Lead scoring case study

An education company, X Education, sells courses online and wants to find out the quality of the leads so it's easy for them to direct their efforts to people who are more likely to convert.

The objective of the analysis is to give each lead a score from 0 to 100 to demonstrate the likelihood of them converting using logistic regression.

The dataset provided is lead data from the past with around 9000 data points. This dataset consists of user and employee-filled data for a potential lead and whether or not they converted.

The analysis begins by removing the "Select" from the dataset, a remnant of user-filled forms. The value is replaced and treated as a null value.
After that, Columns with high null values or low variance were removed. The tags column has data that is filled later and is not available when a lead comes in, this column is removed as well. Some numeric columns were subjected to outlier treatment and missing value treatment followed by dummy variables creation.

The process of Model building starts with a Test-train split where we split the dataset into 70% train and 30% test data. The test data will later be used to evaluate the model.

After the split, numeric data was scaled with scaler fitting only on the train data. We built a baseline model with all the columns along with an efficient column using Recursive feature elimination (RFE).

The new model was then evaluated on various metrics beyond simple accuracy, such as precision, recall, and F1-score. The Receiver Operating Characteristic (ROC) curve was plotted to visualize the trade-off among accuracy, sensitivity and specificity.
The accuracy, precision and recall of the final model with the training dataset were 0.84, 0.79 and 0.81 respectively, while for the test dataset, it was 0.85, 0.79 and 0.82 respectively. This makes sure that we are able to achieve the target of 80% conversion rate.

Some of the most important inferences from the study are -
Lead Quality, Lead Origin and Lead Source are the most important columns in our dataset.
Leads that were deemed to be high in relevance have the highest chance of converting while leads that are deemed the worst have the lowest. Due to this variance in conversion, lead quality is the most important column and employees should be trained to make this assessment.

Leads from the "Lead Add Form" also have a high probability of conversion, along with the Welingak website. Focus should be made on lead coming from these sources. The marketing team can also be suggested to focus more on these channels to get a higher quality lead.

Lastly, the Lead's occupation, last activity and last notable activity are also important columns and should be taken into account when judging a lead's quality.