Project 2 Report
Predictive Models for Energy Demand and Pricing

By Maryam Shayan, Priyanka Garg, Peiling Xian and Vinh Nguyen
https://github.com/SebrinaX/Group8_Assignment2

Aim of Project: To develop two models that predict the maximum daily energy use and pricing based on weather data. So that these models can be used to predict energy demands based on a weather forecast, which can help energy companies understand plans for future usage, and help businesses plan when to conduct energy-intensive operations.

Question 1

1.1 What wrangling and aggregation methods have you applied?

Data Wrangling: It is an important step as it is the process of collecting and transforming raw data into better format for understanding and helps in decision-making.

For this project, we used:

Step 1: Data Exploration - we used the following methods to read and explore the data:

- pd.read_csv('price_demand_data.csv') to read the price_demand_data file
- pd.read_csv('weather_data.csv') to read the weather_data file
- df.head() to see what columns/fields there are in each data files

Step 2: Reshaping the data - In this we merged both the mentioned files into one file.

- df.apply (lambda x: x.split()[0]) to get the date dd/mm/yyyy from 'SETTLEMENTDATE' column in the price_demand_data file so we can join it with the weather_data file
- df.merge method to join the two above mentioned files
- df.groupby and .max() methods to aggregate data by date
- df.reset_index() to make the 'date' as an attribute
- .map to convert object type values to int value

Step 3: Dealing with missing values: In this process, we check if any null values are in data and fill those null values with the mean of the column.

- df.fillna and .mean() methods to replace NaT value with mean value
- df.isnull() to double check that there is no NaT value in the columns used for analysis

1.2 Why have you chosen these methods over other alternatives?

To be honest, we did not know other methods. Due to time constraints, we could not do research and try other methods. We will definitely explore alternatives outside of this course.

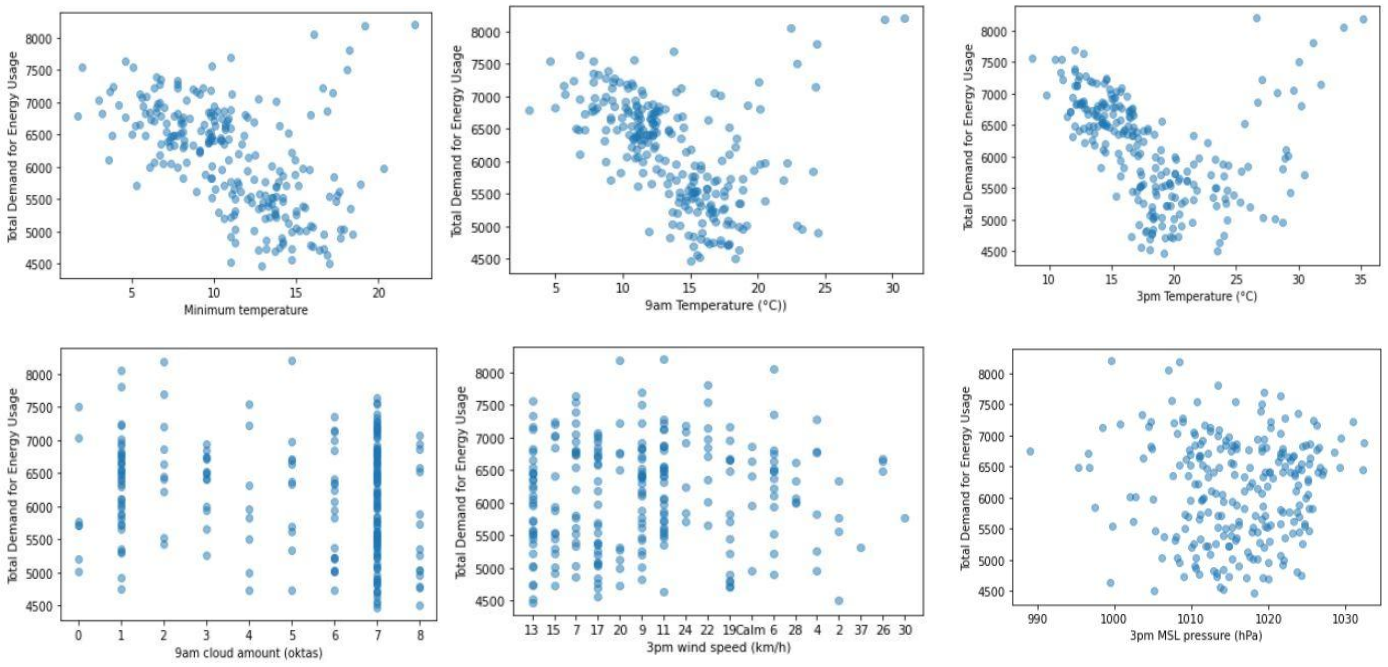
Question 2

How have you gone about building your models and how do your models work?

Model 1: Predicting the maximum daily energy usage based on the provided weather data (the energy spot price)

- **Step 1: Selection of Predictive Model -** (Based on the target variable) In this our target variable is continuous so we used Regression Model.
- **Step 2: Select attributes/predictors-** (Based on the correlation between target variable and other independent variable). In this model, the target variable is the daily energy usage (Y). We plotted the relationship between the target variable and each of the attributes to find out which ones have strongest correlation with the former.

  As shown below, Minimum temperature, Maximum temperature, 9am Temperature, and 3pm Temperature attributes seem to have a linear relationship with the Total Demand for Energy Usage, whereas others don't.



Graph 1: Relationship between each feature and daily total energy consumption

We also used **Pearson Correlation Coefficient** to find the correlation with Total Demand and other attributes and direction of their relationship.

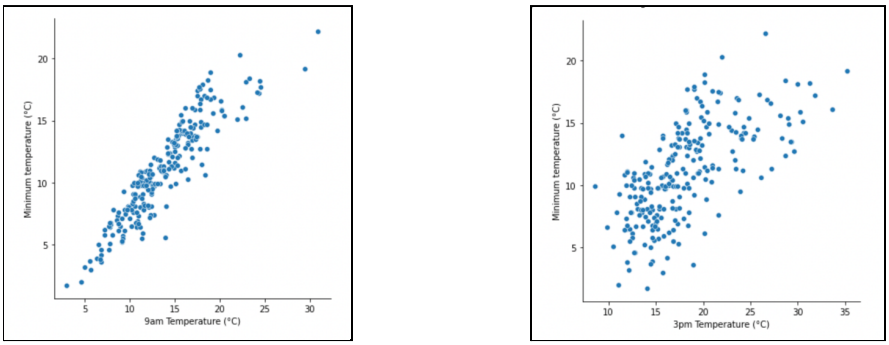
	TOTALDEMAND	Minimum temperature (°C)	Maximum temperature (°C)	Rainfall (mm)	Evaporation (mm)	Sunshine (hours)	Speed of maximum wind gust (km/h)	9am Temperature (°C)	9am relative humidity (%)	9am cloud amount (oktas)	9am MSL pressure (hPa)	3pm Temperature (°C)	3pm relative humidity (%)	3pm cloud amount (oktas)	3pm MSL pressure (hPa)
TOTALDEMAND	1.000000	-0.488245	-0.290005	-0.072715	-0.264008	-0.139581	0.081024	-0.390844	0.103267	-0.167373	0.051993	-0.325252	0.064301	0.070940	-0.005067

Based on the above values, we concluded that:

Table 1: Correlation

Attribute Name	Pearson Coefficient	Direction of Relationship
Minimum Temperature	-0.488245	Negative Medium Correlation
Maximum Temperature	-0.290005	Negative Small Correlation
9am Temperature	-0.390844	Negative Medium Correlation
9am Relative Humidity	0.103267	Positive Small Correlation
3pm Temperature	-0.325252	Negative Medium Correlation

We also explored the relationship between the variables to establish if they were independent from each other. As you can see in the graph below, they seemed to have a strong correlation between them, which violated the independence rule of the regression model. However, when we removed any of these variables, the value of R^2 decreased. So we kept these variables in the model.



Graph 2: Relationship between independent variables

Finally, we selected Minimum temperature, 9am Temperature, and 3pm Temperature as the predictors of the daily energy usage because it looks like they have a linear relationship with the target variable, their Pearson coefficient values are also high as compared to other values and are not highly correlated to each other.

- **Step 3 : Data segregation into Training and Test subsets -** by using train-test split technique. We split the aggregate and cleaned dataset into 2 groups: train subset and test subset with the training subset used to build the model and the test subset used to validate it.

- **Step 4 : Apply the Regression Model and check efficiency -** We fed the train subset data to the linear regression model. Then we compared the sum of squared difference between actual values and the predicted values (MSE), and the proportion of the variance in the target variable which was explained by the linear regression model (r^2 scores), which was generated from different combinations of predictor features. Below is the summary of the comparison:

Table 2: Feature combination with MSE and R^2 value

Feature combinations	MSE	R^2
Minimum temperature (°C) 9am Temperature (°C) 3pm Temperature (°C)	565,697.44	0.209
Minimum temperature (°C) Maximum temperature (°C)	596,100.16	0.166
Minimum temperature (°C)	598,290.66	0.162
Minimum temperature (°C) 3pm Temperature (°C)	601,137.92	0.159
Maximum temperature (°C) Speed of maximum wind gust (km/h)	844,000.83	-0.096
Maximum temperature (°C)	581,910.39	0.089

As shown above, the first combination of features showed the lowest MSE and the highest r^2 score. So we selected these features for our prediction model.

- **Step 5: Find Intercept and the Slope Coefficient of the regression equation.**

$$Yi = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$$

In our model, β0 (Intercept) = 7,286. It is the estimated average value of total Energy usage when the value of 'Minimum temperature', 'Temperature at 9am' and 'Temp at 3pm' are zero. Which means without the temperature's influence, the daily basic energy consumption is 7286 MW, which is the amount of energy needed for this purpose in a defined period of time, namely, the basal metabolic rate (BMR). and β1 = -165.338, β2 = 76.98, β3 = -21.83 are the three coefficients with regards to 'Minimum temperature', 'Temperature at 9am' and 'Temp at 3pm' respectively. These slope coefficients β1, β2 and β3 measure the estimated change in the average value of the Maximum Daily Energy Usage as a result of one unit change in each of the variables (i.e. 'Minimum temperature', 'Temperature at 9am' and 'Temperature at 3pm'), keeping the other two variables held constant.

- **Step 6: Construct the Regression Equation** - To predict the maximum daily energy usage

  Max_Energy_Usage = 7,286 - 165.338*MinTemp + 76.98*Temp_at_9am -21.83*Temp_at_3pm

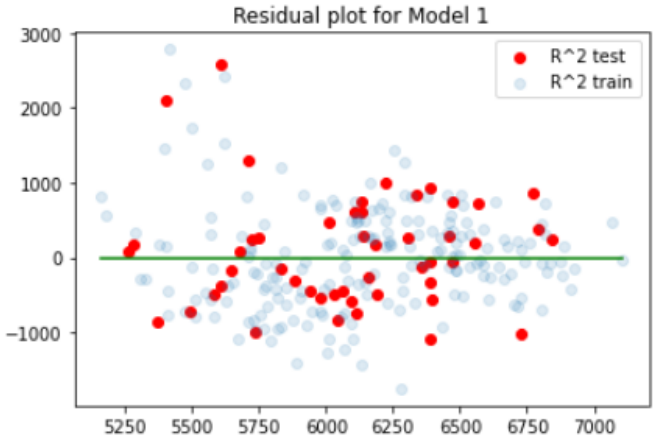
- **Step 7: Compare the Predicted Data with the Actual Data** - Using the above equation, we predicted the maximum energy usage and compared it with the given actual data to see how the model works.

Date	Predicted_Usage	Actual_Usage
1/01/2021	5571.4607	5019.64
1/02/2021	5948.0278	5228.29
1/03/2021	5852.2511	5225.37
1/04/2021	5717.9433	5807.02
1/05/2021	6053.0514	5261.09
...	...	...
9/04/2021	5831.7791	5688.63
9/05/2021	6096.3205	5222.89
9/06/2021	6474.6221	7224.91
9/07/2021	7069.3570	7536.11
9/08/2021	6483.0607	6675.69

243 rows × 2 columns

Graph 3: Predicted Energy Usage V.S. Actual Energy Usage

- **Step 8: Perform Residuals Analysis**. We plotted the residuals of the train subset and the test subset.



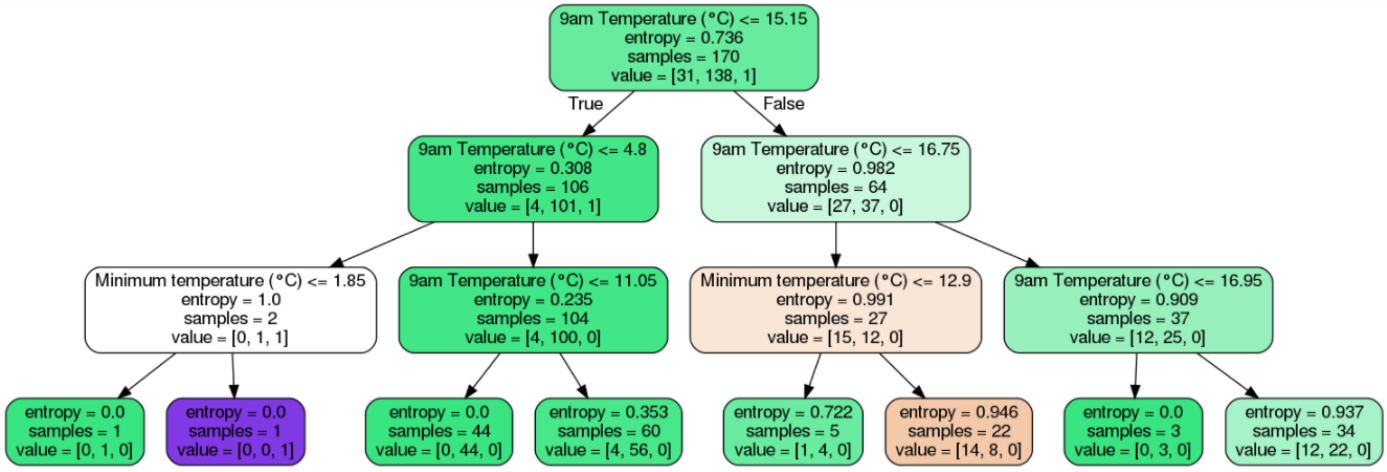
Graph 4: Residual Analysis between Train and Test Subset

- **Step 9:** Check the Assumptions of Regression by examining the residuals. Our observations are as follows:
    - The differences between the actual values and estimated values are 0.
    - The residuals doesn't fit a line or a pattern, which indicate that there is no bias or trend
    - There seems to be no strong correlation in the residuals.

Model 2: Predicting the Maximum Daily Price Category Based on the Provided Weather Data (Category of Usage)

- **Step 1: Model Selection -** (Based on the target variable) In this our target variable is finite so we used Classification technique and use Decision Tree Model.

- **Step 2: Examine the Target Variable:** The values of the daily price category are “low”, “medium” and “high”. We converted them into numerical values for ease of training of the model.
- **Step 3: Data Cleaning and Selection of Predictor Variables-** We removed all the variables that have string-type and null values, e.g. direction of maximum wind gust, and used the rest for our model.
- **Step 4: Data Segregation into Training and Test Subsets -** by using train-test split technique. We used the same approach as explained in Model 1
- **Step 5: Create the Decision Tree.** We tried 2-way split, multi-way split, max depth of between 3 and 6, different random state numbers and different combinations of features. The highest accuracy score we could get was with 2-way split i.e. 0.835, which means 83.5% of the predictions are accurate.



Graph 3: Decision tree hierarchy structure

Question 3

How effective are your models? How have you evaluated this?

For Model 1, we experimented with different variables, and variable combinations as training data input to feed the model, our final model result has the highest efficiency compared to our previous ones. To evaluate our model we used MSE values and R^2 score. And the final model we have, has the smallest MSE value, which is 565,697.44. The R^2 Value is 0.209, which means 20.9% of the variation in Daily Total Energy Demand is explained by variation in ‘Minimum temperature (°C)’, ‘9am Temperature (°C)’, and ‘3pm Temperature (°C)’.

As we were taught the R^2 score should be high but in our model it is very low and MSE seems to be very high as well. So we think Model 1 is not so effective**.**

Whereas, Model 2 seems to be more effective. To evaluate we used an accuracy score which is 0.835, which means 83.5% of the predictions by our model are accurate. This is a very good score.

Question 4

What insights can you draw from your analysis? For example, which input variables are most valuable for predicting energy usage/price?

For model one: we use linear regression method to predict the maximum daily energy usage. We got feature ranking by running correlation functions for each feature. From the correlation chart we got 'Minimum temperature (°C)', '9am Temperature (°C)'and '3pm Temperature (°C)' has larger negative correlation values with "Total Energy Demand", figures as -0.488, -0.39 and -0.325 respectively. These figures show there might be a moderate negative linear relationship in between these features with "Total Energy Demand". Based on the intercept and coefficient values that we calculated from our model, we built a function:

$$\text{Max Energy Usage} = 7286.33 - 165.34 * \text{MinTemp} + 76.98 * \text{Temp}_{\text{at}_{9\text{am}}} - 21.83 * \text{Temp}_{\text{at}_{3\text{pm}}}$$

Intercept value 7286.33 MW means when those three features are zero, the basic daily energy usage is 7286.33 MW. Coefficient values means one unit change of those features correspond to 'Total Energy Demand' change.

In our model, the most valuables for predicting energy usage and price category are the Minimum Temperature, 9am Temperature and 3pm Temperature. We felt they may have a strong correlation between them, which violated the regression assumptions. However, when we removed one of the variables, the r2 score in Model 1 and accuracy score in Model 2 were lower. But because they also had strong linear correlation with the target variable, we left them in the models.

Variables like Rainfall, Direction of Wind, Speed of Wind Gust, etc. did not have much impact in the model.

For model two: Because we use the Classification method - decision tree model to predict the price, the important result/value after we run the model are: Decision tree shows us the hierarchy importance of each feature we send through to the model. Here in our model, the most important feature is: Total Demand > Evaporation > Minimum Temperature and so on.

By analyzing the Information gain of this decision tree: At root node: H(TOTALDEMAND) - entropy of the class label ‘TOTALDEMAND’, we can read directly from the graph, which is 1.969 at the root node. Sample = 194 means there were 194 instances of training data which have been put to the model. Where (194/243)100% = 80%, that is exactly 80% of the dataset's data as training data. where 243-194 = 49, means 49 instances as testing data, and (49/243)100% = 20%. Value[33,60,52,49] means there are four groups which refer to the four price categories. In these training data, there are 33 in Low price category, 60 in Median, 52 in High, and 49 in Extrem. When adding these four group's number is 33+60+52+49 = 194. We can calculate the Information Gain(IG) from this graph: when 5354.67 < TOTALDEMAND <= 6017.23, we need add feature Evaporation(mm) <= 6.5 to split down the tree. and the IG of these two levels is IG = 1.969-1.529 = 0.44. which is ok.

According to our research, for the Decision Tree Model the highest accuracy was achieved when we split the training data and test data in the ratio of 70% and 30%. And by doing that the entropy at the majority leaf nodes reached almost to zero.

Question 5

Why are your results significant and valuable?

We think the two models are valuable in helping energy companies to understand plans for future usage, and help businesses plan when to conduct energy-intensive operations. They also might be useful for the government and organizations promoting sustainability to formulate suitable policies to encourage businesses and households to reduce energy costs.

Question 6

What are the limitations of your results and how can the project be improved for future?

Model 1 is not very effective. The difference between the predicted values and actual values seems to be too big. As we searched the internet, we found several methods to transform the data to improve the effectiveness of the model. But we were worried that it would go beyond the requirements of this assignment. In real world situations, we would want to conduct desktop research and consult different stakeholders to get their insights into the variables that have strong impacts on the demand for energy usage. From there, we might need to add more variables to the analysis.

Model 2 seems to be more effective based on the accuracy score. However, the predicted price categories based on the test subset doesn't look right as they included only 'low' or 'medium' or only 'medium' labels. We suspected that there was an overfitting issue here. So we tried to change the combination of variables, train/test subset ratio, random state but the outcomes were almost the same. Due to time constraints and limited mention in the lectures and workshops of practical solutions to these types of issues, we were unable to confirm if there was a problem and if it was, what problem was that. Also, we did not know how to fix the problem so that our model is technically sound.

Both data sets were small, having less than a year of records. A project that has access to more comprehensive data may offer more conclusive results.

Future research can improve the prediction accuracy of our existing model. There are various other methods that can be used like Normalisation, K-fold Cross Validation, AIC and many more.