International Journal of Computing and Digital Systems
ISSN (2210-142X)
Int. J. Com. Dig. Sys. 9, No.5 (Sep-2020)

http://dx.doi.org/10.12785/ijcds/090503

A Study of the Demand Forecasting Model for Publishing Business using Business Analysis

Mee Hwa Park[1], Jong Sup Lee[2] and Ill Chul Doo [3]

[1] *Software Focused University, Korea*
[2] *Development of smart community policing system(Googi) Center, Dongguk University, Seoul, Korea*
[3] *Hankuk University of Foreign Studies, Korea*

Received 25 Nov. 2019, Revised 9 Jul. 2020, Accepted 1 Aug. 2020, Published 1 Sep. 2020

Abstract: Demand forecasting is the activity of predicting the future using historical data and establishing a model that can grasp trends. Demand forecasting is widely used in a variety of business areas, including production and inventory planning as well as process management. The goal of a company that publishes and sells books is to accurately predict sales volume, thereby increasing book sales, generating more revenue, and reducing losses from inventory management. Using data analysis to predict accurate publishing demand and establishing countermeasures against factors that may cause returns can reduce the amount of losses incurred due to inventory control and returns. The purpose of this study is to identify the factors affecting the sale and return of specific books and to create a model to forecast sales demand. For this purpose, we used the sales data of the books sold for 5 years (2012 ~ 2016) by A publishing company. In addition, we collected the data related to books in the Internet portal system and SNS site. We hypothesized the factors that affect the sale and return of books and collected the variables needed for hypothesis testing from web pages and SNS sites. As a result of this study, it was possible to identify the factors affecting the return and sales of a specific book, and it was possible to establish a sales order prediction model. Because the available data is limited in the study, the scope of this study was limited to forecasting the sales demand of some books. If we apply the proposed analytical procedure and method directly from the company, we can expect better prediction results. It is also expected to be applicable to various business processes of book publishing or sales companies.

Keywords: Business Analysis, Demand Forecasting, Linear Regression Model., Vector Machine Regression,

INTRODUCTION

Demand forecasting is an activity that uses historical data to establish a model that can identify trends and forecasts the future by using this model. Demand forecasting is widely used in a variety of business areas [1]. Demand forecasting is a field of interest for business owners of all sizes, from department stores to large marts to convenience stores to street vendors. Interest in demand forecasting is also high, regardless of the type of business, from agriculture, construction and manufacturing to services such as restaurants and hospitals. Accurately predicting sales volume in an industry that handles fresh foods such as vegetables and fruits is an important factor in minimizing inventory management costs and losses incurred by discarding unsold goods. In addition, in the manufacturing sector, which requires large-scale production lines, and in the construction sector, which takes a long time from ordering to completion, it is essential for the decision maker to predict how the demand will change in the long term and how the market will be supplied. Demand forecasting is a key element for setting sales targets, capital investment, production planning, inventory management, supply chain management and marketing strategy development [2].

To become a sustainable company in today's changing and competitive environment, companies must recognize the importance of establishing a demand forecasting system and invest enough to raise the level of forecasting. Good demand forecasting models should be able to combine changing information with new information. And it must provide a variety of ways to analyze the relevant data. Demand forecasting should aim not only to obtain forecasts but also to derive strategic implications through forecasts. Given this, organization's decision makers or data analysts must have the capability to present the insights and alternatives that align demand forecasts with strategies.

The popularization of mobile terminals and the Internet and the resulting changes in the lifestyle of modern people are replacing existing businesses at a very

E-mail: dic@hufs.ac.kr, maia.mhpark@gmail.com, jsleearmy@dongguk.edu

rapid pace. With the advent of electronic books, the paper book market has been shrinking, which has shrunk the publishing market. In the publishing industry, there were concerns that the development of e-book devices and software to help readers would threaten traditional book publishing. However, the magazine market, which was more news-oriented than the depth of information, was replaced by e-books. Demand for paper books did not decrease from literature to literary books, various practical books and study reference books [8].

The change in the publishing market due to digital is that the market share of online bookstores that appeared 20 years ago is increasing exponentially [9]. Internet bookstores are contributing greatly to the activation of the publishing market. Internet bookstore services are similar to online shopping malls. In online shopping malls, customers get information about products they are interested in by chatting with other customers, bulletin boards, and consumer reviews. Consumers make judgments based on various information provided to them when purchasing goods. Based on these facts, there is a study that verifies that users' reviews of book sales are helpful by examining whether Internet bookstore reviews have a direct effect on book purchase decisions of book buyers using Internet bookstores [10]. In the publishing business, various problems such as inventory management and financing, loss of customers due to uncertain demand, and customer complaints due to delayed delivery can occur when the demand for the distribution stage is not predicted. In other words, if the demand is not properly predicted and the number of publications is arbitrarily calculated, the company will be struggling to deal with inventory due to overproduction. In addition, bookstores that do not have the books they want can lose their customers to competitors.

A company that publishes and sells books tries to generate more revenue by increasing book sales and reducing inventory losses. Using data analysis to predict accurate publishing demand and establishing countermeasures against factors that may cause returns can reduce the amount of losses incurred due to inventory control and returns. The goal of this study is twofold: ① Identify factors that affect the increase or decrease of sales rate of books. ② Develop a predictive model to predict the publication demand of books. For this purpose, this study used the sales data of books sold by A publishing company as experimental data. Additional data needed for forecasting were collected from Internet portal systems and SNS sites.

The composition of this paper is as follows. In Chapter 2, we examine existing prior studies and forecasting algorithms for forecasting analysis in the publishing market. In Chapter 3, we present the demand analysis and forecasting methods for books, and In Chapter 4, we demonstrate the usefulness of the proposed method based on empirical analysis. In the last chapter, conclusions and limitations of this study and future research directions are presented.

Related Research

Existing research on forecasting analysis in publishing business focuses on analysis of sales revenue and marketing mainly in internet bookstore. Boon Do Jeong and Mi Seon Hong [11] studied the effect of online shopping business on consumers' intention to repurchase and their intention to speak. They found that convenience of online shopping and empathy affect customer satisfaction and affect repurchase intentions. Here, empathy means that the online shopping mall offers a variety of services besides the products, or grasps the interests of the customers and cares for the customers well. Kim Hyun-Chul Kim [12] conducted a survey to investigate the effect of customer's beliefs on reputation, quality, and certainty of online shopping malls. As a result, he found that trust in shopping mall has a significant effect on commitment and purchase intention.

Consumers who want to buy books cannot read all the books, so the reader decides whether to buy the book in a variety of ways. Ok Ryun Jung [13] found that readers selected different books depending on their social status or information processing capacity. And they found that readers tend to take into account both previous reading experiences, their knowledge, and information from social networks such as friends and family when choosing books. Many studies have found that the expertise of a reviewer generally has a positive impact on the usefulness of reviews [14-16]. In particular, it was confirmed that the reputation of the reviewer is the dominant factor in determining the usefulness of the review. In addition, when a reviewer writes a lot of reviews only for a specific product line, online consumers can trust that the reviewer's expertise and consider the review useful. Previous studies have examined the purchase intention analysis and customer review activities necessary for marketing based on the online bookstore, but no study has predicted the sales demand of the book. The purpose of this study is to identify the factors affecting the sale and return of specific books and to create a model to forecast sales demand.

Many companies use demand forecasting techniques to adapt to changes in the business environment, such as changes in circumstances, consumption trends, or emergence of new technologies. The prediction can be classified into long-term prediction, mid-term prediction, and short-term prediction according to the target prediction period. In general, long-term forecasting can be considered to be a case where the forecasting period is more than two years, and is used mainly in the case of strategic decision-making such as product planning, capability planning, and location determination. Since the prediction period is long, subjective judgment based on environmental prediction is widely used and accuracy is relatively low. Mid-term forecasts are typically 6 months to 2 years, and quantitative approaches are available and expert opinions are helpful. Short-term forecasts are quarterly, monthly, weekly, and daily forecasts, usually within six months, and allow relatively accurate forecasts.

It is also classified into qualitative method, quantitative method, and systematic method depending on whether numerical calculation method is central or not [17]. The qualitative prediction technique is mainly applied to the mid - and long - term prediction. As market potential changes with changes in external environmental factors such as economy, politics, society, and technology, qualitative prediction techniques predict demand based on subjective judgment or opinion. Qualitative techniques can also be useful if past data are not sufficient. Qualitative prediction techniques include Delphi method, market research method, and panel consensus method. In qualitative prediction techniques, information from experts and external organizations can be used as important. Qualitative techniques are useful in identifying the needs of a product's features or attributes, or predicting market response to new products, by taking advantage of expert opinions with experience and knowledge of the product or similar product market. Qualitative techniques generally have a high time and cost of applying the prediction techniques and can be applied to mid- and long-term strategy decisions such as product development, technology forecasting, market strategy, and factory location selection.

The quantitative prediction method is classified into causal forecasting method and time series analysis. The causal prediction technique is a technique for predicting future demand by identifying the environmental factors that affect demand and identifying the causal relationship between demand and factors. Therefore, causal prediction techniques use demand as a dependent variable and factors that affect demand as independent variables. In the causal prediction method, factors related to changes in the external environment such as GNP, sales policies of competitors, birth rate, and internal factors such as advertising, promotional activities, quality, and credit policy can all be reflected in the model. The causal prediction methods include regression analysis, industry association analysis, input output analysis, and leading indicator method.

Regression analysis is a statistical model of demand as a dependent variable and a function of independent variables as independent variables. The modeled function is called the regression equation. When the value of the independent variables is given, the predicted value of the demand is calculated through the regression equation. At this time, the demand response for the independent variables can be linearly modeled. If there is one independent variable, it is called simple regression analysis, and if more than two, it is called multiple regression analysis. Nonlinear regression analysis is used when the response of demand to independent variables is modeled nonlinearly [18].

The time series forecasting method is a method of analyzing past demand, grasping the pattern of demand according to time, and predicting future demand on the extension line. In other words, as a method of projecting future demand from past demand flow, market stability is required as a basic assumption that past demand patterns will persist in the future. However, since the demand pattern of the past cannot always be maintained, the time series prediction technique is mainly used for the mid-term prediction. The time series prediction method has an advantage that relatively accurate prediction can be made even with a small amount of data. Time series prediction techniques include moving average method, exponential smoothing method, least squares method, and Box-Jenkins method [19]. Systematic techniques are methods for collecting and predicting information based on automated systems. Systematic techniques include system dynamics, which models the chain causal relationship between variables, and analysis of the change process through simulation, and a method using machine learning algorithms such as artificial neural networks [20-25].

In this study, we use past sales data and various information collected from the website to identify the factors affecting sales to the book and forecast demand. Therefore, we made various prediction models using quantitative and systematic techniques and compared the accuracy of each prediction model. The multiple linear regression model with multiple independent variables was used as the base model to reflect various factors affecting demand. In order to select a forecasting model with high accuracy of demand forecasting, a prediction model was established using Gaussian Process Regression, Support Vector Machine Regression, and Random Forest algorithm.

Demand Forecasting Procedures and Methods

3.1 Demand Forecasting Process

The purpose of this study is to identify the factors affecting the sale and return of specific books and to create a model to forecast sales demand. To do this, we first identify the factors that affect the sale and return of books, and develop a predictive model for predicting book demand for books using historical sales data and factors affecting sales.

We establish a hypothesis to find factors affecting the sale and return of books, collect data necessary for hypothesis testing, and analyze the correlation between the number of sales and the number of returns. The variables used for prediction were used after standardization. The prediction models used in this study are Linear Regression, Gaussian Process Regression, Support Vector Machine Regression, and Random Forest. Finally, to select the best fit model, we compared the accuracy of each predictive model.

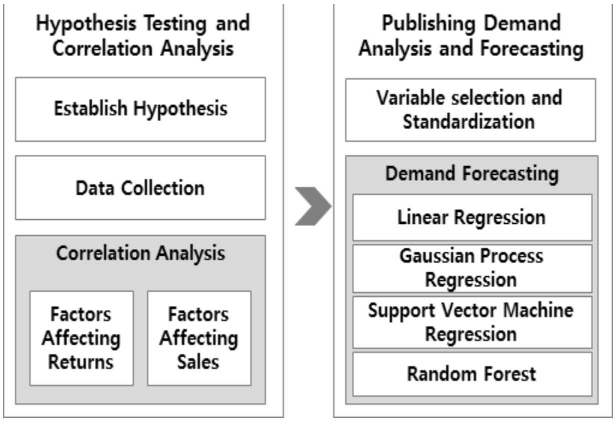
Hypothesis Testing and Correlation Analysis
Establish Hypothesis
Data Collection
Correlation Analysis
Factors Affecting Returns
Factors Affecting Sales
Publishing Demand Analysis and Forecasting
Variable selection and Standardization
Demand Forecasting
Linear Regression
Gaussian Process Regression
Support Vector Machine Regression
Random Forest

Figure 1. The Process of the Demand Analysis and Forecasting Research.

3.2 Demand Forecasting Algorithm

3.2.1 Linear Regression Model

We selected the linear regression model, which is a statistical method, as the first model for predicting sales demand. The Linear Regression Model is a model for predicting dependent variables by combining several independent variables. If the data has linearity, the Linear Regression Model alone can produce good results. Linear regression is a regression technique that models the linear correlation between dependent variable y and one or more independent variables X. Simple linear regression is based on one explanatory variable, and multiple linear regression is based on two or more explanatory variables. Linear regression models the regression equation using a linear prediction function, and unknown parameters are estimated from the data. This regression equation is called a linear model. The developed linear regression equation can be used to predict y for x values without y.

The simplest form of linear regression is the relationship between one scalar independent variable x and one scalar dependent variable y. This is called simple linear regression. Multiple linear regression is an extension of several independent variables in a simple linear regression model. Most of the problems in the real world include multiple independent variables, so multiple linear regression techniques are generally used. The multiple linear regression is expressed as:

$$\mathrm{Y} = \alpha + \beta 1\,\mathrm{X1} + \beta 2\,\mathrm{X2} + \cdots . + \beta \mathrm{k}\,\mathrm{Xk} + \varepsilon \qquad (1)$$

The data used in the polynomial regression model consists of n observations for the dependent variable Y and k predictive variables X1, X2, ..., Xk. Where X1, X2, ..., Xk are input or predictive variables. The relationship

between Y and X1, X2, ..., Xk is shown in Equation (1). Where α, β1, β2, ..., βk are the coefficients of each independent variable and are constants called regression coefficients. ε is a random variable that is not observed as an error term and means the error between the X variable and the Y variable that can not be explained by the regression coefficient. The multiple linear regression model with multiple values is shown in Equation 2

$$\mathrm{Yi} = \alpha + \beta 1\,\mathrm{X1i} + \beta 2\,\mathrm{X2i} + \cdots . + \beta \mathrm{k}\,\mathrm{Xki} + \varepsilon \mathrm{i} \qquad (2)$$

Where xi1, xi2, ..., xip represent the values of the I-th predictor for X1, X2, ..., Xk, respectively. And εi represents the approximate error of the dependent variable yi. In the standard linear regression analysis model, if there is a correlation between the independent variables (predictive variables), multi-collinearity occurs and the accuracy of the regression model decreases. Multicollinearity is when a part of an independent variable can be represented by a combination of other independent variables. It occurs when the independent variables are not independent of one another but the cross-correlation is strong. If the independent variables are dependent on each other, the so-called over-fitting problem occurs and the stability of the regression results is deteriorated.

The most basic way to eliminate multicollinearity is to eliminate variables that depend on other independent variables. The Variance Inflation Factor (VIF) can be used to select the most dependent independent variables.

VIF is the linear regression of independent variables to other independent variables. The more dependent on the other variables, the larger the VIF. Generally, if the VIF exceeds 10, it is determined that multi-collinearity occurs. In this study, if GVIF ^ (1 / (2 * Df)) value is more than 4, it is interpreted that there is multi collinearity. In order to improve the prediction accuracy of the regression model, the optimal model was selected using the Stepwise method and then the multi - collinearity test was performed to complete the final model. Stepwise regression is a method of fitting regression models in which the choice of predictive variables is carried out by an automatic procedure. In each step, a variable is considered for addition to or subtraction from the set of explanatory variables based on some prespecified criterion.

3.2.2 Gaussian Process Regression

In order to compare the prediction accuracy with the linear regression model, a prediction analysis using the Gaussian process regression which can detect the nonlinearity was additionally performed. The Gaussian process regression model is an analytical method that infer the regression function based on the preliminary distribution of the unknown regression function as the Gaussian probability process in the Bayesian regression model. The Gaussian process obtains the distribution of the relational function f of the data set X, Y with nonlinear regression relation. When we know this distribution, the estimate for Y 'for new data X' is the mean of the distribution for function f. Gaussian Process Regression is a kind of supervised learning. That is, after tracing the labeled data, it estimates the label corresponding to the new data when it arrives. It is possible to generate a predictive model quickly and accurately even when there is a large amount of learning data.

3.2.3 Support Vector Machine Regression

SVM (support vector machine) is an algorithm that can handle both regression and classification. Support Vector regression adjusts to place the data as much as possible within a certain range from the straight line and its straight line. In some cases, there is data that can not be linearly regressed. For this purpose, SVM can map low dimensional data to high dimensional space through 'kernel' technique and then find the hyperplane in the high dimensional space and regress. The support vector machine (SVM) is the field of machine learning. It is a supervised learning model for pattern recognition and data analysis. It is mainly used for classification and regression analysis. Given a set of data separated into two categories, the SVM algorithm builds a non-stochastic binary linear classification model that determines which categories the new data belongs to. The generated classification model is represented as a boundary in the space where data is mapped. The SVM algorithm is an algorithm for finding the boundary having the largest width.

3.2.4 Random Forest

Random forest is a type of ensemble learning method used for classification and regression analysis in machine learning. Random forests classify new values or calculate average predictions using multiple decision trees constructed during data training. The Decision Tree constructs the tree to produce optimal results with optimal selection at each moment. If the depth of the decision tree is made large, overfitting is likely to occur. Random forest is a technique to prevent overfitting. Create multiple Decision Trees and generate the results in an ensemble of the trees. Each tree learns only some of the randomly selected data from all features. In the case of the linear regression model, when there are many variables used in the prediction model generation, the correlation between the variables is high, so that multiple lines can be issued. In case of the random forest, since the variables are randomly selected, can do. Random Forest has created a number of trees that simplify the model, reducing the likelihood of overfitting and enhancing predictive performance in an ensemble fashion. The Random Forest also has the advantage of being able to selectively use variables by calculating the importance of the variables.

DEMAND FORECASTING MODEL CREATION AND EVALUATION

4.1 Preparing Verification Data

For the experiment, we used the sales status data of the books sold by A publishing company. In order to identify the factors influencing the sales and return of the books, additional data related to the books were collected from Internet portal system and SNS site. A total of 386 volumes of sales data of the books were used, 256 of which were used as learning data for the prediction model generation and 130 were used for the prediction accuracy test of the model. The dependent variable used in the forecasting was sales volume by book and the independent variable was selected by using independent variables that affect the sale and return of books. Hypotheses were constructed to select independent variables and validated through testing. In other words, hypotheses about factors affecting the sale and return of books were established, and variables necessary for hypothesis testing were collected from three web pages A, B, and C. The variables used in the demand forecasting model were selected through hypothesis testing. The variables used in the experiment consist of a total of 29 variables including continuous variables and categorical variables.

TABLE 1. DEPENDENT VARIABLES AND INDEPENDENT VARIABLES USED IN THE PREDICTION MODEL

No	Variables	Description
1	Number_of_Publications	Number of copies
2	Number_of_returns	Number of returned books
3	Return_rate	Ratio of number of returns to publications
4	Year	The year the book was published
5	Category	Field of books
6	Series	Series of serialized books
7	Subject	Book Topics
8	BestSeller	Whether it's a bestselling book
9	Price	The price of the book
10	Popular_authors	Whether it's a popular writer
11	Number_of_Reviews_A	The number of reviews in this book published on website A.
12	Number_of_Search_A	Number of searches for this book on website A
13	Score_of_Reviews_A	The average score of reviews in this book published on website A.

14	Number_of_Posts_B	Number of posts for this book on website B
15	Number_of_Views_B	Sum of views on this book's posts for this book on website B
16	Number_of_Search_B	Number of searches for this book on website B
17	Number_of_Posts_C	Number of posts for this book on website C
18	Number_of_Views_C	Sum of views on this book's posts for this book on website C
19	Number_of_Search_C	Number of searches for this book on website C
20	Sum_of_Search_ABC	Sum of searches for this book on website A, B, C
21	Avg_of_Search_ABC	Averge of searches for this book on website A, B, C
22	Past_Year_Youth_population	the number of Youth people who can read books in the previous year, Population aged from 10 to 29
23	Past_Year_Middleage_population	the number of Middle-aged people who can read books in the previous year, Population aged from 30 to 59
24	Past_Year_Elderly_population	the number of Eldered people who can read books in the previous year, Population aged from 60 to 79
25	This_Year_Youth_population	the number of Youth people who can read books in this year, Population aged from 10 to 29
26	This_Year_Middleage_population	the number of Middle-aged people who can read books in this year, Population aged from 30 to 59
27	This_Year_Elderly_population	the number of Eldered people who can read books in this year, Population aged from 60 to 79
28	Previous_year_population	Previous year's population to read books, Population aged from 10 to 79
29	This_year_population	This year's population to read books, Population aged from 10 to 79

4.2 Analysis of Influencing Factors affecting Sales and Returns

Recently, books are being purchased more and more through Internet bookstores. Research shows that even if offline bookstores are used, the purchasing rate of books that are well-read by readers is high. We set up some hypotheses about the sale and return of the book, taking into account that readers will obtain information about the book through consumer reviews, purchase reviews, and postings on the website and online bookstores, and will decide the purchase intention of the book.

Hypothesis 1). The greater the number of posts, reviews, and searches on a portal site and an online bookstore, the more sales will increase for that book and the fewer the returns will be.

Hypothesis 2) As the number of people who read books increases, sales volume for books will increase and returns will decrease.

Hypothesis 3) If a writer who writes a book is a famous writer or a popular writer, sales volume for books will increase and returns will decrease.

The data needed to verify established hypotheses were collected through web page crawl and search. Hypothesis testing was performed through correlation analysis between net deliveries and return rates. Correlation coefficient is a measure of the degree of correlation between two variables, X and Y, and refers to the relationship between two variables as one of the two variables increases or decreases as the other increases or decreases. Correlation coefficients can be expressed in numerical values different from causal relations. If the absolute value of the correlation coefficient is closer to 1, the correlation is stronger. If the correlation coefficient is larger than 0, the correlation is positive. If the correlation coefficient is smaller than 0, the correlation is negative. The absolute value of the correlation coefficient represents the correlation between two variables, and generally divides the degree of correlation by 0.5. In this study, if the absolute value of the correlation is greater than 0.5, the hypothesis is adopted; if it is less than 0.5, the hypothesis is rejected.

4.2.1 Test for Hypothesis 1

In order to test the hypothesis No.1, three sites among the portal and online bookstore sites were selected for data collection, and the crawl program was used to collect data for each book. For each site, data collected on the number of posts, views, searches, reviews and review scores, and search volume per book were standardized, and correlation coefficients were calculated for each of the sales quantity and return rate. In statistics, standardization refers to the process of transforming the value of a variable such that the mean is 0 and the standard deviation is 1. By standardization, the data collected at each site have different ranges, which makes it difficult to compare. Each data collected from websites A, B, and C was standardized and grouped into 11 variables. Table 2 shows the correlation coefficient between each variable and the number of shipments. Although all 11 variables are slightly different, all correlation coefficient values are larger than 0.5, indicating that book sales have a significant impact on sales. However, these variables have a weak negative relationship to returns.

TABLE 2. COMPARISON RESULTS OF ACCOMMODATIVE RESPONSE BETWEEN 2.50 D AND HOLOGRAM STIMULUS

Variables	**Correlation coefficient for Number_of_Publicatios**		**Correlation coefficient for Number_of_Returns**	
Number_of_Reviews_A	0.74	The correlation coefficient is larger than 0.5, which is considered to have a significant effect on	- 0.31	Since the correlation coefficient is less than 0.5, it is considered to
Number_of_Search_A	0.71		- 0.24	

Score_of_Reviews_A	0.70	sales of the book	-0.25	have no significant impact on book returns.
Number_of_Posts_B	0.68		-0.24	
Number_of_Views_B	0.71		-0.24	
Number_of_Search_B	0.66		-0.29	
Number_of_Posts_C	0.61		-0.27	
Number_of_Views_C	0.61		-0.26	
Number_of_Search_C	0.54		-0.26	
Sum_of_Search_ABC	0.61		-0.28	
Avg_of_Search_ABC	0.66		-0.28	

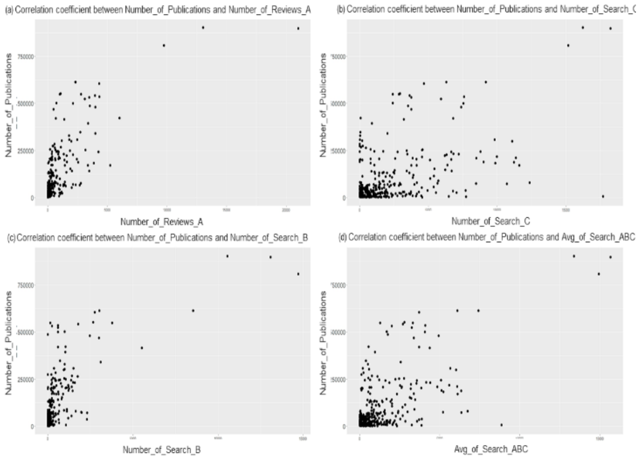

Figure 2. Graph of correlation coefficient for sales and returns.

(a)Correlation coefficient between

Number_of_Publications and Number_of_Reviews_A

(b)Correlation coefficient between

Number_of_Publications and Number_of_Search_C

(c)Correlation coefficient between

Number_of_Publications and Number_of_Search_B

(d) Correlation coefficient between

Number_of_Publications and Avg_of_Search_ABC

4.2.2 Test for Hypothesis 2

In the case of a specific book, sales will vary depending on the number of readers who read the book. It is obvious that demand for books will vary depending on the number of school-age population, especially in the case of study or reference books. Based on this fact, the number of populations by age group was collected by the National Statistical Office and the correlation between sales and returns was calculated.

The population used the voluntary 10- to 79-year-old domestic population data and calculated the correlation coefficient between the previous year's population and this year's population to see if it relates to yearly sales. As a result of the calculation, it was found that there was a weak correlation with 0.428 for the previous year's population and 0.419 for this year's population. As a result of calculating the correlation coefficient with the return, it was found that there is a negative correlation with -0.526 for the previous year's population and -0.537 for this year's population.

TABLE 3. THE RESULTS OF POPULATION-BASED SALES AND RETURN-RELATIONSHIP TESTS

Variables	**Correlation coefficient for Number_of_Publicatios**		**Correlation coefficient for Number_of_Returns**	
Previous_year_population	0.4.	Since the correlation coefficient is less than 0.5, it is considered to have no significant impact on sales of the book	-0.53	Because the absolute value of the correlation coefficient is greater than 0.5, it is considered to have a significant impact on book returns.
This_year_population	0.42		-0.54	

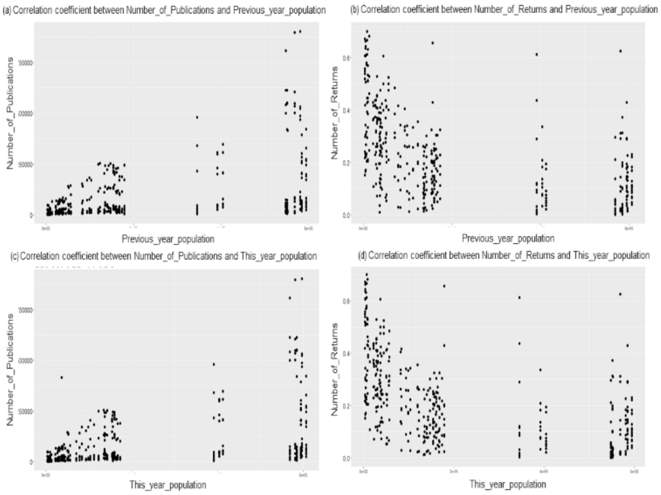

Figure 3. Graph of population-based correlation coefficient for sales and returns.

(a)Correlation coefficient between

Number_of_Publications and Previous_year_population

(b)Correlation coefficient between

Number_of_Returns and Previous_year_population

(c)Correlation coefficient between

Number_of_Publications and This_year_population

(d)Correlation coefficient between

Number_of_Returns and This_year_population

4.2.3 Test for Hypothesis 3

To verify the hypothesis that books written by famous authors or those written in the past are best sellers, we collected data on the popularity of authors, celebrities, and bestseller authors at internet bookstores. As a result of the correlation, it was found that the relation between the sales of the book and the return was 0.146 and 0.025, respectively.

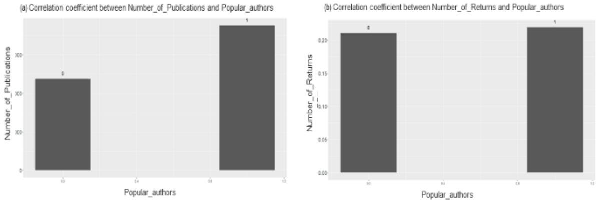

Figure 4. Graph of Correlation coefficient between author's popularity and sales and returns.

(a) Correlation coefficient between

Number_of_Publications and Popular_authors

(b) Correlation coefficient between

Number_of_Returns and Popular_authors

4.3 Sales Demand Forecasting Model

4.3.1 Sales Demand Forecasting Model Using Linear Regression

Linear Regression is a model that predicts dependent variables by combining various independent variables with traditional statistical techniques. In this study, all the data collected during the hypothesis testing step were selected as independent variables candidates. Then, the optimal parameters were selected using the Stepwise method, and the final parameters were selected for the multi - collinearity test. Finally, a regression equation for predicting book sales demand was derived using the selected variables. The dependent variable used in constructing the regression model is the number of sales for the book, and the original data does not follow the normal distribution, so the log is taken to be transformed to follow the normal distribution. The variables selected by the Stepwise method are as follows.

TABLE 4. THE FINALLY SELECTED INDEPENDENT VARIABLES

No	Variables	Description
4	Year	The year the book was published
5	Category	Field of books
6	Series	Series of serialized books
7	Subject	Book Topics
11	Number_of_Reviews_A	The number of reviews in this book published on website A.
12	Number_of_Search_A	Number of searches for this book on website A
14	Number_of_Posts_B	Number of posts for this book on website B
16	Number_of_Search_B	Number of searches for this book on website B
17	Number_of_Posts_C	Number of posts for this book on website C
29	This_year_population	This year's population to read books, Population aged from 10 to 79

The regression equations before and after the stepwise selection method were compared using the coefficient of determination R2. The decision coefficient R2 is a measure of the linear relationship between the independent variable and the dependent variable. In other words, the coefficient of determination means the rate at which the independent variable X describes the dependent variable Y, and 1 - R2 represents the rate of variation derived from other factors or chance other than the independent variable X. The modified coefficient of determination before applying the stepwise selection method is 0.9591, and the modified coefficient of determination after application is 0.9614, which means that the fitness is higher.

Next, we checked whether there is multicollinearity among the variables selected by the stepwise selection method. In this study, if GVIF ^ (1 / (2 * Df)) is more than 5, it is interpreted as having multi - collinearity.

TABLE 5. RESULT OF THE MULTI-COLLINEARITY TEST

No	Variables	Variables	Description
1	Year	1.344	
2	Category	1.107	
3	Series	1.455	
4	Subject	3.523	
6	Number_of_Reviews_A	7.212	Exclude from a variable
7	Number_of_Search_A	7.001	

			Exclude from a variable
8	Number_of_Posts_B	2.795	
9	Number_of_Search_B	4.380	
10	Number_of_Posts_C	3.412	
11	This_year_population	3.918	

Finally, selected variables are divided into continuous variables and categorical variables. In the case of categorical variables, the variables are subdivided according to the category value, and 38 variables are used in the final demand forecast regression model. The table below shows each variable and regression coefficient of the final model.

TABLE 6. RESULT OF THE MULTI-COLLINEARITY TEST.

Independent variables	Description	Regression coefficient
x1	Year	-0.06
x2	This_year_population	0,24
x3	Number_of_Posts_B	0.01
x4	Number_of_Search_B	0.09
x5	Number_of_Posts_C	0.09
x6	first Category	0.48
x7	second Category	0.73
x8	third Category	0.97
x9	fourth Category	0.33
x10	fifth Category	0.66
x11	sixth Category	0.14
x12	seventh Category	0.38
x13	eighth Category	0.78
x14	ninth Category	0.86
x15	tenth Category	0.33
x16	eleventh Category	0.71
x17	twelfth Category	0.18
x18	zerot Subject	0.35
x19	first Subject	0.87

Using the regression coefficients in Table 6, the final regression equation can be expressed as:

$$y = 3.75 - 0.06x_1 + 0.24x_2 - 0.01x_3 + 0.09x_4 - 0.09x_5 + \cdots + 0.64x_{38} + e \quad (3)$$

The fit (R2) of the final regression model is 0.9582, and the significance probability of the individual variables is generally close to zero, indicating that the selected variable is suitable for the model. Here, the significance probability is a criterion for judging whether the individual variable is suitable for the regression model. The range of significance probabilities is from 0 to 1, meaning that if the significance level is less than 0.05, the variable is considered to be appropriate for the regression model, and if it is greater than 0.05, it is considered to be less suitable for the regression model.

In order to investigate the model defects of the final regression equation, we compared the Residuals vs Fitted plots of the final regression equation. In Fig. 5, the residuals are distributed horizontally evenly on the horizontal line with Residuals = 0, so that the relationship between the dependent variable and the predicted variable has a linear relationship and the variance of the error term is the same. We looked at the Normal Q-Q plot to check the normality of the residuals. If the residuals satisfy the normality, the points on the graph are on a straight line at a 45-degree angle. Fig. 5 shows the normal Q-Q plot of the final regression equation. As shown in Fig. 5, it can be seen that the points are on a 45-degree straight line, but the shapes are disarranged toward both ends. Thus, we cannot be certain that the residuals are perfectly normal

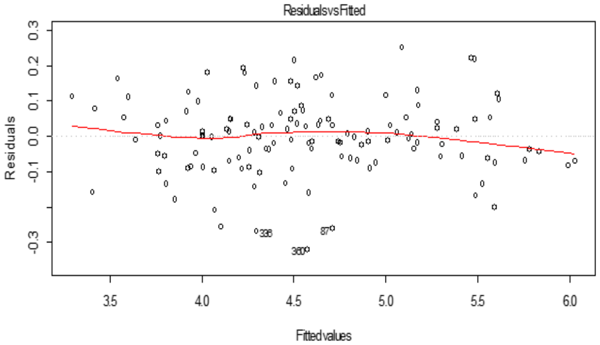

Figure 5. Residuals vs fitted plot of the final regression equation.

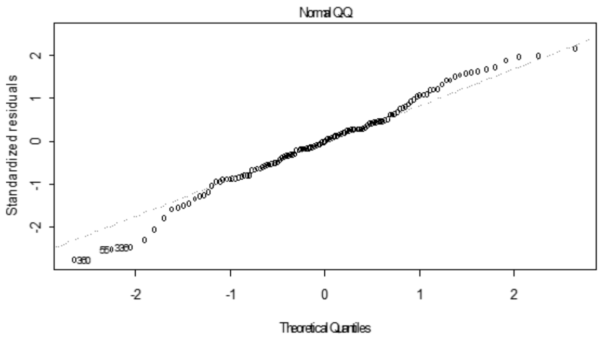

Figure 6. Normal Q-Q plot of the final regression equation.

4.3.2 Sales Demand Forecasting Model Using Gaussian Process Regression Model

We have developed a model of sales demand forecasting of a book using machine learning method which can search nonlinearity beyond linear search. The variables used were the entire data set collected during the hypothesis verification process. In this study, Gaussian process regression model was established using isotropic kernel. Since the number of learning data is small at only 300, 5-folds cross validation is used to prevent overfitting. The mean square root error (RMSE) of the final model was 0.09, and the error value was low. The R2 was 0.97, which was relatively high. RMSE is the mean square root mean square error. The lower the value, the more accurate the prediction model. Fig 7. (a) is a graphical representation of the Gaussian process regression results. This graph is a net output forecast value graph for actual net deliveries. It is judged that the closer the point is to the straight line of y = x shape, the more accurate the prediction is.

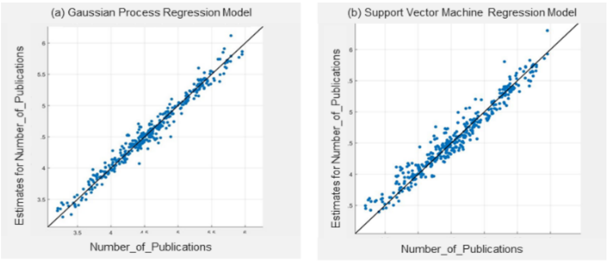

Figure 7. Comparison Graph of the Actual and Forecasted Values.

(a) Gaussian Process Regression Model

(b) Support Vector Machine Regression Model

4.3.3 Sales Demand Forecasting Model Using Support Vector Machine

SVM is an algorithm that can handle both regression and classification. Support Vector regression adjusts to place the data as much as possible within a certain range from the straight line and its straight line. However, there are cases where it is not possible to linearly regress according to the data. For this purpose, SVM can map low dimensional data to high dimensional space through 'kernel' technique and then find the hyperplane in the high dimensional space and regress. In this study, we used a regression algorithm of Support Vector Machine to model the sales demand of books. The mean square root error (RMSE) of the final model was 0.14, and the error value was low. The R2 was 0.95, which was relatively high. Fig7. (b) is a graphical representation of the predicted results of the SVM regression model. This graph is a net output forecast value graph for actual net deliveries. It is judged that the closer the point is to the straight line of y = x shape, the more accurate the prediction is.

4.3.4 Sales Demand Forecasting Model Using Random Forest

Random forest is a tree method that can prevent overfitting. Create multiple Decision Trees and select the results with the majority of the trees (ensemble). Here, each tree learns only a part of the entire feature at random. In the GPR or SVM, we used the variables selected through the Stepwise method and the multi-collinearity test in the regression analysis part. However, since Random Forest can calculate the importance of the variables themselves, we used all the collected variables. The explanatory power (R2) of the random forest model is 0.93. The predicted accuracy of the random number with random forest is 98.1%.

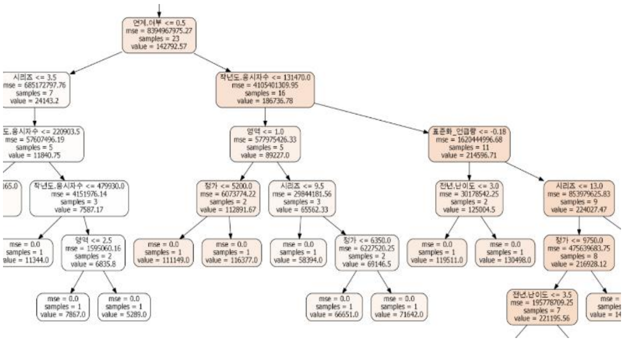

Figure 8. Part of the fourth tree in the random forest.

4.4 Comparison and Evaluation of Sales Demand Forecast Model

The prediction models used in this study are Linear Regression, Gaussian Process Regression (GPR), Support Vector Machine (SVM), and Random Forest. The learning data used to generate the forecasting model is 256 books sold in 2014 and 2015, and the test data used to evaluate the accuracy of the model is 130 books in 2016. A total of 29 variables were used for the variables used in the prediction, including the continuous variables and the categorical variables collected during the hypothesis testing. The prediction model and accuracy are shown in Table 7.

TABLE 7. THE COMPARISON OF ESTIMATION RESULTS OF FORECAST MODELS

Forecast model	Explanation(R^2)	Net sales Value	Predicted value	Accuracy
Linear Regression	0.96	11,603,352	11,953,805	97.0%
Gaussian Regression	0.93	11,603,352	11,214,397	96.6%
SVM Regression	0.87	11,603,352	11,017,538	94.9%
Random Forest	0.93	11,603,352	11,379,641	98.1%

CONCLUSIONS

In the linear regression model, the explanatory power and prediction accuracy of the model are superior to those of the other models. The accuracy of the random forest model, which is a machine learning method, is the best.

Demand forecasting is the activity of predicting the future using this model by establishing a model that can grasp trends using historical data. Using data analysis to predict accurate publishing demand and identify factors that affect sales and factors that cause returns, we can reduce losses and increase sales from inventory management and returns.

The purpose of this study is to identify the factors affecting the sale and return of specific books and to create a model to forecast sales demand. For this purpose, wd used the sales data of the books sold for 3 years (2014 ~ 2016) by A publishing company. In addition, we collected the data related to books in the Internet portal system and SNS site. In this study, we hypothesized the factors affecting the sale and return of books. The variables needed for hypothesis testing were collected from web pages and SNS sites, and correlation analysis was performed to confirm whether the variables were related to actual demand forecasting.

We also developed a forecasting model of book sales by using linear regression and machine learning algorithms in artificial intelligence.

The results of this study can be expected as follows. First, return loss cost decreases. Using company internal data will allow us to design a more accurate forecasting model and perform more accurate forecasting, which will reduce the cost of loss due to returns when used in publishing projects. Second, it can support the publishing and sales process. By analyzing various data, it is possible to understand the tendency of consumers (students, etc.) and suppliers (distributors, etc.). You will be able to identify your favorite textbooks and use them for marketing, and in some cases, see what returns are happening to improve your publishing and sales processes more efficiently. Third, it can be used to forecast demand for various services. Based on the basic model proposed in this study, various services can be extended to the demand forecasting model. Using this information, we can find out the possibility of improving service by utilizing it in various business of company as well as publishing business.

In this study, it is found that if the additional factors affecting book sales can be secured, basic prediction model can be used to make a practical prediction model. In fact, the Random Forest used in this study took about 2 seconds to learn. As the number of features increases and the number of books to be analyzed increases, the learning time will also increase, but the reliability of the model will increase and it will be applicable to the real industry. Because the available data is limited in the study, the scope of this study was limited to forecasting the sales demand of some books. If we apply the proposed analytical procedure and method directly from the company, we can expect better prediction results. It is also expected to be applicable to various business processes of book publishing or sales companies.

ACKNOWLEDGMENT

This research was supported by Hankuk University of Foreign Studies Research Fund (of 2020)

This research was supported by National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. 2018R1A5A7023490).

REFERENCES

[1] Schaer, O., et al., Demand forecasting with user-generated online information. International Journal of Forecasting (2018), https://doi.org/10.1016/j.ijforecast.2018.03.005.

[2] Fildes, R., & Ord, J. K. (2002). Forecasting competitions: Their role in improving forecasting practice and research. In M. P. Clements, & D. F. Henry (Eds.), A Companion to Economic Forecasting (pp. 322–353).

[3] Geva, T., Oestreicher-Singer, G., Efron, N., & Shimshoni, Y. (2017). Using forum and search data for sales prediction of high/involvement products. MIS Quarterly, 41(1), 65–82.

[4] Elshendy, M., Colladon, A. F., Battistoni, E., & Gloor, P. A. (2018). Using four different online media sources to forecast the crude oil price. Journal of Information Science, 44(3), 408–421.

[5] Fantazzini, D., & Toktamysova, Z. (2015). Forecasting German car sales using Google data and multivariate models. International Journal of Production Economics, 170, Part A, 97–135.6. Schneider, M. J., & Gupta, S. (2016). Forecasting sales of new and existing products using consumer reviews: A random projections approach. International Journal of Forecasting, 32(2), 243–256.

[6] Cao, L. and Tay, F. E. H. (2001). Financial forecasting using support vector machines. Neural Computing and Applications, 10, 184-192.

[7] Byeong Won Geun. (2010). Changes in e-book and publishing markets. Artist World, 22 (2), 371-384.

[8] Cassidy, William B. & Reynolds Hutchins(2016), “The Amazon Effect,” Journal of Commerce, 17(7), 10-1

[9] Seung - Hee Baek, Soo - Yeon Son, Ju - young Lee, (2015). A Study on the Influence of Purchasing of books on Internet Bookstore Review. Proceedings of the Korea Information Management Society,, 109-114.

[10] Boon Do Jeong, Mi Seon Hong, "The Effect of Online Shopping Business on Repurchasing Intention and Oral Intention of Korean Consumers - Focused on AMAZON.COMBoon", The e-Business Studies Volume 19, Number 1, February, 2018 : pp. 39~53

[11] Hyun-Chul Kim. (2018). "A Study on the effect of perceived online shopping mall attribute on trust, commitment, purchasing intention". Journal of the Korea Computer Information Society, 23 (9), 123-132.

[12] Ok Ryun Jung, "Exploring the social factors related to the classification of literature and the selection of reading material " Reading Research Vol. 26, No. 0, Korea Reading Society 2011

[13] Cho, Sin-Hee·Yi, Mun Yong, "Business Implications of the Factors that Determine Online Review Helpfulness", Entrue Journal of Information Technology 13.1 (2014): 29-40.

[14] Babic Rosario, A., Sotgiu, F., De Valck, K., & Bijmolt, T. H. (2016). The effect of electronic word of mouth on sales: A meta-analytic review of platform, product, and metric factors. Journal of Marketing Research, 53(3), 297–318.

[15] Dellarocas, C., Zhang, X. M., & Awad, N. F. (2007). Exploring the value of online product reviews in forecasting sales: The case of motion pictures. Journal of Interactive Marketing, 21(4), 23–45.

[16] S. Park, et. al., 2012, "SERI Issue Paper: Effective Sales Predict and Examples," Samsung Economic Research Institute, [Online] Available: https://www.slideshare.net/girujang/seri20120303.

[17] Myers, Raymond H., and Raymond H. Myers. Classical and modern regression with applications. Vol. 2. Belmont, CA: Duxbury press, 1990.

[18] Gabor MR, Dorgo LA. Neural Networks Versus Box-Jenkins Method for Turnover Forecasting: a Case Study on the Romanian Organisation. Transformations in Business and Economics. 2017;16(1):187–211.

[19] Zhang G, Eddy Patuwo B, Hu Y M. Forecasting with artificial neural networks: The state of the art. International Journal of Forecasting. 1998;14(1):35–62.

[20] Robinson C, Dilkina B, Hubbs J, Zhang W, Guhathakurta S, Brown MA, et al. Machine learning approaches for estimating commercial building energy consumption. Applied Energy. 2017;208(Supplement C):889–904. https://doi.org/10.1016/j.apenergy.2017.09.060.

[21] Voyant C, Notton G, Kalogirou S, Nivet ML, Paoli C, Motte F, et al. Machine learning methods for solar radiation forecasting: A review. Renewable Energy. 2017;105(Supplement C):569-582. https://doi.org/10.1016/j.renene.2016.12.095.

[22] Kock AB, Teräsvirta T. Forecasting Macroeconomic Variables Using Neural Network Models and Three Automated Model Selection Techniques. Econometric Reviews. 2016;35(8–10):1753–1779.

[23] Schölkopf B, Smola AJ. Learning with kernel: Support Vector Machines, Regularization, Optimization and Beyond. The MIT Press; 2001. Available from: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.167.5140{&}rep=rep1{&}type=pdf.

[24] Rasmussen CE, Williams C. Gaussian Processes for Machine Learning. The MIT Press; 2006.

[25] Ahn, H. C., Kim, K. J. and Han, I. G. (2005). Purchase prediction model using the support vector machine. Journal of Korea Intelligent Information Systems Society, 11, 69-81.

FIGURE CAPTIONS

Mee Hwa Park Software Focused University, Korea. Areas of Interest: Big Data Analytics, IoT Service, Multimedia Database.

Jong Sup Lee Development of smart community policing system(Googi) Center, Dongguk University-Seoul, Korea. His Ph.D. degree in computer science from Daejeon University, Korea. He is currently a Research Professor of Dongguk University, Korea. His interests are C4I, Big Data, NUI/NUX, RFID/USN, IoT etc

Ill Chul Doo Hankuk University of Foreign Studies, Korea. His Ph.D. degree in Digital Culture & Contents from Hanyang University, Korea. He is currently a Professor at Hankuk University of Foreign Studies, Korea. His interests are mobile contents, and cultural industry, and cultural technology.