Solar Energy Output Forecasting from SolarGIS Data for Connected Grid Station

Submitted by

Saradindu Sengupta

(Reg no: 91616012)

In partial fulfillment of the requirements for the award of
Master of Science in Computer Science with
Specialization in Machine Intelligence

COCHIN UNIVERSITY OF SCIENCE AND TECHNOLOGY, KOCHI

Conducted by

Indian Institute of Information Technology and Management-Kerala
Technopark Campus
Thiruvanathapuram-695 581

May 2018

Declaration

I, *Saradindu Sengupta, MSMI-1613*, hereby declare that this report is substantially the result of my own work, except, where explicitly indicated in the text and has been carried out during the period December 2017-April 2018.

Place: Thiruvananthapuram, Kerala

Date: 17.05.2018

Saradindu Sengupta

Acknowledgment

I would first like to thank my thesis advisor Dr. Asharaf S of the Indian Institute of Information Technology & Management , Kerala . The door to Prof. Asharaf S's office was always open whenever I ran into a trouble spot or had a question about my research or writing. He consistently allowed this paper to be my own work, but steered me in the right the direction whenever he thought I needed it.

I would also like to thank the experts who were involved in the validation survey for this research project. Without their passionate participation and input, the validation survey could not have been successfully conducted.

Finally, I must express my very profound gratitude to my parents for providing me with unfailing support and continuous encouragement throughout my years of study and through the process of researching and writing this thesis. This accomplishment would not have been possible without them.

Thank you.

Place: Thiruvananthapuram, Kerala

Date:17.05.2018

Saradindu Sengupta

Abstract

Using random forest regression method, daily mean solar output generation can yield promising result rather than conventional NWP model for forecasting. Using that in practice also the goal was to create a user-friendly application , with easy access, to provide accurate forecasting regarding saving and conservation.This goal has been accomplished in three stages:

- Building a machine learning model that predicts the annual energy production of a prospective solar installation.
- Building a model that predicts installation cost.
- Implementing these models on a user-friendly web app that shows users how much they should expect to save on their energy bill each year by switching to solar.

The random forest model out-performs its other rivals and also conventional models , thus providing a better suited model to run with for forecasting.

Table of Contents

LIST OF TABLES

Table

LIST OF FIGURES

Figure

CHAPTER I

INTRODUCTION

An electrical operator should ensure a precise balance between the electricity production and consumption at any moment. This is often very difficult to maintain with conventional and controllable energy production system, mainly in small or not interconnected (isolated) electrical grid (as found in islands). Many countries nowadays consider using renewable energy sources into their electricity grid. Photo voltaic systems are becoming important sources of energy in electricity networks. However, electric utility companies are required to guarantee electricity supply within certain ranges which is difficult given the fluctuating nature of weather conditions.

With the dwindling fossil fuel resources, research in renewable energy has gained significant impetus. The leading source for renewable energy is solar power generation mainly through photo-voltaic cells. Advantages of using solar energy include its immunity to imitative circumstances like the oil prices, a clean source of energy, and reduction of imports and dependability on external resources. Though photo-voltaic cells are considered as a major source for future energy generation, their return on investment and upfront cost is hindering their deployments. One of the reasons for this is lack of predictable supply because of changing weather conditions. Since photo-voltaic cells generate electricity by converting solar energy to electric current, the amount of solar energy being provided in a day is very important to size the photo-voltaic system. Therefore, the amount of electricity produced depends upon solar irradiance in a particular day, which itself depends on various parameters such as location, time, and weather patterns. Solar irradiance is defined as the power per unit area received from the Sun in the form of electromagnetic radiation in the wavelength range of the solar cell being used. Unfavorable weather reduces the output of the solar plant to a large extent. Therefore, in order to fulfill the energy requirements, a power supply company needs to supplement the remaining amount by purchasing power from different power generation companies running on costlier fossil fuels. The power rates charged by these companies not only depend upon the amount of the power required but also on the timeliness of the order. A timely order placement with these companies not only helps to meet the promised power supply goals, but also helps reduce the cost. Hence prior knowledge of the power produced plays a crucial role in maintaining quality of service and reducing cost. With solar power, its possible to predict the production knowing current and the past information about the weather and the irradiance. Various researchers have proposed forecasting mechanisms with good results however a room for improvement still exists. There are two orthogonal avenues of improvement in this domain, one is more efficient algorithm design for forecasting and second is identification and quantification of the effect of parameters on forecast. In this paper we attempt to improve the state of the art in both the dimensions.

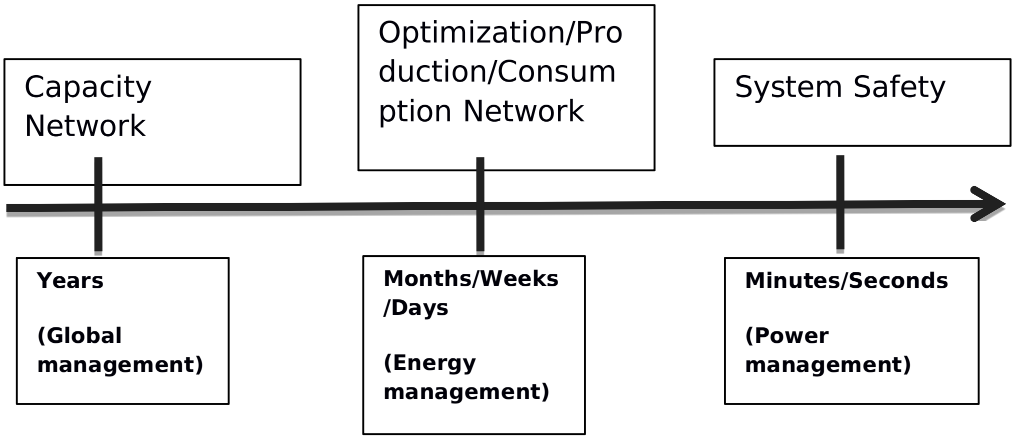


Fig: 1 Prediction scale for energy management in an electrical network

1.1 The necessity to predict solar radiation or solar production

One of the most important challenge for the near future global energy supply will be the large integration of renewable energy sources (particularly non-predictable ones as wind and solar) into existing or future energy supply structure. An electrical operator should ensure a precise balance between the electricity production and consumption at any moment. As a matter of fact, the operator has often some difficulties to maintain this balance with conventional and controllable energy production system, mainly in small or not interconnected (isolated) electrical grid (as found in islands). The reliability of the electrical system then become dependent on the ability of the system to accommodate expected and unexpected changes (in production and consumption) and disturbances, 5 while maintaining quality and continuity of service to the customers. Then, the energy supplier must manage the system with various temporal horizons. The integration of renewable energy into an electrical network intensifies the complexity of the grid 5 management and the continuity of the production/consumption balance due to their intermittent and unpredictable nature [1, 2]. The intermittent and the non-controllable characteristics of the solar production bring a number of other problems such as voltage fluctuations, local power quality and stability issues . Thus forecasting the output power of solar systems is required for the effective operation of the power grid or for the optimal management of the energy fluxes occurring into the solar system . It is also necessary for estimating the reserves, for scheduling the power system, for congestion management, for the optimal management of the storage with the stochastic production and for trading the produced power in the electricity market and finally to achieve a reduction of the costs of electricity production. Due to the substantial increase of solar power generation the prediction of solar yields becomes more and more important. In order to avoid large variations in renewable electricity production it is necessary to include also the complete prediction of system operation with storage solutions. Various storage systems are being developed and they are a viable solution for absorbing the excess power and energy produced by such systems (and releasing it in peak consumption periods), for bringing very short fluctuations and for maintaining the continuity of the power quality.These storage options are usually classified into three categories:

- Bulk energy storage or energy management storage media is used to decouple the timing of generation and consumption.

- Distributed generation or bridging power - this method is used for peaks shaving - the storage is used for a few minutes to a few hours to assure the continuity of service during the energy sources modification.

- The power quality storage with a time scale of about several seconds is used only to assure the continuity of the end use power quality.

Category	**Discharge power (MW)**	**Discharge Time**	**Stored Energy**	**Representative Application**
Bulk energy	10 - 1000	1 - 8 H	10- 8000 MWh	Load leveling, generation capacity
Distributed generation	0.1 - 2	0.5 - 4 H	50 – 8000 kWh	Peak shaving, transmission deferral
Power quality	0.1 - 2	1 - 30 S	0.03 – 16.7 kWh	End-use power quality /reliability

Table 1. Specification of various storage options

Table 1 shows these three categories and their technical specifications. As shown, every type of storage is used in different cases to solve different problems, with different time horizon and quantities of energy.Table 1 shows that the electricity storage can be widely used in a lot of cases and applications as a function of the time of use and the power needs of the final user. Finally, it shows that the energy storage acts at various time levels and their appropriate management requires the knowledge of the power or energy produced by the solar system at various horizons: very short or short for power quality category to hourly or daily for bulk energy storage. Similarly, the electrical operator needs to know the future production (Figure 1) at various time horizons from one to three days, for preparing the production system and to some hours or minutes for planning the start-up of power plants (Table 2). Starting a power plant needs between 5 min for a hydraulic one to 40 hours for a nuclear one. 9 Moreover, the rise in power of the electrical plants is sometimes low, thus for an effective balance between production and consumption an increase of the power or a starting of a new production needs to be anticipated sometimes well in advance.

Type of electrical generator	Power size MW	Minimum power capacity percentage of peak power	Rise speed in power per min percentage of peak power	Starting time Hour
Nuclear Power Plant	400–1300 per reactor	20%	1%	40 h (cold)-18 h
Steam thermal plant	200–800 per turbine	50%	0.5%-5%	11-20 h (cold)-5 h
Fossil-fired power plants	1–200	50-80%	10%	10 min-1 h
Combined-cycle plant	100–400	50%	7%	1-4 h
Hydro power plant	50–1300	30%	80%-100%	5 min
Hydro power plant	25	30%	30%	15-20 min
Internal combustion engine	20	65%	20%	45-60 min

Table 2. Characteristics of electricity production plants

Furthermore, the relevant horizons of forecast can and must range from 5 minutes to several days as it 9 was confirmed by Diagne et al. [6]. Elliston and MacGill [10] outlined the reasons to predict solar radiation for various solar systems (PV, thermal, concentrating solar thermal plant, etc.) insisting on the forecasting horizon. It therefore seems apparent that the time-step of the predicted data may vary depending on the objectives and the forecasting horizon. All these reasons show the importance of forecasting, whether in production or in consumption of energy. The need for forecasting lead to the necessity to use effective forecasting models. In the next section the various available forecasting methodologies are presented.

1.2 Related Work

The current eras' accelerating advances in all fields rely on the constant supply of the electric power. The fossil fuels are key source of the reliable and consistent energy provision, but they have high cost associated with them, emit dangerous gases, and are subject to uncertainty of the international oil prices. On the other hand, solar power is cheap, clean source of energy that can be produced by every country, but its dependence on weather conditions make it less reliable. To cope with this unreliability the prediction of the solar power contributes towards a consistent supply. The prediction of solar power is a multidisciplinary research that needs contribution from meteorology, solar cell engineering, electrical engineering, and machine learning computation. used the radial basis function (RBF) to model the solar radiation based on sunshine duration and air temperature data. They use Multi-Layer Perceptron (MLP) to model the hourly forecast of the solar radiation using present values of air temperature and mean daily solar irradiance. [2] compare different models including Fuzzy Logic, Neural Networks and Autoregressive to predict the half daily values of solar radiation. In [3], 3-day forecast of the solar irradiance has been modeled using the forecast data provided by European Center for Medium-Range Weather Forecasts (ECMWF). In [4], different weather parameters such as solar hours, latitude, longitude, elevation, maximum and minimum air temperature, humidity, and rainfall has been used to predict the solar power in Indonesia. It also used ANN to generate a model that could predict the solar power of a particular region.A recent addition is the use of computer algorithm based on fuzzy logic control (FLC) to estimate the wind and solar energies in a hybrid renewable energy system from natural factors [5]. The solar power was estimated using the temperature and the lighting as input parameters. A detail introduction to the current research on forecasting solar irradiance is presented in [6]. It is a in-depth review to facilitate selection of the appropriate forecast method according to needs. They also comment on the statistical approaches and techniques based on images from satellite imagery. They also discuss numerical weather prediction (NWP) and hybrid models. Online forecasting of power production from PV systems is discussed in [7], to predict hourly values of solar power for horizons of up to 36 h. A two-stage method of using adaptive linear time series models for clear sky model and autoregressive (AR) models NWPs is elaborated. [8] compare multiple regression techniques for generating prediction models for solar power, including linear least squares, and support vector machines using multiple kernel functions.

1.3 Available Forecasting Methodologies

1.3.1 NWP Model

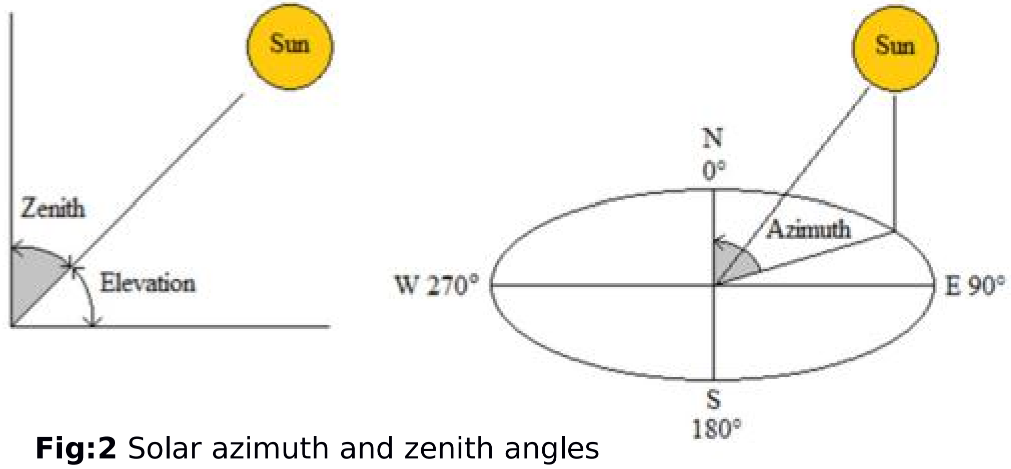
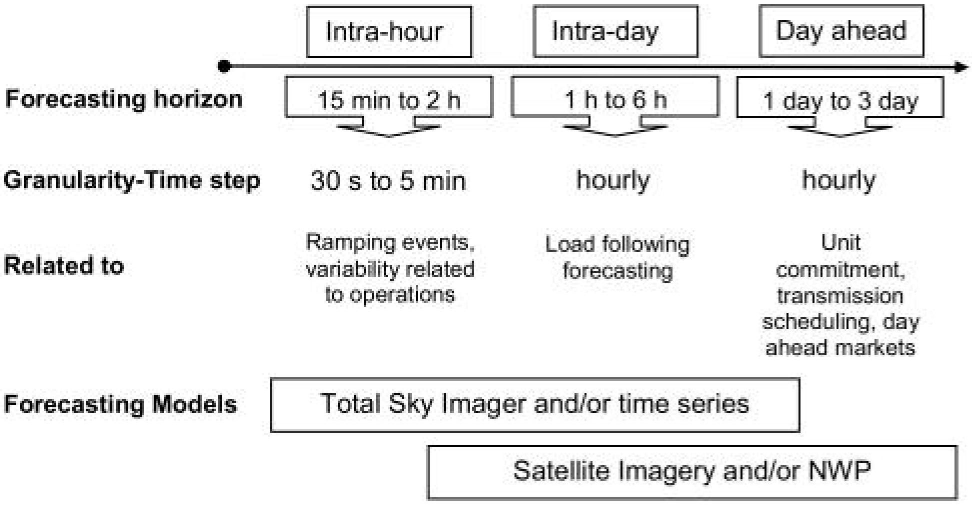


Fig:2 Solar azimuth and zenith angles

The solar power forecasting can be performed by several methods; the two big categories are the cloud imagery combined with physical models, and the machine learning models. The choice for the method to be used depends mainly on the prediction horizon; actually all the models have not the same accuracy in terms of the horizon used. Various approaches exist to forecast solar irradiance depending on the target forecasting time. The literature classifies these methods in two classes of techniques:

- Extrapolation and statistical processes using satellite images or measurements on the ground level and sky images are generally suitable for short-term forecasts up to six hours. This class can be divided in two sub-classes, in the very short time domain called “Now- casting’’ (0–3 h), the forecast has to be based on extrapolations of real-time measurements [5]; in the Short-Term Forecasting (3–6 h), Numerical Weather Prediction (NWP) models are coupled with post-processing modules in combination with real-time measurements or satellite data [5, 11].

- NWP models able to forecast up to two days ahead or beyond [12, 13] (up to 6 days ahead [13]). These NWP models are sometimes combined with post-processing modules and satellite information are often used [2].


Intra-hour
Intra-day
Day ahead
Forecasting horizon
15 min to 2 h
1 h to 6 h
1 day to 3 day
Granularity-Time step
30 s to 5 min
hourly
hourly
Related to
Ramping events, variability related to operations
Load following forecasting
Unit commitment, transmission scheduling, day ahead markets
Forecasting Models
Total Sky Imager and/or time series
Satellite Imagery and/or NWP

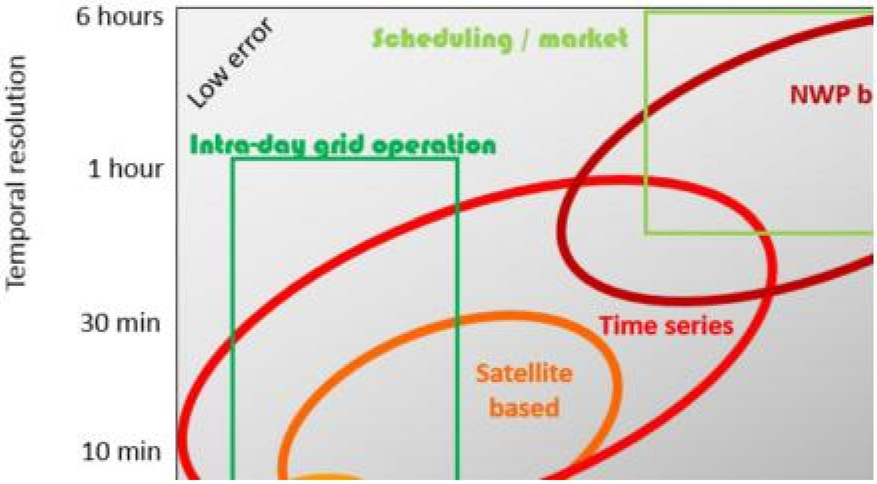


Fig: 3 a) Forecasting error versus forecasting models (left) [6,14]. b) Relation between forecasting horizons, forecasting models and the related activities (right) [6, 14]

The NWP models predict the probability of local cloud formation and then predict indirectly the transmitted radiation using a dynamic atmosphere model. The extrapolation or statistical models analyze historical time series of global irradiation, from satellite remote sensing [15] or ground measurements [16] by estimating the motion of clouds and project their impact in the future [6, 13, 17]. Hybrid methods can improve some aspects of all of these methods [6, 14]. The statistical approach allows to forecast hourly solar irradiation (or at a lower time step) and NWP models use explanatory variables (mainly cloud motion and direction derived from atmosphere) to predict global irradiation N-steps ahead [15]. Very good overviews of the forecasting methods, with their limitations and accuracy can be found in [1, 5, 6, 10, 12, 14, 18]. Bench-marking studies were performed to assess comparisons and the accuracy of the results produced, as shown in this paper, must be carefully evaluated in selecting the right method to use. As part of COST Action ES1002. (European Cooperation in Science and Technology) [22] on Weather Intelligence for Renewable Energies (WIRE) a literature review on the forecasting accuracy applied to renewable energy systems mainly solar and wind is carried out. In this paper an overview on the various methodologies available for solar radiation prediction based on machine learning is presented. A lot of review papers are available, but it is very rare to find a paper which is totally dedicated to the machine learning methods and that some recent prediction models like random forest, boosting or regression tree be integrated. In the next section the different methodologies used in the literature to predict global radiation and the parameters used for estimating the model performances are presented.

1.4 Proposed Methodology

Instead of using NWP model for weather simulation and solar power, irradiance forecasting, the proposed method uses Random Forrest Regression method as a bench-mark for daily mean power forecasting.

The final model used for production was a random forest regression model, which has several benefits in this context. It is a non-parametric model, which means it can predict a variable that is non-normally distributed. Because there are a wide range of solar panels in the OpenPV dataset, with some utility-scale installations producing thousands of times more energy per year than small, residential panels, the data is very positively skewed. This means that in order to use a model like linear regression without having biased results, it would be necessary to log transform the data, or use a generalized linear model like a Poisson regression. Random forest models also do well with categorical features, and in this case there were a few such features, including technology type and tracking type.

CHAPTER II

Literature Review

1. Direct Normal Irradiance (DNI):

Direct Normal Irradiance (DNI) is the amount of solar radiation received per unit area by a surface that is always held perpendicular (or normal) to the rays that come in a straight line from the direction of the sun at its current position in the sky. Typically, you can maximize the amount of irradiance annually received by a surface by keeping it normal to incoming radiation. This quantity is of particular interest to concentrating solar thermal installations and installations that track the position of the sun.

2. Diffuse Horizontal Irradiance (DHI):

Diffuse Horizontal Irradiance (DHI) is the amount of radiation received per unit area by a surface (not subject to any shade or shadow) that does not arrive on a direct path from the sun, but has been scattered by molecules and particles in the atmosphere and comes equally from all directions.

3. Direct Radiation:

Direct radiation is also sometimes called "beam radiation" or "direct beam radiation". It is used to describe solar radiation traveling on a straight line from the sun down to the surface of the earth.

4. Diffuse radiation:

Diffuse radiation, on the other hand, describes the sunlight that has been scattered by molecules and particles in the atmosphere but that has still made it down to the surface of the earth.However, since diffuse radiation is generally pretty equally distributed throughout the sky, the most diffuse radiation is gathered when your solar panels are laying down horizontally.

The steeper your solar panels are tilted, the less of the sky they are facing and the more of the sky's diffuse radiation they miss out on. If, for example, your solar collectors are tilted at a 45° angle, they are facing away from about a quarter of the sky and would only collect about three-fourths of the diffuse radiation in the sky.

Still, because direct radiation is much more intense than diffuse radiation, the amount of radiation missed by tilted solar panels is generally more than compensated for by the extra radiation gained by tracking the sun.

5. Global Horizontal Irradiance (GHI):

Global Horizontal Irradiance is the total amount of shortwave radiation received from above by a surface horizontal to the ground. This value is of particular interest to photovoltaic installations and includes both Direct Normal Irradiance (DNI) and Diffuse Horizontal Irradiance (DHI).

Global Horizontal (GHI) = Direct Normal (DNI) X cos(θ) + Diffuse Horizontal (DHI)

6. Random Forest:

The Random Forest is one of the most effective machine learning models for predictive analytic, making it an industrial workhorse for machine learning.

Background:

The random forest model is a type of additive model that makes predictions by combining decisions from a sequence of base models. More formally we can write this class of models as:

g(x)=f0(x)+f1(x)+f2(x)+...

where the final model g is the sum of simple base models fi. Here, each base classifier is a simple decision tree. This broad technique of using multiple models to obtain better predictive performance is called model ensembling. In random forests, all the base models are constructed independently using a different sub-sample of the data.

7. NWP Model:

Numerical Weather Prediction (NWP) data are the form of weather model data we are most familiar with on a day-to-day basis. NWP focuses on taking current observations of weather and processing these data with computer models to forecast the future state of weather. Knowing the current state of the weather is just as important as the numerical computer models processing the data. Current weather observations serve as input to the numerical computer models through a process known as data assimilation to produce outputs of temperature, precipitation, and hundreds of other meteorological elements from the oceans to the top of the atmosphere.

8. Azimuth Angle:

The azimuth angle is the compass direction from which the sunlight is coming. At solar noon, the sun is always directly south in the northern hemisphere and directly north in the southern hemisphere. The azimuth angle varies throughout the day as shown in the animation below. At the equinoxes, the sun rises directly east and sets directly west regardless of the latitude, thus making the azimuth angles 90° at sunrise and 270° at sunset. In general however, the azimuth angle varies with the latitude and time of year.

9. Elastic Net Regression:

Elastic net regularization method includes both LASSO (L1) and Ridge (L2) regularization methods.

Overfitting : The core idea behind machine learning algorithms is to build models that can find the generalised trends within the data. However, if no measures are taken, sometimes the models tend to rote learn the data instead of learning the patterns. During such cases, although the model fits well to the training data (model yields accurate results when evaluated on training data), however, it evaluated poorly on the test data. This is called overfit.

Regularization is used to prevent overfitting the model to training data. This is achieved by slightly perturbing (adding noise) the objective function of the model before optimizing it (optimising a model means to find the model parameters w* such that the argmin /argmax of the objective function is found- in other words, it is to find the global optima of the objective function) . In L1 Regularisation, a noise of magnitude lambda .|w*| is added while in L2 Regularisation, noise of magnitude lambda.|w*|. |w*| is added. where |w*| is the magnitude of the optimal parameter vector.

10. **Pyranometer:**

A pyranometer is a type of actinometer used for measuring solar irradiance on a planar surface and it is designed to measure the solar radiation flux density (W/m2) from the hemisphere above within a wavelength range 0.3 μm to 3 μm. A typical pyranometer does not require any power to operate. However, recent technical development includes use of electronics in pyranometers, which do require (low) external power.

CHAPTER III

Materials & Methodologies

In global horizontal irradiance forecasting the models can be used in three different ways [24]:

- structural models which are based on other meteorological and geographical parameters;

- time-series models which only consider the historically observed data of solar irradiance as input features (endogenous forecasting);

- hybrid models which consider both, solar irradiance and other variables as exogenous variables (exogenous forecasting). The proposed method uses the last mentioned way,i.e., hybrid method.

1.1 Materials

1.1.1 Data Sources and Viability

Data for this project came from three sources, both managed by the NREL. The first is The OpenPV Project, which contains data related to over one million solar panel installations across the U.S. This dataset includes the following:

1. Annual energy production
2. Installation cost
3. Size
4. Orientation
5. Tilt
6. Installer
7. Technology type

etc.

The second dataset comes from the National Solar Radiation Database (NSRDB) API. This dataset includes hourly measures of:

8. Radiation
9. Temperature
10. Wind speed
11. Position of the sun

The third dataset comes from CleantechSolar Energy Corporation Pvt Ltd(R) ‘ s own rooftop solar plant from Singapore and US west coast. The data at Cleantech Solar’s sites are recorded using proprietary data logger provided by SERIS(Solar Energy Research Institute , Singapore, NUS) which contains data in the following way

12. Global Horizontal Irradiance
13. Global Tilted Irradiance
14. Total Power pushed to the Grid
15. Humidity
16. Wind Speed

As a failsafe process to ensure unhindered availability of data, satellite data also use here provided as commercial use by SolarGIS which shows as follows:

17. Global Diffusion Irradiance
18. Humidity
19. Wind Speed

1.1.1.1 Data Source Description

The data provided by NREL is in plain CSV file format provided according to ZIP code and contains over one million solar system data logged on bi-monthly basis.

The data provided by NSRDB is in form of API and can be accessed on daily basis at once. This data contains mainly sensor data rather than actual

system output, load generation and grid pumping data.
The third data provided by CleantechSolar Energy Corporation Pvt Ltd comes from company owned cloud storage which contains time series data for sensors and electric power generation as well.

1.1.2 Data Collection

The data for NREL's OpenPv project came as a simple CSV download-able from their website here [https://openpv.nrel.gov/search], accessible by ZIP code in amount of millions of solar power station installation of either rooftop or industrial.

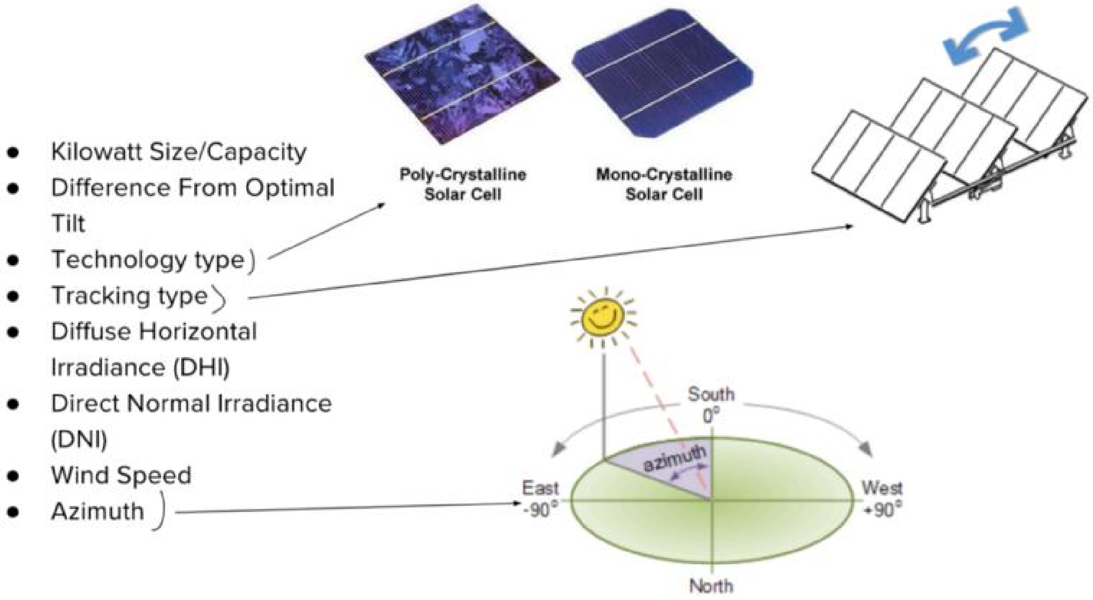
The NSRDB API only allows one thousand daily queries, so in order to gather local radiation data for the roughly fifteen thousand ZIP codes in the OpenPV dataset, used a python script, set it to run every 24 hours, and deployed it on a remote Amazon Web Services EC2 instance. The script pulled hourly data from all of 2017 (the most recent data available), averaged it for the year, and saved it to a local Neo4J database. Once pulled down radiation data for all one million solar panel installations, merged the two datasets according to ZIP code, the most granular location measure available.

The data from CleantechSolar Energy Corporation Pvt Ltd, came from Digital Ocean hosted server where every station's data comes on certain time of the day.

1.2 Methodologies

1.2.1 Parameters

Based on exploratory analysis, the following variables have been chosen to build a model for annual energy production.Direct Normal Irradiance (DNI) is the amount of radiation that travels directly from the sun to the earth, whereas Diffuse Horizontal Irradiance (DHI) is the amount of radiation that reflects off particles in the air before hitting the surface of the earth. Diffuse irradiance is therefore higher in places with more cloud cover or more dust in the air to block the travel of sunlight to the earth.



- Kilowatt Size/Capacity
- Difference From Optimal Tilt
- Technology type
- Tracking type
- Diffuse Horizontal Irradiance (DHI)
- Direct Normal Irradiance (DNI)
- Wind Speed
- Azimuth

In the northern hemisphere, the further north an installation is located, the more it ought to be tilted. The optimal tilt difference is simply the difference between the tilt of a panel and its latitude. Technology type refers to what type of silicon is used in the panel. Some panels move to track the sun, and tracking type refers to whether a panel is fixed or has some kind of tracking.

1.2.2 Exploring the Data

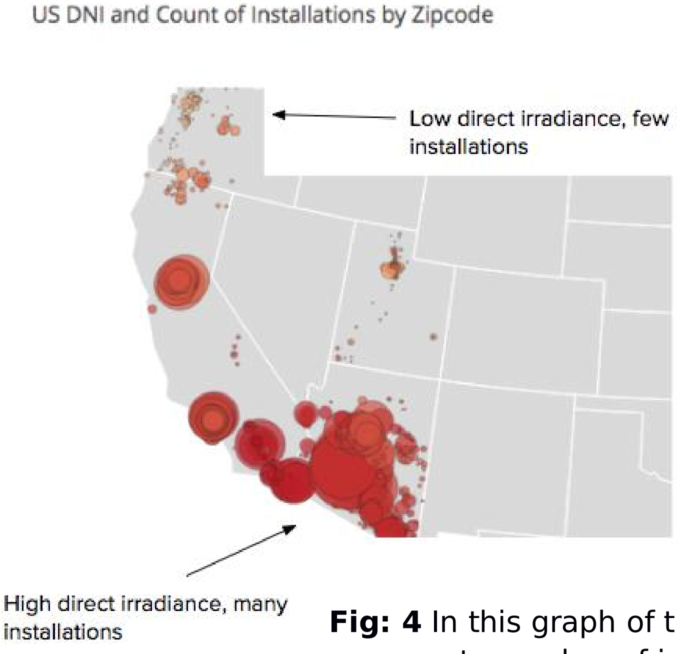

US DNI and Count of Installations by Zipcode
Low direct irradiance, few installations
High direct irradiance, many installations

Fig: 4 In this graph of the western U.S., size represents number of installations per ZIP code, and color represents amount of direct radiation.

Though most solar installations are in places where direct irradiance (DNI) is high, it appears the size of an installation actually plays the biggest role in determining a panel’s energy output. Size is highly correlated with annual energy production. When we look at a scatter plot of size and annual production, the linear relationship between the two is clear. The larger an installation, the higher its capacity for transforming solar radiation into electricity, and thus the observed relationship.

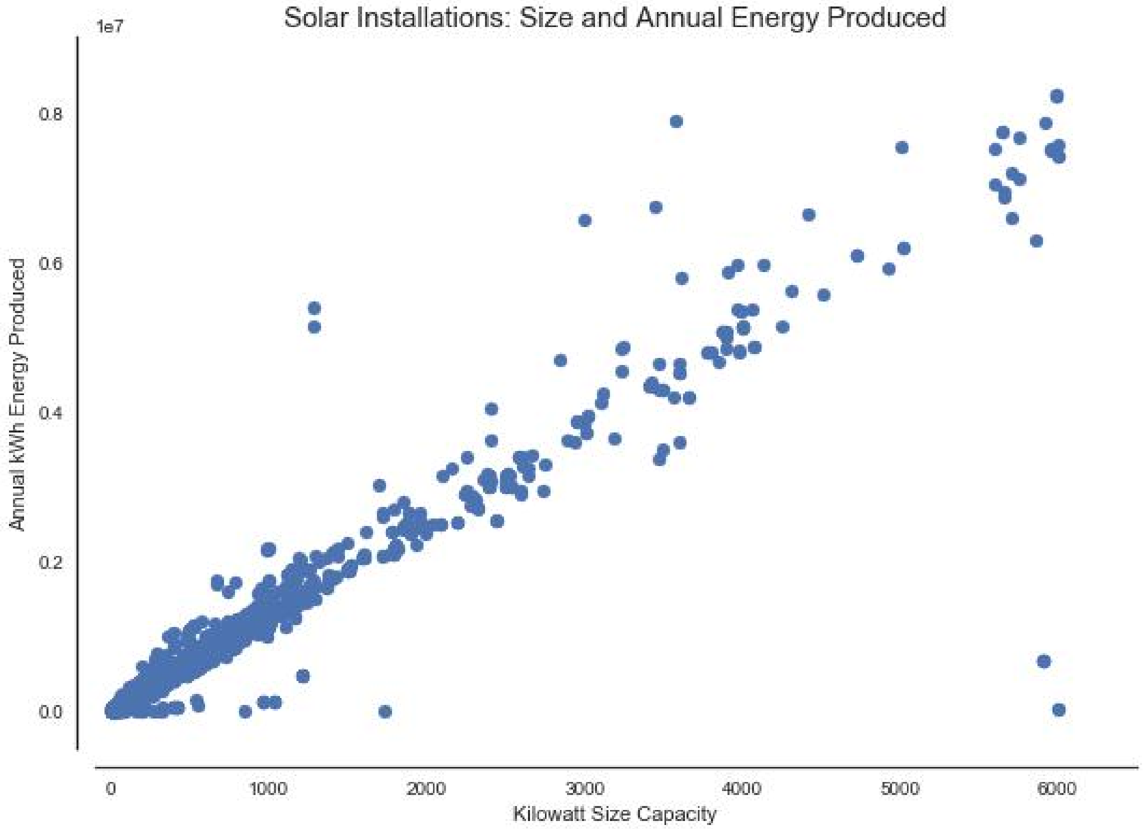


Fig: 5 A scatter plot of the size of solar panels and their annual

1.2.3 Partial Correlation

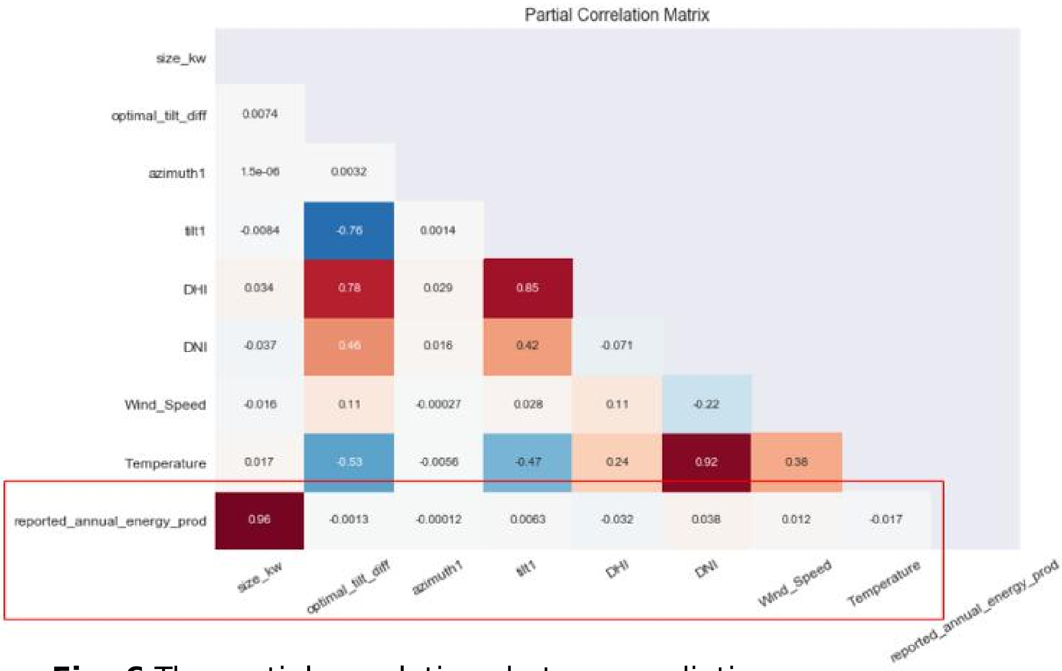


Fig: 6 The partial correlations between radiation factors and installation factors. The red box highlights partial correlations with annual energy production.

The above graph shows the partial correlations between installation factors and radiation factors. The red box highlights the partial correlations between these factors and annual energy production. Partial correlations are useful because they show the correlation between two variables when the effects of other variables are held constant. Dark red indicates high positive correlation, dark blue indicates high negative correlation, and white indicates no correlation.

Surprisingly, there appears to be no more than very slight correlations between the radiation measures (DNI and DHI) and energy production. We might expect—given that we're talking about solar energy after all—that these factors would play a large role in determining how much energy a solar panel generates. In the U.S., this may not end up being the case because solar radiation is high enough that panels reach their capacity for energy production, and are unable to produce more energy even with more sun exposure. If we were to compare the performance of solar panels in the U.S. to solar panels in the arctic, we might see that radiation is more highly correlated with energy production than we see here. An alternative explanation is that the level to which size correlates with output is biasing the results of the partial correlations.

Like in the scatter plot, there’s again a very strong positive correlation between size and energy production. It is interesting to note that there is a slight negative partial correlation between temperature and energy production. Though temperature obviously correlates highly with areas of high sun exposure, when controlling for the effect of radiation, high temperatures actually cause solar panels to produce energy less efficiently, and degrade more quickly.

1.2.4 Baseline Model

The NREL OpenPV dataset includes two measures of annual energy production, one that is a self reported value, and another predicted value based on a user’s inputs. Using these two values, Established a baseline R^2 score of .915. An R^2 score measures how much of the variance around the mean a prediction captures. So, if predictions were as good as guessing the average of the predicted variable every time, the R^2 score would be zero, and if the predictions were perfect, the R^2 would be one. At an R^2 score of .915, the baseline predictions are very accurate. Above is a scatter plot of the predicted and reported values, showing how closely they correlate.

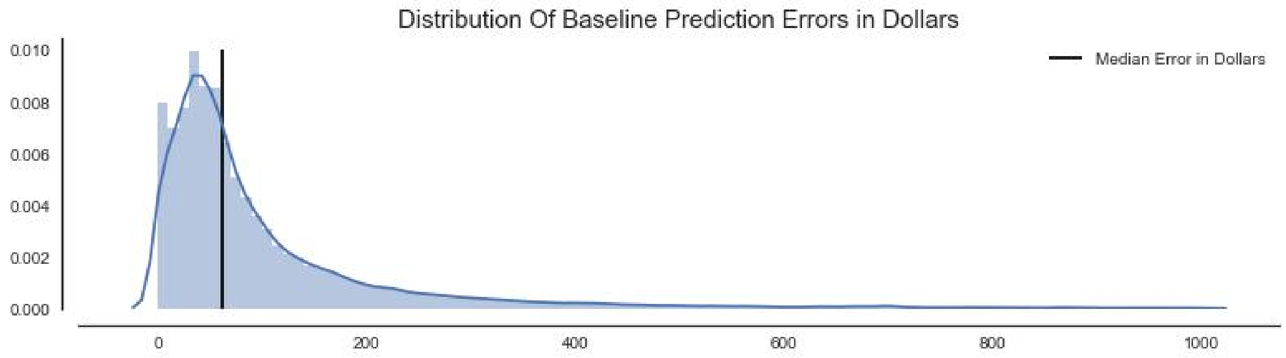


Fig: 7 A distribution of the differences in predicted return from actual return shows a very strong positive skew, due to large prediction errors for utility scale solar panels.

At an average return rate from a utility company of $0.1024 per kilowatt hour of energy generated, the median error in the baseline is equivalent to +/- approximately $60.00 in predicted savings on energy. The above chart shows the distribution of the differences between predicted and actual output for the baseline model, in dollars. The median is a more robust measure of central tendency than the average in this situation, given the large positive skew in the data. It's also a better measure of the prediction error for small-scale, residential solar customers. Over the course of the lifetime of a solar panel, this error compounds, amounting to thousands of dollars difference between the predicted and actual return.

1.2.5 Model Selection

The final model used for production was a random forest regression model, which has several benefits in this context. It is a non-parametric model, which means it can predict a variable that is non-normally distributed. Because there are a wide range of solar panels in the OpenPV dataset, with some utility-scale installations producing thousands of times more energy per year than small, residential panels, the data is very positively skewed. This means that in order to use a model like linear regression without having biased results, it would be necessary to log transform the data, or use a generalized linear model like a Poisson regression. Random forest models also do well with categorical features, and in this case there were a few such features, including technology type and tracking type.

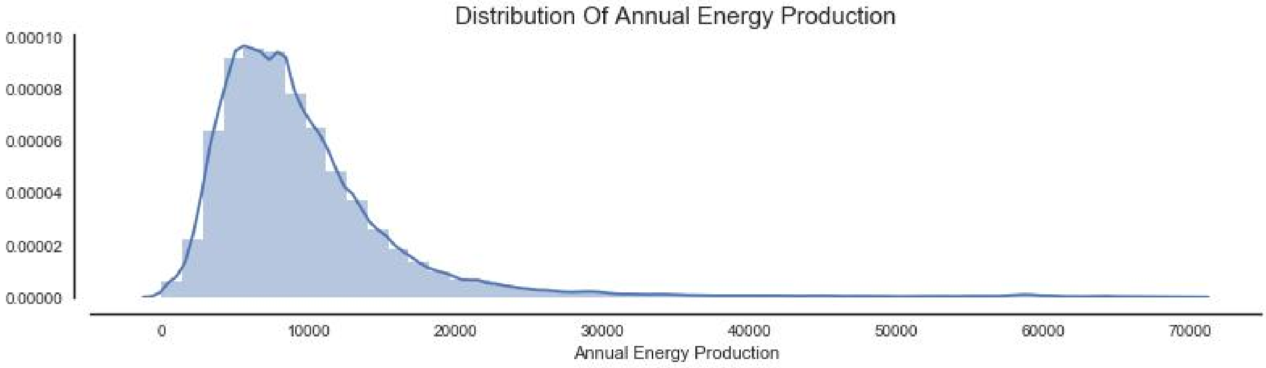


Fig: 8 The distribution of annual energy production is highly positively skewed, with some utility-scale installations producing many hundreds of times more energy per year than the median.

The performance of this model was compared against three others, including:

1. Ordinary Least-Squares Regression
2. Elastic Net Regression
3. 2-Layer Feed Forward Neural Network

Built the regression models and the random forest model using Scikit-Learn, and the neural network using Keras with a Tensorflow backend.

CHAPTER IV

Findings

Using the random forest model, achieved an R² value of .973 on validation data. The validation data were the same values used to generate the baseline score, so the comparison is completely like-to-like. In terms of annual savings, the median error drops from the baseline of $60.00 to +/- approximately $15.00 in predicted savings on energy.

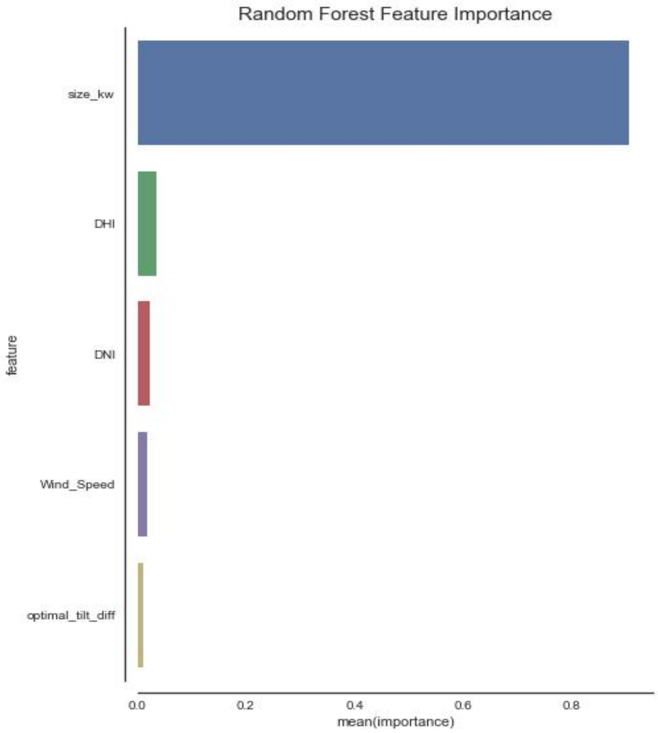
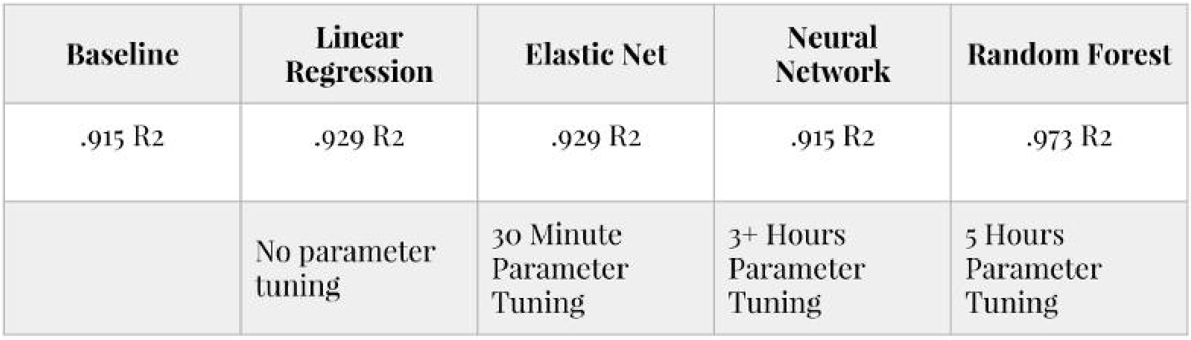


Fig: 9 Feature importance of the random forest model

Feature importance of a random forest model are a clear and easy way to interpret how much different variables contribute to predictions. More specifically, in Figure 9. they show the percent increase to mean squared error were a variable to be excluded from the model. We see that size contributes the most to predicting energy production, with an 87% importance. Direct and diffuse irradiance also play a role. These findings indicate that while it’s most important to build as large an installation as possible, building in places with high direct irradiance and low diffuse irradiance will help to produce more energy. Wind speed also plays a role, likely because high wind speeds correlate with areas of less shade, and high winds keep debris from collecting on panels. The last feature here is optimal tilt difference, which plays a comparatively small role in determining energy output. The more off a solar panel is from its optimal tilt, the less energy it will generate. Other factors in the model have a negligible importance.

1.1 Comparing Model Performance

While the random forest model ultimately performs best, it has several trade-offs. It took the longest time to hone in on the optimal parameters, and the interpret-ability of its results is lacking compared to linear regression and elastic net. The latter two models give a clear relationship between changes to factors like size and radiation and expected annual energy production, whereas random forest only provides feature importance. The neural network model suffers from both the drawback of time and a lack of interpret-ability. Given the relative simplicity of predicting energy output, a neural network appears to be an unnecessarily complex model for this situation. With more training time and parameter tuning, the neural network model would likely match the performance of the random forest model, but the marginal benefits are clearly limited.



Baseline	Linear Regression	Elastic Net	Neural Network	Random Forest
.915 R2	.929 R2	.929 R2	.915 R2	.973 R2
	No parameter tuning	30 Minute Parameter Tuning	3+ Hours Parameter Tuning	5 Hours Parameter Tuning

Table 3. Model Comparison

1.2 Model Output/Size

Because size dominates the performance of the models when used as a predictive factor, I built a regression model predicting output per unit size to gain a more granular understanding of the relationship between other factors and energy production. Though this model performs more poorly than when size is included as a feature, it helps demonstrate the effect of radiation and installation factors on output. The model scores an R² of .683 on validation data.

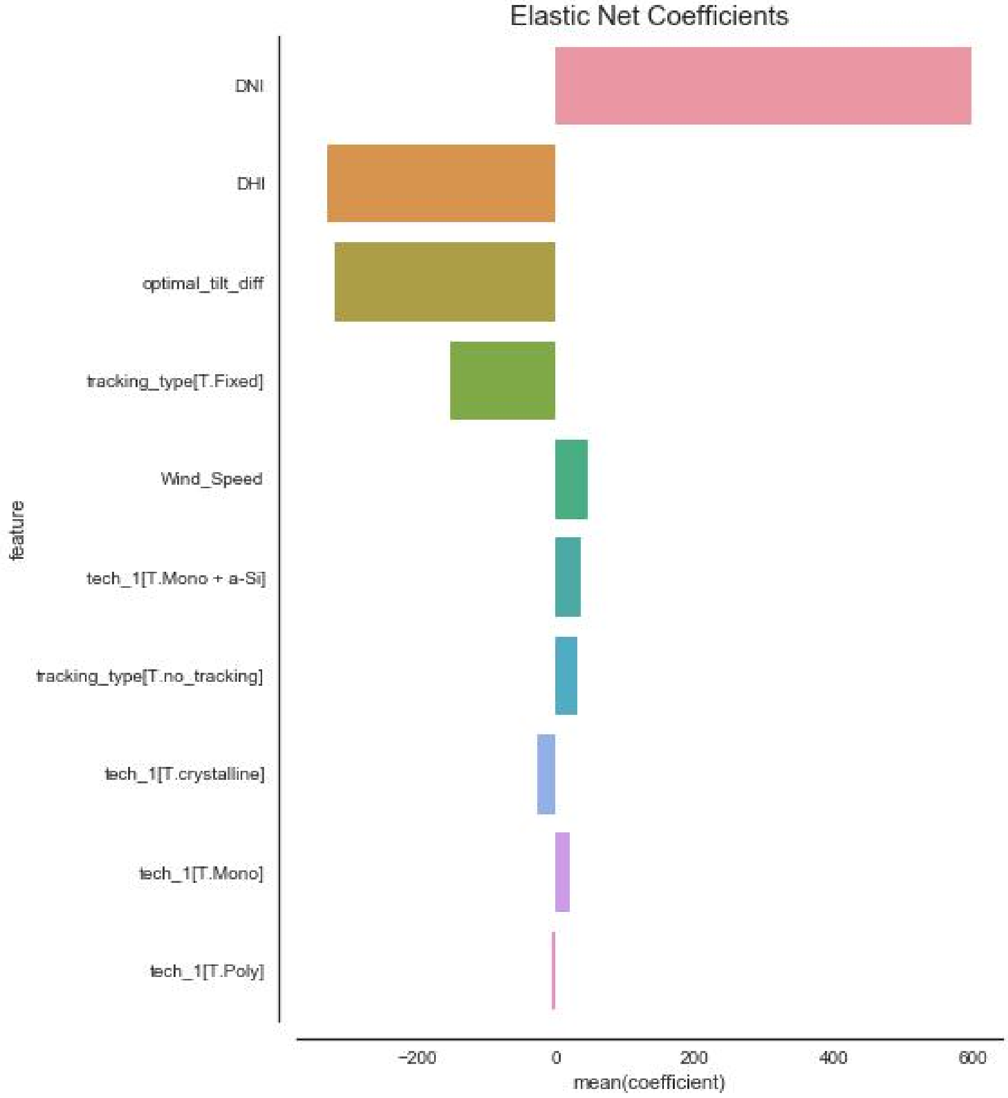


Fig: 10 The coefficients on an elastic net regression model predicting output per unit size. These values show the positive or negative effect a change to these factors has on the predicted variable.

Again Scikit-Learn has been used to build the elastic net regression for this model. Elastic net uses two types of regularization, lasso and ridge, which help to prevent over-fitting due to outliers and inconsequential features. Put simply, over-fitting is when a model learns the intricacies of its training data so specifically that it does not generalize well to other data.

The results of the model show how much positive or negative effect an increase in each factor has on annual energy production per unit size of an installation. Direct irradiance is the most important feature, with a large positive impact. Both diffuse irradiance and optimal tilt difference have substantial negative impacts on energy output, which is a fairly intuitive result. We see that a solar panel with fixed tracking will produce less energy as well, because a panel that does not track will have less exposure to radiation. Presumably it requires energy to operate a tracking solar panel, and this finding suggests that generally, the energy generated from tracking outweighs the energy required to operate the tracking system. Finally, mono-crystalline panels outperform poly-crystalline. We see this result because mono panels contain purer silicon, which leads to increased efficiency.

1.3 Creating Web App

Using the random forest model, designed a web app that allows users to input information about where they want to build solar panels, and learn how much they would save on their energy bill. The web app also uses a second model to predict installation cost based on the following factors:

1. Size
2. Installer
3. Tracking
4. Technology

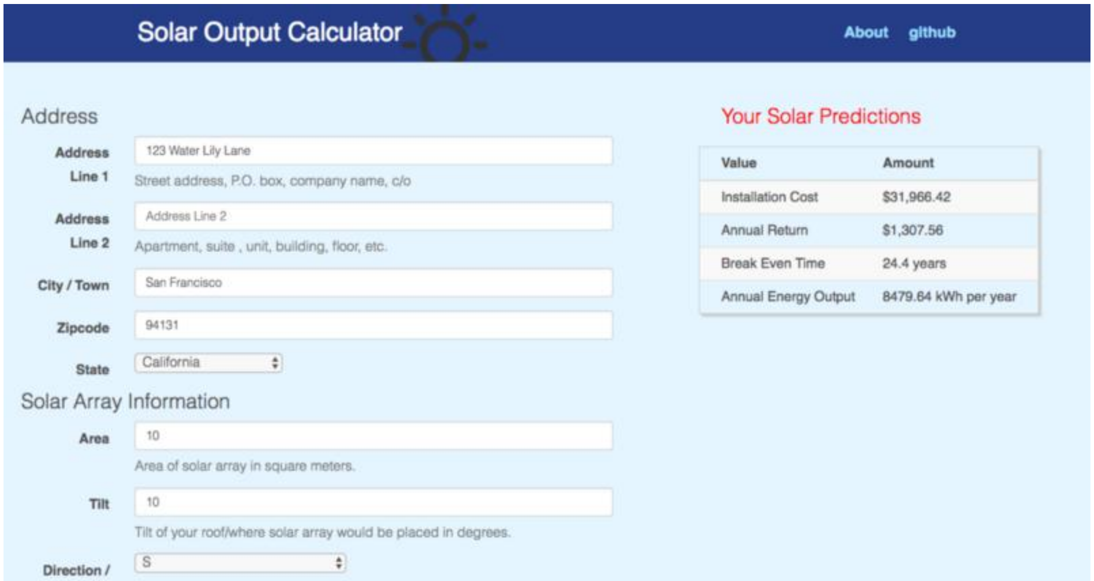
Built the website using Django, a python-based web framework. then set it up to run on an Amazon Web Services EC2 instance with a Gunicorn HTTP server. When a user inputs their information, the app queries the NREL solar radiation database to find their local radiation data, and calculates their expected annual energy production and installation. A Scrapy web crawler that finds their local average per-kilowatt return from a utility company, and uses that to calculate their average savings per year, as well as how long it would take to pay off the installation cost is also in use. The app can be checked here[http://www.solarcalculator.xyz/].

1.3.1 Using the Web App

The web app can be accessed from here [http://www.solarcalculator.xyz].
To access the system, initially it is without any restriction. After opening the website,

Step 1. Input the desired address line 1, 2 and city along with Zip code,
for US and Singapore both.
Step 2. Input the total covered area in square meters preferably and
the tilt angle of the solar system
Step 3. Azimuth value determines the direction of the PV solar station.

The result will be shown in the right hand corner, in a way like showing optimal tilt angle as well as total initial cost and annual return.



Address

Address Line 1: 123 Water Lily Lane
Street address, P.O. box, company name, c/o

Address Line 2: Address Line 2
Apartment, suite , unit, building, floor, etc.

City / Town: San Francisco

Zipcode: 94131

State: California

Solar Array Information

Area: 10
Area of solar array in square meters.

Tilt: 10
Tilt of your roof/where solar array would be placed in degrees.

Direction /: S

Your Solar Predictions

Value	Amount
Installation Cost	$31,966.42
Annual Return	$1,307.56
Break Even Time	24.4 years
Annual Energy Output	8479.64 kWh per year

CHAPTER V

Conclusion

Over the past 10 years, installation costs for solar energy technology have dropped an astonishing 60%. This form of renewable energy is more accessible now than ever before. Yet over that same period, soft costs, such as sales and marketing, have remained almost completely stagnant. According to the Solar Energy Industries Association (SEIA), in Q4 2016, soft costs accounted for 67% of installation costs for residential solar[2]. This has moved the impetus for growth in the solar industry from the development of cheaper technologies to a focus on ways to attack soft costs by more efficiently spreading information to potential solar customers.

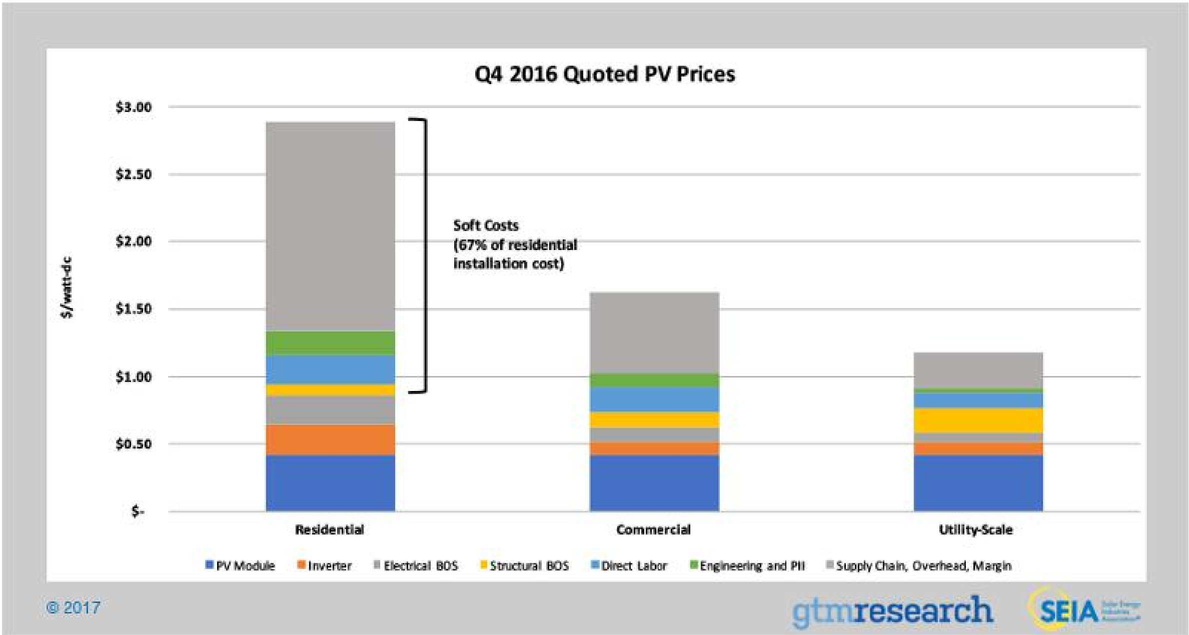


Fig 11: PV solar system cost and forecasting

If there is a choice between paying to install tracking, more tilt, etc., versus installing additional panels, the additional panels will almost always be the right move.

Using machine learning, the model gives highly accurate predictions of the expected return on energy generated by a prospective solar panel, and made it easily accessible at *solarcalculator.xyz*. Tools such as this, which use the machine learning techniques described above, will make information regarding one’s ability to switch to solar more widely available, ultimately bringing down soft costs of installation and accelerating the transition to renewable energy. But there are some hiccups in the process mentioned below.

1.1 Problems

The main problem arising before handling the data was the availability which is one of the reason that the project now is mainly aimed at US and Singapore. The unusual communication hiccups causes serious damage to the data modeler.

1.2 Future Implementation

In future implementations, the learning model can be improved using better sourced data and meaningful parameters. Also the amount of output measurement treated here are very low thus increasing parameters involving output generation is a positive step towards robust system.

This system not only work for solar power forecasting or return on investment showing but also belongs to a suit of operations where the same model of solar power forecasting and NWP modeling is used indicating cleanliness drives of solar panels as solar panels need regular washing, specially in desert areas. It will decrease human intervention in mundane jobs and increase efficiency if the whole solar PV system.

REFERENCES

[1] B. Espinar, J.-L. Aznarte, R. Girard, A.M. Moussa, G. Kariniotakis, Photovoltaic Forecasting: A state of the art, in: OTTI - Ostbayerisches Technologie-Transfer-Institut, 2010: p. Pages 250-255- ISBN 978-3-941785-15-1.https://hal-mines-paristech.archives-ouvertes.fr/hal-00771465/document (accessed April 14, 2018).

[2] V. Lara-Fanego, J.A. Ruiz-Arias, D. Pozo-Vázquez, F.J. Santos-Alamillos, J. Tovar-Pescador, Evaluation of the WRF model solar irradiance forecasts in Andalusia (southern Spain), Sol. Energy. 86 (2012) 2200–2217. doi:10.1016/j.solener.2011.02.014 (accessed April 14, 2018).

[3] A. Moreno-Munoz, J.J.G. De la Rosa, R. Posadillo, F. Bellido, Very short term forecasting of solar radiation, in: 33rd IEEE Photovolt. Spec. Conf. 2008 PVSC 08, 2008: pp. 1–5. doi:10.1109/PVSC.2008.4922587 (accessed April 14, 2018).

[4] D. Anderson, M. Leach, Harvesting and redistributing renewable energy: on the role of gas and electricity grids to overcome intermittency through the generation and storage of hydrogen, Energy Policy. 32 (2004) 1603–1614. doi:10.1016/S0301-4215(03)00131-9 (accessed April 14, 2018).

[5] M. Paulescu, E. Paulescu, P. Gravila, V. Badescu, Weather Modeling and Forecasting of PV Systems Operation, Springer London, London, 2013. http://link.springer.com/10.1007/978-1- 4471-4649-0 (accessed April 14, 2018).

[6] H.M. Diagne, P. Lauret, M. David, Solar irradiation forecasting: state-of-the-art and proposition for future developments for small-scale insular grids, in: n.d. https://hal.archives- ouvertes.fr/hal-00918150/document (accessed April 14, 2018).

[7] A. Hammer, D. Heinemann, E. Lorenz, B. Lückehe, Short-term forecasting of solar radiation: a statistical approach using satellite data, Sol. Energy. 67 (1999) 139–150. doi:10.1016/S0038- 092X(00)00038- 4 (accessed April 14, 2018).

[8] E. Lorenz, J. Remund, S.C. Müller, W. Traunmüller, G. Steinmaurer, D. Pozo, J.A. Ruiz-Arias, V.L. Fanego, L. Ramirez, M.G. Romeo, others, Benchmarking of different approaches to forecast solar irradiance, in: 24th Eur. Photovolt. Sol. Energy Conf. Hambg. Ger., 2009: p. 25. http://task3.iea-shc.org/data/sites/1/publications/24th_EU_PVSEC_5BV.2.50_lorenz_final.pdf (accessed April 14, 2018).

[9] M. Saguan, Y. Perez, J.-M. Glachant, L'architecture de marchés électriques : l'indispensable marché du temps réel d'électricité, Rev. Déconomie Ind. (2009) 69–88. doi:10.4000/rei.4053 (accessed April 14, 2018).

[10] B. Elliston, I. MacGill, The potential role of forecasting for integrating solar generation into the Australian national electricity market, in: Sol. 2010 Proc. Annu. Conf. Aust. Sol. Energy Soc., 2010. http://solar.org.au/papers/10papers/10_117_ELLISTON.B.pdf (accessed April 14, 2018).

[11] E. Lorenz, J. Kühnert, D. Heinemann, Short term forecasting of solar irradiance by combining satellite data and numerical weather predictions, in: Proc. 27th Eur. Photovolt. Sol. Energy Conf. Valencia Spain, 2012: pp. 4401–440.

http://meetingorganizer.copernicus.org/EMS2012/EMS2012-359.pdf (accessed April 14, 2018).

[12] D. Heinemann, E. Lorenz, M. Girodo, Forecasting of solar radiation, Sol. Energy Resour. Manag. Electr. Gener. Local Level Glob. Scale Nova Sci. Publ. N. Y. (2006). https://www.uni-oldenburg.de/fileadmin/user_upload/physik/ag/ehf/enmet/publications/solar/conference/200 5/Forcasting_of_Solar_Radiation.pdf (accessed April 14, 2018)).

[13] R. Perez, S. Kivalov, J. Schlemmer, K. Hemker Jr., D. Renné, T.E. Hoff, Validation of short and medium term operational solar radiation forecasts in the US, Sol. Energy. 84 (2010) 2161–2172. doi:10.1016/j.solener.2010.08.014 (accessed April 15, 2018).

[14] M. Diagne, M. David, P. Lauret, J. Boland, N. Schmutz, Review of solar irradiance forecasting methods and a proposition for small-scale insular grids, Renew. Sustain. Energy Rev. 27 (2013) 65–76. doi:10.1016/j.rser.2013.06.042 (accessed April 15, 2018).

[15] E. Lorenz, A. Hammer, D. Heinemann, Short term forecasting of solar radiation based on satellite data, in: EUROSUN2004 ISES Eur. Sol. Congr., 2004: pp. 841–848. https://www.uni-oldenburg.de/fileadmin/user_upload/physik/ag/ehf/enmet/publications/solar/conference/2004/eurosun/short_term_forecasting_of_solar_radiation_based_on_satellite_data.pdf (accessed April 15, 2018).

[16] G. Reikard, Predicting solar radiation at high resolutions: A comparison of time series forecasts, Sol. Energy. 83 (2009) 342–349. doi:10.1016/j.solener.2008.08.007 (accessed April 15, 2018).

[17] Mohammed, Azhar Ahmed,Aung, Zeyar, Ensemble learning approach for probabilistic forecasting of solar power generation (accessed April 15, 2018).

[18] Cyril Voyant, Gilles Notton, Soteris Kalogirou, Marie-Laure Nivet, Christophe Paoli, Fabrice Motte, Alexis Fouilloy, Machine Learning methods for solar radiation forecasting: a review (accessed April 15, 2018).

[19] Jawaid, Faizan,NazirJunejo, Khurum, Predicting daily mean solar power using machine learning regression techniques (accessed April 15, 2018).

[20] Haupt, Sue Ellen,Kosovic, Branko, Big Data and Machine Learning for Applied Weather Forecasts: Forecasting Solar Power for Utility Operations. (accessed April 15, 2018).

[21] J. Remund, R. Perez, E. Lorenz, Comparison of solar radiation forecasts for the USA, in: Proc 23rd Eur. PV Conf., 2008: pp. 1–9. http://ww.w.iea-shc.org/data/sites/1/publications/Comparison_of_USA_radiation_forecasts.pdf (accessed April 15, 2018).

[22] COST | About COST, (n.d.). http://www.cost.eu/about_cost (accessed April 15, 2018).

[23] R.H. Inman, H.T.C. Pedro, C.F.M. Coimbra, Solar forecasting methods for renewable energy integration, Prog. Energy Combust. Sci. 39 (2013) 535–576. doi:10.1016/j.pecs.2013.06.002 (accessed April 15, 2018).

[24] S.K. Aggarwal, L.M. Saini, Solar energy prediction using linear and non-linear regularization models: A study on AMS (American Meteorological Society) 2013–14 Solar Energy Prediction Contest, Energy. 78 (2014) 247–256. doi:10.1016/j.energy.2014.10.012 (accessed April 15, 2018).

[25] Z. Şen, Ş.M. Cebeci, Solar irradiation estimation by monthly principal component analysis, Energy Convers. Manag. 49 (2008) 3129–3134. doi:10.1016/j.enconman.2008.06.006. [26] M. Zarzo, P. Martí, Modeling the variability of solar radiation data among weather stations by means of principal components analysis, Appl. Energy. 88 (2011) 2775–2784. doi:10.1016/j.apenergy.2011.01.070 (accessed April 15, 2018).

[27] C. Paoli, C. Voyant, M. Muselli, M.-L. Nivet, Forecasting of preprocessed daily solar radiation time series using neural networks, Sol. Energy. 84 (2010) 2146–2160. doi:10.1016/j.solener.2010.08.011 (accessed April 15, 2018).

[28] D.J. Hand, Data Mining: Statistics and More?, Am. Stat. 52 (1998) 112–118. doi:10.1080/00031305.1998.10480549 (accessed April 15, 2018).