Predicting the Solar Energy of Illinois

Goal

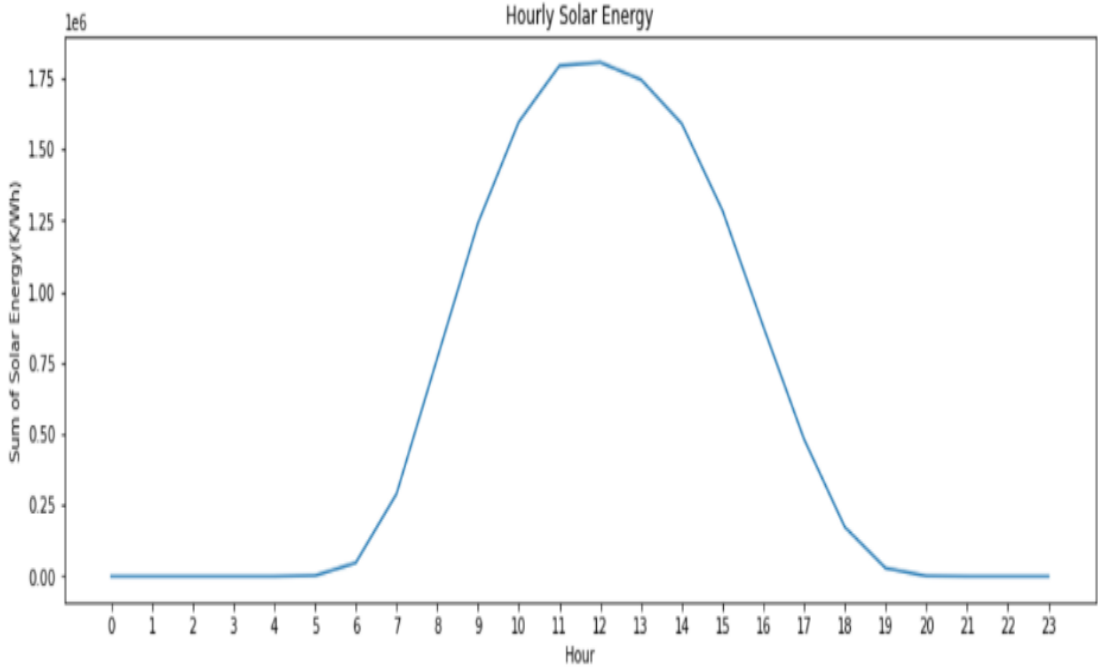
- **Goal:** The goal of this project was to use Regression models to predict the solar energy amount from a solar power plant at the University of Illinois in order to help improve grid management and maintenance. I worked with data provided by NOAA and University of Illinois, leveraging numerical feature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

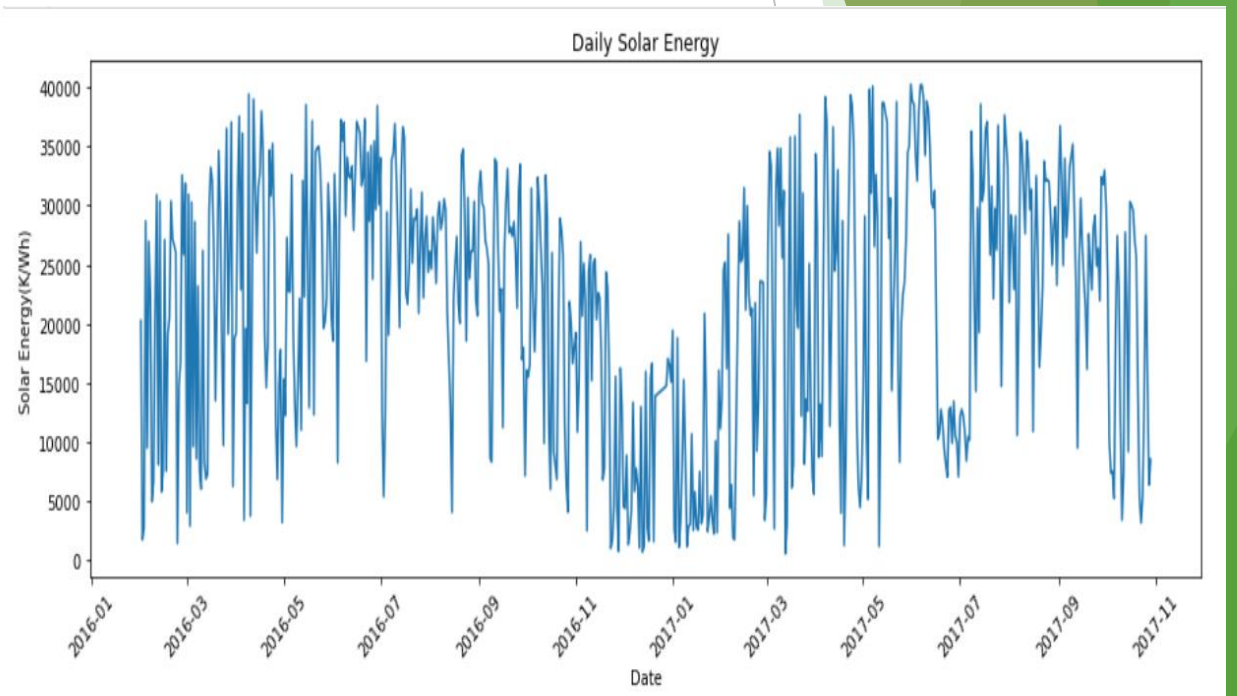
Data sets

- 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

Weather features	Unit	Weather features	Unit
Cloud coverage	% range	Relative humidity	%
Visibility	Miles	Wind speed	Mph
Temperature	°C	Station pressure	inchHg
Dew point	°C	Altimeter	inchHg

Solar Energy hourly and daily





Algorithms

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 0.49The RMSE (root mean squared error) on the test data is 7429.17

- The figure here shows that True or Actual values vs Predicted values and it can be improved for best result of prediction.