Lead Scoring Case Study Summary

Ramya D & Mounica N

In this report we provide a summary on the various steps in our modelling approach to find hot leads among the given pool of visitors.

Data Cleaning and Transformation:

1. A quick inspection on the dataset revealed that we have 9240 observations for 37 columns. The columns Lead Number are Prospect ID are unique for each observation and there are quite a few columns with missing values and incorrect data categories.
2. Since the columns Lead Number and Prospect ID are not relevant for modelling, we dropped these columns in this phase. Also, there are few columns which have only a constant value. We dropped these columns as well as they do not have any predictive information in them.
3. In the next step we observed that there are quite a few columns with missing percentages in the range of 20% to 70% and decided to drop all those columns who have a missing percentage greater than 35% . Prior to this step we replaced “Select”, “select” values in the dataset with nan so that those observations are properly addressed in the data cleaning stage.
4. For a few of the remaining columns that have a missing percentage values less than 35%, we decided to create a level called “not provided” so that we do not lose information either due to dropping or imputing. For the rest we have imputed the missing values with the mode of the respective categorical columns.
5. After this step, we reduced the levels in categorical columns with an exceedingly high number of levels to a reasonable number of levels with appropriate data transformations.
6. Next, we converted categorical columns with binary responses to numeric columns (0-1 mapping) and data types of the rest of object columns to category.
7. For the categorical variables that have more than 3 levels, we use one hot encoding to create dummy variables for each of those variables.
8. After all the above data processing steps, we did a outlier analysis on the three numerical variables present in the dataset. We found that out of these 3 columns, there are not outliers present in one column, and the rest of the two columns have outliers in the order of 200. Since this is a significant number of observations, instead of dropping these rows we decided to transform the numerical variable to categorical variable by binning. For this purpose, we used the quantiles of the column to come with the relevant boundaries for the level. After creating these bins, we again transformed these two categorical variables using one hot encoding

Splitting the cleaned dataset into train and test datasets.

1. Split the dataset into train and test dataset with a a split of 70% : 30% and scaled the train dataset using the standard transformer .
2. Did a detailed analysis on correlations and plotted heatmaps to fine the dependencies between the variables.

Model building:

1. Used the standard logistic regression model with recursive feature elimination (RFE) to select the top 20 variables.
2. After feature selection step, we started iteratively updating the model by removing a variable at time so that there is no multi collinearity present in the model and all the VIF values are < 2 in the model.
3. Next, we removed all the variables which have an insignificant coefficient. For achieving this we had a threshold of 5% of p-values.
4. At the end of eight such model building iterations our final model has 14 variables.

Model Evaluation

1. Our final model has AUCROC of 87%
2. To get the final predictions we plotted sensitivity, specificity, and accuracy of the model with thresholds starting from 0 to 0.95 with a step size of 0.05. From the graph we found that 0.35 as the optimal threshold for our model.
3. For this optimal threshold of 0.35 the model showed a precision of 0.72 and recall of 0.77. Overall accuracy for the train dataset stands at 80%
4. From the precision recall analysis we did, we found that both these values coincide almost at a threshold of 0.4.

Predictions on Test Dataset:

1. To evaluate the model on the test dataset, we used the fitted scalar from the train dataset to transform the numeric variable in the test dataset.
2. On the test dataset also, the accuracy stands at 80% showing that the final model was neither under fit nor overfit.
3. For the test dataset, we have sensitivity, specificity, precision and recall as 0.77, 0.82, 0.74 and 0.77 respectively.

Conclusion:

1. The Accuracy, Precision and Recall score we got from test set in acceptable range.
2. We have high recall score than precision score which we were exactly looking for.
3. In business terms, this model has an ability to adjust with the company's requirements in coming future.
4. This concludes that the model is in stable state.
5. The top three variables in our model which contribute most towards the probability of a lead getting converted are Total Time Spent on Website, Lead Origin_Lead Add Form, Last Notable Activity_Modified with absolute z values of 27.296, 18.782 and 18.008 respectively. Their coefficients are 0.9343, 3.4568 and -1.3824.