Summary

Problem Statement: X Education gets a lot of leads, but its lead conversion rate is very poor at around 30%. 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 have a higher conversion chance. CEO's target for lead conversion rate is around 80%.

1. Reading and Understanding the data

- Importing the data and checking rows/columns and necessary information.

2. Data Cleaning:

- Checking if there are any duplicate values present in the datasets.
- Dropping the unwanted columns which are not required for analysis.
- Checked null values and dropped the columns having > 40% of null values. For the remaining null values, imputed Categorical Variable with the modes (If imputation doesn't bias the result) and numerical values with a median.
- Removed the Variables having very high data imbalance.
- Outlier treatment, fixing invalid data, grouping low-frequency values and mapping binary categorical values were carried out.

3. EDA

- Performed univariate, bivariate, and multivariate analyses for categorical and numerical variables. 'TotalVisit', 'Total Time Spent on Website', 'Current occupation', 'Lead Source', etc. provide valuable insight for the target variable.
- Checked for Data imbalance- A total of 38.36 percent of lead is converted.
- Time spent on the website shows a positive impact on lead conversion.

4. Model Building

- Created dummy features (one-hot encoded) for categorical variables
- Splitting the data into Train-Test Sets (70:30 ratio)
- MinMax Scaling is used on the numerical columns of the dataset excluding the dummy variables.
- Manual Feature Reduction process was used to build models by dropping variables with p-value > 0.05.
- After iterating through three models, the third model was the best fit with all the p-values <0.05 and no sign of multi-collinearity with VIF <5.
-

5. Model Evaluation

- A confusion matrix was made and a cut-off point of 0.30 was selected based on an accuracy, sensitivity, and specificity plot. This cut-off gave accuracy, specificity, and precision all around 90%. Whereas the precision-recall view gave fewer performance metrics around 75%.

- As to solving the business problem CEO asked to boost the conversion rate to 80%, but metrics dropped when we took a precision-recall view. So, we will choose a sensitivity-specificity view for our optimal cut-off for final predictions
- Lead score was assigned to train data using 0.30 as the cut-off.

Observations on Train & Test Data sets			
S.No	**Parameters**	**Train Data**	**Test Data**
1	Accuracy	92%	92.39%
2	Sensitivity	91.80%	91.80%
3	Specificity	92.71%	92.71%
4	Precision	88.77%	88.77%
5	Recall	91.80%	91.80%

6. Conclusions & Recommendation

The model is predicting the conversion rate very well and can be recommended to the CEO to make calls basis on it. To increase the overall lead conversion, the following points can be helpful:

- More leads from API and Landing Page Submission need to convert and more leads can be generated from Lead Add Form, Lead Import, and Quick Add Form.

- More leads need to be converted from Sources such as Google, Organic Search, Olark Chat, and Direct Traffic whereas more leads need to be generated from Referral Sites, References, and Welingak Website.

- Converting more leads from the Unemployed population and increasing the number of leads from Students and Working Professionals.

- Generating more leads from cities other than Mumbai and converting more leads from Mumbai.

- More leads need to generate from other specializations and most of the leads generated are from Management Specialization.

Thank You!