# Predictive Model for PV Output

While solar cells are great renewable energy sources, there is a bit of unknown as they are dependent on unpredictable weather and solar emissions. SPC has been having issues with back up power when output is low so they have set out to have a model built to predict output so the company can be better prepared for power shortages.

# Data

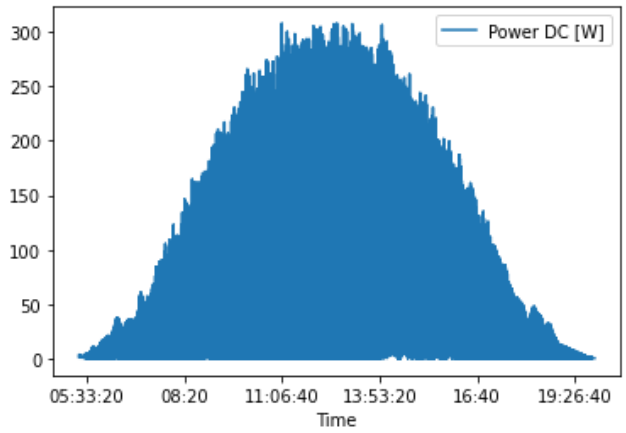
The data sets are provided by Sunlab in Faro, Spain and include PV cell performance as well as weather station data. Many aspects of the weather (temperature, radiation, UV, etc.) and PV temperatures are taken into account along with the angle of the PV cell will also be used. The data covers 3 years, 2 different PV types, and different angles to consider when building the model. The link is here:
https://opendata.edp.com/explore/?refine.keyword=visible&sort=modified

# Methods

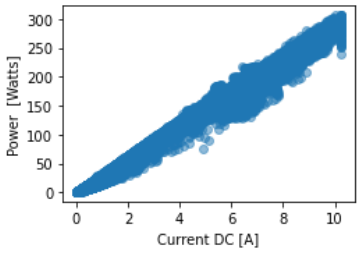
The goal of the project is to predict the voltage output given different characteristics of weather and the solar cell itself. Three models will be initially considered and, if they are not sufficient, others will be explored. First, a very basic mean model will be deployed to ensure the weather and PV features do influence the Voltage. Second, a linear regression and random forest will be considered. A logistic regression will be added if the linear regression and random regression do not provide sufficient predictive power.

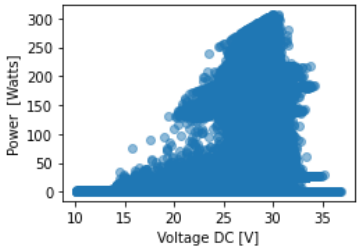
# Data Cleaning & EDA

Data was generally in good shape, but provided a challenge due to the size. Three years of data with ~200,000 entries each meant 600,000 rows of minute by minute PV cell data. Each year did contain missing time frames and some outliers. Only PV cell type A was considered in order to cut down on the number of rows for the model. As a note, the measurements run from 5:33 in the morning to 19:26 in the evening (~7:30 PM at night). January 1: sunrise 7:35 AM and sunset at 5:33 PM. July 15 : 6:24 AM, 8:51 PM. The daily power output is plotted:




An additional issue is, by definition, Power = Voltage * Current, both of which are provided in the PV data. Current is more linearly correlated, but voltage also has a correlation. In order to focus on weather and solar influence of Power, these two features were removed when building the models.

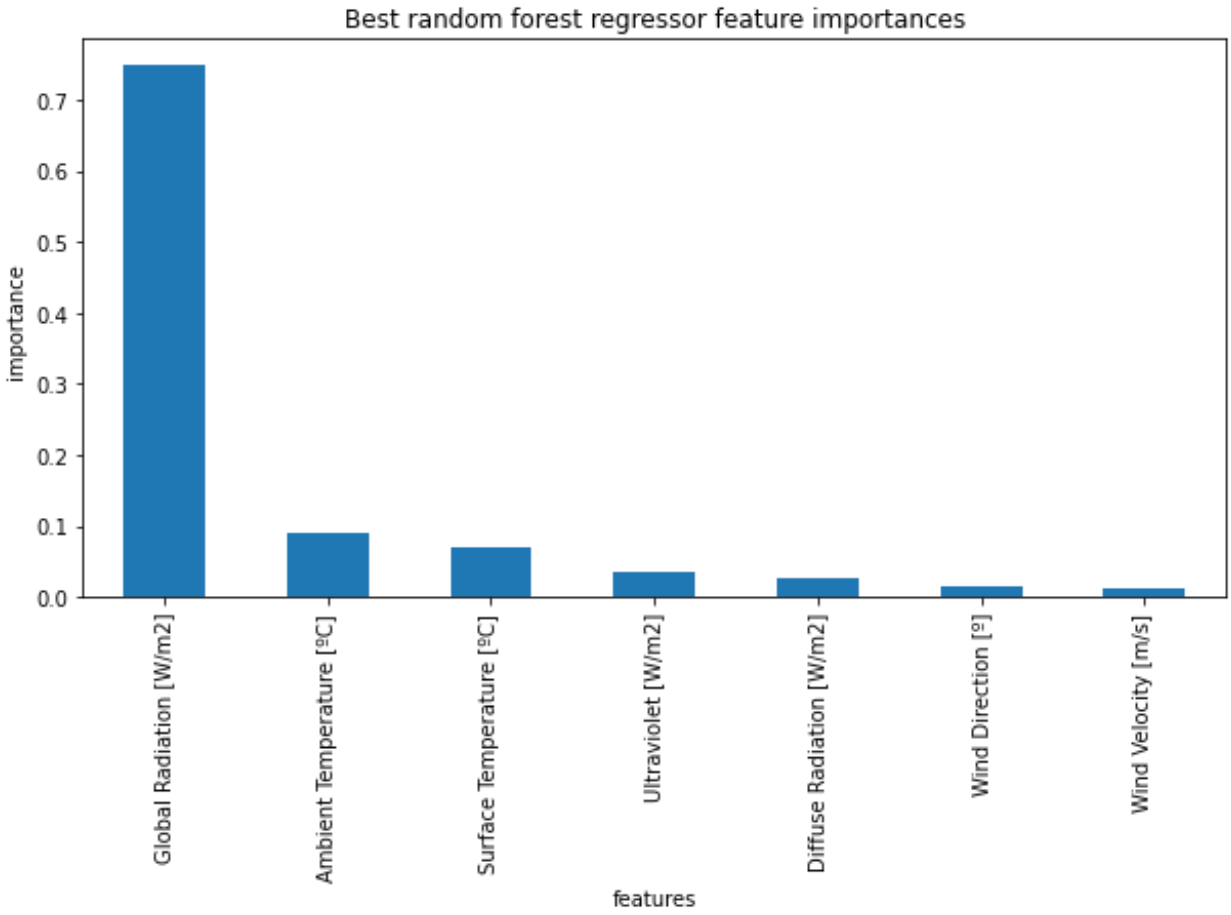



The weather features did not provide a clear correlation with the Power output.

# Algorithms & ML

Sci-ki-tlearn package was used to create all models. First, a mean DummyRegressor was used to ensure weather does indeed influence Power output. The results were off by about 64.4 Watts, which is rather significant. For the linear regression model, k=7 gave the best results, and indicated Global Radiation was the most substantial influence on Power Output, followed by PV Cell Surface Temperature. Interestingly Ultraviolet Radiation was not important. The random forest model was the best, but with extremely similar results as the linear model. Global radiation dominated the feature importance, followed by ambient and surface temperatures, and then ultraviolet at less than 10% importance.

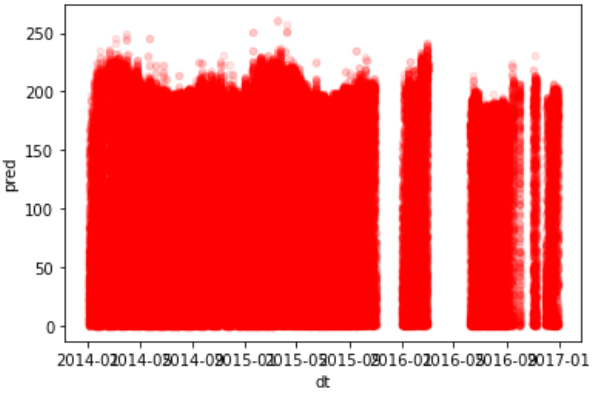


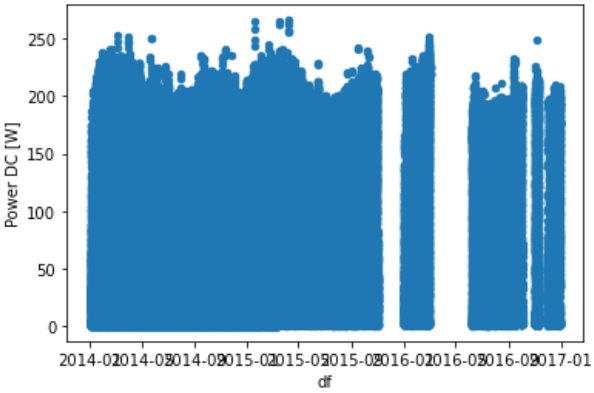

The winner was the Random Forest model. However, the best RF model uses 1530 n_estimators and comes at an extremely high computational cost compared to that of the linear regression with similar results. When moving into production, this computational cost must be taken into consideration.

| Algorithm | MAE |
|---|---|
| Mean | 63.188 |
| Linear Regression | 19.887 |
| Random Forest | 11.533 |

# Predictions

MAE = 2.61
MSE = 19.68
R2 = 0.996


pred
250
200
150
100
50
0
2014-01
2014-05
2014-09
2015-01
2015-05
2015-09
2016-01
2016-05
2016-09
2017-01
dt


Power DC [W]
250
200
150
100
50
0
2014-01
2014-05
2014-09
2015-01
2015-05
2015-09
2016-01
2016-05
2016-09
2017-01
df