Energy Demand Forecasting Based on Weather Data

Course: Data Analytics with Python

Prepared by: Xiaohan (Brian) Yang

Disclaimer

This work was submitted as a course assignment for DATA0006 - Data Analytics with Python at The University of Melbourne (April 2023),
and is revised and published as a portfolio item of the author.
Please kindly do not misuse this work for any other purposes.

Table of Contents

INTRODUCTION

Australia is renowned for its unpredictable often extreme weather conditions, which have a significant impact on the energy consumption patterns of its inhabitants. Melbourne, a city that is famous for its erratic weather patterns, is no exception to this phenomenon. This report presents a comprehensive analysis of the relationship between daily weather conditions and energy usage in Melbourne. Our objective is to develop a model that can accurately predict the maximum daily energy usage based on weather data. The goal of this model is to aid energy companies in planning and managing future energy usage effectively.

To achieve this objective, our team of data analysts will perform extensive data cleaning on two provided datasets: weather data and energy price demand data. The analysis will also include a brief research component, which will evaluate the current weather and energy price data in Melbourne and how it can be leveraged to improve this project's accuracy. The report will also discuss the effectiveness and limitations of the model and provide valuable insights into the relationship between weather and energy usage. Ultimately, this analysis aims to contribute to more effective and sustainable energy management strategies in Melbourne.

DATA CLEANING / PREPROCESSING

Data cleaning is critical step in ensuring the accuracy and reliability of data analysis. In this study, we utilised several methods to clean the data, including converting data/time columns to datetime format, converting categorical values to numerical values, dropping non-required columns and rows, grouping data by date and aggregating values, feature engineering by adding a variable to identify weekdays vs. Weekends, and filling in missing values using interpolation and forward filling.

To ensure consistency and compatibility across datasets, we converted date/time columns to datetime format using the 'pd.to_datetime' method. This enables us to easily manipulate and visualise the data by grouping data by date or plotting data over time.

To perform mathematical operations or analysis on the data, it is essential to convert categorical values to numerical values. We converted the 'PRICECATEGORY' and wind direction values to integers using predefined dictionaries mapping each category to a numerical value. Additionally, we also applied the same method to convert the 'Calm' value to a numerical value of zero. This step ensured that the data was compatible with mathematical operations and allowed for easier analysis and modelling.

Simplifying the data by dropping non-required columns and rows is a useful method to remove unnecessary information that may hinder analysis or visualisation. Therefore, we dropped the 'REGION' and 'TOTALDEMAND' columns, as well as the last row of energy price demand data, which contains data for only one date.

Grouping data by date and aggregating values is a helpful method to obtain a summary of the data at daily level. We grouped the data by date and calculated the 'DAILY ENERGY USAGE' and then mean of 'ENERGY PRICE' for each date. Please find the scatter plot of daily energy usage vs date in the figure 1 below, noting the size of each point is proportional to the energy pricing multiplied by 50.

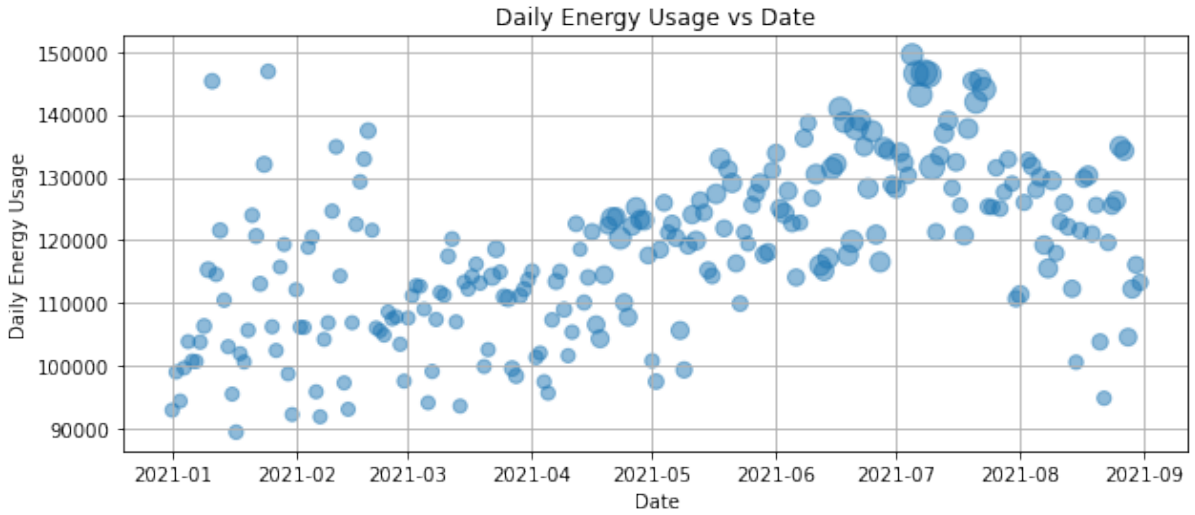

Figure 1: Scatter plot of daily energy usage vs date

Feature engineering involves creating new variables or transforming existing ones to provide additional information that may be useful in analysing the data. In this study, we added a new variable 'WEEKEND' to the dataset to distinguish between weekday (0) and weekends (1). This variable allows for more refined analysis of patterns in the data and can help identify energy demand differences in behaviour between weekdays and weekends.

Filling in missing values using interpolation [1] and forward filling can help ensure the completeness of the dataset and avoid errors in analysis or modelling. We filled in missing values in the wind direction and wind speed columns using linear interpolation and forward filling.

Overall, the chosen data cleaning methods are appropriate for the given dataset and allow for easier analysis and visualisation. Alternative methods could have been used, such as filling missing values using different techniques or dropping columns based on other criteria, but the chosen methods are appropriate for the given dataset and help ensure the quality of the data.

CORRELATION ANALYSIS

After completing Data Cleaning and Pre-processing, the next step is to conduct Correlation Analysis. This determines the strength and direction of the connection between Daily Energy Usage and various weather features like Minimum Temperature, Maximum Temperature, Rainfall, Evaporation, Sunshine Hours, Direction of Maximum Wind Gust, Speed of Maximum Wind Gust, Energy Price, and Weekend. The chosen method for Correlation Analysis was Pearson and a correlation heatmap was plotted (refer to Figure 2) using the Pandas library.

Notably, most of the Pearson Correlation values were below 0.5, suggesting a weak relationship. Negative values indicated an inverse relationship. The weather features that demonstrated correlation values closest to 0.5 were Maximum Daily Temperature, Minimum Daily Temperature, Sunshine Hours, Evaporation, Energy Price, and WEEKEND. These features were further visualized in the next session to check for linearity.

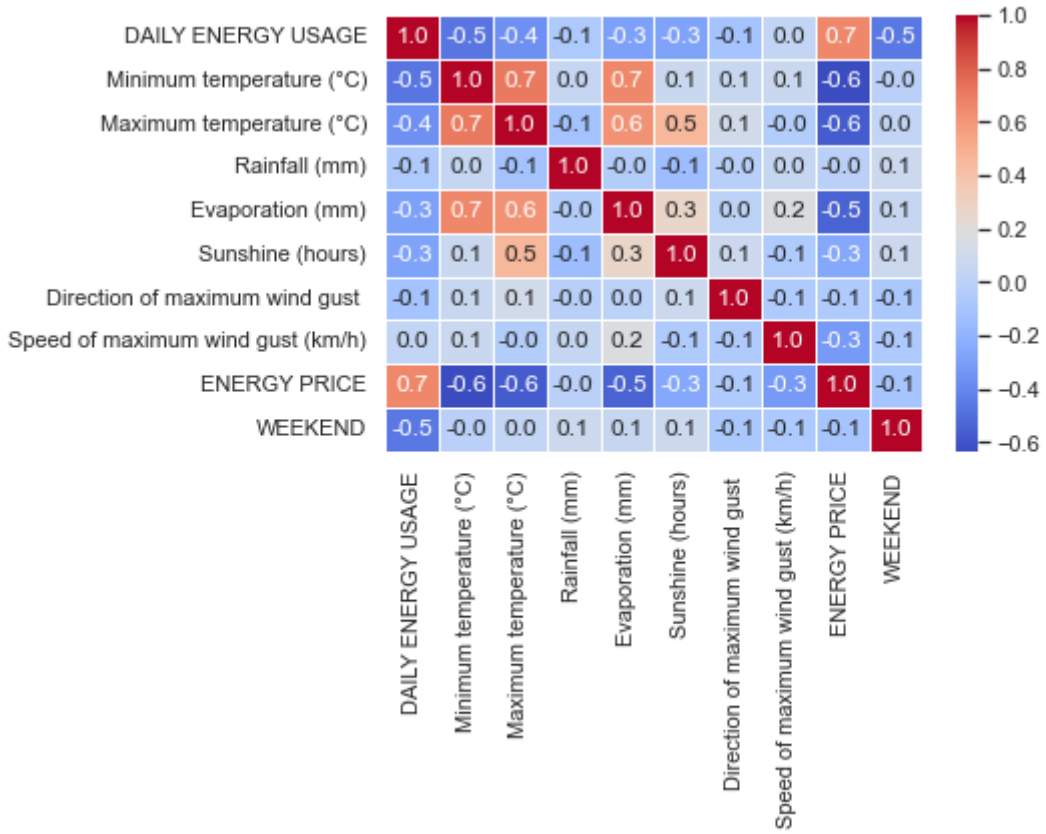

Figure 2: Correlation heatmap between Daily Energy Usage and different features of the weather

VISUALISATION

Following the Correlation Analysis, the next step involves Data Visualisation which is a powerful technique to represent data in an easy-to-understand manner. This technique involves plotting 2D/3D graphs between different variables to showcase the relationship between them. Depending on the requirement and scenario, different types of graphs can be plotted such as Line graph, Scatter plot, Bar chart, Pie charts, Histograms, and Box plot.

In our assignment, we have utilized Scatter plots and Pair plot to understand the relationship and patterns that exist between dependent and independent variables. Pair plot was implemented using seaborn library in section 3.1 to map Daily Energy Usage against different weather features such as Maximum Daily Temperature, Minimum Daily Temperature, Sunshine Hours, and Evaporation. The results of Pair plots revealed some of these relationships to be non-linear.

Based on the low Pearson Correlation values and non-linear relationships observed through visualisation, we assume a non-linear relationship between the variables. Therefore, we will use Regression Models for our research as the dependent variable, i.e. Maximum Daily Energy Usage, has continuous data. We will start with the Linear Regression Model and subsequently explore non-linear Regression Models as part of our research.

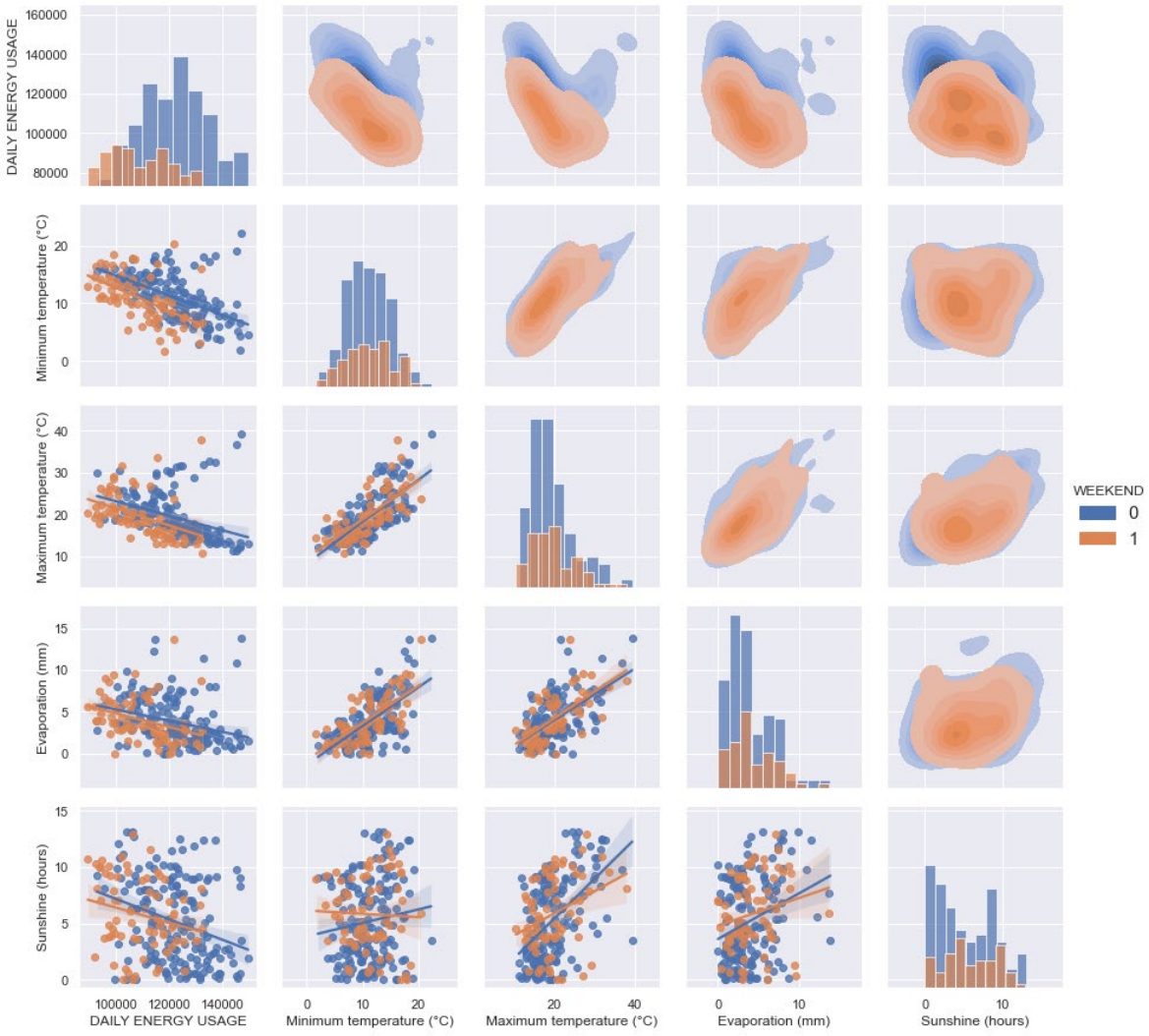

Figure 3: Pair plots between maximum daily energy usage and weather conditions

MODEL BUILDING AND EVALUATION

Following the data visualisation and correlation analysis, we built a linear regression model using k-fold cross-validation to predicting the maximum daily energy usage in Melbourne based on weather conditions and energy pricing. The 'build_model' function was used, which takes the feature matrix, target variable, and number of folds as input and returns a list of R-squared scores for each fold, the average R-squared score and the average root-mean-squared-error (RMSE) across all folds.

The following features were selected:

- Minimum temperature (°C)
- Maximum temperature (°C)
- Rainfall (mm)
- Evaporation (mm)
- Sunshine (hours)
- Direction of maximum wind gust
- Speed of maximum wind gust (km/h)
- ENERGY PRICE
- WEEKEND

Our linear regression model achieved an average R-squared score of 0.66 and an average RMSE of 7488.58, indicating that the selected features can be used to predict the total energy demand with reasonable accuracy.

The data scatter plots as well as the Pearson's correlation coefficients in the last session illustrate relative low correlation between selected features and target variable. To further explore their relationship, we also investigated non-linear regression models, namely the Random Forest Regressor [2] and Extra Trees Regressor [3]. Both models are part of the scikit-learn machine learning regression models. Interestingly, the non-linear regression models outperformed the linear regression model in predicting daily energy usage, in terms of RMSE, as shown in the figure 4 bar chart. The Random Forest Regressor model achieved an R-squared score of 0.78 and an RMSE of 5846.49, while the Extra Trees Regressor model achieved an R-squared score of 0.81 and an RMSE of 5556.56. One reason for this is that non-linear models are better able to capture the complex relationships between dependent and independent variables that are often present in energy usage analysis [4]. For example, energy usage may be influenced by a combination of weather, energy price, and time of the day, and the relationship between these variables may not be linear. Non-linear models are better equipped to capture these complex relationships than linear models, which assume a linear relationship between the dependent and independent variables.

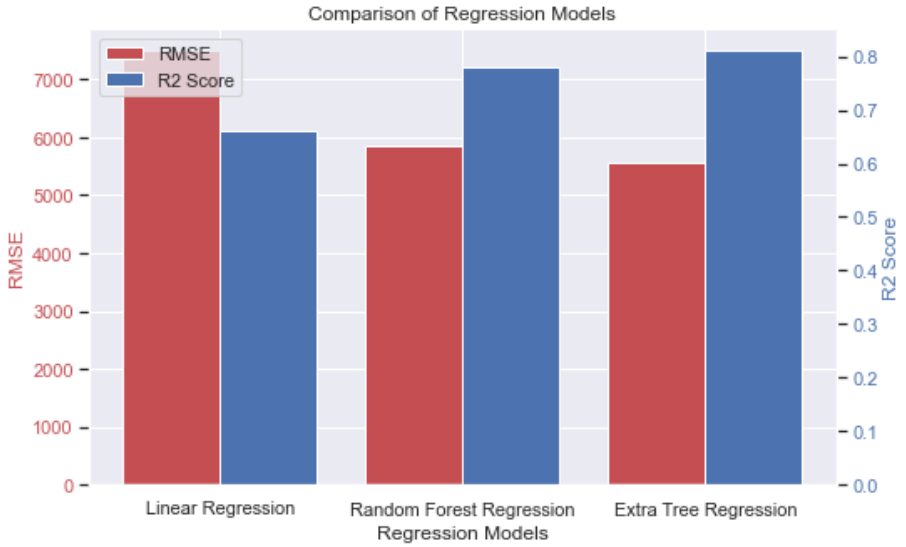

Figure 4: Comparison of different regression models

Overall, the superiority of non-linear regression models in predicting daily energy usage can be attributed to their ability to capture the complex relationships and patterns that exist in energy usage and weather data. Random Forest Regressor or Extra Trees Regressor are particularly effective non-linear models in this context, and their use can lead to more accurate predictions of maximum daily energy usage.

INSIGHTS ABOUT WEATHER AND DAILY ENERGY USAGE

Our analysis of the relationship between weather and daily energy usage revealed several significant insights. We found that temperature, evaporation, and sunshine had low negative correlations with daily energy usage. This means that as the temperature, levels of sunshine, and evaporation increase,

daily energy usage tends to decrease. On the other hand, we found that maximum wind gust and rainfall had very weak correlations with daily energy usage, suggesting that energy usage is less likely related to wind and rainfall conditions.

Furthermore, our analysis revealed a negative correlation between weekend and daily energy usage, indicating that energy usage during weekdays is higher than on weekends. This finding is consistent with previous research [4] that has identified higher energy usage during weekdays, likely due to increased demand from commercial and industrial activities. This result underscores the importance of taking into account the day of the week when predicting and managing energy usage patterns.

In addition, we found a moderate positive correlation between energy price and daily energy usage. This suggests that as energy prices increase, daily energy usage also increases. This finding has important implications for energy providers and policymakers in managing energy supply and demand as well as pricing strategies. This is likely due to energy companies tends to buy more electricity to meet the energy demand, which boosts the energy price.

In summary, our analysis highlights the significant impact of weather data, energy price and weekday/weekend on daily energy usage. These findings could be helpful for energy providers and policymakers in managing energy supply, demand and pricing. By taking these factors into account, energy providers can better predict and manage energy usage patterns, ensuring a more reliable and efficient energy supply for consumers.

LIMITATIONS/ CURRENT WEATHER ANALYSIS

While our research yielded valuable insights, it is important to note the limitations of our findings. One major constraint was the incomplete and outdated nature of the data, as the original dataset only included information from January to August 2021, with missing data that imputed with data interpolation and forward filling methods. These factors could potentially skew results and bias underlying data analysis.

Furthermore, some inaccuracies were inherent in our analysis due to the categorical-to-integer conversion of price category values, posing potential issues with accuracy. Also, this research only examined energy usage patterns in Melbourne, so the results may not be applicable to other areas of Victoria with different weather and energy usage patterns. Finally, our analysis was limited by the exclusion of relevant data such as renewable energy sources, which could influence energy consumption and prices.

To mitigate these limitations, we attempted to improve the data by using current datasets from Bureau of Meteorology and Australian Energy Market Operator covering the period from October 2022 to March 2023. By including the more current data, we aimed to improve the up-to-datedness of the analysis. We analysed the correlation between daily energy consumption and energy prices with weather features at the same time stamps each day.

Despite efforts to enhance the accuracy of the analysis, our findings revealed little correlation between weather patterns and energy pricing. At 9am, a low positive correlation existed between relative humidity and energy pricing in both the 2021 and current datasets. However, at 3pm, a negative correlation between temperature and pricing was observed in the 2021 dataset, while the current data showed a positive correlation. The wind speed remained a constant, indicating a low negative correlation to energy prices in both datasets.

CONCLUSION

In conclusion, this report has presented a comprehensive analysis of the correlation between daily weather conditions and energy usage in Melbourne, with the objective of developing a model that can accurately predict the maximum daily energy usage based on weather data. Through extensive data cleaning, we were able to ensure the accuracy and reliability of the data, and through data visualisation and correlation analysis, we identified the most significant features that influence energy usage in Melbourne.

We built a linear regression model using k-fold cross-validation, which achieved a reasonable accuracy level with an average R-squared score of 0.66 and an average RMSE of 7488.58. However, we found that non-linear regression models, such as the Random Forest Regressor and Extra Trees Regressor, outperformed the linear regression model in terms of predicting daily energy usage. It is proved that non-linear models are better suited to capturing the complex relationships between dependant and independent variables that are often present in energy usage analysis.

Furthermore, this analysis has provided valuable insights into the relationship between weather and energy usage, which can help energy companies in planning and managing future energy usage effectively. Specifically, we found that temperature, sunshine hours, and evaporation have low negative correlations with the energy usage, while energy price and weekday have stronger correlations.

However, this study also has some limitations. First, the datasets used in this study only cover a specific period of time, and future research could expand the analysis to include more years of data to improve the accuracy of the model. Second, while our model achieved reasonable accuracy, there may be other significant factors that influence energy usage that were not included in the analysis, such as renewable energy sources or special events. Finally, while our analysis provides valuable insights, it should be noted that the model is specific to Melbourne, and results may not be applicable to other cities or regions.

REFERENCES

[1] Pandas Library: pandas.DataFrame.interpolate https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.interpolate.html?highlight=interpolation

[2] Scikit-learn Library: RandomForestRegressor https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html

[3] Scikit-learn Library: Extra Trees Regressor https://scikit-learn.org/stable/modules/generated/sklearn.tree.ExtraTreeRegressor.html

[4] Joaquín Amat Rodrigo, Javier Escobar Ortiz (2021) Forecasting electricity demand with Python. https://www.cienciadedatos.net/documentos/py29-forecasting-electricity-power-demand-python.html