­­­­­­­General to Particular or Particular to General, Best Approach to Train a Model

# **Problem**

## **Main Objective:**

Developing a machine learning algorithm, “training data” with known results of ROI will be needed. After the training of algorithm, it will generate ROI for any location with data available.

## **Target Output:**

ROI of a certain area by implementing renewable energy.

## **Input Data:**

Solar, Wind, Water data sets with relevant information for renewable energy generation.

## **Decision making:**

Decide which data will be better to use as the training data for the algorithm. Should we use Australian data to train and generalize to global scale? Or use global data to train the model and narrow down to a particular area, like Australian.

# **Collect data**

Collecting the data of different types of energy is mainly through the energy department of government and online sources which including the database site need to pay. From the past experience of data sourcing, we noticed that there are plenty of government open source data for Australian, which is easier to find as compared with the global scale while the amount of free global data is less, so hard to find.

We would encounter the possibility of incomplete database and extra cost for accessing the database separately for different countries due to the lack of complete global scale data. As most of the machine learning algorithms often fail to deliver expected results because of lack of (suitable) data, lack of access to the data, privacy problems, Australian-based data is more suitable to use.

# Preparing the data

## **Reliability of value**

From the data we collected so far, the solar data in a global-scale is from 2015, while the wind data of Australia region is 2018. In order to implement an efficient and good performed algorithm, the data we used to train must be reliable.

Therefore, we prefer more recently data with good quality will be used to train the algorithm. The higher the quality and the more representative, the easier it will be for the model to learn the unknown and underlying “true” mapping that exists from inputs to outputs. Results will be different when using past data instead of current data. Since past data contains less reliable values which have changed in the recent years, this may influence the accuracy of the trained model, leading to an incorrect prediction of ROI for other location.

## **Underfitting**

While the Australian data has more parameters, and the global scale data provides only general information.

Data with more features will also avoid underfitting, which usually occurs when a model cannot adequately capture the underlying structure of the data. An under-fitted model is a model where some parameters that would appear are missing. Like some specific parameters needed for calculating ROI may not include in the global scale data. Such a model will tend to have poor predictive performance.

# **Evaluation Method**:

Since we are predicting the ROI of a certain area, it considered as a regression problem, which can use certain evaluation metrics such as mean squared error (MSE) to measure the performance.

As we obtained the data of potential energy storage in terms of pumped hydro from a research conducted by ANU, we can evaluate the performance and turn the hyperparameters of the model to seek more accurate prediction. While on the global scale data, there is no estimated calculation for any type of renewable energy, therefore, it’s hard to evaluate the model we built for further improvement.

Additionally, the Australian-based data is large enough to be split into different subsets for training and testing processes, even for the validation set to adjust the hyperparameters, there are enough patterns to be used for a better model.

# **Conclusion:**

As a result, We believe the Australian data will be more suitable to be the training data since it’s more recent and more specific as compared with the global-scale data. Meanwhile, it seems easier to find from reliable data sources.

Then, generalize from Australia to the global scale for the unseen ROI in the other location, it will make more accurate predictions with a well-trained algorithm.