Weather forecasting with Deep Learning

Team Members - Anagha Mayasandra Vinaya Simha and Naveen Asokan

Executive Summary

Decisions to be impacted : In the context of weather forecasting, machine learning plays a very important role in helping us predict f uture weather conditions and also the accuracy of our predictions. These predictions play a crucial role in many real-world applications. Some of them include:

Agriculture:

Crop planting: Decisions on when to plant a crop is highly dependent on the weather. If the region is expecting famine during crop planting, then the farmer might undergo huge losses.

Irrigation: Farmers can decide on how much water would be needed based on future weather data.

Pest Control: Many pests thrive on warmer climates where there is high CO2 levels in the soil. Controlling when to use pesticides and how much pesticides should be used is critical to crop survival.

Transportation:

Flight and Shipping schedules: it is important for flight and shipping companies to know the weather conditions so that they can reroute in case of bad weather conditions.

Public transport Operations: In case of severe weather, the public might need more access to public transport and knowing weather conditions beforehand can help them be prepared.

Emergency and Disaster Management:

Resource allocation: The government can provide the necessary resources to the public when they have managed procuring large quantities of food, water etc.,

Evacuations: The government and the public can know about the evacuation plans and can better prepare their home and shelter to bear the brunt of the disaster.

Business values:

The aim of this project is to help the society to be better prepared in terms of future weather. As seen in the past, hurricanes and snowfall can cause a lot of property damage and have a wrongful

impact on the everyday activities of a human being. This project aims to minimize these situations and improve the current weather forecasting applications.

Data Assets:

While we did consider utilising the NOAA data as our original dataset, the hurricane Helene caused major outages due to which we had to change our data to Open Meteo historical data API. O ur dataset now containes 32 features and 1086120 samples.

Predictive analytics: the data from open Meteo contains past data(from the year 2000 to Oct 11 2024) which forms the core of our features which are required for modelling. An irrigation scheduling machine can predict rain in the next few days, asking a farmer to use lesser water than usual.

Descriptive Analytics: Since we have used past data, we can easily catch the trends and seasonality in our models. For example, the transportation department can assess the past weather data patterns to determine when bad weather conditions can be expected, so that rerouting could be done in advance.

Prescriptive Analytics: The government can use prescriptive analytics to determine how many bags of grains will be required during a disaster by analyzing previous trends.

Questions

1) Los Angeles seems to be very different from all the other cities by weather codes. How did you handle this?

Answer:

In our LSTM model we will be having an feature embedding layer which could take in encoded city names. This would make sure that we handle Los angeles different than to St louis. We chose all the cities based on the fact that they have very different weather conditions. We wanted to consider these cities because they represent the variation in temperature across different regions in the US. We dont need to do any seperate preprocessing for this. In additinal we have some features based on geography and topography, like wind and soil temperature in various elevations.

2) Would you isolate data into cities that have snowfall occur throughout the year? There might be some insights towards predicting inclement weather that is not present in the other cities.

Answer:

As far as our data processing goes, we do not have any cities where snowfall occurs throughout the year. The cities that we have considered are all different in terms of their weather. Our aim is to use a general model that can predict the weather patterns of all the 5 cities and we are still in the process of seeing how this goes. But in future if we are considering cities like with heavy rainfall and snowfall, then yes that should be modelled serperately. Example Columbia in South america, it rains almost everyday and weather patter in alaska are significantly different.

3) Can your model account for long-term changes in weather patterns due to climate change?

Answer:

We will be using LSTM and 1D-CNN for weather forecasting. Either of the models cannot predict long-term changes in data. LSTMs and CNNs are effective for short- to medium-term weather prediction as they capture temporal and spatial patterns present in historical data. For this project, our aim is to predict weather for 1-2 days ahead. But if we want model to account for loing term changes like an year ahead , then we need to consider climate simulation models.

4) What kind of post-hoc analysis did you perform to understand model behavior after training?

Answer:

We're still in the process of training. But we are considering doing these

- Residual analysis
- Temporal analysis
- Feaature importance
- Uncertainty Quantification

5) What method do you use for outlier detection?

Answer:

We have used 4 methods for outlier detection: IQR, STL Decomposistion, MCD and Mahalanobis distance. A combination of STL and Mahalanobis distance gives us the outliers present in our data.

6) How are you accounting for the microclimate effect in Los Angeles? That is an extremely sprawling area, and being situated in between mountains and the ocean causes a great deal of variability in weather conditions across the city.

Answer:

For our weather prediction model, we treat Los Angeles as a single entity rather than dividing it into sub-regions, focusing on capturing general patterns across the city.This simplifies the model design and is suitable for capturing general weather patterns across the city.

To account for the variability across cities, including Los Angeles and its neighbors, we use a single model that incorporates a dedicated embedding layer to encode city names, enabling the model to distinguish city-specific weather patterns.

However in phase 2 of this project that we are currently working, For regions with significant microclimatic variability, such as Los Angeles and its neighboring cities, we build a single city model. We have taken new york in our case. It could be done for cities with microclimatic varaition. We enhance the model with topographical features like elevation, proximity to the coast, and geographical coordinates (latitude and longitude) to better represent spatial dependencies. As with the combined model, we use an embedding layer to encode city names. This allows the model to understand and account for city-specific differences within the combined dataset.

7) Can you please indicate a little bit about what feature selection or feature engineering technic you used in the data preprocessing?

Answer:

We have used correlation matrix to see which of the features are highly correlated and which of them do not have a lot of correlation. Features which had high correlation were dropped. We have also performed log transformation on the vpd feature values to check for skewness. We also have a scatterplot of the different wind speeds and soil temperatures. Since they exhibited high correlation, one of the values were dropped.

8) Are you intending to train one model with data from all these cities or train 5 models for 5 cities? Given the micro climates in all of these cities, is that going to be accounted for in model development?

Answer:

We are going to train on both of the above methods. But we are more leaning towards the a single model for all 5 cities. For that we will be having an additional layer both in LSTM and CNN, feature

embedding layer where the cities would be encoded and fed as an additional input

The micro climates will be accounted on using a single model for a weather station.

9) For weather data, sometimes it can be extreme while also sometimes majority of data are 0. IQR might not work well for some features.

Answer:

Yes, this is the reason we have not considered IQR as the best outlier detection method for our dataset. Also when we did IQR , we didnt did for fetaures like precipitation since we almost had 0's in it.

10) How's you plan about feature engineering for your DL model based on your descriptive analysis?

Answer:

We will be using STL decomposition along with SARIMA model to make future predictions on our data. As shown in the presentation, our model has shown seasonality and we have also established stationarity with respect to trend. I believe these factors can help model our descriptive analytics.

11) In the box plots, rain for example, I noticed that the range for data points in the box has a very tight range and many sparse points are shown as outliers in boxplots. Is it indicating that 'rain' has a very fixed value?

Answer:

The reason for this is because, all of the cities that we have considered do not have daily rainfall. On the few days that rainfall occurs, it is not a very significant amount. Since we have considered hourly data, most of our values for rain is 0. This would be the common trend right if you consider countries like USA. However if you take countries like Columbia (Most rainfall), you would be able to see different levels of precipitation there.

12) What are you going to do with those high correlated variables?

Answer:

We will be dropping one of the highly correlated variables. Having the correlated features can Make coefficient estimates unstable and highly sensitive to small changes in the data. In our case for

wind_speed_10m and 100m were correlated. Similarly soil_temp_depth10m and 100m were correlated. So we opt to remove one of them which ahs the least corrleation with the target temperature.

13) How do you decide which feature to keep when multiple features have high correlation? What criteria do you use to select the most representative feature?

Answer:

As seen from the covariance matrix (Heatmap), only few of our data points had high correlation. For example, wind_speed_10m and wind_speed_100m were two different variables but they had high correlation so we decided to drop one of those values. Whenever we observe high correlation, we can just drop either of the features. High correlation implies that both the features contribute equally to the prediction.

14) Do you plan to look into whether the weather is easier to predict for some of your locations rather than others? As you might know, people like to joke that the weather in St. Louis is particularly difficult to predict, so I was wondering if there is any evidence for that.

Answer:

Like all weather prediction models, our model can appproximate the weather conditions for 1-2 days ahead but an absolute prediction might be hard to make. With respect to St Louis, we have observed some patterns and seasonality in our data... although the winters in St louis are quite out of box. It sometimes even drops beyond chicago's weather. Chcicago which is very close to St louis, however has very cold climate. It might be due to the lake or mountain that passes through. So the best model to fit all this would be a unique model for some cities (like stl and chicago). We are into building a single model and also unique model for each cities

Some locations like LA have almost the same weather conditions which might make it easier to predict, but there are other factors which might not allow for perfect predictions.

15) why did you choose these area as samples?

Answer:

These cities were chosen due to their high variation in different weather patterns. For example, St Louis experiences extreme weather conditions whereas Houston remains warm for most parts of the year. Likewise, LA is known for its pleasant weather and Chicago, New York are known for their

harsh winters. All these weather patterns are different from each other which makes our problem more challenging.

16) Do you think there is predictable features across countries? Like could you predict something about the weather in St Louis by looking at the weather data in Chicago? It would be interesting to see if there is, and if so, how it changes with the distance between cities, etc.

Answer:

The differences in local geography (e.g., mountains, coastlines) and microclimates starts playing a larger role. In winter ST louis and chicago are almost similarly cold (They may both be influenced by a cold front moving through the Midwest), however in summer St louis stays very hot. But cites which are very closer to each other might experience similar wetaher pattern. Hoever for this project, we took a different cities which are widely apart. So we couldnt find any correlation between the cities.

In the second phase of the project, we are leveraging the neighboring city model to predict the weather patterns of closest cities. There we could plot correlation coefficients between cities as a function of distance to determine how predictability changes geographically.

17) What are the key weather variables can be analyzed deeper and what weather elements make you pick those cities.

Answer:

We believe that temperature, rain, snowfall are some of the variables that allow for deeper analysis due to their prevalence in everyday life and the inconvenience that they can bring to everyday life makes them important. We picked the above cities due to the different weather patterns each of them follow.

References

- Hochreiter, S., & Schmidhuber, J. (1997). "Long Short-Term Memory." Neural Computation, 9(8), 1735-1780. This paper introduces LSTMs, explaining how they effectively capture temporal dependencies, making them suitable for time-series weather prediction tasks.
- Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep Learning. MIT Press. Provides a foundational understanding of neural networks, including CNNs and their application in spatial

data analysis.

- Miller, A., Rohling, E. J., & Schmittner, A. (2012). "Temperature and Salinity Variability in the Ocean: Microclimates and Regional Weather Patterns." Nature Climate Change, 2(8), 588-592. Discusses the impact of geographical and topographical features on microclimates and regional weather variations.
- Benedetti, A., Morcrette, J.-J., Boucher, O., et al. (2009). "Aerosol Forecasting with the ECMWF Integrated Forecast System." Atmospheric Chemistry and Physics, 9(3), 743-758. Details how incorporating spatial and regional dependencies improves weather forecasting accuracy, supporting your inclusion of geographic and topographic features.
- Rasp, S., Dueben, P. D., Scher, S., et al. (2020). "WeatherBench: A Benchmark Dataset for Data-Driven Weather Forecasting." Journal of Advances in Modeling Earth Systems, 12(11). Explains machine learning-based weather prediction and the importance of embedding spatial data for effective regional modeling.
- Shumway, R. H., & Stoffer, D. S. (2017). Time Series Analysis and Its Applications. Springer. A detailed resource on time-series data modeling, including methods for spatial and temporal analysis relevant to microclimates.