Predicting Solar Panel Output - A Case Study

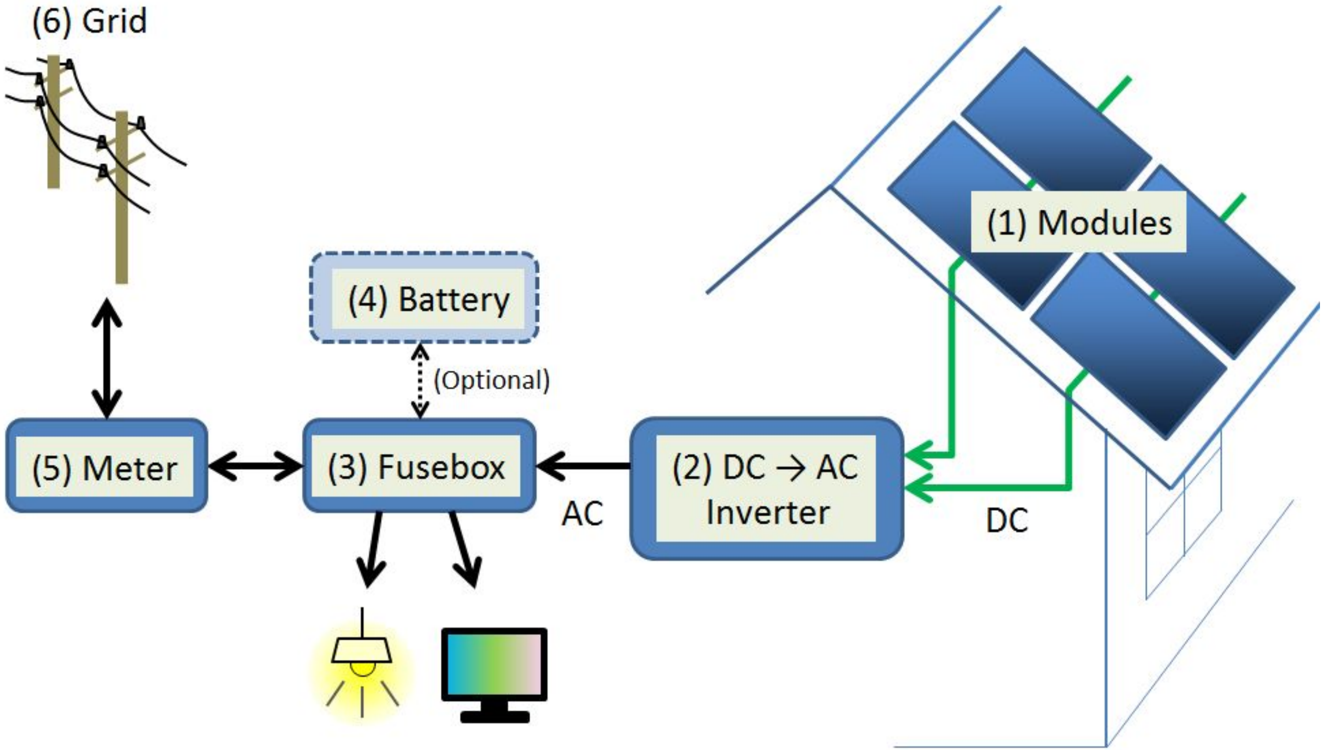
Corina Lentz
Garrett Hart
Joseph Hicks
7/27/2021

Table of Contents

Problem Statement

Forecast the energy output of a residential solar array located in Antwerp, Belgium using previous output data from the array combined with weather data.

Background



Residential PV System Schematic courtesy of Wikipedia. Image found here:
https://en.wikipedia.org/wiki/Photovoltaic_system

The Project Solar Array

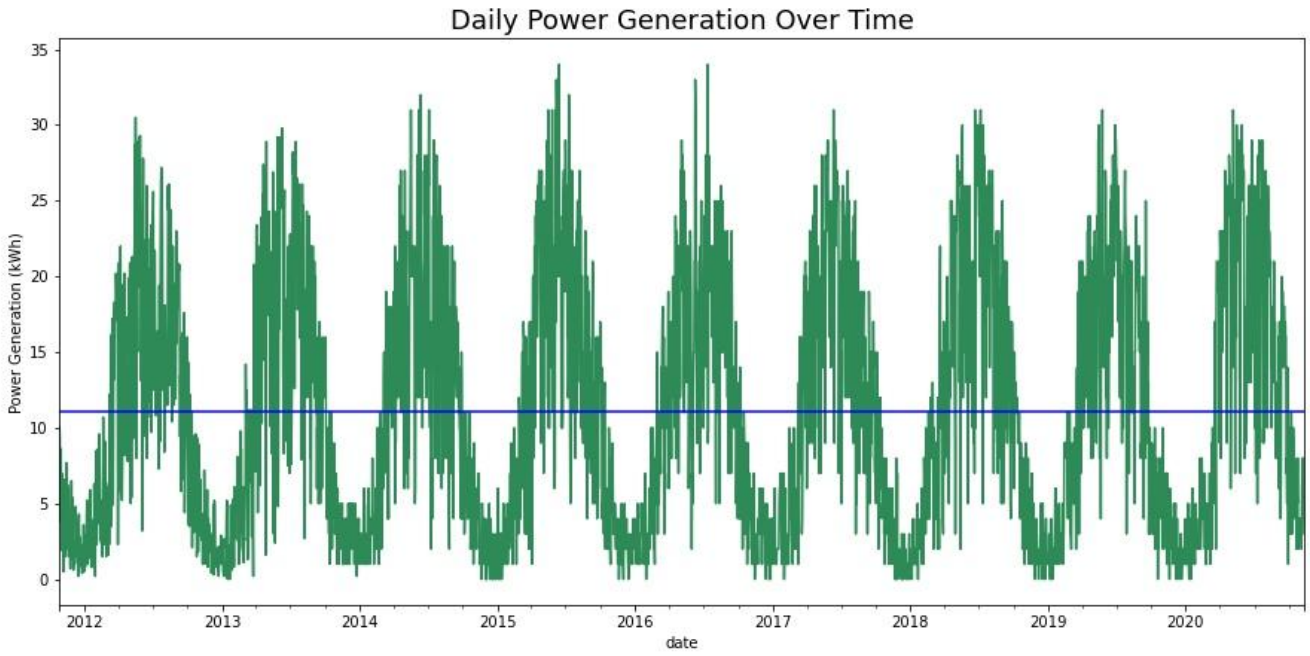
- Residential photovoltaic (PV) system with an array of 24 polycrystalline panels made by S-Energy.
- Antwerp, Belgium has temperatures averaging 32-77 ° F and 2.2-3.1" of precipitation a month.
- Belgium has a temperate maritime climate.

Data

- Electrical output data was recorded everyday from 10-26-2011 through 11-10-2020 creating 3,304 observations.

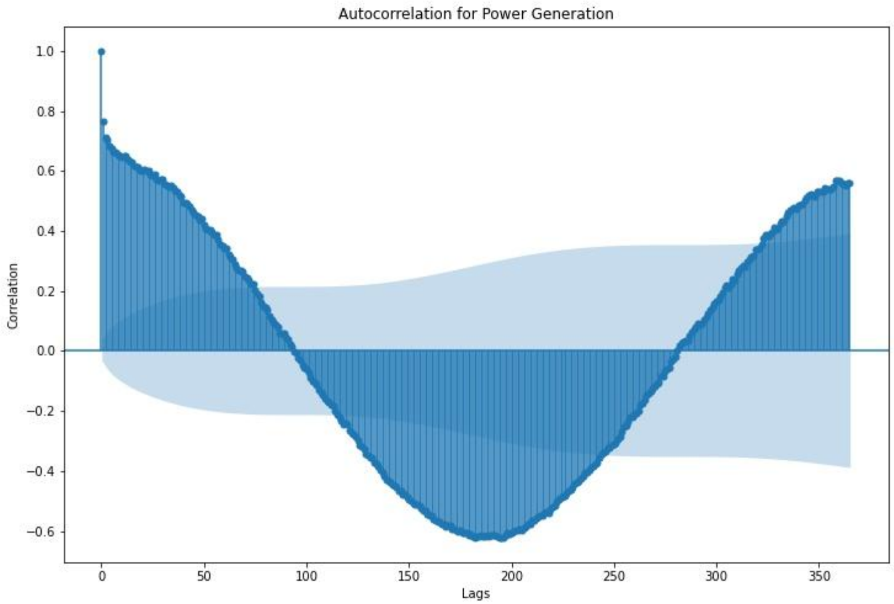
- Weather data was scraped from World Weather Online over the same timeframe.

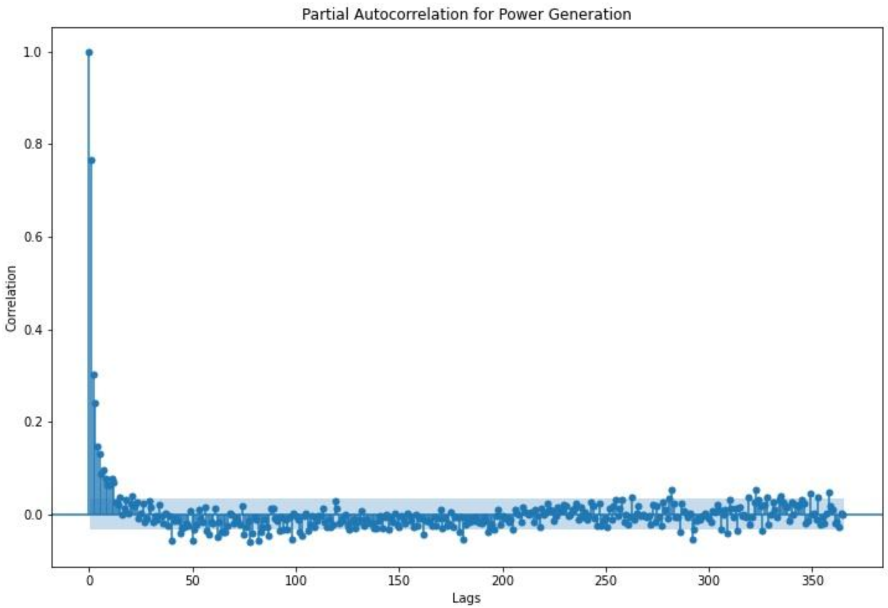
Exploratory Data Analysis



- There is a large amount of seasonality in the data as the hours of sunlight per day and weather changes
- Average power generated per day: 11.04 kWh

Exploratory Data Analysis

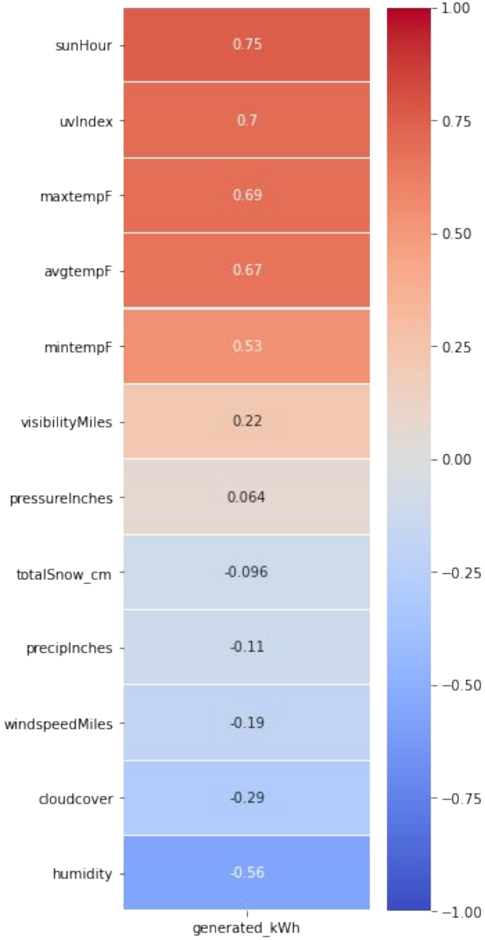




- The plot for autocorrelation over a year's time confirms there is significant seasonality and that there is no clear trend otherwise
- The partial autocorrelation plot shows only the most recent lag terms are statistically significant

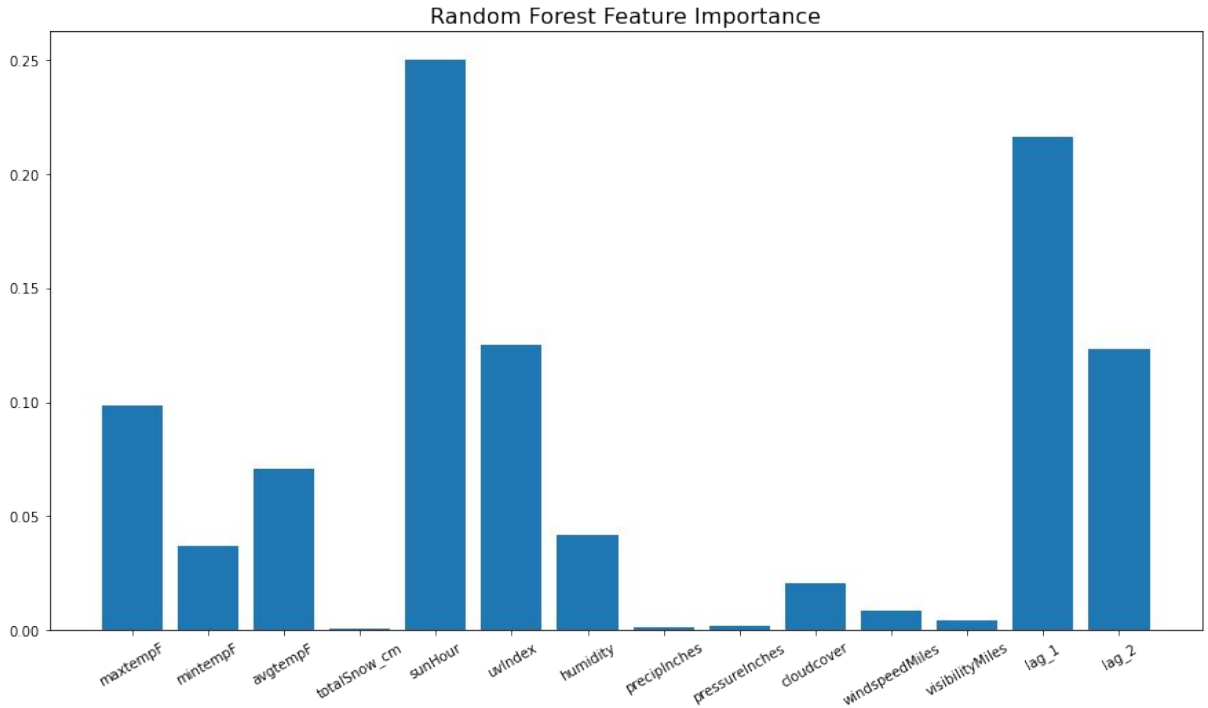
Exploratory Data Analysis

- The heatmap on the right depicts the correlation of the weather features with our target variable, Daily Power Generation
- The hours of sun per day has the largest positive affect on our target, and humidity has by far the most negative impact



Modeling - Random Forest Regression

- Random Forest Regression models are great predictive models that don't require much data pre-processing
- Baseline RMSE: 8.44
  Training RMSE: 3.83
  Testing RMSE: 4.56
- Training R squared: 0.78
  Testing R squared: 0.71
- The feature importance on the right shows the impact of each feature on our best model - the time series lag terms have a large affect



Modeling - Linear Regression

- Created Linear Regression model for interpretability
- Created the following coefficients:

Feature	UV Index	Precip (Inches)	Minimum Temperature	Average Temperature	Hours of sunlight
Coefficient Value	1.28	1.17	-0.48	0.55	0.42

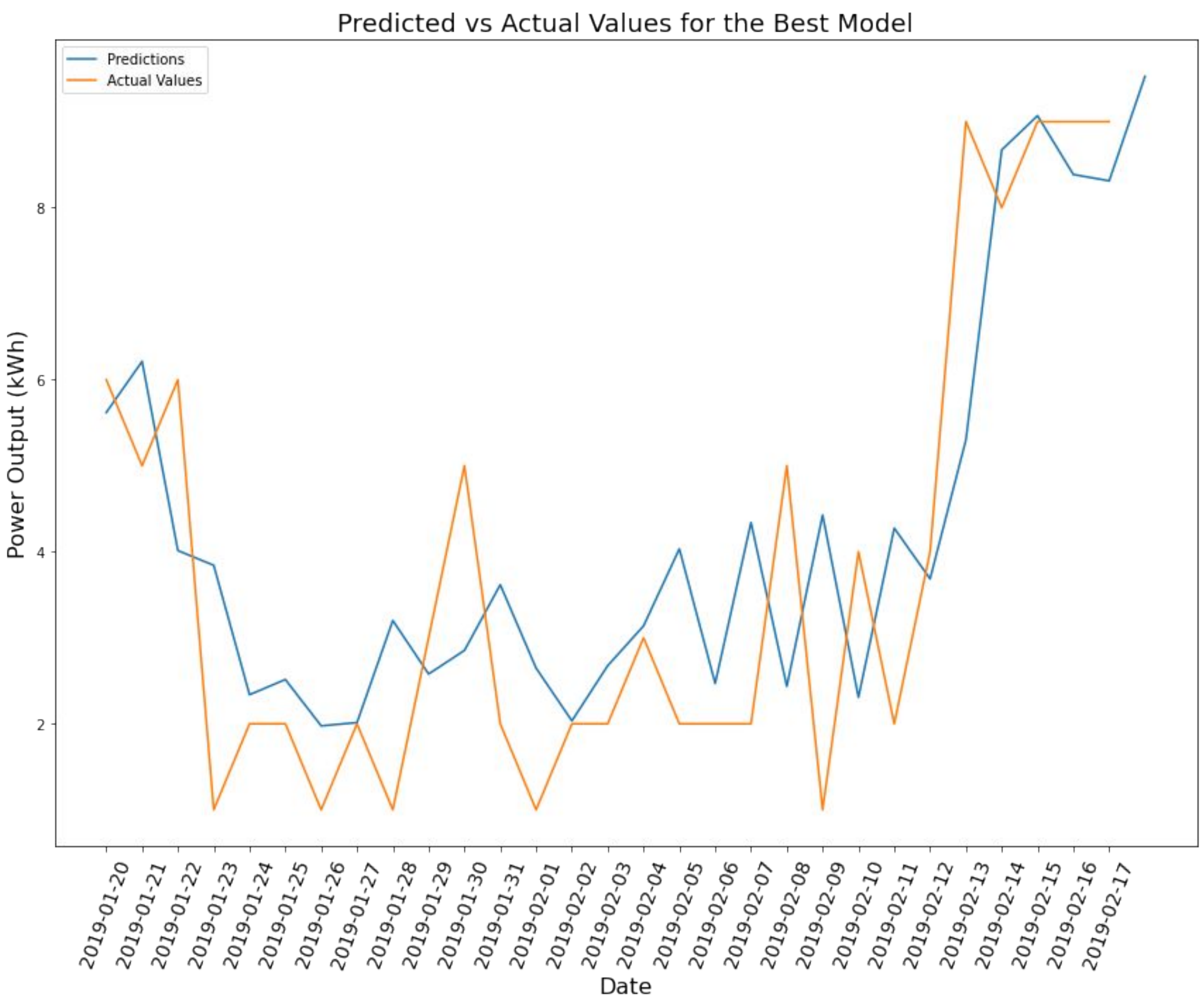
- Test RMSE was 4.58
- Testing set R squared score was 0.7
- Training set R squared score was 0.71

Modeling - Bagging Regressor

- Base model was a Decision Tree Regressor
- Bagging aggregates the decisions of many models over different random sets of the data
- Performed slightly better than the last two models
- Test RMSE was 4.40
- Testing set R squared score was 0.73
- Training set R squared score was 0.70

Conclusions

- Created a Bagging Regressor model with 73% accuracy
- Created interpretable Linear Regression and Random Forest models
  - Sunlight hours, UV Index , high temperature and humidity affect output the most
- Can account for 73% of variance in output
- Model can take in weather forecasts to make predictions on power output


Predicted vs Actual Values for the Best Model
Predictions
Actual Values
Power Output (kWh)
8
6
4
2
Date
2019-01-20
2019-01-21
2019-01-22
2019-01-23
2019-01-24
2019-01-25
2019-01-26
2019-01-27
2019-01-28
2019-01-29
2019-01-30
2019-01-31
2019-02-01
2019-02-02
2019-02-03
2019-02-04
2019-02-05
2019-02-06
2019-02-07
2019-02-08
2019-02-09
2019-02-10
2019-02-11
2019-02-12
2019-02-13
2019-02-14
2019-02-15
2019-02-16
2019-02-17

Next Steps & Recommendations

- Find more data from arrays with the same conditions
  - More data for the model overall
- Make new models for different variables
  - Equatorial solar arrays
  - Different panel technologies
- Recall assumptions made for future models
  - Polycrystalline Silicon panels
  - Latitude and climate
  - Averaging weather conditions

Any Questions?