

|  |
| --- |
| **Group Members:** |
| BONG LAU  FAN LI  GEORGE KOROVESIS |

|  |  |
| --- | --- |
| The University of Melbourne | |
| K1 PRELIMINARY DESIGN REPORT | |
| DATA0006 Data Analytics with Python |

|  |
| --- |
| The University of Melbourne  31/03/2022 |

Table of Contents

[1. Introduction 2](#_Toc99627200)

[2. Requirements and Getting Started 2](#_Toc99627201)

[2.1. Data Sets Provided 2](#_Toc99627202)

[2.2. Programming Environment 2](#_Toc99627203)

[3. Data Wrangling and Aggregation 3](#_Toc99627204)

[3.1 Data Aggregation - Obtain Daily Maximum 3](#_Toc99627205)

[3.2 Data Aggregation - Combine Energy and Weather Data 3](#_Toc99627206)

[3.3 Data Augmentation - Date Derived Variables 3](#_Toc99627207)

[3.4 Data Wrangling - Non-Numeric Variables 3](#_Toc99627208)

[3.5 Data Wrangling - Drop Null Data 3](#_Toc99627209)

[4. Model Development 3](#_Toc99627210)

[4.1. Feature Selection using Correlation 3](#_Toc99627211)

[4.2. Building the Model for Energy Demand - Linear Regression 4](#_Toc99627212)

[4.3. Building the Model for Price Category - Decision Tree Classification 5](#_Toc99627213)

[4.4. Building the Model for Price Category - KNN 6](#_Toc99627214)

[5. Further Improvements 6](#_Toc99627215)

[5.1. Feature Selection 6](#_Toc99627216)

[5.2 Energy Demand Model - Beyond Linear Regression 6](#_Toc99627217)

[6. Appendix A: GitHub 7](#_Toc99627218)

# Introduction

This project is to extract, explore and analyze the maximum daily energy use, pricing category and weather data. The objective of this project is to develop models which can predict the maximum daily energy use and pricing based on the weather data. Hopefully these models can help energy companies understand plan for future usage, and help businesses plan when to conduct energy-intensive operations.

**Date: 25/03/2022**

**Programming Environment**: Python 3.8 and Jupyter Notebook

# Requirements and Getting Started

## 2.1. Data Sets Provided

**2.1.1. Energy Demand Data (CSV format)**

* Date/time (in half-hour interval)
* Energy demand
* Price category (LOW, MEDIUM, HIGH, EXTREME)

**2.1.2. Weather Data (CSV format)**

* Date
* Minimum temperature (°C)
* Maximum temperature (°C)
* Rainfall (mm)
* Evaporation (mm)
* Sunshine (hours)
* Direction of maximum wind gust
* Speed of maximum wind gust (km/h)
* Time of maximum wind gust
* 9am Temperature (°C)
* 9am relative humidity (%)
* 9am cloud amount (oktas)
* 9am wind direction
* 9am wind speed (km/h)
* 9am MSL pressure (hPa)
* 3pm Temperature (°C)
* 3pm relative humidity (%)
* 3pm cloud amount (oktas)
* 3pm wind direction
* 3pm wind speed (km/h)
* 3pm MSL pressure (hPa)

## 2.2. Programming Environment

* Jupyter Notebook
* Python 3.8
* Libraries
  + import pandas as pd
  + from sklearn.metrics import normalized\_mutual\_info\_score
  + from sklearn.feature\_selection import mutual\_info\_classif as mi
  + from sklearn.tree import DecisionTreeClassifier
  + from sklearn import tree, neighbors, preprocessing
  + from sklearn.model\_selection import train\_test\_split, KFold
  + from sklearn import linear\_model
  + from sklearn.metrics import mean\_squared\_error, r2\_score, accuracy\_score
  + import matplotlib.pyplot as plt

# Data Wrangling and Aggregation

## 3.1 Data Aggregation - Obtain Daily Maximum

As the models are to predict the maximum half-hourly energy demand and pricing category for a day, we need to obtain the daily maximums from the given data set which contains half-hourly data. This is done by splitting the date/time into separate date and time, and use the Pandas 'groupby' method on date and Pandas 'agg' method to obtain the maximum data within a day. The pricing category is in string and thus an ordered category data type needs to be defined to calculate the maximum.

## 3.2 Data Aggregation - Combine Energy and Weather Data

With daily maximum energy data in place, an inner join of the energy data with the weather data (which is already daily) gives a single dataframe / CSV file of energy and weather data.

## 3.3 Data Augmentation - Date Derived Variables

It is stipulated that the models may depend on dates in some way. So these variables are added to the data set based on the date:

* Month
* Day of week
* Weekend

## 3.4 Data Wrangling - Non-Numeric Variables

As some algorithms (Pearson correlation, linear regression) to be used later would require numeric data, non-numeric variables would need to be converted to numeric.

**3.4.1 Wind Directions**

The wind directions ('N', 'NNE', 'NE' etc.) are converted to the their bearing values (0, 22.5, 45, etc.) Null wind direction is converted to -1.

**3.4.2 Calm Wind**

The word 'Calm' is used in the data and this is converted to 0.

**3.4.3 Time**

Time is converted to number of minutes.

## 3.5 Data Wrangling - Drop Null Data

Any rows with null data is dropped using the 'dropna' method.

# Model Development

Two models are to be developed:

1. Maximum half-hourly energy demand
2. Maximum pricing category

Energy demand is continuous data. So linear regression model would be used.

Pricing category is discrete data. So decision tree classification model or knn classification model can be used.

But before developing any models, we need to select the features/variables to be included in the models.

## 4.1. Feature Selection using Correlation

We select the features by examining their correlation with the target variable and picking the most correlated ones. Two correlation algorithms are used - Pearson correlation and mutual information.

We also compute the pair-wise correlation between the features so as to detect high correlation between certain features and only pick a representative one among them to avoid the 'curse of dimensionality'.

**4.1.1. Feature Selection - Pearson Correlation**

For the continuous dependent variable of energy demand, we apply Pearson correlation with all independent variables. The variables of top correlations are:

|  |  |
| --- | --- |
| Variable | Correlation |
| Month | 0.566494 |
| Minimum temperature (°C) | 0.482917 |
| 9am Temperature (°C) | 0.375774 |
| 9am wind direction | 0.326448 |
| 3pm Temperature (°C) | 0.307012 |
| Maximum temperature (°C) | 0.278702 |
| Weekday | 0.248948 |
| Evaporation (mm) | 0.240583 |

They are then examined with regard to the correlation between them (see the output of pairwise correlations). It is found that 'Minimum temperature' highly correlate with '9am Temperature' (0.91) while 'Maximum temperature' highly correlate with '3pm temperature' (0.97). It is determined to drop 'Minimum temperature' and 'Maximum temperature'.

**4.1.2. Feature Selection - Mutual Information**

For the discrete dependent variable of price category, we apply mutual information correlation with all independent variables. The variables of top correlations are:

|  |  |
| --- | --- |
| Variable | Correlation |
| Maximum temperature (°C) | 0.314163 |
| 3pm Temperature (°C) | 0.296882 |
| Minimum temperature (°C) | 0.293136 |
| Month | 0.282424 |
| 9am Temperature (°C) | 0.280045 |
| Evaporation (mm) | 0.255480 |

They are then examined with regard to the correlation between them (see the output of pairwise correlations). The correlations between them are not high (<=0.6) and thus they are all accepted.

## 4.2. Building the Model for Energy Demand - Linear Regression

Using the features selected above (Section 4.1.1), a linear regression model is built with 80% of data set as training data and 20% as testing data. The result is as follows:

Coefficients: 'Month': 201.1913, 'Weekday': -95.7112, 'Evaporation (mm)': 32.9096, '9am Temperature (°C)': -8.7438, '9am wind direction': -1.3716, '3pm Temperature (°C)': 12.859 Intercept: 5425.1093

**r2: 0.5785**

**4.2.1. Evaluation**

From the r2 value, we can tell that it is not an obvious linear regression relationship. Then we draw the scatter plot between the energy demand and the maximum temperature as follows:

**Figure 1**: The relationship between the maximum daily energy usage and the maximum temperature

Chart, scatter chart

Description automatically generated

From figure 1, we can tell that the energy demand is lowest between 20 ℃ and 25 ℃. This is reasonable because this is the most comfortable temperature range. People will need heater in colder days and air conditioner in hotter days, which result in an increase of the energy usage.

**Figure 2**: Residuals plotting (left) and their distribution histogram (right).

Chart, scatter chart

Description automatically generated

Figure 2 shows the residual distribution of the Train (blue) and Test (green) datasets. The Test points distribution around the horizontal axis is not well dispersed which indicates that the linear model is not very appropriate for our data.

## 4.3. Building the Model for Price Category - Decision Tree Classification

Using the features selected above (Section 4.1.2), a decision tree classification model is built with 80% of data set as training data and 20% as testing data, and a maximum depth of 3. The result is as follows:

**Accuracy: 0.5417**

**Figure 3**: Decision Tree Classification.

Graphical user interface, application, Teams

Description automatically generated

**4.3.1. Evaluation**

1. According to the decision tree for predicting price category, the top level node is testing against the month (Month <= 3.5). This implies that the **month** is the most valuable variable in predicting price - summer months (January to March) are usually of lower price, while the other months (April onwards, until August as available in the given data set) are usually of higher price.
2. To ascertain the true performance of the decision tree model, the model is run through the K-Fold cross-validation algorithm (with k = 10) . The result is **an average accuracy of 0.38.**

## 4.4. Building the Model for Price Category - KNN

Using the features selected above (Section 4.1.2), a KNN model is built with 80% of data set as training data and 20% as testing data, and k = 5. The result is as follows:

**Accuracy: 0.5**

**4.4.1. Evaluation**

To ascertain the true performance of the KNN model, the model is run through the K-Fold cross-validation algorithm (with k = 10) . The result is **an average accuracy of 0.45.**

# Further Improvements

## 5.1. Feature Selection

We have used correlation to guide our selection of features to be included in the models.

To seek further improvement in feature selection to potentially improve the models, we can use chi-squared test to test whether the class label feature/dependent variable is really dependent on each of the feature/dependent variables in the weather data set.

Furthermore, PCA can be used to synthesize a few variables from the given weather data that should produce models that perform better, though the downside is that the synthesized variables are not easily comprehensible.

## 5.2 Energy Demand Model - Beyond Linear Regression

**5.2.1 Non-linear Regression**

It is apparent that the relationship between energy demand and weather data (be it a single variable or combinations) is not linear. We need some algorithms of non-linear regression to more accurately model the relationship.

**5.2.2 Classification**

Meanwhile, we can discretize the energy demand and build classification models. The following two programs build the decision tree model and KNN model respectively.

The key finding is that the finer the discretization, the less accurate is the model. And this is true for both classification methods. This places a severe limitation on the usefulness of the models: if we want to have a more accurate prediction result, the range of energy demand is broader; we cannot give a more specific energy demand value while being accurate in the prediction.

Having said that, KNN model is better than decision tree in general for this dataset, with a higher accuracy score at the same level of discretization (number of bins).

Graphical user interface, application, Teams

Description automatically generated

# 6. Appendix A: GitHub

GitHub Link: <https://github.com/GeoKoro13/DATA0006_Assignment-2-.git>