Forecast of Solar Power Generation

October 14, 2020

Team 35 - Hawraa Salami, Aishwarya Sanjay Bhangale, Arusha Kelkar, Fairy Gandhi

1 Introduction

Why using solar energy? With the increase in awareness of climate change and how it is affecting natural disasters, it is becoming important for many countries, to transition from the coal and oil gas industry to an industry that uses renewable and more sustainable energy sources, such as solar energy. According to the International Atomic Energy Agency, the amount of CO2 emitted by coal (natural gas) energy is 16 (7) times greater than the amount emitted by solar energy. By helping in reducing greenhouse gases around the globe, solar energy is known to have a positive impact on the environment. This is why many countries are increasing their efforts toward integrating and producing more solar power.

Importance of forecasting solar power generation: Unlike conventional sources of energy, there is an uncertainty in the generation of solar power due to its dependency on weather conditions. One of the most effective ways to cope with the uncertainty of solar power production, is to forecast its expected power. The solar forecasts can then be used by many stakeholders including grid operators, electric utilities, energy traders and facility managers. For instance, grid operators can use the forecasts to balance out the supply and demand for electricity, by estimating how much solar energy will feed into the electric grid. Accurate forecasts can then help grid operators avoid any possible costs related to the deviation between the agreed power (to be generated) and the actual one. Moreover facility managers can use the forecasts to schedule maintenance, energy traders can use the forecasts to predict the influence of solar energy on market price (especially during ramping period), and utility companies can use the forecasts to perform power trading and process planning [1,2,3]. For a summary of the main application areas of forecasting, we extracted the following table from this online reference.

Time scale of forecast	Area of application	Stakeholder
Shortest-term (0 – 6 h)	Trading on intraday energy market Control of curtailment due to negative market price Correct activation of regulation power (secondary and tertiary reserve)	Traders
	Influence of vRE on market price	Speculators
	Balancing Unit re-dispatch Curtailment of power plants	Grid operators, load dispatch centers, independent system operators
Short-term (6 – 48 h)	Trading on day-ahead energy market Participation in regulation market Influence of vRE on market price	Traders
	Unit dispatch Load flow calculations DACF congestion forecast	Grid operators, load dispatch centers, independent system operators
	Day-ahead planning of maintenance	vRE operators
Medium-term (2 – 10 days)	Trading on long-term markets	Traders
	2DACF congestions forecast Week-ahead planning	Grid operators, load dispatch centers, independent system operators
	Medium-term planning of maintenance	vRE operators

Project goal: In this project, we aim at exploring various methods for forecasting solar power generation. We focus on short-term forecasting (1 hour or 1 day ahead), using the dataset of aggregated solar power generated in Germany, a country that has been implementing an aggressive policy of energy transition (the Energiewende) with a goal of producing all energy from renewable source [4]. Since solar power generation depends on the solar radiation and sky clarity, we also examine how weather features can help enhancing our short term forecast.

Methods of solar energy forecasting: Multiple approaches have been proposed in the literature to perform short term forecasting of solar energy production. While some approaches are based on time-series analysis with or without exogenous inputs, other methods used machine learning algorithms by focusing on weather features and/or incorporating lagged variables, or used neural networks models (LSTM). Hybrid approaches that combine the outputs from various models were also considered. For a review of all the proposed methods, see [5, 6, 7]. There is no clear agreement in the literature on the existence of one right approach, which really depends on the data source, the considered horizon, and the availability of weather features. In our project, we consider machine learning approaches and time series analysis, compare between their performances and conclude on the best approach for the data we have. By developing the forecasting model, we wish to answer what features help in the short-term forecast of solar power production. Do we only need weather features? Or does incorporating the historical measurements enhance the quality of forecasting? What if we change the horizon period?

2 Datasets: Hourly Solar Power Generated & Weather Data

The dataset used in this project is the time series of the total solar power produced by Germany from January 1, 2005 to May 1, 2019 (with hourly resolution). We downloaded it from the website of open power system data. It consists of the following fields:

- utc_timestamp: start of timeperiod in Coordinated Universal Time
- DE_actual_solar_generation: actual solar generation in MW (mega Watts)
- DE_solar_capacity: electrical capacity of solar in Germany in MW
- DE_solar_profile: share of solar capacity producing in Germany

We also used the scripts provided by the same website to download the time series (with hourly resolution) of the geographically aggregated weather data in Germany. It contains the following fields:

- utc_timestamp: start of time-period in Coordinated Universal Time
- DE_temperature: temperature in Celsius
- DE_radiation_direct_horizontal: direct solar radiation in W/m^2
- DE_radiation_diffuse_horizontal: diffuse solar radiation in W/m^2
- DE_precipitation: total precipitation in mm/hour
- DE_cloud_cover: fraction of cloud cover [0-1] scale
- DE_air_density: air density in Kg/m^3

2.1 Data Cleaning

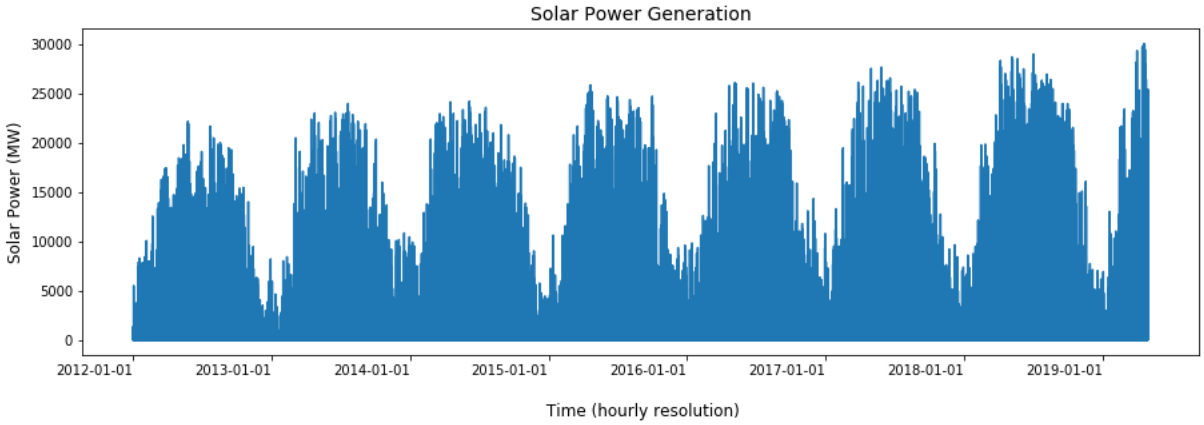
Both datasets were mostly clean and did not require a lot of pre-processing steps. However, we noticed that the values of the solar powers were only reported after 2012 (and till April 2019), the remaining values that correspond to 2005 till 2011 were empty. We extracted the solar data and weather data that correspond to the same time period (from 2012 till April 2019). We also checked if we have all the values of all hours from 2012 till April 2019 in both datasets; we noticed that the values of solar power that correspond to March 31, 2013 and March 29, 2014 are missing. To make sure that both datasets are consistent, we removed from the weather data the rows that correspond to these two missing days.

3 Data Exploratory Analysis

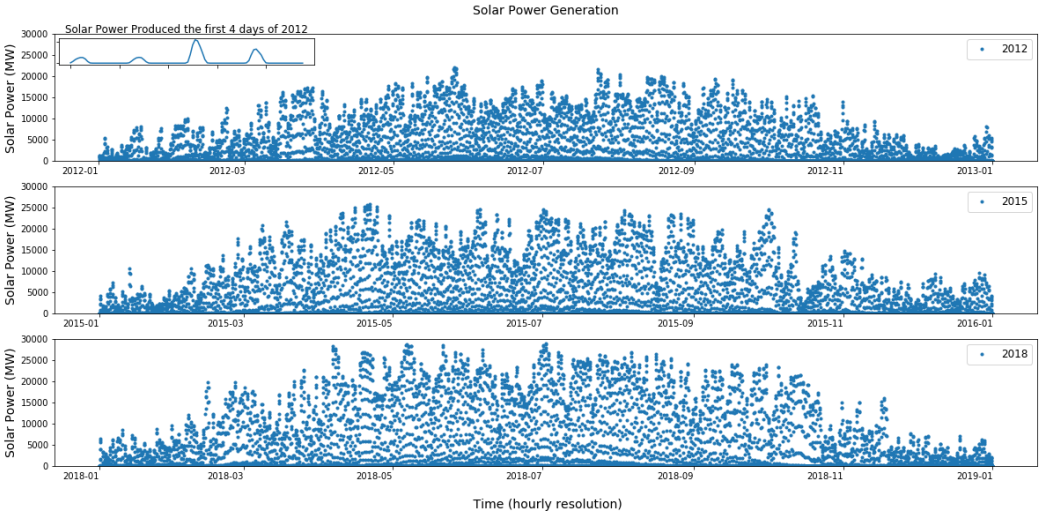
Before we model the generation of solar power, we start with exploring the data. This exploratory analysis allows us to understand the temporal structure of the data and how the generation of solar power varies with time. In particular, we want to look at the trends and seasonal components that exist in our data. How does solar generation change with respect to the hour of the day, month of the year, or in a given season? We address here all of these questions. We also examine how solar power production change with respect to different weather features and how it correlates with its past values.

3.1 Solar Power Generation over the years (2012-2019)

We begin with plotting the time series of the solar power generated in Germany with respect to time.



We clearly observe a yearly pattern in solar power production: each year starts with a low production in solar power, then this production starts to increase to reach its peak at the half of the year (summer months) and then it decreases again to reach its lowest value. This is expected as the solar power production depends on solar energy received, which in its turn depends on the season. We also observe more solar power that is being produced over the years, which translates the increasing effort of Germany to produce more solar power.
Let's select some years (2012, 2015, 2018) and plot for each year its corresponding time series.

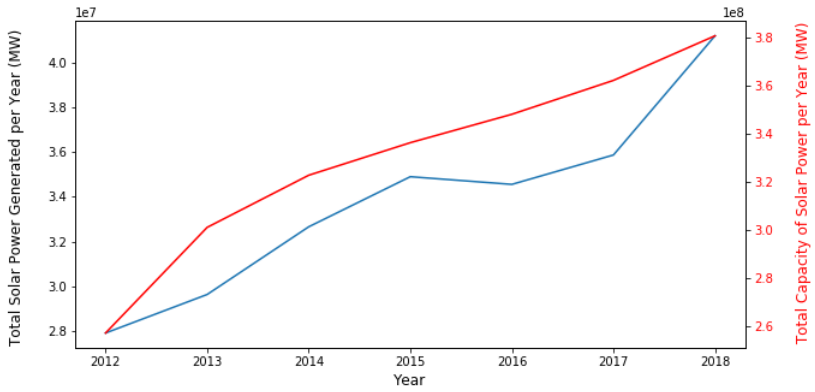


For each year, we observe closely the yearly pattern (dome-shaped); we also notice that, with respect to 2012, higher amounts of solar power were produced in 2018 and these higher values are dispersed over many months. From the zoomed-in plot for the first four days of 2012, we also observe another daily pattern, which we will address in the next sections.

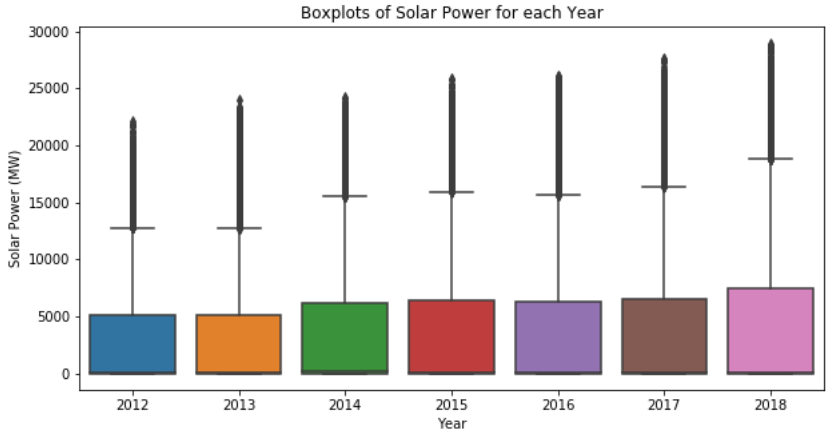
3.2 Increasing Production over Years

We observed form the previous plots, that there is an increase in the total solar power produced over the years. To examine this increasing trend, we plot the total power produced for each year

and on the same plot we also show the total capacity for each year (capacity means the maximum output (generation) of a power plant).



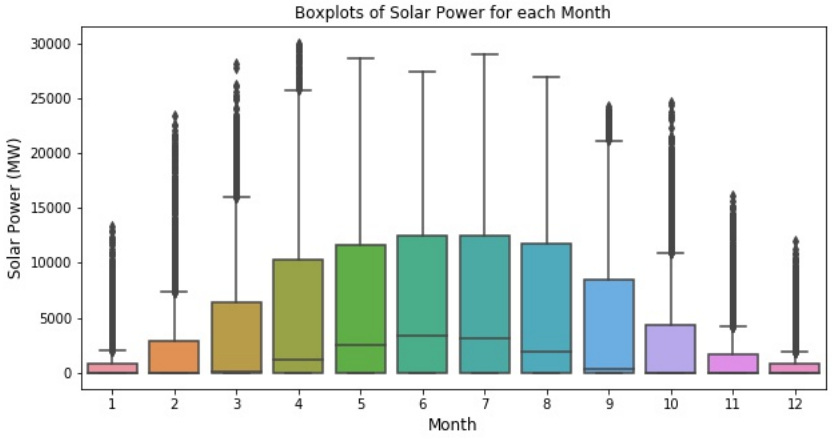
As already observed, the total power produced per year is increasing, however we see a slight decrease in 2016. This overall increase in total power produced is aligned with the increasing trend in the total solar capacity in Germany, which reflects the increasing efforts of Germany to become less independent on conventional sources of energy.
To compare between the distributions of the solar power produced over the years, we plot the boxplots of the values for each year.



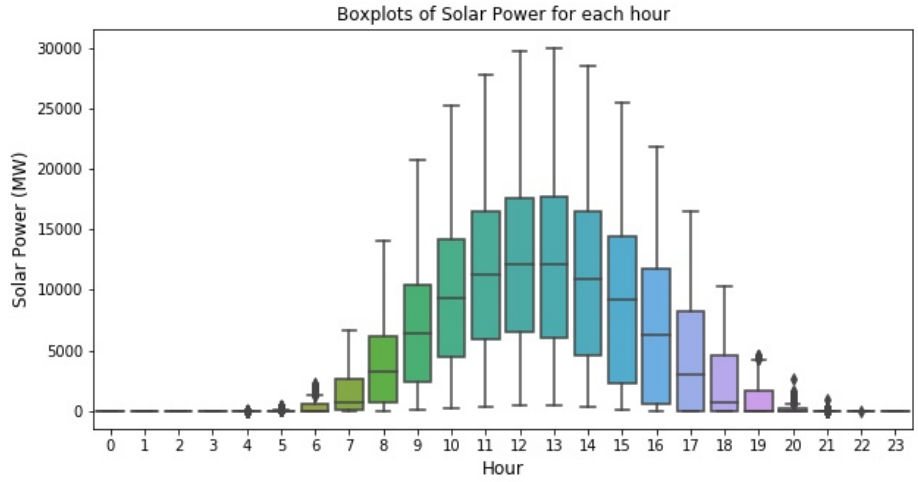
The range of values increased from 2014 to 2013, did not change much in 2014, 2015 and 2016, but it increased again in 2017 and 2018. We next explore the variations in solar power production for each month of the year and each hour of the day.

3.3 Solar Power Generation by Month and Hour

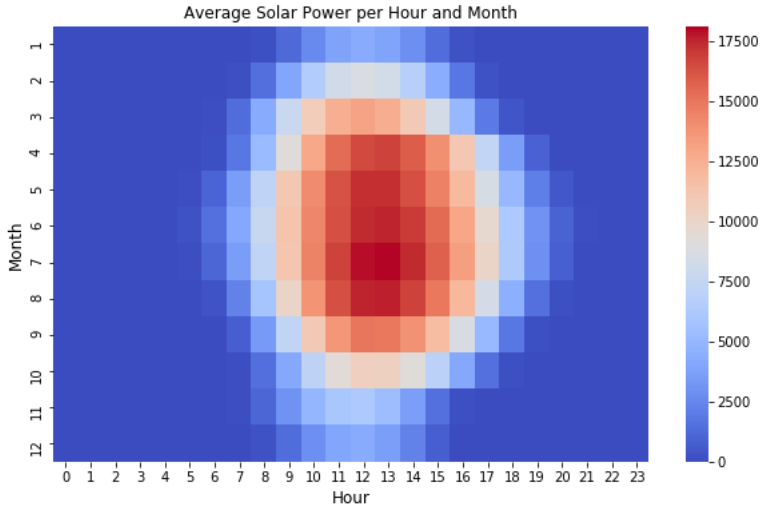
We now examine how the generation of solar power changes with respect to the month of the year. For this sake, we group the values by each month and then plot the corresponding boxplots as shown below:



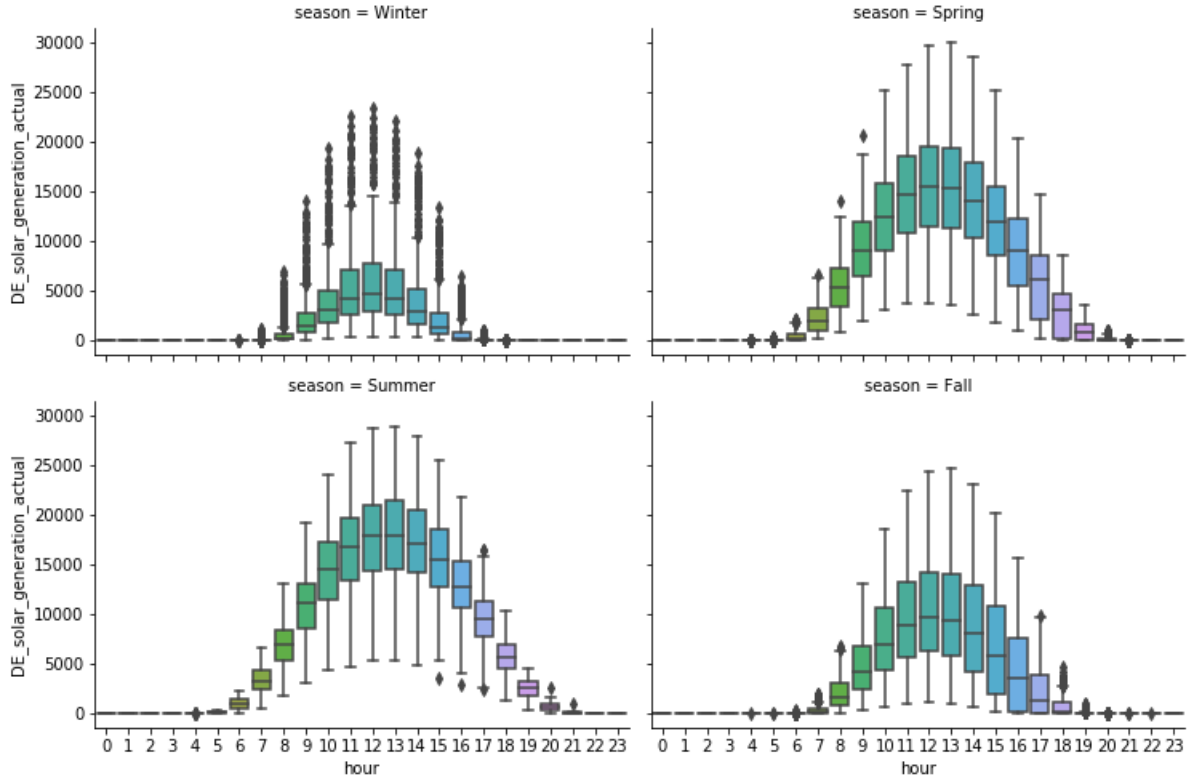
We clearly observe how the range of values continuously increases from January to June and July. Then it starts to decrease again till December. We also observe more outliers for the months of January, February, March, October, November and December (Fall and Winter months), and more high values for the remaining months (Summer and Spring). We now examine how the generation of solar power changes with respect to the hour of the day. We again group the values by each hour and then plot the corresponding boxplots as shown below:



We clearly observe how the range of values starts to increase from early morning to noon (where the highest values of solar power are produced), then it decreases again during afternoon.
Let's now consider both time features: hour and month, and check the average power produced in a given month and hour.



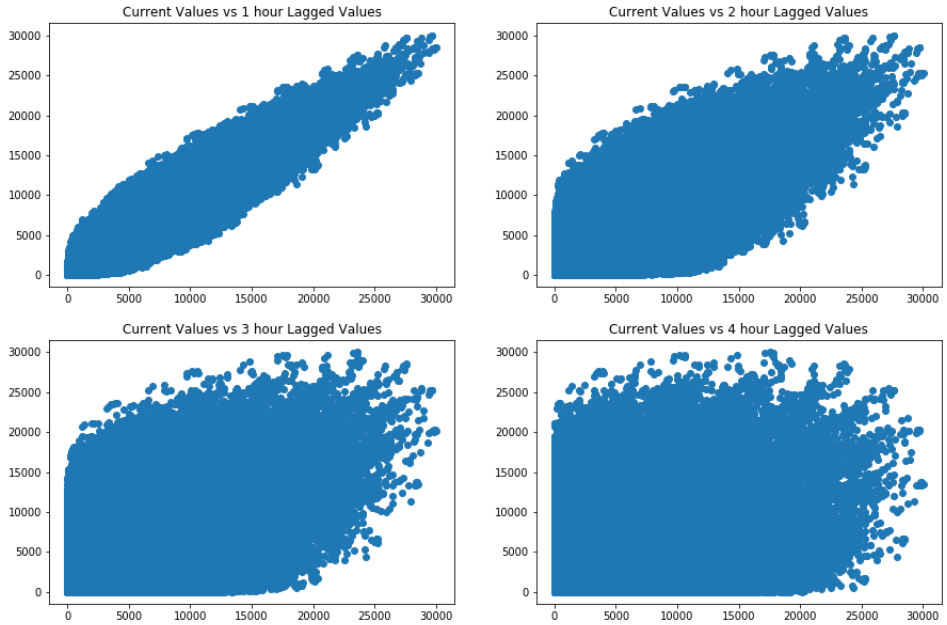
Since the summer days are longer than winter days, solar power production spans over a longer period of time in a summer day than in a winter day. We can also observe the same pattern in the following figures, which show the boxplots of the total power produced for each hour according to the season of the year.



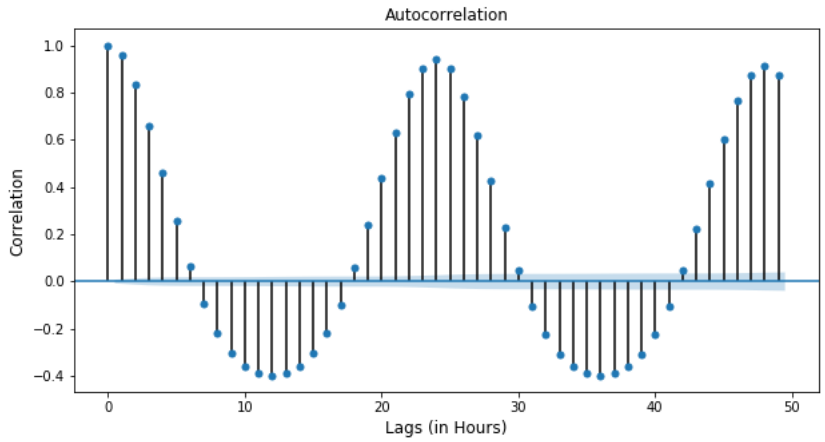
We explored so far the pattern observed in solar power production in terms of hours, months and seasons. We found strong seasonality components in our time series that need to be considered when modeling for solar power generation.

3.4 Autocorrelation of the time series of solar power generation

Another important factor to consider when analyzing a times series, is to understand how a given value at a given time is associated with its past values. This can help us identify which lagged values to consider when doing short-term forecasting.
The figure belows shows the scatter plots between solar power values and some of its lagged values.



We see a strong linear relationship when the lag is 1 hour and this relationship becomes weaker as the lag value starts to increase. We can further check the relationship between the current and lagged values by plotting the autocorrelation function of the time series.

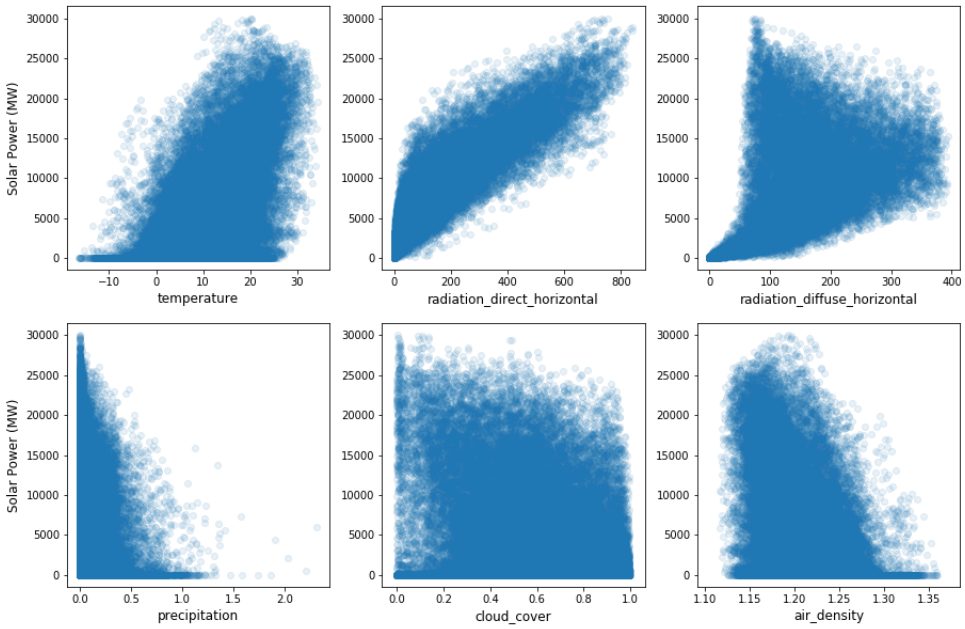


We see a strong statistically significant relationship between a current value and its closed lagged values (1 or 2 hours). This relationship becomes less stronger when we increase the lag in hours (6 to 18 hours), but then it starts to increase again (19 to 24 hours). This re-increase is mainly due

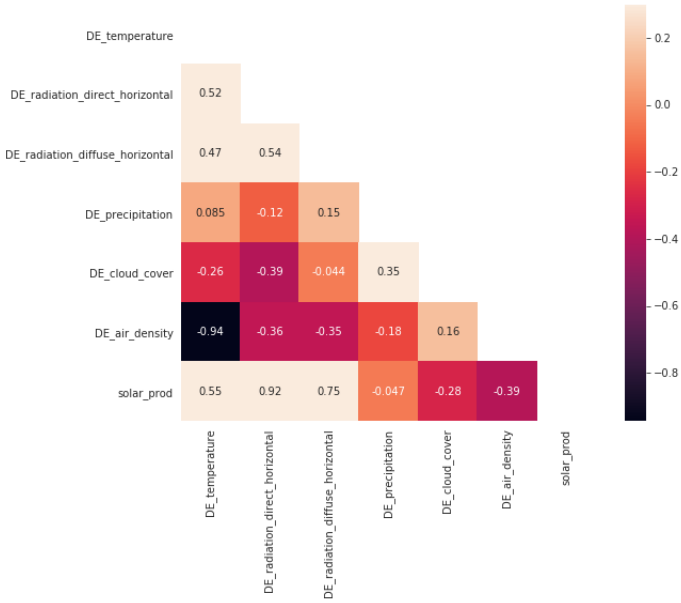
to the strong daily component in our time series.

3.5 Solar Power Production and Weather Features

We now examine the relationship between the total solar power generated and weather features, we check the scatter plots of the solar power produced with respect to each weather feature:

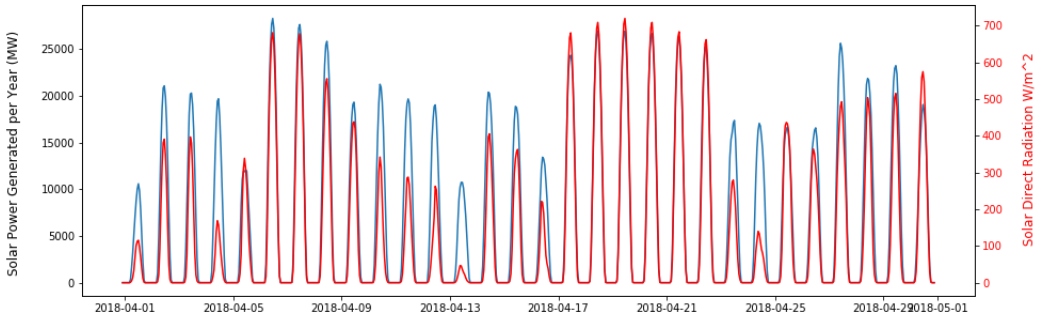


We also check their Pearson's correlation coefficient:



As expected, we notice a very strong positive linear relationship between solar power production and direct solar radiation, as well as diffuse solar radiation. Air temperature also has stronger

relationship with solar production than the remaining weather features. Note also the strong negative relationship between air density and temperature, this is because air density highly depends on temperature (we will only consider in our modeling the temperature feature without air density, because from the scatter plots we see very similar behavior (but in inverse way) between these two weather features and solar power generation). Precipitation has the weakest relationship with solar power generation.
Let's superimpose solar power generated with direct solar radiation and plot as an example the month of April in 2018.



Both time series have same daily seasonality and varies daily in a parallel way. However, if for instance we take two days that have received similar solar radiation, this fact does not guarantee that the same amount of solar power will be produced in both days, which means that we need to take into account for other factors.
We explored so far the temporal structure in our time series data and observed how its generation depends on the hour of the day and month of the year, how it correlates with its past values and weather features.

4 Short-Term Forecast of Solar Power Generation - 1 Hour Ahead

After having explored our data, we now focus on building a predictive model for solar power generation. In this section, we address the following question: given the historical data of power generation, can we predict the total solar power that will be generated for the next hour? To answer this question, we focus on machine learning approaches; we first only consider weather features and examine the performance of different machine learning models, we then incorporate some lagged values in our models and check their performance and we finally remove weather features to see how the models can perform without weather features. What we wish to answer is what features and model can help in predicting future solar power generation.

Chosen Metrics and Method used for Model Comparison (Nested Cross Validation) To compare between the models' performances, we use as error metrics: root mean square (RMSE) and R-squared; the root mean square measures how large our errors are and it penalizes large errors, R-squared gives us an idea of how well our model fit the data (in terms of variance reduction). To perform model selection and validation, we perform nested-cross validation:

- split the data into training and testing sets; we use the testing set to evaluate the final model (we reserved the last year of data as testing set);
- the training set is used to train and validate our models: we divide it into sub-training and

validation sets (we also took the last year from training set as validation); we then train and test for our models in four iterations:

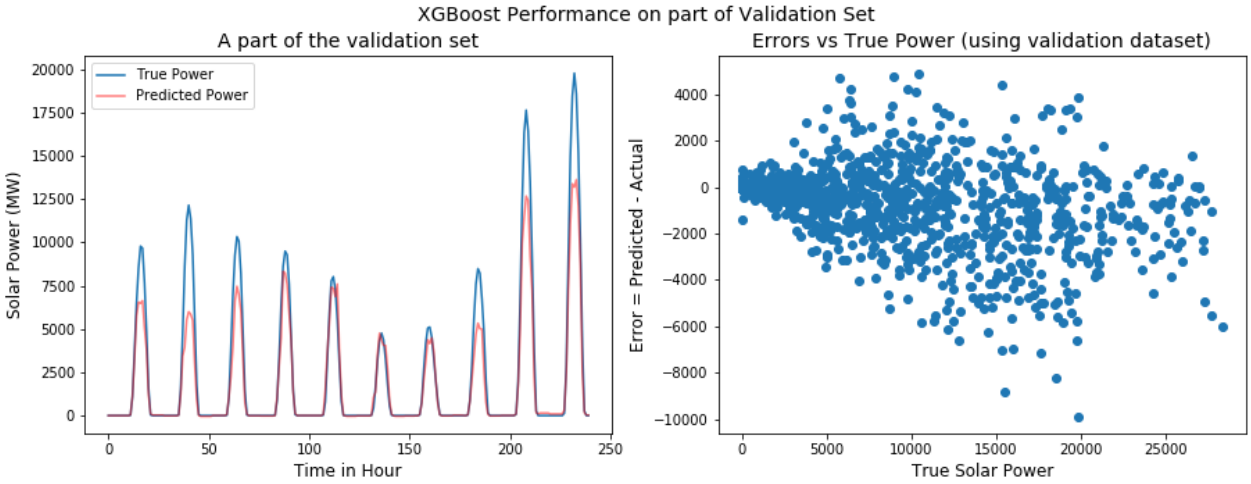
1. first iteration: train on the sub-training data and test on the first quarter of the validation set;
2. second iteration: expand the sub-training data to include the first quarter of the validation set, and then test on the second quarter of the validation set;
3. subsequent iterations: expand the sub-training data to include the previous quarter of the validation set, and then test on the next quarter of the validation set. When we reach the last iteration, we compute the average performance of all models, then we choose the best model. Once the final model is selected, we train it using the whole training set and then test it on the testing set. Also to provide a lower bound for the performance of the proposed models, we compute the performance of "persistence algorithm", which predicts the value of the next hour using the value of the current hour.

4.1 Forecast with Only Weather Features

We first start with exploring how our given weather features can help in forecasting the 1 hour ahead of solar power generation, without including any lagged values. In other words, we address the following question, if at a given hour t, we know the following weather features: solar radiation (direct and diffuse), temperature, cloud cover and precipitation, can we estimate the total solar power that will be produced at t? To enrich our model, we add the following time features: hour, year, month and day.
We test the following models: linear regression, lasso, support vector machine (SVR), random forest and extreme gradient boosting (XGBoost) models. Note that we treated the time features: hour, day, year and month as ordered categorical variables, and we standard-scaled the remaining numerical features. We obtain the following results (using nested cross validation):

Performance (Nested CV)	Persistence Algorithm	Linear Reg.	Lasso	SVR	Rand. Forest	XGBoost
Average RMSE	1797.77	1480.71	1481.10	1796.27	1156.18	1079
Average R-squared	0.879	0.886	0.886	0.881	0.940	0.94

We see that the random forest and xgboost performed the best between all the models, and had similar performances in terms of average R-squared. Let's check a zoomed-in part of the last validation set on which we trained XGBoost.



We notice that the model predicts low power values better than predicting higher values. For higher values, we see that the model in general underestimates them. We can also check this underestimation for high values from the right plot, where most of the obtained errors are negative for high solar power. Let us also check its features importance:

Features	Importance
DE_radiation_direct_horizontal	0.8275
DE_radiation_diffuse_horizontal	0.0980
year	0.0197
hour	0.0127
DE_temperature	0.0108
DE_cloud_cover	0.0105
DE_precipitation	0.0094
month	0.0076
day	0.0037

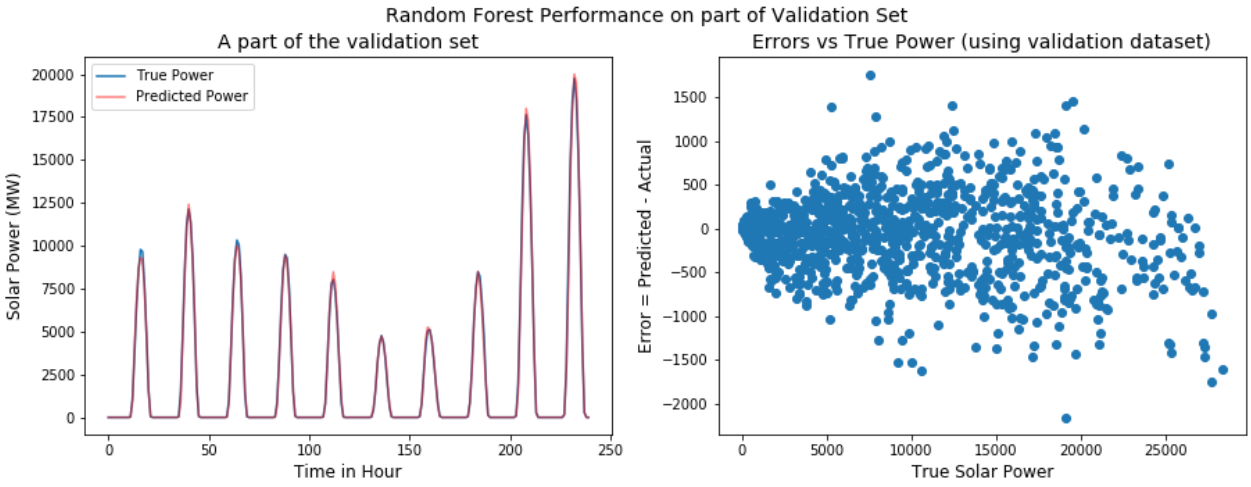
As expected, solar radiation features have the highest importances.

4.2 Forecast with Weather Features and two Lagged Values (t-1, t-2)

We now explore enriching our models with more features: lagged values. We already observed how a value at time t highly correlates with some of its past values. We again use the same features of the previous model and add to them two lagged values. We now focus on the two lagged values (at t-1 and t-2 because they showed the strongest correlation with solar power generation), we will mention in a later section if the addition of more lagged values can help or not.
We again test the same models and obtain the following results (using nested cross validation):

Performance (Nested CV)	Persistence Algorithm	Linear Reg.	Lasso	SVR	Rand. Forest	XGBoost
Average RMSE	1797.77	581.26	581.71	1199.63	297.88	300.94
Average R-squared	0.879	0.979	0.979	0.945	0.996	0.994

The performance of all models improved and random forest is the winning model in terms of both metrics. Let's check a zoomed-in part of the last validation set on which we trained the random forest model.



We notice that the model now is doing a better job in predicting higher values, hence the decrease in RMSE. We can also check it from the right plot, where most of the obtained errors are randomly spread between negative and positive values. Let us check how it weighted the importance for each feature:

Features	Importance
lag_1	0.8446
DE_radiation_direct_horizontal	0.0222
DE_radiation_diffuse_horizontal	0.0868
hour	0.0414
lag_2	0.0027
DE_precipitation	0.0006
DE_temperature	0.0005
month	0.0003
DE_cloud_cover	0.00023
day	0.0002
year	0.0001

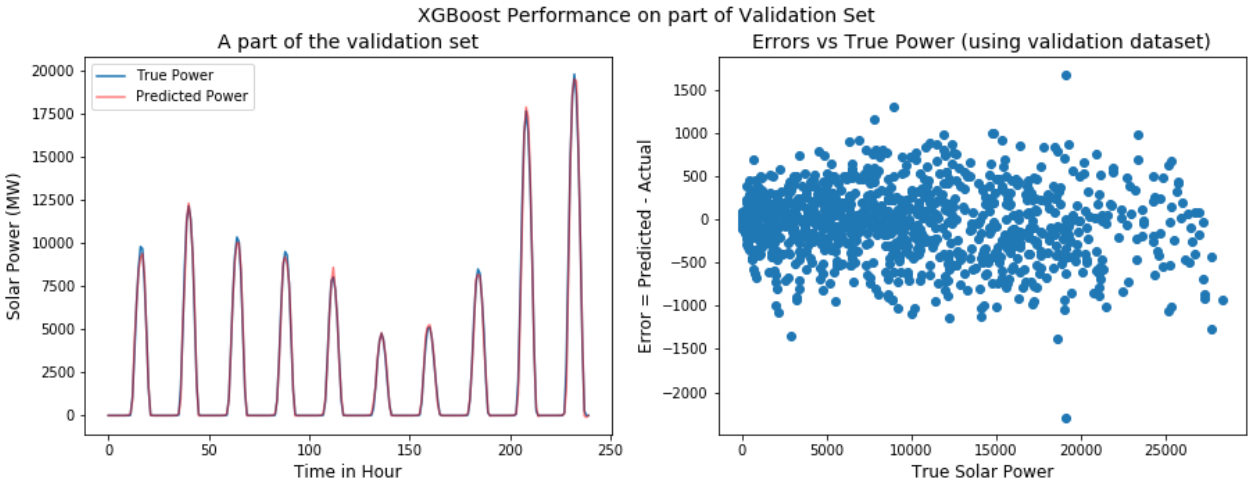
We clearly observe how the previous lagged values are now the most important feature.

4.3 What if we remove Weather Features?

We want now to check if we don't have weather features, will time features and lagged values still able to forecast the next hour solar power production? The reason we're considering removing the weather features is because in real life, we do not have the real weather features, we have instead their forecasts.
We remove the weather features, keep the time features and test our models. We obtain the following results:

Performance (Nested CV)	Linear Regression	Lasso	SVR	Random Forest	Xgboost
Average RMSE	679.45	679.42	1533.98	295.98	271.51
Average R squared	0.968	0.968	0.915	0.996	0.996

Surprisingly, we see how random Forest and XGBoost have very similar performance (even slightly better here) to when we had weather features. This might suggest that for short term forecasting (1 hour ahead), the lagged values as well as the time features contain enough information to predict for the next hour, so that we don't rely on weather features. Let's check a snapshot of the performance of XGBoost on a part of the validation set.



We again observe very similar performance to when we used "weather features + Lagged Values". Let's check the features' importances:

Features	Importance
lag_1	0.8429
hour	0.1475
lag_2	0.0054
month	0.0034
day	0.0004
year	0.0004

Note that we also tried including additional lagged values, but the performance of all algorithm did not get better and we concluded that the last two lagged values are enough.

4.4 Final model Performance

The final model we chose is XGBoost with two lagged values and time features without weather features. We train it on the whole training set and test it on the testing set. We obtain the following results: RMSE: 321.79, and R2:0.998. On the other hand, the performance of persistence algorithm is: RMSE: 2170 and R2:0.912.

Features (Final Model)	Importance
lag_1	0.8444

Features (Final Model)	Importance
hour	0.1459
lag_2	0.0059
month	0.0033
day	0.0004
year	0.0003

We conclude how for 1-hour ahead short term forecasting, an XGBoost with lagged values and time features can be used to perform 1-hour short term forecasting without the need for weather features.

5 Multi-Step Solar Power Forecast - 24 hours ahead

The algorithms we proposed so far predict for one hour ahead. What if we want to predict for some more future hours? If we are using machine learning approaches that incorporate lagged values, we need to take into account that to predict for any of the hours ahead, these previous lagged values might not be available. One way is to sequentially predict for each hour of the 24 hours span, i.e. we start with first predicting the first hour, then the second hour and so on till we reach the last hour. For any hour t, if we want to use a past value within the last 24 hours (between t-23 and t-1) , we assume it is not available and use its predicted value from the past forecasting step. Another way to is to build a different model for each hour. In this project due to time limitation, we only focus on the first approach, which is using the predicted value of a past hour as our lagged values for our next hour forecasts.

5.1 Random Forest: with vs without Predicted Lagged Values

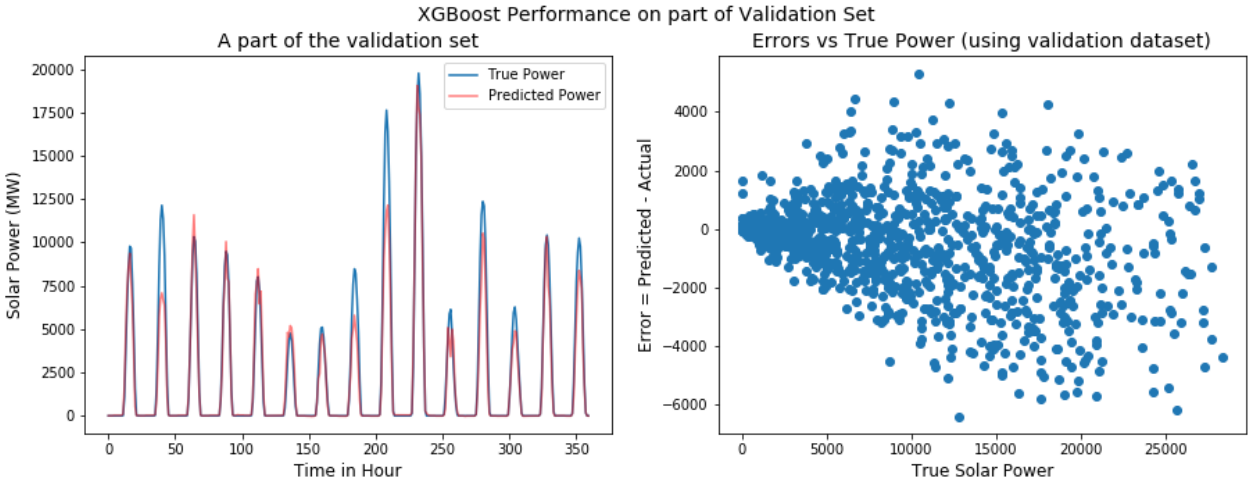
To predict for the next 24 hours, we try the following models: 1) Random Forest (RF) with two lagged values (t-1, t-2), predicted when not available, as we assumed that the actual values are not available, 2) XGBoost with only weather features, 3) XGBoost with weather features + two lagged values (t-23, t-24), predicted when not available. We compare their performance with a persistence algorithm that uses values from the past 24 hours to predict the next 24 values. We again use nested cross-validation and we obtain the following results.

Performance (Nested CV)	Persistence Algorithm	RF - Two Lagged Values (t-1, t-2)	XGB - Weather	XGB - Weather + Lagged Values (t-23, t-24)
Average RMSE	2042	2765	1079	1046
Average R squared	0.827	0.526	0.940	0.945

We notice how when increasing the time horizon, we need to rely on weather features again. Using the two predicted lagged values showed a bad performance; this is because predicted lagged values might include errors that start to propagate when used to predict the solar power production for the next hours. We also observe that when using XGB with weather features, enriching this model with the past values (at t-23 and t-24) slightly enhanced its performance (in terms of RMSE and R-squared). Let's check its features' importance:

Features	Importance
lag_24	0.7982
DE_radiation_direct_horizontal	0.150
DE_radiation_diffuse_horizontal	0.0155
year	0.0081
DE_precipitation	0.0052
hour	0.0051
DE_temperature	0.0046
DE_cloud_cover	0.0046
month	0.0034
lag_23	0.0025
day	0.0022

Let's check a snapshot of its performance on some part of the validation set.



We see that the model is not always performing well in estimating the high values of solar power, as it underestimates many of them. The 24-hour multi-step forecast might need some additional features that can help with enhancing the overall performance of the model.

5.2 Note regarding traditional Time Series Models

We also considered traditional time series models, in particular we tried SARIMA from the library statsmodels. However, since we have long historical data (at hourly resolution), training such models was taking a lot of time. To try this approach, we trained some SARIMA models on a portion of our sub-training data and validate them using the walk forward testing. For a 24 hour forecast, we used a SARIMA model (order (p,d,q)=(2,1,0), seasonal_order (P,D,Q,S)=(1,0,1,24)), we obtained an RMSE of 1351 on the portion of validation set that we used. To use this approach, we need to perform more fine-tuning or reformulate the problem differently, since having this long historical data might not be suitable for the algorithm used for SARIMA in statsmodels library.

6 Conclusion and Future Improvements

In this project, we considered machine learning approaches to perform short term forecasting (1-hour and 24-hour ahead). We noticed that for 1-hour ahead forecasting, time features and lagged values can be used to perform 1-hour ahead forecasting without the need of weather features. However, this is not the case when we wanted to perform multi-step forecasting, where we needed to incorporate weather features.
For future works, we would like to investigate the following:

- enhance the multi-step forecasting model by examining what additional features could be added or by proposing different models for each hour; consider 6-hour ahead forecasting as well;
- try LSTM (a recursive neural network model) or a hybrid modeling approach;
- explore the solar power generation for each station in Germany and examine if our findings are still valid when we consider solar power generated at each station;
- use the forecast values of the weather features instead of the actual weather features;
- explore how the predicted values can be used in price or load forecast.

7 References

1. https://energypedia.info/images/2/2a/Discussion_Series_06_Technology_web.pdf
2. https://publications.jrc.ec.europa.eu/repository/bitstream/JRC106897/emhirespv_gonzalezaparicioetal2017_newtemplate_corrected_last.pdf
3. https://www.cleanenergywire.org/factsheets/volatile-predictable-forecasting-renewable-power
4. https://www.agora-energiewende.de/fileadmin2/Projekte/2019/Jahresauswertung_2019/A-EW_German-Power-Market-2019_Summary_EN.pdf
5. https://www.math.kth.se/matstat/seminarier/reports/M-exjobb18/180601f.pdf
6. http://coimbra.ucsd.edu/publications/papers/2013_Inman_Pedro_Coimbra.pdf
7. https://www.sciencedirect.com/science/article/abs/pii/S0038092X1630250X