University of Notre Dame

# Modeling for Temperature Forecasting and Prediction Markets

Group #15 - Alex Heck (aheck3), Tim McDonald (tmcdona5), Cooper Foster (cfoster3)

Time Series Forecasting MSBR-70320-SS-1F
Professor Sriram Somanchi
February 2025

## Introduction and Data Description

Polymarket is an event prediction market founded in 2020. The company experienced a surge in popularity in the last year as it offered a range of betting markets on the 2024 U.S. elections. At the same time, some of Polymarket's competitors became formally legalized for U.S. bettors and regulated by the Commodity Futures Trading Commission. As of early 2025, Polymarket is still barred from operating in the U.S., due in part to Polymarket's business being based on the blockchain. Polymarket specifically hosts all markets, contracts, trades, and resolutions on the Polygon blockchain (where it derives its name), a technology built off the second-largest cryptocurrency, Ethereum. While Polymarket's use of web3 technology draws caution from U.S. regulators, it also adds a unique layer of transparency, as every single Polymarket transaction is stored in an open blockchain. Our group did not perform historical queries of that blockchain data, but in the spirit of that transparency, Polymarket makes it very simple to acquire a limited range of historical market data.

Ultimately, our group focused on studying a weather-prediction market that Polymarket offers. Specifically, we picked a market that tracks the daily high temperature in New York City. We will discuss more about why we picked this market below, but for the purposes of acquiring data, we were able to look back to January 22nd of this year to pull daily market price data. It is worth noting that this "market" is not a single market; rather, it is a continuous offering of distinct daily markets. For each market, seven outcome contracts are offered. Five of these are temperature intervals of two degrees Fahrenheit each, with the other two set lower (x degrees or below) and upper (x degrees or above) temperature limits. Markets typically open a day before their expiration, so their average lifespan is around forty-eight hours. Our team was able to pull hourly price snapshots from each of these markets going back from February 21st to January 22nd. Markets were not offered for January 27th and February 10th, so we collected twenty-nine market datasets in total. In the appendix, see **Figure 1** for a visualization of how contract prices move throughout the trading period.

Collecting this data was necessary for validating our model's prediction, but we did not involve this data with actual modeling. Instead, we were interested in the variable that determines the outcome of this market: the daily high temperature in New York City. Polymarket resolves these markets using data from the National Weather Service's KLGA weather station at LaGuardia Airport in Queens, New York. Our team pulled a dataset for this weather station containing the daily high temperature looking back over five years to January 1st, 2020. See **Figure 2** for a visualization of those daily high temperatures across the five years of the dataset.

## Context and Objectives

Pricing for event contracts on Polymarket will always be determined by the “wisdom of the crowd”, where bids and asks will eventually meet to make a transaction and set a price. Correctly predicted contracts have a nominal value of one dollar, and because these contracts have binary outcomes, the price of a contract, in cents of one dollar, should roughly reflect the implied probability of that event occurring. In Polymarket’s most popular and liquid markets, the price should nearly equal the implied probability.

The most important question for our group’s analysis is whether the implied probability set by market pricing for an event accurately reflects the actual probability of that event occurring. For a given contract, if the implied probability and real probability are equal, the expected value of that contract is zero and an informed trader would not participate in the market. If the priced probability and the real probability diverge, then the expected value turns positive, and informed traders should purchase contracts accordingly.

Many markets on Polymarket offer contracts for events where it is extremely difficult, if not impossible, to accurately estimate the real probability of the event occurring. Take the market for “Taylor Swift and Travis Kelce engaged in 2025?”, for example. There is no dataset that our team is aware of to reliably estimate the probability of this event happening. In cases like these, of which there are many on Polymarket, the best estimation of the real probability of an event is likely the implied probability of the market pricing itself. This is thanks to the effect of the “wisdom of the crowd”. In theory, the crowd’s average forecast should prove even more accurate since each member of the crowd has a financial incentive to be correct in their individual forecast.

Other markets offer contracts for events where the real probabilities can be estimated. Our team thinks that the market for the daily high temperature in New York City is one of these cases, since daily high temperatures fall in an expected frequency distribution. This leads our team to believe that they may have predictability. See **Figure 3** for a density plot of daily high temperatures for our full time series.

For the density function in **Figure 3**, we see that the high temperatures follow a bimodal distribution. From examining this shape, we know that the distribution of these temperatures adheres to the central limit theorem, and the distribution of future temperatures should follow this density function.

Our team believes that this daily temperature variable should have a degree of predictability, and we will test that belief below. If our team can make informed and accurate predictions about what the high temperature will be for the next day as expressed by $Y_{t+1}$, we could trade based on those forecasts. If we can identify cases where a contract pricing's implied probability does not match our forecast of that event's probability, then, given enough transactions in this state, we would be almost guaranteed to generate a profit from trading activity.

## Evaluating Predictability

Before modeling our temperature data and producing our own forecasts, it was important to test our assumption that the temperature data was in fact predictable and not a random walk. By plotting the autocorrelations of the daily high temperature data, we observe a high correlation between periods of lag, as shown in **Figure 4**.

Next, we can employ two methods to assess our data for predictability. First, we will plot the autocorrelations of the differenced daily high temperature time series, plotted in **Figure 5**.

Then, we will fit an AR(1) model to the time series to get a coefficient of 0.911 and a standard error of 0.0096. We divide the difference of the coefficient and one by the standard error to calculate the t-stat: -9.27. Because the t-stat falls outside of the range of -2 to 2, we reject the null hypothesis that the temperature data is a random walk.

Now that we have a formal confirmation that the daily high temperature is not a random walk, we can proceed to modeling the data to produce forecasts.

## Modeling:

To forecast the daily high temperature, we examined five different models. Following is a list of these models and how we constructed them.

### *Seasonal Naive:*

For our first modeling exercise, we decided to benchmark some more sophisticated models against a seasonal naive model. To prepare our data, we filtered out two leap days (2/29/2020 and 2/29/2024) present in the data and set a frequency of 365 periods for the daily data. We used the same period that we pulled Polymarket data as our validation data: the month from January 22nd to February 21st of 2025. With a baseline set, we can move to more advanced methods.

### *Linear Regression:*

Using the same training and validation split as above, we trained a linear regression model. As shown in the visualization above (cite Figure), our data shows strong seasonality but little if any observable trend. For this reason, we only modeled seasonality in our linear regression.

### *ARIMA:*

To build our ARIMA model, we used several rolling validation periods in 2023 and 2024. We looped through these validation periods, saving the error metrics, and then examined the average of these error metrics to initially validate our model and tune the parameters. For our final tuned ARIMA model, we used the parameter set (5,0,0), where we use 5 lagged observations, no differencing is applied, and there are no lagged errors included in our forecasts. After tuning the model parameters, we recombined the data from the rolling validation periods in 2023-2024 with the preceding training data and retrained the model using the same data set as used in the S-Naive and Linear Regression models. Then, we validated the ARIMA model over the same period as those models.

### *Neural Net:*

In a similar fashion to how we tuned our ARIMA model, we also used rolling validation periods in 2023 and 2024 to select the best parameters. The best combination ended up as inputting fourteen lagged observations, two seasonally lagged input observations, five neurons in the hidden layer, and twenty model repeats per forecast. Like the ARIMA model building process, after selecting our parameters, we retrained the model on the complete training set and validated over the select period of 2025 observations.

### *Naive:*

After building and validating the previous four models, we noted a strong discrepancy in the performance of the S-Naive model, and to some extent the linear regression as well. We created a Naive forecast to produce a set of predictions that would ignore the seasonality of the data, and the analysis of this benchmark greatly contributed to our conclusions below.

## Model Forecast Accuracy

We calculated the error metrics for our various models and included these in our appendix, shown in figures **8-12**. Across the metrics, we noted generally better performance for forecasts that minimized or ignored seasonal values. The S-Naïve and Linear Regression forecasts, both utilizing seasonal values, fared worse than their model peers.

Along with the error metrics that our team generated, we sought to validate our models in another way: by examining our hypothetical performance had we traded contracts on Polymarket based solely on these forecasts. The same validation period of January 22nd to February 21st was used, and we pulled historical Polymarket data from this period. To make these comparisons we looked at our continuous forecasted temperatures and classified them as one of the seven contracts offered on these daily markets. We then assigned each model’s daily prediction as either correct or incorrect based on how that contract resolved.

For our hypothetical validation, we assumed that we bought $100 of the model’s predicted contract at the best available price at or near 6pm on the day preceding the market's expiration. For the days that our models correctly predicted a winning contract, we calculated our profit as the final value of our contracts minus the initial $100 they cost. For the days that our models failed to predict a winning temperature, contracts resolved valueless, so this was simply a loss of $100. This table can be seen by viewing **Figure 6**.

Looking across the five models that we generated, the Naive and ARIMA models generated the best net profit/loss. The seasonal naive model also yielded a profit. The linear model and the neural network model generated losses.

Additionally, our group wanted to investigate the frequency with which our models were able to correctly forecast correct temperatures and winning contract intervals. See **Figures 7a** and **7b**.

The naive model performed the best, generating correct predictions eight times out of our twenty-nine validation days. The neural network, naive, and linear models performed the worst, with correct prediction on only three out of twenty-nine days.

As a final benchmark of our models' performance, consider that there are seven contracts offered each day that can be randomly selected. Over twenty-nine markets, we would expect a blindfolded chimpanzee to make just over four correct predictions. Given that our validation window was not extensive, this does not automatically invalidate the forecasting methods that failed to outperform our meteorologist primate friend, but it is not encouraging for the predictive ability of these underperforming methods moving forward.

## Conclusion: Summarizing Results and Their Impact

Temperature is a quintessential example of seasonal data. Paradoxically, seasonality did not help to predict temperatures as much as we had thought it would during the validation period we selected. During this validation period, the contiguous United States experienced a far colder month than is normally expected. We observed newsworthy events such as New Orleans, Louisiana experiencing a blizzard and the Presidential inaugural ceremony being moved indoors due to frigid temperatures in Washington, D.C.. While January was a warm month globally, the United States was a considerable exception, as shown by the temperature anomaly in **Figure 13**.

Our team believes this is why our ARIMA model, which does not account for seasonal values, performed better than the linear regression model and the neural network. When comparing the two benchmark models, we also see a much stronger performance from the Naive model over the

S-Naive. As a potential further analysis, our team would be interested in revalidating these models over a period where temperatures followed a more typical seasonal pattern.

As we consider the implications of our analysis, our team believes we have answered our primary question. It seems that the market prices for these contracts do not always reflect the real probabilities of their respective outcomes, and so we believe that there are opportunities for informed traders to capitalize on mispriced contracts.

Like other markets for financial securities, these daily markets will generally be rewarding to those participants with better information at the expense of participants operating with less information. While forecasting using models like ours will not provide a “magic bullet” for zero-risk returns, they can provide an information advantage which could be the difference between being in the black or red for traders in these competitive markets.

We also think that the implications of our analysis extend beyond temperature betting markets on the blockchain. When we notice our seasonal models underperforming, we see a real effect of a changing climate. As seasons shift and temperature variability increases, temperature data will become less and less stationary over time, and historical datasets like the one we have used to build our models could become obsolete for forecasting. In this case, meteorologists and climatologists will need to retrain and rebuild their models for forecasting tomorrow’s weather–and beyond.

## References:

LaGuardia Aiport Weather Station Data
*Meteostat*.net
https://meteostat.net/en/station/72503

Highest Temperature in NYC on Feb 25? *(example market)*
*Polymarket.com*
https://polymarket.com/event/highest-temperature-in-nyc-on-feb-25?tid=1740717871269

The polar vortex is acting weird and the US is paying the price this winter
*Mary Gilbert, cnn.com*
https://www.cnn.com/2025/02/21/weather/polar-vortex-cold-winter-climate/index.html

## Appendix:

**Figure 1:** Polymarket price screen illustrating how contract prices move with time

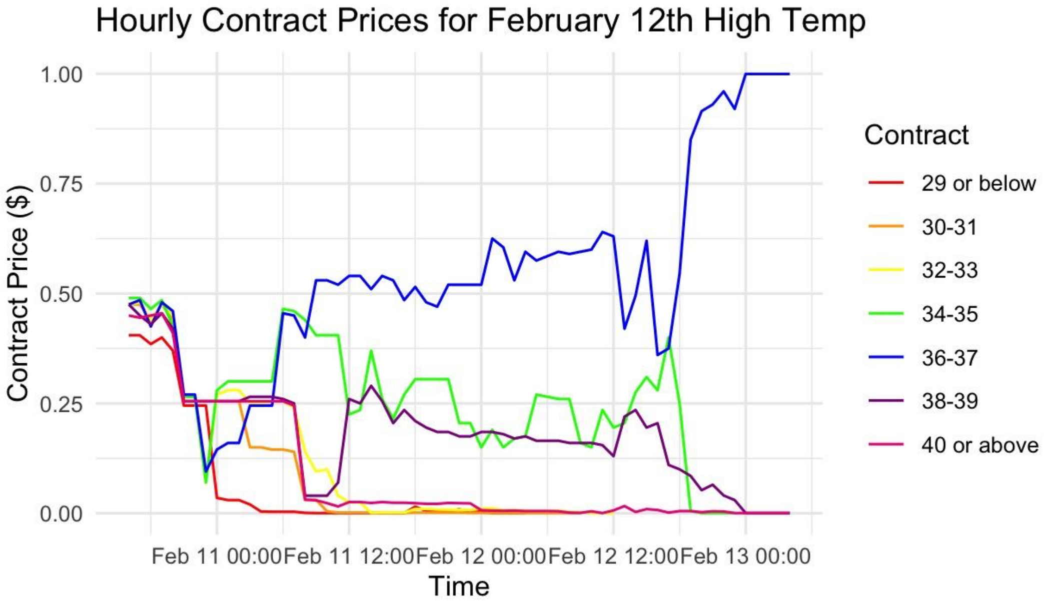


**Figure 2:** Time plot of daily high temperatures across the five year span of the dataset

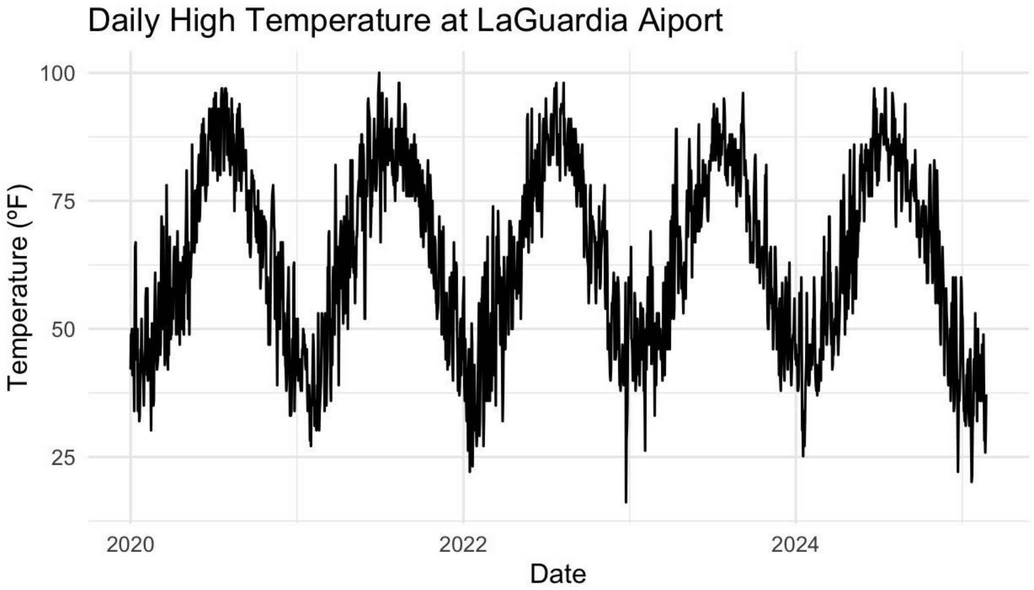


**Figure 3:** Illustrated Density Plot of daily high temperature for our full time series

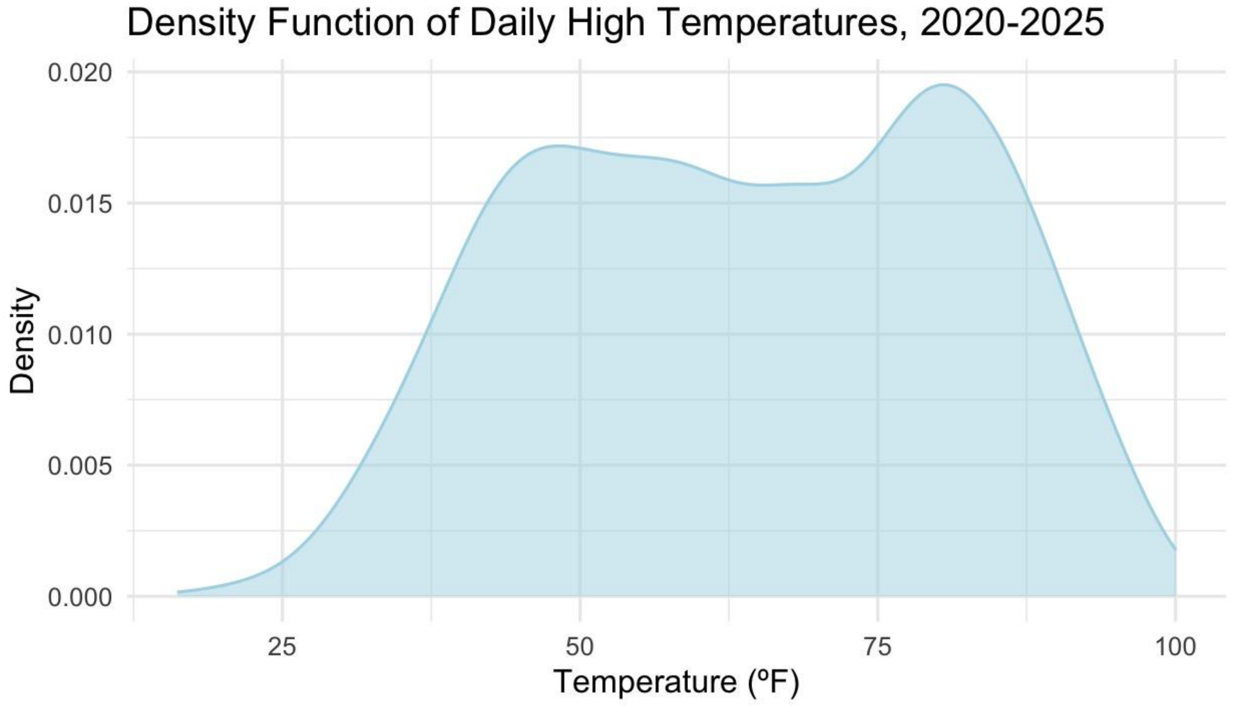


**Figure 4:** ACF Plot evaluating for random walk; Autocorrelations of daily high temp. data

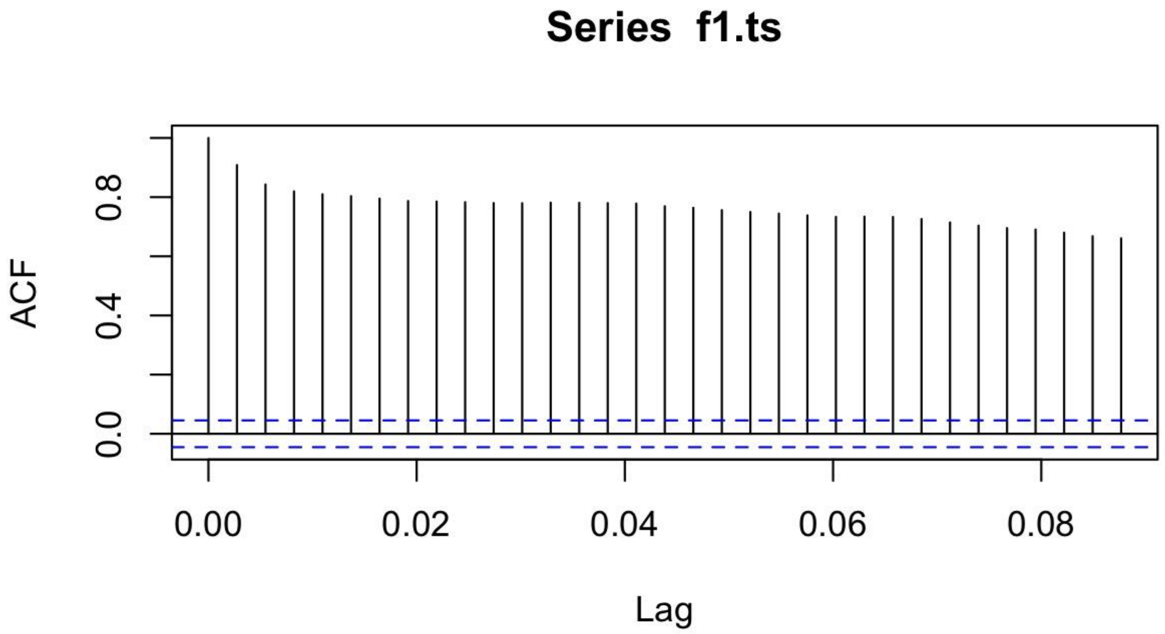


**Figure 5:** Autocorrelations of the differenced daily high temperature time series

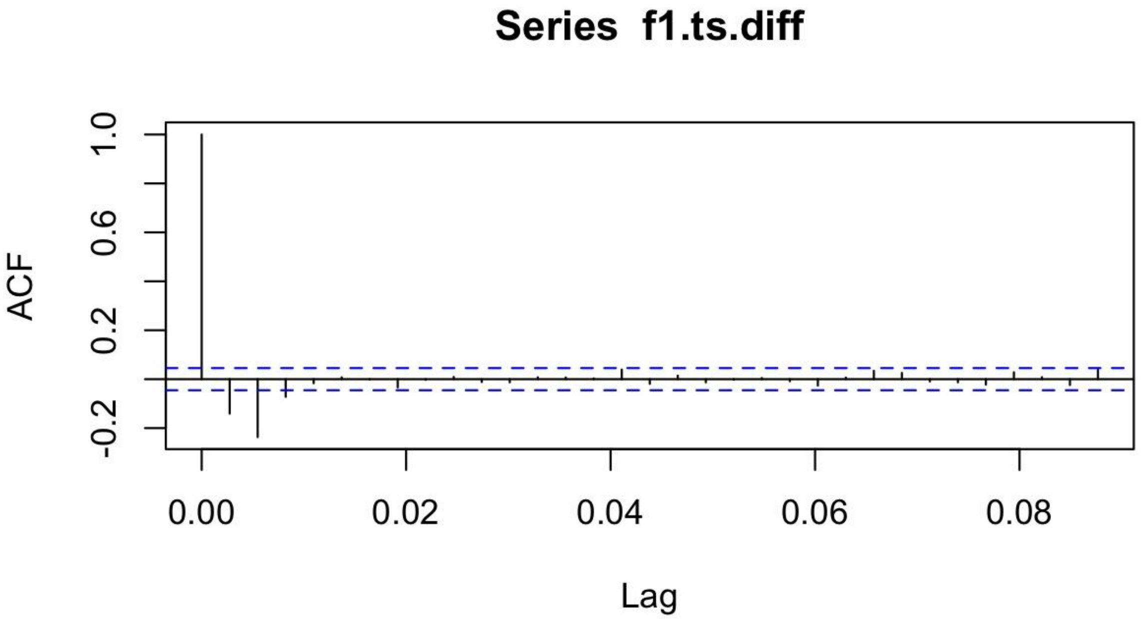


**Figure 6:** Profit/Loss of each model

| | NAÏVE | SNAÏVE | LinearMod | ARIMA | NeuralNet |
|---|---|---|---|---|---|
| **Winnings** | $12,471 | $9,382 | $231 | $12,380 | $1,123 |
| **Losses** | $(2,100) | $(2,600) | $(2,600) | $(2,200) | $(2,600) |
| **Net P/L** | $10,371 | $6,782 | $(2,369) | $10,180 | $(1,477) |

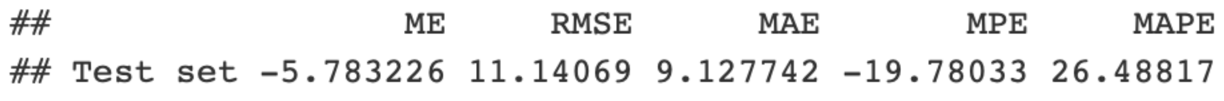
**Figure 7a:** Calendar of correct predictions by model

| | | | | | | |
|---|---|---|---|---|---|---|
| 1/22 | 1/23<br>ARIMA | 1/24 | 1/25<br>NAIVE<br>ARIMA | 1/26 | 1/27<br>*No market* | 1/28<br>NAIVE<br>LM<br>NN |
| 1/29 | 1/30<br>SNAIVE<br>LM | 1/31 | 2/1 | 2/2<br>NAIVE<br>SNAIVE<br>ARIMA | 2/3 | 2/4<br>NAIVE<br>ARIMA |
| 2/5 | 2/6<br>ARIMA | 2/7<br>ARIMA<br>NN | 2/8 | 2/9<br>NAIVE | 2/10<br>*No market* | 2/11 |
| 2/12<br>NAIVE | 2/13 | 2/14 | 2/15<br>NAIVE<br>ARIMA | 2/16<br>LM | 2/17<br>SNAIVE | 2/18 |
| 2/19 | 2/20<br>NAIVE | 2/21<br>NN | | | | |

**Figure 7b:** Count of correct predictions by model

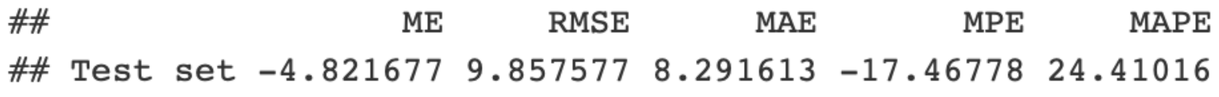
| Model | NAIVE | ARIMA | LM | SNAIVE | NN |
|---|---|---|---|---|---|
| # Correct | 8 | 7 | 3 | 3 | 3 |

**Figure 8:** Seasonal Naïve Forecast Accuracy



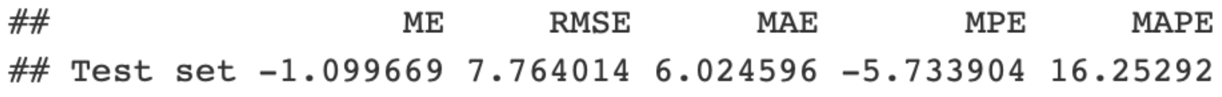
```
##                   ME     RMSE      MAE       MPE     MAPE
## Test set -5.783226 11.14069 9.127742 -19.78033 26.48817
```

**Figure 9:** Linear Regression Model Forecast Accuracy



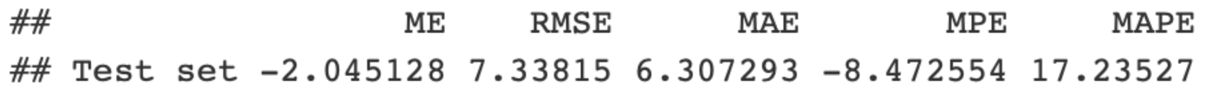
```
##                   ME     RMSE      MAE       MPE     MAPE
## Test set -4.821677 9.857577 8.291613 -17.46778 24.41016
```

**Figure 10:** ARIMA Model Forecast Accuracy



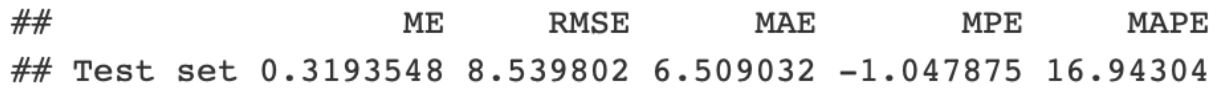
```
##                   ME     RMSE      MAE       MPE     MAPE
## Test set -1.099669 7.764014 6.024596 -5.733904 16.25292
```

**Figure 11:** NeuralNet Model Forecast Accuracy



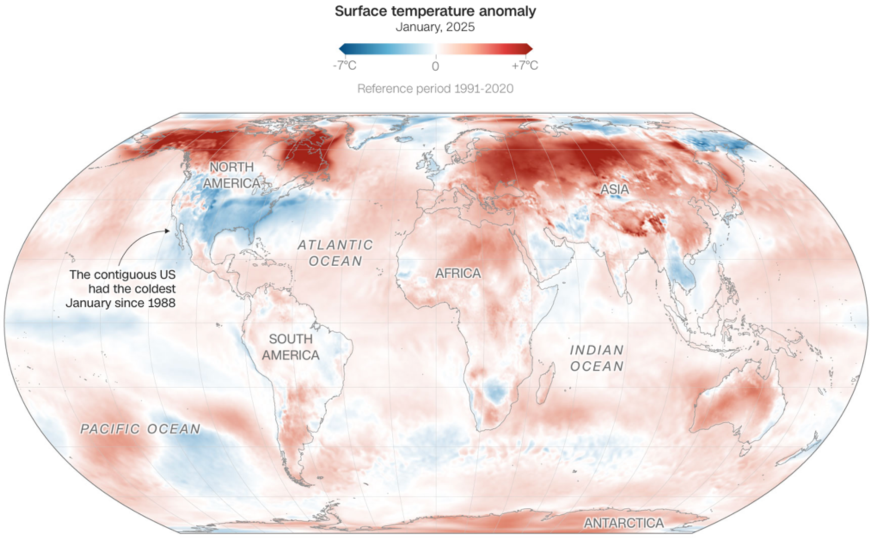
```
##                   ME    RMSE      MAE       MPE     MAPE
## Test set -2.045128 7.33815 6.307293 -8.472554 17.23527
```

**Figure 12:** Naïve Forecast Accuracy



```
##                   ME     RMSE      MAE       MPE     MAPE
## Test set 0.3193548 8.539802 6.509032 -1.047875 16.94304
```

**Figure 13:** Heatmap of temperature anomaly in January 2025, via cnn.com


Surface temperature anomaly
January, 2025
-7°C
0
+7°C
Reference period 1991-2020
NORTH AMERICA
ATLANTIC OCEAN
AFRICA
ASIA
SOUTH AMERICA
INDIAN OCEAN
PACIFIC OCEAN
ANTARCTICA
The contiguous US had the coldest January since 1988