SUMMARY REPORT

Data Analysis Assignment Process and Key Learnings

INTRODUCTION:

This summary report provides an overview of the data analysis assignment conducted for X Education, an online education company that sells courses to industry professionals. The assignment aimed to improve the company's lead conversion rate by identifying the most potential leads, also known as 'Hot Leads'. By focusing on these leads, the company aimed to increase the efficiency of their sales team and improve the overall lead conversion rate.

Data Cleaning and Preparation:

In the data cleaning and preparation phase, various steps were undertaken to ensure data quality.

- Values in certain columns were converted from "Yes/No" to "1/0" for consistency. Variables with "yes/no" labels were encoded, and categorical values were converted to NaN.
- Missing values were addressed by dropping columns with missing values greater than or equal to 70%.
- Additionally, columns with low variability and no informative value were dropped.
- The remaining missing values were imputed based on most of the data, resulting in a cleaner dataset and handle outliers.

Exploratory Data Analysis (EDA):

- EDA was conducted to gain insights into the dataset and inform subsequent analysis. The spot chart was uitilized to identify patterns, relationships, and potential factors influencing lead conversion.
- Correlation analysis was performed to uncover highly correlated variables and avoid multicollinearity issues. These exploratory insights guided subsequent data cleaning and feature selection steps.

Train-Test Split & Scaling:

- Split the dataset into a 70% training set and a 30% test set.
- Applied min-max scaling to the variables ['TotalVisits', 'Page Views Per Visit', 'Total Time Spent on Website'] for uniform scaling.

Model Building:

- Used Recursive Feature Elimination (RFE) for feature selection.
- Identified the top 15 relevant variables using RFE.
- Manually eliminated remaining variables based on VIF values and p-values.
- Created a confusion matrix to evaluate the model's overall accuracy, which was found to be 91.77%.

Model Evaluation:

- Creating a data frame to store the actual converted flag and the predicted Lead Score probabilities.
- Reshaping the predicted probabilities array for further analysis.
- Classifying the prospects as likely to convert (1) or not likely to convert (0) based on a cutoff probability of 0.5.
- Creating a confusion matrix to evaluate the model's performance.
- Calculating the accuracy score of the model.
- Checking the Variance Inflation Factors (VIF) for the predictor variables to identify multicollinearity.
- Calculating sensitivity, specificity, false positive rate, positive predictive value, and negative predictive value.
- Plotting the Receiver Operating Characteristic (ROC) curve to evaluate the model's performance.
- Finding the optimal cutoff point by calculating accuracy, sensitivity, and specificity for different probability cutoffs.
- Making the final predictions using a cutoff of 0.3.

RESULTS ACHIEVED:

The final model achieved an accuracy score of 91.77% on the training dataset, indicating a high level of accuracy in classifying leads as potential or non-potential. This evaluation metric provided confidence in the model's predictive capabilities. The model's precision, recall, and F1 score were also assessed, indicating its effectiveness in identifying 'Hot Leads' and reducing false positives or false negatives.

CONCLUSION:

This summary report outlined the methodology employed in the data analysis assignment for X Education. By leveraging data cleaning, exploratory analysis, feature engineering, and model building, model evaluation the company aimed to improve its lead conversion rate by identifying 'Hot Leads'. The insights gained from the analysis, along with the achieved results, can help X Education optimize their sales efforts, focus on potential leads, and increase the overall conversion rate.