주택 가격 요소에 따른 부동산 가격 예측 모델

- 프로젝트 설명 : 주어진 데이터를 사용하여 부동산 가격을 예측하는 모델 구현
  - 주어진 주택 가격요소에 따른 부동산 가격에 대한 데이터를 분석한 후, 가격 예측을 해야함
    - 주어진 데이터에 대해 적절히 Train, test set을 알맞은 비율로 나누어 train set에 대해 가격 예측 진행

- 제출해야 할 내용
  1. 구현 코드 (python, c, c++)
     1. 주석이 들어가야 함 **(없으면 감점)**
  2. 프로젝트 report
     1. 보고서 내용: 접근 방법 설명, 데이터 분석 내용, model 설명, 성능평가 진행 및 지표 설명, 구현 코드의 한계, 결론 등
     2. 주의사항
        - ✓보고서는 반드시 PPT로 만들어야 함
        - ✓PPT to PDF 변환 후 제출 필수
  3. 훈련 후 저장된 모델

주택 가격 요소에 따른 부동산 가격 예측 모델

- 프로젝트 평가 기준

① 데이터 분석 방법

✓ 주어진 주택 가격 요소 데이터에 대해 얼마나 잘 분석하였는가

- 여러 가정을 세우고 **시각화** 또는 **분석을 통해 가정을 증명한 뒤 모델 성능에 반영되는 과정**에 평가 점수 부여
- 예시) Model Train 시에 지역 범죄율이 높은 곳은 주택 가격이 낮은 경향 을 보임 → Test set 에 적용 시 어떤 결과를 도출할지 예측

② 분석한 데이터를 이용하여 적절한 모델을 선정하고 훈련

✓ 여러 모델을 활용하여 적절한 모델을 선정

✓ 딥러닝 모델은 활용하지 말아야 하며, 모델에 대해 본인의 이해도를 입증할 수 있게 끔 관련 설명 필요

③ 성능 평가

✓ 자체적으로 성능 평가 진행 후 평가 지표에 대한 설명 서술

✓ 성능 평가 지표는 RMSE, R-Squared 등 선택 자유 **(여러 metric을 사용할 수록 가점 – 최대 3개까지)**

유의사항

- 라이브러리 사용가능, 그러나 **라이브러리를 사용하지 않고 regression 모델을 직접 코드로 구현 시 가점**
- **단순 성능의 차이에는 점수를 부여하지 않음 (너무 예측을 못할 경우에만 감점)**

주택 가격 요소에 따른 부동산 가격 예측 모델

- Dataset 정보: Boston Housing 1970 데이터
  - Boston Housing 1970 데이터의 일부 변수를 추출한 데이터
    - 8602개의 데이터로 구성
      - 506 rows (ID) × 17 columns (가격 요소 변수)로 이루어진 csv 파일
      - 예측 target column ($\hat{y}$): CMEDV
        - ✓해당 지역의 주택 가격 중앙 값을 의미
      - 데이터 셋 주소
        - ✓파이썬 사용자는 해당 코드로 csv 포맷의 데이터 불러오면 됨

```
df=pd.read_csv("https://raw.githubusercontent.com/yoonkt200/FastCampusDataset/master/BostonHousing2.csv")
```

        - ✓C/C++ 사용자는 해당 첨부 파일 이용 (단, 첨부파일은 23개의 변수가 있으므로 4page의 해당 가격 요소 변수만 이용할 수 있도록 전 처리 필수)

	TOWN	LON	LAT	CMEDV	CRIM	ZN	INDUS	CHAS	NOX	RM	AGE	DIS	RAD	TAX
0	Nahant	-70.955	42.2550	24.0	0.00632	18.0	2.31	0	0.538	6.575	65.2	4.0900	1	296
1	Swampscott	-70.950	42.2875	21.6	0.02731	0.0	7.07	0	0.469	6.421	78.9	4.9671	2	242
2	Swampscott	-70.936	42.2830	34.7	0.02729	0.0	7.07	0	0.469	7.185	61.1	4.9671	2	242
3	Marblehead	-70.928	42.2930	33.4	0.03237	0.0	2.18	0	0.458	6.998	45.8	6.0622	3	222
4	Marblehead	-70.922	42.2980	36.2	0.06905	0.0	2.18	0	0.458	7.147	54.2	6.0622	3	222

데이터 셋 형식 (캡처로 인해 일부 column 생략)

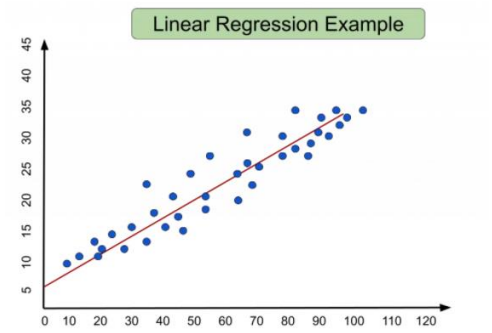
      - 주어진 데이터를 머신 러닝 기반의 regression 방식으로 학습 및 평가
      - 해당 데이터를 본인이 임의의 비율로 학습 및 테스트 셋으로 나눠야 함
- Dataset 관련 유의사항
  ① 데이터 전처리 및 특성 삭제를 할 경우 **해당 코드를 따로 만들어 함께 제출**해야 평가 가능

주택 가격 요소에 따른 부동산 가격 예측 모델

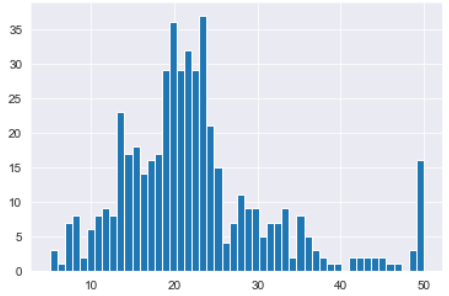
- Dataset 정보: Boston Housing 1970 데이터
  - 데이터 셋 특성
    - **TOWN**: 소속 도시 이름 -**LON, LAT**: 해당 지역의 경도(Longitudes) 위도(Latitudes) 정보
    - **CMEDV**: 해당 지역의 주택 가격 (중앙값) -**CRIM**: 지역 범죄율 (per capita crime)
    - **ZN**: 소속 도시에 25,000 제곱 피트(sq.ft) 이상의 주택지 비율
    - **INDUS**: 소속 도시에 상업적 비즈니스에 활용되지 않는 농지 면적
    - **CHAS**: 해당 지역이 Charles 강과 접하고 있는지 여부 (dummy variable)
    - **NOX**: 소속 도시의 산화질소 농도 -**RM**: 해당 지역의 자택당 평균 방 개수
    - **AGE**: 해당 지역에 1940년 이전에 건설된 주택의 비율
    - -**DIS**: 5개의 보스턴 고용 센터와의 거리에 따른 가중치 부여
    - **RAD**: 소속 도시가 Radial 고속도로와의 접근성 지수 -**TAX**: 소속 도시의 10000달러당 재산세
    - **PTRATIO**: 소속 도시의 학생-교사 비율
    - **B**: 해당 지역의 흑인 지수 (1000(Bk - 0.63)^2), Bk는 흑인의 비율
    - **LSTAT**: 해당 지역의 빈곤층 비율

주택 가격 요소에 따른 부동산 가격 예측 모델

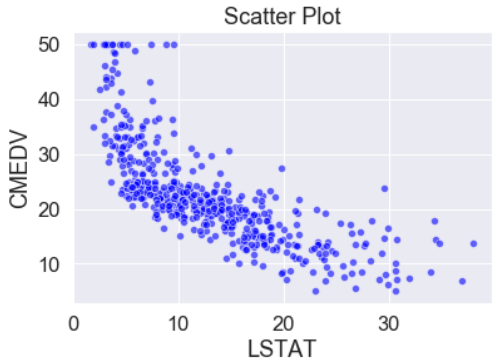
- 데이터 분석 예시
  - ① 데이터 분포 막대 그래프 출력
  - ② 각 가격 요소들에 대해 주택 가격과의 상관관계를 heatmap으로 표현
    - 오른쪽 그림 첨부
  - ③ 일부 변수 에 대한 plot 출력 예시
- 모델 Regression 결과 시각화 필수
  - **다음과 같은 형태로 시각화 가능해야 함**



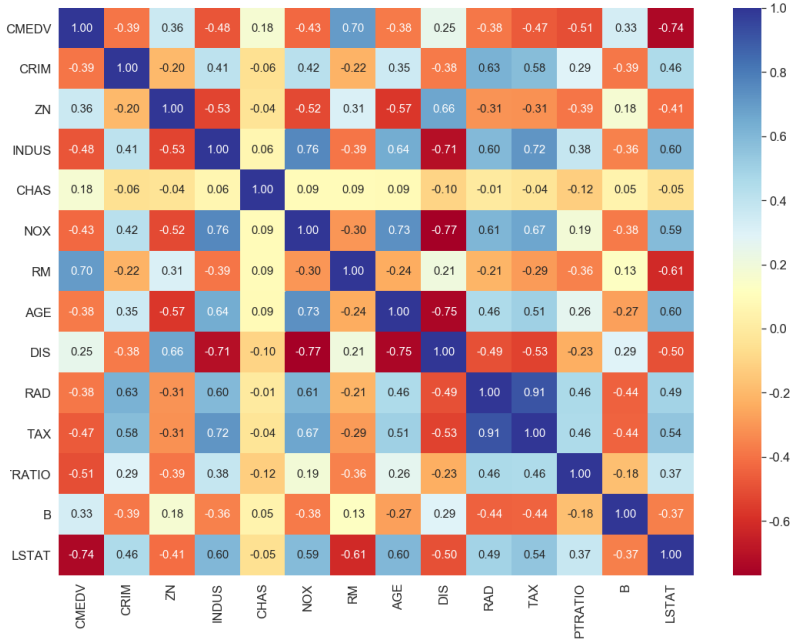
  - 해당 예시는 과제 데이터 셋 예시는 아님



데이터 분포 그래프



LSTAT에 대한 plot



상관 관계 heatmap