Lead Conversion Analysis – Summary Report

We started by examining a dataset of 9,240 leads with 37 variables to predict lead conversion. Our first step was data cleaning. We carefully reviewed missing values and dropped columns with excessive null entries—such as Country and other fields with over 3,000 missing values. There are some unique ID such as the Prospect ID and Lead Number which will not support us to predict the probability so we have to eliminate it. There are some critical features have been missing values such as Total Visits, Lead Source, then we need to keep them with only 69% original data to move forward

In the next phase, we execute the EDA to evaluate the correlation among features. There are some features has shown the high correlation such as Total Visit and Page Views per Visit look it makes sense as the more they visit the more page they views and . However, the more views clients are having does not correlate with the more clients being converted. Also, drop features have many values of 'Select'. Dummy variables were created for all categorical features using one-hot encoding.

For feature engineering, there are some metrics have been added Avg_Time_Per_Visit, Avg_Time_Per_View for better understanding; after that, we use MixMaxScaler for Avg_Time_Per_Visit, Avg_Time_Per_View, Total Time Spent on Website to fit with the model. We then used Recursive Feature Elimination (RFE) with logistic regression to narrow down the predictor set to the most relevant 15 variables. The selected features included crucial engagement metrics like Total Time Spent on Website and TotalVisits, as well as important categorical indicators such as Lead Origin_Lead Add Form, Lead Source_Olark Chat, and Lead Source_Welingak Website.

We built our predictive model using a Generalized Linear Model (GLM) with a Binomial family (logistic regression) implemented in statsmodels. The training set of 4,461 observations was used for model fitting, and through iterative refinement, removing predictors with high p-values or excessive multicollinearity—we finalized our model. The final model achieved a log-likelihood of –2,067.5 and a R-squared near 0.3675.

When we evaluated the model on the training data with the default cutoff of 0.5, we obtained an accuracy of about 79.1%, sensitivity near 74.1%, and specificity around 83.7%. ROC curve is 0.86 which confirms strong discriminative performance. Base on the chart, the intersection of 0.42 offered the optimal values among accuracy, sensitivity, and specificity. On the test set, using this 0.44 threshold resulted in an accuracy of about 78.8%, with precision of around 78.2% and recall near 77.2%.

Findings:

Data preparation is crucial for improving model performance by removing non-informative variables and managing missing data effectively. We discovered that user engagement is a key predictor of conversion, specifically the amount of time spent on the website and the frequency of visits. Leads source has contributed the main game to change the conversion rate. And based on subjective questions, adjusting the categorization threshold is one of the most effective strategies to adapt the model to different business conditions. For example, lowering the threshold while more manpower is available (such as during the intern phase) can enhance recall, whilst raising it when seeking to decrease unnecessary calls can improve precision.