Summary

X Education, an online course provider for professionals, seeks to enhance its 30% lead conversion rate. This case study aims to build a logistic regression model to score leads from 0 to 100, prioritizing potential conversions.

Here are the steps followed to achieve desired outcomes -

Data Cleaning:

- Identified columns with > 40% missing values and analysed those columns to take appropriate actions for handling missing values.
- Fields which seems significant for business are handled by imputing missing values with “No Information” and fields which no intrinsic significance for the model are dropped e.g. Tags.
- The value "Select", which seems to be a default option in a dropdown menu, so "Select" is equivalent to a NULL.
- Removed skewed data by setting threshold of 95% and if 95% of all the values in a field fall into the same category, we will drop the column as it won't be useful for analysis and model building.
- Outliers in the numerical fields are removed by dropping the records due to low no. of outliers.

Exploratory Data Analysis:

- Data imbalance checked, no notable difference in leads converted vs not converted.
- Checked correlation between the Numeric Variables and the Target Variable - no significant correlation between any of the numeric columns to the target variable found.
- Visualized the distribution of the categorical variables with respect to the Target Variable and derived insights and their effects on target variable.

Data Preparation:

- Encoded & created dummy features for categorical variables.
- Splitting Train & Test Sets: 80:20 ratio
- Feature Scaling using Standardization.

Model Building:

- Used RFE to reduce variables to 15.
- Manual Feature elimination was performed to build models by dropping variables with p – value > 0.05. and high VIF > 5. Total 3 iterations before reaching final Model.
- Used Hyperparameter tuning on final model using ROC AUC, to select most optimal parameters for the model.

Model Evaluation & Predictions on Test Data:

- ROC curve plot showed 92% AUC which indicates good model.
- Confusion matrix was made and cut off point of 0.34 was selected based on accuracy, sensitivity, and specificity plot. This cut off gave accuracy, sensitivity & specificity all around ~83%
- As to solve business problem CEO asked to boost conversion rate to 80%, but metrics dropped when we took precision-recall view. So, we choose sensitivity-specificity view for our optimal cut-off for final predictions.

Interpretation and Results:

- **Total Time Spent on Website** : This is the most important variable for the model, shows that the more time the lead has spent on the website, more likely to convert.
- **Lead Origin_Lead Add Form** : Positive coefficient indicates that the lead originated from "Lead Add Form", are likely to convert.

- **Lead Quality_Worst** : Negative coefficient here means lead labelled as "Worst", should not be prioritized.
- **Last Notable Activity_SMS Sent** : If lasty notable activity was sending SMS, are likely to convert.