Predicting Residential Energy Usage Based on Meteorological Data

Siena Moca, Amber Chau, Justin Wong, Juhyun Lee, Kamsey Agu

I. Abstract

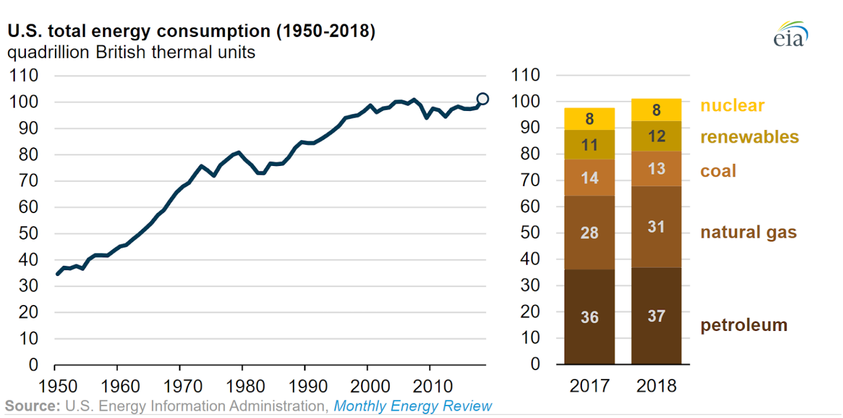
Global climate change continues to increase the demand of energy consumption over time as temperatures and weather patterns are more extreme. Increasing demands have led to the need for increasing energy capacity and distribution for customers. In this study, we developed a machine learning model to predict residential energy consumption in the Bay Area geographic zip codes based on meteorological data, such as temperature, humidity, near infrared, and weather. Historic residential energy consumption data from PG&E and NREL National Solar Radiation data served as our validation data. We sought to develop a generalizable model so that such inputs can be applied to other geographic locations outside the Bay Area. The KNN model performed the best based on the evaluation of mean squared error (MSE) values. Such a model serves as key to ensuring efficient operations of residential energy systems as climate change takes its course.

II. Introduction

Motivation & Background

It is commonly accepted that energy consumption will increase in the future as a result of climate change. Reasons driving this trend include economic development, technological developments, and extreme climate (increase and decrease in temperatures) and weather events.

Energy consumption in the United States has been continuously increasing, as seen in Figure 1. According to the U.S EIA (United States Energy Information Administration), the amount of energy use in the U.S. recorded the historical peak. This growth of the amount of energy consumption is causing a vicious circle: the increase of energy consumption causes climate change which again leads to greater demand for energy because the majority of energy sources are still from fossil fuels generating carbon dioxides. Statistics show that 80% of the energy consumed in the U.S is from fossil fuels and only 20% is from non-fossil fuels such as wind, biofuels, and other renewables (Figure 1). It is apparent that energy consumption will continue to grow in the future unless the vicious circle breaks.



Source: U.S. Energy Information Administration, *Monthly Energy Review*

Figure 1. U.S Total energy consumption and energy portfolio (U.S. EIA 2019)

Forecasting energy consumption increase is vital in determining whether or not electricity grids can sustain increased demands and prevent large outages and how it is going to impact climate change. With a better forecast, utilities and stakeholders can plan ahead and invest in improving resiliency and environmental sustainability of their energy system. Recently in Texas, a massive snowstorm has caused power outages not only due to the damage to utility systems but also the very high demand for households to heat their homes as a result of extreme cold weather (Ferman 2021).

Challenges associated with managing this particular energy system includes planning ahead for unexpected climate change events such as wildfires, snowstorms, etc. We also do not know what kind and impact new technologies will help mitigate climate change and may produce a different trend regarding energy consumption.

Relevant Literature

Multiple studies have been conducted in attempting to predict energy consumption through using past energy consumption sets with machine learning methods but have used various input variables such as building characteristics and climate data.

Williams et. al used statistical learning methods including linear regression, regression trees (RT) and multivariate adaptive regression splines (MARS) to predict future monthly energy consumption in residential homes in Bexar County, TX. The authors inputted monthly energy consumption for 36 months and historical temperature and humidity values from the past 30 years. While the MARS model had significantly better performance in predicting future monthly consumption, there were still limitations since there was uncertainty about climate data and regression trees were better for aggregate predictions. Different fuel types were also unknown about space heating.

Zhao and Magoules in addition to regression models, included Artificial Neural Networks (ANNs) and Support Vector Machines or Regression to evaluate the significance of weather conditions determining building energy usage such as temperature, humidity, solar radiation and wind speed. ANNs prediction was useful for solving nonlinear problems and had the highest prediction accuracy when combined with simplified engineering methods.

In New Zealand, Salcedo et. al also used Support Vector Regression (SVR) but added on separately Multi-layer Perceptron (MLP) to predict monthly mean air temperature based on historical monthly mean air temperature in Australia and New Zealand. The researchers found that the SVR algorithm performed the best out of all the other approaches and concluded that Machine Learning methods are appropriate for energy consumption prediction.

Lastly, Wang et. al analyzed models and previous work relating to building energy prediction based on the principle of algorithm integration. They reviewed previous prediction models, such as Random Forest, Gradient Boosting, auto-regressive integrated and moving average (ARIMA) and k-nearest neighbors (KNN). The authors further developed integration prediction models by comparing using "integration learning", which essentially combines the advantages of single models to improve overall performance. The core idea of this “stacking model” was to collect the differentiation of various base algorithms (each of which can observe data from different perspectives) by constructing an integrated framework. The paper sought robust methods to reduce overfitting and evaluate model performance based on accuracy, generalization, and robustness. Their stacking model was more accurate than any of the individual base models.

Focus of This Study

This study used machine learning and modeling methods to predict the scale of energy consumption increase based on past climate data. For this study, we specifically looked at residential energy usage in kilowatt-hours (kWh). We predicted the scale of energy consumption applicable to the Bay Area, California which can be used to plan increasing energy capacity across utilities, storage, etc.

III. Technical Description

Data Sources

The PG&E dataset contained customer electric (kWh) usage data in the Bay Area. The data is reported by zip code, month, year and four customer types including residential, commercial, agricultural and industrial. The reports from the dataset are publicly available with each report containing 3 months of usage data through the end of the calendar quarter and are provided in CSV file format. For this study, electric usage data was extracted for 2013 to 2020 from the PG&E public database website. Data cleaning was performed by selecting only residential customers and grouping the data by months. The parameters in the data include the zip code, total number of customers, and average and total electric usage in kilowatt-hour.

The NREL National Solar Radiation Database is a collection of half-hourly and hourly values of meteorological data with three common solar radiation measurements, which include global horizontal, direct normal, and diffuse horizontal irradiance (DHI). The dataset majorly covers the United States and certain international locations, and the data was collected at adequate temporal and spatial scales to represent regional solar radiation climates effectively. For this study, the solar radiation dataset was selected for the Bay Area and grouped into monthly data to match the PG&E dataset. The dataset was cleaned by extracting the direct normal irradiance (DNI), diffuse horizontal irradiance (DHI), global horizontal irradiance (GHI), solar zenith angle, temperature and wind speed. The DNI, DHI, and GHI represent the amount of solar radiation from the direction of the sun, the solar radiation on a horizontal surface received from the sky excluding the solar disk, and the solar radiation on a horizontal surface received from the sky, respectively, given in Watt per square meter (W/m^2). The solar zenith angle, temperature, and wind speed are given in degrees, degree Celsius ($^\circ C$), and meter per second (m/s). The temperature values were converted to degree Fahrenheit ($^\circ F$). After extracting these variables, the cleaned NREL dataset was merged with the PG&E dataset for analysis and prediction.

The NREL National Solar Radiation Database provides API instructions on how to download their data in Python. According to the instructions, specified API parameters, such as year, longitude and latitude, are passed via a query string for the URL. After defining all variables, a special url is now to be declared. This url is the query string that is used to fetch data from the NREL API, which returns csv data. Then the command “pd.read_csv(url, skiprows=2)” will load the data which consists of 46 columns including city, state, country, latitude, longitude, time zone, DHI, DNI, GHI, and Solar Zenith Angle.

Zip codes mapped to longitude and latitude coordinates were merged using opendatasoft public dataset. This was used so that the zip codes specified in the PG&E dataset corresponded to a coordinate pair for the NREL API. Since we were interested in solar irradiance data for the PG&E dataset, we first grouped the PG&E data by zip code, coordinate, and year; these served as parameters for the NREL API. At the time of this project, the NREL API was not updated to reflect 2020 solar radiation measurements. Furthermore, there were 6474 unique zip codes-year pairs in the entire PG&E data from 2013-2019. The resulting dataset consisted of 77,689 rows, where each row represented a zip code (with a coordinate pair) at a month and year (between 2013-2019) with additional columns from the PG&E measurement (in kWh) and NREL API (GHI, DHI, DNI, etc. as mentioned previously). We then filtered the PG&E dataset into a bounding box for Bay Area; the South West corner was (36.897966, -123.433313), while the North East corner was (38.593263, -121.381268). Below you can see the bounding box we created in Figure 2.

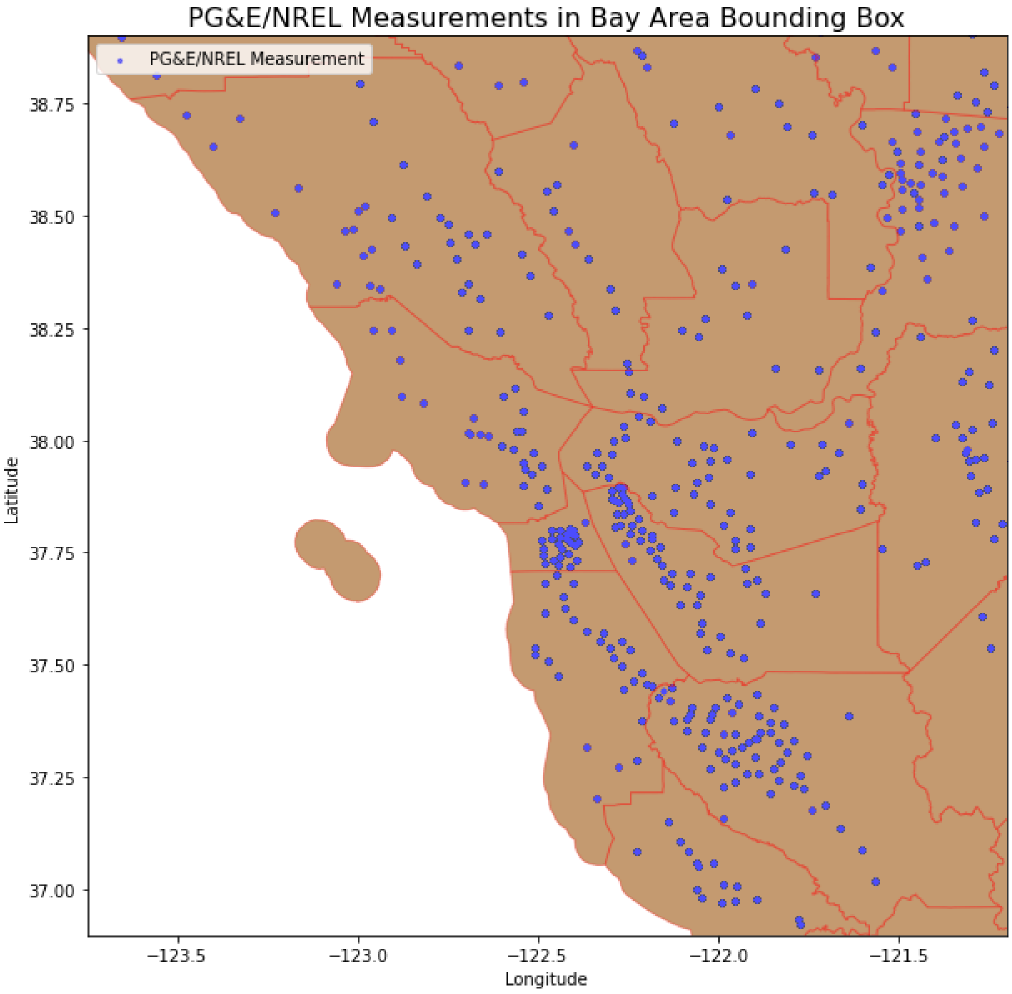


Figure 2. PG&E measurements within our bounding box.

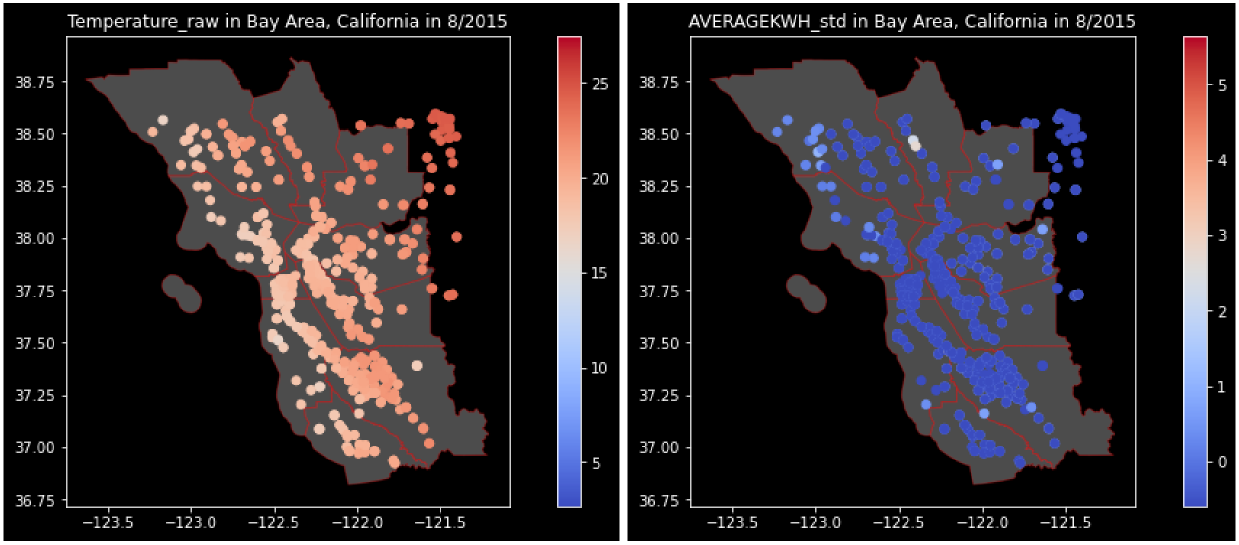


Figure 3. Plot of temperature and average kWh during August in 2015 in the Bay Area

In Figure 3, we plotted the temperature over the region of the Bay Area and the average kWh reported by PG&E for a zip code. We plotted the counties of the Bay Area using red lines on the graph and used a heat map style to plot the relative intensities of each of these measurements. We made similar graphs for each month from 2013 to 2020, in order to visualize the environmental metrics and related average kWh for each region over time.

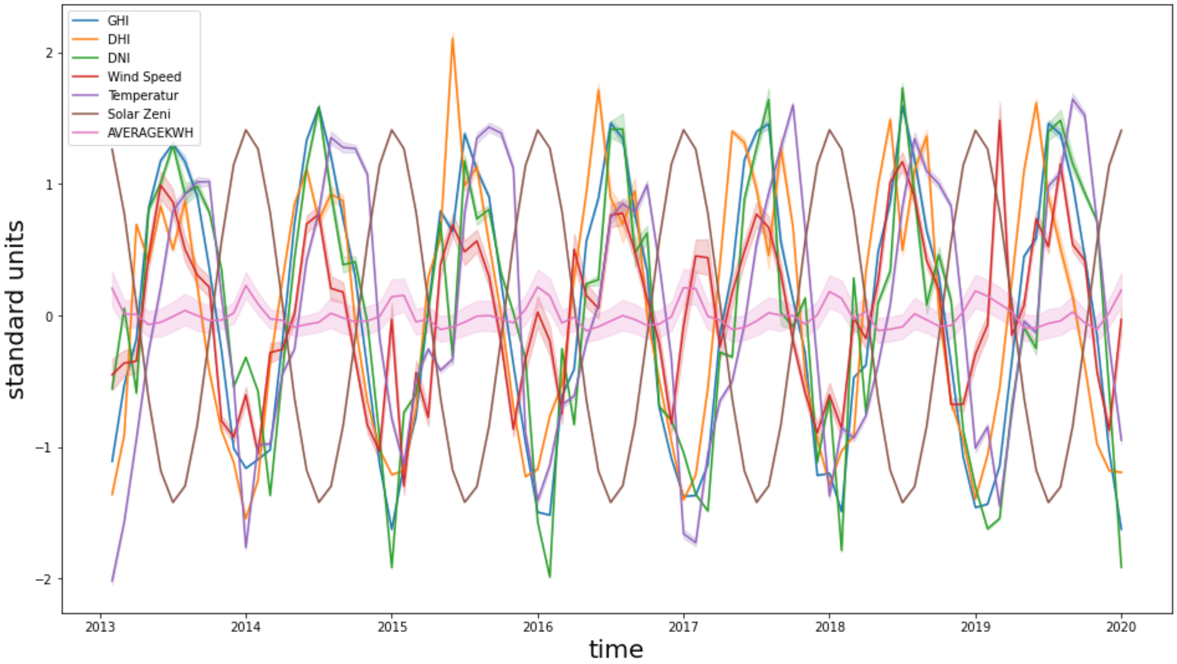


Figure 4. Plot of environmental metrics from NREL and PG&E Data in standard units over time

In Figure 4, we standardized each of our environmental metrics and plotted each of them using the line plot function to see how they changed with respect to time. This step was performed to visualize the relationship between these variables and average kWh and to see if there was a trend and to help guide us in which variables to include in our initial model. After reviewing this, we decided to include all of the variables from the NREL dataset in our model. Figure 5 shows the PG&E average electricity usage over time from 2013 to 2020 for residential customers.

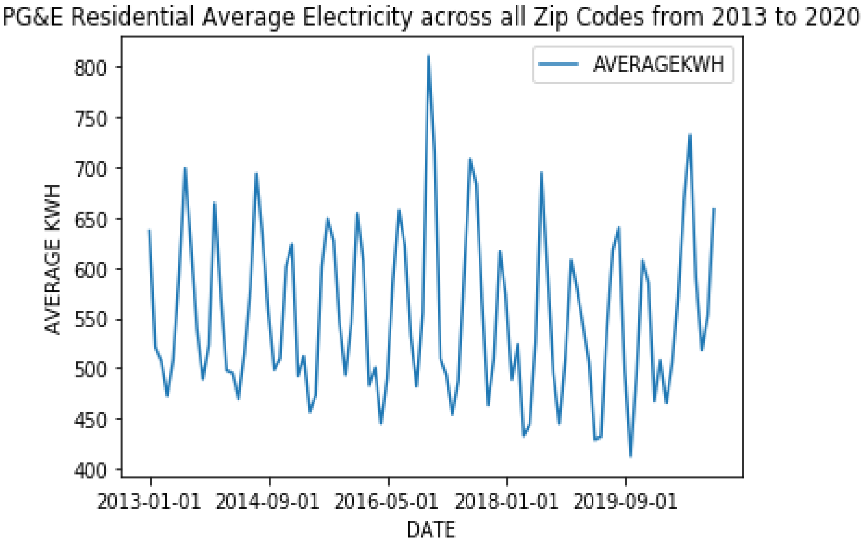


Figure 5. PG&E Average Electricity Usage Over Time

Data Cleaning

We also noticed a few outliers after plotting the standard units for the mean kWh consumed. As a result we decided to only include all values that were within 8.5 standard units away from the mean total kWh consumed from the PG&E data. Additionally, gaps in the data existed especially for certain zips code. Thus, we excluded the zip codes that had less than 84 monthly temperature measurements over the past 7 years or only had measurements for a singular year. The comparison of average electricity consumption of the Bay Area PG&E data before and after data cleaning shown in Figure 6 and 7.

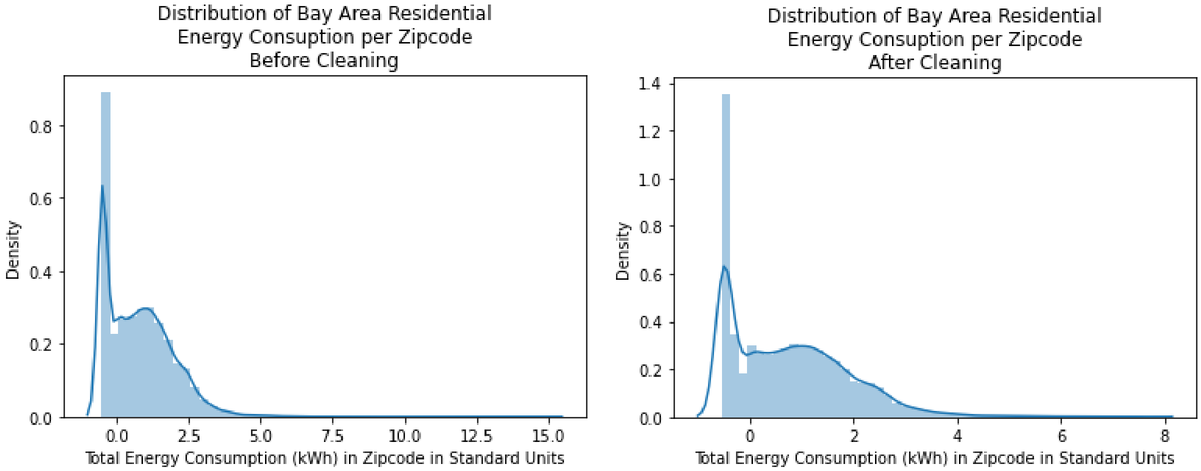


Figure 6. Distribution of mean kWh consumed in standard units before and after data cleaning

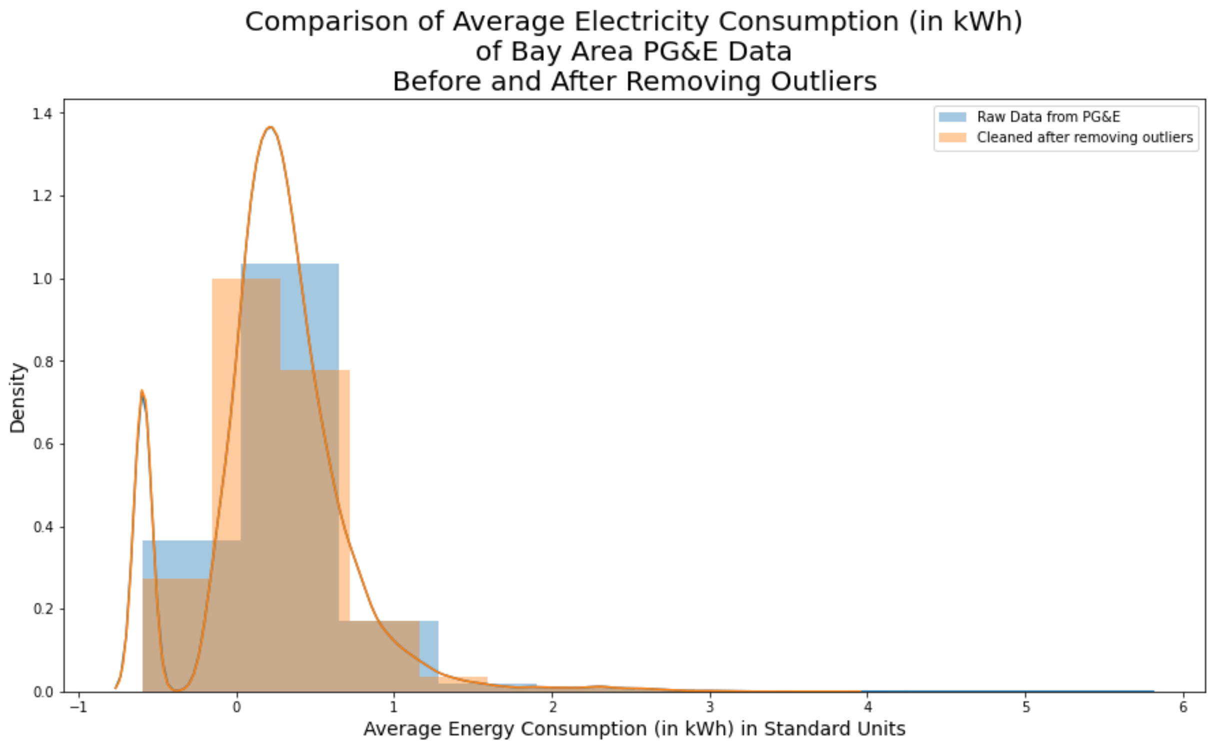


Figure 7. Plot of mean kWh consumed in standard units over time before and after data cleaning

Model Features

Each of the machine learning methodologies used the same X matrix, which consisted of 23 different features. From the PG&E's residential energy consumption dataset, we used month, year, and historical average energy consumption in kilowatt-hours. From NREL's Solar Irradiance Database, we used the following variables: GHI, DHI, DNI, wind speed, temperature, solar zenith angle. Additionally, we lagged certain features by zip code to incorporate the previous three months of data, which made our predictions significantly more robust. The lagged features included the average electricity consumption (in kWh) and all of the NREL solar irradiance variables.

Since our data measurements were as granular as zipcodes, the process of lagging these features was performed using the following:

```
Start with a blank Pandas DataFrame df.
For each unique zipcode Z:
    Using only measurements corresponding to Z, sort the measurements chronologically (using month and year) and call this S.
    Lag each variable (GHI, DHI, DNI, Wind speed, Temperature, Solar Zenith Angle, and Average KWH) by 1, 2, and 3 months to create an additional 21 columns in S.
    Concatenate S to df.
Resulting df now consists of 23 columns.
```

In summary, we lagged seven features (six from NREL and one from PG&E) for three months and included two variables from PG&E, for a total of 23 features to generate our X matrix, which was ultimately a 23136 x 23 matrix.

Machine Learning Methodology Employed

Ordinary Least Squares

The method of linear least squares is to find the best model by finding the model that minimizes the sum of the squares of the error between the observed data and the computed data. By minimizing the squared error, this is expected to reduce the effect of noise on the model.

The data is fit using these parameters, in the form:

$$f(x, \beta) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n$$

The β values represent the parameter estimates that are fit to the data, and the x variables are the explanatory variables. Some of the limitations of linear least squares regression is that it is sensitive to outliers, limitations in the shape of the regression over long ranges, collinearity among the features in the X matrix, and poor extrapolation properties. These limitations are solved somewhat by using other machine learning methods, which we will discuss next.

KNN Regression

K-nearest neighbours regression is non-parametric, which means it makes no assumptions about the underlying mathematical model. It is a rudimentary unsupervised learning technique that makes predictions of a new test data point based on the K nearest distance to surrounding data points from the training set. It's also important to note that there's no perfect number of neighbours to use that fits all data sets perfectly. By using hyperparameter tuning techniques, we found that the best hyperparameter, *n_neighbors*, was 10. When training the KNN model, the "y variable" that is essentially matched with the corresponding X vector. This leads to a MSE of 0 for our training data, because the model waits until the testing data in order to do any real work of predicting the classes.

The drawback of using KNN is that it is computationally expensive because in order to predict the "y variable" of a given input, it must compute the distance between the input and all of the training data then take the average y values of the surrounding K points. Additionally, our KNN model belongs to a 23-dimensional vector space. Figure 8 displays some interesting visualizations that show 12 different clusters in the data:

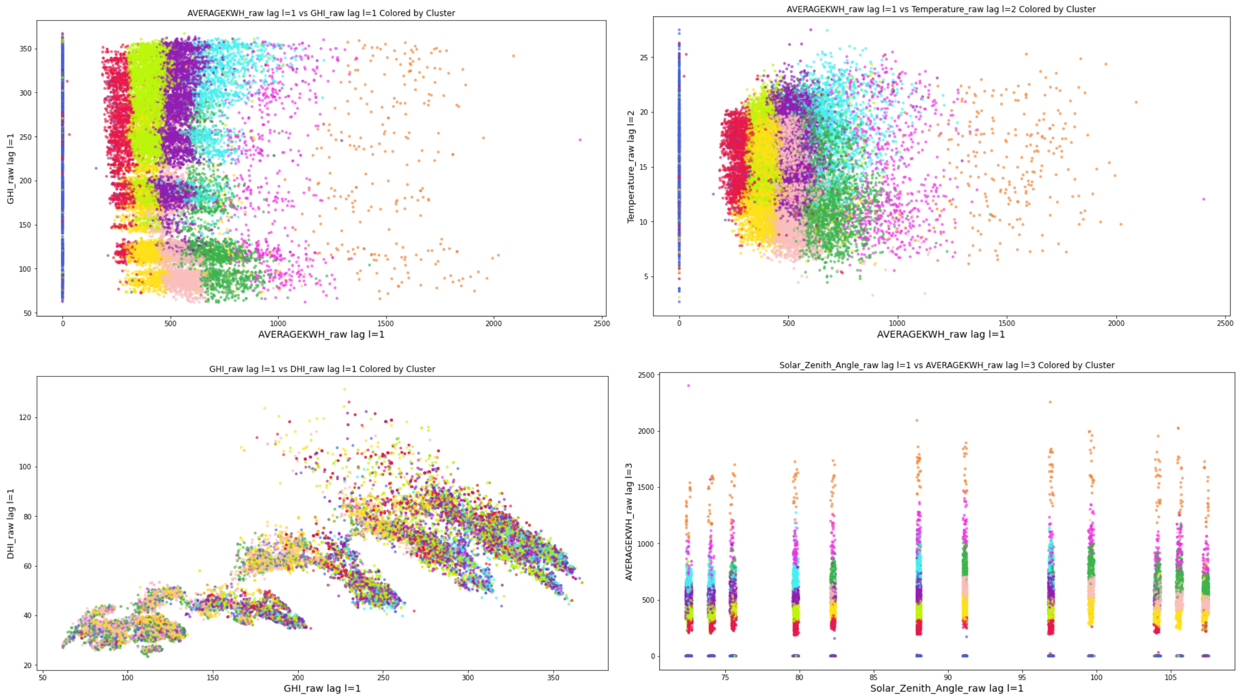


Figure 8: Subplots of 12 Clusters in the Cleaned Dataset

Because of the potentially encountering the curse of dimensionality, we performed principal component analysis (PCA) to reduce the dimensionality. Using the first 17 principal components accounted for at least 99.77% of the original variance in the X matrix. Using the first 9 principal components accounted for at least 95% of the original variance in the X matrix. Further investigation proved that using 23 dimensions had a better test MSE than the dimensionality reduced X matrix (Figure 9).

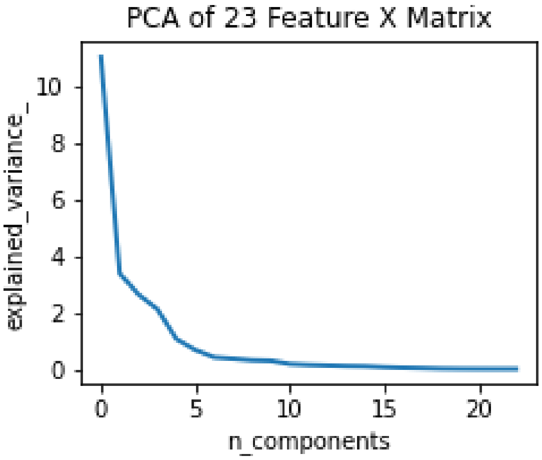


Figure 9: PCA Scree plot of X Matrix

Table 1: KNN MSE values for different X Matrix dimensions

KNN using D dimensions in X matrix	*Training MSE*	*Test MSE*
D = 23 (100% variance)	0.0	5243.805577161207
D = 17 (99.7% variance)	3.140670787779e-22	6217.155193178626
D = 9 (95% variance)	4.005471751402e-22	6242.230004593895

Support Vector Regression (SVR)

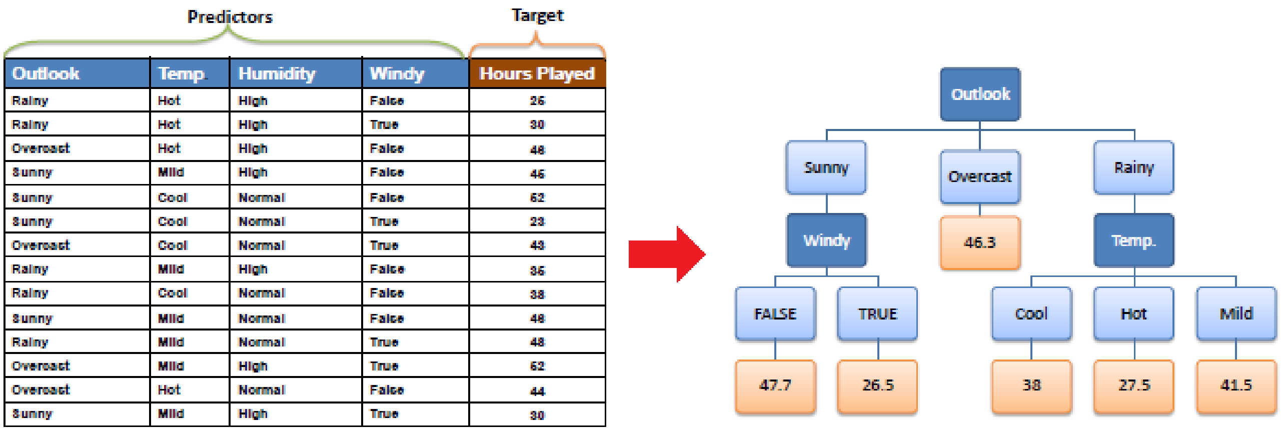
The objective function of SVR minimizes the l2-norm of the coefficient vector rather than the squared error. Thus, samples are penalized if the prediction is at least ε (epsilon) away from their true target. ε can be tuned in order to increase accuracy of the model. We used the following formulation:

$$min\ \frac{1}{2}||w||^2 \qquad with\ the\ constraint \qquad |y_i - w_i x_i| \le \varepsilon$$

Decision Tree (Randomized Tree)

The Decision Tree model is a non-parametric (like KNN) machine learning model, that has a goal of creating a model that produces the target variable (monthly average electricity consumption) by learning simple decision rules inferred from the data features (which are our environmental metrics).

Decision tree model is one of the supervised learning methods that uses a structure similar to a flowchart which consists of conditions (internal nodes), branches (edges) and decisions (leaves) to predict the value of a target variable. Although this method seems best for classification, it is also widely used for regression. Figure 10 is one of the examples that show the usage of the decision tree model for regression. The target variable predicted in this model is hours played and the predictor variables have roles in making decisions on each branch.



Predictors				Target
Outlook	Temp.	Humidity	Windy	Hours Played
Rainy	Hot	High	False	26
Rainy	Hot	High	True	30
Overcast	Hot	High	False	48
Sunny	Mild	High	False	46
Sunny	Cool	Normal	False	62
Sunny	Cool	Normal	True	23
Overcast	Cool	Normal	True	43
Rainy	Mild	High	False	36
Rainy	Cool	Normal	False	38
Sunny	Mild	Normal	False	48
Rainy	Mild	Normal	True	48
Overcast	Mild	High	True	62
Overcast	Hot	Normal	False	44
Sunny	Mild	High	True	30

Figure 10. Decision tree model for playing hours (Sayad 2021)

One of the key advantages of this model is that it is simple to understand, interpret and visualize. This model can handle both numerical and categorical data and predict multi-target variables while little effort for data preparation is required. However, this model has a so-called overfitting problem. This problem happens when tree models are too complex that they do not predict well in real cases.

Results

We used functions in the sklearn library to employ each of these machine learning algorithms, and produced a plot showing the training and testing data result after implementing each of these models. Figure 11 and 12 show the estimated results as the thick line and the confidence interval as the band surrounding the lines of the same color.

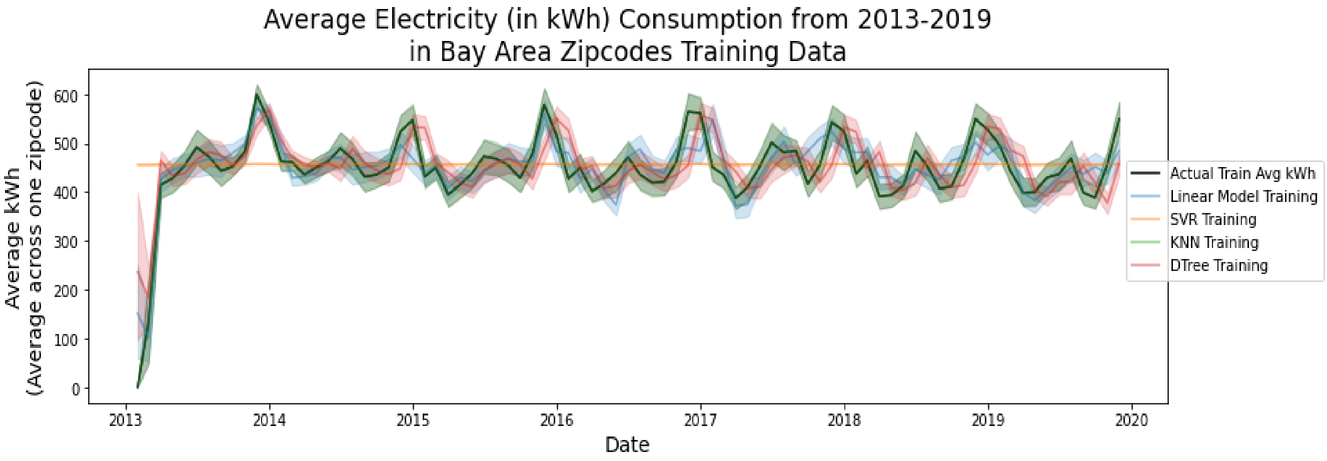


Figure 11. Plotted predicted average electricity consumption for all models using training data

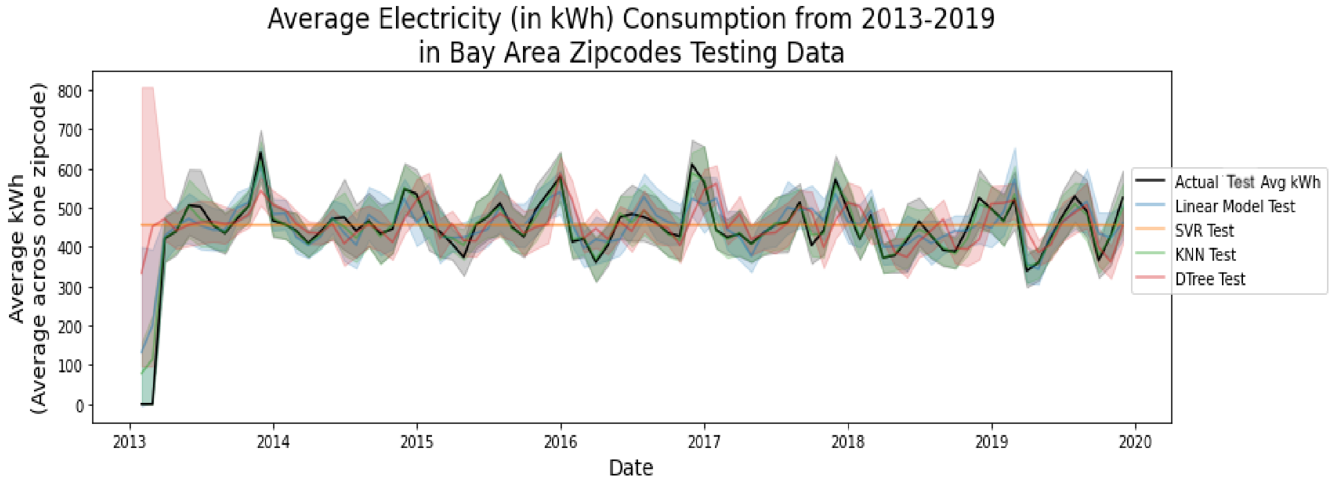


Figure 12. Plotted predicted average electricity consumption four all models using testing data

We also found the training and testing mean squared error for each of the models (Table 2), and found the model with the lowest testing MSE to be the KNN model, and SVR to have the highest MSE. We were thrown off at first when we calculated the MSE of KNN to be 0, however we learned this is because KNN is a lazy learner meaning it does not do the work of classifying the points until the testing stage.

Table 2. Computed MSE after using four different selected models

	Training MSE	*Testing MSE*
Linear Model prediction	11512.800185	12299.512756
SVR prediction	60582.595559	64781.677366
KNN prediction	0.00000	5243.805577
DTree prediction	22361.924220	24123.355030

IV. Discussion

The goal of this project was to create a machine learning model that could predict future residential electricity loads, and it was successfully conducted using NREL and PG&E monthly data from residential consumers after exploring four different machine learning models. A possible application of this model could be to predict future residential energy consumption, based on environmental metrics which will become more extreme as climate change continues. This can help electric utilities to understand when they may experience too much demand and can help them adapt and eventually predict blackouts more accurately and farther in advance to give customers who may be especially vulnerable more warning, such as senior citizens during summer seasons who may experience heat exhaustion if their power and subsequently A/C is turned off.

Some limitations were identified in our project including that household energy consumption loads may be dependent by other variables not related to climate change such as charging an electric vehicle, running an electric appliance, etc. Furthermore, a huge limitation can incur if the NREL and PG&E data is insufficient or inaccurate. Thus, it would be important to explore other energy consumption and climate data sets.

In addition, while the purpose of this study was to predict energy consumption in the Bay Area based on meteorological data such as temperature, humidity, and weather, the key parameter used in the models that increased the accuracy more dramatically than any other variables turned out to be the average monthly electricity usage in kilowatt-hour from the previous three months. The model accuracy suddenly improved after adding those energy consumption data to the X matrix used to train for the machine learning models. The reason for this could be that energy usage tends to increase when the weather is extreme on both sides. Due to this nonlinearity or complexity, it might be difficult for the machine learning to accurately predict future energy consumption without previous months' energy consumption data.

Another limitation of these models is that it could be difficult to apply in real time. Some of the data needed for the X matrix in the models may be difficult or impossible to obtain in a timely manner. For example, to predict energy consumption in June 2021, energy consumption in May

2021 would be needed to run the models. However, the energy consumption data of May 2021 cannot be obtained before June begins, which would present issues. In order to use the model, it is recommended to use data two or three months behind of the target month of prediction. Obtaining this data may be easier if there was a close relationship with PG&E established, meaning potentially preliminary data could be obtained.

Finally, because our model could only learn from data obtained for the years 2013 to 2019, only three months lagging was applied to prevent cutting too much training data. Twelve months lagging could be a better option looking at the seasonal trends of average kilowatt-hour data from Figure 4. However if we were to apply 12-month lagging, the data from the year 2013 would have been removed from the X matrix, which would have cut our available data by about 1/7 which was not ideal. For this reason, we kept the three month lagging which we still felt captured some of the seasonality.

The results were promising; however, there could be some improvements to this model. Firstly, to add more granularity to our models and predictions, it would be preferable to use daily data than monthly data. Moreover, due to the COVID-19 pandemic, it would have been interesting to include 2020 data as well because most residents spent more time in their homes in 2020 compared to previous years. Additionally, a prediction of future electricity usage would have provided more depth to this study. For example, it would have been insightful to predict May or June 2021 electric usage by utilizing weather/environmental forecasts and comparing them with the actual data, but due to data limitations this was not possible as described earlier.

V. Summary

In our project, we predicted monthly energy use in residential buildings and compared estimations based on different machine learning models and methods. Historical energy consumption data and solar radiance data from the Bay Area was used to train and validate our models. Data was taken from PG&E and NREL for the models. The data from PG&E included monthly residential energy consumption while data from NREL included the monthly values of the direct normal irradiance (DNI), diffuse horizontal irradiance (DHI), global horizontal irradiance (GHI), solar zenith angle, temperature and wind speed.

The machine learning models that were used to forecast average monthly energy consumption for residential homes in the Bay Area with climate data included ordinary least squares, KNN, support vector regression, and decision trees. The KNN model performed the best prediction results based on comparing all four MSE values computed for each of the models

Further improvements can be made to ensure more accurate machine learning models, such as utilizing daily data rather than monthly data to add more granularity to our model and predictions. We can also expand our research by incorporating future weather forecasts to predict future electricity, which can be compared later with actual data when available.
However, our energy forecasts can serve as an initial evaluation of the overall impact of climate change on the total demand on the energy grid in the future. This forecast can also be translated to the energy demand on a grid in a region/area. Thus, utilities and organizations can plan ahead to supply and demand imbalance and present power outages.

References:

1.10. Decision Trees—Scikit-learn 0.24.1 documentation. (n.d.). Retrieved March 12, 2021, from https://scikit-learn.org/stable/modules/tree.html

4.1.4.1. Linear Least Squares Regression. (n.d.). Retrieved March 17, 2021, from https://www.itl.nist.gov/div898/handbook/pmd/section1/pmd141.htm

4.1.4.2. Nonlinear Least Squares Regression. (n.d.). Retrieved March 12, 2021, from https://www.itl.nist.gov/div898/handbook/pmd/section1/pmd142.htm

API Instructions—NSRDB. (n.d.). Retrieved March 17, 2021, from https://nsrdb.nrel.gov/data-sets/api-instructions.html

Bay Area (California)—Wikitravel. (n.d.). Retrieved March 30, 2021, from https://wikitravel.org/en/Bay_Area_(California)

Decision Tree Regression. (n.d.). Retrieved March 12, 2021, from http://www.saedsayad.com/decision_tree_reg.htm

Gupta, P. (2017, May 17). *Decision Trees in Machine Learning | by Prashant Gupta | Towards Data Science*. Towards Data Science. https://towardsdatascience.com/decision-trees-in-machine-learning-641b9c4e8052

Kretchmer, H. (2020, July 3). *Chart of the day: How US energy consumption has evolved since independence*. World Economic Forum. https://www.weforum.org/agenda/2020/07/united-states-energy-consumption-since-independence/

Limón, M. F., Sami Sparber and Elvia. (2021, February 15). *2 million Texas households without power as massive winter storm drives demand for electricity*. The Texas Tribune. https://www.texastribune.org/2021/02/15/rolling-blackouts-texas/

Marohl, B. (2020, April 28). *In 2019, U.S. energy production exceeded consumption for the first time in 62 years—Today in Energy—U.S. Energy Information Administration (EIA)*. Today in Energy. https://www.eia.gov/todayinenergy/detail.php?id=43515

McFarland, A. (2019, April 19). *In 2018, the United States consumed more energy than ever before—Today in Energy—U.S. Energy Information Administration (EIA)*. Today in Energy. https://www.eia.gov/todayinenergy/detail.php?id=39092

Moura, S. (2019). *Chapter 4: Machine Learning*.

National Solar Radiation Data Base—OpenEI Datasets. (n.d.). Retrieved March 17, 2021, from https://openei.org/datasets/dataset/national-solar-radiation-data-base

Navlani, A. (2018, August 2). KNN Classification using Scikit-learn. DataCamp Community. https://www.datacamp.com/community/tutorials/k-nearest-neighbor-classification-scikit-learn

PG&E's Energy Data Request Portal. (2021). PG&E Energy Data Request-Public Data Sets. https://pge-energydatarequest.com/public_datasets

Salcedo-Sanz, S., Deo, R. C., Carro-Calvo, L., & Saavedra-Moreno, B. (2016). Monthly prediction of air temperature in Australia and New Zealand with machine learning algorithms. *Theoretical and Applied Climatology*, *125*(1), 13–25. https://doi.org/10.1007/s00704-015-1480-4

Sharp, Tom. *(2020, May 6). An Introduction to Support Vector Regression (SVR). Medium. https://towardsdatascience.com/an-introduction-to-support-vector-regression-svr-a3ebc1672c2*

U.S. EIA. (2019, April 16). *In 2018, the United States consumed more energy than ever before—Today in Energy—U.S. Energy Information Administration (EIA).* https://www.eia.gov/todayinenergy/detail.php?id=39092

US Zip Code Latitude and Longitude. (n.d.). Retrieved March 30, 2021, from https://public.opendatasoft.com/explore/dataset/us-zip-code-latitude-and-longitude/

Wang, R., Lu, S., & Feng, W. (2020). A novel improved model for building energy consumption prediction based on model integration. *Applied Energy*, *262*, 114561. https://doi.org/10.1016/j.apenergy.2020.114561

Williams, K. T., & Gomez, J. D. (2016). Predicting future monthly residential energy consumption using building characteristics and climate data: A statistical learning approach. *Energy and Buildings*, *128*, 1–11. https://doi.org/10.1016/j.enbuild.2016.06.076

Yadav, P. (2018, November 13). *Decision Tree in Machine Learning*. Towards Data Science. https://towardsdatascience.com/decision-tree-in-machine-learning-e380942a4c96

Zhao, H., & Magoulès, F. (2012). A review on the prediction of building energy consumption. *Renewable and Sustainable Energy Reviews*, *16*(6), 3586–3592.