*Predicting Solar Power Generation*

Archies Desai  
*AIMT*  
 *Lambton College*Toronto, Ontario c0827794@mylambton.com  
Ashwin Krishna Raparthy   
*AIMT*  
 *Lambton College*Toronto, Ontario  
 c0826091@mylambton.caSanthi Priya Yalla   
*AIMT*  
 *Lambton College*Toronto, Ontario  
 c0796614@mylambton.caShivam Bhatt   
*AIMT*  
 *Lambton College*Toronto, Ontario  
 c0827381@mylambton.com

*Abstract* — The world has grown warier of fossil fuels and has seen a large push towards renewable sources of energy. Solar energy is one of the biggest areas for renewable energy. As such many new solar power plants are being built or upgraded. This raises the question of which sites to select and whether it would be beneficial to upgrade an existing solar power plant. This project looks to address these questions by predicting total solar energy generation based on the local environment variables.

# Introduction

The importance and demand for energy from renewable sources have been on the rise. In this area, solar energy has seen the most focus. Governments across the world are leading this change by becoming energy independent through solar power plants, either implemented on rooftops or within city utility sites.

This project aims to help with this move by using local weather variables to predict annual solar energy produced. The hypothesis proposed is that weather parameters like wind speed, temperature, solar radiation, day length, vapour pressure, precipitation, snowfall, and the altitude of the solar power plant impact the daily solar energy generated.

Hence, these variables are used to train a model for the prediction of solar energy production for the power stations present in Calgary. This model can then be used for capacity prediction of potential power plant locations.

# Objectives

For this project, we are planning on using a fully connected ANN. The input for our model would be different environment variables and the output would be generated solar power.

The dataset would be collected from various sources. We are planning to perform the prediction on the City of Calgary. The location and power generation data will come from the official City of Calgary website and the environment data will be taken from the Daymet dataset. The Daymet dataset is a NASA initiative that provides weather data for 1km grids in North America.

# Importing necessary libraries

Importing required Python libraries and methods for data collection and analysis is the intial step of our project. Pandas, Matplotlib, Scikit-learn, and ArcGIS for geographic information are few important libraries that we are importing as part of our project. To fetch the geographical data it’s necessary to connect with ArcGIS.

# Dataset access & visualization

Collecting a dataset for building a model is the primary task in any project. The accuracy and usability of model depends on type of collected dataset.

The following are the key data sources for this sample:

Eleven solar photovoltaic power facilities in the city of Calgary were chosen for the research.

There are two parts to the dataset:

1. Production of daily solar energy for each power plant in the period of September 2015 and December 2019.
2. Weather measurements for the given sites on a daily basis.

The datasets were collected from a variety of sources and then preprocessed to create the main dataset for this project. This dataset was then used to generate two feature layers, one for training and the other for verifying/validating.

## Training Dataset

The training dataset contains data from ten solar locations that will be used to train the algorithm. The data feature layer is accessed through the Arcgis portal and visualised as follows:

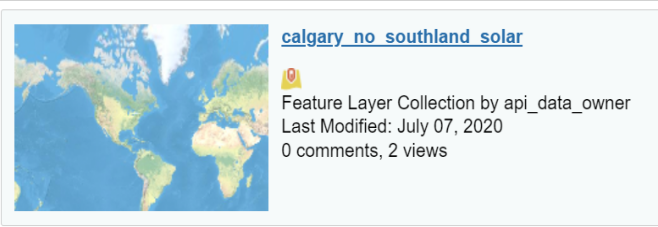


Fig. a) solar dataset feature layer

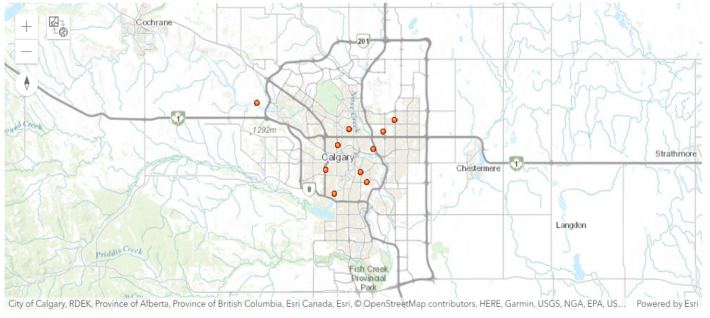


Fig. b) location of 10 solar sites in calgary

The map above depicts the ten power plant locations where the training data was collected.

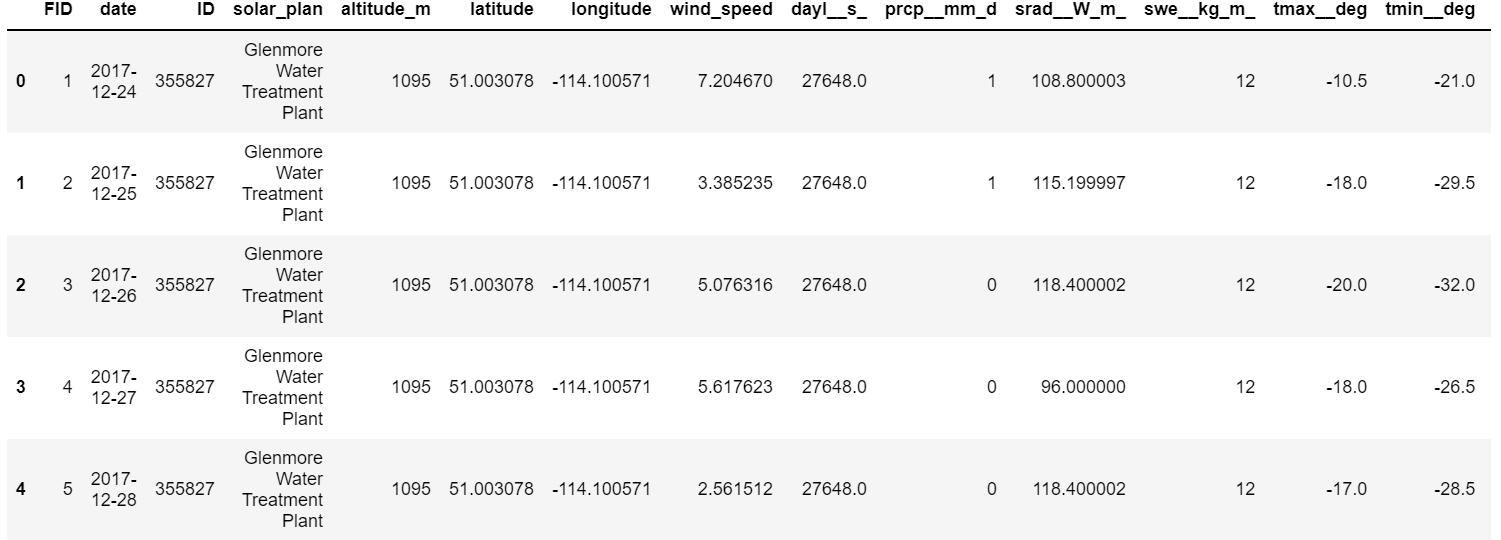


Fig.c) Dataframe visualization

Each row in the table above represents a single day from September 2015 to December 2019, with the associated date in the date field and the name of the solar site in the solar\_plan field.

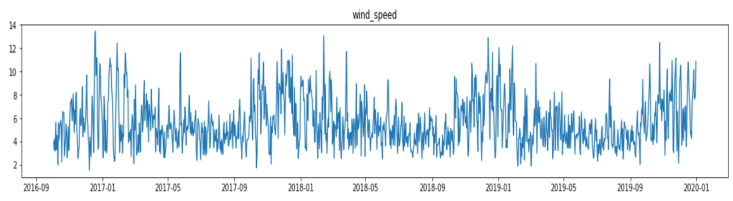
The core data consists of the daily energy generation in kilowatt-hours (KWh) for each of the 10 solar photovoltaic power plants in Calgary, which is stored in the field kWh\_filled. The capacity factor, which is generated after normalising the kWh\_filled by the peak capacity of each solar photovoltaic site, and will be utilised as the dependent variable below, is indicated by the field capacity\_f.

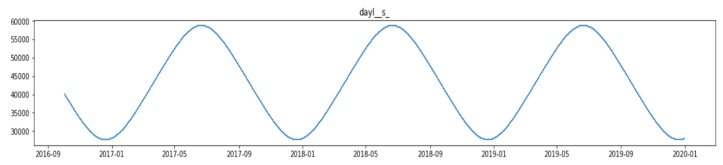
Furthermore, It also contains data on meteorological factors for each day for the connected solar plant, all of which were gathered from MODIS, Daymet observations except wind speed.

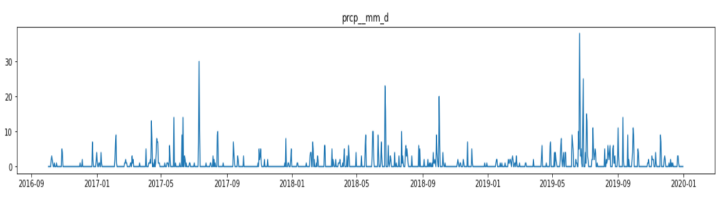
The following are the variables:

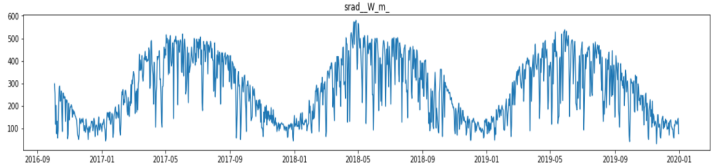
* wind\_speed : wind speed (m/sec)
* dayl\_\_s\_: Length of the day (sec/day)
* prcp\_\_mm\_d: Precipitation (mm/day)
* srad\_\_W\_m : Shortwave radiation (W/m^2)
* swe\_\_kg\_m\_ : Snow water equivalent (kg/m^2)
* tmax\_\_deg : Maximum air temperature (degrees C)
* tmin\_\_deg: Minimum air temperature (degrees C)
* vp\_\_Pa: Water vapor pressure (Pa)

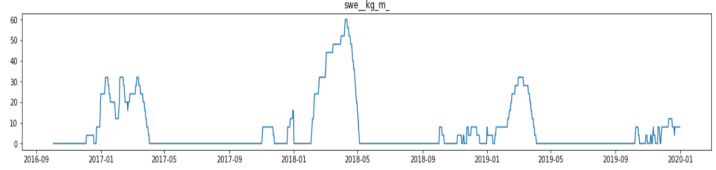
Data from one of the stations is plotted in the following way to understand the circulation of the variables over the last few years and their relationship with the dependent variable of daily energy produced for a station.

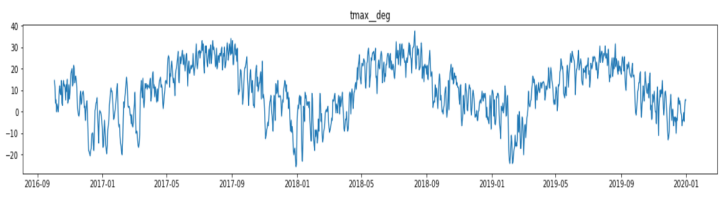


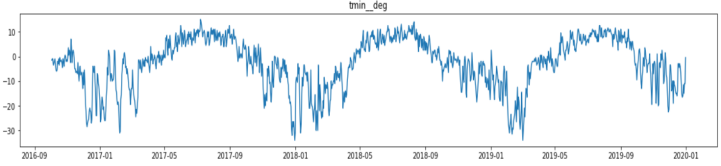


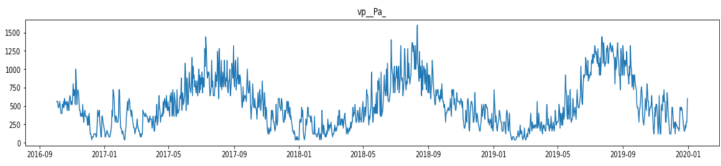


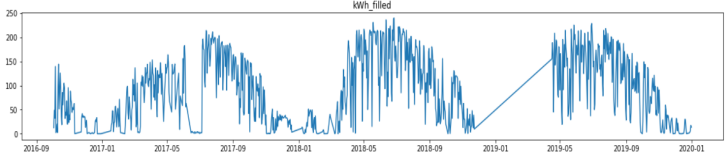












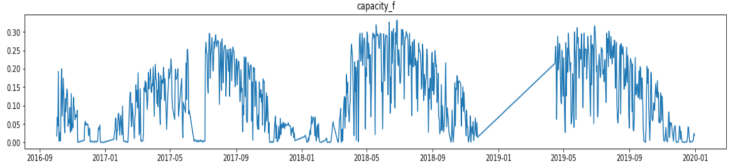


Fig.d) Training set weather variables visualization

From the all above plots, it’s evident that each variable has a high seasonality, and there appears to be a relationship between the dependent variable kWh\_filled and the explanatory factors. As a result, a correlation diagram should be made to see if the variables are related.

Table

Description automatically generated

Fig.e) correlation matrix between the predictors and the dependent variable of capacity\_factor

The resulting correlation diagram demonstrates that the dependent variable of total solar energy produced expressed in terms of capacity factor (capacity\_f) has the strongest correlation with the variable of shortwave radiation per meter square (srad\_\_Wm). In addition, The variable of day length(dayl\_s) is then followed, as longer days are tend to produce more solar energy. The daily temperatures of max(tmax\_\_deg) and min(tmin\_\_deg) are closely followed by the remaining variables with lesser correlation values.

## Validation Dataset

The validation Dataset consists of daily solar generation data for one solar site, known as Southland Leisure Centre, from September 2015 to December 2019. It will be used to validate the trained model.

# Building the Model

The training and validation datasets are ready for modelling after they have been processed and analysed.

In this project two methods are used for sampling. They are:

1. *FullyConnectedNetwork:* In first method, a deep learning framework called FullyConnectedNetwork is used, which is available in the arcgis.learn module of the ArcGIS API for Python.
2. *MLModel:* In the second method, a scikit-learn regression model is implemented in arcgis.learn using the MLModel framework. By passing the name of the algorithm and its relevant parameters as keyword arguments, this framework can deploy any regression or classification model from the library.

In the end, the two methods' performance in terms of model training and validation accuracy will be compared.

## **1 — FullyConnectedNetwork**

This is an Artificial Neural Network model from the arcgis.learn module that is used for modelling in this case.

### Data preparation

First, a list of the feature data that will be used to predict daily solar energy generation is created. It will receive continuous variables by default, and the True value should be passed inside a tuple alongside the variable in the case of categorical variables. All of the variables in this example are continuous.

Once the explanatory variables have been identified, the prepare\_tabulardata method from the arcgis.learn module in the ArcGIS API for Python handles the bulk of the data preprocessing. This function accepts a feature layer or a spatial dataframe containing the dataset as input and returns a TabularDataObject that can be fed into the model.

The tool requires the following input parameters:

* input\_features : feature layer or spatial dataframe containing the primary dataset
* variable\_predict : The name of the field containing the y-variable from the input feature layer/dataframe
* explanatory\_variables : list of field names as 2-sized tuples containing the previously mentioned explanatory variables

### Initializing the Model

The data is ready to be passed to the ANN for training after it has been prepared by the prepare\_tabulardata method. First, the ANN which is also called as FullyConnectedNetwork is imported from arcgis.learn and initialised.

### Search for Learning Rates

lr\_find is used to search for an optimal learning rate before passing it to the final model fitting.

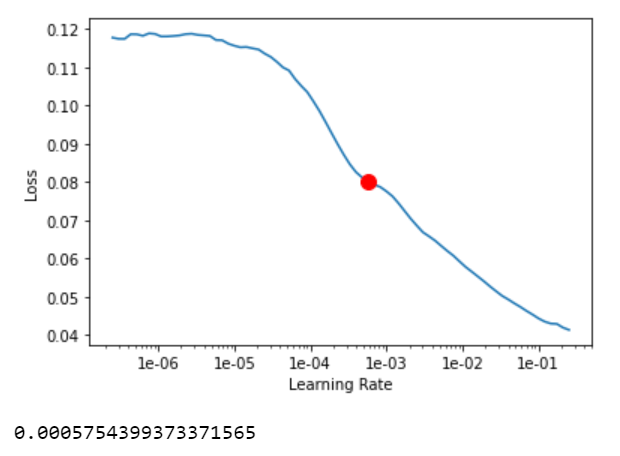


Fig.f) learning rate vs loss

The lr\_find method suggests a learning rate of around 0.0005754 in this case. The automatic lr\_finder will use a conservative estimate of the learning rate, but some experts can interpret the graph more accurately and find a better learning rate to use for final model training.

### Model Training

Finally, the model is complete and ready for training. To train the model, the model.fit function is invoked and given the number of training epochs and the estimated learning rate suggested by lr\_find in the previous step.

To determine whether the model is over-fitting, the train loss and valid loss are plotted. The resulting plot shows that the model was properly trained and that the losses are gradually but not significantly decreasing.

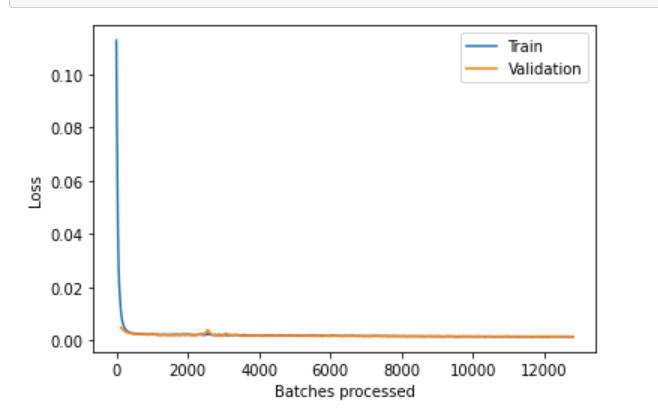


Fig. g) Train loss vs valid loss

Later, the training results are printed to evaluate the trained model's prediction on the test set.

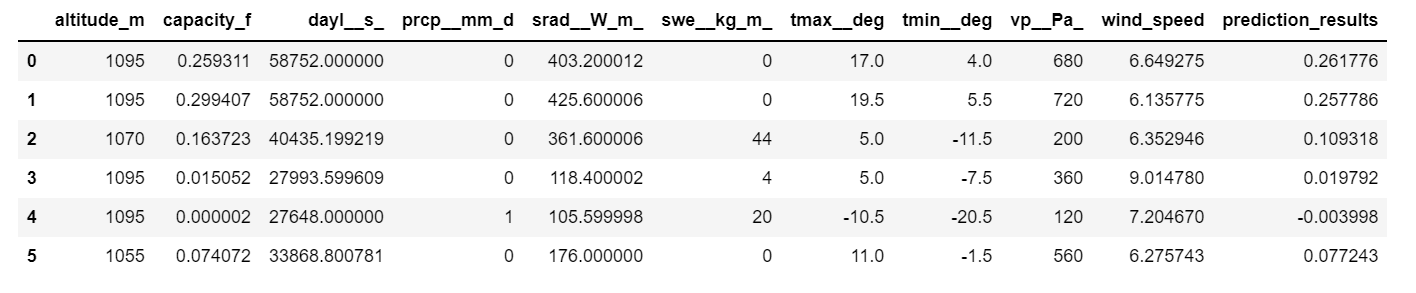


Fig.h) predicted values by trained model

The values predicted by the model when applied to the test set, prediction\_results, in the table above, are similar to the actual values of the test set, capacity\_f.

As a result, the trained model's model metrics can now be estimated using the model.score function, which returns the fitted model's r-squared metric.

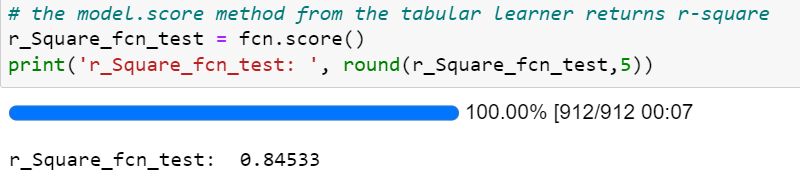


Fig.i) r-squared metric of the fitted model

The model's high r-square value indicates that it has been well trained.

### Forecasting and Validation of Solar Energy Generation

The trained model (FullyConnectedNetwork) will now be used to forecast the daily lifetime solar energy output for the solar plant installed at the Southland Leisure Centre in 2015. The goal is to evaluate the trained model and assess its effectiveness in estimating solar output using just meteorological data from the Southland Leisure Center.

As a result, the model.predict method from arcgis.learn is used with the daily weather variables from September 2015 to December 2019 as input for the mentioned site to predict daily solar energy output in KWh for the same time period. The trained model selects predictors from the input feature layer of southland layer without explicitly identifying them because their names are identical to those used to train the model.

The predicted values for the Southland photovoltaic power plant are saved in the field prediction\_results, which contains the model's estimated daily capacity factor of energy generation, whereas the actual capacity factor is stored in the field capacity\_f.

The capacity factor is a normalized number that will be rescaled back to the original unit of KWh by utilising the Southland solar power plant's peak capacity of 153KWp.

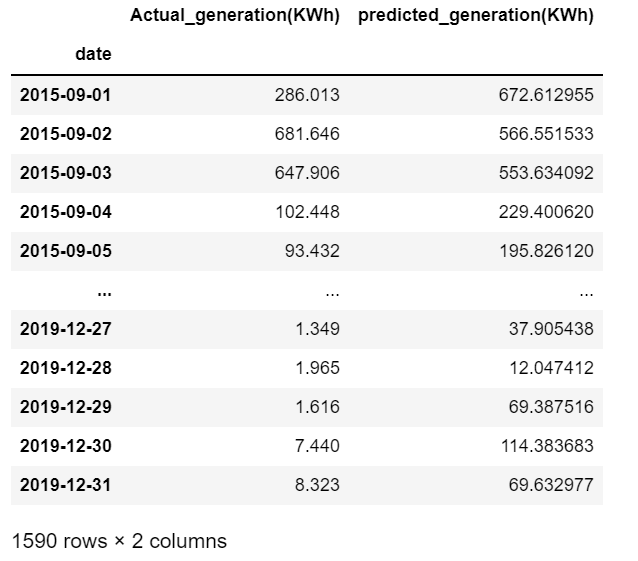


Fig.j) actual vs model predicted daily solar energy geneartion

The table above compares the actual versus expected daily solar energy output for the Southland facility from late 2015 to the end of 2019. These values are now utilized to estimate the various model metrics in order to assess the model's predictive power.

The comparison yields an r-square of 0.86, indicating a high degree of similarity between the actual and anticipated values.

To summarise, the actual average annual energy created by the solar plant is 170.03 MWh, which is close to the anticipated annual average generated energy of 170.34 MWh, demonstrating a good level of accuracy.

### Visualization of Results

Eventually, the actual and expected values are shown to show their distribution across the power plant's full lifetime.

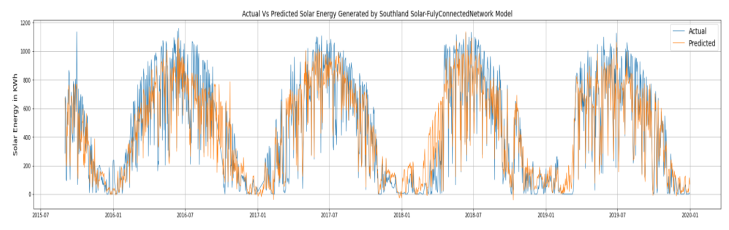


Fig.k) actual and predicted solar energy generation values

The blue line in the graphic above indicates actual generation levels, while the orange line represents expected generation values. The two have a high degree of overlap, showing that the model is quite predictive.

## **Gradiant Boost Regressor**

The second technique uses the MLModel framework from arcgis.learn to model the same data with a machine learning model. This framework may be used to import and apply any machine learning model from the scikit-learn library on data supplied by the arcgis.learn prepare tabulardata function.

### Data Preprocessing

As with the neural network data preparation method, a list of feature data that will be used to estimate daily solar energy output is created initially. It will receive continuous variables by default, however for categorical variables, the True value needs be given within a tuple with the variables. The RobustScaler function from scikit-learn is then used to modify these variables by sending it, along with the variable list, into the column transformer function.

Once the list of explanatory variables has been created and the preprocessors have been computed, they are used as input for the prepare\_tabulardata function in arcgis.learn. The function accepts a feature layer or a spatial dataframe that contains the dataset and returns a TabularDataObject that can be fed into the model.The tool's input parameters are identical to the ones discussed previously.

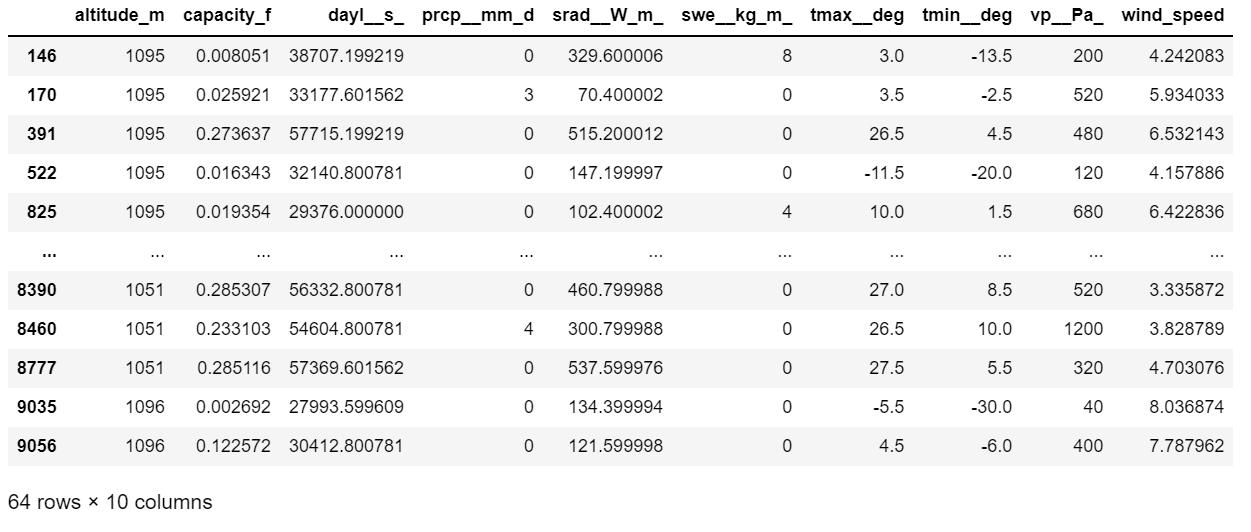


Fig. l) Training data

### Initializing the Model

Once the data has been prepared using the prepare\_tabulardatamethod, it is ready to be fed into the chosen machine learning model for training. The GradientBoostingRegressor model from scikit-learn is used here, and its parameters are provided into the MLModelfunction.

### Training the Model

Finally, The model is now ready for training, and the model.fit method is used to fit the machine learning model with the previously set parameters.

The training results are presented in order to compute several model metrics and evaluate the trained model's quality.

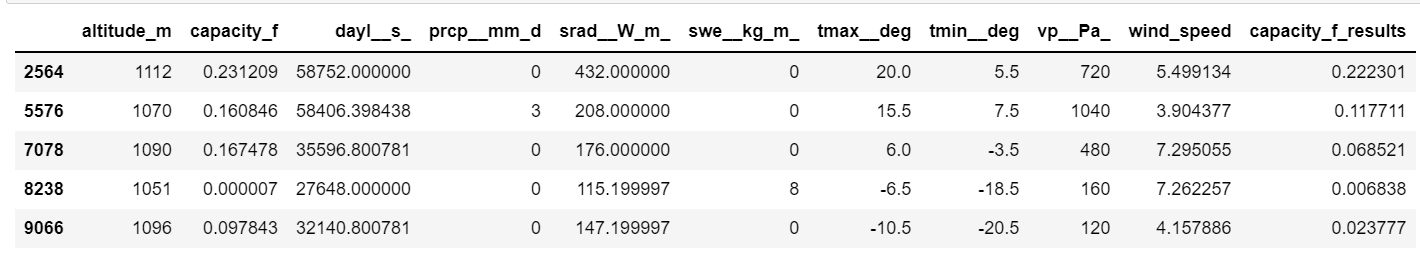


Fig.m) Training results

The capacity\_f\_results column in the table above returns the model's projected values, which are comparable to the actual values in the target variable column, capacity\_f.

Following that, the trained model's model metrics are estimated using the model.score() method now returns the model fit's r-squared.

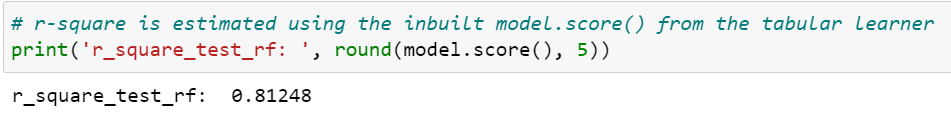


Fig.n) R-squred value of model

The model's high R-squared value shows that it has been successfully trained.

### Explaining the Importance of Solar Energy Generation as a Predictor

It would be fascinating to understand the model's explanability, or the elements that are responsible for forecasting solar energy output from the many variables utilised in the model, once it has been fitted. For this, the feature\_importances method is utilized.

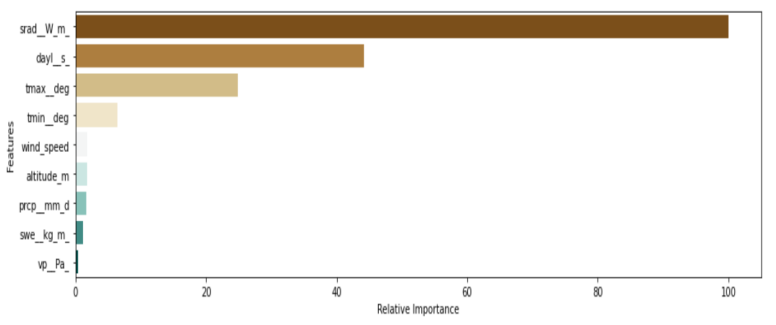


Fig.o) Feature importances returned by feature\_importances method

The feature\_importances method of the trained Gradient Boosting regressor algorithm implemented via the MLModel framework returns the feature importances seen in the graph above. The quantity of solar radiation received by the site in watts per square metre (srad\_Wm) is demonstrated to be the most important predictor for estimating solar energy generation at the location. This is followed by the duration of the day in seconds, the maximum temperature, and the minimum temperature, all in the same sequence as the correlation diagram.

### Forecasting and Validation of Solar Energy Generation

Since its installation in 2015, the trained GradientBoostingRegressor model, implemented via the MLModel, has been utilized to estimate the daily lifespan solar energy output for the solar plant situated at the Southland Leisure Centre. The goal is to evaluate and validate its results against those of the FullyConnectedNetwork model produced previously in this course.

To summarize, the model.predict method from arcgis.learn is used with weather variables on a daily basis as input for the mentioned site from September 2015 to December 2019, to estimate daily solar energy output in KWh for that particular time period. The trained model chooses the predictors from the input feature layer of southland\_layer without saying them explicitly because their names are the same as those used to train the model.

The anticipated values for the Southland plant are placed in the field prediction while the actual capacity factor is stored in the field capacity\_f.

The capacity factor is a normalized number that will be rescaled to the original unit of KWh using the Southland solar power plant's peak capacity of 153KWp.

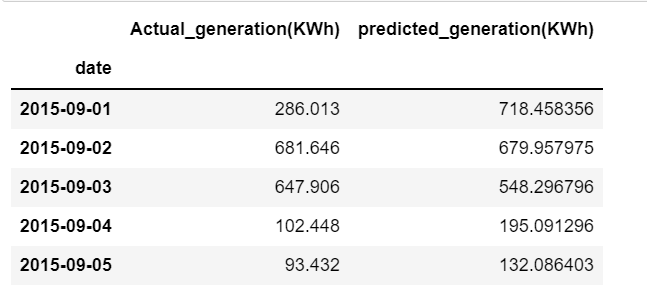


Fig.p) actual versus the MLModel predicted daily solar energy generation

The table above compares the actual daily solar energy generated for the Southland plant to the MLModel expected daily solar energy generated for the period late 2015 to the end of 2019. These data are now utilized to calculate the different model metrics in order to determine the MLModel's prediction power.

The comparison yields a high R-squared of 0.84, suggesting that the real and projected values are quite comparable.

The actual average yearly energy created by the solar plant is 170.03 MWh, which is close to the expected annual average generated energy of 171.48 MWh when the data are added together. This denotes a high level of accuracy.

### Visualization of the Results

Eventually, the actual and expected values are shown to show their distribution across the power plant's lifetime.

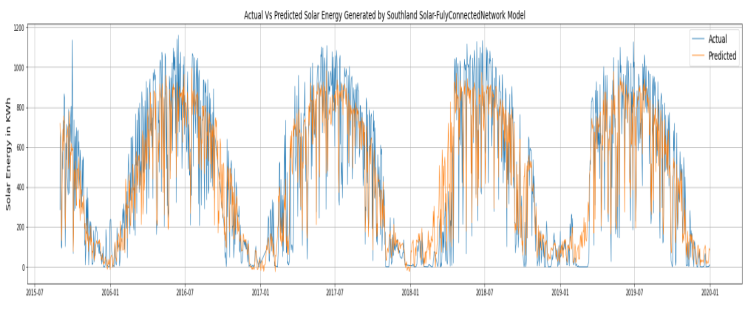


Fig.q) Actual and predicted solar energy generation

# conclusion

The purpose of this project was to develop a model that could forecast daily solar energy efficiency and, as a result, estimate the actual production of a photovoltaic solar energy plant at a given location based on daily weather factors. It did so by demonstrating the newly developed artificial neural network, FullyConnectedNetwork, as well as machine learning models, MLModel, both of which are accessible in the arcgis.learn module of the ArcGIS API for Python.

As a result, data from ten solar energy installation sites in the Canadian city of Calgary was utilised to train two separate models: the FullyConnectedNetwork model and the MLModel framework from the arcgis.learn module. These were later used to forecast the daily solar output of a separate solar facility in Calgary, which was not included in the training set. The notebook outlines the procedures for implementingthese models, which include data preparation, model training, and final inference*.*

The FullyConnectedNetwork and the MLModel algorithms both effectively forecasted the solar energy production of the test solar plant, with projected values of 170.34 MWh and 171.48 MWh, respectively, compared to the actual value of average yearly solar generation of 170.03 MWh for the station*.*

Finally, in order to improve this model in the future, it would be interesting to apply it to additional solar production facilities across other geographies and track their performance in order to understand the model's generalizability.

##### References

1. M. M. Thornton, R. Shrestha, Y. Wei, P. E. Thornton, S. Kao, and B. E. Wilson, “Daymet: Daily Surface Weather Data on a 1-km Grid for North America, Version 4.” ORNL Distributed Active Archive Center, 2020. doi: 10.3334/ORNLDAAC/1840.
2. Data.calgary.ca, 2022. [Online]. Available: https://data.calgary.ca/Environment/Solar-EnergyProduction/ytdn-2qsp. [Accessed: 23- Mar- 2022].
3. Data.calgary.ca, 2022. [Online]. Available: https://data.calgary.ca/dataset/City-of-Calgary-SolarPhotovoltaic-Sites/vrdj-ycb5. [Accessed: 23- Mar- 2022]