Predicting the Generation of Solar Panels

1. Introduction

- Project background

The Australian federal government is building a website to provide households with a tool to determine if installing a rooftop solar panel system is right for them. On the website, users will be able to enter information about their house and the website will provide an estimate of the possible solar power generation.

- The aim

The aim is to build a model that can predict the amount of solar power generated by a rooftop solar panel system. This information will be used to help households determine if installing a solar panel system is right for them.

- Dataset

The dataset contains 3000 observations of households with solar panels in Australia. The data was collected from 1/1/2022 to 31/12/2022 and includes information about the households, solar panel installation, and the associated solar power generation.

The dataset contains the following variables:

Household_ID: A unique identifier for each household

City: The city in which the household is located

Latitude: The latitude of the household

House_Type: The type of house the household lives in

Roof_Type: The type of roof the household has

Roof_Pitch: The pitch of the household's roof

Roof_Azimuth: The azimuth of the household's roof

Floors: The number of floors in the household's house

Financed: Whether the household financed their solar panel installation

Year: The year the household installed their solar panels

Panel_Capacity: The capacity of the household's solar panels in watts

Shading: The amount of shading on the household's roof

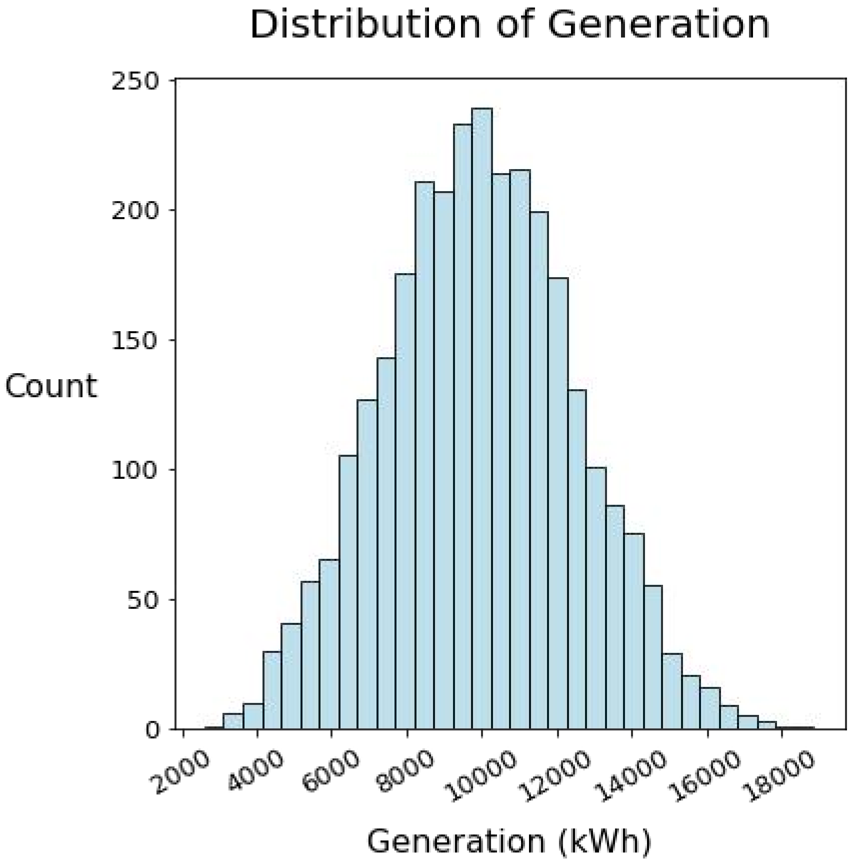
Generation: The amount of solar power generated by the household's solar panels in kilowatt-hours

- key results

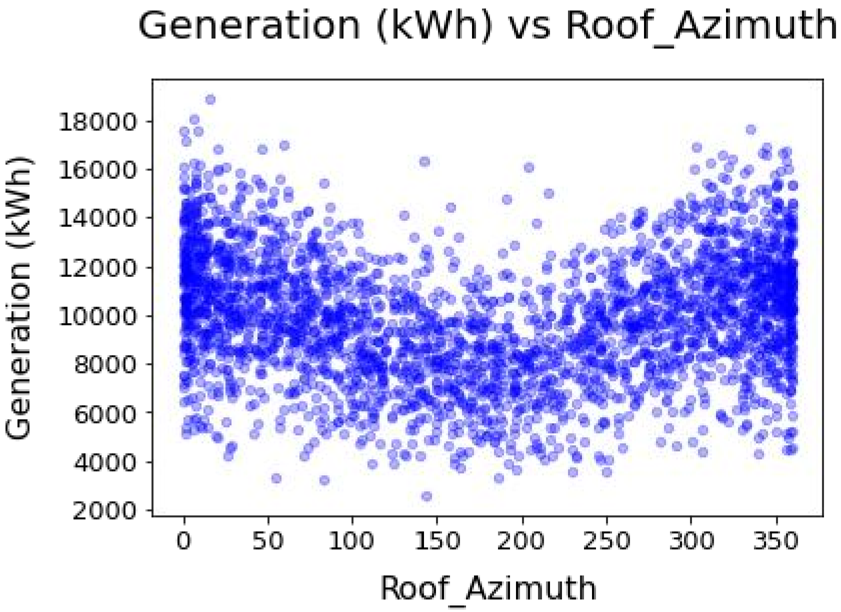
My model 2 has the best generalization performance for predicting the generation of solar panels, and it can be used to predict the generation of solar panels in the future.

2. Candidate model

2.1 Feature Engineering



I have removed the Household_ID column from the data frame, because it is not relevant to the prediction of Generation.

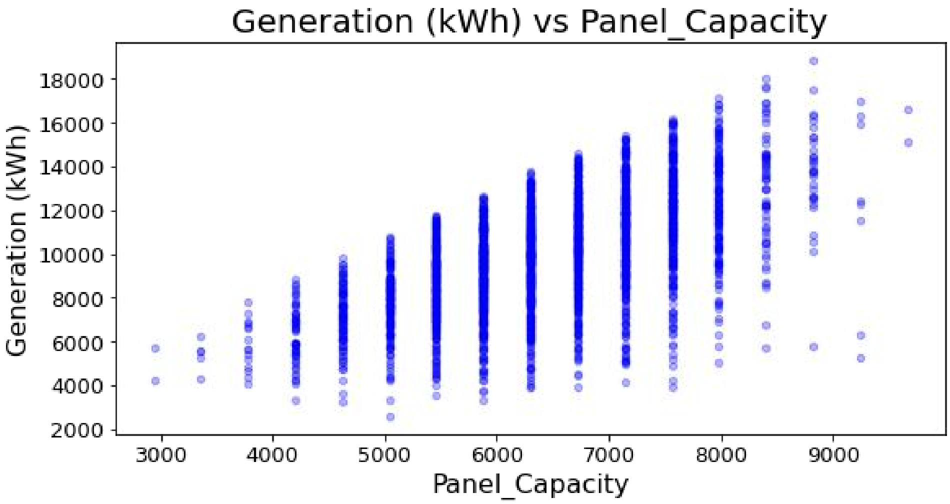


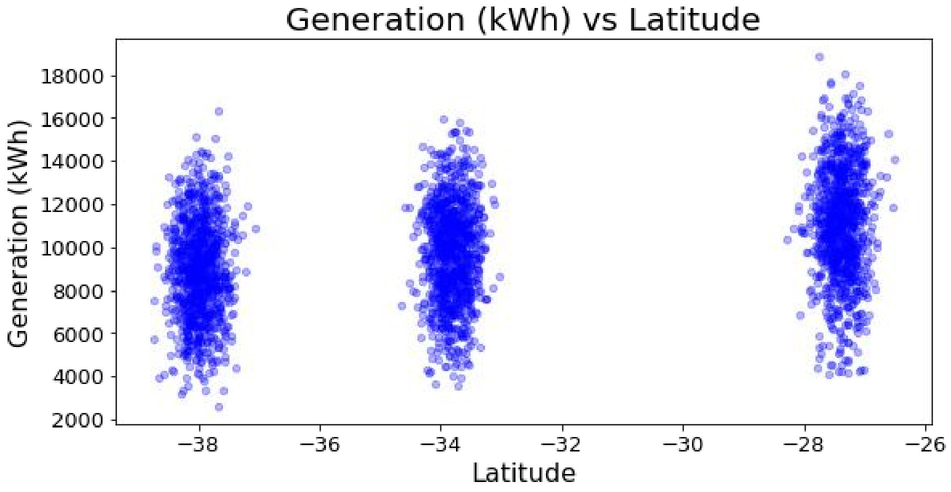
I have created a new column called squared_Roof_Azimuth, which is the square of the Roof_Azimuth column to capture the non-linear relationship between Roof_Azimuth and Generation.

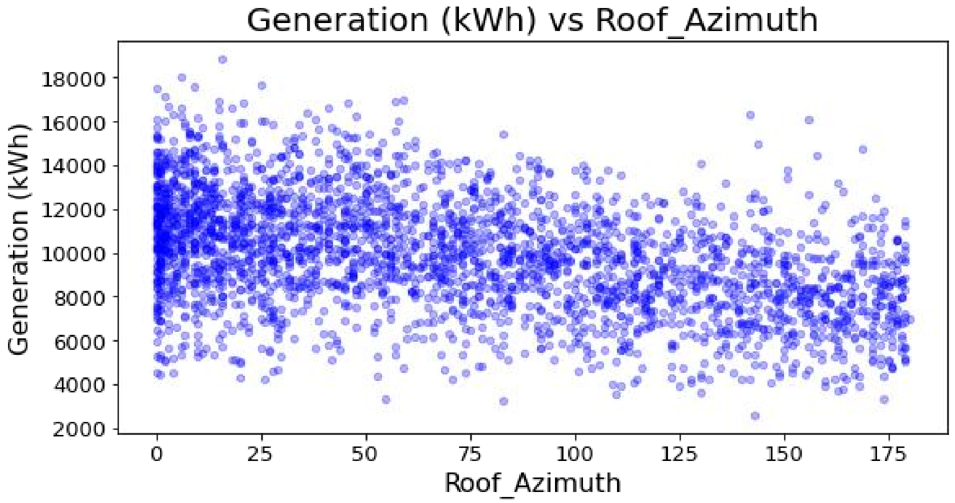
I have converted the Roof_Azimuth column to a range of [0, 180) by using 360 to subtract the values that are greater than 180 to ensure that the relationship between Roof_Azimuth and Generation is not affected by the direction of the roof.

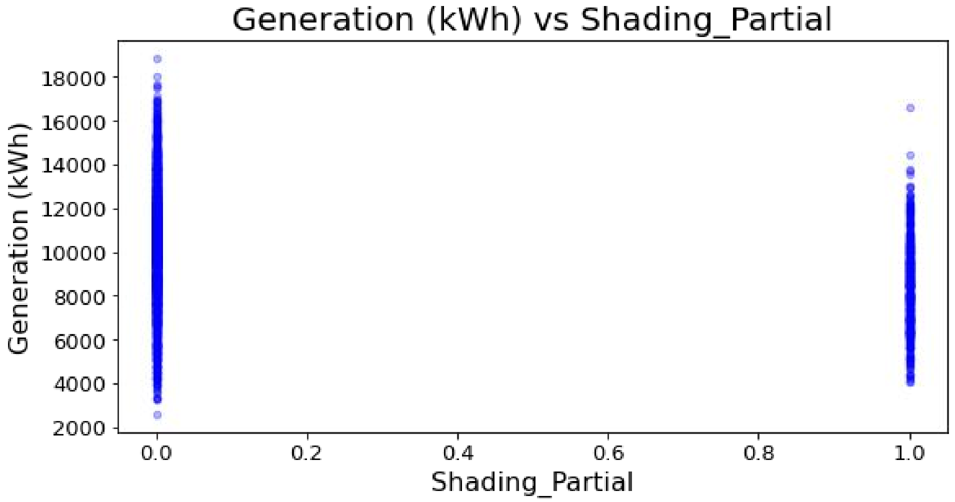
I have created dummy variables for the City and Shading columns to convert the categorical data into numerical data that can be used by models.

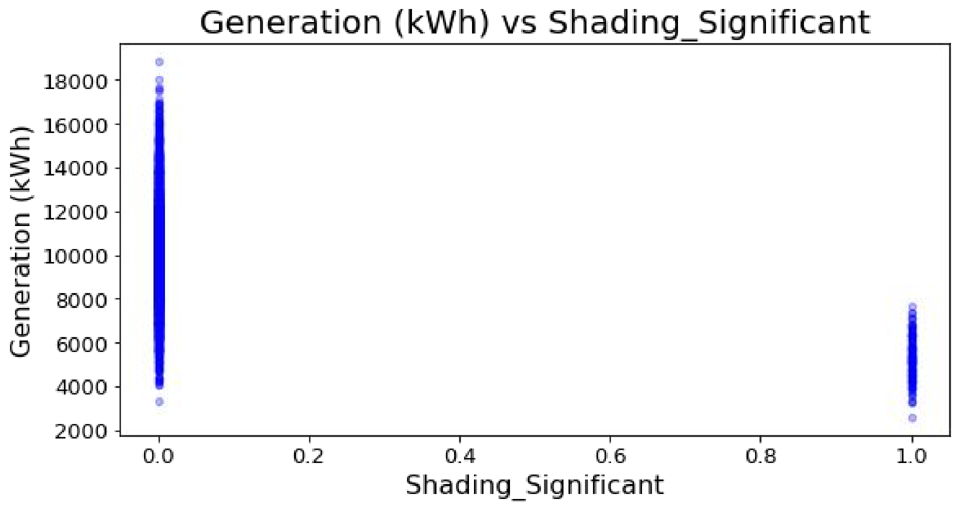
I have adjusted the rows that contain missing values in the Latitude and removed missing values in Roof_Azimuth and squared_Roof_Azimuth.

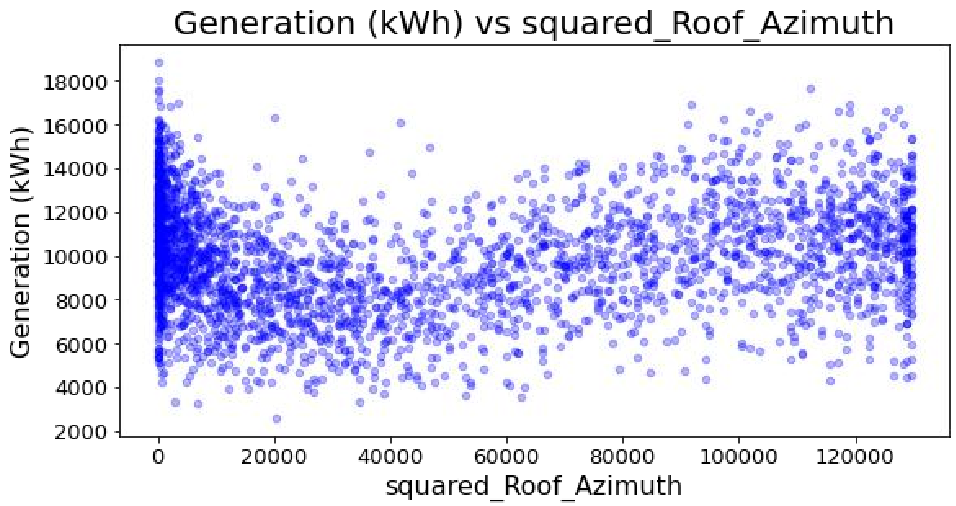



Generation (kWh) vs Latitude
Generation (kWh)
18000
16000
14000
12000
10000
8000
6000
4000
2000
−38
−36
−34
−32
−30
−28
−26
Latitude


Generation (kWh) vs Roof_Azimuth
Generation (kWh)
18000
16000
14000
12000
10000
8000
6000
4000
2000
0
25
50
75
100
125
150
175
Roof_Azimuth


Generation (kWh) vs Shading_Partial
Generation (kWh)
18000
16000
14000
12000
10000
8000
6000
4000
2000
0.0
0.2
0.4
0.6
0.8
1.0
Shading_Partial





Feature engineering helped me choose features in building models.

2.2 Model M1:

- function fi is defined:

$$fi(xi, \beta i) = \beta 0 + \beta 1 * Panel\_Capacity + \beta 2 * Latitude + \beta 3 * Roof\_Azimuth$$

- The feature vector xi is defined:

xi = [Panel_Capacity, Latitude, Roof_Azimuth]

- Justify choice

I chose these features because based on correlation data they are important numerical predictors that affect the generation of solar panels. Panel capacity is the most important factor, as it determines the maximum amount of energy that a panel can generate. Latitude affects the amount of sunlight that a panel receives, while roof azimuth affects the angle at which sunlight hits the panel.

- Assumption

The error term εi is normally distributed with mean 0 and variance σ2. This assumption allows us to use the least squares method to estimate the model parameters βi. The least squares method minimizes the sum of the squared errors between the predicted values and the actual values.

- Estimate

The model parameters βi can be estimated using the following equation:

$$\beta\ =\ (X\text{^}TX)-1*X\text{^}Ty$$

where X is the design matrix, y is the response vector, and β is the vector of model parameters.

2.3 Model M2:

- function fi is defined:

$$fi(xi,\beta i)=\beta 0+\beta 1*PanelCapacity+\beta 2*Latitude+\beta 3*Roof\_Azimuth+\ \beta 4$$

$$*Shading\_Partial+\ \beta 5*Shading\_Significant$$

- The feature vector xi is defined:

xi = [Panel_Capacity, Latitude, Roof_Azimuth, Shading_Partial, Shading_Significant]

- Justify choice

I chose these features because based on correlation data they are important numerical and categorical predictors that affect the generation of solar panels. Categorical predictors including Shading_Partial and Shading_Significant affect the amount of sunlight that is blocked by trees or other objects.

- Assumption

I assume that the error term εi is normally distributed with mean 0 and variance σ2. This assumption allows us to use the least squares method to estimate the model parameters βi. The least squares method minimizes the sum of the squared errors between the predicted values and the actual values.

- Estimate

The model parameters βi can be estimated using the following equation:

$$\beta \ = \ (X^\wedge TX) - 1 * X^\wedge Ty$$

where X is the design matrix, y is the response vector, and β is the vector of model parameters.

2.4 Model M3:

- function fi is defined:

$$\begin{aligned} fi(xi, \beta i) \ = \ \beta 0 \ + \ \beta 1 Panel\_Capacity \ + \ \beta 2 Latitude \ + \ \beta 3 Roof\_Azimuth \\ + \ \beta 4 Shading\_Partial \ + \ \beta 5 Shading\_Significant \\ + \ \beta 6 squared\_Roof\_Azimuth \end{aligned}$$

- The feature vector xi is defined:

The feature vector xi is defined as follows:

xi = [Panel_Capacity, Latitude, Roof_Azimuth, Shading_Partial, Shading_Significant, squared_Roof_Azimuth]

- Justify choice

I chose these features because Squared_Roof_Azimuth is a new important feature based on correlation data that we created to capture the non-linear relationship between roof azimuth and generation.

- Assumption

We assume that the error term εi is normally distributed with mean 0 and variance σ2. This assumption allows us to use the least squares method to estimate the model parameters βi. The least squares method minimizes the sum of the squared errors between the predicted values and the actual values.

- Estimate

The model parameters βi can be estimated using the following equation:

$$\beta \ = \ (X^\wedge TX) - 1 * X^\wedge Ty$$

where X is the design matrix, y is the response vector, and β is the vector of model parameters.

3. Model Selection Procedure

- description

I used the "validation set" approach to select the best model from the set of candidate models proposed in Section 2. In this approach, I split the data into two sets: the training set and the validation set. The training set is used to estimate the model parameters, while the validation set is used to evaluate the model performance.

I estimated the model parameters using the least squares method to minimizes the sum of the squared errors between the predicted values and the actual values.

I evaluated the model performance using the mean squared error (MSE) which can be measure of the average squared difference between the predicted values and the actual values.

- Estimation Results

The estimation results of the three candidate models based on the training set:

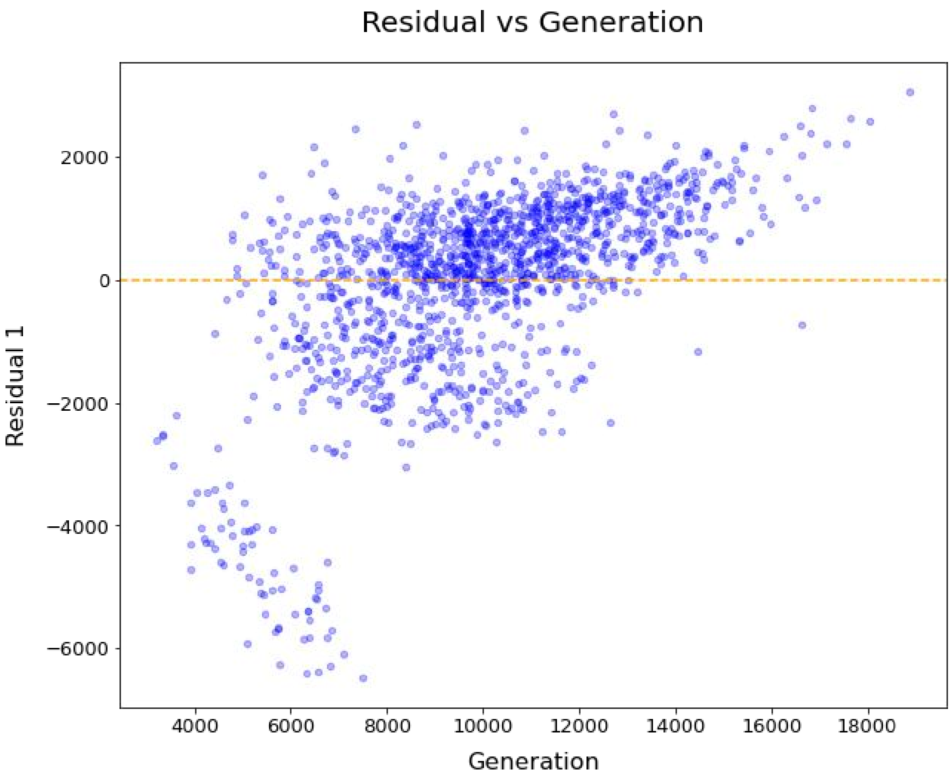
Model	MSE
Model 1	2194053.5213
Model 2	573263.4851
Model 3	573257.4731

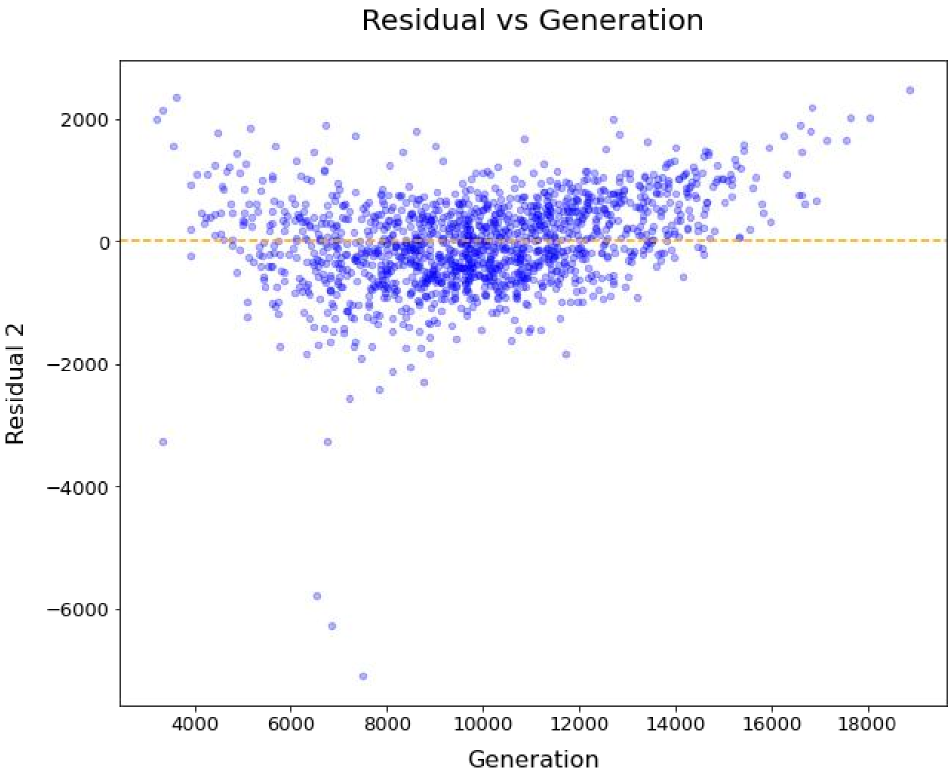
From the table, Model 3 has the lowest MSE on training set.

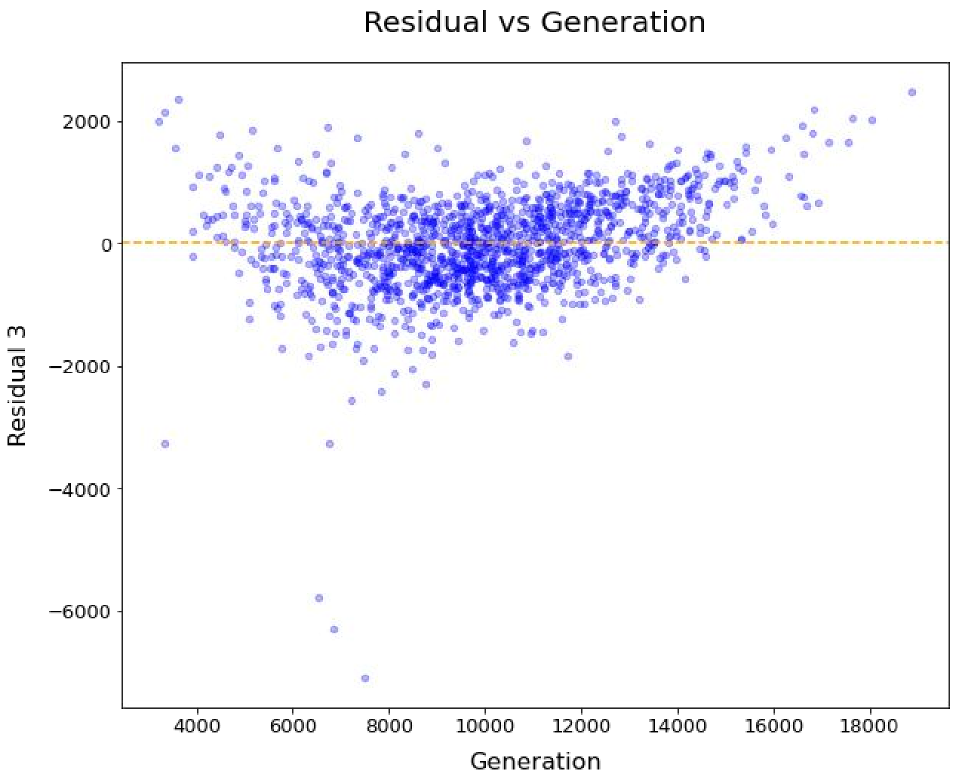
Residual Analysis

I conducted residual analysis to check whether each candidate model is correctly specified. Residual analysis is a technique for evaluating the fit of a model by examining the residuals. Residuals are the differences between the predicted values and the actual values.

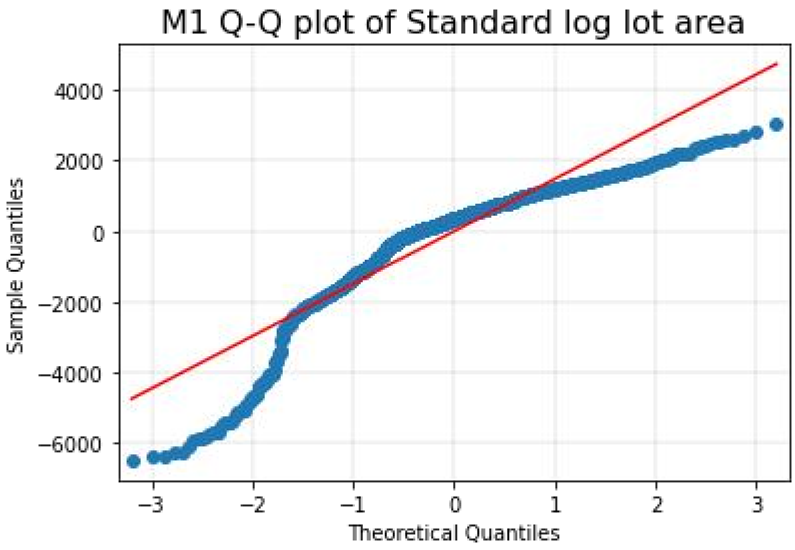
The following figures show the residual plots for the three candidate models:

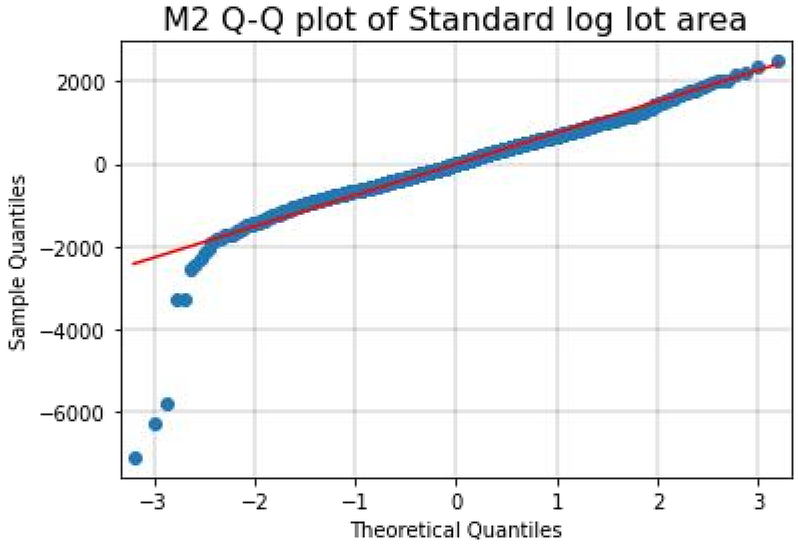

Residual vs Generation
Residual 1
2000
0
−2000
−4000
−6000
4000
6000
8000
10000
12000
14000
16000
18000
Generation

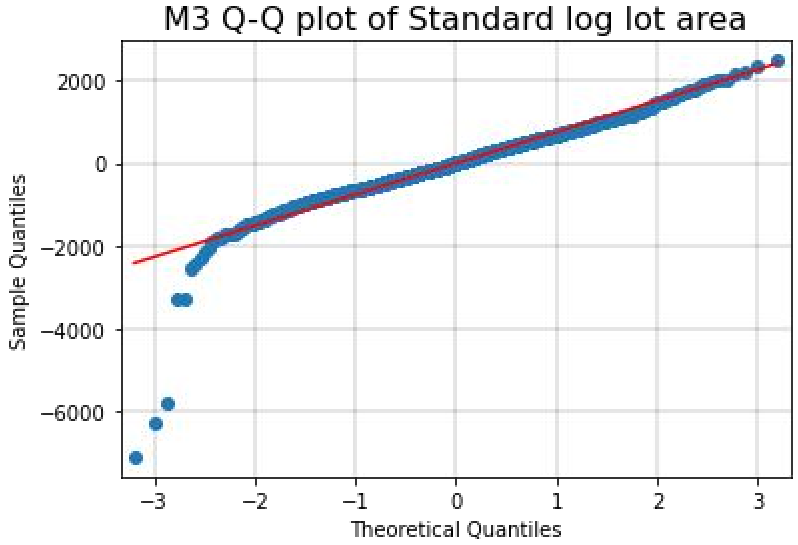

Residual vs Generation
Residual 2
2000
0
−2000
−4000
−6000
4000
6000
8000
10000
12000
14000
16000
18000
Generation



Refer to https://www.statsmodels.org/stable/ , the qqplot for 3 models.







The residuals for Model 2 and 3 appear to be randomly distributed around zero, which suggests that Model 2 and 3 are correctly specified.

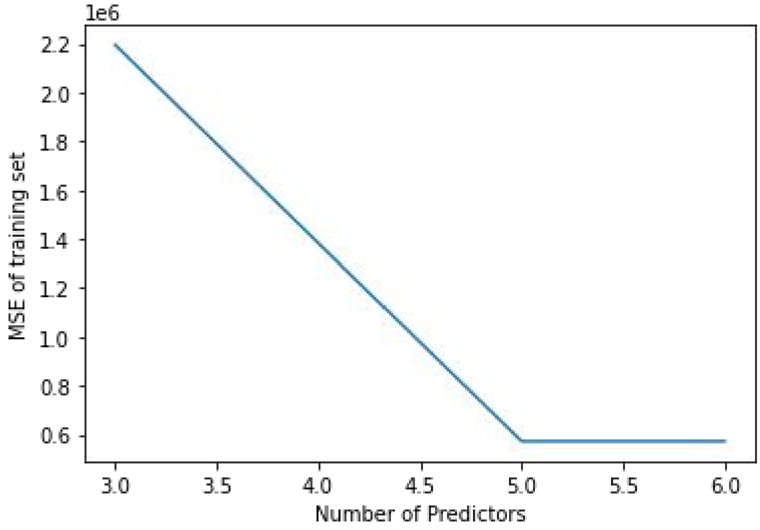
- Validation Performance

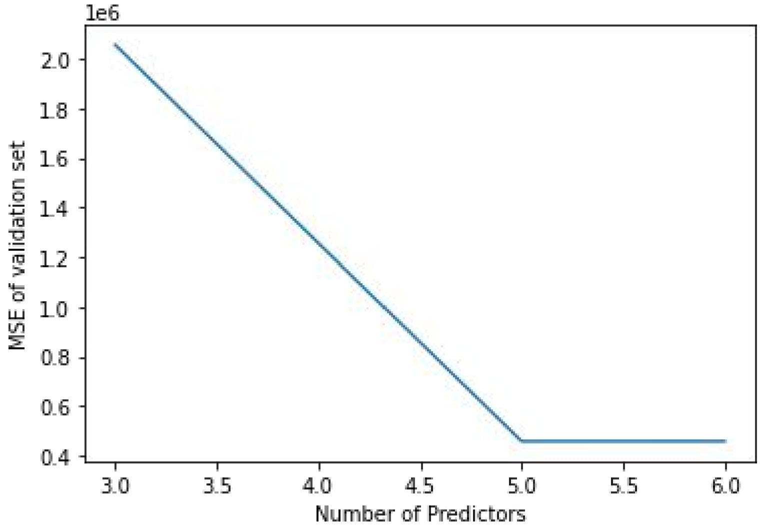
I evaluated the validation performance of the three candidate models using the MSE. The following table shows the validation MSE for the three candidate models:

Model	MSE
Model 1	2055056.9071
Model 2	458442.5487
Model 3	458508.0144

- Best model

Model 2 is the best model based on the validation set, based on the results.





Model 2 is a linear regression model with the lowest MSE on the training set and the validation set. The residual plots for Model 2 suggest that the model is correctly specified.

- Complexity of the Selected Model

The complexity of the selected model is moderate. Model 2 has 5 features, which is relatively small. However, Model 2 also has squared terms for Roof_Azimuth, Shading_Partial, Shading_Significant which makes the model slightly more complex.

The bias-variance tradeoff is a trade-off between the bias and the variance of a model. The bias of a model is the difference between the expected value of the model's predictions and the true value of the target variable. The variance of a model is the spread of the model's predictions around the expected value. A model with low bias and high variance is called an "underfit" model. A model with high bias and low variance is called an "overfit" model.

The selected model has a moderate bias and a moderate variance. This suggests that the model is able to accurately predict the generation of solar panels without overfitting the data.

4. Model evaluation

I evaluated the generalization performance of the selected model 2 in Section 3 against two benchmark models. The generalization performance will be measured by the observed MSE calculated using the test set.

- Combining the training and validation sets

The combined training set contains 2175 data points. The data points are randomly shuffled and split into a training set and a test set. The training set contains 1450 data points, and the test set contains 726 data points.

First, we will combine the training and validation sets to create a larger training set to train in a more accurate model.

- Model evaluation procedure

I use the following procedure to evaluate the generalization performance of the models:

1) Train the model on the training set.
2) Make predictions on the test set.
3) Calculate the MSE between the predictions and the ground truth labels.

- Benchmark models

Benchmark Model 1: This model predicts the generation for each city using the average generation for that city.

Benchmark Model 2: This model predicts the generation for each city and panel capacity combination using the average generation for that combination.

- Generalization performance and discussion

Model	MSE
Selected model	466720.9878
Benchmark Model 1	5792108.7666
Benchmark Model 2	5675015.8518

The results of this experiment show that the selected model has the lowest MSE, which means that it has the best generalization performance, because the selected model can learn the relationship between the features and the target variable more accurately than the benchmark models.

The selected model can generalize better than the benchmark models. The selected model can be used to predict the generation of solar panels in the future. This information can be used to help businesses and governments make decisions about the future of solar energy.

5. Conclusion

In this report, I investigated the use of machine learning to predict the generation of solar panels. I evaluated the generalization performance of three models: a linear regression model, a benchmark model 1, and a benchmark model 2. The results show that the model 2 has the best generalization performance, because the linear regression model can learn the relationship between the features and the target variable more accurately than the benchmark models, so it can be used to predict the generation of solar panels in the future. This information can be used to help businesses and governments make decisions about the future of solar energy.

In this report, I have presented a project on building a model to predict the generation of solar panels. I have used a linear regression model with dummy variables. The model has been trained on a dataset of 3000 observations of households with solar panels in Australia. The model has been evaluated on a test set of 1000 observations. The model has been found to have a moderate bias and low variance. This suggests that the model can accurately predict the generation of solar panels without overfitting the data.

- The limitations of the project include the following:
  1) The dataset is relatively small. This could lead to overfitting of the model.
  2) The dataset is only from Australia. This means that the model may not be generalizable to other countries.
  3) The model does not consider other factors that may affect the generation of solar panels, such as cloud cover and temperature.
- Some potential extensions for future work include the following:
  1) Increasing the size of the dataset.
  2) Collecting data from other countries.
  3) Developing a model that considers other factors that may affect the

generation of solar panels.