Summary Report: Lead Scoring Assignment

The objective of this assignment was to develop and validate a Logistic Regression Model for predicting lead conversion and optimizing business decision-making for the Institute. The process involved steps explained below.

1. Data Cleaning and Preparation
We started by preparing the dataset for analysis. This step included identifying and addressing missing or null values, which were handled either through imputation techniques (mean, median, or mode) or by removing entries with more than 30% missing data. Additionally, there was some features which had less than 30% missing value, but we could not impute using any of the technics, so we have assigned those values to new category under 'Not Given'. Also, the column which has 'Select' as values in them we replaced those with null as they are as good as user did not fill any value in them and then processed those columns. Additionally, we identified outliers and restricted our data to not include those outliers in the process. We also dropper highly imbalance columns.

2. EDA
To better understand the dataset, we conducted univariate and multivariate analysis. We plotted both single feature and multifeatured charts using matplotlib and seaborn to understand relations between various feature combinations.

3. Feature Engineering
Dummy variables were created for categorical features, enabling their use in the model. Feature scaling was applied to numerical features using Standardization method to ensure all variables contributed proportionally to predictions.

4. Data Splitting
The data was divided into training and testing sets. The training set was used to build the logistic regression model, while the test set validated its generalizability.

5. Model Development
We used Recursive Feature Elimination(RFE) to select only top 15 relevant features and then used manual selection methods to identify the most significant predictors by maintaining the p value under .05 and VIF under 5.

6. Model Evaluation and Validation
The model's performance was measured using metrics such as accuracy, sensitivity, specificity, precision and recall and the ROC curve. We also optimized the cutoff value by comparing accuracy, sensitivity and specificity on the chart. By doing this we achieved satisfactory result as per the requirements.

7. Prediction
Prediction was made on the test data using the most optimized model and cutoff values. The model performed as par with the train model to give satisfactory results.

8. Conclusion

By forming the above analysis and looking at the result of various metric on test and train data we can conclude that model is performing quite well. It is consistent on Test and Train data which shows that model is not overfitting or underfitting. Below are the main features that can help the Institute to target the potential customer to increase leads.

Institute can target user who

- Have spend more time on Website
- Lead origin from Lead add form
- lead source from welingak website
- last activity as SMS
- User who are working profession
- User who responded to phone conversations