Solar Photovoltaic Prediction Analysis with Machine Learning Techniques

Jose Garcia

School of Engineering and Applied Science, The George Washington University

EMSE 6401 Data Science Introduction and Practicum

Dr. Benjamin Harvey

December 15, 2020

School of Engineering and Applied Science

The George Washington University

Abstract

Forecasting solar energy is an important task to manage solar power systems, ensure power supply, and reduce operational risks. In this study, several machine learning models were compared in order to predict solar power from a power plant in Berkeley, California. The methodology in this study was applied using a distributed computing framework on Apache Spark. Data was collected from public databases and includes weather measurements recorded in 2008 and 2009. The results show that the random forest regression algorithm performed the best for predicting solar power, with an R-squared value of 89%. Finally, this model allowed for the identification of which weather parameters are most important in predicting solar power.

1. Introduction:

Solar photovoltaic (PV) is a technology that converts sunlight from solar irradiation into electricity under the photovoltaic effect phenomenon. In the coming years, Global PV is expected to grow at an accelerated rate due to technological advances, policy support, and cost reductions. Additionally, solar energy is an alternative source of power with the potential to cover energy demand in a world with an increasing population, while also combating climate change and reducing greenhouse emissions (IEA, 2020). Although it comes with many environmental benefits, this technology is strongly dependent on weather conditions. Consequently, PV presents challenges and limitations due to its intermittency, instability and unreliability (Amarasinghe & Abeyguanawardane, 2018).

The purpose of this project is to predict solar power generation from a power plant located in Berkeley, CA, by applying several machine learning and data analysis methodology. Predicting solar power helps to manage energy systems, optimize decision making processes, reduce operational costs, balance electrical loads, and reduce financial risks (Xu et al., 2019).

Several authors propose prediction models to forecast solar energy from measured weather records. Amarasinghe and Abeyguanawardane (2018) propose a comparison study of the application of machine learning techniques versus traditional methods to predict solar power in a power plant located in Sri Lanka. Similarly, Carrera and Kim (2020) present a comparison framework for forecasting the solar power generated from a solar power plant located in South Korea. This framework consists of: data collection, data preprocessing, cross-validation, ten-fold CV, and selection subset. In this model, machine learning algorithms are analyzed and compared by their metrics for optimal model selection. In addition, Kim et al., (2019) propose a machine

learning model that predicts solar power using weather information provided by weather agencies. The results show that a random forest regression algorithm performed efficiently, achieving an R-squared value of 70.5%. Finally, Xu, Liu and Long (2019) study a hybrid distributed computing framework on Apache Spark to forecast wind speed. This framework improves computation speed and forecasts wind speed with accuracy. Even though this study is related to a different technology, the application of distributed computing and the use of weather data is a valuable reference.

In this study, available datasets were identified and collected to apply machine learning models and predict solar power generated in a plant located in California Berkeley. This paper is divided into three sections: Methodology, Results and Conclusion.

2. Methodology

The methodology in this study includes data identification and acquisition, exploratory data analysis, data cleaning and preprocessing, machine learning modeling, and evaluation of results. Pandas, Scikit-Learn and Spark-MLlib libraries were used in a parallel distributed computed environment to increase computational speed.

2.1 Data identification and Acquisition

Two datasets were identified and collected in .csv format from different public data sources: Kaggle.com and the National Solar Radiation Database, NSRDB. First, PV energy generated in a solar power plant located in Berkeley, CA was collected from Kaggle.com database. This dataset is referred to as 'Power Plant Dataset' in further analysis, and its data was collected in a three-hour resolution, from August 2008 to August 2009. Table 1 describes the variables collected and their respective units (Kaggle, 2020).

Variable	Units	Variable	Units
Year	-	**Wind Speed**	Not specified
Month	-	**Sky Cover**	Categorical
Day of Year	-	**Visibility**	Not specified
Hour	-	**Relative Humidity**	-
Is Daylight	True / False	**Barometric Pressure**	Not specified
Distance to Solar Noon	Not specified	**Wind Direction**	-
Temperature	Fahrenheit	**Power Generated**	Watts

Table1. Power Plant Dataset, units.

Next, data from the NSRDB was used...The NSRDB is a collection of solar radiation measurements and meteorological data. This database enables the estimation of the amount of solar irradiation at a given time and location. Through the NSRDB's Application Programming Interface, a dataset was collected by filtering Berkeley, CA as the location and the time range from Power Plant Dataset. This dataset is referred to as 'Weather Dataset' in further analysis, and its data was collected in a fifteen minute resolution. Table 2 details the variables collected and their respective units (NSRDB, 2020).

Variable	Units	Variable	Units
Year	-	**Cloud Type**	Categorical
Month	-	**Dew Point**	Celsius
Day	-	**Solar Zenith Angle**	Degree
Hour	Hour	**Surface Albedo**	-
Minute	Minutes	**Wind Speed**	m/s
DHI	Watts/m2	**Relative Humidity**	%
DNI	Watts/m2	**Temperature**	Celsius
GHI	Watts/m2	**Pressure**	mbar

Table 2.: Weather Dataset, units

2.2 *Exploratory Data Analysis*

The strategy in this step consisted of studying the different weather variables collected in the datasets (features), as well as the power output (target). Time series statistical analyses were

evaluated to describe trends, seasonality, distribution, range, and standard deviation. Fig.1 shows the Power Generated vs Time in a 1 month sample of the ‘Power Plant Dataset’.

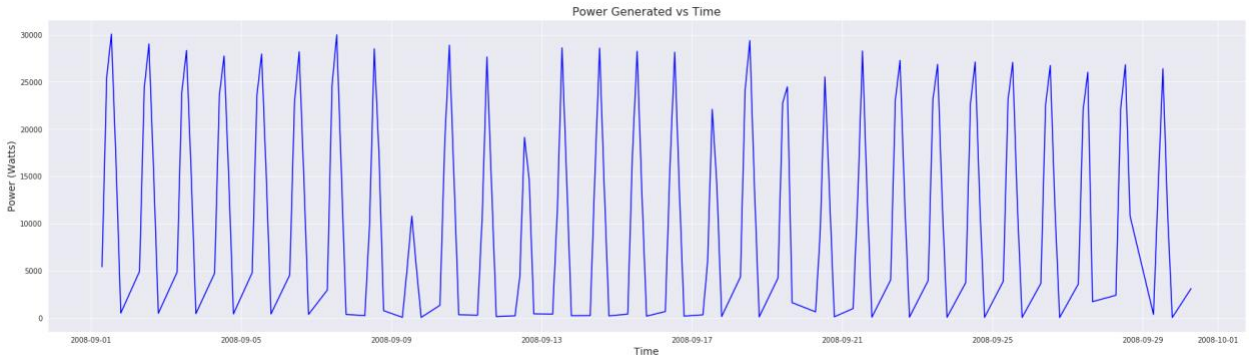

Fig.1 Power Generated in September 2009.

‘Power Generated’ and solar irradiance ‘GHI’ were decomposed into their trend, seasonality, and residual components. This statistical analysis was performed to study the cyclical behavior of these variables (trend), and their seasonality. Fig.2 shows a strong daily seasonality, and a trend related to the different weather seasons. Winter months are related with lowest energy output, while summer months with highest.

Fig.3 shows the difference in the distribution of some of the weather variables, while Table 1 shows basic statistics of the collected data. Features consists of different weather measurements with their respective range, standard deviation and average values. These differences are considered in the cleaning and preprocessing section.

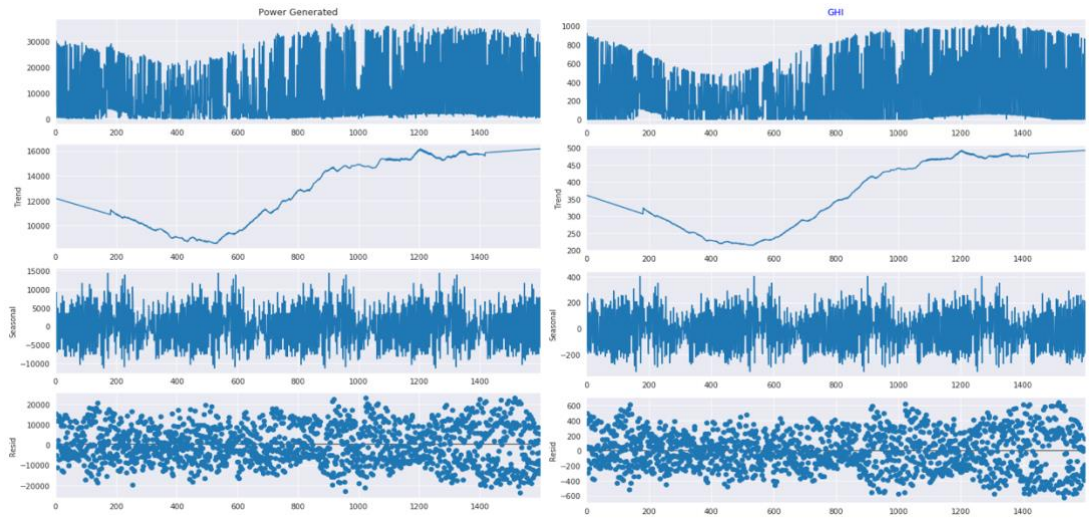

Fig.2 Trend and seasonal Decomposition of Power Generated and Solar Irradiance GHI

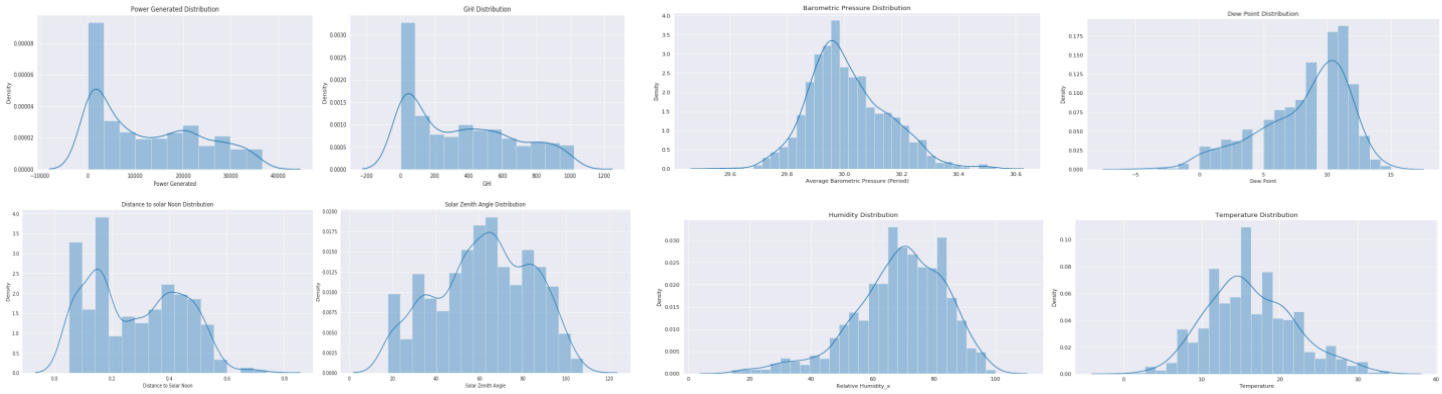

Fig.3 Distribution of Power Generated and Weather parameters

	Power Generated	Solar Noon	Visibility	Humidity	Pressure	DHI	DNI	GHI	Dew Point
Mean	12738	0.28	9.49	68.75	30	108.64	433.33	364	8.18
Std	10987	0.16	1.47	14.99	0.13	98.12	361.11	309	3.52
Min	1	0.05	0.25	14	29.56	0	0	0	-5
Max	36580	4	10	100	30.56	477	970	1021	15

	Zenith Angle	Albedo	Wind Speed	Temperature
Mean	61.95	0.11	2.28	16.30
Std	21.83	0.04	0.98	5.64
Min	18	0.1	0.1	0
Max	108	0.12	6.6	34

Table 3. Statistical summary of collected Data

2.3 Data Cleaning and Preprocessing

A. Missing values

One missing value was identified and eliminated from the 'Power Plant Dataset' dataset. Additionally, night records were discarded by filtering recordings where the variable 'Is Daylight' equals true and 'Power Generated' equals zero.

B. Encoding categorical variables

'Power Plant Dataset' and 'Weather Dataset' presented variables already encoded. The 'Cloud Type' feature was encoded as follows: Clear: 0, Probably clear: 1, Fog: 2, Water: 3, Supercooled water: 4, Mixed: 5, Opaque ice: 6, Cirrus: 7, Overlapping: 8, Overshooting: 9, Unknown: 10, Dust: 11 and Smoke: 12. In the same way, the 'Sky Cover' variable was encoded as follows: Clear: 0, Partially clear: 1, Partly cloudy: 2, Cloudy: 3 and Over casted: 4.

C. Rearranging Datasets

Most of the weather parameters from the 'Power Plant Dataset' were discarded because their units were not specified. However, 'Power Generated', 'Sky Cover', 'Humidity', 'Visibility' and 'Barometric Pressure' variables were kept and merged into 'Weather Dataset.' This final 'Weather Dataset' consists of 1600 rows and 16 columns of power and weather recordings in a 3-hour resolution. From this point, the analysis was done in the final 'Weather Dataset.'

D. Data Normalization

To optimize and reduce computational costs when training models, features and target variables in the final 'Weather Dataset' were normalized between 0 and 1. Fig. 4 shows the resulting distribution of the features and the target variable (Amarasinghe & Abeyguanawardane, 2018).

In the last step, the final ‘Weather Dataset’ was divided into ‘Training dataset’ and ‘Test dataset’, 70% and 30% respectively from the ‘Weather Dataset.’ Finally, Training and Test datasets were transformed to a parallelized Spark DataFrame in order to apply Apache Spark Machine Learning capabilities.

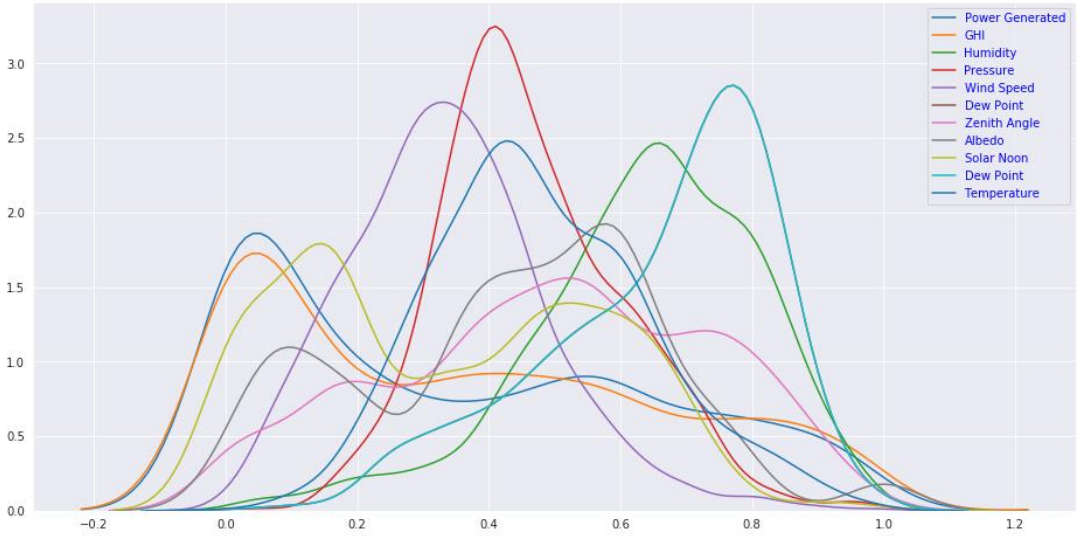

Fig. 4. Target variable and Features Normalized Distribution Plot

2.4 Machine Learning Modelling

In this section, regression models from Spark Machine Learning Library (MLlib) and Scikit-Learn library were applied to predict the solar power generated. Moreover, ML-Flow and Cross validation Spark methodologies were used to track experiments and to optimize models’ accuracy. In these models, the features used to make predictions were selected by analyzing the correlation heatmap in Fig.5. ‘Wind Speed’, ‘Barometric Pressure’, and ‘Surface Albedo’ were discarded since they are not correlated with the target variable.

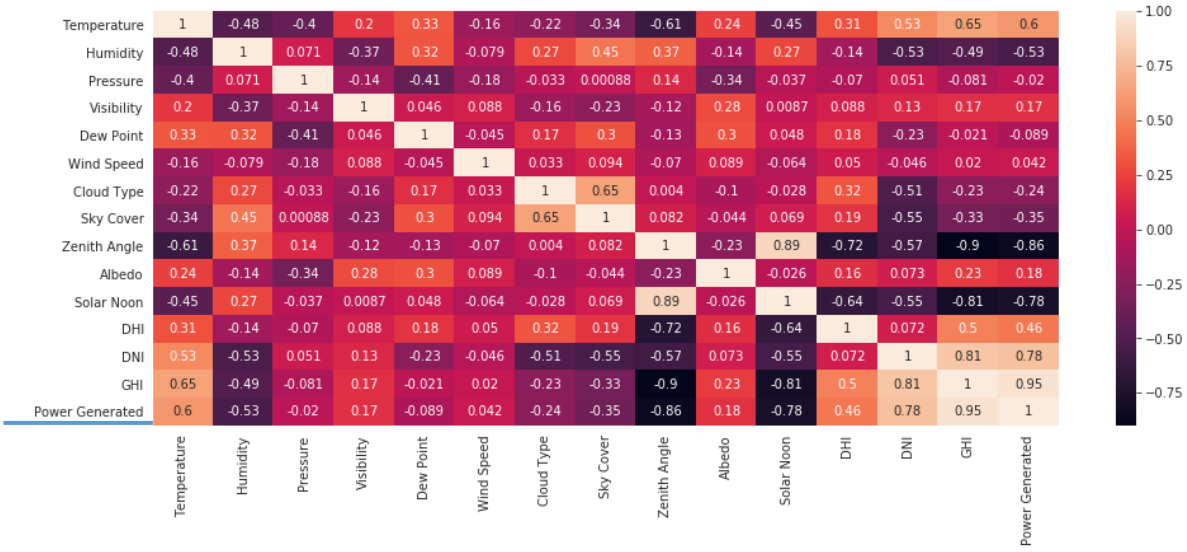

Fig. 5: Heatmap correlation plot.

After selecting the features, several models from Spark MLlib were trained on the Train dataset. These models included *Decision Tree Regression* and *Random Forest*. In addition,, *Support Vector Machine* model from the Scikit-Learn library was trained as well. All these models were trained by using default parameters, and their prediction performance is shown in the next section. With the lowest MAE value, the *Random Forest Regression Model* was selected to be optimized. Finally, A 3-fold cross validation and grid-search methodologies were applied over the 'Training Dataset' to optimize model accuracy, reduce model overfitting, and obtain the best hyper parameters (Carrera and Kim, 2020).

3. Results and Discussion

This section presents the results of the machine learning methods described in the previous section. Each of the models was tested by predicting Power Generated from the 'Test Dataset'. Their performance was evaluated by comparing the Root Mean Square Error (RMSE), Mean Absolute Error (MAE) and the regression score (R2). A summary of these values is shown in Table 4.

Results show that a *Random Forest* model with *maxDepth*=10 and *number of trees = 100* predicted Power Generated with *RMSE = 0.09*, *MAE =0.05* and *R2= 0.92*. Fig.6 shows a sample of the prediction of the last two days, Fig.7 shows the high correlation of the predicted values and the real values and Fig.8 shows the feature importance in the model. The Feature Importance graph shows that power generated in this dataset is influenced the most by solar conditions, rather than atmospheric weather parameters. Due to its location, this solar plant operates under stable weather conditions. Therefore, this model should be applied under similar weather parameters.

Model	RMSE	MAE	R2
Decision Tree	0.125	0.067	0.86
Random Forest	0.11	0.067	0.89
SVM	0.012	0.069	0.89
RF Optimized	0.09	0.05	0.92

Table 4: Model Evaluation Metrics Results

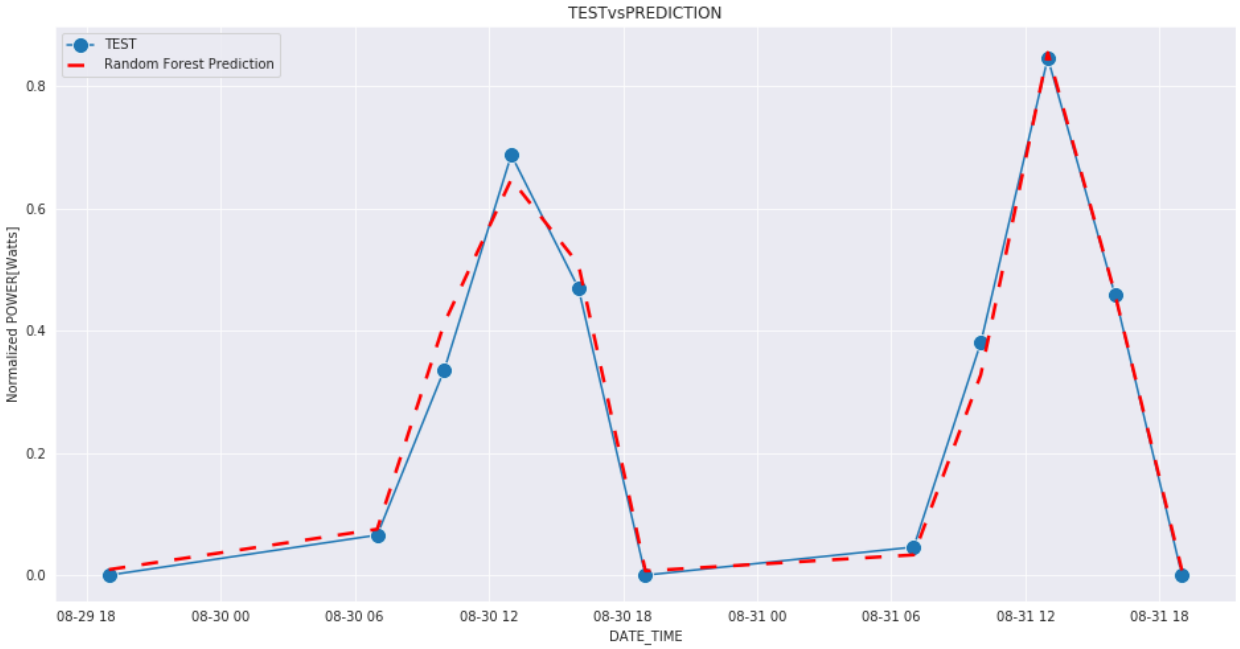

Fig.6 Two last days prediction using *Random Forest* model

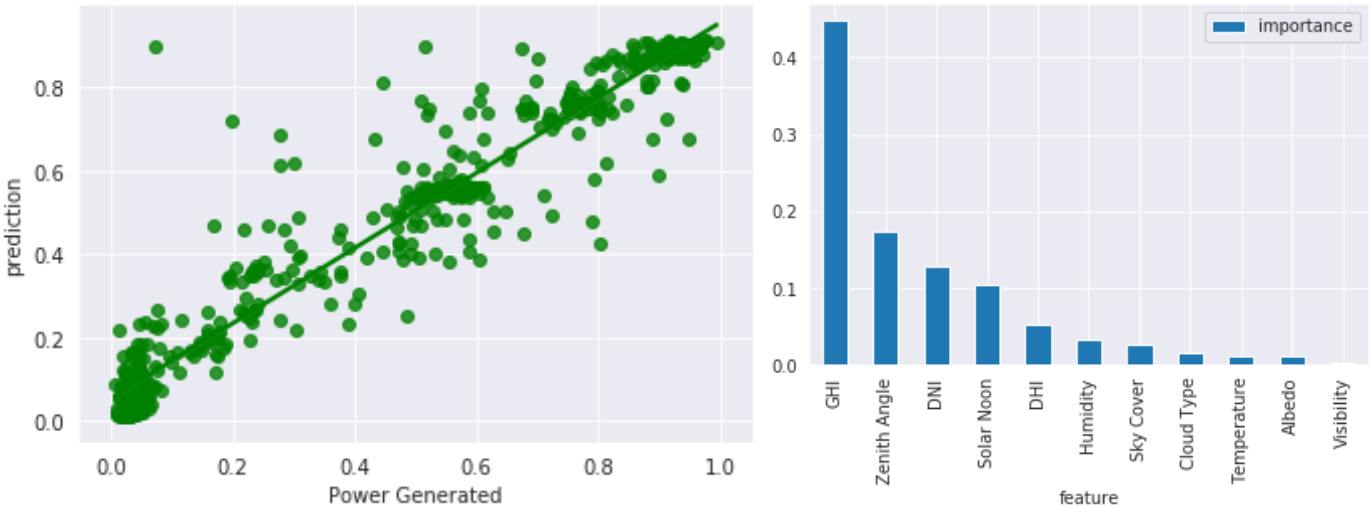

Fig.7 Power Generated vs Prediction and Features Importance

4. Conclusion

In this study, data was collected, cleaned, preprocessed and analyzed in order to predict Solar Power output from Berkeley power plant. From the evaluation of the models, it was found that a Random Forest regression algorithm performs the best, with an R-squared value of 89%, RMSE = 0.11 and MAE= 0.05. For this power plant, solar power is mostly influenced by Global Horizontal Solar Irradiance, and Zenith Angle. However, this relationship is true under similar weather conditions. The data analysis methodology and Spark capabilities applied in this study were critical to process big data, extract meaningful information, and to make predictions of solar power. Further analyses might include: building a data streaming framework to store in real time new weather and power records, predicting solar power, and detecting and predicting power plant equipment failure by using Spark capabilities.

References

Amarasinghe P.A.G.M, Abeyguanawardane S.K., (2018, September 28). *Application of machine learning algorithms for solar power forecasting in Sri Lanka.* 2018 2nd International Conference on Electrical Engineering (EECon). doi: 10.1109/EECon.2018.8541017

Cady, F., (2017). *The Data Science Handbook*. John Wiley & Sons, Inc.

Carrera, B., Kim, K., (2020 June 1). *Comparison Analysis of Machine Learning Techniques for Photovoltaic Prediction Using Weather Sensor Data.* Department of Industrial and Management Engineering, Incheon National University. doi: 10.3390/s20113129.

International Energy Agency, IEA (2020, December). *Renewables 2020 Analysis and Forecast to 2025.* iea.org. https://www.iea.org/reports/renewables-2020

Kaggle, (2020). *Solar power Generation: A model that predicts the output of a solar power system.* https://www.kaggle.com/vipulgote4/solar-power-generation

Kim, S., Jung, J., Sim, M., (2019 March 12). *A Two-Step Approach to Solar Power Generation Prediction Based on Weather Data Using Machine Learning.* Department of Industrial & Management Systems Engineering, Kyung Hee University. doi: 10.3390/su11051501.

McKinney, W., (2018). *Python for Data Analysis. Data Wrangling with Pandas, Numpy, and IPython.* (2nd Edition). 1005 Gravenstein Highway North, Sebastopol: O'Reilly Media, Inc.

NSRDB, (2020). *National Solar Radiation Database*. https://nsrdb.nrel.gov

Xu, Y., Liu H., Long Z., (2019, November 11). *A distributed computing framework for wind speed big data forecasting on Apache Spark.* Institute of Artificial Intelligence & Robotics (IAIR). doi: https://doi.org/10.1016/j.seta.2019.100582