IEOR 4650 Business Analytics

Professor Daniel Guetta
Group 9
Jiachen Liu, Jo Wu, Yuzhi Yao,
Shifan Zhang, Jeanie Zhao
December 11, 2018

Columbia University in the city of New York
116 St & Broadway
New York, NY 10027

Project Report: Repeat Buyers of Tmall.com

How was your Black Friday? Did you finally purchase that product you've been eyeing for at a discounted price? As the holiday season approaches, this is perhaps one of the most often asked questions. This is also one of the most important problems for merchants, as they need to decide which product to discount for, by how much, and how to direct their marketing efforts.

In China, similar to Black Friday, on November 11, also known as "Double 11" or "Single's Day", merchants participate in the year's largest online shopping festival by offering heavy discounts and cash coupons. Sales transaction volume double, or even triple on this date, capturing an expanded customer base rushing to place their orders to empty their shopping carts at discounted prices. Tmall.com, one of the biggest online B2C retail platforms in China, is undoubtedly the most popular destination for these buyers on Double 11. Operated by the Alibaba group, Tmall.com has been extremely successful in becoming the largest market player in terms of B2C retail platforms, with a market share of over 50% and annual gross merchandise volume (GMV) of $340 billion. Almost 10% of Tmall's annual GMV was generated on Double 11 in 2018.

However, although merchants invest tremendous amounts in advertising and marketing campaigns on Double 11, they worry that these will only attract one-time deal hunters and fail to secure repeated buyers who will contribute to regular business, thus leading to disproportionate future gain compared to the amount of discounts given on Double 11. This project, therefore, seeks to identify and categorize buyers in terms of their likelihood to become return buyers, allowing merchants to leverage such data to differentiate their marketing campaign strategies when targeting different groups to achieve optimal profit.

Problem Statement and Value Proposition

The project seeks to leverage historical 6-month anonymized customer shopping logs before and on Double 11 from Tmall.com to predict repeat customers in order to

1) Develop a more comprehensive understanding of customer traffic: Results can be utilized to categorize and group customer behaviors and identify statistically significant factors that contribute to or predict a repeat buyer
2) Reduce inefficient marketing spending through differentiation strategies: repeat customers contribute to a loyal customer base which is an easily-targeted audience and maintain traffic on a regular basis even in the absence of major promotions. Appropriately identifying this group will allow merchants to offer different tiers of promotion to capture the most benefits out of each group.
3) Increase Tmall.com's overall Return on Investment (ROI): Results can be used to position each merchant more accurately and drive up total revenue of the platform

$$\frac{Net\ Return}{Initial\ Investment}$$

Data

To target the problem statement, the team will use a data set on Ali Cloud, which 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 repeated buyers. Data fields include:

1. User_id: unique id for shopper

2. Age_range: age of shopper; divided into 8 groups, each group ranges ~5 years (0: Age Unknown; 1: <18; 2: 18-24; 3: 25-29; 4: 30-34; 5: 35-39; 6: 40-49; 7 & 8: >= 50)

3. Gender: gender of shopper (Female: 0; Male: 1; Unknown: 2)

4. Merchant_id: unique id for merchant

5. Action_type: used to identify which action serves as the most valid predictor of a potential repeat buyer (Click: 0; Add to shopping cart:1; Purchase: 2; Add to favorite: 3)

6. Label: used to benchmark the team's predictions to see if our models accurately identify repeat buyers as recorded by Tmall (Non-repeat buyer: 0; Repeat buyer: 1)

Graphical illustration of exploratory data visualization can be found in Appendix 1.

Initial Hypothesis: Model 1 Logistic Regression

In order to test the relationship between a shopper's demographic information such as age and gender (independent variable), and the likelihood of that shopper to become a repeat buyer (dependent variable), the team decides to conduct the first model using logistic regression, given our dependent variable is a binary response (0 to represent not a repeat buyer, and 1 vice versa). To perform the logistic regression, the team first factors each age and gender group as a categorical variable, with the base variables as age_group 0 (unknown), and gender 0 (female). The results are shown in Appendix 2.

The summary shows that on the age side, age_range4, age_range5, and the age_range6 are highly significant with coefficients of 0.132, 0.172, and 0.168 respectively, which means consumers who are in their 30s and 40s have more possibility of becoming repeat buyers than consumers in their 20s, since age group 2 has a negative coefficient and all these age groups are being compared to the same benchmark group of age_range 0. This is intuitive in practicality as well, as younger generations typically have less loyalty to specific online merchants because they have a broader browsing range and are more open to change.

On the gender side, gender 1 is highly significant with a negative coefficient of -0.01, which means when compared to gender 0 which is female buyers, male consumers have less

repeated buying actions. This again is consistent with the team's initial hypothesis, as female consumers are a more active buying force online and are more likely to become repeat buyers.

Model 2 Logistic Regression with Customer Behavior Data

Now that the team has performed a logistic regression to identify how a shoppers' age and gender contribute to their likelihood of becoming a repeat buyer, the team can add more variables such as shoppers' action type to decide if their actions are predictive of their becoming repeat buyers. Under the original data set activity_log, every action happened between a shopper and a merchant is recorded under a certain action type, where 0 stands for click, 1 stands for add-to-cart, 2 is for purchase and 3 is for add-to-favorite.

To process the data for an enhanced Model 2 based on customer behavior, the team needs to calculate how many times each action happened for every user under each merchant. Therefore, by using Python Dask Dataframe, the team counts the number of each action type under every user_id and merchant_id combination.

After merging dataset from the previous model with users' behavioral data, the model now has a more comprehensive set of independent variables. The team now uses gender (factor), age group (factor), and action_type as independent variables. By including the new variable action_type, the team now can examine the relationship between customers' current actions and their potential to become repeat buyers. Using the same reasoning behind Model 1, the team runs another enhanced logistic regression to examine the relationship between explanatory variables and probability of being a repeat buyer. Results can be found in Appendix 3.

The results are similar to Model 1 in the age and gender variables. After adding user behavioral data, our model's performance improved. The Area Under the Curve (AUC) of

Model 2 is 0.6066, compared to 0.53 of Model 1. AUC is a common method to assess accuracy of logistic models, as it indicates the probability of a True Positive (sending coupon to a repeat buyer) is ranked more highly than False Positive (sending coupon to a one-time deal hunter). The higher the AUC, the better the performance of the model.

Moreover, the summary from Model 2 indicates that customers who click on the product and bookmark it as their favorite are more likely to purchase it later and become a repeat buyer, and customers who add the item to their shopping carts are more likely to be a one-time deal hunter. This ties back to real-world intuition as well. Customers who bookmark a product as their favorite tend to be very interested in the product even without major promotions. Adding to favorite also means the customers are browsing the merchant regularly, are typically loyal customers of the brand, and know that they will return later. On the other hand, items added to customers' shopping cart tend to sink into the cart along with many other items or may have just been added to the shopping cart by one-time deal hunters on Double 11.

Business Implications & Next Steps

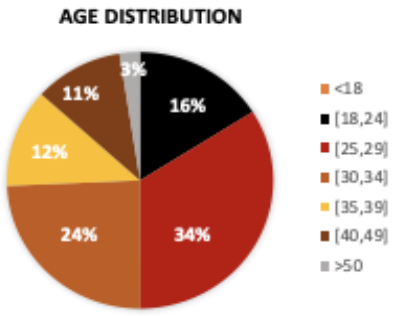
An effective model is not just statistically significant, but should also generate practical economic benefit when placed in a real-world context. Therefore, the final stage of this project is for the team to evaluate the actual savings our models can generate for Tmall and its merchants. Tmall's current methodology is to not categorize repeat buyers and send coupon discounts to all active shoppers. The team will use this as the benchmark cost to compare with the potential savings generated by our models. By conducting market research, the team has found out that the total sales of Tmall on Double 11 this year is $31B, and the number of transactions is 1.48B[1]. Moreover, marketing/cash coupon cost is approximately 15% - 20% of

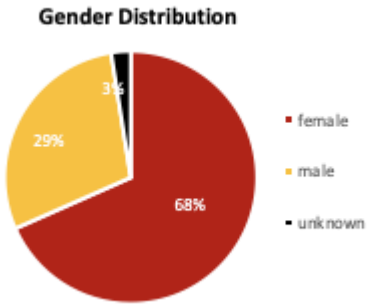
[1] Statistics from Alibaba Group

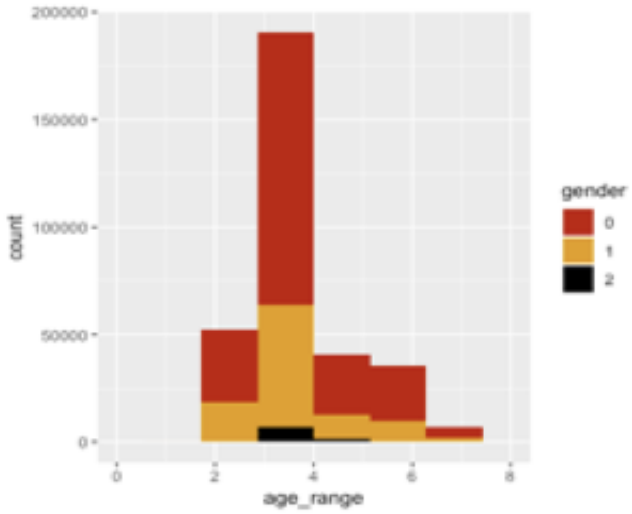
total revenue, and the value of a repeat buyer is approximately 40% of total revenue[2]. Therefore, the coupon cost per transaction is roughly $2.1, and the value for each repeat buyer is $8.4. Net profit can thus be calculated as revenue from repeat buyers minus the cost from coupons sent to buyers. For the benchmark cost, where Tmall is not categorizing repeat buyers and sending coupons to every customer, the net profit is -$406,954. The team then applies the same profit matrix to Model 2 to determine the best p-value by maximizing amount saved compared with benchmark. The confusion and cost matrices can be found in Appendix 4. By implementing our team's model, Tmall can effectively categorize customers based on their age, gender, and actions to identify repeat buyers, and enable merchants on Tmall.com to differentiate discounting strategies to capture the maximum benefit out of each consumer group. This will achieve the team's original value proposition of reduced marketing expenses, increased ROI, and a more established understanding of consumer traffic, leading to a maximum savings of $278,094.

[2] *The ROI from Marketing to Existing Online Customers*, Adobe Digital Index Report, 2012

Appendix 1: Exploratory Data Analysis







Appendix 2: Model 1 Logistic Regression Results

```
> summary(model1)

Call:
glm(formula = label_factor ~ age_range_factor + gender_factor,
    family = "binomial", data = user_final_tr)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.3955  -0.3820  -0.3558  -0.3292   2.4978

Coefficients:
                  Estimate Std. Error  z value Pr(>|z|)
(Intercept)       -2.71292    0.02146 -126.421  < 2e-16 ***
age_range_factor1 -7.83113   42.22708   -0.185    0.853
age_range_factor2 -0.17547    0.03680   -4.768 1.86e-06 ***
age_range_factor3 -0.01500    0.02808   -0.534    0.593
age_range_factor4  0.13196    0.02933    4.499 6.81e-06 ***
age_range_factor5  0.17182    0.03529    4.869 1.12e-06 ***
age_range_factor6  0.16774    0.03731    4.496 6.94e-06 ***
age_range_factor7  0.05096    0.07760    0.657    0.511
age_range_factor8  0.15310    0.17393    0.880    0.379
gender_factor1    -0.18598    0.02194   -8.477  < 2e-16 ***
gender_factor2     0.03182    0.05620    0.566    0.571
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 89218  on 192855  degrees of freedom
Residual deviance: 89011  on 192845  degrees of freedom
AIC: 89033

Number of Fisher Scoring iterations: 9
```

Appendix 3: Model 2 Logistic Regression Results

```
> summary(model2)

Call:
glm(formula = label_factor ~ age_range_factor + gender_factor +
    click + cart + purchase + favourite, family = "binomial",
    data = user_final_tr)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.6658  -0.3571  -0.3342  -0.3203   2.6070

Coefficients:
                    Estimate Std. Error  z value Pr(>|z|)
(Intercept)       -3.1133644  0.0246853 -126.122  < 2e-16 ***
age_range_factor1 -7.8609003 42.0732432   -0.187 0.851787
age_range_factor2 -0.1783821  0.0370490   -4.815 1.47e-06 ***
age_range_factor3 -0.0469337  0.0282748   -1.660 0.096932 .
age_range_factor4  0.0894640  0.0295477    3.028 0.002464 **
age_range_factor5  0.1285385  0.0355501    3.616 0.000300 ***
age_range_factor6  0.1363110  0.0375908    3.626 0.000288 ***
age_range_factor7  0.0227267  0.0781620    0.291 0.771231
age_range_factor8  0.1318337  0.1746073    0.755 0.450231
gender_factor1    -0.1379408  0.0221131   -6.238 4.43e-10 ***
gender_factor2     0.0246422  0.0565733    0.436 0.663141
click              0.0068957  0.0004172   16.527  < 2e-16 ***
cart              -0.1120931  0.0503655   -2.226 0.026041 *
purchase           0.2219706  0.0080860   27.451  < 2e-16 ***
favourite          0.0365387  0.0049033    7.452 9.20e-14 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 89218  on 192855  degrees of freedom
Residual deviance: 87500  on 192841  degrees of freedom
AIC: 87530

Number of Fisher Scoring iterations: 9
```

Appendix 4: Confusion and Profit Matrices

	Predicted		Total
Actual	Repeat	Non-repeat	
Repeat	1,659	14,179	15,838
Non-repeat	9,624	231,679	241,303
Total	11,283	245,858	257,141

Predicted

Actual	Repeat	Non-repeat
Repeat	6.3	-8.4
Non-repeat	-2.1	0

```
> max(profit)
[1] -128862.3
> benchmark = - cost * nrow(user_mer_final) + value_repeat_buyer * sum(user_mer_final$label == 1)
> benchmark
[1] -406956.9
> amount_saved = max(profit) - benchmark
> amount_saved
[1] 278094.6
```