COMP20008 SEM2 2022

Assignment 2 – Final Submission

[Can –Teddy]

Research Question + Aim

Can we predict the energy demand given a weather forecast?

The goal of our study is to find whether weather data is correlated with energy demands. Our underlying hypothesis is that energy usage fluctuates depending on the weather, for reasons such as climate control, or increased amounts of people staying indoors, leading to higher energy usage.

We will also transform our data to include weather data from a previous day, and including time data to see if these extra attributes can further increase our model's performance.

Therefore, we aim to create a model that can use the attributes of weather and time to gauge the demand for a day.

Audience

The outcome of this study will be highly beneficial to energy companies, as predicted the energy usage and adjusting the power produced will not only reduce the prices of energy, but also save the environment from reduced coal-plant usage.

Additionally, city planners and engineers may also look at pre-existing weather data to gauge the fluctuations in energy demand for future development.

Datasets

Two sets of data of excel type were used in this study, weather data gathered from the Bureau-of-Meteorology, and energy/electricity demand data from AEMO, in the period of 2021-2022.

The weather dataset includes various types of indicators of the weather, ranging across a few cities in Australia each day, over a period of 13 months. It is to be noted that the weather data for each city is located in a separate excel file.

Entire Day	**3pm / 9pm**
• Minimum Temperature • Maximum Temperature • Rainfall (mm) • Evaporation (mm) • Sunshine (hours) • Direction of Max Wind Gust • Speed of Max Wind Gust • Time of Max Wind Gust	• Temperature • Relative Humidity • Cloud Amount • Wind Direction • Wind Speed • MSL Pressure (Mean Sea Level)

The demand dataset includes the energy demand, and whether there is a price surge (Boolean) in different states of Australia. Both datasets range in the same time period of time, and have matching state-city pairs.

There are no data quality issues to be concerned with, as both institutions the datasets came from follow a consistent data format, ensuring accuracy, completeness, consistency, timeliness, believability, and interpretability.

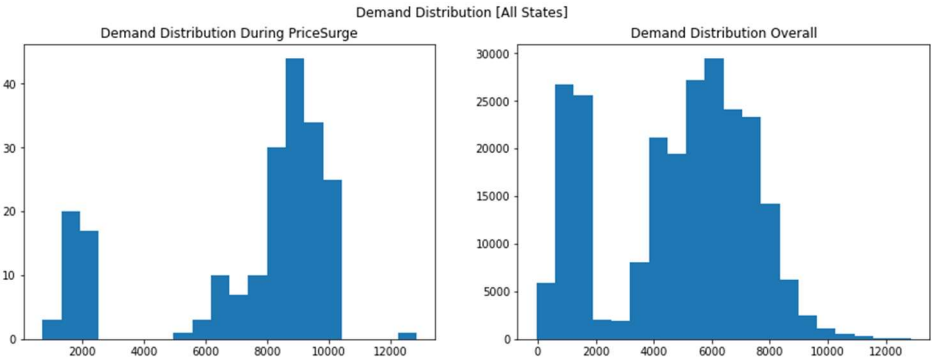
Pre-processing / Visual Analysis

There were no data-quality processing steps required due to reasons previously mentioned.

Each column for both the weather and demand datasets were renamed into something more code-friendly (i.e., filter non-letters, lower-case, abbreviate). This was done using a set of regular expression rules.

For the demand dataset, we had opted to split the demand data of each state into separate dataframes. We had noticed that the distribution of the entire demand dataset was unusual, and so concluded that each state has a different demand distribution.

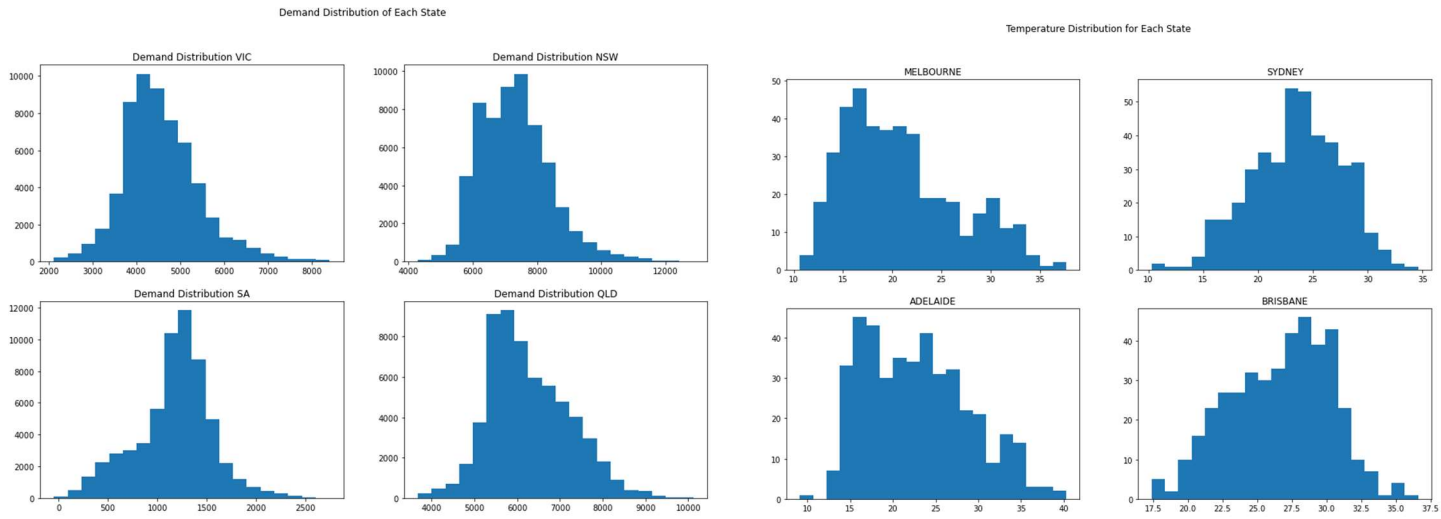
Analysis of Demand



	total_demand		
	min	max	mean
REGION			
NSW1	4286.0	12864.0	7320.0
QLD1	3672.0	10119.0	6276.0
SA1	-46.0	2742.0	1172.0
VIC1	2123.0	8406.0	4569.0

We then visually analysed the demand of each state individually, to find out whether each state behaves differently in terms of demand. We also hypothesise that the cause for different central tendencies for the demand of each state is likely due to outside factors such as population.

Looking at the demand for each state, we noticed that the shape of distributions was also largely different, so we also compared them to our main attribute, temperature.



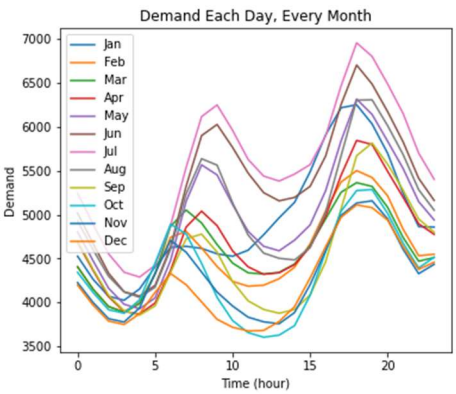
We can see that the temperature distribution of each state also varies widely from each other. The demand for both VIC and QLD were largely similar, yet the temperature distribution reflected each other.

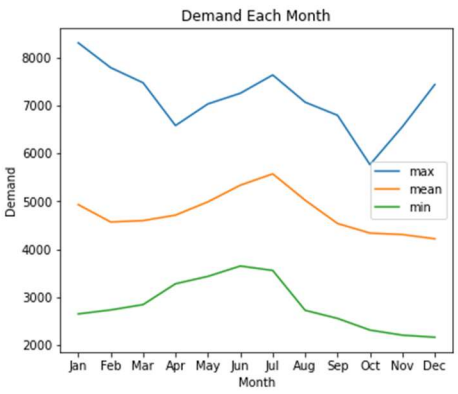
Under our hypothesis, if Brisbane is consistently at its higher range of temperatures, we would expect to see the central tendency of demand at the higher points, but that is not the case. Thus, as we do not have enough domain knowledge about the geographical influences on our datasets, we have opted to create separate learning models for each region.

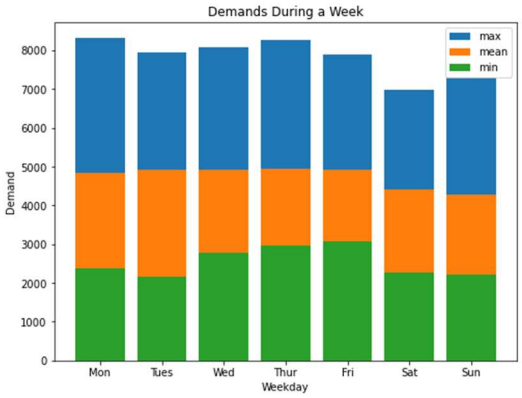
We should note that there is definitely a possibility that linking the regions together would create a more comprehensive model, but we have opted not to take the risk as we do not have enough time to analyse both.

Analysis of Time

Additionally, we also looked at the demand behaviour of each month, and each weekday. This was done because we had also hypothesised that people might be more likely to use heating/aircon during specific months of the year, or that energy usage might be higher when people are at home (i.e., during weekends).







It is very noticeable that demand varies from month to month, and that there is a significant increase in the energy used on weekends. Thus, we had used some data transformation techniques to include the new attributes: month, and day.

After normalising the demand of each state, we also decided on 3 measures of demand that our model will learn and predict. Because the weather data was based on individual days, and the demand data was accurate to 30 minutes, some information had to be lost when linking the data.

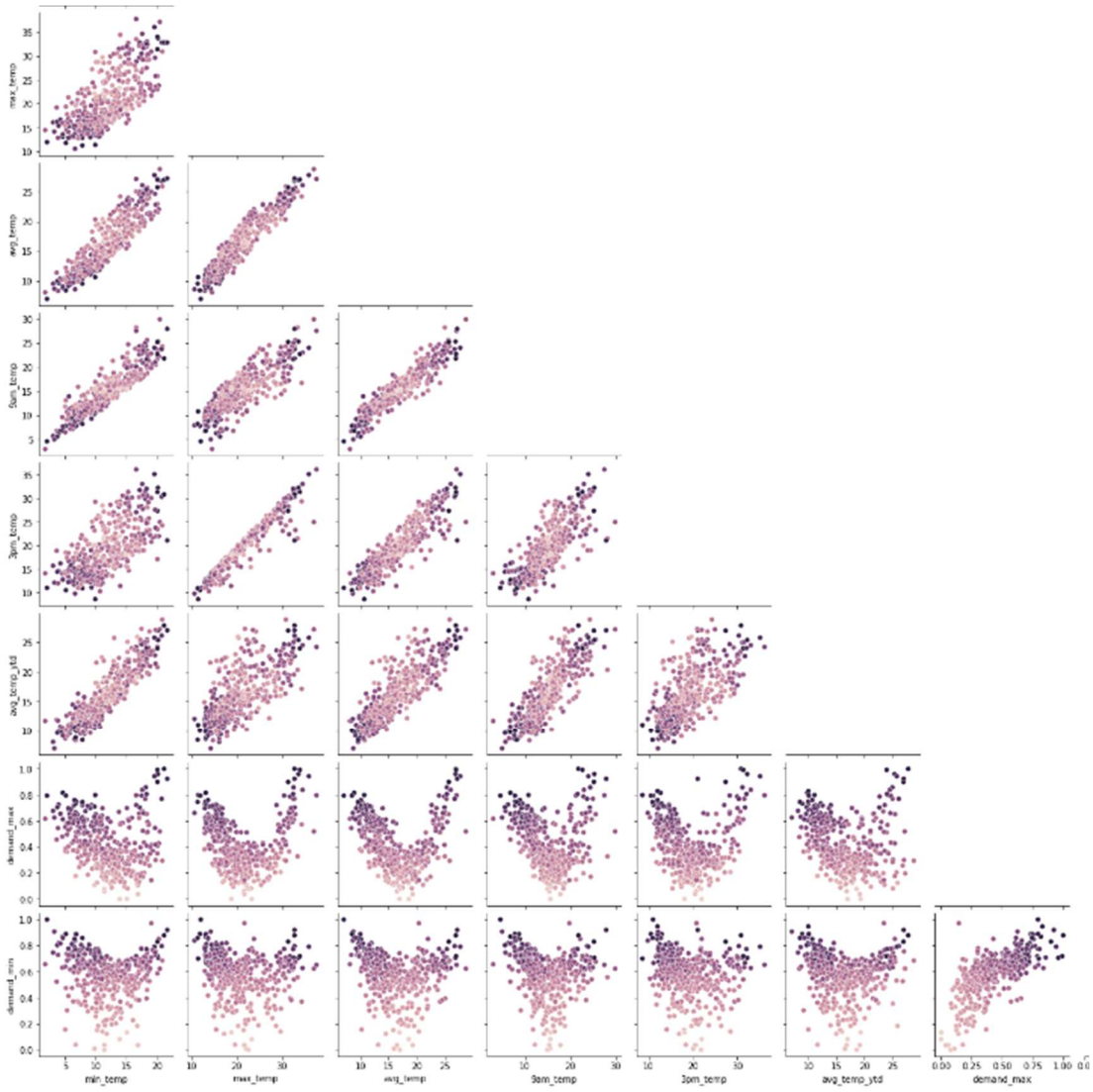
Data linkage was performed joining the demand dataset to the weather dataset, with demand being split into the mean, and the minimum and maximum value of the day.

Feature Selection

Attributes that are not able to be predicted by advanced weather forecasts were omitted from our study, as our goal is to be able to predict demand ahead of schedule. Further, we used a simple seaborn pair-plot of each variable (continuous) against each other, and against demand to further understand their relationships.

This pair-plot provided us with knowledge that the independent continuous variables (temperatures) we all correlated with each other. Meaning that if linear regression were to be used, we would only be able to use one of the variables, or a measure of all of them.

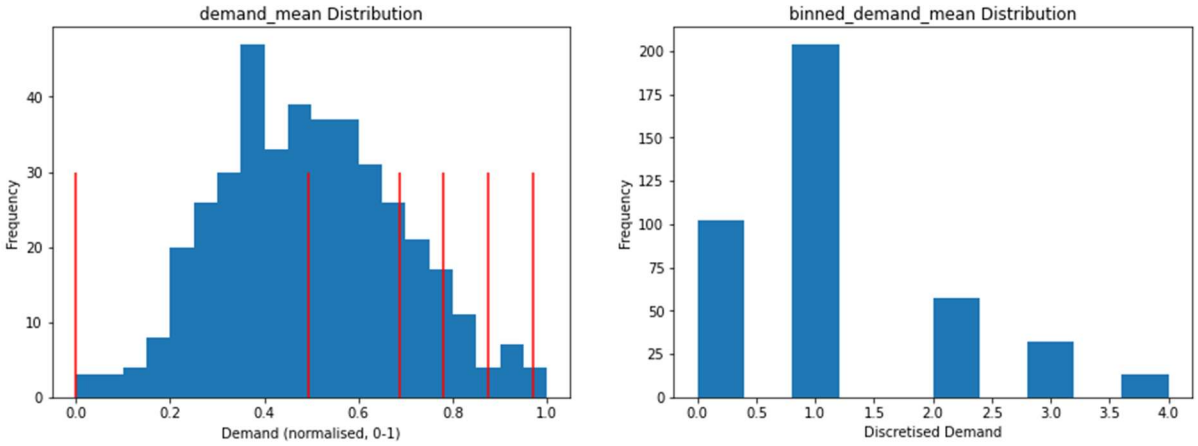
However, this would remove too much information and reduce the complexity by a large margin, so we will opt to use classification techniques instead. Further, it is clearly visible that the correlation between demand and temperature is not linear, thus even if linear regression was performed the residual analysis would not align with the rules of linear regression.



Discretisation of Demand

Because we will be using classification type modelling, we first have to bin the demands. The way we have discretised the demand is through domain knowledge by looking at the distributions. We have split the demand into 5 class labels: low, normal, high, very high, and extreme.

Our first bin will be between 0 and the Q1 (i.e., 1st interquartile range), followed by Q1 – Q3, which is followed by 3 equal width bins between Q3 and 1 (the maximum value, as we have normalised the demand).



Analysis – Methods

As previously mentioned in feature selection, we will not be using linear regression due to the nature of the dataset. We will be exploring two types of classification models, kNN (k nearest neighbour) and decision trees.

We will first split the dataset into a train-test set of 8:2. We will use k-fold cross analysis with k=4 on each of our analysis methods with varying attributes used and bin sizes to find the best performing model.

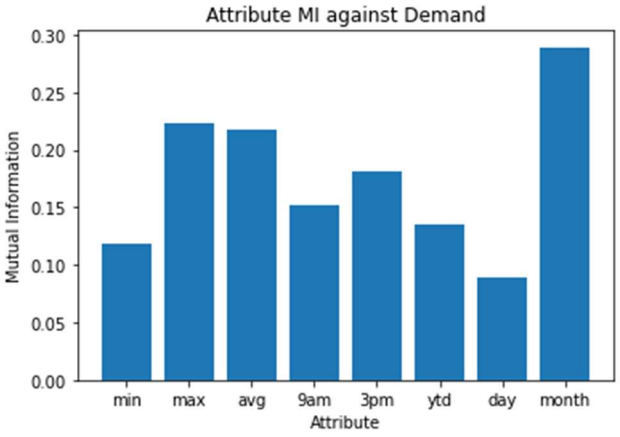
The chosen model will then be tested against the test set, and this will be our final model for discussion.

Analysis – Feature Selection

Under both decision tree and kNN models, we noticed that performance was highest when we only used the attributes of minimum and maximum temperature, followed by the month and day. It should be noted that all attribute's p-values measured against demand were below 0.00, meaning that the null-hypothesis that they are not correlated is unlikely.

Even though correlation analysis show that the other attributes are highly correlated with demand, including them only increased noise. We predict that the cause of this is due to the fact that these attributes are too similar, and that only including the minimum and maximum values is best because they are the two least correlated value pairs.

Due to kNN models only being able to use continuous variables, we were also unable to include the day type (weekday/weekend), but we were able to substitute the month as a continuous variable, which unexpectedly resulted in better performance.



We can see that both maximum temperature and average temperature share the highest MI (apart from month). But contrary to expectations using both maximum temperature and average temperature performs worse than using minimum and maximum temperatures.

This is highly likely due to the fact that the average temperature is just the average of minimum and maximum temperatures.

Analysis – Variable Binning/Discretisation

Two methods of binning the continuous variables for the decision tree was used. The first one being the same method used to bin the demand, and the second being equal width bins.

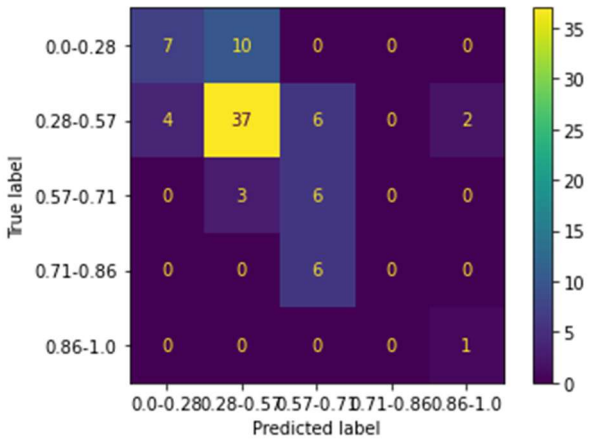
We found that the use of equal width bins, and a setting of 4 bins maximised performance.

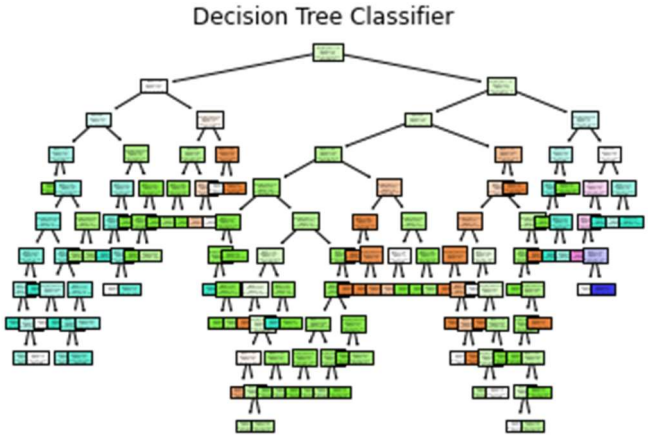
Analysis – Best Model

We ran the 4-fold cross validation 5 times, and obtained an average accuracy of .73 using a decision tree model, and an average accuracy of .66 for kNN model. We also ended up only using the models to predict the maximum demand, as not only was performance significantly worse, but it also created a lot of information that we could not completely explain in this research.

After choosing the decision tree model with the best fitted parameters, we trained the model on the entire training dataset, and measured the results.

Class	n (truth)	n (classified)	Accuracy	Precision	Recall	F1 Score
Low	11	17	82.93%	0.41	0.64	0.50
Medium	50	49	69.51%	0.76	0.74	0.75
High	18	9	81.71%	0.67	0.33	0.44
Very High	0	6	92.68%	0.0	0.0	0.0
Extreme	3	1	97.56%	1.0	0.33	0.50





Interpretation of the Results

The final average accuracy turned out to be 0.62, much worse than the 0.73 we got off of the k-fold analysis. The accuracy for the medium class (IQR-1, IQR3) was surprisingly the worst, this is likely due to the fact that it has the most true values. The F1 score for the medium class shows that its performance is still acceptable.

It should also be noted that in the test dataset there were zero occurrences of the very-high class label, a small issue due to the way we've worked with the dataset (individual states).

But what we are concerned about is when the demand is higher than normal, to which looking at the precision scores (just for high and extreme), they seem to be quite good. With a bigger dataset we believe that there is a strong likelihood of being able to predict when the demand will be higher.

As a whole, we would not recommend this model for our audience to predict demand (reasons being stated after).

Limitations and Future Improvement

There have been many limitations caused by our approach. The impact of not linking the datasets together was severely underestimated. This can clearly be seen in our training, validation, and test section.

It turns out that the risk we tried to avoid by separating the states increased the volatility of our results. Additionally, we were also not able to include the results we got from the other states (so far it has only been VIC) as the study would have been twice as long.

However, this is not to undermine the results and findings that we have gathered in this study, as there is an almost definite correlation between temperature, and energy demand (and also time).

If further analysis were to be performed, two methods would be considered. The first being linking all states together via normalised and standardised demand, which looking back, we should have done in the first place.

The second being an extension of the first, but rather than linking demand with weather, we link weather to demand instead, so for each half hour in the demand dataset the weather data will just be substituted for the entire day.

The both of these methods allow for a much greater amount of data to be processed, and upon analysis we may then choose to further reduce the amount of data as we see fit.