Summary

X Education receives a substantial number of leads, but its lead conversion rate remains low at approximately 30%. The company seeks our assistance in developing a mode=l to assign lead scores, prioritising leads with a higher chance of conversion. The CEO aims for an 80% lead conversion rate.

Data Cleaning:

- Columns with over 40% null values were dropped. Value counts in categorical columns were examined to determine appropriate actions such as dropping the column, creating a new category (e.g., "others"), imputing the most frequent value, or removing columns with no value.
- Numerical categorical data were imputed using the mode, and columns with only one unique response from customers were dropped.
- Additional activities included treating outliers, fixing invalid data, grouping low-frequency values, and mapping binary categorical values.

EDA:

- Data imbalance was verified, with only 38.5% of leads converting.
- Univariate and bivariate analysis conducted for categorical and numerical variables. Variables like 'Lead Origin,' 'Current occupation,' and 'Lead Source' offered valuable insights into their impact on the target variable.
- Time spent on the website demonstrated a positive influence on lead conversion.

Data Preparation:

- One-hot encoded dummy features were created for categorical variables.
- The dataset was split into training and test sets using a 70:30 ratio.
- Feature scaling was applied using standardisation.
- Certain columns were dropped due to high correlation with each other.

Model Building:

- RFE was employed to reduce variables from 49 to 23, improving dataframe manageability.
- In the Manual Feature Reduction process, variables with p-values > 0.05 were dropped to build models.
- Six models were tested before selecting Model 7, which exhibited statistical significance (p-values < 0.05) and no multicollinearity (VIF < 5).
- The final model, lr7, with 17 variables, was chosen for making predictions on both the train and test sets.

Model Evaluation:

- The confusion matrix was created, and a cutoff point of 0.386 was selected based on accuracy, sensitivity, and specificity plots. This cutoff achieved around 81% accuracy, specificity 81% and sensitivity of 79.5%. Recall value is about 79.5% however precision value is low (73%).
- To address the business problem of increasing the conversion rate to 80%, sensitivity-specificity view was chosen for the optimal cutoff for final predictions.
- Using the cutoff value of 0.386, lead scores were assigned to the train data.

Making Predictions on Test Data:

- Predictions on the test set were made by scaling and using the final model.
- Evaluation metrics for both train and test data were close, achieving our target of around 80%.
- Lead scores were assigned based on the model predictions.
- The top 3 features contributing to lead conversion are
    - Lead origin - Lead Add Form
    - Current Occupation- Housewife
    - Last Activity - SMS Sent

Recommendations

- Focus on aggressively targeting housewives by offering incentives or discounts as their lead conversion rate is high
- Target people whose last activity involves sending of SMS as their lead conversion rate is very high.
- Emphasis on features with positive coefficients such as Welingak Website, working professionals and total time spent on website