Executive Summary

In this Lead Scoring Case Study, we received the data set from X Education containing about 9000+ records of people who were contacted or who responded to surveys hosted by X Education.

We inspected the data set and found that there were several columns that did not have proper values or were null. We cleaned up the data by dropping largely missing columns and dropped the rows for which some of the columns were null. There were several variables which were highly skewed, so we dropped those as well as they won't be of much use in building a model for predictions.

After cleaning up the data, we transformed it into a form suitable for building a model. We scaled some of the features and introduced dummy variables for categorical variables. Once all the features were ready, we dropped all the redundant columns.

Then we began the process of building a model. We started off with an initial model and then iterated upon it until we arrived at a model that had only useful features that could be used to predict the leads with greater accuracy.

The accuracy of the final model was 79.9%, while the sensitivity was 67.7% and the specificity was 87.5%. Using the ROC curve, we found that the optimal cutoff for the probabilities of conversion was 0.35.

The final model that was built performed really well on the test data as well achieving an accuracy of 78.6%.

The variables that were found to be contributing most towards getting leads were:

- Lead Source_Welingak Website
- Lead Source_Reference
- Last Notable Activity_Email Link Clicked