Summary Report

- by Harsh Bajpai

X Education receives a lot of leads, but only about 30% of those leads are converted. According to the company's requirements, we must create a model in which each lead is given a lead score, increasing the likelihood that a consumer would convert. The CEO wants the lead conversion rate to be about 80%.

1. Data Cleaning:

- Columns that had more than 40% nulls were removed. To determine the best course of action, value counts within categorical columns were examined. If imputation resulted in skew, the column was discarded; otherwise, a new category (others) was created; high frequency values were imputed; and columns that added no value were dropped.
- Columns containing just one distinct customer answer were eliminated, and numerical categorical data were imputed using mode.
- Additional actions, such as treating outliers, fixing erroneous data, mapped binary category values, and grouped low frequency values.

2. EDA :

- A examination of the data revealed that just 38.5% of the leads converted.
- Analyzed numerical and categorical variables using univariate and bivariate methods. "Lead Origin", "Current Occupation", "Lead Source" and so forth offer important information about how they affect the target variable.
- Website time has a beneficial effect on lead conversion.

3. Data Preparation:

- For categorical variables, (one-hot encoded) dummy features were created.
- Train and test sets are split in a 70:30 ratio.
- Standardization for Feature Scaling
- A few columns were dropped, and they showed a strong correlation with one another.

4. Model Building:

- Used RFE to decrease variables from 48 to 15. Dataframes will become easier to handle as a result.
- To construct models, variables having a p-value greater than 0.05 were eliminated using the manual feature reduction approach.
- Before arriving at the final Model 4, which was stable with p-values < 0.05, a total of three models were constructed. VIF < 5 indicates no multicollinearity.
- With 12 variables, logm4 was chosen as the final model, and we used it to make predictions on both the train and test sets.

5. Model Evaluation:

- Accuracy, sensitivity, and specificity plots were used to create a confusion matrix and choose a cutoff point of 0.345. Accuracy, specificity, and precision were all approximately 80% at this cutoff. On the other hand, the precise recall view provided less performance measurements, about 75%.
- As for resolving commercial issues When we adopted a precision-recall view, metrics declined, despite the CEO's request to increase the conversion rate to 80%. For our ideal cut-off for our final forecasts, we will hence select the sensitivity-specificity view.
- The train data was given a lead score with a cutoff of 0.345.

6. Predicting on Test Data:

- Predicting on Test: Using the final model to scale and make predictions.
- Train and test evaluation metrics are quite near to 80%.
- The lead score was determined.
- The three best features are:
- Welingak Website as the Lead Source
- Reference for the Lead Source
- Present Employment Status: Working Professional

7. Recommendations:

- The Welingak website might use more funding for advertising and other purposes.
- Rewards or discounts for submitting references that result in leads, which motivates you to submit more.
- Because they have a greater conversion rate and are in a better financial position to pay higher fees, working professionals should be actively targeted.