Team 9: Yuchi Shih, Wendy Huang, Zoe Cheng, Jessy Yang

Predicting Loyal Customers
to Increase Merchants' Income

天猫 TMALL.COM

Business Goal

Stakeholder Sellers on TMall

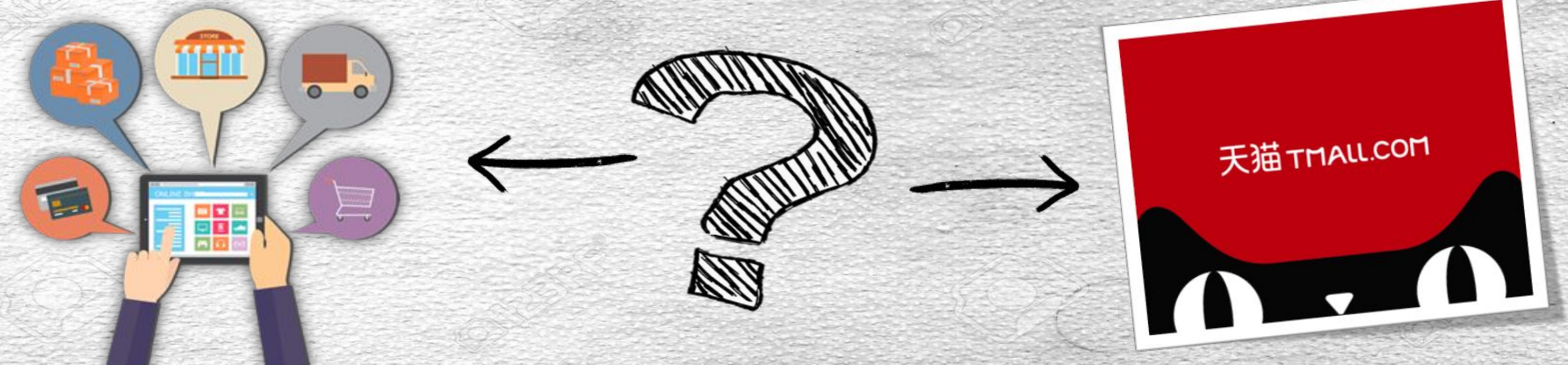
Problem One-time buyers attracted by coupons and discounts (Low ROI)

Goal Reduce the promotion cost & increase the rate of the repeated customers

Data Mining Goal

Predicting whether this client will be the loyal customer to this seller.

Outcome variable: binary outcome(0,1)



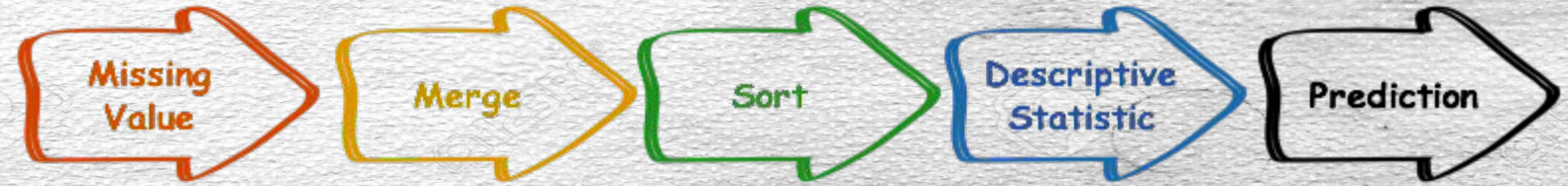
Data

- Source: Tmall.com (A huge E-commerce company in China)
- Data: 1470911 obs. based on actual user activity on the platform
- Population: customers who bought on 1111
- Loyal customer: customers bought again in the next 6 month

Preprocessing

Profile	Behavior	label
user_id age_range gender	user_id item_id cat_id merchant_id brand_id time_tamp action_type	user_id merchant_id label

Categorical variables	Output
user_id item_id cat_id merchant_id brand_id time_tamp action_type gender age_range	label



Data(Data cleaning)

	user_id	item_id	cat_id	seller_id	brand_id	time_stamp	action_type
1	328862	323294	833	2882	2661	829	0
2	328862	844400	1271	2882	2661	829	0
3	328862	575153	1271	2882	2661	829	0
4	328862	996875	1271	2882	2661	829	0
5	328862	1086186	1271	1253	1049	829	0
6	328862	623866	1271	2882	2661	829	0

	user_id	age_range	gender
1	376517	6	1
2	234512	5	0
3	344532	5	0
4	186135	5	0
5	30230	5	0
6	272389	6	1
7	281071	4	0
8	139859	7	0
9	198411	5	1
10	67037	4	1
11	149002	5	2
12	7468	4	0

	user_id	merchant_id	label
1	34176	3906	0
2	34176	121	0
3	34176	4356	1
4	34176	2217	0
5	230784	4818	0
6	362112	2618	0
7	34944	2051	0
8	231552	3828	1
9	231552	2124	0
10	232320	1168	0
11	232320	4270	0
12	167040	671	0

0	Click
1	Add-to-chart
2	Purchase
3	Add-to-favorite

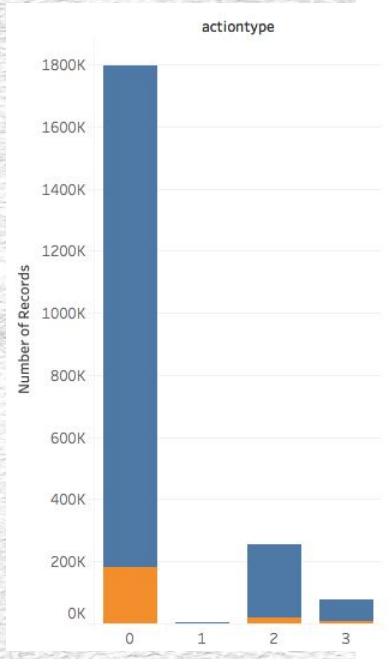
Data

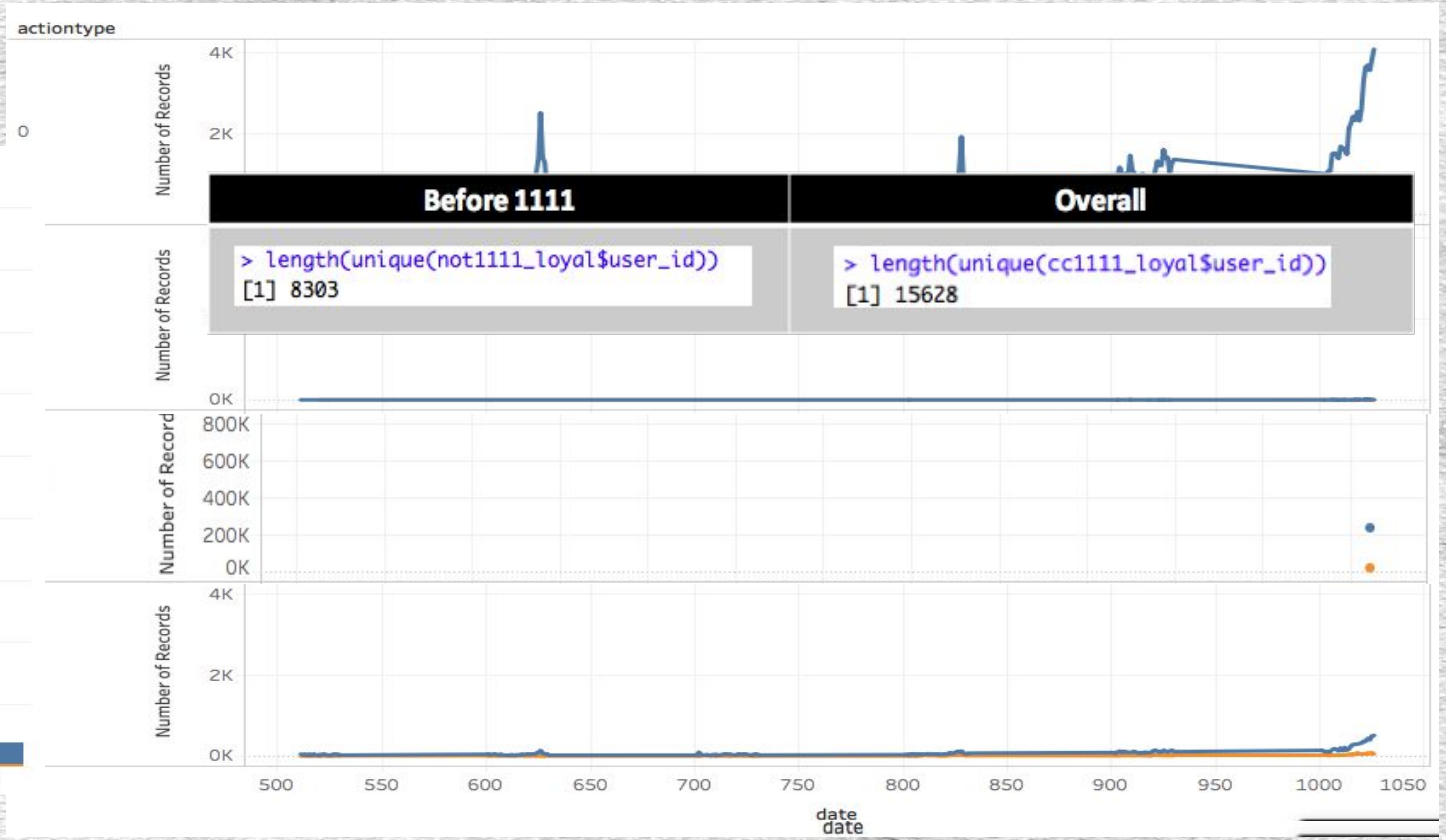
- Source: Tmall.com (Top B2C E-commerce platform in China)
- Data: 1476911 obs. (those customers all bought on 1111)
- Definition of Loyalty: Customers bought again at the same seller in the next 6 month after 1111
- Preprocessing:
    1. **Missing value**: replace missing value by MICE (tree)
    2. **Derived variable**:
        a. separate action_type variable into action_0, action_1….
        b. sum up action_0, action_1…. according to user, seller ID, time_stamp
    3. **Variable selection**: age, gender, action_0~3

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

Data visualization

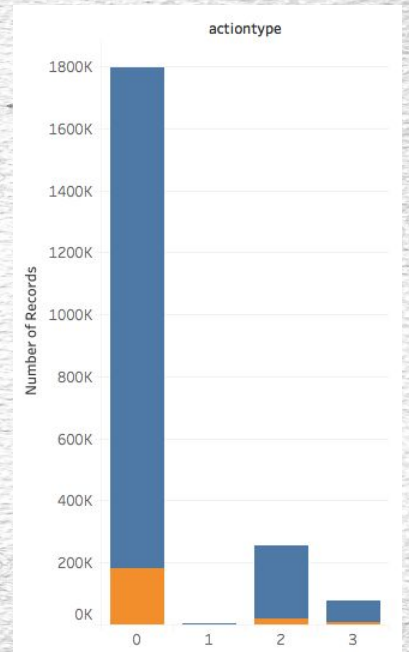
Click

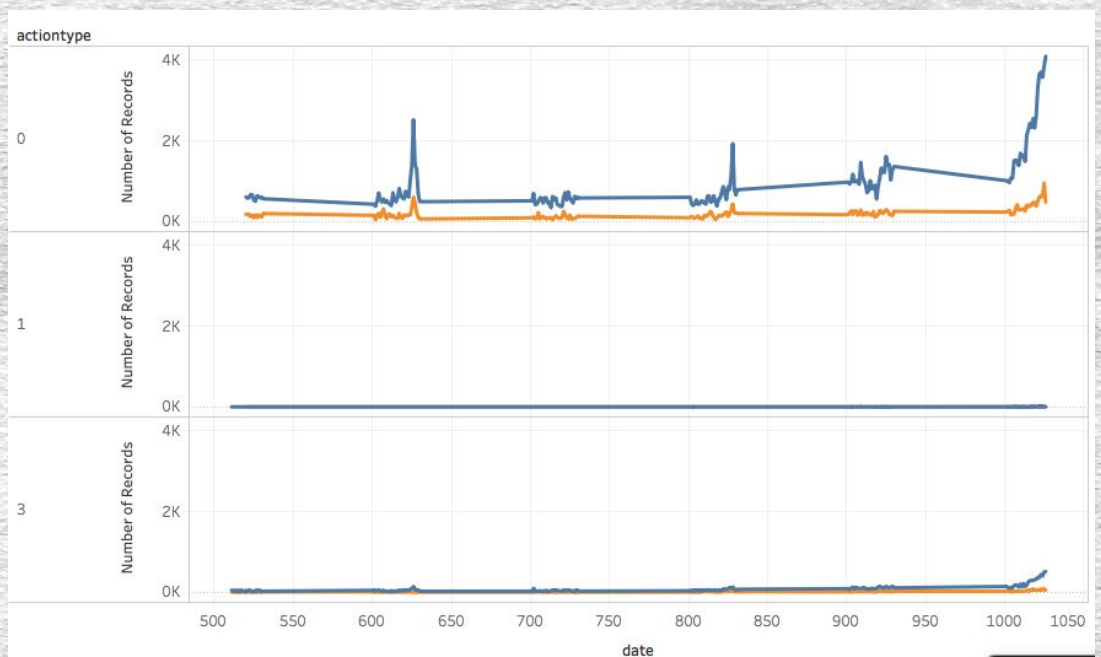




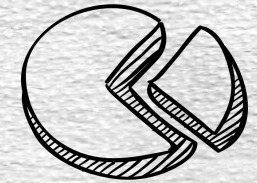
Visualization (before November)

Before 1111	Overall
`> length(unique(not1111_loyal$user_id))` `[1] 8303`	`> length(unique(cc1111_loyal$user_id))` `[1] 15628`

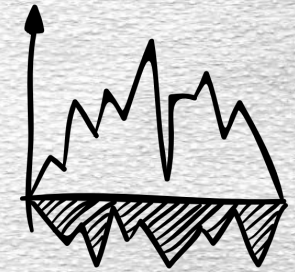




Methods & Performance Evaluation



- **Data mining methods:** Regression, Random forest, XGBoost
  a. K-NN and Hierarchical clustering are not suitable in this case because of the huge dataset.
- **Performance measure:** lift chart
- **How to map our business goal?**
  a. Use the K-means clustering method to cluster the training dataset.
  b. Build different model for each group by decision tree and regression.
  c. Predict the outcome and evaluate the performance by confusion matrix.
  d. The results allow us to predict the potential repeated buyers in each group.



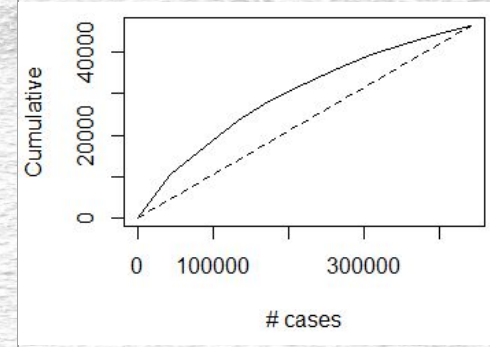
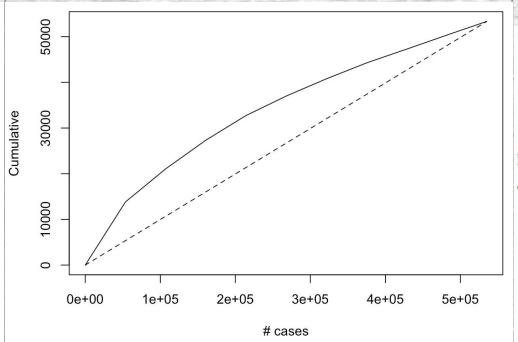
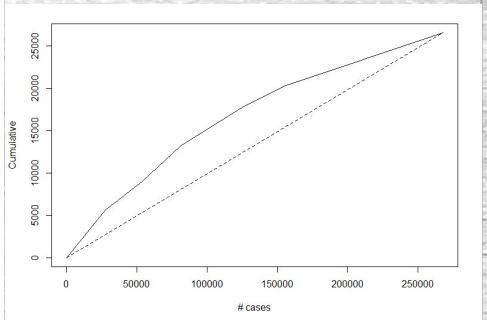
Data

ALL	Original		Undersampling	
	# Records	loyal (%)	# Records	loyal (%)
Training (50%)	738,455	10.47%	738,455	50%
Validation (30%)	443,073	10.46%	443,073	10.46%
Testing (20%)	295,383	10.53%	295,383	10.53%

Result (data=all)

Benchmark: Naive Rule

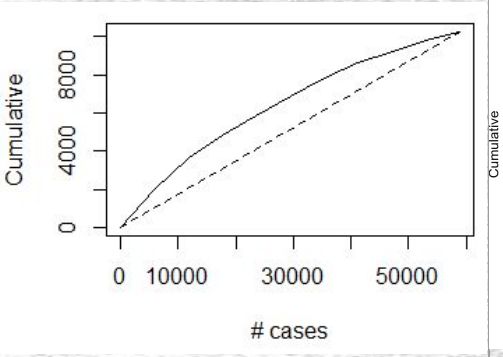
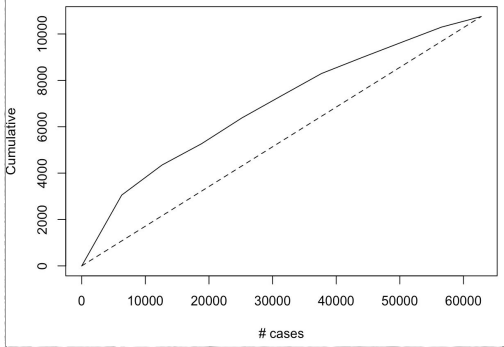
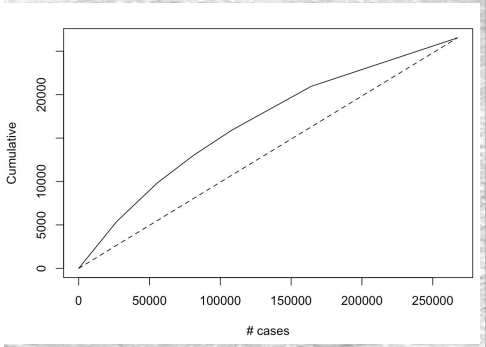
	Accuracy of top 10%
Best Model:	Random Forest
Naive Rule	10 %

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

Result(data=0511-1025)

Benchmark: Naive Rule

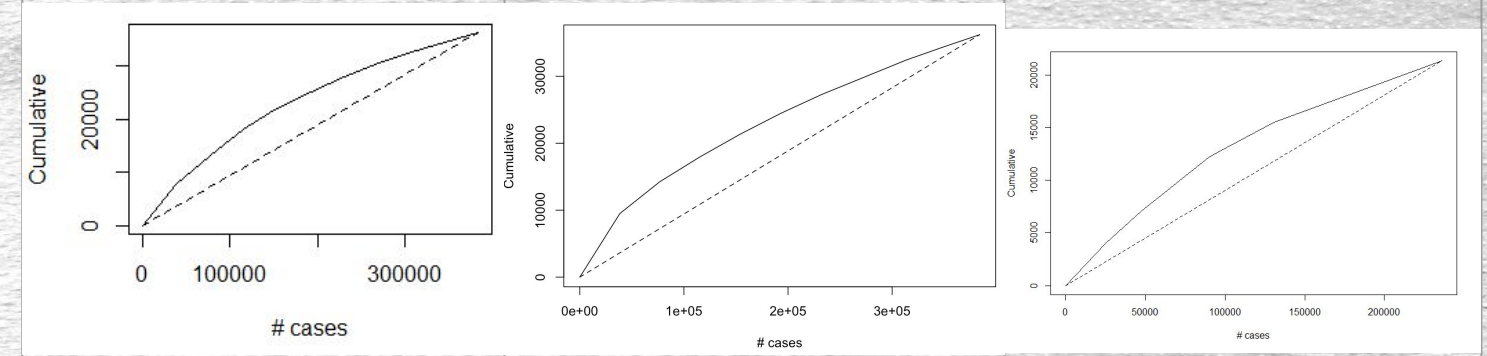
	Accuracy of top 10%
Best Model	Random Forest
Naive Rule	10 %

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

Result(data=1026-1111)

Benchmark: Naive Rule

	Accuracy of top 10%
Best Model	Random Forest
Naive Rule	10 %

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

Recommendations

For Implementation	For Model Risk	For Business Policy
1. Models might require updates 2. When encountering missing value, we have to understand what happened so that we can properly deal with it.	1. Low accuracy of top 10% 2. The total error rate would be higher if undersampling. 3. The dataset exists some missing value such age and gender. 4. Since the samples increase enormously and rapidly in the last two weeks, the bias might affect the outcome.	1. Collect more data about characteristics of the sellers. 2. Increase the customer's willingness to add their product into favorite. 3. Lurkers are potentially loyal customers, which the sellers should not ignore these people.

Implementation/Production Considerations

Time difference: This dataset was collected from 2 years ago.

A certain period of data:

- The dataset only contains half year of the business data. Therefore the model might require updates as time goes, in case features changes.
- For instance, if there is a new brand not included in this dataset, it might not be able to predict the results.

Lack of information: Not included merchants' or discounts' data.