Summary

X Education, an online seller of educational courses, gets a lot of leads on several websites and search engines like Google. The company offers its courses to Industry Professionals. Typically, the company's lead conversion rate is around 30%, and they wish to enhance the efficiency. The company requires us to build a model wherein we need to assign a lead score to each of the leads such that the customers with a higher lead score havehigher conversion chance. CEO's target for lead conversion rate is around 80%.

Data Cleaning:

- Columns with greater than 40% nulls were eliminated. Value counts in categorical columns determined actions: drop if imputation causes skew, create new category (others), impute high-frequency value, or drop columns with no added value.
- Numerical categorical data were imputed with the mode, and columns with only one unique response from customers were dropped.
- Additional tasks included treating outliers, rectifying invalid data, grouping low-frequency values, and mapping binary categorical values.

EDA:

- Verified data imbalance; only 38.5% of leads converted.
- Conducted univariate and bivariate analysis for categorical and numerical variables. 'Lead Origin,' 'Current occupation,' 'Lead Source,' etc., offer valuable insights into the effect on the target variable.
- Time spent on the website exhibits a positive impact on lead conversion.

Data Preparation:

- Created dummy features (one-hot encoded) for categorical variables
- Splitting Train & Test Sets: 70:30 ratio
- Feature Scaling using Standardization
- Dropped few columns, they were highly correlated with each other

Model Building:

- Employed Recursive Feature Elimination (RFE) to reduce variables from 48 to 15 for better manageability.
- Utilized a Manual Feature Reduction process by dropping variables with p-value > 0.05 during model building.
- Developed a total of 3 models before arriving at the stable final Model 4, where all p-values were < 0.05 and no signs of multicollinearity with VIF < 5.
- Selected 'logm4' as the final model with 12 variables and applied it for predictions on both the train and test sets.

Model Evaluation:

- Created a confusion matrix and opted for a cutoff point of 0.345, considering accuracy, sensitivity, and specificity plot. This choice yielded approximately 80% accuracy, specificity, and precision, while the precision-recall view showed slightly lower performance metrics around 75%.
- Despite the CEO's goal to achieve an 80% conversion rate, metrics dropped in the precision-recall view. Therefore, the sensitivity-specificity view was chosen for the optimal cutoff in the final predictions.
- Assigned lead scores to the train data using 0.345 as the cutoff.

Making Predictions on Test Data:

- Making Predictions on Test: Scaling and predicting using final model.
- Evaluation metrics for train & test are very close to around 80%.
- Lead score was assigned.

- Top 3 features are:
    - Lead Source_Welingak Website
    - Lead Source_Reference
    - Current_occupation_Working Professional

Recommendations:

- Allocate additional budget/spending on Welingak Website for advertising, etc.
- Offer incentives/discounts for successful lead-converting references, encouraging individuals to provide more references.
- Implement aggressive targeting of working professionals due to their high conversion rate and potentially better financial capability to afford higher fees.
