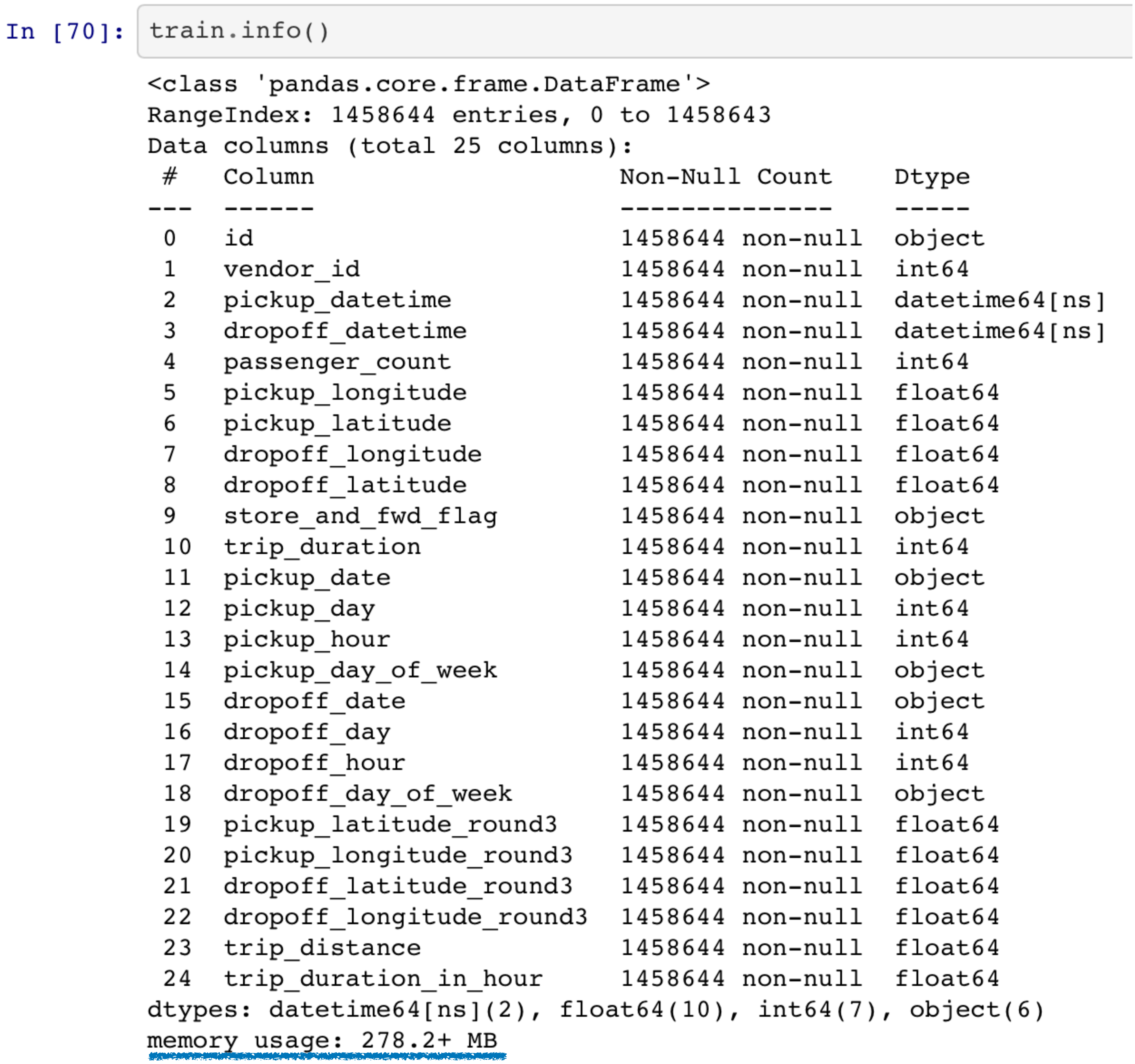
이 Random Forest을 계속 쓰면 안될까요?

2021.11.14 김은희

많은 데이터를 다룰 때?

- PC 사양 : 16GM RAM
- train 데이터 로딩에 278.2 MB 메모리 사용

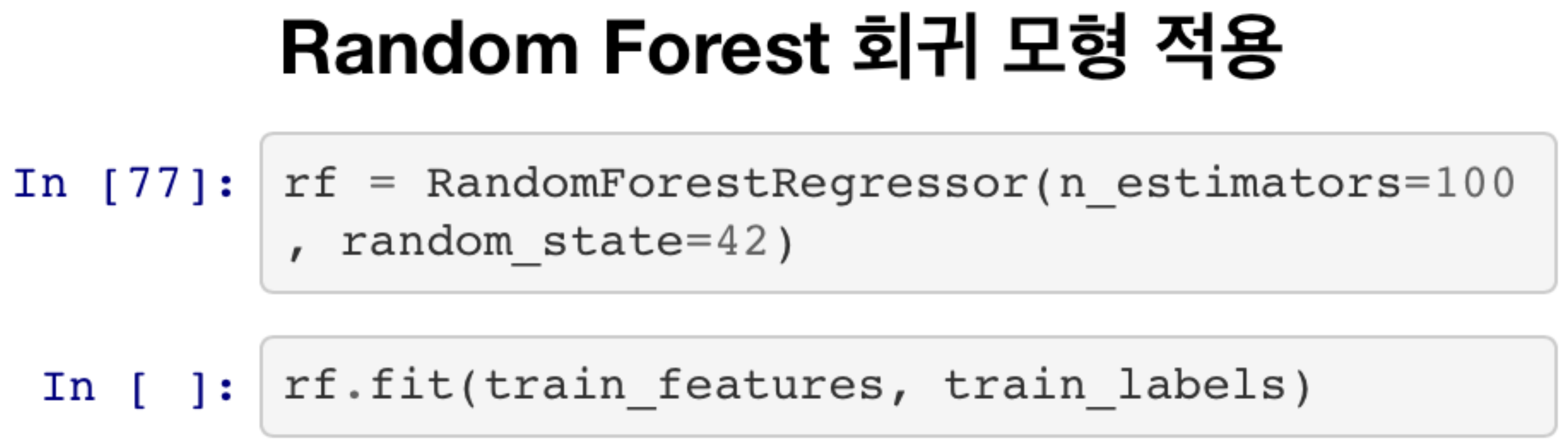


```
In [70]: train.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1458644 entries, 0 to 1458643
Data columns (total 25 columns):
 #   Column                      Non-Null Count    Dtype
---  ------                      --------------    -----
 0   id                          1458644 non-null  object
 1   vendor_id                   1458644 non-null  int64
 2   pickup_datetime             1458644 non-null  datetime64[ns]
 3   dropoff_datetime            1458644 non-null  datetime64[ns]
 4   passenger_count             1458644 non-null  int64
 5   pickup_longitude            1458644 non-null  float64
 6   pickup_latitude             1458644 non-null  float64
 7   dropoff_longitude           1458644 non-null  float64
 8   dropoff_latitude            1458644 non-null  float64
 9   store_and_fwd_flag          1458644 non-null  object
 10  trip_duration               1458644 non-null  int64
 11  pickup_date                 1458644 non-null  object
 12  pickup_day                  1458644 non-null  int64
 13  pickup_hour                 1458644 non-null  int64
 14  pickup_day_of_week          1458644 non-null  object
 15  dropoff_date                1458644 non-null  object
 16  dropoff_day                 1458644 non-null  int64
 17  dropoff_hour                1458644 non-null  int64
 18  dropoff_day_of_week         1458644 non-null  object
 19  pickup_latitude_round3      1458644 non-null  float64
 20  pickup_longitude_round3     1458644 non-null  float64
 21  dropoff_latitude_round3     1458644 non-null  float64
 22  dropoff_longitude_round3    1458644 non-null  float64
 23  trip_distance               1458644 non-null  float64
 24  trip_duration_in_hour       1458644 non-null  float64
dtypes: datetime64[ns](2), float64(10), int64(7), object(6)
memory usage: 278.2+ MB
```

Random Forest Regressor?

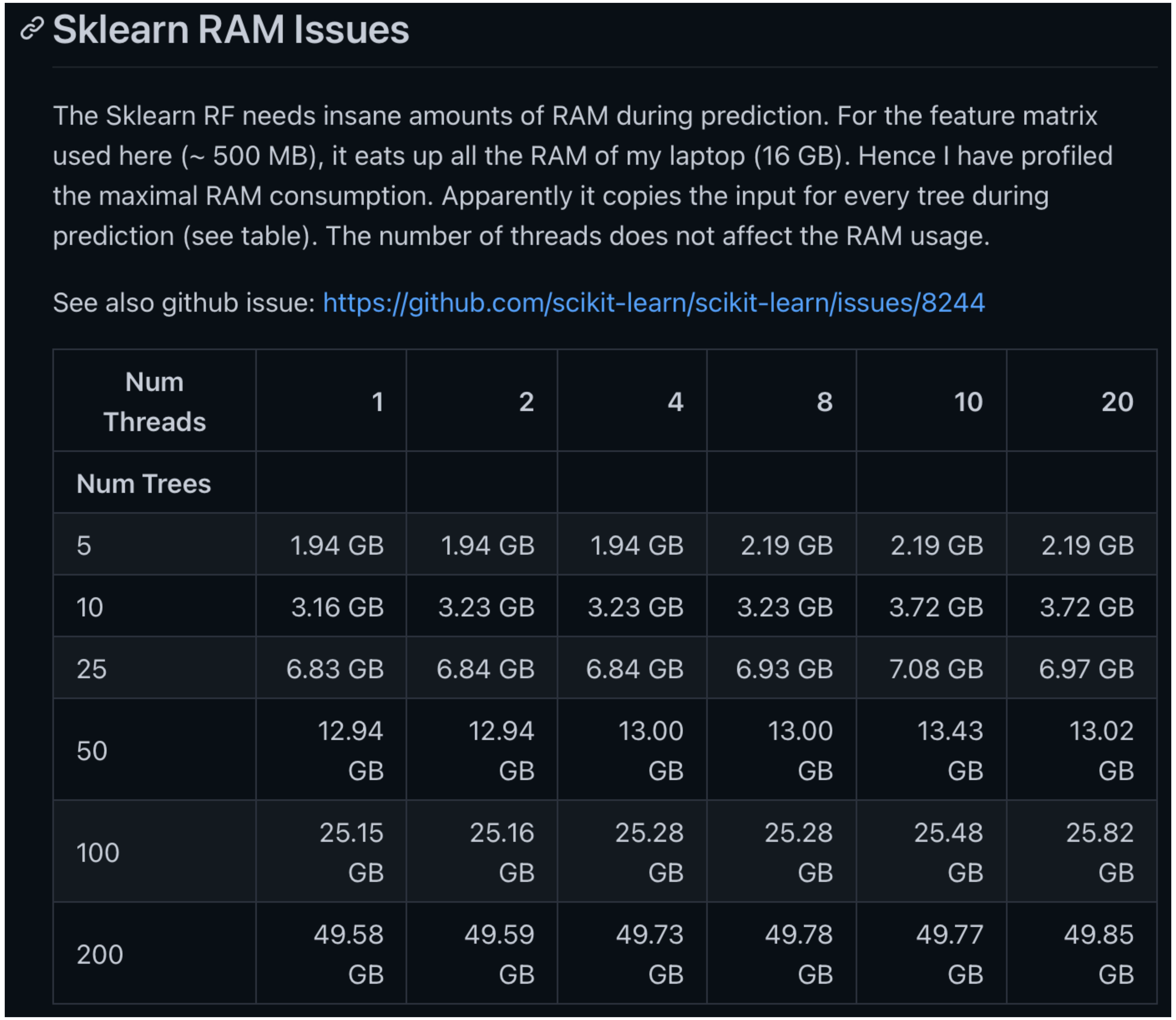
- RAM 사용량이 90% 이상 올라가는 문제가 발생
- 너무 느리고, PC가 터질 것 같다! (포기)



Random Forest 회귀 모형 적용

```
In [77]: rf = RandomForestRegressor(n_estimators=100
         , random_state=42)

In [ ]:  rf.fit(train_features, train_labels)
```



Sklearn RAM Issues

The Sklearn RF needs insane amounts of RAM during prediction. For the feature matrix used here (~ 500 MB), it eats up all the RAM of my laptop (16 GB). Hence I have profiled the maximal RAM consumption. Apparently it copies the input for every tree during prediction (see table). The number of threads does not affect the RAM usage.

See also github issue: https://github.com/scikit-learn/scikit-learn/issues/8244

Num Threads	1	2	4	8	10	20
Num Trees						
5	1.94 GB	1.94 GB	1.94 GB	2.19 GB	2.19 GB	2.19 GB
10	3.16 GB	3.23 GB	3.23 GB	3.23 GB	3.72 GB	3.72 GB
25	6.83 GB	6.84 GB	6.84 GB	6.93 GB	7.08 GB	6.97 GB
50	12.94 GB	12.94 GB	13.00 GB	13.00 GB	13.43 GB	13.02 GB
100	25.15 GB	25.16 GB	25.28 GB	25.28 GB	25.48 GB	25.82 GB
200	49.58 GB	49.59 GB	49.73 GB	49.78 GB	49.77 GB	49.85 GB

https://github.com/constantinpape/rf_benchmarks#sklearn-ram-issues

Random Forest의 대안?

- 방법 1 : n_estimators 파라미터를 적절히 조절한다
- 방법 2 : Support Vector Machine 을 사용할 수도 있다

Support Vector Regressor

- 회귀식 추정 이후 +- ϵ 만큼의 마진을 생성 (상한선, 하한선)
  - 마진 안에 값이 있다면 loss function의 penalty : 0
  - 마진 밖에 값이 있다면 loss function의 penalty : C
- 마진 안에 가능한 많은 샘플이 포함되도록 학습하는 것이 목표
- sklearn.svm의 SVR 을 이용하여 학습 가능