Historical Weather Data for Weather Prediction

1]Ricardo Trujillo

Abstract

Presented with weather data of 2022 and 2023 for Los Angeles, this time series analysis focused on key variables including temperature, pressure, sky cover, and rainfall. Explorations of these variables revealed cyclical flucations corresponding to seasonal changes, with variations in local extremas. Practical applications of this data range from crop yield prediction using SARIMAX to precision agriculture and risk reduction strategies.

Keywords: Weather Prediction, Forecasting, SARIMAX, Time Series Analysis

1 Background and Introduction

The data we have consists of weather in the Los Angeles area for the years 2022 and 2023. The data originally had various amounts of variables, but we distilled this information to what we thought was essentially the most important variables pertaining to the weather. These variables include temperature, pressure, sky cover, and rainfall. Each of these variables are measured through their averages, and correspond to their specific date within each respective month for the years 2022 and 2023.A practical application of this data would be for crop yield, as weather parameters can use machine learning regression techniques. Another would be precision agriculture, since time series analysis can aid in precision agriculture for farming sustainability on crop yield modeling. Overall risk reduction is another practical application of this data because we can apply it for something like weather index-based insurance. If the weather parameters reach a certain threshold, the insurance pays out. This allows for an additional protective barrier for farmers who might have lower crop yields due to adverse weather effects.

2 Components

In the exploration of the Temperature component, the same timeframe of months for 2022 and 2023 experienced a gradual rise and drop in temperatures — reinforced by the cyclical nature of seasons. One thing to note, the magnitude of the peak temperatures in each season varied. The data was tested for stationarity. With a p-value of 0.0467, the null hypothesis was rejected to assume stationary data. Then, the temperature values were fed into a SARIMAX model with (p,d,q) = (1,1,0,) and (P,D,Q,M) = (1,1,0,90). With these choices, the previous day's temperature will have an effect on the current day's temperature as well as assuming that the seasonal periods occur over 90 days (i.e. 365 divided by 4 seasons is roughly 90 days). Wit the model then being fitted and tested against the recorded Temperatures in the first quarter of 2024 (January-April), we obtained an MSE as shown in the Results section table.

The pressure variable is important for predictions as it can be used to predict incoming weather systems such as wind, rain, heat waves, or even hail. In the LA area it is very consistent hovering from 1.05 to 1.2 atmospheres. A SARIMAX model was also used with parameters of (2,0,1) and (1,1,0,30). 90 days was initially tried to account for the seasons however it was not quite able to predict the spikes of lower and higher pressure systems. 30 days was instead used to mimic a monthly cycle. The model for pressure returned an MSE score of 18.24 signifying that model is not the most accurate but still usable. Perhaps with an expanded dataset of more than 2 years it could be more accurate.

Viewing the Rainfall component, the Los Angeles area is very sporadic when it comes to when exactly the rainfall occurs. There are long periods of time where there is no rainfall, and there are others where there is a large spike in the amount of rainfall that occurs. Firstly, the data was split into test and train sets, and the last 30 days were used for testing. The SARIMAX model was then defined, with (p,d,q) = (1,1,1), and (P,D,Q,M) = (1,1,0,12). The model was then fit and predictions were made, and the RMSE was calculated which became 3.849. In the actual data, there are two spikes, with one being fairly large, compared to the model's prediction being fairly linear hovering around the 0.0 range.

For the Sky Cover component, we can note similar months express the same behavior in sky coverage. For example, spring months contain the steady increase in sky coverage where as the opposite occurs in the summer months onward. The SARIMAX model is defined as (1, 1, 1)x(1, 1, 0, 90)). The model fits in closely with our research as it's able to predict the cloudier days in the model relative to the least cloudy days. The model's MSE resulted in 9.384. While it may seem to appear well, the model still needs more work for predictive power, given that sunlight plays a big role in agriculture.

3 Concluding Remarks

By evaluating the data and obtaining these insights, we can extrapolate it to the central valley and wine regions. These findings are not just applicable to Los Angeles, but can also be extended to other regions. Understanding this project fully can allow for a broader understanding of agricultural trends and practices, therefore making implementation and planning more effective in the agricultural sector of these regions. The use of weather prediction technology ultimately improves decision-making, lead-

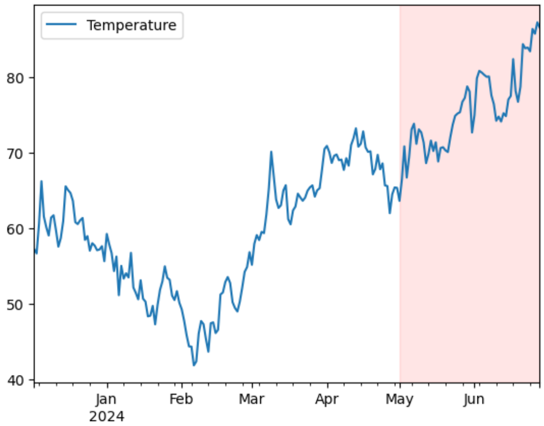
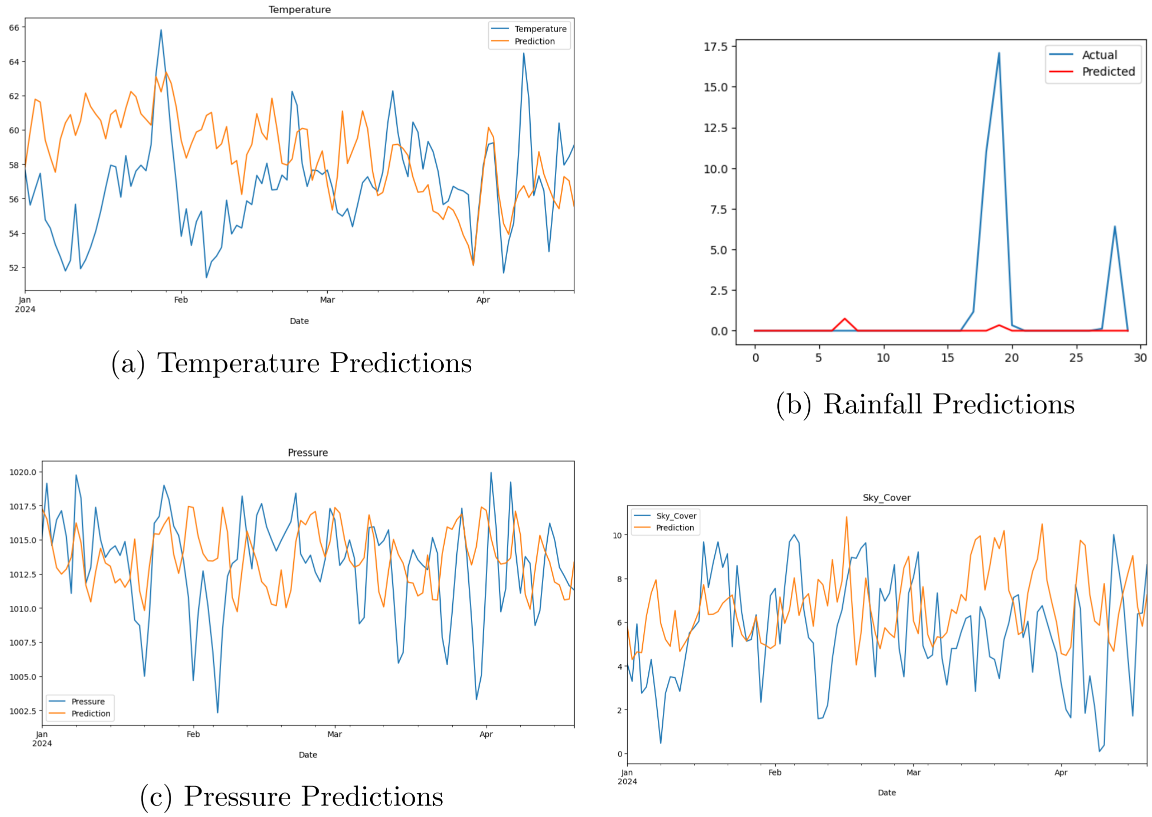


Figure 1: Temperature Forecast

ing to optimized crop yields as well as efficient resource management. Although the rainfall was not as easy to forecast due to the lack of data, accurate weather forecasts with these variables enable farmers to better adapt their practices to changing weather patterns, and promote resilience in the face of adversity caused by unpredictable weather.

4 Results



(a) Temperature Predictions

(b) Rainfall Predictions

(c) Pressure Predictions

(d) Sky Cover Predictions

Temperature MSE	Rainfall MSE	Pressure MSE	Sky Cover MSE
15.08	14.82	9.38	9.98

References

[1] California Measurement Advisory Council - California Weather Files. (n.d.). https://www.calmac.org/weather.asp