2. Bike Sharing Demand

◯ Data

- train.head() → (10886, 12)
- test.head() → (6493, 9)
- 컬럼 설명
  - ① datetime : 시간. 연-월-일 시:분:초
  - ② season : 봄(1), 여름(2), 가을(3), 겨울(4)
  - ③ holiday : 1이면 공휴일, 0이면 공휴일이 아님
  - ④ workingday : 1이면 근무일, 0이면 근무일이 아님
  - ⑤ weather : 1 ~ 4 사이의 값
    1(깨끗한 날씨), 2(약간 안개와 구름이 끼어있는 날씨),
    3(약간 눈, 비가 오거나 천둥), 4(아주 많은 비 또는 우박)
  - ⑥ temp : 온도. 섭씨(Celsius)
  - ⑦ atemp : 체감 온도. 섭씨(Celsius)
  - ⑧ humidity – 습도.
  - ⑨ windspeed – 풍속.
  - ⑩ casual : 비회원(non-registered)의 자전거 대여량
  - ⑪ registered : 회원(registered)의 자전거 대여량.
  - ⑫ count : 총 자전거 대여량으로
    비회원(casual) + 회원(registered)

◯ Environment

- Python 3.73 with `jupyter==1.0.0`
- pandas, numpy, matplotlib, seaborn
- scikit-learn

➢ Perpose

정 시간대에 얼마나 많은 사람들이 자전거를 대여하는지 예측하는 것

➢ Methodology

- 알고리즘 : RandomForest, lightGBM(Coarse & Fine Search)
- 모델평가 : Root Mean Squared Logarithmic Error (RMSLE) $\sqrt{\frac{1}{n}\sum_{i=1}^{n}(\log(p_i+1)-\log(a_i+1))^2}$
  ※ 정답(aiai, actual)과 예측값(pipi, predict)의 차이가 크면 클수록 페널티를 덜 주는 방식

➢ Solution

- Exploratory data analysis
  - overfitting을 피하기 위해 datetime 컬럼 중 year, hour만 사용
  - 알고리즘 성능향상을 위해 요일(dayofweek) feature 생성하여 학습시킴
    ※ 근무일(workingday) column과 시너지 효과가 난다고 판단
  - seaborn의 distplot을 통해 데이터 왜곡(skewed) 현상 발견
    ✓ 자전거 대여량이 1 ~ 20대인 비중이 굉장히 높음
    ✓ 자전거 대여량이 1,000대에 근접하는 경우도 있음 (977대)
    ■ 해결 Idea : Evaluation에 인사이트를 얻어 자전거 대여량(count) colum을 log transformation → 정규분포화 함으로써 예측 정확도를 높이고자 함
- Feature : "season", "holiday", "workingday", "weather", "temp", "atemp", "humidity", "windspeed", "datetime-year", "datetime-hour", "datetime-dayofweek"
- Label: log transformation한 자전거 대여량(log_count)
- 스코어 : RandomForest 0.51037, LightGBM 0.37502 예측(상위2.9% Rank)