LEAD SCORING CASE STUDY
Executive Summary

1. Background

X Education is an online education company that offers courses to industry professionals. The company markets its courses on several websites and search engines like Google. Once people land on the website, they may browse the courses, fill out a form for a course, or watch some videos.

2. Business Problem & Objective

When people provide their email address or phone number, they are classified as a lead. The leads are then contacted by the sales team to convert them into customers. The typical lead conversion rate at X Education is around 30%, which considers poor conversion rates.

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 higher conversion chance. CEO's target for lead conversion rate is around 80%.

The pre-processing dataset contains missing/null values, unnecessary values, outliers, etc. Therefore, as a data analytics, we need to approach the following steps:

3. Data Cleaning:

- Checking dataset to find out what types of data contain missing/null values, duplicated, values, skewed values, etc
- Dropping columns with >40% nulls, skewed values, unnecessary values, only one unique response from customer
- Imputing the missing data using mode/median/mean
- Creating new category (others)
- Fixing outliers' treatment, invalid data, wrong format, mapping binary categorical values

4. EDA:

- Data imbalance checked- only 38.5% leads converted.
- Univariate analysis with categorical and numeric variables

5. Data Preparation:

- Created dummy features (one-hot encoded) for categorical variables.
- Splitting Train & Test Sets: 70:30 ratio
- Feature Scaling using Standardization.

6. Model Building:

- To reduce the dimensionality of the dataset and select only the important features, Recursive Feature Elimination (RFE) (from 48 to 15) to make dataset more manageable.
- Manual Feature Reduction process was used to build models by dropping variables with p – value > 0.05.
- Total 3 models were built before reaching final Model 4 which was stable with (p-values < 0.05). No sign of multicollinearity with VIF < 5.

7. Model Evaluation:

- Confusion matrix was made and cut off point of 0.345 was selected based on accuracy, sensitivity and specificity plot. This cut off gave accuracy, specificity and precision all around 80%.
Whereas precision recall view gave less performance metrics around 75%.
- The target of lead conversion rate is 80%, but metrics dropped when we took precision-recall view. So, we will choose sensitivity-specificity view for our optimal cut-off for final predictions.
-Lead score was assigned to train data using 0.345 as cut off.

8. Conclusion:

- Evaluation metrics for train & test are very close to around 80%.
- Top 3 features are:
  - Lead Source_Welingak Website
  - Lead Source_Reference
  - Current_Occupation_Working_Professional

9. Recommendations:

- Focus on features with positive coefficients for targeted marketing strategies.
- Develop strategies to attract high-quality leads from top-performing lead sources.
- Engage working professionals with tailored messaging.

- Optimize communication channels based on lead engagement impact.

- More budget/spend can be done on Welingak Website in terms of advertising, etc.

- Incentives/discounts for providing reference that convert to lead, encourage providing more references.

- Working professionals to be aggressively targeted as they have high conversion rate and will have better financial situation to pay higher fees too.

- Analyze negative coefficients in specialization offerings.

- Review landing page submission process for areas of improvement.