INF592: Internship in data science

Mrs. I. Manolescu & Mr. A. Astolfi

Internship Report

14th of June 2021

Ardian Infrastructure

Côme de Germay

Contents

1 Introduction

1.1 The Company

Ardian is a worldwide leader in Private Equity that has currently over $112 billion in assets under management from the public sector and private sector. Although originally French, Ardian has over 15 offices across the world and currently employs over 710 people (including 265 investment professionals). Ardian is divided into various divisions investing in different assets (e.g. Growth invests in profitable startups, Fund of Funds invests in funds holding stakes in different companies, ...) [1].

1.2 Infrastructure within Ardian

I am interning in this division (40 people worldwide) from March 2021 to August 2021. Ardian Infrastructure oversees over $17 billion in assets, including:

- transports: world's second-largest toll road operator by network size, and with assets covering France, Italy, Spain, Portugal, Brazil, and Chile; operates high-speed rail lines; and airport holding company with interests in six Italian airports including Milan Malpensa and Linate, Turin and Naples.
- energy: invests significantly in Europe and America and is building major clean energy companies in the US (Skyline Renewable) and Scandinavia (eNordic); operates wind and solar assets in Spain, Italy, Chile and Peru, as well as oil, gas and petrochemical distribution networks in Europe and the US (+3.5GW of sustainable energy by 2020).
- telecommunications: investments include a 30.2% stake in the joint control of INWIT, Italy's leading telecoms tower operator, and a 26% stake in EWE, one of Germany's largest utilities and a leading provider of telecommunications services.
- social infrastructure: investments include water and waste treatment assets; majority stake in Holding di Investimenti in Sanità e Infrastrutture (Hisi), which indirectly holds a concession to operate 1,000-bed hospitals in Lombardy and the Piedmont region of Italy over the long term.

Since most of these investments are performed in the long term (on average for 15 years) and that these sectors are recording more and more industrial data, Ardian Infrastructure decided to launch various research projects, aiming at improving productivity and profitability of its assets including:

- Car Carbon: ESG initiative, leveraging airports' proprietary live stream of traffic data as well as external data on aircraft carbon emission could allow building a carbon emission calculator for the airports, running in real-time.
- Electricity Price Forecasting: relies on the extensive volume and diversity of data communicated by the electricity market regulator, as well as other data sources we may find, to get a better view of the price level at a short-to-long-term horizon. Anticipating extreme price levels would allow Ardian to hedge the risk beforehand.

1.3 My work within Ardian

While originally my project was to build a company analysis tool that would allow for a fast overview of a targeted business, upon discussion with the team at Ardian, we concluded I could

work on another research project additionally (I had started the company analysis tool before joining Ardian).

The overall goal of this project is to develop a predictive analysis tool that will forecast the price of electricity in Italy. It is a project that was launched on my arrival, hence I started it from scratch and was conducted as a research project (with the aim to exploits the results on Ardian's Italian renewable assets). See Figure 1 for the project outline:

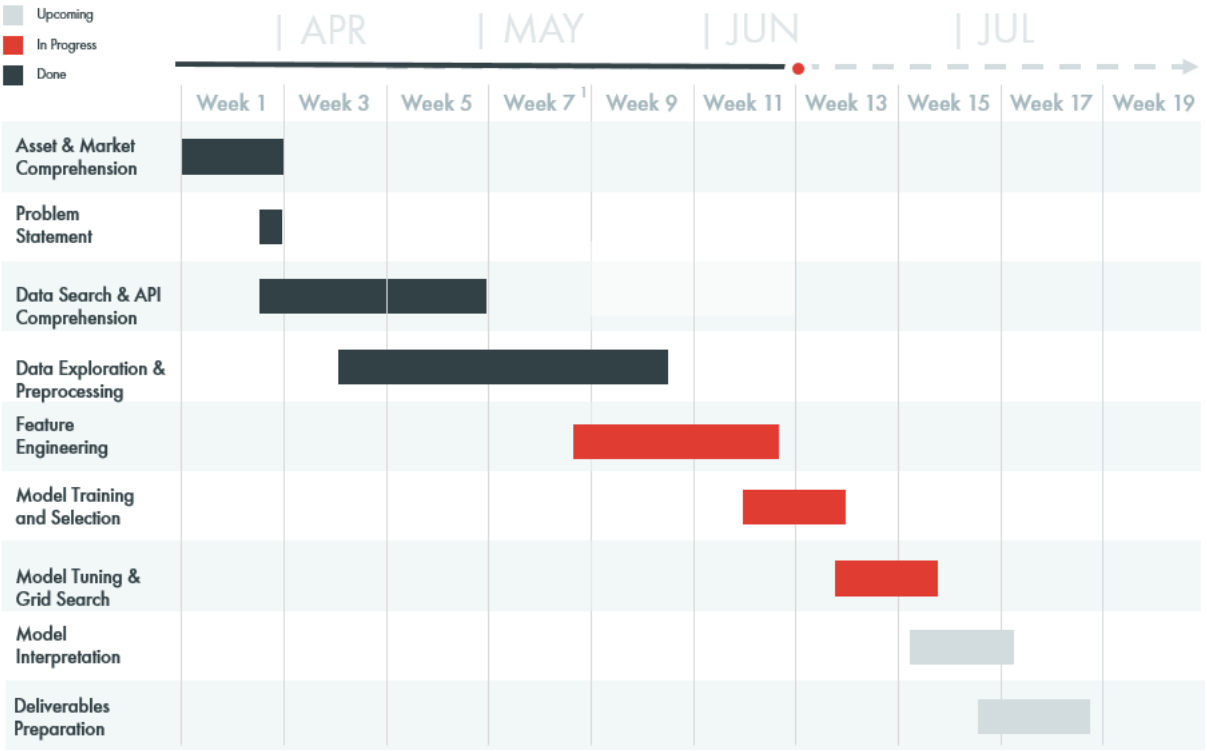

Figure 1: Project's Gant chart

Since these project's results proved to be more interesting, I decided that this project will be the subject of the following report and the upcoming presentation.

2 Context & Problem framing

2.1 Asset Overview

As part of its energy investment portfolio, Ardian has invested in three wind farms in the Basilicata region (South of Italy) that have been under-profitable recently. Part of it is due to large imbalance costs paid by the generators. Indeed, as will be detailed in the next section, generators can either receive (or pay) a premium (or fee) depending on their production and the imbalance sign (if the production of the macrozone is superior or not to the demand in the macrozone) of the macrozone (to facilitate electricity markets, the country has been divided into 2 macrozones (North and South)).

Figure 2: Location of Ardian's three wind farms in the South macrozone

2.2 Electricity production and demand in Italy

As per Figure 3, around 75% of the electricity consumption depends on the country's economic health and activity, hence analyzing both the Italian macroeconomics and seasonality seems relevant when investigating the electricity price dynamics.

As part of the European Green Deal, the National Energy and Climate Plans forecasts that renewable energy will account for over 55% of the total demand for electricity by 2030 (vs 36% in 2019). Furthermore, photo-voltaic and wind demand should be multiplied by 3 and 2 respectively.

This will reduce the contestable market for traditional thermal plants, that however will still represent 35% of the production in 2030.

Investigation of the production by generator types (in 2018, last available data) proved that 40% of the Italian electricity is produced by renewable plants (wind, hydro, biomass, and solar).

However, among those, hydro, solar, and wind productions are highly dependent on key weather factors (radiation, pressure, temperature, cloud coverage, wind, precipitation, ...) which makes them challenging to predict, which in turn explains why the actual production often differs from the expected one. Thus, since demand is increasingly met by RES, which have unstable and unpredictable supply patterns, renewable generators will face stronger fees in the future (if no change of regulation is done).

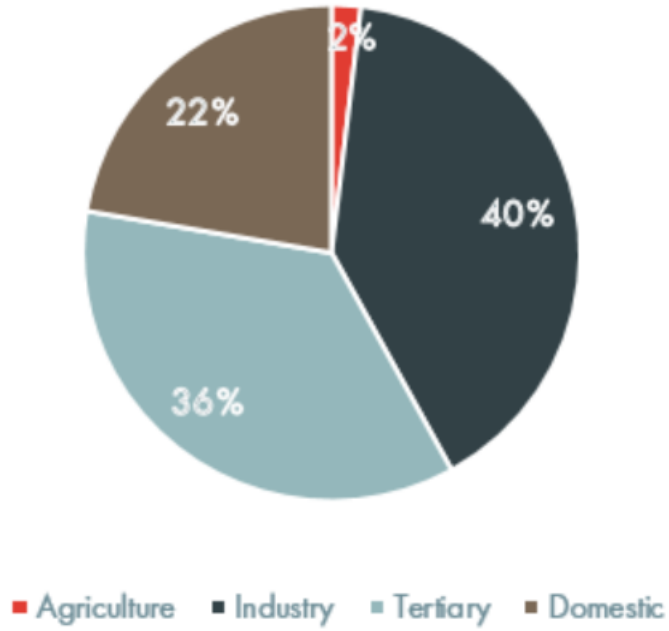

Figure 3: Italy: countrywide electricity consumption by sector (2018)

2.3 Market overview

The day ahead market (MGP):
The MGP opens 9 days before delivery time and closes 24 hours before it (the delivery time) [2]. It is the main electricity market in Italy. Based on an auction principle, each member is allowed to place a bid (maximum/minimum price and volume) for which he/she will buy/sell electricity to other parties. Computed hourly, the MGP sets the price of electricity at the end of each session by determining the last accepted price on the market (i.e. the Marginal Price). Due to this auction mechanism, prices are highly volatile and thus strongly impact profitability on the generator side.

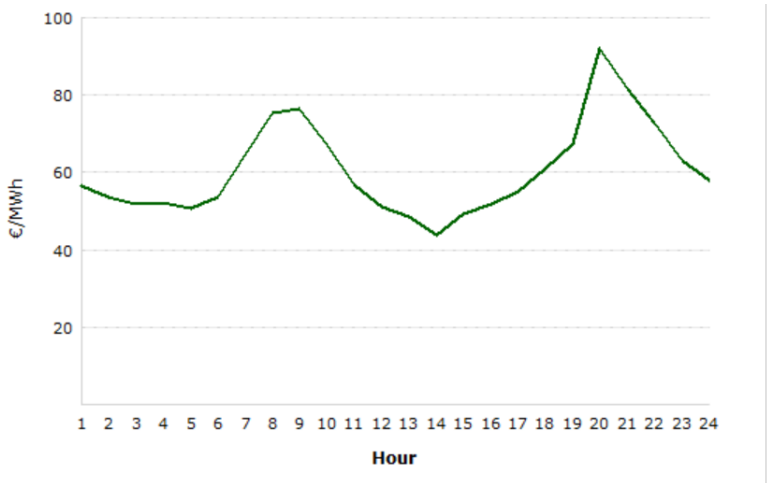

Figure 4: MGP Prices on March 31, 2021

The intraday market (MI):

Once the MGP closes, the intraday opens. It is subdivided into 5 markets: MI1 to MI5 (upon discussion with experts in the field, we concluded we could regroup some intraday sub-markets together for simplification purposes (actual total of 7 intraday submarkets)). The intraday allows for each market user to add bids/offers closer to the delivery time, hence when it has a better prediction of its production. It acts as a way to modify the bids/offers submitted on the MGP. Since it also follows the same auction mechanism, the prices are set and available at the end of each session.

The dispatching service's market (MSD):

While GME (the state-owned organization responsible for operating power, gas, fuel, and environmental markets in Italy) operates the Italian energy markets, Terna (a Transmission System Operator) manages the reliable transmission of electricity across the grid. Terna has to ensure that the consumer's demand is matched at all times and that the network is not unbalanced (i.e. that no particular area is in oversupply or under-supply of electricity). As explained above, due to lack of storage, operators have to predict their production and since 40% of them are producing renewable energy, their predicted production often differs from the actual production, leading to imbalance. Hence, the MSD acts as the first layer of balancing the market. There, operators can help Terna to balance the grid by either inputting more electricity than they were supposed to (when the market is in electricity shortage) or by requiring more electricity (when the market is in oversupply of electricity) [3].

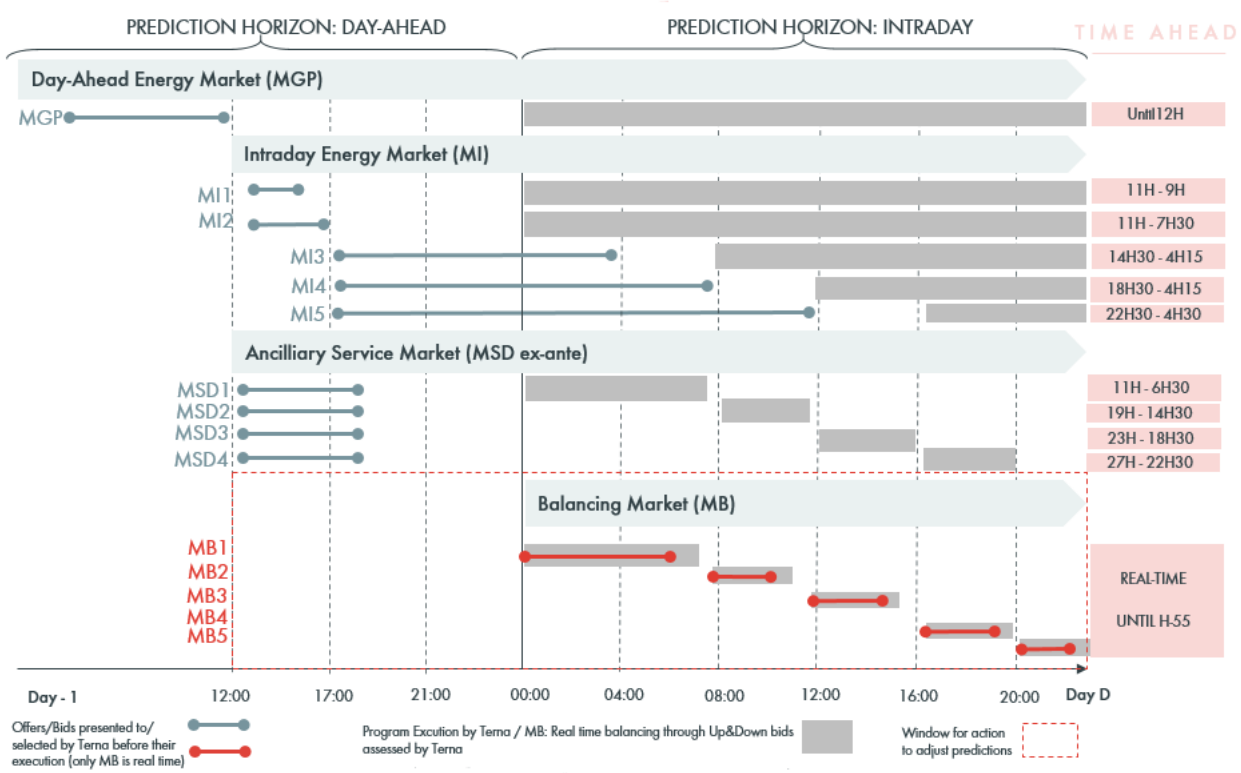

Figure 5: Electricity spot market organization

2.4 The balancing market & the imbalance price in Italy

As discussed above, since Terna is in charge of securing the system, every time the system is imbalanced it relies on the MSD market to restore the macrozone's equilibrium and then in real-time, it uses the balancing market (MB) to make final adjustments to the market to satisfy demand.

After real-time, Terna charges or pays generators depending on whether they are in undersupply or oversupply of electricity, respectively. These payments or charges (known as imbalance settlement prices and computed hourly) also depend on the macrozone imbalance state. Imbalance settlement prices are the prices perceived on the delta of production between forecasted and realized production.

Hence, Terna defines the following four configurations for the generator [4]:

Figure 6: Imbalance price matrix

- Case A: The generator, given it is in oversupply, contributes to the positive imbalance of the macrozone. As a result, the imbalance price is penalizing since they usually receive a price lower than the market price.
- Case B: The generator, given it is in undersupply, helps the macrozone to exit from the positive imbalance situation. Even though the generator was imbalanced, the imbalance price is not penalizing since the generator is charged for the quantity it was not able to produce at a price lower than the market.
- Case C: The generator, given it is in oversupply, helps the macrozone to exit from the imbalance unbalancing situation. Even though the generator was imbalanced, the imbalance price is not penalizing since the generator can sell its surplus quantity at a higher price.
- Case D: The generator, given it is in undersupply, contributes to the negative imbalance of the macrozone. As a result, the imbalance price is penalizing since the generator is charged at a price higher than the market price for the quantity it was not able to produce.

2.5 Problem framing

Since Ardian is an investment firm that promises its investors high returns it always aims at improving the profitability of its assets. This philosophy also applies to its renewable assets, including these wind farms.

As discussed above, the profitability of electricity producers is impacted either positively or negatively depending on the imbalance costs.

Hence, this project was launched to minimize these imbalance costs through data science techniques. Since it is a wide topic and there are various ways of improving imbalance costs, first we had to determine the target variable of this project.

2.5.1 Target variable

A direct prediction of the hourly value of the imbalance price, or an approximated value of it for each hour (since the market is traded on an hourly basis) through a regression would be more suited for this task as it would allow a direct optimization of these costs. However, due to data and time constraints (since the project is supposed to last my internship's length), a classification task should be considered first: predicting the sign of the imbalance of the macrozone, while the regression could be the object of a following study. Based on the predicted imbalance sign, the management could already alter its production forecasts on the day ahead market and benefit from a strong reduction in imbalance costs, improving profitability at the asset level.

Upon completion of my internship the following two projects will be undertaken to complete the wider task: first a prediction of the imbalance price through regression and then an optimization project (using game theory) to adapt production strategy according to the predicted imbalance costs.

Since the imbalance price is traded on an hourly basis, we had to define a prediction time (i.e how much time before delivery time) for the imbalance sign of the macrozone.

2.5.2 Prediction hour

One could determine that the appropriate horizons for prediction and bringing higher value on the asset side would be from 1-days-ahead (when the MGP is still open) to intraday (when the MI is still open). In practice, long-term (more than 2 days) forecasts might be too uncertain to give reliable results (the closer the horizon, the safer the weather forecast), hence we chose to perform the predictions 24H before delivery time (on the MGP), at 11 am of the D-1. This will allow for the generator to adapt its bid before the closure of the MGP market (at midday of D-1).

2.5.3 Classification considerations

A macrozone imbalance consists of a difference between the sum of the scheduled productions of a macrozone's generators and their actual production, hence, from a mathematical point of view, a three-way classification task should be envisioned to reflect negative imbalance, positive imbalance, and neutral (when production scheduled equals actual production). However, following both a data analysis of the distribution of the volume of imbalance and discussions with renewable professionals, we chose not to consider neutral as an option. Furthermore, this design choice highly decreases the model's complexity and widens the potential model choices (e.g. logistic regression).

Upon conclusion of this project framing, we concluded that the task is to first, forecast the imbalance sign of the market, and then in another study, minimize balancing costs at the asset level through a modified production forecasting strategy.

3 Litterature review

Since predicting the imbalance sign has an impact on the profitability at the generator side, it has been the subject of various papers. To obtain the best possible output we chose to leverage these few past analyses by observing which inputs were selected and which type of model architecture was chosen.
However, there are few relevant past studies focusing on the Italian market and adapted to today's regulation. Among the papers we found focusing on the prediction of the imbalance of the electricity market, a few were focusing on other European markets (the UK [5, 6], Belgium [7], Netherlands [8], ...), the following two were focusing on Italy:

3.1 Novel approaches to the energy load unbalance forecasting in the Italian electricity market

This research paper published in early 2017 (February) by Luca Di Persio, Alessandro Cecchin & Francesco Cordoni [11] analyzes statistical properties of the Italian daily electricity load market by evaluating the following models:

- Exponential smoothing model: To observe the impact of past values on the following ones, the team tested exponential smoothing: a model that predicts based on exponentially-weighted average of past observations.

- ARMA-ARIMA model: Since the present analysis is done on time series, the researchers initially studied the AutoRegressive Moving Average (ARMA). As they correctly describe them: "they play a central role because they are capable of describing weakly stationary stochastic processes with a rather restricted set of assumptions, being mainly based on the use of two polynomials: the first one takes into account the autoregressive character of the data set, while the second takes into account the moving average. It is worth to mention that such a method results as a combination of the Moving Average method (MA) together with an AutoRegressive (AR) one".
  In a second time, they studied the ARIMA model: ARMA that allows for non-stationary models to be studied.

- ARIMA-GARCH model with additional features: Version of ARMA where the random noise components are modeled with a Generalized Autoregressive Conditional Heteroscedasticity (GARCH) with other features (ex: temperature, day of the week, ...).

The overall goal of the research is to predict the sign of the macrozone imbalance one day before the delivery time.

The outcome of their analysis leads to the fact that the ARIMA-GARCH performed better than the other ones and hence should be taken into account during our study. They state that further improvement could be done by adding more features as inputs: they only added the day of the week and the temperature. Finally, they raise a point regarding the availability of data and its quality: since their research was done in 2017, before the regulation change, we should not face a similar issue since both GME & Terna changed their reporting standards.

3.2 Analyzing and Forecasting Zonal Imbalance Signs in the Italian Electricity Market

This research paper was submitted by Francesco Lisi & and Enrico Edoli in 2018 [12] (still using data from before the changes in regulation). Their research was directed towards building "a

suitable model for zonal sign dynamics", used to perform "an out-of-sample forecasting exercise concerning the probability of a positive imbalance sign".

The team focused on the following model:

- Binary auto-regressive with exogenous variables: Auto-regressive model that is adapted for binary targets (considered here as the sign of the imbalance: positive or negative) where input data also contains other features: periodicity (hourly, weekly, monthly, yearly, and bank holidays) and the production mix (renewable: wind, solar, hydro vs regular)

At the term of their analysis, the team obtained accuracy above 80% when predicted the next day's imbalance sign in the South macrozone. While it is encouraging for our study, there is no mention of the distribution of their training set (see a future section of this report) and since regulations have changed in 2017, their dataset (using data from c. 2014) does not reflect ours. Nonetheless, it seems that including the energetic mix has a great impact on the outcome: hence it will be one of the features of interest in this study.

Further, comparing these two studies, we noted that past values of the imbalance should be added to our model, either through an auto-regression model or by simply feeding them to another architecture. Thus, this will be a starting point for our model architecture.

4 Data Analysis

As per the Gant chart, since this is a research project, I spent a significant amount of time understanding and selecting the data at hand to make sure that no essential data source was missed. Due to changes in regulations in the Italian market, the latest available data were from September 2017, hence our analysis (for the entire project) begins at this moment.

4.1 Analysis of the target variable

As a starting point, a deep analysis of the historical values of our target variable was performed.

Imbalance data (called Sbil from the Italian sbilanciamento) can be found on Terna's website on an hourly basis by macrozone, since our asset is located in the South, this project focuses on this macrozone.

Performing this analysis, we observed two key elements:

- In the South, the negative imbalance is the minority class and occurs c.30% of the time (vs. 40% in the North), hence leads to class imbalance for the classification task.
- The Sbil (south macrozone) has a mean of 241,81 MWh (positive imbalance refers to a positive Sbil while negative imbalance the opposite). As per Figure 8, one could expect that it follows a normal distribution but it failed all statistic tests and is not a symmetric distribution either.

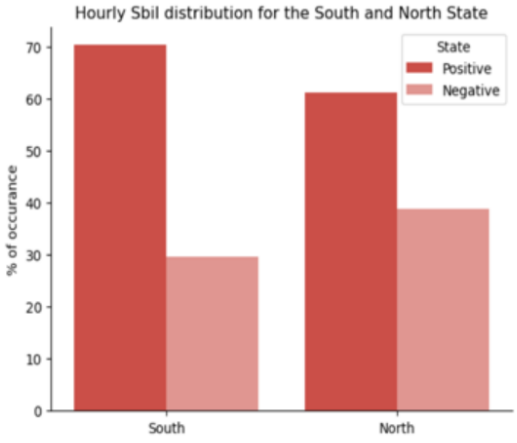

Figure 7: Proportion of time when each macrozone is in positive or negative Imbalance (2017-2021)

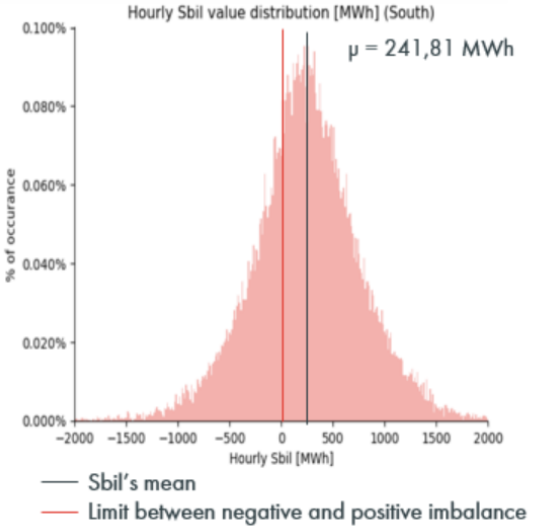

Figure 8: Hourly Sbil distribution for the south macrozone (2017-2021)

Since this project also involves a business strategy component, investigations were made to better understand why is there such an imbalance between the two classes. We concluded that since Terna already performs a first layer of balancing in the MSD before the balancing market, and that it prefers oversupply versus undersupply (from a practical point of view, the grid operator needs to provide electricity for its customer, at the risk of overproducing), the actual behavior seems logical.

4.2 Data Search

Once we performed an analysis of the target variable, we began to perform the data search process to identify which features should be relevant for our task.

4.2.1 Analysis of the seasonality

As per the literature review, the Sbil seems to have a seasonality, hence we chose to first focus on that.

Since the target horizon is a day in advance (on an hourly basis), and that the dataset starts in 2017, we concluded that seasonality could be analyzed at the monthly, weekly, and hourly level (yearly would be not significant since there are only 4 values).

Monthly: To witness this seasonality, we aggregated the hourly values of the Sbil at the monthly level (by average, to discard differences in months' lengths). Due to the results of the production/consumption/demand analysis done previously, we expected to witness differences between each month, reflecting the evolution in economic activities during the year. As per Figure 9, our prediction was correct:

- Negative imbalance occurs most in July and August while positive imbalance is recurrently peaking in April
- But the impact is still minor (-1.29%) compared to the Sbil's average (the black line represents the July and August's mean, while the red line represents the limit between over and undersupply)

Indeed, during the summer months, the demand is more predictable since most of the industrial production is stopped (and does not have the same requirement on electricity). However, we concluded that due to such a small difference compared to the Sbil's average, monthly seasonality is a weak predictor of imbalance variations.

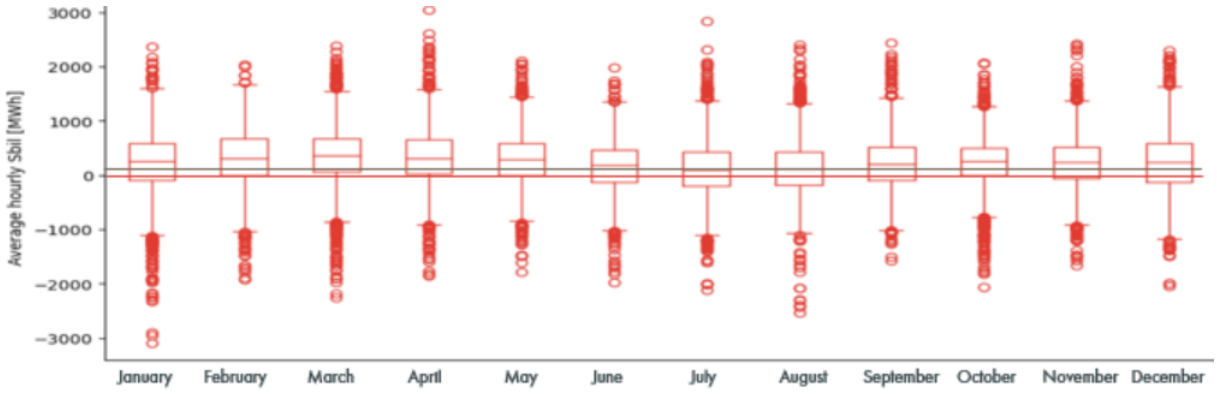

Figure 9: Average hourly Sbil value for each month (in MWh) (South)

Weekly: To witness this seasonality, we followed the same strategy as the one for the monthly seasonality (aggregation of the hourly values of the Sbil at the weekly level through average) for coherence purposes. Regarding the weekdays, no true expectations were expressed before the results were achievable. From the results in Figure 10, we concluded:

- Weekdays seem to have a slight impact on the market imbalance: concentrated increase in positive imbalance in the weekend
- But the impact is still minor (0.85%) compared to the Sbil's average (the black line represents Sunday's mean while the red line represents the limit between over and undersupply)

While for the same reasons as before, a higher positive imbalance on Sunday could be explained by the fact that economic activity is significantly stopped on that day (and does not have the

same requirement on electricity). To confirm these assumptions, we chose to perform further analysis on working vs non-working days seasonality.

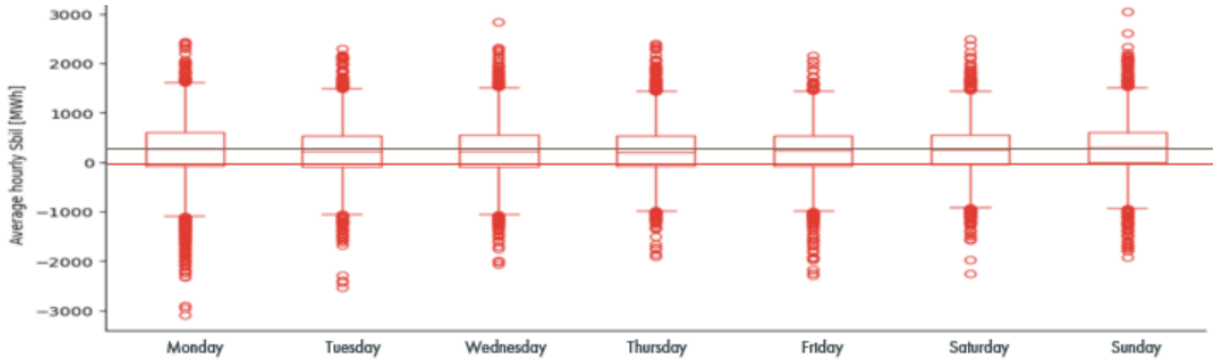

Figure 10: Average hourly Sbil value for each day of the week (in MWh) (South)

Working vs Non-working days: Although the impact of the weekly seasonality is limited, we chose to analyze how different is the Sbil distribution depending on whether it is a:

- Working day: from Monday to Saturday (excluding Italian public holidays)
- Non-working day: Sundays and public holidays

It is interesting to note that during holidays we observed more positive imbalance and fewer negative imbalance severe peaks than during regular working days, which is intuitive as again on holidays and Sundays most of the industry is not producing.

Further, from Figure 11, we concluded that there is a shift between the two distributions (mean of non-working days is 26.31% more than the mean of regular days) which confirms our assumption.

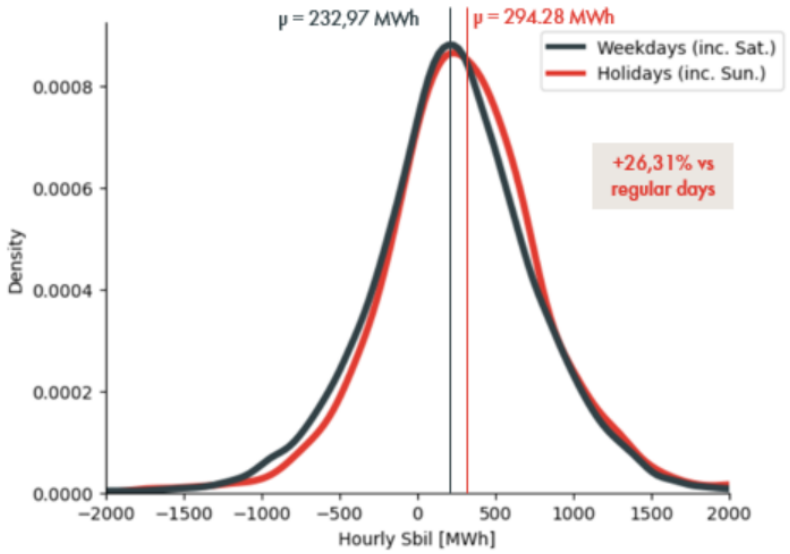

Figure 11: Sbil's distribution between working and non-working days (in MWh) (South)

Hourly: Regarding the lowest level of seasonality, the hours, we aggregated them the same way as for both weekly and monthly seasonality. We obtained the following conclusions from our analysis (see Figure 12):

- Spread of the Sbil is smaller at night and increases during day-time, but the distribution of imbalance looks overall even at a constant 70/30 split at all hours of the day (see the circle on Figure 12)
- Weak hourly seasonality apart from a positive imbalance peak at Midnight (which impact is minor (1.38%) compared to the Sbil's average) (the black line represents midnight's mean while the red line represents the limit between over and undersupply)

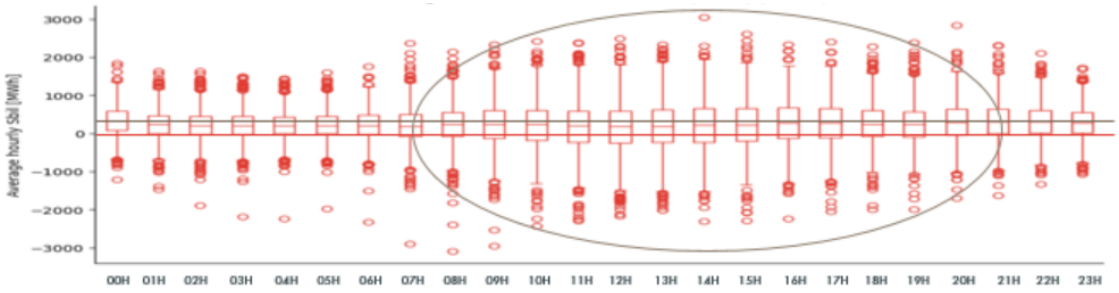

Figure 12: Average hourly Sbil value for each day of the week (in MWh) (South)

Upon conclusion of the analysis on seasonality, we concluded that overall the impact is limited on the evolution of the Sbil. However, we noted that certain key features seem to be relevant (and hence should be included in the model's input list):

- Working and Non-working split
- Hourly division between day-time and night-time

4.2.2 Analysis of the macroeconomics

As discussed previously, the country's economic state seems to have an impact on the level of imbalance at the macrozone level. Hence, we decided to perform an analysis on the relation between the main macroeconomic indicators of Italy and the Sbil level.

The World Bank and the FRED (Federal Reserve Economic Data) provide real-time data points on historical key economic indicators for most countries across the world, including Italy. The data is provided by various sources including Eurostat, OECD, or UNESCO. Employment, consumer confidence, price inflation, housing data, energy, and others are available are examples of indicators available for download.

Indicators selected While there are thousands of indicators available, we searched for low granularity (maximum quarterly) due to the small size of our dataset, and for indicators linked to either energy production/consumption or the general level of the country's economic state.

- Consumer Confidence Index
- Price Deflator
- Consumer Price Index
- Unemployment
- Consumer price index growth rate for energy
- Consumer price index for gas
- Consumer price index for electricity
- Interest rate
- Production and distribution of energy
- Rental prices for residential properties
- Residential properties prices

- Total production of industry
- Total retail trade
- Total exports in goods and services Net savings

Correlation analysis As part of our analysis, we did a correlation analysis of each of the indicators with the aggregated Sbil (at the indicator's granularity).

We observed that their correlation with the Sbil is very limited (at most 19% in absolute value for the Production and Distribution of energy, see Figure 13, the rest is under 10% in absolute value).

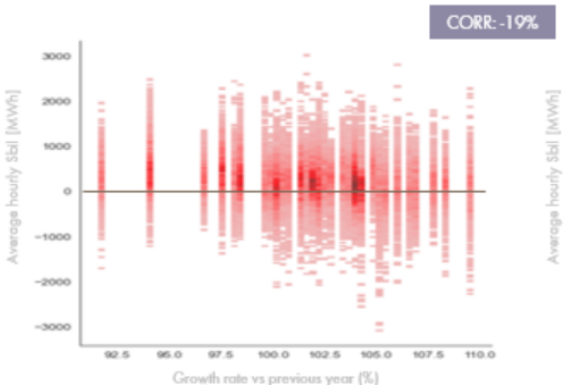

Figure 13: Production and distribution of energy with hourly Sbil correlation plot

Distribution analysis This analysis represents the difference in the distribution of the selected macroeconomic indicators between the top and bottom 50% data points (separated by the median). As per Figure 13 (green boxes refers to over the Sbil median, and red below it), although not strictly statistically significant, a couple of key economic indicators seem to have distinct distributions depending on the imbalance state. However, because next-day predictions are performed, we realized that macroeconomic indicators (even with either monthly or quarterly granularity) will not be fed to the model.

Nevertheless, following our literature review and how performing is auto-regression seems to be, the model will be fed historical Sbil's value, hence it will indirectly take these macroeconomic trends into account. In parallel, the team aims to find proxies (at lower granularity: daily or hourly) to include this information in the model.

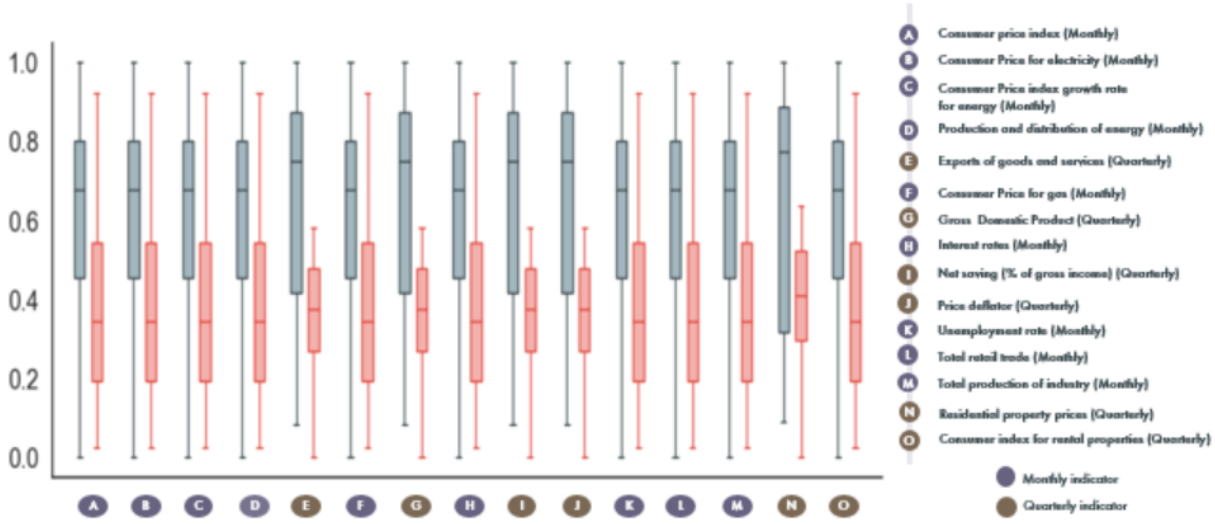

Figure 14: Macroeconomic Indicators value by Imbalance state (aggregated at each indicator's granularity, indicator values normalized to 0 - 1)

4.2.3 Analysis of the market, production, & demand

Before beginning the analysis of the predictive power of each feature, we analyzed the behavior of the imbalance price depending on the sign of the imbalance of the macrozone. This analysis came following the consideration that since negative imbalance occurs only in 30% of the cases, its occurrence could be linked with specific imbalance price patterns.

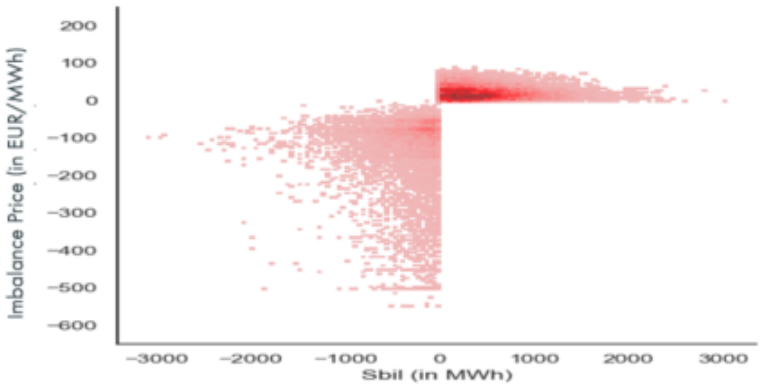

Figure 15: Hourly Sbil vs hourly imbalance price heat-map

As per Figure 15 (that should be read a heat-map where the x-axis represents the volume of imbalance and the y-axis the price associated with the balancing market), imbalance prices are at the maximum when the macrozone is in slight negative imbalance. While this seems counterintuitive (a higher negative imbalance should imply higher prices), upon discussion with people in the industry and further research, this is due to an incentive mechanism created by the grid operator to push generators not to be in negative imbalance). This analysis proves that a correct prediction of the negative class is more important than that of a positive class. This will be taken into account when designing the model structure.

As part of our analysis of the data available, we investigated various key elements of the electricity market and production/demand data.

Market As discussed above, there are various markets before the balancing market and analyzing all these indicators' evolution would have been too time-consuming for the length of the project, hence we decided to focus on the ones that have a direct impact on the value of the imbalance price (the price of the Balancing market).

Initially, a correlation analysis was performed on each of the indicators selected and the imbalance volume. Since our model will predict 24H ahead of delivery time we decided to perform correlation analysis with the 24H shifted Sbil to better reflect the most realistic approach.

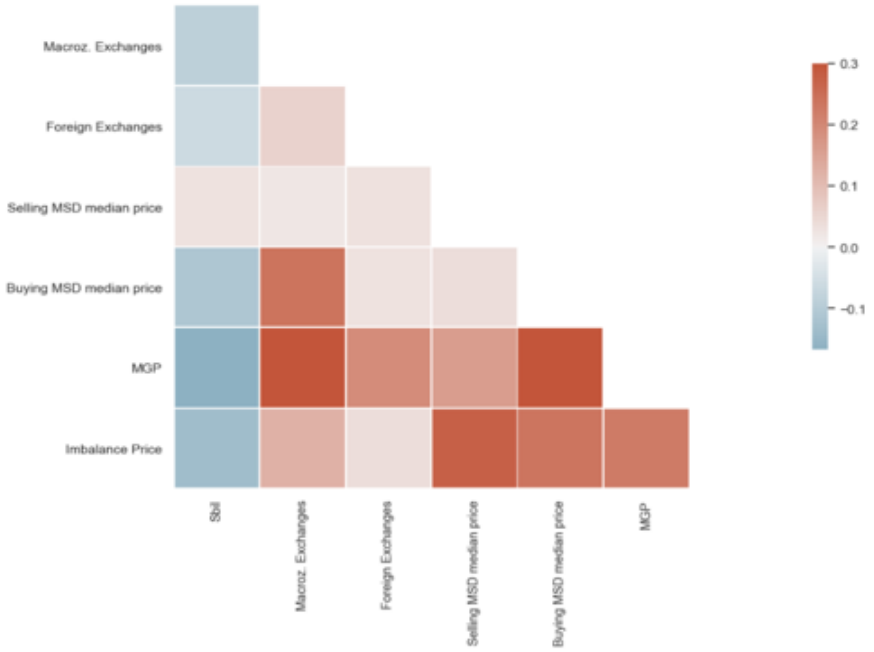

Figure 16: Correlation matrix between key market data and the 24H shifted (ahead) Sbil

From the correlation matrix we notice that overall correlations are very low with the 24H shifted (ahead) Sbil:

- MGP: correlation with the 24H shifted Sbil at -0.1
- As final adjustments to ensure the demand is matched, Terna allows exchanges between macrozones and between countries, hence it should be studied to get a complete view of the market. However, their correlation with the 24H shifted Sbil are both at -0.1
- MSD: correlation with the 24H shifted Sbil at 0.1

Although correlation values do not reflect the impact of a variable in a predictive model, it gives insights on the type of relationship between two variables, we decided not to include these variables in our dataset as we believe that these variables will only induce noise.

Production Since the imbalance volume refers to the delta between actual production and schedule production, production data is a relevant indicator to analyze. While ideally, we would have used the macrozone production data, since it was unavailable, we used the countrywide data as a proxy.

Since the reliability of the generator highly depends on its type, analyzing the production of electricity by the source is relevant as it impacts the macrozone imbalance. Indeed knowing for instance how much of the production is dependent on the wind will be valuable when adding the forecasts of wind in the dataset as wind and imbalance are closely linked. As per Figure 17, we noted two main elements: that hydro production varies across the year following a seasonality pattern and that wind production seems highly unpredictable, hence could impact the imbalance volume largely (although its share of the total production is limited).

Since forecasts are not available, a correlation analysis with the previous day's value is performed to observe if the model can be fed actual value (24H before delivery time) instead of the forecasted value:

- Self consumption: 82%
- Thermal 84%
- Photo-voltaic 94%
- Hydro 93%
- Geothermal 82%
- Wind 55%

As expected all productions are strongly correlated (except for the wind due to its volatility) to their previous day's value and hence can be included in the model's inputs as forecasts.

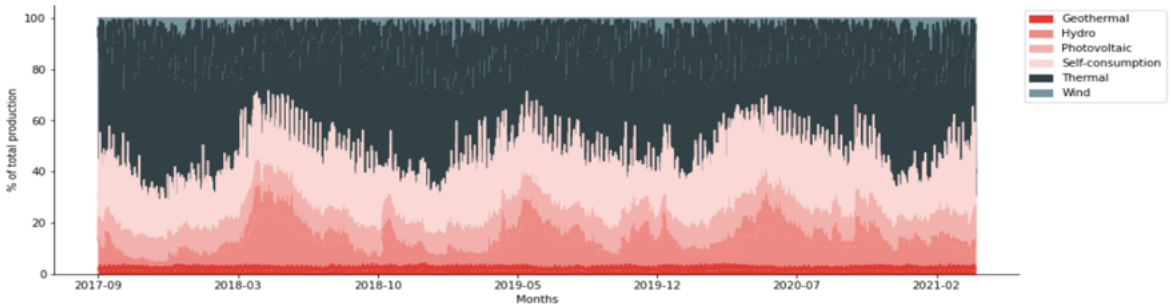

Figure 17: Country-wide production of electricity by source

Demand Regarding the demand, Terna publishes in the morning the load and its forecast (forecast has a 99% correlation with the actual load) (until the end of the day) for each macrozone. Since our prediction is performed at 11 am of the day before delivery, we will not be able to use the latest available forecast for our target timestamp. As for the production, we chose to analyze the correlation with the last day's load and obtained 86%, hence we chose to add it to the model's input list.

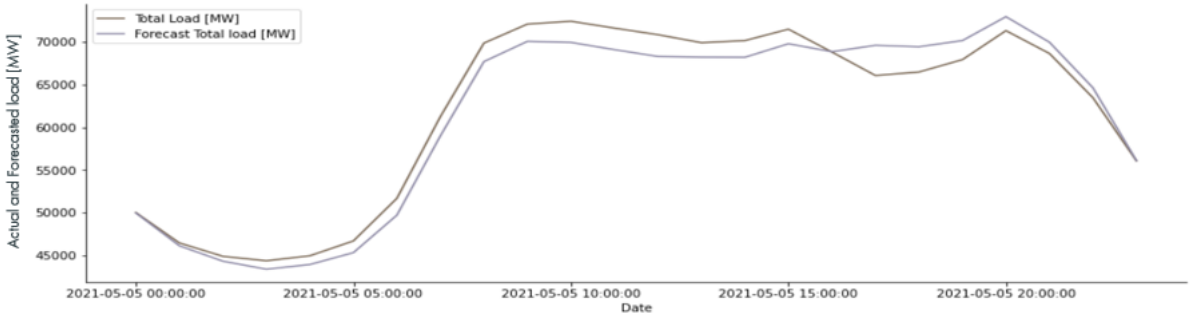

Figure 18: South macrozone actual and forecasted load on the 5th of May 2021

4.2.4 Weather data

As discussed above, most of the imbalance volume derives from the lack of accuracy of production forecasts for renewable generators, which base their predictions on weather forecasts. Hence we decided to add to our model key weather indicators that could either impact the production or the demand. Data was extracted from MARS (service provided by the European Centre for Medium-Range Weather Forecasts) which provides hourly estimates of a large number of atmospheric, land, and oceanic climate variables. It combines vast amounts of historical observations into global estimates using advanced modeling and data assimilation systems. Since the area to cover is the full South macrozone, we defined grids of 110km x 110km where we took the average value for each of the following indicators (if stated otherwise, the indicator is used for production purposes):

- 2m temperature (demand)
- 2m dewpoint temperature
- 10m U wind component
- 10m V wind component
- 100m U wind component
- 100m V wind component
- Total cloud cover
- High cloud cover
- Medium cloud cover
- Low cloud cover
- Snow depth (demand)
- Total precipitation
- Total column ozone
- Surface pressure

Extraction-Preprocessing Weather data is obtained for 96 different square grids across the South macrozone, leading to 1536 values (to become the same amount of columns in the dataset) for each hour for the full macrozone. Depending on whether we qualified the feature as a demand driver or a production driver, we adopted a different strategy to reduce the number of dimensions while retaining most information:

- Demand: PCA was performed using 85% of explained variance to ensure that enough information was kept using a reasonable number of components (see Figure 19 for snow depth)
- Production: performed a weighted average of each weather indicator's value with the amount of electricity produced in the respective grid for the related production type (e.g. for wind production, as per Figure 20 a higher production is observed in the south of Italy, hence wind data should be averaged with a higher weight for Southern regions of Italy than for other regions)

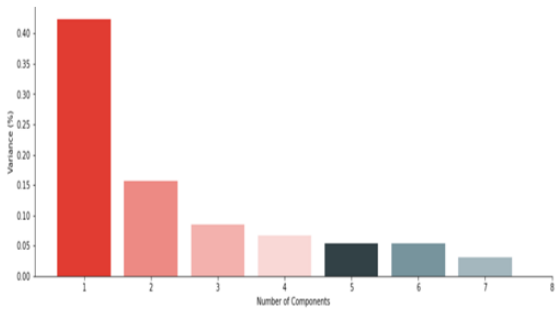

Figure 19: % of explained variance by PCA component for snow depth

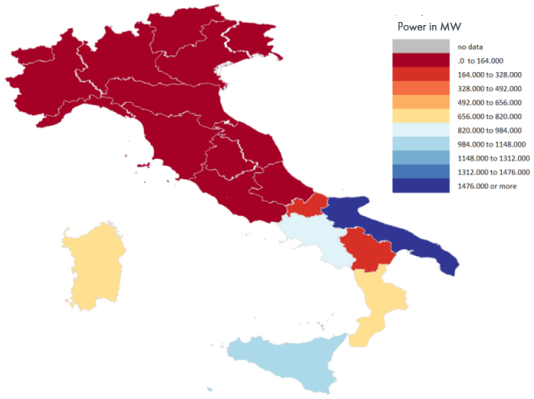

Figure 20: Heat map of the wind production of Italy for each regions

Distribution analysis To have an idea of the behavior of each weather indicator, we analyzed the difference in the distribution of each forecast weather measurements between positive and negative imbalance (see in Figure 21) Although not strictly statistically significant, a couple of key weather measurements seem to have distinct distributions depending on the imbalance state, hence we decided to include weather data in the model's input list.

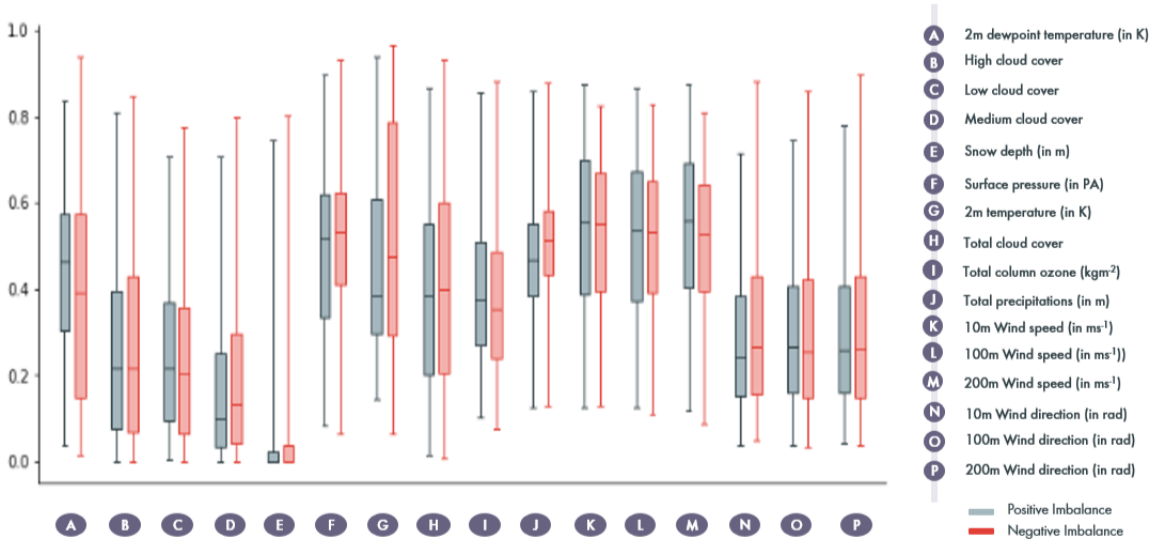

Figure 21: Weather measurements value by Imbalance state (whiskers at 5 and 95 percentiles)

4.2.5 Summary of data analysis

Upon the conclusion of this section, not only do we know which features should be included (seasonality, weather, production, and demand) in the future model but we also have a clearer view of the dynamics of the market. We believe that this deep analysis was required for a thorough feature engineering and a wise model selection and tuning.

5 Models

5.1 Baseline

Initially, we began with a very simple model architecture to have performances to compare with when improving the features/models quality. Regarding the baseline cases, we chose to focus on the accuracy as well as the confusion matrix since we noticed that a correct prediction of the negative class is mandatory (and more important than that of the positive class) to improve profitability at the asset level.

5.1.1 Seasonality model

Upon the analysis based on seasonality, we chose to assess the predictive impact of seasonality on the sign of the macrozone imbalance. This was performed using Prophet (a forecasting tool developed by Facebook that performs time-series seasonality analysis and regression). As per Figure 22, we noted two important elements:

- Seasonality has a weak impact on the Sbil since the model predicts volumes close to the mean of the Sbil
- The training set should be balanced since the current accuracy is biased as there are 71.2% of positive imbalance in the test. Indeed simply predicting the mean of the Sbil outputs the same confusion matrix and accuracy

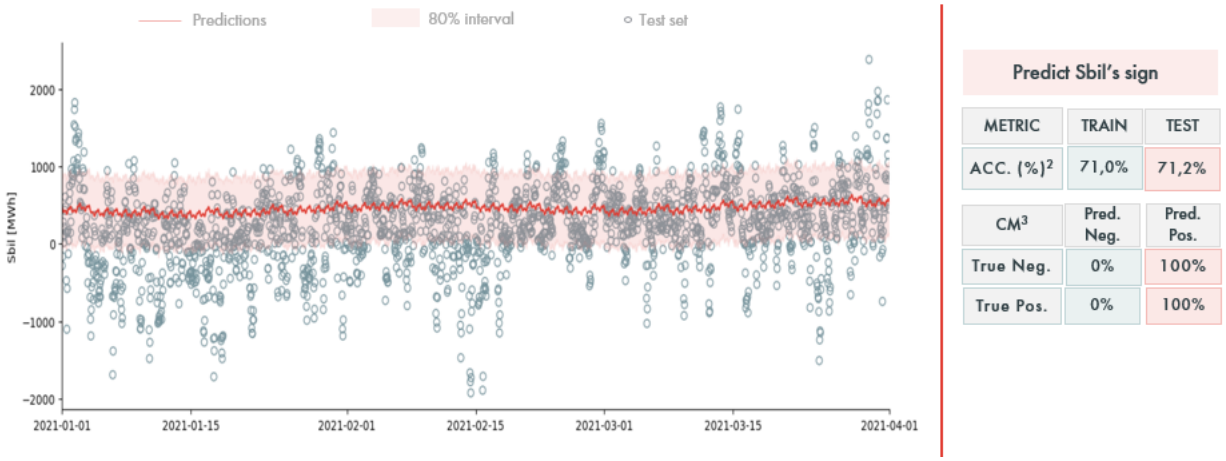

Figure 22: Macrozone imbalance predictions and their respective performances for January to March 2021

5.1.2 Autoregression

As per the literature review, we tested an autoregressive model called Autoreg (based on the Statsmodel library) [14].

Autoregression models an output value based on a linear combination of input values and makes an assumption that the observations at previous time steps are useful to predict the value at the next time step. It could be reduced to the following equation:

$$Y_t = a + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + ... + \beta_p Y_{t-p} + \eta_t \quad (1)$$

Where a is a constant, β are the coefficients of the lags of Y of order p and η_t the error.

This implementation requires the users to input the lag value (i.e. the number of relevant data points to predict the next ones) – p in the above equation. Then, it trains and defines coefficients based on the provided dataset using the Ordinary Least Squares estimation. Finally, it predicts the next value by feeding $\beta_1 Y_{t-1}$ to $\beta_p Y_{t-p}$ (last p available data points in the train set) to the above equation.

By running a first version of an optimization algorithm the value of p was set to 72 hours and we obtained a meaningful baseline:

- The predictions now include negatives, of which 28% are correct: this will be a relevant element of comparison with the following models
- Accuracy has now increased to 73.5% but should not be compared to the following models since the model still trains on a non-balanced dataset

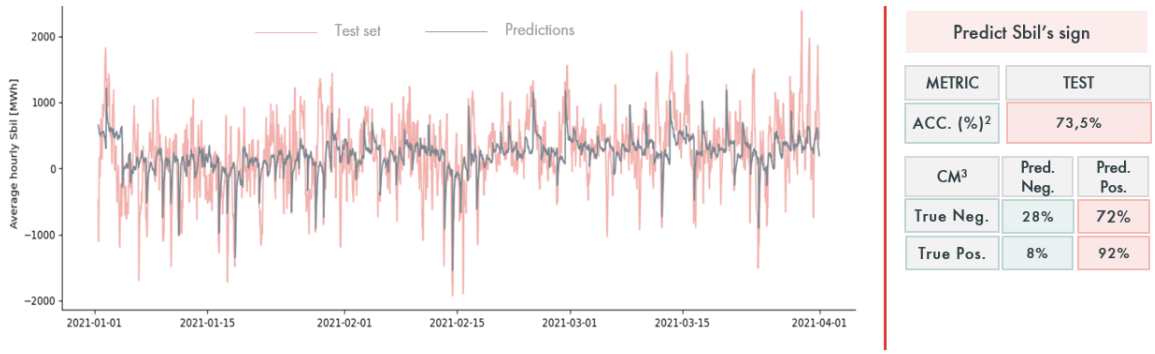

Figure 23: Autoreg model predictions and performances for January to March 2021

5.1.3 Dataset balancing

As discussed above, we believed that the above accuracy is biased due to the current distribution of the dataset, hence we decided to balance it.
There were two possible options to do so: first performing over-sampling which consists of creating new samples for the underrepresented class (negative imbalance in our case) but this is a technique that often results in performances drop or performing under-sampling. It consists of reducing the over-represented class to reach the same amount of samples in each class. Since we have a large dataset, we chose to reduce the positive class by randomly removing positive samples from the dataset.

5.1.4 Including data sources (excluding weather) to the model

From the performances obtained using an autoregression, the team chose to define a first dataset only with the previous values of the Sbil as features. Since the model will have to be trained using non-autoregressive architecture, we concluded it would act as a valuable proxy.
In addition to the past 12 hours of the Sbil (refinement of the number of past values to use), the model was fed the relevant data found during the data analysis part of the project:

- Seasonality: the month, weekday, hour, and holiday vs working day Boolean as one-hot encoding
- Demand: the forecasted demand in electricity for the macrozone
- Generation mix: the split in percentage by production type for Italy (countrywide level)

Since the task is a binary classification, we chose as first model a Logistic Regression (other regressions were tried including Lasso, Ridge, and Linear, however, Logistic obtained the highest results).
As per Figure 24, we obtained the following results:

- 58.22% of accuracy, based on a balanced dataset
- 60% of correct prediction of the negative class (57% for the positives)

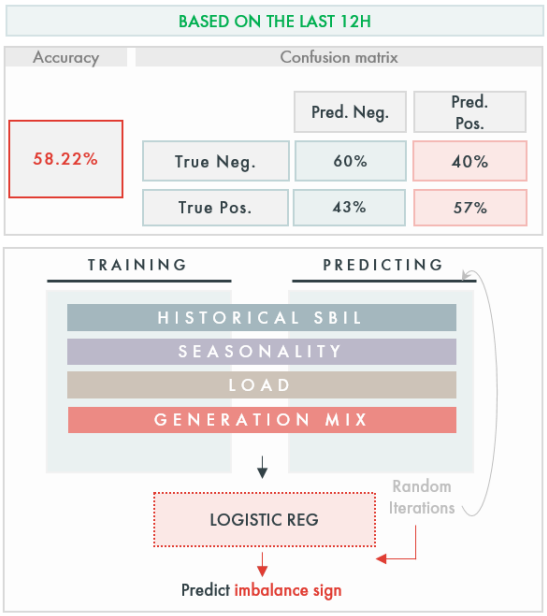

Figure 24: Logistic Regression model predictions and performances for January to March 2021

Performances obtained on this model confirmed that the selected data source (excluding weather) have an impact on the Sbil. It is worth noting that currently, the asset cannot predict the sign of the imbalance of the macrozone, hence it should be assimilated to 50% accuracy. Therefore, business-wise, any improvement past this threshold is a valuable help for the business.

5.1.5 Feature engineering and feature selection

To achieve better performances in a data science task, performing feature engineering is a mandatory step. Although some prepossessing was done before feeding data to the previous model, it would not qualify as a strong feature engineering.

Seasonality From the data analysis part we noted that hours and holidays vs working days seemed to have the most impact on the variations of the Sbil. Regarding hours, two groups were identifiable: day-time (7 am to 9 am) vs night-time, thus, we decided to only keep those two information as well as the working days/holidays feature, hence reducing the number of features fed to the model.

Sbil past values Although 12 hours was the optimal number of hours with a Logistic Regression model, we decided to add three additional features to reflect the overall trend of the last 72 hours: the CAGR, the mean, and the variance of the last 72 hours. Furthermore, we included the number of negative imbalance cases in the last 12 hours (based on the assumption that negative imbalance occurs often in groups). Adding these features gave more valuable information to the model.

Load and Generation mix Regarding the load, we decided to include it in the model using the value of the Sbil: the Sbil divided by the load. By doing so, the model will be able to account for the percentage of demand the imbalance represents. For the generation mix, we decided to keep the split into the 6 different sources as a percentage of the total production of electricity in Italy.

Weather Weather data has already gone through PCA/weighted average, thus we decided not to include further feature engineering since the data is already highly compressed.

5.1.6 Other metrics

Ultimately, the aim is to maximize the correct predictions of the negative without completely neglecting the positive class, with the complexity being: how much more weight to give to the negative class.

Since accuracy does not reflect these constraints we decided to add two additional metrics.

Weighted average accuracy From a business perspective, the simple prediction of the macrozone sign does not take into account that predicting two signs correctly does not have the same impact on the PL of the asset. Since imbalance prices are extremely volatile, a correct prediction of the sign when the imbalance price is at its max/min would be key to knowing the true performances of the model.
To include the volatility of the market, a weighted average of the accuracy of the macrozone sign and the difference between the MGP and the imbalance price (imbalance cost) could be envisaged. It would translate the hedging to be done between negative and positive classes to a more business-wise approach.

$$Y = \sum(\gamma_i - \beta_i)\|P_{MGP} - P_{imbalance}\|_2 \tag{2}$$

While it is an interesting metric, one has to be careful since the metric can be too biased towards the negative sign (since higher deltas in prices are observable in the negative imbalance).

AUC-ROC curve A threshold is the probability value that separates two classes. A key parameter is to set the threshold according to the use case. ROC is a performance metric for the classification problems to test all threshold settings. AUC represents how much the model is capable of distinguishing between classes [15]. Higher the AUC, the better the model is at predicting 0s as 0s and 1s as 1s. By analogy, the higher the AUC, the better the model is at distinguishing between positive and negative macrozone signs.

5.1.7 Latest model architecture

Since this report is written in June and the expected end of the internship is in August, the model is supposed to further evolve.

Once the feature engineering was concluded, other model architectures were designed to further improve performances. Random Forests were tested without large improvements compared to the Logistic Regression, however, Gradient Boosting provided an improvement in performances.

Gradient Boosting parameters Since it is a very powerful model, the XGBoost version allows the user to modify certain parameters. Among those, 4 are key to our task:

- Learning rate: it has a significant impact on the overfitting and thus should be dealt with carefully, not to decrease performances on the test set.
- Number of estimators: it refers to the number of trees used in the model. It also leads to overfitting if the number is too high.
- Maximum depth: it refers to the dept of the trees used in the model: too important leads to overfitting.
- Weight of the positive class: this is the most significant parameter, it gives the amount of weight that should be attributed to each class when training. For our case, we chose to specify that the positive imbalance should be attributed a third of its current weight, hence increasing that of the negatives.

The value of each of these parameters was determined using a Grid-Search (optimization method to find the most performing combination of parameters) to optimize the AUC-ROC.

Model results Including the feature engineered data and feeding it to the tuned Gradient Boosting led to obtaining the following results:

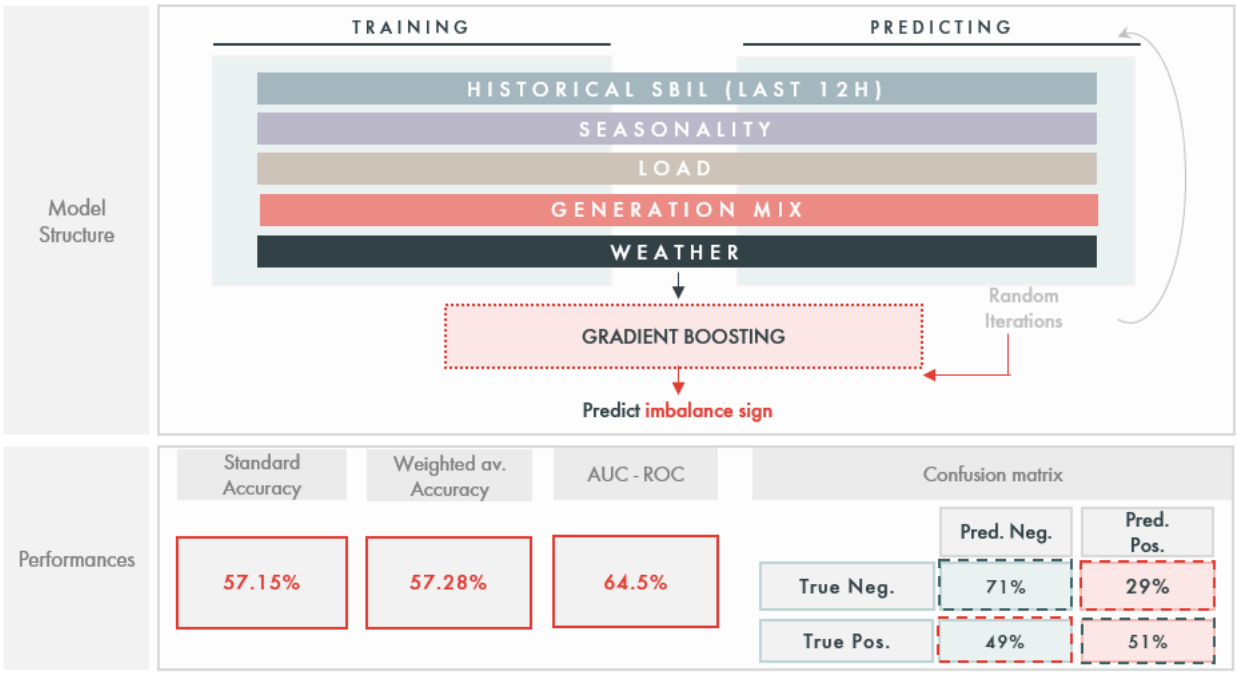

Figure 25: Gradient Boosting model predictions and performances for January to March 2021

Although the accuracy dropped, the correct prediction of the negative class increased to 71% while keeping the correct predictions of the positives above 51%. Further, both our two other metrics increased by adding feature engineering. Finally, we expect these metrics to continue to improve towards the end of the internship.

6 Project Status & Next steps

6.1 Current project status and results

Currently, all the data has been found and analyzed, a first version of the feature engineering was performed, and a model architecture was defined. Feature engineering will be refined within the next weeks to further improve performances leading to potential refinement of the hyper-parameters of the Gradient Boosting architecture.
When applying the results of our model to the electricity selling process of our asset, we obtained a revenue increase of c. 6% which proves that our project has a significant impact on the profitability of our asset. Although this is not the direct goal of the project, within an investment company, it is important to apply theoretical research to improve the Internal Rate of Return (key metric in Private Equity that measures the returns on investment).

6.2 Next steps

This report was written in the first half of June, hence, as per the Gant chart described in the first parts of the report, the next steps include: further feature engineering and model tuning. In addition, the last weeks of the project will be dedicated to the analysis of the model itself: understanding its decisions and observing the weight attributed to each feature. This is a key step as it will allow, not only to understand the model and get a better view of its architecture but also to explain it to the industrials that will use it afterward. If they cannot understand the model fully, they will not be able to trust its predictions.
Since this project is part of a wider project, at the end of my internship, follow-up projects will be launched leveraging my work:

- Predicting the imbalance price
- Predicting the imbalance costs
- Defining a production strategy

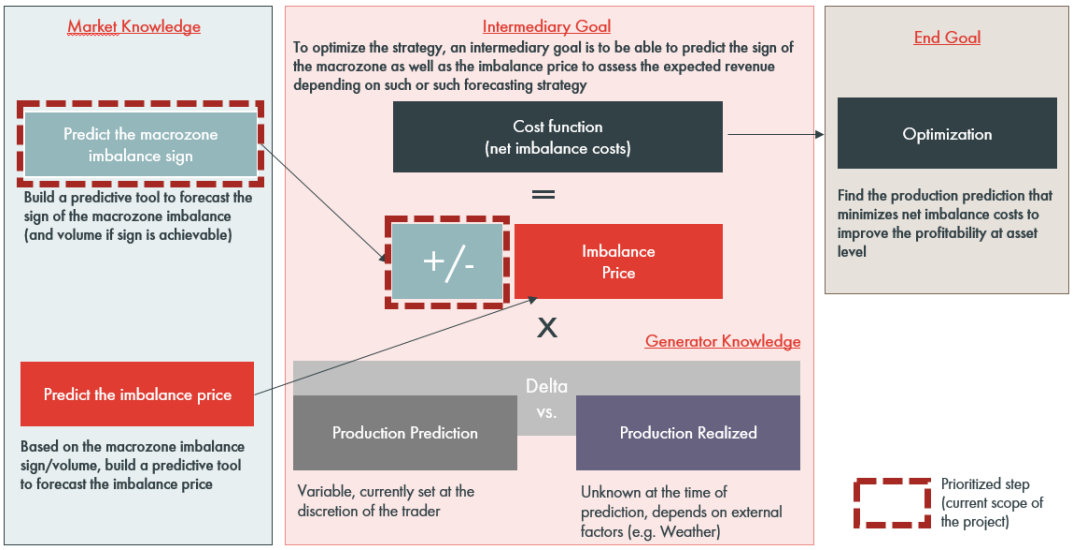

Figure 26: Global project overview

6.3 Comparison to the State of the Art

Although the project is not over yet, currently it offers a very different approach than the two main papers described earlier since:

- Period studied: Both papers are based on outdated market structure and thus do not reflect the current structure of the Italian Electricity market. Thus, our study offers an analysis closer to the actual dynamics of the market.
- Data sources used: both studies only used a limited amount of data to perform their predictions. They both used past Sbil values and respectively days of the week & temperature and seasonality & generation mix. Our solution uses a wider range of data sources (used in the model: market data, increased weather data, refined seasonality, load and generation mix or not: population, macroeconomics, ...).
- Models investigated: both studies focused on autoregressive models and studied neither statistic models (Linear Regression, Ridge, Lasso, Logistic Regression, ...) nor Machine Learning models (Decision Trees, Random Forests, and Gradient Boosting).

Overall, we believe our approach to diverges significantly from theirs. Regarding the results obtained, it is hardly comparable since, first, we were not able to see if their dataset was balanced and second, Sbil itself changed significantly after September 2017 (hence the accuracy cannot be compared).

7 Conclusion & Acknowledgement

Upon the first half of this internship, I believe I have learned much not only on core project organization skills (how to define a project's scope, how to present results to a non-technical public,...) but also core data science skills (where to find relevant data, handle non-perfect datasets, which data to investigate, extract analyze, feature engineering skills, ...). Further, doing a research project while being able to see the true impact of the research on something measurable (such as the income of a business) makes the project more practical. Finally, I was glad to be able to apply most of the teaching I received from both at Imperial College London and at Ecole Polytechnique. I look forward to improving my skills in data science, first while pursuing the end of this internship, and then in the future. I believe this experience as an intern at Ardian made me realized how impactful can tech be in the business sector.

I would like to thank both Ioana Manolescu and Alessandro Astolfi who have made this internship possible. Further, I would like to thank the Infrastructure team at Ardian and specifically Pauline Thomson, Michaël Brouard, Olivier Hamot, Louise Badarani, and Skander Kamoun who have supervised me during this project.

8 References

References

[1] Ardian's website. https://www.ardian.com/

[2] Gestore Mercati Energetici's website. https://www.mercatoelettrico.org

[3] Giorgia Oggioni and Cristian Lanfranconi. *Empirics of Intraday and Real-time Markets in Europe: ITALY*. 2015.

[4] Terna's website. https://www.terna.it/en

[5] Alexandre Lucas, Konstantinos Pegios, Evangelos Kotsakis and Dan Clarke. *Price Forecasting for the Balancing Energy Market Using Machine-Learning Regression*. 2020.

[6] Patrick Avis and Geo Lee. *Evaluating System Imbalance Forecasting Models for the United Kingdom Electricity Market*. 2021.

[7] Jonathan Dumas, Ioannis Boukas, Miguel Manuel de Villena, Sébastien Mathieu, Bertrand Cornélusse. *Probabilistic Forecasting of Imbalance Prices in the Belgian Context*. 2019.

[8] Nils Terpstra. *Day-ahead and imbalance price forecasting on the Dutch Electricity Market a comparison between time series and artificial neural networks models* . 2020.

[9] Andrea Cervone, Ezio Santini, Sabrina Teodori, Donatella Zaccagnini Romito. *Electricity Price Forecast: a Comparison of Different Models to Evaluate the Single National Price in the Italian Energy Exchange Market*. 2014.

[10] Federica Davo, Maria Vespucci, A. Gelmini, Paolo Grisi, and Dario Ronzio.*Forecasting Italian electricity market prices using a Neural Network and a Support Vector Regression*. 2016.

[11] Luca Di Persio, Alessandro Cecchin, and Francesco Cordoni. *Novel approaches to the energy load unbalance forecasting in the Italian electricity market*. 2017.

[12] Francesco Lisi and Enrico Edoli. *Analyzing and Forecasting Zonal Imbalance Signs in the Italian Electricity Market*. 2018.

[13] Moto Dei. *Facebook Prophet: (Almost) everything you should know to use Facebook Prophet like a pro, with many example Python codes and cheat sheet*. 2020·

[14] Ajay Tiwari. *Let's Forecast Your Time Series using Classical Approaches A Gentle Introduction to 14 Classical Forecasting Techniques and their Implementation in Python*. 2020.

[15] Sarang Narkhede. *Understanding AUC - ROC Curve*. 2018.