Solution Architecture

Algorithms and Techniques

For both questions, Prediction of heat waves and AQI, we plan to implement the following two techniques.

1. A RNN with LSTM and GRU based on the work done by Anderson et.al [1]

Motivation:

In the referenced paper, the authors have combined LSTM and GRU, the two most advanced time series predictors, and have obtained better results than any of the techniques applied individually. LSTM and GRU are state-of-the-art models to predict time series data such as ours. We aim to combine the prowess of both LSTM and GRU in this model to predict Time series data.

Fig. 3. Proposed Model Architecture

LSTM:

LSTM (Long Short-Term Memory) is a type of Recurrent Neural Network (RNN) that is widely used for processing sequential data.

Motivation for usage of LSTM:

1. Handling long-term dependencies: LSTMs are designed to capture long-term dependencies in sequential data, making them well-suited for time series data where patterns and trends can persist over long periods of time.

2. Handling vanishing gradients: LSTMs are equipped to handle the vanishing gradient problem, which is a common issue in traditional RNNs when processing long sequences. This makes LSTMs a good fit for time series data, where patterns and trends can persist over long periods of time.

3. Flexibility: LSTMs have a more complex structure than GRUs, which allows them to handle a wider range of complex relationships in the data. This makes LSTMs well-suited for time series data where the relationships between variables and patterns in the data can be complex.

4. Improved performance: LSTMs have been shown to outperform other types of RNNs, including GRUs, on a wide range of time series prediction tasks, making them a popular choice for many practitioners.

5. Widely used: LSTMs have been widely used and studied for over two decades, and a large body of research and best practices has been developed around their use, making them a well-understood and established tool for processing time series data

GRU:

Gated recurrent units (GRUs) are a gating mechanism in recurrent neural networks introduced in 2014 by Kyunghyun Cho et al.

Motivation for usage of GRUs:

1. Handling long-term dependencies: Just like LSTM, GRUs are designed to capture long-term dependencies in sequential data, making them a good fit for time series data, where patterns and trends can persist over long periods of time.

2. Efficient computation: GRUs are computationally more efficient than LSTMs, making them well-suited for processing large time series datasets, which can be computationally intensive.

3. Easy to train: GRUs have fewer parameters than LSTMs, making them easier to train and less prone to overfitting.

4. Robust performance: GRUs have been shown to perform well on a wide range of time series prediction tasks, including forecasting stock prices, sales, and energy consumption, making them a popular choice for many practitioners.

5. Interpretability: GRUs have a simpler structure than LSTMs, making them easier to interpret and understand, which can be important when working with time series data, where the relationships between variables and patterns in the data can be complex.

The model will be built using TensorFlow in Python, and hyperparameters such as the number of neurons and dense layers will be modified if necessary.

Anderssen et. al. produced exemplary results using this technique in their paper

2. A neural network based on Neural Prophet [2]

Motivation:

NeuralProphet is a successor to Facebook Prophet, which set an industry standard for explainable, scalable, and user-friendly forecasting frameworks. With the proliferation of time series data, explainable forecasting remains a challenging task for business and operational decision making. Hybrid solutions are needed to bridge the gap between interpretable classical methods and scalable deep learning models.

NeuralProphet is a hybrid forecasting framework based on PyTorch and trained with standard deep learning methods. Local context is introduced with auto-regression and covariate modules, which can be configured as classical linear regression or as Neural Networks. Otherwise, NeuralProphet retains the design philosophy of Prophet and provides the same basic model components.

Our results demonstrate that NeuralProphet produces interpretable forecast components of equivalent or superior quality to Prophet on a set of generated time series. NeuralProphet outperforms Prophet on diverse collection of real-world datasets. For short to medium-term forecasts, NeuralProphet improves forecast accuracy by 55 to 92 percent.

Heat Wave Forecasting:

Feature Extraction

As discussed earlier in the introduction,

Studying literature on the causes for Heat Waves brings to the forefront a number of factors that cause Heat Waves. Some of these factors are well known local factors such as humidity, precipitation, and windspeed, other lesser-known local factors such as surface soil moisture and a combination of cloud cover and daylight hours, as demonstrated by Zeppetello et.al.[3], show promising results by using these factors as part of larger statistical models such as the proposed SEMB model (Soil Energy and Moisture Budget). We plan to use these factors as input to powerful neural networks.

As we know, The Earth is a continuous system and phenomena in one part of the earth especially oceans have a significant impact on the weather of a station. Thus, factors such as ONI (Oceanographic Nino Index), as demonstrated by Rohini P. et. al. [4] play an important role in determining the climactic conditions of a station in any given year. MJO (Madden-Julian Oscillation) amplitude and as evidenced in the work of Guigma K.H. et. al. [5] and SST (Surface Sea Temperatures) from the world's major oceans (Indian, Pacific, Atlantic) at various pressure levels will also be considered.

We plan to collect daily data for the below-mentioned features using a number of techniques including but not limited to Data scraping using crawlers, polling public APIs and Reverse engineering PUBLIC government APIs.

<table>
<tr><th>FEATURES</th><th>DATASET</th></tr>
<tr><td>Cumulative Rainfall</td><td rowspan="6">Telangana Daily Weather Data 2018 - 2022([Telangana Weather Data 2022 | Telangana Open Data Portal](https://data.telangana.gov.in/dataset/telangana-weather-data-2022))</td></tr>
<tr><td>Temp Max</td></tr>
<tr><td>Humidity Max</td></tr>
<tr><td>Humidity Min</td></tr>
<tr><td>Wind Speed Max</td></tr>
<tr><td>Wind Speed Min</td></tr>
<tr><td>Cloud Cover</td><td rowspan="2">National Information System for Climate and Environmental Studies (Satellite Data from various satellites such as CARTOSAT and INSAT) accessed through the Bhuvan Portal)https://bhuvan-app3.nrsc.gov.in/data/download/index.php</td></tr>
<tr><td>Surface soil moisture</td></tr>
<tr><td>Daylight Hours (Number of hours between sunrise and sunset)</td><td>Using python library suntime</td></tr>
<tr><td>ONI (Oceanographic Nino Index)</td><td>NOAA Oceanic Nino Index (ONI) [Climate Prediction Center - ONI (noaa.gov)]</td></tr>
<tr><td>MJO amplitude (RMM1 + RMM2)</td><td>NOAA MJO Index Data 1971-2022 [NOAA][https://www.psl.noaa.gov/mjo/mjoindex/]</td></tr>
<tr><td>SST Tropical Pacific at various pressure levels</td><td rowspan="3">[International Comprehensive Ocean-Atmosphere Data Set (ICOADS)][(https://icoads.noaa.gov/products.html)]</td></tr>
<tr><td>SST Tropical Atlantic at various pressure levels</td></tr>
<tr><td>SST Indian Ocean at various pressure levels</td></tr>
</table>

After the data is collected, it will be cleaned. We have no considered any Categorical Data hence encoding is not imperative, however each datapoint will be scaled using Standard Scaling, using sklearn's StandardScaler

Usage

We plan to implement two models for each of the 5 districts, using both the previously mentioned techniques.

The results will be compared and the best model will be chosen for each city. The comparison metrics will be the standard comparison metrics of MAE, MAPE and R^2 score. The first four years' worth of data will be used as a training data and the last year (i.e. 2022) will be used as testing data.

The models' prediction will then be stored as either a CSV or Parquet file for further use.

AQI:

Different models for time series data for each of the pollutants For time frame 2018 - 2022

- PM 2.5
- PM 10
- NO2
- SO2
- CO
- O3

Then, as per India AQI rules, the highest scaled pollutant will determine the overall AQI.

Features Extracted:

From literature review of the problem of Air Quality prediction, we decided on the factors for which data should be collected. Again, some were prominent and obvious factors affecting the air quality such as precipitation, number of vehicles and forest cover. This is also backed by relevant studies such as work done by Dongsheng Zhen et.al. [7]. However, there were surprising results in the literature such as the strong positive correlation of Per Capita Income [6] with AQI, i.e., the more the average income of a place, worse is the air quality. According to the work done in the same publication, Construction has been determined as a leading cause for worsening AQI. Thus, SGDP attributed to construction is also chosen as a feature [6]. Another such revelation was that Urbanization Level is not a significant contributor to the Air Quality of a place. Daily temperature is not considered as it is not a causal factor, rather it is an effect of the Air Quality.

FEATURE	DATASET
Monthly AQI	Monthly AQI for districts in Telangana (2018-2022)
Daily precipitation (Cumulative Rainfall)	
Motor Vehicle Tax collected	Telangana SEO (Socio-Economic Outlook Report 2018-2022) [Telangana State Portal Reports] (https://www.telangana.gov.in/downloads/reports)
Construction contribution to GSDP	
Personal Income per Capita	
Relative Humidity	

Usage

Similar to the Heat Wave models, we create a model for each pollutant for each of the 5 districts. The predicted concentration for each pollutant is then scaled to the AQI indices then choosing the worst pollutant, which is considered to be the prevailing AQI.

Libraries to be used:

- PyTorch
- TensorFlow
- Suntime
- Python Base packages
- Matplotlib
- Seaborn
- Sklearn
- Flask
- Vue3

References

1. Hossain, E., Shariff, M.A.U., Hossain, M.S., Andersson, K. (2021). A Novel Deep Learning Approach to Predict Air Quality Index. In: Kaiser, M.S., Bandyopadhyay, A., Mahmud, M., Ray, K. (eds) Proceedings of International Conference on Trends in Computational and Cognitive Engineering. Advances in Intelligent Systems and Computing, vol 1309. Springer, Singapore. https://doi.org/10.1007/978-981-33-4673-4_29
2. NeuralProphet: Explainable Forecasting at Scale (arXiv:2111.15397)
3. Zeppetello, Lucas R. Vargas, David S. Battisti, and Marcia B. Baker. "The Physics of Heat Waves: What Causes Extremely High Summertime Temperatures?", Journal of Climate 35, 7 (2022): 2231-2251, accessed Feb 2, 2023, https://doi.org/10.1175/JCLI-D-21-0236.1
4. Rohini, P., Rajeevan, M. & Srivastava, A. On the Variability and Increasing Trends of Heat Waves over India. Sci Rep 6, 26153 (2016). https://doi.org/10.1038/srep26153
5. Guigma, K.H., Guichard, F., Todd, M. et al. Atmospheric tropical modes are important drivers of Sahelian springtime heatwaves. Clim Dyn 56, 1967–1987 (2021). https://doi.org/10.1007/s00382-020-05569-9
6. Jun Ma, Yuexiong Ding, Jack C.P. Cheng, Feifeng Jiang, Yi Tan, Vincent J.L. Gan, Zhiwei Wan, Identification of high impact factors of air quality on a national scale using big data and machine learning techniques, Journal of Cleaner Production, Volume 244, 2020, 118955, ISSN 0959-6526, https://doi.org/10.1016/j.jclepro.2019.118955.
7. Dongsheng Zhan, Mei-Po Kwan, Wenzhong Zhang, Xiaofen Yu, Bin Meng, Qianqian Liu, The driving factors of air quality index in China, Journal of Cleaner Production (2018), doi: 10.1016/j.jclepro.2018.06.108