Summary

We developed a Logistic Regression model to help X Education Company, which aims to enhance its lead management system to increase the conversion rate, optimize the sales team's efforts, and ultimately improve overall business profitability.

The steps we followed for building this model:

1. Preparation of the data: Importing all libraries, read and clean the input data, and carried out EDA like dropping the Prospect ID, Lead Number as the column is not needed for building the model and has no impact on the target. Dropped columns that are majority nulls. We used subplots for analyzing numerical columns and boxplot for categorical columns. Dropped those categorical columns which show no variance with the target variable (Converted).

2. Train-Test Split: The dataset was split into 80% and 20% for training and testing of the model respectively.

3. Feature Scaling: After creating the dummy variables the categorical variables showing variance with variance in Target variable, we scaled the numerical features using Standard Scaler technique, using 'StandardScaler' from 'sklearn'.

4. Checking Conversion Rate: We have a conversion rate of 38% approximately in the original data frame.

5. Feature Selection using RFE and Correlation: Used RFE feature selection technique to attain top relevant variables. Looking at correlations, we dropped dummy variables that showed higher correlation i.e., coefficients of correlation greater than 50% by observing correlation matrix.

6. Model Building: We used a Logistic Regression model. The rest of the insignificant variables were removed on the basis of their significance in the model as shown in model summary. The VIF for the features of the model were calculated, to handle multi collinearity in the model.

7. Model Evaluation: We plotted ROC Curve, which was tending towards the top left corner and ROC Curve value stood at 0.87, which is close to 1, indicating a strong predictive model performance.

8. Finding Optimal Cut-off: Plotted "accuracy", "sensitivity" and "specificity" for various probability (ranging from 0 to 1) to get optimal cut-offs, which observed to be 0.3 approximately.

9. Making predictions and calculating scores on train and test data: model has an accuracy of 79% with 82% sensitivity on the train data whereas an accuracy of 77% and sensitivity of 82% on the test data.

As the sensitivity is the measure of the power of a model of predicting true positive, our model has a very strong sensitivity on both train and test data i.e., sensitivity of approximately 82%. At overall level, our model shows a good ability to predict Conversion of lead which was asked by CEO to get it around 80%. We multiply the probabilities obtained by our model with 100 to score the leads, and falls under the range 0 to 100 as stated in the problem statement.

Below is the list of the features that are significant in scoring a lead using our model:

- Lead Source_Olark Chat
- Last Activity_Olark Chat Conversation
- Last Notable Activity_Modified
- Total Time Spent on Website
- Last Notable Activity_SMS Sent
- Lead Source_Reference
- Do Not Email
- Lead Source_Welingak Website
- Last Notable Activity_Email Link Clicked
- Lead Origin_Lead Import
- Last Notable Activity_Had a Phone Conversation
- Last Notable Activity_Unreachable