CS 519 Project Report

Randall Woodall
New Mexico State University
hossrw@nmsu.edu

Emrah Sariboz
New Mexico State University
emrah@nmsu.edu

Aya Elsayed
New Mexico State University
aynasser@nmsu.edu

ACM Reference Format:
Randall Woodall, Emrah Sariboz, and Aya Elsayed. 2020. CS 519 Project Report. In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 PROBLEM STATEMENT

Prediction of power production for solar panels from weather information.

1.1 Motivation

Since starting up generators is costly, having an idea of how much power a utility would need to produce can help the utility to reduce costs of generation. There exists an abundance of data in relating to solar production, and an even greater abundance relating to weather and forecasting. Combining weather history, weather forecasting, and information about solar production could lead to a schedule of solar power production that would allow, in the long run, utilities to create more optimal generator schedules and power rates. This type of scheduling would help the consumer to save money by allowing the producer to avoid startup costs and idling costs. Because of this, new and more accurate methods will be valuable to all parties involved in electrical power. Power scheduling is not a new problem, and there have been some successful models to describe renewable schedules [1]. Our project is to attempt to use regression models to minimize the error from the model, which could then be used with higher confidence in conjunction with more accurate weather models.

1.2 Direct Problem Definition

Given solar production data in the form of power, and also given historical weather data including wind, temperature, humidity and pressure, can we make semi-accurate predictions about the solar data at the time of the weather data? To restate this problem in fewer words, can you use current state of the weather to predict current output of a solar panel? If this is the case, then a day-ahead power forecast should be directly correlated to a day-ahead weather forecast. The goal of this project is to show that, given current data, we can predict what a solar panel produces, and by logical extension, if we can predict what it will produce tomorrow.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
Conference'17, July 2017, Washington, DC, USA
© 2020 Association for Computing Machinery.
ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

1.3 Problem Limitations

We acknowledge that we have a hard limit on the accuracy of any day-ahead power predictions, based in the uncertainty of day-ahead weather predictions. Going back to statistics, this limit on accuracy and, therefore, the introduced error, goes back to the statement, "Garbage in, garbage out". Due to this, we will eliminate the space for introduced error by restraining our semester project to include only the historical data, where we have known values and actual measurements. If we can assume a normal distribution around the predicted weather values wherein the actual weather values will fall, then we can incorporate this into our prediction for power produced to give a normal distribution showing what power will probably be.

1.4 Related Work

We were able to find one similar work from 2017 which attempted to predict power in the form of irradiance using SVR [2]. We attempted to replicate these results with a much larger dataset, but obtained significantly lower accuracy than was reported.

In a second paper published in 2018, we found compared power prediction using Linear Regression and a regression tree [3]. This paper helped confirm that relatively few weather features go into accurate predictions for PV power. This paper only used two machine learning methods.

2 DATA SOURCES

2.1 Solar Power Data

Our solar data is being collected by direct monitoring of solar panels on the roof of Thomas and Brown hall. This data is being collected on a 5 second interval for use with another experiment, which is a far higher frequency than we can retrieve weather data at. Specifically, our data is the power produced by a set of four panels. Since there are four panels hooked up to a three phase output, we simply added the three phase output power of all the panels together to get the total power out. To get hourly three phase powers to line up with weather data, we took a rolling average over the hourly values to pair up with weather data.

2.2 Weather Station Data

By using our own weather station on Thomas and Brown hall's roof, We are getting a time stamped weather data. Along with the time, we are getting other 26 attribute each five minutes. The twenty- six attributes are solarRadiation, uvHigh, winddirAvg, humidityHigh, humidityLow, humidityAvg, qcStatus, ,tempLow,tempAvg, windspeedHigh, windgustLow, windspeedAvg, dewptHigh, dewptLow, dewptAvg, windchillHigh, windchillAvg, heatindexHigh, heatindexLow, heatindexAvg, pressureMax, pressureMin, pressureTrend, precipRate and precipTotal. For all of our data, we have utilized the

solar panels and weather station situated on top of Thomas and Brown hall.

2.3 Data Preprocessing

Both data sets (solar power data and weather station data) have timestamp as the first column. For each day there is a file named under the date of the day for both data sets. To merge the two data sets together. Considering each two data files having the same we compare each timestamp in the weather station data with the timestamp with solar power data.

Before we apply any pre-processing technique, the shape of the dataset was (12949, 28) including the time stamp. Also, we disregarded the time column for now. At some point in the day, a solar panel couldn't get the power data. Thus, some values of the power attribute had NaN values. The first thing we did was to remove these rows and the shape of the resulted dataset was (10004, 28). And then, we used the Sklearn SimpleImputer library to impute NaN values with the mean of the corresponding columns. The reason we decided to impute rather than dropping is only 10 percent of this column was missing. Dropping these columns could result in serious information losses. The details of the preprocessing steps can be found in the GitHub repository under Stage 3.

3 ALGORITHMS TESTED

3.1 Linear Regression

For linear regression analysis, we first used a correlation matrix to find those features that have a high correlation with the power attribute as shown in Figure 1. In other words, we first conducted a feature section. Upon this process, we have left with two features. To evaluate the model on the unseen dataset, we have split the dataset into training and testing (30 percent) using the sklearn model selection library. By applying linear regression, we were able to get an r^2 score of 0.79 on the training dataset and 0.80 on the testing dataset. The distributions of the solarRadiation and tempHigh columns with the power attribute is plotted in Figure 2. It turns out that both features are negatively correlated with the power feature. R2 solely is not a sufficient way to evaluate the regression model. Thus, we have added 2 more evaluation metrics, namely MAE and RMSE. Using a linear regression model, we were able to get 74.4 MAE and 129.89 RMSE on the training dataset and 78 MAE and 135 RMSE on the testing dataset.

3.2 Polynomial Regression

To use polynomial regression, we first applied PCA to reduce down to 2-dimensions, this is the only algorithm in which we reduced the features. This was simply to reduce computation time, noting that most information will probably be contained in few features. This theory is proven with the correlation matrix showing a -.89 correlation between solarRadiation-power and uvHigh-power. Using this technique, we were able to get an r^2 score of 0.819 on testing data, with .821 on training data. Although this is better than standard linear regression, we are hoping to do better than this with a different algorithm. MAE was 75.557 on training data, and 73.535 on testing data. RMSE was 122.832 on training data, and 118.641 on testing data.

3.3 Random Forest Regression

Applying Random Forest Regressor from scikit learn library with 100 estimator and a random state of one. By splitting the data by 30 percent for testing data. Achieving a R^2 score for train data of 0.997 and a r2 score for test data of 0.981.

3.4 Decision Tree

Applying Decision Tree Regressor from scikit learn library With max depth 15 and a zero random state came out slightly worse than the Random Forest Regressor. This is entirely expected, but also had the unintentional side-effect of showing us that there is very little gain to training many trees in this case. This in itself could be an entire project, and for our purposes it suffices that the Random Forest Regressor is somewhat better than the Decision Tree Regressor.

3.5 Neural Network Multi-layer Perceptron Regression

We can see from Table 1 that this basic Neural Network is effective, but not the most effective method that we tried. It is valuable to note that this method had very little drop in all accuracy metrics when compared to some of the others. This is especially true versus the Random Forest Regressor, which doubled MAE and RMSE.

3.6 Support Vector Regression

We tested SVR since we saw good results in a paper from several years ago [2]. Our results showed that this was not a particularly good method, more succintly that it was the worst method we tested. This would suggest that any of the other methods would be a better choice for this analysis, and the SVR was probably used in this paper due to popularity or familiarity.

4 RESULTS

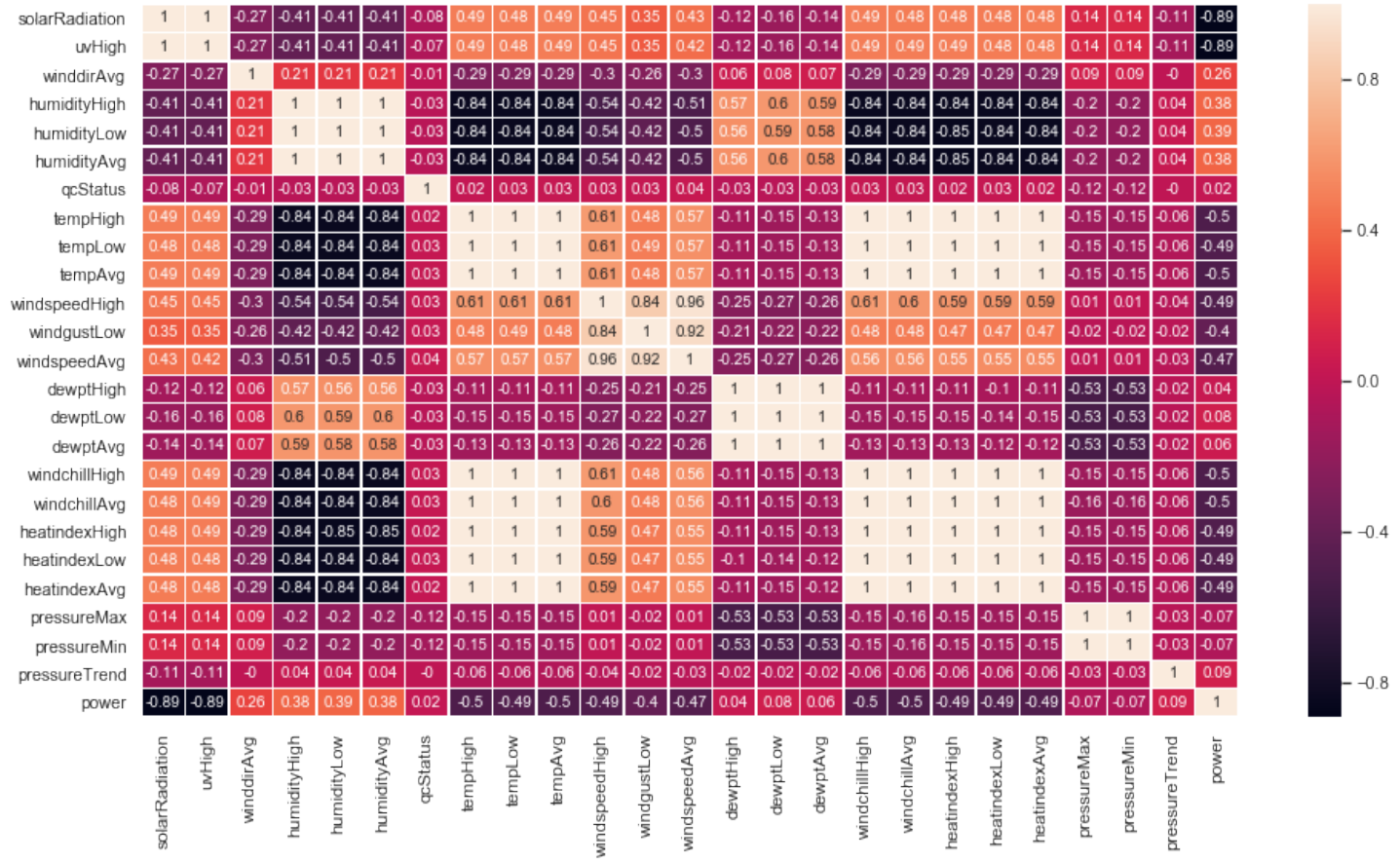
To evaluate a regression model, Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) are the two most popular metrics for continuous variables. Using R squared (R^2) along with MAE and RMSE as metrics to analysis the performance of the different used regression models. MAE equation 1 shows the average of absolute error where n is the number of instance, y_j is true value and $\hat{y}_j$ is the predicted value. RMSE equation 2 shows the sample standard deviation of the differences between predicted values and true values. R^2 Equation 3 where The numerator is MSE (average of the squares of the residuals) and the denominator is the variance in predicted values.

$$MAE = \frac{1}{n}\sum_{j=1}^{n}|y_j - \hat{y}_j| \tag{1}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(y_j - \hat{y}_j)^2} \tag{2}$$

$$R^2 = 1 - \frac{\sum_{j=1}^{n}(y_j - \hat{y}_j)^2}{\sum_{j=1}^{n}(y_j - \bar{y_j})^2} \tag{3}$$

Figure 1: Correlation Matrix for Linear Regression Analysis



A good regression model has a high R^2, low MAE and low RMSE. Table 1 discloses MAE, RMSE and R^2 for Linear, polynomial(Poly), Random Forest(RF), Decision Tree(DT), Neural Network Multi-layer Perceptron (MLP) and Support Vector Regressor(SVR).

Metrics	Models					
	Linear	Poly	RF	DT	MLP	SVR
	(1) For training Data					
MAE	70.369	75.557	10.479	6.657	24.342	66.265
RMSE	125.58	122.8	24.395	26.374	52.037	125.35
R^2	0.877	0.821	0.995	0.995	0.979	0.8766
	(2) For testing Data					
MAE	69.713	73.535	28.802	31.073	31.525	68.335
RMSE	126.33	118.6	64.956	83.987	65.948	129.46
R^2	0.877	0.819	0.968	0.946	0.967	0.872

Table 1: Mean Absolute Error, Root Mean square Error and R^2 for different regression models .

Random forest regression models is the best model with R^2 0.968, MAE 28.802 and RMSE 4219.338 for the testing data.

5 FUTURE WORK

We would like to expand this in the future to include a complete study encompassing weather forecasting including error analysis and a total comparison of predicted values for power produced versus actual power produced. We expect that a complete analysis would be of substantial benefit to the field, especially with complete descriptions of steps taken to mitigate error for the weather forecasts.

REFERENCES

[1] S. Xia, Z. Ding, T. Du, D. Zhang, and H. Yin. Multi-time scale coordinated scheduling for the combined system of wind power, photovoltaic, thermal generator, hydro pumped storage and batteries. In *2019 IEEE Industry Applications Society Annual Meeting*, pages 1–8, 2019.

[2] M. Z. Hassan, M. E. K. Ali, A. B. M. S. Ali, and J. Kumar. Forecasting day-ahead solar radiation using machine learning approach. In *2017 4th Asia-Pacific World Congress on Computer Science and Engineering (APWC on CSE)*, pages 252–258, 2017.

[3] K. Benhmed, F. Touati, M. Al-Hitmi, N. A. Chowdhury, A. S. P. Gonzales, Y. Qiblawey, and M. Benammar. Pv power prediction in qatar based on machine learning approach. In *2018 6th International Renewable and Sustainable Energy Conference (IRSEC)*, pages 1–4, 2018.