UNIVERSITAT POLITÈCNICA DE VALÈNCIA

Milestone M2

Project III

Degree in Data Science (GCD) - Academic Year 2023/2024

Javier Luque, Daniel Garijo, Ángel López, Andrea Sánchez, Claudia Martínez and Pablo Parrilla

Index

List of Figures

1 Introduction

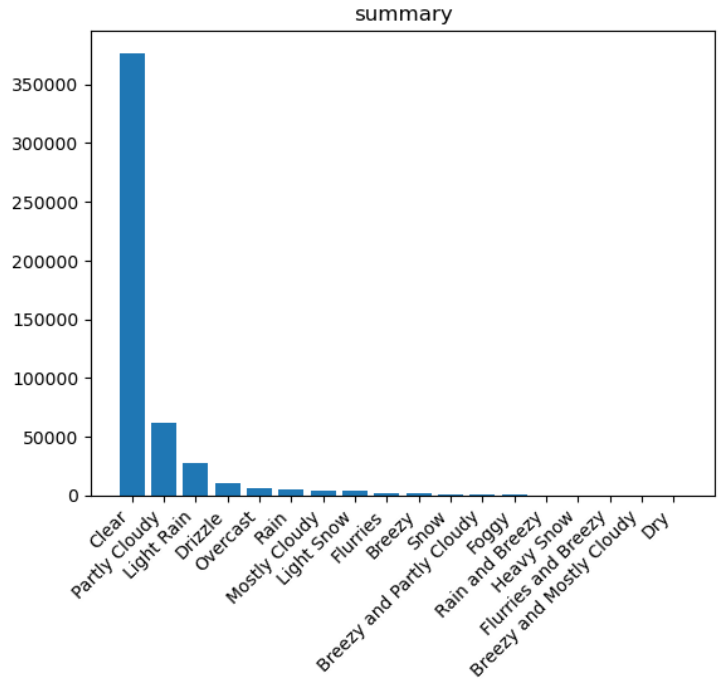
In this *"Digital Twin"* project, we delve into the world of smart homes with the aim of optimizing energy consumption and reducing costs. To achieve this, we have gathered comprehensive data on appliance usage and weather conditions. Following a meticulous process of data cleaning and preparation, we have employed advanced statistical and machine learning techniques to build predictive models. Now, we embark on the evaluation phase to ensure that these models fulfill our primary purpose: enhancing energy efficiency in smart homes.

2 Data Preparation

The data used to carry this project out is a database containing the kilowatts spent by each appliance from a house, each minute throughout a year. It is 503911 rows and 32 columns. There are different group of variables. On one hand, we have variables related to the household appliances' expenditure, while on the other hand, variables concerning the meteorological conditions of the area where the house is located. In order to know the house location since it is unknown,

2.1 Reduction of categories in weather

In our data, we had a variable named "summary" with the distribution represented in the following graphic.



zz

Figure 1: Summary distribution before

As we can see, the majority of observations are classified as "Clear", and some of them as "Partly Cloudy" and "Light Rain"; while all the other categories are very low in observations. Furthermore, the categories represent similar meteorological phenomena, e.g. "Light Rain", "Drizzle" and "Rain".

Keeping this in mind, we grouped all these categories and fit them into five new categories: “Clear”, “Cloudy”, “Rain”, “Snow” and “Others”, which contemplates phenomena as wind or fog. We can see the new distribution in the following graphic.

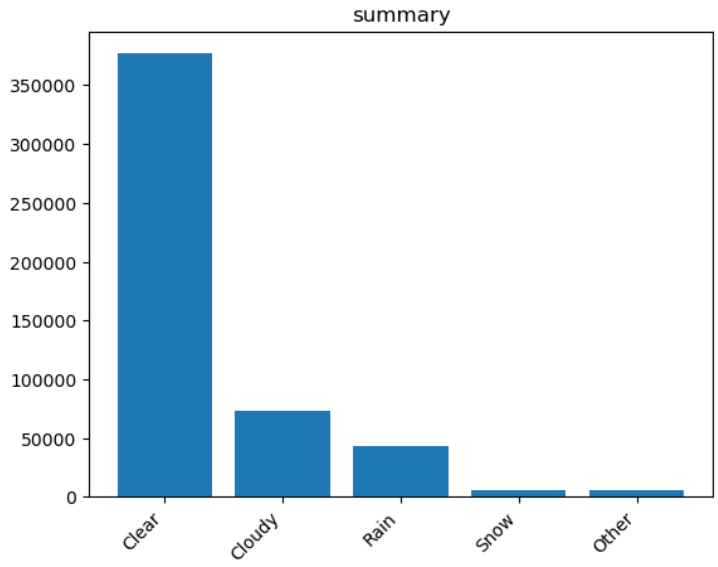


Figure 2: Summary distribution after

With this transformation, we ease the creation of dummy variables for the summary variable; given that we only need 4 dummy variables, so much better than the previous 17 needed dummy variables.

2.2 Creation of “ON” binary variables

Some of the appliance considered in the data are machines that are usually off and people turn on in order to use them. Specifically, we consider this type of appliance the microwave, the dishwasher and the two ovens. Despite the appliance being off most of the time, it still consumes energy; but when the appliance is turned off, there’s a highlight on its consumption.

Studying this four variables, we decide to impose a threshold to considerate if the appliance is on and off and keep this information in a new variable for each appliance.

In order to select the threshold for each appliance, we visualize the consumption of each appliance during two days, and draw a line of the selected threshold right above the line of values that clearly represent when the appliance is off.

In the next graphic, we can see how the black line is above the red line formed by the big quantity of low values (when the appliance is off) and represents when the appliance is turned on. The concrete threshold for each appliance are: 0.03 (dishwasher), 0.06 (furnace 1), 0.09 (furnace 2) and 0.04 (microwave).

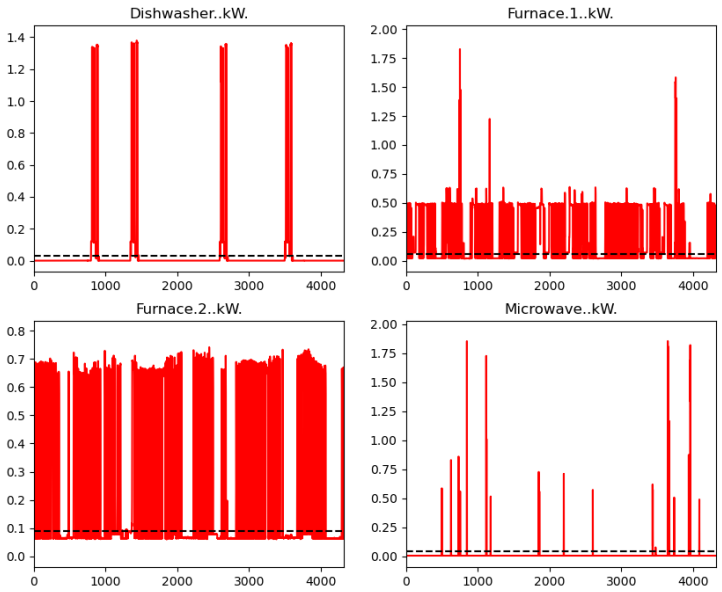


Figure 3: Consumption per appliance

Finally, we create these new variables for each appliance in which, if the value of the original variable is greater than the threshold, the value of the new variable will be 1; otherwise, it will be 0.

2.3 Change of date

Originally, the date data was in UNIX format. Initially, working with the date in this format was not feasible, so it had to be converted to "datetime" type. Additionally, the returned format was incorrect, as it matched the date (January 1, 2016) but not the time, which should have been 5 o'clock in the morning.

After replacing the dates, we had all the data in the appropriate format to proceed with the project.

3 Task Description

To continue with the report, we will undertake a detailed exploration of the three crucial aspects to outline the task description of it: *'Business Goals and Analysis Objectives'*, *'Minable View from Historic Data'* and *'Bussiness Metrics'*.

3.1 Business Goals and Analysis Objectives

The project objectives range from understanding consumption patterns to implementing practical measures to reduce energy usage in households. To achieve these objectives, prototype models will be developed to simulate the energy behavior of homes and allow for detailed analysis. These prototype models will be essential for investigating and better understanding the relationships between energy consumption, meteorological variables, and other relevant factors. Additionally, they will be valuable tools for testing and validating different energy efficiency and cost reduction strategies.

3.2 Minable View from Historic Data

Now is time to talk about the historical data, which refers to the records or logs of past events, activities, or measurements within the smart house environment. In this way we can recognize two types of variables:

Key Variables

- Historical Energy Consumption Data: Records of electricity consumption over time, measured in kilowatt (kW).
- Temperature Variations: Time-series data showing fluctuations in indoor and outdoor temperatures recorded by temperature sensors.

Additional Variables to Consider

- Weather Data: Historical weather conditions such as temperature, humidity, precipitation, and wind speed, tell us how can impact that in the energy usage and indoor comfort levels.
- Appliance Usage: Records of when specific appliances were turned on/off or their power consumption levels, what it can provide us insights into usage patterns and energy demand.

3.3 Bussiness Metrics

The main metrics of the proyect will be two: energy and money. As we want to reduce the use of energy, one of the main metrics must be the energy, measured in kW, as well as the data. The used energy in a period of time is important to learn the behavior of the devices which use energy so we can attempt to reduce it.

The second key metric is money. Our second goal is to try to reduce the cost of energy usage for buyers, so the money our clients can save is both a metric and a vital aspect of the project. Ultimately, by optimizing energy efficiency, we aim to provide substantial financial benefits to our customers.

4 Model prototype and evaluation

4.1 Temporal series

Before proceeding to the modeling prototype, it is necessary to have a deeper understanding of the data. Thus, an analysis using correlations has been conducted to identify which variables may cause fluctuations in energy consumption. At first glance, strong correlations between weather variables and device energy consumption have not been observed, beyond slight correlations between temperature and the consumption of the cellar or oven (in this case inverse). Although it is true that it seems that higher energy expenditure occurs during warm periods than in cold seasons. Perhaps, by conducting a more detailed analysis, groups of energy consumption variables that behave similarly when weather values change could be identified.

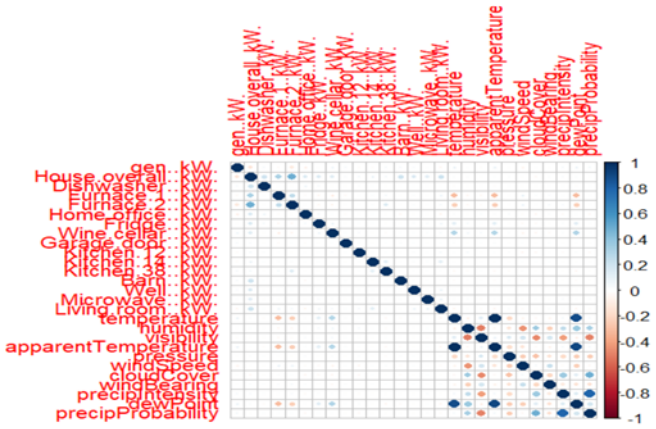


Figure 4: Covariance matrix

Once significant relationships between variables have been reviewed, the distribution of energy consumption based on its usage will be studied. This could prioritize reducing energy expenditure in activities that may be less essential and allocating expenses to essential household functions that are almost unavoidable. For example, basic appliances like the refrigerator or washing machine are grouped into Essential Services, Work Area corresponds to office or work areas, while Living and Entertainment encompasses the home's relaxation area such as the living room.

The resulting distribution of consumed energy could be considered expected, and once understood, further exploration can begin. As observed, the majority of household energy consumption is attributed to essential appliances, while non-leisure areas consume almost twice as much energy as areas designated for relaxation or entertainment.

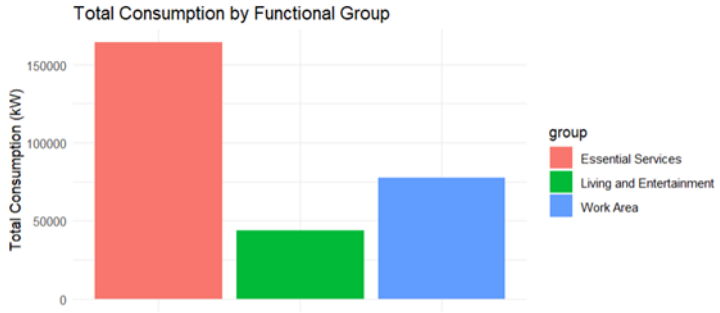


Figure 5: Total Consumption

Firstly, an estimated prediction of energy expenditure is sought without adding weather conditions to the model. Therefore, the model will attempt to predict the total energy consumption of the house based on historical data. In this case, to serve as a basis for further iterations and deepening of the model, an attempt will be made to predict 15 days of the already studied period to check the operation and corroborate the potential of the model. The study is chosen to be conducted on a daily basis because conducting it on a weekly basis would greatly reduce the sample and complicate the development of the model, while conducting it hourly would increase costs and probably a prediction on an hourly basis would not be more useful to consumers. Therefore, this specific number of days has been chosen to balance both the effectiveness and usefulness of the model.

Thus, after data preparation, the dataset is divided into a training set and a test set, which will be used to verify the accuracy of the model. Since it is a time series, cross-validation is not possible. However, an ARIMA model has been chosen to perform the task, especially effective in contexts where data exhibit trends and seasonal patterns, as is common in energy consumption records. In Figure 6, we can observe that there are indeed strong seasonal patterns, by weeks.

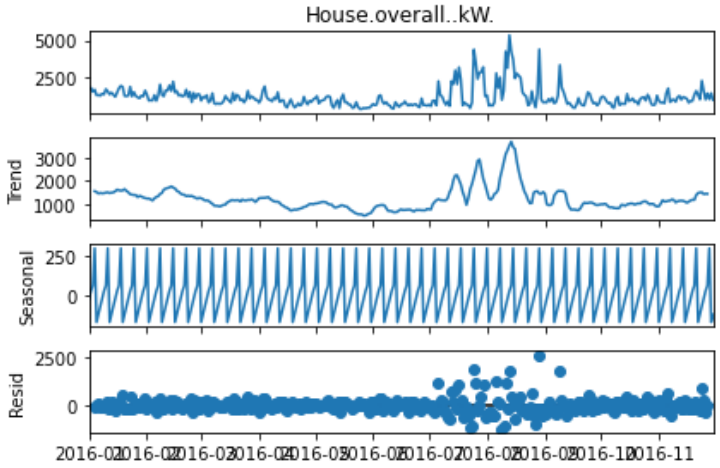


Figure 6: House Overall series components

This line graph compares the actual and predicted energy consumption by ARIMA model for both training and test data sets. The x-axis represents the time period, and the y-axis represents energy consumption in kilowatts (kW). We can see that the model effectively captures the patterns in the training data, while there is a higher error for the test data.

There are some more discrepancies between the real and predicted lines for peak consumption periods, maybe caused because of factors like extreme weather events or unexpected changes in consumption patterns. The prediction generally follows the trend of the test real line. However, we deduce the model is overfitted due to the small amount of samples. We will solve this problem the following weeks.

To assess the adequacy of the ARIMA model, we will measure it using RMSE. This allows us to understand the error in the same units as the time series and to give more weight to larger errors, which is ideal for our case where energy consumption peaks are critical. We have obtained an RMSE of 266.29 kW for the ARIMA model on the test set.

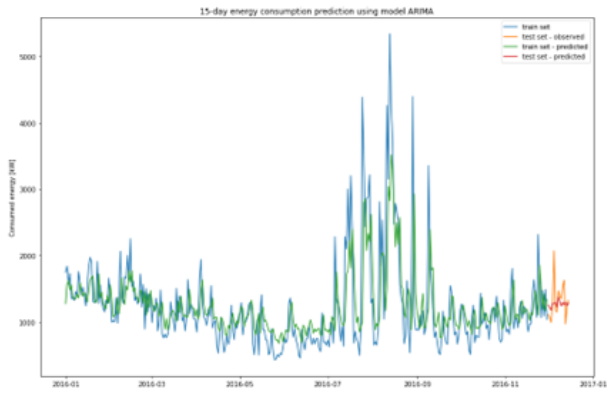


Figure 7: Predicted consumption

5 Discussion

It should be noted that there might be some overfitting to this prototype model, evidenced by the less accurate estimation of values corresponding to the test set. This will be one of the aspects to improve in the following iterations.

Thus, once this initial contact with the model has been made, along with its corresponding evaluation, further exploration of the various concepts mentioned will be carried out to enhance it and contribute effectively to the final product. We will aim to detail the appliances or locations that contribute most to consumption, so that when attempting to reduce a home's potential consumption, the model focuses on those items that are most significant in that consumption.

Additionally, we will attempt to add the option to the model to predict energy expenditure among possible groupings of devices by areas of the house or function, in order to make isolated predictions, assuming an improvement in the quality of the final product.

Similarly, by introducing climatic variables, a model can be generated that combines the time series with corresponding weather values, which could enhance effectiveness in distinguishing between seasons and consumption patterns depending on the weather. Thus, generating a unique product that could provide future energy consumption for customers considering both historical energy expenditure data and expected weather conditions.

6 Mockup

To solve our objectives, we will design a mockup prototype. It provides a tangible way to illustrate the predictions through a user interface (UI) prototype. This helps stakeholder to understand how the model would be used in a real-world application. In essence, a deployment mockup acts as a bridge between the technical aspects of the model and its potential real-world application. It helps assess the feasibility, value proposition, and user experience of the final product before significant development efforts begin.

Our prototype consists of a mobile application in which the user has to answer different questions about the house members' routine. Once the program has analysed the answers, it will predict the kilowatts the house will consume the next day. The mockup can be visualized in [3].

7 Appendix

[1] https://github.com/PROYIII/Appliances

[2] https://github.com/PROYIII/Summary

[3] https://github.com/PROYIII/MockUp

[4] https://github.com/PROYIII/Analysis