Journal of Knowledge Information Technology and Systems(JKITS), Vol. 14, No. 6, pp. 625~634, December 2019
https://doi.org/10.34163/jkits.2019.14.6.005

Journal of Knowledge Information Technology and Systems

ISSN 1975-7700

http://www.kkits.or.kr

A Study on Comparison of the Machine Learning Models for the Trip Distance Prediction of the Seoul Public Bike Sharing Service

Jangwoo Park*, Chang-Sun Shin

Dept. of Information and Communication, Sunchon NationalUniversity

ABSTRACT

Cities in many countries offer public bicycle sharing services to help solve the health and traffic jams of citizens and to solve environmental problems caused by automobiles. Using the machine learning models, the public bike service in Seoul have been analyzed. To service the bike sharing efficiently, the prediction of trip distance or trip duration will be needed. The prediction of trip distances and durations has important roles to help the proper operations and improvement of the bike rental services. The trip distances are deeply related to the trip durations, and could be calculated accurately when including the positions of pickup and return places, environment information such as temperature, humidity, fine dust density, et., al. To build models, linear regression, Random Forest, XGBoost and deep learning techniques have been used. Random Forest and XGboost provides important features among features. Especially, XGBoost being interested in many data manipulating and anlysing areas shows improved accuracy and can utilized the GPU to boost speed. To apply the deep learning in analysis of structured data(tabula data), embedding of categorical features will be done and the remaining continuous features and embedded features put into the fully connected neural nets. The deep learning neural net model shows the best accuracy and then XGBoost model, Random Forest models followed.

KEYWORDS: Linear regression, XGboost, Random forest, Embedding, Haversine distance

ARTICLE INFO: Received 8 October 2019, Revised 30 October 2019, Accepted 7 December 2019.

*Corresponding author is with the Dept. of Information and Communication Engineering, Sunchon National University, 255 Jungang-ro, Sunchon, Jeollanamdo, KOREA.

E-mail address: jwpark@sunchon.ac.kr

1. 서 론

자전거는 건강과 휴식을 위한 운동 기구일 뿐만 아니라 친환경 이동 수단으로 각광받고 있다. 세계적인 대도시들에서는 교통 난 해소와 환경 문제를 해결하기 위해서 공공 자전거 서비스를 시행하고 있다. 우리나라도 서울시를 비롯하여 각 지방 자체 단체들에서 여러 가지 이름의 공공 자전거 대여 서비스를 제공하고 하고 있다. 서울시의 경우 서울시 열린데이터광장[1]을 통하여 서울시에서 발생하는 각종 데이터들을 제공하고 있다. 자전거의 이용 현황[2]에 관한 데이터를 분석하면, 서울 시민의 자전거 이용 형태와 자전거 이용 목적 등을 알 수 있다. 또한 요일별, 시간 별 이용 양상은 서울 시민의 삶을 설명하는데 이용할 수 있다[3].

자전거는 출퇴근을 위해 단독으로 이용될 수도 있지만 대중 이용 교통수단과 연계하여 이용하는 경우도 있다. 공원이나 한강 둔치 등에서 레저 및 운동을 위하여 이용되는 것도 많이 있다. 이들은 자전거 이용을 요일, 시간 등으로 분석한 패턴에서 잘 나타난다. 자전거 이력 데이터를 살펴보면 많은 경우가 단거리 이용자들이지만 장거리 혹은 장시간 이용자들도 상당히 많다. 또한 자전거 대여 장소와 반납 장소가 같으며 이동거리는 긴 경우들도 있다. 자전거는 야외에서 이용하는 기구이기 때문에 기온, 습도, 미세 먼지 등 환경적인 영향을 많이 받는다.

자전거 이용 형태 분석 및 예측은 자전거 수요를 미리 알고 사전에 대처하는데 의미가 있다. Kaggle[4]에서 기상 상태를 요소로 포함하여 자전거 수요 예측[5]을 한 경우가 있다. 또한, 뉴욕 택시의 이동 시간[6]과 이용 요금을 여러 조건에서 예측하는 것들도 있다[7]. 이와 같은 예측은 실제 환경에서 발생할 수 있는 일들을 미리 알고 대처하는데 도움이 된다. 서울시 자전거 이용 형태는 앞서 설명했듯이 요일과 시간에 따라서 상이하다. 또한 이동 거리를 설명하는데 이용 시간이 중요하기는 하지만 이용거리가 비례하는 관계에 있지 않다. 물론 이용 시간을 이동 거리로만 설명 못하는 것도 마찬 가지이다. 이는 택시 요금, 혹은 택시의 이동 시간이 이동거리에 전적으로 의존하지 않는 것과 유사하다.

논문에서는 서울시 자전거 이력 데이터를 이용하여 자전거 이동 거리를 예측하는 모델을 만든다. 이때, 고려하는 것은 자전거 대여소의 위치, 반납 장소의 위치, 온도, 습도, 미세먼지 등 환경 요인, 이용 시간, 이용 요일, 이용 월 등의 시간 정보이다. 모델에 사용한 방법은 선형 회귀, 랜덤 포레스트, XGBoost, 딥러닝 기법 등이다. 이 방법을 이용하여 데이터를 모델링하고 계산한 결과의 정확성을 비교하였다.

2 장에서는 데이터 셋을 만드는 과정과 이상 데이터 처리에 대하여 소개한다. 3장에서는 사용한 모델에 대하여 간단히 설명한다. 4장에서는 모델의 정확성을 확인하기 위한 결과들을 제시한다. 끝으로 5장에서 마무리를 짓는다.

2. 데이터 셋과 이상 데이터

자전거 이력 데이터 셋은 2018년 1년간의 자전거의 대여 및 반납에 관한 것이다. 데이터 샘플은 천만 건 이상이며, 대여 일시, 대여 장소 아이디, 반납일시, 반납장소 아이디, 운행시간, 운행거리 등을 포함하고 있다. 데이터 셋에서 이상 데이터를 제거하기 위하여 운행 시간과 이동거리가 영인 데이터 샘플들을 제거한 후 속력을 계산해 보았다. 속력에서 수용할 수 없을 정도로 매우 큰 값을 보이는 경우가 많이 있으며 속력이 50km/h 이상인 샘플들은 제거하였다.

자전거 이력 데이터 셋의 경우 자전거 대여소는

아이디와 명칭으로만 주어져 있다. 따라서 자전거의 실제 위치를 위도와 경도를 이용해서 표시하는 것이 필요하다. 데이터 구성에 필요한 자전거 대여소 정보는 별도의 데이터 셋에 주어져 있다. 자전거 대여소 정보 데이터 셋은 자전거 대여소 아이디, 자전거 대여소가 있는 구의 명칭, 자전거 대여소의 위도와 경도를 제공하고 있다. 자전거 대여소 데이터 셋으로 부터 확인한 2019년 5월 현재 자전거 대여소의 개수 1523개이고 강남구가 100개, 송파구가 98개 순으로 많았다.

자전거 대여 장소와 반납 장소의 위도, 경도를 이용하여 대여 장소와 반납 장소 사이의 거리를 계산하였다. 두 장소 사이의 거리는 지구의 곡률을 고려한 Haversine 공식[8]을 이용하였다. 실제로 서울시의 면적이 크지 않기 때문에 Haversine 거리와 유클리드 거리는 거의 비슷하다. 계산한 거리는 두 지점 사이의 최단 거리이다. Haversine 거리가 영이지만 데이터 셋에 기록된 거리는 영이 아닌 것이 있다. Haversine 거리가 영이 라는 것은 대여 장소와 반납 장소가 같다는 것이다. 또한 실제 기록 거리가 Haversine 거리보다 작은 경우도 있는데 이러한 것은 실제로 존재할 수 없기 때문에 실제 모델링에서는 제거하였다.

이상 데이터를 제거한 후 이동 거리의 분포를 <그림 1>에 나타내었다. 이동 거리는 로그를 취하여 표시하였다. 이동 시간 데이터를 그려보면 정규분포에서 많이 벗어나 있다. 일반적으로는 평균 시간이내의 짧은 거리 이용이 대부분이지만 30분에서 1시간 사이의 샘플들도 많다.

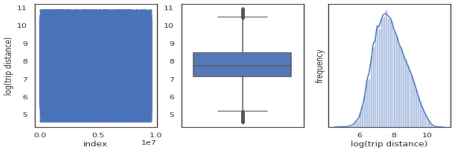


그림 1. 이동거리의 분포
Figure 1. The distribution of log riding distances

자전거 타기는 옥외 활동이기 때문에 기상 환경이 자전거 운행에 영향을 미칠 것이다. 따라서 2018년 서울시의 기상 정보와 미세 먼지 정보를 자전거 이력 데이터 셋에 포함시켰다. 기상 데이터는 매 시간별 여러 기상 정보를 포함하고 있다. 이들 중 실제 분석에 사용한 것은 '일시', '기온(°C)', '강수량(mm)', '풍속(m/s)', '습도(%)', '일조(hr)', '일사(MJ/m^2)', '적설(cm)', '지면온도(°C)', '30cm 지중온도(°C)' 등이다. 데이터 셋에는 약간의 결측치가 포함되어 있다. 온도와 습도 같은 양의 결측치는 직전 시간 측정값으로 처리하였으며 강수량, 강설량, 일조, 일사 등은 영으로 처리하였다.

기온, 습도, 미세 먼지가 자전거 이용 거리에 영향을 미치는지 알아보기 위하여 이들 데이터를 이용하여 Kmeans 클러스터링을 하였다. 기온, 습도, 미세 먼지가 자전거 이용 거리 다섯 개의 클러스터로 구분하였다. 온도와 습도에 따라서 클러스터들이 분리되는 것을 확인할 수 있었다. 온도가 영도에서 30도 사이이고, 습도가 60이하이며 미세 먼지가 70 이하, 이동거리가 10km인 경우의 클러스터가 가장 많다. 온도가 높으며, 습도가 높고 미세 먼지가 낮으며 이동 거리가 10km 이하인 경우가 그 다음이다. 온도의 낮으며 이동거리가 10km 이하인 경우가 세 번째를 차지한다. 온도와 습도가 활동하기에 적절한 경우 미세 먼지가 높은 경우에도 자전거의 이용이 상당히 많지만 다른 클러스터들에 비하여 많지 않다. 온도, 습도, 미세 먼지 등은 자전거 이용 패턴에 영향을 미치고 있으며, 자전거 이용자들이 미세 먼지에 영향을 받는다는 것을 알 수 있었다.

3. 사용한 머신러닝 모델

연구에 사용한 모델은 선형 회귀, 랜덤 포레스트, XGboost, 임베딩을 이용한 딥러닝등이다. 선형

회귀는 가장 기본적인 모델로 타 모델과 비교를 위하여 사용되었다.

1) 랜덤 포레스트[9, 10]

랜덤 포레스트는 분류, 회귀 등에 이용되는 앙상블 학습 기법으로 훈련과정에서 여러 개의 의사 학습 트리를 만들고 그 결과 들에서 다수결로 결과를 한다. 훈련 과정에서 매우 많은 노드들을 생성하며 다수결에 의한 투표 방식에 의해서 과 적합을 막는다. 트리를 생성할 때 무작위 특성은 데이터를 부트스트래핑(bootstrapping)해서 얻는다. 즉, 전체 데이터를 모두 이용하는 것이 아니라 샘플링 한 데이터를 학습에 이용함으로써 무작위 특성을 얻는다. 이 과정을 배깅(bootstrap aggregating, bagging)이라고 한다.

2) XGBoost

XGBoost(Extreme Gradient Boosting) [11]는 구조화된 데이터를 처리하기 위한 기계학습 혹은 Kaggle 경합에서 최근 주요한 학습 기법이다. XGBoost와 GBM(Gradient Boosting Machine)은 약한 분류기를 확장하는 원리를 이용하는 앙상블 트리 방법이다. xgboost는 약한 분류기를 선택할 때 “greedy” 알고리듬을 사용하고 분산 처리를 이용하여 적합한 가중치를 찾는다. XGBoost는 시스템 최적화와 알고리즘 강화를 통하여 기본 GBM 프레임워크를 개선하였다[12]. XGBoost는 효율적이고 유연하며, 이식성이 좋은 최적 분산 그래디언트 부스팅 라이브러리이다. XGBoost는 빠르고 정확하게 문제를 해결할 수 있는 병렬 트리 부스팅을 제공하고 있다. 또한 같은 코드가 여러 주요한 분산 환경에서 돌아갈 수 있으며 수십억 샘플이상의 문제도 해결할 수 있다.

3) Deep Learning

딥러닝은 이미지와 자연어 처리 등에 주로 적용하는 방법으로 알려져 있다. 그러나 최근 구글, Pinterest, Instacart 등에서 범주형 데이터의 임베딩을 만든 후 딥러닝을 구조화된 데이터(tabular data)에 딥러닝을 적용하였다.

임베딩은 이산 데이터를 연속 실수의 벡터로 표현하는 방법으로 자연어 처리에서 단어 임베딩으로 알려져 있으며, 범주형 변수의 개체 임베딩[13, 14, 15]에 응용되었다. 딥 러닝의 관점에서 임베딩은 이산 데이터를 낮은 차원의 학습된 실수 벡터 표현으로 바꾸는 것이다. 임베딩은 범주 형 데이터의 차원을 낮추고 변환된 공간에서 범주 형 샘플들을 의미 있게 표현할 수 있다. 딥 러닝에 임베딩을 사용하는 목적은 세 가지이다. 첫째, 임베딩 공간에서 가장 가까운 이웃을 찾는다. 이것은 사용자의 관심 혹은 사건의 분류에 기반을 둔 추천 시스템을 만드는데 사용될 수 있다. 둘째로는 지도학습 문제의 기계학습 모델의 입력으로 사용된다. 셋째, 범주들 사이의 관계와 개념의 시각화를 위해 사용될 수 있다.

일반적으로 기계학습에서 범주 형 변수를 처리하는 방법은 “one-hot encoding”이다. One-hot 인코딩은 각 범주를 다른 벡터에 대응시키는 매우 간단한 방법이다. 즉, 하나의 범주형 샘플을 모든 원소가 영이고 이 특별한 샘플이 속하는 범주에 해당하는 값만 1을 갖는 벡터를 만드는 방법이다. one-hot 인코딩은 두 가지 단점이 있다. 첫째, 범주의 종류가 많은(cardinality가 높은) 변수의 경우 변환된 벡터의 차원이 매우 커지게 된다. 둘째, 변환된 결과에서 어떠한 정보도 얻을 수 없다. 유사한 범주의 변수들도 임베딩 공간에서 서로 이웃하지 않는다.

자전거 이력 데이터 셋에서 자전거 대여소는

1500개 이상 있고 각각은 고유 인덱스를 갖고 있다. 대여소 인덱스는 대여소의 위치 정보를 포함하고 있다고 할 수 있다. 자전거 대여소 인덱스를 one-hot 인코딩하려면 1500 차원의 벡터가 필요하다. 그러나 이렇게 인코딩한 벡터는 서로 완전히 무관하여 cosine 유사도를 구하면 영이 된다. 또한 인코딩한 벡터는 학습되지 않는다.

대여소 아이디는 각 대여소의 위치와 연관되어 있고 지리적 특성과 이용자들의 특성을 고려하는데 중요할 것이다. 토요일과 일요일은 유사한 특징을 보일 것이고 금요일은 주중의 평균적인 특성을 갖을 것이다. 본 연구에서는 대여소의 아이디 뿐만 아니라 시계열 정보인 월, 일, 요일, 시간, 분 등에도 임베딩을 적용하였다.

전체적인 딥러닝 모델의 구조는 <그림 2>와 같다. 우선 시계열 데이터와 자전거 대여소 아이디의 임베딩을 구한다. 이렇게 임베딩한 것들과 나머지 연속적인 데이터는 함께 완전 연결 네트워크의 입력이 된다. 여러 단계의 완전 연결 네트워크로 구성되며, 활성함수(비선형 함수)로는 ReLU (Rectified Linear Unit)를 사용한다. 과 적합을 막기 위하여 Dropout을 사용하였다.

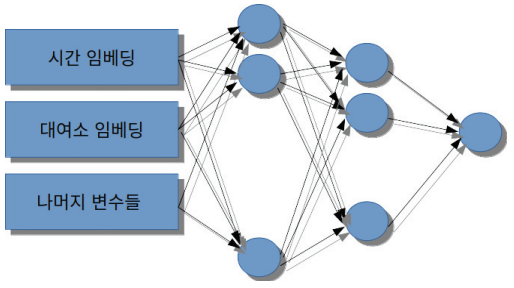


그림 2. 딥러닝 모델
Figure 2. Deep learning model

4. 결과

자전거 이력 데이터셋은, 자전거의 대여와 반납 장소에 관한 정보인 반납장소와 대여 장소의 아이디, 명칭이 있으며, 대여 및 반납 시각, 이동거리, 이용 시간 등이 기본적으로 기록되어 있다. 2절에서 설명하였듯이 자전거 대여소의 위도와 경도를 자전거 대여소 정보를 이용하여 추가 하였다. 또한, 매 시각의 온도, 습도, 미세먼지 등의 기상 환경 데이터를 포함시켰다.

랜덤 포레스트와 XGBoost 해석을 위하여 자전거 데이터 셋의 시계열 분석을 위하여 대여시각과 반납 시각은 월, 일, 요일, 시간, 분 등으로 분리하였고 이들 값들은 범주 형 데이터로 처리하여 one-hot 인코딩하였다. 자전거 데이터 셋에서 대여소의 아이디와 이름은 같은 정보이다. 자전거 대여소의 수는 1500개 이상이기 때문에 이를 one-hot 인코딩하면 cardinality가 너무 커서 자전거 대여소의 위도 경도 정보를 이용하였다. 우선 자전거 대여소의 위도, 경도와 반납소의 위도 경도를 KMeans 클러스터링 기법을 이용하여 20개의 그룹으로 그룹화 하였다. <그림 3>은 KMeans 클러스터링 결과이다. 그리고 각 클러스터 인덱스를 one-hot 인코딩하였다.

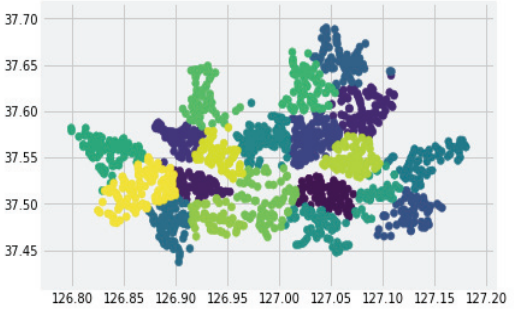


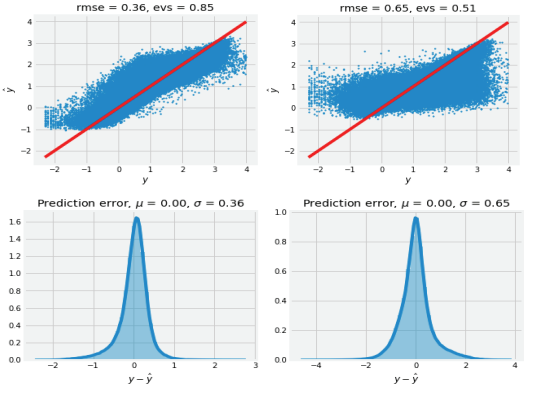
그림 3. 자전거의 대여소와 반납 장소의 KMeans 클러스터링
Figure 3. KMeans clustering results of bike rental and return locations

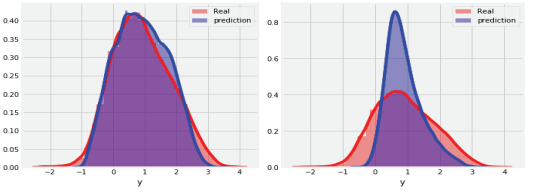
1) 선형 회귀

파이썬의 대표적인 기계학습 패키지인 scikit-learn을 사용하였다. 우선, 데이터를 one-hot 인코딩하고 정규화 한 후에 단순 선형 회귀를 적용하여 로그-이동 시간을 예측하였다. 독립 변수로 이동 시간을 포함하는 경우와 그렇지 않은 경우를 고려하였다. 이동시간은 이동거리와 밀접한 관계가 있기 때문에 이동시간을 고려하면 그렇지 않은 경우보다 RMSLE도 줄어들고 explained- variance score 증가한다. 공간 문제로 결과는 생략한다.

2) 랜덤 포레스트

선형 회귀와 같은 데이터 셋을 이용하여 랜덤 포레스트를 적용하였다. 랜덤 포레스는 조정해주어야 하는 하이퍼 파라미터가 많다. 따라서 그리드 서치를 통하여 하이퍼 파라미터를 구한 후 약간의 조정을 하였다. <그림 4>는 랜덤 포레스트를 이용하여 구한 모델의 결과이다. 그림 4(a)은 이동 시간를 모델을 만들 때 고려한 것이고 그림 4(b)는 고려하지 않은 것이다. 이동 시간과 이동거리는 선형적인 관계를 갖는 것은 아니지만 그림4에서 보듯이 이동시간은 이동거리를 예측하는데 가장 중요한 독립 변수임을 확인할 수 있다.



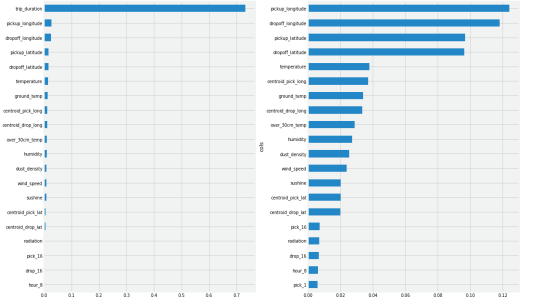


(a) (b)

그림 4. 랜덤 포레스트를 이용하여 구한 예측 결과(이동거리 포함(a), 이동거리 미포함(b))

Figure 4. The prediction and ground truth comparisons

<그림 5>은 결과에 기여하는 중요한 독립 변수들을 그림으로 표시하였다. 당연히 이동 거리가 모델에 포함되면 이것이 이동 거리를 계산하는 데 있어서 가장 중요하다는 것을 알 수 있다. 이동 거리 다음에는 대여소와 반납 장소의 경도, 위도 등이 중요한 요소이다. 위도와 경도는 거리를 예측하는 중요한 지표이다. 위치 데이터 다음에 중요한 것은 온도, 습도, 미세 먼지 등이다.



(a) (b)

그림 5. 중요한 독립 변수 (a) 이동 시간 포함 (b) 이동 시간 포함하지 않음

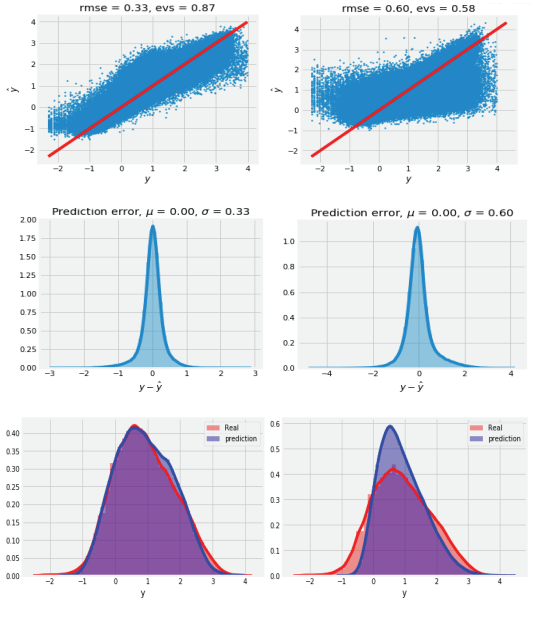
Figure 5. Feature importance (a) with riding duration (b) without riding duration

3) XGBoost

랜덤 포레스트를 이용하면 선형 회귀 모델에 비하여 중요한 개선을 얻을 수 있다. XGBoost는 최근 Kaggle 등을 포함한 여러 분야에서 사용되고 있으며, GPU를 사용하여 연산 시간을 단축할 수 있

다. XGBoost도 랜덤 포레스와 마찬가지로 많은 하이퍼 파라미터가 있다. 이들 하이퍼 파라미터들을 hyperopt 패키지를 이용하여 최적화 하였다. <그림 6>은 XGBoost를 이용하여 구한 모델의 결과 이다. 이동 시간을 고려한 경우와 그렇지 않은 경우 모두 랜덤 포레스트를 능가하는 결과를 얻고 있다. 실제 데이터와 예측 결과의 분포를 비교한 세 번째 그림을 보면 이동 시간을 고려한 경우 두 분포가 거의 비슷하다. 이동 시간을 고려하지 않는 경우에는 짧은 이동 거리 혹은 긴 이동 거리의 예측이 어렵다.

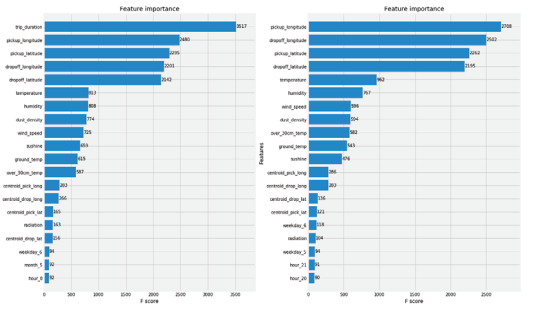
<그림 7>에서는 XGBoost를 이용하여 구한 독립변수의 중요도이다. 이용 시간이 모델에 구현에서 가장 중요하다. 이용 시간 다음에는 대여소와 반납 장소의 위치가 당연히 중요하다. 또한, 이러한 위치 정보 다음에는 온도, 습도, 미세 먼지 등이 자전거 이용 거리 예측에 영향을 미치고 있다.



(a) (b)

그림 6. XGBoost 결과 (a) 이동 시간 포함 (b) 이동시간 불포함

Figure 6. Results from XGBoost



(a) (b)

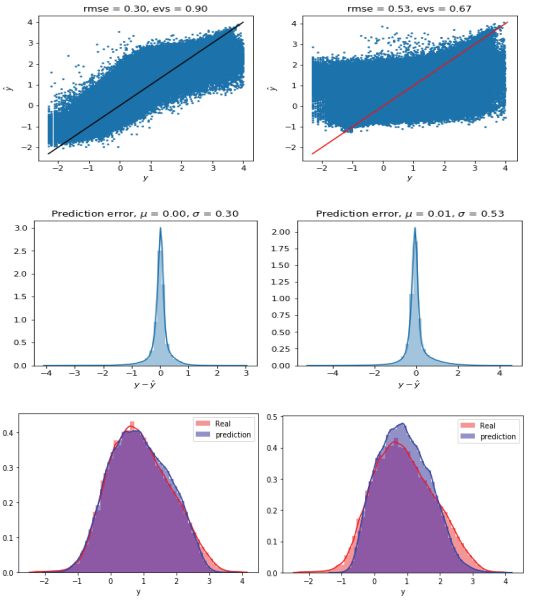
그림 7. XGBoost 모델의 중요한 독립변수. (a)이동거리 포함, (b)이동거리 미포함

Figure 7. feature importance of XGBoost

3) 딥러닝

딥러닝의 데이터 셋에서 대여소의 아이디를 사용한다. 대여소의 아이디, 월, 일, 시간, 분 등을 임베딩하고 이를 완전 연결 신경망에 입력하였다. 완전 연결 신경망의 넓이와 깊이는 hyperparameter 이다. 활성층으로는 ReLU를 이용하였고 regularization 방법은 dropout을 이용한다. 딥 러닝 프레임워크는 Pytorch를 기반으로 하는 Fastai[15]를 사용하였다. Fastai는 빠르고 사용하기 편리한 프레임워크이다.

자전거 대여소의 아이디와 위도 경도는 같은 정보를 표시하기 때문에 여기에서는 대여소의 아이디만 포함시킨다. 실제 위도와 경도를 포함하여도 결과는 거의 차이가 없다. <그림 8>의 결과는 자전거 대여소의 아이디와 위도 경도를 같이 포함하여 계산한 결과이다. 실제 자전거 아이디와 위도, 경도는 같은 정보를 포함하고 있다. 또한, 이용 거리를 구하기 위하여 이용시간을 포함하였다. 이용거리가 이용 시간에 완전히 비례하는 것은 아니지만, 이용시간은 이용거리를 구하는데 있어서 매우 중요한 역할을 한다. <그림 8>에서 보듯이 딥러닝 모델은 앞에서 설명한 랜덤 포레스트나 XGBoost보다 좋은 결과를 얻었다.



(a) (b)

그림 8. 딥러닝 결과. (a) 이동거리 포함, (b) 이동거리 미포함
Figure 8. The results from deep learning

표 1. 모델들의 비교
Table 1. Comparison among models

	RMSLE		Explained-variance score	
	With trip duration	Without trip duration	With trip duration	Without trip duration
Linear Regression	0.61	0.89	0.57	0.07
Random Forest	0.36	0.65	0.85	0.51
XGBoost	0.33	0.60	0.87	0.58
Deep Learning	0.30	0.53	0.90	0.67

<표 1>에 선형 회귀, 앤덤 포레스트, XGBoost, 딥러닝 모델의 결과를 비교하였다. 모델에서 알 수 있듯이 선형 회귀를 제외한 나머지 세 모델의 결과는 거의 비슷하지만 딥러닝의 결과가 가장 좋다.

5. 결 론

논문에서는 서울시 공공데이터 중에서 자전거 이력 데이터를 이용하여 자전거 이동 거리를 예측하는 모델을 만들었다. 자전거 이용시간은 자전거 이동 거리 계산에서 중요한 요소이다. 그러나 자전거의 이동 거리는 단순히 이동시간에 의해서 결정되지 않는다. 자전거의 이동거리에 영향을 미치는 중요한 요인들은 자전거 대여소의 위치, 반납 장소의 위치 등 위치 정보가 중요하다. 또한 외부 환경인 온도, 습도, 미세 먼지등도 중요한 역할을 한다.

연구에서 사용한 모델은 선형 회기 모델, 랜덤 포레스트 모델, XGBoost 모델, 딥 러닝 모델이다. 딥 러닝 모델은 임베딩 계층을 이용하여 범주 형 데이터의 임베딩을 구하고 임베딩 결과와 나머지 연속적인 데이터를 함께 완전 연결 층에 입력하여 결과를 얻는다. 사용한 자전거 데이터와 같은 구조화된 표형식의 데이터에는 최근 들어서 사용되기 시작했다.

모델링 결과는 <표 1>에 요약하였고 선형 회귀를 제와한 랜덤 포레스트, XGBoost, 딥 러닝 결과는 거의 비슷하다. 그러나 딥 러닝 결과가 다소 좋은 값을 보였다. 현재 사용하고 있는 것과 같이 대용량의 데이터를 이용하는 경우 XGBoost와 딥러닝은 GPU에서 연산이 가능하기 때문에 하이퍼 파라미터 조절 등에서 유리하다.

References

[1] Seoul metropolitan public bike information, http://data.seoul.go.kr/, Apr. 2019.

[2] Seoul metropolitan public bike information, http://data.seoul.go.kr/dataList/datasetView.do?infId=OA-15245&srvType=F&serviceKi, Apr. 2019.

[3] Jangwoo Park, Yongyun Cho, *Anaysis for public bike utilization status of Seoul city*,

on publishing, Dec. 2019.

[4] https://www.kaggle.com/, Apr. 2019.

[5] Bike Sharing Demand, https://www.kaggle.com/c/bike-sharing-demand, Apr. 2019.

[6] New York City Taxi Trip Duration, https://www.kaggle.com/c/nyc-taxi-trip-duration/overview, Apr. 2019.

[7] New York City Taxi Fare Prediction,https://www.kaggle.com/c/new-york-city-taxi-fare-prediction, Apr. 2019.

[8] Haversine formula, https://en.wikipedia.org/wiki/Haversine_formula, Apr. 2019.

[9] Random Forest,https://en.wikipedia.org/wiki/Random_forest, Apr. 2019.

[10] Andreas Müller, Sarah Guido, Introduction to Machine Learning with Python, 2017.

[11] XGBoost, https://en.wikipedia.org/wiki/XGBoost, Apr. 2019.

[12] V. Morde, *XGBoost algorithm: long may she reign!,* https-medium-com-vishalmorde-xgboost-algorithm-long-she-may-rein-edd9f99be63d, Apr. 2019.

[13] C. Guo, and F. Berkhahn, *Entity embeddings of categorical variables,* https://arxiv.org/abs/1604.06737, 2016.

[14] W. Koehrsen, *Neural network embeddings explained,* https://towardsdatascience.com/neural-network-embeddings-explained-4d028e6f0526, 2018.

[15] fastai, https://www.fast.ai/2018/04/29/categorical-embeddings/, Apr. 2019.

서울시 공공자전거 운행 거리 예측을 위한 머신 러닝 모델의 비교 연구

박장우, 신창선

순천대학교, 정보통신 멀티미디어 공학부 교수

요 약

여러 나라의 여러 도시들에서 시민의 건강과 교통 정체를 해소하고 자동차 등으로 인한 환경 문제를 해결하기 위하여 공공 자전거 서비스를 실시하고 있다. 기계학습모델을 이용하여 서울시 공공 자전거 대여 서비스의 데이터를 분석한다. 자전거 공유 서비스를 효과적으로 운영하기 위해서는 운행 거리 및 운행 시간의 예측이 필요하다. 자전거 운행 거리와 운행 시간 예측은 자전거 대여 서비스 운영과 개선에 중요한 역할을 것이다. 자전거 운행 거리는 운행 시간과 밀접한 관계를 갖고 있으며, 자전거 대여소의 위치, 대여소와 반납 장소의 관계, 기온, 습도, 미세 먼지 등에 의해서 결정된다. 사용한 기계학습모델은 선형 회귀, 랜덤 포레스트, XGBoost, 딥러닝 기법 등이다. 랜덤 포레스와 XGBoost는 중요한 변수의 선택이 가능하여 특히 XGBoost는 다양한 데이터 분석 분야에서 즐겨 사용하는 방법으로 정밀도가 높으며, GPU(Graphic Process Units)를 사용하여 연산 속도를 빠르게 할 수 있다. 정형화된 데이터에 딥러닝을 적용하기 위해서 범주형 데이터를 임베딩 하였고 변환한 데이터와 연속 변수 값을 완전 연결 신경망에 적용하였다. 모델의 정확성은 딥러닝 모델이 가장 우수하며, 다음은 XGBoost, 랜덤 포레스트, 선형 회귀 순이었다.

Jangwoo Park received the B.S., M.S. and Ph.D. degrees in Electronic engineering from Hanyang University, Seoul, Korea in 1987, 1989 and 1993, respectively. In 1995, he joined the faculty member of the Sunchon National University, where he is currently a professor in the Department of Information & Communication engineering. His research focuses on Localization and SoC and system designs and RFID/USN technologies.

E-mail address: jwpark@sunchon.ac.kr

Changsun Shin received the Ph.D. degree in Computer Engineering at Wonkwang University. Currently, he is a professor of the Dept. of Information & Communication Engineering in Sunchon National University. His researching interests include Distributed Computing, IoT, Machine Learning, and Agriculture/ICT Convergence.

E-mail address: csshin@sunchon.ac.kr