ADP 20회 기출문제

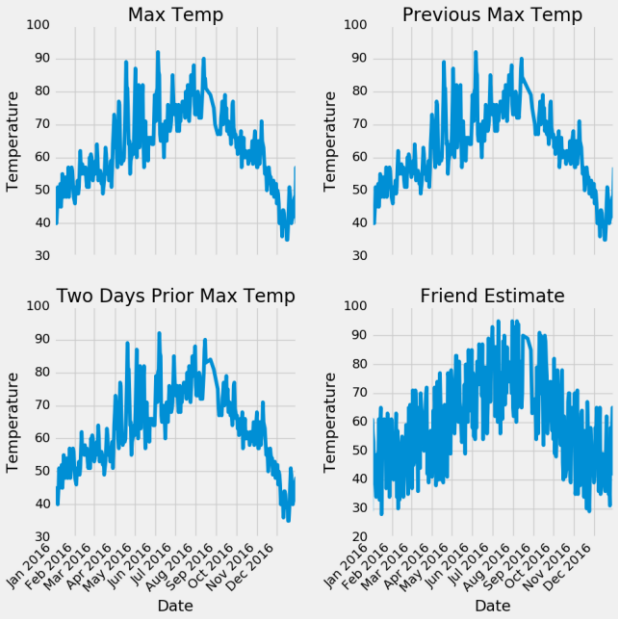
1. 기계학습 - 날씨데이터 : temps.csv (50점)

year	2016
month	1~12
day	1~31
hour	0~23
week	요일 (character)
temp_1	1일 전의 온도
temp_2	2일 전의 온도
actual	실제 최대 온도(label)
average	전년도 평균 최고 온도
friend	친구의 예측값

- 회귀분석(연속형 종속변수), 시계열 문제로 푼 사람도 있음.
- 유사 문제 : https://towardsdatascience.com/random-forest-in-python-24d0893d51c0

1-1. (10점)

- EDA
- 전처리 : 결측치 있다면 처리하고, 요일을 뜻하는 범주형 변수 원핫인코딩
- 전처리 된 데이터셋의 품질에 문제 없음을 주장 : 정량적 통계와 그래프 이용. 명확한 이상치 없으며, 결측치가 있지만 분석을 방해할 수준이 아님.
- 데이터 요약 : 2016년은 366일인데 348개의 행이 존재함. 모든 값이 있는 것은 아님. 하지만 각 열에 0이 없으며, 변칙적으로 적용될 이상치 없음.
- 그래프



- 학습-시험 데이터 분할 방법 설명 : 단순 랜덤으로 나눔. 모든 시간 포인트의 feature를 사용하기 위해 (만일 연도의 처음 9개월에 훈련하고 마지막 3개월을 예측에 사용한 경우, 우리 알고리즘은 지난 3개월 동안의 데이터를 학습할 수 없어 성능이 안나올 수 있음.)

1-2. Random Forest 모델링 (15점)
- 해당 모델의 **예측 기준선** 설정하는 방법 설명하고, 그 중 방법 선택 & 제시
- Random Forest 모델 학습하고, 시험데이터에서 성능 확인
- 예측 결과 검정 해석, 주요 변수 도출.
- 변수 중요성을 분석하고 결과 시각화 rf.feature_importances_

1-3. SVM 모델링 (15점)
- 해당 모델의 예측 기준선 설정하는 방법 설명하고, 그 중 방법 선택 & 제시
- SVM 모델 학습하고, 시험데이터에서 성능 확인
- 예측 결과 검정 해석, 주요 변수 도출.
- 변수 중요성을 분석하고 결과 시각화. svm.coef_(Python에서 SVM kernel을 linear로 하지 않는 이상 변수 중요도 제공을 안함.)

1-4. (10점)
- Random Forest, SVM 모델 성능 비교하여 최종 우수한 모델 선택 (RMSE, MAE 등 측정)
- 두 모델의 장단점 분석, 향후 운영에 있어서 어떤 모델을 선택하는게 좋을지
- 모델링 관련 추후 개선 방향 제시

풀이 :
예측 기준선? 지문에서 정의가 주어지지 않았으나, “유사 문제” 페이지에서 baseline에 속하는 듯.
(전년도 평균온도(average) – 실제 최고 온도(actual)) 의 절대값. 평균 기준선 오차 : 5.06도
모델 적용 시 평균 loss가 5도를 넘는지 확인.

성능 지표 : MAPE (오차가 예측값에서 차지하는 정도를 나타내는 지표. 수식 = 오차/예측값)
변수 중요성 : 막대그래프로 시각화

2. 통계분석 - 전력사용량 데이터 분석 : elec_use.csv (25점)

가구코드	가구번호
Date	2020-09-28
Hour	0~23
Minute	00, 15, 30, 45
P	전력사용량

- 군집 분석

2-1.
- 주어진 데이터를 가구별, 일자별 15분 간격의 데이터로 전력 사용량의 합을 구하고, 5개 그룹으로 클러스터링(group 이라는 신규 컬럼 생성) (아래와 같이 표를 완성해야함)

가구코드	Date	Total_P(총 사용량)	group
			클러스터링 한 그룹

풀이 : 날짜와 시간을 합쳐 ymd_hm 같은 형태로 만들고, K-means같은 군집화 알고리즘 사용, 15분 간격 전력 사용량의 합은 애초에 raw data가 15분 단위로 되어있었음..

2-2.
- 군집한 5개 그룹을 15분 간격의 시간에 대한 요일 별 평균으로 만들고, 히트맵 그리기
- 그룹별로 히트맵 그리기(x축은 15분 간격의 시간, y축은 요일) (아래와 같이 히트맵 나와야함)

월									
화									
수									
목									
금									
토									
일									
	1:15	1:30	1:45	2:00	2:15	2:30	2:45	3:00	3:15

풀이 : Date 컬럼을 이용해서 요일 컬럼 생성, 히트맵 작성

- groupby 다음 바로 히트맵 그릴 수 없어 pivot table로 변환 후 그림.

3. 기계학습 - sun_power.csv (25점)

- 시간별 발전량(PV) 예측
- 데이터셋 분할 및 결과 검증 (필요 시 파생변수 추가 가능)
- 모델 성능 검증 : RMSE, R^2(R-Squared), 정확도 계산
  (정확도 산식은 문제에서 제시
  → '예측값이 실제값보다 크다면 1-(실제값/예측값), 실제값이 예측값보다 크다면 1-(예측값/실제값)')
- 훈련-테스트 데이터 랜덤으로 7:3 분할

풀이:
알고리즘 상관 없이 회귀 분석. 시간 안배를 위해서 랜덤포레스트 사용(1번에서 사용한 코드여서)

RMSE : 평균 제곱근 오차
R-Squared : 결정계수

*** 참고 자료**

- 1번문제와 유사문제 : https://towardsdatascience.com/random-forest-in-python-24d0893d51c0

- 20회 문제 복기 및 후기 :
- https://ysyblog.tistory.com/221
- https://bluemumin.github.io/adp/2021/03/27/ADP-20adp/
- https://cafe.naver.com/sqlpd/18890
- https://cafe.naver.com/sqlpd/18900

- ADP Python 준비 코드 github
- https://github.com/bluemumin/ADP_certificate_preperation