Report - Solar power Generation

ing. Wojciech Pribula

2021-05-12

Contents

1 Introduction **2**
1.1 Data structure 2
1.2 Main objective 2

2 Data analysis **3**
2.1 Night 3
2.2 Corelation 3
2.3 Closer examination of some variables 4
2.4 Summarise of data analysis 6

3 Method for finding best prediction model **7**
3.1 Issues with some methods 9
3.2 Importance of variables 9

4 Results **11**

5 Conclusion **12**

1 Introduction

Data set used in this project contains data from SOlar power generation facility in Berkeley, CA. Source of these data is: https://www.kaggle.com/vipulgote4/solar-power-generation.

Data set contains some weather and environment measurements as temperature, wind speed and direction or visibility. Variable of interest is generated power, which should be predicted based on other variables.

Data set contains 2920 observations from September 2008 to May 2009. Each observation contain day or period averages of measured values.

The data set needs to be divided into training (80%) and validation (20%) sets and validation set should not take part in training to allow validation of trained models.

1.1 Data structure

```
## 'data.frame':	2920 obs. of  16 variables:
##  $ Day.of.Year                         : int  245 245 245 245 245 245 245 245 246 246 ...
##  $ Year                                : int  2008 2008 2008 2008 2008 2008 2008 2008 2008 2008 ...
##  $ Month                               : int  9 9 9 9 9 9 9 9 9 9 ...
##  $ Day                                 : int  1 1 1 1 1 1 1 1 2 2 ...
##  $ First.Hour.of.Period                : int  1 4 7 10 13 16 19 22 1 4 ...
##  $ Is.Daylight                         : logi  FALSE FALSE TRUE TRUE TRUE TRUE ...
##  $ Distance.to.Solar.Noon              : num  0.8599 0.6285 0.3972 0.1658 0.0656 ...
##  $ Average.Temperature..Day.           : int  69 69 69 69 69 69 69 69 72 72 ...
##  $ Average.Wind.Direction..Day.        : int  28 28 28 28 28 28 28 28 29 29 ...
##  $ Average.Wind.Speed..Day.            : num  7.5 7.5 7.5 7.5 7.5 7.5 7.5 7.5 6.8 6.8 ...
##  $ Sky.Cover                           : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Visibility                          : num  10 10 10 10 10 10 10 10 10 10 ...
##  $ Relative.Humidity                   : int  75 77 70 33 21 20 36 49 67 49 ...
##  $ Average.Wind.Speed..Period.         : int  8 5 0 0 3 23 15 6 6 0 ...
##  $ Average.Barometric.Pressure..Period.: num  29.8 29.9 29.9 29.9 29.9 ...
##  $ Power.Generated                     : int  0 0 5418 25477 30069 16280 515 0 0 0 ...
```

1.2 Main objective

Main task is to predict generated power better than use of average from previous records. To compare results RMSE (Root-mean-square deviation) can be used:

$$RMSE = sqrt(\frac{\sum_{i=1}^{N}(x_i - \hat{x}_i)^2}{N})$$

N ... number of observations

x_i ... original value

$\hat{x}_i$ ... predicted value

RMSE when using average is 1.0151131×10^4 and this precision should be beaten by trained model.

2 Data analysis

Data should be analyzed first to provide some basic idea about relationships of predictors to each other and to predicted value.

2.1 Night

Firs what can be noticed is that data set contains information if it is day or not ($Is.Daylight$). Both average and standard deviation of $Power.Generated$ is 0 when filtered for $Is.Daylight = FALSE$. This is expected as there is not sunlight at night so solar power station can't generate power. Following this all night data can be filtered out from the data set as prediction for this is always 0.

Filtered data set than has 1805

2.2 Corelation

What should be examined next is correlation between data.

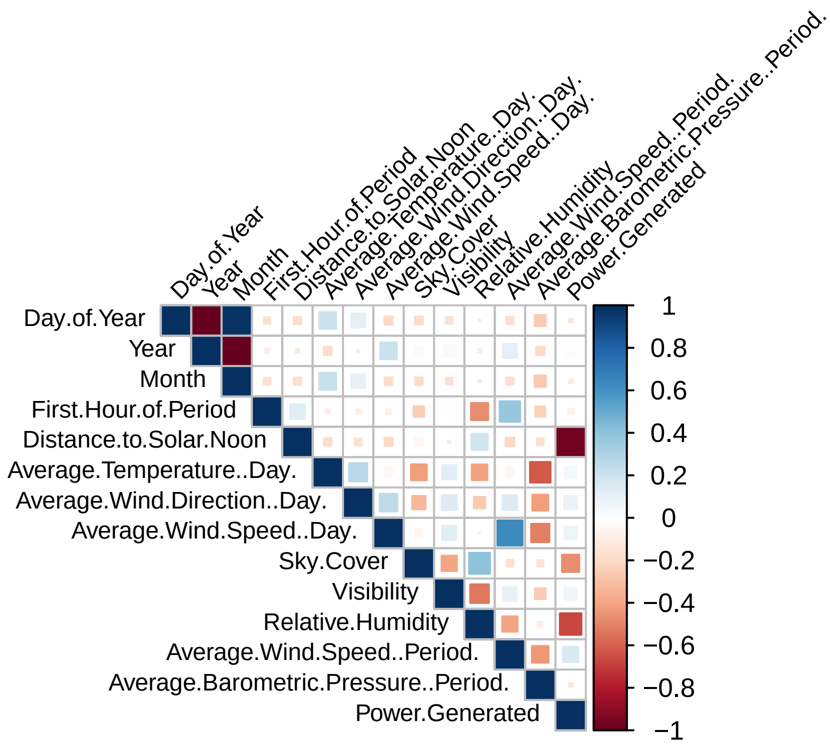

Figure 1: Correlation between variables.

What can be observed is that there is very strong negative correlation between generated power and Distance to solar noon, Relative humidity and Sky cover, so it is expected that these three variables should have strong influence on prediction. However as Sky cover and Relative humidity are positively correlated, only two of these may by main influencers. Visibility has positive correlation with with power generation, however it can be observed that it is practically opposite of Sky cover. What may surprise is that there is positive correlation for power generation and wind direction and speed, this can be explained by wind influence on weather. Wind has positive correlation with temperature. What can be deducted is that faster wind from specific direction moves could away and does not bring new clouds.

2.3 Closer examination of some variables

Negatively correlated variables: Distance to solar noon and Relative humidity. Relationship is not that clear from point data, however smooth line shows that some relationship exists and it should be very strong.

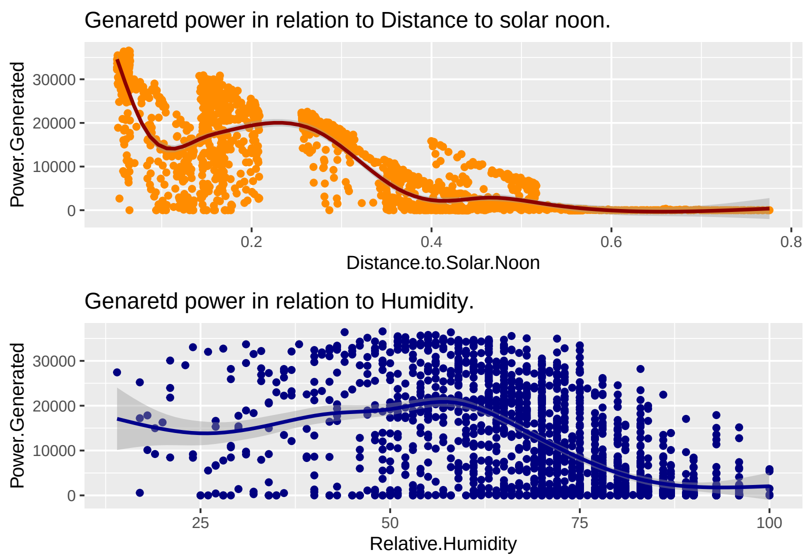

Figure 2: Genrated power and distance to the solar noon and humidity.

It can be seen from following plots that Sky coverage and visibility have negative correlation to each other and that some relation to power generation exists. Sky coverage seems to have strong relation to generated power what is logical as more clouds means less sunlight so less power generated.

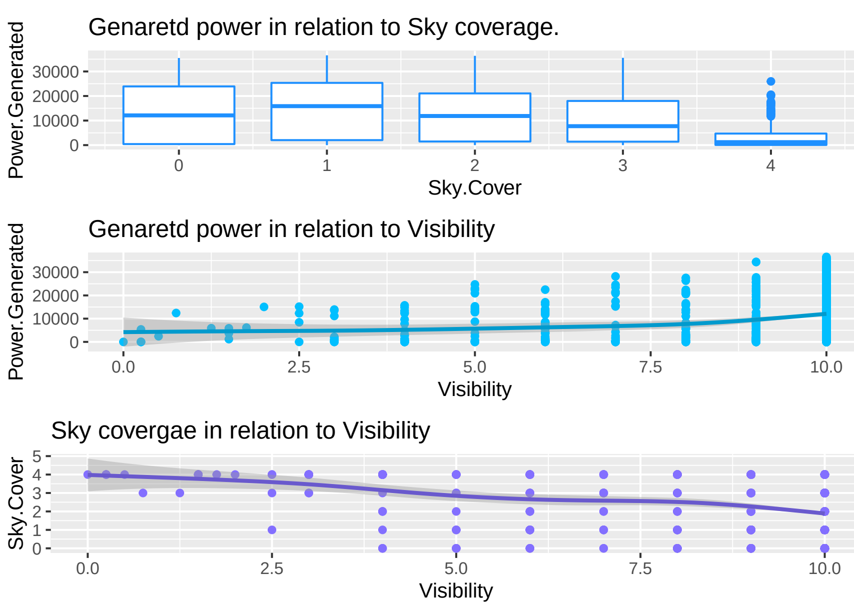

Figure 3: Genrated power and Sky coverage and visibility.

Wind speed, wind angle and temperature should have some influence on predicted generated power. This can be examined from following plots. What can be observed on points is that these relations are not strong. It can be seen that some relations to generated power exists, however these won't be strong predictors due to big spread of individual observations. Relationship between angle and temperature is clearly visible too.

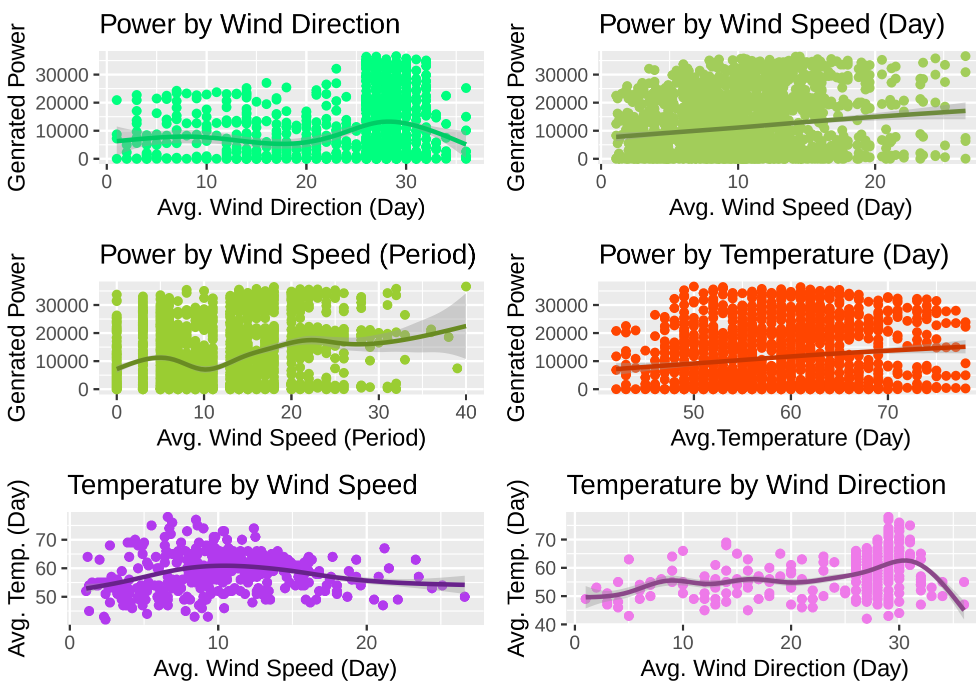

Figure 4: Genrated power realtion to wind speed, wind angle and temperature.

To make weather analysis full barometric pressure should be examined too. Relation is not very clear, however it may help a little with prediction.

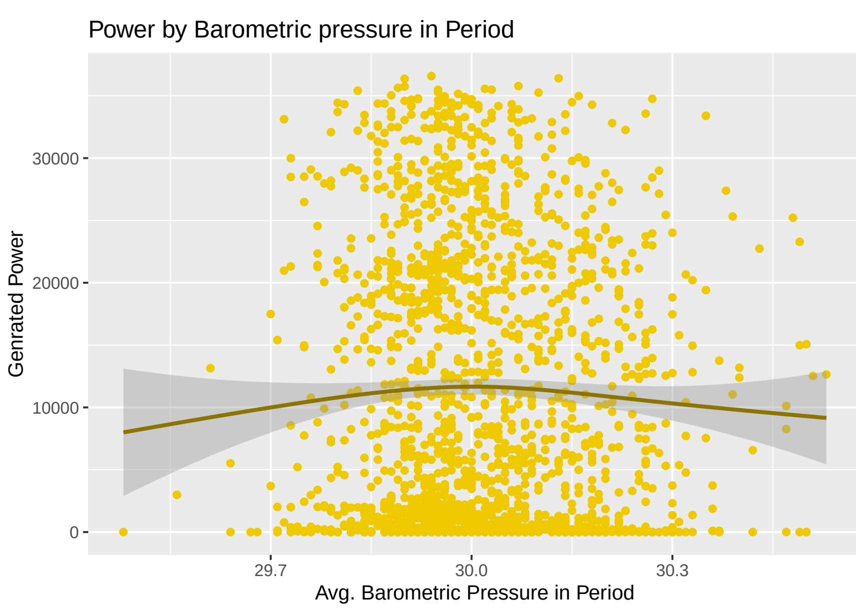

Figure 5: Power by Brometric pressure in period.

The last area which needs to be examined is influence of time. Weather works in year cycles so it can be expected that day of the year and month should have influence on generated power. Day of the year and month are highly correlated and may be unnecessary to include both of them.

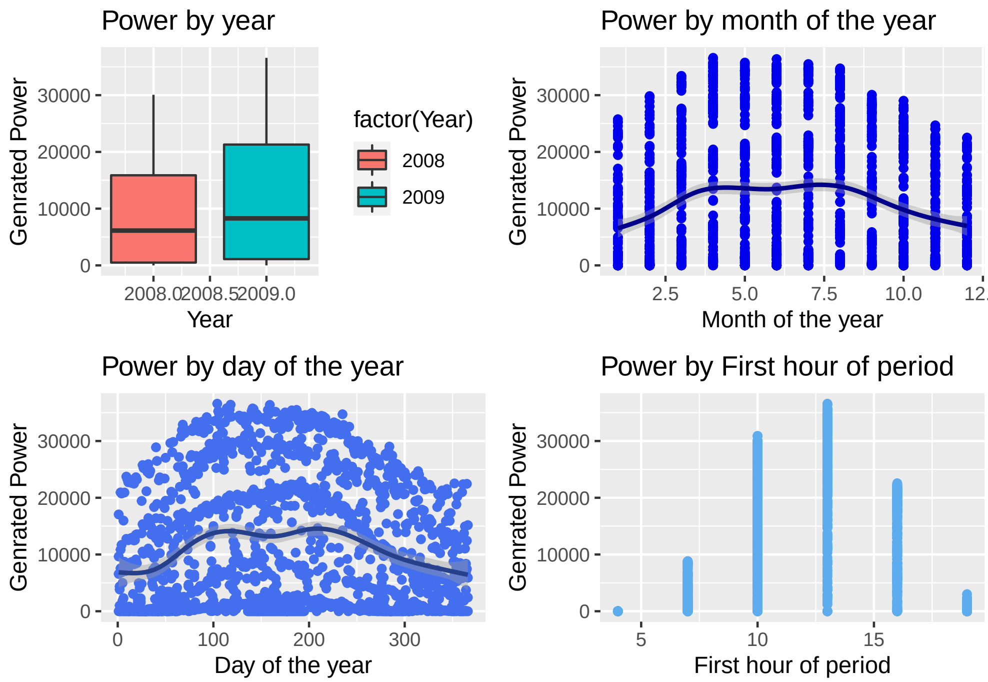

Figure 6: Genrated power in time.

2.4 Summarise of data analysis

It seems that each parameter has some influence on generated power however it is not always very straight as spread of values is wide (visible when examining point plots). It seems that it would be necessary include most of variables to do prediction of generated power with smaller error. Most variables are depended on each other in some degree, however that dependency is mostly complicated so all parameters should be used for training. There is few straight dependencies between variables, however even these seems to be influence in some degree by other factors. It can be said that all measured values have sense and were well chosen for this data set to allow prediction of generated power.

3 Method for finding best prediction model

There are many methods how to train model, however it is not obvious which one should be the best. Here is list of possible methods:

knn - k-Nearest Neighbors

glm - Generalized Linear Model

treebag - Bagged CART (Classification And Regression Tree)

ctree2 - Conditional Inference Tree

rf - Random Forest

rpart - CART - Classification And Regression Tree

rpart2 - CART - Classification And Regression Tree

bridge - Bayesian Ridge Regression

ppr - Projection Pursuit Regression

gaussprLinear - Gaussian Process

gamSpline - Generalized Additive Model using Splines

brnn - Bayesian Regularized Neural Networks

Then algorithm for whole process can look like this:

1. Load data and divide to training data (80
2. Remove night rows.
3. Use train data to train all models 5 times.
   (a) Divide training data to training data set (80
   (b) Train all models and calculate RMSE using test data set.
4. Calculate mean RMSE for each method.
5. Find the best performing method.
6. Use original training data to train best performing method.
7. Validate on validation data and calculate RMSE

This algorithm should allow cross-validation of all models and should provide good prediction on averall models performance on this data set.

Here are results for all methods:

Table 1: RMSE for each cycle and average RMSE for all cycles.

Method	1	2	3	4	5	Average
rf	3117.839	2866.210	2816.187	2712.214	3066.466	2915.783
brnn	3416.774	3219.982	3436.214	2949.618	3198.228	3244.163
ppr	3525.907	3229.523	3627.853	3299.664	3230.152	3382.620
ctree	3747.966	3366.620	3551.698	3100.934	3491.831	3451.810
ctree2	3747.966	3388.962	3551.698	3100.934	3491.831	3456.278
rpart	4032.585	3643.192	3256.717	3172.810	3292.871	3479.635
treebag	4017.673	3881.077	3822.727	3467.655	3722.782	3782.383
rpart2	4114.679	4162.770	4037.537	3725.252	3844.473	3976.942
gamSpline	4243.812	4039.831	4035.049	3699.718	4020.932	4007.868
gaussprLinear	6209.562	6124.718	6255.596	6038.510	6222.876	6170.252
glm	6209.913	6124.807	6256.004	6040.776	6223.452	6170.990
bridge	6217.222	6115.794	6244.535	6048.608	6231.845	6171.601
KNN	6507.164	6870.138	6430.034	6405.163	6841.681	6610.836
Average	10463.593	10347.705	10234.160	10573.829	10602.901	10444.438

The best performing method seems to be Random Forrest followed by Neural network with one internal layer of neurons. **This can be confirmed in following plot too:**

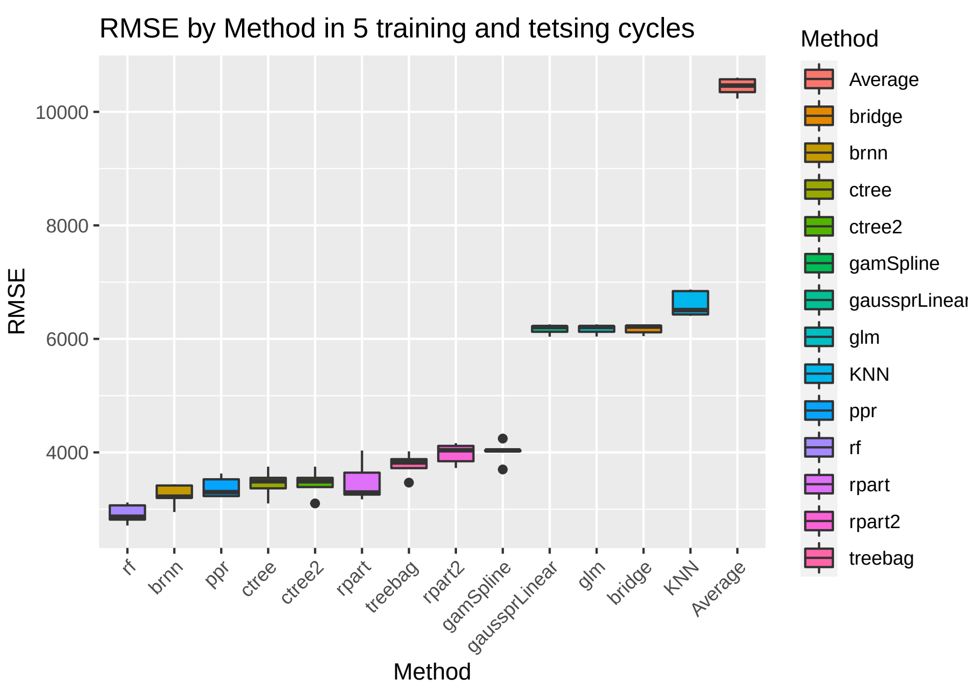

Figure 7: RMSE by Method in 5 training and tetsing cycles.

3.1 Issues with some methods

Some methods as KNN work better with standardized data (centered around average value and divided by standard deviation). However these methods do not perform better than winning methods even if standardized data are used. So this does not need to be taken in count as winning method wins in both cases, with standardized and original data.

Some methods allow some tuning parameters. Ranges for these were configured during algorithm testing and development to cover ranges which allow good tuning for this data.

3.2 Importance of variables

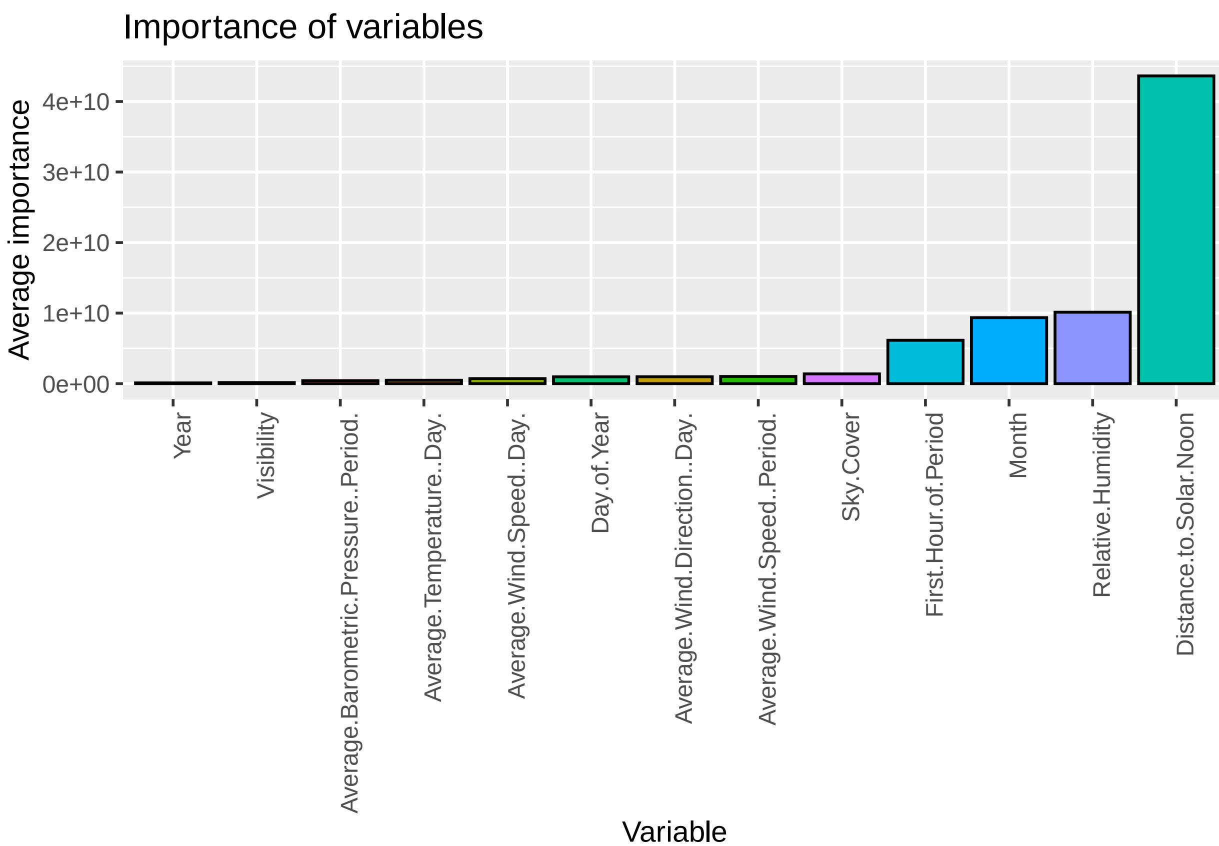

Figure 8: Importance of variables.

RMSE for all models when using only four most important variables which stand out in above plot is in following table and it is clear that results are worse, so if training time is not important it is better to include all variables or at least more than four.

Table 2: RMSE for each cycle and average RMSE for all cycles for top 4 most important variables

Method	X1	X2	X3	X4	X5	Average
rf	5027.517	4667.483	3960.930	4343.939	3838.085	4367.591
ctree2	4886.040	4901.064	4170.918	4567.309	3939.628	4492.992
brnn	5170.960	4933.179	4309.480	4353.803	4009.452	4555.375
rpart	5055.533	5096.980	4200.923	4554.839	4130.811	4607.817
ppr	5278.774	4927.842	4115.452	4564.463	4244.460	4626.199
gamSpline	5033.445	5497.219	4719.981	4679.778	4607.509	4907.586
treebag	5319.380	5693.398	4421.183	4709.248	4509.195	4930.481
rpart2	5678.298	6182.477	4888.287	5030.247	4763.702	5308.602
KNN	6538.937	5839.266	5607.404	5743.306	5968.089	5939.400
glm	5970.804	6546.222	6025.592	5805.526	5642.619	5998.153
gaussprLinear	5970.354	6545.578	6025.773	5806.041	5643.528	5998.255
bridge	5969.010	6544.853	6025.917	5806.927	5645.999	5998.541
Average	10793.651	11083.114	11224.953	11346.733	10771.546	11043.999

4 Results

Best method to train model for these data seems to be Random Forrest which produces best results if compared with RMSE function.

Using more models and then averaging them is not good option as this does not improve results.

Final results with variables importance follow what was discovered in data analysis chapter, that Distance to Solar Noon, Humidity, Hour of the day and Sky coverage have big influence on final prediction.

Final $RMSE = 2841.67$ is much better in compare to situation when only Average is used - RMSE = 10151.13.

Here is plot of actual values and predicted values

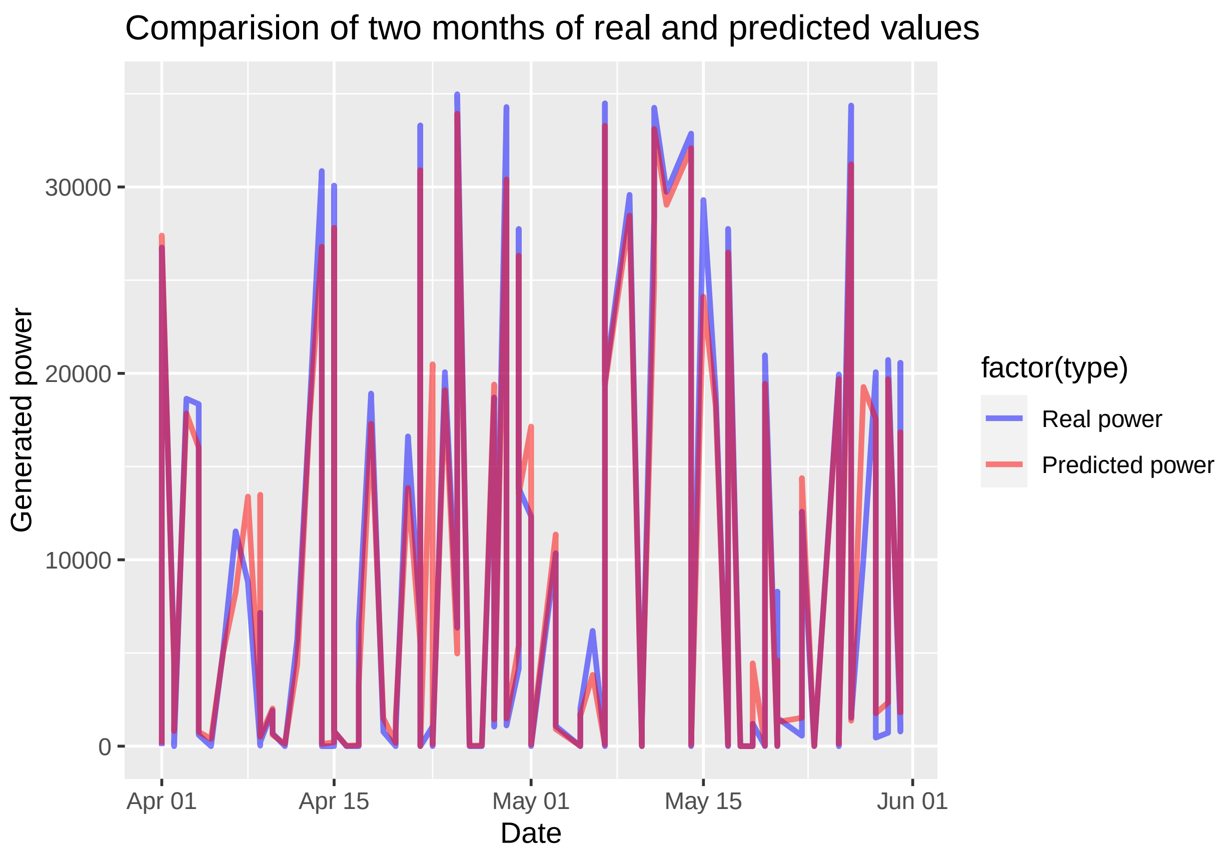

Figure 9: Comparision of two months of real and predicted values.

5 Conclusion

Final prediction model seems to be performing well and prediction is close to real values.

Measured data were selected well and allowed good prediction.

Number of predictors did not play big role in this case as number of observations is low so training time is not very long. If more observations should be introduced then it would be good to consider reduction of predictors to allow faster training. However this would need to be done with more research on relationships between values.

More data from more years could provide better or worse predictions, that would require more research and more data.

Second best method was neural network with one internal layer with ± 8 neurons. This is good result and number of neurons is expected as more neurons could lead to nonconverging training of the network

Best method is Random Forest, which seems to be good choice for this kind of task, however requires significantly longer time for training.

It would be interesting to compare these two methods on larger set of data. It is possible that neural netowork could be better than random forest.

Other models may perform well too, here is comparision of 5 best performing models on training data when run on validation data:

Table 3: RMSE for 5 best performing models when run on validation data.

Method	RMSE
rf	2850.955
brnn	2946.773
rpart	3265.290
ctree2	3396.210
ppr	3620.026