시계열 특성을 고려한 강수량 예측 모델링

벤치클라이머

정정룡 정성훈 황영우

Contents 목차

공모 배경 및 목표

2023.07.20 BBC 코리아
매년 반복되는 침수 사고, 오송 지하차도 참사 막을 수 없었나?

2024.07.23 연합뉴스
경기 개시 직전...수원 SSG-kt 경기, 우천 취소

2024.07.20 조선일보
싸이 '흠뻑쇼' 과천 공연, 1시간만에 취소... 안전상 이유

강수 예측 → 인명/재산 피해 감소

목표

3년간 AWS 지점별 강수량 데이터를 바탕으로 구간별 강수량 예측 모델 구축

- 기간 : 5월 ~ 9월
- 데이터 : 지점별 1시간 강수자료 및 앙상블 모델의 구간별 강수량 누적확률값

Flow Chart

STEP.1

데이터 수집

- 기상청 앙상블 데이터

STEP.2

EDA

- EDA
- 변수 시각화

STEP.3

데이터 전처리

- 변수 변환 및 정렬
- 결측치 제거
- 파생변수 생성
- 학습데이터 분할

STEP.4

모델링

- Custom Loss function
- Layer 설계 및 학습

STEP.5

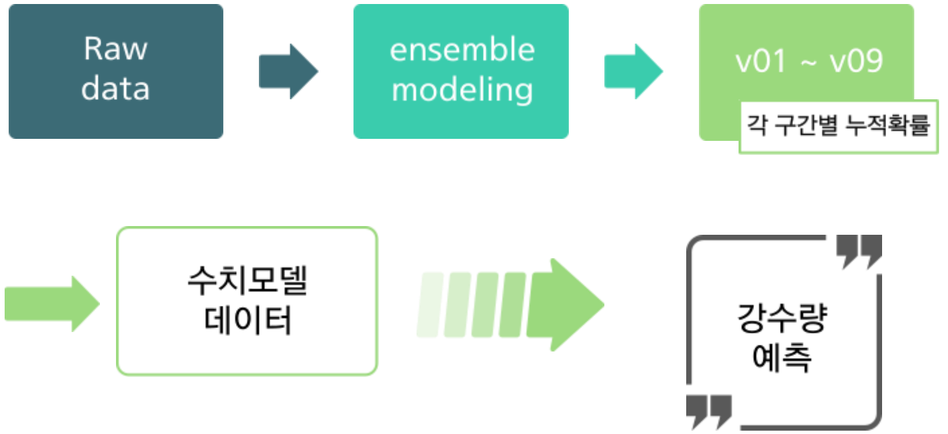
모델 검증

- 최종 모델 성능 검증
- 모델 결과 해석

변수 정의

변수	설명
TM_FC	기준 발표시각(년, 월, 일, 시)
TM_EF	예측시간(년, 월, 일, 시)
DH	기준시각 - 예측시각
STN	AWS 지점 코드
V01~V09	앙상블 모델의 구간별 강수량 누적확률
V00_ind~V09_ind	구간별 강수량 확률
basis_index	예측시각 및 관측소 (ef_year + month + dat + hour + stn)
V_expect	강수 기댓값
VV	지점별 1시간 강수자료의 3시간 누적 실강수량
class_interval	강수 계급

Ensemble



basis_index

- 정렬 기준 변수
- 예시 : 2021-05-01 12:00:00_STN001

v_expect

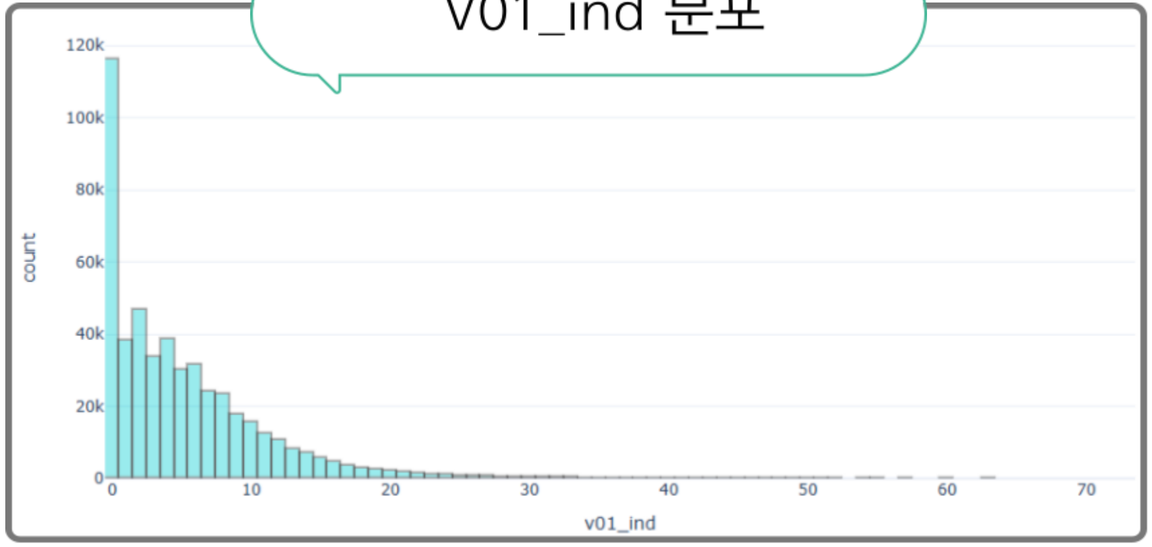
- (각 구간별 확률값 x 각 구간의 중앙값) 총합

EDA 시각화

연도별 강수량 분포

연도별 강수량 분포 상이

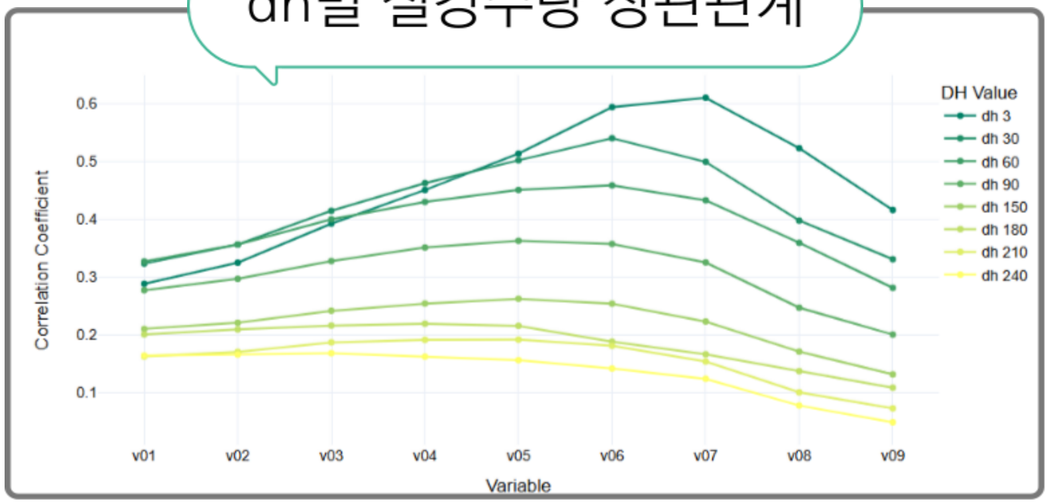
V01_ind 분포



변수들의 분포 skewed 상태

EDA+

dh별 실강수량 상관관계



dh가 작을수록 상관계수 증가

Attention Mechanism

특정 시점의 입력에 더 많은 가중치 부여

성능 저하

행 가중치

dh값이 클 수록 loss에 페널티 증가
dh가 작은 데이터를 더욱 신뢰

$$w = e^{-\alpha * dh / dh_{\max}}$$

성능 향상 미비 + 시간 소요

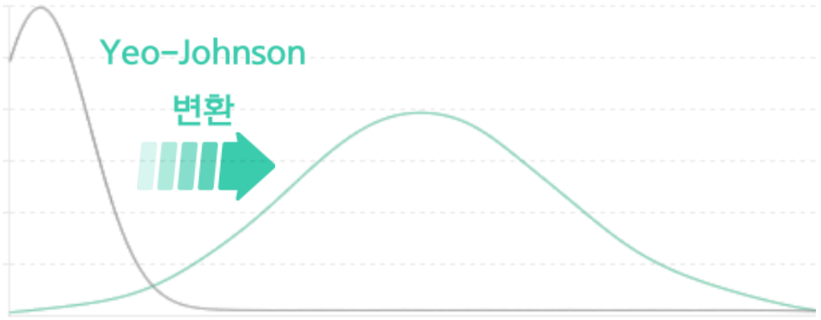
데이터 전처리

데이터 정렬

- 학습과정에서 시계열 요소를 유지
- basis_index 기준 내림차순 정렬
- 정렬 후 dh 기준 오름차순 정렬

Yeo-Jonhson 변환

- 모델에 사용하는 변수 대부분 skewed된 분포
- 표준화 작업 진행 ➡ 성능 향상



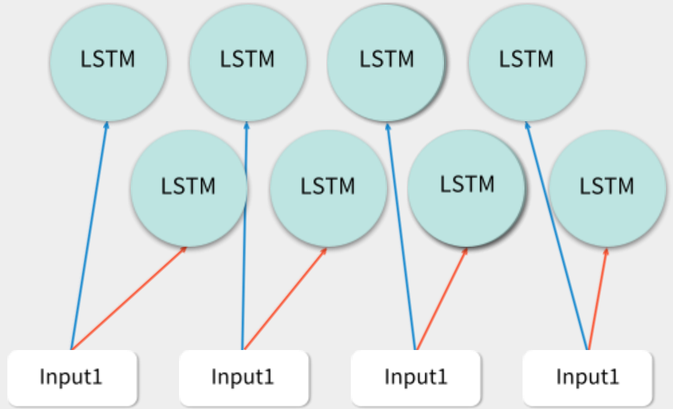
결측치 제거

실강수량 값이 -999인 행 제거

모델 구조 설계

1. Bidirectional LSTM

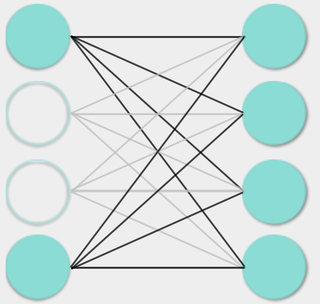
- LSTM 양방향 학습 => 성능 향상
- 시퀀스 증가 => 더 많은 정보 추출



2. Dropout

- 설정 비율에 따라 뉴런 비활성화
- 과적합 방지
- LSTM 모형 내부에 위치 => 시간 종속성 유지

ex)
dropout = 0.5



Custom Loss function

vv 예측 후 계급 변환 : label 순서에 민감

계급 구간 예측 : 불균형에 민감

batch

실제

vv	class_interval
0.8	3
0.17	2
13	7
1.5	4
0.11	2

예측

vv	class_interval
1.2	4
0.08	1
14	7
0.9	3
0.13	2

기존

CSI
0
0
1
0
1

행 단위로 계산

배치 단위로 계산

변경

CSI : 0.4

기존

RMSE: 0.55

Custom Loss function

RMSE와 CSI를 결합하여 손실함수 생성

$$RMSE + \beta(1 - CSI)$$

학습 방법

교차 검증

- A, B, A/B 앙상블 3개모델로 C년도 예측비교

▶ 직전 연도로 학습할 때 성능 향상

연도 분리

- 학습 : C년도
- 검증 : A,B년도 Average Loss

실강수량(vv) 예측 후
계급(class_interval) 변환

파라미터 설정

파라미터	값
epoch	60
Sequence Length	50
Batch Size	64
Beta	40
Learning Rate	0.0007
Optimizer	Adam
Active Function	sigmoid, tanh (Dense Layer : linear)

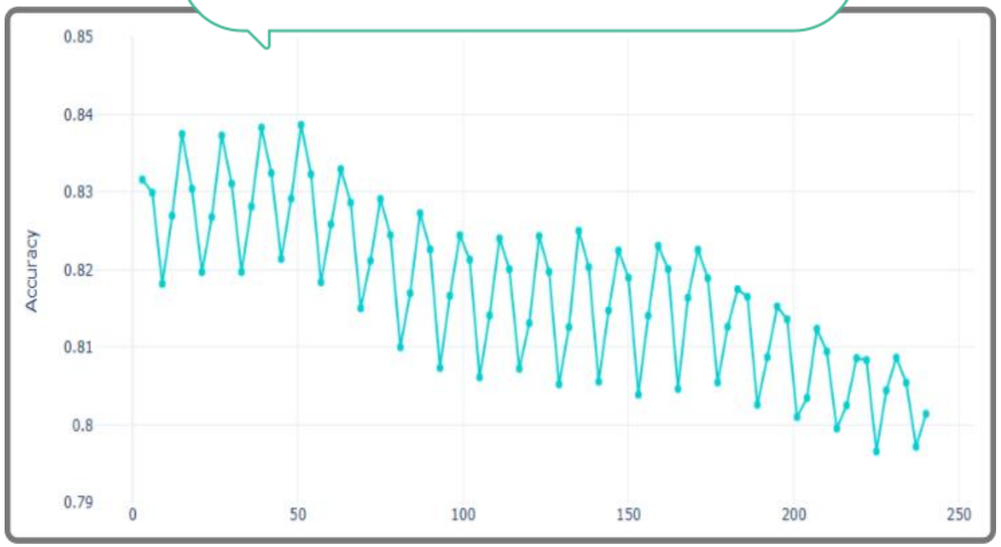
모델링 결과

Validation Set	CSI	RMSE	Loss	Average Loss
A년도	0.08917	4.760	41.1929	40.7288
B년도	0.08073	3.494	40.2648	

검증 결과의 CSI는 0.075이다.

모델 결과해석

dh별 정확도



"dh가 증가해도 높은 정확도를 보임"

구간별 평가지표

빗방울	0
약한비	1,2,3,4
보통비	5,6
강한비	7,8,9

출처. 기상청 날씨누리

"무작위 모형 대비 Recall, Precision 향상"

Recall : 미탐지 평가지표
Precision : 오탐지 평가지표

활용방안 및 기대효과

기상청 앙상블 모델 + Bidirection LSTM(+ Custom Loss function)

기술적 활용

- 시계열 모델을 사용함으로써 보다 정확하고 먼 시일의 강수 예측이 가능
- Custom Loss function을 활용하여 연속, 이산형 평가기준을 동시에 고려

실용적 활용

- 작물 피해예방 및 스마트 농업시스템 구축(자동 관개시스템 등)
- 외부 작업 및 행사 기획시 일정 조정 / 대체 계획 마련
- 오탐지로 인한 불필요한 비용 감소 + 경보 피로감 감소

참고문헌

Zaremba, Wojciech, Ilya Sutskever, and Oriol Vinyals. "Recurrent neural network
regularization." arXiv preprint arXiv:1409.2329 (2014).

Shi, Jimeng, Mahek Jain, and Giri Narasimhan. "Time series forecasting (tsf) using
various deep learning models." arXiv preprint arXiv:2204.11115 (2022).

홍성재, et al. "기계학습의 LSTM 을 적용한 지상 기상변수 예측모델 개발." 대기 31.1
(2021): 73-83.