LEAD SCORING CASE STUDY REPORT

X Education is an edu-tech that has generated a lot of leads from multiple markets but has had a low conversion rate. Our goal was to identify the most promising leads, i.e. the leads that are most likely to convert into paying customers.

With the given dataset, we first worked on understanding the data. Then we proceeded towards data cleaning in which we dropped the columns that were highly skewed, replaced all ‘Select’ fields with nulls, dropped columns with >40% missing values, dropped rows of columns with < 2% missing values and imputed all other missing values, performed outlier treatment and also dropped Sales team columns. There were some columns with a lot of fields which would result in a large number of dummy variables - we dealt with this challenge by manipulating those columns to bucket < 2% values into the category ‘Others’ using replace function.

We then performed Exploratory Data Analysis using univariate analysis, bivariate analysis, and multivariate analysis. Using various plots such as box plots, dist plots, pairplot, heatmap we gained insights about data.

The dummy variables were created out of categorical variables with more than 2 levels. The dataset was then split into the training and test data set so as to later test the performance of the model on a data set unseen to the model. We then performed scaling so that the coefficients obtained from the modeling would indicate the true weights that the said variables have with respect to their impact on the target variable.

For the modeling, first RFE (Recursive Feature Elimination) was used to identify most important features. Next the p-values were judged and co-linearity of the variables were ranked using Variance Inflation Factor to further eliminate variables. Threshold for p-values was 0.05 and for VIF it was 5.

The focus of the model was to have a high sensitivity so that all potential leads could be identified for certain. It would be acceptable to identify some cold leads as hot but not vice-versa.

Then, the optimal probability cut off value was obtained by making use of accuracy-sensitivity-specificity curve. An appropriate optimal cutoff was thus obtained so that the conversion rate of leads at X Education improves. A trade-off on sensitivity, at a cut-off value of 0.24 resulted in the best case scenario, therefore it was chosen as the cutoff value for marking leads as ‘hot leads’.

Lead score was calculated by the formula *conversion probability * 100*
The Accuracy and Precision values using the final model on the test data were about 75% and 64% respectively.
The Sensitivity was observed to be 79% and Specificity 71%.
The False Positive Rate was observed at 29%.

The model helped us identify the ‘Hot Leads’ which will help X Education achieve a high conversion rate of 79%. The sales team can now focus on the Hot Leads and increase customer conversion. The sales team could also contact the customers who have a low lead score but have Lead Origin as ‘Lead Add Form’.