Lead Score Assignment Subjective Question Answers

- ***Submission by Akileshwari DJ and Vivek Sharma***

Problem statement:
X Education markets its courses on several websites and search engines like Google and get a lots of leads but conversion rate of the company is around 30%.

Objective:
To help X Education to identify the most promising leads out of leads that they receive from people filling a form on website or through past referrals. These hot leads should categorize users that are most likely to convert into paying customers. The CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80%. Also, the model should be able to adjust if the company's requirement changes in the near future.

Approach:
To build a logistic regression model to predict lead conversion probability
Decide on a probability threshold value above which a lead will be predicted as converted, whereas not converted if it is below it.
Assign a lead score value between 0 and 100 to each of the leads which can be used by the company to target potential leads.

Summary of the solution:

1. **Understanding & Cleaning the Data:** We have used the Jupyter Python Notebook to load and understand the Leads datasets. We have gone through the breadth & the depth of the features present in the datasets along with their definition to assess the data quality & it's spread at a high level. As part of the Data Cleaning process, we had found & treated all the irregularities in the dataset such as Missing Values; Outliers; Skewed Categorical features & Invalid data points.

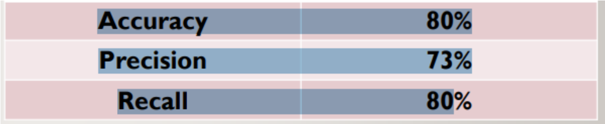
2. **Exploratory Data Analysis:** After cleaning the date, we performed various types of Univariate, Bivariate & Multivariate analysis by plotting appropriate graphs with respect to the Target variable. This helped us to draw relevant insights & correlations present within the dataset.

3. **Data Preparation:** Here, we had firstly created the dummy variables of all the categorical features in the dataset. Then, we had split the dataset between training & test sets & after that, we performed the Standard Scaling of the independent features.

4. **Model Building:** Since the count of features were high, we first started with RFE to eliminate redundant features & then built the first Model. Over multiple iterations of refinement(through p-value & VIF), we concluded this step with a final model.

5. **Model Evaluation:** Here, with the final model, we first obtained the lead score and plotted the ROC Curve. After that, we determined the optimal cut-off to proceed further with the Evaluations Metrics.

6. **Making Predictions:** After evaluating the final model, we ran the model over the test dataset to make predictions & then, we reviewed the predicted results with the actual records. We finally concluded with our analysis findings & recommendations to the business.



Accuracy	80%
Precision	73%
Recall	80%

7. **Recommendations:**

   a. Model performed equally well on the test dataset as it had on the training dataset. This shows that the **model is quite stable** and has a very good Accuracy & Recall.

   b. Based on changes in probability cut-off, the model is adjusting with the change in company’s requirements in the near future.

   c. Below are the top 3 Important Features that the company should focus on to further increase the conversion rate of the leads:

      i. Lead Source as “WelingakWebsite”

      ii. Lead Source as “Reference”

      iii. Current Occupation as “Working Professional”