Lead Scoring Case Study Summary Report

This case study examines the analytical steps undertaken to build a lead scoring model that can predict the likelihood of a sales lead converting into an actual customer. By accurately scoring leads, the model enables the sales team to focus efforts on promising leads.

The first and most critical step was thoroughly comprehending the business challenge and objectives. Additionally, we needed to understand the dataset itself - the meaning of each variable and how they relate to the conversion target. Developing this initial grasp of the problem and data provides the foundation for an effective analysis.

With this background knowledge, data cleaning commenced by importing necessary libraries and packages into our workspace. We scanned for missing observations, finding over 70% of cases missing in a few non-critical columns. As those variables were inconsequential for the analysis, we removed them entirely. For the remaining features, we filled in missing values using median imputation for continuous variables and mode imputation for categorical factors.

After preparing a complete dataset, we moved on to exploratory analysis to understand the variable distributions and relationships. All variables were encoded into numerical formats, with 1 representing "yes" and 0 as "no" for binary factors like ‘Converted’. Using distributions, correlations and data visualizations, we drew several insights and made notes of non-informative features to discard later. Plotting the target variable against other metrics provided preliminary confirmation that patterns exist within the dataset.

With a grasp of the data landscape, we then prepared the dataset for modeling. Several categorical variables were one-hot encoded to capture their effects. We split the data into train, test and scaled sample sets for appropriate model building and evaluation. Correlation heatmaps illuminated high correlations, causing us to remove one of each related variable pair to avoid multicollinearity issues. We then standardized all independent features through scaling to facilitate interpretation.

On the prepared dataset, we constructed a logistic regression model using statsmodels with a Recursive Feature Elimination approach to reach an optimal set of 15 predictive variables. We generated predicted conversion probabilities on the train set and evaluated different probability cutoffs for optimizing accuracy, specificity, sensitivity and other metrics. Additionally, we plotted the ROC curve, finding it skewed closer to the left edge, indicative of strong predictive accuracy. With multiple evaluation metrics analyzed, we also assessed the precision-recall tradeoff, finding 0.4 probability as the optimal cutoff. We then scored the holdout test dataset, achieving accuracy, sensitivity and specificity measures within acceptable ranges. By tuning between precision and recall, the model can shift to suit business needs.

Through this analytical process, the most indicative lead conversion features were total time spent on website, lead source referrals and lead source social. The model performance metrics were reasonably consistent between train and test sets. As more data accrues, the model can easily re-train to improve predictiveness. In summary, the systematic workflow from data cleaning to exploratory analysis to modeling and evaluation resulted in an actionable lead scoring model for the business.