Data Analytics with Python – Assignment 2

Group 6 Members:
Ma Estela Arenas
Sean Howman
Yuxiao Liu
Feng Nie

1. Background

It has often been observed that energy consumption tends to be at its highest on days with hotter temperatures. In this project, our group will develop models that predict the maximum daily energy usage and pricing category based on provided weather data.

2. Purpose

These models can be used to predict likely energy demands based on a weather forecast, which can help energy companies understand plan for future usage, and help businesses plan when to conduct energy-intensive operations.

3. Process

3.1 Data Mining

Two datasets were provided in csv format, one for weather data with 243 rows and 21 columns, which has blanks and columns with float, integer and string data types. The other dataset contains the price-demand data with 11,664 rows and 4 columns. Demand usage given is within the 30-minute time interval daily. Date range used in this project is between 1st of January and 31st of August 2021.

3.2 Data Cleaning

What wrangling and aggregation methods have you applied? Why have you chosen these methods over other alternatives?

SQL, Python, Excel and OpenRefine were used to wrangle and aggregate the datasets. SQL, Excel and OpenRefine were chosen for ease of use in data cleaning while Python enabled us to apply our learnings from this course.

The price-demand dataset was modified in such a way that each day was represented by one instance (one row) with each day containing the maximum energy demand and the maximum price category for that day. The column containing the information on region was removed as superfluous and the date format was modified to match the date column in the weather dataset.

Using numeric and text facets in OpenRefine, the weather dataset was explored for blanks, outliers and non-numeric/non-text data. The data was found to contain several values which were missing or did not match the prevalent data type in the column.

Where the wind direction and wind speed data columns contained the wrong data type, “0” was substituted for “Calm”, and “Calm” for “0” in these columns respectively. The columns were then transformed to numeric or text data as necessary.

Missing rainfall and wind direction values were imputed by inferring similar values from related data on that day and surrounding days, while missing times of maximum wind gusts were filled using the average time of maximum wind gusts. This method was chosen as opposed to simple imputation or linear regression as it was thought values were likely to correspond with surrounding values rather than values from the larger feature dataset.

For the remaining missing values, scatterplot facets were generated in OpenRefine as a quick-look correlation with other features. A feature that correlated with the missing value feature was then selected and the two features plotted against each other in Excel with a linear relationship determined and the resulting equation used to impute the missing value for each instance. This method was chosen as simple imputation using the whole feature dataset would give a figure calculated across 8 months. An example in Figure 1 and Figure 2 below show a strong correlation between the minimum temperature and the 9am temperature and the equation which was used to impute a missing minimum temperature value. One row in particular (data from 08/07/2021) contained several missing values. Some of the imputations for these data were based on lower confidence correlations, however after discussing whether the row should be deleted from the dataset, it was decided that the imputations were robust enough not to greatly affect the model output.

After cleaning the data, the price-demand dataset and weather dataset were combined, with the date as the common feature.

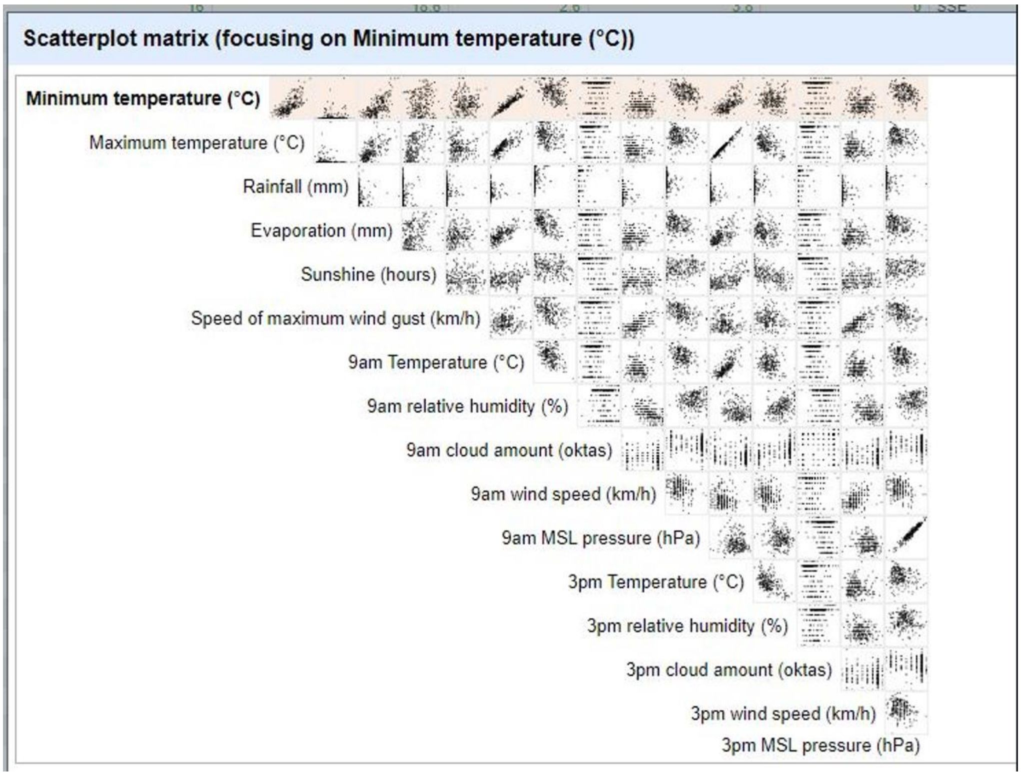


Figure 1: Scatterplot from Google OpenRefine showing correlations between features in the weather dataset. These plots used for data imputation and model feature selection.

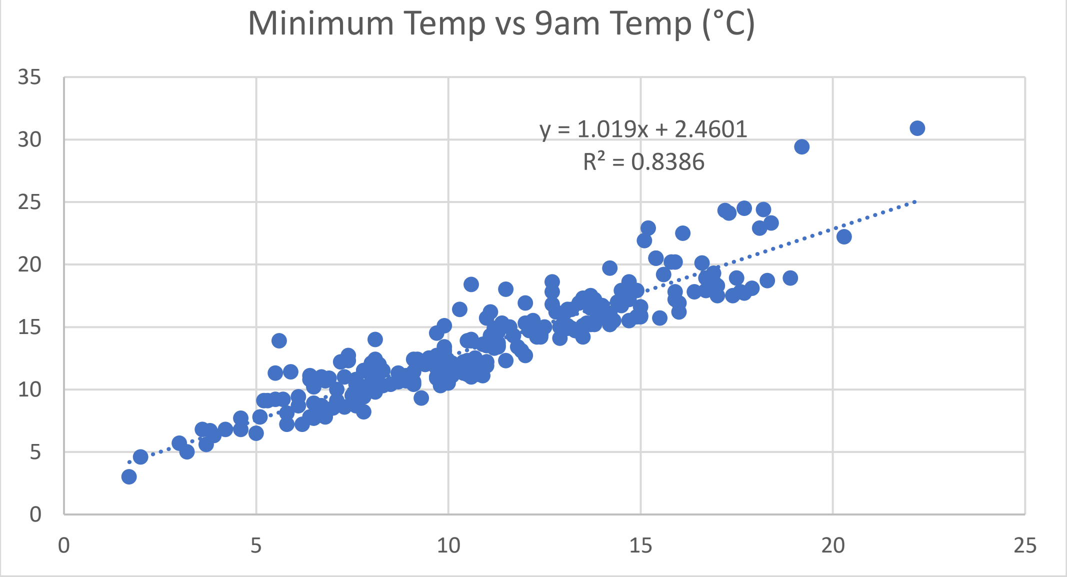


Figure 2: Excel plot showing linear relationship and equation used for imputation of missing minimum temperature value.

3.3 Data Exploration

We’ve selected features by observing scatterplots in OpenRefine, and where features appeared to be correlated with other features, one of those features was removed (usually the 9am or 3pm “snapshot” data). We’ve also removed wind direction features as they are discretized features and even if we change it to other data formats it will not add value to our model and by domain knowledge, we think they are unnecessary. We also chose minimum temperature as it works better than any other temperature features.

3.4 Model Building and Prediction

How have you gone about building your models and how do your models work? How effective are your models? How have you evaluated this?

3.4.1 Linear Regression Model

The goal of this model is to predict the maximum daily energy usage based on provided weather data. The output is numerical data thus we will be using linear regression to build our model.

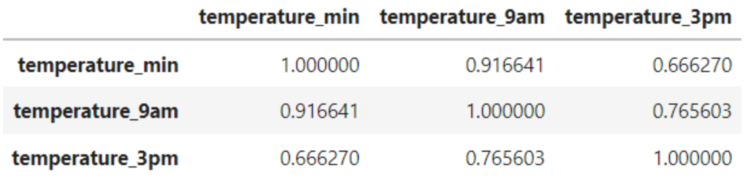
Our assumptions for using linear regression are:

- The dependent and independent variables are both numerical.
- There is a linear relationship between the dependent and independent variables.
- There are no significant outliers.
- There is independence of observations.
- The data shows homoscedasticity, which is where the variances along the line of best fit remain similar as you move along the line

In order to create the model using linear regression algorithm, we did the following:

1. Import required libraries.
2. Load the combined data set.
3. Select the features.

   We’ve used Pearson correlation coefficient to see which features are relevant to this model. We’ve checked the correlation between the highest demand and the numerical weather features. We’ve also reviewed the independence of the possible features. We’ve selected the features with an absolute Pearson correlation coefficient over 0.3, which are “temperature_min”, “temperature_9am”, and “temperature_3pm”. To check the independence of these three features, Pearson correlation analysis was conducted again, the results showed that all these three features were strongly correlated over 0.6.



	temperature_min	temperature_9am	temperature_3pm
temperature_min	1.000000	0.916641	0.666270
temperature_9am	0.916641	1.000000	0.765603
temperature_3pm	0.666270	0.765603	1.000000

Figure 3: Pearson correlation analysis results

   Finally, we’ve chosen the one with the highest correlation with the output, which is the “temperature_min”.

4. Separate the dataset into 90% train and 10% test parts.
5. Train the model and predict the result with test data.

   We’ve done k-fold cross validation with k value of 10 and we have observed that the average r2 score raised while k increasing, but it went down while k increased too big. Here we chose k-10 for experimental design.

6. Evaluate the result.

   Based on the results of our linear regression model, with an average r^2 score of 0.23, we can say that there is some linear relationship existing between the minimum temperature and the max demand usage. However, the linear relationship is not strong enough to use this model to predict maximum demand usage.

For one specific data-split model in K folder technique, the r^2 score is 42.99%. The test result followed the regression line.

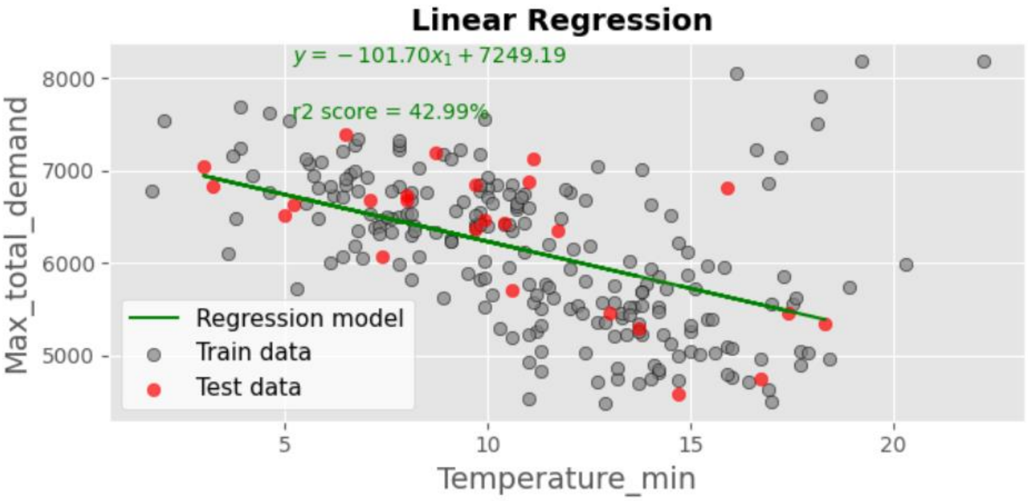


Figure 4: Linear Regression Plot for Max_total_demand vs Temperature_min

However, for the specific data split model as below, the r2 score is 17.54%. We thought that was caused by the tested result which was far away from the regression line when temperature was around 17 degrees.

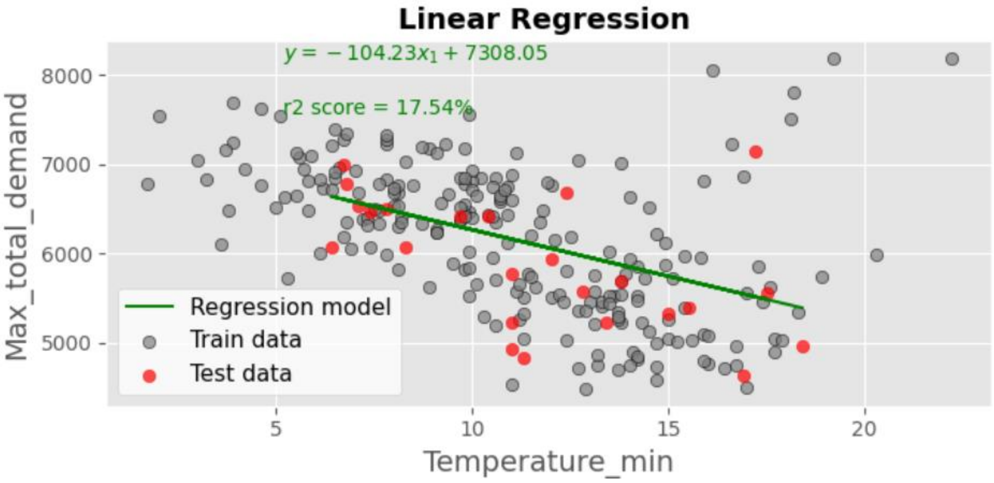


Figure 5: Linear Regression Plot for Max_total_demand vs Temperature_min

Also, from the original data set, some of the maximum demand usage points are not always going down while the minimum temperature increases. There were 2 trends that happened when minimum temperature over 15 degrees, some parts went down, another part went up.

This model could be improved by 2 possible ways:

1. Feature selection: Other features may also affect the maximum demand usage output, which were not selected during our feature selection process, or even other features which may not even in the original data source. Therefore, feature selection part can be enhanced and consider other features out of the original source as well.

2. Separate feature range: Since we observed that there were 2 different trends between minimum temperature and max demand usage when the minimum temperature was over 15 degrees. Therefore, current prediction model could be used to predict when the minimum temperature is under 15 degrees. And for the cases over 15 degrees, it is

possible to use the subset data to reconduct above process again to get another linear regression model.

3.4.2 Classification Model

The goal of this model is to predict the maximum price category based on provided weather data. The output is categorical data thus we will be using either Decision Tree or K nearest neighbor (KNN) classifier to build our model. Our group has decided to try both algorithms and see later which one of them will produce a better model.

Now that the data is totally prepared, the classifier is instantiated and the model is fit onto the data. The criterion chosen for this classifier is entropy. Once our model fits the data, we tried predicting values using the classifier model. This is often done in order to perform an unbiased evaluation and get the accuracy score of the model. We have done parameter tuning to select the best model.

3.4.2.1 Decision Tree

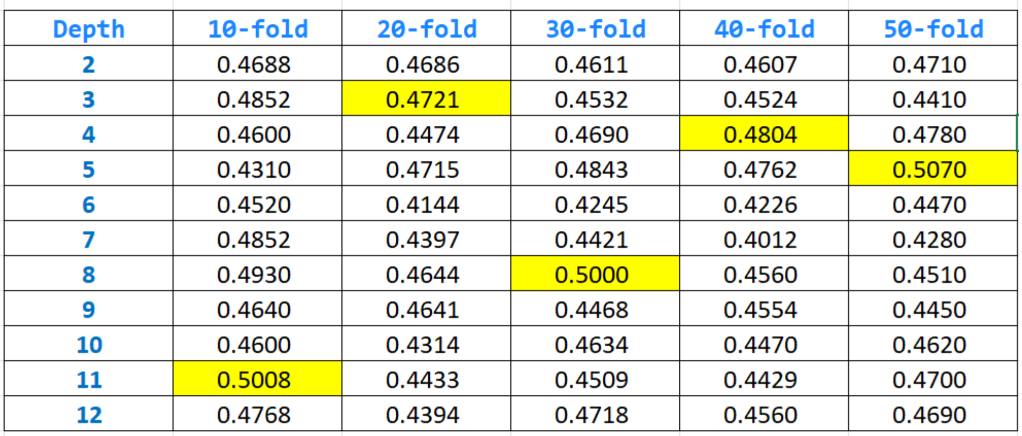
To create the model using the Decision Tree algorithm, we did the following:

1. Import the required libraries.
2. Load the combined data set.
3. Select the features.

   We have experimented by using chi2 and mutual information for feature selection but this resulted to a lower accuracy score. We've chosen minimum temperature, rainfall, sunshine and max wind speed to be the features that will predict the maximum price category. These features were selected because they improved our model's performance, they were correlated with the class label, they were dependent of the class label and they were not correlated with the other features.

4. Separate dataset using experimental design.
5. Train the model and predict the result with test data.

   We've done k-fold cross validation with a variety of k-folds and tree depths, with sample accuracy scores as shown below. We have observed that the k-folds and tree depths doesn't affect the average accuracy score. We have chosen the k value with the highest accuracy score. With that, we've used k-fold cross validation with 20 k-folds and maximum depth of 3.



Depth	10-fold	20-fold	30-fold	40-fold	50-fold
2	0.4688	0.4686	0.4611	0.4607	0.4710
3	0.4852	0.4721	0.4532	0.4524	0.4410
4	0.4600	0.4474	0.4690	0.4804	0.4780
5	0.4310	0.4715	0.4843	0.4762	0.5070
6	0.4520	0.4144	0.4245	0.4226	0.4470
7	0.4852	0.4397	0.4421	0.4012	0.4280
8	0.4930	0.4644	0.5000	0.4560	0.4510
9	0.4640	0.4641	0.4468	0.4554	0.4450
10	0.4600	0.4314	0.4634	0.4470	0.4620
11	0.5008	0.4433	0.4509	0.4429	0.4700
12	0.4768	0.4394	0.4718	0.4560	0.4690

Figure 6: Sample accuracy scores for a variety of k-folds and max depths variation in our Decision Tree prediction model

6. Evaluate the result.

   Based on the results of the Decision Tree algorithm, an accuracy score of approximately between 0.40 - 0.50 indicates that our model can somewhat predict the maximum price category using the selected features. Using different random states, tree depths, k-folds and train/test splits will not result to significantly different accuracy scores.

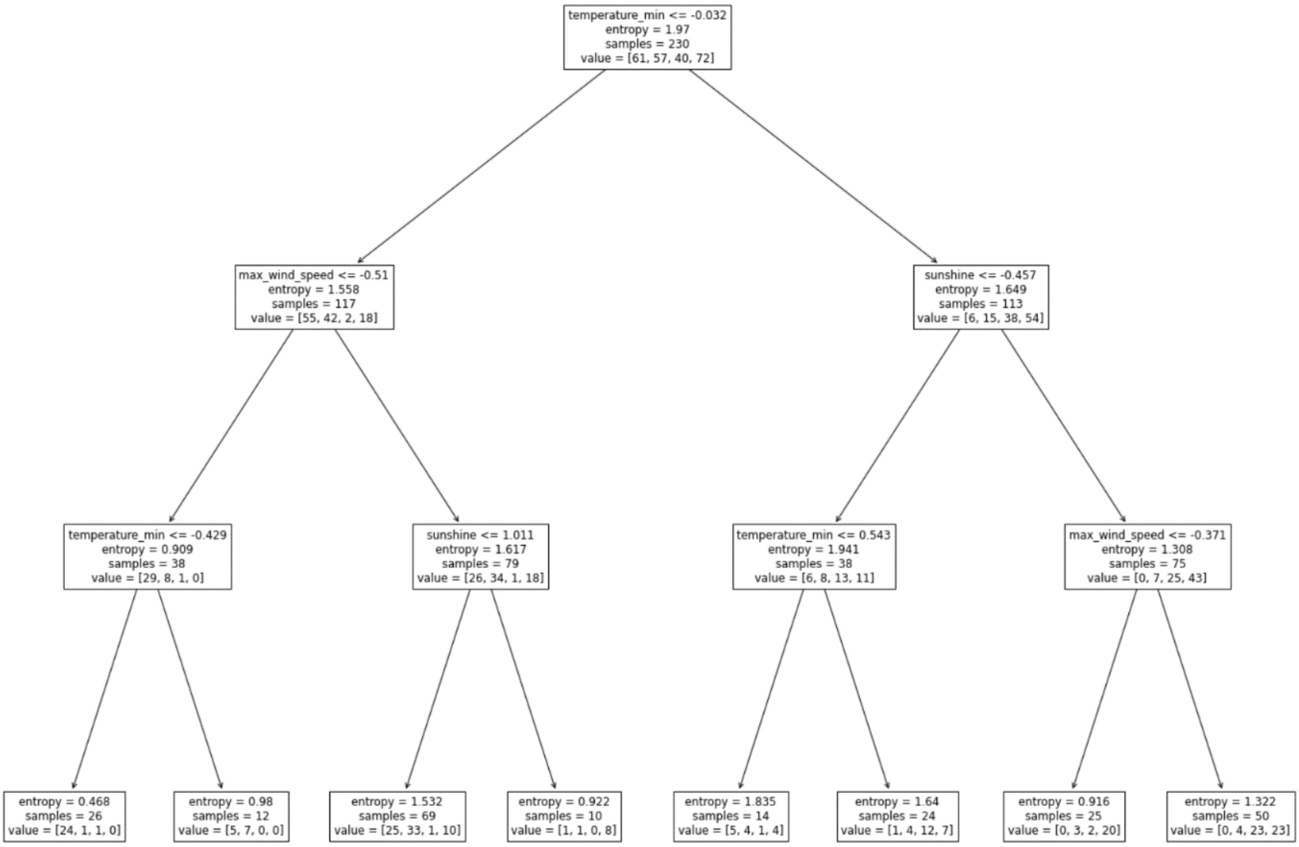
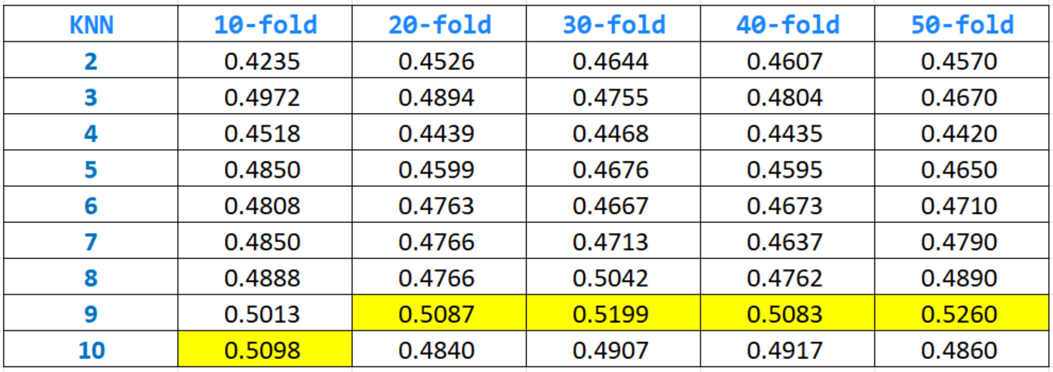


Figure 7: Decision Tree model with max depth of 3

3.4.2.2 K-nearest neighbour (KNN)

To create the model using the KNN algorithm, the steps will be the same as the Decision Tree algorithm.

We have done k-fold cross validation with a variety of k-folds, with accuracy scores as shown below. We think that if k was too small it's sensitive to noise points while if it's too big the neighbourhood may include points from other classes, we have observed that 10-fold cross validation with a k value of 9 produces the highest average accuracy score.



KNN	10-fold	20-fold	30-fold	40-fold	50-fold
2	0.4235	0.4526	0.4644	0.4607	0.4570
3	0.4972	0.4894	0.4755	0.4804	0.4670
4	0.4518	0.4439	0.4468	0.4435	0.4420
5	0.4850	0.4599	0.4676	0.4595	0.4650
6	0.4808	0.4763	0.4667	0.4673	0.4710
7	0.4850	0.4766	0.4713	0.4637	0.4790
8	0.4888	0.4766	0.5042	0.4762	0.4890
9	0.5013	0.5087	0.5199	0.5083	0.5260
10	0.5098	0.4840	0.4907	0.4917	0.4860

Figure 8: Sample accuracy scores for different k-folds and k-neighbours in our KNN prediction model

Based on the results of the KNN algorithm, an accuracy score of approximately between 0.42 - 0.52 indicates that our model can somewhat predict the maximum price category using the selected features, namely, minimum temperature, rainfall, sunshine, and max wind speed.

4. Discussion and Conclusion

What insights can you draw from your analysis? Which input variables are most valuable for predicting energy usage/price? Why are your results significant and valuable? What are the limitations of your results and how can the project be improved for future?

Given the results of our maximum demand usage prediction model, we can confirm that the underlying relationship existing between the minimum temperature and the demand is somewhat linear but not strong enough, especially when the minimum temperature is over 15^{o}C. The r^2 score for this model indicates that some but not all of the variation in the demand is explained by variation in the minimum temperature.

On the other hand, based on the results of our maximum price category prediction model, we therefore conclude that a change in the selected features will somewhat result in a change in the price category. Both Decision Tree and KNN produced similar results to validate this. Like the demand usage prediction model, it seems that the minimum temperature is the most valuable predictor of the price category. We've tried using chi2 and mutual information to validate our feature selection but this result can be explored better if given more time.

The results are significant insofar as they can help to predict the energy demand and price category for energy usage planning and costing. For a given weather forecast, we may be able to make some assumptions about energy demand and cost at a future date.

The models however cannot be used to predict outside of the range of weather data contained in the models. Also, the low accuracy scores may indicate that there are other factors beyond localized/regional weather data that affect energy demand and price, such as industry energy consumption or gas prices which fluctuate depending on supply constraints, transport prices and gas export demand.

The project could potentially be improved by using a larger dataset for modelling and perhaps incorporating other features outside of weather data that are found to correlate with energy demand and price. With larger datasets, other methods for data cleaning, feature selection, modelling and perhaps machine learning may become more practical. We could use Principle Component Analysis if we have a larger dataset.