J SYST SCI SYST ENG
Vol. 31, No. 2, April 2022, pp. 247–264
DOI: https://doi.org/10.1007/s11518-022-5520-1
ISSN: 1004-3756 (paper), 1861-9576 (online)
CN 11-2983/N

Item-level Forecasting for E-commerce Demand with High-dimensional Data Using a Two-stage Feature Selection Algorithm

Hongyan Dai,[a] Qin Xiao,[a] Nina Yan,[a] Xun Xu,[b] Tingting Tong[c]

[a]Business School, Central University of Finance and Economics, Beijing 100081, China
daihy@cufe.edu.cn, 2017211037@email.cufe.edu.cn, yannina@cufe.edu.cn

[b]Department of Management, Operations, and Marketing, College of Business Administration, California State University, Stanislaus, Turlock, CA, 95382, United States
xxu@csustan.edu

[c]Institute of Supply Chain Analytics & International Business College, Dongbei University of Finance and Economics, Dalian 116025, China
tongtingting@dufe.edu.cn (✉)

Abstract. With the rapid development of information technology and fast growth of Internet users, e-commerce nowadays is facing complex business environment and accumulating large-volume and high-dimensional data. This brings two challenges for demand forecasting. First, e-merchants need to find appropriate approaches to leverage the large amount of data and extract forecast features to capture various factors affecting the demand. Second, they need to efficiently identify the most important features to improve the forecast accuracy and better understand the key drivers for demand changes. To solve these challenges, this study conducts a multi-dimensional feature engineering by constructing five feature categories including historical demand, price, page view, reviews, and competition for e-commerce demand forecasting on item-level. We then propose a two-stage random forest-based feature selection algorithm to effectively identify the important features from the high-dimensional feature set and avoid overfitting. We test our proposed algorithm with a large-scale dataset from the largest e-commerce platform in China. The numerical results from 21,111 items and 109 million sales observations show that our proposed random forest-based forecasting framework with a two-stage feature selection algorithm delivers 11.58%, 5.81% and 3.68% forecast accuracy improvement, compared with the Autoregressive Integrated Moving Average (ARIMA), Random Forecast, and Random Forecast with one-stage feature selection approach, respectively, which are widely used in literature and industry. This study provides a useful tool for the practitioners to forecast demands and sheds lights on the B2C e-commerce operations management.

Keywords: Forecasting, e-commerce, high-dimensional feature, feature selection

1. Introduction

Demand forecasting is a key component of supply chain management, which affects replenishment decisions, manufacturing planning, inventory management, logistics, and other aspects of enterprises (Abolghasemi et al. 2020, Dai et al. 2019, Goltsos et al. 2019). Facing the fierce competition in the e-commerce industry, e-commerce enterprises need to perceive the demand signal in advance to have a quick response to customer orders. With the development of data acquisition, storage, and processing technology in recent years, demand forecasting is becoming the dominant force driving the development of enterprises (Nikolopoulos 2021). E-merchants have accumulated a large number of high-dimensional data, such as price, customer comments, and the information about the competitive products, which are conducive to learn the demand fluctuation and improve the prediction accuracy (Leung et al. 2020).

However, e-merchants are also having greater challenges facing these massive data. First, e-merchants need to find appropriate approaches to leverage the large amount of data and exact forecast features to capture various factors affecting the demand. Second, it is important for e-merchants to efficiently identify the most important features to improve the forecast accuracy and better understand the drivers for demand change. The huge amounts of data leads to dimension curse (Li et al. 2006). Excessive features inevitably consume a lot of data processing and prediction time, and the irrelevant features may interfere with model training, causing problems such as overfitting.

To solve the above challenges, feature selection can be used to filter out the redundant and irrelevant features to reduce feature dimensions while retaining important information (Chen et al. 2017). It is an important data preprocessing process in forecasting with several advantages including data collection cost and processing time reduction and learning efficiency improvement (Maldonado et al. 2017). Previous studies conducting feature selection (Omuya et al. 2021, Ot et al. 2021, Sun et al. 2019) mainly used three common approaches – filter, wrapper, and embedded methods. Although these three methods have proved to have good forecasting performance, each of them has their own disadvantages such as high computational complexity, high computational cost and time, and non-interaction with the demand model (Kim et al. 2021).

The aforementioned disadvantages are amplified for the demand forecasting for e-commerce products due to the fact that these products have a short life cycle, low industry entry and exit barriers, and rapid market environment changes (Yan and Baowen 2011). Therefore, it is essential to update the forecasting model monthly, weekly, or even daily and make predictions on a daily level. Consequently, it is required that the demand forecasting has both high forecast accuracy and short computational time, which is the objective of this study.

In detail, this study designs five categories of features according to the characteristics of the e-merchants in a holistic view, in order to capture all the possible factors affecting the item-level demand. We further develop a two-stage feature selection algorithm to reduce to computational burden and in the meanwhile improve the forecast accuracy. We test our proposed algorithm with a large-scale dataset from the largest e-commerce platform in China. The numerical results with 21,111 items and 109 million sales observation show that our proposed algorithm delivers 11.58%, 5.81% and 3.68% forecast accuracy improvement, compared with ARIMA, Random Forecast, and Random Forecast with one-stage feature selection approach, respectively, which are widely used in literature and industry.

The main contributions of our study lie in the following two aspects. First, we utilize various features including historical demand, price, page view, reviews, and competition to conduct the demand forecasting using the high-dimensional data. Second, we develop a two-stage feature selection algorithm to efficiently identify important features to improve item-level forecast accuracy and computational efficiency. Our study can help e-merchants to utilize the big amount of historical data to forecast future customer demand in a more efficient way by improving the forecasting accuracy and reducing the computational burden. In this way, e-merchants can leverage accurate and effective item-level demand forecasting to improve operations management efficiency, such as inventory management and logistics management.

The rest of the study is organized as follows. Section 2 reviews the literature. Section 3 introduces our two-stage feature selection algorithm and demand forecasting framework.

Section 4 presents the data preprocessing procedures and descriptive statistics. Section 5 displays demand forecasting results. Finally, Section 6 concludes this study and proposes directions for further research.

2. Literature Review

Demand forecasting has been widely researched in previous literature (Chou et al. 2020, Narayanan et al. 2019, Yu et al. 2019). Companies can use demand forecasting, combined with price optimization to achieve a better financial performance (Ferreira et al. 2016). The key factors to forecast demand can be categorized as follows.

The first category of factors refers to the historical data of the sales (Pannakkong et al. 2018, Petropoulos et al. 2018). The historical sales can be fluctuating that depends on various variables incorporated in the demand model in previous studies, such as seasonal index (Choi et al. 2014), trend (Hyndman et al. 2002), holiday index (Li and Lim 2018), temporal aggregate sales (Athanasopoulos et al. 2017), promotion sales (Fildes et al. 2019), perishability (Van Donselaar et al. 2016), and intermittent demand (Li and Lim 2018).

The second category of factors includes price and promotion of the product. Price directly affects demand (Huang and Liu 2006, Wang et al. 2020, Wu et al. 2020). Compared with brick-and-mortar stores, e-commerce companies have lower labor and rental costs, and they have a higher extent of promotions to attract consumers (Kamakura and Kang 2007). Scholars have studied the impact of price and promotion on customer demand (Neto et al. 2016, Pang et al. 2015, Xu et al. 2017). Promotion features are widely used in the forecasting (Ali et al. 2009, Ramanathan and Muyldermans 2010, Trapero et al. 2015).

The third category of factors refers to customers' behavior. For example, the sales time during the day and the week influences customers' demand (Subramanian and Subramanyam 2012). Clickstreams and customers' willingness to engage in the purchase process also affect customer demand (Andersen et al. 2000, Besbes et al. 2016).

The fourth category of factors refers to customers' evaluation of past sales, which mainly refer to their reviews. With the online shopping rapid developing, customer online reviews play an important role to predict the future demand (Chong et al. 2016 2017). Online reviews can generate the electronic word-of-mouth (eWOM) effect, which affects future customers' purchase intention and behavior (Cantallops and Salvi 2014). Fan et al. (2017) used sentiment analysis extracting customer emotion from their online reviews and based on this information to forecast future demand. Zhu and Zhang (2010) found that online customer reviews have a significant impact on sales, and the relationship is moderated by product and customer characteristics.

The fifth category of factors is competition. The importance of product competition on demand has been confirmed by previous research (Cao et al. 2019, Ding and Liu 2021, Lu et al. 2016). Competitive information at both the brand and category level has been used in demand forecasting (Divakar et al. 2005, Ma et al. 2016).

The aforementioned various categories of influential factors of demand leads to the hundreds of features that are extracted from high-dimensional data. This can cause the issue of excessive features. To handle this challenge, various demand forecasting methods are proposed by previous studies, which include regressions-based models (Korobilis 2017), decision trees (Martínez et al. 2020), neural networks (Wu et al. 2016), and support vector machine (Xie et al. 2021).

However, noisy features may affect forecast accuracy, as they may result in over-fitting and increase computational time and costs. Fea-

ture selection, as a method to reduce data dimension and identify effective features from the massive data, is proved to be an efficient approach to solve this issue and is studied in previous research (Abasabadi et al. 2021, Chandrashekar and Sahin 2014, Jiménez-Cordero et al. 2021). For example, Ma et al. (2016) proposed a four-step procedure including identifying influential categories, building of explanatory variable space, using multistage regression to select variables and estimate models, and using rolling schemes to generate forecast to deal with the high dimensional data to forecast demand.

Filter, wrapper and embedded methods are among the three most common feature selection methods. In terms of the filter method, it is based on the evaluation criteria of the intrinsic properties of data set, and thus it has a relatively high computational efficiency and low time consumption (Guyon and Elisseeff 2003, Navarro and Muñoz 2009, Peng et al. 2005). However, the cons of filter lie mainly in the fact that it does not interact with the demand model. Regarding the wrapper method, it uses the performance of the forecasting algorithm to evaluate the performance of the features (Nakariyakul and Casasent 2009, Reunanen 2003). However, the downside of the wrapper method is that it has a high computational complexity because training and testing are needed for each subset evaluation. In particular, a larger data set will cause a longer execution time. In terms of the embedded method, it is named for embedding the feature selection algorithm into the forecasting algorithm as a component (Guyon et al. 2002, Maldonado et al. 2011). Similar to the wrapper method, the embedded method also obtain a high forecasting performance with the expense of computational cost and time (Kim et al. 2021). The high computational cost and time cost make these selection procedures difficult and even infeasible to be implemented in a dynamic and large-volume item-level forecasting context. Therefore, some researchers (Chiew et al. 2019, Giang et al. 2019, Nakariyakul 2018) are attempting to combine the filter method with other methods, such as the two-stage selection process. However, most of these methodologies select the relevant features in the first stage using relatively subjective evaluation criteria, such as entropy and mutual information, rather than forecasting accuracy (Got et al. 2021).

Our study extends previous studies about demand forecasting through two aspects. First, we identify more features to accommodate the item-level forecasting. Second, we propose a two-stage feature selection algorithm embedding two wrapper methods and integrate it into the forecast process to increase the forecast accuracy and reduce the computational burden.

3. Framework

In this study, we first extract forecast features from the high-dimensional online and offline data, which aims to accommodate item-level forecasting. Then, we propose a two-stage feature selection algorithm to identify the important features from a large feature pool. Based on the selected feature set, we train the demand forecast model and test with a large-scale real-world data from the largest e-commerce platform in China.

3.1 Feature Extraction

The main challenges of the item-level forecasting in the e-commerce context lie in the facts of the low amount of, high variation of, and intermittent demand. Thus, high-dimensional features are needed to forecast item-level demand. We collect five categories of influential features that can affect the demand of e-merchants, including historical demand, price, page view, reviews, and competition. We elaborate these five specific features in detail below.

3.1.1 Historical Demand Features

Past customers' purchase behavior and past demand affect future online shoppers' purchase decision (Dong et al. 2017). Thus, we collect the historical demand as one of the influential factors for customer demand on a certain day. Examples are as follows:

- Demand of the SKU for the past 1 days, 7 days, and 30 days, respectively

- Demand of the SKU's subcategory for the past 1 day, 7 days, and 30 days, respectively

- Demand of the SKU's merchant for the past 1 day, 7 days, and 30 days, respectively

3.1.2 Price Features

The demand of products can be significantly influenced by price and promotion (Lee and Charles 2021). The e-commerce platforms often mark down price to promote sales during the holiday seasons. In addition, it is possible that price changes during the middle of a day, which results in different prices among different orders during that day. However, from our data set, we find this situation is very rare in our data, and thus we use the price of an item in its first order during a day to be the item price for that day. In addition, we use historical prices to reflect the impact of price change on demand. Examples are as follows:

- Price of the SKU every day of the past 30 days

- Price change of the SKU every day of the past 30 days

- The mean, maximum, and minimum of the SKU's price for the past 7 days and 30 days, respectively

3.1.3 Page View Features

This is the highlight of our features. Page views can serve as a signal on the website to reflect the popularity of the products and past customers' searching and shopping preference (Koehn et al. 2020, Yeo et al. 2018). We obtain the number of page views and the number of unique online visitors of an item each day on PC and APP, respectively, and then obtain the descriptive statistics of page views and unique visitors including mean, average, minimum, maximum, and standard deviation. We forecast the demand of the item on a certain day based on the calculated statistics of page views and unique visitors of the item for the previous 1 day, 3 days, and 7 days. We did not collect the number of page views more than a week ago due to the diminishing effect of past customers' behavior on feature customers' behavior with longer time periods (Ye et al. 2009). Examples are as follows:

- Total, maximum, and minimum number of the SKU's page views on PC and APP, respectively

- Total, maximum, and minimum number of the SKU's unique visitors on PC and APP, respectively

- The ratio between page views on PC and APP

3.1.4 Online Review Features

One of the advantages the e-merchants can utilize is to acquire customer online evaluations of their products and services with a relatively low cost (Hanna et al. 2019). After completing the transaction, customers have the opportunities to write and post their reviews online. These online reviews serve the eWOM functions to influence future customers online purchase intention and behavior (He et al. 2020). Thus, we collect and include customer ratings and performance about product quality, delivery, and service as features. Examples are as follows:

- Ratings for the product quality, delivery, and service, respectively

- Average, maximum, and minimum ratings for the SKU's subcategory

- Number of times of delayed delivery

- Number of times of early delivery

- Average time length of shipping

- Fulfiller identity – whether the products are delivered by the platform's logistics

3.1.5 Competition Features

One of the features of online shopping is easier comparison, where the merchants are just mouse-clicks away. The first stage of online shopping is information search, in which customers collect information to compare and make decisions of what product to buy (i.e., product selection issue) and from whom (i.e., merchant selection issue, (Bauer et al. 2006)). Thus, to reflect customers' information search and comparison process, we examine the competition level of the items. We divide competition features into two groups – number features and rank features.

In terms of the number features, an increasing number of merchants and items in a subcategory may result in stronger competition. Such number features can reflect how severe the competition is in the category or subcategory, which eventually affects demand. Examples of number features are:

- Number of SKUs in the SKU's subcategory
- Number of SKUs in the SKU's merchant

In terms of the rank features, during online shopping, buyers will be given numerous choices for their purchase. Choosing items can be an item ranking process of the buyer. For many buyers, comparison of historical demand is an important selection criterion during purchase, and thus the construction of ranking features is useful for demand forecast (Tang et al. 2021). Examples are as follows:

- Rank and proportion of demand of the SKU in its subcategory for the past 1 day, 7 days, and 30 days, respectively. - Rank and proportion of demand of the SKU in its merchant for the past 1 day, 7 days, and 30 days, respectively

3.2 Feature Selection

The aforementioned high-dimensional features provide an available pool of factors that may influence demand. A large number of features can be extracted from these categories. However, too many features may lead to overfitting, reducing forecast accuracy, and increasing computational burden. Moreover, the demand patterns of different items in various time periods may differ. Feature selection is an important step of machine learning (Kursa and Rudnicki 2010). Thus, a mechanism is needed to select the most relevant explanatory variables. One of the most popular methods is one-step stepwise selection, which starts with a null set and adds explanatory variables, step-by-step (Huang et al. 2014).

However, the one-step process is often time-consuming. We choose Boruta algorithm as the one-stage feature selection approach, which is built based on the random forest algorithm and select feature sets that are correlated with the demand (Kursa and Rudnicki 2010). The Boruta algorithm consists of the following steps. First, it builds shadow features by creating shuffled copies of all features and combining features and shadow features. Then, it trains a random forest with the combined features and evaluates the importance of each feature. Next, it confirms a feature that has a higher importance than the best of shadow features and rejects a feature that is deemed highly unimportant at every iteration. Last, the algorithm stops either when all features get confirmed or rejected. $T = 1, 2, \cdots, N$ is the index set of training period. For a given demand vector $D_T = \{d_t\}_{t \in T}$ and feature matrix $X_T = \{x_t\}_{t \in T}$ with m features, the complexity of Boruta algorithm is approximately $O(m \times N)$, where m and N are the numbers of features and objects, respectively (Kursa and Rudnicki 2010). This suggests that one-step Boruta has a high complexity cost for high-dimensional features in e-commerce.

In this study, we propose a two-stage feature selection algorithm to address this issue, which includes feature group selection stage and individual feature selection stage (algorithm 1). Following the stepwise method, the two-stage feature selection algorithm is to se-

lect the important feature groups, which can improve the accuracy of forecasting model.

In the first stage, features are grouped into L groups as $X_T = \{X_T^1, X_T^2, \cdots, X_T^L\}$ according to the information contained in the features as follow. First, features are classified into 5 groups according to the category. Second, we subdivide these groups by time. With the increase of time interval, the effect of historical information gradually decreases. Therefore, we subdivide the groups into short-term groups (features that describe the information within the latest 3 days), medium-term groups (within the latest 7 days), and long-term features (within the latest 30 days). Third, we differentiate value features from statistical features, based on which we categorize these features into 2 groups. The value features refer to the original data such as the price of the day, price of the previous day, and price of 2 days before. The statistical features refer to the statistical values such as the average and minimum price of the first 3 days. The value feature groups and statistical feature groups are substitutes for each other and may be redundant in model training.

To test the performance of each feature group, we divide the data set $\{D_T, X_T\}$ into two subsets: training subset $\{D_{(T_1)}, X_{(T_1)}\}$ with the first N_1 periods and testing subset $\{D_{(T_2)}, X_{(T_2)}\}$ with the rest $N - N_1$ periods. Then we train the forecasting model by training subset with and without the feature group, respectively, and compare the errors to see the performance of this feature group. To reduce the possibility of discarding important features, we change the sequence of feature groups and repeat the process of group selection. In this way, we can acquire the performance of the feature groups interacting with different features. For the feature groups with different results in the repeated cycles, groups that have a significant improvement on forecast accuracy in some cycles and are not redundant (no other substitute feature groups retained) will be retained. Otherwise, they will be removed.

Algorithm 1 Two-Stage Feature Selection Algorithm

Input: training dataset $\{D_T, X_T\}$, cycle index R
Output: relevant features X_T^r

```
Separate dataset {D_T, X_T} into two subsets {D_T1, X_T1} and {D_T2, X_T2}
Train demand model d̂_t = f(x_t) with {D_T1, X_T1}
Forecast D̂_T2 = f(X_T2) and calculate the mean error ξ_Y
X_T^c = φ and X_T^s = φ
while r ⩽ R do
    Randomly arrange the sequence of L groups X_T = {X_T^1, X_T^2, ··· , X_T^L}
    for each l in 1 : L do
        Train demand model d̂_t = f_l(x_t) with {D_T1, X_T1^c, X_T1^{l+1}, ··· , X_T1^L}
        Forecast D̂_T2 = f_l(X_T2^c, X_T2^{l+1}, ··· , X_T2^L
        Calculate the forecasting error ξ_N
        if ξ_Y < ξ_N then
            X_T^c = {X_T^c, X_T^l}
        else
            ξ_Y = ξ_N
        end if
    end for
    Update X_T^s according to X_T^c
    r = r + 1
end while
Put {D_T, X_T^s} into the Boruta algorithm to select relevant features X_T^r
```

The second stage of the two-stage feature selection method is feature selection stage with the Boruta algorithm, which aims to select the individual features related to demand. We pool all the $m^s (m^s < m)$ features in the remaining feature groups X_T^s and select the most relevant features in this pool with the Boruta algorithm.

The time complexity of two-stage algorithm is $O(L)+O(m^s \times N)$, where $O(L)$ and $O(m^s \times N)$ are the complexity of the first and second stage, respectively. And $O(L)$ can be regard as a small constant compares with $O(m^s \times N)$. Thus, this

two-stage algorithm significantly reduces the computational time compared with the one-stage Boruta algorithm.

3.3 Forecast

Random forest (RF) achieves a satisfactory prediction accuracy over a wide range of applications (Lohrmann and Luukka 2019, Mueller 2020, Yıldırım et al. 2021) and fits unknown nonlinear and complex interactions of features with manageable computational complexity (Biau and Scornet 2016). Therefore, we apply RF in this study. Random forest is an ensemble method, which combines predictions from several classification and regression trees (Breiman et al. 1984, Breiman 2001). The "random" are reflected by the following two aspects. First, samples are randomly extracted from the training dataset to train regression trees. Second, features are randomly selected to decide how the leaf nodes split. Forest means the combination of numerous tree models. Let K be the number of regression trees and $g_{kx_t}, k = 1, 2, \cdots, K$ represent the demand model of tree index set of training periods, which is trained by the random sample set.

The predicted demand by random forest is the average of K trees:

$$\widehat{d}_t = \frac{\sum_{k=1}^{K} g_k(x_t)}{K} \tag{1}$$

3.4 Evaluation

Regarding the forecasts accuracy measurements, mean absolute error (MAE) and mean absolute percentage error (MAPE) are two classical measurements. However, the downsides for these two measurements are that MAE is scale-dependent and MAPE is infinite when actual demand is zero (Kim and Kim 2016). To solve this issue, some other measurements are proposed, such as the symmetric mean absolute percentage error (SMAPE, (Makridakis 1993)), the mean absolute scaled error (MASE, (Hyndman and Koehler 2006)), the mean arctangent absolute percentage error (MAAPE, (Kim and Kim 2016)). Different from MAPE, the divisor of SMAPE is half of the sum of the actual and forecast values. This study follows Makridakis (1993) to use the adjust mean absolute percentage error (AMAPE) to measure the errors of forecasting due to its advantage of reducing the impact of outliers (extremely large percentage errors). The range of AMAPE is from 0 to 1.

$$AMAPE = \frac{\sum_{t=1}^{N} |\frac{d_t - \widehat{d}_t}{d_t + \widehat{d}_t}|}{N} \tag{2}$$

where N is the number of samples.

4. Data

We collect data from one of the largest online retailing platforms in China. We obtain data about customer orders, inventory history, and detailed order fulfillment logistics of hundreds of merchants. The dataset contains 13.73 billion order records from January 1, 2017 to July 31, 2017 (211 days), with 537 merchants and 272,339 SKUs. The total number of orders of each merchant during this time period can be found in Figure 1.

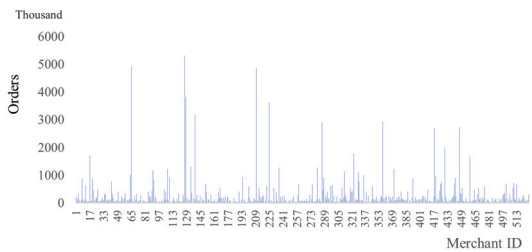

Figure 1 Distribution of Total Number of Orders by Merchant

Our target is to forecast the demand on item-level. However, because the original dataset is based on each order, we do not have the item order information directly. Thus, we extract item order information from its corresponding order and then summarize item order information (including price and demand) each day. By matching each item with its corresponding order, we then collect the related information for each item including page views,

ratings, and logistics. We train the data and test models based on the sales period of each item, which is defined as the time period starting from the first day when the demand volume is non-zero, and ending on the last day when the demand volume is non-zero. This avoids the situation that we analyze the demand of an item before it is on sale or after it has exited the market.

We summarize the demand volume of each item from each merchant, and thus obtain the total demand volumes (i.e., herein after "TSV") of that merchant. As can be found from Figure 2, the distribution of TSV is highly right skewed: the largest 30 merchants contribute to nearly 60 % of the TSV of all of the merchants. Therefore, for simplicity and without loss of generality, we focus on the demand forecasting for the largest 30 e-merchants in this study.

The total items sold by these 30 merchants are 43,783. Among these items, we choose 21,111 items with 109 million sales observations that were sold for more than 100 days. The number of items sold by each merchant can be found from Figure 3. In addition, the ratio of no-sale days and sales period among these items are shown in Figure 4 and Figure 5, respectively, which shows that most of the items are available to sell during our entire study period (i.e. approximately 200 days). Around 5% of the items contribute nearly 80% of the total demand volumes, as can be found in Figure 6.

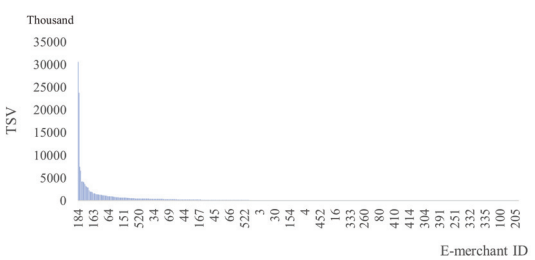

Figure 2 Distribution of TSV by Merchant

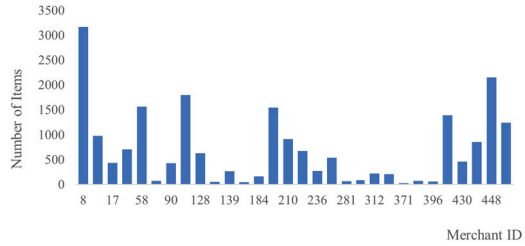

Figure 3 Distribution of Number of Items Sold by Merchants

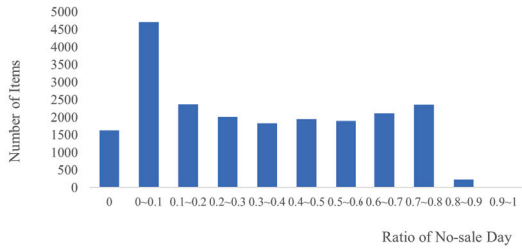

Figure 4 Distribution of Ratio of Days with No Demand

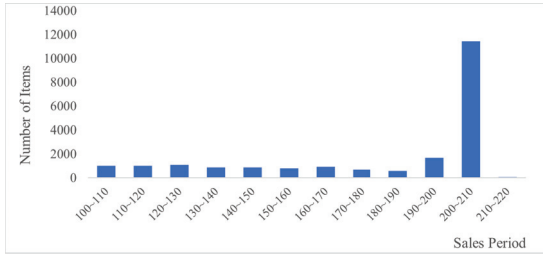

Figure 5 Distribution of Item Sales Period

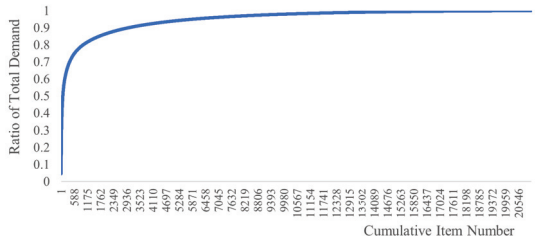

Figure 6 Distribution of Item Demand

5. Results

We extract more than 300 features and then divide them into 27 groups. In the first stage, 12 groups are removed. Then, on average, 10 important features are left after Boruta. To better understand the relative importance of features in demand forecasting, we calculate the frequency of selection of each feature during the demand forecast of 21,111 items and present the top 15 features in a descending order in Table 1. Table 1 shows that the most frequently

Table 1 Rank of Feature Selection Frequencies

Features	Category	Frequency
Price today	Price	16,219
Price difference between today and the past 1 day	Price	10,388
Page views on APP in the past 1 day	Page view	10,097
Number of unique visitors on APP in the past 1 day	Page view	10,044
Average page views on APP in the past 3 days	Page view	9,479
Cumulative page views on APP in the past 3 days	Page view	9,439
Average unique visitors on APP in the past 3 days	Page view	9,218
Cumulative unique visitors on APP in the past 3 days	Page view	9,201
Number of SKUs in SKU's category in the past 1 day	Competition	7,234
Number of unique visitors of SKU's subcategory on APP in the past 1 day	Page view	6,792
Cumulative page views on APP in the past 7 days	Page view	6,501
Average page views on APP in the past 7 days	Page view	6,465
Page views of SKU's subcategory on APP in the past 1 day	Page view	6,377
Average number of unique visitors on APP in the past 7 days	Page view	6,286
Cumulative number of unique visitors on APP in the past 7 days	Page view	6,276

selected features are 2 price-related features and 12 page view-related features. This is expected because price directly affects demand. In addition, we find that among the top 15 features, 12 features are page view related. Page view of an SKU is the prerequisite of purchase, and a higher number of page view and unique visitors represent a larger buyer pool, which will eventually affect the demand. In addition, the increased number of page views and unique visitors may reflect the presence of promotion, which has been recognized as an important factor affecting demand. Moreover, all of the most frequent chosen page views and unique visitor features are on APP, which is as expected considering most buyers use APPs to make the purchases.

After identifying the important features, we use RF to train demand model. During the training process, we utilize the sliding window strategy to update the model, which enables that the daily demand forecasts is based on the most recent data. In detail, for each forecast, the dataset we used is the most recent 160 days, and the RF model is continuously iterated for forecast.

Based on the above procedure, we test the forecast accuracy of the RF algorithm with a two-stage feature selection algorithm (two-stage RF), and compare it with three common algorithms: ARIMA, RF, and RF with one-stage feature selection (one-stage RF). The numerical results in Table 2 show that our proposed two-stage RF delivers 11.58 %, 5.81% and 3.68% forecast accuracy improvement, compared with ARIMA, Random Forecast, and Random Forecast with one-stage feature selection approach, respectively. The underlying reasons for these improvement achievements are elaborated as follows. First, compared with ARIMA, applying the RF algorithm with high-dimensional features improves the forecast accuracy by 5.77 %. This suggests that understanding how the various factors affect the e-commerce demand and modeling their nonlinear interactive relationship through machine learning techniques are important. Second, compared with RF, the RF with one-stage feature selection approach further improves the forecast accuracy by 2.13 %. This shows that it is critical to identify the important features that can avoid the noisy information deteriorating the forecast model performance. Third, the forecast accuracy is improved by 3.68 % using

Table 2 Forecast Accuracy of Various Algorithms

Algorithm	AMAPE
ARIMA	39.75%
RF	33.98%
One-stage RF	31.85%
Two-stage RF	28.17%

our proposed RF-based forecasting framework with a two-stage feature selection algorithm compared with RF with one-stage feature selection. This verifies that our proposed feature selection approach enhances both the forecast accuracy and the computational efficiency.

In the following section, we show the details of the forecast results for item-, merchant-, and category-levels, respectively, which can shed more managerial insights for the operations management of e-commerce.

5.1 Results for Item Level

The investigation of forecast accuracy from the item level includes different demand amount, sales period, and no-sale day's percentage, and the results show that feature selection can significantly improve the forecast accuracy. The forecast accuracy with different demand amount is presented in Table 3, and it is visually shown in Figure 7. From Table 3 and Figure 7, we can find that as total demand increases, AMAPE is reverse U-shaped. Compared to regular items with intermediate demand (i.e., 101-100,000), forecast accuracy is relatively high when the demands are sufficiently low (below 100) or high (above 100,000). The possible reason is that when the demand is low, we have a larger proportion of no-sale days. It is likely to have a zero-demand forecast and meanwhile, the actual sale is also zero, which significantly enhances the forecast accuracy.

In addition, Table 3 and Figure 7 also display the relatively high forecast accuracy of top-selling items (more than 100,000). The forecast accuracy of best sellers can be affected by different factors. On one hand, when the total demand is sufficiently high, we have enough information for the machine learning, which can increase the accuracy. On the other hand, the best-sellers are usually daily consumables, and most of the demand of these products comes from the regular or planned purchase behavior. The sporadic demand that have no relation to demand pattern and features may account for a low proportion in total demand, which can positively affect the forecast accuracy. Based on our results, we find that the forecast accuracy of top-selling items tend to be relatively high, which indicates that high number of total demand is beneficial for the demand forecast.

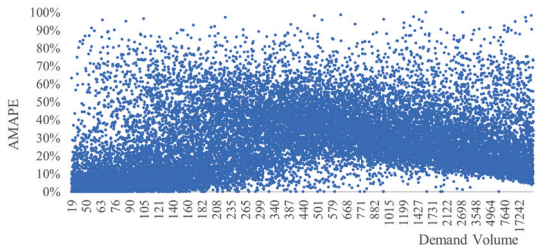

Figure 7 AMAPE with Different Demand Volume

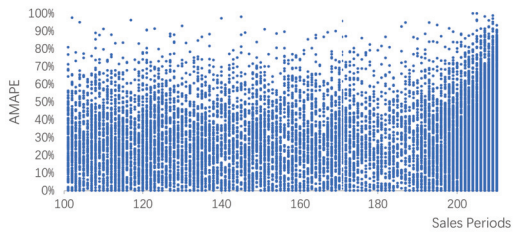

Figure 8 AMAPE with Different Sales Periods

Table 4 reveals that the forecast accuracy with different sales periods, which is visually shown in Figure 8. From Table 4 and Figure 8, we can find that forecast accuracy remains relatively stable for different sales periods. Specifically, AMAPE is the largest (30.12%) when the sale period is the longest (201-220 days). Because of the minor increase of AMAPE with the longer sales period, as found in Table 4, we can find a sales period of approximately 100 days can provide enough information for the learning. Longer sales periods increase the fluctuation of demand, which may bring unnecessary obstacles to the forecast.

Table 3 Forecast Accuracy with Different Total Amount of Demand

Total Demand Volume	Number of Items	RF AMAPE	Two-stage RF AMAPE	Improvement
0-100	3,101	19.23%	12.37%	6.87%
101-1,000	11,343	40.33%	32.06%	8.27%
1,001-10,000	5,653	30.96%	9.81%	1.15%
10,001-100,000	880	25.18%	24.17%	1.01%
> 100,000	134	23.88%	22.58%	1.30%

Table 4 Forecast Accuracy with Different Sales Periods

Sales Period	Number of Items	RF AMAPE	Two-stage RF AMAPE	Improvement
100-120	22,037	233.59%	24.67%	8.92%
121-140	21,985	234.70%	26.12%	8.59%
141-160	21,690	234.09%	26.43%	7.66%
161-180	21,608	234.42%	26.72%	7.70 %
181-200	22,277	233.50%	25.60%	7.90 %
201-220	211,514	233.95%	30.12%	3.83 %

Table 5 Forecast Accuracy with Different No-sale Days' Percentage

No-Sale Days' Percentage	Number of Items	RF AMAPE	Two-stage RF AMAPE	Improvement
0-0.2	8,684	28.97 %	28.98 %	–0.01%
0.2-0.4	3,856	45.42 %	38.16%	7.26%
0.4-0.6	3,840	42.01%	29.44%	12.57%
0.6-0.8	4,464	27.76%	17.83%	9.94%
0.8-1	267	20.64%	12.59%	8.06%

Table 5 shows that the forecast accuracy with different no-sale days' percentages, and it is visually shown in Figure 9. From Table 5 and Figure 9, we can find that AMAPE is the smallest when the percentage of no-sale days is the highest. When no-sale percent increases to 80%, AMAPE decreases to 12.59%, which is significantly lower than 38.16% when the no-sale percentage is 20%. The proposed algorithm has a relatively high forecast accuracy when the percentage of no-sale days is relatively high. This is because high no-sale days' percentage indicates a large number of continuous zero, rather than intermitted zero, in terms of the demand. Thus, it is easier to learn and predict the demand when the percentage of no-sale days is relatively high. This is evidenced by the findings from Table 5 that when the proportion of the no-sale days reaches more than 80%, the forecast error is lower.

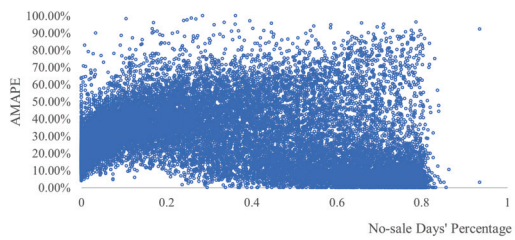

Figure 9 AMAPE with Different No-sale Days' Percentage

5.2 Results for Merchant Level

The evaluation of forecast accuracy from the merchant level includes different number of items and average demand. Figure 10 and Figure 11 display the forecast accuracy with different number and different average demand of items sold by merchants, respectively. The performance of the forecasting model is quite stable across merchants as forecast accuracy is not significantly affected by the number of

items or the average demand of items sold by merchants.

Figure 10 AMAPE with Different Number of Items Sold by Merchant

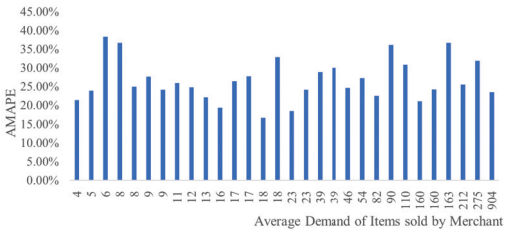

Figure 11 AMAPE with Different Average Demand of Items Sold by Merchant

5.3 Results for Category Level

The results of forecast accuracy from the category level are presented in Figure 12 and Figure 13. We find that AMAPE is higher when the number and average demand of items in a category are lower. A small number of items and low average demand in a category indicate that the category is unpopular. Compared to other items, the demand of unpopular items may exhibit higher variation, which is not conducive to result in a high demand forecast accuracy.

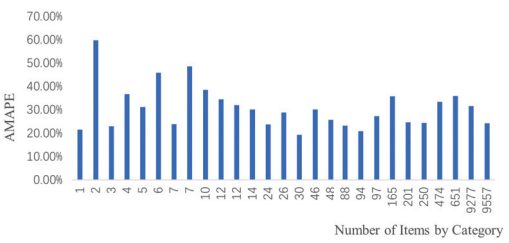

Figure 12 AMAPE with Different Number of Items by Category

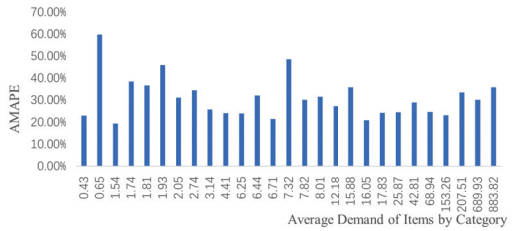

Figure 13 AMAPE with Average Demand of Items by Category

6. Conclusing Remarks

This study forecasts the daily demand of 21,111 items of the top 30 e-merchants from 512 e-merchants on the largest B2C e-commerce platform in China. This study adds values to the previous forecasting studies through several aspects. First, this study builds an extensive set of features and divides them into five categories including historical demand, price, page view, reviews, and competition. Our results show that among all the feature categories, page view is the most important category in predicting demand, which accounts for 12 of the 15 most frequently selected features. Based on our findings, online vendors should pay attention to increase the number of page views and attract more visitors, especially from APPs considering most buyers use smart phones during purchases. Possible strategies include advertisement, promotion, and a better design of the webpage.

Second, we design a two-stage feature selection algorithm and demonstrate that the algorithm is a powerful tool in identifying the appropriate number of effective features from high-dimensional online e-merchant dataset in the online environment. Our results show that the demand forecast based on RF performs significantly better than the traditional time series methods ARIMA and demand forecast based on RF with the two-stage feature selection algorithm has lower forecast error and computational cost than that with the one-stage feature selection algorithm.

Third, we conduct an in-depth analysis of the forecast accuracy from the item-, merchant-

, and category-levels, which provides guidance on e-commerce management such as production, inventory, and sales efforts. For instance, our results show that compared to normal items, the forecast accuracy for best-sellers is higher.

This study has several limitations. First, due to data availability issues, we are not able to include features such as inventory and promotion. Future studies can extend this study through the collection of these features to further improve the demand forecasting accuracy. Second, our study focuses on a B2C e-platform. To test the validity and generalizability of our results, a comparative study examining other online B2B and C2C e-platforms and conducting the corresponding demand forecasting can be interesting.

Acknowledgments

This work has been supported in part by the National Natural Science Foundation of China under Grant Nos. 72172169, 71903024, 91646125 and Program for Innovation Research at the Central University of Finance and Economics. The authors sincerely thank the editors and two anonymous referees for their constructive comments to significantly improve the paper.

References

Abasabadi S, Nematzadeh H, Motameni H, Akbari E (2021). Automatic ensemble feature selection using fast non-dominated sorting. *Information Systems* 100: 101760.

Abolghasemi M, Beh E, Tarr G, Gerlach, R (2020). Demand forecasting in supply chain: The impact of demand volatility in the presence of promotion. *Computers & Industrial Engineering* 142: 106380.

Ali Ö G, Sayın S, Van Woensel T, Fransoo J (2009). SKU demand forecasting in the presence of promotions. *Expert Systems with Applications* 36(10): 12340-12348.

Andersen J, Giversen A, Jensen A H, Larsen R S, Pedersen T B, Skyt J (2000). Analyzing clickstreams using subsessions. *In Proceedings of the 3rd ACM international workshop on Data warehousing and OLAP*. ACM, November, 25-32.

Athanasopoulos G, Hyndman R J, Kourentzes N, Petropoulos F (2017). Forecasting with temporal hierarchies. *European Journal of Operational Research* 262(1): 60-74.

Bauer H H, Falk T, Hammerschmidt M (2006). eTransQual: A transaction process-based approach for capturing service quality in online shopping. *Journal of Business Research* 59(7): 866-875.

Besbes O, Gur Y, Zeevi A (2016). Optimization in online content recommendation services: Beyond click-through rates. *Manufacturing & Service Operations Management* 18(1): 15-33.

Biau G, Scornet E (2016). A random forest guided tour. *Test* 25(2): 197-227.

Breiman L (2001). Random forests. *Machine Learning* 45(1): 5-32.

Breiman L, Friedman J, Stone C J, Olshen R A (1984). *Classification and Regression Trees*, CRC press.

Cantallops A S, Salvi F (2014). New consumer behavior: A review of research on eWOM and hotels. International *Journal of Hospitality Management* 36: 41-51.

Cao P, Zhao N, Wu J (2019). Dynamic pricing with Bayesian demand learning and reference price effect. *European Journal of Operational Research* 279(2): 540-556.

Chandrashekar G, Sahin F (2014). A survey on feature selection methods. *Computers & Electrical Engineering* 40(1): 16-28.

Chen Q, Zhang M, Xue B (2017). Feature selection to improve generalization of genetic programming for high-dimensional symbolic regression. *IEEE Transactions on Evolutionary Computation* 21(5): 792-806.

Chiew K L, Tan C L, Wong K, Yong K S, Tiong W K (2019). A new hybrid ensemble feature selection framework for machine learning-based phishing detection system. *Information Sciences* 484: 153-166.

Choi T M, Hui C L, Liu N, Ng S F, Yu Y (2014). Fast fashion sales forecasting with limited data and time. *Decision Support Systems* 59: 84-92.

Chong A Y L, Ch'ng E, Liu M J, Li B (2017). Predicting consumer product demands via Big Data: The roles of online promotional marketing and online reviews. *International Journal of Production Research* 55(17): 5142-5156.

Chong A Y L, Li B, Ngai E W, Ch'ng E, Lee F (2016). Predicting online product sales via online reviews, sentiments, and promotion strategies: A big data architecture and neural network approach. *International Journal of Operations & Production Management* 36(4): 358-383.

Chou M C, Sim C K, Yuan X M (2020). Policies for inventory models with product returns forecast from past demands and past sales. *Annals of Operations Research* 288: 137-180.

Dai A, Zhang Z, Hou P, Yue J, He S, He Z (2019). Warranty claims forecasting for new products sold with a two-dimensional warranty. *Journal of Systems Science and Systems Engineering* 28(6): 715-730.

Ding Y, Liu J (2021). Joint pricing strategies of multi-product retailer with reference-price and substitution-price effect. *Journal of Data, Information and Management* 3(1): 49-63.

Divakar S, Ratchford B T, Shankar V (2005). Practice prize article - CHAN4CAST : A multichannel, multiregion sales forecasting model and decision support system for consumer packaged goods. *Marketing Science* 24(3): 334-350.

Dong J, Hu Z, Liang C (2017). E-commerce supply chain coordination under demand influenced by historical sales rate. *2017 3rd International Conference on Information Management (ICIM)* 61-71, IEEE.

Fan Z P, Che Y J, Chen Z Y (2017). Product sales forecasting using online reviews and historical sales data: A method combining the Bass model and sentiment analysis. *Journal of Business Research* 74: 90-100.

Ferreira K J, Lee B H A, Simchi-Levi D (2016). Analytics for an online retailer: Demand forecasting and price optimization. *Manufacturing & Service Operations Management* 18(1): 69-88.

Fildes R, Goodwin P, Önkal D (2019). Use and misuse of information in supply chain forecasting of promotion effects. *International Journal of Forecasting* 35(1): 144-156.

Giang N L, Ngan T T, Tuan T M, Phuong H T, Abdel-Basset M, de Macêdo A R L, de Albuquerque V H C (2019). Novel incremental algorithms for attribute reduction from dynamic decision tables using hybrid filter–wrapper with fuzzy partition distance. *IEEE Transactions on Fuzzy Systems* 28(5): 858-873.

Goltsos T E, Syntetos A A, van der Laan E (2019). Forecasting for remanufacturing: The effects of serialization. *Journal of Operations Management* 65(5): 447-467.

Got, A, Moussaoui A, Zouache D (2021). Hybrid filter-wrapper feature selection using Whale Optimization Algorithm: A Multi-Objective approach. *Expert Systems with Applications* 183: 115312.

Guyon I, Elisseeff A (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research* 3: 1157-1182.

Guyon I, Weston J, Barnhill S, Vapnik V (2002). Gene selection for cancer classification using support vector machines. *Machine Learning* 46(1): 389-422.

Hanna R C, Lemon K N, Smith G E(2019). Is transparency a good thing? How online price transparency and variability can benefit firms and influence consumer decision making. *Business Horizons* 62(2): 227-236.

He J, Wang X, Vandenbosch M B, Nault B R (2020). Revealed preference in online reviews: Purchase verification in the tablet market. *Decision Support Systems* 132: 113281.

Huang G, Liu L (2006). Supply chain decision-making and coordination under price-dependent demand. *Journal of Systems Science and Systems Engineering* 15(3): 330-339.

Huang T, Fildes R, Soopramanien D (2014). The value of competitive information in forecasting FMCG retail product sales and the variable selection problem. *European Journal of Operational Research* 237(2): 738-748.

Hyndman R J, Koehler A B (2006). Another look at measures of forecast accuracy. *International Journal of Forecasting* 22: 679-688.

Hyndman R J, Koehler A B, Snyder R D, Grose S (2002). A state space framework for automatic forecasting using exponential smoothing methods. *International Journal of Forecasting* 18(3): 439-454.

Jiménez-Cordero A, Morales J M, Pineda S (2021). A novel embedded min-max approach for feature selection in nonlinear Support Vector Machine classification. *European Journal of Operational Research* 293(1): 24-35.

Kamakura W A, Kang W (2007). Chain-wide and store-level analysis for cross-category management. *Journal of Retailing* 83(2): 159-170.

Kim J, Kang J, Sohn M (2021). Ensemble learning-based filter-centric hybrid feature selection framework for high-dimensional imbalanced data. *Knowledge-Based Systems* 220: 106901.

Kim S, Kim H (2016). A new metric of absolute percentage error for intermittent demand forecasts. *International Journal of Forecasting* 32(3): 669-679.

Koehn D, Lessmann S, Schaal M (2020). Predicting online shopping behaviour from clickstream data using deep learning. *Expert Systems with Applications* 150: 113342.

Korobilis D (2017). Quantile regression forecasts of inflation under model uncertainty. *International Journal of Forecasting* 33(1): 11-20.

Kursa M B, Rudnicki W R (2010). Feature selection with the Boruta package. *Journal of Statistical Software* 36(11): 1-13.

Lee L, Charles V (2021). The impact of consumers' perceptions regarding the ethics of online retailers and promotional strategy on their repurchase intention. *International Journal of Information Management* 57: 102264.

Leung K H, Mo D Y, Ho G T, Wu C H, Huang G Q (2020). Modelling near-real-time order arrival demand in e-commerce context: A machine learning predictive methodology. *Industrial Management & Data Systems* 120(6): 1149-1174.

Li C, Lim A (2018). A greedy aggregation–decomposition method for intermittent demand forecasting in fashion retailing. *European Journal of Operational Research* 269(3): 860-869.

Li J, Manry M T, Narasimha P L, Yu C (2006). Feature selection using a piecewise linear network. *IEEE Transactions on Neural Networks* 17(5): 1101-1115.

Lohrmann C, Luukka P (2019). Classification of intraday S&P500 returns with a Random Forest. *International Journal of Forecasting* 35(1): 390-407.

Lu L, Gou Q, Tang W, Zhang J (2016). Joint pricing and advertising strategy with reference price effect. *International Journal of Production Research* 54(17): 5250-5270.

Ma S, Fildes R, Huang T (2016). Demand forecasting with high dimensional data: The case of SKU retail sales forecasting with intra-and inter-category promotional information. *European Journal of Operational Research* 249(1): 245-257.

Makridakis S (1993). Accuracy measures: Theoretical and practical concerns. *International journal of Forecasting* 9(4): 527-529.

Maldonado S, Pérez J, Bravo C (2017). Cost-based feature selection for support vector machines: An application in credit scoring. *European Journal of Operational Research* 261(2): 656-665.

Maldonado S, Weber R, Basak J (2011). Simultaneous feature selection and classification using kernel-penalized support vector machines. *Information Sciences* 181(1): 115-128.

Martínez A, Schmuck C, Pereverzyev Jr S, Pirker C, Haltmeier M (2020). A machine learning framework for customer purchase prediction in the non-contractual setting. *European Journal of Operational Research* 281(3): 588-596.

Mueller S Q (2020). Pre-and within-season attendance forecasting in Major League Baseball: A random forest approach. *Applied Economics* 52(41): 4512-4528.

Nakariyakul S, Casasent D P (2009). An improvement on floating search algorithms for feature subset selection. *Pattern Recognition* 42(9): 1932-1940.

Nakariyakul S (2018). High-dimensional hybrid feature selection using interaction information-guided search. *Knowledge-Based Systems* 145, 59-66.

Narayanan A, Sahin F, Robinson E P (2019). Demand and order-fulfillment planning: The impact of point-of-sale data, retailer orders and distribution center orders on forecast accuracy. *Journal of Operations Management* 65(5): 468-486.

Navarro F F G, Muñoz L A B (2009). Gene subset selection in microarray data using entropic filtering for cancer classification. *Expert Systems* 26(1): 113-124.

Neto J Q F, Bloemhof J, Corbett C (2016). Market prices of remanufactured, used and new items: Evidence from eBay. *International Journal of Production Economics* 171: 371-380.

Nikolopoulos K (2021). We need to talk about intermittent demand forecasting. *European Journal of Operational Research* 291(2): 549-559.

Omuya E O, Okeyo G O, Kimwele M W (2021). Feature selection for classification using principal component analysis and information gain. *Expert Systems with Applications* 174, 114765.

Ot A, Ttn B, Sm C (2021). A novel wrapper-based feature subset selection method using modified binary differential evolution algorithm. *Information Sciences* 565, 278-305.

Pang G, Casalin F, Papagiannidis S, Muyldermans L, Tse Y K (2015). Price determinants for remanufactured electronic products: A case study on eBay UK. *International Journal of Production Research* 53(2): 572-589.

Pannakkong W, Sriboonchitta S, Huynh V N (2018). An ensemble model of arima and ann with restricted boltzmann machine based on decomposition of discrete wavelet transform for time series forecasting. *Journal of Systems Science and Systems Engineering* 27(5): 690-708.

Peng H, Long F, Ding C (2005). Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 27(8): 1226-1238.

Petropoulos F, Hyndman R J, Bergmeir C (2018). Exploring the sources of uncertainty: Why does bagging for time series forecasting work? *European Journal of Operational Research* 268(2): 545-554.

Ramanathan U, Muyldermans L (2010). Identifying demand factors for promotional planning and forecasting: A case of a soft drink company in the UK. *International journal of production economics* 128(2): 538-545.

Reunanen J (2003). Overfitting in making comparisons between variable selection methods. *Journal of Machine Learning Research* 3(Mar): 1371-1382.

Subramanian R, Subramanyam R (2012). Key factors in the market for remanufactured products. *Manufacturing & Service Operations Management* 14(2): 315-326.

Sun L, Zheng X, Jin Y, Jiang M, Wang H (2019). Estimating promotion effects using big data: A partially profiled LASSO model with endogeneity correction. *Decision Sciences* 50(4): 816-846.

Tang L, Sun L, Guo C, Zuo Y, Zhang Z (2021). A Simulation Research Towards Better Leverage of Sales Ranking. *Journal of Systems Science and Systems Engineering* 30(1): 105-122.

Trapero J R, Kourentzes N, Fildes R (2015). On the identification of sales forecasting models in the presence of promotions. *Journal of the operational Research Society,* 66(2): 299-307.

Van Donselaar K H, Peters J, de Jong A, Broekmeulen R A(2016). Analysis and forecasting of demand during

promotions for perishable items. *International Journal of Production Economics* 172: 65-75.

Wang P, Du R, Hu Q (2020). How to promote sales: Discount promotion or coupon promotion? *Journal of Systems Science and Systems Engineering* 29(9): 381-399.

Wu M, Ma L, Xue W (2020). Order timing for manufacturers with spot purchasing price uncertainty and demand information updating. *Journal of Systems Science and Systems Engineering* 29(6): 631-654.

Wu W, Liu M, Liu Q, Shen W (2016). A quantum multi-agent based neural network model for failure prediction. *Journal of Systems Science and Systems Engineering* 25(2): 210-228.

Xie G, Qian Y, Wang S (2021). Forecasting Chinese cruise tourism demand with big data: An optimized machine learning approach. *Tourism Management* 82: 104208.

Xu X, Zeng S, He Y (2017). The influence of e-services on customer online purchasing behavior toward remanufactured products. *International Journal of Production Economics* 187: 113-125.

Yan T, Sun B (2011). A study on statical and dynamical characteristics model of e-commerce competitive environment. *2011 International Conference on Business Management and Electronic Information* IEEE 4: 573-580.

Ye Q, Law R, Gu B (2009). The impact of online user reviews on hotel room sales. *International Journal of Hospitality Management* 28(1): 180-182.

Yeo J, Hwang S W, Koh E, Lipka N (2018). Conversion prediction from clickstream: Modeling market prediction and customer predictability. *IEEE Transactions on Knowledge and Data Engineering* 32(2): 246-259.

Yıldırım M, Okay F Y, Özdemir S (2021). Big data analytics for default prediction using graph theory. *Expert Systems with Applications* 176: 114840.

Yu H, Chen X, Li Z, Zhang G, Liu P, Yang J, Yang Y (2019). Taxi-based mobility demand formulation and prediction using conditional generative adversarial network-driven learning approaches. *IEEE Transactions on Intelligent Transportation Systems* 20(10): 3888-3899.

Zhu F, Zhang X (2010). Impact of online consumer reviews on sales: The moderating role of product and consumer characteristics. *Journal of Marketing* 74(2): 133-148.

Hongyan Dai is currently working as a professor in Business School, Central University of Finance and Economics, Beijing, China. Dr. Dai received her BSc from Beijing University of Posts and Telecommunications in 2004, MSc from Tsinghua University in 2006, and PhD from Hong Kong University of Science and Technology in 2011. Dr. Dai's current research interest includes sharing economy, O2O logistics network design, data-driven optimization etc. She published over 20 papers in reputational international journals, including European Journal of Operational Research, International Journal of Production Economics, International Journal of Production Research, and Annals of Operations Research. She is the Principal Investigator for several National and Provincial level projects, including the Major Research Plan of the National Natural Science Foundation of China; and several industry projects, such as Jingdong-to-home, State Grid, and Siemens. She serves as the research fellow of China Society of Logistics, review specialist of National Natural Science Foundation of China, and referee of over ten international journals.

Qin Xiao is a PhD candidate in Business School, Central University of Finance and Economics, Beijing, China. She received her bachelor degree from Central University of Finance and Economics in 2016. Her current research interest includes O2O logistics network design, data-driven optimization etc. She published several papers in international journals, such as International Journal of Production Economics. She has participated in National and Provincial level projects, including the Major Research Plan of the National Natural Science Foundation of China; and industry projects, such as Jingdong-to-home.

Nina Yan is currently working as a full professor in Business School, Central University of Finance and Economics, China. She received her PhD in management science from Northeastern University in March 2007. Her current research interests include supply chain finance, operations management, platform economics, etc. She has published over twenty papers in such journals, including Decision Sciences, Decision Support Systems, European Journal of Operational Research, International Journal of Production Economics, International Journal of Production Research, Journal of Business Research, and Omega.

Xun Xu holds a PhD in operations management from Washington State University. He is currently an associate professor in the Department of Management, Operations, and Marketing in College of Business Administration at California State University, Stanislaus in the United States. His research interests include service operations management, supply chain management and coordination, sustainability, e-commerce, data and text mining, and hospitality and tourism management. He has published over fifty papers in such journals as Annals of Tourism Research, Computers and Industrial Engineering, Decision Sciences, Decision Support Systems, European Journal of Operational Research, Journal of Business Research, Journal of the Operational Research Society, Journal of Travel Research, International Journal of Hospitality Management, International Journal of Contemporary Hospitality Management, International Journal of Information Management, International Journal of Production Economics, and International Journal of Production Research.

Tingting Tong is an associate professor in Dongbei University of Finance and Economics, China. She received her PhD in Economics from Georgia Institute of Technology in August 2016. Her research focus on operations manage-

ment, labor economics, and applied econometrics. Her articles have been published by journals such as China Economic Review, Decision Sciences, Decision Support Systems, International Journal of Production Economics, Journal of Business Research, and Journal of Transport Geography.