IMF Business School y Universidad de Nebrija

# Predicción del precio del pool eléctrico español

**Trabajo Fin de Máster en Big Data & Business Analytics**

**Autor: Carlos Alonso Salcedo**

**Tutor: Juan Manuel Moreno Lamparero**

**Junio 2022**

# Índice de contenidos

## *AGRADECIMIENTOS*

Antes de comenzar, me gustaría agradecer, con mención especial, a Lorenzo y Mario, por darme la oportunidad, enseñarme e introducirme en el mundo del mercado eléctrico y energía, a todo el departamento de Analytics de Deloitte, por darme la oportunidad de seguir aprendiendo y hacer lo que me gusta, y a mis compañeros de proyecto Inés, Miguel Ángel, Antonio y Miguel por su apoyo e ideas dentro y fuera del trabajo.

Dedicado a la memoria de Marta, no hay día que pase en el que no te recuerde, nada me gustaría más en el mundo que estuvieras con nosotros, espero que estés orgullosa.

# 1. RESUMEN

El precio de la electricidad es un tema controvertido por la actual escalada de precios en la que estamos inmersos. El objetivo del presente trabajo es el de la predicción del precio del pool eléctrico español durante todo el año de 2021. Para ello, se hará uso de diferentes bases de datos y páginas web con información financiera, de producción de energía renovables y datos meteorológicos, entre otros.

Para predecir esta variable objetivo, se hará uso de diferentes algoritmos y métodos, desde los más sencillos, basado en un modelo de persistencia, hasta modelos complejos basados en arquitecturas de redes neuronales. El resultado final será de un *benchmark* de los modelos usados para ver cuáles se comportan mejor en este periodo de tiempo caracterizado por el periodo con más volatilidad de la historia del precio de la electricidad en España.

Para ello, todos los modelos se compararán mediante un *backtesting out of sample* para todo el año 2021 y se compararán con las mismas métricas para poder realizar una comparación directa y así poder determinar cuál es el mejor modelo para resolver nuestro problema.[1]

[1] Todo el código desarrollado en el presente trabajo y los gráficos mostrados pueden ser encontrados en el siguiente repositorio de GITHUB: **https://github.com/caralosal/TFM-Big-Data-Business-Analytics**

# 2. INTRODUCCIÓN Y ANTECEDENTES

## 2.1. INTRODUCCIÓN Y AGENTES DEL SECTOR ELÉCTRICO

La electricidad se trata de un pilar en el que se basa el mundo actual. Prácticamente todo lo que se usa día a día como ordenadores, electrodomésticos, teléfonos móviles, iluminación, industria y un largo etcétera, dependen de la electricidad. De hecho, la energía es un bien básico de primera necesidad cuyo acceso debe ser garantizado como servicio público (1).

Esto se traduce en que la cantidad de energía que se consume es altísima pues está presente en prácticamente todos los sectores y en el día a día de cada persona. Por ejemplo, la cantidad de energía eléctrica que se consumió en 2021 en España fue de 256.387 GWh, comparado con el consumo en 1980 que fue de 97.231 GWh, se trata de un consumo de más del doble, debido al aumento de la población y al uso cotidiano de cada vez más dispositivos electrónicos (2).

Sin embargo, pese a que cada vez hay más población y se usan más dispositivos electrónicos, en Europa se redujo el consumo de electricidad más de un 10 % entre 2005 y 2015 debido a las mejoras en la eficiencia energética, el aumento de la proporción de energías procedentes de la fuentes hidráulica, eólica y solar fotovoltaica, los cambios estructurales en la economía y la recesión económica de 2008. También ha contribuido el hecho de que los inviernos hayan sido más cálidos, lo que ha permitido reducir la cantidad de energía destinada a calefacción (3).

El sector eléctrico es altamente complejo, pero al mismo tiempo fundamental para poder mantener el estilo de vida contemporáneo. En el *pool* del mercado es donde se compra la energía que llega a nuestros hogares y donde se vende la electricidad producida en las centrales. (3)

Un **sistema eléctrico** es el conjunto de elementos que operan de forma coordinada en un determinado territorio para satisfacer la demanda de energía eléctrica. Este sistema, en España, tiene 7 componentes:

1. Centros o plantas de generación (donde se produce y eleva la tensión de la electricidad para transportarla)
2. Líneas de transporte de la energía de alta tensión, gestionada, desarrollada y mantenida por Red Eléctrica de España (REE).
3. Las estaciones transformadoras, que reducen la tensión.

4. Líneas de distribución de media y baja tensión que llevan la electricidad hasta los consumidores.
5. Las instalaciones de los clientes o consumidores de energía eléctrica
6. Los centros de control de las empresas generadoras, distribuidoras y comercializadoras.
7. Un centro de control eléctrico nacional desde el que se gestiona, coordina y opera el sistema eléctrico, y que está gestionado también por REE.

Sin embargo, el negocio del sector eléctrico no sería tal sin la participación de los siguientes actores:

- Generadores: producen la energía eléctrica, independientemente del tipo de tecnología utilizada para ello.
- Transportista: transmiten la energía de la red de transporte a los puntos de consumo donde se entrega a los distribuidores. En España, por ley, hay un transportista único y es REE.
- Distribuidores: son las compañías que llevan la electricidad hasta los clientes finales. Aunque hay numerosas empresas que ejercen esta actividad, en cada área de España sólo puede haber un distribuidor.
- Operador del sistema: es la empresa que se encarga de que todo el proceso de la operación del sistema eléctrico funcione correctamente. La clave es cumplir la siguiente ecuación y regla de oro en el negocio de la electricidad.

$$\boldsymbol{Generacion = Demanda} \quad (1)$$

- Comercializadoras: venta de energía eléctrica a los consumidores según la potencia contratada, comprándola a su vez a los generadores en el llamado '*pool eléctrico*', una especie de subasta que determina el precio de la energía eléctrica en la Península Ibérica.

## 2.2. HISTORIA DEL SECTOR ELÉCTRICO EN ESPAÑA

Durante años, el sector eléctrico en España funcionó como un oligopolio en el que el precio de la electricidad dependía de pocas compañías eléctricas en España. El país estaba dividido en diversas áreas geográficas, cuyo suministro de electricidad se adjudicó en exclusiva a cinco grandes empresas. Cada una de estas eléctricas gestionaba las 4 fases del suministro de energía de su ámbito de actuación, dejando a los consumidores sin posibilidad de escoger su propia compañía eléctrica.

En 1997, con el objetivo de fomentar la competencia en el sector de la energía eléctrica y mejorar el conocimiento que los usuarios tenían del mismo, se aprobó la primera ley de liberalización del sector eléctrico nacional. La normativa prohibía que una misma compañía opere en más de una de las fases del proceso de suministro y transfiera la gestión del transporte a distintas redes eléctricas en España.

Como consumidores, no podemos escoger qué empresa distribuye la energía eléctrica que consumimos, pero sí podemos elegir a quién se la compramos.

Como se ha visto, la electricidad forma parte de la vida y de la actividad económica de las personas. Realizando un cálculo rápido y sencillo, el precio medio de la electricidad en España en 2021 fue de 111,93 €/MWh. Esto indica que, únicamente en el negocio de compra y venta de la electricidad en el mercado del pool, se trató de un negocio que movió cerca de 29 mil millones de euros en España en 2021.

$$256.387\ \cancel{GWh} \cdot \frac{1.000\ \cancel{MWh}}{1\ \cancel{GWh}} \cdot \frac{111{,}93\ €}{1\ \cancel{MWh}} = 28.697\ M€$$

Esto lo convierte en un negocio que mueve mucho dinero y que, debido a lo mucho que se consume, su precio es muy importante ya que afecta a todas las personas y todos los negocios.

## 2.3. ¿Cómo se establece el precio de la electricidad?

Una vez se ha visto lo importante que es la electricidad, qué agentes participan en el negocio y la cantidad de energía y dinero que mueve este negocio cabe preguntarse, ¿cómo se establece el precio de la electricidad?

La forma de establecer el precio sigue el **algoritmo Euphemia** que surgió en la iniciativa "Price Coupling of Regions" (PCR) por parte de siete mercados eléctricos europeos, entre los que se encuentra el español. Este algoritmo calcula los precios de la energía eléctrica de forma eficiente persiguiendo la maximización del bienestar, que se define como el excedente o beneficio tanto de los compradores como de los vendedores, al tiempo que optimiza el uso de capacidad disponible en las interconexiones.

En resumen, las empresas encargadas de la generación hacen sus ofertas (cantidad de energía y precio) y las empresas encargadas de la venta al por menor, consumidores directos, etc., demandan la energía necesaria a un

precio determinado. Una vez realizadas las ofertas, se ordenan según el precio, en orden creciente en el caso de la venta y decreciente en el caso de la compra. La intersección de las curvas de oferta y demanda se denomina **punto de casación**.

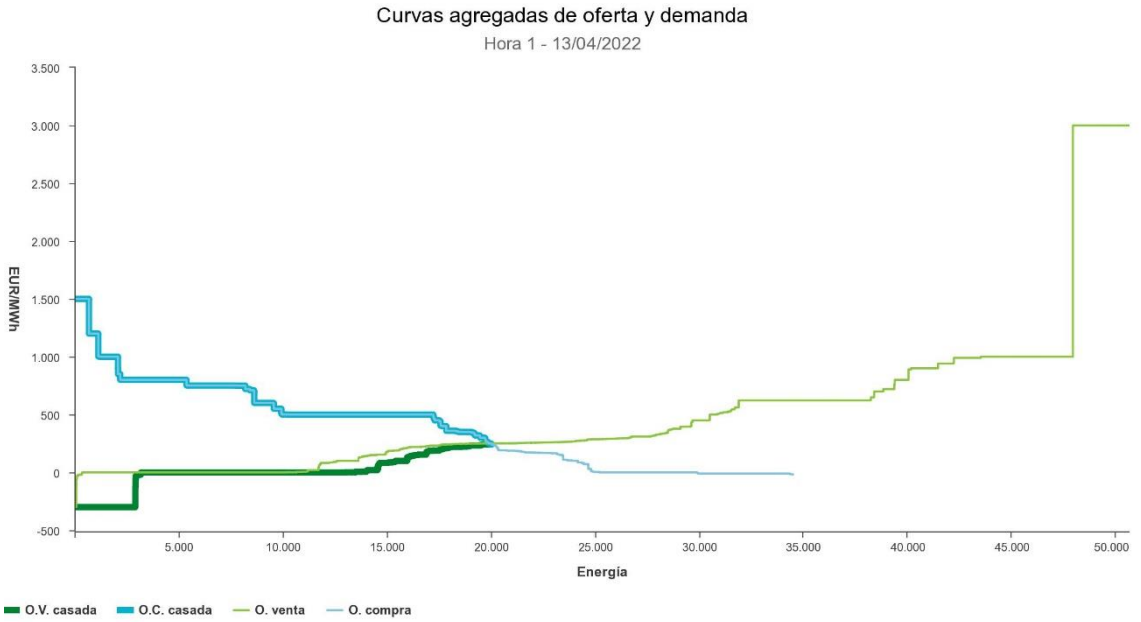


Imagen 1. Curva de oferta y demanda para el 13 de abril de 2022 en la hora 1. Imagen procedente de OMIE. https://www.omie.es/es/market-results/daily/daily-market/aggragate-suply-curves

Por tanto, este punto de corte es el punto que optimiza el bienestar y, por tanto, establece el precio de la energía para esa hora concreta. Toda la energía ofrecida y demandada a un precio inferior al punto de casación se intercambiará a ese precio, mientras que la que tenga un precio superior no lo hará. Este proceso se repite para cada una de las 24 horas de un día, estableciendo el precio de la electricidad para cada una de las horas de cada día.

La energía que se envían a las casas suele ser una mezcla de varias tecnologías (eólica, fotovoltaica, hidráulica, ciclos combinados de gas, carbón…). La generación de electricidad por cada energía tiene un precio asociado, sin embargo, a la hora de venderla a mercado, todas se venden al mismo precio que es el que marca el punto de casación, que es en lo que se basa un **sistema marginalista**.

Para ilustrarlo, se expone un ejemplo completamente ficticio: a una hora determinada hay una demanda de 20.000 MW, y se tiene ofertando la energía nuclear con 7.000 MW a precio cero; eólica, 12.000 MW a precio también cero. Luego, los ciclos combinados (gas), 10.000 MW a precio 40

€/MWh y por último carbón, 8.000 MW a precio 60 €/MWh. Al casar la oferta resultan los 7.000 MW de la nuclear, los 12.000 MW de la eólica y solo 1.000 MW de ciclo combinado, el carbón se queda fuera. **Pero a todos se les paga el precio de casación, 40 €/MW**.

Es por ello que al sistema se le denomina marginalista, porque a todos se les paga el precio marginal de casar la oferta y la demanda, independientemente de lo ofertado.

## 2.4. Potencia instalada en España.

Como se ha adelantado, la energía que llega a nuestras casas y negocios es una mezcla de la electricidad producida por diversas tecnologías. Actualmente en España se dispone de una potencia instalada de 114.290 MW, distribuida en las siguientes tecnologías. (6)

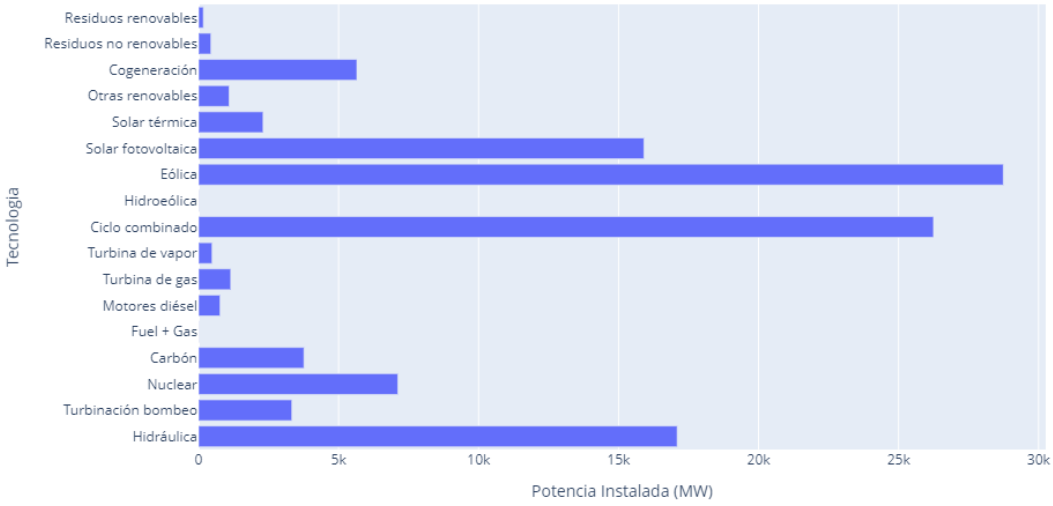


Imagen 2. Distribución por tecnología de la potencia instalada en España (6). Elaboración propia

De ella se deduce que las 4 tecnologías con mayor capacidad instalada en España son, en primer lugar, la eólica (28.743 MW), seguida del ciclo combinado (26.250 MW), la hidráulica (17.094 MW) y la solar fotovoltaica (15.098 MW)

Sin embargo, estos números hay que tratarlos con cuidado. Estos valores de generación hacen referencia a la potencia total que podría producir el sistema si todas las distintas tecnologías produjesen simultáneamente el máximo del que están preparadas. Poniendo un ejemplo más concreto, la

energía eólica tiene 28.743 MW de potencia instalada lo que quiere decir, que si todos los aerogeneradores situados en España produjesen el máximo de electricidad al que están preparados, producirían esta cantidad de energía, pero esto es bastante complicado que ocurriese, ya que en todas las plantas de España debería estar soplando un viento suficientemente fuerte como para producir la máxima energía de la que está preparada una central en el mismo momento. Lo mismo pasaría con las demás tecnologías.

Tampoco indica que en esta proporción se produce electricidad, por ejemplo, de solar hay 15.098 MW de potencia instalada, pero es evidente pensar que el peso de esta energía respecto a la total en verano e invierno será muy distinto, ya que en verano la potencia generada estará más cerca de la potencia instalada fotovoltaica que en invierno y el % de la generación ocupará un mayor peso en los meses con mayor radiación como los de verano.

En el siguiente gráfico (7) se muestra, para cada día del periodo comprendido entre el 01/01/2022 y 01/05/2022. El % de energía producida por cada tecnología con respecto al total de la energía producida. A simple vista se observa que las barras verdes (correspondiente a la eólica) es muy volátil, pues depende del viento, que producirá cuanto más haya, en caso de que no exista, se observa como crece la barra color naranja claro correspondiente al ciclo combinado, que trata a hacer frente a corregir cuando la eólica no produce en niveles altos.

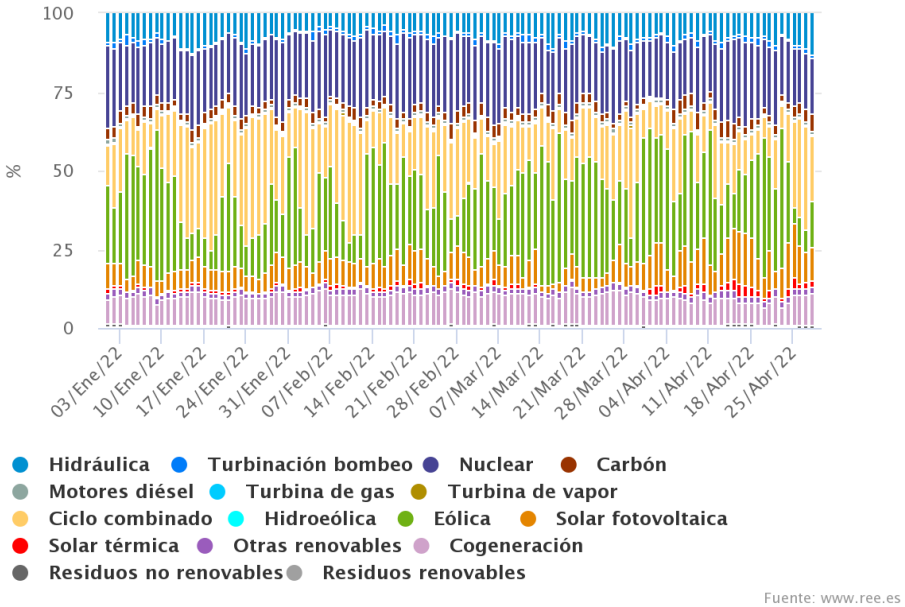


Imagen 3. Porcentaje de energía producida por cada tecnología con respecto al total (7).

Por poner números concretos, comparamos un miércoles de enero con un miércoles de abril y vemos la producción de las tecnologías más importantes.

| Tecnología | 05/01/2022 (GWh) | 27/04/2022 (GWh) |
|---|---|---|
| Hidráulica | 57 | 79 |
| Fotovoltaica | 35 | 68 |
| Eólica | 295 | 48 |
| Ciclo Combinado | 92 | 227 |
| Total | 767 | 695 |

Tabla 1. Comparativa entre las tecnologías producidas el 05/01/2022 y el 27/04/2022.

Como se observa, la fotovoltaica, como era de esperar, produce el doble en abril que en enero. Sin embargo, el 27/04/2022 parece que fue un día en el que no hizo mucho viento, ya que la producción fue muy pequeña comparada con la que se produjo el 05/01/2022. Para contrarrestar esta disminución, se observa un aumento muy claro de producción en el ciclo combinado de la fecha de abril con respecto la de enero.

Todo esto pone de manifiesto lo complicado que es de gestionar este ejercicio de demanda, pues la producción de fuentes renovables como la eólica o fotovoltaica depende de las condiciones climáticas las cuales no son constantes y, si estas no favorecen, se han de utilizar otras fuentes (más caras) para satisfacer la demanda, que impacta directamente en el precio de la energía.

### 2.5. ¿Cómo se consume en España?

En un día cualquiera, el inicio de la jornada laboral, el cierre de los comercios durante el mediodía o la mayor ocupación de los hogares en las horas finales del día, explican por qué la demanda no es idéntica en las distintas horas del día. (8)

Nuestra sociedad demanda más energía en algunos momentos del día (**horas punta**). En invierno, estas horas punta se dan entre las 11:00 y 12:00 o bien entre las 19:00 y 20:00, debido a la confluencia entre actividad comercial y ocupación de los hogares. Sin embargo, en verano las horas punta se producen en las horas centrales del día, coincidiendo con los momentos de mayor temperatura. (8).

Durante las horas nocturnas se produce la demanda mínima diaria (**horas valle**). A estas horas, únicamente la demanda industrial mantiene un consumo importante, ya que las grandes fábricas consumen las 24 horas del día, aprovechando las horas nocturnas, cuando la energía se puede contratar más barata (8).

Un perfil de consumo de un día normal (ya que los festivos pueden no seguir este perfil) es el siguiente. (8)

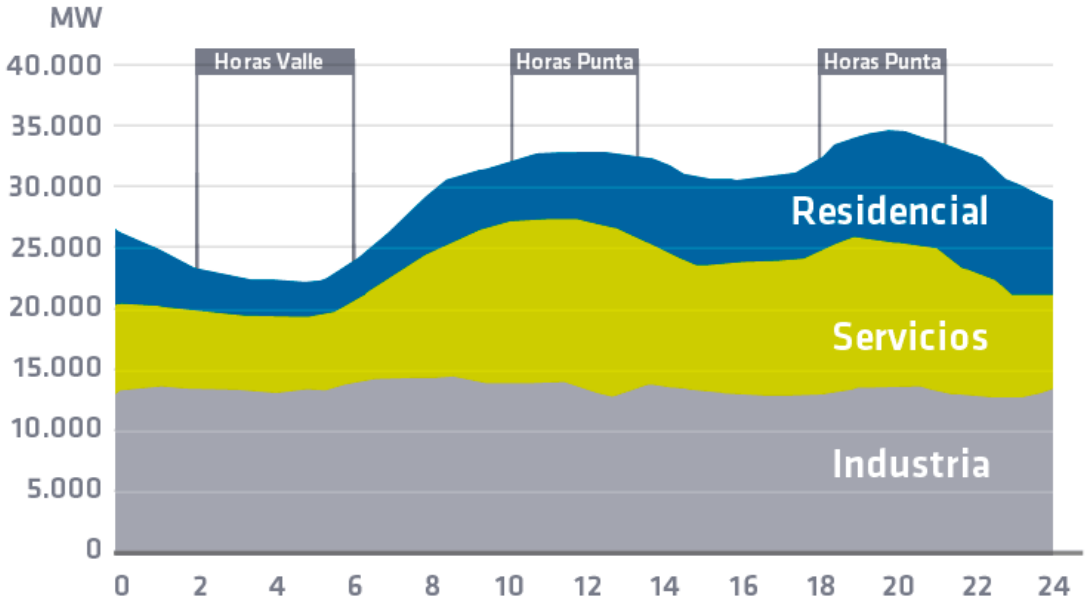


Imagen 4. Perfil de consumo en un día para el sistema eléctrico español (8).

El máximo de consumo en una hora de un día puede oscilar entre los 35.000 y 40.000 MW, mientras que la potencia instalada con la que cuenta España es cercana a los 114.290 MW, es decir, si todo funcionase al máximo de eficiencia en todas las tecnologías, podríamos producir 3 veces más energía de la que normalmente se demanda en un día.

Sin embargo, esto nunca se alcanza, pues las energías renovables dependen de las condiciones meteorológicas y no se ha dado el caso de que se produzca el máximo de todas las renovables en España.

## 3. HIPÓTESIS DE TRABAJO Y OBJETIVOS

En el presente trabajo se va a predecir el precio de la electricidad en España. Para obtener esta variable objetivo, se ha hecho uso del histórico guardado de REE (Red Eléctrica España) en la página de ESIOS [2]. Cada día, alrededor de las 14:00, se publica el valor del precio de la electricidad para cada una de las 24 horas del día siguiente.

El objetivo es el de la predicción del precio de cada una de estas 24 horas antes de que sea publicado, haciendo uso de una serie de variables explicativas que serán explicadas más adelante.

En el siguiente gráfico se muestra la evolución del precio de la electricidad en España entre 2014 y 2021. En ella se observa que hasta enero de 2021, el precio no había superado los 140 €/MWh (en esta fecha se alcanzó por el temporal Filomena) y, a partir de ahí y que recuperase su precio normal, se incrementó hasta los niveles de finales de año rondando los 300 – 400 € /MWh, un cambio en la tendencia histórica muy drástico y sin ningún precedente.


Imagen 5. Evolución del precio de la electricidad en España

Como hipótesis del trabajo (que más adelante se validarán) se proponen como posibles variables explicativas las siguientes:

[2] https://www.esios.ree.es/es/analisis/600?vis=1&start_date=01-01-2021T00%3A00&end_date=30-11-2021T23%3A00&compare_start_date=31-12-2020T00%3A00&groupby=hour&geoids=3

- Retardos de la variable objetivo, es decir, tiene sentido pensar que el precio de la electricidad para cada una de las horas de mañana va a tener bastante correlación con el precio de las horas del precio de la electricidad de hoy o de las horas del mismo día de la semana anterior. Por ejemplo, el retardo de la variable objetivo de 1 día y de una semana.
- El gas es la materia prima que entra en el proceso de generación de electricidad por ciclo combinado y es una de las tecnologías que entran en el proceso de casación, por lo que esta variable también podría dar una idea de tendencia o niveles del precio de la electricidad.
- Variables financieras que son un reflejo del estado económico de la sociedad, por ejemplo, el IBEX35 (al estar en España) o el BRENT (petróleo crudo) son variables que pueden estar también relacionadas.
- Hueco térmico: las energías más limpias y baratas de conseguir son las generadas a partir de energías renovables. El problema, es que la demanda, pocas veces es cubierta íntegramente por energías renovables y ésta tiene que ser cubierta por otros procesos no renovables como la energía generada en centrales térmicas o ciclos combinados. A esta energía que no es cubierta por las energías renovables, y que es necesaria de procesos térmicos convencionales o ciclos combinados se denomina hueco térmico, que puede estar bastante correlacionada con el precio pues cuanto mayor sea el hueco térmico, más uso se deberán hacer de estas tecnologías que son las más caras para generar electricidad y por tanto, el precio suba. Se define como:

$$\boldsymbol{HT = D - R}\ (\boldsymbol{MW})\quad (2)$$

Denotando, HT, D y R, hueco térmico, demanda y renovables, respectivamente, en unidades de potencia. En la producción de renovables se incluyen las energías generadas por las tecnologías eólica, fotovoltaica, hidráulica y nuclear. Sin embargo, dado que la publicación de la previsión de las energías hidráulica y nuclear para el día D+1 es a partir de las 16:00, no son incluidas en la ecuación, pues el precio es publicado aproximadamente 2 horas antes. Por el contrario, la previsión horaria de las energías eólica y fotovoltaica para el día D+1 es publicada a las 11:00, por lo que éstas son las que entrarán en la ecuación del cálculo del hueco térmico.

- **Día de la semana y festivos**, pues tanto los negocios como los propios consumidores domésticos no consumen igual en fin de semana o festivo que en un día laboral. Dado que la demanda varía y está relacionada con el precio, esto debería afectar también al precio de la electricidad.
- **Datos climatológicos,** ya que están las renovables íntimamente ligadas a las condiciones climáticas, y esta generación de renovables impacta en el precio de la electricidad, valores como temperatura, radiación solar o viento, pueden estar relacionadas con el precio de la electricidad.

El objetivo es crear un modelo y testearlo out-of-sample para todo el periodo de 2021, es decir, 365 días x 24 horas = 8760 horas de predicción y comparación con el valor real.

### 3.1. Métricas

Para evaluar los modelos, se usarán las siguientes métricas, disponibles y usadas en el script *metricas.py* del repositorio adjunto al trabajo.

Se harán uso y se estudiarán todas las métricas, pero la que se utilizará como métrica principal será el **WMAPE**. Se optó por no incluir la variable de MAPE (*Mean Absolute Percentage Error*, que es la media de los errores porcentuales en valor absoluto de cada observación), debido a la naturaleza de la variable objetivo ya que, en algunos periodos de tiempo en el que las renovables producen mucha energía (y el hueco térmico es cercano a 0), el precio es menor a 1 €/MWh, por lo que, aun realizando una predicción de 2 €/MWh, el valor de MAPE para esa predicción es del 100 %, lo que desvirtuaba la métrica en algunos periodos y no era buena elección, problema que se soluciona usando su versión ponderada (WMAPE).

Siendo 'y' el valor predicho, 'x' el valor real y 'N' el número de observaciones realizadas (número de horas consideradas en el estudio):

- **MAE** (*Mean Absolute Error)*. Función que calcula el error absoluto medio entre las predicciones y los valores reales.

$$MAE = \frac{1}{N}\sum_{i}^{N}|y_i - x_i| \quad (3)$$

- **WMAPE** (*Weight Mean Absolute Percentege Error).* Es el MAPE ponderado por el peso de la variable real. De esta forma, se minimizan

los efectos de productos con grandes errores, pero cuyo valor es pequeño.

$$WMAPE = \frac{\sum_i^N |y_i - x_i|}{\sum_i^N x_i} \cdot 100\ (\%) \quad (4)$$

- **RMSE** (*Root Mean Squared Error)*. Se trata de la desviación estándar de los residuos (errores de predicción). Los residuos son una medida de cómo de lejos de la predicción están los valores reales.

$$RMSE = \sqrt{\frac{\sum_i^N |y_i - x_i|^2}{N}} \quad (5)$$

- **% Tendencia**: es una variable que define únicamente si el precio de la electricidad va a aumentar o disminuir en la hora siguiente. Es una medida para determinar si el modelo está captando la tendencia de la serie horaria.

$$\%\ Tendencia = \frac{Nº\ horas\ con\ tendencia\ acertada}{Nº\ horas} * 100\ (\%) \quad (6)$$

El "Nº horas con tendencia acertada" se define como:

$$Num\_acierto(x_i, y_i) \begin{cases} 1 & si\ x_i - x_{i-1} < 0\ ;\ y_i - y_{i-i} < 0 \\ 1 & si\ x_i - x_{i-1} > 0\ ;\ y_i - y_{i-i} > 0 \\ 0 & en\ otro\ caso \end{cases}$$

$$Nº\ horas\ con\ tendencia\ acertada = \sum_i^N Num\_acierto(x_i, y_i) \quad (7)$$

# 4. Material y métodos

## 4.1. Procedimiento

Todo el trabajo se realizó con Python 3.8.3 en distintos Jupyter Notebooks en el entorno de Anaconda (versión 2020.07), teniendo cada uno de ellos una función determinada, y los cuales vienen recogidos en el repositorio en GitHub adjunto al trabajo para facilitar la consulta del mismo. Las distintas versiones de las librerías utilizadas a lo largo del trabajo están recogidas en el Anexo A.

El procedimiento fue el siguiente:

1. Creación del dataset de predicción del precio de la electricidad. A partir de las hipótesis postuladas anteriormente, el primer paso es recoger todas las variables y alinearlas con la variable objetivo, asegurarse de que se puede disponer de todas ellas en el momento de la predicción.
2. Estudio de las variables explicativas y su correlación con la variable objetivo.
3. Creación de un modelo baseline. Antes de comenzar un ejercicio, se plantea un modelo sencillo. Por ejemplo, un modelo sencillo podría ser que la predicción del precio de la electricidad sea el valor medio del histórico que tengamos. En este caso, se propuso como modelo baseline que el precio de las 24 horas de la electricidad de mañana es igual al precio de las 24 horas de la electricidad de hoy. Este tipo de modelos tienen importancia antes de usar modelos más complejos pues si no hay variabilidad en los datos y este modelo tan sencillo tiene ya un MAPE muy bajo, no es atractivo usar modelos complejos pues, probablemente ni lo mejoren o si lo mejoran no compensa su utilización.
4. Definición de los modelos a usar. El objetivo del trabajo es resolver el problema de predicción del precio mediante distintos métodos (algoritmos) desde los más sencillos hasta los más complejos. Los definidos en el trabajo son los siguientes:
    a. Modelo baseline
    b. Modelo de regresión lineal
    c. Modelo de regresión Ridge y Lasso
    d. Modelo SARIMA
    e. Modelo SARIMA de los residuos de la regresión lineal
    f. Modelo Random Forest
    g. Modelo XGBoost

    h. Modelo de redes neuronales LSTM
    i. Modelo Ensemble
    j. Modelo Stacked
5. Evaluación de los modelos en el mismo periodo y con las mismas métricas para poder determinar el que mejor se adapta al precio real de la electricidad.
6. Conclusiones y análisis de los resultados obtenidos.

Una vez establecido el procedimiento, se describe la construcción del dataset usado para la predicción. Todos los datos se descargaron de las urls que se mencionarán a continuación y se descargaron los datos y se incluyeron todos en la carpeta 'Data' disponible en el repositorio.

## 4.2. Construcción de las variables explicativas

La construcción de las variables explicativas se realiza en dos notebooks incluidos en el repositorio: "Preprocesamiento.ipynb" y "Preprocesamiento Climatologia y Finanzas.ipynb". En el de preprocesamiento se tratan las variables de materias primas (precio de electricidad, gas, demanda energética, hueco térmico, dummies, festivos…) y en el de climatología y finanzas, los datos climatológicos y financieros como temperatura y BRENT, respectivamente.

Ya se ha hablado de la variable objetivo, así que se describe una a una las variables usadas y el origen.

### 4.2.1. Previsión MIBGAS D+1

El MIBGAS es el responsable de la gestión del Mercado Organizado de Gas y es el nombre de la cotización del precio del gas en la Península Ibérica. En la página del MIBGAS se puede acceder al precio spot y varios precios forward (a 1 día, 1 mes, 1 año, entre otros). Como se va a predecir el precio del día D+1, se obtiene esta variable en la siguiente url[3] y cabe destacar que se trata de una variable de cotización diaria, no horaria como la electricidad. La serie obtenida es la siguiente (comienza en enero de 2016 porque es cuando empezó a cotizar este producto financiero). Se observa un pico muy alto en 2021 (correspondiente al temporal provocado por Filomena, con unos 60

[3] https://www.mibgas.es/es/file-access

€/MWh gas) y, tras recuperar sus valores normales, se observa la escalada de precios en la que estamos inmersa de precios, con precios superando todos los días los 120 €/MWh gas.

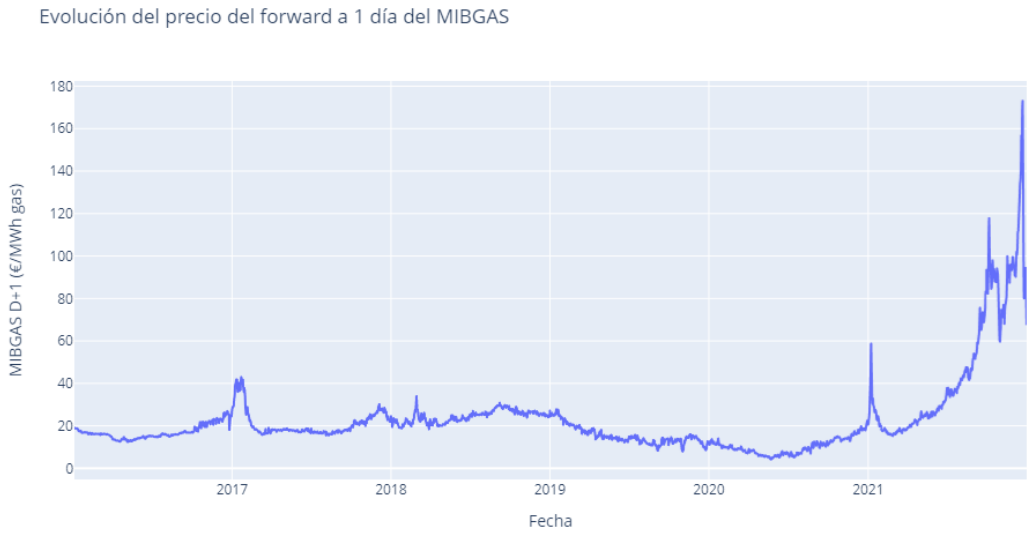


Imagen 6. Evolución del precio forward a 1 día del MIBGAS

### 4.2.2. Previsión de la demanda horaria D+1

Dato también proporcionado por REE[4]. Consiste en una serie a nivel horaria en el que el dato corresponde con la previsión de la demanda de energía en España. Se observa una tendencia estacional en la que marca máximos de demanda en los meses de climas extremos, ya sean fríos (diciembre y enero) o los más cálidos (julio y agosto). Se observa entre marzo y julio de 2020 un periodo atípico de bajo consumo que corresponde al periodo de confinamiento debido al Covid-19.

[4] https://www.esios.ree.es/es/analisis/1775?vis=1&start_date=06-10-2016T00%3A00&end_date=31-12-2021T23%3A50&compare_start_date=05-10-2016T00%3A00&groupby=hour

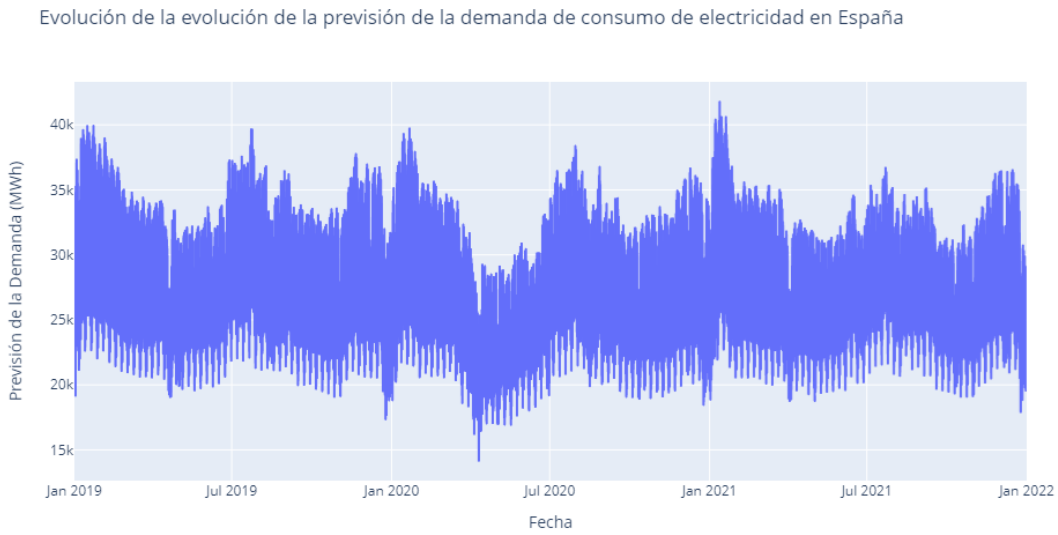


Imagen 7. Evolución de la previsión de la demanda D+1 desde 2019

### 4.2.3. Previsión de la generación eólica + fotovoltaica D+1

Dato también publicado en la web de REE[5]. Se trata de otra serie a nivel horaria cuyo dato representa la cantidad de energía, en MWh, generada mediante las tecnologías eólica y fotovoltaica de forma conjunta. Como está la suma de ambas tecnologías, y la eólica tiene mucha más potencia instalada (y genera más que la fotovoltaica), no se observa la estacionalidad de que en verano haya más producción, es más, agregando mensualmente, los meses en los que más producción hay son octubre, noviembre, diciembre y enero.

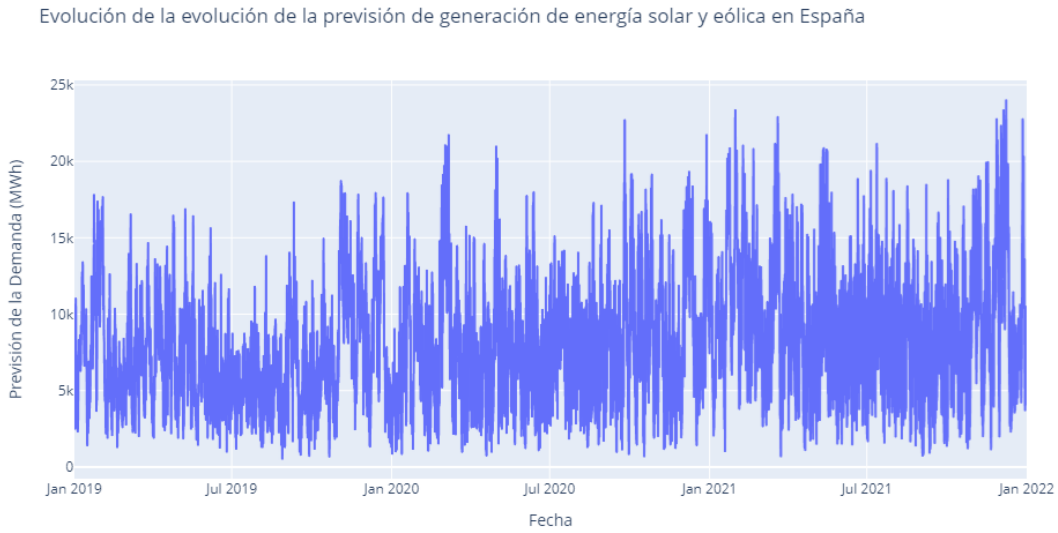


Imagen 8. Evolución de la generación de eólicas + fotovoltaicas D+1 desde 2019

[5] https://www.esios.ree.es/es/analisis/10358?vis=1&start_date=01-10-2015T00%3A00&end_date=31-12-2021T23%3A50&compare_start_date=30-09-2015T00%3A00&groupby=hour

### 4.2.4. Hueco Térmico

Calculado mediante (2). Se trata también de una serie con desglose horario. Esto nos da la cantidad de demanda que hay que cubrir mediante centrales térmicas y de ciclo combinado por lo que, a menor valor de esta variable, quiere decir que la demanda ha sido cubierta en mayor medida por las renovables, por lo que el precio de la electricidad debería ser menor.

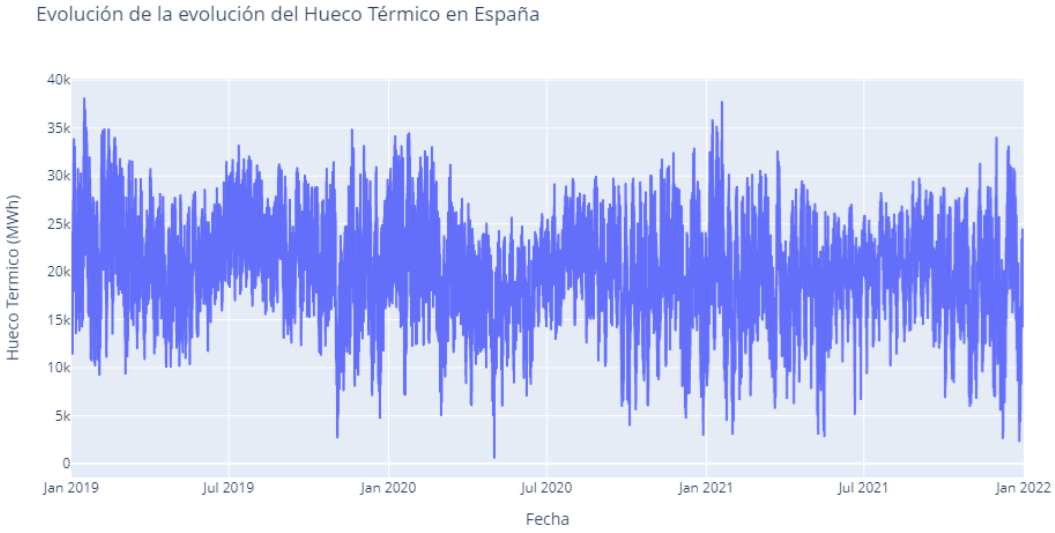


Imagen 9. Evolución del hueco térmico previsto D+1 desde 2019

### 4.2.5. Dummies, festivos y otras

Como variables dummies se escogieron la hora del día (desde las 0:00 hasta las 23:00) y el de la semana en que nos encontremos (de lunes a domingo). Estas variables valen 1 cuando cumplan estén en la hora o día de la semana correspondiente y 0 en el resto de los casos.

La misma idea ocurre con los festivos (distinguiendo entre festivos nacionales y regionales por comunidad autónoma introducidos a mano en un Excel). La fecha a predecir que se encuentre en este Excel tendrá en todas las horas de ese día la variable 1 y, en caso contrario, de 0.

Otras variables consideradas han sido el precio máximo y mínimo. Para predecir D+1 se usa como variable para todas las horas del día de D+1 el precio máximo y precio mínimo del día en el que nos encontremos, definidas dos variables distintas con el objetivo de poner unos niveles entre los que, más o menos, debería moverse el precio. También se incluye la diferencia entre estas dos como otra variable (llamada Spread_precio) que recogería si

el día anterior tuvo mucha volatilidad (la diferencia entre el precio máximo y mínimo del día anterior fue muy alta) o poca volatilidad (estas dos variables tuvieron valores similares).

### 4.2.6. Climatológicas

Se incluyeron datos acerca de condiciones meteorológicas, ya que éstas están directamente relacionadas con la producción de las renovables. Se obtuvieron de la web de datos meteorológicos de Madrid[6]. En ella se encuentran un montón de datos para distintas estaciones meteorológicas. Se escogió la estación meteorológica de Alcobendas y se obtuvieron las variables de velocidad del viento, temperatura, humedad relativa, radiación solar y precipitación (más información en el pdf en la ruta del repositorio 'Data/documentación_datos_climatologicos.pdf'). Los datos son históricos y a nivel horario, por lo que para usarlos como predicción se cogió su retardo de 24 horas. También se crearon las variables temperatura máxima, mínima y spread de temperatura con el mismo objetivo y procedimiento que el descrito para el precio en la sección 4.2.5.

### 4.2.7. Financieras

Las variables financieras se obtuvieron mediante el módulo *yfinance* de Python, una api que carga directamente los datos financieros de *Yahoo Finance*. Las variables que se obtuvieron fueron el IBEX35 (por encontrarnos en España), el BRENT (petróleo crudo) y el API2 (cotización del carbón) que es la que está directamente relacionada con las centrales térmicas. Otras variables que podrían haberse incluido son el TTF (precio del gas que se suele comerciar en Europa, que se decidió no incluir por tener ya el gas que se comercializa en España) o el EUA $CO_2$ (derechos de emisiones de $CO_2$ a la atmósfera, que tampoco se incluyó por no disponer tampoco de previsiones y no encontrarse en *Yahoo Finance*). Estas cotizaciones son diarias e históricas, por lo que, para usarlas para predecir y alinearlas al precio de la electricidad, también se tomaron sus retardos de 24 horas.

[6]https://datos.madrid.es/portal/site/egob/menuitem.c05c1f754a33a9fbe4b2e4b284f1a5a0/?vgnextoid=fa8357cec5efa610VgnVCM1000001d4a900aRCRD&vgnextchannel=374512b9ace9f310VgnVCM100000171f5a0aRCRD&vgnextfmt=default

## 4.3. Relación entre las variables

En toda la sección 4.2. se han descrito todas las variables que se recogieron como hipótesis que podrían tener sentido en la predicción del precio de la electricidad. Esta sección se basa en estudiar las relaciones de estas variables con la variable objetivo y seleccionar/descartar algunas, aunque muchos de los algoritmos que se usan en la actualidad ya tienen en cuenta la importancia de las variables y descartan (o no le dan ningún peso en la predicción) las variables que menos relación tienen con la variable a predecir.

En la sección 3 se mostró la serie temporal a nivel horario de la variable a predecir, en la que se intuía un periodo de alta volatilidad y precios altos en el último año. Esta información se ve de una forma mucho más clara en un gráfico de cajas o de violín como los siguientes.

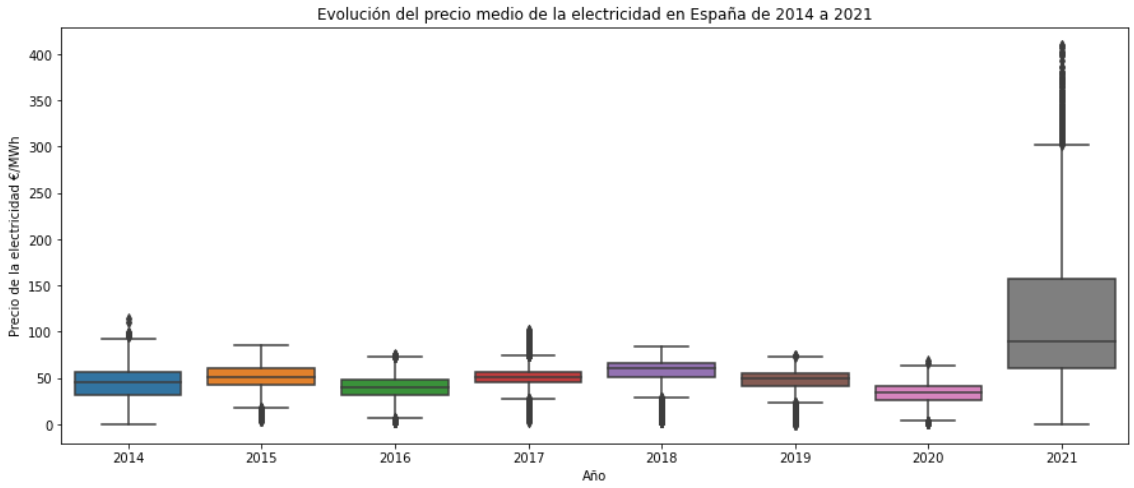


Imagen 10. Boxplot del precio de la electricidad desde 2014 hasta 2021. Elaboración propia.

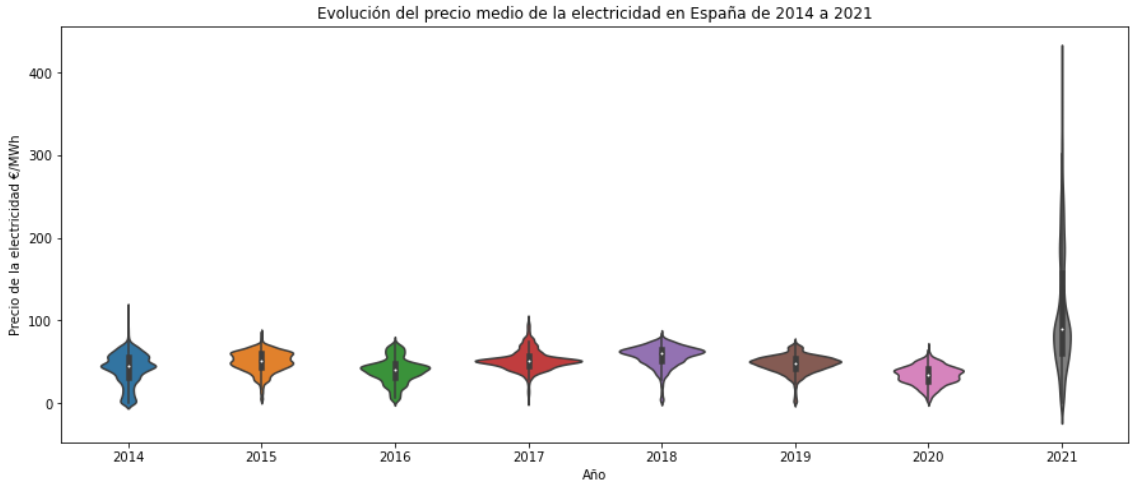


Imagen 11. Gráfico de violín del precio de la electricidad desde 2014 hasta 2021. Elaboración propia.

Resulta que la mediana del valor de la electricidad en 2021 es más alto que el máximo de cada uno de los años que le preceden, y que es sin duda el más volátil, tiene gran cantidad de valores que escapan del límite de para la detección de valores atípicos (que es, tanto superior como inferiormente, 1.5 veces el rango intercuartílico, calculado como la diferencia entre el primer y tercer cuartil de la distribución). También, por la posición de la mediana en el rango intercuartílico de 2021, entendemos que la distribución no es simétrica, a diferencia de como sí parece serlo en 2020.

Respecto los gráficos de violín, se contempla que las distribuciones están muy centradas en la mediana, excepto en 2021.

Para representar la relación entre las variables, se usa un pairplot del módulo seaborn, que crea una cuadrícula en la que cada variable numérica se comparte en los ejes x e y, de manera que escogiendo una fila y una columna puedes comparar una variable con otra (con un gráfico de dispersión). Los elementos de la diagonal, como son una comparación de una variable consigo misma, se representa mediante un histograma de dicha variable o una estimación de la densidad de distribución de la misma en caso de usar un parámetro para desglosar los datos.

Para no incluir demasiados gráficos, se muestran los resultados de este estudio en el Anexo B, desglosado con tres tipos de variables, las relacionadas con REE, las climatológicas y las financieras, para que el gráfico fuese lo suficientemente grande.

Tras esto, se realiza un estudio de las correlaciones, también por cada uno de los años de los que disponemos datos para todas las variables.

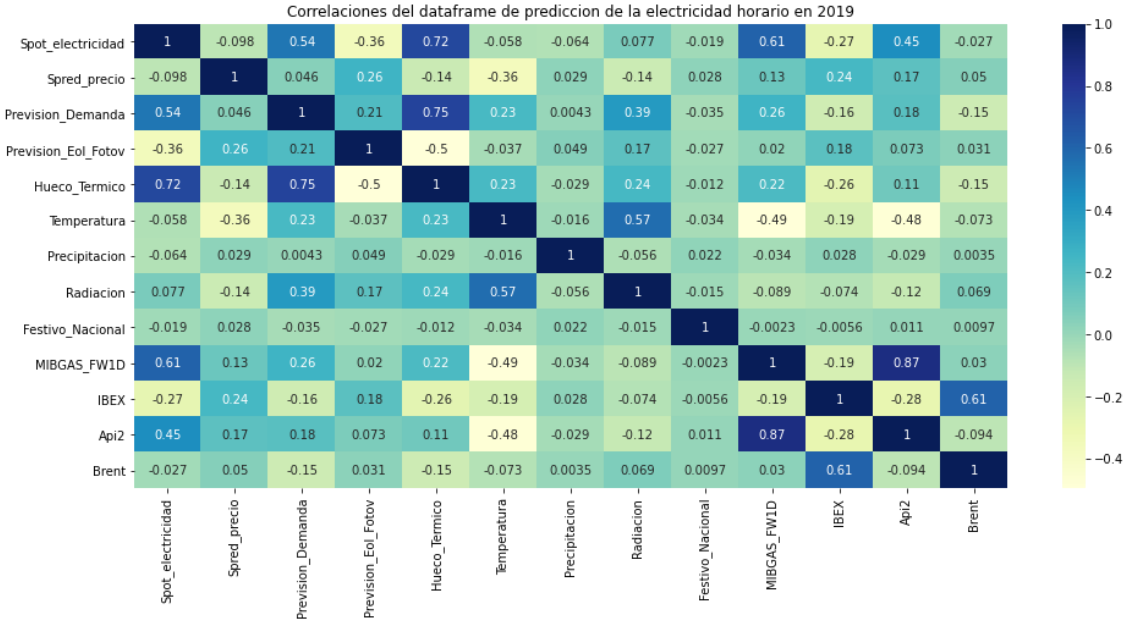


Imagen 12. Correlación entre las variables en el año 2019. Elaboración propia.

En el año 2019, se observa que las variables más importantes son el Hueco Térmico, la previsión de la demanda, el MIBGAS y el API2, en ese orden. Como se adelantó en las hipótesis, la variable del hueco térmico iba a ser bastante importante, con este estudio lo corroboramos que esta variable tiene un gran impacto en la variable a explicar. Las variables meteorológicas, al igual que los festivos no parecen seguir una correlación alta con la variable a predecir, estas pueden servir para afinar alguna predicción, como se mencionó en el Anexo B.

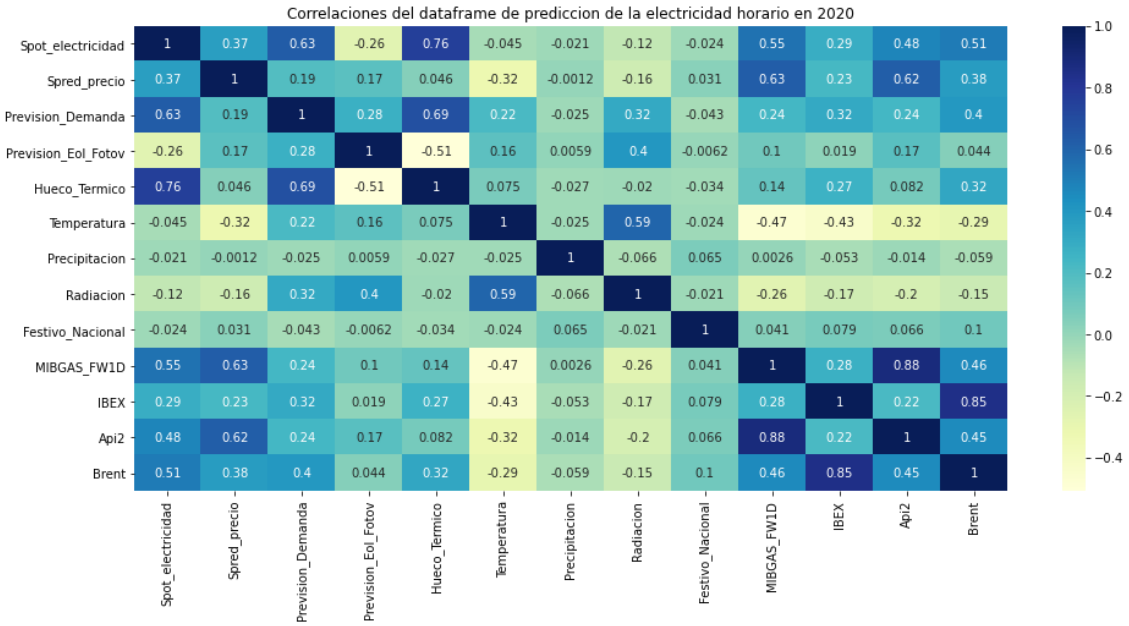


Imagen 13. Correlación entre las variables en el año 2020. Elaboración propia.

En el año 2020, las variables nombradas anteriormente siguieron manteniendo la misma correlación con la variable objetivo. Sin embargo, han aparecido otras que han cambiado de forma drástica su tendencia. Por ejemplo, el IBEX ha pasado de tener una correlación negativa de 0.27 a una correlación positiva de 0.29, por lo que ha cambiado totalmente su tendencia. Otra variable que parecía no estar para nada correlada con el precio de la electricidad en 2019, el Brent, ha pasado de prácticamente 0 a un valor de 0.51 en 2020, al igual que la variable del spread del precio, que ha evolucionado de -0.1 a 0.37.

Esto está dando una pista y es que, en el pasado, la variable de hueco térmico era indiscutiblemente la que más relación tenía con el precio de la electricidad, pero desde el periodo del Covid, las variables financieras son las que han tomado el protagonismo, el cual se acentuará aún más en el siguiente año a estudiar, en el que se han disparado los valores de las cotizaciones de casi todos los productos financieros.

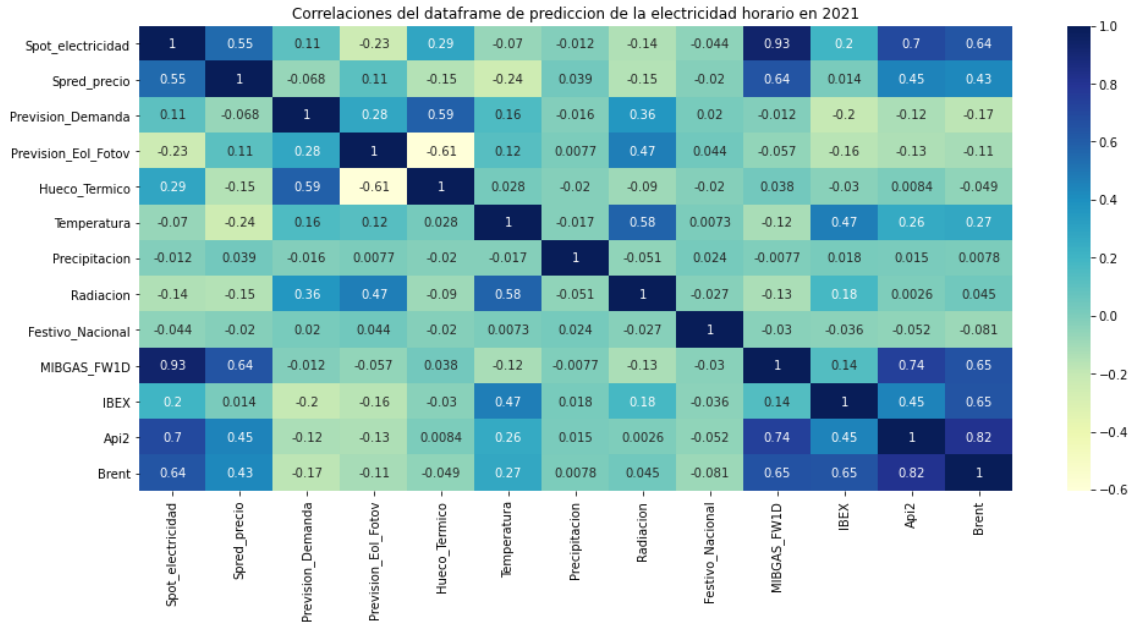


Imagen 14. Correlación entre las variables en el año 2021. Elaboración propia.

En el año 2021 se observa un claro cambio en la tendencia. La variable de hueco térmico ha pasado de tener una correlación de 0.72 y 0.76 en 2019 y 2020, respectivamente, a un valor de 0.29. Por su parte, la variable de MIBGAS ha pasado de 0.61 y 0.55 en 2019 y 2020, respectivamente, a 0.93. Lo mismo ocurre con la correlación del Api2, Brent y valor de spread de con el spot de la electricidad en este año.

Esto manifiesta un cambio de tendencia en la función de distribución de las variables y un cambio en la relación existente entre ellas, por lo que va a dificultar aún más el backtesting, pues si vamos a entrenar con los datos de

2019 y 2020 que seguían una relación entre las variables más o menos estables y se va a usar como predicción el año 2021 que tiene otra relación entre las variables, es de esperar que no se obtengan tan buenos resultados.

Una manera de solucionar esto es entrenar los modelos con ventanas de entrenamiento móviles fijas (o Rolling Windows) fijas, de manera que el modelo se vaya adaptando en todo momento a los cambios de tendencia que hay entre las relaciones de las variables.

Por ejemplo, aunque se tengan datos desde 2019, tras este estudio de la relación de las variables, es de esperar que un modelo entrenado con los datos de los 4 o 5 meses más recientes sea mejor que un modelo que tenga en cuenta los datos desde 2019 y, pensándolo, tiene más sentido que para predecir los valores del precio de la luz de mañana se consideren únicamente los valores más recientes. Estos diferentes estudios se realizarán con los distintos modelos a probar.

## 5. Resultados

Para todos los modelos que se van a mostrar se realiza un backtesting horario out-of-sample desde el 2021-01-01 hasta el 2021-12-31, calculando todas las métricas definidas en la sección 3, provenientes del script disponible del repositorio *métricas.py*.

El mejor modelo resultante de cada sección exportará sus resultados en un archivo .csv a la carpeta 'Resultados' del directorio desde junio de 2020 hasta diciembre de 2021, con el objetivo de usar los meses de predicción de 2020 como variables regresoras de un stacked model.

Se han probado modelos con todos los datos posibles y con ventanas móviles fijas de entrenamiento de 7, 14, 30, 60, 90, 150, 220 y 365 días.

También, para este mejor modelo de cada sección, se realiza un gráfico por cada mes de 2021 en el que se muestra la serie predicha con la real para ver los resultados de las predicciones y el valor del WMAPE resultado, que se encuentra disponible en el notebook destinado a cada modelo.

### 5.1. Modelo baseline

Este modelo se puede encontrar en el notebook *Modelo_Baseline.ipynb* que se puede encontrar en el repositorio.

La idea de este modelo es que la predicción para las horas del día de mañana sea igual al valor de las horas de hoy. Matemáticamente se formula, siendo 'y' el valor predicho y 'x' el valor real:

$$y(t) = x(t - 24) \quad (8)$$

Es un modelo muy sencillo pero que sirve para establecer unos valores de métricas iniciales e intuir si merece la pena aplicar algún modelo para mejorarlo o si este modelo ya es suficientemente bueno y es complicado de mejorar.

Los resultados obtenidos por este modelo en todo 2021 son los siguientes:

| MODELO | MAE | MAE (median) | WMAPE (%) | RMSE | % TREND |
|---|---|---|---|---|---|
| BASELINE | 16.07 | 8.92 | 14.36 | 27.38 | 74.84 |

Tabla 2. Métricas obtenidas del modelo baseline

Las métricas fueron las que se definieron y se calculan en el script *metricas.py*. Se obtiene el valor del MAE (medio) y MAE (mediana) y se

observa que hay valores de errores absolutos muy altos, pues la media es mucho mayor que la mediana, lo que indica la existencia de ciertos valores muy altos. El resto de métricas establecen un punto de partida a mejorar para el resto de modelos a desarrollar.

## 5.2. Modelo de regresión lineal múltiple

Este modelo se puede encontrar en el notebook *Modelo_Regresion_Lineal.ipynb* del repositorio.

El siguiente modelo probado también es bastante sencillo y consiste en una regresión lineal. Estos modelos suponen una relación lineal entre las variables explicativas y la variable a predecir. Consiste en un término independiente y unos coeficientes que multiplican a las variables explicativas. Estos coeficientes miden el efecto promedio que tiene el incremento en una unidad de la variable predictora 'x' sobre la variable dependiente 'y', manteniéndose constantes el resto de variables. Se conocen como coeficientes parciales de regresión.

Matemáticamente puede formularse como sigue, siendo 'y' la variable a predecir, 'x' las variables explicativas, 'N' el número de variables explicativas, 'a' los coeficientes de regresión, 'b' el término independiente y '$\varepsilon$' el residuo o error, la diferencia entre el valor observado y el estimado por el modelo:

$$y_j = b + \sum_{i}^{N} a_i * x_{ij} + \varepsilon_j \quad (9)$$

El objetivo de este método es obtener los parámetros 'a' y 'b' que hagan el cuadrado de los residuos mínimo (método de los mínimos cuadrados). Hay que elevar al cuadrado porque si la diferencia entre el valor estimado y el real es aleatoria, su media será o estará muy próxima a cero.

Para evitar multicolinealidad, se incluyó solo el hueco térmico (las previsiones de demanda y generación de renovables no se incluyeron) y se escogió tanto el spread de precio como el de temperatura, eliminando los máximos y mínimos de esas variables, ya que se encuentran incluidos ya en la construcción del spread.

A las variables ya consideradas se le sumaron los retardos a 24, 48 y 168 horas (1, 2 y 7 días) y con ello se propuso el modelo de regresión múltiple. No se incluyen el resultado de todos los coeficientes por ser tan numerosos,

pero el resumen de la regresión fue el siguiente (se puede consultar completo en el notebook):

OLS Regression Results

| | | | |
|---|---|---|---|
| **Dep. Variable:** | Spot_electricidad | **R-squared:** | 0.954 |
| **Model:** | OLS | **Adj. R-squared:** | 0.954 |
| **Method:** | Least Squares | **F-statistic:** | 1.325e+04 |
| **Date:** | Sat, 14 May 2022 | **Prob (F-statistic):** | 0.00 |
| **Time:** | 16:41:16 | **Log-Likelihood:** | -1.0239e+05 |
| **No. Observations:** | 26256 | **AIC:** | 2.049e+05 |
| **Df Residuals:** | 26214 | **BIC:** | 2.052e+05 |
| **Df Model:** | 41 | | |
| **Covariance Type:** | nonrobust | | |

Imagen 15. Resultado de la regresión lineal múltiple. Elaboración propia.

Resulta un $R^2$ de 0.954, lo que muestra que el 95.4 % de la varianza de la variable a explicar se explica mediante las variables explicativas, lo que parece un buen modelo.

Además de este modelo que, a priori tendría más sentido, al disponer de todas las variables preparadas y cargadas, se probaron más modelos quitando algunas variables y probando el backtesting y se encontraron modelos con mejores métricas resultantes en el backtesting.

Para diferenciarlo en los resultados, a este primer modelo propuesto se le denominará con el grupo de variables **Var 1**. Otro modelo probado se realizó sin quitar ninguna de las variables que consideramos al principio, incluyendo todo el dataframe de predicción, al que denominaremos grupo de variables **Var 2**. Por último, otro grupo de variables, que fueron todas excepto Brent, Demanda, Eólica, Festivo Regional, Humedad Relativa, Velocidad del viento, Radiación y Precipitación, que denominamos como grupo de variables **Var 3**. Se escogieron algunas de estas porque eran, analizando los test de importancia de cada variable, los que menos significativos eran o que tenían muestras de multicolinealidad.

Los valores del campo Rolling Window, cuando es distinto de 'No' (que no hay Rolling window), el número hace referencia al número de días hacia atrás que se está seleccionando como entrenamiento.

| Grupo Variables | Rolling window | MAE | MAE (median) | WMAPE | RMSE | % TREND |
|---|---|---|---|---|---|---|
| Var 1 | No | 12.23 | 8.25 | 10.93 | 20.30 | 83.42 |
| Var 1 | 7 | 17.99 | 9.32 | 16.08 | 41.04 | 81.45 |
| Var 1 | 14 | 60.36 | 18.8 | 53.93 | 221.43 | 81.62 |
| Var 1 | 30 | 11.43 | 7.73 | 10.21 | 18.03 | 83.79 |
| Var 1 | 60 | 10.88 | 7.46 | 9.72 | 16.88 | 84.22 |
| Var 1 | 90 | 10.57 | 7.41 | 9.45 | 16.58 | 84.20 |
| Var 1 | 150 | 10.83 | 7.43 | 9.68 | 17.00 | 84.11 |
| Var 1 | 220 | 10.77 | 7.63 | 9.62 | 16.97 | 84.06 |
| Var 1 | 365 | 11.49 | 7.84 | 10.26 | 18.68 | 83.75 |
| Var 2 | No | 11.54 | 7.16 | 10.31 | 19.83 | 83.81 |
| Var 2 | 7 | 21.21 | 11.00 | 18.95 | 44.50 | 80.39 |
| Var 2 | 14 | 38.45 | 15.49 | 34.35 | 86.00 | 81.86 |
| Var 2 | 30 | 12.70 | 8.36 | 11.35 | 20.20 | 84.05 |
| Var 2 | 60 | 10.89 | 7.24 | 9.73 | 17.02 | 84.45 |
| Var 2 | 90 | 10.41 | 7.10 | 9.30 | 16.40 | 84.29 |
| Var 2 | 150 | 10.27 | 6.76 | 9.17 | 16.46 | 84.29 |
| Var 2 | **220** | **10.21** | **6.78** | **9.12** | **16.59** | **83.99** |
| Var 2 | 365 | 10.76 | 6.68 | 9.61 | 18.13 | 83.56 |
| Var 3 | No | 12.02 | 8.06 | 10.74 | 19.99 | 82.76 |
| Var 3 | 7 | 19.99 | 9.78 | 17.86 | 48.60 | 81.45 |
| Var 3 | 14 | 173.29 | 24.26 | 154.82 | 1707.78 | 80.88 |
| Var 3 | 30 | 12.01 | 8.14 | 10.73 | 18.86 | 83.61 |
| Var 3 | 60 | 10.69 | 7.24 | 9.55 | 16.81 | 84.13 |
| Var 3 | 90 | 10.59 | 7.36 | 9.46 | 16.58 | 84.22 |
| Var 3 | 150 | 10.52 | 6.95 | 9.40 | 16.76 | 84.25 |
| Var 3 | 220 | 10.50 | 7.00 | 9.38 | 16.84 | 84.39 |
| Var 3 | 365 | 11.25 | 7.39 | 10.05 | 18.52 | 84.16 |

Tabla 3. Métricas obtenidas del modelo de regresión lineal múltiple.

Este primer modelo de regresión lineal múltiple, a pesar de ser también bastante sencillo, nos aporta bastante información. La principal es que el modelo baseline lo ha mejorado considerablemente en todas las métricas que se están estudiando, en concreto, en la que más nos fijaremos en este trabajo, el WMAPE ha bajado de un 14.36 % hasta un 9.12 % que se ha encontrado con el mejor modelo considerado. También muestran estos resultados que, como se adelantó en el estudio de las variables, como la relación entre estas está cambiando con el tiempo, un Rolling Window obtiene mejores resultados que considerar todo el histórico. En este caso, se han optimizado las métricas para un Rolling window de 220 días para el modelo que emplea todas las variables.

## 5.3. Modelo de regresión Ridge y Lasso

Los modelos se pueden encontrar en el notebook *Modelo_Regresion_Ridge.ipynb* y *Modelo_Regresion_Lasso.ipynb* incluidos en el repositorio.

Al retener un subconjunto de predictores y descartar el resto, la selección de subconjuntos produce un modelo interpretable y con un error de predicción posiblemente menor que el modelo completo. Sin embargo, al tratarse de un proceso discreto (las variables se vuelven a entrenar o se descartan), suele presentar una alta varianza, por lo que no reduce el error de predicción del modelo completo. Los métodos de penalización/reducción son más continuos y no sufren tanto la alta variabilidad.

La **regresión Ridge (o regularización L2)** reduce los coeficientes de regresión imponiendo una penalización a su tamaño. Los coeficientes de la regresión minimizan una suma de cuadrados residual penalizada.

$$y_i = argmin\left\{\sum_{i=1}^{N}\left(y_i - b - \sum_{j=1}^{p} x_{ij}a_j\right)^2 + \lambda\sum_{j=1}^{p} a_j^2\right\} \quad (10)$$

Aquí, $\lambda \geq 0$ es un parámetro de complejidad que controla la cantidad de penalización: cuanto mayor sea el valor de lambda, mayor será la cantidad de penalización. Los coeficientes se reducen hacia cero (y entre sí). La idea de penalizar por la suma de cuadrados de los parámetros también se utiliza en las redes neuronales.

Una forma equivalente de escribir el problema de regresión Ridge es;

$$y_i = argmin\sum_{j=1}^{p}\left(y_i - b - \sum_{j=1}^{p} x_{ij}a_j\right)^2 \qquad sujeto\ a{:}\ \sum_{j=1}^{p} a_j^2 \leq s \quad (11)$$

que hace explícita la restricción de tamaño en los parámetros. Existe una correspondencia uno a uno entre los parámetros lambda en las ecuaciones. Cuando hay muchas variables correlacionadas en un modelo de regresión lineal, sus coeficientes pueden quedar mal determinados y presentar una alta varianza. Un coeficiente positivo muy grande en una variable puede ser cancelado por un coeficiente negativo igualmente grande en su prima correlacionada. Al imponer una restricción de tamaño a los coeficientes, se evita que se produzca este fenómeno.

La **regresión Lasso (o regularización L1)** es un método de regularización como el Ridge, pero sutiles pero importantes diferencias:

$$y_i = argmin \sum_{j=1}^{p} \left( y_i - b - \sum_{j=1}^{p} x_{ij} a_j \right)^2 \qquad sujeto\ a: \sum_{j=1}^{p} |a_j| \leq s \quad (12)$$

Al igual que en la regresión ridge, podemos reparametrizar la constante b estandarizando los predictores; la solución para 'b' es la estimación de 'y', y a partir de ahí ajustamos un modelo sin ordenada.

Obsérvese la similitud con el problema de la regresión Ridge. Esta última restricción hace que las soluciones no sean lineales en $y_i$, y se utiliza un algoritmo de programación cuadrática para calcularlas. Debido a la naturaleza de la restricción, hacer t suficientemente pequeño hará que algunos de los coeficientes sean exactamente cero. Por lo tanto, el Lasso hace una especie de selección de variables.

Por tanto, a grandes rasgos, la diferencia fundamental entre ambos algoritmos es que en la regularización Lasso los coeficientes pueden ser exactamente cero (eliminando variables del modelo), mientras que en la Ridge nunca serán exactamente cero, esto hace los modelos Lasso más interpretables y son un modelo de selección de variables. En estos modelos entrarán todas las variables, ya que el propio algoritmo realizará la selección de las variables. Para calcular estos valores de lambda se realiza validación cruzada usando mean squared error como métrica.

Tras la introducción de los métodos de regularización, se realiza un backtesting en el mismo periodo y usando todas las variables, con el objetivo de que el propio algoritmo vaya reajustando las penalizaciones en cada predicción realizada.

| Regularización | Rolling window | MAE | MAE (median) | WMAPE | RMSE | % TREND |
|---|---|---|---|---|---|---|
| Ridge | No | 11.55 | 7.16 | 10.32 | 19.84 | 83.0 |
| Ridge | 7 | 13.86 | 8.52 | 12.38 | 22.82 | 80.28 |
| Ridge | 14 | 24.41 | 12.0 | 21.81 | 43.64 | 81.40 |
| Ridge | 30 | 12.34 | 8.13 | 11.03 | 19.71 | 82.83 |
| Ridge | 60 | 10.80 | 7.16 | 9.65 | 16.96 | 83.20 |
| Ridge | 90 | 10.37 | 7.04 | 9.27 | 16.38 | 83.30 |
| Ridge | 150 | 10.26 | 6.76 | 9.17 | 16.46 | 83.45 |
| Ridge | **220** | **10.21** | **6.76** | **9.12** | **16.59** | **83.16** |
| Ridge | 365 | 10.77 | 6.71 | 9.62 | 18.14 | 82.92 |
| Lasso | No | 11.9 | 7.63 | 10.63 | 20.05 | 83.85 |
| Lasso | 7 | 12.84 | 7.48 | 11.47 | 22.68 | 82.48 |
| Lasso | 14 | 11.84 | 7.18 | 10.58 | 20.09 | 83.23 |

| | | | | | | |
|---|---|---|---|---|---|---|
| **Lasso** | 30 | 11.00 | 7.17 | 9.83 | 18.06 | 84.94 |
| **Lasso** | 60 | 10.61 | 7.15 | 9.48 | 16.80 | 84.87 |
| **Lasso** | **90** | **10.12** | **6.86** | **9.04** | **16.25** | **84.62** |
| **Lasso** | 150 | 10.36 | 7.03 | 9.25 | 16.58 | 84.04 |
| **Lasso** | 220 | 10.46 | 6.98 | 9.34 | 16.92 | 84.28 |
| **Lasso** | 365 | 11.18 | 7.18 | 9.99 | 18.46 | 83.85 |

Tabla 4. Métricas obtenidas de los modelos de regularización Ridge y Lasso.

Como se observa, se han obtenido métricas muy similares para las mismas ventanas de entrenamiento que en la regresión lineal. En general, vuelve a ocurrir lo que se había anticipado en el estudio de la relación entre las variables y es que el modelo se va a comportar mejor con una ventana de entrenamiento fija que usando todos los datos disponibles.

Se obtienen valores muy parecidos al mejor modelo de la regresión lineal, es más, parecen prácticamente las mismas métricas para la regresión Ridge. Sin embargo, para la ventana de entrenamiento de 90 días, resulta una pequeña mejora de la regresión Lasso respecto a la regresión lineal múltiple y la regresión Ridge. Esto puede ser debido a que este tipo de regularización elimina directamente las variables menos importantes y sea mejor realizar esto que penalizarlas con un peso cercano a cero, pero sin eliminarlas completamente de la predicción.

### 5.4. Modelo SARIMA

El modelo se puede encontrar en el notebook *Modelo_Sarima.ipynb* que se puede encontrar en el repositorio.

Los modelos ARIMA ofrecen otro enfoque para la predicción de series temporales. Estos modelos se basan en las autocorrelaciones de los datos. Los modelos ARIMA (Autoregresive Integrated Moving Average), como sus siglas indican, son modelos con componente autorregresiva, integrada y de media móvil. Se suelen expresar como ARIMA (p,d,q) haciendo referencia las letras a la parte autorregresiva, integrada y de media móvil, respectivamente.

En un **modelo de autorregresión**, se pronostica la variable de interés utilizando una combinación lineal de valores pasados de la variable. El término autorregresión indica que se trata de una regresión de la variable contra sí misma. Así, un modelo autorregresivo de orden 'p' puede escribirse como:

$$y_t = c + \varphi_1 y_{t-1} + \varphi_2 y_{t-2} + \cdots + \varphi_p y_{t-p} + \varepsilon_t \quad (13)$$

Donde el término $\varepsilon_t$ es ruido blanco. Es como una regresión múltiple, pero con valores retardados de $y_t$ como predictores.

En los **modelos de media móvil**, en lugar de usar los valores pasados de la variable de previsión en una regresión, este modelo usa los errores de previsión pasados en un modelo de tipo regresivo. Así, un modelo de media móvil de orden 'q' puede escribirse como:

$$y_t = c + \theta_1 y_{t-1} + \theta_2 y_{t-2} + \cdots + \theta_q y_{t-q} + \varepsilon_t \quad (14)$$

Donde el término $\varepsilon_t$ es ruido blanco. No se observan los valores de $\varepsilon_t$ por lo que no es realmente una regresión en el sentido habitual. Notar que cada valor de $y_t$ puede considerarse como una media móvil ponderado de los últimos errores de previsión (aunque los coeficientes no sumen normalmente uno).

Normalmente no es posible saber, simplemente a partir de un gráfico temporal, qué valores de 'p' y 'q' son apropiados para los datos. Sin embargo, a veces es posible utilizar el gráfico ACF y el gráfico PACF estrechamente relacionado, para determinar los valores adecuados de 'p' y 'q'.

Un **gráfico acf** muestra las autocorrelaciones que miden la relación entre $y_t$ e $y_{t-k}$ para diferentes valores de 'k'. Ahora bien, si $y_t$ e $y_{t-1}$ están correlacionados, entonces $y_{t-1}$ e $y_{t-2}$ también deben estarlo. Sin embargo, entonces $y_t$ e $y_{t-2}$ podrían estar correlacionados, simplemente porque ambos están conectados con $y_{t-1}$, y no por cualquier información contenida en $y_{t-2}$ que pudiera utilizarse en la previsión de $y_t$.

Para superar este problema, podemos utilizar autocorrelaciones parciales (**gráfico pacf**). Éstas miden la relación entre $y_t$ e $y_{t-k}$ después de eliminar los efectos de los retardos 1,2, 3, ..., k-1. Por tanto, la primera autocorrelación parcial es idéntica a la primera autocorrelación, porque no hay nada entre ellas que eliminar. Cada autocorrelación parcial puede estimarse como el último coeficiente de un modelo autorregresivo. En concreto, $\alpha_k$, el coeficiente de autocorrelación parcial, es igual a la estimación de $\varphi_k$ en un AR(k). En la práctica, existen algoritmos más eficientes para calcular $\alpha_k$ que ajustar todas estas autorregresiones, pero dan los mismos resultados.

El **término integrado 'd'** muestra una forma de hacer estacionaria una serie temporal no estacionaria: calcula las diferencias entre observaciones consecutivas, procedimiento conocido como **diferenciación**. La diferenciación puede ayudar a estabilizar la media de una serie temporal

eliminando los cambios en el nivel de una serie temporal y, por tanto, eliminando (o reduciendo) la tendencia y la estacionalidad.

Una **serie temporal estacionaria** es aquella cuyas propiedades estadísticas no dependen del momento en el que se observa la serie. Así, las series temporales con tendencias o con estacionalidad no son estacionarias: la tendencia y la estacionalidad afectarán al valor de la serie temporal en diferentes momentos (10).

Algunos casos pueden ser confusos, ya que una serie temporal con comportamiento cíclico (pero sin tendencia ni estacionalidad) es estacionaria. Esto se debe a que los ciclos no tienen una duración fija, por lo que antes de observa la serie no podemos estar seguros de dónde estarán los picos y los valles de los ciclos.

En este caso, la serie (imagen 5) en el periodo a predecir (2021) parece seguir un periodo cíclico diaria, tendencia y un aumento de varianza respecto a los años anteriores.

Sin embargo, los modelos ARIMA también pueden modelar datos estacionales, con los Seasonal ARIMA (SARIMA), que serán los tratados en esta sección. Un ARIMA estacional se forma incluyendo términos adicionales en los modelos ARIMA no estacionales. Se escribe como sigue:

$$\mathrm{SARIMA} = \mathrm{ARIMA}(\mathrm{p},\mathrm{d},\mathrm{q}) \cdot \mathrm{ARIMA}(\mathrm{P},\mathrm{D},\mathrm{Q})_{\mathrm{m}} \quad (15)$$

Siendo m = periodo estacional, en nuestro caso, m = 24 (tenemos datos horarios y la serie sigue un parecido ciclo diario). Se utilizan las mayúsculas para las partes estacionales del modelo y las minúsculas para las partes no estacionales. La parte estacional del modelo consiste en términos similares a los componentes no estacionales del modelo, pero que implican desplazamientos hacia atrás del periodo estacional. Los términos estacionales adicionales se multiplican simplemente por los términos no estacionales.

En la siguiente imagen se muestran la serie original, diferenciada regular y diferenciada regular más estacional.

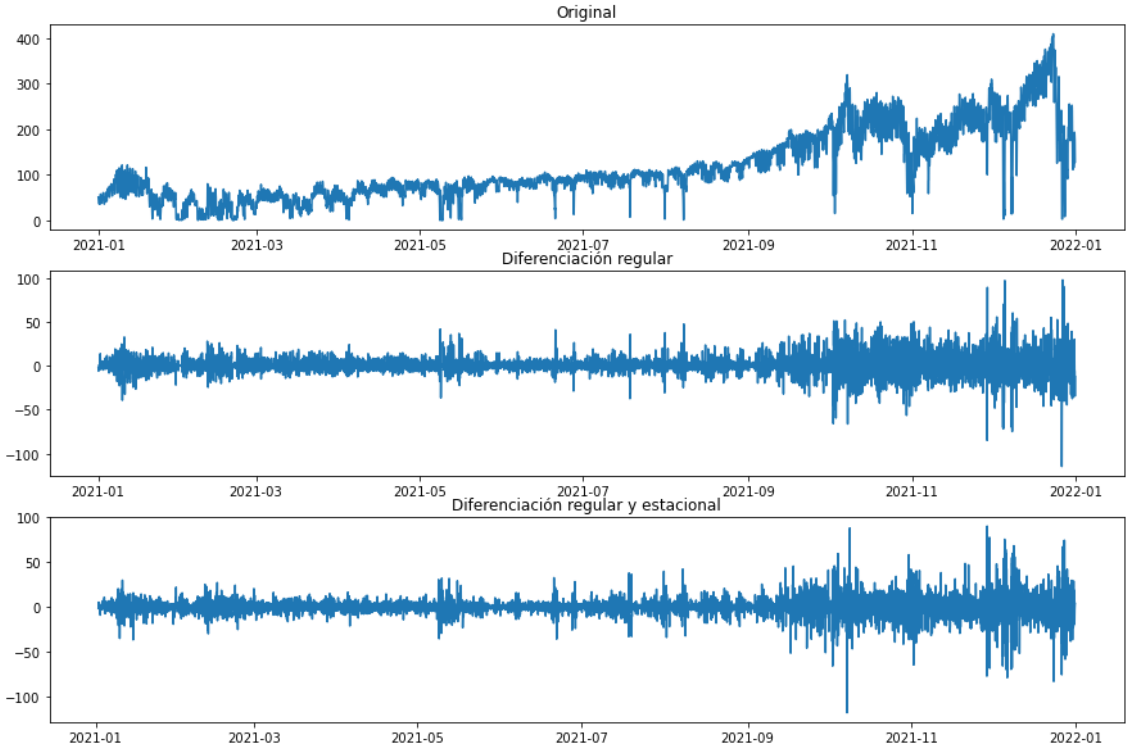


Imagen 16. Series temporales originales y diferenciadas del precio de la electricidad

La ACF y PACF de la serie original es la siguiente:

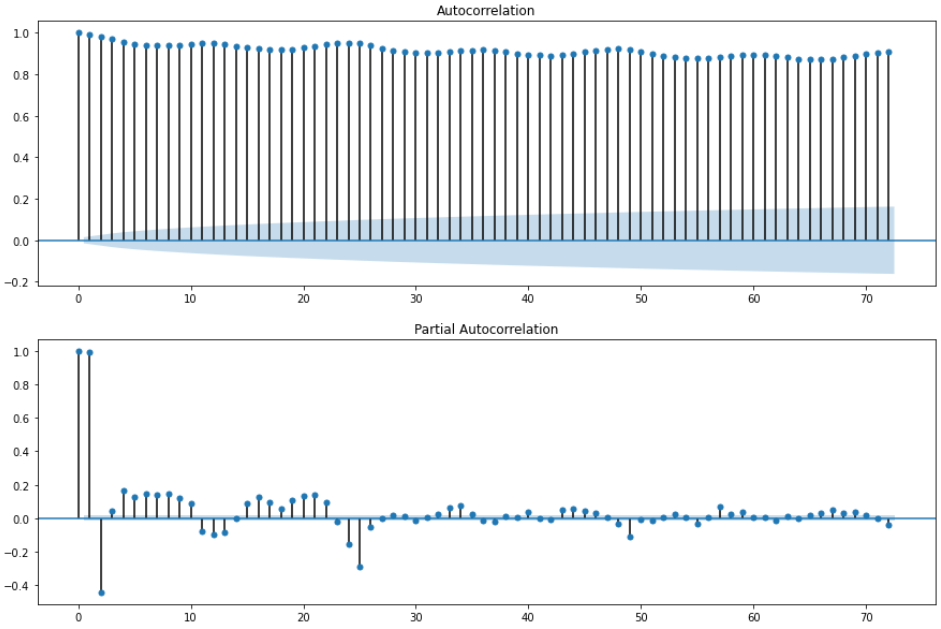


Imagen 17. Resultado de la ACF y PACF de la serie original

Analizando la serie original de la imagen 14, se observa una tendencia al alza y un aumento de la varianza, por lo que parece razonable pensar en una diferenciación regular o estacional, como las que aparecen inferiormente en

la misma imagen, en la que ya no observamos tendencia y se encuentra la serie centrada con media cero.

Con un modelo SARIMA(2,1,1)(0,1,1,24) (que hace referencia a las índices (p,d,q)(P,D,Q,m), los residuos del modelo son los siguientes:

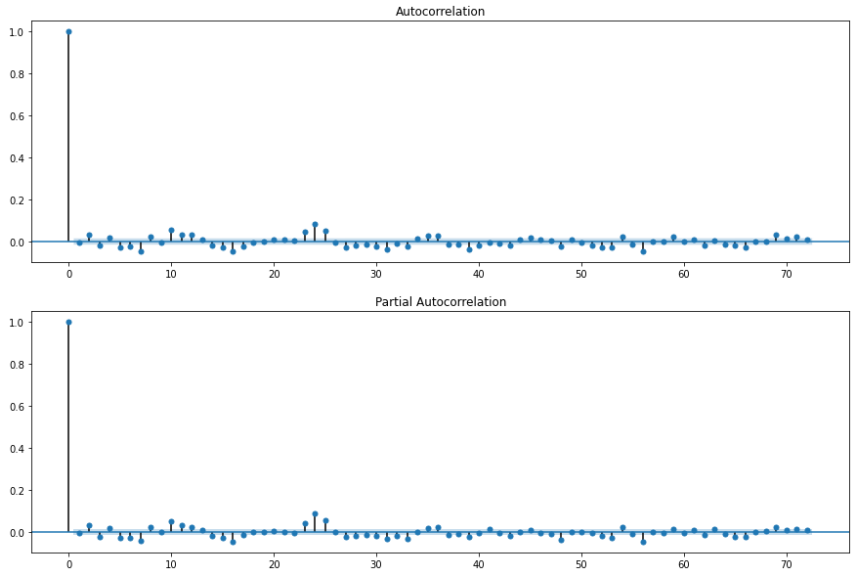


Imagen 18. Resultado de la ACF y PACF de los residuos del modelo SARIMA(2,1,1)(0,1,1,24)

No parece que exista más tendencia que capturar con el modelo, no hay retardos significativos salvo quizás, un retardo en el lag = 24. Esto puede solucionarse aumentando el término P, ya que coge los retardos estacionales. Obviando también la diferenciación estacional, resultan, la ACF y PACF de los residuos del modelo SARIMA(2,1,2)(2,0,1,24):

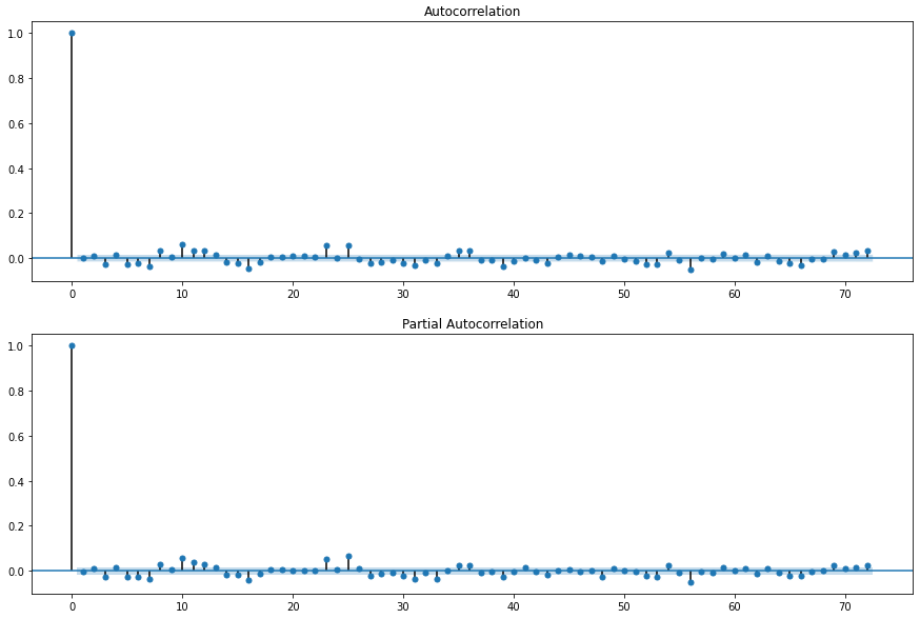


Imagen 19. Resultado de la ACF y PACF de los residuos del modelo SARIMA(2,1,2)(2,0,1,24)

Con estos modelos, se realizaron las mismas pruebas de backtesting que los modelos anteriores. Los resultados se recogen en la siguiente tabla:

| SARIMA (p,d,q)(P,D,Q,m) | Rolling window | MAE | MAE (median) | WMAPE | RMSE | % TREND |
|---|---|---|---|---|---|---|
| (2,1,1)(0,1,1,24) | No | 15.61 | 10.28 | 13.94 | 24.10 | 83.27 |
| (2,1,1)(0,1,1,24) | 7 | 15.29 | 9.22 | 13.66 | 24.19 | 81.64 |
| (2,1,1)(0,1,1,24) | 14 | 14.74 | 8.97 | 13.17 | 23.49 | 82.73 |
| (2,1,1)(0,1,1,24) | 30 | 14.16 | 8.72 | 12.65 | 22.78 | 82.91 |
| (2,1,1)(0,1,1,24) | **60** | **14.15** | **8.66** | **12.64** | **22.72** | **83.10** |
| (2,1,1)(0,1,1,24) | 90 | 14.30 | 8.79 | 12.78 | 23.11 | 82.76 |
| (2,1,1)(0,1,1,24) | 150 | 14.58 | 8.73 | 13.03 | 23.82 | 82.68 |
| (2,1,1)(0,1,1,24) | 220 | 14.42 | 8.64 | 12.88 | 23.46 | 82.70 |
| (2,1,1)(0,1,1,24) | 365 | 15.59 | 10.21 | 13.93 | 24.10 | 83.29 |

Tabla 5. Métricas obtenidas de los SARIMA.

No se obtienen buenos resultados usando únicamente los retardos regulares y estacionales de la propia variable. Según los gráficos ACF y PACF de los modelos considerados, no puede explicarse más a través de su pasado, son necesarias variables exógenas para conseguir una predicción más cercana a la realidad.

### 5.5. Modelo SARIMA de los residuos de la regresión lineal

El modelo se puede encontrar en el notebook *Modelo_RL_Sarima_Residuos.ipynb* del repositorio.

Este modelo consiste en, una vez realizada la regresión lineal múltiple de la sección 5.2, aplicarle un modelo SARIMA a los residuos, de manera que la predicción final sería, además de la ecuación 9, aplicarle un SARIMA al término $\varepsilon_j$.

$$y_j = b + \sum_{i}^{N} a_i * x_i + \varepsilon_j \quad (9)$$

Si observamos los residuos de la regresión lineal de la sección 5.2, hacemos un gráfico de la serie, un histograma de los residuos y un gráfico de la ACF y PACF, resulta:

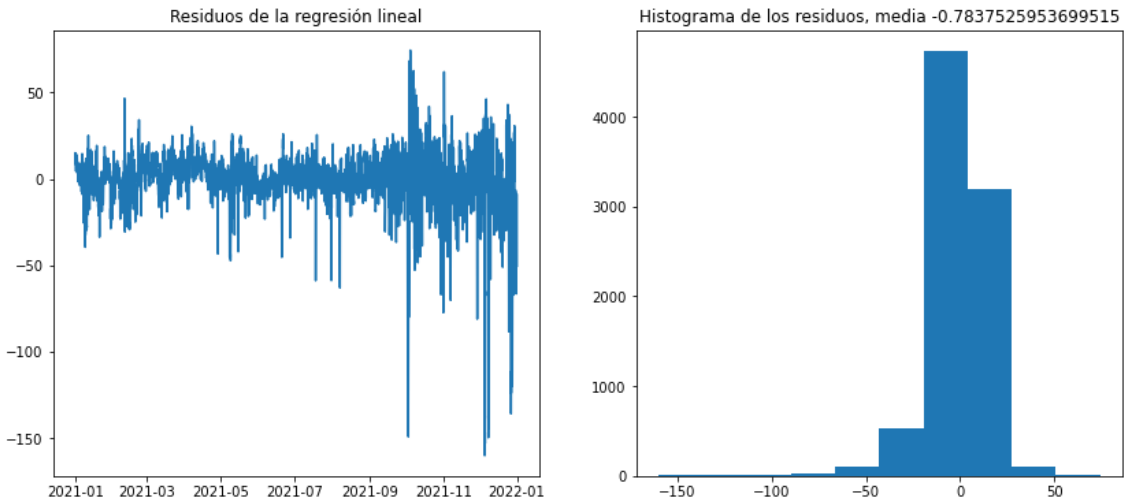


Imagen 20. Gráfico e histograma de los residuos del mejor modelo de regresión lineal múltiple

Los residuos están más o menos centrados en cero, pero parece apreciarse cierta tendencia que podría ser capturada por otro modelo. Esto se podría ver mejor con la ACF y PACF de los residuos de la regresión lineal.

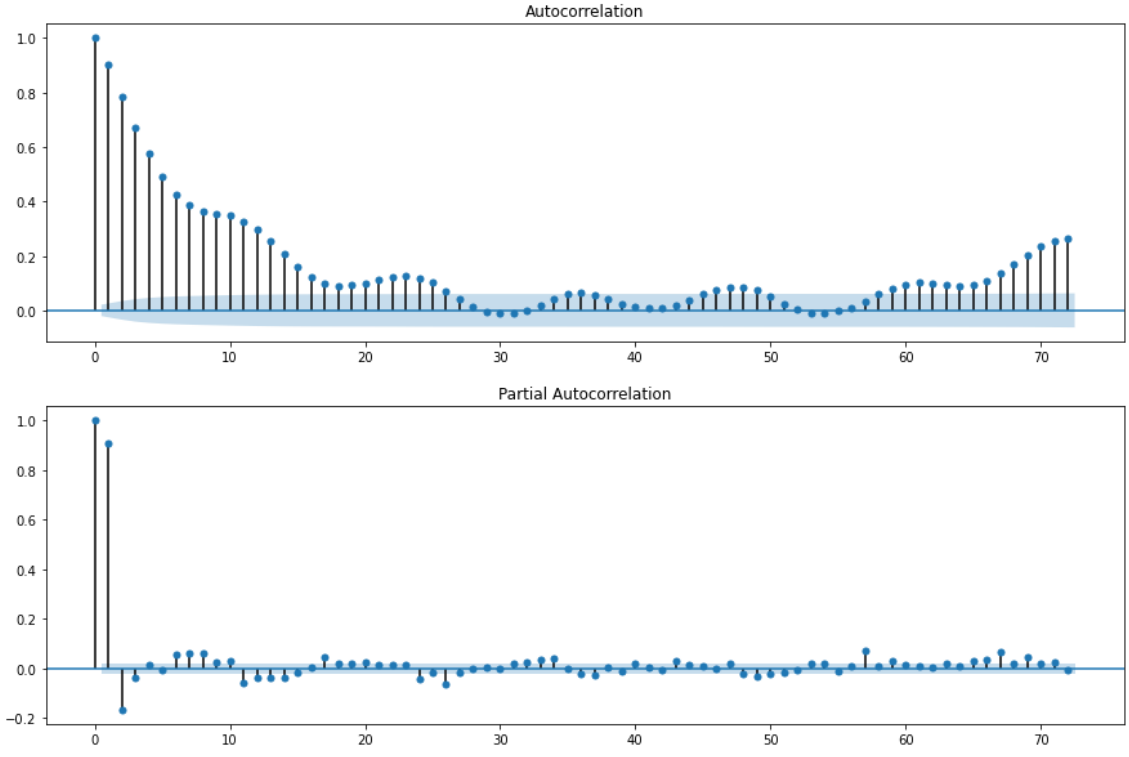


Imagen 21. ACF y PACF de los residuos del mejor modelo de regresión lineal múltiple

Como se observa, aún quedan autocorrelaciones y tendencia que puede seguir siendo capturada por otro modelo, como en este caso, un ARIMA o un SARIMA. Este es el objetivo de este modelo, predecir la serie de los residuos.

Los resultados del modelo aplicado, realizando el Rolling window

| SARIMA | Rolling window | MAE | MAE (median) | WMAPE | RMSE | % TREND |
|---|---|---|---|---|---|---|
| (7,0,0) (1,0,0,24) | 7 | 10.43 | 6.81 | 9.31 | 17.05 | 83.43 |
| (7,0,0) (1,0,0,24) | 14 | 10.33 | 6.72 | 9.23 | 16.99 | 83.62 |
| (7,0,0) (1,0,0,24) | 30 | 10.38 | 6.70 | 9.27 | 16.95 | 83.75 |
| (7,0,0) (1,0,0,24) | 60 | 10.29 | 6.66 | 9.20 | 16.69 | 83.80 |
| (7,0,0) (1,0,0,24) | 90 | 10.27 | 6.67 | 9.17 | 16.86 | 83.82 |
| (7,0,0) (1,0,0,24) | **150** | **10.24** | **6.62** | **9.15** | **16.85** | **83.83** |
| (7,0,0) (1,0,0,24) | 220 | 10.26 | 6.60 | 9.17 | 16.89 | 83.89 |
| (7,0,0) (1,0,0,24) | 365 | 10.26 | 6.63 | 9.17 | 16.91 | 83.95 |

Tabla 6. Métricas obtenidas del SARIMA de los residuos del mejor modelo de regresión lineal.

En este caso, se observa que las predicciones han empeorado ligeramente respecto a considerar únicamente la regresión lineal como modelo explicativo. Esto puede ser debido a que la corrección que intenta implementar este SARIMA posterior lo que está es desviando un poco más la estimación del valor real, por lo que no realiza una mejora sobre el modelo en que se basa.

### 5.6. Modelo Random Forest

El modelo se puede encontrar en el notebook *Modelo_RF.ipynb* que se puede encontrar en el repositorio.

El **modelo random forest** es un método basado en árboles usado para regresión y clasificación. Estos incluyen estratificar o segmentar el espacio de predicción en una serie de regiones simples. Para hacer una predicción de una determinada observación, se suele utilizar la media o la moda del valor de respuesta de las observaciones de entrenamiento en la región a la que pertenece. Dado que el conjunto de reglas de división utilizadas para segmentar el espacio de predicción puede resumirse en un árbol, este tipo de enfoques se conocen como métodos de árbol de decisión.

Este enfoque implica la producción de múltiples árboles que luego se combinan para obtener una única predicción de consenso. La combinación de un gran número de árboles puede dar lugar a menudo a mejoras drásticas en la precisión de la predicción, a expensas de una cierta pérdida en la interpretación.

Los modelos random forest son una modificación sustancial del *bagging* que construye una gran colección de árboles descorrelacionados y luego los promedia. En muchos problemas, el rendimiento de los bosques aleatorios es muy similar al del *boosting*, y son más sencillos de entrenar y ajustar. En consecuencia, son algoritmos muy populares en este tipo de problemas.

La idea esencial del *bagging* es promediar muchos modelos ruidosos, pero aproximadamente insesgados y, por lo tanto, reducir la varianza. Los árboles son candidatos ideales para el *bagging*, ya que pueden capturar estructuras de interacción complejas en los datos y, si crecen lo suficiente, tienen un sesgo relativamente bajo. Dado que los árboles son notoriamente ruidosos, se benefician en gran medida del promediado.

Además, dado que cada árbol generado en el embolsamiento está idénticamente distribuido, es decir, la expectativa de un promedio de B de tales árboles es la misma que la expectativa de cualquiera de ellos. Esto significa que el sesgo de los árboles ensacados es el mismo que el de los árboles individuales, y la única esperanza de mejora es la reducción de la vairanza. Esto contrasta con el *boosting*, en el que los árboles crecen de forma adaptativa para eliminar el sesgo y, por lo tanto, no son idénticos.

Una media de B variables aleatorias, es decir, con una varianza $\sigma^2$, tiene una varianza $\frac{1}{B}\sigma^2$. Si las variables son idénticamente distribuidas, pero no necesariamente independientes, con correlación positiva entre pares $\rho$, la varianza de la media es:

$$varianza = \rho\sigma^2 + \frac{1-\rho^2}{B}\sigma^2 \quad (16)$$

Cuando B crece, el segundo término desaparece, pero el primero permanece, y por lo tanto el tamaño de la correlación de los pares de árboles embolsados limita los beneficios del promedio.

La idea en los bosques aleatorios es mejorar la reducción de la varianza del embolsado reduciendo la correlación entre los árboles, sin aumentar demasiado la varianza. Esto se consigue en el proceso de crecimiento de los árboles mediante la selección aleatoria de las variables de entrada.

Específicamente, cuando se hace crecer un árbol en un conjunto de datos Bootstrap:

- Antes de cada división se selección m $\leq$ p de las variables de entrada al azar como candidatas para la división.

Los resultados del random forest, en el backtesting del 2021 son los siguientes, para un modelo con los siguientes parámetros:

- Conf_1 → n_estimators = 150, criterion = "mae", max_depth = None, max_features = X_train.shape[1]- 1, n_jobs = -1, warm_start = False, random_state = 123
- Conf_2 → mismos parámetros que conf_1 pero cambiando max_features = X_train.shape[1] – 6 y usando target mean encoding.
- Conf_3 → mismos parámetros que conf_1 pero cambiando max_features = X_train.shape[1] – 10 y usando target mean encoding
- Conf_4 → misma configuración que conf_2 pero quitando las siguientes variables: "Festivo_Regional", "Humedad_Relativa", "Precipitacion", "Radiacion", "Velocidad_Viento",'Precio_max', 'Precio_min','Spred_precio', 'Temperatura', 'Temperatura_max', 'Temperatura_min','Spred_temperatura'.
- Conf_5 → misma configuración que conf_4 pero además sin las variables predictivas de previsión de demanda y previsión eólica y fotovoltaica.

| CONFIGURACIÓN | Target Mean | Rolling window | MAE | MAE (median) | WMAPE | RMSE | % TREND |
|---|---|---|---|---|---|---|---|
| Conf_1 | No | No | 11.12 | 6.89 | 10.13 | 19.24 | 80.28 |
| Conf_1 | No | 7 | 10.45 | 6.16 | 9.33 | 17.88 | 82.09 |
| Conf_1 | No | 14 | 10.22 | 5.94 | 9.13 | 17.67 | 83.08 |
| Conf_1 | No | 30 | 10.36 | 6.18 | 9.26 | 16.95 | 83.00 |
| Conf_1 | No | 60 | 10.60 | 6.31 | 9.47 | 17.49 | 83.29 |
| Conf_1 | No | 90 | 10.54 | 6.39 | 9.42 | 17.09 | 83.19 |
| Conf_1 | No | 150 | 10.48 | 6.38 | 9.36 | 17.1 | 83.15 |
| Conf_1 | No | 220 | 10.58 | 6.41 | 9.45 | 17.3 | 83.17 |
| Conf_1 | No | 365 | 10.75 | 6.50 | 9.61 | 17.56 | 82.87 |
| Conf_1 | Si | 30 | 10.36 | 6.25 | 9.26 | 16.97 | 83.64 |
| Conf_2 | Si | 30 | 10.19 | 6.08 | 9.11 | 16.77 | 84.01 |
| Conf_3 | Si | 30 | 10.34 | 6.20 | 9.23 | 17.1 | 84.47 |
| Conf_4 | **Si** | **30** | **10.02** | **6.19** | **8.95** | **16.34** | **83.59** |
| Conf_5 | Si | 30 | 10.20 | 6.29 | 9.11 | 16.76 | 83.87 |

Tabla 7. Métricas obtenidas del Random Forest Regressor

Inicialmente, en la conf_1 se estaban tratando las variables dummies, por lo que si se seleccionaban como número de variables por árbol los parámetros por defecto (como la raíz cuadrada de las variables predictivas o la mitad de las variables por árbol) resultaban valores bastante inferiores a los de la regresión lineal, por lo que, la primera configuración tiene como max_features el número de variables consideradas menos uno, lo que da lugar a pocos árboles descorrelados. Para solventarlo, a partir de la conf_2 (incluida) se sustituyen las variables dummies por el *target encoding*. Esta

técnica se basa en el proceso de reemplazar una variable categórica por el valor medio de la variable objetivo. Es decir, en vez de tener 24 columnas binarias, tenemos 1 en la que, para cada hora, tenemos la media de la variable objetivo en el dataset de entrenamiento para esa hora. Hay que tener cuidado porque tiene que calcularse con la media **únicamente del dataset de entrenamiento** ya que, si entra la parte de test o a predecir en el backtesting, se le estaría pasando una información que el modelo no podría conocer, por lo que no sería una predicción completamente *out-of-sample*.

De esta manera, se sustituyen las 24 variables categóricas de la hora y las 7 del día de la semana por 2 variables a través de este proceso. De esta manera, ya se pueden usar menos variables para cada árbol y obtener realmente lo atractivo de este modelo, que es realizar la media de muchos *weak learners*. Así se obtienen, como era de esperar, mejores resultados que mediante la primera configuración y, si además, se eliminan manualmente las variables que a priori son menos importantes, como en la configuración 4, se obtiene el mejor modelo hasta ahora con un valor de WMAPE del 8.95 % con una ventana de entrenamiento móvil de 30 días.

## 5.7. Modelo XGBoost

El modelo se puede encontrar en el notebook *Modelo_XGBoost.ipynb* que se puede encontrar en el repositorio.

El *boosting* es una de las ideas de aprendizaje más potentes introducidas. Se diseñó originalmente para problemas de clasificación, pero, también puede extenderse de forma provechosa a la regresión. La motivación del *boosting* fue un procedimiento que combina las salidas de muchos clasificadores 'débiles' para producir un poderoso output.

El algoritmo XGBoost (eXtreme Gradient Boosting) utiliza el algoritmo del árbol de decisión que aumenta el gradiente. El método de aumento de gradiente crea nuevos modelos que hacen la tarea de predecir los errores y los residuos y luego realiza la predicción final (13), es decir, está formado por un conjunto de árboles de decisión individuales, entrenados de forma secuencial, de forma que cada nuevo árbol trata de mejorar los errores de los árboles anteriores. La predicción de una nueva observación se obtiene agregando las predicciones de todos los árboles individuales que forman el modelo. En cada árbol individual, las observaciones se van distribuyendo por bifurcaciones (nodos) generando la estructura del árbol hasta alcanzar un

nodo terminal. Para entender cómo funcionan es necesario conocer los conceptos de *ensemble* y *boosting* (14).

Todos los modelos de aprendizaje estadístico y machine learning sufren el problema de equilibrio entre bias y varianza. El término bias (sesgo) hace referencia a cuánto se alejan en promedio las predicciones de un modelo respecto a los valores reales. El término varianza hace referencia a cuánto cambia el modelo dependiendo de los datos utilizados en su entrenamiento. El mejor modelo es aquel que consigue un equilibrio óptimo entre bias y varianza.

Los métodos de *ensemble* combinan múltiples modelos en uno nuevo con el objetivo de lograr un equilibrio entre bias y varianza, consiguiendo así mejores predicciones que cualquiera de los individuales. Los dos tipos de *ensemble* más utilizados son el *bagging* (estudiado y usado con el Random Forest en la sección 5.6) y el *boosting* (que ocupa esta sección y el modelo XGBoost). Los algoritmos más empleados de boosting son *AdaBoost*, *Gradient Boosting* y *Stochastic Gradient Boosting*. Todos tienen muchos hiperparámetros, pero tres de los más importantes son: (14)

- Número de *weak learners* o número de iteraciones: a diferencia del random forest, un alto número aquí puede producir *overfitting*. Para evitarlo, se emplea regularización:
- *Learning rate*: controla la influencia que tiene cada *weak learner* en el conjunto del *ensemble*, es decir, el ritmo al que aprende el modelo. Suele estar entre 0.001 y 0.01 este valor.
- Sil os *weak learners* son árboles, el tamaño máximo permitido de cada árbol, suelen emplearse valores pequeños, entre 1 y 10.

Aunque el objetivo final es el mismo, lograr un balance óptimo entre bias y varianza, existen dos diferencias fundamentales:

- Forma en que consiguen el error total: el error total de un modelo puede descomponerse como bias + varianza + e.
    - En *bagging* se emplean modelos con poco bias pero mucha varianza.
    - En *boosting* se emplean modelos con poca varianza pero con mucho bias
- Forma en que se introducen variaciones en los modelos que forman el ensemble.

- En *bagging* cada modelo es distinto al resto porque cada uno se entrena con una muestra distinta obtenida mediante *bootstrapping*.
- En *boosting* los modelos se ajustan secuencialmente y la importancia (peso) de las observaciones va cambiando en cada iteración, dando lugar a diferentes ajustes.

De forma resumida, el algoritmo de *Gradient Boosting* (una generalización del algoritmo *AdaBoost*) es el siguiente:

Se ajusta un primer *weak learner* $f_1$ con el que se predice la variable respuesta *y*, y se calculan los residuos $y - f_1(x)$. A continuación, se ajusta un nuevo modelo $f_2$, que intenta predecir los residuos del modelo anterior, en otras palabras, trata de corregir los errores que ha tenido el modelo $f_1$. (14)

$$f_1(x) \approx y \quad (17)$$

$$f_2(x) \approx y - f_1(x) \quad (18)$$

En la siguiente iteración, se calculan los residuos de los dos modelos de forma conjunta $y - f_1(x) - f_2(x)$, los errores cometidos por $f_1$ y que $f_2$ no ha sido capaz de corregir, y se ajusta un tercer modelo $f_3$ para tratar de corregirlos.

$$f_3(x) \approx y - f_1(x) - f_2(x) \quad (19)$$

Este proceso se repite *M* veces, de forma que cada nuevo modelo minimiza los residuos (errores) del anterior.

Dadao que el objetivo de *Gradient Boosting* es ir minimizando los residuos iteración a iteración, es susceptible de *overfitting*. Para evitarlo, se usa el parámetro *learning rate*, que limita la influencia de cada modelo en el conjunto del *ensemble*. Como consecuencia, se necesitan más modelos para formar el *ensemble* pero se consiguen mejores resultados:

$$f_1(x) \approx y \quad (20)$$

$$f_2(x) \approx y - \lambda f_1(x) \quad (21)$$

$$f_3(x) \approx y - \lambda f_1(x) - \lambda f_2(x) \quad (22)$$

$$y \approx \lambda f_1(x) + \lambda f_2(x) + \cdots + \lambda f_m(x) \quad (23)$$

Los resultados del XGBoost en el backtesting del 2021 son los siguientes, para un modelo con los siguientes parámetros:

Conf_1 → n_estimators = 1000, max_depth = None, eta(learning rate) = 0.03, subsample = 0.7, colsample_bytree = 0.8 con todas las variables

Conf_2 → igual que la configuración 1 pero usando target mean encoding.

Conf_3 → igual que la configuración 1 pero usando las variables de la configuración 4 del random forest que fue la que obtuvo las mejores métricas.

| VARIABLES | Rolling window | MAE | MAE (median) | WMAPE | RMSE | % TREND |
|---|---|---|---|---|---|---|
| Conf_1 | None | 10.10 | 6.19 | 8.96 | 16.48 | 83.45 |
| Conf_1 | 7 | 10.64 | 6.27 | 9.51 | 17.78 | 78.89 |
| Conf_1 | 14 | 10.01 | 5.99 | 8.95 | 16.58 | 81.33 |
| Conf_1 | 30 | 10.00 | 6.22 | 8.93 | 16.10g | 82.25 |
| Conf_1 | 60 | 10.09 | 6.31 | 9.02 | 16.14 | 82.61 |
| Conf_1 | 90 | 9.99 | 6.27 | 8.92 | 15.73 | 82.67 |
| Conf_1 | 120 | 9.93 | 6.25 | 8.87 | 15.50 | 83.13 |
| Conf_1 | **150** | **9.83** | **6.18** | **8.78** | **15.26** | **82.83** |
| Conf_1 | 220 | 9.96 | 6.22 | 8.90 | 15.82 | 82.94 |
| Conf_1 | 365 | 9.90 | 6.12 | 8.85 | 15.69 | 83.39 |
| Conf_2 | 150 | 9.85 | 6.26 | 8.80 | 15.33 | 83.26 |
| Conf_3 | 150 | 9.89 | 6.21 | 8.83 | 15.53 | 83.08 |

Tabla 8. Métricas obtenidas del XGBoost

En este modelo, a diferencia del random forest, por la naturaleza del algoritmo que lo entrena y cómo realiza la predicción, no importa dejar las variables dummies, es más, usar éstas o el target *mean* para la hora y el día de la semana resultan en prácticamente las mismas métricas obtenidas. En este caso la configuración de las mejores variables para el random forest no se encuentran para el XGBoost, en este caso se obtiene mediante la configuración 1 y una ventana móvil de entrenamiento de 150 días el mejor resultado con un WMAPE del 8.78 %.

Un resultado curioso de este modelo se encuentra a la hora de realizar el backtesting en el mes de junio. En cada uno de los notebooks se encuentra visualizada, para ese modelo, la mejor configuración y su backtesting mes a mes en todo 2021, para visualizar las predicciones de una manera más limpia. Comparando un modelo como la regresión lineal y este XGBoost resulta lo siguiente.

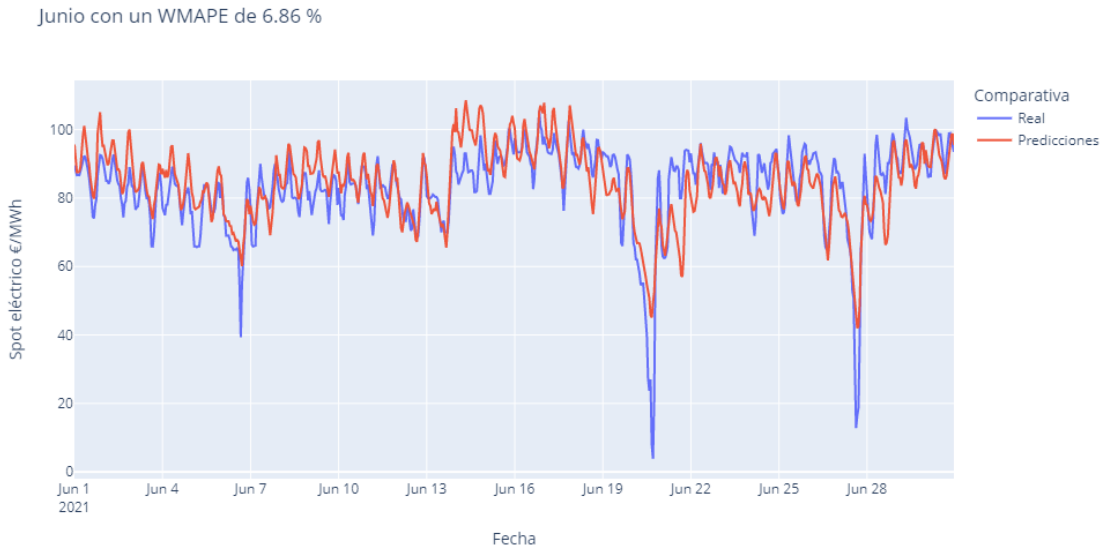


Imagen 22. Backtesting de junio 2021 realizado por la regresión lineal múltiple

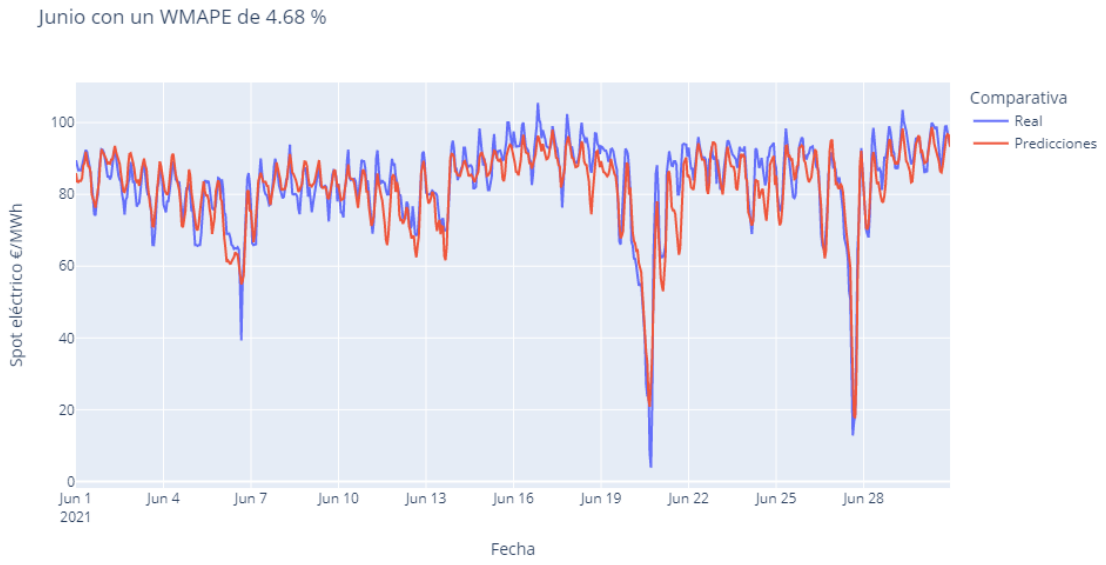


Imagen 23. Backtesting de junio 2021 realizado por el modelo XGBoost.

Se observan dos caídas muy bruscas en junio, una sobre el día 21 y otra sobre el día 28. La regresión lineal parece capturar esta tendencia correctamente, pero, si se observan las predicciones del XGBoost resulta, en concreto, para estos dos días atípicos de la serie, que este modelo más sofisticado se adecúa perfectamente a la tendencia. En concreto, para este mes, se obtiene un 6.86 % de WMAPE para la regresión lineal y un 4.68 % para el XGBoost.

## 5.8. Modelo de redes neuronales LSTM

El modelo se puede encontrar en el notebook *Modelo_LSTM.ipynb* que se puede encontrar en el repositorio[7].

Otra forma de proceder en la solución del problema de predicción del precio de la electricidad es mediante algoritmos de redes neuronales. Una característica importante de muchas de estas redes, como las densamente conectadas y las convnets, es que no tienen memoria. Cada entrada que se les muestra se procesa de forma independiente, sin mantener ningún estado entre las entradas. Con este tipo de redes, para procesar una secuencia o una serie temporal de puntos de datos, hay que mostrar toda la secuencia a la red de una vez: convertirla en un único punto de datos. (15)

Para resolver esto, nace la *Recurrent Neural Network* (RNN), que es una clase de red neuronal artificial en la que las conexiones entre unidades forman un gráfico dirigido a lo largo de una secuencia. Esta estructura le permite mostrar un comportamiento temporal dinámico para una secuencia de tiempo. A diferencia de las redes neuronales feedforward, las RNN pueden utilizar su estado interno para procesar secuencias de entradas. (16)

Sin embargo, las SimpleRNN tienen un problema importante: aunque teóricamente debería ser capaz de retener en el tiempo 't' información sobre entradas vistas muchos pasos de tiempo antes, en la práctica, tales dependencias a largo plazo son imposibles de aprender. Esto se debe al problema del desvanecimiento del gradiente, un efecto similar al que se observa en las redes no recurrentes (redes *feedforward*) que tienen muchas capas de profundidad: a medida que se van añadiendo capas a una red, ésta acaba por volverse intentrenable. (15)

Para solventar este problema, nacen las redes LSTM (y también las GRU). La LSTM es también una red neuronal recurrente, que se ha utilizado para resolver muchos problemas de secuencias temporales.

Esta capa es una variante de la capa SimpleRNN; añade una forma de llevar la información a través de muchos pasos de tiempo. Imagina una cinta transportadora que corre paralela a la secuencia que estás procesando. La información de la secuencia puede saltar a la cinta transportadora en cualquier punto, ser transportada a un paso de tiempo posterior, y saltar fuera, intacta, cuando la necesites. Esto es esencialmente lo que hace la

[7] Este código se ejecutó en Kaggle aprovechando la capacidad de GPU para entrenar más rápidamente este tipo de algoritmos.

LSTM: guarda la información para más tarde, evitando así que las señales más antiguas se desvanezcan gradualmente durante el procesamiento (15).

La estructura de la LSTM se muestra en la Imagen (en comparación con la SimpleRNN), y su funcionamiento se ilustra con las siguientes ecuaciones:

$$f_t = \sigma\left(W_f \cdot [h_{t-1}, x_t] + b_f\right) \quad (24)$$

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \quad (25)$$

$$\tilde{C}_t = tanh(W_c \cdot [h_{t-1}, x_t] + b_c) \quad (26)$$

$$C_t = f_t \cdot C_{t-1} + i_t \cdot \tilde{C}_t \quad (27)$$

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \quad (28)$$

$$h_t = o_t \cdot \tanh(C_t) \quad (29)$$

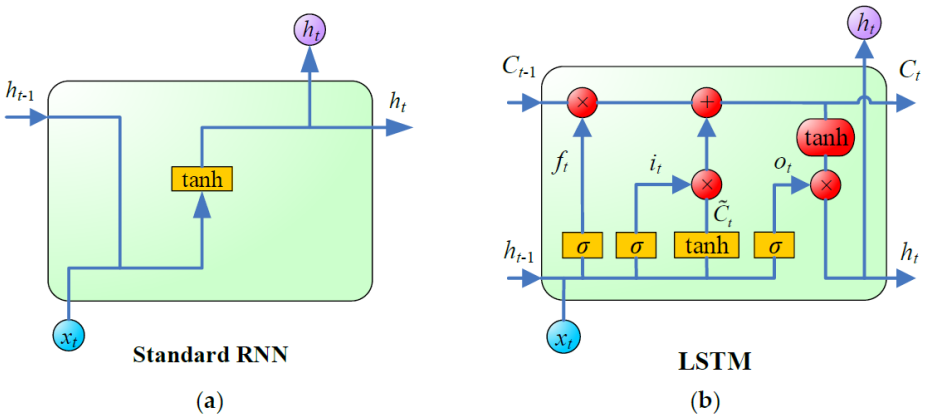


Imagen 24. Estructuras de las redes RNN y LSTM (16)

Donde $x_t$ es la entrada de la red, $h_t$ es la salida de la capa oculta, $\sigma$ denota la función sigmoide, $C_t$ es el estado de la célula, y $\tilde{C}_t$ denota el valor candidato del estado. Además, hay tres puertas en la capa LSTM: $W_f$, $W_i$, $W_o$ y $W_c$ son los pesos de la puerta de olvido, la puerta de entrada, la puerta de salida y la céclula, respectivamente. $b_f$, $b_i$, $b_o$ y $b_c$ son los sesgos de la puerta de olvido, la puerta de entrada, la puerta de salida y la célula, respectivamente.

La puerta de entrada decide si la información de entrada será reservada o no, y la puerta de olvido determinará si la información será descartada o no. El estado de procesamiento se registrará en la celda, y los valores de salida de la LSTM serán entregados por la puerta de salida. Gracias a este diseño inteligente mencionado anteriormente, la LSTM puede aprender las dependencias a largo plazo de los datos secuenciales de tiempo. $i_t$ es la

puerta de entrada, $o_t$ es la puerta de salida, y $f_t$ es la puerta de olvido. El LSTM está diseñado para resolver el problema de la dependencia a largo plazo a largo plazo. En general, la LSTM proporciona buenos resultados de previsión.

Dicho esto, se realizó el mismo backtesting en 2021 para este modelo. En este tipo de modelos es un poco más complicado porque el algoritmo se entrena en los años 2019 y 2020, mientras que se predice 2021. Se ha visto como el reentrenamiento es muy importante ya que el retardo de 24 horas es una variable muy importante, por lo que una vez entrenado, se realiza la predicción y se reentrena con cada día con los 24 valores reales para disponer de la máxima información en todo momento.

Los resultados obtenidos se recogen en la siguiente tabla.

| LSTM, DROPOUT | Training | Retrain / steps_epochs | Lookback | MAE | MAE (median) | WMAPE | RMSE | % TREND |
|---|---|---|---|---|---|---|---|---|
| (128, 64, 32, 16; 0.15) | (50, 50) | (2, 24) | 28 | 14.78 | 7.95 | 13.18 | 25.32 | 81.32 |
| (128, 64, 32, 16; 0.15) | (75, 75) | (2, 24) | 28 | 15.07 | 8.39 | 13.44 | 25.47 | 80.14 |
| (128, 64, 32, 16; 0.15)[8] | (50, 50) | (2, 24) | 28 | 14.74 | 8.00 | 13.15 | 25.04 | 80.98 |
| (128, 64, 32, 16; 0.15) | (50, 50) | (2, 24) | 14 | 16.00 | 8.60 | 14.27 | 27.22 | 77.54 |
| (128, 64, 32, 16; 0.15) | (50, 50) | (2,24) | 49 | 14.66 | 8.38 | 13.08 | 24.74 | 81.71 |
| (128, 64, 32, 16; 0.15) | (50, 50) | (4,24) | 49 | 14.43 | 7.78 | 12.87 | 25.34 | 81.04 |
| (64,32, 16; 0.05)[9] | (100,100) | (4,24) | 49 | 14.06 | 7.97 | 12.54 | 23.90 | 81.16 |

Tabla 9. Métricas obtenidas del backtesting de la configuración LSTM.

Los resultados arrojados con este modelo y las distintas configuraciones del mismo son bastante pobres comparadas con los modelos anteriores. Esto era de esperar viendo la relación entre las variables y estudios realizados por otros expertos, como en (18) y (19), que exponen que las series temporales no pueden predecirse en estos modelos porque los datos utilizados para la estimación suelen limitarse a una ventana temporal reciente. Los datos financieros pueden ser 'no estacionarios' y propensos a cambios de régimen, lo que puede hacer que los datos más antiguos sean menos relevantes para la predicción. Esto puede estar ocurriendo en este caso, ya que, al analizar

[8] Empezando a entrenar en marzo de 2020, evitando el periodo de 2019 y primer trimestre de 2020
[9] Modelo usando las variables seleccionadas para el Random Forest.

las relaciones entre las variables y el cambio de tendencia con el paso de los años, entrenar el modelo desde 2019 puede estar empeorando las predicciones de 2021 pues se está entrenando el modelo con una relación entre las variables que no se cumplen en 2021. Sin embargo, entrenando solo a partir de marzo de 2020 (tercera línea de la tabla anterior) tampoco resultan buenos valores, lo que refleja una limitación de este tipo de algoritmos en la predicción del precio de la electricidad.

## 5.9. Modelo *ensemble* de los modelos propuestos

El modelo se puede encontrar en el notebook *Modelo_Ensemble.ipynb* incluido en el repositorio del trabajo.

Este modelo se basó en usar las predicciones de los modelos anteriores como variables predictoras y coger la media de todas las combinaciones posibles de los modelos y seleccionar el que menor WMAPE resulte.

Realizando esto, la mejor combinación es la obtenida mediante la media de: los modelos estimados por el Random Forest, XGBoost, Regresión Lineal y Regresión Lasso.

| Métricas | MAE | MAE (median) | WMAPE | RMSE | % TREND |
|---|---|---|---|---|---|
| Valores | 9.19 | 5.89 | 8.25 | 14.94 | 85.04 |

Tabla 10. Métricas obtenidas del mejor modelo ensemble

Este modelo, que tiene en cuenta todas las predicciones de los modelos anteriores, reduce en gran medida el error obtenido por cada uno de ellos por separado. Este es el mayor poder de este tipo de modelos, y el WMAPE obtenido para la mejor combinación de ellos en todo el periodo de backtesting de 2021 es del 8.25 %, mejorando el mejor modelo individual de todos los anteriores que consistía en el XGBoost.

## 5.10. Modelo Stacked model (modelos apilados)

El modelo se puede encontrar en el notebook *Modelo_Stacked.ipynb* incluido en el repositorio del trabajo.

En los *model stacked* (o apilamiento de modelos) no se utiliza un único modelo para hacer las predicciones, sino que se hacen predicciones con varios modelos diferentes y luego se usan esas predicciones como

características para un metamodelo de nivel superior. Puede funcionar especialmente bien con varios tipos de modelos de nivel inferior, todos ellos contribuyendo con diferentes puntos fuertes al metamodelo. Los modelos apilados pueden construirse de muchas maneras, y no hay una forma "correcta" de utilizar el apilamiento. De forma gráfica, la estructura que sigue es la siguiente:

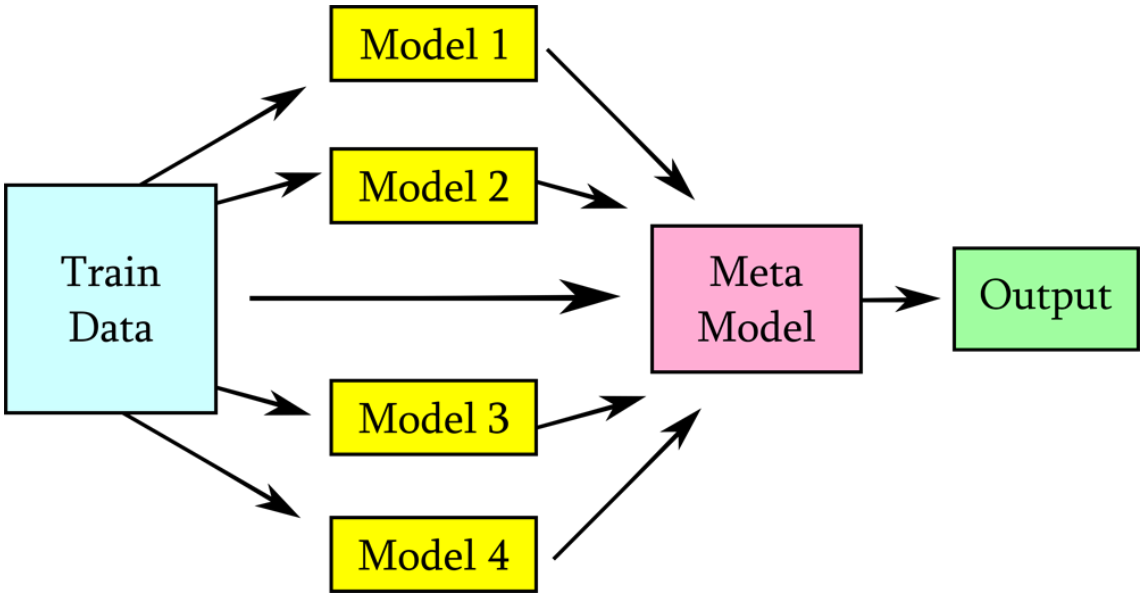


Imagen 25. Estructura de un *model stacking* (17)

Se han probado distintos Meta Model, según la imagen 22, entre ellos varios de los probados durante el trabajo, como son la regresión lineal múltiple, Ridge, Random Forest y XGBoost. Se realizaron varias pruebas, aunque en la tabla solo se recoge el modelo obtenido con la mejor ventana móvil y otro con todos los datos disponibles.

Para realizar este tipo de modelo y seguir testeando el backtesting sobre 2021 y, por tanto, los resultados sigan siendo comparables, se realizaron predicciones de los mejores modelos de las secciones anteriores desde junio de 2020 hasta fin de 2021, con el fin de usar los 6 meses de 2020 como entrenamiento y empezar a testear el modelo desde el 1 de enero de 2021 hasta el 31 de diciembre de 2021.

Cada modelo usa de variables regresoras las de los modelos estimados por la Regresión Lineal, Lasso, Ridge, XGBoost y Random Forest.

| Modelo | Rolling Window | MAE | MAE (median) | WMAPE | RMSE | % TREND |
|---|---|---|---|---|---|---|
| Regresión Lineal | No | 9.04 | 5.63 | 8.07 | 14.80 | 85.00 |
| Regresión Lineal | 90 | 9.04 | 5.66 | 8.07 | 14.82 | 84.86 |
| Ridge | No | 9.04 | 5.64 | 8.08 | 14.81 | 85.03 |
| Ridge | **120** | **8.98** | **5.60** | **8.02** | **14.77** | **84.95** |
| XGBoost | 30 | 10.84 | 6.41 | 9.69 | 18.43 | 69.36 |
| XGBoost | No | 10.08 | 6.25 | 9.01 | 16.22 | 74.29 |
| RandomForest | 30 | 10.22 | 6.11 | 9.13 | 17.35 | 73.96 |
| RandomForest | No | 9.99 | 6.22 | 8.93 | 16.02 | 75.15 |

Tabla 11. Métricas obtenidas del modelo stacked model

Los resultados arrojados con este modelo son mejores comparados con todos los anteriores. Esto tiene sentido, pues las variables que se usan para predecir ya tienen unas métricas bastante bajas por sí solos, así que los modelos, como mínimo, deberían tener una métrica muy parecida que la del mejor de los modelos individuales.

Con esto, para las variables anteriormente comentadas, de los modelos que recogen la tabla 11, el mejor Meta Model es el que consiste en la regresión Ridge y ventana móvil de 120 días, con un WMAPE del 8.02 %, que es el mejor modelo de todos los probados.

## 6. Discusión

Tras probar y analizar todos los modelos de la sección 5, recogemos en la siguiente tabla las métricas de cada uno de los mejores modelos de cada sección.

| Modelo | MAE | MAE (median) | WMAPE | RMSE | % TREND |
|---|---|---|---|---|---|
| Persistencia | 16.07 | 8.92 | 14.36 | 27.38 | 74.84 |
| Regresión Lineal | 10.21 | 6.78 | 9.12 | 16.59 | 83.99 |
| Ridge | 10.21 | 6.76 | 9.12 | 16.59 | 83.16 |
| Lasso | 10.12 | 6.86 | 9.04 | 16.25 | 84.62 |
| Sarima | 14.15 | 8.66 | 12.64 | 22.72 | 83.10 |
| Sarima (regresión de los residuos) | 10.24 | 6.62 | 9.15 | 16.85 | 83.83 |
| Random Forest | 10.02 | 6.19 | 8.95 | 16.34 | 83.59 |
| XGBoost | 9.83 | 6.18 | 8.78 | 15.26 | 82.83 |
| LSTM | 14.06 | 7.97 | 12.54 | 23.90 | 81.16 |
| Ensemble | 9.19 | 5.89 | 8.25 | 14.94 | 85.04 |
| Stacked | 8.98 | 5.60 | 8.02 | 14.77 | 84.95 |

Tabla 12. Resumen de todas las métricas de cada uno de los mejores modelos

Como ya se ha comentado, el mejor modelo resultante es el stacked model usando como variables explicativas las predicciones de los modelos de regresión lineal, ridge, lasso, random forest y xgboost.

Respecto a la diferencia entre los modelos, no hay demasiada entre los modelos lineales y los no lineales. Aun así, se refleja ligera mejora en los modelos no lineales respecto a los lineales, como vemos en las imágenes 22 y 23, en la que observamos que modelos no lineales como el XGBoost se adaptan mejor a las tendencias que modelos como la regresión lineal.

Hay que destacar que el periodo de backtesting es todo el año 2021 que, como hemos visto, es el año en el que los precios de electricidad han sido los más volátiles de la historia y hay determinados cambios de tendencia que ningún modelo habría podido predecir, por lo que las métricas finalmente se igualan y no hay demasiada diferencia entre los modelos, ya que, todos usan las mismas variables.

Lo que sí parece claro es que, al considerar modelos ensemble o stacked, la predicción mejora, pues se compensarán los errores de los modelos considerados resultando en un modelo final mucho más robusto. Por tanto, aunque las métricas de cada modelo por separado sean similares, parece indicar que cada modelo, ligeramente, se ha centrado en algunos periodos.

Respecto al baseline (modelo de persistencia) encontramos que los modelos estudiados representan una mejora considerable, por lo que parece interesante la implementación de estos modelos.

A continuación, estudiamos para el mejor modelo encontrado (el stacked) las métricas obtenidas por día de la semana, hora del día y mes del año durante todo 2021.

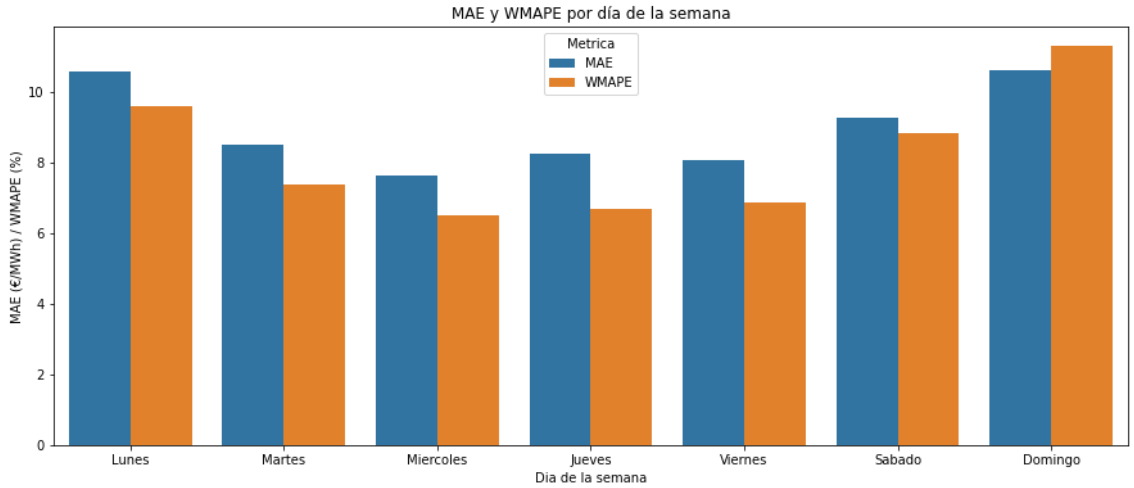


Imagen 26. Métricas MAE y WMPAE para el modelo stacked por cada día de la semana.

De aquí deducimos que, en promedio, para el backtesting de 2021, los días de la semana en los que más se desvían las previsiones son los domingos y los lunes. Esto puede ser debido a que, al pasar de un día festivo a un día laboral, sea el cambio más brusco de tendencia y les cueste más a los modelos capturar este cambio.

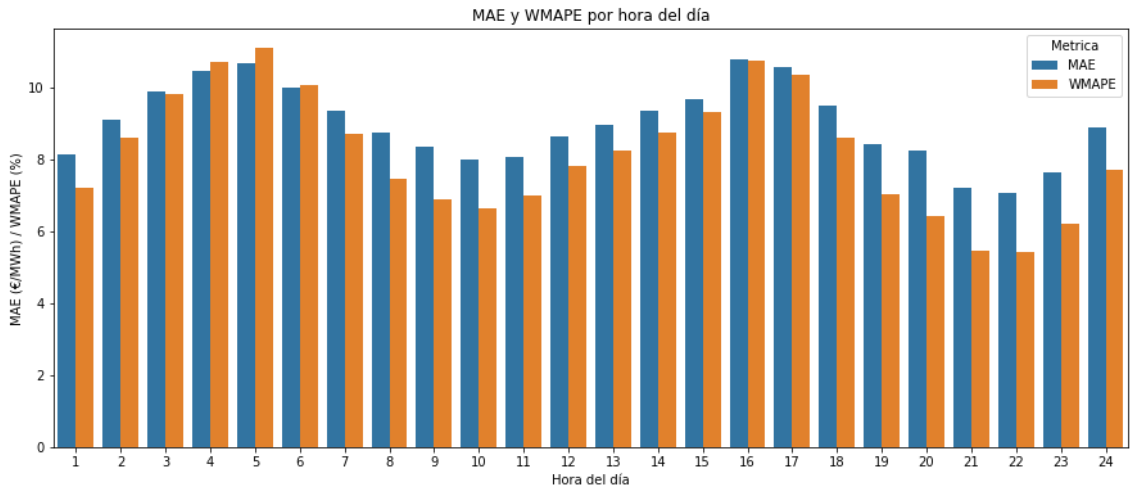


Imagen 27. Métricas MAE y WMPAE para el modelo stacked por cada hora del día

Respecto a las horas en las que el modelo más se desvía del valor real, encontramos dos zonas, una de madrugada entre las 3:00 y las 6:00, y otra por la tarde, entre las 15:00 y las 17:00. Por contra, las horas en las que menos error se encuentra son por la mañana entre las 8:00 y las 11:00 y por la noche entre las 20:00 y las 23:00.

Esto da una idea de que hay zonas horarias en las que el modelo se ajusta mejor que otras, esto podría dar una ligera indicación de que podría ser otra opción realizar modelos en función de franjas horarias, ya que vemos que el modelo actual se ajusta mejor a unas franjas que a otras.

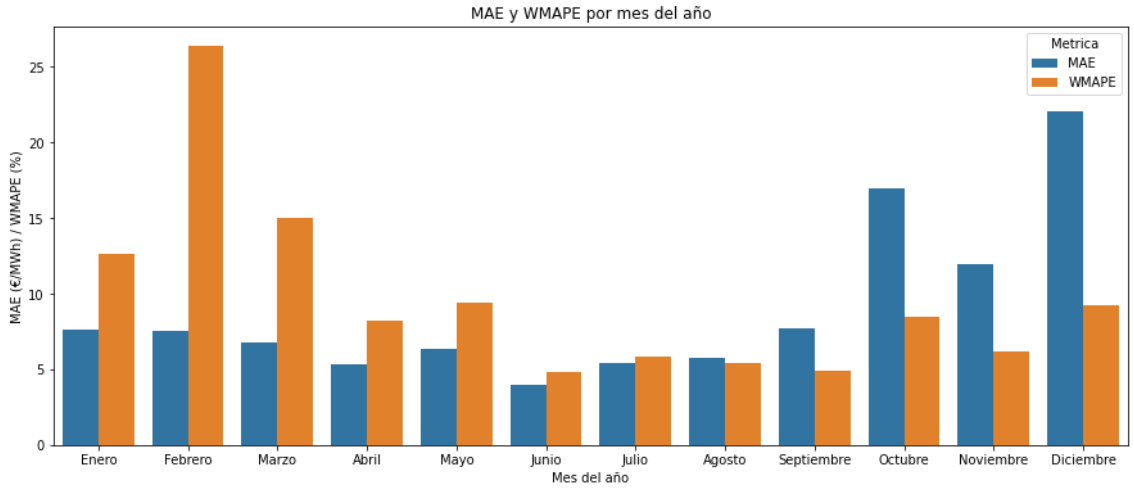


Imagen 28. Métricas MAE y WMPAE para el modelo stacked por cada mes del año

Por último, analizando las métricas por cada uno de los meses de 2021, resulta un valor de MAE bastante más alto en los últimos meses que respecto los primeros, lo cual es debido a que, observando la serie de la imagen 5, los últimos meses del año son los que tienen los precios más altos, y dado que esta métrica es el error absoluto medio, esta es mayor en los meses de precios más altos que en los de los precios más bajos.

Para evitar este efecto se suelen usar métricas que midan % de variación, que relativizan esta diferencia, como podría ser el MAPE, pero como hay precios reales cercanos a 0, pierde la información esta variable, por lo que se usa esta métrica. Según esta métrica, los meses de peor predicción son el primer trimestre de 2021. Lo que se obtiene es en los meses más calurosos (junio, julio, agosto y septiembre) unas métricas bastante estables y bajas.

# 7. Conclusiones

En el presente trabajo se han aplicado distintos modelos con algoritmos de diferente naturaleza para resolver el mismo problema y encontrar qué aproximación es la mejor para resolver este problema de predicción de series temporales.

Lo primero que se realizó fue la construcción del dataset de predicción, con la variable objetivo y las variables explicativas. Esta es la parte más importante, pues hemos visto que modelos autorregresivos (SARIMA) se quedan lejos del rendimiento de cualquier otro modelo que usa variables explicativas. En problemas de este tipo es la parte que pasa más desapercibida, pero es indispensable.

Tras ello, se ha visto que la relación entre las variables ha ido cambiando con el paso del tiempo, lo que ha complicado más el ejercicio de predicción pero, que a su vez, lo ha hecho más interesante, exponiendo el uso de las ventanas de entrenamiento como vital para encontrar un gran resultado, adaptándose a los cambios de tendencias del modelo. Esto le ha dado un valor añadido pues no solo hay que testear los modelos en periodos de estabilidad, sino que para analizar su robustez hay que probarlos en todo tipo de contextos.

Finalmente, de todos los modelos considerados, el mejor es un stacked entre varios modelos complejos. También es de destacar los buenos resultados del ensemble.

# 8. Referencias


(1) La energía es un bien básico de primera necesidad cuyo acceso debe ser garantizado como servicio público (2022). Consultado el 12 de abril de 2022 en: https://www.energias-renovables.com/panorama/la-energia-es-un-bien-basico-de-20191119

(2) España – Consumo de electricidad (2022). Consultado el 12 de abril de 2022. https://datosmacro.expansion.com/energia-y-medio-ambiente/electricidad-consumo/espana

(3) ¿Cómo funciona el sector eléctrico en España? (2021). Consultado el 12 de abril de 2022. https://www.aura-energia.com/como-funciona-el-sector-electrico-en-espana/#:~:text=Una%20comercializadora%20de%20electricidad%20compra,trav%C3%A9s%20de%20un%20precio%20pactado

(4) Ley 24/2013, de 26 de diciembre de 2013, del Sector Eléctrico. *Boletín Oficial del Estado*. Madrid, 27 de diciembre de 2013, núm 310, 46-48: https://www.boe.es/buscar/act.php?id=BOE-A-2013-13645&p=20211027&tn=1#a30

(5) El sistema marginalista de fijación de precios eléctricos, a debate. Alejandro Nieto Gonzalez (2021). Consultado el 3 de mayo de 2022 en: https://www.elblogsalmon.com/sectores/sistema-marginalista-fijacion-precios-electricos-a-debate

(6) Generación en España. (2022). Consultado el 3 de mayo de 2022 en: https://www.ree.es/es/datos/generacion/potencia-instalada

(7) Generación en España (2022). Consultado el 4 de mayo de 2022 en: https://www.ree.es/es/datos/generacion/estructura-generacion

(8) Nuestros hábitos de consumo (2021). Consultado el 4 de mayo de 2022 en: https://www.ree.es/es/red21/eficiencia-energetica-y-consumo-inteligente/nuestros-habitos-de-consumo

(9) Hastie, T., Tibshirani, R. y Friedman, J. (2001) *The Elements of Statistical Learning. Data Mining, Inference, and Prediction* (2ºed). Nueva York. Springer

(10) Hyndman, R.J., y Athanasopoulos, G. (2021) *Forecasting: principles and practice* (3º ed), Melbourne, Australia. Otexts, OTexts.com/fpp3. Acceso 15 mayo 2022

(11) James, G., Witten, D., Hastie, T. y Tibshirani, R. (2021) *An introduction to Statistical Learning with applications in R* (2º ed) Nueva York. Springer.

(12) Grus, J. (2019) *Data Science from Scratch. First Principles with Python* (2º ed) Sebastopol. O' Relly Media.

(13) Algoritmo XGBoost (2020). Consultado el 15 de mayo de 2022 en: https://es.education-wiki.com/4794400-xgboost-algorithm

(14) Amat Rodrigo, J. (2020). Gradient Boosting con Python. Consultado el 13 de mayo de 2022 en: https://www.cienciadedatos.net/documentos/py09_gradient_boosting_python.html

(15) Chollet, F. (2018) *Deep Learning with Python*. Nueva York. Manning Publications Co.

(16) Ping-Huan, K., Chiou-Jye,H. (2018) An electricity Price Forecasting Model by Hybrid Structured Deep Neural Networks. *Sustainability*, 10.

(17) Wadkings, J. (2021) Simple Model Stacking, Explained and automated. Consultado el 16 de mayo de 2022 en:

https://towardsdatascience.com/simple-model-stacking-explained-and-automated-1b54e4357916#:~:text=In%20model%20stacking%2C%20we%20don,a%20higher%2Dlevel%20meta%20model.
(18) Lemus, G. (2018) Why Financial time series LSTM Predictions Fails. Consultado el 29 de mayo de 2022 en: https://medium.datadriveninvestor.com/why-financial-time-series-lstm-prediction-fails-4d1486d336e0
(19)Culurciello, E. (2018) The fall of RNN / LSTM. Consultado el 29 de mayo de 2022 en: https://towardsdatascience.com/the-fall-of-rnn-lstm-2d1594c74ce0

# 9. ANEXOS

## 9.1. ANEXO A.

Versiones de los paquetes más importantes utilizados en el trabajo.

| Paquete | Versión |
|---|---|
| Pandas | 1.0.5 |
| Numpy | 1.19.5 |
| Matplotlib | 3.2.2 |
| Statsmodels | 0.12.2 |
| Plotly | 5.4.0 |
| Scikit-learn | 0.23.1 |
| Tensorflow | 2.4.0 |
| Keras | 2.4.3 |
| Yfinance | 0.1.70 |
| Xgboost | 1.3.3 |
| Scipy | 1.5.0 |
| Seaborn | 0.10.1 |
| Pmdarima | 1.8.2 |

Tabla 13. Módulos y versiones usados en el presente trabajo.

## 9.2. ANEXO B.

En este anexo se muestran los resultados del estudio de la relación entre las variables que servirán para explicar el precio de la electricidad en España mediante un gráfico pairplot del módulo seaborn desglosado por los 3 años con los que contamos datos: 2019, 2020 y 2021.

Las variables a estudiar se han dividido en 3 grupos:

- Grupo 1: variables relacionadas con REE:

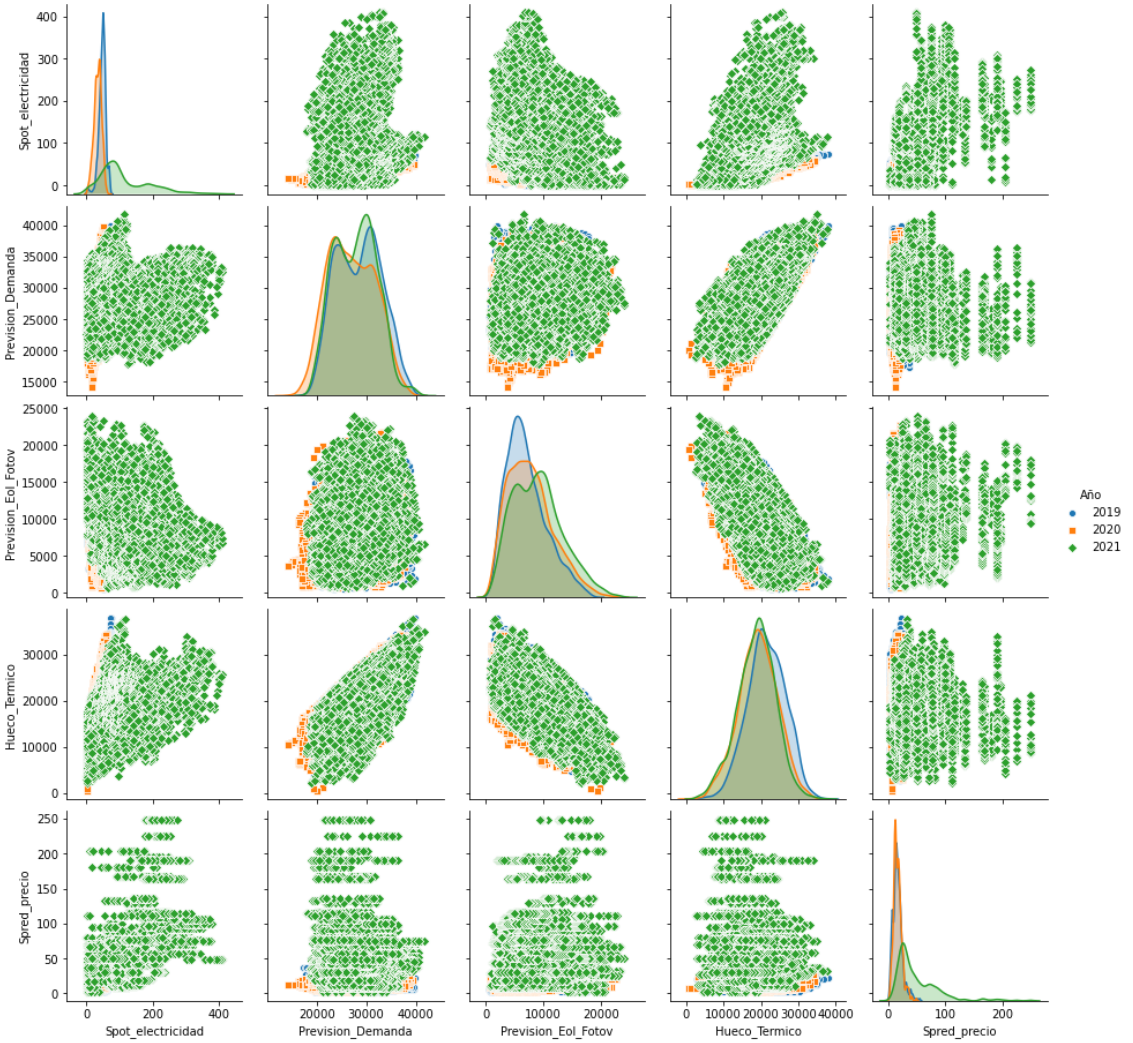


Imagen 29. Gráfico de violín del precio de la electricidad desde 2014 hasta 2021. Elaboración propia.

Este grupo incluye las variables de previsión de demanda, previsión de generación de renovables, hueco térmico y el spread del precio del día anterior (diferencia entre el precio máximo y el mínimo del día anterior). La primera columna o la primera fila son en las que hay que fijarse si se quiere observar la relación entre las variables explicativas y la variable a predecir, que es la llamada **Spot_electricidad**. A grandes rasgos, parece existir una correlación positiva entre la variable objetivo y las variables hueco térmico y previsión de la demanda, es decir, estas variables evolucionan en el mismo sentido que la variable a predecir, mientras que la previsión de renovables parece seguir una correlación negativa, lo cual tiene sentido, ya que, a mayor generación de renovables, el precio de la electricidad debería ser menor. No parece que exista diferencia al observar estas variables en los 3 distintos años que se muestran.

- Grupo 2: variables meteorológicas.


Imagen 30. Gráfico de violín del precio de la electricidad desde 2014 hasta 2021. Elaboración propia.

Este grupo muestra la relación de la variable objetivo Spot_electricidad con las variables meteorológicas seleccionadas, que son, con sus unidades, las siguientes:

- Temperatura: °C
- Spread temperatura: °C
- Precipitación: $l/m^2$
- Radiación: $W/m^2$
- Velocidad viento: m/s

No parece existir ninguna relación clara entre ellas, lo cual tiene sentido, estas variables puede que sirvan para afinar alguna predicción o justificar algunos atípicos, por ejemplo, alguna variable extrema puntual como una temperatura muy baja o muy alta puede justificar un precio algo más alto que los días cercanos a esa fecha.

- Grupo 3: variables financieras

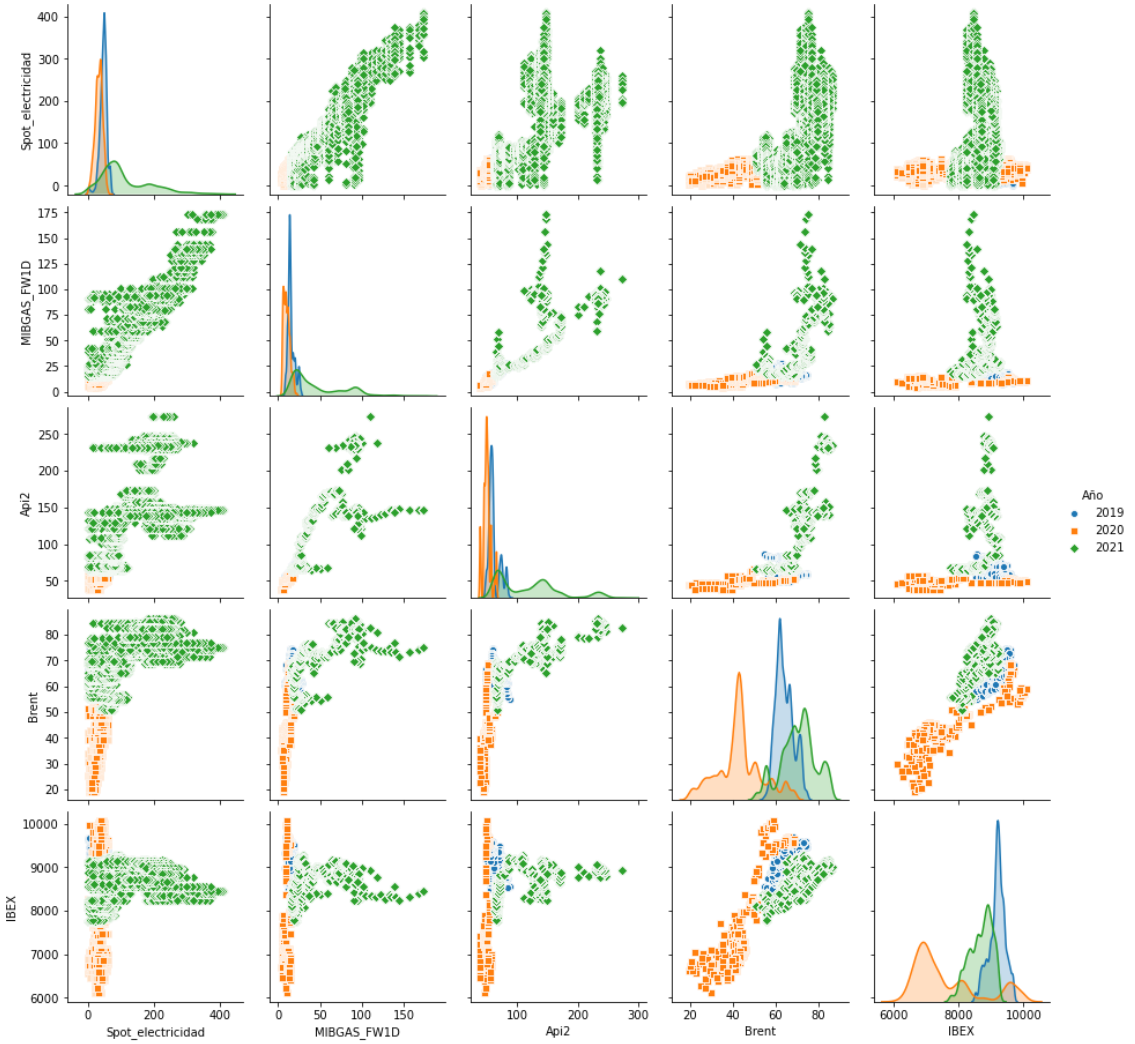


Imagen 31. Gráfico de violín del precio de la electricidad desde 2014 hasta 2021. Elaboración propia.

Este grupo incluye las variables del precio del gas, carbón, brent e IBEX. Estas variables parecen dar más información, por ejemplo, la del gas en 2021 parece seguir una correlación positiva bastante fuerte con la variable objetivo, al igual que la del API2. El BRENT e IBEX también parecen seguir una correlación positiva con la variable objetivo en 2019 y 2020, pero, se observa que esta tendencia ha cambiado completamente en 2021, ha habido un cambio en la correlación entre estas variables, esto es algo que habrá que tener en cuenta, pues se van a usar los datos de 2019 y 2020 para entrenamiento del modelo y el de 2021 para backtesting. Los gráficos mostrados en el presente anexo sirven para ver si las funciones de distribución de las variables son constantes en el periodo considerado, y ya estamos viendo que muchas de ellas no.