Summary report on Lead Scoring Case study

CASE STUDY OBJECTIVE: The goal of the case study is to Build a logistic regression model to assign a lead score between 0 and 100 to each of the leads which can be used by the company to target potential leads. A higher score would mean that the lead is hot . i.e. is most likely to convert whereas a lower score would mean that the lead is cold and will mostly not get converted.

APPROACH AND LEARNINGS FROM CASE STUDY

1. Once the given data are imported , ,initial understanding of data is done. (i.e) nowing the total number of rows and columns , the statistical aspects , understanding the datatypes of each column etc.
2. The insight obtained from them is that , Some variable has missing data. And some variables have high data variability.
3. From the value_counts, performed on each column , we could see that the below variables has highly IMBALANCED data. So these will not significantly contribute to our model results. Hence we will tag them to be dropped. - Magazine - Newspaper Article - X Education Forums - Newspaper - Digital Advertisement - Through Recommendations - Receive More Updates About Our Courses - Update me on Supply Chain Content - Get updates on DM Content - I agree to pay the amount through cheque - What matters most to you in choosing a course - Search - Do Not Call - Do Not Email - Country - What is your current occupation
4. After this , exploratory data analysis is performed on the columns.
5. Initially we Calculated the % of Non Converted and Converted Leads in the Dataset
6. After calculating the Imbalance Percentage in lead DataFrame , we understood that Our Target variable is having a 62:38 ratio, and seems to be properly balanced with respect to the conversion ratio.
7. To further EDA analysis , Data Cleaning & Treatment is performed.
8. Handling **'Select'** values in the data:During initial analysis (value_counts) , we could see that there are many variables having 'Select' as the

categorical values. These are values that customer has clearly missed to add, so we will consider them as NULL values. We will convert all 'Select' to **'NaN'**

9. Then we checked unique value counts of city , specializations , How did you hear about X Education columns and then removed ‘select’ values.
10. Then we checked percentage of null values in each columns.
11. Then we dropped columns having nans of more than 40 %.
12. Then we explored categorical variables and merging less frequent values of categorical columns to 'Others' category.
13. Then we explored Numerical Variables & Outlier Handling.
14. For outlier handling we have deployed the **IQR CAPPING Method**. That is we will identify the outlier in each of the numeric variable and impute them with IQR * 1.5. This will help us to remove the Outlier also retain the rows.
15. All outliers are treated using IQR method and they are treated.
16. After this the final data frame is created by excluding the 'to_drop_list' columns.
17. The null values are checked again and the insights obtained is we retained ~99% of the data.
18. Then dummy is created.
19. Afrer this data is divided for train-test split.
20. After dropping highly correlated dummy variables , feature scaling is done.
21. Then we build the model. - **Logistic regression model**
22. Then Feature Selection Using RFE is done.
23. Then we Check for the VIF values of the feature variables.
24. Then we build the model after removing the variable with Insignificant P Value.
25. Then we predicted Probability Calculation.
26. Then we created a dataframe with the actual Converted flag and the Predicted probabilities.
27. Then we created Confusion Matrix.
28. Then we measured Accuracies.
29. Then we plotted the ROC Curve.
    - It shows the tradeoff between sensitivity and specificity (any increase in sensitivity will be accompanied by a decrease in specificity).
    - The closer the curve follows the left-hand border and then the top border of the ROC space, the more accurate the test.

- The closer the curve comes to the 45-degree diagonal of the ROC space, the less accurate the test.

30. Optimal Cutoff Point is calculated.
31. Then Precision and Recall trade-off is calculated.
32. Then Overall Accuracy of the Model is calculated to be 87 %.
    Sensitivity – 83
    Specificity – 90
    Precision – 84
    Recall - 83
33. For Maximizing the Conversion rate , **Lift & Gain Method** is used.

Conclusion:

~84% of Recall value indicates that our model is able to predict 84% of actual conversion cases correctly

~85% of Precision value indicates that 85% of the conversions that our model predicted is actually converted.

Top 3 variables which contribute most towards the probability of a lead getting converted.

1 . Lead origin_Lead Add Form.

2. Tags- will revert after reading the email.

3. Total time spend on website.

The major learnings from the case study are

1. Perform EDA for a very large number of columns.
2. Eliminating less frequent values together into “Others” category.
3. Handling outliers efficiently using IQR capping method.
4. Perform logistic regression on a dataset.
5. Calculate accuracy , precision , recall using confusion matrix.
6. Increasing the conversion rate using Lift and gain method.
7. Predicting the 3 top lead conversion columns.