빅데이터 프로젝트 분석보고서

서울시 공공도서관 대출정보 및 기상정보를 활용한
도서관 일일 대출자 수 예측 모델 구현

과 목 명	:	빅데이터
제 출 일	:	2024.06.16
학　　과	:	컴퓨터공학전공
학　　번	:	20200818
이　　름	:	유수연

1. 분석 배경

1.1. 문제에 대한 설명과 목적

도서관은 이용자에게 도서 대출 및 반납, 예약과 상호대차 서비스 등 다양한 도서 관련 서비스를 제공한다. 또한, 이용자들이 편리하게 도서를 이용할 수 있도록 최신 도서와 다양한 분야의 자료를 지속적으로 확보하고, 방문자들이 쾌적한 환경에서 독서를 즐길 수 있도록 조용하고 편안한 열람 공간을 조성한다. 또한, 도서관 이용자들에게 풍부한 독서 경험을 제공하기 위해 프로그램을 기획하고, 원활한 운영을 위해 열람실 내에서 유동적으로 직원을 배치한다. 직원은 도서 대출 및 반납, 자료 검색, 이용 안내 등 다양한 업무를 담당하며, 방문자들의 질문에 친절하게 응대한다. 그러나 도서관 이용자 수가 많아지면 직원 인력이 부족하여 도서 관련 업무 처리, 프로그램 진행에 있어 지연이 발생할 수 있다.

본 프로젝트는 일별 기상 정보를 활용하여 도서관에서 도서 대출인원수를 예측하는 모델을 구축하고자 한다. 기상 정보는 날씨, 온도, 강수량 등의 다양한 요소를 포함하며, 이러한 요소들이 도서관 대출자 수에 어떤 영향을 미치는지 분석한다. 이를 통해 도서관은 요일별 적절한 직원 인력을 배치하여 도서관 이용자에게 원활한 서비스를 제공할 수 있도록 한다.

1.2. 문제의 중요성

도서관에서 직원 인력이 부족하면 대출 및 반납, 자료 검색 등 도서 관련 업무에 지연이 발생할 수 있다. 이는 이용자들에게 불편을 주며 도서관 이용 만족도가 떨어질 것이고, 인력이 부족한 상황에서 직원들이 과도한 업무를 처리해야 하기 때문에 업무 효율성이 떨어질 것이다. 이와 같이 도서관 이용자수와 이용자 수에 따른 직원 인력 배치는 도서 관련 업무 처리, 이용자 만족도, 직원 복지, 도서관 운영 등 여러 측면에서 영향을 미치기 때문에 도서관 이용자 수를 예측하여 대응하면 원활한 도서관 운영과 서비스를 제공할 수 있다는 점에서 이 문제는 중요하게 다뤄야 한다.

2. 데이터 처리

2.1. 사용 데이터셋

본래 데이터셋을 가공 및 통합하는 과정에서 사용된 컬럼을 첨부했다.

- 문화 빅데이터 플랫폼: 전국 공공도서관 정보(2023.5)

도서관코드	도서관명	도서관주소	1지역명	2지역명
도서관유형명				

- 문화 빅데이터 플랫폼: 전국 공공도서관 일별 대출현황(2023.01~2023.12)

대출일자	회원일련번호	도서관코드	도서관이름

전국 공공도서관 정보 데이터셋의 '1지역명', '2지역명', '도서관유형명' 컬럼을 사용하여 서울시 공공도서관 데이터를 추출했으며, 2023년 '일별 대출현황' 데이터셋 중 78개의 데이터셋에서 서울시 공공도서관 대출정보에 대한 데이터를 추출했다.

- 기상청 기상자료개방포털: 서울시 일별 기온, 강수량, 습도, 풍속 데이터셋(2023)

  1) 기온

지점명	일시	평균기온	최고기온	최저기온

  2) 강수량

지점	지점명	일시	강수량(mm)

  3) 습도

지점명	일시	최저습도

  4) 풍속

지점명	일시	평균풍속

- 서울 열린 데이터광장: 서울시 일별 평균 대기오염도 정보(2023)

측정일시	측정소명	미세먼지	초미세먼지

2.2. 데이터 가공 및 기본 변수 설명

위 5개의 데이터를 통합하여 아래와 같은 칼럼을 갖는 데이터셋을 생성했다. 파란색으로 칠해진 칸은 파생변수이다.

도서관코드	시	구	날짜	연도
월	일	요일	일일 대출인원수	미세먼지농도
초미세먼지농도	평균풍속	평균기온	최고기온	최저기온
평균습도	강수량	도서관이름		

Row: 61724, Column: 18

2.3. 파생변수 생성

데이터셋의 '날짜' 컬럼에서 연도, 월, 일, 요일을 추출하여 데이터셋에 추가하였다. 또한, 전국 공공도서관 일별 대출현황 데이터셋에서 '대출일자', '회원일련번호', '도서관코드' 칼럼을 기준으로 그룹화한 후 각 그룹의 행수를 카운트해서 각 도서관에서의 일일 대출자수를 구하여 새로운 변수 '대출인원수' 컬럼을 데이터셋에 추가했다.

2.4. 결측값 처리

가공한 데이터에서 강수량, 풍속에서 존재하는 결측값 NaN을 확인 후 0으로 처리했다. 또한, 미세먼지 농도, 초미세먼지 농도에 존재하는 결측값 NaN을 0으로 처리했다.

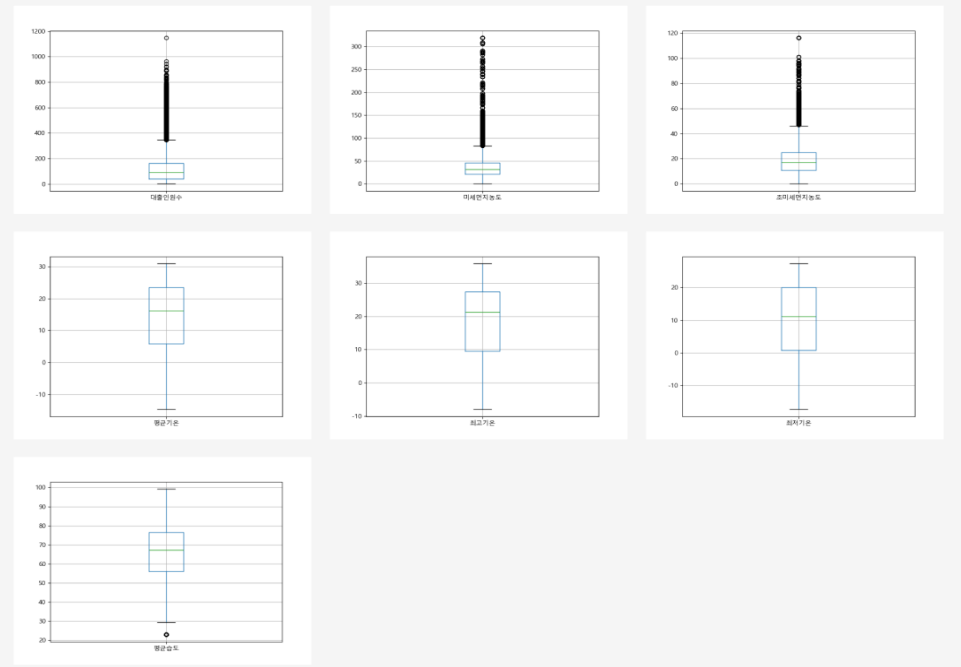
2023년의 법정 공휴일에 해당하는 날 도서관이 운영하지 않으므로 법정공휴일에 등록된 대출정보를 결측값이라고 보았으며, 해당 데이터는 필터링 후 데이터셋에서 제거했다.

```
weather_seoul_data.강수량.fillna(0, inplace=True)
weather_seoul_data.평균풍속.fillna(0, inplace=True)
air_person_seoul_data.미세먼지농도.fillna(0, inplace=True)
air_person_seoul_data.초미세먼지농도.fillna(0, inplace=True)
data.isnull().sum() # 결측값 여부 확인
```

```
도서관코드        0
월              0
일              0
요일             0
미세먼지농도        0
초미세먼지농도       0
평균풍속          0
평균기온          0
최고기온          0
최저기온          0
평균습도          0
강수량           0
대출인원수         0
dtype: int64
```

2.5. 이상치 처리

데이터에서 나타나는 이상치를 검출하기 위해 IRQ 방식을 사용했다. IRQ는 데이터 분포의 분위 중 0.25와 0.75 구간 사이에 포함되는 간격이다. 이때 0.75에 해당하는 값(Q3)과 0.25에 해당하는 값(Q1)을 사용하여 아래와 같이 이상치를 구하고 식을 통해 이상치가 포함된 행을 데이터셋에서 제거한다. 본 데이터셋에서는 미세먼지농도, 초미세먼지농도, 평균풍속, 평균기온, 최고기온, 최저기온, 평균습도 등의 기상변수와 대출인원수에 대해 이상치가 있을 것이라 예상했으며 이상치를 검출하고 제거했다.



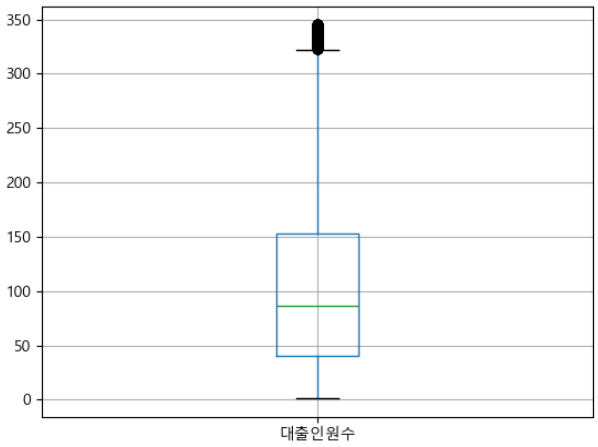
<그림1> 이상치 처리 전 칼럼별 데이터 분포, 박스 플롯

아래는 대출인원수에 대한 이상치를 구하는 코드이다.

```
data.boxplot(column='대출인원수', return_type='both')
Q1_borrow = data['대출인원수'].quantile(q=0.25)
Q3_borrow = data['대출인원수'].quantile(q=0.75)
IQR_borrow = Q3_borrow-Q1_borrow
```

IRQ값을 구하고 데이터셋에 이상치를 포함하지 않는 코드를 작성한다.

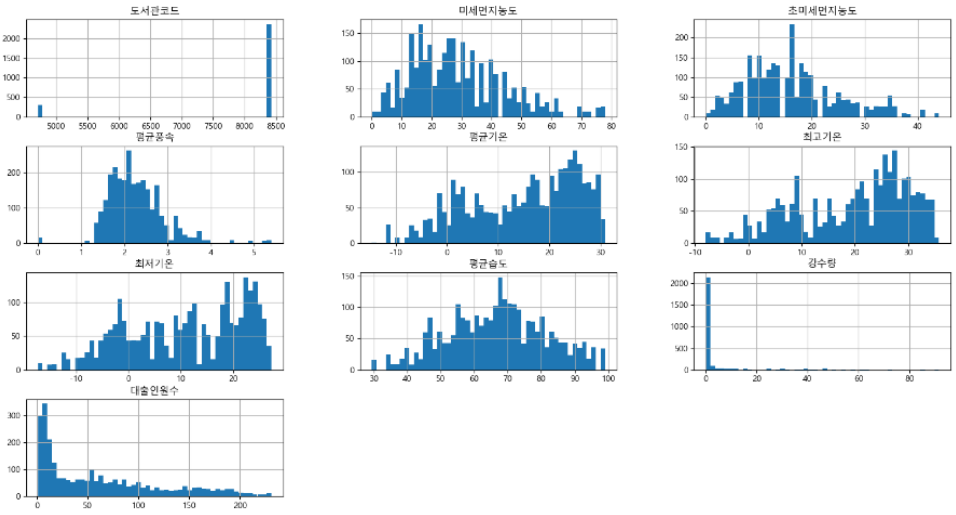
```
cond_borrow = (data['대출인원수']<Q3_borrow+IQR_borrow*1.5)
& (data['대출인원수']>Q1_borrow-IQR_borrow*1.5)
data_IQR=data[cond_borrow]
```



<그림2> 이상치 처리 후 대출인원수의 분포, 박스 플롯

2.6. 모델을 학습할 데이터 범위 축소

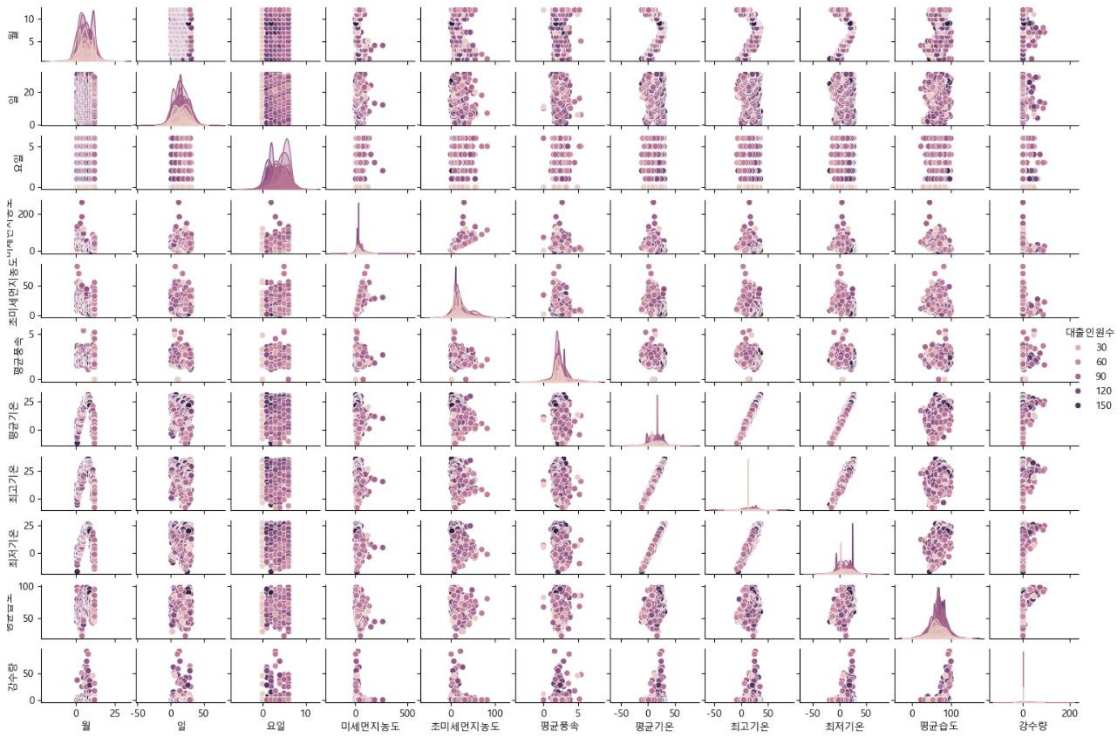
공공도서관의 대출자수는 기상요소 뿐만 아니라 도서관 주변 인구 밀도, 도서관 규모, 도서관 행사 등의 요소에 의해 그 값이 상이할 것으로 예상되어, 모델을 학습하기 위한 데이터셋의 범위를 시에서 지역구로 변경했으며, 추후에 모델의 파이프라인을 구축하고 데이터를 확보하면 모델을 학습할 데이터의 범위를 확장할 예정이다. 현재는 도봉구에 속한 공공도서관의 일일 대출내역을 모델 학습 데이터로 사용한다.



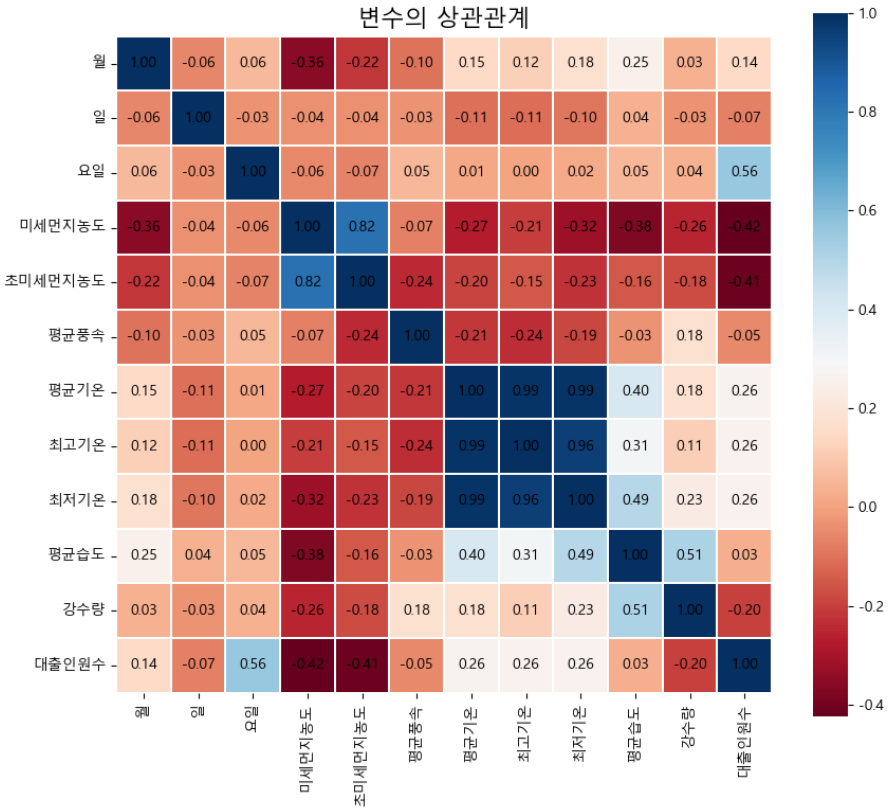
<그림3> 도봉구 일일 도서대출내역 컬럼별 데이터 분포

2.7. 컬럼 간 상관관계 분석, 모델 학습 변수 설정

corr() 메서드를 사용하여 데이터의 각 변수에 대해 상관관계를 분석했다. 이때 파라미터로 numeric_only=True를 적용하여 숫자형 타입의 변수에 대해서만 상관관계를 보이도록 했다. 아래는 상관관계 분석 결과에 대해 히트맵과 산점도를 그렸다.



<그림4> 데이터 상관관계 분석 시각화: 산점도



<그림5> 데이터 상관관계 분석 시각화: 히트맵

2.8. 데이터 분할

sklearn.model_selection에서 제공하는 train_test_split으로 데이터를 분할했다. 이때 데이터는 학습 데이터(train), 검증 데이터(val), 테스트 데이터(test)를 얻기 위해 2번의 분할 과정을 거쳐 5:2:3으로 분할했다. 데이터셋의 변수와의 상관관계를 확인 후, 학습에 도움이 될만한 변수를 선택한 데이터셋을 재구성했다.

2.9. 데이터 인코딩 및 스케일링

'요일' 칼럼을 replace() 메서드를 사용하여 숫자 범주형으로 인코딩한다. '요일' 칼럼에서 월, 화, 수, 목, 금, 토, 일을 0, 1, 2, 3, 4, 5, 6으로 인코딩한 후, 변수의 타입을 object로 변경했다.

데이터셋을 분할한 뒤, sklearn.preprocessing에서 지원하는 StandardScaler를 사용하여 데이터셋을 표준화했다. 이때, 학습 데이터셋은 fit_transform을 사용하여 StandardScaler를 학습 및 스케일링하고, 학습 데이터셋으로 학습된 StandardScaler로 transform을 사용하여 검증 데이터셋과 테스트 데이터셋을 스케일링한다.

	0	1	2	3	4	5	6	7	8
0	-1.79636	1.20596	0.0570932	-0.31411	0.337903	-1.97561	-2.21247	-1.81136	-0.299757
1	0.320144	-1.36335	0.568976	-0.877272	-0.786723	1.40112	1.40302	1.39398	-0.352642
2	0.320144	-0.312264	1.59274	-0.31411	0.000515489	1.11973	1.08822	1.04185	-0.352642
3	0.92486	0.855604	-1.47856	-0.0638163	0.000515489	0.0129125	0.162879	0.00349715	-0.352642
4	-1.494	0.154883	0.568976	1.12508	2.1373	-1.0939	-0.94371	-1.13417	-0.352642
5	-1.19165	-0.195477	-0.966673	0.749639	-0.67426	-0.549876	-0.409494	-0.700774	-0.352642
6	0.320144	1.32275	1.59274	-0.376684	-0.111947	1.03531	0.945124	1.05088	-0.352642
7	-0.58693	-1.59692	-0.966673	-0.251537	-0.561797	0.15361	0.382289	-0.177085	-0.352642
8	-0.284572	1.55632	0.0570932	-0.12639	0.225441	0.857095	0.59216	1.09602	3.86495
9	0.92486	0.27167	-0.45479	-0.12639	0.000515489	0.0598115	0.181959	-0.168056	-0.352642
10	-1.79636	1.20596	0.0570932	-0.31411	0.337903	-1.97561	-2.21247	-1.81136	-0.299757
11	-1.494	-0.779411	0.0570932	0.0613306	0.450366	-1.03763	-0.93417	-1.25155	-0.352642
12	0.320144	-0.429051	1.08086	-1.00242	-0.786723	0.772677	0.420448	1.03282	-0.253482
13	1.52958	-0.429051	-0.966673	-1.56558	-1.57396	-0.906309	-0.991408	-0.682716	-0.326199
14	0.92486	1.43954	1.08086	0.0613306	0.225441	-0.0808856	0.0961025	-0.149998	-0.352642

<그림67> StandardScaler를 사용한 데이터 스케일링

3. 알고리즘 적합 및 평가

3.1. 예측 모델 선정 및 분석 도구

예측하려는 타겟(대출자수)은 연속형 변수로 회귀 분석 기법을 채택한다. '일일 대출자 수' 변수를 예측하기 위해 회귀 분석에서 사용되는 LinearRegression, DecisionTreeRegression, RandomForestRegression을 사용한다. 3개의 모델을 학습하고 평가를 통해 좋은 평가를 보이는 모델을 예측 모델로 선정한다. LinearRegression의 경우, 교차검증 후 추출된 정확도들의 평균을 내어 평가한다. 또한, 회귀분석의 평가지표인 MAE(mean absolute error), MSE(mean squared error), RMSE(root mean squared error), R2 score를 통해 모델의 성능을 평가한다.

Python 3.10, Spyder IDE의 환경에서 코드를 실행한다. 주요 패키지는 데이터 프레임을 다루기 위한 pandas, 데이터 분석과 모델 학습, 평가를 위한 sklearn, 데이터 시각화를 위한 matplolib.pyplot과 seaborn을 사용한다.

3.2. 분석 방법

데이터의 상관관계를 분석하여 '대출인원수'와 모델에 학습시킬 변수(칼럼)을 선정한다. corr() 메서드를 사용하여 데이터셋 변수별로 상관관계를 분석한다.

학습 데이터로 모델을 학습, 검증 데이터로 모델의 하이퍼파라미터 조정 및 평가, 테스트 데이터로 모델을 예측하여 결과값을 확인한다.

1) LinearRegression

Sklearn.metrics에서 cross_score_value 메서드를 사용하여 교차검증을 5번 진행한다. 교차검증 후 MAE, MSE, RMSE, R2 Score를 구한다.

2) DecisionTreeRegression

데이터셋을 사용하여 의사결정트리 회귀 모델을 학습한다. 의사결정트리의 하이퍼 파라미터인 max_depth, max_features의 최적값을 찾기 위해 Grid search를 사용하여 모든 하이퍼파라미터의 조합을 탐색 후 적합한 하이퍼파라미터를 설정한다.

3) RandomForestRegression

데이터셋을 사용하여 랜덤포레스트 회귀 모델을 학습한다. 랜덤포레스트의 하이퍼 파라미터인 n_estimator, max_depth, features의 최적값을 찾기 위해 Grid search를 사용하여 모든 하이퍼파라미터의 조합을 탐색 후 적합한 하이퍼 파라미터를 설정한다.

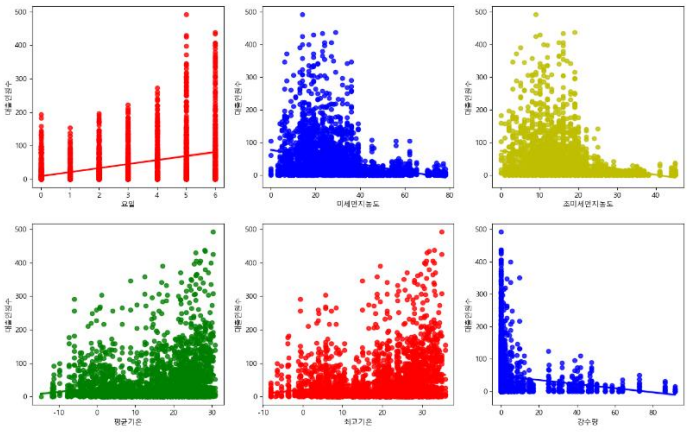
3.3. 분석 결과 및 모델 평가

3) 모델 학습 몇 평가 결과

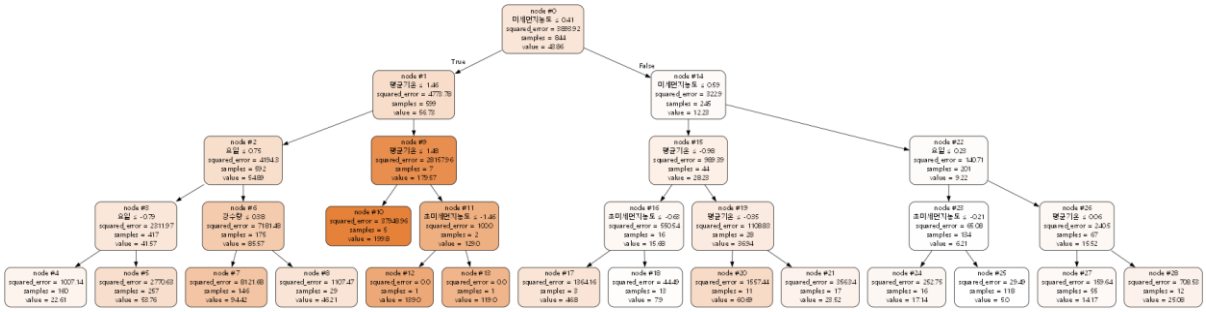
LinearRegression, DecisionTreeRegression, RandomForestRegression 3가지 모델을 사용하여 데이터를 학습시켰으며, 평가지표에 따른 모델의 성능은 아래와 같다. 모델 중에서 Test set score은 RandomForerstRegression이 0.314로 가장 높은 값을 가진다. 오차와 관련된 평가지표에 대해서도 가장 낮은 값을 갖는다.

LinearRegression	**DecisionTreeRegression**	**RandomForestRegression**
1.MAE: 36.576	1.MAE: 35.289	1.MAE: 33.248
2.MSE: 3036.19	2.MSE: 3188.878	2.MSE: 2796.146
3.RMSE: 55.101	3.RMSE: 56.47	3.RMSE: 52.878
4.R2 score: 0.255	4.R2 score: 0.218	4.R2 score: 0.218
5.Test set Score: 0.27	5.Test set Score: 0.218	5.Test set Score: 0.314

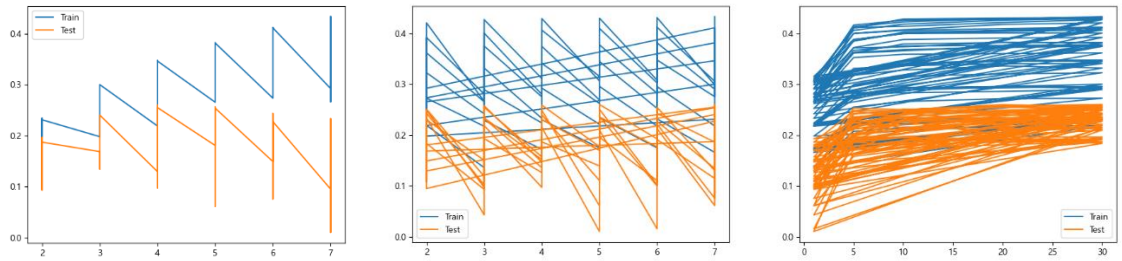
<그림7>은 LinearRegression의 회귀분석 결과를 산점도로 시각화했다. <그림8>은 트리의 결과를 시각화했으며, <그림9>는 트리의 best params의 값을 그래프로 시각화했으며, 왼쪽부터 max_depth, max_feature, n_estimator에 대한 그래프이다..



<그림7> 회귀분석 결과 산점도



<그림8> RandomForestRegression 트리



<그림9> RandomForestRegression 모델의 best params

4. 프로젝트 요약 및 결론

4.1. 모델 요약 및 결론

문화 빅데이터 플랫폼의 전국 도서 대출정보 데이터셋과 기상청의 기상요소 데이터셋을 활용하여 도서관의 일일 대출자 수 예측 모델을 구현했다. 도서관 대출자수는 기상요소 뿐만 아니라 휴관일, 건물 규모, 주변 인구 밀도 등등 다른 요소에도 영향을 미칠 것을 감안하여, 위해 서울시 내에 있는 한 도서관으로 특정하여 기상요소와 대출인원수의 상관성을 분석했다.

데이터셋의 변수 간의 상관관계를 분석하고 대출인원수와 상관성을 보이는 요일, 미세먼지농도, 초미세먼지농도, 평균기온, 최고기온, 강수량을 학습 데이터셋의 컬럼으로 선택했다. 선택한 변수로 모델이 예측하고자 하는 변수는 일일 '대출인원수'이다. 연속형 변수에 대한 값을 예측하는 모델을 구축해야 하므로 회귀분석모델 LinearRegression, DecisionTreeRegression, RandomForestRegression을 사용했고, 의사결정트리의 경우, GridSearch를 사용하여 모델의 높은 성능을 보이는 하이퍼파라미터를 탐색했다. 테스트 데이터로 정확도를 측정했을 때 3개의 모델 중 RandomForestRegression이 가장 높은 성능을 보였고, 그 값은 0.314이다.

4.2. 개선 사항

1) 모델 학습을 위한 분석 요소 추가: 도서관 행사일 데이터 반영

대출인원수는 기상요소 뿐만 아니라 도서관 행사 등의 이벤트 대해 상관관계를 보일 것으로 예상된다. 각 공공도서관에서 개최하는 행사일 데이터셋을 확보하여 모델 학습시에 활용하면 도서관에서 일일 대출인원수 예측에 대한 정확도를 높일 수 있을 것이다.

2) 모델 학습을 위한 분석 요소 추가: 유동적인 법정 공휴일, 도서관 휴관일 반영

법정공휴일과 도서관 휴관일에 도서관이 운영하지 않기 때문에 법정공휴일과 도서관 휴관일에 대한 데이터셋을 확보하고 모델의 학습 변수에 추가하면 보다 높은 성능을 보일 것이다.

3) 모델 학습을 위한 분석 요소 추가: 도서관 주변 인구 밀도, 도서관 건물 규모

도서관 주변 인구 밀도와 도서관 규모는 도서관 이용자 수와 상관성이 있을 것으로 예상되지만, 데이터셋을 확보하지 못해 이를 가공한 데이터셋에 반영할 수 없었다. 추후에 도서관 주변 인구 밀도, 도서관 규모에 대한 데이터를 확보할 방안을 마련하고, 본 데이터셋에 추가하여 모델을 학습하면 보다 높은 성능을 보일 것이다.

4) 시계열 데이터인 기상정보 데이터를 활용하여 모델의 성능 높이기

기상정보 데이터는 시계열 데이터로, 시간 순서에 따라 관측된 데이터이다. 반복 패턴을 보이는 계절성, 특정 날짜와 이전 날짜의 관계를 측정하는 자기상관성을 고려하여 기상정보 데이터와 도서 대출자수의 상관성을 파악하고 데이터를 분석하고 모델을 학습시키면 도서관 대출자수를 예측하는 데 높은 성능을 보일 것이다.

4.3. 활용 방안

1) 예측 모델을 적용할 도서관 범위 확장

예측 모델은 도봉구에 속한 공공도서관 데이터를 사용하여 구현했지만 서울시에 속한 공공 도서관 데이터셋으로 모델을 학습하고 모델의 정확도를 높여 예측모델의 사용 범위를 확장할 수 있을 것이다.

2) 적절한 직원 인력 배치로 원활한 도서관 운영

예측된 일일 대출자수를 통해 도서관은 적절한 직원 인력을 배치하여 도서관 이용자에게 원활한 서비스를 제공할 수 있도록 한다.

5. 참고문헌

5.1. 데이터셋

기상청 기상자료개방포털, 기후통계분석 2023(기온, 강수량, 풍속, 습도)
https://data.kma.go.kr/climate/RankState/selectRankStatisticsDivisionList.do?pgmNo=179

문화빅데이터 플랫폼, 공공 도서관 대출정보 2023,
https://www.bigdata-culture.kr/bigdata/user/data_market/detail.do?id=d77fa66b-6944-4d8f-b85d-79df6f5ba59e#!

문화 빅데이터 플랫폼, 도서관 정보 2023,
https://www.bigdata-culture.kr/bigdata/user/data_market/detail.do?id=7461f23e-8958-417b-be03-7ede86ab760b

서울 열린데이터 광장, 서울시 일별 평균 대기오염도 정보 2023,
https://data.seoul.go.kr/dataList/OA-2218/S/1/datasetView.do