The Journal of the Convergence on Culture Technology (JCCT)
Vol. 8, No. 6, pp.861-866, November 30, 2022. pISSN 2384-0358, eISSN 2384-0366

http://dx.doi.org/10.17703/JCCT.2022.8.6.861

JCCT 2022-11-106

최적화된 Gradient-Boost를 사용한 서울 자전거 데이터의 결정 요인 예측

Predicting Determinants of Seoul-Bike Data Using Optimized Gradient-Boost

김차영*, 김윤**

Chayoung Kim*, Yoon Kim**

요 약 서울시에서는 공유 자전거 시스템, "따릉이"를 2015년부터 도입, 운영하여, 교통량 감축과 대기오염 해소를 위해 노력하고 있다. 하지만 공유 자전거 시스템, "따릉이"의 운영전략 미흡으로 인해 많은 문제가 발생하고 있어 이를 해결하고자 다양한 연구들이 제시되고 있다. 이들 연구의 대다수는 수요와 공급의 불균형을 해결하고자 하는 전략적 "자전거 배치"에 집중되어 있으며 또한 이들 중 다수가 날씨나 계절과 같은 특징을 그룹화함으로써 수요를 예측하고 있다. 그리고 이전에는 이들 예측방법이 주로 시계열 분석을 기반으로 하고 있었으나 최근에는 딥러닝/머신러닝으로 수요를 예측하는 연구들이 속속 등장하고 있다. 본 논문에서는 기존에 제시된 다양한 특징들을 기반으로 하면서, 새로운 특징을 발견하고 선택된 특징들의 중요도를 비교, 이를 순서화함으로써, 보다 정확한 수요 예측이 가능함을 보인다. 그리하여, 우리는 기존의 딥러닝/머신러닝 및 시계열 분석을 그대로 사용하면서 비교적 정확한 결정계수를 획득하고 이를 이용해 개선된 수요예측이 가능하도록 한다.

주요어 : 그라디언트 부스트, 특징 변수, 결정 계수, 공유 자전거 시스템, 배치 전략

Abstract Seoul introduced the shared bicycle system, "Seoul Public Bike" in 2015 to help reduce traffic volume and air pollution. Hence, to solve various problems according to the supply and demand of the shared bicycle system, "Seoul Public Bike," several studies are being conducted. Most of the research is a strategic "Bicycle Rearrangement" in regard to the imbalance between supply and demand. Moreover, most of these studies predict demand by grouping features such as weather or season. In previous studies, demand was predicted by time-series-analysis. However, recently, studies that predict demand using deep learning or machine learning are emerging. In this paper, we can show that demand prediction can be made a little better by discovering new features or ordering the importance of various features based on well-known feature-patterns. In this study, by ordering the selection of new features or the importance of the features, a better coefficient of determination can be obtained even if the well-known deep learning or machine learning or time-series-analysis is exploited as it is. Therefore, we could be a better one for demand prediction.

Key words : Gradient-Boost, Feature Pattern, Coefficient of Determinant, Seoul Public Bike, Bicyle Rearrangement

*정회원, 경기대학교 교양학부 조교수 (제1저자 및 교신저자)
**정회원, 국립한국복지대학교 컴퓨터정보보안과 정교수 (참여저자)
접수일: 2022년 10월 5일, 수정완료일: 2022년 10월 25일
게재확정일: 2022년 11월 1일

Received: October 5, 2022 / Revised: October 25, 2022
Accepted: November 1, 20202
*Corresponding Author: kimcha0@kyonggi.ac.kr

Ⅰ. 서 론

도시환경문제 개선을 위해 도시 교통시스템에 대한 무동력 교통수단을 활용하고자 서울시에서 공공 자전거 대여 시스템 “따릉이”를 약 2,000대를 시작으로 매년 공급하고 있다. 그림 1에 보이는 바와 같이 공유 자전거 시스템 “따릉이”는 무료 또는 유로로 단기간, 개인에게 자전거를 대여해주는 서비스로서 대도시의 교통 혼잡과 환경 오염 문제를 개선하기 위한 해결책 중 하나로 제시되었다. 최근에는 IT 기술의 발전으로 인해 무인시스템으로 누구나 손쉽게 자전거의 대여와 반납이 가능해지면서 급속도로 공유자전거 시스템을 도입한 도시도 서울 외에 증가하고 있다. 이러한 “따릉이”의 운영 효율성은 다양한 운영전략, 지역별 특성, 해당 지역 이용자의 이용특성에 따라 많은 영향을 받는다. 공유 자전거 시스템은 운영상의 많은 문제점이 발생되고 있는데, 그 중 가장 큰 문제가 시간대 및 지역별 자전거 배치의 불균형이다. 특정 시간대 혹은 전체적으로 한쪽 방향으로의 이동이 잦은 경우가 많아서 특정 대여소에 자전거가 부족하거나 과다하게 배치될 수 있다. 즉, 수요/공급의 불균형이다 [1]. 다양한 이용특성에 따라 대여소 별 자전거 과잉/부족현상은 항상 일어나고 있다. 하지만, 극복할 수 없는 현상으로 할 수 만은 없는 이유가 “부적절 자전거 재배치에 의한 자전거 이용 수요 저조”가 “따릉이”의 실패로 귀결될 것이기 때문이다. 이에 서울시는 각 대여소 별 자전거 재고 상태를 실시간으로 확인하고, 트럭을 이용해서 직접 자전거를 재배치하고 있다. 현재까지 이러한 자전거 배치의 불균형을 해소하기 위한 다수의 연구가 제시되어 왔으나 대부분이 기존 시스템의 변화를 반영하지 못하고 있다. 그럼에도 불구하고 서울시에서는 다각도의 노력을 기울이고 있으며 이러한 서울시의 노력이 시스템 최적화의 측면에서 운영 효율성으로 이어지도록 다양한 그룹에서 이용 효율을 연구하고 있다. 그 중 첫 번째는 대여 및 반납을 예측하는 연구이다. 전체 시스템 또는 대여소 단위로 정해진 시간 간격으로 미래의 대여 및 반납 횟수를 예측하여, 불균형이 발생할 가능성이 높은 대여소를 미리 추정하는 것이다. 두 번째는 자전거 재배치를 최적화하는 것이다. 트럭 등을 이용해서 자전거를 재배치하는 경우 비용이 발생하게 되므로, 이를 최소화하는 재배치 알고리즘을 도출하고 있다. 이러한 첫 선행 연구 중 Holt-Winters 모형과 같은 시계열 분석 및 군집분석과 딥러닝은 공공자전거 대여량 예측에 초점이 맞추어 있다 [2-3]. 기존 연구 는 부스팅 방법을 적용하고 있는데, 특징을 잘 뽑아내기 위한 데이터 셋의 새로운 모니터링 방법을 제시하고 있다 [4]. 기존 연구는 설문지 결과를 가지고 다시 설문지의 문항을 셋팅하는 방법으로 특징에 대한 새로운 관점을 제시하고 있다 [5]. 본 논문에서는 기존 데이터에 있는 특성 외에 새로운 특징 선택(Feature Set)에 대한 분석을 더하여, 기존의 어떤 모형을 활용하더라도 결정계수의 신뢰도가 좀 더 나아질 수 있도록 하는 특징 선택에 관한 것에 관심을 두었다. 이에 본 연구에서는 기존의 데이터에서 ‘불쾌지수’(Temperature Humidity Index)라고 하는 이용자에 대한 행태 분석을 우선 고려하여, 특징의 중요도(Feature-Importance)를 불쾌지수로 다시 설정한 후, 결정계수가 0.01 향상되었음을 보이고 있다.

그림 1. 서울 공유 자전거 시스템, “따릉이”
Figure 1. Seoul Public Bike

Ⅱ. 관련연구

시계열 분석의 대표적인 Holt-Winters 은 지수평활법 (Exponential Smoothing)을 추세, 수준, 계절성의 세 가지 요소로 구분하여 예측하는 방법이다 [2]. 기존 연구 은 재배치 전략에 대해 통계적 방법을 주로 활용하고 있으며, 최근에는 연구 와 같이 시계열 분석에 기반한 수요예측 모델링에 포커스를 맞추어서 활성화하고 있다 [1-2, 6]. 공공자전거, “따릉이”의 대여량을 예측하는 딥러닝 연구도 있다 [3]. 기존 연구은 공공자전거 대여량, 기상 자료 와 지하철 이용량을 수집하여, 지수평활법, Autoregressive Integrated Moving Average (ARIMA) 및 Long Short-Term Memory(LSTM)으로 딥러닝 모형을 구축하여, 평균제곱오차와 평균 절대

오차를 활용한 예측 오차를 비교하였다 [3]. 기존 연구은 시스템의 확장을 고려하면서 신규 대여소 후보 지역 탐색에 주안점을 두었다 [7]. 기존 연구은 딥러닝 모형보다 랜덤 포레스트(Random Forest)라는 머신러닝의 앙상블 모형이 더 나은 정확도를 보인다는 결과를 보여주고 있다 [7].

본 연구는 위의 다양한 연구들을 기반 하여, 딥러닝(LSTM)이나 머신러닝(랜덤 포레스트) 등의 기존 방법을 사용할 때, 데이터의 전처리나 데이터가 가진 특성에 보다 더 집중해야 함을 인지하였고, 이에 연구의 초점을 Feature Set의 선택에 조금 더 맞추고 있다.

III. 본 론

3.1. 데이터 및 전처리

본 연구는 google.colab의 SeoulBikeData.csv 데이터를 기반으로 한다. 데이터가 가진 특징은 그림 2에 나타나 있다. 그림 3에 보이는 것과 같이 전처리를 위하여 python의 pairplot(), corr()등을 사용해 특징들 간의 상관관계를 분석하였고, 이상치 처리를 위해 boxplot()을 사용하였다.

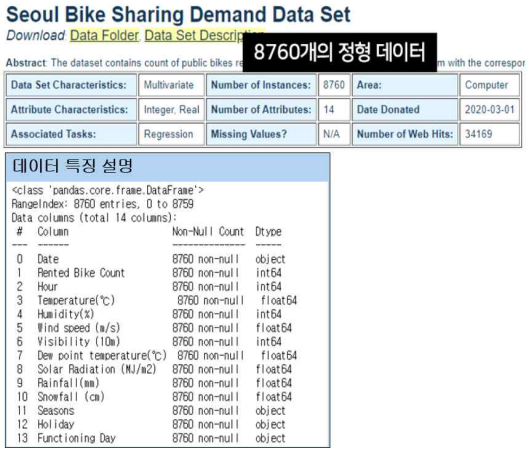

그림 2. 서울 공유 자전거 데이터 세트
Figure 2. Seoul Public Bike Data Set

다양한 전처리를 통한 분석 결과, 대여는 여름철인 5 ~ 7월에 가장 많았고, 겨울철인 12월 ~ 2월에는 다소 저조하며, 월 초에는 낮았다가 둘째 주부터 증가하는 추세를 보이며, 출퇴근 시간에는 증가한다. 그리고, 그림 4에서 보는 바와 같이, "기온에 따른 평균 대여량"을 기반으로 한 예측은 매우 불안정하지만, 기온과 대여량은 밀접한 상관관계가 있음을 확인할 수 있었다.

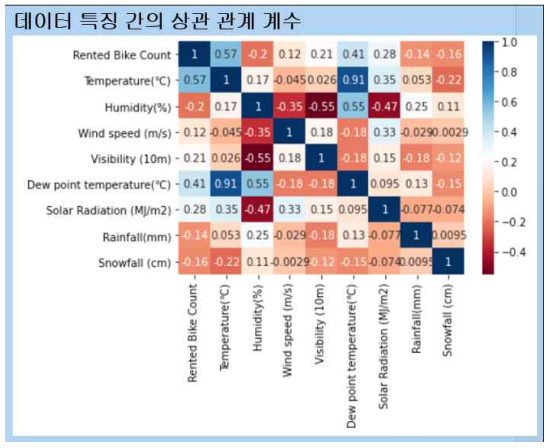

그림 3. 데이터 특징 간의 상관관계
Figure 3. Correlation of Feature Set

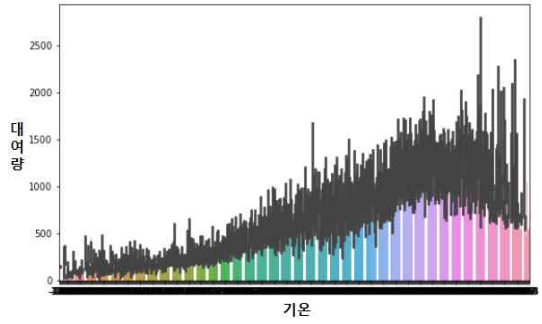

그림 4. 기온과 대여량의 밀접한 상관관계
Figure 4. Close correlation of Rented Bike Count and Temperature

기존의 연구들에 의하면, 시간 및 기온이 주요 입력 특징으로 활용될 수 있다 [2, 6]. 연구들에 근거한 본 연구의 상관관계 분석에 의하면 기온은 주요 특징 지표이긴 하나 다소 불안정하다는 결론에 이르게 되어, 기온이라는 특징을 좀 더 안정적으로 사용하기 위한 Feature Set의 선택에 집중하기로 하였다 [2, 6].

3.2. 특징 선택 및 다양한 방법들

그림 5, 그림 6 및 그림 7에 나타난 바와 같이 본 연구에서는 선형회귀, 랜덤 포레스트 및 그라디언트-부스트를 각각 사용해보았고, 이 중에서 그라디언트-부스트의

결정 계수가 가장 높음을 확인하였다 [7-8].

선형 회귀 Score : 0.66

```
from sklearn.linear_model import LinearRegression as lr
from sklearn.metrics import roc_auc_score, accuracy_score, mean_squared_error, r2_score

model=lr()
model.fit(train_x, train_y)

print("모델의 회귀계수는: ", model.coef_, "이고 모델의 절편은 :", model.intercept_)

pred_y=model.predict(test_x)
print("RMSE on Test set : {0:.5f}",format(mean_squared_error(test_y, pred_y)**0.5))
print("R-squared Score on Test set : {0:.5f}".format(r2_score(test_y,pred_y)))
```

```
모델의 회귀계수는:  [-1.62196701e+01 -1.76923552e+01  2.82066877e+00 -1.48637122e-03
  4.31804609e+01  1.51665116e+02 -4.74864556e+01  3.75748562e+01
 -2.13695070e+01 -1.25435750e+02 -2.31395596e+02 -3.23485304e+02
 -3.76490188e+02 -3.75928684e+02 -1.97614663e+02  1.16917188e+02
  4.64823959e+02 -5.60696144e+01 -2.57429458e+02 -2.51784575e+02
 -2.62704467e+02 -2.21756791e+02 -2.58902906e+02 -1.41646237e+02
 -2.21988550e+01  2.96621468e+02  5.07897714e+02  4.31176770e+02
  4.52222835e+02  4.20501365e+02  3.52143022e+02  8.19082740e+01
  1.55852687e+02  1.11570461e+01  3.92997560e+00 -1.70939708e+02
 -5.54433823e+01  5.54433823e+01] 이고 모델의 절편은 : 1677.0286305749992
RMSE on Test set : {0:.5f} 352.28241417121404
R-squared Score on Test set : 0.66761

R-squared 값으로 0.66761이 나왔다.
```

그림 5. 선형회귀 및 결정계수
Figure 5. Linear Regression and Coefficient Determinant

Random Forest Score : 0.65

```
from sklearn.ensemble import RandomForestRegressor as rfr
from sklearn.metrics import roc_auc_score, accuracy_score, mean_squared_error, r2_score
---

ravel()

다차 배열 => 1차원 배열


rfr()

n_estimators => 결정트리 개수
default = 10
늘어날수록 성능은 증가하나, 시간도 늘어난다.

max_depth => 트리의 깊이
default = 할당은 완벽하게 클래스 값이 결정 될 때까지 분할 됨
너무 깊으면 과적합

min_samples_split
노드를 분할하기 위한 최소한의 샘플 데이터
가본 => 2
작게 설정할 수록 분할 노드가 많아져 과적합 가능성이 높아진다.

min_samples_leaf
leaf노드가 되기 위해 필요한 최소한의 샘플 데이터
"""
train_y = np.ravel(train_y, order='C')

model=rfr(n_estimators=100, max_depth=5, min_samples_split=30, min_samples_leaf=15)
model.fit(train_x, train_y)

pred_y = model.predict(test_x)
print("RMSE on Test set : {0:.5f}",format(mean_squared_error(test_y, pred_y)**0.5))
print("R-squared Score on Test set : {0:.5f}".format(r2_score(test_y,pred_y)))
```

```
RMSE on Test set : {0:.5f} 358.4969179300232
R-squared Score on Test set : 0.65578

R-squared 값이 0.65578로 선형회귀보다 조금 낮게 나왔다.
```

그림 6. 랜덤 포레스트 및 결정 계수
Figure 6. Random Forest and Coefficient Determinant

본 연구에서는 트리(Tree) 기반 앙상블 방법을 사용하는 부스팅 기법인 Gradient Boosted Regression Trees, (GBRTs)를 고려하였는데, 추후에는 XGboost 및 LightGBM이라는 GBRTs의 후속 모델을 사용하는 것을 고려하고 있다. 이러한 기계학습기법에서 발생하는 가장 큰 문제점 중에 하나는 과적합(overfitting)이다 [9-11]. 부스팅 기법은 트리기반 데이터 셋을 학습하지만, 순차적으로 오차가 큰 부분에 가중치를 부여하는 방식으로 트리를 학습하여 과적합에 매우 취약하다. 따라서, 그리드 서치(Grid Search)나 랜덤 서치(Random

Gradient Boosting Score :0.83

gradientboosting

```
from sklearn.ensemble import GradientBoostingRegressor as grb
from sklearn.metrics import roc_auc_score, accuracy_score, mean_squared_error, r2_score


train_y = np.ravel(train_y, order='C')

---
grb
---
model=grb(n_estimators=100, learning_rate=0.1,max_depth=5, min_samples_split=30, min_samples_leaf=15)
model.fit(train_x, train_y)

pred_y = model.predict(test_x)
print("RMSE on Test set : {0:.5f}",format(mean_squared_error(test_y, pred_y)**0.5))
print("R-squared Score on Test set : {0:.5f}".format(r2_score(test_y,pred_y)))
```

```
RMSE on Test set : {0:.5f} 245.1885760473574
R-squared Score on Test set : 0.83898

R-squared 값이 0.83898으로 랜덤 포레스트보다 높게 나왔다.
```

```
#Feature 중요도 확인

import matplotlib.pyplot as plt
import seaborn as sns

grb_importances_values = model.feature_importances_
grb_importances = pd.Series(grb_importances_values, index=train_x.columns)
grb_top10 = grb_importances.sort_values(ascending=False)[:10]

plt.figure(figsize=(8,6))
plt.title('Top 10 Feature Importances')
sns.barplot(x=grb_top10, y=grb_top10.index, palette="RdBu")
plt.show()
```

그림 7 . 그라디언트 부스트 및 결정 계수
Figure 7. Gradient Boosting and Coefficient Determinant

Search)와 같은 방법을 사용하여 다양한 하이퍼-파라미터들을 최적화시킴으로써 과적합 문제를 예방할 수 있다 [9-11]. 또한 AutoML을 지원하는 pycaret을 사용하여 다양한 모델간의 비교를 통하여 이를 극복할 수 있다.

본 연구에서는 그리드 서치를 사용하여 하이퍼-파라미터를 튜닝하였으며, 앙상블 모형에 대응되는 학습 속도(learning_rate)와 트리 개수 (n_estimators), 그리고 부스팅에 사용되는 결정트리 파라미터인 최소 샘플 분할 (min_samples_split) 과 최소 샘플 리프 (min_samples_leaf)를 각각 그룹화하여 모형을 개선하였다. 표1의 결과를 보면, GBRTs의 결정계수는 0.848이고, 하이퍼-파라미터 튜닝 후의 결정 계수는 0.849이다. 테스트 셋의 정확도가 하이퍼-파라미터 튜닝을 계속할 경우, 조금 더 나은 결과를 보일 수 있을 것으로 생각된다.

본 논문은 Feature Set에 초점을 맞추고 있고 "기온"이라는 특징이 보다 더 안정적으로 사용될 수 있음에 착안하여, 이에 "불쾌지수"라는 새로운 특징을 Equation 1.으로 사용하였다.

data['Temperature Humidity Index'] = round(1.8*data['Temperature(Celsius)'] - 0.55*(1-data['Humidity(%)']/100)*(1.8*data['Temperature(Celsius)']-26)+32) ---------- Equation 1.

Ⅳ. 실험 및 결과

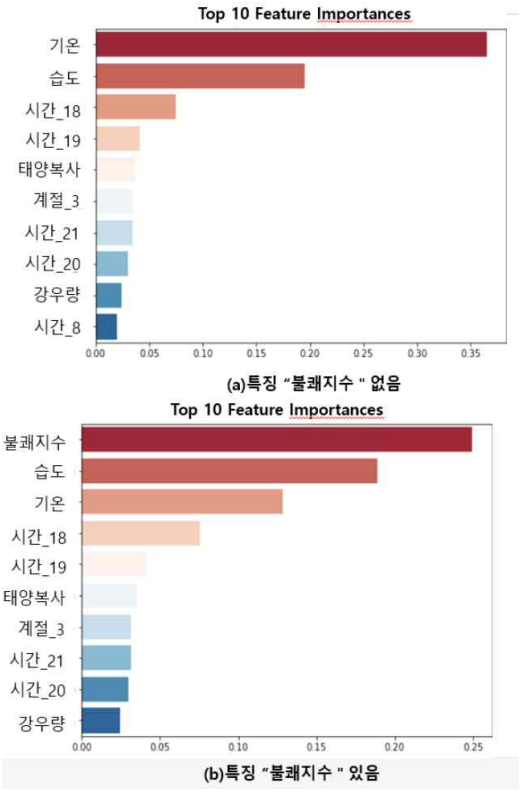

그림 8. "불쾌지수"라는 특징을 포함한 특징 간의 중요성 비교
Figure 8. Comparison of Importance of Feature Set including "Temperature Humidity Index"

표 1. 다양한 방법 간의 결정계수 비교
Table 1. Comparison of coefficient of determination

방법	결정계수
GBRTs	0.848
Hyper-Parameter Tuning after GBRTs	0.849
New Feature Set after Tuning	**0.854**

표 1에서 보는 바와 같이, 기존의 방법을 사용하더라도, Feature Set을 적절히 선택할 경우, 다소 나은 정확도를 보임을 알 수 있다. 표 1에서 사용한 방법 중에서 첫 번째 행은 트리 기반의 앙상블 방법 중 GBRTs이며, 두 번째 행은 하이퍼 파라미터 튜닝을 한 GBRTs이며, 세 번째 행은 "불쾌지수"라는 특징을 더 포함하여 하이퍼 파라미터 튜닝을 한 GBRTs이다. 표1의 제일 마지막 행이 본 논문에서 제안한 방법이며, 결정계수 0.854로 기존 GBRTs인 첫 번째 방법보다 본 논문에서 제안한 세 번째 방법이 가장 결정 계수가 나아졌음 알 수 있다. 그림 8은 그러한 Feature Set의 정확도(Importance)에 대한 비교이다.

Ⅴ. 결 론

본 연구에서는 기존의 다양한 연구들에서 제안된 딥러닝이나 머신러닝의 수요 예측모델링을 기반으로 하여 적절한 Feature Set을 선택함으로써 보다 나은 결정계수를 획득할 수 있음을 볼 수 있었다. 이로써, 수요예측의 정확도는 다양한 모델링을 통해서도 높일 수 있지만, 적절한 특징 선택에 기반 해서도 다소 높일 수 있음을 알 수 있다.

References

[1] E. LEE and B. SON, "Optimal Rebalancing Strategy for Public Bike-sharing System in Seoul", J. Korean Soc. Transp. Vol.37, No.1, pp. 28-38, 2019. https://doi.org/10.7470/jkst.2019.37.1.027
[2] C. Chatfield, "The Holt-Winters Forecasting Procedure," Journal of the Royal Statistical Society, Series C(Applied Statistics), Vol.27, No.3, pp.264-279, 1978. https://doi.org/10.2307/2347162
[3] K. CHO. S. S. Lee, and D. H. Nam, "Forecasting of Rental Demand for Public Bicycles Using a Deep Learning Model," J. Korea Inst. Intell. Transp. Syst., vol.19, no.3, pp.28-37, 2020. https://doi.org/10.12815/kits.2020.19.3.28
[4] S. Park, M. Kim, and J. Im, "Estimation of Ground-level PM10 and PM2.5 Concentrations Using Boosting-based Machine Learning from Satellite and Numerical Weather Prediction Data," Korean Journal of Remote Sensing, Vol.37, No.2, pp.321~335, 2021. https://doi.org/10.7780/kjrs.2021.37.2.11
[5] T. Park and C. Kim, "Predicting the Variables That Determine University (Re-)Entrance as a Career Development Using Support Vector Machines with Recursive Feature Elimination: The Case of South Korea," Sustainability, MDPI, Vol.12, No.18, pp.1-11, 2020. https://doi.org/10.3390/su12187365
[6] H. Lim and K. Chung, "Development of Demand Forecasting Model for Seoul Shared Bicycle,"

Jour. of KoCon.a, vol.19, no.1, pp.132-140, 2019. https://doi.org/10.5392/JKCA.2019.19.01.132

[7] Kyung-Ok Kim, "A Study on the Characteristics of Shared Bike Use for Operation of a Shared Bike System Considering Bicycle Imbalance," 2018 Seoul Research Paper Competition, Seoul Research Institute, 16th 2018. https://www.si.re.kr/node/61093

[8] S. M. Woo., G. Y. Kim. and H. C. Kim, "How to Improve Suitability of Irradiation Utilization in Development of Linear Regression Model for Estimating Paprika Productivity", The Journal of the Convergence on Culture Technology (JCCT), Vol. 7, No. 4, pp.779-783, November 30, 2021. https://doi.org/10.17703/JCCT.2021.7.4.779

[9] C. Kim, and T. Park, "Predicting Determinants of Lifelong Learning Intention Using Gradient Boosting Machine (GBM) with Grid Search," Sustainability MDPI, Vol.14, No.9, Article Number. 5256, 2022. https://doi.org/10.3390/su14095256

[10]S. H. Moon, "Analysis of AI-Applied Industry and Development Direction," The Journal of the Convergence on Culture Technology (JCCT), Vol. 5, No. 1, pp.77-82, February 28, 2019. https://doi.org/10.17703/JCCT.2019.5.1.77

[11]J Son, C. Kim, and M. Jeong, "Unsupervised Learning for Anomaly Detection of Electric Motors," Vol.23 Issue 4, pp.421-427, April 2022. http://doi.org/10.1007/s12541-022-00635-0