Smarcle Data Analysis Study

3장. 퇴근시간 버스 승차인원 예측

Team 2 _ 구범준, 박지하, 최태규, 장윤정

목차

01.1 문제 정의

1. 문제점

- 2019년 11월 기준 제주도 인구 외국인과 관광객 포함 90만명 추정
- 약 50만명 제주시에 거주
- 인구밀집으로 인한 심각한 교통 체증 발생

2. 목적

- 퇴근시간 버스 승차인원 예측 -> 효율적인 버스 운행

01.2 평가 척도

평균 제곱근 편차

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(\hat{y}_i - y_i)^2}{n}}$$

$\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_n$ are predicted values

$y_1, y_2, \ldots, y_n$ are observed values

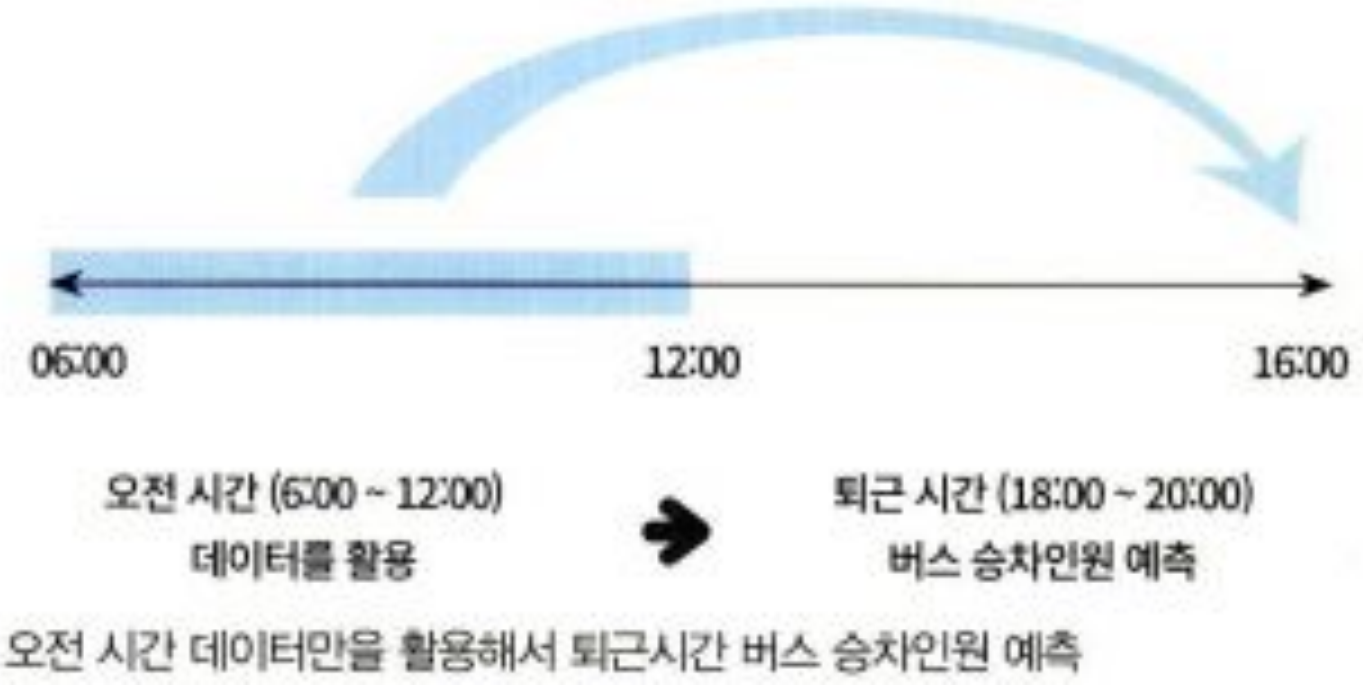
n is the number of observations

- 예측한 퇴근시간 버스 승차인원이 실제 퇴근시간 버스 승차인원과 유사할수록 RMSE가 낮음

01.3 문제 해결을 위한 접근 방식

주의할 점

1) 오전(06:00~12:00) 데이터 활용 -> 퇴근시간(16:00~20:00) 승차인원 예측



01.3 문제 해결을 위한 접근 방식

주의할 점

2) (9/1 ~ 9/30)일간의 데이터 -> (10/1 ~ 10/15)일 간의 데이터 예측

주의할 점2: 9월 1~30일의 데이터로 10월 1~15일의 퇴근시간 버스 승차인원 예측

01.3 문제 해결을 위한 접근 방식

주의할 점

3)
train, test , bus_bts 데이터
버스에서 하차를 할 때, 버스카드를 찍지 않는 경우, 해당 기록이 비어 있는 상태. 따라서, 승차 인원수와 하차 인원수가 동일하지 않고 다소 차이가 있음

train, test csv
같은 정류장 이름이지만 위도와 경도가 서로 다른 경우가 존재.
해당 경우는, 같은 정류장 이름을 가지고 있는 길 건너편의 정류장.

머신러닝 모델

단일 모델

앙상블 모델

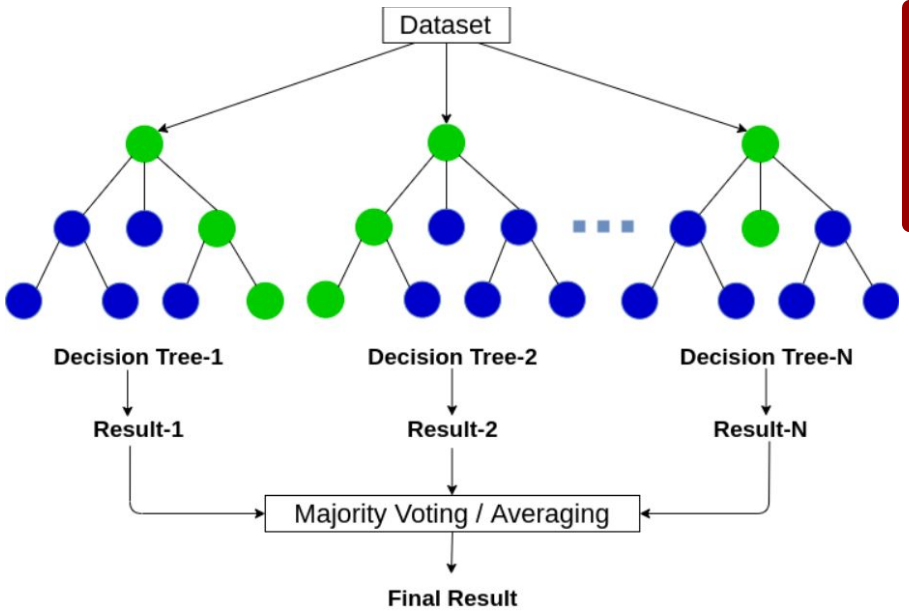
배깅 방식 앙상블 모델

랜덤 포레스트

부스팅 방식 앙상블 모델

XGBoost

LightGBM



SMARCLE

- **변수의 수가 무조건 많다고 좋은 것이 아님!**
- **유용한 변수를 찾아야 모델이 간단해지고 일반화 성능을 높임(과적합 피함)**

- **변수 선택 방법 : A/B 테스트**

  1. 트리 계열의 모델과 기본 변수 설정

  (e.g. RandomForestRegressor(random_state=1217))

  2. 기본 변수만 가지고 설정한 모델의 성능 확인

  (e.g. 교차검증)

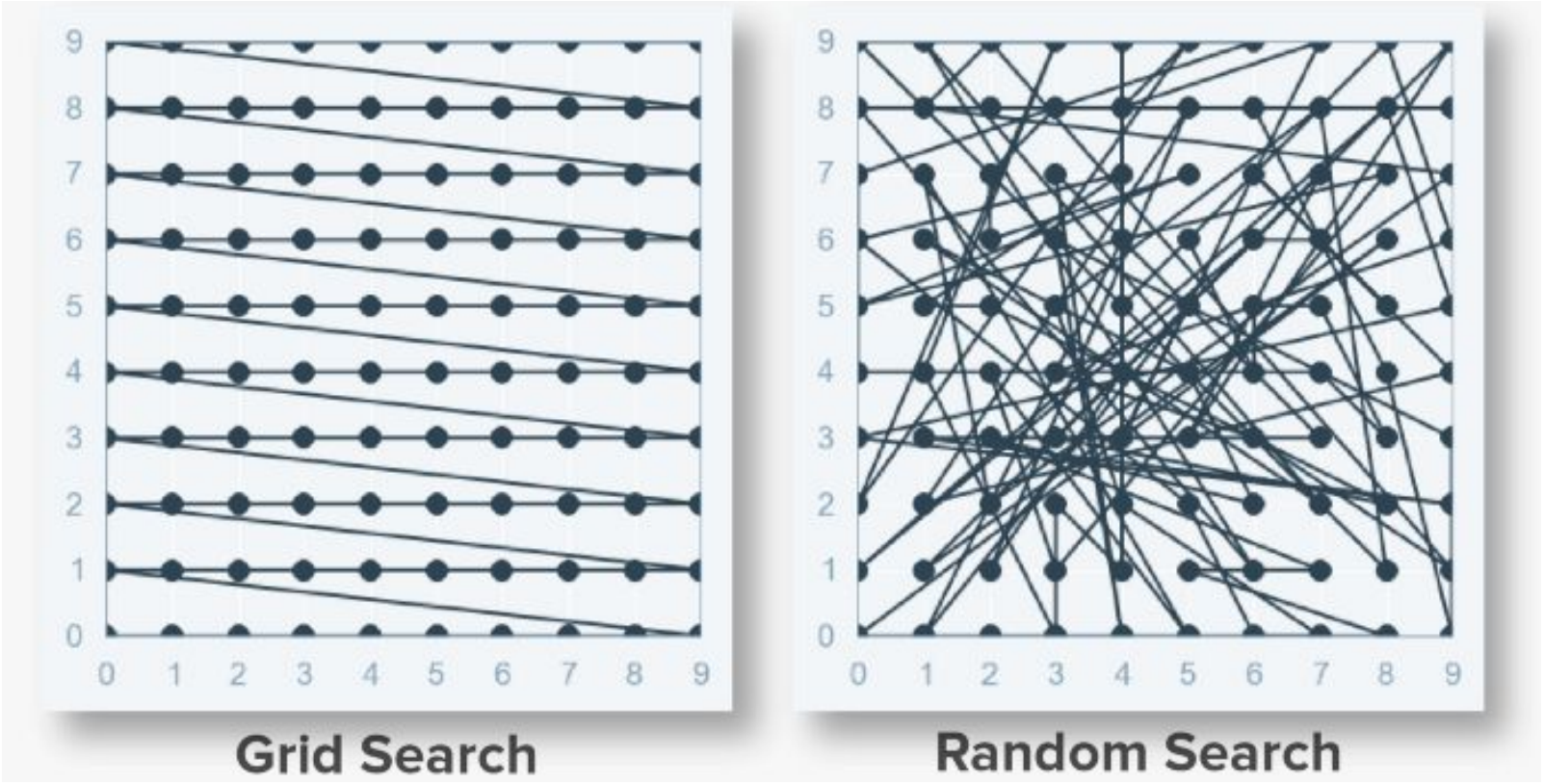
  3. 여기에 변수를 하나씩 추가해 성능이 향상되면 그대로 두고, 향상되지 않으면 제거 -> 반복

  (e.g. 처음 86개 변수에서 68개로 축소)

```
model = RandomForestRegressor(random_state=1217, max_features=7, min_samples_leaf=2, n_estimators=400, n_jobs=-1)
```

그리드 탐색

임의 탐색



04 성능 향상을 위한 방법

- 대회(DACON)의 목표? = 문제에서 원하는 결괏값과 가장 근접하게 예측하기!

Submission 간 앙상블

구축이 끝난 머신러닝 모델에서 도출된 결괏값을 결합함으로써
더 정확한 예측값을 도출하는 기법

→ 임시 스코어와 파일 간 상관관계를 파악

(임시 스코어 : 노트북 상 RSME 값(X) / 대회에서 요구하는 실제 값과 차이를 구한 임시 점수)

04 성능 향상을 위한 방법

ex) submission 파일이 5개인 경우라고 가정

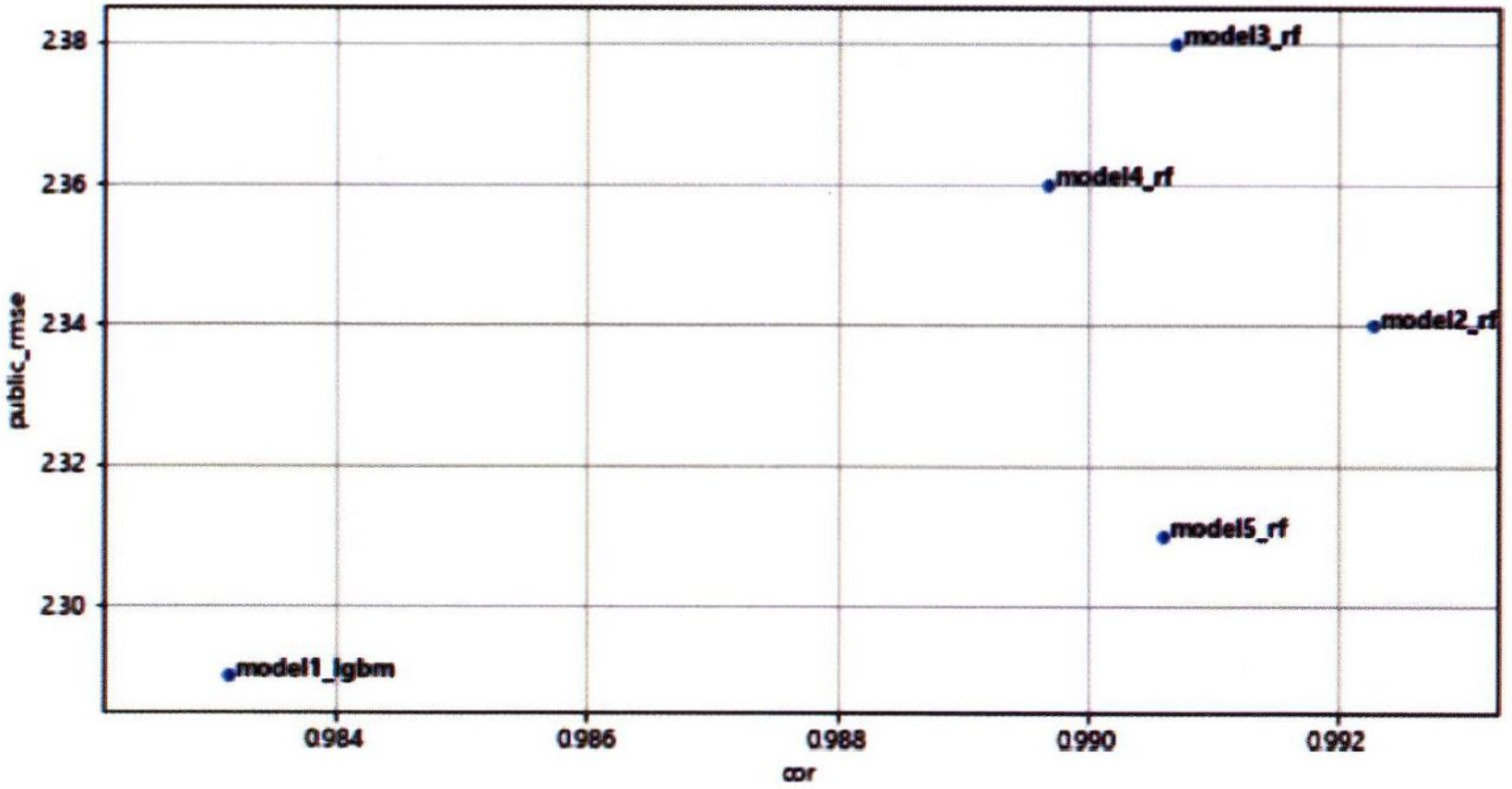
- 편의를 위해 각 파일 명에 임시 스코어 표시 (ex. model1.csv -> model1_2.29.csv)

1. 상관계수 행렬 확인

	model1_lgbm=2.29.csv	model2_rf=2.34.csv	model3_rf=2.38.csv	model4_rf=2.36.csv	model5_rf=2.31.csv
model1_lgbm=2.29.csv	1.000000	0.979854	0.974709	0.974184	0.986977
model2_rf=2.34.csv	0.979854	1.000000	0.995806	0.993388	0.992342
model3_rf=2.38.csv	0.974709	0.995806	1.000000	0.995072	0.987900
model4_rf=2.36.csv	0.974184	0.993388	0.995072	1.000000	0.985777
model5_rf=2.31.csv	0.986977	0.992342	0.987900	0.985777	1.000000

ex) submission 파일이 5개인 경우라고 가정

2. 임시 스코어를 한눈에 알아보기 위한 산점도



	model	public_rmse	cor
0	model1_lgbm	2.29	0.983145
1	model2_rf	2.34	0.992278
2	model3_rf	2.38	0.990697
3	model4_rf	2.36	0.989684
4	model5_rf	2.31	0.990599

※ cor = 상관계수 행렬의 각 행의 평균

ex) submission 파일이 5개인 경우라고 가정

3. 여러가지 앙상블 기법

– 가중산술평균

: 산술평균에 가중치를 반영한 평균값

ex) 고속도로 구간 단속
최고 제한속도는 110km/h 이라고 가정. 100km/h로 1분, 115km/h로 4분을 달린 경우

1. **산술평균 : (100 + 115) / 2 = 107.5km/h –> 구간 단속에 걸리지 않음 (No!!)**
2. **가중산술평균 : (100 x 1 + 115 x 4) / (1 + 4) = 112km/h –> 구간 단속에 걸림**

ex) submission 파일이 5개인 경우라고 가정

3. 여러가지 앙상블 기법

– 멱 평균

: 평균식을 일반화한 식

$$M_p(x_1, \cdots, x_n) = \left(\frac{1}{n} \cdot \sum_{i=1}^{n} x_i^p \right)^{1/p}$$

※ p 값이 1일 때 = 산술평균
→ 적절한 p값을 조정하며 각 결괏값에 다양한 가중치 부여

ex) submission 파일이 5개인 경우라고 가정

4. 앙상블 과정

1. model 1번과 5번 파일을 p = 21로 하여 멱 평균을 취한다.
2. model 2번과 4번 파일을 p = 21로 하여 멱 평균을 취한다.
3. 1.의 결과 파일과 2.의 결과 파일, model 3번 파일 총 3개의 파일을 가중산술평균을 이용해 앙상블
   (1.의 결과 파일에는 0.22, 2.의 결과 파일에는 0.30, model 3번 파일에는 0.48의 가중치 부여)
4. 최종 예측값 도출

감사합니다 :)

Team 2 _ 구범준, 박지하, 최태규, 장윤정