**경진대회 코드 설명자료**

**팀 명 : Sevento11 (닉네임: platypusdata)**

1. **브라이틱스 프로젝트 내 모델 설명**

* Sevento11 (닉네임: platypusdata).json 파일을 오픈할 시 보이는 메인 화면과 모델 화면

|  |
| --- |
| 브라이틱스 메인 화면    풍력 테양광 예측 통합 모델 화면     * 편의성을 위하여 기압예측모델, 풍력예측모델, 태양광예측모델을 생성하였음으로 우측 상단에 있는 러닝 버튼 실행 시 결과까지 한번에 도출 가능 |

1. **Input 파일 (활용 데이터) 설명**

* Sevento11\_Load\_data.zip : Load 데이터
* Sevento11\_Raw\_data.zip : 원천데이터
* wind\_power.csv : 풍력 발전량 데이터
* wind\_power 업로드 시 target은 double형으로 업로드

텍스트이(가) 표시된 사진

자동 생성된 설명

* wind\_test.csv : 풍력 test 데이터

텍스트이(가) 표시된 사진

자동 생성된 설명

* pressure\_train.csv : 기압 train 데이터

텍스트이(가) 표시된 사진

자동 생성된 설명

* pressure\_test.csv : 기압 test 데이터

텍스트이(가) 표시된 사진

자동 생성된 설명

* solar\_power.csv : 태양광 발전량 데이터
* solar\_power 업로드 시 target은 double형으로 업로드

텍스트이(가) 표시된 사진

자동 생성된 설명

* solar\_train.csv : 태양광 train 데이터

텍스트이(가) 표시된 사진

자동 생성된 설명

* solar\_test.csv : 태양광 test 데이터

텍스트이(가) 표시된 사진

자동 생성된 설명

|  |  |
| --- | --- |
| **외부 데이터에 대한 설명** |  |
| * 외부 데이터 사용 * 기상청 기상자료개방포털 동네예보 단기예보   - 기간 : 2020.08.01 ~ 2022.06.30  - <https://data.kma.go.kr/data/rmt/rmtList.do?code=420&pgmNo=572>  동면 구예보    동면 신예보    청라3동 구예보    청라3동 신예보     * 공공데이터포털 한국천문연구원\_음양력 정보   - 기간 : 2020.08.29 ~ 2022.06.30  - <https://www.data.go.kr/data/15012679/openapi.do>   * 공공데이터포털 한국천문연구원\_출몰시각 정보   - 기간 : 2020.08.29 ~ 2022.06.30  - <https://www.data.go.kr/data/15012688/openapi.do>   * 기상청 기상자료개방포털 동네예보 단기예보   - 기간 : 2016.08.01 ~ 2022.06.30  - <https://data.kma.go.kr/data/rmt/rmtList.do?code=420&pgmNo=572>  동면 예보 데이터 (예시)     * 기상청 기상자료개방포털 종관기상관측   - 기간 : 2017.05.01 ~ 2021.06.30 (5,6월만)  - <https://data.kma.go.kr/data/grnd/selectAsosRltmList.do?pgmNo=36>  - 광주 기압 데이터     * 외부 데이터 사용 시 원본 설명 * 데이터의 정제/ 통합/ 정리/ 변환 등 데이터 전처리 결과에 대한 내용 서술 | |
| * 동면 구예보 (9 feature) , 동면 신예보(9 feature) 에서 ‘forecast\_time’ 14시 데이터 중 ‘forecast’>=10 & ‘forecast’ <=33인 데이터만 추출, 통합하여 풍력 예보데이타 생성, 음양력 정보와 일출일몰정보를 API로 가져와 풍력 예보데이터에 feature로 ‘lunYear’, ‘lunMonth’, ‘lunDay’, ‘sunrise’, ‘sunset’, ‘sunshine’ 생성 후 날짜를 2022-04-30일까지wind\_train.csv 생성, 2022-05-01~2022-06-30까지 wind\_test.csv 생성 * 청라3동 구예보, 청라3동 신예보도 위와 동일하게 가공하여 solar\_train.csv, solar\_test.csv 생성 * 동면 예보 데이터 2017년~2022년 데이터 통합 후 관측 기압 데이터를 ‘datetime’으로 결합 후 5,6월 데이타만 뽑아서pressure\_train.csv , pressure\_test.csv 생성 | |

1. **탐색적 자료 분석 (Exploratory Data Analysis)**

* 풍력
* Heatmap으로 변수간 상관관계를 파악하여 상관관계가 서로 높은 feature별 조합
* 일정기간 동안 지속적으로 traget값이 0인 구간 삭제
* Target과 상관관계가 가장 높은 풍속을 기준으로, 풍속이 매우 강한데 target값은 0 또는 너무 낮은 값들 제거 (df['windspeed']>8)&(df['target']<=2500))
* 태양광
* Heatmap으로 변수간 상관관계를 파악하여 상관관계가 서로 높은 feature별 조합
* 야간인데 target값이 0이 아닌 값들 0으로 대체(df.loc[((df['hour']<df['sunrise'])|(df['hour']>df['sunset']))&(df['target']>0.0), 'target']=0.0

|  |
| --- |
| * 풍력 (EDA 데이터 시각화)          * 태양광 (EDA 데이터 시각화)   텍스트이(가) 표시된 사진  자동 생성된 설명 |

1. **변수 선택 및 모델 구축 (Feature Engineering & Initial Modeling)**

* **풍력**

|  |  |
| --- | --- |
| **풍력 데이터 feature engineering** |  |
| * Pressure : Random Foresst Predict를 활용하여 기압데이터 생성 * WW : winddirection 과 windspped를 곱한 데이터 생성 * HT : humidity와 temperature을 곱한 데이터 생성 * PP : precipitationprob과 precipitationform을 곱한 데이터 생성 * fTemp : 섭씨온도에서 화씨온도로 바꾸기 위해 공식을 대입한 데이터 생성 * density : 기압 데이터에 화씨온도로 나눈 값 * pt : 기압데이터에 섭씨온도를 곱한 값 * pt\_1 : 기압데이터에 화씨온도를 곱한 값  |  |  | | --- | --- | | **텍스트이(가) 표시된 사진  자동 생성된 설명** | **code** |  * wind4 : 풍향을 4가지의 category 데이터로 변환하는 사용자 함수 | | |

|  |  |
| --- | --- |
| **변수 및 모델 선정** |  |
| * Adaboost Regression 모델 선정 * 최종 feature로 month, day, hour, HT, WW, density, wind04, pt, temperature, pt\_1 * Max Depth : 6 * Number of Estimators : 138 * Learning Rate : 1.0 * Loss : LINEAR * Seed : 22 | |

* **태양광**

|  |  |
| --- | --- |
| **태양광 데이터 feature engineering** |  |
| 텍스트이(가) 표시된 사진  자동 생성된 설명  텍스트이(가) 표시된 사진  자동 생성된 설명   * temp : temperature 데이터를 반올림한 값 * fTemp : 섭씨온도에서 화씨온도로 바꾸기 위해 공식을 대입한 데이터 생성 * HsT: humidity 데이터에서 fTemp를 나눈 값의 데이터 * temp : temperature 데이터를 반올림한 값 * WW : winddirection 과 windspped를 곱한 데이터 생성 * PP : precipitationprob과 precipitationform을 곱한 데이터 생성 * HT : humidity와 temperature을 곱한 데이터 생성 * wind4 : 풍향을 4가지의 category 데이터로 변환하는 사용자 함수 | | |

|  |  |
| --- | --- |
| **변수 및 모델 선정** |  |
| * MLP Regression 모델 선정 * 최종 feature로 hour, lunMonth, lunDay, temperature\_standard, humidity\_standard, cloud\_standard, windspeed\_standard, winddirection\_standard, precipitation\_standard, precipitationprob\_standard, sunset\_standard, sunrise\_standard, WW\_standard 선택 * Hidden Layer Sizes   : Node 수 10 layer 4     * Activitation Function : relu * Solver : adam * Alpha : 0.001 * Batch Size Auto : True * Learning Rate : adaptive * Learning Rate Initial : 0.001 * Max Iteration : 1000 * Seed :22 | |

1. **모델 학습 및 검증 (Model Tuning & Evaluation)**

* 풍력
* AdaBoost Regression 모형을 활용하여 parameter 조절을 통해 검증 셋에서 NMAE 확인
* seed는 22로 고정 후 max depth 6, Estimator 138, learning rate 1,0(default값)의 가중치로 적합하였을 때 NMAE 11.62 확인

|  |  |  |
| --- | --- | --- |
| |  |  | | --- | --- | |  | **code** | |

* 태양광
* MLP Regression 모델 사용
* Hyperparameter로는 hidden layer 층, Learning rate, Max Iteration 및 seed evaluation하여 Hidden Layer Sizes=[10,10,10,10], Learning Rate=0.001, Max Iteration=1000, Seed=22로 고정
* 검증 셋에서 NMAE 확인: 8.7432

|  |
| --- |
| * 태양광   모델    검증  테이블이(가) 표시된 사진  자동 생성된 설명    모델 test  테이블이(가) 표시된 사진  자동 생성된 설명 |

1. **결과 및 결언 (Conclusion & Discussion)**

< 결과 >

* 풍력

- 화순 동면의 기상청 단기예보를 이용하여 14시 기준으로 train data를 생성.

- 데이터 분석을 통해 이상치 확인 후 처리

- 기상청 예보 데이터+기압 관측 데이터를 학습하여 기압 회귀 모델 생성하고 2022년 5,6월 기압 예측하여 feature로 추가.

- 풍력 발전기 발전량에 큰 영향을 미치는 요인 중 하나인 에너지 밀도는 기압에 비례하고, 온도에 반비례함. 이를 feature로 사용하기 위해 섭씨온도를 화씨온도로 변환 후 에너지 밀도 feature를 생성하여 사용. (에너지 밀도=기압/온도).

- feature engineering을 통해 다양한 feature를 추가하여 학습에 사용함.

* 태양광

- 인천 청라3동의 기상청 단기예보를 이용하여 14시 기준으로 train data를 생성.

- 데이터 분석을 통해 이상치 확인 후 처리

- 일출, 일몰 데이터를 활용하여 야간 시간(일출 전, 일몰 후)에 target값이 0보다 큰 값은 이상치로 간주하여 0으로 치환했을 때 성능이 향상됨을 확인함.

- 태양과 24절기의 상관관계에 따라, 양력 대신 음력 데이터로 feature를 사용하여 모델 성능을 개선함.

- feature engineering을 통해 다양한 feature를 추가하여 학습에 사용함.

< 모델 보완 및 추가 연구>

- target 데이터에서 연속된 값이 일정기간 지속될 경우 이상치로 간주하여 제거했으나, 성능이 좋아지지 않았음.

- brightics 제공 기본 모델인 Random Foreset, XGB, Linear Regression 모델로 학습해보았으나, 성능 차이가 심하여 앙상블하지 못함.

- keras, tensorflow 등의 라이브러리 설치가 불가능하여 lstm 같은 시계열 모델을 다양하게 시도하지 못한 점이 아쉬움.

- 태양광 데이터가 경향성을 띄기 때문에 향후 jupyter 환경에서 keras, pytorch 라이브러리를 활용하여 신경망 모델을 구축하여 앙상블 하여 진행 예정. 성능 향상이 기대됨.