Predicting the Solar Energy of Illinois

Reem Aldraiwaish

Abstract

The goal of this project was to use Regression models to predict the solar energy amount from a solar power plant at the University of Illinois in order to help improve grid management and maintenance. I worked with data provided by NOAA and University of Illinois, leveraging numerical feature engineering along with a linear regression model to achieve results for this continuous quantity problem. Some visuals for Solar Energy through date and time.

Design

This project build for data science Course. The data is provided by NOAA and University of Illinois, and presents a continues amount of Solar energy on hourly and daily basis. Predicting the amount solar accurately via machine learning models would enable the University of Illinois to take action to improve grid management and maintenance of solar plant, allocate resources more quickly to needed areas, and ensure solar energy is access to as many people as possible.

Data

I have two data set the first is have weather and solar energy on an hourly basis which contains 15,072 observation with 16 features, all are numerical where date and time feature combined into a single date time feature, It has a daily resolution 6am until 5pm starting from 02/01/2016 - 09/31/2017. Since I found that the weather has no big differences among the day itself, Then I choice better to work with the daily basis instead of hourly bases and this data set contains 637 observation with 10 features all are numerical.

Algorithms

Feature Engineering

- I processed and merge weather data inputs with the solar energy output file in order with to get meaningful numeric values with hourly resolutions.

- Dropping Inverters column since it is all contain null value and have the same meaning of solar energy.

- I run the correlation analysis between the weather features and the energy output and I figure out that the highest positive correlation with the target "Solar Energy" is Visibility by 0.468649 and the highest negative correlation with the Target "Solar Energy" is Cloud coverage by -0.687789 Therefore, If we have a high visibility we could have high solar power unless if we have a high cloud coverage.

Models

Linear regression, k-nearest neighbors, linear regression as the model with highest R squared 49 %. K-nearest neighbors found that the best number of neighbors is 18 neighbors with R squared 42%.

Model Evaluation and Selection

The entire training dataset of 637 records was split into 80/20 train vs. holdout

Training:

The highest R squared is 0.65
The RMSE (root mean squared error) on the training data is 6392.41

Testing:

The highest R squared is 0.49

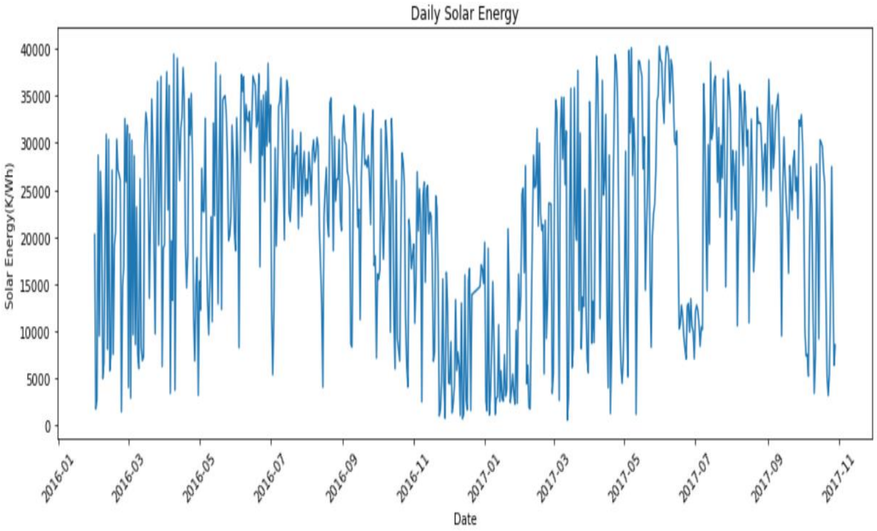
The RMSE (root mean squared error) on the test data is 7429.17

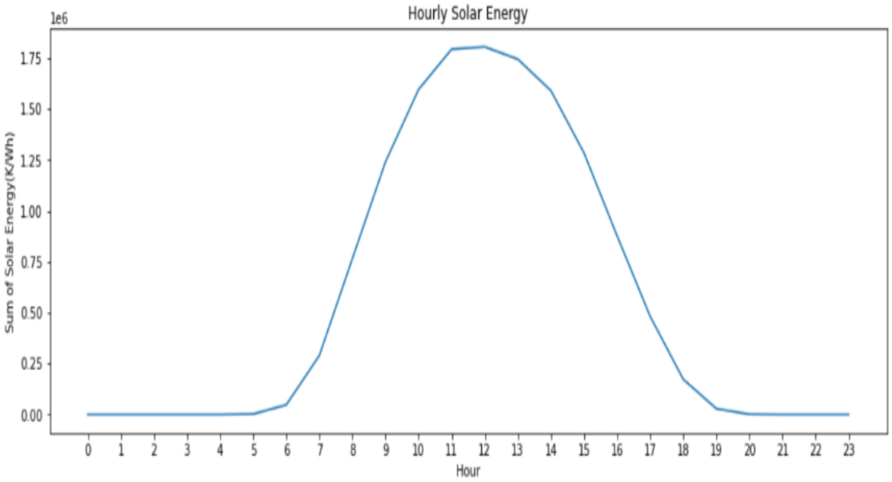
Tools

The main tool I have used is Python and Jupyter Notebook, which include

- Numpy and Pandas for data manipulation
- Scikit-learn for modeling
- Matplotlib and Seaborn for plotting

Communicate


Daily Solar Energy
Solar Energy(K/Wh)
Date


Hourly Solar Energy
1e6
Sum of Solar Energy(K/Wh)
Hour