Lead Scoring Case Study

Summary Report

Objective of the case study:

Lead scoring assignment is to help X Education improve its lead conversion rates by building a logistic regression model to score leads based on their conversion. The X Education is an online education platform which receives a large number of leads daily but struggling with a low conversion rate of around 30%. The case study is to identify 'hot leads' that are quite likely to convert and advising the sales team with data driven decisions for better conversion rate.

Approach to the case study:

- Our approach to the case study involved the following steps: Data Exploration and Cleaning, Feature Engineering, and Model Building.

- To begin with we loaded the dataset which contained around 9,000 leads with various attributes, including lead source, time spent on the website, and last activity. Initial data exploration involved checking the data types, identifying missing values, and assessing the distribution of the target variable ('Converted').

- The presence of 'Select' was identified in many categorical variables, which was treated as missing data. We dropped columns with more than 40% missing values and replaced 'Select' entries with NaN. The missing values were addressed by either dropping or imputing them with appropriate replacements, we handled duplicates to ensure data integrity.

- We performed feature engineering by consolidating categories in variables like 'Specialization' and 'Lead Source' to reduce complexity and enhance the predictive power of the model. For instance, many management specializations were grouped under a single 'Management' category. Categorical variables were encoded into dummy variables to prepare the data for further modelling.

- The logistic regression model was selected due to its simplicity and effectiveness for binary classification tasks. We split the data into training and testing sets with a 70:30 ratio and scaled the numerical variables using Min-Max scaling. To select the most impactful features, we used Recursive Feature Elimination (RFE), which helped the variables in lead conversion.

- Through iterative model building, we examined the p-values and Variance Inflation Factor (VIF) to refine the feature set by ensuring only statistically significant and non-collinear variables were included. This process resulted in a final model with proper accuracy, sensitivity, and specificity.

Insights of the case study:

- Our final model performance was attained using accuracy, sensitivity, specificity and ROC-AUC scores. The model demonstrated a good ability to differentiate between converting and non-converting leads, with sensitivity being particularly important given the business goal of identifying hot leads.

- We also observed that the Total Time Spent on Website, Lead Origin_Landing Page Submission and Last Activity_SMS Sent were the top predictors of conversion suggesting that engagement metrics are key drivers of customer interest.

- One key learning from this assignment was the importance of thorough data cleaning and preparation, especially handling missing values and redundant categories in categorical variables. Proper feature engineering significantly impacts model performance by reducing noise and focusing on the most relevant information.

Recommendations:

From a strategic perspective, the findings suggest that X Education should increase its engagement through channels like Olark Chat and ensure prompt follow-ups for leads showing high web engagement. When additional resources like interns are available, the strategy should involve lowering the cutoff probability to capture a wider range of potential leads. During quieter periods when targets are met, the company should raise the cutoff to focus only on the highest probability leads, minimizing unnecessary efforts. This assignment shows the importance of data-driven decision-making in sales strategies and highlighted the need for continuous model evaluation and adjustment based on changing business needs and trends. The learnings we gathered are valuable for future data science projects. Understanding how to balance model complexity with interpretability and actionable insights in particular.