It's Always Sunny in San Diego: Forecasting Power Consumption with Weather Patterns

Saba Alemayehu, Hunter Blum, Mackenzie Carter

San Diego University

ADS-506: Applied Time Series Analysis

Prof. Rick Sanchez

December 6, 2022

Abstract

Keywords- Energy consumption, Time Series Forecasting, Autoregressive Integrated Moving Average (ARIMA), Neural Networks, Root Mean Square Error (RMSE).

San Diego Gas and Electric (SDGE) plays an essential role in providing power to the city. Nearly half of SDGE's power is produced from renewable energy, with the rest coming from natural gas (SDGE, 2022). Because of this, it is essential to have highly accurate and timely power forecasts to reduce power waste and prevent blackouts. Combining weather predictors with multivariate time series models has proved highly effective for past daily data (Zhao and Magoulès, 2012). Our goal was to provide accurate hourly power consumption forecasts for SDGE utilizing local weather data. After exploratory analysis, we determined that *HourlyStationPressure* and *HourlyWetBulb* temperature had the strongest relationship with power consumption. We then created a naive forecast to establish a baseline RMSE of 149 MWh. Next, we constructed two non-weather and four weather-included models and evaluated the forecasts to the baseline. The only model that performed better than the baseline was the 2-hour moving average, which had an RMSE of 74 MWh. The neural network performed the best for models with weather predictors but had a poor RMSE of 422 MWh. The weather predictors likely failed to prove useful due to San Diego's stable climate causing low variance predictors. In the future, it could prove useful to utilize non-weather predictors, such as using cellular data as a proxy for human activity.

It's Always Sunny in San Diego Forecasting Power Consumption with Weather Patterns

Power plants are an integral part of every modern society. In San Diego, CA, San Diego Gas and Electric (SDGE, n.d.) is the main power provider, servicing 3.7 million citizens across 4,100 square miles. In 2021, SDGE generated 29.6% percent of its electricity from natural gas generators and another 23.9% from "unspecified sources," meaning power was purchased from outside sources (SDGE, 2022). However, most of the outsourced power likely comes from natural gas as well, with the top five sources all being natural gas plants ("Stefan Covic Testimony", 2021). Natural gas turbines must produce the power at the same time it is consumed, so having accurate control over the system is vital (Kim and Cho, 2019). Not producing enough power will result in blackouts. Generating too much power can also have negative effects. In 2021, 46.3% of SDGE power came from renewable or hydroelectric sources (SDGE, 2022). So, any excessive natural gas production will force cleaner sources to dump their electricity, often running it through resistors to dissipate the extra electricity as heat (Saffar and Ghasemi, 2021). Therefore, our study's main objective was to forecast hourly energy consumption in San Diego, allowing for more precise control of natural gas generators, making clean energy more efficient, and reducing the chances of a blackout.

One of the most promising predictors of energy consumption has been weather patterns. This is especially apparent in the United States, with the USA having some of the highest fluctuations in energy consumption due to weather (Fikru and Gautier, 2015). The weather has a particularly large impact on the energy used for HVAC systems, with one study finding a 6-10% decrease in heating and a 5-20% increase in cooling demand per 1.6°F increase in average temperature (Fikru and Gautier, 2015). In San Diego, temperatures below freezing rarely occur, however, days above 90°F are much more common (NOAA, 2021). This leads to San Diego County energy production spiking during the heat of late summer and from sporadic cold days in the winter (University of San Diego, 2022). Our study's secondary objective is to capture these fluctuations in energy usage through its relationship to weather variables, especially temperature.

In order to accurately forecast hourly energy consumption and capture fluctuations in usage, our study focuses on statistical modeling of these predictions in comparison to engineering models. Engineering models of energy consumption are often highly accurate but can be highly complex to perform, computationally expensive, and hard to understand by those not directly involved in the predictive modeling (Zhao and Magoulès, 2012). For our statistical modeling, time series forecasting methods were used, including naive forecasts, moving average forecasts, and various versions of autoregressive integrated moving average (ARIMA) and neural networks.

Exploratory Data Analysis

Data Collection

The first dataset focused on energy data and was sourced from the California Independent System Operator (CAISO, 2022). The data spanned from January 2019 to September 2022. The dataset contained hourly energy usage in megawatt hours (MWh) for the four energy companies in CAISO's area. However, this project only focused on the San Diego Gas and Electric data. The original data contained separate date and hour features which were combined into a single date-time feature.

The second dataset focused on hourly weather data and was sourced from the National Oceanic and Atmospheric Administration (NCEI, 2021). This dataset spanned from January

2013 to November 4, 2022. In the dataset, there are features for monthly, daily, and hourly averages, including multiple measures based on temperature, precipitation, pressure, wind, and sky conditions. Since the CAISO data is hourly, we kept the hourly weather features. Hourly variables were collected at the 54th minute of each hour. We rounded this time down to provide a one-hour and six-minute difference between the energy and weather data. This window should allow ample time for SDGE to take action on the forecasts.

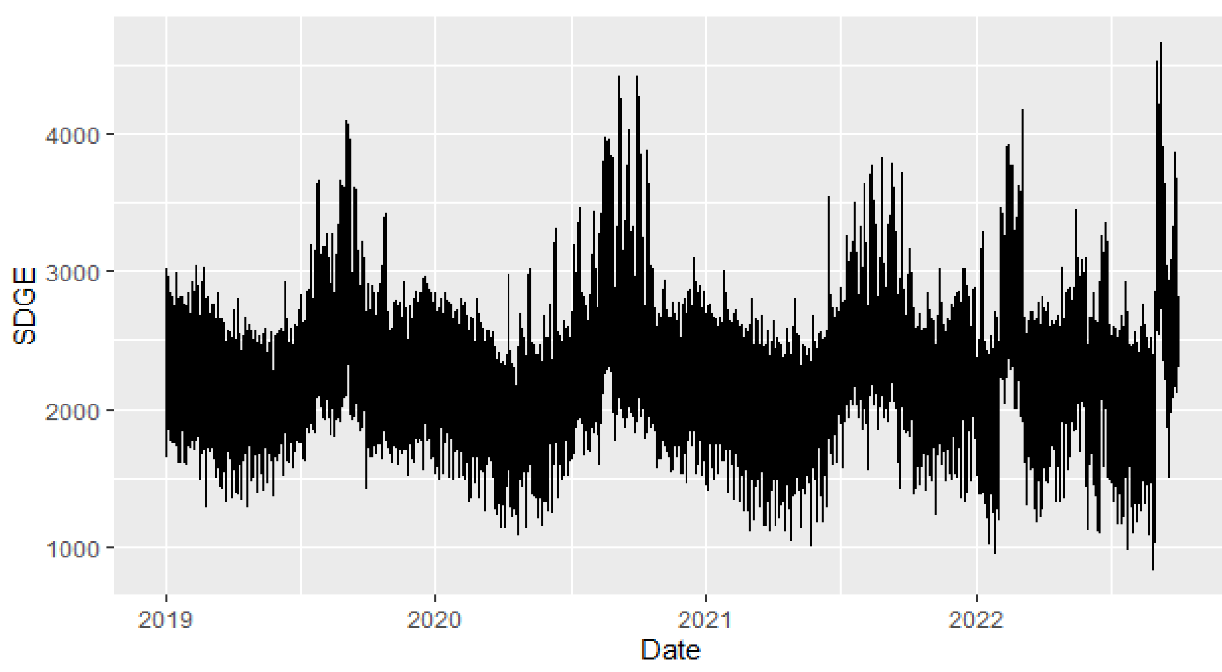
The NOAA dataset was left joined to the CAISO data, leaving us with a final dataset consisting of 39,485 observations and 18 features. Since we joined based on the CAISO data, our final dataset spans from January 1, 2019, to September 30, 2022.

Target Feature

The target variable is the San Diego Gas and Electric power consumption in MWh (SDGE). As shown in Figure 1.7, *SDGE* showed yearly seasonality with a polynomial trend in each season. The yearly seasonality indicates that *SDGE* has regular and predictable changes that occur every calendar year.

Figure 1.7

Time Series for SDGE



Numeric Features

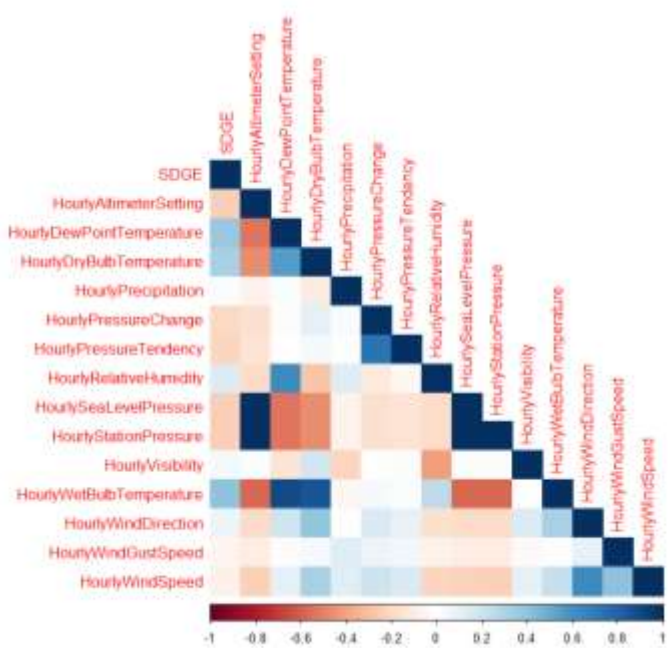
We divided the data frame into numeric and categorical categories for the purpose of feature selection. We had 15 numeric features including the target, SDGE, as well as various hourly pressure, temperature, wind, precipitation, and visibility variables.

Outliers

To identify outliers, we used the *grubbs.tes*t function from the *outliers* package (Komsta, 2022). The Grubbs test showed our max value for wind speed, 33, was an outlier. This makes sense because the wind speed typically ranges in single digits, so this value was likely a typo or a major storm. If wind speed is utilized in a model, we will replace this outlier with the last known value.

Figure 1.2

Correlation Between Numerical Features



Correlation

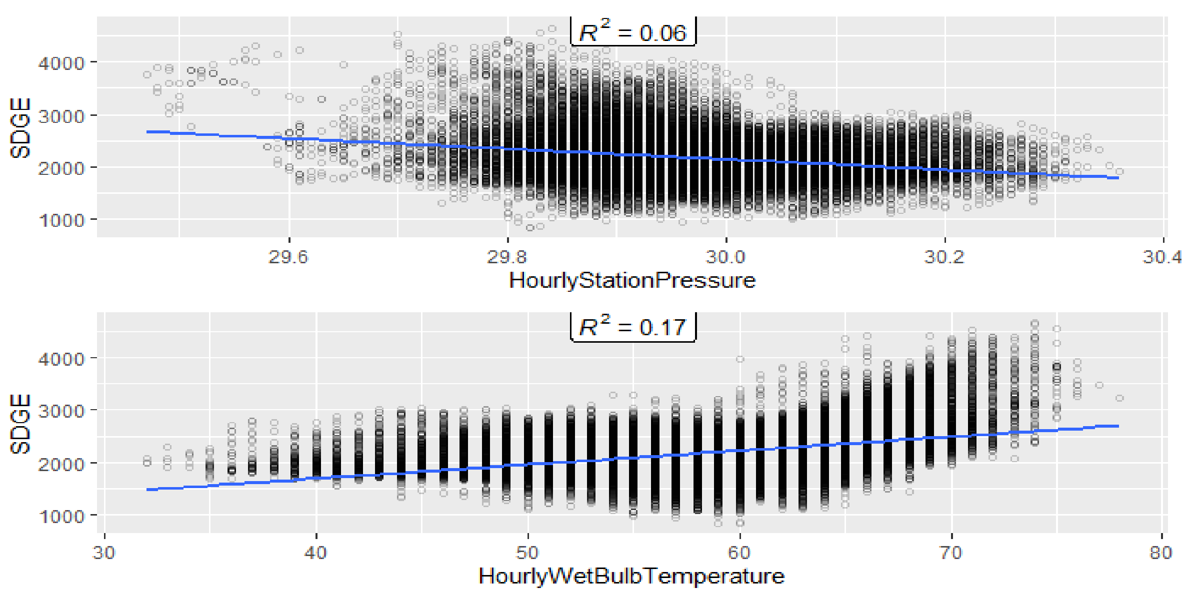
As shown in Figure 1.2, *HourlyStationPressure*, *HourlyAltimeterSetting*, & *HourlySeaLevelPressure* were 100% correlated. HourlyDryBulb, HourlyWetBulb, and HourlyDewPointTemperature were closely correlated as well. For the correlated temperature features, we kept the wet bulb temperature as that is analogous to what humans feel. We kept the *HourlyStationPressure* for the correlated pressure features since it had the most understandable name.

Scatter plots

We created scatter plots for each numeric predictor versus SDGE (Figure 1.5 in appendix). The only predictors that exhibited a coefficient of determination above 0.06 were *HourlyStationPressure and HourlyWetBulbTemperature*. *HourlyStationPressure* had a very small negative linear relationship with SDGE. Whereas the *HourlyWetBulbTemperature* and SDGE exhibited a slight positive linear relationship (Figure 1.6).

Figure 1.6

Scatter plot for Selected Predictors vs. SDGE



Distributions

Figure 1.13

Histogram for Numerical Variables

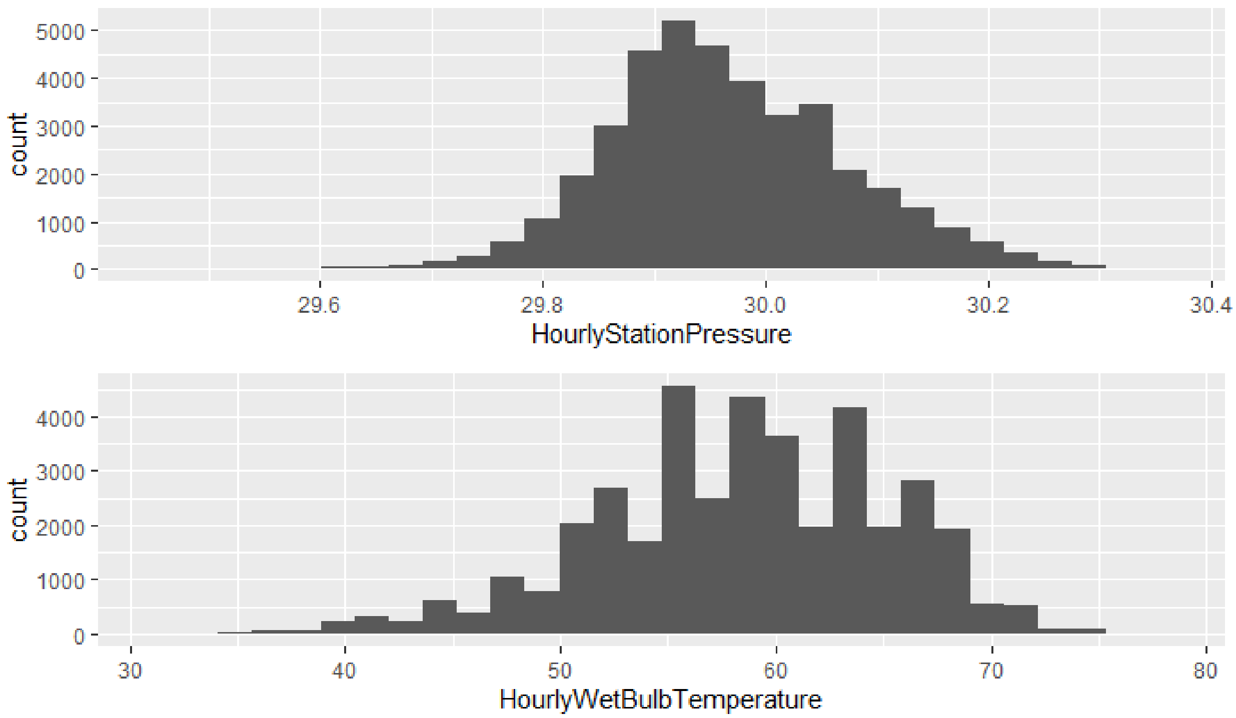


Figure 1.13 shows the distribution of our two highest correlated predictors. *HourlyStationPressure* was relatively normal and spread from 29.6 to 30.3 inches Hg. *HourlyWetBulbTemperature* was slightly skewed to the right and spread from 35 to 75°F. Neither distribution had evidence of large outliers.

Categorical Features

The categorical variables were *HourlyPresentWeatherType*, and *HourlySkyConditions*. To assess their effects on SDGE, an ANOVA was performed on each predictor, comparing their categories to the target. Each ANOVA returned a significant p-value (nearly zero), however, each predictor had so many levels that these p-values were likely hyper-deflated to zero. Because of these many levels, these predictors were excluded from modeling to reduce the dimensions of our data and prevent overfitting.

Data Cleaning and Preparation

Imputation

The variables *Date*, and *SDGE* had no missing values. However, most of the variables in the datasets had 304 missing values, except for the features *HourlyWindGustSpeed* (37640), *HourlyPressureTendency* (28670), *HourlyPressureChange* (28670), and *HourlySeaLevelPressure* (6956). The missing values in wind gusts were associated with low wind speeds, so each missing value was set to zero. The remaining missing values were filled, prior to EDA, by the last observation carried forward imputation with the *na.locf* function from the *zoo* package (Zeileis and Grothendieck, 2005).

Splitting the Dataset

We split our data into a training and validation period, with September 1, 2022, as the splitting point. Our final training dataset had 38,632 observations while the validation dataset had 852 observations.

Modeling

Naive

Naive forecasts were generated to establish a baseline model. The forecasts were made by taking the last hour’s energy consumption using the *lag* function in *dplyr* (Wickham, 2019). The naive forecasts had a root mean square error (RMSE) of 148.92 MWh on our validation dataset. We evaluated the success of each proceeding model based on the RMSE of these naive forecasts.

Moving Average

Next, we created trailing moving average forecasts, evaluating window sizes from 1 to 24 hours and 2 to 7 days. Each window size, k was run through *rollmean* function in *zoo* (Zeileis and Grothendieck, 2005). We then calculated each k's RMSE on the training set and selected the top performing window size, 2 hours, as our final moving average model.

Exponential Smoothing

We produced Holt-Winter's exponential forecasts utilizing the *ets* function in *forecast* (Hyndman et al., 2022). Our time series for the target, SDGE, exhibited additive seasonality with no clear trend, so we selected "ZNA" for our modeling parameter (Figure 1.10). Our final model consisted of a multiplicative error, no trend, and additive seasonality and had an alpha (learning rate) of 0.9972. Although an alpha so close to one may be concerning, it makes sense in this scenario. Higher alphas lead to our model giving more weight to more recent values. Since we have such high-frequency data, our forecasts will likely perform the best with these more recent observations, which we have also seen in the moving-average forecasts.

ARIMA

Manual ARIMA

An ARIMA model was used to predict the hourly power consumption of San Diego for the next 30 days. We selected the *HourlyWetBulbTemperature* and *HourlyStationPressure* as predictors for the model since they had the highest R^2 with SDGE (Figure 1.6) These predictors will be utilized in the remainder of our models. The data were examined for non-stationarity through ACF and PACF plots. The ACF and PACF plot revealed that the data had p = 3 (the number of autoregressive terms), d = 2 (the number of times the series was differenced), and q = 2 (the number of moving average components). After trial and error of different manual ARIMA models, the model with the lowest AIC (453028.3) was selected. The final model had normally distributed residuals with no clear pattern over time. However, the ACF plot of the residuals had large significant spikes, especially at seasonal intervals. So, we did not fully capture all seasonal components with this model.

Automatic

We trained an ARIMA model with the *auto.arima* function in the *forecast* package (Hyndman et al., 2022). Because we also observed additive seasonality in SDGE and in the residuals of the manual ARIMA, we set the seasonal option to true in hopes to account for it. The final model had p = 2, d = 1, and q = 3. The residuals exhibited no clear patterns and were normally distributed. However, the ACF plot of the residuals still showed significant spikes, especially at the 24-hour mark. Our Ljung-Box test also returned a p-value of zero, so there is very strong evidence that we did not capture all of the seasonal components in the data.

Manual Seasonal ARIMA

In another attempt to capture the seasonality in our ARIMA, we tuned a manual ARIMA with seasonal components. To remove the daily seasonal component, we took the difference at the 24th lag. A one-lag difference was also taken, to account for potential trends in the data. In the ACF and PACF for this differenced time series, we saw strong autoregressive and moving average terms at the first lag. So, our first manual seasonal model used one for p, d, and q in both the seasonal and non-seasonal components. Our residuals were normally distributed, however, we still had large spikes in the residuals ACF and PACF plots, especially at the seasonal level. We further tuned this model by increasing the seasonal and non-seasonal p’s and q’s. Much like the last manual ARIMA, we used a process of trial and error to find the lowest AIC and best diagnostic plots. Our best manual ARIMA had p = 3, d = 1, and q = 3 for the non-seasonal parameters and P = 2, D = 1, and Q = 2 for the seasonal parameters. The residuals for this model were also normally distributed and the ACF/PACF plots had slight significance after the fifth lag. However, we still had very significant seasonal dependence in the ACF and PACF plots for the residuals. So our ARIMA model still struggled to capture some seasonal components, even after deseasonalizing the daily pattern.

Neural Network

Following the ARIMA models, neural networks were explored as another model to predict energy consumption based on weather activity. Neural networks base their function on the function of the human brain. Neurons are processing units, connected with many other units, with an input, activation function, and outputs. These networks create multiple 'layers' through which an output is produced (Camara et al., 2016). Neural networks are used often in time series forecasting in a variety of fields such as financial forecasting. However, with little input completed by the data scientists, neural networks are often considered a 'black box' model. Neural networks can also be resource intensive (Shmueli and Lichtendahl, 2018). Despite these factors, neural networks are one of the most used models in time series forecasting due to their ability to find patterns in complex data sets, and their consistent accuracy over other statistical models (Chakraborty et al., 1992).

Average Neural Networks

Before moving to autoregressions, generic neural networks were briefly explored to get a general understanding of how a neural network model would perform compared to previously modeled ARIMA and other forecasting methods. The Averaged Neural Networks (*avNNet*) function was utilized via the R *caret* package (Kuhn, 2022). Continuous high RMSE values determined that the inclusion of predictors unique to the data would be required. Due to this, the RMSE values are not included in Figures 3.1 & 3.2 as final models were not created to allow focus on feed-forward neural network autoregressions.

Neural Network Autoregression

To try to improve accuracy over generic predictor neural networks, lagged values of the time series can be used, as well as the addition of external predictor values. This method is known as neural network autoregression and was modeled through the *nnetar* function via the R *forecast* package (Hyndman et al., 2022). The model is denoted as NNAR(p,P,size), with p referring to non-seasonal lags, P being seasonal lags, and size being the number of nodes within the hidden layer ((p+P+1)/2) (Shmueli and Lichtendahl, 2018). As before, a process of trial and error was used to find the best performance. A p=3, P=2, and a calculated size of 3 had the best performance, with normally distributed residuals.

A key to the success of this model was the inclusion of external regressors. We continued to use *HourlyWetBulbTemperature* and *HourlyStationPressure* for this, which are passed to the model through the *xreg* argument. This improved RMSE values from 791 MWh to our final RMSE value, 431 MWh, a nearly 75% improvement.

Results

Our exponential smoothing model was the only model that had a worse training RMSE than our baseline naive forecasts. This RMSE is essentially equal to the baseline, only being 0.78 MWh higher than the naive forecasts (Figure 3.1). However, the results for the validation data tell a different story. The only model that performed better than the baseline on the validation data was the 2-hour moving average forecast, with an RMSE of 74.46 MWh compared to 148.92 MWh for the baseline. The best model with weather features was the neural network, with an RMSE of 430.8 MWh, while the worst was the non-seasonal manual ARIMA with an RMSE of 1,022.23 MWh (Figure 3.2).

Figure 3.1

Model Results for Training Data

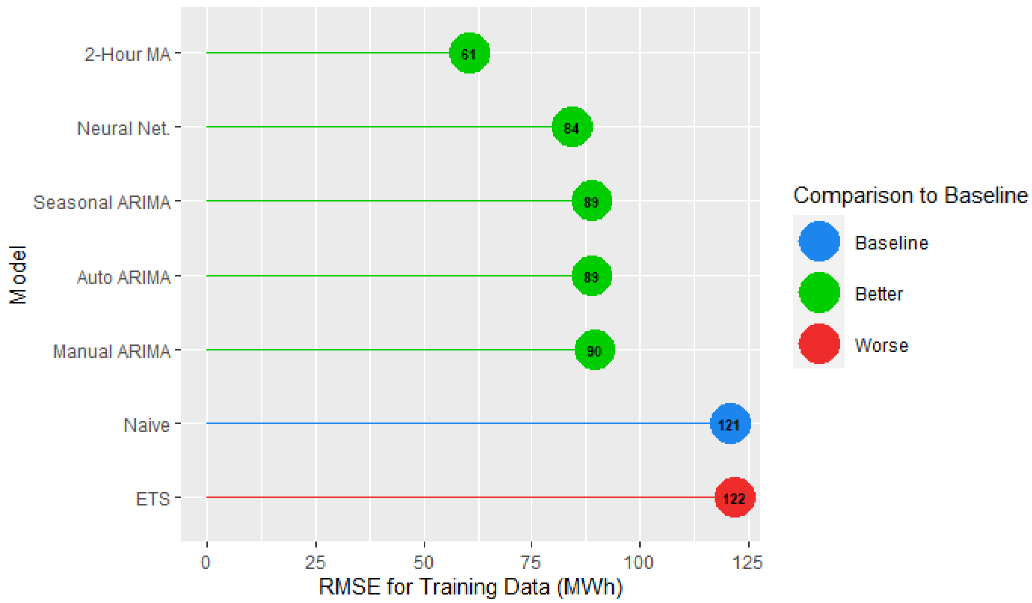
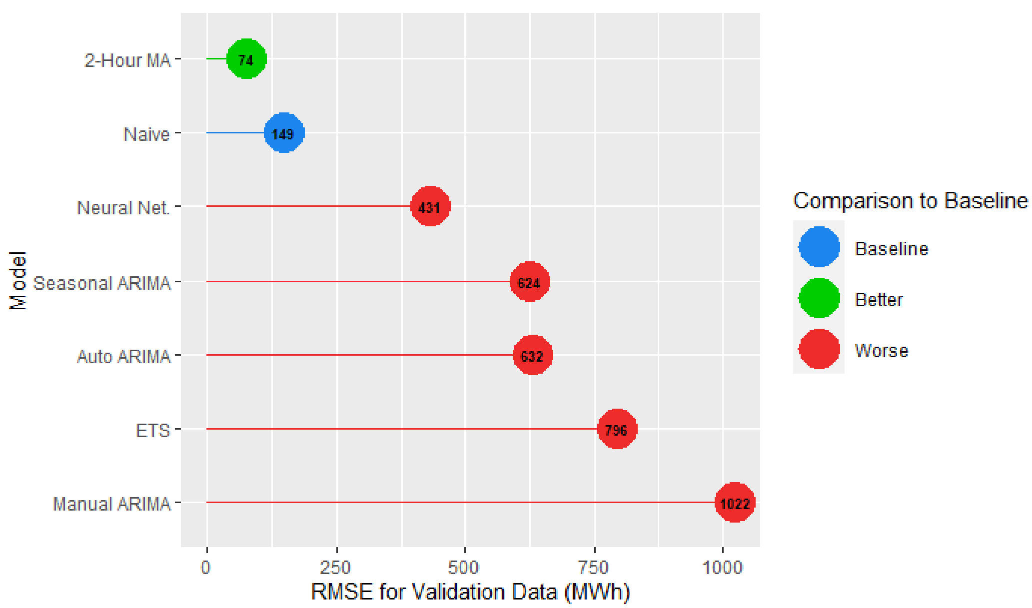


Figure 3.2

Model Results for Validation Data



Discussion

Power production and consumption are an ongoing concern in the city of San Diego. San Diego's millions of citizens rely on accurate consumption predictions to ensure enough power production to avoid loss of power, while simultaneously avoiding overproduction. Statistical modeling can be used to make these predictions, without the challenges of engineering models. This study primarily compared two multivariate models, ARIMA and neural networks to predict energy consumption in San Diego based on previous usage patterns and external weather conditions. These methodologies were applied to a merged dataset of energy consumption data and weather condition data, sourced from the California Independent System Operator (CAISO) and National Oceanic and Atmospheric Administration (NOAA).

Although it does not include weather features, we would recommend the 2-hour moving average forecasts for SDGE. The moving average model had less than half of the RMSE on the validation data than our baseline naive forecasts (Figure 3.2). Based on RMSE, we observed that neural networks outperform ARIMA when including weather conditions in energy consumption prediction. Despite this improved performance, neural networks still gave high prediction errors, calling for further evaluation to determine if weather conditions such as wet bulb temperature and atmospheric pressure are useful for a city like San Diego. Despite some of the warmest weather conditions across the United States, there is little variability in this data, which may explain why adding in weather conditions to our time series forecasting caused a stark decrease in accuracy.

Future Directions

In the future, we would first like to explore different predictors, particularly the categorical variables. These variables were excluded due to their high number of levels, however many of these levels can likely be manually combined. Since weather predictors did not display promise for San Diego, we would also want to explore other predictors. A recent study by Bogomolov et al. (2016), utilized anonymous cellular network data to build random forests predicting daily energy consumption. This strategy could prove even more useful, as it would allow for SDGE to hourly adjust each of its microgrids, rather than our current strategy which generalizes to the entire system. However, Bogomolov evaluated a large province in Italy with daily data. So, evaluating hourly data in the densely populated San Diego may not be more effective than the current moving average forecasts.

References

Bogomolov, A., Lepri, B., Larcher, R., Antonelli, F., Pianesi, F., & Pentland, A. (2016). Energy consumption prediction using people dynamics derived from cellular network data. *EPJ Data Science*, *5*, 1-15.

CAISO. (2022, November 11). Reliability requirements. California ISO - Reliability Requirements. Retrieved December 4, 2022, from http://www.caiso.com/planning/Pages/ReliabilityRequirements/Default.aspx

Camara, A., Feixing, W., & Xiuqin, L. (2016). Energy Consumption Forecasting Using Seasonal ARIMA with Artificial Neural Networks Models. International Journal of Business and Management, 11:5, 1833-8119.

Chakraborty, K., Mehrotra, K., Mohan, C. K., & Ranka, S. (1992). Forecasting the behavior of multivariate time series using Neural Networks. *Neural Networks*, *5*(6), 961–970. https://doi.org/10.1016/s0893-6080(05)80092-9.

Fikru, M. G., & Gautier, L. (2015). The impact of weather variation on energy consumption in residential houses. Applied Energy, 144, 19-30.

Hyndman R, Athanasopoulos G, Bergmeir C, Caceres G, Chhay L, O'Hara-Wild M, Petropoulos F, Razbash S, Wang E, Yasmeen F (2022). _forecast: Forecasting functions for time series and linear models_. R package version 8.18, <URL: https://pkg.robjhyndman.com/forecast/>.

Kim, T. Y., & Cho, S. B. (2019). Predicting residential energy consumption using CNN-LSTM neural networks. Energy, 182, 72-81.

Komsta, L. (2022). outliers: Tests for Outliers. R package version 0.15. https://CRAN.R-project.org/package=outliers

Kuhn M (2022). _caret: Classification and Regression Training_. R package version 6.0-93, <https://CRAN.R-project.org/package=caret>.

National Centers for Environmental Information (NCEI), N. O. A. A. (2021). Data tools: Local climatological data (LCD). Local Climatological Data (LCD) | Data Tools | Climate Data Online (CDO) | National Climatic Data Center (NCDC). Retrieved December 4, 2022, from https://www.ncei.noaa.gov/cdo-web/datatools/lcd

Prepared Direct Testimony of Stefan Covic on Behalf of San Diego Gas and Electric Company. Public Utilities Commission of the State of California. (2021) (testimony of Stefan Covic). https://www.sdge.com/sites/default/files/regulatory/PUBLIC%20Covic%202022%20Testimony_Redacted.pdf

Saffar, A., & Ghasemi, A. (2021). Energy management of a renewable-based isolated micro-grid by optimal utilization of dump loads and plug-in electric vehicles. Journal of Energy Storage, 39, 102643.

SDGE. (n.d.). About Us. About Us | San Diego Gas & Electric. Retrieved November 6, 2022, from https://www.sdge.com/more-information/our-company/about-us

SDGE. (2022). 2021 Power Content Label. San Diego, CA; SDGE.

Shmueli, G. & Lichtendahl Jr., K.C. (2018). Practical time series forecasting with R: A hands-on guide (2nd ed.).

University of San Diego. (2022). Residential Energy Use. Residential Energy Use - School of Leadership and Education Sciences - University of San Diego. Retrieved December 4, 2022, from https://www.sandiego.edu/soles/hub-nonprofit/initiatives/dashboard/electricity.php#:~:text=per%20household%20(kWh)-,End%20of%20interactive%20chart.,5%25%20to%2011%25%20increase.

US Department of Commerce, N. O. A. A. (2021, April 14). San Diego climate. National Weather Service. Retrieved November 14, 2022, from https://www.weather.gov/sgx/san-san

Wickham et al., (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686, https://doi.org/10.21105/joss.01686

Zhao, H., & Magoulès, F. (2012). A review on the prediction of building energy consumption. Renewable and Sustainable Energy Reviews, 16, 3586–3592.

Zeileis, A. and Grothendieck, G. (2005). zoo: S3 Infrastructure for Regular and Irregular Time Series. Journal of Statistical Software, 14(6), 1-27. doi:10.18637/jss.v014.i06