CAPÍTULO II

Adelantando el consumo de las administraciones públicas: *big data* a través del *BOE*

Carlos Cuerpo Caballero*
Teresa Morales Gómez-Luengo

Los avances en las técnicas de análisis de big data, junto con la creciente disponibilidad de grandes repositorios de datos, están permitiendo novedosas aplicaciones en el campo de la macroeconomía, en especial en la previsión de los principales agregados macroeconómicos. Este artículo presenta una posible aplicación, a través de la previsión del consumo público para España utilizando la plataforma de contratación del sector público. El análisis recoge la práctica totalidad de la actividad contractual del sector público, cubriendo más de 1.185.337 licitaciones, realizadas desde 2018 por más de 15.000 órganos de contratación distintos.

Palabras clave: big data, consumo público, machine learning, contratación pública.

* Los autores agradecen las ideas y los comentarios de Israel Arroyo Martínez y Raquel Losada Muñoz. Las ideas reflejadas en el artículo corresponden a los autores y no representan a las instituciones en las que trabajan.

1. BASE DE DATOS: PLATAFORMA DE CONTRATACIÓN DEL SECTOR PÚBLICO

1.1. Definición y características generales

Los datos utilizados en este capítulo proceden de la plataforma de contratación del sector público, mantenida y puesta a disposición de los usuarios de forma gratuita por el Ministerio de Hacienda[1]. La plataforma representa la puerta de entrada a la actividad contractual del sector público, facilitando información relevante sobre las licitaciones públicas, desde su convocatoria, hasta los resultados de las mismas.

Su primera gran ventaja, además de ser pública, reside en su completitud, puesto que recoge las licitaciones del sector público entendido en sentido amplio[2], incluyendo, entre otras:

- La Administración General del Estado, las administraciones de las comunidades autónomas y las entidades que integran la Administración Local.
- Las entidades gestoras y los servicios comunes de la Seguridad Social.
- Los organismos autónomos, las entidades públicas empresariales, las universidades públicas y las agencias estatales, entre otras entidades de derecho público vinculadas a o dependientes del sector público.
- Las sociedades mercantiles en cuyo capital social la participación pública, directa o indirecta, sea superior al 50 %.
- Las mutuas de accidentes de trabajo y enfermedades profesionales de la Seguridad Social.

Las licitaciones registradas recopilan información sobre al menos seis dimensiones de interés. En primer lugar, sobre la duración del contrato, incluyendo fechas de publicación, adjudicación y número de meses de duración. En segundo lugar, sobre la tipología de los contratos, destacando las categorías de servicios, suministros, obras y concesión de obras, entre otras. En tercer lugar, información geográfica y sectorial sobre el órgano contratante. En cuarto lugar, información sobre el procedimiento mediante el cual se lleva a cabo la contratación, ya sea Abierto, Restringido, Negociado con y sin publicidad, Acuerdo Marco, Simplificado, etc. En quinto lugar, el tipo de tramitación, separando la Ordinaria de la Urgente y de la de Emergencia. En sexto lugar, información sobre la competencia entre licitadores, destacando el número de concurrentes y el precio de licitación y adjudicación final, que permite analizar las bajas o ahorros en el precio debidas al proceso de competencia entre empresas licitadoras.

[1] Puede accederse a través de su página web: https://contrataciondelestado.es/wps/portal

[2] Tal y como recoge el artículo 3.1 del Real Decreto Legislativo 3/2011, Texto Refundido de la Ley de Contratos del Sector Público.

La base de datos original incluye información desde 2012 si bien se ha seleccionado un periodo de análisis de 2018 a 2020[3]. Con esta muestra se dispone de información sobre 1.185.337 licitaciones, realizadas por más de 15.000 contratantes distintos. Para llegar a este número se ha realizado un proceso de depuración previa de la base de datos original, eliminando las entradas duplicadas[4], los valores extremos y errores detectados en los precios de adjudicación , duración y número de licitadores.

1.2. Análisis descriptivo

1.2.1. Tipo de contrato

La plataforma clasifica los contratos en función de su objeto en las siguientes categorías: (a) Servicios, (b) Suministro, (c) Gestión de servicios públicos, (d) Concesión de servicios, (e) Obras, (f) Concesión de obras públicas, (g) Privado, (h) Administrativo especial, (i) Colaboración entre sector público y privado, (j) Patrimonial y (k) Otros. Para facilitar la interpretación, las once categorías se agregan en tres más genéricas: (i) Consumo público, que incluye las cuatro primeras, (ii) Inversión pública, que incluye Obras y Concesión de obras públicas; y la categoría (iii) Otros, que incluye al resto.

Tal y como refleja la figura 1, los contratos cuyo objeto es el consumo público son mayoritarios, tanto en número como en presupuesto, seguido de los contratos de inversión y, en último lugar y con una presencia residual, del resto. En concreto, los contratos de consumo representan en torno al 90 % del total de los contratos, por un 8 % de los contratos de inversión. Sin embargo, la cuantía promedio de los contratos de inversión es mayor y esto hace que, en términos del presupuesto total anual (panel b) su peso relativo supere el 20 %.

1.2.2. Tipo de procedimiento

La plataforma clasifica los contratos también en función del procedimiento utilizado para su adjudicación. Se incluyen las siguientes categorías: (a) Abierto, (b) Abierto simplificado, (c) Restringido, (d) Negociado sin publicidad, (e) Negociado con publicidad, (f) Diálogo competitivo, (g) Normas internas, (h) Acuerdo marco, (i) Concurso de proyectos, (j) Asoc. para la innovación, (k) Sist. dinámico adquisición, (l) Licitación con negociación, (m) Otros, y (n) Menores.

---

[3] Los contratos menores están únicamente disponibles desde 2018. Para el resto de contratos se seleccionan licitaciones publicadas desde 2017 pero adjudicadas como pronto en 2018.

[4] Para cada licitación se encuentran varias entradas pues la información va actualizándose conforme se suceden las distintas fases en el proceso, desde la publicación original hasta la adjudicación. El número de entradas previo a esta depuración era de más de 2.150.000.

Como norma general, y con arreglo a la Ley 9/2017 de Contratos del Sector Público, los contratos que celebren las administraciones públicas se adjudicarán utilizando (a) el procedimiento abierto o (c) el procedimiento restringido.

En el caso de los procedimientos abiertos, toda empresa interesada podrá presentar una propuesta, sin posibilidad de negociación de los términos del contrato. Existe la posibilidad de recurrir a la modalidad (b) abierta simplificada en los contratos de obras, suministro y servicios cuando su valor estimado no supere unas cantidades determinadas legalmente y no haya criterios de adjudicación evaluables mediante juicio de valor que superen el 25 % del total[5]. Esta modalidad permite realizar las adjudicaciones en el plazo de un mes desde la licitación, agilizando el proceso.

En cuanto a los procedimientos (c) restringidos, solo podrán presentar proposiciones aquellas empresas que sean seleccionadas por el órgano de contratación en atención a su solvencia. Tal y como especifica la Ley 9/2017, este procedimiento está particularmente indicado para servicios intelectuales de especial complejidad.

El resto de procedimientos solo pueden darse en los casos previstos en la ley. Entre ellos, destacan los contratos (n) menores y los negociados sin (d) y con (e) publicidad. Los menores permiten realizar adjudicaciones de forma directa, sin previa licitación, hasta un importe máximo (15.000 euros en caso de servicios y suministros). En cuanto a los negociados, su principal característica radica en que las condiciones del contrato se negocian previamente con uno o varios licitadores. La ley habilita esta modalidad para los contratos de obras, suministros, servicios, concesión de obras y concesión de servicios cuando se cumplan ciertos supuestos tasados en los artículos 167 y 168, como que la prestación incluya soluciones innovadoras o así lo exija la complejidad de la prestación, entre otras. Por defecto, los procedimientos negociados se harán con publicidad de la licitación, excepto en los casos tasados en el artículo 168 de la Ley 9/2017, como por ejemplo aquellos procesos en los que no se hubiera presentado ninguna oferta[6].

Los menores son los contratos más comunes representando el 62 % del total, tal y como puede verse en la figura 2, panel a. Esta característica se mantiene tanto para los contratos de consumo público como para los contratos de inversión, representando el 66 % y el 40 %, respectivamente. La segunda categoría más importante la constituyen los contratos abiertos (18 %) y abierto simplificado (10 %). Si añadimos el procedimiento negociado sin publicidad, que representa el 5 % del total, quedarían recogidos el 95 % de los contratos en estas cuatro categorías.

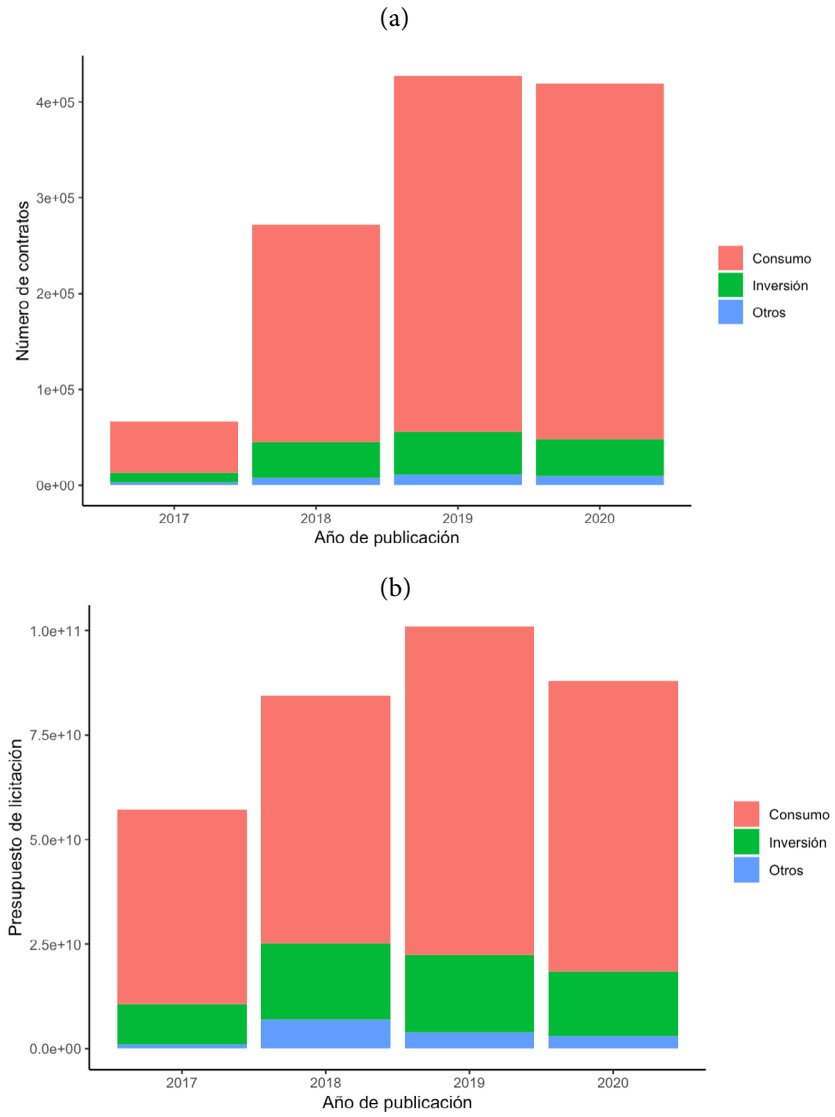
Pese a ser los contratos más comunes, los menores apenas suponen el 1 % del presupuesto movilizado, como refleja el panel b de la figura 2. Los contratos abiertos reinan en esta clasificación, alcanzando el 69 % de la cantidad total movilizada. Este liderazgo se mantiene

[5] Salvo para los contratos de prestaciones de carácter intelectual, en que su ponderación no podrá superar el cuarenta y cinco por ciento del total.

[6] Para más detalles sobre los distintos procedimientos, ver Royo (2018).

Figura 1.

Número y presupuesto de los contratos, por tipología y por año de publicación

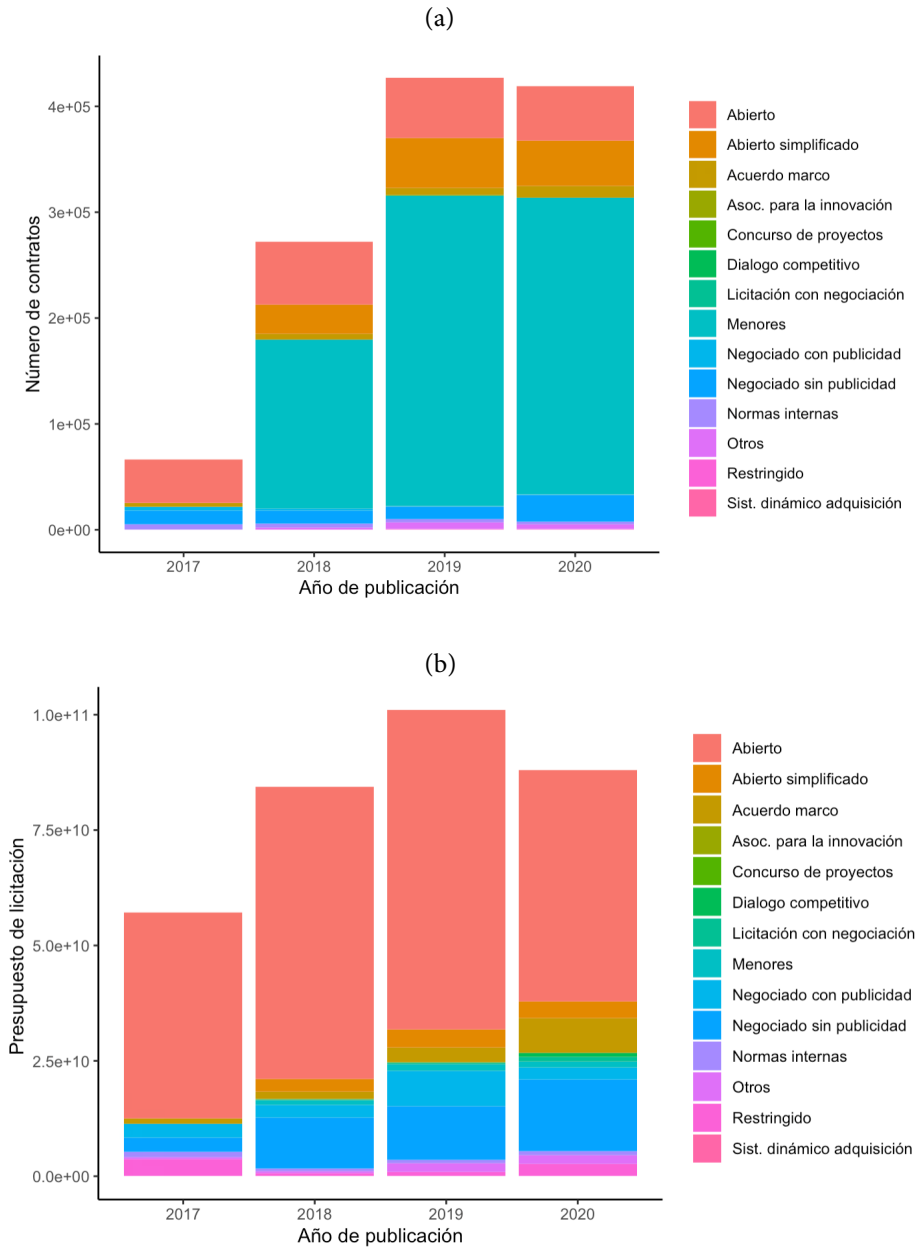


Fuente: Elaboración propia y plataforma de contratación del sector público.

tanto para los contratos de consumo como para los de inversión. Los contratos negociados se encuentran en segundo lugar de importancia en esta clasificación, con un 17 % del total movilizado. Cabe destacar la importancia de las licitaciones abiertas simplificadas en el caso de los contratos de inversión, llegando a representar un 12 % de los mismos.

Figura 2.

Número y presupuesto de los contratos, por procedimiento y por año de publicación



Fuente: Elaboración propia y plataforma de contratación del sector público.

1.2.3. Tipo de tramitación

La Ley 9/2017 prevé, más allá de (a) la tramitación ordinaria, dos mecanismos para acelerar las licitaciones: la tramitación de urgencia y (c) la tramitación de emergencia.

La tramitación de urgencia, regulada por el artículo 119 de la Ley 9/2017 supone la reducción de los plazos de licitación, adjudicación y formalización a la mitad. Solo será de aplicación a aquellos contratos cuya celebración responda a una necesidad inaplazable o que sea preciso adjudicar de forma acelerada por causa de interés público.

La tramitación de emergencia, regulada en el artículo 120, se limita excepcionalmente para aquellas situaciones en que las administraciones públicas tengan que actuar de manera inmediata como consecuencia de acontecimientos catastróficos, de situaciones que supongan grave peligro o de necesidades que afecten a la defensa nacional. Ante estas circunstancias de extrema gravedad, el órgano de contratación podrá ordenar la ejecución sin obligación de tramitación del expediente.

Por norma general, el procedimiento ordinario es el más común superando el 70 % del total en términos de número de contratos y el 60 % en términos de presupuesto movilizado. La tramitación de emergencia se ha limitado a menos del 1 % del total en años ante-

Figura 3.

Número y presupuesto de los contratos, por tramitación y por año de publicación

(a)

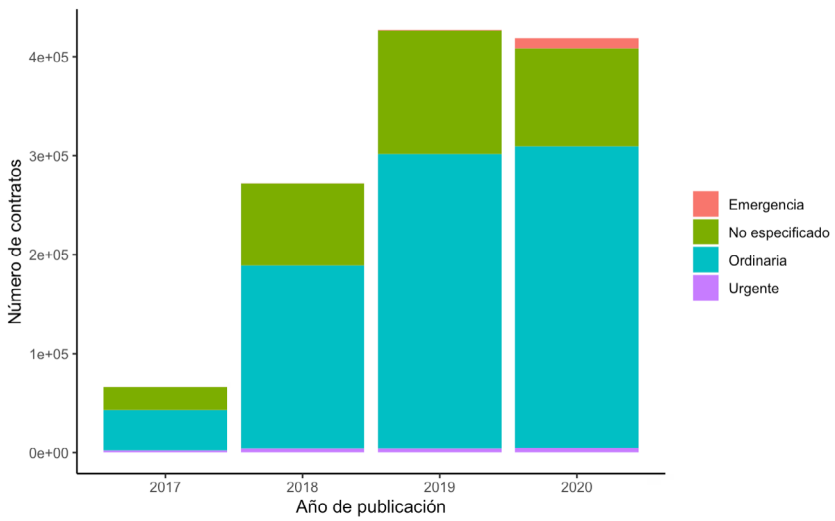
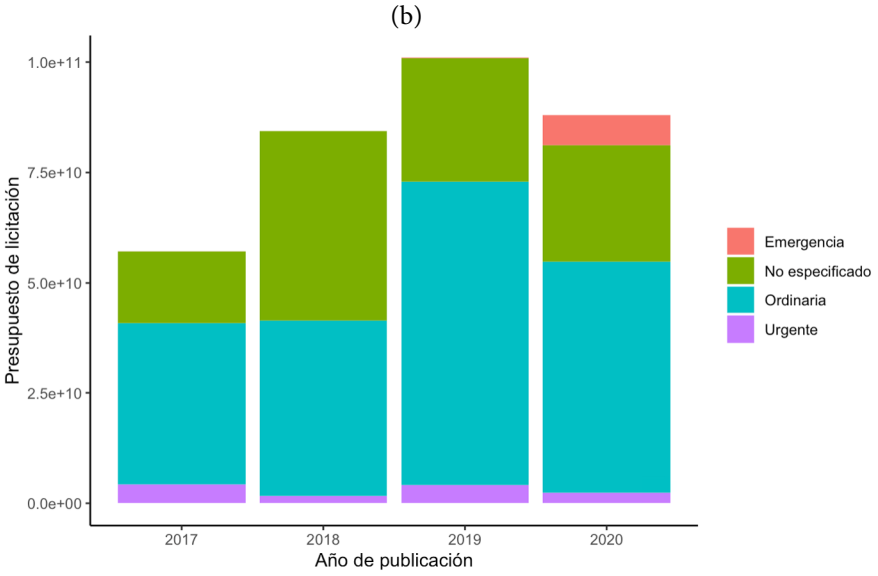


Figura 3. *(continuación)*

Número y presupuesto de los contratos, por tramitación y por año de publicación



Fuente: Elaboración propia y plataforma de contratación del sector público.

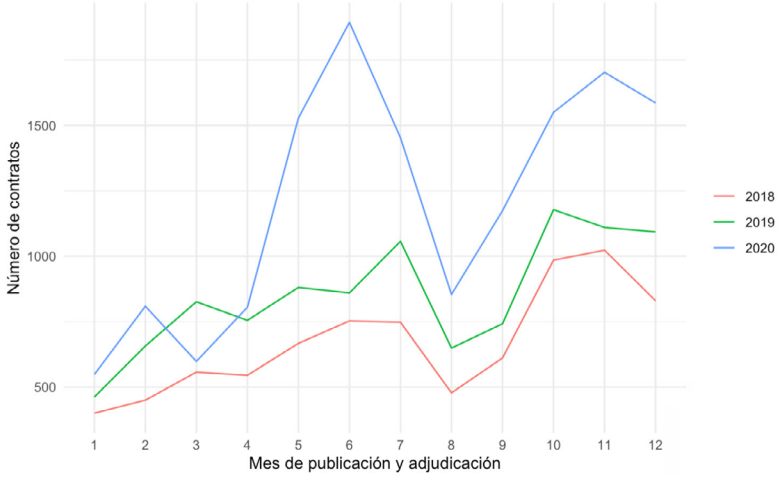
riores a 2020. Sin embargo, la declaración del estado de alarma como consecuencia de la COVID-19 ha supuesto un incremento sustancial del peso de los contratos de emergencia en 2020, pasando a representar el 3 % del total de contratos de consumo y el 9 % del presupuesto de los mismos.

Esta circunstancia puede verse también en la figura 4, que recoge el número de contratos adjudicados y publicados en el mismo mes en 2018, 2019 y 2020. Puede observarse como en el año 2020 en los meses de abril, mayo y junio, existe un repunte en los contratos publicados y adjudicados dentro del mismo mes (principalmente a través de la utilización de los procedimientos de emergencia).

La importancia de la COVID-19 en la contratación de emergencia puede verse a través de un análisis de la frecuencia de las palabras más repetidas en el objeto del contrato. Si miramos la totalidad de los contratos de consumo en 2020 (panel a, figura 5), destaca la importancia de los contratos de mantenimiento, o de equipamiento informático, tanto *hardware* como *software*. Si por el contrario nos centramos en los procedimientos de emergencia, aparece la motivación de la pandemia con los suministros relacionados con el material de protección frente al virus como principal objeto.

Figura 4.

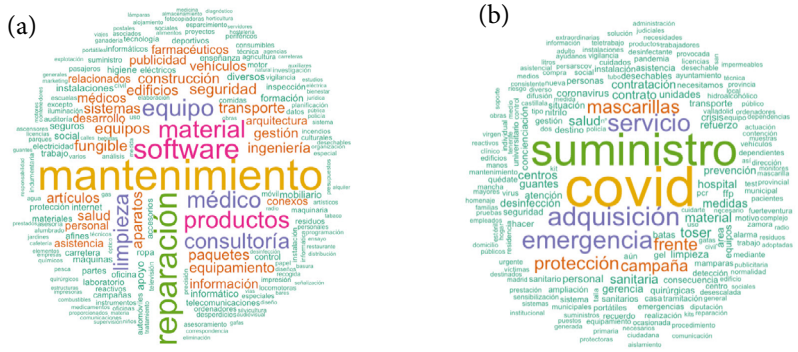
Evolución mensual del número de contratos adjudicados en el mismo mes
(Tiempo de adjudicación = 0)



Fuente: Elaboración propia y plataforma de contratación del sector público.

Figura 5.

Frecuencia de palabras en contratos de consumo y contratos de consumo de emergencia en 2020



Fuente: Elaboración propia y plataforma de contratación del sector público.

1.2.4. Concurrencia, baja de precio en la adjudicación y duración

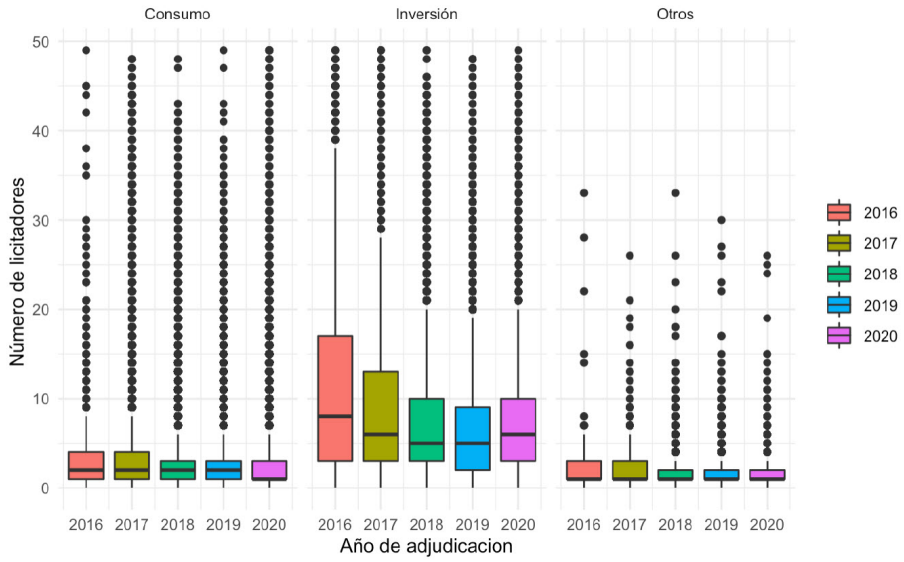
La competencia en los procesos de licitación permite optimizar la eficiencia en el uso de los recursos públicos, disminuyendo el riesgo de prácticas lesivas como la colusión entre operadores o la corrupción[7].

[7] Ver CNMC (2011) para una visión más profunda de la importancia de la competencia como principio inspirador de la normativa en materia de contratación pública.

La figura 6 refleja la situación en términos de concurrencia por tipo de contrato, excluyendo los contratos menores, menos relevantes en términos de cuantía presupuestada. Destacan los contratos de inversión con un promedio de más de siete licitadores, muy por encima de los de consumo, que en su mayoría se deciden entre tres empresas.

Figura 6.

Número de licitadores por año y tipo de contrato
(Contratos no menores)



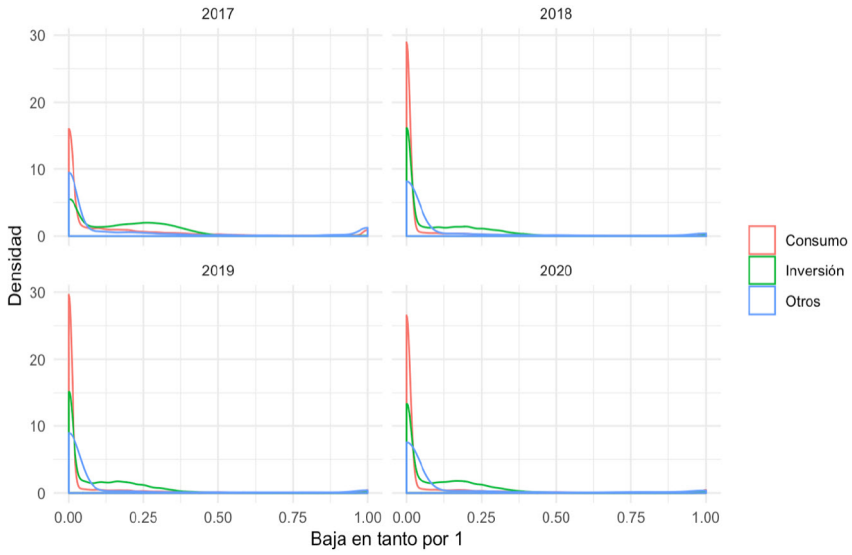
Fuente: Elaboración propia y plataforma de contratación del sector público.

La baja de precio desde el precio presupuestado o licitado hasta el precio final de adjudicación, depende también del tipo de contrato, como puede verse en la figura 7 y, por lo tanto, de la concurrencia en la licitación. Los contratos de inversión, de mayor concurrencia, presentan también una mayor baja porcentual en el precio, del 10 %, frente a los de consumo, que apenas llegan al 5 % en promedio.

La duración de los contratos sigue ciertos patrones comunes a lo largo de los años. En primer lugar, se observa una concentración de los contratos de obras por cortos períodos de tiempo (menores de seis meses). Los contratos de consumo se concentran en periodos más largos y tienen una periodicidad marcada por la frecuencia anual, fecha clave también para la renovación de los mismos (ver figura 8). Finalmente, se observan algunos cambios en 2020 en comparación con el resto de años, como un aumento de la densidad en los contratos menores de tres meses, sobre todo para las de consumo.

Figura 7.

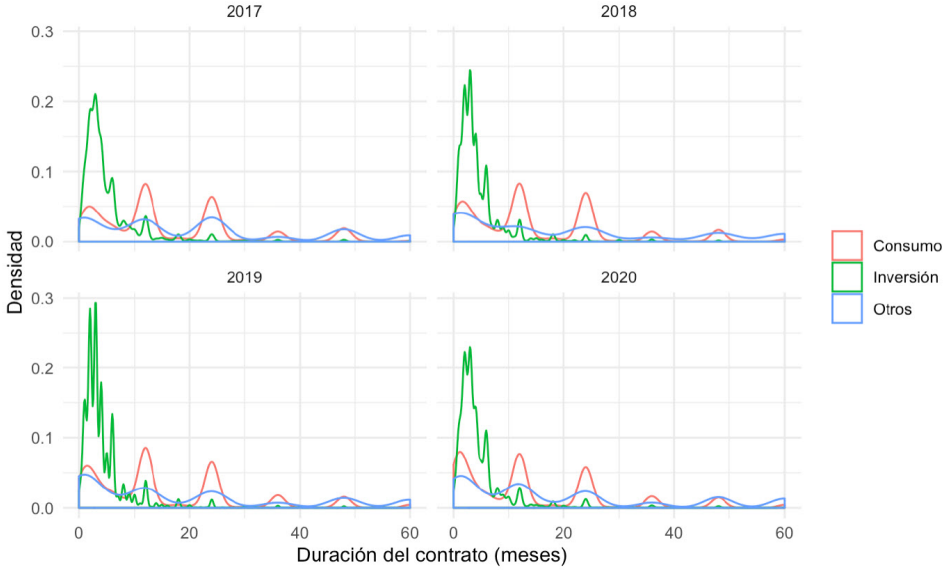
Densidad de las bajas de todos los contratos



Fuente: Elaboración propia y plataforma de contratación del sector público.

Figura 8.

Densidad de la duración de contratos no menores



Fuente: Elaboración propia y plataforma de contratación del sector público.

2. EJERCICIO DE PREDICCIÓN DEL CONSUMO PÚBLICO

2.1. Planteamiento del ejercicio

El objetivo final de este ejercicio es utilizar la plataforma de contratación descrita en el apartado anterior para realizar una previsión del consumo público en tiempo real.

El consumo público es uno de los componentes más importantes del producto interior bruto (PIB), suponiendo en torno a un 20 % del mismo en promedio. Su provisión se encuentra altamente descentralizada, correspondiendo un 60 % a las comunidades autónomas (educación y salud) y apenas un 20 % a la Administración Central. La remuneración de asalariados supone en torno al 60 % del total, seguido de los consumos intermedios (un 25 %), y de otros elementos como las ventas, el consumo de capital fijo, otros impuestos sobre la producción y las transferencias sociales en especie (TSE), adquiridas en el mercado. Esta última categoría puede llegar a presentar hasta el 15 % del total[8]. La disparidad de conceptos se traduce en una elevada heterogeneidad en sus determinantes principales, que pueden ir desde factores tan estructurales como la población para el componente salarial, hasta variables de coyuntura o el *stance* de política fiscal.

Estas características dificultan la previsión del consumo público con modelos estructurales, particularmente de sus componentes más pequeños y volátiles, como las TSE adquiridas en el mercado.

Para sortear estas dificultades y poder realizar una previsión del gasto público en tiempo real se plantea la posibilidad de proyectar el gasto comprometido utilizando los datos disponibles en la Base de Datos del Sector Público en el momento de la publicación de cada licitación. Esto permitiría hacer una previsión mensual que incorpore los contratos licitados hasta ese mes, independientemente de que se hayan adjudicado o no. Como hemos visto, la información disponible acerca de cada licitación es muy amplia, incluyendo información sobre el tipo de procedimiento, el objeto del contrato, el sector, la geografía, etcétera.

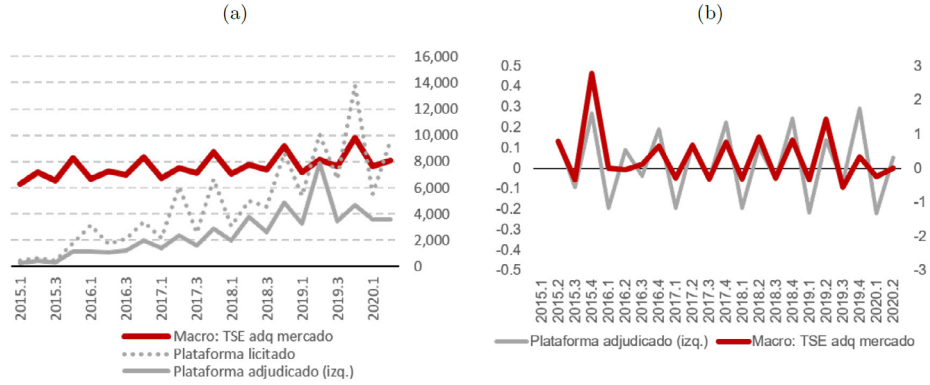
Sin embargo, si lo que se busca es proyectar el gasto en el momento de la licitación, existen dos datos fundamentales con los que no contamos: el precio de adjudicación, que aparecerá registrado en la mayoría de los casos en un momento posterior y el tiempo que tardará la licitación en adjudicarse, que puede ser muy variable en función de diversas circunstancias. Estas dos variables son fundamentales a la hora de conseguir un buen ajuste entre las licitaciones y el consumo público. En efecto, tal y como se observa en la figura 9 (panel a), la evolución de las TSE y la información de los contratos de consumo licitados sigue un perfil marcadamente distinto. Sin embargo, una vez que incorporamos la información final sobre la adjudicación, la dinámica de ambas series se vuelve altamente correlacionada, como puede verse en la evolución de sus tasas de crecimiento en el panel de la derecha.

[8] Para un análisis completo y pormenorizado del consumo público y sus componentes ver Losada (2017).

Figura 9.

Contratos de consumo como *proxy* de las transferencias sociales en especie adquiridas en mercado

(M€, panel a) y (% variación trimestral, panel b)



Fuentes: INE, elaboración propia y plataforma de contratación del sector público.

2.2. Metodología

Lo que planteamos en esta fase de nuestro estudio es analizar si dichos factores de incertidumbre (baja porcentual resultante en la adjudicación y tiempo de adjudicación) pueden ser estimados utilizando modelos de aprendizaje automático.

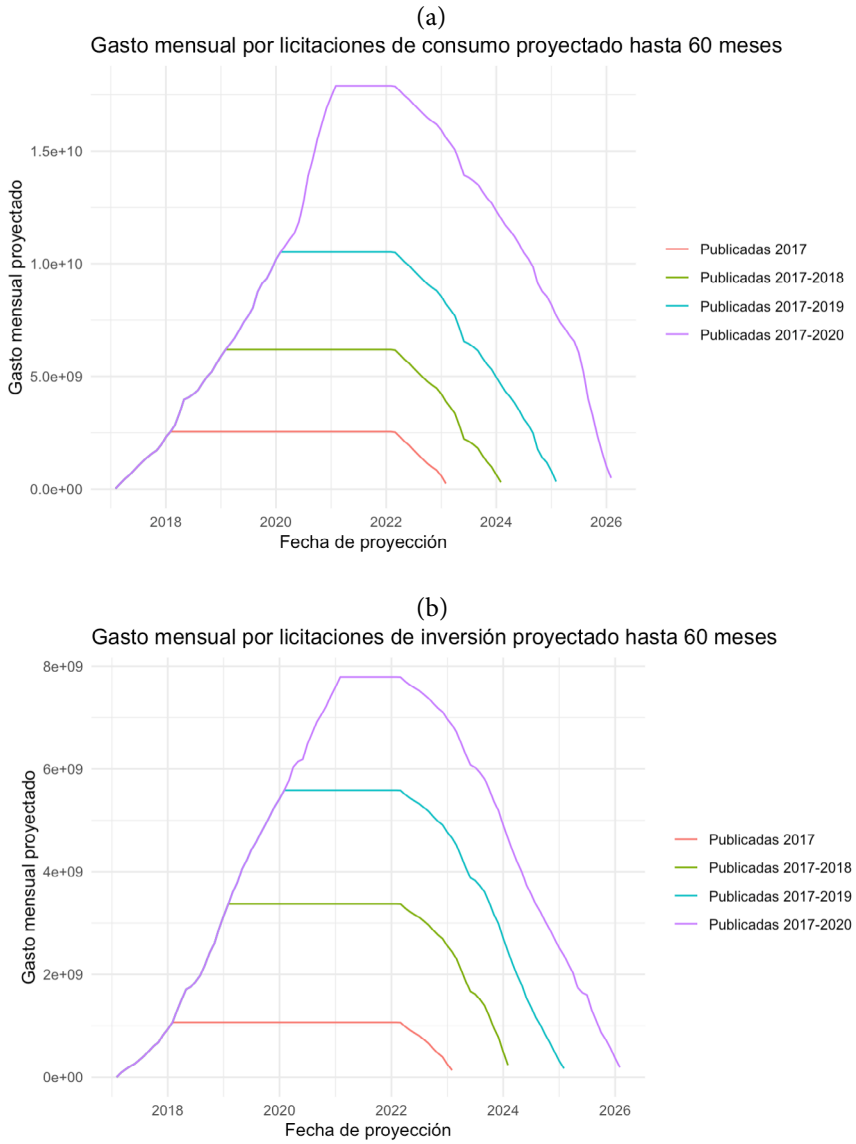
A continuación explicamos en primer lugar, los resultados que obtenemos cuando hacemos un ejercicio de proyección mensual de los montantes licitados que no tiene en cuenta los elementos de incertidumbre relacionados con la adjudicación. En segundo lugar explicamos como ir más allá, utilizando modelos de aprendizaje automático que nos permitan predecir el tiempo de adjudicación y la baja porcentual prevista cuando se publica cada licitación. En un ejercicio posterior habría que incorporar estas predicciones a la proyección de gasto público junto con otros elementos que pudieran incorporarse con datos como los de subvenciones, remuneraciones de los empleados públicos, etcétera.

2.2.1. Ejercicio de proyección del gasto

En la figura 10 se muestra una proyección del gasto mes a mes realizada para los contratos clasificados como consumo (panel a) y los contratos clasificados como inversión (panel b). Como vemos en los gráficos se realiza un ejercicio de proyección diferente, que incluye la información disponible desde 2017 hasta el año indicado. Es decir, la línea más alta muestra la proyección del gasto de todos los contratos publicados entre 2017 y 2020, la siguiente más baja los publicados entre 2017 y 2019 y así sucesivamente.

Figura 10.

Proyección del consumo con la información disponible sobre presupuesto de licitación y duración del contrato



Fuente: Elaboración propia y plataforma de contratación del sector público.

Para realizar esta proyección se calcula el gasto mensual asociado a cada contrato y se proyecta ese gasto mensual durante el plazo de duración del mismo. De este modo se está asu-

miendo que todos los contratos se ejecutarían mes a mes y que se empezaría a ejecutar el gasto comprometido al mes siguiente de la publicación de la licitación. Además, por simplicidad, se fija un límite temporal para la proyección de cada contrato de 60 meses.

Como podemos observar, la información disponible en la base de datos de contratación del sector público nos permite identificar una cuantía considerable de gasto ya comprometido en ejercicios futuros, tan lejanos como 2026. Además, con la información disponible a día de hoy, es posible prever un incremento del gasto mensual ya comprometido que alcanza su volumen máximo a finales de 2021 y no comienza a descender hasta finales de 2022. Esto es así tanto en contratos de consumo como en contratos de inversión.

Por otro lado, conforme se replica este ejercicio año a año se observan patrones comunes que parecen permitir identificar anomalías en la evolución del gasto. Así, la distancia entre las curvas de licitaciones de consumo publicadas hasta 2017, 2018 y 2019 es homogénea y contrasta con un salto que parece producirse en la curva de consumo de los contratos publicados hasta 2020. Este salto podría indicar un incremento considerable del gasto en consumo probablemente asociado, como hemos visto, a los contratos asociados a la crisis de la COVID-19.

2.2.2. Modelos de machine learning

Las proyecciones del montante licitado, sobre el que se ha practicado una reserva de crédito son sin duda informativas. Sin embargo, como decíamos, resultaría interesante poder contar con una predicción más cercana a la realidad del gasto que se va a ejecutar utilizando predicciones del tiempo de adjudicación y del montante de adjudicación (o baja presentada por la empresa adjudicataria). Para ello planteamos utilizar modelos de aprendizaje automático que, gracias al uso de una gran volumetría de datos y variables, pueden ser más precisos que los modelos de regresión estadística tradicionales.

En esta primera fase de nuestro análisis hemos utilizado el modelo de *Random Forest regression*. Se trata de un modelo relativamente sencillo en comparación a las redes neuronales o *deep learning,* pero requiere de menos datos y ha mostrado ser igual de efectivo en algunos casos, gracias a la agregación de muchos modelos sencillos entrenados con submuestras. En particular el Random Forest lleva a cabo una agregación de árboles de decisión que utilizan muestras distintas extraidas de la muestra principal por muestreo aleatorio con reemplazo (*bagging*)[9].

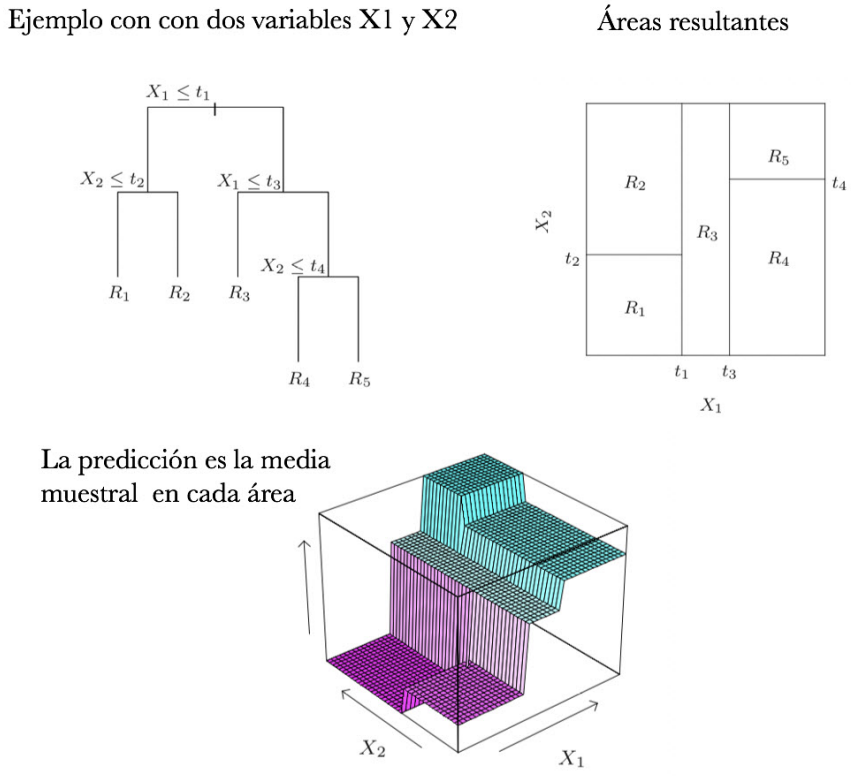
Los árboles de decisión son modelos simples que permiten ir clasificando la muestra en grupos basados en reglas de decisión binarias. Así, cada árbol de decisión elige la variable que permita dividir dos grupos de la muestra con la máxima diferencia en la variable objetivo

[9] Ver Breiman (1996) para una descripción detallada de por qué la técnica de *bagging* permite obtener buenos resultados.

posible. Para cada uno de esos grupos repite el ejercicio y así sucesivamente, hasta contar con una predicción basada en la media de una submuestra lo suficientemente pequeña de características similares. A continuación se presenta un ejemplo gráfico de cómo funciona un árbol de decisión para un caso de regresión:

Figura 11.

Representación de un árbol de decisión



Fuente: Hastie,Tibshirani y Friedman (2001).

Si se permite que el algoritmo persiga una predicción perfecta, el árbol de decisión extendería sus hojas hasta el límite de manera que, en cada extremo, solamente habría una observación y la predicción sería igual a la variable objetivo. Sin embargo, el árbol así entrenado no funcionaría para predecir muestras que no hubiera observado (muestra de validación o cualquier muestra que se encuentre cuando entre en producción). Para evitar este efecto, que se denomina *overfitting* se deben introducir elementos de regularización que permiten simplificar el algoritmo resultante para que sea generalizable a otras observaciones. Entre las técnicas que permiten regularizar los árboles de decisión se encuentran el denominado *tree*

prunning o limitación de la profundidad del árbol, la fijación de un número mínimo de muestras previo a la división en un nodo o la limitación del número de variables utilizadas para la división de las ramas (ver James *et al.,* 2014).

Como decíamos, el algoritmo de Random Forest, lo que hace es entrenar un número elevado de árboles de decisión. Como resultado de dicho entrenamiento con muestras diferentes, cada árbol de decisión habrá utilizado variables distintas y tendrá un número diferente de nodos de decisión. La predicción final es la media de las predicciones de todos los árboles que, por agregación, ha demostrado tener una precisión mucho más alta que un árbol de decisión simple individual.

Para construir un modelo de machine learning que permita realizar una predicción se utilizan técnicas de entrenamiento de los modelos con el objetivo de elegir los valores de sus parámetros e hiperparámetros. En el caso del Random Forest las elecciones que deben extraerse de este ejericio son la secuencia de variables y valores límite que permitirá ir dividiendo la muestra para cada árbol, el número de árboles que utilizará el algoritmo y las variables de regularización a las que nos hemos referido como la profundidad máxima del árbol o el tamaño mínimo de muestra de entrenamiento para la división.

Para entrenar el modelo se selecciona de forma aleatoria una muestra de entrenamiento con el 70 % de las observaciones y una muestra de validación con el 30 % restante. Adicionalmente, para asegurar que se seleccionan los parámetros e hiperparámetros que permiten generalizar mejor el modelo se utiliza la técnica de validación cruzada (*cross validation*). Esta técnica consiste en dividir la muestra de entrenamiento en *n* partes y entrenar el modelo en *n* iteraciones, dejando fuera cada vez una parte distinta. Después de entrenar cada modelo posible con las *n* iteraciones se elige el modelo que menor error cuadrático medio tiene en promedio. En nuestro caso la muestra de entrenamiento se ha dividido en tres partes por lo que cada uno de los modelos que se han probado se han entrenado tres veces, dejando fuera del proceso de entrenamiento cada vez una parte distinta, con la que se calcula el error cuadrático medio de predicción. Para cada modelo que se quiere considerar, se obtienen tres errores cuadráticos medios estimados, correspondientes a las tres iteraciones, con los que se calcula el promedio global para determinar qué modelo es el más adecuado.

Por último, otra técnica habitual en machine learning es considerar un amplio abanico de valores posibles para los hiperparámetros con el objetivo de elegir el modelo que mejor predicción haga cuando se enfrente a la realidad desconocida. Para ello, hemos utilizado una búsqueda aleatoria, que permite entrenar un número razonable de modelos al tiempo que se contemplan miles de combinaciones posibles ya que se van eligiendo hiperparámetros lo suficientemente diferentes de entre las opciones iniciales que se planteen.

2.3. Resultados

Se han construido dos modelos para predecir, por un lado, el tiempo de adjudicación previsto y por otro, la baja porcentual aplicada a una licitación. Las variables utilizadas son

el presupuesto de licitación, la duración del contrato y variables ficticias para cada categoría de código CPV, tipo de contrato (clasificación detallada y agregada), Comunidad Autónoma, tipo de procedimiento, tipo de tramitación y mes de publicación. Considerando cada variable ficticia por separado el número total de variables utilizadas es de 5.565.

El tamaño total de la muestra que excluye observaciones con valores omitidos en alguna de esas variables o en la variable objetivo es 75.403 para el modelo de tiempos de adjudicación y 194.941 para el modelo de bajas porcentuales. Se utiliza validación cruzada en bloques de tres y múltiples combinaciones de hiperparámetros.

Los modelos seleccionados a través del procedimiento de validación cruzada con búsqueda aleatoria de hiperparámetros tienen las siguientes características[10]:

a) Modelo de predicción ganador para tiempos de adjudicación

Número de árboles de decisión	100
Número mínimo de observaciones en un nodo para poder dividirlo	446
Número máximo de observaciones utilizado en cada árbol	25.000
Número máximo de variables consideradas para la división	1.000
Máxima profundidad de cada árbol (niveles)	70

b) Modelo de predicción ganador para baja porcentual

Número de árboles de decisión	2.000
Número mínimo de observaciones en un nodo para poder dividirlo	8.000
Número máximo de observaciones utilizado en cada árbol	72.500
Número máximo de variables consideradas para la división	4.000
Máxima profundidad de cada árbol (niveles)	200

Para evaluar estos modelos elegidos en el proceso de entrenamiento se utiliza el 30 % de la muestra reservada como muestra de validación. Por tanto, se trata de datos que el modelo nunca ha procesado por lo que permiten aproximar el error que se obtendría si se utilizara el modelo con datos nuevos, no disponibles hasta la fecha. Los resultados muestran que el modelo construido para la predicción del tiempo de adjudicación es muy útil, mientras que el modelo de predicción de las bajas porcentuales presenta una peor calidad.

El modelo de predicción de tiempos de adjudicción permite predecir cuándo será adjudicado un contrato con un error de en torno a un mes. En particular, la raíz del error cuadrático medio (*RMSE,* por sus siglas en inglés) es de 1,23 meses, muy por debajo del error que obtendríamos con un modelo de regresión lineal con regularización de *Ridge* (2,52

[10] Para la construcción de los modelos se ha utilizado la librería Scikit-learn de Python (ver Pedregosa *et al.,* 2011).

meses)[11]. El error cuadrático medio pondera más las predicciones muy alejadas de la realidad entre las que es posible que se encuentren valores atípicos. Para evitar que los atípicos pesen demasiado, se suele considerar también el error absoluto medio (*MAE,* por sus siglas en inglés) que da el mismo peso a todas las predicciones. El error absoluto medio para el modelo de predicción de tiempos de adjudicación es de tan solo 0,5 meses, frente a un valor de 0,7 para la regresión lineal con regularización de Ridge. Por lo tanto, en general se considera que este modelo de predicción de tiempos de adjudicación es bastante satisfactorio.

En cambio, las predicciones resultantes del modelo construido para predecir la baja porcentual asociada al precio de adjudicación son malas predicciones de la realidad. En particular, si bien el modelo exige mucho más tiempo de entrenamiento y mucha más capacidad computacional que el modelo de regresión lineal, se acaban obteniendo resultados muy similares. Ambos modelos (Random Forest y regresión lineal con regularización de Ridge) predicen la baja presentada por la empresa ganadora del concurso con una raíz del error cuadrático medio de 23 puntos porcentuales y un error absoluto medio de en torno a 15 puntos porcentuales, resultados que, como se observa en la figura 13, son bastante pobres.

Modelo	*Tiempo de adjudicación (meses)*		*Baja porcentual (tanto por 1)*	
Random Forest Regressor	RMSE = 1,23	MAE= 0,50	RMSE = 0,23	MAE = 0,16
Regresión lineal	RMSE = 2,52	MAE = 0,7	RMSE = 0,23	MAE = 0,15

En los siguientes gráficos se puede ver una muestra aleatoria de 100 observaciones de las muestras de validación ordenadas de menor a mayor por su valor real.

En el modelo de tiempos de adjudicación (a), una gran parte de las predicciones mostradas en rojo acierta el tiempo de adjudicación real observado, a diferencia de lo que sucede con el modelo lineal (b), en donde se observan discrepancias generalizadas entre lo que predice el modelo y lo que sucede en la realidad. Además el modelo de aprendizaje automático parece detectar saltos que podrían estarse produciendo debido a la existencia de distintos tipos de procedimiento, diferentes tipos de tramitaciones y otros aspectos contemplados en las variables incluidas. En cambio, el modelo de regresión lineal no parece ser capaz de detectar este tipo de patrón. Cuando se analiza qué variables están influyendo más en esta predicción aparece, en primer lugar, el tipo de contrato y, en segundo lugar, el código CPV, es decir el sector.

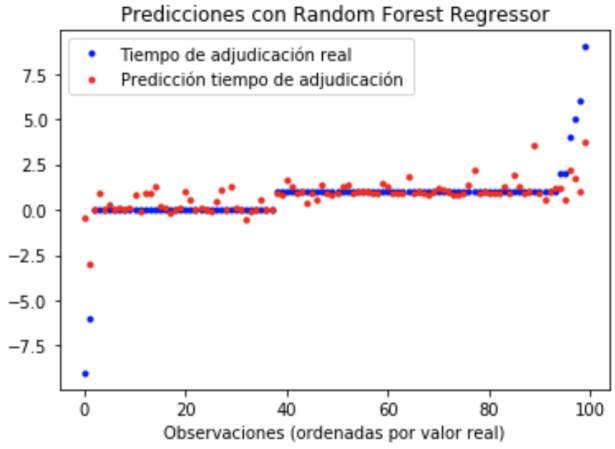
En la figura 13 se observa que el modelo de Random Forest de predicción de bajas porcentuales del precio de adjudicación es poco eficiente. Si bien parece comportarse algo mejor que un modelo de predicción lineal en los valores más reducidos, en términos generales el modelo no supone una mejora particularmente relevante respecto de un modelo lineal, tal y como confirman los valores de error cuadrático medio y error absoluto medio mostrados anteriormente. En su conjunto ninguno de los dos modelos utilizados para predecir la baja

[11] Se utiliza una regularización de Ridge para contar con un modelo de referencia básico que incluya, aunque sea con una influencia pequeña, todas las variables contempladas en el modelo de Random Forest

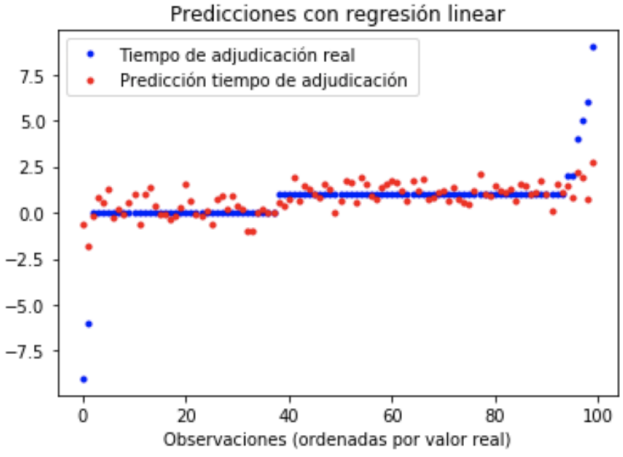
Figura 12.

Tiempos de adjudicación. Predicciones y observaciones de la muestra de validación
(Para 100 observaciones aleatorias)

(a)



(b)



Fuente: Elaboración propia.

porcentual parece útil, por lo que sería necesario buscar modelos alternativos que permitan mejorar estos resultados.

En el caso de la baja porcentual, las variables más relevantes son en primer lugar el sector (CPV) y en segundo lugar, el tipo de contrato. Una de las limitaciones importantes de este análisis es la imposibilidad de incluir el número de empresas concurrentes, variable que

probablemente sea muy significativa a la hora de determinar la baja ganadora. No obstante, no consideramos la posibilidad de incluir esta variable ya que lo que se pretende es predecir el gasto derivado de cada licitación en el momento en que esta se publica es decir, cuando no se cuenta con información sobre el número de empresas que concurrirán. Una alternativa que se podría explorar es construir primero un modelo que permita estimar el número de empresas concurrentes y, posteriormente, utilizar esa predicción como *input* de un segundo modelo que prediga la baja porcentual ganadora. Adicionalmente, se deberían explorar otro tipo de modelos más sofisticados e incorporar información no estructurada, como el texto descriptivo del objeto del contrato que se encuentra también disponible en esta base de datos.

Conviene apuntar que, al margen de los resultados obtenidos, los modelos de aprendizaje automático tienen ciertas limitaciones que deben tenerse en cuenta a la hora de hacer predicciones futuras sobre gasto público. En primer lugar, dado que estos modelos son muy intensivos en el uso de la evidencia empírica, es posible que se queden desactualizados a menudo. Así, si las relaciones entre variables cambian con el tiempo por circunstancias como cambios en la gestión pública de los contratos o situaciones excepcionales como la de la COVID-19, la representatividad de los datos utilizados para construir el modelo acaba siendo limitada. Por lo tanto, los modelos deben reentrenarse con cierta frecuencia. Por otro lado, estos modelos también sufren del problema estadístico tradicional de datos omitidos. Si las observaciones no consideradas por no tener toda la información no son aleatorias, podría haber un problema de generalización del modelo con datos futuros.

Figura 13.

Baja porcentual. Predicciones y observaciones de la muestra de validación
(Para 100 observaciones aleatorias)

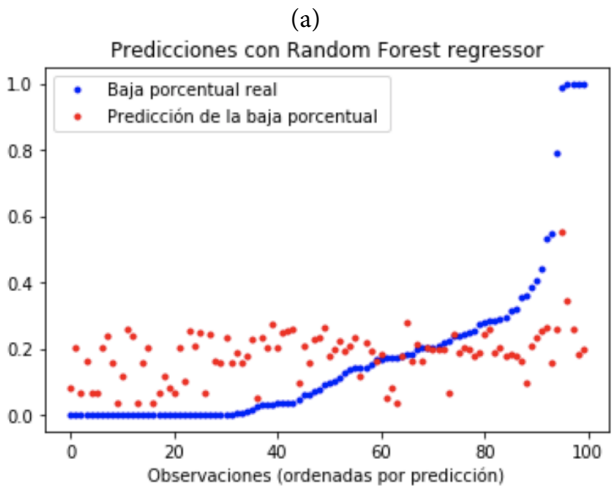
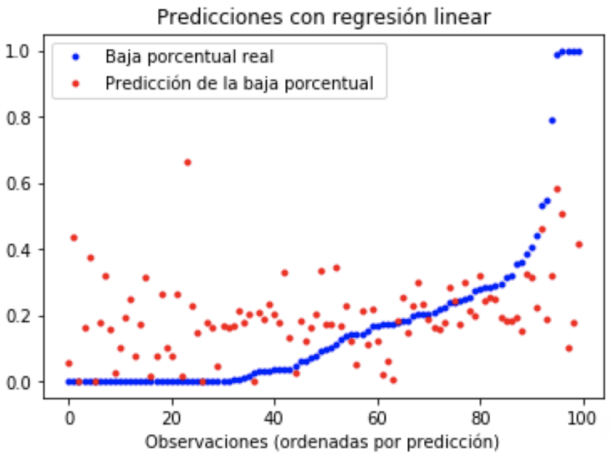


Figura 13. *(continuación)*

Baja porcentual. Predicciones y observaciones de la muestra de validación
(Para 100 observaciones aleatorias)

(a)



Fuente: Elaboración propia.

3. CONCLUSIÓN Y APLICACIONES ADICIONALES

La Base de Datos de Contratación del Sector Público tiene un amplio potencial para el análisis del gasto público tanto desde un punto de vista macroeconómico como desde un punto de vista microeconómico y de gestión pública. El análisis desarrollado ha mostrado el potencial de utilizar toda esa información para predecir el gasto público en tiempo real. No obstante, el potencial de uso de estos datos para otras finalidades es enorme. En particular consideramos que la información disponible permitiría por ejemplo detectar cuellos de botella en la absorción de fondos europeos, ayudar a mejorar la concurrencia de empresas en licitaciones públicas o detectar irregularidades en la contratación basadas en conexiones anómalas entre empresas y órganos de contratación o el fraccionamiento de contratos.

En lo que respecta a la posibilidad de contar con una previsión del gasto público actualizada con alta frecuencia, las avenidas abiertas en términos de investigación se centran en dos cuestiones fundamentales:

En primer lugar, mejorar los modelos que nos permitan predecir la baja porcentual de las licitaciones públicas. Para ello se recurrirá a modelos más sofisticados como modelos encadenados que permitan incorporar una previsión del número de empresas licitantes o redes neuronales. Sobre todo, se procurará explotar información adicional como la contenida en el objeto del contrato a través de técnicas de procesamiento del lenguaje natural.

En segundo lugar, se hará extensivo el análisis a otros conceptos de consumo e inversión públicas y a las subvenciones, utilizando toda la información pública disponible, incluida por ejemplo la Base de Datos Nacional de Subvenciones.

Una vez se desarrollen los necesarios modelos predictivos de los distintos factores que determinan el gasto público, se podrá contar con un indicador adelantado que agregue toda esta información y permita prever la evolución del gasto público con mayor precisión y anticipación.

Referencias

Breiman, L. (1996). Bagging predictors. *Machine Learning,* Volume 24, pp. 123–140.

CNMC (2011). *Guía sobre Contratación Pública y Competencia*. https://www.cnmc.es/file/123708/download

Hastie, T., Tibshirani, R. y Friedman, J. (2001). *The Elements of Statistical Learning*. New York Inc: Springer.

James, G., Witten, D., Hastie, T. y Tibshirani, R . (2014). *An Introduction to Statistical Learning: With Applications* in *R*. Springer Publishing Company.

Losada, R. (2017). ¿A qué nos referimos al hablar de consumo público? *AIReF, documentos de trabajo,* 2/2017.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M. y Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. *JMLR,* 12, pp. 2825-2830.

Royo, M. T. (2018). Los procedimientos de adjudicación de contratos públicos. *Monografías de la Revista Aragonesa de Administración Pública,* pp. 365-373