Heliyon 8 (2022) e10708

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon

Research article

Analysis of changes in geographical factors affecting sales in commercial alleys after COVID-19 using machine learning techniques

Lee Kangjae *

Department of Convergence and Fusion System Engineering, Kyungpook National University, 2559 Gyeongsang-daero, Sangju-si, Gyeongsangbuk-do 37224, Republic of Korea

ARTICLE INFO

Keywords:
Random forest
Extreme gradient boosting
Geographic information system (GIS)
Feature importance
Shapley additive explanations (SHAP)

ABSTRACT

Social restrictions, such as social distancing and self-isolation, imposed owing to the coronavirus disease-19 (COVID-19) pandemic have resulted in a decreased demand of commodities and manufactured products. However, the factors influencing sales in commercial districts in the pre- and post-COVID-19 periods have not yet been fully understood. Thus, this study uses machine learning techniques to identify the changes in important geographical factors among both periods that have affected sales in commercial alleys. It was discovered that, in the post-COVID-19 period, the number of pharmacies, age groups of the working population, average monthly income, and number of families living in apartments priced higher than $600k in the catchment areas had relatively high importance after COVID-19 in the prediction of a high level of sales. Moreover, the percentage of deciduous forests appeared to be a important factor in the post-COVID-19 period. As the average monthly income and worker population in their 60s and numbers of pharmacies and banks increased after the pandemic, sales in commercial alleys also increased. The survival of commercial alleys has become a critical social problem in the post-COVID-19 era; therefore, this study is meaningful in that it suggests a policy direction that could contribute to the revitalization of commercial alley sales in the future and boost the local economy.

1. Introduction

The outbreak of the severe acute respiratory syndrome coronavirus 2 has significantly changed the pattern of our daily life. Owing to the coronavirus disease-19 (COVID-19), most people prefer to stay at home and are reluctant to go outside to reduce the risk of infection (He et al., 2020). COVID-19 has affected the paradigm of our daily lives in several different ways, including dietary intake (Bertrand et al., 2021), travel habits (Anzai et al., 2020; J. Li et al., 2021), recreational activities (Ainsworth and Li, 2020; Craig, 2021; Hammami et al., 2022), and has also had a psychological impact (Hagger et al., 2020; Rodríguez-Rey et al., 2020).

Social restrictions, such as social distancing, travel restrictions, and self-isolation, imposed owing to the COVID-19 pandemic have resulted in a decreased demand of commodities and manufactured products (Nicola et al., 2020). The restaurant industry was one of the sectors most vulnerable to the restrictions imposed during the pandemic. Dube et al. (2021) discovered that the number of dine-in guests dropped to zero in many countries owing to the restrictions imposed. Therefore, the unexpected crisis brought about by the recent COVID-19 pandemic has inflicted an economic blow on businesses and, in all respects, changed consumption patterns and business environments. Hence, businesses are forced to change their strategies to overcome the economic effects of COVID-19 (Verbeke and Yuan, 2021).

After the pandemic, few studies regarding COVID-19 and its impact on sales in commercial alleys in South Korea have been conducted (Lee and Kim, 2021; Yu, 2021). For example, Yu (2021) analyzed the long-term changes in commercial alley sales using time-series cluster analysis and associations between geographical factors by including the occurrence of COVID-19 as a dummy variable, and the sales of each resulting cluster using panel analysis. However, recent studies have not suggested any guidelines for growing sales of commercial alleys based on the COVID-19 situation. The factors influencing the growth of sales in commercial alleys must be identified, which might have changed after the pandemic.

Thus, this study focused on the changes in crucial geographical factors in the pre- and post-COVID-19 periods that have affected sales in commercial alleys. Machine learning techniques were used to predict sales using various influential contexts and explore the differences in important factors in the pre- and post-COVID-19 periods. Random forest

* Corresponding author.
E-mail address: kasbiss@knu.ac.kr.

https://doi.org/10.1016/j.heliyon.2022.e10708
Received 23 May 2022; Received in revised form 21 July 2022; Accepted 16 September 2022

(RF) and extreme gradient boosting (XGB) are two popular tree-based algorithms. Among all machine learning algorithms, tree-based algorithms offer excellent performance when structured data, such as geographic information system (GIS) data, are used (Rudin, 2019). The two algorithms helped detect the differences in influential factors during both periods and were compared using 10-fold cross-validation to determine the best model. The importance of influential factors in pre- and post-COVID-19 periods was explored, and the results were interpreted using Shapley additive explanations (SHAP). Because the SHAP approach is specialized for tree-based models (Rodríguez-Pérez and Bajorath, 2020), RF and XGB models were used in this study. Tree-based models combined with SHAP can obtain complex information from data, and some studies using tree-based algorithms with SHAP have been conducted in different research domains (Y. Kim and Kim, 2022; Mangalathu et al., 2020; Zhang et al., 2022). Owing to the importance of sales clusters suggested by Yu (2021), we included a variable from the cluster analysis using an approach based on distances between commercial alleys. Because clusters created based on sales can vary among different years, we selected a more invariant method based on distance. The distance also indicates that closer commercial alleys have similar cultural, physical, and socioeconomic characteristics, which may affect their sales. Additionally, owing to the increasing importance of green spaces after COVID-19 (Lee and Kim, 2021), we included some variables related to green spaces. Instead of the number of green spaces, the density of different green spaces, such as deciduous forest, coniferous forest, mixed forest, natural grass, and artificial grass, was included to discover more profound evidence of their role based on their type.

The rest of this paper is organized as follows: Section 2 is a summary of the published work regarding the conceptual categories for geographical factors and the impact of COVID-19 on commercial alleys. Section 3 introduces data processing and machine learning model tuning. Section 4 compares the performances of RF and XGB models, selects the best model, and shows the changes in the important variables before and after COVID-19. In Section 5, we discuss the associations between the selected important variables and sales in commercial alleys. We conclude this study in Section 6.

2. Related work

2.1. Conceptual categories for geographical factors

Turhan et al. (2013) suggested the following six categories of factors that influence sales: population structure, economy, competition, saturation level, store characteristics, and magnet. Population structure includes population size, age, gender, education level, occupation, number of households, political orientation, traveling time, and shopping habits (Y. Li and Liu, 2012). Economic factors indicate the population economic status and represent part of the population structure (Turhan et al., 2013). Competition includes elements related to the effects of a competitive retail environment on store performance, such as size and number of competitors (Reinartz et al., 1999). The saturation level indicates the attractiveness of a particular market for determining whether a higher profit can be achieved (Dunne et al., 2013). Store characteristics include important aspects of stores, such as accessibility and store image attributes (Turhan et al., 2013). The magnet factors suggested by Kuo et al. (2002) include magnet shops or facilities that can attract more trade, including cultural and educational institutions, vehicle maintenance facilities, and leisure activities.

These categories were used to strategically select the location of retail store in the research area. Many studies have considered various types of influential factors to maximize profitability using the analytical hierarchy process (Akalin et al., 2013; Erbıyık et al., 2012; Harwati and Utami, 2018; Hsu and Chen, 2007; KoÃ & Burhan, 2015; Kuo et al., 2002; Manowan et al., 2022; Tzeng et al., 2002) and techniques for order preference by similarity to ideal solution (Chang and Hsieh, 2014), fuzzy models (Chou et al., 2008), analytic network process (Cheng et al., 2005), hybrid models (Singh et al., 2020), and machine learning models (Fu et al., 2022). Recently, Fu et al. (2022) included residential and business areas, pedestrian flow, competition, store characteristics, and the number of households as factors for training a model using machine learning techniques to predict daily turnover for eventual use in convenience store site selection.

2.2. Impact of COVID-19 on commercial alleys

Various studies have been conducted to understand the impact of COVID-19 on small businesses. From April to June 2020, the number of active business owners in the U.S. abruptly fell from 15.0 to 11.7 million (Fairlie, 2020). Although small businesses employ approximately 50% of American workers (A. Bartik et al., 2020), they are more vulnerable to economic crises than large businesses (=Kennickell et al., 2015). A survey identified that many small businesses were likely to fail without financial support during the pandemic (A. W. Bartik et al., 2020). Moreover, financial difficulties caused by the pandemic reduced the employers' and employees' capacities to endure stress (Isabelle et al., 2022). One of the ways that small business owners can overcome the negative impact of the pandemic is to stay connected with the community and industry to acquire sources required for their entrepreneurial ecosystems (Liguori and Pittz, 2020). Despite previous studies, there is insufficient information regarding the changing nature of the surrounding influential environments after the COVID-19 pandemic that can influence the performance of small businesses in commercial districts to guide them for growing their performance, such as sales.

In South Korea, interest in commercial alleys has recently increased; hence, some studies have focused on the revitalization of commercial alleys and their performance prediction, such as sales, considering numerous geographical factors in the pre- and post-COVID-19 periods. Particularly, commercial alleys in Seoul do not have large distribution facilities, are located near densely populated areas, and are commercial districts comprising businesses related to daily life, such as wholesale, retail, restaurants, and services (Kang and Lee, 2018). Small business stores are primarily located close to residential areas and are easily affected by changes in the surrounding environment (Yu, 2021). Kim and Lee (2019) examined the associations between sales in commercial alleys and influential factors, including business characteristics, spatial structure of the city, and catchment characteristics, using multiple linear regression. The sale rates in the 20–40 age group and periods when people are active, such 6–11 am and 5–9 pm, were found to be statistically significant and positively associated with total sales of commercial alleys. The monthly average income in the catchment areas was also significantly and positively associated with sales. Kang and Lee (2018) used geographically weighted regression to identify the customer characteristics that can affect sales in commercial alleys. They discovered that the proportion of female customers had the largest effect on sales in the northwestern region of Seoul. Moreover, they discovered that the proportion of customers in the 20–30 and 40–50 age groups significantly affected the sales in southeastern and northeastern regions. They used logistic regression to identify the factors that contribute to sales growth in commercial alleys. Their results showed that growth in sales shares of females and population aged between 20s–30s, and building density had significantly positive effects on the sales growth of commercial alleys, whereas the income level of catchment areas and proximity to subway stations had significantly negative effects.

After the COVID-19 pandemic, few studies regarding COVID-19 and its impact on commercial alley sales were conducted. Yu (2021) analyzed the long-term changes in commercial alleys sales using time-series cluster analysis and the associations between geographical factors, including the occurrence of COVID-19 as a dummy variable, and the sales of each resulting cluster using panel analysis. Their results indicated that the COVID-19 pandemic had a negative effect on all clusters, which had relatively different sales levels. Lee and Kim (2021) focused on the role of urban parks in the revitalization of commercial alley sales post-COVID-19

and discovered that a higher number of parks near commercial alleys and parks with better visual features were associated with lower sale losses.

However, recent studies have not suggested guidelines for growing sales in commercial alleys by considering the unusual situation of COVID-19. Thus, this study attempts to provide insights into the revitalization of commercial alley sales using machine learning techniques.

3. Methods

3.1. Datasets

This study used the commercial alley data of Seoul, South Korea, published in the Seoul Open Data Plaza (https://data.seoul.go.kr/), to identify geographical factors. We selected 1,008 out of the 1,009 commercial alleys in Seoul; one was omitted because its data regarding the factors associated with apartments was missing (Figure 1). Data collected in 2018 and 2019 were used for pre-COVID-19 period, whereas those collected in 2020 and 2021 were used for the post-COVID-19 period. Because 2018 and 2019 were the years immediately preceding the COVID-19 pandemic, we used the data of these two years to represent the pre-COVID-19 period. In terms of the timeline, comparing data of previous years, such as 2018 and 2019, is more plausible than comparing data of other years with those of 2020 and 2021, because the trend of sales over time and other influential factors might not have changed considerably and we can focus solely on the effect of COVID-19 on the commercial alley sales.

All factors provided in the dataset were calculated based on the catchment areas of commercial alleys from where businesses or services attract their customers (Figure 2). A catchment area is a 200 m circular buffer around each commercial alley, and previous studies have noted that the characteristics of catchment areas can affect the commercial alley sales (Kang and Lee, 2019; Lee and Kim, 2021). However, the sales used in this study were estimated based on the areas of commercial alleys.

Land cover data from 2019 and 2020 provided by the Ministry of Environment, Korea were used to calculate the densities of different types of green spaces. The middle level of the land cover map had a 5 m resolution and 22 classes, including residential, industrial, transportation, forest, and grass areas, and water bodies. As no data were available for 2018 and 2021, we used the land cover data for 2019 and 2020 as substitutes to determine the densities of green spaces in 2018 and 2021, respectively.

3.1.1. Response variable and factors

This study aimed to classify sales into three levels: low, medium, and high, and it specifically focused on the influential factors that determined high level of sales in the pre- and post-COVID-19 periods, and the prediction of sales in the post-COVID-19 period using these factors. Therefore, the response variable indicates the level of total sales of all businesses in each commercial alley (Table 1). The abbreviated factor names used in this study and their descriptions are summarized in Table 2. Sales below the 25th percentile level indicate low level, those above the 75th percentile indicate high level, and those between the 25th and 75th percentiles are indicate medium level. The number of influential factors used in this study was 75, as presented in Table 1, and all factors can be grouped into the economy, magnet, and population structure categories based on the six different categories proposed by Turhan et al. (2013), except for time and connectivity. As there was no data available for the competition, saturation level, and store characteristics categories, these three categories were not considered. Among the two newly added categories, connectivity describes how nearby commercial alleys are socially and culturally connected and the time category defines temporal units, such as quarters or months, which can help evaluate the variability of sales over time. The data of the influential factors were not normalized or transformed, which allowed us to interpret the associations between the factors and sales predictions using their original values and units.

The quarters in the time category indicate the sales variations in different seasons, whereas the clusters contribute to the better prediction

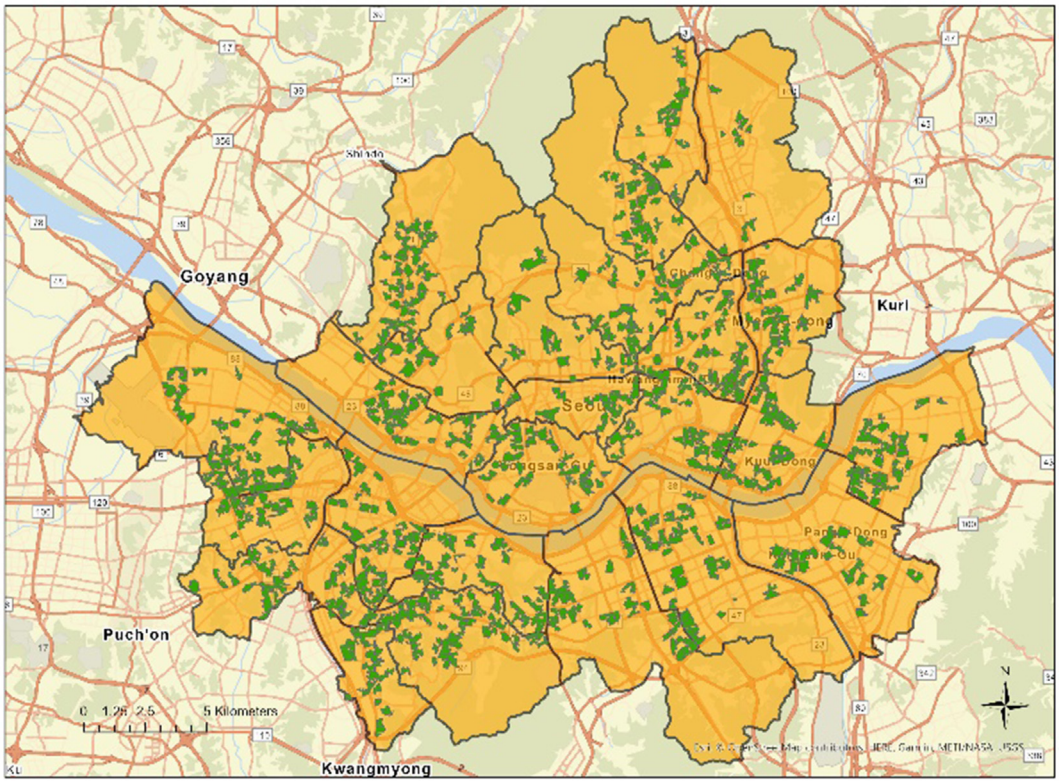

Figure 1. Commercial alleys (green color) in Seoul.

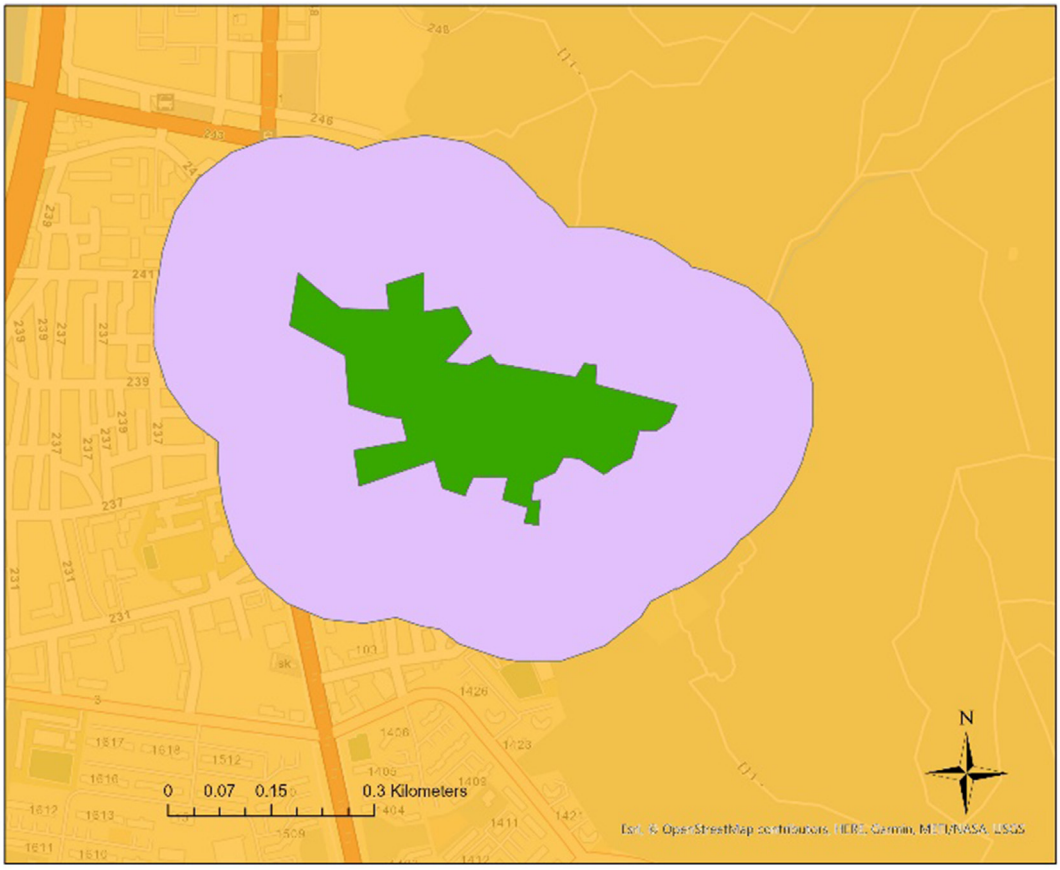

Figure 2. An example of a catchment area (pink color) around a commercial alley (green color).

of different levels of sales by accounting for the connectivity of nearby commercial alleys, which indicate that commercial alleys are more related to nearby commercial alleys based on Tobler's first law of geography (Tobler, 1970), both socially and culturally. Particularly, k-means cluster analysis (Lloyd, 1982) was conducted to group commercial alleys, and a silhouette (Rousseeuw, 1987) was used to evaluate the quality of clusters and determine the number of clusters (Figure 3). To determine the clusters, we used the x and y coordinates that can help geographically define clusters of commercial alleys. As a result, the optimal number of clusters was discovered to be two, indicating the highest average silhouette coefficient value. The two clusters separating the western and eastern areas of Seoul are shown in Figure 4. The number of commercial alleys in the western cluster was 495, whereas that in the eastern cluster was 513.

3.2. Machine learning techniques

3.2.1. RF and XGB

We used two machine learning models, RF and XGB, to predict different levels of sales in commercial alleys based on economy, magnet, population structure, connectivity, and time factors. All machine learning processes, including training and validation, were implemented on Python. An appropriate algorithm between the two was selected based on the evaluation results for further analysis to examine the important factors and associations between some factors and sales.

Both RF and XGB are tree-based ensemble machine-learning algorithms that combine the predicted results of multiple classifiers to perform more accurate predictions. Particularly, RF (Breiman, 2001) generates numerous decision trees in parallel and predicts different classes based on the votes of the generated trees. Each tree is built using a different bootstrap sample, which is called bagging, and a subset of variables is randomly selected for each node. The best split is then selected within a random set of variables to split the node. In contrast, XGB (Chen and Guestrin, 2016) trains several weak classifiers sequentially to generate a strong classifier. It adopts the notion of gradients such that the loss function is minimized. Particularly, XGB supports parallelization during the construction of each tree for efficient computation. In this study, we used the scikit-learn and xgboost packages of Python to implement the RF and XGB algorithms, respectively.

For each model, hyperparameter optimization was performed through a random search using the entire four-year data, and the models were tuned using the optimized hyperparameters. Random search defines a search space using hyperparameter value ranges and randomly selects combinations from the space for validation. To avoid overfitting, we conducted a three-fold cross-validation during hyperparameter tuning. Table 3 lists the hyperparameter ranges used for tuning the RF and XGB models and the optimized hyperparameter combinations for each model. Training and validation of the two models were conducted using the resulting combinations.

For the RF, the minimal cost-complexity pruning alpha parameter was used for pruning to prevent overfitting. The largest cost-complexity, which was smaller than the defined parameter, was used to select the subtree. The default parameter is not pruned. When the parameter value increases, the number of pruned nodes increases. In XGB, the gamma parameter is a regularization parameter, which defines the threshold of gain improvement to maintain a split. XGB starts pruning the tree backward and removes splits beyond which there is no positive gain. In each split, the gamma subtracted from the gain is calculated, and pruning is performed when its value is negative.

A part of the first decision tree constructed via the RF using tuning parameters is shown in Figure 5. Important factors used in the prediction

Table 1. Details of the response variable and influential factors.

	Name	Description	Notes
Response	Sales	Level of total sales of all businesses in each commercial alley.	Seoul commercial alley data
Economy	Household related to the area of the apartment	Number of families living in apartments smaller than 66 m^2 or larger than 66, 99, 132, or 165 m^2.	Seoul commercial alley data
	Household related to the price of the apartment	Number of families living in apartments priced under $100,000 or over $100,000, $200,000, $300,000, $400,000, $500,000, or $600,000.	Seoul commercial alley data
	Income (currency: KRW)	Average monthly income.	Seoul commercial alley data
Magnet	Facility	Number of facilities including all facilities, public facilities, banks, hospitals, clinics, pharmacies, kindergarten, elementary schools, middle schools, high schools, colleges, department stores, supermarkets, theaters, accommodations, airports, railway stations, bus terminals, subway stations, and bus stops.	Seoul commercial alley data
	Green space	Percentage of deciduous forest, coniferous forest, mixed forest, natural grass, and artificial grass.	Land cover data
Population structure	Dynamic population	Total, male, female, 10s, 20s, 30s 40s, 50s, or over 60, Monday, Tuesday, Wednesday, Thursday, Friday, Saturday, Sunday.	Seoul commercial alley data
	Resident population	Total, male, female, 10s, 20s, 30s, 40s, 50s, or over 60.	Seoul commercial alley data
	Worker population	Total, male, female, 10s, 20s, 30s, 40s, 50s, or over 60.	Seoul commercial alley data
Connectivity	Cluster	Two clusters based on distance.	K-means
Time	Quarter	Quarter of a year.	Seoul commercial alley data

of sales, including the working and dynamic populations, are illustrated with the values that split the incoming data.

3.2.2. SHAP

Over the past few years, the machine learning research domain has endeavored to develop methods for interpreting the internal logic of predictions, which remains a complex black box (Carvalho et al., 2019). The SHAP approach can provide interpretability tools (Lundberg and Lee, 2017). It was developed based on the theoretically optimal Shapley values of the game theory. It aims to explain the contribution of each variable to the prediction, and can provide information regarding the importance of variables in the prediction and the associations between each variable and predictions using the Shapley value. The Shapley value determines the average marginal contribution of each variable to the predicted results. The following equation defines the contribution Φ of each variable. S is a subset containing the variables, whereas N denotes all variables. The value function v takes S as the input, and $v(S\cup\{i\}) - v(S))$ indicates the marginal contribution of variable i to the outcome. $\frac{|S|!(N-|S|-1)!}{N!}$ is the weight used to calculate the

Table 2. Factor names and their descriptions.

Factor name	Description
A_a_#	Number of families living in apartments smaller than 66 m^2 or larger than 66, 99, 132, or 165 m^2.
A_p_#	Number of families living in apartments priced under $100,000 or over $100,000, $200,000, $300,000, $400,000, $500,000, or $600,000.
IC	Average monthly income.
F_t, F_pf, F_ba, F_hos, F_clin, F_pha, F_kin, F_ele, F_mid, F_high, F_col, F_ds, F_supm, F_thea, F_acco, F_air, F_railsta, F_buster, F_substa, F_busstp	Total number of all facilities, public facilities, banks, hospitals, clinics, pharmacies, kindergarten, elementary schools, middle schools, high schools, colleges, department stores, supermarkets, theaters, accommodations, airports, railway stations, bus terminals, subway stations, and bus stops, respectively.
LC_df, LC_cf, LC_mf, LC_ng, LC_ag	Percentages of deciduous forest, coniferous forest, mixed forest, natural grass, and artificial grass, respectively.
DP_t, DP_m, DP_f, DP_#, DP_M, DP_Tu, DP_W, DP_Th, DP_F, DP_Sa, DP_Su	Dynamic population total, male, female, 10s, 20s, 30s 40s, 50s, and over 60, Monday, Tuesday, Wednesday, Thursday, Friday, Saturday, and Sunday, respectively.
RP_t, RP_m, RP_f, RP_#	Resident population total, male, female, 10s, 20s, 30s, 40s, 50s, and over 60, respectively.
WP_t, WP_m, WP_f, WP_#	Worker Population total, male, female, 10s, 20s, 30s, 40s, 50s, and over 60s, respectively.
Cluster	Cluster.
Qs	Quarter.

Shapley value based on the weighted average of all marginal contributions over the varying S.

$$\Phi_i(v) = \sum_{S\subseteq N\setminus\{i\}} \frac{|S|!(N-|S|-1)!}{N!}(v(S\cup\{i\}) - v(S))$$

Python contains a package for SHAP. Using SHAP, we examined the importance of influential factors, impact of these factors on model output, and dependence of some factors on the predictions.

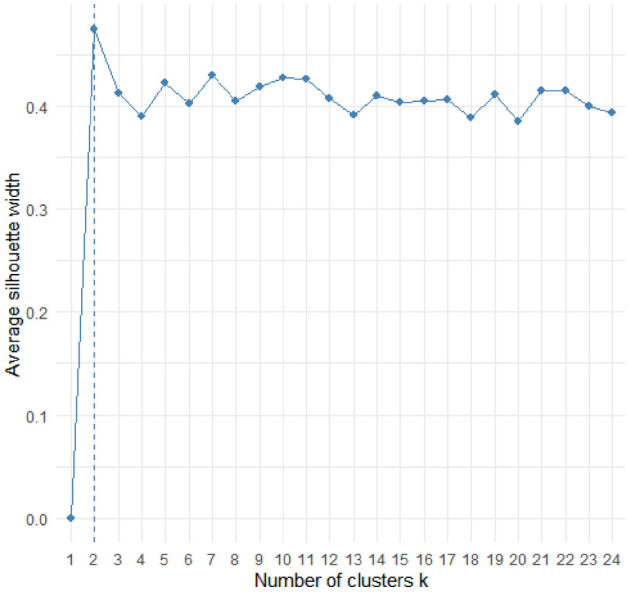

Figure 3. Silhouette plot to determine the optimal number of clusters via k-means.

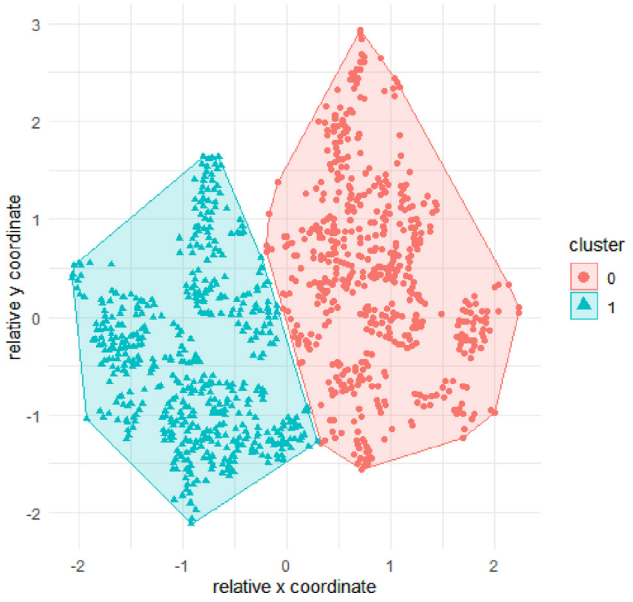

Figure 4. Result of k-means cluster analysis of commercial alleys with the optimal number of clusters; x and y coordinates are relative coordinates, which represent the locations of commercial alleys and clusters.

Table 3. Hyperparameter ranges and optimized hyperparameter combinations.

Hyperparameter ranges used for tuning	RF	*maximum depth* ∈ [5, 7, 10, 15, 20], *minimum samples leaf* ∈ [1, 2, 4], *minimum samples split* ∈ [2, 5, 10], *number of estimators* ∈ [1000, 1100, 1200, 1300, 1400, 1500, 1600, 1700, 1800, 1900, 2000] *minimal cost-complexity pruning alpha* ∈ [0.01, 0.05, 0.1, 0.2]
	XGB	*gamma* ∈ [0.5, 1, 5, 10], *learning rate* ∈ [0.001, 0.01, 0.1], *maximum depth* ∈ [5, 7, 10, 15, 20], *minimum sum of instance weight in a child* ∈ [5, 10, 15, 20], *number of estimators* ∈ [1000, 1100, 1200, 1300, 1400, 1500, 1600, 1700, 1800, 1900, 2000], *subsample* ∈ [0.6, 0.8, 1.0], *subsample ratio of columns for each tree* ∈ [0.6, 0.8, 1.0]
Optimized hyperparameter combinations	RF	*maximum depth* = *70*, *minimum samples leaf* = 1, *minimum samples split* = *2*, *number of estimators* = 1900, *minimal cost-complexity pruning alpha* = 1900,
	XGB	*gamma* = 0.5, *learning rate* = 0.1, *maximum depth* = 15, *minimum sum of instance weight in a child* = 10, *number of estimators* = 1700, *subsample* = 1.0, *subsample ratio of columns for each tree* = 0.8

4. Results

4.1. Prediction performance

The performances of RF and XGB models were evaluated via 10-fold cross-validation. Precision, recall, and F1-score were calculated for each class, and all averages were weighted considering class imbalance because the numbers of instances with low and high level of sales were much lower than that with the medium level. We compared their performances by separating the four-year data into 2018–2019 and 2020–2021 datasets, because we assumed that the two models may work differently with the datasets containing information from pre- and post-COVID-19 periods. As evident from Table 4, XGB performed better than RF, achieving higher predictive accuracy, precision, recall, and F1-score for both datasets. Interestingly, although RF and XGB were tuned using the entire four-year data, they exhibited better performances for the 2020–2021 dataset. Their prediction accuracies with errors using 2- to 10-fold cross-validations are shown in Figures 6(a) and (b). The mean accuracy of RF increased until 5-fold cross-validation and fluctuated slightly thereafter. In contrast, the mean accuracy of XGB increased until 7-fold cross-validation and remained almost unchanged thereafter. For both RF and XGB, the difference between the minimum and maximum accuracies was the smallest in 2-fold cross-validation, whereas it was the highest in 10-fold cross-validation. We decided to employ XGB for further analyses using SHAP because a higher number of folds indicates less bias in measuring the prediction error (Rodriguez et al., 2010), and XGB showed higher accuracy than RF in 10-fold cross-validation.

4.2. Importance of influential factors

The relative importance of the top 20 influential factors at all and high levels of sales with mean Shapley values are shown in Figures 7(a) and (b), respectively. For the prediction of sale level, worker population in their 50s, quarters, the total number of all facilities and banks, number of families living in larger-area or higher-priced apartments, resident population in their 60s, and others were found to be relatively important in the pre-COVID-19 period. In contrast, in the post-COVID-19 period, the worker population in their 60s became the second most important influential factor, which was not among the top 20 influential factors in the pre-COVID-19 period. Additionally, the importance of other age groups, such as 10s, 30s, and 40s, in the worker population increased compared to that in pre-COVID-19 period. The percentage of artificial grass and average monthly income were also found to be important in the post-COVID-19 period.

For only predicting high level of sales, the importance of the number of banks, pharmacies, and subway stations, and male and worker population in their 20s increased compared to the prediction of all levels in the pre-COVID-19 period. In the post-COVID-19 period, the number of pharmacies, worker population in their 30s, 40s, and 60s, average monthly income, and the number of families living in apartments priced higher than $600k had higher importance than in the period pre-COVID-19 period. Additionally, the percentage of deciduous forest appeared to be a relatively important influential factor in the post-COVID-19 period.

4.3. Impact of important influential factors on the high level of sales

The impact of the 20 most important influential factors on the prediction of a high level of sales in the pre- and post-COVID-19 periods is shown in Figure 8. In the summary plot, each point represents a Shapley value that indicates an instance of a factor. As the Shapley value of a factor increases, its contribution toward the prediction of the high level of sales increases. The red and blue colors indicate high and low values of factors, respectively. In the pre-COVID-19 period, as the total number of all facilities, banks, subway stations, and male and worker population in their 20s increased, sales in commercial alleys were higher. In contrast, low values of the dynamic population in their 60s and low percentages of natural grass and deciduous forest had a higher contribution toward the prediction of the high level of sales. Larger numbers of families living in apartments priced higher than $100k, $200k, $300k, or $600k aided in the accurate prediction of the high level of sales, whereas smaller number of families living in apartments priced under $100k contributed more toward the high level of sales.

In the post-COVID-19 period, sales in commercial alleys were higher as the numbers worker population (in their 20s, 40s, and 60s), pharmacies, banks, and total facilities and average monthly income increased (Figure 9). However, the low percentage of deciduous forests and low dynamic population on Sundays contributed to the higher prediction of

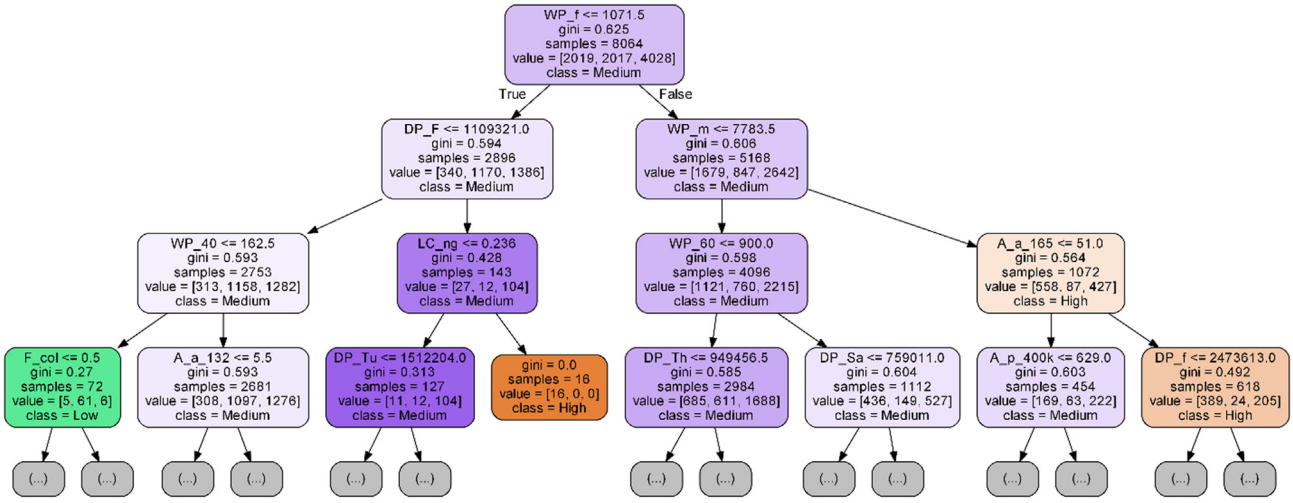

Figure 5. Example of a single decision tree constructed via RF.

Table 4. Performances of the RF and XGB models for the two datasets: 2018–2019 and 2020–2021.

	RF		XGB	
	2018–2019	2020–2021	2018–2019	2020–2021
Accuracy (%)	89.34	91.29	89.89	91.58
Precision (%)	89.46	91.34	89.97	91.61
Recall (%)	89.34	91.29	89.89	91.58
F1 (%)	89.32	91.29	89.88	91.57

the high level of sales. Further, a larger number of families living in apartments priced higher than $200k or $600k aided in the accurate prediction of the high level of sales, whereas a smaller number of families living in apartments priced lower or higher than $100k contributed to the higher prediction of the high level of sales.

Next, we examined the impact of a single factor on predictions with each data instance using the dependence plots shown in Figures 10 and 11 for the pre- and post-COVID-19 periods. We selected clusters and the percentage of deciduous forests, which were newly created for this study, as well as the top seven most important influential factors. As the worker population in their 30s, 40s, and 60s and numbers of pharmacies, banks, and total facilities increased, the prediction of a high level of sales increased until some points in both periods. For example, the prediction increased until the total number of all facilities was 100 in the post-COVID-19 period and then remained constant or gradually decreased. Among the sales quarters, the second quarter showed higher prediction values than the other three quarters in both periods, whereas the fourth quarter aided more in predictions in the pre-COVID-19 period. Additionally, among the clusters, cluster 0, which groups the commercial alleys on the east side of Seoul, achieved a higher prediction value than cluster 1, and this pattern was more evident in the pre-COVID-19 period. For deciduous forests, the prediction value remained constant or decreased as their percentage reached 10 in the pre-COVID-19 period, whereas it increased until their percentage reached 10 in the post-COVID-19 period. A percentage of more than 10 for the deciduous forest in the catchment areas did not aid in the accurate prediction of the high level of sales.

5. Discussion

A noticeable trend in the importance of geographical factors was that the importance of some factors decreased in the post-COVID-19 period. It was discovered that the relative importance of the numbers of all facilities (F_t), families living in apartments priced higher than $600k (A_p_600k), and male worker population (WP_m) in the pre-COVID-19 period dramatically decreased in the post-COVID-19 period, considering the high level of sales in commercial alleys. Particularly, the relative importance of the number of facilities became seventh in the post-COVID-19 period, from first in the pre-COVID-19 period. This indicates that people no longer visit retailed stores or restaurants after reaching their destinations, such as subway stations, hospitals, or schools. The importance of the number of families living in apartments priced higher than $600k and male worker population in the pre-COVID-19 period was also reduced to 10th and 18th, respectively, from 4th and 5th, respectively, in the post-COVID-19 period.

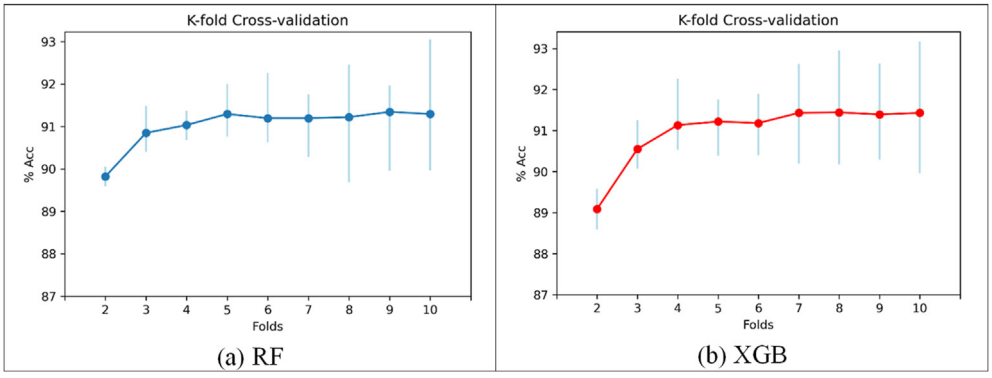

Figure 6. Prediction accuracies with errors of RF and XGB after 2- to 10-fold cross-validations.

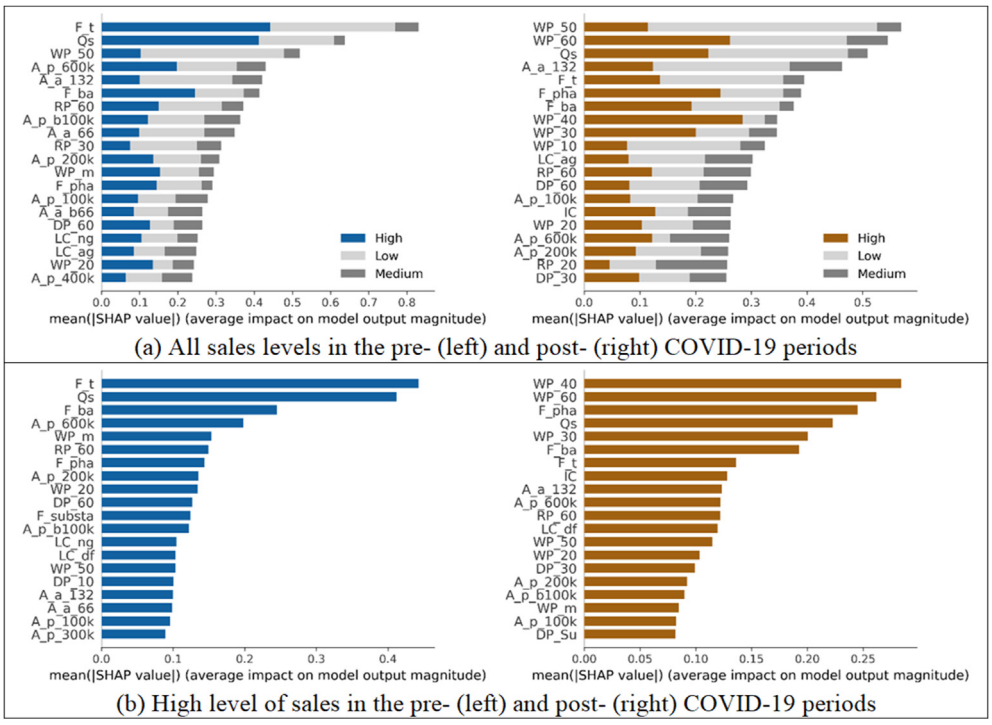

Figure 7. Importance of influential factors in the XGB model.

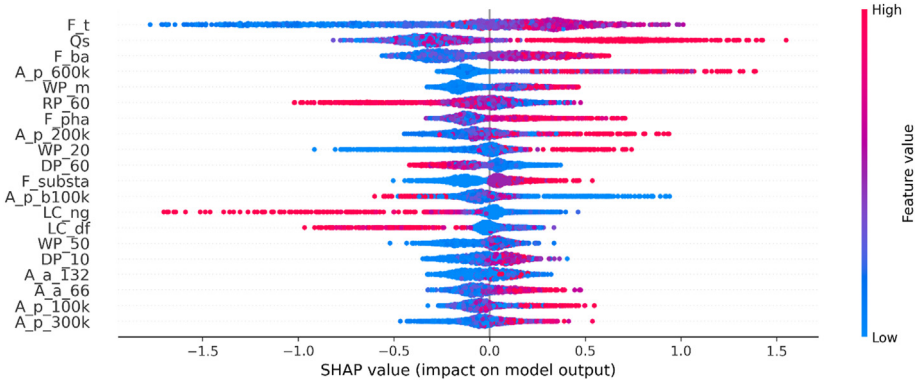

Figure 8. Impact of important influential factors on the prediction of high level of sales in the period before COVID-19.

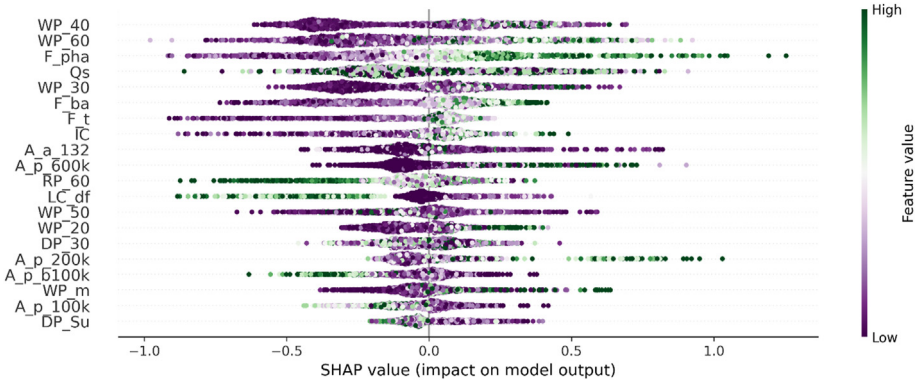

Figure 9. Impact of important influential factors on the prediction of a high level of sales in the period after COVID-19.

However, the importance of some factors did not significantly change in the post-COVID-19 period. The numbers of pharmacies and banks retained high importance for predicting the high level of sales in the post-COVID-19 period. This corroborates the importance of banks and pharmacies in the service strategy of supermarkets in the U.S., as traditional supermarkets in the U.S. have been adding convenience services,

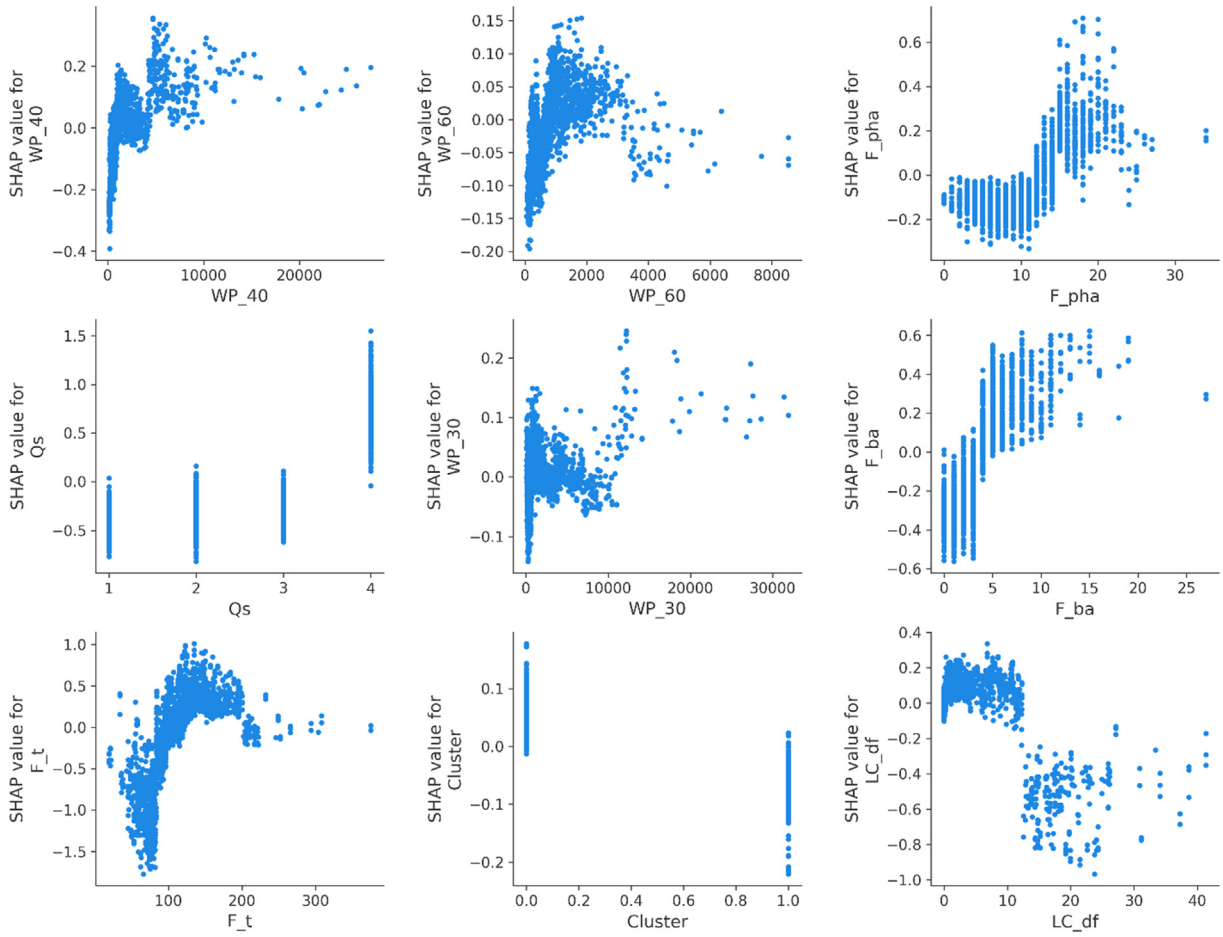

Figure 10. Dependence of influential factors on the prediction of the high level of sales in the pre-COVID-19 period.

including banks and pharmacies, for several decades to offer one-stop shopping (Kinsey and Senauer, 1996). Interestingly, regarding the high level of sales, the worker population comprising middle-aged or older adults and average monthly income were two emergent important factors in the post-COVID-19 period. Regarding the older adult worker population, as they had spent most of their life in an offline world, they remained loyal customers of nearby commercial alleys in the post-COVID-19 period. Average monthly income was determined to be one of the important influential factors in new shopping patterns owing to the change in consumer behavior during the pandemic (Valaskova et al., 2021).

The impact and dependence of important influential factors on the accurate prediction of the high level of sales were interpreted using Shapley values. The worker population in their 60s and total facilities were the two factors that showed considerable changes in Shapley values. The prediction was highest when the worker population in their 60s reached 1,500 in the pre- and post-COVID-19 periods, and then it diminished by degrees. However, when the worker population in their 60s reached 1,500, its Shapley value was much higher in the post-COVID-19 period than pre-COVID-19 period and remained high even after the worker population increased. This indicated the changed importance of worker population in their 60s during the pandemic period. Before the pandemic, the Shapley value of the number of facilities reached 1.0 when the number of facilities reached 150, whereas it increased to only 0.2 when the number of facilities was 100 in the post-pandemic period. This indicates that the number of facilities around commercial alleys is not as important anymore, and a higher number does not contribute to a dramatically high prediction of the high level of sales.

Additionally, some noticeable Shapley values were those of sales quarters and economic status. In the pre-COVID-19 period, the fourth quarter had the highest Shapley value, reaching 1.5, while the second quarter had the highest Shapley value in the post-COVID-19 period. This indicates that there was a large rebound in services and retail sales in Korea in the second quarter of 2021 (Statistics Korea, 2022). The large recovery in retail sales in the second quarter of 2021 was also observed in other countries, such as Canada (Retail Insider, 2021). High levels of economic status, including monthly income and the number of families living in apartments priced higher than $600k, in the surrounding areas of commercial alleys aided in the higher prediction of the high level of sales. Similarly, when the number of families living in lower priced apartments priced (A_p_b100k) was low, the prediction of the high level of sales became accurate. The fact that families with a high level of economic status account for a higher percentage of total annual spending (U.S. Department of Labor Bureau of Labor Statistics, 1998) is also true during the pandemic, although its importance was diminished.

Even though the clusters were found to be not relatively important, the contribution of sales in commercial alleys on the east side of Seoul was higher than the contribution of those on the west side. However, such a pattern was not as clearly observed in the post-COVID-19 contribution compared to that in the pre-COVID-19 period, which indicated that the geographical location of a commercial alley did not significantly affect sales in the post-COVID-19 period, as the sales of several commercial alleys scattered all over Seoul were simultaneously impacted negatively during COVID-19 (Yu, 2021).

Although the role of green space was found to be important during the pandemic period, as proven in a previous study (Lee and Kim, 2021), this study was the first to discover the details of their association. The

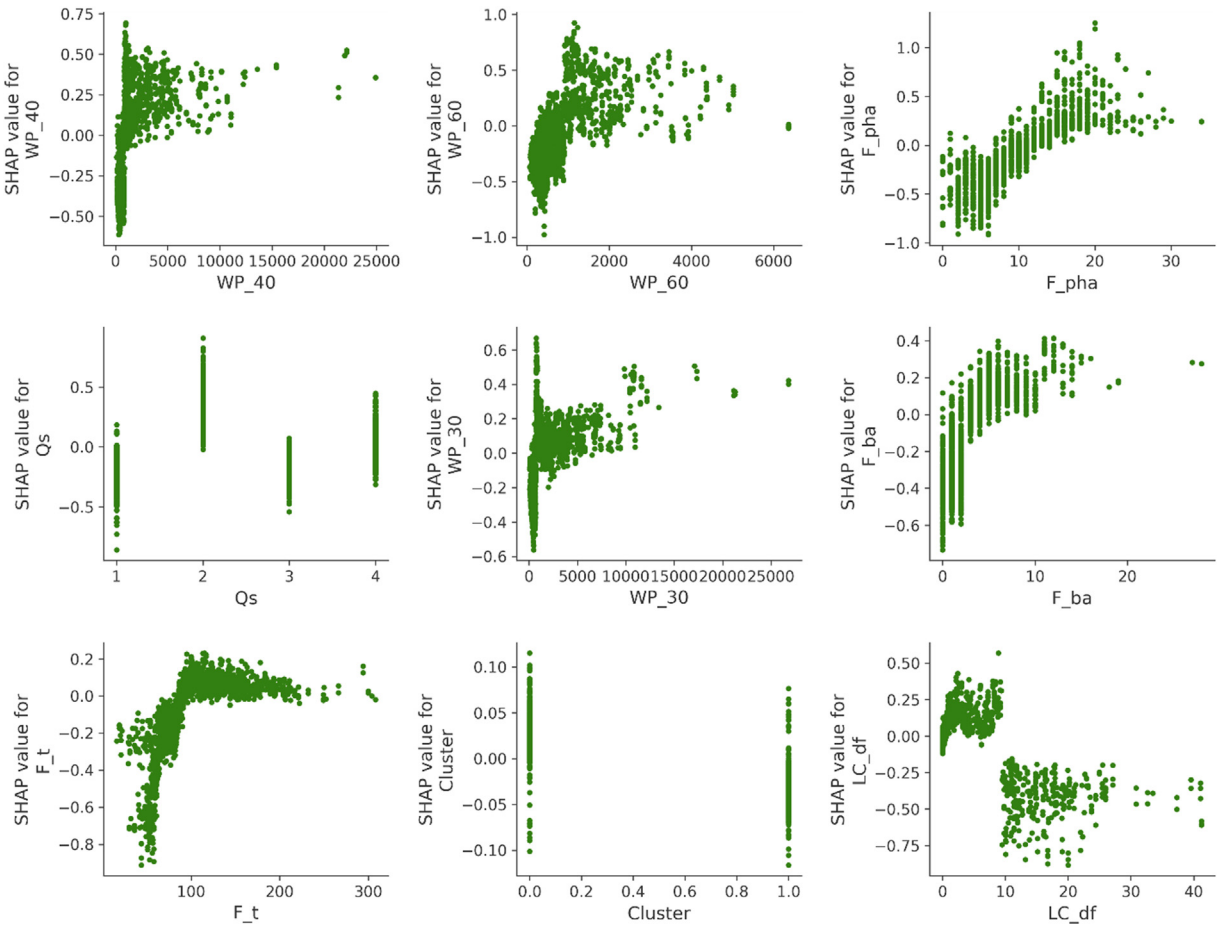

Figure 11. Dependence of influential factors on the prediction of the high level of sales in the post-COVID-19 period.

summary plot indicated that as the percentage of deciduous forests decreased, the prediction increased. However, the dependence plot revealed complex associations in that the Shapley value increased as the percentage of deciduous forests reached 2–3%. Thereafter, the Shapley value decreased as the percentage of deciduous forests increased to 10 in the pre-COVID-19 period, whereas it remained unchanged or increased in the post-COVID-19 period. The Shapley value was below zero after the percentage of deciduous forests reached 10, which indicates that a higher percentage of green environments did not help increase sales in either period. This may be related to the fact that a small percentage of greenery is sufficient to impart a relaxing effect on people, and they do not prefer an excessively high percentage of surrounding greenery (Choi et al., 2016).

6. Conclusions

7. This study examined the effects of changes in important influential factors on commercial alley sales in the pre- and post-COVID-19 periods using machine learning techniques. XGB was found to be the best algorithm among the RF and XGB algorithms that were tested to predict the level of sales, and it showed higher prediction accuracy (91.58%) for the post-COVID-19 dataset than for the pre-COVID-19 dataset (89.89%).

8. As the survival of commercial alleys has become a critical social problem in post-COVID-19 era, this study makes a significant contribution in that it suggests a policy direction that could contribute to the revitalization of commercial alleys in the future and boost the local economy. However, this study has some limitations that should be addressed in future research. First, it did not consider the changing consumption trends post-COVID-19. E-commerce and home delivery services have rapidly emerged during the COVID-19 pandemic, along with a significant increase in Internet usage and social media (Donthu and Gustafsson, 2020; Unnikrishnan and Figliozzi, 2020). Specifically, online sales in California grew by 180% during the pandemic (Fairlie and Fossen, 2022). Home delivery can be an influential factor in store characteristics because it provides an ease of access to services (Turhan et al., 2013). In future studies, it can be used as a salient factor for predicting sales in commercial alleys. Second, this study did not explain how each influential factor affects the sales of different types of businesses. All sales data used in this study included the total sales of different types of businesses, such as retail stores, restaurants, and services, in commercial alleys; thus, the resulting impacts were averaged associations and did not specify the contribution of each influential factor toward each type of business. Therefore, future research should focus on specific types of businesses or consider the business types as influential factors. Third, the factors within the catchment areas were estimated using a pre-defined 200 m circular buffer. However, we need to test different distances or buffer types to validate the reliability of the size and shape of the catchment areas. Therefore, sensitivity analyses should be conducted to investigate the varying associations between influential factors and sales, according to different conditions. Finally, other machine learning algorithms or advanced ensemble learning techniques, such as stacking ensembles, can be used in future research. CatBoost and light gradient boosting machine algorithms are some competitors to XGB. In future studies, we need to compare the performances of these algorithms with that of XGB. Further, we will develop a new innovative machine learning algorithm specialized for GIS data to contribute to the machine learning research domain, rather than using existing traditional machine learning algorithms. A stacking ensemble can combine the predictions of multiple machine learning algorithms to achieve a better prediction performance. By applying this technique, we can increase the performance of the trained model.

Declarations

Author contribution statement

Kangjae Lee: Conceived and designed the experiments; Performed the experiments; Analyzed and interpreted the data; Contributed reagents, materials, analysis tools or data; Wrote the paper.

Funding statement

This work was supported by the Korea Institute of Energy Technology Evaluation and Planning (KETEP) and the Ministry of Trade, Industry & Energy (MOTIE) of the Republic of Korea (No. 20224000000150).

Data availability statement

The data used in this study are available via the Seoul Open Data Plaza website (https://data.seoul.go.kr/).

Declaration of interest's statement

The authors declare no conflict of interest.

Additional information

No additional information is available for this paper.

References

Ainsworth, B.E., Li, F., 2020. Physical activity during the coronavirus disease-2019 global pandemic. Journal of Sport and Health Science 9 (4), 291–292.
Akalin, M., Turhan, G., Sahin, A., 2013. The application of AHP approach for evaluating location selection elements for retail store. International Journal of Research in Business and Social Science 2 (4), 2147–4478, 1–20.
Anzai, A., Kobayashi, T., Linton, N.M., Kinoshita, R., Hayashi, K., Suzuki, A., Yang, Y., Jung, S., Miyama, T., Akhmetzhanov, A.R., Nishiura, H., 2020. Assessing the impact of reduced travel on exportation dynamics of novel coronavirus infection (COVID-19). J. Clin. Med. 9 (2), 601.
Bartik, A., Bertrand, M., Cullen, Z., Glaeser, E., Luca, M., Stanton, C., 2020. How are small businesses adjusting to COVID-19? Early Evidence From a Survey. National Bureau of Economic Research, w26989.
Bartik, A.W., Bertrand, M., Cullen, Z., Glaeser, E.L., Luca, M., Stanton, C., 2020. The impact of COVID-19 on small business outcomes and expectations. Proc. Natl. Acad. Sci. USA 117 (30), 17656–17666.
Bertrand, L., Shaw, K.A., Ko, J., Deprez, D., Chilibeck, P.D., Zello, G.A., 2021. The impact of the coronavirus disease 2019 (COVID-19) pandemic on university students' dietary intake, physical activity, and sedentary behaviour. Appl. Physiol. Nutr. Metabol. 46 (3), 265–272.
Breiman, L., 2001. Random forests. Mach. Learn. 45 (1), 5–32.
Carvalho, D.V., Pereira, E.M., Cardoso, J.S., 2019. Machine learning interpretability: a survey on methods and metrics. Electronics 8 (8), 832.
Chang, H.-J., Hsieh, C.-M., 2014. A TOPSIS model for chain store location selection. Review of Integrative Business and Economics Research 4 (1), 410–416.
Chen, T., Guestrin, C., 2016. XGBoost: a scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining 785–794.
Cheng, E.W.L., Li, H., Yu, L., 2005. The analytic network process (ANP) approach to location selection: a shopping mall illustration. Construct. Innovat. 5 (2), 83–97.
Choi, J.-Y., Park, S.-A., Jung, S.-J., Lee, J.-Y., Son, K.-C., An, Y.-J., Lee, S.-W., 2016. Physiological and psychological responses of humans to the index of greenness of an interior space. Compl. Ther. Med. 28, 37–43.
Chou, T.-Y., Hsu, C.-L., Chen, M.-C., 2008. A fuzzy multi-criteria decision model for international tourist hotels location selection. Int. J. Hospit. Manag. 27 (2), 293–301.
Craig, C.A., 2021. Camping, glamping, and coronavirus in the United States. Ann. Tourism Res. 89, 103071.
Donthu, N., Gustafsson, A., 2020. Effects of COVID-19 on business and research. J. Bus. Res. 117, 284–289.
Dube, K., Nhamo, G., Chikodzi, D., 2021. COVID-19 cripples global restaurant and hospitality industry. Curr. Issues Tourism 24 (11), 1487–1490.
Dunne, P.M., Lusch, R.F., Carver, J.R., 2013. Retailing. Cengage Learning.
Erbıyık, H., Özcan, S., Karaboğa, K., 2012. Retail store location selection problem with multiple analytical hierarchy process of decision making an application in Turkey. Procedia - Social and Behavioral Sciences 58, 1405–1414.
Fairlie, R., 2020. The impact of COVID-19 on small business owners: evidence from the first three months after widespread social-distancing restrictions. J. Econ. Manag. Strat. 29 (4), 727–740.
Fairlie, R., Fossen, F.M., 2022. The early impacts of the COVID-19 pandemic on business sales. Small Bus. Econ. 58 (4), 1853–1864.
Fu, H.-P., Yeh, H.-P., Chang, T.-H., Teng, Y.-H., Tsai, C.-C., 2022. Applying ANN and TM to build a prediction model for the site selection of a convenience store. Appl. Sci. 12 (6), 3036.
Hagger, M.S., Keech, J.J., Hamilton, K., 2020. Managing stress during the coronavirus disease 2019 pandemic and beyond: reappraisal and mindset approaches. Stress Health 36 (3), 396–401.
Hammami, A., Harrabi, B., Mohr, M., Krustrup, P., 2022. Physical activity and coronavirus disease 2019 (COVID-19): specific recommendations for home-based physical training. Managing Sport and Leisure 27 (1–2), 20–25.
Harwati, Utami, I., 2018. Quantitative analytical hierarchy process to marketing store location selection. MATEC Web of Conferences 154, 01075.
He, F., Deng, Y., Li, W., 2020. Coronavirus disease 2019: what we know? J. Med. Virol. 92 (7), 719–725.
Hsu, P.-F., Chen, B.-Y., 2007. Developing and implementing a selection model for bedding chain retail store franchisee using delphi and fuzzy AHP. Qual. Quantity 41 (2), 275–290.
Insider, Retail, 2021. Rocky Rebound for Canadian Retail Sales Q2 into the Summer. Strapagiel. https://retail-insider.com/retail-insider/2021/08/rocky-rebound-for-canadian-retail-sales-q2-into-the-summer-ed-strapagiel/.
Isabelle, D.A., Han (Jade), Y., Westerlund, M., 2022. A machine-learning analysis of the impacts of the COVID-19 pandemic on small business owners and implications for Canadian government policy response. Canadian Public Policy, e2021018.
Kang, H.M., Lee, S.-K., 2018. An analysis of the effects of customer characteristics on sales of alley market area using geographically weighted regression. Journal of the Korean Society of Surveying, Geodesy, Photogrammetry and Cartography 36 (6), 611–620.
Kang, H.M., Lee, S.-K., 2019. Analyzing growth factors of alley markets using time-series clustering and logistic regression. Journal of the Korean Society of Surveying, Geodesy, Photogrammetry and Cartography 37 (6), 535–543.
Kennickell, A.B., Kwast, M.L., Pogach, J., 2015. Small businesses and small business finance during the financial crisis and the great recession: new evidence from the survey of consumer finances. Finance and Economics Discussion Series 2015 (39), 1–94.
Kim, Y., Kim, Y., 2022. Explainable heat-related mortality with random forest and SHapley Additive exPlanations (SHAP) models. Sustain. Cities Soc. 79, 103677.
Kim, H., Lee, S., 2019. A study on the factors affecting the revenue in seoul's side street trade areas. The Seoul Institute 20 (1), 117–134.
Kinsey, J., Senauer, B., 1996. Consumer trends and changing food retailing formats. Am. J. Agric. Econ. 78 (5), 1187–1191.
KoÃ, E., Burhan, H.A., 2015. An application of analytic hierarchy process (AHP) in a real world problem of store location selection. Adv. Manag. Appl. Econ. 5 (1), 1–4.
Kuo, R.J., Chi, S.C., Kao, S.S., 2002. A decision support system for selecting convenience store location through integration of fuzzy AHP and artificial neural network. Comput. Ind. 47 (2), 199–214.
Lee, J.H., Kim, H.W., 2021. Examining the role of urban parks in the post-COVID-19 era through the assessment of alley market district sales. Journal of the Korean Urban Management Association 34 (3), 135–157.
Li, Y., Liu, L., 2012. Assessing the impact of retail location on store performance: a comparison of Wal-Mart and Kmart stores in Cincinnati. Appl. Geogr. 32 (2), 591–600.
Li, J., Nguyen, T.H.H., Coca-Stefaniak, J.A., 2021. Coronavirus impacts on post-pandemic planned travel behaviours. Ann. Tourism Res. 86, 102964.
Liguori, E.W., Pittz, T.G., 2020. Strategies for small business: surviving and thriving in the era of COVID-19. Journal of the International Council for Small Business 1 (2), 106–110.
Lloyd, S., 1982. Least squares quantization in PCM. IEEE Trans. Inf. Theor. 28 (2), 129–137.
Lundberg, S., Lee, S.-I., 2017. A unified approach to interpreting model predictions. Adv. Neural Inf. Process. Syst. 30.
Mangalathu, S., Hwang, S.-H., Jeon, J.-S., 2020. Failure mode and effects analysis of RC members based on machine-learning-based SHapley Additive exPlanations (SHAP) approach. Eng. Struct. 219, 110927.
Manowan, D., Manowan, V., Hengmeechai, P., 2022. Using the AHP method to evaluate laundromat store location selection: a case study in Bangkok Metropolitan Region. ABAC Journal 42 (1), 121–141.
Nicola, M., Alsafi, Z., Sohrabi, C., Kerwan, A., Al-Jabir, A., Iosifidis, C., Agha, M., Agha, R., 2020. The socio-economic implications of the coronavirus pandemic (COVID-19): a review. Int. J. Surg. 78, 185–193.
Reinartz, W.J., Kumar, V., Reinartz, W.J., Kumar, V., 1999. Store-, market-, and consumer-characteristics: the drivers of store performance. Market. Lett. 10 (1), 5–23.
Rodriguez, J.D., Perez, A., Lozano, J.A., 2010. Sensitivity analysis of k-fold cross validation in prediction error estimation. IEEE Trans. Pattern Anal. Mach. Intell. 32 (3), 569–575.
Rodríguez-Pérez, R., Bajorath, J., 2020. Interpretation of machine learning models using shapley values: application to compound potency and multi-target activity predictions. J. Comput. Aided Mol. Des. 34 (10), 1013–1026.
Rodríguez-Rey, R., Garrido-Hernansaiz, H., Collado, S., 2020. Psychological impact and associated factors during the initial stage of the coronavirus (COVID-19) pandemic among the general population in Spain. Front. Psychol. 11, 1540.
Rousseeuw, P.J., 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. J. Comput. Appl. Math. 20, 53–65.
Rudin, C., 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat. Mach. Intell. 1 (5), 206–215.

Singh, J., Tyagi, P., Kumar, G., Agrawal, S., 2020. Convenience store locations prioritization: a fuzzy TOPSIS-GRA hybrid approach. Modern Supply Chain Research and Applications 2 (4), 281–302.

Statistics Korea, 2022. The Index of Services and Retail Sales Index by Province in the Fourth Quarter of 2021. https://kostat.go.kr/portal/eng/pressReleases/14/1/index.board?bmode=read&bSeq=&aSeq=417423&pageNo=1&rowNum=10&navCount=10&currPg=&searchInfo=&sTarget=title&sTxt=.

Tobler, W.R., 1970. A computer movie simulating urban growth in the detroit region. Econ. Geogr. 46 (sup1), 234–240.

Turhan, G., Akalın, M., Zehir, C., 2013. Literature review on selection criteria of store location based on performance measures. Procedia - Social and Behavioral Sciences 99, 391–402.

Tzeng, G.-H., Teng, M.-H., Chen, J.-J., Opricovic, S., 2002. Multicriteria selection for a restaurant location in Taipei. Int. J. Hospit. Manag. 21 (2), 171–187.

Unnikrishnan, A., Figliozzi, M.A., 2020. A study of the impact of COVID-19 on home delivery purchases and expenditures. Working Paper, Civil and Environmental Engineering, Portland State University.

U.S. Department of Labor Bureau of Labor Statistics, 1998. Spending patterns of high-income households. Issues in Labor Statistics.

Valaskova, K., Durana, P., Adamko, P., 2021. Changes in consumers' purchase patterns as a consequence of the COVID-19 pandemic. Mathematics 9 (15), 1788.

Verbeke, A., Yuan, W., 2021. A few implications of the COVID-19 pandemic for international business strategy research. J. Manag. Stud. 58 (2), 597–601.

Yu, H.J., 2021. A study on COVID-19 and effects factors concerning the sales of side-street trade areas in Seoul. Journal of The Korean Regional Development Association 33 (3), 45–75.

Zhang, X., Liu, L., Lan, M., Song, G., Xiao, L., Chen, J., 2022. Interpretable machine learning models for crime prediction. Comput. Environ. Urban Syst. 94, 101789.