# Lead Scoring Case Study

## Summary

*This summary is based on an analysis of X Education's leads dataset from the past with around 9000 data points.*

### Problem Context:

X Education sells online courses to industry professionals. The company markets its courses by various methods. It gets a lot of leads but its lead conversion rate is very poor (around 30%). For example, for every 100 leads acquired, only 30 of them convert. The company wants to make the process more efficient and improve the lead conversion rate. To do so, the company wants to identify the most potential leads, also known as 'Hot Leads'.

Upon doing so, the company can focus more on acquiring leads classified as 'Hot Leads'. This will make the process more resource-efficient and business more profitable.

### Business Objective:

1. Identify the most promising leads, i.e leads that are most likely to convert into paying customers.
2. Identify the variables that have a significant effect on the outcome/ variables that are indicative of converting.
3. Build a logistic regression model to assign a lead score (0 - 100) to every lead such that a lead with a higher lead score will have a higher chance of conversion.
4. Achieve a target lead conversion rate of around 80%.

### Solution Methodology:

1. **Reading Dataset and Basic inspection:**
   a. Read the data.
   b. A basic inspection of columns, datatypes, statistical summary, etc.
2. **Data Cleaning:**
   a. Check for missing values. Dropping columns with high percentages of missing values. Treated other columns appropriately by imputing median in case of numerical variables, mode for categorical variables, and in some cases created a new level, 'Unknown' or added the values to the level, 'Other'.
   b. Performed Outlier analysis, identifying outliers. The presence of these extreme outliers can affect the performance of the model. We can get rid of these.

   c. Performed a check on columns and their levels again. Dropped columns that had only one level or had one highly dominant level such variables are useless for modeling.

3. **Exploratory Data Analysis:**
   a. Performed Univariate, bivariate analysis to visualize the variables with respect to the target. Derived insights based on observations here.
   b. Looked at the distribution of numerical variables with respect to the target. Derived insights from observations.
   c. Created a common level for levels with low frequency for some of the categorical variables.
4. **Preprocessing Data for Modeling:**
   a. One-hot encoding categorical features.
   b. Split the data into training and test sets.
   c. Scaling numerical variables using standard scaler.
5. **Model Building (Logistic Regression):**
   a. Built the first model using statsmodel to check p-values corresponding to all variables.
   b. Used Recursive Feature Elimination (Coarse tuning) to get down to the 18 most important features. Further used manual VIF and p-value checks (fine-tuning) to get the final 15 variables that have an impact on the final outcome. Also, made sure that there was no multicollinearity in data by checking VIF.
   c. Cut-off tuning: After building the final model, used the ROC curve, sensitivity, specificity, precision-recall, and other advanced evaluation metrics to tune the cut-off.
   d. Made predictions on the test set (cross-validation), obtained confusion matrix and other evaluation metrics, and verified the goodness and stability of the model.

## Final Observations and Model Interpretation:

1. Lead Origin: Company should try to bring in more leads by Lead Add Form.
2. Lead Source: The company should invest more resources into acquiring leads from the following sources-
   a. welingak website
   b. References
3. The company should try to acquire people who happen to be working professionals.
4. Lead who spent more time on the website should be pursued more. Also, websites can be made more engaging and user-friendly to improve the numbers.

Most important variables to consider:

- 'Tags_Closed by Horizzon'
- 'Lead Source_Welingak Website'
- 'Tags_Will revert after reading the email'
- 'Tags_switched off'
- 'Tags_Already a student'
- 'Tags_Ringing'

Finally, the model has performed well and is generalizing enough. It was able to identify around 90% of the leads that convert and could also deliver a precision score of about 85% in the validation set, making the new conversion rate well above 80%.

=======xxxxxxxxx=======