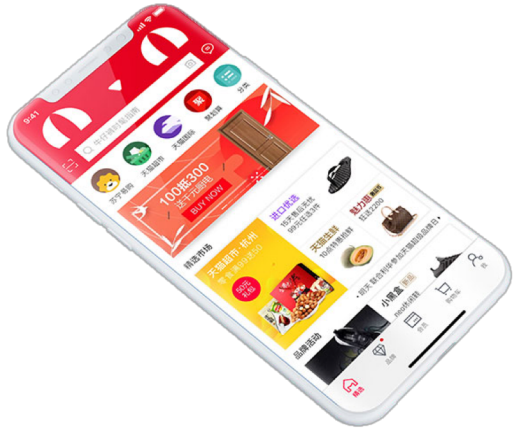
Tmall.com Repeat Buyers

Business Analytics IEOR4650
Fall 2018

Jiachen Liu	jl4981
Jeanie Zhao	jz2894
Jo Wu	jw3567
Shifan Zhang	sz2712
Yuzhi Yao	yy2933

Agenda

- Background Information
- Data Visualization
- Initial Hypothesis: Model 1
- With Customer Behavior Data: Model 2
- Business Implications
- Q&A
- Appendix



Background

- Tmall.com, operated by Alibaba Group, is one of China's largest B2C online retail platforms
- > 50% market share
- > 500 million users
- $340b annual GMV
- 10-20% marketing
- Double 11
  10% annual GMV

天猫 TMALL.COM

TMALL SINGLE'S DAY 2018

KEY HIGHLIGHTS

¥ 214 billion ($31 billion) Gross Merchandise Volume	**27 %** YoY growth
> 1 billion Parcels shipped	**19 k** Oversea brands

Source: Alibaba

Problem Statement

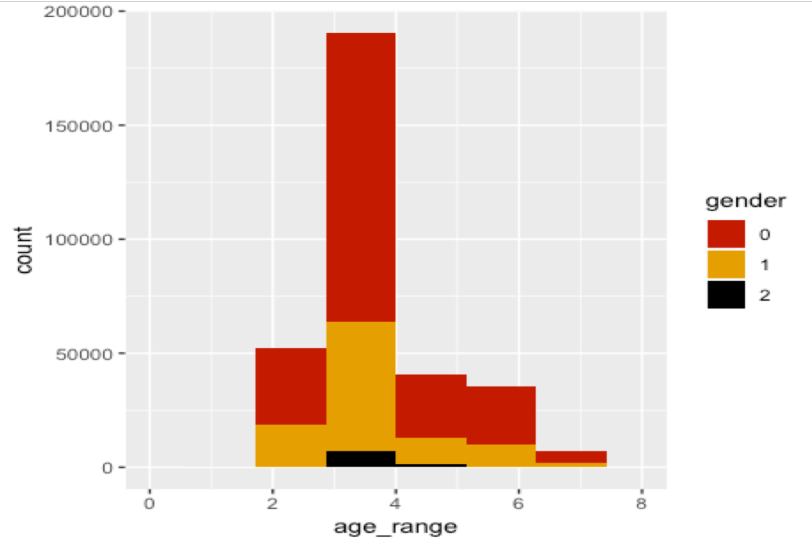
How to leverage past anonymized users' shopping log data to identify potential repeat buyers

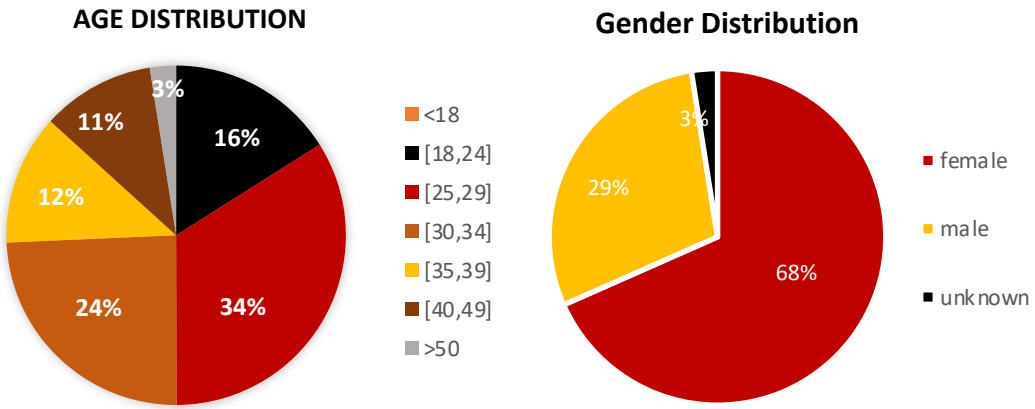
Value Proposition

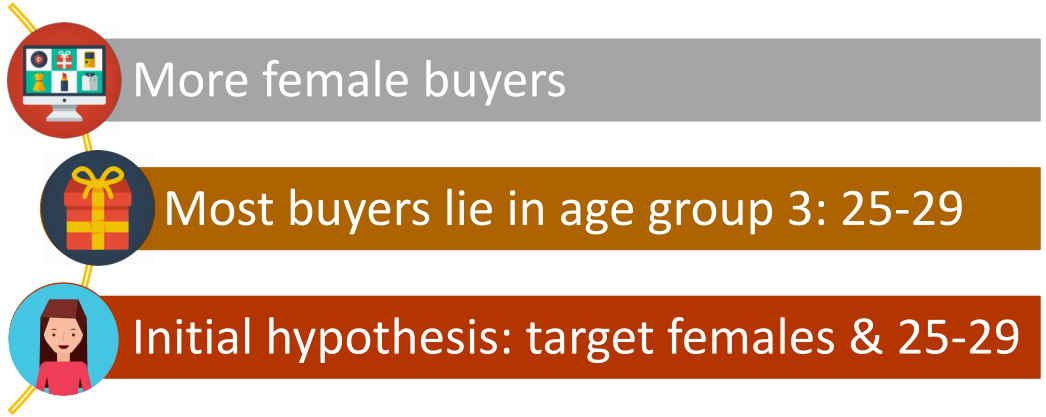
Increased comprehensive understanding of customer traffic
Reduced inefficient marketing spending
Enhanced Return on Investment (Return/Investment)

Dataset Visualization

- 6 month historical anonymized users' shopping log data before and on 11/11
- Different age groups & gender
- Unique user/merchant id
- Action: click, cart, purchase, favorite
- Label: Repeat buyers (6.12%)



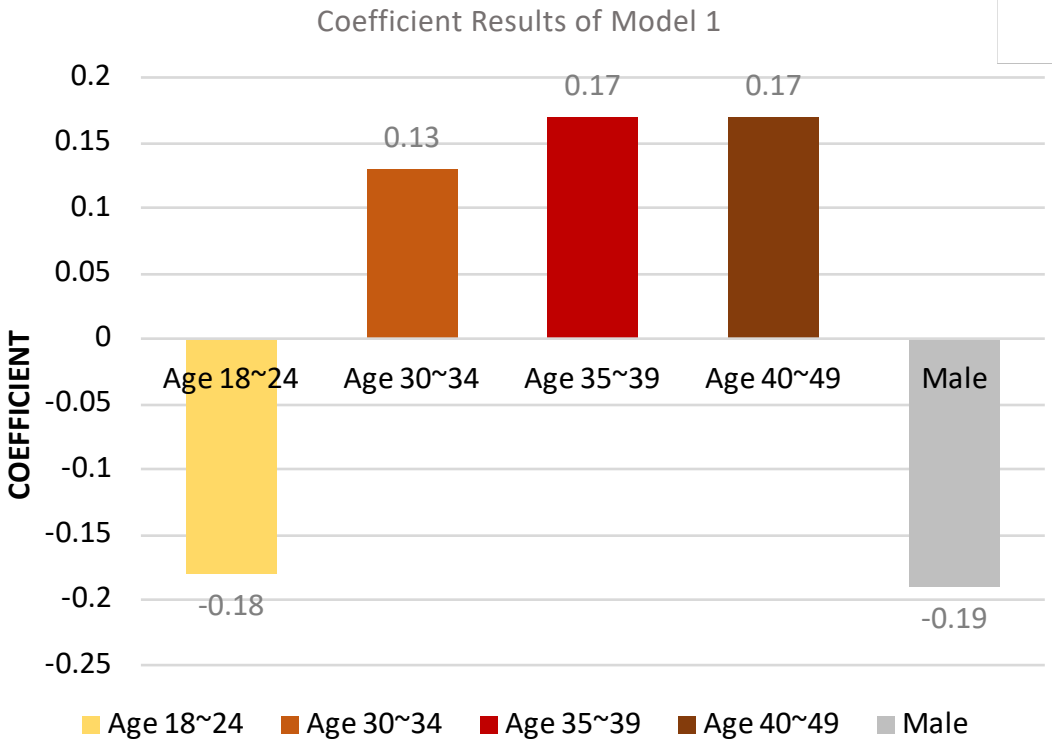




Initial Hypothesis: Model 1

- Logistic Regression
- Label ~ age + gender
- Results:

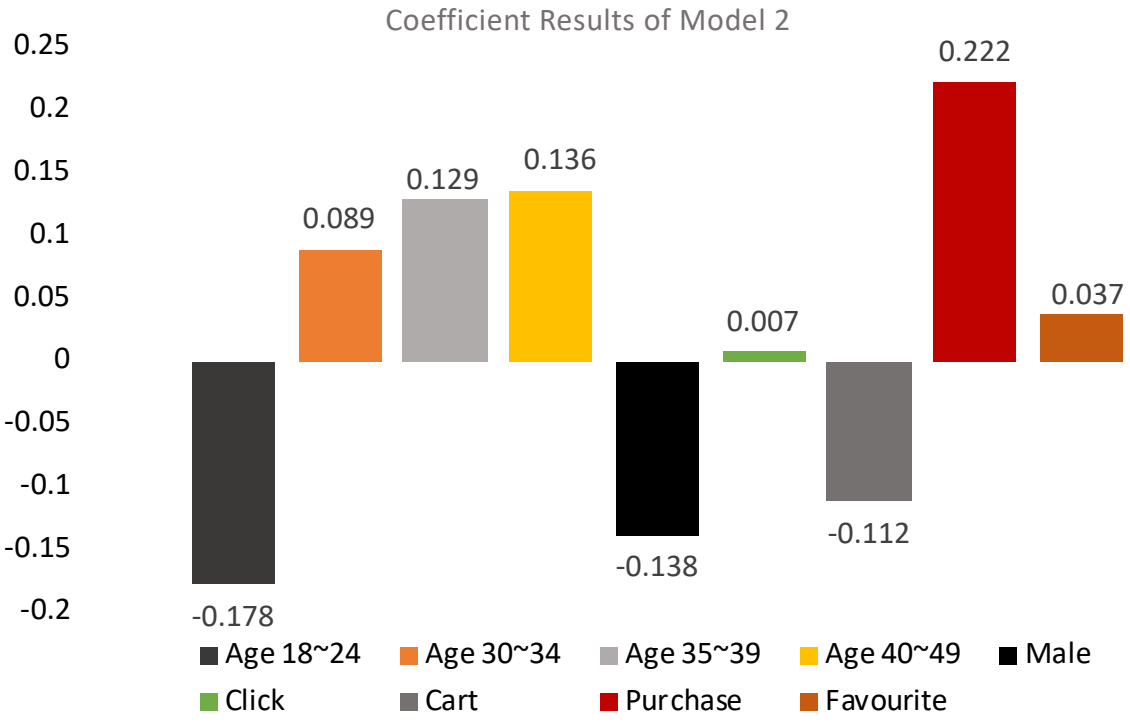
MORE LIKELY	LESS LIKELY
• Age 30~50 • Female	• Age 18~24 • Male



- Conclusions
  - Female have more frequent buying actions online
  - Younger generations have a boarder browsing range and are more open to change; thus less likely to become loyal repeat buyers

With Customer Behavior: Model 2

- Action type: Click, Add-to-Shopping Cart, Purchase, Add-to-Favorite
- Data Processing: Action type counts per user per merchant



Repeat Buyers

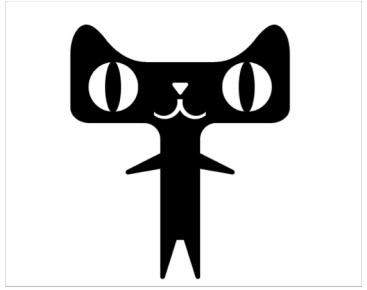
Add to Favorite
 # Clicks

One-time Deal Hunters

Add to Shopping Cart

	AUC
Model1	0.53
Model2	0.6066

Business Implications



- How much savings can this model generate in practicality?

  Amount saved = Net Profit Using the Model – Net Profit from Benchmark

  = **$378,924**

Confusion Matrix

	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

Profit Matrix

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

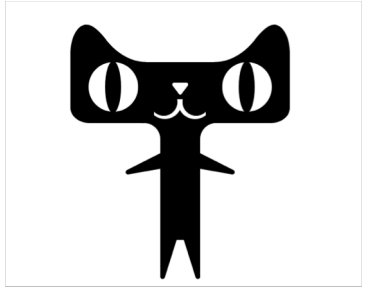
Conclusion

By implementing Model 2, merchants can save approximately $0.8 per transactions on Double 11. In total, merchants on Tmall can save $1.2B on Double 11, approximately 3.9% of the total sales.

Assumptions: coupon cost is 10% of the revenue, value of repeat buyer is 40% of revenue, total # of 2018 Double 11 transactions: 1.4B

Source: Medium Article; Adobe Study

Next Steps



- Continuously monitor and update model assumptions with actual merchant data to enhance prediction

- Conduct pilot program with Tmall.com to test savings efficiencies on sample merchants

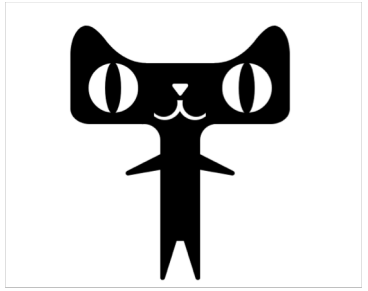
Q&A

Appendix 1: Model 1 & 2 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
```

```
> 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 2: Confusion & Cost Matrix



truth	predict FALSE	TRUE
0	231679	9624
1	14179	1659

```
# estimate coupon cost & value_repeat_buyers
# double 11 sales = 31,000,000,000 USD
# double 11 # transactions = 1,480,000,000
# avg_paid = 21
# coupon cost = 0.1 * 21 = 2.1
# value_repeat_buyers = 0.4 * 21 = 8.4
cost = 2.1
value_repeat_buyer =  8.4
profit  = findprofit(user_mer_final,cost,value_repeat_buyer)
```

```
> 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
```