

SUMMARY FOR LEAD SCORE CASE STUDY

Problem Statement

An education company named X Education sells online courses to industry professionals.

X Education needs help in selecting the leads that are most likely to convert into paying customers. The company requires you to build a model wherein you need to assign a lead score to each of the leads such as lead score have a higher conversion and vice versa.

The CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80%.

Our solution/ approach:

Steps:-

- 1. Reading and understanding the data :inspect**
- 2. Data cleaning and preparation:**
 - Changing to required format, removed unwanted columns.
 - Dropped columns with more than 40%missing values.
 - Changed columns containing “select” category as null value.
 - Imputed missing values based on value counts ; such as mode value for categorical columns.
 - Dropped unwanted columns after checking its value count and plots.
 - Segregated variables according to its types – numerical and categorical.
- 3. Analysis:**
 - Univariate, Bivariate and multivariate analysis done on both numerical and categorical columns – Boxplots, barplot, countplot, scatterplot, histogram created.
 - Pairplots and heat maps created to study the correlation.
 - Bivariate analysis with target variable “converted” done.
 - Many useful insights drawn from the plots and graphs.
- 4. Data transformation:**
 - Encoding of variables to binary form “0” and “1”.
 - Creation of dummy variables for categorical columns.
- 5. Data split into train and test:**
 - Data Split in 70-30 ratio (train-test).

6. Rescaling of features:

- Numerical columns rescaled using minmax scaler.

7. Model building:

- Logistic regression model building done by using Recursive Feature Elimination (RFE) and selected top 18 features.
- Assessed the model using statsmodel summary and checked the p-values and dropped the insignificant p-values > 0.05 .
- The Variance inflation Factor was also assessed to explain the predictor variable against all other predictors.
- Finally ended up with predictor variables with significant p value and good VIF.
- On train set ,
 - Optimal cutoff found on using ROC curve and the curve obtained was good with area under coverage (AUC) 95%.
 - Confusion matrix plotted and check the precision, accuracy, recall and specificity; calculated the conversion probability.
 - A trade-off between precision and recall was plotted to visually see the intersecting at optimal cutoff.
- Implemented the learning to test data and run the model to measure its predictions on test data.
- Performed all the functions of train data on test data and calculated the conversion probability.

8. Comparison of metrics for train and test data:-

parameters	train data	test data
accuracy	0.8913	0.8821
Precision /Positive predictive value	0.8400	0.8180
sensitivity / Recall	0.8867	0.8685
specificity	0.8942	0.8898
false postive rate	0.1057	0.1101
Negative predictive value	0.9265	0.9222

9. Conclusion:

- Important positive predictors of the case with decreasing coefficient values (top 5):-
 - Tags_Closed by Horizzon 8.2871
 - Tags_Lost to EINS 8.2116
 - Total Time Spent on Website 4.3958
 - Lead Origin_Lead Add Form 3.8077

5. Tags_Will revert after reading the email 3.6800

- Important negative predictors of the case with decreasing coefficient values:-
 1. Do Not Email -1.8844
 2. Tags_Ringing -1.6169
 3. What is your current occupation_Unemployed -1.5669
 4. Tags_switched off -1.0927
- **Conversion percentage of more than 80% obtained(86.85) as per the problem statement and requirement from CEO.**