LEAD SCORING CASE STUDY – SUMMARY

Topic: Linear Regression
Name 1: RAJAT DIXIT
Name 2: THERA BHUVANA CHANDRA

Summary of Case Study:

This analysis is done for X Education and to find ways to get more professionals to join their courses. The basic data provided gave us a lot of information about how the potential customers visit the site, the time they spend there, how they reached the site and the conversion rate.

Basic steps:

Import data and all the libraries, inspect the data frame

The following are the steps used:

1.Data Understanding:

a. Routine Data Check: No of rows, columns, data type of each column, distribution, mean and median for all numerical columns etc.
b. b. Missing value analysis
c. c. Duplicate rows check.

2. Data Cleaning:

- Check and handle duplicate data Check and handle NA values and missing values
- Drop columns which are having large missing values
- Imputation of values if necessary, Check and handle duplicate data
- Check and handle NA values and missing values
- Drop columns which are having large missing values
- Imputation of values if necessary

3. EDA:

- **Univariate data analysis:** value count, distribution of variable etc.
- **Bivariate data analysis:** correlation coefficients and pattern between the variables etc.

4.Dummy Variables:

- The dummy variables were created concatenated (joined) the results to the master data frame.

5.Train-Test split:

- The split was done at 70% and 30% for train and test data respectively. Perform scaling Divide the data into X and y.

6.Model Building:

- Use RFE For Feature Selection.
- Building A Model by Removing the Variable Whose p-value Is Greater Than 0.05 And VIF

Value Greater Than 5

- Predictions On Test Data Set
- Overall Accuracy 82%

7. Model Evaluation:

- A confusion matrix is made.
- Check overall accuracy found 82%
- Optimum cutoff value 0.5(ROC Curve) was used to find the accuracy
- The area under ROC curve is 0.89 which is very good value.

8. Prediction:

- Prediction was done on the test data frame and with an optimum cut off as 0.37 with accuracy 81.26%, sensitivity 80.69% and specificity of 81.60%.

9. Precision and recall tradeoff:

- With the current cut off as 0.43 we have Precision around 75.40% and recall around 76.99%

10. Conclusion & Recommendation

From model, we can conclude following points:

Focus on

- The Total time spends on the website.
- Total number of visits.
- When the lead source was:
    a. Google
    b. Direct traffic
    c. Organic search
- We must majorly focus on leads whose last activity is SMS sent or Email opened
- The customer/leads who fills the form are the potential leads.
- We must majorly focus on working professionals
- It's better to focus least on customers to whom the sent mail is bounced back.
-

Keeping these in mind the X Education can flourish as they have a very high chance to get almost all the potential buyers to change their mind and buy their courses.

Focus Less

- If the lead source is referral, he/she may not be the potential lead.
- If the lead didn't fill specialization, he/she may not know what to study and are not right
  people to target.
- Page Views Per Visit.

So, it's better to focus less on such cases.