서울과학기술대학교
SEOULTECH Seoul National University of Science & Technology

서울과학기술대학교

Seoul National Univ. of Science & Technology

과 목 명 : 머신러닝
담당 교수 : 김성은 교수님
제 출 일 : 2025년 5월 13일
학 과 : 컴퓨터공학, 산업정보시스템전공
학 번 : 21101164, 23101936
성 명 : 곽용진, 은채웅

머신러닝 기반의 다음날 평균 기온 예측 모델 개발

개요

본 연구는 '머신러닝 기반의 다음날 평균 기온 예측 모델 개발'의 사례를 바탕으로 '데이터 전처리의 중요성'과 '앙상블 기법을 통한 예측 성능 향상'을 시행함으로써 머신러닝에 의한 기온 예측 전략이 기상 예보의 정확도 및 실용성에 어떻게 영향을 미치는지를 판별하고자 한다. 즉, 본 연구는 현대 기상 예측의 문제성을 직시하여 연구의 다양한 관점을 다각도에서 분석하고 그에 알맞은 해결방안을 제시하는 데 의의를 가진다. 본 연구는 국가기상청에서 제공하는 2019년부터 2024년까지의 기상 데이터를 활용하여 기온, 대기, 강수, 시간 관련 변수들을 분석하였다. 결측치 처리와 특성 공학, 시계열 특성 반영을 통한 데이터 전처리 과정을 거쳐 Random Forest, XGBoost, SVM 알고리즘을 비교 분석하였으며, Optuna를 활용한 하이퍼파라미터 최적화와 앙상블 기법 적용을 통해 모델 성능을 향상시켰다. 성능 평가 결과, XGBoost 앙상블 모델이 RMSE 1.2440°C로 가장 우수한 성능을 보였다. 그러나 극단적 기상 현상 예측의 어려움, 지역 특수성 반영의 한계, 단기 예측에 더 적합한 시간적 제약 등의 한계점도 존재한다. 향후 연구에서는 딥러닝 기반 시계열 모델과의 융합, 지역 특화 모델 개발, 위성 데이터 등 다양한 데이터 연계를 통해 더욱 정교한 예측 시스템을 구축할 필요가 있다.

서론

본 프로젝트의 목표는 머신러닝 기법을 활용하여 **다음날 평균 기온을 예측**하는 것이다. 수업에서 제공된 기상 데이터를 바탕으로, 평균 기온, 강수량, 습도, 이슬점 등 다양한 입력 변수를 사용하여 기온을 예측한다.

본 보고서는 다음과 같은 구조로 구성된다.

- **데이터 세트 설명**: 사용된 기상 데이터의 특성과 출처 소개
- **데이터 전처리 과정**: 결측치 처리 및 특성 공학
- **예측 모델 설계**: 선택된 머신러닝 알고리즘과 선정 이유
- **성능 평가**: 모델의 예측 정확도 및 성능 분석
- **팀원 역할 및 수행 내용**: 프로젝트 내 각자의 역할과 기여 내용

이 프로젝트를 통해 머신러닝 기술이 기상 예측에 어떻게 적용될 수 있는지, 그리고 데이터 과학의 실제 응용 사례를 학습하고자 한다.

데이터 세트 설명

본 프로젝트에서는 **국가기상청**에서 제공하는 공식 기상 관측 데이터 세트를 활용하였다. 데이터는 2019년부터 2024년까지의 일일 기상 관측 기록으로, 전국 8개 관측소에서 수집되었다. 주요 변수는 크게 네 가지 카테고리로 분류된다.

- **기온 관련 변수**: 직접적인 온도 측정값 또는 열적 특성과 연관된 변수로, 지면 온도와 이슬점 온도가 포함된다. 특히, 이슬점은 공기의 수분 함량을 온도로 환산한 값이지만, 열역학적 특성을 고려해 기온 관련 변수로 분류하였다.
- **대기 관련 변수**: 대기 상태를 나타내는 변수로, 증하층운량, 습도, 현지기압, 해면기압, 최저운고, 증기압, 시정, 풍속, 풍향, 일조 등이 포함된다.
- **강수 관련 변수**: 강수 및 적설과 직접적으로 관련된 변수로, 강수량과 적설이 해당된다.
- **시간적 변수**: 시간이나 계절적 특성을 반영하는 변수로, 날짜와 평균 기온이 포함된다. 평균 기온은 과거 동일 시기의 평균값으로, 시간적 맥락을 반영하는 특성으로 해석하였다.

이 데이터 세트는 다양한 기상 요소를 포함하고 있어, 머신러닝 기반의 기온 예측 모델 개발에 적합하다

데이터 전처리 과정

본 연구에서는 기상 데이터의 신뢰성과 예측 모델 성능 향상을 위해 다음과 같은 체계적인 전처리 과정을 수행한다.

결측치 처리

기상 데이터의 결측치는 대부분 관측 누락이나 센서 오류 등의 문제로 발생한다.

- **0으로 대체**: 일조량, 적설, 강수량, 시정, 풍속, 증하층운량 결측치는 '측정값 없음' 또는 '해당 현상 없음'을 의미하므로 0 처리한다.
- **최댓값 기반 보정**: 최저운고의 경우 구름이 전혀 없는 맑은 날에는 "관측할 구름이 없어서" 최저운고 값이 기록되지 않고 결측치로 남는 경우가 있으므로 해당 피처의 최댓값보다 조금 더 큰 값으로 처리한다.

특성 공학

- **총량 변수 생성**: 적설과 강수량, 시정, 일조의 경우에는 0시부터 21시까지의 총량을 갖는 새로운 변수를 추가한 뒤 기존 변수는 제거한다. 평균값은 분포가 넓거나, 일시적 극단값이 있을 때 정보를 희석시킬 수 있기 때문에 평균보다 총량(합계)을 선택한다. 또한 21시 이후의 시간은 다음 날과 가깝기에 다음 날 온도 예측에 영향이 높을 것으로 판단하여 해당 시간대는 있는 그대로 사용한다.
- **평균 변수 강조**: 운량과 풍속의 경우에는 평균을 새로운 변수로 추가한다. ECMWF 등 주요 기상 기관의 예측 모델 및 검증 연구에서 운량은 일평균값으로

다루는 것이 예측력과 해석력 모두에서 더 유리하다는 것을 확인할 수 있다. 풍속 또한 평균값이 기온 변화에 미치는 영향을 더 잘 설명하기에 평균값으로 한다.

- **과적합 방지 효과**: 위와 같은 전처리의 경우 변수의 수가 줄어들기에 과적합을 방지할 수 있는 효과도 기대할 수 있다.

시계열 특성 반영

- **주기성 변환**: 날짜는 선형적인 정수이지만, 실제 계절/날씨 패턴은 주기적이다. 즉, 1월 1일과 12월 31일은 숫자상 멀지만 날씨는 유사하다. 머신러닝 모델은 1월 1일과 12월 31일이 연결되지 않은 독립적인 값으로 인식하므로 날짜를 sin/cos 변환으로 두 날짜가 좌표 평면에서 가까운 위치에 배치되어 모델이 주기성을 자연스럽게 학습하도록 처리한다.

$$\text{Date}_{\sin} = \sin\left(\frac{2\pi \cdot \text{day\_num}}{365}\right) \tag{1}$$

$$\text{Date}_{\cos} = \cos\left(\frac{2\pi \cdot \text{day\_num}}{365}\right) \tag{2}$$

- **Magnus 공식 활용**: 이슬점 온도를 추정하는 Magnus 공식을 활용하여 이슬점과 상대습도로부터 기온을 역산해 시간대별 대기 기온 피처를 추가하며 보강하고, 모델의 입력 피처 다양성을 높인다. 이슬점(T_d)과 상대습도(RH)로부터 시간대별 기온(T) 역산하는 공식이다.

$$T = \frac{b \cdot \gamma(T_d, RH)}{a - \gamma(T_d, RH)}, \tag{3}$$

$$\gamma(T_d, RH) = \frac{a \cdot T_d}{b + T_d} + \ln(RH/100) \tag{4}$$

[그림 1]는 전처리 후 SHAP 값을 통해 분석한 상위 10개 피처 중요도를 보여준다. 평균 기온(1.38), Magnus 기반 시간대별 기온 중 23시가 가장 높은(0.74), sin/cos 날짜 변환이 각각 (0.29), (0.2) 등 주요 영향 요인으로 확인되었다.

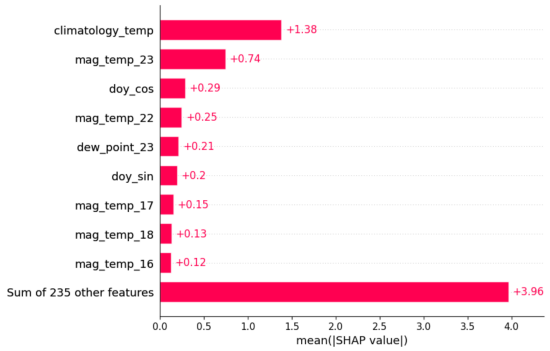

Figure 1: 상위 10개 피처 중요도 (SHAP 기준)

예측 모델 설계

정확한 기온 예측을 위해 다양한 머신러닝 알고리즘을 평가하고 최종 모델을 선정한다. 기상 데이터의 복잡한 비선형 특성을 고려하여 (Random Forest, XGBoost, SVM) 알고리즘들을 비교 분석한다.

알고리즘 비교

- **Random Forest**
  - 장점: 랜덤 포레스트는 비선형성 처리에 뛰어나며 과적합에 강하고, 특성 중요도 또한 제공해준다.
  - 단점: 고차원 데이터에서 계산 복잡도 증가한다.
- **XGBoost**
  - 장점: 높은 예측 정확도와 효율적인 학습, 결측치 처리에 뛰어난 능력을 가지고 있다.
  - 단점: 하이퍼파라미터 튜닝에 많은 시간이 소요된다.
- **SVM**
  - 장점: 차원 높은 데이터에서 우수한 성능을 보인다.
  - 단점: 대규모 데이터셋에서 학습 시간이 길어지는 특성을 가진다.

하이퍼파라미터 최적화

하이퍼파라미터의 튜닝은 optuna 라이브러리를 활용한다. optuna는 Bayesian Optimization 기반의 효율적인 탐색 전략과 scikit-learn과의 원활한 통합 지원을 제공한다.

- **튜닝 도구**: Optuna(Bayesian Optimization 기반)
- **주요 파라미터**:
  - 학습률(learning rate): 0.01–0.3
  - 트리 깊이(max_depth): 3–12
  - 부스팅 라운드(n_estimators): 100–1000
  - 정규화 파라미터(alpha, lambda)
  - 분할 최소 손실 감소량(gamma)

앙상블 기법 적용

단일 모델의 예측 변동성 감소와 일반화 성능 향상을 위해 앙상블 기법을 적용한다.

- **다양성 확보**: 다양성 확보를 위한 시드 기반 앙상블을 사용하며, 동일한 모델에 10개의 서로 다른 random seed를 적용한다.
- **결합 방식**: 앙상블 기법 과적합을 방지하고 예측 성능을 향상키시기 위해 Soft Voting을 이용하여 예측값 산술평균으로 최종 결과를 도출한다.

최종적으로 **XGBoost**가 RMSE 기준 최우수 성능을 보였으며, 앙상블 기법 적용 시 추가 성능 향상을 확인하였다.

성능 평가

본 연구에서는 모델 성능 평가를 위해 **평균 제곱근 오차(RMSE)**를 주요 지표로 사용하였다. RMSE는 다음과 같은 특징으로 인해 기온 예측 평가에 적합하다.

- **직관적 해석**: 예측 오차의 평균적인 크기를 실제 기온 단위(°C)로 바로 보여주기에 직관적인 해석에 용이하다.
- **대오차 강조**: 오차를 제곱해서 평균을 내기 때문에, 큰 오차에 더 민감하게 반응하는 특징이 있다. 이는 실제 기온 예측에서 큰 오차가 중요한 영향을 미치는 경우, 모델의 성능을 더 엄격하게 평가할 수 있게 해준다.
- **산업 표준**: 실제로 기온, 강수량 등 연속적인 수치 예측에서 RMSE는 국내외 기상청, 논문, 산업 현장에서 널리 쓰이는 표준 평가 지표이다.

모델별 성능 비교

- **XGBoost 앙상블**: RMSE 1.2440 °C (최우수)
- **Random Forest**: RMSE 1.4120 °C
- **SVM**: RMSE 2.8543 °C

교차 검증 결과

모델이 다양한 데이터 서브셋에서 일관된 성능을 보이는지 혹인하기 위해 5-폴드 교차 검증을 진행한다. 그 결과 평균 RMSE: 1.2916 ± 0.0268로 이는 일관된 성능을 보인다고 해석할 수 있다.

성능 평가 결론

XGBoost 앙상블 모델의 RMSE 1.2440 °C는 기상 예측 분야에서 우수한 수준으로 평가되며, 교차 검증을 통해 일반화 가능성을 확인하였다.

팀원 역할 및 수행 내용

본 프로젝트는 2명의 팀원이 협력하여 수행하였으며, 각자의 역량과 노력을 최대한 발휘하여 머신러닝 기반 기온 예측 모델을 개발하였다. 주요 역할 분담과 수행 내용은 다음과 같다.

공통 수행 항목

- 데이터 도메인 지식에 대한 학습 및 전체적인 코드 작성을 수행하였다.
- 주어진 데이터에 모델을 최적화하기 위해 지속적인 협업 및 피드백을 수행하였다.

21101164 곽용진

주요 역할: 코드 구현 & 최적화, 데이터 시각화, 하이퍼파라미터 최적화, 과적합 방지를 담당했다.

- **시각화 시스템 구축**: SHAP 값을 기반으로 피처 중요도를 분석하고, matplotlib 라이브러리를 활용하여 결과를 시각화한다.
- **코드 최적화**: 전체적인 코드 속에서 세부적인 코드 작성 및 수정을 진행하며, 전처리 & 모델링 파이프라인 최적화한다.
- **하이퍼파라미터 튜닝**: 팀원이 선택한 최적의 모델을 바탕으로 optuna 라이브러리를 통해 Bayesian Optimization 기반 탐색으로 하이퍼파라미터 튜닝을 진행한다.
- **과적합 진단**: . 교차 검증 진행 후 검증 데이터 세트와 훈련 데이터 세트의 평가 지표(RMSE)의 차이를 확인하여 과적합을 판단한다.

23101936 은채웅

주요 역할: 데이터 전처리 및 피처 공학, 최적의 모델 선택을 위한 비교분석을 담당했다.

- **결측치 처리 전략**: 자연적인 결측치의 처리와 모델의 복잡도 감소를 위해 피처의 합계와 산술평균을 활용한다.
- **시계열 피처 설계**:
  - 다음날의 평균 기온 예측이라는 시계열 특성을 고려하여 날짜를 sin/cos 변환하여 주기성 반영한다.
  - Magnus 공식을 통해 이슬점과 습도를 기반으로 기온을 역산하는 등의 피처 공학을 진행한다.
- **모델 선정**: 평균 기온 예측이라는 문제를 해결하기 위한 최적의 모델을 선택하기 위해 SVM과 Random forest, XGBoost 모델들의 비교분석을 통해 최종적으로 XGBoost 모델을 선택한다.
- **차원 축소**: 총량/평균 계산을 통한 피처의 수를 감소시킨다.

협업 전략 & 성과

- **주요 도전 과제**:
  - 복잡한 기상 데이터 도메인 이해
  - 개인별 코딩 스타일 차이 조율
  - 제한된 시간 내 고성능 모델 개발
- **해결 방안**:
  - 주 1회 이상 정기적인 팀 미팅
  - Github을 통한 서로의 코드 공유 및 추가적인 작업 기록
  - 명확한 역할 분담과 지속적인 소통

결론 및 한계점

연구 성과

- XGBoost 기반 앙상블 모델을 통해 **RMSE 1.2440 °C**의 우수한 성능 달성했다.
- 머신러닝 기반 기온 예측 모델 개발 프로세스에 대한 이정표를 제공한다.

한계점

- **극단적 기상 현상 예측**: XGBoost는 과거 데이터에 기반한 통계적 학습에 의존하므로 급격한 기후 변화나 극단적인 기상 현상을 완벽하게 예측하기 어렵다.
- **지역 특수성**: 지역적 특수성을 완전히 반영하지 못하는 한계가 있다.
- **시간적 제약**: 장기 예측보다는 단기 예측에 더 적합한 구조를 가지고 있다.

대안 제안

- **융합 모델 구축**: 딥러닝 기반의 시계열 모델(LSTM)을 통합한다.
- **지역 특화 모델**: 다양한 지역의 기후 특성을 반영하는 세분화된 모델을 구축한다.
- **다중 데이터 연계**: 위성 데이터나 레이더 관측 정보 등의 추가 데이터 소스를 활용한다.

향후 연구 방향

- 향후 연구는 머신러닝과 기후과학의 융합을 통해 더욱 정교하고 신뢰할 수 있는 **예측 시스템**을 구축하기 위해 집중해야 한다.

고찰

이번 프로젝트에서는 다음날 평균 기온을 예측하는 모델을 개발하고, 실제 데이터를 기반으로 예측 성능을 평가하였다. 모델 개발 과정에서 다양한 기상 변수와 데이터 전처리, 그리고 모델링 기법을 적용하며 여러 시행착오를 겪었다. 모델의 성능을 높이기 위해 결측치 처리, 변수 선택, 새로운 변수 생성 등의 다양한 전처리 과정을 거쳤다. 특히, Magnus 공식을 활용해서 기온을 역산하는 변수 생성을 통해 모델의 성능을 크게 개선하는 경험을 했고, 이를 통해 도메인 데이터에 대한 이해와 특성 공학의 중요성에 대해서 실감했다. 또한 feature의 순서와 구성이 일치하지 않아 ValueError가 발생하는 등 데이터의 일관성 확보의 중요성을 깨달았다. 예측 결과를 분석해보면, 기온 예측 자체에서는 비교적 안정적인 성능을 보였으나, 극단적인 상황에서는 오차가 커지는 경향이 나타난다. 이를 보안하기 위해서 추가적인 외부 변수 도입과 같은 추가적인 연구가 필요하다.

References

[1] Kaggle. *Next Day Air Temperature Forecast Challenge 2*. https://www.kaggle.com/competitions/next-day-air-temperature-forecast-challenge-2/data

[2] Patrick Urbanek. *Feature Engineering for Time Series*. https://www.kaggle.com/code/patrickurbanke/feature-engineering-for-time-series

[3] scikit-learn. *Cyclical Feature Engineering Example*. https://scikit-learn.org/stable/auto_examples/applications/plot_cyclical_feature_engineering.html

[4] meteoblue. *Forecast API Documentation*. https://docs.meteoblue.com/en/weather-apis/forecast-api/forecast-data

[5] ECMWF. *Skill of ECMWF Cloudiness Forecasts*. https://www.ecmwf.int/sites/default/files/elibrary/2015/17326-skill-ecmwf-cloudiness-forecasts.pdf

[6] 기상청. *지상기상관측지침*. https://data.kma.go.kr/resources/images/publication/%EC%A7%80%EC%83%81%EA%B8%B0%EC%83%81%EA%B4%80%EC%B8%A1%EC%A7%80%EC%B9%A8.pdf

[7] XGBoost. *Parameter Reference (v3.0.0)*. https://xgboost.readthedocs.io/en/release_3.0.0/parameter.html

[8] Optuna. *Official Tutorial*. https://optuna.readthedocs.io/en/stable/tutorial/index.html

[9] scikit-learn. *Random Forest Regressor*. https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html

[10] scikit-learn. *Support Vector Regressor (SVR)*. https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVR.html