Prediction of Hourly Subway Ridership based on Artificial Intelligence Algorithms Using Weather Information

San Gwon
Departmet of Computer Science
Andong National Univ.
Andong, Gyeongsang, South Korea
jmt30269@gmail.com

Hyeon-Jeong Kim
Departmet of Computer Science and Engineering
KonKuk University
Seoul, South Korea
b.bhj0817@gmail.com

Ye-Jin Lee
Departmet of Bio-medical engineering
Keimyung Univ.
Daegu, South Korea
ofxxcrax@gmail.com

Jo-Eun Kim
Departmet of Computer Science
Andong National Univ.
Andong, Gyeongsang, South Korea
joenkim200212@gmail.com

Seokheon Cho
Qualcomm Institute
University of California, San Diego
La Jolla, USA
justinshcho@gmail.com

***Abstract*— This paper presents a model for predicting hourly subway ridership based on weather conditions. The goal is to provide the efficient service for Seoul's main transportation system. To achieve this, we applied several artificial intelligence algorithms defined by KNIME, a workflow-based data analysis tool, to compare and analyze the performance of our prediction model about hourly subway ridership. Three artificial intelligence algorithms, namely Multiple Linear Regression (MLR), Random Forest Regression (RFR), and Multi-Layer Perceptron (MLP), were implemented for prediction. Rather than merely developing a prediction model for a single subway station, we aimed to optimize the performance of the prediction model across two subway stations with diverse regional characteristics. For each subway station, we selected the different independent variables, according to the correlation with the dependent variable and feature importance of the RF to compare the prediction results. The outcomes indicated that the subway station that is heavily used by office workers is almost unaffected by weather variables, while the subway station that is heavily used by people who enjoy leisure activities is affected by weather variables. Across all datasets, the model using MLP algorithm outperformed the alternative models in predicting subway ridership.**

Keywords*— *artificial intelligence, machine learning, time-series data, and hourly subway ridership prediction

I. Introduction

The subway is an essential mean of public transport in today's society. As reported by the Ministry of Land, Infrastructure, and Transport, in Seoul, the percentage of subway ridership (67.9%) is more than twice as high as that of other means of transport (32.1%) [1]. Ensuring precise ridership data is imperative for the effective scheduling of trains according to demand. There are numerous factors that impact public transport usage. Among them, the weather and surrounding environment characteristics of the station extensively influence people's behavioral patterns and accordingly impact subway ridership. Therefore, this study aims to develop an accurate prediction model for subway ridership by analyzing the correlation between various weather data, subway ridership, and the characteristics of the area around subway stations.

Hee-Jin Kim et al. used Gaussian Process, Bagging, and Random Forest algorithms to predict monthly public transportation ridership in Seoul using weather data, including temperature, precipitation, and fine dust. Among the three algorithms, Gaussian Process showed the lowest error rate with an average of 15.49% in the test set [2]. Sang-Gi Choi et al. also used daily weather data and public transportation ridership in Seoul, Korea to investigate the impact of weather conditions on public transportation demand using Least Square Simple Regression [3]. The weather data included precipitation, unpleasantness index, snowfall, and temperature. The results of the Least Square Simple Regression model analysis using precipitation and temperature were 0.77 and 0.75, respectively. Chuan Ding et al. proposed the use of a Gradient Boosting Decision Tree (GBDT) model to improve the prediction performance of short-term subway ridership [4]. The dataset used is Beijing subway ridership data. Also SVM, BP-neural Network, and Random Forest are used as comparison models. Three stations in Beijing were selected for prediction, and the average R^2 values of the models were between 0.9636 and 0.9871. Jingyi Liu used a Linear Regression model to predict Hangzhou subway ridership by time of day. After six rounds of training, the model with the highest R^2 value of 0.9313 was selected to predict the ridership of stations on three lines [5].

Most studies have predicted subway ridership at a single station using weather data. However, there are no studies that compare the performance of subway ridership prediction models for multiple subway stations with different characteristics of the surrounding environment and what should be considered for performance improvement. In addition, previous studies used monthly and daily data to predict monthly and daily ridership, so the prediction performance was generally high. However, previous studies have not performed hourly ridership predicts, which are necessary to suggest optimal subway dispatch times according to subway ridership. Therefore, this study aims to derive an optimal model for predicting subway ridership at a subway station by using hourly weather data and subway ridership data and considering the environmental characteristics around the subway station.

This paper is organized as follows. Section II introduces the initial dataset and describes the dataset and preprocessing techniques considered in this analysis. Section III presents the algorithms and performance metrics used in this study. Section IV compares the performance of the algorithms discussed in Section III based on the performance metrics. Section V concludes with the results of this study and future directions.

II. Data preprocessing and dataset construction

A. *Original Data Set*

In this paper, we process and use data provided by the Open Government Data Portal [6] and the Weather Data Open Portal [7]. The subway data contains 'information on the number of people getting on and off the train by station and time of day'. Both weather data and subway data are daily hourly data for four years, from January 1, 2016 to December 31, 2019. The reason why we did not consider data after 2020 is that the global pandemic caused the loss of the unique characteristics of each subway station's surroundings, which makes predicting ridership by subway station largely meaningless. In addition, we used pre-pandemic data because it is relatively stable and consistent, and mixing it with pro-2020 data would worsen data training and testing results. Table 1 summarizes the data required for this study from two datasets: the Open Government Data Portal and the Open Weather Data Portal. Among the various weather data, we selected four factors that affect subway ridership: 'temperature', 'rainfall', 'wind speed', and 'humidity'. Most of the data listed in Table 1 are hourly measurements. 'Boarding Flag' is an identifier that distinguishes between boarding and exiting. The dependent variable, 'Ridership', is the number of passengers who used the subway. Specifically, if a 'Boarding Flag' represents a boarding, then 'Ridership' represents the number of passengers who got on at that station. While if the 'Boarding Flag' indicates getting off, then 'Ridership' indicates the number of passengers who got off at that station.

B. *Data Preprocessing and Creation of Dataset #1*

Given that subway ridership patterns differ between weekdays and weekends, this study exclusively utilized weekday data, excluding weekends and holidays, to enhance the performance of the prediction model. Furthermore, the time range provided by subway data is from 6 to 23 o'clock, so we removed the weather data except for that time. The sensible temperature will be utilized due to our belief that it has a greater effect on subway ridership than the 'Temperature' in Table 1.

TABLE I. ORIGINAL DATASET

	Name	**Type**
Independent Variables	Date & Time	Year-Month-Date Hour:Minute
	Temperature [°C]	Double
	Rainfall [mm]	Double
	Wind speed [m/s]	Double
	Humidity [%]	Int
	Snowfall [cm]	Double
	Station name	String
	Boarding Flag	String [Boarding, Exiting]
Dependent Variable	Ridership	Int

The formula for the sensible temperature can be determined from the temperature data and is categorized by summer (May to September) and winter (October to April) [8]. The formula for the summer temperature, $t_s^S(\mathrm{kT})$, at time kT is as follows:

$$t_s^S(\mathrm{kT}) = -0.2442 + 0.55399 * t_w(\mathrm{kT}) \; +0.45535 * \mathrm{t(kT)} - 0.0022 * \; t_w(\mathrm{kT})^2 + 0.00278 * \; t_w(\mathrm{kT}) * \mathrm{t(kT)} \; + 3.0, \quad (1)$$

where $t_w(\mathrm{kT})$ and t(kT) are the wet bulb temperature and air temperature at time kT, respectively. The wet bulb temperature is the temperature measured by wrapping the tip of a mercury bar with a cotton ball soaked in water, and the wet bulb temperature at time kT can be defined as follows:

$$t_w(\mathrm{kT}) = \; \mathrm{t(kT)} * \tan^{-1}\left(0.151977 * \sqrt{\mathrm{r(kT)} + 8.313658}\right) + \tan^{-1}\left(\mathrm{t(kT)} + \mathrm{r(kT)}\right) \; - \; \tan^{-1}(\mathrm{r(kT)} - 1.67633) + 0.00391838 * \mathrm{r(kT)}^{\frac{2}{3}} * \tan^{-1}\left(0.023101 * \; \mathrm{r(kT)}\right) - 4.686035, \quad (2)$$

where r(kT) is the relative humidity (%) at time kT. Unlike summer, winter uses the 10-minute average wind speed (km/h) instead of the wet bulb temperature. The formula for the winter sensible temperature, $t_s^W(\mathrm{kT})$, at time kT is as follows:

$$t_s^W(\mathrm{kT}) = 13.12 + 0.6215 * \; \mathrm{t(kT)} \; - \; 11.37 * \mathrm{v(kT)}^{0.16} + 0.3965 * \; \mathrm{v(kT)}^{0.16} * \; \mathrm{t(kT)}, \quad (3)$$

where v(kT) is the 10-minute average wind speed in km/h. v(kT) is a processed value of the wind speed in the dataset. The wind speed of the data set is the 10-minute average wind speed calculated from the average of 10 one-minute average wind speeds from 10 minutes before the time to the hour. Since the original data is presented in m/s, we converted it to km/h by multiplying by 3.6. Subways are less affected by rainfall and snowfall compared to other modes of transportation because passengers board and exit at stations. Therefore, we decided that even 0.1 mm of rainfall or snowfall at a specific time would be deemed significant, regardless of the quantity. For Table 1, we created a new Rain Flag variable of String kind. It was set to 1 if there was any rainfall or snowfall, or 0 if the amount of precipitation and snowfall was lesser than 0.1 mm.

The dependent variable, ‘Ridership’, is a time series data. To improve the performance of the prediction results for this dependent variable, we added data from one hour ago ('_1hr') and one week ago ('_1wk') as independent variables for some of the independent variables listed in Table 1 except ‘Hour’. Since we are only using data from non-holiday days, if the data from a week ago was a holiday, we used the data from two weeks ago. And if it was also a holiday two weeks ago, we deleted all data from that day.

Table 2 shows ‘Dataset #1’, which we will consider for subway ridership prediction after this preprocessing. It consists of 15 independent variables and one dependent variable. We selected two subway stations in Seoul with different surroundings. 'E station' is characterized by a large concentration of companies in the area, resulting in a spike in ridership during rush hour. On the other hand, 'T station' is connected to a residential area and Han River Park, so its ridership is almost constant throughout most of the weekdays. It is also characterized by the fact that weather affects ridership due to park users. This paper analyzes the number of people getting on and off the subway from 6:00 to 23:00 at the two stations mentioned above.

C. Creation of Dataset #2 for Performance Improvement

To determine the significant independent variables in predicting subway ridership at two stations among multiple independent variables in 'Dataset #1', We want to use Linear Correlation and Random Forest Regression 's Feature Importance between the dependent variable, 'Ridership' of boarding and exiting per station, and the rest of the independent variables.

Table 3 shows the absolute value of each correlation in order to prioritize the correlations. The eight independent variables with the highest absolute value of correlation in the boarding and exiting data at each station are shaded green. Of the independent variables shaded in green, we can see that five variables are the same across stations and boarding and exiting, while one to three variables vary across stations and boarding and exiting. For both stations, the current time ('Hour'), wind speed one hour ago ('Wind speed_1hr'), humidity one hour ago ('Humidity_1hr'), ridership one hour ago ('Ridership_1hr'), and ridership one week ago ('Ridership_1week') are highly correlated with the dependent variable, regardless of boarding and exiting.

TABLE II. DATA COMPOSITION OF DATASET #1

	Variable	Unit
Independent Variables	Hour	-
	Rain Flag, Rain Flag_1hr, Rain Flag_1wk	{0,1}
	Humidity, Humidity_1hr, Humidity_1wk	[%]
	Sensible Temp, Sensible temp_1hr, Sensible temp_1wk	[℃]
	Wind speed, Wind speed_1hr, Wind speed_1wk	[m/s]
	Ridership_1hr, Ridership_1wk	-
Dependent Variable	Ridership	-

TABLE III. LINEAR CORRELATION BETWEEN DEPENDENT VARIABLE AND INDEPENDENT VARIABLES

Independent Variables	**Stations**			
	E Station		***T Station***	
	Boarding	***Exiting***	***Boarding***	***Exiting***
Hour	0.612	0.58	0.459	0.487
Rain Flag	0.028	0.011	0.040	0.073
Wind speed	0.183	0.119	0.141	0.153
Humidity	0.141	0.122	0.156	0.114
Sensible temp	0.048	0.079	0.056	0.179
Rain Flag_1hr	0.033	0.014	0.041	0.074
Wind speed_1hr	0.269	0.169	0.160	0.228
Humidity_1hr	0.243	0.198	0.192	0.199
Sensible temp_1hr	0.086	0.099	0.052	0.214
Ridership_1hr	0.644	0.51	0.641	0.332
Rain Flag_1wk	0.030	0.016	0.012	0.043
Wind speed_1wk	0.193	0.13	0.138	0.181
Humidity_1wk	0.154	0.142	0.224	0.077
Sensible temp_1wk	0.053	0.083	0.030	0.154
Ridership_1wk	0.975	0.993	0.889	0.849

In particular, 'Hour', 'Ridership_1hr', and 'Ridership_1wk' are highly correlated, with averages of 0.5345, 0.53175, and 0.9265, respectively, across boarding and exiting at the two stations. The high correlation is due to the characteristic of the subway, where there are fixed users. In particular, the correlation value for 'Ridership_1wk' is the highest because the dependent variable, subway ridership at a given time, is similar to subway ridership at the same time one week ago. The correlation values for rainfall ('Rain Flag'), rainfall one hour ago ('Rain Flag_1hr'), and rainfall one week ago ('Rain Flag_1wk') are very low because 97% of the total data is rain-free and there is a fixed number of subway users regardless of rain or snow. Comparing boarding and exiting at ‘E station’, 'Wind speed' and 'Humidity' are highly correlated among the 12 weather data in 'Dataset #2'. In the case of boarding and exiting at 'T Station', it is affected by 'Wind Speed' and 'Humidity' like the ‘E station’, but 'Sensible Temp' also has a high correlation. In particular, the correlation value between ‘Sensible temp’ and ‘Sensible temp_1hr’ is high at ‘T station’, which is near the Han River park, because the number of subway passengers using that station changes closely according to the sensible temperature. Table 4 shows the results of the feature importance obtained by applying the Random Forest Regression algorithm to the 15 independent variables included in Dataset #1.

The tree depth was set to 10 and the total number of tree models was set to 1,500. In particular, when applying the Random Forest Regression algorithm, the data from 2016 to 2018 and 2019 were separated from the total data for training and evaluation. Independent variables with high feature importance have a relatively greater impact on ridership prediction. The eight independent variables with high feature importance values in the boarding and exiting data at each station are shaded in green. Among the independent variables shaded in green, for 'E station', the considered independent variables for boarding and exiting are the same.

TABLE IV. FEATURE IMPORTANCE BETWEEN DEPENDENT VARIABLE AND INDEPENDENT VARIABLES

Independent Variables	**Stations**			
	E Station		***T Station***	
	Boarding	***Exiting***	***Boarding***	***Exiting***
Hour	252	215	256	283
Rain Flag	0	1	4	4
Wind speed	89	68	53	65
Humidity	62	51	114	22
Sensible temp	11	8	34	63
Rain Flag_1hr	5	1	1	3
Wind speed_1hr	177	150	93	190
Humidity_1hr	128	175	140	64
Sensible temp_1hr	35	36	21	150
Ridership_1hr	213	278	248	211
Rain Flag_1wk	3	2	0	0
Wind speed_1wk	124	103	68	119
Humidity_1wk	71	105	175	11
Sensible temp_1wk	17	11	7	24
Ridership_1wk	313	296	286	291

However, 'Wind speed', 'Humidity', 'Humidity_1hr', and 'Sensible temp_1hr' are the variables that differ between boarding and exiting at 'T station'. This shows that the weather conditions that subway users consider when getting on and off the train are different. While getting on and off at 'E station' and getting on at 'T station' have similar feature importance patterns overall, getting off at 'T station' currently seems to have relatively high feature importance values for 'Sensible temp' and 'Sensible temp_1hr'. Many of getting off at 'T station' are leisure activities, and it is possible to reach 'T station' within an hour from most stations in Seoul. Therefore, the number of people getting off at 'T station' is affected by the weather conditions when they leave their location, which is why 'Sensible temp_1hr' has a high feature importance value. Even when they arrive at the station, they are likely to move to another station again depending on the weather conditions, so 'Sensible temp' is currently the main variable that determines the number of exiting.

For these reasons, in order to compare the performance of the two stations' boarding and exiting prediction models, this study defines a new ‘Dataset #2’ that includes only 8 independent variables from the 15 independent variables in ‘Dataset #1’ based on Table 3 and Table 4. The eight independent variables included in ‘Dataset #2’ are shown in Table 5.

TABLE V. DATA COMPOSITION OF DATASET #2

	Variable	**Unit**
Independent Variables	Hour	-
	Humidity_1hr Humidity_1wk	[%]
	Wind speed, Wind speed_1hr, Wind speed_1wk	[m/s]
	Ridership_1hr, Ridership_1wk	-
Dependent Variable	Ridership	-

The selection criteria are 'Hour', 'Ridership_1hr', weekly ridership ('Ridership_1wk'), 'Wind speed_1hr', weekly wind speed ('Wind speed_1wk'), and 'Humidity_1hr', which have high correlation values and feature importance values in common. In addition, we selected 'Wind speed' and humidity one week ago ('Humidity_1wk') because they are included in most cases. However, we excluded 'Sensible temp_1hr', which was exceptionally high at 'T station'

III. AI ALGORITHMS AND PERFORMANCE METRICS

In this paper, we used three algorithms: Multiple Linear Regression, Random Forest Regression, and Multi-Layer Perceptron.

A. *Multiple Linear Regression (MLR)*

Multiple Linear Regression (MLR) represents the linear relationship between a single dependent variable and several independent variables. Unlike simple linear regression, MLR considers multiple independent variables to account for interactions among variables, providing an advantage. The predicted value, y, for MLR is determined as follows (4).

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \cdots + \beta_i x_i + \varepsilon, \qquad (4)$$

where ε and β_0 are the error and y-intercept, respectively, and β_i is the value representing the increment of the dependent variable with respect to the independent variable x_i.

B. *Random Forest Regression (RFR)*

Random Forest Regression (RFR) is an ensemble machine learning algorithm that combines multiple decision trees to perform accurate predictions for regression tasks. It has the advantage of providing stable prediction performance across a variety of data and is useful in complex datasets with nonlinear relationships. Additionally, RFR is less likely to overfit than other algorithms because it only uses a subset of features for training. In this study, we found the optimal tree depth and total number of tree models that avoid overfitting and exhibit the best performance through multiple training and testing. The optimal tree depth and total number of tree models are 10 and 1,500, respectively.

C. *Multi-Layer Perceptron (MLP)*

Multilayer Perceptron (MLP) is an artificial neural network composed of multiple layers of perceptron. MLP is composed of an input layer, a hidden layer, and an output layer, and the hidden layer can have one or more layers. Each hidden layer is composed of multiple neurons. The hidden layer applies weights to the input and transmits the output through an activation function. To improve the performance of MLP, the number of hidden layers and the number of neurons should be adjusted appropriately. In this study, MLP using RProp is implemented in the workflow-based data analysis tool called KNIME. RProp is an abbreviation for ‘resilient propagation,’ a learning algorithm created to overcome the limitations of gradient descent [9]. Unlike the conventional gradient descent, it calculates weight updates using only the sign of the derivative. This allows for faster and more efficient learning than before. However, it is difficult to implement due to the complexity of adjusting the learning rate, and the performance can vary greatly depending on the learning rate adjustment. In this study, we were able to find the following optimal settings through various training and testing to have the best model performance. The optimal number of model iterations, number of hidden layers, and number of neurons per layer are 500, 4, and 36, respectively. We conducted training and testing for the subway passenger number prediction model using these values.

D. Performance Evaluation Metrics

For each algorithm, four performance evaluation metrics are used: R^2 (R-squared), Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE). R^2 is a coefficient of determination that shows the explanatory power of independent variables on the dependent variable. The range of R^2 is from 0 to 1, and the closer the value is to 1, the better the fit. MAE is the average of the absolute values of the differences between the algorithm's predicted and actual values. It is the most intuitive metric to understand because it takes the absolute value. However, it has the disadvantage of being unable to tell whether the algorithm is overfitting or underfitting because it takes the absolute value. MSE is different from MAE in that it is the sum of the values squared instead of the absolute value. It has the advantage of being easier to detect outliers because it is squared. RMSE is the value of MSE with the root taken, and it has the advantage of being intuitive because the metric is the same as the predicted variable. Both MSE and RMSE are affected by the size of the predicted variable.

IV. SUBWAY RIDERSHIPS PREDICTION ANALYSIS AND RESULTS

We use three algorithms, MLR, RFR, and MLP, to predict subway ridership over time by station for 'Dataset #1' and 'Dataset #2' and compare the results. Table 6 shows the results of the boarding and exiting ridership prediction models at 'E station'. As shown in Table 6, both ‘Dataset #1’ and ‘Dataset #2’ of ‘E station’ show higher prediction results of RFR and MLP than MLR. For boarding, MLR outperforms the other algorithms by 0.012 to 0.017 in R^2 values.

For exiting, the difference between the three algorithms is insignificant, ranging from 0.001 to 0.005, suggesting that they perform equally well. The performance comparison between RFR and MLP is as follows: both algorithms show similar R^2 values. However, the RMSE values obtained by applying RFR are smaller than the RMSE values of MLP for 'Dataset #1' boarding and 'Dataset #2' boarding and exiting, so the RFR model performance is slightly better. In addition, the RMSE and R2 values for 'Dataset #2' are better than those for 'Dataset #1' for both RFR and MLP, which show good predictive model performance. This indicates that ‘Dataset #1’, which consists of the required weather data, has a positive impact on the subway ridership prediction performed with ‘Dataset #2’, which consists of only a few key independent variables selected through correlation and feature importance obtained through the RFR.

TABLE VI. PERFORMANCE RESULT S AT ‘E (EULJIRO 1 GA) STATION’

	E Station			E Station		
	MLR	***RFR***	***MLP***	***MLR***	***RFR***	***MLP***
	Dataset #1					
R^2	0.949	0.968	0.963	0.984	0.986	0.987
MAE	256.465	271.369	237.053	188.545	245.759	203.057
MSE	463428.24	288197.65	336083.547	193202.85	170837.37	150667.59
RMSE	680.756	536.840	579.727	439.548	413.325	388.159
	Dataset #2					
R^2	0.948	0.976	0.973	0.984	0.989	0.989
MAE	250.907	230.990	227.549	188.821	204.479	171.435
MSE	471987.41	211695.46	245445.490	193339.99	128374.14	133599.98
RMSE	687.013	460.104	495.425	439.704	358.293	365.513

However, the insignificant performance difference between the two datasets considered, ‘Dataset #1’ and ‘Dataset #2’, is due to the fact that ‘E station’ is a station that many office workers use regardless of the weather conditions. In other words, 'E station' is a station with a very strong fixed subway demand regardless of weather conditions, which is why 'Dataset #2' with less weather data performs better in predicting subway demand.

Table 7 shows the results for the boarding and exiting ridership prediction model for 'T station'. As shown in Table 7, when considering all four performance metric results across datasets and without distinguishing between boarding and exiting, MLP, RFR, and MLR perform best overall. Except for MLP for boarding, there is a performance improvement for other algorithmic-based models for subway ridership prediction using ‘Dataset #2’ compared to ‘Dataset #1’. Similar to the results observed in Table 7, this result indicates that the performance improvement of the subway ridership prediction model performed with ‘Dataset #2’, which consists of only a few key independent variables selected from ‘Dataset #1’ through correlation and feature importance obtained through the RFR. The large performance difference between the two datasets, as shown in Table 7, is due to the fact that T station's main ridership is more influenced by weather data than ‘E station’. In other words, 'T station' is a station that does not have a large demand for regular subway ridership regardless of weather conditions.

When comparing ‘E station’ and ‘T station’ in Tables 6 and 7, it becomes clear that ‘E station’ has a higher overall prediction performance. for subway ridership at different times of the day.

TABLE VII. PERFORMANCE RESULTS AT ‘T (TTUKSEOM PARK) STATION’

	T Station Boarding			T Station Exiting		
	MLR	***RFR***	***MLP***	***MLR***	***RFR***	***MLP***
	Dataset #1					
R^2	0.836	0.927	0.944	0.798	0.899	0.894
MAE	82.976	53.481	43.976	83.666	60.393	58.609
MSE	17360.718	7703.532	5909.195	21294.975	10677.527	10374.777
RMSE	131.760	87.770	76.871	145.928	103.332	101.857
	Dataset #2					
R^2	0.827	0.931	0.989	0.822	0.899	0.920
MAE	84.815	51.595	171.435	79.238	57.304	53.478
MSE	18224.725	7313.890	133599.980	17471.379	9888.176	7821.455
RMSE	134.999	85.521	365.513	132.179	99.439	88.439

This can be explained by the characteristics of the main users at each station. In the case of 'E station', the subway ridership prediction model performs well because the main users, office workers, are less affected by the weather and there is not much variation in the demand. In other words, the ridership at 'E station' varies in a regular pattern over time without significant deviations. By way of contrast, since the main users of 'T station' are people who want to spend leisure time in the park, the station ridership is more affected by the weather, so there is a large variation in the number of passengers using the subway at different times of the day, and the performance of the subway ridership prediction model is low. To put it another way, the deviation of subway ridership at 'T station' is rather large due to the continuous change in weather conditions across seasons and time. Due to this phenomenon, 'E station' has a higher subway ridership prediction model performance than 'T station'.

V. Conclusion

The purpose of this study is to propose a model to predict the hourly ridership of the subway, a representative public transportation system in Korea, and to evaluate its performance. Due to the characteristic of subway, we believe that weather has a strong influence on the ridership, so we used hourly weather data to predict hourly subway ridership. Since subway ridership is a time series data collected over a continuous period, we additionally used data from one hour ago and one week ago to improve the performance of the prediction model. We created Dataset #1 including 15 of these independent variables. Among them, we selected only 8 independent variables with high correlation values and high importance of the characteristics obtained when applying the Random Forest Regression algorithm. We created 'Dataset #2' consisting of these 8 independent variables. Compared to 'Dataset #1', 'Dataset #2' has mainly less weather data. The correlations and feature importance show that each station has different weather data to consider, and that the weather data to consider are related to the environmental characteristics around the station. The AI algorithmic models used to predict hourly subway ridership are Multiple Linear Regression (MLR), Random Forest Regression (RFR), and Multi-Layer Perceptron (MLP). In this paper, we consider two stations with different station surroundings characteristics. At 'E Station,' subway riders primarily consist of office workers, resulting in subway demand closely mirroring historical patterns regardless of weather conditions. Therefore, there is no significant improvement in the performance of the prediction model using 'Dataset #2' compared to 'Dataset #1' because the historical subway demand is a more important factor than the weather data in predicting subway ridership. On the other hand, in the case of 'T station', the main subway users are those who come to enjoy the nearby outdoor amusement park, which means that the number of customers at 'T station' varies greatly depending on the weather conditions. Therefore, there is a significant improvement in the performance of the prediction model when using 'Dataset #2' compared to 'Dataset #1' because the weather data as well as the historical subway demand is a major factor in predicting the number of subway users. Also, due to these characteristics of the two stations, the subway usage prediction results for 'E station' are higher than for 'T station'.

Since both stations have fixed ridership, MLR's results were not as good as the other two algorithms when using weather data. Among the three models, RFR performed the best for E station and MLP performed the best for T station.

In this study, we predicted the subway ridership at the current time. For more efficient subway operation, it is necessary to develop a model that accurately predicts subway ridership after a certain period, and to develop a management system that enables flexible distribution based on this. In addition, we will study a model that predicts usage by line, considering line characteristics and weather data, which is necessary for optimal distribution by subway.

Acknowledgment

This work (or result) was supported by the Ministry of Education and the National Research Foundation of Korea through the LINC 3.0 project, a third-phase industry-academic collaboration initiative, and the MSIT(Ministry of Science ICT), Korea, under the National Program for Excellence in SW, supervised by the IITP (Institute of Information & Communications Technology Planning & Evaluation) in 2023. (No. 2018-0-00213) We appreciate for mentoring from Dr. Changhee Han at Gyeongsang National University, South Korea.

References

[1] Digital Communication Team. "Public transportation ridership status at a glance 2020." Ministry of Land, Infrastructure and Transport of Kore a. https://www.molit.go.kr/USR/NEWS/m_35045/dtl.jsp?lcmspage=1&id=95085966 (accessed Jul. 16, 2023).

[2] Hee-Jin kim, Sujin OH and Ung-Mo Kim, "A Study on the Prediction o f Public Transportation Consumption in Seoul by Weather," in KIPS 20 17 Autumn Academic Presentation, Seoul, Korea, Nov. 2017, pp. 656-6 59.

[3] Sang Gi Choi, Jong Ho Rhee, and Seung Hwoon Oh, "The Effect of W eather Conditions on Transit Ridership," J. Korean Civil society, vol. 3 3, no.6, pp. 2447-2453, Nov. 2013, doi: http://dx.doi.org/10.12652/Ksc e.2013.33.6.2447.

[4] Chuan Ding, Donggen Wang, Xiaolei Ma and Haiying Li, "Predicting Short-Term Subway Ridership and Prioritizing Its Influential Factors U sing Gradient Boosting Decision Trees," in Sustainability, Volume 8, Is sue 11, Oct. 2016, [Online]. Available: https://doi.org/10.3390/su81111 00

[5] Jingyi Liu, "Analysis and Prediction of Subway Ridership -- Take a Sta tion in Hangzhou as an Example," in CMLAI 2023, 2023, pp. 315-320, doi: https://doi.org/10.54097/hset.v39i.5333

[6] Seoul Transportation Corporation, Seoul Transportation Corporation_I nformation on the number of people getting on and off by daily time zo ne by station, 2016 to 2019, Seoul Transportation Corporation, Seoul T ransportation Corporation_Information on the number of people getting on and off by daily time zone by station, May. 2023. [Online]. Availab le: https://www.data.go.kr/data/15048032/fileData.do#/layer_data_info mation

[7] Synoptic meteorological observation (ASOS), 2016 to 2019 Temperatu re, wind speed, rain by hour in Seoul, Jul. 2023. [Online]. Available: htt ps://data.kma.go.kr/data/grnd/selectAsosRltmList.do?pgmNo=36

[8] "Sensible Temperature." Korea Meteorological Administration Weathe r Data Service Open MET Data Portal. https://data.kma.go.kr/climate/w indChill/selectWindChillChart.do?pgmNo=111 (accessed Jul. 16, 2023).

[9] M. Riedmiller. and H. Braun, "A Direct Adaptive Method for Back pro pagation learning: The RPROP algorithm," in Proceedings of the IEEE International Conference on Neural Network, pp. 586-591, April. 1993