Journal of Cases on Information Technology
Volume 23 • Issue 4

Forecasting Solar Radiation:
Using Machine Learning Algorithms

Pankaj Chaudhary, North Carolina A&T State University, USA
Rohith Gattu, Indiana University of Pennsylvania, USA
Soundarajan Ezekiel, Indiana University of Pennsylvania, USA
James Allen Rodger, Slippery Rock University of Pennsylvania, USA*

ABSTRACT

Renewable energy, such as solar and wind, has been increasing in popularity for over a decade. This is especially true in rural, underdeveloped areas and urban households that desire energy independence. Renewable energy sources, such as solar, provide enhanced environmental benefits while simultaneously minimizing the carbon footprint. One popular technology that can capture solar energy is solar panels. The demand for solar panels has been on the rise due to increases in energy conversion efficiency, long-term financial advantages, and contributions to decreasing fossil fuel usage. However, solar panels need a steady supply of sunlight. This can be challenging in many situations, geographies, and environments. This paper uses multiple machine learning (ML) algorithms that can predict future values of solar radiation based on previously observed values and other environmental features measured without the use of complex equipment with methods that are computationally efficient so that forecasting can be done on consumer premises.

KEYWORDS

Linear Regression, Random Forest Regression, Renewable Energy, Solar Radiation Forecasting, Support Vector Regression

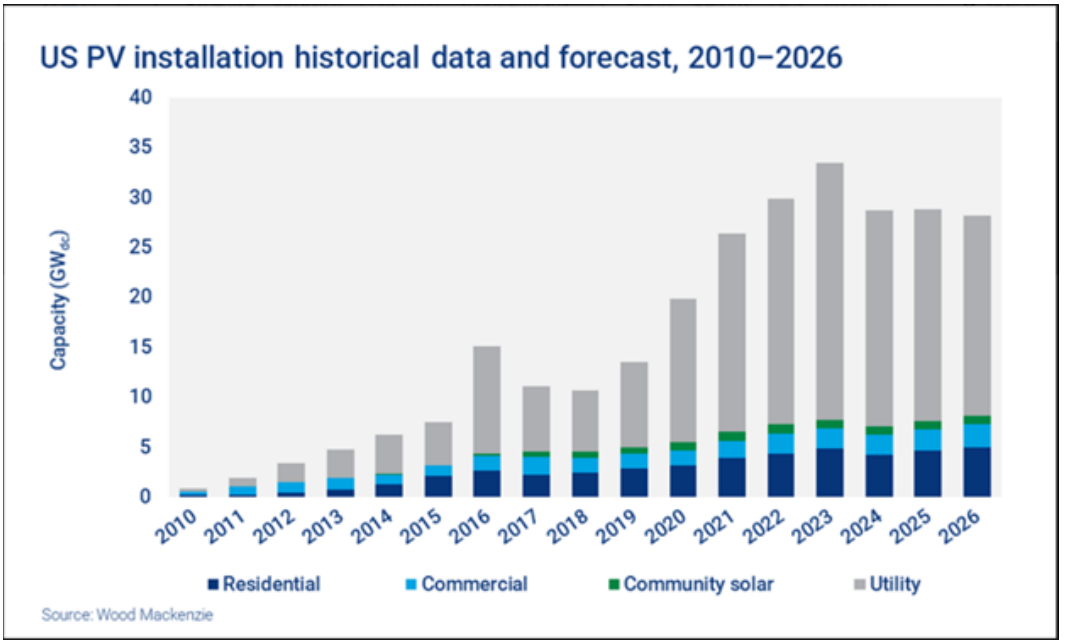
INTRODUCTION

Most of the worldwide energy is derived from fossil fuels. In 2018, the world's electricity consumption amounted to approximately 23.4 trillion kilowatt-hours (Sönnichsen, 2020). Various countries employ different methods of electricity generation such as steam, nuclear, biomass, geothermal, coal, etc. Along with the United States, China is one of the highest per capita consumers of electricity in the world. In 2019, the China consumed 5,564 billion kWh of electricity and US consumed 3,902 billion kWh of electricity (Sönnichsen, 2020). With the exponential increase in electricity consumption, countries have implemented new ways to decrease emissions stemming from power generation through the use of fossil fuels. Carbon dioxide emissions from fossil fuel sources grew by 2.7% in 2018 which was much faster compared to the growth of 1.6% in 2017 (Levin, 2018). In 2019 the growth somewhat slowed to 0.6% for the first 6-10 months (Friedlingstein et. al., 2019). This decreased by -7% in 2020 due to the pandemic (Friedlingstein et. al., 2020). The dip though is seen a temporary, and steps are

DOI: 10.4018/JCIT.296263
*Corresponding Author

needed to reduce the carbon emissions. Renewable energy sources like solar energy will make a significant contribution to this effort.

Figure 1. US Photo Voltaic Installation (Wood Mackenzie, 2021)



In recent years, solar energy has been increasing in popularity. Figure 1 shows the photo voltaic installations in the use both in terms of history and future projections (Wood Mackenzie, 2021). The residential share of this market can be seen as steadily increasing. Power generation from solar increased 22% in 2019 and this constituted 3% of the total electricity generation in 2019 (IEA, 2020). The demand for solar systems has increased due to added interest from homeowners and businesses. Also, the likely pairing with battery storage to extend electricity availability throughout the off-hours has increased demand for the solar systems. By 2025, more than 25% of all behind-the-meter solar systems will be paired with battery storage, compared to under 5% in 2019 (SEIA, n.d.). Strong federal policies and incentives, such as Solar Investment Tax Credit, have increased the likelihood that consumers will become interested in solar power. Increased solar energy usage offers a myriad of benefits including environmental safety, decreased pollution, and financial benefits, such as decreased utility bills (Stevović, 2017). Solar power has become more affordable, accessible, and prevalent in many parts of the world. Solar panels are usually integrated with smart grids. The main goal of smart grids is to substantially increase the adaptability of renewable energy. Due to off-time challenges and dependency on nature such rainy or cloudy days, it becomes challenging to run renewable energy integrated smart grids efficiently in the absence of any predictive analysis. The problem with substantial renewable integration is that the electricity generated from renewables is not easily predictable and will vary based on weather conditions and site-specific conditions (Jolliffe, 2017). This is where predictive analysis can add significant value to the operation of a smart grid and allow for a steady and balanced supply of electrical power. The predictive analysis carried out in this manuscript can be employed at consumer locations which are not covered by projects such as the Google Sunroof Project (google.com/get/sunroof). Even for those locations that are covered by such projects the granularity

of prediction at consumer sites can provide some additional value. The analysis presented in this manuscript may be used to provide a more accurate prediction of solar radiation using feature set from a consumer grade weather station in conjunction with day length forecast from an Internet weather website of choice. The feature set used does not require measurements using complex equipment as usually done in several studies discussed later in this manuscript.

In addition to smart grid integration, solar radiation forecasting is an essential tool for the operation and management of solar power plants. It is also important for future solar panel installations on independent buildings and residences. Solar radiation forecasting anticipates the solar radiation transience and power production of solar energy systems. This allows for the setup of contingency mechanisms to mitigate any deviation from the required production (Lorenz et. al., 2017). For solar farm owners and consumers who are in the process of installing solar panels or are future solar panel owners, providing predictive capabilities can help alleviate many uncertainties due to environmental factors. An accurate forecast of available solar resources and power is essential for managing the electric grid, market operations, and reducing the cost of solar energy (Kumler et. al., 2018). Many data repositories of information about solar activity in various areas and predictive models exist, but these areas are limited to high populations and large installations. This study uses regression models to forecast radiation energy using data that can be gathered from a reasonably simple consumer grade weather station and data about day length that can be obtained from several weather sites on the Internet. Such forecasting can then serve as an aid to assist further growth and spread of solar energy by providing easily implementable forecasts of solar radiation.

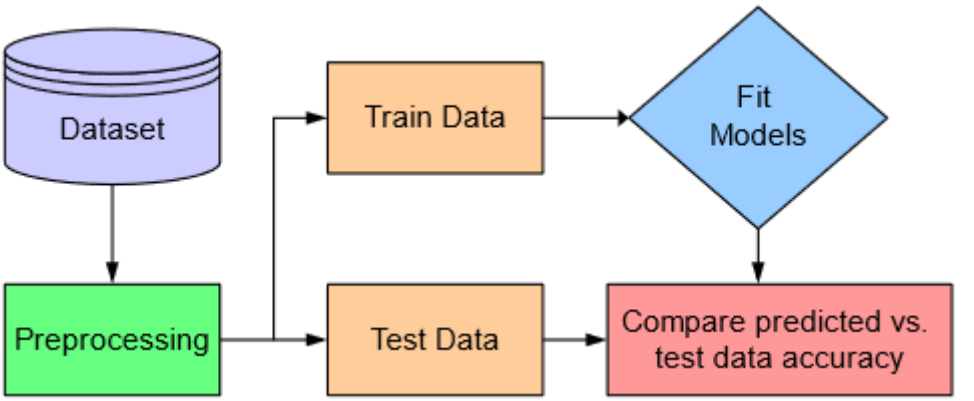
The remainder of this paper is organized as follows: Section II presents background research on related topics that apply to solar prediction. It also covers background information on the solar radiation prediction and its evolution. Section III summarizes the prediction techniques and concepts that are used in this paper. Section IV provides details of data preprocessing done including data sources used in this manuscript. Section V discusses the results of running linear regression, random forest regression, multi perceptron regression, and support vector machine regression models with visuals that show the accuracy of the machine learning algorithms as well as patterns in the dataset. Section VI discusses conclusions, limitations, and future work that can be undertaken following this study. It also discusses the practical applications of this study.

BACKGROUND

Solar radiation that reaches Earth's surfaces provides a crucial role in the balance of energy of various physical, biological, and chemical processes. Changes in the amount of solar radiation greatly influence the fluxes of sensible and latent heat, the hydrological cycle, terrestrial ecological ecosystems, and the climate (Islam et. al., 2009). Solar radiation can be captured and employed to generate electrical energy that has a much lower environmental impact in comparison to fossil fuels and nonrenewable energy. The future of energy production relies heavily on renewable resources due to the negative environmental effects caused by nonrenewable energy, such as global warming. While some may claim that the science is inconclusive, sufficient evidence can be found in the popular media and academic journals (Botzen et. al., 2021). The use of Machine Learning (ML) for solar energy prediction can provide many benefits. Accurate prediction will lead to better estimate of electrical energy during a day and better planning and integration with the electrical grid. This translates to benefits such as increasing the efficiency of the solar panels, reducing overall costs, increasing employment in the solar energy sector, and decreasing carbon emissions. Thus, the benefits associated with solar prediction can lead to increased use of solar energy. Figure 2 provides an overview of the machine learning process.

There are many studies that have been conducted for prediction of solar radiation using different methods including traditional regression models, artificial neural network (ANN) and machine learning (ML) models. In one of the early studies of its kind, Halouani et. al (1993) have done a comparison of various methods of calculation of the average global radiation using Gariepy's (1980), Hay's

Figure 2. Machine Learning Process



(1979), lqbal's (1979), and Rietveld's (1978) models of prediction. All these models are based on the Kimball-Angstrom-Page equation. The study shows the result that Gariepy's model provides the best performance. However, all these models require complex measurement like total precipitation in Gariepy's model to Hay's model requiring measurement of ground albedo, cloud sky albedo, and cloud albedo. Iqbal's model requires monthly average beam and diffuse solar radiation. This is in contrast to the ML methods used in the current study using simple features that can be obtained from the consumer grade weather station and weather information available on the Internet for a location or zip code. All require information on monthly average daily hours of bright sunshine. Another model by Matzarkis and Katsoulis (2006) uses a formula consisting of distance from nearest coast, height above sea level, percentage of land cover around measuring station, and latitude and longitude of the station. Okundamiya and Nzeako (2010) present a model for Nigeria using a regression with monthly mean daily data set for minimum and maximum ambient temperatures. This model used historical temperature data and is simple in nature. However, the estimation is done for monthly mean daily or sub-hourly solar radiation. This study estimated a different model for different cities. In the Halouani et. al (1993) study, the models performed worst in the Northeast region of Canada characterized by severe and particular climactic conditions with long days and long solar nights. In addition, the models had different performance accuracies in different regions signifying that one model is not sufficient for different regions and possibly different models are needed for different regions. Due to variation of the climactic conditions radiation prediction models may be more specific to location rather than being applicable globally though such attempts have been made (Yin, 1999).

Kandirmaz et. al. (2014) employed an ANN to estimate monthly sunshine duration in Turkey using cloud cover, day length, and month. In this study sunshine duration was predicted since solar radiation is highly correlated to sunshine duration (Suehrcke et. al., 2013). Global solar radiation measurements are generally made with the actinographs which are often not reliable due to the need of routine calibration of thermal sensitivity of the mechanical components of their sensors (Kandirmaz et. al., 2014). More accurate measurements can be done by constructing networks with calibrated modern pyranometers but these are expensive (Kandirmaz et. al., 2014). Sunshine duration can be more accurately determined using cheaper instruments. Day length is a function of the latitude and longitude and solar declination and provides the maximum duration of sunshine in a day (Kandirmaz et. al., 2014). In this study the generalized regression neural network (GRNN) and multilayer perceptron (MLP) neural network performed better than the Radial Basis Function (RBF) neural net. The bright sunshine hours were recorded using a Campbell-Stokes type sunshine recorder. 21 years of data was used for training and last six years of data was used for testing. This is a long time horizon which more likely constrains the practical widespread use of the model. There are other several ANN based studies

for predicting solar radiation using variables such as sunshine duration, temperature, cloud cover, relative humidity, wind speed, vapor pressure, precipitation, elevation, latitude, longitude, month, and satellite recorded or derived variables (Yin, 1999). Karasu et. al. (2017) used machine learning with linear and gaussian regression to estimate solar radiation using wind speed, temperature, pressure, and humidity for Zonguldak province in Turkey. In this study gaussian regression performed better than the linear regression for prediction purposes. Qazi et. al. (2015) and Yadav and Chandel (2014) provide a comprehensive review of various studies in the area of prediction of solar radiation using ANN. Yadav and Chandel (2014) found that ANN techniques predict solar radiation more accurately in comparison to the conventional methods like regression analyses. The prediction accuracy of ANN models is found to be dependent on input parameter combinations, training algorithm and architecture configurations. Qazi et. al. (2015) came to a similar conclusion that ANN is one of the reliable and accurate methods for prediction of solar radiation. ANNs provide good accuracy with prediction error of less than 20%. The model accuracy was found to be dependent on input parameters and algorithms that were utilized. Though both studies do not refer to locational specificity of the models, that may be an additional consideration when using ML and ANN techniques to predict solar radiation as it was found in earlier conventional studies (Halouani et. al., 1993; Okundamiya and Nzeako, 2010).

Short-Term Forecasting of Solar Radiation

Reliability is the key factor for an acceptable standard and amount of electrical energy. Reliability is needed and desired and it is even more important for renewables which can vary due to environmental conditions. Ensuring reliability in the case or renewables is a difficult thing however coupling them with the traditional sources and battery storage can ensure reliability provided predictability can be provided. The underlying system reliability indices in the power distribution system are the "load interruption frequency", "expected duration of load interruption events", and "magnitude of the load interruption" (Moreno-Munoz et. al., 2008). Frequency, duration, and magnitude have a significant effect on electric supply. Multiple benefits come from forecasting variations of solar irradiances which also include a better expected value for these indices. Forecasting increases predictability of the availability of the system that impacts load supply points through metrics such as total expected interruption time per year and expected demanded but unsupplied energy per year. This manuscript tests various classes of models to assess which specific class of models provides the best predictive analysis for sub-hourly solar irradiation. Sub-hourly prediction is important since the sub-hourly quality of power is crucial for assessing the availability and reliability indices. Also, most protection, monitoring, and control devices are designed based on reliable sub-hourly supply. The reliability of sub-hourly supply has become more important as distributed generation (which includes solar power) has become more popular in the energy market due to various economic and environmental issues (El-Khattam and Salama, 2004). The predictive analysis models can also be easily incorporated and used for a maximum power point tracker (MPPT) for radiation measurement.

There are several studies that have attempted to predict hourly solar radiation using ML or Artificial Intelligence models like ANN. Khosravi et al. (2018) estimated hourly solar radiation using local time, temperature, pressure, wind speed, and relative humidity as input variables of the models and a time-series prediction model. They employed multilayer feed-forward neural network (MLFFNN), radial basis function neural network (RBFNN), support vector regression (SVR), fuzzy inference system (FIS) and adaptive neuro-fuzzy inference system (ANFIS) for the two predictions. Their results obtained an $R = 0.9999$ and 0.9795 SVR and MLFFNN models. For the time series models SVR, MLFFNN and ANFIS models reported the correlation coefficient more than 0.95 for the testing dataset. Khatib et. al. (2012) used eight geographical and climatic variables of hour, day, month, latitude, longitude, temperature, humidity, and daily sunshine hours ratio (i.e., measured sunshine duration over daily maximum possible sunshine duration) to estimate both hourly global and diffused solar radiation using GRNN, feed forward back propagation neural network (FFNN),

cascade forward back propagation neural network (CFNN) and Elman back propagation neural network (ELMNN). Their analysis showed GRNN had the best performance.

There are also several studies for forecasting sub-hourly solar radiation which is the target of this study. Hocaoglu and Serttas (2017) used Mycielski-Markov hybrid method with accurate results. Zhang [31] used a computational statistics-based approach for solar radiation reconstruction at 1 minute temporal resolution with a normalized root means square error of 23.4%.

Chen et. al. (2019) outline the state-of-the-art solar radiation forecasting methods into five clusters. These clusters are Numerical Weather Prediction (NWP), Statistical Methods, Top-down methods, Bottom-up methods, and Hybrid methods. Of all these methods statistical methods are further categorized into model-based methods and data driven methods. The data driven methods based on ML have good elasticity in spatial and temporal dimensions. As such the use of machine learning methods for sub-hourly forecasting of the solar radiation is an appropriate methodology to explore. In addition, use of features that are readily available without the use of complex measuring equipment is another distinction of the study in this manuscript. Thus, the research question explored in this study is if a computationally friendly data driven ML algorithm can accurately forecast (R^2 of 90% or more) sub-hourly solar radiation.

PREDICTION TECHNIQUES

Several techniques are available for predictive analysis in ML. They range from using Artificial Neural Networks (ANN) of various kinds to traditional techniques of time series analysis. Techniques that combine different procedures in one model usually perform better at predictive analysis. Once a model has been trained, it can be used in the future unless there is a significant shift in the relationships between independent and dependent variables. For weather parameters such shifts are not large from year to year though they are present due to the phenomena of global warming which introduces more extreme variations in the weather conditions. Different models are advised since they vary in accuracy depending on the data. Accuracy is of vital importance for surface energy budget, climate change, and energy applications (Shrama et. al., 2017). As part of this research, four different regression models were evaluated which are discussed below. The choice of these regression models was based on these being amongst the popular models which were appropriate for the task at hand.

Linear Regression

Linear regression is a supervised ML algorithm. Linear regression models provide a simple and easy to understand estimation procedure on a modular level. However, most real-world problems are not linear, and thus a significant negative of the algorithm is its oversimplicity (Copas, 1997). Though it should be stressed that when comparing models with similar accuracy simple models are always preferred over complex models.

In statistics, regression analysis aims to construct mathematical models that describe or explain relationships that may exist through the variables (Bayindir et. al., 2011). Most regression models target a specific value based on independent variables. Still, regression models can differ based on the relationship between the dependent and independent variables as well as the number of independent variables being used. Usually, a single dependent, continuous variable called the response (predicted or dependent) variable is studied in terms of how it depends on a set of variables called the explanatory variables (regressors or independent variables) (Seber and Lee, 2012). The dependent variable is expressed as a linear function of independent variables, corresponding regression parameters, and a random error term. The error term represents the variation in the dependent variable unexplained by the function of the independent variables and coefficients. The regression function is determined only by the parameters estimated by the regression technique. Multiple methods have been used to determine various parametric relationships between the response variable and independent variables.

Such methods typically depend on the form of the parametric regression function and the distribution of the error within the model. A linear regression line has an equation of the form:

$$Y=a+\beta_1 X_1+ \beta_2 X_2+\cdots+\varepsilon \quad (1)$$

In this equation, Xi is the explanatory variable and Y is the dependent variable. The slope of the hyperplane is dictated by βi and a is the intercept on that hyperplane.

Random Forest Regressor

Random forest regressor (RFR) is a supervised learning algorithm. An RFR typically uses the ensemble learning method for purposes of classification and regression. It is a meta estimator that fits several classifying decision trees on various sub-samples of the dataset. Then, it uses averaging to improve the predictive accuracy and control over-fitting (Scikit, n.d.). A particular tree predicts the output according to the features that are part of that tree. Predictions from multiple trees are averaged to provide a final prediction. RFR provides a good balance on imbalanced datasets that have high volume (Speiser et. al., 2019). There are several benefits of RFR (Scikit, n.d.). The random forest algorithm is not biased as there are multiple trees and each tree is trained on a subset of data. The random forest algorithm relies on the power of "the crowd", which reduces the overall biases of the algorithm. It is also very stable. Even if a new data point is introduced into the dataset, the overall algorithm is minimally affected. This is because although new data may impact one tree, it is very hard for it to impact all the trees. It works well with both categorical and numerical features. Lastly, the random forest algorithm functions properly when data has missing values, or it has not been scaled well (Malik, 2018).

The central concept is to use a large ensemble of decision trees. Each of the trees exhibits a very low-quality classification/regression result, but due to their large quantity, the result becomes sufficiently accurate (Speiser et. al., 2019). Random forest model is represented through a class of models:

$$g(x)=f_0(x) + f_1(x) + f_2(x) +\ldots.. \quad (2)$$

The result of the final model g is the sum of the simple base models fi, with each base model being a classifier of a simple decision tree. The results depend on the base learners chosen for a given situation and problem.

Multilayer Perceptron Regressor

Multilayer perceptron (MLP) regressor is based on Artificial Neural Network (ANN). ANNs are information processing systems inspired by the biological neural network that has been abstracted into a mathematical model (Khalyasmaa et. al., 2019). A perceptron is mathematical counterpart of a neuron that classifies input by separating two categories with a straight line. An MLP model consists of an input layer, one or more hidden layers, and a traditional output layer. For each hidden layer and the output, multiple nodes are interconnected to the previous layer. Information is passed through to train the network to predict the outcome and then predictions for the outcome are made. MLPs are suitable for classification and prediction problems where inputs are assigned a class or label. MLPs are very flexible and are advantageous when using tabular data.

The number of nodes and hidden layers can vary for each specific problem. More nodes and hidden layers result in higher sensitivity to the data being used for the prediction. This can lead to an increased risk of overfitting the data. Also, computational power requirements increase as the number of nodes and hidden layers increase. Through a combination of several perceptrons in an

MLP architecture, nonlinear classification or regression problems can be addressed by distinguishing data that is not linearly separable (Wang et. al, 2016).

Support Vector Regression

Support vector regression (SVR) is a forecasting model that has an exceptional rate in enhancing nonlinear prediction performance. SVR is a memory-efficient algorithm. Effectiveness comes into play when the number of dimensions is greater than the number of sample items. It is also effective in high dimensional spaces. The support vector regression function can be approximated as follows:

$$f(x)=w\cdot v(x)+b \tag{3}$$

In the equation, b is the bias, w is the weight vector, and v is the vector array. Maximizing the flatness of the function leads to a much smoother function in the input space. This is important in the formulation of the optimization problem used to construct the SVR approximation. The two main components that make up the SVM are a hyperplane and a decision boundary. The hyperplane acts as a separating line between two or more data classes in the support vector regression. The decision boundary is a demarcation line that lies on both the positive and negative sides.

The accuracy and results of an SVR model depend on the kernel function as well as the parameters. Kernel options include linear kernel, polynomial kernel, and radial basis function (RBF) kernel (Leonard and Kramer, 1991). Regressions based on these kernels are strongly correlated with each other. SVR models may contain redundant information, which manifests into an accuracy decrease. Principal component analysis (PCA) is a popular method for removing the redundant pieces of information from an input dataset, thereby reducing its dimensionality (Leonard and Kramer, 1991). With the implementation of PCA, the SVR noise can be decreased, which leads to better results in terms of prediction accuracy.

DATA PREPROCESSING

This data for this study was obtained from the HI-SEAS weather station for four months (September through December 2016) between Mission IV and Mission V (Python 3.8.5 Documentation, n.d.). The data was downloaded from Kaggle (Andrey, 2017). This was a NASA mission of a simulated mars environment that took place in Hawaii (HI-SEAS, n.d.). The fields in the dataset are: “UNIXTime” which is the UNIX epoch time, “Data” is a date in yyyy-mm-dd format, “Time” is the local time of day in a hh:mm:ss 24-hour format, “Radiation” is the solar radiation measured in watts per meter squared, “Humidity” is measured in percent, “Pressure” is the barometric pressure measured in Hg, “Temp” is the temperature measured in degrees Fahrenheit, “WindDirection” is the wind direction measured in degrees, “Speed” is the wind speed measured in miles per hour, “TimeSunRise” is the sunrise time, and “TimeSunSet “ is the sunset time. A partial data set is shown in Table 1. The total number of data points in this data set is 32,686. It is worth mentioning at this point that this data is something that most consumer grade weather station should be able to measure with sufficient accuracy. Hence the analysis employed in this research may be employed with data gathered from a normal weather station.

A time-series regression analysis can be done on the UNIX epoch time “UNIXTime”. The UNIX epoch time is used by UNIX operating system to track time as a running total in seconds. The count starts at the Unix Epoch on January 1st, 1970 at UTC. Therefore, the Unix timestamp is the number of seconds between a particular date and the Unix Epoch (Murali, 2013). The dataset contains data from autumn and winter of 2016 at approximately five-minute intervals. The .csv file was loaded in a python pandas data frame. Seaborn module in Python was used for data visualization. Sunrise and sunset values were converted to Python DateTime objects, which are stored as time-zone naïve

Table 1. Snapshot of the data set

UNIXTime	Data	Time	Radiation	Temp	Pressure	Humidity	Wind Direction (Degrees)	Speed	TimeSunRise	TimeSunSet
1475229326	9/29/2016	23:55:26	1.21	48	30.46	59	177.39	5.62	6:13:00	18:13:00
1475229023	9/29/2016	23:50:23	1.21	48	30.46	58	176.78	3.37	6:13:00	18:13:00
1475228726	9/29/2016	23:45:26	1.23	48	30.46	57	158.75	3.37	6:13:00	18:13:00
1475228421	9/29/2016	23:40:21	1.21	48	30.46	60	137.71	3.37	6:13:00	18:13:00
1475228124	9/29/2016	23:35:24	1.17	48	30.46	62	104.95	5.62	6:13:00	18:13:00
1475227824	9/29/2016	23:30:24	1.21	48	30.46	64	120.2	5.62	6:13:00	18:13:00
1475227519	9/29/2016	23:25:19	1.2	49	30.46	72	112.45	6.75	6:13:00	18:13:00
1475227222	9/29/2016	23:20:22	1.24	49	30.46	71	122.97	5.62	6:13:00	18:13:00
1475226922	9/29/2016	23:15:22	1.23	49	30.46	80	101.18	4.5	6:13:00	18:13:00
1475226622	9/29/2016	23:10:22	1.21	49	30.46	85	141.87	4.5	6:13:00	18:13:00
1475226323	9/29/2016	23:05:23	1.23	49	30.47	93	120.55	2.25	6:13:00	18:13:00
1475226025	9/29/2016	23:00:25	1.21	49	30.47	98	144.19	3.37	6:13:00	18:13:00
1475225720	9/29/2016	22:55:20	1.22	49	30.47	99	139.8	6.75	6:13:00	18:13:00

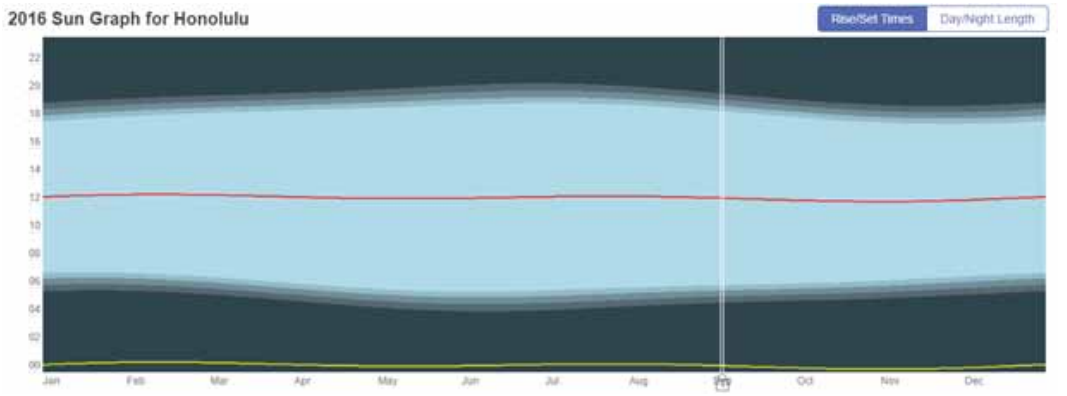
UNIX time values. Python's DateTime module was used for manipulating dates and times during the preprocessing routine. "UNIXTime" was also converted to a Python DateTime object. Time objects and data objects can be categorized as either "aware" or "naïve." A naive object does not contain enough information to unambiguously locate itself relative to other date/time objects (Murali, 2013). Also, a naïve object can be represented by Coordinated Universal Time (UTC). However, it becomes aware of the presence of a specific time zone. "UNIXTime" was transformed from UTC to Hawaii Standard Time, then data was sorted by "UNIXTime". A new variable "DayLength" was calculated as the difference between the "TimeSunSet" and "TimeSunRise". "TimeSunRise" and "TimeSunSet" were removed from the data frame since the information was subsumed to a limited extent into the "DayLength".

"DayLength" may serve as a proxy for seasonal variation in solar radiation. The length of the day at the location of the station situated at an approximate latitude of 19.8968° N will vary with the season. This pattern can be visually confirmed from the sun graph in Figure 3. Within the time horizon of the collected data the "DayLength" reduces as one progresses towards the December month. Table 2 show the descriptive statistics for "DayLength" (units are in seconds). As the season moves from September to December, the intensity of solar radiation also changes. "DayLength" may also serve as proxy for intensity.

Table 2. Descriptive statistics for "DayLength"

mean 41433.032491
std 1795.873502
min 39360.000000
25% 39720.000000
50% 41040.000000
75% 42900.000000
max 45060.000000

Figure 3. 2016 Sun Graph for Honolulu



Source:

Correlations were calculated and are shown in Figure 4. “Radiation” appears to be highly correlated with temperature and moderately correlated with “Pressure”. “Radiation” has a negative correlation of -0.23 with “Humidity” and a correlation of 0.12 with “Pressure”. There appears to be a small correlation between the “DayLength” and “Radiation”, however for reasons explained before “DayLength” was considered in the analysis. Wind direction and speed were not considered especially relevant for the purpose of predicting solar radiation. They may be more pertinent to windmill power generation. Therefore, they were dropped from the analysis. Scientifically, the sun affects the wind, but the wind does not affect the sun. Even though both are characteristics of local weather, they may not be predictors of radiation.

There are several timescales to consider for prediction, such as monthly, daily, and hourly. Radiation is known to be a volatile variable because many variables affect it. Seasonal weather changes such as autumn to winter will be accompanied by reduction in solar radiation. For this study, the granularity of the downloaded data was retained given the arguments presented earlier about electricity supply reliability and predictability. The granularity of the dataset is an interval level of approximately 5 minutes. Given the granularity of 5 minutes “WeekOfYear” which is at a higher aggregated level was dropped from the final prediction analysis. “WeekOfYear” can also be a proxy for seasonal variation though at a higher aggregation level. “DayLength” was considered to be sufficient proxy for the purpose of the analysis for the given granularity level. The final set of dependent variables for prediction in the data frame included “Temperature”, “Humidity”, “Pressure”, and “DayLength” (as a proxy for seasonal variation). The pattern of variation between radiation and the first three variables (referred to as features hereafter) were charted on a scatterplot and kdeplot. These are shown in Figure 5.

The scatterplots and kdeplots do not show any distinct relationship pattern, except for a slight relationship of “Radiation” with “Temperature”. This relationship is confirmed by a Pearson R-value of 0.70 shown in Figure 4. Further analysis of the relationships between the radiation and the features was performed. The variation of radiation and feature was charted on average hourly values for the entire period that the data was available. The variation was also charted as a weekly time series graph. This was done to study to get micro and macro view of variation of radiation in related to the feature being considered. Weekly variation through a time series chart was chosen to smoothen the pattern and the consequent ability to better discern any patterns, if any. Due to the geographical region where the data came from the weekly variations are not expected to be dramatic. Variations may be more drastic in other parts of the United States than in Hawaii which is in the equatorial region. In such cases a daily time series pattern chart may also be relevant. Month-to-month variation was considered too

broad to capture variation given that prediction is at 5-minute intervals. Thus, average daily variation and the week of the year variation was used to study the variation between the "Radiation" and other feature variables. These graphs are shown in figures 6, 7, and 8.

Figure 4. Correlations

Examining the mean variation between "Radiation" and the features during day and over each week for the complete time horizon, it was observed that when the "Temperature" increases, there is an increase in "Radiation" and vice versa both for mean daily variation and over weeks. "Humidity" with a Pearson R-value of -0.23, may be predicted to have some impact on radiation. The variation in "Humidity" and "Radiation" appears to have little relation on a daily level. However, for the weekly variation over the time horizon, there is a small negative relationship between the two. In addition, "Humidity" may serve as proxy for rain or cloud cover. "Pressure" appears to vary slightly in sync with radiation for the weekly graph, but not in the daily graph. It can be inferred that the "Temperature" variation would be a good predictor for variation in "Radiation". "Humidity" and "Pressure" were retained in the model due to a weak relationship in the weekly trend. The three features along with "DayLength" were used in the analysis for purposes of prediction of radiation at 5-minute interval duration.

Figure 5. Kdeplots and Scatterplots of features with radiation

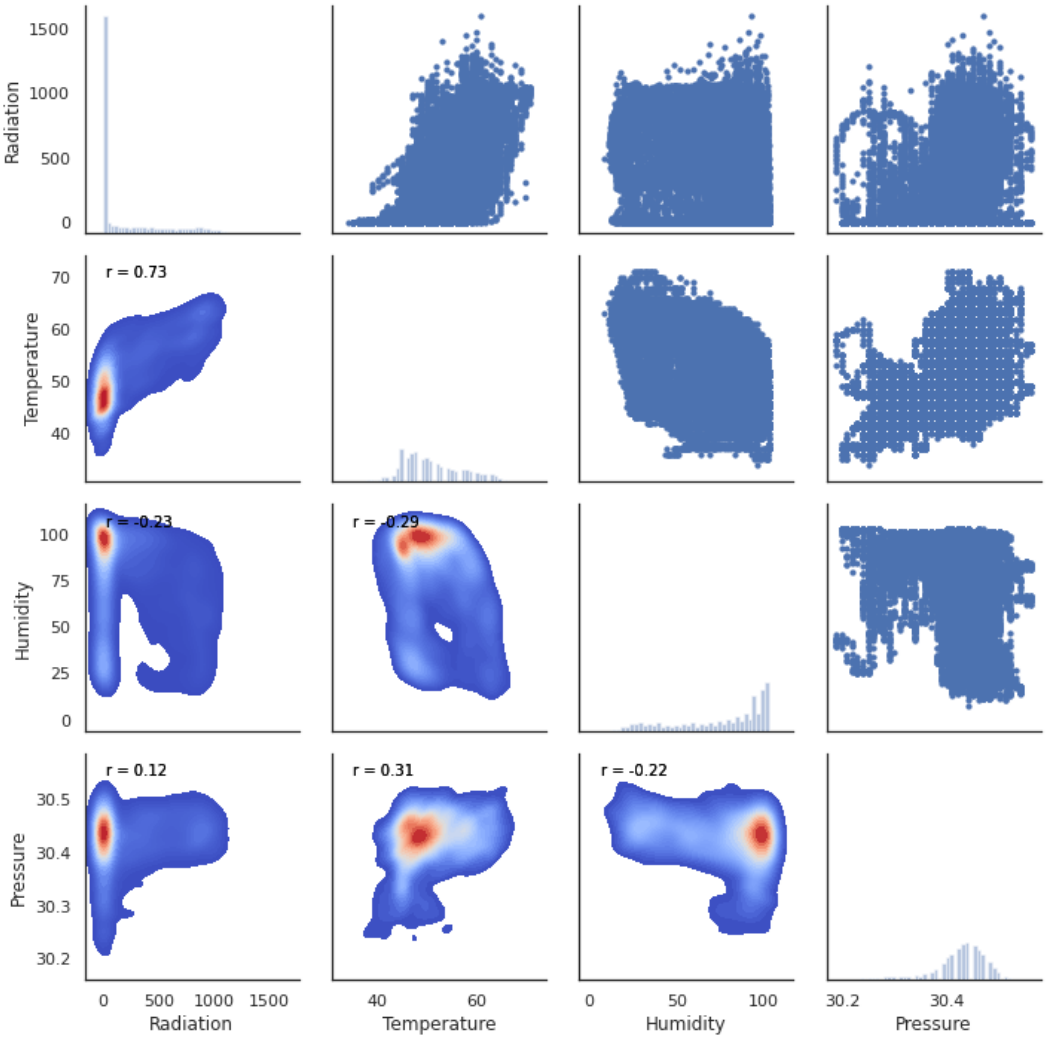


Figure 6. (a) Mean hourly variation of temperature and radiation; (b) Mean weekly variation of temperature and radiation

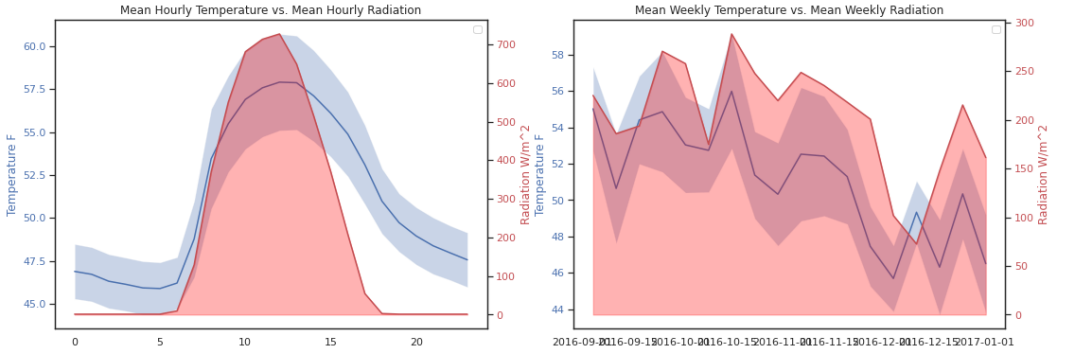


Figure 7. (a) Mean hourly variation of humidity and radiation; (b) Mean weekly variation of humidity and radiation

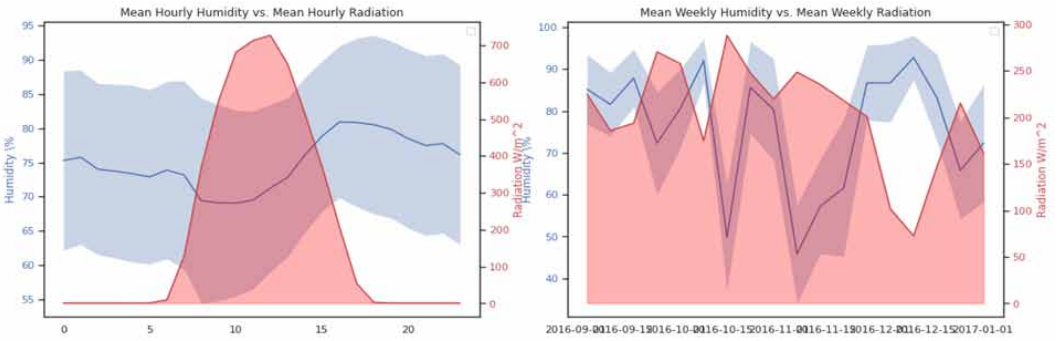
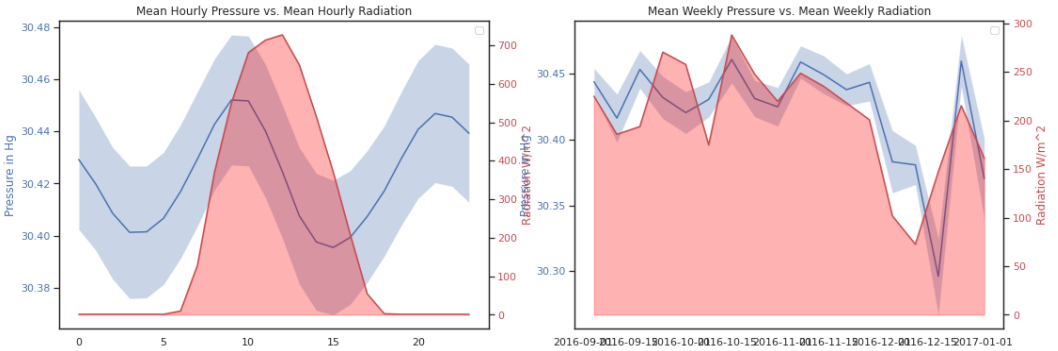


Figure 8. (a) Mean hourly variation of pressure and radiation; (b) Mean weekly variation of pressure and radiation



ANALYSIS AND RESULTS

Machine Learning Model Processing

Scikit-learn, which is a module available in Python, was used to run the four ML models of interest. Sci-kit is an open-source toolbox used to provide advanced tools to solve predictive data analysis problems. With a wide variety of protocols in place, calling each method using scikit-learn yielded the accuracy results for linear regression, MLP regressor, RFR, and SVR. The libraries in Scikit-learn allow for efficient and accurate data processing. Furthermore, the parameters of the algorithms used can be fine-tuned. Information about the use of the algorithms and fine-tuning is aided by an extensive Application Programming Interface (API) and documentation provided. Fine-tuning the parameters can lead to increased prediction accuracy, but too much fine-tuning can overfit the data to the noise or random variation accompanying the data.

The prediction models were run with "Radiation" as the dependent variable and "UNIXTIME", "Temperature", "Pressure", "Humidity", and "DayLength" as the independent variables. The data was split into a training set and a test set using the usual norm in the ML processing. 80% of the data was in the training set and the remaining 20% of the data was in the test set. This split was done randomly to prevent bias in the learning algorithms. Since the data is split randomly, the test data is not necessarily continuous over time, but rather a variation of points from the original dataset. This is not seen as a major limitation. If the prediction performs well over the random test data, then the likelihood of it performing well on continuous time data may be better. Linear regression, RFR,

MLP regression, and SVR were run on the training set. Data was fed into the function with defined parameters for each algorithm and it output a graph of the actual vs. prediction for both training and validation. Algorithm 1 in Figure 9 shows the general approach and example usage of the linear regression procedure to compute the output. The error statistics of Mean absolute error (MAE), Mean Squared Error (MSE), Median Absolute Error, R2, Mean Absolute Percentage Error (MAPE), and Mean Bias Error (MBE) were also computed and output along with graphs for training and validation. MBE is the average forecast error representing the systematic error of a forecast model to under or over forecast (Kato, 2016). If the model consistently under forecasts, then this value will be more negative and if it consistently over forecasts this value would be more positive. The MAE is the mean absolute error and is the average of the absolute difference between the forecasted and actual value. A low value is desired. The MSE or variance, gives more weight to the largest errors.

Algorithm

```
procedure Accuracy
 x ¬ Procedure Array
 y ¬ Label Array
 clf ¬ classifier
 define_train_model (x, y, clf)
 model ¬ classifier.fit
 acuracy ¬ classifier.score
 result ¬go(x, y, LinearRegression)
end
```

Figure 9. Linear regression algorithm using Sci-kit learn

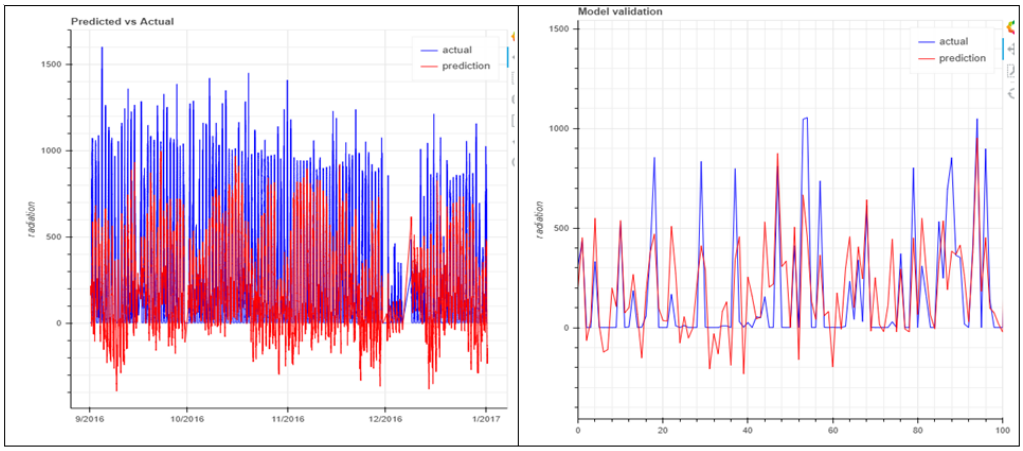
Algorithm

```
procedure Accuracy
    x ← Procedure Array
    y ← Label Array
    clf ← classifier
      define_train_model (x, y, clf)
         model ← classifier.fit
         acuracy ← classifier.score
       result ←go( x, y, LinearRegression)
end
```

Results

Results showed that various ML algorithms had different levels of accuracy as would be expected. The Random Forest Regressor (RFR) however showed the best accuracy of 95.50% with the time series analysis and the included features of "Temperature", "Humidity", "Pressure", and "DayLength". The RFR also performed the best of all the other error measures too. Accuracy is defined in terms of R^2 or percent of variance explained. Different error measures for different methods are detailed in Table 3. The models are trained on the training set and then forecasting is done on the test set. The difference

Figure 10. (a) Linear regression training; (b) Linear regression validation



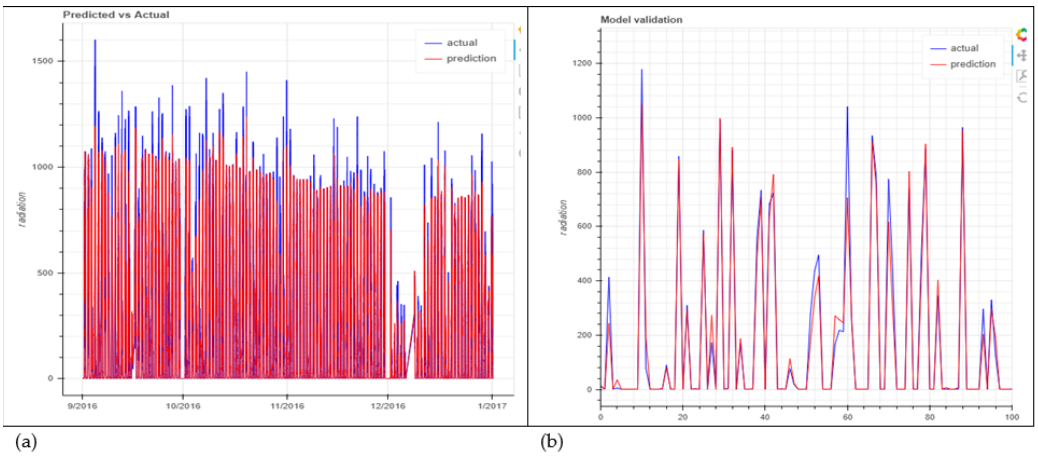
(a) (b)

between actual and forecasted values are used to calculate various error metrics. The training and validation charts along with accuracy are detailed in Figure 10, 11, 12 and 13.

Linear Regression

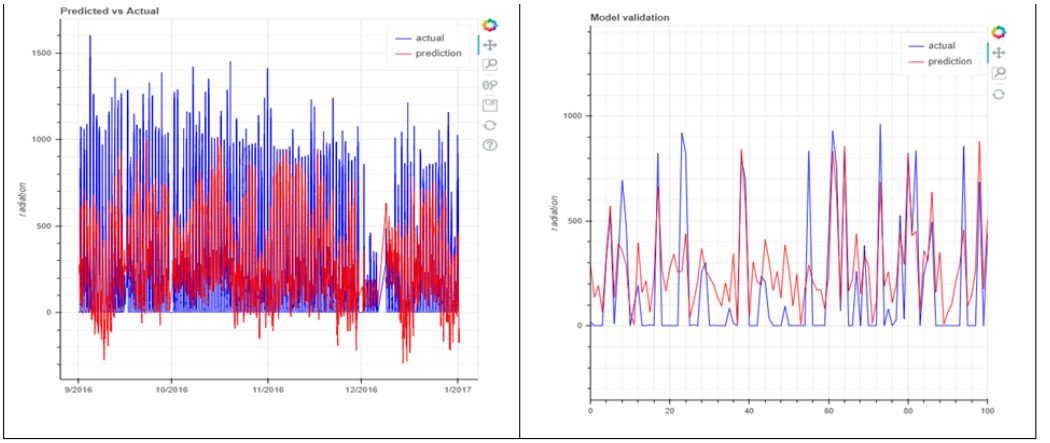
Linear regression has an R^2 of 58.68%. The training and validation charts are shown in Figure 9. Given the simple nature of linear regression this appears a reasonable performance. Other error metrics for the model are shown in Table 3. This accuracy of the model may be explained on account of the explanatory power of the "Temperature" for "Radiation" as was observed in the variation charts in Figures 6a and 6b.

Figure 11. (a) Random forest regression training; (b) Ransom forest regression validation



(a) (b)

Figure 12. (a) Multilayer perceptron regression training; (b) Multiplayer perceptron regression validation



(a) (b)

Random Forest Regressor

RFR has an accuracy of R2 of 95.50%. This accuracy is really good given that the features used are limited in number and optimizations were not done to keep the model simple. Due to the accuracy fine tuning of the RFR was not undertaken. Fine tuning of the control parameters may increase the accuracy through experimentation but given the high accuracy, it was determined that such an exercise was not needed. The training and validation charts are shown in Figure 11. Other metrics of performance are shown in Table 3.

MLP Regressor

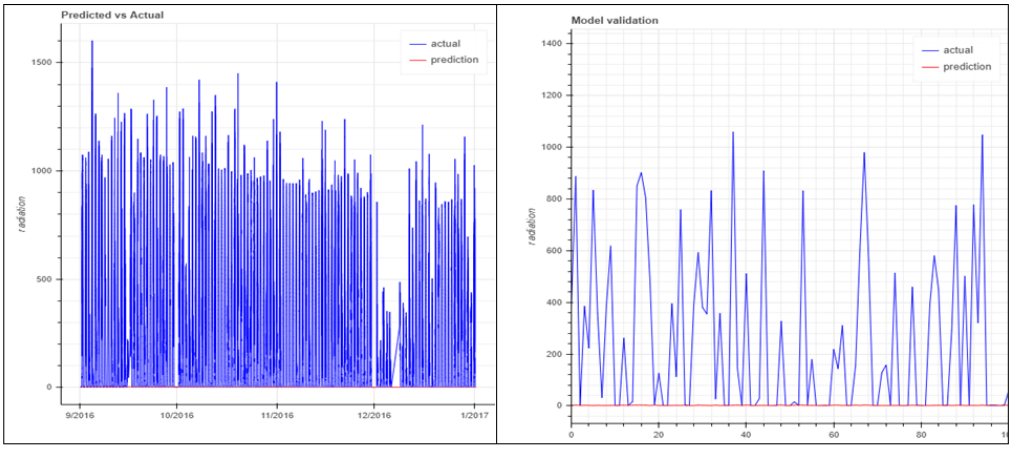
MLP Regressor has an accuracy of R2 of 57.80%. This is lesser that that of linear regression which is a much simpler and a less computationally complex algorithm. The training and validation charts are shown in Figure 12.

Support Vector Regression

Support vector machine has an R^2 of -42.05%. The accuracy is measured as the R-squared value. A negative R-squared implies that the model fits worse than a horizontal line and no variance is explained. In the case of solar radiation, which is time series regression, the support vector machine regression does not follow the trend of the data and may not be appropriate for this time series regression. For the analysis Radial Basis Function (RBF) kernel with default parameters was used. The implementation of SVM in sklearn learn is based on libsvm, where the fit time complexity is more than quadratic with the number of samples (Sklearn.svm.svr, n.d.). As such it makes SVM hard to scale to datasets with more than a couple of 10000 samples which was the case in this dataset. Linear and polynomial kernel SVM were run however they did not produce results in a span of about 3 hours. Given the purpose here is to forecast solar radiation using simple to use methods at consumer sites the SVM regression was considered to not add any value for prediction purposes. The training and validation charts are show in Figure 13.

A look at the error metrics demonstrates the superior performance of the RFR regression. The MBE of 0.07 indicates that there is no systematic bias in forecasting. The linear regression also compares favorably in this regard. RFR is also characterized by lowest MAE, MSE, Median Absolute Error, and MAPE.

Figure 13. (a) Support Vector Regression (RBF) training; (b) Support Vector Regression (RBF) validation



(a) (b)

Table 3. Error Metrics for different methods

Metric	Linear Regression	Random Forest Regressor	MLP Regressor	SVM (Radial Basis Kernel)
Mean absolute error	152.90	28.09	152.59	205.78
Mean Squared Error	41233.09	4494.60	42898.80	141769.27
Median absolute error	117.63	3.33	114.24	1.88
R^2 Score	58.68%	95.50%	57.80%	-42.05%
Mean absolute percentage error	43.06%	2.16%	40.44%	0.78%
Mean Bias Error	-0.24	0.07	27.87	204.94

Significance of "DayLength"

"DayLength" as discussed earlier was included as proxy for seasonal variation in spite of its low correlation with "Radiation" 0.073 (Figure 4). To assess if this "DayLength" was indeed acting as a proxy for seasonal variation the analyses with run with the four ML models without including "DayLength". The accuracy of prediction changed for all ML algorithms as one would expect. Linear regression accuracy dropped to 54.85%. The accuracy of the RFR had a significant drop and came out to 67.29%. The accuracy of MLP Regressor increased to 63.07% which is close to that of RFR. The accuracy of SVR increased significantly from a -43.44% to 44.40%. Further analyses with exclusion of different features are most likely to change accuracy further in different ways. These were not explored given the RFR accuracy of 93.42% is a very good accuracy and the model can still be considered as a simple model.

CONCLUSIONS, LIMITATIONS, AND FUTURE WORK

Based on the results of the prediction analysis, it appears that RFR provides the best accuracy for predicting solar radiation in a time series regression analysis for the data set under question that includes "Temperature", "Humidity", and "Pressure" features and "DayLength" as a proxy for seasonal variation. Further fine-tuning of the random forest regressor was not undertaken. The advantages of extensive tuning are not practical given the high accuracy of 93.42% for RFR. One may state with some confidence that historical data for this and surrounding locations can be fed into an RFR to predict the solar radiation for consequent grid planning, and battery support for the solar panels. This data on "Temperature", "Humidity", and "Pressure" can be obtained from consumer grade weather stations which capture the features used in this study or may even be compiled from the internet based on the latitude and longitude of a location. The "DayLength" can also be computed from data available at various weather sites. The simplicity and the feasibility of obtaining the meteorological measurements of temperature, humidity, and pressure from consumer grade weather stations, are the biggest advantages of this model and this work. Scripts can be written to automate the loading of data from a weather station mounted on a house or building into a text file and then process it to make a compatible .csv file that can be loaded into a Pandas DataFrame. Length of the day data can be added, and projections can then be done using the open-source, freely available tools employed in this research. Scripts can also be written to train the model quarterly to account for seasonal variations. Since the model building is not resource intensive this exercise can be done frequently on a consumer grade computer or workstation. Application in the equatorial regions would be beneficial, especially in areas closer to the equator since the variations in the features are not expected to be extensive and day length would be fairly stable. The need for retraining the model in such situations may be reduced. Such areas would likely benefit the most due to a constant supply of solar radiation throughout the year.

One limitation of this study is that the models were trained on data from one location. A more extensive exercise may be to use the same methodology within different equatorial regions. Given some evidence of efficacy of use of length of the day as proxy for seasonal variation the exercise may also be undertaken to regions north of the equatorial region. Comparing the results may lead to a better judgment of the efficacy of using a RFR with the temperature, pressure, and humidity features, and day length. There is always the possibility of an RFR model only fitting well to the particular data set analyzed. However, in the opinion of the authors, the model may be trained and localized to any location due to its simplicity and number of independent variables involved.

Future work may explore the use of other algorithms and modification of various parameters to increase accuracy. However, the accuracy obtained in the model for this study is sufficiently high. This may need to be balanced against the computational complexity of the various algorithms, the complexity of fine-tuning the model, and the possibility of overfitting the model. In general, one may prefer simple models with reasonable accuracy and complexity, over highly complex models. Such models can be run at more places including at homes, small business, and on the edge. Another idea for future work may be to include more variables in the model. Factors such as cloud cover, angle of the sun, and precipitation may lead to a more accurate, robust model. However, as explained earlier, the introduction of such meteorological variables may demand the use of more complex weather stations. This may put the predictive analysis outside the reach of simple households and small businesses. The simple model with temperature, humidity, pressure, and day length can also be implemented using IoT sensors for quick output of results. This could be done by using edge computing devices and enabling real-time analysis. Such implementations can lead to faster and more time-realistic predictions of solar radiation. Another important future work would be to run this exercise on dataset that is from different location however at the same latitude level and/or with similar sun graphs to provide validity to the model.

REFERENCES

Andrey. (2017). *Solar Radiation Prediction (Version 1)* [Data set]. Kaggle. https://www.kaggle.com/dronio/SolarEnergy

Bayindir, R., Gok, M., Kabalci, E., & Kaplan, O. (2011). An Intelligent Power Factor Correction Approach Based on Linear Regression and Ridge Regression Methods. *10th International Conference on Machine Learning and Applications and Workshops*, 2, 313-315. doi:10.1109/ICMLA.2011.34

Botzen, W. J., Nees, T., & Estrada, F. (2021). Temperature Effects on Electricity and Gas Consumption: Empirical Evidence from Mexico and Projections under Future Climate Conditions. *Sustainability*, *13*(1), 305. doi:10.3390/su13010305

Chen, L., Du, H., & Li, Y. (2019). Scoping Low-Cost Measures to Nowcast Sub-Hourly Solar Radiations for Buildings. *IOP Conference Series. Earth and Environmental Science*, *329*(1), 012041. doi:10.1088/1755-1315/329/1/012041

Copas, J. B. (1997). Using Regression Models for Prediction: Shrinkage and Regression to the Mean. *Statistical Methods in Medical Research*, *6*(2), 167–183. doi:10.1177/096228029700600206 PMID:9261914

El-Khattam, W., & Salama, M. M. (2004). Distributed Generation Technologies, Definitions and Benefits. *Electric Power Systems Research*, *71*(2), 119–128. doi:10.1016/j.epsr.2004.01.006

Friedlingstein, P., Jones, M. W., O'sullivan, M., Andrew, R. M., Hauck, J., Peters, G. P., Peters, W., Pongratz, J., Sitch, S., Le Quéré, C., Bakker, D. C. E., Canadell, J. G., Ciais, P., Jackson, R. B., Anthoni, P., Barbero, L., Bastos, A., Bastrikov, V., Becker, M., & Zaehle, S. et al. (2019). Global carbon budget 2019. *Earth System Science Data*, *11*(4), 1783–1838. doi:10.5194/essd-11-1783-2019

Friedlingstein, P., O'Sullivan, M., Jones, M. W., Andrew, R. M., Hauck, J., Olsen, A., Peters, G. P., Peters, W., Pongratz, J., Sitch, S., Le Quéré, C., Canadell, J. G., Ciais, P., Jackson, R. B., Alin, S., Aragão, L. E. O. C., Arneth, A., Arora, V., Bates, N. R., & Zaehle, S. et al. (2020). Global carbon budget 2020. *Earth System Science Data*, *12*(4), 3269–3340. doi:10.5194/essd-12-3269-2020

Gariepy, J. (1989). *Estimation de rayonment sloaire global. Internal Report, Service of Meterology*. Government of Quebec.

Halouani, N., Nguyen, C. T., & Vo-Ngoc, D. (1993). Calculation of monthly average global solar radiation on horizontal surfaces using daily hours of bright sunshine. *Solar Energy*, *50*(3), 247–258. doi:10.1016/0038-092X(93)90018-J

Hawai'i Space Exploration Analog and Simulation. (n.d.). *HI-SEAS*. https://hi-seas.org/

Hay, J. E. (1979). Calculation of Monthly Mean Solar Radiation for Horizontal and Inclined Surfaces. *Solar Energy*, *23*(4), 301–307. doi:10.1016/0038-092X(79)90123-3

Hocaoglu, F. O., & Serttas, F. (2017). A novel hybrid (Mycielski-Markov) model for hourly solar radiation forecasting, *Renewable Energy*, *108*, 635-643. 10.1016/j.renene.2016.08.058

IEA. (2020). *Solar PV*. IEA. https://www.iea.org/reports/solar-pv

Iqbal, M. (1979). Correlation of Average Diffuse and Beam Radiation with Hours of Bright Sunshine. *Solar Energy*, *23*(2), 169–173. doi:10.1016/0038-092X(79)90118-X

Islam, M. D., Kubo, I., Ohadi, M., & Alili, A. A. (2009). Measurement of Solar Energy Radiation in Abu Dhabi, UAE. *Applied Energy*, *86*(4), 511–515. doi:10.1016/j.apenergy.2008.07.012

Jolliffe, I. T. (2002). *Principal Component Analysis* (2nd ed.). Springer-Verlag. doi:10.1007/b98835

Kandirmaz, H. M., Kaba, K., & Avci, M. (2014). Estimation of monthly sunshine duration in Turkey using artificial neural networks. *International Journal of Photoenergy*, *2014*, 2014. doi:10.1155/2014/680596

Karasu, S., Altan, A., Sarac, Z., & Hacioglu, R. (2017). Prediction of solar radiation based on machine learning methods. *The Journal of Cognitive Systems*, *2*(1), 16-20.

Kato, T. (2016). Prediction of photovoltaic power generation output and network operation. In *Integration of Distributed Energy Resources in Power Systems* (pp. 77–108). Academic Press. doi:10.1016/B978-0-12-803212-1.00004-0

Khalyasmaa, A., Eroshenko, S. A., Chakravarthy, T. P., Gasi, V. G., Bollu, S. K. Y., Caire, R., & Karrolla, S. (2019). *Prediction of Solar Power Generation Based on Random Forest Regressor Model. In 2019 International Multi-Conference on Engineering, Computer and Information Sciences (SIBIRCON) (pp. 780-785)*. IEEE., doi:10.1109/SIBIRCON48586.2019.8958063

Khatib, T., Mohamed, A., Sopian, K., & Mahmoud, M. (2012). Assessment of Artificial Neural Networks for Hourly Solar Radiation Prediction. *International Journal of Photoenergy*. 10.1155/2012/946890

Khosravi, A., Koury, R. N. N., Machado, L., & Pabon, J. J. G. (2018). Prediction of hourly solar radiation in Abu Musa Island using machine learning algorithms. *Journal of Cleaner Production*, *176*, 63-75. 10.1016/j.jclepro.2017.12.065

Kumler, A., Xie, Y., & Zhang, Y. (2018). *A New Approach for Short-Term Solar Radiation Forecasting Using the Estimation of Cloud Fraction and Cloud Albedo* (No. NREL/TP-5D00-72290). National Renewable Energy Lab (NREL). <ALIGNMENT.qj></ALIGNMENT>10.2172/147644

Leonard, J. A., & Kramer, M. A. (1991). Radial Basis Function Networks for Classifying Process Faults. *IEEE Control Systems Magazine*, *11*(3), 31–38. doi:10.1109/37.75576

Levin, K. (2018). *New Global CO2 Emissions Numbers Are In. They're Not Good*. World Resources Institute. https://www.wri.org/blog/2018/12/new-global-co2-emissions-numbers-are-they-re-not-good

Lorenz, E., Ruiz-Arias, J., & Wilbert, S. (2017). Forecasting Solar Radiation. In Best Practices Handbook for the Collection and Use of Solar Resource Data for Solar Energy Applications: Second Edition. NREL (Technical Report, NREL/TP-5D00-68886).

Malik, U. (2018). *Random Forest Algorithm with Python and Scikit-Learn.* Stack Abuse. https://stackabuse.com/random-forest-algorithm-with-python-and-scikit-learn/

Matzarakis, A. P., & Katsoulis, V. D. (2006). Sunshine duration hours over the Greek region. *Theoretical and Applied Climatology*, *83*(1–4), 107–120. doi:10.1007/s00704-005-0158-8

Moreno-Munoz, A., De la Rosa, J. J. G., Posadillo, R., & Bellido, F. (2008). Very Short-Term Forecasting of Solar Radiation. In *33rd IEEE Photovoltaic Specialists Conference* (pp. 1-5). IEEE. doi:10.1109/PVSC.2008.4922587

Murali, A. (2013, December 29). *What is a Unix timestamp and why use it?* Stack Overflow. https://stackoverflow.com/questions/20822821/what-is-a-unix-timestamp-and-why-use-it

Okundamiya, M. S., & Nzeako, A. N. (2010). Empirical model for estimating global solar radiation on horizontal surfaces for selected cities in the six geopolitical zones in Nigeria. *Research Journal of Applied Sciences, Engineering and Technology*, *2*(8), 805–812.

Python 3.8.5 Documentation. (n.d.). *Datetime - Basic Date and Time Types.* https://docs.python.org/3/library/datetime.html

Qazi, A., Fayaz, H., Wadi, A., Raj, R. G., Rahim, N. A., & Khan, W. A. (2015). The artificial neural network for solar radiation prediction and designing solar systems: a systematic literature review. *Journal of Cleaner Production*, *104*, 1-12. 10.1016/j.jclepro.2015.04.041

Rietveld, M. R. (1978). A new method for estimating the regression coefficients in the formula relating solar radiation to sunshine. *Agricultural Meteorology*, *19*(2-3), 243–252. doi:10.1016/0002-1571(78)90014-6

Scikit. (n.d.). *Learn Sklearn.ensemble.RandomForestRegressor*. https://scikitlearn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html

Seber, G. A., & Lee, A. J. (2012). *Linear Regression Analysis* (Vol. 329). John Wiley & Sons.

SEIA. (n.d.). *Solar Industry Research Data*. Solar Energy Industries Associates. https://www.seia.org/solar-industry-research-data

Sharma, N., Sharma, P., Irwin, D., & Shenoy, P. (2011). Predicting Solar Generation from Weather Forecasts Using Machine Learning. In *2011 IEEE international Conference on Smart Grid Communications (SmartGridComm)* (pp. 528-533). IEEE. doi:10.1109/SmartGridComm.2011.6102379

Sklearn.svm.svr. (n.d.). https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVR.html#sklearn.svm.SVR

Sönnichsen, N. (2020, March 24). *Electricity Consumption Globally, 2017*. Statista. https://www.statista.com/statistics/280704/world-power-consumption/

Speiser, J. L., Miller, M. E., Tooze, J., & Ip, E. (2019). A Comparison of Random Forest Variable Selection Methods for Classification Prediction Modeling. *Expert Systems with Applications*, *134*, 93–101. doi:10.1016/j.eswa.2019.05.028 PMID:32968335

Stević, I. (2017). Strategic Orientation to Solar Energy Production and Long Term Financial Benefits. *Arhiv za tehničke nauke/Archives for Technical Sciences*, *1*(17), 1-12. . 10.7251/afts.2017.0917.001S

Suehrcke, R. H., Bowden, S., & Hollands, K. G. T. (2013). Relationship between sunshine duration and solar radiation. *Solar Energy*, *92*, 160–171. doi:10.1016/j.solener.2013.02.026

Wang, L., Kisi, O., Zounemat-Kermani, M., Salazar, G. A., Zhu, Z., & Gong, W. (2016). Solar Radiation Prediction Using Different Techniques: Model Evaluation and Comparison. *Renewable & Sustainable Energy Reviews*, *61*, 384–397. doi:10.1016/j.rser.2016.04.024

Wood Mackenzie. (2021). *US Solar Market Insight, Q3: 2021*. https://www.woodmac.com/reports/power-markets-u-s-solar-market-insight-q3-2021-526090/

Xiongwen, Z. (2014). A statistical approach for sub-hourly solar radiation reconstruction. *Renewable Energy*, *71*, 307-314. 10.1016/j.renene.2014.05.038

Yadav, A. K., & Chandel, S. S. (2014). Solar radiation prediction using Artificial Neural Network techniques: A review. *Renewable and Sustainable Energy Reviews*, *33*, 772-781. 10.1016/j.rser.2013.08.055

Yin, X. (1999). Bright sunshine duration in relation to precipitation, air temperature and geographic location. *Theoretical and Applied Climatology*, *64*(1-2), 61–68. doi:10.1007/s007040050111

Pankaj Chaudhary is Associate Professor Business Information Systems and Analytics at North Carolina A&T University.

Soundarajan Ezekiel is a Professor of Computer Science at IUP.

James A. Rodger is a Professor of Healthcare and Information Systems at Slippery Rock University of Pennsylvania (SRU). He received his Doctorate in MIS from Southern Illinois University at Carbondale in 1997. Dr. Rodger has published several journal articles related to these subjects. His work has appeared in the following journals: Computers in Human Behavior, Journal of Computer Information Systems, Issues in Information Systems, IEEE Transactions on Software Engineering, International Journal of Hybrid Intelligent Systems, Information and Software Technology, Information Technology and Management. Annals of Operations Research, Communications of ACM, Computers & Operations Research, Decision Support Systems, Expert Systems with Applications, Lecture Notes in Computer Science, International Journal of Human-Computer Studies as well as several other journals.