SUMMARY

X Education receives a lot of leads, but only about 30% of those leads really become customers. The business wants us to create a model in which we score each lead individually so that leads with higher scores have a higher likelihood of converting. The CEO aims to convert leads at a rate of about 80%.

Therefore I have created a logistic regression model to predict the lead conversion rate and which all factors needs to be taken care of as our model predicts which is highly correlated with the model and which all are the factors negatively affects the lead conversion rate and as per the variables company needs to formulate the policies and procedures to take corrective as well as preventive measures by taking into consideration all the intricacies of the market factors , sales factors , financial factors and others.

The following steps I have taken while creating the model.

Data cleaning:

- Columns with more than 40% null values were removed. Value counts within categorical columns were reviewed to determine the best course of action: eliminate the column, create a new category (others), impute high frequency values, and drop columns that don't contribute any value if imputation creates skew.
- Numerical categorical data were imputed with mode, and columns containing just one distinct customer response were eliminated.
- Other tasks included handling outliers, correcting inaccurate data, grouping low frequency values, and mapping binary category values.

EDA:

- Only 38.5% of leads were converted when data imbalance was evaluated.
- Conducted categorical and numerical variable univariate and bivariate analyses. 'Lead Origin', 'Current occupation', 'Lead Source', etc. give important information about the impact on the target variable.
- Time spent on a website has a favourable effect on converting visitors into leads.

Data Preparation:

- 80:20 split between the train and test sets; one-hot encoding of dummy features for categorical variables; feature scaling using standardization
- A couple columns were dropped because they were very associated with one another.
- Model construction: RFE was used to condense 51 variables to 20. Dataframe will be easier to manage as a result.
- By excluding variables with a p-value greater than 0.05, models were constructed manually using feature reduction and also we have considered the VIF value and which have high VIF i.e VIF>5 has been eliminated.

- Before arriving at the final Model 7, which was stable with (p-values 0.05), a total of 7 models were constructed. With VIF 5, there is no indication of multicollinearity.
- We utilized the final model, logmf, which included 14 variables, to make predictions on both the train and test sets.

Model Evaluation:

- Based on the accuracy, sensitivity, and specificity plot, a confusion matrix was created and a cutoff point of 0.372 was chosen. Accuracy, specificity, and precision were all around 80% at this cutoff. The precise recall view, however, provided performance values that were less than 75%.
- CEO requested an increase in conversion rate to 80% in order to solve a business challenge, but metrics declined if we adopted a precision-recall perspective. Therefore, sensitivity-specificity view will be our top candidate for the cut-off for final forecasts.
- Lead score was applied to train data with a cutoff of 0.372.

Making Predictions on Test Data:

- Scaling and making predictions using the final model.
- Evaluation metrics for both the train and test phases are very close to 80%.
- Score for the lead was given.
- Top 3 characteristics are:
    - Total Time Spent on Website
    - Lead Source_Welingak Website
    - Lead Origin_Lead Add Form

Recommendations:

- The Welingak Website should provide more funds for advertising, etc.
- Discounts or incentives for supplying references that result in leads, which motivates submitting more references.
- Working professionals should be aggressively targeted because they convert well and are more likely to have the money to pay higher fees.