Solar Radiation Prediction

Kai Ye ye000064

1. Background and introduction

Developing renewable energy has raised more and more public awareness since the situation of global warming is getting worse while the industrial demand for energy is still increasing. One of the biggest obstacles in the effective utilization of renewable energy is the uncontrollability of renewable energy resources like solar energy. Solar energy is commonly used due to its abundance and easy accessibility. However, the utilization of solar energy completely depends on the radiation from the sun thus it’s impossible to manually control the energy output like traditional power generation methods. Therefore, it would be useful to predict the radiation in a certain area and time by analyzing previous data.

I used a data set named “Solar Radiation Prediction” from Kaggle[1] and tried to build a model to predict the radiation using weather data such as wind speed, temperature and so on.

2. Analysis and preprocessing of the data set

The data set is given with the radiation of an unknown area in four months (September to December in 2016) with several other parameters about weather conditions: temperature, pressure, humidity, wind speed and direction, sunrise time, and sunset time.

From the data of December 2016, the daytime is shorter than night-time thus the area is at the North Hemisphere. And the day time on September 1st is around 12.5 hours, which is the same as the daytime at Lahaina in Hawaii based on my search. Thus the place is located at around the same latitude as Lahaina, that is, 20.88° N.

The data are recorded at a certain time respectively. Therefore, I used the ‘Chron’ library in R to process times and dates. I transformed these times and dates to the fractions of a day or year, which is more meaningful for the prediction of solar radiation. And I added another column showing the length of daytime by taking the difference of sunset and sunrise time. I think it would be more straightforward and useful than a specific sunrise or sunset time. For the requirement of linear independence, only sunrise time and the daytime length is used in my modeling. What’s more, I added another tag to tell if the data is recorded during day time since there is no direct sunshine during night-time and the radiation is close to zero.

3. Development of the full backward elimination model based on all predictors

I randomly took 800 (5%) samples and used the “pairs()” function to visualize the relationships. And the plot is shown in Figure 1.

As shown in the plot, all these variables might be useful for the prediction of radiation. One anomaly in the figure is that there exist some irregular values in “ws” (wind speed). It shows that there are around five data points in my samples with unusually large wind speed. By analyzing the distribution of wind speed in the complete data set, the mean is 6.285 and the third quarter is 7.87. However, the maximum wind speed is 40.5, much larger than the mean. Thus the exclusion of data points with large windspeed may be useful in later analysis.

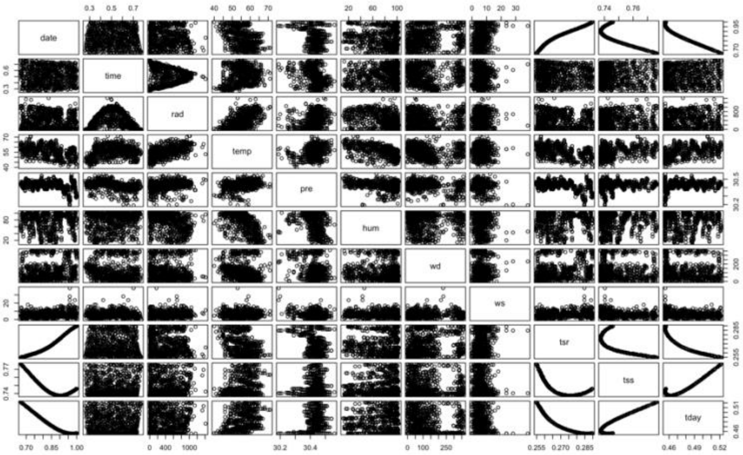

Figure 1. All of the pairwise comparisons for daytime solar data

Then I built a multi-factor regression model using all the available parameters.

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  1.540e+04  1.401e+03  10.989  < 2e-16 ***
date         1.163e+03  3.709e+02   3.134  0.00173 ** 
time        -8.267e+02  1.619e+01 -51.048  < 2e-16 ***
temp         5.202e+01  4.954e-01 105.006  < 2e-16 ***
pre         -5.363e+02  4.308e+01 -12.448  < 2e-16 ***
hum          2.140e+00  9.849e-02  21.730  < 2e-16 ***
wd          -2.432e-01  1.835e-02 -13.254  < 2e-16 ***
ws           7.835e+00  5.221e-01  15.006  < 2e-16 ***
tsr         -5.540e+03  1.828e+03  -3.030  0.00245 ** 
tday        -1.561e+03  8.592e+02  -1.816  0.06932 .  

Residuals:
    Min      1Q  Median      3Q     Max 
-807.73 -165.79  -17.58  143.42  963.42 

Residual standard error: 227.9 on 15598 degrees of freedom
Multiple R-squared:  0.5369, Adjusted R-squared:  0.5366
```

The p-value of ‘tday’ is close to our common threshold p=0.05, thus backward elimination is not required for this data set. However, the performance of this model is suboptimal. The residual standard error is large and the adjusted R-squared value is small.

Then I partitioned the data into training and testing sets using portions of 60% for training and 40% for testing. Based on the analysis of the effects of varying partition fraction on data processing, the same partition fraction is also used for predictability tests later in this report. The RMSE of the prediction from this model is around 228, which is averaged on 100 times of simulations.

4. Further analysis and segmentation modeling

Inspired by the segmentation model we built last time, it might be useful to segment the data to achieve a better model with an adjusted R_2 value close to 1.

First, as mentioned in the previous part, we can exclude large wind speed values. Based on the distribution of wind speed as shown in Figure 2, I extracted the data points with wind speed smaller than 15 for further analysis.

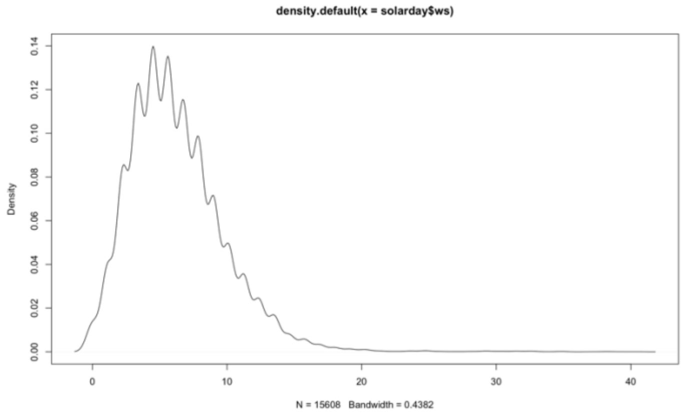

Figure 2. The density plot of the distribution of wind speed in daytime solar data

Then model after refitting is shown below:

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  1.750e+04  1.454e+03  12.032  < 2e-16 ***
date         1.031e+03  3.741e+02   2.756  0.00585 ** 
time        -8.216e+02  1.628e+01 -50.453  < 2e-16 ***
temp         5.192e+01  4.990e-01 104.055  < 2e-16 ***
pre         -6.040e+02  4.479e+01 -13.485  < 2e-16 ***
hum          2.032e+00  9.941e-02  20.439  < 2e-16 ***
wd          -2.457e-01  1.839e-02 -13.360  < 2e-16 ***
ws           1.058e+01  6.008e-01  17.615  < 2e-16 ***
tsr         -4.889e+03  1.839e+03  -2.658  0.00787 ** 
tday        -1.790e+03  8.682e+02  -2.062  0.03927 *  

Residuals:
    Min      1Q  Median      3Q     Max 
-816.87 -164.94  -15.98  142.77  956.32 

Residual standard error: 227.5 on 15320 degrees of freedom
Multiple R-squared:  0.5388, Adjusted R-squared:  0.5385 
F-statistic:  1989 on 9 and 15320 DF,  p-value: < 2.2e-16
```

RMSE = 227.15

The model is slightly improved and we need further segmentation. As stated before, the data are recorded in only 4 months, which is too short to be divided by seasons. And from our knowledge of solar radiation, it's stronger around noon. Thus I divided the data set into two sets according to the time in a day. One set consists of data recorded in 3 hours before or after the local noon, that is, 6 hours centered at local noon. The other is the data left since the daytime is generally around 12 hours per day.

1) For the peak hours, my final model is:

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  9.412e+03  2.109e+03   4.462 8.22e-06 ***
date         5.967e+03  5.290e+02  11.279  < 2e-16 ***
time        -7.206e+02  4.257e+01 -16.927  < 2e-16 ***
temp         2.601e+01  9.981e-01  26.056  < 2e-16 ***
pre         -3.471e+02  6.514e+01  -5.330 1.01e-07 ***
```

```
hum          -1.705e+00  1.802e-01  -9.466  < 2e-16 ***
wd           -1.922e-01  2.314e-02  -8.308  < 2e-16 ***
ws            9.988e+00  8.405e-01  11.884  < 2e-16 ***
tsr          -3.396e+04  2.623e+03 -12.947  < 2e-16 ***
tday          1.022e+04  1.227e+03   8.329  < 2e-16 ***

Residuals:
    Min      1Q  Median      3Q     Max 
-741.87 -163.30    0.46  153.97  958.54 

Residual standard error: 230.3 on 7943 degrees of freedom
Multiple R-squared:  0.421,  Adjusted R-squared:  0.4203
```

RMSE = 230.275

2) For the valley hours,

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  1.525e+04  1.337e+03  11.409  < 2e-16 ***
time        -5.361e+02  1.221e+01 -43.920  < 2e-16 ***
temp         3.335e+01  4.823e-01  69.149  < 2e-16 ***
pre         -5.399e+02  4.179e+01 -12.920  < 2e-16 ***
hum          7.997e-01  8.378e-02   9.546  < 2e-16 ***
wd          -1.889e-01  1.984e-02  -9.525  < 2e-16 ***
ws           3.018e+00  5.686e-01   5.308 1.14e-07 ***
tsr          1.746e+03  4.541e+02   3.845 0.000122 ***
tday        -1.297e+03  2.270e+02  -5.714 1.15e-08 ***

Residuals:
    Min      1Q  Median      3Q     Max 
-403.55 -104.94  -18.05   90.30  805.91 

Residual standard error: 143.6 on 7368 degrees of freedom
Multiple R-squared:  0.4879, Adjusted R-squared:  0.4873
```

RMSE = 143.7

3) Based on the idea of this segmentation and daytime filtering in part 2, I defined a new period from 1.5 hours after sunrise to 1.5 hours before sunset as the “effective working hours” for solar devices. The radiation at daytime excluding this period is usually weak according to Figure 3.

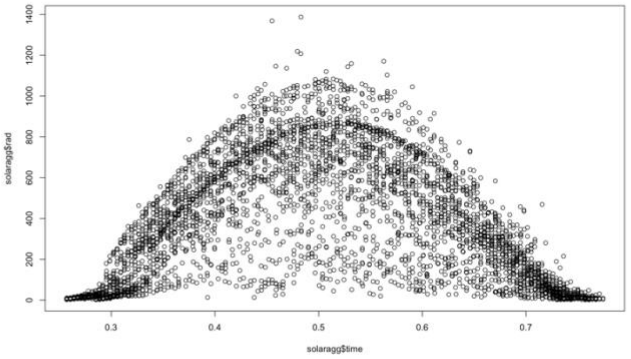

Figure 3. Average radiation versus time at daytime (time shown as the fraction of a day)

And my final “effective-working-hour” model is:

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.000e+04  1.774e+03   5.637 1.77e-08 ***
date         3.725e+03  4.461e+02   8.351  < 2e-16 ***
time        -6.850e+02  2.427e+01 -28.222  < 2e-16 ***
temp         3.604e+01  5.073e-01  71.042  < 2e-16 ***
pre         -3.622e+02  5.473e+01  -6.618 3.81e-11 ***
wd          -1.951e-01  2.019e-02  -9.662  < 2e-16 ***
ws           1.185e+01  7.039e-01  16.837  < 2e-16 ***
tsr         -2.048e+04  2.192e+03  -9.343  < 2e-16 ***
tday         4.651e+03  1.039e+03   4.476 7.68e-06 ***

Residuals:
    Min      1Q  Median      3Q     Max 
-736.68 -171.92   -6.31  151.19  980.92 

Residual standard error: 230.1 on 11303 degrees of freedom
Multiple R-squared:  0.4191, Adjusted R-squared:  0.4187
```

RMSE = 230.0

4) Conclusion

By segmentation, the adjusted R_2 value is decreased. However, model (1) fits better with residuals almost distributed in the normal distribution. Thus the segmentation in (1)&(2) is a good attempt. Similarly, model (3) is a compromise between model (1) and the original model, and the improvement is moderate.

5. Conclusion

In summary, the segmentation model based on time is a convincingly good trial. However, the performance improvement is not very successful with lower adjusted R_2 values. As analyzed before, this area is located at around 20.88° N, where generally the radiation can be strong all year. And the climate also matters, which is not given in the data set. What’s more, the prediction results are suboptimal since some specific weather conditions like clouds are hard to be analyzed from the given data, which may also have strong effects on solar radiation. Hopefully, this project gives us some insights into the uncontrollability of solar radiation.

[1] Solar radiation prediction (Task from NASA Hackathon), data provided by NASA, accessed 28th November 2019, <https://www.kaggle.com/dronio/SolarEnergy >