Predicting electricity demand for the PG&E territory through multiple linear regression

Maya Bruguera

Abstract

Developing accurate short-term electricity demand forecasts is critical to power planning and ensuring adequate power supply, and is the subject of ongoing research. This project investigates whether electricity demand in Northern California can be accurately predicted through a multiple linear regression model, using historical electricity demand data, temperature, and temporal characteristics (i.e., day of week) as inputs to forecast electricity demand. In order to account for real-world data constraints facing electric utilities, the project also evaluates changes in model error when the most recent demand data available are from the previous week. The project uses electricity demand data for the Pacific Gas & Electric (PG&E) territory and temperature data retrieved from the National Oceanic and Atmospheric Administration (NOAA) Global Historical Climatology Network (GHCN) weather stations. The project applies the Scikit-learn Python library and its linear model module to fit five multiple linear regression models. When evaluated on the testing dataset, the models produced an RMSE of 264 MWh – 798 MWh (2.2% – 6.9%), and multiple R-squared values of 0.92 – 0.98. The model that included all three feature categories (demand, temperature, and temporal) and included demand features from the previous hour and previous day performed best.

1. Introduction

This project aims to determine whether a multiple linear regression model can be built that accurately predicts electricity demand for the Pacific Gas & Electric (PG&E) service territory using temperature, historical electricity demand data, and temporal characteristics (e.g., month, day of week, hour of day). To account for real-world data constraints impacting utilities, the project also aims to consider how model accuracy changes demand data from earlier in a given day and from the previous day are unavailable, and the most recent demand data available are from the previous week.

Short-term electricity demand forecasts are essential to enabling electric utilities to balance electricity supply and demand, particularly with the expansion of variable renewable generation resources (Gajowniczek & Zabkowski, 2017; Zhang et al., 2019). The accuracy of short-term electricity demand forecasts has substantial financial implications for electric utilities; one study found that a mere 1% increase in short-term demand forecast error can increase annual operating costs by $10 million (Zhang et al., 2019). As a result, researchers and practitioners continue to work to develop and improve the accuracy of short-term electricity demand forecast models.

Various methods have been developed to forecast electricity demand in the short-term, most commonly employing multiple linear regression and time series models (e.g., autoregressive moving average models) (Ahmed et al., 2018; Lee & Tong, 2011). Multiple regression models typically offer improved accuracy relative to time series models, as times series models exclusively rely on demand data, while linear regression models incorporate additional features (e.g., weather data) (Singh et al., 2019; Zhang et al., 2019). More recently, researchers have begun to develop methods based in artificial intelligence (e.g., artificial neural network, support vector regression) (Bedi & Toshniwal, 2019; Heydari, 2020; Singh et al., 2019). These models are advantageous as they can more effectively fit non-linear input variables (Singh et al., 2019); however, they are beyond the scope of the Data 100/200 course.

Selecting appropriate variables is key to developing accurate short-term demand forecasts (Zhang et al., 2019). Models typically apply meteorological variables (namely temperature), historical electricity demand (e.g., from the previous hour or previous day), and temporal variables (e.g., time of day or day of week) (Zhang et al., 2019).

This project aims to complement existing research by contributing a model for short-term electricity demand forecasting using temperature, historical electric demand data, and temporal variables specifically for the Pacific Gas & Electric (PG&E) service territory, which is located in Northern California.

2. Methodology

2.1 Data Collection

This project uses three main datasets, one of which provides global temperature data, while the other contains electricity demand data for the PG&E territory, while the last contains the spatial boundaries of the PG&E territory. The temperature data were retrieved from the National Oceanic and Atmospheric Administration (NOAA) Global Historical Climatology Network (GHCN), and reflect average daily temperature measured by GHCN weather stations across the globe from January to October 2020 (NOAA, n.d.); this dataset was retrieved, processed, and made available by the Data 100/200 course staff. The demand data were retrieved from the U.S. Energy Information Administration (EIA), and reflect hourly electricity demand in megawatt-hours (MWh) for the PG&E service territory from July 2018 to April 2021 (EIA, 2021). A geospatial shapefile of California electric utility service areas was used to identify the PG&E service territory boundaries, and was retrieved from the California State Geoportal (CEC, 2020).

2.2 Data Pre-processing & Exploratory Data Analysis

After retrieving the datasets, processing was done to align them spatially and temporally. First, the global temperature data were processed to create a data frame of temperature within the PG&E territory. To accomplish this, the temperature data were read in as a data frame, and were grouped by station name to create a new data frame of unique weather stations. The shapefile of the electric utility territories was read in to the Jupyter Notebook as a geopandas GeoDataFrame, and the PG&E service area was extracted as a new GeoDataFrame. In order to visualize GHCN weather stations in or nearby PG&E's territory, then the minimum and maximum latitude and longitude of the PG&E territory were extracted and used to apply a Boolean mask to the weather station data frame, creating a GeoDataFrame of weather stations in or nearby the territory (left plot in Figure 1).

Based on this plot, to create a dataset of weather stations relatively evenly distributed across the territory, and based on the high concentration of weather stations outside of the territory in the Sierra region, a new data frame of only stations within the PG&E territory was created. To accomplish this, the GeoDataFrame of stations in or nearby the territory was spatially joined with the PG&E GeoDataFrame, and a Boolean mask was applied, forming a GeoDataFrame of only weather stations within the PG&E territory. These weather station names were then used to create a Boolean mask for the original temperature data, creating a data frame of all temperature data from weather stations that lie within the PG&E territory (middle plot in Figure 1). [1] The dataset was searched for null values; none were identified.

To explore the temperature data for the territory, a scatterplot was created of each temperature reading by month (right plot of Figure 1). Based on this plot and guidance from Data 100 staff, the temperature data were transformed from the original tens of degrees Celsius to Celsius by dividing by 10. Based on the plot, it also appears that all of the temperature data seem to be plausible readings. However, potential outliers appeared to exist, particularly in the month of February and in the month of June. As a result, a second data frame was created with outliers removed by applying the interquartile range (IQR) method. The IQR method was applied separately for each month in order to account for seasonal temperature trends (ensuring removal of low outliers from warmer months, and high outliers from cooler months).

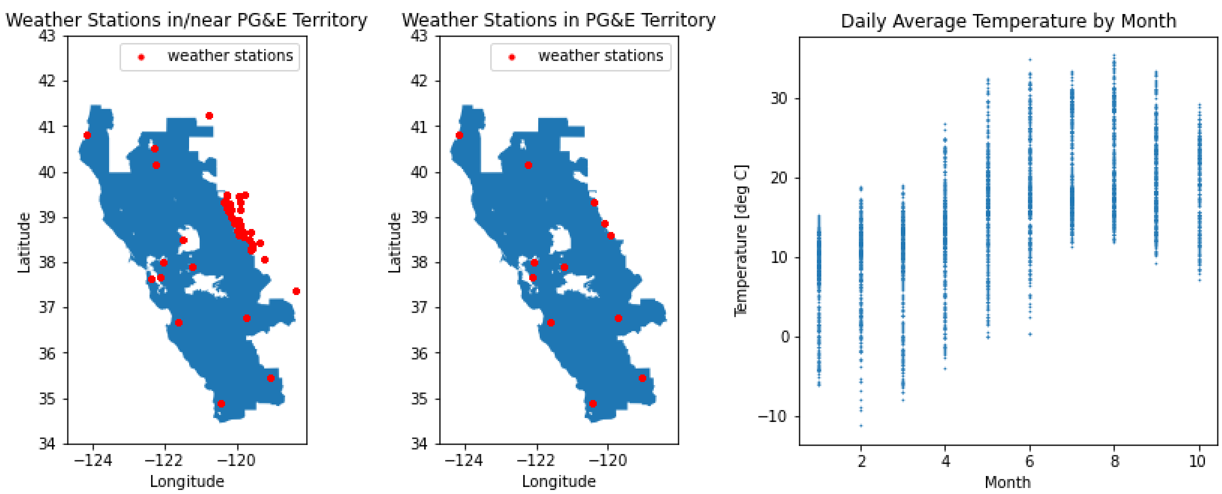

Figure 1. Left: PG&E territory (blue) and weather stations (red) in or nearby the territory. Middle: PG&E territory (blue) and weather stations (red) in the territory. Right: Daily average temperature grouped by month for all weather stations within the PG&E territory, for all days from January 1, 2020 to October 22, 2020.

In order to create a temperature metric for the entire PG&E territory at each time step (i.e., each day), the data frame was grouped by the datetime reading, and aggregated by the mean, minimum, and maximum of the weather station average daily temperature values. To prepare the dataset for the regression model, columns were added for the previous day's mean, minimum, and maximum temperature value, and for the day before that.

To allow for temporal analysis, in both the temperature and electricity demand data frames, the date column was converted to datetime, localized to Pacific Time, then the localtime method was applied to extract year, month, week of year, day of week, hour, and day of year. These columns were also used to create a column indicating whether the day was a weekday or weekend.

To explore the PG&E demand data, hourly electricity demand from 2020 was plotted by weekend and weekday (Figure 2), and by each day of the week (Figure 3), with line color indicating the month. As shown in Figure 2 and Figure 3, demand is typically higher during

---

[1] Initially, the full temperature data were directly spatially joined with the PG&E GeoDataFrame. However, this was very computationally expensive, therefore the process was modified to only join unique weather stations, then identify temperature entries associated with PG&E based on the corresponding weather station.

warmer months, and lower during cooler ones. Additionally, demand curves vary by day of week, with weekends experiencing lower morning peaks relative to weekends.

The PG&E demand data were then processed to prepare them for the machine learning model. Columns were added for hourly demand from the previous hour, from the same hour during the previous day, and from the same hour and same day of the week during the previous week.

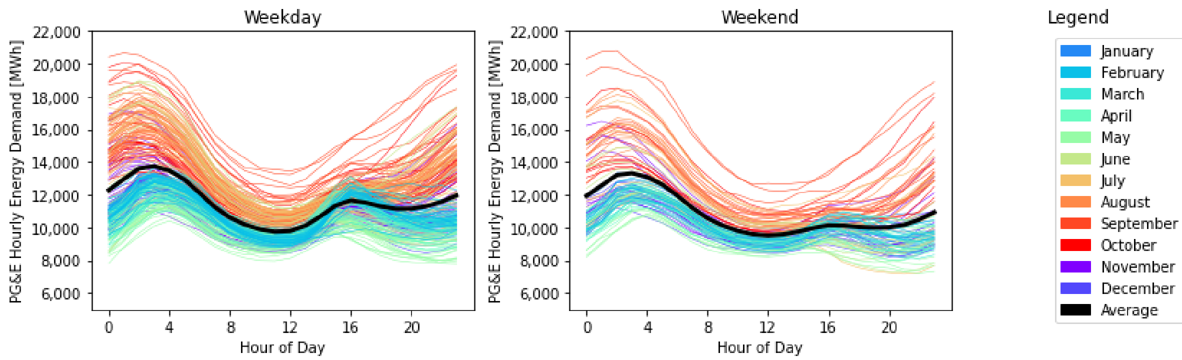

Figure 2. PG&E total hourly energy demand profile by weekday and weekend, for each day of the year in 2020.

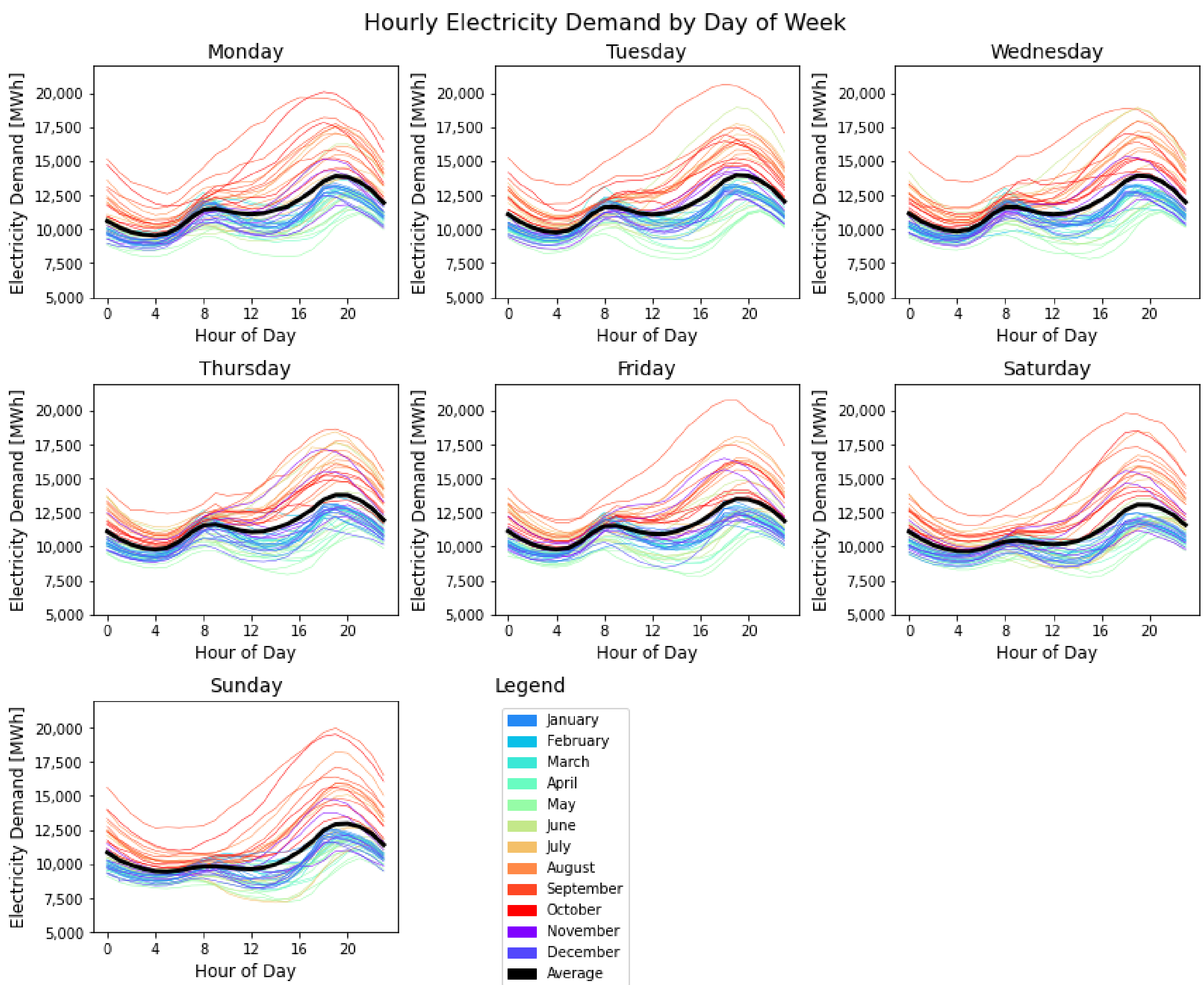

Figure 3. PG&E total hourly energy demand by day of the week, for each day of the year in 2020.

Based on these graphs, all data appeared to be plausible demand values, therefore no data were removed. The dataset was also searched for null values, and none were identified. Based the graphs, outliers appeared to exist, therefore a data frame was created without outliers. As with the temperature data, outliers were calculated using the IQR method separately for each month.

The temperature and demand data were then combined into a single data frame for the machine learning model. Because the temperature data were available over a shorter timeframe (1/1/2020 to 10/22/2020) relative to the electricity demand data (7/1/2018 to 4/25/2021), a Boolean mask was applied to create a demand data frame with a timeframe matching that of the temperature data. The temperature and demand data frames were then merged based on their date, retaining all hourly demand values and adding columns for the associated temperature values for the same day. Because in both the temperature and demand data, though outliers existed, all values appeared to be plausible, and predicting demand is particularly important during abnormally high-demand days in order to prevent outages, one combined data frame was produced using all data points and one was produced excluding days considered outliers either due to the temperature or the demand data.

2.3 Multiple Linear Regression Modeling & Feature Engineering

Multiple linear regression was selected for this analysis as it is a model designed to predict an output based on a range of input features, making it an appropriate choice for predicting hourly electricity demand based on temperature, historical demand data, and day of week. Furthermore, as described in the introduction, it is a commonly used model for predicting electricity demand. For this model, the data frame of combined temperature and electricity demand data was split into training (80%) and testing (20%) data frames.

The potential features—variables that might predict electricity demand—can be broken down into three categories: historical demand, temperature, and temporal. These were selected due to their known relationship to demand, based on the literature review. Potential quantitative features (i.e., historical demand and temperature variables) were plotted against electricity demand using the training data. These plots were used to determine whether the two appeared to have a relationship, and whether that relationship was linear. These plots indicated that hourly demand has a linear relationship to hourly demand from the previous hour, hourly demand for the same hour of the previous day, and hourly demand for the same hour and day of the previous week. Non-quantitative features (i.e., the temporal variables such as hour-of-day) had already been plotted against demand during the exploratory data analysis process, and therefore were not re-plotted. Both quantitative and qualitative features considered and the rationale for why they were considered (i.e., how they relate to demand) are listed in Table 2.

The plots indicated that hourly demand has non-linear relationships to various temperature variables, including the mean, minimum, and max of the weather station readings for that day's average temperature, and the mean, minimum, and max of the weather station readings for the previous day's average temperature. The shape of each of these plots roughly approximated that of the lower right quadrant of the Tukey-Mosteller Bulge Diagram. Intuitively, this makes sense, as temperature primarily induces electricity demand by increasing air conditioning use when temperatures are very hot, therefore at low temperatures one would expect minimal relationship with demand, while once temperatures begin increasing beyond temperatures comfortable to humans, one would expect to see a positive relationship between temperature and demand.

As a result, transformed plots were created to identify transformations that might linearize the relationships between hourly demand and these variables. Based on the Tukey-Mosteller Diagram, the $x^2, x^3, \log Y,$ and $\sqrt{Y}$ transformations were applied to the graphs of hourly demand versus each temperature variable. For all of the variables, the x-squared and x-cubed transformations appeared to most effectively linearize the relationship with hourly demand. As a result, within the train and test data frames, columns were added for cubed versions of four of these variables (same day and previous day mean and maximum of average daily temperature across the weather stations) and for squared versions of two of the variables (same day and previous day minimum of average daily temperature across the weather stations).

Based on the exploration of relationship between hourly demand and the demand, temperature, and temporal features, five linear regression models were designed. This first three models consisted of variations of feature groups, to investigate how the different variable types impacted model accuracy. The first model included demand and temperature variables as features, the second included demand and temporal variables, and the third included all three (demand, temperature, and temporal variables). A second iteration of the third model was run, but used the data that excluded outliers. As mentioned earlier, it is helpful for utilities to be able to accurately predict electricity demand even on abnormally hot or high-demand days, therefore most model iterations were run including outliers. This model run was included to consider how model accuracy might change when only considering non-outlier days. Two additional models were run which were also based on the third model, but excluded the previous hour demand feature, and the previous hour and previous day demand features, respectively. These were run in order to consider how model accuracy changes when the model lacks same-day or previous-day demand information, which is often the case in utility operations (Zhang et al., 2019). Table 1 summarizes the feature groups included in each model, and whether outliers were included or excluded; Table 2 lists the specific features within each of these groups.

Table 1. Linear regression models created, including feature groups included (marked in ***dark green****) and whether outliers were included or excluded.*

Model	Feature Groups Included			Includes Outliers?
	Demand	Temperature	Temporal	
Model 1				Yes
Model 2				Yes
Model 3				Yes
Model 3 w/o Outliers				No
Model 3 w/o Same-Day Demand data				Yes
Model 3 w/o Same- or Previous-Day Demand data				Yes

For each model, cross-validation was performed in order to refine feature selection to prevent overfitting, and optimize the model's complexity and the bias-variance trade-off. A function was defined that splits the training data into four subsets, and iteratively selects three subsets as the training set to fit the model, then uses the last subset as the validation set to calculate the model's root mean square error (RMSE), then takes the average of the validation RMSE values from the four iterations. This function was applied repeatedly using the first N features in the model, where N ranged from 1 to the total number of features. The number of features that produced the lowest RMSE was selected for the final model.

For each model, the featured identified through cross-validation were used to create design matrices for the training and test datasets, while the hourly demand data were used to create the true y-value vectors. For each model, Table 2 lists the features included in the final cross-validated model and those included in the initial model but removed due to cross-validation.

Table 2. Features included in linear regression model, rationale for consideration, and transformation applied (if any). Model inclusion columns is highlighted in ***dark green*** *if feature was included in initial and cross-validated model;* ***light green*** *if included in initial model but removed based on cross-validation;* ***green*** *if included in initial model but partially removed (for one-hot encoded values) based on cross-validation; grey if excluded from initial (and cross-validated) model.*

Feature Group	Feature	Reason Considered	Transfor-mation	Included in Models:					
				1	2	3	3.2	4.1	4.2
Demand	Previous hour's demand	Demand from previous hour closely predicts demand for the next hour (Zhang et al., 2019)	N/A						
	Demand from same hour, previous day	Demand data from the previous day is often used to help predict demand, particularly when same-day data are unavailable (Zhang et al., 2019)	N/A						
	Demand from same hour of the same day of the previous week	When same- and previous-day demand is unavailable, demand data from the previous week is often used to help predict demand (Zhang et al., 2019)	N/A						
Temperature	Mean of the weather station readings for that day's average temperature (T_{avg})	It is well documented in literature that temperature is closely related to electricity demand, as high temperatures typically drive increased use of air conditioning units, which are highly energy-intensive. Both same-day temperature and temperature from previous days (e.g., during a heatwave) can influence cooling-driven electricity demand.	Cubed						
	Maximum of the weather station readings for that day's T_{avg}		Cubed						
	Minimum of the weather station readings for that day's T_{avg}		Squared						
	Mean of the weather station readings for previous day's T_{avg}		Cubed						
	Maximum of the weather station readings for previous day's T_{avg}		Cubed						
	Minimum of the weather station readings for previous day's T_{avg}		Squared						
Temporal	Hour of day one-hot encoded	It is well documented in the literature that demand profiles differ by hour-of-week (Zhang et al., 2019; Figure 3).	N/A						
	Day-of-week one-hot encoded	It is well documented in the literature that demand profiles differ by day-of-week (Zhang et al., 2019; Figure 3).	N/A						
	Month one-hot encoded	It is well documented in the literature that demand profiles differ by month (Zhang et al., 2019; Figure 3)	N/A						

For each model, the Scikit-learn linear model module was used to fit a linear model to the training design matrix and true y-value vector, then used to predict hourly demand using the training and testing design matrices. Root mean squared error (RMSE) and R-squared values were calculated and used to evaluate the models.

Notably, linear regression is a probabilistic prediction model, which predicts outcomes assuming that we continue to draw from the same population. Meanwhile, a causal inference model predicts outcomes assuming that variables are made to take on particular values. It is important to keep in mind that the models developed in this report, therefore, does not demonstrate causality (i.e., a cause-and-effect relationship) between any of the features in the design matrix and hourly demand.

3. Results

The modeling results indicate that of the first three models, Model 3—which included demand, temperature, and temporal variables—performed the best, producing an RMSE of 264 MWh (2.3% of mean hourly demand) and a relatively high R-squared value of 0.98. Model 2 (composed of demand and temporal variables) performed nearly as well as Model 1 (Model 2 RMSE was just 2% higher than Model 1 RMSE, and had the same R-squared value of 0.98). Meanwhile, Model 1 performed the worst of the first three models (Model 3 RMSE was 65% higher than that of Model 1, and had an R-squared value of 0.97). This indicates that the combination of the demand and temporal variables selected are more powerful demand predictors than the combination of demand and temperature variables selected. However, as expected, combining all three feature categories results in the best performance.

Using a dataset that excluded outliers improved Model 3's accuracy even further, reducing RMSE by 4%. Conversely, removing recent demand data, including same-day demand data and previous-day demand data both drastically worsened model performance. Removing previous hour demand as a feature more than doubled RMSE relative to Model 3, leaving it at nearly 5% of mean hourly demand. Similarly, removing both previous hour demand and previous day same hour demand more than tripled RMSE, leaving it at nearly 7% of mean hourly demand. That said, all three of these models exhibited relatively high predictive power, with R-squared values of 0.92–0.95.

The metrics used to evaluate each model are summarized in Table 3. Figure 4 visually depicts model predictive power and error, with the points from models with higher performance more closely clustered around the $x = y$ line.

Table 3. Metrics used to evaluate models, including RMSE, RMSE contextualized as a percentage of mean hourly demand, R-squared, and the percent difference between a given model's RMSE and the Model 3 RMSE.

Model	RMSE [MWh]	RMSE Percent of Mean Hourly Demand	R-Squared	% Diff. between RMSE & Model 3 RMSE
Model 1 (Demand & Temperature)	436	3.8%	0.97	+65%
Model 2 (Demand & Temporal)	270	2.3%	0.98	+2%
Model 3 (Demand, Temperature, & Temporal)	264	2.3%	0.98	0%
Model 3 w/o Outliers	253	2.2%	0.92	-4%

Model	RMSE [MWh]	RMSE Percent of Mean Hourly Demand	R-Squared	% Diff. between RMSE & Model 3 RMSE
Model 3 w/o Same-Day Demand data	548	4.8%	0.95	+108%
Model 3 w/o Same- or Previous-Day Demand data	798	6.9%	0.92	+202%

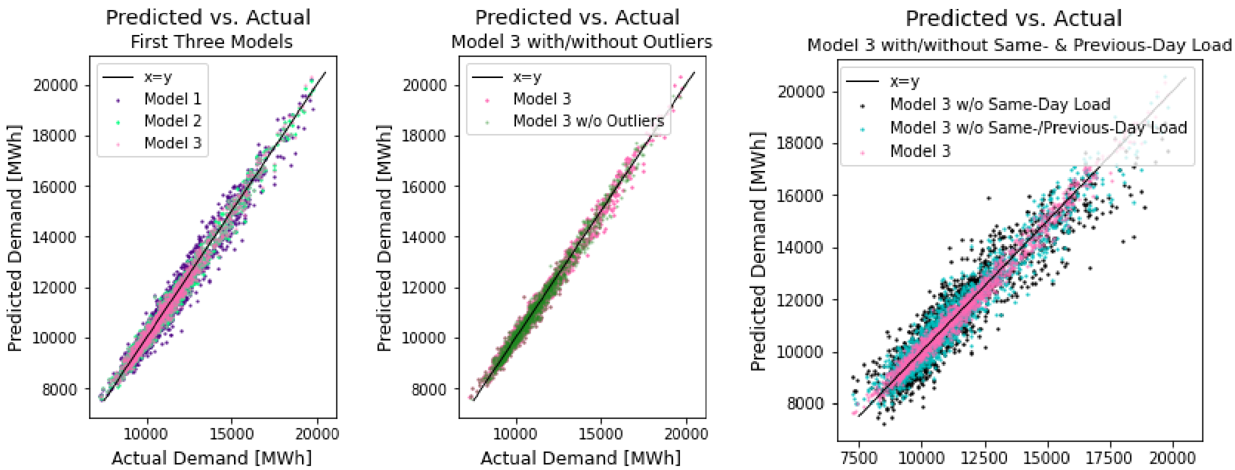

Figure 4. Predicted vs. actual hourly electricity demand for the PG&E Territory for Models 1, 2, & 3 (left), Model 3 with and without outliers (center), and Model 3 with and without same-day demand and without same- and previous-day demand (right).

4. Discussion

The evaluation metrics resulting from the linear regression models indicate that hourly electricity demand in the PG&E service territory can be relatively accurately predicted through a multiple linear regression model, using temperature, historical electricity demand, and day of week data. All five models were able to explain the vast majority of the variance in hourly electricity demand in the PG&E territory (R-squared of 0.92 – 0.98), with relatively low error (RMSE of 2.2% – 6.9% of mean hourly load). However, the model that included all three categories of variables (demand, temperature, and temporal) and included same-day and previous-day demand data (the baseline Model 3) performed the best.

The improved performance when removing outliers highlights the difficulty in predicting demand on anomalously low- and high-demand days. This creates a challenge for utilities, as high-demand days are those when the greatest supply is needed, and underpredicting demand on these days can lead to a shortage of electricity supply, and potentially force demand shedding (i.e., blackouts).

Similarly, the substantial increase in error when recent historical demand data—namely same-day and previous-day data—are excluded from the model demonstrate the utility of recent historical demand data in short-term demand forecasting, and the heightened error in predicting demand when these data are unavailable.

While the models hold relatively high predictive power, there is opportunity to take additional steps to address limitations and improve the model. One limitation of this study is that the temperature features used in the model only reflect the average of temperature measurements at locations with GCHN weather stations. To address this, temperature data could be collected from a set of weather stations chosen to reflect the range of climate zones throughout the territory. Additionally, the mean daily temperature values could be recalculated such that they weight each weather station's temperature value to better reflect the spatial distribution of PG&E customers and their demand. Alternatively, the territory mean temperature features could be replaced with unique temperature features representing readings from each weather station within the territory.

The study is also limited by the use of a single year—2020—to train and test the model. Given that the COVID-19 pandemic has altered electricity demand patterns (Agdas & Barooah 2020) the model could be made more generalizable to other years by incorporating training and test data from additional years. Furthermore, the model is trained on data from the PG&E territory, and could be made more generalizable to demand for other utilities by incorporating data from other utility territories.

Furthermore, additional features that are or may be related to hourly demand data could be incorporated. For example, data on the minimum and maximum daily temperature[2] (rather than the average daily temperature) and heatwave indicators (e.g., consecutive days where temperatures exceed 90°F) could be retrieved and incorporated; Li et al. (2015) found that incorporating high temperature improved accuracy of electricity demand forecast models. Additional relevant meteorological variables typical of short-term demand forecasts could also be integrated, such as humidity, hours of sunshine, cloud cover, and wind speed (Zhang et al., 2019). Including behind-the-meter solar generation (i.e., electricity generated by residential solar installations which can serve customers directly and reduce the demand that the utility sees) has also been found to improve demand forecasts (Zhang et al. 2019). Additionally, Dahl et al. (2018) found that encoding holidays can strengthen short term demand forecasts.

The model could also be modified to operate with different electricity demand features. For instance, while this (and other) models use hour-before and day-before demand data as an input to forecast demand, in practice, these data are often unavailable to utilities (Zhang et al. 2019). The model could instead be trained to forecast hourly demand based on the previous week's hourly demand profiles and/or additional features to better reflect the data that utilities typically do and do not have at their disposal. Along the same lines, the model could be fitted to forecast demand for smaller geographic regions, as utilities must not only consider the balance of total electricity supply and demand, but must also consider the demand and generation resources in particular regions in order to ensure that ensure that power can be transmitted to where it is needed while accounting for infrastructure and power flow constraints (e.g., losses, frequency regulation, voltage drops, substation capacities, etc.).

In terms of ethical concerns, no adverse impacts from this demand forecasting have been identified. Accurate load forecasting can in fact help improve likelihood of adequate electricity generation supply, and prevent utilities from procuring insufficient supply and adversely impacting the public through blackouts.

Improved electricity demand forecasts can also help utilities avoid unexpected electricity supply shortages, and the excessive costs that can arise from these situations. More importantly,

---

[2] Note that this is not to be confused with the minimum and maximum of the set of weather station average daily temperature values, which was available and incorporated into the model

improved forecasts have the potential to prevent power outages and improve the reliability of electricity supply, an essential public service.

References

Agdas, D., & Barooah, P. (2020). Impact of the COVID-19 pandemic on the US electricity demand and supply: an early view from data. IEEE Access, 8, 151523-151534.

Ahmed, T., Vu, D. H., Muttaqi, K. M., & Agalgaonkar, A. P. (2018). Load forecasting under changing climatic conditions for the city of Sydney, Australia. Energy, 142, 911-919.

Bedi, J., & Toshniwal, D. (2019). Deep learning framework to forecast electricity demand. Applied energy, 238, 1312-1326.

California Energy Commission (CEC). (2020). California Electric Utility Service Areas map available on the California State Geoportal. Accessed April 25, 2021. https://gis.data.ca.gov/datasets/b95ca182aa254c3db8ad4d92bd32a73c_0?geometry=-161.087%2C31.071%2C-77.459%2C43.276

Dahl, M., Brun, A., Kirsebom, O. S., & Andresen, G. B. (2018). Improving short-term heat load forecasts with calendar and holiday data. Energies, 11(7), 1678.Gajowniczek, K., & Ząbkowski, T. (2017). Two-stage electricity demand modeling using machine learning algorithms. Energies, 10(10), 1547.

Energy Information Administration (EIA). (2021). Demand for California Independent System Operator (CISO), Pacific Gas and Electric, hourly - local time. EIA Open Data API Query Brower. Accessed April 25, 2021. https://www.eia.gov/opendata/qb.php?category=3390291&sdid=EBA.CISO-PGAE.D.HL

Heydari, A., Nezhad, M. M., Pirshayan, E., Garcia, D. A., Keynia, F., & De Santoli, L. (2020). Short-term electricity price and load forecasting in isolated power grids based on composite neural network and gravitational search optimization algorithm. Applied Energy, 277, 115503.

Lee, Y. S., & Tong, L. I. (2011). Forecasting time series using a methodology based on autoregressive integrated moving average and genetic programming. Knowledge-Based Systems, 24(1), 66-72.

Li, Y., Bao, Y. Q., Yang, B., Chen, C., & Ruan, W. (2015). Modification method to deal with the accumulation effects for summer daily electric load forecasting. International Journal of Electrical Power & Energy Systems, 73, 913-918.

National Oceanic and Atmospheric Administration (NOAA). (n.d.). Datasets: Global Historical Climatology Network (GHCN). https://www.ncdc.noaa.gov/data-access/land-based-station-data/land-based-datasets/global-historical-climatology-network-ghcn

Singh, P., Dwivedi, P., & Kant, V. (2019). A hybrid method based on neural network and improved environmental adaptation method using Controlled Gaussian Mutation with real parameter for short-term load forecasting. Energy, 174, 460-477.

Zhang, N., Li, Z., Zou, X., & Quiring, S. M. (2019). Comparison of three short-term load forecast models in Southern California. Energy, 189, 116358.