**Kellogg School of Management**
Northwestern University


# Tuango:
# RFM Analysis for Mobile App Push Messaging

Tuango is one of the major "deal-of-the-day" websites in China. The website's business model is similar to that of Groupon, promoting discounted gift certificates that can be used at local or national retailers. The pronunciation of "Tuango" in Chinese sounds similar to "group buying," which refers to the fact that customers are buying as a big group for each "deal."

Susan Liu had been working as a data analyst in Tuango's Internet Marketing group for three years and had recently been appointed as Chief Data Scientist to the newly founded Mobile Marketing group. As Chief Data Scientist, Liu managed a small but highly competent analytics team. The Mobile Marketing group managed Tuango's marketing campaigns on mobile apps installed on customers' Android and iOS devices. A conservative estimate by Tuango in 2018 was that the company had about 14 million active mobile customers. More importantly, the smartphone user base in China surpassed 700 million in 2018, which meant that there was great potential for growth in Tuango's business.

Tuango had been experimenting with promotional push message campaigns through mobile apps for several months. These campaigns followed a common pattern. It always started with a deal that the company wanted to offer. Tuango then selected customers who had expressed an interest in the product category the deal was for, either when they signed up with Tuango, or because they had already purchased a deal in the same category during the last 12 months. Finally, if the deal was tied to a physical store, Tuango made sure to only target customers that lived sufficiently close to the promoted store. Once customers were identified, the offer was pushed out using Tuango's app on customers' mobile devices.

When Liu became Chief Data Scientist, she decided to reevaluate how mobile campaigns were executed. In particular, she was bothered by the view in the company that the cost of pushing deals onto customers' phones was essentially zero. However, Liu knew that the true marginal cost of each message was much higher. If customers received too many deal offers that were not relevant to them, customers could block future messages in the app, thereby preventing Tuango from contacting them.

Liu's first task for her analytics team was to determine the true marginal cost of sending a push message. The team needed two key metrics to determine marginal cost. First, what was the loss in lifetime value associated with a customer blocking deal messages from Tuango? Second, by how much did an incremental pushed deal increase the probability that a customer would block future deals from Tuango?

---

*Professors Song Yao and Florian Zettelmeyer prepared this case to provide material for class discussion rather than to illustrate either effective or ineffective handling of a business situation. The names and the data used in this case have been disguised to assure confidentiality and some events are fictionalized. The case is partially based on the Tuscan RFM case by Professor Charlotte Mason.*

Getting at the first metric was easy. Tuango currently had two types of customers. Many customers used the mobile app but there were also many web-only customers. Liu decided to approximate the loss in value of a customer who refused deals on their mobile app by assuming that they would subsequently behave like web-only customers.

Getting at the second metric was harder. Luckily, the analytics team found that there was a lot of variation in the number of messages customers had received from Tuango in the past. Since this variation seemed to be largely random, the team could approximate the probability of blocking deal messages using the average fraction of customers who blocked deals on their mobile devices across groups of customers who received fewer or more deals from Tuango.

By multiplying the change in customer value from blocking deal messages with the probability of blocking deal messages, Liu's team determined that 2.5RMB was a good approximation for the true marginal cost of sending an additional deal. Liu knew that Tuango should send deal offers only to those customers for whom Tuango's expected value exceeded this marginal cost.

Liu did not have much first-hand experience in mobile marketing. However, for years she had been applying a variety of targeting techniques in direct marketing campaigns, including for direct mail, emails, banner ads, online video ads, and so on. She believed that some of the techniques she had used so far should be applicable for mobile marketing as well. To test this, she planned to try some techniques for deal targeting. A good place to start would be RFM analysis, a simple and very popular targeting technique.

RFM has a long and successful history in database marketing to target customers. RFM stands for Recency, Frequency, and Monetary value. The fundamental premise underlying RFM analysis is that customers who purchased more recently, made more purchases, and made larger purchases are more likely to respond to an offer than customers who purchased less recently, less often, and for less money. RFM and its derivatives are still popular in the digital age, particularly because they are easy to implement even with the huge customer databases that many Internet companies have.

**RFM Classification**

Typically, firms group customers into quintiles – or five groups – when performing a RFM analysis. This leads to, at most, 125 (5 x 5 x 5) RFM cells or segments.

There are three typical approaches to determine RFM segments:

- **Independent quintiles**

  This approach computes quintiles independently for recency, frequency, and monetary value. That is, the entire customer list is sorted based on recency and divided into recency quintiles. Then, the entire customer list is re-sorted based on frequency and divided into frequency quintiles. Finally, the entire customer list is sorted one more time based on monetary value and divided into monetary quintiles. The three quintile variables are then combined to form a RFM index. For example, a customer in the 1$^{st}$ recency quintile, 2$^{nd}$ frequency quintile, and 4$^{th}$ monetary quintile – is assigned a RFM index of 124.

- **Sequential quintiles**

  This approach first computes quintiles for recency. Then quintiles for frequency are computed *within* each of the five recency quintiles, resulting in a total of 25 recency/frequency combinations. Finally, *within* each of these 25 groups, quintiles for monetary value are computed.

- **Intuitive groupings**

  This approach uses intuitive splits rather than quintiles to form groups. For example, customers may be grouped into recency groups determined by: (1) purchase in the last 6 months, (2) last purchase 6-12 months ago, (3) last purchase 12-24 months ago, (4) last purchase 24-36 months ago, and (5) last purchase more than 36 months ago. This approach is 'intuitive' in that it is easy to see what is meant if a customer is in the second recency group. However, it does rely on the analyst's judgment to know where to 'draw the lines.'

**Testing the performance of RFM for mobile deal offers**

Liu decided to test the performance of RFM on a deal for a 1, 2, or 3-hour Karaoke session at one of Hangzhou's leading Karaoke chains. The deals were priced at 129RMB, 209RMB, and 259RMB, respectively. Tuango's fee was 50% of the deal price when a deal was sold to customers (similar to Groupon in the US).

To guide her team on how to perform a RFM analysis to target deals to Tuango's mobile customers, Liu wrote down the key steps:

1. Select all mobile customers that expressed interest in the Karaoke category (i.e., 278,780 customers in Hangzhou).
2. Categorize the customers into RFM cells.
3. Randomly select a 10% sample of mobile customers and offer all of them the deal. This meant that a total of 27,878 customers received the Karaoke deal. This would be the data used for analyses.
4. Track response in each RFM cell (i.e., response rate, order size, etc.).
5. Assess response across recency, frequency, and monetary quintiles to get a feel for the data before performing a RFM analysis.
6. Assess response, profitability, and return on marketing expenditure per RFM cell.
7. Use the test results to determine if RFM analysis can improve profits and return on marketing expenditures in mobile deal targeting campaigns.
8. If so, use the profitable RFM cells to target the remaining 250,902 customers.

**The data**

After the 10% random sample had been offered the deal, Liu's analytics team received a dataset with the results. The dataset contains the information needed for the RFM analysis (see Exhibit 1 for a summary and definitions of key variables found in the *tuango_pre.pkl* dataset).

**Exhibit 1**
Variable Names and Descriptions
(tuango_pre.pkl dataset)

| Name | Description |
|---|---|
| userid {object} | Unique user ID |

*Recency, Frequency, and Monetary variables*

| | |
|---|---|
| recency {integer} | Days since last purchase (excluding the Karaoke deal offer) |
| frequency {integer} | Number of deals purchased during the one-year period prior to the Karaoke deal offer |
| monetary {float} | Average amount spent per order (in RMB) during the one year period prior to the Karaoke deal offer |

*RFM index:*

| | |
|---|---|
| rfm_iq_pre {object} | Independent RFM indices |

*Response to the customized push message*

| | |
|---|---|
| buyer {category} | Bought one of the three Karaoke deals? (“yes” or “no”) |
| ordersize {float} | Amount spent on the Karaoke deal (in RMB) |

*Other observed variables in the dataset*

| | |
|---|---|
| platform {category} | Platform used by customer to register with Tuango |
| category {integer} | Category of last purchase before current targeting |
| mobile_os {category} | Customer’s mobile OS |
| training {integer} | Splits the dataset into training (1) and test (0) data. The training sample is used to collect responses |

**Assignment**

To complete this assignment, you should use the lecture notes on RFM and the Python code examples that contain the calculations for the Bookbinders RFM analysis. You can get the folder with the BBB RFM analysis by running the command below from a terminal in the docker container.

```
usethis "https://www.dropbox.com/sh/zjl0kxegyyoivgv/AACaiiHgMR7JyEKHVK33KyE3a?dl=1"
```

Data for the Tuango case is included in the GitLab repo (data/tuango_pre.pkl). For question 14 you will need the data/tuango_post.pkl

***Part I: Preliminary and Quintile Analysis***
***(Q1 to Q6, 3 points each)***

Use the "tuango-pre.ipynb" file in the repo you cloned from GitLab to answer questions 1-13.

1. What percentage of customers responded (i.e., bought anything) after the push message?

2. What was the average amount spent on the Karaoke deal by customers that bought one (or more)? Use the **ordersize** variable for your calculation.

3. Create independent *quintile* variables for recency, frequency and monetary.

4. Create bar charts showing the *response rate* (i.e., the proportion of customers who bought something) for this deal per (independent) recency, frequency, and monetary quintile (i.e., 3 plots).

5. Create bar charts showing the *average amount spent (in RMB)* (i.e., **ordersize)** per independent recency, frequency, and monetary quintile *using only those customers who placed an order after the push message.* **Hint**: constrain the data used for the plot with a "filter" (i.e., 3 plots).

6. What do the above bar charts reveal about the likelihood of response and the size of the order across the different recency, frequency, and monetary quintiles?

***Part II: Profitability Analysis***
***(Q7, 2 points; Q8 to Q13, 6 points each, Q14, 10 points)***

Create two RFM indices. First, create a variable ***rfm_iq*** using the independent quintile approach. As a way to check you have done this correctly, compare your index to the variable **rfm_iq_pre** already in the dataset. Next, create a variable ***rfm_sq*** using the sequential quintile approach. To compute this index, review the steps described for the Bookbinders case discussed in class.

The following questions will ask you to use data to predict the profit and the return on marketing expenditures from offering the deal to the remaining 250,902 potential customers (i.e., 278,780 – 27,878).

To calculate profit and return on marketing expenditures assume the following:

Marginal cost to offer a deal is 2.5RMB
Fee on each deal sold is 50% of sales revenues

7. What is the breakeven response rate?

8. What is the projected profit in RMB and the return on marketing expenditures if you offer the deal to ***all*** 250,902 remaining customers?

9. Evaluate the performance implications of offering the deal to only those customers (out of 250,902) in RFM cells with a response rate greater than the breakeven response rate.

Generate your result based on **both** sequential and independent RFM. Determine the **projected** profit in RMB and the return on marketing expenditures for each approach.

**Hint:** As you will be evaluating multiple models in this assignment, you should define a function (call it "perf_calc") that can calculate the response rate, profit, ROME, etc. and that can be re-used for each model. Even better, would be if the function also prints text for each model with all information about response rates, costs, profit, ROME, etc. See the tuango_pre.ipynb notebook for an example of what the call to such a function should look like in python. "profit_sq" and "ROME_sq" can be used in performance plots (see question 13 below). Note that the function does not need to generate the "smsto_sq" variable. Rather, the function should use information contained in the "smsto_sq" variable that you added to the dataset to calculate the response rate, ROME, etc.

**Hint:** Check that your *rfm_iq* variable is exactly equal to *rfm_iq_pre* using the code snippets below

**Python:**

```
(tuango.rfm_iq_pre != tuango.rfm_iq).any()
(tuango.rfm_iq_pre == tuango.rfm_iq).sum()
```

10. What do you notice when you compare the ***rfm_iq*** and ***rfm_sq*** variables? That is – do the two approaches generally yield the same RFM index for any given customer? What do you see as the pros and cons of the two approaches (from a statistical as well as logical perspective) and why?

11. The answer to question 9 assumes a single breakeven response rate that applies across all cells. Redo your analysis for sequential RFM based on a breakeven response rate calculated for each RFM cell. What implications can you draw from the difference in predicted performance compared to question 9?

    Hint: Imagine how you would calculate breakeven if you only had one cell/bin. What would your cost and margin numbers be? Once you have determined how to do this for one cell, extend your approach to apply to all cells separately.

    Note: You only need to calculate smsto_pcsq for this approach. You can use your perf_calc function to determine the performance implications without any adjustments

12. The answer to question 9 does not account for the fact that the response rate for each cell is an estimated quantity (i.e., it has a standard error). Redo your analysis from question 9 for independent RFM, adjusting for the standard error of the response rate in each cell.

13. Create a bar chart with profit information and a bar chart with ROME numbers for the analyses conducted in questions 9, 11, and 12

14. You also have access to a dataset with the results from the SMS roll-out (tuango_post.pkl). Tuango actually contacted all 250,902 customers. The data has a "training" variable (training = 1 for the data used in the test, training = 0 for the remaining customers). You can use this variable to help evaluate the actual performance for each

of the different RFM approaches. Re-create the plot in question 13 based on this new dataset.

Copy your “tuango-pre.ipynb” file to a new file “tuango-post.ipynb”. Instead of using the tuango_pre.pkl data, load tuango_post.pkl. You should be able to re-use most of your code and text. Create a new function perf_calc_actual that calculates the actual performance for each targeting approach on the 'roll out' sample (i.e., training == 0). Also, keep your perf_calc and use it to check that this still calculates the same values you were seeing before based on the tuango-pre.pkl data. The perf_calc_actual function will be similar to perf_calc but you will need to make some changes to determine the actual outcomes. Also make sure to check that your rfm_iq variable is the same as rfm_iq_pre in the tuango_post.pkl dataset.

**Hint:** It is important that you do NOT use any information about buyers that were in the 'roll out' sample (i.e., training == 0) when calculating the break-even response rate etc. for targeting.