Solar Power Generation Forecast

Jayam Sutariya

CS 458

Introduction

Solar power has been seen to have great potential. It is already in use in many areas. It is most likely to take over fossil fuels as the next source of energy. It is renewable, meaning it is environmentally friendly, unlike fossil fuels. Currently, authorities and organizations are trying to expand it to a very large scale much like the fossil fuel plants. However, there are many problems that come with. Solar generation already is an issue. Generating power through the sun seems very easy at first glance, however, with current methods, the generation efficiency is not optimal at all. Not only that, storing solar power is extremely difficult. In fact, making solar energy economical is one of the 14 Engineering Grand challenges today [NAE].

That begs the question, what does have to do with solar power generation forecasting? It turns out that high levels of energy distribution introduce several challenges in power system operation because of the intrinsic intermittent and uncertain nature of such energy distributions plants. This is especially the case for solar plants, as there are so many variables that go into the power generation. In that case, it is extremely fundamental that development occurs which grants the ability to accurately forecast energy production for such large-scale solar plants.

So, what exactly are the benefits that these energy generation forecast provides? There are three main ones: Dispatchability, Efficiency, and Monitoring. Dispatchability is beneficial because energy plants in general, solar, or not, and their daily operation relies heavily upon day-ahead dispatches of power plants. These dispatches can only happen if accurate day-ahead predictions happen for solar plants. Earlier it was mentioned that solar power generation efficiency is not optimized. Efficiency is important because it solar plants can show high efficiency, it will help them take over the traditional fossil fuel plants. Some countries have even introduced penalties for not accurately predicting solar power generation. This means that efficiency can be improved upon if the forecasting can be improved as well. As time goes on, power plants start to age and perform at rate much less than they originally performed at. If solar forecasting is accurate, it would be easy to monitor the performance of a solar plant. This would be done by comparing the forecasts with the actual data and if the actual data is lower, then the power plant is not performing at its peak.

Background and related work

Mentioned earlier is the reason solar power generation forecasting is such a huge issue. In the past decade, it has become a widely known phenomenon amongst the power generation community. Individuals who posses the skills to forecast accurately are actively sought out by the big power providers. For the past few years, many researchers have tried to accurately predict the power generation using different models and shared their results. Listed here are a few attempts at accurately predicting solar power generation using different models.

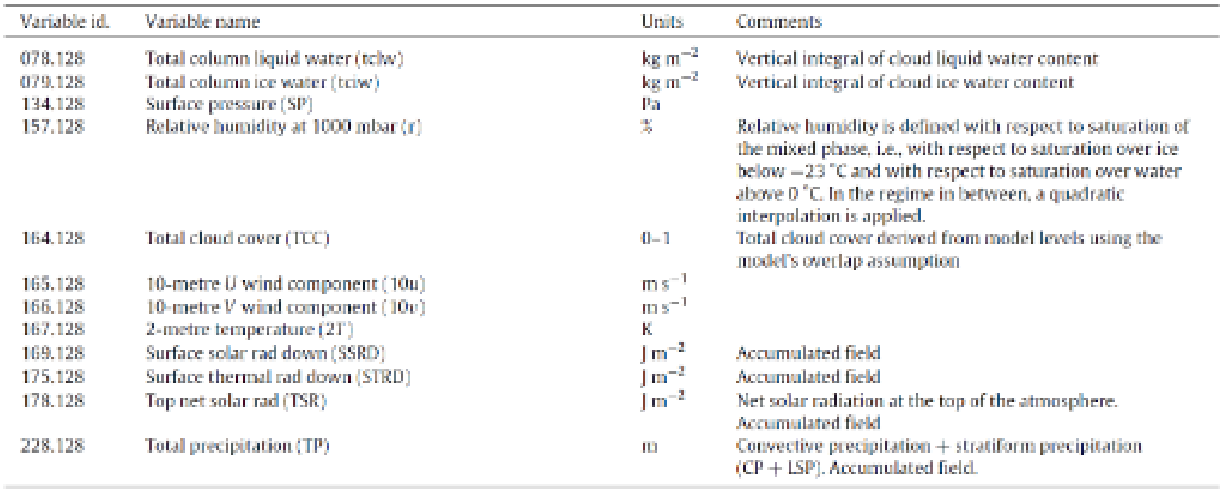
- The most obvious choice would be to use a Neural Network as it is very commonly used today for classification and regression problems. Researchers at Texas Tech University tried to do exactly that for the solar power generation forecast problem in 2016 using attributes such as cloud cover, temperature, wind speed, etc. They extracted results from the neural network model and additional regression models and found that the neural networks performed slightly better due to the attributes selected. [Verma]
- Another paper also addresses the short-term problem of renewable energy forecasting. This research took place in North China in 2020, and they used the Random Forest algorithm to accurately predict the power generation. The paper addresses the problem of feature identification and appropriate error metrics to use which seem to be really helpful for the power forecasting community. Their model result in a 93% accuracy on average and they concluded with many issues that need to be resolved to improve the forecasting in general. [Khalyasmaa]
- A different group of researchers went a step further in analyzing many different classification and regression methods for day ahead solar power generation forecasting. They illustrate the importance of forecasting in today's world. They gathered data from 12 solar plants and used many attributes for their models. After comparing 6 different model and careful analysis on them using useful metrics, they found that an ensemble of all of the models provided the best results. [Gigoni]

Taking into consideration these research studies and more happening actively, it is safe to assume that solar forecasting is very important avenue that needs to be improved upon using machine learning techniques. All of the different studies provide various lessons that can be learned. The data collection for example needs to be improved. Feature selection needs to be improved. Meteorological forecasting also needs to be improved along with its data. Perhaps an ensemble of all the models needs to happen. Last lesson but the not the least that can be learned from these studies is that the correct error metrics need to selected for analysis of the model.

Analysis of the Dataset

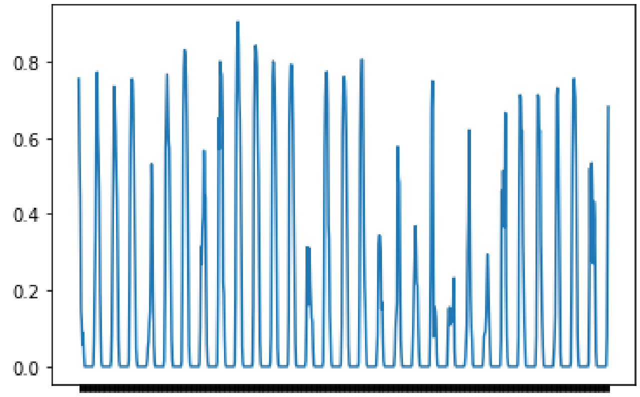
For my model, I acquired the dataset from Global Energy Forecasting Competition from 2014 in Australia. It is refined version of that dataset. It captures data from 3 different solar plants and utilizes the following variables to provide with a solar power generation figure. The data naturally has a category to identify the solar plant. The dataset itself is distributed amongst two years from 2012 to 2014 and provides hourly data for each of these variables and the final solar power generation at that given time.

Table 1. 12 weather variables



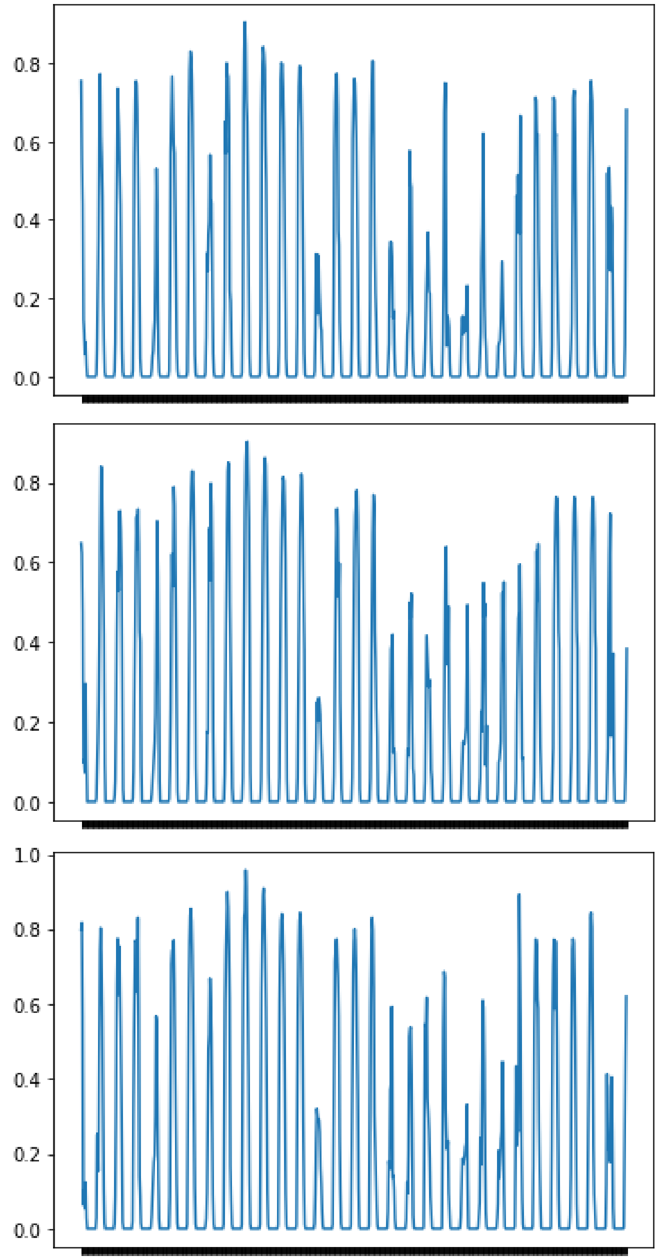
Variable id.	Variable name	Units	Comments
078.128	Total column liquid water (tclw)	kg m^{-2}	Vertical integral of cloud liquid water content
079.128	Total column ice water (tciw)	kg m^{-2}	Vertical integral of cloud ice water content
134.128	Surface pressure (SP)	Pa	
157.128	Relative humidity at 1000 mbar (r)	%	Relative humidity is defined with respect to saturation of the mixed phase, i.e., with respect to saturation over ice below −23 °C and with respect to saturation over water above 0 °C. In the regime in between, a quadratic interpolation is applied.
164.128	Total cloud cover (TCC)	0–1	Total cloud cover derived from model levels using the model's overlap assumption
165.128	10-metre U wind component (10u)	m s^{-1}	
166.128	10-metre V wind component (10v)	m s^{-1}	
167.128	2-metre temperature (2T)	K	
169.128	Surface solar rad down (SSRD)	J m^{-2}	Accumulated field
175.128	Surface thermal rad down (STRD)	J m^{-2}	Accumulated field
178.128	Top net solar rad (TSR)	J m^{-2}	Net solar radiation at the top of the atmosphere. Accumulated field
228.128	Total precipitation (TP)	m	Convective precipitation + stratiform precipitation (CP + LSP). Accumulated field.

Provided below is the graph for solar power generation for the first month of April 2012 for the first solar plant:



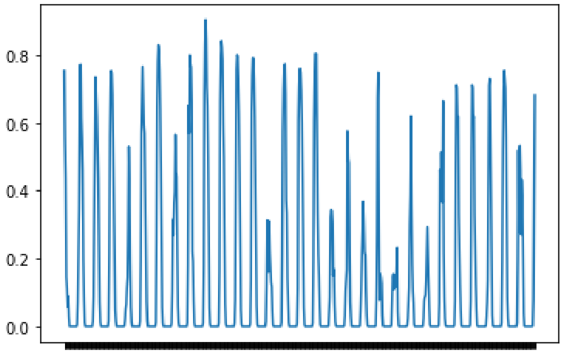
As it can be clearly seen, there are spikes during certain time of the day and no power generation during the rest of the day. The more important thing to notice here is that are many irregularities even among one month. Some days, the power generation is high, some days it low, meaning the data already has complexity.

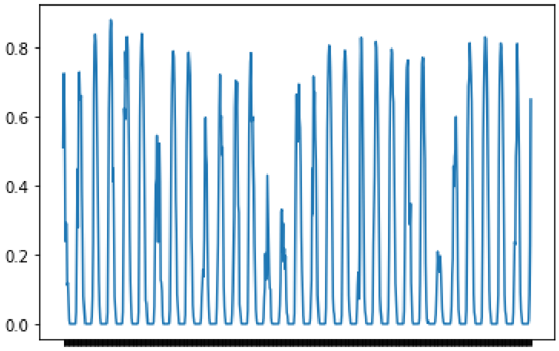
The graphs below are for the solar power generation comparing the three different plants on the same month of April 2012:



Again, it can be observed that even though there many similarities between the three power generation values of the 3 different plants, there are still many discrepancies between the three plants. Meaning that there is more complexity with the data, which is not good.

If we go a step further and look at the solar power generation graphs for the months of April 2012 and November 2012:





Obviously, there are going to be discrepancies between the solar power generation values between the months of April and November in the same year. However, there does not seem to be a pattern here to illustrate that. In some cases, the solar power generation was higher for the month of November. In any case, again there is more complexity to the data that is yet to be understood.

Approach

The approach that I took was to first split the data into two parts, training, and testing. This was done by simply splitting the dates into half, meaning each set contains one year of data. I then decided to further split the data into 3 different parts relating to the three different solar plants. The reason for this was to reduce the complexity in the data. The data was already so complex, there did not need to be more. Making the model simple is always desirable as it reduces time and memory use while allowing us to analyze the solar plants separately to try to understand the formation of the data.

Data preprocessing

I decided to use several different models for this problem ranging from classification and regression models. For the classification models, I quickly found that the models will not work with the continuous data that is available. To resolve that issue, I had to encode and transform the data from continuous type to multiclass so that the classifier models can be used.
After trying on a few classification models, I realized that they all took a really long time to classify. Most of them did not converge. To resolve this issue, I used PCA to reduce the dimensionality from 12 to lower dimensions. I tried using the models for classification using 9, 6, 4, and even 2 dimensions and it turned out that most of them still did not converge purely because of the multiclass approach and the huge number of classes. This meant that SVM, Neural Network, and the Random Forest classifier all did not function even with and without PCA.

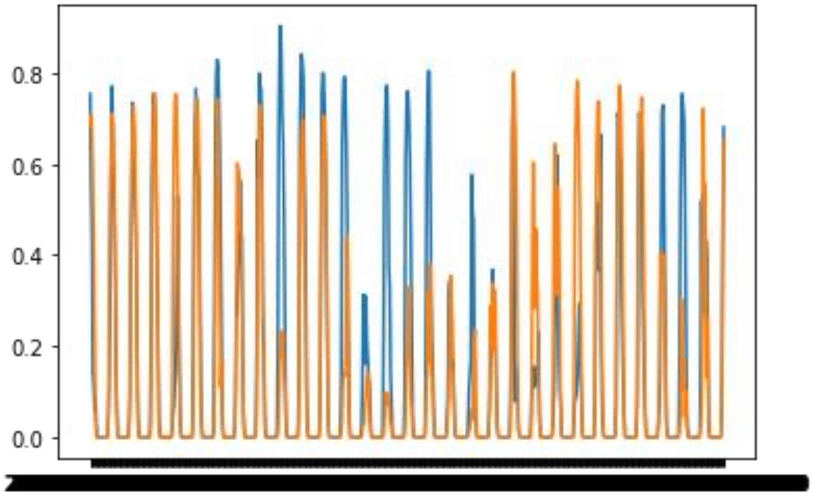
Model selection

The only classifier that worked was the K-nearest neighbor classifier. All the other classifiers mentioned have high or medium complexity to train which is why they took so long and did not converge. However, the KNN is extremely easy to train, and it is also very insightful. It does unfortunately, have slow predictions, but the predictions were really quickly to my surprise. The only problem with the KNN was that the results were terrible. Both the error rates that I used were extremely high for all three solar plants.

I finally decided to move onto the regression models and the first model I tried was the Random Forest Regressor. The primary reason I originally tried the Random Forest Classifier was because of its many benefits such as medium training complexity, medium paced predictions, and compatibility with data much like the one being used based on the research studies mentioned earlier. With the regressor, though, no data encoding and transformation needed to be done as the regressor model accepts continuous data. The parameters that were used were very simple. I simply set the random_state to 0 and the max_depth to 2 and obtained desirable results. These parameters made sure the model was simple and naturally provided quick training and predictions. I further adjusted the max depth to increase the accuracy and finally landed on max_depth = 10. Increase the max depth anymore would just overfit the data so I left the max depth at that.

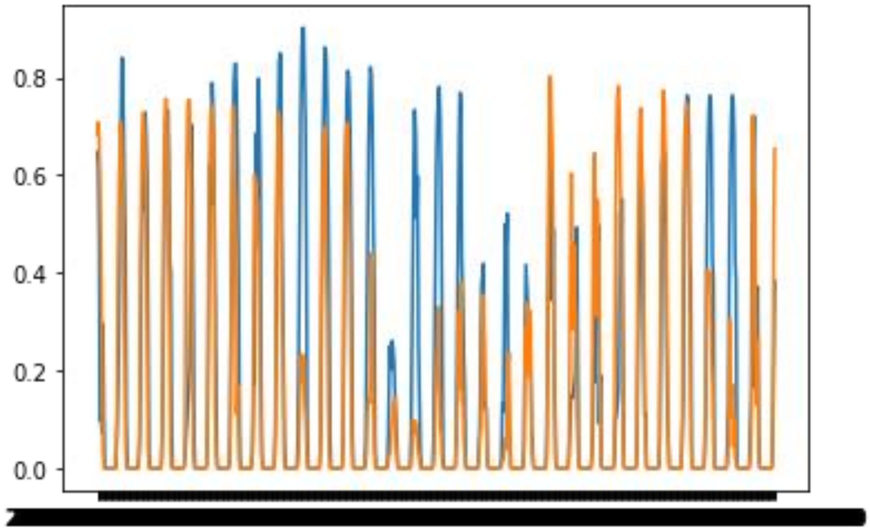
Results

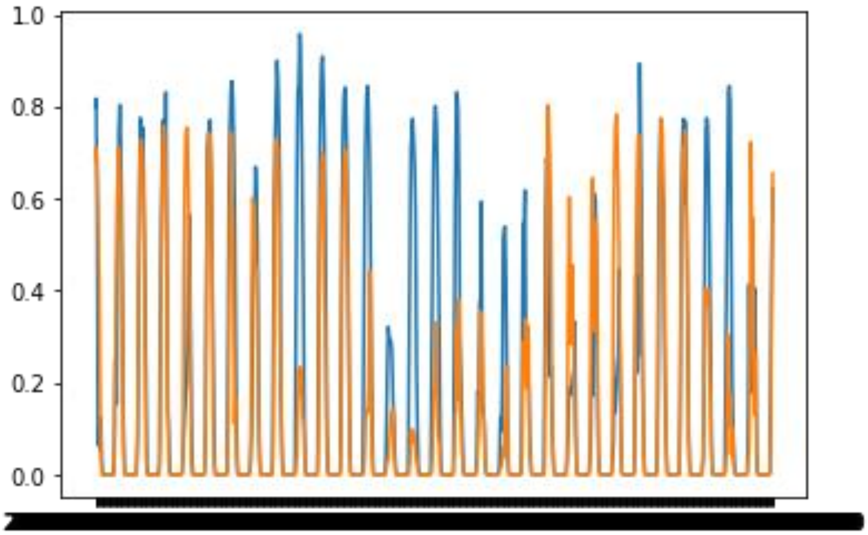
Provided below is the comparison of the solar power generation for April 2012 for the first solar plant:



As it can be clearly seen, the prediction in orange very nearly matches the original power generation values in blue. There are days where the prediction is way off compared to the actual power generation values.

Similarly, provided below are the solar power generation comparison graphs for the month of April 2012 for the second and the third plant, respectively:





There is a similar pattern with the second and the third plants as well. It is mostly accurate, however there are days when the prediction is way off.

To look more deeply into this issue, we must calculate some metrics. Mean absolute error and the root mean squared error were used as metrics for evaluating the success of the model.

The equations for each are:

$$MAE = \frac{1}{number\ of\ points}\sum_{t}|P_t - \hat{P}_t|$$

$$RMSE = \sqrt{\frac{1}{number\ of\ points}\sum_{t}|P_t - \hat{P}_t|^2}$$

They are both great metrics that were seen to be used in the research studies analyzed earlier.

ZONEID	1	2	3	Overall
MAE	0.05	0.05	0.06	**0.0547**
RMSE	0.10	0.10	0.11	**0.1046**

It can be seen that the error rates are very low. They are very close to the values in the research studies that use these exact same metrics. However, when taking into consideration the graphs where the predicted values do not match the actual, there is some issue.

Conclusion

The small error rate can be accredited to only some of the dissimilar values from predicted and actual values. I believe most of the dissimilarity can be designated to the fact that a lot of the solar power generation values are 0 or close to 0. Which means that predicting these values would be much easier.

To fix this, perhaps we must split the dataset in a way that puts less emphasis on the times when the solar power generation is 0 or close to 0. Perhaps a different mode with ideal parameters could fix this, however, based on my research and experimentation, Random Forest Regression model is the best. Perhaps an ensemble of different models would provide optimal results.

Nonetheless, the solar power generation forecast model provided here is not optimal as it does not predict the important values correctly all the time, even though, at the surface level they seem correct because of the low error rates.

References

Make solar energy economical. Grand Challenges - Make Solar Energy Economical. (n.d.). Retrieved December 15, 2021, from http://www.engineeringchallenges.org/challenges/solar.aspx

T. Verma, A. P. S. Tiwana, C. C. Reddy, V. Arora and P. Devanand, "Data Analysis to Generate Models Based on Neural Network and Regression for Solar Power Generation Forecasting," 2016 7th International Conference on Intelligent Systems, Modelling and Simulation (ISMS), 2016, pp. 97-100, doi: 10.1109/ISMS.2016.65.

A. Khalyasmaa et al., "Prediction of Solar Power Generation Based on Random Forest Regressor Model," 2019 International Multi-Conference on Engineering, Computer and Information Sciences (SIBIRCON), 2019, pp. 0780-0785, doi: 10.1109/SIBIRCON48586.2019.8958063.

L. Gigoni *et al*., "Day-Ahead Hourly Forecasting of Power Generation From Photovoltaic Plants," in *IEEE Transactions on Sustainable Energy*, vol. 9, no. 2, pp. 831-842, April 2018, doi: 10.1109/TSTE.2017.2762435.