8-2020

Solar Irradiance Prediction Using Xg-boost With the Numerical Weather Forecast

Pratyusha Sai Kamarouthu
Utah State University

Recommended Citation

Kamarouthu, Pratyusha Sai, "Solar Irradiance Prediction Using Xg-boost With the Numerical Weather Forecast" (2020). *All Graduate Theses and Dissertations*. 7896.
https://digitalcommons.usu.edu/etd/7896

SOLAR IRRADIANCE PREDICTION USING XG-BOOST WITH THE NUMERICAL WEATHER FORECAST

by

Pratyusha Sai Kamarouthu

A thesis submitted in partial fulfillment
of the requirements for the degree

of

MASTER OF SCIENCE

in

Computer Science

Approved:

Nicholas Flann, Ph.D.
Major Professor

John Edwards, Ph.D.
Committee Member

Dan Watson, Ph.D.
Committee Member

Richard S. Inouye, Ph.D.
Vice Provost for Graduate Studies

UTAH STATE UNIVERSITY
Logan, Utah

2020

ABSTRACT

Solar irradiance prediction using xg-boost with the numerical weather forecast

by

Pratyusha Sai Kamarouthu, Master of Science

Utah State University, 2020

Major Professor: Nicholas Flann, Ph.D.
Department: Computer Science

There is a need to transition on carbon-free energy sources like solar energy and wind energy to avoid catastrophic global climate change. However, because wind and solar energies are intermittent, its supply must be integrated with storage and load management to satisfy the ever-growing energy demand. This thesis applies machine learning to predict solar energy production based on weather predictions and evaluates how predictions can improve the performance of a microgrid. Solar energy production is directly proportional to the solar irradiance available at the Earth's surface. Solar irradiance depends on many atmospheric parameters like temperature, cloud cover, and relative humidity, that are predicted in weather forecasts. This work applied regression algorithms - support vector regression and gradient boost algorithms like, xg-boost and cat-boost to weather forecast dataset from GFS (Global Forecast System), to predict solar energy production anywhere in the world up to 10 days with a resolution of three hours. Results show that prediction is reliable for two days. The learning model r-square error varies from 0.85 to 0.94 at different locations. Predicted solar energy is combined with predicted loads into a reinforcement based microgrid optimizer that schedules the charge and discharge of the battery and calculates the profit of an electric vehicle charging station. Results show that the profit made for 48 hours ahead by predictions is roughly proportional to the accuracy of solar irradiance.

(53 pages)

PUBLIC ABSTRACT

Solar irradiance prediction using xg-boost with the numerical weather forecast

Pratyusha Sai Kamarouthu

To defeat global warming, the world expects to look at renewable energy sources. Solar energy is one of the best renewable energy sources which causes no harm to the environment. As solar energy changes with atmospheric parameters like temperature, relative humidity, cloud coverage, dewpoint, sun position, day of the year, etc. It is difficult to understand its nature by science. Predicting solar irradiance which is directly proportional to solar energy using atmospheric parameters is the main goal of this work. Powerful artificial intelligence algorithms that won many coding competitions have been used to predict it. Using these methods and numerical weather forecast datasets one can predict solar irradiance up to ten days with the resolution of three hours. Two-day prediction is more reliable as error after that increases.

As solar energy is not available all day there is a need to pre-plan the storage and utilization. From an electric charge station perspective, if he knows the energy generated by solar and the amount of load he needs to supply, he can take a wise decision to supply the maximum load with the available power. This will make him get more profits. This experimental study has been executed by driving solar energy predictions along with load predictions to an algorithm that gives an optimum charge and discharge schedule of the battery considering the profit of the electric vehicle charging station. Profit is calculated with solar predictions in different scenarios with the consideration of the price of the energy at a given time.

To all the little people....

ACKNOWLEDGMENTS

I am so glad that I am part of Utah State University and its support to complete my graduation. I am thankful to the USU School of Graduate Studies, USU College of Engineering, USU College of Science, and USU computer science department for providing me financial support.

I am so grateful to my major professor Dr. Nicholas Flann who helm me in the right direction in my research. I admire his patience and time he has given to make things clear. I had wonderful learning from his great ideas. I have a great appreciation for his knowledge and his mastery to guide students and his encouragement to try new things. I also want to thank professor Edwards and professor Watson for being my committee chair

I had a wonderful and enthusiastic team that worked with me to produce great work. Saju Saha has been a wonderful human being and the person who explore the details and the discussion we had was so informative and wonderful. Ashit Neema was so helpful in running my experiments and complete my task. I want to acknowledge Qi Luo for sharing her project findings which helped me in my work.

Aditi, Agnib, Sheril, Rejoy, and Wasim for making my stay at Utah State University wonderful. I appreciate Aditi Jain, Lasya Alavala, Chelsi Gupta, and Manish for their advice and guidance. I want to thank my brother Gnanendera Varma who supported me financially and emotionally at times of distress. I appreciate my dad Nageswara Rao who encouraged me to pursue my dreams and my mom Maha Lakshmi who second him. A special mention of my aunt Rukmini and my friends Anuj Verma, Sruthi, Srikanth for supporting me emotionally all the time.

Pratyusha Sai Kamarouthu

CONTENTS

LIST OF TABLES

Table Page

LIST OF FIGURES

Figure Page

ACRONYMS

GHI	Global horizantal irradiance
DHI	Diffuse horizantal irradiance
DNI	Direct normal irradiance
MAE	Mean absolute error
GFS	Global forecast system
ECMWF	European center for medium range weather forecast
ARMA	Auto-regressive moving averages
ARIMA	Auto-regressive integrated moving averages
NOAA	National oceanic and atmospheric administration
SVR	Support Vector Regression
Xg-boost	extreme gradient boost
Cat-boost	Categorical boosting
EV	Electric Vehicle

CHAPTER 1
INTRODUCTION

1.1 Introduction

According to the United States Environment Protection Agency Electricity, transportation, and industries contribute seventy-nine percent to the greenhouse gas emissions. Transportation sector topped followed by electricity production. Currently, around sixty-three percent of electricity is generated by fossil fuels either coal or natural gas. Increment of greenhouse gasses concerns future climate changes. Hence Paris agreement has been adopted by 197 nations to keep global temperature below 2° Celsius. [4] Decarbonizing the energy sources is one of the prominent steps to achieve this goal. This makes us explore carbon-free energy production options. Many countries decided upon integrating clean energy sources into their grid as 100% renewable energy system is impractical at the moment. [5] [6]

Integrating clean energy sources is not an easy task to achieve because of two reasons: 1.its intermittent and uncontrollable nature [5] 2. ever-growing demands. All renewable energy sources like solar energy, wind energy, tide energy, geothermal energy are available from no reliable sources. Many unknown parameters contribute to affects energy production and make them unpredictable and uncontrollable. There would be a 40% increase in the demand until 2040 [6]. But if one can overcome those challenges by reliable forecasting and load shifting algorithms, then they no need to import oils and natural gas from other countries. Energy production in every country will become independent of coal reserves. This shift would greatly benefit the country's economy in the long run.

Due to the development of artificial intelligence and the availability of high-end computer sources, it is possible to predict solar and wind energies. Solar irradiance is directly proportional to the amount of solar energy generated. Though solar irradiance data is cyclic in general nonlinearity occurs because of various atmospheric parameters like temperature, relative humidity, dew point, etc. [7]

Solar Irradiance is defined by three components. Those are GHI (Global Horizontal Irradiance), DHI (Diffuse Horizontal Irradiance) and DNI (Direct Normal Irradiance). GHI is defined as the total amount of radiation received by the surface horizontal to the ground per unit area, θ is 90 at the vertical sun.

$$GHI = DNI \times cos\theta + DHI$$

DHI is the amount of diffused solar energy received from all directions per unit area other than direct sun rays. Generally, radiation is scattered by all the particles and things for example clouds. [8]

DNI is the amount of radiation received per unit area that is normal to the rays from the sun. In this work, GHI is predicted and GHI is addressed as solar irradiance in the rest of the document.

1.2 Problem Description

Solar irradiance is directly proportional to solar energy produced. Hence predicting solar irradiance will help to predict solar energy by a mathematical formula. [9] Solar irradiance is defined as the amount of energy received from the sun for one-meter square. The measure of solar irradiance is highly dependent on sun position diurnal cycle and weather parameters like cloud cover, relative humidity, temperature, dew point, precipitation, etc.. [8]. To forecast it, a model needs to be built using machine learning algorithms, that can predict solar irradiance that is GHI(Global Horizontal Irradiance) by weather parameters and position of the sun which ultimately mapped to solar energy.

Given that latitude and longitude of a location and historical solar irradiance of a particular location, global horizontal irradiance is predicted using numerical weather forecast, for ten days with the resolution of three hours.

1.3 Previous work

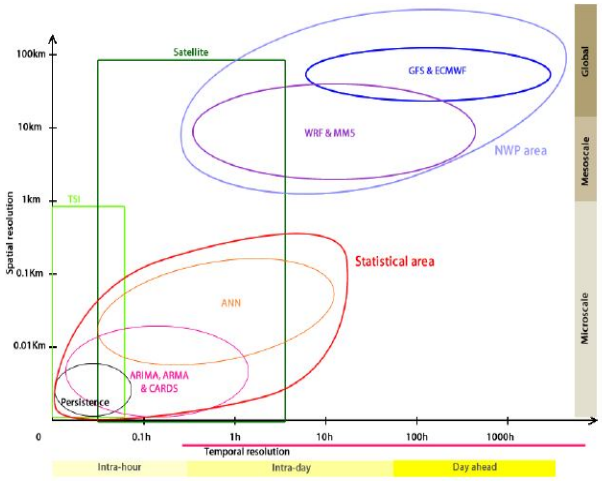


Fig. 1.1: Different prediction methods used for different spatial and temporal resolutions [1]

Based on different scenarios different methods are used to predict solar irradiance. In figure 1.1, different methods for different time and spatial resolutions are shown. There are six different classes of methods for the prediction. They are persistence models, classical statistical models, machine learning techniques, cloud motion tracking from ground or satellite, numerical weather prediction models, hybrid models [8] [1].

Persistence models mainly state that current climatic conditions will be similar to past climatic conditions. Instead of using it individually, it is used with the combination of the model to accurately measure the solar. [8]

Classical statistical methods like ARMA, ARIMA are used for intra hour prediction and sometimes up to three hours ahead. [8] These methods consider the past values of solar irradiance to predict the future. They don't consider the present weather conditions. [8]These methods capture the sharp transitions in solar irradiance associated with diurnal cycle. [1]

Cloud motion tracking from ground or satellite methods are very accurate and can predict the fluctuations better. These methods require a camera that would capture the cloud motion. By processing these images by a convolution network prediction of solar power is made. As ground camera covers just a few meters and clouds move rapidly prediction for short periods holds good. [8] If satellite images are used instead of ground images it can predict for intraday. [9] [1] In either way, predicting this way gives a high resolution.

Numerical weather forecast models are a mathematical model that takes the present weather conditions and predict weather conditions in the future. After predicting climatic conditions solar irradiance can be derived by a formula [8]. NWP methods are good to predict for two days and can be extended up to 6 days. [9]

Machine learning techniques are used to understand the relationship between inputs and output [9]. These techniques are better to understand the variance of solar irradiance using weather data [9]. Many regression algorithms like an artificial neural network, decision tree learning, support vector machines, K-means clustering, ensemble learning, etc methods have been used for prediction. [8] [9]. Because of the lack of uniformity in data performance comparison is impossible. [9].But prediction is dependent on the weather forecast available.

The method proposed in this work is a hybrid model that uses machine learning algorithms on the numerical weather forecast data to address the uniformity in data throughout the world and exploit the weather forecast.

Chapter 2 describes the numerical weather forecast and its validity along with feature selection. Chapter 3 describes the different learning algorithms used and talks about the performance of each learning method. Chapter 4 talks about the error analysis and experimental study of solar prediction on the optimizer to make a profit out of this solar irradiance prediction.

The scope of this work limited to the accuracy of solar irradiance and understand how numerical weather prediction forecast influences the accuracy of solar irradiance. But no efforts are made to correct the numerical weather forecast data in case of errors.

CHAPTER 2
NUMERICAL WEATHER PREDICTION MODELS

2.1 Dataset selection

Considering this project will be deployed in a physical system, a dataset is required that can predict solar irradiance in real-time. Additionally, one can predict at any given location in the entire world. Numerical weather forecast data gives us the flexibility to work on any data set at any location. For example, if one has load data somewhere in New York he should be able to pull out the solar irradiance prediction at that place and carry on with the research. When different datasets are considered other than numerical weather prediction data there is no synchronization between historical weather parameters and forecasted weather parameters. For example, NOAA provides historical cloud data in octanes but forecasted cloud data is in percentage. There is no single source that provides both historical and forecasted weather parameters. As the main task is to build a machine learning model, historical data is very important for training the model.

The numerical weather prediction dataset addresses all the issues. Each file generated from the mathematical model contains weather parameters all over the world. As one can get historical data and forecasted data from a single source it makes training and predicting easy. Numerical weather prediction file also contains many weather parameters that would help us to choose all the attributes that contribute to solar irradiance.

2.2 Description

Forecasts of atmospheric parameters can be done using physical laws. As per this approach forecasting should be looked at as an initial value problem of mathematical physics. By integrating the initial climate conditions with partial differential equations to predict future values. [2] When this idea is proposed the world has very limited resources but now after 100 years, it has the most powerful computers which can solve all the complex problems. The physical-chemical processes that happen in the atmosphere are more accurately measured and considered in the mathematical model. If one look back 40 years there is a huge improvement in forecasting skills by numerical weather predictions as in figure 2.1 Forecasting skill is the correlation between the forecasts and verified data. If predictions are 60% accurate it is a useful prediction were 80% accurate it can be said highly accurate. There have been continuous efforts to make these NWP models accurate. [2]

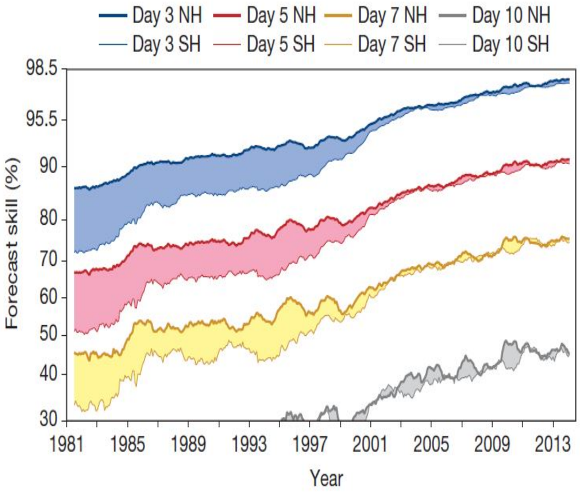


Fig. 2.1: Improvement of forecasting skill over 40 years for 3 days, 5 days, 7 days and 10 days ahead by NWP model [2]

2.3 Different numerical weather prediction models

There are different NWP models but mainly classified as global models and mesoscale models. Global models give forecasted parameters for the entire world while mesoscale models provide forecasted parameters for a limited area. Meso-scale models are built on the output of global models and add local variables and make them accurate. NAM(North American Meso Scale model) and NDFD(National Digital Forecast Database) are examples of the mesoscale model. To measure solar irradiance at any given place in the world one needs to consider a global model. GFS(Global Forcast System) and ECMWF(The European Center for Medium-range Weather Forecast) are two important global NWP models. Though ECMWF does better predictions than GFS, ECMWF data is not freely available. Hence this work is carried on using GFS forecasts.

2.4 Understanding GFS data

GFS(Global Forecast Model) is built and maintained by NCEP(National Centers for Environmental Prediction).GFS runs four times in a day at 00,06,12 and 18 hours in UTC. In each run, it can predict 384 hours with a resolution of three hours for 240 hours that is 10 days and for 12 hours from 240 to 384 hours with a spatial resolution of 0.5 latitudes and longitude. [10]

All the files generated are in the GRIB format and the past one year data is available for immediate download but data previous than one year will be uploaded on the request. Each time the system also provides a zero-hour forecast. Ideally, the zero-hour forecast would be the initial atmospheric state which is collected from weather stations. Hence in the given work that data is considered as historical data and used to train the model. The latest file will be updated after each run this enables us to predict solar irradiance in real-time.

There are 354 weather parameters given in the output of the GFS model and each parameter is a matrix of (361,720) as there are 180 latitudes and 360 longitudes. Locating the exact location among this matrix is one of the crucial steps. One can directly map latitude value to the 2-d matrix. But there is a need to change the notation of longitude from (-180,180) to (0,360) to accurately take the required location. Each point represents the intersection of latitude and longitude in a 2-D array which is then averaged value around 50 km. Few parameters are given at different pressure levels of the atmosphere.

2.5 Data Extraction

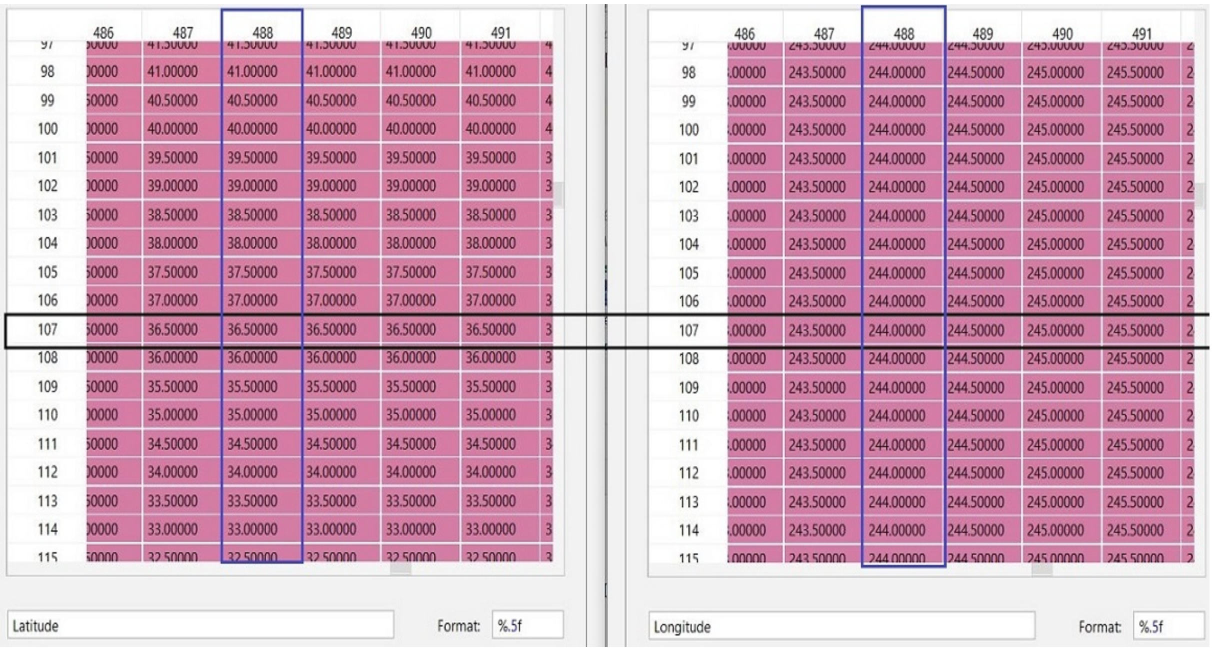
The output of the GFS model is gridded data with a resolution of 0.5 latitudes and longitudes. There is a python package named Pygrib which is written in C helps to read grib file. EcCodes setup is required to install Pygrib. This is recommended to use in ubuntu in the given documentation but it can be set up in windows as well. This work is carried on a python miniconda environment in windows.

Each atmospheric parameter is given in 361 by 720 2-D matrix as there are 180 latitudes and 360 longitudes. In GFS latitudes vary from 90 to -90 with 0.5 resolution and longitudes from 0 to 360 with 0.5 resolution. The exact location should be located based on the latitude and longitude of a specific place. Consider the location Desert Rock in Nevada which has latitude 36.624 and longitude -116.019. The first step is that one needs to approximate the latitude and longitude to the nearest latitude and longitude that is 36.5 and -116 in this scenario as the data has a resolution of 0.5.

In figure 2.2 left side window shows the 2-D matrix of latitudes and right side window shows 2-D matrix of longitudes. 107$^{\text{th}}$ row represents 36.5 latitude. To locate longitude one needs to convert -180 to 180 notation to 0 to 360 notation. This conversion can be done by longitude(-180 to 180) mod 360 which in 244 in this scenario hence 244 longitude is located in 488$^{\text{th}}$ column. [107,488]$^{\text{th}}$ value in the 2-D array correspond to Desert Rock in Nevada. 2.3 shows temperature matrix measured in Kelvin. The highlighted cell gives the temperature value of Desert rock in Nevada. To extract data one needs to select [107,488]$^{\text{th}}$ value in each atmospheric parameter.

GFS model output is available every six hours that is at 00,06,12 and 18 hours of universal time zone but forecasts with an interval of 3 hours. The live data is also available with NOAA which helps us to predict the real future.

Under SUFRAD program NOAA has established the SUFRAD stations at seven places. These stations measured the solar irradiance components with high accuracy using a pyranometer. This data is helpful for the training of the model. Data is available for every minute. To match this data to the weather data which is available for every six hours, downsampling is required. Downsampling can be done in many ways. One among them is picking up the instantaneous value. This does not give accurate measure because of the sudden cloud at that minute. Hence averaging about an hour data will give a better measure while downsampling. For example, if one needs a solar irradiance measure at 6 then



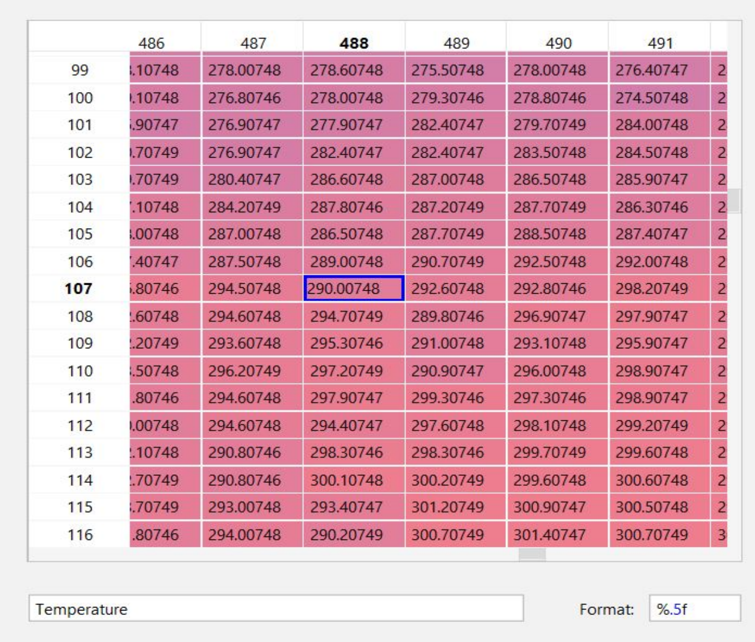
	486	487	488	489	490	491	
97	50000	41.50000	41.50000	41.50000	41.50000	41.50000	4
98	00000	41.00000	41.00000	41.00000	41.00000	41.00000	4
99	50000	40.50000	40.50000	40.50000	40.50000	40.50000	4
100	00000	40.00000	40.00000	40.00000	40.00000	40.00000	4
101	50000	39.50000	39.50000	39.50000	39.50000	39.50000	3
102	00000	39.00000	39.00000	39.00000	39.00000	39.00000	3
103	50000	38.50000	38.50000	38.50000	38.50000	38.50000	3
104	00000	38.00000	38.00000	38.00000	38.00000	38.00000	3
105	50000	37.50000	37.50000	37.50000	37.50000	37.50000	3
106	00000	37.00000	37.00000	37.00000	37.00000	37.00000	3
107	50000	36.50000	36.50000	36.50000	36.50000	36.50000	3
108	00000	36.00000	36.00000	36.00000	36.00000	36.00000	3
109	50000	35.50000	35.50000	35.50000	35.50000	35.50000	3
110	00000	35.00000	35.00000	35.00000	35.00000	35.00000	3
111	50000	34.50000	34.50000	34.50000	34.50000	34.50000	3
112	00000	34.00000	34.00000	34.00000	34.00000	34.00000	3
113	50000	33.50000	33.50000	33.50000	33.50000	33.50000	3
114	00000	33.00000	33.00000	33.00000	33.00000	33.00000	3
115	50000	32.50000	32.50000	32.50000	32.50000	32.50000	3

Latitude Format: %.5f

	486	487	488	489	490	491	
97	.00000	243.50000	244.00000	244.50000	245.00000	245.50000	2
98	.00000	243.50000	244.00000	244.50000	245.00000	245.50000	2
99	.00000	243.50000	244.00000	244.50000	245.00000	245.50000	2
100	.00000	243.50000	244.00000	244.50000	245.00000	245.50000	2
101	.00000	243.50000	244.00000	244.50000	245.00000	245.50000	2
102	.00000	243.50000	244.00000	244.50000	245.00000	245.50000	2
103	.00000	243.50000	244.00000	244.50000	245.00000	245.50000	2
104	.00000	243.50000	244.00000	244.50000	245.00000	245.50000	2
105	.00000	243.50000	244.00000	244.50000	245.00000	245.50000	2
106	.00000	243.50000	244.00000	244.50000	245.00000	245.50000	2
107	.00000	243.50000	244.00000	244.50000	245.00000	245.50000	2
108	.00000	243.50000	244.00000	244.50000	245.00000	245.50000	2
109	.00000	243.50000	244.00000	244.50000	245.00000	245.50000	2
110	.00000	243.50000	244.00000	244.50000	245.00000	245.50000	2
111	.00000	243.50000	244.00000	244.50000	245.00000	245.50000	2
112	.00000	243.50000	244.00000	244.50000	245.00000	245.50000	2
113	.00000	243.50000	244.00000	244.50000	245.00000	245.50000	2
114	.00000	243.50000	244.00000	244.50000	245.00000	245.50000	2
115	.00000	243.50000	244.00000	244.50000	245.00000	245.50000	2

Longitude Format: %.5f

Fig. 2.2: Identifying the exact location for Desert Rock in Nevada with latitude 36.624 and longitude -116.019



	486	487	**488**	489	490	491	
99	.10748	278.00748	278.60748	275.50748	278.00748	276.40747	2
100	.10748	276.80746	278.00748	279.30746	278.80746	274.50748	2
101	.90747	276.90747	277.90747	282.40747	279.70749	284.00748	2
102	.70749	276.90747	282.40747	282.40747	283.50748	284.50748	2
103	.70749	280.40747	286.60748	287.00748	286.50748	285.90747	2
104	.10748	284.20749	287.80746	287.20749	287.70749	286.30746	2
105	.00748	287.00748	286.50748	287.70749	288.50748	287.40747	2
106	.40747	287.50748	289.00748	290.70749	292.50748	292.00748	2
107	.80746	294.50748	290.00748	292.60748	292.80746	298.20749	2
108	.60748	294.60748	294.70749	289.80746	296.90747	297.90747	2
109	.20749	293.60748	295.30746	291.00748	293.10748	295.90747	2
110	.50748	296.20749	297.20749	290.90747	296.00748	298.90747	2
111	.80746	294.60748	297.90747	299.30746	297.30746	298.90747	2
112	.00748	294.60748	294.40747	297.60748	298.10748	299.20749	2
113	.10748	290.80746	298.30746	298.30746	299.70749	299.60748	2
114	.70749	290.80746	300.10748	300.20749	299.60748	300.60748	2
115	.70749	293.00748	293.40747	301.20749	300.90747	300.50748	2
116	.80746	294.00748	290.20749	300.70749	301.40747	300.70749	3

Temperature Format: %.5f

Fig. 2.3: Locating temperature at Desert Rock Nevada

averaged values of solar irradiance from 5:30 to 6:30 is considered. In figure 2.4 blue line shows the solar irradiance varying every 10 minutes and the red line shows the averaged solar irradiance over an hour at each point on the curve. This red curve has been used to match the weather parameters for the particular timestamp.

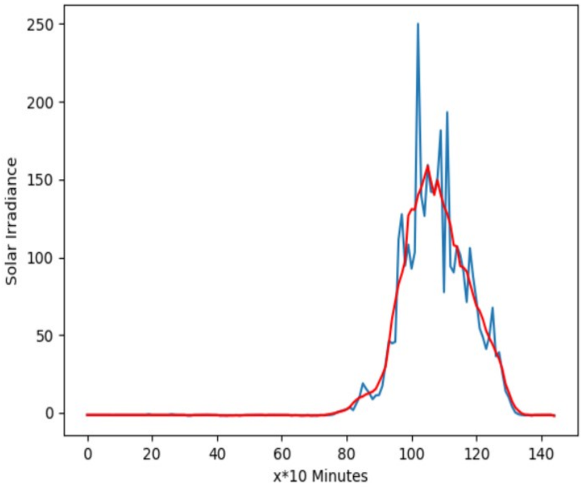


Fig. 2.4: Solar irradiance curve for a day by averaging the values over one hour

2.6 Re-evaluating the forecast

Error in prediction is highly dependent on numerical forecast error. Hence it is necessary to re-evaluate how close is the forecast value to the true data. Figures 2.5 2.6 2.7 helps to understand the deviation of real value to the predicted value as the predicted timestamp is approaching. Where the x-axis shows the timestamp of true data and the y-axis shows the number of hours before which the value has been predicted. This has been plotted for all the considered weather parameters.

In figures 2.5 2.6 2.7 each block in the heat map shows the difference between forecasted weather parameter and recorded weather parameter. A randomly selected period of November is considered for the evaluation. On the x-axis it shows the timestamp of a particular column which is from November 10th zeroth hour to November 20th the eighteenth

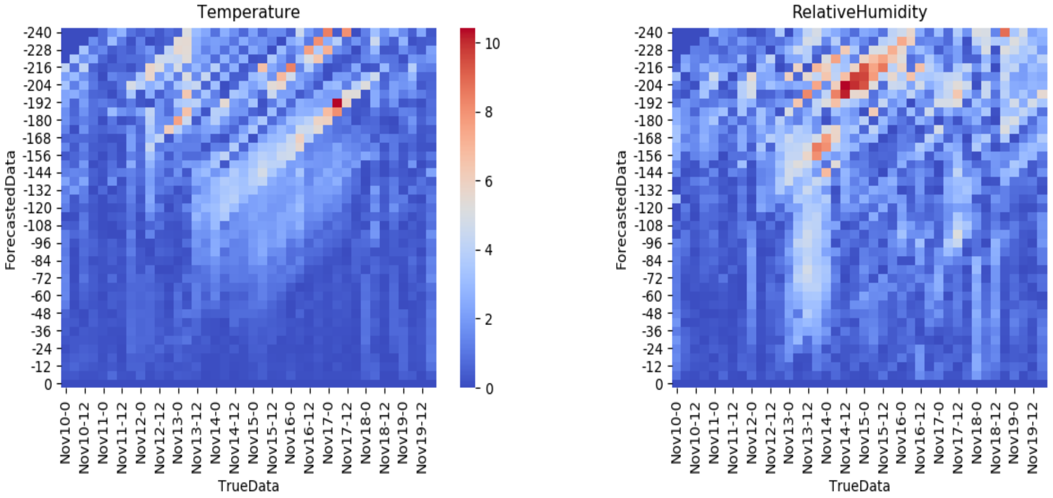


Fig. 2.5: Absolute difference of forecasted versus verified temperature and relative humidity

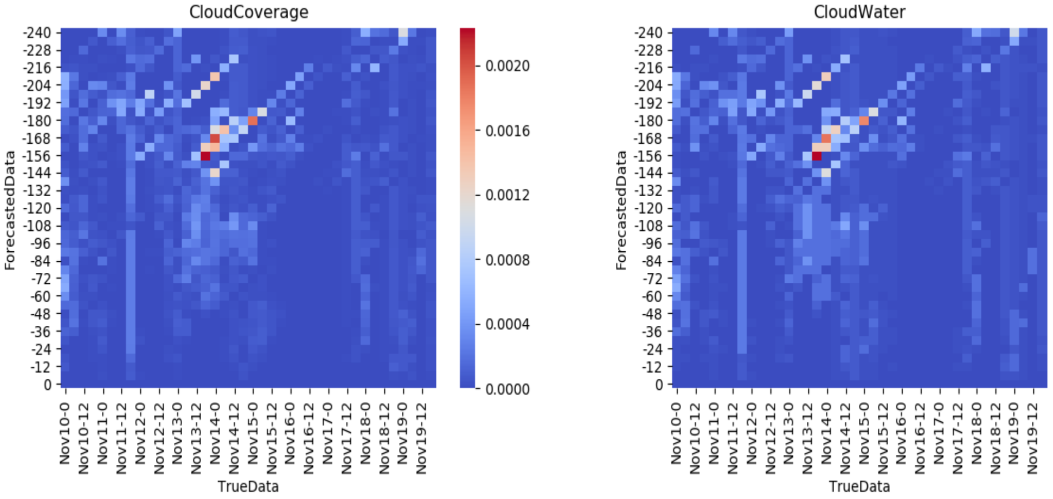


Fig. 2.6: Absolute difference of forecasted versus verified cloud coverage and cloud water

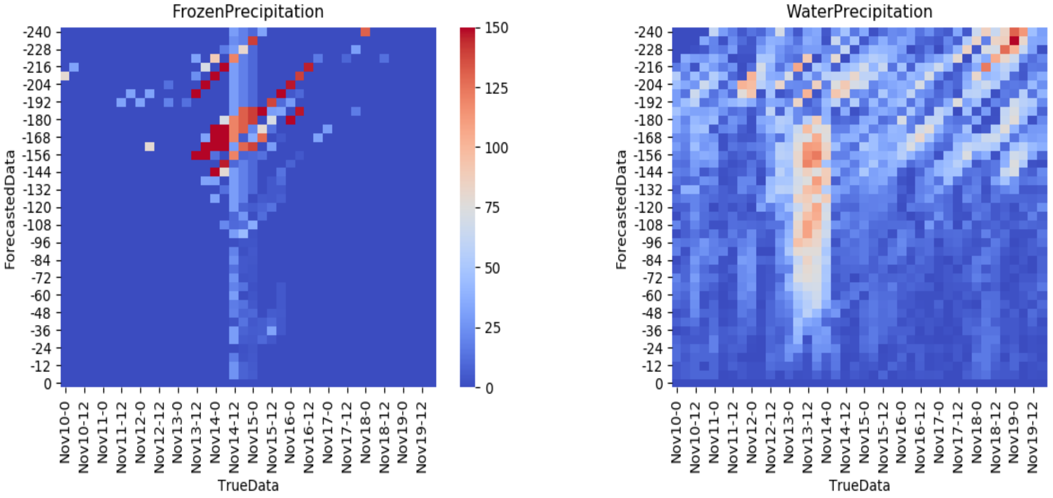


Fig. 2.7: Absolute difference of forecasted versus verified frozen and water precipitation

hour for every six hours. The top row in the visualization shows the absolute difference between predicted value 240 hours ahead to the true data for the particular time stamps. The second top row shows the difference between 234 hours ahead predicted value to true data. Likewise, it goes down by six hours until it reaches zero. Dark blue shades show the minimum error. As one can see through, the error is more and varying for long forecasts and it is forecasting good for short forecasts which are until 72 hours ahead.

Few observations can be made looking at these pictures. One is the visible cross lines in the heat maps this shows carrying on of error for future forecasts. For example fig 2.5 for temperature there are many diagonal lines in the upper part of the heat map. One among them is forecast on November 7th 18th hour for November 14th 6th hour temperature value has an error. In other words temperature forecast made for November 14th 6th hour 156 hours ahead has an error when compared with true data. So the error is carried on to the further temperature forecast. As November 7th 18th has error when forecasted for November 14th 6th hour this results in error for November 14th 12th hour temperature forecast and November 14th 18th hour temperature forecast and so on. In general, the diagonal line shows that forecast error at a particular time is carried on to other future forecasts.

Observing red blocks in figures 2.6 2.7 for cloud coverage, cloud water, frozen precipitation, and water precipitation shows that frozen precipitation has not been forecasted properly for November 13 18 th hour 150 hours ahead but as the forecast is approaching near to zero the error has been reduced. This shows forecasting is getting better as it becomes a short-range. Though it was able to predict better for short ranges there are few exceptions. For example in figure 2.7 water precipitation error is more for 48 hours ahead prediction during unexpected precipitation time. Apart from that numerical weather forecast is doing a good job to be considered for solar irradiance prediction.

2.7 Feature Selection

The correlation matrix in Fig 2.8 gives how weather parameters are correlated with the GHI. Blue color shows a positive correlation while red shows a negative correlation. There is a total of 354 variables available from the GFS output. Among them, few weather

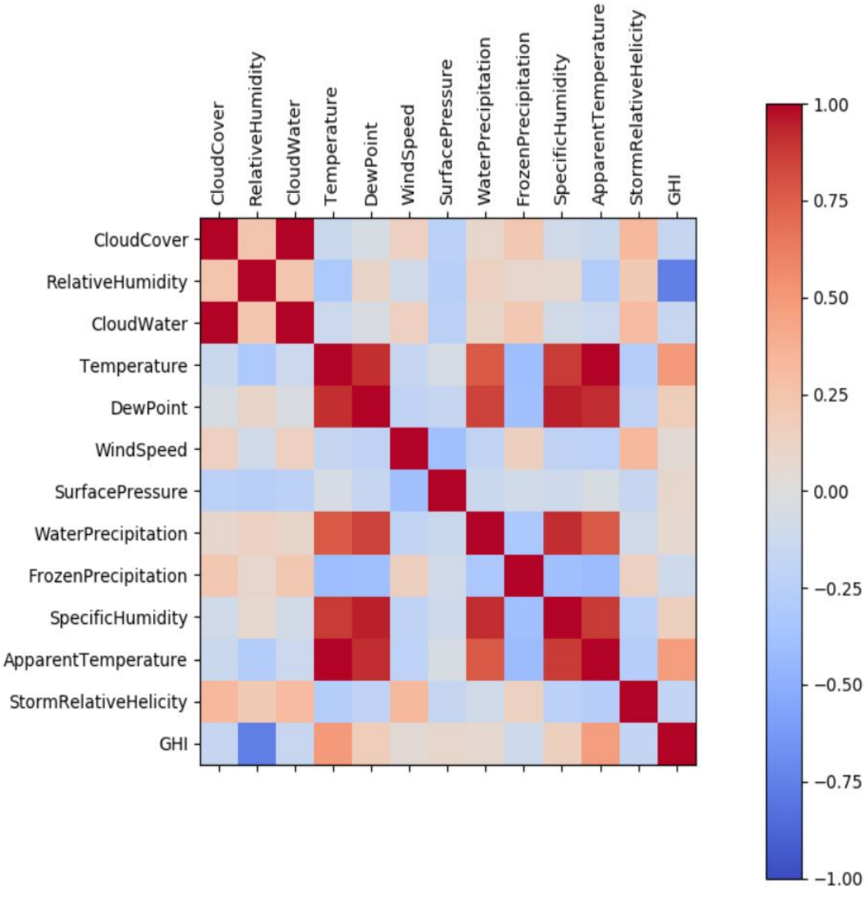


Fig. 2.8: Correlation matrix plot with different possible features

parameters are available at different pressure levels. For example wind speed, U component of wind speed, V component of wind speed, vertical velocity, relative humidity, temperature, etc. So the first set of parameters is selected by looking at the previous work and general analysis. For example, geopotential height and wind components would not have any effect on solar irradiance. So among 354 a set of 33 are selected as the second set.

Solar irradiance is indirectly proportional to the density of the cloud. A total amount of cloud density throughout the atmosphere of the earth should be calculated to see how much cloud is obstructing the sun's intensity. To find out the total cloud density the cloud mixing ratio which means the mass of the cloud divided by the mass of the air is considered. As it gives values amount of cloud material that is obstructing the sun rays. So a simple addition of cloud mixing ratio at different levels gives total cloud density. Hence cloud density of 20 different levels has been added to know the entire cloud cover. With that 14 parameters have been plotted to see their correlation.

Temperature and Relative humidity above two-meter are available. These parameters are also available at different pressure levels. The pressure level's values are not considered because this parameter is fixed in terms of height where value is calculated. But as the altitude of the place changes this value doesn't make much sense. Hence for temperature, two-meter temperature and apparent temperature are considered and for relative humidity, two-meter relative humidity and specific humidity are considered as the altitude does not affect these values. Apparent temperature is considered as the feature because it gives how much temperature a person feels this gives a generalized temperature in the atmosphere irrespective of altitude of the location. There is a total of three different kinds of features that have been used for training the dataset. One is meteorological parameters like apparent temperature, cloud water, relative humidity, frozen precipitation, water precipitation, Specific humidity, dew point, cloud cover, temperature. Second is sun position parameters and the third is the diurnal cycle parameters.

Sun position is the key to determine the variation of sun intensity on the given day. Sun position can be defined by the zenith angle and azimuthal angle. So these two parameters have been considered as features. The third set of features are diurnal cycle parameters like the day of the year, month, and hour are considered.

CHAPTER 3
METHODS

Machine learning is a part of artificial intelligence. Due to technological advancements, the world has an enormous amount of data available in all industries. This made machine learning as an indispensable tool in every industry to draw formative conclusions [8]. Predicting solar irradiance has also been treated as a machine learning problem because calculating solar irradiance is the complex environmental physics it is very difficult to calculate numerically as it has many variables to be considered. Machine learning algorithms can draw linear and nonlinear relationships between inputs and output. Solar irradiance has a nonlinear relationship with the environment variables. Support Vector regression with radial bias kernel is helpful to extract relations as per previous work. And gradient boosting algorithms which are tree-based machine learning algorithms are also powerful. These algorithms won many coding competitions. xg-boost, in particular, has won many coding competitions. In KDD cup 2015 all the top 10 winners used the xg-boost algorithm [11]. While training day data has been considered. 2016 to 2019 august weather data is considered as a training dataset at all locations and for all methods.

3.1 Support vector regression

Support Vector Machine algorithm can be modified and can use as a regression algorithm. The modified algorithm is known as SVR(Support vector regression). It produces accurate results with less computation. It fits a hyperplane in multidimensional space to categorize data. This hyperplane is the regression line in the case of SVR and works as a segregation for classification problems. Support vectors which are the data point close to the hyperplane help to maximize the margin for efficient classification. For complex data where hyperplane cannot be fit, these data points are transformed into the higher dimensional plane where it fits a hyperplane and again inverse the transform. These transformations are called kernel functions. These three types of kernel functions which are the linear kernel, polynomial kernel, and radial bias kernel. I have used the radial bias kernel which helps to fit a nonlinear curve for the given data. [12]

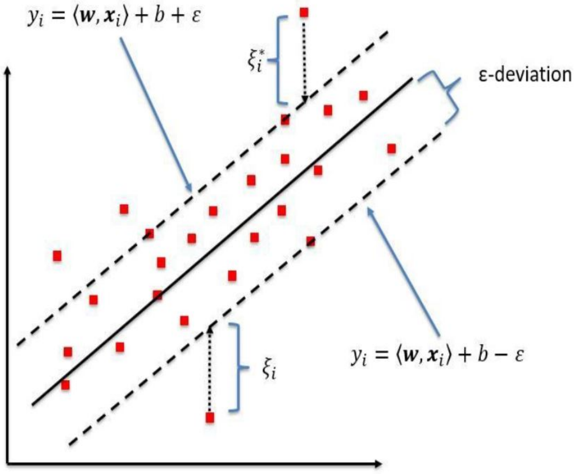

$y_i = \langle \boldsymbol{w}, \boldsymbol{x}_i \rangle + b + \varepsilon$
ξ_i^*
ε-deviation
ξ_i
$y_i = \langle \boldsymbol{w}, \boldsymbol{x}_i \rangle + b - \varepsilon$

Fig. 3.1: Support vector regression method representation [3]

In Fig 3.1 the red points are the data points and a line is a hyperplane. The dotted line denotes the margin around the hyperplane

The basic difference between linear regression and support vector regression is the error. In linear regression, the algorithm tries to fit the line with as much less error as possible. But in support vector regression, the algorithm fixes the error and try to fit the best possible line between that error. SVR uses ϵ insensitive loss function. If the error is between -ϵ and ϵ the cost function is not penalized. If there are data points above that margin it penalizes the loss symmetrically. To avoid the overfitting regularization parameter is used. Regularization is penalizing the loss function and make it insensitive to the loss and avoid over-fitting the curve [3].

Fig 3.2 shows the solar irradiance prediction in Nevada using SVR

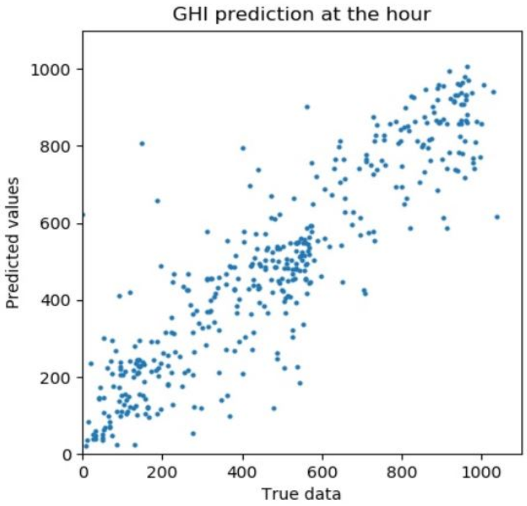


Fig. 3.2: True data Vs predicted data using SVR in Nevada

3.2 Gradient boosting methods

Machine learning algorithms that come in this category are really powerful. These algorithms have proven records in many Kaggle data science competitions. After deciding upon data a single predictive model either by ANN(Artificial Neural Network) or SVR(Support Vector Regression) and use a particular model to do prediction. But ensembling has brought a different idea of building a model. Many models will be built by different parameters and a final decision is made. This is the idea to build strong learners using many weak learners. In random forests has simple averaging of predicted models but gradient boosting algorithms follow a different strategy to combine and make a strong learner. The gap between predicted and true data will be further processed by another model. This process is repeated until one gets a good accurate model. The gap between predicted output and true data is the loss function. Generally, the mean square error(MSE) is used. Gradient descent has been used for minimizing the loss function. [13]

Suppose for the given data $\mathrm{F_o(x)}$ is the model to estimate with a minimum loss function.

$$F_{\mathrm{o}}(x) = argmin_{\gamma} \sum_{n=1}^{n} (y_{\mathrm{i}}, \gamma)$$

$\mathrm{F_o}$(x) gives the initial stage of the model and now residual error for each instance ($\mathrm{y_i}$-$\mathrm{F_o}$(x)).This helps to predict $\mathrm{h_1}$(x) which is not predicting value instead it helps to predict $\mathrm{F_1}$(x). $\mathrm{h_1}$(x) is the addictive model that computes the mean of the residuals at each leaf of the tree.$\mathrm{F_1}$(x) is obtained by the addition of $\mathrm{h_1}$(x) with $\mathrm{F_o}$(x). This way $\mathrm{h_1}$(x) learns from residuals of $\mathrm{F_o}$(x) and suppress it in $\mathrm{F_1}$(x) in order to obtain better model. This process is repeated many times. The residual at each stage will be useful to predict the next model and it observes the pattern in the residual errors. When it obtains maximum accuracy or there won't be any pattern in the residual error finally training would be stopped. The number of splits that a tree can make can be specified by the user.

3.2.1 xg-boost

Xg-boost(eXtreme gradient boosting) is one of the powerful machine learning algorithms in recent times. It has taken over the world in terms of accuracy and speed. It operates on parallel and distributed computing which makes learning very fast compared to other ensemble algorithms. This algorithm is a modified version of the generalized gradient boosting algorithm. [14] xg-boost algorithm builds a different kind of tree from the gradient boosting algorithm. In xg-boost, the split is found by using similarity score and gain. The regularization parameter is used to avoid over-fitting of the split. When the regularization parameter is zero it falls into the traditional gradient boosting algorithm. Along with regularization, two other techniques avoid overfitting. One is shrinkage scales which modify the weight after each step by a factor η. Its idea is to reduce the influence of an individual tree on the model. The second way is to use column subsampling this also improves the training time. The other important step is finding the best split by using an approximate algorithm. [14]

An efficient xg-boost tool has been built which can be used with Python, R, Julia, and scala. This work has been integrated with a python application. This library has built with the best system optimizations to push the computational limits.

Feature score can be generated by considering many times a particular feature appears for the split. This does not necessarily mention the importance of a particular feature. For

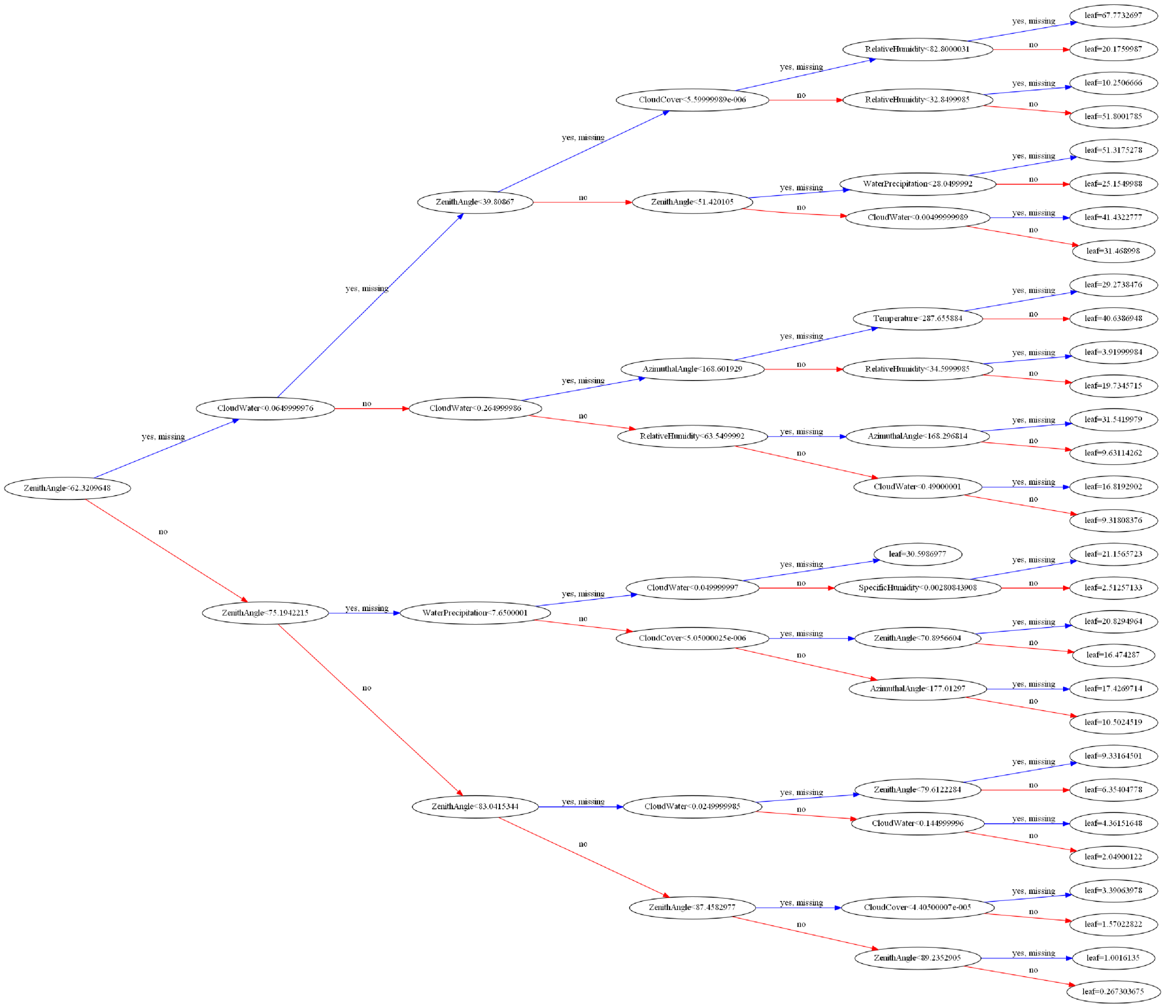


Fig. 3.3: One of the decision tree generated in xg-boost algorithm

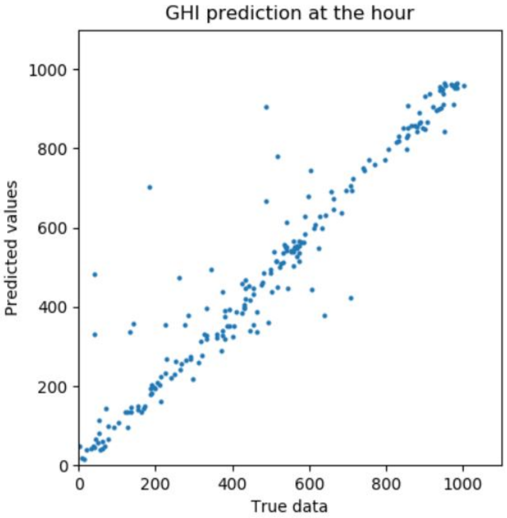


Fig. 3.4: True data Vs predicted data using Xg-boost in Nevada

example, based on an hour of the day a split can be made based on which algorithm will understand the weather is morning or afternoon or evening. But the feature score of the hour is less but still, it is an important feature in the prediction. Feature scores should be looked at as the number of key decisions made based on a particular feature.

Hyper parameter	value
objective	squared error
learning rate	0.09
eta	0.1
early stopping rounds	10
maximum dept	5
subsample	0.9
colsample by tree	0.7
scale pos weight	1

Table 3.1: Hyper parameters for the given xgboost model

In table 3.1 the designed model hyperparameters are shown for the given problem. In objective, one can specify what they want to achieve in the given problem like to build a regressor with a squared error. This also specifies squared error as an evaluation metric. The learning rate defines the influence of the new tree on the existing model values. There

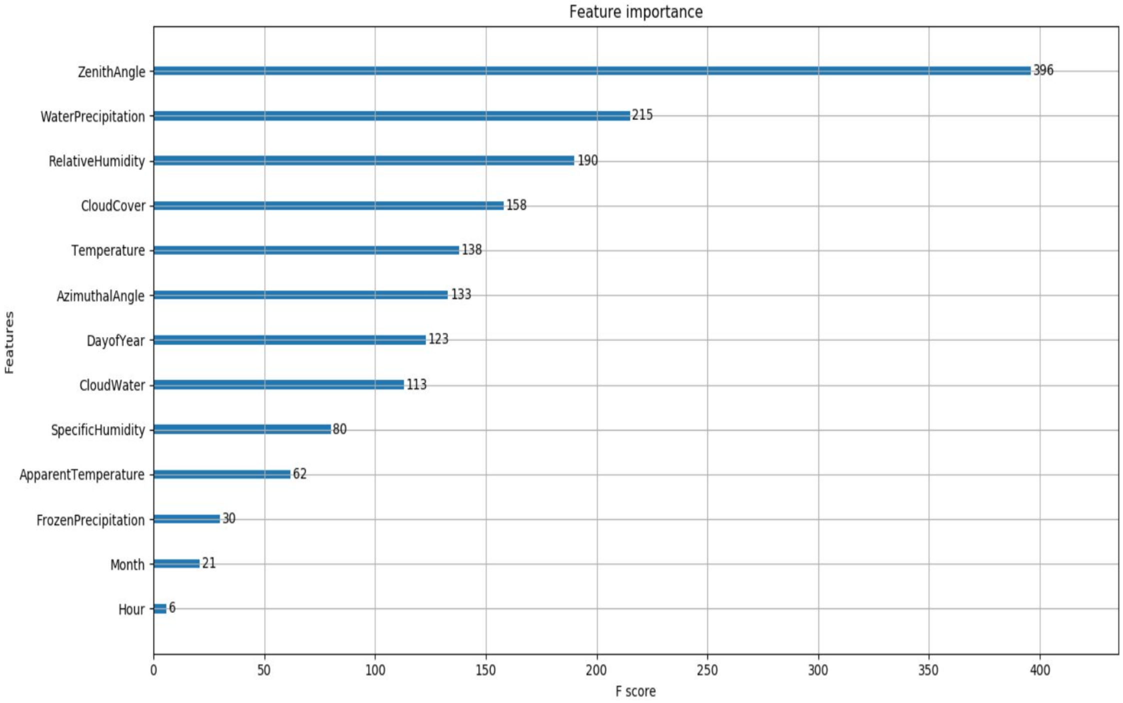


Fig. 3.5: Feature score for xg-boost algorithm for Pensylvania

are different parameters to avoid the overfitting of the function. For example eta, max depth, gamma, min child weight which will be used at different stages of building tree. eta is used to shrink weights, max depth specifies the maximum dept a tree can go to avoid overfitting, gamma is regularization parameter and min child weight is the minimum sum of weights of all observations required in a child. sub-sample is used to choose the random samples from the given training data set before building each tree.col sample by the tree is used to select a fraction of features to be selected randomly.

An efficient set of parameters should be selected among all. The grid search cross-validation method is used for hyperparameter tuning. All the possible values are considered for each parameter. First, the model complexity(eta, gamma, max dept, min child weight) is fixed and then adjusted randomness(sub-sample and col sample by the tree) of the model [15].

3.2.2 cat-boost regression

Cat boost regression algorithm is an advanced version of gradient boost algorithms. A drawback of prediction shift caused by target leakage is identified in gradient boost algorithms including xgboost. To overcome that, a new concept of ordered boosting with a permutation driven alternative is explained [16]. This changes the way how samples are selected to make a tree in each step.

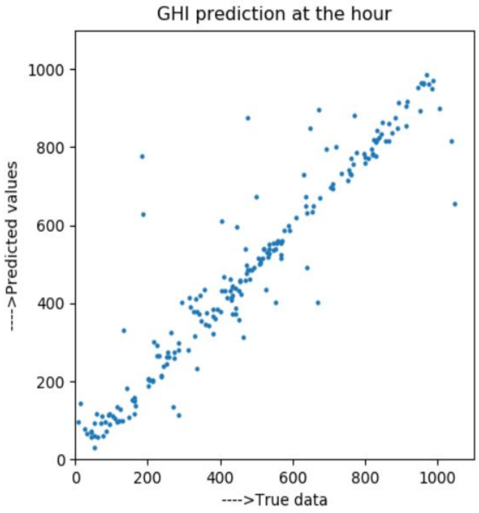


Fig. 3.6: True data Vs predicted data using cat boost in Nevada

Hyper parameter	value
evaluation metric	RMSE
learning rate	0.02
random seed	23
early stopping rounds	10
dept	12
bagging temperature	0.2
od type	Iteration
metric period	25

Table 3.2: Hyper parameters for the given cat model

Table 3.2 shows the hyper parameter selected for cat boost algorithm.Figure 3.7 shows the feature importance by loss function. The value for each feature gives the difference between the metric obtained by the model and the metric obtained with out that feature. This helps to understand the importance of each feature in the algorithm. This more relevant feature importance plot than xg-boost model.

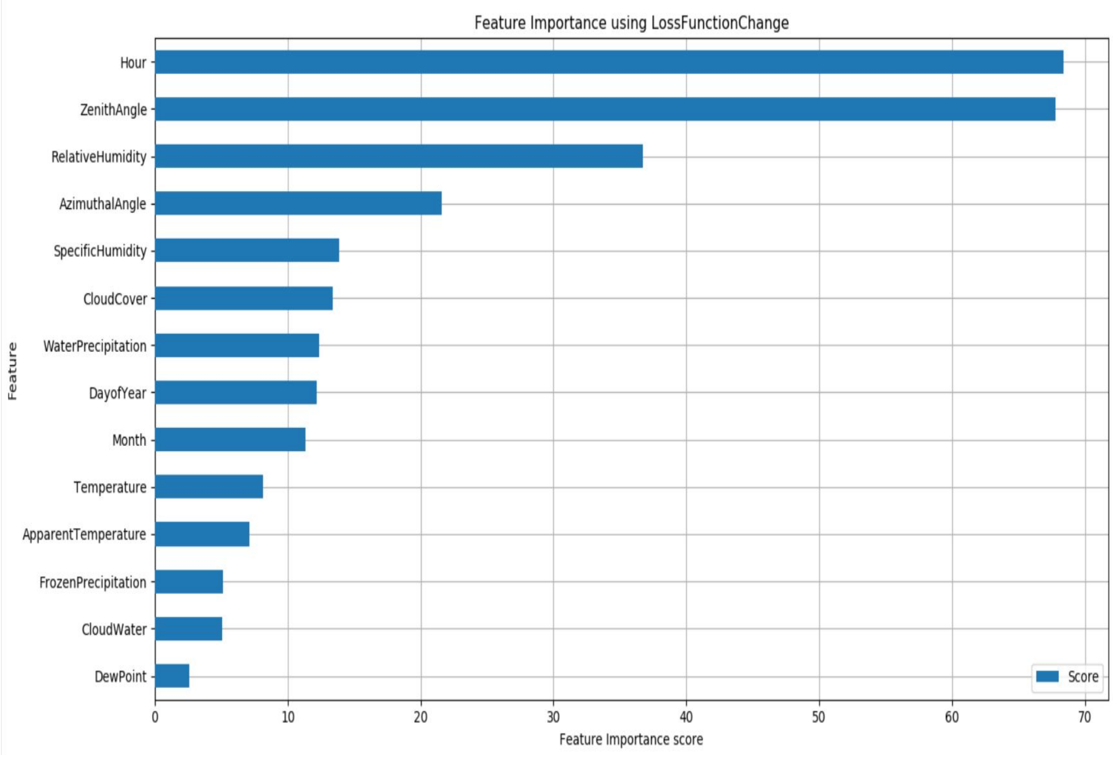


Fig. 3.7: Feature Importance by loss value change in Pennsylvania

Figure 3.8 shows the feature importance by prediction value changes. This gives the how much is the variation on an average if particular feature changes. This plot tells the sensitivity of the feature in the given model.

cat boost algorithm is doing equally well compared to the xg-boost model in terms of accuracy. But the xg-boost model is being used for the experimental study because the amount of training time for the xg-boost model is very less when compared to the cat boost model. Table 3.4 shows the training time for both algorithms.

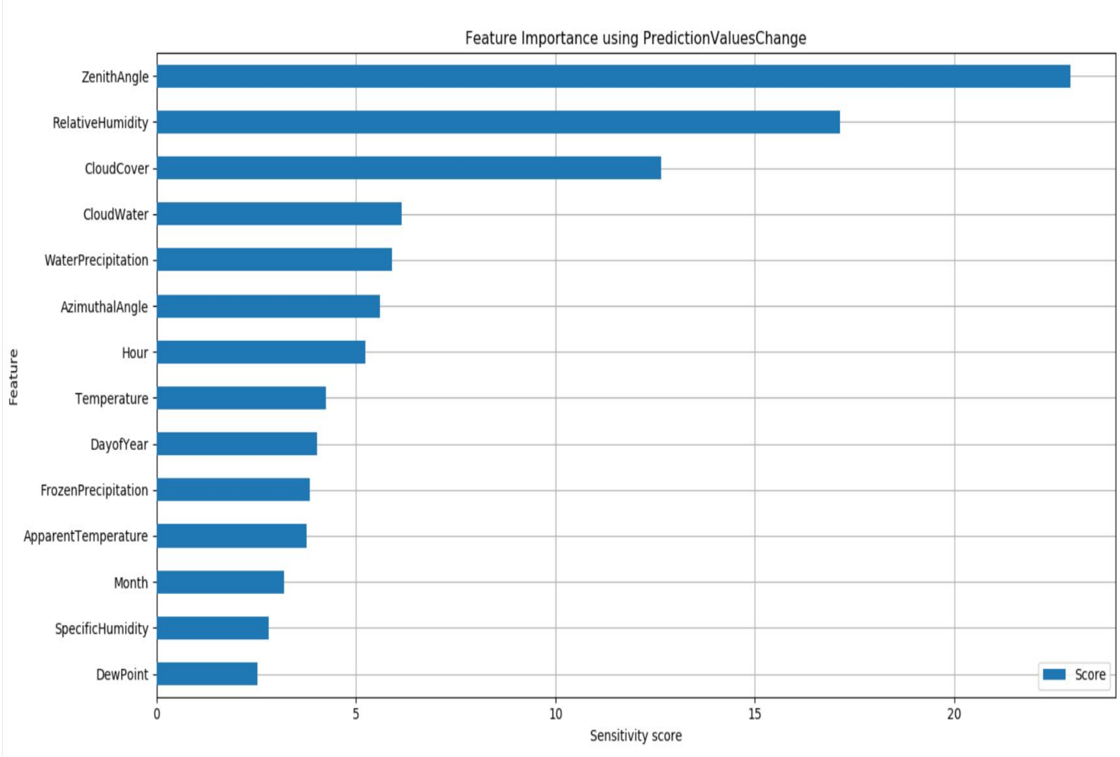


Fig. 3.8: Feature Importance by prediction value changes in Pennsylvania

Location	Algorithm	r-square score	Mean absolute error	Rms error
Rock Springs,Pennsylvania	SVR	0.816	81.2	126.4
Rock Springs,Pennsylvania	xg-boost	0.850	68.4	111.4
Rock Springs,Pennsylvania	cat-boost	0.856	68.0	111.7
Sioux Falls,South Dakota	SVR	0.816	81.2	126.4
Sioux Falls,South Dakota	xg-boost	0.853	67.0	112.9
Sioux Falls,South Dakota	cat-boost	0.854	68.0	112.7
Desert Rock,Nevada	SVR	0.824	87.4	118.2
Desert Rock,Nevada	xg-boost	0.93	42.4	73.9
Desert Rock,Nevada	cat boost	0.93	42.7	74.6

Table 3.3: Comparing errors for different algorithms and different locations

Algorithm	Training time in seconds
xg-boost	0.14
cat-boost	144

Table 3.4: Comparing training time for two algorithms

CHAPTER 4
RESULTS

In the previous chapter, it has been proven that the xg-boost algorithm could predict solar irradiance. In this chapter validation of the findings in different scenarios is understood. It also shows how different parameters influence the prediction. One is to check whether the prediction accuracy is the same in all the locations around the US. Another is how long prediction is reliable. And an experiment study needs to be conducted on how solar irradiance prediction is helping to make a profit for an EV charging station.

4.1 Prediction analysis

Solar irradiance accuracy is checked in different scenarios in the following section.

4.1.1 Predicting at different locations

A one can predict solar irradiance at any location in the world if the historical GHI is available to train the model. This developed xg-boost model has experimented at three different locations. Those are Desert Rock in Nevada, Sioux Falls in South Dakota, Rock Springs in Pennsylvania. Nevada produces the best results with an r-square score of 0.94 followed by South Dakota and Pennsylvania with 0.85. As the weather parameters from GFS are available with 0.5 latitude and longitude resolutions, the values have been averaged over the entire square. So the accuracy of GHI depends on how much particular location climate will correlate with the average value of 50 km. Nevada got good accuracy and compared it to Pennsylvania and South Dakota. Figures 4.1, 4.2 and 4.3 shows prediction for random two hundred timestamps which is not seen in training dataset.

4.1.2 Predicting hours ahead

As weather forecast data is available for 16 days. It is important to know how long

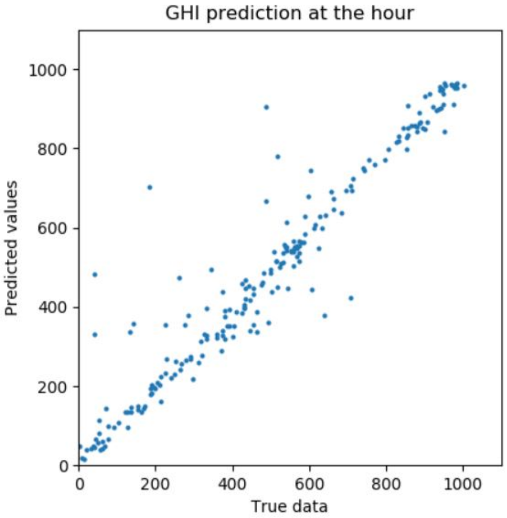


Fig. 4.1: True data versus Predicted data in Nevada

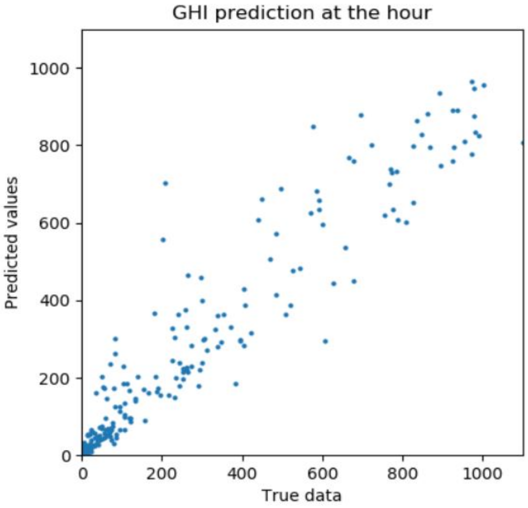


Fig. 4.2: True data versus Predicted data in Pennsylvania

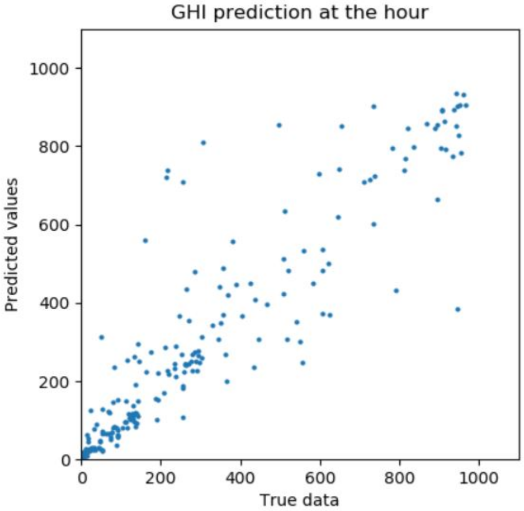


Fig. 4.3: True data versus Predicted data in South Dakota

the model can predict GHI accurately. Figure 4.4 is plotted to know the trends of average mean absolute error for as long as 10 days ahead prediction. Each dot in the figure shows the mean absolute error of the averaged value of 100 timestamps prediction. That is the first blue dot in the figure show the averaged mean absolute error for 100 timestamps for zeroth hour prediction. The second blue dot shows the average mean absolute error for three hours ahead prediction and so on until 240 hours for every three hours. The x-axis shows how many hours ahead the value is predicted and the y-axis shows the averaged mean absolute error. There is an exponential growth of error, as one is predicting further ahead. The training dataset is for every six hours interval but practically weather forecast is available with a resolution of three hours this helps to do a solar irradiance prediction for every three hours. The model has been trained with the sun position which helps it to predict for every three-hour resolution though trained for every six hours. This is clearly shown in the figure, the error is little high for three, nine, fifteen hours ahead compared to zeroth, six, twelve hours ahead prediction. When a polynomial curve is fit the error is increasing exponentially. So to optimize the error 48 hours prediction is considered for further experimental study.

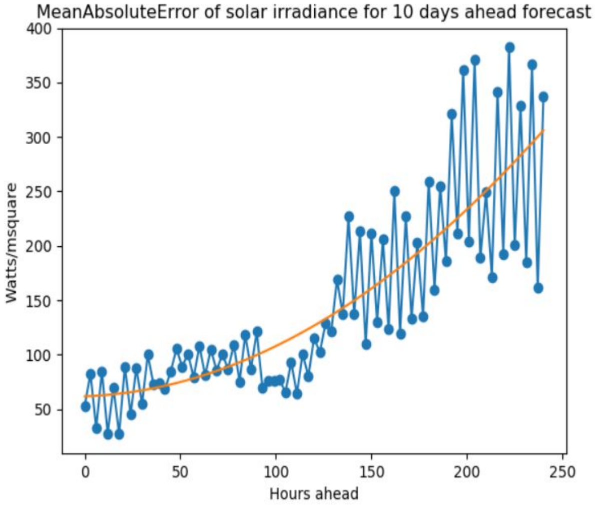


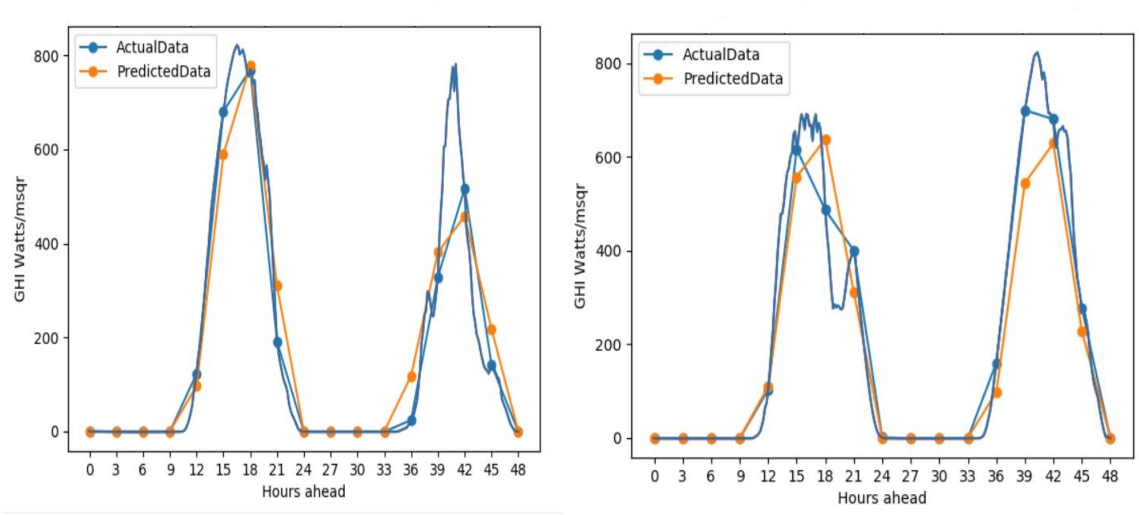
Fig. 4.4: Trends of mean absolute error of solar irradiance from October 1-17 days in Sioux Falls to hours ahead forecast

This prediction will be helpful to get the best estimation of solar energy available in the near future. In figures 4.5 4.6 different scenarios of prediction for 48 hours are shown. Prediction is compared with the recorded GHI values. Blue lines represent the true values of the GHI and the orange line shows the predicted values. The y-axis shows GHI in Watts/msquare and the x-axis shows the number of hours ahead that is zero to forty-eight hours from the given time. In figure 4.5a predicted curve is completely different from the true curve on the second day as the day has sunlight for only 3 hours. The model misses recording any value between those three hours. This is one of the scenarios where the prediction would give completely different results. These kinds of scenarios are unavoidable as the weather forecast resolution is three hours. A sudden drop in solar irradiance can occur because of the passing cloud. In figure 4.5b is an example of one such scenario. In 4.5b there is a drop in solar irradiance for about an hour. This can not be included in the predicted solar irradiance curve because of low resolution. The given prediction cannot exactly trace the ramps of the solar curve. But if the cloud is for a long period it predicts decently as in figure 4.6a. In this figure, it is a cloudy day and prediction has been a good estimate. 4.6b shows the perfect prediction on a sunny day.

Weather forecast is available at 0,3,6,9,12,15,18,21 hours of the day in UTC. Figure 4.5 plotted in Eastern timezone and figure 4.6 plotted in central timezone. In eastern time zone prediction can be done at 8 am, 11 am, 2 pm, 5 pm during the day. There is a high chance of missing a peak at noon. In central timezone weather forecast is available at 7 am, 10 am, 1 pm 4 pm. In central timezone, there are fewer peak cutoffs compared to other timezones. Timezone also plays a role in accuracy when one samples the solar irradiance curve for high frequency.

4.2 Solar energy prediction applications

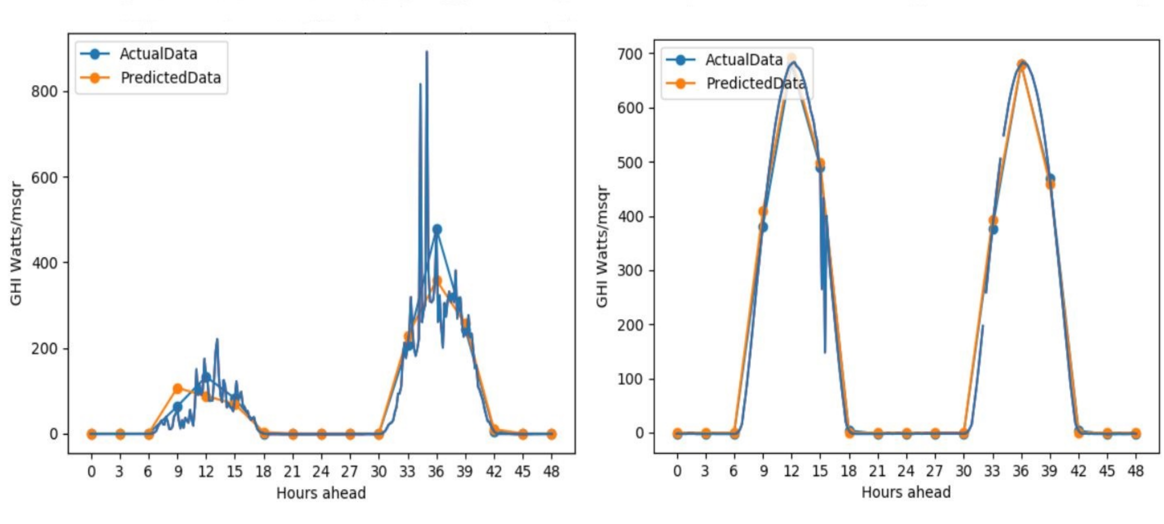
Solar energy is not self-sufficient to address all the energy demands as it is available only for eight to ten hours a day. Hence integrated systems are required to handle the energy storage and demand-supply. Demand supply could be anything either we are powering an industry or electrical vehicles. Wherever there is a requirement in power solar energy



(a) September 1,2019 at 8 pm

(b) September 9,2019 at 8 pm

Fig. 4.5: GHI prediction for 48 hours in Rock springs, Penn State



(a) October 2,2019 at 2am

(b) October 6,2019 at 2am

Fig. 4.6: GHI prediction for 48 hours at 2 am in Sioux Falls, South Dakota

storage should be planned to utilize the power efficiently.

One such integrated system is microgrid. To operate microgrid efficiently solar energy prediction and load prediction is needed. One prediction is provided a learner like reinforcement learner can give optimal charge or discharge schedule to minimize the cost. While making decisions it would consider the price of energy at a given point and look forward to the energy available and energy demands it decides on when to buy electricity and when sell to the power grid. These intelligent decisions by reinforcement learners manage microgrid with fewer maintenance costs.

Solar energy can also be helpful to set the prices of energy in the future. If the amount of solar energy production may be estimated a week beforehand along with estimated demand, an energy trader can set a price for that energy and increase their likelihood of profit. As a demonstration, an experimental study has been conducted on one of the applications of solar energy prediction and load prediction for an EL charging station.

4.3 Experimental study on EV load optimization

To understand the use of predicting solar energy a specific example is considered. This study is conducted from the perspective of an electric vehicle charging station.

Consider an electric vehicle charging station with a solar panel that has a capacity of 60 kiloWatts. It is designed to serve the demands of electric vehicles by only solar energy. A wise decision should be made about storing power in the battery to maximize profit. A reinforcement learning system is implemented using the value iteration algorithm [17] which helps to make these decisions.

A prediction of solar energy and load is provided for 48 hours with a one-hour resolution. Solar prediction and load prediction is feed to the optimizer to determine actions after every one hour, which are fed to the power management system to execute them. Figure 4.7 shows the integrated system.

The load is measured in kWh and is predicted using historical data. Rocky mountain power data is used for this study based on data collected from EV charging stations deployed around the Salt Lake City area. [18]

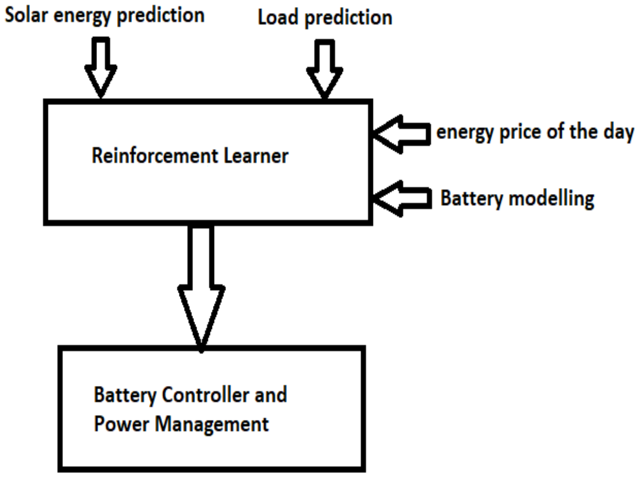


Fig. 4.7: Integrated system for an EV fuel station optimization

Solar energy generated should be support as many EVs as possible to gain profits is the main goal of this analysis. The algorithm optimizes the action sequence of battery charging and discharge based on the source and demand predictions. Time of day price of electricity and battery modeling are also considered to generate an action sequence. Action sequence dictates when to charge the battery and when to discharge the battery. In 4.8 third row shows the action sequence for 48 hours. The amount of profit from the generated action sequence has also been calculated. Action is generated for every one hour and solar and load predictions are also with one-hour resolution.

Figure 4.8 illustrates the operation of the microgrid optimizer as four plots aligned on the same horizontal axis of time. The first plot shows predicted solar energy generated (top in orange) and predicted demand (in grey as negative values). The second plot shows the amount of energy stored in the battery for future needs at a given point of time. The third plot shows the action sequence to charge or discharge the battery. The fourth plot shows the amount of energy wasted because of lack of battery capacity and it also shows the amount of energy charging station failed to supply. The fourth plot shows the profit EV charging station earns by following this optimized action sequence until that moment.

Experiment study is carried on this algorithm by mapping different scenarios of solar prediction to the dollar value and studied how much is the difference between profits for predicted solar and load and true solar and loads.

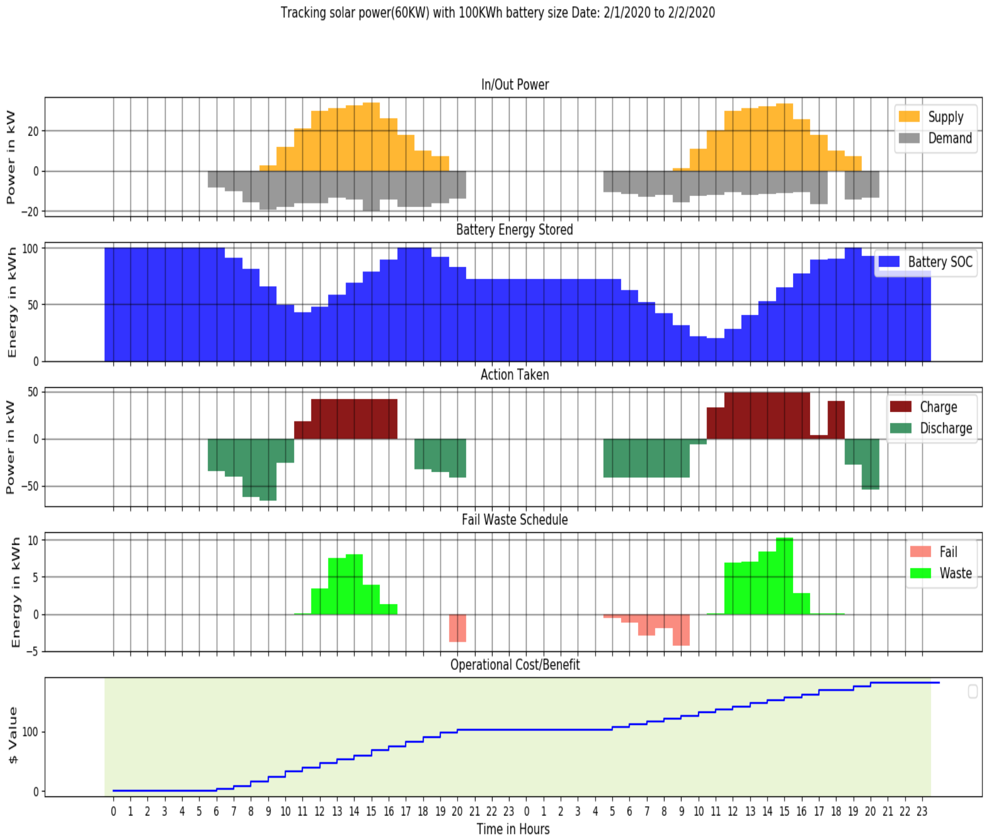


Fig. 4.8: Optimized charge-discharge schedule of battery suggested by RL-learner for a one-time stamp with 60 kW tracking solar and load predictions and also shows the optimized profit generated, by specifying the times when the charging station fails to supply the demand

4.3.1 Experimental setup

There two types of solar panels one is fixed solar panels and the other is tracking solar. Tracking solar changes its rotation to face the sun throughout the day. Different configurations of solar panel configurations and battery capacities are considered. For the given experiment study 60 kW tracking solar and 100kW battery has been used. This

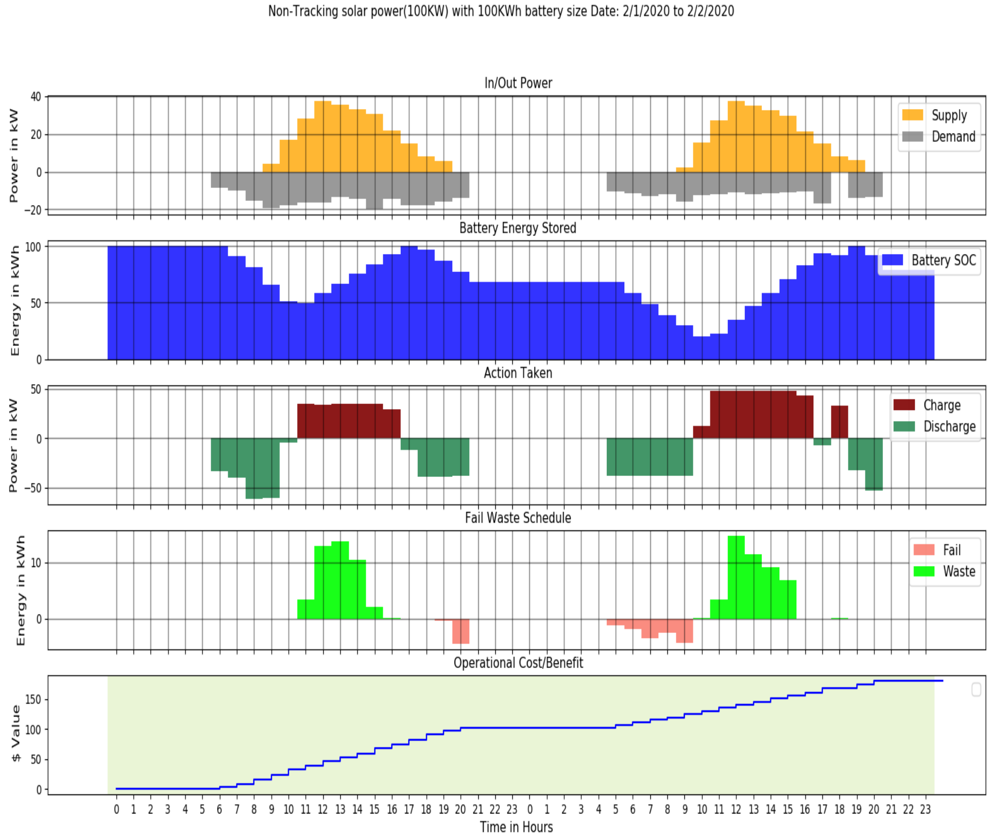


Fig. 4.9: Optimized charge-discharge schedule of 100kW battery suggested by RL-learner for a one-time stamp with 100 kW non tracking solar and load predictions and also shows the optimized profit generated, by specifying the times when the charging station fails to supply the demand

configuration was compared to an alliterative configuration using a 100kW fixed solar and 100kW battery. Figure 4.9 shows the charge-discharge schedule of 100kW battery with 100kW fixed solar panel and 4.8 shows the charge-discharge schedule of 100kW battery with 60kW tracking solar. Considering the power wastage and failing to supply load both are working equally well. Hence for this experiment study 60kW tracking solar as installation cost will be less.

4.3.2 Profits made in different seasons of the year in Nevada Desert Falls

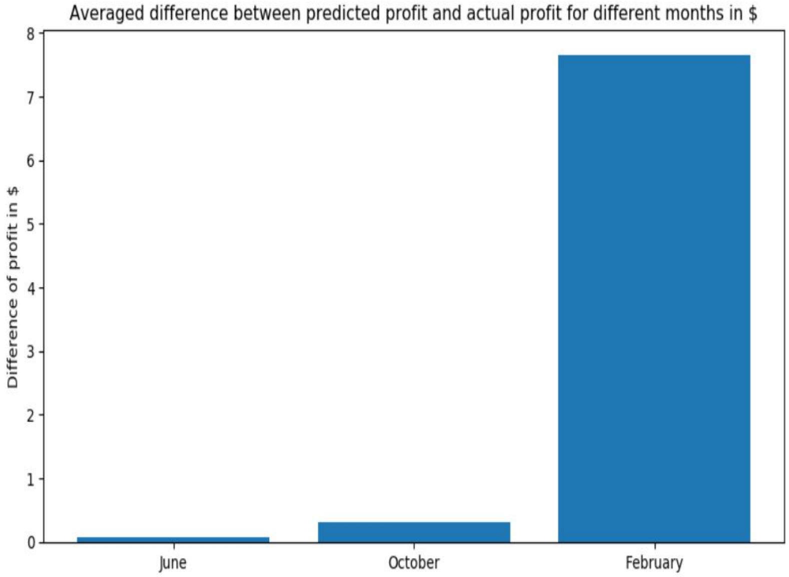


Fig. 4.10: Averaged difference in profit between forecasted profit and true profit in different months in Nevada

Forecasted profit with predicted load and solar has been compared with the true profit that is obtained for the true data in the given months. The difference between the two profits is averaged over one month is plotted in Figure 4.10. Different months are considered to show the effect of the season's summer, fall, and winter. Results show that summer and fall months are doing better in predicting the actual profit compared to winter months.

4.3.3 Profit made by EV charging station at different locations in United States

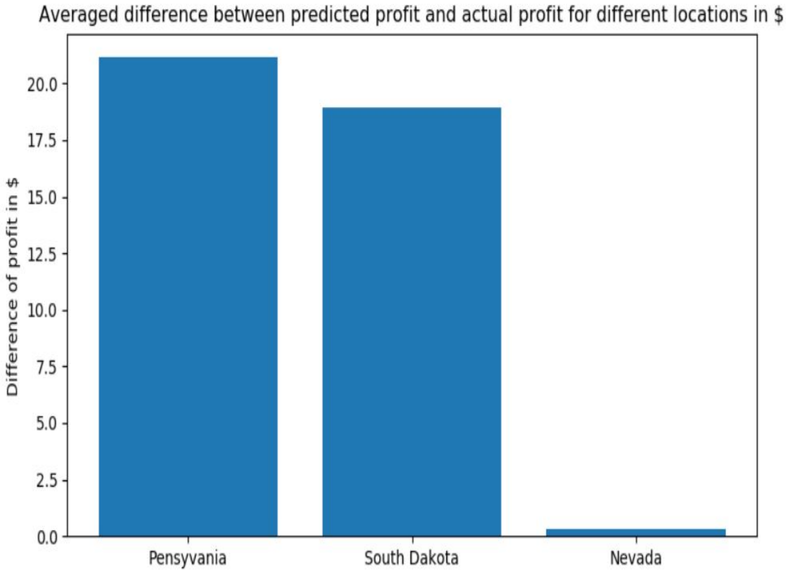


Fig. 4.11: Averaged difference in profit between forecasted profit and true profit at different locations in United States

Figure 4.11 show the averaged difference between predicted profit and true profit for different locations in the United States in the month of October. Profit is calculated considering the same load at all the locations to understand how solar irradiance accuracy is affecting profit. Solar irradiance prediction has a high accuracy of 94% in Nevada and it around 85% in South Dakota and Pennsylvania. The prediction of profit is directly proportional to the accuracy of solar irradiance.

CHAPTER 5
CONCLUSION AND FUTURE WORK

This chapter concludes with the findings in the work and future work to improve the prediction of solar irradiance at a higher time and spatial resolution.

5.1 Conclusion

This work is helpful to predict solar irradiance for about two days ahead with three hours resolution. As numerical weather forecast data is used it is available all over the world and can predict at any location. As solar irradiance varies with many environmental parameters this problem is treated as a regression problem. SVR and gradient boost algorithms like xg-boost and cat-boost were applied to the given problem. xg-boost performs well in terms of accuracy and speed with an accuracy of between 85% and 94%. Error analysis shows that error is less until 48 hours ahead and after that, the error tends to increase.

The accuracy of the model is dependent on the location. As the spatial resolution considered is 0.5-degree latitude and longitude. The atmosphere parameters value is averaged over 50 km. So the variability of atmospheric parameters of a particular location determines that possible accuracy at that location. Highly accurate predictions were obtained in Nevada compared to Pennsylvania and South Dakota because of this. Additionally, an artifact of the phase of the 6 hour sampling period influenced accuracy in eastern time zones. Here, the samples missed the time of peak solar energy production (around noon), so overall energy prediction accuracy was lower.

The sampled predicted solar irradiance curve is converted to solar power and fed to the microgrid optimizer along with predicted load that gives optimizing charge-discharge schedule of a battery to make profits. Solar irradiance is mapped to solar power by two different methods depending on whether the panels are fixed or track the sun. A configuration of a 60kW tracking solar panel was shown to perform equally well compared to the 100kW fixed solar panel in terms of supplying the demand. For 60kW tracking solar variation in profit, an error has been observed at different locations and different seasons. By 4.11, 4.10 we can conclude that profit is prone to error in winter months compared to summer and fall. The profit prediction is also proportional to solar irradiance accuracy. As Nevada has high accuracy, RL learner predicted the profit two days ahead with less than a dollar error.

5.2 Future work

This work has been done with the mesoscale numerical weather forecast, which has a spatial resolution of about 12km. Training data considered in this work is over three and a half years, which can be extended so that the predicted model would consider more weather patterns over a longer period. The error appears to increase significantly after 48 hours mainly because of the numerical weather forecast errors grow similarly. So as weather forecasts improve, then the predictions will also improve. Figure 2.1 shows the continuous efforts to make the weather forecast better that will result in better solar irradiance prediction in the future.

Experiments were run with electric vehicle optimizer and may be extended to microgrid optimization that will greatly benefit the industrial sector and increase the solar energy market that would ultimately reduce the dependency on non-renewable energy sources. This will reduce the emission of greenhouse effect gases and help humanity reduce the impact of global climate change.

REFERENCES

[1] P. L. J. B. N. S. Hadja Maïmouna Diagne, Mathieu David, “Review of solar irradiance forecasting methods and a proposition for small-scale insular grids,” in *Renewable and sustainable energy reviews*, Dec. 2014.

[2] P. B. Gilbert Brunet and A. Thorpe, “The quiet revolution of numerical weather prediction,” 2015.

[3] Indresh Battacharya. (2018) Support vector regression or svr. [Online]. Available: https://medium.com/coinmonks/support-vector-regression-or-svr-8eb3acf6d0ff

[4] K. A. R. H. K. T. Keii Gi, Fuminori Sano, “Potential contribution of fusion power generation to low-carbon development under the paris agreement and associated uncertainties,” pp. 1–2, 2019.

[5] N. S. J. L. John Mathe, Nina Miolane, “Pvnet: A lrcn architecture for spatio-temporal photovoltaic power forecasting using numerical weather predictions,” Apr. 2019.

[6] H. J. Elise Duponta, Rembrandt Koppelaarb, “Global available solar energy under physical and energy return on investment constraints,” 2020.

[7] N. J. N. Sujan Ghimire, Ravinesh C.Deo, “Global solar radiation prediction by ann integrated with european centre for medium range weather forecast fields in solar rich cities of queensland australia,” 2019.

[8] N. S. C. D. M. P. Robert Blaga, Andreea Sabadus, “A current perspective on the accuracy of incoming solar energy forecasting,” in *Progress in energy and combustion science*, Oct. 2018.

[9] S. K. M.-L. N. C. P. F. M. A. F. Cyril Voyant, Gilles Nottan, “Machine learning methods for solar radiation forecasting: A review,” 2017.

[10] K. Bhargava, “Estimation and adaptive online correction systamatic errors in the global forecast system using analysis increments,” Master’s thesis, University of Maryland, College Park, 2019.

[11] D. Neilsen, “Tree boosting with xg-boost,” Master’s thesis, Norwegian university of science and technology, 2019.

[12] D. I. Navin Sharma, Pranshu Sharma and P. Shenoy, “Predicting solar generation from weather forecasts using machine learning,” 2011.

[13] Alexey Natekin, Alois Knoll. (2013) Gradient boosting machines, a tutorial. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fnbot.2013.00021/full

[14] C. G. Tianqi Chen, “Xg-boost: A scalable tree boosting system,” 2016.

[15] xgboost-developers. (2019) xg-boost library documentation. [Online]. Available: https://xgboost.readthedocs.io/en/latest/index.html

[16] A. V. A. V. D. A. G. Liudmila Prokhorenkova, Gleb Gusev, "Cat boost: unbiased boosting with categorical features," 2019.

[17] R. S.Sutton and A. G.Barto, *Reinforcement Learning - An Introduction.*

[18] A. Neema, "Load forecasting analysis using contextual data and integration with microgrids used for off grid ev charging stations," Master's thesis, Utah State University, 2020.