Bike Sharing 대본

1 장

안녕하십니까? Bike Sharing Demand에 대해 발표하게 된 김동현입니다. 발표 시작하겠습니다.

2 장

목차는 이와 같이 구성되었습니다.

3 장

주제 목적으로 Bike Sharing의 수요량 예측을 해볼 것입니다.

4 장

TrainDataSet의 속성은 총 12개, TestDataSet의 속성은 총 9개로 구성되어 있습니다.

5 장

수요량을 예측하기 위해서 저는 회귀 모델로 예측하는 모델을 만들어볼 것입니다.

6 장

결측치를 처리하기 위해 결측치를 확인하였습니다. 확인 결과 두 개의 DataSet 모두 결측치가 없는 것을 확인할 수 있었습니다.

7 장

Datetime 속성을 좀 더 세분화하여 year, month, day, hour, dayofweek 속성을 추가하였습니다.

8 장

DataSet을 보다 보니 Month 속성의 이상한 점을 확인할 수 있었습니다. 워싱턴은 우리나라와 계절의 차이가 거의 없어 3월에 봄을 시작으로 3개월씩 계절을 나누었습니다.

9 장

시간을 기준으로 수요량을 확인해 보았습니다. 계절을 보면 추운 겨울이 수요량이 적고, 날씨가 좋은 날이 수요량이 많고, 좋지 않은 날은 수요량이 거의 없는 것을 확인할 수 있었습니다. 요일별요 보았을 때 평일이 출퇴근 시간으로 수요량이 가장 많고, 주말, 10 ~ 18시까지 수요량이 많고, 월별로 보았을 때도 계절 그래프와 별 차이가 없는 것을 확인할 수 있었습니다.

10 장

다음으로 범주화할 속성들을 범주화해주었습니다. 사용하지 않을 변수로는 회원, 비회원 변수는 사용하지 않았습니다.

11 장

온도, 체감온도, 습도, 바람세기, 수요량의 이상치를 제거해 주고, 풍속 값이 0에 많은 것을 확인할 수 있었습니다. 풍속은 값이 0인게 없어 random forest를 사용하여 0인 값을 조정하였습니다. Random forest를 사용한 이유는 누락된 데이터를 추정하는 효과적이 방법을 가지고 있어 사용하게 되었습니다.

12 장

Count 속성의 큰 수를 작게 만들기 위해 log를 취해주었습니다. 선별된 속성으로는 계절, 날씨, 온도, 체감온도, 습도, 연도, 시간, 요일 별, 공휴일, 근무 여부의 속성들을 선별하였습니다.

13 장

제가 사용할 모델들입니다. Radom Forest, Linear Regression, SVR 세 가지 모델을 사용하여 수요량을 예측해볼 것입니다.

14 장

모델에서 나온 결과의 평가 방법입니다. RMSLE는 RMSE에 로그를 적용해 준 지표입니다. 아웃라이어, 이상치가 있더라도 값의 변동폭이 크지 않고, 상대적 Error, 절대적 크기가 변하더라도 상대적 크기가 동일합니다. 낮은 예측에, 예측값이 실제 값보다 작을 때 더 높은 페널티가 주어집니다.

15 장

여러 번의 예측을 하였지만 그나마 가장 잘 나온 세 번의 예측을 설명하겠습니다. 첫 번째 예측의 속성은 전에 선택한 모든 속성으로 예측해 보았습니다. Windspeed 속성과 count 속성은 처리하기 전의 속성값으로 예측을 해 보았습니다. 세 가지 모델 중 Random Forest가 RMSLE 점수가 가장 낮게 나온 것을 확인할 수 있었습니다.

16 장

두 번째 예측의 속성은 season, weather, temp, humidity, year, hour, dayofweek, holiday, workingday 들을 선택하였고, windspeed 속성과 count 속성은 값을 조정한 후에 예측한 결과입니다. 이번에도 Random Forest가 가장 낮은 것을 확인할 수 있었습니다.

17 장

세 번째 예측입니다. Season, weather, temp, humidity, hour, dayofweek, holiday, workingday의 속성들을 가지고 예측하였습니다. 이번에도 Random Forest 점수가 가장 낮은 것을 확인할 수 있었습니다.

18 장

Kaggle에서 예측점수를 확인한 결과 두 번째 시도가 가장 낮게 나왔습니다.

19 장

제가 선택한 가장 좋은 모델은 세 번째 모델입니다. 온도와 체감온도는 상관관계가 높게 나와 체감온도보단 온도를 선택하였고 월별 수요량과 계절의 수요량이 비슷하다 판단하여 계절의 속성을 선택하였습니다. 시간 때에 따라 수요가 있으므로 시간의 속성도 선택하고, 습도와, 날씨도 영향이 있다 판단하였습니다. 요일 별, 공휴일, 근무 유무 속성 또한 영향력이 있다 판단하여 세 번째 예측을 선택하게 되었습니다. 이러한 속성들 중 가장 영향력이 있는 속성은 시간으로 나왔고, 나머지는 비슷하지만 가장 높은 두 개 온도와 근무 유무로 나왔습니다.

20 장

결과입니다. 근무를 하는 사람들의 평일 출퇴근 시간에 가장 수요량이 많고, 주말은 10 ~ 18시 사이, 기온은 15도 이상이 되면 수요량이 늘기 시작하는 것을 확인할 수 있었습니다.

21 장

이상으로 발표를 마치겠습니다.