2024 날씨 빅데이터 콘테스트

[과제 1] 수치모델 앙상블을 활용한 강수량 예측

팀명	Love MAE	접수번호	240011

1. 분석 주제

수치 모델 (3 시간 단위) 앙상블 강수 확률 자료를 활용하여 5 월에서 9 월의 누적 강수량(계급구간) 예측을 한다. 10 개의 강수계급(mm)을 반영하여 모델 구성 및 예측 수행을 목표로 하여 분석을 진행하였다.

2. 분석 데이터 처리 및 변수 설정

2-1. 기본 제공 데이터 분석

공모에 제시된 예측 자료는 수치 앙상블 계급별 강수 확률로 3 시간 단위 누적 강수량의 관측 값을 갖고 있다. 학습데이터의 기간은 A 년부터 C 년의 각 5 월에서 9 월까지의 데이터고 검증 데이터의 기간은 D 년 5 월에 9 월까지의 데이터로 특정 5 지점의 확률자료이다. 수치모델 데이터의 변수는 <표 1>과 같이 구성되어 있다.

변수	설명	변수	설명	변수	설명
TM_FC	기준 발표시각	V02	0.2mm 이상 누적 확률	V07	10.0mm 이상 누적 확률
TM_EF	예측 시간	V03	0.5mm 이상 누적 확률	V08	20.0mm 이상 누적 확률
DH	기준시각-예측시간	V04	1.0mm 이상 누적 확률	V09	30.0mm 이상 누적 확률
STN	AWS 지점 코드	V05	2.0mm 이상 누적 확률	VV	실강수량
V01	0.1mm 이상 누적 확률	V06	5.0mm 이상 누적 확률	class_interval	강수계급

<표 1> 수치모델 변수 구성

2-2. 파생 변수 설정

제공된 데이터를 기반으로 생성한 파생변수는 <표 2>와 같다.

변수	설명	변수	설명	변수	설명
fc_season	계절 변수	fc_hour_sin	기준 시각 시간 사이클릭	fc_hour_squared	시간 변수 제곱
ef_season		fc_hour_cos		ef_hour_squared	
fc_day_sin	기준 시각 'day' 사이클릭	ef_hour_sin	예측 시간 시간 사이클릭	fc_month_squared	월 변수 제곱
fc_day_cos		ef_hour_cos		ef_month_squared	
ef_day_sin	예측 시간 'day' 사이클릭	prob_sum	누적 확률 합계	fc_month_hour_interation	시간, 월 상호작용
ef_day_cos		prob_mean	누적 확률 평균	ef_month_hour_interation	

<표 2> 파생 변수

3. 데이터 파생 변수 설정

제공된 데이터를 기반으로 예측의 성능을 높일 수 있는 파생변수를 추가하였다. 추가적으로 결측치 처리도 진행하였다.

3-1. 결측치 처리

타겟 변수인 'rainfall_train.vv' 열의 결측치인 -999 를 모델 학습을 위해, vv 열의 최빈값인 0 으로 변경하였다.

3-2. 계절 변수 추가

계절(season) 정보를 생성하기 위해 변수를 추가하였다. 4 월, 5 월, 6 월 봄은 '2', 7 월, 8 월, 9 월 여름은 '3', 10 월, 11 월, 12 월 가을은 '4', 1 월, 2 월, 3 월 겨울은 '1' 로 각각 지정하여 'fc_season', 'ef_season' 변수를 생성하였다.

3-3. 사이클릭 변수 추가

주기적인 특성을 강조하기 위해 사이클릭(cyclic) 변수를 생성하였다. '월(month)과 시간(hour) 변수를 각각 sin 및 cos 변환을 해주었다.

3-4 누적 확률 합계 및 평균

누적 확률 값 v01 부터 v09 를 이용하여 각각 합(sum)과 평균(mean)을 구하여 'prob_sum', 'prob_mean' 변수를 생성하였다.

3-5 제곱 변수 추가

월(month)과 시간(hour) 변수를 제곱하여 변수들의 비선형성을 모델에 추가로 제공하였다. 특정 월이나 시간대가 강수량에 더 큰 영향을 미치는 경우 필요한 변수이다.

3-6 상호작용 변수

각 예보 발표 시간의 월(month)과 시간(hour)을 곱하여 새로운 상호작용(interaction) 변수 'interaction'을 생성하였다. 이 변수를 생성함으로써 월과 시간의 조합에 따른 특정한 패턴이나 변동을 모델이 학습할 수 있도록 도와준다.

4. 모델링 및 결과

랜덤 포레스트(Random Forest)는 앙상블 학습 알고리즘의 하나로, 다수의 결정 트리(decision tree)를 사용하여 예측 성능을 향상시키는 방법이다. 여러 트리를 결합하므로 개별 트리의 과적합을 줄여준다는 특징이 있다. 본 공모안에서는 당시 엑스트라트리(extratree)모델의 RMSE 가 2.37 로 더 낮았으나 검증 CSI 는 랜덤 포레스트 모델이 더 낮아 채택하게 되었다. 전처리 및 파생변수를 넣은 후 Train 세트를 8:2 로 split 한 후 랜덤포레스트 모델로 예측을 수행한 결과의 RMSE 이다. 'n_estimator'을 100 으로 설정하여 예측을 수행하였다. 결과, RMSE 는 2.837 로 확인되었다.