LEAD SCORING CASE STUDY SUMMARY

BUSINESS PROBLEM STATEMENT:

X Education, an online course provider, seeks to boost its lead conversion rate, which currently stands at 30%. The company attracts professionals through various channels, including website visits, form submissions, and referrals. To optimize this process, they intend to pinpoint high-potential leads, known as 'Hot Leads,' to enhance their lead conversion rate. Aiming for an 80% conversion rate, the company aims to build a lead scoring model that assigns scores reflecting the likelihood of conversion, allowing the sales team to focus on prospects with the highest conversion potential.

GOAL OF THIS CASE STUDY:

Build a logistic regression model to assign a lead score between 0 and 100 to each of the leads which can be used by the company to target potential leads. A higher score would mean that the lead is hot, i.e., is most likely to convert whereas a lower score would mean that the lead is cold and will mostly not get converted.

SUMMARY SOLUTION:

We have done step by step approach to build the final module, below is the explanation of the summary steps one by one.

STEP1: IMPORTING THE LIBRARIES, READING AND UNDERSTANDING THE DATA:

- Imported essential libraries, silenced warnings, and configured the display settings for seamless analysis and module construction.
- After importing the libraries, the data was read and comprehensively inspected to gain a clear understanding of its contents.
- Several columns in the data frame exhibit a substantial number of missing values (null values).

STEP2: DATA CLEANING:

- As indicated in the Problem Statement, numerous categorical variables feature a 'Select' category that needs attention due to its equivalence to a null value. To facilitate module construction, all 'Select' variables were converted to null values.
- After evaluating the null values, it was determined that columns with over 40% null values should be dropped.
- Following the column drop, measures were taken to manage null values. Non-contributory columns were eliminated, while imputation was carried out where necessary to enhance the module's efficacy.
- Upon null value management, an assessment of skewness was conducted.
- The subsequent columns exhibit substantial skewness: 'Do Not Call', 'Search', 'Newspaper Article', 'X Education Forums', 'Newspaper', 'Digital Advertisement', 'Through Recommendations'. Due to their limited contribution to model enhancement and the potential to distort logistic regression models, these columns will be excluded. Skewed variables can influence parameter estimates, potentially leading to biased or inaccurate results.
- Moving on to outlier treatment, both "TotalVisits" and "Page Views Per Visit" display outliers. These outliers have been addressed by applying capping and flooring techniques.
- Necessary steps include rectifying invalid values and standardizing data across columns. Invalid values were resolved, and infrequent value levels were grouped as "Others".

STEP3: EXPLORATORY DATA ANALYSIS (EDA)

- Binary variables were identified and subsequently encoded as 0s and 1s to facilitate module construction.
- The data exhibits an imbalance ratio of 1.59:1.
- Conducted both univariate and bivariate analyses to gain insights into the distribution of values within each variable.
- Evaluated feature correlations to address multicollinearity.

STEP4: DATA PREPARATION

- For categorical variables with multiple levels, dummy features were generated through one-hot encoding. As binary-level categorical columns were already mapped to 1s and 0s in prior steps, the process involved creating dummy variables and subsequently dropping the original variables.
- Following dummy variable creation, the data was divided into test and train datasets using a 70:30 ratio.
- Post-split, x and y variables were generated to facilitate model construction.
- Standard scaling was applied to feature data using the Standard Scaler.

STEP5: MODEL BUILDING

- The model construction commenced by employing the Recursive Feature Elimination (RFE) method for feature selection, with the selection of 15 features using the parameter n_features_to_select.
- The P-values were iteratively examined to eliminate features in a recursive manner, ultimately leading to the final model.
- Model 4 exhibited significant p-values within the specified threshold (p_value > 0.05), and the Variance Inflation Factor (VIF) values were also favourable, all being less than 5.

STEP6: MODEL EVALUATION

- The final model's predicted values were applied to the training set.
- By setting the sensitivity-specificity threshold at 0.345, the model achieved the desired 80% target for the True Positive Rate, aligning effectively with the business objectives.
- Our final decision was to opt for the sensitivity-specificity approach as the optimal cutoff point, ensuring accurate predictions.

STEP7: PREDICTIONS

- The model exhibited robust predictive power, evident from its ROC curve area of 0.87 out of 1.
- By incorporating converted probabilities, the final predicted values were obtained, ready to be leveraged for lead conversion efforts.

STEP8: CONCLUSION

- Successfully attained the desired sensitivity of 80.05% in the training dataset and 79.82% in the test dataset, employing the threshold of 0.345. This accomplishment closely aligns with the CEO's stipulated target sensitivity of approximately 80%.
- The model's accuracy of 80.46% directly aligns with the predefined study objectives, confirming its effectiveness in lead prediction.
- The model highlights four pivotal features that significantly contribute to predicting hot leads:
  - Lead source welingak Website
  - Lead Source_Reference
  - Current_occupation_Working Professional
  - Last Activity_SMS Sent