Springboard Capstone Project 1 – Machine Learning
Predicting Short Term Solar Energy Production

Connor McAnuff

September 18, 2019

1. Overview

This report discusses machine learning model building for predicting solar energy production. This report should be read in conjunction with the accompanying Jupyter Notebook.

The goal of the machine learning models is to predict the daily solar energy availability at each of the 98 stations given daily weather forecasts. It is a regression task. Multiple models have been constructed for evaluation and comparison:

1) Ordinary Least Squares (OLS) linear regression
    a. No Feature Selection
    b. Reduced Features
    c. Further Reduced Features
2) Stochastic Gradient Descent (SGD)
3) Gradient Boosting Regressor (GBR)

2. Model Evaluation

The data is first split into training and test data – the training data is then further split into 5 contiguous folds for cross-validation (CV). The folds and train/test split are done at year end/beginning to account for autocorrelation in the energy data from year-to-year:

- 1994-1995: Contiguous Fold 1
- 1996-1997: Contiguous Fold 2
- 1998-1999: Contiguous Fold 3
- 2000-2001: Contiguous Fold 4
- 2002-2003: Contiguous Fold 5
- 2004-2007: Test Data

Models will be evaluated using the Mean Absolute Error (MAE) which is commonly used in the energy industry to measure prediction accuracy according to the client. First, an average CV MAE will be calculated by running each of the 5 contiguous folds as the CV test data, leaving the remaining 4 folds as the CV train data. Next, all 5 contiguous folds are used together as the training data, and the 2004-2007 test data is used as a final validation (evaluated using MAE). Additionally, Pearson's Correlation Coefficient (R^2) will be used for model evaluation.

3. Linear Regression (OLS)

3.1 Model Overview

Linear regression fits a hyperplane to a set of points in N-dimensions (N = number of features). It generally requires the following assumptions:

- Linear relationship between the target variable and features exists.
- Data should not exhibit multicollinearity.
- Homoscedasticity (constant variance) of residuals.
- Normally distributed residuals.

3.2 OLS Model 1a – No Feature Selection

The baseline model will use no feature selection. Each of the 15 weather forecast variables at 5 forecast hours are used as 75 different features. Additionally, latitude, longitude, elevation, and the month of the year are used as features to total 79 features. The baseline model evaluation criteria scores are given in Table 1.

Table 1: OLS Model 1a – No Feature Selection Evaluation Criteria Scores

Evaluation Criteria	Score
CV MAE (W/m^2)	2,353,201
Test MAE (W/m^2)	2,195,538
Train Data R^2	0.834

Assumption #1: Linear relationship between the target variable and features exists.

The coefficients of the regression are non-zero and thus show that there are linear relationships between the target variable and the features. The p-value of the t-test scores are close to 0, indicating that it is very likely that these features have predictive power. The relationship between a single feature (current temperature at 2m above the ground) and energy is shown in Figure 1. The data can by modelled by a linear relationship but would have significant and biased residuals.

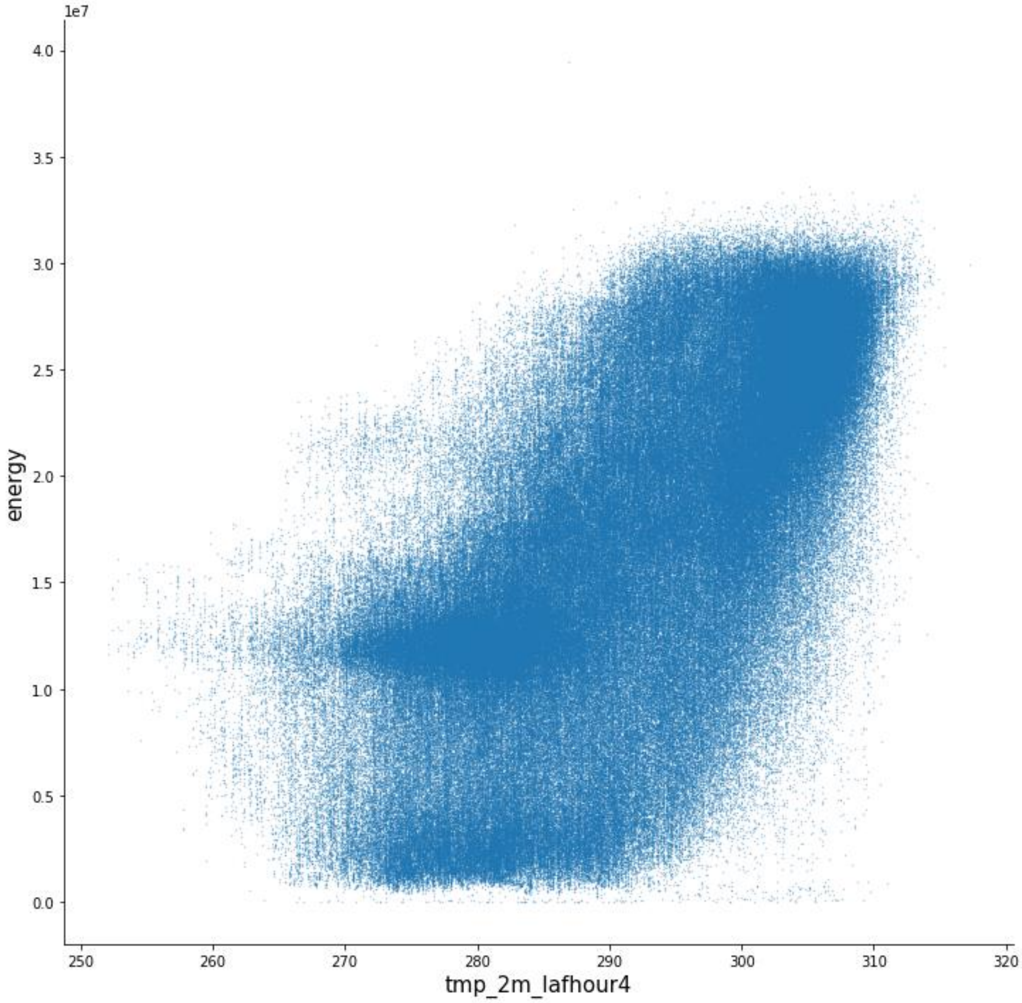

Figure 1: Current temperature at 2m above the ground forecasts vs energy.

Assumption #2: Data should not exhibit multicollinearity.

Features highly correlated with other features can cause an increase in the standard errors of the coefficients. In turn, coefficients for other features may not be found to be significantly different from 0, and thus they will be labelled statistically insignificant. Multicollinearity can be tested using a correlation matrix (calculating Pearson’s Correlation Coefficient between each set of features and between each feature and the target variable). To start, only forecast hour 0 will be checked for multicollinearity between features – the correlation matrix is shown in Figure 2.

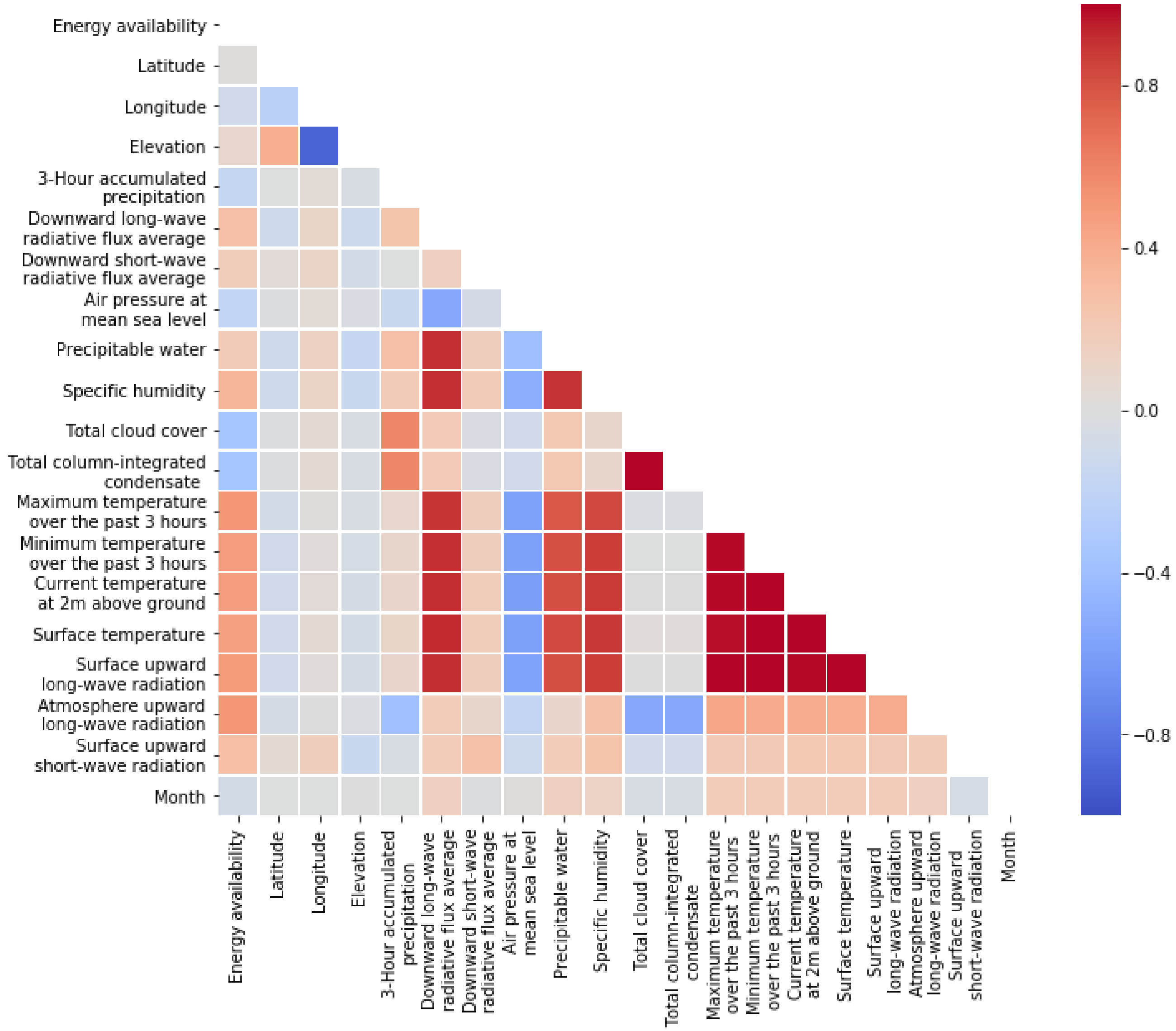

Figure 2: Correlation Matrix for all features (weather variables forecast hour 0 only)

Many features exhibit multicollinearity. A reasonable cut-off for labelling a correlation ‘strong’ is $|R^2| > 0.7$. When choosing which of a set of features to remove, it makes sense to remove the feature that has a weaker correlation with the target variable, energy. This logic has been used to determine that the following features should be removed from the data to lower the risk of multicollinearity causing issues with the linear regression:

- Elevation
- Downward long-wave radiative flux average at the surface (all forecast hours)
- Precipitable Water over the entire depth of the atmosphere (all forecast hours)
- Specific Humidity at 2 m above ground (all forecast hours)
- Total column-integrated condensate over the entire atmosphere (all forecast hours)
- Minimum Temperature over the past 3 hours at 2 m above the ground (all forecast hours)
- Current temperature at 2 m above the ground (all forecast hours)
- Temperature of the surface (all forecast hours)
- Upward long-wave radiation at the surface (all forecast hours)

Assumption #3: Homoscedasticity (constant variance) of residuals.

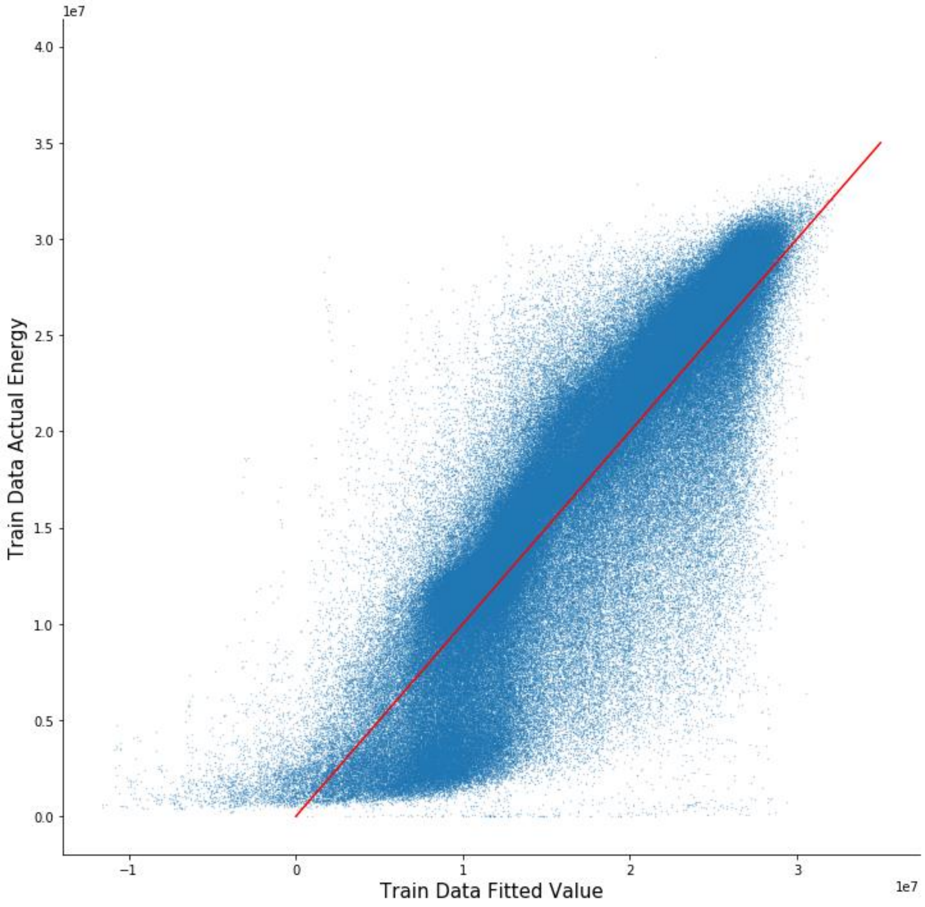

Figure 3: Model 1a train data fitted vs actual energy values.

Visually, it can be seen in Figure 3 that the residuals are not distributed evenly along the line of a theoretical perfect prediction. Low actual energy values are over-predicted while higher actual energy values are under predicted. Statistically, homoscedasticity can be tested using the Breusch-Pagan (BP) test, which takes the residuals and features as inputs. The null hypothesis of this test is that there is no violation of homoscedasticity. The BP test gave a p-value of near 0, indicating the null hypothesis is rejected, and there is likely a violation of homoscedasticity, confirming what was seen visually.

Assumption #4: Normally distributed residuals.

A histogram of residual values is shown in Figure 4. The distribution appears somewhat normally distributed, but it appears that the left side has more mass than the right side. The Anderson-Darling (AD) Normality test has the null hypothesis that the data is normally distributed. The p-value of the test statistic is near 0, thus we reject the hypothesis that the residuals are normally distributed.

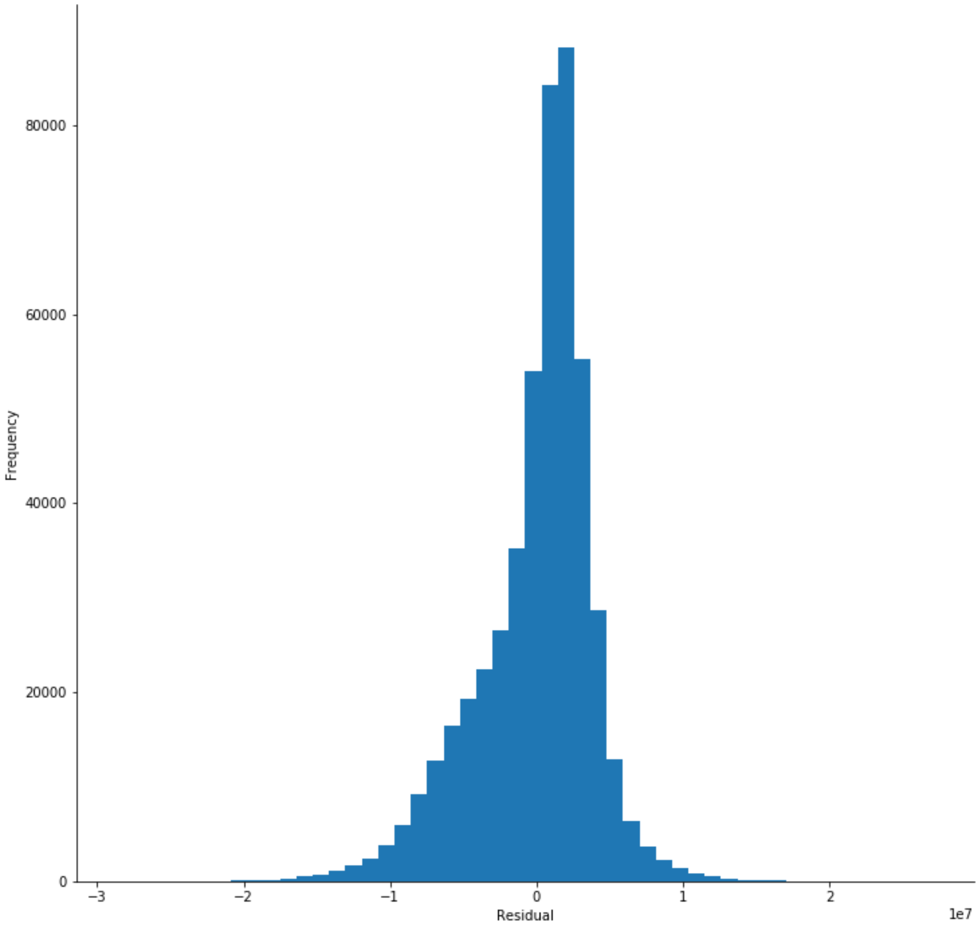

Figure 4: Model 1a residual histogram

Non-normality of residuals can also be visualized using Quantile-Quantile (Q-Q) plots. Ideally, the Q-Q trace will be along the red line (Figure 5). The trace clearly strays from the red line at the tails, indicating the linear regression is a poor fit at low and high values of energy.

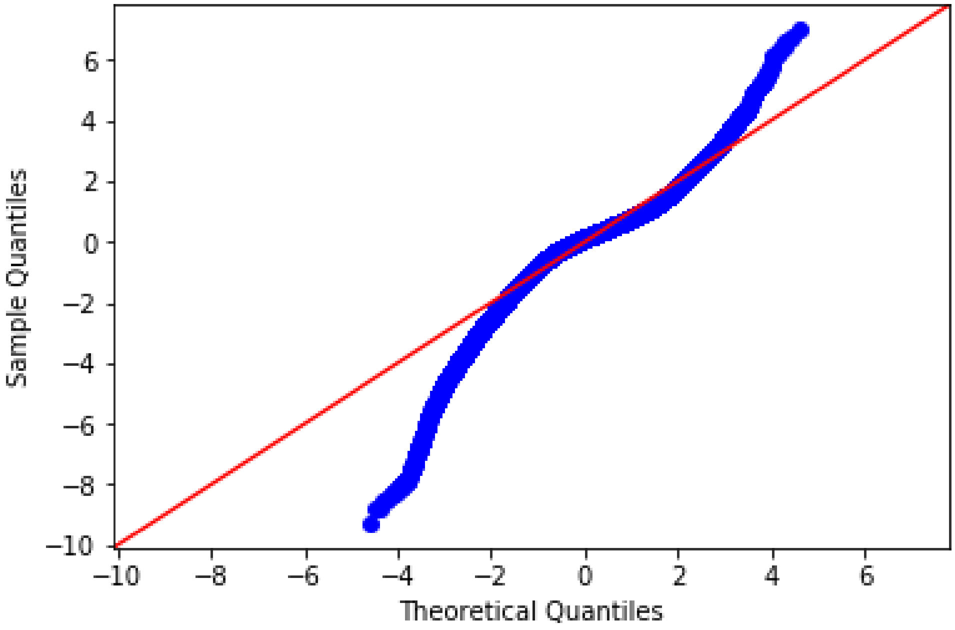

Figure 5: Model 1a Quantile-Quantile plot

3.3 OLS Model 1b – Reduced Features

The features identified for removal in model 1a discussion to reduce multicollinearity have been removed to construct model 1b. The model evaluation criteria scores are given in Table 2. CV MAE and test MAE are both greater than for model 1a. Additionally, train data R^2 has increased.

Table 2: OLS Model 1b – Reduced Features Evaluation Criteria Scores

Evaluation Criteria	Score
CV MAE (W/m^2)	2,552,359
Test MAE (W/m^2)	2,359,393
Train Data R^2	0.810

This model exhibits the same violations of the basic assumptions as model 1a. Homoscedasticity and normally of residuals are both violated. Additionally, there is still multicollinearity present, as only forecast hour 0 was checked previously. A correlation matrix for all features in model 1b is shown in Figure 6. This correlation matrix allows us to identify further features for removal from the model to account for multicollinearity.

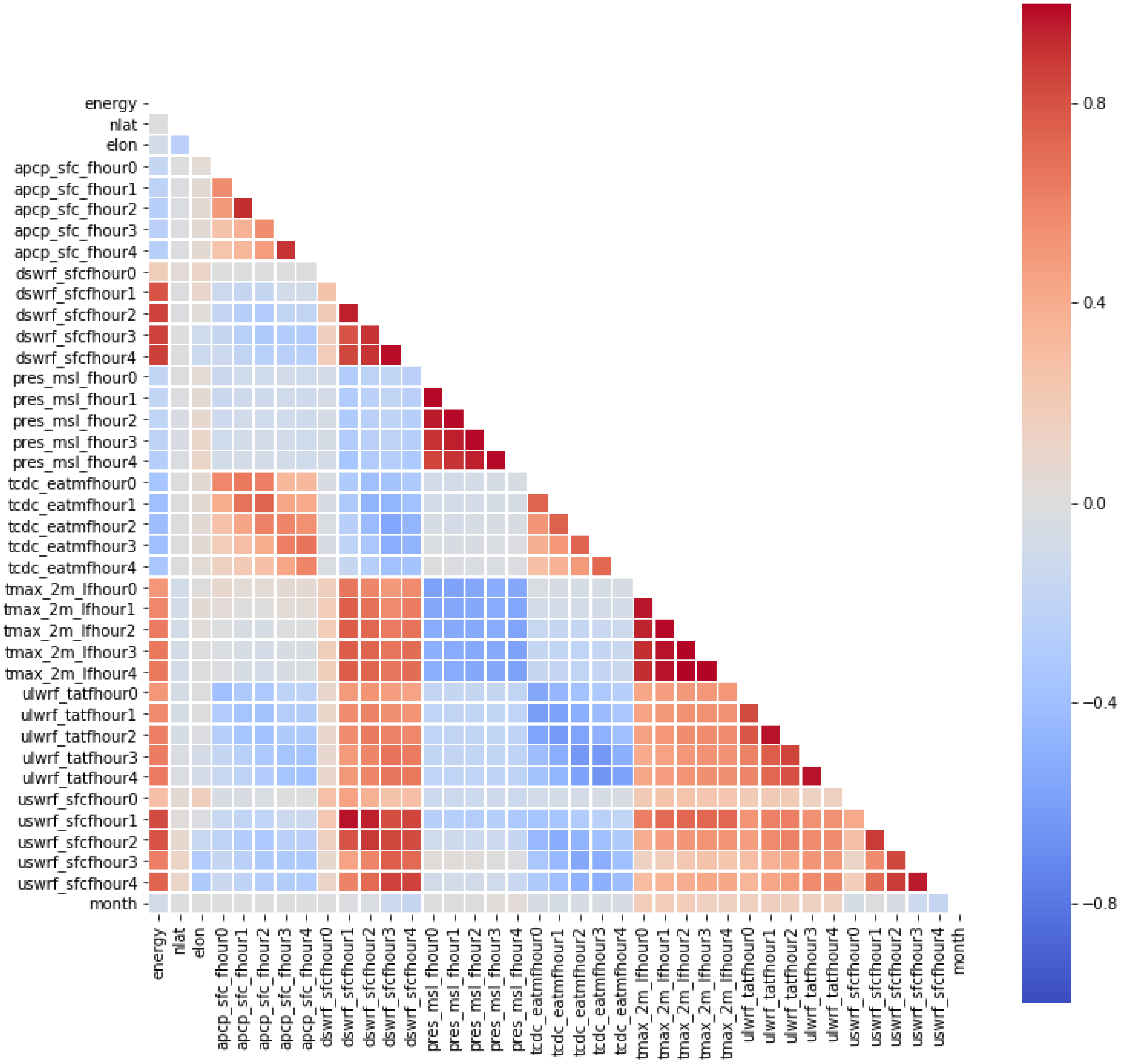

Figure 6: Correlation matrix for all features in model 1b.

3.4 OLS Model 1c – Further Reduced Features

The features identified for removal in model 1b discussion to reduce multicollinearity have been removed to construct model 1c. The model evaluation criteria scores are given in Table 3. CV MAE and test MAE are both greater than for model 1a and model 1b. Additionally, train data R^2 has increased.

Table 3: OLS Model 1c – Further Reduced Features Evaluation Criteria Scores

Evaluation Criteria	Score
CV MAE (W/m^2)	3,021,995
Test MAE (W/m^2)	2,844,458
Train Data R^2	0.750

This model again violates homoscedasticity and normality of residuals, although there are no longer any 'strong' correlations between features.

3.5 OLS Model Comparison

Homoscedasticity and normality of the residuals could not be achieved by any OLS model. When multicollinearity was reduced, the predictive power of the model also decreased. This result implies that multicollinearity was not a problem for this dataset. Overall, the linear regression models do have predictive power, with model 1a having a test MAE of 2,195,538 (W/m^2). The mean daily energy is 16,567,964 (W/m^2), resulting in a test MAE that is ~13.3% of the mean daily energy. The models all have decreased prediction power when predicting at very low and high values.

4. Stochastic Gradient Descent

4.6 Model Overview

Stochastic Gradient Descent (SGD) is an iterative process that can be used to optimize the coefficients of a linear regression by minimizing a cost function. 'Stochastic' means that the process is linked with random probability – samples of the dataset are randomly selected for each iteration, as opposed to using the entire dataset, reducing computational requirements. SGD is sensitive to feature scaling – the constructed model uses a pipeline to scale the features prior to fitting.

4.7 Hyperparameter tuning

SGD models have several hyperparameters to be tuned. For this model, the following hyperparameters have been tuned using GridSearchCV - the resulting best parameters are also listed:

- Alpha (0.00001)
- Max iterations (60)
- Epsilon (0.1)
- Loss function (sum of squared differences)

There are additional hyperparameters that could be tuned, but for the purposes of this project, only those listed above were tuned – the remaining parameters have been left as the default values.

4.8 Results

The model performs very similarly to the simple OLS model (1a) with the same features. Again, the homoscedasticity and normality of residuals have been violated.

Table 4: SGD Model Evaluation Criteria Scores.

Evaluation Criteria	Score
CV MAE (W/m^2)	2,360,498
Test MAE (W/m^2)	2,204,851
Train Data R^2	0.833

5. Gradient Boosting Regressor

5.9 Model Overview

Gradient Boosting Regression (GBR) is an ensemble prediction process (i.e., it uses a collection of predictors to give a final prediction). The aim of this process is to reduce noise, variance, and bias. The process uses boosting, meaning the ensemble of predictors are made sequentially – observations with higher errors from previous predictors are more likely to appear in subsequent predictors. Each predictor is a regression tree.

5.10 Hyperparameter tuning

GBR models have several hyperparameters to be tuned. For this model, the following hyperparameters have been tuned using GridSearchCV - the resulting best parameters are also listed:

- Number of estimators (1000)
- Max features (10)
- Max depth (6)

The loss function used was least absolute distance. A subsample of 0.5 was used for each estimator. There are additional hyperparameters that could be tuned, but for the purposes of this project, only those listed above were tuned – the remaining parameters have been left as the default values. Computation time for fitting GBR models with many datapoints is high – the grid search was performed on an Amazon Web Service EC2 instance.

5.11 Results

The GBR model performs better than all other models. Again, the homoscedasticity and normality of residuals have been violated.

Table 5: GBR Model Evaluation Criteria Scores.

Evaluation Criteria	Score
CV MAE (W/m^2)	2,123,177
Test MAE (W/m^2)	1,975,215
Train Data R^2	0.846

6. Model Comparison

The three best-performing models have their evaluation criteria compared in Table 6. GBR has the lowest CV and Test MAE values, and a highest R^2. Additionally, Figure 7 shows that the residuals of the GBR model have a much more even distribution. Figure 8 compares distributions of fitted and actual values for each of the models. The GBR model produces values more similar to the actual data than the OLS model.

Table 6: Model evaluation criteria comparison.

	OLS Model 1a	SGD	GBR
CV MAE (W/m^2)	2,353,201	2,360,498	2,123,177
Test MAE (W/m^2)	2,203,998	2,204,851	1,975,215
Train Data R^2	0.833	0.833	0.846

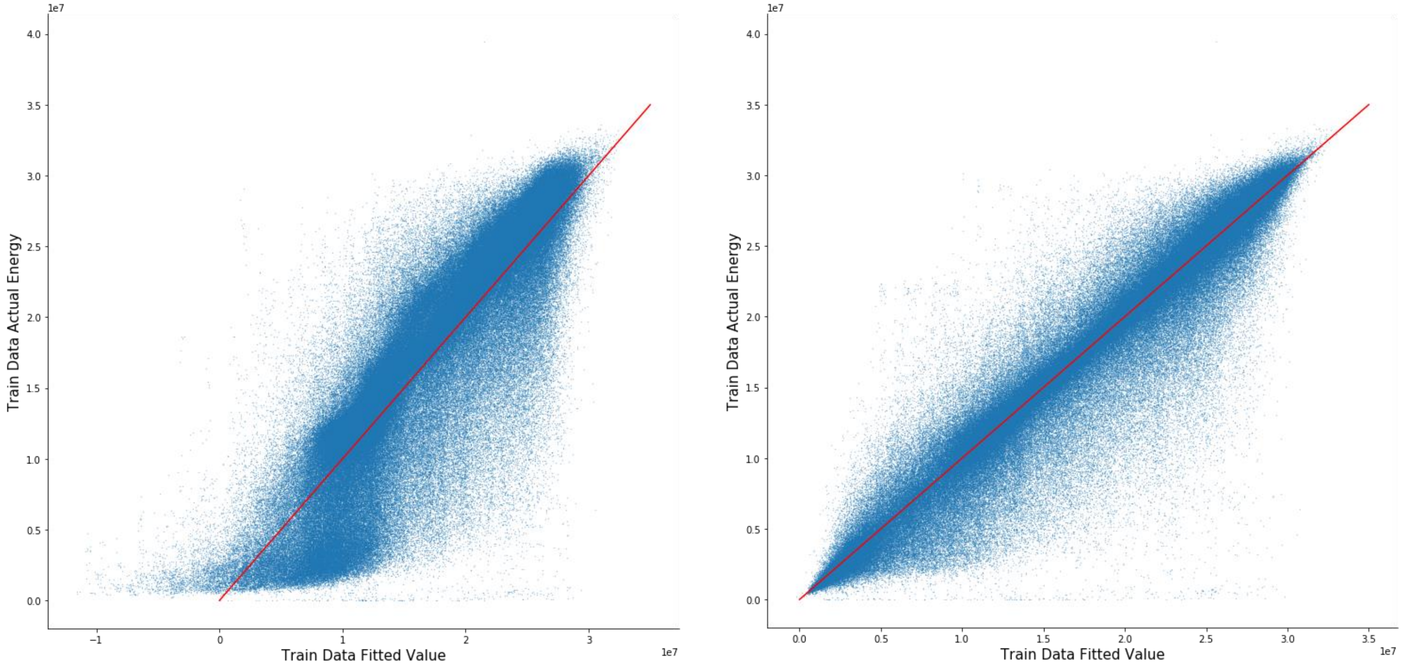

Figure 7: Fitted vs actual energy value scatterplots for (left) OLS model 1a and (right) GBR model

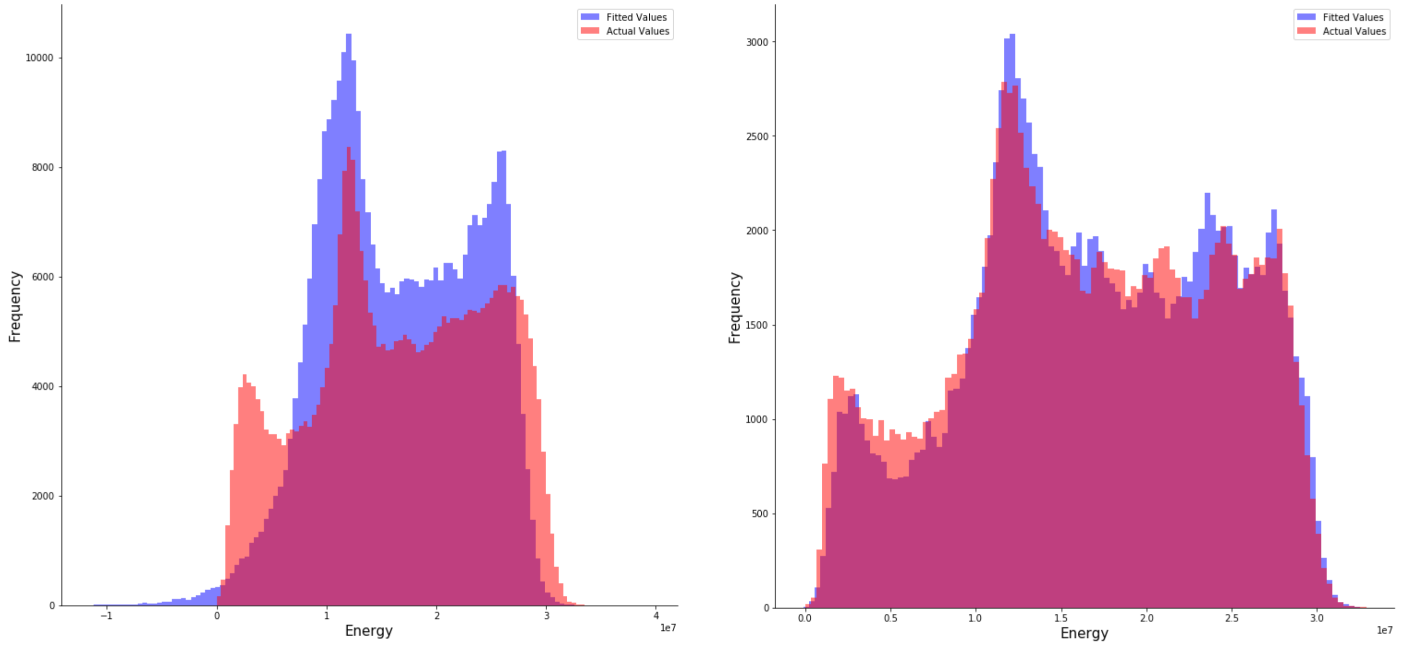

Figure 8: Fitted and actual energy value histograms for (left) OLS model 1a and (right) GBR model

7. Conclusions and Further Work

Linear regression resulted in models with predictive power but many shortcomings. Low values were consistently over-estimated while higher values were consistently under-estimated. The residuals therefore violate homoscedasticity and normality, indicating that linear regression may not be the best choice of model for this dataset. Additionally, the linear regression models predict negative values, as an implicit artifact of the method.

The GBR model performed better than the linear regression models according to visual assessment of the data and the evaluation criteria of CV MAE, Test Data MAE, and Train Data R^2. The GBR model was able to predict on a test dataset (separate from fitting and hyperparameter tuning) the total daily energy with a mean absolute error of 1,975,215 (W/m^2), which is approximately 11.9% of the mean total daily energy value.

Further work on this project would be to spend more time tuning the GBR hyperparameters using randomized and grid searches. Additionally, it would be interesting to use all 11 of the weather forecast models to train 11 regression models and average the predictions. Lastly, it could be explored to use a distance-weighted average of weather forecasts (at multiple grid points) as the weather forecast variable features.