SUMMARY REPORT

The analysis was conducted for X Education, with the objective of identifying strategies to attract more industry professionals to enroll in their courses. The available data provided valuable insights into the behavior of potential customers, including their website visit patterns, duration of visits, referral sources, and conversion rates.

We achieved the objective by following the below mentioned steps :-

1. **Data Preparation**: - In this process, we began by identifying variables that were redundant and subsequently removing them. We also converted the "Select" variable to null values, since they were essentially the same. This was because "Select" indicated that the user had not entered any data during the time of data entry. Next, we assessed the percentage of null values present in each column and decided to drop any columns that had more than 45% null values. Additionally, we dropped some rows to handle the null values in the dataset.
2. **Exploratory Data Analysis**: - In this step, we conducted a brief exploratory data analysis (EDA) to assess the quality of our data. We discovered that many of the elements in the categorical variables were not significant. Furthermore, we observed outliers in the "TOTALVISITS" variable, which we handled by removing the top 1% and bottom 1% of values from the column.
3. **Creating Dummy Variables**: - In this step, we created dummy variables with the help of pandas get_dummies function.
4. **Performing Train-Test Split**: - In this step, we split the data into X_train, X_test and y_train, y_split the ratio of split we chose for train and test was 70% and 30% respectively.
5. **Feature Scaling**: - Here, we scaled the data using StandardScaler from sklearn.preprocessing library.
6. **Model Building:**- In this process, we build 2 machine learning model at first we did RFE to select 15 most relevant variables, then we dropped variables based on p-value and VIF score, here we chose the threshold to be 5 and 0.05 for VIF and p-value respectively.

7. **Model Evaluation:** - We created a confusion matrix and to find out the best cut off point so that the sensitivity and specificity of a classification model are balanced, after finding the optimal cut off point i.e., 0.35 in this case we checked for the accuracy, specificity and sensitivity in the test set which came out to be approximately 80%, 81% and 81% respectively.
8. **Precision and Recall: -** After plotting precision-recall curve we found out an optimum cutoff of 0.42 and after setting that as our cut off we headed towards making prediction over test dataset and found out precision and recall value of 74% and 76% respectively.
9. **Conclusion:** -The Variables which are most significant in deciding if the lead is a potential buyer or not are as follows:-
    - Total Time Spent on Website.
    - When the Lead is a Working Professional.
    - If the lead had a Phone Conversation with the employee from the sales team.
    - If the lead source was :-
        a) Welingak Website
        b) Reference

----------XXXXXXXXXXXXXXXXX----------