UNIVERSIDAD POLITÉCNICA DE MADRID

ESCUELA TÉCNICA SUPERIOR DE INGENIEROS INDUSTRIALES

GRADO EN INGENIERÍA EN TECNOLOGÍAS INSUTRIALES

Aplicación de algoritmos Boosting a la predicción de precios de energía eléctrica

Autor:

Juan Carlos Durante Beleña

Tutor:

José Manuel Mira McWilliams

TRABAJO FIN DE GRADO

JUNIO 2021

Agradecimientos

Quisiera agradecer a mi familia, en especial a mi madre y mis abuelos, la educación que me han otorgado ya que ello ha formado parte de que haya luchado y me haya esforzado para conseguir ser ingeniero industrial.

Del mismo modo dar las gracias, José Manuel Mira McWilliams que como tutor me ha ayudado y guiado en el proyecto, sin su tiempo y dedicación este trabajo habría resultado extremadamente más complicado

Por último, dar las gracias a mi novia que ha sido un apoyo incondicional a lo largo de los años de carrera y en los momentos más complicados tanto académicamente como fuera de ello.

Índice

1. Resumen ejecutivo

Este Trabajo de Fin de Grado tiene como título: "Aplicación de algoritmos Boosting a la predicción de precios de energía eléctrica" y en a lo largo del proyecto he contado con la ayuda de mi tutor, el profesor José Manuel Mira Mcwilliams.

El trabajo se adentra en el mundo del Machine Learning, también conocido como "aprendizaje automático". Machine Learning es una disciplina del campo de la inteligencia artificial que, a través de algoritmos, dota a los ordenadores de la capacidad de identificar patrones en datos masivos para hacer predicciones.

En este proyecto, se trata de combinar dos campos distintos. El primero de ellos es el mercado eléctrico de España donde encontramos el precio de la energía eléctrica entre muchas otras variables y el segundo de ellos es el ya comentado mundo del Machine Learning y esta combinación se realiza mediante la predicción del precio de la energía eléctrica.

Dentro del mundo de Machine Learning tenemos numerosos y variados algoritmos que se desarrollan de distintas maneras, los más famosos son Random Forest, redes neuronales, análisis cluster o boosting. Concretamente el que se aplicará en este proyecto es el boosting.

Al ser un ámbito tan extenso tenemos algoritmos muy parecidos como por ejemplo son el algoritmo bagging y el algoritmo boosting pero no son iguales ya que en el boosting se actúa de forma secuencial a diferencia del bagging.

Boosting está clasificado dentro del grupo de algoritmos de aprendizaje supervisado. El aprendizaje supervisado es aquel que tiene como objetivo crear una función capaz de predecir el valor correspondiente a cualquier objeto de entrada válida después de haber observado una serie de ejemplos que son llamados datos de entrenamiento.

El algoritmo boosting intenta conseguir un clasificador fuerte a partir de muchos clasificadores débiles comúnmente son conocidos como weak learners o base learners. Cada uno de esos clasificadores débiles son modelos de distintos tipos: modelos de regresión lineal simple, modelos de regresión lineal múltiple o incluso, modelos binarios donde solo se trabaje con dos tipos de datos diferentes.

Para comprender bien el concepto de Boosting debemos aplicar secuencialmente diferentes base learners con versiones modificadas de los datos para producir una secuencia de base learners que consigan finalmente dar con un clasificador fuerte. Para conseguir este procedimiento el algoritmo sigue los siguientes pasos generales:

1. Al principio se les otorga el mismo peso absolutamente a todos los datos del conjunto de entrenamiento. Este peso se puede formular como $w_i = 1/n$; donde n es el número de datos del conjunto e i se extiende desde 1 hasta n, siendo n un número finito.
2. Se produce el entrenamiento del modelo con esos datos de entrenamiento y con los pesos iniciales obteniéndose así el primer base learner.
3. Se calcula el error del modelo en el conjunto de entrenamiento con los pesos anteriores

4. Se produce un incremento de peso en aquellos datos que el modela ha clasificado de forma errónea. Esto se hace para conseguir que en la siguiente iteración obtener un base learner que tenga en mayor consideración estos datos y se acerque más a ellos
5. Se vuelve a entrenar un nuevo modelo con un nuevo base learner con el conjunto de entrenamiento con los pesos modificados
6. Se repite la secuencia continuamente desde el paso número 3 hasta el punto de iteraciones fijadas
7. Cada uno de los modelos obtenidos que conlleva un base learner s le habrá asignado un peso. De tal forma que el modelo final se calcula con una votación ponderada por los pesos de todos los modelos.

Antes de llegar a lo que es en sí la metodología del trabajo y el cálculo de nuestro modelo. Explicaremos una breve y concisa introducción sobe el mercado eléctrico y su funcionamiento.

El precio de la energía eléctrica fluctúa, es un hecho, pero no siempre ha sido así, esta fluctuación se da desde el momento que liberalizaron el mercado eléctrico español en el año 1997, ya que antes los precios estaban predeterminados por el estado.

El OMIE (Operador del Mercado Ibérico de la Energía) es el encargado de gestionar este mercado de la energía tanto en nuestro país como en el país vecino Portugal.

La variación de los precios de la energía eléctrica puede ser causado por múltiples factores, en este trabajo se trata de averiguar que variables o factores son importantes y pueden afectar a esa variación de precio para posteriormente introducir estas variables en una base de datos con la que trabajar y conseguir un modelo boosting predictivo que arroje cierta fiabilidad.

Se observa que por lo general los precios de la energía eléctrica no tienen grandes variaciones de una hora para otra, mediante el visionado de distintas gráficas en oficial del OMIE se aprecia que la variable precio en la página el instante anterior al que se realiza en la predicción puede ser un variable que afecte notablemente a la predicción. No obstante, a lo largo de semanas o meses sí que vemos grandes o notables fluctuaciones en el precio.

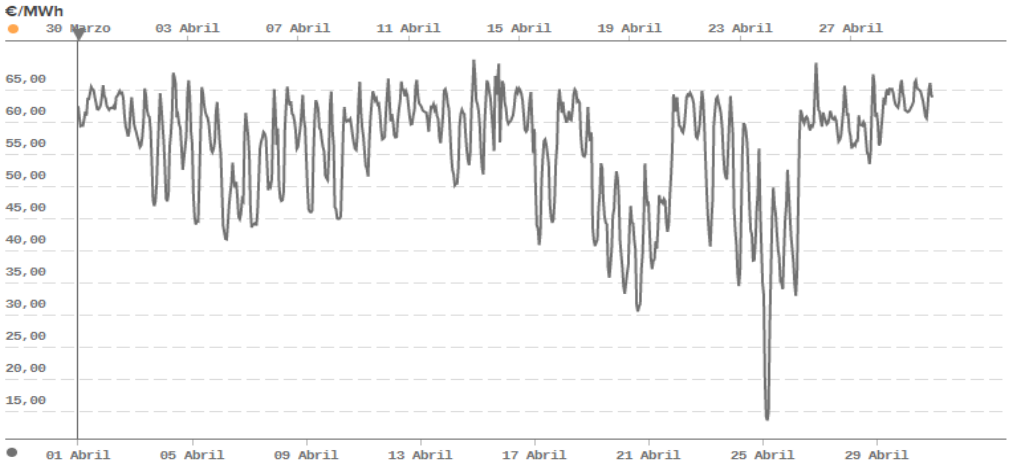

Figura 1: Gráfica del precio de la energía eléctrica durante Abril de 2019

También es importante la cantidad de energía renovable que se genera ya que no se produce al mismo precio la energía renovable que la no renovable, y del mismo modo los habitantes del país no consumen la misma cantidad de energía cuando frío que cuando hace calor, cuando están de vacaciones o cuando no lo están. Por esto, para la realización de nuestro proyecto, se deciden escoger las variables demanda, energía generada sin emisiones de CO_2 y las temperaturas máxima mínima y media y ver cómo pueden afectar o no a nuestro modelo predictivo.

Para empezar la realización del modelo tenemos que obtener una base de datos, con una cantidad de volumen de datos decente, por ello se propone la realización de una base de datos en el programa Microsoft Excel donde tendremos 8670 filas y 7 columnas. Cada columna representará una variable incluida nuestra variable que queremos predecir y cada fila representa 1hora del año puesto que un año no bisiesto como es el año elegido para este proyecto, el 2019, tiene ese número determinado de horas.

Se ha utilizado el programa Rstudio en el cual se trabaja con el software estadístico R para la implementación de un modelo predictivo con el algoritmo boosting.

Para el manejo del programa puesto que se ha realizado una exhaustiva búsqueda de información acerca de su programación, sus funciones, interpretación de código etc. en la que la gran mayoría has ido en inglés.

Finalmente, se ha conseguido el objetivo de obtener un modelo predictivo con cierta fiabilidad gracias a la función gamboost que accedemos a ella mediante el paquete mboost en Rstudio, también se ha estudiado la función blackboost y mediante el cálculo d errores y visualización por pantalla de diversas gráficas que se pueden generar en el propio programa RStudio se han podido comparar los distintos resultados que nos arrojaban.

Uno de los papeles más importantes a la hora de la fiabilidad del modelo es el MAPE que es conocido como el error medio en valor absoluto y en porcentaje que junto a las gráficas de los resultados de la predicción que se comparan con los valores reales, han sido herramientas muy útiles para dictaminar sí tiene o no coherencia el modelo.

Por otro lado, en el proyecto se han realizado experimentos con distintos horizontes de tiempo, hay predicciones para t+1, t+3 y finalmente t+8, es decir, predicciones para una hora en adelante, tres horas en adelante y 8 horas en adelante.

En líneas generales se ha observado que aquel modelo que se ha realizado con gamboost nos aporta mayor coherencia que aquel realizado con blackboost. Esto se debe a que gamboost realiza la boosting de tal forma que optimiza el peso de los errores gracias a muestras aleatorias que utilizan los diferentes datos como base learners a través del análisis de árboles aditivos.

En la parte de análisis de resultados se han plasmado numerosas gráficas que permiten visualizar y comparar unas predicciones con otras, así como se ha calculado el valor de correlación entre predicción y realidad ya que puede resultar interesante.

A continuación, se muestra una comparación donde tenemos el modelo predictivo a nuestra izquierda, concretamente el que mejor fiabilidad ha tenido de todo el proyecto, predicción gamboost para t+1, y los valores reales del precio a la aparecen en la gráfica de la derecha.

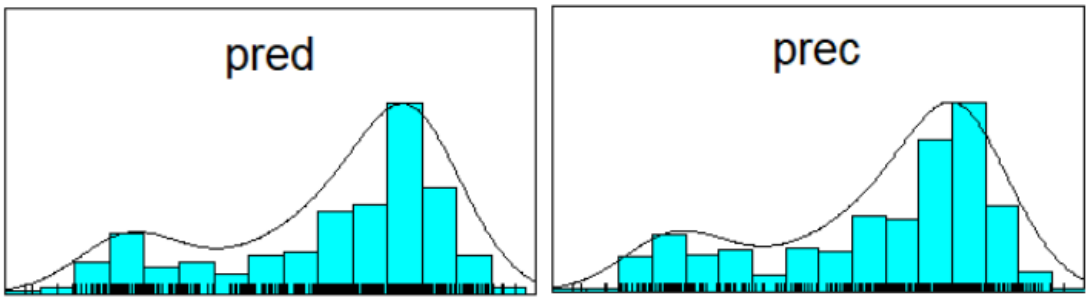

Figura 2: Gráficas comparativas de predicción gamboost en+1

Tras la realización del análisis de los resultados en todas las predicciones se han revelado cierta serie de conclusiones donde destacan las siguientes: como era de esperar el error de la predicción aumenta cuanto más aumente el horizonte de nuestra predicción en t+1 nos situamos en valores de MAPE del 5% y finalmente en t+8 nos vamos a valores superiores al 10%. El modelo nos revela la importancia de las variables seleccionada y cual de ellas es la que más afecta a la predicción del precio de la energía eléctrica, como se podía prever es el precio en el instante anterior y a continuación la variable más relevante es la demanda. Por otro lado, el algoritmo blackboost que optimiza los errores utilizando árboles de regresión como bases de aprendizaje a medida que avanza el horizonte de predicción denota una ligera mejora.

Para finalizar el trabajo se han incluido un capítulo de líneas futuras en el que se comentan propuestas interesantes que se podrían realizar en un futuro y también se ha calculado la planificación temporal y presupuesto que teóricamente, habría supuesto la ejecución del TFG

Por último, se han recopilado al final del trabajo en el Anexo el código R resumido que se ha desarrollado para la realización de este trabajo de fin de grado, un índice de figuras que existen a lo largo del trabajo y un índice de tablas.

2. Introducción

2.1 Evolución en el sector

Los primeros indicios de una aplicación práctica de electricidad en España se dieron en Barcelona en 1852 cuando un hombre fue capaz de iluminar su botica. En el mismo en Madrid se hicieron pruebas de iluminación en la plaza de la Armería y en el Congreso de los Diputados. Tendrán que pasar una cierta cantidad de años para que en 1875 1875 se instala una dinamo en Barcelona que logró iluminar las Ramblas, la Boquería, el Castillo de Montjuic y parte de los altos de Gracia. Esto propició que al año siguiente comience la electrificación industrial de nuestro país. Se forma la primera empresa española de electricidad en Barcelona bajo el nombre de la Sociedad Española de Electricidad. El impulso de la electricidad creció en el país a raíz de que en 1885 se publicara un primer decreto que ordenaba las instalaciones eléctricas y, posteriormente, tras 3 años de debates se publicó una Real Orden regula el alumbrado eléctrico de los teatros, prohibiendo expresamente el alumbrado con gas y dejando las lámparas de aceite sólo como sistema de emergencia. Tras estas leyes se notó un aumento del uso de electricidad en la península ya que en 1889 había 23 poblaciones con electricidad y finalmente en 1901 la cifra llegaba a las 571 poblaciones.

Con el comienzo del siglo, el auge de la electricidad fue un factor determinante en revolución industrial y del mismo modo también lo fue el hecho de que Nicola Tesla se impusiera en la "Guerra de Corrientes" en 1879 propiciando así el desarrollo de la corriente alterna la cual tiene un mayor porcentaje de culpa en conseguir la sociedad que tenemos hoy en día frente a la corriente continua. Como todos sabemos, la corriente alterna permitió la generación de electricidad a larga distancia y concretamente en España derivó en el desarrollo de grandes centrales hidroeléctricas. Este hecho permitió disponer de más recursos de cara a la industrialización y ayudo a realizar grandes avances en el mundo de la tecnología hasta que en 1970 la producción de energía alcanzó una notable cifra de 56.500 GWh, triplicando así a la que teníamos a comienzos de la década anterior.

En 1987 con un Real Decreto, se alcanza la regulación del sistema de ingresos de modo que se establecen una serie de normas que fijan los precios. Aun así, no será hasta 1997 cuando termine el monopolio existente en este sector con la Ley 54/97 aparecida en el BOE que dice textualmente: "La presente Ley se asienta en el convencimiento de que garantizar el suministro eléctrico, su calidad y su coste no requiere de más intervención estatal que la que la propia regulación específica supone. No se considera necesario que el Estado se reserve para sí el ejercicio de ninguna de las actividades que integran el suministro eléctrico. Así, se abandona la noción de servicio público, tradicional en nuestro ordenamiento pese a su progresiva pérdida de trascendencia en la práctica, sustituyéndola por la expresa garantía del suministro a todos los consumidores demandantes del servicio dentro del territorio nacional"

La liberalización del sector acarreó consecuencias como la clarividente eliminación del monopolio y, por tanto, la aparición de la competitividad en el sector que trajo consigo

los intentos de implantar el mejor servicio posible con unos precios lo más competitivos posibles.

Figura 3: Imágenes de Gran Vía a principios del siglo XX y a principios del siglo XXI

2.2 Actualidad de la electricidad en España

Para este proyecto se estudiará el precio de la energía eléctrica en España, es por ello que debemos conocer el sistema eléctrico español, y esto podemos hacerlo gracias a los datos facilitados en la página web de la red eléctrica de España.

En primer lugar, vamos a analizar la oferta y demanda que ha tenido el país en estos últimos años. Puesto que nuestro trabajo estudiará los precios de la energía eléctrica, los parámetros de oferta y demanda cobran una gran importancia a lo hora de entender las fluctuaciones de precio.

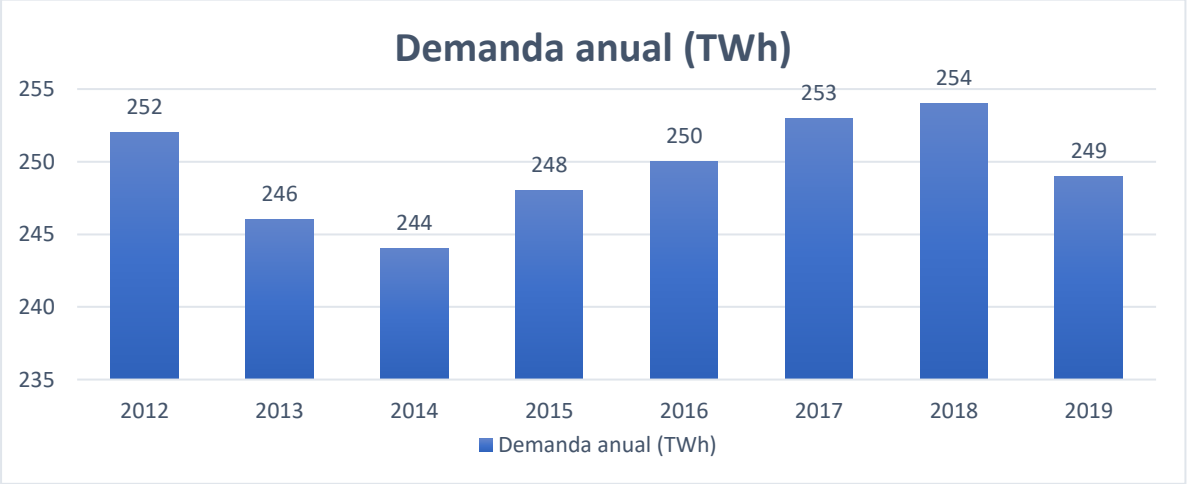

Figura 4: Evolución de la demanda de energía eléctrica en la Península de 2012 a 2019

Nuestro gráfico comienza en el 2012, año en el que todavía España esta sumergida en la crisis económica, la cual provocó un descenso notable de demanda ya que en años anteriores al comienzo de la misma se rondaban los 260TWh de consumo anual. A partir que el país empieza a salir a flote y consigue recuperarse, del mismo modo,

aumenta la demanda anual de energía eléctrica como se puede apreciar en el gráfico: tenemos una tendencia ascendente desde el 2013 hasta nuestro penúltimo año de estudio que es el 2018.

Se observa un valor mínimo en 2014 con 244TWh y un máximo que se alcanza con gran diferencia en 2018 con 254TWh. Tras 4 años de crecimiento, se rompe este crecimiento positivo ya que en 2019 sólo se alcanzan 249TWh. Por último, vemos un intervalo de 10TWh en los que ha fluctuado la demanda.

A continuación, analizaremos la oferta, es decir, la producción de energía que hemos en el mismo período de tiempo.

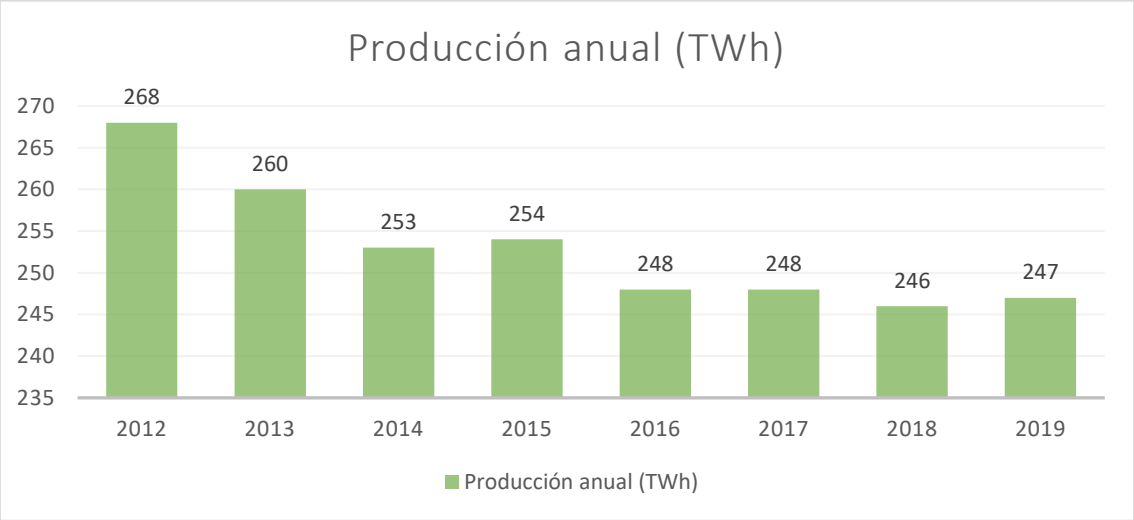

Figura 5: Evolución de la producción anual de energía eléctrica en la Península de 2012 a 2019

Analizando la oferta vemos una mayor fluctuación que la que teníamos en la demanda ya que hay 21 TWh de diferencia a lo largo de los años. Por otro lado, podemos ver como en los años de crisis desciende de forma notable la generación de electricidad, hasta 2015 no era necesario importar energía eléctrica puesto que la producción era mayor que la demanda.

En 2016 vemos que esto cambia con una estabilización de la producción en torno a 247 TWh durante 4 años en los cuales sí que será necesario la importación de energía de países vecinos puesto que no tenemos suficiente para abastecer satisfactoriamente la demanda.

Comparando ambos gráficos de demanda y oferta nos damos cuenta de que España pasa de ser un país exportador de energía eléctrica a ser un país que necesita importar esa energía eléctrica.

Otro de los factores importantes en la evaluación del precio de la energía eléctrica es su origen, es decir, energía renovable o no renovable, por lo que vamos a observar unos gráficos que nos detallan cómo se genera la energía renovable o no renovable durante nuestro período de tiempo de 7 años.

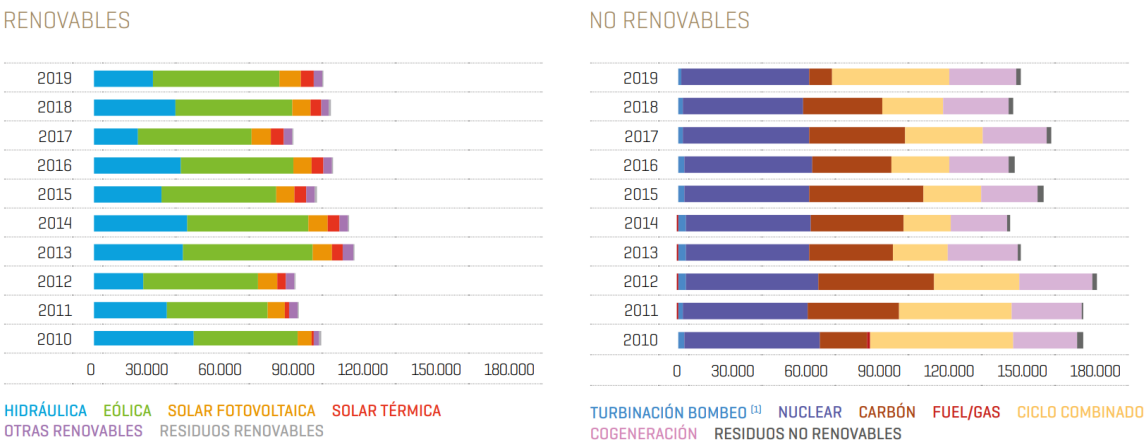

Figura 6: Fuentes de energía renovables y no renovables durante la última década

A simple vista del gráfico de energías renovables se aprecia la trascendencia que tiene la energía eólica puesto se sitúa como la segunda fuente de generación eléctrica durante los últimos 4 años. La energía hidráulica es la que más variaciones tiene a lo largo de los años y la solar fotovoltaica está en pleno auge alcanzando su máxima con un 9,2% de generación en el ámbito de renovables.

En la otra cara de la moneda nos encontramos a las energías no renovables, predomina la generación nuclear con valores muy constantes y la que más varía es el carbón, el cual ha visto reducido su generación notablemente en el último año de estudio 2019 y del mismo modo las centrales de ciclo combino han duplicado su generación a pesar de que los últimos años venían con una tendencia decreciente.

Para finalizar este apartado, en la gráfica que observamos a continuación vemos la evolución de las energías renovables frente a las no renovables y sus respectivos porcentajes de participación en la generación eléctrica:

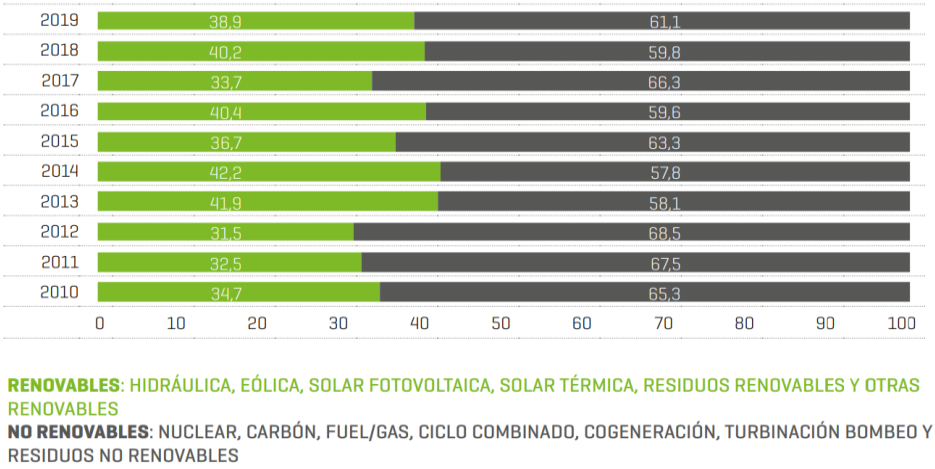

Figura 7: Generación de energía durante del año 2010 -2019

En la década se experimenta el mayor cambio en 2013, ya que nos situamos en valores en torno al 40% cuando anteriormente nos colocábamos en poco más del 30 %. Desde ese momento hasta el 2019 tendremos valores en torno al 39% con altibajos dependiendo del año. Se refleja de forma obvia la predominancia de las fuentes de energía no renovables.

2.3 Mercado eléctrico en España

Como ya hemos mencionado, en 1997 se liberaliza el mercado eléctrico español y esto conlleva a la compra y venta de energía eléctrica en plazos de meses, semanas, días e incluso horas.

Tenemos dos tipos de actividades en el país: las regulas por el Estado como son el transporte y la distribución de energía eléctrica; y las actividades liberalizadas que son generación y comercialización de energía. Las actividades liberalizadas son realizadas por las cinco grandes empresas eléctricas que tenemos en España: Endesa Iberdrola, Naturgy, EDP España y Viesgo.

El mercado eléctrico está dividido en dos sectores:

- Mercado minorista: comprende a la mayor parte de consumidores
- Mercado mayorista: tiene un modo de operación similar al que tenemos en la bolsa de valores

Mercado minorista

En este caso los consumidores compran su energía a las empresas que hacen su correspondiente oferta, que a su vez estas empresas obtienen su energía del mercado mayorista y transfieren a los pequeños compradores el coste establecido previamente en la administración.

La factura se divide en dos partes: el coste de la energía y el coste regulado. El coste de energía fluctúa según el precio de la energía eléctrica y también depende de los servicios de ajuste. El coste regulado es aquel que conforma la tarifa de acceso, esta

tarifa es la que asume el coste de redes de transporte y de subsidios a alas renovables entre otros costes.

La competencia en el mercado, entre las grandes empresas eléctricas por querer más clientes se basa en el coste de la energía ya que el otro tipo de coste está fijado por el gobierno. Es por ello que las grandes eléctricas buscan tener acuerdos con las generadoras más baratas para de ese modo poder sacar el máximo beneficio posible.

El precio que se aplica a los clientes pequeños es el PVPC que se calcula a partir de la media de los precios horarios de la electricidad ponderados correspondientemente por el consumo del cliente.

Mercado mayorista

Se conoce como el MIBEL (Mercado Ibérico de la Electricidad) y está formado por aquellos países que forman la Península Ibérica: Portugal y España. Es el resultado de un proceso de cooperación desarrollado por ambos países con la finalidad de promover la integración de los sistemas eléctricos de los países. Los resultados que se derivan de ello establecen una contribución relevante, no sólo a la consecución del mercado de la electricidad en la Península Ibérica, sino también a nivel europeo; se trata de un paso importante hacia la construcción del mercado interior de la energía.

En este mercado se producen intercambios de todo tipo de productos como puede ser un acuerdo cerrado de distribución de electricidad a una fábrica por hora para un trimestre con meses de antelación. Hay acuerdos de especies de entregas de electricidad a una hora especifica con pocas horas de antelación.

Encontramos dos operadores principales en el mercado: OMIE y OMIP. La diferencia entre ambos es que el OMIE es para productos en corto plazo y el OMIP para productos más a largo plazo.

Lo que más relevancia tiene en este mundo es el mercado diario, ya que hay grandes incentivos para que todos los generadores presentes sus correspondientes ofertas. El mercado funciona de la siguiente manera; los generadores presentan ofertas de venta a un determinado precio y con una cantidad concreta de energía para cada una de las 24h del día siguiente, llegando incluso a poder presentar hasta 25 ofertas distintas para cada hora del día siguiente. El procedimiento que se realiza para determinar el precio consiste en estimar la demanda y el precio inicial será el precio del último megavatio hora que satisfaga las condiciones de demanda.

Por otro lado, existe una regulación interna que asegura que las centrales no tienen que hacer frente a grandes subidas o bajadas de potencia. También hay un límite en el propio sistema eléctrico ya que las redes de transporte en la Península están limitadas.

Las ofertas pueden tener grandes diferencias entre ellas dependiendo de la fuente de energía. Por ejemplo, las centrales nucleares ofertan energía a bajo precio ya que tienen un coste elevado de arranque y un precio bajo de combustible, en cambio, si la oferta proviene de una central de carbón, puesto que el combustible es más caro, la oferta será más alta.

El precio medio aritmético diario en España de 2019 se situó en los 47,68€/MWh, valor inferior en un 16,8% al del año anterior (57,29€/MWh) y ligeramente superior al de Portugal que fueron 47,87€/MWh. Como es lógico, puesto que comparten el mismo mercado eléctrico, los precios rondaran los mismos valores.

A continuación, vemos un gráfico de la Generación de energía eléctrica en España en los dos últimos años y el precio del MWh que tenemos en cada mes. Se observa un precio más elevado durante el 2018 que alcanza su máximo en Septiembre de 2018 y su mínimo en Diciembre de 2019. Una observación importante para este tipo de gráfico es que un aumento en la generación renovable produce una bajada en el precio del MWh.

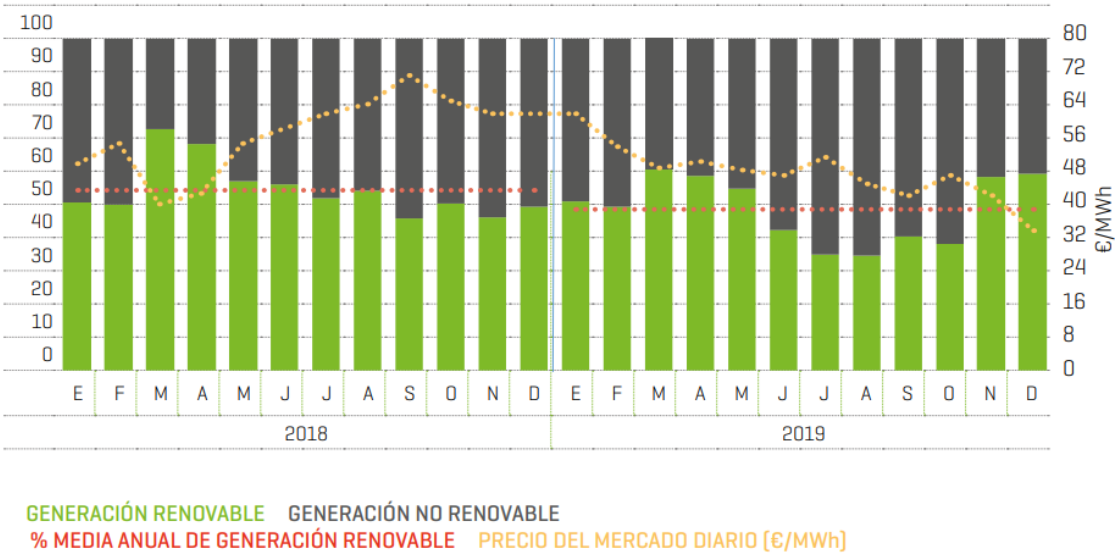

Figura 8: Variación del precio de la electricidad mensualmente y el porcentaje de generación renovable a lo largo del año 2018 y 2019

La gráfica que acabamos de analizar reproduce datos del mercado diario, el cual protagoniza gran parte del mercado eléctrico, pero no es el único componente del mismo ya que también existen otros tipos de mercado u organismos que influyen en el conjunto global que conocemos como MIBEL como son el mercado intradiario, los servicios de ajuste, los pagos por capacidad y los servicios de ininterrumpibilidad.

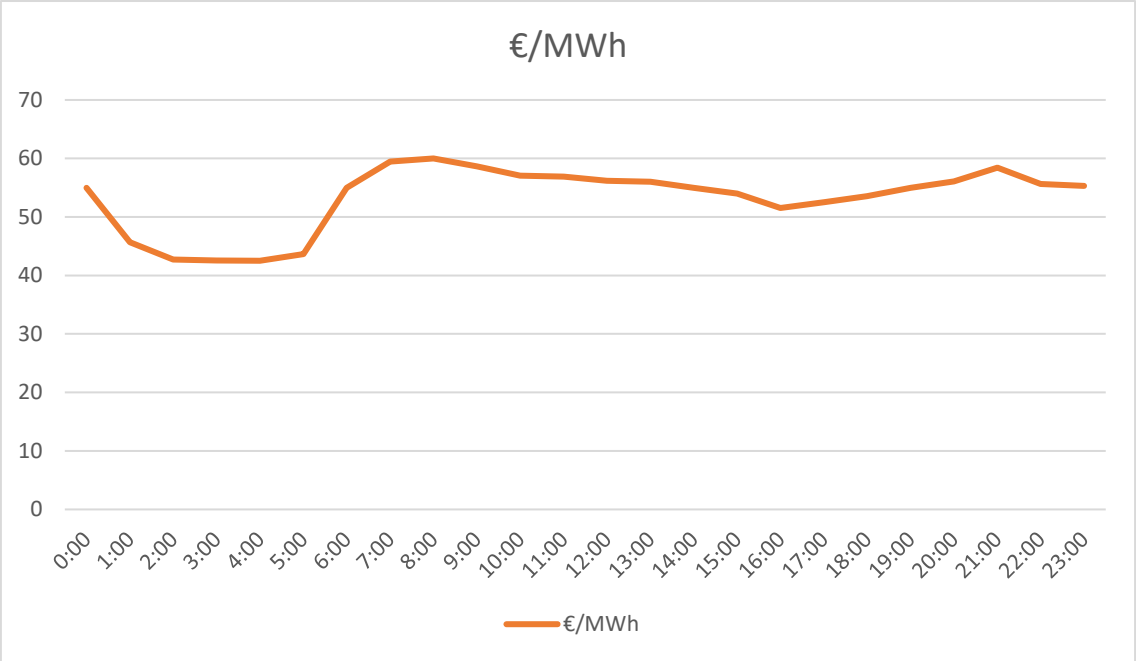

Figura 9: Precio del MWh durante el día 8/04/2019

Se ha seleccionado un día al azar del 2019 para observar la fluctuación del precio del Megavatio-hora en el mercado diario. La cuota que ocupan los comercializadores libres en este mercado ronda el 85-90% y del resto de porcentaje se encarga las comercializadoras de referencia.

En la gráfica se observa que el coste del MWh es más caro cuanto mayor demanda de luz haya en el país, es decir, a primeras horas de la mañana 7:00, 8:00 o 9:00 observamos el coste más elevado pues la mayoría de la población está en sus casas y utiliza electricidad al levantarse, y a su vez nos encontramos la misma temática a las horas finales del día cuando comienza a irse el sol y se ve una tendencia creciente que alcanza el pico sobre las 21:00-22:00. En el lado opuesto, vemos que el menor precio es por la noche cuando la mayoría de la población está durmiendo y no demanda el uso de la energía eléctrica.

A lo largo de las 24h hay una variación de 17.49€/MWh entre el mínimo valor de la gráfica que se alcanza a las 4:00 y el máximo que encontramos a las 8:00 con 60.00€/MWh

3. Objetivos

El objetivo esencial del proyecto de fin de grado se focaliza en el estudio de uno de los algoritmos más novedosos en el mundo de los árboles de predicción, concretamente se buscará un modelo que pueda predecir cierto nivel de coherencia los precios de la energía eléctrica.

En el mundo del DataMining y Machine Learning existen muchas formas y algoritmos que se utilizan analizar y trabajar con un cierto volumen de datos a los cuales se les pueden realizar clasificaciones, predicciones, árboles de regresión etc. cada día cobran más importancia ya que el mundo a medida que avanza temporalmente siempre va aumentar el volumen de datos. Por esto, las técnicas para analizar y sacar conclusiones en son esenciales en el mundo de la estadística, en el que actualmente el sueño del mismo es el BigData. En nuestro proyecto se tratará en concreto el algoritmo llamado boosting que permite analizar nuestros datos de forma que consigue crear una regla de predicción latamente precisa combinando numerosas reglas de predicción relativamente débiles o imprecisas.

Para conseguir nuestra predicción, necesitaremos conocer un poco el funcionamiento del mercado eléctrico y posteriormente nos enfocaremos en analizar distintas variables que pueden influir en mayor o menor medida en la fluctuación del precio de la electricidad. Nuestra obtención de datos se ha realizado gracias al historial de datos que hemos obtenido de distintas fuentes oficiales sobre variables como temperatura, precio de energía eléctrica en tiempos anteriores o la influencia actual de las energías renovables o demanda a lo largo de un año, de tal modo que se puede construir un modelo predictivo del precio de la energía eléctrica, del mismo modo podremos ver cómo afecta cada variable a la variación de precios de la energía eléctrica.

Un objetivo secundario que pretende alcanzarse con el correcto desarrollo del estudio reside en la capacidad de dar consciencia al consumidor, de modo que conozca y pueda estar familiarizado con los precios de energía eléctrica y con ello conseguir y un coste más bajo con el consumo de electricidad o incluso ser capaz de realizar una predicción del precio de su consumo eléctrico. También es importante que se conozcan aquellos factores más influyentes en las variaciones del precio. Así el sector podrá ir enfocándose en aumentar la competencia y del mismo modo conseguir precios más competitivos tanto para el consumidor como para los agentes encargados de la compra y venta de energía, por supuesto sin inducir grandes pérdidas económicas en el sector.

4. Fundamento teórico

4.1 Machine Learning

Machine Learning es también conocido en nuestro idioma como aprendizaje automatizado. Es un de las ramas de la computación y está relacionada con el concepto de Inteligencia Artificial (IA), que se utiliza a menudo para la creación de sistemas capaces de forma autónoma.

Está tecnología permite automatizar una gran cantidad de operaciones de forma que la actividad humana se convierte en innecesaria; lo cual aporta una gran ventaja a la hora de procesar grandes cantidades de información de una forma mucho más eficiente.

El concepto de *Machine Learning* es un aprendizaje debido a que consiste en la habilidad de un sistema para identificar múltiples patrones complejos determinados por una gran cantidad de variables o parámetros. Como es lógico, la máquina o sistema no aprende por iniciativa propia, sino que existe un algoritmo en su programación que se modifica constantemente con la entrada de datos en la interfaz, de tal manera que puede predecir escenarios futuros o tomar decisiones de manera autónoma bajo ciertas condiciones. Como todo este proceso se realiza sin intervención humana, por ello se dice que el aprendizaje es automático.

En los orígenes de la programación, la forma de conseguir que un sistema nos obedeciera, era escribir un algoritmo que detallara por completo el contexto y los detalles de cada acción a seguir por el sistema. En cambio, en la actualidad, los algoritmos empleados en el *Machine Learning* realizan un gran porcentaje de estas acciones por su cuenta. Consiguen sus propios cálculos partiendo de los datos que han recopilado del sistema, es por ello que cuantos más datos se aporte al sistema, mejores y más precisas serán las acciones resultantes.

4.2 Tipos de Machine Learning

Dentro del mundo de *Machine Learning* tenemos varios modelos o técnica distintas en las cuales los procedimientos varía, pero siempre están fundamentadas en el aprendizaje automático.

Se clasifican en 3 grandes ramas: aprendizaje por refuerzo, aprendizaje no supervisado y aprendizaje supervisado. En este proyecto utilizaremos una técnica de aprendizaje supervisado.

El aprendizaje por refuerzo consiste en la máquina aprenda a actuar en un entorno determinado, es u aprendizaje basado en los hallazgos y se utiliza para maximizar el número de hallazgos.

El aprendizaje no supervisado se utiliza para la realización de modelos descriptivos mediante un cierto procedimiento en el cual la evaluación es cualitativa o indirecta. No se utiliza para realiza predicciones, sino que se usa para búsquedas específicas.

El aprendizaje supervisado consiste aportar una cantidad razonable de datos que se definen al detalle con etiquetas. Cuando hayamos aportado esa cantidad, se podrán introducir nuevos datos sin necesidad de etiquetas, en base a patrones distintos que se han ido registrando durante el proceso de entrenamiento, esto se conoce como clasificación.

Otro método consiste en predecir un valor cuantitativo, utilizando distintas variables que al combinarse junto con la introducción de datos permite predecir un determinado resultado. Esto se conoce como regresión y permitir definir a nuestra técnica del proyecto: Boosting.

En resumen, en el aprendizaje supervisado se usan distintos ejemplos concretos a partir de los cuales se busca generalizar para nuevos casos. Este tipo de metodología es utiliza para realizar modelos de predicción y es capaz de resolver problemas de clasificación y regresión.

4.3 Boosting

El algoritmo Boosting tiene su origen en el trabajo de Kearns y Valiant (1989) que se plantearon si existía la posibilidad de estimular o "boostear" un algoritmo teniendo en cuenta sus errores. Desde este primer momento que se planteó el boosting, se han seguido desarrollando y mejorando hasta el día de hoy.

El algoritmo Boosting ha sido la herramienta más potente en el mundo de la creación de árboles introducida en los últimos años. El Boosting es una técnica de Machine Learning que consiste en la idea de crear una regla de predicción lo más precisa posible combinando muchas reglas débiles e imprecisas, estas reglas débiles se conocen en inglés como base learners o weak learners. Concretamente, dentro del mundo del boosting, existe un algoritmo complementario llamado AdaBoost que significa AdaptativeBoost que se ha convertido en el algoritmo más estudiado y popular de la familia. Este algoritmo nació en 1997 y fue formulado por Robert Schafire y Yoav Freund. La misión de este algoritmo es entrenar iterativamente una serie de predictores base o también conocidos como base learners, de modo que en cada iteración se tengan en cuenta los errores que han aparecido y se dé más importancia o peso a aquellos resultados con menor error emitido, de este modo se logrará una predicción óptima en forma de árboles de regresión. También destacan el Gradient Boosting que detallaremos más adelante o XGBoost que es una modificación del anterior en el que existe una velocidad de ejecución menor, pero a cambio tenemos una maximización del rendimiento.

A la hora de hablar de Boosting es también importante hablar del concepto ensemble. Un ensemble es una combinación de modelos, y cada modelo trata un base learner. Por ejemplo, si se tienen como base learners diferentes modelos de regresión múltiple, el ensemble estará formado por n modelos de regresión múltiple, en los que obviamente encontraremos distintos valores de salida. Por tanto, se asume que el boosting es uno ensemble methods que existen con el que se consigue un clasificador fuerte.

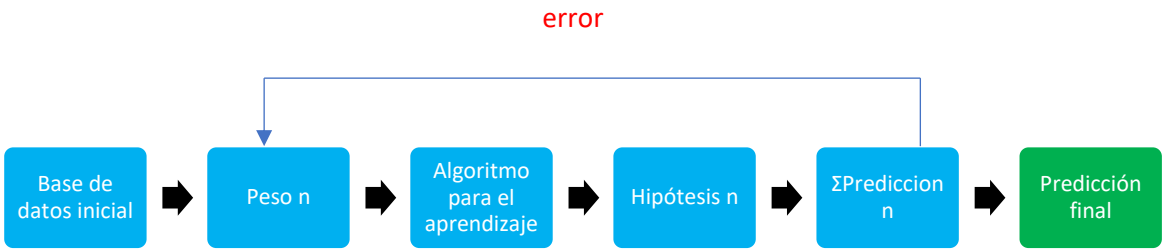

Figura 10: Flujograma resumido del algoritmo Boosting

En la zona superior vemos un esquema muy resumido de lo que sería el funcionamiento del algoritmo de boosting. Se comienza con una base de datos en la cual tenemos ciertos pesos, que en la primera iteración son exactamente iguales para todas las variables, el algoritmo realiza su aprendizaje y de ahí sale una primera hipótesis que luego formará parte del sumatorio total de predicciones y de este punto casi final del esquema vemos que sale el error que se retroalimenta y es una información muy valiosa para que el sistema pueda volver a actualizar los pesos de los datos que hay en el sistema en función de los resultados obtenidos, hasta que finalmente se consigue obtener una predicción final.

Un error común es confundir el boosting con el *Bagging*, ambos son técnicas de Machine Learning, pero tienen sus diferencias que voy a explicar a continuación con una imagen muy representativa para esta ocasión.

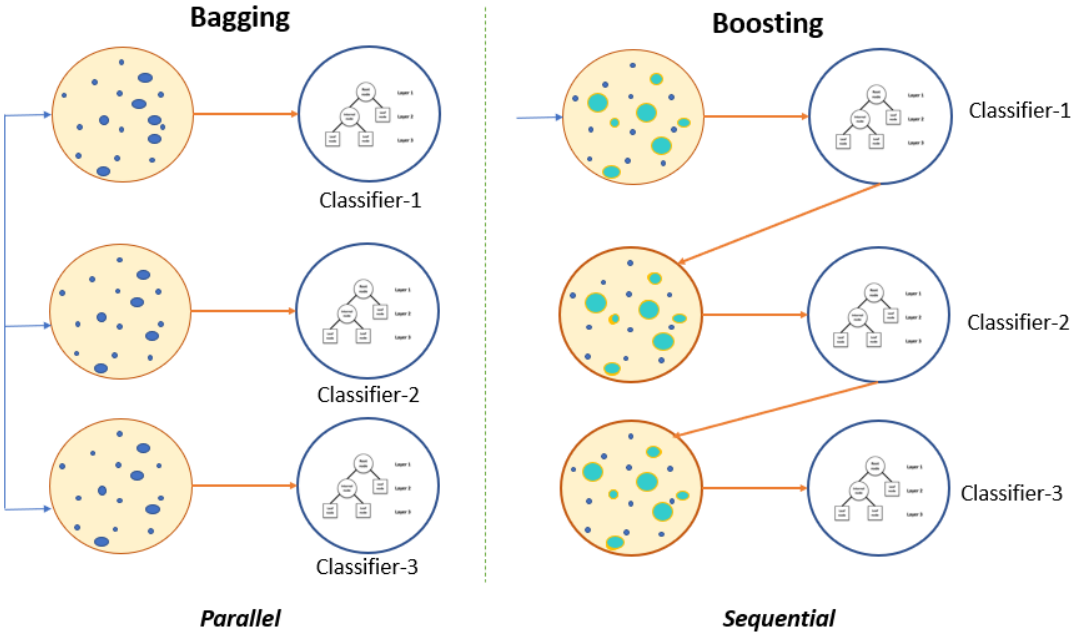

Figura 11: Comparativa de Bagging y Boosting

Ambos métodos son capaces de generar algoritmos fuertes mediante la combinación de algoritmos débiles. Los algoritmos en los métodos bagging son empleados en paralelo como se aprecia en la imagen. De esta manera, el objetivo es aprovechar la independencia que hay entre algoritmos simples, ya que el error se reduce notablemente al hacer el promedio de las salidas de los modelos simples. En términos coloquiales, este método daría como bueno aquello que eligiera la mayoría, es decir, los resultados similares, aunque independientes que más se repitieran.

En el otro lado de la imagen tenemos el Boosting que no actúa de manera paralela si no que su modo de actuación es secuencial. Cada modelo simple ocupa un lugar anterior o posterior con respecto al otro. El objetivo esencial que vemos en este tipo de métodos es aprovechar la dependencia entre estos modelos simples de tal forma que el rendimiento general mejora haciendo que un modelo posterior otorgue mayor trascendencia a los errores generados en modelos anteriores. En términos coloquiales, a la hora de resolver un problema el sistema Boosting observaría los errores que han cometido los anteriores sistemas o personas que hayan intentado resolver el problema y se fijaría en sus errores para tratar de no cometerlos.

La gran diferencia entre ambos métodos es que en Boosting los algoritmos no se entrenan de forma independiente, sino que se ponderan según los errores de los anteriores.

4.4 Gradient boosting

Es una parte importante dentro del ámbito del boosting puesto que es la técnica con la que más nos podemos asemejar a una optimización del modelo. Su esencia consiste en buscar una función de error cuyo gradiente tenemos que minimizar, los componentes de este gradiente serán los errores.

Según el tipo de problemas que se quiera tratar se utilizarán un tipo de errores u otros. Por ejemplo, si queremos tratar un problema de clasificación utilizaremos errores logarítmicos en cambio si queremos tratar un problema de regresión se utilizan errores cuadráticos.

El algoritmo débil que se utiliza en Gradient Boosting es el de árboles de decisión. Concretamente tiene la siguiente manera de actuar: se utilizan los árboles de regresión, los cuales generan valores reales para las divisiones cuyas salidas se pueden sumar, permitiendo de este modo que los resultados de los siguientes modelos sean agregados y corrijan los errores promediando las predicciones. Es un modelo aditivo ya que los árboles de decisión se van implementando uno por uno y los existentes no son modificados. Para saber que parámetros tendrá cada uno de los árboles se utiliza el proceso de gradiente descendente que minimizará la función de pérdida (error). Así se irán agregando árboles distintos de modo que la combinación entre ellos minimiza la pérdida en el modelo y mejora el valor de la predicción final.

Por todo esto, el uso de Gradient Boosting junto con los árboles de decisión es una de las técnicas más utilizadas para problemas de aprendizaje supervisado, aún así también presenta algún que otro inconveniente como el hecho de que se requiere un ajuste

cuidadoso de los parámetros para evitar el "overfitting" y esto puede requerir mucho tiempo de entrenamiento. El overfitting o sobreajuste se produce cuando un modelo está sumamente ajustado a los datos de entrenamiento que se provoca una mala generalización de los datos de test. El aprendizaje automático se basa en obtener patrones de ciertos datos para luego ser capaz de predecir de forma correcta los datos nuevos. Para evitar el overfitting se tiene que intentar suministrar el mayor volumen de datos posibles porque así aumentan las posibilidades de que el algoritmo sea capaz de generalizar mejor o simplificar los parámetros de los algoritmos como se puede hacer por ejemplo reduciéndola profundidad en un árbol de decisión.

4.5 Procedimiento matemático de AdaBoost

A continuación, se va a explicar el procedimiento matemático que sigue el algoritmo AdaBoost puesto que dentro del ámbito del Boosting es el más estudiado, como ya hemos mencionado anteriormente.

Dada una base de datos de entrenamiento (x_1, y_1), …, (x_m, y_m); donde m es la cantidad de muestras, $xi \in \boldsymbol{X}$ es el objeto o muestra a clasificar, $y_i \in \{-1, +1\}$ es la clasificación.

T se refiere al número de clasificadores a utilizar.

Inicializar la función de distribución de probabilidad $D_1(i) = 1/m$ para todo $i = 1, \ldots m$.

Para $t = 1, \ldots, T$.

1. Entrenar un clasificador débil teniendo en cuenta la distribución D_t
2. Seleccionar una hipótesis débil $h_t : \boldsymbol{X} \rightarrow \{-1, +1\}$ con el menor error ε_t

$$\varepsilon t = \sum_{m=1}^{m} Dt(i) * Pt(i)$$

donde $P_t(i) = 0$ si $y_i = h_t(i)$;

donde $P_t(i) = 1$ si $y_i \neq h_t(i)$;

3. Calcular el coeficiente α_t, que corresponde al peso del aprendiz débil de la iteración t

$$\alpha t = 0.5 * \ln\left(\frac{1 - \varepsilon t}{\varepsilon t}\right)$$

4. Actualizar D_t para i = 1, …, m:

$$D_{t+1}(i) = \frac{Dt(i) * \exp\left[-\alpha_t y_i h_t(x_i)\right]}{Zt}$$

donde Z_t es un factor de normalización escogido para que D_{t+1} sea una función de distribución de probabilidad.

5. Calcular la salido del clasificador fuerte H_{final} (x)

$$H_{final}(x) = signo\,[\sum_{n=0}^{\infty} \alpha_t h_t(x)]$$

5. Metodología

5.1 Software estadístico R

Existen muchos tipos de software y programas en la actualidad con los que se pueden realizar numerosas técnicas de Machine Learning, y en este proyecto en concreto vamos a utilizar el software estadístico R y el programa Rstudio para aplicar la técnica de boosting y conseguir una predicción en el precio de la energía eléctrica, que se adecue correctamente a los valores reales del precio.

R es un lenguaje libre especializado en programación estadística que se actualiza gracias a la Fundación para la Programación Estadística en R, la cual es una organización no gubernamental que actualiza y mejora el programa de forma continua. R es un GNU, es decir, es un sistema en el que los usuarios tienen la libertad de compartir y mejorar el software integrado. Fue creado en la Universidad de Auckland por Ross Ihaka y Robert Gentleman en el año 1993, pero no fue hasta el año 2000 cuando finalmente se hizo público y la población pudo comenzar a utilizarlo.

R tiene una importante cantidad de técnicas estadísticas como análisis de series temporales, análisis cluster, clasificaciones, modelos lineales y no lineales etc. Su ventaja esencial es la facilidad que existe para generar diferentes tipos de gráficos y diagramas, junto con la posibilidad de implementar diferentes paquetes que nos permiten formular múltiples algoritmos y estudios estadísticos.

Rstudio es el programa que se utilizara, es gratuito al igual que el programa R, pero se ha desarrollado para facilitar la navegación por el programa y tiene ciertas facilidades al ahora de meter los códigos necesarios para la programación. A pesar de ser un programa gratuito, tiene una muy buena capacidad de análisis y de cálculo a lo que hay que sumar la posibilidad de introducir tus propias fórmulas o extender las ya creadas para llegar aún más lejos en el análisis de resultados en el programa.

Como hemos mencionado anteriormente, existe la posibilidad de descargarse ciertos paquetes en caso de que se quiera profundizar en algún ámbito concreto. En nuestro proyecto esto va ser necesario y puesto que vamos a necesitar algoritmos de boosting, tendremos que descargarnos dos paquetes fundamentales para realizar los análisis convenientes, estos paquetes son mboost y psych. Aunque no se utilice la función mboost como tal ya que se optará por otras opciones como explicaremos en el siguiente epígrafe, pero estas funciones están dentro del paquete mboost.

A pesar de haber estudiado y aprendido el manejo de este programa a niveles básicos en la Universidad concretamente el segundo año de la carrera, para la realización del proyecto deberemos refrescar esos conocimientos y profundizar para poder alcanzar los objetivos propuestos

Del mismo modo también necesitaremos la instalación del paquete readxl para poder importar nuestros datos de Excel a Rstudio y que el programa los interprete de forma correcta.

5.2 Generación del modelo

5.2.1 Base de datos

Tras todo el estudio e investigación necesaria, se obtuvieron los conocimientos suficientes del software estadístico R y del programa en concreto que íbamos a utilizar que ha sido Rstudio, del mismo modo, nos ayudamos de ciertos trabajos y documentos que se han realizado con anterioridad al nuestro ya sean de boosting o de predicción de precios de energía eléctrica y una vez estamos preparados comenzamos con la base de datos.

Para la posible realización del modelo nos hemos ceñido a datos del año 2019 recopilados en una base de datos en el programa Microsoft Excel. Esto se debe a que el año 2020 fue año de pandemia mundial que tuvo datos muy atípicos respecto a consumos y precios de energía eléctrica causados por la cuarentena que se vivió en España durante 2 meses en los que la gran mayoría de la población española estaba en su casa y se dio un colosal aumentó del teletrabajo por lo que esto provocó unos valores para nada acordes con los valores a los que estamos acostumbrados en esa época del año en años anteriores.

Hemos seleccionado las variables: precio de la energía eléctrica en €/Wh, precio de la energía eléctrica en el instante t-1, energía generada sin contaminación de emisiones CO_2 medido en MW, temperaturas máxima en grados centígrados ºC, temperatura mínima en grados centígrados ºC mínima y la temperatura media entre estas dos últimas variables que por supuesto, también se mide en grados centígrados ºC; y por último la demanda medida en MW.

Variables
Precio de la energía eléctrica
Precio de la energía eléctrica en el instante t-1
Energía generada sin CO_2
Temperatura máxima
Temperatura mínima
Temperatura media
Demanda

Tabla 1: Tabla con los nombres de las variables

Los datos se han obtenido de diversas fuentes, para conseguir los precios no ha sido de gran ayuda la página del OMIE (Operador del mercado Ibérico de Energía), para las temperaturas nos hemos ayudado de AEMET (Agencia Estatal de Meteorología) y por último de la REE (Red Eléctrica de España) hemos obtenido las variables de generación y demanda.

La variable que queremos predecir es el precio de la energía eléctrica, para ello utilizaremos el resto de variables y generaremos un modelo, en el cual, mediante

algoritmos de boosting que nos reflejará la importancia de las variables para nuestra predicción final.

Utilizaremos una base de datos de 8760 filas y 7 columnas, donde las columnas son las variables y las filas son el número de horas que hay en 2019, es decir, cada fila representa una hora concreta y en esa hora tenemos siete variables.

5.2.2 Boosting en Rstudio

En el ambiente de Rstudio hemos utilizado esencialmente el paquete mboost para la realización de un modelo boosting, pero también se utiliza para realizar otros tipos de estudios de clasificación, regresión etc. El paquete es capaz de realizar árboles en los cuales se da mayor importancia a las variables más significativas de este modo se consigue que las futuras predicciones sean lo más exactas posibles. El paquete presenta distintos algoritmos para la generación de un modelo u otro, pero tiene un procedimiento común que sigue los siguientes pasos:

1. Al principio se les otorga el mismo peso absolutamente a todos los datos del conjunto de entrenamiento. Este peso se puede formular como $w_i = 1/n$; donde n es el número de datos del conjunto e i se extiende desde 1 hasta n, siendo n un número finito.
2. Se produce el entrenamiento del modelo con esos datos de entrenamiento y con los pesos iniciales obteniéndose así el primer base learner.
3. Se calcula el error del modelo en el conjunto de entrenamiento con los pesos anteriores
4. Se produce un incremento de peso en aquellos datos que el modela ha clasificado de forma errónea. Esto se hace para conseguir que en la siguiente iteración obtener un base learner que tenga en mayor consideración estos datos y se acerque más a ellos
5. Se vuelve a entrenar un nuevo modelo con un nuevo base learner con el conjunto de entrenamiento con los pesos modificados
6. Se repite la secuencia continuamente desde el paso número 3 hasta el punto de iteraciones fijadas
7. Cada uno de los modelos obtenidos que conlleva un base learner s le habrá asignado un peso. De tal forma que el modelo final se calcula con una votación ponderada por los pesos de todos los modelos.

Dentro de este paquete existen varias funciones que se pueden encargar de realizar boosting como son gamboost, mboost, blackboost, glmboost. Concretamente, nos hemos centrado en dos funciones para la realización de nuestro proyecto, gamboost y blackboost.

Para realizar cualquier boosting, por definición se necesitan unos datos de entrenamiento con los que el algoritmo pueda trabajar para posteriormente optimizar los errores en la propia predicción con los datos de test. Concretamente para este trabajo de finde grado se ha actuado de la siguiente manera: para el entrenamiento se han seleccionado los 8160 primeros datos y luego el test se ha realizado con con las 600 últimas filas de la base de datos.

Como hemos mencionado anteriormente, se ha optado por elegir la función gamboost frente a mboost porque en la primera se pueden elegir el grado de libertades y es un boosting más generalizado. Es una función que optimiza el peso de los errores a través de muestras aleatorias que utilizan los diferentes datos como bases de aprendizaje a través de árboles aditivos y por otro lado blackboost optimiza los errores utilizando árboles de regresión como bases de aprendizaje. Tras la realización de distintos modelos con ambas funciones veremos cual se adecúa mejor.

Para calcular el mayor o menor ajuste que se ha realizado y también el comportamiento del modelo para todas las horas del año, se ha utilizado el MAPE que significa Mean Average Percentage, en español quiere decir: error absoluto medio en porcentaje y al estar en valor absoluto no nos transmite si el error es por defecto o por exceso, pero sí que nos transmite una información vital sobre la coherencia del modelo.

El error mide el tamaño del error en términos absolutos y porcentuales. La fórmula que se utiliza para calcular este tipo de error es la siguiente:

$$\boldsymbol{MAPE = 100}\left(\sum_{\boldsymbol{k=1}}^{\boldsymbol{n}} \frac{|\boldsymbol{Ai-Fi}|}{|\boldsymbol{Ai}|}\right) \boldsymbol{/}\ \mathrm{n}\ \boldsymbol{;}$$

Puede ser resultar interesante para el lector que en el software estadístico R se permite realizar el cálculo del MAPE puesto que hay una función denominada mape que nos permite calcular el mape entre una predicción y los valores reales sin necesidad de utilizar la fórmula, esta función se encuentra dentro del paquete MLmetrics, no obstante, en nuestro caso hemos aplicado la fórmula directamente.

La correspondiente programación que hemos utilizado para la realización del proyecto estará adjunta en el Anexo 1: Código de R en la cual se podrá observar la realización de ciertos comentarios de carácter explicativo en el propio script de Rstudio, para que así, sea más fácil el entendimiento del código programado y poder seguir de esa forma los pasos necesarios para la comprensión del modelo.

En este proyecto se ha estudiado la predicción del precio de la energía eléctrica para distintos horizontes de tiempo. La primera predicción se ha realizado para obtener el precio de la energía eléctrica en el instante t+1, la segunda de ellas se ha realizado para un instante t+3 y por último hemos reflejado una predicción para t+8.

Para no extendernos en exceso en la escritura durante la realización el proyecto, lo escribiré coloquialmente, lo indicaré como t+3, t+3 etc. pero realmente lo que se indica es que la predicción es para 1,3 u 8 horas siguientes, ya que nuestra base de datos de datos está dividida por horas, concretamente 8760h.

Cómo es lógico, a medida que vayamos avanzando en el tiempo obtendremos resultados menos precisos y aumentaremos el error, lo analizaremos y comentaremos con la ayuda de gráficas en el siguiente capítulo.

Se observa también la influencia de las variables independientes que en nuestro caso son 6, precio en el instante anterior, energía generada sin emisiones de CO_2, demanda de energía eléctrica, temperatura media, temperatura máxima y temperatura mínima, en la variable que se quiere predecir que es el precio de la energía eléctrica.

También se consideró implementar en el proyecto el boosting con la función gbm para realizar predicciones, pero tras la realización de una serie de prueba se observó que no se obtuvieron buenos resultados y lo errores eran mucho mayores que los obtenidos por tanto debido a la falta de coherencia en las predicciones de prueba realizadas, se decidió no seguir adelante con este algoritmo y descartarlo del proyecto para centrarnos en gamboost y blackboost.

6. Análisis de los resultados

En este apartado nos centraremos en el análisis de resultados, visualizaremos gráficas comparativas del error calculado en las predicciones del modelo con los valores reales de la variable precio de la energía eléctrica, comentaremos la cercanía de las predicciones a la realidad observando sus errores, concretamente su MAPE y también se apreciará una gráfica comparativa de la predicción y la realidad. Se realizarán predicciones en tres horizontes distintos: previsión en horizonte t+1, es decir, predicción a una hora después, luego tendremos una predicción a t+3 es decir, 3 horas después y por último terminaremos este capítulo con la predicción a t+8 que indica 8 horas después y comparando y discutiendo los datos obtenidos en cada una de las predicciones.

Cada horizonte tendrá dos tipos de predicciones, una predicción con la unción gamboost y otra predicción con la función blackboost

6.1 Predicción para t+1

A continuación, vamos a analizar los resultados obtenidos en el estudio para t+1. En este estudio se toman las variables de: energía sin CO_2, demanda y temperaturas en el instante t+1 y el precio en el instante t, con esta combinación realizaremos nuestra predicción para el precio de la energía eléctrica en el instante t+1.

Se realizarán dos modelos con dos tipos de funciones distintas, gamboost y blackboost y veremos cual de ellos se adecúa mejor a los resultados.

Gamboost

A continuación, se observa un gráfico que muestra la correlación entre el precio real y la predicción realizada en este modelo, se ha calculado el valor de la correlación entre precio y predicción y se obtiene un valor de 0,89

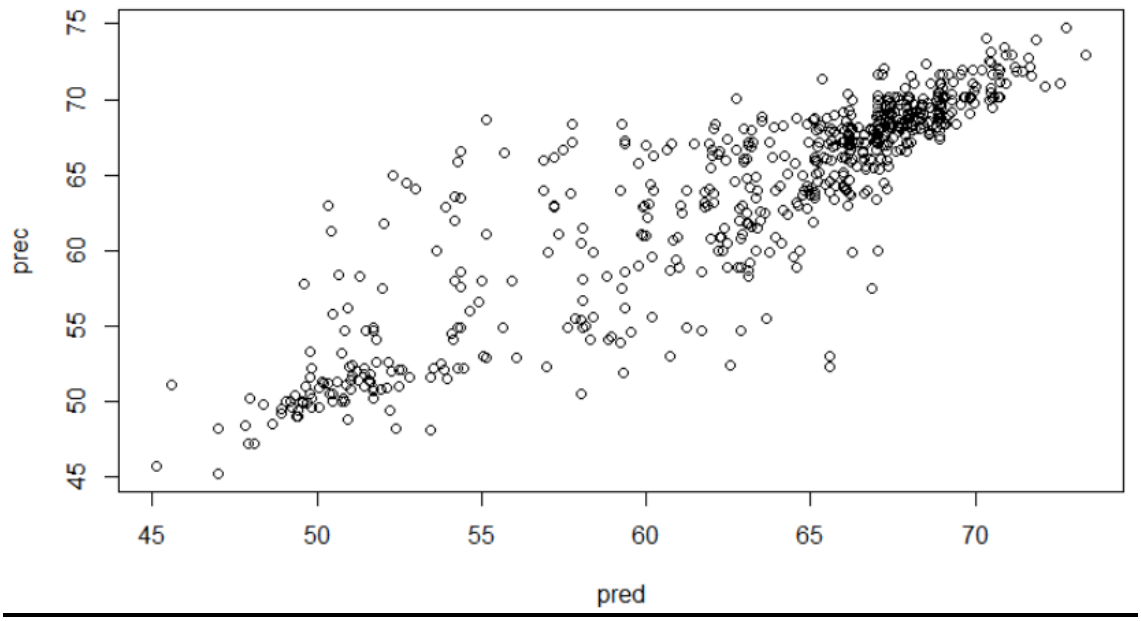

Figura 12: Correlación entre predicción y el precio real de la energía eléctrica en t+1

Tenemos un valor del MAPE inferior al 4% lo que nos indica que sí puede ser un buen modelo competitivo y fiable para este tipo de predicciones

Error mínimo	Error máximo	MAPE
0.000166%	20.43%	**3.59%**

Tabla 2: Errores en predicción gamboost en t+1

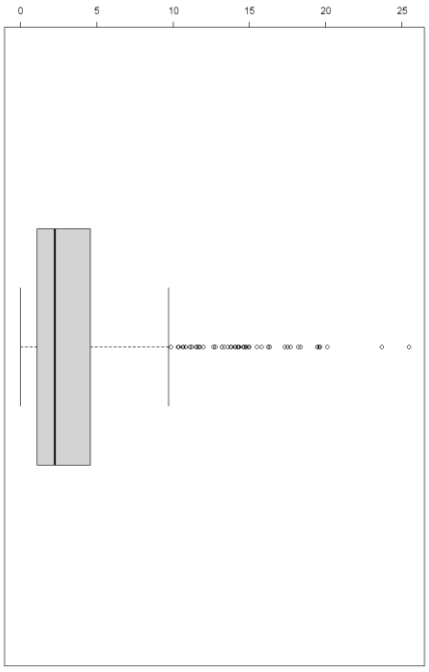

Figura 13: Gráfica de caja y bigotes de predicción gamboost+1

Como podemos apreciar gracias a este gráfico, se observa que el 50% de los errores cometidos con esta función gamboost para t+1 se encuentran entre el 2,27% y el 4,56%

Por último, para terminar este apartado con gamboost, vamos a ver un dibujo de la predicción junto a otro de los valores reales. Observaremos en las mismas que el modelo diseñado sigue una coherencia notable con los precios reales del precio de la energía eléctrica, a nuestra izquierda tenemos la predicción y a la derecha tenemos los valores reales.

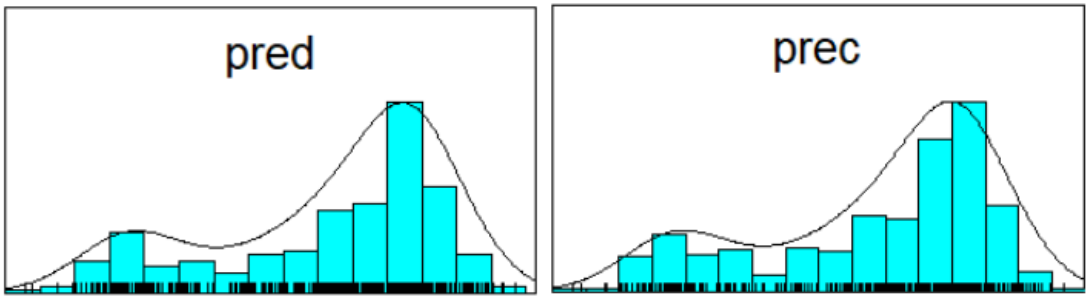

Figura 14: Gráficas comparativas de predicción gamboost en+1

Blackboost

Se observa en el siguiente gráfico que sigue existiendo un importante nivel de correlación, concretamente 0,87 entre la predicción y el valor real del precio de la energía eléctrica, pero vemos también que aumenta la variabilidad con respecto a la otra predicción con la función gamboost.

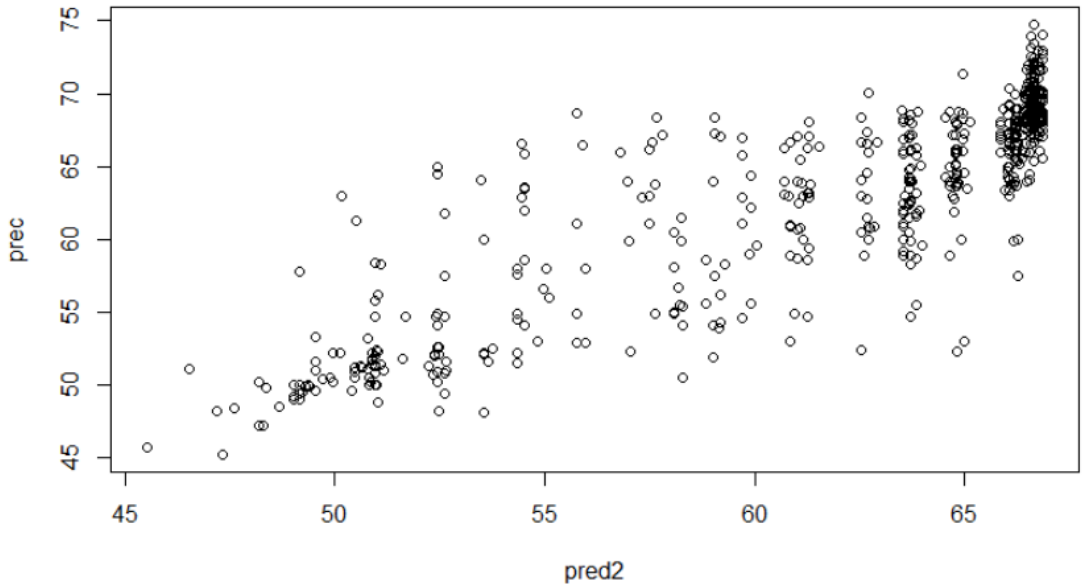

Figura 15: Correlación entre predicción y el precio real de la energía eléctrica en t+1

Nos encontramos con un valor MAPE de 4,4% ligeramente superior al anterior.

Error mínimo	Error máximo	MAPE
0.0015%	20.98%	**4.4%**

Tabla 3: Errores en predicción blackboost en t+1

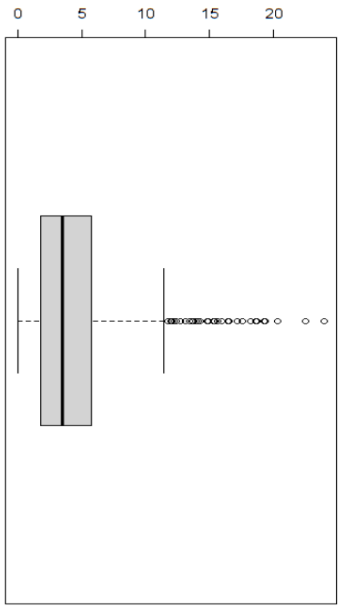

Figura 16: Gráfica de caja y bigotes de predicción blackboost en t+1

Se aprecia en el gráfico de caja y bigotes que el 50% de los errores cometidos se encuentran entre el 1,76% y 5,73%.

Por último, como hemos hecho anteriormente en gamboost, vamos a ver a continuación una comparación entre la predicción blackboost en t+1 y el precio real. Observaremos que a pesar de tener un modelo con un buen MAPE, la función blackboost en esta ocasión no ofrece una buena respuesta predictiva.

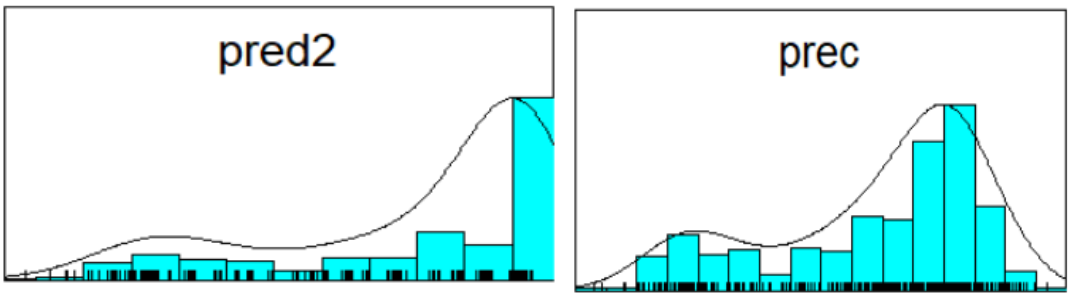

Figura 17: Gráficas comparativas de predicción blackboost en t+1

Se demuestra con claridad, que para este tipo de predicciones de 1 hora más adelante del tiempo actual, la función gamboost cumple mucho mejor con la realización de una predicción correcta.

6.2 Predicción para t+3

Tras la predicción de t+1, hemos decidido ir un poquito más allá y en vez de calcular t+2 que sería la siguiente predicción posible, hemos calculado el comportamiento y predicción del modelo para t+3. Al ser una predicción más lejana en el tiempo que la primera que hemos realizado, saldrán mayores niveles de errores. Para predecir el precio en t+3 energía sin CO_2, demanda y temperaturas en el instante t+3 y el precio de la energía eléctrica en el instante t.

Al igual que en el instante anterior, tenemos dos funciones gamboost y blackboost.

Gamboost

A continuación, se observa un gráfico que muestra la correlación entre el precio real y la predicción realizada en este modelo, se ha calculado el valor de la correlación entre precio y predicción y se obtiene un valor de 0,61, por lo que el gráfico se ve distinto respecto de la misma función gamboost en t+1 realizada anteriormente.

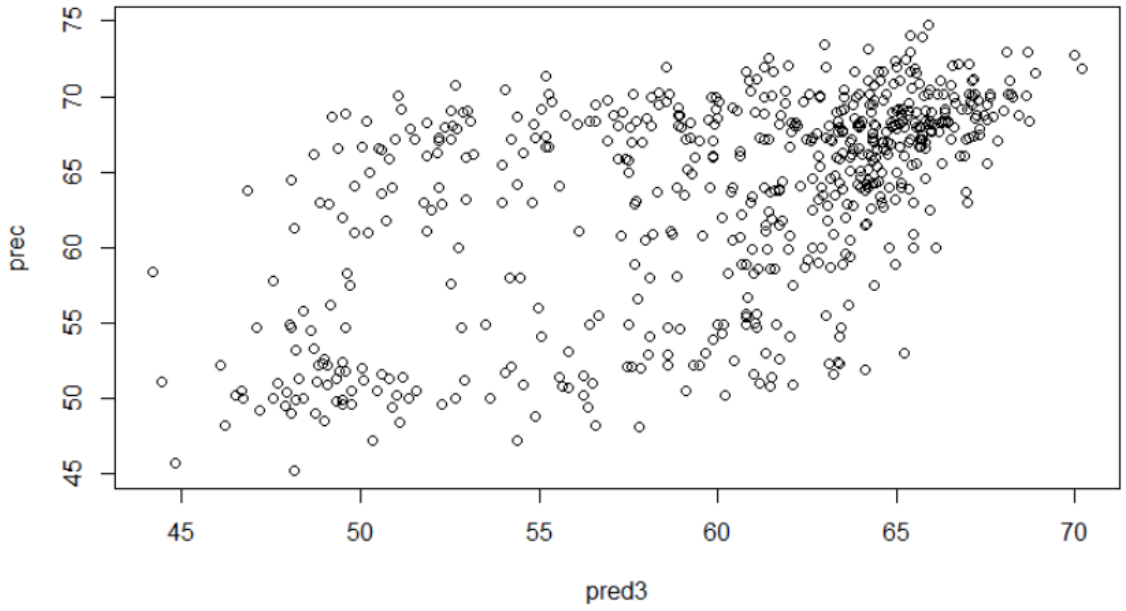

Figura 18: Correlación entre predicción y el precio real de la energía eléctrica en t+3

En esta predicción a t+3 se produce un aumento del MAPE incrementando su valor en torno al 5% y llegando al valor de 8.54%

Error mínimo	Error máximo	MAPE
0.0036%	26.06%	**8.54%**

Tabla 4: Errores en predicción gamboost en t+3

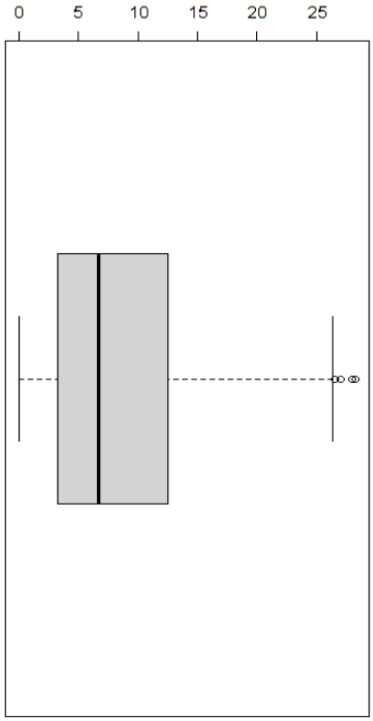

Figura 19: Gráfica de caja y bigotes de predicción gamboost en t+3

Se aprecia en el gráfico de caja y bigotes que el 50% de los errores cometidos se encuentran entre el 3.25% y 12.54%.

Lo cual amplía notablemente los límites de la caja respecto de nuestras predicciones anteriores, como es normal debido a que estamos en un horizonte de 3 horas más respecto de los valores reales del precio.

Para terminar este apartado con gamboost en t+3, vamos a ver un dibujo de la predicción junto a otro de los valores reales. Observaremos en las mismas que el modelo diseñado sigue una coherencia muy buena y similar a la que veíamos en t+1 con los precios reales del precio de la energía eléctrica, a nuestra izquierda tenemos la predicción y a la derecha tenemos los valores reales.

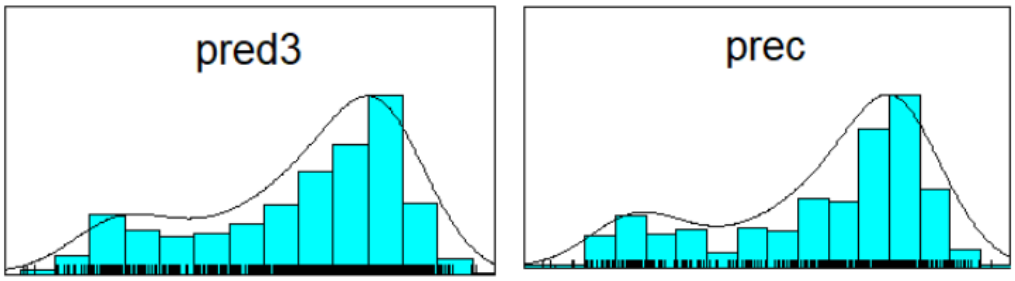

Figura 20: Gráficas comparativas de predicción gamboost en t+3

Blackboost

Se observa en el siguiente gráfico que tenemos un valor de correlación, concretamente de 0,61 entre la predicción y el valor real del precio de la energía eléctrica, y una gráfica muy similar al modelo gamboost en t+3, por lo que se puede decir que para este horizonte ambos modelos se comportan igual a pesar de tener distintos algoritmos.

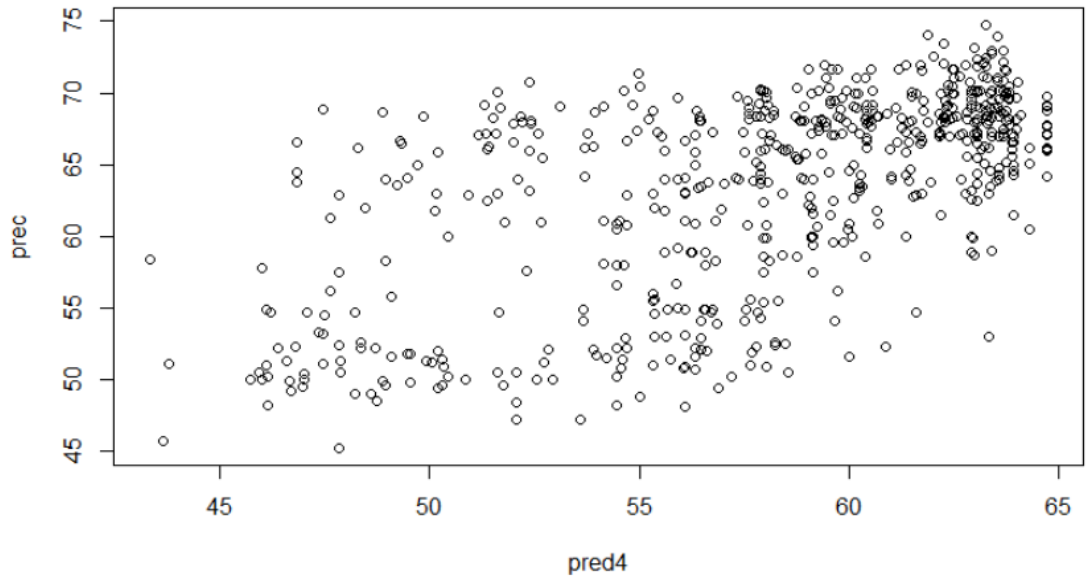

Figura 21: Correlación entre predicción y el precio real de la energía eléctrica en t+3

En esta predicción con la función blackboost a t+3 se produce un aumento del MAPE incrementando su valor en torno al 6% concretamente 9,92% y es un poco mayor del que hay en gamboost a t+3

Error mínimo	Error máximo	MAPE
0.03%	26.89%	**9.92%**

Tabla 5: Errores en predicción blackboost en t+3

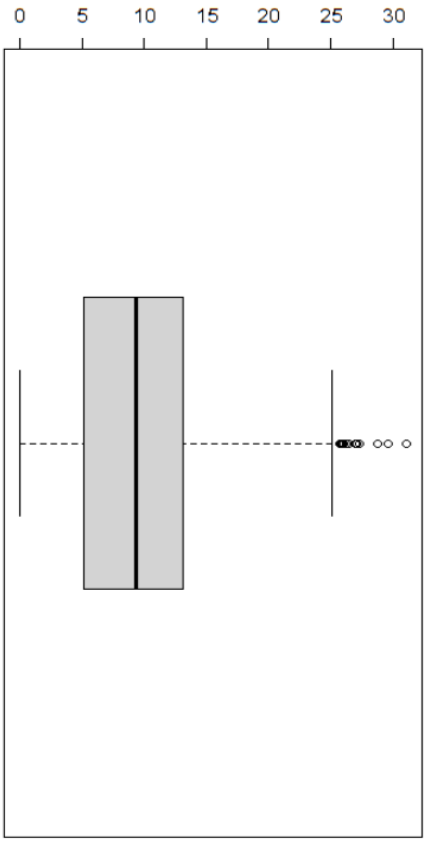

Figura 22: Gráfica de caja y bigotes de predicción blackboost en t+3

Se aprecia en el gráfico de caja y bigotes que el 50% de los errores cometidos se encuentran entre el 4.94% y 13.11%.

Es un gráfico de caja y bigotes muy similar al que teníamos anteriormente en t+3 con gamboost.

Para terminar este apartado con predicciones en t+3, vamos a ver un dibujo de la predicción junto a otro de los valores reales. Observaremos que esta predicción blackboost ha mejorado respecto a la realizada ent+1 ya que aquella gráfica no tenía relación con la realidad, en cambio en esta ocasión, sí que vemos como la gráfica se va adaptando a los valores reales lo que indica cierta mejoría, a nuestra izquierda tenemos la predicción y a la derecha tenemos los valores reales

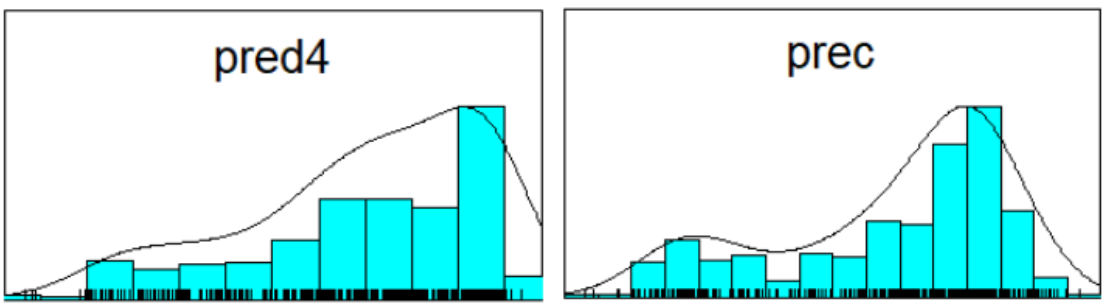

Figura 23: Gráficas comparativas de predicción blackboost+3

Tas observar el panorama de t+3 se puede apreciar que gamboost y blackboost se igualan en rendimiento respecto a las predicciones de t+1, pero sigue siendo mejor el modelo gamboost ya que se obtienen mejores resultados y se acerca ligeramente más a la realidad que buscamos.

6.3 Predicción para t+8

Nos presentamos en el horizonte de 8 horas más, las últimas predicciones que sacaremos serán en este horizonte y repetiremos el proceso de análisis que hemos llevado hasta ahora. Por ello, para predecir el precio en t+8 energía sin CO_2, demanda y temperaturas en el instante t+8 y el precio de la energía eléctrica en el instante t. Del mismo modo obtendremos una predicción para gamboost y otra para blackboost y veremos que resultados y errores arrojan.

Gamboost

Se calcula el valor de correlación que es 0,55 por lo que podemos decir que se asienta un poco alrededor de estos valores ya que no se experimenta la misma bajada de valor que vimos al pasar del campo de t+1 a t+3. A continuación, se observa un gráfico que muestra la correlación entre el precio real y la predicción realizada en este modelo.

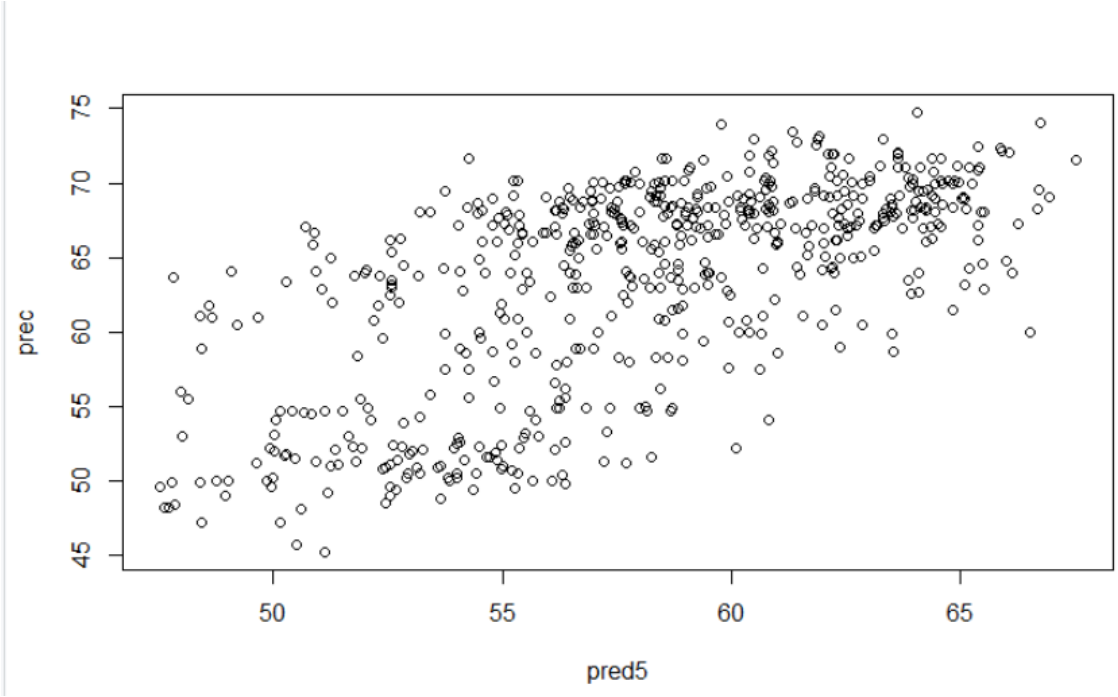

Figura 24: Correlación entre predicción y el precio real de la energía eléctrica en t+8

En esta predicción a t+8 observamos que hay un ligero aumento del MAPE y llega al valor de 10.55%, también aumentan los errores mínimo y máximo con respecto a la anterior predicción en gamboost

Error mínimo	Error máximo	MAPE
0.12%	28.29%	**10.55%**

Tabla 6: Errores en predicción gamboost en t+8

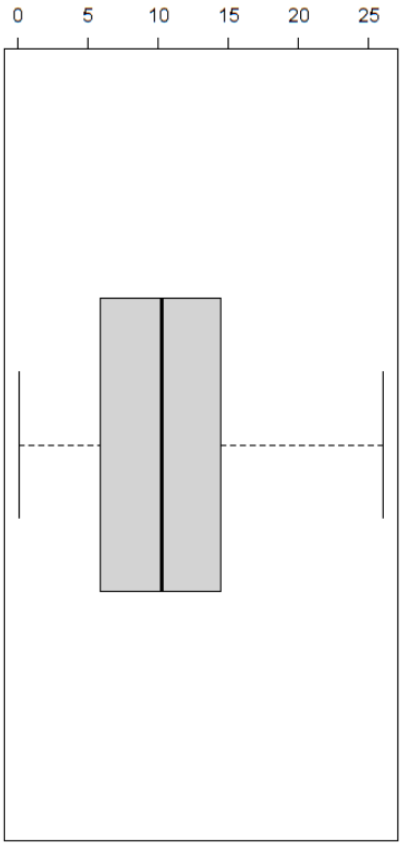

Figura 25: Gráfica de caja y bigotes de predicción gamboost en t+8

Se aprecia en el gráfico de caja y bigotes que el 50% de los errores cometidos se encuentran entre el 5,94% y 14.47%. Esto se traduce en un ligero aumento de la anchura de la caja del gráfico superior

Al igual que anteriormente, procedemos a mostrar el comportamiento de la predicción, en este caso, gamboost en t+8, reflejado en un gráfico para poder compararlo con los valores reales. A nuestra izquierda tenemos la predicción y a la derecha tenemos los valores reales

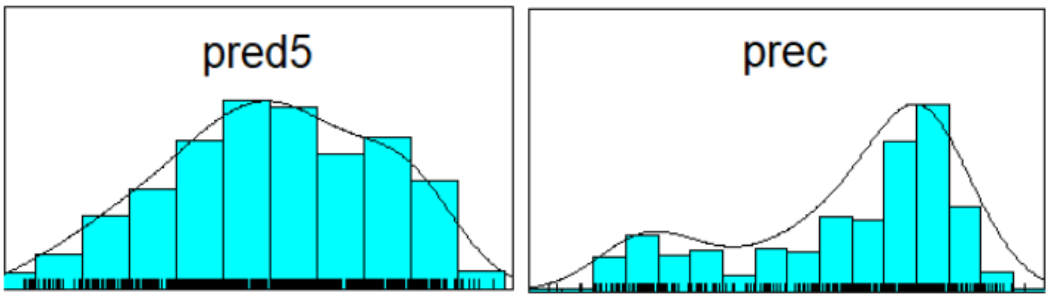

Figura 26: Gráficas comparativas de predicción gamboost en t+8

Se observa como debido al aumento del horizonte de predicción hasta 8 horas, el modelo de predicción gamboost para precios de la energía eléctrica pierde su coherencia poco a poco.

Blackboost

Existe un valor de correlación de 0,54 en el grafico que vamos a ver a continuación por lo que podemos decir que se asienta un poco alrededor de estos valores cercanos a 0.6 ya que no se experimenta la misma bajada de valor que vimos al pasar del campo de t+1 a t+3

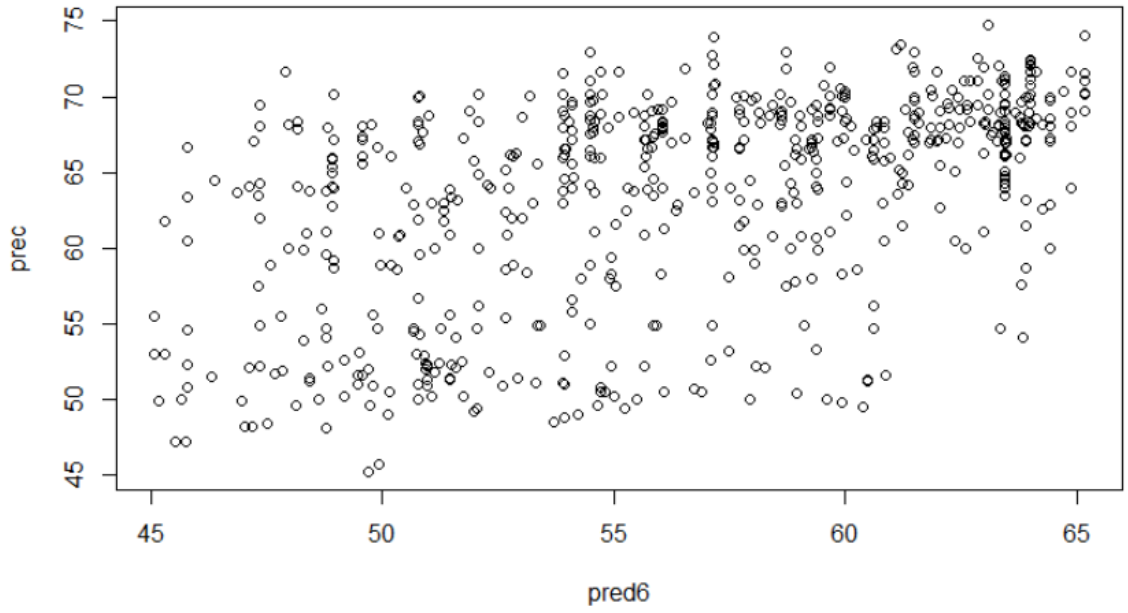

Figura 27: Correlación entre predicción y el precio real de la energía eléctrica en t+8

En esta predicción a t+8 observamos que hay un MAPE de 11,38%, al igual que en gamboost para t+8, son valores altos del MAPE por lo que se ve como a medida que avanzamos en el tiempo, empeora la predicción.

Error mínimo	Error máximo	MAPE
0.03%	31,08%	**11.38%**

Tabla 7: Errores en predicción blackboost en t+8

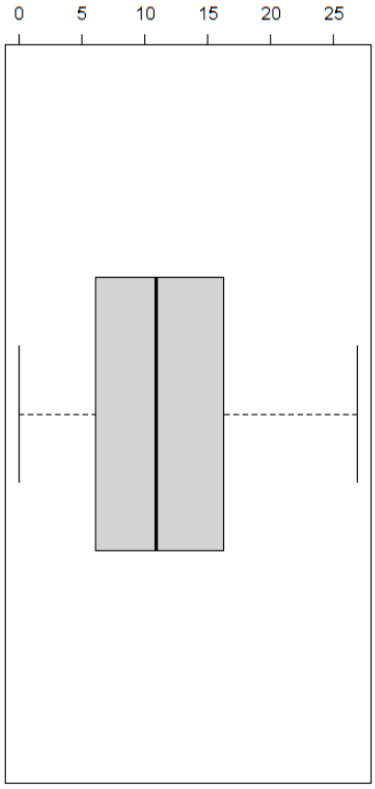

Figura 28: Gráfica de caja y bigotes de predicción blackboost en t+8

Se aprecia en el gráfico de caja y bigotes situado justo en la zona superior de esta hoja de texto, se ve con facilidad que el 50% de los errores cometidos se encuentran entre el 6.11% y 16,3%.

Para finalizar el análisis de esta última predicción blackboost para t+8 tenemos la comparativa de gráficas. En la izquierda colocamos la predicción y en la derecha la gráfica real.

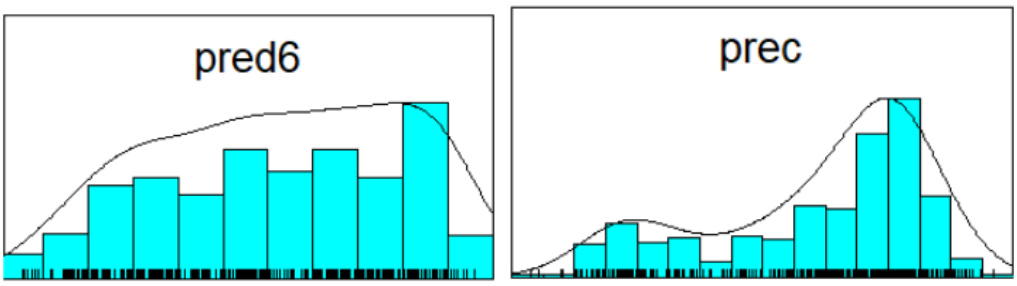

Figura 29: Gráficas comparativas de predicción blackboost en t+8

Al igual que en el caso anterior de gamboost para t+8, observamos que el modelo empeora frente a las predicciones de t+1 y t+3, no obstante, vemos una ligera mejoría respecto al

algoritmo gamboost ya que sigue y refleja mejor gráfica predictiva que el otro caso de t+8.

6.4 Comparación y discusión de los resultados

Una vez realizadas todas las predicciones se refleja que para t+1 hay una clara superioridad en cuanto a mejores resultados del gamboost frente al blackboost, la primera predicción gamboost tiene un menor MAPE inferior incluso al 4% lo que resulta un modelo competitivo debido a este bajo porcentaje de error absoluto, el modelo blackboost por su parte no tiene un mal valor de MAPE, pero en la gráfica comparativa si se refleja que no cumple tan bien como el modelo predictivo anterior. De hecho, se obtiene nuestra predicción óptima para gamboost en t+1 entre todas las predicciones realizadas en el proyecto Se cumple lo esperado en cuanto a que estos son los menores errores que vamos a encontrar en cualquiera de las predicciones por ser los que predicen a corto plazo.

Siguiendo el mismo orden que en el epígrafe anterior nos encontramos la predicción de t+3, donde se igualan en cierto modo el rendimiento de blackboost y gamboost ya que muestran valores de MAPE muy similares y si observamos las gráficas de sus predicciones no se adecúan tan bien a la realidad tan bien a la realidad como como la predicción gamboost en t+1 pero sí que se corrige la mala gráfica que se observaba en blackboost+1.

Esto quiere decir que a pesar de que al principio el algoritmo de la función blackboost no se adecuaba del todo bien, a medida que ampliamos el horizonte de la predicción va mejorando su rendimiento. En el lado contrario, gamboost en t+3 se mantiene con un rendimiento correcto y sigue siendo la mejor opción para la predicción de precios de energía eléctrica. Se observa en ambas predicciones, un error cercano al 8 o 9%, es decir, se dobla el MAPE respecto a las predicciones anteriores, como es lógico pues estas predicciones tienen un horizonte mayor.

Continuamos con la predicción a t+8 en laque de nuevo vemos que vuelve a aumentar el MAPE con valores del 10% y 11%, lo que nos revela que a pesar de aumentar mucho más el tiempo desde la segunda predicción a la tercera que de la primera a la segunda, el MAPE no tiene un aumento tan significativo. Pasa exactamente lo mismo con el valor de correlación entre las predicciones y los valores reales del precio de la energía eléctrica Ambas gráficas no se corresponden mucho con la realidad debido al notable aumento en el tiempo de la predicción, aun así, se confirma que el blackboost mejora para predicciones que n son inmediatamente después del valor real, es decir, horizontes medios y lejanos.

Predicción	MAPE
Gamboost t+1	3.59%
Blackboost t+1	4.4%
Gamboost t+3	8.54%
Blackboost t+3	9.92%
Gamboost t+8	10.55%
Blackboost t+8	11.38%

Tabla 8: Tabla comparativa de los errores MAPE en las predicciones

Como balance global, se podría decir que el gamboost funciona mejor que el blackboost, pero que a medida que avanzamos en el tiempo blackboost va mejorando su rendimiento. El boosting realizado en gamboost es mejor debido a que sus base learners son additive models y otorgan mejor rendimiento que los árboles de regresión que tenemos en blackboost, debido a estos árboles es lógico que se mejore con el tiempo puesto que para t+1 tenemos menor ramificación que a medida que vamos pasando el tiempo.

Por último, también tenemos que indicar la importancia de las variables, durante las predicciones se va analizando que variables afectan más a nuestra variable que queremos predecir, es decir, al precio. Se ha seleccionado el gráfico para la predicción que más se adapta a una buena predicción que es la predicción realizada para t+1 con función gamboost. Como era de esperar, la variable que más afecta es el precio de la energía en el instante anterior, con casi un 0,8 del peso, seguido de la demanda que se sitúa en torno al 0,1 y el restante 0,1 se lo reparten entre las temperaturas y la energía generada sin emisiones de CO2, pero no lo hacen a partes iguales si no que en tercer lugar se coloca esta energía generada y luego las temperaturas que apenas tienen peso en el proceso, destaca la temperatura media por encima del resto de temperaturas. A continuación, tienen una gráfica de lo que se acaba de lo que acabamos de comentar

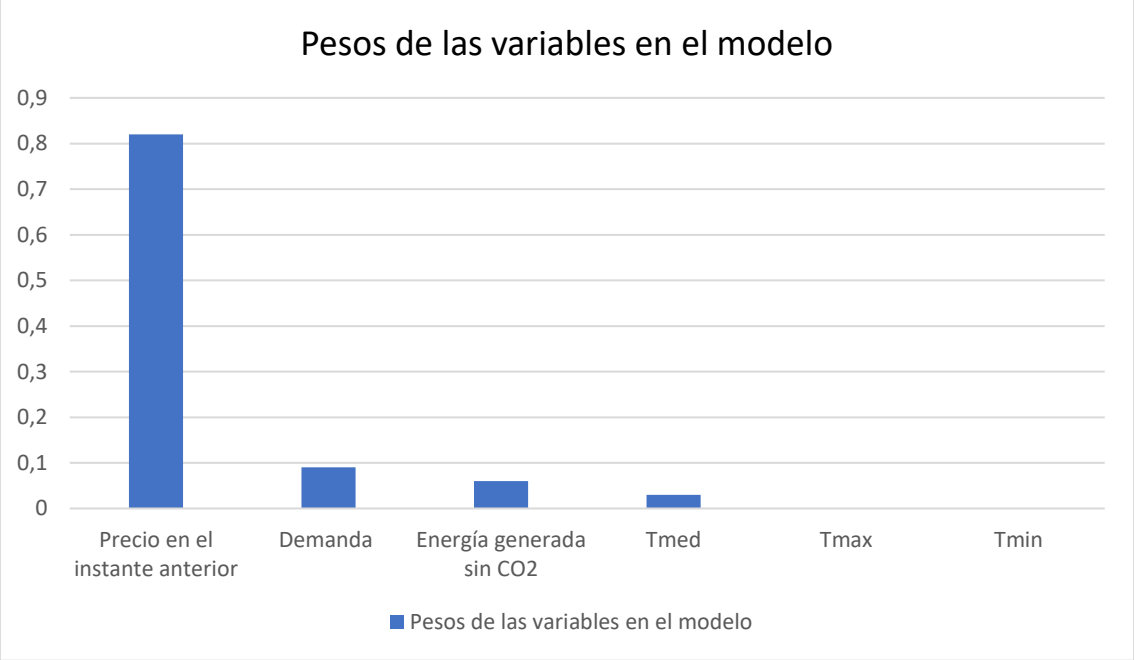

Figura 30: Pesos de las variables en el modelo de predicción gamboost en t+1

7. Conclusiones

Durante este Trabajo de Fin de Grado, se han introducido y explicado los algoritmos Boosting y se ha ido viendo en las distintas predicciones como se adecúan al precio del mercado eléctrico.

Se ha intentado reflejar la gran dificultad que conlleva el trabajo continuo con un gran volumen de datos, es decir, en el mundo del Data Mining, tanto en términos de computación como a la hora de conseguir conclusiones fiables.

La complejidad de la realización del proyecto se basa en distintos pilares. Por un lado, considero que tanto este como cualquier otro tipo de trabajo de fin de carrera tiene gran parte de autoaprendizaje en el que tendremos que sumergirnos para ser capaces de desarrollar nuestras propias ideas, ve si llegan a puerto o no, toma de decisiones sobre si plasmar unos resultados u otros y manejo de ciertos programas como en mi caso puede ser Rstudio en un nivel más allá de lo estudiado propiamente en la carrera. No obstante, la ayuda del tutor es vital en cualquier trabajo de esta índole.

En mi proyecto considero que la metodología ha sido el grueso más difícil de obtener porque conlleva dominio de ciertos campos e incluso el propio entendimiento del boosting puede ser ambiguo en ocasiones. Aún así se han conseguido varias predicciones interesantes que nos revelan ciertos datos en el mundo del Machine Learning.

Las conclusiones más importantes a las que se ha llegado a lo largo de la realización del trabajo son:

- Las variables más significativas para cualquier tipo de algoritmo ya sea gamboost o blackboost son los precios en horas anteriores, la demanda y la energía generada sin CO_2, están ordenadas de más a menos relevante.
- El aumento del número de observaciones produce un aumento de la fiabilidad de la predicción independientemente de la técnica de Machine Learning que se haya utilizado.
- Para predicciones en corto plazo siempre va a funcionar mucho mejor una predicción tipo gamboost que una tipo blackboost
- En el medio y largo plazo aumenta la fiabilidad de la predicción tipo blackboost, pero no llega a ser tan competitivo como gamboost. Esto se debe principalmente a la forma de calcular sus base learners.
- A medida que comienza a ampliarse el horizonte de la predicción se ve como os modelos comienzan a tener más errores y resultar menos fiables.

Este es un trabajo realizado por un estudiante en el cual se ha intentado hacer lo mejor posible con el máximo rigor que me ha sido posible, aun así, se muestra que no es posible explicar toda la variabilidad que existe detrás del precio de la energía eléctrica ya que existen otras muchas variables y factores que se deberían tener en cuenta en modelos más avanzados. Además, los modelos obtenidos con la herramienta estadística utilizada buscan un modelado genérico por lo que nunca van a ser una adaptación perfecta. Si nos ceñimos a la realidad del mercado eléctrico; el precio de la energía eléctrica depende de muchos más factores y datos que los que se han presentado aquí en este proyecto y existen modelos mucho mejores y fiables en las predicciones.

Desde el punto de vista del mercado eléctrico, se observa que los resultados obtenidos en cuanto al peso de las variables, son lo esperado puesto que el precio en horas anteriores suele mandar ya que en las graficas de precios de energía eléctrica no se ven grandes fluctuaciones de una hora para otra si o que para ver este tipo de variación tiene que pasar varios meses. La importancia de la demanda también hay que remarcarla ya que como en la gran mayoría de productos de una economía capitalista, el hecho de que aumente o disminuya la demanda, tiene un efecto en el precio del producto que en nuestro caso es energía eléctrica.

8. Líneas futuras

Todo este tipo de técnicas de Machine Learning tanto boosting que es la que hemos utilizado como cualquier otra son bastante modernas por lo que viendo su utilidad e interpretabilidad frente a un gran volumen de datos se puede deducir que van a seguir evolucionando.

Una de las líneas futuras interesantes que es imposible realizar actualmente sería aplicar boosting para predecir el precio de la energía eléctrica con los datos que tengamos en 2023 o 2024 o años superiores, esto se debe a que actualmente en el país se ha cambiado la tarifa de la luz de forma que se ha aumentado notablemente el precio de la misma por lo que creo que esto puede generar cambios importantes en la demanda y sería muy interesante ver hasta que punto puede afectar a la predicción mediante el algoritmo boosting.

Se podría analizar dos modelos distintos, uno con los días de Lunes a Viernes y otro para Sábados y Domingos, bien es cierto que para ello se deberían recopilar daos de varios años para los días de fin de semana puesto que si no se nos quedaría un volumen de datos muy pobre que probablemente provoque que tengamos un modelo que no nos aporte ningún beneficio respecto al general; y como hemos mencionado anteriormente, a mayor volumen de datos, mejor fiabilidad tendrá nuestro modelo predictivo.

Otra posible mejora sería la búsqueda y análisis de nuevas variables que pudieran afectar a nuestra futura predicción de precio de la energía eléctrica y que optimizarán y ajustarían más nuestro modelo a la realidad.

Para terminar, también resultaría interesante aplicar estas técnicas a otro tipo de problemáticas como puede ser la de accidentabilidad en carreteras o emisión de contaminantes, áreas de estudio en la que está involucrado el Departamento de Estadística de la ETSII UPM.

9. Planificación temporal y presupuesto

9.1 Planificación temporal

La duración estimada de este proyecto de fin grado es casi 7 meses, se comenzó a principios de Enero tras haber seleccionado el tutor y el tema en el primer cuatrimestre del año y se terminó el 23 de Junio de 2021. A continuación, se explicará en varios subapartados como se ha desarrollado cada fase del proyecto y el tiempo dedicado a la misma.

9.1.1 Investigación y estudio del proyecto

La fase comenzó el 4 de enero de 2021 y finalizó el 1 de Febrero de 2021. En esta fase se realizó una investigación profunda sobre artículos, libros, informaciones acerca del Machine Learning, más concretamente acerca del Boosting que es el método de análisis utilizado en el proyecto.

Se trató de leer e informarse en la máxima medida de lo posible para comprender el uso del software R, a pesar de haber estudiado este software en alguna asignatura a lo largo del grado, he tenido que profundizar mucho más en este ámbito para la correcta realización del proyecto.

Una vez que teníamos la suficiente información del Boosting y de R procedemos a entrar en la siguiente fase.

9.1.2 Base de datos: creación y tratamiento

Como ya hemos explicado en alguna ocasión a lo largo del proyecto, se necesita una contundente base de datos para lograr la obtención de unos resultados con cierta coherencia.

En este período nos hemos familiarizado con la utilización de las bases de datos, que han sido almacenadas en el programa Excel. Se ha estudiado ciertas variables que podrían afectar de un modo u otro a la predicción del precio de la energía eléctrica como pueden ser variables de temperatura, demanda, energía renovable producida o el precio en el instante anterior t-1.

La base de datos en sí consta de 8760 filas y 7 columnas donde cada columna corresponde a valores de una cierta variable y las 8760 se corresponden con el valor de esa variable desde la primera hora de 2019 hasta la última hora del mismo año.

Esta fase transcurre entre el 2 de Febrero de 2021 y el 14 de Marzo de 2021.

9.1.3 Estudio y generación del modelo

En esta fase tendremos que mezclar lo aprendido en la primera fase con nuestra base de datos generada en la segunda fase. Se realiza mediante el programa Rstudio en el cual se introducen distintas variables para conseguir crear un modelo en el que se obtengan unos resultados de predicción coherentes.

Es la fase más complicada y larga del proyecto debido a la dificultad de la implementación de un modelo boosting que trabaja con algoritmos que requieren cierto conocimiento para trabajar con ellos.

La duración es desde el 15 de Marzo de 2021 hasta el 10 de Mayo de 2021

9.1.4 Análisis y estudio del modelo

La duración se extiende del 11 de Mayo hasta el 4 de Junio de 2021.

Se observa el correcto funcionamiento del modelo realizado y a su vez se estudian los resultados que son plasmados en gráficas para hacer más fácil la interpretación de los mismos.

Por otro lado, se pueden sacar conclusiones del proyecto realizando un análisis de la predicción conseguida, y se plasman en el proyecto.

9.1.5 Repaso general y corrección de errores

Se realiza un último repaso general para supervisar y revisar todos los pasos seguidos en el Trabajo de Fin de Grado y que no haya errores en los mismos, ya sean errores del propio proyecto o errores de ortografía que se hayan producido involuntariamente en este documento.

Comienza el 5 de Junio y termina el 22 de Junio de 2021

9.2 Presupuesto

Este proyecto ha podido ser realizado gracias a una serie de recursos que he podido utilizar. En este apartado procedo a realizar un cálculo estimativo sobre le coste de este trabajo. Tendremos ciertos costes materiales y otros inmateriales.

Ordenador portátil: el proyecto se ha realizado en un ordenador portátil Lenovo S145-15AST cuyo coste ha sido de 350€, la vida útil de portátiles de semejantes características es aproximadamente de 5 años y nuestro proyecto se ha desarrollado en aproximadamente 7 meses, esto quiere decir que tenemos un coste de amortización de aproximadamente 40€.

Software estadístico R: existe una versión de pago del programa que hemos utilizado, pero no hemos utilizado esa versión, sino que hemos realizado el proyecto con la versión gratuita por lo que en este apartado tenemos un coste total de 0€

Licencia Microsoft Office 2019: su coste anual es de 129,99€ y hemos utilizado durante 7 meses el programa Excel para la base de datos y el Word para redactar nuestro proyecto, su coste son casi 76€

Impresión del proyecto: se ha realizado una impresión a color de la totalidad del proyecto lo que ha costado unos 25€

Encuadernación del proyecto: se ha realizado una encuadernación especial reforzada la cual ha tenido un precio de 40€

Tiempo dedicado por el estudiante: actualmente estoy cursando prácticas extracurriculares en una empresa con un sueldo de becario que es lo que actualmente tengo puesto que todavía no tengo el grado por lo que aproximadamente mi sueldo son 5€/hora con lo que eso es lo que costará la mano de obra del estudiante. Si el proyecto dura 7 meses y en cada mes tenemos 30 días, estimamos 210 días y personalmente suelo realizar períodos de trabajo de entre 2h y 3h, por lo que tomaremos 2,5h por día. De este modo salen unas 525 horas dedicadas por parte del estudiante a5€/hora, nos da un total de 2625€

Tiempo dedicado por el tutor: el coste de una hora de trabajo del tutor se estima en 30€. Entre llamadas, tutorías, correos etc. José Manuel Mira McWilliams, mi tutor, ha empleado más o menos unas 20 horas, por lo que nos queda un coste de total en este apartado de 600€.

Gastos varios: en este apartado se incluye la bibliografía, material de oficina etc. supone un coste total de 30€

Tras la explicación de cada apartado en esta sección de presupuesto obtenemos la suma total de todos los costes. El precio total del proyecto asciende a 3436€

Concepto	**Cantidad**	**Precio unitario**	**Precio total**
Ordenador portátil	1	40	40
Software estadístico R	1	0	0
Licencias Microsoft Office 2019	1	76	76
Impresión y encuadernación	1	65	65
Horas del estudiante	525	5	2625
Horas del tutor	20	30	600
Gastos varios	1	30	30

TOTAL	**3436€**

Tabla 9: Tabla del presupuesto total del proyecto

Anexos

Anexo 1: Código de R

A continuación, se mostrará un código de Rstudio resumido en el cual se muestra cómo se ha realizado el análisis y estudio de este proyecto. El código que aparece está escrito para una predicción gamboost, para blackboost simplemente tendríamos que cambiar la función y se desarrollaría otro tipo de predicción.

```r
#Descargamos la librería que necesitamos, tras haber instalado su correspondiente paquete e introducimos la base de datos en Rstudio#
rm(list=ls())
library(mboost)
library(readxl)
library(psych)
dataproy <- read_excel("C:/Users/juanc/OneDrive/Escritorio/dataproy.xlsx")
attach(dataproy)
#Se selecciona de donde a donde va el conjunto de prueba y el conjunto de entrenamiento#
observaciones=nrow(dataproy)
seqobsentrenamiento=seq(from=1,to=8159,by=1)
seqobstest=seq(from=8160,to=observaciones,by=1)
#Se definen el conjunto de prueba y el conjunto de entrenamiento#
entrenamiento=dataproy[seqobsentrenamiento,]
test=dataproy[seqobstest,]
#Aplicación del método Boosting para predecir la variable precio en función de el precio anterior, la energía generada sin CO2,demanda y temperaturas#
modelo=gamboost(precio~precio8efore+energiasinCO2+Tmax+Tmin+Tmed+demanda,baselearner ="bbs", data = entrenamiento, control=boost_control(mstop=200))
print(modelo)
#Buscamos el numero de iteraciones óptima para evitar el sobreajuste#
set.seed(123)
mod=cvrisk(modelo)
mod
plot(mod)
mstop(mod)
modelo[mstop(mod)]
#Realizamos la predicción#
pred<-predict(modelo,test)
#Calculamos el MAPE y obtenemos gráficas#
mape=100*abs((precio[8160:observaciones]-pred)/precio[8160:observaciones])
summary(mape)
boxplot(mape)
prec=precio[8160:observaciones]
corre=data.frame(prec,pred)
plot(corre)
pairs.panels(corre)
#Obtenemos la importancia de las variables en el modelo#
plot(varimp(modelo))
```

Figura 31: Captura de pantalla del script resumido para la realización de boosting

Anexo 2: Índice de figuras

Anexo 3: Índice de tablas

Bibliografía

Alayo J. Ub.edu. (2021). Visitado 24 de Enero 2021, http://www.ub.edu/geocrit/Electricidad-y-transformacion-de-la-vida-urbana/JoanAlayo.pdf

Aldo R. (2021). Introducción al Machine Learning con BigML. Slideshare.net. (2021). 30 de Enero 2021, https://www.slideshare.net/AldoRamiro/introduccin-al-machine-learning-con-bigml.

Base de datos meteorológica. Consulta de Datos de temperatura. Datosclima.es. (2019). Visitado 1 de Marzo de 2021, https://datosclima.es/Aemethistorico/Meteostation.php.

Boosting - Wikipedia, la enciclopedia libre. Es.wikipedia.org. (2021). Visitado 17 de Enero de 2021, https://es.wikipedia.org/wiki/Boosting.

Boosting Algorithms Explained. Medium. (2021) Visitado 5 de Abril de 2021, https://towardsdatascience.com/boosting-algorithms-explained-d38f56ef3f30

Boosting algorithms: Regularization, predictions and model fitting Web.stanford.edu. (2021). Visitado 13 de Mayo 2021, https://web.stanford.edu/~hastie/Papers/buehlmann.pdf.

Briega, R. (2021). Boosting en Machine Learning con Python. Relopezbriega.github.io. Visitado 23 de Marzo, https://relopezbriega.github.io/blog/2017/06/10/boosting-en-machine-learning-con-python/.

Datanalytics.com. (2021). Visitado 6 de Junio de 2021, https://datanalytics.com/libro_r/_main.pdf.

Demanda y generación de energía | ESIOS electricidad · datos · transparencia. Esios.ree.es. (2021). Visitado 18 de Febrero https://www.esios.ree.es/es.

Generalized boosted models Cran.r-project.org. (2021). Visitado 3 de Mayo de 2021, https://cran.r-project.org/web/packages/gbm/vignettes/gbm.pdf.

Kearns, Michael & Valiant Leslie (1989) ’Cryptographic limitations on learning Boolean formulae and finite automata’ Visitado 23 de Enero de 2021

mboost function - RDocumentation. Rdocumentation.org. (2021). Visitado 20 de Abril 2021, from https://www.rdocumentation.org/packages/mboost/versions/2.9-5/topics/mboost.

Mercados y precios | ESIOS electricidad · datos · transparencia. Esios.ree.es. (2019). Visitado 8 de Abril de 2021, from https://www.esios.ree.es/es/mercados-y-precios?date=08-04-2019.

Precio horario del mercado diario | OMIE. Omie.es. (2019). Visitado 27 de Febrero 2021 https://www.omie.es/es/market-results/daily/daily-market/daily-hourly-price?scope=daily&date=2019-01-01.

Reyes Fajardo, Laura Marcela (2021). Repository.udistrital.edu.co. (2021) Visitado 14 de Febrero de 2021 https://repository.udistrital.edu.co/bitstream/handle/11349/5232/ReyesFajardoLauraMarcela2017.pdf?sequence=1&isAllowed=y.

SpainML. Spainml.com. (2021). Visitado el 3 de Marzo 2021, https://spainml.com/blog/como-funciona-gradient-boosting/.

What is Boosting and AdaBoost in Machine Learning? Knowledgehut. Knowledge.com (2021) Visitado 18 de Marzo de 2021 https://www.knowledgehut.com/blog/data-science/boosting-and-adaboost-in-machine-learning