Understanding How Different Variables Impact New York's Solar Project Output

M. Kwawu, G. Munger, J. Aleiner, S. Kaltalioglu

Columbia University

Image from: https://emagazine.com/how-solar-power-works-to-energize-your-home/

Background and Research Question

The excessive use of non-sustainable energy sources, particularly fossil fuels, is the primary contributor to greenhouse gas emissions – the greatest factor driving climate change. As the severity of the climate crisis intensifies, and its impacts start to magnify, we must begin to change our societies and economies to mitigate its consequences. One necessary shift is to move towards more renewable forms of energy such as solar energy. Residential and industrial solar energy programs, such as mega grids, solar cooperative models, and home solar arrays, have been steadily gaining popularity in the past two decades. In this research project, we pay attention to residential solar because of its importance in gauging the public's enthusiasm about solar energy and also because our dataset pushed us in that direction. The average citizen is relatively far removed from large commercial solar projects, but the decision to put solar panels on their home, or invest in a community solar farm is a much better indicator that someone supports solar. Widespread residential solar also has the potential to decrease our reliance on fossil fuel power plants, since its growth is far outpacing fossil fuels.[1]

In examining residential solar we choose to look specifically at New York State solar projects – recently, many of the projects are smaller in scope since larger projects have gradually slowed down leaving smaller projects to boom.[2] This means that the number of projects, by itself, is not a good measure of New York's residential solar capacity. Thus, combining annual trends in the output of individual projects with trends in the annual number of projects would be vital in getting a fairly accurate picture of New York's solar output. With all this in mind, our project seeks to answer the following questions:

1. How effective has New York State been at increasing its residential solar output from 2003 to 2017?
2. Using data and trends from those years, how much residential solar output is New York State expected to generate by 2019?

Data Description [Link to Dataset]

Data for this study, provided by the state of New York through the New York State Research and Development Authority, outlines features of solar electric projects in New York, spanning from January 2003 to February 2018. The set contains 84721 observations (rows) and 30 features (columns) (see Appendix A). Each element of this dataset provides unique insight into the changes and progress in solar projects over the last two decades. By monitoring such changes, we can model which variables allow for the best prediction of expected solar output in NYC. This will ultimately tell us which variables contribute most to the output of NYC solar. The original ("solar-electric…") and cleaned ("new_data") dataset, as well as the jupyter notebook can be found here:

https://drive.google.com/drive/u/0/folders/1bS6gEVO0dQG9NxnwLpRCFyC76gHK6xk-

[1] https://www.energymonitor.ai/renewables/weekly-data-why-growth-in-solar-and-wind-is-truly-unprecedented/
[2] https://www.utilitydive.com/news/solar-growth-breaks-streak-tariffs-uflpa/638696/

Data Cleaning

The first thing we did in our project was clean the data, narrowing it down to only information we deemed relevant. Data cleaning was also important to ensure that data for each feature was in a suitable data type for easy extraction for our exploratory data analysis and predictive analysis. To clean the data, we changed the data type of all date columns to datetime; stripped all white spaces in column names; stripped off "$" signs and changed data to *int* type in the *Incentive* and *ProjectCost* columns; removed observations with missing *Incentive* and *DateCompleted* values; changed entries from "Yes" and "No" to 0, 1 (int type) in the *AffordableSolar* and *CommunityDistributedGeneration* columns (to enable potential use in predictive analysis); added the *duration* feature (*DateApplicationReceived* - *DateCompleted*); removed observations with negative entries for *duration* (projects completed before application to join the program); removed rows with information about commercial solar this was to enable us focus on residential solar projects (there were only about 300 of those entries); removed entries from 2018 as the dataset was completed in February that year so 2018 had relatively few observations, and removed 8 columns that we did not find useful. This brought our clean data to 74192 observations and 23 features.

Exploratory Data Analysis

To get a better understanding of our dataset, we performed some exploratory data analysis on our data. Some preliminary analysis included visualizing the relationship between the output of projects and their associated incentives, how many projects were completed each year, and incentive's effect on effective cost. We learned that the number of completed projects per year increased steadily until 2017, but that the output of projects and effective cost were not impacted by incentives. We learned that from 2003 to 2010, as average incentives increased, out-of-pocket costs increased as well. But from 2011 to 2017, as incentives decreased, out-of-pocket costs also decreased. From a layered graph of the *total* incentive, cost, and expected output, we learned that all 3 parameters (cost, incentive, output) rose fairly steadily from 2004 to 2015 and dropped from 2016 to 2017 (see Figure 1). Projects with a large total cost expected output were generally given higher total incentives. From a layered graph of the *mean* incentive, cost, and expected output, we learned that average cost, incentive, and expected output generally increased from 2003 to 2010 and then decreased from 2011 to 2017 as opposed to the drop in 2016 using the total values (see Figure 2). To understand how expensive it is to produce every kWh, we created a variable called *cost/kW*. A graph of the average cost/kWh over time showed that cost/kWh increased slightly from 2003 to 2008 and then decreased rapidly from 2009 to 2013 and then decreased steadily till 2017 (see Figure 3). Additionally, we aimed to understand the effect of incentives on the cost of the projects, so we made a very similar graph that was called *effcostperkW.* This graph demonstrates how the average effective cost/kWh increased till around 2007, then steadily declined until 2013, (potentially with an increase in the program's popularity and breadth) with a small increase until 2015, and finally a decrease in cost into 2018 (see Figure 4). We defined a new variable called *efficiency* that corresponds to the total kWh output per PV module used. A graph of the average efficiency over time

(2003 - 2017) showed a steadily increasing line which means that the solar panels became more efficient with time (see Figure 5). Finally we wanted to understand how well our predictors were correlated to *ExpectedKWhAnnualProduction*. To do this we used a simple correlation measurement (see Table 1) which shows that *ProjectCost* and *TotalPVModuelQuantity* are highly correlated with *ExpectedKWhAnnualProduction* while *Incentive* and *Duration* are less correlated.

Predictive analysis

Our overall goal was to build an accurate model to predict expected solar output (ExpectedKWhAnnualProduction), for 2018 and 2019, based on 5 key variables; Duration, Incentive, ProjectCost, TotalPVModuleQuantity & the previous year's output. We determined that if these variables created an accurate model in which each was statistically significant, we could conclude that those variables do affect output and that a similar prediction method could be applied to predict solar output for future years.

To add the previous year's output column, we found the average output for each year and entered those into a separate dataset. We then shifted those observations down by one year and merged that dataset with our original dataset to add this column. This gave each observation its associated previous year's average output. After that, we divided our dataset, based on each project's year of completion, into a training set (40218 observations from 2003 to 2015) and a testing set (33908 observations from 2016 and 2017). We then used the OLS linear regression model ("model1") to predict expected output using *Duration*, *Incentive*, *ProjectCost, TotalPVModuleQuantity*, *previous_years_output* as predictors. The model (see Table 2) is as follows;

(*predicted*) *ExpectedKWhAnnualProduction* = -2842.092957 + *Duration*(-2.859735) + *Incentive*(-0.061599) + *ProjectCost*(0.098877) + *TotalPVModuleQuantity*(192.155886) + *Previous_Years_Output*(0.269916).

We then ran a cross-validation by using our model to predict the *ExpectedKWhAnnualProduction* for the entries in our testing dataset (later year data) in a new column called *predict*. We evaluated a very strong correlation between predicted values and the actual values (correlation= 0.9895097222629332). This shows us that the model we created based on data from earlier years is very accurate for predicting future output values. To determine if all our predictors are statistically significant, we ran a 95% confidence interval on model1 (see Table 3). Considering all values are of the same sign for each side of the interval, we can be assured that each predictor in our model is significant in predicting output.

We also tried another model created by using a nested method of predicting individual variables, and then using those predictions as variables in an OLS regression. As shown in Appendix B, the correlations between the actual output and the predicted values were very close for Model 1 and 3, the mean squared error was lowest for Model 1 and the individual predicted values in each model varied on their level of similarity to the actual values. Therefore, we chose to continue using Model 1 for our predictive analysis.

Now that we have proved that *Duration*, *Incentive*, *ProjectCost*, *TotalPVModuleQuantity*, and *previous_years_output* are good predictors for output, we want to extend our model to predict total output in future years. Our idea was that we could use model1 to predict the average output of any project completed in a certain year in the future. If we multiplied that output by the number of projects we predict to be completed that year, we could get the total output constructed that year.

To begin, we evaluated the correlation between each of our variables and time. The very low correlation values for *duration* and *TotalPVModuleQuantity* showed that time was a poor predictor for duration and number of solar panels. Thus, we can keep these variables constant (at their average value over the whole data set) when creating a model that will make predictions for future years. To run a polynomial regression model to predict the number of projects completed in each year, we created a new feature, $year^2$ ($year\ squared$). We then used *groupby* to find the total number of projects completed in each year and then used our polynomial regression model to predict the total number of projects completed in future years. Using the predicted values of this model, we run a for loop using the beta values from "model1" to predict expected outputs for 2018 and 2019. To find the total expected output for 2018 and 2019 individually, we multiplied each year's predicted average output with its predicted number of completed projects. We then added the predicted total outputs for 2018 and 2019 to the total outputs from 2003 to 2017. According to our model, by the end of 2019, New York will have had about 150.8 MW total output for residential solar.

Discussion

According to the Solar Energy Industry Association, New York's solar energy output for residential projects in 2019 was about 150 MW (see Figure 6), which is very similar to our model's prediction of about 150.8 MW. This demonstrates a capable model to predict and anticipate New York's solar energy output for residential projects. To extend this model to predict 2030's output(and therefore identify whether New York will hit its solar energy goals), we would need data from more recent years, as any errors would compound over such a long period of time. Due to the lack of data for commercial and industrial solar projects, we were unable to predict the energy output from those projects, and therefore a complete total for NY's solar output. Another limitation of our dataset is that it does not have any information about the construction sites other than their general location. If certain projects are constructed in suboptimal sites, that could affect expected output, project cost, and even duration. For example, a certain site in a mountainous area might be very expensive to construct in and have less direct sunlight leading to less output. Another limitation of this dataset is the lack of information on the long-term health of these projects. We assumed that once built, these projects continue to operate at peak efficiency, and do not break down or come offline. This likely led us to overestimate NY's solar capacity.

APPENDIX A: Variables that are struck out were deleted in our data cleaning process

Column name	Description
~~Reporting Period~~	~~The period of time the data was collected. (Date)~~ *all reports were made on the same day
Project Number	Unique identifier for each project. (Number)
City	The city where the project is located. (Text)
County	The county where the project is located. (Text)
~~State~~	~~The state where the project is located. (Text)~~ *useless because all are in new york
Zip Code	The zip code of the project location. (Number)
~~Sector~~	~~The sector the project is associated with. (Text)~~ *Sector and program type describe very similar ideas and have very similar data, however, program type contains more descriptive data than sector.
Program Type	The type of program the project is associated with. (Text) residential 0 v industrial 1
Solicitation	The solicitation the project is associated with. (Text)
Electric Utility	The electric utility (electricity generator and distributor) associated with the project. (Text)
Purchase Type	The type of purchase associated with the project. (Text)
Date Application Received	The date the application for the project was received. (Date)
Date Completed	The date the project was completed. (Date)
~~Project Status~~	~~The status of the project, either completed or non completed(Text)~~ *only looking at completed projects
Contractor	The contractor associated with the project. (Text)
Primary Inverter Manufacturer	The primary inverter manufacturer associated with the project. (Text) the manufacturer that produced the majority, if not all of the solar panels on the project,
Total Inverter Quantity	The total number of inverters associated with the project. (Number) the number of inverters on the project photovoltaic system
Primary PV Module Manufacturer	The primary PV module manufacturer associated with the project. (Text)
Total PV Module Quantity	The total number of PV modules associated with the project. (Number)
Project Cost	The total cost of the project. (Number)
Incentive	The total incentive amount associated with the project. (Number)the monetary incentive the project received
Total Nameplate kW DC	The total nameplate capacity of the project in kW DC. (Number)
Expected KWh Annual Production	The expected annual production of the project in KWh. (Number) *Assuming expected KWh is actual annual production
~~Remote Net Metering~~	~~The remote net metering (an indicator of use for solar cooperative model or not)~~ *Deemed unnecessary in analysis
~~PV Module Model Number~~	~~The primary PV module model number associated with the project. (Number)~~ *Deemed unnecessary in analysis
~~Primary Inverter Model Number~~	~~The primary inverter model number associated with the project. (Number)~~ *Deemed unnecessary in analysis
Affordable Solar	subsidized solar for lower income areas
Community Distributed Generation	subscribe to solar farm
Green Jobs Green New York Participant	new york incentive
~~Location 1~~	~~latitude and longitude~~ *this specificity is unnecessary

Relevant Graphs/Images Referenced

Figure 1:

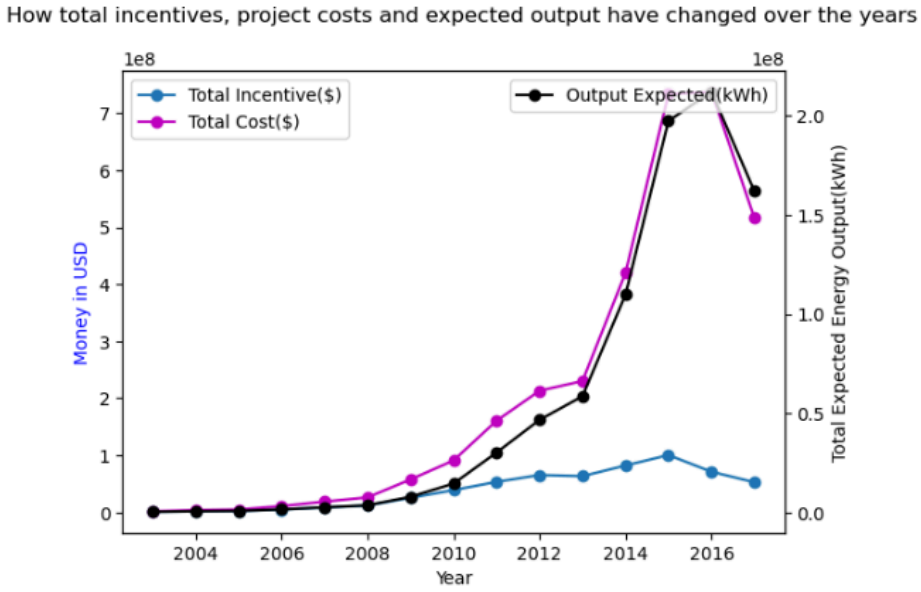


Figure 2:

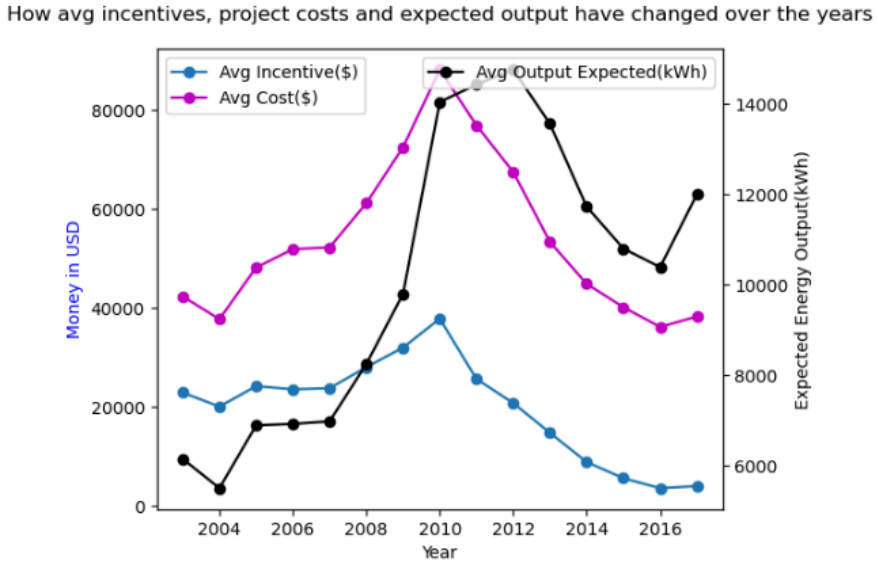


Figure 3:

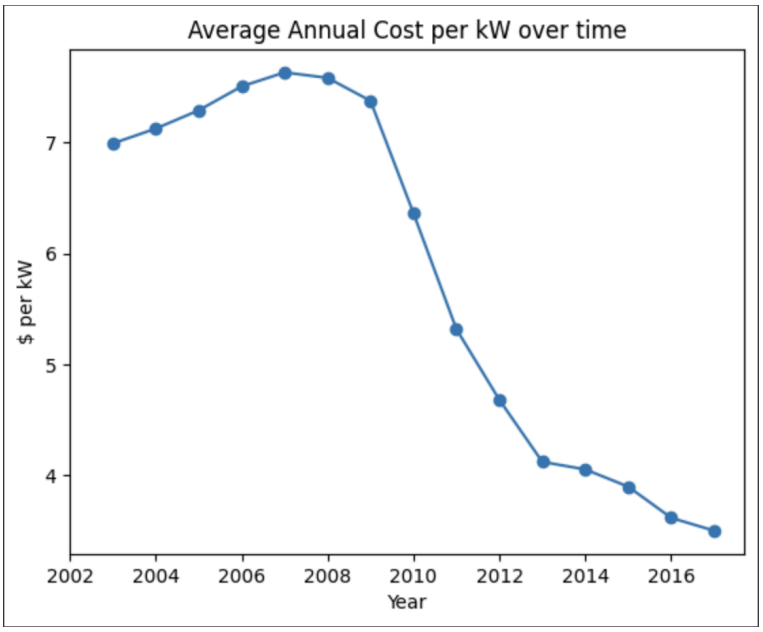
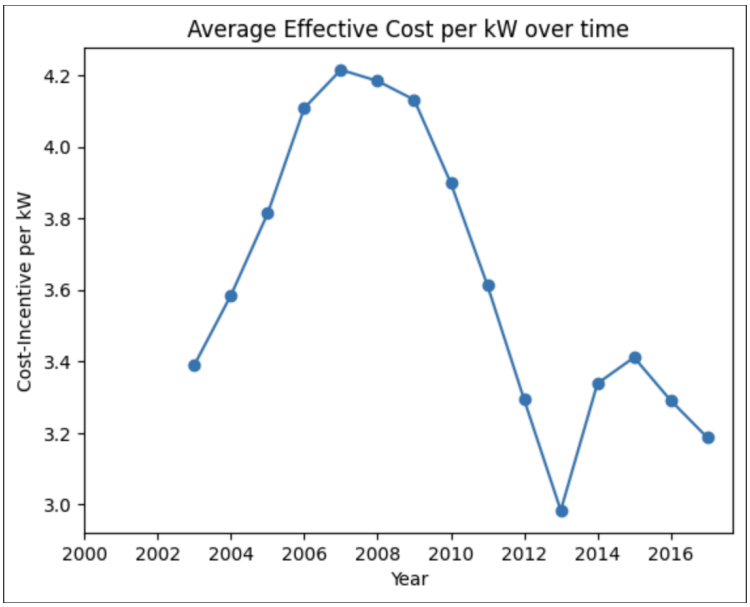


Figure 4:



*effective cost is project cost minus its incentive amount

Figure 5:

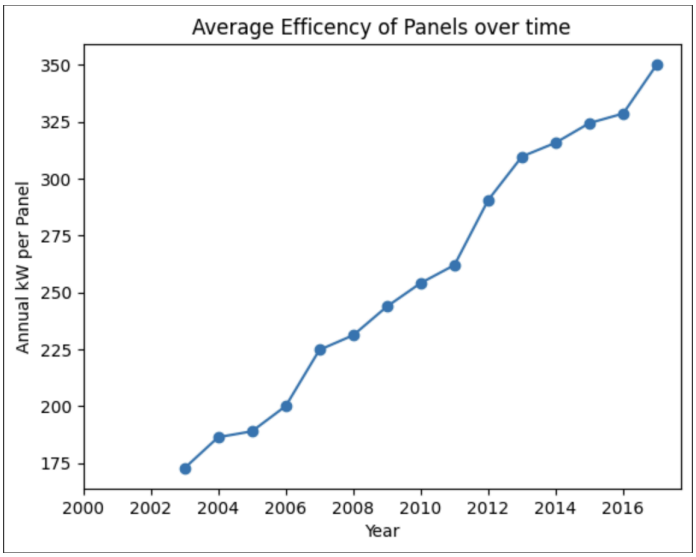


Figure 6: This figure depicts New York Annual Solar installations per year. Since we are only considering residential communities, our prediction for 2019's average MW output is very close to the actual 2019 MW output. https://www.seia.org/state-solar-policy/new-york-solar

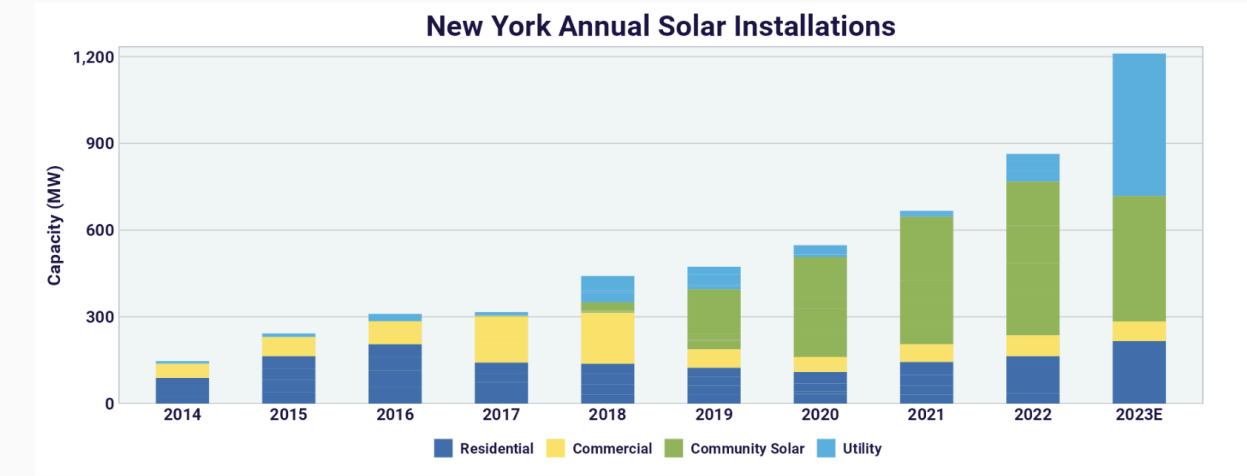


Table 1: Correlation between *ExpectedKWhAnnualProduction* and predictive variables *ProjectCost*, *TotalPVModuelQuantity*, *Incentive,* and *Duration*

Parameters	*ExpectedKWhAnnualProduction*
Total PV Module Quantity	0.951046
Project Cost	0.900064
Incentive	0.634686
Duration	0.203408

Table 2: Coefficients from Model 1

Parameters from Model 1	Value
Intercept	-2842.092957
Duration	-2.859735
Incentive	-0.061599
Project Cost	0.098877
Total PV Module Quantity	192.155886
Previous Year Avg Output	0.269916

Table 3: Confidence Intervals of predictors used in Model 1

Parameters	Lower bound [0]	Upper bound [1]
Intercept	-3289.134101	-2395.051812
Duration	-3.349720	-2.369749
Incentive	-0.066535	-0.056664
Project Cost	0.096898	0.100855
Total PV Module Quantity	190.377441	193.934331
Previous Year Avg Output	0.235543	0.304288

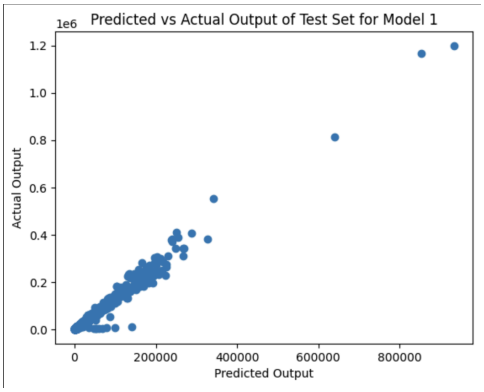
APPENDIX B: Comparing the Performance of 3 Prospective Models

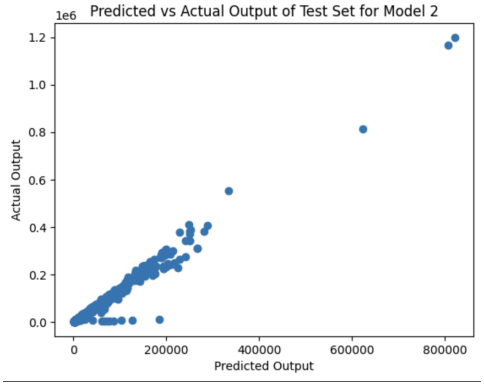
In order of appearance: Model 1, Model 2, Model 3

Table B.1. Correlations and Mean Squared Error between Actual Values versus Predicted Expected KWh Annual Production (Highest Correlation and Lowest Error Bolded)

Model	Correlation	Mean Squared Error
Model 1	**0.9895097222629332**	**38099654.49398671**
Model 2	0.9323865078507847	46650767.29721333
Model 3	0.9892387914017943	42513854.016383916

Figure B.2. Visualization of Correlations between Actual Values versus Predicted Expected KWh Annual Production





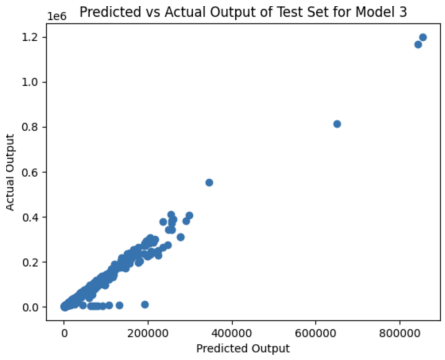


Figure B.3. Comparison of Actual Values and Predicted Expected KWh Annual Production

	ExpectedKWhAnnualProduction	predict1	predict2	predict3
3	5758.0	4704.894275	5098.048753	4761.004354
22	13793.0	12107.024515	12757.916154	13122.906050
51	48080.0	41605.192312	34790.047611	35008.566892
67	5374.0	4755.207544	4732.320700	4471.159952
79	7248.0	100090.365123	126206.530876	132693.410783
...	...	...	...	...
74113	11198.0	9973.395831	10747.148729	10421.308832
74116	10576.0	9099.219848	10152.972024	9883.542285
74119	9643.0	8424.661625	9271.926276	9047.415879
74123	9332.0	8241.930066	9159.219260	8825.040109
74125	11504.0	10146.550114	10511.693912	10170.111511

There is some clear inconsistency in which model is better at predicting individual values.