Lead Score Case Study Summary

Submitted by,

Mohana Raja Manohar

Chinmayee Shreevatsa

DS C6 INTL May 2022

Problem Statement

An education company named X Education sells online courses to industry professionals. Although X Education gets a lot of leads, its lead conversion rate is very poor. The problem is to increase the conversion rate of potential leads.

Business Goal

- To build Logistic Regression Model to predict the potential leads.
- A lead score between 0 and 100 is to be assigned to each of the leads, which the company can use to target potential leads.

Processes followed and the results obtained

1. Imported necessary libraries, read and understood the data with basic analysis.

2. **Data cleaning**

   The data was partially clean except for the few null values and the option 'Select' has to be replaced with NaN. Few of the null values were updated with other categorical values so as to not lose much data.

3. **Missing value treatment**

   Features having missing percent of more than 45% and less than 2% of the data are dropped from the data set. The rest of the missing values are imputed with the most frequent values.

4. **Outlier Analysis**

For categorical features, there is no need for outlier treatment. Still, the categories containing fewer counts are grouped together to form a new category.
For the numerical features, the outliers are treated and proceeded with the model building.

5. **Data Preparation**

- Binary variables are converted to 0 or 1.
- Dummy variables are created for the other categorical features.

6. **EDA**

- The conversion rate of the hot leads is 38% of the total data.
- Few features were removed as they were the least important and had high variance.

7. **Train Test Split**

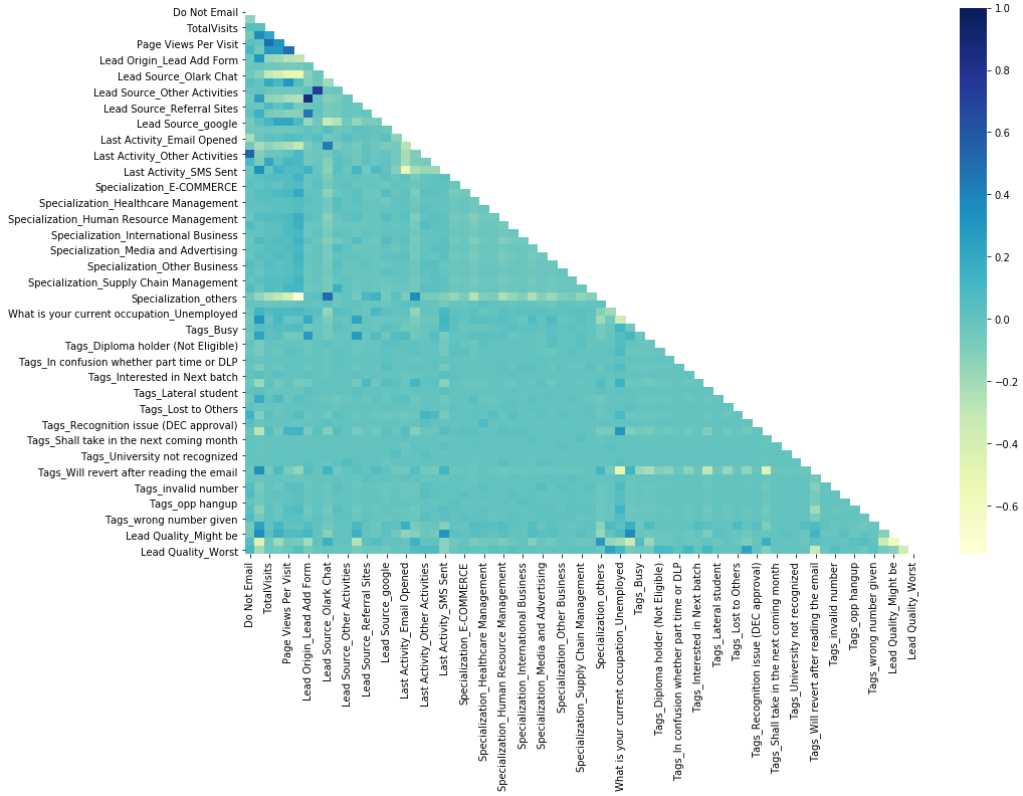
Train test split is done in the ratio of 70% train set and 30% test set.

8. **Feature Scaling**

Feature scaling is handled by applying Standard Scaler for the continuous features.

9. Correlation Matrix

A few features were found to be multi-collinear and were removed.



10. Base Model Creation

Created a base model using a Generalized linear model to understand the detailed summary and check for the p-values.

11. Logistic Regression Model

Model creation is done using the Logistic Regression from the scikit learn library.

12. Recursive Feature Elimination

Applied RFE for feature selection of top 15 relevant features.

13. Variance Inflation Factor

Applied the VIF to find the multicollinearity between the features.

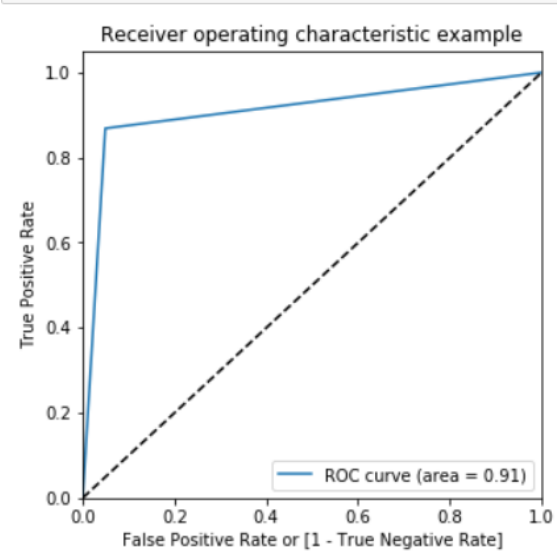
14. Feature Elimination

Feature elimination was done manually if the p-value of the features in the GLM report was greater than 0.05 or VIF is greater than 5.

15. The procedure (10-14) is repeated until the criteria is met for p-value and VIF.

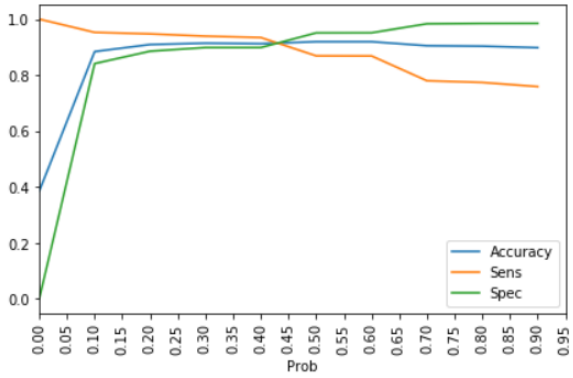
16. AUC-ROC

The AUC and ROC is obtained by plotting FPR vs TPR.



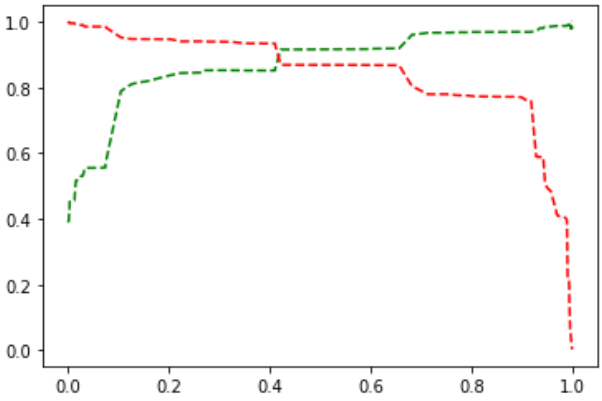
17. Finding Optimal threshold

After the model is built, the model is verified for the various threshold values for the conversion.



18. Precision-Recall Tradeoff

Precision-Recall trade-off is plotted with the obtained threshold value.



19. Model Evaluation

The model has evaluated for the below metrics and the result is as follows.

Metrics	Train set	Test set
Accuracy	91.21%	91.22%
Precision	91.71%	89.64%
Recall	86.88%	85.74%
Specificity	95.08%	94.35%

20. The top 3 variables

1. Tags_Lost to EINS
2. Tags_Closed by Horizzon
3. Tags_Will revert after reading the email

21. Hot Leads Conversion Rate

The model estimates that the leads with more than a 3.5 conversion threshold value have a 90% chance of being converted and the leads with higher threshold value of 9.5 have 98% chance of being converted.