Find the most likely repurchasing customers next month:
For targeted promotions

BADM Group 6:
Rona Lu-Lai(108078504), Vivian Lu(108078512),
Joanne Hsueh(108078508), David Hung(107077503)

Executive Summary

Business Problem

Our business goal is based on the current marketing strategies of Me-come. Most of the current marketing strategies of Me-come were not targeted at specific groups of customers. For example, Me-come sends all customers text messages or send flyers to nearby residents, which is costly and inefficient.

Data

Our analytical objective is to sort VIPs by their potential to repurchase next month and to identify the top 10% of VIPs who most likely to return. VIPs with at least one transaction record in the next month will be classified as returning customers(1), the others will be classified as not returning customers(0). We used **VIP, Transaction and Product data** from Me-come and selected 77820 VIP customers as our data.

Analytics Solution

We found that **total sales points, purchase times Dec-May, pur times morning-evening(all purchase times), min/mean/max sales amount, category 1/5/7** are important factors. These factors are related to the purchase frequency and sales amounts of each VIP. It implies that VIPs who purchase more times or amounts will have a higher probability to return next month.

Best Model on test data

The best model is a classification method called Random Forest with important factors mentioned above. By using our model, Me-come can catch 1.83 times of repurchasing VIPs than randomly selecting by only sending promotions to 10% of VIPs.

Random Forest Lift Chart on test data	**Random Forest Gain Chart on test data**

Recommendations

Implementation

We shall forward the result to our marketing team after our monthly expected results so that they can establish marketing strategies. The marketing team will provide the sales team with tactics for the implementation and will deliver promotions to potential customers.

Production issues

Since we use historical data, we will not run our solution in real-time. We need to collect and update new purchase information on a monthly basis.

Detailed Report

Business Goal

To improve the marketing strategy of Me-come and reduce the cost, we plan to select the top 10% customers who will most likely return in the next month and promote them with marketing strategies. This idea could lead to increased loyalty and retention of customers and boost the rate of consumption. Our model would help Me-come send promotions to the top 10 % VIP who have high chance to return next month. If the selected VIP came back with the promotion, we would consider it as a success.

Data Mining Goal

Our analytical objective is to sort VIPs by their potential to repurchase next month and to identify the top 10% of VIPs who most likely to return. VIPs with at least one transaction record in the next month will be classified as returning customers(1), the others will be classified as not returning customers(0).
Our solution is a supervised, predictive task and a forward-looking prediction.

Data Description

We used **VIP, Transaction and Product tables** from Me-come's database. We selected VIPs who bought products from Me-come in the past 7 months(we filtered the 'LBUY_DATE' from 2018/12-2019/6) and predicted the probability of VIP returning next month. We derived the outcome from the last buying date(return: 1; not return: 0). Our data size is 77820 Rows*39 Columns, and each row means a VIP customer.
Belows are four parts of out input & output variables:

- **VIP profile:** Basic profile of VIP

vip_no	sex	age_binned	age_binary	birth_mon	tot_spont	most_freq_store	mean_sale_amt	min_sale_amt	max_sale_amt
0006004464	0	2	1	6	39	1006	257	20	643
005024655	0	3	1	5	17	1005	150	150	150
010010939	1	*NA*	0	*NA*	107	1010	50	50	50
011010306	1	4	1	9	465	1011	5638	5638	5638
016002621	0	3	1	4	474	1016	215	72	529

- **Promotion Types:** Different promotion type of the products that VIP bought

promo_type_1	promo_type_2	promo_type_3	promo_type_6	promo_type_7	promo_type_A	promo_type_B	promo_type_V	promo_type_Z
7	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0
0	0	1	0	0	0	0	4	6
4	0	0	0	0	0	0	3	0

- **Products categories:** Different categories of the products that VIP bought

catg_1	catg_2	catg_3	catg_4	catg_5	catg_6	catg_7	catg_8	catg_9	catg_10
5	0	0	0	11	4	0	0	0	0
1	0	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0	0
5	1	2	0	0	5	1	1	0	0
5	0	0	0	9	0	1	0	0	0

- **Purchase Time:** VIP shopping times
- **Outcome:** repurchase_or_not: No (0) / Yes (1)

pur_times_morning	pur_times_afternoon	pur_times_evening	pur_times_Dec	pur_times_Jan	pur_times_Feb	pur_times_Mar	pur_times_Apr	pur_times_May	repurchase_or_not
1	5	6	0	5	1	4	2	0	0
0	0	1	1	0	0	0	0	0	0
0	2	0	1	0	0	0	0	1	0
0	1	0	0	1	0	0	0	0	0
1	9	0	3	3	1	1	1	1	0

Data Preparation

1. Dealing with outliers & unusual values:

- age (with negative values): Set possible VIP age range from 12 to 100
- sale_amount (with negative values): Set negative sales amout to 0

2. Dealing with missing Values

- sex (2 rows missing): Delete two rows
- age (31586 rows missing): Change into categorical variable by data binning (missing & unusual values/12-18/18-30/30-65/above 65)
- birth_month (31586 rows missing): Change into categorical variable by data binning (missing / 1-12 months)

3. Data Visualization: (Appendix 1)

4. Data partition: We use 60% in Training, 30% in validation, and 10% in testing.

Method

We built 8 different models (Appendix 2: Important steps for building models) :

1. K-nearest-neighbors
2. Logistic regression
3. Lasso regression
4. Ridge regression
5. Stepwise regression
6. Classification tree
7. Random Forests
8. Ensemble

Variable Importance Plot using Random Forest

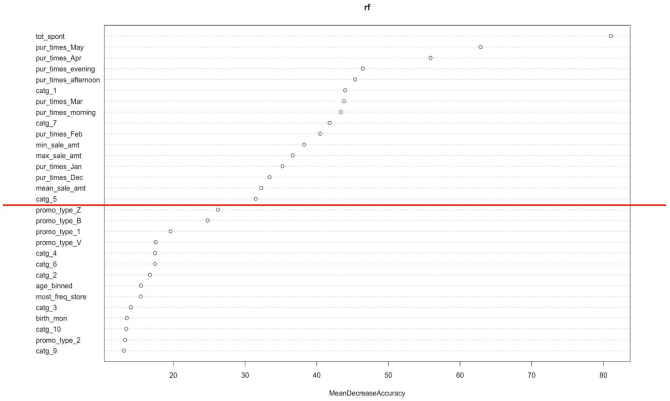


Figure 1. Variable Importance Plot using Random Forest

We ran random forest and ranked the most important variables to find the relationship between predictors and the outcome variable. In the end, we selected some important variables: **total sales points, purchase times Dec-May, pur times morning-evening(all purchase times), min/mean/max sales amount, category 1/5/7.**

We found that those important variables are related to the purchase frequency and sales amounts of each VIP. It implies that VIPs who purchase more times or amounts will have a higher probability to return next month.

Performance Evaluation

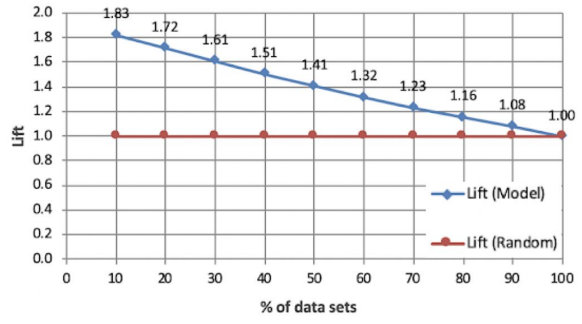
We used **Lift and Gain charts** to compare the performance of models. We computed the percentage of 1 in the first decile of each model and divided them by the naive benchmark which in our case is the propensity of 0.514(24040/46758= 0.514, in the training set we have 24040 1s and 22718 0s). At last choose the model with the highest lift in the first decile (Highest possible lift: 1/0.514 = 1.945). (See charts from appendix 3)

	Logistic	Lasso	Ridge	KNN	Stepwise	Classification Tree	Random Forest
All Variables	1.82	1.82	1.80	1.18	1.72	1.49	1.82
Selected Variables	1.81	1.81	1.80	0.86	1.70	1.49	1.82

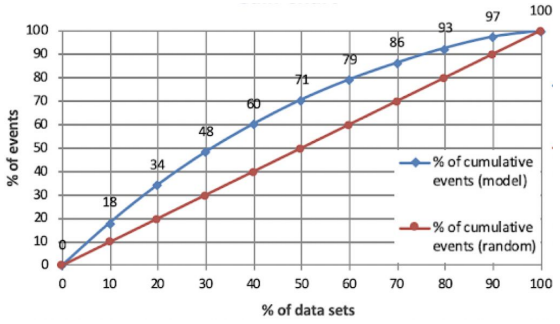
Table 1: Top decile lift of each model (All variables and Selected variables)

Ensembles (Combine results from all models with Top Decile > 1.8)

The first decile lift of the ensembles is 1.83, which is a bit better than the Random Forest with selected variables. However, the model training and computational cost are much higher. Hence, we still use Random Forest with selected variables as our best fit model.



Ensembles Lift Chart



Ensembles Gain Chart

Best Model(Random Forest with selected variables) on test data

In the end, we select Random Forest with selected variables as our best model. The first decile is **1.83,** which performed well and without overfitting.

Random Forest Lift Chart on test data

Random Forest Gain Chart on test data

Recommendations

Implementation

We shall forward the result to our marketing team monthly so they can establish marketing strategies. The marketing team will provide the sales team with tactics for the implementation and will deliver promotions to potential customers.

Production issues

Since we use historical data, we will not run our solution in real-time. We need to collect and update new purchase information on a monthly basis.

Appendix 1: Data Visualization

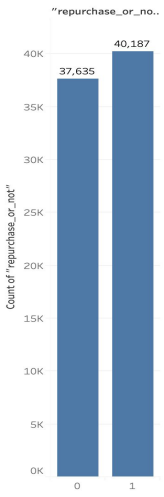


Figure 1

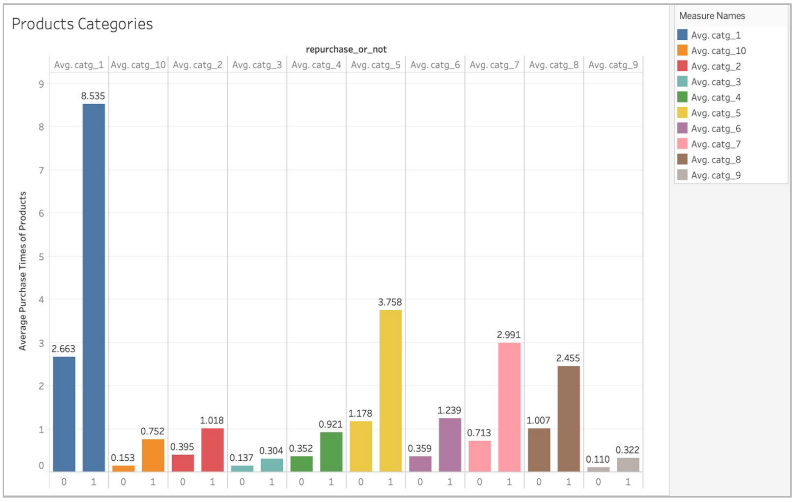


Figure 2

Figure 1: outcome (0 as not repurchasing,1 as repurchasing; a balanced outcome)
Figure 2: Average purchase times of each product categories over past six months

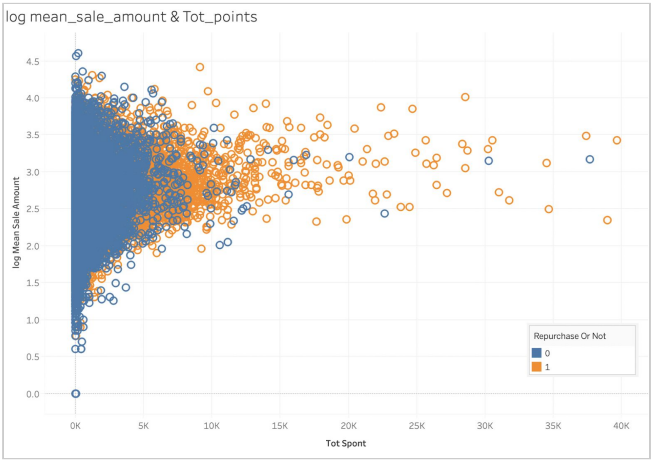


Figure 3

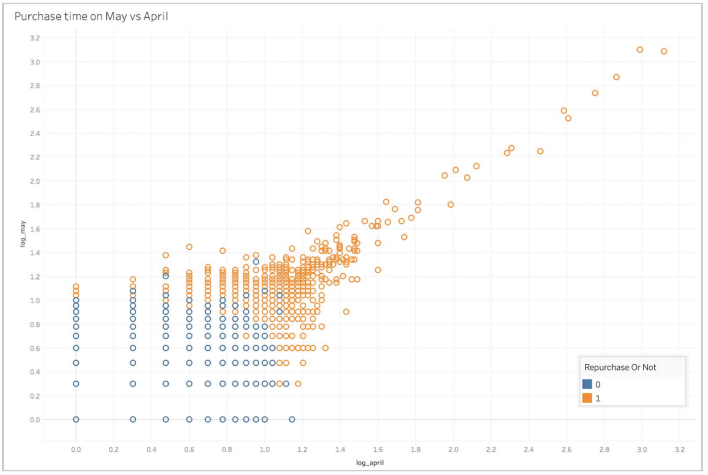


Figure 4

Figure 3: Mean_sale_amount & total points (log scale on y axis)
Figure 4: purchase_time_Apr & urchase_times_May (log scale on x & y axis)