LEAD SCORING SUMMERY REPORT

Submitted by:

1. Saikat Chowdhury
2. Aman Diwakar

CASE STUDY OBJECTIVE:

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.

The company markets its courses on several websites and search engines like Google. Once these people land on the website, they might browse the courses or fill up a form for the course or watch some videos. When these people fill up a form providing their email address or phone number, they are classified to be a lead. Moreover, the company also gets leads through past referrals. Once these leads are acquired, employees from the sales team start making calls, writing emails, etc. Through this process, some of the leads get converted while most do not. The typical lead conversion rate at X education is around 30%.

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. Importing Libraries and Checking the Data
2. Missing Value Treatment
3. Outlier Treatment
4. EDA
5. Multivariate Analysis
6. Data Preparation
7. Train Test Split
8. Feature Scaling
9. Feature Selection Using RFE & Further by Manual Inspection
10. Finding Optimal Cutoff Point Using ROC
11. Metrics beyond simply accuracy & Plotting the ROC Curve
12. Precision,Recall and Specificity
13. Making predictions on the test set
14. Over all model accuracy
15. Top three variables in model which contribute most

Description:

1. Importing Libraries and Checking the Data:

This involves importing the necessary libraries and understanding the data, its dimensions, and the data types of the columns. Identifying the target variable and the variables out of which feature selection will be done at a later stage.
Duplicates were also checked.

2. Missing Value Treatment:

The insight obtained from them is that, some variable has missing data. And some variables have high data variability.
The missing values in the case of different columns were replaced by either mean, median, or mode depending on the type of data in the column, i.e., numerical or categorical, and keeping in mind their business significance.

3. Outlier Treatment:

After looking at the percentile values in the columns outliers were removed from the Total Visits and Page Views Per Visit columns. Boxplots were used to confirm the data after cleaning.
Some outlier data is deleted which may skew our results by making our model less accurate.
Some columns which will not contribute to our analysis have been dropped.
Example - What matters most to you in choosing a course, columns that have more than 40% null values have been dropped, etc.
After cleaning data 98% data has been retained. There are 8953 entries and 14 columns.

4. EDA:

Data were analyzed after grouping according to the target variable. The relation of some of the variables with the Converted data was seen using count plots and cat plots.

5. Multivariate Analysis:

Pair plots and heat maps were created to understand the correlation between the numerical variables.
Total Visits and Page Views Per Visit has a strong correlation. In such a case, one may be dropped before analysis starts

6. Data Preparation:

Based on the EDA analysis it is seen that many columns are not adding any information to the model, hence we drop them before further analysis. After dropping the unnecessary columns for categorical variables with multiple levels, dummy features using one hot encoding was done.

7. Train Test Split:

Data is divided into Train and Test.

8. Feature Scaling:

Features that require scaling are done so using the Standard Scaler.

9. Feature Selection Using RFE & Further by Manual Inspection:

After EDA, Logistic Regression Model is built in python using GLM() function, under stats model library.
The model contained all the variables, some of which had insignificant coefficients.
Such variables are removed using Automated Approach: RFE (Recursive feature elimination) with number of features = 15.
Manual approach based on VIFs and p values.
The final tally of variables with their respective values. Significant p-values near to zero VIFs < 3

10. Finding Optimal Cutoff Point Using ROC:

Our final model gave us the following performance metrics.
We used Recall/Precision trade-off graph to derive the optimal threshold value.
Cut-off point 0.3

11. Metrics beyond simply accuracy & Plotting the ROC Curve:

ROC Curve demonstrates Trade-off between sensitivity and specificity.
Closer the curve follows the left-hand border and then the top border of ROC space, the more accurate the test closer the curve comes to 45°diagonal of the ROC space, the less accurate the test.

For our model, ROC curve is towards the upper left corner, and area under the curve is more. Thus, our model is an optimal choice to move forward with the analysis.

12. Precision, Recall and Specificity:

Precision, Recall and Specificity values of test set are around 88%, 91% and 92% which are approximately closer to the respective values calculated using trained set.

13. Making predictions on the test set:

The confusion matrix is created Precision, Recall and Specificity values of test set are around 89%, 91% and 93% which are approximately closer to the respective values calculated using test set based on various probability cutoffs are calculated.

14. Over all model accuracy:

Train data set over all model accuracy: 92.29%
Test data set over all model accuracy: 92.77%
Hence, overall this model seems to be good.

15. Top three variables in model which contribute most:

1. Lead Origin_Lead Add Form
2. Tags_Will revert after reading the email
3. Last Activity_SMS Sent

<u>The major learnings from the case study are</u>:

- Perform EDA for a very large number of columns.
- Eliminating less frequent values together into “Others” category.
- Handling outliers efficiently.
- Perform logistic regression on a dataset.
- Calculate accuracy, precision, recall using confusion matrix.
- Predicting the 3 top lead conversion columns