Project Report

Project code: EP001

Project title: Prediction of Solar Power Potential in Singapore using Numerical Weather Prediction Models and Machine Learning

Introduction

Solar radiation [1] [2]

When it comes to energy generation, solar power remains the most promising renewable energy source in Singapore. With an average solar irradiance of 1,580 kWh/(m^2*year) and about 50% more solar radiation than temperate countries, solar photovoltaic (PV) generation has the greatest potential for wider deployment in Singapore. [1] Moreover, a report by National Renewable Energy Lab [2] has indicated that from 2010 to 2017 there has been approximately a 70% reduction in the total PV system hardware cost. Combining great potential and decreasing cost, there is a strong demand to deploy solar power widely in Singapore. Precise solar radiation prediction will be the key to successful integration of solar power.

Available prediction tools [3] [4] [5] [6]

Numerical Weather Prediction (NWP) models, which solve differential equations in order to describe the dynamics in the atmosphere, are the most popular type of prediction approaches. The second type is statistical learning, also known as machine learning, which can derive functional dependencies directly from the observations.

Currently, one of the most widely used weather prediction models based on NWP is known as Weather Research and Forecast (WRF) model [7]. The model has been used in the weather forecast in the study [8] [9] and has proved its reliability by its successful power output forecast in the context of Smart Grids and Renewable Energy on [10]. The WRF model can produce simulations based on actual atmospheric conditions, both from observations and analyses, or idealized conditions. [11] By analyzing how the atmospheric variables will interact with one another based on physical equations, the model will simulate the whole atmospheric environment and make predictions based on different requirements. For example, the WRF model can simulate the movement and thickness of the cloud and based on the real-time simulation, it can predict the solar potential of the region with very high accuracy because the environment simulated is very close to the actual world and they are based on the same physical principles.

For example, the WRF model can simulate the real-time atmospheric condition based on the data collected and hence can predict the solar potential of the region with high accuracy since the environment simulated is very close to the actual world based on the same physical principles.

The WRF model is commonly used to make short-term weather predictions and long-term climate predictions. One reason that restricts the WRF's performance is the model requires vast amounts of computing power. Manipulating massive amounts of data and performing complex calculation requires some of the most powerful supercomputers in the world and it may still take huge amounts of time to complete the prediction. Despite it is highly time-consuming and requires numerous computing power, the WRF model requires a huge amount of prior knowledge to interpret the computer forecast and because of this, very few people can operate the model to make the predictions as they want.

Due to the limitations of the WRF model, there is a strong demand to develop alternative methods to make the predictions [12] and that is where machine learning comes in. Machine Learning [13], that research different algorithms, can learn information from the dataset and make a useful tool to predict the solar radiation with the much shorter time taken. [6] [14] Essentially, various algorithms can accept a given set of training examples and derive a general trend over time. [15] There are different machine learning algorithms, such as Support Vector Machine (SVM), Random Forest (RF), K-Nearest Neighbours and Time series and so on. [16] The focus of the study is to learn to build prediction models with four algorithms mentioned above and employ them to make the prediction of solar radiation in a fast and more importantly, accurate manner.

Methodology

Data pre-processing

A dataset was generated using WRF for the year 2018. The data was saved at 62 different locations across Singapore. The original dataset included 20 different climatic variables. To make it easier to use, it was reduced to 1 four column dataset. The four columns are Time, Direct Solar Radiation (SWDIR), Diffuse Solar Radiation (SWDIF) and Long-wave radiation (GLW). When both SWDIR and SWDIF are zero, it corresponds to the night time when there is no solar radiation at the location.

In this study, the solar potential is measured by the clear-sky ratio (CSR), defined as

$$\mathrm{CSR} = \frac{\mathrm{SWDIR}}{\mathrm{SWDIR} + \mathrm{SWDIF}} \tag{1}$$

wheere CSR ranges from 0 to 1 by its definition. As Singapore has the very limited land area, if there is no cloud cover, the total amount of radiation from the sun is rather constant across the country and hence CSR can be an applicable measure of solar potential across the country. In order to represent the solar radiation intuitively, the CSR value has been divided into ten continuous even groups (Group 1 to Group 10) from 0 to 1, namely 0.0-0.1, 0.1-0.2, 0.2-0.3 ... 0.9 - 1.0. The CSR values at night, which are expected to be NaN by its definition, are classified to Group 0, in order to maintain the continuity of the dataset.

The whole dataset has been normalized such that its mean is 0 and standard deviation is 1. The formula is given as below:

$$X_{\mathrm{norm(i,\,j)}} = \frac{X_{(i,j)} - \mathrm{mean}(X_j)}{\mathrm{std}(X_j)} \qquad (2)$$

where X_j is the jth column of matrix X. After normalising the dataset, the mean becomes zero and has unit variance along each feature and can be less sensitive to the scale of features.

Various Models Used

Machine learning is the study of algorithms that can contribute to progressively improving the performance of the computer on certain tasks. By building a mathematical model with sample data, known as the training set, the computer is able to make predictions or decisions without being explicitly programmed to perform the task. In this study, by building four machine learning models based on different algorithms and feeding them with the training set, we can derive the prediction of solar radiation data and proceed to the further evaluation process.

Machine learning models are able to learn from data based on different algorithms. In supervised learning, computers are usually presented with sample input and their desired outputs and the goal is to learn a general function that maps inputs to outputs [17]. There are generally two types of models in supervised learning: classification and regression. Classification aims to identify which of a set of categories the new observation belongs to, on the bias of a training set of data containing observations with their category membership known and the purpose of regression models is more about estimating the relationship between a dependent variable and one or more independent variables, which means it can help people understand how the dependent variable changes when any one of the independent variables is varied. Both of the two kinds can contribute to making predictions of the solar radiation data and this section will introduce four different models with their algorithms.

Support Vector Classifier [18] [19] [20] [21] [22] [23] [24]

Support Vector Machine algorithms (SVMs) are a set of supervised learning methods used for classification, regression and outlier detection. The objective of the algorithm is to find a hyperplane in an N-dimensional space (N - the number of features) that distinctly classifies the data points.

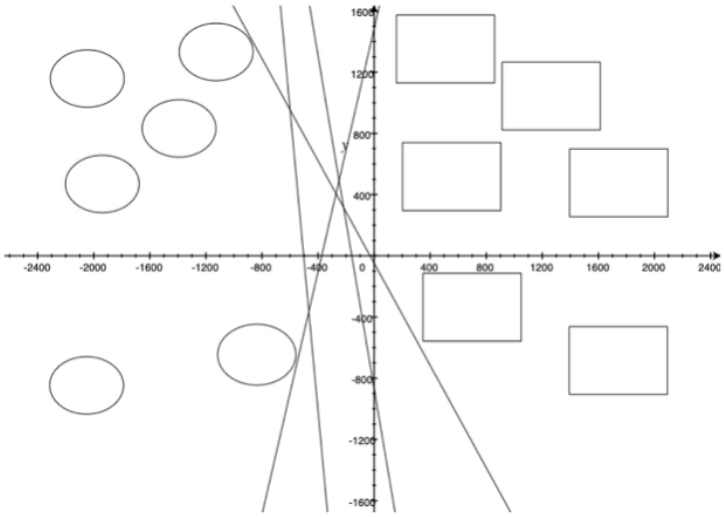


Figure 1: Possible hyperplanes

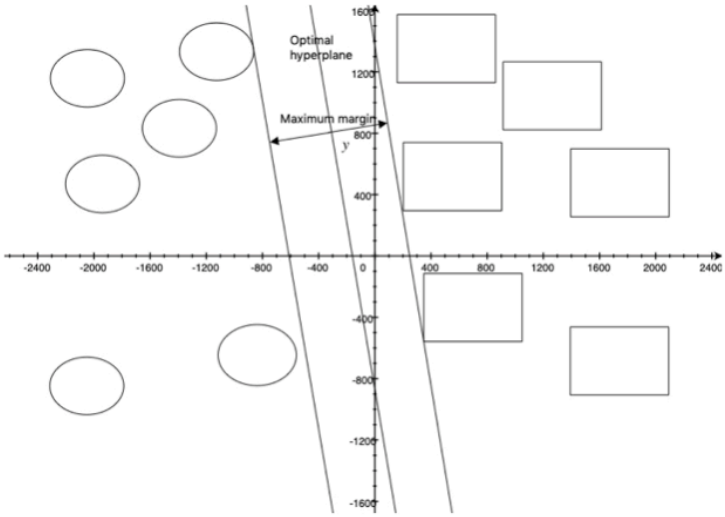


Figure 2: Optimal hyperplanes with the maximum margin

As Figure 1 suggests, there can be many possible hyperplanes available and the one with the maximum margin is to be found, i.e the maximum distance between data points of both classes shown in Figure 2. Maximizing the margin distance reinforces the model and makes the future classification more reliable.

Hyperplanes refer to decision boundaries that contribute to classifying the data points. Data points falling on either side of the hyperplane can be attributed to corresponding classes. The dimension of the hyperplane is equal to N-1, where N is the number of features. Support vectors are data points that are closer to the hyperplane and can influence the position and orientation of the hyperplane. With these support vectors, the margin of the classifier can be maximized.

Random Forest [3] [4]

Random Forest algorithm usually includes three parts: Decision Tree, Feature Bagging and Ensemble of the result.

In the formation of a Decision Tree, the subsets of training data are randomly sampled with replacement and the target variable is fitted to regression models using each of the independent variables. The data is split at several split points for each independent variable. At each split point, the Sum of Squared Errors (SSE) between the predicted values and the actual values is calculated and compared against one another. The variable or point yielding the lowest SSE is hence chosen as the split point, also called the root node. The process, known as "bagging" method, recursively continues. Many regression Decision Trees are formed with several nodes in each tree and the ensemble of these decision trees is known as the "random forest".

The Random Forest method introduces more randomness and diversity by applying the bagging method to the features. Rather than search for the best predictors greedily to find new split points, it randomly samples elements from the predictors, thus increasing the diversity and reducing the variance of the trees while keeping bias high or even increasing. The process is known as “Feature Bagging” and can help make the model more robust.

In the prediction process, each new data point goes from the root node to the bottom until it is fitted to a function with all the different trees in the ensemble and the average prediction of each tree is used as the final result. The ensemble learning allows multiple algorithms to obtain better predictive performance than that from any constituent learning algorithm alone. By combing result of all the "trees" in the "forest" with ensemble learning, the more accurate prediction can be obtained.

K-Nearest Neighbours Regressor [25]

The K-nearest neighbor regressor (KNN) is an algorithm that stores all available cases and makes the numerical prediction based on a similarity measure, e.g. distance function. KNN identifies the K-nearest neighbors by some distance metric and then average the value of these neighbors. The prediction is given by:

$$\hat{Y} = \frac{1}{K} \sum_{i \in N_0} y_i \tag{3}$$

Commonly for continuous variables, the distance used is the Euclidean distance given by:

$$D = \sqrt{\sum_{i=1}^{k} (X_i - y_i)^2} \tag{4}$$

Usually, it can be very helpful to give higher weights to the nearby points compared with faraway points and this can be achieved by assigning weights proportional to the inverse of the distance.

Time-series model [26] [27] [28] [29] [30]

An important aspect of the time-series model is the concept of stationary processes. That means the autocorrelation and expectation of the overall dataset do not vary with time. Known the solar radiation has its seasonal variation pattern, the time-series can be employed to make the prediction. The model is known as Seasonal Auto-Regressive Integrated Moving Average with eXogenous regressors model (SARIMAX) and is generally a combination of the Auto-regressive model (AR) and the Moving Average model (MA) with additional seasonal terms.

The auto-regressive (AR) process is a regression term that the explanatory variable depends on historical values. The formula is given below:

$$X_t = \sum_{i=1}^{p} \phi_i X_{t-i} + c + \epsilon_t \tag{5}$$

In the equation, p is chosen depending on the autocorrelation functions and an AR(1)-model correspond to $p = 1$. ϕ corresponds to the auto-regressive parameter and X refers to the time series under investigation. c is a constant and ϵ_t is the error term.

The Moving Average (MA) process is another regression form where the prediction parameter depends on external shocks rather than historical data. Formally, the process is defined as:

$$X_t = \mu + \epsilon_t * (1 + \theta_1 B + \cdots + \theta_q B^q) \tag{6}$$

Here, μ is the mean of the series and $\theta_1 \ldots \theta_q$ are the parameters of the model and the $\epsilon_t \ldots \epsilon_{t-q}$ are white noise error terms. The q term is called the order of the MA model. B is the backshift operator.

The seasonal ARIMA model incorporates both non-seasonal and seasonal factors in a multiplicative model. One shorthand notation for the seasonal ARIMA model is

$$\text{ARIMA}(p, d, q) \times (P, D, Q)S \tag{7}$$

with p = non-seasonal AR order, d = non-seasonal differencing, q = non-seasonal MA order, P = seasonal AR order, D = seasonal differencing, Q = seasonal MA order, and S = time span of repeating seasonal pattern.

Analysis and Evaluation of the results [23] [31] [32] [33]

A common performance metric is the Mean Absolute Error (MAE), given by:

$$MAE = \frac{\sum_{i=1}^{n} |y_i - \hat{y}_i|}{n} = \frac{\sum_{i=1}^{n} |e_i|}{n} \tag{8}$$

where y_i and $\hat{y}_i$ corresponds to the actual value of y and predicted value of y and $e_i = y_i - \hat{y}_i$. As the MAE sums up the absolute errors, it can reflect the overall error from the actual trend where all individual terms have equal weights.

Another common performance metric is the Mean Squared Error (MSE), given by:

$$MSE = \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n} \tag{9}$$

Same as the MAE, the terms in the formula represent the exact same meanings. By summing up the squared errors, MSE gives the outliers a higher weight and can reflect the variation of each individual term.

Similar to the MSE, Root Mean Squared Error (RMSE) gives higher weight to outliers and is given by:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n}} \tag{10}$$

By finding the square root of MSE, the RMSE can be directly interpreted in terms of measurement units, and hence is a better measure of accuracy than a correlation coefficient or MSE. While giving higher weights to terms with large errors, which are particularly not desired outcomes, RMSE has a tendency to be increasingly larger than MAE as the sample size increases. This can be problematic and restricts the application of RMSE in this study given the total dataset which has the size of nearly 10^7.

Table of metrics and Analysis

Table 1: metrics of the SVC, RFD, KNN, and SARIMXA models

Model Name	Mean Absolute Error (MAE)	MAE percentage (%)	Mean Squared Error (MSE)	Root Mean Squared Error (RMSE)	RMSE percentage (%)
Support Vector Classifier (SVC)	0.23	2.32	1.80	1.34	13.40
Random Forest Classifier (RFC)	0.08	0.75	0.44	0.21	2.10
K-Nearest Neighbour Regressor (KNN)	1.10	10.99	3.82	1.96	19.55
Time-series (SARIMAX)	0.83	8.30	3.75	1.94	19.35

From Table 1, we can clearly see that the SVC and RFC models have significantly higher performance than the KNN and SARIMAX model in their MAE error. In order to gain a deeper insight, visualization can be of significant help. As plots of a year can be too big as a time interval, the time interval is restricted to one day. The predictions of the 1st January at the location Tuas South are randomly picked and plotted as Figure 3. In order to make the analysis more convincing, performances of four models at extreme cases are studied and plotted as Figure 4 and 5, namely the sunniest day and the cloudiest day at the randomly selected location Bukit Timah.

Figure 3: CSR prediction by SVC, RF, KNN, and SARIMAX models on 1st January 2018 at Tuas South

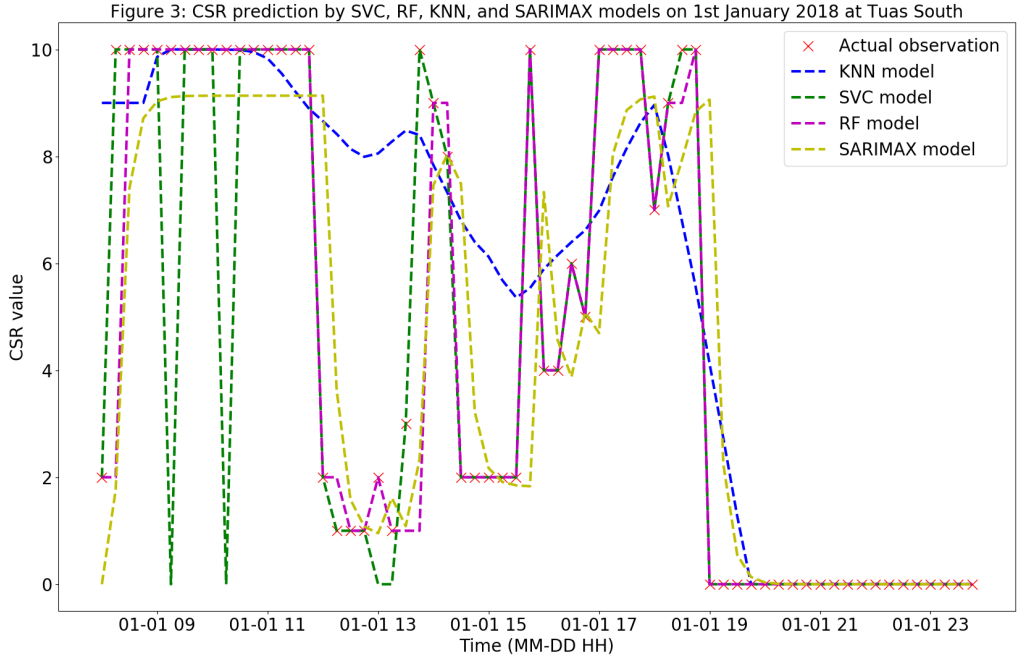


Figure 4: CSR prediction by SVC, RF, KNN, and SARIMAX models on 7th August 2018 at Bukit Timah (A sunny day specially picked)

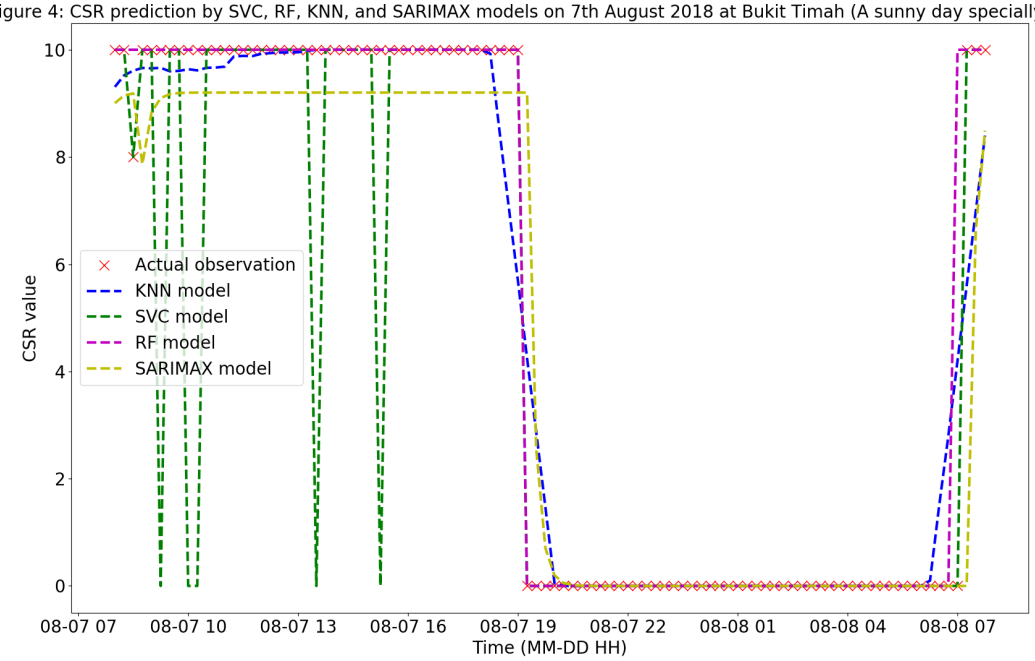
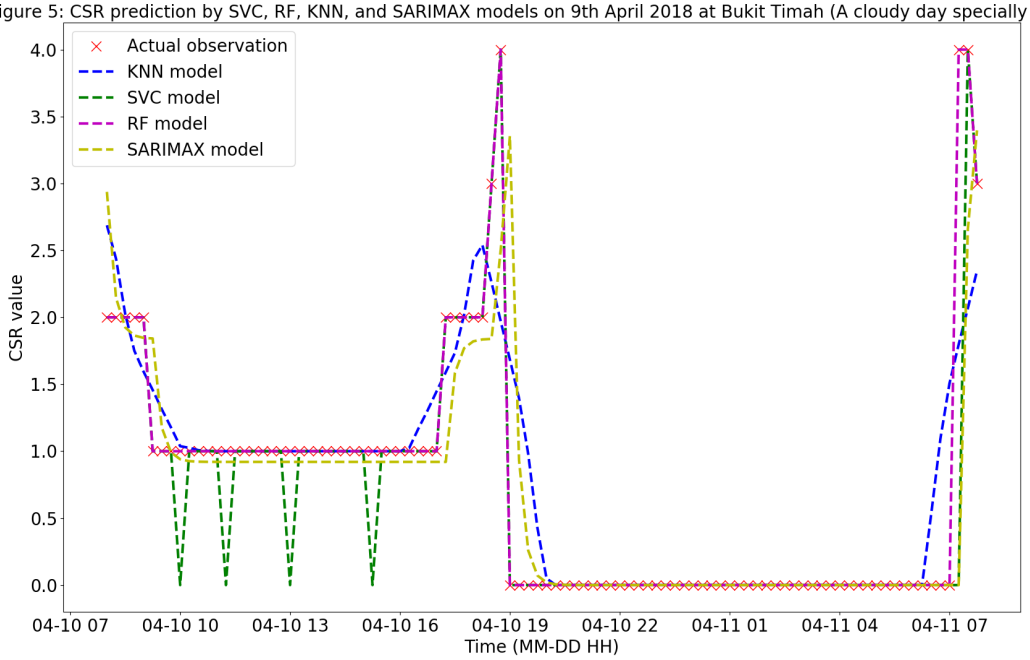


Figure 5: CSR prediction by SVC, RF, KNN, and SARIMAX models on 9th April 2018 at Bukit Timah (A cloudy day specially picked)



From the plots, we can clearly see that the SVC (green) and RF (magenta) model better fit the actual observations (red cross) compared with the KNN (blue) and SARIMAX (yellow) model, just as the metrics suggest. The KNN model seems to have the worst performance of all. In general, the CSR remains high in the morning, drops significantly during the early afternoon, rises up again and goes down to 0 at night time.

The SVC and RF models look quite similar generally while the SVC plot looks more complex than the RF. This can be achieved by a much more complex decision function behind the classification algorithm of SVC. While the complex function can help make the classification better fit the training set, it can also lead to the problem of over-fitting and cause the function unable to find the general patterns between features and labels. That means the function used by SVC is so powerful that it even validates some odd values rather than consider them as the outliers. This can help the SVC model to restimulate the solar radiation data to the greatest extent, while it also restricts the SVC model's application in other cases. In other words, the RF model can be more suitable to be employed to predict the solar radiation data in the next year as its decision function is more likely to draw the general patterns. However, by the metrics, the RF model has better performance than the SVC model and as the figure suggests, the steep change in the morning in the SVC plot is neither necessary nor accurate. This can be caused by interference from the future data and cause the decision function to detect them as the outliers. On the other side, the prediction by the RF model looks much smoother and very close to the line graph connecting the red points except certain

outliers. Combining the metrics and the figures, the RF model is a better choice than the SVC model, in both complexity and accuracy.

The KNN has very poor performance based on the figure as it totally ignores certain parts of the data in Figure 3. For example, the drop during the early afternoon is not completely reflected in the trend and this could lead to inaccuracy of the model and hence lowers the scores of the KNN model. This is partly due to the restriction of computing power. While building the model, the number of estimators, which is also the "K" in the KNN, can be decided. After testing a few numbers, 400 is finally chosen as this is the maximum number the memory of the computer running the program can accommodate. With a larger number of estimators, better precisions can be achieved as the function is more complex and capable to indicate the complex relationship between the dependent and independent variables. Yet a large number of estimators would also require larger memory space and the computer system would automatically kill the process if this goes beyond the memory space of the computer. The large memory space required is one of the reasons that restrict the performance of the KNN model. However, the MSE and RMSE values of the model are very close to the time series model, with the MAE value much larger. This can be the result of the algorithm as the Euclidean distance function is the same as the RMSE in their Mathematical definitions. As smaller Euclidean distance values are chosen, smaller MSE and RMSE values of the KNN model are achieved.

The SARIMAX model has a moderate estimation of the CSR values compared with the other three. While it less fits the actual observations compared with the SVC and RF model, its plot is much smoother than the RF model and can hence better represent the general trend of a day at the cost of losing some precisions. In other words, the SARIMAX model seldom makes extreme predictions and detect the extreme values 10 to be the outlier. This would certainly reduce the accuracy of the model. However, it also makes sure the SARIMAX model always reflects the general trend. As a result, in terms of the prediction of future solar radiation data, the SARIMAX model can be a good choice. However, as the focus of this study is to simulate the solar radiation data based on the WRF data provided, the RF model meets the purpose of the research and is still the best choice.

Overall, the RF model is the best choice in this study as it best simulates the solar radiation data in this certain year with the provided dataset and the engineering goal of this study is achieved with this model.

Conclusion and outlook

This study has explored different machine learning algorithms in order to find quick alternative methods to make solar radiation prediction rather than the currently most popular Weather Research and Forecast models. Several popular machine learning techniques such as Support Vector Machine, Artificial Neural Network, K_Nearest Neighbours and Time-series, have been studied and their performance in making the prediction has been evaluated based on different metrics including Mean Absolute Error and Root Mean Squared Error and plots of their trends. The greatest achievement of the study is to build a model based on Random Forest that is capable of completing the whole process including learning and prediction within only half an hour with relatively high accuracy. More machine learning techniques such as ensemble learning, which can combine two models to be a model with better performance, can be further explored and applied to the study and can be the future work of the research.

Reference and Bibliography

1. https://www.ema.gov.sg/solar_photovoltaic_systems.aspx (accessed on 20 Nov 2018) ↩↩

2. Ran Fu, David Feldman, Robert Margolis, Mike Woodhouse, and Kristen Ardani. U.S. solar photovoltaic system cost benchmark: Q1 2017. Sep.↩↩

3. Ernst, B., Oakleaf, B., Ahlstrom, M., Lange, M., Moehrlen, C., Lange, B., Focken, U.,Rohrig, K. Predicting the Wind. In Power and Energy Magazine, 5(6), 78-89, 2007.↩↩

4. Nils Andre Treiber, Stephan Spath, Justin Heinermann, Lueder von Bremen and Oliver Kramer. Comparison of Numerical Models and Statistical Learning for Wind Speed Prediction.↩↩

5. Sotiris Vardoulakis, Bernard E.A. Fisher, Koulis Pericleous, Norbert Gonzalez-Flesca. Modelling air quality in street canyons: a review. Atmospheric Environment, Elsevier, 2003, 37 (2), pp.155-182.↩

6. Aoife M.Foleyabd, Paul G.Leahyab, AntoninoMarvugliac, Eamon J.McKeoghab. Current methods and advances in forecasting of wind power generation ↩↩

7. Numerical Weather Prediction (Weather Models) https://www.weather.gov/media/ajk/brochures/NumericalWeatherPrediction.pdf (accessed on 15 Nov 2018)↩

8. T.M.Giannaros, V.Kotroni, and K.Lagouvardos. Predicting lightning activity in Greece with the weather research and forecasting (wrf) model. Atmospheric Research, vol. 156. 1 - 13, 2015.↩

9. Terren-Serrano, Guillermo. "Machine Learning Approach to Forecast Global Solar Radiation Time Series."(2016).[http://digitalrepository.unm.edu/ece_etds/249] (Stephan Rasp and Sebastian Lerch. Neural networks for post-processing ensemble weather forecasts. arXiv: 1805.09091v1 [star.ML] 23 May 2018)↩

10. Haupt, Sue & Kosovic, Branko. (2015). Big Data and Machine Learning for Applied Weather Forecasts Forecasting Solar Power for Utility Operations. 10.1109/SSCI.2015.79. ↩

11. https://www.mmm.ucar.edu/weather-research-and-forecasting-model (accessed on 15 Dec 2018)↩

12. U. Divya, Chitra Pasupathi. A machine learning approach to predict solar radiation for solar energy based devices. By International Journal of Computer Science & Engineering Technology (IJCSET)↩

13. http://fortune.com/2016/09/14/data-machine-learning-solar/ (accessed on 10 Nov 2018)↩

14. Ricardo Aler, Ricardo Martin, Jose M. Valls, and Ines M. Galvan. A Study of Machine Learning Techniques for Daily Solar Energy Forecasting using Numerical Weather Models.↩

15. Aoife M.Foleyabd, Paul G.Leahyab, AntoninoMarvugliac, Eamon J.McKeoghabCurrent methods and advances in forecasting of wind power generation ↩

16. Sanyam Gupta, Infumath K, Govind Singhal. Weather Prediction Using Normal Equation Method and Linear regression Techniques. (IJCSIT)↩

17. 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.↩

18. https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47(accessed on 30 Nov 2018)↩

19. https://scikit-learn.org/stable/modules/svm.html(accessed on 30 Nov 2018)↩

20. https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html#sklearn.svm.SVC(accessed on 30 Nov 2018)↩

21. I. Guyon, B. Boser, V. Vapnik - Advances in neural information processing 1993. "Automatic Capacity Tuning of Very Large VC-dimension Classifiers". ↩

22. C. Cortes, V. Vapnik - Machine Learning, 20, 273-297 (1995). "Support-vector networks"↩

23. Cyril Voyant, Gilles Notton, Soteris Kalogirou, Marie-Laure Nivet, Christophe Paoli, Fabrice Motte, Alexis Fouilloy. Machine Learning methods for solar radiation forecasting: a review. Horizon 2020 project (H2020-LCE-2014-3 - LCE-08/2014 - 646529) TILOS "Technology Innovation for the Local Scale, Optimum Integration of Battery Energy Storage".↩↩

24. Wei-Zhen Lu, Wen-Jian Wang. Potential assessment of the "Support Vector Machine" method in forecasting ambient air pollutant trend↩

25. https://www.saedsayad.com/k_nearest_neighbors_reg.htm(accessed on 1 Dec 2018)↩

26. Peter J Brockwell. Introduction to Time Series and Forecasting. Springer Texts in Statistics. Springer, 3rd ed. 2016.. edition, 2016. ↩

27. https://www.statsmodels.org/dev/generated/statsmodels.tsa.statespace.sarimax.SARIMAX.html(accessed on 1 Dec 2018)↩

28. Xie-Kang Wang, Wei-Zhen Lu. Seasonal Variant of air pollution index: Hong Kong case study↩

29. http://www.forecastingsolutions.com/arima.html (accessed on 15 Dec 2018)↩

30. https://newonlinecourses.science.psu.edu/stat510/node/67/ (accessed on 15 Dec 2018) ↩

31. https://medium.com/human-in-a-machine-world/mae-and-rmse-which-metric-is-better-e60ac3bde13d(accessed on 1 Dec 2018)↩

32. https://www.vernier.com/til/1014/(accessed on 1 Dec 2018)↩

33. Seckin Karasu, Aytac Altan, Zehra Sarac, Rifat Hacioglu. Prediction of solar radiation based on machine learning methods. Bülent Ecevit University (BAP Project No: 2012-17-15-01 and 2014-75737790-01) and International Engineering Research Symposium-UMAS 2017(Duzce University) ↩