X Education Lead Conversion ML Assignment - Executive Summary

By: *Shayak Majumder* | *Shreya Dutta* | *Shweta*

The primary objective was to build a logistic regression model to predict the likelihood of lead conversion for a lead scoring system for X Education, enabling the company to prioritize leads based on the probability of conversion. This would help optimize resource allocation, enhancing sales efficiency.

Dataset Overview: The dataset used for this project consisted of 9,240 leads, with 37 columns at first, which we whittled down to 14 features including Lead Number, demographics, and behavioural attributes.

The target variable, 'Converted' (0 or 1), indicates whether a lead was successfully converted.

Top dataset attributes:

- **Demographic Data**: Includes categorical variables like city and country.
- **Behavioural Data**: Includes web activities (like time spent on website), last activity, and last notable action.
- **Historical Data**: Includes the source of the lead, special tags, and interaction history.

Preprocessing Steps:

1. **Data Cleaning**: Missing values were handled by removing columns or imputing based on the distribution. Columns with over 40% missing values were excluded from the dataset. We also dropped columns with no variance.
2. **Feature Encoding**: Categorical variables were transformed using dummy encoding.
3. **Feature Scaling**: Continuous variables were standardized to improve model convergence.
4. **Train-Test Split**: The data was divided into training and testing sets to ensure a robust evaluation.

Model Development:

1. **Base Model**: An initial logistic regression model was developed using all features to evaluate baseline performance.
2. **Feature Selection**: Recursive Feature Elimination (RFE) was used to identify the most significant predictors, reducing the number of features to a manageable set without compromising accuracy.
3. **Manual Feature Removal**: After RFE, multiple models were created (a total of 8) by dropping features with high p-value, until all p-values were brought under 0.05. The final VIF scores for features were all under 5 as well.

Performance Metrics: The following metrics were used to evaluate the model:

- **Accuracy**: Proportion of correctly classified instances.
- **Sensitivity**: Ability of the model to correctly identify converted leads.
- **Specificity**: Ability of the model to identify non-converted leads.
- **Confusion matrix:** This was done to identify positive and negative cases and the nuances in between.

The above was done by randomly selecting a cutoff of 0.5 first and then using ROC to decide a cutoff of 0.3.

Key Results:

- The final model achieved an **accuracy of 78.89% (~79%)**, sensitivity of **77.6% (~78%)**, and specificity of **79.7% (~80%)** on the training set.
- The lead scoring system assigned scores from 0 to 100 based on conversion probabilities, helping prioritize high-potential leads.

Recommendations For Company:

1. **Deployment**: Integrate the lead scoring system into CRM platforms to guide sales teams. High-priority leads should, of course, be targeted aggressively.
2. **Monitoring**: Regularly update and retrain the model with new data to maintain performance. This will be possible based on feedback from lead reachouts.
3. **Feature Expansion**: Incorporate additional data sources, such as customer feedback or competitor analysis.

The logistic regression model provides a practical and interpretable solution for lead prioritization. By focusing on high-probability leads, businesses can maximize conversion rates and optimize resource allocation, ensuring long-term growth and profitability.