CS542 PRINCIPLES OF MACHINE LEARNING

COMMON TASK – REPORT

Submitted By – Manushi Munshi (U25855816) (manushi@bu.edu)

1. Task Overview

The objective of this task was harnessing machine learning methodologies to anticipate daily climate occurrences in four prominent U.S. urban cities: New York, Chicago, Austin, and Miami. The different phases of the task included manual predictions, data extraction, model predictions and automated predictions. Trades were made on the demo Kalshi account depending on the predictions for the maximum temperature for the four cities. Throughout the first week, manual trades were executed, while the second week was dedicated to refining the model's predictive accuracy through training on historical data. The third week saw the transition from manual to automated trading, using the model's daily predictions. This report aims to detail the methodologies employed, analyze the outcomes, and evaluate the model's performance, with particular attention to the trades executed and the machine learning strategies utilized to navigate the complexities of climate event prediction.

2. Data Collection

The first task for the second week was to identify weather data sources for the four cities. I explored several weather APIs like –

(1) Open Meteo
(2) Visual Crossing
(3) Meteostat
(4) Meteomatics
(5) NCEI/NOAA

There were restrictions on the number of free API calls that could be made to these sources for getting historical weather data. After thorough study, I decided to stick with three data sources, i.e., Open Meteo, Visual Crossing and Meteostat which gave me the most flexibility and efficiency in data retrieval. The data retrieved was for 4 months before the current date which was used to train the model. This duration of 120 days was decided by assessing the rate limits of the APIs and this was the maximum duration for which the data could be efficiently extracted. Python scripts were used for data retrieval to make requests to each data source API endpoint using the desired location and date range details for all the four cities. Upon receiving the API response, the data was processed and transformed into a structured pandas DataFrame format for further analysis, ensuring accuracy and consistency across the collected datasets. The weather variables extracted were daily maximum and minimum temperatures, precipitation, wind speed, wind gusts, and snowfall.

3. Data Cleaning and Feature Aggregation

The weather data extracted from multiple sources was prepared for further analysis by data cleaning. To ensure consistency across datasets, we standardized column names and handled missing values uniformly. Specifically, missing values were replaced with zeros to maintain dataset integrity. The date column was converted to the pandas datetime format, and the date part was extracted from the timestamp. Additionally, we performed unit conversions to ensure uniformity in measurement units across temperature (F) , precipitation (inches), wind speed (mph), wind gusts (mph), and snowfall (inches) variables.

The next step was aggregation of features to form one consolidated DataFrame for model training. This included averaging values for key features like temperature, precipitation, wind speed, and gusts to smooth out discrepancies between forecasts. This method enhances the reliability of the data by offsetting potential biases or inaccuracies inherent in individual datasets. The aim is to provide a more accurate and consensus-based view of weather conditions, laying a solid foundation for the model training. This approach ensures our analyses are both robust and reflective of a comprehensive understanding of weather patterns.

4. Exploratory Data Analysis

For the exploratory data analysis phase, two graphs were plotted for each of the four cities' datasets. The first plot was a time series variation for the maximum temperature value for the three individual datasets and the consolidated dataset. These line plots vividly illustrate the daily variation of maximum temperature readings across the four cities for the three data sources and the aggregated data. Each line represents a different source, showing how closely the datasets track with one another, while also highlighting where discrepancies arise, potentially due to differing collection methods or interpolation techniques used by each source.

The second plot presents a comparative boxplot visualization of maximum temperature readings from the three datasets and the consolidated dataset, illustrating the range and distribution of temperatures captured by each source for each city. These boxplots help in analyzing the consistency of data within each source and evaluate the central clustering of values around the median, offering insights into typical temperature ranges and extremities reported by each service.

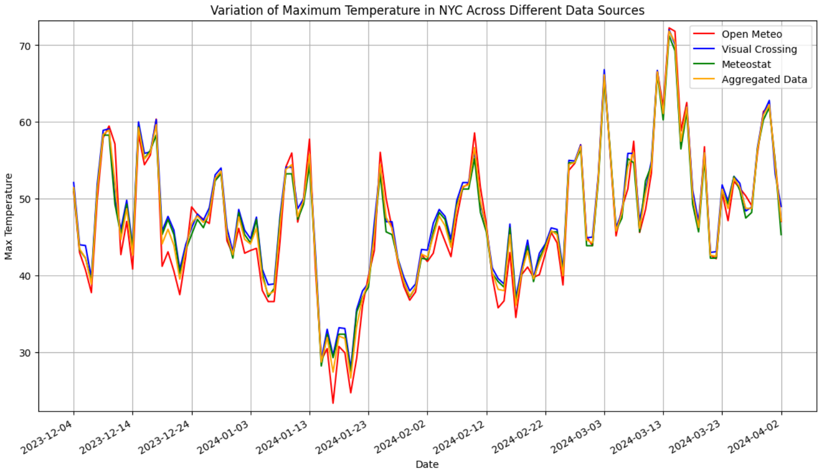
The line plot for New York City reveals a moderate variation between sources, with the aggregated data smoothing out the spikes and dips, offering a balanced view of the temperature trends. The boxplot displays a good level of consistency in the temperature readings, with all sources indicating similar medians, though with some variation in the interquartile ranges and outliers.

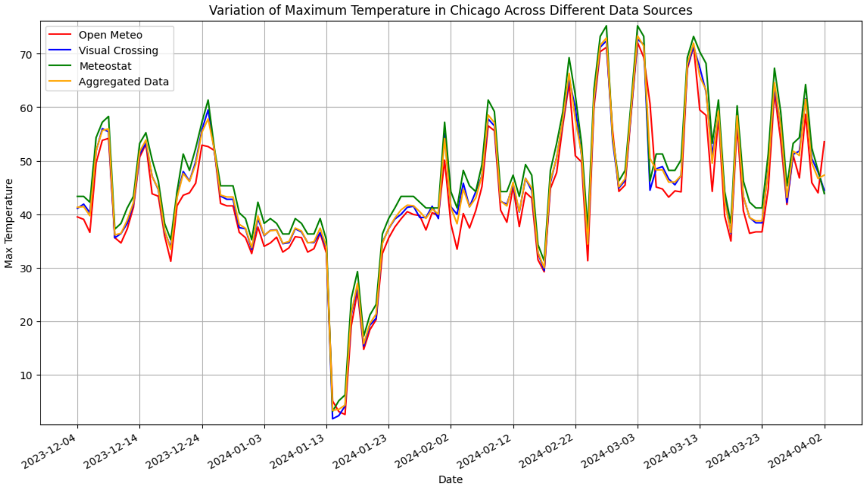
The temperature in Chicago, as per the line plot, exhibits greater variability among sources compared to Austin, with noticeable divergence on certain days. This could be due to microclimatic variations that different stations capture. The boxplot reveals a wider range, particularly for Meteostat, which suggests this source records a broader spectrum of temperature variations, and potential outliers indicate sporadic temperature events that are not as consistently recorded by other sources.

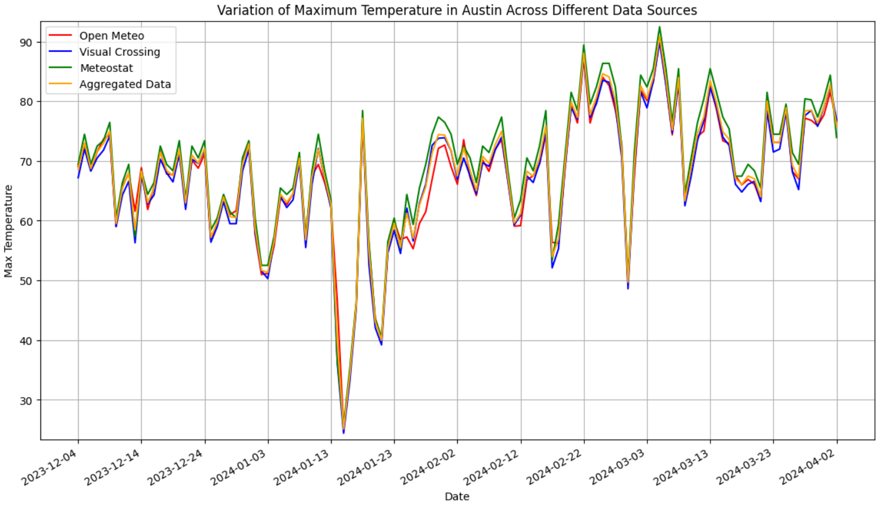
The line plot for Austin shows a high level of agreement among the different data sources regarding maximum temperature trends, with only slight variations. This suggests that in Austin, the data is consistent and reliable across platforms. The boxplot reinforces this, showing similar medians and interquartile ranges, albeit with some outliers, indicating occasional deviations from the typical temperature range.

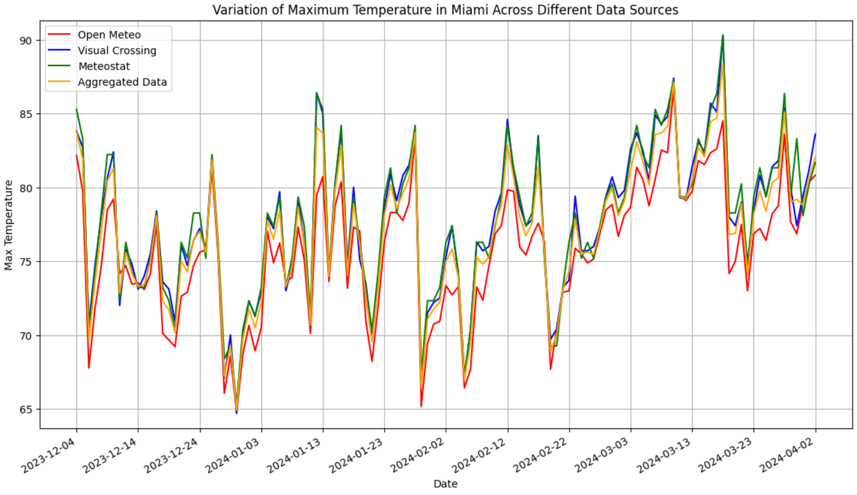
In Miami, the line plot shows a tight clustering of temperature readings, suggesting that the data sources concur closely on the maximum temperatures, reflective of the less variable, warmer climate. The boxplot's tighter interquartile ranges and fewer outliers across sources corroborate the less volatile nature of Miami's climate as reported by the data providers.

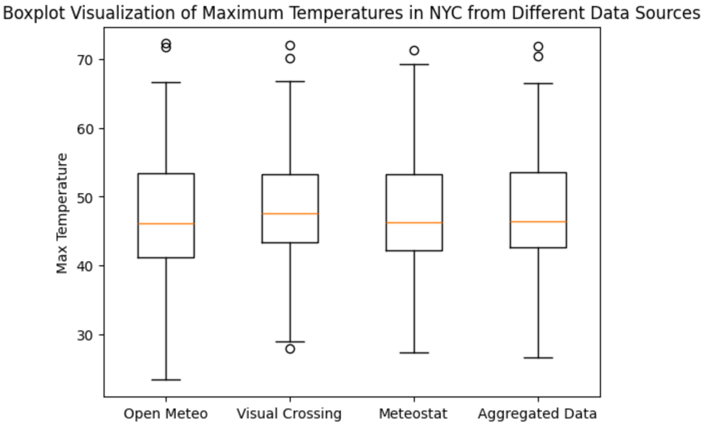
Overall, the plots suggests that the temperature readings are consistent across different sources for each city, with some expected variability. The outliers and variations in the boxplots may indicate differences in measurement precision or data collection methods. The aggregated data generally provides a middle ground that could potentially offset individual source anomalies, giving a more stable picture of the temperature trends across these cities.

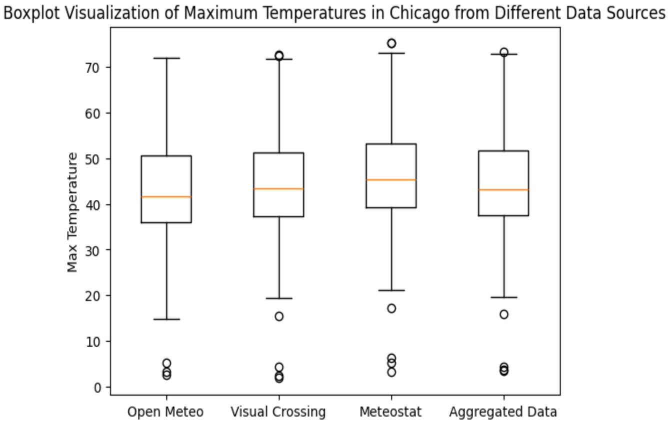


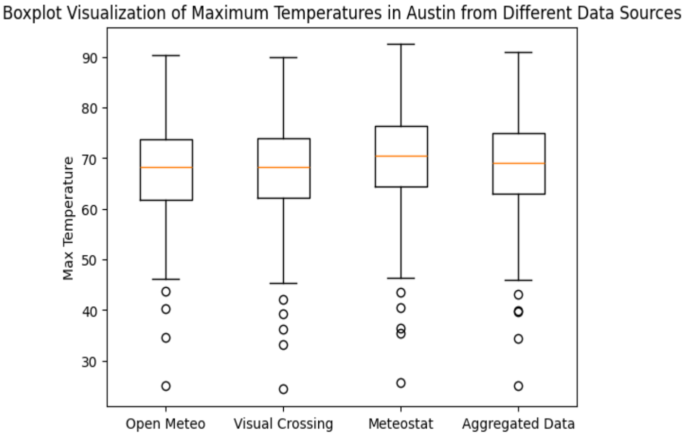

Variation of Maximum Temperature in Chicago Across Different Data Sources

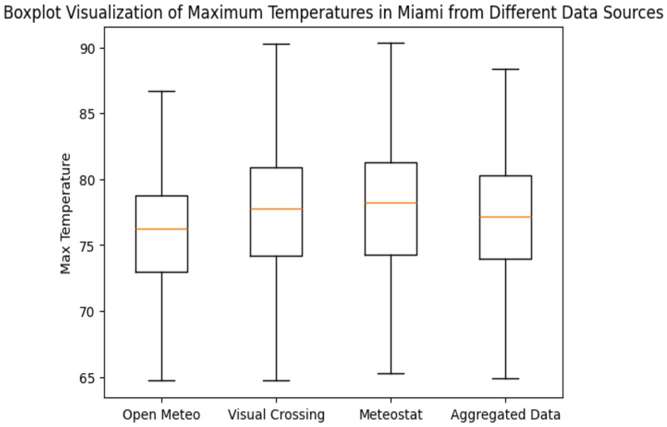

Variation of Maximum Temperature in Austin Across Different Data Sources


Variation of Maximum Temperature in Miami Across Different Data Sources









5. Data Pre-Processing and Model Development

The next step involved pre-processing the consolidated dataset and preparing it for model training. Firstly, the feature set was standardized, excluding the 'date' and target variable max_temp, using a Standard Scaler to normalize the data and reduce potential biases in the models due to varying scales. This normalization was performed to ensure that each feature contributed proportionately to the predictive process. The last row of the DataFrame, which contained the variables for the current date, was excluded from the training and testing set to validate the model's predictive accuracy on unseen data. The normalized data was then split into train and test sets with an 80-20 split using the `train_test_split` function, enabling the evaluation of the model's performance on unseen data.

For model training, a variety of regression models were implemented. The first model was a traditional Linear Regression, to assess the performance of regression on the data. The next models implemented were Ridge and Lasso Regression models, which introduced L2 and L1 regularization, respectively. These regularization techniques were implemented to prevent overfitting by penalizing large coefficients, thus enhancing the models' generalizability. The models were evaluated using the Mean Squared Error(MSE).

To optimize the Ridge and Lasso models, cross-validated versions—RidgeCV and LassoCV—were implemented, which methodically searched for the best regularization parameters, alpha. This process not only refined the models but also helped find the most optimal alpha values that minimized cross-validated MSE. With the aim of further improving the performance, XGBoost Regressor, an ensemble learning method known for its performance and speed, was then implemented. Hyperparameter tuning was performed on the XGBoost Regressor using GridSearchCV to search through a defined hyperparameter space, identifying the combination that minimized the MSE. The XGBoost was then trained on the best hyperparameters obtained.

An ensemble model, constructed by averaging the predictions of the individual models, was established as a strategic approach to harness the collective predictive strength and to offset individual model biases. This model's performance, indicated by its MSE, served as a comparative benchmark against the individual models. The model with the lowest MSE was then used for the final trade prediction.

6. Result Analysis

For NYC, the linear regression models showed relatively high Mean Squared Error (MSE) values, reflecting challenges in capturing the complex weather patterns of the region. The Ridge and Lasso Regression models offered slight improvements over linear regression due to their capability of handling overfitting. The XGBoost model was the better choice out of the other three as it gave least MSE thereby showcasing its ability to capture nonlinear relationships and intricate patterns in NYC's dynamic climate. The ensemble model further enhanced predictive accuracy and was used to make the final trade prediction.

In Chicago, linear regression models displayed significantly higher MSE compared to other cities, possibility because of the challenges posed by the region's variable and often extreme weather conditions. While Ridge and Lasso Regression models provided modest improvements, the XGBoost model demonstrated superior performance, suggesting its effectiveness in capturing complex interactions among weather variables. The ensemble model showcased a notable reduction in MSE compared to standalone models and was thus used to make final trade prediction.

Miami has a relatively stable and predictable climate, and thus exhibited lower MSE values across all models of the other cities. The linear regression models performed reasonably well, reflecting the city's comparatively stable weather dynamics. While Ridge and Lasso Regression models further refined predictions, the XGBoost model did not yield significant improvement over linear models in Miami. Nevertheless, the ensemble model provided a slight reduction in MSE, showcasing its potential to enhance predictions even in relatively stable climates like Miami. The Ridge Regression gave the minimum MSE and was used to make final trade prediction.

In Austin, the linear regression models displayed relatively high MSE values, indicating challenges in accurately capturing the intricate interactions among weather variables in city’s climate. While Ridge and Lasso Regression models provided marginal improvements over linear regression, the XGBoost model demonstrated a significant reduction in MSE, suggesting its effectiveness in capturing nonlinear relationships and complex weather dynamics. Surprisingly, the ensemble model did not yield a notable improvement over individual models, indicating that the combined predictive strength of the models did not contribute significantly to enhancing predictive accuracy in Austin's weather forecasting. The XGBoost was used to make the final trade predictions.

The discrepancies in MSE values and the varying optimal models across different cities can be attributed to the diverse climatic characteristics inherent to each location. Factors such as geographical features, proximity to water bodies, altitude, and local atmospheric conditions contribute to distinct weather patterns. Cities like New York and Chicago experience pronounced seasonal changes, leading to higher variability in weather variables. In contrast, cities like Miami and Austin have milder climates with less variability. Regression models trained on data from cities with more volatile weather may struggle to capture underlying relationships accurately, resulting in higher MSE values. Moreover, the optimal model choice may differ due to the unique interplay of factors influencing weather dynamics in each city, highlighting the importance of selecting models tailored to specific climatic conditions.

7. Trade History

Week 1 Manual Trades

6 out of 19 predictions are correct (approx. 32%)

Sr. No.	Date	Market	Direction	Settlement
1	Feb 27	NYC high on Feb 27	No	No
2	Feb 27	Austin high on Feb 27	Yes	No
3	Feb 27	Miami high on Feb 27	Yes	No
4	Feb 27	Chicago high on Feb 27	Yes	No
5	Feb 29	NYC high on Feb 29	Yes	No
6	Feb 29	Austin high on Feb 29	Yes	No
7	Feb 29	Miami high on Feb 29	No	No
8	Feb 29	Chicago high on Feb 29	No	No

9	March 1	NYC high on March 1	Yes	No
10	March 1	Austin high on March 1	Yes	No
11	March 1	Miami high on March 1	Yes	Yes
12	March 1	Chicago high on March 1	Yes	No
13	March 6	NYC high on March 6	Yes	Yes
14	March 6	Austin high on March 6	Yes	No
15	March 6	Miami high on March 6	Yes	No
16	March 6	Chicago high on March 6	No	Yes
17	March 7	NYC high on March 7	No	No
18	March 7	Austin high on March 7	Yes	No
19	March 7	Miami high on March 7	Yes	No

Week 2 Model Prediction Trades

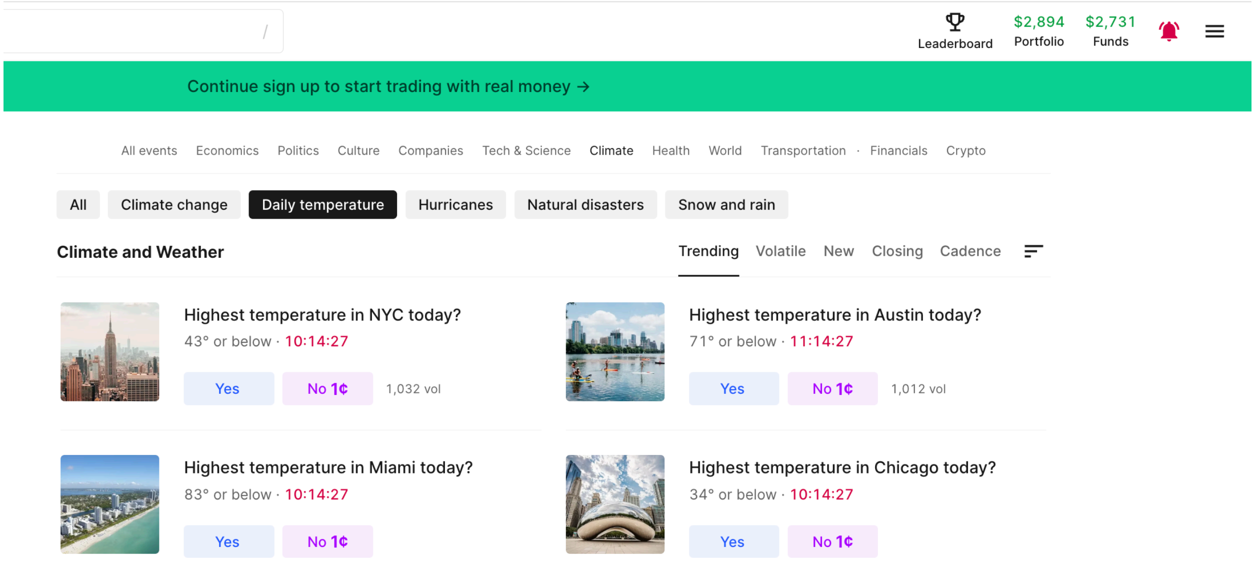
15 out of 24 predictions are correct (approx. 62%)

Sr. No.	Date	Market	Direction	Settlement
1	March 19	Austin high on March 20	Yes	No
2	March 19	Chicago high on March 20	No	No
3	March 20	NYC high on March 20	Yes	No
4	March 20	Austin high on March 20	Yes	Yes
5	March 20	Miami high on March 20	Yes	No
6	March 20	Chicago high on March 20	Yes	No
7	March 21	NYC high on March 20	No	No
8	March 21	Miami high on March 20	Yes	No
9	March 21	Miami high on March 20	No	No
10	March 21	Chicago high on March 20	Yes	No
11	March 22	NYC high on March 22	Yes	Yes
12	March 22	NYC high on March 22	No	No
13	March 22	Austin high on March 22	No	No
14	March 22	Austin high on March 22	Yes	No
15	March 22	Miami high on March 22	Yes	Yes
16	March 22	Miami high on March 22	No	No
17	March 22	Chicago high on March 22	Yes	Yes
18	March 22	Chicago high on March 22	No	No
19	March 23	NYC high on March 23	No	No
20	March 23	NYC high on March 23	Yes	No
21	March 23	Austin high on March 23	Yes	Yes
22	March 23	Austin high on March 23	No	No
23	March 23	Miami high on March 23	Yes	No
24	March 23	Chicago high on March 23	No	No

Week 3 Automatic Trades

19 out of 27 predictions are correct (approx. 70%)

Sr. No.	Date	Market	Direction	Settlement
1	March 29	Chicago high on March 29	No	No
2	March 29	Chicago high on March 29	Yes	No
3	March 29	NYC high on March 29	No	No
4	March 29	NYC high on March 29	Yes	Yes
5	March 29	Miami high on March 29	No	No
6	March 29	Miami high on March 29	Yes	No
7	March 29	Austin high on March 29	No	No
8	March 29	Austin high on March 29	Yes	No
9	March 30	Chicago high on March 30	No	Yes
10	March 30	NYC high on March 30	No	No
11	March 30	NYC high on March 30	Yes	No
12	March 30	Miami high on March 30	No	No
13	March 30	Miami high on March 30	Yes	No
14	March 30	Austin high on March 29	No	No
15	March 30	Austin high on March 29	Yes	Yes
16	March 30	Chicago high on March 30	No	Yes
17	April 1	NYC high on April 1	No	No
18	April 1	NYC high on April 1	No	No
19	April 1	Miami high on April 1	No	No
20	April 1	Miami high on April 1	Yes	Yes
21	April 1	Austin high on April 1	No	No
22	April 1	Austin high on April 1	Yes	Yes
23	April 2	NYC high on April 2	No	No
24	April 2	NYC high on April 2	Yes	No
25	April 2	Miami high on April 2	No	No
26	April 2	Austin high on April 2	No	No
27	April 2	Chicago high on April 2	No	No



Leaderboard $2,894 Portfolio $2,731 Funds

Continue sign up to start trading with real money →

All events Economics Politics Culture Companies Tech & Science Climate Health World Transportation · Financials Crypto

All Climate change Daily temperature Hurricanes Natural disasters Snow and rain

Climate and Weather

Trending Volatile New Closing Cadence

Highest temperature in NYC today?
43° or below · 10:14:27
Yes No 1¢ 1,032 vol

Highest temperature in Austin today?
71° or below · 11:14:27
Yes No 1¢ 1,012 vol

Highest temperature in Miami today?
83° or below · 10:14:27
Yes No 1¢

Highest temperature in Chicago today?
34° or below · 10:14:27
Yes No 1¢

8. Challenges

Several challenges were encountered throughout the task, primarily related to data collection, model development, and trading execution. One of the main challenges was sourcing reliable and comprehensive weather data for the four cities from various APIs while adhering to API call limits and rate restrictions. The variability in data quality and consistency across different sources posed challenges in standardizing and aggregating the datasets for model training. Additionally, selecting the appropriate machine learning models and optimizing their hyperparameters posed challenges due to the diverse and dynamic nature of weather patterns in each city. Balancing model complexity with interpretability was another challenge, especially considering the trade-off between model performance and computational resources required for training and inference. Lastly, transitioning from manual to automated trading presented challenges in deploying the predictive models effectively in real-time trading environments, including integrating the models with trading platforms and ensuring seamless execution of trades based on model predictions.

9. Conclusion

In conclusion, the task provided valuable insights into leveraging machine learning techniques for climate event prediction and automated trading. By harnessing historical weather data and employing regression and ensemble learning models, the daily temperatures in the four cities were anticipated. Despite the challenges encountered, the models demonstrated promising predictive accuracy, with varying degrees of success across different cities. The automated trading system showed potential in translating model predictions into profitable trades, albeit with room for further optimization and refinement. Moving forward, continued research and development in machine learning methodologies and their application to weather forecasting and trading automation hold promise for enhancing decision-making processes in financial markets and other domains influenced by environmental factors.