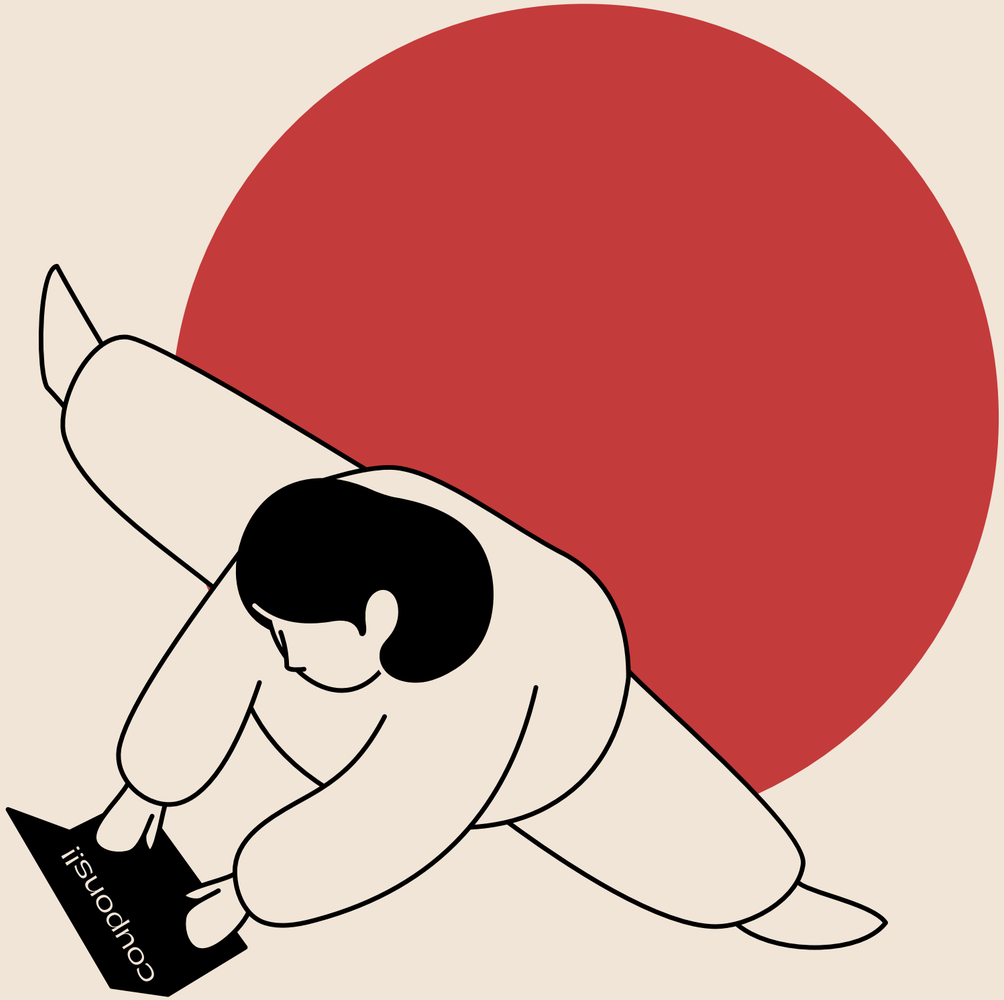
Campaign Response Model



TIMPIKA AUNGNOI 6310412025

Response Model

predicts the probability that a customer is going to respond to a promotion or offer. Response models are typically built from historical purchase data.

Response modeling is usually formulated as a binary classification problem. The customers are divided into two classes, respondents and nonrespondents.

The response modeling procedure consists of several steps such as data collection, data preprocessing, feature construction, feature selection, data balancing, classification, and evaluation.

References:
https://www.sciencedirect.com/science/article/pii/B9780124115118000062

Objective

The objective of this assignment is to improve model accuracy and select the best model based on AUC. In addition, the goal of this model is to predict whether an existing customer will purchase in the next campaign by using historical data. The independent variables include RFM and CLV variables.

The response model is a classification model. In this analysis, classification methods were used for response modeling including Logistic regression and XG Boost.

Data Analysis Process

1.Data loading

Dataset is Retail_Data_Response.csv
Retail_Data_Transactions.csv

2.Data preparing

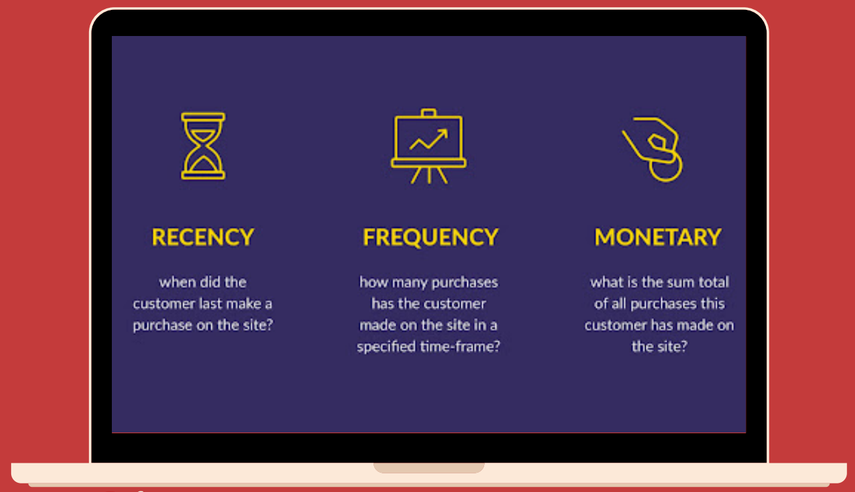
Dividing the data into two groups of independent variables to predict campaign response:

Group 1 (RFM variables)

- Recency
- Frequency
- Monetary

Group 2 (CLV variables)

- Recency
- Frequency
- Monetary
- AOU
- Ticket size



References:
http://rrtechguru.blog/rfm-analysis-customer-segmentation/

Data Analysis Process

3. Split dataset into train and test sets

4. Fixing data imbalance by resampling techniques

- Undersampling
- Oversampling
- SMOTE

5. Model selection

- Logistic regression
- XG Boost

6. Model evaluation

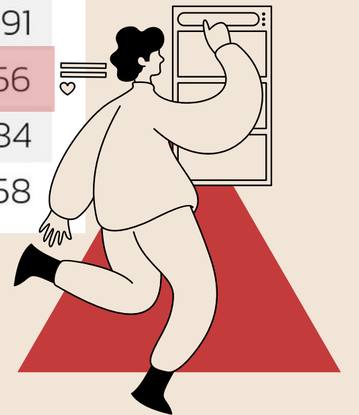
- Confusion matrix

Result

Comparing model by using AUC score. The best model is XG Boost with CLV variables by fixing imbalance with Undersampling method. AUC train is 0.76691 and AUC test is 0.74235 (maximize) . Moreover, We will choose a model where the difference between these two values does not more than 0.03 which is 0.02456 (0.76691-0.74235)

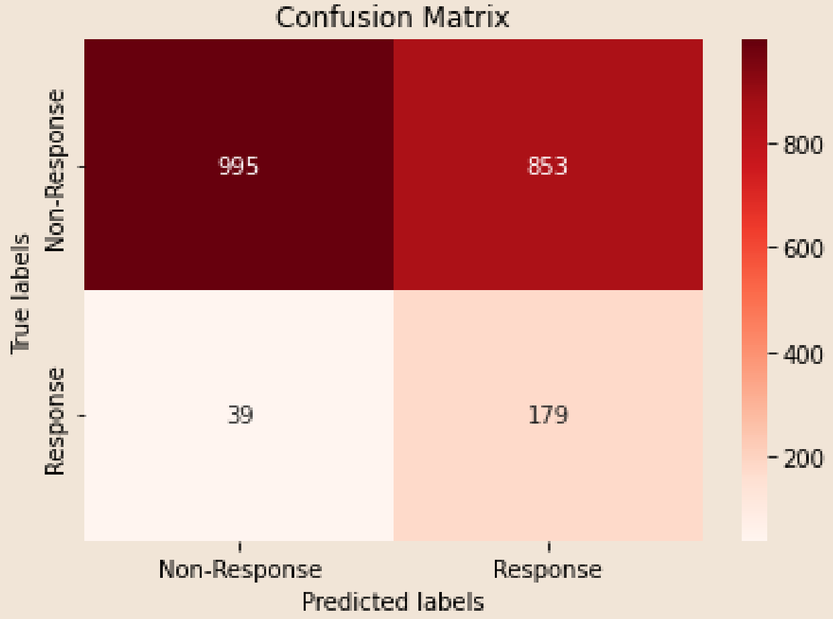
Model	Variables	Classification Method	Resampling	AUC train	AUC test	The difference of AUC train and test
1	RFM	Logistic regression	Undersampling	0.73004	0.72017	0.00987
2	RFM	Logistic regression	Oversampling	0.71814	0.72561	0.00747
3	RFM	Logistic regression	SMOTE	0.72559	0.70788	0.01771
4	CLV	Logistic regression	Undersampling	0.73173	0.73329	0.00156
5	CLV	Logistic regression	Oversampling	0.71737	0.73329	0.01592
6	CLV	Logistic regression	SMOTE	0.74016	0.70104	0.03912
7	RFM	XG Boost	Undersampling	0.75986	0.72476	0.0351
8	RFM	XG Boost	Oversampling	0.74828	0.73897	0.00931
9	RFM	XG Boost	SMOTE	0.74754	0.73563	0.01191
10	CLV	XG Boost	Undersampling	0.76691	0.74235	0.02456
11	CLV	XG Boost	Oversampling	0.75273	0.74089	0.01184
12	CLV	XG Boost	SMOTE	0.76304	0.72346	0.03958

RFM variables include Recency Frequency Monetary
CLV variables include Recency Frequency Monetary AOU Ticket size



Confusion Matrix

The predicted result of the best model is shown by the confusion matrix. it contains information about actual and predicted classifications.



It can be seen in the confusion matrix, 1034 corresponds to the number of customers that the model classified as non-respondents and 1032 corresponds to the number of customers that the model classified as respondents.

995 customers out of 1848 customers were predicted as non-respondents who were actually non-respondents (True negative), also 179 customers out of 218 customers were predicted as respondents who were actually respondents. (True positive)

853 out of 1848 customers were predicted as respondents while they were nonrespondents (False positive). There were also 39 out of 218 customers predicted as nonrespondents who were actually respondents (False negative)

References:
https://www.sciencedirect.com/science/article/pii/B9780124115118000062
https://www.sciencedirect.com/topics/computer-science/confusion-matrix