Lead Scoring Assignment Report

In this assignment, we are tasked with making recommendation about the lead generated on the website which have higher chance of converting into final customer.

To accomplish this task, we have been provided with data regarding the source of these leads, their activity on the website and their details and preferences.

We will be analyzing these data to derive a probability score for each lead and based on this score, find the most promising leads and target them to enroll for the course.

We will be using Logistic Regression for deriving the probability score with a value between 0 and 1, where 1 means highly likely to convert and 0 means least likely to convert. We do this by fitting a sigmoid curve with the parameters which has low p-value and low VIF i.e., low probability that the outcome obtained is at random and also low collinearity.

Steps involved in model building is as follows: -

1. We start our case study by looking at the data set provided.
2. The data set provided has 9240 data points with 37 columns.
3. Prospect ID has no valuable information and can be dropped.
4. Converting "Select" value to null as it makes business sense.
5. Removing columns with null values more than 25%.
6. Handling outliers for numeric columns by capping the upper limit at 95th percentile value.
7. Converting Yes/No values to 1 and 0 respectively.
8. Removing columns with single values as it adds no extra information.
9. Visualizing numeric variables with distribution plots to look at spread of values.
10. Creating dummy variables from categorial variables for model training.
11. Performing train test split to obtain datasets for training and testing purposes.
12. Creating first model and looking at p-value and VIF for variables.
13. Performing RFE to obtain top 20 features.
14. Iterating and fine-tuning model to get final features which has p-value less than 0.05 and VIF less than 5.
15. Obtained final 15 features to evaluate model performance and predict on test set.
16. With the current cutoff of 0.4, we have Accuracy of 79%, Precision of around 72% and Recall of around 76%
17. Performing predictions on our test set, with the current cutoff of 0.4, we have Accuracy of 80%, Precision of around 73% and Recall of around 75%.

With the model created and the performance parameters observed to have acceptable score for both the training and test set, with very marginal difference we can conclude our model building step. The top 5 features influencing the probability score is as follow:

1) Lead origin is Lead add format
2) Total time spent on Website
3) Lead Source is Welingak Website
4) Last Activity was Had a Phone Conversation
5) Lead Origin as Lead Import