Individual Portfolio

By Ryan Ngo

**Table of contents:**

[**Week 2: 1**](#_cib6ov6u3oyj)

[**Week 3: 1**](#_uulbpfpfh8ot)

[**Week 4: 2**](#_s5zn7sx9w6m)

[**Week 5: 2**](#_hh01gas1vhlx)

[**Week 6: 2**](#_i4zf1b8w4p5o)

[**Reflective of Week 2 - 6: 2**](#_tfpd968cym4)

[**Week 7: 3**](#_anko0fkwpvhf)

[**Week 8: 3**](#_6c8w0tbj6kl9)

[**Week 9: 3**](#_wlyzjzzb807n)

[**Week 10: 3**](#_jpy5ipnfyqha)

[**Week 11: 3**](#_637i7bq1910u)

[**Week 12: 4**](#_xwowi6iduxzx)

[**Reflective of Week 7 - 12: 4**](#_a944b7yzm4br)

# Week 2:

This week, I looked through the detailed description of our group project, looking through all the CSV files provided. However, the weather CSV file is missing many columns and is deemed not enough for the project. After talking with Zac, he put all the datasets on Moodle, including the new weather dataset and different datasets such as the testing dataset and training dataset. After that, we split the work into 6, with each of us doing a part of EDA from 3 datasets. Next week we will bring our work and combine it.

# Week 3:

This week, I looked through all the files that were given to our group. From the description, we noticed that the data given to us is a txf file, but we were able to convert the given data into CSV based on the given code. However, Max said that he could convert the txf file into a real CSV file, so we assigned him to convert the txf file. At this stage, I only loaded the data and looked through the training and testing dataset to see any basic information about the dataset. Although the weather dataset covers the data from 01-01-2010, we decided to take the data from 25-04-2020 to match it with the training and testing datasets. I have pushed my progress on the progress on weathering dataset, with a smaller dataset that suits the testing and training dataset, dropping all the NaN in the dataset. Next week we will start EDA for this project.

# Week 4:

This week, I started EDA for this project. First I loaded the training and testing dataset, and I observed that the starting time for each solar panel is different, with Solar 2 starting in 2019 while other solar power started in 2020. After further analysis, I noticed that if we reduce the Solar 2 data, dropping all data from 2019 to 2020, there would be many important missing data values, as the highest power output of Solar 2 lay in the range of 2019 to 2020. Therefore, our team decided to keep the whole Solar 2 data, and we separated the training and testing datasets by categorise them into 5 different solar stations, hence keeping all important data. We also observed that there is a constant power output between April and June of 2020, hence indicating that the dataset had been interpolated by the provider, so we decided to convert it to NaN and drop those rows. Next time we will merge the solar station datasets with the weather datasets and start some basic analysis with the new merged dataset.

# Week 5:

This week, I merged the solar power output dataset and the weather dataset. We recognised that the time interval for the weather dataset was 1 hour, while the time interval for solar power output was 15 minutes. We decided to interpolate the weather dataset into 15-minute intervals because if we interpolate the solar power output dataset into 1-hour intervals, our power output might have a large margin of error while interpolating the weather datasets would only slightly affect other factors. We also plotted out the correlation heatmap between every factor and we observed that the power output has the highest correlation with surface thermal radiation. We also observed that the average power output is correlated with the weather of the year, with summer having higher power output compared to winter. Next time we will try modelling the datasets with some basic machine learning models.

# Week 6:

This week, I tried multivariate linear regression on the combined dataset. I observed that the model had a small accuracy score between the testing and training datasets. Although solar surface radiation had the highest correlation with power output, the R^2 score was only 0.31, leading to the acquired low accuracy score. We will try different types of machine-learning models next week to find if we can find a higher accuracy score as this score is not satisfied for our project theory.

# Reflective of Week 2 - 6:

For the past 5 weeks, we worked on a Solar Power project. We had many discussions in class about our topic, which direction to go, and what we should do for our project. Everyone participated in the discussion, and we divided the work evenly among each other. We encountered some difficulties with the project as it was a txf file at first but we managed to convert it to a CSV file. I did some EDA with the weather dataset and discovered that it had a different time interval from the solar power output dataset, so I interpolated it into 15-minute intervals. I also started linear regression on the Solar 2 dataset and got an accuracy of 0.32. I feel happy because everyone volunteered to contribute to the project, with a positive response from each team member.

Overall this was a positive experience because despite the low accuracy score of the initial model, we had identified all basic correlations between all the factors, and each team member knew what they needed to do for this project.

For next time, we can apply different machine learning models to find the factor that is most important for the solar power output, hence perfecting our project.

# Week 7:

This week, we managed to fix the linear regression model. Due to the negative value in the dataset, when we tried linear regression, it gave a low R^2 score. We fixed this by taking the absolute value of these power outputs, and trained the model again with this absolute value, therefore improving the R^2 score to 0.8.

# Week 8:

This week, I tried random forests on all features from the weather dataset. I observed that the model had a high R^2 score between the testing and training datasets. With this model, we achieve a 0.83 R^2 score from all the features available from the weather dataset, hence this model can support our aim. We will try feature selection to find if we can improve the accuracy score as this score might be improved further.

# Week 9:

This week, we finished modeling the random forest model, with a high R^2 score. With the feature selection, although the R^2 score did not change much, we can see that the run time is significantly shorter than when utilising all features. We compared this model with linear regression and found out that they have similar R^2 scores. Next, we will try a neural network on this project to see if this model will help us achieve our aim.

# Week 10:

This week, I tried neural on selected features from the weather dataset based on feature selection. I observed that the model had a high R^2 score between the testing and training datasets. With this model, we achieve a 0.82 R^2 score from all the features available from the weather dataset, hence this model can support our aim. We will try an early stop to find out if we can improve the run time of this model.

# Week 11:

This week, we finished modeling the neural network model, with a high R^2 score. Combining with linear regression model and random forest model, we had 3 different approaches for this project, with each model achieving a high R^2 score. We compared each model and finished the group presentation. We also had a team meeting on Zoom, where we had a mock presentation to estimate the final time frame.

# Week 12:

This week we had our final presentation. Overall our presentation was good, with a clear structure and we managed to answer most of our questions. With the presentation finished, we only have to finish the final report for this project, and we can conclude a 12-week-long project.

# Reflective of Week 7 - 12:

For the past 5 weeks, we worked on a Solar Power project. We had many discussions in class about our topic, which direction to go, and what we should do for our project. Everyone participated in the discussion, and we divided the work evenly among each other. We fixed the linear regression problem, therefore we achieved a high accuracy score on the model. We also utilised the random forest model and neural network model, with the accuracy score of both models being high, and we compared the advantages and disadvantages of each model.

Overall this was a positive experience because we managed to fix the previous problem with the low accuracy score for linear regression, and we also achieved a high accuracy score for random forest and neural network models. Every team member contributed to model making, presentation slides, and final report so the workload was manageable.

For next time, we can apply time series and seasonal predictions, hence perfecting our project.