서울시 따릉이 자전거 이용 예측

김동영(소프트웨어) , 서준영(AI), 윤다인(소프트웨어), 이승연(소프트웨어), 이주영(소프트웨어), 이효근(소프트웨어)

2021 CUAI 중앙대학교 인공지능 학회 동계 컨퍼런스
2021 Chung-Ang University Artificial Intelligence Society's Winter Conference

요약

서울시 마포구의 날짜 별, 시간 별 기상상황 데이터를 통해서 서울시 마포구의 따릉이 대여 수를 예측하는 것이 본 연구의 목적이다. 시각화로 따릉이 대여 수 추이와 상관계수가 높은 컬럼을 확인할 수 있다. 학습에 사용하기로 결정한 변수를 데이터 전처리로 결측치 처리를 수행한 후 랜덤 포레스트 모형으로 기계 학습하여 따릉이 대여 수를 예측하고 모형 별 성능을 비교하였다.

연구방법 1: EDA(탐색적 데이터 분석)

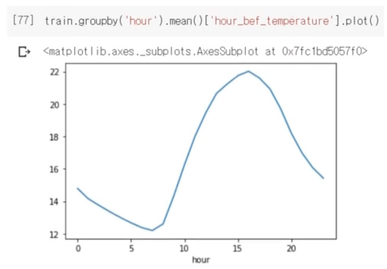
ID	날짜와 시간 별 ID
hour_bef_temperature	1시간 전 기온
hour_bef_precipitation	1시간 전 비 정보, 비가 오지 않았으면 0, 비가 오면 1
hour_bef_windspeed	1시간 전 풍속(평균)
hour_bef_humidity	1시간 전 습도
hour_bef_visibility	1시간 전 시정, 시계(특정 기상 상태에 따른 가시성을 의미)
hour_bef_ozone	1시간 전 오존
hour_bef_pm10	1시간 전 미세먼지(머리카락 굵기의 1/5에서 1/7 크기의 미세먼지)
hour_bef_pm2.5	1시간 전 미세먼지(머리카락 굵기의 1/20에서 1/30 크기의 미세먼지)
count	시간에 따른 따릉이 대여 수

<표1>따릉이 데이터 컬럼의 종류

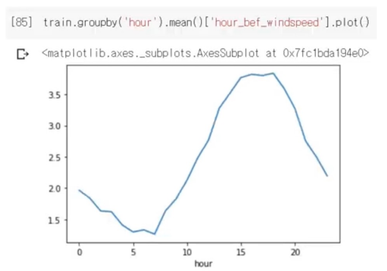
- X축을 시간, Y축을 따릉이 대여 수로 놓은 그래프를 측정했을 때 오후 6시가 가장 높고 아침 시간에 사용량이 저조한 것을 볼 수 있다. 이를 통해 출근시간에 사용량이 급감함을 간단하게 추측해 볼 수 있다.
- 학습을 위한 컬럼은 count값과 상관 관계가 높은 컬럼으로 결정하였다.
  (hour, hour_before_temperature, hour_bef_windspeed)

연구방법 2: 데이터 전처리

- hour_bef_temperature, hour_bef_windspeed 변수는 결측값(Null)을 가지고 있으며, 이는 모델링을 하기 전에 반드시 수정되어야 한다.



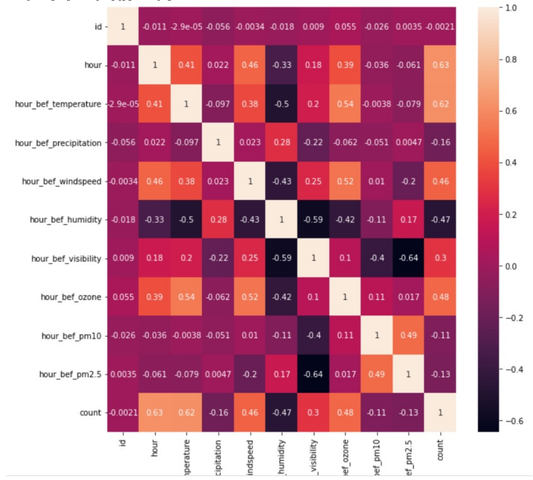
<그림1>시간에 따른 온도 그래프



<그림2>시간에 따른 풍속 그래프

- 시간에 따른 온도 변수를 관찰하면 새벽에는 낮은 온도를, 낮에는 높은 온도를 보인다. 이어서 시간에 따른 풍속 변수를 관찰하면 온도 변수와 유사하게 새벽에는 낮은 풍속을, 낮에는 높은 풍속을 보인다.

- 따라서 본 연구에서는 온도와 풍속이 시간에 따라 규칙적인 흐름을 보인다는 사실을 기반으로 결측값이 발생한 그 시간의 온도 또는 풍속의 변수를 각각 모아 평균을 산출하고, 그 평균값을 결측값에 대입하여 데이터가 시간 관계성을 보존할 수 있도록 결측치 처리를 수행하였다.



<그림3>각 컬럼 관의 상관관계 hitmap

연구방법 3: 기계학습

- 본 연구에서 사용된 기계학습 모형은 분류, 회귀 분석 등에 자주 이용되는 랜덤 포레스트를 사용하였다.
- 본 연구에서 사용된 주요 파리미터는 결정트리의 개수를 지정하는 'n_estimators', 트리의 최대 깊이를 지정하는 'max_depth'이다. 이외에 주요 파라미터는 min_samples_split, max_features 등이 있다.

<표2>와 같이 결정 트리의 개수, 트리의 최대 깊이에 의한 성능 비교를 위해 총 세 가지 모델을 구축하고자 하였다.

모델명	특징
model100	n_estimators = 100
model100_5	n_estimators = 100 max_depth = 5
model200	n_estimators = 200

<표2> 모델명과 각 모델의 특징

연구결과

본 연구에서는 평가 지표로서, RMSE(평균 제곱근 오차)를 이용하였다. RMSE 값이 낮게 나올수록 높은 성능을 보인다.

모델명	RMSE 값
model100	50.1905420613
model100_5	51.7719422024
model200	49.2887916522

<표3> 각 모델별 RMSE 값

결론

1. 성능
- Model200 49.3, Model100 50.2, Model100_5 51.8
- 세 가지 렌덤 포레스트 모델 모두 RMSE 50 정도를 기록하였음

2. 하이퍼 파라미터 변화
- n_estimators, max_depth에만 변화를 줌
- 성능이 높은 모델에 도달하기 위해 다양한 조건으로 튜닝해 볼 필요가 있음

3. 결측값
- 온도와 풍속이 시간에 따라 규칙적인 흐름을 보인다는 사실에 기반하여 그 시간의 온도 또는 풍속의 평균값을 결측값에 대입함
- 0으로 처리, 전체 평균값을 대입하는 방법 보다는 시간과의 관계성을 고려하였지만 데이터 전처리 방법을 연구할 필요가 있음