Weather Prediction

Anuj Tanwar

Bellevue University

Abstract

Term project I have picked for DSC630 is weather prediction system. 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: Weather, Prediction, US historical data

Weather Prediction

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.

Datasets

1. Flat File/CSVs:

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

- **GlobalLandTemperaturesByMajorCity:**

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

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

WeatherEvents_Jan2016-Dec2020.csv dataset is a countrywide weather events dataset that includes 6.3 million events, and covers 49 states of the United States. Examples of weather events are rain, snow, storm, and freezing condition. Some of the events in this dataset are extreme events (e.g. storm) and some could be regarded as regular events (e.g. rain and snow). The data is from January 2016 to December 2020, using historical weather reports that were collected from 2,071 airport-based weather stations across the nation. Below are the fields in dataset:

- **EventId**: This is the identifier of a record.
- **Type**: The type of an event; examples are rain and snow.

- **Severity**: The severity of an event, wherever applicable. The severity of an event, wherever applicable.
- **StartTime(UTC)**: The start time of an event in UTC time zone.
- **EndTime(UTC)**: The end time of an event in UTC time zone.
- **TimeZone**: The US-based timezone based on the location of an event (eastern, central, mountain, and pacific).
- **AirportCode**: The airport station that a weather event is reported from.
- **LocationLat**: The latitude in GPS coordinate of airport-based weather station.
- **LocationLng**: The longitude in GPS coordinate of airport-based weather station.
- **City**: The city in address record of airport-based weather station.
- **County**: The county in address record of airport-based weather station.
- **State**: The state in address record of airport-based weather station.
- **ZipCode**: The zipcode in address record of airport-based weather station.

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.

Type of Model

I am planning to use time series plotting with linear regression for future predictions.

How do you plan to evaluate your results?

Below are the steps I plan to execute to 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

What do you hope to learn?

I am hoping to explore and learn about my dataset during the project execution.

Assess any risks with your proposal

Linear Regression itself cannot handle outliers. Outliers have be searched and filtered.

Contingency plan

If liner regression modelling does not work then I will try to build LSTM (Long-Short Term Network) model. LSTM is able to 'remember' long-term dependencies/information from a feature. Information runs across the networks and passes down such that earlier information from a sequence could influence prediction of later input. This passing of information is crucial for time-series weather data where each input is dependent on other input from varying time points.

Potential Challenges

Main anticipated 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.

Will I be able to answer the questions I want to answer with the data I have?

Predicting weather is always tricky but I believe I will be able to solve the questions and be able to predict temperature in reasonable range.

What visualizations are especially useful for explaining my data?

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:

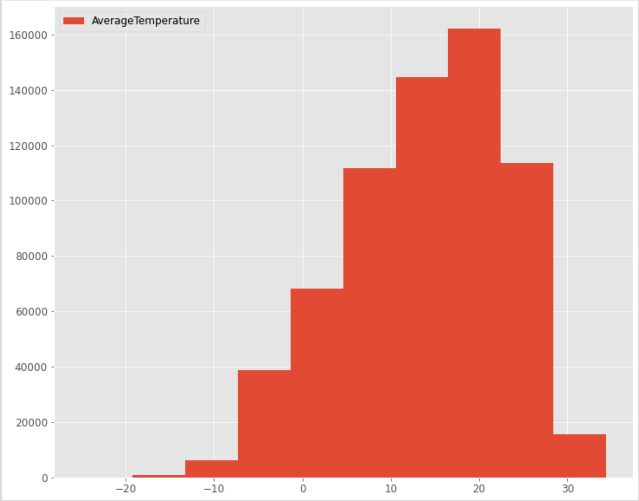

Figure 1: Histogram Average Temperature Counts

2. Box Plot

   This can help us in identifying and eliminating outliers

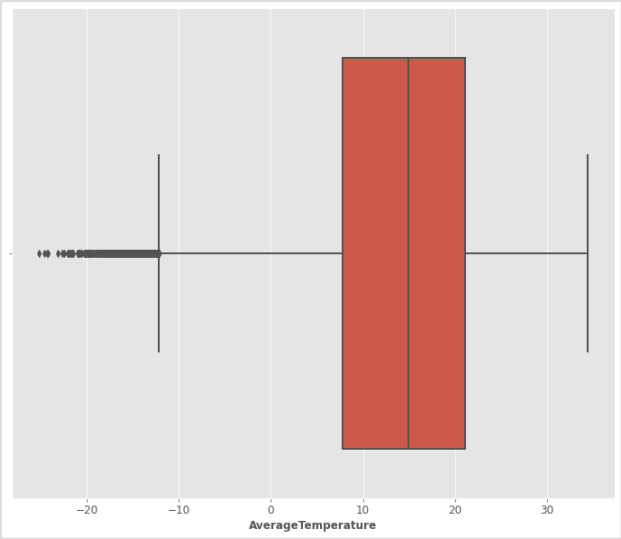

Figure 2: Boxplot showing outliers

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.

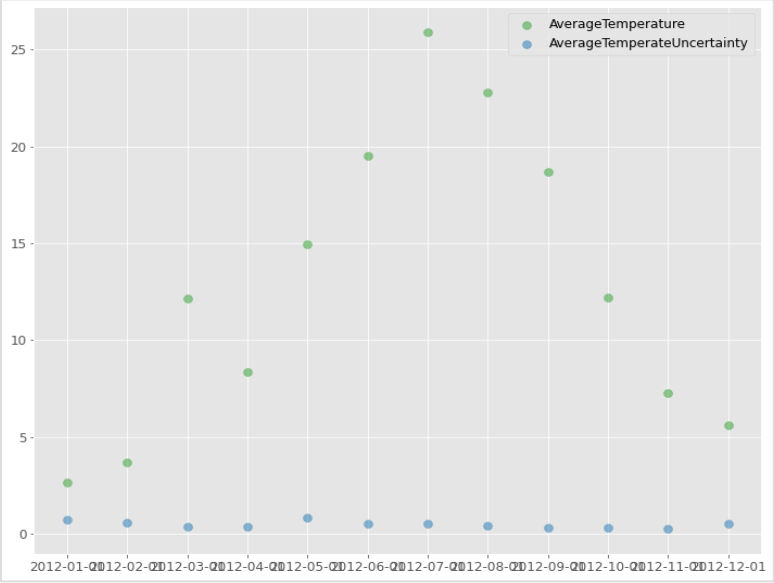

Figure 3: Scatter plot showing no correlation between avg. temp. and avg. temp. uncertainty

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.

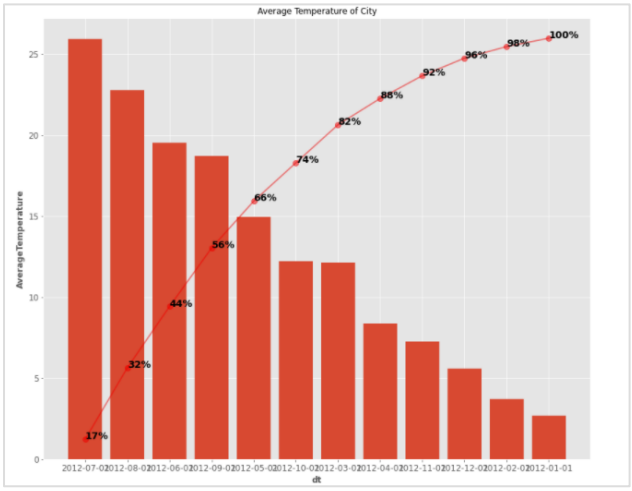

Figure 4: Pareto Distribution showing how temperature changes over time

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

Do I need to adjust the data and/or driving questions?

I will absolutely need to adjust and clean the data to make it work. Some of the steps are:

1. Filter on a particular city on which we need to perform analysis.
2. Remove Duplicates.
3. Taking care of outliers.
4. Convert date field into ordinary format as linear regression does not work on dates.

Do I need to adjust my model/evaluation choices?

Depending on the accuracy of prediction results I might have to adjust my model or might have to chose a different type of modeling technique. This question can be answered as I will make more progress in model evaluation.

Are my original expectations still reasonable?

Yes, I believe my original expectations are still reasonable. I am using more than 100 years of historical data to predict future data. I believe we have sufficient amount of data to make predictions for near future.

Finalizing results – Prepping the data (Milestone 4)

1. Read the dataset.
2. Rename fields to relevant names.
3. Filter data for the proposed city.
4. Use fuzzy matching to set City parameter and validating user input.
5. Clean the data by dropping the duplicates.
6. Find and filter out outliers using box plot, taking care of Nulls.
7. Join both the datasets.
8. Create visualizations to get insight of the data.
9. Created an additional field to record the difference between consecutive month's temperatures using shift method.

10. Splitting test and train data by filter out the last year in original dataset (2013 for Chicago) for test and rest as training dataset.

Finalizing results – Building and Evaluating Model (Milestone 4)

- 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

Finalizing results – Result Interpretation (Milestone 4)

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.

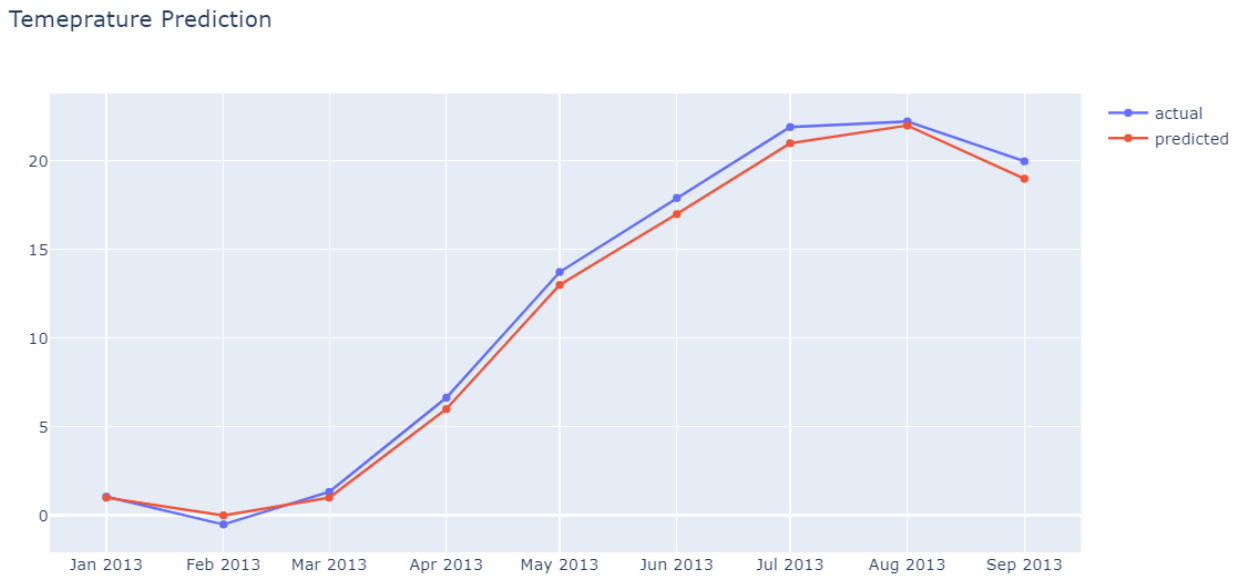

Figure 7: Predicted Values Vs Real Values Plot

- **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

Finalizing results – Formulating Conclusion (Milestone 4)

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.

References

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

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