Lead Scoring Case Study

Business Objective:

An education company named X Education sells online courses to industry professionals.

Currently company's conversion rate is low (30%). The Company needs your help to identify **Hot Leads** for a more **focused marketing and conversion strategy**. Thus, aiming for a **target lead conversion rate of 80%**

Objective:

To build a **logistic regression** model and assign lead score which can be used to predict lead conversion

Approach:

1. **Understanding Past Data**
   - We first understood past data to identify data structure, main variables etc.

2. **Initial Data Cleaning**
   - Identified columns with high percentage of **null and select values**. We **dropped the columns which have more than 30% null values**

3. **Exploratory Data Analysis**
   - **Identified and handled outliers** in Numerical Columns. We removed top 1% outliers in - TotalVisits and Page Views Per Visit
   - Some columns had values with **negligible frequency. We grouped them together** in other category to avoid too many features
   - Identified **not-important columns** which do not add value. Example Do Not Call has 100% No values.
   - **Checking Data imbalance**, the data has 38% records as Converted. Thus data is imbalanced

4. **Feature Engineering and Data Preparation**
   - **For Numerical Variables** – we have done **Standardization** that they are on similar scale and not impact model coefficients
   - **For Categorical Variables –** we have created dummy variables **using pd.get_dummies()**
   - We have created two sets of Data – Training Data (70%) and Test Data (30%). This will help us **evaluate the model performance on unseen data**

5. **Building Logistic Regression Model**
   - We have used **RFE to pick the top 20 significant variables**.
   - We did **multiple iterations to refine** using model statistics – Pvalue and VIF. We have removed variables suing below criteria
     - **Pvalue more than 5%** - means variable is not significant
     - **VIF value more than 5** – means variable shows multi-collinearity
   - Finally , after 9 iterations we arrived at final model

6. Identifying the correct value of Probability threshold

- Basis **ROC curve** we got a rough idea on Probability cutoff
- Plotting Accuracy, Sensitivity and Specificity we arrived at **probability cutoff of 0.35** which give best balance among - Accuracy, Sensitivity and Specificity

7. Making Predictions

- Use used final model (9th iteration) to make predictions on training and test data and used 0.35 cutoff to map Converted leads

8. Model Evaluation

- We created Confusion matrix on training and test predictions
- Training Accuracy: Accuracy = 80.6% , Sensitivity = 80.6% , Specificity = 80.5%
- Training Accuracy: Accuracy = 80.6% , Sensitivity = 79.6% , Specificity = 81.2%
- All these were found to meet the 80% standard set on accuracy

9. Conclusion and Recommendations:

Top Features which determine higher conversion are

- Lead Source_Welingak Website (Coefficient = 6.48)
- Lead Source_Reference (Coefficient = 4.08)
- What is your current occupation_Working Professional (Coefficient = 2.22)
- Basis availability of resources we should focus on these variables to maximize conversion.