Journal of Computers Vol. 32 No. 1, 2021, pp. 212-225
doi:10.3966/199115992021023201018

Sales Forecasting of Promotion Activities Based on the Cross-Industry Standard Process for Data Mining of E-commerce Promotional Information and Support Vector Regression

Yiman Zhang*

College of Oujiang, Wenzhou University, Wenzhou, China
714261545@qq.com

Received 10 June 2019; Revised 21 August 2019; Accepted 6 January 2020

Abstract. The purpose of the study is to analyze and forecast the effect of e-commerce promotional activities and sales volume, and improve the scientific and intelligent degree of enterprise management and decision-making through data mining and analysis. The short-term effect of e-commerce promotional activities was analyzed. In terms of effect of sales activities, the sales volume of goods during activities will be significantly increased. Meantime, as the activity time goes on, its ability to stimulate sales is gradually weakening, while the sales of goods before and after the activity will decline to a certain extent. The passenger flow will increase significantly during the activity period, while the passenger flow will not change significantly before and after the activity. The decrease of sales is mainly due to the low conversion rate during this period. The results show that the comprehensive prediction model has a certain improvement in prediction accuracy compared with the single support vector machine (SVM) prediction model. Combining with interest, effective association rules can be found better, and the decision-making model considering loss of profits before and after activities is helpful for enterprises to make better management decisions. Thus, under the guidance of current methodology, excellent prediction results and rich management inspiration can be acquired, which proves that the forecasting process proposed has great application value.

Keywords: E-commerce promotion, sales forecasting, data mining, data-based decision-making, support vector regression

1 Introduction

E-commerce is a business activity centered on information exchange technology and is based on the commodity exchange. It can also be understood as an activity of conducting transactions and related services by means of electronic transactions on the Internet, intranets, and value-added networks, which is the electronic, online, and informationized form of each section of the traditional transactions [1]. Internet-mediated business practices are all in the category of E-commerce. The participants generally include Agents, Businesses, and Consumers (ABC). There are 8 common models, which are respectively Business-to-Business (B2B), Business-to-Consumer (B2C), Consumer-to-Consumer (C2C), Business-to-Government (B2G), Online To Offline (O2O), Business To Family (B2F), Provide to Demand (P2D), Online to Partner (O2P) [2], among which the B2B and B2C models are the most commonly seen E-commerce models. With the development of E-commerce, consumer-to-business (C2B) has also begun to rise and is considered to be the future of E-commerce. At present, with the rapid development of Internet technology and the rise of E-commerce shopping models, traditional enterprises have begun to vigorously develop online business. Well-known E-commerce platforms and enterprises at home and abroad include Jingdong, Alibaba, NetEase, Amazon, Sephora, Farfetch, etc. Given the new business

* Corresponding Author

model of "Internet + retail", the competition among companies is becoming increasingly fierce. In order to increase sales, many companies would participate in or organize various promotional activities from time to time. At the same time, with the rapid development of enterprise information, the data accumulation of enterprises has also reached a certain scale.

However, since 2016, the growth rate of the major E-commerce websites represented by Jingdong and Alibaba has decreased from 40% to 50% compared with the previous 3-digit growth rate [3], which indicates that the growth rate of E-commerce begins to slow down. The reason is that E-commerce has begun to enter the new stage of the fine business and exploration of new business fields. The intensive work of the existing business is mainly the promotion of marketing methods and the optimization of back-end service capabilities such as supply chain, logistics, and distribution, and the use of E-commerce data could help empower the entire value chain. The new business is mainly the O2O model, cross-border E-commerce, rural E-commerce, and smart supply chain represented by the Hema Fresh of Alibaba Group and the Seven Fresh of Jingdong Group. In terms of the E-commerce industry, companies are paying much attention to customer identification and marketing. Sales forecasting is an important task of E-commerce and has an important impact on making informed business decisions. It can help E-commerce companies manage labor, cash stream, and resources, and optimize the supply chain of manufacturers, etc. [4]. Traditional sales forecasting techniques rely primarily on historical sales data and are limited in predicting future sales and their accuracy. Since 2012, "Big Data" and "Data Mining" technologies have been widely concerned by the business community [5]. The famous "Beer and Diaper" theory and Amazon's personalized recommendation technology have become popular application cases. The practical value of related technologies has been proven by numerous studies and practices. But at that time, there were only theories and was in lack of practices in China. The reason is mainly the gap in data mining talents and the lack of corporate data. By 2017, almost all large-scale enterprises and companies have contributed a relatively complete enterprise database, which uses information management systems provided by SAP, Oracle, etc. [6]. It can be said that the current data shortage problem has been basically solved. Although the data quality needs to be improved, the data mining talents have already had a lot of room to play. Therefore, from the perspective of corporate profitability, based on the support vector regression (SVR) for comprehensive modeling analysis, the benefits of the commodities in the promotion activities, as well as before and after the promotion activities, have been predicted by the model. The loss and the commission fee charged by the E-commerce platform are used for predictive analysis so that the company can arrange its next stage of production, logistics, promotion activity, and new business expansion, etc. The results have shown that the comprehensive forecasting model has a certain improvement in prediction accuracy compared with the single SVR prediction model, and it helps the E-commerce companies improve the scientific degree and intelligence degree of their management and decision-making, which is of certain theoretical and practical significance.

2 Literature Review

2.1 E-commerce

Because E-commerce has accumulated a large amount of data, it has unique advantages in forecasting and has great similarities with traditional retail. However, there are many differences between E-commerce and physical retail, such as the lack of access to physical objects, the faster dissemination of information on the Internet, and the difference between E-commerce and traditional retail model. In 2017, Sadasivam et al. believed that accurate and reliable sales forecasting is the key to E-commerce business and proposed a systematic pre-processing framework to overcome the challenges in E-commerce environment [8]. In 2017, Tricahyadinata and Za proposed a new method to automatically learn effective features from structured data using a convolutional neural network (CNN). In 2016, Gallego et al. selected GDP, per capita consumption expenditure of urban residents, total retail sales of consumer goods, and the number of netizens as economic prediction indicators [9]. According to the grey correlation degree, they selected closely related indicators and established a multi-variable grey prediction model [10]. In 2017, Kudyba and Lawrence calculated the reliable forecasting quantity of clothing E-commerce by iteration fitting of the grey model and analyzed it with the example of clothing E-commerce in a province [11]. In 2016, See-To and Ngai proposed a data mining framework for sales forecasting based on online user behavior data. Under this framework, the relationship between sales data and online user

behavior data was modeled, and the optimal lag of online user behavior data in sales forecasting was determined [12]. In 2016, Tolstoy et al. used a regression model to forecast and test the corresponding comprehensive index, and finally got the daily sales forecast results of different categories of goods [13]. In 2016, Aulkemeier et al. established the LS-SVM online prediction model based on time factor rejection samples, and input the observation data into the network sales example [14]. In 2016, Antipov and Peckryshevskaya took manufacturing E-commerce as an example to analyze it and formed a conceptual model of E-commerce sales service [15].

2.2 Data mining

With the development of computers and the internet, the amount of data is increasing. Data mining is the process of finding anomalies, patterns and correlations within large data sets to predict outcomes [15]. By using a broad range of techniques, data mining techniques could be used to increase revenues, cut costs, improve customer relationships, and reduce risks, etc. The process of digging through data to discover hidden connections and predict future trends has a long history. Sometimes referred to as "knowledge discovery in databases"; the term "data mining" wasn't coined until the 1990s. The foundation of data mining comprises three intertwined scientific disciplines, i.e. statistics, artificial intelligence, and machine learning [16]. Data mining technology keeps evolving to keep pace with the limitless potential of big data and affordable computing power. Over the last decade, advances in processing power and speed have enabled human to move beyond manual, tedious, and time-consuming practices to quick, easy, and automated data analysis. The more complex the data sets collected, the more potential there is to uncover relevant insights [17]. Retailers, banks, manufacturers, telecommunications providers, and insurers, among others, are using data mining to discover relationships among everything from price optimization, promotions, and demographics to how the economy, risk, competition and social media are affecting their business models, revenues, operations and customer relationships [18]. Large customer databases hold hidden customer insight that can help retailers and companies improve relationships, optimize marketing campaigns, and forecast sales. Through more accurate data models, retailers and companies can offer more targeted campaigns and find the offer that makes the biggest impact on the customer [19]. Mining large reservoirs of data in online transactions involve committing to key processes and technologies. In order to extract value from vast data stores and change the way decisions are made, many E-commerce companies have turned to advanced data mining techniques along with real-time analytical and data processing capabilities [20]. In 2019, Ghimire et al. showed that, in order to obtain the relevant input characteristics of the model, the MODIS variables were screened by particle swarm optimization (PSO) algorithm. Secondly, the sensitivity analysis of all selected variables was used to determine their relative role in predicting GSR. In order to solve the related non-stationary problem, before merging the variables selected by PSO into SVR, the maximum overlapping discrete wavelet transform is used to decompose them, and then a three-phase PSO-W-SVR hybrid model is constructed, in which the hyperparameters are obtained by evolution (i.e. PSO and genetic algorithm) and grid search method.

2.3 E-marketing and Promotions

The research of network marketing can be divided into the study from the perspective of consumer buying behavior, the study from the impact of enterprise marketing decisions on consumers, or the study of a comprehensive theory. In 2017, Kudyba and Lawrence improved the key prior association rules and K-means clustering core algorithm on the basis of studying several Web mining algorithms in an E-commerce environment and verified the effectiveness of the improved algorithm through experiments [21]. In 2017, Shaikh et al. analyzed and summarized the current research status and development trend of E-commerce system framework, and pointed out the existing problems of traditional E-commerce Web mining system [22]. In 2017, Hernández et al. made an in-depth and simple analysis of several marketing modes in E-commerce and made a clear interpretation of creative online marketing modes such as "one-second price reduction", "one-dollar auction", "buy-on-behalf" and "collective purchase" in E-commerce [23]. In 2018, Xue and Jarvis analyzed the current situation of the development of E-commerce industry under the background of big data and proposed a new E-commerce marketing model based on big data analysis and processing [24]. In 2016, García et al. introduced the concept of Web mining, described the process of Web data mining in detail, summarized the relationship between Web

data mining and E-commerce, and how to apply Web mining technology in E-commerce [25]. In 2017, Escobar-Rodríguez and Bonsón-Fernández introduced the application of ontology technology in the construction of a knowledge base of the E-commerce system of direct clothing network. The experimental results show that the application technology has high application efficiency [26]. In 2017, Kim and Peterson used SWOT analysis model to understand the basic characteristics of international e-tourism marketing and put forward the strategy of seeking a balance between environmental data and digital data [27]. In 2017, according to the needs and current situation of enterprises, Hong designed an enterprise network commodity marketing platform, which can realize the functions of commodity display, customer online shopping, commodity information release and so on [28]. In 2017, Hajli and Featherman discussed the role of online social networks in E-marketing and their relationship with E-commerce [29]. In 2017, Orman analyzed the causal relationship among direct network externalities, indirect network externalities, compatibility and other factors using system dynamics theory from the perspective of consumer choice, and established the system flow chart and system dynamics model of B2C E-commerce platform competition [30]. In 2018, the research of Han and Bian showed that the intelligent model based on SVM combined with PSO (PSO-SVM) technology is used to predict the sales of goods. The accuracy and reliability of the proposed model are evaluated by input variables of the proposed PSO-SVM model assisted with grey correlation analysis through 34 data sets collected in public literature, and compared with PSO-BP neural network - an empirical method of e-commerce promotion.

3 Method

Under the environment of online shopping, consumers can easily and cheaply grasp all kinds of information through the network, and the cost of transforming other businesses is much lower than in the traditional market. Therefore, it is necessary to study online consumer behavior. Only by understanding the motivation and behavior characteristics of online consumers and grasping the factors affecting consumers' purchase decision-making behavior, can enterprises formulate targeted marketing strategies to maximize consumer satisfaction, win customer trust and win in the competition.

Consumers' shopping behavior can be regarded as a black box, which makes purchase decisions under the stimulation of consumers' own needs, business marketing, and external environment. Whether physical stores or online businesses, if they want to attract and motivate consumers' own stores to complete purchase decisions and achieve the purpose of sales, they must first understand the characteristics of target consumers. From the perspective of consumer behavior, the stage of online shopping is the same as that of ordinary consumer behavior. It is possible to describe online shopping with the behavior model of ordinary consumers. Among the current consumer models, the EBM model is the most complete one in the context of consumer-driven business, which includes the five stages shown in Fig. 1.

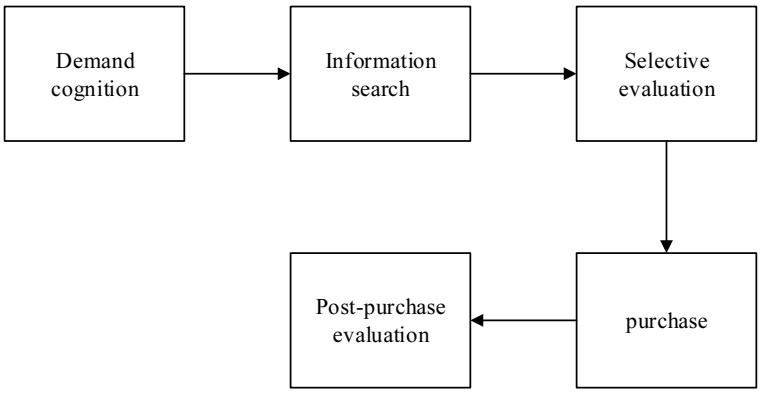

Fig. 1. Consumer shopping behavior

From the perspective of sales, a classical theoretical model is "sales funnel", which means that sales opportunities are declining in the process of purchasing behavior if the shape is expressed as a funnel in graphics. From the perspective of the whole transformation of consumer behavior, merchants use this concept for reference and decompose sales into "sales = browsing volume * conversion rate". Flow

reflects the interaction between enterprises and external demand, and the conversion rate reflects the ability of enterprises to convert consumption intention into purchasing behavior. These two indicators are especially important and widely used in the operation of e-commerce enterprises. By observing the relationship among the three indicators of e-commerce enterprises, people can have a better understanding of the current operational links and the overall operating environment of e-commerce enterprises, as shown in Fig. 2.

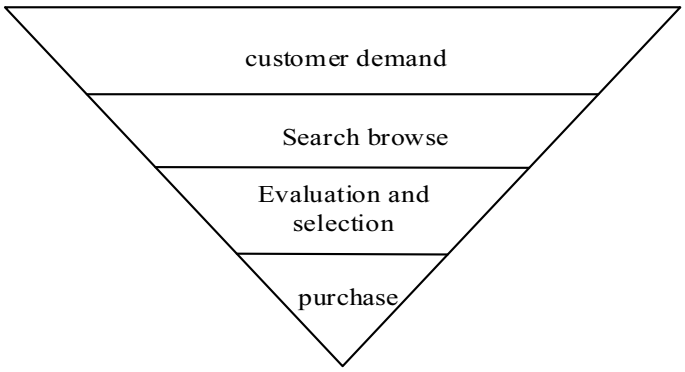

Fig. 2. "Sales Funnels" in consumer behavior

Data mining is a process of discovering valuable patterns, relationships, and trends by analyzing data, in which data is stored in the form of a database. Data mining is not only the synthesis of some pattern recognition technologies but also a systematic method system based on the principle of "discovering knowledge from data sets". Data mining comes from application and orients application. CRISP-DM, a cross-industry standard process of data mining, is one of the most common standards in the data mining industry nowadays. It emphasizes the application in a business environment and solves the problems in business, rather than confining data mining to the research field. CRISP-DM model is business application-oriented and data-centered. The process includes business understanding, data understanding, data preparation, model building, model assessment, and result released, and continuously circulates and optimizes, as shown in Fig. 3.

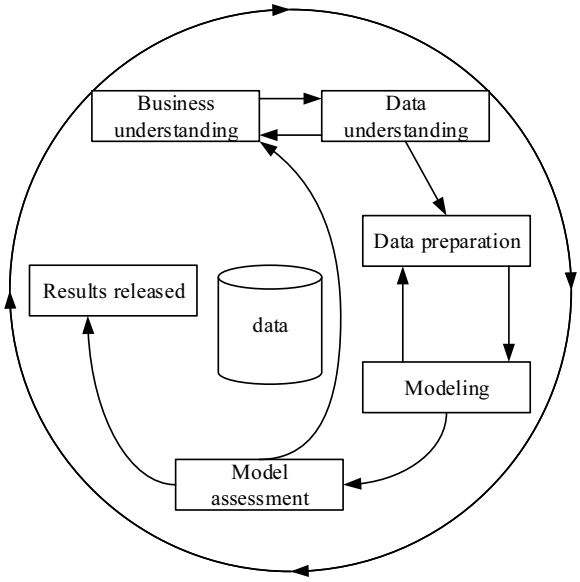

Fig. 3. Inter-industry Data Mining Process Diagram

Causal prediction method is used to build data mining models such as SVM to predict, and the classical multiple linear regression is compared. Using the method of 10-fold cross validation to evaluate the prediction results of the model, it is found that under the current data set, SVM based on radial basis kernel has the best prediction effect, with an error rate of 7%, and the generalization performance of the model is the most stable. However, the prediction effect of multiple linear regression is also excellent,

and the prediction effect is close. It shows that in the current business scenario, the influence mode of influencing factors on sales is mainly linear. Under the guidance of current methodologies, excellent forecasting results and rich management inspiration can be obtained, which proves that the forecasting process proposed in this paper has great application value.

4 Results and Discussion

4.1 Experimental Design and Environment

The effect of E-commerce promotional activities includes short-term effect and long-term effect. This paper mainly focuses on the short-term effect of the activities, starting from the two aspects of passenger flow and sales that the businessmen pay more attention to. Considering the passenger flow effect of the activities and the sales effect of the activities synthetically, the sales effect is analyzed from the commodity dimension of the activities and the commodity dimension of the non-activities, respectively, and the passenger flow effect is analyzed from home page and commodity details. Sales effect is mainly analyzed by sales volume index, and passenger flow effect is mainly analyzed by visitor number index. The overall analysis is shown in Fig. 4:

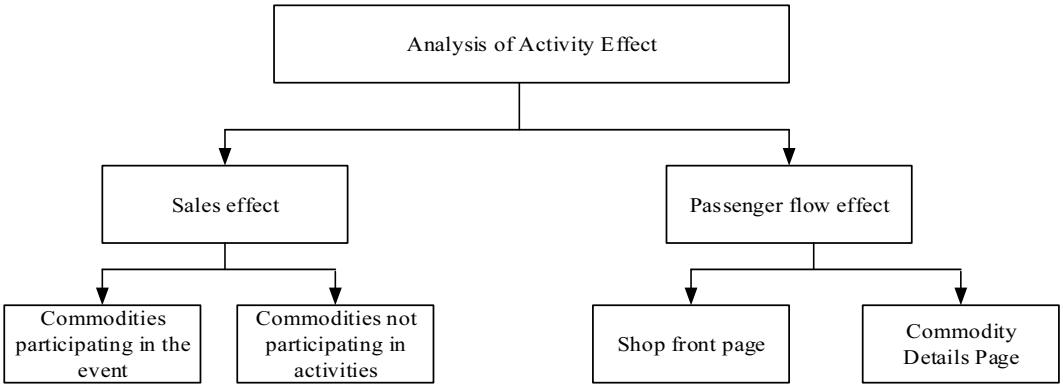

Fig. 4. Analysis chart of promotion effect of E-commerce

As for the sales effect of activities, as a whole, it includes not only the effects of activities on participating commodities but also the effects of activities on non-participating commodities. Therefore, the activity sales effect defined mainly includes the following four parts: the positive sales effect of the activity on the goods participating in the activity, the negative sales effect before and after the activity, the positive sales effect of the goods not participating in the activity during the activity, and the negative sales effect before and after the activity. The specific activity sales effect is expressed as follows:

$$E_K^{sales} = \sum_{i=1}^{n}\sum_{t=1}^{m} E_{ikt} + \sum_{i=1}^{n}\sum_{w=1}^{u} E_{ik}\left(t_{s/e} \pm w\right) + \sum_{j=1}^{g}\sum_{t=1}^{m} E_{jkt} + \sum_{i=1}^{g}\sum_{w=1}^{u} E_{jk}\left(t_{s/e} \pm w\right). \quad \textbf{(1)}$$

In Equation (1), E_K^{salses} indicates the overall sales effects of the k-th promotion activity, $\sum_{i=1}^{n}\sum_{t=1}^{m} E_{ikt}$ suggests the sales effects of the k-th activity on the commodities participating in the activity during the period of the activity, n is the number of commodities participating in the activity, m is the number of days lasting for the activity, $\sum_{i=1}^{n}\sum_{w=1}^{u} E_{ik}\left(t_{s/e} \pm w\right)$ represents the sales effects of the k-th activity on the commodities participating in the activity before and after the activity, and u is the number of days lasting for the activity before and after the activity. $\sum_{j=1}^{g}\sum_{t=1}^{m} E_{jkt}$ indicates the sales effect of the kth activity on the commodities not participating in the activity, g is the quantity of the commodities not participating in the activity, m is the duration of the activity, $\sum_{i=1}^{g}\sum_{w=1}^{u} E_{jk}\left(t_{s/e} \pm w\right)$ is the sales effects of the k-th activity on the

goods for participation in the activities in the early and late periods, and u is the duration of the activity before and after the activity.

For each of the above effects, the difference between the actual sales of goods and the benchmark sales during that period should be taken into account. Taking Eikt as an example, the effect can be expressed as follows:

$$E_{ikt} = Q^{r}_{t,ik} - Q^{b}_{t,ik} . \tag{2}$$

$Q^{r}_{t,ik}$ refers to the actual sales volume of i-th goods on the t-th day of the k-th activity; $Q^{b}_{t,ik}$ indicates the base sales volume of i-th goods on the t-th day of the k-th activity.

4.2 Data Collection and Data Processing

In the analysis of activity effect, combined with its definition and expression, the sales and passenger flow effects before and after the activity and the sales and passenger flow effects during the activity are separately analyzed. For this reason, the daily sales and passenger flow are divided into three samples, i.e. the activity period, the pre-activity and post-activity period, and the daily period. Then, the T-test analysis method is used to analyze the samples in the pre-activity and post-activity period, in the activity period, and in the daily period, respectively for hypothesis tests. The mathematical expressions of the T-test method are as follows:

$$t = \frac{\overline{X_1} - \overline{X_2}}{\sqrt{\frac{S_p^2}{n_1} - \frac{S_p^2}{n_2}}} . \tag{3}$$

$$S_P^2 = \frac{\sum\left(X_1 - \overline{X_1}\right)^2 + \sum\left(X_2 - \overline{X_2}\right)^2}{\left(n_1 - 1\right) + \left(n_2 - 1\right)} . \tag{4}$$

The samples to be tested are S_1 and S_2, S_P is the weighted average of the variance of S_1 and S_2, $\overline{X_1}$ and $\overline{X_2}$ are the average of S_1 and S_2 of the samples to be tested, and n_1 and n_2 are the number of S_1 and S_2 of the samples to be tested. In the process of T-test, this paper firstly performs T-test analysis on whether there is a difference between test samples. When there is no difference between the test samples, the test ends. When there is a difference between test samples, then carry out a positive difference test and negative difference test analysis, respectively. Finally, it draws the necessary conclusion. The detailed inspection process is shown in Fig. 5 below.

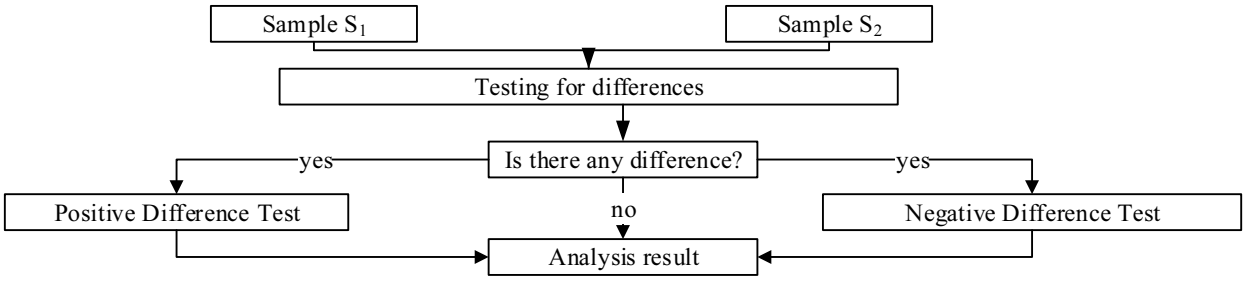

Fig. 5. Model diagram of effect T-test and analysis

For the analysis of the sales effect of activities, hypothesis test analysis is made on the sales volume of all commodities, the effects of the sales volume of participating commodities during the activities, and the effects of the activities before and after the activities, as well as the interactive effects of the sales volume of non-participating commodities during the activities and before and after the activities. For the analysis of the effects of active passenger flow, hypothesis test analysis is performed on the effect of home page visits and commodity details pages visits during and before and after activities.

Each promotion activity usually lasts for 2 to 3 days. The time-varying analysis of activity effect is

mainly aimed at the changing law of its effect during the activity period and along with the activity time. Therefore, a SVR model is established. The specific form of the model is as follows:

$$\ln(Y) = \beta_0 + \beta_J X_J + \varepsilon . \quad \textbf{(5)}$$

Y is the dependent variable, considering the fluctuation of effect change, the dependent variable is logarithmically changed to reduce the impact of fluctuation on the results. X_j is a p-dimensional row vector, representing the set of independent variables, and p is the number of independent variables.

$$X_J = (X_1, \cdots, X_p) . \quad \textbf{(6)}$$

At the same time, β_0 and β_j are the coefficients for regression estimation, β_0 is intercepted term, and β_j is a p-dimensional column vector, representing the set of independent variable coefficients:

$$\beta_j = (\beta_1, \cdots, \beta_p)^T . \quad \textbf{(7)}$$

In addition, ε is a random error term, which satisfies the basic assumptions such as $E(\varepsilon \mid X_j) = 0$. This paper mainly analyses the time-varying sales effects of participating commodities during the activity, and the selected independent variables include the activity time and price discount. The time of promotion activity, i.e. the number of days accumulated from the beginning of activities to k-th days of activities, is expressed in an incremental way from 1 to k. Price discount is the percentage of price discount for commodities. The dependent variable is the logarithmic value of the sales effects of the participating commodities during the activity.

4.3 Performance Evaluation and Discussion

Through the information of the relevant personnel of the enterprise, during the activity, the enterprise will choose some commodities to participate in the activity. For the commodities involved in the activity, they can be divided into two categories, including drainage promotion commodities and ordinary promotion commodities. The overall diagram of the commodities is shown in Fig. 6.

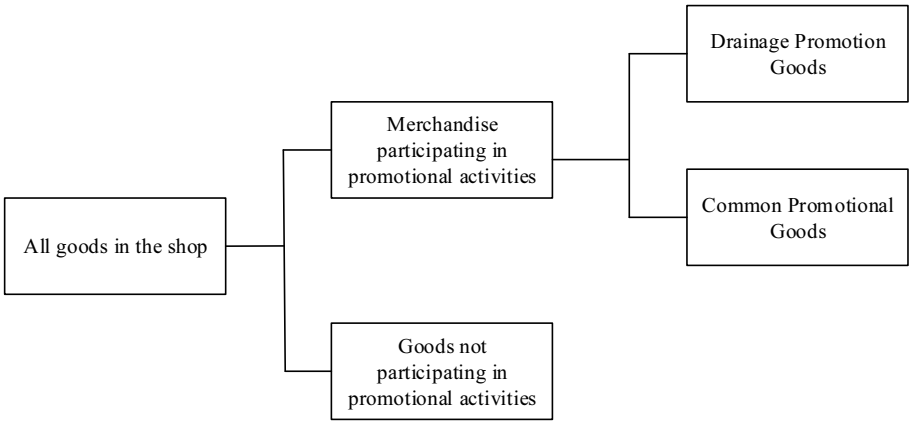

Fig. 6. Overall schematic diagrams of commodities

For diversion promotional commodities, enterprises often provide such commodities as diamond booth and home page display to promote sales. Under the E-commerce environment, different booths will bring different click-through rates and traffic, thus affecting the potential purchasing power of commodities. Common promotional commodities mainly rely on price discounts to promote sales. For the above two types of commodities, the same prediction model can be used to predict, but the influencing factors need to be selected differently. In the analysis of influencing factors, the analysis of ordinary promotional commodities is mainly considered. The commodities mentioned are not special annotations, referring to ordinary promotional commodities.

This paper synthetically considers the two aspects of easy collection and quantification at the present stage and chooses the appropriate influencing factors to forecast and analyze the sales volume of the commodities participating in the activities. Generally speaking, the influencing factors of sales during its

activities mainly include two aspects: product factors and the activity itself factors, among which product factors include its own product factors and competitive product factors.

For other factors, such as external macroeconomic policies, individual differences of consumers, etc., will also have an impact on the activity sales of products. However, these factors are difficult to quantify or collect effectively, and the existing literature often does not analyze and discuss them. Therefore, it mainly considers the price discount of the target commodity during the activity period, the emotional score of the target commodity comments in the pre-activity period, the monthly seasonal index of the target commodity during the activity period; the price discount index of the target commodity relative to the commodity within the category during the activity period, the commodity sales index of the target commodity relative to the category in the pre-activity period, the activity duration, the activity interval time and the activity daily sales index and so on 8 kinds of influencing factors. To sum up, the factors influencing the volume of sales activities considered here are shown in Fig. 7.

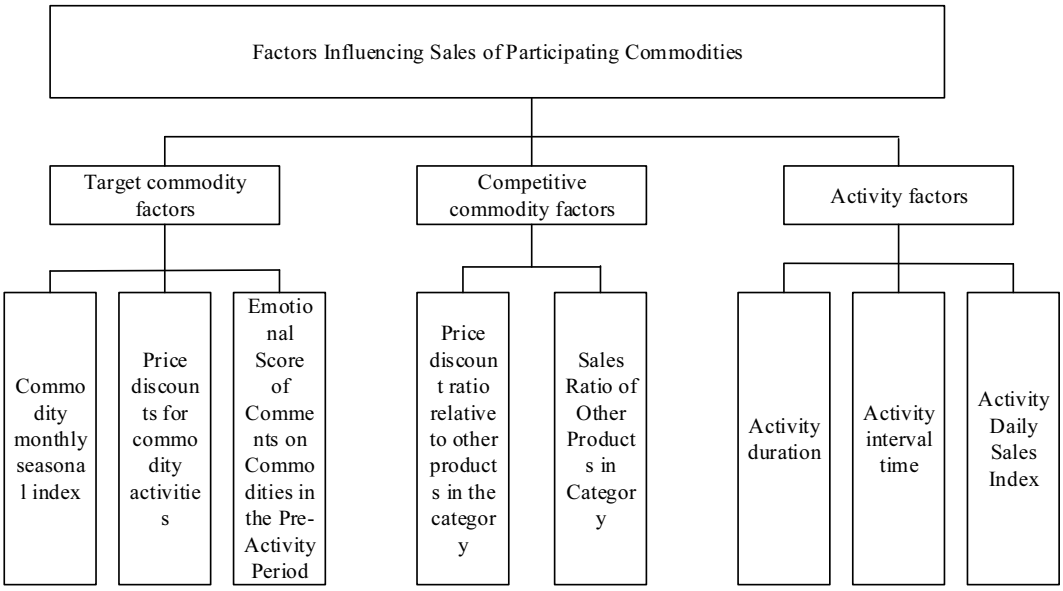

Fig. 7. Factors influencing sales of participating commodities

In order to improve the prediction efficiency of the model and the prediction performance of the system, the prior experience is adjusted combined with the method of data mining. Therefore, SVR machine is used as the basis of the prediction model, and grey comprehensive correlation analysis is applied to the modeling process of SVR machine. According to the similarity or dissimilarity of the changing trend between influencing factors and activity sales volume, the main influencing factors are screened out by grey comprehensive correlation analysis, and then these main influencing factors are used as the input of SVR machine prediction model to solve support vector.

The quantification, availability, and cost in the business environment are determined to be used as the feature variables of model. Based on the quantitative analysis of influencing factors, the data are compared and missing fields are required from enterprise operationist. The fields obtained are shown in Table 1.

Table 1. Data set of influencing factors of e-commerce sales

Subject of influence	Influencing factor	Quantitative index
Online retailers	Product richness	SKU kinds
Online retailers	Consumption intention	Number of collectors per person/piece
Online retailers	Consumption intention	Number of additional purchases per person/piece
Online retailers	Direct drop	Direct drop in amount/order/sales
Online retailers	Reduction	Reduction in amount/order/sales
Online retailers	Coupon	Coupon discount amount/order/sales
Online retailers	Stock	Stock amount
Social / market environment	Festival	Holiday label
Social / market environment	Week	Week label

The sales volume, browsing volume, and promotional data obtained are SKU-level data, and the inventory data are SKU-level data, but the number of collectors and purchasers are day-level data. Festivals and weeks are additional labels in days.

Data need to be pre-processed before activity sales forecasting. The form and quality of data will have a certain impact on the final results. Data pre-processing process generally includes the following aspects: outlier processing, missing value processing, data discretization, and data standardization. Outlier processing, missing value processing, and data standardization are analyzed. Through the analysis of the sales volume of the existing data, it is concluded that the main reason for the outliers is the large-scale promotional activities of E-commerce platforms, such as "double eleven" and "double twelve". This kind of activity has become the national shopping festival of E-commerce platforms. During this period, the sales volume and activity effect will increase significantly. If not eliminated, the accuracy of the prediction model will be affected. In the process of data collection and processing, due to some subjective or objective reasons, the data set has some deficiencies. For example, when a commodity participates in an activity for the first time, it will lack the indicator of the interval between activities. Data standardization is a method to narrow the range of sample data, which avoids the influence of too large difference among the order of magnitude of variables in the input model, or the large difference of data size in spite of the difference of numerous orders of magnitude. [-1, 1] standardization method is adopted. The formulas for calculating standardization and anti-standardization are as follows. X' and X are the normalized values and the original values, and X_{max} and X_{min} are the maximum and minimum values, respectively.

$$X' = \frac{X - \frac{1}{2}\left(X_{\max} + X_{\min}\right)}{\frac{1}{2}\left(X_{\max} - X_{\min}\right)}. \tag{8}$$

$$X = \frac{1}{2}\left(X_{\max} - X_{\min}\right)X' + \frac{1}{2}\left(X_{\max} + X_{\min}\right). \tag{9}$$

SVM is a machine learning method based on statistical learning theory. Considering the principle of structural risk minimization rather than empirical risk minimization, it can solve practical problems such as non-linearity and small sample size. SVM is a new technology in data mining, which is often used to deal with many problems such as regression (SVR machine) and can be extended to prediction and other fields. In the process of modeling SVR machine, it is necessary to consider the setting of kernel function and its parameters. The generalization ability of the model depends to a certain extent on the kernels and their parameters and penalty factors. Kernel function realizes the non-linear mapping relationship from sample input space to feature space. Choosing different kernel functions will have some influence on the final results of the SVR machine. Therefore, it is necessary to adapt the kernel function according to the characteristics of different problems. Because the use of Gauss kernels not only has better approximation performance for high-frequency nonlinear systems, but also has a smaller range of effective parameters, and the space complexity will not be too large when the parameters change in the effective range. Therefore, the commonly used Gauss kernel function is used as the kernel function of the model. The expression of the Gauss kernel function is as follows:

$$k\left(x, x_i\right) = \exp\left(-\frac{\left\|x - x_i\right\|}{2\sigma^2}\right). \tag{10}$$

For SVM, besides the selection of kernel function, its penalty parameters and kernel function parameters will also have an important impact on the learning ability and generalization ability of the model, and their changes will produce certain differences in the final results. These parameters only perform well in a certain range of areas. Once deviated from this area, the generalization ability and prediction ability of the model will be dramatically reduced. The global optimization ability and parallel searchability of PSO are used, and the cross-validation method is used to optimize the parameter selection of SVR machine. Each particle in POS algorithm corresponds to a set of parameters of SVR machine, and the optimal parameters are solved by minimizing the generalization error. The overall optimization process is divided into the following eight steps. The forecasting model is evaluated on

datasets collected from Alibaba.com. Initially, the model is evaluated on a subset of 1,724 items that belong to the product household category, which consists of 15 different sub-categories. Next, the number of products to 18,254 is scaled up by extracting a collection from a single super-department, which consists of 16 different categories. A total of 190 consecutive days of sales data in 2018 are used. The last 10 days of data are reserved for model testing. The optimization flow of the algorithm is shown in Fig. 8 below.

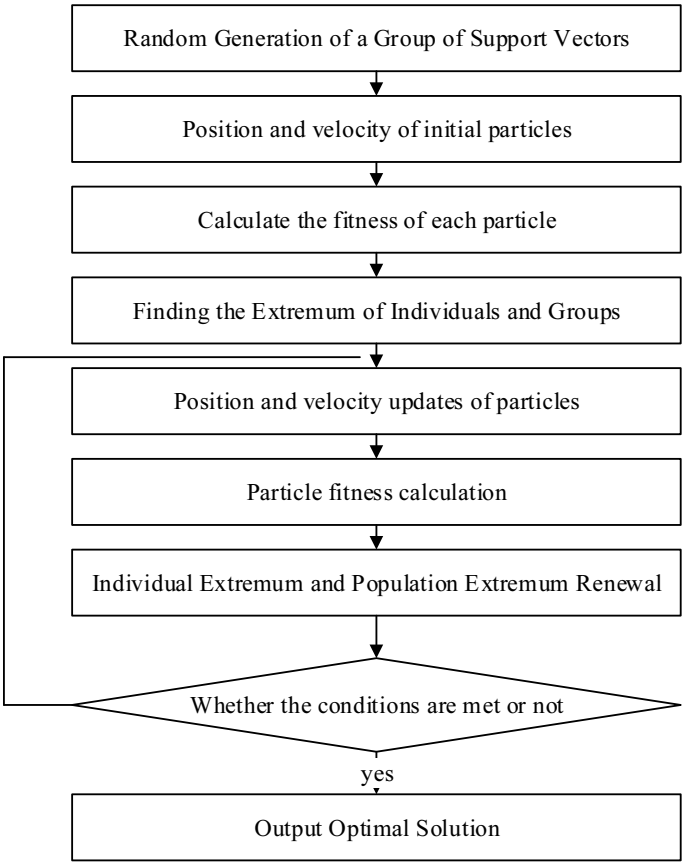

Fig. 8. Parameter flow chart of SVR machine based on PSO

Table 2. Final model forecasting results

Activity number	Actual Value (piece)	Comprehensive prediction model			SVM single model		
		Predicted values (piece)	Absolute error	Relative error rate	Predicted values (piece)	Absolute error	Relative error rate
15	112	137	25	22.32%	134	22	19.64%
16	110	119	9	8.18%	126	16	14.55%
17	135	125	-10	7.41%	124	-11	8.15%
18	144	140	-4	2.78%	132	-12	8.33%
19	108	97	-11	10.19%	93	-15	13.89%
Average error rate				10.17%	Average error rate		12.91%

A schematic diagram of the forecasting results is shown in Fig. 9 below.

It is seen that the predicted value and the actual value of the model are relatively close, and the accuracy of the prediction is about 89.83%. Compared with the methods of index smoothing and moving average, the forecast of sales during the activity has been improved. Meantime, it is known that enterprises mainly use expert judgment method for activity sales prediction, which relies heavily on experts. To sum up, this method has certain practical significance.

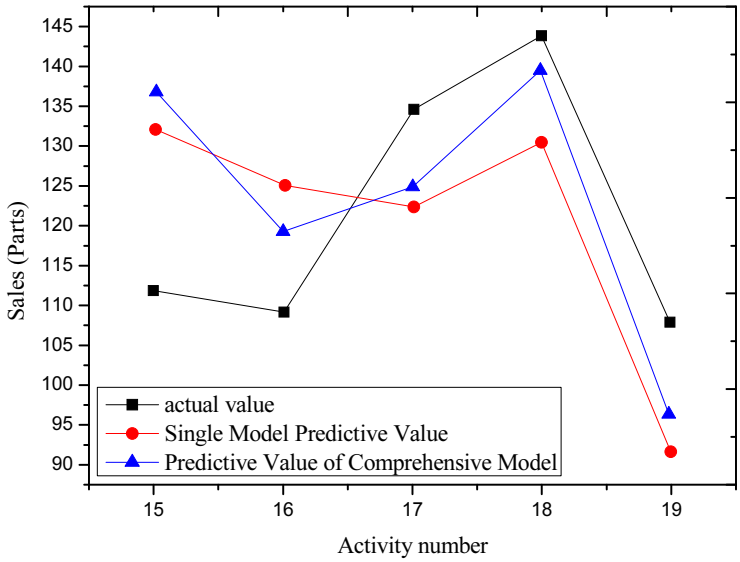

Fig. 9. Forecasting results

All the commodities sold during the enterprise activities are modeled and analyzed according to the sales forecast of the participating activities and the non-participating activities. The prediction model PSO-SVR based on grey comprehensive correlation analysis is studied and applied to the sales forecast of participating goods. The grey comprehensive correlation analysis is used to calculate the correlation degree of the influencing factors of sales volume. The grey relative correlation degree and grey absolute correlation degree are considered comprehensively. The main influencing factors extracted are taken as input. Then, the parameters of SVR machine are optimized by using PSO algorithm, and the PSO algorithm is used to calculate the parameters of SVR machine. The group optimal extremum obtained by the PSO method is used as the parameter value of the vector regression machine to improve the prediction accuracy and applicability of the algorithm. Additionally, the indexes of association rules are analyzed to determine the calculation method of sales of non-participating goods. Considering the algorithm design, the association rules generated based on interest degree are adopted to prevent the generation of invalid rules.

5 Conclusion

In order to solve the current problems of sales volume of E-commerce, the ideas and concepts of data mining are adopted in the research technology and methodology of the study. In addition, to ensure the validity of modeling and the application of real scenarios, the research attaches great importance to the analysis and collection of sales-related factors. Therefore, the Cross-Industry Standard Process for Data Mining (CRISP-DM), which is commonly known by its acronym CRISP-DM and is a data mining process model that describes commonly used approaches that data mining experts use to tackle problems, is applied to build the forecasting model. In addition, data collection, data processing, data analysis, and modeling are carried out.

Besides, based on the characteristics of E-commerce sales, the sales effects and customer flow effects of E-commerce promotion activities are comprehensively considered, defined, and quantified. Through the statistical analysis, it is found that the sales volume during the promotion activity period would be significantly larger than that during the normal times, while the sales volume before and after the activity period would be smaller than that during the normal times. At the same time, the analysis of passenger flow and conversion rate during, before, and after the activity period, as well as during the normal times, has shown that the conversion rate before and after the activity period is lower. The forecasting process based on the CRISP-DM process focuses more on the combination of data and business knowledge than the traditional prediction process, which is very effective under the background of big data. Moreover, in this process, through a variety of data processing methods, more information can be learned, which is conducive to the later optimization. Practically, the forecasting model has been approved by managers of various companies and is finally retained in the form of a program, which proves the application value of

the forecasting process. From the perspective of corporate profitability, based on the SVR for comprehensive modeling analysis, the benefits of the commodities in the promotion activities, as well as before and after the promotion activities, have been predicted by the model. The loss and the commission fee charged by the E-commerce platform are used for predictive analysis so that the company can arrange its next stage of production, logistics, promotion activity, and new business expansion, etc. The results have shown that the comprehensive forecasting model has a certain improvement in prediction accuracy compared with the single SVR prediction model, and it helps the E-commerce companies improve the scientific degree and intelligence degree of their management and decision-making, which is of certain theoretical and practical significance.

References

[1] F. Dai, S.T.T. Teo, K.Y. Wang, Network Marketing Businesses and Chinese Ethnicity Immigrants in Australia, Journal of Small Business Management 55(3)(2017) 444-459.

[2] H. Rui, J. Wu, H.S. Du, Customer social network affects marketing strategy: A simulation analysis based on competitive diffusion model, Physica A Statistical Mechanics & Its Applications 469(2017) 644-653.

[3] R.S. Sadasivam, S.L. Cutrona, T.M. Luger (Eds.), Share2Quit: Online Social Network Peer Marketing of Tobacco Cessation Systems, Nicotine & Tobacco Research 19(3)(2017) 314.

[4] I. Tricahyadinata, S.Z. Za, An Analysis on the Use of Google Adwords to Increase E-Commerce Sales, Social Science Electronic Publishing 4(1)(2017) 60.

[5] S. Kudyba, K. Lawrence, Enhancing information management through data mining analytics to increase product sales in an e-commerce platform, International Journal of Electronic Marketing & Retailing 2(2)(2017) 97-104.

[6] Z. Shaikh, A. Mohadikar, R. Nayak (Eds.), Security and Verification of Server Data Using Frequent Itemset Mining in Ecommerce, International Journal of Synthetic Emotions 8(1)(2017) 31-43.

[7] S. Hernández, P. Álvarez, J. Fabra (Eds.), Analysis of Users' Behavior in Structured E-Commerce Websites, IEEE Access 5(99)(2017) 11941-11958.

[8] J. Xue, S. Jarvis, Mining association rules for admission control and service differentiation in e- commerce applications, Wiley Interdisciplinary Reviews Data Mining & Knowledge Discovery 8(11)(2018) e1241.

[9] T. Escobar-Rodríguez, R. Bonsón-Fernández, Analysing online purchase intention in Spain: fashion e-commerce. Information Systems and e-Business Management 15(3)(2017) 1-24.

[10] Y. Kim, R.A. Peterson, A Meta-analysis of Online Trust Relationships in E-commerce, Journal of Interactive Marketing 38(2017) 44-54.

[11] X. Hong, To compete or to take over? An economic analysis of new sellers on e-commerce marketplaces, Information Systems and e-Business Management 16(3)(2017) 1-13.

[12] N. Hajli, M.S. Featherman, Social commerce and new development in e-commerce technologies, International Journal of Information Management 37(3)(2017) 177-178.

[13] L.V. Orman, Information markets over trust networks, Electronic Commerce Research 16(4)(2017) 1-23.

[14] W.J. Ding, H.B. Duan, S.J. Ye, Research on the Promotion of E-commerce Website of Small and Medium-sized Enterprises in China, Global Finance Review 1(1)(2018) 31-38.

[15] A. Valarezo-Unda, T.P. Amaral, T. Garín-Muñoz, I. Herguera, Drivers and barriers to cross-border e-commerce: Evidence from Spanish individual behavior, Telecommunications Policy 42(6)(2018) 464-473.

[16] D. Di Fatta, D. Patton, G. Viglia, The determinants of conversion rates in SME e-commerce websites, Journal of Retailing and Consumer Services 41(2018) 161-168.

[17] M. Tang, H.C. Liao, Multiple Criteria Group Decision-Making Based on Hesitant Fuzzy Linguistic Consensus Model for Fashion Sales Forecasting, Artificial Intelligence on Fashion and Textiles 849(2018) 329-336.

[18] R. Hermanto, E. Junaeti, R. Wirantika, Implementation of Automatic Clustering Algorithm and Fuzzy Time Series in Motorcycle Sales Forecasting, IOP Conference Series: Materials Science and Engineering 288(2018) 012126.

[19] E.W.K. See-To, E.W.T. Ngai, Customer reviews for demand distribution and sales nowcasting: a big data approach, Annals of Operations Research 270(1-2)(2018) 415-431.

[20] S. Tandon, A. Govindaraj (Eds.), Decoding Digital Consumer Feedback: Customer Intelligence Insights Through Unstructured Data Mining, Social Media Marketing 1(2018) 113-120.

[21] W. Li, Q. Zhou, J. Ren (Eds.), Data mining optimization model for financial management information system based on improved genetic algorithm, Information Systems and e-Business Management 4(2019) 1-19.

[22] H. Zhang, H.G. Rao, J.Z. Feng, Product innovation based on online review data mining: a case study of Huawei phones, Electronic Commerce Research 18(1)(2018) 3-22.

[23] W.B. Chang, X.L. Yuan, Y.L. Wu, S.H. Zhou, J.S. Lei, Y.Y. Xiao, Decision-Making Method based on Mixed Integer Linear Programming and Rough Set: A Case Study of Diesel Engine Quality and Assembly Clearance Data, Sustainability 11(3)(2019) 1-21.

[24] C. David, M. Salamó, Data-driven decision making in critique-based recommenders: from a critique to social media data, Journal of Intelligent Information Systems 51(1)(2018) 1-22.

[25] H. Golpîra, A novel Multiple Attribute Decision Making approach based on interval data using U2P-Miner algorithm, Data & Knowledge Engineering 115(2018) 116-128.

[26] D. Arunachalam, N. Kumar, Benefit-based consumer segmentation and performance evaluation of clustering approaches: An evidence of data-driven decision-making, Expert Systems with Applications 111(2018) 11-34.

[27] T. Sultana, Z.A. Khan, N. Javaid (Eds.), Data Analytics for Load and Price Forecasting via Enhanced Support Vector Regression, 7th International Conference on Emerging Internet, Data, & Web Technologies (EIDWT-2019) 29(2019) 259-270.

[28] S. Petruševa, D. Car-Pušić, V. Zileska-Pancovska, Support Vector Machine Based Hybrid Model for Prediction of Road Structures Construction Costs, IOP Conference Series: Earth and Environmental Science 222(1)(2019) 012010.

[29] D. Oreški, N.B. Ređep, Data-driven decision-making in classification algorithm selection, Journal of Decision Systems 27(1)(2018) 248-255.

[30] C.Q.X. Poh, C.U. Ubeynarayana, Y.M. Goh, Safety leading indicators for construction sites: A machine learning approach, Automation in Construction 93(2018) 375-386.