DATA 698 Capstone Project

Accurate Solar Energy Generation Predictions Using Weather Forecasts

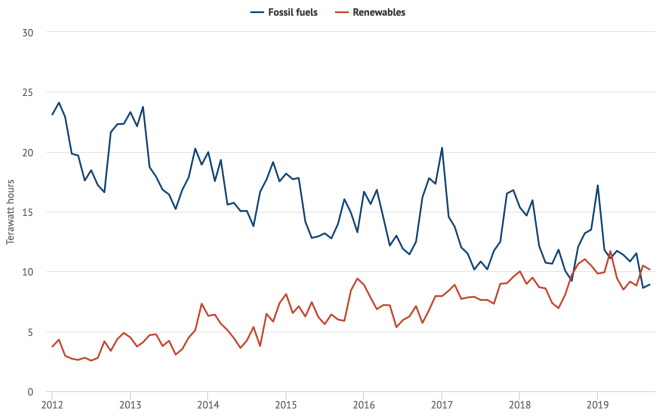
Jose A. Mawyin
December 16, 2020

Abstract

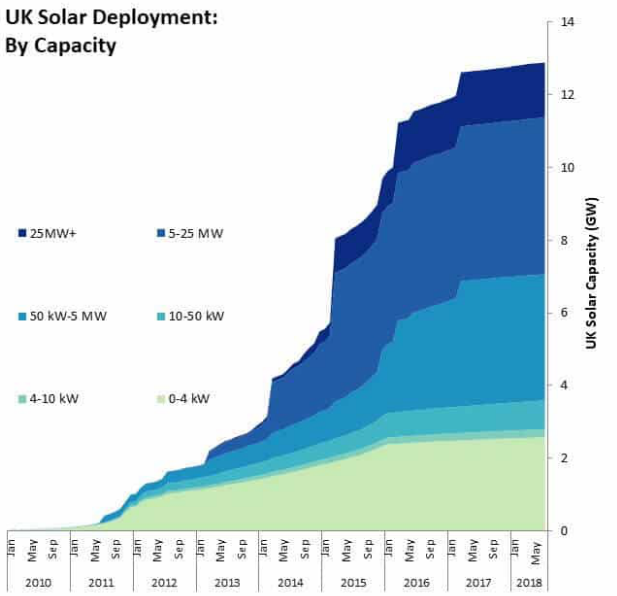
Accurate forecasting of solar electricity generation in the United Kingdom at different levels of spatial granularity was accomplished using supervised machine learning algorithms on weather parameters.

Project Aim

In the UK as in the rest of the worlds, the total generation capacity of solar panels has been increasing continuously in the last decade. At the moment, there is enough solar electricity capacity deployed able to generate up to 6% of the annual electricity demand of the country. However, solar generation depends not only on the time of the day but also on rapidly varying weather conditions. For example, solar generation can ranges from 0% at night to up to 30% of the total electricity demand of the country during some days of year.



https://www.carbonbrief.org/analysis-uk-renewables-generate-more-electricity-than-fossil-fuels-for-first-time

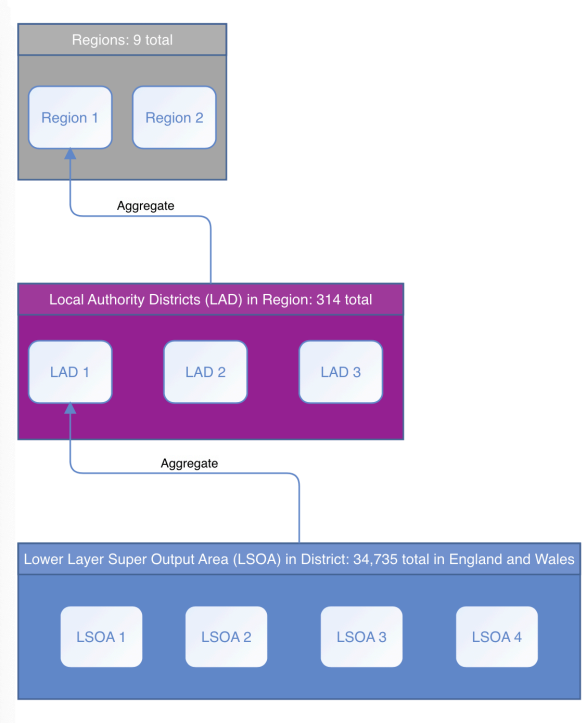


https://ukbusinessenergy.co.uk/uk-solar-pv-market-report/

Since solar generation has a large variability, this makes full use of the resource difficult. A homeowner may not know when to schedule energy intensive usage and an electricity provider may need to keep fossil fuel generators on standby to keep up with lower than expected solar generation. The aim of this project is to generate a highly accurate machine learning model that can use weather parameters such as wind speed, cloud coverage and temperatures as predictors for solar yield (Watts per watt peak) or solar generation (Watts) as a response variable.

Geographic Levels of Details

This project will generate predictions at different levels of spatial granularity in England and Wales. These different levels of spatial granularity are defined by works on statistical human geography. At the lowest level we have a unit called Lower Layer Super Output Area (LSOA) that is defined as an area with a mean population of 1500 people. The next level is called a Local Area District (LAD) that can be up to city size. The final level is the region of which there are nine in the are of England and Wales.

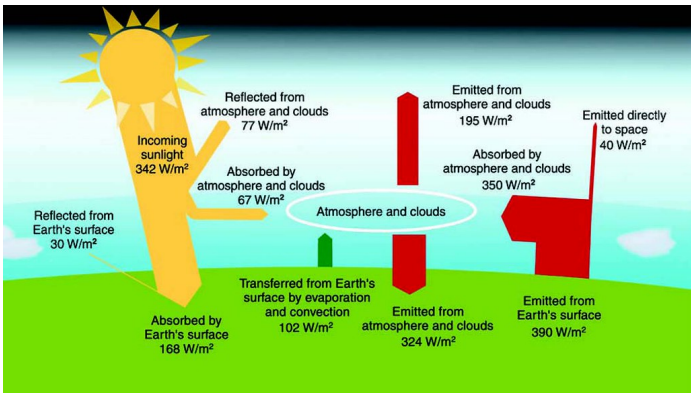


Levels of statistical geography in England and Wales used in this study.

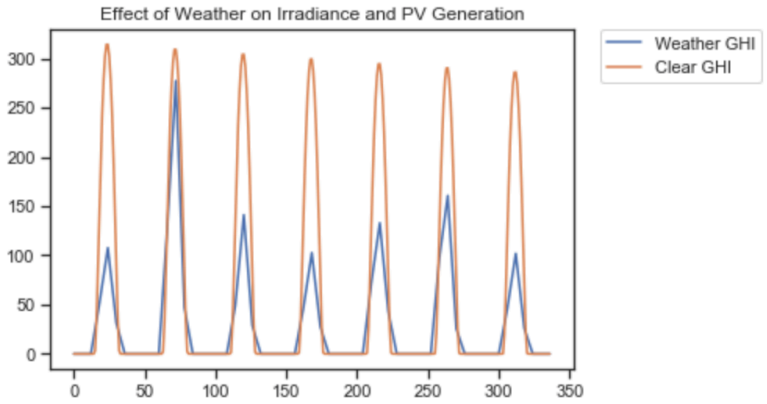
Across England and Wales, every district and region have their own local mix of non and renewable generators. Also, across the national geography there is a lot of variability in renewable resources. Therefore, a single forecast across the country or even for each region does not take into account this variability. This is the reason for the need of a highly granular (at the district level) solar electricity generation forecast.

Effect of Weather on Irradiance and PV Generation

Solar panels convert solar radiation into electricity. This is why there is no generation at night and generation profile follows a curve that peaks at midnight when the amount of sunlight also peaks. The quantity of solar radiation reaching ground level reaches has a maximum under clear sky conditions. Weather conditions such as cloudy skies will always reduce the ground level solar radiation from its peak under under clear sky conditions.



Energy budget of the planet

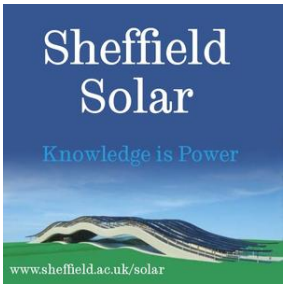


Sunlight reaching ground level under clear sky and under different weather conditions

Data Sources

This study will source data from different sources. Some of the data is statistical and has been gathered across many years (e.g. pv deployment, geographical boundaries) while other sources are collected at a realtime level (e.g. regional generation, weather parameters).

- Office for National Statistics: Population Density, LSOA boundaries, electricity demand, # of meters, District boundaries and centroid.
- Ofgem: Number of deployed photovoltaic systems and DC capacity.
- NOAA: Using the Python PVLIB library hourly weather conditions can be accessed from the Global Weather Model for any latitude/longitude location.
- Sheffield Solar Farm: Hourly generation and yield is sourced from the site at https://www.solar.sheffield.ac.uk/pvlive/



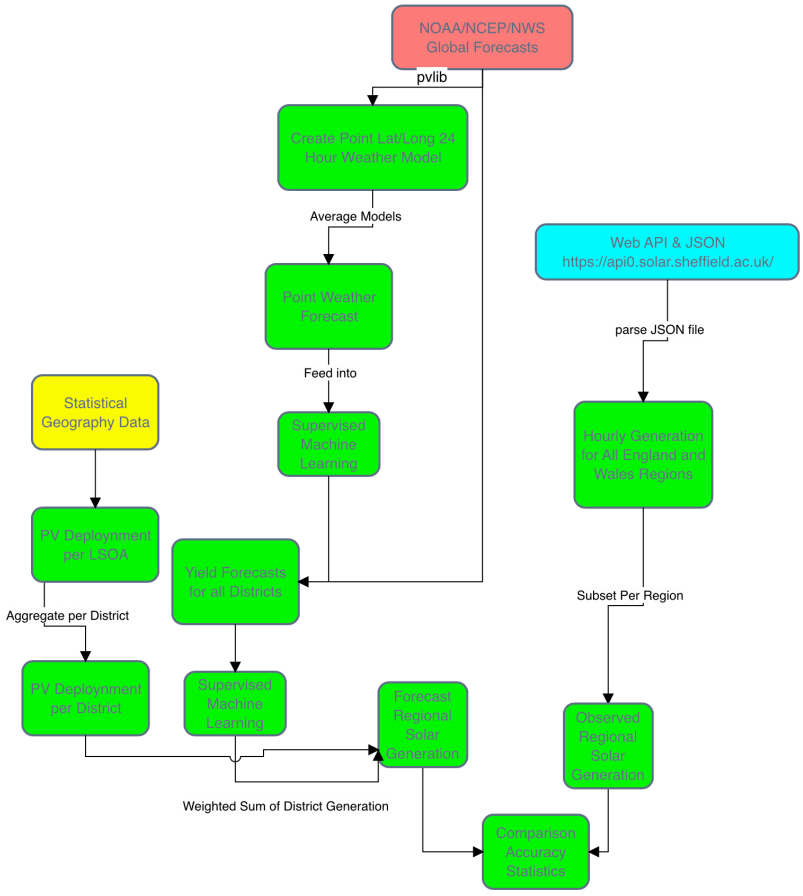
ofgem

Office for National Statistics

Sources of data used in this study

Process Workflow

The workflow diagram below indicates how the different types of data: statistical geography (yellow), Web API (red) and live data (blue) flow together to train a supervised machine learning model and generate the district yield and regional power generation predictions.



Project Workflow

Machine Learning for Electricity Forecasting

Some of the geographic statistical information (e.g. solar deployment and reported power output) at the output level will be aggregated to the local area district level. For each of the 331 local district areas, weather predictions will be collected with a temporal resolution of 30 minutes.

This information will be used to create a data frame containing the following predictors: observation time stamp, district name, region name, latitude, longitude, air temperature, wind speed and cloud coverage at different heights.

These predictors will be used to train for and generate the sub-hourly (30 minute resolution) solar panel yield. Yield tells us to how much energy (Wh) is produced for every Wp of module capacity over the course a time period - in this case a 30 minute period.

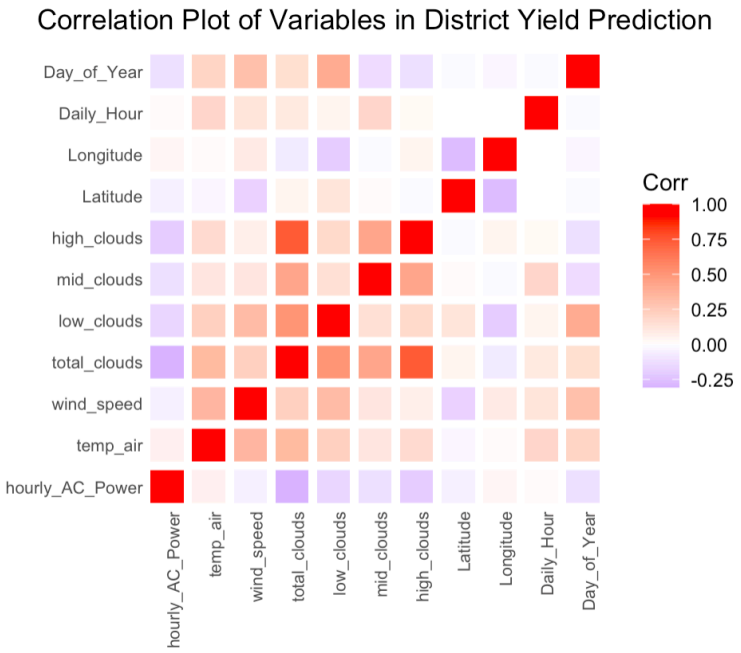
Many linear, non-linear and tree-based machine learning algorithms were tested in this study. However, the 3 best performing per-category were:

Linear Regression:

k - Nearest Neighbors:

Random Forest:

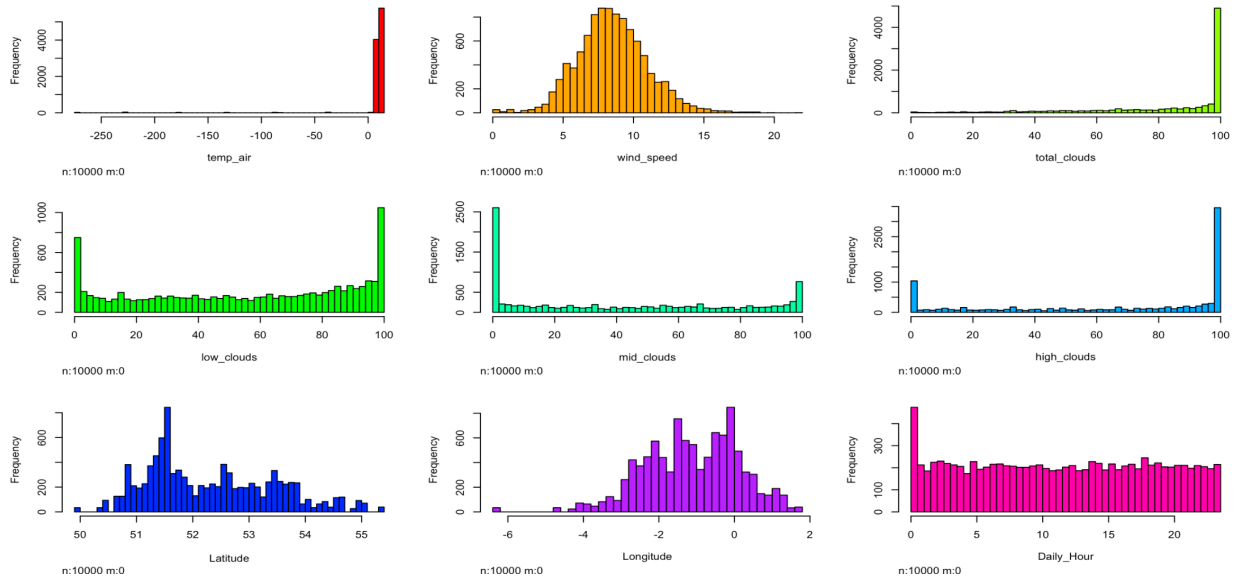
Data Exploration



Correlation plot of the predictors and response variable

Above we can see a plot showing the correlation between the response and predictor variables. The hourly AC yield is negatively correlated to the cloud coverage, day of the year. The more cloud coverage the less sunlight there is for generation. The time of the day correlation is negative in this set of observations because it only covers the month of November when with increasing day of the year, the amount of sunlight per day decreases. Hourly AC yield is also positively correlated to the air temperature. The higher the temperature, the higher the performance of a solar panel.

All of the predictors in the model are numeric. Only wind speed follows a normal distribution while appears that for the rest of the predictors, there are almost equal number of observations across the range of possible values. We should note that in the cloud coverage predictors there



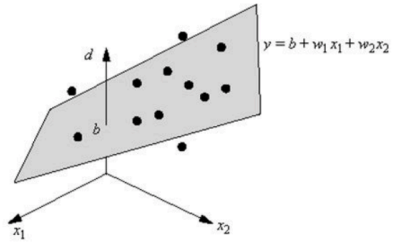
Histograms of predictors in machine learning model

seems to be an abundance of min and max values over those values in between these extreme ranges.

Supervised Machine Learning Algorithms

The three top-performing algorithms per category are the following:

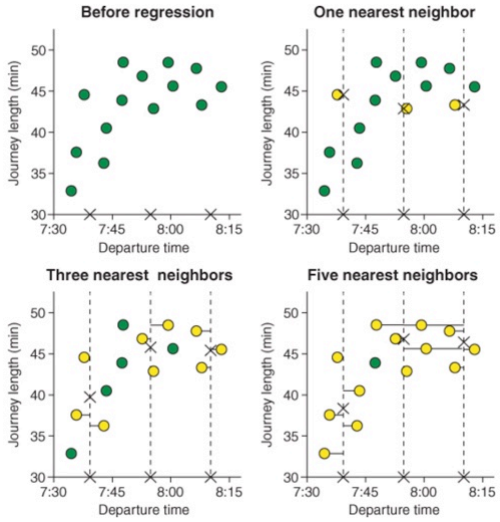
Linear Regression:



Ordinary linear regression equation can be written as
$y_i = b_0 + b_1 x_{i1} + b_2 x_{i2} + \ldots + b_j x_{ij} + e_i$
where yi represents the numeric response for the i_{th} sample, b0 represents the estimated intercept, bj represents the estimated coefficient for the t_{th} predictor, xij represents the value of

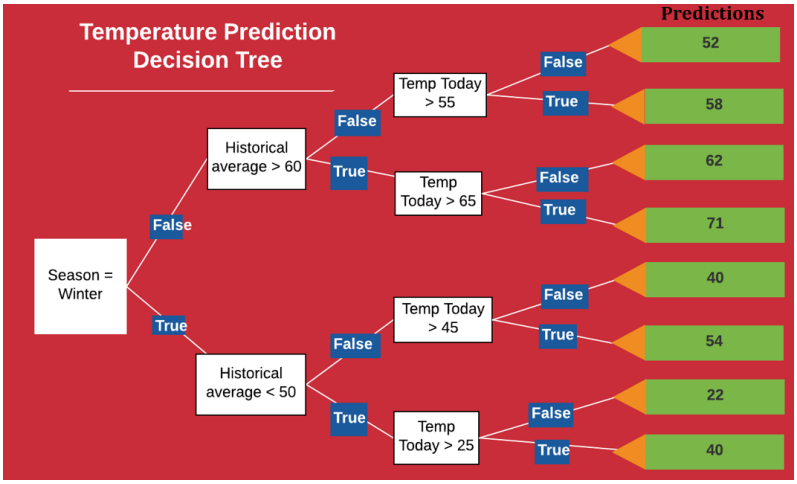
the j_{th} predictor for the i_{th} sample, and ei represents random error that cannot be explained by the model.

k Nearest Neighbor:



It assumes that similar data points are grouped together and it calculates the distance between a 'k' number of points. In the case of regression, the algorithm then calculates the mean of the 'k' neighbors around an observation.

Random Forest:



Random forest combines an ensemble of multiple decision trees. It uses the ‘bagging’ technique in that it combined the results of each decision tree to increase the accuracy of the overall result.

Model Performance

The data set used to train the model had two weeks of data and consisted of more than 100000 observations. Initially, the models were trained on all the available data using the DoParallel library on a desktop computer using 4 cores. However, the training time for the non-linear model was close to half a day while the training time for the tree-based model ran for more than a day without completion. Therefore, sampling of the data set was used to quicken model training completion and test the results without too long of a wait. After training, the models were tested on a complete new data set a week forward from the time period used for training.

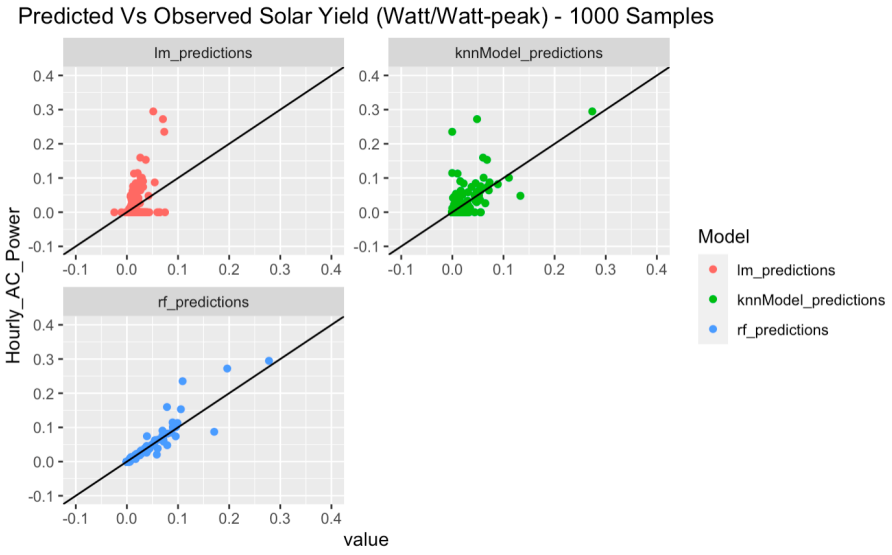
```
Linear Regression training time:
0.667 sec elapsed

knn training time:
1.488 sec elapsed

Random Forest training time:
227.694 sec elapsed
```

Model <chr>	RMSE <dbl>	Rsquare <dbl>
Linear Regression	0.03955913	0.1399861
k-Nearest Neighbors	0.03229901	0.4289623
Random Forest Grid	0.01504436	0.8834463

Training time and Model Accuracy at 1000 Sample Observations



Observed Vs. Predicted Performance at 1000 Sample Observations

The results above are from a sample of 1000 observations or 1/100th of the total. We see that the random forest algorithm is the top performer able to predict the observations with an R_{square} score of 0.88. Looking at the observed vs predicted plot, we see how the linear regression algorithm is not able to predict observations with zero value while also predicting negative values not present or possible in the data.

Below we see the model performance results when we increase the sampling size to 10000 observations or 1/10th of the total. The R_{square} of our top-performer algorithm increases from 0.88 to 0.98. The observed vs predicted plot shows how the random forest algorithm has excellent predictive power along the full range of solar yield values. The k nearest neighbor algorithm is second best at 0.87 R_{square} but it still has problems accurately predicting low Yield values.
We achieve an increase in the predictive power of the algorithms by increasing the sample size. However, we achieved this through a significant increase in training time. The 10 fold increase in sample size from 1000 to 10000 observations increase the training time of the random forest algorithm by 43 times.

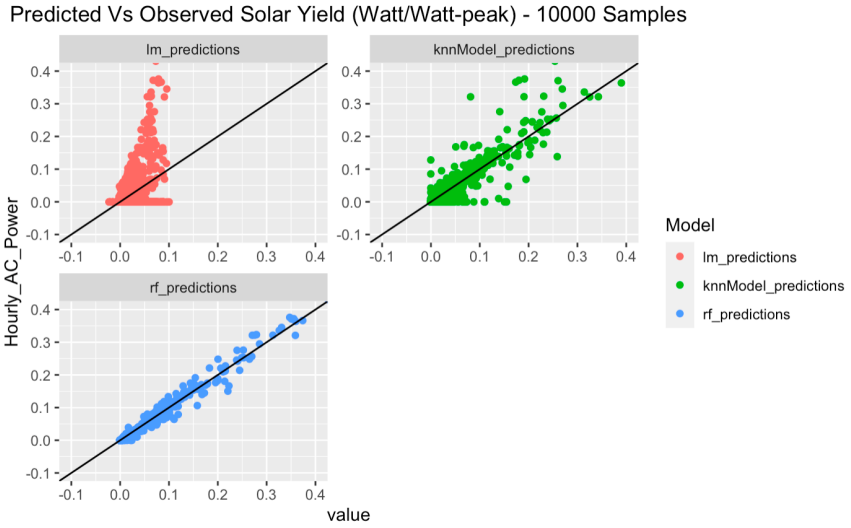
```
Linear Regression training time:
1.17 sec elapsed

knn training time:
24.579 sec elapsed

Random Forest training time:
4027.059 sec elapsed
```

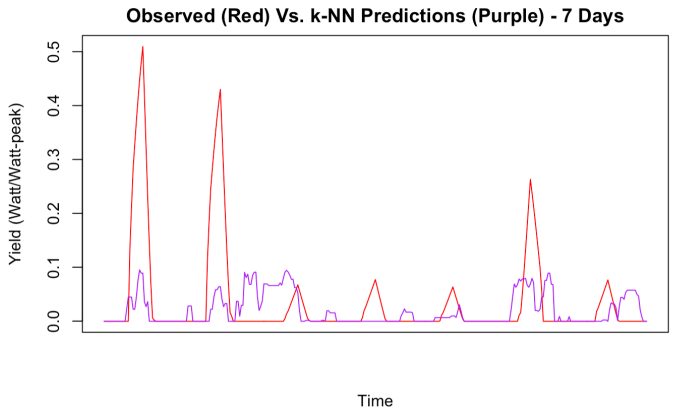
Model <chr>	RMSE <dbl>	Rsquare <dbl>
Linear Regression	0.049338757	0.1460002
k-Nearest Neighbors	0.020040717	0.8617034
Random Forest Grid	0.006896239	0.9843703

Training time and Model Accuracy at 10000 Sample Observations

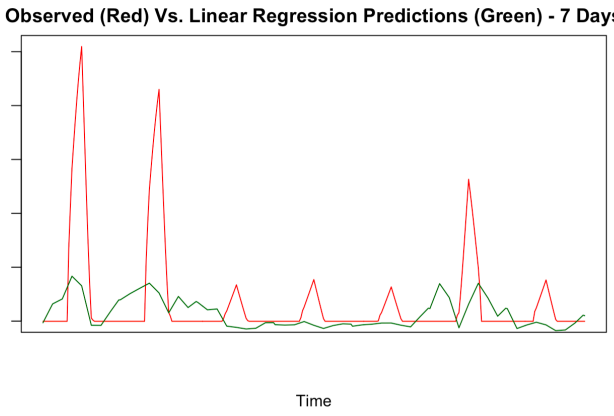


Training time and Model Accuracy at 10000 Sample Observations

The predictions are not independent from each other but are part of a time series. Below we can see how linear regression and k-NN predictions fit a s a time series. Neither the linear regression or k-NN algorithm can generate the time dependent peaks in generation. Linear regression performs worse as it predicts negative values that are not present in the data set or physically possible.

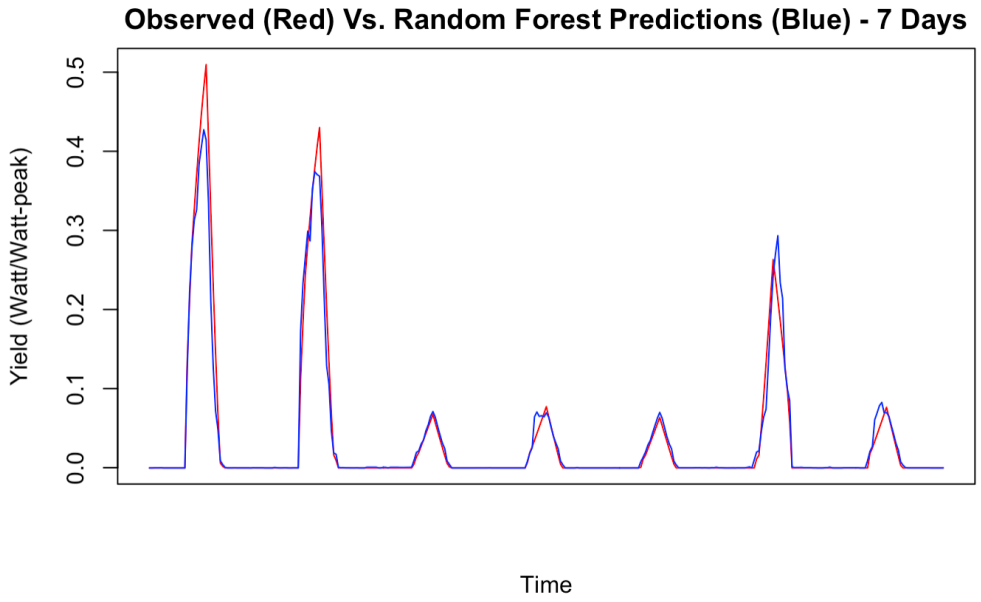


k-NN Time Series Comparison



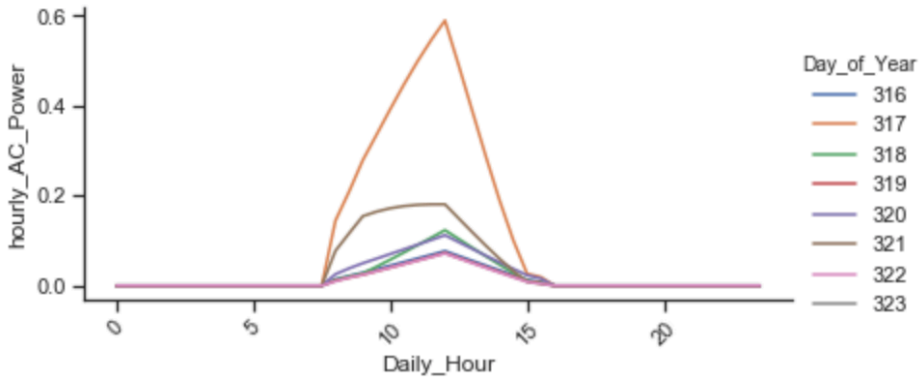
Linear Regression Time Series Comparison

The time series predictions from the best performing algorithm, random forest, are shown below. We can see how the algorithm is able to forecast the day to night seasonality of the time series as well as the changing maximum yields caused by the changing weather conditions.



Random Forest Time Series Comparison

Benefits of Accurate Solar Generation Forecasting

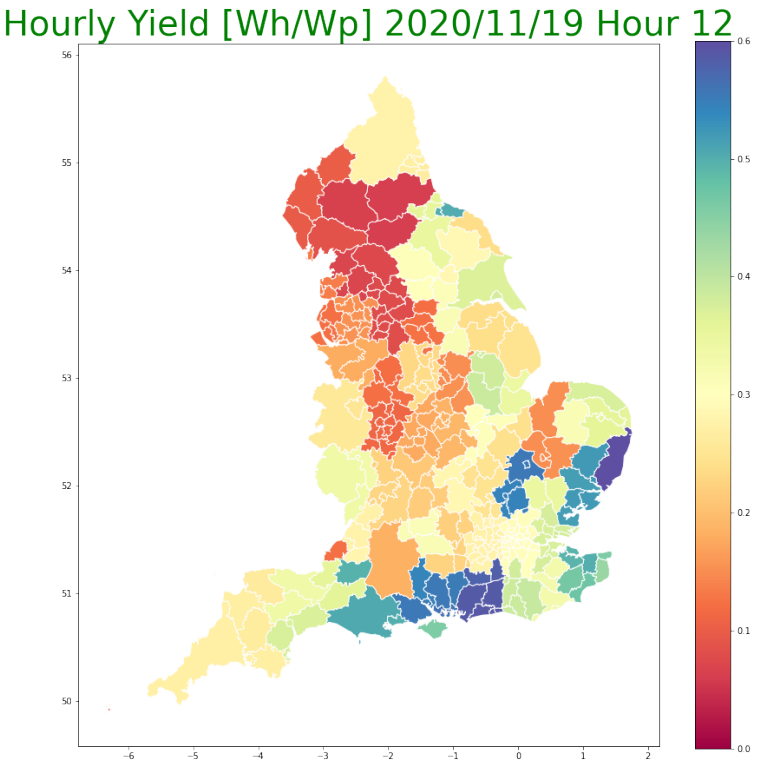


Variability of Solar Yield Across Hours and Days

For the Homeowner: Having accurate solar generation forecast will enable homeowners to schedule energy intensive usage to when the panels are generating the most energy. As we can see above, the peak of generation depends not only on the hour of the day but also can

change across different days. Energy ownership is important for some home generators of solar electricity. Maximum use of electricity generated locally can be achieve thanks to accurate forecasting.

For the Electricity Provider: The local or regional electricity provider wants to know how variable is the generation across its service area. As we can see below, there is a lot of variability across the geographic space of England and Wales. Accurate generation forecast will inform the electricity provider with confidence if it's necessary to keep fossil fuels generators on standby to account for shortage on solar generation during daytime hours.

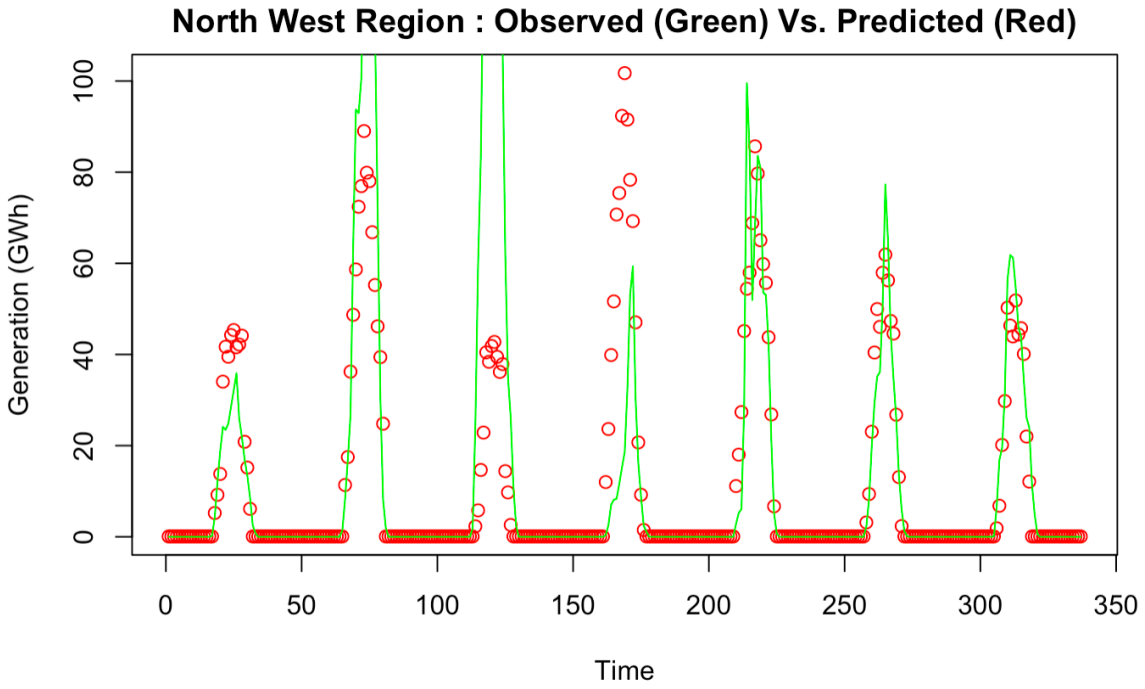


Variation of Solar Yields at Noon Time

Regional Forecasting

We generated a second supervised machine learning model using as predictors all the district yields (Watt/Watt$_{Peak}$)inside a region. The response variable was the reported regional generation (Megawatt hours) sourced from the live feed broadcast by the University of Sheffield PV Live site at https://www.solar.sheffield.ac.uk/pvlive/.

The best performing algorithm was again Random Forest with a calculated R_{Square} performance parameter of 0.9. However, there is a mismatch between the prediction and observed peaks during some days. Other than model limitations, the mismatch could be because the regional areas defined by the electricity providers are slightly different from the regions areas as defined by the UK's Office for National Statistics.



Caption

Final Thoughts

This study has shown the predictive power of supervised machine learning when applied to a solar electricity generation time series. Just by using weather parameters, it has been possible to trained a model able to predict solar panel yield at the district level. Prediction at the regional level shows promise but can be improved if the area match between the regions used for the predictors and response is more accurate.

This results are useful to the homeowners wanting to make better use of the electricity generated locally at their home but also to electricity providers wanted to make decision about the large scale deployment of solar plants. Better use of the resource is a necessity as the percentage of total electricity used is many countries around the world coming from renewable resources such as wind and sunlight keeps increasing.

The model that we shown in this report only used data from part of the year. Even then it showed good predictive power when used from a time period different from the training period.

A more accurate model will need data from at least a whole year to take into account annual seasonality.

This work is a good step forward in making better solar generation predictions and to reduce the carbon footprint of the electricity that we consume.

Things that I learned during this project

R and Python were used in this project. Python was used to wrangle data into a form useful for this study and to source data from the NOAA servers and the PV Live API. Python was also used with the geopandas library to generate the UK choropleth maps. R was used to some data wrangling, model training and performance plots.

Most of the work was spent aggregating, filtering, reshaping and joining the different data sets with different naming conventions and levels of granularity. This project provided a learning experience in how to source data from Python libraries as well as from web API's providing data in JSON files.

Towards the end of the project, a simplified pipeline going from data input to model performance output was created to simplify the time it takes to analyze different data sets. This project also gave me the opportunity to try out new an interesting ways to showcase geographic and time series data.