**Interim Project Report**

**On**

**Solar Irradiance Prediction**

****

Submitted in the partial fulfillment of the requirement for the award of degree

**Post graduate program in Data Science Engineering**

**Submitted By –**

Vikas Katoch

Vinit Kumar

Kajal Sharma

Rishabh Bhardwaj

**TABLE OF CONTENT**

|  |  |  |
| --- | --- | --- |
| **Chapter no.** | **Particulars** | **Page no.** |
| **1** | **Introduction** |  |
| **2** | **Dataset Description**  **&**  **Exploratory Data Analysis** |  |
| **3** | **Modelling** |  |

**Solar Irradiance Prediction**

# Submitted to Great Lakes Institute of Management Gurugram

In partial fulfillment of the requirements for the award of degree

**Post graduate program in Data Science Engineering**

## Submitted by

**Group Number - 7**

|  |  |
| --- | --- |
| **Name** | **Signature** |
| Vinit Kumar |  |
| Vikas Katoch |  |
| Rishabh Bhardwaj |  |
| Kajal Sharma |  |

**Mentor**

**Animesh Tiwari**

**Abstract**

Solar radiation is an essential source of energy that has yet to be fully utilized. This energy can be converted into another form of more usable energy, electricity, by using photovoltaic power generation systems in order to fight against global warming. When the photovoltaic power generation systems are connected to an electrical grid, predicting near-future global solar radiation is important to stabilize the entire network. Global solar radiation (GSR) is an essential parameter for the design and operation of solar energy systems. Long-standing records of global solar radiation data are not available in many places because of the cost and maintenance of the measuring instruments. The major objective of this work is to develop a Machine learning model for accurately predicting solar radiation. Meteorological data collected for the last 4 months from September to December have been used to train the models. The best machine learning model are identified based on minimum mean absolute error (MAE) and root mean square error (RMSE) and maximum linear correlation coefficient (R). Further, the present study confirms that prediction accuracy of the regression model depends on the complete set of data being used for training the network for the intended application. The developed regression model has a low mean absolute percentage error (MAPE) which ascertains the accuracy and suitability of the model to predict the monthly average global radiation so as to design or evaluate solar energy installations.

**SECTION - 1**

**INTRODUCTION**

1. **Introduction**

Solar irradiance is the power per unit area received from the Sun in the form of electromagnetic radiation as reported in the wavelength range of the measuring instrument. Solar resource forecasting is very important for the operation and management of solar power plants. Solar radiation is highly variable because it is driven mainly by local weather patterns. This high variability presents challenges to meeting power production and demand curves, notably in the case of photovoltaic (PV) power plants, which have little or no storage capacity. However, temporally and spatially varying irradiance introduces thermal stress in critical system components and plant management issues that can result in the degradation of the overall system’s performance and reduction of the plant’s lifetime. The variability can also result in lower plant efficiencies compared to operation in stable conditions because optimally operating the plant is more challenging. For PV power plants that have battery storage, forecasts are helpful to schedule the charging process of the batteries at the most appropriate time, optimize the fractions of electricity delivered and stored at any instant, and thus avoid the loss of usable energy. Solar radiation forecasting anticipates the solar radiation transients and the power production of solar energy systems, allowing for the setup of contingency mechanisms to mitigate any deviation from the required production. With the expected integration of large shares of solar power, reliable predictions of solar power production are becoming increasingly important as a basis for efficient management and operation strategies as well as for solar energy trading. Today, solar power prediction systems are an essential part of electric grid management in countries that have substantial shares of solar power generation. Since Machine learning is one of the newest approaches to this challenge and aims to bring major changes to short-term solar forecasting.

* 1. **Need For Data Science**

Worldwide demand for a reliable and sustainable supply of renewable energy, including solar, is growing. Accurate estimates of solar energy production and insights into solar equipment performance help solar plant owners and operators optimize inspections, schedule maintenance, improve the operational performance of their equipment, and maximize the environmental benefit of their investments in renewable energy. However, due to the uncertainties inherent in the unpredictable nature of this renewable resource, many challenges are associated with estimation of solar power production and detection of performance issues. In this study, our goal is to explore how predictions of solar radiation can be improved by applying data science techniques, and how machine learning models can be applied to correctly predict the solar radiation. Our results show that regional weather data can be used to estimate (and potentially predict) solar energy production for some applications; that a hybrid machine learning model based on historical data, temperature, and information from physical models outperforms predictions from state-of-the-art physical models; and that environmental factors such as lightning and ambient temperature, as well as grid operating conditions, can influence device reliability.

**SECTION-2**

**DATASET DESCRIPTION**

**AND**

**EXPLORATORY DATA ANALYSIS**

This chapter gives the insight of the dataset such as description of the dataset with all features and also through Exploratory Data Analysis which makes it easy to understand about the dataset.

**2.1 Solar Irradiance Dataset**

NASA HI-SEAS missions act as a testbed and training ground for humans as we develop the capability to explore Mars. A recent NASA Space Apps Challenge hackathon asked participants to use data collected from the HI-SEAS site to predict solar radiation given a set of measurable meteorological conditions. Knowing when conditions are most favorable for incident solar radiation is crucial for deciding when and where to deploy solar energy harvesting equipment, especially for colonists or astronauts on the surface of Mars.

These datasets are meteorological data from the HI-SEAS weather station from four months (September through December 2016) between Mission IV and Mission V.

For each dataset, the fields are:

* A row number (1-n) is useful in sorting this export's results
* The UNIX time, date (seconds since Jan 1, 1970). Useful in sorting this export's results with other export's results.
* The date in YYYY-MM-DD format. The local time of day in HH:MM:SS 24-hour format.
* The numeric data, if any (may be an empty string)
* The text data, if any (may be an empty string)

The units of each dataset are:

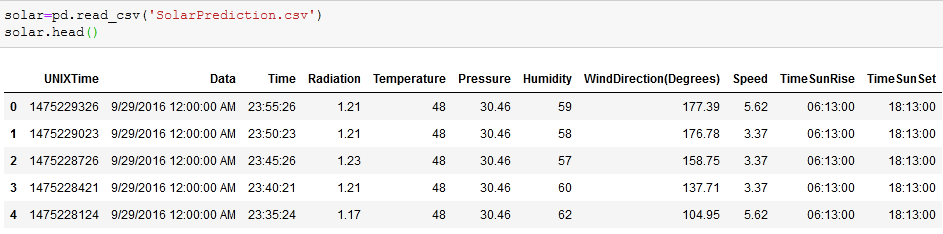
* Solar radiation: watts per meter^2
* Temperature: degrees Fahrenheit
* Humidity: percent
* Barometric pressure: Hg
* Wind direction: degrees
* Wind speed: miles per hour
* Sunrise/sunset: Hawaii time

**2.2 Exploratory Data Analysis**

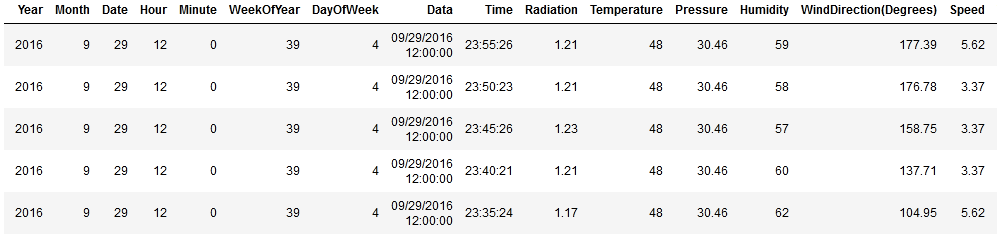
**Understanding the Data**

Shape of the Data

The Dataset has 32686 rows and 12 columns.



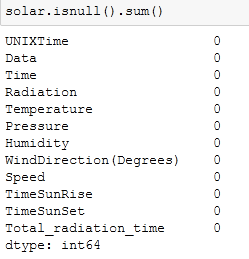
This fig shows, the data column has Date and Time in a single column which needs to be splitted into Date, Month & Year and Time needs to be splitted into Hours, Minutes & Seconds so as to get some better inferences.



**Missing Values**

These are the values which are not available in the dataset. Following are the reason responsible for it-

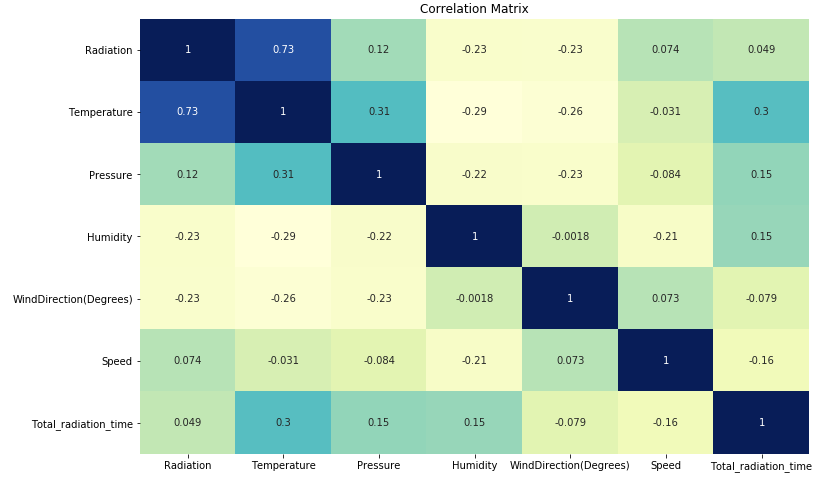
1. The person who filled the data did not want to share that part of information.
2. The person who entered the data into the system mistakenly forgot to enter it.
3. While transferring the data to the other system some portion of data was lost.



**Inference:** As the fig shows, there are no missing values in the dataset.

**Correlation Matrix**

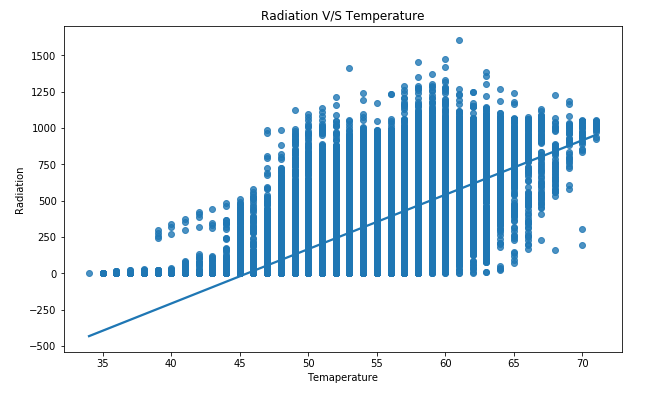
A correlation matrix is a table showing correlation coefficients between variables. Each cell in the table shows the correlation between two variables. A correlation matrix is used to summarize data, as an input into a more advanced analysis, and as a diagnostic for advanced analyses.



**Inference:** There is a high correlation of Radiation with Temperature is 0.73 which means with the increase in the Temperature there will be increase in the Radiation. Apart from it there is no such any major correlation.

**Scatterplot**

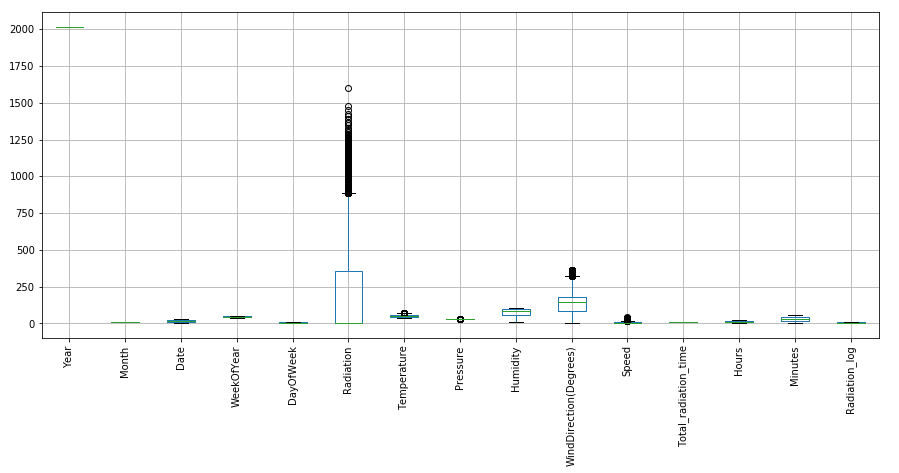
A scatter plot uses dots to represent values for two different numeric variables. The position of each dot on the horizontal and vertical axis indicates values for an individual data point. Scatter plots are used to observe relationships between variables.



**Inference:** The scatterplot has a positive Linear Relationship. But on the upper side there are more number of data points as compared to the Lower side of the best fit line. However, the best fit line is showing the positive linear relationship between both the variables.

**Outliers**

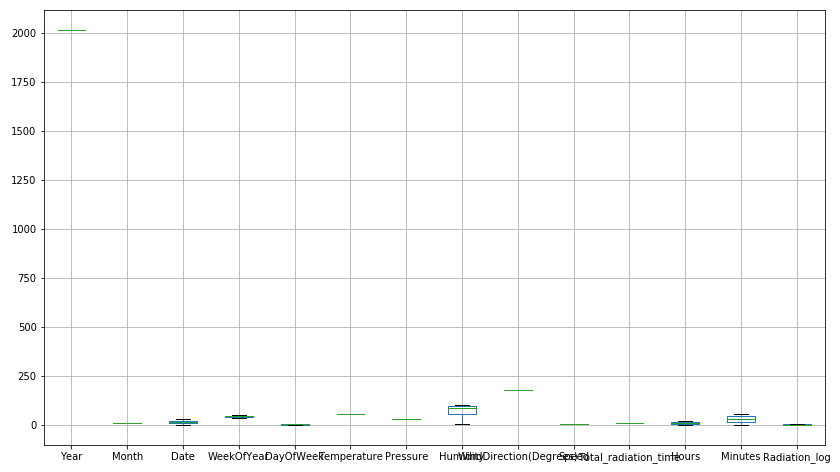
An outlier is an observation that lies an abnormal distance from other values in a random sample from a population.



**Inference:** There are outliers in Radiation, Temperature, Pressure, Wind Direction, and Speed. Since Radiation is our dependent variable therefore, we need to Log transform it & other column needs to be treated with the outlier treatment. The technique we will be using is Winsorization. Winsorization is the transformation of statistics by limiting extreme values in the statistical data to reduce the effect of possibly spurious outliers.

**Outliers Treatment**

By performing the IQR technique remove the outliers. We were losing 15% of the data. So in order to resolve that issue we have replaced the upper extreme values with upper boundary of the boxplot while we have replaced the lower extreme values with lower boundary of boxplot.



Shape of the data after treating the Outliers



**Inference:** Outliers have been removed from the data without losing the original shape.

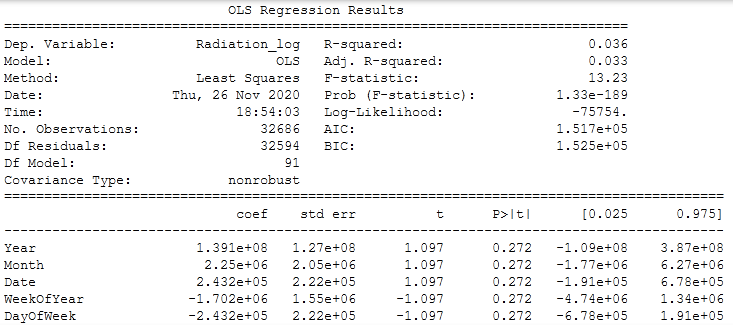
**SECTION – 3**

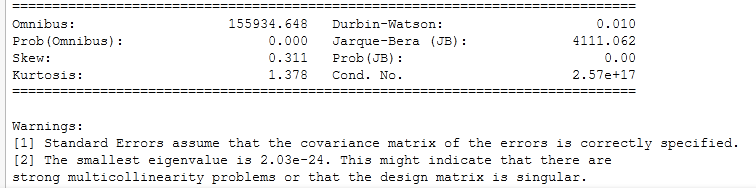
**MODELLING**

In this section, we are using 6 models which are as follows –

1. OLS Model
2. Linear Regression Model
3. Ridge Regression
4. Decision Tree Regression
5. K Neighbors Regression
6. Random Forest Regression
7. **OLS Model**

Ordinary Least Square (OLS) is a type of linear least squares method for estimating the unknown parameters in a linear regression model. OLS chooses the parameters of a linear function of a set of explanatory variables by the principle of least squares: minimizing the sum of the squares of the differences between the observed dependent variable (values of the variable being observed) in the given dataset and those predicted by the linear function.





**Inference:**

* R2 value is low which determines this model is not good. Condition number is high which shows there is a multicollinearity in the data.
* Durbin-Watson test value is 4.111 which represents the strong autocorrelation in the data.

1. **Linear Regression Model**

Simple linear regression is a linear regression model with a single explanatory variable. That is, it concerns two-dimensional sample points with one independent variable and one dependent variable (conventionally, the x and y coordinates in a Cartesian coordinate system) and finds a linear function (a non-vertical straight line) that, as accurately as possible, predicts the dependent variable values as a function of the independent variable. The Equation for linear regression model is -

**Linear Regression = b0 + b1x**

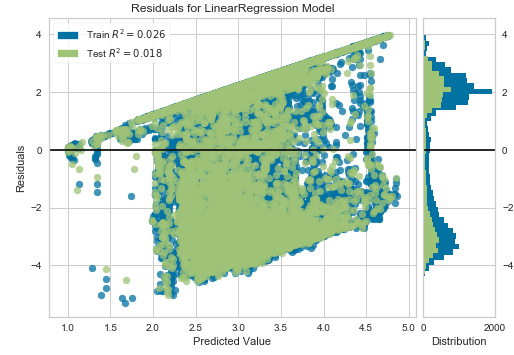
The overall accuracy of the model is-



R2 and RMSE score is –



Residual plot for Linear Regression is –



**Inference:**

* The residuals are unevenly distributed. This shows that a non-linear regression might perform better on the dataset.
* Majority of the residuals seem to be negative, indicating that in most of the cases, our predictions are lower than actual answers
* Residuals are widely spread in both directions, this may indicate that our model is under-fitted to the dataset.
* Linear Regression model is a bad choice for this dataset.
* The dataset shows Heteroscedasticity.

1. **Ridge Regression**

Ridge Regression is a technique used to analyze multiple regression data are multicollinear. When Multicollinearity takes place, estimates of the least squares are unbiased, but their variances are large so that they can be far from being the Real Value. By adding a degree of bias to regression estimates, the standard errors are reduced by ridge regression. It is hoped that the net effect will be to give more reliable estimates.

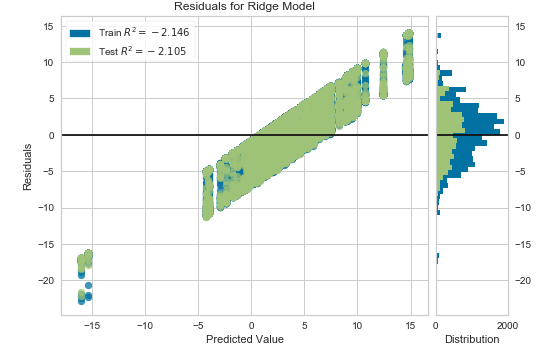
The overall accuracy of the model is-



R2 and RMSE score is –



Residual plot for Ridge Regression is –



**Inference:**

* Majority of the residuals seem to be negative, indicating that in most of the cases, our predictions are lower than actual answers
* Residuals are widely spread in both directions; this may indicate that our model is under fitted to the dataset.
* Linear Regression model is a bad choice for this dataset.

1. **Decision Tree Regression**

Decision Tree is a decision-making tool that uses a flowchart-like tree structure or is a model of decisions and all of their possible results, including outcomes, input costs and utility. Decision-tree algorithm falls under the category of supervised learning algorithms. It works for both continuous as well as categorical output variables. Decision tree regression observes features of an object and trains a model in the structure of a tree to predict data in the future to produce meaningful continuous output. Continuous output means that the output/result is not discrete, i.e., it is not represented just by a discrete, known set of numbers or values.

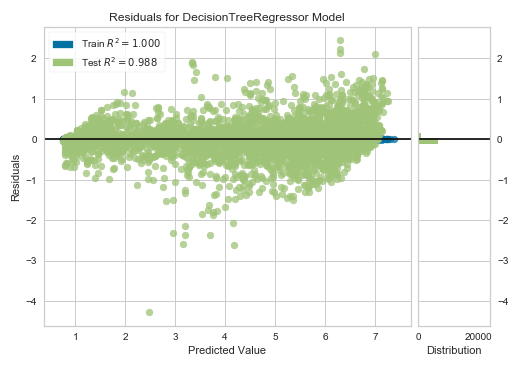
The overall accuracy of the model is-



R2 and RMSE score is –



Residual plot for Decision Tree Regression is –



**Inference:**

* The residuals for training data is evenly distributed but for the testing data they are randomly distributed.
* Test residuals are spread in both directions.

1. **K Neighbors Regression**

K nearest neighbors is a simple algorithm that stores all available cases and predicts a numerical target based on a similarity measure (e.g. distance functions). The simple implementation of KNN regression is to calculate the mean numerical target of the nearest K neighbors. Another approach uses the inverse distance weighted average of the nearest K neighbors. The KNN regression uses the same distance function as the KNN classification.

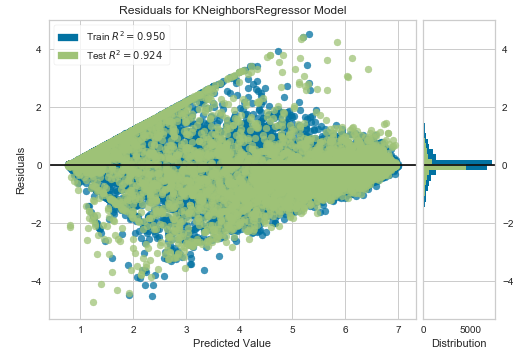
The overall accuracy of the model is-



R2 and RMSE score is –



Residual plot for K Neighbors Regression is –

****

**Inference:**

* The residuals are unevenly distributed. This shows that a non-linear regression might perform better on the dataset.
* Residuals are widely spread in both directions; this may indicate that our model is under fitted to the dataset.
* Test Residuals are negative in the first half of the predictions but are positive in the second half of test predictions

1. **Random Forest Regression**

Random forests or random decision-making forests are a group learning method for classification, regression and other tasks that operate by constructing a multitude of decision-making trees at training time and producing a class that is the class mode (classification) or mean/average prediction (regression) of individual trees. Random forests generally outperform decision-making trees, but their accuracy is lower than gradient boosted trees.

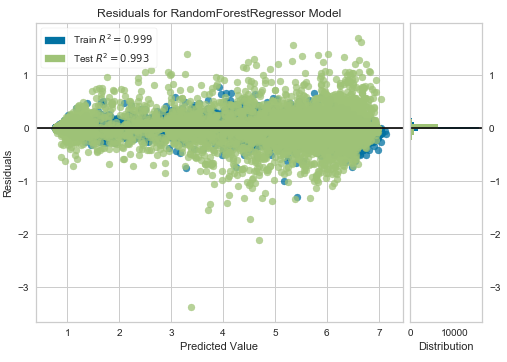
The overall accuracy of the model is-



R2 and RMSE score is –



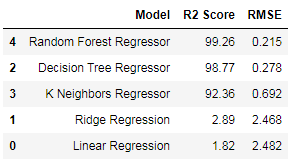
Residual plot for Random Forest Regression is –

****

**Inference:**

* The model is performing really well on the Train dataset
* The residuals are not randomly distributed. This shows that a non-linear regression such as this one is a good choice for this dataset
* Majority of the residuals are around the 0 line, indicating that the model is making lessor Errors in Prediction
* The residuals on Test dataset are negative in the mid-range of predictions, but positive when the predicted value is higher

**Selecting the best model-**



**Inference:**

* Among all the models Random Forest regression has the highest r2 score and lowest RMSE value. Therefore Random Forest Regression is the best performing model.