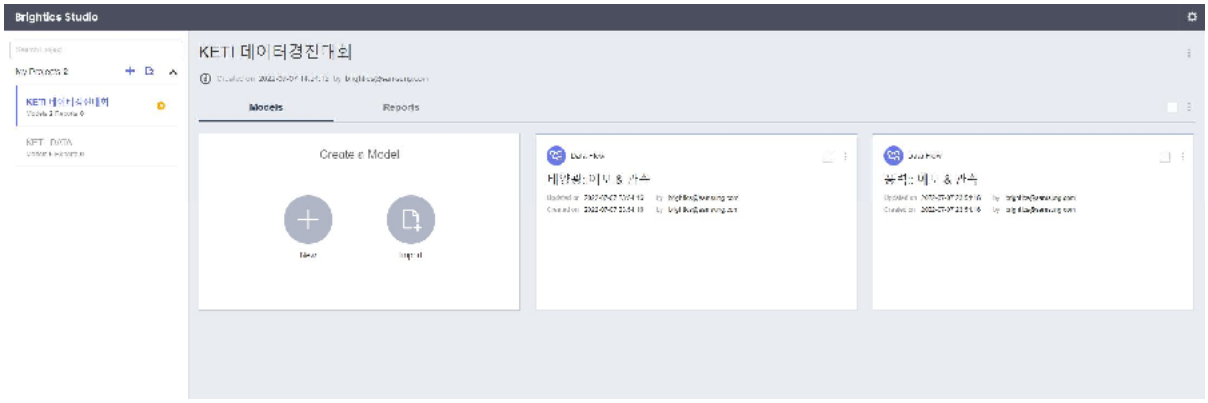
KETI 지속가능한 에너지 활용을 위한 인공지능 경진대회

이 문서는 브라이틱스를 활용하여 KETI 데이터 경진대회에 참여하는 분들을 위한 문서입니다. 이 문서의 내용은 하나의 예시일 뿐입니다. 대회에서 가이드하는 제출 양식만 잘 지킨다면, 이 문서의 분석 방법/순서를 따를 필요는 없습니다.

브라이틱스 스튜디오 모델 설명

제공되는 브라이틱스 프로젝트 파일 'KETI_데이터경진대회.json'을 import 하면 'KETI 데이터경진대회' 프로젝트에 아래처럼 두 개의 모델이 생성됩니다.

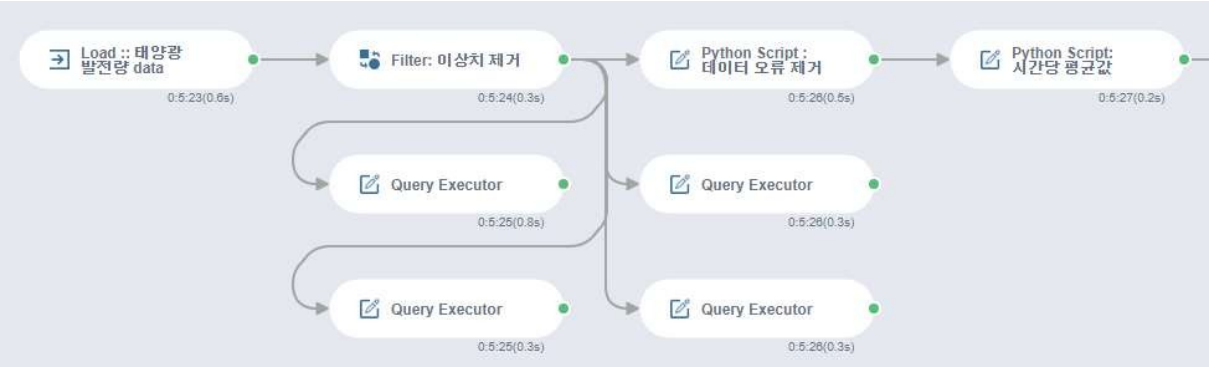


태양광 발전량 예측 문제

태양광 발전량 예측 문제 예시 모델인 '태양광:예보&관측' 모델을 클릭하면 아래와 같은 화면이 열립니다.

발전량 데이터 load 및 EDA

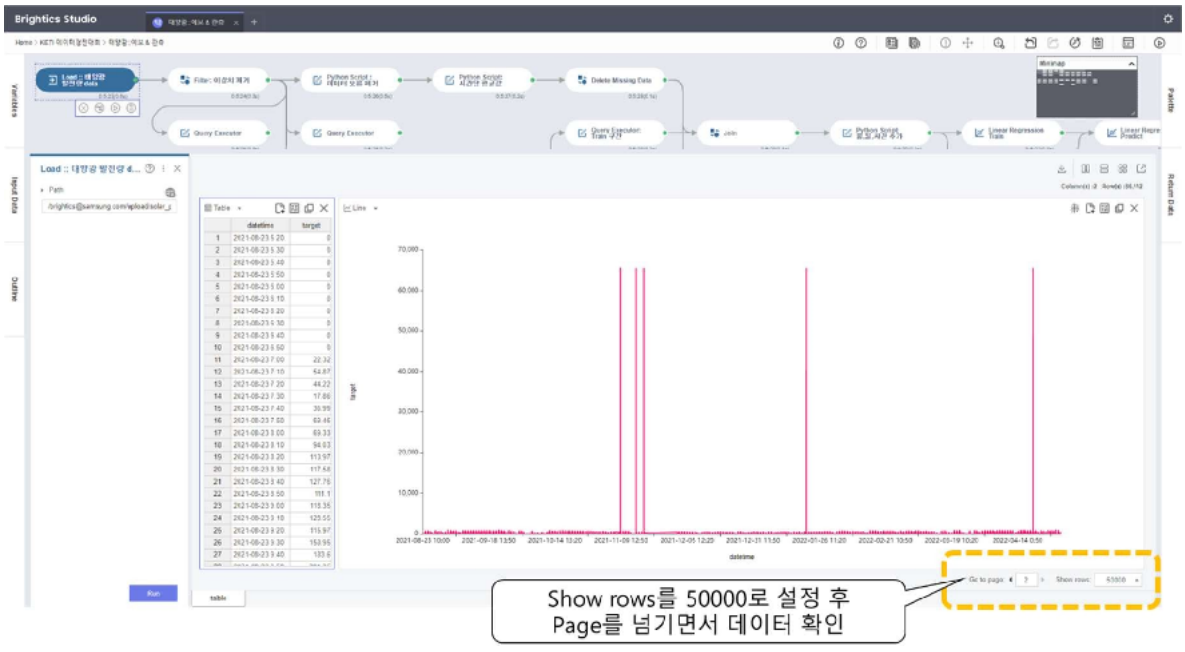
모델의 좌측 상단에 아래처럼 발전량 데이터를 load 하는 부분과 이상치 처리하는 부분이 있습니다.



Load: 태양광 발전량 data

대회에서 제공하는 'solar_power_2204.csv' 파일을 load 합니다.

우측 하단에 Show rows를 조정하면 더 많은 데이터를 한번에 볼 수 있습니다. Page 버튼을 클릭하면서 이상한 데이터가 있는지 확인합니다. 평소보다 지나치게 높게 기록된 데이터가 있습니다.

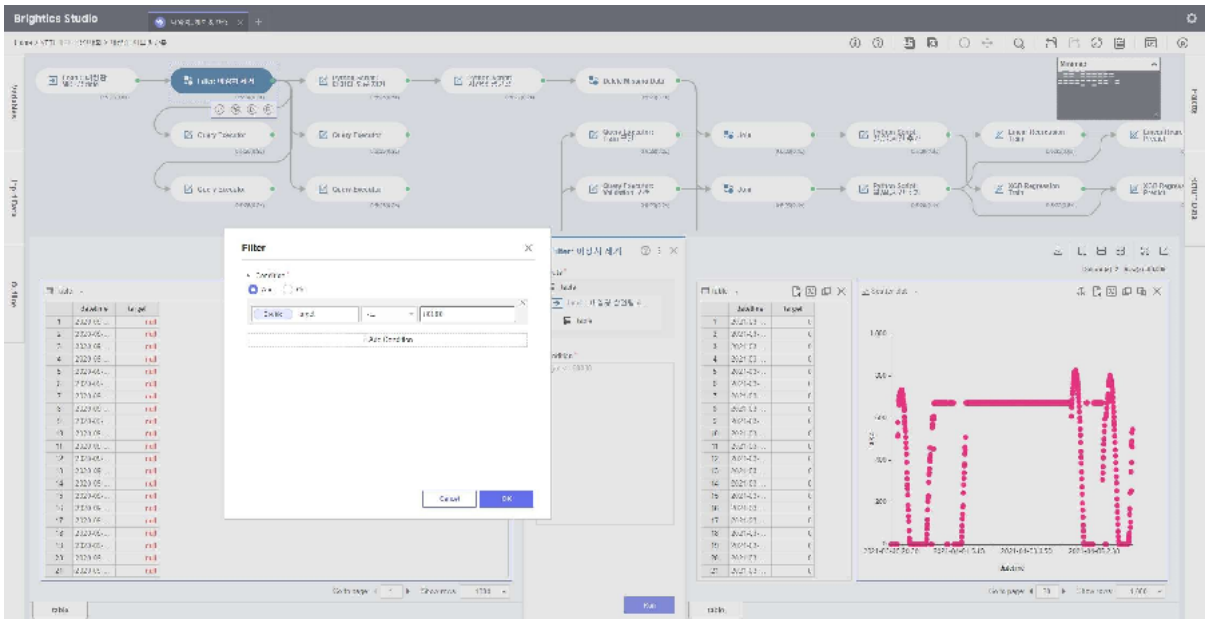


Filter 이상치 제거

브라이틱스는 기본적으로 이상치를 자동으로 찾아 제거하는 기능을 제공합니다. 하지만, 여기서는

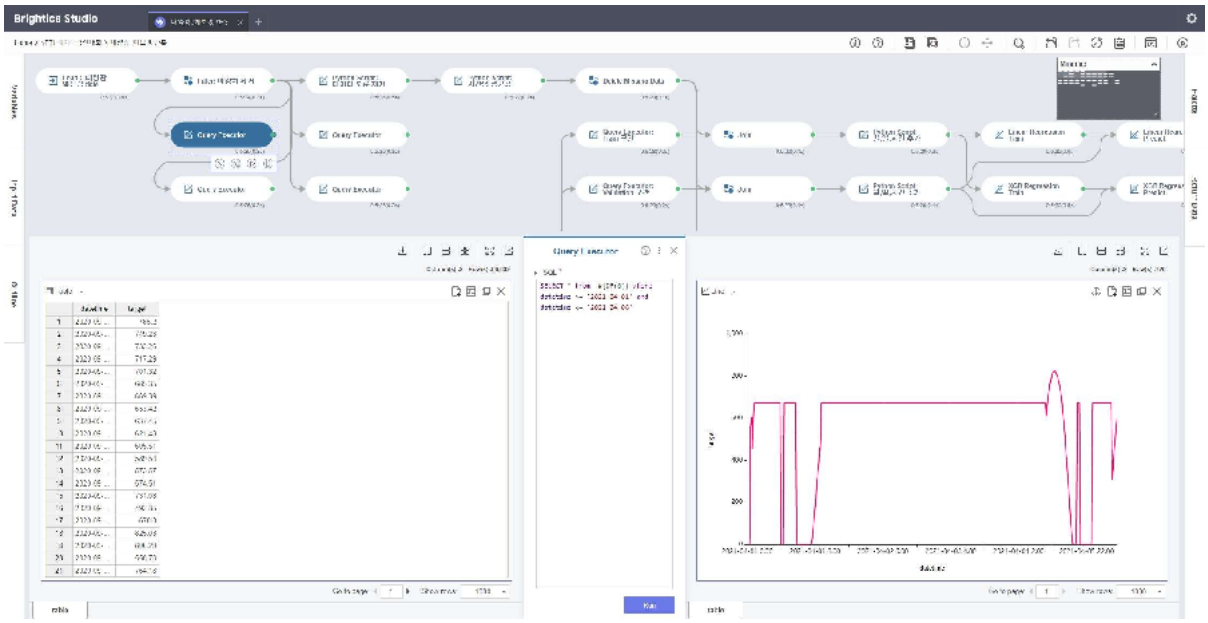
단순히 60,000을 초과하는 값은 이상치로 간주하고 버리기로 했습니다.

아래 캡쳐화면에는 없지만, filter를 통해 이상치(60,000을 초과하는 값)를 제거한 후에도 이상한 데이터가 있는지 살펴보세요. 연속된 값으로 들어오는 구간이 있습니다.



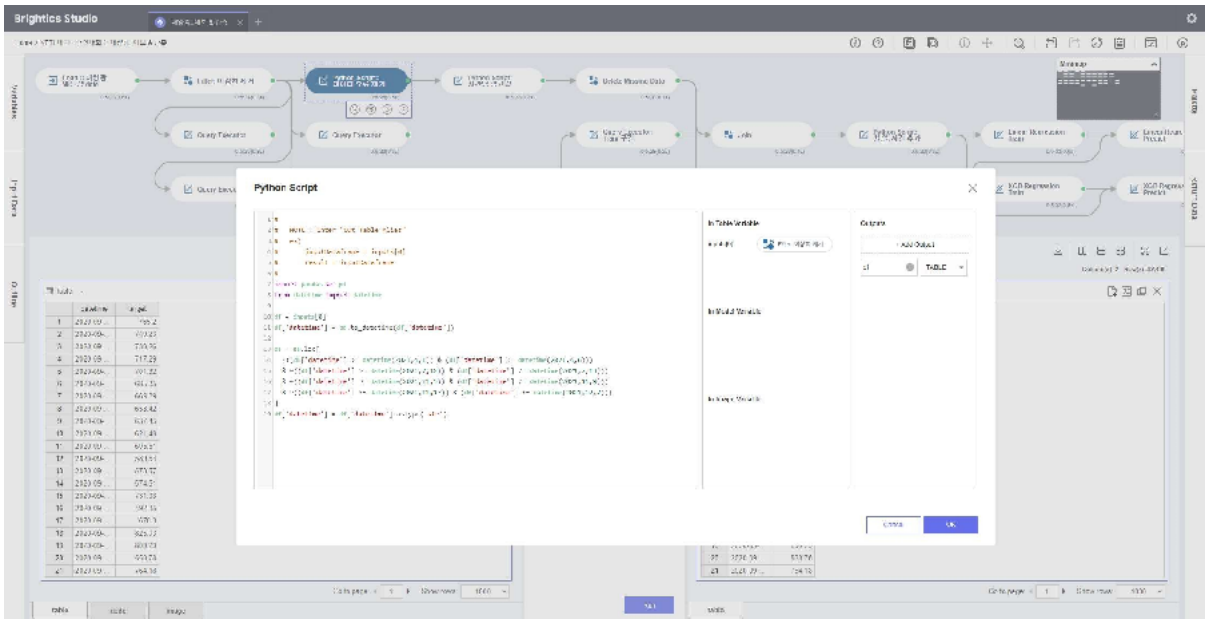
Query Executor

Query Executor를 이용하면, SQL문을 이용해 데이터프레임을 특정 조건에 따라 쉽게 선택할 수 있습니다. 여기에서는 연속된 값을 그래프를 그려 확인하는 용도로 활용했습니다.



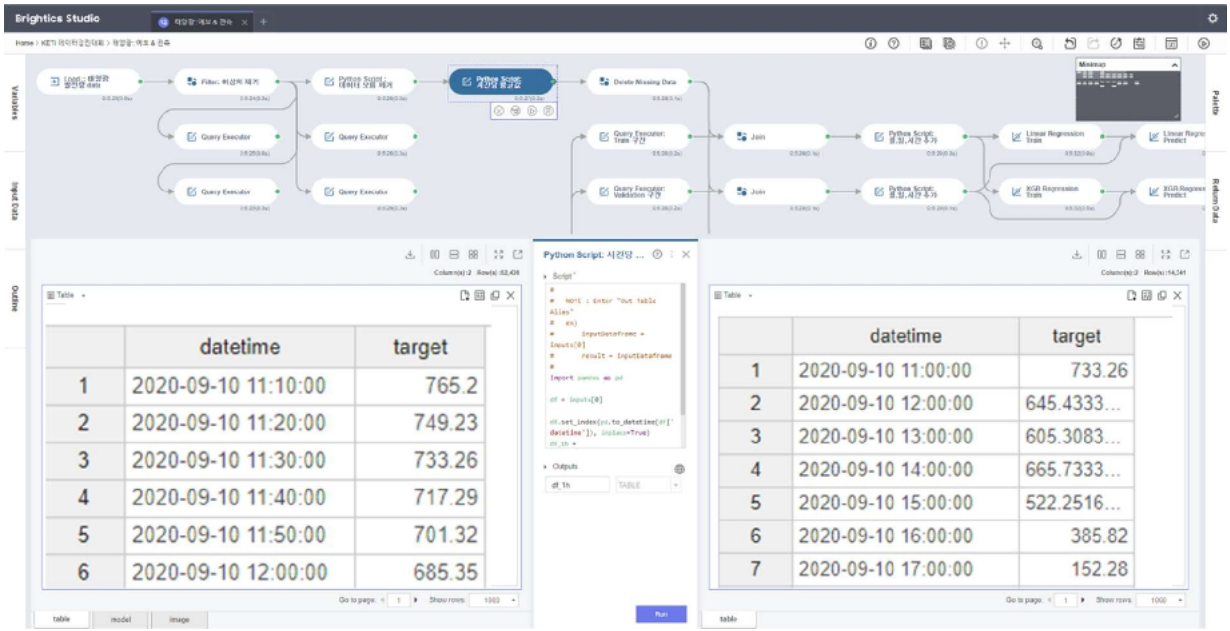
Python Script: 데이터 오류 제거

파이썬 스크립트를 이용해 잘못된 데이터라고 판단되는 구간을 버렸습니다.



Python script: 시간당 평균값

이 문제는 시간당 평균 발전량을 예측하는 문제입니다. 파이썬 스크립트를 활용하여 10분 단위 데이터를 시간당 평균값으로 변경했습니다.

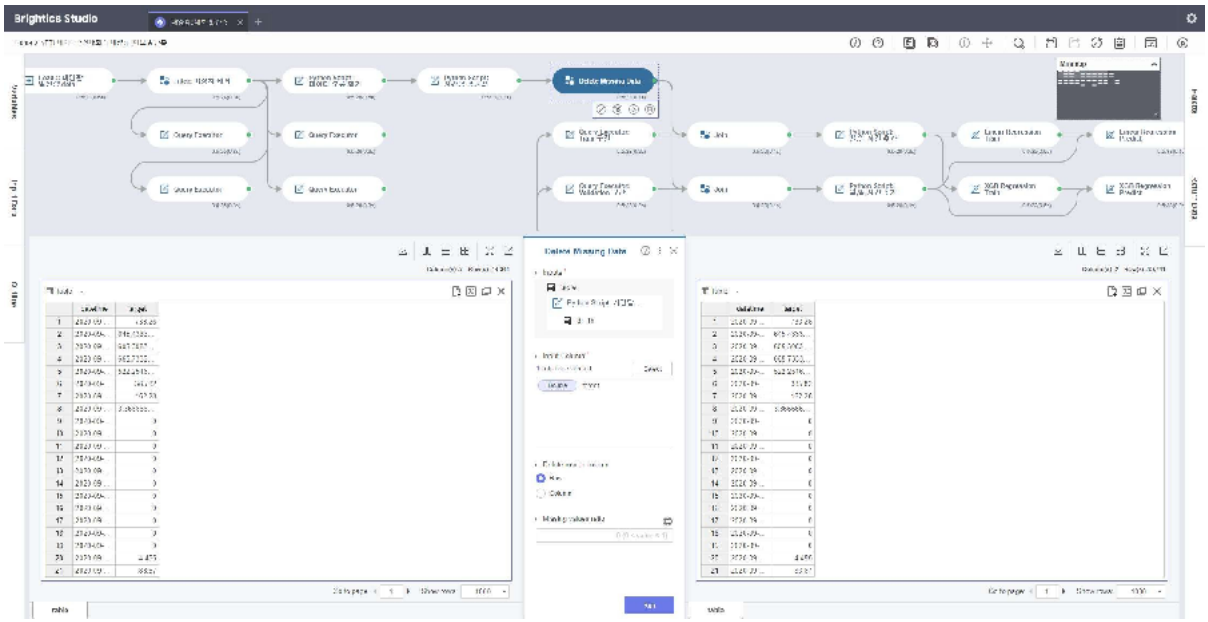


	datetime	target
1	2020-09-10 11:10:00	765.2
2	2020-09-10 11:20:00	749.23
3	2020-09-10 11:30:00	733.26
4	2020-09-10 11:40:00	717.29
5	2020-09-10 11:50:00	701.32
6	2020-09-10 12:00:00	685.35

	datetime	target
1	2020-09-10 11:00:00	733.26
2	2020-09-10 12:00:00	645.4333...
3	2020-09-10 13:00:00	605.3083...
4	2020-09-10 14:00:00	665.7333...
5	2020-09-10 15:00:00	522.2516...
6	2020-09-10 16:00:00	385.82
7	2020-09-10 17:00:00	152.28

Deleting Missing Values

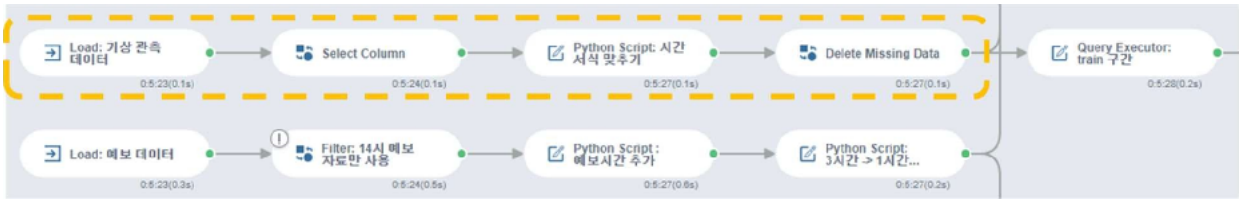
발전량 값이 누락된 경우가 있습니다. 이런 데이터를 삭제합니다.



기상 관측 데이터 Load 하고 서식 맞추기

예측을 하기 위한 인자로 기상 관측자료를 사용하고자 합니다. 제공된 csv 파일(weather_solar_actual.csv)을 로드합니다.

Select Column으로 필요한 컬럼을 선택합니다. 그리고 나중에 발전량 데이터와 시간 단위로 결합할 수 있도록 시간 양식을 파이썬 스크립트로 맞춰줍니다. 마지막으로 관측값이 누락된 행은 삭제합니다.



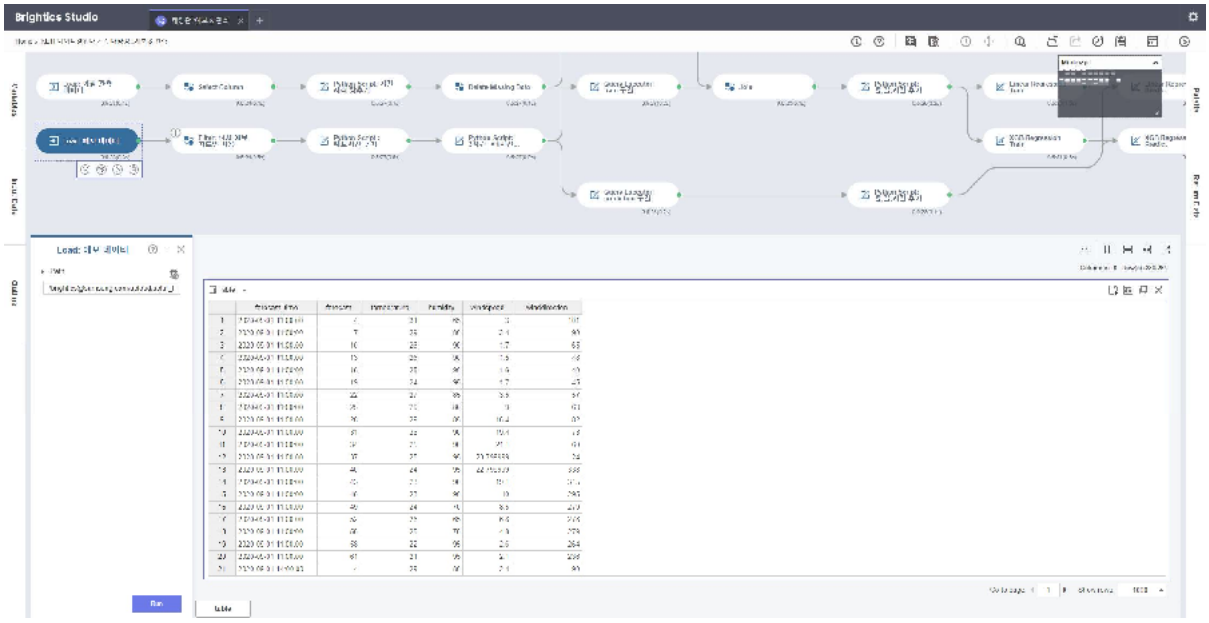
기상 예보 데이터 Load 하고 서식 맞추기

본 문제는 다음날의 발전량을 예측하는 문제입니다. 현재 시점이 4월 30일이라면, 내일인 5월 1일의 발전량을 예측할 때, 5월 1일의 기상 관측값을 사용할 수 없습니다. 따라서 본 대회에서는 예측하는 날짜의 **하루 전 17시 이전에 발표된 자료만 사용**해야 합니다. 참가자께서 본 대회에서 기본적으로 제공하는 데이터 이외의 데이터를 사용할 때, 이 점을 꼭 유의하시기 바랍니다.

Load: 예보 데이터

태양광 발전량 예측을 위해 기본으로 제공해드리는 예보 데이터는 'solar_forecast_weather.csv' 입니다.

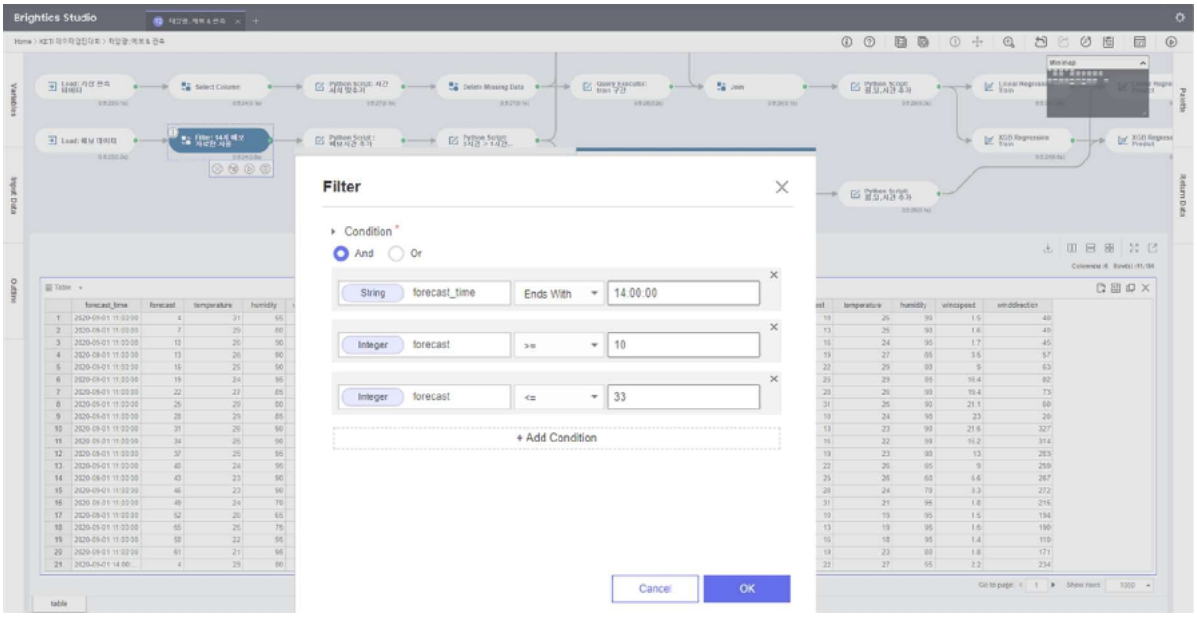
- forecast_time: 예보가 발표된 시각
- forecast: 예보 시점부터 몇시간 뒤를 예측한 것인지
  (예를 들어, forecast_time이 2020-09-01 11:00:00 이고, forecast가 4인 경우, 2020-09-01 15시를 예측한 값입니다)
- temperature, humidity, windspeed, winddirection: 기온, 습도, 풍속, 풍향



Filter로 사용 가능한 예보데이터 선택하기

Filter로 forecast_time이 14:00:00 이면서, forecast가 10 이상이고, 33 이하인 값만 취합니다.

앞서 설명했듯이 예보자료는 전일 **17시 이전 자료만 사용**할 수 있으므로, 17시에 발표된 자료는 쓸 수 없고, 그 중 가장 최신자료인 14시 자료를 사용할 수 있습니다. Forecast를 10~33인 값만 사용한 이유는 다음날 0시(14+10=24)부터 다음날 23시까지 예측하는데 사용하기 위해서입니다.



Python script: 예보시간 추가

파이썬 스크립트를 이용해 forecast_time과 forcast 값을 이용해 datetime으로 예보시간을 만듭니다.

Python script: 3시간 -> 1시간 보간

예보자료는 세시간 단위로 발표됩니다. 우리는 매시간마다 데이터가 필요하므로, 파이썬 스크립트를 이용해 1시간 단위로 보간하여 데이터를 만듭니다.

모델 만들기

우리는 2022년 4월 30일까지의 데이터를 활용해, 2022년 5월과 6월의 발전량을 예측해야 합니다.

더 좋은 방법이 많겠지만, 본 예시에서는 간단히 2022년 2월 28일까지의 데이터를 이용하여 학습하고, 3~4월을 예측하여 실제 발전량과 비교한 뒤, 잘 맞는 모델을 선택하고자 합니다. 예시를 위해 Linear Regression과 XGBoost를 사용했습니다.

우선 Query Executor를 활용해, 모델의 Train과 validation을 위한 구간을 나눕니다. Train을 위한 데이터는 Join 함수를 이용해 기상 관측값과 실제 발전량 데이터를 결합하여 만듭니다. Validation을 위한 데이터는 예보자료와 실제 발전량 데이터를 결합하여 만들었습니다.

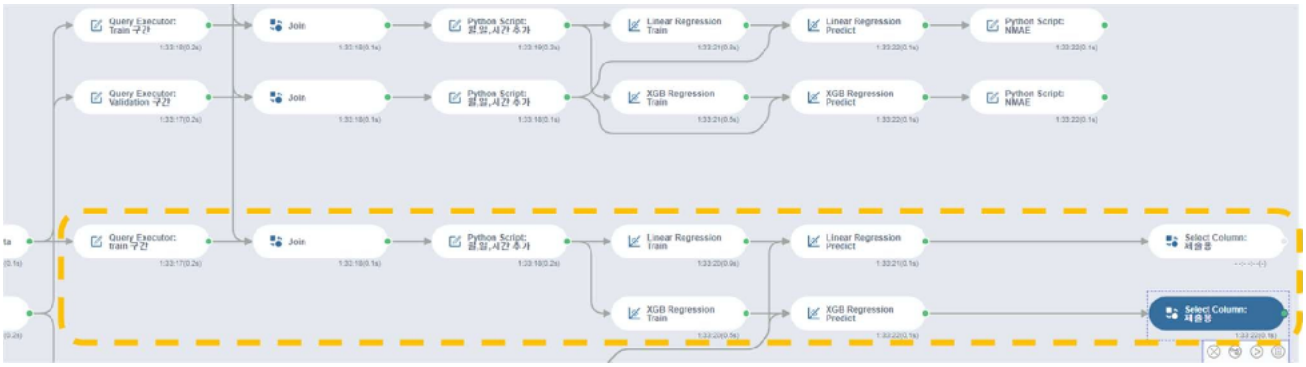
두 데이터셋 모두 파이썬 스크립트를 이용해 월, 일, 시간 컬럼을 추가로 생성했습니다.

그리고 Linear Regression과 XGBoost를 이용해 3~4월의 발전량을 예측했습니다.

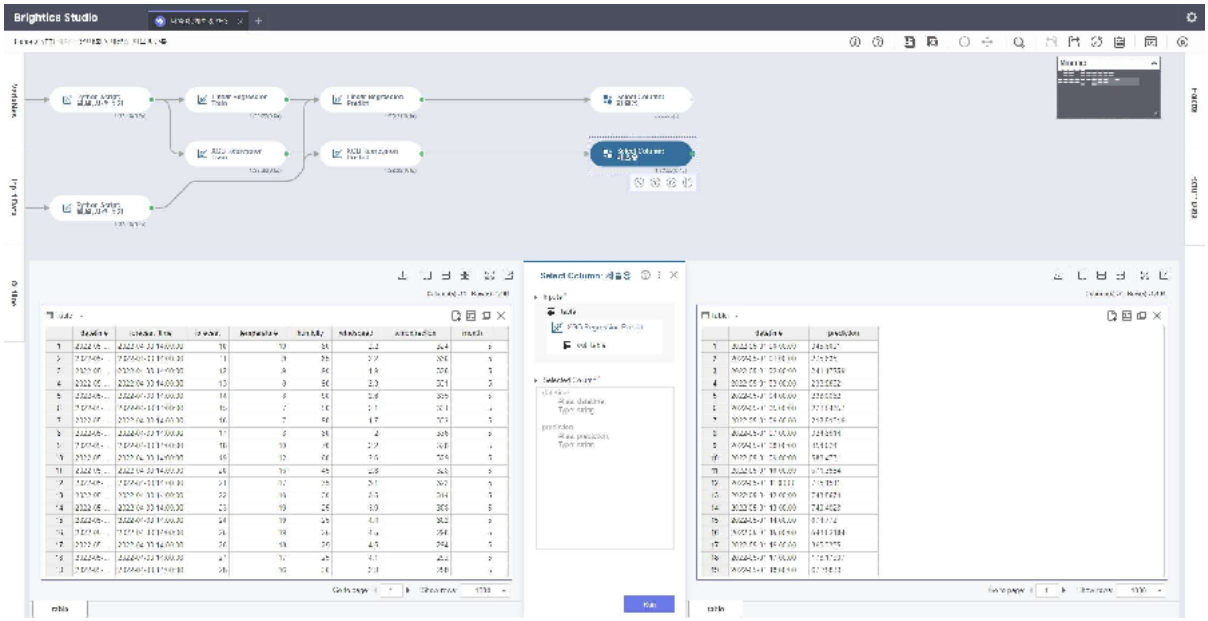
맨 마지막 파이썬 스크립트는 NMAE (Normalized Mean Absolute Error)를 계산하기 위한 함수입니다. 비교 결과, Linear Regression은 16.7%, XGBoost는 12.8%가 나왔습니다. 따라서 XGBoost 모델 을 선택하기로 했습니다.

예측하기. 제출용 CSV 생성하기

5월부터 6월까지 예측하는 문제이므로, 이제 4월 30일까지 학습을 하고, 5월 1일부터 6월 30일까지를 예측하고자 합니다. 모든 내용은 위의 '모델 만들기'의 내용과 동일합니다.

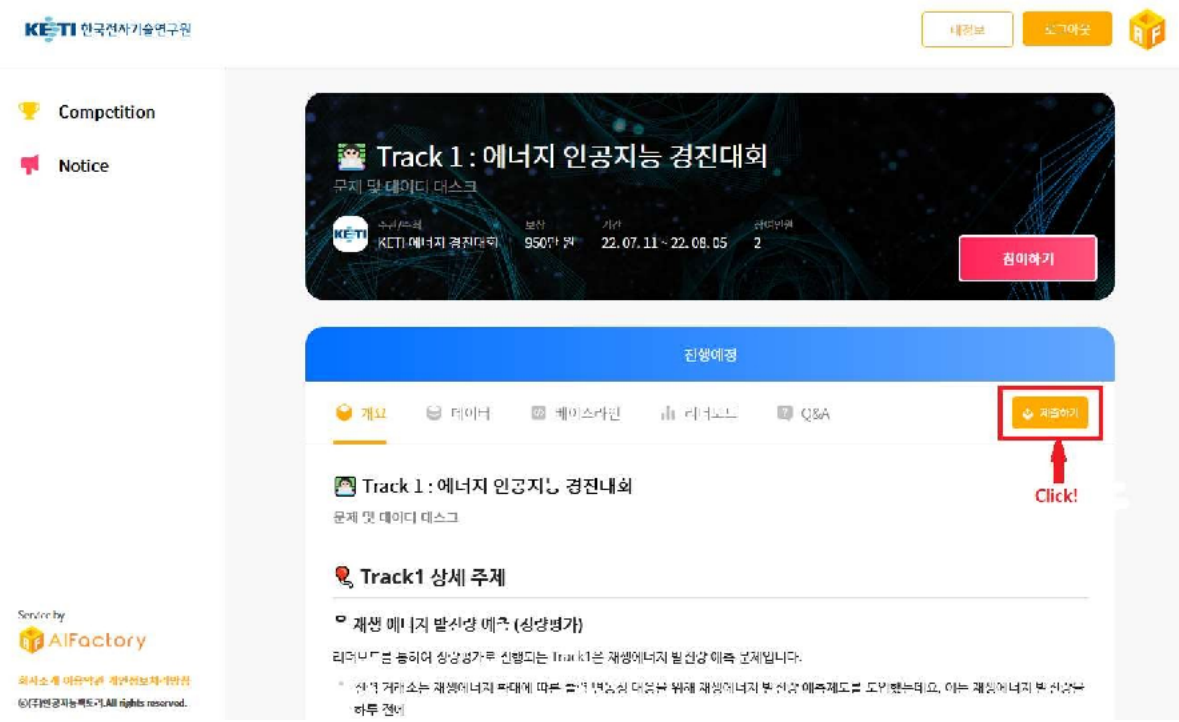


제출을 위한 csv 생성을 위해, 맨 마지막에 Select Column 함수를 활용했습니다. datetime과 prediction 두개 컬럼을 갖는 csv 파일을 생성합니다. 우측 상단에 다운로드 버튼을 클릭하여 각각의 결과를 취득합니다.



대회 페이지에 업로드할 최종 양식은 다음과 같습니다. (answer_sample.csv 참조) 다운로드한 두 csv 결과를 양식에 맞게 하나로 합한 다음 AIFactory의 태스크 페이지의 제출하기를 통하여 리더보드로 제출합니다.

datetime	solar	wind
2022-05-01 0:00	0	0
2022-05-01 1:00	0	0
2022-05-01 2:00	0	0
2022-05-01 3:00	0	0
2022-05-01 4:00	0	0
2022-05-01 5:00	0	0
2022-05-01 6:00	0	0
2022-05-01 7:00	0	0
2022-05-01 8:00	0	0
2022-05-01 9:00	0	0
2022-05-01 10:00	0	0
2022-05-01 11:00	0	0
2022-05-01 12:00	0	0
2022-05-01 13:00	0	0
2022-05-01 14:00	0	0
2022-05-01 15:00	0	0


KETI 한국전자기술연구원
Compctition
Notice
Track 1 : 에너지 인공지능 경진대회
문제 및 데이터 데스크
KETI 에너지 경진대회
950만 원
22.07.11 ~ 22.08.05
2
참여하기
진행예정
개요
데이터
Q&A
제출하기
Click!
Track 1 : 에너지 인공지능 경진대회
문제 및 데이터 데스크
Track1 상세 주제
재생 에너지 발전량 예측 (정량평가)
Service by
AIFactory