 80%.



these evaluation metrics offer a comprehensive understanding of the model's performance and its suitability for lead classification and targeting.

RECOMMENDATION



Lead Source - Welingak Website: The conversion rate is notably higher for leads originating from the 'Welingak Website', making it a focal point for attracting potential leads.



Lead Origin - Lead Add Form: Leads generated through the 'Lead Add Form' exhibit a higher conversion rate, making it imperative for the company to emphasize this method to increase lead acquisition.



Current Occupation - Working Professional: Leads identified as 'Working Professionals' demonstrate a higher likelihood of conversion, prompting the company to concentrate efforts on targeting this demographic for lead generation.



Last Activity - SMS Sent: Leads whose last recorded activity is receiving an SMS have shown potential for conversion, making them a priority for engagement strategies.



Total Time Spent on Website: Leads spending more time on the website exhibit characteristics of potential conversion, underscoring the importance of engaging and retaining visitors on the platform.