# **2021 데이터사이언스 ‘나만의 프로젝트’ 보고서**

# **Bike Sharing Demand**

|  |  |
| --- | --- |
| 제1저자 : 장희권  소속1 : 컴퓨터공학부  학번1 : 202001549  jang010505@inu.ac.kr | 제2저자 : 지은미  소속2 : 컴퓨터공학부  학번2 : 202001555  jum0624@inu.ac.kr |

**목차**

1. **서론**
2. **관련 연구**
3. **데이터**
4. **데이터 선택**
5. **결측치 업데이트**
6. **방법**
7. **실험**
8. **결론**

**참고문헌**

**요약**

캐글에 있는 Bike Sharing demand라는 대회에 대한 보고서이다. 대회는 외부 요인에 따른 자전거 대여 량을 예측하는 모델을 설계하는 것이 목표인 대회이다. 우리는 결정 트리를 기반으로 한 랜덤 포레스트로 모델링 하여 하이퍼파라미터를 찾는 것이 목표이다.

**I.서론**

이 프로젝트는 2년 간 미국 워싱턴 D.C에서 날씨 및 기온, 시간, 습도, 풍속 등의 정보를 기반으로 1시간 간격의 자전거 대여 횟수를 기록한 데이터를 기반으로 학습 모델을 모델링하여 자전거 수요를 예측하는 것이 목표이다.

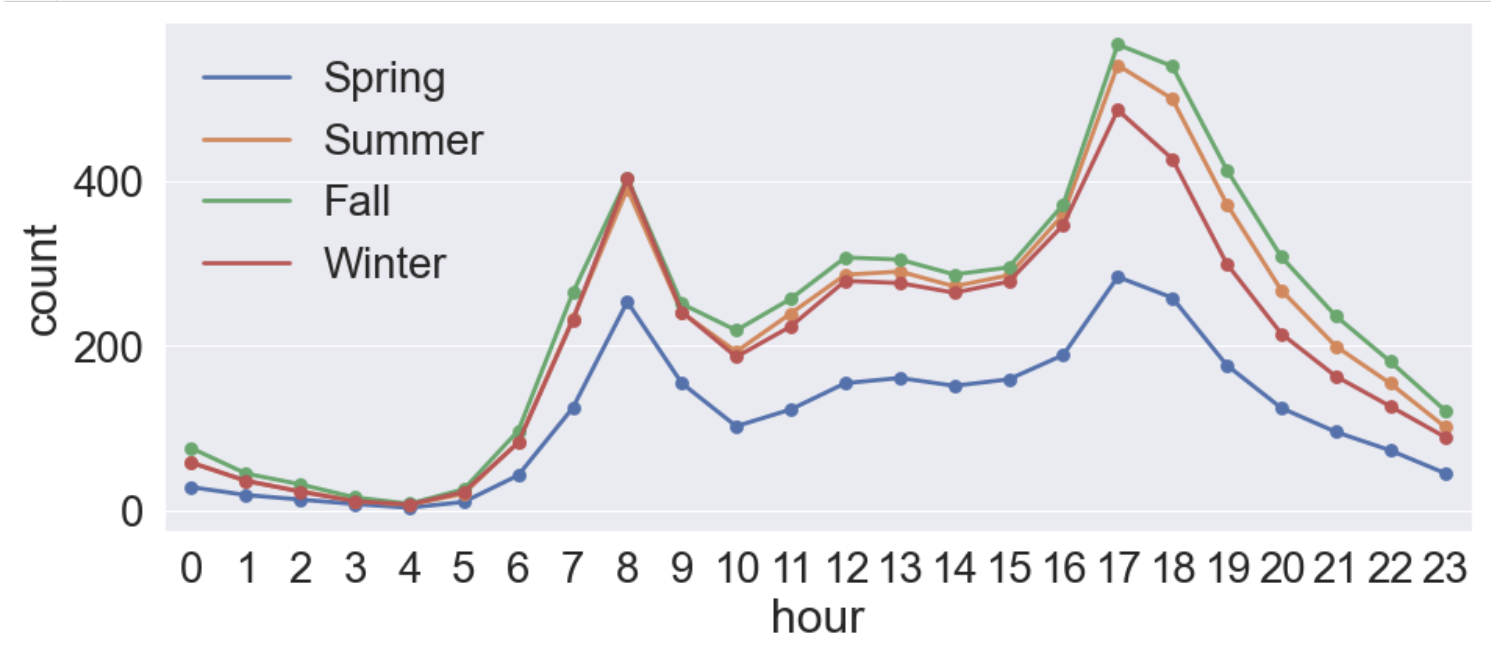
이 프로젝트를 통하여 외부 환경에 따라 자전거를 배치하여 고객의 만족도를 올릴 수 있으며 효율적으로 자전거를 관리할 수 있다.

**II. 관련연구**

결정 트리(Decide Tree)를 기반으로 한 랜덤 포레스트(random forest)를 구현하려 했으나 구현 자체가 너무 어려웠으며 또한 인터넷에 자료가 너무 희박하여 sklearn을 이용하여 모듈을 불러와 모델링 하였다.

**III. 데이터**

1. **데이터 선택**

**텍스트, 영수증이(가) 표시된 사진

자동 생성된 설명**여기서 우리는 random forest가 학습할 수 없는 데이터를 제거해야 한다. 그 중 하나가 datetime 즉 날짜 및 시간이다. 그러나 이 데이터는 학습시킬 때 빼기엔 너무 의미 있는 데이터이다. 즉 우리는 datetime을 제거하는 게 아니라 datetime에서 필요한 데이터를 추출해 random forest에 적용할 수 데이터 타입으로 바꾼다는 것이 목표이다.

시간대 별 자전거 대여량

데이터 타입

 Pandas에 있는 datetime을 이용하여 train셋과 test셋에 있는 시간정보를 추출하여 다시 넣어주었다.

train, test 데이터 셋 Datetime 시간정보 추출 및 삽입

텍스트, 영수증이(가) 표시된 사진

자동 생성된 설명

테이블이(가) 표시된 사진

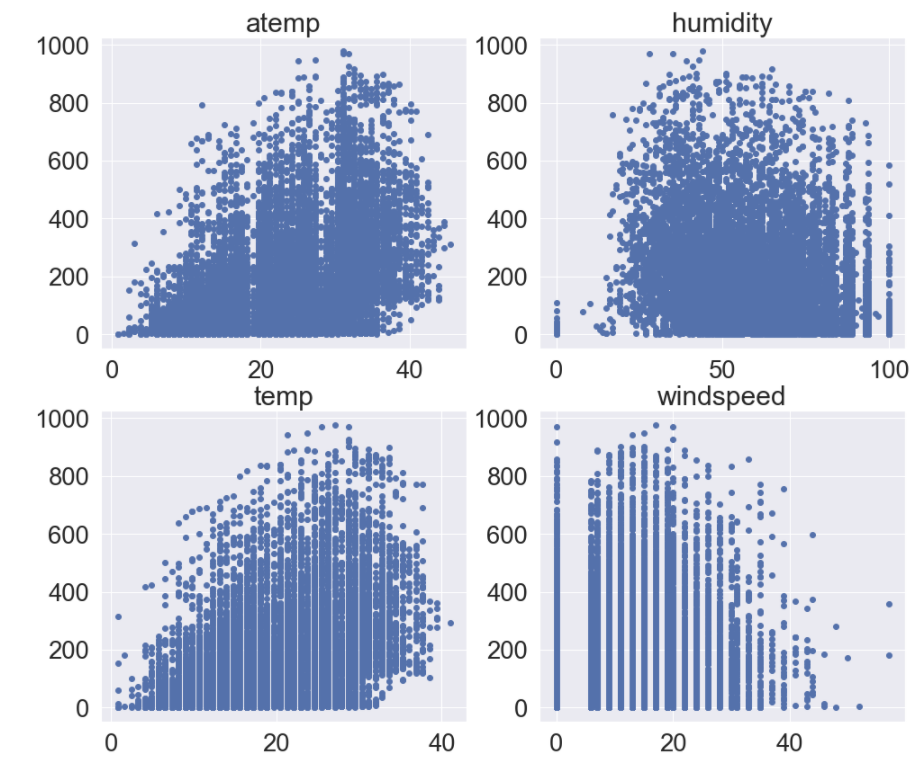
자동 생성된 설명

제거 후 train 셋

제거 전 train 셋

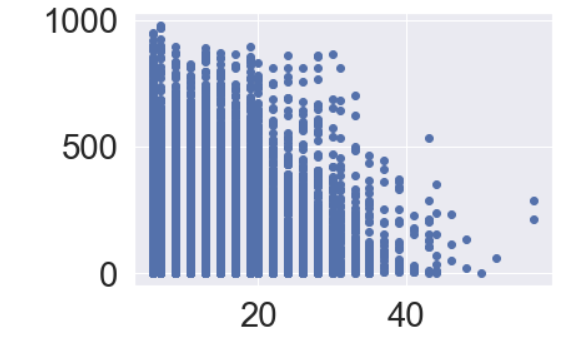
데이터 셋에서 필요한 데이터를 저장하였다. 그러면 학습에 필요한 데이터를 선택해야한다. 데이터 셋을 보면 계절, 공휴일, 업무일, 날씨, 온도, 체감온도, 습도, 풍속, 비회원 대여 량, 회원 대여 량, 총 대여 량, 시간 정보가 있다. 여기서 예측할 값인 비회원 대여 량, 회원 대여 량, 총 대여 량은 학습시킬 때 제외해야 한다. 또한 시간 정보를 따로 빼냈음으로 datetime도 제외하고 학습시켜야 한다.

**2. 결측치 업데이트**

 옆의 데이터를 보면 풍속에서 0이라는 값을 많이 가지는 것을 볼 수 있다. 이 소리는 진짜 풍속이 0인 경우가 많이 존재하는 것과 아니면 측정되지 않아 0으로 기록되는 현상이다. 그러나 이 데이터에서는 측정 되지 않아 0으로 기록되는 것이다. 이것을 수정하기 위해 2가지 방법을 있다.

결측치 확인

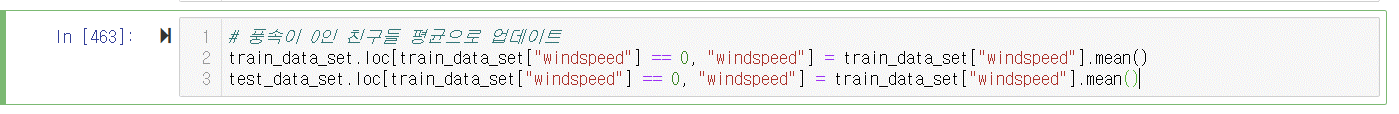
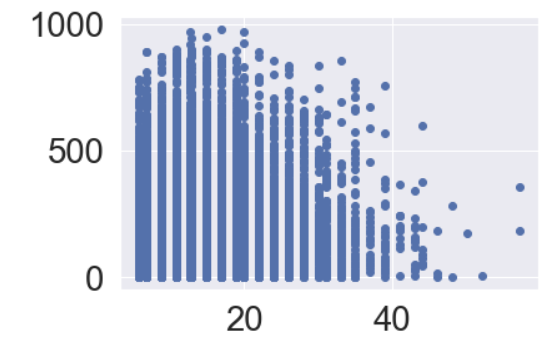
첫번째 방법은 다른 변수들로 풍속을 예측하는 방법이다. 그리고 두 번째 방법은 평균을 구해서 업데이트 하는 방법이다.

텍스트이(가) 표시된 사진

자동 생성된 설명

바뀐 데이터

풍속 예측

 다른 변수들을 이용하여 풍속을 예측하여 업데이트 해준 결과이다.

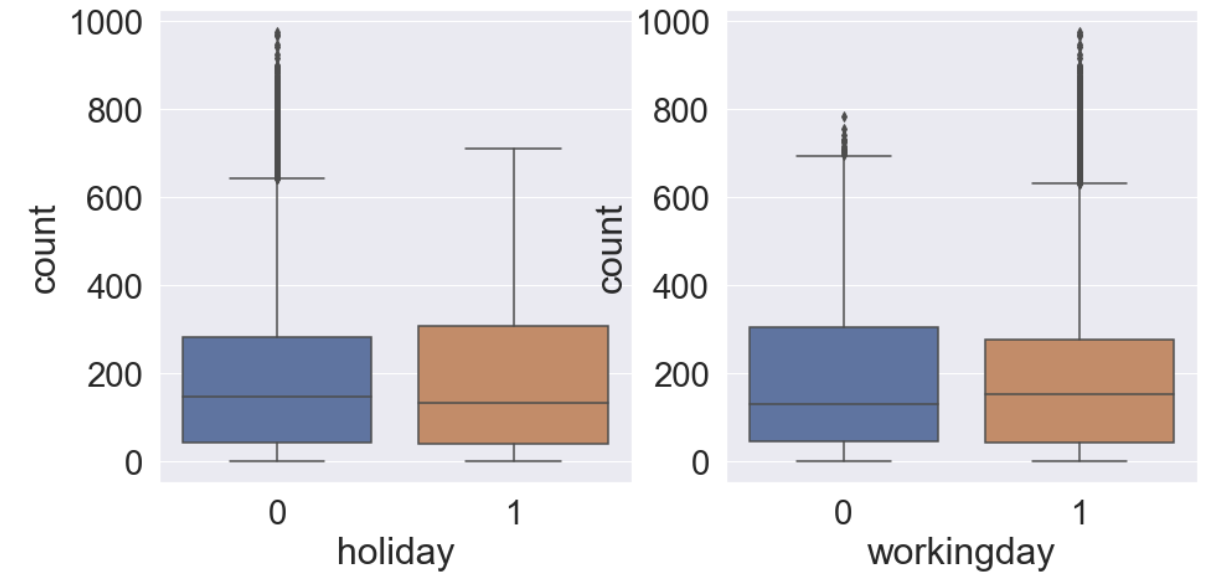
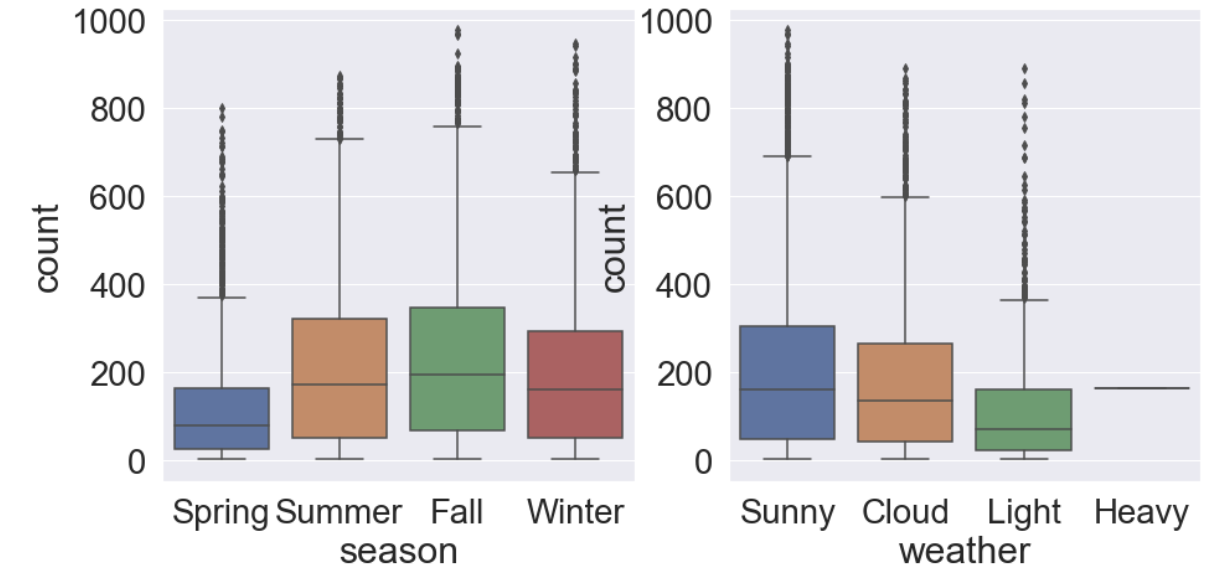
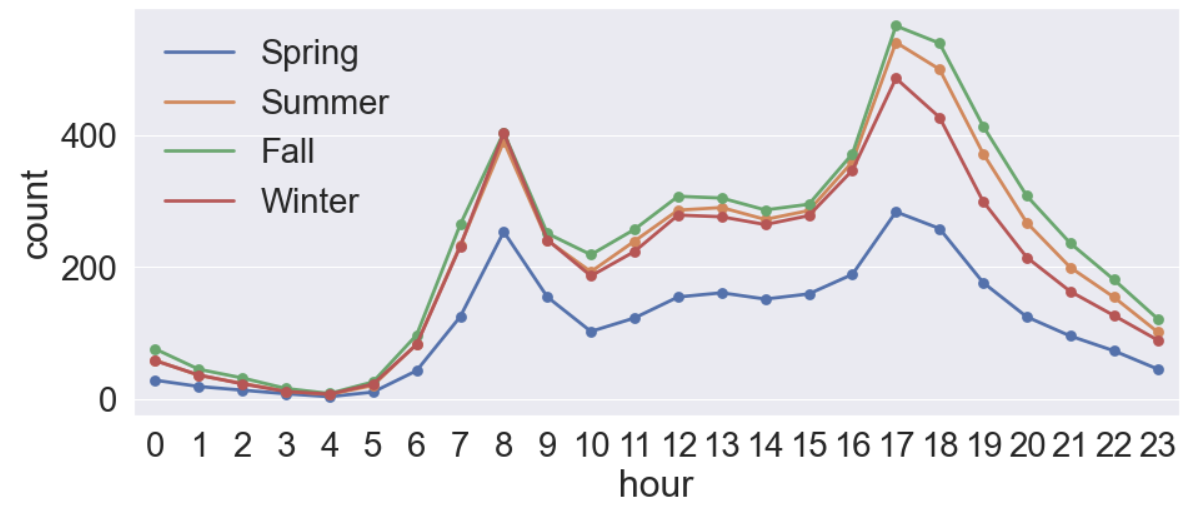
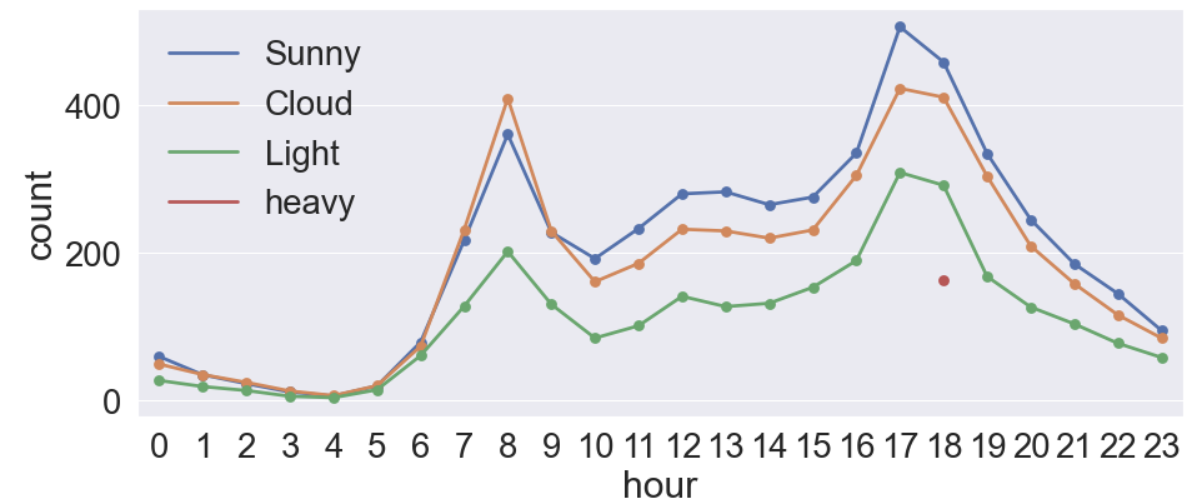
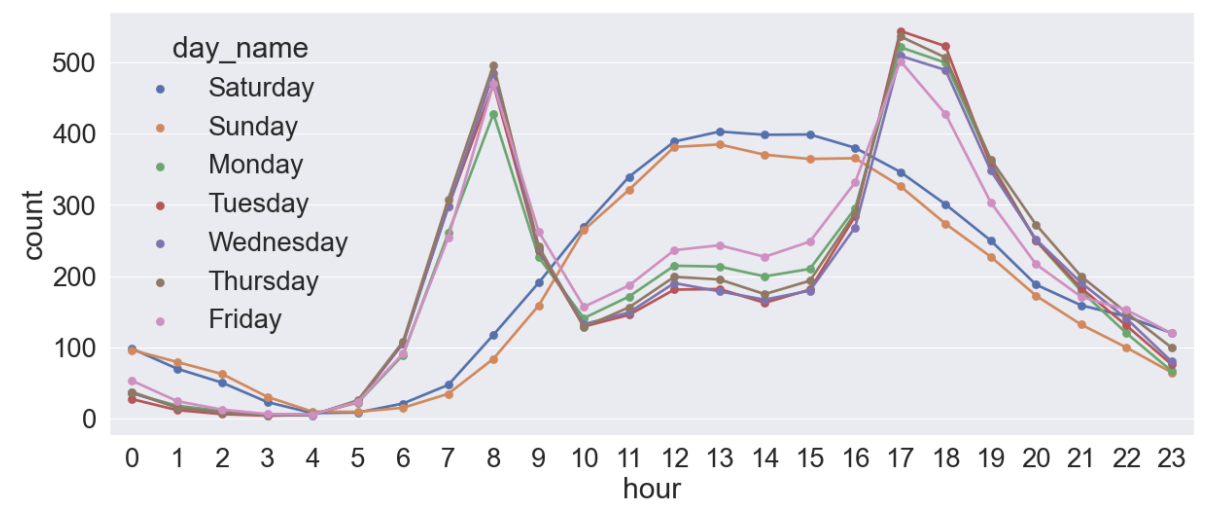
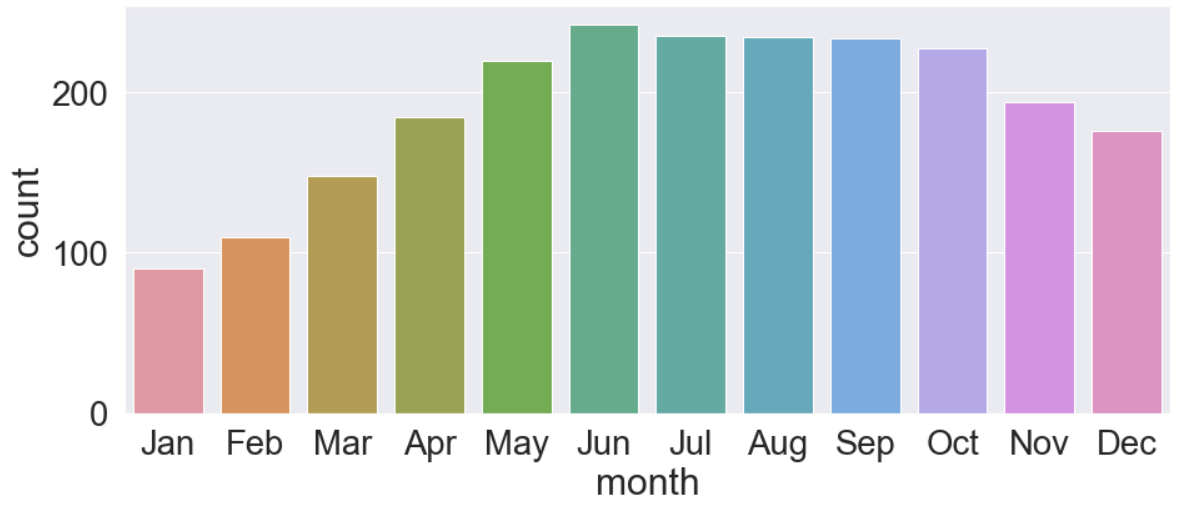
바뀐 데이터

풍속 업데이트

전체 평균을 사용하여 풍속아 0인 값들을 업데이트 해준 결과이다.

두 가지 방법 중 퍼포먼스가 더 높은 평균으로 풍속을 업데이트해주는 방법으로 모델링을 할 것이다.

**IV. 방법**



시간-요일 빈도수

시간-날씨 빈도수

시간-계절 빈도수

월 별 빈도수

계절, 날씨에 따른 빈도수

휴일, 업무일에 따른 빈도수

차트로 통해 “휴일과 출근일은 대여량에 큰 영향을 안 준다.”, “계절과 날씨는 대여량에 영향을 준다.”, “날씨가 안 좋은 날에는 대여 량이 거의 없다.”, “월은 대여량에 영향을 준다. 6월부터 10월까지 이용 고객이 많다. 즉 여름과 가을에 이용 고객이 많다.”, “계절은 대여량에 영향을 준다. 주로 출퇴근 시간에 이용을 많이 한다. 다른 계절에 비해 겨울에는 이용 고객의 수가 적다.”, “날씨는 대여량에 영향을 준다. 주로 출퇴근 시간에 이용을 많이 한다. 날씨가 매우 안 좋은 날에는 이용고객이 없다.”, “요일은 대여량에 영향을 준다. 평일에는 주로 출퇴근 시간 주말에는 주로 낮 시간대에 대여량이 많다.”

텍스트이(가) 표시된 사진

자동 생성된 설명

모델 학습 후 예측측

Random forest가 구현 자체가 어려워서 sklearn의 모듈을 불러와서 학습을 시켰따.

**V. 실험**

데이터를 전처리 할 때 풍속이 0인 데이터를 어떻게 방식으로 전처리 할 것인지를 다루었는데 풍속이 0인 데이터를 예측을 하냐, 평균으로 설정하는가를 선택을 했어야 한다. 그 중 평균으로 설저하고 돌리는 것이 더 높은 결과를 내놓았다.

**VI. 결론**

우리는 모델링을 통하여 날씨와 평일, 휴일에 따라 자전거 수요량을 예측하여 수요가 많을 거 같은 날은 자전거를 많이 배치하여 고객의 만족도와 자전거 대여의 효율성을 높일 수 있다.

**참고문헌**

1. PyTorch로 시작하는 딥러닝 입문 CAMP (2017.7~2017.12) 강의자료, 2018. <https://github.com/GunhoChoi/PyTorch-FastCampus/> [Accessed: 2020-05-05]