딥러닝 모형을 활용한 공공자전거 대여량 예측에 관한 연구

Forecasting of Rental Demand for Public Bicycles Using a Deep Learning Model

조 근 민* · 이 상 수** · 남 두 희***

* 주저자 : 아주대학교 교통연구소 연구원
** 교신저자 : 아주대학교 교통공학과 교수
*** 공저자 : 한성대학교 사회과학부 교수

Keun-min Cho* · Sang-Soo Lee** · Doohee Nam***

* Researcher, Ajou Transp. Research Institute
** Professor, Dept. of Transportation Eng., Ajou Univ.
*** Professor, School of Social Science., Hansung Univ.
† Corresponding author : Lee Sangsoo, sslee@ajou.ac.kr

Vol.19 No.3(2020)
June, 2020
pp.28~37

pISSN 1738-0774
eISSN 2384-1729
https://doi.org/10.12815/kits.2020.19.3.28

요 약

본 연구는 공공자전거의 대여량을 예측하는 딥러닝 모형을 개발하였다. 이를 위하여 공공자전거 대여량 자료, 기상 자료, 그리고 지하철 이용량 자료를 수집하였다. 지수평활 모형, ARIMA 모형과 LSTM기반의 딥러닝 모형을 구축한 후 MSE와 MAE 평가 지표를 사용하여 예측 오차를 비교·평가하였다. 평가 결과, 지수평활 모형으로 MSE 348.74, MAE 14.15 값이 산출되었다. ARIMA 모형으로 MSE 170.10, MAE 9.30 값을 얻었다. 그리고 딥러닝 모형으로 MSE 120.22, MAE 6.76 값이 산출되었다. 지수평활 모형의 값과 비교하여 ARIMA 모형의 MSE는 51%, MAE는 34% 감소하였다. 그리고 딥러닝 모형의 MSE는 66%, MAE는 52% 감소하여 딥러닝 모형의 오차가 가장 적은 것으로 파악되었다. 이러한 결과로부터 공공자전거 대여량 예측 분야에서 딥러닝 모형의 적용시 예측 오차를 크게 감소시킬 수 있을 것으로 판단된다.

핵심어 : 딥러닝, 자전거, 예측, LSTM, 수요

ABSTRACT

This study developed a deep learning model that predicts rental demand for public bicycles. For this, public bicycle rental data, weather data, and subway usage data were collected. After building an exponential smoothing model, ARIMA model and LSTM-based deep learning model, forecasting errors were compared and evaluated using MSE and MAE evaluation indicators. Based on the analysis results, MSE 348.74 and MAE 14.15 were calculated using the exponential smoothing model. The ARIMA model produced MSE 170.10 and MAE 9.30 values. In addition, MSE 120.22 and MAE 6.76 values were calculated using the deep learning model. Compared to the value of the exponential smoothing model, the MSE of the ARIMA model decreased by 51% and the MAE by 34%. In addition, the MSE of the deep learning model decreased by 66% and the MAE by 52%, which was found to have the least error in the deep learning model. These results show that the prediction error in public bicycle rental demand forecasting can be greatly reduced by applying the deep learning model.

Key words : Deep learning, Bicycle, Forecasting, Long short-term memory, Demand

Received 26 May 2020
Revised 15 June 2020
Accepted 17 June 2020

Ⅰ. 서 론

공공자전거는 교통 혼잡을 완화하고 접근성을 높여주는 친환경적인 교통수단으로 인식되고 있으며, 이에 따라 많은 지자체에서 선도적인 교통사업으로 추진되고 있다. 서울시에서는 2015년 10월, '따릉이'라는 명칭으로 공공자전거 사업을 시작하였으며 현재는 약 1,500개의 대여소, 약 2만대의 자전거를 운영 중이다. 이 사업은 2017년과 2018년 서울시 정책 만족도 1위를 차지하였으며 2019년 11월까지 이용 3,000만 건을 돌파하였다. 등록된 회원 수는 약 166만 명으로, 서울시민 6명 중 1명꼴로 가입했다고 볼 수 있다. 서울시는 이러한 인기에 힘입어 '세계에서 가장 많은 공공자전거를 소유한 도시'를 목표하고 있으며 그에 따른 적정 규모의 산정 및 사업 효율화 방안을 마련 중이다.

그러나 이용 수요가 많아지는 만큼 문제점들도 발생하고 있다. 이 중 이용자 측면에서 가장 큰 문제점은 수요와 공급의 불균형이 존재한다는 것이다. 또한 운영자 측면에서는 정확한 수요예측 방법이 확립되지 않고 운영되고 있다는 것이다. 예를 들어, 강북 지역의 공공자전거 대여량은 강남 지역보다 약 3만 건 많은데 비하여 공급되는 자전거 댓 수는 오히려 406대가 적었다. 이런 문제는 지역별 대여량 예측 값의 큰 오차 때문에 발생된 것으로 자전거 수요에 맞는 적정 공급량을 예측하지 못하여 이용자에게 불편을 제공하는 것은 물론 사업 운영에 따른 적자도 함께 증가하는 원인이 되고 있다. 또한 향후 사업 확대 시 보다 정확한 정량적 근거를 제시할 수 없는 한계점도 갖고 있다.

국내에서는 공공자전거 대여량 예측을 위하여 Holt-Winters 모형이나 시계열 및 군집분석 모형 등을 사용한 연구들이 수행되었다. 그러나 최근 국내외적으로 패턴 인식과 예측분야에서 활발하게 적용되는 딥러닝 모형을 활용한 공공자전거 대여량 예측 연구는 없는 실정이다. 딥러닝 모형은 빅데이터를 사용할 수 있으므로 전통적인 기법보다 예측력이 높을 것으로 예상하고 있다. 반면 국외에서는 딥러닝 모형을 활용하여 교통량 및 자전거 수요 예측을 위한 연구가 활발하게 진행되고 있다.

본 연구의 목적은 공공자전거의 대여량을 예측하는 딥러닝 모형을 개발하여 평가하는 것이다. 이를 위하여 공공자전거 대여량 자료를 활용하였고, 또한 이와 상관관계가 높은 변수인 기상 자료, 지하철 이용량 자료를 수집하였다. 이를 가공하여 지수평활 모형, ARIMA 모형과 딥러닝 모형을 각각 구축한 후 평가 지표를 선정하여 예측 오차를 비교· 평가하였다. 이러한 결과로부터 공공자전거 대여량 예측 분야에서 딥러닝 모형의 적용 가능성을 평가하여 제시하고자 한다.

Ⅱ. 선행연구 고찰

1. 공공자전거 관련 특성 연구

Kim et al.(2012)은 공공자전거 대여량과 연관성이 있을 수 있는 많은 요인들에 대한 분석을 통하여 날씨, 회원 여부, 휴일 여부 등이 어떠한 영향을 미치는지를 분석하였다. 여러 요인 중 기온과 강수량, 구름의 양이 공공자전거 대여량에 영향을 주었고, 또한 회원 및 휴일 여부에 따른 대여량을 비교한 결과, 비가 오는 날엔 회원의 대여량이 비회원의 대여량보다 높은 결과를 나타내는 것을 도출하였다.

Do and Noh(2014)는 대전시 공공자전거'타슈'를 대상으로, 이용자들의 공공자전거 대여 패턴과 특성을 조사하였다. 다중회귀분석으로 분석한 결과, 공원 주변의 대여량이 비교적 높았고, 평일보단 주말의 대여량이 높은 것으로 드러났다. 이 외에도 주변 인구수와 자전거 도로의 길이, 버스 승·하차 인원수, 지하철 출입구로부터의 거

리, 수변공간까지의 거리 등이 영향력 있는 변수로 분석되었다. 선정된 모형의 결정계수 값은 0.74로 나타났다.

Faghih-Imani et al.(2014)은 캐나다 몬트리올의 공공자전거인 'BIXI'를 대상으로 대여소에서의 대여 특성과 반납 특성에 관한 기상 요인, 시간대 요인, 주변 지역의 인프라 요인, 토지 이용과 주변 환경의 요인을 조사하였다. 분석 결과, 대여소 주변의 상업 시설 요인과 교육 시설 요인, 업무 시설 요인은 공공자전거의 대여 특성과 반납 특성, 대여소 개수 등에 큰 영향을 미치는 것으로 분석되었다.

Lee et al.(2016)은 대여소의 위치와 기상 조건이 공공자전거 대여 및 이용패턴에 미치는 영향을 파악하였다. 선형회귀분석방법으로 분석한 결과, 대여량은 평균 기온이 높아질수록 증가하는 것으로 나타났다. 강수량 10mm 이상, 평균기온 29도 이상, 그리고 풍속이 7 m/s 이상이면 대여량이 감소하는 것으로 제시하였다. 또한 지하철 출입구 인근 대여소의 대여량은 퇴근 시간대에 높았으며, 공원과 일반상업지역의 경우 오후 시간대대여량이 높은 것으로 분석되었다.

Sa(2019)는 서울시 공공자전거'따릉이'를 대상으로 공공자전거 대여소로부터 떨어진 거리에 따른 토지 이용 요소들의 영향력 차이를 분석하였다. 다중회귀모형을 사용하여 분석한 결과, 대여소 인근 토지 이용 특성, 즉 주거·업무 지역 여부, 교육·상업 지역 여부, 공원이나 지하철 출입구, 자전거 도로 등의 시설 특성에 따라 공공자전거 대여량에 영향을 미치는 요인인 것으로 확인되었고, 거리 역시 영향력의 강도에 유의한 변수로 확인되었다.

2. 예측 모형 관련 연구

Kaltenbrunner et al.(2010)은 바르셀로나의 공공자전거 'Bicing'을 대상으로 특정시간대 사람들의 이용 특성을 분석하였다. 운영자 홈페이지의 자료를 기반으로 도시내 시·공간적 이용자 형태를 예측하고자 하였다. 이를 위하여 시계열 분석 기법인 ARMA(Auto Regressive Moving Average) 모형을 적용하였으며 시간별 예측 대여량과 실제 대여량을 비교하여 평균제곱오차와 평균절대오차를 도출하였다. 이러한 결과로부터 자전거 이용자의 수요를 미리 예측할 수 있음을 확인하였다.

Fu et al.(2016)은 교통류가 확률적이고, 비선형적인 특성을 가지고 있기 때문에 ARMA와 ARIMA(Auto Regressive Integrated Moving Average) 같은 기존 시계열 모형보다는 딥러닝 모형을 이용한 교통류 예측이 더욱 적합하다고 판단하였고, 단기 교통류를 예측하기 위하여 LSTM(Long Short-Term Memory)과 GRU(Gated Recurrent Units) 두 가지 신경망을 이용하여 비교 평가하였다. 평가 결과, LSTM과 GRU같은 순환신경망 기반 딥러닝 모형이 ARIMA 모형보다 예측 오차가 작다는 것을 보여주었다.

Min et al.(2017)은 대전광역시의 공공자전거 자료로 대여량을 예측하고자 2013년과 2014년의 대여 정보 및 기상 정보를 가공하여 공공자전거 이용패턴을 분석하였다. 랜덤 포레스트 알고리즘을 이용하여 2015년의 대여량을 예측하였고 실제 대여량과의 오차를 산출하였다. 평가 결과로부터 평균제곱근오차는 낮게 관측되어 향후 공공자전거 재배치 작업의 효율성을 높일 수 있을 것으로 전망하였다.

Yang et al.(2015)은 시단위의 공공자전거 이용량을 예측하기 위하여 컨벌루션 신경망 기반의 딥러닝 모형을 개발하였다. 과거의 자료에 기상자료를 추가하여 구축한 모형은 기존의 모형과 비교하여 우수한 예측 성능을 보였다. 또한 연구에서는 제안한 모형의 성능이 인접한 정류장수, 패치 크기, 학습 비율 등과 같은 다양한 변수를 포함함에 따라 변할 수 있음을 보여주었다. 따라서 사용되는 변수에 대한 정산이 필요하며, 이를 통하여 딥러닝 모델의 활용 가능성을 제시하였다.

Lim and Chung(2019)은 시계열 분석 기법인 Holt-Winters 모형을 이용하여 공공자전거의 대여량을 예측하였고 지수평활 모형과 비교하여 평가하였다. 1년 6개월 동안의 실제 대여량 자료를 이용하였고, 평가 지표로

평균제곱오차를 사용하였다. 분석 결과 Holt-Winters 모형은 적정한 오차값을 나타내었고, 이러한 예측모형을 이용하여 대여소별 수요예측을 실시한 후 공유자전거 재배치에 활용할 수 있다고 판단하였다.

3. 시사점

공공자전거의 이용 특성과 관련한 연구에서는 주로 자전거 대여량에 영향을 미치는 독립변수를 조사하였으며 기온과 풍속, 강수량, 지하철 출입구, 주말 여부, 토지 이용 특성 등이 유의한 것으로 나타났다. 그리고 예측 모형 관련 연구에서는 공공자전거 대여량이 시계열 자료임을 공통적으로 인식하여 국내 연구는 주로 ARIMA같은 시계열 분석 기법을 적용하고 딥러닝을 적용한 연구는 전무하였으며 국외에서는 컨벌루션 신경망을 이용한 딥러닝 모형을 적용한 연구 결과를 제시하고 있다.

본 연구에서는 기존 연구에서 언급된 기상 특성이외에 지하철 이용 및 시간대 특성을 함께 사용하여 대여량 예측과정에 보다 국내 환경에 적합한 변수들을 포함하도록 하였다. 이때, 토지 이용 특성에 따른 영향을 상쇄하기 위해서 주거 시설과 업무 시설, 근린 시설이 균일하게 분포되어있는 곳을 분석 지점으로 선정하였다. 또한 시계열 자료에 효과적인 LSTM 기반 순환신경망을 사용한 딥러닝 모형을 구축하여 예측력이 향상되도록 구성하였다.

Ⅲ. 자료 수집 및 분석

1. 대상 지점 선정 및 자료 수집

본 연구의 자료는 크게 자전거 대여량 자료와 기타 자료의 2개로 구성되었다. 먼저, 자전거 대여량 자료를 구성하기 위하여 운영되는 각 대여소별로 대여량 값을 추출하였다. 두 번째로, 기타 자료를 구성하기 위하여 선정된 지점의 기상 자료와 지하철 이용량 자료를 수집하였다. 기존 연구결과에 근거하여 공공자전거 대여량과 상관성이 높은 변수로 기상 변수를 추가하였고, 또한 지하철 이용량 변수를 추가하였다. 기상 자료는 기온, 풍속, 강수량의 자료를 수집하였고, 지하철 이용량 자료는 대여소 인근에 위치한 지하철역의 이용량을 수집하였다.

대여량 자료를 수집하기 위하여 서울시의 공공데이터 개방 포털인 열린 데이터 광장에서 '공공자전거 대여이력' 자료를 이용하였다. 대여이력 자료의 시간적 범위는 2017년 1월 1일부터 2019년 5월 31일까지로 설정하였고, 공간적 범위는 서울시의 모든 공공자전거 대여소를 포함하도록 설정하였다. 분석 단위는 대여소로 설정하여 대여소별 대여량을 분석하였다. 이는 사업의 확장 시 대여소별로 공공자전거를 추가하거나 새로운 대여소를 신설하는 등 대여소 단위로 진행되기 때문이다.

이러한 공간적 범위에서 본 연구의 대상 지점은 다음과 같은 몇 가지 특징을 검토하여 선정하였다. 먼저, 대여량이 큰 대여소를 우선적으로 선정하였다. 이는 대여량이 많은 대여소의 분석 결과는 대여량이 적은 곳과 비교하여 전체적인 운영 효율이나 오차에 미치는 영향이 크기 때문이다. 또한 반경 100m 이내에 주거 시설, 업무 시설, 근린 시설이 모두 존재하는 토지 이용 특성을 갖는 대여소를 선정하고자 하였다. 이는 각 시설의 편중에 따른 불균형을 제외하기 위함이다(Sa, 2019). 그리고 기상 자료 및 지하철 이용량 자료의 수집이 가능한 지점이다.

2017년 1월 1일부터 2019년 5월 31일까지의 대여량을 조사한 결과, '여의나루역 1번 출구 앞'대여소의 대여량이 약 16만 건으로 가장 많았고, '뚝섬 유원지역 1번 출구 앞' 대여소가 13만 건, '홍대입구역 2번 출구

앞’ 대여소 12만 건 순으로 나타났다. 그리고 ‘여의나루역 1번 출구 앞’ 대여소 100m 이내로 한국전력공사 남서울지역본부와 서울아파트가 위치하고 여의동로를 건너면 여의도 한강공원이 위치하여, 주거·업무 시설 및 근린 시설이 모두 존재하는 토지이용 특성을 만족하였다. 그리고 기상자료 및 여의나루역 지하철 이용량 수집이 가능하였다. 따라서 본 연구에서는 최종적으로 ‘여의나루역 1번 출구 앞’대여소를 대상 지점으로 선정하였다. 선정된 지점의 대여량 자료는 공공자전거 대여이력 자료에서 추출하여 사용하였다. 기상 자료는 기상 자료 개방 포털을 이용하여 대상 지점의 기온, 풍속, 강수량 자료를 포함하도록 하였다. 지하철 이용량 자료는 서울 열린 데이터 광장을 이용하였고, 여의나루역 지하철 이용량 자료를 추출하였다.

2. 최종 데이터 셋 구성

수집한 대여량 자료와 기타 자료를 딥러닝 모형에 입력하기 위하여 하나의 데이터 셋으로 가공하였다. 이때, 공공자전거 대여량은 출퇴근 시간 등의 영향을 받기 때문에 1시간 단위 분석이 가장 적합하다고 판단하여 데이터 셋을 1시간 단위로 구성하였다. 또한 대여량 자료는 시계열 특성을 가지고 있으며, 해당 특성을 딥러닝 모형에 입력하기 위하여 기상 자료와 지하철 이용량 자료 외에도 연, 월, 일, 시, 요일 등 시간 정보를 추가하였다.

가공된 수집 자료의 최종적인 형식은 다음 <Table 1>과 같다. 대여량 자료는 2017년 1월 1일 00시부터 2019년 5월 31일 23시까지 1시간 단위로 구성되었다. ‘2017-01-01 00:00’행의 대여량은 2017년 1월 1일 00시 00분부터 00시 59분까지의 대여량을 의미한다. 또한 기타 자료와 결합하기에 앞서 연, 월, 일, 시, 요일의 시간 정보를 먼저 입력하였다. 계절, 시간대, 주말의 여부 등 시간 관련 특성은 공공자전거 이용에 영향을 미치며 이를 통해 시간 정보를 딥러닝 모형에 따로 입력할 수 있게 된다. 다음으로는 기상 자료의 1시간 평균 기온과 풍속, 1시간 총합 강수량을 추가하였다. 그리고 지하철 이용량 자료와 자전거 대여량의 관계를 명확하게 설명하는 연구결과가 없는 관계로, 본 연구에서는 분석 대여소의 입지 특성 상 ‘대여’자료는 지하철의 ‘하차’자료와 상관이 있다고 가정하고 하차했을 때의 이용량을 직접적으로 한정하여 적용하였다. 따라서 지하철 운행을 종료하는 오전 1시부터 오전 5시까지의 이용량 자료는 부재하므로 값 ‘0’을 기입하였다. 최종적으로 가공된 자료는 21,145개의 행과 11개의 열로 정리되었고, 약 22만개의 데이터를 가지는 데이터 셋을 구성하였다.

<Table 1> Final Data Set

Rental Date & Time	Rental Demand	Year	Month	Date	Hour	Day	Temperature	Wind Speed	Precipitation	Subway Usage
2017-01-01 00:00	0	2017	1	1	0	Sun.	3.0	1.6	0	0
2017-01-01 01:00	0	2017	1	1	1	Sun.	2.8	1.1	0	0
2017-01-01 02:00	0	2017	1	1	2	Sun.	2.5	1.7	0	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
2019-05-31 21:00	36	2019	5	31	21	Fri.	18.5	1.7	0	487
2019-05-31 22:00	44	2019	5	31	22	Fri.	17.9	2.5	0	293
2019-05-31 23:00	22	2019	5	31	23	Fri.	17.4	1.7	0	150
Total	161,708	-	-	-	-	-	-	-	-	11,005,395

3. 상관관계 분석

Pearson 상관분석을 통해 본 연구에서 수집된 대여량과 기상 자료, 지하철 이용량 자료의 관계성을 분석하였다. Pearson분석은 명목형 변수를 사용하지 않고 연속형 변수를 사용하며, 대여량, 지하철 이용량, 기온, 강수량, 풍속을 선정하여 분석하였다. 분석 결과, 대여량과 지하철 이용량간의 상관계수가 0.5로 가장 높게 나타났으며 이는 두 변수가 비교적 강한 선형관계에 속한다는 것을 의미한다. 대여량과 기온 간의 상관계수는 0.42로 두 번째로 높았다. 지하철 이용량의 경우와 마찬가지로 기온은 대여량과 비교적 강한 선형관계에 속하였다. 대여량과 강수량의 상관계수는 -0.1로, 약한 부적인 선형관계에 속하며, 풍속과의 상관계수는 0.06으로 선형관계가 거의 없는 것으로 나타났다. 그러나 이는 두 변수가 선형적인 관계가 없다는 뜻일 뿐, 관계가 아예 없다는 것은 아니다(Kim et al., 1989). 기존의 연구에서 대여량과 풍속의 관계가 유의하다는 결과가 있으므로 본 연구에서도 독립변수로 사용하였다.

Ⅳ. 모형 구축 및 평가

본 연구에서 공공자전거의 대여량을 예측하기 위한 딥러닝 모형을 구축하였고 기존 연구에서 많이 사용된 시계열 분석 모형과 비교하였다. 이때 시계열 분석 기법은 지수평활 모형과 ARIMA 모형 두 가지를 선정하였다. 평가 지표로는 평균제곱오차(Mean Square Error, MSE)와 평균절대오차(Mean Absolute Error, MAE)를 선정하였다. 지수평활 모형, ARIMA 모형, 딥러닝 모형을 각각 구축하기 위하여 2017년 1월 1일 00시부터 2019년 4월 30일 23시까지 1시간 단위의 데이터 셋을 이용하였다. 파라미터를 다양한 값으로 적용하며 각각의 모형을 구축하면서, MSE와 MAE가 가장 작은 모형을 최적 모형으로 선정하였다. 다음으로는 5월 1일 00시부터 5월 31일 23시까지 1시간 단위의 자료를 이용하여 예측모형의 평가를 실시하였다. 즉, 각 모형별 예측 대여량과 실제 대여량을 비교하여 평균제곱오차와 평균절대오차를 산출하고 이를 이용하여 각 모형을 비교 평가하였다. 딥러닝 모형 구축을 위해 모델링에 우수한 Keras를 이용하며 빅데이터 분석 및 프로그래밍에 적합한 Python 3.7을 이용하였다.

1. 지수평활 모형 구축

지수평활 모형은 단변량 예측 방법으로서, 독립변수와 종속변수는 동일하다. 즉 대여량 외에 다른 변수를 입력하지 않는다. 따라서 20,400개의 1시간 단위 대여량을 모형을 구축하는 자료로 사용하였다. 지수평활 모형을 구축하기 위해서는 먼저 초기값을 적절하게 설정해야 하며, 이는 초기에 발생한 오류의 누적을 줄이기 위해서다. 본 연구에서는 초기 1일 동안의 대여량을 평균한 값을 사용하였다. 다음으로 지수평활계수(Alpha)를 설정해야한다. 지수평활계수는 시계열자료에서 기존 실제값에 대해 부여하는 가중치며 보통 0.05에서 0.30까지의 값으로 설정한다. 따라서 본 연구에서도 0.05부터 0.30까지의 지수평활 계수를 사용하여 각 계수별로 모형의 오차를 평가하였고, 결과는 <Table 2>와 같다. 지수평활계수가 0.30일 때 평균제곱오차 61.57, 평균절대오차 4.78로 오차가 가장 적은 것으로 나타났다. 따라서 0.30의 지수평활계수를 갖는 지수평활 모형을 최종 모형으로 선정하였다.

<Table 2> Results from Exponential Smoothing Models

Index	Smoothing parameter, Alpha			
	0.05	0.10	0.20	0.30
MSE	92.57	86.92	73.60	61.57
MAE	6.01	5.79	5.28	4.78

2. ARIMA 모형 구축

ARIMA 모형도 지수평활 모형처럼 단변량 예측 방법을 사용하므로 모형 구축자료는 지수평활 모형과 동일하게 적용하였다. ARIMA 모형은 분석하기에 앞서 수집된 자료의 정상성을 확인해야 한다. 따라서 ADF 검사(Augmented Dickey-Fuller test)를 시행하였다. ADF 검사는 분석할 시계열 자료에 단위근(unit root)이 존재하는지 확인하고, 이를 통해 자료가 정상적인 자료인지 판별한다. 귀무가설은 '자료에 단위근이 존재하며 정상적이지 않음.'이며, 대립가설은 '자료에 단위근이 존재하지 않으며 자료가 정상적임.'이다. 추세성을 상쇄하기 위해 1차 차분한 자료를 확인한 결과 p-value는 1.4e-14로, 귀무가설을 기각하였다. 다음으로 ARIMA의 p, q파라미터를 결정하여 모형을 구축하였다. 이를 위하여 자기상관함수(Autocorrelation Function, ACF)와 편자기함수(Partial Autocorrelation Function, PACF)를 이용한 자동 탐색을 진행하였다. 그 결과 p=1, q=1의 값이 산출되었으며, ARIMA 모형은 최종적으로 ARIMA(1,1,1)로 선정되었다. 최종 모형의 평가 지표를 산출한 결과 MSE는 92.57, MAE는 6.01이었다.

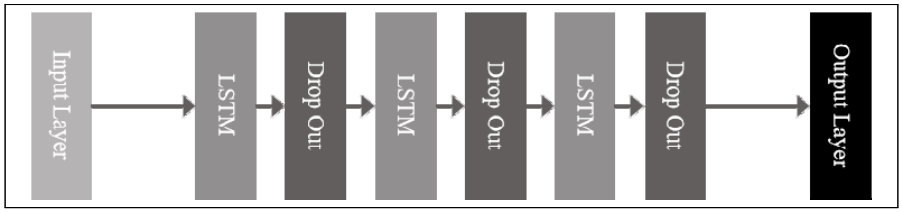
3. 딥러닝 모형 구축

1) 데이터 구성

모형 구축에 앞서, 약 22만개의 데이터로 구성된 데이터 자료를 훈련·검증·시험 데이터 3가지로 분류하였다. 먼저, 2017년 1월부터 2019년 3월까지의 196,800개 데이터를 훈련데이터로 설정하여 모형의 학습에 사용하였다. 검증 데이터는 2019년 4월에 해당하는 데이터 자료로 설정하여 과적합을 효과적으로 방지하도록 하였다. 그리고 시험 데이터는 다른 모형과 마찬가지로 2019년 5월에 해당하는 데이터 자료를 이용하였다.

2) 모형 및 파라미터 구성

본 연구에서는 많은 신경망 모형중에서 시계열 자료 분석에 효과적이라고 알려진 LSTM을 활용한 순환신경망을 사용하였다. LSTM의 메모리 셀은 32개로 구성하여 예측 성능을 높이고 드롭아웃을 0.3으로 설정하여 과적합을 방지하도록 구성하였다. 또한 상태유지 모드를 활성화하여 먼저 산정된 가중치가 다음 학습 시 초기 상태로 입력이 되도록 하였다. 활성화함수는 현재까지 가장 보편적이고 우수하며 수렴속도가 빠르다고 평가되어진 Adam을 사용하였다. 다음 <Fig. 1>은 본 연구에서 설정되어 사용된 딥러닝 모형의 구성도를 나타낸다. 먼저 입력층을 구성한 후, 상태유지 순환신경망과 드롭아웃을 3겹으로 설정하여 상태유지 스택 순환신경망으로 구성하여 더 깊게 추론하도록 설정하였다. 이후에는 Dense 레이어를 출력층으로 설정하여 1개의 예측값을 산출하도록 구성하였다.



<Fig. 1> Deep learning model structure

LSTM 모형 구성시 input_dimension값은 입력할 속성 개수로서 10으로 설정하였고, time_step은 LSTM에 입력할 자료의 길이로, 4와 12, 24로 각각 설정하였다. Epoch를 500으로 설정하여 분석한 결과는 다음 <Table 3>과 같다. 검증 데이터의 평균제곱오차는 각각 91.18, 101.28, 96.16, 평균절대오차는 6.13, 6.29, 6.43으로 산출되어 time_step값은 4로 결정되었다. 즉 4시간의 대여량을 입력한 후 다음 1시간의 대여량을 예측하는 모형의 오차가 가장 낮은 것으로 나타났다.

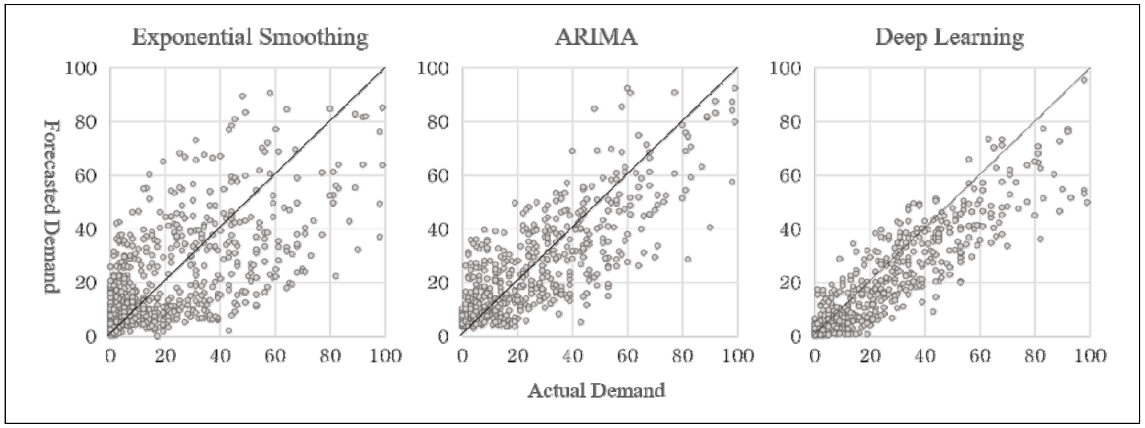
<Table 3> Results from Deep Learning Models

Index	Time_step		
	4	12	24
MSE	91.18	101.28	96.16
MAE	6.13	6.29	6.43

4. 평가 결과

최종적으로 선정된 세가지 모형을 활용하여 2019년 5월 동안의 1시간 단위 대여량을 예측하고, 산출된 예측 대여량과 실제 대여량에서 평균제곱오차와 평균절대오차를 계산하여 각 모형을 비교 평가하였다. 지수평활 모형으로 예측한 결과, MSE는 348.74, MAE는 14.15가 산출되었다. ARIMA 모형으로 예측한 결과 MSE는 170.10, MAE는 9.30이 산출되었다. 딥러닝 모형으로 예측한 결과, MSE는 120.22, MAE는 6.76이 산출되었다. 지수평활 모형의 결과값과 비교하여 ARIMA 모형의 MSE는 약 51%, MAE는 약 34% 감소하였다. 반면 딥러닝 모형의 MSE는 약 66%, MAE는 약 52% 감소하여 딥러닝 모형의 오차가 가장 적은 것으로 파악되었다.

이러한 오차 결과를 시각적으로 표현하기 위하여 x축이 실제 대여량, y축이 예측 대여량인 산점도를 <Fig. 2>에 제시하였다. 산점도에는 45도로 기울여진 직선이 그려져 있는데, 해당 직선과 점들 간의 거리가 짧을수록 예측 대여량이 실제 대여량과 유사하다는 것을 의미한다. 지수평활 모형은 점들의 분포가 가장 넓게 되어 있으며 직선과의 거리가 멀다는 것을 확인할 수 있다. 반면 딥러닝 모형은 점들의 분포가 비교적 좁으며 직선과의 거리가 짧게 나타나 딥러닝 모형의 오차가 가장 적은 것을 확인할 수 있다. 전체적으로 한 시간 단위 Rental Demand의 평균이 21.55로 나타났고, 따라서 각 모형별로 하위값에 해당하는 경우의 빈도가 높게 나타난 것으로 판단된다. 그리고 각 모형별 결과 차이는 지수평활 모형이나 ARIMA 모형의 경우 타 변수의 활용이 없는 자기 변수 중심의 시계열 모형인데 반하여, 딥러닝 모형은 자기 변수 이외에 추가적인 변수를 활용하여 구성된 모형이므로 오차가 상대적으로 적게 나타난 것으로 생각될 수 있다.



<Fig. 2> Scatterplot of actual demand and forecasted demand

V. 결 론

본 연구는 공공자전거의 대여량을 예측하는 딥러닝 모형을 개발하여 평가하였다. 이를 위하여 공공자전거 대여량 자료와 이와 상관관계가 높은 변수인 기상 자료, 지하철 이용량 자료를 수집하였다. 자료는 서울 열린 데이터 광장과 기상 자료 개방 포털에서 수집한 대여량 자료와 기상 자료, 지하철 이용량 자료 등으로 약 22만개의 데이터를 포함한 데이터 셋을 구성하였다. 이를 가공하여 지수평활 모형, ARIMA 모형과 LSTM기반의 딥러닝 모형을 각각 구축한 후 MSE와 MAE 두가지 평가 지표를 사용하여 예측 오차를 비교· 평가하였다.

평가 결과, 지수평활 모형으로부터 MSE는 348.74, MAE는 14.15가 산출되었다. ARIMA 모형으로부터 MSE는 170.10, MAE는 9.30이 산출되었다. 그리고 딥러닝 모형으로 예측한 결과 MSE는 120.22, MAE는 6.76이 산출되었다. 지수평활 모형의 결과값과 비교하여 ARIMA 모형의 MSE는 약 51%, MAE는 약 34% 감소하였다. 그리고 딥러닝 모형의 MSE는 약 66%, MAE는 약 52% 감소하여 딥러닝 모형의 오차가 가장 적은 것으로 파악되었다. 이러한 결과로부터 공공자전거 대여량 예측 분야에서 딥러닝 모형의 적용시 예측 오차를 크게 감소시킬 수 있을 것으로 판단된다.

본 연구는 공공자전거 대여량 예측에 딥러닝 모형을 활용하였고, 다양한 변수를 적용하였다는 의의를 갖는다. 또한 딥러닝 모형의 활용이 공공자전거 대여량 예측의 정확도를 높일 수 있음을 확인하였다. 향후에는 대여량과 관련성 있는 다양한 변수를 추가하여 모형 구축에 사용한다면 오차를 보다 감소시킬 것으로 예상된다. 또한 peak와 off-peak로 분석 시간을 구분하여 각각의 모형을 구축한다면 예측 오차를 보다 감소할 수 있을 것으로 생각된다.

ACKNOWLEDGEMENTS

이 논문은 2020년도 정부(경찰청)의 재원으로 도로교통공단의 지원을 받아 수행된 연구임.(POLICE-L-00002-02-202, 자율주행을 위한 AI 기반 신호제어 시스템 개발)

REFERENCES

Do M. and Noh Y.(2014), “Analysis of the Affecting Factors on the Bike-sharing Demand focused on Daejeon City,” *Journal of the Korean Society of Civil Engineers*, vol. 34, no. 5, pp.1517-1524.

Faghih-Imani A., Eluru N., El-Geneidy A. M., Rabbat M. and Haq U.(2014), “How land-use and urban form impact bicycle flows: Evidence from the bicycle-sharing system (BIXI) in Montreal,” *Journal of Transport Geography*, vol. 41, pp.306-314.

Fu R., Zhang Z. and Li L.(2016), “Using LSTM and GRU Neural Network Methods for Traffic Flow Prediction,” *31st Youth Academic Annual Conference of Chinese Association of Automation*, p.329.

Kaltenbrunner A., Meza R., Grivolla J., Codina J. and Banchs R.(2010), “Urban cycles and mobility patterns: Exploring and predicting trends in a bicycle-based public transport system,” *Pervasive and Mobile Computing*, vol. 6, no. 4, pp.455-466.

Kim D., Shin H., Park J. and Im H.(2012), “The Impact of Weather on Bicycle Usage-Focus on Usage of Bike-sharing System in Goyang,” *Journal of Transport Research*, vol. 19, no. 3, pp.77-88.

Kim Y., Kim W., Park B., Park S., Park T., Oh H., Lee S., Lee Y., Lee J., Lim Y., Jeon J. and Cho S.(1989), *Introduction to Statistics*, Yongji-Munhwa, Korea.

Lee J., Jeong G. and Shin H.(2016), “Impact Analysis of Weather Condition and Locational Characteristics on the Usage of Public Bike Sharing System,” *Journal of the Korean Society of Transportation*, vol. 34, no. 5, pp.394-408.

Lim H. and Chung K.(2019), “Development of Demand Forecasting Model for Seoul Shared Bicycle,” *Journal of the Korea Contents Association*, vol. 19, no. 1, pp.132-140.

Min J., Mun H. and Lee Y.(2017), “Demand Forecast for Public Bicycles ("Tashu") in Daejeon using Random Forest,” *Proc. of the Korea Information Science Society Congress*, p.969.

Sa K.(2019), *A Study on the Characteristics of Physical Environments Affecting the Usage of Public Bike in Seoul, Korea*, Master's Thesis, Hanyang University.

Yang H., Xie K., Ozbay K., Ma Y. and Wang Z.(2015), “Use of Deep Learning to Predict Daily Usage of Bike Sharing Systems,” *Transportation Research Record*, vol. 2672, no. 36, pp.92-102.