**Melbourne’s Daily Energy Use and Pricing Based on Weather Data**

**Executive Summary**

**Key Insights**

<TBA>

**Key Recommendations**

<TBA>

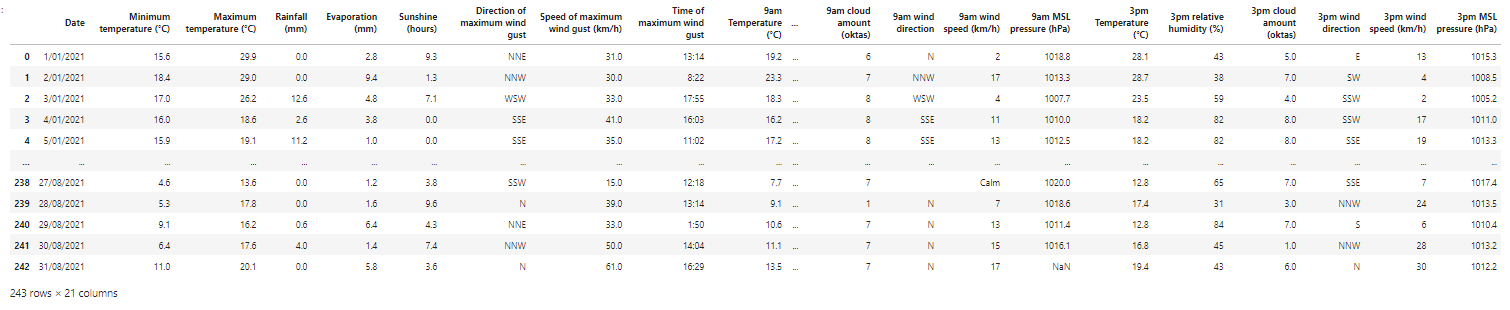
**Data Preparation**

**Datasets**

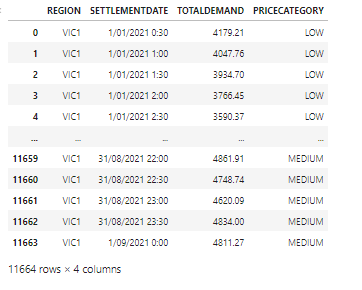
The following datasets were provided for this exercise.

*Table A – Overview of datasets*

|  |  |  |
| --- | --- | --- |
| **Dataset name** | Weather data | Price Demand Data |
| **Description** | Key weather indicators for the City of Melbourne for each day between January to August 2021 inclusive from the Bureau of Meteorology. | Energy price and demand figures for the State of Victoria for each half-hour period between January and August 2021 from the Australian Energy Market Operator. |
| **Format** | csv | csv |
| **Number of columns** | 21 | 4 |
| **Number of rows** | 243 | 11,663 |



*Figure 1 – Snapshot of weather data (raw)*



*Figure 2 – Snapshot of price demand data (raw)*

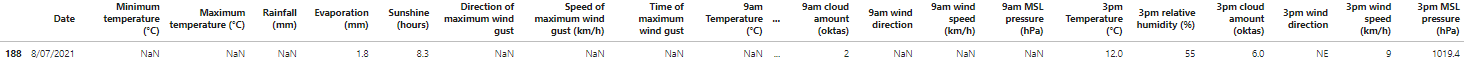
**Data Wrangling**

Analysis of both datasets indicates that there are no unexpected values (e.g. negative values where only positive values are expected) or outliers (e.g. unusually high temperature beyond 39°C).

*Missing data*

However, analysis of the weather dataset indicates that there are missing observations from multiple variables across the weather data. The missing observations impact different variables on different dates. For example, the value for the *9 am wind direction* is missing on 23/01/2021 whereas the value for the *3 pm wind direction* is missing on 21/03/2021. This suggests that the missing values are missing completely at random (MCAR).

Crucially, 11 values are missing on 08/07/2021 including temperature data for the day. For Model 1, the median maximum temperature in °C across the weather dataset has been inputted to replace the missing value. Given that this missing data represent a small subset of the total data population (1 of 243 rows), it is unlikely that this substitution will have a material impact.



*Figure 3 – Weather data on 08/07/2021 before the substitution of missing information*

For Model 2, the missing values were substituted as follows:

* For string datatype, the missing values were substituted with the mode of the corresponding variable.
* For integer or float datatype, the missing values were substituted with the mean of the corresponding variable.

*Inconsistent formatting*

The sole common denominator identified for both datasets is the date values. The date format in the weather dataset is formatted as “dd/mm/yyyy” (see column *Date* - Figure 1) whereas the format in the price demand dataset is “dd/mm/yyyy hh:mm” (see column *Settlement Date* - Figure 2). To enable data aggregation, the date format in both datasets was converted so they are consistent.

*Additional variable – Variance between maximum temperature to average maximum temperature*

For model 1, the independent variable selected for the regression model is the variance between the maximum temperature of the day against the average maximum temperature across the provided dataset. Therefore, a new variable containing the absolute value (i.e. negative values are converted into positive values) of this is added to the weather data. The reason why absolute value is used is that it is expected that energy consumption will increase during both during coldest and hottest days.

*Additional variables – average temperature, average date temperature and average humidity*

For model 2, the below variables were added as they scored highly in the correlation matrix with the maximum price category:

* Average temperature – average between maximum and minimum temperature of the day.
* Average day temperature – average between 9 am and 3 pm temperature.
* Average humidity – average between 9 am and 3 pm relative humidity.

**Data Aggregation**

*Aggregating daily energy demand*

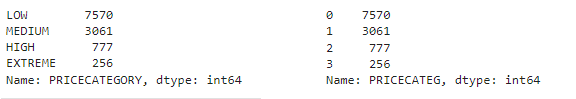
For model 1, the Total Demand values for every half-hour are summed up so a daily total demand is obtained for each settlement date. This is to enable the model to predict total daily energy demand/usage (the dependent variable) based on the selected independent variable which is *the var to avg max temp*.

As a result of the date conversion method employed, there is a half-hour data for 01/09/2021. This was removed there is no corresponding weather data. Given that there is still a population of 243 records remaining once this is removed, the removal is immaterial.

The total demand data is then joined up with the weather dataset based on the date value.

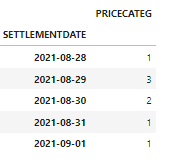
*Assigning maximum daily price category*

As noted above, the price demand dataset has data for every half-hour including price category for every half-hour. To obtain the maximum daily price category for Model 2, a numerical value of 0-3 is first assigned to each price category:



*Figure 4 – Maximum daily price category data and assignment of numerical value.*

Then, the relevant dataset for Model 2 (date and price category) is grouped by the maximum value of the price category for each date.



*Figure 5 – Maximum value of price category for each date*

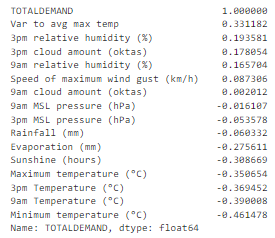
For model 2, the price category is then joined up with the weather dataset based on the date value.

**Model Building**

**Model 1 – Prediction of Daily Total Energy Usage Based on Provided Weather Data**

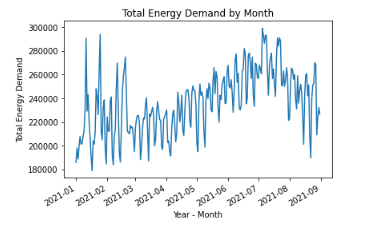
For Model 1, a simple linear regression model is employed to predict total energy use based on *var to avg max temp*. This method is selected as both variables contain numerical data and regression is a useful statistical tool to quantify the relationship between two variables.

This variable has a Pearson correlation coefficient of 0.3316. This suggests that there is a moderate correlation between this variable and total energy demand. There are other variables with higher Pearson correlation coefficients (albeit, they are negative figures which suggest negative correlation) such as Minimum temperature. However, visual inspection of line charts of these variables suggests that the correlation may be inconsistent and these variables are not appropriate for model 1. For example, it is expected that the minimum temperature during winter months would be lower than during the warmer months but energy demand seems to peak during both the hottest and coldest days (see Figure 8).

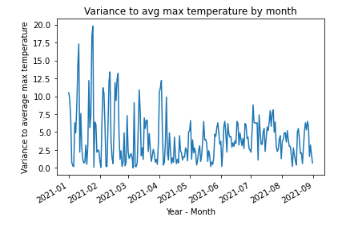


*Figure 6 – Person correlation coefficient of variables in weather data against the total demand*

A visual inspection of line charts created on the dependent (Figure 5) and independent variables (Figure 6) also suggests that there is a strong correlation. In particular, the correlation appears to be stronger during the warmer months (January 2021 to April 2021).

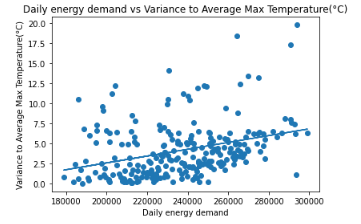


*Figure 7 – Line chart on daily total energy demand*



*Figure 8 – Line chart on daily variance to average maximum temperature*

A scatter plot created also provides a visual hint that there are potential causal relationships between the two variables.



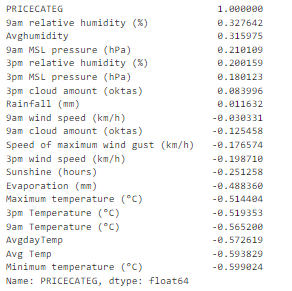
*Figure 9 – Scatter plot of daily energy demand vs variance to average max temperature*

For Model 1, 80% of the dataset is assigned as the training dataset with the remaining 20% assigned as the test dataset to avoid overfitting. The random state is not specified to ensure that the model is non-deterministic. Based on its learning from the training dataset, Model 1 predicts total energy demand based on the input of variance between maximum temperature °C for the day to average maximum temperature.

**Model 2 – Prediction of Maximum Daily Price Category Based on Provided Weather Data**

For Model 2, two types of models are employed – A K-Nearest Neighbour classifier and a Decision Tree. The classification models are selected because the target is finite which are the 4 price categories (low, medium, high and extreme).

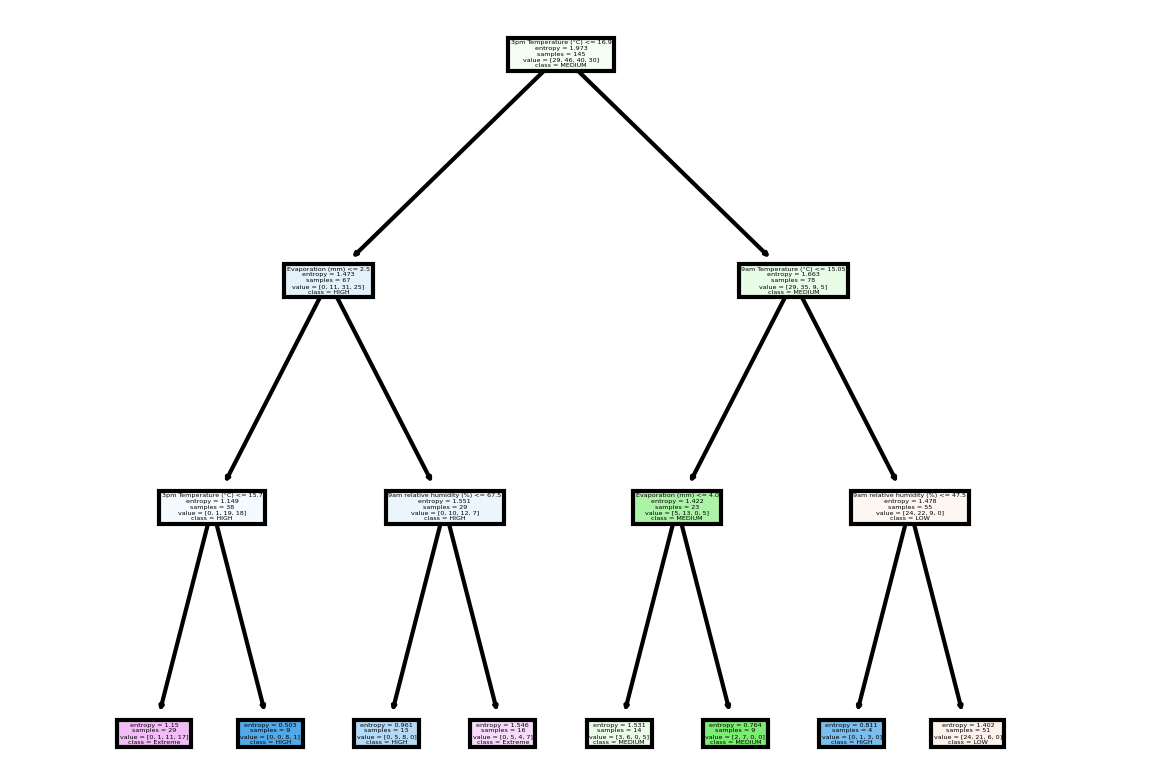
First, the features in the aggregated dataset are reviewed using the correlation matrix. Features with correlation scores of +/- 0.2 and/or above are selected.



*Figure 10 – Correlation matrix of the features against price category*

For Model 2, 60% of the dataset is assigned as the training dataset with the remaining 40% assigned as the test dataset to avoid overfitting. Based on its learning from the training dataset, Model 2 predicts the classification of the maximum price category for the day based on the feature data.

For the Decision Tree, below is an outline of the model’s operations:



*Figure 11 – Decision tree model operations*

**Results**

**Assessment of Model Effectiveness**

*Model 1 – Effectiveness*

To test the effectiveness of Model 1, the following tests were performed:

* The first test is to use the model to predict total energy demand on the 16th variance to average max temperature. The results were as follows:

|  |  |
| --- | --- |
| Model prediction – total energy demand | 240,407 |
| Actual – total energy demand | 251,081 |

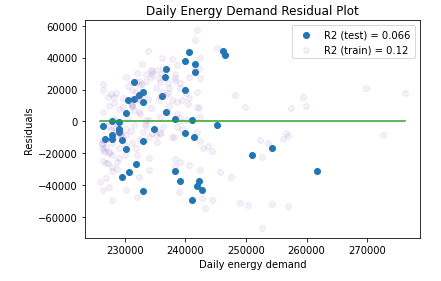
Although the result produced by the model appears to be pretty close to the actual, it is not efficient to manually test the model this way.

* The next test performed was to calculate the coefficient of determination (r2) values and the result is as follows:

|  |  |
| --- | --- |
| r2 -Training dataset | 0.1149 |
| r2 -Test dataset | 0.0661 |

The r2 scores suggest that the model is not effective in predicting the total energy demand as only 10.23% - 11.14% of the variation in these is explained by the variation of variance between maximum temperature against average maximum temperature.

* The final test performed to assess the effectiveness of the model is a residual analysis.



*Figure 11 – Residual Analysis*

The residual analysis (refer to Figure 11) suggests that the residuals are independent. However, it is observed that the residuals are not linear and the variances are not constant. This suggests that there are violations of the regression assumptions and that the linear regression model is not necessarily appropriate for two variables.

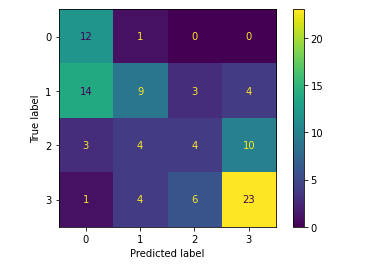
*Model 2 – Effectiveness*

To test the effectiveness of the K-NN, the following tests were performed:

* The K-NN model has an accuracy score of 0.4591 which means that the model makes a correct prediction approximately 45.91% of the time.
* To further test the accuracy of the model, the K-fold cross-validation is employed. First, the dataset is split into 10 partitions. The first fold will be used as test data and the remaining fold will be used as training data. In the next iteration, the second fold will be the test data, and so on until the last fold. Using this method, the accuracy score achieved for the K-NN model ranges from 0.4198 to 0.4692.

To test the effectiveness of the decision tree, the following tests were performed:

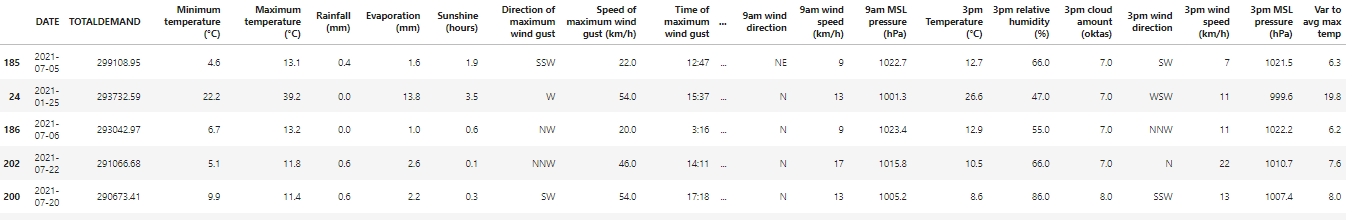
* The decision tree has an accuracy score of 0.4898 which means that the model makes a correct prediction approximately 48.98% of the time. This suggests that the decision tree is slightly more accurate than the K-NN model.
* The accuracy score achieved by the model can also be summarized in a confusion matrix (see Figure 12 below).



*Figure 12 – Confusion Matrix*

**Insights from Data Analysis**

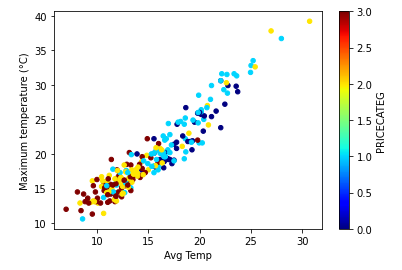
Across the dataset, total energy demand peaks on either the hottest days or the coldest days (see Figure 9 below). This is consistent with the initial conclusion that there is at least a moderate correlation between the variance between the maximum temperature to the average maximum temperature across the dataset.



*Figure 13 – Merged dataset ranked by total energy demand*

However, the use of this variable in model 1 does not lead to highly accurate predictions of total energy demand. The variability in daily energy demand differs between colder and hotter days.

An observation of the distribution of the maximum daily price category (see Figure 14 below) suggests that Extreme price categories are more likely to occur when the temperature is colder. As observed below, the red dots representing the Extreme price category are more prominent where the average temperature is less than 15°C.



*Figure 14 – Distribution of maximum daily price category*

For Model 2, the use of temperature and other weather data as features to predict the maximum price category achieved a better result than Model 1.

Observation of all of the data suggests that energy prices seem to be at their highest during colder weather when the energy demand is high. Interestingly, high energy demand during warmer weather does not seem to yield the same outcome.

An improved model that can provide more accurate predictions of energy demand/usage based on weather data can be used by energy providers to plan for peak demand periods and prevent outages. From a profitability perspective, energy companies may also wish to increase their pricing when energy demand peaks during warmer months.

**Limitations & Improvements**

*Model 1 – Limitations & Improvements*

For model 1, a linear regression model was utilized and therefore only a single independent variable was considered. The model did not consider additional variables which may improve the accuracy of predictions such as other weather-type variables (e.g. rainfall) or the impact of the different types of days weekdays vs weekends, school holidays, public holidays) to energy usage.

The following improvements may improve the model’s ability to predict energy demand:

* This report has not explored the use of a multiple regression model. From the analysis above, we know that the energy demand moves differently during colder and hotter months. The inclusion of additional variables may lead to a more accurate result.
* Principal component analysis (PCA) may also be employed to find a new set of features that better capture the variability in energy demand.

*Model 2 – Limitations & Improvements*

As noted above, Model 2 achieves a better predictive outcome in comparison to Model 1. However, the Model’s achieved accuracy score is below 0.5 which means that the model can be improved.

The following improvements may improve the model’s ability to predict the maximum price category:

* Whereas features with lower correlation have been filtered out, it is possible that dimensionality reduction techniques can be applied to narrow down the number of features and improve the model’s performance.
* PCA may also be employed to find a new set of features which improves classification accuracy.