WGU Data Analytics Graduate Capstone

Neural Network Prediction of NYSERDA PC Incentive Funds

By: Christopher Kamper

April 2024

**Research Question**

**Question:** Do the Project Cost and Expected KWh Annual Production variables have a significant impact on incentive amounts to be used to predict a high or low amount for a PV project under the NYSERDA program?

The research question driving the analysis comes from a need to be able to predict the possible funding a photovoltaic (PV) project may receive under the New York State Energy Research and Development Authority (NYSERDA) incentive program. Recent interest in utilizing renewable energy sources has created an increase in the number of individuals and businesses looking at using PV panels as a solution. It was reported that in 2016, 4% of homeowners considered installing these panels on their homes, and this number increased to 6% by 2019. 92% of these respondents noted that saving money was their driving factor since most costs associated with this solution are one-time and upfront. (“Home Solar Panel Adoption Continues to Rise in the U.S.”) With money being such an important factor, the cost could be the difference between a project being completed or not. Government agencies such as NYSERDA have begun these programs to help provide project funding that can play a large role in reducing the out-of-pocket costs that an individual or small business may have to utilize a PV solution. It has been reported that 64% of individuals looking into PV solutions were interested due to government credits that could help reduce their expenses. (“Home Solar Panel Adoption Continues to Rise in the U.S.”) Considering the importance of the overall project cost, it would be helpful for contractors and their clients to understand what the costs would be taking these incentive dollars into account. Being able to know if a project is expected to receive a large or small amount of funding could very well determine the viability of that project.

**Hypothesis:** A statistically significant predictive model can be built from the Project Cost and Expected KWh Annual Production variables to predict the incentive amount as high or low.

**Null Hypothesis:** A statistically significant model cannot be built from the Project Cost and Expected KWh Annual Production variables to predict the incentive amount as high or low.

The hypothesis developed supports the idea that an artificial neural network model (ANN) can be developed based on the data set to predict if a project will receive high or low funding with statistical significance. This is based on the idea that a neural network can be created to provide a single classification outcome based on multiple variables. Considering the size of the data set, there is an optimistic assumption that there is enough data to train the model well enough to make an accurate prediction. This does not directly mean the model is expected to be significant though as the features chosen may not have a direct enough effect on the outcome of the classification. If this is the case, the null hypothesis will be confirmed.

**Data Collection**

The data was sourced from the NY Open Data website here: <https://data.ny.gov/Energy-Environment/Solar-Electric-Programs-Reported-by-NYSERDA-Beginn/3x8r-34rs/about_data>. A CSV file was downloaded that included all features and project data that was available as of April 2nd, 2024. This included the Project Cost, Total NYSERDA Incentive, and Expected KWh Annual Production features that are required for the analysis. The CSV file format allows for the data to easily be imported to a Pandas data frame for analysis using Python. This is an advantage due to the ease of use and ability to access the data without concern with the format. Downloading and then loading a CSV file makes the process clean and simple compared to other methods that may require other formatting or challenges before getting the data to the data frame. This method does come with the concern over importing quite a bit of extra data though. Since the file exported includes all features available, there are quite a few that are not needed for this specific analysis. Therefore, there are cleaning steps that will be needed to remove the features that are not going to be required.

**Data Extraction and Preparation**

Python was used for this due to its ability to easily maintain a data frame through the preparation process with Pandas and provide good support for any modifications that are needed for this analysis. This includes the ability to remove unnecessary data, calculate new values, and ensure the data type of each value is as expected. (GfG) One possible concern using Python and the Pandas data frame though would be complications in adding data easily as more information may become available.

Once the data set has been downloaded from the NY Open Data website as a CSV file, it is loaded to a Pandas data frame and reviewed to ensure it looks as expected.

A screenshot of a computer

Description automatically generated

Since it is known that there are features in the data set that are not required for this analysis, these will be dropped to aid in processing. The unique identifier "Project Number" column will also be kept ensuring that is available for reference if needed. This will leave the "Project Number", "Project Cost", "Total NYSERDA Incentive", and "Expected KWh Annual Production" columns. The data frame will then be reviewed to ensure all looks as expected.

A screenshot of a computer program

Description automatically generated

Since this analysis utilizes only a few columns from the larger data set, there is quite a bit of data that is removed. An advantage is reducing possible compute time since there is less to reference, but it also means data that could be valuable to further analysis would be lost.

The "Project Number" column is a unique ID for each project included in the data set. As a result, it can be used to ensure that there are no duplicated rows. After using drop duplicates, the data frame will again be reviewed to see if any rows were dropped.

A screenshot of a computer

Description automatically generated

There are no duplicated values in that column, so no rows were dropped based on the non-null count for each column.

Any null value in the remaining columns will be removed for this analysis so a count of any by each feature will be reviewed.

A screenshot of a computer code

Description automatically generated

There are some null values in the "Total NYSERDA Incentive" and "Expected KWh Annual Production" columns so those corresponding rows need to be dropped.

Any row containing a null value will be removed using drop na. The data frame will then be reviewed to ensure this has been done.

A screenshot of a computer

Description automatically generated

Comparing the previous non-null counts shows that those rows have not been dropped so only rows with all features available are in the data set.

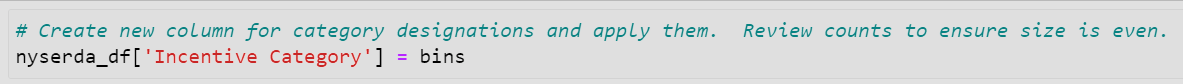
Multiple methods can be used to deal with having null values in a data set and for this analysis, it is chosen to drop any of these rows. This is useful in that one knows the lack of meaningful value is not going to have a negative impact on the calculations in the analysis. However, this does mean that some values from other features that could be useful may be dropped.

The "Incentive Category" needs to be added to the data set based on the value in the "Total NYSERDA Incentive" column. The values will be split into two bins that categorize each as a "low" or "high" amount. These will be represented in the data set as "0" for low and "1" for high. This is completed using the qcut function with two bins and the labels previously mentioned based on the "Total NYSERDA Incentive". The data frame will be reviewed to ensure this column has been added with expected values.

A screen shot of a computer

Description automatically generated

The values from this calculation need to be added to the data frame as a new column. The column "Incentive Category" will be created and filled in with those values.



The research question references predicting a binary categorical value for the incentive amount. Since the "Total NYSERDA Incentive" is available, the values can be broken down into these desired categories. This simplifies the outcome for the model which can improve accuracy. The drawback to utilizing this instead of an actual numeric value or more categories is that some granularity in knowing the actual value predicted is lost. More categories could provide extra detail on what that incentive value could be but add more complexity to the model that may not improve accuracy or significance.

The final data frame will be reviewed, and the first 5 rows looked at to ensure all is looking as expected.

A screenshot of a computer

Description automatically generated

The cleaned and prepared data frame is exported as a CSV file.

A close-up of a computer screen

Description automatically generated

The final preparations for the model analysis come by splitting the data set into multiple groups for training and testing. The train test split function allows for this to easily be done. This is used to split the set based on the 80/20 rule that is commonly used. (Roshan) A new data frame will be loaded containing just the columns for the features needed from the original data frame. The category type will be set to ensure it is a float value before being split into the training and test sets. This will help ensure that there is sufficient data to train the model while still holding some for testing. This does mean the full data set won't be available to train on as much data as possible, but the testing is an important part of the process and fresh data is needed to do this properly.

A screenshot of a computer program

Description automatically generated

**Analysis**

Exploratory analysis was used to determine some basic information about the data set being used for the model.

First, the summary statistics of the data frame are viewed. This information includes the mean, min, and maximum values for each of the features to be used in the model. These values do not directly give insight into the model itself, but they do provide an idea of the spread between the higher and lower values for each while including the average that can give guidance about where a project falls even if the model isn't significant.

A screenshot of a computer

Description automatically generated

Since there are two bins being used for the categorical values, checking to see that the count of each category is important. It is expected that these should include 50% of the data in each category, but there is a chance of repeated incentive values. This means the median value will act as the limit for determining each bin. The median is found and compared to the counts that are above and below.

A screenshot of a computer

Description automatically generated

The count values match those from the previous calculation based on the median value showing that is the limit where values below are in the low category and values greater are in the high category. These values are plotted to give a visual representation of how the categories are close to having equivalent counts.

A screenshot of a computer

Description automatically generated

Since two features are being used to determine the categorical outcome in this model, it can be useful to see if there is a strong relationship between the independent variables and the dependent. The "Total NYSERDA Incentive” values are plotted against the "Project Cost" while being colored based on the "Incentive Category" to see if there is a clear relationship between them.

A screen shot of a graph

Description automatically generated

The plot shows that there is some correlation of the variables as the low points do not show on the higher value areas of the graph but there is some overlap between the categories since the higher category values cover the majority of the lower category.

A similar plot is created for the "Total NYSERDA Incentive" values against the "Expected KWh Annual Production" with the same coloring.

A screenshot of a computer screen

Description automatically generated

The plot shows that there is some correlation of the variables as the low points do not show on the higher value areas of the graph but there is some overlap between the categories since the higher category values cover the majority of the lower category.

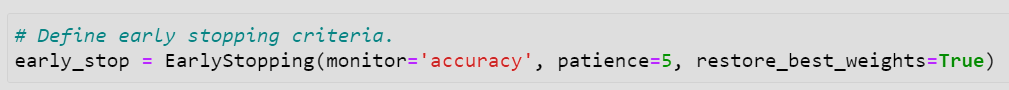
These graphs help see if a variable will have a strong effect on the outcome of a model, but it is not an exact method that would provide a calculation that could help aid in the model itself. The advantage comes from getting an idea of how effective these variables may be through the visual medium, but the issue is a lack of detailed information that could be used to affect the outcome of the model.

The goal of this analysis is to create the actual model to prove support for the hypothesis. The parameters for this model need to be set. Since the goal is to have one outcome from the model, the output dimensions will be set at 1. The outcome is a binary value, so the loss function chosen is binary cross-entropy. Adam has been chosen as the optimizer due to its efficiency in working with larger data sets while combining the benefits of gradient descent methodology and root mean squared propagation (RMSP). (GfG)

A screenshot of a computer code

Description automatically generated

Early stopping criteria is going to be used as a method of limiting the run time of the model while also helping minimize overfitting of the model during training. It will be based on the 'accuracy' of each epoch with a patience of 5. A possible disadvantage of utilizing this could be missing an improved model based on the criteria but, the patience helps ensure this risk is minimized as is set as 10% of the total epochs in this analysis.



The model itself is designed to input both features using the training set previously created. It will input to dense layers of the model going from 128 nodes to 64, 32, and then 16 before going to the final output layer giving the one output value. The relu activation function is chosen for the input and hidden layers since it adds nonlinearity to the model and helps solve any vanishing gradient issues. It sets any positive value as the maximum of the function while setting any negative value as 0. A disadvantage to this approach is that any negative values are set to zero meaning these values are not properly addressed by the model. (“An Introduction to the ReLU Activation Function”) The sigmoid function is used in the output layer due to the binary output since it will interpret the input as a range between 0 and 1. (GfG) The model is then compiled utilizing these parameters.

A screenshot of a computer

Description automatically generated

Now that there is a model constructed, the training data can be used to train the model. This is set to utilize the number of epochs from the parameters, a 20% validation split of the test set, and utilize the early stopping criteria.

A screenshot of a computer

Description automatically generated

The model stops training after the 8th epoch due to the stopping criteria.

The model is reviewed by checking on the loss and accuracy of the results of the training.

A screenshot of a computer program

Description automatically generated

This shows that while the model is 72.15% accurate, there is concern due to the loss value at 0.5575. This signifies that there is a larger amount of error in the model than what would be considered reasonable.

The metrics of the model accuracy are plotted to give a visual representation of how the training and validation steps for each epoch perform.

A screen shot of a graph

Description automatically generated

The metrics of the model loss are plotted to give a visual representation of how the training and validation steps for each epoch perform.

A graph with text on it

Description automatically generated with medium confidence

**Data Summary and Implications**

The goal of the analysis is to test the hypothesis by using this model to make predictions of the incentive category based on the features chosen. To prove the hypothesis, correct, the area (AUC) under the Receiver Operator Characteristic (ROC) curve. This curve shows the true positive predictions of the model and the false positives to show how well a model is performing. This means that a higher AUC, the better performing the model is. (Bhandari) This is accomplished by plotting the true positive rate and the false positive rate. The AUC score is the value of what area exists under the plotted curves. To interpret this value, the chart below from K2 Analytics will be followed to determine if the model is significant using the fair model limit as the threshold.

A yellow and white rectangular box with black text

Description automatically generated

First, predictions from the model need to be made then they can be stored as a flat list of integer values to be used in the confusion matrix.

A screenshot of a computer program

Description automatically generated

A confusion matrix is created to view the actual vs. the predicted values that the model made so the performance can be determined. This method plots to show the number of correct and incorrect predictions made by the model for each of the possible categorical outcomes.

A screenshot of a computer

Description automatically generated

The prediction probabilities need to be calculated and stored. These values will help create the ROC curve.

A screenshot of a computer

Description automatically generated

Once the probabilities are found, the ROC curve can be created and then plotted.

A screen shot of a graph

Description automatically generated

Finally, the area can be calculated using the auc function.

A screenshot of a computer

Description automatically generated

The AUC score is 0.7214 which means the model falls into the Fair category on the K2 Analytics chart. This means that the model is significant thus supporting the hypothesis.

Considering the result is a significant model, it is recommended that it be used to help contractors or individuals interested in installing PV systems on their properties in New York to determine what amount they could expect to receive from the NYSERDA program.

While this model is significant, it should be noted that this is a lower AUC score than what may be desired for a model to be implemented. The model is limited due to the low number of available features that were used in comparison to the full data set that is available.

This leaves a great option for further analysis to be completed on the same data set but looking at other features available. A different selection or combination may result in a stronger model or could be incorporated into the current model to help improve it. Another method for analyzing a similar use case could be looking at the K-nearest neighbors’ method (KNN). Distances could be calculated between the values used in this analysis through this method to classify if the incentive amount would be considered low or high.

Just using the Project Cost and the Expected KWh Annual Production values, this model can determine if their project could expect to receive $0 to $2,545 (Low) or $2,545 to $200,000 (High).

**Sources**

“An Introduction to the ReLU Activation Function.” Built In, 2022, builtin.com/machine-learning/relu-activation-function#:~:text=The%20rectified%20linear%20unit%20(ReLU)%20or%20rectifier%20activation%20function%20introduces,activation%20functions%20in%20deep%20learning. Accessed 12 Apr. 2024.

Bhandari, Aniruddha. “Guide to AUC ROC Curve in Machine Learning : What Is Specificity?” Analytics Vidhya, 15 June 2020, [www.analyticsvidhya.com/blog/2020/06/auc-roc-curve-machine-learning/](http://www.analyticsvidhya.com/blog/2020/06/auc-roc-curve-machine-learning/). Accessed 1 Apr. 2024.

“Classification Accuracy & AUC ROC Curve | K2 Analytics.” K2 Analytics, 9 Aug. 2020, [www.k2analytics.co.in/classification-accuracy-auc-roc-curve/](http://www.k2analytics.co.in/classification-accuracy-auc-roc-curve/). Accessed 1 Apr. 2024.

GfG. “Activation Functions in Neural Networks.” GeeksforGeeks, GeeksforGeeks, 29 Jan. 2018, [www.geeksforgeeks.org/activation-functions-neural-networks/](http://www.geeksforgeeks.org/activation-functions-neural-networks/). Accessed 12 Apr. 2024.

GfG. “SAS vs R vs Python.” GeeksforGeeks, GeeksforGeeks, 12 June 2023, [www.geeksforgeeks.org/sas-vs-r-vs-python/](http://www.geeksforgeeks.org/sas-vs-r-vs-python/). Accessed 28 Mar. 2024.

GfG. “What Is Adam Optimizer?” GeeksforGeeks, GeeksforGeeks, 22 Oct. 2020, [www.geeksforgeeks.org/adam-optimizer/](http://www.geeksforgeeks.org/adam-optimizer/). Accessed 12 Apr. 2024.

“Home Solar Panel Adoption Continues to Rise in the U.S.” *Pew Research Center*, 14 Oct. 2022, [www.pewresearch.org/short-reads/2022/10/14/home-solar-panel-adoption-continues-to-rise-in-the-u-s/](http://www.pewresearch.org/short-reads/2022/10/14/home-solar-panel-adoption-continues-to-rise-in-the-u-s/). Accessed 26 Mar. 2024.

Roshan, Joseph V. “Optimal Ratio for Data Splitting.” Statistical Analysis and Data Mining, vol. 15, no. 4, 4 Apr. 2022, pp. 531–538, onlinelibrary.wiley.com/doi/full/10.1002/sam.11583#:~:text=A%20commonly%20used%20ratio%20is,optimal%20for%20a%20given%20dataset., <https://doi.org/10.1002/sam.11583>. Accessed 12 Apr. 2024.