An Internet-of-Things Approach to Minimal Feature-Based Machine Learning Photovoltaic Output Predictions

Abstract

Not only is there is much interest in the accurate prediction of energy production from zero-carbon energy sources like solar and wind power to provide flexibility for basic usability, but also for the other possible practical applications such as energy deployment-on-demand and high energy production contracts. Estimates based on complex multivariate forecasting equations or formulae tend to err in any capacity for continuous prediction because of the inflexibility inherent in fixed equations when dealing with variability in climatic conditions and in the complexity of energy production patterns. To counter this issue, I took a cloud based approach to photovoltaic predictions, using Microsoft Azure Machine Learning services coupled with an internet of things approach for data collection from the photovoltaic sets that were tested. Using a low-feature, many-model approach with available weather data, I was able to achieve a 91% prediction accuracy for the real power output. Using Azure Machine Learning based operationalized web services, and cloud integration to make bulk predictions possible, photovoltaic predictions should only require a connection between the various aspects of the data flow, allowing both large photovoltaic plants and small household sets to accurately predict power output using minimal weather data.

An Internet-of-Things Approach to Minimal Feature-Based Machine Learning Photovoltaic Output Predictions

1. Introduction

1.1. Overview of Solar Power

The limited supply and negative impact of hydrocarbon sources on the environment via CO_2 emissions drive the continued quest for a viable replacement energy source. Zero-carbon sources in use today such as solar photovoltaic cells or wind turbines are rapidly increasing.[1] Most of the work in this area is focused on photovoltaic plants in the United States and Europe, with large specific endeavors to create a single model that works perfectly, while Asia-Pacific and China are observing the most growth in past two years for new photovoltaic systems, with significant amounts of data that could easily be used to teach a machine and pick a model or models from a large selection.[2] Photovoltaic cells – semiconductor devices which convert solar irradiance into electrical energy – contain a total global capacity of roughly 139 gigawatts by the end of 2013: a 38% increase from 100.5 gigawatts by the end of 2012, and a significant increase from the 23 gigawatt capacity by the end of 2009.[3] Individual solar photovoltaic plants have also increased in size; average photovoltaic plants in the United States increased in capacity from 8.78 kilowatts in 2009 to 24.10 kilowatts in 2012 and 30.65 kilowatts in 2013.[4] The new installed capacity of photovoltaic cells for the period from 2014 to 2020 is predicted to be 513 gigawatts, with solar power leading the growth among all other renewable and conventional energy sources.[5] However, at the end of 2013, the theoretical global output of the world's photovoltaics was approximately 160 terawatt hours, enough energy to power the world for only half a day.[6]

Despite the significant increase in photovoltaic capacity and output that is ongoing and predicted, the application of photovoltaic energy as a renewable zero-carbon energy resource remains in question, as it lacks the capability of traditional energy sources such as coal, natural gas, petroleum, and oil. For photovoltaic energy to function as a viable resource, the following must be achieved:

a. **The production of energy must be scalable.**[7]

   Like traditional hydrocarbon sources, photovoltaic sources must also be able to provide as much or as little power as needed. To this end, photovoltaic sets may be placed in strategic locations to maximize power achieved; improvement of collection, storage, and distribution chains constantly make photovoltaic production more viable in terms of scalability.

b. **The production of energy must be predictable or controllable.**[8]

   For carbon-based sources, such as coal burning electricity production or petroleum based vehicles, energy production is predictable due to the control inherent in managed burning. Thus, even if the process is inefficient, the output can be predictable. However, for photovoltaic production, control is lost due to the factors that affect the amount, intensity, and spread of solar energy that reach photovoltaic sets.

Not only is there is much interest in the accurate prediction of energy production from zero-carbon energy sources like solar and wind power to provide flexibility for basic usability, but also for the other possible practical applications such as energy deployment-on-demand and high energy production contracts. However, estimates based on complex multivariate forecasting equations or formulae tend to err in any capacity for continuous prediction because of the inflexibility inherent in fixed equations when dealing with variability in climatic conditions and in the

complexity of energy production patterns.[9]

1.2. Overview of Solar Forecasting

Photovoltaic arrays are both variable and intermittent in their energy output, making integration with the power grid challenging; for example, photovoltaic output is affected by temporal factors such as the time of day and day of the year and environmental factors such as cloud cover[10,11], temperature, and air pollution[12]. The degree of environmental variability, lack of accurate prediction capability, and the unreliability of power output negates the benefits that result from the possibly reduced costs, higher efficiency, and increased "greenness"[13] compared to current solutions. Thus, an advanced forecasting of photovoltaic power output to provide reliability and to increase efficiency of collection and distribution is necessary for a practical photovoltaic grid system.

Often, predictive models are created using long equations and some data to simulate a particular photovoltaic set[14]. Although these are somewhat effective, they require a newly developed and tested model for each and every photovoltaic set. Since these models are created for a specific photovoltaic set, which consists of certain manufacturer and operating parameters, they are generally rendered unusable for application to other photovoltaic sets.

Solar irradiance forecasting is the basis of photovoltaic power prediction.[15] Cloud movement, cloud formation, and dissipation are primary causes for solar irradiance variation, and are therefore required to predict the solar radiation of the next day.[16] A variety of solar power forecasting techniques have undergone extensive research such as the use of Numerical Weather Prediction (NWP) models[17], which track cloud movements from satellite images[18], and cloud movements from direct ground observations with sky cameras[19].

1.3. Problems with current approaches

Numerical weather prediction (NWP) models are consistently prone to errors for solar irradiance predictions and thus suffer from instability; the physical models typically form around complex non-linear equations formulated around mathematical equations describing the various measurable attributes of the atmosphere.[20] In addition, performance curves, meteorological conditions (solar radiation, wind direction, and humidity), and a variety of operating strategies need to be factored for the simulation model to provide accurate estimates and forecasts.[21,22,23] Although it is possible to develop an accurate model of a PV plant, the efforts and resources required to determine the best-fit algorithm and regression models are restrictive. This kind of modeling and prediction traditionally requires complex software, high-end computers, and seasoned data scientists who understand the comprehensive scope of photovoltaic radiation, along with sufficient funding to carry out such an expensive endeavor. A vast majority of current solutions and services require complex processes[24] involving the creation of detailed physical and statistical models, which are needed to predict expected power output from PV at any time based on the system input variables and a large number of parameters, optical, thermal, and hydraulic characteristics of every component.[25]

1.4. Approach towards Problem

Machine learning is the science of programming computational systems to automatically detect patterns and correlations in data.[26] Machine learning converts data sets into pieces of software, known as "models," that can represent the data set and generalize to make predictions on new data. Through rigorous training, a machine learning model can be "taught" a large set of previous data, then predict data based on the inputs, without explicit programming. Machine learning uses data sets of related factors to aid prediction and the outputs to train a regression,

usually a certain regression picked for an application that can best represent the data set and make predictions on new data.

Microsoft Azure Machine Learning is a cloud based services implementation of machine learning; using a web interface, users can drag and drop predefined components, or “modules”, and create a data flow to create predictions, score them, and statistically evaluate the predictions’ accuracy. The web service API also exposes these models that have been trained to allow secured remote access and use of the model. Along with these tools, other configurable tools provided allow the user to format and edit datasets to make them comparable and usable together, and allow the importation of data from several sources. In this project, the goal was to provide a basis for future prediction models by creating a data flow using sensor to cloud communication – an internet-of-things implementation - that would be applicable and available to other photovoltaics through cloud-based web services. Thus, to that end, Microsoft Azure Machine Learning was used for modeling and prediction, allowing access and sufficient applicability to provide an easily interchangeable and widely deployable backend.

2. Materials and Methods

2.1. Experimental Design

Using Azure Machine Learning, several data flow graphs were created with parallel paths, allowing execution of complex experiments and side-by-side comparisons, because of the lack of the usual local computational constraints. Several regression modules were used simultaneously to find the best possible regression model for the input data, making this viable for use with photovoltaic sets. The project used real-time live data feeds direct to Microsoft Azure cloud

storage from the large 125 kW onsite array of photovoltaic solar panels at the Phipps Conservatory[27] in Pittsburgh, PA and from the co-located Carnegie Mellon University weather stations. The goal of this project was not only a high accuracy, but also, wide applicability to a variety of photovoltaic sets, from the application of just one or two factors. With an accuracy of around 90%, the predictions are provided with a published web service in 1-4 seconds, and most importantly, is accessible anywhere with an internet connection.

2.2. Data Collection Approach

For data collection, photovoltaic power output and weather data collection sensors were directly connected to the Azure storage tables using an internet-of-things connection strategy, where the collector's side is responsible for pushing data. Once the data was available on Azure tables, it was incorporated inside the Azure Machine Learning studio.

Using the Azure Machine Learning service and Azure cloud storage tables to store historical data and receive real time updates directly from the sensors, trained models achieved a high degree of convergence predicting real power using solar radiation and weather data. The approach to create the optimal model of the photovoltaic sets at the Phipps Conservatory– both the ESRC and ground-mounted sets– was to collect data that was as granular as possible, and then use the data to train as many different models as possible to find the best fits. This approach was selected as the idea behind the implementation not only to create a specific model type, but also to establish a widely applicable model that could work given data from different photovoltaic sets. The data collected were characterized as such; the photovoltaic data collected included: Energy Delivered, Energy Received, Power Factor, Reactive Power, Apparent Power, Real Power; the weather data collected included: Outdoor Air Temperature, Wind Direction, Wind Speed, Solar Radiation, Rain Fall Sensor, Outdoor Air Humidity.

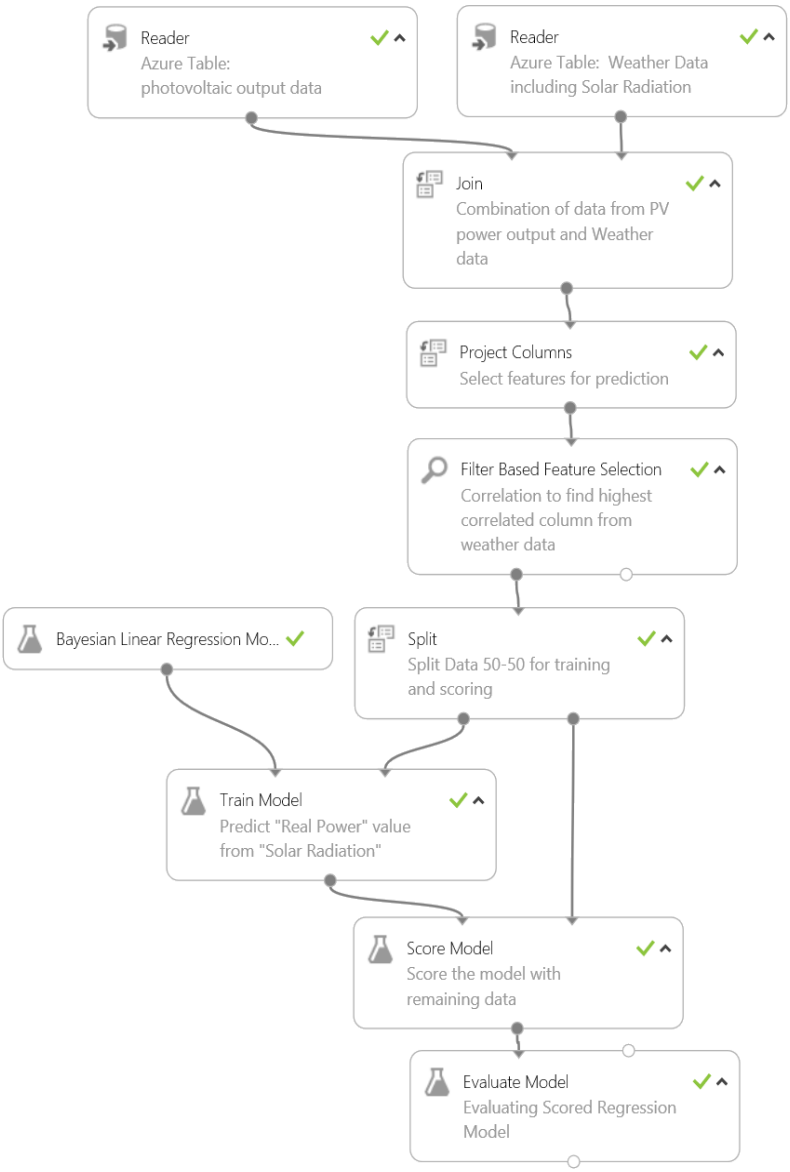

Figure 1: The diagram to the right provides an example of the data flow used to create the series of models used throughout this project.
Readers: the readers take input from Azure Cloud Storage, one from a photovoltaic set (either ESRC or the ground mounted) and the other from the weather station.
Join: This data feed is then joined using the timestamps as row keys using the "Join" module. It automatically discards values that have no matching value in the other dataset.
Project Columns: the "Project Columns" module then removes the miscellaneous photovoltaic columns that are not to be predicted and are not useful in making predictions.
Filter Based Feature Selection: This module takes a column to correlate to and a set of input columns and returns a user specified amount of the most correlated input columns using a user-specified correlation method.
Split: The "Split" module then randomly partitions the data; half is fed to the training module. The other half is used to evaluate the trained model using "Score Model".
Train Model: This module takes trains the selected models– in this case, the Bayesian linear regression – and trains it against the data given. The output is a link to the trained model.
Score Model: This module uses the trained model to create predictions. This dataset, which now contains a column of predictions called "Scored Labels" is given to the "Evaluate Model" module.
Evaluate Model: This module performs statistical analyses on the predictions, such as the coefficient of determination, mean absolute error, root mean squared error, relative absolute error, and relative squared error.

Data extraction from the photovoltaic sets and weather data requires varied interfacing methods. Early implementations of physical collection made the usage of a cloud-based machine learning pointless due to the inefficiency and increase in apparent processing time. Thus, the solution implemented to allow data input to the machine learning environment was a sequential timed upload to the Azure Table servers for an indeterminate period of time. The wireless modules for the both the photovoltaics and the Intelligent Workplace (IW) overhead weather station

were previously attached and set; the available wireless hooks and interfaces were used to connect these wireless upload mechanisms directly to open Azure Storage Tables.

Azure Tables allow single directional modification – in this case from the source – enabling the wireless upload to be set at maximum resolution or granularity without the need of user-modification-error checking in the Azure Storage Table. However, sampling rates differed among the two photovoltaic sets and the weather station, creating an asynchronous set of data. This was not accounted for during collection of the data; the data was simply uploaded to create the most granular dataset possible using the existing wireless upload capabilities. Thus, for each of the datasets created, the following properties were true:

a. **The dataset remained devoid of user introduced error.**

   This holds importance as this ensures that it is guaranteed that any errors that are detected are sourced from either the machine or wireless chain's side, and could be accounted for and dealt with as such. Such an approach holds each object in the chain responsible for its own data but creates a much more fluid and safe data collection approach than a normal collection approach – whether manually or by computer based querying.[28]

b. **The dataset was as granular as possible through the wireless upload chain.**

   This allowed the highest resolution possible – the resolution of the least granular set – while at the same time, keeping the data available for reference. It also allowed for the possibility that the highest possible granularity is used, and that some form of mean function is used to compute in-between values.

Therefore, the datasets for the photovoltaic sets were updated every minute, while the weather data could only be collected every 5 minutes. Around a third of the way through our data collection during which the models were being continuously trained, the ESRC data upload

failed, and the models for the ESRC, although operational, were no longer learning after 15 days into the 50 day trial period.

Using the preconfigured feature set of the Azure ML environment: the Reader object class modules, resulting datasets were imported to the machine learning environment from the Azure Table every time the experiment was run (runs were done after every change made to the data flow). Readers take input from a set of predetermined and preconfigured sources: a native HTTP table set, an Azure SQL table, an Azure Table, Azure Blob storage, the Apache HiveQ query style, and Microsoft PowerQuery. However, the source used was the Azure Table, so the readers throughout the experiment were selected to use AzureTables. The key, specific table names, and generic account set names were all entered for each reader instance. Each reader module was assigned a table to parse: one for the ESRC photovoltaic data set, one for the ground mounted photovoltaic data set, and one for the weather data streamed from the IW overhead. After the data was set to be collected by individual readers, and the readers set to refresh when run, the next step was to fix the misalignment of the data caused by the difference in maximum granule size among the different data sources.

2.3. Alignment and Selection of Relevant Columns

The module used to fix this misalignment was the “Join” module; the “Join” module takes a key column, available in two datasets, and merges the two datasets by matching keys – if there is no matching key for one value, the value is omitted from that table. This was used to combine these data meaningfully; the data rows were all partitioned using the timestamps as keys. These timestamp keys were aligned in the “Join” module, and rows’ timestamps that were present in both the photovoltaic data and the weather data were kept; data that did not fit this paradigm were discarded. This meant that the dataset with the lowest granularity was the limiting

factor; in our case, the photovoltaic data was this limiting factor. Inner join was selected which is the typical join operation and returns the combined rows only when the values of the key columns match. After the join module provided alignment and discarded all unusable data, the relevant data was selected from the merged dataset; the column to be predicted, real power, as well as all weather related data columns were kept, but all columns from the photovoltaic sets other than real power, were discarded as they would provide a high correlation when the data was tested even though those columns were not to be used for predictions using the "Project Columns" module, as seen in Figure 1.

2.4. Filter Based Feature Selection

With the complete weather data set and the real power column merged using the join module, a filter based feature selector was used to determine which weather data feature was most correlated to the real power output of the photovoltaic cell.

The "Filter Based Feature Selection" module uses different statistical tests to determine a subset of features with the highest predictive capability in relation to a single column, as seen in Figure 1. This does not guarantee any meaningful relationship among the variables, but rather simply proves that the outputs are in some way correlated. In this case the input was the merged or "joined" dataset that contained two or more feature columns. These columns were tested using five correlation methods that are included in the Azure Machine Learning environment, each of which returned solar radiation as the most related feature. Two methods are discussed here out of the three in Table 1, as these methods provided the highest correlation between solar radiation and real power.

2.4.1. Pearson Correlation

This correlation, known formally as the Pearson Product Moment Correlation, refers to a

linear regression that relates two sets of data. It attempts to create a line – a linear function – to represent the relationship between the two variable columns given. Since it simply finds a linear relationship, the strongest relationship would be a 1 (a positive 100% correlation) or a -1 (a negative 100% correlation). It is however, very unlikely that a correlation of 1 is found, since a perfect linear relationship is rare in real world testing. Similarly, a complete lack of correlation (0) is also impossible as even a completely random distribution would create some amount of correlation.

Correlation Method	Solar Radiation Correlation
Pearson Correlation	0.944635
Spearman Correlation	0.911432
Kendall Correlation	0.747661

Table 1: A table of a sample of different correlation scores received by solar radiation using different correlation methods. The correlations that were available were: Pearson Correlation, Mutual Information, Kendall Correlation, Spearman Correlation, Chi Squared, Fisher Score, and Count Based. For all correlations, solar radiation received the highest score. The top three scores are shown here (closest to 1 is best). The Kendall Correlation was not used.

$\boldsymbol{R} = \frac{\boldsymbol{n}(\sum \boldsymbol{xy}) - \sum \boldsymbol{x} \sum \boldsymbol{y}}{\sqrt{[\boldsymbol{n}\sum \boldsymbol{x^2} - (\sum \boldsymbol{x})^2][\boldsymbol{n}\sum \boldsymbol{y^2} - (\sum \boldsymbol{y})^2]}}$ In this equation, n is the number of data points, x and y are the two variable columns being checked for correlation, and R is the Pearson Product Moment Correlation[29]. The Pearson Product Moment Correlation effectively increases as the linearity of the data's relationship does.

2.4.2. Spearman Correlation

The Spearman correlation, or the Spearman's rank correlation coefficient (ρ or r_s) measures correlation based on monotonic relation. Therefore, even if a correlation is not completely linear, the Spearman correlation can actually be perfect, given that the function is always increasing. $\rho = \mathbf{1} - \frac{\boldsymbol{6}\sum \boldsymbol{d^2}}{\boldsymbol{n}(\boldsymbol{n^2}-\boldsymbol{1})}$ The above equation gives the process through which the Spearman correlation is determined; first, the ranks of the data are computed, and a Pearson Product Moment Correlation is run between the ranked variables to establish the Spearman ranking.[30] Once the feature selector was run with each of the above correlation methods, the most related features

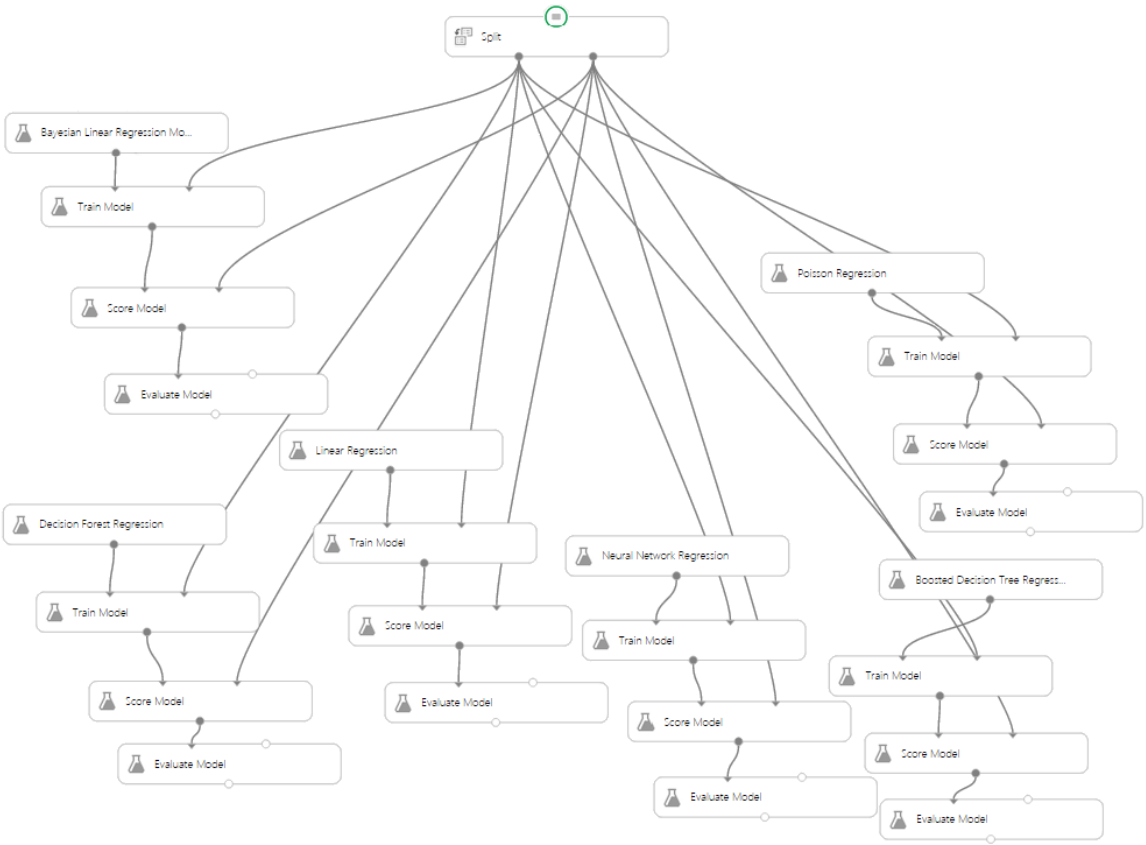

Figure 2: After using the "Join" module and selecting the columns needed and the data split using the "Split" module – as seen in Figure 1 – six models were trained and scored using the same data in parallel. Using several "Add Rows" modules, the best model for the data inputted was selected. At the last "Add Rows" module, the output was a table of correlations - the ones used to create Table 1. The green circled dot on the top of the "Split" module is actually an input for an operable web service, allowing data inputted from the web service to be analyzed and have the best model selected. Not pictured here: "Reader" modules take input from Azure Cloud Storage, one from a photovoltaic set (either ESRC or the ground mounted) and the other from the weather station, along with the "Join" and "Project Columns" modules.

were selected to be used in the model. As in Table 1, all correlations suggested that solar radiation was the most correlated to real power.

2.5. Training

The final step was to train the models; each of the six models – Bayesian Linear regression, Boosted Decision Tree regression, Decision Forest regression, Linear regression, Neural Network regression, and Poisson regression – were attached to their individual training module, and trained, as shown in Figure 2.

3. Results and Discussion

To score and evaluate the models, two modules, named "Score Model" and "Evaluate Model" were used. "Score Model" performs a prediction using an attached model, creating a column titled "Scored Labels" containing these predictions. Scored labels can then be graphed against the actual data to create a graph that shows the relationship between the predicted and actual data, providing a visual impression of how well the prediction model performs, and therefore, how applicable the feature used to make the predictions is. In our specific case, the scored labels are the machine's prediction of power (measured in watts), while the real power is the actual output at that time (in watts). If the model were perfect, then we would see a straight line when the scored labels were graphed against real power since the data would be exactly 1:1; an exact correlation between solar radiation and real power would allow a perfectly linear graph. As seen in Figure 4, the relation between scored labels and real power is consistent for the regressions used. This graphical comparison was confirmed using the "Evaluate Model", as seen in Figure 1. To select the best model, the main factor of selection was the coefficient of determination. The coefficient of determination was calculated and compared for different models using the "Evaluate Model" module: this module took the scored dataset and provided several statistical factors, including the coefficient of determination. According to the calculations performed, all models achieved a statistically non-weak correlation, as all coefficients of determination were greater than 0.5;[31] this is logical due to the graphical fidelity that each of the corresponding graphs have to a line. Indeed, if the Poisson regression is omitted, all other models achieved a statistically strong correlation – a coefficient of determination above 0.8,[32] and the Boosted Decision Tree regression predictions, although graphically confusing, still maintained a high coefficient of correlation, suggesting that it is still a viable option. Thus, the evaluation module then

confirmed the strong visual correlation seen in the Scored Label vs. Real Power graphs.

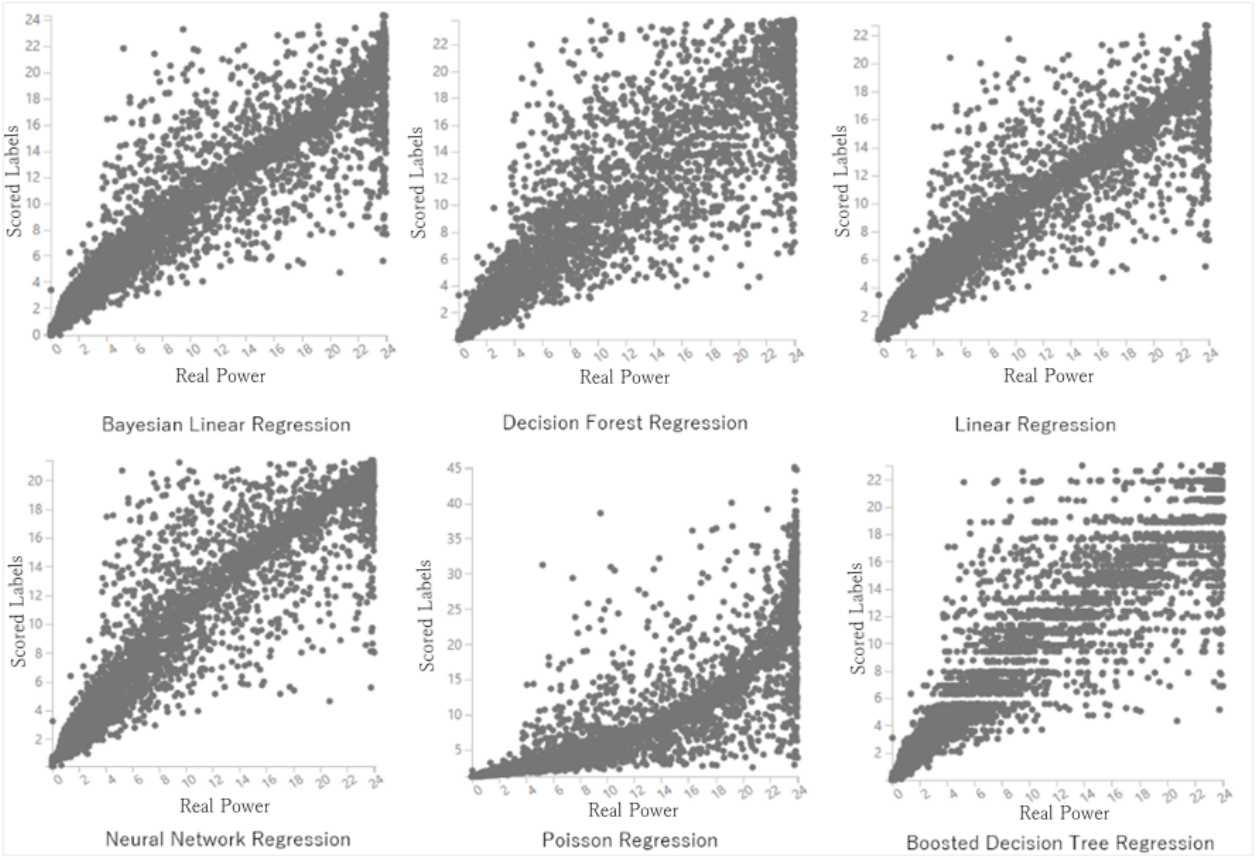

Figure 2: Scored Labels vs. Real Power for six regressions, labeled underneath the respective graph, for the full set of ground mounted photovoltaic data (collected at the end of 50 days - July 12th to August 30th). The Bayesian Linear, Decision Forest, Linear, and Neural Network appear fairly linear, with outliers along the sides. Both the Poisson and Boosted Decision Tree regression appear to be not as linear when graphed.

3.1. Discussion

During the trial period which lasted from July 12th to August 30th (a total of 50 days), the leading two models in terms of the coefficient of determination were the Bayesian linear regression and the neural network regression. On July 26th, the ESRC collection module failed (fifteen days into the trial period), and no more data was received; however, the ground mounted continued collecting data until the end of the trial period. For the ESRC, it was clear that the Bayesian linear regression model was the clear best choice. The Bayesian model, clearly better than the

alternatives when compared graphically to the other models, also had a high coefficient of determination – 89.69%. Even with only fifteen days of full data, the model was the most linear in relationship when compared to the other graphs of the real power and scored labels. However, the other models also had similarly high results, suggesting that much of the prediction's accuracy came from the "perfectness" of the data – solar radiation is very highly correlated with real power output. At the time that the ESRC data collection failed, the ground mounted data was also best learned with the Bayesian linear regression model. Indeed, at that time, the models performed similarly for both the ESRC and ground mounted.

	Coefficient of Determination		
	ESRC (Roof Top)	Ground Mounted Solar (PV) Cells	
Regression Model	15 days		50 days
Bayesian Linear	0.89689	0.897863	0.9062
Boosted Decision Tree	0.869047	0.869323	0.9039
Decision Forest	0.866368	0.863596	0.8762
Linear	0.858693	0.850529	0.8896
Neural Network	0.869005	0.881559	0.91
Poisson	0.651891	0.682093	0.7558

Table 2: Coefficients of determination for the ESRC and ground mounted sets' models after 15 days and for the ground mounted cells after 50 days. The Neural Network regression slowly increased in accuracy to become the most accurate on day 50, compared to the Bayesian Linear regression, which was superior in the first 15 days.

However, as more data was collected for the ground mounted photovoltaic,

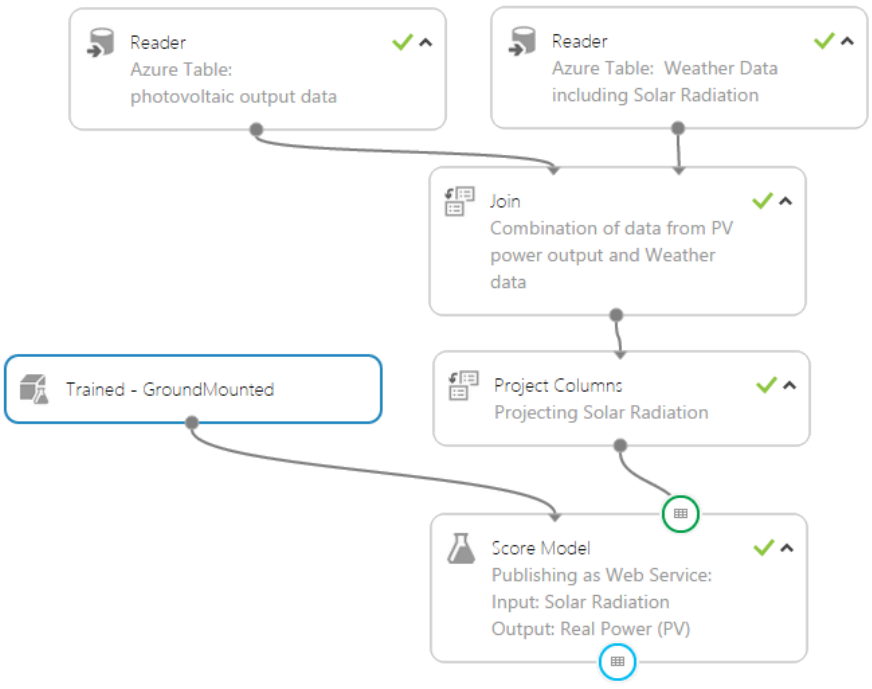

Figure 3: A saved model is seen here, named "Trained - GroundMounted". This model is used with the "Score Model" module to provide a score for the newly tested data from the reader. At the same time, the green circle on the "Score Model" module – symbolizing an input – is the input of the web service, while the blue circle is the output.

the neural network regression increased in accuracy to 91% compared to the Bayesian linear model's accuracy of 90.62%. If the pattern seen in the first 15 days had held true, then the implication would be that for the two photovoltaics tested, the starting predictions when there is less data should be performed with a Bayesian linear regression, and then as the amount of data increases, the neural network regression should be utilized.

4. Conclusion

Azure Machine Learning's widely variable modules allow the front end web service to change the model based on its suitability to the data, thus providing the variability and predictability necessary for making solar a viable replacement for hydrocarbon sources while simultaneously requiring low overhead for calculations – the only required data for a typically sunny location with a photovoltaic set are the real power and the solar irradiation. Thus, this model, although not as precise as carefully modified single model equations for a specific photovoltaic sets, still provides reasonable accuracy with low model creation time and applicability to any photovoltaic sets. From this research, I conclude that the viability of a generic photovoltaic model with easily changed inputs has been ascertained, and it is clear that the predictions made are accurate enough to increase reliability of solar and for use in other calculations.

4.1. Changes

Given the nature of the created models, it is not difficult to add more features to the prediction. During research, the filter based feature selector monitored all weather related columns and selected the column with the greatest relation to real power to make predictions of real power output. However, as seen in these results, the relationship between solar radiation and real power, although certainly very strong, is not the only factor that affects photovoltaic output. One

important feature, the temperature of the back of the photovoltaic cells or T_m, could not be obtained due to the limitations of the onboard measuring equipment of the studied photovoltaic cells. When the temperature at the back of the cells increases too much, even with maximum solar radiation, the output will always be significantly lower. A model for this feature, along with solar radiation, would allow for almost a 100% accuracy. Another factor that would have improved accuracy would have been having an integrated solar radiation column that would take into account the amount of cloud cover. Our data used for training the model was real solar radiation data – therefore taking into account all possible factors that affect solar radiation including cloud cover. Although our solar radiation prediction model made use of cloud cover forecasts, the model was not truly integrated as the cloud cover forecasts were intended for the entirety of the city and not just our location. For future work, I would model T_m and use cloud cover data from the location of the photovoltaic set to increase the accuracy of our model. An increase in trial locations, although not possible in the original experiment, would serve to increase accuracy. Photovoltaic sets from around the world along with their weather data would allow this model to be refined, and perhaps make generalized improvements to some of the regressions for all photovoltaic sets. However, by my results, it is clear that the method of using multiple models and selecting the best of them as more data is received is an approach that has immediate applicability within the research community.

4.2. Future Work

Azure Machine Learning allows trained models to be saved and used as a permanent scorer. In Figure 3, the trained model for the ground mounted photovoltaic is shown saved as "Trained - GroundMounted". This trained model is then used to predict values: once the experiment is set to use the saved trained model, then a set of input and output ports can be selected as

ports for a web service that can run with inputted data to output a prediction. In this case, the input data was the solar radiation and the output of the web service the real power. Using this accessible web service and a utility to allow the bulk influx of data, a much wider set of data could be used to train models, thereby allowing not only wide availability, but also higher accuracy for predictions made.

The approach taken in this project, where the photovoltaics communicated autonomously to the cloud in addition to data access, is an internet of things implementation of photovoltaic output prediction. Using data collection methods that are self-reliant in accord with prediction strategies that require minimal input for high accuracy while still being widely available means the only requirement is a solid connection between the various parts of this system. With this approach, not only was overhead minimized, but also allowed for wide extendibility; with the web services API built into Azure Machine Learning, implementation on a wide variety of platforms, using R, C#, and Python become a very real possibility. Another step forward would be to increase the amount of communication be-tween devices. Perhaps a cloud computing initiative could take place between actual photovoltaic sets, and predictions could be made without a computer, and allow photovoltaics to make their own adjustments to maximize output. My current work has obtained a research grant in the amount of $56,000 from Microsoft BizSpark to create a web service to utilize the selection features created in this research, which will allow models to be distributed and used more effectively. The goal is to enable any research scientist to readily test and optimize a particular model that would best suit a specific photovoltaic set. Although this is a step in the direction of integration, solar power is still unable to replace hydrocarbon power due to the cheap effectiveness that it embodies. Until then, we can hope to continue make solar power more viable through widely applicable and interpretable predictability.

References

[1]Zhang, Jie, Bri-Mathias Hodge, and Anthony Florita. *Investigating the Correlation Between Wind and Solar Power Forecast Errors in the Western Interconnection*. Technical rept. no. 3. Oak Ridge: NREL, 2013. Print. ASME 7.

[2] Zhao, Zhen-yu, et al. *Impacts of Renewable Energy Regulations on the Structure of Power Generation in China - A Critical Analysis*. Adelaide: University of South Australia, 2010. Print.

[3] Masson, Gaëtan, Sinead Orlandi, and Manoël Rekinger. *Global Market Outlook for Photovoltaics (2014-2018)*. Västerås: EPIA, 2013. Print.

[4] Kaufmann, K. *Utility Solar Trends Executive Summary*. N.p.: SEPA, 2014. Print.

[5] *Global Renewable Energy Report 2014*. Beijing: China New Energy Chamber of Commerce, 2013. Print. Global Renewable Energy Report 2.

[6]Zhang, Jie, Bri-Mathias Hodge, and Anthony Florita. *Investigating the Correlation Between Wind and Solar Power Forecast Errors in the Western Interconnection*. Technical rept. no. 3. Oak Ridge: NREL, 2013. Print. ASME 7.

[7]Kleissl, Jan, and Carlos F. M. Coimbra. *Solar Energy Forecasting and Resource Assesments*. Oxford: Academic Press, Elsevier, 2013. Print.

[8] Ibid.

[9] Hoff, Thomas E., and Richard Perez. *Modeling PV Fleet Output Variability*. Napa: Clean Power Research, 2010. Print.

[10]Sharma, Navin, et al. *Cloudy Computing: Leveraging Weather Forecasts in Energy Harvesting Sensor Systems*. N.p.: UMass, n.d. Print.

[11] Mathiesen, Patrick, and Craig Collier. *Characterization of Irradiance Variability Using a High-Resolution, Cloud-Assimilating NWP*. San Diego: n.p., 2011. Print.

[12] Mathiesen, Patrick, and Jan Kleissl. *Evaluation of Numerical Weather Prediction for Intra-Day Solar Forecasting in the Continental United States*. Research rept. no. 85. La Jolla: ScienceDirect, 2010. Print.

[13] Lizen, Sebastien, et al. *Life Cycle Analyses of Organic Photovoltaics: A Review*. Diepenbeek: Royal Society of Chemistry, 2013. Print.

[14] Gow, J. A., and C. D. Manning. *Development of a Photovoltaic Array Model For Use In Power-Electronics Simulation Studies*. Loughborough: IET, 1999. Print. Electric Power Applications, IEE Proceedings 2.

[15] Wang, Fei, et al. *Short-Term Solar Irradiance Forecasting Model Based on Artificial Neural Network Using*
Statistical Feature Parameters. Research rept. no. 5. Hebei: Energies, 2012. Print. Energies 2012.

[16]Zhang, Jie, Bri-Mathias Hodge, and Anthony Florita. *Investigating the Correlation Between Wind and Solar Power Forecast Errors in the Western Interconnection*. Technical rept. no. 3. Oak Ridge: NREL, 2013. Print. ASME 7.

[17] Mathiesen, Patrick, and Craig Collier. *Characterization of Irradiance Variability Using a High-Resolution, Cloud-Assimilating NWP*. San Diego: n.p., 2011. Print.

[18] Hammer, A., Heinemann, D., Lorenz, E., and Ckehe, B. L., 1999, "Short-Term Forecasting of Solar Radiation: A Statistical Approach Using Satellite Data, Solar Energy, 67(1–3)

[19] Chow, W. C., Urquhart, B., Lave, M., Dominquez, A., Kleissl, J., Shields, J., and Washom, B., 2011, "Intra-Hour Forecasting with a Total Sky Imager at the UC San Diego Solar Energy Testbed," Solar Energy, 85(11)

[20] Mathiesen, Patrick, and Craig Collier. *Characterization of Irradiance Variability Using a High-Resolution, Cloud-Assimilating NWP*. San Diego: n.p., 2011. Print.

[21] Cococcioni, Marco, Eleanora D' Andrea, and Beatrice Lazzerini. *24-hour-ahead forecasting of energy production in solar PV systems*. Pisa: University of Pisa, 2011. Print. International Conference on Intelligent Systems Design and Applications 11.

[22] Chakraborty, Prithwish, et al. *Fine-grained Photovoltaic Output Prediction Using a Bayesian Ensemble*. Blacksburg: HP Labs, 2013. PDF file.

[23] Sharma, Navin, et al. *Cloudy Computing: Leveraging Weather Forecasts in Energy Harvesting Sensor Systems*. N.p.: UMass, n.d. Print.

[24] Kaufmann, K. *Utility Solar Trends Executive Summary*. N.p.: SEPA, 2014. Print.

[25] Hoff, Thomas E., et al. *Behind-The-Meter PV Fleet Forecasting*. Napa: Clean Power Research, 2010. Print.

[26] Mitchell, Tom M. *Machine Learning*. New York [etc.]: MacGraw-Hill, 1997. Print.

[27] *Center For Sustainable Landscapes*. Pittsburgh: Phipps, 2014. Print.

[28]Witten, I. H., and Eibe Frank. *Data Mining: Practical Machine Learning Tools and Techniques, Second Edition*. 2nd ed. San Diego: Elsevier Science & Technology, 2005. Print.

[29] Starnes, Daren S., et al. *The Practice of Statistics*. 4th ed. New York: W.H. Freeman, 2012. Print.

[30] Ibid.

[31] Ibid.

[32] Ibid.