Data Science Test 2 - Temperature Forecast

October 3, 2024

1 Introduction

This exercise focuses on setting up a time series forecast algorithm using a publicly available dataset. We have chosen the Earth Surface Temperatures dataset, which records temperatures from 243 countris from 1743 to 2013. Early data were collected by technicians using mercury thermometers, where any variation in the visit time impacted measurements. In the 1940s, the construction of airports caused many weather stations to be moved. In the 1980s, there was a move to electronic thermometers that are said to have a cooling bias.

We have explored two approaches; using a Seasonal Autoregressive integrated moving average (SARIMA) model and a Long Short Term Memory (LSTM) model. However and after some testing, we are only supporting the latter.

The report is structured as follows. Section 2 presents the dataset along with the trends and features identified during the Exploratory Data Analysis (EDA). Then Sect. 3 explains the methods implemented in this study, and finally Sect. 4 discusses the performance of the methods and analyzes the results.

2 Data

The dataset comprises monthly Earth surface temperature records for 243 countries from 1743 to 2013. Given the distinct temperature trends for each country, we narrow our focus to data from the United States.

Due to the large number of entries, visualizing the entire dataset to capture trends is challenging. Therefore, we first visualize data from the years 1900, 1950, 2000, and 2010, which were randomly selected to provide a representative view across different periods. The top panel of Fig. 1 illustrates the expected seasonal temperature pattern for the Northern Hemisphere: temperatures start low in January, gradually rise to a peak in summer, and then decrease again. This pattern is consistent across the four selected years and follows a cyclical, seasonal trend (bottom panel of Fig. 1).

The previous plots revealt a seasonal trends over a year. Figure 2 aims to reveal any underlying trend in average temperatures over the years. At first glance, we can observe an upward trajectory, which is further confirmed by the linear fit. This increase in temperature aligns with the effects of global warming.

Stationarity refers to the property of a time series where the statistical properties such as mean and variance. From the previous plots, we can claerly see that the mean temperature is not constant over time. As expected, it ciclely increases and decrases again in correspondance with the time/weather seasons. Furthermore, there is a warming trend over years due to global warming.

2.1 Data Preprocessing

For this exercise, minimal data preprocessing is required. As previously mentioned, we focus exclusively on temperature data from the United States. The dataset is split into two subsets: the training set includes records from 1743 to 2004, while the test set contains data from the remaining years, 2005 to

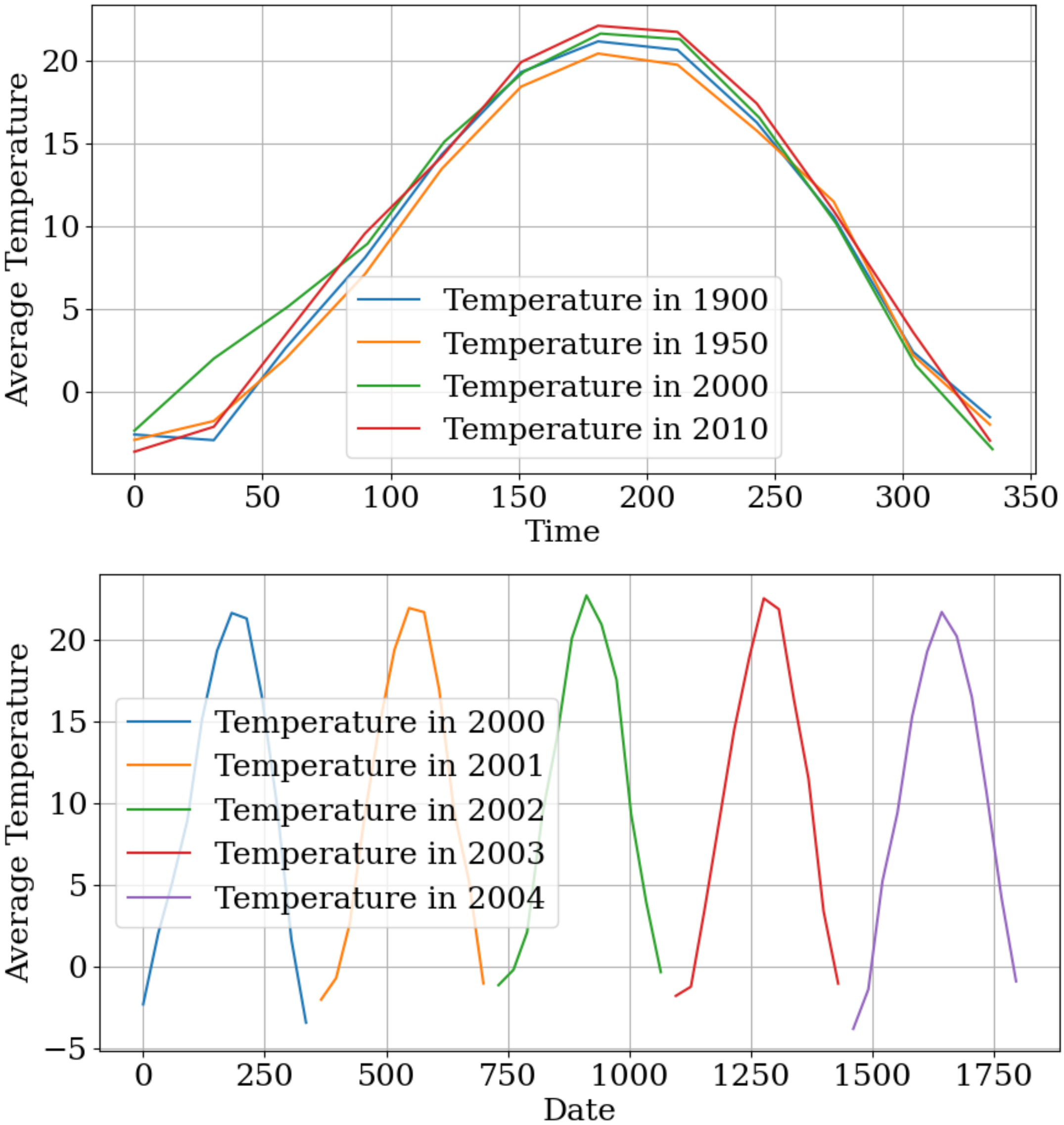

Figure 1:
caption*Top*: Seasonal trend over a year, for years 1900, 1950, 2000, and 2010. *Bottom*: Seasonal trends for five consecutive years from 2000 to 2005.

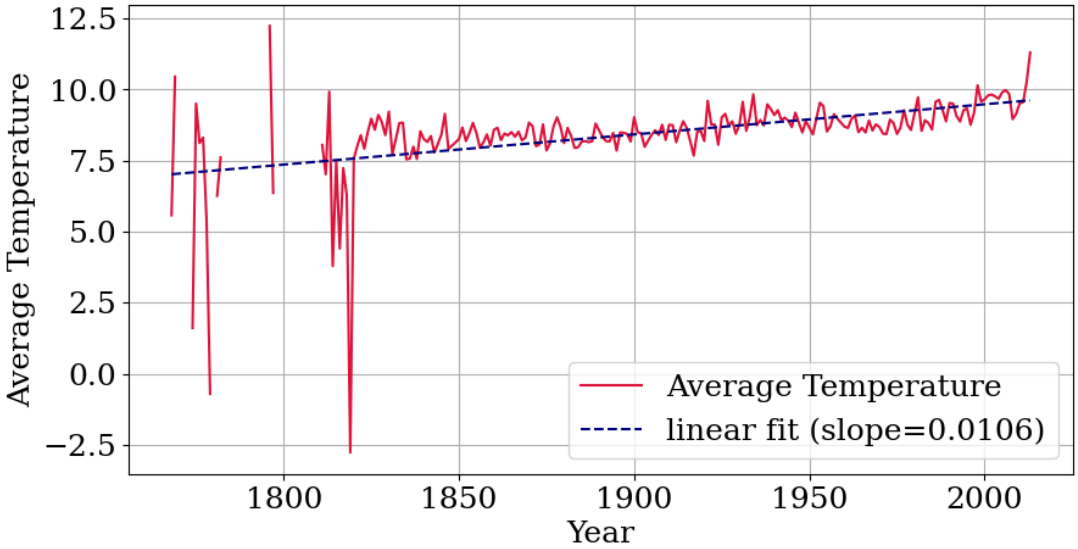

Figure 2: Long term temperature evolution. We average temperature over years. The plot depics a global increasing term potentially related with global warming.

2013. To ensure consistency, years with missing records, particularly in the earlier periods, are excluded from the training dataset.

3 Methodology

We initially tested two methods for time-series forecast: SARIMA and an LSTM network. We finally decided to go for the latter, since it is more flexible and is better supported for monitoring and deploying with MLFlow. Therefore, in this section we focus on the the LSTM implementation.

We began with a simple LSTM network comprising two LSTM layers. During the exploration phase, various architectures were tested and compared based on relative and mean absolute error in forecasting temperatures. This included experimenting with LSTM networks containing additional layers or more hidden units. The final model in production consists of two LSTM layers with 100 hidden units. Expanding the number of layers introduced complexity without yielding better forecasts. Increasing the number of hidden units improved accuracy but also extended training time. The choice of 100 hidden units represents a balance between performance and computational cost.

We also fine-tuned the time step for the LSTM network, which determines the time window the model uses to learn patterns and make forecasts. We found that excessively long time steps increased both forecast errors and computational costs, while very short time steps failed to capture the seasonal trends effectively. After testing various options, we settled on a time step of 48, corresponding to four years, which provided a good balance between accuracy and the ability to capture long-term seasonal patterns.

Adding bidirectionality to the LSTM network significantly improved the model's forecasting performance. This enhancement allows the network to learn from both past and future temperature sequences during training, leading to more accurate predictions. Additionally, we incorporated a 5% dropout between the LSTM layers to prevent overfitting. Higher dropout probabilities were tested, but they hindered the network's ability to train effectively, reducing overall performance.

In conclusion, our deployed LSTM model is bidirectional, comprising two LSTM layers with 100 hidden units in the fully connected layers and incorporating a 5% dropout probability. This architecture

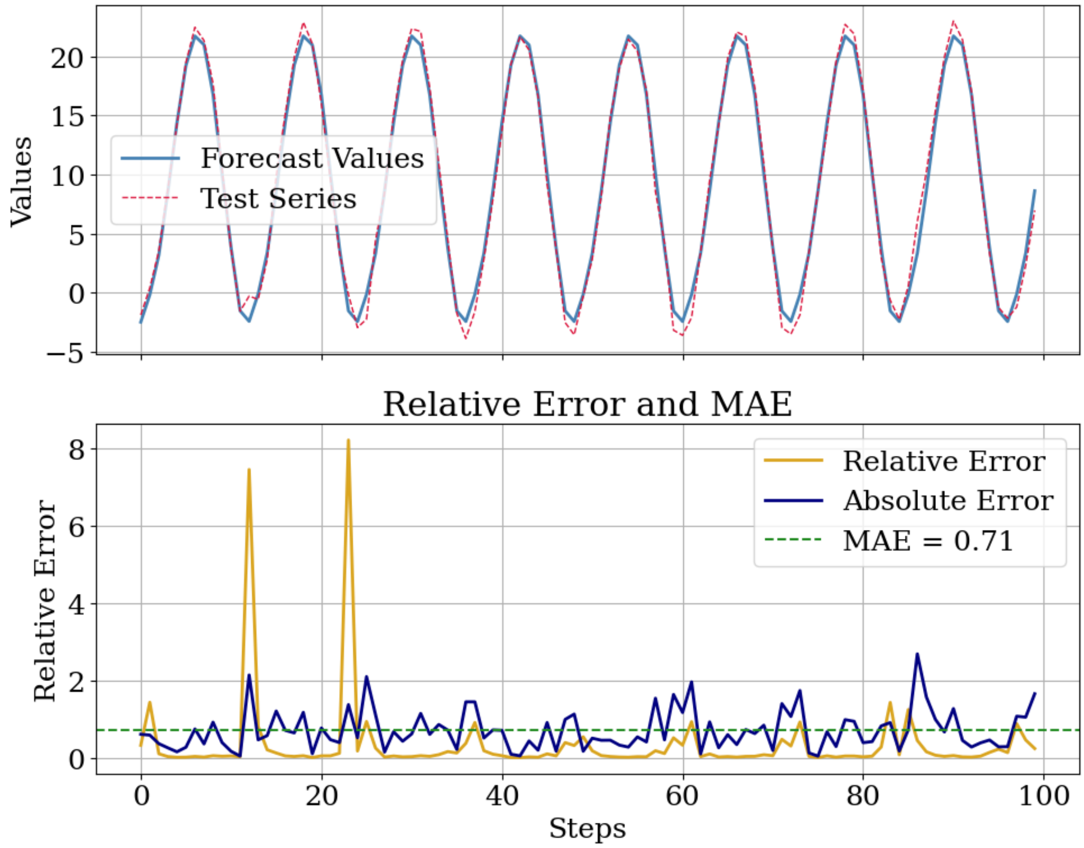

Figure 3: *Top*: Collected and forecasted tempeatures from 2005 to 2013. *Bottom*: Relative (yellow) and absolute error (blue) for the predictions in the top panel.

strikes a balance between forecast accuracy and computational efficiency.

The model is made accessible through a Docker container, which includes a notebook for generating temperature forecasts. In the following section, we will proceed with validating the model's predictions

4 Temperature forecasts

As introduced in Sect. 2, we use temperature data up to 2004 for training, with the period from 2005 to 2013 reserved for testing. Figure 3 compares the forecasted temperatures with the recorded values for this test period. To evaluate the model's performance, we examine both the relative error and the absolute error. The absolute error fluctuates around zero without significant deviations, indicating overall stable performance. However, we observe two noticeable peaks in the relative error. These peaks occur at temperatures near zero, where the relative error becomes less meaningful as an indicator of accuracy. The mean absolute error for the predictions is 0.71 and an $r^2 = 0.99$, suggesting that the model performs reasonably well overall.

Figure 4 provides additional insights into the LSTM forecast. The left panel displays a 1:1 scatter plot of the forecasted versus recorded temperatures, with the background density scatter representing the distribution of training values. It's worth noting that these values are not evenly distributed.

This uneven distribution pattern is likely influenced by the monthly averaging of the temperature data. Such aggregation smooths out less frequent intermediate temperatures, causing the more extreme

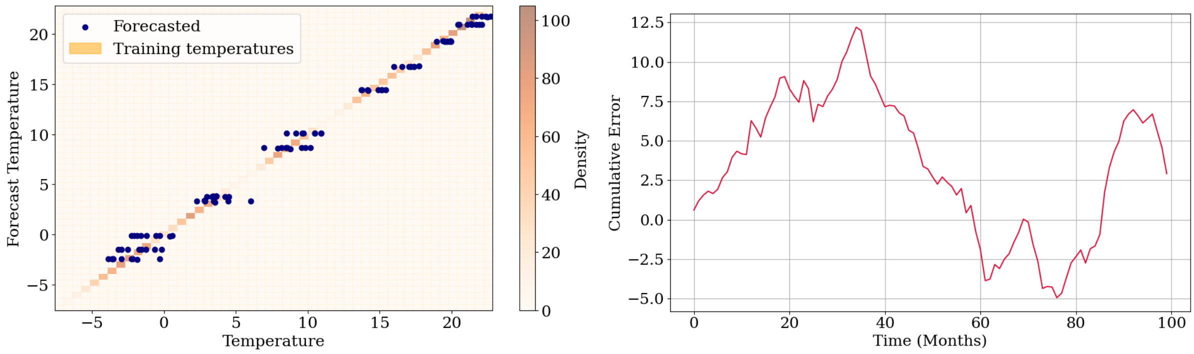

Figure 4: *Left*: Scatter plot of the forecasted vs the collected temperatures (blue). The orange density points indicate the training temperatures distribution, which is shown to be uneven. *Right*: Cumulative bias on the forecasted temperatures.

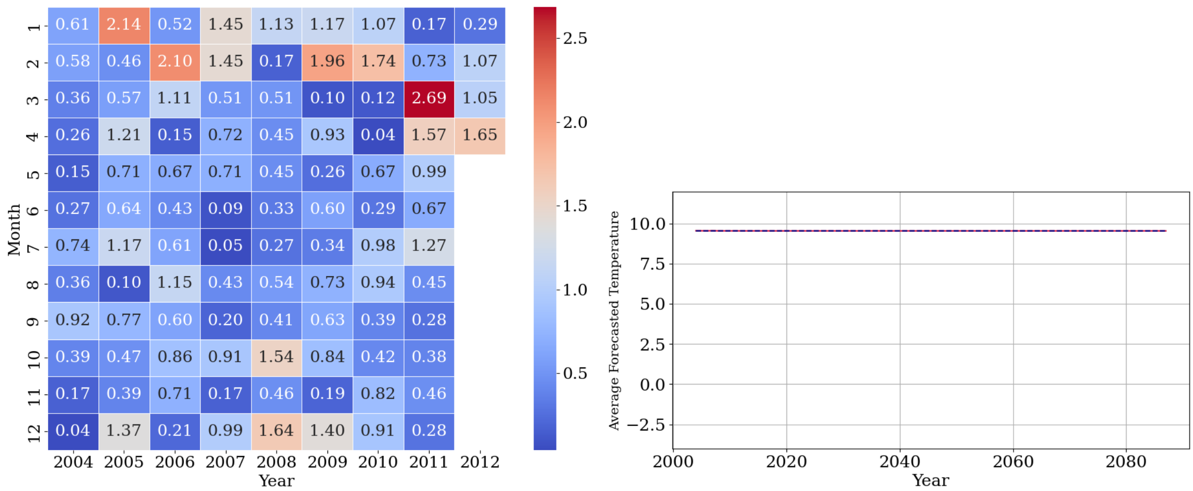

Figure 5: *Left*: Heatmap of the error on the forecasted temperatures in a monthly grid. The plot aims to search seasonal error correlations. *Right*: Long-term trend of forecasted temperatures.

or common temperatures to dominate, while smaller variations are diminished.

The scatter plot also shows that the forecasted temperatures are not smoothly distributed, but tend to cluster within specific ranges. This indicates that temperature variations are not gradual over time, with abrupt changes more prominent in the trends observed in the training set. These sharp transitions are especially clear in the density plot, where forecasted values cluster around densely populated regions. Additionally, the presence of stripes in areas of clustered predictions is a concerning anomaly that warrants further investigation.

The right panel on 4 shows the cumulative bias, where the error is estimated as

$$\mathrm{AE} = T_{\mathrm{True}} - T_{\mathrm{forecast}}\,. \tag{1}$$

The model tends to underestimate the forecasted temperature, although we do not see a clear concerning bias.

Another crucial evaluation involves examining whether patterns exist in forecast accuracy across different time periods, particularly with respect to seasonal biases. This entails investigating whether the model consistently overpredicts or underpredicts temperatures during specific months. The left panel of Fig. 5 shows the absolute error estimated as the absolute value of Eq.1 in a montly grid. From this visualization, we observe no significant correlation between errors and seasons.

During data exploration, we established that, in addition to seasonal trends, the data exhibits a subtle increasing trend potentially linked to global warming. Capturing this long-term trend is more challenging than identifying seasonal patterns. The long-term trend is not only more subtle but also requires recognizing patterns over significantly extended time windows.

In the left panel of Figure 5, we forecast the next 1,000 temperatures for which no recorded values exist, allowing us to compare the long-term trend with that of the training data (see Fig. 2). When we average the predictions over the years, the resulting mean appears flat and fails to capture the underlying global warming trend. Further investigation will be necessary to address this issue.

5 Discussion

In this study, we have developed and validated a bidirectional LSTM model for forecasting monthly Earth surface temperatures using a dataset spanning from 1743 to 2013 in the United States. Our model achieves a mean absolute error of 0.71 and an $r^2 = 0.99$ during the testing phase, indicating robust predictive performance. The visualizations of forecasted versus recorded temperatures reveal clustering of predicted values which necessitate further investigation. While the model effectively captures seasonal temperature patterns, it struggles to account for long-term trends, which also requires further analysis.