Multi-Plant Photovoltaic Energy Forecasting Challenge with Regression Tree Ensembles and Hourly Average Forecasts

Kathrin Bujna[1] and Martin Wistuba[2]

[1] Paderborn University
[2] IBM Research Ireland

Abstract. In this paper, we present the winning solution to the ECML-PKDD Discovery Challenge 2017 on Multi-Plant Photovoltaic (PV) Energy Forecasting. The goal of the challenge is to utilize the historic data of three different PV plants in Italy regarding meteorological conditions and production in order to forecast their energy production. A major problem is that the data contains many missing value for most sensors and especially for the period of the year for which predictions shall be made in the subsequent year. We investigate two approaches: a regression tree ensemble whose hyperparameters where tuned via Bayesian optimization and a simple rule that just predicts the hourly averages that were observed during the previous year. The latter approach is the winning solution of the challenge.

1 Introduction

Due to the urgent need to reduce pollution emission, clean and renewable energy sources have become a strategic European Union and international sector. In particular, photovoltaic (PV) power plants, which are already widely distributed in Europe, are becoming a major source of renewable energy. The main challenges faced by the renewable energy market are grid integration, load balancing and energy trading. In order to face these challenges, we need reliable tools for the prediction of the energy production.

Challenge. In the ECML PKDD 2017 Multi-Plant Photovoltaic Energy Forecasting Challenge, we are given multiple time series regarding meteorological conditions and production collected by sensors on three closely located PV plants in Italy. Given the data for the period from January to December 2012, the goal is to forecast the energy production of each plant for the period from January to March 2013.

Related Work. [1] assess several approaches for PV power forecasting. In contrast to the given challenge, the authors consider a larger number of features, for instance, the weather forecast provided by numerical weather prediction systems and the geographic coordinates of the plants. Moreover, they consider the data collected from 18 power plants, instead of the data from just 3 plants.

2 Task

For detailed information on the given data, we refer to [1].

Time Series. The given data set consists of time series data regarding weather conditions and production collected by sensors on three closely located PV plants in Italy. The elements of these time series are hourly aggregations obtained as the average of all the measures available during a specific hour at a specific day of the year. That is, for each plant, day, and sensor, we are given a time series of 19 values representing hourly aggregated observations (plants are active from 2am to 8pm).

The data was gathered locally, from sensors available on plants and from the closest meteorological stations. There are three time series that describe the plant: irradiance, temperature, and the power production. There are seven time series that describe the weather: temperature, cloudcover, dewpoint, humidity, pressure, windbearing, and windspeed.

Train and Test. The given training data spans over a period of 12 months (year 2012) and includes the target time series, i.e. the power production observed for each plant. The testing data consists of 3 months (January to March 2013) for which the target time series is not provided. According to the challenge website, min-max scaling (between 0 and 1) on power variables needs to be performed on the training and testing data before applying the prediction model.

Additional Information. According to the challenge website, the data can contain reasonable missing values or outliers. Missing values are indicated by *zero* measurements. Besides that, the PV plants have a maximal power rate of 1 000 kW/h.

3 Exploration

Some of the power measurements are much higher than the maximal power rate of 1 000 kW/h (e.g. 885 124 kW/h). Therefore, in the following figures, we clean the power measurements by setting all power values larger than 1 000 kW/h to exactly 1 000 kW/h.

3.1 Single Time Series vs. Power Production Time Series

In Fig. 1 we consider the impact of the single variables with respect to the power production and, additionally, the average monthly and hourly power production. The results confirm our expectation: First and foremost, we expect that the power production is higher during summer, during the middle of a day, and while the plant irradiance is higher. We expect that, if the weather is sunny, then the plant irradiance is higher. We expect sunny weather if the weather

temperature is higher, the pressure is rather high, there are few clouds, and the humidity is low.

Most notably, there is a strong correlation between the plant irradiance and the power production. The correlation of the power production with the weather temperature, pressure, cloud cover, and humidity is less striking. Besides that, in Fig. 2, we see that these variables also correlate with the plant irradiance.

The dewpoint, the wind speed, and wind bearing do not seem to be correlated to the power production, respectively. One might suspect that the informativeness of the wind speed and bearing might depend on the location. However, except for the wind bearing measured at the plant with index 1 (see Fig. 1), these location dependent figures are similar to the averages (over all plants) depicted in Fig. 1.

Last but not least, apparently, there is a significant number of missing values, which have been set to zero by default (see Sec. 2).

3.2 Daily Measurements of the Plant Sensors

We considered the plant irradiance and the plant temperature the most promising variables. Fig. 3 depicts the daily averages values of these variables during the year 2012.

For plant 1, a lot of the plant irradiance and plant temperature measurements are probably missing between day 30 and day 58. During this interval, both measurements are constantly 0, respectively. Besides that, we observe that the plant power of plant 1 drops significantly during the very same time. That is, in comparison to the other plants, we would expect roughly 150 kW/h, but observe about 100 kW/h.

Regarding plant 2, there seem to be even more missing values in the first 90 days of the year 2012. During the first 90 days, the plant irradiance and plant temperature measurements are constantly 0. Besides that, the plant power measurements behave as expected. Still, there seems to be a slight drop of the power production between day 25 and 50.

Finally, consider the measurements regarding plant 3. Here, there seem to be no missing measurements. However, we observe a strange behavior of the plant irradiance measurements: Approximately between day 50 and day 150, the plant irradiance increases much faster compared to the irradiance measurements for plant 1 and 2, respectively. At day 150, the measurements decrease to values which are similar to the values measured for plant 1 and 2, respectively. Again, there seems to be a slight drop of the power production between day 25 and 50.

To sum up, the time series of all three plants exhibit an unusual behavior in the first 90 days of the year 2012. Unfortunately, the test data consists of measurements for the first 90 days of the subsequent year. On a side note, the *test* data does *not* seem to contain missing plant irradiance or plant temperature measurements.

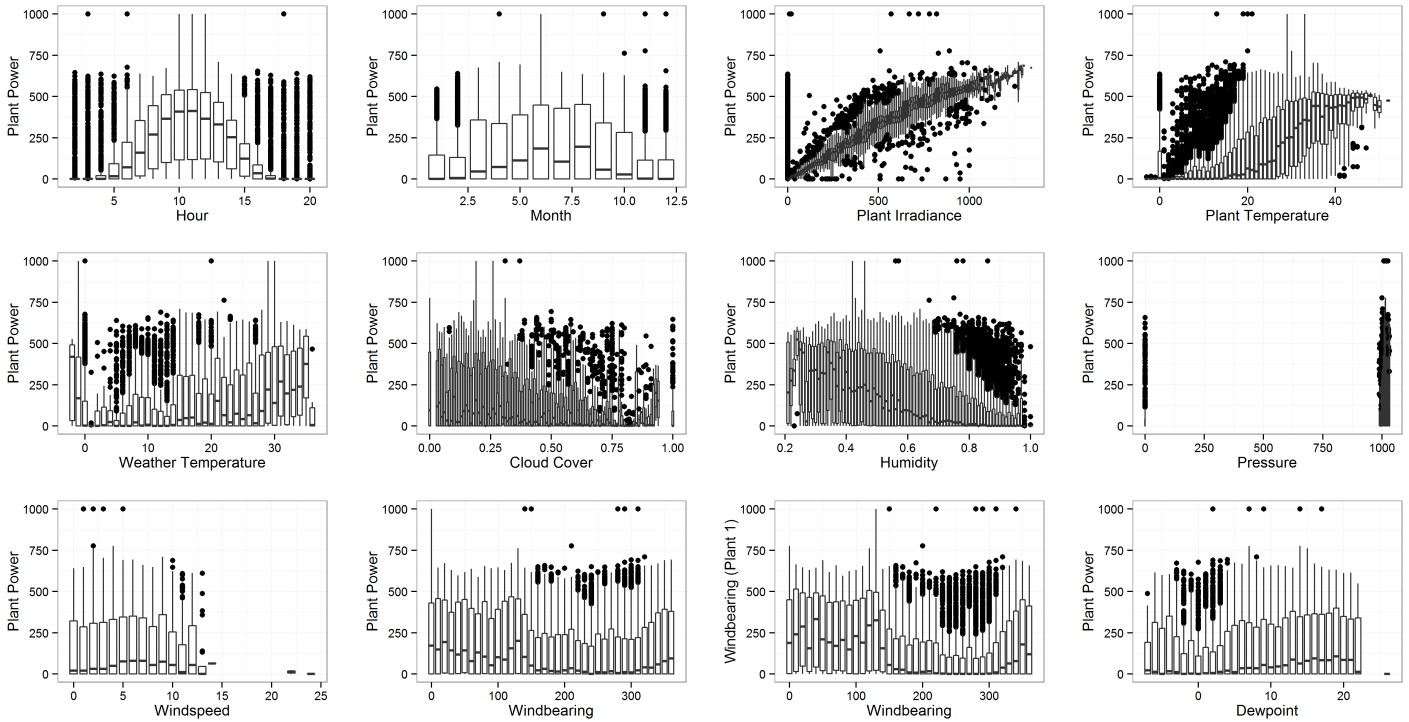

Fig. 1. Relation between single variables and the power production of the plants. Note that we collected the measurements for all three plants. In the plots depicting the plant irradiance, wind speed, wind bearing, dewpoint, and pressure, we rounded the measured values to obtain reasonable box plots.

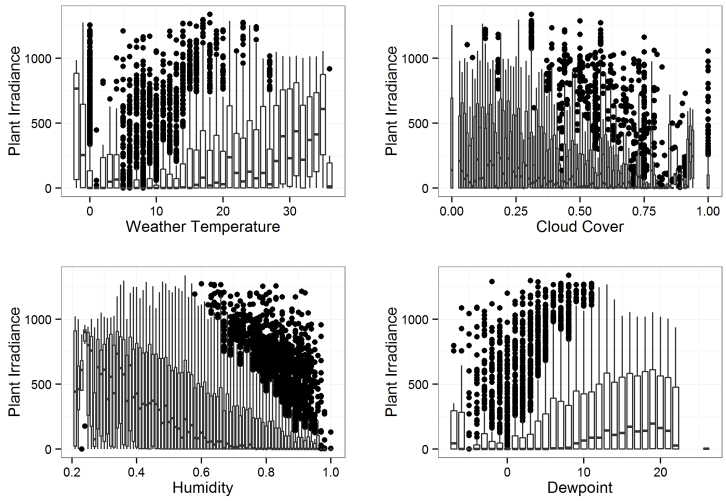

Fig. 2. Relation between single variables and the plant irradiance. Note that we collected the measurements for all three plants.

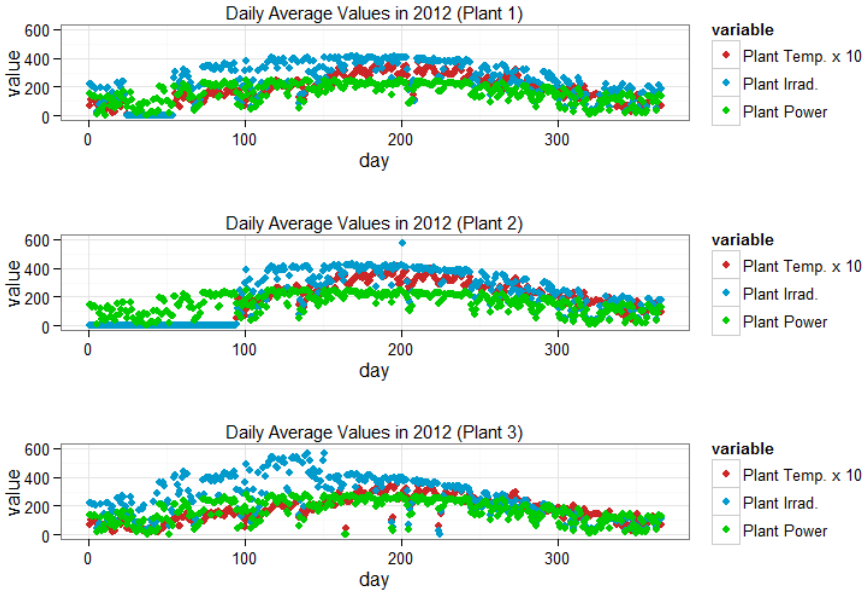

Fig. 3. Daily average measurements of plant power, plant irradiance and plant temperature per plant.

4 Data Preprocessing

From Sec. 3, we know that the data set most probably contains a lot of missing values, which are indicated by zero measurements, and outliers.

Missing Data. Recall that missing values are indicated by *zero* measurements. However, a zero measurements does not necessarily indicate a missing value. Therefore, we did *not* mark all zero measurements as missing: In the plant irradiance time series, we marked a zero measurement as missing if, at the same time point, the plant temperature was also zero (i.e. we marked 2577 of 7921 zero measurements as missing) (cf. exploration of data regarding plant 2 in Sec. 3.2). After that, we marked all zero measurements in the plant temperature, weather pressure, and weather pressure time series as missing.

Outlier Removal. First, we removed unusual values. In the given training data, the minimum plant temperature is -3°C, while in the test data we observe plant temperature measurements of up to **-202°C**. This is why we decided to remove (i.e. mark as missing) all plant temperature measurements of less than -3°C from the test data.

Second, we used the outlier removal proposed by [1]. That is, given the time series $(v_t)_{t=1,\dots,T}$ (i.e. the plant irradiance time series), we consider the value v_t at time t an outlier if $v_t \notin [\bar{v} \pm 4 \cdot \sigma_v]$, where $\bar{v} = \frac{1}{T}\sum_{t=1}^{T} v_t$ denotes the sample mean and $\sigma_v^2 = \frac{1}{T-1}\sum_{t=1}^{T} \|v_t - \bar{v}\|^2$ denotes the (unbiased) sample variance. To be precise, we conduced this outlier removal separately for the time series given by the training data and by the test data, respectively.

Normalization of Power Measurements. After our outlier removal there were still some power measurements larger than the maximal power value of 1 000 kW/h. We simply replaced these measurements with the maximal power value (i.e. 1 000 kW/h).

Additional Date Features. We created one training and test data set that contains the plant ids, the plant sensor features, weather features, and additionally, three date features. The date features indicate the hour of the day (by values $1, \dots, 19$), the day of the year (by values $1, \dots, 366$), and the month (by values $1, \dots, 12$).

5 Prediction

In the following, we first describe two approaches: a regression tree ensemble and an evaluation of hourly averages. The latter has turned out to be the winning solution of the challenge.

5.1 Training Data Splits

In order to estimate the performance of our approaches, we split the training data in two splits: The first split T_{1-23} contains all data from the training data set concerning the first 23 days of each month in 2012. The second split T_{24+} contains the remaining data, i.e. all training data containing the 24th day through the last day of each month in 2012. In the following, we train our models using the larger split T_{1-23}, and evaluate the resulting models with respect to T_{24+}.

Our motivation for these splits is twofold: First of all, we want to ensure that there is enough information about each month contained in each split. Second, we want the splits to correspond to rather independent time series. In other words, we want that the number of measurements v_t for which the subsequent measurement v_{t+1} is contained in a different split than v_t is as small as possible. This property would be violated if, for example, we would choose the assignment of a measurement to one of the splits uniformly at random.

5.2 Approaches: Regression Tree Ensembles and Hourly Averages

Our first approach is to use gradient boosted regression trees (GBRT) [2], which can handle heterogeneous data with missing values. To tune the hyperparameters of the GBRT technique, we apply a black-box optimization method [3]. That is, we apply Bayesian optimization to find the hyperparameters λ which minimize the RSME of gradient boosted regression trees, which have been trained with hyperparameters λ and training data T_{1-23}, with respect to the validation split T_{24+}. As the resulting regression tree ensemble predicted rather strange time series (regarding the testing data), we also tried the following simple baseline.

Our second approach is to predict hourly average values: For each plant p, hour of the day h, and for each month m, we computed the average power production of plant p during the month p at this specific hour h of the day. For instance, our prediction of the power production of plant 1 for February 1st between 1pm and 2pm is the average power production of plant 1 between 1pm and 2pm in February 2012.

5.3 Evaluation

First, let us consider the performance of our models with respect to the validation split T_{24+}: Our resulting regression tree ensemble yields a root mean squared error (RSME) of 0.0685. In comparison, our hourly average approach performs poorly and only achieves an RSME of 0.1297. However, our models perform differently with respect to the testing data.

Regarding the testing data, our simple hourly average approach turned out to be the winning approach. Tab. 1 depicts the temporary and the final leaderboard, which have been published on the challenge website. Our simple hourly average approach already achieved the second best score on the temporary leaderboard, which contains 10% of the testing data. Regarding the complete testing data, the hourly approach outperformed all the other competing solutions.

Table 1. Leaderboards: The testing data consists of 3 months (January to March 2013) for which the target time series (power) is not provided. The temporary leaderboard is obtained with respect to only 10% of the testing data, while the final leaderboard takes into account the complete testing data. Only the best submission per team is listed. Our hourly average approach is marked in bold.

Temporary		**Final**	
position	RSME	position (previous)	RSME
1	0.16697563293016	**1 (2)**	**0.20797269178412**
2	**0.16864488931773**	2 (4)	0.22345258018878
3	0.17726390671562	3 (6)	0.22633730458759
4	0.18583154100811	4 (3)	0.25383336655979
5	0.19462844662109	5 (1)	0.25432166920657
6	0.21447590228825	6 (10)	0.26261669303547
7	0.22291339472308	7 (7)	0.26687963564708
8	0.22307448875416	8 (9)	0.27442245421399
9	0.22404629864761	9 (8)	0.27587241955111
10	0.2267988609685	10 (11)	0.28524305512214
11	0.23031635550655	11 (5)	0.30189592552655

5.4 Conclusion

As already explained, the given training data contained many missing values, especially, during the time of the year, for which we have to predict the power production in the subsequent year (see Sec. 3.2). We expect that, given more reasonable data, a machine learning approach will outperform our simple hourly average approach.

References

1. Ceci, M., Corizzo, R., Fumarola, F., Malerba, D., Rashkovska, A.: Predictive modeling of pv energy production: How to set up the learning task for a better prediction? IEEE Transactions on Industrial Informatics 13(3), 956–966 (June 2017)
2. Chen, T., Guestrin, C.: Xgboost: A scalable tree boosting system. In: Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 785–794. KDD '16, ACM, New York, NY, USA (2016), `http://doi.acm.org/10.1145/2939672.2939785`
3. Snoek, J., Larochelle, H., Adams, R.P.: Practical bayesian optimization of machine learning algorithms. In: Advances in Neural Information Processing Systems 25: 26th Annual Conference on Neural Information Processing Systems 2012. Proceedings of a meeting held December 3-6, 2012, Lake Tahoe, Nevada, United States. pp. 2960–2968 (2012), `http://papers.nips.cc/paper/4522-practical-bayesian-optimization-of-machine-learning-algorithms`