Predicting customers' health supplement purchase propensities for precision marketing at MeCome pharmacy

Created by Group5 William Feng, Silvia Yang, Vivian Lee, Thompson Lin

|Executive summary

Surviving in this competitive pharmaceutical market is not as simple as one would think it to be. Faced with a range of strong competitors, MeCome's main business problem is to find a solution for its low customer loyalty. Therefore, our group proposes to use naive, XGBoost, and feature engineering as our primary data mining tools for precision marketing at MeCome. After deciding our tools, we moved on to data partition, data preprocessing, and RFM scores for better exploration. Furthermore, to simplify our search on a pharmacy that provides a variety of products, we have narrowed our research by building a predictive model for a sub-categorical product "2-06-01" (Natural enzyme products). The outcome variable is binary, 1 represents customers who will buy health supplements in the following month, while 0 means will not buy for the next month.

Through rigorous cleaning of the data, we then tested the naive, XGBoost, and feature engineering. After our explorations with different data mining strategies, we discovered that for data analysis, feature engineering performed poorly and naive has the best prediction performance. Additionally, the best tool to use to target potential customers that have high purchase propensity but does not shop in MeCome is by using machine learning through different features.

In terms of identifying customers with high health supplement purchase propensities, our recommendation to MeCome is to focus on deployment and future implementations. For deployment, we advise the company to send text messages to the high purchase propensity customers identified in our analysis. Additionally, for future implementation, a one-time analysis should be conducted before each promotion by the frequency of 2-3 months.

|Business Goal and Data Mining goal

MeCome is a pharmacy that sells daily necessities, health supplies, and prescription drugs to customers primarily in the Taipei metropolitan, the New Taipei city, and Taoyuan area. Their goal is to be able to sell a variety of products ranging from so-called "birth to death"; so that customers can purchase health-related supplies when they need it. The greatest challenge they are currently facing is low customer loyalty. The company suspects that local customers might not regard MeCome as their first choice due to the intense competition from the alternative stores. Even the company distributes advertising coupons regularly to the local communities. However, due to a lack of knowledge and skills for digital marketing, the company has never conducted any personalized marketing to targeted customers, which makes the marketing cost high, ineffective, and inefficient.

Our goal is to help MeCome Pharmacy to increase customer loyalty and retention by predicting potential customers who have a high probability to respond to predefined personalized marketing campaigns. We believe that our solution can help the company increase both sales and profit margin by:

1. reducing advertising costs and developing more effective marketing strategies,
2. targeting a group of customers who are loyal and contributing a large portion of sales to the company

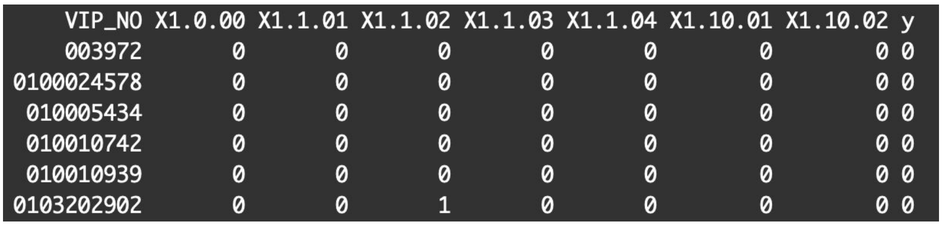
3. optimizing the value extraction from the customer lifecycle.

To achieve the business goal mentioned above, our data mining goal is to **find out the potential customers who will purchase health supplements in the following months**. The outcome variable is binary, 1 represents customers who will buy health supplements in the following month, while 0 means will not buy for the next month. Since MeCome has thousands of products, it's hard to build a predictive model for every individual product. Hence, we only targeted a subcategory - "2-06-01" (Natural enzyme products), which belongs to health supplements that are known to have high gross profit and the short customer repurchase cycle.

|Data description & Brief data preparation details

We collected data from MeCome's MS SQL database. We found 7 data tables are related to our goal, and the three main data tables are sales transactions (30,845,110 raws), products (71,718), and VIP members data (247,242). The data range from 2009 to 2019, and we used the first 6 of the seven months to find customer behavior and use the last month to compute the outcome variables. In data partition, we were using purchase history (2018.11.1-2019.5.30) to predict whether customers will buy health supplements in the following months (2019.6.1-7.17). We separate the data into 70% training data and 30% testing.

In data preprocessing, we transform transaction data into a sparse matrix (Figure 1). Each row means the purchase history of the VIP member, and each column indicates the buying amounts of that particular products. We then also filter sales prices and quantities that are over 0. Since some customers might use free points to buy things or take their stored products from the store. Lastly, we also used RFM scores as features to help improve our predictions and explore the data we have (Figure 2). From figure 2, we can easily see those different types of customers, and this gives us a brief picture of our customers and help us adjust strategies in the future.



```
    VIP_NO X1.0.00 X1.1.01 X1.1.02 X1.1.03 X1.1.04 X1.10.01 X1.10.02 y
    003972       0       0       0       0       0        0        0 0
0100024578       0       0       0       0       0        0        0 0
 010005434       0       0       0       0       0        0        0 0
 010010742       0       0       0       0       0        0        0 0
 010010939       0       0       0       0       0        0        0 0
0103202902       0       0       1       0       0        0        0 0
```

Figure 1. Sample of several rows after data preprocessing.

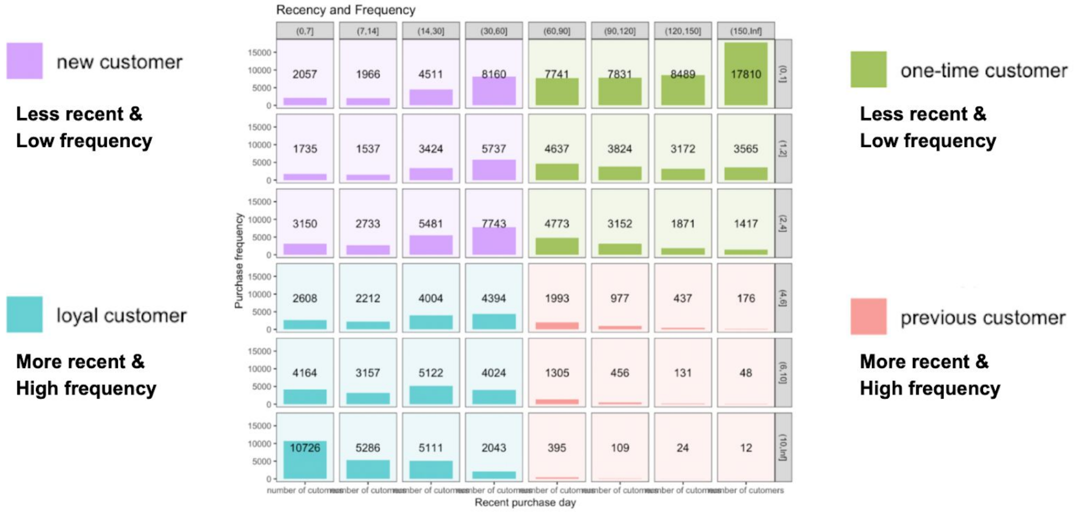


Figure 2. Exploratory data analysis using recency and frequency.

|Data mining solution

In this project, we used the majority vote as our benchmark and tried naive forecast, XGBoost, and feature engineering with several models. In the end, we compare different methods of performance by their confusion matrix and lift chart.

- **Naive forecast**

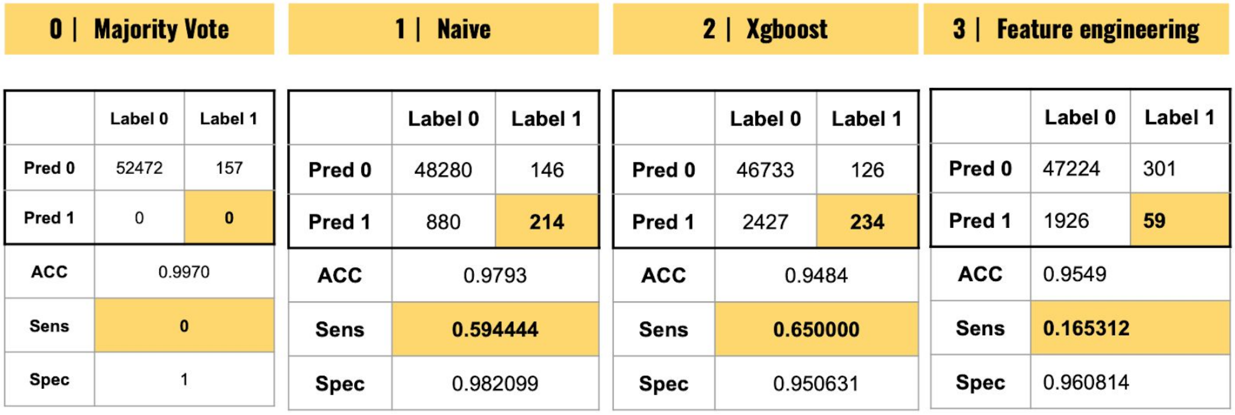
  We predict customer who has bought the target product before (in the training dataset) as 1. For example, if the customer buys the target products in the last 6 months (the training dataset), the naive rule will predict that the customer is going to buy the same products in the next month (the testing dataset).

- **XGBoost**

  We tried different ways to clean the data to see which had better performance, including purchase history with the sum of quantity, binary purchase history (1 for purchase, 0 for not purchase), and purchase history with RFM scores. For the methods and parameters, we tried Tree with oversampling parameters and the random forest. We found out that purchase history with the sum of quantity and the random forest has the best performance in XGBoost.

- **Feature engineering with several models**

  For the feature engineering, we calculated the cosine similarity of all purchased items in the training dataset. We selected the items that have highly correlated with our target item. Finally, we got about 80 correlated items as predictors. Then we trained and predicted via different machine learning models, such as Naive Bayes, decision tree, and random forest with the same training dataset, features, and testing dataset. We found that feature engineering with Naive Bayes results in the best performance.



0 | Majority Vote

	Label 0	Label 1
Pred 0	52472	157
Pred 1	0	0
ACC	0.9970	
Sens	0	
Spec	1	

1 | Naive

	Label 0	Label 1
Pred 0	48280	146
Pred 1	880	214
ACC	0.9793	
Sens	0.594444	
Spec	0.982099	

2 | Xgboost

	Label 0	Label 1
Pred 0	46733	126
Pred 1	2427	234
ACC	0.9484	
Sens	0.650000	
Spec	0.950631	

3 | Feature engineering

	Label 0	Label 1
Pred 0	47224	301
Pred 1	1926	59
ACC	0.9549	
Sens	0.165312	
Spec	0.960814	

Figure 3. Confusion matrix of each method.

Overall, the three confusion matrices above show that all our approaches have better performance than the benchmark. Among these three methods, the naive forecast and XGBoost work better than feature engineering. As you can see that we highlight the true positive and the sensitivity to help compare the evaluation. In terms of true positive and sensitivity, XGBoost performs better than naive. However, if we look at false positive, naive performs better than XGBoost.

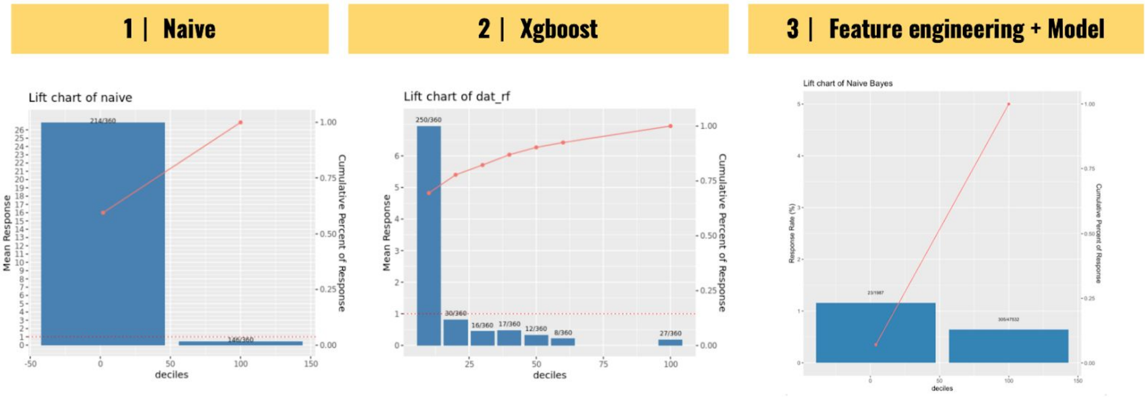


Figure 4. Lift chart of each method.

If we only compared the lift charts above, we saw that the naive prediction provides better lift performance than others. One interesting finding is that the feature engineering plus traditional machine learning algorithms approach does not perform well as it was expected; the XGBoost trained with all features outperformed the feature engineering.

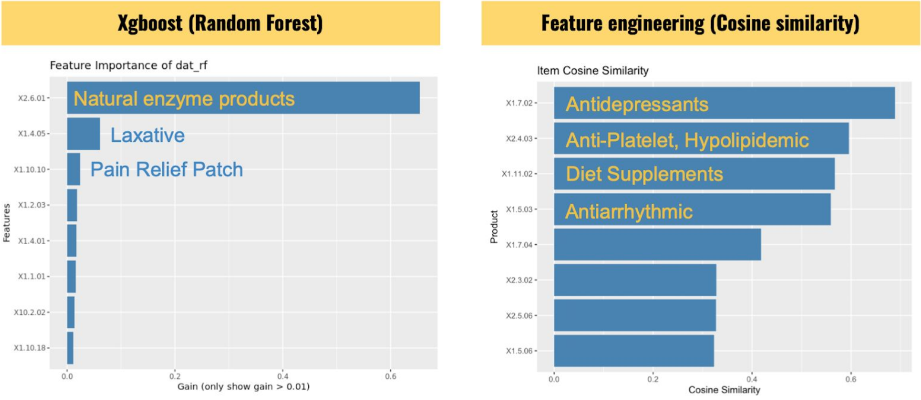


Figure 5. (left) XGBoost Feature importance, (right) Cosine Similarity Feature importance. These two graphs show some interesting findings by different method approaches.

XGBoost random forest feature importance:

From figure 6 (left), the y-axis represents the important features, and the x-axis represents the importance score. The most important feature is the item “2-6-01”, which represents the natural enzyme products, and it is also our target item. This information told us that the random forest prediction by XGBoost is similar to the naive forecast. Two-thirds of the buyers did purchase the product, so that’s why the model tends to believe the naive forecast. Nonetheless, the model also captures some new buyers, and that’s why we have other important features here. The second feature importance is the laxative pill, and it is a close-correlated product to our target, the natural enzyme product because both have the effect of losing weight.

Cosine similarity for feature engineering:

The cosine similarity scores show different results than the XGBoost on the ranking of the features. The top three item categories found by the cosine similarity approach, the antidepressants, antiplatelet, and diet supplements, which are entirely different compared to XGBoost random forest.

|Conclusions (advantages and limitations) and operational recommendations

This project **identifies customers with high health supplements purchase propensities** using data mining algorithms to achieve precision marketing and thus increase customer loyalty. In terms of data analysis, **feature engineering didn't help much on prediction in this situation**. Overall, "Naive" has the best prediction performance. However, machine learning can still help us capture some potential customers. Here, what we mean "the potential customers"are those customers who might have high purchase propensities with particular product but didn't make their purchase with MeCome. In this case, "naive" cannot help us, but machine learning can help us identify these potential customers.

As for the deployment, we suggest that MeCome can send text messages to those customers that we have identified with high purchase propensities. Depends on what budgets are, we can target on the top decile of customers first, and so on. For future implementation, this one-time analysis should be conducted before each promotion, or by every 2-3 months. Besides the health supplements, the same analyzing and predicting process can also be used on any other product category with a small adjustments.