Assignment 2: Data Science Project

Group 4

University of Melbourne

Authors

Anuradha Mani

Basaran Ozal

Brandon Louey

Camilo Lancheros

Word count: 2464 (excludes any figures, tables, and appendices)

Objective

The objective is to create a model that can predict the maximum daily energy use and pricing based on weather data. Modeling is based on two data sets taken between November 2022 and April 2023: (1) key weather indicators for the city of Melbourne at daily intervals, and (2) energy price and demand figures for Victoria at half hour intervals. Weather data is taken from Bureau of Meteorology (BoM) and price demand data is taken from Australian Energy Market Operator (AEMO).

Contents

Authors..... 0
Objective..... 1
1. Data cleaning..... 3
2. Building of Model..... 3
2.1 Pre-analysis..... 3
2.1.1 Pre- analysis of price and demand..... 4
2.1.2 Pre- analysis of weather..... 5
2.1.3 Pre- analysis of combined dataset..... 6
2.2 Regression Model..... 7
2.2.0 Feature Selection for Regression..... 7
2.2.1 Linear Regression..... 8
2.2.2 Decision Tree Regressor..... 8
2.2.3 kNeighborsRegressor..... 9
2.3 Classification Model..... 9
2.3.1 Bin data and feature selection..... 9
2.3.2 k-Nearest Neighbors..... 9
2.3.3 Decision Tree Classifier..... 10
3. 1 Model Effectiveness..... 11
3.1.1 Evaluation of Regression Models..... 11
3.1.1.1 Linear Regression Model..... 11
3.1.1.2 Decision Tree Regressor..... 11
3.1.1.3 K Neighbors Regressor..... 11
3.1.2 Classification..... 12
4. Weather Analysis and Insights: Exploring the Impact on Energy Consumption..... 14
5. Limitations..... 15
5.1 Internal..... 15
5.2 External..... 16
References..... 18

1. Data cleaning

Before starting data cleaning, data files were uploaded using the encoding **'latin1'**. It is important to use the encoding when reading files with non-ASCII characters to ensure that all characters are accurately represented. All data cleaning methods were implemented using Python programming language and Pandas library.

The following data cleaning methods were applied to the datasets:

1. The `info()` method was used to find out what needs to be fixed, such as changing the data type.
2. A `column_names` list was used to name Price_demand columns. Names are taken from AEMO.
3. Empty columns and rows were removed using the drop() method. In **weather data**, 4 empty columns were removed: Evaporation (mm), Sunshine (hours), 3pm cloud amount (oktas), 9am cloud amount (oktas), and Location.

   For **Price_Demand** data, 'Area' and 'Name' columns are dropped as they contain only one data point.
4. Checking for missing values is important to ensure that the dataset is complete and accurate. The isna() method was used to detect missing values in a dataset and row 174 was removed.
5. The `pd.to_datetime()` function was used to convert a string to a datetime object and specified as '%d/%m/%Y %H:%M' converts a string to a datetime object.
6. The `replace()` method was used in the **weather** dataset for windspeed columns. "Calm" is equivalent to 0 according to AEMO and replaced.
7. The astype() method was used to convert the data type of a variable. .astype(float) was used to convert it to a float data type. For example, from the **weather dataset**, the '9am relative humidity (%)' column was converted from an object data type to a float data type.
8. The Date is set as an index to resample the data by different time intervals.A new dataframe was created to include minute, hour, day, month and quarter
9. The two datasets were merged into one file and saved before starting the analysis.

Due to the data set being relatively small and simple, Python and Pandas functions were sufficient to clean the data. For detailed coding steps, please refer to the Python coding file.

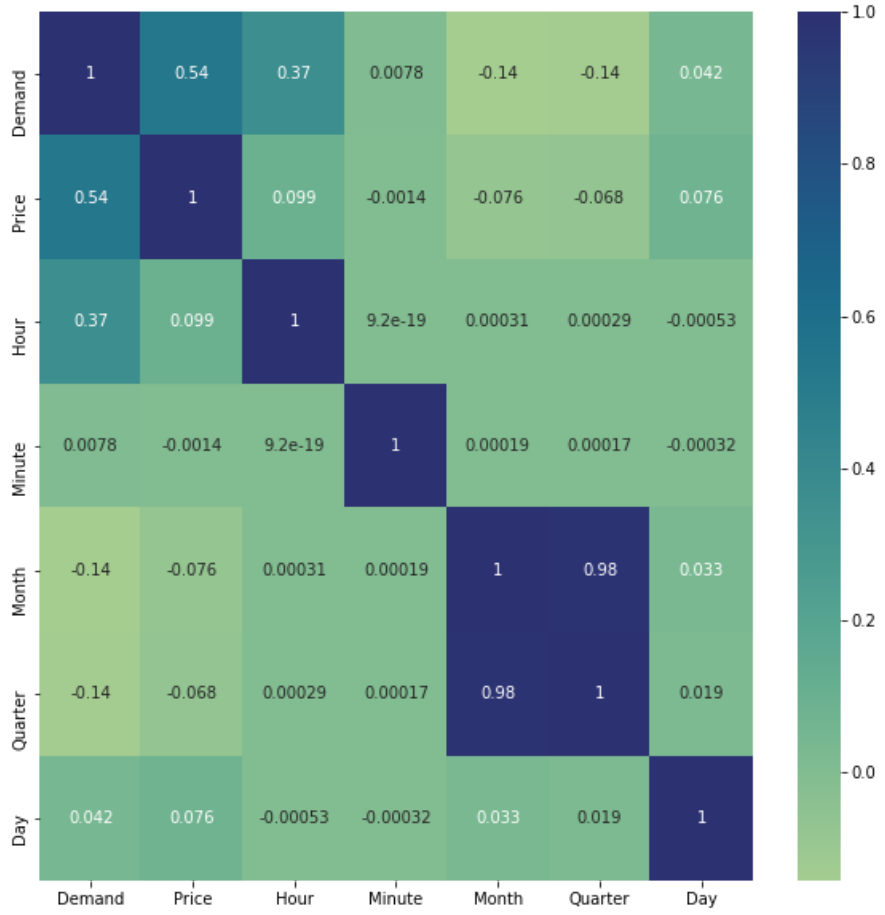
2. Building of Model

2.1 Pre-analysis

A pre-analysis was performed on both datasets to quickly draw insights into potential patterns and relationships between features.

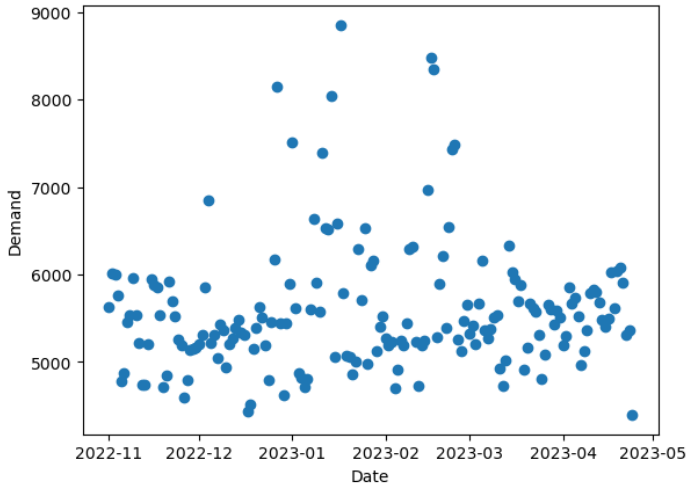
2.1.1 Pre- analysis of price and demand

First, Seaborn was used to produce a heat map to visualise the relationship between demand, price and date broken into increments. Notably, some intervals indicate a relatively high correlation to demand, while some have close to no correlation (Eg. demand/hour has a Pearson correlation of 0.37, while demand/minute is only 0.0078). This indicated that breaking the date into intervals would provide valuable information to the model.

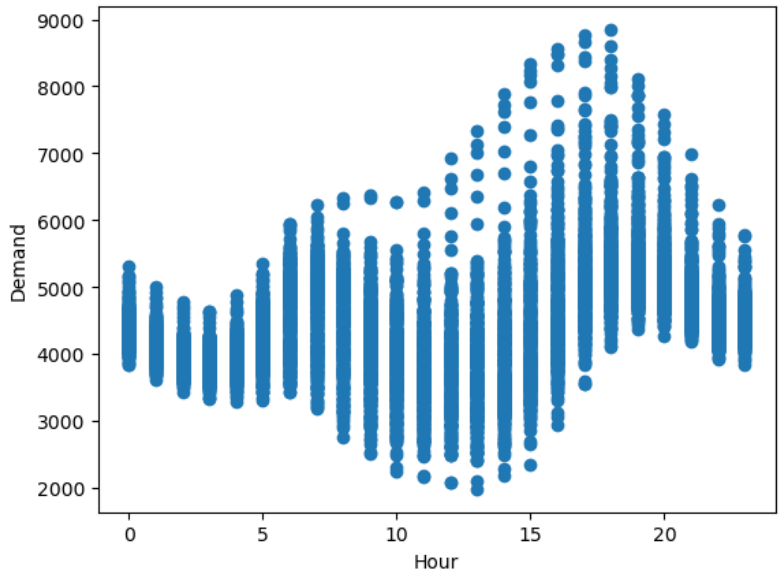


Further modeling was done to investigate the date and time interval relationship to demand.

A scatter plot of demand against date showed that demand is highest in the months of Dec, Jan, and Feb, which matches with the summer time in Australia.



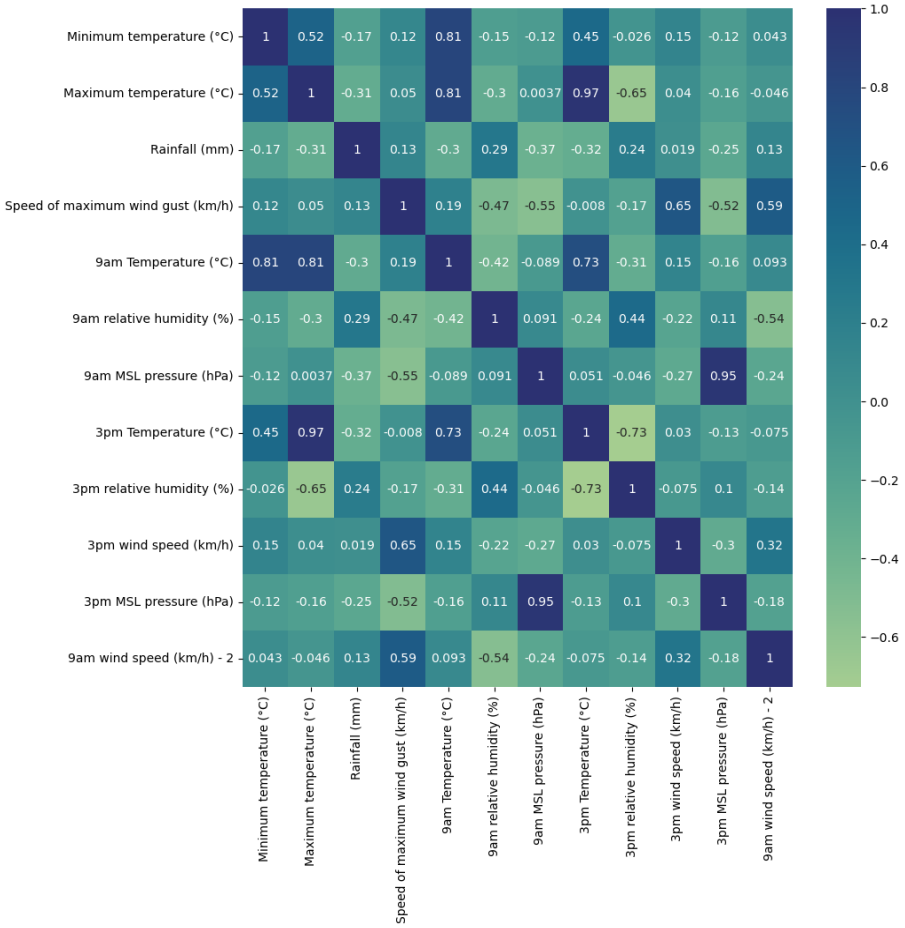
A scatter plot of demand against hour produced a familiar energy demand curve, known as the duck curve. This usage pattern follows curves based on morning, midday and afternoon energy usage and production. Solar energy is understood as the primary driver to the dips at midday where it is the sunniest.



2.1.2 Pre- analysis of weather

Similarly to the price and demand date, a heat map of weather features was produced. Notable relations are how the minimum temperature is closely correlated to the 9am temperature, and

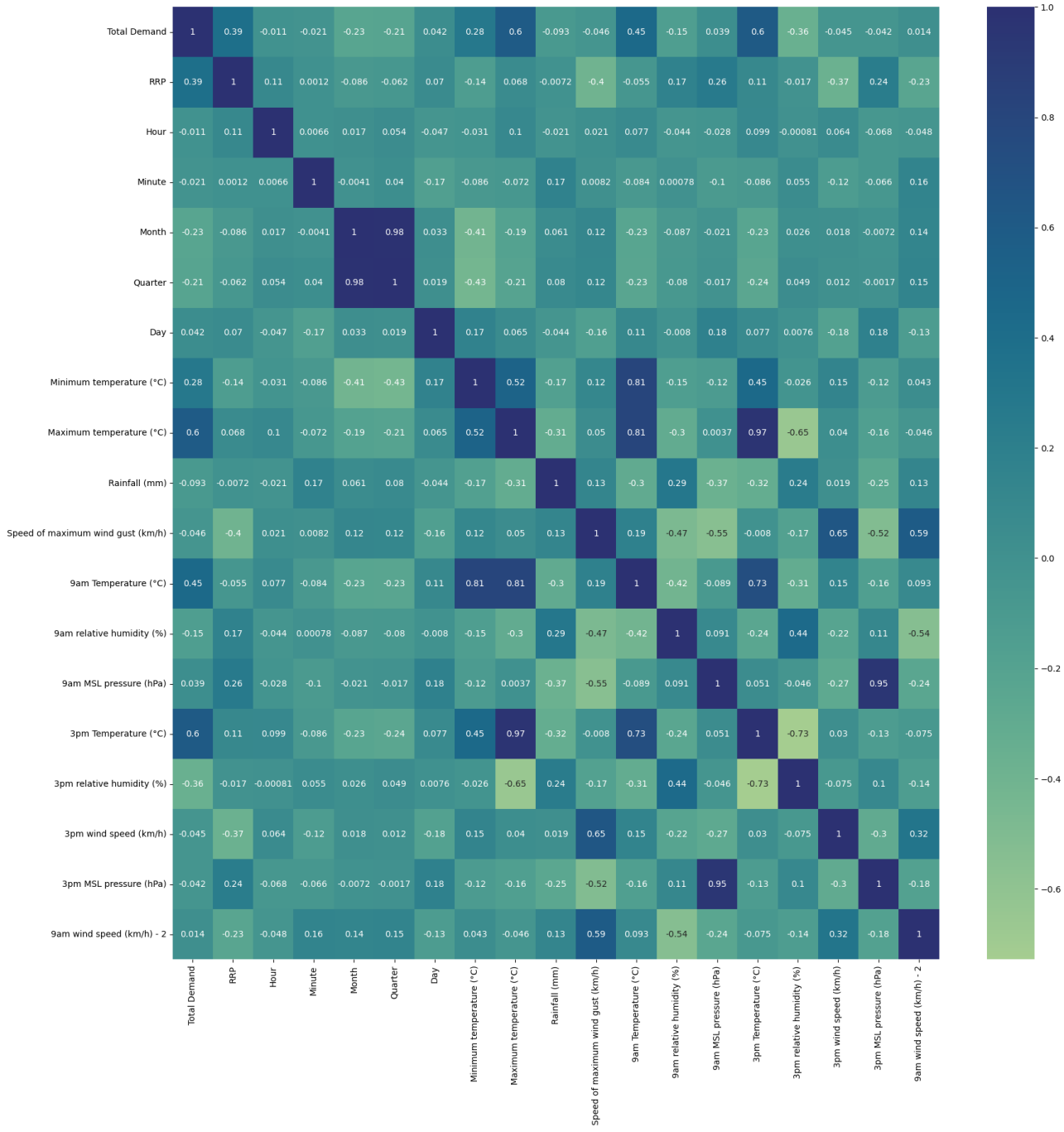
similarly with maximum temperature at 3pm. Close correlation between features can inform which features may be able to be merged or removed in the final model.



2.1.3 Pre- analysis of combined dataset

Finally, a heat map of a combined dataset of weather, price and demand is produced.

See Section 4 for analysis and insights found.



2.2 Regression Model

To build a predictive model, several regression and classification methods were tested and described in the below sections.

2.2.0 Feature Selection for Regression

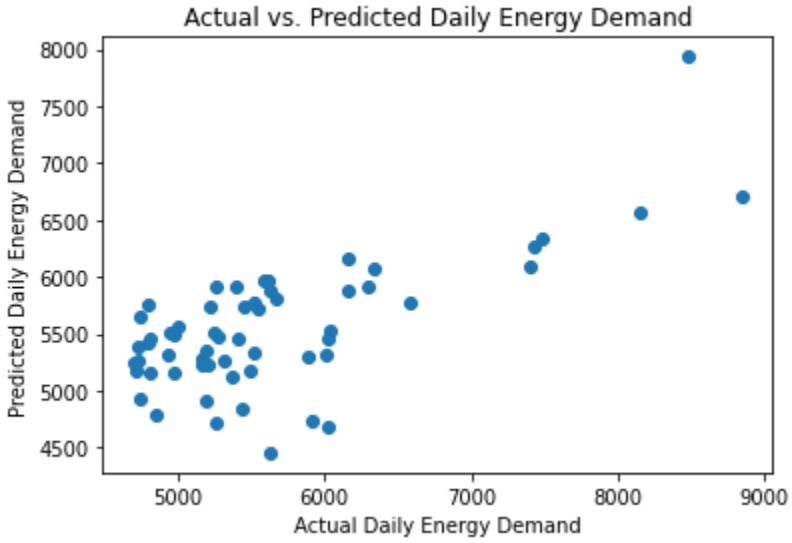
First, a number of feature sets were created with varying features included in each. Each set was created using correlating features informed by the heatmap in Section 2.1.3. Feature sets were experimentally swapped in and out to build models and analyse the response to outputs. See code for detail on feature sets.

'SelectKBest' method is used to eliminate redundant features and is experimentally adjusted.

'PCA' method is also to further reduce the dimensionality and noise.

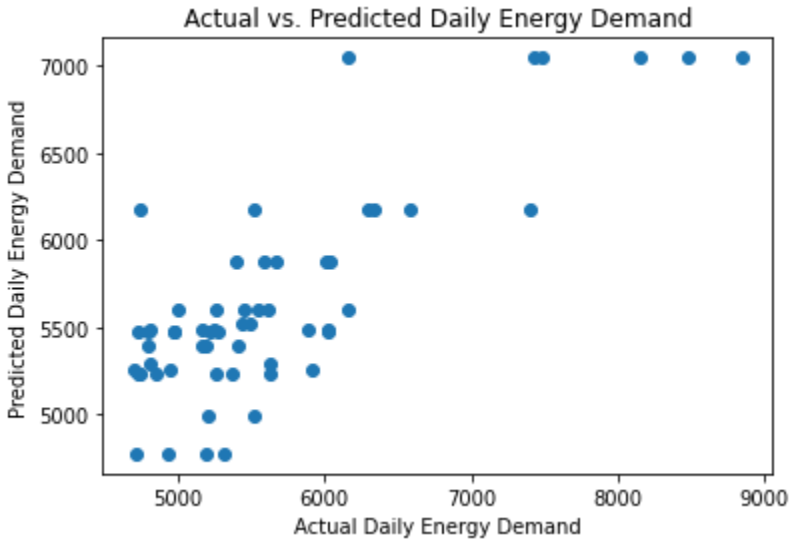
2.2.1 Linear Regression

For linear regression the data is split to train on ⅔ of the data. 'Mean_absolute_error', 'mean_squared_error' and 'R2 score' are used to evaluate the effectiveness of predicting demand. To test the model, it was run 100 times and averaged to get stable measurements



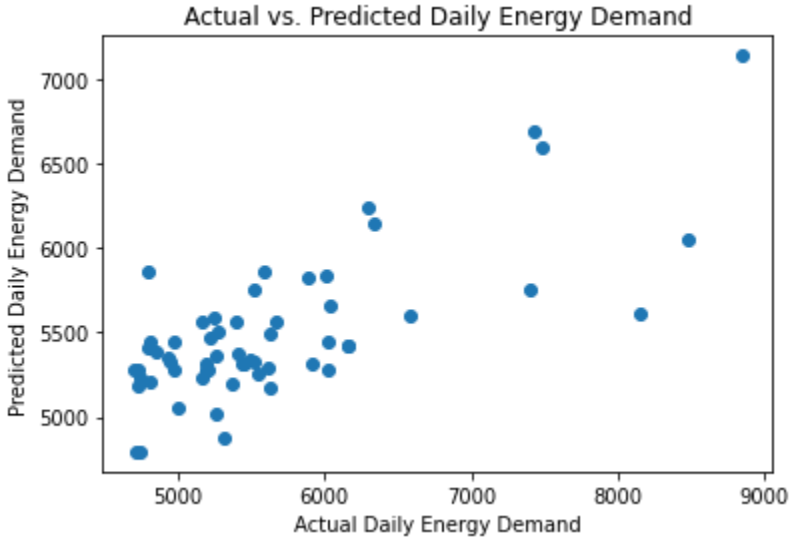
2.2.2 Decision Tree Regressor

A decision tree regressor using a similar method was tried.



2.2.3 kNeighborsRegressor

A kNeighbors regressor using a similar method was tried.



2.3 Classification Model

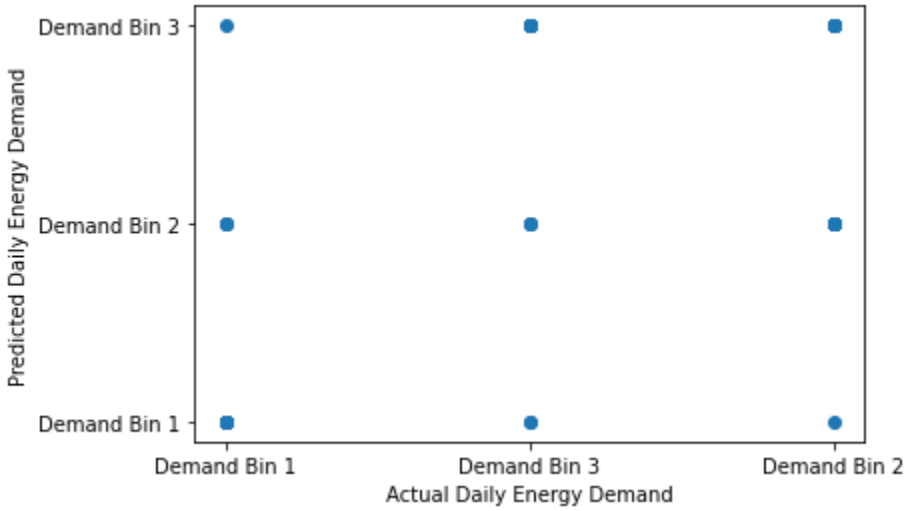
2.3.1 Bin data and feature selection

The data is divided into three bins, representing classifications of ‘low’, ‘medium’ and ‘high’ demand.

Before running the k-NN algorithm, data is scaled to ensure consistent magnitudes.

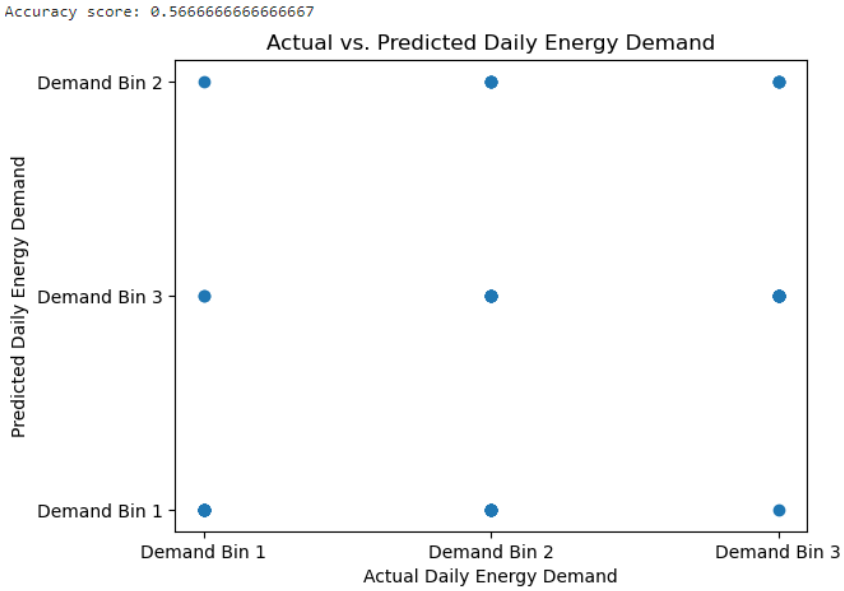
2.3.2 k-Nearest Neighbors

A k-NN model is run using mean imputation with 'fit-transform' and K-fold cross-validation. The hyperparameters ('n_splits', 'n_neighbors', and 'n_components') are manually tuned to optimise for a stable and high accuracy score. A repeat is applied for an average scoring.



2.3.3 Decision Tree Classifier

A model for decision tree classifier is run similarly to the k-NN method.



3. 1 Model Effectiveness

This section analyses each model, providing an analysis of their accuracy and efficiency scores.

3.1.1 Evaluation of Regression Models

Since the primary features are predominantly continuous values, our focus for assessment will be on regression models. Feature selection was carried out across six distinct groups, as detailed in Section 2.2 Machine Learning Analysis.

In the case of the linear regression model, features0 were applied to the merged_data.

3.1.1.1 Linear Regression Model

- Mean Absolute Error: 525.64

- Root Mean Squared Error: 673.38

- R2 score: 0.48

The results indicate that the linear regression model accounts for 48% of the observed variance in the target variable. Notably, the correlation between variables and the target is moderate.

3.1.1.2 Decision Tree Regressor

- Mean Absolute Error: 449.26

- Root Mean Squared Error: 577.60

- R2 score: 0.62

The R2 score of the Decision Tree Regressor suggests a better performance compared to the linear regression model in explaining the target variable. With a score of %62, there is above moderate linear relation.

3.1.1.3 K Neighbors Regressor

- Mean Absolute Error: 473.29

- Root Mean Squared Error: 698.55

- R2 score: 0.44

The K Neighbors Regressor demonstrates a similar correlation as the Linear Regression Model, explaining approximately 44% of the variance in the target variable.

3.1.2 Classification

For the classification models, as explained in Section 2.3, we divided the data into three distinct bins.

Both K Neighbors and Decision Tree models have been evaluated with kfold and own features selection respectively.

For all classification models, as explained in Section 2.3, 3 bins have been created. Kneighbor and Decision Tree models have been evaluated with both kfold and own features selection respectively.

Model	**Accuracy**
Kneighbor - Own Features Selection	%60.3
Kneighbor - KFold	%60
Decision Tree - Own Features Selection	%54.4
Decision Tree - KFold	%57.3

1. K-Nearest Neighbors (KNeighbor) - Own Features Selection:

Accuracy: 60.3%

This accuracy has an above average value and suggests that the model is performing moderately in classifying the data. There might be issues with the selected features or the hyperparameters of the K-Nearest Neighbors algorithm. Further investigation and tuning can be done to improve the model's performance.

2. K-Nearest Neighbors (KNeighbor) - KFold:

Accuracy: 60%

Similar to the previous case, this accuracy is also in average range. The use of K-Fold cross-validation suggests an attempt to mitigate overfitting, but the performance improvement is not observed. Again, further tuning and potentially exploring different algorithms might be necessary.

3. Decision Tree - Own Features Selection:

Accuracy: 54.4%

This accuracy is relatively low and indicates that the Decision Tree model, along with the selected features, is not capturing the underlying patterns in the data as efficient as the other models. Decision Trees can be sensitive to the quality of features and might easily overfit. Feature selection and hyperparameter tuning could potentially help improve the model's performance.

4. Decision Tree - KFold:

Accuracy: 57.3%

This model offers a slightly better accuracy comparing to decision tree with own selected features. Also, we can see a similar outcome with the KNN Model. The use of K-Fold cross-validation has provided better data sets and as an outcome, model has achieved a better result than own selected features.

In brief, considering the provided accuracy and R2 Score metrics, it appears that all the models have achieved above average to moderate outcomes and none of the models are managing to attain high

performance levels on the dataset. It seems that classification models may present a more viable approach for predicting the desired outcome variable ("Demand" in our specific context).

4. Weather Analysis and Insights: Exploring the Impact on Energy Consumption

Weather can have a significant impact on daily energy usage in various ways. Our comprehensive analysis reveals that specific weather features exert a certain influence on energy demand and consumption patterns.

Temperature

Warmer periods stimulate augmented energy expenditure due to the operation of air conditioning and cooling systems. Evidence of this is represented by the plot of demand by month in section 2.1.1. Additionally in Section 2.1.3, the maximal temperature emerges as the most influential parameter on demand, exhibiting a Pearson Correlation coefficient of 0.6 with energy demand. The severity of temperature fluctuations directly impacts the workload of these systems, resulting in heightened energy demand.

Humidity

As analyzed in Section 2.1.3, humidity appears as the second most related feature to the energy demand as per the heat map by Pearson Correlations. The relationship between humidity levels and energy consumption becomes apparent upon closer examination. Elevated humidity levels during summertime can intensify the perception of heat, thereby compelling air conditioning systems to operate with heightened intensity. It's worth mentioning that certain scenarios may necessitate the activation of dehumidification systems, further increasing energy consumption.

Seasonal Dynamics

Long-term weather patterns, such as seasonal changes, influence overall energy consumption. As different seasons bring different weather conditions, energy usage can vary throughout the year.

The data set which has been provided for this study, comprise data from November 2022 to April 2023. Heatmap shows medium-above average relation between the demand and price, and weak relation between the hour of day and demand. Although month and/or quarter was expected to be in high relation with the demand due to seasonal changes, no strong relation observed. This may be a result of not having a dataset covering the whole year cycle.

Conclusion

In summary, our comprehensive analysis between weather conditions and energy consumption yields valuable insights. Temperature emerges as a main determinant, with its fluctuations distinctly impacting energy demand. Humidity's subtler influence and the overarching effect of seasonal variations further underscore the intricate relationship between weather and energy utilization.

5. Limitations

The model developed analyses the relation between weather conditions and the dynamics of energy demand and pricing. There are internal, external factors and limitations affecting the accuracy and efficiency of the models.

5.1 Internal

Data reliability is the primary concern as whole modeling has been built on provided datasets. Possible weak points are listed as follows:

Accuracy:

The data used in this report is gathered by the BoM and AEMO. Their accuracy is relative to the quality of the devices gathering the information. AEMO shares a Forecast Accuracy Report yearly which details their data output and demand forecast, which is required as part of the National Electricity Rules. As part of this report, the indicators from Victoria performed as expected, except for energy consumption

in the year 2022, explained by the forecasting underlying business mass market consumption. Also, delays in commissioning of new installed generators led to more inaccuracies in the data obtained.

Completeness:

Both datasets were taken from original larger datasets. Therefore, the completeness is not good for our data, in terms of data points available.

Reliability:

The model does not consider an understanding of the data logging systems in place and how accurate they are. Further study on these systems would give more understanding of the reliability of the data.

Relevance:

As more renewable energy is deployed, weather becomes a key area to evaluate the demand of electricity as its fluctuation affects the necessity of more production by other sources.

Other data however is part of the demand forecast such as population, housings stock, electric vehicles, outages accidents among others. This data was not considered in this report.

Timeliness:

The data is useful for the period assessed but it is not time complete as it was mentioned earlier. However it is useful for this exercise of forecasting the energy demand in the Victoria area in the period studied.

Machine learning model

Further models are available which could provide more accuracy than the obtained and were not assessed as part of this report are another key limitation.

5.2 External

A complex web of factors thread through the fabric of global energy consumption and supply. In this section, we aim to assess several of these influential elements.

At the forefront of these considerations lies the geopolitical landscape, a paramount determinant of the global energy market. In the present era, economies hinge upon energy resources. Despite aspirations towards renewable energy sources, there is still a major reliance on fossil fuels. The consequence of this reliance is evident in the struggles among nations for control over vital energy

reservoirs. These geopolitical clashes, with their far-reaching impacts, cause substantial fluctuations in energy prices, distinct from the conventional impact of weather-related variables.

Technological and environmental limitations emerge as formidable challenges cast a shadow over the energy market. While there is a certain allure of renewable energy, its practical implementation struggles with issues of efficiency and consistency. Consider implementation of renewables to the grid which requires predictive analytics to ensure grid stability. Accuracy and effectiveness of such systems can influence local energy price and demand.

Additionally, an array of other factors influences both energy demand and pricing. Historical and socio-economic underpinnings, the march of industrialization, economic advancement, population growth rates, and more, all contribute to the complex structure of energy dynamics. The insights of this study reveal a moderate to low correlation between weather and energy demand, suggesting that the gaps in this correlation can be bridged by the other variables explained upon in this discourse.

References

https://www.synergy.net.au/Blog/2021/10/Everything-you-need-to-know-about-the-Duck-Curve

Australian Energy Market Operator, Forecast Accuracy Report (2022)

Australian Energy Regulator, (2022) State of the Energy Market

https://www.aer.gov.au/system/files/State%20of%20the%20energy%20market%202022%20-%20Full%20report.pdf