Summary [Lead Scoring Case Study]

X Education gets a lot of leads for its online courses. But its lead conversion rate is quite poor, close to 30%. The company requires us to build a model each lead is assigned a score such that a higher lead score is indication of conversion chance for that customer.

Data Sourcing & Reading:

- Importing the required libraries
- Reading the given dataset "Leads.csv"
- Cursory Data Check: No of rows, columns, data type of each column, mean and median for all numerical columns, distribution etc.
- Missing value analysis.
- Duplicate rows check.

Data Cleaning:

- "Select" value is replaced with NAN.
- Columns with more than 35% of null values were dropped.
- Numerical data columns were imputed with their median value.
- Columns with only one unique response from customer were dropped.
- Other activities like outliers' treatment , grouping low frequency values, fixing invalid data, mapping binary categorical values were carried out.

EDA:

- Data imbalance checked : around 39% leads converted.
- Performed univariate and bivariate analysis for both categorical and numerical variables.

Data Preparation:

- Created dummy features for categorical variables
- Splitting dataset into Train & Test Sets: 70:30 ratio
- Feature Scaling using Standardization for numerical columns
- A few columns were dropped ,as they were highly correlated with each other

Data Modelling :

- Used RFE to reduce variables to 15.
- Manual Feature Reduction was applied by dropping variables with P value greater than 0.05. to improve model viability.
- Total 6 models were built before reaching final Model 7 which was stable with (p-values < 0.05).
- No sign of multicollinearity was present in the final model with VIF < 3.
- Final model had 9 variables , it was used for making prediction on train and test set.

Model Evaluation:

- Confusion matrix was made and cut off point of 0.3 was selected based on accuracy, sensitivity, and specificity plot.
- As to solve business problem CEO asked to boost conversion rate to 80%, but metrics dropped when we took precision-recall view. So, sensitivity-specificity view was used for optimal cut-off for final prediction model.
- Lead score was assigned to train data using 0.3 as cut off.

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 scores were assigned.
- Top 3 features influencing the model are:
  1. Lead Origin_Lead Add Form
  2. Total Time Spent on Website
  3. What is your Current Occupation_Working Professional

Learnings & Recommendations:

● More budget/spend can be done on Welingak Website, Olark chat and Google in terms of advertising, etc.

● Prioritize leads nurturing through conversion channels like SMS or Emails as they have a higher likelihood of getting converted.

● Working professionals to be aggressively targeted as they have high conversion rate as well as better financial situation to pay adequate fee.