DACON
DATA TO VALUE

Dacon 13회

퇴근시간 버스 승차인원 예측 경진대회

제주감귤팀

남윤주 류예나 이채연 최승희 황태용

목차

1 데이터 전처리 & EDA

1.1~1.8 EDA를 통한 변수 생성 과정

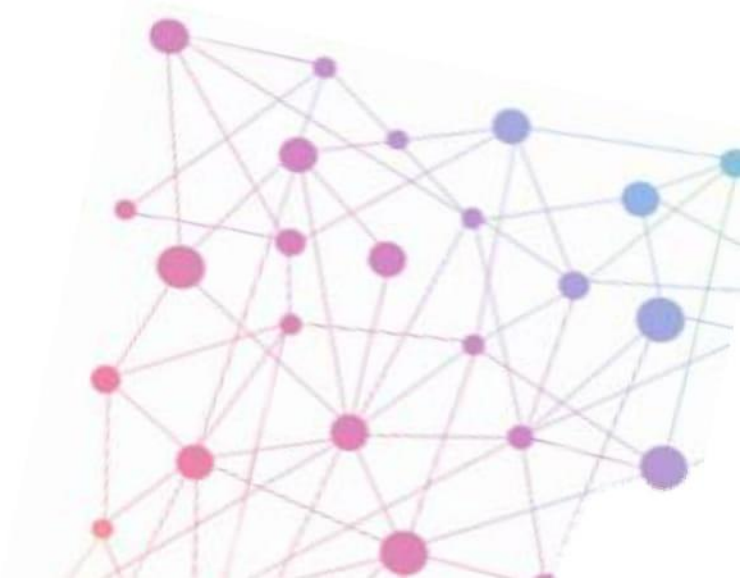
2 모델 구축 & 검증

2.1 전체적인 Process

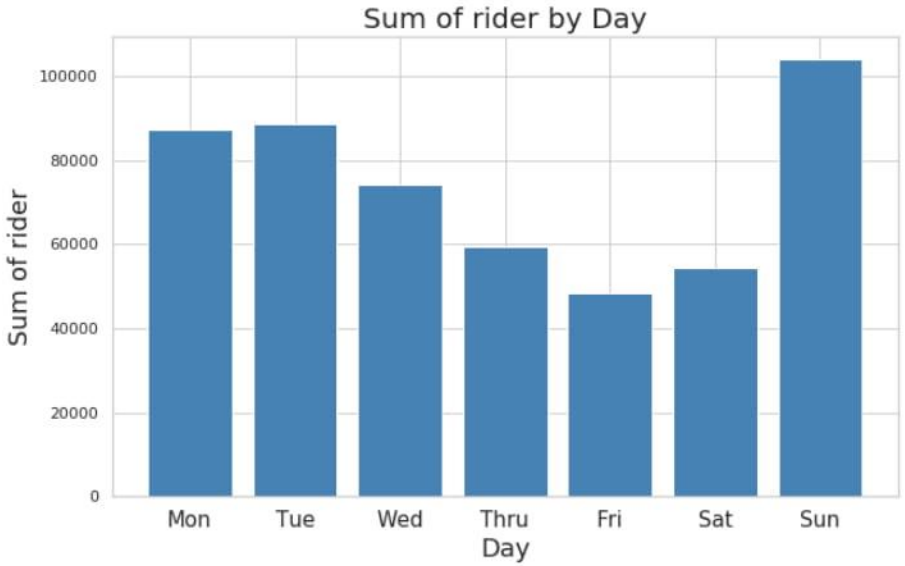
2.2 Adapted model - RF

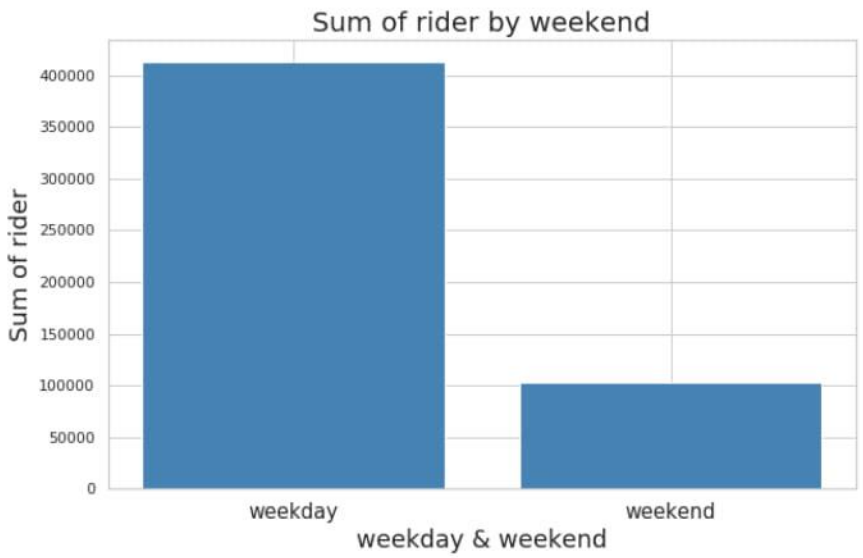
2.3 Ensemble

3 결과 및 결언



1-1. 요일에 따른 18~20시 탑승객 수



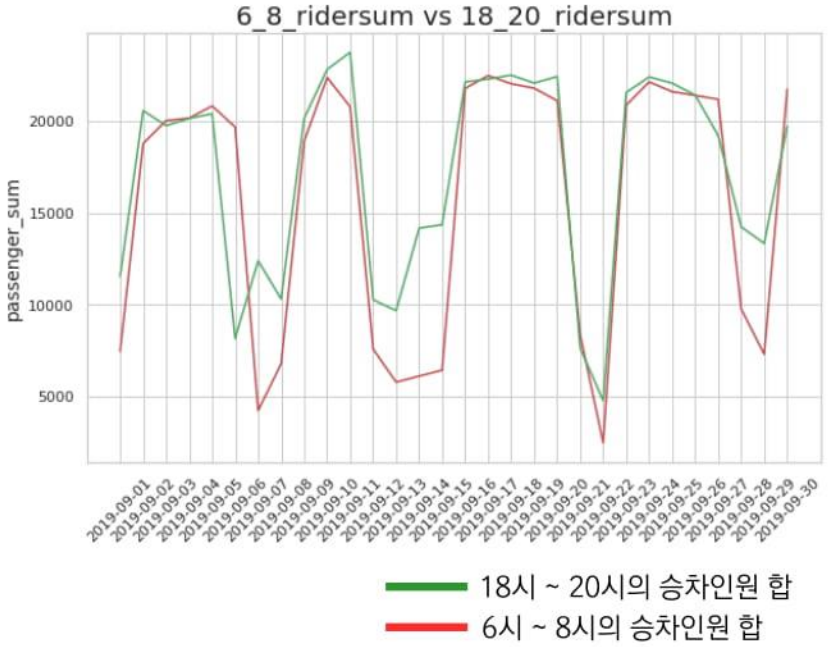


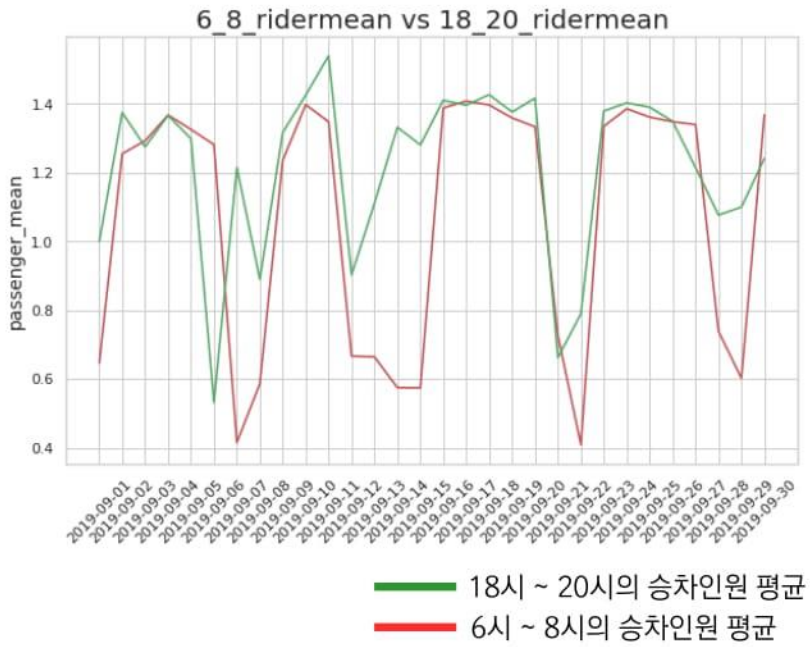
- **분석 결과** : 요일에 따라 탑승객 수가 확연히 차이가 나고, 특히 주중과 주말이 큰 차이가 남.
  월요일에서 금요일로 갈수록 탑승객 수가 줄어들다가, 다시 주말에 탑승객 수가 점점 올라감.
- **피쳐 생성** : weekday – date변수를 datetime으로 type변경 후, 요일을 dummy 변수 생성
  weekend – weekday변수에서 주중 0, 주말 1
  1820_w_mean – 요일별 18~20시의 탑승인원의 평균
  1820_w_sum – 요일별 18~20시의 탑승인원의 합

1-2. 출근시간 승차인원과 퇴근시간 승차인원

데이터 다운로드

DACON
DATA TO VALUE



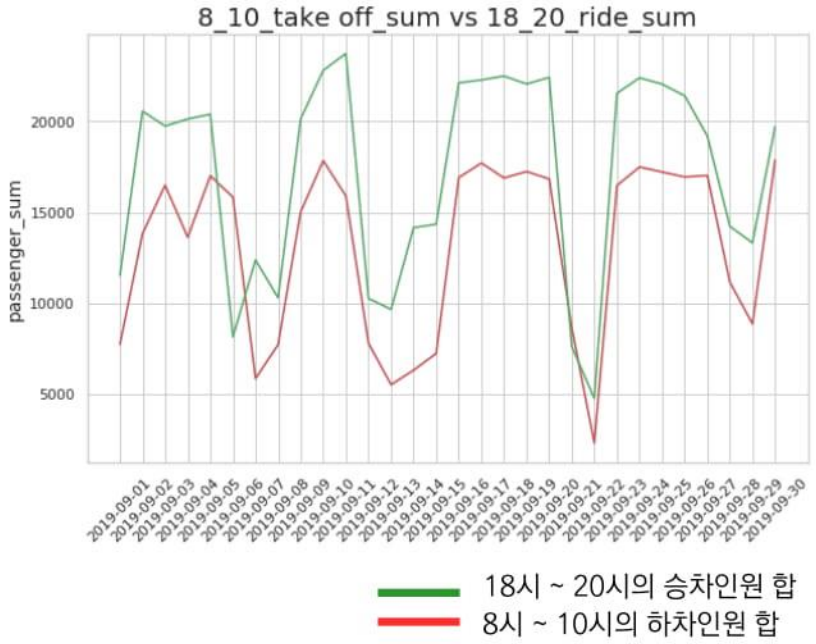


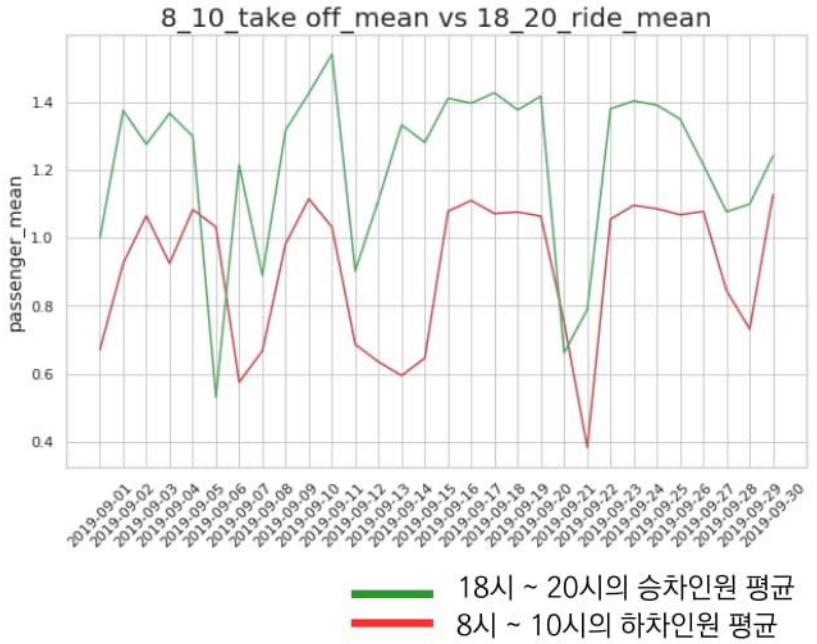
- **분석 결과** : 출근시간의 승차인원 합과 평균 / 퇴근시간 승차인원의 합과 평균이 비슷한 양상을 띠고 있음.
  즉, 버스를 타고 출근한 인원은 퇴근 후에도 버스를 타고 귀가함.
- **피쳐 생성** : 68a, 810a, 1012a– t, t+2 기준 출근시간 승차인원의 정보를 담고 있음.
  69a, 912a – t, t+3 기준 출근시간 승차인원의 정보를 담고 있음.

1-3. 출근시간 하차인원과 퇴근시간 승차인원

데이터 다운로드

DACON
DATA TO VALUE

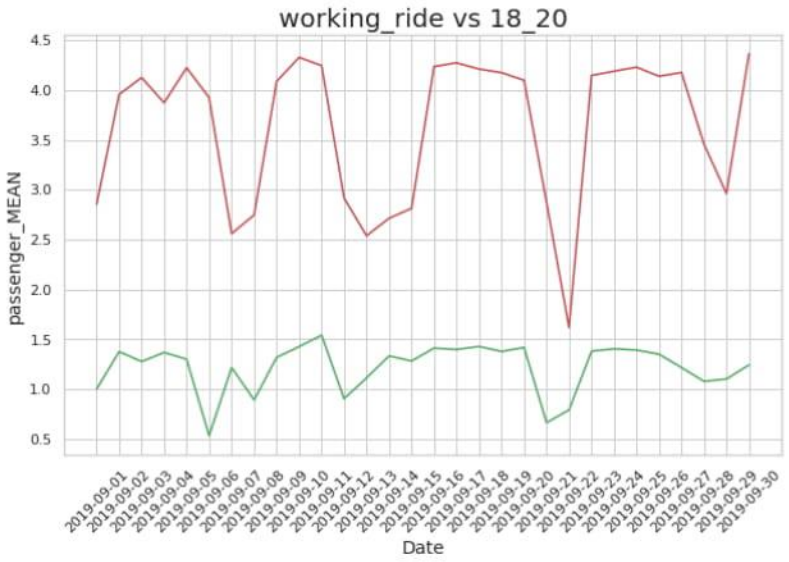


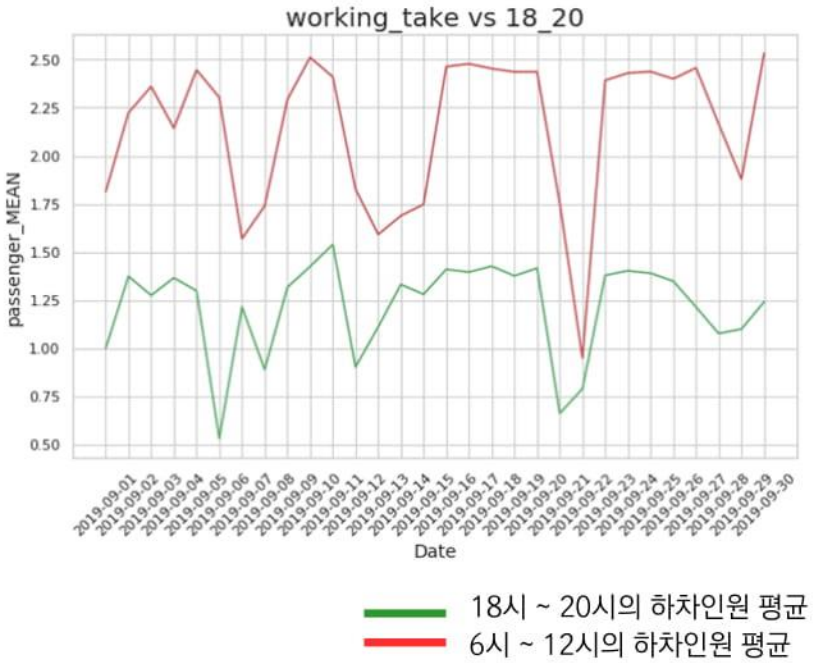


- **분석 결과** : 출근시간 하차인원의 합과 평균 / 퇴근시간 승차인원의 합과 평균이 비슷한 양상을 띠고 있음.
  출근시간 하차인원이 많은 곳에서는 퇴근시간 승차인원이 많을 것으로 예상됨.
- **피쳐 생성** : 68b, 810b, 1012b– t, t+2 기준 출근시간 하차인원의 정보를 담고 있음.
  69b, 912b – t, t+3 기준 출근시간 하차인원의 정보를 담고 있음.

1-4. 오전시간대 승/하차인원과 퇴근시간 승차인원

데이터 다운로드

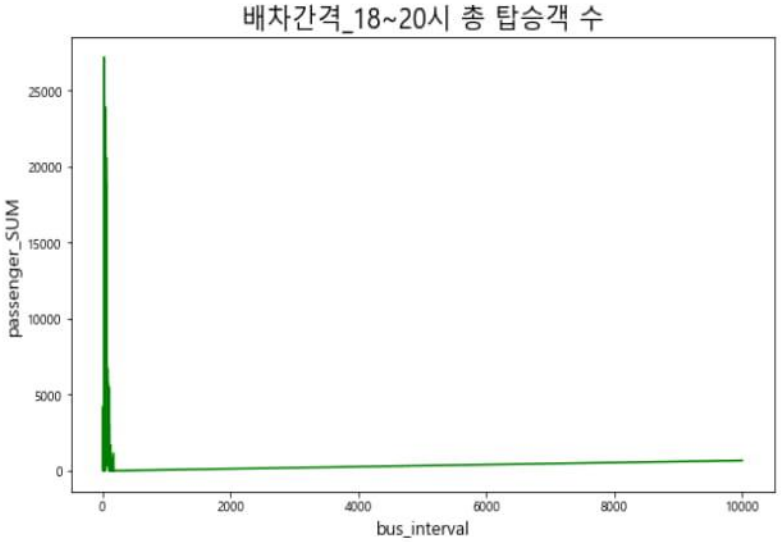




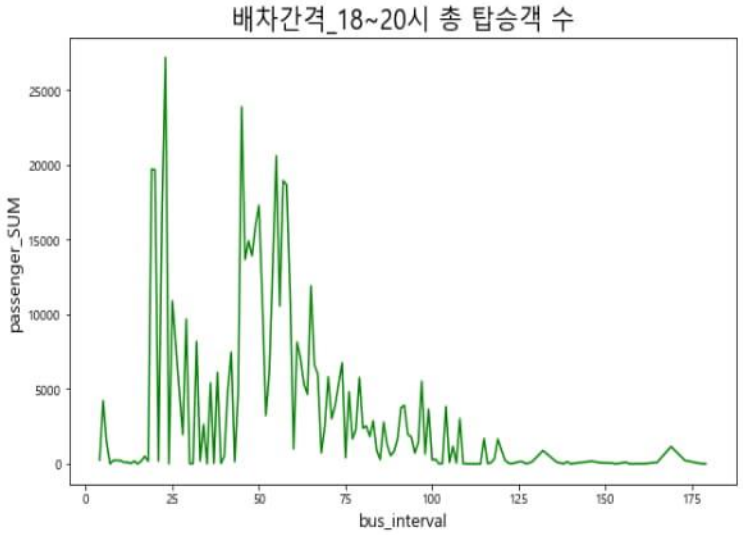
- **분석 결과** : 오전시간에 승/하차인원이 많다는 것은 그날의 유동인구가 많다는 것을 뜻함.
  주말이나 연휴에는 유동인구 자체가 적고, 주중에는 상대적으로 유동인구가 많음.
- **피쳐 생성** : ride_sum – 오전시간 승차인원 합
  takeoff_sum – 오전시간 하차인원 합

1-6. 배차 간격에 따른 퇴근시간 탑승객 수

데이터 다운로드



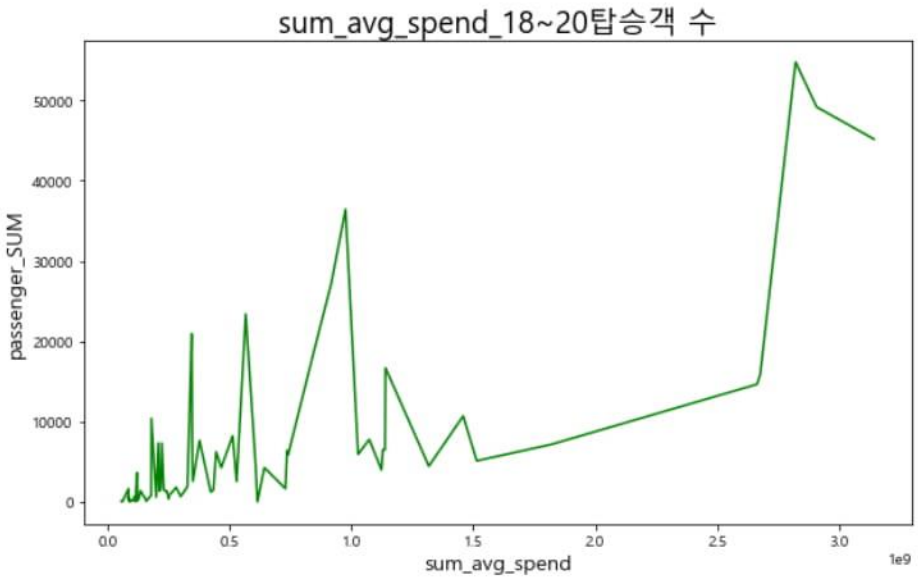
로그변환



- **분석 결과** : 배차 간격에 따른 퇴근시간 탑승객 수 사이의 패턴을 명확히 보기 위해 로그변환을 함.
  수요가 많은 버스일수록 배차간격이 짧고, 그에 따라 배차 간격이 짧을수록 퇴근시간 탑승객 수가 많음.
- **피쳐 생성** : bus_interval – bus_route_id별 배차간격의 평균값.
  Scale에 무관한 Tree기반 모델을 사용했기 때문에 실제 변수는 log변환을 하지 않음.

1-7. 소비 비율에 따른 18~20시 탑승객 수

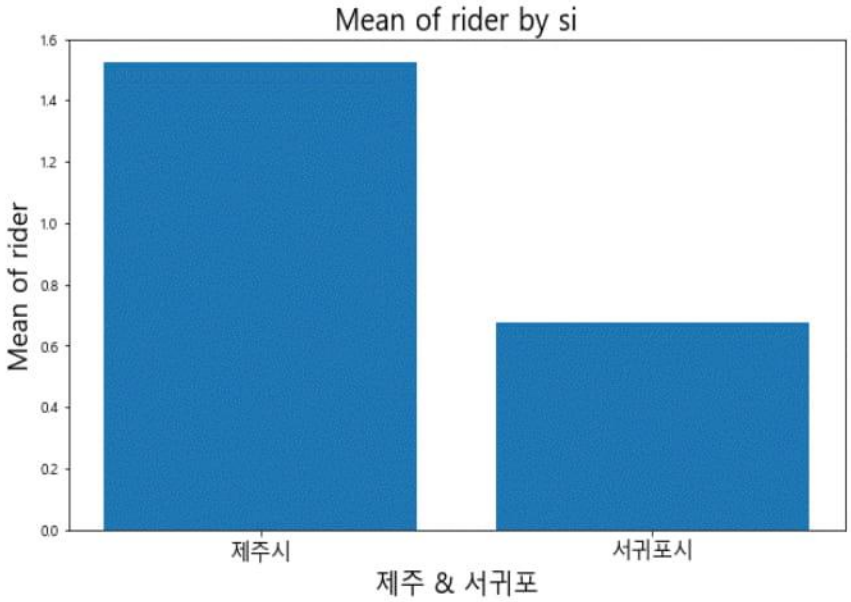
데이터 다운로드



- **분석 결과** : 평균 소비액이 높다는 것은 그만큼 소비자, 즉 유동인구가 많다는 것을 의미하기도 함.
  위의 그래프는 평균 소비액이 높을수록 대체로 탑승객 수가 많음을 보여줌.
- **피쳐 생성** : mean_avg_spend – 동별 소비액의 평균
  sum_avg_spend – 동별 소비액의 합
  rate_avg_spend – 동별 소비액의 비율

1-8. 제주특별자치도의 시별 탑승객 수

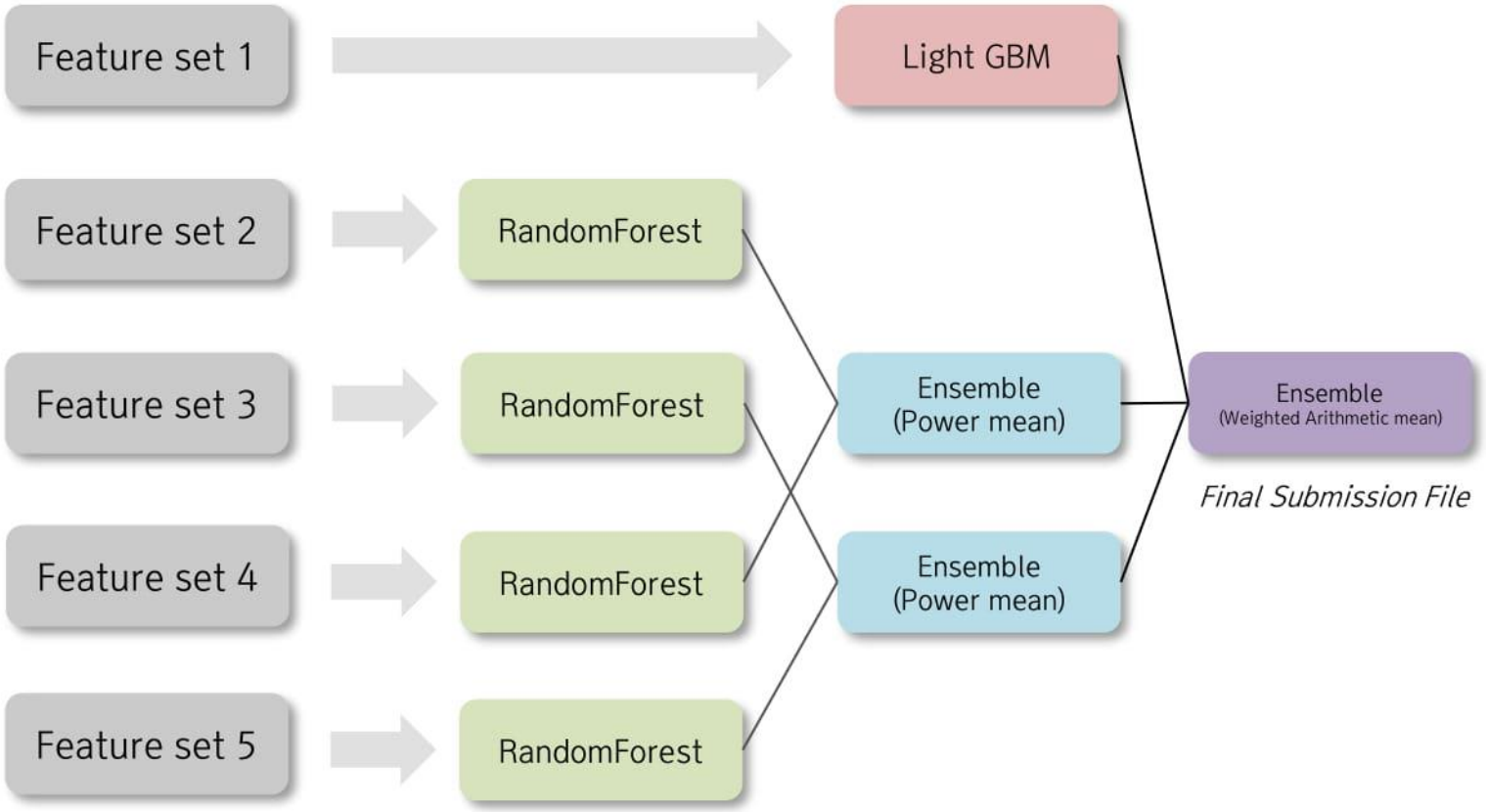
데이터 다운로드



- **분석 결과** : 제주시의 평균 탑승객 수와 서귀포시의 평균 탑승객 수는 거의 두 배 차이임.
  따라서 제주시에 가까운 버스정류장일수록 탑승객이 많을 것이며, 서귀포시에 가까운 버스정류장일수록 탑승객이 적을 것으로 예상됨.
- **피쳐 생성** : dis_jejusi – 위도 경도 정보를 통해 제주시의 중심으로부터의 거리 정보가 담겨 있음.
  si – 외부 프로그램을 사용하여 주소를 추출한 후 제주시와 서귀포시를 더미 변수 생성

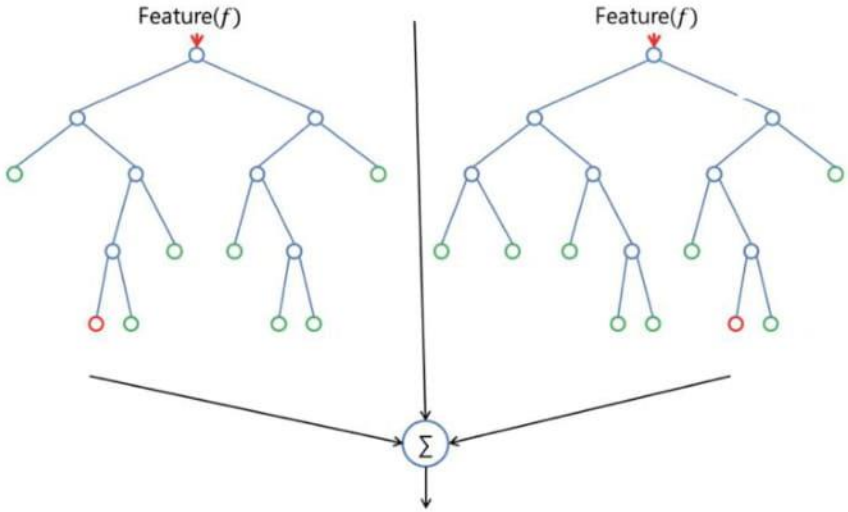
2-1. 전체적인 Process

데이터 다운로드



2-2. Adapted model – Random Forest

데이터 다운로드



- ✓ 일반적으로 Boosting 모델이 RandomForest 보다 성능이 높다. 하지만, Overfitting의 위험성이 더 큼.
- ✓ 우리의 문제는 9월의 한달 간의 데이터로 10월 중순까지의 데이터를 예측하는 문제임. 9월의 데이터 만으로 Boosting 모델을 학습하는 것은 Overfitting의 위험이 있으며, 실제로도 CV rmse와 public rmse score 사이의 큰 차이가 있는 것을 확인함. 이에 우리 팀은 Randomforest를 주 모델로 채택
- ✓ 앞선 EDA 과정에서 데이터가 요일에 따른 시계열성을 띠는 것을 확인해볼 수 있었다. 시계열 모델을 사용하지 않은 이유는 9월과 10월에 공휴일이 많고 짧은 학습 데이터 만으로 일반화된 모델을 만들기 어렵기 때문임.

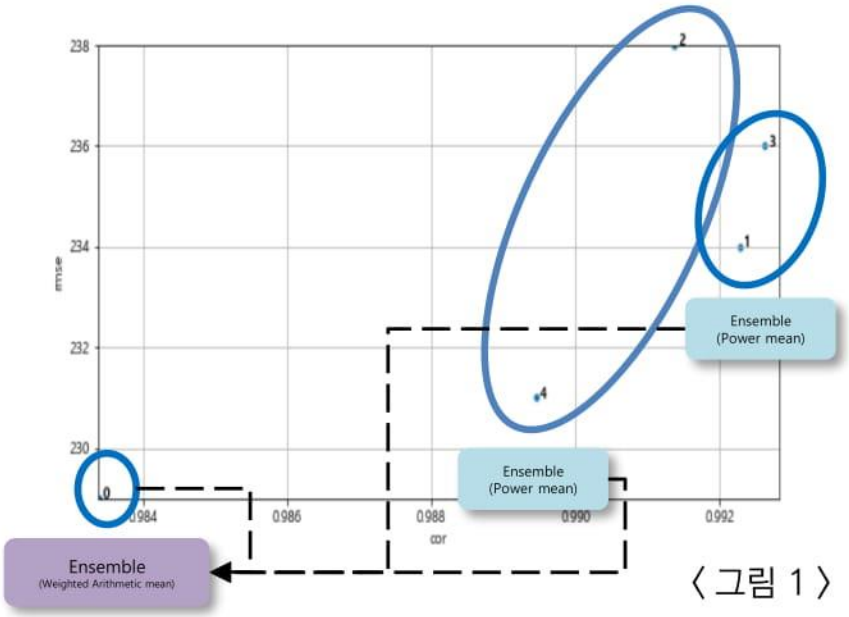
2-3. Ensemble

데이터 다운로드

DACON
DATA TO VALUE

	lgbm0=229.csv	rf1=234.csv	rf2=238.csv	rf3=236.csv	rf4=231.csv
lgbm0=229.csv	1.000000	0.978485	0.975649	0.978608	0.984244
rf1=234.csv	0.978485	1.000000	0.996976	0.997258	0.988642
rf2=238.csv	0.975649	0.996976	1.000000	0.998541	0.985677
rf3=236.csv	0.978608	0.997258	0.998541	1.000000	0.988685
rf4=231.csv	0.984244	0.988642	0.985677	0.988685	1.000000

〈 표 1 〉



〈 그림 1 〉

- ✓ 〈표 1〉 Feature Set 1 ~ 5로 만든 최종 예측 파일의 상관계수 행렬
- ✓ 〈그림 1〉 최종 예측 파일들의 성능과 파일들 간의 평균적인 상관관계를 나타낸 그림
- ✓ 상관 관계가 높은 모델 끼리는 멱 평균(Power mean)을 수행 함 (멱 평균 – 1 & 3, 2 & 4)
- ✓ 멱평균을 통해 나온 결과는 Feature Set – 0과 가중 산술 평균(Weighted Arithmetic mean)을 수행 함.

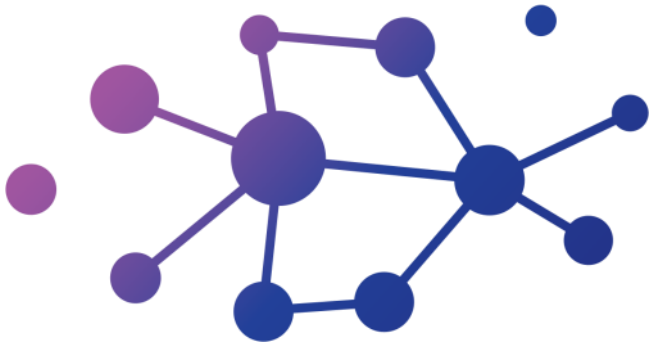
3. 결과 및 결언

데이터 다운로드

DACON
DATA TO VALUE

- ✓ 모델 구축의 목적에 맞게 오전에 발생할 수 있는 데이터만 활용 함 (data leakage 차단)
- ✓ 여러 단일 모델을 앙상블 → 안정성 있는 모델 구축
- ✓ Overfitting 가능성이 상대적으로 높은 Boosting 모델 지향 → 안정성 있는 모델 구축
- ✓ Id당 데이터 발생 건수가 적어 Overfitting의 가능성이 큼.
  모든 의사결정을 Cross validation rmse를 기준으로 함 → Shake, Overfitting 방지
- ✓ 위의 과정을 통해 Variance가 적은 일반화된 모델 구축
  (CV rmse, Public rmse, Private rmse 사이의 분산이 적음)

THANK YOU



대회 참가해보기