Lead Scoring Case Study
Project Summary

Problem Statement:
An education company named X Education sells online courses to industry professionals. The company gets a lot of leads, some of the leads get converted while most do not. The typical lead conversion rate of the company is around 30%. To make this process more efficient, the company wishes to identify the most potential leads, also known as 'Hot Leads'. If they successfully identify this set of leads, the lead conversion rate should go up as the sales team will now be focusing more on communicating with the potential leads rather than making calls to everyone.

Business Objective:
The business objective of this case study is to identify the most promising leads , that is the leads that are most likely to convert into paying customers for the X Education company.
We need to build a model and assign a lead score to each of the leads such that the customers with a higher lead score have a higher chance of conversion than the customers with a lower lead score.
The ballpark target given by the company's CEO is to get the lead conversion rate to be around 80%

Solution Summary:

Step-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.

Step-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.

Step-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.

Step-4: Model Building
Since the count of insignificant features were very high, we first started with RFE to eliminate redundant features & then built the first Model with 15 features. Over a few iterations of refinement(through p-value & VIF), we concluded this step with a final model with 14 features.

Step-5: Model Evaluation
Here, with the final model, we first obtained the lead score and plotted the ROC Curve to determine the model stability with AUC (Area Under the Curve) Score. We found that this score is 0.88 which is a great score & represents quite stable & reliable model.

In order to find the most optimal cut off, we had plotted the graph between 'Accuracy', 'Sensitivity', and 'Specificity' at different probability/lead_score values. Finally, we found 0.35 as the most optimal probability cut-off value.
With this cut off value, we had created the Confusion Matrix (see table at the to check the Accuracy, Sensitivity, Specificity, Precision & Recall of the model.

We calculated & obtained the below Evaluation Metrics from the Confusion Matrix:

EVALUATION METRICS	SCORE (rounded)
Accuracy	80%
Sensitivity	79%
Specificity	81%
Precision	72%
Recall	79%

Step-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. The results showed that our model is very much stable even on unknown datasets.

Step-7: Conclusion & Recommendations

We finally concluded with our analysis findings & recommendations to the Business.

As we had seen that our 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.

Also, by changing the probability cut off, the model has the ability to adjust with the change in company's requirements in the near future

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

-> Lead Source as " Welingak Website"
-> Lead Source as "Reference"
-> Current Occupation as " Working Professional