LEAD SCORING CASE STUDY
SUMMARY REPORT

Prasasti Choudhury, Satadhriti Chakrabarty
DS C17 Group 2 February 2020

Summary Report

DEALING WITH SELECT VALUES

The **Leads** dataset have some id columns which have been dropped as they don't add any value to the model. There were four variables with 'Select' values and we impute them with meaningful categories or nan values depending on the **sense** they make.

MISSING VALUE TREATMENT AND GROUPING OF FIELDS

We move on to the treatment of missing values and any variable with missing percentage more than 40% have been dropped from the dataset. Then the categorical variables with still high missing percentages are treated separately and **grouping** of some categories in the variables have been done to reduce the number of categories in a variable and for better interpretation. Consequently, in this step, some variables have been dropped due to high **skewness** towards a field which wouldn't have added any value to our model. The continuous variables have few missing values and those **rows** have been removed from our data.

DROPPING VARIABLES BASED ON THEIR DISTRIBUTION

In the next step, we again drop a few variables based on their**distribution** (maximum entries in one field).

OUTLIER TREATMENT

Next, we perform some outlier treatments for our continuous variables and mostly capped the outliers at **99 percentiles**.

EXPLORATORY DATA ANALYSIS

We do some EDA on the **categorical** and **continuous** variables separately and recommended some measures to increase the conversion rate along those lines.

MODEL BUILDING

We find that our target variable is not imbalanced and then proceed to the model building part.We divide our data into the **training** and **testing** sets and start working on building a robust model on the training data.

RFE AND BUILDING A ROBUST MODEL USING P-VALUES AND VIF

We run some obvious steps like feature scaling for the continuous variables. We perform a **Recursive Feature Elimination (RFE)** to select the **15 strongest feature variables**. We build a logistic regression model with those variables and 'Converted' as our target variable. We look at **p-values** and drop some variables which turn out to be insignificant. Then we run a check for multicollinearity using **VIF** and when we are satisfied with the p-values and VIF's of our feature variables, we generate the regression equation.

GENERATING LEAD SCORE COLUMN TO IDENTIFY HOT LEADS

Next, we create a new column for conversion probability ('Conversion_prob') using our predicted Y values from the train set. We check some evaluation metrics like accuracy, sensitivity, specificity and plot the ROC curve to find the optimal cutoff point. We multiply our conversion probability values by 100 to get to a generic **Lead Score** that will help us identify the **potential hot leads** on a 1-100 scale. We again check the basic evaluation metrics along with precision.

MAKING PREDICTIONS ON THE TEST SET

We then work on the test set similarly to get a Lead_Score. We check the evaluation metrics on the test set and find them to be **nearly like our metrics from the train set**. We conclude that our model is robust and will provide the best of results to X Education.