2024-1 빅데이터 기말 프로젝트

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

컴퓨터공학전공 20200818 유수연

1. 문제 설명

도서관

- 도서관 이용자에게 도서 대출 및 반납, 예약, 상호대차 등의 서비스 제공
- 최신 도서, 다양한 분야의 자료 확보
- 독서를 즐길 수 있는 열람 공간 제공

예상치 못한 이용자 수의 증가

도서 서비스 제공, 프로그램 진행 등의
업무가 지연되는 문제 발생

도서관 일일 대출인원수를 예측하는 모델

예측된 대출인원수를 통해 도서관에서
적절한 직원 수를 배치하여
도서관 이용자에게 서비스를 원활하게 제공

가설:

1. 도서관 대출자수와 기상정보는 상관성을 가질 것이다.
2. 도서관 대출자수와 평일, 주말 여부는 상관성을 가질 것이다.

2. 사용 데이터

데이터

- 2023년 전국 공공도서관 대출내역
- 전국 공공도서관 정보
- 2023년 일별 기온, 강수량, 습도, 풍속, 미세먼지 등 기상정보 데이터

데이터 가공 및 통합

2023년 서울시 공공도서관
일일 대출내역 데이터셋

데이터 컬럼

rows: 61724, columns: 18

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

빨간색으로 색칠된 부분은 파생변수

날짜 컬럼에서 **연도, 월, 일, 요일** 추출 후 데이터셋에 추가

2023년 전국 공공도서관 대출내역 데이터에서 '도서관코드', '대출일자', '회원코드' 로 그룹화 후
각 그룹의 행의 수를 카운트하여 **도서관별 일일 대출인원수** 추출

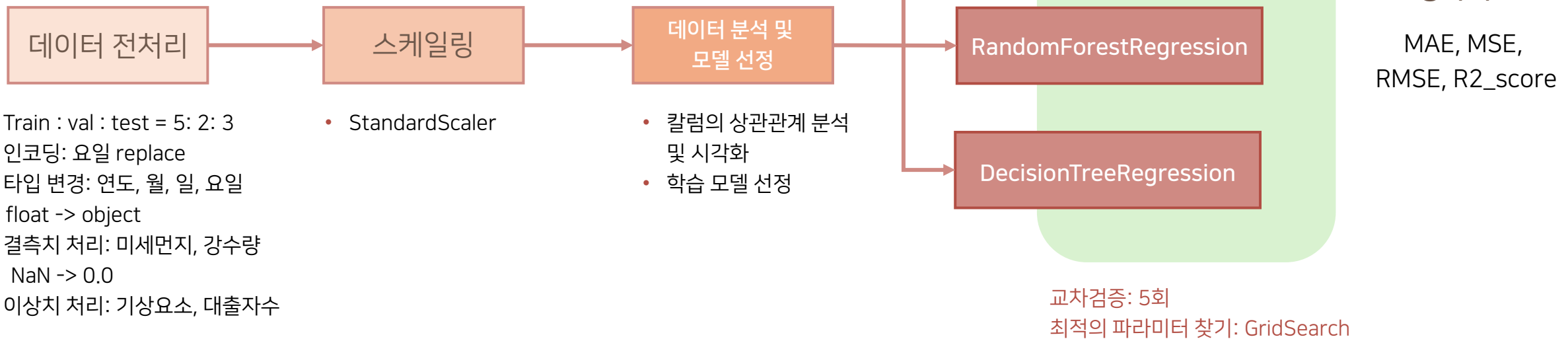
3. 분석 방법

실제 학습 시 사용한 데이터는 '도봉구'에 속한 공공도서관 중 한 도서관의 대출내역을 학습 데이터로 사용하여 학습 후 성능이 좋은 모델을 결정

실행환경

Python 3.10
IDE: Spyder
Package: pandas, sklearn, pyplot, seaborn

분석 방법



- Train : val : test = 5: 2: 3
- 인코딩: 요일 replace
- 타입 변경: 연도, 월, 일, 요일
  float -> object
- 결측치 처리: 미세먼지, 강수량
  NaN -> 0.0
- 이상치 처리: 기상요소, 대출자수

- StandardScaler

- 칼럼의 상관관계 분석 및 시각화
- 학습 모델 선정

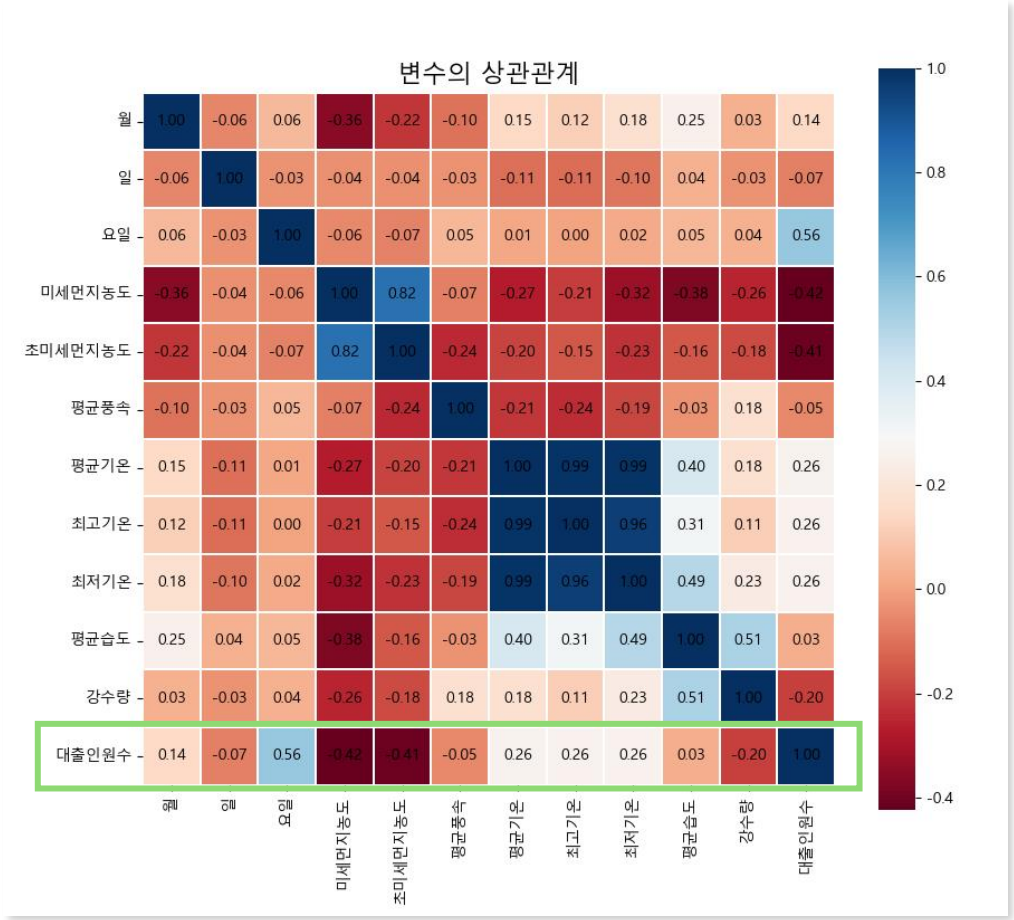
교차검증: 5회
최적의 파라미터 찾기: GridSearch

평가지표

MAE, MSE,
RMSE, R2_score

4. 분석 결과

상관관계 분석



모델 평가

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

가장 높은 성능을 보인 모델

RandomForestRegression

Score: 0.314

Best parameter: max_depth: 4, max_features: 5, n_estimators: 30

5. 프로젝트 요약 및 결론

문화 빅데이터 플랫폼의 '전국 공공 도서관 대출정보 데이터셋'과 기상청의 '기상정보 데이터셋'을 활용하여
공공 도서관의 일일 대출자수 예측 모델 구현

- 모델이 예측하고자 하는 값: 대출자수
- 사용 모델 및 선정 모델: LinearRegression, DecisionTreeRegression, **RandomForestRegression**

RandomForestRegression
1.MAE: 33.248 2.MSE: 2796.146 3.RMSE: 52.878 4.R2 score: 0.218 5.Test set Score: 0.314

개선 사항

- **모델 학습을 위한 분석 요소 추가**
  도서관 행사일, 도사관 휴관일, 유동적인 법정 공휴일, 도서관 주변 인구 밀도, 도서관 규모 등
- **시계열 데이터를 활용한 모델의 성능 높이기**
  시계열 데이터인 '기상정보 데이터셋'의 계절성과 자기상관성을 고려하여 기상정보와 대출자수의 상관성 파악

활용 방안

- **예측 모델을 적용할 도서관의 범위 확장**
  학습 변수 추가: 도서관 행사일, 도사관 휴관일, 유동적인 법정 공휴일, 도서관 주변 인구 밀도, 도서관 규모 등
- **도서관 대출자수에 따른 직원수 배치 및 도서관에서의 원활한 서비스 제공**
  예측된 일일 대출자수를 통해 도서관은 적절한 직원 인력을 배치
  → 직원들의 업무를 분산시켜 업무 효율성을 높이고 도서관 이용자에게 서비스를 원활하게 제공할 수 있다.

6. 참고문헌

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