sustainability

MDPI

Article

Prediction of Shipping Cost on Freight Brokerage Platform Using Machine Learning

Hee-Seon Jang [1], Tai-Woo Chang [1,*] and Seung-Han Kim [2]

[1] Department of Industrial & Management Engineering/Intelligence & Manufacturing Research Center, Kyonggi University, Suwon 16227, Republic of Korea
[2] Hwamulman Co. Ltd., Gwangju 12777, Republic of Korea
* Correspondence: keenbee@kgu.ac.kr; Tel.: +82-31-249-9754

Abstract: Not having an exact cost standard can present a problem for setting the shipping costs on a freight brokerage platform. Transport brokers who use their high market position to charge excessive commissions can also make it difficult to set rates. In addition, due to the absence of a quantified fare policy, fares are undervalued relative to the labor input. Therefore, vehicle owners are working for less pay than their efforts. This study derives the main variables that influence the setting of the shipping costs and presents the recommended shipping cost given by a price prediction model using machine learning methods. The cost prediction model was built using four algorithms: multiple linear regression, deep neural network, XGBoost regression, and LightGBM regression. R-squared was used as the performance evaluation index. In view of the results of this study, LightGBM was chosen as the model with the greatest explanatory power and the fastest processing. Furthermore, the range of the predicted shipping costs was determined considering realistic usage patterns. The confidence interval was used as the method of calculation for the range of the predicted shipping costs, and, for this purpose, the dataset was classified using the K-fold cross-validation method. This paper could be used to set the shipping costs on freight brokerage platforms and to improve utilization rates.

Keywords: machine learning; shipping cost; freight; price prediction; prediction interval

Citation: Jang, H.-S.; Chang, T.-W.; Kim, S.-H. Prediction of Shipping Cost on Freight Brokerage Platform Using Machine Learning. *Sustainability* **2023**, *15*, 1122. https://doi.org/10.3390/su15021122

Academic Editors: Zhiyuan Liu, Xinyuan Chen and Di Huang

Received: 11 November 2022
Revised: 3 January 2023
Accepted: 4 January 2023
Published: 6 January 2023

1. Introduction

Internet transactions are increasing, and the logistics market is also activated. Many logistics centers have been built, and the parcel forwarding service has grown. As of 2020, the volume of general freight has been continually increasing [1]. The importance of domestic freight transportation using roads has been emphasized, even with the outbreak of COVID-19. The traffic volume of small- and medium-sized vehicles used for freight transportation increased between January and August 2020, after the outbreak of COVID-19, compared with 2019 [2]. Therefore, domestic freight transportation using roads is quite important for stimulating the logistics market. However, there is no accurate standard for the shipping costs in the domestic freight industry. Currently, the criteria for setting the shipping costs simply consider distance and vehicle tonnage. This is only a guideline for new market entrants because it cannot consider various characteristics of freight and is difficult to use in practice. Shipping costs are set based on the shipper's know-how. Shippers set the shipping costs by considering the shipping costs of similar freights in the past and the current market price. Shipping costs are undervalued relative to labor and are unreasonable from the perspective of vehicle owners, and some transportation agents use their high market position to charge excessive commissions [3]. Due to this situation, which is made up of strong disputes between shippers and vehicle owners, current vehicle owners have a strong dissatisfaction.

This paper proposes a machine learning-based shipping cost prediction method for a domestic freight transportation environment using data from a freight brokerage platform. It also shows that predictive models can set the shipping costs appropriately, and it compares the predictive power to present the best predictive model.

We used transportation-related data for 6 months from the freight brokerage platform. To identify the major factors, new factors were added, and various preprocessing methods were applied. Correlational analysis and a step selection method were used to derive the major factors. After that, we developed a fare prediction model using the derived factors with a machine learning algorithm. The machine learning algorithms we used were multiple linear regression (MLR), deep neural networks (DNNs), extreme gradient boosting (XGBoost) regression, and light gradient boosting machine (LightGBM) regression. LightGBM is a model that reduces the learning time compared to the XGBoost model.

We present a method for setting the range of predicted fares considering realistic usage behaviors; the fares should be presented as a range rather than as a single value to the user. A total of 30 training sets were generated using k-fold cross-validation. We trained the sets and predicted the test set for each iteration. Assuming that the 30 derived predicted values follow a normal distribution, a confidence interval was calculated, and an appropriate fare range was presented. R-squared was used as the performance evaluation index for the predictive model.

The structure of this paper is as follows. Section 2 explains the theoretical background and previous research. Section 3 describes the results of the data collection and preprocessing, and Section 4 describes the derivation of the major factors. In Section 5, the model construction and results are explained, and, finally, in Section 6, conclusions and future research directions are presented.

2. Literature Review

2.1. Prior Studies

Kovács [4] calculated road freight shipping costs, which previously had only been estimated. Transport-related factors such as "distance," "fuel," "price," and "highway toll" were selected to calculate the cost. A predictive model based on multiple regression analysis was built using the selected factors, and it demonstrated an excellent predictive performance.

Sternad [5] attempted to extract the major factors that affect road freight shipping costs. Fixed costs related to vehicles and drivers, and variable costs such as "fuel cost," "toll fee," and "mileage" were derived as characteristic factors. Next, the coefficient values for each characteristic factor were derived through multiple regression analysis. As a result, "fuel cost," "travel cost," and "working cost" were found to be major factors that affect the shipping costs.

Li et al. [6] used characteristic factors, such as "vehicle capacity," "delivery location," and "cargo volume," with the mixed constant planning method to optimize the matching and pricing of multidelivery services for the cargo O2O (online to offline) platform. As a result of using the developed optimization technique on data from the Chinese cargo O2O platform, the pickup distance was improved by 75–81%, and the shipping costs were reduced by 60–93%.

Lindsey et al. [7] investigated factors that affect truck fare rates in North America and found that factors such as "distance" and "truck type" were the most important factors for determining the shipping costs.

Price predictions in other fields are studied. Jo et al. [8] selected factors related to housing prices, such as "total lump-sum housing lease price index," "increase in KOSPI (Korea Composite Stock Price Index)," and "consumer price index," to predict changes in housing sale prices. The collected factors were used for logistic regression and random forest algorithms, and an appropriate prediction accuracy was achieved in the dataset. Jang and Park [9] predicted art prices based on eight factors whose correlation with art prices was verified. The algorithms used for the prediction were linear regression and k-nearest neighbor (KNN). The KNN algorithm, a nonparametric model capable of flexible

fitting to the data, showed a better performance in that there were not many variables that were relevant to the art, and it was difficult to assume the distribution of the data due to insufficient information.

As a result of previous studies, factors that affect the shipping cost setting include freight information factors such as "distance," "vehicle type," and "car volume," and additional cost factors such as "delivery location," "fuel cost," and "highway fee." Various algorithms have been used for price prediction. In the case of freight shipping cost forecasting, most studies have used traditional analysis models such as multiple regression analysis and mixed integer programming. In fields other than the shipping cost, research on price prediction methods has been conducted using traditional analysis models and machine learning such as MLR, random forest, and KNN algorithms.

There have been many studies where advanced optimization algorithms have been applied as solution approaches, such as online learning, scheduling, multiobjective optimization, data classification, and others. The effectiveness of these advanced optimization algorithms in the various domains, such as transportation and logistics, and their potential applications for the decision problem have been addressed in the studies [10–13].

Currently, predictive research on freight shipping costs needs further research considering more factors and algorithms. In order to set freight shipping costs, not only freight characteristics but also environmental factors must be considered. Therefore, in this study, factors such as "distance," "vehicle type," and "car volume", that have been considered in previous studies, and environmental factors such as precipitation are included to derive factors that affect how the shipping costs are set. Furthermore, currently, most studies on freight shipping cost prediction are conducted using traditional regression models. Looking at cost prediction studies in other fields, there are many studies using artificial intelligence algorithms. AI algorithms often have a higher accuracy than traditional models. Therefore, it is necessary to advance research by applying artificial intelligence algorithms to the field of freight shipping cost prediction. Therefore, we build a shipping prediction model using the derived factors and artificial intelligence regression algorithm. For this process, we use the k-fold cross-validation and the confidence interval to predict the range of the shipping costs and to increase applicability in the field.

2.2. Theoretical Background

Machine learning is a field of artificial intelligence that analyzes and learns data using algorithms, and it determines or predicts the dependent variables based on what has been learned [14]. According to the learning method, machine learning is categorized as supervised learning and unsupervised learning. Supervised learning is a learning algorithm that learns data with input and output values, and it predicts output values for unseen data or future data. It is used for classification or regression analysis [15].

In this study, MLR, and the supervised learning algorithms DNN, XGBoost regression, and LightGBM regression were used. The definition and characteristics of each algorithm are shown in Table 1.

Table 1. Definition and characteristics of machine learning algorithms used in this study.

Algorithm	Definition	Characteristic
multiple linear regression (MLR)	A statistical technique for estimating a predictive target using a linear relationship between two or more predictive factors (independent variables) for one predictive target (dependent variable) [16].	Predicts the dependent variable using multiple independent variables.
Deep neural network (DNN)	An artificial neural network consisting of many hidden layers between an input layer and an output layer [17].	Like general artificial neural networks, this algorithm models complex nonlinear relationships, and it contains multiple hidden layers [14].

Table 1. *Cont.*

Algorithm	Definition	Characteristic
XGBoost regression	Abbreviation for extreme gradient boosting, an improved algorithm based on the gradient-boosting algorithm [18]. The boosting algorithm is one of the machine learning algorithms used for classification and regression problems. It is an ensemble technique that uses multiple decision trees in combination. Gradient boosting is a method for improving performance by sequentially combining weak learners in the direction of reducing the value of the loss function of the model [19]. The XGBoost algorithm implements gradient boosting to support parallel learning.	It has excellent efficiency, flexibility, and portability, and it can prevent overfitting, which is a disadvantage of the gradient-boosting algorithm.
LightGBM Regression	LightGBM is a gradient boost-based algorithm that includes two techniques, gradient-based one-side sampling (GOSS) and exclusive feature bundling (EFB). GOSS is a new sampling method of the gradient boost algorithm, and it offers the advantage of reducing the number of data instances and maintaining the accuracy of the trained tree. EFB is a dimensional reduction technique to improve efficiency while maintaining a high level of accuracy by bundling exclusive functions as a lossless method of reducing the number of factors.	As one of the ensemble techniques, it is an algorithm that uses a leaf-wise tree partitioning method. In addition to the advantage of the XGBoost algorithm, it has the advantages of reducing the number of instances and factors, which results in a faster calculation speed and lower memory usage. However, there is also the disadvantage that overfitting problems can easily occur when a small dataset is used [20].

3. Data Collection and Preprocessing

3.1. Data Collection

For this study, we collected freight brokerage data that were registered on the freight brokerage platform within the 6 months from April to September 2020. The dataset consists of 1,885,033 data observations and 78 variables used by freight brokerages, such as cargo information, vehicle type, vehicle tonnage, loading date and time, and unloading date and time.

3.2. Data Preprocessing

3.2.1. Creating and Removing Variables

In the dataset, variables that were related to the personal information of the cargo owners and vehicle owners, such as name, vehicle number, and phone number, were deleted, as they were judged to be irrelevant to the shipping cost prediction.

To derive factors that affect the shipping cost and to increase the predictive power of the shipping cost prediction model, variables that were expected to affect the shipping cost were added. The latitude and longitude of the upper and lower location were calculated using the haversine distance formula and were added as a "linear distance." It was judged that detailed date and time information could affect the cost setting, so the arrival and departure dates were subdivided into the month, day, day of the week, and time, and new variables were created for each. In addition, we added the precipitation amount as a new factor, considering that the weather conditions at the time of the cargo transport would affect the shipping cost. The precipitation data of the Korea Meteorological Administration were used, and the precipitation value was added by considering the loading and unloading locations and dates.

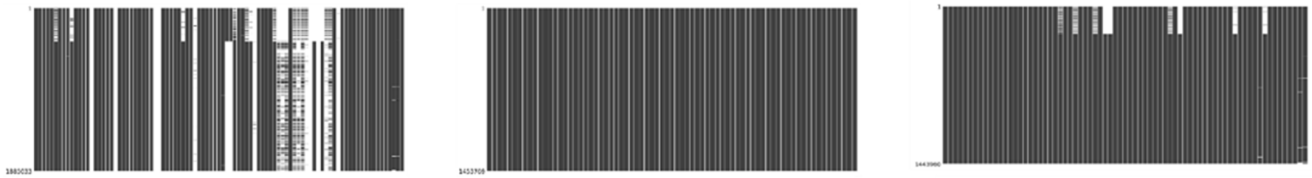
3.2.2. Removing Data Outliers

The interquartile range (IQR) was used to remove outliers in the data. Outlier removal was applied only to continuous variables.

3.2.3. Handling Missing Data

To predict accurate shipping costs, it was important to manage missing values in the input data. After applying two methods for managing missing values, we compared which method was more useful. Before the processing of missing values, factors for which more than 50% of the data were missing were determined to be factors that did not have a great influence on the prediction and were, thus, removed. We removed 20 factors, including "load/unload name address," "summary," and "order number." For the missing value treatment, listwise deletion and the mean imputation were used, and a dataset was created to which each treatment for missing values was applied. The listwise deletion removed all data with missing values, and the mean imputation replaced the missing values with the average value of each factor. After we processed the missing values, the listwise deletion dataset consisted of 73 factors and 1,353,543 data observations, and the mean imputation dataset consisted of 73 factors and 1,442,036 data observations.

Figure 1 shows the dataset before and after preprocessing. Figure 1a is the data form before preprocessing, while Figure 1b,c are the data form after preprocessing. The white part in the figure indicates the missing values, and the black part indicates the data.



(a) Before removing missing values (b) After listwise deletion (c) After mean imputation

Figure 1. Dataset after preprocessing.

4. Derivation of Key Factors

From the 73 factors obtained through the data collection and preprocessing, we attempted to derive the factors that affect the shipping costs. Correlational analysis and step selection were applied as a means to derive the major factors.

4.1. Correlational Analysis

Correlational analysis is a method for analyzing a linear relationship between two variables. When a dependent variable is predicted through several independent variables, a meaningful variable can be selected by considering the correlation between the independent variable and the dependent variable, and the correlations between the independent variables [21]. In this study, independent variables with a correlation coefficient of 0.1 or higher, which is judged to indicate a linear relationship between the independent variable and the dependent variable, were judged to be the main factors.

Table 2 shows the variables that had a linear relationship with the dependent variable "shipping cost," as well as the values of each Pearson correlation coefficient. As a result of the correlational analysis, we found seven significant factors in the dataset, to which the listwise deletion was applied. There were eight significant factors in the dataset, to which the mean imputation was applied, and the factor "phase difference" was added to the significant factors for the listwise deletion. Additionally, for both datasets, it can be observed that the "linear distance" and "actual distance" factors have a high linear relationship with the "shipping cost." Therefore, a shipping cost prediction model was constructed using the significant variables from each dataset.

Table 2. Pearson's correlation coefficient values by factor.

Factor	Listwise Deletion	Mean Imputation
linear distance	0.7322	0.7258
actual distance	0.7062	0.7035
freight weight	0.2760	0.2582
vehicle tonnage	0.2700	0.2372
type of unloading	0.2535	0.2323
standard fare	0.2015	−0.1999
unloading time	−0.1943	0.1562
loading time		0.1023

4.2. Stepwise Method

The stepwise method was one of the methods used for selecting several independent variables to be included in the regression model. It is a method that is used to find the variable constituting the optimal regression model by repeating the addition and removal of variables. The selected variable was judged to be a strong predictor in the prediction model [22]. Table 3 shows the variables selected through the stepwise method. As a result of applying the stepwise selection method to the listwise deletion of the dataset, 35 variables were selected. After we applied the mean imputation to the dataset, there were a total of 33 selected predictors, which were the same predictors found after we applied the listwise deletion and removed the "total cost" and "arranging fee" factors. Thus, a shipping cost prediction model was constructed using the significant variables of each dataset.

Table 3. Independent variables selected by stepwise method.

Mean Imputation Dataset				Listwise Deletion Dataset (Added)
• loading month • loading day • loading time • loading day of the week • loading location • loading latitude • loading longitude • sort sequence • company code	• unloading month • unloading day • unloading time • unloading day of the week • unloading location • unloading latitude • unloading longitude	• linear distance • actual distance • vehicle tonnage • vehicle type • freight weight • type of loading • type of unloading • standard fare • primary key • serial number	• shipping cost payment • payment method • loading classification • dispatch status • share state • shipper number • registrant key	• total cost • arranging fee

5. Model Construction and Analysis Results

5.1. Data Preparation

To ensure the accuracy of the model, all variables were normalized to the same scale. Min–max normalization, which converts all continuous variable data to values between 0 and 1, was used for normalization. We attempted to derive the shipping cost as a range using the confidence interval. In the case of freight, some characteristics were not fully expressed in the data. Therefore, if the recommended cost is presented as a single value, it has limited means to reflect the volatility of reality. To derive the shipping cost as a range, a 95% confidence interval was calculated for the cost value predicted by the shipping cost prediction model. To ensure that the distribution of the predicted values follows a normal distribution, we increased the predicted values (number of samples) using K-fold cross-validation.

For the suitability of the model, 80% of the collected data were allocated to a training set and 20% to a test set, and the datasets were then used for the model construction and verification. At this time, to derive the cost range, the training set was divided into 30 folds

through K-fold cross-validation, and 30 predicted values were derived by predicting the test set for each iteration. For each iteration, 29 training sets and 1 validation set were used. After that, the model was trained on the training set of each fold, and the process of predicting with the test set was repeated until 30 predicted values were derived. Afterward, the maximum and minimum values in the confidence interval were determined as the upper and lower limits of the prediction interval to estimate the predicted cost range. Overall configuration of a dataset for range prediction is shown in Figure 2.

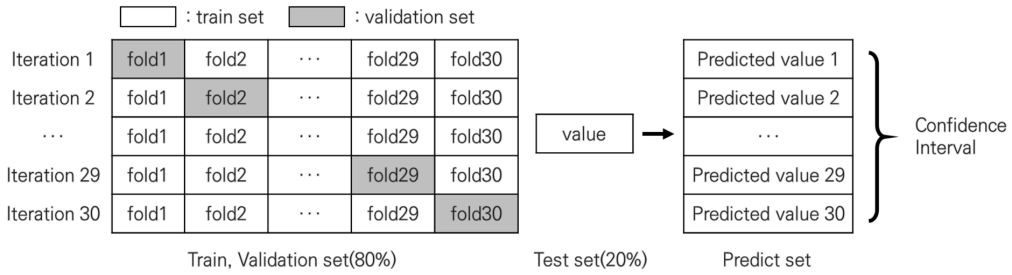


Figure 2. Configuration of a dataset for range prediction.

This study was conducted using Python version 3.9. Python language-based TensorFlow and the scikit-learn machine learning algorithms were also used.

5.2. MLR

Multicollinearity, where two or more independent variables have a high correlation, is among the many assumptions made about the regression analysis model. If there is a correlation between independent variables, the standard error increases, and the variance of the independent variable coefficient increases. Therefore, the process of diagnosing multicollinearity is important, and the variance inflation factor (VIF) is used for this diagnosis. The VIF is a tool that measures and quantifies how inflated the variance is. In general, when the VIF is 10 or more, there is a high correlation between independent variables [23]. In this study, the VIF was confirmed for four cases obtained through data preprocessing and variable selection processes. Table 4 shows the VIF for variables judged to have a high correlation with independent variables for each case.

Table 4. Multicollinearity results.

	Listwise Deletion		Mean Imputation	
	Variable	**VIF**	**Variable**	**VIF**
Correlation analysis	actual distance	45.118	actual distance	44.007
	vehicle tonnage	12.671	vehicle tonnage	11.423
			type of unloading	10.284
Stepwise method	loading longitude	19,820.792	loading longitude	20,049.129
	unloading month	4308.540	unloading month	4526.561
	arranging fee	3071.275	loading latitude	1237.245
	unloading latitude	1277.074	share state	689.888
	share state	677.228	unloading longitude	324.679
	unloading longitude	198.595	primary key	124.909
	primary key	98.354	unloading latitude	93.472
	actual distance	47.556	actual distance	46.726
	type of unloading	28.909	type of unloading	29.301
	loading location	26.450	loading location	26.017
	Company code	24.813	Company code	21.881
	loading latitude	22.948	loading day	20.398
	loading day	21.098	loading month	20.154
	loading month	20.884	unloading location	17.221
	registrant key	19.298	registrant key	16.748
	unloading location	17.219	type of loading	12.950
	vehicle tonnage	14.252	vehicle tonnage	12.703
	type of loading	13.042	unloading day of the week	10.339
	unloading day of the week	10.527		

As a result of confirming multicollinearity, as shown in Table 4, variables with a high multicollinearity were removed from each dataset. Two factors were removed from the dataset to which correlation analysis and listwise deletion were applied. In addition, three factors were removed from the dataset using correlation analysis and mean imputation. Both datasets were analyzed using five factors. The dataset with the stepwise method and listwise deletion was applied and analyzed with 16 factors after removing 19 factors. A total of 18 factors were removed from the dataset using the stepwise method and mean imputation. After that, 15 factors were used for analysis.

The process of learning and predicting was repeated 30 times to obtain enough samples so that the predicted value could approximately follow a normal distribution. For the learning step, 30 training sets obtained through K-fold cross-validation were used, and the prediction was carried out on one test set. The R-squared value of the resulting MLR model is shown in Table 5. The explanatory power of the model was calculated as the average of the R-squared for each predicted value. The fare prediction results for the MLR model show that the average explanatory power of the model was approximately 63.4%, and the R-squared value of the model obtained by processing the missing data by listwise deletion and processing the factors using the stepwise method was the highest.

Table 5. Multiple linear regression model results (R-squared).

	Listwise Deletion	**Mean Imputation**
Correlation analysis	0.617	0.607
Stepwise method	0.668	0.644

Table 6 shows the five values with the smallest error between the actual value and the predicted value among the results obtained for predicting the range of the cost using the multilinear regression model. The fare range was calculated using the confidence interval of the predicted value obtained by the model. The predicted fare range was estimated by

judging the maximum and minimum values in the confidence interval as the upper and lower limits of the prediction interval.

Table 6. Prediction results—multiple linear regression.

	Listwise Deletion						Mean Imputation					
	Correlation Analysis			Stepwise Method			Correlation Analysis			Stepwise Method		
	Fee	Min	Max	Fee	Min	Max	Fee	Min	Max	Fee	Min	Max
1	290,000	289,991	290,008	230,000	229,988	230,011	260,000	259,989	260,010	240,000	239,984	240,015
2	300,000	299,993	300,005	220,000	219,987	220,012	200,000	199,995	200,005	160,000	159,985	160,014
3	360,000	359,981	360,017	270,000	269,991	270,009	350,000	349,987	350,010	160,000	159,988	160,009
4	120,000	119,993	120,005	110,000	109,989	110,011	250,000	249,993	250,004	400,000	399,989	400,012
5	180,000	179,992	180,009	230,000	229,984	230,014	230,000	229,995	230,007	180,000	179,899	180,103

5.3. DNN

In this study, a DNN model with five hidden layers was constructed. The number of hidden layers and neurons was empirically determined after testing various combinations. The parameters of the DNN model are shown in Table 7.

Table 7. DNN model component.

Model Parameters	Values
Hidden Layers (Number of Nodes)	5 Layers (256→128→64→32→16)
Optimizer	Adam
Epochs	500
Batch size	256
Learning rate	0.001

The DNN model repeated the same learning and prediction process 30 times. The R-squared value of the DNN model after the learning process is shown in Table 8. The fare prediction by the DNN algorithm showed that the average explanatory power of the model was about 73.2%. When the variables obtained through the stepwise method were applied, it was found that the difference in the predictive power was large, depending on the preprocessing method. In addition, it was confirmed that the R-squared value of the model was the highest when the mean imputation method and the stepwise method were used. Table 9 shows the five values with the smallest error between the actual value and the predicted value among the results for the fare range prediction using the DNN model.

Table 8. DNN model results (R-squared).

	Listwise Deletion	Mean Imputation
Correlation analysis	0.718	0.738
Stepwise method	0.628	0.843

Table 9. Prediction results—DNN model.

	Listwise Deletion						Mean Imputation					
	Correlation Analysis			Stepwise Method			Correlation Analysis			Stepwise Method		
	Fee	Min	Max	Fee	Min	Max	Fee	Min	Max	Fee	Min	Max
1	340,000	338,019	341,980	240,000	238,856	241,143	350,000	348,456	351,543	430,000	428,056	431,943
2	420,000	418,148	421,852	290,000	288,245	291,754	170,000	168,614	171,386	270,000	267,670	272,329
3	180,000	178,672	181,327	290,000	288,245	291,754	190,000	189,010	190,988	300,000	297,243	302,756
4	180,000	178,518	181,480	290,000	288,245	291,754	360,000	358,718	361,280	240,000	237,869	242,131
5	170,000	168,955	171,046	140,000	138,700	141,299	270,000	263,806	276,191	250,000	248,674	251,326

5.4. XGBoost Regression

The XGBoost model was built using the basic form of the XGBoost algorithm, which consists of 400 weak learners and a maximum tree depth of three levels. The learning rate was set to be a default value of 0.3. The results of the XGBoost model are shown in Table 10. The results for predicting the shipping cost using the XGBoost model show that the model has an average explanatory power of about 74.6%. A significant difference in the explanatory power according to the variable selection method was found. Table 11 shows the five values with the smallest error between the actual value and the predicted value among the results for the predicted cost range using the XGBoost model.

Table 10. XGBoost regression model results (R-squared).

	Listwise Deletion	Mean Imputation
Correlation analysis	0.710	0.695
Stepwise method	0.802	0.776

Table 11. Prediction results—XGBoost regression model.

	Listwise Deletion						Mean Imputation					
	Correlation Analysis			Stepwise Method			Correlation Analysis			Stepwise Method		
	Fee	Min	Max	Fee	Min	Max	Fee	Min	Max	Fee	Min	Max
1	490,000	489,071	490,929	420,000	419,503	420,497	380,000	379,596	380,403	260,000	258,756	261,243
2	300,000	299,530	300,469	160,000	159,196	160,804	260,000	257,363	262,637	220,000	219,471	220,528
3	200,000	199,767	200,231	240,000	239,558	240,442	140,000	139,419	140,581	230,000	229,502	230,496
4	230,000	229,372	230,625	240,000	239,109	240,890	170,000	169,501	170,497	170,000	169,370	170,627
5	270,000	269,040	270,958	270,000	269,249	270,749	220,000	219,700	220,301	290,000	289,053	290,947

5.5. LightGBM

Table 12 shows the analysis results of the model built using the basic form of the LightGBM algorithm. The learning rate was set to be a default value of 0.1. The results for predicting the shipping cost using the LightGBM model show that the model has an average explanatory power of about 78.6%. The LightGBM model also shows a difference in the explanatory power depending on the variable selection method, but the explanatory power of all the models was 0.7 or greater. Table 13 shows the five values with the smallest error between the actual value and the predicted value among the results for predicting the cost range using the LightGBM model.

Table 12. LightGBM model results (R-squared).

	Listwise Deletion	Mean Imputation
Correlation analysis	0.734	0.727
Stepwise method	0.851	0.831

Table 13. Prediction results—LightGBM model.

	Listwise Deletion						Mean Imputation					
	Correlation Analysis			Stepwise Method			Correlation Analysis			Stepwise Method		
	Fee	Min	Max	Fee	Min	Max	Fee	Min	Max	Fee	Min	Max
1	110,000	109,598	110,401	270,000	269,160	270,839	210,000	208,954	211,043	110,000	109,166	110,833
2	380,000	379,248	380,751	170,000	169,384	170,616	230,000	228,996	231,001	210,000	209,462	210,537
3	200,000	199,433	200,567	300,000	298,726	301,273	180,000	179,621	180,380	180,000	178,262	181,738
4	150,000	149,636	150,362	350,000	348,639	351,360	270,000	269,154	270,847	340,000	339,243	340,756
5	230,000	229,362	230,638	120,000	119,126	120,872	270,000	269,154	270,847	180,000	178,818	181,180

5.6. Model Comparison

Table 14 shows a comparison of the explanatory power of all the analysis methods. When the predictive power was compared based on the preprocessing method, there appears to be no significant difference between the models, except for the DNN model. In addition, the model that predicted the shipping cost using the variables selected through the step selection method has a higher explanatory power than the model to which the correlation analysis was applied. It was confirmed that there was a big difference in the case of the boosting model. Compared to the other models, the learning time was short, and the predictive power was high. The reason why the model to which the stepwise selection method was applied has a higher explanatory power is thought to be because relatively more factors are considered by the model. In addition, the variables obtained through the stepwise method include all the variables obtained through the correlational analysis.

Table 14. Cost prediction model performance comparison.

	Correlation Analysis				**Stepwise Method**			
	Listwise Deletion		**Mean Imputation**		**Listwise Deletion**		**Mean Imputation**	
	R-Squared	**Learning Time(s)**	**R-Squared**	**Learning Time(s)**	**R-Squared**	**Learning Time(s)**	**R-Squared**	**Learning Time(s)**
MLR	0.617	21.98	0.607	7.22	0.668	30.60	0.644	23.39
DNN	0.718	97,500	**0.738**	93,780	0.628	115,500	**0.843**	219,780
XGBoost	0.710	492.41	0.695	492.22	0.802	1775.50	0.776	1393.70
LightGBM	**0.734**	67.60	0.727	66.49	**0.851**	176.90	0.831	143.94

The results show that the boosting model has an excellent predictive power. The LightGBM model has the best predictive power, followed by the XGBoost and DNN, and the MLR model has the lowest predictive power. Machine learning has the characteristic of iteratively learning and improving the model to increase the probability of success in the prediction. The traditional analysis model has the characteristic of making predictions with a fixed model through a single analysis. This is thought to be due to the differences between the machine learning and traditional analysis models.

The time required for model learning is also an important factor to consider for field applications. If a model takes a long time to learn, even if it shows a high accuracy, it may be difficult to apply to a field where rapid decision-making is required. The learning time of the correlation analysis method with few variables to consider was short, and the model with the shortest required time was the MLR model. The second shortest required time model was the LightGBM. The DNN model had a high predictive power, but it took a long time to learn compared with the other models, so it was judged to be difficult to apply in the field.

Considering the above results, it is evident that the machine learning models show a higher predictive power than the traditional analysis method. Considering both the speed and performance among the machine learning models, the LightGBM model was judged to be the most suitable for predicting the shipping cost. If the model performance is optimized in the future, the accuracy of the model could be further increased.

5.7. Variable Importance

Another purpose of this study was to derive factors that affect the cost setting. In order to derive these factors, the variable with the greatest contribution to the prediction of the shipping cost was identified using Shapley Additive exPlanations (SHAP). SHAP comprehensively calculates the contribution of each variable by comparing all combinations of variables in the model. In this study, the SHAP value of the LightGBM model with the highest predictive power was measured, and the analysis results are shown in Figure 3.

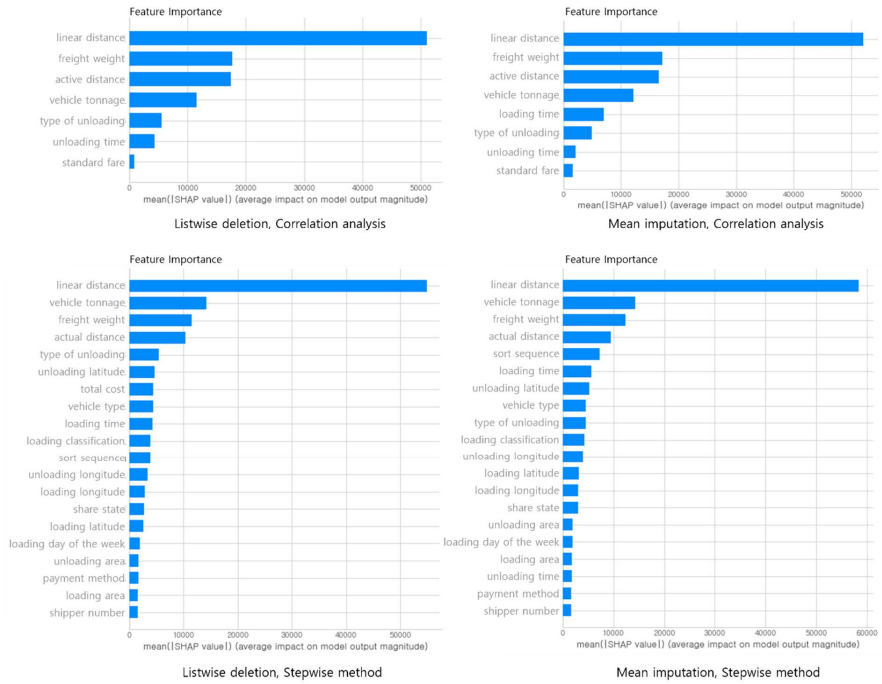


Figure 3. Variable importance for cost prediction.

The factors that make a high contribution to the shipping cost prediction are "linear distance," "actual distance," "freight weight," and "vehicle tonnage," which are highly related to transportation distance and freight characteristics. In particular, "linear distance" showed a high contribution of greater than 50%. It was judged that "linear distance" has a greater influence than "actual distance" in determining the shipping cost.

6. Conclusions and Future Research

To solve the problem of fare setting on a freight transportation brokerage platform, where there is no standardized shipping cost, the main factors that affect the shipping cost setting were derived in this study, and a price prediction model was built using machine learning. Factors that affect the shipping cost were selected using correlational analysis and the stepwise method from a total of 73 factors, including factors that were obtained from the freight brokerage process and environmental factors such as precipitation. The selected factors were cargo characteristic factors, vehicle owner characteristic factors, and environmental factors. Using these factors, a shipping cost prediction model was built, and the performance of each model was compared. Cargo characteristic factors included "freight weight," "loading/unloading time," and "loading/unloading location." "Vehicle tonnage" and "vehicle type" were included as characteristic factors of the owner. Precipitation was an environmental factor. The results of the analysis showed that the DNN, XGBoost, and LightGBM models, which are machine learning models, performed better than the linear regression, which is a traditional analysis method. The model that showed the best predictive power among the models used was the LightGBM model.

In addition, this study explored factors that affect the cost setting. Factor exploration was conducted using the LightGBM model, which had the highest predictive power. The factor that contributed the most to the cost setting was "linear distance," and "actual distance," "freight weight," and "vehicle tonnage" were also found to be major variables that influence the cost setting. No valid results were obtained for "precipitation," which was thought to affect the forecast of the shipping costs. This is believed to be due to the characteristics of the freight market. In the actual freight market, cargo transport volume decreases on rainy days. Since the data used in this study are the data of cargoes that have completed freight transportation, these market conditions do not appear. However, major variables obtained through research can be a quantitative indicator for determining the

shipping costs. This is expected to solve the problem of setting the shipping costs based on the shipper's experience. Because factors that were not previously considered could be considered in the future, the appropriateness of the shipping costs are expected to improve. In this study, daily precipitation data were added as an environmental factor, but a significant correlation between precipitation and the shipping costs could not be confirmed. A more accurate model could be presented if research is conducted by additionally considering other factors that reflect the actual situation.

Machine learning has been in the spotlight because it has shown excellent performance in forecasting for many fields, but there have been no case studies in the field of shipping cost prediction. In this study, a model with a high predictive power was presented by introducing a machine learning algorithm for fare prediction. This model has a higher accuracy than currently existing freight rates because it considers more cargo characteristics. Therefore, this model could be used in future cost prediction research and for setting the standard shipping costs on freight transport brokerage platforms. However, there are factors that cannot be confirmed with data, since the actual shipping costs are determined by the know-how of the shipper. Because of this, there is a limit to accurately predicting the transportation costs. However, if more factors are considered through future research and the model is advanced, a more accurate model can be built.

In future studies, we will generate and analyze meaningful data by changing the preprocessing method of environmental factors. In addition, other factors such as "highway fee" and "fuel cost" will be added to determine which environmental factors affect the shipping costs. In addition, it is necessary to optimize the model performance in future studies to increase its accuracy. Freight is divided into a range of fares according to various characteristics. A more accurate model could be built if the shipping cost is predicted after the data are filtered with consideration for these data characteristics. Exploring more advanced optimization algorithms or metaheuristics for this decision problem could be provided. In the future research, the proposed approach could be compared to the advanced optimization or metaheuristic algorithms.

Author Contributions: Conceptualization, H.-S.J.; Funding acquisition T.-W.C.; Investigation, H.-S.J.; Project administration, H.-S.J. and T.-W.C.; Validation, T.-W.C. and S.-H.K.; Writing—original draft, H.-S.J. and T.-W.C.; Writing—review and editing, S.-H.K. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by a Korea Institute for Advancement of Technology (KIAT) grant funded by the Korea Government (MOTIE) (P0008691, HRD Program for Industrial Innovation) and the GRRC program of Gyeonggi province [(GRRC KGU 2020-B01), Research on Intelligent Industrial Data Analytics].

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable. Due to trade secret concerns, the raw data are kept confidential, and you may request some data from the authors or Hwamulman Co. Ltd., Gwangju, Republic of Korea.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Lee, T.-H.; Heo, J.-S. *2021 Logistics Industry Outlook*; Issue Paper 2021-02; The Korea Transport Institute: Sejong, Korea, 2021.
2. Ko, Y.-S. Post COVID-19, the Change of Road Transport System and Logistics. *Transp. Technol. Policy* **2021**, *18*, 12–16.
3. Do, K.-H. *A Study on the Problems and Improvement of the THC, CAF & BAF in Korea Ocean Freight*; Konkuk University Graduate School: Seoul, Republic of Korea, 2009.
4. Kovács, G. First cost calculation methods for road freight transport activity. *Transp. Telecommun. J.* **2017**, *18*, 107–117. [CrossRef]
5. Sternad, M. Cost Calculation in road freight transport. *Bus. Logist. Mod. Manag.* **2019**, *19*, 215–225.
6. Li, J.; Zheng, Y.; Dai, B.; Yu, J. Implications of matching and pricing strategies for multiple-delivery-points service in a freight O2O platform. *Transp. Res. Part E Logist. Transp. Rev.* **2020**, *136*, 101871. [CrossRef]

7. Lindsey, C.; Frei, A.; Babai, H.; Mahmassani, H.; Park, Y.; Klabjan, D.; Reed, M.; Langheim, G.; Keating, T. Modeling carrier truckload shipping costs in spot markets. In Proceedings of the 24 Annual Meeting of the Transportation Research Board, Washington, DC, USA, 13–17 January 2013.
8. Jo, S.-H.; Kang, M.-G.; Kim, G.-E.; Ban, J.-H.; Lee, J.-H.; Kang, T.-W. Predicting changes in housing prices nationwide through machine learning. In Proceedings of the KIIT Conference, Bhubaneswar, India, 17–18 December 2021.
9. Jang, D.-R.; Park, M.-J. Price Determinant Factors of Artworks and Prediction Model Based on Machine Learning. *J. Korean Soc. Qual. Manag.* **2019**, *47*, 687–700.
10. Zhao, H.; Zhang, C. An online-learning-based evolutionary many-objective algorithm. *Inf. Sci.* **2020**, *509*, 1–21. [CrossRef]
11. Dulebenets, M.A. An adaptive polyploid memetic algorithm for scheduling trucks at a cross-docking terminal. *Inf. Sci.* **2021**, *565*, 390–421. [CrossRef]
12. Pasha, J.; Nwodu, A.L.; Fathollahi-Fard, A.M.; Tian, G.; Li, Z.; Wang, H.; Dulebenets, M.A. Exact and metaheuristic algorithms for the vehicle routing problem with a factory-in-a-box in multi-objective settings. *Adv. Eng. Inform.* **2022**, *52*, 101623. [CrossRef]
13. Rabbani, M.; Oladzad-Abbasabady, N.; Akbarian-Saravi, N. Ambulance routing in disaster response considering variable patient condition: NSGA-II and MOPSO algorithms. *J. Ind. Manag. Optim.* **2022**, *18*, 1035–1062. [CrossRef]
14. Lee, Y.-S.; Moon, P.-J. A Comparison and Analysis of Deep Learning Framework. *J. Korea Inst. Electron. Commun. Sci.* **2017**, *12*, 115–122.
15. Bae, S.-W.; Yu, J.-S. Predicting the Real Estate Price Index Using Machine Learning Methods and Time Series Analysis Model. *Hous. Stud. Rev.* **2018**, *26*, 107–133. [CrossRef]
16. Wilks, D.S. *Statistical Methods in the Atmospheric Sciences*; Academic Press: Cambridge, MA, USA, 2011; Volume 100.
17. Schmidhuber, J. Deep learning in neural networks: An overview. *Neural Netw.* **2015**, *61*, 85–117. [CrossRef] [PubMed]
18. Chen, T.; Guestrin, C. Xgboost: A scalable tree boosting system. In Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016.
19. Natekin, A.; Knoll, A. Gradient boosting machines, a tutorial. *Front. Neurorobotics* **2013**, *7*, 21. [CrossRef] [PubMed]
20. Ke, G.; Meng, Q.; Finley, T.; Wang, T.; Chen, W.; Ma, W.; Ye, Q.; Liu, T.-Y. Lightgbm: A highly efficient gradient boosting decision tree. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 3815895.
21. Hall, M.A. Correlation-Based Feature Selection for Machine Learning. Ph.D. Dissertation, The University of Waikato, Hamilton, New Zealand, 1999.
22. Wang, D.; Zhang, W.; Bakhai, A. Comparison of Bayesian model averaging and stepwise methods for model selection in logistic regression. *Stat. Med.* **2004**, *23*, 3451–3467. [CrossRef] [PubMed]
23. Daoud, J.I. Multicollinearity and regression analysis. *J. Phys. Conf. Ser.* **2017**, *949*, 012009. [CrossRef]