Water Consumption Prediction with Machine Learning

A. Blackwell, K. Castillo, J. Huang, K. Zhao

ABSTRACT

Four time-series machine learning models were used on our data: exponential smoothing, ARIMA, SARIMA, and LSTM. The seasonality of the data was represented best by Holt-Winter's seasonal extension to exponential smoothing and the seasonality captured in SARIMA. The ARIMA model failed to determine a pattern and suffered from underfitting. The LSTM model had issues generalizing to the test set and resulted in poor performance on our data. The best fit model to our data was SARIMA, achieving the lowest median absolute error of all the considered machine learning models.

1 INTRODUCTION

The recent frequency of climate issues has risen, and with this topical issue is the concern of droughts. Every summer season in the New York area comes with the possibility of water sparseness. Often, water restrictions must be invoked, causing civil unrest and raising concerns over whether or not the constraints are of an appropriate scale. Our project forecasts water consumption so that civilians have more time to prepare for water restrictions. Additionally, our prediction software could provide insight regarding the most suitable magnitude of restrictions.

2 PRELIMINARIES

2.1 Time Series Machine Learning

The opportunity to apply time series machine learning presents itself when there exist discrete quantities occurring over time. Based on the previously observed values, a prediction can be made. Different time series machine learning models consider different aspects of the time series data but are able to provide a prediction. Forecasting a prediction at time t entails estimating a value at $t + h$ with only the information available at t.

Many of the most popular time series machine learning utilize supervised learning by continuously partitioning a data set into training and testing data. A prediction can be made by the model within a known time span. The prediction can then be tested against the ground truth yielding an error that can be minimized.

2.2 Autoregressive Models

An autoregressive model (or AR model) is a model commonly applied in the analysis of time-series data. An AR model is without latent variables: all variables are directly observed rather than inferred. In an AR model, each predicted element x_i in a data vector X is based on other elements in the same data vector X. An AR model of order k may predict a new data point x_t given $x_{t-1}, x_{t-2}, ..., x_{t-k}$. As shown by the formula, the order k gives the number of immediately preceding values to consider in determining the next number [3].

3 EXPLORATORY DATA ANALYSIS

Our dataset, shown in Figure 1, includes water consumption statistics in Hundred Cubic Feet (HCF) for the New York City boroughs:

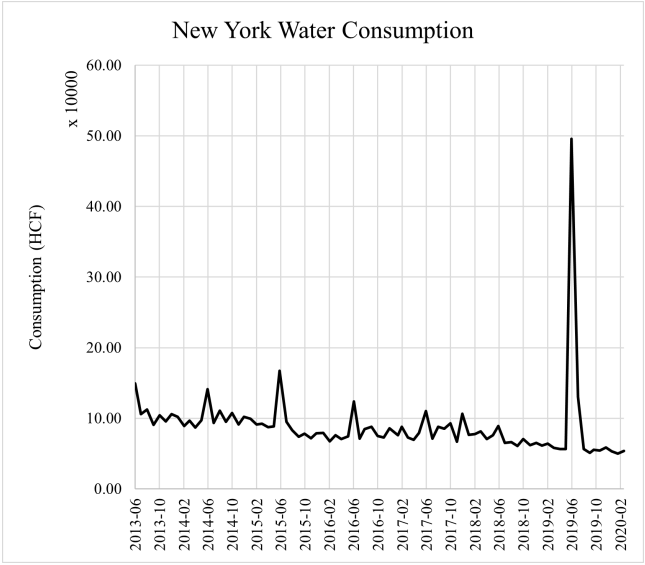

Figure 1: New York water consumption data over six years, eight months.

the Bronx, Brooklyn, Manhattan, Queens, Staten Island, and the Federal Housing Administration (FHA). Further, from the water consumption of the boroughs, New York City's water consumption is derived. The sequence of water consumption data ranges from 2013-06-23 to 2020-03-11.

A rough seasonal pattern can be observed. Summers in New York render a high water consumption before falling again during the winter. In July of 2019, an abnormally large spike in water consumption reveals itself in Manhattan. With further research, a leading explanation suggests the spike may be the result of distinguishing a fire that left 73,000 civilians without power in Manhattan [2]. The data in its entirety represents neither a positive nor negative overarching trend. From the observable patterns in the data, seasonality will play a role in the prediction of future dates.

4 APPLYING MACHINE LEARNING MODELS

Different machine learning models forecast values with varying degrees of accuracy. The models evaluated against the New York water consumption data are exponential smoothing, ARIMA, SARIMA, and LSTM. Exponential smoothing and ARIMA were selected as they are among the most widely used approaches to time series forecasting. SARIMA (Seasonal ARIMA) is an extension of ARIMA and adds a linear combination of seasonal past values or forecast errors. An LSTM model was additionally selected to be evaluated as it serves as a state-of-the-art model regarding time series machine learning.

4.1 Exponential Smoothing

The variant of exponential smoothing we used in the forecasting of our data was the Holt-Winters' seasonal method. The data is known to possess seasonal qualities. So, the inclusion of the seasonal component s_t and a seasonality frequency m allowed for a more accurate prediction. Exponential smoothing generally makes predictions using the linear weighted sum of the most recent observations; however, where exponential smoothing branches from ARIMA models is in the weights. The weights of exponential smoothing models decrease exponentially as observations are further in the past. Exponential smoothing models put an emphasis on more recent observations compared to old ones.

4.2 ARIMA and SARIMA

AutoRegressive Integrated Moving Average (ARIMA) is a form of regression analysis that considers the differences between values in a series. ARIMA is an autoregressive model, meaning it forecasts values corresponding to a linear combination of the variable's past values. The model is integrated as it may differ the raw observations to allow the time series to become stationary. That is: a stationary time series is void of trends and seasonality. The model additionally has a moving average which corresponds to a linear combination of past forecast errors.

ARIMA allows for three parameters to be tuned: p, d, and q [1]. The parameter p is the number of lag observations in the model to consider. The parameter d determines the number of times raw observations are differenced. The parameter q represents the size of the moving average window. Using grid search, optimal values of 2, 1, and 0 were found for p, d, and q respectively.

Seasonal AutoRegressive Integrated Moving Average (SARIMA) shares all the qualities ARIMA does; however, SARIMA adds a linear combination of seasonal past values or forecast errors [4].

The inclusion of seasonal consideration brings with it four additional parameters: P, D, Q, and s. The parameters P, D, and Q are seasonal equivalents of their corresponding lower-case parameters found in the ARIMA model. The fourth additional parameter s is the length of a cycle. The New York water consumption data has a seasonal cycle length (s) of 12 months—the cyclic pattern repeats every year. Using grid search, optimal values for P, D, and Q were found to be 0, 1, and 1 respectively.

4.3 LSTM

Long Short Term Memory Networks (LSTM Networks) extend recurrent neural networks (RNNs). RNNs differ from traditional feed-forward networks by how the input neurons take in data. In traditional networks, input neurons are set directly and propagated forwards through the network. Recurrent networks, however, feed their intermediate or final outputs back into their inputs. By effect, RNNs form an internal state or memory. A shortcoming of recurrent networks is their tendency to forget long-term knowledge. LSTM networks branch from RNNs and add functionality for recalling long-term patterns, as well as short-term ones.

To improve the performance of the LSTM network, hyperparameters including lookback, batch size, and LSTM units were hand-tuned. Lookback is the number of immediately previous observations considered by the network. The batch size corresponds to the number of samples trained for each step, and LSTM units are neurons. A neuron in an LSTM network is composed of a cell, and an input, an output, and a forget gate. The cell is responsible for remembering values over arbitrary time intervals and gates determine information retention.

With the lookback parameter tuned to 4, hidden units totalling 6, and training with 100 epochs, the LSTM network rendered its best performance. Increasing the number of units comes with a steep fall from the initial loss rate; however, after the steep fall, varying the number of neurons alone did not result in learning past the first 5 epochs. The loss curve plateaus after only a few iterations of the training data indicating overfitting. Increasing the look-back parameter resulted in a slower and more gradual descent in loss. To fit the model, 100 epochs were run on the LSTM network before a persisting plateau was observed.

5 EVALUATION

Table 1: Error Statistics on Test Data

Model	Mean AE	Median AE	RMS
Holts-Winter	31465.83	5352.43	96374.67
ARIMA	37282.08	10202.17	105706.43
SARIMA	34887.35	4535.55	103371.30
LSTM	65684.11	18488.60	129717.00

To compare different models, the same data was used, partitioning training and testing data equivalently. The time span of the training data ranged from the earliest observation, being in June of 2013, to October of 2018. The period of our training data represented approximately 80% of the total data, leaving the remaining 20%—November 2018 to February 2020—as testing data.

To determine what model performed the best, mean absolute error, median absolute error, and root-mean-square error (RMS error) were recorded and contrasted for each model Table 1. The median absolute error provides the greatest insight into the generalizing abilities of each model because the median is robust against outliers. Within the test data exists the abnormal spike in water consumption observed in June of 2019. The mean absolute error and the RMS error are greatly swayed by the outlier. Any predicted increase during the time of the spike results in a considerable reduction to the errors given including a mean representation of the average. The median absolute error, however, shares no such fluctuation to the outlier.

6 CONCLUSION

The Holts-Winter implementation of exponential smoothing achieved the lowest score in both mean absolute error and RMS. The SARIMA outperformed all other models in the remaining error calculation: median absolute error. ARIMA served as a baseline as it underfitted the training data and provided little more than a naive straight line forecast relaying the last observed value. The LSTM failed to generalize to the test set and performed the worst in all three error calculations.

While the Holts-Winter exponential smoothing model achieved the best performance in two of the three error assessments, the

SARIMA model is the best of the tested models. The anomaly in June of 2019 skewed the error calculations involving arithmetic means, because the median absolute error does not share this quality, it provides the most valuable insight to the best model. The SARIMA model was observed through the test data to be the best considered model, and as such, it was the model utilized in the future prediction of New York City water conservation from March 2020 to December 2025.

ACKNOWLEDGMENTS

This work was supported by the Borealis AI Undergraduate Mentorship Program *Let's SOLVE it*.

REFERENCES

[1] James Chen. 2021. Autoregressive Integrated Moving Average (ARIMA). *Investopedia* (April 2021). https://www.investopedia.com/terms/a/autoregressive-integrated-moving-average-arima.asp

[2] Mihir Zaveri James Barron. 2019. Power Restored to Manhattan's West Side After Major Blackout. *New York Times* (July 2019). https://www.nytimes.com/2019/07/13/nyregion/nyc-power-outage.html

[3] Peter Norvig Stuart Russell. 2021. *Artificial Intelligence: A Modern Approach* (4th ed.). Pearson. (book).

[4] Adith Yavegi. 2020. Time Series in Machine Learning. *Medium* (Nov. 2020). https://medium.com/analytics-vidhya/time-series-in-machine-learning-c3299742b2e1