Summary

X Education is an online course provider education company which is looking to improve it's lead conversion efficiency. CEO has targeted a conversion rate of around 80% but the current conversion is only around 30% which is very poor.

Data Cleaning:

- Features with >35% null values are dropped.
- Null values are treated for both categorical and numerical columns.
- Unwanted columns which do not have any impact on case study are dropped.

EDA:

- Data imbalance for target variable is checked.
- Univariate analysis of categorical columns and numerical columns
- Bivariate analysis of categorical columns is also performed to understand the data.

Data Preparation:

- Binary level categorical columns are mapped to 1/0
- Dummy variables are created for categorical columns with more than 2 levels.
- Split train and test data in 70:30 ratio.
- Use Min-Max scaling to scale features which also handles outliers in numerical columns.

Model Building

- Recursive Feature Elimination is used to select top 20 features.
- Then the model is manually tuned to end up with 16 features.
- Features with p-value higher than 0.05 are dropped one by one and model is built again.
- VIF is withing range (<5) for all the features.
- Model 6 is the final model.

Model Evaluation

- CEO of the company has set target to boost lead conversion from 30% to around 80%
- Probability cutoff of 0.35 is chosen as per sensitivity-specificity curve.
- Sensitivity is the metric which indicates the correct identification of the leads who will convert.
- Sensitivity of 81% is achieved which satisfies the CEO's requirement.
- 0.4 was probability cutoff which was obtained from precision recall curve which gave sensitivity of 77% which was less compared to the previous cutoff, so this threshold was not chosen.
- AUC (area under curve) of 0.88 is obtained which is really good number and curve is closely hugging y-axis.

Making predictions on test data

- Scaling is done for test data and predictions are made using model 6.
- Test set yielded sensitivity of 80% and area under curve of 0.88 which is similar to values obtained on train data.
- Lead score is assigned to both train and test data.

Recommendations:

- Develop strategies to attract high-quality leads from top-performing lead sources.
- Total visits and total time spent on website has very high positive co-efficients. So, increasing the number of ads in websites like welingak would help in landing lot of leads.
- Analyze negative coefficients for the features to make improvements.