Summary- Lead Scoring

X Education, an online education company, faces a challenge with its low lead conversion rate despite receiving lots of leads daily. To enhance efficiency, they want to identify the most promising leads to increase the conversion rate from the current 30% to a target of around 80% and assign a lead score between 0-100 to them. This led to the development of a logistic regression model, following a sequence of comprehensive steps for data exploration, preprocessing, model construction, and evaluation.

The initial phase encompassed an exploration of the data with a focus on adhering to the Data Dictionary's specifications. Data cleaning involved the removal of columns with over 40% missing values, imputing missing values in columns with lower rates, and eliminating rows with a mere 1% of missing values. Unnecessary variables were discarded, low-frequency categories were grouped, duplicates were checked, and outlier treatment was administered. Subsequently, the data underwent exploratory data analysis, including visualization through plots, and correlation analysis.

The logistic regression model construction commenced by generating dummy variables for categorical columns while dropping other/unknown categories if present; otherwise, the first category was dropped. The dataset was then divided into training and testing sets at a 70/30 ratio, with numerical variables scaled for best performance. Using the training data, a generalized linear model was developed, integrating Recursive Feature Elimination (RFE) and Variance Inflation Factor (VIF) techniques to refine feature selection and model effectiveness. The final model encompassed 12 features alongside a constant. The crated model can be described as follows:

Logit(P) = -1.3352 + 7.2024 * Tags_Closed by Horizzon + 5.8929 * Tags_Lost to EINS + 4.7345 * Tags_Will revert after reading the email + 4.4264 * Lead Source_Welingak Website + 2.0169 * Last Activity_SMS Sent - 1.5555 * Last Notable Activity_Olark Chat Conversation - 1.8582 * Last Notable Activity_Modified - 1.9641 * Tags_Interested in other courses - 2.7328 * Tags_Already a student - 3.1794 * Tags_Ringing - 3.8167 * Tags_invalid number - 4.2308 * Tags_switched off

Evaluation of the model on the training data revealed an overall accuracy of 91.9% at a cutoff point of 0.28. Sensitivity measured at 86.4%, and specificity reached 95.2%. Other assessments, including ROC curve analysis, Precision, and Recall measurements, further solidified the model's performance. Applying the same logistic regression model to the test data yielded similar results, with an overall accuracy of 92.8%, sensitivity of 87.1%, and a specificity of 96%. Moreover, the assignment of a lead score between 1-100 facilitates the identification of hot leads.

By concentrating efforts on the most promising features such as "Tags_Closed by Horizzon" and avoiding segments with lower conversion rates, such as "Already a student", X Education can optimize resource allocation. Embracing these insights, the company should be able to realize the target conversion rate of above 80%, ensuring a more efficient and effective lead conversion process. The lead scoring system will further aid in promptly identifying the most promising leads.