

Lead Scoring Case Study

Summary Report

The analysis conducted for X Education aimed at optimizing strategies to attract more industry professionals to enrol in their courses. The initial dataset provided comprehensive information regarding customer site visits, duration spent on the site, sources of site access, and the conversion rate.

A lead scoring case study utilizing a logistic regression model was conducted to align with business requirements. The majority of leads are at an initial stage, with a relatively low conversion rate to paying customers. The predominant lead source is India, particularly from Mumbai. To enhance data quality, columns with the 'Select' level were adjusted for compulsory selection, including Customer Occupation and Specialization.

Higher total time spent on the platform seems to correlate the most with increased chances of lead conversion. Leads expressing interest in courses for better career prospects, especially those specializing in Finance Management, show a higher probability of conversion. Notably, leads from HR, Finance, and Marketing Management specializations demonstrate a strong likelihood to convert.

Improving customer engagement, particularly through emails and calls, is crucial. Leads engaging with emails exhibit a higher probability of conversion. Sending SMS can also be beneficial. Unemployment is prevalent among leads, suggesting a focus on strategies tailored to unemployed individuals for increased conversion rates.

The analytical process involved several key steps. In the data cleaning phase, efforts were made to address null values and replace the "select" option with null values, maintaining data integrity. Geographical data was standardized, categorizing elements into 'India', 'Outside India', and 'n/a.'

Exploratory Data Analysis (EDA) provided insights into the dataset's condition, revealing irrelevant elements in categorical variables while confirming the absence of outliers in numeric values.

Binary mapping was done and dummy variables were created for the categorical variables. The MinMaxScaler was applied for numeric values. The dataset was then split into 70% for training and 30% for testing.

Model building involved Recursive Feature Elimination (RFE) to identify the top 15 relevant variables. After observing the initial result, the number of relevant

variables dropped to 6 after RFE and Manual removal based on p-values. Remaining features have $VIF < 5$ and $p\text{-value} < 0.05$.

The model was evaluated using a confusion matrix, and an optimal cutoff value, determined through ROC curve analysis, yielded around 85% accuracy, sensitivity, and specificity. Predictions on the test dataset using a cutoff of 0.3 resulted in similar performance. Precision-recall analysis revealed a cutoff of 0.3 with approximately 80% precision and 82% recall on the test dataset.

The analysis identified critical variables influencing potential buyers, including total time spent on the website, total number of visits, specific lead sources, last activity, lead origin, and the current occupation as a working professional.

Leveraging these insights can significantly enhance X Education's success in attracting and converting potential buyers for their courses, providing a roadmap for targeted marketing and engagement strategies. The emphasis on precision-recall analysis underscores the importance of a nuanced approach to maximise the conversion of leads.

The End