NOVA IMS Information Management School

MDSAA

Master Degree Program in

Data Science and Advanced Analytics

Predicting Individual Guarantee Based on Energy Consumption

Maria Leonor Santiago Neves Morgado

Internship Report

presented as partial requirement for obtaining the Master Degree Program in Data Science and Advanced Analytics

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

PREDICTING INDIVIDUAL GUARANTEE BASED ON ENERGY CONSUMPTION

by

Maria Leonor Morgado

Internship report presented as partial requirement for obtaining the Master's degree in Advanced Analytics, with a Specialization in Business Analytics

Supervisor / Co Supervisor: Bruno Damásio

February 2023

STATEMENT OF INTEGRITY

I hereby declare having conducted this academic work with integrity. I confirm that I have not used plagiarism or any form of undue use of information or falsification of results along the process leading to its elaboration. I further declare that I have fully acknowledge the Rules of Conduct and Code of Honor from the NOVA Information Management School.

Lisbon, 28 February 2023

ABSTRACT

Towards having a successful business, a company must be compliant with all legal and regulatory demands, namely liabilities and obligations payment. The non-fulfilment of these rules can severely harm an entity, either in its finances and reputation or even in its business performance. Hence, EDP Comercial wants to comply with regulation and ensure that the Individual Guarantee liability is covered. Since this responsibility is secured trough a bank guarantee and changing it requires a long process with higher management's consent there is a need to determine the likely amount required ahead of time. The Individual Guarantee is based on the agent's client's energy consumption thus, in order to determine its future value will be we need to predict energy consumption.

Several models were used to try and predict energy consumption. Prophet was the time-series forecasting model that best addressed our problem. Trough the described approach we found that the guarantee in place is enough to be compliant with the regulations for, at least, the next three months and so, no action is needed in what concerns the Individual Guarantee. In addition to answering the research question, with this project we also delivered the tools to measure and control this liability progression over time. The project was appreciated and well received by the company which already has in sight its application to other markets.

KEYWORDS

Energy Consumption; Time series forecasting; Prophet

INDEX

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS AND ACRONYMS

ANN	Artificial Neural Networks
ARIMA	Autoregressive Integrated Moving Average
CRISP-DM	Cross-Industry Standard Process for Data Mining
ERSE	Energy Services Regulatory Authority
GIG	Gestor Integrado de Garantias
LSTM	Long Short-Term Memory
MAE	Mean Absolute Error
MAPE	Mean Absolute Percentage Error
ML	Machine Learning
MLR	Multiple Linear Regression
MMC	Multivariate Markov Chains
MSTL	Multiple Seasonal-Trend Decomposition using LOESS
MSTS	Multi Seasonal Time Series
REN	Redes Energéticas Nacionais
RMSE	Root Mean Square Error
SARIMA	Seasonal Autoregressive Integrated Moving Average
SEATS	Seasonal Extraction in ARIMA Time Series
STL	Seasonal-Trend Decomposition using LOESS

1. INTRODUCTION

Modern societies use power for everything and that is shown in daily activities, from basic tasks like cooking to more complex ones such as transportation. Everything requires electrical power to operate and that means modern societies depend on energy to function. Being such an important resource, it needs to be managed, regulated, and monitored carefully by governing bodies in each country. In Portugal, those responsibilities fall under "Entidade Regulatória dos Serviços Energéticos" (ERSE). They are responsible for setting the rules that govern energy production and usage and also ensuring those rules are followed by all the agents in the market.

One of the standards set by ERSE is the need for each energy trader to provide a bank guarantee to ensure compliance with liability. That means energy providers need to transfer an amount to the governing body based on their expected use of the network. The availability of that amount on a daily basis is critical to enable the use of the network. In case there is a breach, the provider can be severely punished by either a monetary penalty or the inability to acquire new clients.

In the context of an internship at EDP, there was a need to determine if the amount currently available as a guarantee suffices for the next 90 days or if, on the other hand, that amount needs to be increased in order to comply with the regulation put forward by ERSE. Having this information ahead of time is of the utmost importance for the business given the process to increase the amount is time consuming and it would be difficult to comply with the 8-day requirement put forward by ERSE if that need was only identified when the regulator reached out with a warning.

In order to complete the project and determine if the amount currently available as a guarantee suffices for the next 90 days, we need to estimate the amount that will be requested as a guarantee in 90 days' time. That amount will include a solidarity component and an individual component. The solidarity component is used as a contribution to the emergency fund and is stable over time. The individual component is driven by energy consumption and varies greatly over time given it is impacted by many external factors. To answer the question put forward in this project, we need to predict energy consumption, because that is the component that determines the amount that needs to be added as an individual guarantee. The solidarity component is disregarded given it is stable over time.

This internship report aims at explaining how this problem was addressed in the context of a Data Science internship at EDP. In the second chapter we describe the energy market in Portugal and describe the guarantee and how it is calculated in greater detail. The third chapter is focused on existing literature concerning the prediction of energy consumption. The fourth chapter describes the methodology leveraged in the project and the remaining chapters analyse the results.

2. BUSINESS CONTEXT AND INDIVIDUAL GUARANTEE

Before describing the guarantee, it is worth spending some time describing the energy market in Portugal and how energy is produced. The energy sourcing process starts with its generation, a sector that is liberalized in Portugal. Energy is produced in specialized centrals owned by energy providers. Following from that, energy is taken to the population via transmission towers, in a process that can be better characterized as a regulated monopoly managed by "Redes Energéticas Nacionais" (REN), that is also regulated by ERSE. Distribution comes next, after being transported and transformed, energy is ready to be distributed for domestic use, this process is also regulated by ERSE. The final step is the actual supply whereby, if you have a contract with an energy supplier, you can use it.

As explained previously, this process needs to be heavily regulated. The service must be accessible to everyone and costs need to be reasonable. Also, transparency, efficiency and the continuity of the energy supply must be ensured to protect the society and everyone's interests. Hence, a set of standards regarding the quality of the supply needs to be ensured by market agents for them to be a part of this sector and their remuneration is based on their performance and accomplishment on those standards. So, in sum, the energy sector is regulated to minimize its inefficiencies and protect society in its entirety.

The regulations imposed on the sector are set by ERSE and monitored by the same regulating body. ERSE is also responsible for setting sanctions in case of non-compliance and for solving conflicts in case they occur. Therefore, ERSE's intervention is most seen in commercial relations, service quality and in the tariffs applied (ERSE, 2022.)

Towards assuring market agents are compliant with their duties and to soften the risk of any problems in energy supply, the energy services regulatory entity has in place a norm which states that market agents must hold a daily guarantee. This entity also nominated an organization called "Gestor Integrado de Garantias" (GIG) to manage and secure the national electric and gas systems guarantees (EDP Comercial, 2015; GIG, 2022). This duty is fully explained in Diretiva n.o 7/2021, de 15 de Abril | DRE, published by ERSE, the latest update in the law decree on this matter.

According to this document, the guarantee can be placed in several ways such as a cash deposit, a credit line with a financial entity or a bank guarantee. It is composed by two elements, a solidarity guarantees and an individual one. The first one, concerns coverage of overall market agents activities as the last one is based solely on the agent itself. For this study, we will focus on the individual guarantee for the National Electric System which has a higher interest and so weights more in traders duties.

The calculation for the individual guarantee is mainly dependent on the agent responsibilities and is given by

$$G_i^{IND} \geq RespIND_i \times k_i \times z \quad (1)$$

where G_i^{IND} is the value to be hold by the market agent i,$RespIND_i$ is value of its responsibilities, k_i a multiplicative factor determined by the compliance with duties over the last 90 days and z, a variable of partition of the value which is approved by ERSE.

The mathematical expression to compute the value of the individual responsibilities is the following

$$RespIND_i = ContUR_i + GGS_i$$

(2)

Being that ContUR regards the access and usage of the networks and is given by

$$ContUR_i = F_i \times (d_i + a_i)$$

(3)

where F_i is the daily average of the billings in the prior ninety days, d_i is the average number of days where credit was granted and a_i has the value of two ways to justify the guarantee values update for the market agent i.

Finally, GGS concerns the market adhesion as stated by

$$GGS_i = F_i \times (d_i + a_i)$$

(4)

where F_i is the daily average value of payment obligations from in the prior ninety days, d_i is the average number of days where credit was granted and a_i has the value of two days to justify the delay in guarantee values update.

The decree also mentions that this individual guarantee has a daily frequency and if the guarantee on hold is less than the responsibilities GIG notifies immediately the agent in question, which has eight days to update the guarantee otherwise, the broker is unable to acquire new clients.

To prevent this situation from happening GIG warns the market agent when the guarantee in place represents 75% of their responsibilities and also when that duty takes the value of 95%. However, if they are not compliant once all deadlines expire and the guarantee is not updated the infrastructure company is notified and the guarantee in place is liquidated.

A company can be seriously impaired if their duties are not being tackled, the bigger the company is the more it can lose. Furthermore, the time established to make an update of the guarantee is short in case it is needed to reach out to many people in order to have permission to increase the guarantee and then for that to be in place so, is very important to be one step ahead of time and control the guarantee progression across time.

3. LITERATURE REVIEW

This chapter contains the literature review conducted with the aim of understanding which approaches were taken by others when confronted with the challenge of predicting energy consumption. We will review the different techniques used and their particularities as well as the context in which they were implemented.

Energy consumption forecasting is like any time-series forecast in that historical data collected over time is used to predict a future value of the target variable. Time-series data is processed with different levels of granularity (e.g., daily, hourly) (Ghalehkhondabi et al., 2017; Hyndman & Athanasopoulos, 2015) and with the assumption that the future can be predicted leveraging past values (Hyndman & Athanasopoulos, 2015). Beyond this, other explanatory variables are often used based on their relation with the target variable. In energy consumption typical explanatory variables include population, gross domestic product, industrial sector, temperature, and humidity(González-Romera et al., 2007; Li, 2019). This can be explained by the fact that energy consumption reflects technological developments across industry, population growth and economic progression in societies (Li, 2019; Murti et al., 2022). One interesting proposal to make use of both historical and explanatory variables is the usage of Multivariate Markov Chains (MMC) in a regression model. This can be used as covariates on time dependent variables and can provide an upgrade in model performance (Damásio & Nicolau, 2014).

As for techniques used, some studies invest on more traditional models which deduce their estimate through past observations, being based on statistics and econometrics. While other researches explore bolder approaches that apply Machine Learning (ML) to get an accurate estimate, like Artificial Neural Networks (ANNs) and Long Short-Term Memory (LSTM) (Chaturvedi et al., 2022; Ghalehkhondabi et al., 2017). Multiple Linear Regression (MLR) and Autoregressive Integrated Moving Average (ARIMA) are part of the most traditional models. The first model is easy to implement and to understand, that combined with the fact that it does not need a very large period of analysis are its strongest points. Considering its simplicity, it also has good accuracy in non-linear situations and with external effects, which can be observed in the model estimates for its accuracy and performance (Amber et al., 2018; Ghalehkhondabi et al., 2017). Regarding ARIMA, a popular approach in forecasting for its broad scope due to the possibility of tunning it to best fit the data through its variations such as Seasonal Autoregressive Integrated Moving Average (SARIMA). Here, time series data is decomposed into trend, seasonality, and error to better understand each component and its effect. However, this model lacks accuracy in predicting non-stationary time series and in detecting all features in energy consumption data, known for its high variation (Ghalehkhondabi et al., 2017; Goudarzi et al., 2019).

An alternative to MLR and ARIMA is Prophet, a recent model created by Facebook which takes the form of an additive equation between trend, seasonality and holidays. As a consequence of accounting for outliers in behavior and also multiple seasonality and holidays, very important factors in energy since they clearly define energy demand, this method retrieves higher accuracy. Besides that, prophet also gains for its faster performance and process over large datasets (Chaturvedi et al., 2022; Murti et al., 2022).

Moreover, on the ML options ANNs are one of the most seen when conducting a forecast for energy consumption. These networks are favorite when forecasting with unpredictable and unknown factors due to its adaptability to the problem. They are composed by several neurons which are all linked following a specific architecture and can be divided into many layers. ANNs map the output using the inputs given with weights and biases which are constantly adjusting and learning as the inputs are added and the model trained until the error is minimized. Although they require some computation capability, ANNs are acknowledged to forecast hourly and daily energy consumption (Amber et al., 2018; Rodrigues et al., 2014).

Similar to ANNs, LSTM also requires a large training set to be properly executed and to process, classify and estimate the data. This one easily understands the sequence and contexts of the series, hence its increased usage and popularity in time series forecasting (Murti et al., 2022).

The different techniques implemented are explained by distinct contexts of the studies mentioned and the data that was available. Regardless, it is unanimous that energy consumption varies greatly and is influenced by many external factors. That means a wide range of methods needs to be tested with the view of solving the problem we have before us.

4. METHODOLOGY

In order to reach the goal of this assignment, several steps were performed following the Cross-Industry Standard Process for Data Mining (CRISP-DM) methodology. Thus, as the Figure 4.1. displays, in this chapter we start by presenting the view on research understanding aligned with the business and supported by context, following the energy sector. Then, we introduce the approach on data understanding, which includes, data collection, exploratory analysis and data quality check. Furthermore, in data preparation is exposed how data is transformed to accommodate the next phases. Next, the modelling techniques applied are described and after that, in evaluation phase, evaluation metrics on those models are explained to realize if the defined objective is reached. Finally, the deployment stage is implemented, which is in the next section of the study.

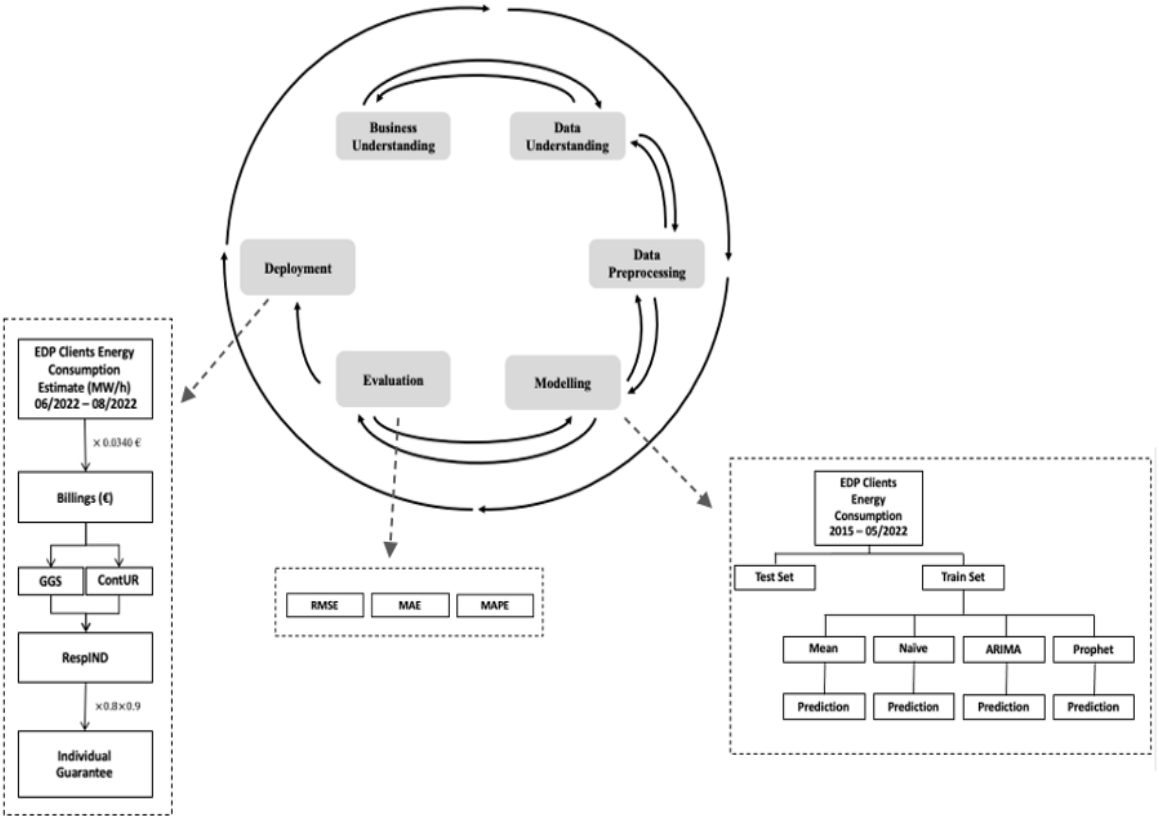

Figure 4.1 - Overview on Methodology workflow

4.1. RESEARCH UNDERSTANDING

This study intends to calculate an estimate of the Individual Guarantee that EDP Comercial will need to have in place for the next ninety days. Based on our research on this obligation and its calculation, described in Business Contextualization chapter, this parameter is mainly given by the company responsibility of paying for the network's usage, which they employ to supply their customers. Such obligation is calculated based on the billings average during the previous ninety days, hence, the guarantee depends on EDP's daily billings with is directly connected to their customers consumption of energy as shown below

$$F_i = C_i \times u_r$$

(5)

where Fi is the daily value of the billings, Ci stands for the daily consumption in MWh and u_r is the networks unit cost per MWh which, for the time we are considering, takes the value of 0.0340 €.

Thereby, by applying a predictive model to determine the consumption of EDP clients for the next ninety days we can reach the desired goal, the daily individual guarantee value.

4.2. DATA UNDERSTANDING

In this subsection, is detailed the steps performed on the data employed for this study. First, we give a general overview on the data and its variables, after that some exploratory data analysis was conducted with the purpose of extracting some information from the data. To conclude, we assessed the quality of the data and ensured its virtue.

4.3. DATA COLLECTION

The dataset used in the context of this project contains hourly consumption of EDP Comercial customers between January 2015 and May 2022. This dataset is composed by two variables, Timestamp and Consumption across 64 992 observations. Timestamp states the day and time of the year of an observation and is a time series variable of hourly frequency that goes from 01-01-2015 00:00 to 31-05-2022 23:00 and energy consumption in MWh is a continuous variable. Therefore, each observation represents the total EDP Comercial clients energy consumption for a single day and in a specific hour of that day, as shown in Table 4.1.

As this project is carried out within EDP, data from their clients is private and so confidentiality has to be ensured during this thesis. To comply with the company rules, a delta was applied to the data that was used in the context of this thesis. This does not affect either methodology or the conclusions described in this work.

Table 4.1 - Top 10 records on EDP clients energy consumption

Timestamp	Consumption (MW/h)
01/01/2015 00:00	1 812 793
01/01/2015 01:00	1 712 141
01/01/2015 02:00	1 567 271
01/01/2015 03:00	1 437 003
01/01/2015 04:00	1 362 408
01/01/2015 05:00	1 317 304
01/01/2015 06:00	1 290 811
01/01/2015 07:00	1 266 532
01/01/2015 08:00	1 307 256
01/01/2015 09:00	1 445 969

4.4. Exploratory Data Analysis

Before manipulating the data and getting in more thorough analysis, some exploratory data analysis was conducted with the purpose of analyzing the raw data before making any assumption that could bias our output as well as for assessing data quality and possibly detect some incongruences or noise in the data.

Figure 4.2 was achieved just by placing the data as it was given across two axes, x axis which represents the timeline and y axis has the values of consumption therefore, this graphic describes the consumption between 2015 and May 2022.

By visualizing the displayed below, is possible to see obvious pattern over the years as well as a heavy seasonality. The consumption has highs, then decreases and then raises back again, repeating this cycle behavior and moving between the values of 1M and 3M MW/h.

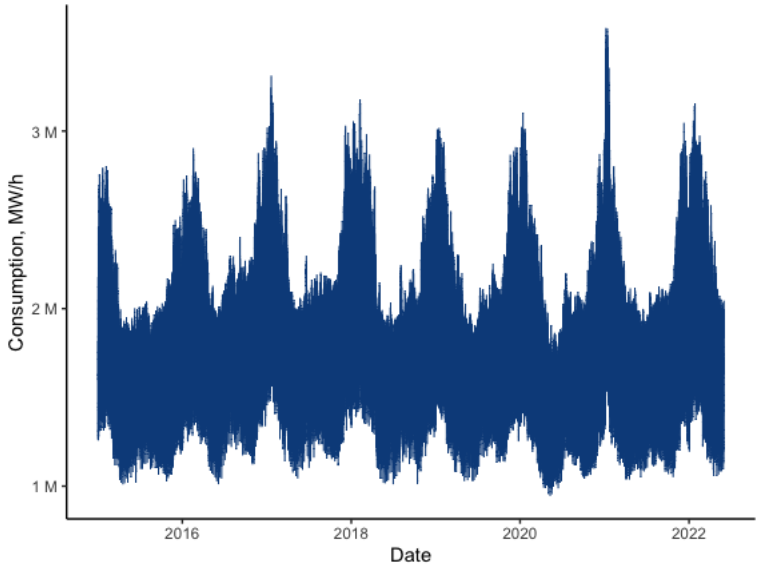

Figure 4.2 - EDP customers energy consumption between 2015 and May 2022

As shown in Figure 4.3, taking a closer look at the data and considering only the year of 2021, we can say that consumption assumes its highest values in January and February and then towards November and December as well. This behavior is expected since concerns winter months, when is cooler so it is logical that people use more electricity. In July we can also see a slight increase in power usage comparing to the previous months, this can be due to the high temperatures verified in summer season that led to need of cooling equipment. Also, through this visualization we can also support our previous comment, in fact it confirms the presence of both trend and seasonality components.

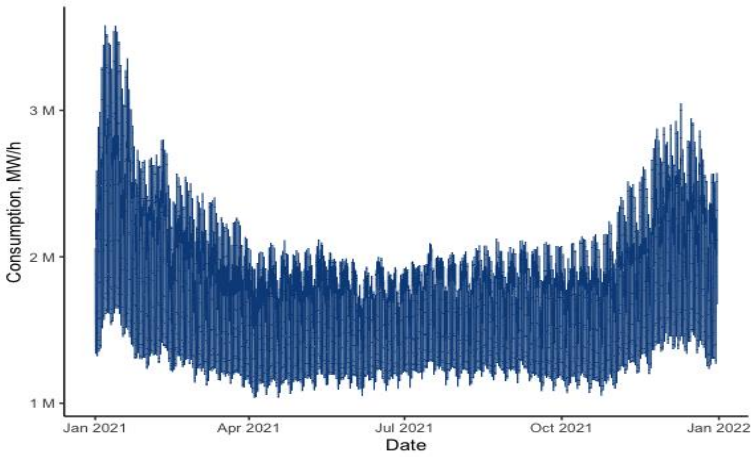

Figure 4.3 - EDP customers energy consumption in 2021

With the two previous figures we realize that the Timestamp variable highly affects the consumption of energy and its behavior. So, with the intent of extracting more knowledge on this dataset we deepen our exploratory research on this variable by creating a binary feature characterizing the energy consumption considering its week period. Hence, DayOfWeek takes the value 'weekday' if the timestamp fits a business day during the week or 'weekend' if otherwise. In computing this new feature by hour of the day we enable more insights on the energy consumption and its daily course, as can be seen below in the plot that illustrates the average consumption by hour of the day and by day of the week.

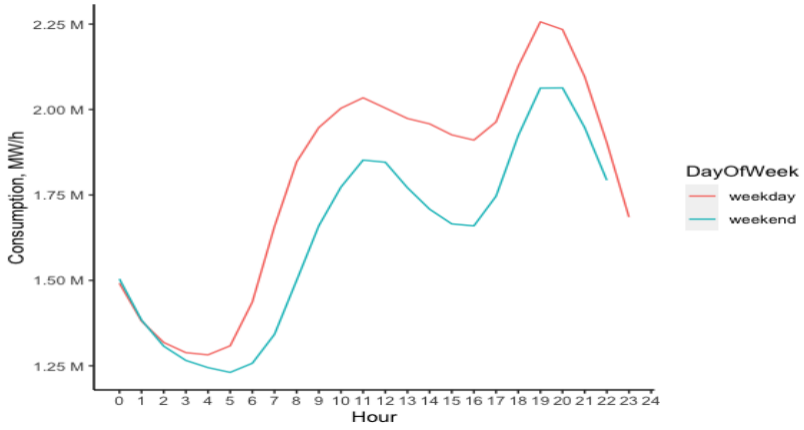

Figure 4.4 - Hourly EDP clients energy consumption by day of the week

Based in this we can affirm that consumption peaks are around lunch and dinner times, about 12p.m and 20p.m. both during weekdays and in the weekend. Also, besides similar hourly behavior, at the weekend the energy consumption is significantly higher which makes sense since people have more free time and spend it at home.

A clear cycle pattern is shown by this figure, at dawn power usage is very low and starts increasing as closer we get to the morning, until around lunch time. Then decreases somewhat and goes up again towards afternoon and dinner time where has its highest value. After that consumption is reduced and this cycle repeats itself. That said, cycle patterns are present in this data and trend and seasonality are part of it as well so, we will consider that moving forward with this assignment.

4.5. DATA QUALITY

Databases are susceptible to have some discrepancies, namely company databases and as a matter of fact, poor quality data will retrieve poor and mistaken results. As mentioned by Han et al., 2012, p.84, "Data have quality if they satisfy the requirements of the intended use. There are many factors comprising data quality, including accuracy, completeness, consistency, timeliness, believability, and interpretability". Missing values is one of many incongruences that can affect our data quality and is very common when working with data. This dataset is no exception and missing energy consumption relative to dates was found. Since this absence of value was due to time changes because of daylight save in Portugal that specific timestamp cannot have a consumption thus, the values were removed from the dataset. Another frequent occurrence in data are outliers, observations that are significantly distant from other observations. Those observations can negatively impact our analysis and reading of the data as they do not represent or translate a usual behavior, they are considered exceptions in conduct. Therefore, they should be removed from the analysis.

To sight the outliers in our data some graphics were set up. First on the range and frequency of the consumption variable and then considering both timestamps, as month, and consumption variables to see detached observations at a specific time.

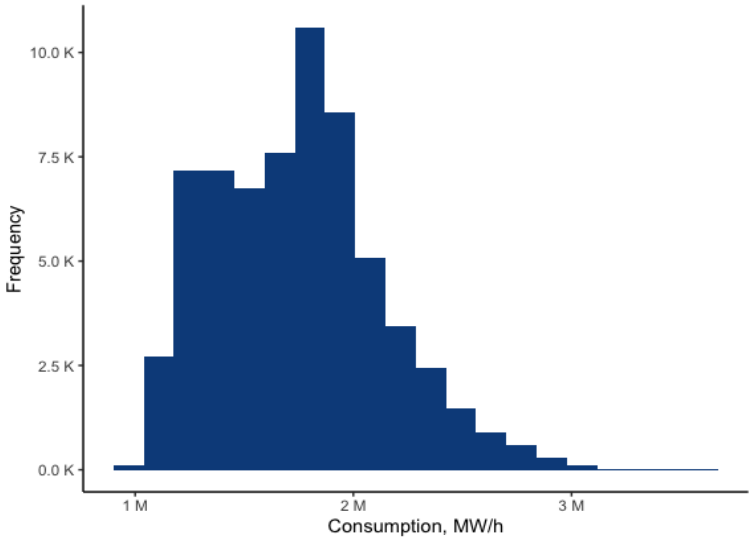

Figure 4.5 - Energy consumption frequency

Regarding solely consumption, the Figure 4.5 shows that this feature ranges between around 1M MWh and 3M MWh, having the majority of observations between 1M MWh and 2.5M MWh. So, from this figure we can claim that are a couple of observations we can exclude from this analysis for being lower or higher than all others.

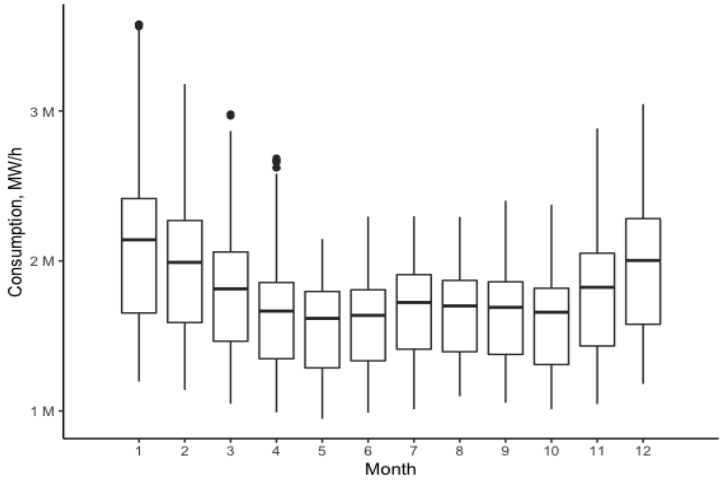

Figure 4.6 - Energy Consumption distribution over month

Looking at the data by month through a boxplot is possible to see some detached observations denominated outliers, as previously mentioned. In this figure, the observations are analysed following the Interquartile Range method, meaning that all the observations inside the rectangle are values between Q0.25 -1.5 IQR and Q0.75 + 1.5 IQR, where Q0.25 and Q0.75 are the first and third quartile respectively and IQR is the difference between them. This criterion measures the spread of the data and uses it to identify data points which have lower contribution on identifying a pattern in the data with the purpose of excluding them from the study. These observations are displayed as black points and were eliminated from our dataset since they introduce biased judgement of the data.

4.6. DATA PREPROCESSING

Time series data has some particularities as it captures data points over a frequency in time. In this dataset because of its high frequency the data is more susceptible to hold hidden relevant information. So, in order to make best use of the data we stored it in a time series object with frequency of 24 since each rows represents hourly observation.

After that, the series was decomposed into seasonality, trend and random noise, with the aim of confirming and seeing these components clearly.

Because we are dealing with multi-seasonal data Multi-Seasonal Time Series (MSTS) was used to perform the time series decomposition hence, this method allows us to specify the frequencies we want to consider for our data. In this case we accounted for hour, week and year commonness. Below, we are able to examine the data trend and hourly, weekly and yearly seasonality respectively across the past 7 years.

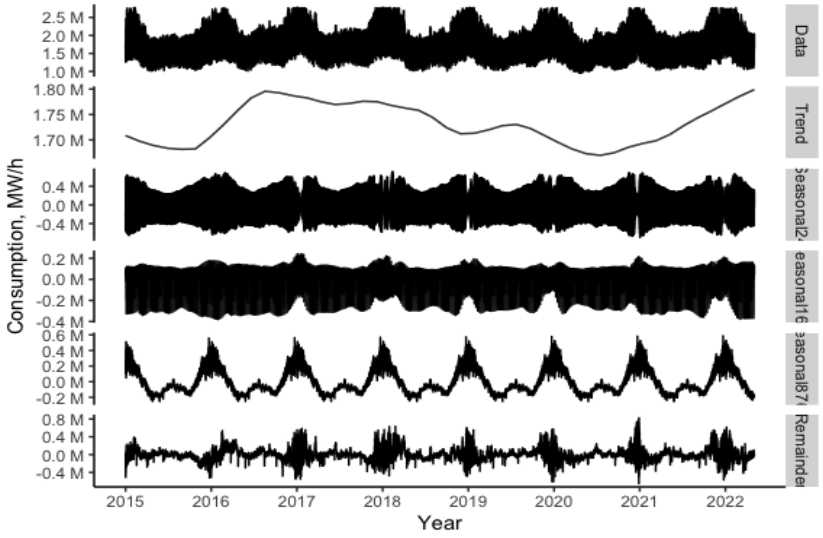

Figure 4.7 - Energy consumption time series decomposition

4.7. MODELLING

A quantitative forecast can be applied when numerical data regarding the past is available, and it is acceptable to take on that some past patterns will continue onto the future. This type of forecast fits our problem and so, we will apply predictive modelling to time series data to best estimate the future outcome on energy consumption.

In this section we start by describing time series analysis, its main characteristics, components and the analysis that can be retrieved and then move on to forecasting methodologies and their particularities.

4.7.1. Time series characteristics

Time series is a sequence of observations taken at regular intervals of time (Peña and Tsay, 2001), for instance hourly, daily, weekly, monthly, or annually. This series are univariate if contains records of a single variable or is denominated multivariate if considers records of more than one variable (Adhikari K. & R.K., 2013).

Modeling and analysis of a time series is performed with the purpose of uptaking the time-dependent structure of a univariate time series of observations and also can be fulfilled having in sight to explore the relationship among several series, in case of a multivariate time series analysis (Peña and Tsay, 2001). The observations can be measured at every instance of time and so we talk about continuous time series or they a can be discrete otherwise, when they are only collected at specific times (Adhikari K. & R.K., 2013).

When working with a time series is convenient to separate its components since each can represent a pattern in the data. This way, the insights on the series are richer and the accuracy of the forecast will also be improved (Hyndman, R.J., & Athanasopoulos, 2015). Trend, Seasonal, Cycle and other Irregular components are comprised in a time series. While Trend is a long term movement in series, Cycle represents medium-term changes that repeat there selves over time. The majority of economic time series present this cycle component and has usually four phases, prosperity, decline, depression and

recovery. The seasonal component represent fluctuations in the series within a year and can be caused by several factors. Finally we have anything else producing variations in the data that are unpredictable and not regular as they do not show a pattern (Adhikari K. & R.K., 2013). To simplify the trend and cycle components are usually combined into a single one, named trend. Therefore, having three components in a time series, Trend, which includes trend and cycle, Seasonal and the remainder (Hyndman, R.J., & Athanasopoulos, 2015).

For example in a time series explaining the customers of an ice cream brand, we can have a positive trend or a negative one, if the number of customers is increasing or decreasing over five years, respectively. Also, the clients might increase in summer time which shows the seasonality component, which can repeat itself in a cycle. Using this example, we can also have irregular components due to the lack of milk in the market for instance.

Having these four components in mind, they can be completely independent of each other or unlike that they can have some influence on each other, generating two different models, additive and multiplicative, respectively (Adhikari K. & R.K., 2013).

If the decomposition is additive,

$$y_t = S_t + T_t + R_t$$

(6)

otherwise, is multiplicative and given by

$$y_t = S_t \times T_t \times R_t$$

(7)

where y_t represents the dataset S_t is the seasonal component, T_t is the trend and cycle components combined, and Rt is the remainder, all in period t.

In case that the magnitude of the seasonality and the trend do not vary with time, a additive decomposition is more appropriate, if those criteria do not apply a multiplicative decomposition should be applied (Hyndman, R.J., & Athanasopoulos, 2015).

Several methods to decompose a time series can be applied, such as Seasonal Extraction in ARIMA Time Series (SEATS), Seasonal and Trend decomposition using Loess (STL) or Multiple Seasonal and Trend decomposition using Loess (MSTL). The first method has the downside that only handles monthly or quarterly data and does not consider the possibility that the seasonal component might change over time as STL does. On the other hand, STL does not consider multi-seasonality, a more complex seasonal pattern often present in time series with higher frequency. MSTL applies that multi seasonality to the STL method, thought this method the user can ensure to capture the all existing seasonality that best fits the data.

4.7.2. Forecasting methods

With the purpose of estimating future events, a mathematical model is developed based on past observations, this is mostly applicable when there is not more data available besides historic behaviour on the variable in study. There are several models that can be applied in time series forecast thereby,

considering the data available, our major goal and the business and company context we applied models that better fitted our situation.

Before trying some forecasting algorithms, the data was split into two subsets, train and test. The training set contains data from 2015 to 2020, which represents around 81% of the total data, while the test dataset remaining 19% of the whole data, starting in 2021 until May 2022.

4.7.2.1. Mean Method

The simplest path when making a prediction for the future is to assume that all future values are equal to mean of all observation occurred in the past. So, a future observation is given by,

$$y_{T+h|T} = (y_1 + \cdots + y_T) \div T$$

(8)

Where T is the number of observations and y_T an already observed value.

This method is very straightforward, easy in application and in interpretation of the results, it gains for being understandable for anyone and also effortless, as it does not demand many resources. This is very important in a company, where the goal is to maximize gains while using the least resources possible.

4.7.2.2. Naïve Method

Additionally, another light assumption is to consider all future observations the same as the last observation available. Such method is written as

$$y_{T+h|T} = y_T$$

(9)

This method is also a basic forecast in which we give the last value observed to all future observations. Similarly to the mean method, is easy to explain and apply however, is less accurate because it bases all its forecast in one observation.

Thereby, the Naïve method does not consider seasonality and its effects on data and, since in our study the observations have high variation, is one of the worst models for our problem.

4.7.2.3. Seasonal Naïve Method

Like the mindset prevailing in the previous method, this one makes the same supposition but counts for the existence of seasonality. Hence, the forthcoming events will be equal to the historic ones registered in the same season. Using this method is possible to specify seasonality that best fits the data. This method is given by

$$y_{T+h|T} = y_{T+h-m(k+1)}$$

(10)

where m is the seasonal period, k the number of complete years in the forecast period prior to $T + h$.

Besides taking seasonality into consideration, this model lacks in estimating yearly evolution of events and since energy consumption is not static from year to year and changes while society and industry evolves it will not retrieve the best results.

4.7.2.4. ARIMA Methods

ARIMA is one of the most seen techniques applied when time series forecasting. A time series model can be linear or non-linear depending on whether the values of the series are a linear function of past observations or not (Adhikari K. & R.K., 2013). Linear models are widely used due to their simplicity, both in understanding and implementation.

Focusing on the autocorrelation present in the data, this model is based on the combination of AR and MA models but for non-stationary time series, meaning time series that contain seasonal or trend components and so their value depends on the time at which they were observed. This method as the mathematical function of $ARIMA(p, d, q)$ and is obtained by applying differencing to the data trough p, the order of the autoregressive part, d indicating the degree of differencing used and lastly q, refers to the moving average part.

Since the ARIMA model is applied on non-seasonal data, a variation of this model was created to deal with seasonality. The seasonal ARIMA is obtained by adding the seasonal terms to ARIMA hence, is given by $ARIMA(p, d, q)(P, D, Q)_m$ where we have P, D and Q as in ARIMA but for m, the seasonal period considered. Here, the seasonal terms are only multiplied by the non-seasonal ones.

ARIMA has several variations and since we are dealing with multiple seasonality in our data, a dynamic regression was applied. This application is computationally more expensive as it is of slow processing and on the consequences side is also harder to explain which is not convenient in a company like EDP. On the other hand, this algorithm is good when working with short term forecast like ours and it only requires historic data like we have to perform our forecast.

4.7.2.5. Prophet Method

Created by Facebook, Prophet is one of the most recent approaches which can be used in time series forecasting. This method was originally thought to forecast daily data affected by weekly and yearly seasonality and also created to analyse holidays effects. Prophet works better with long historic seasons datasets and series with heavy seasonality.

This method can be looked at as a non-linear regression model and takes the form of

$$y_t = g(t) + s(t) + h(t) + \varepsilon_t$$

(11)

where $g(t)$ characterizes the trend in growth, $s(t)$ contains the several seasonal patterns effects, $h(t)$ concerns the holiday factor, and $\varepsilon_t$is an error term.

As mentioned, Prophet was originally created to predict daily observations based on large historic datasets which fits exactly our case, we want to predict daily consumption based on past records of

this variable. Besides requiring the data to be in a specific format, this method as the advantage of being very simple to use and interpret as it combines trend and seasonality components.

4.8. Evaluation Metrics

It is decisive to consider forecast accuracy when selecting the forecasting model to implement. This element is usually determined by the size of the error term. So, to compute it and understand how a model performs the dataset was divided into two sets, the training data and test data. The first one has the mission of fitting the data and estimating any parameters of that forecast method while the test one is used as an indication of how the model will perform on future data. The comparison between the results obtained for the test data applying the model and the real test data indicates the model accuracy. Therefore, the error term is given by the difference between the observed value and the predicted one. Thus, after training the model we see what it would predict for our test set timeline and since we have that data we compare its forecast to the real data and so, we have the model accuracy.

The most common metrics to determine model performance quantify the error, what was not well predicted and the most known are Root Mean Square Error (RMSE) and Mean Absolute Error (MAE) which depend on the data scale.

While in MAE we move towards a simpler approach by computing the median as shown by

$$MAE = mean(|e_t|)$$

(12)

with RMSE we look more into the mean value and that why this last one is most used.

$$RMSE = \sqrt{mean(e_t^2)}$$

(13)

Anyway, it is also possible to compute the error term without being dependent on the scale of the data, Mean Absolute Percentage Error (MAPE) is a form understanding the model performance by giving us the percentage of error in the data and is written as

$$MAPE = mean(|p_t|)$$

(14)

This last metric provides an easier way is to understand the part that the error takes in the forecast performed.

5. RESULTS AND DISCUSSION

The ultimate step when answering a business question is to make use of all research and methodology discussed and deliver the best solution in response to the problem which lead off all this study. In this chapter, the chosen model is applied to estimate EDP clients energy consumption, from there, we calculate the company billings and responsibilities with the aim of reaching the individual guarantee needed. After that we confront that value to the one EDP has already guaranteed in order to perceive if any action is needed.

After several models are studied and applied, a choice must be made to determine the model which best fits our data and the question we are trying to answer. In table below the performance of each model is exposed given the metrics selected to evaluate our predictive modelling in the context of this project.

Table 5.1- Comparison of forecast models applied

Method	Mean	Naïve	SNaïve	Arima	Prophet
RMSE	396 730	418 323	568 075	415 734	257 010
MAE	323 631	341 648	489 869	349 975	208 369
MAPE	19.3	22.1	29.1	20.0	11.6

Mean and Naïve approaches are usually used to establish a benchmark on the models used, having 396 730MW/h and 418 323 MW/h in RMSE and a MAE of 323 631 MW/h and 341 648 MW/h which translates in a 19.3% and 22.1% error, respectively. Prophet clearly stands out for returning the best results when compared with the other methods. With 257 010 MW/h in RMSE and a MAE of 208 369 which renders in 11.6% of error.

Given the results in Table 5.1 we decided to implement the Prophet method. Besides granting the best accuracy, which aligns with providing the best solution to this study, this approach considers very well both trend and multiple seasonality which is very important when predicting hourly consumption. Moreover, this model is easily understandable and not complex as requested by the team. The management required a solution to the problem that would be comprehensible for all and that would answer our question in an efficient manner.

Based on historical energy consumption data regarding the EDP clients, the Prophet model was trained and applied. Given this model, from June 1st to August 31st, we got consumption values between around 35M MWh and 45M MWh, as shown in the Figure 5.1 below. Being that we are predicting power consumption for a warmer time of the year these values are as expected and a similar behavior can be seen for the same period in the past.

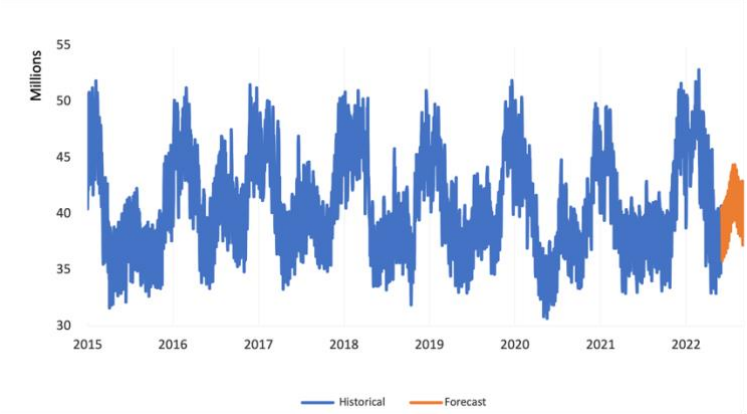

Figure 5.1 - EDP Clients Energy Consumption from 2015

Taking a closer look into the predicted consumption, below we focused only on the previous three years. The forecast behavior is not exactly the same of those years however, is has the same trend of 2020 and the consumption is similar to the one registered in the previous year.

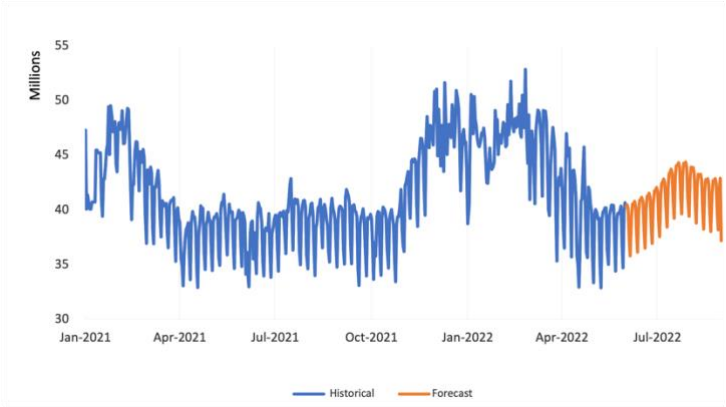

Figure 5.2 - EDP Clients Energy Consumption from 2021

Having predicted the consumption, we are closer to reach our goal and get a value for the individual guarantee. By multiplying consumption by the networks usage unit cost, which takes the value of 0.0340 € for the considered period we get EDP's billings for the same period.

Following the Diretiva n. º 7/2021, discussed in Literature Review where the guarantee calculation is fully explained, we calculated a 90 days moving average of the billings and multiplied it by 38 days, which represents the period the credit is granted, plus the days to deliver the guarantee, getting ContUR.

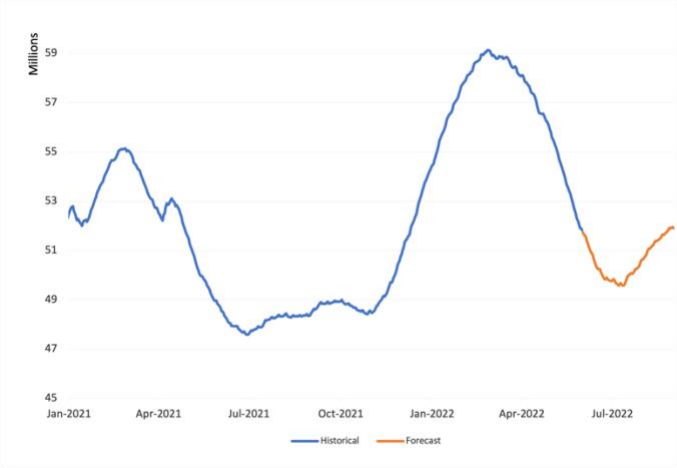

Figure 5.3 - ContUR

As mentioned in Literature review chapter, the Individual Guarantee is given by the sum of ContUR and GGS. The first one is related to the networks employed to deliver energy to the EDP clients which placed a contract of supply with the company for that period. Besides ContUR, which proves that networks are crucial to the company and support the business, this guarantee is also composed by a portion regarding the access to the market since is a regulated market with high importance to society.

Besides depending also on consumption, the calculation of GGS is not entirely clear by the team and the company and so we were asked to use a simpler approach for that prediction. To estimate the GGS for the time considered we went with same values as last month hence, this variable was also based on its historical values, provided by ERSE. As observed in Figure 5.4, GGS is very uncertain and varies a lot, even assuming some negative values. Since this piece is a smaller fraction of the whole Individual guarantee, it does not impact our prediction as much as ContUR.

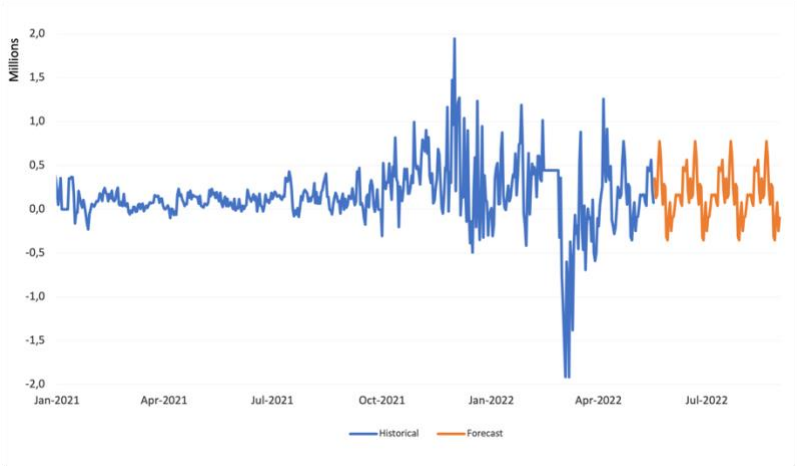

Figure 5.4 - GGS

Adding both components, ContUR and GGS, we finally get the Individual Guarantee and our main subject for this exercise, for the time considered.

The figure below, illustrates this guarantee behavior from June to August, with its forecast highlighted in orange. Within this period, this calculation hits its highest between February and April, taking values around 43M €. This is due to being one of the coolest epochs of the year where consequently the energy consumption is higher and, besides that by looking at 2021 data we can see that is a recurring pattern. From June to November is where the guarantee is lower assuming values between 34M€ and 36M €, around 80% of the maximum value registered.

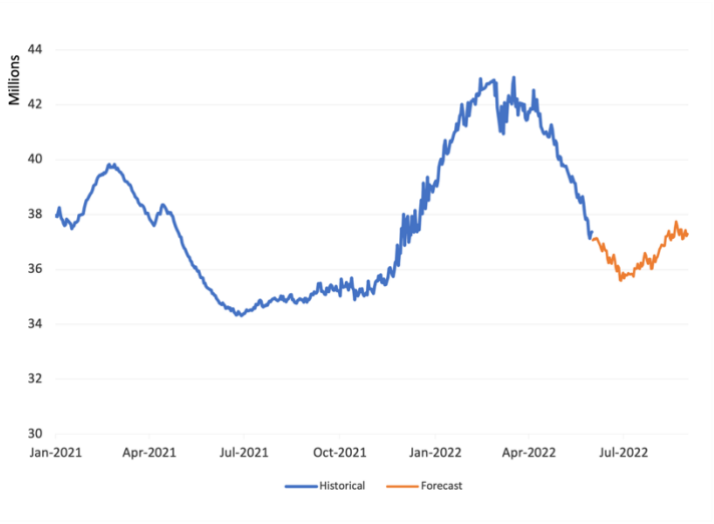

Figure 5.5 - Individual Guarantee

After having estimated the value which the Individual Guarantee will take for the next 90 days, we need to confront that value with the one the company has already in place to satisfy this payment obligation. That way we will be able to answer the major question of this study, if any action on the amount guarantee is needed, at least for the next 3 months.

For the time of this analysis EDP has in place a guarantee of 60M €, once the highest value we predict is around 43M € that leaves us with extra 17M €. Figure 5.6 considers the both the Individual Guarantee predicted in this exercise and the value EDP has already granted and shows the percentage of guarantee being used and the space we have for any unpredictable event.

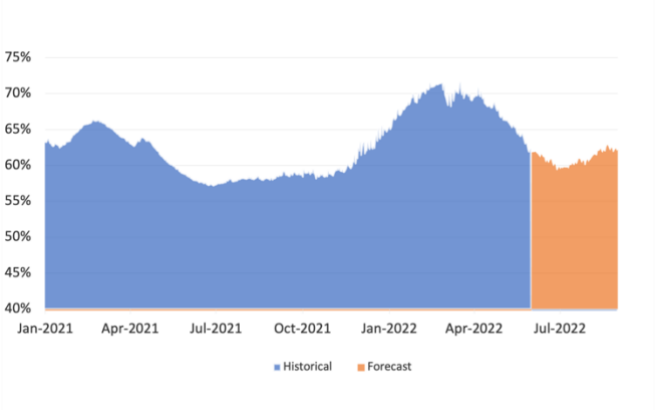

Figure 5.6 - Percentage of Individual Guarantee in use

The Enablement and Analytics team defined some key aspects to analyze this graphic and guide our insights. Thereby, to be more comfortable and have peace of mind the guarantee in use should always be less than 75% of guarantee in place. If this value reaches 85% it is a warning and we should keep an eye on it to ensure it does not get to 95% otherwise it is signed as a red flag, meaning that we should start the internal process of requesting an increase of the guarantee in order to be compliant with the industry rules and avoid being punished for delays in payment.

Currently we are in the clear and only using around 60% to 65% of the guarantee provided also, according to our prediction, for the next 90 days the maximum percentage in use will not be higher than 65% so we can assume that for the time considered no action will be needed on this matter, the Individual Guarantee will not need to be increased.

6. CONCLUSION

In sum, by conducting this study we were able to answer the question made by the management team regarding the need to increase their Individual Guarantee provided to GIG.

Alongside fulfilling that task we also developed and delivered a solution to monitor this settlement continuously for the future and in doing so, answered an important business need in an efficient way.

After realizing how the Individual Guarantee was calculated, we defined the best model to predict consumption, Prophet. From there we applied all the steps for its calculation and obtained the key value desired. To display and communicate our analysis in a way we clearly answered the question on the need of more guarantee, we chose the Power BI visualisation tool.

Through this tool the user can analyse himself the guarantee behaviour since 2015 until 3 months in the future. Both the visualization layer and the forecast which provides the predicted guarantee are weekly updated and communicated to interested ones in order to deliver a continuous analysis and be prepared for any occasion.

Furthermore, with this final product, which is the visualization dashboard, the process of requesting for an increase of the guarantee in place will be much easier. First, the weekly update on this variable will contribute for the people to be more aware of its progression. Secondly, when the request is done we can justify it with our graphic explaining the evolution of the guarantee and its forecast, without much more effort. Finally, with the estimation of the guarantee for the near future we are head of time and can make the request for any modification of the guarantee in place with time to spare and without being impaired.

Having control over this payment obligation, contributed for peace of mind of the team and the management. Employee focus can now be directed to other business needs while he task related to the obligation is secure.

This solution wins for being straight in answering the question that started this project also, the ease of access from everyone in the department and the little time spent on its update contributed for an advancement in analysis and reporting that was well received within the company and opened the possibility in automation of other narratives.

Hence, the implementation of this approach for the gas guarantee in Portugal as well as for the energy guarantee in Italy was requested and is ongoing. For the future, the idea is to extend this guarantee control and report for every market and region, starting with Poland.

In conclusion, with this study we were able to extinguish a business need and provide a efficient and long term solution. Therefore, this project was successful as the main goal was achieved, for now there is no need in increasing the Individual Guarantee in place.

7. LIMITATIONS AND RECOMMENDATIONS FOR FUTURE WORKS

In this final section the limitations and major challenges uncovered while addressing the major question of this project as well as recommendations for future works are presented and discussed.

The first limitation we were confronted with is directly related to the data available to conduct this project and introduced some difficulty in the data extraction process. Energy consumption data on EDP clients is provided by REN in an outdated frontend that, aside from having very restricted access and so only specific users can extract the data, only allows downloads on a very small period of analysis to be completed at once. This happens because consumption data is of hourly periodicity making it very heavy and the platform is not ready for such data features.

Moreover, after the research question was raised, the project was guided to focus solely on the Individual Guarantee and its components that are heavier in the overall guarantee amount, leaving other constraints with inferior estimates. This decision was caused by the lack on information on some elements combined with the alignment on priorities and time this task would take for the gains it could retrieve for the company.

Towards future works, such elements can be studied in more detail in order to achieve an estimate on the overall guarantee as close to the reality as possible.

Another point that could improve this project, is the usage of other variables that influence energy consumption like weather, energy tariffs, GDP and population growth. Through our literature review we acknowledged that energy usage varies with these factors and that if combined with historical data on the output variable in an adequate model they can also improve its estimate for the future.

In sum, this project had some limitations that impacted its development which are due to the context in which the project was proposed. Those are identified and can be addressed in future works on this matter having in sight the best estimate on the future guarantee amount.

8. REFERENCES

Adhikari K., R., & R.K., A. (2013). An Introductory Study on Time Series Modeling and Forecasting Ratnadip Adhikari R. K. Agrawal. *ArXiv Preprint ArXiv:1302.6613*.

Amber, K. P., Ahmad, R., Aslam, M. W., Kousar, A., Usman, M., & Khan, M. S. (2018). Intelligent techniques for forecasting electricity consumption of buildings. *Energy*, *157*, 886–893. https://doi.org/10.1016/J.ENERGY.2018.05.155

Chaturvedi, S., Rajasekar, E., Natarajan, S., & McCullen, N. (2022). A comparative assessment of SARIMA, LSTM RNN and Fb Prophet models to forecast total and peak monthly energy demand for India. *Energy Policy*, *168*, 113097. https://doi.org/10.1016/J.ENPOL.2022.113097

Damásio, B., & Nicolau, J. (2014). Combining a regression model with a multivariate Markov chain in a forecasting problem. *Statistics & Probability Letters*, *90*(1), 108–113. https://doi.org/10.1016/J.SPL.2014.03.026

Diretiva n.º 7/2021, de 15 de abril | DRE. (n.d.). Retrieved February 27, 2023, from https://dre.pt/dre/detalhe/diretiva/7-2021-161433228

ERSE - A ERSE. (n.d.). Retrieved February 27, 2023, from https://www.erse.pt/institucional/erse/a-erse/

Ghalehkhondabi, I., Ardjmand, E., Weckman, G. R., & Young, W. A. (2017). An overview of energy demand forecasting methods published in 2005–2015. *Energy Systems*, *8*(2), 411–447. https://doi.org/10.1007/S12667-016-0203-Y/TABLES/8

González-Romera, E., Jaramillo-Morán, M. Á., & Carmona-Fernández, D. (2007). Forecasting of the electric energy demand trend and monthly fluctuation with neural networks. *Computers & Industrial Engineering*, *52*(3), 336–343. https://doi.org/10.1016/J.CIE.2006.12.010

Goudarzi, S., Anisi, M. H., Kama, N., Doctor, F., Soleymani, S. A., & Sangaiah, A. K. (2019). Predictive modelling of building energy consumption based on a hybrid nature-inspired optimization algorithm. *Energy and Buildings*, *196*, 83–93. https://doi.org/10.1016/J.ENBUILD.2019.05.031

Hyndman, R.J., & Athanasopoulos, G. (2015). Forecasting: Principles and Practice (3rd ed). In *OTexts: Melbourne, Australia. OTexts.com/fpp3*.

Li, Y. (2019). Prediction of energy consumption: Variable regression or time series? A case in China. *Energy Science & Engineering*, *7*(6), 2510–2518. https://doi.org/10.1002/ESE3.439

Murti, M. A., Setianingsih, C., Narendra, I., Angelo, K., Aryomukti, M., Bazwir, A., & Kurniawan, R. N. (2022). Forecasting Electricity Consumption using Long Short Term Memory and Prophet Algorithm. *Proceedings of the 2022 IEEE International Conference on Internet of Things and Intelligence Systems, IoTaIS 2022*, 376–381. https://doi.org/10.1109/IOTAIS56727.2022.9975971

O GIG o Gestor Integrado de Riscos e Garantias no SEN e no SNG | GESTOR INTEGRADO DE GARANTIAS. (n.d.). Retrieved February 27, 2023, from https://www.gigenergia.pt/pt/o-gig-o-gestor-integrado-de-riscos-e-garantias-no-sen-e-no-sng

Rodrigues, F., Cardeira, C., & Calado, J. M. F. (2014). The Daily and Hourly Energy Consumption and Load Forecasting Using Artificial Neural Network Method: A Case Study Using a Set of 93 Households in Portugal. *Energy Procedia*, *62*, 220–229. https://doi.org/10.1016/J.EGYPRO.2014.12.383

APPENDIX

```
# Packages
library(tidyverse)
library(IRdisplay)
library(ggplot2)
library(fpp2)
library(forecast)
library(xts)
library(forecast)
library(lubridate)
library(scales)
library(knitr)
library(data.table)
library(scales)
library(ggplot2)
library(prophet)

# Read data
df = read.csv("ConsumoEDP.csv", header = T, sep = ";", stringsAsFactors
= F)
head(df,10)
summary(df)

# Define data type
df$TIMESTAMP <- as.POSIXct(df$TIMESTAMP, '%d/%m/%Y %H:%M', tz = "GMT")
df$CONSUMPTION.MWh = as.numeric(gsub(",",".",df$CONSUMPTION.MWh))

# Exploratory Analysis
ggplot(data = df, aes(x = TIMESTAMP, y = CONSUMPTION.MWh))+
  geom_line(color = "#0c4c8a", size = 0.3) +
  xlab('Date') + ylab('Consumption, MW/h') + scale_y_continuous(labels
= label_number(suffix = " M", scale = 1e-6))  +
  theme(panel.grid.major = element_blank(), panel.grid.minor =
element_blank(),
        panel.background = element_blank() ,axis.line =
element_line(colour = "black"))

ggplot(data = df[df$TIMESTAMP >= "2021-01-01" & df$TIMESTAMP <= "2021-
12-31",], aes(x = TIMESTAMP, y = CONSUMPTION.MWh))+
  geom_line(color = "#0c4c8a", size = 0.3) +
  xlab('Date') + ylab('Consumption, MW/h') + scale_y_continuous(labels
= label_number(suffix = " M", scale = 1e-6))  +
  theme(panel.grid.major = element_blank(), panel.grid.minor =
element_blank(),
        panel.background = element_blank() ,axis.line =
element_line(colour = "black"))

# Exploratory Analysis - weekdays
weekday <- c('Monday', 'Tuesday', 'Wednesday', 'Thursday', 'Friday')
df$DofWeek <- c('weekend', 'weekday')[weekdays(df$TIMESTAMP) %in%
weekday+1L]
# create mean by group
mean <- df%>%
group_by(DofWeek,hour(TIMESTAMP))%>%summarise(mean_val=mean(CONSUMPTION
.MWh))
colnames(mean)=c('DayOfWeek','x','y')
```

```
ggplot(mean, aes(x=x,y=y,col=DayOfWeek)) + geom_line() +
  scale_y_continuous(name='Consumption, MW/h',labels =
label_number(suffix = " M", scale = 1e-6)) +
  scale_x_continuous(name='Hour', breaks=0:24)  +
  theme(panel.grid.major = element_blank(), panel.grid.minor =
element_blank(),
        panel.background = element_blank() ,axis.line =
element_line(colour = "black"))

#Data Quality
# missing values
df[!complete.cases(df),]
df=df[complete.cases(df),]
df=df[order(df$TIMESTAMP),]
#outliers
ggplot(df) +
  aes(x = CONSUMPTION.MWh) +
  geom_histogram(bins =20L, fill = "#0c4c8a")+ ylab('Frequency') +
  scale_y_continuous(name ='Frequency', labels = label_number(suffix =
" K", scale = 1e-3)) +
  scale_x_continuous(name ='Consumption, MW/h', labels =
label_number(suffix = " M", scale = 1e-6)) +
  theme(panel.grid.major = element_blank(), panel.grid.minor =
element_blank(),
        panel.background = element_blank() ,axis.line =
element_line(colour = "black"))
#-IQR
ggplot(df) + geom_boxplot(aes(x = month(TIMESTAMP), y =
CONSUMPTION.MWh, group = month(TIMESTAMP))) +
  scale_y_continuous(name='Consumption, MW/h',labels =
label_number(suffix = " M", scale = 1e-6)) +
  scale_x_continuous(name='Month', breaks=1:12, limits=c(0, 13)) +
  theme(panel.grid.major = element_blank(), panel.grid.minor =
element_blank(),
        panel.background = element_blank() ,axis.line =
element_line(colour = "black"))

out <- boxplot.stats(df$CONSUMPTION.MWh)$out
out_ind <- which(df$CONSUMPTION.MWh %in% c(out))
length(out_ind)/length(df$CONSUMPTION.MWh)

df <-df[-out_ind,]

##MODELLING
# Storing the data in a time series object
df$TIMESTAMP <- ymd_hms(df$TIMESTAMP) #datetime format
ts_train<-df$CONSUMPTION.MWh %>% ts(freq= 24) #specifying the number of
times that data was collected

## Estimating the trend component and seasonal component of our data
ts_train %>%
  tail(24*7*6.4) %>% #hour, weekly and yearly
  decompose() %>%
  autoplot()

#Decomposing our data using mstl
```

```
msts_power_1 <- msts(df$CONSUMPTION.MWh, seasonal.periods =
c(24,24*7,24*365.25), start = decimal_date(as.POSIXct("2015-01-01
00:00:00")))
msts_power_1 %>% mstl() %>% autoplot()+
  scale_x_continuous(name='Year', breaks=2015:2022) +
  scale_y_continuous(name='Consumption, MW/h',labels =
label_number(suffix = " M", scale = 1e-6)) +
  theme(panel.grid.major = element_blank(), panel.grid.minor =
element_blank(),
        panel.background = element_blank() ,axis.line =
element_line(colour = "black"))

# MODELLING
#Dividing our data into train and test
df_train= df[df$TIMESTAMP <= "2020-12-31",]
df_test= df[df$TIMESTAMP >= "2021-01-01",]

head(df_train,10)
nrow(df_test)/nrow(df)

#### FORECASTING

msts_power <- msts(df_train$CONSUMPTION.MWh, seasonal.periods =
c(24,24*7,24*365.25), start = decimal_date(as.POSIXct("2015-01-01
00:00:00")))
#Mean forecasting
mean_baseline <- meanf(msts_power,h=12383)
summary(mean_baseline)

autoplot(mean_baseline) + ggtitle('Mean - Forecast')+
  scale_x_continuous(name='Year', breaks=2015:2022) +
  scale_y_continuous(name='Consumption, MW/h',labels =
label_number(suffix = " M", scale = 1e-6))

accuracy(mean_baseline,df_test$CONSUMPTION.MWh)
#checkresiduals(mean_baseline)

#Naive Forecasts
fcast_naive <- naive(msts_power,h=12383)
summary(fcast_naive)

autoplot(fcast_naive) +ggtitle('Naive- Forecast')+
  scale_x_continuous(name='Year', breaks=2015:2022) +
  scale_y_continuous(name='Consumption, MW/h',labels =
label_number(suffix = " M", scale = 1e-6))

accuracy(fcast_naive,df_test$CONSUMPTION.MWh)

#Snaive Forecasts
fcast_snaive <- snaive(msts_power,h=12383)
summary(fcast_snaive)

autoplot(fcast_snaive) +ggtitle('SNaive- Forecast')+
  scale_x_continuous(name='Year', breaks=2015:2022) +
  scale_y_continuous(name='Consumption, MW/h',labels =
label_number(suffix = " M", scale = 1e-6))

accuracy(fcast_snaive,df_test$CONSUMPTION.MWh)
```

```
#ARIMA
#model = auto.arima(msts_power,seasonal=TRUE)
#arimaa=forecast(model,h=12383)
#plot(arimaa)

#Dynamic regression with Arima
fourier_power <- auto.arima(msts_power, seasonal=FALSE, lambda=0,
                            xreg=fourier(msts_power, K=c(10,10,10)))

f_fourier <-  forecast(fourier_power, xreg=fourier(msts_power,
K=c(10,10,10), h=12383))
f_fourier
autoplot(f_fourier) +
  ylab("Power Consumption predicted") + xlab("Time")

accuracy(f_fourier,df_test$CONSUMPTION.MWh)

#mstl
mstl_power <- mstl(msts_power)
f_mstl <- forecast(mstl_power, h = 12383)
autoplot(f_mstl)

#Prophet
colnames(df_train) <- c('ds','y')
fit_prophet <- prophet(df_train)
future_df <- data.frame(df_test$TIMESTAMP)
colnames(future_df) <- 'ds'
f_prophet <- predict(fit_prophet,future_df)

plot(fit_prophet,f_prophet)  +
  scale_x_continuous(name='Year', breaks=2015:2022) +
  scale_y_continuous(name='Consumption, MW/h',labels =
label_number(suffix = " M", scale = 1e-6))

prophet_plot_components(fit_prophet,f_prophet)#decomposition


## Comparing the accuracies
mean_results <-accuracy(mean_baseline,df_test$CONSUMPTION.MWh)
naive_results <- accuracy(fcast_naive,df_test$CONSUMPTION.MWh)
snaive_results <- accuracy(fcast_snaive,df_test$CONSUMPTION.MWh)
tbats_results<- accuracy(f_tbats,df_test$CONSUMPTION.MWh)
prophet_results<- accuracy(f_prophet$yhat, df_test$CONSUMPTION.MWh)


Summary_table=
data.table(rbind(mean_results,naive_results,snaive_results,tbats_result
s))
Summary_table[,Split:=c("Train","Test","Train","Test","Train","Test","T
rain","Test")]
Summary_table[,Method:=c(rep("Mean",2),rep("Naive",2),rep("Snaive",2),r
ep("TBATS",2))]
kable(Summary_table)

#########-------------------------------------------------------------
#########
#APPROACH CHOOSEN!
```

```
df_final = df
colnames(df_final) <- c('ds','y')
fit_prophet <- prophet(df_final)

NoOfHours <- as.numeric(ymd_hms("2022-08-31 21:00:00") - ymd_hms("2022-
05-31 24:00:00"))*24
ts=seq(from = as.POSIXct("2022-05-31 24:00", tz="GMT"), length.out =
2205, by = "hours")
next90Days <- data.frame(ds = ts)

f_prophet <- predict(fit_prophet,next90Days)

plot(fit_prophet,f_prophet)  +
  scale_x_continuous(name='Year', breaks=2015:2022) +
  scale_y_continuous(name='Consumption, MW/h',labels =
label_number(suffix = " M", scale = 1e-6))

future = f_prophet %>% select(ds,yhat)
colnames(future) = c('TIMESTAMP','CONSUMPTION.MWh')
final_df = rbind(df,future)
summarise(final_df)
head(final_df)

write.csv(future,"/ForecastConsumoMWh.csv", row.names = FALSE)

#90 DAYS MOVING AVERAGE

#get df by day
final_df$Datee = format(final_df$TIMESTAMP,"%d/%m/%Y", tz="GMT")
Daily_df =
aggregate(final_df$CONSUMPTION.MWh,by=list(Date=as.Date(final_df$Datee,
format = "%d/%m/%Y")),FUN=sum)
Daily_df=Daily_df[order(Daily_df$Date),]
head(Daily_df)
colnames(Daily_df) <- c('Date','MWh')

Daily_df$Billing = Daily_df$MWh*0.0340
#calculate MA 90D
Daily_df <-
  Daily_df %>%
  mutate(MA90D = rollmean(Billing, k=90, fill=NA, align='right'))


Daily_df$CONTUR=Daily_df$MA90D * (30+5+2)
Daily_df_Final = Daily_df

write.csv(Daily_df_Final,"FinalData.csv", row.names = FALSE)
```

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação

Universidade Nova de Lisboa