**Data Analytics with Python – Assignment 2**

**Group 6 Members:**

Ma Estela Arenas

Sean Howard

Yuxiao Liu

Feng Nie

**Data Science Group Project**

**1.0 Business Understanding**

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. The hope is that 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.

**2.0 Data Mining**

Data sets provided were **weather\_data.csv** with 243 rows and 21 columns, has blanks and columns with both float and string, and **price\_demand\_data.csv** with 11,664 rows and 4 columns.

**3.0 Data Cleaning**

**3.1 Changes Made/Assumptions**

The following changes have been in the **price\_demand\_data.csv** file:

* The REGION column has been deleted as it only contains VIC1 in all the rows.
* The SETTLEMENTDATE column has been changed from DD/MM/YYYY<space>HH:MM format to MM-DD format.
* There were no changes made in TOTALDEMAND and PRICECATEGORY columns.

The following changes have been in the **weather\_data.csv** file:

* Edited Wind Speed columns, replaced “Calm” with zero.
* Edited blanks in Wind Speed Direction Columns where blanks related to zero wind speed; replaced blank with “CALM”
* Transformed necessary columns from value to numeric
* Transformed necessary columns from value to text
* Produced numeric facets and scatterplot facets for all numeric columns, to explore blanks, outliers and non-numeric data. Also, to highlight correlation of each feature with other features, in order to explore data and also to ascertain which features could be imputed using a simple linear relationship with other features.
* Impute “12.7” for missing value in row 189 for Minimum Temp, using y = 0.8503x + 8.6687 (from excel plot), where x is Min Temp, y is 3pm Temp.
* Impute “12.7” for missing value in row 189 for Maximum Temp, using y = 0.8966x + 0.6303 (from excel plot), where x is Max Temp, y is 3pm Temp.
* Impute numeric “0” (zero) in row 189 and 190 for Rainfall, by observing high sunshine/pressure, low cloud/humidity for that day and surrounding days had zero rainfall.
* Impute “N” and “NE” respectively in row 188 and 189 for Direction of Maximum Wind Gust, from observations of wind direction for those days and surrounding days.
* Impute “16” and “24” respectively in row 188 and 189 for Speed of Maximum Wind Gust, using y = 0.3886x - 0.1852 (from excel plot), where x is Max Wind, y is 3pm Wind.
* Impute “6.8” for missing value in row 189 for 9am Temp, using y = 0.8796x + 5.997 (from excel plot), where x is 9am Temp, y is 3pm Temp.
* Impute “70” for missing value in row 189 for 9am Humidity, using y = 0.3985x + 27.267 (from excel plot), where x is 9am Humidity, y is 3pm Humidity. Low Confidence.
* Impute “NE” in row 189 for Direction of Maximum Wind Gust, from observations of wind direction for this day and surrounding days.
* Impute “1” for missing value in row 189 for 9am Wind Speed, using y = 0.4741x + 8.4902 (from excel plot), where x is 9am Wind Speed, y is 3pm Wind Speed. Low Confidence.
* Impute “1021.5” for missing value in row 189 for 9am Pressure, using y = 0.9376x + 61.637 (from excel plot), where x is 9am Pressure, y is 3pm Pressure.
* Impute “1013.8” for missing value in row 243 for 9am Pressure, using y = 0.9376x + 61.637 (from excel plot), where x is 9am Pressure, y is 3pm Pressure.
* Impute “34” in row 16 for Speed of Maximum Wind Gust, using y = 0.3886x - 0.1852 (from excel plot), where x is Max Wind, y is 3pm Wind.
* Impute “4” in row 150 for 3pm Cloud Amount, using y = 0.3694x + 3.3874 (from excel plot), where x is 9am Cloud, y is 3pm Cloud. Low Confidence.
* Impute “1028.8” for missing value in row 150 for 3pm Pressure, using y = 0.9376x + 61.637 (from excel plot), where x is 9am Pressure, y is 3pm Pressure.
* Impute “W” in row 16 for Direction of Maximum Wind Gust, from observations of wind direction for this day and surrounding days.
* Impute “12:43” in rows 16, 189. 190 for Time of Maximum Wind Gust, from average of maximum wind gust times.
* Convert columns to text or numbers as necessary.

After cleaning the data, we have selected the daily maximum demand and price category from the **price\_demand\_data.csv**. Then we have combined it with the **weather\_data.csv** to form **1.Combined\_detail\_cleaned.csv**, with the date as the common feature.

**3.2 Limitations**

Date range used in this project is between 1st of January and 31st of August 2021. Demand usage is within the 30-minute time interval daily.

**4.0 Data Exploration** *– form hypothesis about your defined problem by visually analyzing the data*

**5.0 Feature Engineering** *– select important features and construct more meaningful ones using the raw data that you have. What insights can you draw from your analysis? For example, which input variables are most valuable for predicting energy usage/price?*

**6.0 Predictive Modelling** *– train machine for learning models, evaluate their performance, and use them to make 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?*

**6.1 Model 1: Maximum Demand Prediction Model**

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

Our assumptions for using linear regression are:

* The dependent variable is numerical.
* The independent variable is 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, we did the following:

1. Import required libraries
2. Load the combined data set.
3. Check the correlation between the highest demand with the numerical weather features, pick the Pearson correlation coefficient over 0.1 or lower than -0.1 ones.
4. Separate dataset into 70% train and 30% test parts.
5. Train the model and predict the result with test data.
6. Evaluate the result.

**6.2 Model 2: Maximum Price Category Prediction Model**

The goal of this model is to predict the maximum price category based on provided weather data. The output is expected to be a categorical data thus we will be using **decision tree** to build our model.

In order to create the model, we did the following:

1. Import required libraries
2. Load the combined data set.
3. Features were selected by observing scatterplots in OpenRefine, removing features that appear to be correlated with each other. Also removed wind direction features as unnecessary. We tried using max\_temp as feature, but min\_temp seems to work better.
4. Separate dataset into 70% train and 30% test parts.
5. Train the model and predict the result with test data.
6. Evaluate the result.

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 try 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.

**7.0 Data Visualisation** *– communicate the findings with key stakeholders using plots and interactive visualisations. Why are your results significant and valuable? What are the limitations of your results and how can the project be improved for future?*

**7.1 Model 1: Maximum Demand Prediction Model**

Given the results of our maximum demand prediction model, the underlying relationship between X and Y is linear.

Linear regression most often uses mean-square error (MSE) to calculate the error of the model. MSE is calculated by:

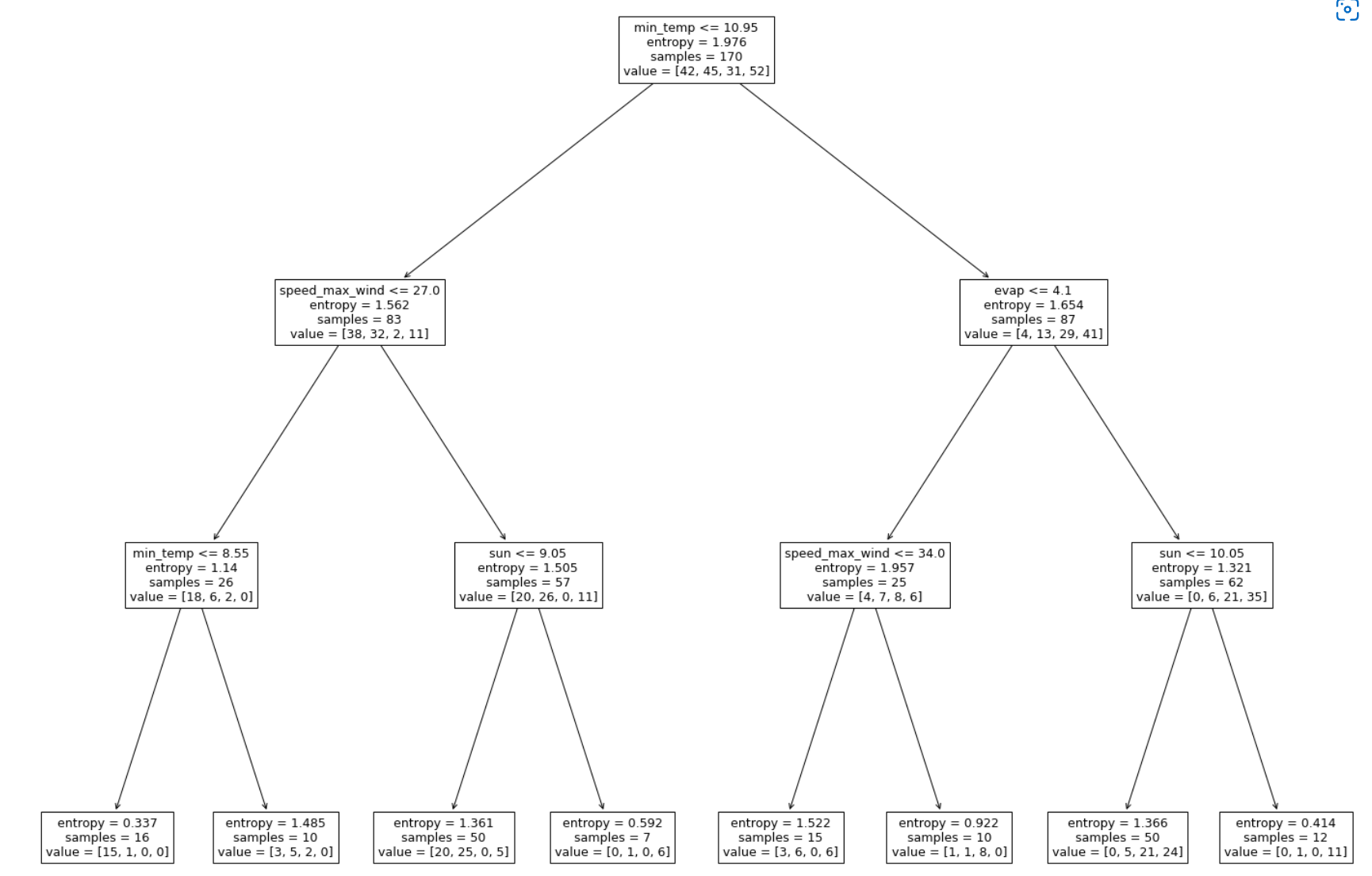
* measuring the distance of the observed y-values from the predicted y-values at each value of x;
* squaring each of these distances;
* calculating the mean of each of the squared distances.

Linear regression fits a line to the data by finding the regression coefficient that results in the smallest MSE. The mean squared error (MSE) tells us how close a regression line is to a set of points. The lower the MSE, the better the forecast. The smaller the MSE, the closer you are to finding the line of best fit.

R-Squared (R² or the coefficient of determination) is a statistical measure in a regression model that determines the proportion of variance in the dependent variable that can be explained by the independent variable. In other words, r-squared shows how well the data fit the regression model (the goodness of fit).

**7.2 Model 2: Maximum Price Category Prediction Model**

Based on the results of our maximum price category prediction model….



**7.3 Limitations of the Results**

**7.4 Further Improvements**