Lead Scoring Case Study Summary

1. First we import the data and do some initial analysis
2. Then we proceed further by cleaning the data. Cleaning involves removal of unnecessary columns which will not have any useful information and removing duplicates.
3. Some columns include a default option called "Select" which is asking the customer to select some option in it, hence we replace that with NaN (Not a Number).
4. Then we drop some columns which have only a single value and no added values since this will not provide any useful insight to us.
5. Then we check for missing values and find the percentage of missing values in each column. We are able to see a few columns with a very high percentage of null values. We decide on a threshold and drop all columns which have a null percentage of more than 45%.
6. We still have some missing values. So for categorical columns, we fill in the values based on findings and observations. In the case of numerical columns, we fill in the missing values with the mode of the respective columns.
7. Then, we proceed to do some analysis on the categorical columns and derive conclusions based on data visualization that a few columns do not add any vital information and we proceed to drop them.
8. Then, we check for outliers and handle them.
9. Now, we proceed towards data analysis and find the percentage of leads converted.
10. We now perform univariate analysis and derive some insights from the data :
    a. Lead Origin: Customers were identified by "Landing Page Submission" in 52.9% of cases and by "API" in 38.7%.
    b. Current_occupation: Its consumers make up 89.7% of the unemployed population.
    c. Do Not Email: 92.1% of the population has chosen not to receive emails regarding the course.
    d. Lead Source: Direct traffic and Google together account for 58.9% of the lead source.
    e. Last activity: 68% of customers participated in the last activity, which was SMS sent and email opened.
11. We now perform bivariate analysis and derive some insights from the data :
    a. **Lead Origin:** With a lead conversion rate of 36%, the "Landing Page Submission" generated almost 52% of all leads. Approximately 39% of clients were identified by the "API" with a 31% lead conversion rate.
    b. **Current_Occupation:** With a lead conversion rate of 34%, around 90% of the clients are unemployed. Working professionals make up only 7.6% of all customers and have a lead conversion rate of over 92%.
    c. **Do Not Email:** 92% of respondents have chosen not to receive emails regarding the course.
    d. **Lead Source:** Google contributes 40% of the LCR from its 31% customers, Direct Traffic contributes 32% of the LCR from its 27% customers, Organic

Search also contributes 37.8% of the LCR from its 12.5% customers, and Reference contributes 91% of the LCR from its only 6% of customers.

e. **Last Activity:** "SMS Sent" had a high lead conversion rate of 63% with a 30% contribution from previous activities; "Email Opened" had a 38% contribution from previous customer activities and a 37% lead conversion rate.

f. **Specialisation:** Marketing Management, Human Resources Management, and Financial Management contribute well.

12. We then perform some analysis on the numerical data and come to find that previous leads who spend more time on the website convert more successfully than those who spend less time.
13. Now, we proceed towards data preparation. We prepare the data by creating dummy variables and removing the original columns.
14. We do the classic 70-30 train-test split of the dataset
15. We scale the features using Standard Scalar.
16. We now plot a heat map of the correlation between the columns and based on the data seen, we decide to drop some columns.
17. We now proceed towards Model Building.First, we perform feature selection using Recursive Feature Elimination (RFE) with a logistic regression model. It starts by initializing the logistic regression model and then applies RFE to select the top 15 most important features from the training data. After fitting the RFE model, it generates a list of tuples containing the feature names, their selection status (True for selected, False for not selected), and their ranking based on importance according to RFE.
18. We also get the Variance Inflation Factor (VIF) for these columns.
19. We now do recursive model building, each time we remove a column either because of high p value or high VIF.
20. After 4 iterations, we come to a stop as the VIF of all columns is less than 5 and p values is also less than 0.05.
21. Then, we perform model evaluation, calculate the confusion matrix, Accuracy and the ROC curve.
22. Then, we try to find the optimal cut off point using the data, the confusion matrix and the accuracy.
23. Then by comparing all indicators from the Precision-Recall and Specificity-Sensitivity views will help us determine a better probability threshold for increasing the conversion rate to the CEO's asked 80% and conclude that 0.347 is the ideal cut off.
24. Then we check with some sample data and find out if the model gives correct predictions.
25. We conclude that, by using a cut-off value of 0.347, the model obtained a sensitivity of 80.05% in the train set and 79.82% in the test set. Sensitivity here refers to the proportion of leads that the model correctly predicts among all possible leads that convert. A target sensitivity of about 80% had been established by the CEO of X Education. Additionally, the model's accuracy, which was 80.46%, met the goals of the study.