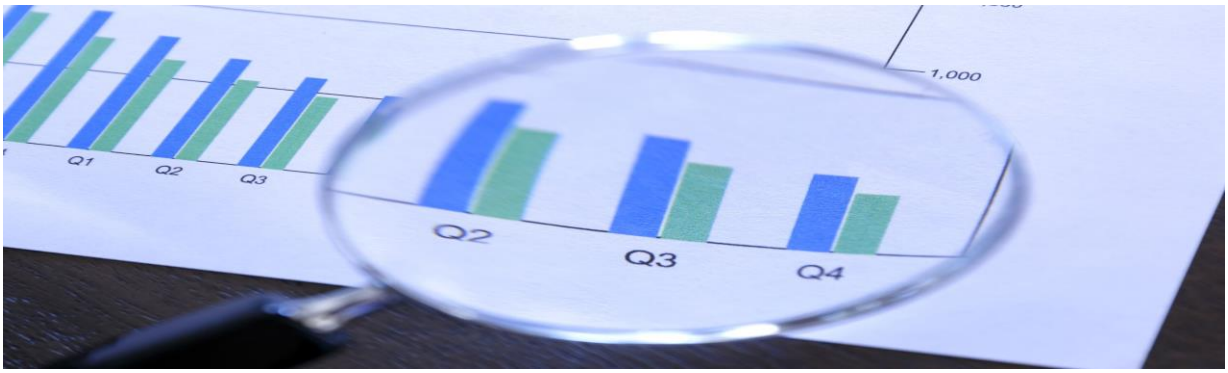

SUMMARY REPORT

An education company named X Education sells online courses to industry professionals. On any given day, many professionals who are interested in the courses land on their website and browse for courses.

Based on the data provided the company wants to find out the most potential customers (hot leads).

Steps taken:

1. Read and Understand data:

Data provided is loaded using appropriate Python libraries. Initial investigations are done (Check size of data, check the columns and description of the columns from the data dictionary provided, Check the data type, description of all numerical variables).

2. Data Cleaning and Analysis:

Data has nulls in a few rows and columns. By checking the percentage of nulls in columns and rows, it is either dropped (if high percentage of missing value) or imputed (less percentage of missing values) with the required values-using mean, median or mode. Many of the categorical variables have a level called 'Select' which needs to be handled. It is replaced with nulls. Also, many categorical columns have only 1 unique category. These columns are dropped.

The percentage of rows retained is calculated.

Analysis includes – Univariate, Outliers check, Bivariate and Multivariate.

Univariate Analysis is done on almost all categorical variables to visualize and understand better. Outlier check is done for numerical variables. Analysis showed 2 numerical variables had huge outliers. Bivariate analysis is done for checking the variables with respect to the target variable. Finally, through heat map relationship among variables are shown.

3. Data Preparation:

Dummy variables are created for all the categorical variables. Train-Test split is done on the data (70-30). Divided data is then scaled using MinMaxScaler. Lastly the data is divided for modelling and prediction.

4. Modelling:

Recursive Feature Elimination (RFE) is used to select 15 features and logistic regression model is built on these feature variables on train data set. P-value and VIF are checked and based on the values features are manually removed one by one. Remove the feature if p-value is greater than 0.05 and/or VIF is greater than 5. Optimal probability cutoff is calculated as 0.38

The above steps are repeated for a stable model with good sensitivity, specificity and accuracy. In train set accuracy is 88%, sensitivity is 87% and specificity is 88%.

5. Model Evaluation:

Confusion matrix is made to find out model's performance using sensitivity, specificity, accuracy, precision and recall. Optimal probability cutoff is calculated (using ROC curve and by plotting sensitivity and specificity for various probabilities).

6. Prediction on Test:

Prediction is done on test data using the optimal cutoff of 0.42. This is calculated by plotting Precision and Recall tradeoff. The accuracy is 89%, sensitivity is 86% and specificity is 90%.

Lead score is assigned based on the probability (probability x100). Based on this score hot lead can be easily extracted.

7. Conclusion:

The feature variables are arranged in the descending order (most significant at the top):

1) Total Time Spent on Website
2) Lead Origin_Lead Add Form
3) Tags_Will revert after reading the email
4) What is your current occupation_Working Professional
5) Last Activity_SMS Sent

Depending upon the above features potential leads can be contacted easily to attract more professionals and better conversions.