**Data 670 Data Analytics**

James Taylor

**Professor Hany Saleeb**

**Assignment 3**

**10/05/2021**

**Executive Summary**

Solar generated electricity has become a viable source to meet power demands in recent years with better technology, investment and society’s increased interest in renewable energy. Unlike other contributors of electrical generation (coal, natural gas, nuclear) in the US (1), solar has little influence over supply of the electrical-generating source (sunlight in this case). The uncertainty caused by solar’s variable electrical output can produce inefficiencies in the electrical grid, both concerning resources and pollutants.

      Building a model that accurately predicts a solar plant’s output in advance would help limit these potential inefficiencies by limiting excess power generation. It will help power grid administrators plan for how much solar energy could be created, whether it be available from the solar facility or an electrical storage facility that stores solar-generated power for use at peak times. Knowing this optimizes how supplemental power sources are used and would likely reduce pollution since those sources are mostly non-renewable. A model would be useful for planning when those supplemental power sources are needed to deliver power to the grid. Plants sometimes take hours to prepare themselves for electrical generation and inefficient timing means either unnecessary run-times or inability to meet the grid’s power demands.

      A variety of models will be made to help power grid administrators predict how much electricity a solar plant will produce. These models (or modeling techniques) could be used at other power grids to plan and optimize their operations.

# Project Scope

Problem Description

The problem is solar is an inconsistent source of electricity. Its availability is affected by the climatic conditions and time of day. That means solar’s ability to provide electricity to the grid is inconsistent. To meet the demand for electricity, grid administrators need to plan and have available sources of electricity to provide for the grid. A model predicting solar’s electrical production depending on climate factors will help grid administrators plan to meet electrical demands and maximize efficiency.

         The problem is important for two reasons, optimization and pollution. A grid administrator’s ability to know in advance what solar’s input would be allows them to plan what sources/plants need to be producing electricity (coal, nuclear, hydropower, etc.). However, these plants can’t be turned on/off instantly and mostly take hours to prepare to effectively produce energy. That means administrators have to predict when to have a plant start its “warm-up” procedures so that it can be producing electricity at a time when the administrators need it.

Because solar produces an inconsistent amount of power, administrators have to be conservative when using solar and have other supplemental electricity sources available in case solar fluctuates and doesn’t meet the electricity demand of the grid. In other words, administrators can’t rely on solar like they can natural gas or coal, so they have to have other sources readily available to funnel into the grid if solar’s electrical production drops. Having to keep those other resources available is inefficient and costly. It also produces unnecessary pollution because supplemental electricity sources skew nonrenewable.

The location of where electricity is generated is crucial as well. Electricity is not free to get from point A to B. There’s all sorts of costs like building and maintaining the lines, and places have varying ways of deciding who gets the bill, but another cost is electricity loss during transmission. For example, there are very productive solar plants in Nevada but that is far from Los Angeles. Every mile of distance is more electricity sent that will not get to its destination.

Business Understanding

         In the US, how energy is bought and sold to be delivered to end-users becomes complicated because it depends on region, state and even city level differences. For most of the US there are three systems on how energy is bought/sold to then be delivered to end-users, which is regulated vertical monopolies, deregulated customer choice, and deregulated wholesale market.

         In every market system, someone is responsible for guaranteeing electrical supply to customers. In the traditional regulated markets, it’s the natural monopoly utility. In the deregulated markets the utility provider has met their customer’s demand by purchasing power from plants. The inefficient costs are in having plants prepared to produce electricity that was never needed. There are also pollution costs since those plants skew nonrenewable.

Organization

There are ten Regional Transmission Organizations (RTO) that are independent and not-for-profit that manage the bids in wholesale electricity markets, which make up about 60% of US electrical supply (2). They are the ‘middle-man’ between electricity generators and the rest of the grid. They are responsible for ensuring the grid has enough power plus the reserve margin. They will need to know how much power solar plants can produce, or the solar plant can tell them the RTO that figure.

Stakeholders

There are a few groups that would be interested in predicting solar’s future output. RTO who manage the grid want to know so they can plan which plants need to be running and prepare to run to ensure they meet the grid's electricity demand. Customer’s and retail utilities will be interested as well because accurate predictions will make solar a more reliable electricity source, meaning the more expensive supplementary electricity sources won’t have to be relied on as heavily, potentially resulting in lower costs. Older plants that take longer to fire up may have prolonged lives as planning when they need to start their warm-up process can be made easier with a better idea of the electrical supply’s future capacity.

Define Business Area

The business area is the energy utilities sector. More specifically it will make solar a more effective source, which currently contributes 1-2% of the total US electricity depending on how it’s measured. Electricity in the US is heavily regulated by the government but for-profit companies still operate much of it. Either way a model could return cost savings.

Business Objectives

The model will help reduce costs and pollution by reducing uncertainty for solar electrical generation. This is good for consumers as this likely means cheaper energy and less pollution to generate it. Solar plants would become a more reliable energy source, minimizing the need for supplemental electricity sources.

         Having a model that would predict solar’s output for the next day, or the next hour would remove some variability. This could lead to more trust in solar to power the grid, potentially spurring more investment in solar.

Business Success Criteria

         A successful model for the business application is an accurate model. There are two probable ways a model can help. One is to make it’s best prediction given the independent variables. This would give the user’s the most likely output. The user, like an RTO, would then have to decide how much to trust this number because the model likely didn’t make a perfect prediction and the actual will be higher or lower than it.

         Another successful model could be one that can give the user a degree of likelihood the solar plant can meet this threshold. For example, a model like this can output the largest kilowatt threshold that it believes will be achieved with 95% confidence. This would give the user an idea of the highest electrical generation level that is very likely to occur.

Background Research

         There is a decent amount of research in the solar output prediction area. It has real monetary value and it’s also an interesting exercise. There are websites (<https://www.renewables.ninja/)> where you can get predictions on a solar system anywhere in the world. One research paper was quite accurate in predicting when solar output would drop by having a camera pointed at the sky, taking in information on clouds (3). There are numerous papers on model solar output with varying techniques.

         The dataset used was publicly available, so there is a significant amount of open source analysis and modeling on it. I have not looked at much of the models done by others, but it appears many different techniques were used.

Gaps in this Problem Resolution

         There are many data sources for aggregate electricity. These would be electrical usage or production across countries or regions. There are less sources of data at the solar plant level, where an individual plant publishes their electrical generation. This may be because individual plants are more likely to be private entities, meaning they probably don’t have to release certain information for reasons like competitive advantage and company privacy.

         Solar plants likely have some predictive models for their equipment. These simpler models don’t have interchangeability between solar plants (a model trained at one plant won’t be good at predicting at another plant). Number of solar panels, their sizes, quality, orientation, and location vary between solar plants.

         There is a decent amount of research on optimizing electrical generation from solar. Not all this information will be able to be utilized however, because there is limited information in our data, like the different types of radiation (direct and diffuse).

Proposed Project

         I’m doing this project for the potential benefit of solar generated electricity being utilized efficiently and I enjoy making future predictions. I’ll use historical data to train a model meant to predict our data’s solar plant’s electrical generation, mostly weather data.

Key Performance Indicators

         The performance indicators will be about the accuracy of the model. The first would be the RSS, or Residual Sum of Squares. The lower the standardized residuals sum the better our model is fitting our data. This technique would be used if the model is producing a point estimate as opposed to a binned prediction.

         If binned predictions are chosen, the accuracy metrics would become focused on accuracy (percent predicted correctly). These are slightly different because it is a classification problem.

         If a model has a confidence level, say a model is 95% confident the generation will be around this point, we will be able to look at the test data, compare it to the actual and see if it’s confidence levels were right. For example, if the actuals from the test are 95% confident the actual will be within +/- 10 kW, but all of the actuals are outside that band then we are misrepresenting how accurate our model is.

Project Insights of your Data Analysis

      I expect radiation and temperature to be highly correlated with electrical output. These are synonymous with direct sunlight which bodes well for solar panels. Sunlight can be blocked by clouds, which decreases both temperatures and irradiation.  The higher intensity on the panels the more electricity will be generated.

         Another likely discovery is that the most productive times will be during mid-day. That is when sunlight coming from its highest point and less light is reflected off the atmosphere, delivering the most intense light to the ground. Obviously at night there isn’t much electricity generated.

         The previous timestamp’s electrical generation may correlate well with the next one’s. The weather and sunlight conditions are likely similar since little time has passed to allow the conditions to change much. A successful model may take the previous timestamps output to help it make a prediction.

Data Set Description

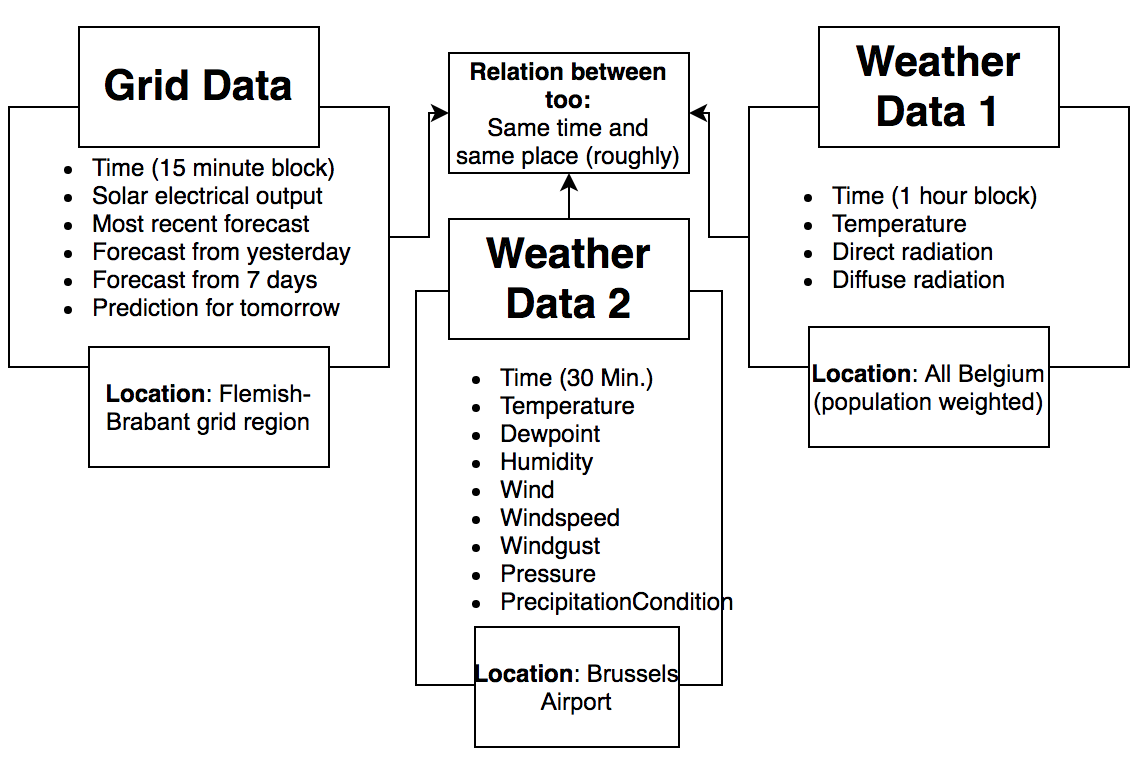
         There are three datasets being used in this dataset. The first is the data from Elia Group, which is an electrical transmission company that the FEBEG (Belgian Federation of Electricity and Gas Enterprises) started an initiative to make more data public. One of those publications is solar power electrical generation by region. The data used was from the Flemish-Brabant region, which surrounds the city of Brussels. It has megawatts produced from solar power in that region every 15 minutes, resulting in about 35,000 observations. It has a timestamp, the solar generation in MW (Megawatts), the most recent prediction for this time, the MW prediction for this time from the day before, the MW prediction for this time from seven days before and the prediction for this time but tomorrow's electrical output. All are in MW’s.

         Another dataset is sourced from Open Power Systems, which gets their data from renewables.ninja which used the NASA MERRA-2 reanalysis. It produces temperature, and both direct and diffuse radiation for every hour in 2019, which is 8,760 observations. These values are country-aggregated, so the values represent the whole of Belgium, but the data is population-weighted. That is beneficial because of the top five populous cities, the top two are from the region we have solar data for, and the fifth one is the region that is surrounded by our region (Brussels has its own region within Flemish-Brabant region).

         The third dataset is scrapped data from the Weather Underground website. It is from the Brussels airport weather station. There was an option to pay for the historic data, but instead I wrote a python script that looped through the dates of 2019 and pulled the data from a table in the HTML. This data was per-30 minutes.

         The datasets will have to be combined in a creative way since one is per-hour, one is per- half hour, and the other is per-15 minutes.

High-Level Data Diagram



The above diagram shows a basic view of the datasets and how they can be related to each other. They are at the same time and the same region (generally).  They do have separate timescales, but creative solutions can overcome this.

Data Definition/Data Profile

         The weather dataset has complicated underpinnings from NASA, but the outputs are simple enough. One important distinction is that ‘radiation’ is probably not the right word for the measurements. Instead, irradiance would be a better word to use. That is the measure of light-energy that is received on a square meter flat surface. Photons carry energy from their source and deliver it to whatever they hit. This is a very important measure for solar generation.

         Another important topic is the direct and diffuse irradiance. It's best to demonstrate the difference. During sunny days, your person projects a shadow on the other side of the light source. Those photons would’ve hit the ground but your body was in the way. That means no light is coming directly from the light source to that spot, which is the direct irradiance. Direct irradiance is energy directly from the source. However, if no light is coming from the light source and to the ground (because the body is in the way), shouldn’t the shadow be completely black? The reason it isn’t and you can still see things in a shadow is because of diffuse light waves.  Those are photons that bounce around our atmosphere and come from random directions. Those are the photons that land inside your shadow, preventing it from being pitch-black.

         The balance between direct/diffuse is telling. On clear blue sky’s at noon the irradiance is overwhelmingly direct. There is still some diffuse, but the atmosphere is minimally redirecting waves. On very cloudy days the irradiance can essentially be 100% since clouds are blocking all light directly from the sun.

|  |  |  |
| --- | --- | --- |
| Variable | Format | Description |
| Temperature | Integer | variable for BE in degrees C |
| Radiation Direct Horizontal | Float | weather variable for BE in W/m2 |
| Radiation Diffuse Horizontal | Float | weather variable for BE in W/m2 |

         The solar electrical generation dataset has more variables on it. Clearly the grid administrators use some predictive model to attempt to guess at the potential output. More analysis will be done to see their accuracy, and potentially use it as a baseline to compare to ours. The main variable we’re interested in is Corrected Upscale Measurement. This is the Megawatts generated by the solar facilities in the Flemish-Brabant region. This is data from the grid operators, not from the solar plants. The plants certainly have data on weather and their electrical generation but it’s likely not available for previously stated reasons. It is still pertinent to the goal of this paper—helping grid administrators predict solar output to better optimize meeting the grids demand.

|  |  |  |
| --- | --- | --- |
| Variable | Format | Description |
| Most Recent Prediction | Float | Most recent prediction of MW generated for this time block |
| Day Ahead Prediction | Float | Yesterday’s prediction of MW generated for this time block |
| Week Ahead Prediction | Float | 7 days ago prediction of MW generated for this time block |
| Corrected Upscale Measurement | Float | Megawatts of energy generated during that time block from the solar panels |
| Day Ahead Prediction (11 Hours) | Float | 11 hours prior’s prediction of MW generated for this time block |

         The Weather Underground scrapped data has the following variables. It is historical data from the Brussels airport where climatic conditions were collected every 30 minutes. This data had to be scrapped from HTML tables using Python. There were 17,498 observations. About 20 data points were missing.

|  |  |  |
| --- | --- | --- |
| Variable | Format | Description |
| Time | String | Date and Time |
| Temperature | Integer | Temperature in F |
| Dew point | Integer | Dew point |
| Humidity | Decimal | In percentage |
| Wind | Integer | Wind in mph |
| Wind Speed | Integer | Speed of wind in mph |
| Wind Direction | Categorical | Compass direction of wind |
| Wind Gust | Integer | Strongest wind gust |
| Pressure | Float | Pressure in inhg |
| Precipitation | Integer | Precipitation in inches |
| Condition | Categorical | Overall weather condition |

Data Preparation/Cleansing/Transformation

Data Preparation

         There was a significant amount of work needed done on the data. There were two separate weather datasets collected that will be combined together. The first was the NASA data that had temperature, direct and indirect sunlight. This data is hourly; however, the entire dataset needs to be in 15-minute blocks to fit the grid data. So, the hourly data had to be mapped onto the 15-minute data. This was done with linear interpolation, which involves taking the two hourly data points and generating numbers between them by assuming a linear relationship between the two. There were no missing data points in the NASA data.

         As a visual aide, look at the table below. The hourly data column is what was collected and known to be a true data point. When putting hourly data in 15-min. data creates missing data points, which are shown as “N/A'' in the table. Linear interpolation was used to generate data for points in-between the known data. We assumed a simple linear relationship between point A and B, then the 25th, 50th and 75th percentiles were used to create the missing data, which the highlighted numbers are in this example. This was done for the direct and indirect irradiance data.

|  |  |  |
| --- | --- | --- |
| Time | Hourly data | 15 min data created by calculation from Hourly data |
| 1:00 | 100 | 100 |
| 1:15 | N/A | 125 |
| 1:30 | N/A | 150 |
| 1:45 | N/A | 175 |
| 2:00 | 200 | 200 |

Another data set was scrapped from the Weather Underground website, where the Brussels airport weather station had data every 30 minutes. This same interpolation was done to get the data to 15 minute intervals, but in that case only the 50th percentile was used to create data, or the mean of the two data points. This scrapped data also had characters like “mph” and others that were captured in the scrape which had to be removed. This dataset collected twice every hour, at the 20th and 50th minute. This does not match other datasets which were for every 0th, 15th, 30th, 45th minute of each hour. This meant the scrapped dataset needed to be shifted by 5 minutes to match the times of other data, which was done, so this data in the final training dataset is actually +5 minutes in the future.

Due to the need for interpolation to match the data sets together, categorical variables were not included in the data set.

         Combining these two datasets provide a variety of climatic conditions. The NASA data has direct and indirect irradiance values which are important to solar productivity because it measure sun’s energy delivery to earth’s surface. The Weather Underground scrapped data provides many of the basic atmospheric conditions. These were things like temperature, humidity, pressure, wind speed, wind gust and dew point. Having all these at a model’s disposal for predictions should be useful because they are general atmospheric conditions. They are all numerical so graphing them and their correlations with others is possible.

There were variables that would’ve been useful for our model, but interpolation doesn’t make sense. These were wind direction and condition (like cloudy, sunny, etc.). Data interpolation does not make sense for categorical variables. Because interpolation was not possible on these variables, they were not usable in the final cleaned dataset because of too many missing values.

The target variable resides in the grid dataset, which will also be merged onto the final dataset by date and time. It is the contribution in Megawatts of the solar plants in the Flemish-Brabant region to the grid in that 15-minute period. The values remained unchanged in this data cleaning, but where the values occur changed and other target variables where created based on these for another modeling approach.

The target variable was unbalanced. About half of all observations had zero electrical generation due to night time. These were removed. This was to alleviate the imbalanced target variable issue. Additionally, using a solar prediction model for times where sunlight is not present does not make sense (like at 1am). It appears all early morning and late evening data was still preserved because there was enough light to produce some electricity (slightly greater than zero).

Data Cleansing

         Python and Excel were used to do the data cleaning.  Some of the open source libraries that make Python powerful were utilized to speed the cleaning process up.

         Not very much cleaning was done in Excel. It is slow and crashes often on my computer when there are 30,000 observations like there are on this project. Excel is very good at visualizing the dataset however. I find it easy to move around and look at the observations and check for issues.

         The actual cleaning and transforming was done in Python. It is much faster to make changes over many rows there. To make sure all the datasets were using the same data, the times were loaded into datetime objects. That way they are easy to join datasets at the same time, adjust times (like the 2 hour difference between Brussels and UTC), and correctly capture the values of date and time of each dataset because each one showed it a different way.

There were some missing values in the scraped data. These were data points that should have had collected actuals but didn’t exist in the data. There was no interpolation used to fill a missing data point that we should have had. It was only used to fill in the points between two known points. There were 20 of these missing data points.

         While creating the interpolation data it was important to determine the trend of the data. If the data was increasing, then the 25th, 50th and 75th percentile we used in that order so the generated data matches the trend of the known data. If the data was trending decreasing, the percentiles were placed in reverse, thereby matching the decreasing trend. The interpolation was done to loop through the actuals, and generate the desired percentiles using the numpy package.

Based on visualizations no outliers were detected and removed. There was no data point that appeared unreasonably high or low enough for its validity to be questioned.

There was some time zone adjustment as well. The NASA data was in UTC time, which is two hours behind Brussels scrapped time. This had to be adjusted for so the data was representing values from the correct time to match with the other datasets. Different datasets also had different definitions of when a day starts. The scraped data had a day starting at 1am instead of midnight. The dates were adjusted to show the correct dates so they all show their intended day starting at midnight.

The purpose of this analysis is to build a model to make predictions in the future. This requires training a model on data where the target variable is actuals some period of time in the future beyond the independent actuals. Thus, the target variable was shifted so that each target value is in observations with independent variables 30 minutes in the past from that observation. In other words, the independent variables contain actuals from a point in time, and the dependent variable have actuals 30 minutes in the future from that point in time.

The datasets were combined on matching date and time after making the corrections mentioned above. That way the observations from each dataset were merged so that they represented the same time. It took some playing with datetime objects in Python but it successfully combined observations from all three datasets (grid, NASA and scrapped) together with no errors or mismatches.

Data Transformation

      The one variable that was created was one that represents the ratio of direct irradiance. It was described earlier, but the higher the proportion direct irradiance makes up between direct and indirect, that is a good measure of cloud cover. Direct sunlight is what’s best for solar electrical generation.

         It is a simple calculation. A script looped through all rows, taking the direct irradiance value and dividing by the sum of direct and indirect value of that observation, then placing that value in a new column on the same row.

This value is important because this value is a good indicator of cloud cover. The more clouds that are present, the more water vapor blocks a direct route for a photon to leave the sun and hit the earth’s surface in a straight line. Cloudy days can achieve nearly 100% indirect sunlight, meaning essentially 0% direct. On non-cloudy days, the percent direct can get close to the 90th-percentile range. There are still particulates in the atmosphere that deflect photons on blue bird, cloudless days that have a measurable effect. Direct sunlight is best for solar because the panels can capture more of the energy the sun is producing, losing less energy to reflection.

One of the modeling approaches required some transformed variables for the logistic regressions. Each threshold (9) had a dummy variable, and if the actual solar generation was above it was marked as true. There was also a binned categorical variable created for the actual solar generation that matched the same 9 threshold binnings. It’s important to note these binning thresholds are not mutually exclusive, an observation can fit into more than one threshold bin if the actuals were large enough.

Data Analysis

      I am going to continue to use Python for the visualizations. Because of its utilization in the Data Science world I want as much practice with it as possible. It is also very customizable, so I can show complex graphs if needed. Much of my attention will be on what correlates well with solar generation, which is the target variable for prediction.

Python will also be the tool for doing the modeling as well. The amount of open source packages is plentiful. There are also many machine learning tools there that may be utilized which are reputable and with lots of documentation online. The real purpose for using it is for myself to get experience with it since its application in the data science world is wide.

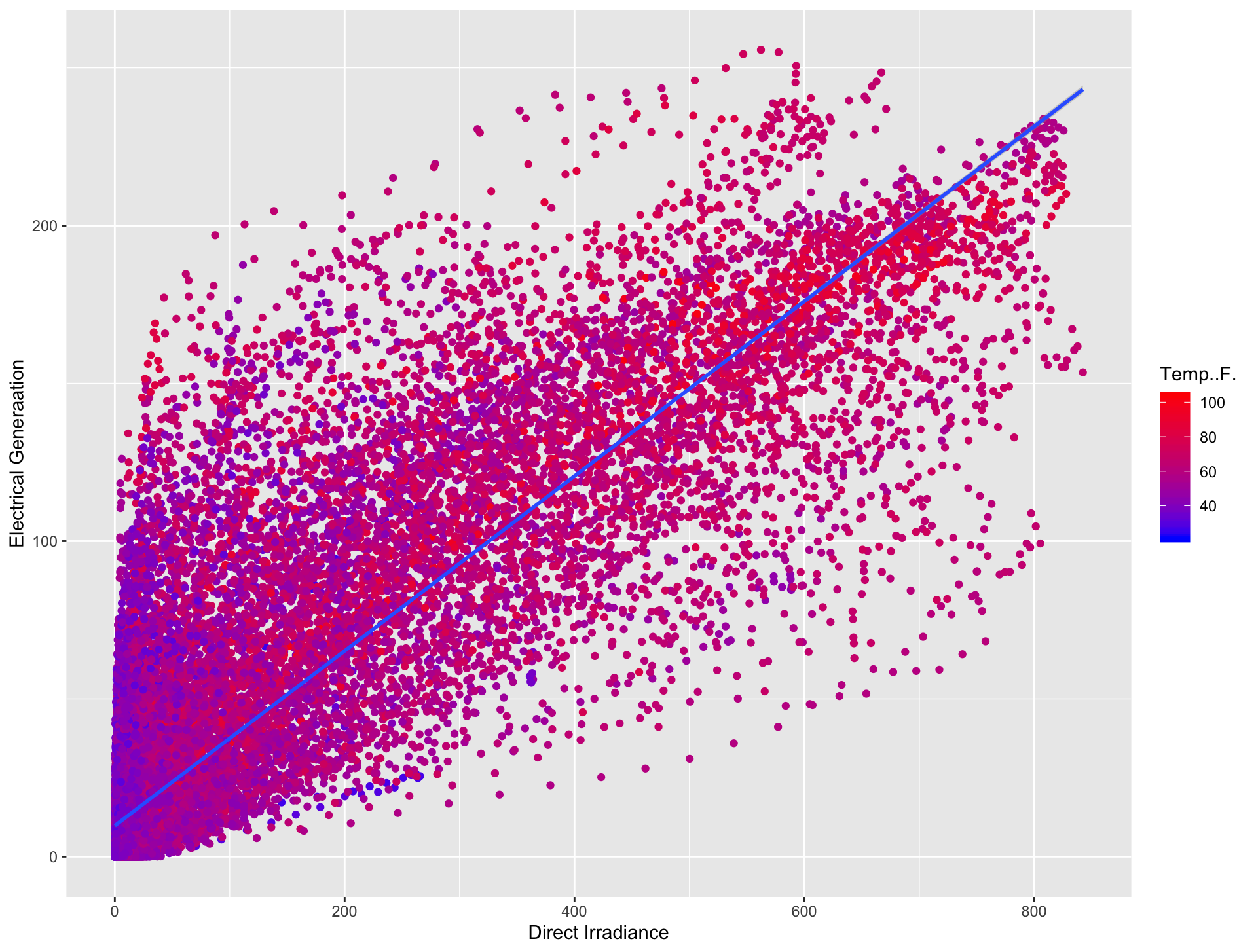
I have had success with neural networks before and hope to use one here. I plan on increasing the size of the NN until it starts showing signs of overfitting the training data.

         A regression would also be useful in showing how each variable affects solar generation keeping all else equal. Controlling for that would help better understand how each variable individually affects solar output for the region. I have had success previously with machine learning techniques and their accuracy, so I plan to explore that option. Another goal is to create a model that produces a sort of confidence level in its prediction. The idea is that grid administrates could look at the predication and take into account how much risk they are willing to accept in their planning around this model’s prediction. If they want a low risk option, they would plan for many back-up energy sources, and few if they went risky. The 9 dummy target variables will be able to achieve this by having a logistic regression for each one.

Data Visualization

The data visualizations where crafted in R using ggplot2 package. The goal was to inspect what where the variables that likely had the largest influence on the target variable, which was Electrical Generation.

Data Visualization 1



         The above visual shows the direct irradiance value vs the solar electrical generation for the region. The color of the dots represents the temperature of the 15-minute period. The data points alone show a clear positive correlation, but the blue line is a linear line of best fit, further showing the relationship between the two variables.

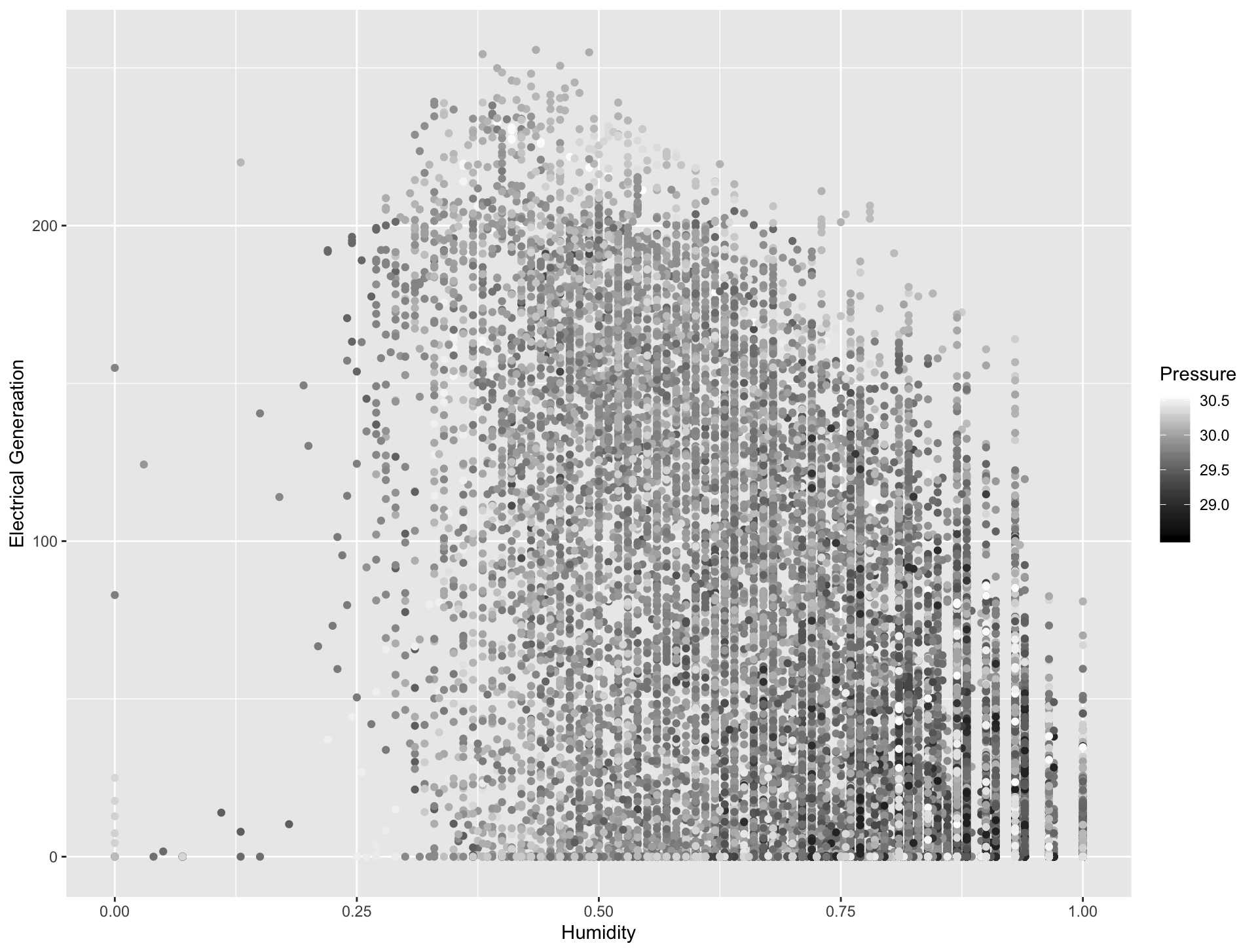
         It shows the distribution of electricity generation for a given direct irradiance level. For example, the largest MW generation during times with irradiance of about 50 is near the average amount of generation seen with irradiance at 650. It appears direct irradiance and temperature are positively correlated. The graph looks redder on the right.

         The first take-away from this graph is an apparent relationship between irradiance and electrical generation. This was expected to be seen.

Another relationship that was seen here is temperature and irradiance. Cold temperatures tend to be with lower irradiance and electrical generation values. This is not surprising since the more energy being delivered to the ground equates to more heat.

This visual was good evidence that direct irradiance was good predictor of electrical generation.

Data Visualization 2



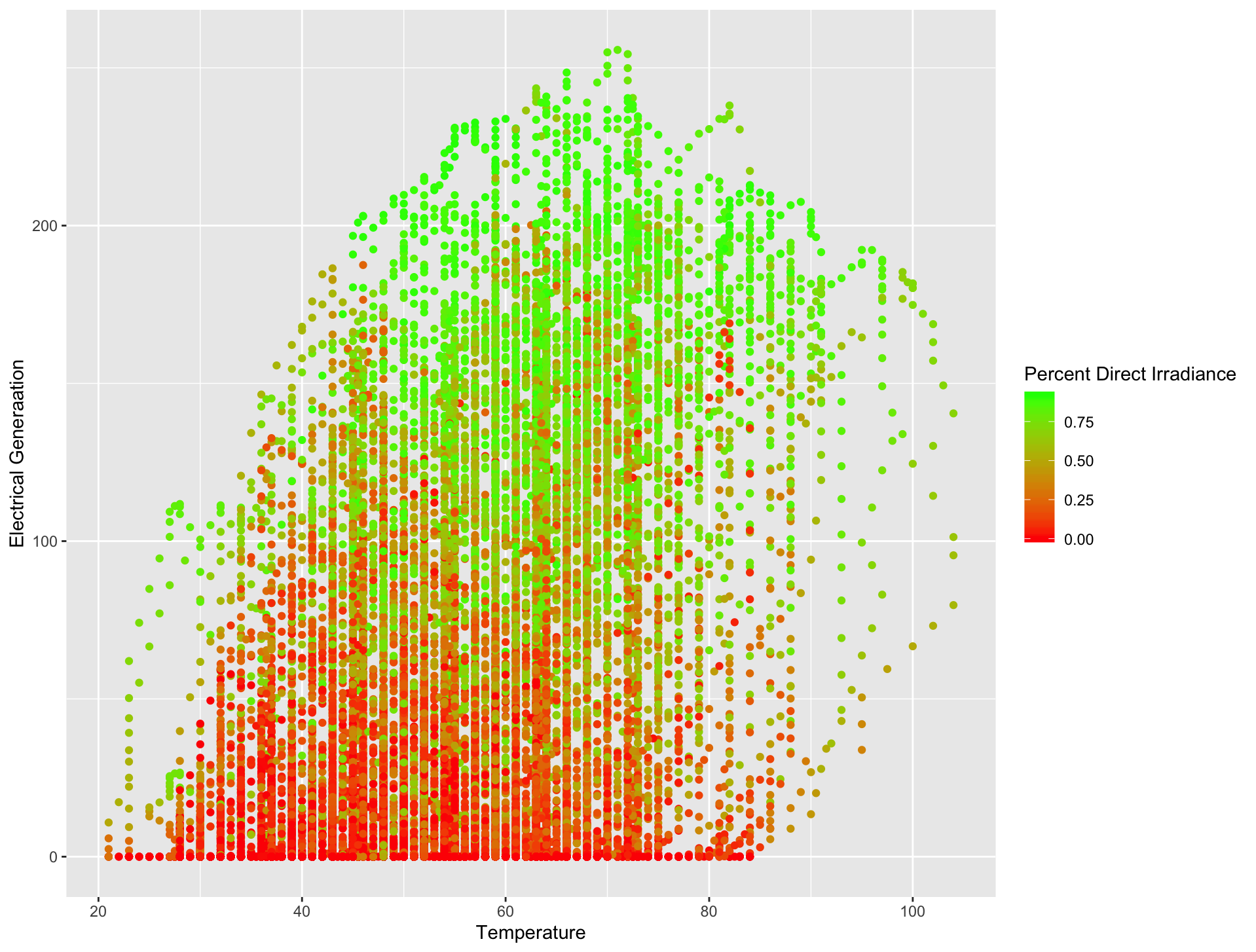
         The graphic above shows humidity vs power generation and is colored by the humidity. The humidity value is a value from 0 to 1. Humidity and power generation look to be negatively correlated. This may not be a surprising result. Humidity represents water vapor in the atmosphere, and the more of it the more diffuse the irradiance becomes. It may also represent cloudy conditions resulting in less direct sunlight.

         It appears most of the low-pressure points are in the high humidity-low power generation part. That is dark data points in the bottom-right section of the graphic. As a rule of thumb, low pressure means cloudy or rainy weather. That is because air is rising and condensing, forming water vapor droplets making clouds or precipitation. That would mean cloud cover which stifles solar panels effectiveness.

         Some things learned were that the humidity of the dataset is frequently between 50 and 80 percent. The lower the humidity the more power is generated generally. High humidity and low pressure seem to occur commonly together.

This visualization showed the role humidity and pressure play in predicating electrical generation. Low pressure and high humidity are not good for solar panels to be effective. This is likely due to how these are correlated with cloudy conditions, which reduce direct sunlight to the panels.

Data Visualization 3



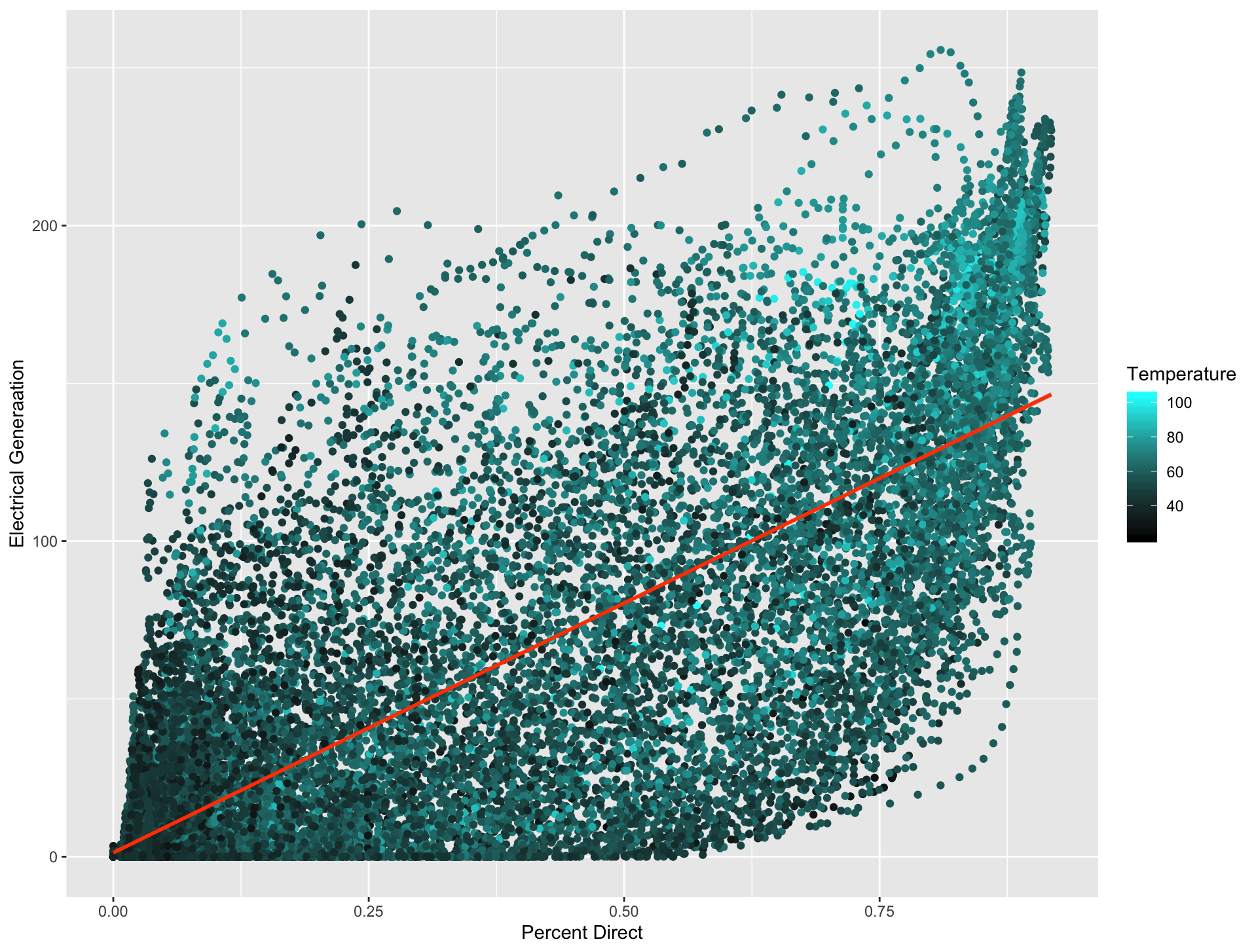
         The visual above shows temperature vs power generation colored by percentage direct irradiance. Green dots represent time blocks where direct is the dominant irradiance type and red is for diffuse.

         It contains the same information as the first graph but shows it a different way. This graph shows a more obvious positive correlation relationship between temperature and power generation. The top half of the data points are overwhelmingly green. Data that is very warm tend to have few points with low percentages of direct irradiance. And the opposite is true, low temperature data has very few points with high percentage direct irradiance.

It is obvious that the best days have a high percentage of direct irradiance. The points highest up on the graphic are all very bright green.

         Most hot days had both intense direct sunlight and good electrical generation. Interestingly, the most productive times were not on the hottest days. This may be because of the temperature distribution, where there are more observations at 70 degrees than 100, so there’s more opportunities for a great solar day to set a generation record. The low-temperature days are overwhelmingly low-direct irradiance. For data with over 200 MW generated by solar, almost none of them had direct irradiance less than 50%. For data less than 50 MW, the opposite is true.

Data Visualization 4



The above visualization shows Percentage Direct Irradiance vs the Electrical Generation of the region. The orange line is the linear relationship between them. The color of each data point represents the temperature.

The visualization shows an obvious positive correlation between the percentage of direct irradiance and electrical generation. The line shows the empirical evidence of it. On data where the direct irradiance is 50%, one should expect the solar panels to produce around 90 Megawatts of power.

There are some days that are very cold but still produce significant electricity. This is good news for solar’s reliability into the winter. There are less daylight hours during winter, but solar is still able to achieve significant generation while temperatures are around 40 degrees Fahrenheit.

The colors are brighter in the top-right section of the data points. High temperature is likely not the main reason the generation since solar panels need light, not temperature. Instead, high temperature must be correlated with conditions that are productive for solar panels. The bottom-left part of the data points is much dark in comparison.

The dots that make an obvious line are likely from the same section of time. Look at line of points that curve up on the bottom-right section. The conditions are gradually changing for every 15-minute period, but not dramatically enough for the next point to be very far on the visualization. One of the clues that this connect-the-dot lines are from the same day is the temperature is very nearly the same across those points. Time is not represented on this visualization, so it is unknown here the direction of these points with respect to chronology. There are also noticeable lines at the top of the data point cloud. This visualization is full of these lines, it just that there are so many points in places that they aren’t discernable.

Proposed Visualizations

         The above graphics proved what a lot of what was suspected. That was conditions that are related to direct sun light have significant correlation with electrical generation. Things that are related to cloud cover are good predictors of lower power generation. Temperature is a trickier relationship because it is related to all the other climatic conditions. It can also be high temperatures in morning or evenings where little sunlight can occur. It is also related to the time of the year. Ina any case I was going to input all the relevant conditions into models and let it find and use the relationships it finds.

Graphics on the time-series characteristic of the data would be of interest. For example, how much can solar productivity fluctuate in a given day? It may be the case that good generating days tend to be good during all times of the day. Good productive mornings may not be a guarantee of good productive afternoons as weather can roll in.

I’d be curious to see what correlates to direct irradiance as well. Humidity appears to influence it, but how much? What time of year is probably a large determinant of a day’s solar production as well. Winter days can have hours less sunlight.

The graphs built above focus on power generation because that is the target variable, but correlations between weather information would be interesting as well. I would be curious if decreases in pressure are correlated with other atmospheric changes later in time, like a leading indicator.

Predictive Models

         The purpose of these models was to help electrical grid administrators plan for the supply of electricity. To do this, these models are built to predict for 30 minutes in the future based on current conditions. So, all results from the models are its effectiveness in predicting 30 minutes in advance.

There were two rounds of models made. The first group, models 1, 2 and 3, used the below inputs to train and make predictions.

|  |  |  |
| --- | --- | --- |
| Temperature | Dewpoint | Humidity |
| Windspeed | Windgust | Pressure |
| Direct Irradiance | Diffuse Irradiance | Percent Direct of all Irradiance |

The other models, 4, 5 and 6, had more inputs. These were current conditions, the conditions 15 minutes ago, and the conditions 30 minutes ago. This was done to give the models more information on the climatic situation. Allowing a model to have a short window of conditions may improve predictions. The below table shows the inputs for these models.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Temperature | Dewpoint | Humidity | Windspeed | Windgust | Pressure | Direct Irradiance | Diffuse Irradiance | Percent Direct of all Irradiance |
| Temperature 15 mins ago | Dewpoint 15 mins ago | Humidity 15 mins ago | Windspeed 15 mins ago | Windgust 15 mins ago | Pressure 15 mins ago | Direct Irradiance 15 mins ago | Diffuse Irradiance 15 mins ago | Percent Direct of all Irradiance 15 mins ago |
| Temperature 30 minutes ago | Dewpoint 30 minutes ago | Humidity 30 minutes ago | Windspeed 30 minutes ago | Windgust 30 minutes ago | Pressure 30 minutes ago | Direct Irradiance 30 minutes ago | Diffuse Irradiance 30 minutes ago | Percent Direct of all Irradiance 30 minutes ago |

Predictive Model 1

         The first model built was the guaranteed-minimum bucket model. Its goal was to output a probability that a certain amount of electricity is going to appear. It was trained on training data (80% of the dataset) and tested on unseen data.

         To have a model output in this desired way, we decided a logistic regression is a simple and interpretable way to get a probability output. Each model was a default LogisticRegression model from sklearn library in Python with max 200 iterations. To achieve the desired output, it was decided that there would actually be 10 models built for this approach (one for each bin). This means 10 model outputs for each observation.

Below is an example of how it was done using less models and example data. Each model corresponds to a certain threshold. They were trained on the training data observations where the target variable was 1 when the actual was over that model’s certain threshold and 0 when it was under. So, when this model is shown a test observation, it returns a probability of this observation being over its threshold. The model with the largest threshold with over 50% probability is the bin that is used for the prediction.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| # | Observation actual (Megawatts 30 minutes in future) | Observation actual threshold bin | Model 1 output (prob) – 50 and above | Model 2 (prob) – 100 and above | Model 3 (prob) – 150 and above | Model 4 (prob) – 200 and above | Models threshold bin prediction (threshold at .5) |
| 1 | 10 | 0 | .21 | .1 | .02 | .001 | 0 |
| 2 | 74 | 1 | .84 | .43 | .33 | .15 | 1 |
| 3 | 115 | 2 | .89 | .61 | .42 | .10 | 2 |
| 4 | 230 | 4 | .94 | .85 | .73 | .65 | 4 |

      Taking the second observation for example, their actual MW value is 74 but there are many model predictions for the many thresholds. The model that predicts if this observation will have at least 50 MW outputs a True value (True was above .5 probability). However, for this same observation, the model for the 100 MW threshold predicts False, so the modeling approach believes it will be above 50 but not 100, which in this example is the right decision because the actual was 74 MW.

Why was this approach taken and not a simpler classification approach?

         Our chosen approach is different from the one-hot encoder or one-vs-all approach. In those approaches, the output would be a model’s best guess at which bin an observation is going to fall under. These bins would be 0-49MW, 50-99MW, 100-149MW and so on. That would mean an observation will fall under one and only one classification or bin.

         Our approach uses thresholds as the bins. This means a single observation can fit into multiple threshold bins if the MW’s are high enough. Take the fourth observation in the table above, the actual data is 230 MW. That means this observation is above 50MW,100MW, 150MW and 200MW, meaning this observation fits into all four threshold bins because it is above those thresholds.

         The benefit in this approach is each observation gets a probability output across all thresholds. That means the user of this approach can be more conservative/risky in what threshold they decide to use. For instance, the table shows us using the largest threshold with .50 probability or above was effective, all the observation actual bins equaled the model bin predictions for each observation.

But, if a user wanted to be more conservative, they could use the .70 threshold. That would mean the model’s bin prediction for 3rd observation would be 1 instead of 2, and the model bin prediction for the 4th observation would be 3 instead of 4.

If a user wanted to be riskier, they could use the .40 threshold. That would mean the model’s bin prediction for 2nd observation would be 2 instead of 1, and the model bin prediction for the 3rd observation would be 3 instead of 2.

This approach allows for this type of threshold selection. A one-vs-all approach does not have this ability to choose based on confidence level.

Results 1

         Below are the results of the approach and the effects of choosing different thresholds on test data. The more confidence one requires to select a threshold, the less likely one is to overestimate the solar generation (predict more electricity is going to be produced than actually was). One can see the slight tradeoff between the two accuracy measures.

|  |  |  |
| --- | --- | --- |
| Threshold on largest bin to make it the prediction | Predicted threshold equals  Actual threshold | Predicted threshold  equals or less than  Actual threshold |
| .4 | .5664 | .7609 |
| .45 | .5741 | .7931 |
| .5 | .5801 | .8221 |
| .55 | .5815 | .8495 |
| .6 | .5855 | .8759 |
| .65 | .5769 | .8953 |

To reiterate, grid administrators need to supply the grid the electricity it demands, otherwise power shortages will occur. Because solar generation is less reliable, grid administrators rely on mostly gas and coal plants to be at the ready in case solar generation is less than expected. That is costly and polluting.

A model that accurately predicts the exact right threshold (middle column) is efficient for both costs and pollution. The grid administrator is not paying for extra, unneeded energy from supplemental power plants and those plants are polluting less because they’re operating less.

A model that accurately predicts less than or equal to the right threshold level (right column) is good for the administrator because it shows a reliable minimum electrical generation they can rely on to ensure the grid has enough power. A low value here means a model overestimated the future solar energy production. This is bad because if a grid administrator trusted the model’s incorrect prediction and planned around it, they would come up short and create a power shortage or need supplemental power from other sources like gas/coal.

         It’s important to remember this is the model’s accuracy on 10 threshold bins, so anything better than 10% is better than random chance. So, at the .50 level the model correctly picks the best bin out of the 10 bins about 60% of the time. The tradeoff between the exact threshold accuracy and the reliable minimum accuracy is obvious from .55 to .60 and then .65 model confidence rounding value, as the exact accuracy decreases but the minimum reliable value increases.

Predictive Model 2

      The second model built was again predicting solar energy production 30 minutes in the future. This model was a Multi-Layer Perceptron Regressor from the sklearn library in Python. There are 3 hidden layers with 20 nodes each and with a ReLU activation function. The data was standardized-scaled so the data better fits the assumptions of the model.

         This model is a point estimate, unlike the classification type approach in the previous model. This model's approach is attempting to predict the numerical value of MW produced by solar plants in the Flemish Brabant region 30 minutes in advance of the current conditions. This approach will produce both positive and negative residuals.

Results 2

        The R-Squared measure can be defined as the percentage of the variation seen in the data that a model can explain for. That means our model successfully accounts for about 80% of the test data.

|  |  |
| --- | --- |
| R-Squared measure | .7949 |

         It is hard to compare this model to the previous one. They are approaching the same prediction from two separate techniques. Because of that they also don’t have the same accuracy metrics to compare. If there was more time, I would have analyzed the point estimate prediction better. For example, I would have shown a distribution of how many predictions were within 5MW of the actuals, and then 10MW, and so on to get a better idea of how close the model was able to get to the true value.

Predictive Model 3

This model is just like the above model but with deeper learning. This model has four hidden layers with 60 nodes in each. The rest of the parameters are the exact same. It is the same model type and point estimate.

|  |  |
| --- | --- |
| R-Squared measure | .8934 |

Results 3

The results were achieved with some tuning, meant to maximize the R-squared value on the test data. Not every conceivable combination of neurons was attempted, but it appears models with more nodes have overfitting issues. The training metrics keep improving but on the validation data they perform worse.

Clearly this model improved on the previous MLPRegressor model. The previous model was too simple and left uncaptured information on the table. The additional complexity to this model allowed it to better fit the underlying relationships present in the training data.

This model may be able to be used in conjunction with the first model. Although it is a different type, it may help in those edge cases (cases where a model can’t confidently choose between two categories). This accurate point estimate model may help decide which of the two categories that are too-close-to-call should be chosen. If I had more time I would have done an approach like this to see if combining these two approaches would benefit accuracy.

Predictive Model 4

The fourth model has everything the same as the first model, except it uses the 27 input variables stated above. All other things remain the same.

Results 4

Although this model had 3 times the amount of input variables, it did not produce major improvements in prediction performance. This model produced the results below.

|  |  |  |
| --- | --- | --- |
| Threshold on largest bin to make it the prediction | Predicted threshold equals  Actual threshold | Predicted threshold  equals or less than  Actual threshold |
| .4 | .5781 | .7669 |
| .45 | .5823 | .7940 |
| .5 | .5927 | .8284 |
| .55 | .5947 | .8513 |
| .6 | .5927 | .8741 |
| .65 | .5881 | .8939 |

This model’s best exact bin prediction was at the .55 rounding threshold, with an accuracy of 59.5%. Compare that to the first model, which used only 9 inputs, that’s best accuracy measure was 58.6% at the .6 rounding threshold. So, the 18 additional inputs, or 2 additional observations at 15 and 30 minutes prior to current conditions, only contributed to an additional 1% of accuracy.

This was an interesting result. It was surprisingly low for 18 additional inputs. The relationship between the current conditions and the 15 and 30-minute lagging conditions may not be a simple one that a coefficient can capture. It may be more complicated.

This model started declining in accuracy measures at the same rounding threshold as the first model, around .6. The model’s accuracy levels of being at or below the actual bin are mixed when compared to the first model’s. This model performed worse on thresholds .6 and .65 relative to the first model.

With this model’s added complexity, and the lackluster improvement in accuracy, I would be unlikely to actually implement it for predictions.

Model 5

|  |  |
| --- | --- |
| R-Squared measure | .9259 |

This model was the exact same as model 3, the NN with 60x60x60x60 node design. The only exception was this model had the same 27 input variables mentioned above. This point-estimate model was able to achieve 92.6% of R2 value. Compare that to model 3, which scored 89.3%. These values mean that this model was able explain 3.3% more of the variation seen in the data. Because the only difference was the two additional lagged conditions (18 input variables), the model was to able to control for 3.3% more variation in the data by including them.

This shows the diminishing returns of including more conditions. It’s hard to tell which had less diminishing returns to complexity, model 4 or 5. This is because they have two separate measures of accuracy, one was R2 and the other was bin accuracy. These are related but not really the same, so an apples-to-oranges comparison.

This neural network was expected to be more likely than model 4 to capture the complex relationships between the conditions. This is due to the NN being larger and more complex, thus able to capture more intricate and otherwise hidden information.

Model 6

|  |  |
| --- | --- |
| R-Squared measure | .9300 |

With the increased complexity of 27 input variables, would a larger NN perform better? In theory, it would be able to capture the complexity in the data. That was model 6’s goal, who has the same design as model 5 but had 4 layers of neurons each with 100 (compared to 4 layers of 60 neurons each in model 5).

The increased complexity was able to increase R2 by about .04. Model 6 had 66% more neurons than model 5. This is a large increase in computation cost for a small return on fitting the data better. This demonstrates that the NN approach is reaching an upper bound on fitting the data. Adding more neurons will eventually be detrimental, overfitting the training data.

Predictive Model Review

         There were two approaches taken in the model selection. One was a non-traditional classification prediction and the other was a point estimate. They are both potentially useful from a grid administrator’s perspective, both give them an idea of how much solar energy to expect in 30 minutes, which helps reduce costs and pollution because the planning for the energy supply becomes more predictable and therefore efficient. There were also two approaches to the model’s inputs. One was just the current conditions, with 9 input variables. The other was to have the current, 15 minutes ago and 30 minutes ago conditions, making 27 inputs in total (the same 9 inputs but over three periods).

         The multi-logit threshold approach had decent results, of the 10 bins it accurately selected the best one a little less than 60% of the time. It was also able to avoid overestimating the solar prediction 80-90% depending on the user’s risk tolerance. The additional input approach showed limited improvement in accuracy. This is likely due to the logit being less complex, capturing less of the information in the data.

         The point-estimate approach proved it can perform well. Simple models were able to explain about 80% of variation seen in the data, and with additional neurons was able to get up to 93%, capturing more information existing in the data. The additional input approach helped explain 3.3% more of the variation, which was a reasonable improvement given how well it was performing already. The complexity of the NN was likely what allowed it to perform well, both in the 9 input and 27 input approach.

         Ultimately the best model would likely be a combination of the point-estimate model and the multi-logit model. It would be an ensemble of the two. I would have done this if there was more time.

The model performance when inputs were increased from 9 to 27 variables did not increase as much as anticipated. Including climate data from 15 and 30 minutes prior to current conditions to predict 30-minute future solar productivity may not have all that useful information. This may be because the current conditions are closest to the prediction event, so they have the best predictive information. The other previous data may have useful information in them, but the current conditions do as well, so the addition of them provides limited improvement in comparison to only the current conditions inputs.

The accuracy metrics between the multi-logit threshold approach and NN approach are not apples to apples, but it is probably reasonable to say the NN performed better. This may be because of the simplicity/interpretability logistic regression offers. Neural networks aren’t hindered by these constraints. Decision makers may be required to use the logit approach for legal reasons because they require inspections of their systems. A NN with its black-box structure makes that very difficult.

         It’s worth noting that another modeling approach could have been taken to predict 30 minutes in the future. A model could use 30 minute predictions of the climatic conditions (independent variables) to then make a prediction of solar energy generation based on them (dependent variable). The reason this approach was not taken was because future predictions on all climatic conditions may not be available or accurate. However, a grid administrator can certainly get the current conditions and know them to be accurate. This approach has more certainty in the independent variables, less noise from other models making predictions of future climatic conditions and more real-world practicality than the other.

Final Results

Analysis Justification

         Countries need to ensure their hospitals, businesses and citizens have the power they demand or electrical blackouts will cause havoc. With the increasing concerns over pollution and its effect on the environment, countries need to balance their environmentally friendly efforts with their power needs. Solar energy is a great tool in countries' toolkits to do this.

         Solar’s downside is its variable electrical output. Unlike other energy sources, it is entirely dependent on the weather and sun. Other sources like coal, gas, nuclear and hydro aren’t at the whim of the weather. Because of this, electrical grid administrators plan back-up power plants to be ready in case solar power drops off or is not as productive as expected. Having extra backup power can be costly when not needed and polluting.

         The goal of this analysis is to help grid administrators maximize efficiency (minimize costly excess power generation and minimize pollution) by letting them plan around the predicted future solar power generation. If grid administrators know how much electricity solar plants will produce in a region, they can plan to maximize solar’s share of power generation, reducing pollution and costs.

Findings

         Accurately predicting a region's solar output 30 in advance showed reasonable success. For the threshold binning approach, the optimal threshold rounding value was around .5 and it was able to pick the optimal bin out of 10 bins just under 60% of the time. There is a trade-off between getting the exact bin correct and over-estimating solar’s output (overestimating solar’s output may risk not having enough energy for consumers). The less risk you take in overestimating, the less accurate about the correct threshold level it becomes. The point estimate approach was able to explain 93% of variation seen in test data using neural networks.

         Direct irradiance appears to be the most important metric in our data that is attributed to productive solar electricity production.

Review of Success

         The analysis was successful in determining that these models were able to predict the future solar outputs somewhat accurately. All the benefits like reduced cost and pollution stem from the ability to predict what the solar output will be. The grid data had the grid company’s own future predictions which could’ve been compared to our model’s, but they did not have a 30-minute prediction but a day ahead and week ahead forecast. Building models that are meant to rival the grid company’s models would have taken significant work because of the need for forecast of climatic conditions at that time.

         It is not known how useful a 30-minute model is to grid administrators. As stated earlier it can take hours for power plants to prepare themselves to produce energy. The farther into the future one has to make a prediction, the more a model would have to rely on climate forecasts, all with their own accuracy troubles.

Including previous climatic observational data showed limited improvement. Models with inputs only from T = 0 were nearly as good as models with inputs from T = 0, -1 and -2. This shows the additional data didn’t provide much predictive power.

Recommendations for Future Analysis

         This paper demonstrated that 30-minute future solar electrical production can have reasonable accuracy using only current climatic conditions (T = 0). More research should be done on how models that only use current conditions perform on farther out predictions. For obvious reasons, the farther in the future the prediction is, the less relevant the current conditions are to that prediction since there has been more time for things to change. I’d be interested to see if the most accurate model at T=+1 is always the best model (better than the other models) at T=+10 (or whichever advance time). A model design that delivers accuracy in near-future may not deliver accuracy in far-future.

         Companies are surely attempting to predict solar output in the future based on weather forecasts, but interesting approaches should be explored. For example, if the current conditions indicate the wind is coming from the west, what are the conditions 10 miles to the west, the idea being the weather there will be here in the future. Also, using satellite imagery of a region to look at cloud cover and cloud movement to the region of interest may be useful. Also, analyzing many forecasts and adjusting them if they tend to be above/below the actual values. A certain forecast source may consistently over-estimate a certain metric, at which point one could adjust it knowing that to get a more accurate value.

References

1. <https://www.eia.gov/energyexplained/electricity/electricity-in-the-us.php#:~:text=Electricity%20in%20the%20United%20States%20is%20produced%20(generated)%20with,diverse%20energy%20sources%20and%20technologies&text=Most%20electricity%20is%20generated%20with,geothermal%2C%20and%20solar%20thermal%20energy>.
2. <https://www.eia.gov/todayinenergy/detail.php?id=790>
3. <https://www.pv-magazine.com/2020/02/04/predicting-solar-power-generation-with-deep-photovoltaic-nowcasting/>