Predicting Solar Radiation

Abstract

With the rise in demands for renewal energies and increased focus on finding an alternative energy sources to fossil fuels, many countries are spending huge budget on solar energies. Solar energies with no doubt offer many environmental advantages over fossil fuels for electricity generation, but the energy produced by them fluctuates with changing weather conditions. Solar energy producers and suppliers need to predict the correct forecast of solar energy production to balance supply and demand. One small error could cause a huge expense to the company as it can lead to borrowing emergency energies from neighboring utilities. The goal of this project is to discover which statistical and machine learning techniques supply the best short-term predictions of solar radiance. This analysis can also be used to predict whether the radiance level at a particular location is safe for the resident's health or not. With the meteorological dataset obtained from HI-SEAS weather station from four months (September through December 2016) between Mission IV and Mission V, weather parameters were examined. Variables surrounding the weather such as temperature, humidity, wind speed, atmospheric pressure, wind direction were examined in relationship to target variable solar radiance. Most key features were identified in relation to target variables. The standard scalar data was then split into different groups. The model was fit to different regressor models such as Linear Regressor, random forest regressor, XGradientboost regressor and Neural Network regressor. The models were then tested on both the training data for their fit and the testing data for their generalization. Key metrics including loss, mean absolute errors etc. were generated. Model was analyzed to improve the performance and to use the optimum parameters.

Background

Solar radiation is radiant energy emitted by the sun from a nuclear fusion reaction that creates electromagnetic energy. It is propagated in the form of electromagnetic waves. Solar radiation includes visible light that allows us to see, ultraviolet light that is invisible to eyes, infrared which is the main source of heat, radio waves, X-rays, and gamma rays. Radiation is one way to transfer heat. At Earth's average distance from the Sun (about 150 million kilometers), the average intensity of solar energy reaching the top of the atmosphere directly facing the Sun is about 1,360 watts per square meter, according to measurements made by the most NASA satellite missions. This energy is in the form of photons with UV photons having the highest energy (con-

sidered as ionization radiation) followed by visible light which have more energy than IR photons. Once the sun's energy reaches earth, it is intercepted first by the atmosphere. A small part of the sun's energy is directly absorbed, particularly by certain gases such as ozone and water vapor. Some of the sun's energy is reflected to space by clouds and the earth's surface. (1)

The angle of sunlight is greater in the Southern Hemisphere during the winter. During the June solstice, the opposite is true. The Northern Hemisphere receives the maximum intensity of the sun's rays, while the angle of sunlight decreases in the Southern Hemisphere.

Above the earth's atmosphere, solar radiation has an intensity of approximately 1380 watts per square meter (W/m2). This value is known as the Solar Constant. At our latitude, the value at the surface is approximately 1000 W/m2 on a cloudless day at solar noon in the summer months.

- **Problem Statement**

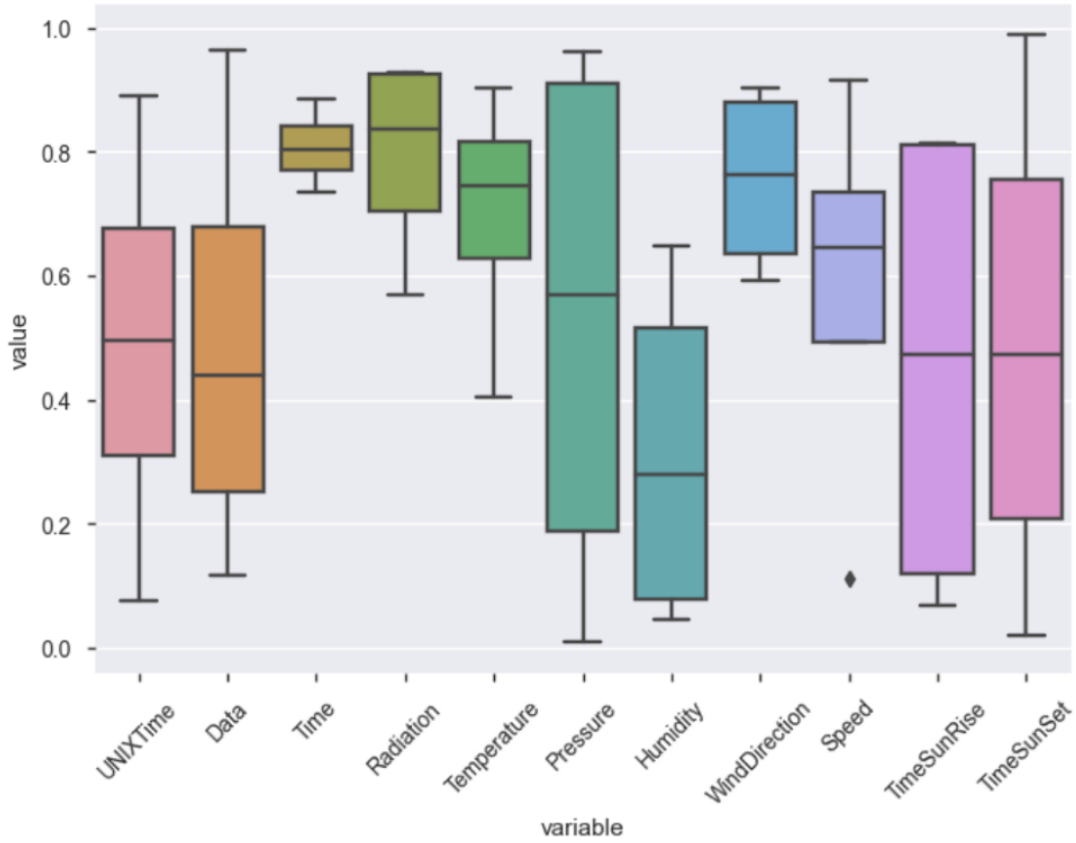
Although the solar radiation incident of the atmosphere of earth is almost constant but the radiation reaching the surface of earth vary widely based upon atmospheric effects like absorption and scattering or local variations of the location like latitude, clouds, pollution and even by the time of the day. This project is to analyze, understand and correlate commonly available weather data that is known in almost all the places these days, and predict the radiation data. The referenced weather data in the project include temperature, humidity, pressure, wind direction, speed of the wind relative to location of sun. The amount of energy reaching the surface of the Earth every hour is greater than the amount of energy used by the Earth's population over an entire year. Thus, considering these factors allows many solar radiation dependent devices and circumstances be planned to harness solar radiation to for many activities ranging from photosynthesis in plants, to creating electricity with photovoltaic (PV) cells, heating water and food, managing solar arrays as well as carefully managing human exposure to solar radiation Since these high intensity radiations can also cause life threatening health problems.

- **Data Exploration**

First step upon importing the dataset was to convert time and date parameters into a more useful format and add some columns that may be useful for visualization, modeling and analysis. Based on the information in dataset, a correlation study was conducted to see how strongly, or weakly variables were related and the magnitude of change in one variable with respect to target variable was calculated. We plotted bar charts to understand the mean value of variables.

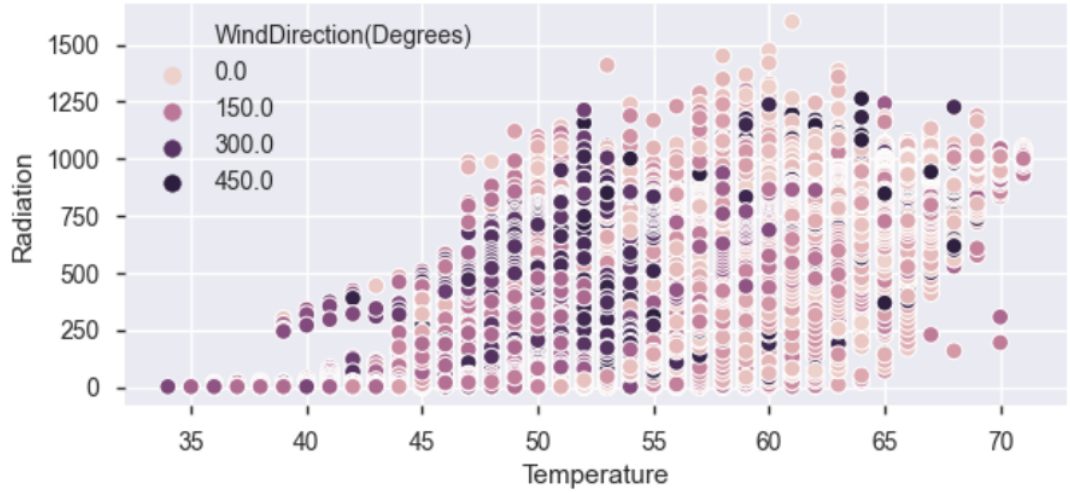
```
df.describe(include='all')
```

	UNIXTime	Data	Time	Radiation	Temperature	Pressure	Humidity	WindDirection(Degrees)	Speed
count	3.268600e+04	32686	32686	32686.000000	32686.000000	32686.000000	32686.000000	32686.000000	32686.000000
unique	NaN	118	8299	NaN	NaN	NaN	NaN	NaN	NaN
top	NaN	12/25/2016 12:00:00 AM	16:20:18	NaN	NaN	NaN	NaN	NaN	NaN
freq	NaN	288	24	NaN	NaN	NaN	NaN	NaN	NaN
mean	1.478047e+09	NaN	NaN	207.124697	51.103255	30.422879	75.016307	143.489821	6.243869
std	3.005037e+06	NaN	NaN	315.916387	6.201157	0.054673	25.990219	83.167500	3.490474
min	1.472724e+09	NaN	NaN	1.110000	34.000000	30.190000	8.000000	0.090000	0.000000
25%	1.475546e+09	NaN	NaN	1.230000	46.000000	30.400000	56.000000	82.227500	3.370000
50%	1.478026e+09	NaN	NaN	2.660000	50.000000	30.430000	85.000000	147.700000	5.620000
75%	1.480480e+09	NaN	NaN	354.235000	55.000000	30.460000	97.000000	179.310000	7.870000
max	1.483265e+09	NaN	NaN	1601.260000	71.000000	30.560000	103.000000	359.950000	40.500000

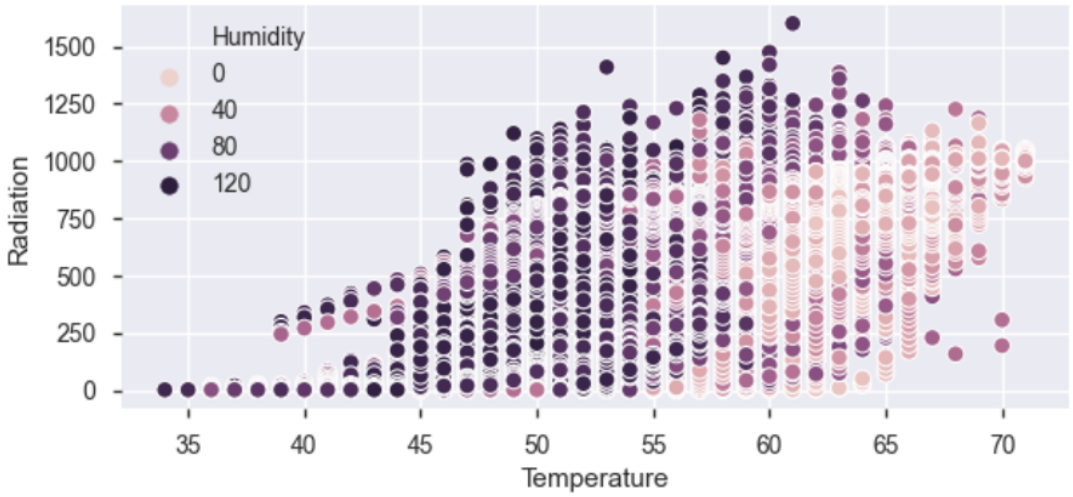


We plotted visualizations to study how was radiation levels varying with different weather parameters. Some of those visualizations were:

```
fig, ax = plt.subplots(figsize=(7,3))
sns.scatterplot(x='Temperature', y='Radiation', hue='WindDirection(Degrees)', data=df)
plt.show()
```




```
fig, ax = plt.subplots(figsize=(7,3))
sns.scatterplot(x='Temperature', y='Radiation', hue='Humidity', data=df)
plt.show()
```

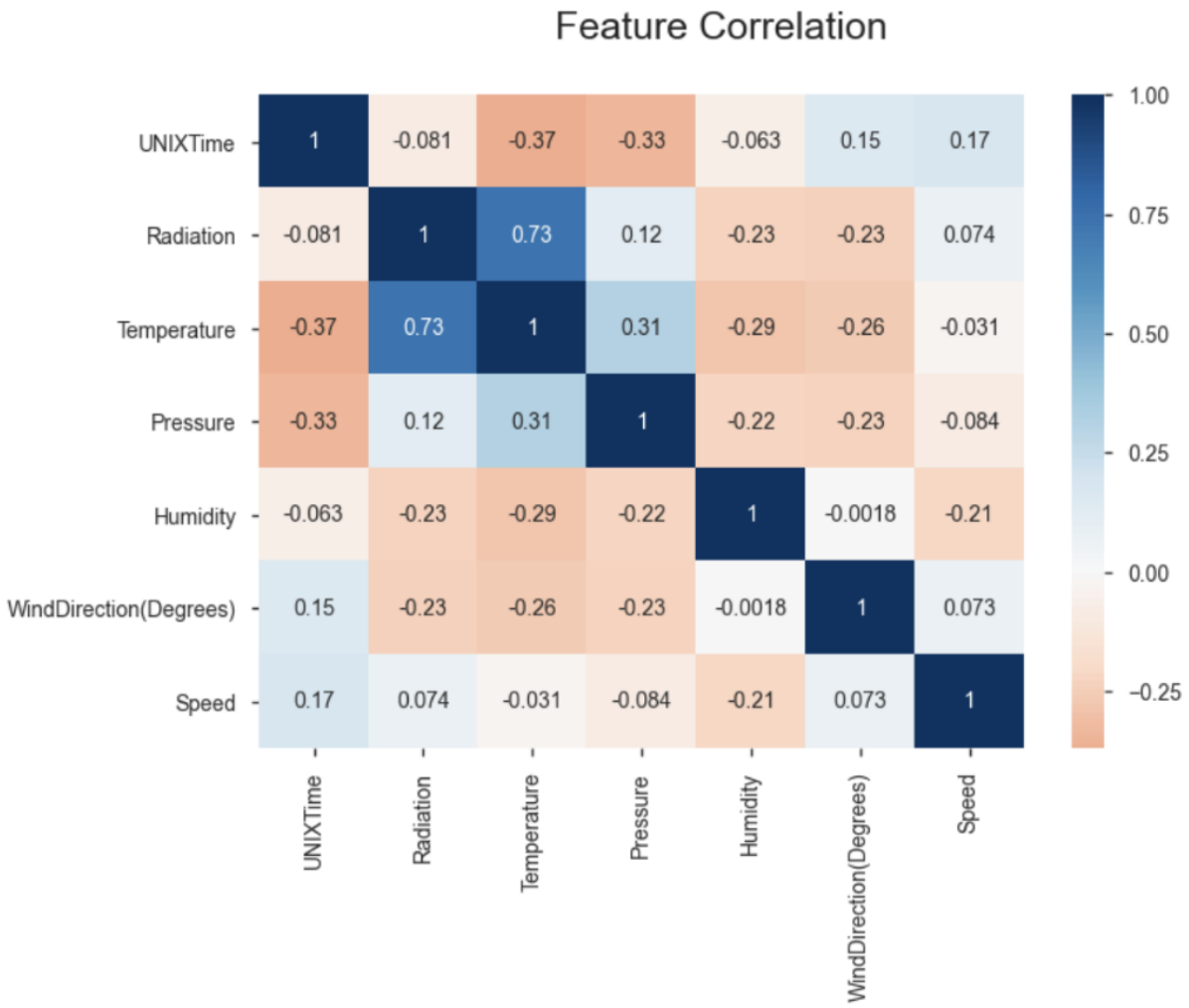


Solar radiation is positively correlated with temperature

1. Radiation vs Temperature -To confirm radiation levels were higher at higher temperatures.
2. Radiation vs Humidity - To confirm radiation levels were higher at low humidity's
3. Radiation vs Month- To see which month recorded highest radiation levels

4. Radiation vs Time of day- To find out that in the time range of 10AM – 3pm radiation levels were highest

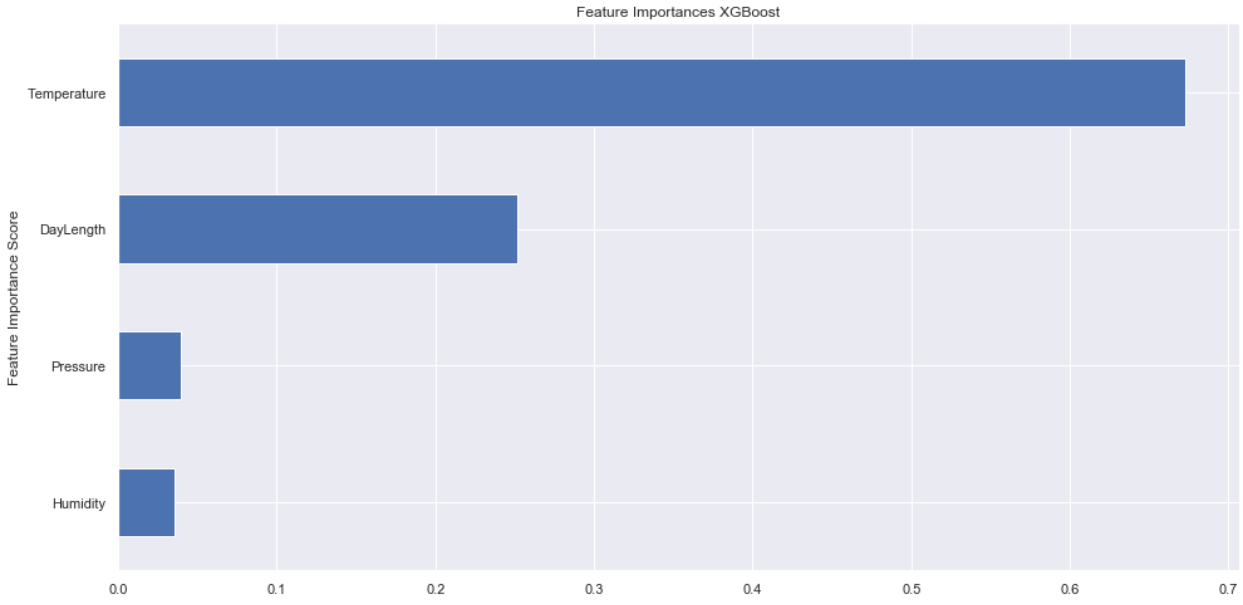
Below diagram shows the correlation plot of different input variables. Radiation levels have positive correlation with temperature and pressure while wind direction and speed seem to very weakly correlated to radiance.



- **Data Preparation**

Our data did not contain any missing values. Input dataset already had numerical values, so less effort was done to prepare the Data. Some of the variables like daylength (sunrise_time – sunset_time) and month were derived from input dataset variables. The dataset contained 11 variables and we derived 3 variables daylength, day of year and month from input variables.

There were no categorical variables in the dataset, so we did not need any special treatment for input variables. Below visualization shows the most important features for the model:



- **Methodology**

This is a regression problem to predict the levels of solar radiations, so we selected different regression models accordingly. Most of our input variables have numerical values, so they were fit to be used for models. For Random forest regression model, we performed feature reduction techniques like PCA to reduce the features. For XGradient boost regressor model we calculated the score for each feature variable to find out the most important features. The data was randomly split into training, testing, and validation datasets. Using Scikit-learn, models explored were Logistic Regression, Random Forest regression and Xtreme Gradient Boosting regression and neural network regression.

- **Regression techniques**

Linear regression is a linear approach to modeling the relationship between a scalar response e.g. solar radiation and one or more explanatory variables (also known as dependent and independent variables) such as weather parameters. The sum of the squared differences between the observed solar radiance predicted by a linear approximation of the forecast weather metrics is minimized by the least squares regression methods. We find a function of that provides ap-

propriate predictions to unseen patterns x′ applying the implemented regression techniques. In the following, we explain linear regression and the basic idea of SVR.

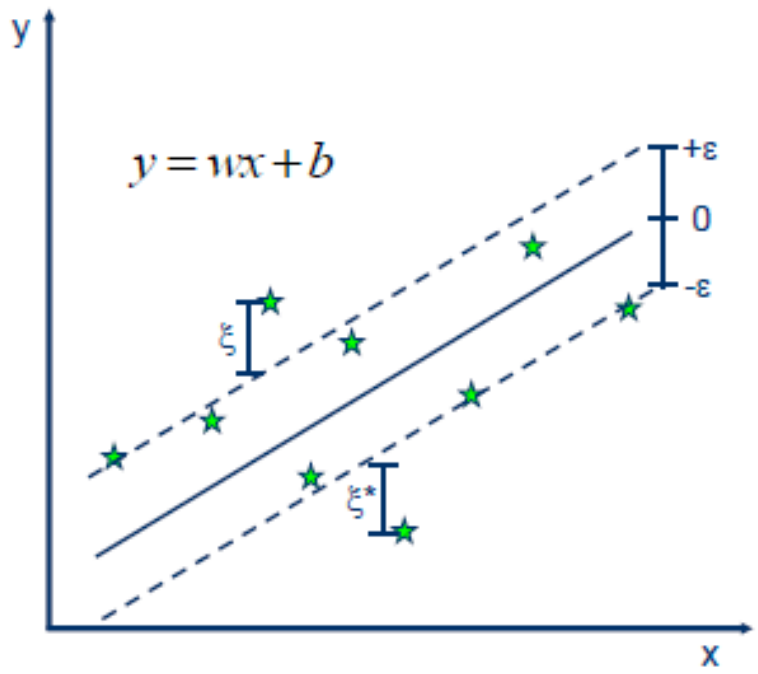
In linear regression model, Suppose that (x_1, y_1), (x_2,y_2), (x_n ,y_n) are realizations of the random variable pairs (X_1,Y_1), (X_2,Y_2) , (X_n,Y_n) . The simple linear regression equation is that the mean of Y is a straight-line function of x. We could write this as:

$$\hat{Y} = \hat{\beta}_0 + \sum_{j=1}^{p} X_j \hat{\beta}_j$$

where $\hat{Y}$ is used to represent the mean value (expected value).

- **Support Vector Regression**

In Support Vector Regression, a margin of tolerance (epsilon) is set in approximation to the SVM which would have already requested from problem.The main target is to minimize the error , individualizing the hyperplane which maximizes the margin keeping in mind the part of error is tolerated.



- Minimize:

$$\frac{1}{2}\|w\|^2 + C\sum_{i=1}^{N}\left(\xi_i + \xi_i^*\right)$$

- Constraints:

$$y_i - wx_i - b \le \varepsilon + \xi_i$$

$$wx_i + b - y_i \le \varepsilon + \xi_i^*$$

$$\xi_i, \xi_i^* \ge 0$$

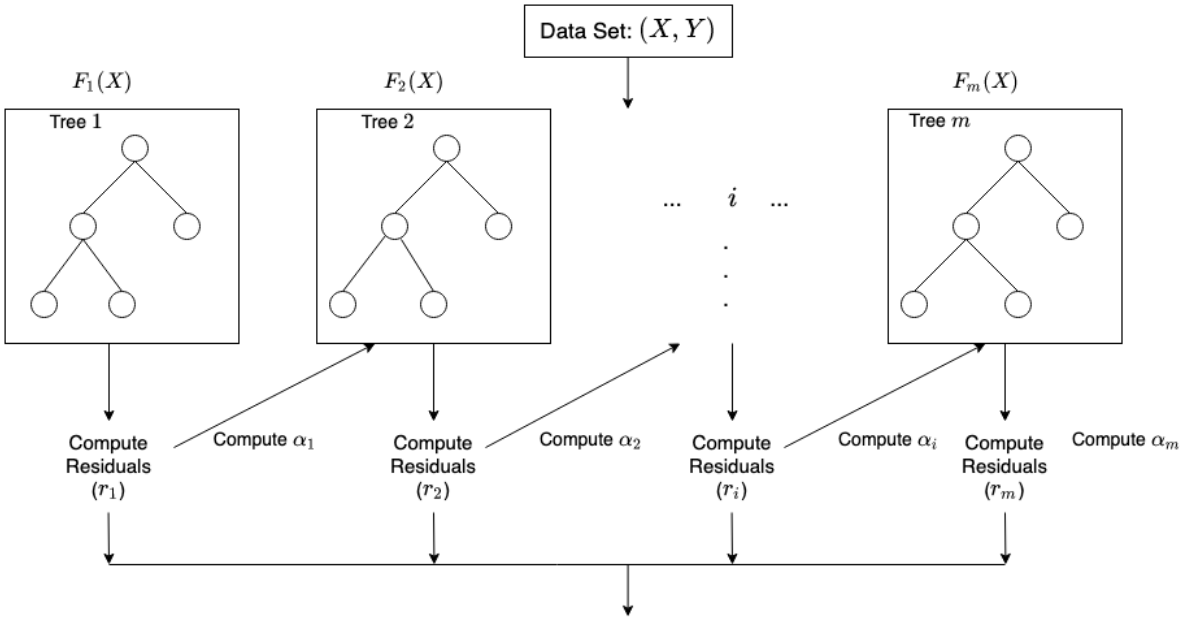
[1]

[1] https://www.saedsayad.com/support_vector_machine_reg.htm

where x_i is a training sample with target value y_i. The inner product plus intercept $(w, x_i) + b$ is the prediction for that sample, and ε is a free parameter that serves as a threshold: all predictions have to be within an ε range of the true predictions. Slack variables are usually added into the above to allow for errors and to allow approximation in the case the above problem is infeasible.

- **XGBoost Regression**

Gradient boosting is a supervised learning algorithm, which attempts to accurately predict a target variable by combining the estimates of a set of simpler, weaker models.
[2] When using XGBoost for regression, the weak learners are regression trees, and each regression tree maps an input data point to one of its leaf's that contains a continuous score. Boost minimizes a regularized (L1 and L2) objective function that combines a convex loss function based on the difference between the predicted and target output) and a penalty term for model complexity. The training proceeds iteratively, adding new trees that predict the residuals or errors of prior trees that are then combined with previous trees to make the final prediction. It is called gradient boosting because it uses a gradient descent algorithm to minimize the loss when adding new models.



[2] https://docs.aws.amazon.com/sagemaker/latest/dg/xgboost-HowItWorks.html

- **Measures of forecast error**

The model that gives the minimum measures of error will be our desired model for forecasting. Suppose for a while that y_i is the actual observation and it is the fitted value for the same time period. The following standard statistical measures can be defined if there are in time periods. The following statistical summary measures of a model's forecast accuracy are defined using the absolute error.

The Mean Absolute Error (MAE):

$$MAE = \frac{\sum_{t=1}^{n} |e_t|}{n}$$

The Root Mean Square Error (RMSE): =

$$RMSE = \sqrt{\frac{\sum_{t=1}^{n} e_t^2}{n}}$$

3

where e_t is the an arithmetic average of the absolute errors i.e. forecast error in time period that is :

$$e_t = \hat{y}_t - y_t$$

y_t is the actual value in the time period t.
n is the number of forecast observations in the estimation period.

The smaller values of MAE and RMSE, the better the model is considered to be.

- **Modeling**

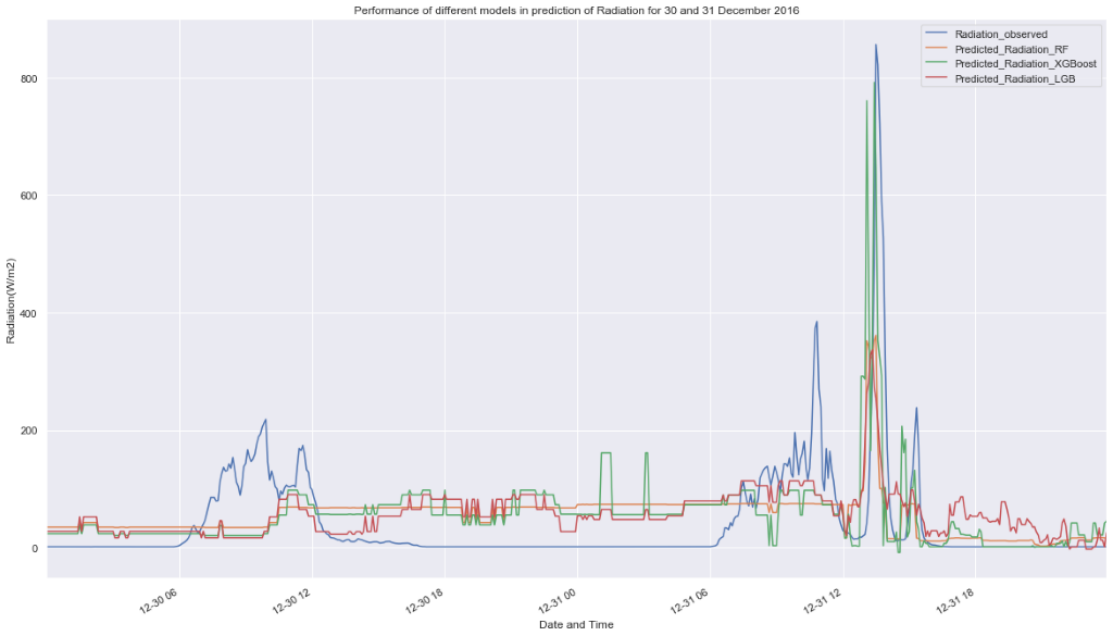
To train the algorithm, we implement a split train/test methodology to prevent bias in the learning. The dataset is split into a randomly sampled pool of data-points. 80% of those points are used for training, the remaining 20% is used for validation of the training data. So, the test data is not necessarily continuous time, but rather a random selection of points from the set. We had plenty of data to train with. We tried several models and compared their performance to evaluate the best algorithm to predict solar radiation. We trained data to fit on 5 different regression models Logistic Regression, XGradient Boost regression , Random Forest Regression, Support Vector Regression and Neural Network Regression model. To evaluate the models we

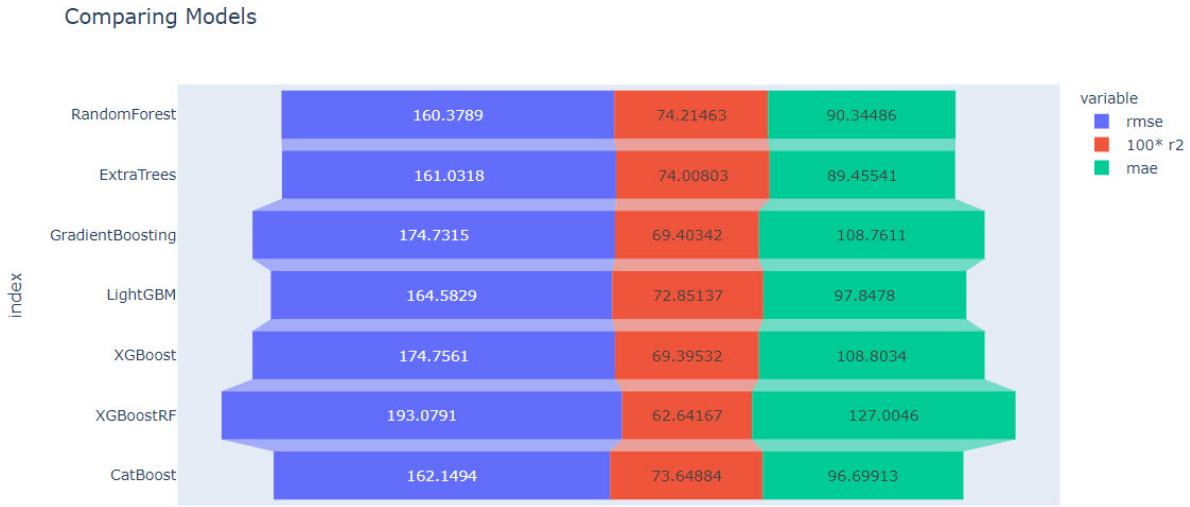
[3] https://www.researchgate.net/publication/329121238_A_Short_Term_Day-Ahead_Solar_Radiation_Prediction_Using_Machine_Learning_Techniques

calculated key metrics like RMSE (Root Mean Square Error), MAE (Mean Absolute Error) and R-Squared (Coefficient of Determination) to compare and see which model was best predictor. Also, we created graphs of predicted model value vs actual model values to see which model was the best fit. For neural network model loss and mean absolute error were plotted against number of Epochs(iterations) for test and train data.

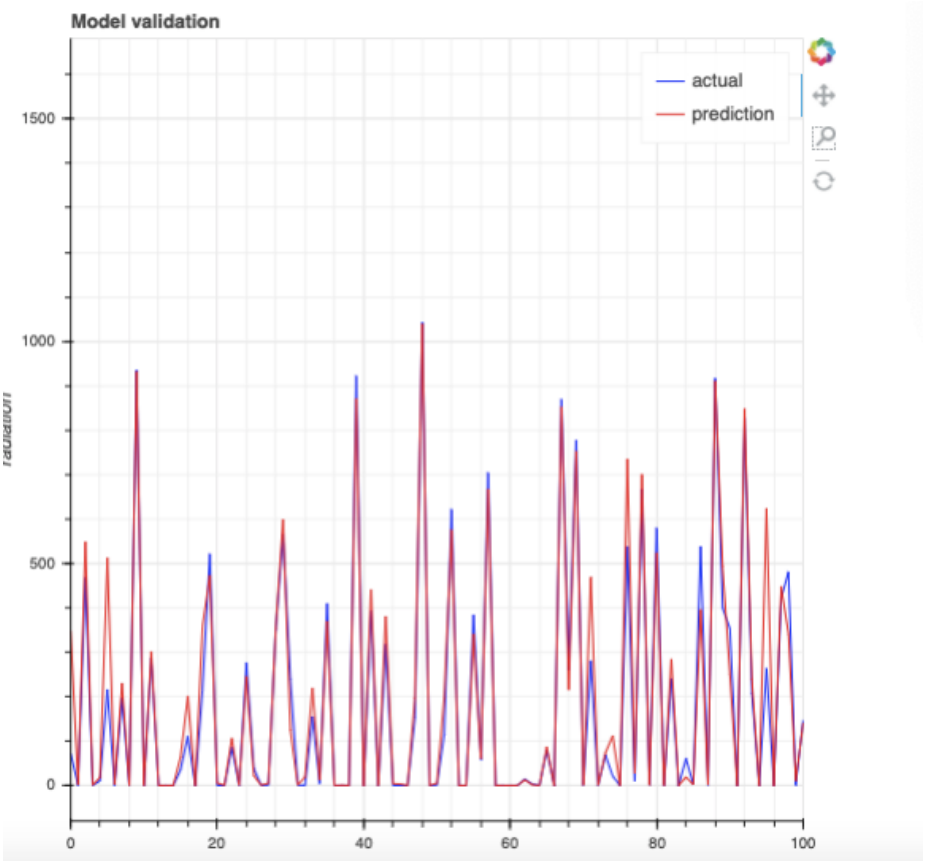
- **Results**

RMSE, R-Squared and MAE values were compared for all models, XGradient Boost seemed to be the best performing model without extracting best features with R-Squared value of 0.81.





However, after tuning the parameters and applying feature reduction for random forest regression model, we were able to achieve the highest matching values (93%) of predicted vs actual- And most of the predicted values feel close to actual values as shown in graph below.



Keras sequential neural network model proved to be 2nd best model with R-Squared value of 0.872.

Model Name	*rmse*	*100*r2*	*mae*
RandomForest	160.378	74.214	90.3444
ExtraTrees	161.031	74.008	89.455
GradientBoosting	174.731	69.403	108.761
LightGBM	164.582	72.851	97.847
XGBoost	174.756	69.395	108.803
XGBoostRF	193.079	62.641	127.004
CatBoost	162.149	73.648	96.699

Conclusion

At this stage three models (Random forest regression, XGradient Boost and Neural Network regression) have proved to be best models for solar radiations prediction with R-Squared value of 0.81 for both training and validation data. The next step may be to apply the model to a new dataset and apply these 3 models again on that dataset to find out the best predictor. We will further need to research and optimize hyper-parameters for each model, test the new models on dataset, and evaluate. The neural network model will need to be optimized for number of epochs(iterations) as well as we will need to change number of hidden layers to see if we can further improve the performance. Some of effects of cloud coverage and rain is ignored directly, they may still be indirectly playing role through temperature and humidity.

Acknowledgements

We want to acknowledge The National Aeronautics and Space Administration for providing us the meteorological data from the HI-SEAS weather station and Kaggle.com from where we downloaded the data for the project. We also want to thank the Philip Linden whose GitHub code repository was used as a reference and his preliminary analysis on the dataset provided us great insights and helped us come up with ideas to complete the preliminary analysis. We would like to thank professor Brett Werner for his support and guidance to complete this project and we would also like to acknowledge the author David Abbott of book 'Applied Predictive Analytics: Principles and Techniques for the Professional Data Analyst' which has given us solid foundations about predictive analytics concepts, principals and methodologies.

References

1. *https://www.kaggle.com/dronio/SolarEnergy*
2. *https://docs.h2o.ai/h2o/latest-stable/h2o-py/docs/_modules/h2o/estimators/random_forest.html*
3. *https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.695.8578&rep=rep1&type=pdf*
4. *https://www.diva-portal.org/smash/get/diva2:1387928/FULLTEXT01.pdf*
5. *https://www.hindawi.com/journals/ijp/2012/946890/*
6. *https://github.com/runphilrun/kaggle-radiation-prediction*
7. *https://docs.aws.amazon.com/sagemaker/latest/dg/xgboost-HowItWorks.html*
8. *https://en.wikipedia.org/wiki/Linear_regression*
9. *https://en.wikipedia.org/wiki/Support-vector_machine*
10. *https://www.researchgate.net/publication/329121238_A_Short_Term_Day-Ahead_Solar_Radiation_Prediction_Using_Machine_Learning_Techniques*
11. *https://mitpress.mit.edu/sites/default/files/titles/content/boosting_foundations_algorithms/chapter007.html*
12. *https://www.saedsayad.com/support_vector_machine_reg.htm*