MAKINA PEBBLES

MACHINE INTELLIGENCE FOR MANUFACTURING

Creatively Designed and Beautifully Engineered

2019. 03

기상 예보를 이용한 태양광 발전량 예측 시스템

알고리즘

기상청 예보 과거 데이터 (Raw Data) → 데이터 전처리 → 예측 모델 → 저차원의 예측값 → 데이터 후처리 (복원) → 태양광 발전량 예측값

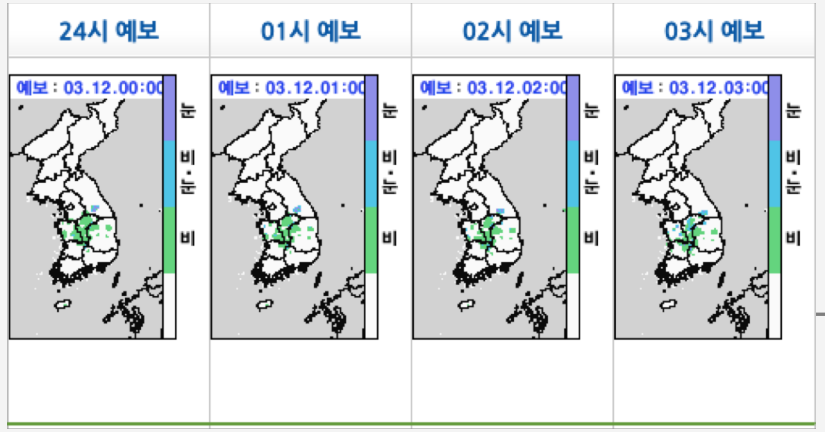
핵심 아이디어

선행연구(도메인 지식)을 반영한 모델 — Singular Value Decomposition (SVD) 통한 차원 축소/복원 — Bayesian Optimization을 통한 최적화

1. 사용 데이터 (X Data)

X Data : 예측을 위해 사용하는 데이터와 트레이닝을 위해 사용하는 데이터는 논리적으로 동일한 유형이어야 함

Training



Training Data
기상청 예보 과거 데이터 (2016.11 ~ 2018.12)

Preprocessing

Training

Inference

Input Data (당일 기상예보)

Data Preprocessing

Model

Data Postprocessing

태양광 발전량 예측(당일, +1일, +2일)

"태양광 발전량 예측을 위해 기상예보를 사용할 것이라면, 모델을 트레이닝 하는 데에도 기상예보 과거 데이터를 사용해야 함"

1. 사용 데이터 (X Data)

사용 변수 : 선행연구(논문) 참고하여 선택

선행연구1

'기상환경 모니터링 데이터를 이용한 태양광발전시스템 발전량 성능 분석'(2018), 권오현 외 1명
- 사용변수 : **일사량, 기온, 습도, 대기 먼지 농도, 적설량**
- 사용모델 : PVsyst (상용 소프트웨어)

선행연구2

'태양광발전설비 원격 관제를 위한 빅데이터 분석 및 처리'(2018), 권준아 외 3명
- 사용변수 : **기온, 습도, 운량, 풍향, 일사량, 모듈온도**
- 사용모델 : ANN(Artificial Neural Network), SVM(Support Vector Machine)

선행연구3

'SolarisNet : A Deep Regression Network for Solar Radiation Problem'(2018), Subhadip Dey et al.
- 사용변수 : **최고기온, 최저기온, 일조 시간**
- 사용모델 : DNN(Dense Neural Network)

사용 변수

- 수치형 변수 : **기온, 습도, 풍속, 풍향**
- 범주형 변수 : **하늘상태, 강수형태**
- 기상현상 관련 변수 : **강수량, 강수확률, 적설량**

"선행연구 참고하여 기상청 예보에서 제공되는 변수 사용"

1. 사용 데이터 (X Data)

기상청 예보(3시간 단위)	태양광 발전량 예측(15분 단위)	Issue
	2019.03.14 12:00	하나의 값으로부터 12개의 값을 뽑아내야 함
2019.03.14 02:00 발표된 12:00 예보	2019.03.14 12:15	15분 단위의 변화에 대한 정보 없음
	⋮	
	2019.03.14 14:45	
	2019.03.14 15:00	Interpolation 하는 과정에서 이를 추측으로 채우게 됨

"하나의 샘플(3시간)로 12개의 결과값을 출력"

1. 사용 데이터 (X Data)

기상청 예보(3시간 단위)	태양광 발전량 예측(15분 단위)	Issue
2019.03.12 23:00 발표된 12:00 예보	2019.03.14 12:00	해당 시간을 예측하는 다수의 예보 존재
⋮	2019.03.14 12:15	몇 시에 발표된 예보를 사용해야 하는가?
2019.03.14 02:00 발표된 12:00 예보	⋮	
⋮	2019.03.14 14:45	모델의 성능에 영향을 미침
2019.03.14 08:00 발표된 12:00 예보	2019.03.14 15:00	

"몇 시 예보를 사용할 것인지 선택 필요"

1. 사용 데이터 (X Data)

기상청 예보 과거데이터 (Raw Data)
(2016.11 ~ 2018.12)

일시	발표시간	예보시간	기온	습도	...	풍향
2016.11.01	2시	+4	5.9	43	...	343
		+7	4.4	50	...	347
		...	...	...	...	...
		+61	7.9	84	...	224
	5시	+4	4.4	50	...	347
		...	...	...	...	...
		+58	7.9	84	...	224
	...	...	...	...	...	...
	23시	+4	13.1	44	...	106
		...	...	...	...	...
		+64	10	73	...	213
...	...	...	...	...	...	...
2018.12.31	2시	+4	4	0		336
	...	...	...	...	...	...
	23시	+64	-4	0		286

X Data (당일 예측)
(02~08시 발표된 당일에 대한 예보)

일시	예보 발표시간	예측하는 시간	예측값		
			기온	...	풍향
2016.11.01	02시	당일 06시	5.9	...	343
		당일 09시	4.4	...	347
		...	...	...	...
		당일 21시	3.0	...	1
	05시	당일 09시	4.4	...	347
		...	...	...	...
		당일 21시	3.0	...	1
	08시	당일 12시	2.9	...	357
		...	...	...	...
		당일 21시	2.4	...	5

X Data (+1일, +2일 예측)
(05~14시 발표된 +1, +2일에 대한 예보)

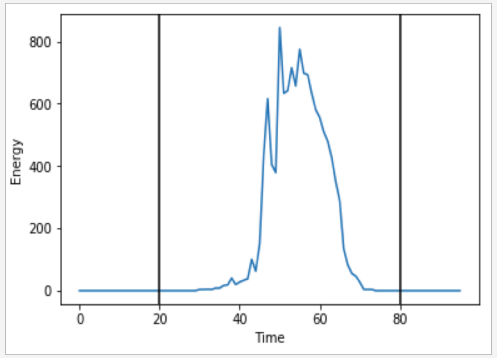
일시	예보 발표시간	예측하는 시간	예측값		
			기온	...	풍향
2016.11.01	05시	+1일 06시	7	...	35
		...	...	...	...
		+1일 21시	9.3	...	310
		+2일 06시	13	...	288
		...	...	...	...
		+2일 15시	7.9	...	224
	08시	+1일 06시	8.1	...	147
		...	...	...	...
		+2일 15시	8.7	...	212
	...	...	...	...	...
	14시	+1일 06시	8.5	...	151
		...	...	...	...
		+2일 15시	8.5	...	230

"예측 목적에 맞게 데이터를 가공"

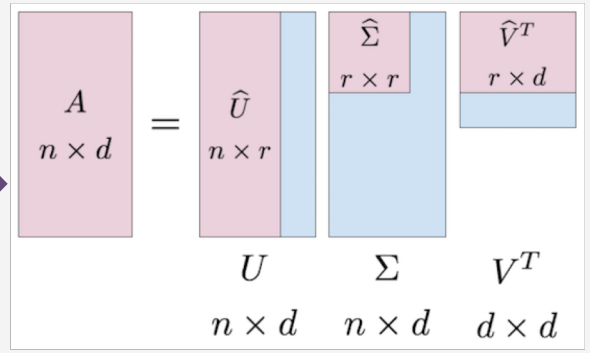
1. 사용 데이터 (Y Data)

Y Data : 15분 단위의 태양광 발전량 출력

일반적으로 모델의 아웃풋 차원(Output Dimension)이 클수록 예측 성능이 떨어지기에, 데이터의 차원을 축소



Y_2 : 24시간동안의 태양광 발전량

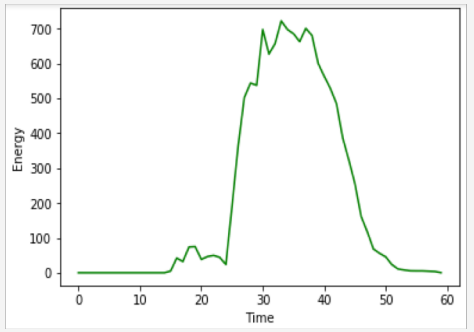


Singular Value Decompoisition(SVD)

Basis($\sum V^t$) 로 Reconstruction

-0.0704 0.0053 -0.0621 -0.0770 -0.0441

Y_1 : 차원축소된 발전량
($U\ Matrix$의 해당 데이터에 해당하는 row)



$\hat{Y_2}$: 복원된 발전량

- 일출/일몰 관측/예보 데이터 (한국천문우주연구원) 사용하여 절삭
- Singular Value Decomposition(SVD) 활용하여 축소/복원

"모델이 저차원인 Y_1을 예측"

1. 사용 데이터 (Y Data)

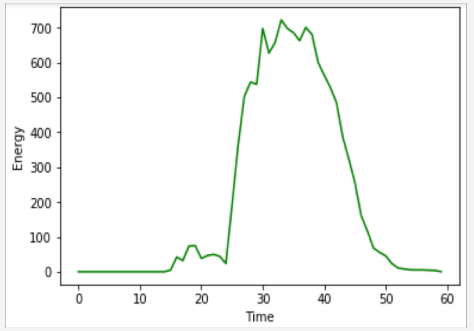
Y Data : 15분 단위의 태양광 발전량 출력

일반적으로 모델의 아웃풋 차원(Output Dimension)이 클수록 예측 성능이 떨어지기에, 데이터의 차원을 축소

기상 예보 → 예측 모델 → -0.0704 0.0053 -0.0621 -0.0770 -0.0441

Y_1 : 차원축소된 발전량
($U\ Matrix$의 해당 데이터에 해당하는 row)
(5, 1)

Basis($\sum V^t$) 로 Reconstruction



$\hat{Y_2}$: 복원된 발전량

- 일출/일몰 관측/예보 데이터 (한국천문우주연구원) 사용하여 절삭
- Singular Value Decomposition(SVD) 활용하여 축소/복원

"모델이 저차원인 Y_1을 예측"

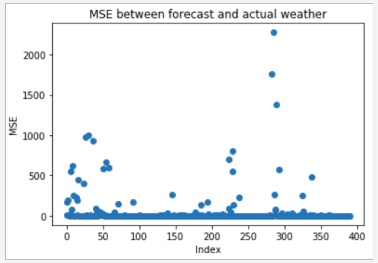
2. 데이터 전처리

1 데이터 클리닝

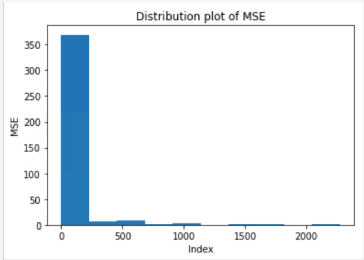
- 관측 오류 값 모두 제거
- 누적합 잘못된 행들 모두 제거
- 누적합 -> 변화량으로 분해
- Site B에 가장 데이터 오류가 많음

2 기상예보 검증

- 틀린 예보를 활용하면 모델은 잘못된 X->Y 관계를 학습함
- 예보와 관측치의 차이를 MSE(Mean Square Error)로 검증
- Bayesian Optimization의 Hyperparameter로 제공하여 모델 성능 최적화



MSE Plot

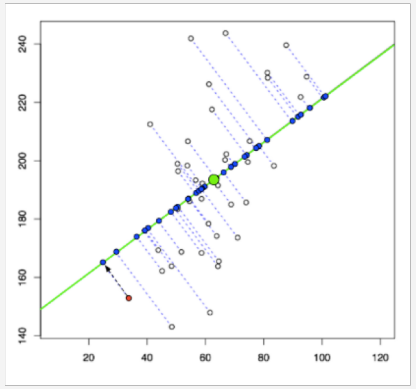


MSE Distribution Plot

3 입력 데이터 차원 축소

PCA로 입력 데이터 차원 축소

- Explained Variance = 0.99
- Bayesian Optimization의 Hyperparameter로 제공하여 모델 성능 최적화



4 피쳐 엔지니어링

Polynomial Features

- 샘플데이터를 분석했을때에 Polynomial Features 활용이 좋은 성능을 보임

[기온, 강수량]

$[1, 기온, 강수량, 기온^2, 기온 * 강수량, 강수량^2]$

모델의 $Input$이 되는 특성(적용 후)
poly_dim=2 일때 Feature Engineering 예시

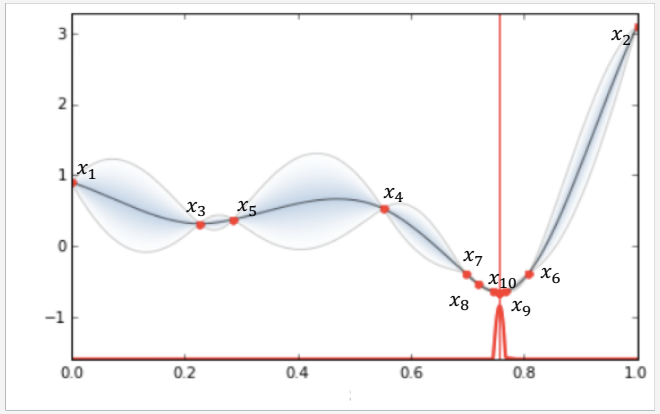
3. 모델 선택 / 트레이닝

XGBoost → Bayesian Optimization → 최적의 모델

Feature수에 비해
부족한 샘플 수
(Site별 300~500개)

Decision Tree
+
Boosting

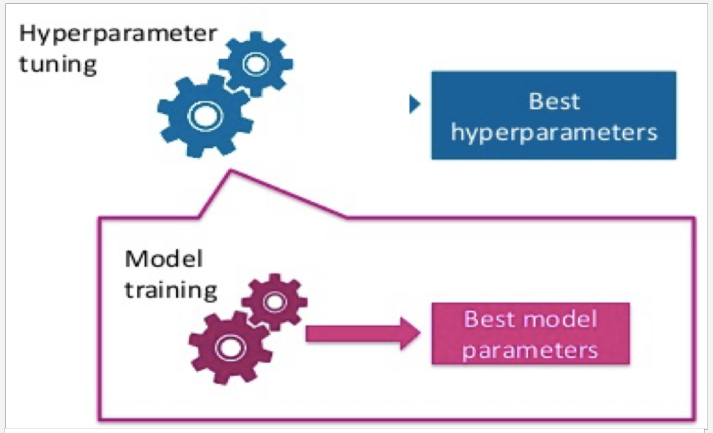
XGBoost Tuning으로
성능 최적화



Hyperparameters

- 전처리 방법에 따른 성능 차이
- Model Tuning을 위한 Hyperparameter

Bayesian Optimization을 통해
전처리 방법, 모델 구조 최적화



"5-Fold Cross Validation을 통해 가장 R^2가 높은
[전처리 방법, 모델]을 선택"

3. 모델 선택 / 트레이닝

Evaluation Metric : 목적에 강건(Robust)한 지표를 사용하여 Training

MAPE	Y의 스케일이 클 때만 |실제값 – 예측값|에 대해 강건함
R^2	Y의 스케일이 작을 때에도 |실제값 – 예측값|에 대해 강건함
Minimize $\|1 - R^2\|$	Y의 스케일이 작으므로 $\|1 - R^2\|$ 을 최소화하도록 최적화

$$MAPE = \frac{100\%}{n}\sum_{t=1}^{n}\left|\frac{\text{실제값} - \text{예측값}}{\text{예측값}}\right|$$

$$\left|\frac{0.002 - 0}{0.002}\right| = 1$$

0을 내놓았을때 잃는 점수

$$\left|\frac{0.002 - 0.8}{0.002}\right| = 390$$

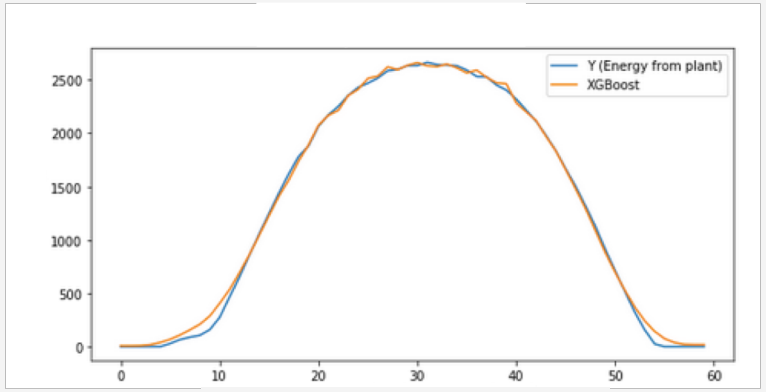
0~1 사이의 예측값을 내놓았을때 잃는 점수

100-MAPE를 최소화하는 방향으로 최적화시키면
모델은 점수를 크게 잃을 위험이 있는 구간에서
0을 제출하여 문제를 풀지 않는 방향으로 최적화됨

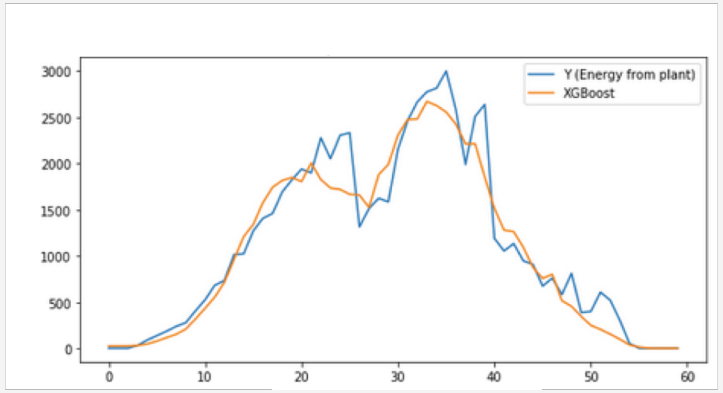
"$|1 - R^2|$을 최소화시킴으로써 R^2가 1에 가까워지도록 최적화"

4. 결과

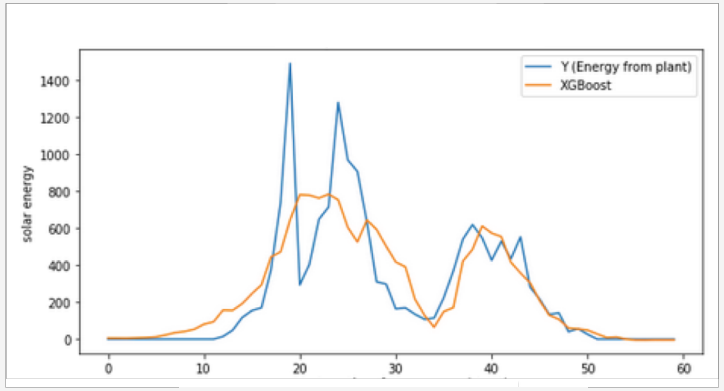
디테일한 변동에 민감하게 반응하지 않으며 발전량 변화의 경향을 성공적으로 예측하는 모델을 얻음



맑은 날



하루 중의 기상 변화



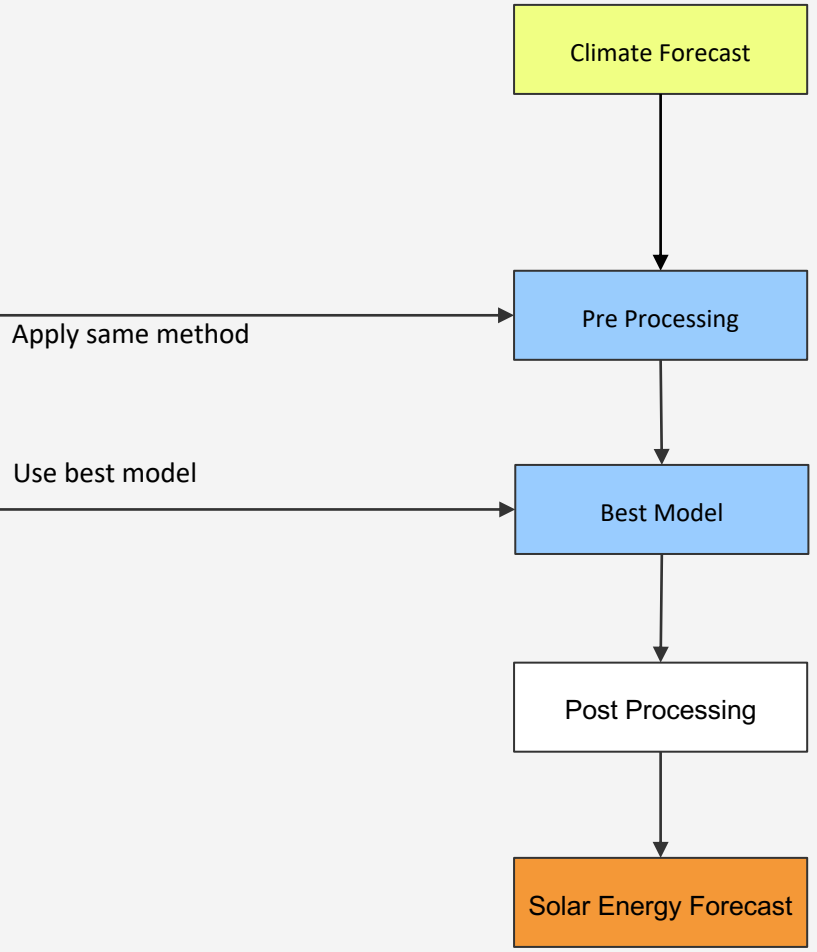
흐린 날

"발전량 변화의 경향을 성공적으로 예측함으로써 성공적인 일일 총 발전량 예측 가능"

5. 알고리즘

1. Training

2. Inference



6. 발전 가능성 / 개선방안

1 추가로 적용할 수 있는 아이디어

- Data Augmentation : 실제 관측값을 트레이닝 데이터에 추가
- 예보 검증 : 강수량 외에 다른 요인도 검증
- 모델 분리 : 현행 2개 모델에서 3개 모델로 분리
- 예보 시간 분리 : 더 적은 수의 예보를 넣었을때 성능 비교
- 미세먼지 데이터 사용 : 미세먼지 예보 데이터 사용하여 성능 향상
- 딥러닝 모델로의 전환 : 현재는 XGBoost를 활용하지만, 더 많은 데이터를 모음에 따라 딥러닝 모델로 전환하여 성능 향상

2 개선 방안

- 기상청과의 협업 : 기상 예보를 위해 사용하는 Raw Data 활용으로 성능 향상
- 출력 차원과 성능 : 일 단위 발전량만을 예측하게 함으로써 모델의 부담 최소화

3 범용성

- 다양한 위치의 발전소에 적용 가능 : 기상예보, 해당 발전소의 발전량 두 가지만 사용하므로, 다양한 지역의 발전소에 적용 가능
- 학습이 완료되고 난 이후에는 예측을 위해 기상예보 외의 다른 데이터를 필요로 하지 않음

This document is strictly confidential