



Predicting Spanish power prices

IE MBD Oct 2020
Section 1 Group 8
Working in Teams

ie
SCHOOL OF
HUMAN SCIENCES
& TECHNOLOGY

Agenda & Management Summary

1. Feature Creation

2. EDA 1: Time & Price

3. EDA 2: Time & Supply

5. Model Evaluation

6. Conclusion

(Additional information included in
Jupyter Notebook)

The main goal of this work was it to predict the day-ahead price for power (€/MWh). We have seen that the demand is highly affected by the time. The supply is affected by the weather and time. The gap between those drives the price. In our analysis we found out that the Linear Regression Ridge model worked the best (RMSE 6.28). However, this must be evaluated in a real-time environment (scoring.csv) now. Important features are the month, thermal gap, and the gap between supply and demand.

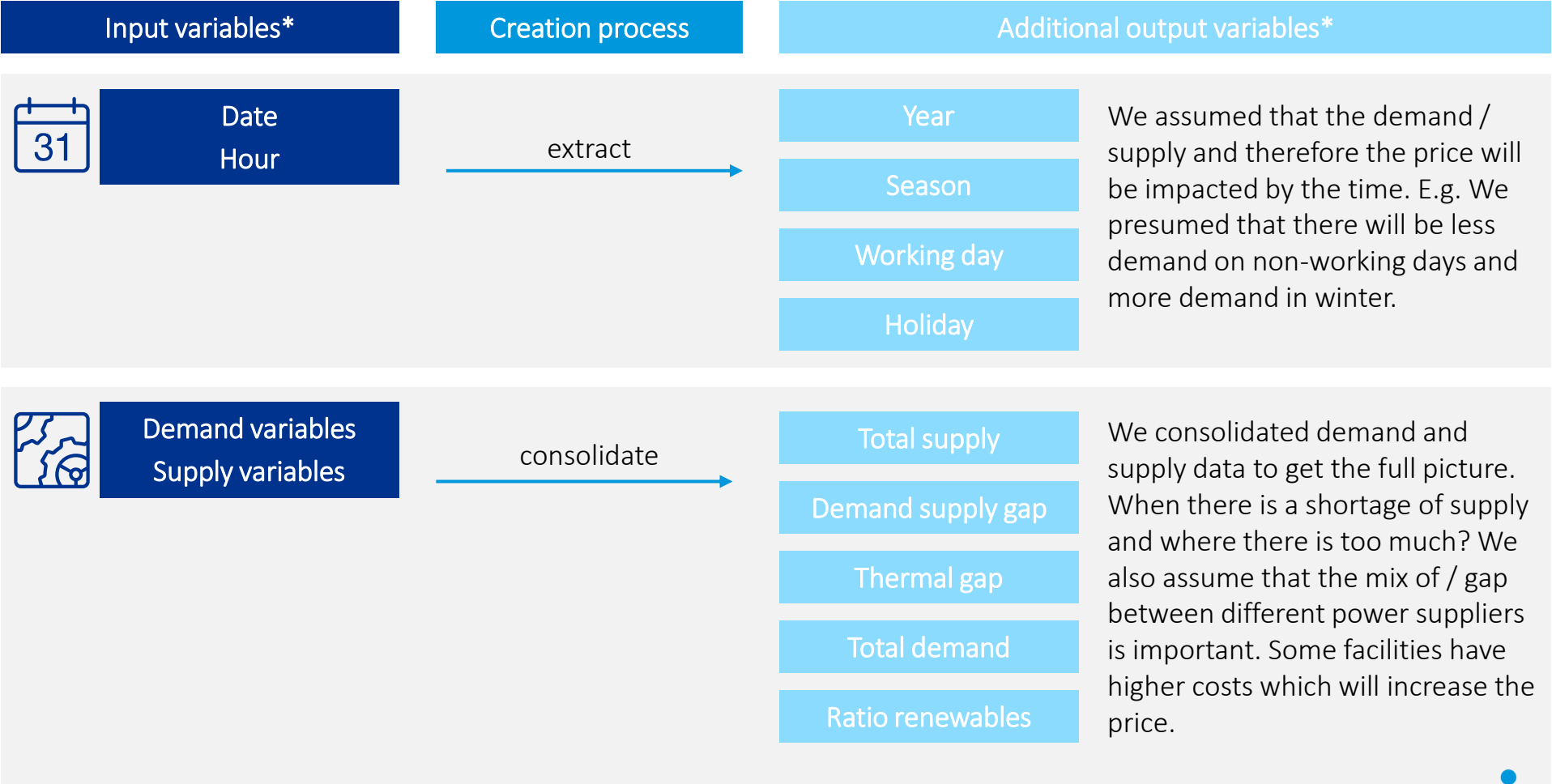
Feature creation

We created additional features to extract hidden information.

Feature Engineering / creation

Feature engineering & creation is one of the most important steps of the process for developing prediction models. It might be considered as an art as it involves human design and some intuition.

We assumed that there could be some hidden information within the data. Therefore, we created additional time-related and usage-related variables



*Not all variables are show and some are consolidated in this presentation, see Notebook for more detailed information

Explanatory Data Analysis 1: Time & Price

The price is highly influenced by several time-related features.

Analysis

After extracting time-related variables, we created graphs to better understand the data and their relationship to price.

Graph 1: Price vs. year

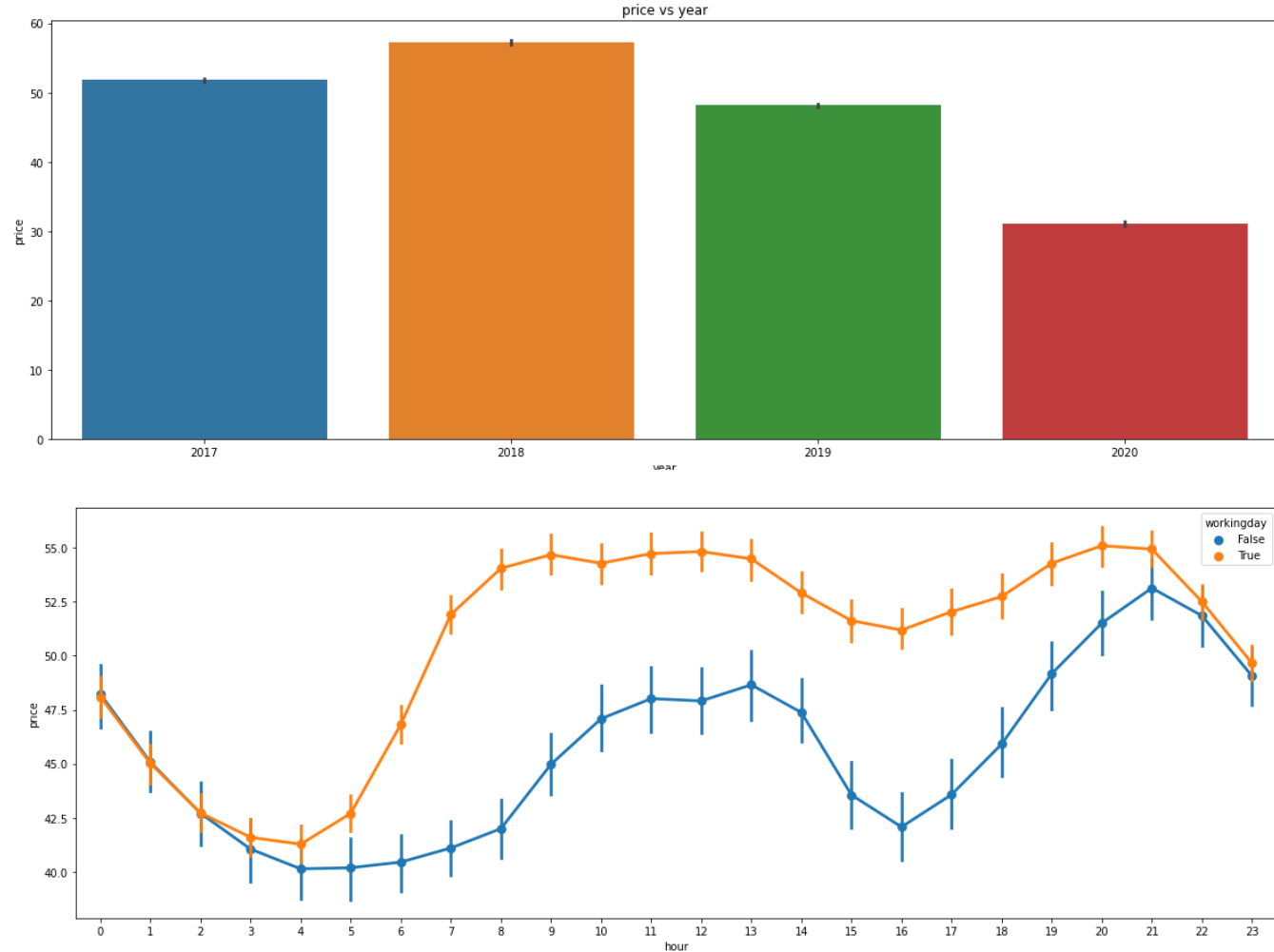
We can see a clear drop in the price ($>0.20\text{€}$) from 2018-2020. This could have several causes, but we can analyze that the year is important factor for our model to forecast the price properly.

Graph 2: Price vs. hour and workday

The hour of the day as well the day itself are seems to be important factors when it comes to determine the price. This makes sense because the consumption/demand is lower when the economy sleeps.

Conclusion

Time-related variables are important factors when it comes to finding the price. Also, the demand is highly infected by the price. The creation of additional features is going to help us to create more accurate predictions.



Explanatory Data Analysis 2: Time & Supply

Renewable energies are dependent on time and weather. Factory-like production facilities are inert.

Analysis

After consolidating supply-and-demand-related variables, we created graphs to better understand the data and their relationship.

Graph 1: Photovoltaic supply

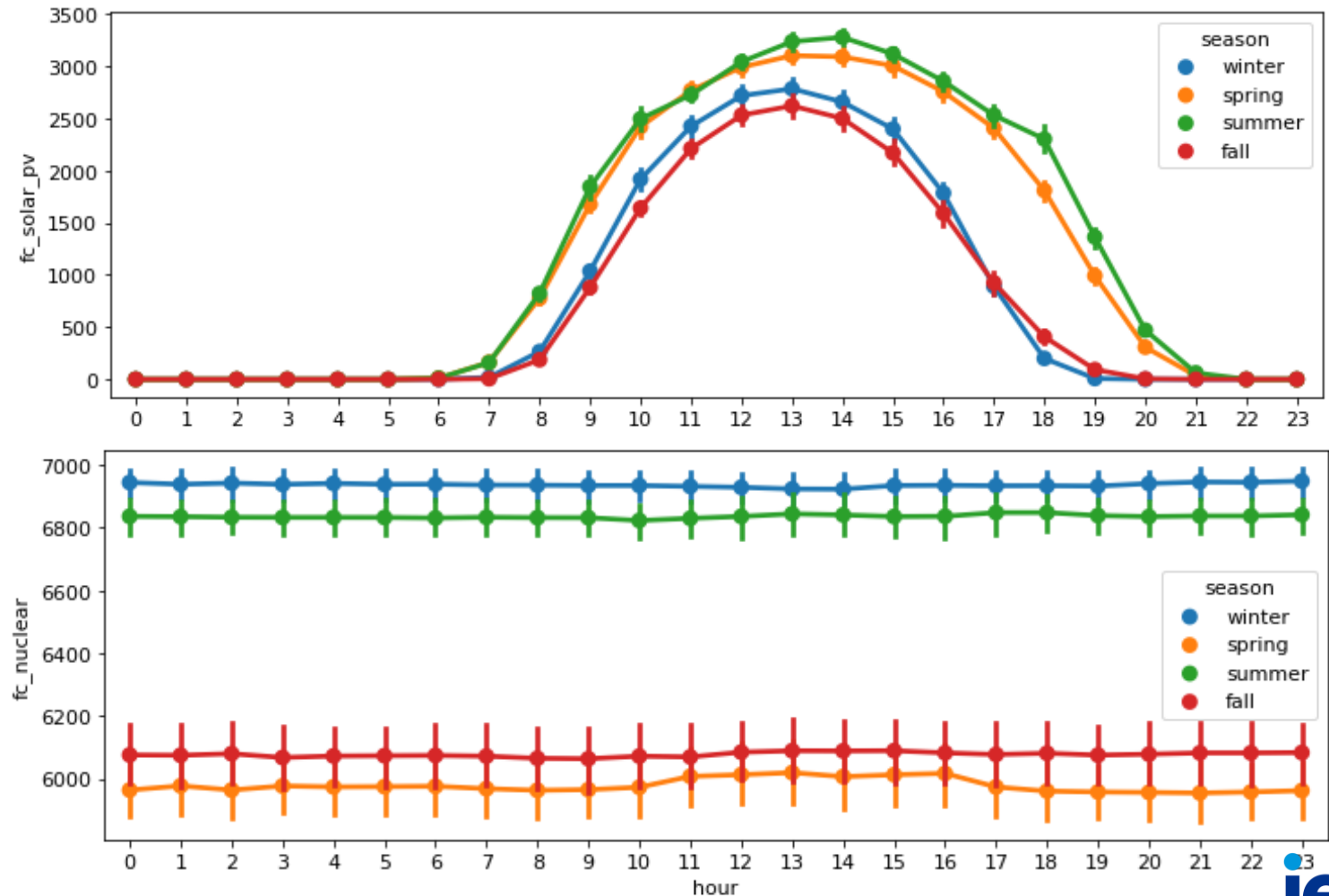
The supply of renewable sources is highly impacted by weather and/or time conditions. We can analyze that solar power has a high variance depending on time and season.

Graph 2: Nuclear supply

Nuclear and other plant production resources stay constant within the day but falls dramatically during fall and spring. It is hard to decrease / increase the production of nuclear plants in short-term.

Some are very rigid and slow, others react strongly to weather influences. This makes it tough to calculate the supply and thus the price. Accordingly, features like thermal gap are important features to determine the price.

Conclusion




Model Evaluation

It looks like Linear Regression Ridge is the best model to predict prices.

Process

1. Split the dataset based on time. We picked a five-month test dataset. The rest is used for training
2. Run different models with features obtained from the EDA
3. Evaluate the models based on the lowest error rate (RMSE)
4. Improve feature selection
5. Run again
6. Evaluate the models based on the lowest error rate (RMSE)

Model	Best RMSE score
Linear Regression Ordinary least squares Linear Regression, who main aim is to minimize the residual sum of squares between the observed targets in the dataset, and the targets predicted by the linear approximation.	7.20
Linear Regression Ridge Technique for analyzing multiple regression data that suffer from multicollinearity. When multicollinearity occurs, least squares estimates are unbiased.	6.28 
Random Forest Regressor A random forest regressor is a meta estimator that fits a number of classifying decision trees on various sub-samples of the dataset and uses averaging to improve the predictive accuracy.	9.02
Gradient Boosting Regressor Gradient boosting is a machine learning technique for regression and classification problems, which produces a prediction model in the form of an ensemble of weak prediction models.	16.93
Support Vector Regressor An SVM model is basically a representation of different classes in a hyperplane in multidimensional space. The goal of SVM is to divide the datasets into classes to find a maximum marginal hyperplane.	Compute time took too long
XGB Regressor XGBoost is a decision-tree-based ensemble Machine Learning algorithm that uses a gradient boosting framework.	9.52

Conclusion



Feature importance

Important features to predict the price are:

- Month
- Thermal gap
- Delta Demand Supply

Model

The best prediction model we could find is:

- Linear Regression Ridge model
- RMSE: 6.28

Business

This is the business value of such a model:

- Supply side: Allocate resources when the price is predicted to be high
- Demand side: Consume power during a time of a low price

Thanks!



IE MBD Oct 2020
Section 1 Group G
Machine Learning II

ie
SCHOOL OF
HUMAN SCIENCES
& TECHNOLOGY