산업통상자원부 공공데이터 활용

비즈니스아이디어 공모전 빅데이터 분석 결과 보고서

팀장 : 전가영

팀원 : 이재원, 권민지

1. **주제**

: 특정품목, 특정국가의 데이터를 분석하여 차년도(2019)에 해당국가가 해당품목을 한국에서 얼만큼 수입할지 예측하는 과제.

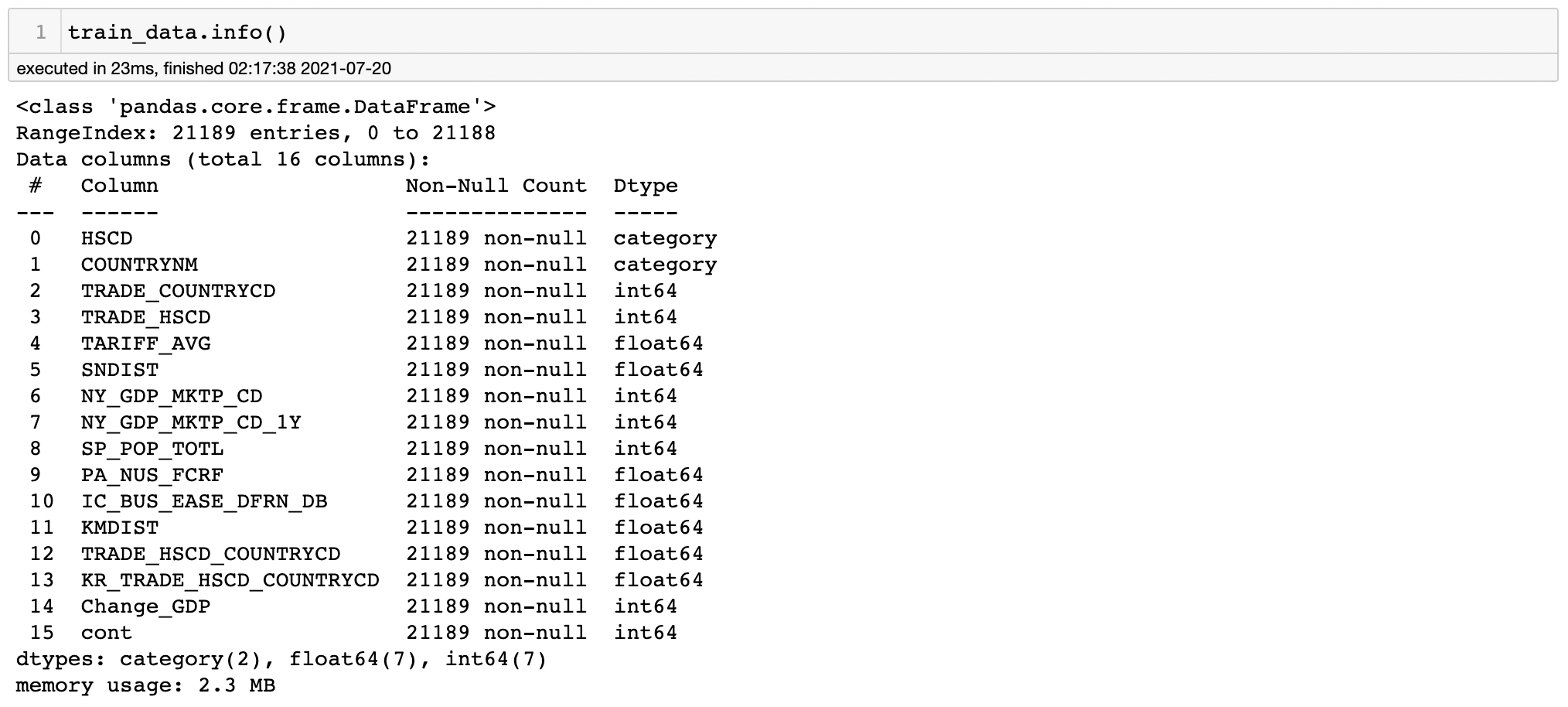
1. **배경 및 필요성**

: 현재 해외 각국으로의 개별 품목의 수출 현황을 분석하여 수출액에 영향을 줄 수 있는 여러 요인을 밝혀 수출량을 늘릴 수 있는 방향성을 제시함

1. **분석 내용 및 분석 결과**
   1. EDA
      1. 기술통계량 및 상자그림
         1. 데이터 타입

총 16개의 변수, category는 범주형, int64는 정수형, float64는 실수형 변수를 의미

범주형 변수는 2개, 정수형 변수는 7개, 실수형 변수는 7개가 있음을 알 수 있다.



* + - 1. 기술통계량 및 상자그림

기술통계량으로 평균, 표준편차, 최소값, 중앙값, 최댓값을 나타내었다.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **TRADE\_COUNTRYCD** | **TRADE\_HSCD** | **TARIFF\_AVG** | **SNDIST** | **NY\_GDP\_MKTP\_CD** |
| **Mean** | 3.424906e+11 | 1.516928e+10 | 3.271472 | 6368.110474 | 1.644687e+12 |
| **Sd** | 4.639194e+11 | 3.915677e+10 | 8.699236 | 2518.061371 | 3.433912e+12 |
| **Min** | 4.337305e+09 | 2.436821e+08 | 0 | 1172.047241 | 1.142576e+10 |
| **Median** | 2.132153e+11 | 4.849947e+09 | 0 | 5624.930017 | 4.453453e+11 |
| **max** | 2.405277e+12 | 3.937064e+11 | 515 | 15134.164110 | 1.951935e+13 |

|  |
| --- |
| TRADE\_COUNTRYCD |
|  |
| TRADE\_HSCD |
|  |
| TARIFF\_AVG |
|  |
| SNDIST |
|  |
| NY\_GDP\_MKTP\_CD |
|  |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **NY\_GDP\_MKTP\_CD\_1Y** | **SP\_POP\_TOTL** | **PA\_NYS\_FCRF** | **IC\_BUS\_EASE\_DFRN\_DB** |
| **Mean** | 1.547985e+12 | 1.225087e+08 | 1993.842026 | 70.723435 |
| **Sd** | 3.257228e+12 | 2.844419e+08 | 6568.366701 | 10.562132 |
| **Min** | 1.118673e+10 | 3.113779e+06 | 0.303350 | 42.671390 |
| **Median** | 4.179836e+11 | 3.797483e+07 | 7.793250 | 73.394083 |
| **max** | 1.871496e+13 | 1.386395e+09 | 33226.298150 | 87.166330 |

|  |
| --- |
| NY\_GDP\_MKTP\_CD\_1Y |
|  |
| SP\_POP\_TOTL |
|  |
| PA\_NYS\_FCRF |
|  |
| IC\_BUS\_EASE\_DFRN\_DB |
|  |

|  |  |  |  |
| --- | --- | --- | --- |
|  | **KMDIST** | **TRADE\_HSCD\_COUNTRYCD** | **KR\_TRADE\_HSCD\_COUNTRYCD** |
| **Mean** | 7951.616743 | 3.083345e+08 | 1.793963e+07 |
| **Sd** | 4054.057558 | 2.130044e+09 | 4.810033e+08 |
| **Min** | 955.651062 | 7.630000e+02 | 0.000000e+00 |
| **Median** | 8262.062500 | 4.082173e+07 | 3.496490e+05 |
| **max** | 18375.181640 | 1.130730e+11 | 6.369533e+10 |

|  |
| --- |
| KMDIST |
|  |
| TRADE\_HSCD\_COUNTRYCD |
|  |
| KR\_TRADE\_HSCD\_COUNTRYCD |
|  |

* + 1. 상관계수 Heatmap



2017년도 데이터셋의 모든 변수들에 대한 상관계수 히트맵 결과이다. 상관계수는 보통 0.6 이상이면 강한 선형상관관계를 갖는다고 할 수 있다.

히트맵을 살펴보면 예측해야 할 타겟변수인 ‘KR\_TRADE\_HSCD\_COUNTRYCD’와의 큰 선형상관관계를 갖는 변수는 없음을 알 수 있다.

* + 1. 결측치 처리
       1. SNDIST(해당 국가와 수입 국가 간 평균 거리)

결측이 있는 국가들의 수입 국가와의 평균거리를 산출해낼 정보가 부족하므로 해당 국가와 지리적으로 가장 근접한 나라 또는 가장 근접한 두 나라의 평균값을 사용하여 결측을 처리하였다.

* + - 1. TARIFF\_AVG(해당 국가에서 해당 품목에 적용되는 평균 관세율)
* 통합무역정보서비스 (<http://tradenavi.or.kr>)홈페이지에서, 해당 국가의 해당품목의 HS-CODE를 직접 검색하여 찾아낸 정보에 대해서 직접 결측값을 채워 넣었다. 홈페이지에서 제공하는 세율은 2020년 기준의 세율이므로 2017년의 세율과 오차가 있겠지만 2017년도 기준으로 FTA체결 여부를 확인하여 값을 채워 넣었다. 2017년도에 FTA가 체결되지 않은 나라들에 대해서는 해당국가의 해당품목에 적용되는 기본세율로 산출하였다.

예측용 데이터셋에 대해서는 마찬가지로 2018년을 기준으로 FTA세율 또는 기본세율로 대체하였다.

* ‘999999’품목에 대해서는 정보가 제공되지 않으므로 ‘0’으로 처리하였다.
* ‘382499’품목은 많은 국가에서 정보가 제공되지 않았다. 따라서 382499품목의 전세계 평균세율을 환산하여 대체하였다.
  + - 1. PA\_NUS\_FCRF(공식환율, 미국 달러에 대한 현지통화단위, 월평균을 기준으로 한 연평균)

공식 환율 결측치에 대한 국가를 출력해보니 독일, 오스트리아, 벨기에, 프랑스, 이탈리아, 네덜란드, 스페인으로 유럽국가 중 7개국에 대해 결측이 있었다. 이는

<https://www.exchangerates.org.uk> 홈페이지를 통하여 2017년도에는 0.8865, 2018년도에는 0.8475의 환율로 처리하였다.

* + - 1. TRADE\_HSCD\_COUNTRYCD(해당연도 해당국가의 해당품목 수입금액)
* 쿠웨이트 : UNcomtrade홈페이지에서 실제 값을 추출할 수 있었다.
* 쿠웨이트를 제외한 나머지 나라들에 대해서는 정확한 값을 찾아내기가 어려워, 해당국가에서 결측인 품목을 제외하고 해당품목의 전세계 총 수입금액(TRADE\_COUNTRYCD)대비 해당국가의 해당품목 수입금액

(TRADE\_HSCD\_COUNTRYCD)의 평균 비율을 각 해당품목의 전세계 총 수입금액(TRADE\_HSCD)에 곱하여 산출하였다.

수식으로 나타내면 MEAN(TRADE\_HSCD\_COUNTRYCD/ TRADE\_COUNTRYCD) \* TRADE\_HSCD

* + 1. 변수생성

전체 훈련용 데이터셋을 사용하여 하나의 예측모델을 사용하는 것 보다, 대륙을 구성하는 국가들 간의 무역적 유사성을 고려하여 7개의 대륙 별로 예측모형을 설정하는 것이 예측력 향상에 더 좋다고 판단하여 7개의 대륙 별 예측모델을 설정하였다. 따라서 43개의 국가들에 대한 대륙변수인 ‘CONTINENT’를 생성하였다.

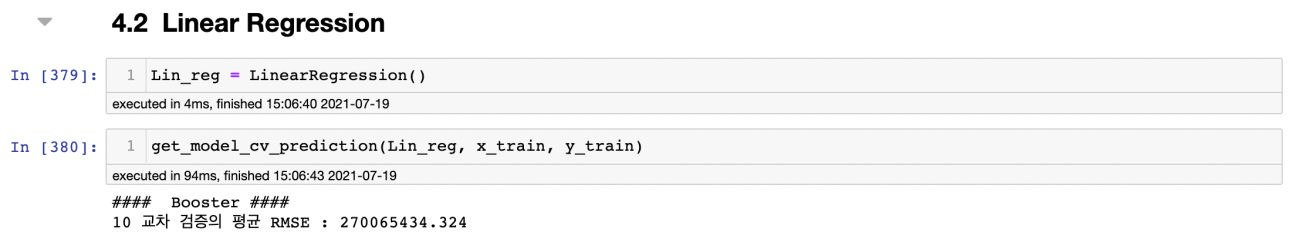
|  |  |  |
| --- | --- | --- |
| **대륙** | **나라** | **CONTINENT** |
| 아시아 | 미얀마, 베트남, 일본, 중국, 홍콩, 인도, 필리핀,  스리랑카, 싱가포르, 몽골, 말레이시아, 인도네시아,  카자흐스탄 | 1 |
| 중동 | 사우디아라비아, 터키, 아랍에미리트,  이란, 쿠웨이트 | 2 |
| 유럽 | 오스트리아, 프랑스, 이탈리아, 벨기에, 독일, 체코,  네덜란드, 노르웨이, 폴란드, 스페인, 영국,  스위스, 러시아 | 3 |
| 아프리카 | 알제리, 이집트, 남아프리카공화국 | 4 |
| 오세아니아 | 호주, 뉴질랜드 | 5 |
| 북아메리카 | 미국, 캐나다 | 6 |
| 남아메리카 | 멕시코, 브라질, 칠레, 과테말라, 페루 | 7 |

* 1. 모형 선택

최적의 모델을 찾아내기 위하여, 각 분석 기법의 최적의 파라미터들을 찾는 GridSearchCV를 통하여 최적파라미터를 찾았다. 2017년도의 훈련용 데이터셋을 10교차검증하여 평균RMSE값을 모형을 선택하는 평가지표로 선택하였다.

* + 1. Linear Regression(선형회귀)

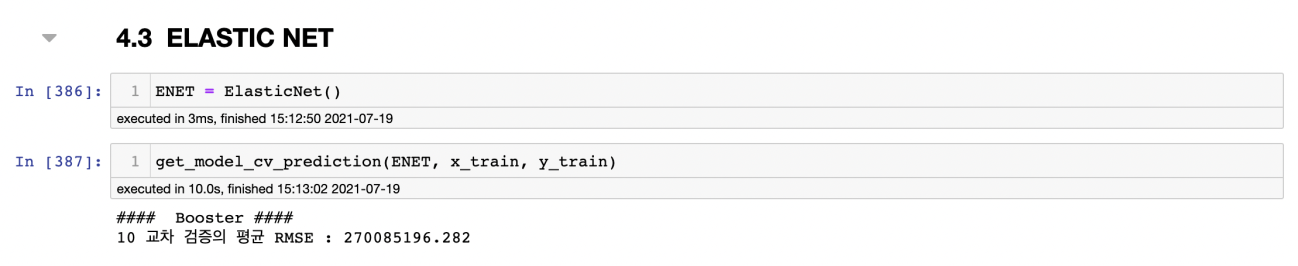
LinearRegression은 지정할 파라미터가 없는 함수이므로, 기본 모델함수를 사용하였다.



그 결과 10교차검증의 평균 RMSE는 약 270065434로 측정되었다.

* + 1. ElasticNet Regression(엘라스틱넷 회귀)

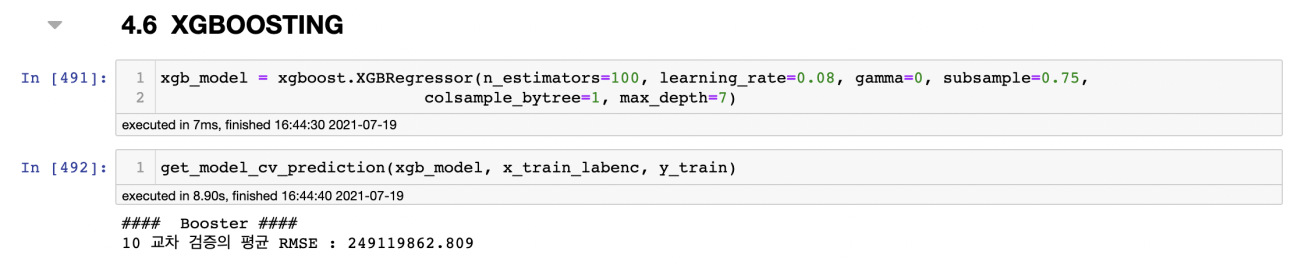
릿지와 라쏘회귀는 선형회귀에 모델의 가중치를 각각 다른 방식으로 규제한다. 엘라스틱넷은 이러한 릿지와 라쏘를 결합한 회귀 모형이다. 엘라스틱넷의 파라미터들에도 특별한 옵션을 주지 않고 모델을 지정하였다.



그 결과 엘라스틱넷의 10교차검증의 평균 RMSE는 약 270085196으로 측정되었다.

* + 1. XGBoosting

다음으로는 모델의 예측력 향상을 위해 앙상블 기법을 고려하였다. 부스팅 알고리즘은 여러개의 약한 학습기를 순차적으로 결합하여 학습-예측을 반복하면서 잘 예측되지 못한 데이터에 가중치를 부여하여 오류를 개선해 나가는 학습방식이다. XGboost는 이러한 boosting기법의 한 종류로서 이전 모델의 오류를 순차적으로 보완해나가는 방식으로 최종 모델을 형성해간다.



xgboosting기법의 주요 파라미터인 n\_estimator, learning\_rate, gamma, max\_depth는

GridSearchCV를 통한 하이퍼파라미터 튜닝을 통하여 위의 최적 파라미터들을 찾아내었고, 이를 통해 xgboosting모델을 적합 시킨 결과 10교차검증의 RMSE는 249119862로 산출되었다.

* + 1. RandomForest

앙상블 기법은 크게 두 가지로 나뉘는데, 위의 xgboosting기법과 같은 boosting과 bagging이 있다. 랜덤포레스트는 bagging의 대표적인 방법으로, 훈련 데이터셋을 통한 훈련 과정에서 생성한 다수의 결정트리로부터 예측된 값의 평균 예측치로 결론을 얻는 방법이다.

트리를 만들 때 사용될 특성의 개수를 제한 함으로서 각 트리들에 다양성을 부여하며, 각 트리마다 전체 특성들 중 일부만을 고려하여 트리를 생성하는 방식이다. 이때 일부의 특성을 선택하는 방식은 무작위로 선택된다. 이 과정은 우리가 선정한 생성할 나무의 개수를 충족할 때까지 반복된다. 랜덤포레스트 기법은 단계마다 모든 특성들을 고려하지 않지만, 결과적으로 볼 때 모든 특성들을 고려하기 위한 방식이라고 할 수 있다. 이러한 과정의 반복으로 모델의 정확도를 향상시킬 수 있다는 장점을 가지고 있다.



랜덤포레스트의 주요 파라미터인 n\_estimators, max\_depth, max\_features는 GridSearchCV를 통한 하이퍼파라미터 튜닝을 통하여 위의 최적 파라미터들을 찾아내었고, 이를 통해 RandomForest모델을 적합 시킨 결과 10교차검증의 RMSE는 약 223875627로 산출되었다.

|  |  |
| --- | --- |
| **model** | **RMSE** |
| Linear Regression | 270065434.324 |
| ElasticNet Regression | 270085196.282 |
| XGboosting | 249119862.809 |
| RandomForest | **223875627.364** |

위의 결과를 종합해보면, 네 개의 생성모델에 대하여 RMSE가 가장 작은 랜덤포레스트 모델이 가장 이 데이터에 적합하다고 판단하여 최종모형으로 RandomForestRegressor를 사용하였다.

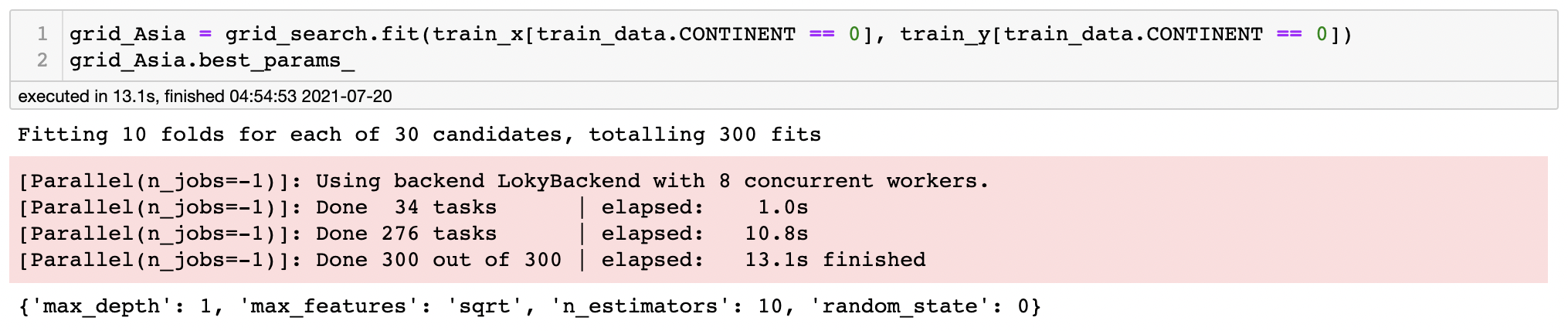
* 1. 최종 모델 분석 과정
     1. 대륙별 GridSearchCV를 통한 파라미터 튜닝

모델의 적합도를 높이기 위해서 하이퍼파라미터를 조정하면서 최적의 성능을 보이는 파라미터를 찾아야 한다. 이를 위해 GridSearchCV함수를 사용하였다. GridSearchCV함수는 교차검증을 통해 미리 설정해둔 하이퍼파라미터의 범위 내에서 최적의 결과를 도출해내는 하이퍼파라미터를 반환하는 함수이다. 이 때 반환된 하이퍼파라미터는 validation error를 최소로 하는 파라미터로 이를 적용한 모델이 새로운 데이터를 이용한 예측에 있어 최선의 결과를 도출하길 기대할 수 있다.

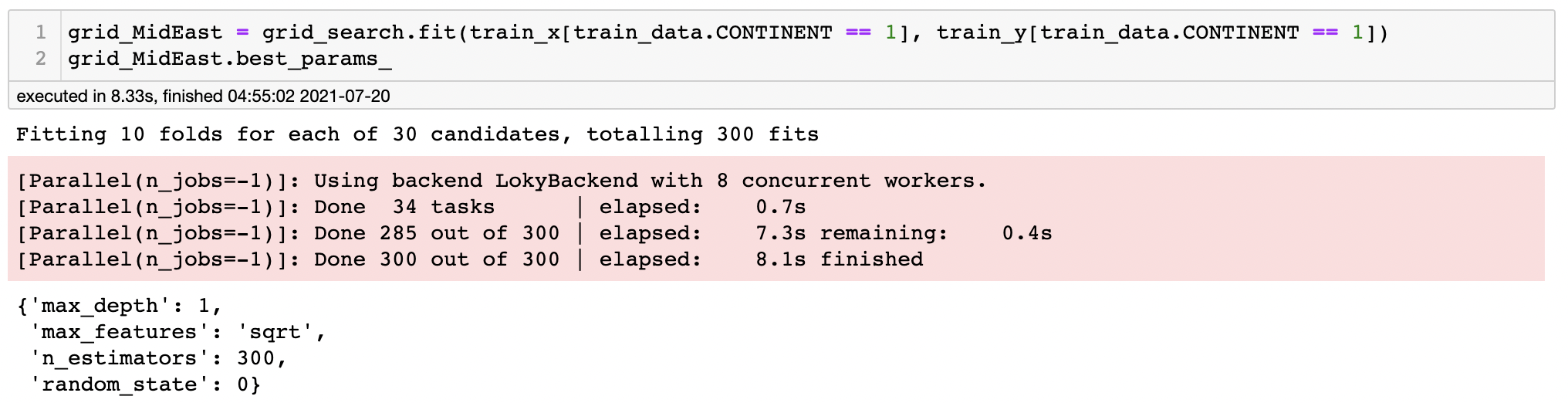
RandomForest의 주요 파라미터인 max\_depth(트리의 깊이), n\_estimators(생성할 결정트리 개수), max\_features(데이터의 여러 특성 중 한 개의 트리를 생성하는데 참조할 특성 개수)에 대해서 GridSearchCV를 시행하였다.

각 대륙에 대한 결과는 다음과 같다.

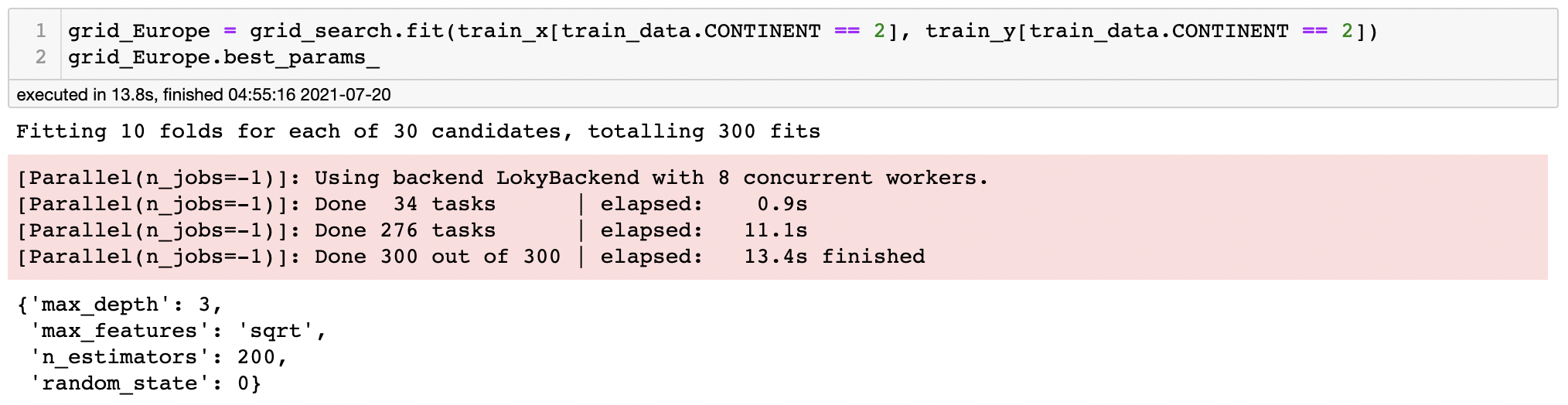
* + - 1. 아시아



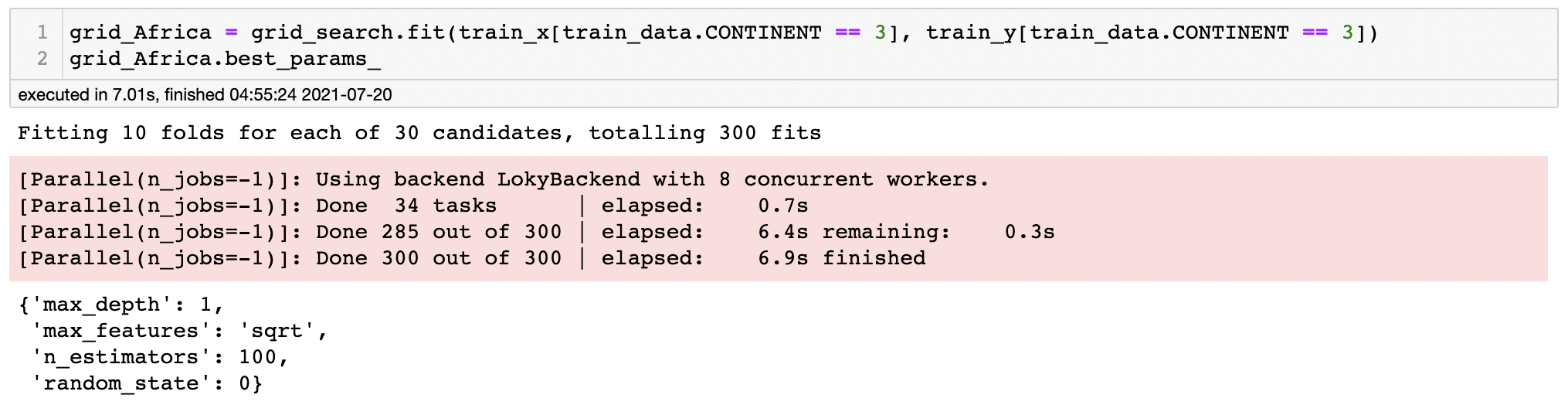
* + - 1. 중동



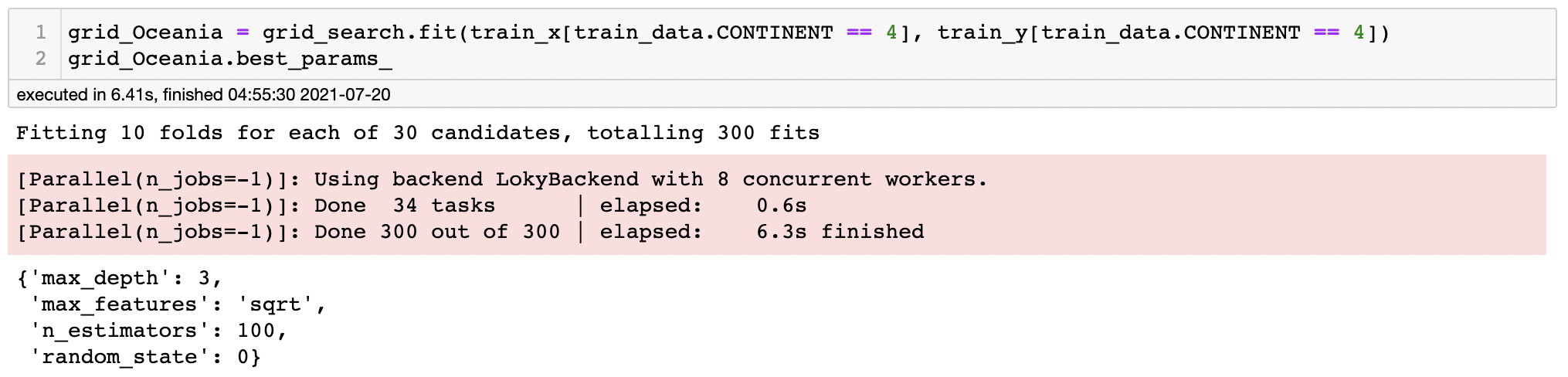
* + - 1. 유럽



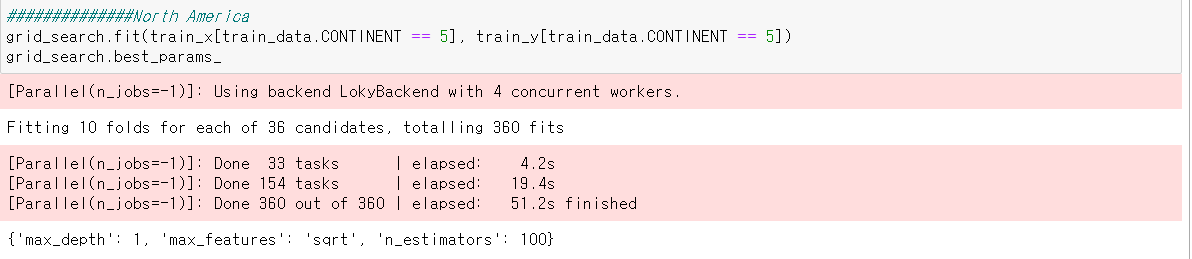
* + - 1. 아프리카



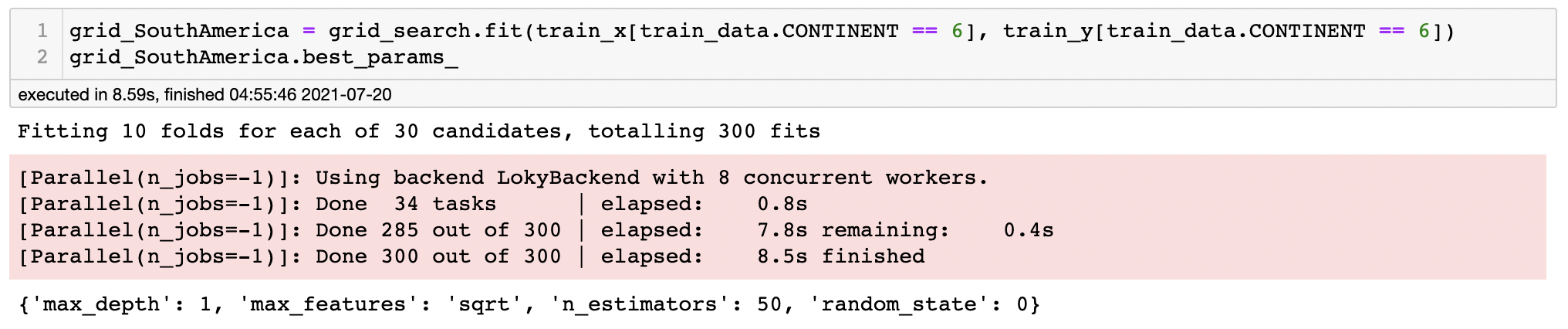
* + - 1. 오세아니아



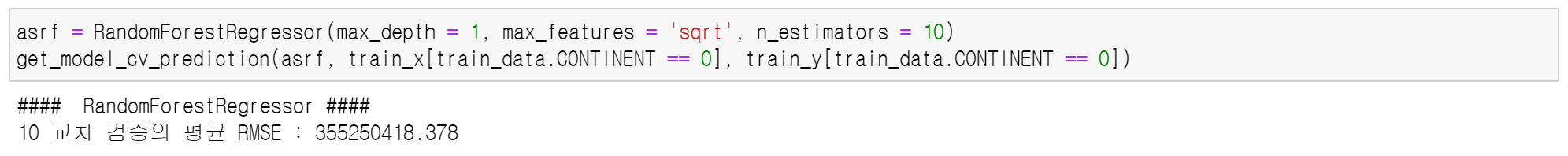
* + - 1. 북아메리카



* + - 1. 남아메리카



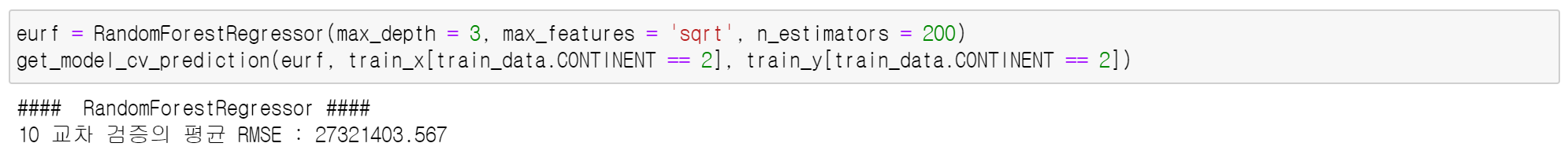
* + 1. 대륙별 모델에 대한 훈련용 데이터셋으로의 10Fold CV 평균 RMSE
       1. 아시아



* + - 1. 중동



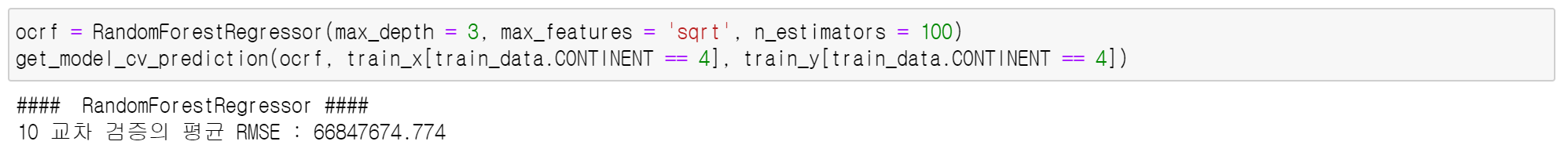
* + - 1. 유럽



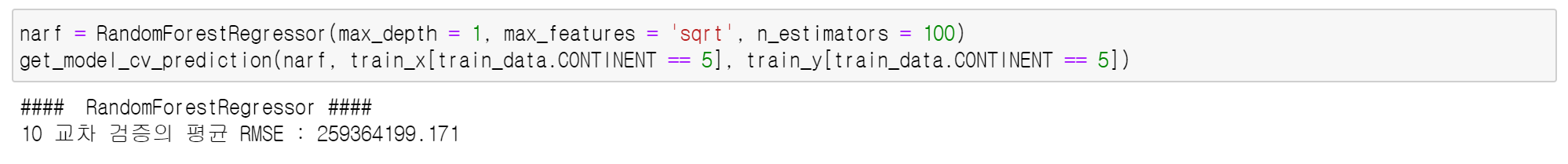
* + - 1. 아프리카



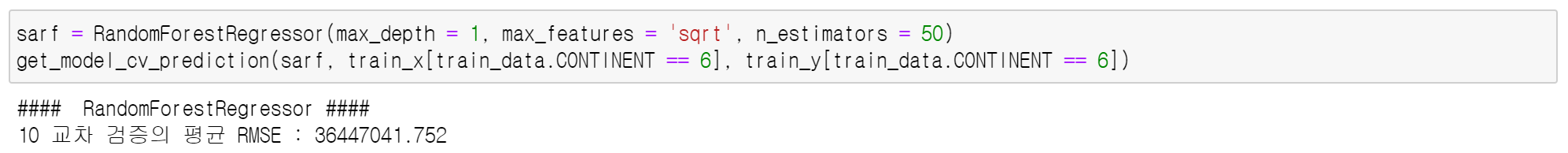
* + - 1. 오세아니아



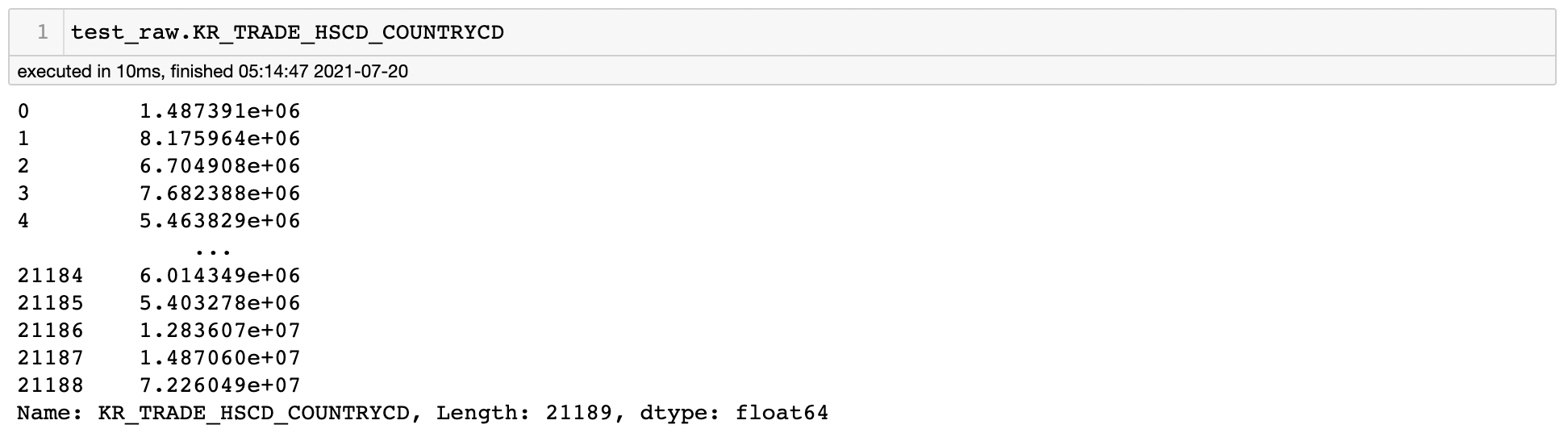
* + - 1. 북아메리카



* + - 1. 남아메리카



* 1. 분석 결과
     1. 예측용 데이터셋을 통한 2019년 ‘KR\_TRADE\_HSCD\_COUNTRYCD’ 예측



모델을 통한 전체 예측값은 엑셀로 첨부하였다.

1. **기대효과**

분석 결과를 활용하여 중소기업이 해외에 자사 제품을 수출하고자 할 때, 해당 국가별 각 품목에 대한 지난해의 수출입 정보를 얻고 수출 대상국과의 관계와 세계시장 정보를 통해 해외 수출에 전략을 새롭게 세울 수 있을 것이다.

1. **활용 데이터**

* TRADE\_COUNTRYCD 결측치 제거 활용 데이터 <https://comtrade.un.org/data>
* 환율 결측치 제거 활용 데이터 <https://www.exchangerates.org.uk>
* 세율 결측치 제거 활용 데이터 <http://tradenavi.or.kr>

1. **참고자료**

(첨부) 국가별 FTA 체결 현황표 <https://www.fta.go.kr/main/situation/kfta/ov/>