Summary

This Analysis is done for X-Education to find more industry professionals join their courses. The basic data provided gave us more information about how the potential customers visit the site.

The steps used in the case study as follows:

1. Reading and cleaning the data:

The data provided had errors with bad quality, below are the steps we had taken to Clean the data

- Reading the data
- Checking for null values in each of the columns and deleting the columns that had nulls more than 35%
- Imputing the nulls with mode, mean or median values to fix the data quality
- Checking and treating the Outliers
- Categorization of some of the values in columns into 'others'
- Checking for data types and fixing the same.

2. EDA:

EDA was done on Target Variable, Categorical Variables and Numerical Variables.

- **Univariate Analysis** – To find the distribution of values in each of the variables

    a. We did a Univariate Analysis for Target variable 'Converted' to understand the conversion rate based on the given data.
    b. We did a Univariate Analysis on Categorical variables such as Lead Origin, Lead Source etc. Based on the distribution and the correlation on conversion rate, we decided to remove some of the columns and keep only significant ones.
    c. Similarly, we carried out the same for Numerical variables to understand the correlation with Conversion Rate. Kept the significant ones and deleted the rest.

- **Bi-Variate/Multivariate Analysis**

    a. Bi variate analysis was done on Numerical Variables against conversion rate by plotting the Heat Map.
    b. The Heatmap plotted gave us a clear picture on the correlation of variables with target variable.
    c. We also plotted box plots with Target variable to understand the median of numerical variables against the target variable. This helped us finalize the Numerical variables required for further analysis.

3. Data Preparation, creating Dummy Variables and Train-Test split:

- Before proceeding to creating dummy variables, we converted the Binary variables with Yes/No entries to 1/0
- We shortlisted the categorical columns for which dummies need to be created
- Dummy variables were created for the categorical variables and dropped the first one using 'drop_first=True'
- The data was split into train and test data with 70:30 ratio. Feature Scaling was done using the Standard Scaler

4. Model Building:

- Model Building was done using Stats Model & RFE
- Using RFE, top 15 relevant variables were shortlisted, and rest were removed
- 3 Models were built by eliminating columns based on P-Values and VIF Scores
- Model 3 was finalized as the Final model to test with the test data
- By balancing Sensitivity and Specificity, we arrived at the Optimal Cutoff point

6. Model Evaluation:

- **Confusion Matrix**

    a. A confusion matrix was used and plotted
    b. Calculated the Overall Accuracy
    c. Sensitivity, Specificity, False Positive Rate, Positive Predictive value and Negative Predictive values were calculated

- **Precision and Recall** – Calculated Precision and Recall values
- **Plotting the RoC Curve** – RoC curve was plotted to understand

    a. The balance between sensitivity and specificity, indicating that an increase in sensitivity results in a decrease in specificity
    b. The more the curve hugs the left and top borders of the ROC space, the more accurate the test is.
    c. The closer the curve is to the 45-degree diagonal in the ROC space, the less accurate the test is.

- **Precision and recall tradeoff** was used to understand, if we can rely on this method to reduce the number of True positives which leads to decrease in Recall/Sensitivity. In our case the results were contradictory and hence, we couldn't rely on this.

7. **Predictions on Test Data:** The model was applied for test data to understand the predictions.

The sensitivity value for both the test and train data is 80%, and the accuracy is approximately 80%. This indicates that the model performs well on the test dataset.

Insights: The variables that mattered the most in the potential buyers were:

- The total time spent on Website
- Total number of visits
- Lead source
- Lead origin

By focusing on the above points' X education can increase potential buyers to buy their courses.

8. Recommendation:

To boost the potential lead conversion rate, X-Education should focus on the key features responsible for a higher conversion rate:

- **Lead Source_Welingak Website**: Leads from the `Welingak Website` show a higher conversion rate, so the company should prioritize this website to attract more potential leads.
- **Lead Origin_Lead Add Form**: Leads who engage through the `Lead Add Form` have a higher conversion rate, so focusing on this method can bring in more high-potential leads.
- **Current Occupation_Working Professional**: Leads who are working professionals have a higher conversion rate. The company should target working professionals to increase the number of leads with a higher likelihood of conversion.
- **Last Activity_SMS Sent**: Leads whose last activity involved an SMS being sent are potentially high value leads for the company.
- **Total Time Spent on Website:** Leads who spend more time on the website are more likely to convert, making them potential high-value leads.