Summary Report

Problem Statement:

X Education sells online courses to industry professionals. X Education gets a lot of leads, its lead conversion rate is very poor. For example, if, say, they acquire 100 leads in a day, only about 30 of them are converted. To make this process more efficient, the company wishes to identify the most potential leads, also known as 'Hot Leads'. If they successfully identify this set of leads, the lead conversion rate should go up as the sales team will now be focusing more on communicating with the potential leads rather than making calls to everyone.

Solution Summary:

1) Reading and understanding the data:

    - Read and analyse the data

2) Data cleaning and data manipulation:

    - Check and handle duplicate data.
    - Check and handle NA values and missing values.
    - Drop columns, if it contains large number of missing values and not useful for the analysis.
    - Imputation of the values, if necessary.
    - Check and handle outliers in data.

3) Data Analysis:

    Then we started with the Exploratory Data Analysis of the data set to get a feel of how the data is oriented.

    - Univariate data analysis: value count, distribution of variable etc.
    - Bivariate data analysis: correlation coefficients and pattern between the variables etc.

4) Dummy Variables and encoding of the data:

    For categorical variables, the dummy data was created.

5) Classification technique:

    The next step was to divide the data set into test and train sections with a proportion of 70-30% values.

6) Feature Rescaling:

    For numeric variables, the Min Max Scaling is used.

7) Feature selection using RFE:

    Recursive Feature Elimination was used to select some top important features.

    From the Statistical data, we tried looking at the P-values in order to select the most significant values that should be present and dropped the insignificant values.

Finally, we arrived at the 15 most significant variables. The VIF's for these variables were also found to be good. We then created the data frame having the converted probability values and we had an initial assumption that a probability value of more than 0.5 means 1 else 0. Based on the above assumption, we derived the Confusion Metrics and calculated the overall accuracy of the model.

We also calculated the 'Sensitivity' and the 'Specificity' matrices to understand how reliable the model is.

8) Plotting the ROC Curve:

We then tried plotting the ROC curve for the features and the curve came out be pretty decent.

9) Finding the Optimal Cut-Off Point:

The intersecting point of the graphs was considered as the optimal probability cut-off point. The cut-off point was found out to be 0.35.

10) Computing the Precision and Recall metrics:

Based on the Precision and Recall trade-off, we got a cut off value of approximately 0.45.

11) Making Predictions on Test Set:

- Test Data Accuracy :81.43%
- Test Data Sensitivity :80.77%
- Test Data Specificity :81.15%
- Test Data F1 Score :0.74

12) Conclusions and Recommendations:

X Education Company needs to focus on following key aspects to improve the overall conversion rate:

- Increase user engagement on Welingak website since this helps in higher conversion.
- Focus on Working Professional which has high conversion certainly.
- Get Total Time Spent on Website increased by advertising and user experience which makes the customer engaging in the website. Since this helps in higher conversion.
- Improve the Olark Chat service since this is affecting the conversion negatively.
- Improving Lead add form also improves the lead conversion with high certainty.