

## Victor Mukora

*Virginia Polytechnic Institute and State University*

## December 31, 2021

**Abstract**

Previous work has been performed on the influence of factors, such as temperature or humidity, on the impact of solar panel efficiency. However, there is a gap in knowledge related to, how the various environmental conditions interrelate to affect solar panel output real-time. By testing, creating, and validating a model that correlates several environmental predictor variables known to impact solar panels, improvements in solar panel adjustments and optimizations can be performed. In this project, prediction models including artificial neural networks (ANN), multiple linear regression (MLR), and regularized linear models like elastic net, ridge, and lasso were used for relating environmental variables such as temperature, humidity, or rain rate to the total solar energy produced. Data containing twenty-nine different weather conditions and their respective solar power outputs were obtained from the UK Power Networks and incorporated as the principal datasets for the models. To check linear model assumptions like normality of residuals or heteroscedasticity for models like MLR, several functions provided from a model performance package in R verified whether the assumptions for the model were met. Using Mean Squared Error (MSE) as the benchmark, cross-validation was utilized for hyperparameter optimization and hold-out validation for overall predictive model selection. MLR resulted in the highest performing model, but the regularized models were the most interpretable models in terms of how conditions affected solar power output. Both predicting and investigating the nature of how weather conditions impact solar energy enable us to determine the most cost-effective optimizations for enhancing solar panels and what environmental factors are the most important in determining future locations for solar panel farms.

# Introduction

Many developing world countries lack access to electricity, resulting in disadvantages in every aspect of their life. In 2014, according to the International Energy Agency (IEA), almost 1.3 billion people had no access to electricity and 2.7 billion individuals relied on use of biomass like wood or charcoal for cooking [1]. An overwhelming 97% of these people reside in sub-Saharan Africa and developing parts of Asia, with 84% living in rural areas [1]. Without adequate access to electricity, the ability to study and pursue an education, to cook, have clean drinking water, and technology to yield good crops are all greatly diminished. Notably, these areas are the ones which receive the strongest sunlight and have the most potential for solar power production. Consequently, because sunlight is readily available, solar panels are being viewed as a much better, reliable, and more promising alternative to other energy-harnessing methods. Moreover, even on a global scale, due to sunlight’s ubiquity, the solar industry is expected to produce a significant portion of total worldwide consumed energy, being one of the main global energy sources in this century [2].

However, despite having tremendous potential, solar panels have only been able to utilize an incredibly minimal portion of the energy they retrieve from the sun. In 2019, crystalline silicon, the most preferred material for solar cells, only had a max threshold efficiency of about 26.7% [3]. Paradoxically, solar panels work best in strong sunlight, but because of the heat generated by the light, the energy yield gained only ends up being wasted [4]. While heat itself does not impact the total amount of energy solar panels receive from the sun, because solar panels operate from the difference in electrons at rest and those excited by the sun, heat reduces their efficiency because it excites electrons and raises the resulting energy of electrons at rest, producing a lower difference and thus minimized energy yield [5]. Intuitively, figuring out the optimum way of minimizing the heat produced through methods like cooling will be a critical aspect of enhancing panel performance [6].

More broadly, when combined with reducing conditions like heat, achieving an ideal balance in the conditions of other variables (i.e., humidity) known to influence solar panel performance will be indispensable in optimizing solar efficiency. Naturally, researching how environmental conditions like temperature, humidity, or atmospheric pressure correlate with each other to impact solar panels holds immense value for achieving this optimum balance. Understanding the output variability of photovoltaic panel (PV) output on any day’s given weather conditions helps transmission system operators find an adequate balance for the entire grid as under or even over-producing electric output results in penalty fees [7]. Using machine learning (ML) techniques to forecast solar energy and to gain insight into the nature of each condition’s relationship play a very critical role in forecasting PV panel energy output accurately [8]. Prior research demonstrates that while time series models are the most effective in short-term solar energy prediction, ML models like ANNs and Gradient Boosting Decision Trees (GBRT) had the most accurate long-term forecasting for solar energy [8]. Generally, models like lasso, ANNs, GBRT, K-Nearest Neighbors (KNN), and Auto Regressive Integrated Moving Average have been known to perform well in prior research on energy forecasting [8]. As a result, this work seeks to utilize predictive models like MLR, lasso, elastic net, ridge regression, and ANNs in assessing the nature of these relationships and using them to predict future solar energy yield.

# Materials and Methods

Predictive analysis was performed on two provided datasets from the [UK Power Networks](https://data.london.gov.uk/dataset/photovoltaic--pv--solar-panel-energy-generation-data) (UPN), with one dataset providing twenty-nine different weather conditions for five sites in the UK and the second having respective solar power measurements at customer endpoints. Both datasets shared three sites. Notably, thirty-two different weather conditions were measured but only twenty-nine were incorporated as three variables had very sparse data. We linked the weather and power measurements according to identical time, site (known as substations for customer endpoints), and date markers. But before doing so, due to the fact the datasets had different notation and format of timestamps and dates, we reformatted them to link them correctly.

Interestingly, while the UPN did provide other customer endpoint datasets that recorded PV generation values in one and ten-minute intervals, the UPN only had weather recordings that were recorded each half hour. And despite weather’s variable nature across larger time intervals, weather values did not widely vary between each immediate half hour. As a result, the most cogent way to join our data was incorporating the dataset with corresponding hourly measurements (ones with the same start times). Further, since minimum and maximum solar power generation values were recorded each hour, we utilized the average of the minimum and maximum power measurements for each hour. While weather monitoring and solar power generation were performed outside, there was also monitoring inside some customer endpoints. Consequently, some variables are indoor conditions and are correspondingly documented in the dataset to differentiate from counterpart outdoor variables. As such, unless noted as being indoor (like indoor air density or indoor humidity), environmental variables mentioned in this paper can be assumed to be outdoor conditions.

Moreover, according to the dataset documentation, because some generation values were multiplied by a negative number for dates before a certain timeline, we filtered non-positive generation values. To ensure normality of residuals, solar power generation values (measured in kW) above 0.5 were also filtered (0.5 was the best cutoff value to adjust for the heavily right-skewed response variable). With 16,469 total linked observations in the dataset, although filtering excluded a rather large portion of data from the analysis, the cutoff still allowed for 3,261 rows of solar power and corresponding weather measurements to be examined.

Markers like site, archival interval (frequency of data logging), hour, date, and various other customer endpoint variables were removed but other factors like wind direction were dummy encoded (numeric values given according to different levels of categorical variable) before being fed in for training and testing. Functions from version 0.7.3 of the model [performance package](https://rdrr.io/cran/performance/) in R assessed linear model assumptions like normality of residuals and outliers, homoscedasticity, multicollinearity, and homogeneity of variance.

With an 80-20 ratio, the data was split into respective training and testing categories (also known as in sample and out of sample respectively) to make the predictive models. For MLR, we made four models in total, the first being a full model with all twenty-nine environmental factors, a second model with variables manually added that had low collinearity (like wind speed, rain, and solar radiation), a third one selected from stepwise regression (this had thirteen variables), and a fourth model that used the variables from stepwise with low collinearity. While ten-fold repeated cross-validation was only used for the regularized models, hold-out validation was used on all predictive models for final model benchmarking. To note, while MSE was the model metric for overall model selection, AIC was the metric for stepwise regression. Data was standardized for the other various ML techniques like lasso, elastic net, ridge regression, and ANNs. The R package [glmnet](https://cran.r-project.org/web/packages/glmnet/index.html), with version 4.1.2, was utilized for implementing lasso, elastic net, and ridge regression. Additionally, instead of R, we incorporated ANNs in Python with version 2.4.3 of the [Keras](https://keras.io/) package (used for deep learning in Python) because of some modeling challenges with implementing neural networks in R.

# Results and Conclusions

Multiple Linear Regression

**Table 1**

*Testing MSE of MLR Models*

|  |  |
| --- | --- |
| MLR Models | Testing MSE |
| Model 1 (full model) | 0.257135 |
| Model 2 (manual selection) | 0.303348 |
| Model 3 (only stepwise) | 0.258881 |
| Model 4 (manual selection from stepwise) | 0.313880 |

**

**Figure 1**

*Depiction of model*

*performances on training sets. Plot adapted from R to show visual summary of model performances across metrics like Bayes Factor (BF), R Squared (R2), R Squared Adjusted (R2Adjust), Akaike Information Criterion (AIC), Bayes Information Criterion (BIC), Root Mean Squared Error (RMSE), and Mean Squared Error (SIGMA). BF assesses and contrasts the likelihood of the input data occurring under a particular model to another model. R2 is a statistical measure of the model’s goodness of fit to the input data. R2Adjust is like R2 but also accounts for the number of input variables as well. BIC, like AIC, enables selection of a model among a finite set of models. RMSE and MSE both benchmark the model’s overall predictive error, with MSE only differing from RMSE in being the square of RMSE. Colors denote which model was used. It is also noteworthy that the metrics were not rendered on the same scale.*

In Figure 1 and Table 1, despite having high multicollinearity (a violation of our MLR model assumptions), model three (only stepwise) performed the best according to our training MSE while model one (full model) had the highest testing MSE. It is worth noting that while model one was the best linear model since it had the highest out of sample MSE, model three outperformed all the other models on every metric (especially BIC) in sample.

To better analyze and interpret our linear models, we wanted to assess how factors like latitude, longitude or elevation would have had on the performance of the solar panels. However, because some locations had different facing solar panels [9] and the fact we could not directly verify which sites were facing the same direction, no effective geospatial analysis could be done across the different sites.

Regularized Models

Due to the violation of certain assumptions in MLR (like multicollinearity), we also utilized other linear models like ridge regression (which is more robust for a violation of multicollinearity) as well as lasso and elastic net. Ridge, lasso, and elastic net models all fall under the category of regularized models and still under the much broader category of linear models. Regularization is a technique that decreases variance at the cost of increasing some bias, which is the accuracy of our predictions [10]. Being able to find an optimum bias-variance balance is what minimizes the model’s total error. Ridge, elastic net, and lasso are all very similar to MLR but have an added penalty term. More specifically, ridge regression penalizes the sum of the squared coefficients, lasso penalizes the absolute values of coefficients, and elastic net incorporates a combination of ridge and lasso penalties. Cross-validation (ten-fold) was used to tune the penalty term to find each model’s best fit.

**Table 2**

*Testing MSE of Regularized Models*

|  |  |
| --- | --- |
| Regularized Models | Testing MSE |
| Elastic Net | 0.266786 |
| Ridge | 0.281489 |
| Lasso | 0.267303 |

As shown from Table 2, though elastic net was the highest performing model, ridge, and lasso both shared very similar MSE values with elastic net, indicating differing adjustments for the dominance of each variable in the models produced negligible changes in overall predictive power. This indicates that incorporating more conditions might be more useful for prediction purposes while knowing the strength and interaction of each variable is more effective for analyzing what adjustments and locations are the most cost effective for solar energy. Beyond just the accuracy of the models, regularized models can provide immensely useful information in variable selection for models (much like stepwise). Regularized models can also help identify the most influential variables in a dataset, which holds promising applications for knowing what regions are most suitable for placing solar panel farms since we know what conditions are most important in affecting solar panel output. Lasso shrinks the estimate coefficients of the variables towards zero so that the most dominant variables influencing the response variable are left remaining [10]. Ridge keeps all the estimate coefficients, but shrinks the ones that are not as influential, whereas the elastic net integrates both approaches [10].

Lasso and elastic net shared a similar selection of the most dominant variables, such as low temperature, wind direction, barometric pressure, solar radiation, wind speed, and cooling degree days. However, lasso did not include variables like heating degree days, indoor humidity, evapotranspiration, or rain. In context, relative to all the other environmental variables, this informs us key conditions like wind speed, wind direction, low temperature, and solar radiation have the largest effect on solar power. It is not surprising that low temperature and solar radiation are dominant characteristics as modern research has already shown, along with wind speed due to its natural cooling effects [11].

Further research would be needed to assess and investigate variables like wind direction, as unlike simple MLR, regularized model coefficients themselves do not indicate the exact effect of one variable on the response given other ones being held constant, it only gives information regarding the strength of that relationship. Also, the sparsity of other estimate coefficients in lasso does not mean they had zero effect on the response, only that the remaining coefficients in lasso carried the most impact. In the case of ridge regression, variables like rain, heating, indoor air density, and cooling degrees exhibited the strongest influence, matching some of the results elastic net and lasso produced.

Neural Networks

ANNs were the last predictive modeling technique used for forecasting future energy yield. They were used to do regression on the models solely for the sake of contrasting prediction accuracy. At the current moment, interpreting ANNs can prove to be a challenging endeavor and resultingly was not our focus. Generally, neural networks are well more suited when prediction is more heavily prioritized than the interpretability of the model. Neural networks would be extremely useful in settings like pure energy forecasting as stakeholders are far more interested in prediction to adjust their energy supply for upcoming demand. After running an ANN with three hidden layers on the data that included all environmental conditions, the model produced a testing MSE of 0.276590, which is comparable with prior results from our simple MLR and regularized linear models. Across the board, while the first MLR model had the highest performance, the MSEs were relatively uniform across all models. Notably, due to our data filtering, the reduced scale of our response variable likely contributed to the small MSE values; however, even in context of our data’s limited scale, our MSEs are still promising.

Conclusions

In future research, going into two different tracks of focus where we have models that are more focused on interpretability and other ones like ANNs solely for prediction will assist in both predicting and investigating how weather conditions influence solar energy. Moreover, utilizing a Monte Carlo method called bootstrap would be needed for stronger statistical inference on the variable selection achieved with regularized linear models. Further, investigating how variable selection impacts the predictive power of more sophisticated models like neural networks might enable us to better optimize solar energy forecasts. Interestingly though, since nearly all the weather conditions known to affect solar panels were already included in the dataset, only cloud cover would need to be incorporated an additional weather condition in future analysis. Ultimately, this work helps energy grids that depend on solar energy to more reliably project and adjust for future demands based on anticipated solar panel production. Having a better understanding of how the variables interrelate to affect solar energy will also help us determine the most optimal adjustments and sites for solar panels.



# References

1. IEA (2014), World Energy Outlook 2014, IEA, Paris https://www.iea.org/reports/world-energy-outlook-2014

2. Bakhiyi, Bouchra & Labrèche, France & Zayed, Joseph. (2014). The photovoltaic industry on the path to a sustainable future — Environmental and occupational health issues. Environment international. 73C. 224-234. 10.1016/j.envint.2014.07.023.

3. Lucio Claudio Andreani, Angelo Bozzola, Piotr Kowalczewski, Marco Liscidini & Lisa Redorici (2019) Silicon solar cells: toward the efficiency limits, Advances in Physics: X, 4:1, DOI: 10.1080/23746149.2018.1548305

4. M. Kumar Panjwani and G. Bukshsh Narejo, “Effect of Humidity on the Efficiency of Solar Cell (photovoltaic),” Int. J. Eng. Res. Gen. Sci., vol. 2, no. 4, pp. 499–503, 2014, [Online]. Available: www.ijergs.org..

5. UCSB Science Line. (n.d.). Retrieved June 17, 2021, from http://scienceline.ucsb.edu/getkey.php?key=2668

6. Moharram, K., Abd-Elhady, M., Kandil, H., & El-Sherif, H. (2013). Enhancing the performance of photovoltaic panels by water cooling. Ain Shams Engineering Journal, 4(4), 869-877. doi:10.1016/j.asej.2013.03.005

7. Isaksson, E., & Karpe Conde, M. (2018). Solar energy Forecasting with Machine Learning Techniques (Dissertation). Retrieved from http://urn.kb.se/resolve?urn=urn:nbn:se:kth:diva-229065

8. Aler R., Martín R., Valls J.M., Galván I.M. (2015) A Study of Machine Learning Techniques for Daily Solar energy Forecasting Using Numerical Weather Models. In: Camacho D., Braubach L., Venticinque S., Badica C. (eds) Intelligent Distributed Computing VIII. Studies in Computational Intelligence, vol 570. Springer, Cham. https://doi.org/10.1007/978-3-319-10422-5\_29

9. UK Power Networks. (2016). Photovoltaic (PV) Solar Panel Energy Generation data – London Datastore. Retrieved December 9, 2020, from https://data.london.gov.uk/dataset/photovoltaic--pv--solar-panel-energy-generation-data

10. (Tutorial) Regularization: Ridge, Lasso and Elastic Net. DataCamp Community. (n.d.). https://www.datacamp.com/community/tutorials/tutorial-ridge-lasso-elastic-net.

11. (PDF) Impact of Environmental Factor's on Solar Photovoltaic Module and Different Material Employed on it. Available from: https://www.researchgate.net/publication/343994758\_Impact\_of\_Environmental\_Factor's\_on\_Solar\_Photovoltaic\_Module\_and\_Different\_Material\_Employed\_on\_it [accessed Dec 09 2020].