Hindawi
Journal of Advanced Transportation
Volume 2022, Article ID 6458371, 12 pages
https://doi.org/10.1155/2022/6458371

Research Article

Exploring Transit Use during COVID-19 Based on XGB and SHAP Using Smart Card Data

Eun Hak Lee

Multimodal Planning & Environment Division, Texas A&M Transportation Institute, 1111 Rellis Pkwy, Bryan, TX 77807, USA

Correspondence should be addressed to Eun Hak Lee; e-lee@tti.tamu.edu

Received 5 May 2022; Revised 15 July 2022; Accepted 30 July 2022; Published 9 September 2022

Academic Editor: Elżbieta Macioszek

As the coronavirus (COVID-19) pandemic continues, many protective measures have been taken in Seoul, Korea, and around the world. This situation has drastically changed lifestyle and travel behavior. An important issue concerns understanding the reasons for giving up transit use and the potential impact on travel patterns during the COVID-19 pandemic. To shed light on these issues that are essential for transit policy, this study explores transit use choice, such as whether users have given-up transit use or not, during the COVID-19 pandemic. Two days of smart card data, before and during the COVID-19 pandemic, were used to look at users who gave up transit use during the COVID-19 pandemic. The choice set of the dataset includes two alternatives, for example, transit use and given-up transit use. An extreme gradient boosting (XGB) model was used to estimate the transit use behavior. Shapley additive explanations were performed to interpret the estimation results of the XGB model. The results for the overall specificity, sensitivity, and balanced accuracy of the proposed XGB model were estimated to be 0.909, 0.953, and 0.931, respectively. The feature analysis based on the Shapley value shows that the number of origin-to-destination trip feature substantially impacts transit use. As such, users tend to avoid transit use as travel time increased during the COVID-19 pandemic. The proposed model shows remarkable performance in accuracy and provided an understanding of the estimated results.

1. Introduction

The global coronavirus (COVID-19) pandemic has profoundly impacted all areas of people's lives around the world. Unlike conventional viruses, the spread of COVID-19 has been difficult to contain, and it is expected to change the appearance of our society permanently rather than temporarily. The first COVID-19 case in Seoul, Korea was reported on January 20, 2020. A year and three months later, in April 2021, 106,898 confirmed cases and 1,756 deaths have been reported in Seoul. To decelerate the spread of COVID-19, the government of Seoul implemented protective measures such as interpersonal distance. The interpersonal distance policy consists of 1~2.5 levels connected with the severity of the spread of COVID-19. Currently, the government of Seoul has announced a 2.5 level which is the strictest lockout measure. This policy significantly changed the lifestyle and travel behavior of local residents in Seoul. The protective measures implemented by the government included closing all facilities, for example, restaurants or gyms, at 10 P.M. and prohibiting gatherings of more than four people. Furthermore, public transport reduced fleet operations by 30% after 9 P.M. According to statistical analysis using smart card data in Seoul, the number of transit trips in 2020 decreased by about 27% compared with the previous year. However, descriptive statistics from smart card data do not provide information on how or why people change their travel behavior, such as given-up transit use. Due to the specificity of the current pandemic situation, little is known about the changes in transit user travel behavior. It is important to understand the reasons for changing travel behavior and the potential impact on transit use during the COVID-19 pandemic.

Early studies addressed the impacts of the COVID-19 pandemic on travel behavior, mode choice, and other activities in different countries worldwide [1–9]. Many studies have focused on travel pattern changes considering work and shopping behaviors. Due to the COVID-19 pandemic, the proportions of telecommuting usage and online

shopping have increased, resulting in a decrease in overall transit and auto trips [10]. To understand this phenomenon, survey and mobile application data were used, however, little work has been published about the change in travel behavior during the interpersonal distance policy period with large-scale data such as smart card data. Many researchers have sought to estimate users' mode choice behavior using smart card data due to its quality and quantity. For example, Kim et al. [11] proposed an express train choice model based on smart card data, and the results of the model showed notable performance in exploring user choice behavior. Lee et al. [12] identified user preference of urban transit modes with the smart card data. Similarly, Jánošíková et al. [13] developed a transit route choice model based on the multinomial logit model (MNL) using smart card data. These previous studies implied that the smart card data was very useful for analyzing mode choice behavior since it accurately provides all transit trip information.

As various data on transit systems are being collected, some studies have explored transit user travel behavior based on a data-driven approach and machine learning techniques [14]. The choice model based on machine learning techniques has an advantage with high accuracy compared to conventional choice models such as the MNL model [15]. One of the major drawbacks of machine learning techniques is the difficulty in interpreting the impact of the inputs on the outputs. However, it has become possible to accurately estimate and analyze various individual travel behaviors with the advent of interpretable machine learning (IML) techniques. For example, Lee et al. [14] used the IML approach to analyze train choice, for example, local and express train, and interpret user preferences. Similarly, Wang and Ross [15] developed an IML-based transit mode choice model and compared it to the MNL model. These studies mentioned that IML provided a more accurate estimation and a better understanding of user preference than other conventional models.

To shed light on these matters that are essential for analysis and transit policy, this study explores transit user's travel behavior, specifically whether or not transit use is given-up, during the COVID-19 pandemic. Two days of smart card data, before and during the COVID-19 pandemic, are used to estimate trips that gave up transit use during the COVID-19 pandemic. The choice set of the dataset includes two alternatives, given-up transit use due to COVID-19 pandemic and transit use. The extreme gradient boosting (XGB) model is used to estimate the transit user's travel behavior. Shapley additive explanations (SHAP) are performed to explain the estimation results of the transit use choice model. Feature importance and relationships between features are investigated by a SHAP summary and dependence plot, respectively. Also, the O-D pairs where the potential for high given-up of transit use were identified in terms of policy implementation.

2. Data Description

2.1. Description of Smart Card Data. The government of Seoul has operated an integrated automatic fare collection system since 2004. The transit fare from the origin to the destination station is charged based on the total distance traveled by transit modes, for example, bus, subway, or both modes (transfer between bus and subway). With a smart card, users can use any combination of transit modes for free up to four transfers. The transit network in Seoul is operated with only a 100% smart card system without any other payment method, for example, cash and ticket, and the smart card data in Seoul provides 99% of transit users' trip information. Thus, it is widely used for microscopic user behavior analysis [16–18].

One of the biggest advantages of the smart card system in Seoul is that users must tap their smart card in or out when they get in or out of transit mode, respectively. If users do not tap in or out their smart card, a double fee will be charged on the next trip as a fine. Thus, the smart card data in Seoul is considered complete and reliable data that records complete transit user information. However, behind these advantages lies the disadvantage of privacy. If someone knows when and where an individual has used transit mode, even roughly, their trip information can be tracked in smart card data. Thus, the government of Seoul implemented a privacy protection policy for smart card data in 2020. The identification of the individual user was deleted to protect the identification of the user's trip sequence and chain. Also, travel time information is recorded every 5 minutes unit, and locations are encrypted with codes that are not identifiable by the general public. Through this privacy protection policy, smart card data has been advanced by recording transit users' information while protecting personal information.

Although the AFC system provides high-quality trip information, limitations of smart card data remain. For example, smart card data typically underestimate ridership owing to possible fare evaders [19, 20]. There also can be anomalies in smart card data due to software problems with the AFC system. These limitations are common that can occur in transit systems around the world. The smart card data used in Seoul also faced this problem, and the government of Seoul estimates anomalies in smart card data to be about 1%. Thus, this study assumed that the smart card data in Seoul contained 99% of transit trips in Seoul without anomalies.

The smart card data from November 14, 2019 and December 10, 2020 were used to analyze the impact of the COVID-19 pandemic on transit mode choice. The smart card data of November 14, 2019 was used as data before the COVID-19 pandemic, and the smart card data of December 10, 2020 was used as data during the COVID-19 pandemic. According to the smart card data, the number of trips before COVID-19 and during the COVID-19 pandemic were 8,196,311 and 4,780,953, respectively. This smart card data indicates that the spread of COVID-19 in Seoul decreased the number of transit trips by about 43% per day. The information for each trip is classified into 16 categories in the smart card data. The description of smart card data and transit network in Seoul is shown in Table 1 and Figure 1, respectively.

2.2. Data Preprocessing. The smart card data of November 14, 2019 and December 10, 2020 were used to estimate the impact of the COVID-19 pandemic on given-up transit use. There are two choice alternatives for transit use, for example, given-up and transit use. However, it is necessary to add

Table 1: Description of the smart card data.

No.	Categories	Description
1	Transaction ID	Unique ID for each transaction
2	Mode code	1: subway, 2: bus, 3: both modes (bus + subway)
3	Line ID	Unique ID for each line
4	Vehicle ID	Unique ID for each bus vehicle (not for subway)
5	Boarding station ID	Unique ID for each station (max five stations are recorded for a trip)
6	Alighting station	Unique ID for each station (max five stations are recorded for a trip)
7	Boarding date/time	Year/month/day/hour/minute/seconds
8	Alighting date/time	Year/month/day/hour/minute/seconds
9	Total travel time	Seconds
10	Total travel distance	Kilometer
11	Number of transfers	0~4 (max four transfers available for a trip)
12	User group	1: general, 2: student, 3: elderly
13	User count	The number of boarding users (for bus trip)
14	Boarding fare	The basic fare for boarding
15	Alighting fare	Additional fare based on the travel distance
16	Zone code	Administrative unit code

Figure 1: Transit network in Seoul.

alternatives such as given-up transit use to take into account the 30% change representing reduced transit trips due to the emergence of the COVID-19 pandemic. Since smart card data only contains the revealed trip information, there is no information about given-up trips due to COVID-19. To obtain information regarding given-up trips, data preprocessing was performed to combine two smart card data sets before and during the COVID-19 pandemic.

Before preprocessing data to obtain the given-up trips for 2020, the data for 2018 and 2019 were compared to identify the yearly change in travel patterns. The results of the comparison showed that the number of trips on November 15, 2018 and November 14, 2019 were 8,268,438 and 8,196,311, respectively. The average travel time and the number of transfers were the same as 30 minutes and 0.24 transfers, respectively. Overall, the difference in travel behavior between November 15, 2018 and November 14, 2019 was less than 0.9%. Thus, the data preprocessing was performed, assuming that the travel pattern in 2019 and 2020 would be the same in the absence of COVID-19. Since the travel time information of smart card data in 2020 is recorded every 5 minutes due to the privacy policy, the travel time information of smart card data in 2019 was also recalculated from seconds unit to 5 minutes unit.

The data preprocessing was performed in two stages. The first stage selected O-D pairs containing given-up trips in 2020. The second stage is to filter the number of given-up trips from the 2019 data and fill it into the 2020 data. Each trip of 2020 was compared to that of 2019 based on departure time, arrival time, mode, number of transfers, and travel time. Departure time, arrival time, and travel time were aggregated by the units of hours when compared for each trip. As a result of the comparison, only trips that existed in the 2019 data were selected as given-up trips.

Figure 2 shows an example of the data preprocessing performed in this study. Firstly, the number of trips of O-D pairs by travel modes was calculated in stage 1. Five O-D pairs were selected as the O-D pairs of trips that are reduced during the COVID-19 pandemic. For the O-D pair from station 1 to station 2, the number of subway trips decreased from 100 to 70. In stage 2, 100 trips from 2019 data and 70 trips from 2020 data were compared based on the departure time, arrival time, mode, number of transfers, and travel time. Among the trips that existed only in the 2019 data, 30 trips were selected as given-up trips. The mode code of a filled trip was set to 0 which refers to the trip that was given-up in the transit use of 2020.

The number of given-up trips was estimated to be 3,415,358. By adding information of trips that were given-up in the transit use of 2020 data, 8,196,311 trips were obtained as the sample. With the preprocessed data, seven features were calculated for each trip to explore changes in transit use choice. The number of O-D trips and the difference between the number of O-D trips were 63.6 and 44.7 trips, respectively, on average. The number of transfers was 0.35 times, on average. The travel time and fare were 27.5 minutes and 1,111 KRW, respectively, on average. The average departure and arrival times were 13:46 and 14:14, respectively. A description and descriptive statistics of the preprocessed data are shown in Table 2.

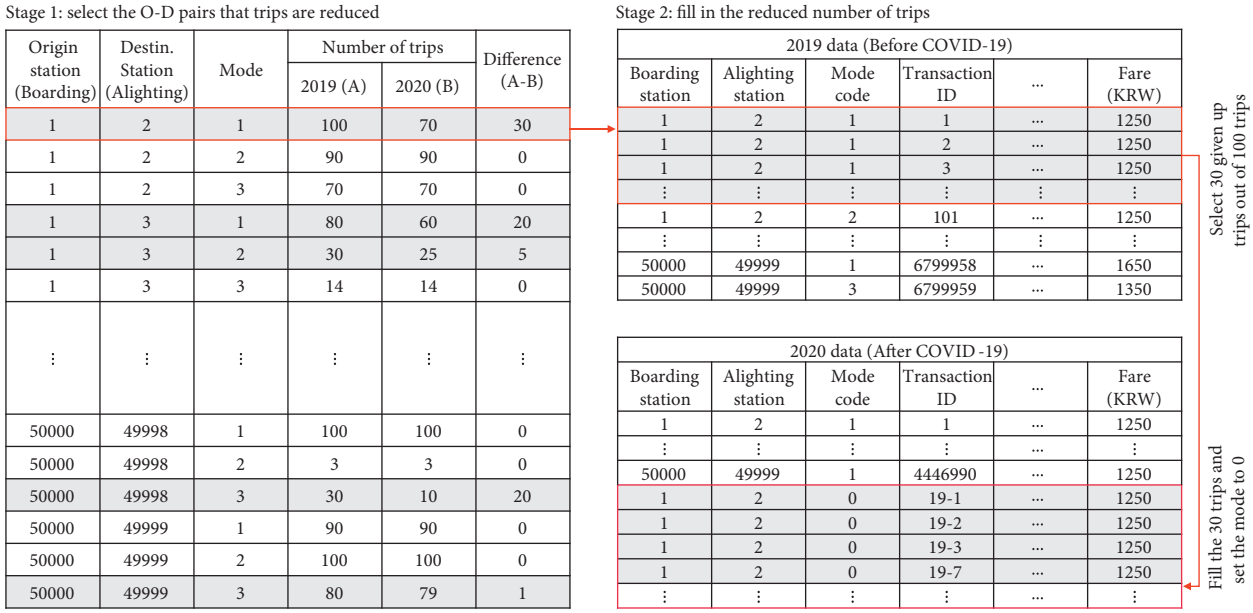

Figure 2: Conceptual illustration of data preprocessing.

Table 2: Description and descriptive statistics of the preprocessed data.

Variable	Category	Count (ratio, %)	Mean
Transit use (choice)	0: given-up transit use	3,415,358 (42.7)	—
	1: transit use	4,780,953 (58.3)	—
	Total	8,196,311 (100.0)	—
Number of O-D trips	Number of trips from origin to destination	—	63.6
Difference between the number of O-D trips	Difference between number of O-D trips before and during COVID-19 pandemic	—	44.7
Number of transfers	Number of transfers from origin to destination	—	0.35
Travel time (minutes)	Travel time from the origin station to the destination station	—	27.5
Fare (KRW)	Fare from origin station to destination station	—	1,111
Departure time (hour)	User's first tap-in time	—	13:46
Arrival time (hour)	User's final tap-out time	—	14:14

3. Methodology

3.1. Extreme Gradient Boosting. Extreme gradient boosting (XGB) proposed by Tianqi Chen and Carlos Guestrin refers to an ensemble machine learning algorithm used for classification or regression predictive modeling problems [21]. XGB is regarded as the most efficient decision tree-based algorithm for data analysis competitions due to its speed and scalability [22]. XGB constructs a sequence of the low-depth decision tree, and each tree is trained to give more weight on the incorrect output of the previous trees. Also, XGB provides parallel tree boosting to solve large-scale problems in a fast and accurate way.

The dataset with 8,196,311 samples includes independent variables x_i and dependent variables y_i, for example, 0 for given-up transit use trip and 1 for transit use trip, $(D = \{(x_i, y_i)\}, |D| = 8,196,311)$. Each x_i has m features therefore $x_i \in \mathcal{R}^m$ ($m = 1$: number of O-D trips, 2: difference between number of trips before and during COVID-19, 3: number of transfers, 4: travel time, 5: fare, 6: arrival time, 7: departure time). These features have corresponding dependent variables such as transit use or given-up ($x_i \in \mathcal{R}^m$, $y_i \in \mathcal{R}$). The tree ensemble model estimates the target value $(\hat{y}_i)$ using f_k which is an Kth independent tree structure with leaf scores as shown in the following equation:.

$$\hat{y} = \phi(x_i) = \sum_{k=1}^{K} f_k(x_i), f_k \in F, \tag{1}$$

where f_k is an independent tree structure with leaf scores and F represent the space of trees. The objective of the model is to minimize $\mathcal{L}(\phi)$ with the loss function l and the mathematical expression of the objective is shown in the following equation:.

$$\mathcal{L}(\phi) = \sum_{i} l(y_i, \hat{y}_i) + \sum_{k} \Omega(f_k). \tag{2}$$

Here, Ω is the term which penalizes the complexity of the model calculated and the mathematical expression of the objective is shown in the following equation:

$$\Omega(f_k) = \gamma T + \frac{1}{2}\lambda w_i^2. \tag{3}$$

In equation (3), w_i is the score of the leaf i and T is the number of leaves. By solving equations (1)–(3), the optimal weight w_i^* and the corresponding value $\widetilde{\mathscr{L}}^t(q)$ are shown the following equations:

$$w_i^* = -\frac{\sum_{i\in I_j}\partial_{\hat{y}^{t-1}} l\left(y_i, \hat{y}^{t-1}\right)}{\sum_{i\in I_j}\partial^2_{\hat{y}^{t-1}} l\left(y_i, \hat{y}^{t-1}\right) + \lambda}, \tag{4}$$

$$g_i = \partial_{\hat{y}^{(t-1)}} l\left(y_i, \hat{y}^{t-1}\right), \tag{5}$$

$$h_i = \partial^2_{\hat{y}^{(t-1)}} l\left(y_i, \hat{y}^{t-1}\right), \tag{6}$$

$$\widetilde{\mathscr{L}}^t(q) = -\frac{1}{2}\sum_{j=1}^{T}\frac{\left(\sum_{i\in I_j} g_i\right)^2}{\sum_{i\in I_j} h_i + \lambda} + \gamma T. \tag{7}$$

It is generally difficult to enumerate all possible tree structures of q. Thus, the greedy algorithm, which branches out a single leaf to many branches iteratively, is used to estimate the optimal solution. The greedy algorithm is usually used to evaluate spilled candidates. $I = I_L \cup I_R$, I_L is the instance set of left nodes after split and I_R is the instance set of right nodes after the split. The mathematical expression is shown in the following equation:

$$\mathscr{L}_s = -\frac{1}{2}\left[\frac{\left(\sum_{i\in I_L} g_i\right)^2}{\sum_{i\in I_L} h_i + \lambda} + \frac{\left(\sum_{i\in I_R} g_i\right)^2}{\sum_{i\in I_R} h_i + \lambda} - \frac{\left(\sum_{i\in I} g_i\right)^2}{\sum_{i\in I} h_i + \lambda}\right] - \gamma. \tag{8}$$

The additional advantage of XGB is that it is not affected by multicollinearity. Thus, several variables can be kept, even if these variables capture the same phenomenon in the same system. This is even desirable since feature analysis using SHAP is conducted in this study.

3.2. Hyperparameter Tuning for XGB. There are several hyperparameters related to the XGB model. Hyperparameter tuning XGB is necessary to avoid the overfitting problem and heavy complexity of the model. A grid search based on cross-validation was performed to set the optimal six hyperparameters, for example, number of iterations, learning rate, subsample, colsample_bytree, alpha, and lambda. The learning rate refers to the scale of the weights of each tree, and it changes the impact of each tree to make a robust model. There are two hyperparameters related to preventing the overfitting problem of the model. The first one is the subsample, which stands for the ratio of randomly selected observations for training instances. The other one is the colsample_bytree parameter which is the fraction of columns when constructing each tree. The alpha parameter is the regulation term on weights of $L1$, and lambda is the regulation term on weights of $L2$. As a result of the grid search based on cross-validation analysis, the hyperparameters of XGB in this study were selected as 622 for the number of iterations, 0.3 for learning rate, 0.9 for subsample, 0.9 for colsample_bytree, 0.4 for an alpha, and 0.3 for lambda, respectively.

3.3. Performance Measures for XGB. Three performance measures, for example, specificity, sensitivity, and balanced accuracy, were selected to evaluate the model performance. These measures are well-known composite classification metrics-based methods for evaluating a multiclass classification model.

Specificity is the number of true-negatives from among the true-negatives and false-positives. Sensitivity stands for the true-positives from among the true-positives and false-negatives. Balanced accuracy is the average of sensitivity and specificity. Balanced accuracy is great for the classification problem when the difference between negative and positive samples is large. In this study, true-positive and false-positive stand that the model estimated transit user as transit use (correct) and given-up (incorrect), respectively. The true-negative and false-negative mean that the model estimated given-up user as given-up (correct) and transit use (incorrect), respectively, where TP is true-positive, FP is false-positive, TN is true-negative, and FN is false-negative. The mathematical expressions of three performance measures are shown in the following equations:

$$\text{specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}}, \tag{9}$$

$$\text{sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \tag{10}$$

$$\text{balanced accuracy} = \frac{\text{specificity} + \text{sensitivity}}{2}. \tag{11}$$

3.4. Shapley Additive Explanations for Model Interpretation. SHAP was used to interpret the results of the transit use choice model proposed in this study. The objective of SHAP is to interpret the contribution of each feature to the output [23, 24]. The Shapley values are estimated based on cooperative game theory. The feature values of each sample act as players in a coalition. The Shapley value helps distribute a payoff for all features when each feature might have contributed more or less than the others. The algorithm repeatedly asks the impact of the feature on each output, and the answer is computed as the Shapley value. With the Shapley value, it is possible to interpret the contribution of each feature [25]. To develop an interpretable mode, SHAP uses an additive feature attribution method, for example, an output model is defined as a linear addition of input features. Assuming a model with input features $x_i = (x_1, x_2, \ldots, x_i, x_m)$, where i is the number of input features (e.g., 1: number of O-D trips, 2: difference between number of trips before and during COVID-19, 3: number of transfers, 4: travel time, 5: fare, 6: arrival time, 7: departure time) and the explanation model $g(z')$ with simplified input z'. For transit use subset $S \subseteq N$ (where N stands for the set of all samples), two models are trained to estimate the effect of feature i. The first model $v(S \cup \{i\})$ is trained with feature i

while the other model $v(S)$ is trained without feature i, where $S \cup \{i\}$ and S are the values of input transit use features. The difference in model outputs $v(S \cup \{i\}) - v(S)$ is estimated for each possible subset $S \subseteq N\{i\}$, equation (12) shows the linear function g which is defined by the additive feature function, and equation (13) shows the mathematical expression of the SHAP.

$$g(x') = \theta_0 + \sum_{i=1}^{m} \theta_i X_i', \tag{12}$$

$$\theta_i = \sum_{S \subseteq N\{i\}} \frac{|S|!\,(n - |S| - 1)!}{n!} [v(S \cup \{i\}) - v(S)]. \tag{13}$$

4. Application

4.1. Feature Selection for XGB Model. With data preprocessing, 13 features were gathered to develop the transit use behavior model during COVID-19. The features related to user behavior and sociodemographics were obtained from the smart card data and the open data portal, respectively. With these datasets, the naïve XGB model was developed to select meaningful features to interpret transit users' behavior. The feature selection process consisted of three steps. Firstly, the features were ranked by importance and frequency scores computed from the naïve XGB model. Then, the importance and frequency scores were clustered by the k-means clustering method. Finally, the features were selected based on the significance of the cluster at 99%.

As a result of feature selection analysis, seven features included in four clusters were selected as significant to the model. Specifically, the number of O-D trips feature was estimated to be the most significant feature, with the highest importance and frequency scores of 0.68 and 0.28, respectively. The difference between the number of O-D trips, number of transfers, travel time, arrival time, and departure time features were analyzed to have a significant impact on the output. However, the six sociodemographic-related features, fo example, population, density, number of households and companies, land-use, and average land price, were estimated to have little impact on output with both importance and frequency scores less than 0.5. The results of the feature selection analysis are shown in Figure 3.

4.2. Performance of the Proposed XGB Model. XGB was trained on 85% of the preprocessed dataset and tested on the remaining 15%. The training and test samples were obtained randomly. The training data included 6,966,864 of 8,196,311 trips, and the test data comprised 1,229,447. The performances of the model were estimated to be over 90% in all measures.

The proposed model was designed as a binary problem to classify given-up or transit use trips. However, three measures were classified by transit modes to identify model performance in detail. For subway users, specificity, sensitivity, and balanced accuracy were estimated to be 0.902, 0.903, and 0.903, respectively. The number of true-positives was 179,611 and the number of true-negatives was 310,191. The number of false-positives was 33,610 and the number of false-negatives was 19,209. For bus users, specificity, sensitivity, and balanced accuracy were estimated to be 0.907, 0.987, and 0.947, respectively. The number of true-positives was 193,715 and the number of true-negatives was 230,070. The number of false-positives was 23,501 and the number of false-negatives was 2,484. For both modes (subway + bus) users, specificity, sensitivity, and balanced accuracy were estimated to be 0.932, 0.980, and 0.956, respectively. The number of true-positives was 114,677 and the number of true-negatives was 111,895. The number of false-positives was 8,185 and the number of false-negatives was 2,299.

Overall, specificity, sensitivity, and balanced accuracy were 0.909, 0.953, and 0.931, respectively. These results indicate that the proposed model showed notable performance with an accuracy of over 93.1%. Moreover, the proposed model was found to be suitable for exploring the impact of COVID-19 on transit mode choice. The performance of the proposed model is shown in Table 3.

To compare the results of XGB with the parametric model, the transit use choice model was developed with the MNL model, the method most widely used for modeling choice behavior [14, 15]. The parameters were estimated with 85% of the dataset and validation was performed with 15% of the dataset. Since a multicollinearity problem between variables, three variables, for example, number of O-D trips, number of transfers, and arrival time, were used to develop the MNL model. The result of the MNL model is shown in Table 4. As a result of estimating MNL, the constants of given-up, subway, bus, and both modes were estimated to be 1.095, 0.434, and 0.673, respectively. These results indicated that many people preferred to use transit even during the COVID-19 pandemic. The parameters of the number of O-D trips, the number of transfers, and arrival time were estimated to be 0.0019, −0.1518, and −0.0299, respectively. These parameters indicated that people preferred to give up transit use as the number of trips, the number of transfers increased, and arrival time increased. The F1 score of MNL was estimated to be about 0.706, which is relatively low compared with that of XGB of 0.931. Specifically, the F1 score of MNL tends to be low, in the order of given-up, subway, bus, and both modes. The MNL model could not estimate users' transit use preferences accurately, with a low F1 score of 0.706. This result implied that the MNL model was suitable for simple problems due to the low flexibility of data distribution assumptions. Also, MNL had a limitation in not being able to interpret the relationship between features. However, the XGB model had high flexibility without distribution assumption and the ability to interpret the relationship between features. Thus, the proposed XGB model accurately estimated the transit use behavior, with a high F1 score of 0.931.

4.3. Feature Analysis of the Transit Mode Choice. Shapley values of seven features of the XGB model are illustrated in Figure 4. The features used in the modeling are ordered by their importance in estimating transit use. If the Shapley value is negative, the preference for transit use is low, and if

Figure 3: Results of feature selection analysis: (a) importance score; (b) frequency score.

the Shapley value is positive, the preference for transit use is high.

The results of the feature analysis showed that the number of O-D trips and the difference between the number of O-D trips had the greatest impact on the transit use choice model. The Shapley values of the number of O-D trips showed that the probability of transit use increased as the number of O-D trips increased. Conversely, the Shapley

Table 3: Performance of proposed XGB model.

Modes	Test set			Performance measure		
	Given-up (trips)	Use (trips)	Total	Specificity	Sensitivity	Balanced accuracy
Subway	198,820	343,801	542,621	0.902	0.903	0.903
Bus	196,199	253,571	449,770	0.907	0.987	0.947
Both modes (bus + subway)	116,976	120,080	237,056	0.932	0.980	0.956
Total	511,995	717,452	1,229,447	0.909	0.953	0.931

Table 4: Results of multinomial logit model.

Variable	Constant	Variable			F1 score
		Number of O-D trips	Number of transfers	Arrive time (minutes)	
Given-up	—	−0.0019***	−0.1518***	−0.0299***	0.785
Subway	1.095***	−0.0019***	−0.1518***	−0.0299***	0.717
Bus	0.434***	−0.0019***	−0.1518***	−0.0299***	0.702
Both modes (bus + subway)	0.673***	−0.0019***	−0.1518***	−0.0299***	0.668
Total	—	—	—	—	0.706

Pseudo R^2: 0.48: **represent $p < 0.05$; ***represent $p < 0.01$.

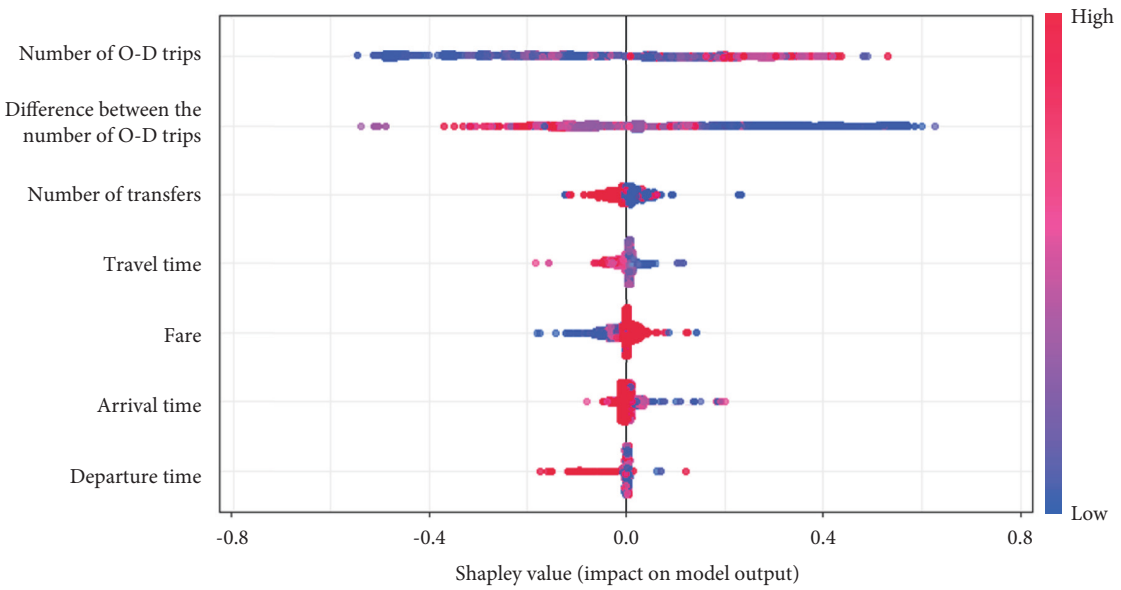

Figure 4: Results of the Shapley values of nine features.

values of the difference between the number of O-D trips indicated that the probability of transit use decreased as the difference between the number of O-D trips increased. The O-D pairs with a high number of transit trips and a low difference between the number of O-D trips had well-equipped transit facilities and had a high level of service (LOS). Thus, users who traveled these O-D pairs did not easily give up their transit use [26].

Among the features related to transit service, for example, number of transfers, travel time, and fare, the number of transfers was found to have the largest impact on transit use behavior. The Shapley value of the number of transfers indicated that the probability of transit use increased as the number of transfers decreased. These results indicated that the users who used both modes gave up transit use since contact with other people could increase as the number of transfers increased. Especially, contact with people was related to the concerns about COVID-19 infection during the pandemic [27]. In the case of the travel time feature, the Shapley value showed that the probability of transit use increased as travel time decreased. The result of the Shapley value for travel time indicated that users avoided long transit times due to concerns about COVID-19 infection [27]. The Shapley value for fare feature showed that users did not prefer transit use as the fare decreased. This result explained that users who use the transit service at a discounted rate, that is, the elderly, disabled, and students, tended to give up transit use during COVID-19. Users who use the transit service at a discounted rate tended to be more health conscious than general users [28]. Thus, elderly, disabled, and student users tended to give up transit use more than general users during the COVID-19 pandemic. The Shapley

values for arrival time and departure time were found to have the least impact on the transit mode choice. These results indicated that the user's departure or arrival time did not significantly affect transit use.

Overall, transit mode preferences were analyzed using eight features, and the impact of each feature was explained. Especially, the presence of COVID-19 had the greatest impact on users that gave up transit use. The impacts of the number of transfers, travel time, and fare on the transit use were also derived from the results, which were consistent with common sense.

4.4. Feature Dependency Analysis of Transit Mode Choice. Travel time was selected as a feature to interpret its impact on transit use since it is one of the most important features to analyze user behavior [1]. Travel time was also the most persuasive to compare the Shapley value by transit modes, since other variables, that is, number of transfers and fare, could vary depending on the modes. The results of feature dependency analysis with travel time and transit use choices, given-up, and transit use were drawn in Figure 5.

In Figure 5(a), subway travel time was selected as a feature to interpret its impact on users who gave up or used the subway during the COVID-19 pandemic. The red points within Circle (1) represent users that gave up transit during the COVID-19 pandemic, and the blue points within Circle (2) represent transit users. Circle (1) described that the Shapley values decreased as subway travel time decreased. The red trips within Circle (1) show the relationship between subway travel time and the Shapley value of subway travel time during the COVID-19 pandemic. The trend for the impact of travel time on subway users was illustrated by a red line. During the COVID-19 pandemic, users tended to give up the subway trip as the travel time increased. The sensitivity of travel time for subway use was estimated to be the highest among that of the transit modes, for example, bus and both modes. The difference between sensitivities of given-up users and transit users was estimated to be the lowest among that of the other modes.

In Figure 5(b), bus travel time was selected as a feature to interpret its impact on users who gave up or used the bus during the COVID-19 pandemic. Circle (3) illustrates the relationship between the travel time of given-up users and Shapley value of travel time. Circle (4) illustrates the relationship between the travel time of transit users and the Shapley value of the travel time. The Shapley value of bus travel time showed that the Shapley value decreased as travel time increased. The trend of the impact of travel time on bus users was illustrated by a red line. During the COVID-19 pandemic, users also tended to give up bus trips as travel time increased. The sensitivity of the travel time for bus use was estimated to be the second-highest among that of the transit modes, for example, subway and both modes.

In Figure 5(c), the travel time of both modes (bus + subway) was selected as a feature to interpret its impact on users who gave up or used both modes during the COVID-19 pandemic. Circle (5) illustrates the relationship between the travel time of given-up users and Shapley value of the travel time. Circle (6) illustrates the relationship between the travel time of transit users and Shapley value of the travel time. Circles (5) and (6) illustrate that the Shapley values decreased as the travel time of both modes increased. The trend of the impact of travel time on both modes users was illustrated by a red line. This result indicated that users did not prefer both modes during the COVID-19 pandemic. However, the sensitivity of the travel time of both modes use was estimated to be very low compared to that of subway and bus.

The overall impact of travel time on transit users is shown in Figure 5(d). Circle (7) illustrates the relationship between the travel time of given-up users and Shapley value of the travel time. Circle (8) illustrates the relationship between the travel time of transit users and Shapley value of travel time. The result of the dependency analysis with the travel time feature indicated that the probability of transit use decreased as travel time increased. Especially, these tendencies in travel time were shown to be more evident for the users that gave up use. Travel time sensitivity was estimated to be high in the order of subway, bus and both modes use. The difference between the sensitivity of given-up and transit users was estimated to be 1.34, 1.69, and 1.88 times, respectively, using linear regression. The difference between the sensitivity of given-up and transit users was high in the order of both modes, bus and subway use. The slopes of the trend line of bus and both modes decreased sharply. These results reflect the behavior of users avoiding spending long travel times in a transit mode due to concerns about infection during the COVID-19 pandemic [5]. These results also implied that the users more easily give up the use of a bus or both modes compared to subway use as travel time increased.

4.5. Discussion. Many countries around the world have implemented policies for the public transportation system as the demand for transit decreased during the COVID-19 pandemic. Specifically, the transit demand in Seoul has been reduced by about 30% during the COVID-19 pandemic. Thus, the government of Seoul considered shortening and reducing the hours of service and dispatching of transit services, respectively. In terms of these practical issues in Seoul, the O-D pairs where the potential for high given-up of transit use was explored using the proposed XGB model. The demand estimation during COVID-19 was performed, and the given-up ratio was calculated for each administrative unit, such as Dong unit. Here, the given-up ratio means a reduction ratio of estimated number of O-D trips during COVID-19 pandemic compared to O-D trips in 2019.

Figure 6 shows the results of the O-D pairs where the potential for high given-up of transit use. The results showed that Jongro and Gangnam areas were the most potential for high given-up of transit use. Specifically, the number of O-D trips between Jongro and Gangnam significantly decreased with a given-up ratio of 0.7~1.0. However, the number of trips from suburban to Jongro or Gangnam was not decreased with a given-up ratio of 0.0~0.2. These results implied that transit use was mostly given-up in the O-D pairs

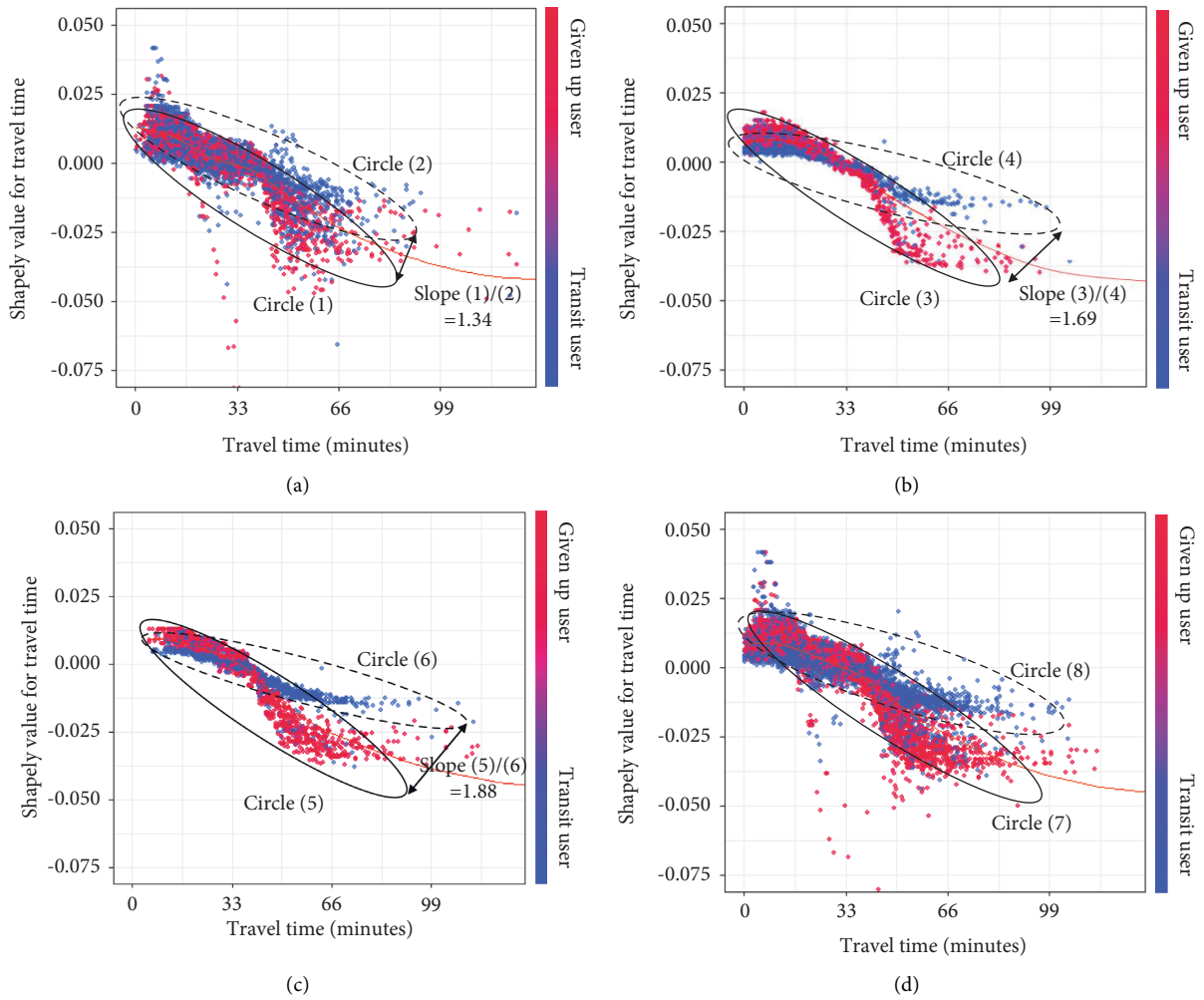

Figure 5: Result of feature dependency analysis. (a) Impact of travel time on subway users during COVID-19 pandemic. (b) Impact of travel time on bus users during COVID-19 pandemic. (c) Impact of travel time on both modes users during COVID-19 pandemic. (d) Overall impact of travel time on transit users during COVID-19 pandemic.

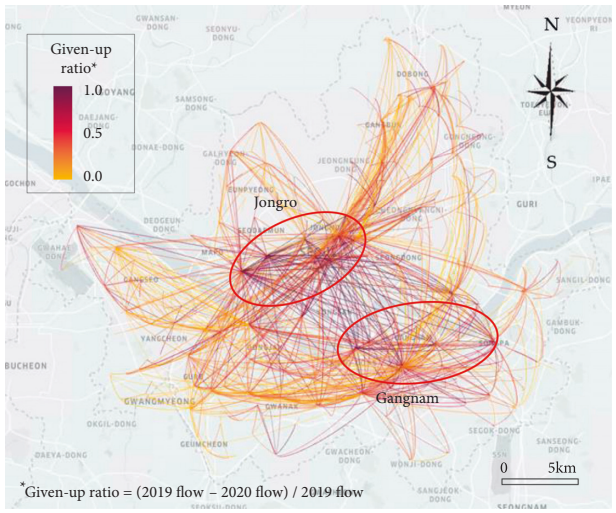

Figure 6: O-D pairs where the potential for high given-up of transit use.

connecting central areas, for example, Jongro and Gangnam, but users still used transit from residential areas to the central areas. From this implication, it could be inferred that users maintain single-purpose trips, such as work trips, but do not have additional business or leisure trips. In terms of transit operation, it is reasonable to reduce hours of service and dispatches of transit services in central areas. Specifically, the transit policies, for example, reduction of hours of service and dispatches of transit services, could be implemented regarding feeder buses in Gangnam and Jongro to improve the operation efficiency. Conversely, the main routes, for example, subway and trunk bus route, connecting the central areas and the suburban areas is not essential to be reduced since the number of trips was not decreased much as inner trips in the central area.

Overall, users tended to give up using transit services when they traveled within the central areas. Thus, it is reasonable to implement transit policies targeting feeder bus routes in central areas to improve operational efficiency.

5. Conclusion

This study aimed to understand the impact of COVID-19 on transit use. Analysis was conducted using two days of smart card data on days, for example, before and during COVID-19 pandemic. With data preprocessing, two alternatives, for example, given-up transit use during the COVID-19 pandemic and transit use, were considered in the choice set. The XGB model was used to train transit preference. Feature analysis based on SHAP was performed to interpret the estimation results from the proposed model. XGB was trained on 6,966,864 of 8,196,311 trips from smart card data and tested on the remaining 1,229,447 trips. The specificity, sensitivity, and balanced accuracy of the proposed model were 0.909, 0.953, and 0.931, respectively. The proposed model was found to be suitable for exploring the impact of COVID-19 on transit use. Feature analysis was performed to explore the impacts of the features on transit use with Shapley values. The number of O-D trips feature was found to impact substantially influence users that gave up transit. Feature dependency analysis was also performed, and the impacts of travel time of the model were identified and interpreted by transit modes. The dependency analysis showed that users gave up transit use as travel time increased. These tendencies in travel time were more evident during the COVID-19 pandemic.

The remarkable performance of XGB supported its ability to estimate the impact of the COVID-19 on transit use. The hyperparameters obtained by the cross-validation conserved the steady low learning error rates in the training of the model. It also derived robust results in estimating transit use. Feature analysis with SHAP provided insights for the proposed model. The Shapley value estimated feature importance and the direction of the impacts. The Shapley value also identified the nonlinear joint impacts of features of the proposed model. There were several interesting findings, such as the COVID-19 pandemic impact on transit use could not be identified by other machine learning techniques. The findings of this study could potentially be helpful and provide implications for policymakers both in mitigating the spread of the disease and establishing appropriate policy that considers travel behavior during the pandemic. With the proposed XGB model, O-D pairs where the potential for high given-up of transit use was identified in terms of policy implementation. As a result, transit use was mostly given-up in the O-D pairs connecting central areas, for example, Jongro and Gangnam. This result implied that it is desirable to implement transit policies targeting feeder bus routes in central areas to improve operational efficiency.

Although the proposed model established notable performance on the estimation of transit use considering users that gave up transit during the COVID-19 pandemic, it would be desirable to consider other external attributes or variables, for example, land-use and sociodemographic features. Understanding additional features would provide a variety of perspectives regarding the impact of the COVID-19 pandemic.

Data Availability

The data used in this research are provided by the Trlab research program conducted at the Seoul National University, Seoul, Republic of Korea. The data are available when readers ask the authors for academic purposes.

Conflicts of Interest

The author declares that there are no conflicts of interest regarding the publication of this article.

Authors' Contributions

Eun Hak Lee conceptualized the study, provided software, wrote the original draft, investigated the study, visualized the study, and validated the study.

References

[1] M. Abdullah, C. Dias, D. Muley, and M. Shahin, "Exploring the impacts of COVID-19 on travel behavior and mode preferences," *Transportation Research Interdisciplinary Perspectives*, vol. 8, p. 100255, 2020.

[2] C. Eisenmann, C. Nobis, V. Kolarova, B. Lenz, and C. Winkler, "Transport mode use during the COVID-19 lockdown period in Germany: the car became more important, public transport lost ground," *Transport Policy*, vol. 103, pp. 60–67, 2021.

[3] N. Askitas, K. Tatsiramos, and B. Verheyden, "Lockdown strategies, mobility patterns and covid-19," 2020, http://arxiv.org/abs/2006.0053.

[4] M. U. G. Kraemer, C. H. Yang, B. Gutierrez et al., "The effect of human mobility and control measures on the COVID-19 epidemic in China," *Science*, vol. 368, no. 6490, pp. 493–497, 2020.

[5] G. Parady, A. Taniguchi, and K. Takami, "Travel behavior changes during the COVID-19 pandemic in Japan: analyzing the effects of risk perception and social influence on going-out self-restriction," *Transportation Research Interdisciplinary Perspectives*, vol. 7, p. 100181, 2020.

[6] J. De Vos, "The effect of COVID-19 and subsequent social distancing on travel behavior," *Transportation Research Interdisciplinary Perspectives*, vol. 5, p. 100121, 2020.

[7] M. de Haas, R. Faber, M. Hamersma, and M. Hamersma, "How COVID-19 and the Dutch 'intelligent lockdown'change activities, work and travel behaviour: evidence from longitudinal data in The Netherlands," *Transportation Research Interdisciplinary Perspectives*, vol. 6, p. 100150, 2020.

[8] A. Shamshiripour, E. Rahimi, R. Shabanpour, and A. K. Mohammadian, "How is COVID-19 reshaping activity-travel behavior? Evidence from a comprehensive survey in Chicago," *Transportation Research Interdisciplinary Perspectives*, vol. 7, p. 100216, 2020.

[9] X. Qu and K. Gao, "Impacts of COVID-19 on the transport sector and measures as well as recommendations of policies and future research: a report on sig-C1 transport theory and modelling," *SSRN Electronic Journal*, pp. 1–12, 2020.

[10] S. Hu, C. Xiong, M. Yang, H. Younes, W. Luo, and L. Zhang, "A big-data driven approach to analyzing and modeling human mobility trend under non-pharmaceutical interventions during COVID-19 pandemic," *Transportation Research Part C: Emerging Technologies*, vol. 124, p. 102955, 2021.

[11] K. M. Kim, S. P. Hong, S. J. Ko, and J. H. Min, "Predicting express train choice of metro passengers from smart card data," *Transportation Research Record*, vol. 2544, no. 1, pp. 63–70, 2016.

[12] E. H. Lee, I. Lee, S. H. Cho, S. Y. Kho, and D. K. Kim, "A travel behavior-based skip-stop strategy considering train choice behaviors based on smartcard data," *Sustainability*, vol. 11, no. 10, p. 2791, 2019.

[13] L. Jánošíková, J. Slavík, and M. Kohání, "Estimation of a route choice model for urban public transport using smart card data," *Transportation Planning and Technology*, vol. 37, no. 7, pp. 638–648, 2014.

[14] E. H. Lee, K. Kim, S. Y. Kho, D. K. Kim, and S. H. Cho, "Estimating express train preference of urban railway passengers based on extreme gradient boosting (XGBoost) using smart card data," *Transportation Research Record*, vol. 34, p. 213214, 2021.

[15] F. Wang and C. L. Ross, "Machine learning travel mode choices: comparing the performance of an extreme gradient boosting model with a multinomial logit model," *Transportation Research Record*, vol. 2672, no. 47, pp. 35–45, 2018.

[16] E. H. Lee, H. Lee, S. Y. Kho, and D. K. Kim, "Evaluation of transfer efficiency between bus and subway based on data envelopment analysis using smart card data," *KSCE Journal of Civil Engineering*, vol. 23, no. 2, pp. 788–799, 2019.

[17] E. H. Lee, H. Shin, S. H. Cho, S. Y. Kho, and D. K. Kim, "Evaluating the efficiency of transit-oriented development using network slacks-based data envelopment analysis," *Energies*, vol. 12, no. 19, p. 3609, 2019.

[18] E. H. Lee, K. Kim, S. Y. Kho, D. K. Kim, and S. H. Cho, "Exploring for route preferences of subway passengers using smart card and train log data," *Journal of Advanced Transportation*, pp. 1–14, 2022.

[19] B. Barabino, C. Lai, and A. Olivo, "Fare evasion in public transport systems: a review of the literature," *Public Transport*, vol. 12, no. 1, pp. 27–88, 2020.

[20] B. Barabino, S. Salis, and B. Useli, "Fare evasion in proof-of-payment transit systems: deriving the optimum inspection level," *Transportation Research Part B: Methodological*, vol. 70, pp. 1–17, 2014.

[21] T. Chen and C. G. Xgboost, "A scalable tree boosting system," in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pp. 785–794, New York, NY ACM 2016, 2016.

[22] A. B. Parsa, A. Movahedi, H. Taghipour, S. Derrible, and A. K. Mohammadian, "Toward safer highways, application of XGBoost and SHAP for real-time accident detection and feature analysis," *Accident Analysis & Prevention*, vol. 136, p. 105405, 2020.

[23] S. M. Lundberg and S. I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems*, pp. 4765–4774, 2017.

[24] E. Štrumbelj and I. Kononenko, "Explaining prediction models and individual predictions with feature contributions," *Knowledge and Information Systems*, vol. 41, no. 3, pp. 647–665, 2014.

[25] L. S. Shapley, "A value for n-person games," *Contributions to the Theory of Games*, vol. 2, no. 28, pp. 307–317, 1953.

[26] S. Hu and P. Chen, "Who left riding transit? Examining socioeconomic disparities in the impact of COVID-19 on ridership," *Transportation Research Part D: Transport and Environment*, vol. 90, p. 102654, 2021.

[27] S. H. Cho and H. C. Park, "Exploring the behaviour change of crowding impedance on public transit due to COVID-19 pandemic: before and after comparison," *Transportation Letters*, vol. 13, no. 5-6, pp. 367–374, 2021.

[28] C. Guida and G. Carpentieri, "Quality of life in the urban environment and primary health services for the elderly during the Covid-19 pandemic: an application to the city of Milan (Italy)," *Cities*, vol. 110, p. 103038, 2021.