Summary Report

X Education, an online course provider, has a low **lead conversion rate** of about **30%**. Despite many daily leads from various channels, few become paying customers. To improve efficiency, they aim to identify and score 'Hot Leads'—those most likely to convert—targeting a conversion rate of **80%**, so the sales team can focus on the best prospects.

Data Understanding and Data Cleaning

- The given data has **9240** rows and **37** columns.
- Presence of **null / missing values**, outliers, which were addressed accordingly.
- "**Select**" levels were converted to null values.
- Columns with null values > 40% were dropped.
- Remaining nulls in categorical columns were filled with the **mode** and columns with only one unique value were removed.
- **Outliers** were treated using **winsorization** (capping) and data standardization included fixing invalid entries, grouping infrequent values and **mapping** binary categorical variables.

Exploratory Data Analysis

- **Data Imbalance** check was done and lead conversion rate = **38.5%**.
- **Univariate** (Categorical), **Bivariate** (Categorical-Categorical and Numerical-Categorical) and **Multivariate analysis** were conducted.
- Comparing Target - '**Converted'** vs '**Lead Origin**', '**Curr_Occupation**', '**Lead Source**', etc. provided valuable insights.
- **Total_time** spent on website positively impacts lead conversion.

Data Preparation

- Created **Dummy variables** (one-hot encoded) for categorical variables
- Splitting data into **70% Training** & **30% Testing** Sets
- Applied **Feature Scaling** using StandardScaler
- Dropped **highly correlated** columns

Model Building

- Used **Recursive Feature Elimination** to reduce variables from 48 to 15 to make it more manageable.
- Employed **Manual Feature Reduction** by dropping variables with p-values > 0.05.
- Built 3 models in total; Model 3 was the final and stable model with all p-values < 0.05 and **no multicollinearity** (VIF < 5)
- The final model, log_model_3, has 12 variables and was used for predictions on the train and test sets.

Model Evaluation

- **Confusion matrix** was created and **cut off point** = **0.35** was selected based on **Accuracy** (80%), **Sensitivity/Recall** (80%), **Specificity** (81%) and **Precision** (72%). Whereas, the **precision-recall** view gave less performance metrics around 75%.
- To meet the CEO's goal of an **80% conversion rate**, we switched to a **sensitivity-specificity approach** for determining the optimal cut-off for final predictions, as the precision-recall metrics declined.
- The **lead score** was assigned to the training data using the 0.35 cut-off point.

Making Prediction on Test Data

- Feature Scaling and predicting using the final model.
- The evaluation metrics for train & test tests were found to be around 80%.
- Lead score was assigned.
- Top 3 features are:
  1. “**Lead Origin_Lead Add Form**”
  2. “**Curr_Occupation_Working Professional**”
  3. “**Lead Source_Welingak Website**”

Recommendations

- By optimizing **form design**, enhancing lead qualification, personalizing **follow-ups** we can systematically increase the probability of lead conversion.
- By tailoring **marketing efforts** and **offers** specifically to address the needs and preferences of Working Professionals as they can pay higher fees due to their better financial situation.
- By optimizing website experience through **targeted content**, budget spent on **advertising** on Welingak and and effective **follow-up strategies** can gain more leads.