Summary Report of Analysis for X Education

A company X Education, wanted to improve their lead conversion rate by classifying leads generated through various sources. We used Logistic Regression (A classification model) to classify these leads and present a solution to the company after they specified a ballpark of the required target lead conversion rate to be around 80%.

Approach

Step 1. Data Preparation:

- Checking for **duplicates** in the data we found that the values were unique
- **Dropping** certain metrics like 'Prospect ID' and 'Lead Number' which would not help in analysis
- Certain fields were not populated and marked as "**select**" they were later converted to "**NaN**" for analysis
- Dropping columns having '**null**' values greater than 40% as part of **Missing Value Handling**

Step 2. Exploratory Data Analysis:

We carried out exploratory data analysis on the categorical variables.

- Identification of missing values in different fields and clubbing with most dominant values
- Creation of graphs to visualize the data better
- Dropping columns from the data where only one value was dominant in order to avoid bias
- Clubbing similar features together to reduce dimensionality
- Creation of new variables like 'Not Tagged' where data was missing
- Dropping missing variables whose % was low
- Analysis of Numerical Values based on Correlation using heatmaps
- Boxplots to visualize and remove outliers

Step 3. Model Building

- **Binary Mapping** - replacing **Yes/No** responses to **0 and 1.**
- Adding **dummy variables**
- Logistic Regression Model building, by first **splitting the data into 70% Train and 30% Test sets**
- Numeric values were standardized using **StandardScaler**
- **RFE for Feature Selection** and built a model with selected features
- Dropping high **P-Values**

- **VIF Variance Inflation Factor to find correlation between the variables**
- Dropping variables with high correlation to maintain **VIF values** below 5
- Interpreting **Probabilities, Lead Score and Predictions on Train Data**.

Observations

- High **accuracy of 92.30%** indicating a good prediction with the model
- **Sensitivity (86.12 %)** and **Specificity (96.11 %)**
- **ROC curve** area of 0.97 also indicating a good predictive model
- **Optimal cut off point** at 0.3 after plotting a graph leading to accuracy of 92.11%
- Details of **training data**
  - Accuracy: 92.11%
  - Sensitivity: 89.85%
  - Specificity: 93.50%
- **Precision** score of **89.47%** and **Recall** score of **89.85%**

Step 4: Model Evaluation

- We proceeded to evaluate our **Testing data**, and following were our observations
  - Accuracy: 92.51%
  - Sensitivity: 91.39%
  - Specificity: 93.2%

Conclusions

By comparing the training and testing data we can conclude that our model is predicting the conversion rate with high accuracy. X Education can use above model to successfully scale up their lead conversion rate.

Key observations for improvement

- **It would be beneficial to increase leads from references and the Welingak website**
- **Working professionals are more likely to become customers**
- **Most of the lead conversion is taking place through google and direct traffic**
- **Lead conversion via SMS sent has the highest conversion rate and lead generation should be increased from the same**
- **Leads who opened the email also have a good conversion rate which could be improved further.**