Summary

X Education sells online courses to industry professionals, who are referred through various websites and search engines. The company's lead conversion rate is around 30%, but it is poor. To improve efficiency, the company aims to identify potential leads, or 'Hot Leads', and focus on communicating with them instead of making calls. This would allow the sales team to focus on converting more leads, resulting in a more efficient lead conversion process.

In the initial stage, many leads are generated but few convert into paying customers. In the middle stage, nurturing potential leads, educating them about the product, and communicating effectively is crucial for higher lead conversion. A model assigning lead scores based on conversion chances is necessary, with a target of 80%.

Data Insights:

The dataset contains **9000 data points** from past leads, including attributes **like Lead Source, Total Time Spent on Website, Total Visits, and Last Activity etc**. The target variable is the **'Converted'** column, which indicates whether a past lead was converted or not. The data dictionary can be found in the zip folder. Categorical variables have 'Select' levels, which must be handled.

Model Used: Logistic Regression

Accuracy Obtained: 91%

Steps Followed: -

0. Imported Libs & Data & Checked the data frame

1. DATA CLEANING -

- Replaced the ‘Select’ cells with Nulls, dropped columns with more than 40% of missing values & check on other columns for replacement.
- Imputed Nulls with others & Reduced Category Reduction, Moved < 100 category to others.
- Deleted Unwanted Columns & More irrelevant Columns.

2. EDA - Exploratory Data Analysis

- Checked for outliers in Numeric Columns, handled those outliers.
- Did Univariate analysis, Univariate analysis w.r.t target variable, Bivariate analysis w.r.t target on numeric variables & categorical data.

3. Data Preparation

- Converted Binary Variables (Yes/No)
- Created Dummy Variables

4. **Test-Train Split**

- 70 / 30 Train / Test Split

5. Feature Scaling

- Used Standard Scaler

6. Model Building - Logistic Regression

- Used RFE - Recursive feature elimination
- Evaluated 5 models

7. Model Evaluation/Prediction

- Confusion Matrix
- 0.910 accuracy on Train set
- 0.911 accuracy on Test set

8. ROC Curve

- **Good AUC - Area Under Curve**

9. Optimal Cutoff point

- **~0.32**

10. Giving the lead score to the final data

Top three variables that contribute towards the probability of a lead getting converted: -

- Will revert after reading the email
- Closed by Horizon
- Lost to EINS

Top three categorical variables that contribute (when Tags not included).

- Time spent
- Lead Origin
- Occupation

____________Best of luck to the X Education's business____________