Lead Score Assignment - Summary Report

Table of Contents

I. PROBLEM STATEMENT 2

II. GOAL 2

III. ANALYSIS APPROACH 2

1. EDA 2
2. Build the logistic regression model 3

IV. FINAL MODEL & EVALUATIONS 5

1. Coefficients of final model 5
3. Find the optimum cut-off 5
2. Evaluate the TRAIN dataset 7
3. Evaluate the TEST dataset 7
4. Conclusion 8

V. BUSINESS RECOMMENDATION 8

I. Problem Statement

X-Education is an education company sells online education courses to professionals and advertises it on the website and search engine. Company gets the data and identify the leads and it conversion. From the collected data and inferences the conversion leads is very low, about 30%. Company wishes to identifying Hot Leads i.e. the most potential leads, also it wants to achieve 80% of the visit and enrollment.

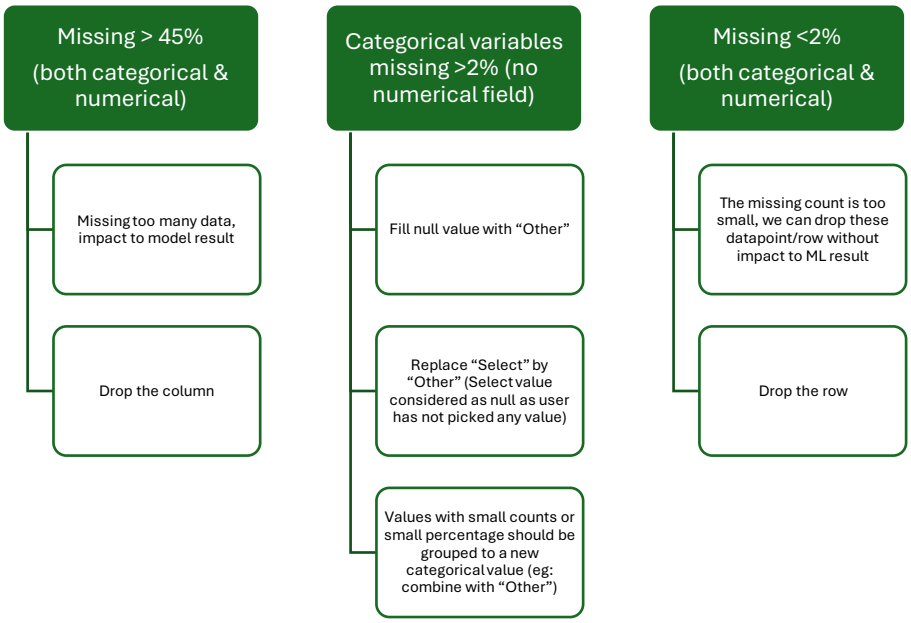
II. Goal

Building logistics regression model to finding lead score between 0 and 100 for Company and help to achieve potential targets 'Hot leads'. For a futuristic change the model should be ready an flexible to accommodate the and predict the outcome.

III. Analysis Approach

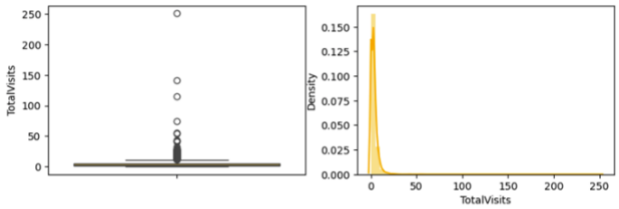
1. EDA

- Handling missing data and categorical data



- For numerical data, we will handle the outliers

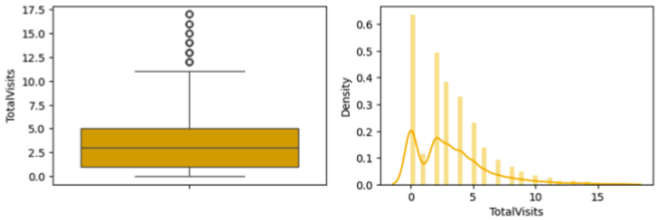
Boxplot showing that there're many outliers datapoints



Quantile 99th is very faraway (17) from max value (251)

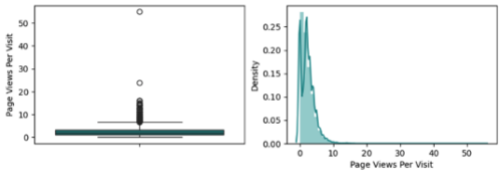
Quantile	Value
0.95	10
0.98	13
0.99	17
1	251

After remove outliers > 99th percentile



Some datapoints seem be outliers but it's not too far vs. original datapoints. This is acceptable!

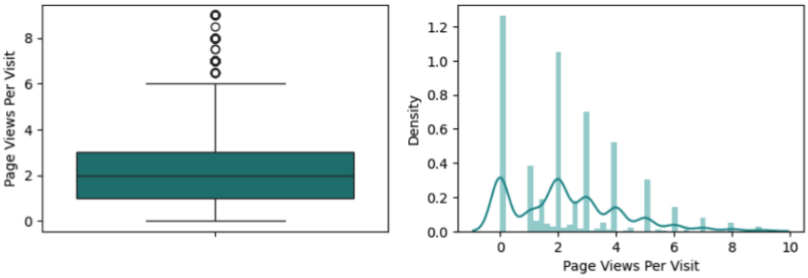
Boxplot showing that there're many outliers datapoints



Quantile 99th is very faraway (9) from max value (55)

Quantile	Value
0.9	5
0.95	6
0.98	8
0.99	9
1	55

After remove outliers > 99th percentile



Some datapoints seem be outliers but it's not too far vs. original datapoints. This is acceptable!

2. Build the logistic regression model

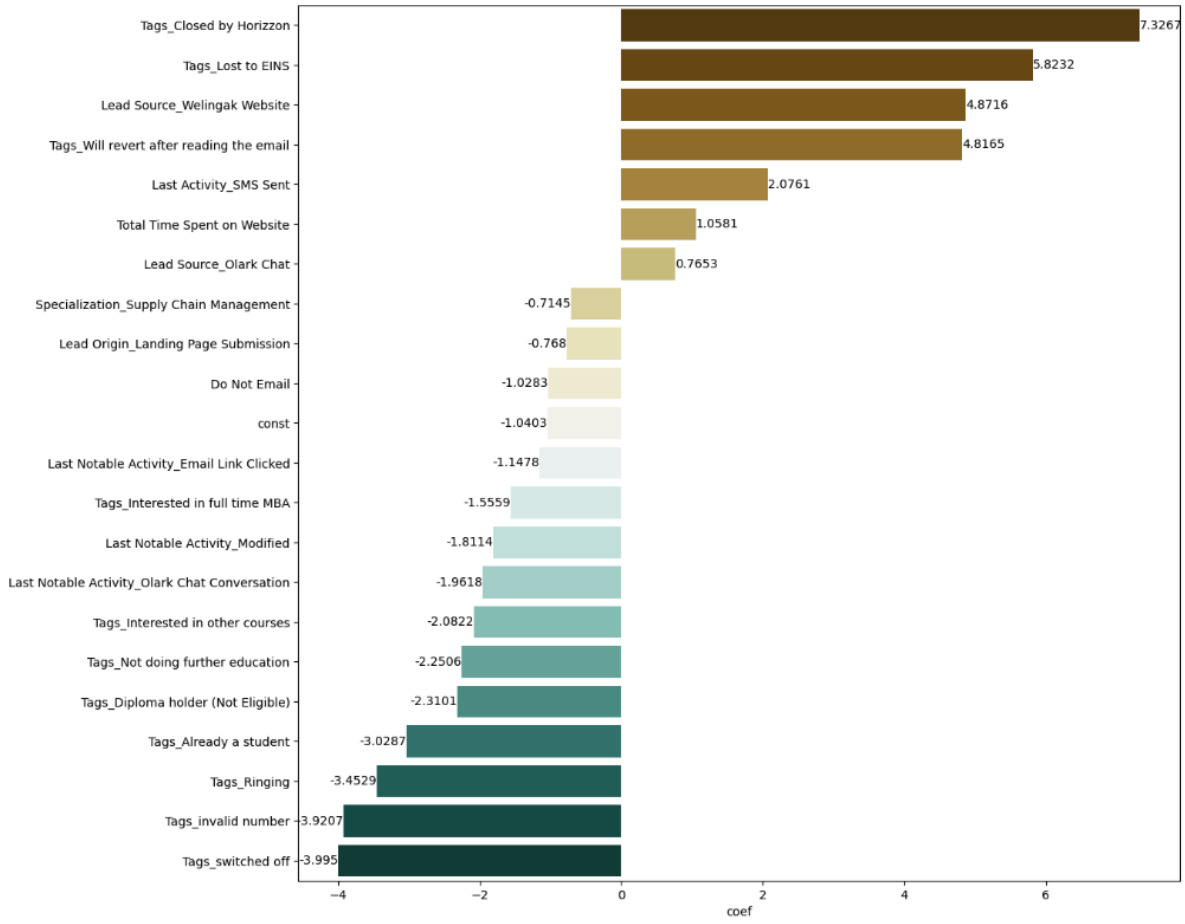
- A. Data Transformations for modelling
    1. Encoding binary categorical (Yes/No --> 1/0)
    2. Encoding categorical with dummy variables
        a. Auto Dropping 1st dummy var
        b. Manual Dropping a selected dummy var
- B. Model Preparation
    1. Check dataset imbalance
    2. Create X and y
    3. Split TRAIN-SET

4. Scale numerical data for TRAIN set

- C. Auto-Eliminate Features with RFE
    1. Start running RFE with number of selected features = 50% to eliminate features
    2. Build logistic regression with statsmodel for interpreting easily
    3. Verify number of insignificant features <= 12 (p-Value>0.05)
    4. If number of insignificant>12, repeat step 1 with new smaller number of selected feature
- D. Manual Eliminate Features
    1. Prepare list of features to build model, initially with selected features from RFE
    2. Build logistic regression model with statsmodel with selected features and fit the model
    3. [Checkpoint 1] If there's insignificant feature (pValue>0.05), remove the the feature with highest pValue from selected feature, and back to step 2
    4. [Checkpoint 2 – No insignificant feature] Calculate VIF score to detect the multicollinearity issue (VIF>5). If there is VIF>5, we remove the feature with highest VIF from selected features and back to step 2
- E. Final model
    1. All features are significant (pValue<=0.05)
    2. All features do not have multicollinearity problem (VIF<=5)
    3. Build the confusion matrix
    4. Calculate model metrics:
        a. Accuracy score
        b. Sensitivity vs. specificity
        c. TPR vs. FPR
        d. Precision vs. Recall
        e. Positive Prediction Ratio vs. Negative Prediction Ratio

IV. Final model & Evaluations

1. Coefficients of final model

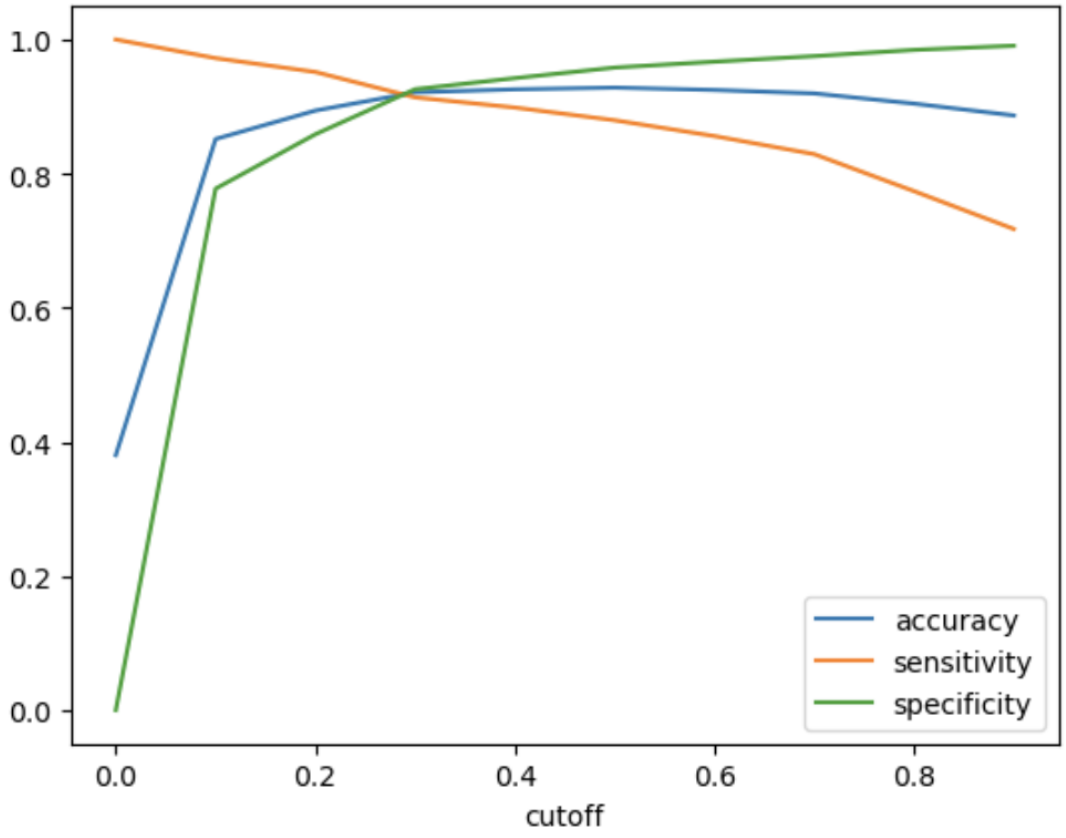


3. Find the optimum cut-off

- Step 1: Metrics for different Cut-off values

cutoff	accuracy	sensitivity	specificity	TPR	FPR	precision	recall	PP_ratio	NP_ratio
0.0	0.380243	1.000000	0.000000	1.000000	1.000000	0.380243	1.000000	0.380243	NaN
0.1	0.851585	0.972211	0.777577	0.972211	0.222423	0.728391	0.972211	0.728391	0.978544
0.2	0.894012	0.951579	0.858693	0.951579	0.141307	0.805130	0.951579	0.805130	0.966560
0.3	0.921230	0.913684	0.925859	0.913684	0.074141	0.883191	0.913684	0.883191	0.945896
0.4	0.925712	0.898526	0.942392	0.898526	0.057608	0.905388	0.898526	0.905388	0.938030
0.5	0.928274	0.879579	0.958150	0.879579	0.041850	0.928032	0.879579	0.928032	0.928411
0.6	0.924752	0.856000	0.966934	0.856000	0.033066	0.940768	0.856000	0.940768	0.916279
0.7	0.919629	0.829053	0.975200	0.829053	0.024800	0.953511	0.829053	0.953511	0.902894
0.8	0.904419	0.773895	0.984500	0.773895	0.015500	0.968388	0.773895	0.968388	0.876495
0.9	0.886808	0.717474	0.990700	0.717474	0.009300	0.979310	0.717474	0.979310	0.851087

- Step 2: Find the optimum cut-off



→ The optimum cutoff is 0.3 (intersect of accuracy, sensitivity and specificity)

2. Evaluate the TRAIN dataset

```
============Confusion Matrix:==========
[[3584  287]
 [ 205 2170]]
============Metrics:==========
  1. Accuracy score:  0.9212295869356388

  2.1 sensitivity (TPR):  0.9136842105263158
  2.2 specificity:  0.9258589511754068

  3.1 TPR:  0.9136842105263158
  3.2 FPR:  0.07414104882459313

  4.1 Precision:  0.8831908831908832
  4.2 recall:  0.9136842105263158

  5.1 Possitive_Prediction_ratio:  0.8831908831908832
  5.2 Negative_Prediction_ratio:  0.9458960147796253
```

3. Evaluate the TEST dataset

```
============Confusion Matrix:==========
[[1555  129]
 [  79  915]]
============Metrics:==========
  1. Accuracy score:  0.9223300970873787

  2.1 sensitivity (TPR):  0.920523138832998
  2.2 specificity:  0.9233966745843231

  3.1 TPR:  0.920523138832998
  3.2 FPR:  0.07660332541567696

  4.1 Precision:  0.8764367816091954
  4.2 recall:  0.920523138832998

  5.1 Possitive_Prediction_ratio:  0.8764367816091954
  5.2 Negative_Prediction_ratio:  0.9516523867809058
```

4. Conclusion

All metrics are good for both TRAIN and TEST dataset, equivalent between TRAIN and TEST. Especially the accuracy score, precision and recall are important metrics for this business:

- Precision: number of predicted YES correctly over Total predicted YES (converted=1)
- Recall: number of predicted YES correctly over Total actual YES (converted=1)

V. Business Recommendation

1. The hot leads are the leads who predicted as "1". Company should priority to make the phone call/emails to these leads to pursuade them to buy the courses.
2. Among the leads who are predicted as "1", the Lead_score (0..100) are important to prioritize. The higher scores, the higher probability to convert to customer. So company should spend more time to call/email and follow up / take care these leads.
3. For the leads who have the same scores, we should focus on the leads who have top contributed features (highest coefficients), as they are likely want to buy the courses. The top 3 features are:
    a. Tags_Closed by Horizzon (coef = 7.3267)
    b. Tags_Lost to EINS (coef = 5.8232)
    c. Lead Source_Welingak Website (coef = 4.8716)
4. For the leads who have negative coefficients (from the bottom to top), company should avoid to call/emails these leads, as they likely don't have any needs to buy a courses. This will help company to save the cost. The bottom 3 features to avoid are:
    a. Tags_switched off
    b. Tags_invalid number
    c. Tags_Ringing