Predicting Loyal Customers for Sellers on Tmall to Increase Return on Promoting Cost

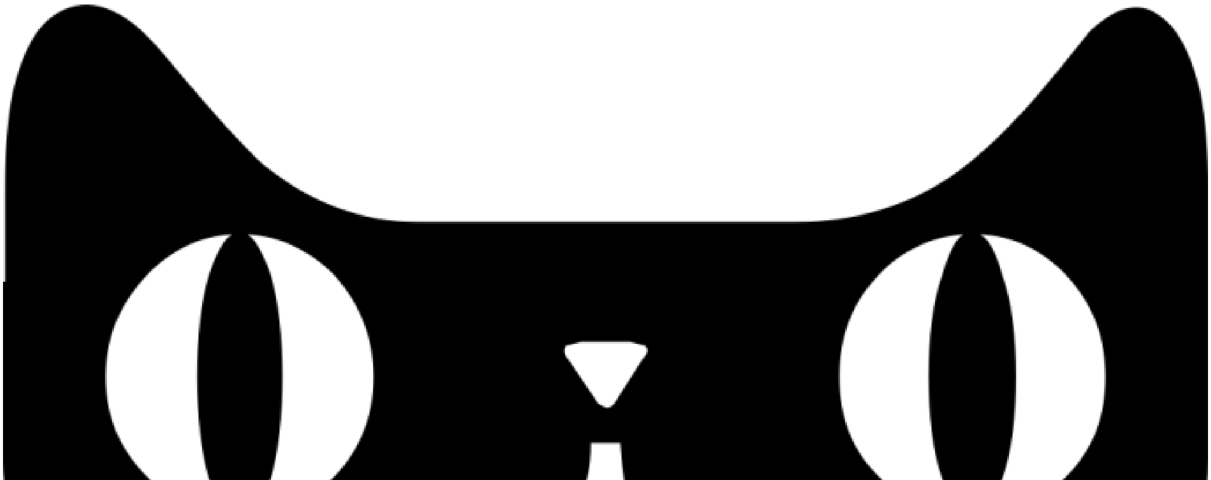
Wendy Huang

Yu-Chih, Shih

Jessy Yang

Zoe Cheng

BADM Team 9



- **Summary**

Sellers on E-commerce platform sometimes run big promotions (e.g., discounts or cash coupons) on particular dates (e.g., Boxing-day Sales, "Black Friday" or "Double 11 (Nov 11th)", in order to attract a large number of new buyers. Unfortunately, many of the attracted buyers are one-time deal hunters, and these promotions may have little long lasting impact on sales. To alleviate this problem, it is important for sellers to identify who can be converted into repeated buyers, in other words, the loyal customers. By targeting on these potential loyal customers, sellers can greatly reduce the promotion cost and enhance the return on investment (ROI). It is well known that in the field of online advertising, customer targeting is extremely challenging, especially for fresh buyers. However, with the 6 months user behavior log accumulated by Tmall.com, we may be able to solve this problem.

To increase the return on promotion cost, it is definitely useful to build a model to predict which new buyers for given sellers will become loyal customers in the future.Therefore, our stakeholders are the sellers on Tmall and our goal is to predict loyal customers for sellers on Tmall to increase the return on promotion cost. We use the data given by Tmall, which is a set of sellers and their corresponding new buyers acquired during the promotion on the "Double 11" day, to to derive a supervised classification model predicting the probability that these new buyers would purchase items from the same sellers again within 6 months.

By trying the logistic regression, Random Forest, and Xgboost to build the model, we finally chose the Random Forest with data from 5/11 to 10/25 and cut off probability 0.588 (which is the probability of top 10% in the outputs) as our final model. However, there are some implementing limitation and potential model risks in the models. We couldn’t guarantee the stablenes to our model. For business policy, we recommend to collect more data about characteristics of the sellers; in addition, increase the customer’s willingness to add their product into favorite; last but not least, lurkers are potentially loyal customers, which the sellers should not ignore these people.

Data source: https://goo.gl/GftTVM

1. Problem description

- **Business Goal**

  Sellers on Tmall usually use discounts and coupons to attract customers, but most of these customers are one-time buyers which causes low return on the promotion cost. By identifying who are the potential loyal buyers to this seller, the seller can spend the promotion cost on potential loyal buyers. However, implementing this idea might lead to the ignorance of the buyers on the long tail. The increase of return on the promotion cost would be considered a success.

- **Analytics/Data Mining Goal**

  To achieve the business goal, we need to predict how likely this buyer will be the loyal customer to this seller after his/her purchase so it is a supervised and predictive task. The outcome variable of interest is the probability of the customer will be loyal to the seller.

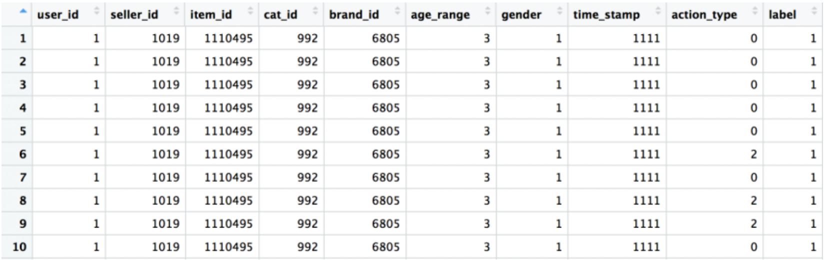
2. Data Description

An open data provided by Tmall which is the largest business-to-consumer (B2C) retail platform in Asia enabling businesses to sell directly to millions of consumers throughout China.

At the first time, there are total 10 columns and over 2,824,241 row data. The data set contains anonymized users' shopping logs in the past 6 months before and on the "Double 11" day,and the label information indicating whether they are loyal customers. The shopping logs columns is based on actual user activity on the platform, where 0 is for click , 1 is for add-to-chart, 2 is for purchase and 3 is for add-to-favorite. And also the other transaction recording, including user_id, seller_id(the same to seller_id), item_id, cat_id, and brand_id, age_range, gender, and timestamp. These columns offer the information about commodities and users.

There are around 10% loyal customers in the records , and we define that the customers who bought again at the same seller in the next 6 month after “Double 11” is a loyal customer.

Table 1: Sample Data (10 rows and 10 columns)

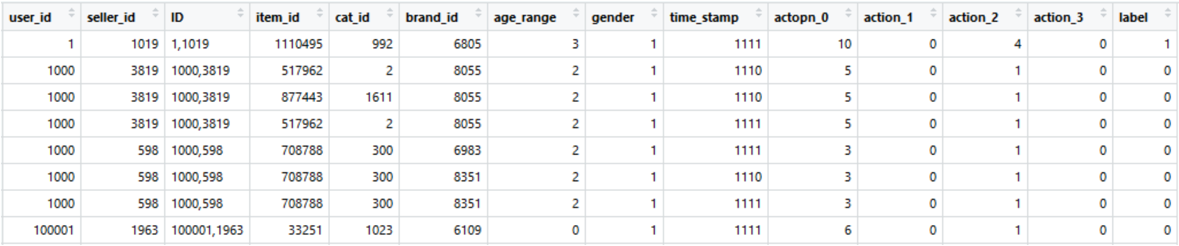


	user_id	seller_id	item_id	cat_id	brand_id	age_range	gender	time_stamp	action_type	label
1	1	1019	1110495	992	6805	3	1	1111	0	1
2	1	1019	1110495	992	6805	3	1	1111	0	1
3	1	1019	1110495	992	6805	3	1	1111	0	1
4	1	1019	1110495	992	6805	3	1	1111	0	1
5	1	1019	1110495	992	6805	3	1	1111	0	1
6	1	1019	1110495	992	6805	3	1	1111	2	1
7	1	1019	1110495	992	6805	3	1	1111	0	1
8	1	1019	1110495	992	6805	3	1	1111	2	1
9	1	1019	1110495	992	6805	3	1	1111	2	1
10	1	1019	1110495	992	6805	3	1	1111	0	1

3. Data Preparation

First, we merge all information into log file by user_id and seller_id and impute the missing value in gender and age by MICE method. Then, we separate actioin_type into action_0, action_1, action_2, action_3 and sum up the frequency of action_0, action_1, action_2, action_3 according to user, seller ID, time_stamp.

Table 2: Processed data

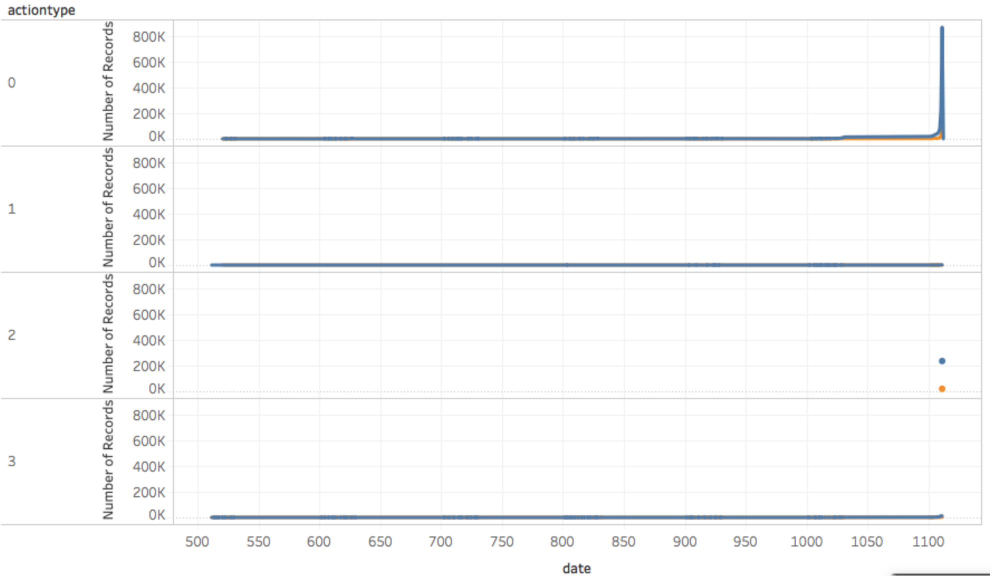


user_id	seller_id	ID	item_id	cat_id	brand_id	age_range	gender	time_stamp	actopn_0	action_1	action_2	action_3	label
1	1019	1,1019	1110495	992	6805	3	1	1111	10	0	4	0	1
1000	3819	1000,3819	517962	2	8055	2	1	1110	5	0	1	0	0
1000	3819	1000,3819	877443	1611	8055	2	1	1110	5	0	1	0	0
1000	3819	1000,3819	517962	2	8055	2	1	1111	5	0	1	0	0
1000	598	1000,598	708788	300	6983	2	1	1111	3	0	1	0	0
1000	598	1000,598	708788	300	8351	2	1	1110	3	0	1	0	0
1000	598	1000,598	708788	300	8351	2	1	1111	3	0	1	0	0
100001	1963	100001,1963	33251	1023	6109	0	1	1111	6	0	1	0	0

The second problem is the variable selection problem. According to the variable importance from Xgboost, we use age, gender, action_0~3 as predictors, and drop out item_id, cat_id, brand_id, and timestamp. The results of variable selection are shown in Appendix.

Lastly, after data exploration, we figure out there are some patterns according to the action type of users based on the timestamp. Therefore we have decided to separate the data into 3 parts by timestamp, which means there are now 3 data sets, including the data in 6 months, between 5/11 to 10/25 and between 10/26 to 11/11. Considering there supposed to have a warm-up activities before Double 11, we pick the date 10/25, which is two weeks before Double 11,to avoid the bias from the timestamp.Some visualizations are shown below.

Figure1: Visualization of action type based on timestamp



4. Data Mining Solution

- **Algorithm**

We partition our data set into 3 subsets: 50% training data, 30% validation data and 20% test data. Firstly, we use the original data set to build the model; however, the accuracy and sensitivity of the results is almost the same to the Naive rules. Therefore, considering there are only around 10% loyal customers in the data set, we decide to use undersampling method to revise it. Since our output to predict *(label)* is categorical, we apply supervised classification models, including logistic regression, random forest and xgboost. The lift charts of each algotithm on different datasets are shown in Appendix A.

- **Model Performance**

We take Naive rule as a benchmark and rank the probability and take the top 10% and set the cutoff value equaling to the min probability of the top 10%, which make our top 10% prediction = 1. Then, we check the accuracy of these top 10% because we assumed that when implementing the model, the seller would send out the vouchers to the top 10% possible customers. That is to say, the accuracy of the top 10% means the ROI when the seller uses this model. According to the tables below, the final model we choose is **random forest with dataset from 5/11 to 10/25**.

Benchmark_Naive Rule: 10%

Table4: data set = All

Model	Logistic	Random forest	Xgboost
Accuracy of top 10%	**23.69%**	**25.78%**	**20.43%**

Table5: data set is from 5/11 to 10/25 (befor double11-related campaigns)

Model	Logistic	Random forest	Xgboost
Accuracy of top 10%	**33.96%**	**48.86%**	**20.56%**

Table6: data set is from 10/26-11/02 (during warm-up and double 11 campaigns)

Model	Logistic	Random forest	Xgboost
Accuracy of top 10%	**20.4%**	**24.61%**	**18.32%**

- **Strategies we failed**

1. We’ve tried to divide sellers into different groups by how many brands one holds. Because we think that customers from different kind of sellers may act differently. However, we haven’t find significant difference.
2. We’ve tried to divide sellers by their transaction amount. Because we thought the top sellers may have more loyal customers. However, the conversion rate of loyal customer is too around 9-10%.
3. We’ve tried Klar package to do categorical clustering but it took for over 8 hours which is nearly impossible to run. We had too unique categorical variables.

5. Conclusion

- **Implementing limitation**

    1. Models might require updates because customer behaviors change as time goes by.
    2. When encountering missing value, we have to understand what happened so that we can properly deal with it.

- **Potential model risk/bias**

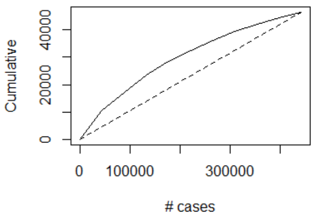
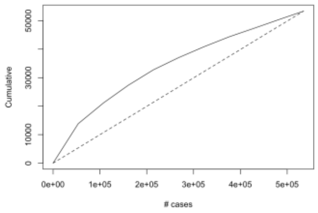
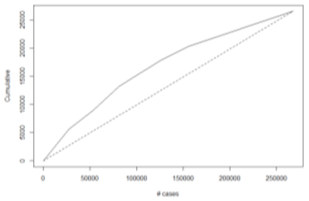
    3. The total error rate would be higher if undersampling.
    4. The dataset exists some missing values such as age and gender, and imputed values might lead to model bias.
    5. The data set we recommend excludes double11-related campaigns so there might be higher bias when predicting the loyal customers from Double 11-related campaigns.

- **Recommendation for business policy**

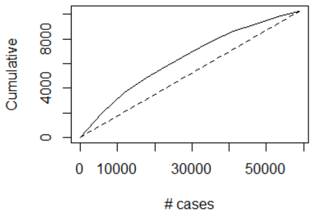
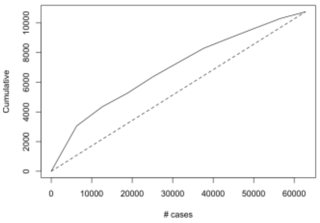
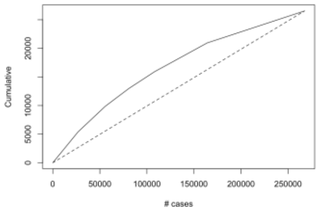
    6. The dataset should add more variables related to the characteristics of the sellers and collect more data on it.
    7. Increasing the customer’s willingness to add their product into favorite is a good idea for sellers to create their loyal customers.
    8. Lurkers are potentially loyal customers, which means that the sellers should not ignore these people and could try to put more efforts on these lurkers.

Appendix A. Lift Charts of Models with Different Data Sets

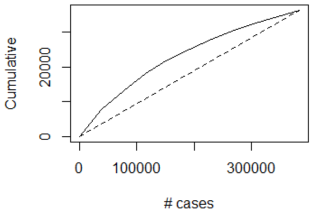
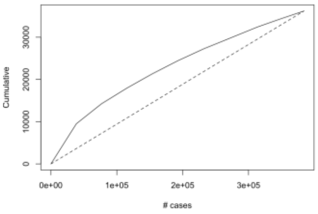
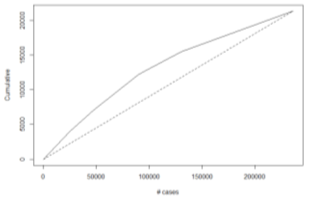
Data set = All

Model	Logistic	Random forest	Xbgoost
Lift chart			

Data set is from 5/11 to 10/25

Model	Logistic	Random forest	Xbgoost
Lift chart			

Data set is from 10/26 to 11/12

Model	Logistic	Random forest	Xbgoost
Lift chart	 Cumulative 0 20000 0 100000 300000 # cases	 Cumulative 0 10000 20000 30000 0e+00 1e+05 2e+05 3e+05 # cases	 Cumulative 0 5000 10000 15000 20000 0 50000 100000 150000 200000 # cases