Lead Scoring Case Study - Summary

This study was conducted on behalf of X Education with the aim of exploring strategies to attract a greater number of industry professionals to enrol in their courses. The foundational data offered valuable insights into the patterns of potential customer visits, their duration on the site, the channels through which they accessed the platform, and the resulting conversion rate.

Steps for the Analysis –

1. **Importing the data**
   In this step the dataset is loaded into Python using the NumPy and Pandas libraries.

2. **Inspecting the data**
   Inspecting the statistical elements of the dataset.

3. **Data Cleaning**
   This is an important step. The data was checked for duplicates. Also, for some records the data was missing, which might be because the user did not provide those data. For data columns where there was more than 30% data missing, the columns were dropped. For the remaining columns the missing values were imputed.

4. **Exploratory Data Analysis (EDA)**
   By performing EDA, it was observed that the many categorical columns contained highly skewed data. Hence, these columns were dropped. For the numerical columns outlier treatment was done.

5. **Data Preparation for Model Building**
   In this step 2 vital operations were done to the dataset –
   - The Binary Columns (Yes/No) columns were converted to 1/0 values.
   - Dummy variables were created for the Categorical columns.

6. **Train-test Split**
   The dataset was split into Train and Test sets with a ratio of 7:3.

7. **Feature Scaling using Standard Scaler**
   The features were brought to one scale using the Standard Scaler.

8. **Checking Correlations among the variables**
   The variables were checked for correlations.

9. **Model Building**
   RFE (Recursive Feature Elimination) was used to attain the top 15 features. Furthermore, manual elimination of the features was performed depending on the p-values and VIF (Variance Inflation Factor). The variables with p-value greater than 0.02 and VIF greater than 5 were ignored.

10. **Prediction on the Train set**
    The model was used to predict the target variable with an arbitrary cut-off probability of 0.5. A confusion matrix against this predicted outcome was created. The key metrics of Logistic regression were calculated –
    - Accuracy – 81.32%
    - Sensitivity – 69.54%
    - Specificity – 88.71%
    - False Positive - Rate – 11.29%
    - Positive Predictive Value – 79.41%
    - Negative Predictive Value – 82.30%

11. **Plotting the ROC Curve**

    An ROC curve was plotted which the following:
    - It shows the trade-off between sensitivity and specificity (any increase in sensitivity will be accompanied by a decrease in specificity).
    - The closer the curve follows the left-hand border and then the top border of the ROC space, the more accurate the test.
    - The closer the curve comes to the 45-degree diagonal of the ROC space, the less accurate the test.

    In this case the ROC curve is 0.89, indicating a good predictive model.

12. **Finding the Optimal Cut-off point**

    Optimal cutoff point is that probability where we get balanced sensitivity and specificity. In this model the cut-off point was determined to be 0.35. With this cut-off point the key metrics of Logistic regression were calculated –
    - Accuracy – 81.32%
    - Sensitivity – 69.54%
    - Specificity – 88.71%
    - False Positive Rate – 11.29%
    - Positive Predictive Value – 79.41%
    - Negative Predictive Value – 82.30%

    Also, the Precision and Recall scores were calculated –
    - Precision – 72.44%
    - Recall – 80.29%

13. **Predictions on the Test Set**

    Prediction was done on the test data frame and with an optimum cut off as 0.35. The key metric values were as follows –
    - Accuracy – 81.32%
    - Sensitivity – 69.54%
    - Specificity – 88.71%
    - False Positive Rate – 11.29%
    - Positive Predictive Value – 79.41%
    - Negative Predictive Value – 82.30%
    - Precision – 70.22%
    - Recall – 77.96%

Conclusion

1. The Test set has optimal key metric values of logistic regression.
2. The most important variables (top 5) are –
    - Lead Origin_Lead Add Form
    - Last Activity_Others
    - What is your current occupation_Working Professional
    - Last Activity_SMS Sent
    - Lead Source_Welingak Website
3. The accuracy and stability of the model is adaptive.