**시계열데이터 분석 보고서**

2022113170 AI융합학부 장정의

1. **데이터셋의 선정 이유 및 시계열 모델이 필요한 이유**

선정한 데이터셋은 ‘서울시 연평균 여름 습도 데이터셋’이다. 우선 여름 습도 데이터는 모든 계절의 데이터를 살펴보는 것이 아니라 습도가 높게 느껴지는 여름의 데이터 추이를 살펴보고자 선정하게 되었다. 선정한 지역인 서울은 대한민국의 수도로, 인구 밀도가 높은 중심지이다. 이에 수도권을 비롯해 국가 전체의 기후 변화를 대표하기에 서울이 가장 적합할 것이라고 생각하였다. 또한, 서울의 기상 데이터는 오랜 기간 체계적으로 수집되어 왔다. 데이터가 부족한 일부 지역과 달리, 서울은 분석하고자 하는 1954년부터 2018년까지의 기상 데이터를 충분히 확보하고 있어 분석에 적합하였다.

습도는 시간의 흐름에 따라 변동하며 일정한 추세를 보였다. 이러한 장기적인 감소 추세를 파악하고, 과거의 값을 기반으로 미래의 값을 예측하기 위하여 시계열 모델이 필요하다. 또한, ADF 테스트의 p-value가 0.97의 높은 값을 가지고 있다는 점과 이동평균 함수의 평균 및 표준편차 시각화 결과를 통해 데이터셋이 비정상성을 가지고 있음을 알 수 있다. 이는 차분을 통해 데이터셋을 안정적인 패턴으로 변화시키며, ADF 테스트 p-value를 0에 가깝도록 바꿀 수 있다. 따라서 위 ‘서울시 연평균 여름 습도 데이터셋’은 시계열 모델을 통해 분석한다.

1. **적합한 모델의 선정**

차분한 데이터셋의 ACF 그래프는 1에서 신뢰구간을 벗어난 값을 보이며, PACF 그래프는 2까지 신뢰구간을 벗어난 값을 보인다. 이러한 자기상관을 모델링하기 위해 ARIMA 모델을 사용하였다. 신뢰구간에 최초로 진입하는 시점을 선택하여 최적의 모델을 ARIMA(2,1,3)으로 예상하였다. 그러나 이는 타 모델에 비해 MSE가 낮지만, 매우 복잡하여 과적합이 발생할 수 있다. 이에, 자동적으로 파라미터를 탐색하여 최적의 모델을 선정해주는 Auto ARIMA를 통해 적합한 모델로 ARIMA(0,1,1)을 선정하였다.

1. **결과에 대한 고찰**

모델을 분석해보면, 주요 지표인 AIC, BIC값이 비교적 낮고, 분산의 일정성 검정 통계값이 유의미하다. 잔차는 평균 0을 중심으로 고르게 분포하고 있으며 히스토그램과 Q-Q 플롯 결과 정규성을 가진다. 또한, 현재 그래프에서 자기상관 계수들이 신뢰 구간 내에 있으므로 잔차는 독립적이다. 결론적으로, ARIMA(2,1,1) 모델은 데이터에 잘 적합되었고, 신뢰할 수 있는 예측을 제공할 가능성이 높다. 예측 결과, 예측된 값은 훈련 데이터의 평균에 가까운 값을 주로 나타내어, 테스트 데이터의 실제 변동성은 반영하지 못하고 있다. 이를 개선하기 위해 한 지점에 대한 예측을 진행하고 모델 업데이트를 반복하여 훈련 데이터에 덜 적합하고 테스트 데이터의 변동성을 더 잘 반영하도록 한다. 예측 모델의 오차는 0.037로 매우 낮다.