SUMMARY

X Education receives a lot of leads, but only about 30% of those leads are converted into sales. The organization wants us to create a model in which we give each lead a score so that clients with higher lead scores are more likely to convert. The CEO wants to convert leads at a rate of about 80%.

Data Cleaning:

- Null Values:
  - Columns with null/missing values ≥ 40% - Dropped from the dataset.
  - Columns with Null Values < 40% - Handled by imputing them with appropriate statistics such as mean, median, mode, or default values. If imputation didn't make sense or add value then the column is dropped from the dataset.
    - Numeric Columns - Imputed using either the mean or median, depending on the distribution of the data.
    - Categorical Columns - Imputed using mode or default values.
- Columns that don't have important information, repeat information or just one unique value are dropped.
- Checked and dropped columns with skewed data

EDA:

- Data imbalance is checked, revealing that only 38.5% of leads converted.
- Univariate and bivariate analysis for categorical and numerical variables is performed. Variables like 'Lead Origin', 'Current occupation', 'Lead Source', etc., provide valuable insights into their effect on the target variable.
- Few Observations:
  - SMS messages can have a high impact on lead conversion.
  - Landing page submissions can be helpful in generating more leads.
  - Specializations in human resources management and marketing management have high conversion rates.
  - References and offers are good sources for higher conversions.
  - Leads from working professionals have a higher conversion rate.

Data Preparation:

- Created dummy features for categorical variables.
- Splitting Train & Test sets in 80:20 ratio.
- Feature Scaling using MinMaxScaler.
- Checking and removing highly correlated columns.

Model Building:

- Used Recursive Feature Elimination (RFE) to reduce the variables from 67 to 15.
- Manually reduced features whose p-value is greater than 0.05 and VIF value is greater than 5.
- The final model with 14 features is evaluated.

Model Evaluation:

- Used accuracy, sensitivity, specificity, precision, and recall curves to select the optimum threshold of 0.345.
- **Train Set** - Accuracy: 80.96% ≈ 81%, Sensitivity: 80.77%, Specificity: 81.08%

Making Predictions on Test Data:

- Test set data is scaled and predicted using the final model.
- **Test Set -** Accuracy: 80.24% ≈ 80%, Sensitivity: 80.19%, Specificity: 80.28%
- Since, the train and test result are similar and they fulfil the business requirement this model is valid.
- Final 14 Feature list for the Model – Time Spent, Lead Origin-Lead Add Form, Current Occupation-Working Professional, Lead Source-Welingak Website, Last Activity-Had a Phone Conversation, Last Activity-SMS Sent, Total Visits, Lead Source-Olark Chat, Page Views Per Visit, Lead Origin-Landing Page Submission, Specialization-Others, Last Activity-Converted to Lead, Last Activity-Olark Chat Conversation

Recommendations:

- Lead Origin-Lead Add Form, Current Occupation-Working Professional, and Time Spent are the three most important variables affecting lead conversion.
- Since working professionals are more likely to take advantage of upskilling possibilities, targeting them seems promising.
- Using text messages or phone calls could increase the conversion rate of customers.
- A significant percentage of website visits do not become leads.
- Moreover, those connected to the most recent Activity-Olark Chat Conversation has a decreased chance of converting.

By Shivani Raut, Shiwani Jamdagni, Shivam Banerjee