RESEARCH ARTICLE

머신러닝 기반의 예측 시장 참여를 위한 태양광 발전량 예측 알고리즘 및 수익성에 관한 연구

김상진[1] · 유재혁[2†] · 장병훈[3] · 우성민[4]

[1]한국전력정보(주), 과장
[2]한국전력정보(주), 사원
[3]한국전력정보(주), 대표이사
[4]충북테크노파크, 책임연구원

A Study on Photovoltaic Prediction Algorithm and Profitability for Machine Learning based Prediction Market Participation

Kim Sang-jin[1] · Yu Jae-Hyeok[2†] · Jang Byung-Hoon[3] · Woo Sung-Min[4]

[1]Senior Researcher, Hankook Electric Power Information Co., Ltd.
[2]Researcher, Hankook Electric Power Information Co., Ltd.
[3]CEO, Hankook Electric Power Information Co., Ltd.
[4]Senior Researcher, Chungbuk Techno Park, Ltd.

[†]Corresponding author: applepy@hepi.co.kr

Abstract

With the introduction of the power generation prediction system, research is being conducted to predict hourly solar power generation through various algorithms and reduce prediction errors. However, increasing settlement revenue when participating in the prediction market is more important than improving prediction accuracy. In this study, we propose a method for predicting solar power generation using forecast and predicted weather data. In addition, the clustering algorithm was used based on solar radiation forecast data, and the causes of low prediction accuracy and profitability were analyzed for each cluster. Through this study, participation in the renewable energy generation prediction market is expected to be activated and opportunities for various business models will be provided.

Keywords: 태양광 발전량(Photovoltaic), 일기예보(Weather forecast), 기상예측(Weather prediction), 기계학습(Machine Learning), 군집화 알고리즘(Clustering algorithm)

Journal of the Korean Solar Energy Society
Vol.42, No.6, pp.173-183, December 2022
https://doi.org/10.7836/kses.2022.42.6.173

pISSN : 1598-6411

eISSN : 2508-3562

Received: 7 November 2022

Revised: 16 December 2022

Accepted: 21 December 2022

기호 및 약어 설명

NMAE : Normalized Mean Absolute Error의 약자로 전력거래소 예측 시장에서 제시한 예측 오차 기준식

PV : photovoltaics의 줄임말로 태양광 발전량으로 발전소의 설비용량으로 나눈 값, 단위 : 0 ~ 1

cGHI : 에너지기술연구원에서 제공받은 데이터의 변수로 수식으로 계산된 청명전 일사량, 단위 : Wh/m^2

Szen : 에너지기술연구원에서 제공받는 데이터의 변수로 수식으로 계산된 태양의 천정각, 단위 : Degree

Sazi : 에너지기술연구원에서 제공받은 데이터의 변수로 수식으로 계산된 태양의 방위각, 단위 : Degree

TMP : NWP 기상 예보 데이터의 변수로 지표기온, 단위 : K

RH : NWP 기상 예보 데이터의 변수로 1.5 m 고도 기온, 단위 : %

HPBL : NWP 기상 예보 데이터의 변수로 경계층고도, 단위 : m

CPRAT : NWP 기상 예보 데이터의 변수로 대류 1시간 누적강수량, 단위 : kg/m^2

DSWRF : NWP 기상 예보 데이터의 변수로 1시간 평균 전일사량, 단위 : W/m^2

DLWRF : NWP 기상 예보 데이터의 변수로 1시간 평균 장파복사량, 단위 : W/m^2

HCDC_rank : NWP 기상 예보 데이터의 상층운량(HCDC)를 범주형으로 표현, 단위 : 0,1

MCDC_rank : NWP 기상 예보 데이터의 중층운량(MCDC)를 범주형으로 표현, 단위 : 0,1

LCDC_rank : NWP 기상 예보 데이터의 하층운량(LCDC)를 범주형으로 표현, 단위 : 0,1

g_temp : 기상자료개방포털 사이트에서 제공하는 종관기상관측 자료의 지면온도, 단위 : $^\circ C$

sunshine : 기상자료개방포털 사이트에서 제공하는 종관기상관측 자료의 일조(hr), 단위 : 0 ~ 1

humidity : 기상자료개방포털 사이트에서 제공하는 종관기상관측 자료의 습도, 단위 : %

cloud : 기상자료개방포털 사이트에서 제공하는 종관기상관측 자료의 전운량(10분위), 단위 : 0 ~ 10

1. 서론

온실가스의 증가 등으로 인한 글로벌 기후위기로 탄소배출 및 에너지 문제의 중요성이 전세계적으로 대두되고 있으며 국내에서도 2050년까지 탄소중립을 달성하겠다는 목표를 발표하였다. 해당 정책에 따라 친환경에너지원인 신재생에너지의 비중은 늘어나는 추세이며 전력거래소의 공개된 정보에 따르면 전력시장에 등록된 발전기를 기준으로 2022년 10월 신재생에너지 설비용량 대비 태양광은 전체의 약 49.6%의 비중을 차지한다[1]. 신재생에너지의 보급이 증가함에 따라 운영하는 기술의 중요성은 더욱 강조되고 있으며 그 중의 핵심적인 기수 중 하나가 재생에너지 예보이다.

산업통상 자원부와 한국 전력 거래소는 2021년 상반기 재생에너지 확대에 따른 출력 변동성 대응을 위해 재생에너지 발전량 예측 제도를 도입하고 전력시장운영규칙 개정안을 확장하였다[2]. 태양광 발전량 예측은 사전에 전력의 가격을 결정하는데 도움이 될 수 있으며 정확한 장기간 태양광 발전량을 예측 하는 것은 태양광 발전 시스템을 운영하는 사업자의 수익을 최대화 하기 위해 매우 중요하다.

발전량 예측제도가 도입되면서, 대표적인 재생에너지 중 하나인 태양광 발전량 예보 기술력 향상을 위한 정확한 모델 구축이 필요해졌으며 다양한 알고리즘을 통하여 시간별 태양광 발전량을 예측하고, 예측 오차를 줄이는 연구가 수행이 되고 있다[3]. 하지만 발전량 예측의 정확도를 올리는 것도 중요하지만 예측 시장에 참여를 하였을 때 정산 수익을 높이는 것도 중요하다. 일사량을 기준으로 예보데이터의 군집화를 통하여 군집별로 예측 정확도와 수익성을 분석하였다.

본 연구에서는 과거의 데이터를 기반으로 기상 예보를 이용하여 실제로 관측된 기상 조건을 예측하여 예보된 기상데이터와 예측된 기상 데이터를 이용하여 발전량 예측 정확도를 향상시킬 수 있는 방법을 제안한다.

본 논문의 구성은 다음과 같다. 2장에서는 제안한 모델에 대한 설명과 발전량 예측에 사용한 데이터에 대해서 설명한다. 3장에서는 발전량 예측에 대한 성능 평가와 예측 시장 참여시 정산 수익금 현황을 분석 후, 마지막으로 4장에서 결론을 기술한다.

2. 발전량 예측 알고리즘과 예측을 위해 사용한 데이터

2.1 발전량 예측에 사용한 알고리즘의 정의

(1) 군집화 알고리즘

군집화(Clustering) 알고리즘은 비지도학습의 한 방법으로서, 데이터 분포들 중에서 유사하다고 판단되는 값들 끼리 묶어 몇 가지의 군집으로 나누는 작업을 의미한다. 가장 대표적인 K-Means 알고리즘은 군집의 개수를 K로 설정하고, 각 군집은 하나의 중심(centroid)을 갖게 되는데, 중심으로부터 데이터가 얼마나 떨어져 있는지를 비교하여 군집을 결정한다. K-Means는 사용 변수들의 거리 기반의 알고리즘으로 다차원의 값들을 비교 분석하기 쉽게 스케일링 과정을 거쳐서 오버플로우(overflow)나 언더플로우(underflow)를 방지를 하여야 한다. 본 연구에서는 중앙값(median)과 IQR (interquartile range)을 사용하여 아웃라이어의 영향을 최소화하는 RobustScalser 기법을 통하여 스케일링을 적용하였다.

K-Means 알고리즘의 군집 수 K는 사용자가 임의로 정하는 값으로 try and error를 통해서 정할 수 있지만 Inertia value를 활용하여 군집 응집도를 탐색하였다. Inertia value는 각 클러스터 중심에서 클러스터에 할당된 데이터 포인트간의 거리를 합산한 것을 의미하며, 군집이 얼마나 잘 응집되었는지 보여주는 지표로써 이 값이 작을수록 응집도가 높게 군집화가 잘 되었다고 평가할 수 있다.

전력수요 예측 연구에도 군집화 알고리즘을 사용하여 패턴을 분석하며 전체 사용량을 예측 하는 방식이 아닌, 군집 분석을 통한 군집별 예측량의 결합을 통하여 정확도를 더 상승시킨 사례가 있었으며[4,5] 본 연구에서도 군집화를 통하여 발전량 예측 정확도 향상과 수익성 향상 분석을 실시하였다.

(2) Lightgbm Regressor 알고리즘

GBM (Graideint Boosting Model)은 서로 다른 개별 모형들을 결합하여 모형의 성능을 높이는 앙상블 기법 중 하나인 부스팅(Boosting) 계열의 알고리즘이다. GBM이 높은 예측력으로 다양한 분석에서 사용되었지만 고차원 변수라 포함된 빅데이터에 적용 시 훈련 속도와 메모리 소비면에서 비효율적이라는 단점을 가지고 있다 . 이를 보완하기 위해 고차원 변수로 인해 데이터 크기가 클 때 학습의 효율성과 확장성을 개선하기 위해 기울기 기반 단측 표본추출법인 GOSS (Gradient-base One-Side Sampling)과 배타적 변수 묶음인 EFB (Exclusive Feature Bundling)을 적용한 새로운 형태의 GBDT인 LightGBM이 Ke et al. (2017)에 의해 제안되었다[6].

2.2 예측을 위해 사용한 데이터의 정의

(1) 기상예보데이터

데이터는 2022년 태양학회 추계학술대회의 태양광 예보 경진대회 참여 관련하여 한국에너지기술연구원(KIER)에서 제공받은 15개 장소의 태양광 발전량 데이터와 기상예보 데이터를 사용하였다. 제공받은 데이터는 2020년 과 2021년의 데이터로 구성되며 2020년 데이터를 통하여 예측 모델을 만들고 2021년에 모델을 적용하여 실제값과 예측값을 비교하였다. 태양광 데이터는 설비용량 대비 발전량으로 0과 1사이로 구성되었고 실제 발전량이 0.1이상 일 때 예측의 정확도를 높이는 것이 중요하다. 예보시간은 전날 21시 기준 으로 다음날의 기상요인을 예측한 데이터를 사용하였다. 상층운량(HCDC), 중층운량(MCDC), 하층운량(LCDC)의 경우에는 수치가 0일때는 값을 0, 그 외에는 1로 구성된 변수를 추가하였다. Table 1은 제공받은 예보데이터에서 예측에 사용한 변수와 발전량과의 상관도를 나타내었다.

Table 1 Configuration table of forecast data provided by KIER

Name	Value	Correlation with PV
cGHI	solar radiation quantity (calculated by formula)	0.77
Szen	zenith angle of the sun (calculated by formula)	-0.7
Sazi	azimuth of the sun (calculated by formula)	0.18
TMP	surface temperature (forecast)	0.28
RH	1.5 m altitude relative humidity (forecast)	-0.45
HPBL	boundary layer elevation (forecast)	0.41
CPRAT	Convection 1 hour cumulative precipitation	0.5
DSWRF	1 hour average total solar radiation	0.84
DLWRF	1 hour average long wave radiation	-0.52
HCDC_rank	Consists of 0 and 1 in HCDC	-0.17
MCDC_rank	Consists of 0 and 1 in MCDC	-0.26
LCDC_rank	Consists of 0 and 1 in LCDC	-0.21

Fig. 1은 훈련집합으로 사용할 2020년 데이터의 발전량과 표1에서 명시한 기상 예보 데이터들의 상관도를 표현하기 위해서 히트맵으로 시각화를 하였다. 발전량(PV)과 상관도가 가장 높은 변수는 예보 일사량(DSWRF), 측정된 일사량(cGHI), 태양의 천정각(Szen)이며 cGHI와 Szen의 경우에는 음의 상관관계가 −1에 근접한 것을 확인할 수 있다.

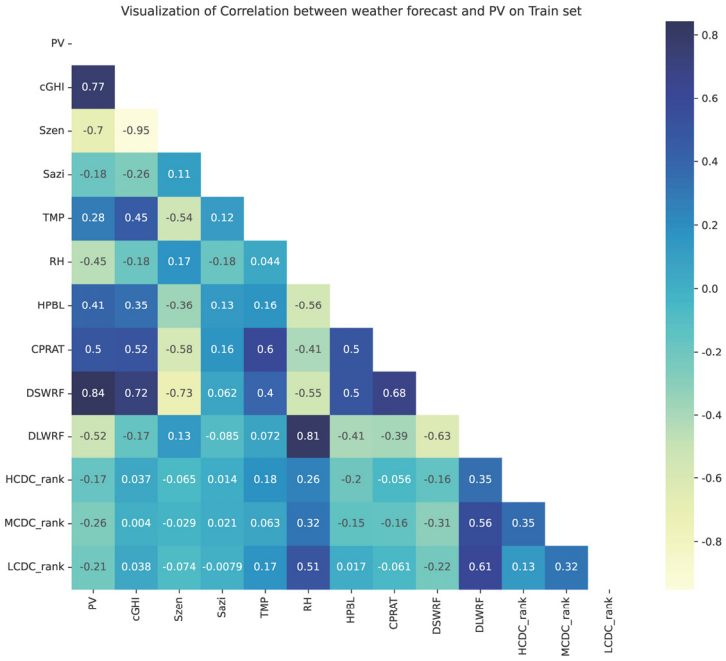

Fig. 1 Visualization of Correlation between weather forecast and PV on Train set

(2) 기상관측데이터

발전량 예측 모델을 만들기 위해서는 예보 데이터를 기준으로 하였을 때 실제 기상요인을 파악하는 것이 중요하다. 훈련집합으로 사용되는 2020년의 발전소 위치에 근거하여 예측 모델 생성을 할 때 기상청의 실제 기상 데이터를 사용하였다.

공공기상청에서 제공하는 종관기상관측자료의 기상요인 중에서 발전량과 상관이 높은 지면온도(g_temp), 일조(sunshine), 습도(humidity), 전운량(cloud)을 예측에 사용하였고 발전량의 상관도를 Fig. 2 (Left)에서 시각화를 하였다. 2020년의 예보 데이터와 실제 기상데이터를 변수로 하여 발전량을 예측하는 모델을 생성하고 2021년의 발전량을 예측하여야 하지만 2021년의 실제 기상데이터는 모른다는 가정하에 발전량을 예측하여야 한다. 따라서 Table 1에서 명시한 예보 데이터 변수를 사용하여 기상요인들을 LGBM Regressor 알고리즘을 사용하여 예측을 수행하였다. Fig. 2 (Right)는 2021년의 예측 된 기상요인과 발전량의 상관도를 표시한 것이다. 2020년 데이터의 발전량과 기상요인의 상관관계 수치와 2021년의 발전량과 예측된 기상요인의 상관관계 수치를 비교해보았을 때 큰 차이가 없는 것을 확인 할 수 있다.

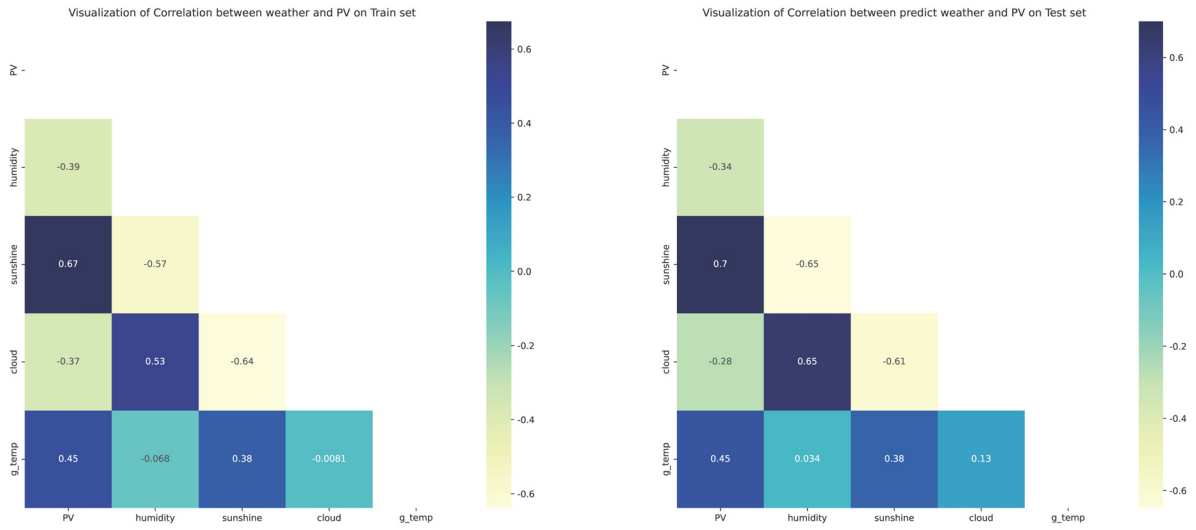

Fig. 2 Visualization of Correlation between predict weather and PV

3. 결과 및 토의

3.1 발전량 예측 알고리즘 적용에 따른 정확도 분석

태양광 발전량 예측은 발전소 별로 2020년에 발전량 데이터를 기준으로 2021년의 발전량을 예측하는 방향으로 진행하였다. 태양광 발전량은 일사량과의 상관도가 크기 때문에 일사량 예보 수치가 높을 때 발전량의 예측 정확도를 높이는 것이 중요하다. 일사량의 유사 패턴을 잡아내기 위해 일사량 예보데이터(DSWRF)와 일사량 측정데이터(cGHI), 태양의 방위각(Sazi)의 변수를 기준으로 군집화 알고리즘을 수행하였고 Fig. 3는 군집화 개수를 설정하기 위해 군집의 수에 따른 응집도를 시각화로 표현하였다.

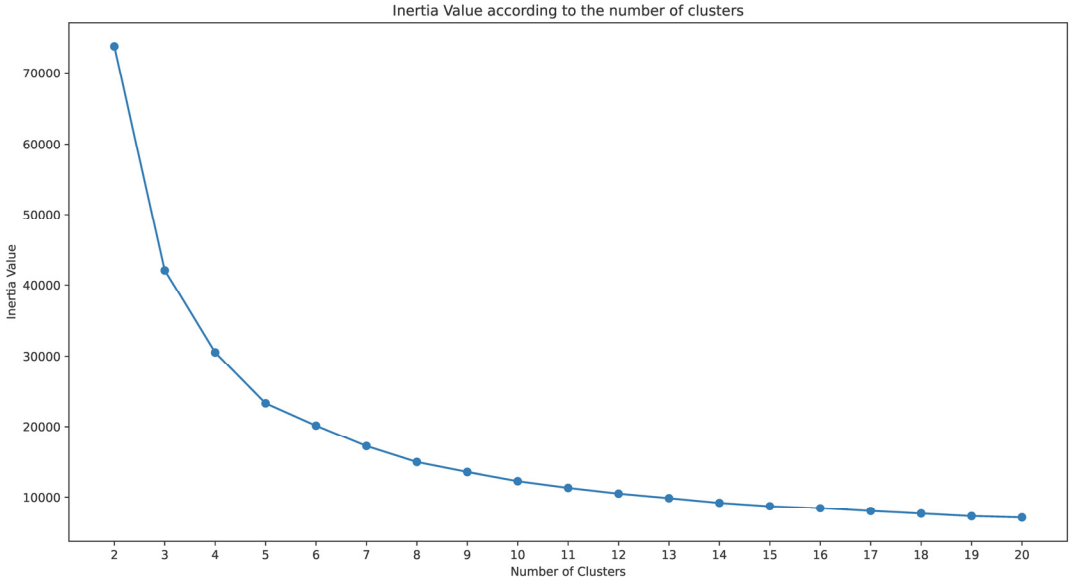

Fig. 3 Inertia value according to the number of clusters

Fig. 3에서 y축에서 명시한 Inertia value의 수치가 작을수록 군집이 잘 되었다고 볼 수 있으며 6개를 넘어가면서 수치가 점차 작아지는 추세를 나타낸다. Table 2는 군집의 개수를 6으로 설정해서 군집별 예보일사량(DSWRF)과 측정 일사량(cGHI)의 평균의 수치를 기준으로 내림차순으로 하여 각 기상예보 데이터의 평균값과 전체 데이터에서 군집이 차지하는 비중(Density)를 나타내었다.

Table 2 Summary table of characteristics of weather forecast data by cluster

Cluster_No	DSWRF	cGHI	CPRAT	RH	HPBL	DLWRF	Density
1	738.4	784.7	0.31	55.2	829.2	-109.7	12.5
2	478.4	703.9	0.19	62.5	627.3	-88.5	12.6
3	413.4	405.6	0.21	59.4	750.7	-85.2	14.4
4	164.7	601.3	0.09	79.4	388.4	-40.9	16.1
5	68.6	54.7	0.08	72.8	329.9	-57.9	25.6
6	25.9	140.1	0.03	81.2	208.9	-51.4	18.7

예측 모델을 만들 때 사용한 기상 변수는 2020년은 예보데이터와 실제 공공기상데이터를 사용하였고 2021년은 예보데이터와 예측된 공공기상데이터를 사용하였다. Fig. 4는 발전량 예측 과정으로 예보 데이터를 군집화하여 패턴을 분석한 다음에 2020년 예보데이터와 실제 발생한 기상데이터를 가지고 2021년의 발생 되어야 될 기상데이터를 각각 예측한 다음에 예측된 기상데이터를 같이 이용하여 2021년의 발전량을 예측하는 방법으로 진행하였다.

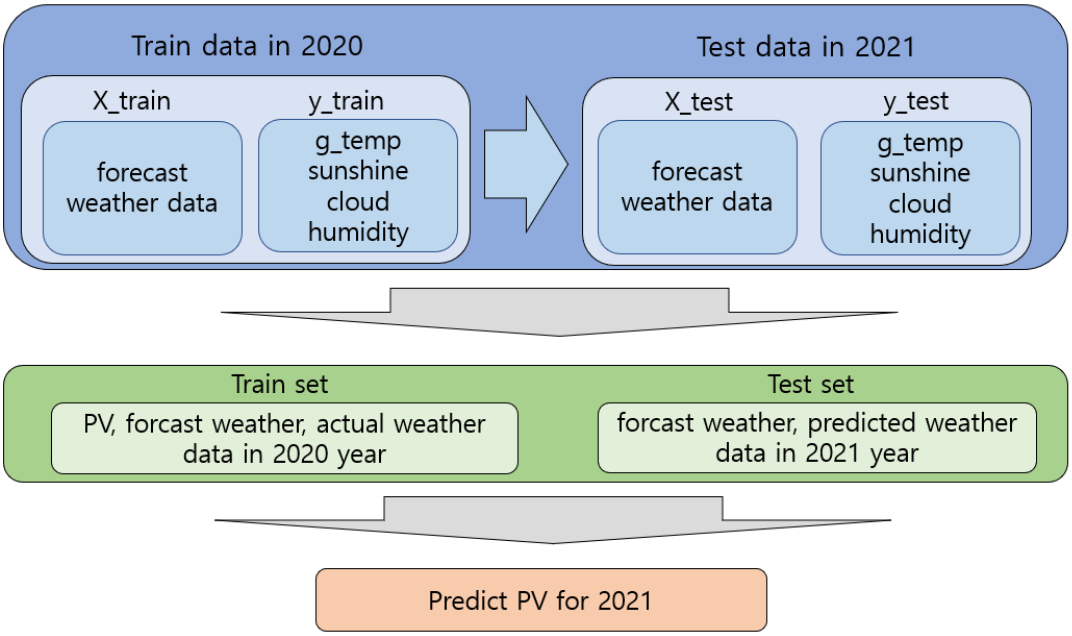

Fig. 4 PV predicting method Process

발전량 예측에 사용된 예측 오차 기준 지표는 전력거래소의 재생에너지 발전예측제도에서 제시한 NMAE (Normalized Mean Absolute Error)이며, 계산하는 방식은 다음과 같다.

$$NMAE(\%) = \frac{100}{n}\sum_{t=1}^{n}\left|\frac{A_t - F_t}{S}\right|, A_t \geq 0.1 \times S \quad (1)$$

여기서 A_t : t 시간대의 실제 발전량, F_t : t 시간대의 예측 발전량, S : 발전량 설비용량을 의미한다. 예측에 사용한 발전량은 실제 발전량에 설비용량을 나눠서 발전량의 수치의 범위는 0과 1사이를 나타내며 Fig. 5는 예보데이터만 변수로 사용한 경우(Case1)와 예보데이터와 기상예측 데이터를 같이 사용했을 경우(Case2)에 대해서 발전소별 평균 정확도를 나타내었다.

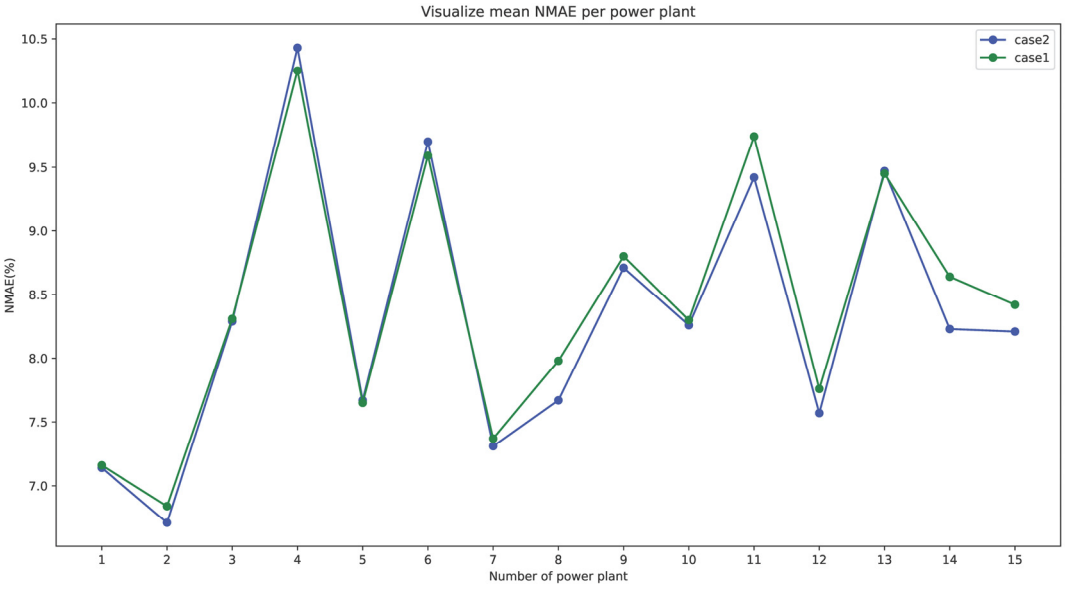

Fig. 5 Visualize mean accuracy per power plant

발전량 예측에 예측된 기상 데이터를 예보데이터와 같이 썼을 때 평균 NMAE는 약 8.31%로 예보데이터만 사용했을 때보다 약 0.1% 향상되었으며 14번 발전소의 경우에는 평균 정확도가 약 0.4% 가까이 향상된 것을 확인하였다.

3.2 발전량 예측 알고리즘 정확도 따른 정산 수익금 분석

전력거래시장에서 태양광 발전량 예측을 통하여 정산을 받기 위해서는 시간대별 실제 발전량이 설비용량의 10%에 해당하는 부분에 대해서 오차율이 8% 이내에 들어가는 것이 중요하다. 오차율이 6% 이하이면 1 kWh당 4원이며 6% 이상 8% 미만이면 1 kWh당 3원의 정산금을 받을 수 있다. 평균 예측 정확도가 높은 것도 중요하지만 발전량이 많은 시간대에 예측 정확도를 높여 예측시장 참여 시 정산수익금을 높이는 것 또한 중요하다.

Fig. 6은 15개의 발전량 설비들의 설비용량을 동일하게 1 MW라고 가정하고 예측시장의 참여 가정시의 수익금(Profits)을 천원 단위(1,000 KRW)로 나타내었다. 또한 설비용량을 동일하게 1 MW라 가정하였을 때 설비용량의 10% 이상을 기준으로 발전소별 연간 누적 발전량을 나타내어 수익금이 떨어지는 원인을 시각화하였다.

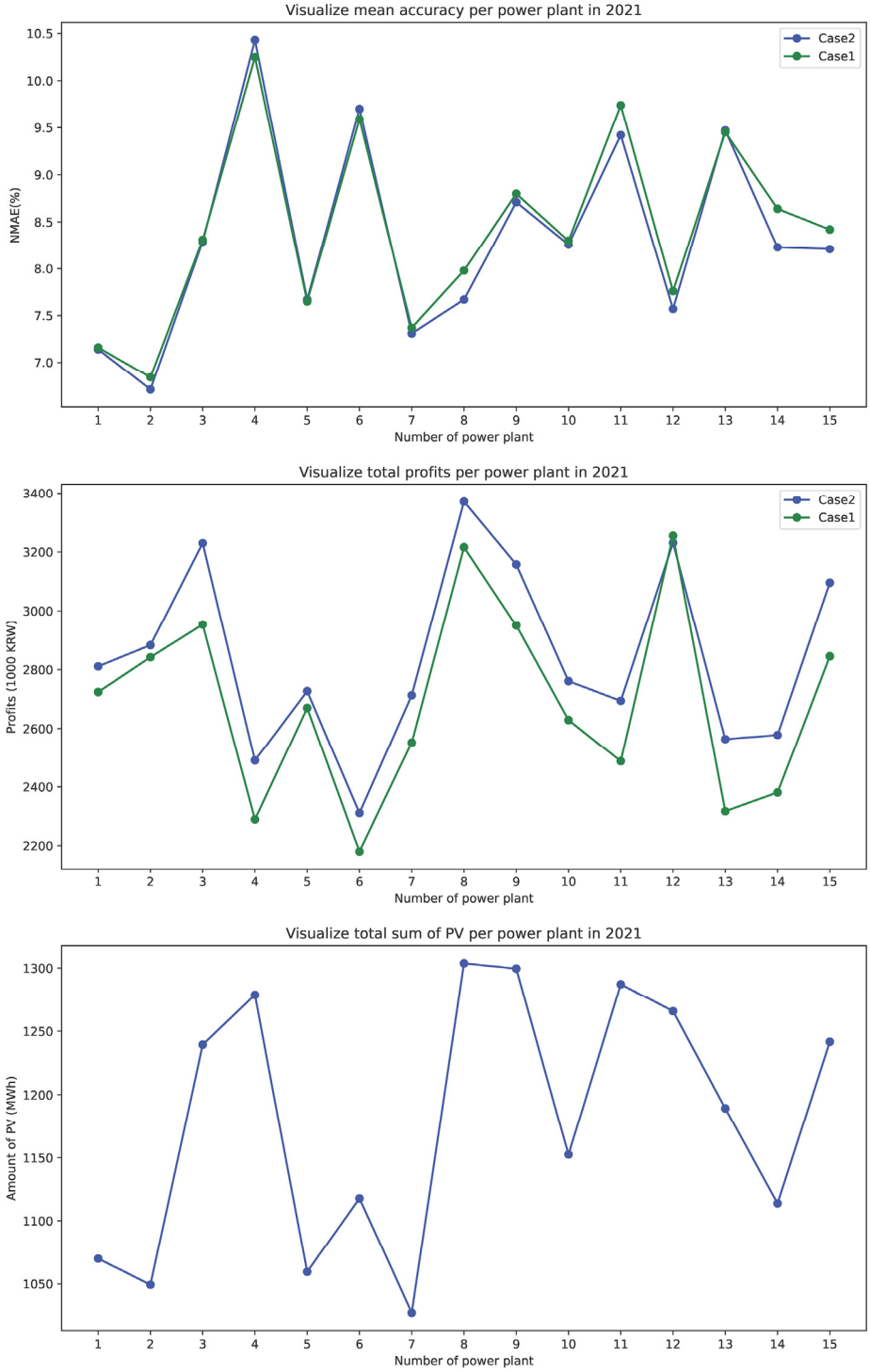

Fig. 6 Visualization of Accuracy and Profits by Power Plant in 2021

또한 3.1절의 발전소별 NMAE 오차율에 근거하여 발전소별 수익금을 나타내어 예측된 기상정보를 같이 사용하였을 때가 더 수익성이 좋은지를 나타내었다. Fig. 6에서 평균 NMAE 에러율이 7.5% 이하인 발전소는 1번, 2번 ,7번의 발전소이며 정산수익금이 320만원 이상인 발전소는 3번, 8번, 12번 발전소로 나타났다.

1번, 2번, 7번의 경우에는 정확도가 높지만 수익성이 상대적으로 떨어지는 원인은 발전량이 상대적으로 적은 것으로 파악하였다. 하지만 8번과 9번을 비교했을 때에는 연간 누적 발전량은 비슷하지만 9번이 8번보다 정확도가 떨어지면서 수익성이 떨어지는 것을 알 수 있다. 9번 발전소는 8번 발전소보다 NMAE 에러가 약 1.1% 높으며 9번 발전소의 정산수익금은 8번 발전소의 정산 수익금 대비 93%로 수익성이 떨어지는 것을 확인하였으며 이는 정확도의 개선이 필요한 것으로 나타났다.

Fig. 7은 연간 누적 발전량이 1150 MWh 이상인 3, 4, 8, 9, 11, 12, 15번 발전소의 군집별로 정산수익금의 합계를 나타내었다. 1번 군집 번호는 상대적으로 측정 일사량(cGHI)와 예보 일사량(DSWRF)가 높은 수치에 해당하는 데이터들의 구성으로 군집의 개수를 늘릴수록 해당 발전량 데이터들의 평균 수치가 높다. 정산수익금은 일반적으로 일사량 예보수치가 상대적으로 높은 1번 군집에서 수익금이 높게 나오는 것을 확인할 수 있다.

하지만 9번 발전소의 경우에는 1번 군집보다 2번 군집에서 정산 수익금이 더 높은 것으로 나타났으며 이는 실제 발전량이 높을 때의 예측 정확도가 상대적으로 떨어지는 것을 알 수 있다. 그리고 NMAE 오차가 가장 높은 4번 발전소는 다른 발전소들과 비교 시 군집 1, 2, 3번에서 정산 수익금의 차이가 많이 벌어진 것을 확인할 수 있다.

따라서 정산수익금을 높이기 위해서는 발전소별 및 지역의 기후특성을 고려하여 상단의 번호의 해당하는 군집의 예측 정확도를 상승시킬 수 있는 모델 개발이 필요하다.

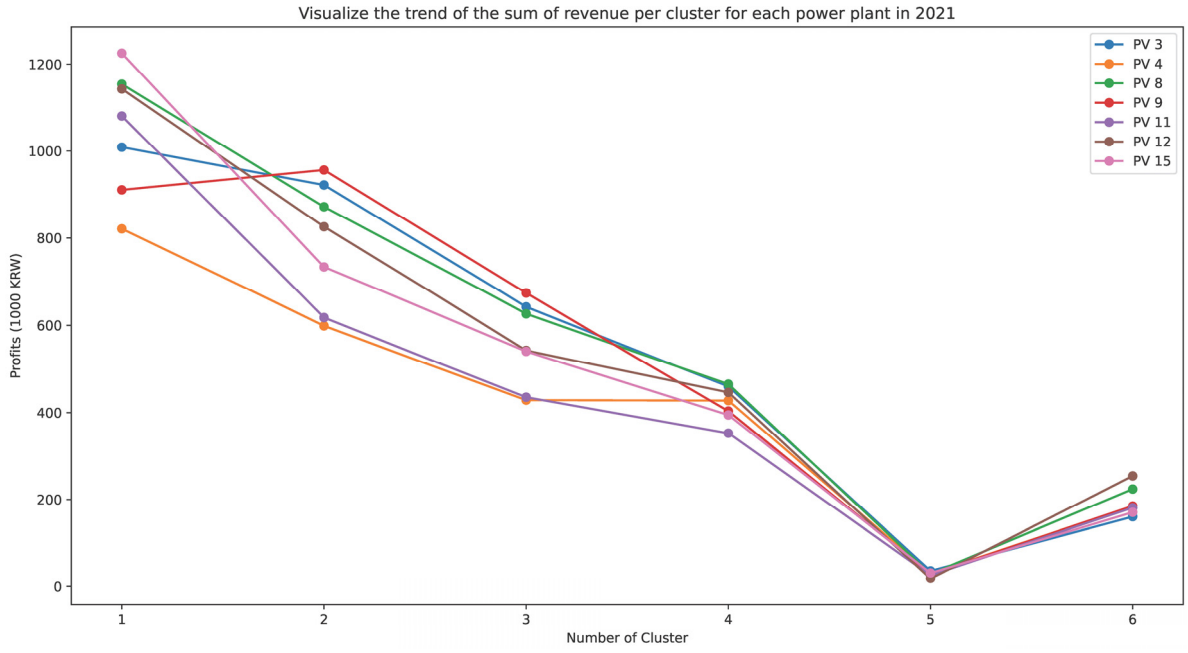

Fig. 7 Visualize the trend of the sum of revenue per cluster for each power plant

4. 결 론

재생에너지의 발전량이 일정하지 못한 특성은 기존 에너지 공급 및 거래시스템에 위해가 되는 문제이며, 이에 따른 제반 문제를 해결하기 위해 다양한 연구들이 진행되어 왔다. 본 연구에서는 에너지기술연구원에서 제공받은 기상 예보데이터와 공공기상데이터를 이용하여 태양광 발전량 예측 분석 기법을 제안하였고 예측 결과에 따른 수익성을 분석하였다.

정확도 측면에서 과거의 공공기상청의 실제 관측된 기상요인을 대상으로 예보 데이터를 통한 예측 시점의 기상요인 예측값을 발전량 예측 변수에 추가하여 발전량 예측시 정확도가 더 향상된 것을 확인하였다. 하지만 수익성 측면에서의 분석 시에는 평균 정확도가 높은 것도 중요하지만 발전량이 높을 때의 예측 정확도를 향상시켜 수익금을 최대화 하는 예측 모델의 개발이 필요하다.

향후 연구에서는 태양광 발전량 예측의 성능을 향상시키기 위해 시간의 흐름에 따라 변화하는 데이터를 학습하면서 데이터의 특징을 잘 포착할 수 있는 태양광 발전량 예측 분석 연구를 진행할 예정이다.

후 기

본 연구는 중소벤처기업부(S3153016) 및 산업통상자원부와 한국산업기술진흥원의 “지역혁신클러스터육성(R&D, P0016222)”의 지원을 받아 수행된 연구결과로써 한국에너지기술연구원-한국태양에너지학회가 공동주최한 경진대회 입상작입니다.

REFERENCES

1. KPX, 20222. https://new.kpx.or.kr/board.es?mid=a10109010700&bid=0082&act=view& list_no=68145. last accessed on the 14th December 2022.
2. Ministry of Commerce, Industry and Energy, Introduction of Renewable Energy Generation Prediction System, 2020.
3. Lee, J., Park, W., Lee, I., and Kim, S., Comparison of Solar Power Prediction Model Based on Statistical and Artificial Intelligence Model and Analysis of Revenue for Forecasting Policy, Journal of IKEEE, Vol. 26, No. 3, pp. 355-363, 2022.
4. Sohn, H., Jung, S., and Kim, S., The Prediction and Valuation of Gas Consumption in Building using Artificial Neural Networks Based on Clustering Method, The Korean Journal of Applied Statistics, Vol. 29, No. 1, pp. 193-203, 2016.
5. Choi, D., Lee, Y., and Ko, M., A Study on Electricity Demand Forecasting Based on Time Series Clustering in Smart Grid, Korea Institute of Ecological Architecture and Environment, Vol. 18, No. 5, pp. 69-74, 2018.
6. Ke, G, Meng, Q., Finely, T., Wang, T., Chen, W., and Ma, W., Lightgbm: A Highly Efficient Gradient Boosting Decision Tree, Advances in Neural Information Processing System, Vol. 30, pp. 3146-3154, 2017.