Lead Scoring Case Study Summary

Problem Statement:

X Education wants to develop a lead scoring model to identify the most promising leads with the highest chance of conversion. The target conversion rate is around 80%, and the model should assign a lead score to each lead based on their conversion potential. To develop a lead scoring model, we can use machine learning algorithms such as logistic regression, decision trees, random forests, or gradient boosting. These algorithms use historical data to identify patterns and create a model that can predict the conversion probability of new leads.

The features that we can consider for the lead scoring model include demographic information such as age, location, job title, and education level, as well as behavioural data such as website activity, email engagement, and social media interactions. To train and evaluate the model, we can use metrics such as accuracy, precision, recall, and F1 score. We can also use techniques such as cross-validation and grid search to optimize the model's hyperparameters and improve its performance.

Once we have developed the lead scoring model, we can use it to prioritize leads and focus the sales team's efforts on the most promising leads. This can help improve the conversion rate and increase revenue for X Education.

Solution Summary:

Step1: Reading and Understanding Data.

1. Obtain the data: The first step is to obtain the data from the relevant source. This could be a database, a spreadsheet, or a data file.
2. Explore the data: Once the data is obtained, the next step is to explore the data and understand its structure. This involves examining the data types, the number of columns, and the number of rows in the dataset.
3. Check for missing values: Missing values can affect the accuracy of the analysis, so it is important to check if there are any missing values in the dataset.
4. Analyse the variables: After checking for missing values, the next step is to analyze the variables in the dataset. This involves looking at the distribution of the variables, identifying outliers, and checking for any correlations between variables.

Step2: Data Cleaning:

We dropped the variables that had high percentage of NULL values in them. This step also included imputing the missing values as and where required with median values in case of numerical variables and creation of new classification variables in case of categorical variables. The outliers were identified and removed.

Step3: Data Analysis:

Then we started with the Exploratory Data Analysis of the data set to get a feel of how the data is oriented. In this step, there were around 3 variables that were identified to have only one value in all rows. These variables were dropped.

Step4: Creating Dummy Variables:

we went on with creating dummy data for the categorical variables.

Step5: Test Train Split:

The next step was to divide the data set into test and train sections with a proportion of 70-30% values.

Step6: Feature Rescaling:

We used the Min Max Scaling to scale the original numerical variables. Then using the stats model we created our initial model, which would give us a complete statistical view of all the parameters of our model.

Step7: Feature selection using RFE:

Using the Recursive Feature Elimination, we went ahead and selected the 20 top important features. Using the statistics generated, we recursively tried looking at the P-values in order to select the most significant values that should be present and dropped the insignificant values.
Finally, we arrived at the 15 most significant variables. The VIF's for these variables were also found to be good.
We then created the data frame having the converted probability values and we had an initial assumption that a probability value of more than 0.5 means 1 else 0.
Based on the above assumption, we derived the Confusion Metrics and calculated the overall Accuracy of the model.
We also calculated the '**Sensitivity**' and the '**Specificity**' matrices to understand how reliable the model is.

Step8: Plotting the ROC Curve

We then tried plotting the ROC curve for the features and the curve came out be decent with an area coverage of 89% which further solidified the of the model.

Step9: Finding the Optimal Cut-off Point

Then we plotted the probability graph for the 'Accuracy', 'Sensitivity', and 'Specificity' for different probability values. The intersecting point of the graphs was considered as the optimal probability cut-off point. The cut-off point was found out to be 0.37
Based on the new value we could observe that close to 80% values were rightly predicted by the model.
We could also observe the new values of the 'accuracy=81%, 'sensitivity=79.8%', 'specificity=81.9%'.
Also calculated the lead score and figured that the final predicted variables approximately gave a target lead prediction of 80%

Step10: Computing the Precision and Recall metrics.

we also found out the Precision and Recall metrics values came out to be 79% and 70.5% respectively on the train data set.
Based on the Precision and Recall trade-off, we got a cut off value of approximately 0.42

Step11: Making Predictions on Test Set

Then we implemented the learnings to the test model and calculated the conversion probability based on the Sensitivity and Specificity metrics and found out the accuracy value to be 80.8%; Sensitivity=78.5%; Specificity= 82.2%.