Temperature Prediction – Data Science Project

Anuj Tanwar

Abstract

In the project, I am using historical data of Global cities that contains average temperature for each day of the year for over a century. I will be using Time Series to plot the average temperature and perform a Linear Regression to predict future temperature.

Keywords: Temperature, Prediction, US historical data

Temperature Prediction – Data Science Project

Model will prompt user to enter the city and search historical data of the city for current date of the month of each historical year. Then, it will clean the dataset and plot various graphs to understand the dataset. I will calculate mean, median, mode, Standard Deviation and Population Variance. I will use linear regression to predict future temperature.

Business Problem

Weather is one that is not just close to us but is essential for our survival. Lot of businesses rely on weather, farmers rely on weather, bad weather can devastate the food on the fields. Sudden change in surface temperature can be harmful for our health as well. Research shows that abnormal weather disrupts the operating and financial performance of 70% of businesses worldwide. Every year, weather variability is estimated to cost $630 billion for the U.S. alone, or 3.5% of GDP. It becomes important to forecast weather in an accurate and timely fashion so that we can take the necessary precautions to minimize weather-associated risks. In the project I will be predicting surface temperature using Long-Short Term Network (LSTM)-based model on more than 100 years of surface temperature recorded data from Kaggle.

Background/History

Predicting weather goes as far back as 650 BC, the Babylonians tried to predict weather based on cloud pattern and astrology. By 350BC, the Aristotle was describing weather patterns in texts, while even Jesus Christ himself had a crack at forecasting in the New Testament. However, science of weather forecasting truly began in 19th Century. Times published first daily weather forecasts in 1861. Since then, technology to understand atmospheric physics has improved drastically. Forecasting techniques include analyzing data relating to pressure, air speed,

precipitation and temperature. These are collected from around the world and fed into supercomputers for analysis.

Data Explanation

Datasets

1. **Flat File/CSVs:**

   Below are the two CSV datasets I am going to use in my term project:

   GlobalLandTemperaturesByMajorCity.csv dataset is Global Land Temperatures By major city from 1743 to 2013 for each date of the year. Below are the fields:

   - **Dt:** Date
   - **AverageTemperature:** Average temperature on the date
   - **AverageTemperatureUncertainty:** Uncertainty on the mean temperature
   - **City:** City from which temperature is recorded
   - **Country:** Country of the City
   - **Latitude:** Latitude of the location where temperature was recorded
   - **Longitude:** Longitude of the location where temperature was recorded

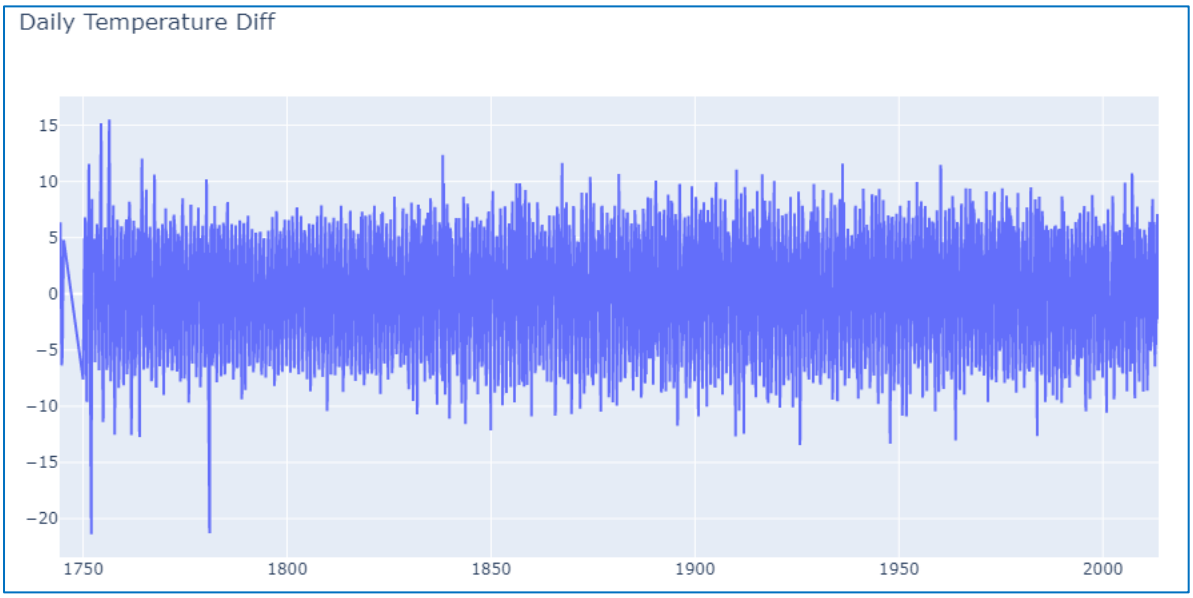
Data Preparation

Below are the steps I plan to execute to evaluate the results:

1. Read the dataset and filter data for the proposed city on which we need to perform analysis.
2. Clean the data by dropping the duplicates using drop_duplicates().
3. Find and filter out outliers using box plot, taking care of NaN.
4. Used box plot to determine temperature less than -12 are outliers.
5. Filtered out records with temperature less than -12.
6. Calculate previous month's temperature for every month by using shift(1).
7. Created an additional field to record the difference between consecutive month's temperatures using shift method.

	Effective_Date	Temperature	Prev_Temp	DIFF
1543353	1744-04-01	8.766	5.436	3.330
1543354	1744-05-01	11.605	8.766	2.839
1543355	1744-06-01	17.965	11.605	6.360
1543356	1744-07-01	21.680	17.965	3.715
1543358	1744-09-01	17.030	21.680	-4.650
1543359	1744-10-01	10.662	17.030	-6.368
1543360	1744-11-01	5.776	10.662	-4.886
1543361	1744-12-01	0.371	5.776	-5.405
1543362	1745-01-01	-0.901	0.371	-1.272
1543363	1745-02-01	-0.422	-0.901	0.479

8. Plot Monthly Temperature Difference.



9. Splitting test and train data by filter out the last year in original dataset (2013 for Chicago) for test and rest as training dataset.
10. Create a model on top of it.
11. Creating time series plot of the clean dataset.

    Finally predicting future values.

Summary of Methods

Project Execution

GlobalLandTemperaturesByMajorCity.csv will be the main driver dataset to find the historical average temperature of a given city for a date and mean of those temperature values will be used as a projected value for current year.

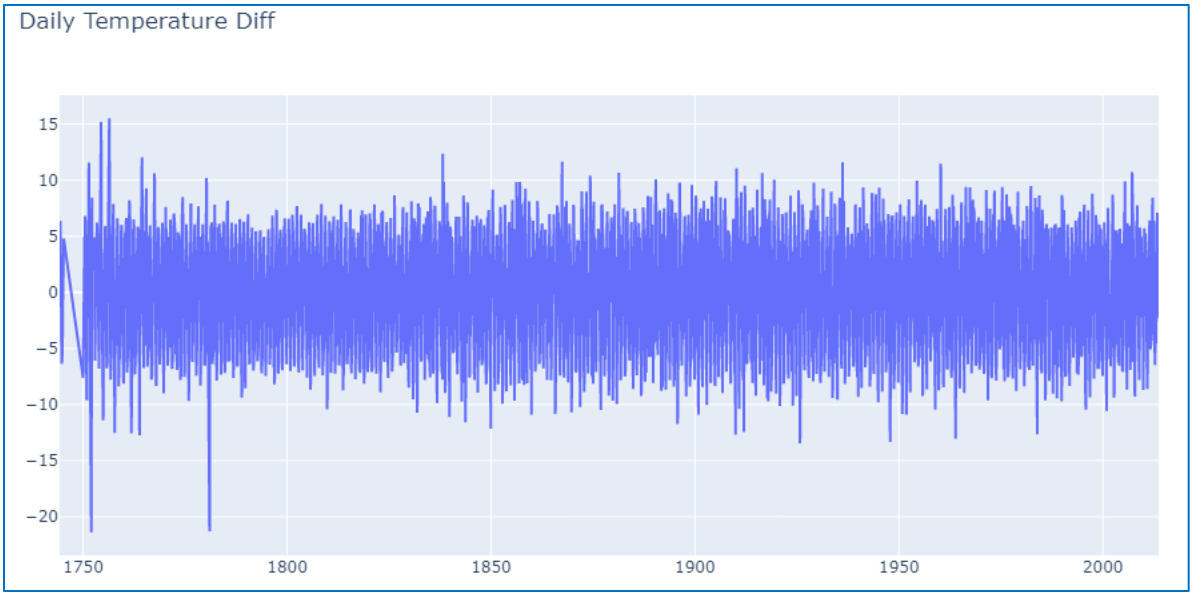
Data Preparation

Below are the steps I plan to execute to evaluate the results:

1. Read the dataset and filter data for the proposed city on which we need to perform analysis.
2. Clean the data by dropping the duplicates using drop_duplicates().
3. Find and filter out outliers using box plot, taking care of NaN.
4. Used box plot to determine temperature less than -12 are outliers.
5. Filtered out records with temperature less than -12.
6. Calculate previous month's temperature for every month by using shift(1).
7. Created an additional field to record the difference between consecutive month's temperatures using shift method.

	Effective_Date	Temperature	Prev_Temp	DIFF
1543353	1744-04-01	8.766	5.436	3.330
1543354	1744-05-01	11.605	8.766	2.839
1543355	1744-06-01	17.965	11.605	6.360
1543356	1744-07-01	21.680	17.965	3.715
1543358	1744-09-01	17.030	21.680	-4.650
1543359	1744-10-01	10.662	17.030	-6.368
1543360	1744-11-01	5.776	10.662	-4.886
1543361	1744-12-01	0.371	5.776	-5.405
1543362	1745-01-01	-0.901	0.371	-1.272
1543363	1745-02-01	-0.422	-0.901	0.479

8. Plot Monthly Temperature Difference.



9. Splitting test and train data by filter out the last year in original dataset (2013 for Chicago) for test and rest as training dataset.
10. Create a model on top of it.
11. Creating time series plot of the clean dataset.
12. Finally predicting future values.

Visualizations

1. Histogram

   Histogram can show count of days for each rounded value of temperature which can help understand the summarization of the distribution. Here is the histogram of average temperature:

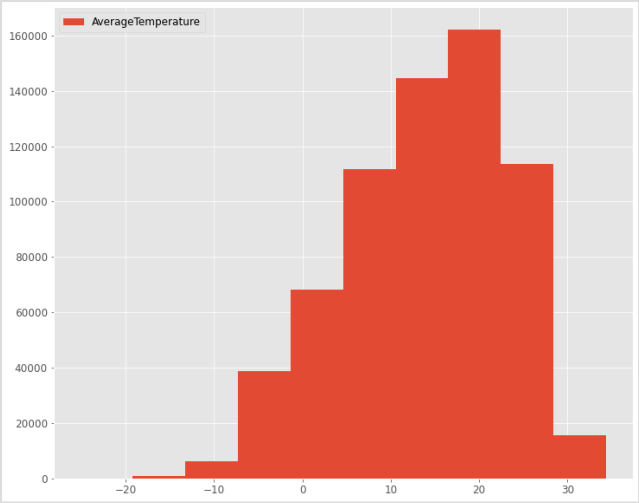


Figure 1: Histogram Average Temperature Counts

2. Box Plot

   This can help us in identifying and eliminating outliers

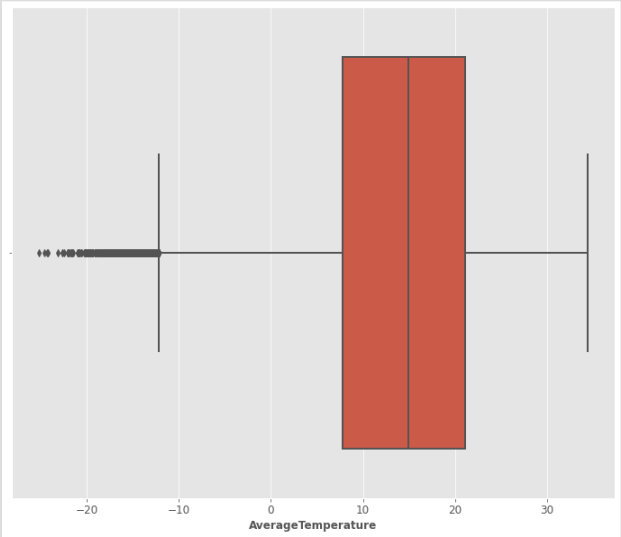


Figure 2: Boxplot showing outliers

3. Scatter Plot

   This can be used to find if 2 variables have a direct correlation. Here is a scatter plot of average temperature and average temperature uncertainty which shows there is no direct correlation between the two.

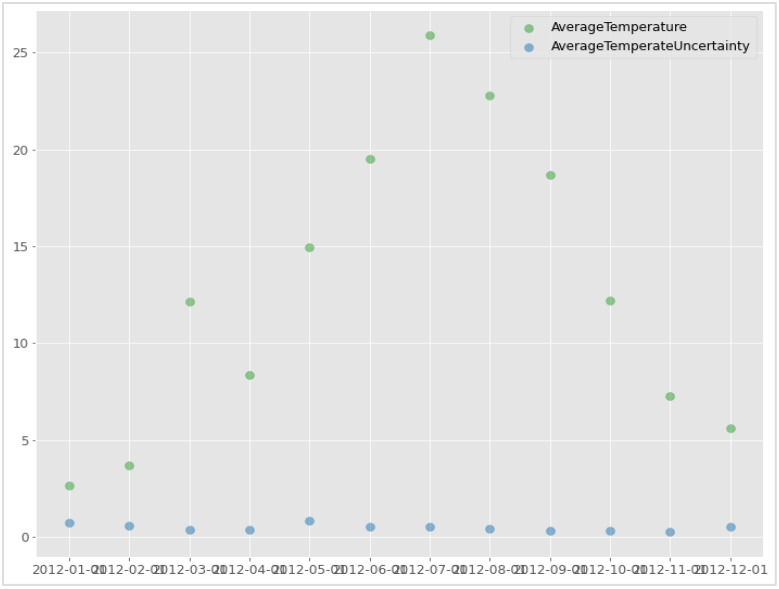


Figure 3: Scatter plot showing no correlation between avg. temp. and avg. temp. uncertainty

4. Pareto Distribution

It can demonstrate how temperature changes over different months for a city. Below plot shows that July to Sep months have good temperature in Chicago and rest of the months are colder.

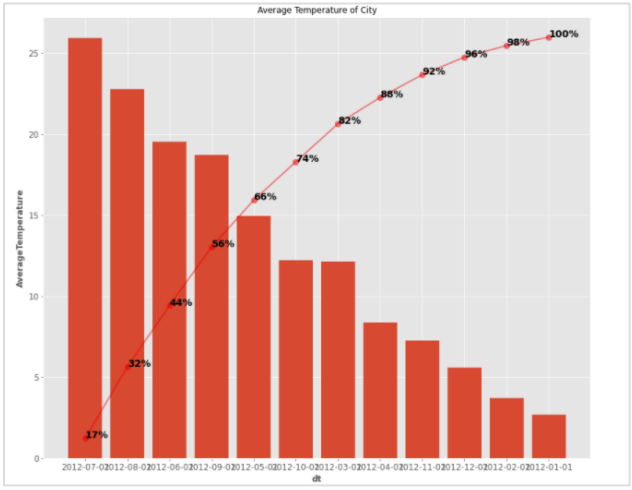


Figure 4: Pareto Distribution showing how temperature changes over time

There will be much more visualizations as we go deep in the project exploration such as actual values vs Predicted values for test dataset.

Analysis – Building and Evaluating Model

- Before building the predictive model, I prepped train and test datasets by using min max scaler that shrinks the data within the given range, usually of 0 to 1 and reshaping the datasets.
- Then I build a LSTM sequential predictive model with 4 neurons and 100 epochs. The mean squared error is being used as the loss function. Additionally, the adam optimizer is used, with training done over 100 epochs.
- Predictions were made on test data using model.predict function to check the accuracy of the data.

```
y_pred = model.predict(X_test,batch_size=1)
print(y_pred)
```

```
[[0.1596923]
 [0.1596923]
 [0.1596923]
 [0.1596923]
 [0.1596923]
 [0.1596923]
 [0.1596923]
 [0.1596923]
 [0.1596923]]
```

Figure 5: Predictions of test data

Conclusion

Result Interpretation

To interpret result of model prediction, I performed the following steps:

- Reshaped the prediction
- Used inverse_transform to scale by the predictions to normal temperature range
- Joined predictions with original test dataset to reflect the values with Effective Date.

	Pred_Temp	Effective_Date
0	1	2013-01-01
1	0	2013-02-01
2	1	2013-03-01
3	6	2013-04-01
4	13	2013-05-01
5	17	2013-06-01
6	21	2013-07-01
7	22	2013-08-01
8	19	2013-09-01

Figure 6: Predicted Values for test dataset.

- Predicted values and real values are then plotted to see how accurate the model is.

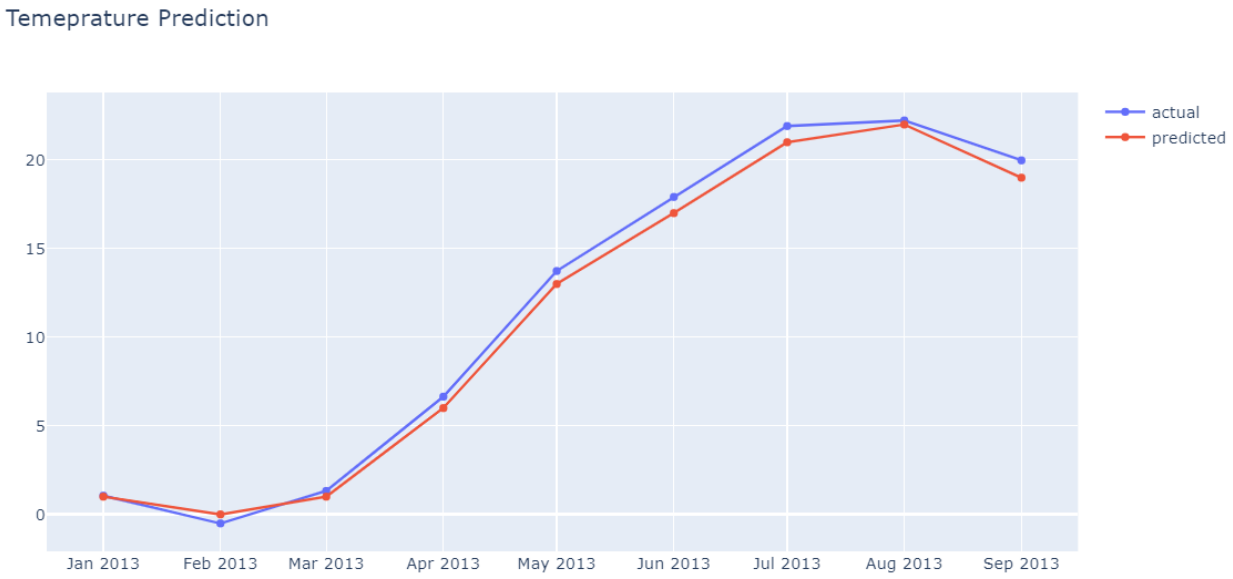


Figure 7: Predicted Values Vs Real Values Plot

- **RMSE**

  RMSE was calculated using mean_squared_error method of sklearn.metrics.

  RMSE came out to be pretty small i.e. 0.67 which means our prediction model is accurate. Same can be seen in the Sales Prediction plot above.

```
: from sklearn.metrics import mean_squared_error

rms = mean_squared_error(test['Temperature'], df_result['Pred_Temp'], squared=False)

print("RMSE : ",rms)
```

```
RMSE :  0.6663692669984118
```

Figure 8: RMSE on predicted values

Conclusion Summary

Looking at the RMSE and Predicted Vs Real Value plot we can state that model predictions are close to accurate. We can conclude that the prediction model can be used for future predictions.

Assumptions

Weather forecasting is a huge challenge, we are trying to predict something which is inherently unpredictable. Atmosphere is a chaotic system, a small change in its state at a location can have huge impact elsewhere, this is called Butterfly affect. A small error in prediction can rapidly grow and cause errors on a larger scale. And since many assumptions must be made when modelling the atmosphere, it becomes clear how easily forecast errors can develop. For a perfect forecast, we would need to remove every single error which is practically not possible. For the project, we are assuming there are no other factors that can influence the prediction and there could be some scope of error between prediction and actual values.

Limitations

Predictive model in the project is built based on the historical data and does not consider other factors such as irregular flows of air form clouds, power storms, and push around cold air that build on each other and form layers. A tiny disturbance in one layer, even one as tiny as a butterfly flapping its wings, can have a domino effect, affecting the other layers and snowballing

into radically different weather patterns. Due to all these variations and uncertainty, there is a limit in predicting the weather.

Challenges

Main challenge will be to clean the data and join different datasets. Datasets might contain null values and outliers which will need to be cleaned. Strategy to handle those null values and outliers need to be analyzed and implemented, which could potentially impact the final prediction. Getting weather prediction is always tricky and there are external factors that impact the changing weathers hence the model is never 100% accurate. Below are some of the other challenges faced in temperature prediction:

- Amount of available data
- Time available to analyze it
- Complexity of weather events

Future Uses/Additional Applications

Below are some of the usages of the Temperature prediction model:

- Can help apparel companies manufacture and supply clothes according to the surface temperature
- Can help retail companies to promote and supply beverages accordingly to the outside temperature
- Can help individual users know when to carry a jacket or wear shorts
- Energy savings to set the HVAC and water heater temperature accordingly to outside temperature. To reduce stress on AC there is a the 20-degree rule which states that you should always keep your AC unit at no more than 20 degrees lower

than the outside temperature. It means that, if the outdoor conditions are at 95 degrees, you should set your thermostat at no less than 75 degrees.

Recommendations

Though LSTM is a good model to see variations and predictions over time, more precision is needed to bring more accuracy in surface temperature predictions. Along with historical data, we can incorporate data collected by other means as well, such as satellites, Barometers, Radar systems, etc., on factors such as hurricanes, sun flares, clouds, atmospheric pressure, to be more precise.

Implementation Plan

Below are the steps I plan to execute to implement the plan and evaluate the results:

1. Read the dataset and filter data for the proposed city.
2. Clean the data by dropping the duplicates, find and filter out outliers using box plot, taking care of NaN
3. Splitting the data in train and test and create a model on top of it
4. Creating time series plot of the clean dataset
5. Finally predicting future values

Ethical Assessment

Weather prediction systems are uncertain and any extreme predictions can cause:

- chaos among public
- Companies can misuse the data to inflate prices of commodities

The ethical solution is to be transparent and “equitable” with regard to forecast interpretation and not get influenced by profit making organizations to create bias in the model.

References

https://www.kaggle.com/berkeleyearth/climate-change-earth-surface-temperature-data?select=GlobalLandTemperaturesByMajorCity.csv

https://www.kaggle.com/datasets/sobhanmoosavi/us-weather-events

https://www.theguardian.com/uk-news/2021/sep/29/from-babylon-to-google-a-history-of-weather-forecasting#:~:text=In%201861%2C%20the%20first%20daily,air%20speed%2C%20precipitation%20and%20temperature.

https://hbr.org/2017/09/severe-weather-threatens-businesses-its-time-to-measure-and-disclose-the-risks