Lead Scoring Case Study Summary Report

The goal of this research is to help X Education attract more business professionals to their courses. We learned a lot about how potential consumers use the site, how long they stay there, how they got there, and the conversion rate from the basic data that was provided.

Steps Involved:

1. **Data Cleansing:**

    - Data given mostly are categorical in nature. Very few numerical columns are present. The 'Select' category which is present in some of the columns is replaced with null values.
    - Columns with null value percentage is more than 70% is dropped
    - Some columns which is not relevant and has more than 45% of null values also dropped
    - Columns which have single category or contains redundant information also dropped
    - To avoid losing more data and columns which has much more business relevance, a small number of the null values were converted to 'UnKnown'.
    - Some category variables have lots of categories which are very less in number. So we did not take that categories, otherwise lots of unwanted dummy variables will be created

2. **Exploratory Data Analysis (EDA):**

    - Countplot for different categorical variable with target variable has ploted. Various inferences such as Working professionals has higher conversion rate, etc. has made based on the EDA
    - For numerical variables, outlier check has been done using boxplot and quantile check. 'Total Visits' column is capped to 95 percentiles to treat the outlier. Otherwise, there are no apparent outliers, and the numerical figures look good.

3. **Dummy Variables:**

    Dummy variables were created and the dummies with **'UnKnown'** elements were removed. For numeric values the StandardScaler is used for feature scaling. Correlation matrix is created to check the multicollinearity between the variables. Dropped the variables with high multi-collinearity

4. **Train-Test split:**

    The 70% and 30% split was done for train and test data respectively.

5. **Model Building:**

Firstly, logistic regression model is built. Some of the features has high p-values. So RFE was done to attain the top 15 relevant variables. Second model was built based on the RFE feature selection. VIF is checked to avoid the multicollinearity of the features and it is less than 2% for all the features.

6. **Model Evaluation:**

A confusion matrix has been created.

- Initially the cut off is taken as 0.5 . Then accuracy score is 84.66%
- Then cut off is taken as .37 based on accuracy, sensitivity, specificity.Then, the accuracy score is 85.33%
- Then cut off is taken as .42 based on recall and precision. Then, the accuracy score is 85.4%

7. **Prediction:**

Prediction was done on the train data frame and with an optimum cut off as 0.42. The metrics:-

8. **Metrics of Train and Test data**

-The model evaluation of the train data

Accuracy : 0.8546685561328925

Sensitivity : 0.8074407195421096

Specificity : 0.8842509603072983

Precision : 0.8137618459002884

-The model evaluation of the test data

Accuracy : 0.847227322805729

Sensitivity : 0.7876643073811931

Specificity : 0.881199538638985

Precision : 0.7908629441624365

-The difference between the metrics of both train and test data is almost 2%.

-It can be inferred that our model is predicting the conversion pretty well on test da ta.

It was found that the most important variables in the potential buyers are:

1. Total time spend on the Website.
2. When the lead source was Welingak website
3. When the lead origin is Lead Import.

4. When their current occupation is as a working professional.

Considering the above variables, X Education can flourish as they have a very high probability of getting almost all the potential buyers to buy their courses.