20XD88 - Data Mining Lab Package Report

Weather Forecasting

Shakthi Mahendira K , Mahitej Reddy K , Rishi Ram S, Santhosh V

20PD10, 20PD14, 20PD22, 20PD25

15 th April, 2024

Introduction

Weather forecasting has undergone significant evolution with the integration of advanced machine learning techniques. By leveraging Seasonal Autoregressive Integrated Moving Average (SARIMA) and Long Short-Term Memory (LSTM) models, accurate predictions can be made. SARIMA captures seasonal patterns while LSTM learns long-term dependencies, enabling precise short to medium-term forecasts.

The synergy of SARIMA and LSTM enhances prediction accuracy, benefiting various sectors reliant on weather forecasts. This sophisticated approach not only improves the accuracy of weather predictions but also enhances our understanding of complex atmospheric dynamics. By leveraging SARIMA and LSTM models, meteorologists can better anticipate weather phenomena, mitigate risks, and support disaster preparedness efforts.

Furthermore, the integration of machine learning techniques fosters continuous innovation in weather forecasting, paving the way for more resilient and adaptive solutions in an ever-changing climate landscape.

Dataset

Area: Indicates the geographical area associated with the data.

Months: Represents the months or time periods covered by the data.

Element: Specifies the element or characteristic being measured or observed.

Y1961 to Y2019 : Contain data values corresponding to the respective years from 1961 to 2019.

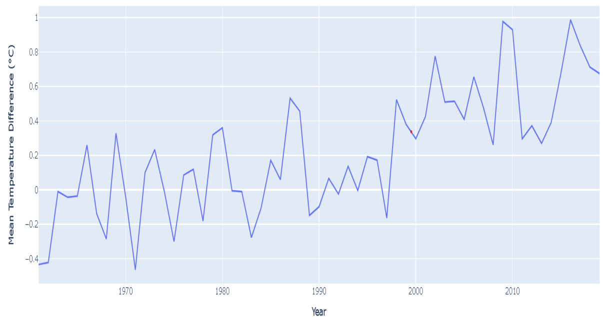
	Area	Months	Element	Y1961	Y1962	Y1963	Y1964	Y1965	Y1966	Y1967	...	Y2010	Y2011	Y2012	Y2013	Y2014	Y2015	Y2016	Y2017	Y2018	Y2019
0	Afghanistan	January	Temperature change	0.777	0.062	2.744	-5.232	1.868	3.629	-1.432	...	3.601	1.179	-0.583	1.233	1.755	1.943	3.416	1.201	1.996	2.951
1	Afghanistan	January	Standard Deviation	1.950	1.950	1.950	1.950	1.950	1.950	1.950	...	1.950	1.950	1.950	1.950	1.950	1.950	1.950	1.950	1.950	1.950
2	Afghanistan	February	Temperature change	-1.743	2.465	3.919	-0.202	-0.096	3.397	0.296	...	1.212	0.321	-3.201	1.494	-3.187	2.699	2.251	-0.323	2.705	0.086
3	Afghanistan	February	Standard Deviation	2.597	2.597	2.597	2.597	2.597	2.597	2.597	...	2.597	2.597	2.597	2.597	2.597	2.597	2.597	2.597	2.597	2.597
4	Afghanistan	March	Temperature change	0.516	1.336	0.403	1.659	-0.909	-0.069	-0.759	...	3.390	0.748	-0.527	2.246	-0.076	-0.497	2.296	0.834	4.418	0.234

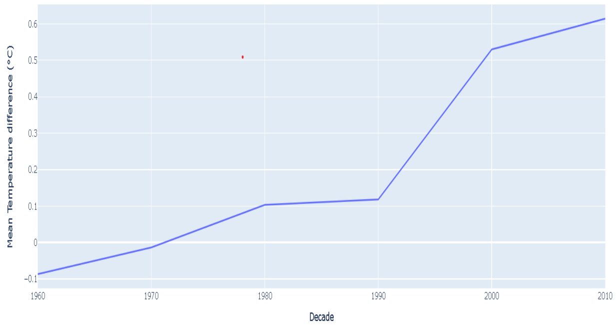
Dataset After Preprocessing

	Area	Months	Element	Year	Temperature
0	Afghanistan	January	Temperature change	1961	0.777
1	Afghanistan	February	Temperature change	1961	-1.743
2	Afghanistan	March	Temperature change	1961	0.516
3	Afghanistan	April	Temperature change	1961	-1.709
4	Afghanistan	May	Temperature change	1961	1.412

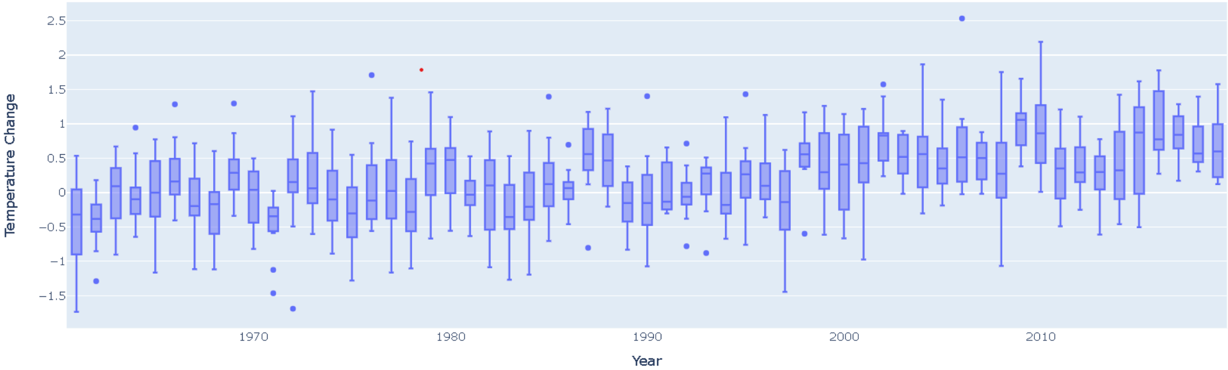
Exploratory Data Analysis

Comparative Analysis of Mean Temperature for India: Decade-Wise Trends





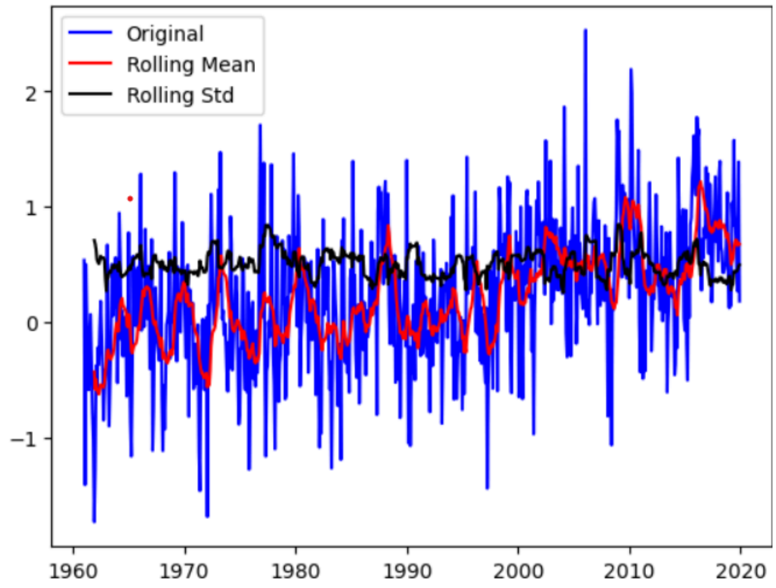
Temperature change Distribution by year in India



Time series analysis

Rolling mean : Rolling mean, a statistical technique, calculates the average of a sequence of data points by creating a series of averages of different subsets of the full dataset.

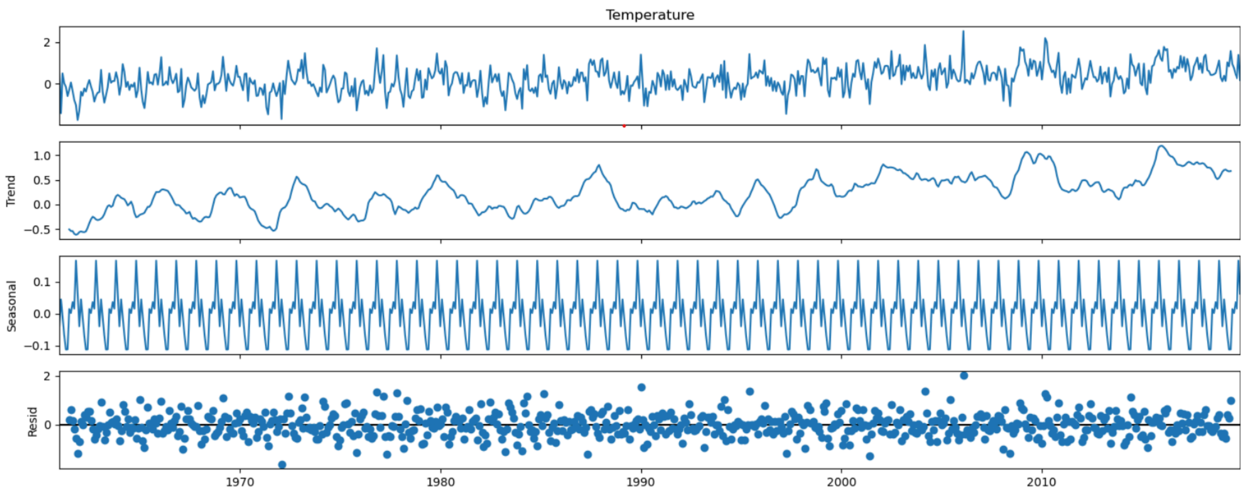
Rolling standard deviation:Rolling standard deviation, a statistical measure, computes the variability of a sequence of data points over a rolling window by determining the standard deviation of different subsets of the full dataset



Residuals capture the deviations between actual and predicted values, aiding in evaluating model accuracy and identifying patterns in the data.

Seasonality highlights recurring patterns or fluctuations that occur at regular intervals within time series data, offering insights into cyclic trends and seasonal effects.

Trend signifies the long-term directional movement observed in time series data, reflecting underlying patterns or structural changes over time.

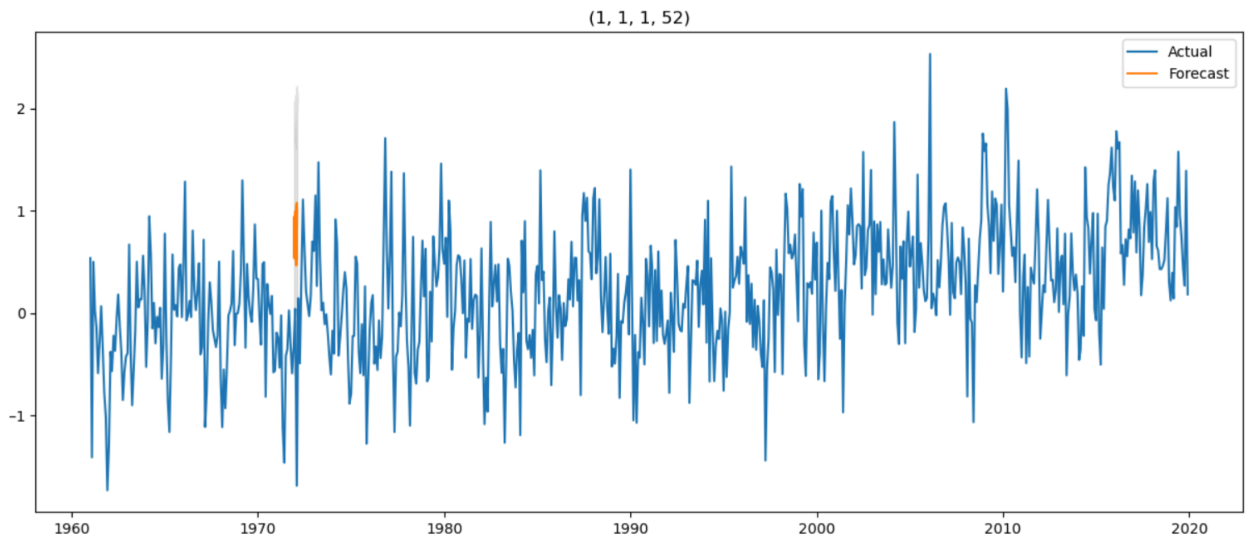


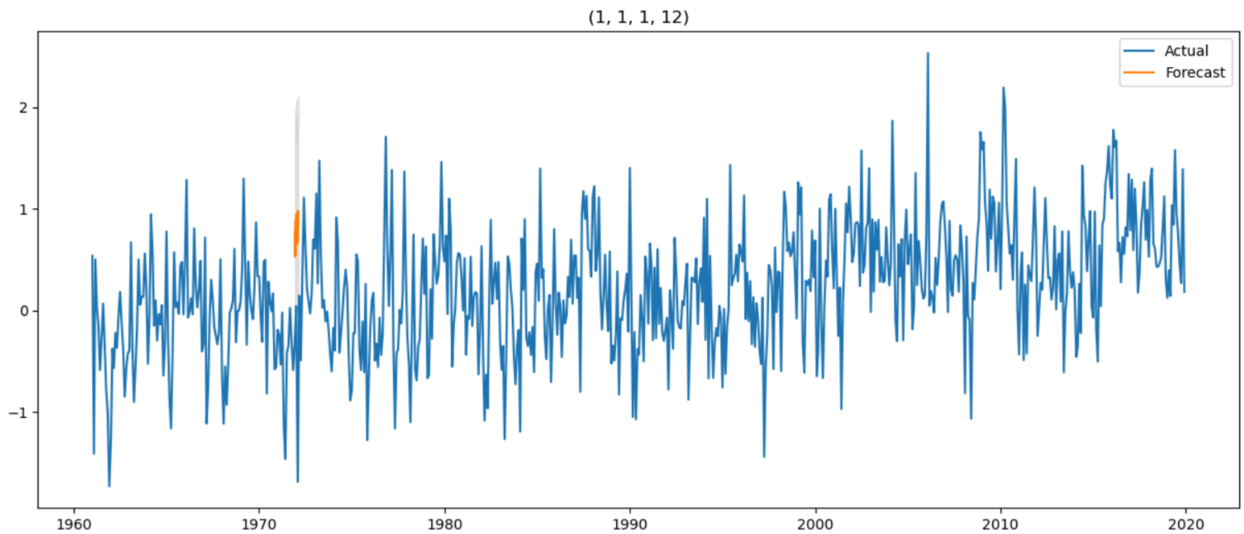
SARIMAX

SARIMAX, an extension of the SARIMA model, incorporates exogenous variables to enhance forecasting capabilities, making it suitable for time series data influenced by external factors.

By integrating exogenous variables, such as weather conditions or economic indicators, SARIMAX captures additional information that may impact the time series, resulting in more accurate and robust forecasts.

The below plots are for different p,d ,q values





From the above the plot with (1,1,1,12) has the best forecast

LSTM

LSTM, a type of recurrent neural network (RNN), excels in capturing long-term dependencies in sequential data.

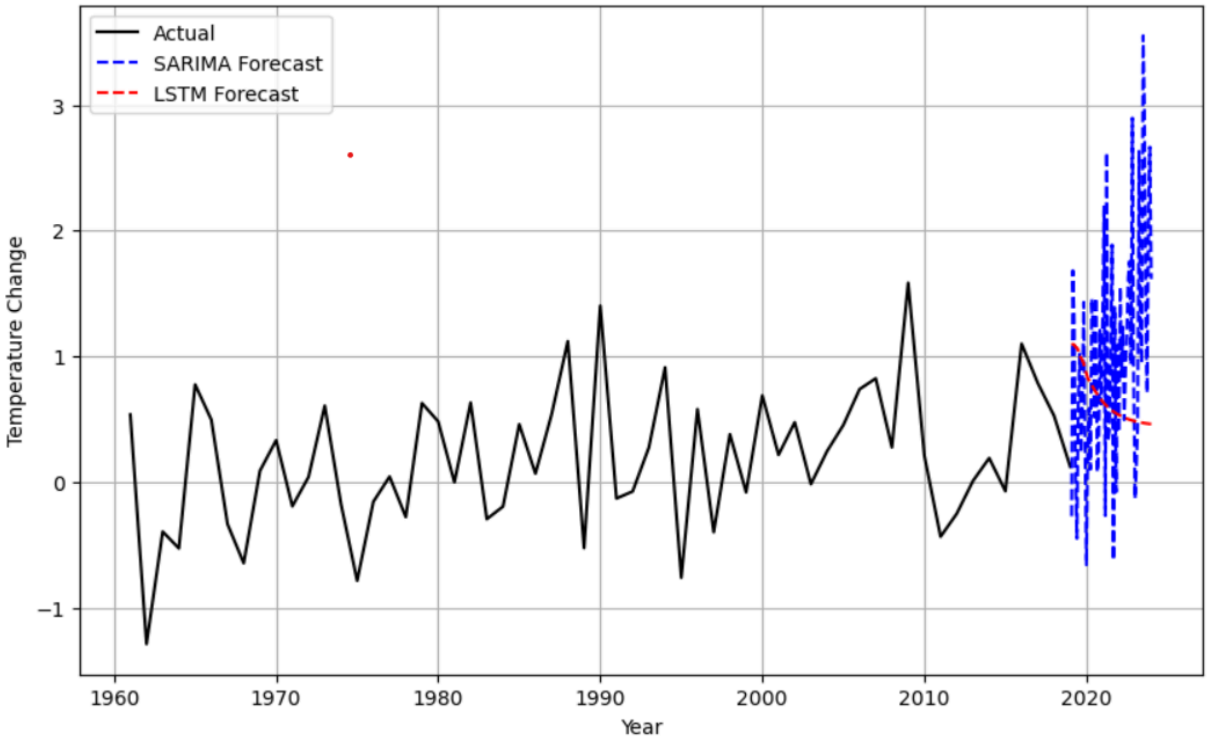
Its specialized architecture includes memory cells that enable it to learn and adapt to different temporal patterns.

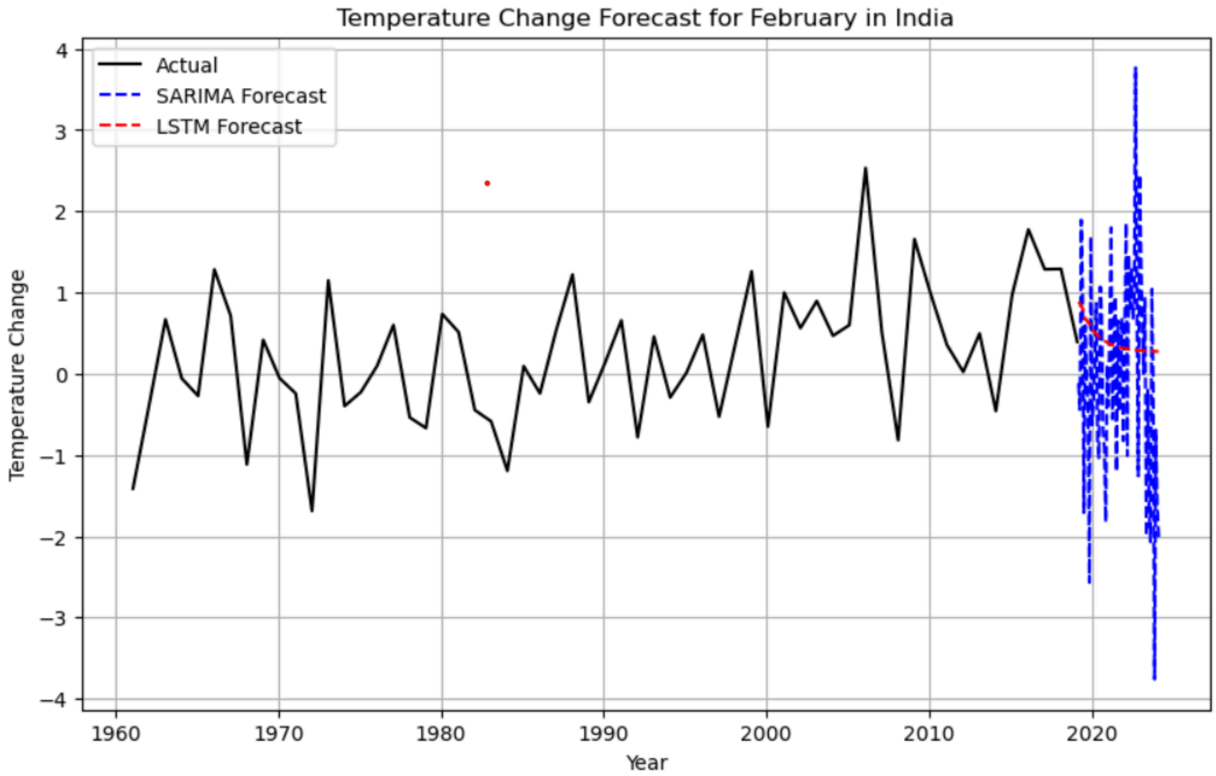
This makes LSTM particularly effective for accurate time series forecasting in domains like stock prices, weather, and demand.

```
model = Sequential()
model.add(LSTM(50, activation='relu', input_shape=(12, 1)))
model.add(Dense(1))
model.compile(optimizer='adam', loss='mse')
model.fit(X, y, epochs=100, verbose=0)
```

Now lets see the comparison between forecasting of SARIMAX and LSTM through plots below

SARIMAX Vs LSTM





Conclusion

In conclusion, both SARIMA and LSTM models offer powerful capabilities for time series forecasting. SARIMA is well-suited for capturing seasonal patterns and short to medium-term trends, while LSTM excels in modeling long-term dependencies in sequential data. By leveraging the strengths of both models, forecasters can benefit from enhanced prediction accuracy across various domains. However, selecting the appropriate model depends on the specific characteristics of the data and the forecasting horizon. Ultimately, the combination of SARIMA and LSTM represents a robust approach to time series forecasting, enabling more reliable predictions and informed decision-making.