RUTGERS

Classification of Retail Customers

Zhen Qian (Martin)
(Quantitative Finance Program of Rutgers Business School)

May 2016

Problem Statement

Data Set Source :

A retail company has collected the data sets of customer behaviors

Data Mining Objectives:

Development algorithms or methods to prediction the customer loyalty

Classification Accuracy Definition:

The target variable is “Active_Customer”, 1 means loyal, 0 means not loyal.

Accuracy=predicted results matches actual results/ total prediction results

***Please see notes for details.

2

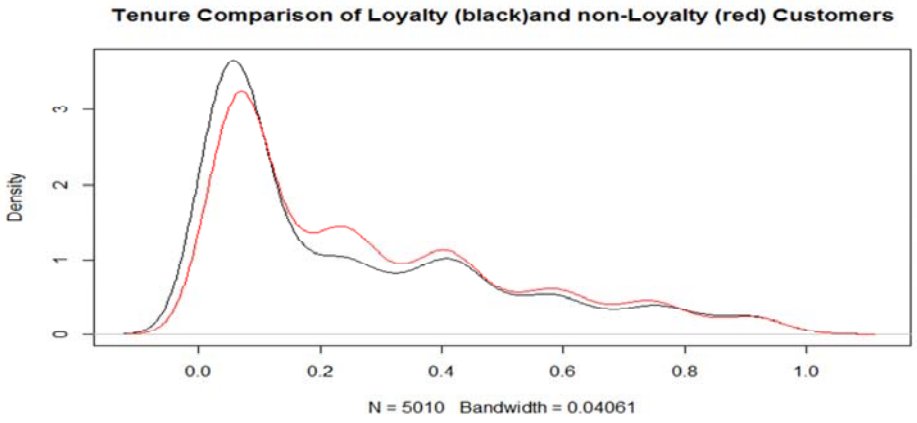
Attached files: (1)Accuracy definition file: accuracy.pdf. (2) train data: Train.csv.(3)test data for prediction: Test.csv. (4) sample submission data: Sample_Submission.csv. You should fill up the Active_Customer results for test data based on your prediction.

Data Mining Analysis: Big

- **Brief Description of Data Set :**
  - (1) Cust_status (Old or New)
  - (2) Cust_Tenure (Income)
  - (3) 41 trans records
  - (4) 164 Food records
  - (5) 48 Promotion records
  - (6) Active_Customer for company loyalty (1 means loyal, 0 means not)
- **Use statistical language to short calculations:**
  - (1) Mean, STD(Standard Deviation), Freq(response times) and SUM (total amount) for three items-Trans, Food, Promotion records.
  - (2) So, the calculation is based on Cust_Tenure, trans_mean, trans_std, trans_freq, Trans_sum, food_mean, food_std, food_freq, food_sum, pro_mean, pro_std, Pro_freq, pro_sum and Active_Customer.

3

Initial Analysis of Data Sets



4

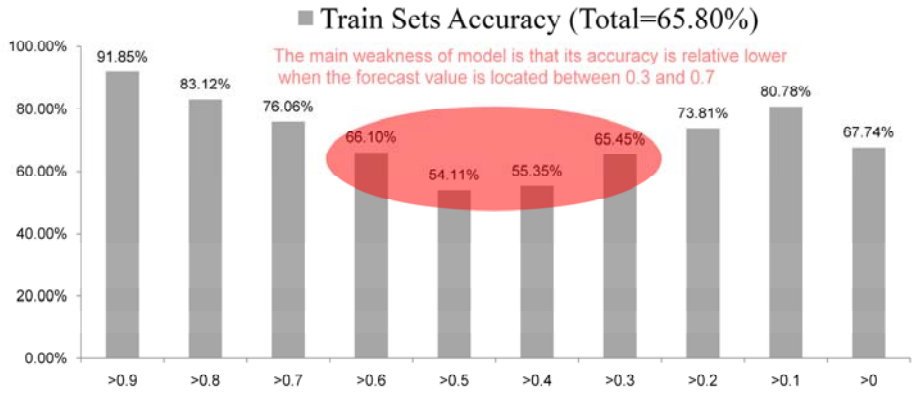
(1) Tenure distribution is used to illustrate the task difficulty, other items have the similar situations.

(2) Non traditional distribution, fat tail, high positive skewness and high standard deviation.

(3) Very close sharp for loyalty and non-loyalty customers.

(4) The task is great challenging and I decide to apply ANN algorithm first.

Predict Target Variable By ANN Algorithm



5

The model based on ANN algorithm will generate a forecast value, and it will round to 1 if it is greater than 0.5 or 0 if it is less than 0.5.

Further Optimization

- Firstly, I need a server to build 300 neural cells rather than 15 neural cells for training and modeling, but it is impossible.
- Second, train more iterations to improve model. However, it cannot breakdown at the accuracy bottleneck.
- Third, use another method to predict data sets that ANN model's prediction value is located between 0.3 and 0.7.

6

Conclusion

- It is a great framework to use ANN algorithm.
- The further optimization process is more difficult than imagining. I have tried to use Gradient boosting algorithm, a machine learning technique to build another model. However, how to hybrid two methods to produce a win-win effect is a great challenging.

Final Results

	Group.Name	University	Score	Time
1	Columbia-ds	Columbia University	0.711207	6/1/2016 21:13
2	Dataminers	North Carolina State University	0.704545	6/1/2016 21:37
3	Olympian	University of Texas at Dallas	0.693182	2016-06-02 18:02:16
4	hundred	DePaul University	0.690168	6/1/2016 20:21
5	CHK	Stevens Institute of Technology	0.689004	2016-06-02 10:44:29
6	White Sox	DePaul University	0.688616	6/2/2016 7:19
41	Arpita	Illinois Institute of Technology	0.667270	2016-06-02 14:48:10
42	BENCH MARK	RANG-KVRA	0.667141	5/27/2016 14:30
43	Hofstra Pride	Hofstra University	0.655886	6/1/2016 22:51
44	SmartSnake	Rutgers	0.653169	2016-06-03 20:11:48
45	Ms. Doudou	University of Southern California	0.634929	6/1/2016 22:46

8

Thank you!

9