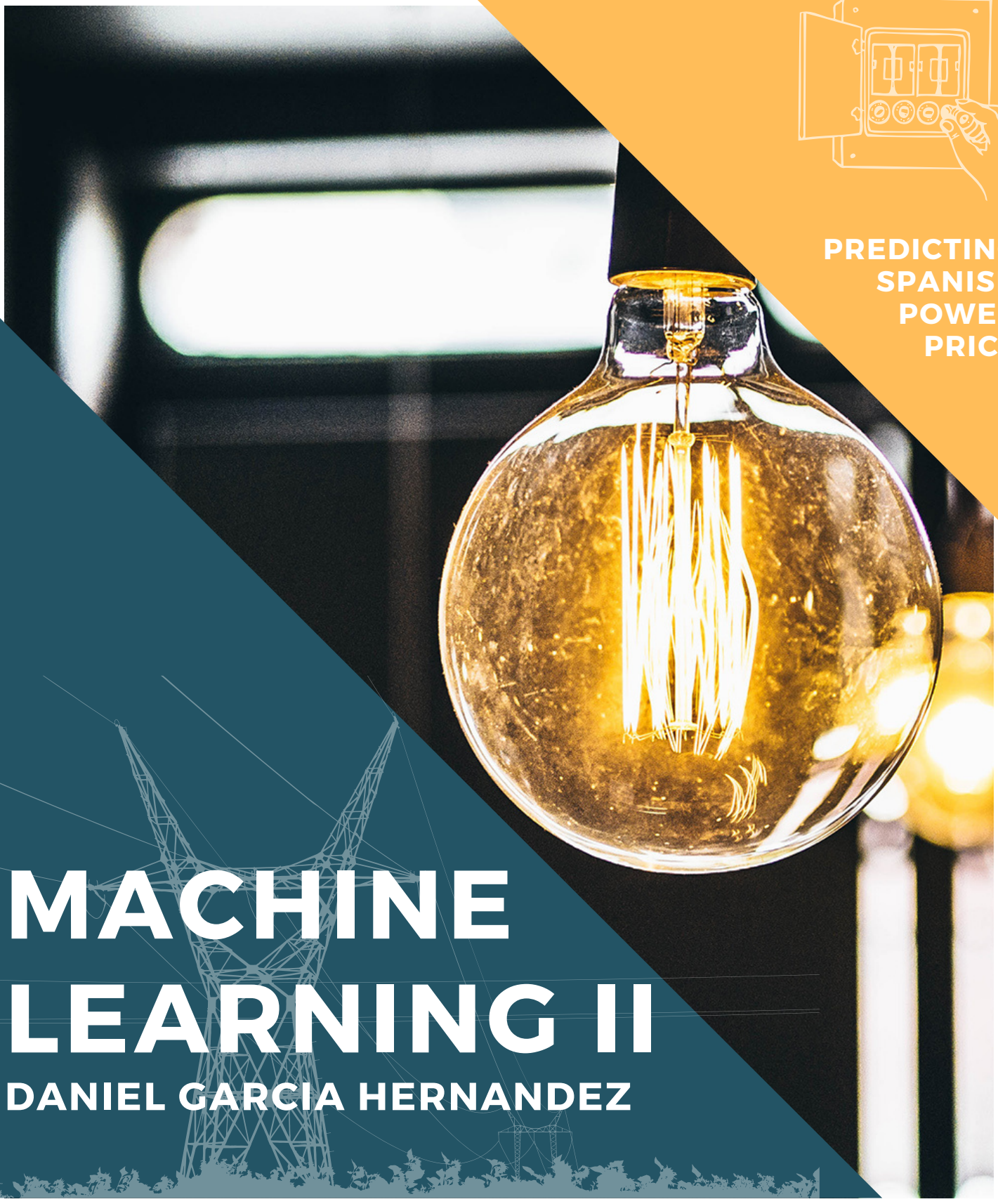


PREDICTING SPANISH POWER PRICE

MACHINE LEARNING II

DANIEL GARCIA HERNANDEZ

ABSTRACT

Predicting power prices is critical to all industry stakeholders in planning regulation and policy in the country as well as financing. The power industry relies heavily on electricity prices. Power price forecasting is made possible by creating a regression model that takes into account all factors that affect electricity demand and supply. Creating and optimizing models to predict electricity prices is expected to have a profound impact on related industries.

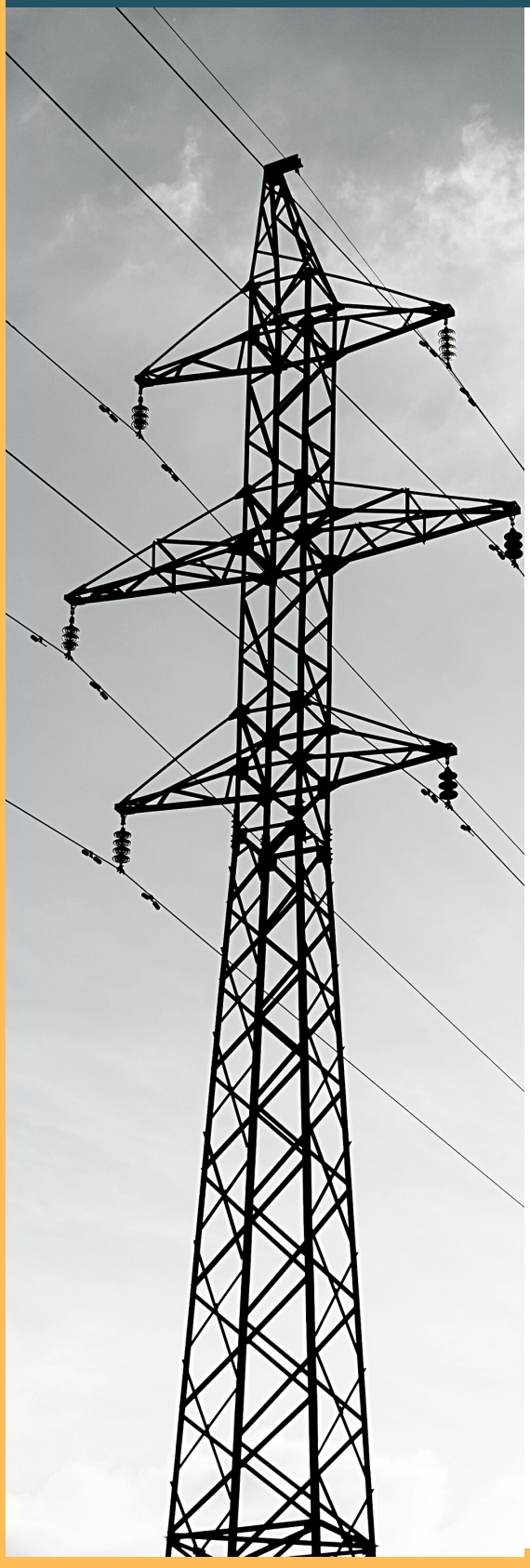
SECTION 1 GROUP H

SAMUEL KURABACHEW, DANIEL MARRANZINI, JUAN BUERO, MATTHIEU DE BEAUCORPS, HA YOUNG LEE, PATRICK PICOLLO, JACOPO SCHLESINGER

TABLE OF CONTENTS

INTRODUCTION

Spanish energy



The nature of pursuing convenience due to the nation's economic growth and increased income levels is responsible for the continued increase in energy demand. Until now, Spain's policy has been prioritized to expand its energy supply to meet growing energy demand in the process of economic growth. Consequently, what are the ways to use limited resources as efficiently and reasonably as possible? An efficient and reasonable way to fit everyone's utility with limited resources in the energy sector can be found in demand management. In particular, energy demand management can be used as an important policy option for sustainable growth in the country.

The goal of this report is to create a model to predict the day-ahead price of power in Spain given some forecast available before the daily auction. We suggest a price forecasting model based on the LightGBM regression model. It performs the best among several models we went through.

This analysis attempts to identify the associated relationship between descriptive variables and predictors in past occurrences. Subsequently, these relationships are utilized to predict future outcomes of variables of interest.

Data overview

OUR EDA

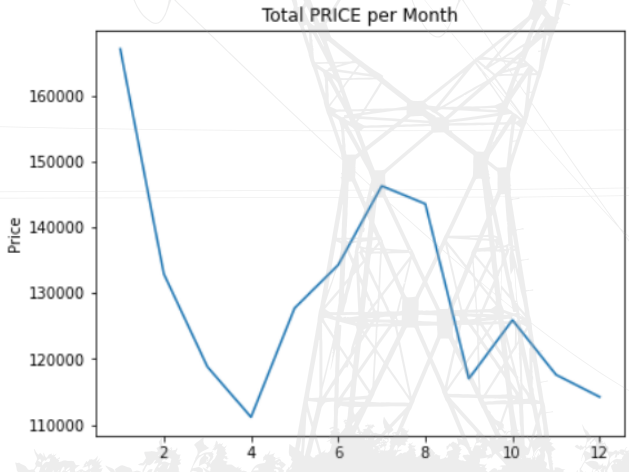
We will not deep dive into the specifics of this section, but instead, give a broad overview of how the thought process developed and which were the insights we gained in this EDA to create a good model at a later stage.

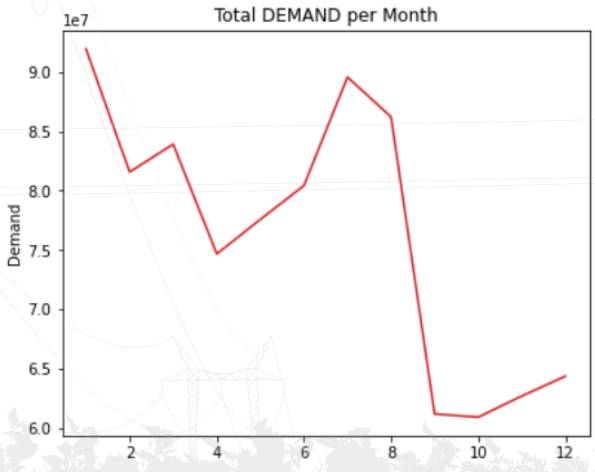
After reading both datasets and having a glimpse of the first rows in each one, we followed the portfolio manager's recommendations and analyzed each variable by itself while also plotting related variables together to visualize their behavior. For instance, we found that demand and price were variables that were clearly correlated as previously outlined in the instructions of this exercise.

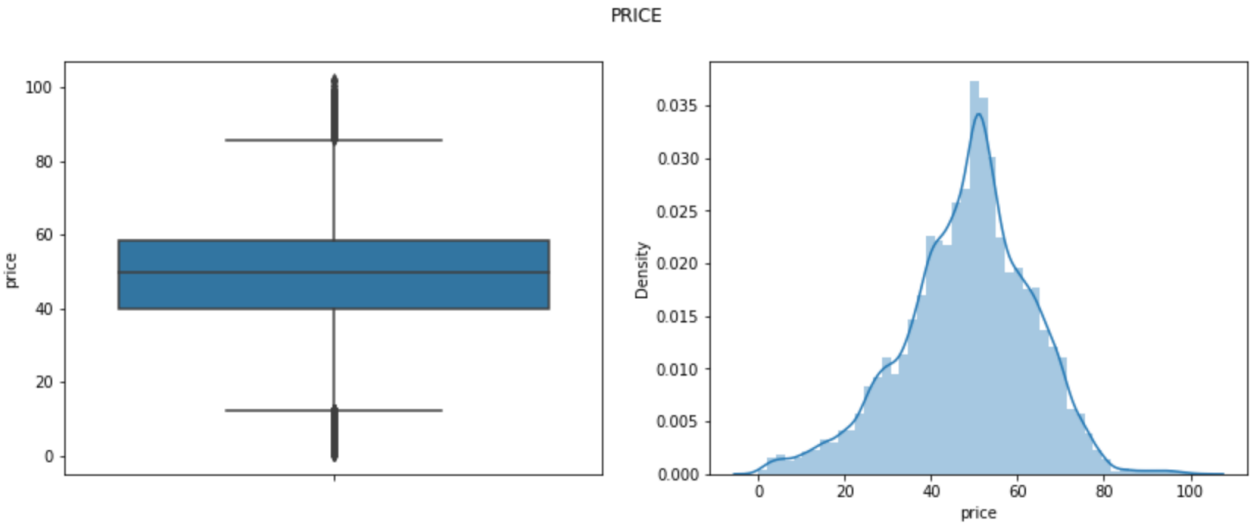
Given the relevance that price has for our analysis, we repeated this process with other measure of time from the data such as month, quarters, and years.

Having looked at Price Vs Demand, we can see that there actually seems to be a linear relationship between the two: as demand increases, price increases. We have also seen that prices surge down at the end of the month, potentially because some months are shorter (28/30 days instead of 31), driving the price to 0 for these periods statistically.

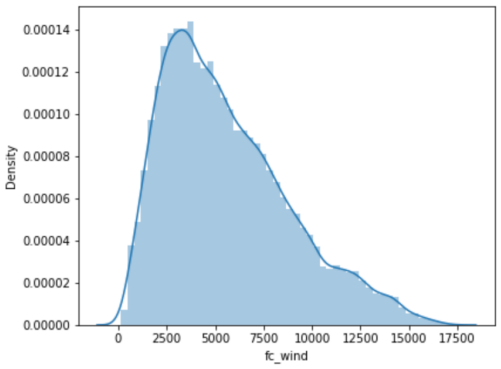
Further on we went on to find that our price was normally distributed. From the box plots, we can see that there are some outliers. We will see later on if these outliers are anomalies or data collection errors.



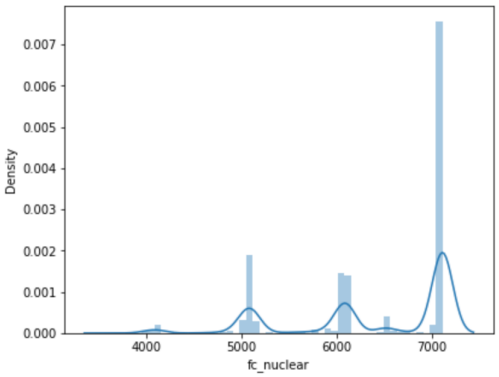




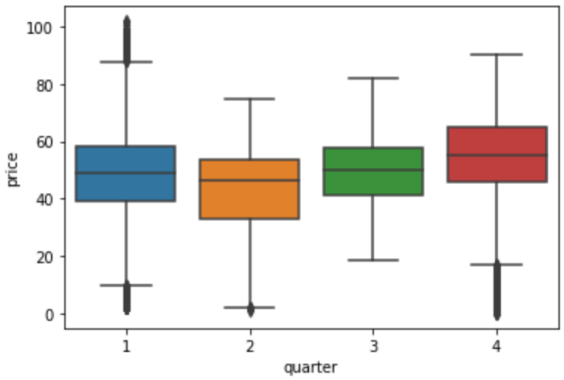
Further on, we went on to analyze the different forecasts according to the different types of energies available. We started by analysing the wind forecast which seemed positively skewed to the right, which indicates most of the wind power forecast lies in the range 2500 - 5000 MWh.

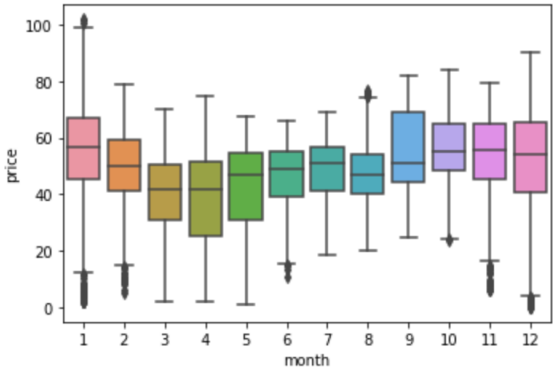


On the other hand, the Nuclear energy forecast clearly showed us that we had higher values at certain points. This might have been due to the fact that Nuclear plants are run for only certain times and the production is related to the amount of time the reactors operate.



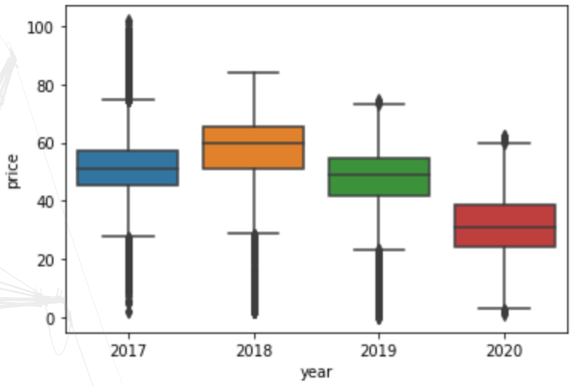
Subsequently, we conducted an analysis on solar power and solar thermal power production. In comparison, these two power sources have lower production. In addition to feature vs target plots, we have also used box plots to indicate patterns in price over the different months, quarters, and years, as shown on the next page.





Initially, we can see a clear distinction in prices, in the four quarters we have analyzed, the 4th one being overall the most expensive and the 2nd one the least expensive. This can be explained in correlations with our second graph. In this one and as for the months, we can see a clear tendency in the increase of price during the "cold" or winter months with an exception or outlier in September, which probably indicates a return to work-life after vacations. This goes in line with the fact that the 4th quarter is also more expensive because it has colder months and therefore a bigger need for energy consumption (heating, less sunshine, more public transport).

Further on, when looking at the yearly patterns for the price, we can see very little correlation between them, 2020 is the period with the cheapest overall price. An observation that could be made for this is definitely Covid-19. As an incentive for people to stay at home, many countries reduced energy prices. Industrial production also decreased a lot and prices, such as those of fossil fuels, plummeted to all-time lows.



Having a feel for the data we were provided with, and visualizing patterns contained within, our team felt much more comfortable with the data and was eager to start building a model.

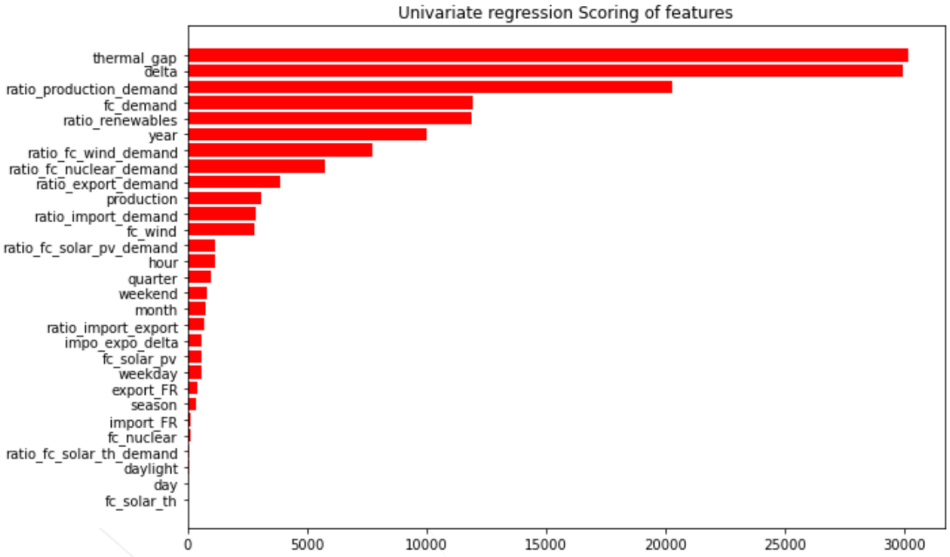
FEATURE ENGINEERING

For our feature engineering process, we used a pipeline to handle all our transformations to the training dataset. The main concept that guided our decisions for what features to engineer was that demand is what sets our price. This is why the first feature we decided to create was the thermal gap, meaning our demand minus our production values for solar, wind, and nuclear energy. The reasoning behind this is that these energy sources are generally the cheapest to produce, at least much cheaper than other fossil fuel-based sources. So if we have let's say 45GW in demand and 20GW of solar/wind power and 20GW of nuclear power, the actual price is going to be decided by those remaining 5GW that either have to be imported or extracted from more expensive sources.

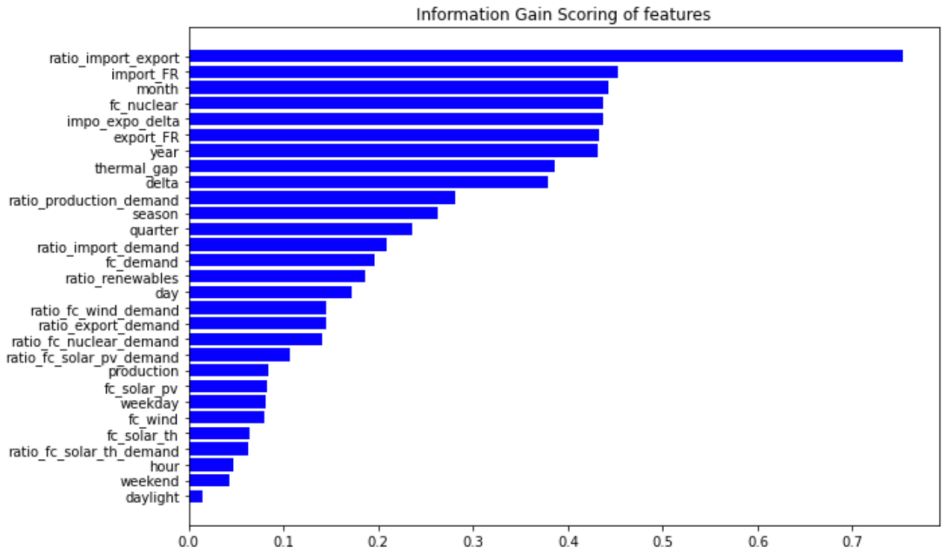
Then going with the idea that demand was our most important feature, we decided to create other variables we thought would have a direct effect on demand. The first step was extracting time values from our date, so month, season, year, weekday, quarter... All of these are values that would have a direct impact on the amount of energy consumed. We also deduced if there was daylight at each given time by making use of the "astral" library which gives us the sunrise and sunset times of a given location. Following this, we extracted a number of ratios we thought might be significant to our prediction. So the ratio of production vs. demand, as well as the ratio of renewables being produced and the import and export ratios among others.

FEATURE SELECTION

Feature Selection is one of the core concepts in machine learning which hugely impacts the performance of your model. The data features that you use to train your machine learning models have a huge influence on the performance you can achieve. Irrelevant or partially relevant features can negatively impact model performance, which was an interesting finding in our notebook. We use two methods to see the importance and capability of our features in explaining the target variable. One method was looking for a linear relationship between our target variable and the feature at hand.



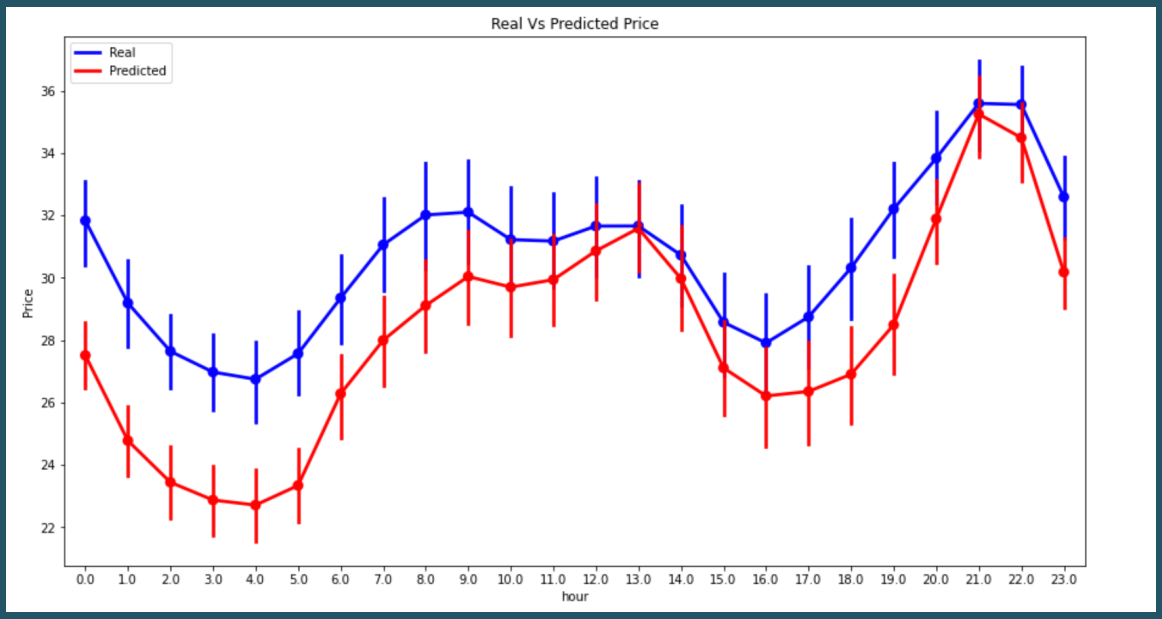
The other one is, trying to find the best features that explain the price in any way: meaning quadratic or any other higher degree relationship.



After getting this information, we started training our models first on the entire feature and then tested the performance against each feature. In the end, we saw that a combination of some features seems was ideal to use. These features are mentioned below in the modeling report.

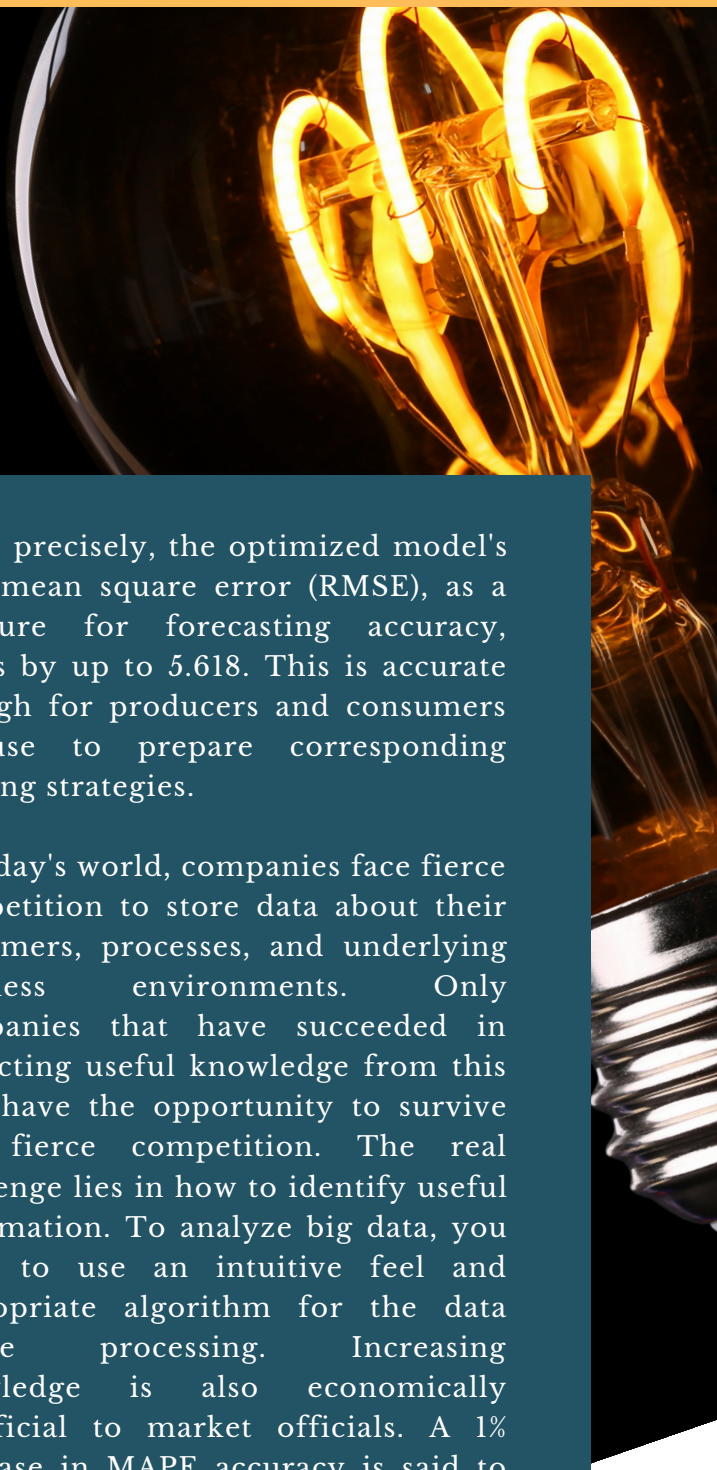
MACHINE LEARNING (MODELING)

We have tested different models starting from simpler ones to more complex and surprisingly most of the relations with the target variable were picked by linear models. Even linear regression and Lasso were performing better than Decision Tree without any parameter tuning. On the other hand, more complex models like LightGBM and XGBoost were giving the best results out of the box. Looking at their performance, we were motivated to do more hyperparameter tuning to get a more accurate and better forecast. Unfortunately, the tuned model didn't perform as our expectations. We arrived at the conclusion that we might be overfitting to our training dataset and we decided to keep the original models. On top of this, our models were better at forecasting only with around 7 features. Looking at the figure above in the feature selection section, the most important features were thermal gap, demand, and the ratio of production vs demand which makes sense because if the renewable power sources are lower the cost of power increases since it has to be met with other costly sources. Also, Year was an important factor since the price seems to decrease as we progress to the end of 2020. We have below the performance of our model versus the real price with hand-picked features.



SAMUEL KURABACHEW, DANIEL MARRANZINI, JUAN BUERO, MATTHIEU DE BEAUCORPS, HA YOUNG LEE, PATRICK PICOLLO, JACOPO SCHLESINGER

CONCLUSION



In this report, we created a model to predict the day-ahead price of power in Spain given some forecast available before the daily auction. First, we conducted an Exploratory Data Analysis (EDA) to explore variables that affect demand and supply that affect electricity prices. In order to create a model to predict the power values a day ahead, we started by analyzing historical data of prices. Later on, and after processing this data, we optimized our model by correlating prices with several variables. For sophisticated predictions, we processed a dataset containing missing values. Since this is time-series data, we used forward fill to impute the missing values.

Furthermore, and to select the best performing model, various types of patterns and parameters were included, among them, Ridge Regression, Lasso Regression, LightGBM, XGBoost, and Random Forest. Of all these combinations of models and parameters, the most precise one was the LightGBM, for which more detailed studies were conducted, including the analysis of the most relevant variables.

More precisely, the optimized model's root mean square error (RMSE), as a measure for forecasting accuracy, drops by up to 5.618. This is accurate enough for producers and consumers to use to prepare corresponding bidding strategies.

In today's world, companies face fierce competition to store data about their customers, processes, and underlying business environments. Only companies that have succeeded in extracting useful knowledge from this data have the opportunity to survive this fierce competition. The real challenge lies in how to identify useful information. To analyze big data, you need to use an intuitive feel and appropriate algorithm for the data you're processing. Increasing knowledge is also economically beneficial to market officials. A 1% increase in MAPE accuracy is said to save $1.5 million per year for mid-sized utilities. As a first step towards helping the electricity market, future experiments conducted will be expanded.