항공사 데이터를 이용한 시계열 예측 모델 비교

1. 서론

시계열 분석은 데이터 과학 및 통계 분야에서 매우 중요하며, 금융, 기상, 경제, 제조 등 다양한 분야에서 널리 활용되고 있다. 시계열 분석 기법은 과거 데이터를 통해 미래를 예측하거나 패턴을 이해하여 다양한 인사이트를 도출해 준다. 고전적인 통계 모델에서부터 최신의 인공신경망 기반 모델에 이르기까지 시계열 예측 기법은 꾸준히 발전해 왔다. 본 보고서에서는 항공사 데이터를 활용하여, 통계 모형인 SARIMA와 여러 딥러닝 모델(RNN, LSTM, GRU)을 적용해 시계열 데이터를 예측하고 각 모델의 성능을 비교하는 과정을 다룬다.

SARIMA 모델은 시계열 데이터의 추세와 계절성을 반영할 수 있는 대표적인 통계적 방법으로, 예측의 안정성과 해석력을 갖춘 모델이다. RNN, LSTM, GRU와 같은 딥러닝 기반 모델들은 복잡한 시계열 패턴을 학습하고, 비선형적인 관계를 학습하여 다양한 유형의 데이터를 더욱 정확하게 예측할 수 있다.

본 보고서의 목표는 항공사의 승객 수 데이터를 이용하여 다양한 모델의 예측 결과를 비교하고, 각 모델이 가지는 장단점을 분석함으로써 가장 우수한 예측 성능을 보이는 모델을 도출하는 것이다. 이를 위해, SARIMA와 RNN, LSTM, GRU 모델을 각각 훈련하고, 이를 기반으로 예측 성능을 평가하였다. 평가 지표로는 평균제곱오차(MSE) 사용하여 각 모델의 정확성을 비교하였다.

2. 시계열 예측을 위한 모델

2.1 SARIMA

계절 자기회귀 누적 이동 평균(Seasonal Autoregressive Integrated Moving Average, SARIMA)은 ARIMA 모델의 확장으로, 계절적인 패턴을 포함한 시계열 데이터를 모델링하기 위해 고안된 통계적 모델임. SARIMA는 자기회귀(AR), 차분(I), 이동 평균(MA) 요소에 계절적 자기회귀, 계절적 차분, 계절적 이동 평균을 추가하여 계절성과 비계절성을 모두 설명할 수 있음. SARIMA 모델은 계절적인 주기와 비정상성을 처리하기 위해 로그 변환과 차분을 포함한 전처리 작업이 필요하며, 계절적 패턴을 포함한 다양한 시계열 데이터에 적합. 이 모델은 선형 추세와 계절적 변동을 효과적으로 캡처할 수 있어 예측에 널리 사용됨.

2.2 RNN

순환 신경망(Recurrent Neural Network, RNN)은 시계열 데이터와 같이 순차적인 의존성이 있는 데이터를 모델링하기 위해 고안된 인공신경망의 한 종류. RNN은 이전 단계의 출력을 현재 단계의 입력으로 사용하는 반복 구조를 가지며, 이를 통해 시간의 흐름에 따른 정보를 반영할 수 있음. 이러한 특성 덕분에 RNN은 순차 데이터의 패턴을 학습하는 데 강점을 가지며, 자연어 처리, 음성 인식, 시계열 예측 등의 다양한 분야에서 활용되고 있음. 그러나 RNN은 긴 시계열 데이터의 경우 장기 의존성을 학습하는 데 어려움이 있을 수 있으며, 이 문제를 완화하기 위해 LSTM 및 GRU와 같은 개선된 구조가 제안됨.

2.3 LSTM

장단기 메모리(Long Short-Term Memory, LSTM)는 RNN의 한 종류로, 시계열 데이터의 장기 의존성을 효과적으로 학습하기 위해 고안된 구조임. LSTM은 정보의 흐름을 제어하는 셀 상태(cell state)와 게이트 메커니즘(입력 게이트, 출력 게이트, 망각 게이트)을 도입하여 중요한 정보는 유지하고 불필요한 정보는 버리도록 함. 이러한 특성 덕분에 LSTM은 긴 시계열 데이터에서도 중요한 패턴을 학습할 수 있어, 자연어 처리, 음성 인식, 금융 데이터 예측 등에서 높은 성능을 보여주고 있음.

2.4 GRU

게이트 순환 유닛(Gated Recurrent Unit, GRU)은 LSTM의 변형된 구조로, 계산 효율성을 높이기 위해 간소화된 게이트 구조를 가지고 있음. GRU는 망각 게이트와 입력 게이트를 결합하여 업데이트 게이트로 사용하며, 셀 상태를 별도로 두지 않는 간단한 구조를 통해 LSTM보다 적은 수의 파라미터로도 유사한 성능을 보임. GRU는 LSTM에 비해 학습 속도가 빠르고, 계산 자원이 제한된 상황에서 장기 의존성을 처리하는 데 강점을 가지고 있음.

3. 데이터 분석

3.1 데이터 소개

본 보고서에서 사용한 데이터는 1949년 1월부터 1960년 12월까지의 국제 항공사 월별 탑승객 데이터를 기반으로 하였음. 이 데이터는 시계열의 특성을 가지고 있으며, 월별 증가하는 추세와 계절성이 존재함. ARIMA 모델에서는 로그 변환과 1차 차분을 통해 추세와 계절성을 제거하고, 신경망 모델에서는 Standard Scaling만을 적용하여 예측을 수행하였음.

OBS	month	x
1	1	112
2	2	118
3	3	132
4	4	129
5	5	121
6	6	135
7	7	148
8	8	148
9	9	136
10	10	119
11	11	104
12	12	118
13	1	115
14	2	126
15	3	141

그림 1 항공사 데이터

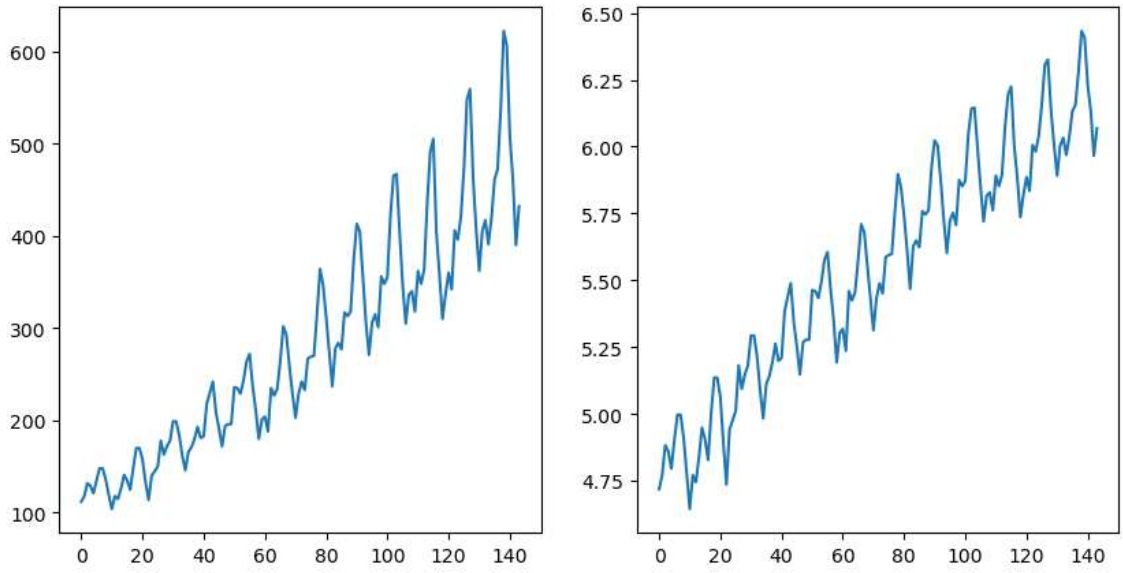


그림 2 데이터의 추세와 계절성

3.2 분석 결과 및 결론

SARIMA 모델은 로그 변환과 1차 차분을 통해 시계열 데이터의 추세와 계절성을 제거하여 정상 시계열로 변환한 후, ACF와 PACF 분석을 통해 AR, MA의 차수를 선정하였음. 1월부터 12월까지의 1년 단위 데이터를 사용하여 계절 주기를 12로 설정했고, 모델 평가 지표로 MSE, AIC, BIC를 사용함.

RNN, LSTM, GRU 모델은 모두 Standard Scaling을 통해 데이터를 표준화한 뒤, 시퀀스를 3으로 설정하여 이전 3개의 시점 데이터를 사용해 다음 시점 값을 예측하도록 함, 전체 데이터셋은 8:2의 비율로 train과 test 데이터로 분리했고, 각 모델의 학습 파라미터로는 epochs를 1000, learning rate를 0.001, optimizer를 Adam으로 설정하여 학습을 진행함. 성능 평가는 MSE를 기준으로 하였음.

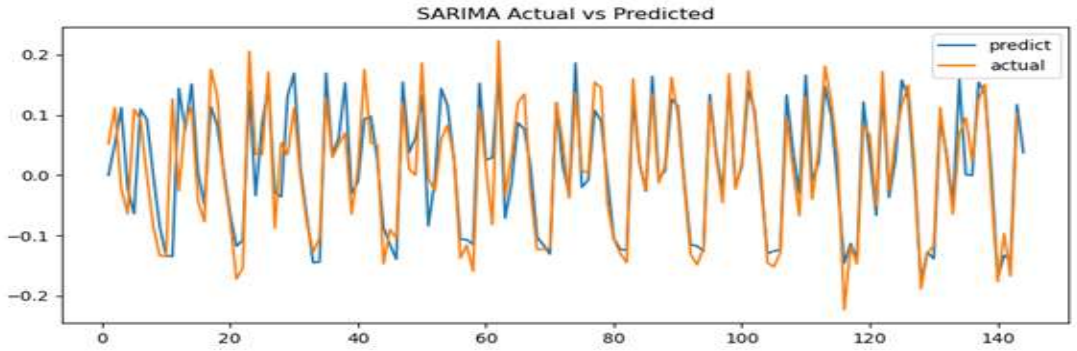
Table 1. 시계열 예측을 통한 각 모델 비교

	MSE	AIC	BIC
SARIMA	0.016	-451.20	-436.82
RNN	0.095	-	-
LSTM	0.112	-	-
GRU	0.094	-	-

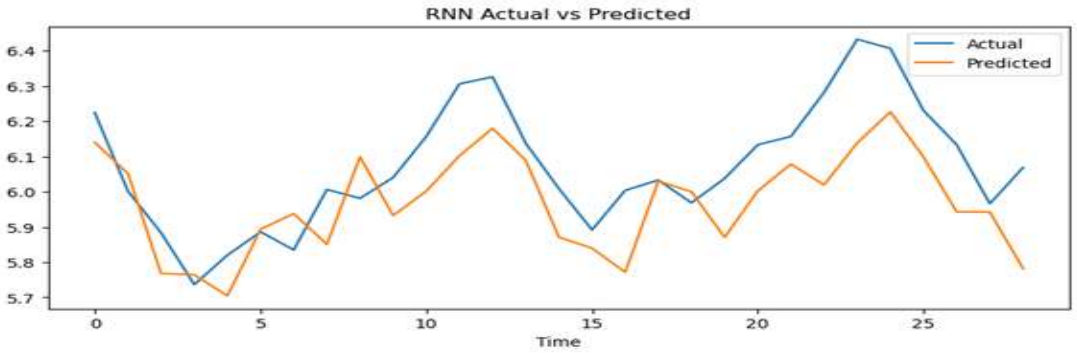
예상과 달리 LSTM과 GRU 모델이 아닌 SARIMA 모델이 MSE 0.016으로 가장 낮은 오차를 보였음.

결론적으로, SARIMA 모델이 이 데이터셋에서는 가장 높은 예측 성능을 보인다는 결과를 도출됨. 이는 데이터의 특성상 명확한 계절성과 추세가 존재해, 이를 효과적으로 반영할 수 있는 SARIMA 모델이 딥러닝 기반의 모델보다 적합했기 때문으로 보임. 반면, RNN, LSTM, GRU와 같은 딥러닝 모델은 비선형성을 다루는 데 강점이 있지만, 비교적 작은 데이터셋과 명확한 계절성을 가진 데이터에서는 오히려 SARIMA보다 성능이 저하될 수 있다는 것을 보임.

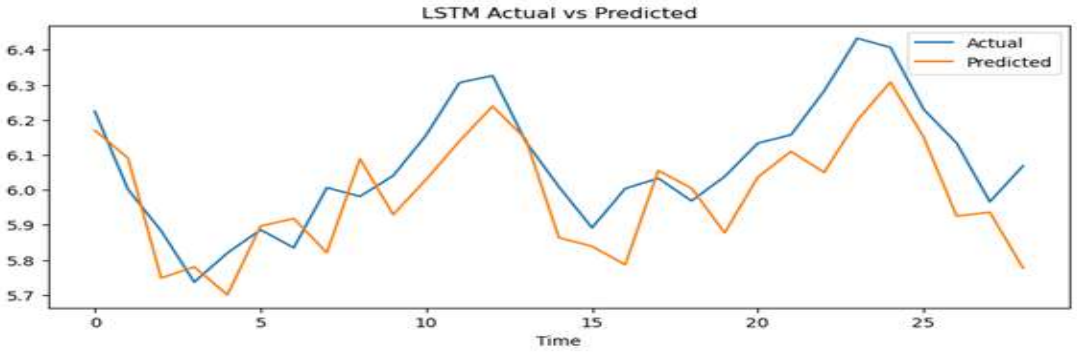
SARIMA 모델의 예측값 실제값 비교



RNN 모델의 예측값 실제값 비교



LSTM 모델의 예측값 실제값 비교



GRU 모델의 예측값 실제값 비교