1. Introduction

NASA Space Apps Challenge hackathon which was held in 2017 asked participants to predict the amount solar radiation using a set of measurable meteorological conditions. The NASA HI-SEAS(Hawaii Space Exploration Analog and Simulation) site where provided the solar radiation dataset simulates a human settlement on Mars. The main power source in the settlement is obtained from a large solar array and battery bank. The prediction of solar radiation will help to decide when or where to deploy solar energy harvesting equipment. This can be found in Kaggle: https://www.kaggle.com/dronio/SolarEnergy

2. About this dataset

The meteorological data from the dataset is from the HI-SEAS station from September through December 2016. The size of the dataset is 2.82 MB and 11 columns. It consists of date and time and numerical formats.

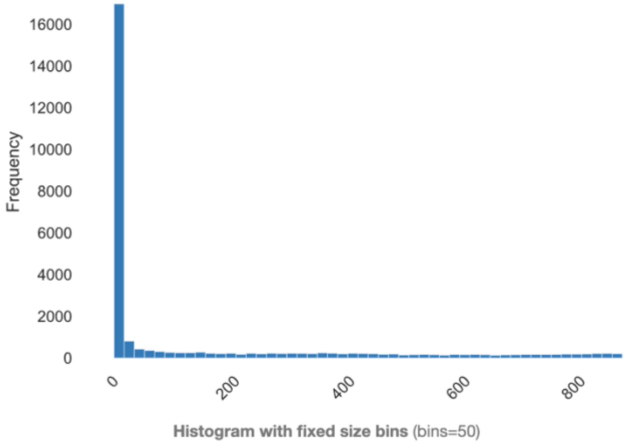
The units and formats of each dataset are:

- UNIX time_t date: seconds since Jan 1, 197
- Date: yyyy-mm-dd format
- Local time of day in hh:mm:ss: 24-hour format
- Solar radiation: watts per meter^2
- Temperature: degrees Fahrenheit
- Humidity: percent
- Barometric pressure: Hg
- Wind direction: degrees
- Wind speed: miles per hour
- Sunrise/sunset: Hawaii time

3. Data Exploration and Wrangling

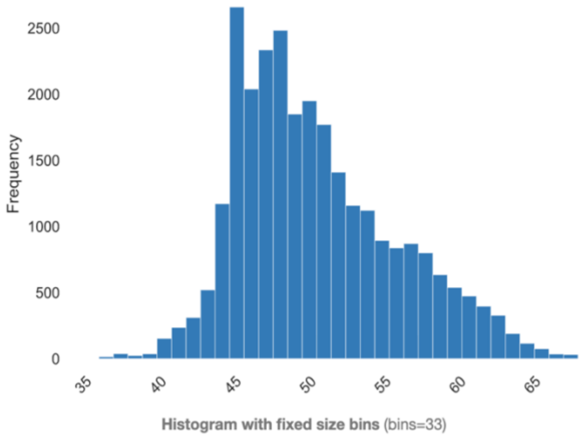
I used pandas profiling to see the distribution of the data and python for data wrangling.

The mean of radiation variable which is a target variable has 148.94m^2 and standard deviation was 249.197. According to the histogram of it, most of the data points were distributed from 0 to 200.

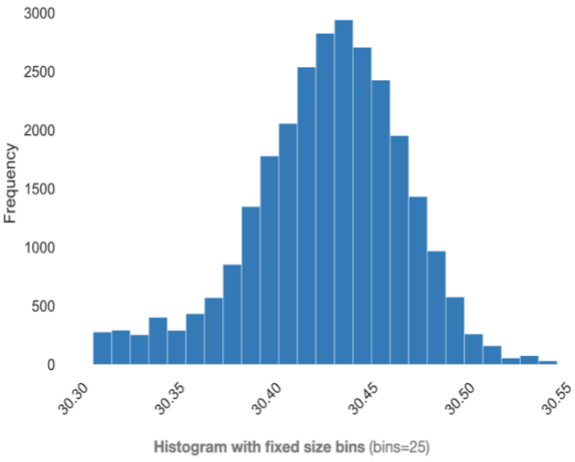


To reduce it skewness and large standardization, I used interquartile range and opted out extreme values.

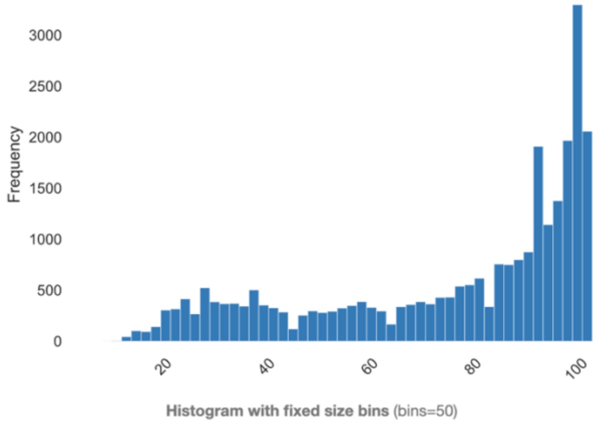
The rest of the meteorological variables which were distributed like below:



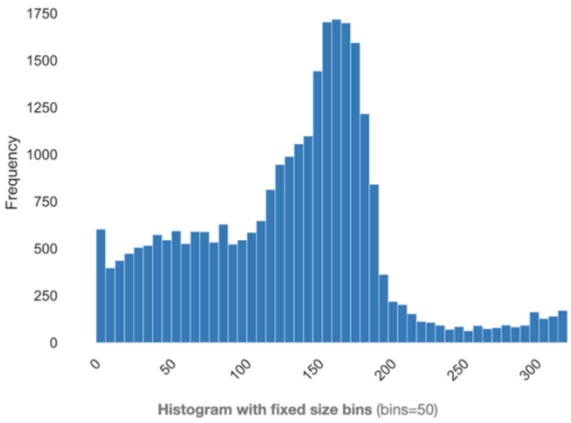
Temperature



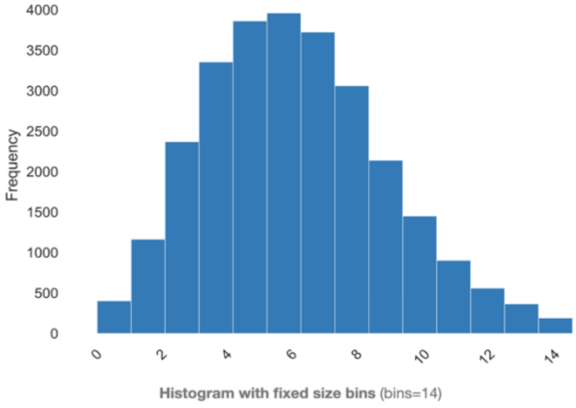
Pressure



Humidity



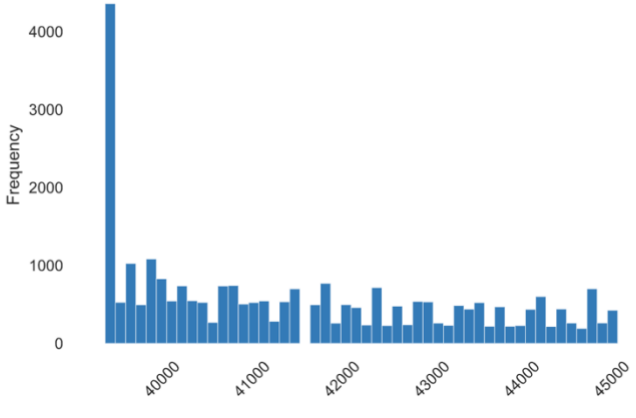
Wind Direction(Degrees)



(Wind) Speed

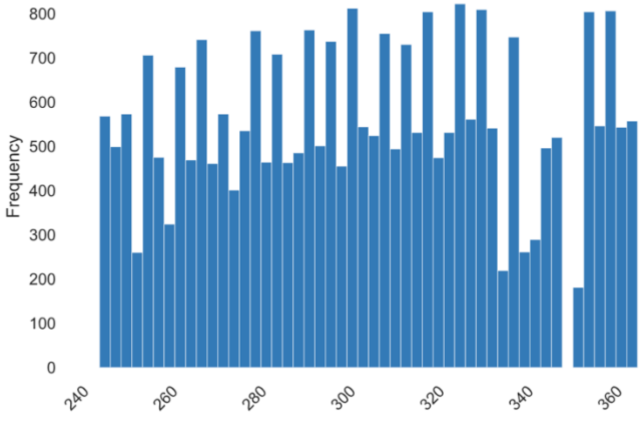
Instead of using the raw date and time variables, a feature engineering was conducted, and 2 variables were created: Length of day and day of year.

Using the sunrise and sunset variables, I created a length of day which is the time period between sunrise and sunset with the seconds unit. The mean of this feature was 41495 seconds which were around 11hrs 52 mins and minimum and maximum are around 10hrs 93mins (39360secs) and 12hrs 51mins (45060secs) respectively.



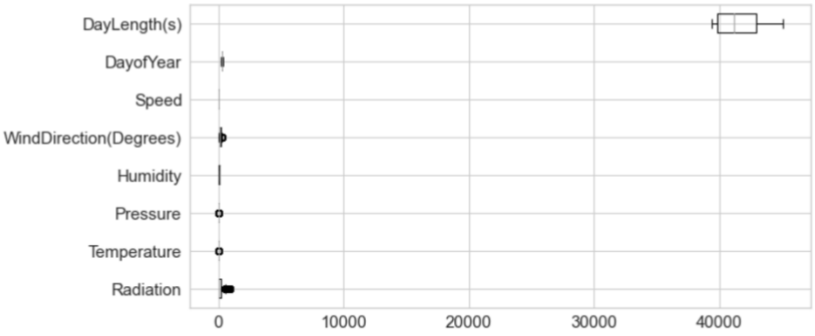
Histogram with fixed size bins (bins=50)

Usually summer has the longer length of a day than winter that we can find out if the solar radiation is getting decreased when it comes to close to winter. In addition, the day of year which is also related to seasons was calculated using the date variable. Since the original dataset was collected from Sept to December, the mean of this feature is 304 days. The minimum and maximum days were 245 and 366 days respectively.



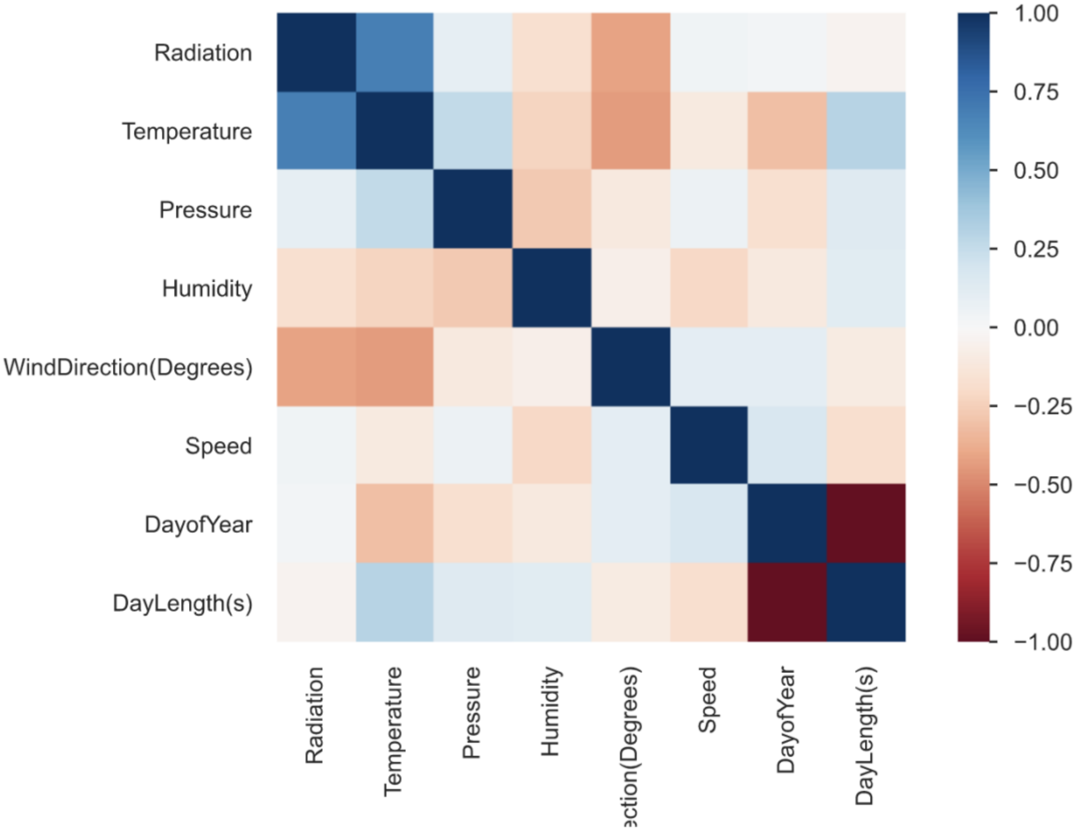
Histogram with fixed size bins (bins=50)

I used boxplots to see the outliers in the dataset and removed them using interquartile range.



The above figure is representing the distribution of the datasets after removing outliers.

Before moving on to the data modeling, I checked the correlation between each variable.



According to this correlation matrix, temperature has positive relationship with radiation which means that high temperature will increase the amount of solar radiation. The rest of the variables have subtle relationship with the radiation except DayLength and DayofYear that I removed the DayofYear variable from the dataset.

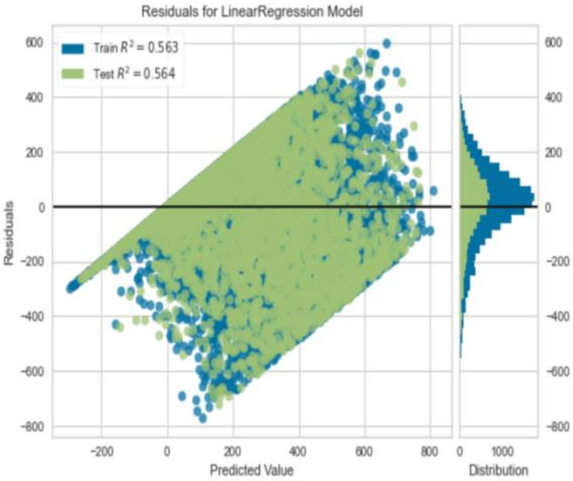
4. Data Modeling

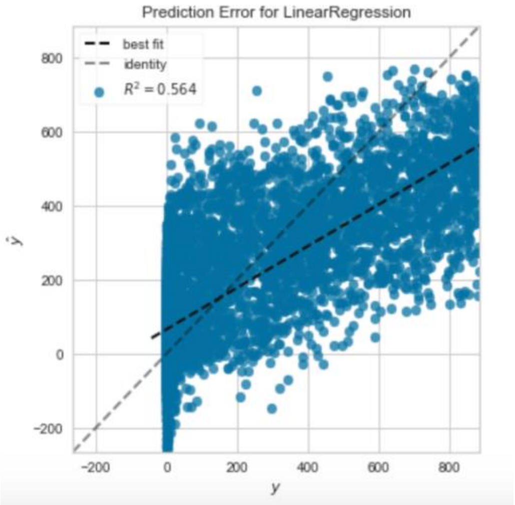
Linear Regression, Lasso, Random Forest and Extra Trees models were used to predict the solar radiation and the scores of MSE and r-squared were utilized to evaluate models.

The target variable is radiation and explanatory variables are temperature, pressure, humidity, wind direction, speed and daylength. All four models have same target variable and explanatory variables. The 30 percent of processed datasets are a test set and the rest of them are a train set. I normalized to X variables to integrate the units of each feature and standardized the X_train and X_test set using scikit-learn. After training the data to a model, I used MSE(mean squared error) and R squared for evaluating the model. MSE measures the average of squared of errors and R squared represents the proportion of variance for a dependent variable that can be contributed to the independent variable.

1) Linear Regression

The value of MSE train and test were around 27115.69 and 27140.58 which were big values. Thus, we could expect the underfitting issue that the model didn't perform well to predict the target variable. The R_squared was 0.5647 which was not really bad.





The left plot is a residual plot that some residuals were distributed away from 0. The right plot represents the prediction error for linear regression that the data points were not really at the x=y line.

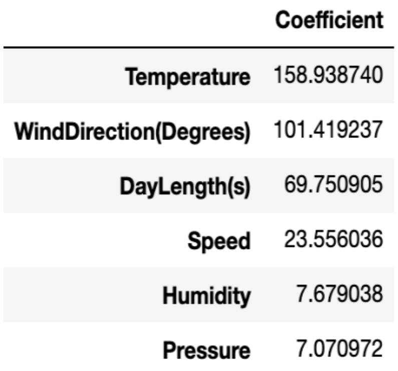
I conducted the hypothesis test to see whether the dependent variables were contributed to the independent variable or not.

Linear regression model: Y = B0+B1x1+B2x2+B3x3+ B4x4+B5x5+B6x6

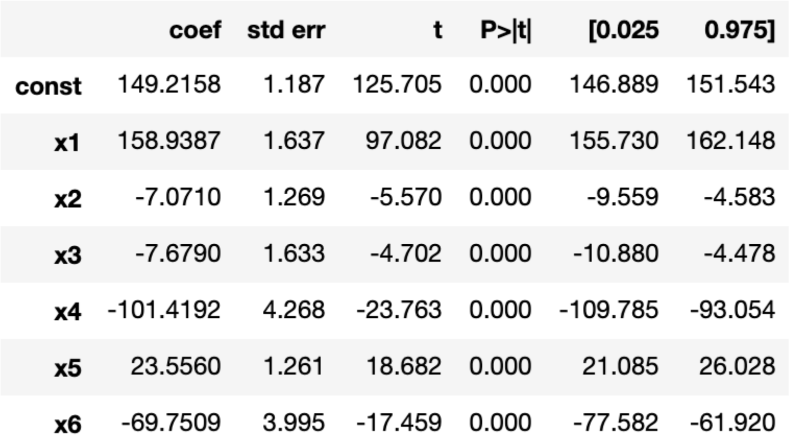
H0: b1=b2=b3=b4=b5=b6=0

H1: At least one of them is not 0

* b1, b2,…b6 were the regression coefficients from dependent variables



	Coefficient
Temperature	158.938740
WindDirection(Degrees)	101.419237
DayLength(s)	69.750905
Speed	23.556036
Humidity	7.679038
Pressure	7.070972

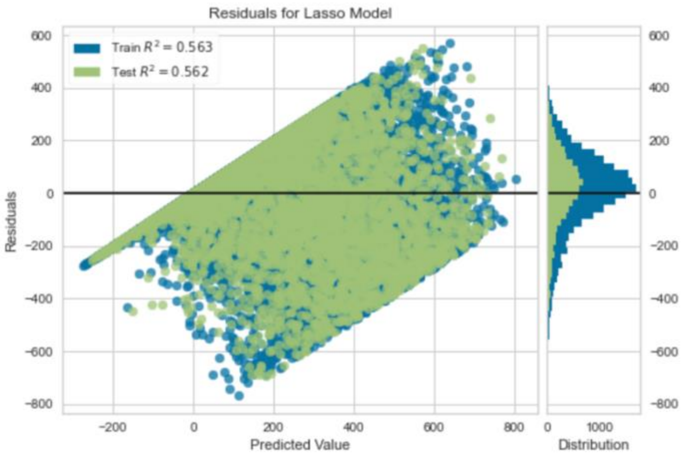


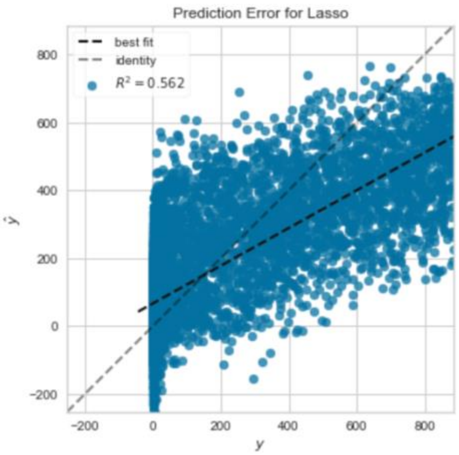
	coef	std err	t	P>|t|	[0.025	0.975]
const	149.2158	1.187	125.705	0.000	146.889	151.543
x1	158.9387	1.637	97.082	0.000	155.730	162.148
x2	-7.0710	1.269	-5.570	0.000	-9.559	-4.583
x3	-7.6790	1.633	-4.702	0.000	-10.880	-4.478
x4	-101.4192	4.268	-23.763	0.000	-109.785	-93.054
x5	23.5560	1.261	18.682	0.000	21.085	26.028
x6	-69.7509	3.995	-17.459	0.000	-77.582	-61.920

According the hypothesis test, the p-values of all variables were less than 0.05 that we can the null hypothesis. Therefore, we can say that they have significant affect to solar radiation. Especially temperature has the highest values which mean if temperature increases 1 unit, radiation will increase around 158.94 m^2. Relatively humidity and pressure have less impact on increasing the radiation.

2) Lasso regression

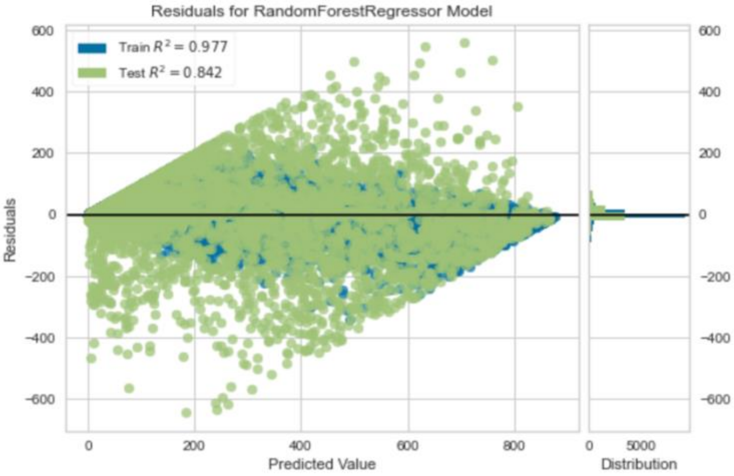
To reduce the MSE and increase the R squared values, I used lasso regression that regularized the linear regression. However, the results were similar with linear regression that the MSE test was 27257.7288, MSE train was 27146.2822 and R squared was 0.5618.

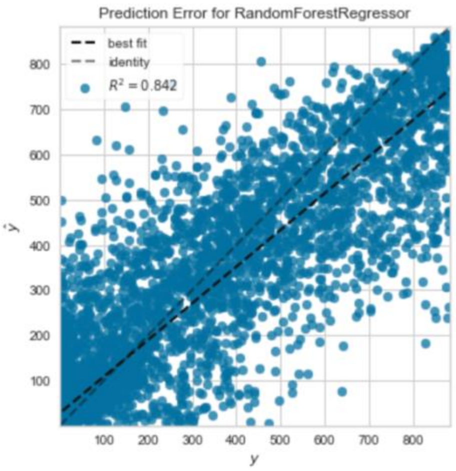




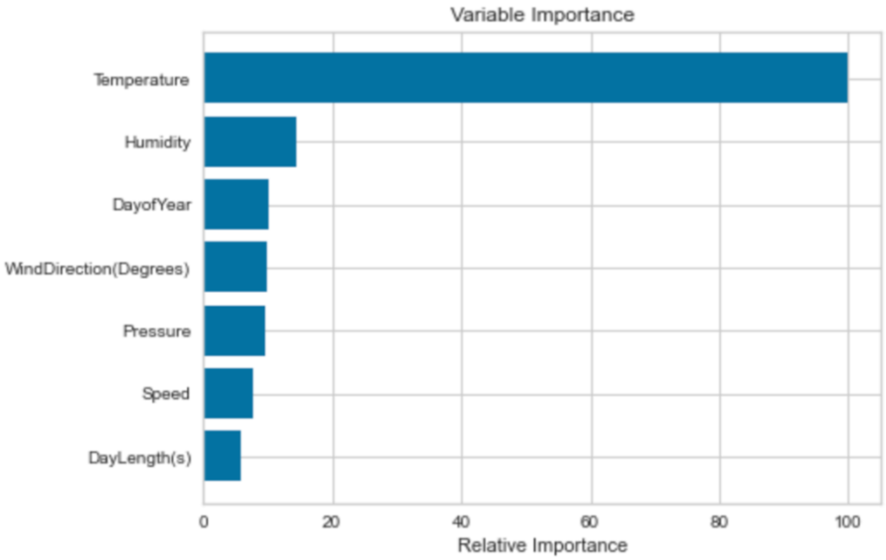
3) Random Forest

The random forest model had better results than the previous models that MSE train was 1366.426, MSE test was 9709.6452 and the R squared was 0.8439. The overall MSE values were lower and r squared was bigger than the linear regression model, however, there was huge gap between MSE train and MSE test which we had to suspect an overfitting issue.





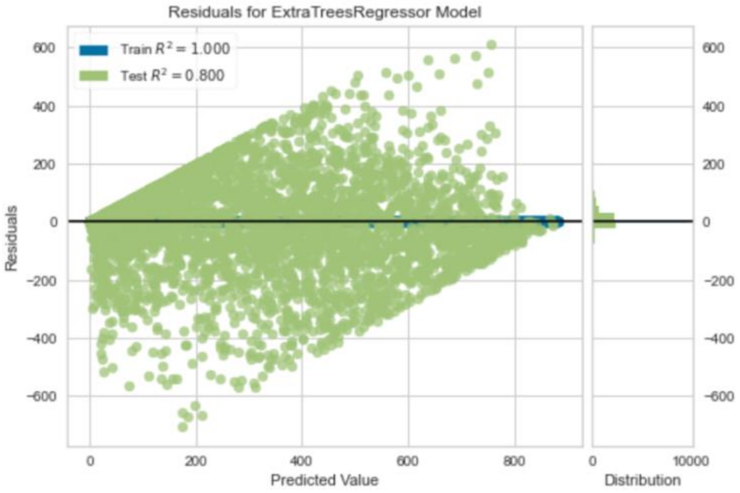
As you can see in these residuals and prediction error plots, the train dataset had less errors than test dataset.

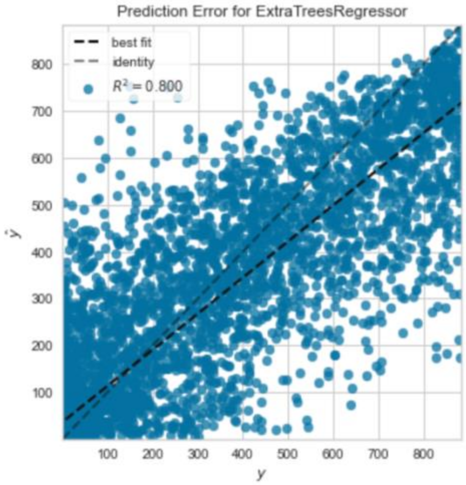


This histogram represents the importance of each variable. As similar as the linear regression model, temperature has the great impact on radiation. The least impact on it is day of length.

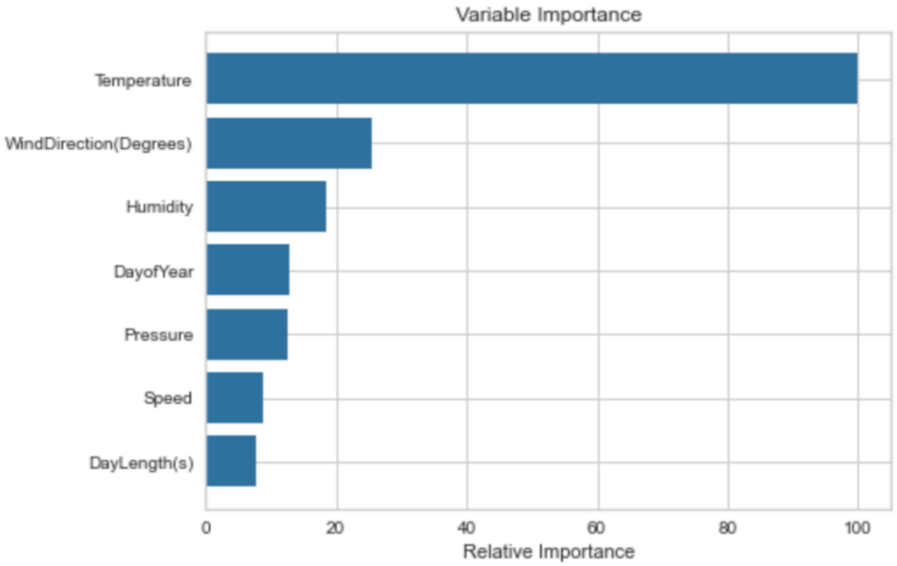
4) Extra Trees model

The last model I used was an extra trees model. The model performed little bit than the random forest model that MSE train was 0.5682, MSE test was 9507.7716 and R squared was 0.8472. However, there is big difference between MSE train and test that we have to suspect about the overfitting issue.





According to the residuals and prediction error plots, the errors were close to 0 and at the x=y line but still there were difference between train and test set.



In the variable importance histogram, temperature has the biggest impact on predicting the solar radiation and day of length has the least impact on it.

5. Model Summary and future study

Model	MSE Train	MSE Test	R2
Linear regression	27115.69	27140.58	0.5637
Lasso	27146.2822	27257.7288	0.5618
Random Forest	1408.84	9844.9473	0.8439
Extra Trees	0.5682	9507.7716	0.8472

For predicting the amount of solar radiation, I recommend using Extra Trees which was the least MSE values and the biggest R squared value among the other models. According to the model summary of extra trees model, temperature, wind direction and humidity were the top three dependent variables which can attribute to radiation. The random forest model had the same results and linear regression had high coefficients of temperature and wind direction and a positive value of humidity. Therefore, HI-SEAS team can consider these three main features when they decide to install additional panels.

However, extra trees model had a risk of overfitting. For the further research, we can consider different machine learning models or hyperparameter tuning to reduce the overfitting.