WGU Data Analytics Graduate Capstone

Executive Summary for

Neural Network Prediction of NYSERDA PC Incentive Funds

By: Christopher Kamper

April 2024

**Neural Network Prediction of NYSERDA PV Incentive Funds**

**Executive Summary**

In recent years, there has been an increase in interest in individuals and small businesses looking at utilizing photovoltaic (PV) systems to decrease recurring expenses. If a PV system is a viable solution for a customer, the main factor that could determine if the project comes to fruition or not is the initial cost. (“Home Solar Panel Adoption Continues to Rise in the U.S.”) Due to this, government agencies have begun programs to help aid in reducing out-of-pocket costs. The New York State Energy Research and Development Authority (NYSERDA) has a program that provides incentive money for these projects. This analysis is intended to develop a model that will take project cost and estimated energy production values to predict a range of how much incentive money a project can be expected to receive. This has led to a hypothesis of: 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. If the hypothesis is proven, that means the model could be used by contractors or individuals involved in a PV project to help determine what level of funding they could get from the NYSERDA program to determine the actual out-of-pocket costs.

Data is available from NYSERDA through the NY Open Data program. (NY Open Data) The set is updated each month and was pulled for this analysis from the March 2024 update. It contains 47 variables with a total of 162,730 rows and can easily be downloaded from the website as a CSV file. This is great for analysis since it is easy to pull the data and get it into a data frame. There are null values though and extra variables that are included when downloading the set so this will require a few extra steps in the preparation phases. It does contain a unique identifier as the "Project Number" along with the "Project Cost", "Total NYSERDA Incentive", and "Expected KWh Annual Production" variables which will all be used for the model. A categorical variable based on the "Total NYSERDA Incentive" variable will be created so the outcome of the model can tell if the incentive amount expected is "Low" or "High".

These are the steps used to prepare the data for the analysis:

* Variables not needed for the analysis are removed.
* Any duplicated values based on the "Project Number" are removed.
* Any null value is removed.
* Calculation for the new category is completed assigning a 0 for the low category and 1 for the high.
* A new column is added to the data frame for this new categorical variable.

The steps above produce the final data set that is ready for analysis as shown below.

A screenshot of a computer

Description automatically generated

An initial analysis on the prepared data was completed to review for any obvious trends or relationships between the variables. Statistics were calculated for each to determine the mean, min, max, and median values.

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Project Cost** | **Total NYSERDA Incentive** | **Expected KWh Annual Production** |
| **Mean** | 9,495 | 1,689 | 5,894 |
| **Min** | 0 | 0 | 0 |
| **Max** | 26,420,000 | 601,250 | 1,165,000 |
| **Median** | 31,164 | 2,545 | 7,773 |

The graphic below shows the breakdown of how many values were categorized to each category.

A graph with blue squares

Description automatically generated

The plots below were used to review the relationship between each independent variable and the dependent variable. They show some correlation on the higher values but shows overlap as one category covers the plotted values of the other on both.

A graph of orange and blue dots

Description automatically generated

A graph of orange dots

Description automatically generated

While this information will not directly affect the model, it gives good insight about what the spread of the incentive amounts within each category could be as well as showing there is at least some relationship between the independent variables and what the outcome for the dependent is.

With the data set ready, it was then split into training and testing sets. The 80/20 rule was used along with the train test split from sklearn. (Roshan)

The model is an artificial neural network that will use the "Project Cost" and "Expected KWh Annual Production" variables to predict the value of the "Incentive Category". This can be done with the model design and parameters shown below.

**Parameters:**

* Loss function: Binary crossentropy
* Optimizer: Adam
* Epochs: 50
* Hidden Layer Activation Function: Relu
* Output Layer Activation Function: Sigmoid
* Compile Metric: Accuracy

A screenshot of a computer

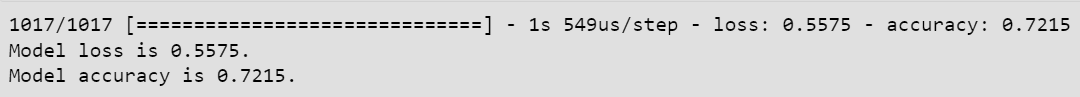
Description automatically generated

Training is completed on the training data set defined with a validation split of 20%. Early stopping criteria was used to aid in reducing overfitting which did stop the process at 8 epochs out of the 50 from the parameters based on the accuracy of the model from a patience of 5.

A screenshot of a computer code

Description automatically generated

The result shows the 3rd model being the best due to having the highest accuracy score at 0.6782. This model was run on the testing set with an outcome of an accuracy score of 0.7215.



The actual metric being used to evaluate the model's significance is the area under the receiver operator characteristic curve (ROC AUC). Predictions were made using the test data set to find the predicted probabilities. The true positive and false positive rates were then plotted to find the resultant area. (Bhandari)

A graph with a line

Description automatically generated



Based on the chart below shared by K2 Analytics, the model is significant and supports the hypothesis.

A yellow and white rectangular box with black text

Description automatically generated

Since the model was found to be significant, it is recommended that it be used. A note of caution should come with the recommendation though as the model's score is on the lower end so it implies there could be some limitations. The most likely is other variables in the original data set may play a stronger role in the incentive money awarded, thus aiding in an improved model.

It would also be recommended that further analysis 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 for 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.

By utilizing this model, contractors and customers looking at PV systems for their homes and businesses could benefit from the information provided by this study since the upfront cost of these systems could determine whether a project will take place. The median value for a project cost reported is $31,164 and the median for the incentive amount is $2,545. This means that 8% of the cost from that project was covered and could have been the difference between the project moving forward if the customer's budget was $30,000. This model gives supports in determining the end cost with guidance if a customer can expect to receive $0 to $2,545 (Low) or $2,545 to $200,000 (High).

**References:**

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.

“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.

NY Open Data. “Solar Electric Programs Reported by NYSERDA: Beginning 2000.” Ny.gov, 27 June 2014, data.ny.gov/Energy-Environment/Solar-Electric-Programs-Reported-by-NYSERDA-Beginn/3x8r-34rs/about\_data. 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.