Lead Score Assignment Summary

The given assignment had goals to funnel the hot leads from the initial pool of leads acquired from various sources. These hot leads are the potential customers that will be converted to students of X education. The intent of study was to identify the parameters to help X education select the most promising leads. As a result of the study, a lead score was assigned to each lead that is the probability of its conversion. A target of 80% was to be achieved.

The following process was followed to achieve the goal:

- ✓ Understanding of data columns with help of data dictionary. Study of overall statistics of the data.
- ✓ 9240 rows and 37 columns found. 32 categorical columns and 5 numeric columns found. Missing values found in the form of 'Select' in 4 columns as these were dropdown values that were not selected by the users while filling a web form.
- ✓ Performed univariate analysis for each categorical column using value count plots and numerical columns using distribution plots.
- ✓ Data cleaning done on categorical columns by techniques used like column removal, clubbing of values with low counts into one with a comparable count in some cases and column conversion in case of columns with highly unstructured data.
- ✓ Numerical columns had 2.4% missing values. These rows were removed from the study because it was a small percentage of total number of rows.
- ✓ We soft capped the outliers on some of the numerical columns. For example, total visits column
- ✓ To prepare data for modelling, we created dummies, performed train-test split of the data into 70:30 ratio. We used standard scaler to normalize the data.
- ✓ We studied the correlation matrix and found the top 10 highly correlated columns.
- ✓ We first did the coarse tuning using RFE (Recursive Feature Elimination) using 25 variables selection. Then used these columns to finetune the model using GLM (Generalized Linear Model) using StatsModels while observing the VIF (Variance Inflation Factor) and the P-value of the parameter at every step.
- ✓ The output value of Converted from the model was checked against the actual and we found the values of True Positive, True Negative, False Positive and False Negative values.
- ✓ We created the ROC (Receiver Operating Characteristics) curve and observed a value of 0.89 area under the curve.
- ✓ We also plotted the optimal cutoff probability of accuracy, sensitivity and specificity values which intersected at 0.3 in our case.
- ✓ We also plotted the precision vs recall tradeoff to check the cutoff value. This matched precisely with our optimal cutoff value found earlier.
- ✓ We found the recall value of 85%, which was higher than our set target value. This particular evaluation metric was chosen because our business goal was to pursue hot leads and reduce cold calls.

Recommendations: Working professionals should be targeted most as they are our hottest leads. A user spending more time and visiting more often on website is a potential hot lead. All the sources of leads should be considered seriously as the smaller values when clubbed together is creating impact. This is re-iterated in other words by the Lead Quality_Low coefficient. Olark Chat is one very good lead source with potential candidates. The leads that have not disclosed their current occupation were found to have negative impact on the lead score.