Lead Scoring Case Study

This case study is done for X Education to identify hot leads, and have a better conversion rate. The model developed for this case study is a classification model which is having accuracy of 77% and sensitivity of 76% for both the test and training datasets.
Major steps performed for the case study are mentioned below:

1. Importing and Inspecting Data

In this step, data set is imported from a csv format file and its size, format, shape, variable descriptions, value ranges are analyzed.

2. Cleaning data

Except for a few null values, the data was mostly clean, and the option select had to be replaced with a null value because it didn't provide any information. To avoid losing too much data, a few of the null values were imputed. Duplicity of rows and uniqueness of columns are also tested in this step.

3. EDA

To check the condition of our data, exploratory data analysis is performed. During this process, many features of categorical type were found to be irrelevant. Outlier analysis was also performed on numerical variables in this step, and the majority of the outliers were capped at particular values. Univariate and bivariate analysis were also used to examine patterns and correlations between the various features, as well as to eliminate irrelevant features.

4. Dummy Variables

Dummy variables were created with the relevant categorical columns like Country etc.

5. Splitting of Data in Training set and Test set

The cleaned/processed data was split into training and test data set randomly. The ratio of dataset was 70 training and 30% test dataset.

6. Scaling

The numerical features were scaled to have a common range so that their values do not influence their relevance in the model.

7. Model Building

Prior to anything in model building, RFE was performed to get the best 15 relevant features.
After RFE, models were built and analyzed iteratively and variables were removed, until the p-value and VIF of every feature was not obtained within the threshold limit (VIF<5, p-value<0.05).

8. Model Evaluation

For model evaluation, target variable was predicted on the basis of final model. Then confusion matrix, accuracy score, sensitivity, specificity, precision and recall scores were calculated. Also, through ROC curve, best threshold value was also calculated.

9. Model Evaluation on Test Data Set

Prior to anything in this step, we scaled the features with the same scaler used for scaling training data set.
Then target variable was predicted as per the model with the same threshold as for training data set. For evaluating the prediction, confusion matrix, accuracy score, sensitivity, specificity, precision and recall scores were calculated.

10. Conclusion

From the above predictions, X company can focus on those leads which are mentioned below:

- • Leads having high number of visits and spending much time on website
- • Leads having Source as Welingak Website, Reference, Olark chat and Google

Also, X company should avoid not focus much on leads which are having current occupation is either Unemployed or Student.