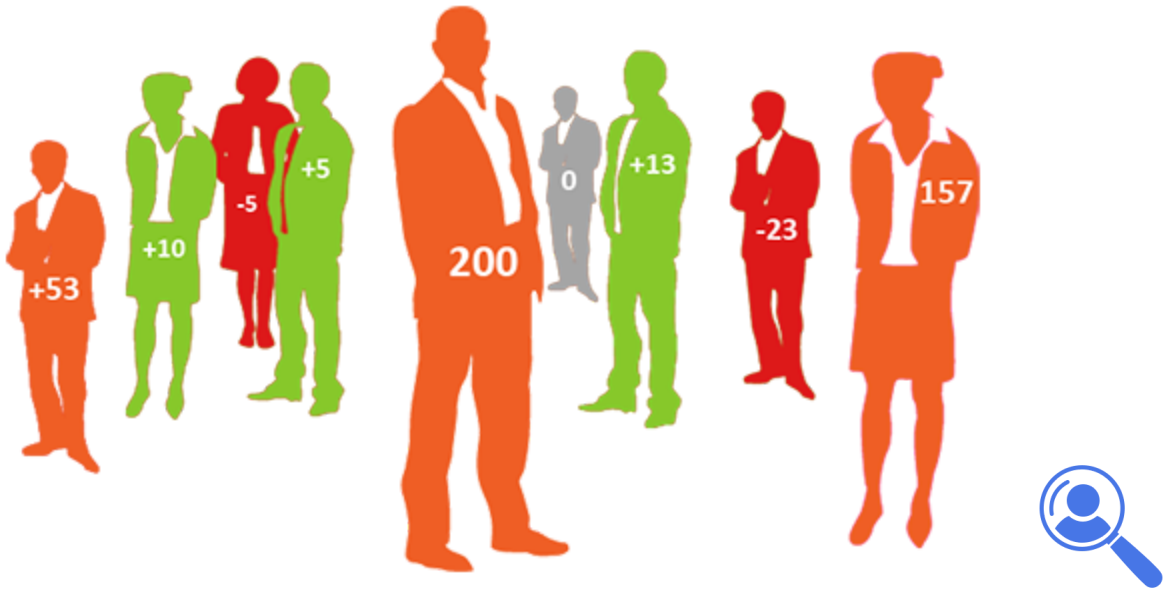
Lead Scoring Assignment

Summary Report



Our objective is to assign the lead score to each lead and find the most promising leads from the final dataset.

In order to find the score of each lead, we need to perform following steps:

1. Understanding Data:
    a. Import necessary libraries
    b. Load the dataset
    c. Check the shape of the dataset: Here we have 9240 rows and 37 columns
    d. Check for data types of all columns
    e. Check for null values: Null values are present in many columns
    f. Check for duplicate values: No duplicate values found
    g. Check for unique values present in each column
    h. Check statistical summary of Dataset: There are few features with very low variance and will not be useful for model generation.
2. Data Cleaning and Preparation:
    a. Dropping columns with null values > 3000
    b. As the brand is selling courses online, country and city doesn't matter. So drop these columns
    c. There are few columns with value as ‘select’. This is because the student has not selected a particular option from the select dropdown menu and hence it

is showing ‘select’ value. These are as good as missing values. So drop these values.
    d. Drop the columns which have very low variance
    e. Rows where values are missing and cannot be predicted at all should also be dropped (eg. Total visits)
3. Data Visualisation:
    a. Plot heatmap to understand collinearity among variables
    b. Plot pair plot of numeric variables to understand spread of data points
    c. Plot box plot to analyse outliers
    d. Plot count plot to analyse categorical variables
4. Data Transformation:
    a. Creating dummy variables for all categorical column
5. Test Train Split:
    a. Split the data set into 70% train set and 30% test set for model building
6. Scale Dataset for better interpretability
7. Model Building using Logistic regression model:
    a. Use RFE to select 15 top features for model building
    b. Drop the other features manually by observing their p values and VIF values
    c. Final model has all features with desirable p and VIF values
8. Making predictions on Train set:
    a. Get predicted values on the train set and compare those with actual values of train set
    b. Plot ROC curve: area under the curve is 0.86 which confirms that the model is good
    c. Plot the accuracy sensitivity and specificity for various probabilities for finding optimum cutoff value which comes out to be 0.42 in this case
    d. Overall accuracy of the model comes out to be 79% and sensitivity/recall value of approximately 80%
9. Making Predictions on Test set:
    a. Get predicted values on the test set
    b. Overall accuracy of the model on test set comes out to be 79% and sensitivity/recall value of approximately 78%
    c. Plot ROC curve: area under the curve is 0.85 which further confirms that the model is good
10. Assign the lead score to each lead, high score indicating hot lead which is worth pursuing
11. Check the features that played important role in efficient model building

Conclusion:

1. Our accuracy and sensitivity is almost same on Train set and Test set which means the model build is stable with adaptive environment skills which will adjust with generic data sets
2. For lead scoring definition, sensitivity is very important from business point of view which comes out to be 78% in our case
3. The top 3 features responsible for building model with good prediction capability are: 'TotalVisits', ' Total Time Spent On Website', 'Lead Origin_Lead Add Form'