Recommendation System of Women's Clothing Stores on Chinese Online B2C Retail Platforms

Yifan Liu, Qiran Zhang, Yuxuan Yang

1 Introduction

Within the last decade, there was an increasing trend in online shopping in China. Especially, a big jump in online demand from 79.1% to 81.6% appeared due to the covid epidemic when most consumers altered their purchasing behaviour from offline to online shopping. Therefore, this study aims to build a recommendation system for Chinese online retail platforms based on a two-stage model to meet this increased demand, prompt transactions and sales, and thus boost consumption and economic growth in China.

Specifically, B2C platforms contributed a huge proportion, around 80%, of online sales in China in 2022. Within them, Tmall led the market by having the largest number of transactions, over 63% of the total market share, thus the model fitted based on data from this platform would be more representative and generalized. Of all the product categories in the online shopping trend, women's clothing is the most popular. Women's clothing is considered a necessity, requiring consistent purchasing. Moreover, its diverse range makes it more likely to be influenced by recommendations. Therefore, a recommendation system focused on the women's clothing sector will be meaningful and flexible.

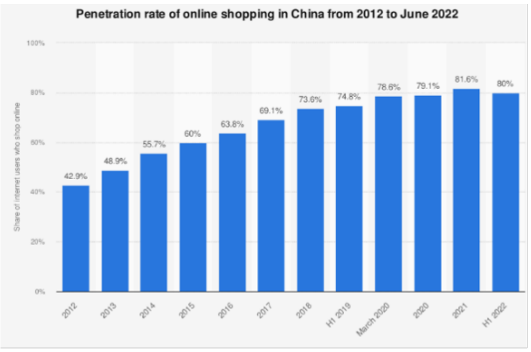


Figure 1: Penetration Rate of Online Shopping

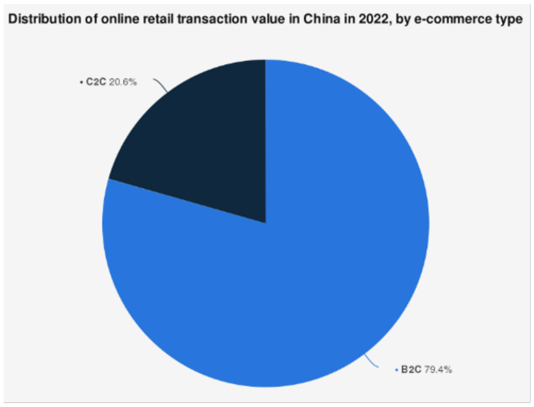


Figure 2: B2C Marketshare

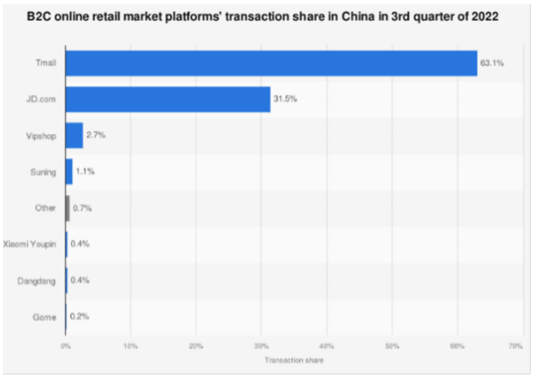


Figure 3: TMall Marketshare

To construct our recommendation system, there are two questions that need to be answered: first, would a certain consumer buy women's clothes on a Chinese B2C online retail platform next month? Second, if she/he is predicted to buy in next month, which sellers would they buy from? These questions are going to be answered in the following sections.

There are some existing papers studying online sales predictions, which share similarities with ours, but some special and unique highlights of our paper stand out and mark the contributions of this paper. We used the time-series cross-validation approach as Yuyu et al did in their study to handle time-dependency and optimize hyperparameters in our time-series models. However, we conducted a Chinese sentiment analysis on sellers' feedback as one of our predictors, which indeed improved the accuracy of the predictions a lot. Hyunwoo et al and our paper both extracted the count of past clicks per consumer as predictors. However, due to limited data, we were not able to access consumer demographics as they were collected. Thus, our two-stage model focuses more on the predictions based on consumers' past operations, including not only clicks but also whether they add potential products to cart and the ratios calculated by each combination between clicks, quantity purchased and "add to cart" actions, and sellers' feedback. Karandeep et al and us all applied random forest and gradient boosting, commonly being used when forecasting sales. Whereas the investigation object of this paper is to Chinese platforms instead of Brazilian ones in their paper. Due to systematic differences in online markets in these two countries potentially caused by different demographics, purchasing behaviours, and platform interface designs, our study will give Chinese platforms insight and solid proof of how they can prompt transactions.

2 Data

2.1 Data Preprocessing

The datasets being used in this paper are the product, review, and customer operations data in Tmall, a popular Chinese online B2C retail platform, collected from April 2014 to September 2014. The customer operations dataset consists of a massive 1.2 billion operations which encompass buy, collect, and click actions. The product dataset contains more than 1 million of detailed information on products and sellers such as product id, seller id, product category, and product brand. The review dataset contains customer feedback on products.

One of the main challenges we encountered was accessing the massive customer operations data in Python. To overcome this issue, we split the dataset into 120 chunks, each chunk containing ten million customer operations data. For each chunk, we only consider the operations on Women's clothes as this is our primary area of interest. Then, we group each chunk by customer id to determine the total number of operations conducted by each customer. This total number of operations is referred to as the ”operation index.” After processing all 120 chunks, we ranked the operation index and identified 5567 highly active users whose operation index ranged between 1500 and 5000. We then selected all the operations of these target users from the original customer operations data. The 1500 cut-off was chosen to focus only on highly active customers with consistent online shopping habits. These customers are more engaged with the online shopping platform, and targeted product recommendations could further stimulate their purchasing. This approach can also improve algorithm efficiency by considering only a smaller proportion of customers. We also set a 5000 cut-off, as we suspected customers with more than 5000 operations were not using the retail platform for shopping purposes. It is unreasonable for a customer to click on more than 30 products every day over the last 6 months. Additionally, we selected only the top 98 sellers in women’s clothing sales for the same reason.

2.2 Feature Extraction

We extracted four types of features from the original datasets: count features, ratio features, dummies, and review features. Count features were used to determine the total number of actions (i.e., buy, collect on the cart, click) of a given customer within a given time span (i.e., the entire 4 or 5 months, within the last month, within the last 10 days). We believe the counts of user actions can be useful measurements of how active users are, contributing to users’ choices in purchasing. Ratio features were created to calculate click-to-buy and cart-to-buy ratios, capturing different user segments as some prefer to browse around while others would purchase directly. Furthermore, the occurrences of some actions might affect future purchases, so we created dummy features to indicate the occurrences of such actions. One example dummy is whether the customer has collected any item in their cart since their last purchase. If so, they are more likely to buy an item than the customers who do nothing from their last purchase. Finally, we performed sentiment analysis on customers’ comments to identify their positive, neutral, or negative sentiments toward sellers, as we believe this could be a strong indicator of their future purchasing behaviour and differentiates our studies from others.

2.3 Train-Test Split

To split our data into training and test sets, we followed the same methodology as Zhang et al. (2014). Specifically, we used records from the first four months of our dataset (April, May, June, and July) to extract features for the training set, and the records from August to serve as the out-

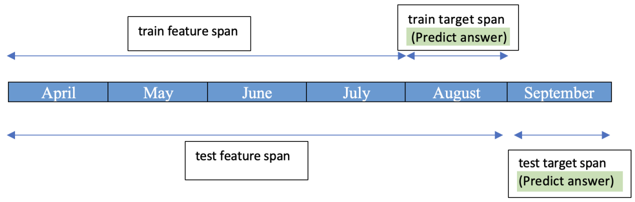


Figure 4: Train-Test Split

come variable. For the test set, we used records from the first five months (April to August) to extract features, while the records from September were used as an invisible outcome to evaluate the model's performance. Figure 4 illustrates the detail.

3 Methods

3.1 Two-Stage Model

We propose a two-stage model to address our research questions. The first stage of the model focuses on predicting whether a target customer will purchase from any target sellers in the next month, based on their past purchasing behaviour. We take individual customer ID as the unit of observation in this stage. The first stage model can be expressed in the following form:

$$y_i = f(HistoricalAction_{ik}) + \epsilon_i$$

where y_i denotes whether the customer i will have any purchase in the next month. $HistoricalAction_{ik}$ denotes the past purchasing behaviour of the customer i within a time span k. The primary objective is to filter out the customers who are unlikely to purchase from any sellers in the next month, thus reducing the total number of observations for the second stage.

In the second stage, we focus on predicting whether a target customer will purchase from a specific seller in the next month, given their historical actions and the fact that they are predicted to purchase from any seller in the first stage. The unit of observation for this stage is a (customer ID, seller ID) pair, where customer ID is taken from the outcomes of the first stage, and seller ID is chosen based on their popularity as discussed earlier. The pairs that are predicted as positive are the targets of our recommendation system. The second stage model can be expressed in the following form:

$$y_{ij} = f(HistoricalAction_{ijk}) + \epsilon_{ij}$$

where y_{ij} denotes whether the customer i will have any purchase from the seller j in the next month. $HistoricalAction_{ijk}$ denotes the past purchasing behaviour of the customer i from the seller j within a time span k.

The key motivation for the two-stage analysis is to improve algorithm efficiency. By considering only customers who are likely to purchase in the next month, we can significantly reduce the total number of observations and improve the speed of the fine-tuning process in the second stage. Additionally, the two-stage approach can mitigate the bias that may arise from different feature dimensions, as we utilize both user-specific features and user-seller-specific features.

3.2 Machine-Learning Algorithms

Feature Selection For each stage, we follow a feature selection process. We begin by selecting the features based on their feature importance and chi-square statistics. We then drop any features that have low importance and low chi-square statistics. This helps us identify the most relevant features and reduces the likelihood of overfitting, leading to a more efficient model.

Fine-Tune In each stage, we utilize two ensemble machine-learning algorithms, random forest classifier and gradient boosting tree, to build our models. To optimize the performance of these algorithms, we fine-tune their hyperparameters using a randomized grid search approach with 500 iterations. We present the best estimators identified by this grid search below. This approach ensures that we are maximizing the accuracy of our models and achieving the best possible results.

Stage-One	Stage-Two
click	click
click_last_m	click_last_m
click_recent	click_recent
purchase	purchase
purchase_last_m	purchase_last_m
cart	cart
cart_last_m	cart_last_m
cart_recent	cart_recent
	click_buy_ratio
cart_buy_ratio	cart_buy_ratio
click_buy_ratio_last_m	
cart_buy_ratio_last_m	
click_buy_ratio_recent	
cart_buy_ratio_recent	
last_purchase	last_purchase
last_cart	last_cart
	last_click
	sentimentscore$_{p}os$

Table 1: Feature Selection

	n_estimator	min_split	min_leaf	max_depth
stage 1	500	5	1	8
stage 2	500	5	1	8

Table 2: Fine-tune results for RF

	n_estimator	learning_rate	gamma	max_depth
stage 1	50	0.1	0.2	5
stage 2	150	0.1	0.0	5

Table 3: Fine-tune results for XGB

4 Results

In both stages, we measure the model performance based on its AUC score and F1 score, with the following formula:

$$AUC_{model} = \int_0^1 ROC(x)dx$$

$$F1Score_{model} = \frac{2 \cdot Preceision \cdot Recall}{Precision + Recall}$$

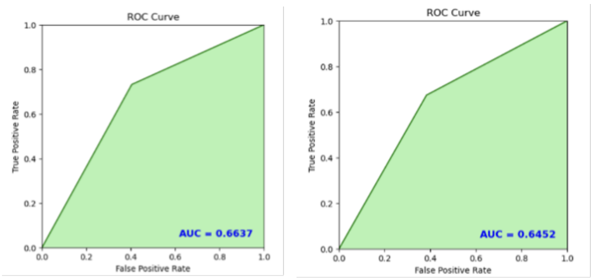


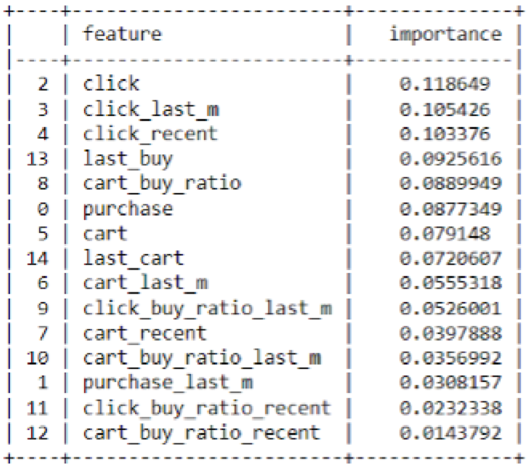
Figure 5: First-Stage AUC

		Predicted Response	
		Not Buy	Buy
True Response	Not Buy	1329	901
	Buy	896	2441

Figure 6: First-Stage Random Forest Classifier Confusion Matrix

		Predicted Response	
		Not Buy	Buy
True Response	Not Buy	1376	854
	Buy	1090	2247

Figure 7: First-Stage Gradient Boosting Tree Confusion Matrix



```
+----+-------------------------+--------------+
|    | feature                 |   importance |
|----+-------------------------+--------------|
|  2 | click                   |    0.118649  |
|  3 | click_last_m            |    0.105426  |
|  4 | click_recent            |    0.103376  |
| 13 | last_buy                |    0.0925616 |
|  8 | cart_buy_ratio          |    0.0889949 |
|  0 | purchase                |    0.0877349 |
|  5 | cart                    |    0.079148  |
| 14 | last_cart               |    0.0720607 |
|  6 | cart_last_m             |    0.0555318 |
|  9 | click_buy_ratio_last_m  |    0.0526001 |
|  7 | cart_recent             |    0.0397888 |
| 10 | cart_buy_ratio_last_m   |    0.0356992 |
|  1 | purchase_last_m         |    0.0308157 |
| 11 | click_buy_ratio_recent  |    0.0232338 |
| 12 | cart_buy_ratio_recent   |    0.0143792 |
+----+-------------------------+--------------+
```

Figure 8: First-Stage Feature Importance

We focus on the AUC and F1 scores because the class are imbalanced in our samples. Especially, in the second stage, the ratio between the positive class and negative class is 1 to 300. Since AUC is calculated based on the TPR and FPR, which are ratios, it is insensitive to the imbalanced distribution. Similarly, F1 scores account for the class imbalance by giving equal weight to precision and recall and ensuring that the models are not biased toward the majority class.

In the first stage, the fine-tuned random forest and gradient boosting tree both achieved a relatively high auc score and F1 score, but the prior did slightly better. Additionally, the random forest classifier reached a recall rate of 73%, whereas the latter model reached 67%. Here, we focused more on the recall rate since we want to capture more consumers who will buy next month and implement the next steps on them but not on people who are predicted to not buy. Since both classifiers had good AUC and F1 scores, we chose the random forest classifier for its superior recall rate.

In the first-stage models, we found that click-related features played a crucial role. One possible explanation for this is that the number of clicks within a given time span is a direct indicator of active customers who are more likely to purchase items within the next month. Additionally, click-related features may have larger variations, providing more information gains than other features.

In the second stage, we also selected the random forest classifier because of its high recall

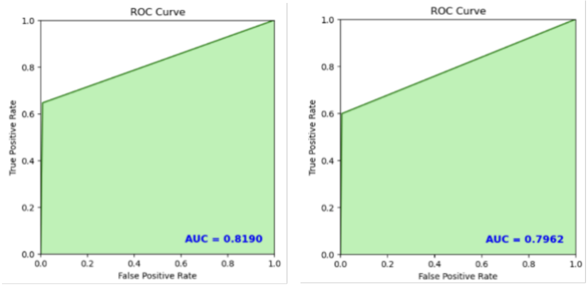
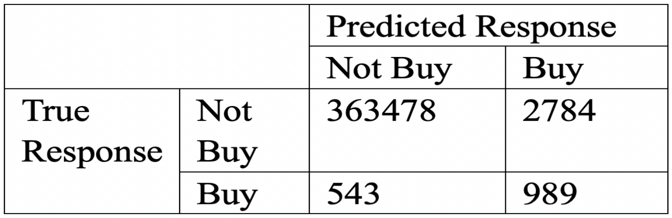
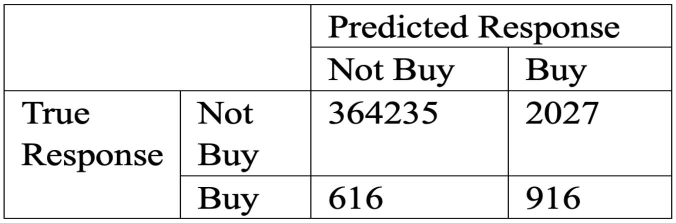


Figure 9: Second-Stage AUC



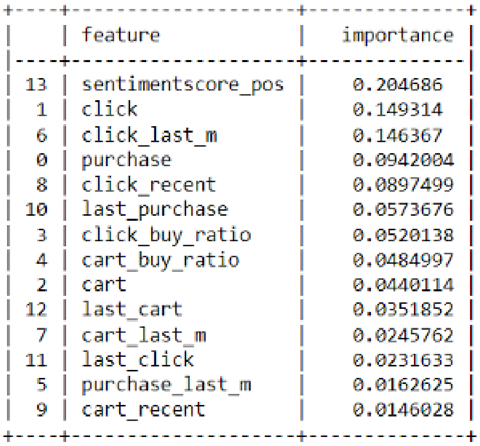
		Predicted Response	
		Not Buy	Buy
True Response	Not Buy	363478	2784
	Buy	543	989

Figure 10: Second-Stage Random Forest Classifier Confusion Matrix



		Predicted Response	
		Not Buy	Buy
True Response	Not Buy	364235	2027
	Buy	616	916

Figure 11: Second-Stage Gradient Boosting Tree Confusion Matrix



```
+----+--------------------+--------------+
|    | feature            |   importance |
|----+--------------------+--------------|
| 13 | sentimentscore_pos |    0.204686  |
|  1 | click              |    0.149314  |
|  6 | click_last_m       |    0.146367  |
|  0 | purchase           |    0.0942004 |
|  8 | click_recent       |    0.0897499 |
| 10 | last_purchase      |    0.0573676 |
|  3 | click_buy_ratio    |    0.0520138 |
|  4 | cart_buy_ratio     |    0.0484997 |
|  2 | cart               |    0.0440114 |
| 12 | last_cart          |    0.0351852 |
|  7 | cart_last_m        |    0.0245762 |
| 11 | last_click         |    0.0231633 |
|  5 | purchase_last_m    |    0.0162625 |
|  9 | cart_recent        |    0.0146028 |
+----+--------------------+--------------+
```

Figure 12: Second-Stage Feature Importance

rate, as the AUC and F1 scores were similar between the two classifiers. Overall, the random forest classifier provided reliable predictions with an AUC score of 0.819, a recall rate of 65%, and an accuracy rate of 0.99. Although the F1 score of the random forest classifier was relatively low at 0.41, this was not a significant concern for our use case. This is because the additional recommendations provided by the model are unlikely to discourage any purchasing. The model's high recall rate, which ensures that we capture as many potential customers as possible, is more important than precision in this context.

In the second-stage models, we found that the most important feature for predicting future purchases was positive sentiment toward the seller. This aligns with our expectation that customers who have positive feelings towards a seller and leave positive comments publicly are more likely to purchase from the seller again.

By combining these two random forest classifiers, a highly reliable recommendation system is now ready to be applied through pairing certain consumer who is predicted to purchase next month, using model 1, with predicted sellers they will buy from, using model 2. The pairing here could be achieved in different ways. For example, platforms can push notifications of products from certain sellers to certain consumers in a way like "check out this you may like/be interested in"; or they can stick these products on the top of the page when certain consumers search for women's clothes.

5 Conclusion

In conclusion, this paper proposes a two-stage model to predict the next month's purchase of a given active customer on a given popular seller in the women's clothes sector and recommend the predicted sellers to the customer on the TMall e-commerce platform. We leverage product, review, and operation datasets to extract 4 types of features, including count features, ratio features, dummies, and sentiment features, to capture different types of customer behaviours that are useful for predicting future purchasing. We fit our models using random forest classifiers and gradient-boosting trees. Overall, both classifiers perform well in both stages. However, the random forest classifier achieves a higher recall rate, which is critical for identifying potential customers for additional recommendations. In addition, we found that positive sentiment towards the seller was the most important feature for predicting future purchases. This suggests that customer satisfaction and loyalty are important factors in predicting future purchasing behaviour.

Nevertheless, we do have some limitations in our studies. For example, our data is from the year 2014 and might not apply to the current trends of the e-commerce market in China. Furthermore, there might be some confounding variables that will affect people's choices of buying or not, e.g. demographic information such as age and gender. Based on the study of Shreya (2022), the majority of E-commerce platform users in China lie between the ages of 25 to 44 in the year 2022. This indicates that middle-aged people are more likely to shop online. However, we don't have such age demographic information in our data from 2014, so our predictions on customers' purchase choices might not be perfectly accurate or complete. Therefore, future studies could build more reliable models by including demographic information, which might be difficult to obtain due to privacy concerns. Lastly, we rely on a two-stage model in our prediction, aiming to reduce the total number of observations and improve the algorithm efficiency. However, this method could introduce bias because the results in the second stage are affected by the results in stage one. However, in reality, a customer with low activity overall but high loyalty to a seller is likely to purchase from that seller in the next month, but he is not likely to be captured by our algorithm in the first stage. To address this limitation, future studies could explore more advanced techniques to build a one-stage model that incorporates a larger number of observations and features from different dimensions, without compromising on computational efficiency.

References

[1] CNNIC. (August 31, 2022). Penetration rate of online shopping in China from 2012 to 2022 [Graph]. In Statista. Retrieved April 10, 2023, from https://www.statista.com/statistics/302071/china-penetration-rate-of-online-shopping/

[2] Hyunwoo, H., Yang, S. K., & Kyung, J. C. (2018 March). Recommendation system development for fashion retail e-commerce. https://doi.org/10.1016/j.elerap.2018.01.012

[3] Karandeep, S., Booma, P.M., & Umapathy, E. (2020). E-Commerce System for Sale Prediction Using Machine Learning Technique. https://iopscience.iop.org/article/10.1088/1742-6596/1712/1/012042/pdf

[4] MOFCOM China. (February 21, 2023). Distribution of online retail transaction value in China in 2022, by e-commerce type [Graph]. In Statista. Retrieved April 10, 2023, from https://www.statista.com/statistics/1346882/china-online-retail-sales-distribution-by-ecommerce-type/

[5] Shreya. (2022, October 7). China ecommerce market in 2022 - all you need to know. AdChina.io. Retrieved April 10, 2023, from https://www.adchina.io/china-ecommerce-market/

[6] Sohu. (November 17, 2022). B2C online retail market platforms' transaction share in China in 3rd quarter of 2022 [Graph]. In Statista. Retrieved April 10, 2023, from https://www.statista.com/statistics/866847/china-b2c-online-retail-market-platforms-by-transaction-share/?locale=en

[7] Yuyu, Z., Liang, P., Lei, S., & Bin, W. (2014, August). Large Scale Purchase Prediction with Historical User Actions on B2C Online Retail Platform. https://arxiv.org/pdf/1408.6515.pdf