Communications for Statistical Applications and Methods
2023, Vol. 30, No. 5, 485–499

https://doi.org/10.29220/CSAM.2023.30.5.485
Print ISSN 2287-7843 / Online ISSN 2383-4757

Comparative analysis of model performance for predicting the customer of cafeteria using unstructured data

Seungsik Kim[a], Nami Gu[b], Jeongin Moon[c], Keunwook Kim[d], Yeongeun Hwang[e], Kyeongjun Lee[1,f]

[a]Department of Statistics, Kyungpook National University, Korea;
[b]Department of Statistics, Pusan National University, Korea;
[c]Department of Statistics, Yeungnam University, Korea;
[d]Daegu Digital Innovation Agency, Big Data Utilization Center, Korea;
[e]Industrial Complex Promotion Department, Korea Industrial Complex Corporation, Korea;
[f]Department of Mathematics and Big Data Science, Kumoh National Institute of Technology, Korea

Abstract

This study aimed to predict the number of meals served in a group cafeteria using machine learning methodology. Features of the menu were created through the Word2Vec methodology and clustering, and a stacking ensemble model was constructed using Random Forest, Gradient Boosting, and CatBoost as sub-models. Results showed that CatBoost had the best performance with the ensemble model showing an 8% improvement in performance. The study also found that the date variable had the greatest influence on the number of diners in a cafeteria, followed by menu characteristics and other variables. The implications of the study include the potential for machine learning methodology to improve predictive performance and reduce food waste, as well as the removal of subjective elements in menu classification. Limitations of the research include limited data cases and a weak model structure when new menus or foreign words are not included in the learning data. Future studies should aim to address these limitations.

Keywords: cafeteria, ensemble model, ESG, food waste, machine learning, menu features, performance improvement, prediction, word embedding

1. Introduction

Accurately predicting the number of diners in a cafeteria is essential for optimal production and cost management. Inaccurate predictions can lead to problems with ingredient supply and storage, decreased customer satisfaction due to inefficient employees, and loss of profits (Cheng *et al.*, 2003). However, many cafeteria operators rely on subjective experience to make estimations due to budget constraints and a lack of experts (Lim, 2016; Jeon *et al.*, 2019). To address this issue, research is needed to objectively and quantifiably anticipate the number of customers, considering the development of big data and artificial intelligence methodology.

Estimating the exact number of diners in a cafeteria is complex due to various variables. Firstly, there is no absolute standard, since the variables affecting the count of diners are diverse and vary

[1] Corresponding author: Department of Mathematics and Big Data Science, Kumoh National Institute of Technology, 61 Daehak-ro, Gumi, Gyeongbuk 39177, Korea. E-mail: leekj@kumoh.ac.kr

Published 30 September 2023 / journal homepage: http://csam.or.kr

depending on the restaurant type, geographical location, and characteristics of the surrounding area. Secondly, individual social activities change over time, making it difficult to update data and models. Finally, since restaurants predominantly manage unstructured data, including text, rather than structured data, researchers require expertise and skills in the preprocessing process (Jeon *et al.*, 2019). Given these challenges, accurately predicting the count of diners necessitates overcoming them through advanced data analysis and modeling methodologies.

Studies on predicting the number of customers can be categorized into statistical models and machine learning models based on the methodology. In most previous studies, variables such as the menu, date, weather, etc., were identified as major predictors in estimating the volume of meals. Additionally, studies have utilized unstructured data to classify menu ingredients and recipes into the Dewey decimal classification (DDC). The Dewey decimal classification system was devised in 1873 to organize the collections and catalogs of the Amherst university library in the United States, and it is currently the most widely used classification system worldwide. However, this approach has limitations, as it involves subjective judgment from the researcher. Therefore, this study aims to minimize the researcher's partial intervention and improve performance by applying the word embedding methodology, commonly used in text mining.

The composition of this paper is as follows. First, we will review previous studies related to estimating the count of customers in the cafeteria to derive significant variables and generate derived variables on unstructured data. Next, we will build a predictive model, validate and compare it with other models. Then, we will present the results and implications and discuss the study's limitations.

The results are expected to be used as a basis for estimating the number of diners at cafeterias operated by both the public and private sectors. Furthermore, the findings can be utilized as research data for cost reduction in companies and as the basis data for establishing policies to reduce food waste in local governments.

2. Prior research

Studies related to predicting diner counts typically involve identifying significant variables (Cheng *et al.*, 2003; Baek *et al.*, 2007; Lim, 2016) and building a prediction model (Ryu and Sanchez, 2003; Blecher and Yeh, 2008; Jeon *et al.*, 2019). These studies use statistical and machine learning methodologies to determine menu items, dates, and weather as the main variables influencing the estimation of diner counts.

2.1. Factor analysis of variables influencing the number of diners

Cheng *et al.* (2003) surveyed university cafeteria managers to investigate their perception of the importance and current state of estimating the number of diners. Cheng found that the previous month's number of customers, menu preferences, weather, and day of the week were the main variables (Cheng *et al.*, 2003). Baek *et al.* (2007) conducted a satisfaction survey on cafeteria menus targeting industrial workers and collected information on menu preferences and cooking methods by gender. The results showed that jeon, steamed, and deep-fried foods were highly preferred, while unleavened foods were less favored (Baek *et al.*, 2007). Lim (2016) identified variables that affect the prediction of the number of meals for catering staff of consigned food service companies. The study found that the day of the week, menu preference, and the number of meals served in the previous week were major variables in predicting the count of meal attendees across all types of cafeterias (Lim, 2016).

Predicting the number of diners is crucial in cafeteria operation. Previous studies have shown that the cafeteria's menu, the number of customers in the previous month or week, the weather, and the day of the week are the main variables affecting meal attendance. By considering these variables, efficient cafeteria operations can lead to improved customer satisfaction and reduced employee turnover. Therefore, alternative solutions for effective cafeteria management need to be developed.

2.2. A study on a model for predicting the number of diners

Ryu and Sanchez (2003) used data from the Texas Tech cafeteria to find a suitable model for predicting diners. They compared and analyzed various models, including moving average and simple linear regression, and used various evaluation indicators represented by MSE (mean squared error) and MAPE (mean absolute percentage error). As a result, they found that the multiple linear regression model was the most appropriate one (Ryu and Sanchez, 2003). Blecher and Yeh (2008) confirmed the superior performance of the moving average model among time series predicting models like moving average and exponential smoothing using cafeteria data of a large university (Blecher and Yeh, 2008). Jeon *et al.* (2019) developed a model to predict the number of diners at the cafeteria in S city hall by simplifying the complex Korean food menus via DDC. Furthermore, the study utilized both traditional statistical models and machine learning methodologies to compare, analyze, and propose improvements (Jeon *et al.*, 2019). Cetinkaya and Erdal (2019) used an artificial neural network (ANN) model to predict the number of customers in the cafeteria of Kirikkale university by applying the main variables that affect the estimation of the number of meals, such as menus, holidays, and events. The study also compared the performance of the model based on the hidden layer structure (Cetinkaya and Erdal, 2019).

The previous studies reviewed in this paper are presented in Table 1. While previous studies had the advantage of using various variables such as menu preference, day of the week, weather, etc., to predict the count of diners, there is a lack of studies harnessing unstructured data including menu information, except for Jeon *et al.* (2019). Additionally, Jeon *et al.* (2019) have some limitations due to the biased classification of menu ingredients and recipes based on DDC. This study attempts to improve the performance of the model by applying the word embedding methodology to representative unstructured data, like recipes and ingredients of Korean menus, which is a distinctive feature from previous studies.

2.3. A study using Word2Vec

Word2Vec is a powerful tool for converting unstructured data into numerical vectors, which can be used for a variety of machine learning methodologies, including regression, classification, and clustering analysis. Previous studies have shown that using Word2Vec and other word embedding methodologies can improve performance in these areas.

Cho (2019) used Word2Vec to convert categorical variables in hospital treatment data into numerical data and then performed cluster analysis. The results showed that using Word2Vec improved clustering performance based on the Silhouette coefficient.

Park and Byun (2021) proposed an item recommendation system using Word2Vec on the online shopping log data and found that it had a positive impact on recommendation accuracy compared to other methods.

In a study by Park and Lee (2018) on sentiment classification with Korean movie review data, Word2Vec combined with machine learning methodologies improved sentiment classification performance.

In this study, Word2Vec will be used to preprocess unstructured menu information into numerical vectors to estimate the number of diners. By doing so, we hope to improve the accuracy of our

Table 1: Research paper related to variables influencing the number of meals and predictive modeling

Category	Author	Purpose of study	Method	Variables
Analysis of variables influencing the number of meals	Cheng *et al.* (2003)	Estimation of number of meals, current status and influencing factor investigation	*t*-test, ANOVA	Number of diners in the previous month, menu preferences, weather, day of the week, etc.
	Baek *et al.* (2007)	Cafeteria menu satisfaction survey	Descriptive analysis	Food quality, food ingredients, food type, etc.
	Lim (2016)	Analysis of influencing variables for reducing error rate of meals	frequency analysis, *t*-test	Menu preference, number of meals in the previous week, weather, day of the week, etc.
Predictive modeling of diners	Ryu and Sanchez (2003)	Exploring the best model for predicting diners	Naive, Moving average, Linear regression	Number of people in the time series
	Bleacher and Yeh (2008)	Explore the most accurate time-series models for predicting diners	Naive, Moving Average, Exponential smoothing	Number of people in the time series
	Jeon *et al.* (2019)	Utilization of menu information and application of machine learning methodologies when predicting the number of diners	Multiple Linear Regression, Lasso, Ridge, Random Forest, Bagging, Boosting	Daily Menu, Weekdays, Holiday, Weather, Seasons, Event
	Cetinkaya and Erdal (2019)	Daily Food Demand Predict with Artificial Neural Networks: Kirikkale University Case	Artificial Neural Network	Daily Menu, Event, Holiday

regression analysis.

3. Data and preprocessing

3.1. Data

To predict the number of diners served in the cafeteria, a study was conducted using data obtained from the Buk-gu office of Daegu. The data included cafeteria card payment details, menus, ingredients, and recipe data from January 2016 to November 2021. Additionally, the number of people who had lunch each day was collected via card payment details, excluding weekends, when the restaurant was closed, event days, and cleaning days. The counts of payments made between 10 : 00 and 15 : 00 were counted, resulting in a total of 1,448 days of data on the daily number of meals. Weather information, including daily maximum and minimum temperature and precipitation in Daegu during the period, was also collected. These data sets is a significant variable in predicting the number of customers based on the findings of Cheng *et al*. (2003) and Lim (2016). The collected data is shown in Table 2.

According to the collected data, the average daily number of diners is 316, with a maximum of 491

Table 2: Composition of collected data

Source	Data	Composition
Daegu Metropolitan City Buk-gu Office	Cafeteria card payment history	Date, payment time, transaction amount
	Diet information by date	Daily meal
	Ingredients	Ingredients by menu (1,193 types)
	Recipe	Recipe by menu (14 types)
	Dining capacity	Number of meals per day
Korea Meteorological Administration	Weather data	Daily high/low temperature, daily precipitation, etc.

Table 3: Stop-words

Stop-words	62 items including Domestic, Imported, Indian, American, Sun, Australian, Inchi, Kilo, etc.

and a minimum of 136. The standard deviation is 54. In terms of the menu, an average of 6 dishes were served per day, with a total of 1,429 unique menus, 14 recipes, and 1,193 ingredients used in the cafeteria. On average, each menu was prepared with 6 ingredients.

3.2. Data preprocessing

3.2.1. Combining and cleaning data

To build a model, it is essential to handle outliers and missing values in the collected data and perform appropriate data processing. Therefore, measures were taken to deal with outliers and missing values in this study. Furthermore, separate preprocessing was performed to refine unstructured data, as described below.

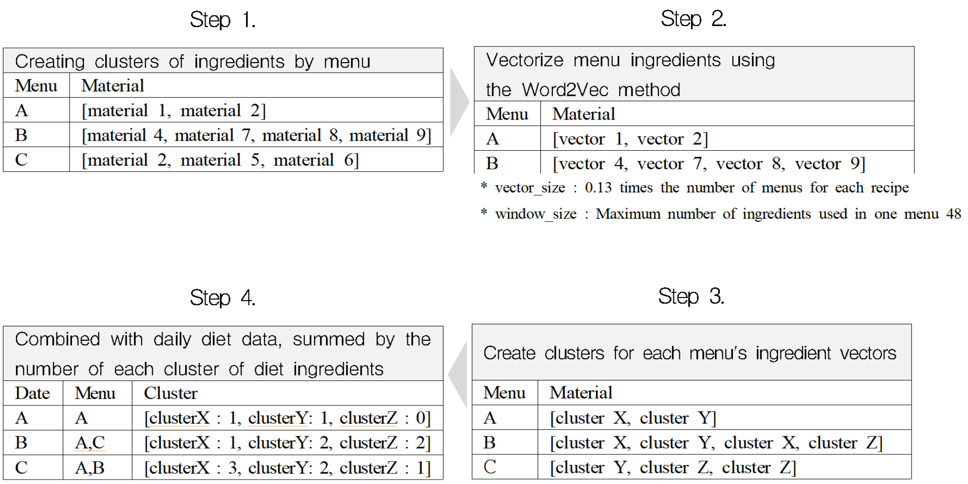
First, the data were combined based on the date and menu to create daily menu data. Menus that were not relevant to the daily diet were removed. For instance, additional side dishes provided when all the prepared ingredients had been exhausted were deemed unrelated to the prepared meal and were thus excluded. This resulted in a total of 902 unique menus.

Second, the food ingredient data for each menu was processed. Words like “unit” and “piece,” which are not semantically related to the ingredient, and the country of origin, “domestic,” were considered stop-words (Table 3). Moreover, all English letters, numbers, and special characters were removed. For example, “domestic pork shoulder 3kg” was refined to “pork shoulder.” Ultimately, there were 1,193 unique ingredients for the 902 menus items.

3.2.2. Creating derived variables

Derived variables were created to capture variables that previous studies have recognized as influencing the number of meals, including the day of the week, weather, and menu characteristics. Menu characteristics were further explored based on Jeon *et al.*’s (2019) classification of menus by recipe. Ingredient data was represented as vectors and clustered to generate derived variables that capture menu properties. Figure 1 illustrates the process of creating the menu property variable. To account for variables that have been found to influence meal consumption, such as the day of the week, weather, and menu characteristics, we created derived variables. Specifically, we extended the work of categorizing menus by recipe and embedding ingredient data in a vector space. We clustered similar ingredients to generate variables that capture menu properties. Figure 1 illustrates the process of creating the menu property variable.

Derived variables were created to capture variables that previous studies have recognized as influencing the number of meals, including the day of the week, weather, and menu characteristics.



Step 1.

Creating clusters of ingredients by menu

Menu	Material
A	[material 1, material 2]
B	[material 4, material 7, material 8, material 9]
C	[material 2, material 5, material 6]

Step 2.

Vectorize menu ingredients using the Word2Vec method

Menu	Material
A	[vector 1, vector 2]
B	[vector 4, vector 7, vector 8, vector 9]

* vector_size : 0.13 times the number of menus for each recipe

* window_size : Maximum number of ingredients used in one menu 48

Step 3.

Create clusters for each menu's ingredient vectors

Menu	Material
A	[cluster X, cluster Y]
B	[cluster X, cluster Y, cluster X, cluster Z]
C	[cluster Y, cluster Z, cluster Z]

Step 4.

Combined with daily diet data, summed by the number of each cluster of diet ingredients

Date	Menu	Cluster
A	A	[clusterX : 1, clusterY: 1, clusterZ : 0]
B	A,C	[clusterX : 1, clusterY: 2, clusterZ : 2]
C	A,B	[clusterX : 3, clusterY: 2, clusterZ : 1]

Figure 1: *Processing of menu property variables.*

Menu characteristics were further explored based on Jeon *et al.*'s (2019) classification of menus by recipe. Ingredient data was represented as vectors and clustered to generate derived variables that capture menu properties. Figure 1 illustrates the process of creating the menu property variable. To account for variables that have been found to influence meal consumption, such as the day of the week, weather, and menu characteristics, we created derived variables. Specifically, we extended the work of categorizing menus by recipe and embedding ingredient data in a vector space. We clustered similar ingredients to generate variables that capture menu properties. Figure 1 illustrates the process of creating the menu property variable.

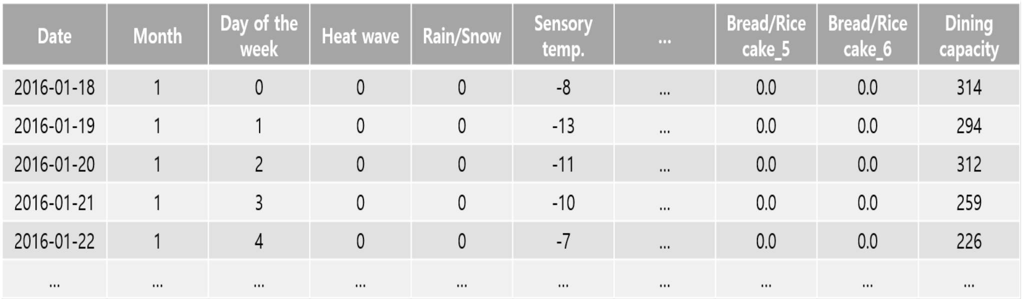
The continuous bag of word (CBOW) method of Word2Vec was used in this study, which uses neighboring words as input values to predict the current word. Unlike the bag-of-words method, CBOW considers the meaning of words, and similar words are represented by similar vectors (Mikolov *et al.*, 2013). After classifying menus according to their recipes, the CBOW method was used to convert the menu ingredients into one sentence. Each food ingredient was treated as a word, and ingredients with similar roles were represented by embedding similar vectors.

As there are approximately 1,200 types of food ingredients in the provided ingredient data, directly reflecting them as variables poses a challenge, similar to the curse of dimensionality. Therefore, the embedding vector generated through Word2Vec was used for performing spherical K-means clustering for each recipe. This clustering method improves the performance of clustering analysis when cosine similarity is used as a distance measure between data. Therefore, spherical K-means clustering is often used for clustering text data (Dhillon and Modha, 2001; Huang, 2008). To determine the number of clusters, the degree of cluster separation was evaluated using the silhouette coefficient, and the results of the cluster analysis are presented in Table 4.

Based on the clustering results, the ingredients in the daily menu were replaced with the cluster number to which the food item belongs. Afterward, the number of menus per cluster was created as a menu characteristical variable. To incorporate various characteristic variables that influence the number of diners, including the day of the week and weather, this study calculated and created derived variables such as perceived temperature, presence of a heatwave, presence of a holiday the following

Table 4: Number of ingredient clusters by recipe

Recipe	Number of clusters	Silhouette factor
Steamed/boiled	12	0.269
Korean soup	15	0.124
Sauteed/fried	3	0.172
Rice	15	0.467
Stewed/pickled	12	0.108
Grilled/stir-fried	7	0.116
Kimbap/rice ball	3	0.462
Seasoned/salad	3	0.101
Fruits/drinks/sweets	12	0.009
Soup	3	0.157
Kimchi	3	0.068
Sauce	15	0.320
Rice bowl/Bibimbap	7	0.512



Date	Month	Day of the week	Heat wave	Rain/Snow	Sensory temp.	...	Bread/Rice cake_5	Bread/Rice cake_6	Dining capacity
2016-01-18	1	0	0	0	-8	...	0.0	0.0	314
2016-01-19	1	1	0	0	-13	...	0.0	0.0	294
2016-01-20	1	2	0	0	-11	...	0.0	0.0	312
2016-01-21	1	3	0	0	-10	...	0.0	0.0	259
2016-01-22	1	4	0	0	-7	...	0.0	0.0	226
...	...	...	...	...	...	...	...	...	...

Figure 2: *Final data set (example).*

day, patronage of the previous day, day of the week, month, and year. Figure 2 illustrates the final data structure used for modeling.

4. Predictive model

4.1. Model

4.1.1. Linear regression

Linear regression is a model used to predict dependent variable based on independent variables. It can be expressed using the following formula:

$$y = \beta_o + \beta_1 x_1 + \beta_2 x_2 + \cdots \beta_n x_n,$$

where the x represents the independent variables, while y represents the dependent variable, whose value depends on x. The β_0 represents the intercept, and β_i represents the slope of each independent variable x_i. Simple linear regression refers to the case where there is only one independent variable. In contrast, multiple linear regression refers to the case where there are two or more independent variables. The least squares method is commonly used to estimate the regression coefficient β. This method aims to minimize the sum of the squared differences between the observed values and the values calculated by the regression model. The difference between the observed value and the value predicted by the regression model is referred to as the residual. The primary goal of the least squares method is to calculate the value of β_i that minimizes the sum of squares of the residuals, as shown in the following equation (Montgomery *et al.*, 2021).

$$S\left(\beta_0,\beta_1\right) = \sum_{i=1}^{n} \left(y_i - \beta_o + \beta_1 x_1 + \beta_2 x_2 + \cdots \beta_n x_n\right)^2 .$$

4.1.2. Random Forest

Random Forest is a bagging ensemble machine learning model that aims to enhance predictive power by aggregating multiple decision trees proposed by Breiman (2001). Bagging, which stands for bootstrap aggregating, is a methodology that involves training models with bootstrap samples and then averaging their results (Segal, 2004). To build a Random Forest model, decision trees are trained on bootstrap samples, and the final estimation is made by combining the predicted values obtained from each tree. Additionally, Random Forest randomly selects variables to be used at each node during the tree model learning process to reduce the correlation between decision trees. This approach improves the performance of the model (Breiman, 2001). One advantage of Random Forest is its ability to handle outliers and noise effectively, as well as prevent overfitting (Yoo, 2015). However, the Random Forest takes a long time to compute results.

4.1.3. GradientBoosting

GradientBoosting, developed by Friedman, is a typical boosting-type ensemble model. Boosting is a machine learning ensemble methodology devised based on the idea that it outperforms a single weak learner. Boosting creates weak learners, trains them sequentially, and weights misclassified values so that the next weak learner can classify better. At each boosting step, the weights are updated to minimize the residual, which is the difference between the actual value and the predicted value. GradientBoosting uses gradient descent for weight updates (Friedman, 2001). Gradient descent is a method to find the minimum value of a function using the derivative of the function, i.e., the gradient. Bagging trains several models at the same time to determine predicted values by voting, but boosting models are different in that they sequentially generate weak learners and reduce residuals. GradientBoosting is useful because it has high accuracy and can be applied without writing a separate loss function. It is also robust against outliers and missing values and is particularly suitable for high-dimensional data with many categorical variables.

4.1.4. CatBoost

The CatBoost model was developed by Yandex company as a specialized model for utilizing categorical variables by transforming GradientBoosting (Dorogush *et al.*, 2018). GradientBoosting has a disadvantage in that target leakage may occur since the gradient of the loss function used in each step is estimated by reusing the target value used when the current learner was created. Unlike existing boosting methods, the CatBoost model updates the weights sequentially whenever the residuals of each data are calculated via an ordered boosting method. The ordered boosting method constructs a random permutation σ as training data, trains the model M_i with only the data from the beginning up to the ith, and uses the j^{th} data as the predicted value calculated through the M_{j-1} model and the j^{th} observation. The residuals for the j^{th} sample are then compared and obtained. This model has the advantage of compensating for the slow learning rate of existing GradientBoosting models and solving the overfitting problem through leaf values. Additionally, it exhibits strong performance when processing multiple categorical variables through the one-hot-encoding method (Dorogush *et al.*, 2018). However, if the dataset consists mostly of numerical data, the learning rate of GradintBoosting may

Table 5: List of variables used for modeling

Type of variable	Category	Variable	Explanation	Data type
Dependent variable	dining capacity	dining capacity	group by date cafeteria dining capacity	Numeric
Independent variable	Menu property variable (105 item)	Fruits/Drinks/Sweets	Number of menus belonging to recipes in daily diet	Numeric
		Grilled/Stir-fried		Numeric
		Korean soup		Numeric
		Kimbap/Rice ball		Numeric
		Kimchi		Numeric
		⋮		⋮
		Sauteed/Fried		Numeric
		Steamed/Boiled_Material Cluster 1	The number of menus belonging to the ingredient group of each recipe in the daily diet	Numeric
		Steamed/Boiled_Material Cluster 2		Numeric
		Steamed/Boiled_Material Cluster 3		Numeric
		⋮		⋮
		Bread/Rice Cake_Material Cluster 6		Numeric
	Time series variable	Number of meals the day before	Number of meals the day before	Numeric
	Date variable	Day of the week	Mon/Tue/Wed/Thu /Fri	Categorical
		Month	January - December	Categorical
		Year	2016-2021	Categorical
		The day before the holidays	Whether it's the day before the holiday	Categorical
	Weather variable	Sensory temperature	Calculated considering temperature and wind speed	Numeric
		Heat wave	Sensible temperature of 33 degrees or more	Categorical
		Rain/Snow	Presence of rain/snow	Categorical

be slow. Furthermore, its performance may not be optimal when working with datasets that contain many missing values.

4.1.5. Ensemble

The Ensemble model is created by learning several individual models and synthesizing the results (Brown, 2010). By generalizing the model through the process of synthesizing the result values, the expected value of the prediction error is smaller than that of a single model, and the predictive accuracy is higher (Polikar, 2006; Zhou, 2012). Ensemble models include voting, bagging, boosting, and stacking. The stacking ensemble methodology creates a dataset by combining the output results of individual sub-models, and then the final meta-model learns and performs predicting. A stacking ensemble is used for several different types of sub-models, unlike boosting and bagging (Syarif *et al.*, 2012).

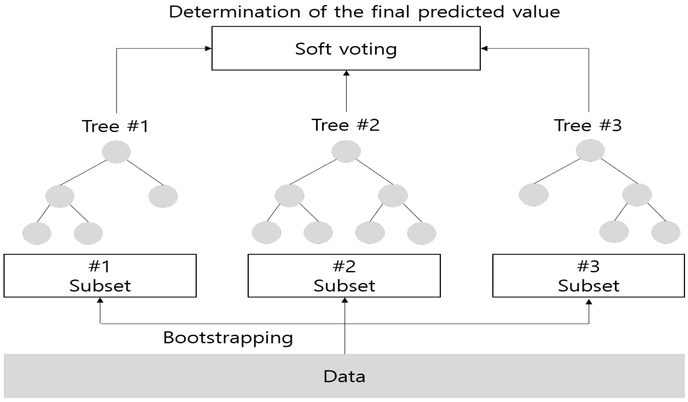


Figure 3: *Concept of random forest.*

Table 6: Performance comparison of models

Model	MSE	MAE	MAPE
Random Forest	1,056.61	24.85	0.083
GradientBoosting	972.41	24.12	0.081
CatBoost	943.59	23.88	0.079
Ensemble	865.28	22.26	0.074

4.2. Modeling

During the period from January 2016 to November 2021, 1,448 days of data were utilized to conduct the modeling. The dependent variable was the number of people dining in the cafeteria, and 152 independent variables were used, including the day of the week, month, number of meals served on the previous day, and menu items. The variables used for modeling are presented in Table 5. 80% of the total data was allocated as training data, and then training data was divided into five folds to prevent overfitting. And the remaining 20% of the total data was used as test data for modeling. One of the five folds was used as a validation dataset for cross-validation, and the other four folds were used as learning data.

The final model was built using the stacking ensemble methodology in this study. While a single model can be less effective, the stacking ensemble model generally delivers superior performance. However, there is a risk of overfitting the training data when operating with the stacking ensemble model. To overcome this issue, sub-models were constructed using Random Forest, GradientBoosting, and CatBoost. Moreover, the possibility of overfitting was reduced by applying 5-fold cross-validation. Lastly, we utilized the meta-model, and the stacking ensemble model is depicted in Figure 3.

4.3. Results of modeling

Table 6 presents a comparative analysis of the performance of the 3 sub-models and the final stacking ensemble model based on MSE, mean absolute error (MAE), and MAPE criteria. Among the 3 sub-models, the CatBoost model demonstrated the highest performance in all evaluation indicators, while the model generated by Random Forest showed the lowest performance. To further enhance the performance, the stacking ensemble methodology was applied, resulting in an MSE of 865.28, which improved by approximately 8 to 12% compared to the existing single models.

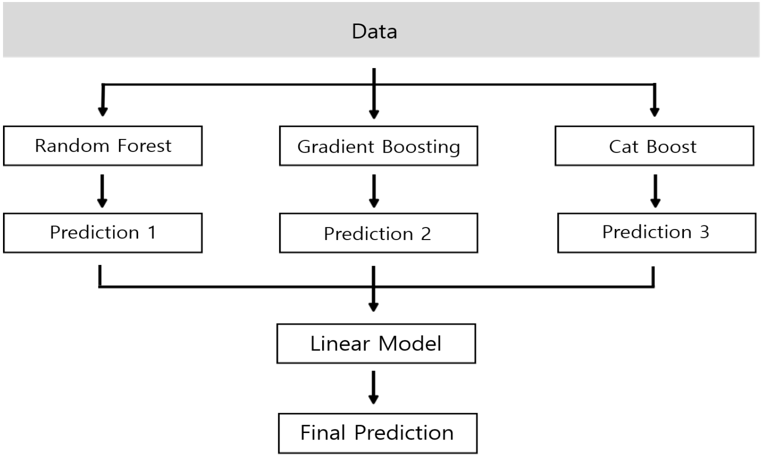


Figure 4: *Stacking ensemble model structure.*

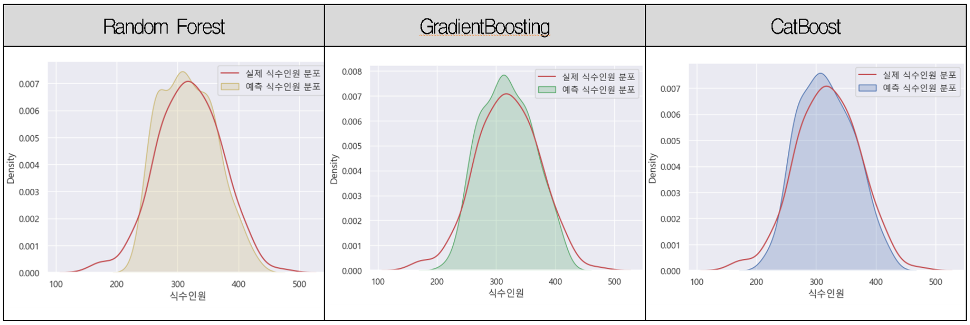


Figure 5: *Kernel density estimation result for each model.*

To visually confirm the difference between the actual and predicted values, the distribution of each was estimated using kernel density estimation, as shown in Figure 4. Kernel density estimation is a non-parametric method of estimating the distribution of observed data by applying kernel functions such as Gaussian and uniform to each data point and summing them to estimate a continuous distribution (Silverman, 2018). For all three models, it can be visually confirmed that the estimated distribution of the actual value and the estimated distribution of the predicted value is similar, except for extreme values. Therefore, it can be concluded that the model fits well in terms of model distribution. Therefore, it can be concluded that the model fits well in terms of distribution of the model (Figure 5).

The model was analyzed to determine the importance of each feature and detect the variables that affect the estimation of the count of people dining. Variables were classified into the date, weather, and menu variables, and their influence was quantitatively identified and presented in Table 7.

The date variable including the year, month, and day of the week, had the greatest influence on predicting the number of diners, followed by the menu variable, which includes ingredients and recipe variables, the weather variable, and other time-series variables related to dining. The date variable reflects the cafeteria's characteristics, such as changes in the number of people allowed to eat each

Table 7: Variable importance by model

Model	Date variable importance	Menu property variable importance	Other variable importance
Random Forest	61.684%	27.117%	11.199%
GradientBoosting	71.188%	15.838%	12.973%
CatBoost	47.713%	35.993%	16.293%
Mean	66.44%	66.44%	66.44%

Table 8: Performance comparison with previous studies

Model	Menu classification method	MSE	MAE	MAPE
Random Forest	Korean food menu classification by Dewey decimal classification	1,055.32	24.83	0.084
	Word2Vec + Spherical K means	1,056.61	24.85	0.084
GradientBossting	Korean food menu classification by Dewey decimal classification	996.43	24.35	0.082
	Word2Vec + spherical K means	972.41	24.11	0.081
CatBoost	Korean food menu classification by Dewey decimal classification	948.21	24.00	0.08
	Word2Vec + spherical K means	943.59	23.88	0.079
Ensemble	Korean food menu classification by Dewey decimal classification	886.09	22.62	0.075
	Word2Vec + spherical K means	865.28	22.26	0.074

day due to vacations, moving in or out, etc. Moreover, the analysis showed that the menu variable had a significant effect on the estimation of diners, with an importance value of 26.32%.

This study compared the effectiveness of word embedding-based menu features in improving predictive performance with the Korean food menu classification method via DDC. The Korean food menus were classified using the same method as in Jeon *et al.* (2019), and other variables were fixed and analyzed (Jeon *et al.*, 2019). As shown in Table 8, when comparing the performance of the three sub-models and the final stacking ensemble model by the MSE, MAE, and MAPE indicators, the methodology used in this study demonstrated some improvement in model performance compared to the menu classification method using DDC.

5. Conclusion

5.1. Results and implications

5.1.1. Analysis results

This study created a menu variables using word embedding methodologies and predicted the number of meals served in a group cafeteria based on a machine learning ensemble model. The Word2Vec methodology was used to perform word embedding for menu variables, and menu characteristic variables were created by clustering them. A stacking ensemble model was then constructed by applying Random Forest, GradientBoosting, and CatBoost as sub-models.

After comparing the performance of CatBoost with Random Forest and GradientBoosting, CatBoost was found to have the best performance among the three sub-models. When the stacking ensemble model was evaluated, it showed an 8% improvement in performance based on the MSE index. Additionally, kernel density estimation was used to visually confirm that the distribution of actual values and predicted values was similar. To understand the contribution of the menu variables within the model, the feature importance of each model was visualized and displayed. The results showed that the menu variables had an importance of 23% in the case of Random Forest, 12% in the case of GradientBoosting, and 32% in the case of CatBoost. Based on the data used in this study, the menu variable processing method used in previous studies was applied to the word embedding and stacking ensemble models, and the models were compared.

5.1.2. Implication

The implications of this study are as follows:

Firstly, the results demonstrate that machine learning methodology can improve prediction performance and can be applied in enterprises to accurately estimate diners. Predicting the number of people dining in a restaurant often depends on the dietitian's experience and intuition. Previous studies have shown that models based on limited variables or statistical methodologies had insufficient performance for real-world applications. However, this study demonstrates that machine learning methodology can improve predictive performance and can be applied in enterprises.

Secondly, the date variable was found to have the greatest influence on the number of diners in a cafeteria, followed by menu characteristics and other variables. These results can serve as basic data for predicting the number of people eating in a company cafeteria and can be used to reduce food waste and establish restaurant support policies by local governments.

Thirdly, Jeon *et al.* (2019) used menu variables to classify menu ingredients and recipes into DDC, but personal elements may have been introduced as the classification was done manually. This study removes subjective elements in menu classification. This can be applied to actual work or future research.

Lastly, the importance of ESG (environmental, social, and governance) has been highlighted in public institutions and large corporations. Food waste accounts for about 30% of total household waste, and cafeterias contribute more than 10% of food waste. An accurate estimation of the number of people eating can contribute to reducing food waste, creating a virtuous cycle structure that reduces social and environmental costs.

5.2. Limitation

This research has the following limitations: Firstly, the menu and ingredient data used in this study were provided by a nutritionist in the cafeteria of the ward office, and there may be environmental conditions or errors that could have influenced the results. If a computerized system for cafeterias is established in the future, data collection conditions can be improved.

Secondly, the number of data cases used in the machine learning model was 1,448, which is relatively small compared to the number of variables. As a result, some of the sufficiency conditions may not have been satisfied in the model learning. Collecting long-term accumulated data could improve model performance and prevent overfitting.

Thirdly, the average number of meals per year varied greatly, with 286 in 2016 and 375 in 2020. This variation may be due to various variables such as the division of workers, district office policies, and infrastructure improvement. Controlling or reflecting various exogenous variables in the model in the future could lead to a higher performance improvement.

Lastly, the model structure may not be robust when a new menu or a mixture of foreign words is not included in the learning data. Although the word embedding methodology used in this study is an improvement over the Korean food menu classification method previously developed, there may still be limitations in handling new menus or foreign words.

Future studies that aim to address these limitations presented in this study need to be conducted.

References

Baek OH, Kim MY, and Lee BH (2007). Menu satisfaction survey for business and industry foddservice workers - Focused on food preferences by gender, *Journal of The Korean Society of Food Culture*, **22**, 511–519.

Blecher L and Yeh RJ (2008). Forecasting meal participation in university residential dining facilities, *Journal of Foodservice Business Research*, **11**, 352–362.

Breiman L (2001). Random forests, *Machine Learning*, **45**, 5–32.

Brown G (2010). Ensemble learning, *Encyclopedia of Machine Learning*, **312**, 15–19.

Cetinkaya Z and Erdal E (2019). Daily food demand forecast with artificial neural networks: Kirikkale University case, In *2019 4th International Conference on Computer Science and Engineering*, Samsun, Turkey, 1–6.

Cheng L, Yang IS, and Baek SH (2003). Investigation on the performance of the forecasting model in university foodservice, *Journal of Nutrition and Health*, **36**, 966–973.

Cho H (2019). Cluster analysis on categorical data using word embedding method : Focused on a high dimensional data (Master's thesis), Kookmin University, Seoul.

Dhillon IS and Modha DS (2001). Concept decompositions for large sparse text data using clustering, *Machine Learning*, **42**, 143–175.

Dorogush AV, Ershov V, and Gulin A (2018). CatBoost: GradientBoosting with categorical features support, Available from: arXiv preprint arXiv:1810.11363

Friedman JH (2001). Greedy function approximation: A GradientBoosting machine, *Annals of Statistics*, **29**, 1189–1232.

Huang A (2008). Similarity measures for text document clustering, In *Proceedings of the Sixth New zealand Computer Science Research Student Conference (NZCSRSC2008)*, Christchurch, New Zealand, 4, 9–56.

Jeon J, Park E, and Kwon OB (2019). Predicting the number of people for meals of an institutional foodservice by applying machine learning methods: S city hall case, *Journal of the Korean Dietetic Association*, **25**, 44–58.

Lim JY (2016). Analysis of forecasting factors affecting meal service in business foodservice (Master's thesis), Yonsei University, Seoul.

Mikolov T, Chen K, Corrado G, and Dean J (2013). Efficient estimation of word representations in vector space, Available from: arXiv preprint arXiv:1301.3781v3

Montgomery DC, Peck EA, and Vining GG (2021). *Introduction to Linear Regression Analysis*, John Wiley & Sons, Hoboken.

Park S and Byun YC (2021). Improving recommendation accuracy based on machine learning using multi-dimensional features of Word2Vec, *Journal of Korean Institute of Information Technology*, **19**, 9–14.

Park SS and Lee KC (2018). Effective Korean sentiment classification method using word2vec and ensemble classifier, *Journal of Digital Contents Society*, **19**, 133–140.

Polikar R (2006). Ensemble-based systems in decision making, *IEEE Circuits and Systems Magazine*, **6**, 21–45.

Ryu K and Sanchez A (2003). The evaluation of forecasting methods at an institutional foodservice dining facility, *The Journal of Hospitality Financial Management*, **11**, 27–45.

Segal MR (2004). *Machine learning benchmarks and random forest regression*, Center for Bioinformatics and Molecular Biostatistics, University of California, San Francisco.

Silverman BW (2018). *Density Estimation for Statistics and Data Analysis*, CRC press.

Syarif I, Zaluska E, Prugel-Bennett A, and Wills G (2012). Application of bagging, boosting and stacking to intrusion detection, In *Machine Learning and Data Mining in Pattern Recognition: 8th International Conference* (pp. 593–602), Springer, Berlin Heidelberg.

Yoo JE (2015). Random forests, an alternative data mining technique to decision tree, *Journal of Educational Evaluation*, **28**, 427–448.

Zhou ZH (2012). *Ensemble Methods: Foundations and Algorithms*, CRC press, Hoboken.

Received March 17, 2023; Revised May 17, 2023; Accepted June 4, 2023