Trabajo Fin de Máster

Desarrollo de un Early Redemption Model

Olamar Benavente Fernández

Máster en Técnicas Estadísticas

Curso 2023-2024

Propuesta de Trabajo Fin de Máster

Título en galego: Desenvolvemento dun Early Redemption Model
Título en español: Desarrollo de un Early Redemption Model
English title: Development of an Early Redemption Model
Modalidad: Modalidad B
Autor/a: Olamar Benavente Fernández, Universidade de Vigo
Director/a: Javier Tarrío Saavedra, Universidade da Coruña; Salvador Naya Fernández, Universidade da Coruña
Tutor/a: Carlos Amor Montañés, ABANCA
Breve resumen del trabajo: Generación de un Early Redemption Model, para el control del riesgo de tipo de interés asociado a los depósitos a plazo, en un contexto económico de tipos elevados y alta demanda de este tipo de producto financiero. Para tal fin, se abordarán técnicas de explotación de bases de datos internas de la entidad, se construirán modelos y métricas, así como, test estadísticos para garantizar la bondad de los modelos. Todo ello siguiendo la normativa bancaria vigente, así como, papers sectoriales y académicos.

Don Javier Tarrío Saavedra, Profesor Titular de la Universidade da Coruña, don Salvador Naya Fernández, Catedrático de la Universidade da Coruña, don Carlos Amor Montañés, Especialista del área de Validación Interna de Modelos de ABANCA, informan que el Trabajo Fin de Máster titulado

Desarrollo de un Early Redemption Model

fue realizado bajo su dirección por doña Olamar Benavente Fernández para el Máster en Técnicas Estadísticas. Estimando que el trabajo está terminado, dan su conformidad para su presentación y defensa ante un tribunal.

En [Vigo], a 03 de [junio] de 2024.

El/la director/a:
Don/doña Javier Tarrío Saavedra

El/la director/a:
Don/doña Salvador Naya Fernández

El/la tutor/a:
Don/doña Carlos Amor Montañés

El/la autor/a:
Don/doña Olamar Benavente Fernández

Declaración responsable. Para dar cumplimiento a la Ley 3/2022, de 24 de febrero, de convivencia universitaria, referente al plagio en el Trabajo Fin de Máster (Artículo 11, Disposición 2978 del BOE núm. 48 de 2022), **el/la autor/a declara** que el Trabajo Fin de Máster presentado es un documento original en el que se han tenido en cuenta las siguientes consideraciones relativas al uso de material de apoyo desarrollado por otros/as autores/as:

- Todas las fuentes usadas para la elaboración de este trabajo han sido citadas convenientemente (libros, artículos, apuntes de profesorado, páginas web, programas,...)
- Cualquier contenido copiado o traducido textualmente se ha puesto entre comillas, citando su procedencia.
- Se ha hecho constar explícitamente cuando un capítulo, sección, demostración,...sea una adaptación casi literal de alguna fuente existente.

Y, acepta que, si se demostrara lo contrario, se le apliquen las medidas disciplinarias que correspondan.

Agradecimientos

Me gustaría expresar mi más profundo agradecimiento a todas las personas que han hecho posible que con esta memoria se cierre un bonito ciclo de aprendizaje y crecimiento, tanto profesional como personal. No ha sido un camino fácil, pero sí gratificante, aprendiendo de grandes profesionales que con tanto tesón y esfuerzo dedican su tiempo a compartir y enseñar todos sus conocimientos para hacer de nosotros grandes profesionales.

Quisiera agradecer profundamente a mis tutores Javier Tarrío Saavedra y Salvador Naya Fernández por su apoyo y paciencia, por compartir conmigo toda su sabiduría y darme las pautas e indicaciones necesarias para poder finalizar con éxito este proyecto. Ha sido muy gratificante poder compartir con vosotros este camino lleno de aprendizaje y poder absorber hasta el último grano de conocimiento que tan generosamente han puesto a mi disposición. Asimismo, quisiera agradecer a ABANCA por abrirme sus puertas, y más concretamente a mi tutor Carlos Amor que con tanta paciencia y dedicación me ha dedicado su tiempo para abrirme las puertas al equipo de trabajo y me ha ayudado tan minuciosamente en cada parte del proyecto, aportándome nuevos conocimientos y habilidades.

No quisiera terminar este apartado sin mencionar a mi familia y a mi pareja que han estado conmigo en todo mi proceso de resiliencia y transformación, que me ha apoyado en cada decisión que he tomado y que me han ayudado a hacer real aquella frase que dice así: "cuando una puerta se cierra, otra se abre", en definitiva, que me han ayudado a superar todos los obstáculos del camino sin importar el esfuerzo que haya supuesto para ellos con tal de no dejarme nunca caer. Gracias por vuestra paciencia, cariño, amor y apoyo diario, no hubiera llegado hasta aquí sin vosotros. Gracias también al equipo médico de neurología del Hospital La Paz (Doctor Tallón, Doctor Torres, Doctora Alba, mi enfermera Bea), que también han estado conmigo, cuidándome en cada paso de este camino desde que empecé el grado hasta el día de hoy que presento mi proyecto de fin de máster. Ellos me han procurado las medicaciones y cuidados necesarios para que haya podido cumplir mis metas y que se alegran conmigo de cada victoria que consigo.

Por último, quiero agradecer a todos los profesores del máster, a mi coordinadora del máster Leticia Lorenzo Picado, así como a todo el equipo de diversidad y apoyo a la discapacidad de la Universidade de Vigo que me habéis acompañado en todo el camino y habéis contribuido a realizar una inclusión real, en igualdad de oportunidades, garantizando que pudiera acceder a una educación de calidad y con totales garantías.

A todos y cada uno de vosotros, infinitas gracias.

Índice general

Resumen

Resumen en español

Ante la subida de tipos de interés acontecida en los últimos años implementada por el BCE como medida de contención de la inflación, muchos productos financieros se han visto directamente afectados, precisando desarrollar e incorporar mecanismos para monitorizar las posibles fluctuaciones y cambios que puedan producirse en estos. En este contexto, la entidad bancaria ABANCA precisó el desarrollo de un modelo para gestionar el riesgo de tipo de interés para sus productos financieros, más concretamente, para los depósitos a plazo, siendo éstos un tipo de producto altamente sensible a las variaciones del tipo de interés. La progresiva e imparable subida de estos, ha hecho que este producto haya cogido peso, derivando en una alta demanda por parte de los clientes, y por consiguiente la necesidad de monitorizar el riesgo que puede suponer para la entidad la cancelación anticipada de este tipo de productos. Para llevar a cabo este cometido se proponen varios modelos de regresión (lineal multivariante, aditivos generalizados, support vector machines y boosting), realizando una comparativa entre ellos mediante el cálculo de diferentes métricas de bondad de ajuste, para tratar de determinar cuál de todos se presta como el mejor modelo para predecir la tasa de cancelación anticipada de depósitos a plazo y dejar implementada en la entidad una herramienta capaz de solventar los posibles riesgos que puedan derivarse de los cambios en los tipos de interés a medio y largo plazo.

English abstract

Due to the rise in interest rates in recent years implemented by the ECB as a measure to contain inflation, many financial products have been directly affected, necessitating the development and incorporation of mechanisms to monitor possible fluctuations and changes that may occur. In this context, the financial entity ABANCA required the development of a model to manage interest rate risk for its financial products, more specifically, for term deposits, as these are a type of product highly sensitive to interest rate movements. The progressive and unstoppable rise in these rates has made this product more significant, resulting in high demand from customers, and consequently the need to monitor the risk that the early cancellation of these products may pose to ABANCA. To achieve this goal, several regression models (multivariate linear, generalized additive, support vector machines, and boosting) are proposed, comparing them by calculating different goodness-of-fit metrics, to try to determine which one is the best model for predicting the early cancellation rate of term deposits and implementing a tool in ABANCA capable of addressing the potential risks that may arise from changes in interest rates in the medium and long term.

Capítulo 1

Introducción

1.1. Motivación y contexto

Los depósitos a plazo son instrumentos financieros por los cuales cualquier persona física o jurídica, puede depositar una cantidad determinada de dinero a un plazo fijo estipulado en el contrato por el cual recibirá a la finalización del periodo de tiempo correspondiente, una remuneración al tipo de interés fijado en el momento de contratación. Es esencial tener en cuenta que la retirada anticipada de fondos es una opción disponible para el cliente, aunque puede implicar una penalización siendo la magnitud de esta penalización diferente según las características específicas del contrato establecido entre el cliente y la entidad financiera.

Estos depósitos representan una modalidad de ahorro atractiva para aquellos que buscan obtener una rentabilidad sin asumir riesgos significativos. Para las entidades bancarias también son instrumentos atractivos dado que durante el tiempo de vida del depósito estas pueden utilizarlos para generar otras operaciones, como por ejemplo, operaciones de préstamos a terceros, proporcionando a la entidad una fuente segura de financiación.

Tras un largo periodo marcado por un entorno macroeconómico de tipos de interés anómalamente reducidos, en el año 2022 el BCE inició un cambio en su política monetaria en respuesta al aumento de la inflación de los últimos dos años. Con el aumento de los tipos de interés del BCE, productos como los depósitos a plazo han recuperado el atractivo perdido en los últimos años (Walsh, 2022, Benigno et al., 2023). Parte de la subida de los tipos de interés se ha trasladado a los clientes en productos como los depósitos a plazo haciéndolos un producto más atractivo para rentabilizar sus ahorros.

En este contexto, y teniendo en cuenta que los depósitos a plazo se consideran un pasivo para las entidades bancarias ya que representan una obligación que estas entidades tienen con sus clientes, ABANCA se enfrenta a la necesidad de implementar metodologías para gestionar el riesgo de las cancelaciones anticipadas en los depósitos a plazo y que ayuden a una gestión adecuada del balance de la entidad en un entorno cambiante de tipos de interés. Estas retiradas, al constituir un riesgo significativo para la entidad, han llevado al banco a buscar estrategias proactivas para gestionar de manera eficaz dicho riesgo.

Es por esta razón que se propone la implementación de los denominados "Early Redemption Models" o "Modelos de Cancelación Anticipada" (Maggi et al., 2017, Wilson, 2022, Bissiri et al., 2014). Estos modelos tienen como finalidad proyectar la tasa de cancelación anticipada de los depósitos a plazo en diferentes escenarios de tipos de interés. En el caso que nos ocupa, estos modelos buscan facilitar la identificación de futuras retiradas, ya sea totales o parciales, de fondos en el marco de los depósitos a plazo. De esta manera, la entidad puede anticiparse y realizar una previsión para gestionar los riesgos

asociados a estas retiradas anticipadas.

Un factor de riesgo clave en la gestión del balance de una entidad financiera es el tipo de interés de mercado. Los depósitos a plazo representan un producto financiero altamente sensible a las fluctuaciones de las tasas de interés. Por tanto, es crucial para la entidad financiera incorporar un indicador que evalúe este riesgo en el modelo final. El comportamiento de los tenedores de los depósitos a plazo puede venir influenciado por los movimientos de los tipos de interés. A modo ilustrativo y desde el punto de vista financiero, ante escenarios de subidas de tipo de interés, se podría esperar un mayor índice de precancelaciones de los depósitos a plazo consecuencia del aumento del coste de oportunidad. Los clientes que contrataron un depósito a un tipo de interés previo al de la subida estarían más incentivados a cancelar el depósito y buscar un producto alternativo con las nuevas condiciones de remuneración ofrecidas en el mercado.

Por ello, en una adecuada gestión del riesgo de tipo de interés de la cartera bancaria, las entidades financieras deberán tener en cuenta la opcionalidad del cliente en el vencimiento de los depósitos a plazo y su relación con el nivel de los tipos de interés.

Estos aspectos también quedan recogidos en el marco normativo. Reguladores y supervisores son conscientes del riesgo que entrañan las opcionalidades, de los depósitos a plazo entre otros, en la gestión del riesgo de tipo de interés en la cartera bancaria. Por ello, han recogido ciertas recomendaciones y directrices al respecto:

- EBA "Final report on guidelines on the management of interest rate risk and credit spread risk arising from non-trading book activities" (EBA/GL/2022/14), 20 October 2022.

- BCBS "Interest rate risk in the banking book", April 2016.

- Banco de España "Circular 2/2016", Norma 50, art. 2(e).

En tanto las directrices de la EBA (Autoridad Bancaria Europea) como los estándares del BCBS (Comité de Basilea de Supervisión Bancaria) resaltan la importancia de tener en cuenta el posible efecto de las cancelaciones anticipadas en la gestión del riesgo de tipo de interés en la cartera bancaria. Estas directrices también destacan la necesidad de valorar de qué manera los cambios en las tasas de interés del mercado pueden influir en las tasas de cancelación anticipada, así como de contemplar aquellos factores macroeconómicos y cambios contractuales que puedan tener un impacto relevante.

1.2. Objetivos

El propósito fundamental de este trabajo será proponer varios modelos capaces de predecir la cantidad de saldo que podría retirarse de manera anticipada, ya sea en retiradas totales o parciales, en el ámbito de los depósitos a plazo ofrecidos por ABANCA. Posteriormente, se llevará a cabo una comparación entre los modelos con el objetivo de identificar cuál de ellos se destaca como el más eficaz. Concretamente, comenzaremos con el modelo propuesto por la entidad, al cual se le realizará una validación detallada de sus supuestos específicos. Adicionalmente, se propondrán varios modelos alternativos que puedan abordar y mejorar cualquier problema que surja con el modelo actualmente propuesto por la entidad. Para ello se presentan los siguientes pasos a seguir:

- **Delimitación del perímetro y configuración de la base de datos:** Esta es una de las partes fundamentales del análisis pues entendiendo la delimitación del perímetro y la estructura y configuración de la base de datos, podremos hacernos una idea global del tipo de datos con el que estamos trabajando, así como poder identificar posibles valores atípicos o datos erróneos que deberemos tratar con posterioridad.

- **Análisis exploratorio y limpieza de muestra.**

- **Estudio de correlación y asociación entre variables:** Este apartado tiene como objetivo realizar una selección inicial de las variables, con el fin de estimar correctamente los modelos de regresión, asegurándonos de que sus parámetros sean interpretables y sus predicciones fiables.

- **Análisis del modelo actual propuesto por la entidad:** Para este punto, se evaluará de manera detallada el modelo estándar que se suele emplear en la industria, evaluando tanto su efectividad como su precisión en la predicción de las retiradas anticipadas en depósitos a plazo.

-**Propuesta de modelos alternativos:** De forma alternativa al modelo estándar, se investigarán y se evaluarán varios modelos alternativos que traten de mejorar aquellas posibles limitaciones que pueda presentar el modelo inicial propuesto por la entidad.

- **Evaluación y comparación de los modelos:** Se evaluará el desempeño de los modelos de regresión, tanto del propuesto por ABANCA como de aquellos que en este trabajo se proponen alternativamente, mediante el cálculo de diversas medidas de bondad de ajuste.

Capítulo 2

Metodología

En este capítulo se muestran y describen brevemente las técnicas estadísticas utilizadas en este trabajo con el objeto de conseguir los objetivos propuestos por ABANCA. En concreto, se describen diversas herramientas del análisis exploratorio de datos como son las medidas de asociación y correlación entre variables, además de técnicas de inferencia estadística como los contrastes de Kruskal-Wallis o Wilcoxon. De igual forma, se introducirán diversos modelos de regresión como son los modelos de regresión lineal, los Modelos Aditivos Generalizados (GAM) y modelos en el marco del aprendizaje máquina como son las Support Vector Machines (SVM) y el método Boosting.

2.1. Medidas de asociación

En esta sección se realiza un análisis de las medidas de asociación entre las variables predictoras. Este análisis tiene como objetivo entre otras cosas, identificar relaciones entre las variables para evitar problemas posteriores de multicolinealidad (en el contexto de los modelos de regresión lineal), identificar asímismo la relación de estas con la variable respuesta, ayudar a la selección de variables para escoger aquellas que sean más relevantes para posteriores modelos predictivos y, en líneas generales, proporcionar una descripción general de nuestros datos y tomar decisiones adecuadas y fundamentadas para el posterior modelado de los datos.

En las siguientes subsecciones se realizará un estudio acerca de la dependencia entre las variables predictoras. Se eliminarán del estudio aquellas variables que presenten una elevada correlación entre ellas para así evitar posibles problemas de multicolinealidad futuros y, además, contribuir a estimar un modelo más sencillo y manejable, con parámetros interpretables desde un punto de vista económico (en el caso de la regresión paramétrica). Además, se eliminarán del modelo aquellas variables que presenten una baja relación con la variable respuesta.

En las diversas subsecciones que se incluyen a continuación, se muestran las diferentes medidas de asociación que se han utilizado para llevar a cabo el análisis.

2.1.1. Correlación de Pearson

El coeficiente de correlación de Pearson (Pearson en 1896), es una medida estadística que se emplea para medir la posible correlación lineal que existe entre dos variables continuas, proporcionando información sobre la dirección (ya sea positiva o negativa) y la fuerza de la relación entre dichas variables. Cabe resaltar que el hecho de que no exista relación lineal entre las variables en cuestión no significa que no pueda existir otro tipo de relación no lineal.

Sean X e Y dos variables aleatorias continuas:

$$\rho(X,Y) = \frac{Cov(X,Y)}{\sigma_X \sigma_Y}. \tag{2.1}$$

Expresión en la cual, el numerador representa la covarianza entre X e Y y el denominador es el producto de las desviaciones típicas de X e Y.

Su cálculo se especifica a través de la expresión (2.2):

$$\rho = \frac{\sum_{i=1}^{n}[(x_i - \bar{x})(y_i - \bar{y})]}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}, \tag{2.2}$$

en la cual, $\bar{x}$ y $\bar{y}$ son las medias aritméticas de x e y respectivamente.

Este coeficiente toma valores entre -1 y 1, de forma que:

1. $\rho = -1$ Existe una correlación perfecta negativa.
2. $\rho = 0$ La correlación es nula.
3. $\rho = 1$ Existe una correlación perfecta positiva.

Se entiende por correlación positiva que siempre que el valor de la X aumente, el valor de Y también tenderá a hacerlo y viceversa, mientras que, cuando la correlación es negativa, se entiende que si el valor de X aumenta, el valor de Y disminuye, y viceversa.

2.1.2. Test de Kruskal-Wallis

Kruskal y Wallis en 1952 propusieron un estadístico no paramétrico destinado a comparar dos o más muestras independientes. Su objetivo principal era realizar inferencia para determinar si las k muestras se originaron a partir del mismo modelo de distribución de probabilidad. Con tal fin, se propone el siguiente contraste:

$$H_0 : F_1 = F_2 = ... = F_k$$

$$H_1 : F_i \neq F_j \quad para \quad al \quad menos \quad un \quad par \quad (i,j), i \neq j$$

El estadístico de contraste viene determinado por la expresión (2.3):

$$H_N = \frac{12}{N(N+1)} \sum_{i=1}^{k} \frac{1}{n_i} \left(R_i - \frac{n_i(N+1)}{2} \right)^2, \tag{2.3}$$

en la cual N es el número de valores de todas las muestras combinadas, R_i es la suma de los rangos de una muestra en particular, y n_i es el número de valores de la suma de rangos correspondiente.

Kruskal demostró además que bajo la H_0 y para valores grandes de n_i se cumple que:

$$\sum_{i=1}^{k} \frac{N - n_i}{N} Z_i = H_N \approx d\chi^2_{k-1}, \tag{2.4}$$

siendo $k-1$ los grados de libertad para el test de Kruskal-Wallis y Z_i una variable aleatoria asociada a la i-ésima muestra.

Se considerará que el test es estadísticamente significativo y por ende, determinaremos que existen diferencias entre al menos dos grupos si el p-valor obtenido es menor que el nivel de significación escogido (α).

En este trabajo en particular este test se aplicará para saber si la tasa de cancelación es significativamente diferente cuando se segmenta la muestra en determinados grupos como puede ser, por plazo, por tipo de persona o por tipo de producto.

2.1.3. Test de la suma de rangos de Wilcoxon

El test de los rangos con signo de Wilcoxon también conocido como prueba de U de Mann-Whitney es un test no paramétrico desarrollado por Henry Mann y Donald Whitney en 1947 como extensión al test propuesto en 1945 por Frank Wilcoxon. Este test es utilizado para comparar dos muestras independientes cuando sus distribuciones no siguen una distribución Normal. La prueba equivalente en un contexto paramétrico sería la prueba t-student para muestras independientes.

La hipótesis a plantear en este contraste es la planteada a continuación:

$$H_0 : F_X(x) = F_Y(x), \quad \text{para todo } x$$

$$H_1 : F_X(x) \neq F_Y(x), \quad \text{para algún } x$$

donde $X_1, \ldots, X_m$ son muestras i.i.d. con distribución F_X y $Y_1, \ldots, Y_n$ son muestras i.i.d. con distribución F_Y, y ambas muestras son independientes entre sí.

Para realizar el contraste, las dos muestras se combinan para después ordenarlas por rangos, juntas. Una vez hecho esto, se determina si los valores de las muestras mezcladas previamente de manera aleatoria siguiendo el orden de los rangos o por el contrario están agrupados en extremos opuestos cuando se combinan. Si se determina finalmente un orden aleatorio, se concluiría que las dos muestras son iguales y, por tanto, no existen diferencias significativas entre ambas. En caso contrario podríamos concluir que sí existen diferencias significativas entre ambas. Con el objeto de ilustrar este procedimiento, se proporciona el ejemplo del Cuadro 2.1.

Cuadro 2.1: Ejemplo visual de resultados del contraste de la suma de rangos de Wilcoxon.

	Comparación
Los valores en la Comparación 1 están ordenados en grupos en extremos opuestos. Esto sugiere que el tratamiento X podría ser superior al tratamiento O.	X X X O X X X X O O O 1 2 3 4 5 6 7 8 9 10 11 12
Los valores en la Comparación 2 están distribuidos a lo largo de toda la distribución. Esto sugiere que no hay una diferencia clara entre los tratamientos.	X O O X X O X O X X X O X 1 2 3 4 5 6 7 8 9 10 11 12

Para determinar el estadístico de contraste se emple la expresión (2.5),

$$U_i = n_1 n_2 + \frac{n_i(n_i+1)}{2} - \sum R_i, \tag{2.5}$$

en la cual U_i es el estadístico de prueba para la muestra de interés, n_i es el número de valores de la muestra de interés, n_1 es el número de valores de la primera muestra, n_2 es el número de valores de la segunda muestra, y $\sum R_i$ es la suma de los rangos de la muestra de interés. Cabe destacar que la distribución de U_i, dentro de la hipótesis nula, depende en exclusiva de los rangos de los datos y no de su distribución, es por ello que es un estadístico de distribución libre.

Una vez se ha construido el estadístico de contraste, el siguiente paso es determinar su significación, para ello se procede a determinar la región crítica de los valores de z a través de la expresión (2.6).

$$z^* = \frac{U_i - \bar{X}_U}{S_U}, \tag{2.6}$$

en la cual $\bar{X}_U$ se define según la expresión (2.7) tal y como sigue,

$$\bar{X}_U = \frac{n_1 n_2}{2}, \tag{2.7}$$

mientras que S_U se define tal y como se indica en (2.8),

$$S_U = \sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}}, \tag{2.8}$$

y cuyo significado es el de la desviación típica, siendo U_i el valor del estadístico de contraste calculado en (2.5).

Finalmente, se determinará que el test es estadísticamente significativo si su p-valor es menor o igual al nivel α fijado por el investigador previamente.

2.2. Modelos predictivos

En esta sección se introducen brevemente los modelos de regresión lineal multivariante, los Modelos Aditivos Generalizados (GAM), los Support Verctor Machines o Máquinas de Soporte Vectorial (SVM) y los métodos tipo Boosting, todos ellos aplicados en este caso de estudio de la empresa ABANCA, con el objeto de realizar predicciones de la variable respuesta, la cantidad de saldo que podría retirarse de manera anticipada en el ámbito de los depósitos a plazo.

2.2.1. Modelo de regresión lineal

Los modelos de regresión son un tipo de herramienta estadística utilizada para estimar la relación existente entre una variable dependiente cuantitativa denotada por Y y una o varias variables independientes denotadas por $X_1, X_2 ... X_n$. Además, este conjunto de técnicas permiten, una vez construido el modelo, realizar predicciones del valor de la variable dependiente Y conocido el valor de las variables independientes X.

Supongamos que el resultado de un proceso puede denotarse por una variable aleatoria Y, denominada también variable dependiente y que va a depender de K variables independientes, las cuales serán determinadas por $X_1, X_2 ... X_K$. Podemos explicar el comportamiento de la variable dependiente Y por la relación existente dada por la función (2.9),

$$y = f(X_1, ..., X_K, \beta_1, ..., \beta_K) + \epsilon, \tag{2.9}$$

en la que f es una función bien definida, mientras que $\beta_1, ..\beta_K$ son los parámetros que deberán de estimarse para poder determinar la contribución de las variables $X_1, ..., X_K$, respectivamente, en la estimación de la variable respuesta. El término ϵ representa la naturaleza estocática de la relación entre las variables independientes y la variable dependiente, es lo que se considera como término de error. Este término recoge todas aquellas variaciones producidas en la variable dependiente que no han podido ser explicadas por las variables independedientes del modelo. Un tratamiento correcto de este término es crucial para la posterior validez y confianza de las técnicas de inferencia estadística (intervalos, contrastes de hipótesis) que se apliquen en el marco del modelo de regresión.

En este contexto, encontraremos que, partiendo de la expresión (1.1) , estaremos ante un modelo de regresión lineal si todas las derivadas parciales de Y con respecto a cada uno de los parámetros $\beta_1, ..., \beta_K$ son independientes de los parámetros, por el contrario, si cualquiera de las derivadas parciales de Y con respecto a cualquiera de los $\beta_1, ..., \beta_K$ no fuera independiente de los parámetros, entonces estaríamos ante un modelo de regresión no lineal.

Respecto a este último apunte, es importante resaltar que la linealidad del modelo, por tanto, no viene definida por la linealidad o no linealidad de las variables del modelo, si no que se refiere a la linealidad o no de los parámetros. Esto explica que, si suponemos lo que se denomina un modelo de regresión lineal polinómico, que viene denotado como se aprecia en (2.10)

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \epsilon, \tag{2.10}$$

en el que encontramos un término cuadrático en el modelo, éste siguirá siendo un modelo lineal, siempre y cuando las derivadas parciales de y con respecto a cada uno de los parámetros del modelo sean independientes de dichos parámetros.

Una labor fundamental dentro de la estimación del modelo de regresión lineal multivariante es desarrollar herramientas e implementar procedimientos para tratar de estimar el valor de los diferentes parámetros β a partir de los valores observados de Y y de $X_1, X_2, ...X_K$.

En lo sucesivo y con el fin de definir tanto parámetros del modelo lineal como sus hipótesis de partida, el presente trabajo se centrará en la explicación del modelo de regresión lineal simple, cuya expresión (recta de regresión) se define en (2.11):

$$y = \beta_0 + \beta_1 X + \epsilon, \tag{2.11}$$

en la cual los parámetros β_0 y β_1 son la constante en el origen y el parámetro de regresión (o pendiente de la recta) respectivamente, mientras que ϵ representa el error o diferencia entre las estimaciones del modelo y los valores reales de Y. Por otro lado, en el marco del modelo de regresión lineal, se parte de la hipótesis de que el término error es una variable aleatoria normalmente distribuida, con media cero y varianza constante σ^2.

Asumiendo un diseño fijo, los valores de la variable independiente X son conocidos por el investigador o experimentador, siendo el valor de la variable dependiente aleatorio, cuyo valor esperado o media es:

$$E(y) = \beta_0 + \beta_1 X, \tag{2.12}$$

siendo su varianza,

$$\text{var}(y) = \sigma^2. \tag{2.13}$$

Por otro lado, bajo un diseño aleatorio, tanto la variable independiente como la variable dependiente son aleatorias, por lo que la media de Y condicionada a cada valor que la X pueda tomar coincide con el modelo de regresión lineal tal y como se muestra en la expresión (2.14),

$$E(y|x) = \beta_0 + \beta_1 x, \tag{2.14}$$

mientras que la varianza condicionada de y dado $X = x$ se define como (2.15),

$$\text{var}(y|x) = \sigma^2. \tag{2.15}$$

El modelo de regresión lineal simple se puede extender a un modelo más complejo conocido como modelo de regresión lineal múltiple, definido por más de una variable independiente o explicativa, y para las cuales se extienden los procedimientos e hipótesis mencionadas anteriormente, así como las que a continuación se describen.

Hipótesis del modelo

El modelo de regresión lineal debe de sustentarse sobre unas hipótesis básicas, las cuales deben cumplirse para poder validar posteriormente el modelo y asumir que la inferencia realizada sobre los parámetros β es correcta y fiable. Dichas hipótesis básicas son las siguientes:

Linealidad. El modelo de regresión lineal simple se representará a través de una línea recta, es decir, Y varía a una tasa constante con respecto a variaciones de X, siendo esta tasa el parámetro de regresión β_1. La expresión de la recta de regresión es la indicada en secciones previas,

$$y = \beta_0 + \beta_1 X + \epsilon. \tag{2.16}$$

La hipótesis de linealidad implica suponer que cuando la variable independiente x toma el valor de cero, el valor promedio de la variable independiente y va a tomar un valor igual a β_0, incrementando dicha media en una cuantía fija igual a β_1 con cada incremento de una unidad de x. Además, asumir la hipotesis de linealidad implica intrínsecamente que estemos ante un modelo paramétrico en el que los valores de los parámetros β_1 y β_0 son desconocidos a nivel teórico y se deben de estimar en base a una muestra de $(X_1, X_1 ... \ X_k, Y_k)$.

Homocedasticidad. La suposición de homocedasticidad implica que la varianza que presenta el error ha de ser constante, con independencia del valor que tome la variable independiente, es decir,

$$\text{var}(\epsilon|X = x) = \sigma^2 \ \ \forall \ \ x. \tag{2.17}$$

Normalidad. Se asume que la distribución del error del modelo de regresión lineal es normal, de forma que,

$$\epsilon \in N(0, \sigma^2). \tag{2.18}$$

Independencia. Como ya habíamos mencionado en párrafos anteriores, se asumen que los errores del modelo $(\epsilon_1, ..., \epsilon_n)$ son independientes e identicamente distribuidos.

El cumplimiento de estas tres últimas hipótesis es necesario para que la inferencia realizada para la estimación de los parámetros β sea válida y el modelo pueda ser validado y pasar a fases posteriores del análisis.

Estimación de los parámetros por mínimos cuadrados ordinarios

Suponiendo que se cumplen las hipótesis de linealidad, normalidad, homocedasticidad e independencia de los errores mencionadas en el apartado anterior, y partiendo de una muestra de I conjuntos de observaciones (x_i,y_i), con ($i = 1, ..., n$), podemos escribir:

$$y_i = \beta_0 + \beta_1 x_i + e_i, \quad \text{con} \quad (i = 1, ..., n). \tag{2.19}$$

El principio sobre el que se sustenta el método de estimación por mínimos cuadrados ordinarios es conseguir que la suma de los cuadrados de la diferencia entre las observaciones y las predicciones sea mínima, por ello, si la predicción de nuestro modelo de regresión lineal viene denotada como:

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i, \tag{2.20}$$

los errores de predicción o residuos se pueden definir como,

$$\hat{\epsilon}_i = Y_i - \hat{Y}_i = Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i, \quad \text{para} \quad i \in \{1, ..., n\}, \tag{2.21}$$

por lo que el método de mínimos cuadrados lo que pretende es encontrar aquellos estimadores $\hat{\beta}_0$ y $\hat{\beta}_1$ que den lugar a un menor error de estimación, es decir, a un valor más bajo, en el conjunto de toda la muestra, de los residuos.

Con tal fin, el presente enfoque minimiza la suma de los cuadrados de los residuos, es decir, se toman los residuos al cuadrado para conseguir evitar que exista una compensación entre residuos positivos y negativos. El criterio de mínimos cuadrados se puede definir a través de la expresión (2.22)

$$\sum_{i=1}^{n} (Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 = \min_{\beta_0, \beta_1} \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 x_i)^2. \tag{2.22}$$

Las derivadas parciales que ayudan a encontrar el mínimo de la suma de cuadrados con respecto a β_0 y β_1 son, respectivamente,

$$\frac{\partial S(\beta_0, \beta_1)}{\partial \beta_0} = -2 \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) \tag{2.23}$$

y

$$\frac{\partial S(\beta_0, \beta_1)}{\partial \beta_1} = -2 \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) x_i. \tag{2.24}$$

El valor de las expresiones (2.23) y (2.24) en el mínimo de la suma de cuadrados es 0. Haciendo esta igualdad, se estiman los valores de los parámetros $beta_0$ y $beta_1$, siendo las expresiones resultantes

$$\hat{\beta}_0 = \bar{Y} - \frac{S_{xY}}{S_x^2} \bar{x} \tag{2.25}$$

y

$$\hat{\beta}_1 = \frac{S_{xY}}{S_x^2}. \tag{2.26}$$

con $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ e $\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$ las medias respectivas de la variable explicativa y la variable respuesta, $S_{xy} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(Y_i - \bar{Y})$ la covarianza y $S_x^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$ la varianza muestral de la variable explicativa.

Finalmente, la recta de regresión estimada por el método de mínimos cuadrados ordinarios será quella que pasa por el punto definido por las medias tanto de la variable idependiente como de la variable dependiente, conocido también como vetor de medias $(\bar{x}, \bar{Y})$, y que además va a presentar una pendiente con valor $\hat{\beta}_1 = \frac{S_{xy}}{S_x^2}$.

Por último, empleando también la suma de los cuadrados de los residuos, podemos estimar la varianza poblacional del error, la cual es desconocida:

$$\hat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n}\hat{\varepsilon}_i^2 = \frac{1}{n-2}\sum_{i=1}^{n}(Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2. \tag{2.27}$$

2.2.2. Modelo GAM

El modelo aditivo generalizado fue desarrollado por Hastie y Tibshirani (1986) como una extensión del modelo lineal generalizado (GLM) propuesto por Nelder y Wedderburn (1972) ya que el Modelo Aditivo Generalizado GAM permite que las predicciones lineales incluyan sumas de funciones suaves de las covariables. Concretamente los moelos GAM reemplazan la forma lineal $\sum \beta_j X_j$ por una suma de funciones suaves $\sum s_j(X_j)$. En general el modelo GAM va a presentar una estructura similar a la que se presenta en un modelo lineal, según la expresión (2.28)

$$g(\mu_i) = X_i^*\theta + f_1(x_{1i}) + f_2(x_{2i}) + f_3(x_{3i}, x_{4i}) + \ldots, \tag{2.28}$$

en la que $\mu_i = \mathbb{E}(Y_i)$ e $Y_i \sim$ según alguna familia de distribución de la familia exponencial, siendo Y_i la variable respuesta, X_i^* una fila de la matriz del modelo para cualquier componente del modelo estrictamente paramétrico (lineal) y, θ el vector de parámetros correspondiente, mientras que f_j son funciones suaves no especificadas de las covariables x_k. Por tanto, este tipo de modelos pueden incluir tanto efectos lineales de las covariables sobre la respuesta como también estimaciones no paramétricas de efectos no lineales de dichas covariables o mismo de sus interacciones.

A diferencia del modelo de regresión lineal visto en el apartado anterior, los modelos GAM son modelos mucho más flexibles dado que no se asume la hipótesis de linealidad entre la variable respuesta y las variables predictoras, por ende son capaces de capturar patrones no lineales y son muy útiles para explorar relaciones complejas entre los datos. Asimismo, son modelos no paramétricos, es decir, no es necesario que se asuma de manera predefinida una forma funcional concreta. El hecho de poder incorporar efectos paramétricos y no paramétricos en su estructura hace que se denominen modelos semiparamétricos.

Si bien es cierto que encontramos bastantes ventajas respecto al modelo de regresión lineal, este modelo no está libre de ciertos problemas teóricos ya que es preciso, por un lado, abordar el modelado matemático de las relaciones no lineales entre la variable respuesta y las variables explicativas y, por otro lado, encontrar el grado de suavidad óptimo, es decir, el nivel de flexibilidad apropiado con el que se va a definir la función suave. Decidir el grado de suavidad de las funciones es de vital importancia para no tener un modelo demasiado flexible o infraajustado donde no se capturen todas las particularidades de nuestos datos o por el contrario obtener un modelo demasiado rígido o sobreajustado donde se capture demasiado ruido aleatorio que no aporta información al modelo y que puede interferir en la interpretación del modelo.

La monografía de Wood (2007) aporta información completa acerca de éste y otros temas relacionados con el ajuste de modelos GAM, proponiendo el ajuste de los efectos suaves de las variables independien-

tes a partir de diferentes tipos de bases de splines, además de algoritmos específicos para la estimación de los parámetros de los modelos.

2.2.3. Modelo Boosting

El modelo Boosting es una de las diversas técnicas que se han ido desarrollando a lo largo del tiempo dentro del aprendizaje autómatico (machine learning). Este modelo fue desarrollado en primera instancia para resolver problemas de clasificación y no fue hasta años después que se extendieron para dar solución a problemas de regresión.

Originariamente, los primeros modelos Boosting vieron la luz en la década de los años 90 de la mano de Schapire (1990) y Freund (1995) cuyas teorías se vieron ampliamente influenciadas por la teoría del aprendizaje de Valiant (1984) y Kearns y Valiant (1989), donde varios clasificadores débiles, esto es, un clasificador que predice ligeramente mejor que el azar se combinan para crear un clasificador de conjunto que presenta una tasa de error de clasificación generalizada superior. Finalmente, tras un trabajo de varios años con el objetivo de implementar un algoritmo que implementara de manera efectiva la teoría boosting, Freund y Schapire colaboran hasta desarrollar el algoritmo AdaBoost, el cual proporcionó una implementación práctica y efectiva para potenciar un aprendizaje débil en un aprendizaje fuerte.

Este algoritmo fue un éxito y se empezó a estudiar la conexión de dicho algoritmo con conceptos estadísticos de funciones de pérdida, modelado aditivo y regresión logística, demostrando que el boosting puede interpretarse como un algoritmo aditivo por etapas que minimiza la pérdida exponencial. En este punto encontramos que tenemos un algoritmo capaz de adaptarse a diferentes tipos de problema y finalmente se consiguió extender el algoritmo, que como se mencionaba anteriormente había sido diseñado para dar respuesta a problemas de clasificación, a modelos de regresión. Friedman, 2001 nombró a este método "gradient boosting machines", los cuales abarcan, ahora sí, tanto la clasificación como la regresión.

Supongamos que como función de pérdida se utiliza la Suma de los Cuadrados de los Residuos (RSS), entonces el coste o pérdida de usar $m(x)$ para predecir la variable respuesta y en la muestra de entrenamiento es:

$$L(m) = \sum_{i=1}^{n} L(y_i, m(x_i)) = \sum_{i=1}^{n} (y_i - m(x_i))^2. \tag{2.29}$$

Teniendo en cuenta la expresión anterior, se desea minimizar $L(m)$ con respecto a m mediante el método de los gradientes, siendo estos los residuos. Si $L(m) = \frac{1}{2}(y_i - m(x_i))$, entonces $-\frac{\partial L(y_i, m(x_i))}{\partial m(x_i)} = y_i - m(x_i) = r_i$, siendo r_i el residuo de la i-ésima observación. Este residuo se espera que sea pequeño, ya que de lo contrario podemos sospechar de que el modelo no estará captando de forma precisa la relación subyacente entre las variables independientes y la variable dependiente. Este residuo será el utilizado en el algoritmo de manera iterativa para ajustar y mejorar el modelo.

Para exponer el funcionamiento del algoritmo, vamos a suponer un problema de regresión utilizando árboles de decisión:

1. Seleccionaremos el número de iteraciones B, así como el parámetro de regularización λ y el número de cortes de cada árbol d.

2. Se establece una predicción inicial constante y se calculan los residuos correspondientes a los datos i de la muestra de entrenamiento:

$$\hat{m}(x) = 0, \quad r_i = y_i. \tag{2.30}$$

3. Para $b = 1, 2, ..., B$, se debe repetir:

   3.1. Ajustar un árbol de regresión $\hat{m}^b$ con d cortes utilizando los residuos como respuesta: (X, r).

   3.2. Calcular la versión regularizada del árbol:

$$\lambda \hat{m}^b(x). \tag{2.31}$$

   3.3. Actualizar los residuos:

$$r_i \leftarrow r_i - \lambda \hat{m}^b(x_i). \tag{2.32}$$

4. Estimar el modelo boosting:

$$\hat{m}(x) = \sum_{b=1}^{B} \lambda \hat{m}^b(x). \tag{2.33}$$

Una vez construido el algoritmo debemos tener en cuenta que contamos con tres hiperparámetros suceptibles de optimizar para tratar de seleccionarlos de forma óptima y tratar de evitar posibles sobreajustes o infraajustes. Dichos parámetros son:

1. B: Este hiperparámetro recoge el número de árboles que vamos a emplear en nuestro algoritmo. Un número muy elevado de árboles puede dar problemas de sobreajuste ya que en este algoritmo los árboles que se construyen no son independientes a diferencia de lo que ocurre en otros métodos como bagging o bosques aleatorios. En cada iteración del algoritmo, cada nuevo árbol que se genera trata de corregir los errores del anterior, y a esta construcción de nuevos árboles se le conoce como proceso "greedy.º "voraz"porque en cada paso que se da se toma la mejor decisiñón local. El problema de esto es que una sucesión de óptimos locales no garantiza la mejor solución u óptimo global posible, por lo que puede conllevar un sobreajuste del modelo.

2. d: Este hiperparámetro controla el número de cortes que debe de tener cada árbol. Debido a que necesitamos un aprendizaje lento, lo conveniente es utilizar un parámetro d pequeño, es decir, que nuestros árboles tengan pocos cortes, ya que esto ayudará a que poco a poco se puedan cubrir zonas en las que la predicción puede ser más compleja. Hay en diversas situaciones en las que lo conveniente es utilizar $d = 1$, es decir, emplear un único corte. Si $d > 1$, podemos interpretarlo como un parámetro que en vez de medir el número de cortes mide el número de iteraciones que se producen entre las diferentes variables del modelo.

3. λ: El parámetro de regularización lambda se puede interpretar como la velocidad a la que aprende el algoritmo. Este hiperparámetro está comprendido entre 0 y 1, y en los inicios del uso de este algoritmo era muy común utilizar $\lambda = 1$ aunque se fue pudiendo observar que dando este valor a λ el algoritmo no acababa de funcionar bien del todo. Se ha sabido que valores pequeños de este hiperparámetro consiguen evitar el sobreajuste, por lo que en la actualidad es común utilizar valores de $\lambda = 0{,}01$ o $\lambda = 0{,}001$, aunque lo ideal sería emplear criterios como puede ser el de validación cruzada para seleccionar el valor óptimo de este parámetro de regularización. Un dato importante a tener en cuenta es que, aunque un menor tamaño de λ ayuda a evitar el sobreajuste tiene como inconveniente que puede empeorar el tiempo computacional ya que a menor valor de λ, mayor lentitud presentará el proceso de aprendizaje y se emplearán más iteraciones para concluir el algoritmo.

Más adelante, el propio Friedman propuso una mejora de su algoritmo apoyandose en la técnica bagging de Breiman (1996). Esta mejora tiene que ver con la incorporación de un esquema de muestreo aleatorio en vez de utilizar toda la muestra de entrenamiento. A este algoritmo mejorado le denominó stochastic gradient boosting (SGB) y es a día de hoy la técnica más utilizada.

El algoritmo SGB presenta, como única diferencia con respecto al algoritmo presentado en líneas anteriores, la selección aleatoria de una fracción de los datos de entrenamiento en la primera línea dentro

del bucle. Esta fracción de datos de entrenamiento, conocida como fracción de bagging incorpora al modelo otro hiperparámetro a optimizar en el algoritmo, sugirendo Friedman una fracción de bagging = 0.5. Nuevamente este parámetro se puee ajustar mediante validación cruzada o por cualquier otro método para tratar de ajustarlo lo mejor posible a nuestro modelo. Cabe resaltar que la introducción de este hiperparámetro mejoró sustancialmente la precisión de predicción del modelo boosting y a la vez redujo de manera significativa el costo computacional.

Como resumen de este modelo podemos determinar que los modelos boosting son modelos de aprendizaje lento, donde los árboles son pequeños y crecen de forma secuencial con el objetivo de mejorar la clasificación o predicción anterior, y entendiendo que a diferencia de metodologías similares como pueden ser los bosques aleatorios o los modelos bagging, puede presentar problemas de sobreajuste. Por último, mencionar que el modelo final se trata de un modelo aditivo ya que representa la media ponderada de las contribuciones individuales de cada árbol al conjunto total.

2.2.4. Máquinas de soporte vectorial

Las máquinas de soporte vectorial (support vector machines, SVM) fueron desalloradas en la década de los 90 por Vapnik, (1995) con el objetivo de dar respuesta a los problemas de clasificación binaria, esto es, problemas de clasificación de dos categorías, empleando hiperplanos para separar los datos. Esta metodología fue ganando popularidad gracias a las ventajas que presentaba debido a ser modelos altamente flexibles y robustos. Más adelante, al igual que sucediera con los modelos boosting explicados en el apartado anterior, estos modelos se extendieron para dar solución a problemas multiclase así como problemas de regregresión o detección de atípicos.

Aunque lo que se ha usado en el presente trabajo son las máquinas de soporte vectorial para la regresión, se va a contextualizar brevemente su uso para clasibicación binaria con el objetivo de entender mejor su posterior extensión a los modelos de regresión.

En clasificación binaria, las SVM parte de la idea de seleccionar un hiperplano que actue como frontera entre los dos conjuntos o clases de datos, debiendo ser elegido no de manera arbitraria sino entendiendo que debe de presentar ciertas propiedades especificas como por un lado, la equidistancia, esto es, el hiperplano debe pasar lo más cerca posible del medio de los dos puntos más cercanos que pertenezan a cada clase y por otro conseguir el margen máximo, esto es, maximizar este margen que no deja de ser otra cosa que la distania entre el hiperplano y el punto más cercano de cada categoría de tal modo que se consiga minimizar el error de generalización del clasificador.

Boser et al (1992) propusieron una modificación en todos los cálculos que conducen a la expresión (2.34):

$$m(\mathbf{x}) = \beta_0 + \sum_{i \in S} \alpha_i \mathbf{x}^t \mathbf{x}_i, \tag{2.34}$$

en la que los productos escalares $\mathbf{x}^t\mathbf{x}_i$, $\mathbf{x}_i^t\mathbf{x}_j$ se estiman mediante funciones alternativas a partir de los datos las cuales reciben el nombre de funciones *kernel*, resultando la máquina de soporte vectorial

$$m(\mathbf{x}) = \beta_0 + \sum_{i \in S} \alpha_i K(\mathbf{x}, \mathbf{x}_i). \tag{2.35}$$

Algunas de las funciones Kernel más empleadas en la práctica son

- Kernel lineal: Definido como el producto escalar entre los vectores x e y:

$$K(\mathbf{x}, \mathbf{y}) = \mathbf{x}^t\mathbf{y}. \tag{2.36}$$

- Kernel polinómico: Este Kernel es una generalización del Kernel lineal, donde encontramos el parámetro d que indica el grado del polinomio, el cual debe ser $d > 1$:

$$K(\mathbf{x}, \mathbf{y}) = (1 + \gamma \mathbf{x}^t \mathbf{y})^d. \tag{2.37}$$

- Kernel radial: El cual depende de la distancia euclídea entre los puntos y su uso es muy útil cuando la separación existente entre categorías no es lineal. A mayor valor de γ mayor flexibilidad, por ello es un parámetro importante y se debe seleccionar su valor con precaución para evitar problemas de sobreajuste:

$$K(\mathbf{x}, \mathbf{y}) = \exp(-\gamma \|\mathbf{x} - \mathbf{y}\|^2). \tag{2.38}$$

- Tangente hiperbólica: Este tipo de kernel emplea una función hiperbólica tangente por la cual una vez calculado el producto escalar entre dos vectores x e y, el resultado de este producto se va a variar añadiendo un parámetro γ de escala y un parámetro c de desplazamiento. Estas variaciones permiten capturar relaciones no lineales entre los vectores:

$$K(\mathbf{x}, \mathbf{y}) = \tanh(1 + \gamma \mathbf{x}^t \mathbf{y}). \tag{2.39}$$

Una vez hemos introducido las SVM apliacadas al contexto de la clasificación binaria, podemos ya adentrarnos en la regresión SVM. En este ámbito de la regresión, ya no disponemos de grupos o clases de datos para separar, por tanto ahora el enfoque de este tipo de modelo se centra en considerar aquel hiperplano que mejor se adapte a los datos de entrenamiento en base a un margen de error predefinido ϵ. Una vez tenemos ese margen de error considerado, la idea es que la gran mayoría de los datos estén a una distancia menor que ϵ del hiperplano. Por otra parte, aquellos datos que disten más de ϵ del hiperplano se llamarán los "vectores de soporte", los cuales ayudan a definir y a ajustar la forma del hiperplano asegurando que el modelo se ajuste óptimamente a la variabilidad de los datos.

El enfoque seguido por Drucker et al. (1997) para definir matemáticamente los modelo SVM para regresión, son característicos por emplear un enfoque robusto. En este enfoque se determina un umbral ϵ que indica la cantidad de error que se está dispuesto a cometer y un parámetro de coste c el cual controla la penalización de aquellos errores que exceden el umbral ϵ. Estos parámetros se pueden fijar usando criterios, como por ejemplo, validación cruzada para tratar de encontrar los valores óptimos que ayuden a construir un modelo que proporcione predicciones precisas y robustas. Emplear este enfoque como alternativa al enfoque de RSS (suma de los residuos al cuadrado) en la regresión SVM evita que todos los datos influyan en el modelo y que además dado que los errores están elevados al cuadrado tengamos un problema con la influencia de los atípicos en el modelo, ya que estos tendrían mucha mñas relevancia que si usáramos el error absoluto.

Una vez definidos los parámetros de coste y de umbral, se procede a fijar la función de pérdida que queda definida como se expresa en (2.40):

$$L_{\epsilon,c}(x) = \begin{cases} 0 & \text{si } |x| < \epsilon \\ (|x| - \epsilon)c & \text{en otro caso} \end{cases} \tag{2.40}$$

Una vez definida la función de pérdida y partiendo del modelo más sencillo de regresión lineal, el modelo SVM estimará los parámetros del modelo minimizando la expresión (2.41) definida por

$$\sum_{i=1}^{n} L_{\epsilon,c}(y_i - \hat{y}_i) + \sum_{j=1}^{p} \beta_j^2. \tag{2.41}$$

De este modo el modelo puede expresarse en función de los vectores de soporte, que como se ha mencionado anteriormente son aquellos datos para los cuales su residuo excede del umbral ϵ como viene especificado en (2.42)

$$m(\mathbf{x}) = \beta_0 + \sum_{i \in S} \alpha_i \mathbf{x}^t \mathbf{x}_i. \tag{2.42}$$

Por último, agregando una función Kernel, el modelo SVM queda finalmente especificado como se muestra en (2.43)

$$m(\mathbf{x}) = \beta_0 + \sum_{i \in S} \alpha_i K(\mathbf{x}, \mathbf{x}_i), \tag{2.43}$$

en la que $m(x)$ es una función que viene definida por la suma de un término constante $beta_0$ y una suma ponderada de la función Kernel (K) aplicada a los pares de vectores (x, x_i), donde para cada una de estas aplicaciones de la función Kernel estará, a su vez, multiplicada por un coeficiente α_i.

Respecto a las ventajas que presenta las máquinas de soporte vectorial respecto a otros modelos se pueden mencionar, entre otras, su robustez frente a valores atípicos gracias al suevizado del margen mediante el ajuste del parámetro de coste C, además de ser modelos flexibles que pueden adaptarse a fronteras no lineales complejas, y por ende pueden proporcionar buenas predicciones. Por el contrario, presentan algunas desventajas como por ejemplo la dificultad que presenta la interpretación de los modelos ajustados bajo esta metodología, así como, el hecho de que pueden conllevar elevados tiempos de computación y están diseñados para predictores numéricos por lo que, si vamos a trabajar con algún modelo que cuente con alguna variable categórica es de vital importancia realizar un preprocesado de dichas variables explicativas categóricas para transformarlas en variables indicadoras.

Capítulo 3

Tratamiento y preprocesado de datos

3.1. Delimitación del perímetro

Para establecer la delimitación del perímetro se extrae mediante, el software SQL, Lans, (1992) la exposición total de la entidad a productos de depósitos a plazo. De acuerdo a las cuentas anuales consolidadas publicadas a cierre de septiembre de 2023, hay un total de 11292 millones de euros en este tipo de instrumento financiero donde se han podido localizar cuatro tipos de productos bien delimitados.

- Depósitos Multiplazo (69 %).
- Depósitos Plazo Clara (7 %).
- Otros depósitos (12 %).
- Depósitos Plazo Empresas Grupo (5 %).

El 7 % restante se incluye en un quinto grupo denominado "Resto", ya que son diferentes tipos de depósitos a plazo cuyos pesos sobre el total de depósitos no son materiales.

Se valoró la posibilidad de trabajar únicamente con los depósitos Multiplazo por ser los que mayor peso tienen sobre el total de depósitos, pero finalmente se decide incluir todos los tipos de depósitos por un lado, para poder tener una visión más fidedigna de los retiros y cancelaciones anticipadas independientemente del tipo de producto contratado, y por otro lado porque se comprobó que computacionalmente era posible manejar la cantidad de datos que suponía trabajar con todos los depósitos (incluido "Resto").

Una vez determinados los productos con los que se iban a trabajar, se ajusta aún más el perímetro de la muestra indicando la ventana temporal, la cual se establece entre el 1-1-2018 y el 31-11-2023, además de escoger solo aquellos contratos cuya divisa es el euro. Además, dado que lo que se obtiene son observaciones diarias y lo que nos interesan son observaciones mensuales, se aplica un último filtro para quedarnos solo con las observaciones referentes al último día del mes.

3.2. Generación y limpieza de muestra

Una vez definido el perímetro sobre el cual se desea trabajar, se procede a la extracción de la muestra. Ayudándonos nuevamente de la herramienta SQL, se acude a la información interna de la entidad para

extraer todos los registros de contratos de depósitos a plazo, donde encontramos diferentes variables (Apéndice) que posteriormente nos servirán para complementar y configurar la base de datos definitiva sobre la que se trabajará para las predicciones de los "early redemption".

Partimos de una base de datos donde se recogen todos los movimientos asociados a cada uno de los contratos existentes en la entidad de depósitos a plazo. Esta es una muestra con 6342339 de observaciones. Concluida la extracción de la muestra, se procede a la limpieza de la misma, eliminando aquellos datos que cumplan las siguientes características:

- Plazo menor que un mes.
- Importes en negativo o cero.
- Contratos con fecha de vencimiento o de constitución missings.
- Fecha de constitución fuera de rango.
- Contratos con un saldo máximo menor que 10.
- Contratos que presenten una tasa de cancelación (ER) anómala.

Una vez aplicados los filtros, la muestra queda reducida a un total de 4624808 de observaciones. Para finalizar este apartado se realizó un pequeño análisis descriptivo de la muestra con el propósito de obtener una comprensión de su configuración y asegurar que efectivamente, no queda ningún dato atípico o incoherencia que no guarde coherencia con el análisis que nos ocupa.

3.3. Posibles factores de riesgo

Se procede, una vez se tiene la muestra limpia y ajustada a los intereses del estudio, a analizar potenciales factores de riesgo que pueden influir en el retiro o cancelación anticipada del monto de los depósitos a plazo.

En base a estudios similares llevados a cabo en la entidad, y atendiendo a la definición de los "Early Redemption" se sabe que ciertos factores de riesgo de tipo contractual que se pueden tener en cuenta para este análisis son:

- **ESTACIONALIDAD:** En cuanto a la estacionalidad, se pretende valorar si hay algún patrón que pueda indicar una periodicidad mensual en la cancelación anticipada de depósitos. Para ello se genera una variable de manera externa llamada ESTACIONALIDAD extraída a partir del mes de la FECHA DE MOVIMIENTO.
- **SEASONING:** El tiempo de vida del depósito podría ser un factor de riesgo ya que se sospecha que dependiendo de cuánto tiempo lleve vivo el depósito puede influir significativamente en la cancelación o retiro anticipado de los depósitos a plazo. Para ello, y dado que esta no es una variable que encontremos de inicio en nuestra muestra, se genera de manera externa. El seasoning se calcula en meses como:

  FECHA DE MOVIMIENTO - FECHA DE CONSTITUCIÓN.

- **PLAZO:** La variable PLAZO indica el tiempo por el que se ha contratado el depósito y, nuevamente, se sospecha que dependiendo del plazo fijado a la hora de contratar el instrumento financiero puede influir en la cancelación anticipada del mismo. Se procede por ende, al cálculo de la variable PLAZO como sigue:

  FECHA DE VENCIMIENTO - FECHA DE CONSTITUCIÓN.

- **TIPO DE PERSONA:** Por último, la variable TIPO DE PERSONA, que bien puede ser física, jurídica u otro tipo de persona, se plantea como otra posible candidata a influir en las

cancelaciones anticipadas. Esta es una variable que podemos emplear del conjunto de datos original y no hace falta crearla de manera externa.

Una vez determinados los factores de riesgo de tipo contractual, se localizan los factores de riesgo de tipo macroeconómico que pueden llegar a tener influencia sobre la tasa de cancelación anticipada. Concretamente, se seleccionan los siguientes:

- **PARO DESESTACIONALIZADO:** Número de desempleados registrados en las Oficinas de los Servicios Públicos de Empleo, en miles. Serie desestacionalizada. Corregida de estacionalidad mediante el X-11 ARIMA. Total España. Fuente: Ministerio de trabajo. Se informa con frecuencia mensual.
- **IPC:** El Índice de precios de consumo (IPC) es una medida estadística de la evolución de los precios de los bienes y servicios que consume la población residente en viviendas familiares en España. Índice 2016=100. Fuente: INE. Se informa con frecuencia mensual.
- **PIB REAL:** Producto Interior Bruto (PIB) generado por la economía española, medido en términos de volumen. Fuente: INE, Contabilidad nacional trimestral de España. Se informa con frecuencia trimestral.
- **PIB NOMINAL:** Producto Interior Bruto en términos nominales generado por la economía española, medido en términos de volumen. Se informa con frecuencia trimestral.
- **EURIBOR 12 MESES:** Media aritmética de la suma del EURIBOR 12M diario durante el mes / entre los días hábiles del EUR (TARGET) de dicho mes.
- **CONSUMO HOGARES:** Consumo realizado por los hogares a final de mes.
- **CONSUMO FINAL:** Gasto en consumo final a final de mes, siendo este el gasto realizado por las unidades institucionales residentes en bienes y servicios que se utilizan para satisfacer directamente las necesidades o carencias individuales o las necesidades colectivas de los miembros de la comunidad.
- **CONSUMO PRIVADO:** Gasto en consumo privado a final de mes, siendo este igual al gasto realizado por las familias más el gasto de las empresas privadas y las instituciones privadas sin ánimo de lucro residentes en un país.
- **GTESP1Y:** Tipo del bono español a un año a final de mes.
- **GTESP10Y:** Tipo del bono español a diez años a final de mes.

Para aquellas variables que tienen periodicidad trimestral se ha procedido a interpolar linealmente los datos para conseguir tener observaciones mensuales. Por otra parte, variables como PIB Nominal/PIB Real, GTESP1Y/GTESP10Y o CONSUMO PRIVADO/CONSUMO HOGARES se valorará cuál es la mejor opción para utilizar en función de nuestra muestra.

Capítulo 4

Análisis exploratorio de datos

Se realizó un análisis exploratorio de todos los factores de riesgo mencionados en el apartado anterior. Este análisis se realiza con el objetivo de tener una mejor comprensión de todas estas variables, obtener información valiosa a cerca del comportamiento de estos factores así como establecer una base sólida para los posteriores análisis que se lleven a cabo.

Para alcanzar estos objetivos se realizaron análisis gráficos detallados de las variables, utilizando herramientas como boxplots, histogramas o diagramas de dispersión. Se examinaron también las densidades de las variables, se analizó su posible distribución, así como la posible relación lineal de estas con la variable respuesta. Esto permitió obtener una visión preliminar de qué variables podrían ser relevantes en los futuros modelos de regresión.

Así mismo se realizó un análisis de medidas centrales, variabilidad, valores máximos y mínimos y cuantiles. Todo esto ayudó a tener una visión completa de las variables con las que se van a trabajar, entender su comportamiento y en líneas generales ayudar a la toma de decisiones durante el posterior análisis estadístico.

4.1. Partición de la muestra

Tras realizar el análisis exploratorio inicial para comprender en profundidad la naturaleza de los datos, se consideró que sería oportuno chequear la posibilidad de partir la muestra en base a alguna variable de interés para la empresa. Para ello se revisaron aquellas variables que podían ser candidatas para particionar la muestra ya que, por la naturaleza de las mismas se sospecha que podrían revelar diferencias estadísticamente significativas en el análisis del comportamiento de las cancelaciones anticipadas, dependiendo de la categoría en la que se sitúen dichas variables.

Se revisó en primera instancia la distribución de las variables PLAZO, TIPO DE PERSONA, SECTOR_ID y REMUNERADA por considerar que estas variables podrían arrojar diferencias significativas entre categorías. Atendiendo a los resultados mostrados en los cuadros 4.1, 4.2, 4.3 y 4.4 determinamos eliminar como posibles candidatas a partición de la muestra las variables TIPO DE PERSONA, SECTOR_ID debido a presentar una distribución totalmente desbalanceada donde aproximadamente el 99 % de las observaciones se encuentran localizadas en una única categoría, presentando el resto de categorías observaciones mínimas o marginales, lo cual no es idóneo para la partición de la muestra y análisis posterior.

Tras analizar la distribución de estas variables, se lleva a cabo un análisis inferencial. En concreto, se emplea el test de Kruskal-Wallis para determinar si efectivamente existen diferencias estadísticamente significativas entre las cuatro categorías para la variable PLAZO en las cancelaciones anticipadas y

Cuadro 4.1: Distribución de PLAZO (Frecuencias Relativas).

PLAZO	Frecuencia Relativa (%)
3 meses	16.11 %
6 meses	30.38 %
12 meses	33.07 %
> 12 meses	20.44 %

Cuadro 4.2: Distribución de TIPO PERSONA (Frecuencias Relativas).

TIPO PERSONA	Frecuencia Relativa (%)
Persona jurídica	0.90 %
Persona Física	99.10 %
Otros	0.00 %

posteriormente el test de la suma de rangos de Wilcoxon para verificar si existen diferencias estadísticamente significativas en las cancelaciones anticipadas para los dos grupos de la variable REMUNERADA.

Tal y como podemos observar en los cuadros 4.5 y 4.6, ambos test resultan estadísticamente significativos ya que presentan un p-valor ambos menor que cualquier α posible, por tanto se determina que existen diferencias significativas entre grupos para las cancelaciones anticipadas, tanto dependiendo del plazo como de si el depósito es o no remunerado.

Teniendo en cuenta estos resultados, se analizaron las prioridades estratégicas de la empresa y se decidió que era importante de cara a los objetivos de la misma analizar y comprender el comportamiento de las cancelaciones anticipadas en relación con el plazo en detrimento de si la cuenta era o no remunerada. Es por este motivo por lo que, finalmente, se decidió que para el análisis final se particionaría la muestra atendiendo al plazo.

Cabe destacar que, concretamente, esta variable fue creada como se explica en el apartado "Generación y limpieza de muestra", pero que posteriormente se discretizó para obtener categorías de tal forma que la primera de ellas (PLAZO = 3) contenga los datos de los contratos que tienen plazo entre 1 y 3 meses, la segunda (PLAZO = 6) se corresponda con los datos de los contratos que presentan plazos entre 4 y 6 meses y, la tercera de ellas (PLAZO = 12) incluya los datos de los contratos que presentan plazo entre 7 y 12 meses y la cuarta (PLAZO >12 meses) abarque todos los contratos que presentan plazos superiores a 12 meses.

Para nuestro estudio y por cuestiones de espacio y tiempo decidimos llevar a cabo el análisis para la muestra de PLAZO = 12, debido a que es la que mayor número de observaciones presenta y a que, por regla general, la mayoría de contratos de depósitos a plazo se firman a 12 meses, por lo que es la más representativa de todas las particiones. En cualquier caso, el procedimiento llevado a cabo para

Cuadro 4.3: Distribución de SECTOR_ID (Frecuencias Relativas).

SECTOR ID	Frecuencia Relativa (%)
Fondos no financieros	0.0011 %
Familias	99.0992 %
Empresas no financieras	0.6618 %
Seguros y fondos de pensiones	0.0314 %
Otros intermediarios financieros	0.0012 %
Aux. financieros	0.0043 %
Inst. sin fines de lucro	0.1661 %
Inst. financieras monetarias	0.0003 %
Org. autónomos comerciales	0.0043 %
Empresas no financieras	0.0023 %
Admon. regional	0.0030 %
Admon. local	0.0113 %
Admon. S.Social	0.00004 %
Org. autónomos admon. central	0.0031 %
Org. autónomos admon. regional	0.0057 %
Org. autónomos admon. local	0.0003 %
Nombre comercial	0.0045 %

Cuadro 4.4: Distribución REMUNERADA (Frecuencias Relativas).

REMUNERADA	Frecuencia Relativa (%)
1	40.82 %
2	59.18 %

la submuestra de PLAZO = 12 sería análogo para las submuestras de PLAZO = 3, PLAZO = 6 y PLAZO > 12.

Cuadro 4.5: Resultados del Test de Kruskal-Wallis.

Test	**Chi-cuadrado de Kruskal-Wallis**	**p-Valor**
Kruskal-Wallis	64834	$< 2{,}2 \times 10^{-16}$

Cuadro 4.6: Resultados del Test de Wilcoxon con Corrección de Continuidad.

Test	**Estadístico W**	**p-Valor**
Wilcoxon	$9{,}3494 \times 10^{11}$	$< 2{,}2 \times 10^{-16}$

4.2. Medidas de asociación

En esta sección se realiza un estudio de correlación entre variables cuantitativas a partir del cálculo de los coeficientes de correlación de Pearson, con el objeto de determinar el grado de dependencia lineal existente entre cada par de variables y así determinar cuál o cuáles de las variables con las que se cuentan pueden ser candidatas para incluir en el modelo de regresión final para predecir la tasa de cancelación anticipada.

4.2.1. Correlación Pearson entre variables continuas

Una vez que tenemos todas las variables que pueden ser de utilidad para predecir la tasa de cancelación anticipada, tanto a nivel contractual como a nivel macroeconómico, se procede a chequear las correlaciones de pearson entre las variables continuas.

Como era previsible, según se aprecia en la Figura 4.1, todas las variables relacionadas con tipos de interés están altamente correladas entre sí. Por lo tanto, en la selección, de entre las diversas propuestas de tasas de interés, optaremos por aquella que exhiba la correlación más fuerte con la variable de respuesta ER. Por otro lado, observamos que el Índice de Precios al Consumo (IPC) también presenta una correlación significativa con las distintas tasas de interés. Esta asociación se debe al ajuste periódico realizado por el BCE de las tasas de interés en respuesta a condiciones económicas, como la inflación. Por consiguiente, es común encontrar una correlación elevada entre estas variables. Dado que nuestro interés en el modelo radica en incorporar una de las variables de tipo de interés para entender la sensibilidad de las cancelaciones anticipadas ante variaciones en las tasas de interés, decidiremos prescindir de la variable IPC.

En lo que se refiere a las variables que tienen en cuenta el consumo, vemos que entre ellas están muy correladas, como es lógico, debido que el cosumo privado tiene en cuenta para su cálculo el consumo de los hogares, y a su vez, el consumo final tiene en cuenta para su estimación el consumo privado. Decidiremos si nos quedamos con alguna de estas tres variables si encontramos una relación significativa con la variable respuesta ER.

En base a estos resultados, se decide optar por trabajar con la variable SPREAD ya que esta tiene en cuenta tanto el EURIBOR como el tipo de interés de referencia fijado en el contrato del depósito a plazo. Es muy buen indicador desde el punto de vista financiero para poder entender las posibles precancelaciones anticipadas, debido a que si los tipos del mercado son superiores a los fijados en el

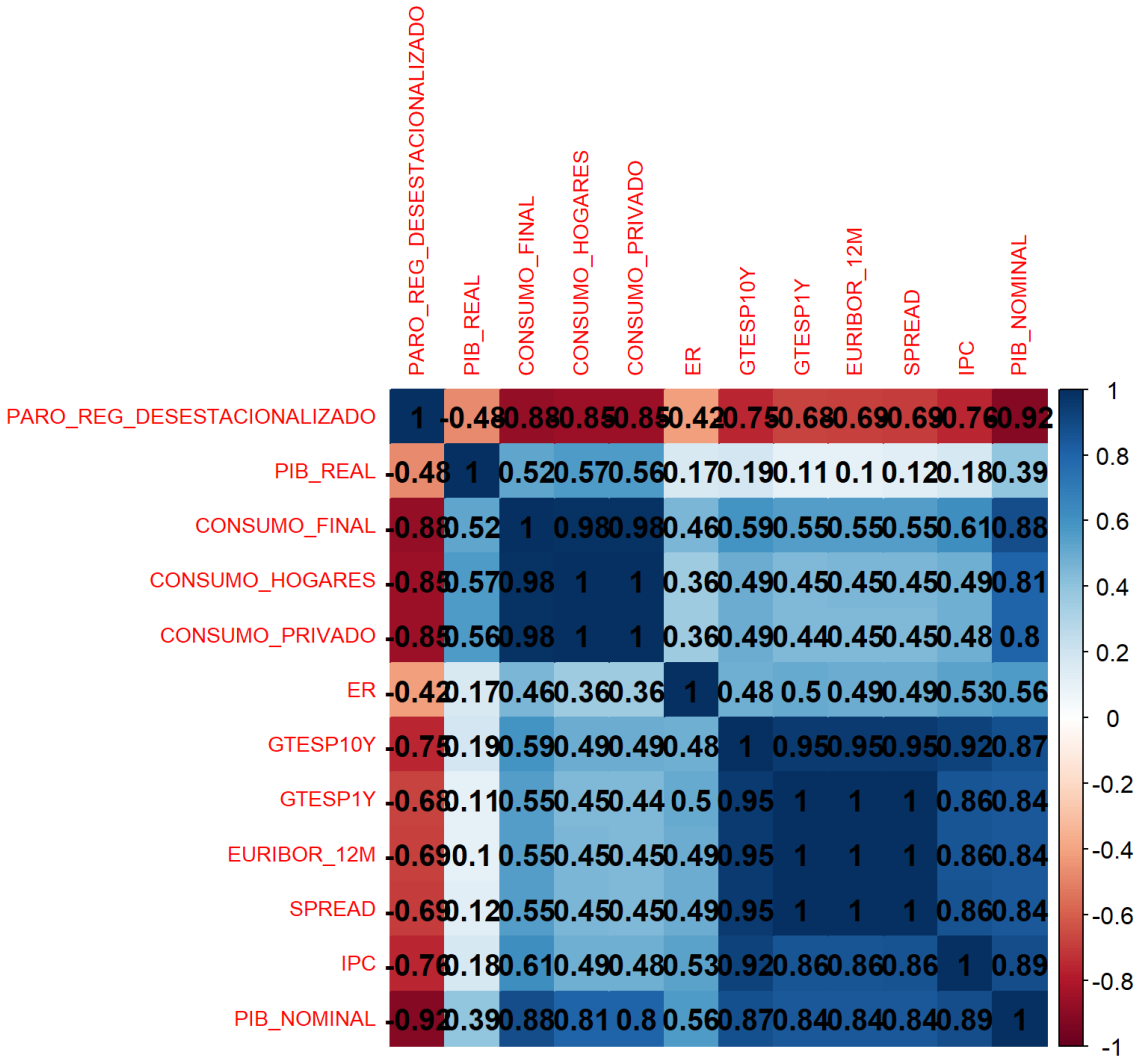

Figura 4.1: Correlograma o matriz de correlaciones (utilizando el coeficiente de correlación lineal de Pearson) para variables continuas.

N°	Variable	Correlación con ER
1	SPREAD	0.49
2	EURIBOR_12M	0.49
3	GTESP10Y	0.48
4	GTESP1Y	0.5
5	IPC	0.53
6	PIB_NOMINAL	0.56
7	PIB_REAL	0.17
8	CONSUMO_FINAL	0.46
9	CONSUMO_HOGARES	0.36
10	CONSUMO_PRIVADO	0.36
11	PARO_DESEST	-0.42

Cuadro 4.7: Correlación de diferentes variables con ER.

contrato del depósito, se espera que se produzcan más cancelaciones ya que se preferirá asumir la penalización por cancelación anticipada a cambio de obtener un mayor rendimiento reinvirtiendo de nuevo a los nuevos tipos de interés. En el supuesto opuesto, en el cual los tipos del depósito sean superiores al EURIBOR, se esperará que suceda lo contrario. Por ello y debido a la alta correlación que ya se comentaba en líneas anteriores, se prescinde de EURIBOR_12M, GTESP10Y, GTESP1Y e IPC.

Respecto al PIB_REAL, se aprecia una correlación muy baja, tan solo del 0.17 con la variable respuesta, por ello también prescindiremos de ella para el análisis. En lo que atañe al PIB_NOMINAL, aunque encontramos una correlación más elevada que PIB_REAL con la variable ER, observamos también una correlación muy elevada con la variable SPREAD, concretamente una correlación del 0.84, lo cual puede ocasionar problemas de colinealidad en el modelo de regresión. Debido al tipo de modelo que se quiere construir, y teniendo sospechas que el tipo de interés (en este caso reflejado por la variable SPREAD) es una variable que afecta a las cancelaciones anticipadas, así como, la necesidad por parte de la entidad en identificar cómo afectan los tipos de interés a las cancelaciones anticipadas (ya que estamos ante un modelo de riesgo de tipo de interés) se decide prescindir de la variable PIB_NOMINAL y de este modo evitar problemas de colinealidad.

Por último, las varaibles PARO_DESEST Y CONSUMO_FINAL muestran una correlación del -0.42 y 0.46 respectivamente, una con una relación inversa y la otra directa. Debido a que la correlación con la variable respuesta no es excesivamente grande y que, tras analizar bien la casuistica de las cancelaciones anticipadas se sospecha que estas variables no tienen una gran relevancia sobre la tasa de cancelación, se decide finalmente prescindir de ellas.

Tras realizar este análisis, se decide incluir únicamente la variable SPREAD en nuestro modelo predictivo, acompañada de la variable SEASONING. La decisión de incluir esta última variable se tomó en

base al modelo estándar que se suele utilizar en la industria para el tratamiento de modelos de riesgo, en los cuales la variable SEASONING (la cual recoge el tiempo vida del depósito) es una variable de interés para predecir la tasa de cancelación anticipada.

Con esto, obtendremos en principio un modelo sencillo con solo dos variables explicativas, con las que se pretende explicar con precisión el comportamiento de las cancelaciones anticipadas para que la entidad pueda ejecutar planes de prevención a futuro debido a la alta demanda de depósitos a plazo en los últimos años en un contexto actual de tipos de interés elevados, pero que tras las últimas previsiones del BCE podría darse una tendencia a la baja en el medio plazo. Este cambio de tendencia podría afectar a este tipo de productos y por ello, como ya se comentó en los objetivos del trabajo, es importante modelizar y plantear una herramienta para poder predecir las cancelaciones anticipadas que puedan suceder en la entidad.

4.3. Agregación de variables

Es importante que se comprenda la agregación que se ha llevado a cabo en las variables debido a que, tras realizar este procedimiento pasamos de contar con un número muy elevado de datos, concretamente 4624808 una vez realizada la limpieza de la muestra a tener menos de 100 observaciones si usáramos la muestra sin particionar.

Esto es debido al tratamiento que se realizó sobre los datos en base al interés y objetivos de la empresa. Los datos extraídos en primera instancia vienen desglosados a nivel contrato, pudiendo un mismo cliente tener asociados distintos números de contrato puesto que para cada renovación del mismo, se generará un nuevo número asociado a esa renovación. Sin embargo, a la empresa lo que le interesa manejar no era el comportamiento de las cancelaciones a nivel contrato ni tampoco a nivel cliente, sino tener una comprensión a nivel agregado de las cancelaciones anticipadas y su evolución mensual.

Es por este motivo por el cual se agregan las variables a nivel mensual. En concreto, la agregación llevada a cabo consiste en agrupar por el número de meses que lleva vivo el contrato y por el año de apertura del mismo. De esta manera se calcula el promedio de las cancelaciones anticipadas, y del SPREAD en función del SEASONING y COSECHA. Tras realizar esta agregación y teniendo en cuenta el espacio temporal en el que están recogidos los datos, resultó en una notable reducción del número de observaciones.

Capítulo 5

Resultados de la aplicación de modelos de regresión

En esta sección se van a analizar los resultados obtenidos por los distintos modelos de regresión empleados para predecir la tasa de cancelación anticipada de los depósitos a plazo. Para ello recordamos que se ha realizado una partición de la muestra, donde en nuestro caso se va a trabajar con la submuestra resultante PLAZO = 12 meses y donde los datos han sido agrupados por las variables COSECHA y SEASONING. Basándonos en esta agrupación, las variables ER (tasa de cancelación anticipada) y SPREAD (diferencial entre el tipo de interés del mercado y el tipo de interés del depósito) han sido tratadas de manera que representen la media en cada combinación de momento de COSECHA y SEASONING.

En la primera etapa, se realiza una partición aleatoria del conjunto de datos, donde el 80 % de las observaciones pertenecen al conjunto de entrenamiento y el 20 % restante al conjunto de prueba o test. Esta división se realiza una única vez. Los modelos seleccionados se entrenan utilizando el conjunto de entrenamiento y posteriormente se evalúan en el conjunto de prueba o test. Los criterios de evaluación incluyen métricas estándar de rendimiento o desempeño, que nos permiten realizar una comparación inicial y determinar el comportamiento de cada modelo bajo esta configuración específica.

En la segunda etapa del análisis, se aplica un enfoque de muestras aleatorias de Monte-Carlo para evaluar la robustez de los modelos (estimando posición y variabilidad de cada una de las métricas de bondad de ajuste). En concreto, se obtienen 100 muestras aleatorias, seleccionando en cada una de ellas un conjunto de entrenamiento y test, según las proporciones 80/20 %. Cada modelo se reentrena y evalúa para cada una de las muestras. Esta cantidad de 100 iteraciones del proceso que dan lugar a 100 muestras, se ha seleccionado cuidadosamente considerando el tamaño reducido de la muestra con la que estamos trabajando, con el objeto de alcanzar un equilibrio efectivo entre evitar problemas de redundancia y, al mismo tiempo, asegurar un número suficiente de iteraciones que permitan obtener estimaciones fiables acerca de la habilidad predictiva de los modelos aplicados, en términos de posición y variabilidad de las medidas de bondad de ajuste calculadas. Con 100 iteraciones, estamos en condiciones de capturar adecuadamente la variabilidad del desempeño de los modelos a través de diferentes subconjuntos de datos, mientras minimizamos el riesgo de sobreajuste y mantenemos la eficiencia computacional. Este procedimiento ayudará a tener una comprensión más sólida y profunda tanto de la estabilidad como, sobre todo, de la fiabilidad del rendimiento del modelo y nos ayudará a ser capaces de determinar de una manera consistente si los resultados obtenidos en la primera parte del análisis se deben al azaar o realmente podemos concluir que el modelo es óptimo y capaz de predecir de manera robusta la tasa de cancelación anticipada.

Por último, recordar que para construir el modelo final, únicamente se considerarán las variables SPREAD y SEASONING por los motivos especificados en el apartado de Análisis de datos.

5.1. Primera etapa

En esta primera etapa como se mencionaba anteriormente, se realiza una primera evaluación de los modelos predictivos realizando una única iteración. Se evaluará la muestra del entrenamiento y la muestra test y finalmente se compararán los resultados para tener una visión general de la capacidad predictiva de los modelos y tener una primera toma de contacto sobre cuál de todos ellos se postula como el mejor modelo para predecir la tasa de cancelación anticipada. Los modelos que se analizarán son: el Modelo de Regresión Lineal, el Modelo GAM, el modelo Boosting y el Modelo de Máquinas de Soporte Vectorial.

A continuación, presentaremos los resultados obtenidos para estos cuatro modelos en la primera etapa del análisis. Se examinarán de manera individual los resultados de cada modelo, teniendo en cuenta que se basan en una única ejecución y recordando que se ha utilizado una división aleatoria del 80 % para el conjunto de entrenamiento y del 20 % para el conjunto de prueba. Este enfoque nos permite establecer una base comparativa inicial y evaluar la eficacia de cada modelo en un escenario controlado.

Modelo de regresión lineal

Empezamos por el modelo más sencillo y el que se usa de manera frecuente en la industria bancaria, ya que se considera el modelo óptimo para modelizar las cancelaciones anticipadas de depósitos a plazo en el ámbito de riesgo de tipo de interés. Se empieza planteando el modelo que queda descrito como aparece en la expresión (5.1)

$$ER = \beta_0 + \beta_1 \text{SPREAD} + \beta_2 \text{SEASONING} + \epsilon. \tag{5.1}$$

Se procede a estimar el modelo con los datos de entrenamiento y a comprobar si los coeficientes son estadísticamente significativos. Para ello ajustamos el modelo utilizando la función `lm` de la librería `stats` de R obteniendo la siguiente salida con la estimación de los parámetros junto con su estudio de significación y medidas de bondad de ajuste:

```
Call:
lm(formula = ER ~ SEASONING + SPREAD, data = train)

Residuals:
       Min         1Q     Median         3Q        Max
-0.0034952 -0.0014058 -0.0001561  0.0012018  0.0071193

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.0147564  0.0007219  20.442  < 2e-16 ***
SEASONING   0.0002993  0.0001082   2.767  0.00852 **
SPREAD      0.0011154  0.0001857   6.006 4.64e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.002211 on 40 degrees of freedom
Multiple R-squared:  0.5461,    Adjusted R-squared:  0.5234
F-statistic: 24.07 on 2 and 40 DF,  p-value: 1.376e-07
```

Figura 5.1: Summary correspondiente al modelo de regresión lineal multivariante ajustado para estimar la variable ER.

Podemos observar que todos los parámetros del modelo son significativamente diferentes de cero pues presentan un p-valor menor que $\alpha = 0.05$ por ello todos los efectos de las variables independiente sobre la respuesta son significativos y, por tanto, relevantes para estimar los valores de ER. Por otra parte, este modelo está caracterizado por un R^2 igual a 54.61 %, es decir, aproximadamente un 54.61 % de la variabilidad en la variable respuesta es explicada por las variables explicativas del modelo de regresión lineal. Aunque esto representa una contribución sustancial, también significa que queda un 45.39 % de la variabilidad sin explicar, lo cual puede considerarse una limitación del modelo, por ello, es prioritario realizar modificaciones adicionales tomando como partida el modelo de regresión lineal anterior, y probando su desempeño mediante el uso de la muestra de prueba, determinando si las diversas modificaciones del modelo mejoran estos resultados o si, por el contrario, tienen dificultades para realizar predicciones precisas. De este modo, además de incluir los efectos lineales de SEASONING (Figura 5.3) y SPREAD (Figura 5.5), se ha añadido el efecto de SEASONING^2, teniendo en cuenta la información mostrada en la Figura 5.4, indicativa de un posible efecto no lineal de tipo parabólico de SEASONING sobre ER. Teniendo esto en cuenta, se ha ajustado un modelo de regresión de tipo polinómico, definido por la expresión (5.2),

$$ER = b_0 + b_1\text{SEASONING} + b_2\text{SEASONING}^2 + b_3\text{SPREAD}, \tag{5.2}$$

con R^2=69.43 % como se puede apreciar en la Figura 5.2.

```
Call:
lm(formula = ER ~ SEASONING + I(SEASONING^2) + SPREAD, data = train)

Residuals:
       Min         1Q     Median         3Q        Max
-0.0040798 -0.0010844 -0.0000427  0.0010378  0.0052920

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     1.810e-02  9.757e-04  18.551  < 2e-16 ***
SEASONING      -1.341e-03  3.880e-04  -3.457  0.00133 **
I(SEASONING^2)  1.410e-04  3.243e-05   4.347 9.57e-05 ***
SPREAD          1.152e-03  1.546e-04   7.451 5.20e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.001837 on 39 degrees of freedom
Multiple R-squared:  0.6943,    Adjusted R-squared:  0.6708
F-statistic: 29.52 on 3 and 39 DF,  p-value: 3.936e-10
```

Figura 5.2: Summary correspondiente al modelo de regresión lineal multivariante con efecto no líneal (tipo parabólico) ajustado para estimar la variable ER.

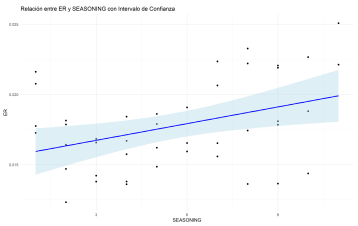

Figura 5.3: Ajuste del modelo ER y SEASONING

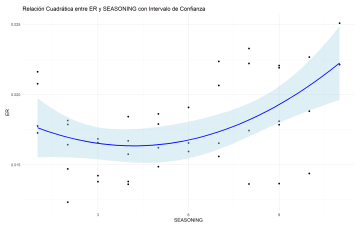

Figura 5.4: Ajuste del modelo ER y SEASONING^2

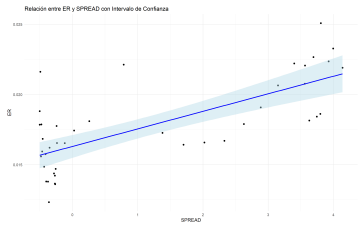

Figura 5.5: Ajuste del modelo ER y SPREAD

A continuación, se muestra el valor de los parámetros del modelo ajustado así como una breve discusión

sobre su significación, bondad de ajuste e importancia de cada predictor. En lo que se refiere a la interpretación del modelo, destacar lo siguiente:

• La constante en el origen es igual a 0.0181, lo cual indica que cuando las variables independientes SEASONING, SEASONING^2 y SPREAD tienen un valor de cero, el valor esperado de la variable de respuesta ER es de aproximadamente 0.0181.

• El coeficiente asociado a la variable SEASONING es -0.001341, lo cual indica que cuando esta variable aumenta una unidad, la variable respuesta ER tiende a disminuir en -0.001341 unidades.

• El coeficiente estimado para el término cuadrático de SEASONING es 0.000141. Específicamente, el valor estimado del parámetro indica que a medida que SEASONING aumenta, el efecto sobre ER aumenta 0.000141 unidades por cada unidad cuadrada de aumento en SEASONING.

• El coeficiente asociado a la variable SPREAD es de 0.001152, lo cual indica que cuando la variable SPREAD aumenta una unidad, la variable respuesta ER se incrementa en 0.001152 unidades.

El estudio de significación de los parámetros estimados se ha realizado a partir de la estimación de intervalos de confianza al 95 % para el valor de cada uno de ellos, mostrados en el Cuadro 5.1 y, de forma más intuitiva, también en la Figura 5.6.

De la observación del Cuadro 5.1 y de la Figura 5.6 se concluye que ninguno de los intervalos contiene al cero, por tanto, todos los efectos de los predictores sobre la respuesta ER son estadísticamente significativos.

Cuadro 5.1: Intervalos de confianza al 95 % para los parámetros del modelo.

Variable	2.5 %	97.5 %
(Intercept)	$1{,}568418 \times 10^{-2}$	0.0197784905
SEASONING	$-1{,}932452 \times 10^{-3}$	-0.0002895253
I(SEASONING^2)	$5{,}121287 \times 10^{-5}$	0.0001904737
SPREAD	$8{,}391099 \times 10^{-4}$	0.0014736211

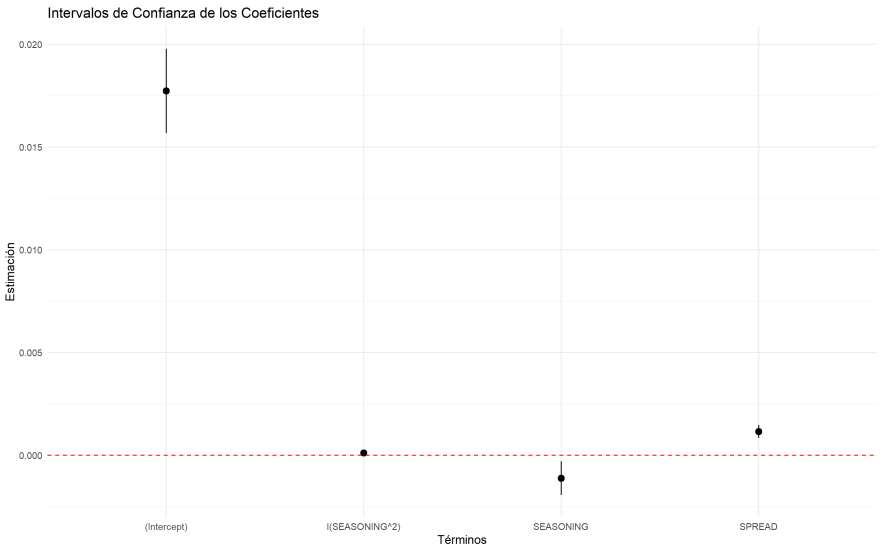

Figura 5.6: Intervalos de confianza modelo de regresión lineal

La Figura 5.7 indica la importancia relativa de SEASONING, SEASONING^2 y SPREAD en la explicación de la variable respuesta ER, en términos de porcentaje sobre el coeficiente de determinación, utilizando la contribución R2 promediada sobre los ordenamientos entre regresores (Linderman, 1980) y la métrica lmg (Grömping, 2006). Se comprueba en la figura 5.7 que la variable SPREAD es la más importante en el modelo suponiendo el 62.52 % de la variabilidad explicada por el modelo (en términos de % sobre el R^2), mientras que la el efecto de SEASONING^2 aporta al modelo el 22.76 % de la variabilidad explicada, siendo el 14.71 % restante el correspondiente a la variable SEASONING. Esto sugiere que cambios en la variable SPREAD tendrán un impacto más significativo en la salida del modelo que cambios en SEASONING^2 o SEASONING.

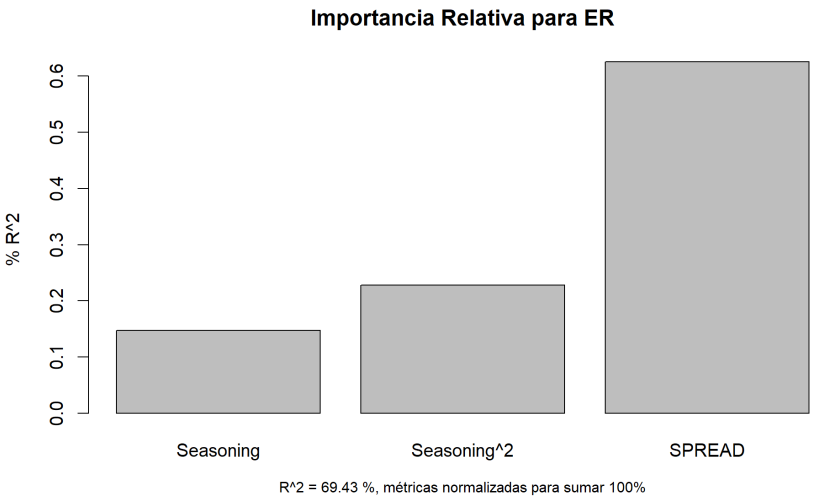

Figura 5.7: Importancia relativa de las variables

Por último, se representa el efecto parcial de cada una de las variables sobre la variable respuesta ER en la Figura 5.8. Observamos que para las variables SEASONING^2 y SPREAD hay una relación positiva con la variable respuesta, mientras que para la variable SEASONING la relación es negativa como ya veíamos en el estudio de significación del modelo.

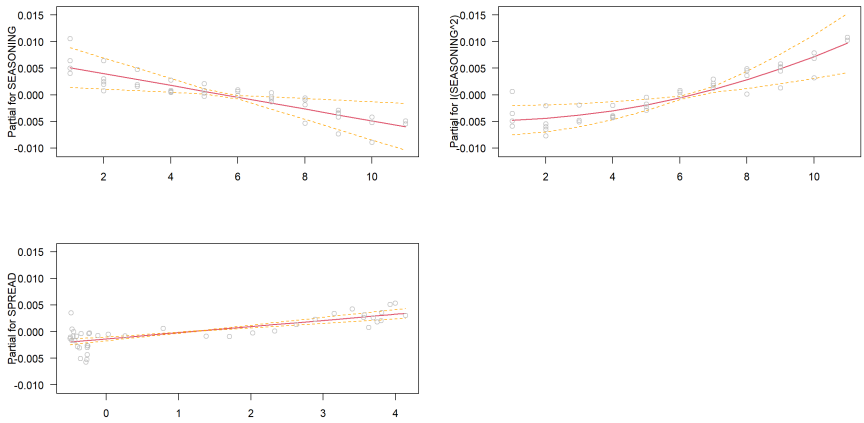

Figura 5.8: Efectos parciales de las variables independientes correspondientes a modelo lineal que explica ER en función de SEASONING, SEASONING^2 y SPREAD.

En lo que respecta a la diagnosis del modelo de regresión lineal, podemos determinar, en base a los resultados que se incluyen en el Cuadro 5.2, las siguientes conclusiones. Fijando un nivel de significación $\alpha = 0.05$, los resultados de los test aplicados para la diagnosis del modelo indican que los p-valores asociados a la normalidad de los residuos, la homocedasticidad y la no autocorrelación son todos superiores al nivel de significación fijado. Por ello, se determina que no hay evidencia estadísticamente significativa para rechazar las hipótesis nulas correspondientes a cada uno de los test y podemos concluir que los residuos del modelo son normales, homocedásticos y no autocorrelados.

Cuadro 5.2: Resultados de los tests diagnósticos para el modelo de regresión en función de SEASONING, SEASONING^2 y SPREAD.

Test	Estadístico	p-valor
Shapiro-Wilk	W = 0.9802	0.6582
Durbin-Watson	DW = 2.1508	0.6355
Breusch-Pagan	BP = 5.0199	0.1703

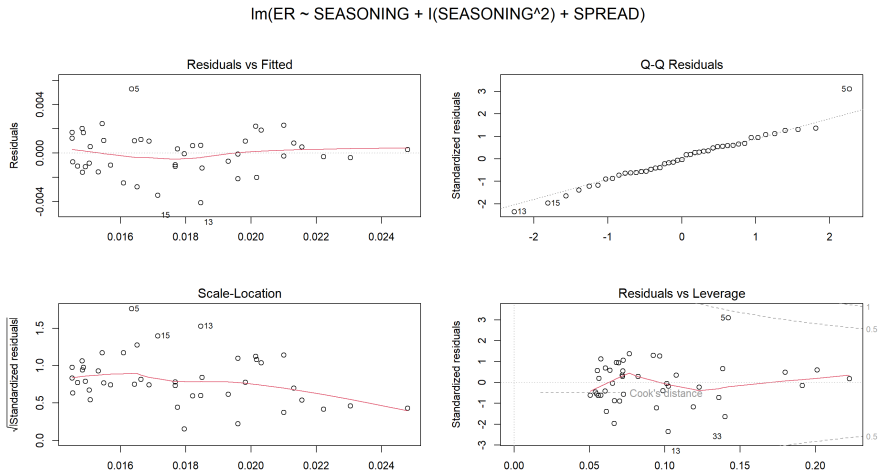

Figura 5.9: Diagnosis del modelo de regresión lineal.

Los gráficos de la Figura 5.9, ayudan a confirmar las conclusiones mencionadas en el párrafo anterior. Adicionalmente, observamos que, aunque el punto etiquetado como cinco podría considerarse un valor atípico debido a su desviación con respecto a los demás datos, no podemos considerarlo como influyente en el modelo debido a que no sobrepasa el umbral de la distancia de Cook.

Por último, se procede a realizar las predicciones sobre la muestra test para poder determinar la capacidad predictiva del modelo, los resultados pueden verse en el Cuadro 5.3.

Cuadro 5.3: Métricas de desempeño, calculadas a partir de la muestra de test, para el modelo de regresión lineal.

Modelo	ME	RMSE	Rsquared
LM	-0.0007949	0.0019743	0.5840509

Observamos que el error medio es negativo, lo cual indica una tendencia a sobreestimar, pero es muy cercano a cero por lo que podemos inferir que no existe un sesgo significativo en las predicciones, y que estas, en líneas generales son bastante cercanas a los valores observados. Por su parte, el RMSE también tiene un valor pequeño, cercano a cero, lo cual nuevamente es un buen indicador, ya que indica que los errores del modelo son pequeños en magnitud y por ende que el modelo presenta un buen ajuste a los datos. Dado que estamos trabajando con porcentajes, un RMSE de aproximadamente 0.19 % refleja una buena en las predicciones del modelo. En lo que respecta al R^2, obtenemos un valor aproximado de 58 %, que es indicativo de que, si bien el modelo resultante es interesante para explicar la relación existente entre ER, por un lado, y SEASONING y SPREAD, por otro, no es así capaz de predecir con alta precisión los valores de ER a partir de nuevos valores de SEASONING y SPREAD. La Figura 5.10 apoya la discusión previa, es decir, el modelo es capaz de estimar un valor relativamente aproximado de ER pero no con una alta precisión ni exactitud. De hecho, existe una diferencia apreciable entre la bisectriz y la línea de regresión lineal obtenida a partir de los valores observados y predichos, además de la existencia de apreciables diferencias entre los pares (observación, predicción) con respecto a la recta. Por todo ello, es interesante el ajuste de modelos de regresión alternativos.

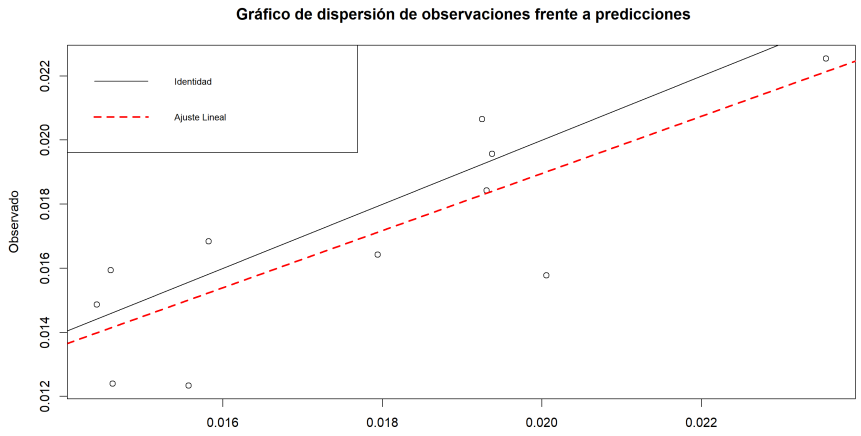

Figura 5.10: Gráfico de dispersión de observaciones vs predicciones.

Modelo GAM

El segundo modelo propuesto es el modelo GAM. Se opta por este modelo para tratar de determinar si existe algún tipo de relación no lineal que el anterior modelo no puede capturar y de este modo mejorar la capacidad predictiva del modelo. En este caso el modelo se ajusta con la función `gam` de la librería `mgcv` Wood, 2007.

El resumen o summary del ajuste, incluyendo la estimación de los parámetros y efectos suaves, así como su estudio de significación y medidas de bondad de ajuste, se incluye en la Figra 5.11.

```
Family: gaussian
Link function: identity

Formula:
ER ~ s(SEASONING) + s(SPREAD, bs = "cr")

Parametric coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.0177705  0.0002195   80.95   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Approximate significance of smooth terms:
               edf Ref.df      F p-value
s(SEASONING) 3.935  4.863  3.564  0.0124 *
s(SPREAD)    7.265  7.773 14.870  <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

R-sq.(adj) =  0.798   Deviance explained = 85.2%
GCV = 2.8934e-06  Scale est. = 2.0724e-06  n = 43
```

Figura 5.11: Resumen del ajuste del modelo GAM.

Observamos que tanto el intercepto como los coeficientes de los términos suavizados son estadísticamente significativos, arrojando p-valores menores a $\alpha = 0.05$. En lo que se refiere a los coeficientes asociados a las variables suavizadas, se observa que la variable SEASONING tiene 3.935 grados efectivos de libertad (edf), lo cual apunta a una complejidad moderada en la relación no lineal. Por el

contrario, la variable SPREAD muestra un edf de 7.265, lo cual indica una mayor complejidad en su ajuste. Además se obtiene un R^2 ajustado de casi el 80 % y un valor para la deviance del 85.2 % lo cual sugiere que el modelo presenta una buena capacidad explicativa.

Esto se puede ver gráficamente en la Figura 5.12, en la que se observa la relación no lineal que existe entre la variable SEASONING y ER, concretamente parece que existe una relación cuadrática entre ambas. Esto implica que tanto valores bajos como altos de SEASONING están asociados con mayores tasas de ER, con un punto de inflexión intermedio. Por otro lado, se muestra una relación claramente no lineal y compleja entre la variable SPREAD y ER.

Además, la zona sombreada representa el intervalo de confianza al 95 % para las funciones suavizadas, es decir, nos indica que a con un 95 % de confianza la verdadera función suavizada se encuentra dentro de la zona sombreada. Esto nos da indicaciones de la incertidumbre en la estimación de la función suavizada, concretamente la amplitud del intervalo de confianza para SPREAD, al ser más ancha, denota una mayor incertidumbre en la estimación de esta relación en comparación con SEASONING, cuyo intervalo es más estrecho y lo cual puede indicar una mayor certeza en la estimación.

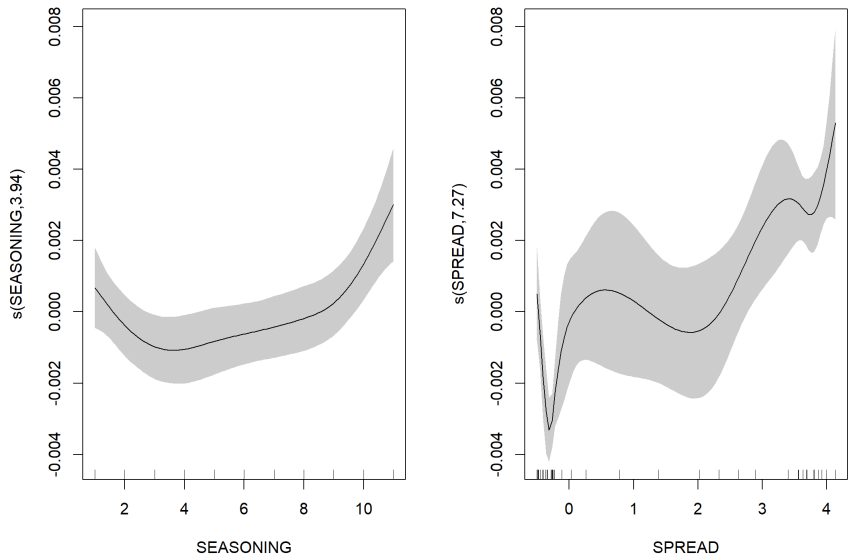

Figura 5.12: Funciones suavizadas e intervalos de confianza

Por otro lado, a la vista de los gráficos de efectos parciales, la relación cuadrática observada para SEASONING indica que las tasas de ER son mayores para valores extremos de SEASONING, lo que podría reflejar una tendencia a una mayor cancelación anticipada en los extremos de la vida de los depósitos (es decir, en su inicio y en momentos cercanos a la cancelación contractual). Respecto a la variable SPREAD, observamos que para valores bajos de esta variable la tasa de cancelación ER es pequeña, pero que estas aumentan rápidamente ante pequeños cambios en el SPREAD. Más difícil es interpretar el efecto de las cancelaciones anticipadas ante valores de SPREAD intermedios ya que existen muchas fluctuaciones pudiendo ser debidas a incertidumbre sobre la variabilidad de esta. Por último, para valores grandes de SPREAD se observan altas tasas de cancelación anticipada.

Se han contrastado las hipótesis del modelo, obteniendo resultados tanto gráficos como numéricos que confirman que se cumplen las condiciones para la validación del modelo tal y como se aprecia en la Cuadro 5.4 y en la figura 5.13 .

Cuadro 5.4: Resultados de los tests de diagnóstico para el modelo GAM.

Test	Estadístico	p-valor
Shapiro-Wilk	W = 0.97729	0.5439
Breusch-Pagan	BP = 2.5158	0.2842
Durbin-Watson	DW = 1.7661	0.4362

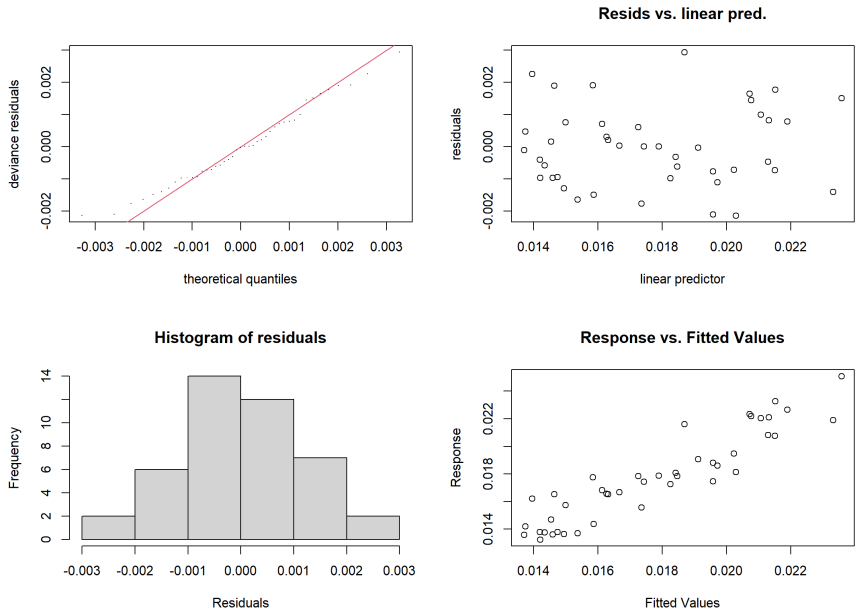

Figura 5.13: Gráficos de diagnosis del modelo GAM.

En última instancia, se procede a realizar las predicciones del modelo GAM sobre la muestra test para poder determinar su capacidad. Los resultados se pueden consultar en la tabla 5.5.

Cuadro 5.5: Métricas de desempeño para el modelo GAM.

Modelo	ME	RMSE	Rsquared
GAM	-0.0010248	0.0012830	0.7443275

Observamos que el error medio en este caso también es negativo, lo cual indica que el modelo GAM también tiene una tendencia a sobreestimar, aunque el valor es muy cercano a cero lo cual es un hallazgo positivo ya que indica que el sesgo en las predicciones es pequeño. En lo que respecta al RMSE podemos decir que presenta un valor de aproximadamente el 0.12 % lo cual indica que en media el modelo se confunde en las predicciones un 0.12 %. Por último, el R^2 ha empeorado ligeramente en comparación con el coeficiente de determinación correspondiente a la muestra de entrenamiento (80 %), ya que ahora es de un 74 % aproximadamente, es decir, aproximadamente un 74 % de la variabilidad en

la variable respuesta puede ser explicada por las variables explicativas del modelo de regresión lineal. En líneas generales, el modelo muestra un buen ajuste, manteniendo un nivel aceptable de precisión en las predicciones, tal y como también se puede ver en la Figura 5.14. De hecho, la bisectriz se encuentra próxima a la recta de regresión construida a partir de los pares (observación, predicción), observándose menores diferencias entre puntos y recta que en el caso del modelo de regresión lineal.

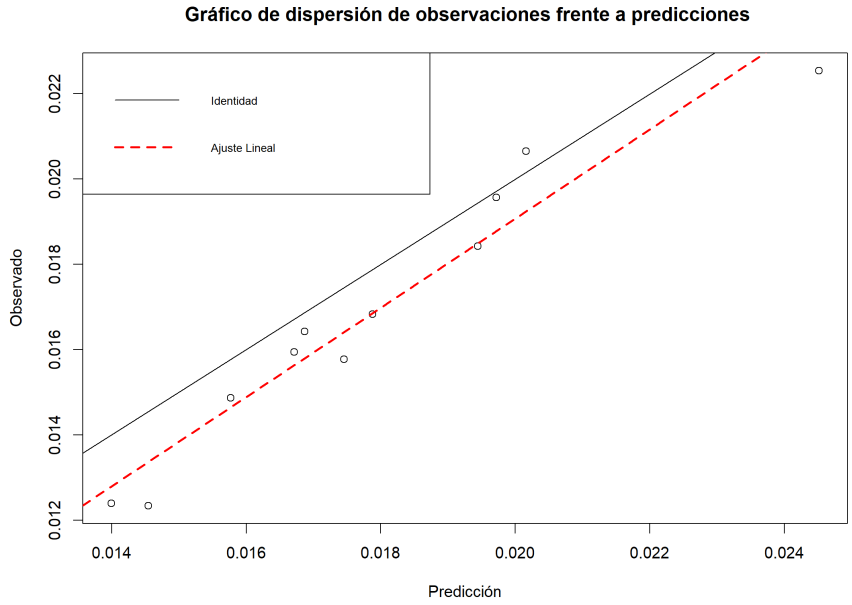

Figura 5.14: Gráfico de dispersión de observaciones vs predicciones.

Boosting

Para ajustar el modelo Boosting, la primera acción que se tomó fue emplear un enfoque combinado de búsqueda en cuadrícula y validación cruzada, con el fin de encontrar los hiperparámetros óptimos para el modelo. Este es un apartado crucial debido a que la selección adecuada de estos hiperparámetros afectan directamente a la eficacia del modelo final. Para nuestro modelo, se define una cuadrícula de hiperparámetros que incluyen los valores que se observan en el cuadro 5.6. En este proceso se ha utilizado el paquete `caret` (Kuhn, 2008), utilizando el argumento `tuneGrid`.

Cuadro 5.6: Valores de parámetros para el modelo boosting.

Parámetro	Valores
Profundidad de Interacción	1, 3, 5
Número de Árboles	100, 500, 1000
Tasa de Aprendizaje	0.01, 0.1
Número Mínimo de Observaciones en Nodos Terminales	5, 10

Además se considera la variación del parámetro `bag.fraction`, debido a que al tener un conjunto de

datos pequeño debemos tener en cuenta la posibilidad de aumentar la fracción de datos utilizados para ajustar cada árbol para tratar de que el proceso de aprendizaje sea mejor. Debido a que este hiperparámetro no se puede optimizar de forma directa utilizando el paquete `caret` se ajusta manualmente a través de un vector de valores predefinido y empleando un bucle donde se itera cada combinación de hiperparámetros de la rejilla para un valor concreto de bag.fraction.

Los valores que se han testeado de `bag.fraction` son: c(0.5,0.6,0.7,0.8). Finalmente nos quedamos con aquella combinación de hiperparámetros que presente el menor RMSE, que en este caso es la combinación compuesta por los valores que aparecen en el Cuadro 5.6 y el código mostrado a continuación.

`interaction.depth = 3`, `n.trees = 500`, `shrinkage = 0.01`, `n.minobsinnode = 5` y `bag.fraction = 0.8`

Esta combinación de hiperparámetros sobre la muestra de entrenamiento devuelve un $RMSE = 0.0012$ y un R^2=0.9335.

En lo que respecta a la importancia de las variables, como se puede ver representado en la Figura 5.15, podemos observar que la variable SPREAD es la más importante en el modelo suponiendo el 76.21 % mientras que la variable SEASONING aporta al modelo el 23.78 % restante. Esto sugiere que cambios en la variable SPREAD tendrán un impacto más significativo en la salida del modelo que cambios en SEASONING.

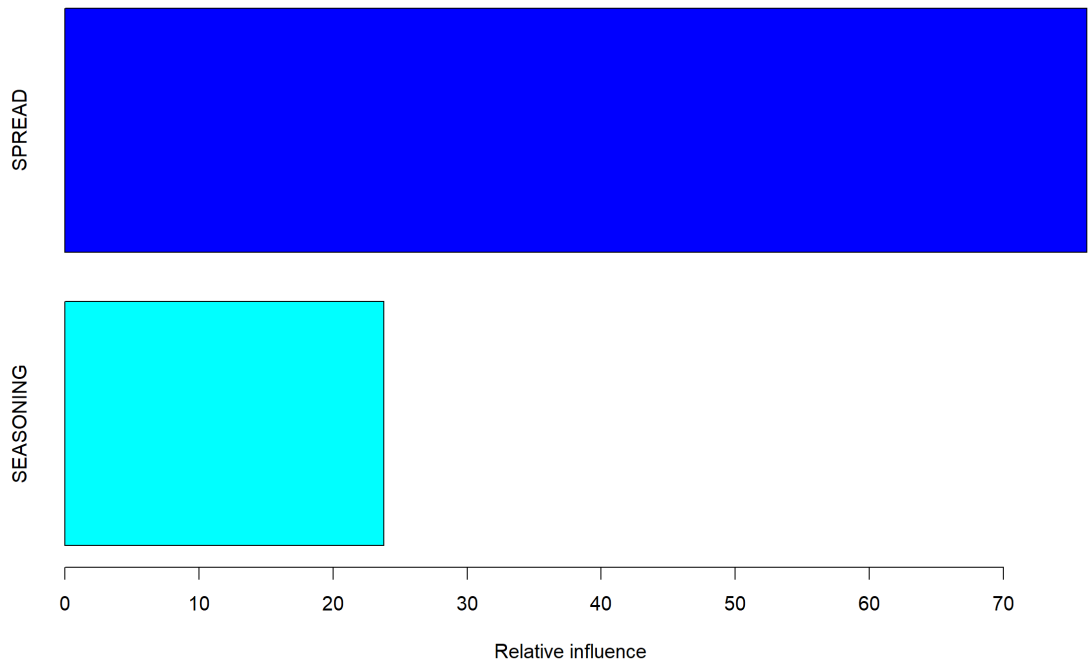

Figura 5.15: Influencia relativa de las variables en el modelo GBM.

Por otro lado, se obtienen los gráficos de dependencia parcial, los cuales quedan representados en la Figura 5.16, que muestra cómo las predicciones del modelo cambian en función de una sola variable predictora, manteniendo el resto constante. Observamos que para el gráfico de la variable SPREAD, la respuesta ER aumenta para mayores valores de la variable explicativa hasta cierto punto donde

parece que se estabiliza (entre 1 y 2.5 aproximadamente). A partir de ese punto, vuelve a aumentar. Esto nos sugiere que para valores extremos de SPREAD la tasa de cancelación aumenta y para valores intermedios parece mantenerse constante y no verse afectada. Para el segundo gráfico (SEASONING) se aprecia que hay un incremento en la respuesta del modelo cuando tenemos valores más altos de SEASONING, aunque la relación no es lineal y muestra ciertos 'escalones' o incrementos discretos. Esto podría indicar que la variable SEASONING tiene un efecto escalonado en la respuesta del modelo, con ciertos rangos de valores de SEASONING que producen predicciones similares.

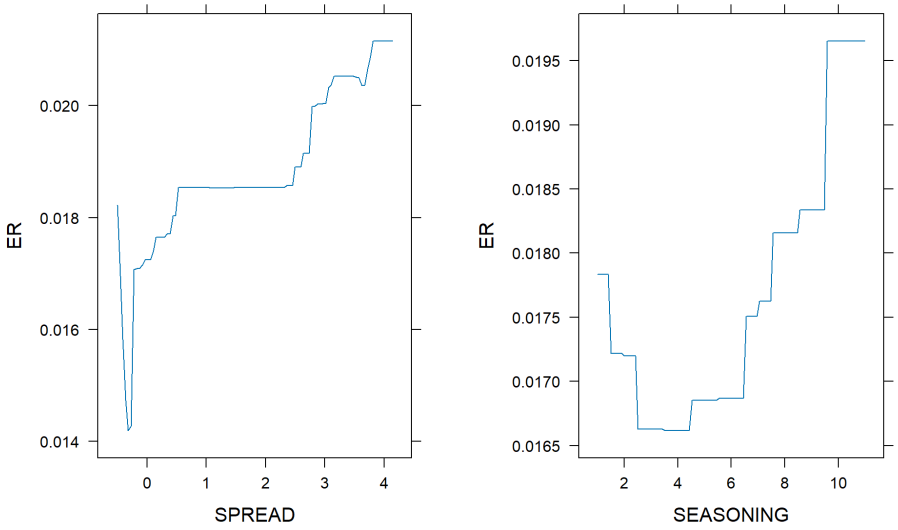

Figura 5.16: Dependencia parcial de las variables.

Pasando a la evaluación de resultados en la muestra de test, obtenemos los resultados expuestos en el cuadro 5.7:

Modelo	ME	RMSE	Rsquared
BOOSTING	-0.0009276	0.0012011	0.8460480

Cuadro 5.7: Métricas de rendimiento o desempeño para el modelo boosting.

Observamos nuevamente que, el error medio vuelve a ser negativo y cercano a cero, lo cual sugiere que el modelo tiende a sobreestimar. El RMSE es de un 0.12 %, lo cual indica un error reducido en las predicciones del modelo. El R^2 es del 84.60 %, lo cual indica que aproximadamente el 84 % de la variabilidad en la variable respuesta puede ser explicada por las variables explicativas del modelo. Estos resultados arrojan que el modelo parece consistente y robusto a la hora de realizar predicciones.

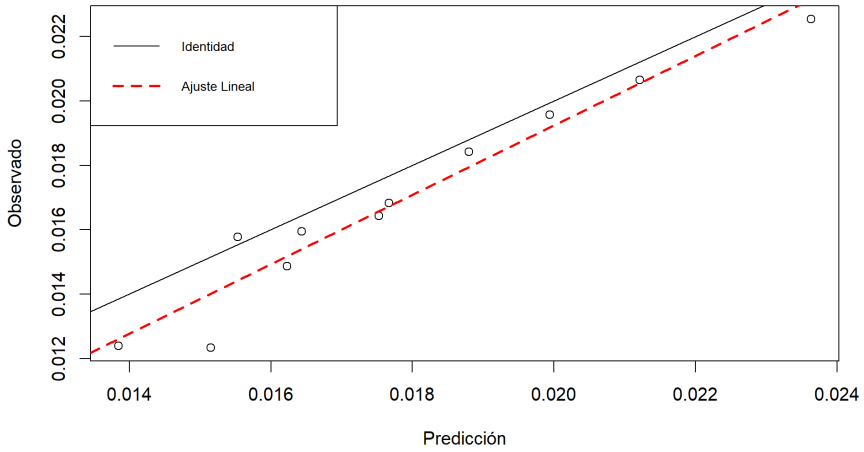

Figura 5.17: Gráfico de dispersión de observaciones vs predicciones.

En la Figura 5.17 se puede comprobar que efectivamente el modelo tiende a sobreestimar en todos los puntos del modelo. Sin embargo, también se muestra una menor diferencia entre la bisectriz del primer cuadrante y la recta de regresión construida a partir de los pares (observación, predicción), además de una menor dispersión de los puntos alrededor de la recta de regresión. Esto es indicativo de un mejor desempeño del modelo boosting con respecto a los modelos lineales y GAM.

SVM

Por último, se ajusta el modelo de Máquina de Soporte Vectorial. Para ello, es importante seleccionar con cuidado el hiperparámetro "C" y el tipo de kernel, que en este caso se utilizó el kernel radial y para este tipo de kernel es importante seleccionar con precisión el hiperparámetro sigma, ya que la elección óptima de estos influirá en la eficienia y rendimiento del modelo final. El hiperparámetro "C" es un parámetro de penalización de los términos de error, concretamente es un término de regularización que controla el equilibrio entre lograr un margen grande y asegurar que cada ejemplo de entrenamiento prediga correctamente. Respecto al kernel radial transforma los datos a un espacio donde las distancias radiales respecto a puntos fijos (centros) son relevantes. Por último, el hiperparámetro "sigma" controla la "suavidad" de la frontera de decisión generada por el modelo. Para ajustar de forma óptima estos hiperparámetros (σ y "C"), se genera una rejilla de valores utilizando el paquete `caret` y se utilizará la validación cruzada con 5 pliegues para seleccionar aquella combinación de hiperparámetros que arrojen un menor RMSE.

Los valores que se han tenido en cuenta para la rejilla son: C = c(0.5, 1, 5, 10) y σ = c(0.01, 0.1, 1, 10).

En base a estos valores de los parámetros C y sigma, se ajusta el modelo y se obtienen los resultados expuestos en la Figura 5.18.

```
Support Vector Machines with Radial Basis Function Kernel

43 samples
 2 predictor

Pre-processing: centered (2), scaled (2)
Resampling: Cross-Validated (5 fold)
Summary of sample sizes: 35, 34, 35, 34, 34
Resampling results across tuning parameters:

  C     sigma  RMSE         Rsquared   MAE
   0.5   0.01  0.002778113  0.6048296  0.002252768
   0.5   0.10  0.002015600  0.6576963  0.001608693
   0.5   1.00  0.002000631  0.6651898  0.001490541
   0.5  10.00  0.002486389  0.4930314  0.002023574
   1.0   0.01  0.002527636  0.6097805  0.002062137
   1.0   0.10  0.001958297  0.6651207  0.001549720
   1.0   1.00  0.002024814  0.6337150  0.001556597
   1.0  10.00  0.002477778  0.4227395  0.002034179
   5.0   0.01  0.002176185  0.6125323  0.001703496
   5.0   0.10  0.001879870  0.6684196  0.001443153
   5.0   1.00  0.002190137  0.5621176  0.001730674
   5.0  10.00  0.002463035  0.4404320  0.002091409
  10.0   0.01  0.002153935  0.6184902  0.001679203
  10.0   0.10  0.001877416  0.6695500  0.001451869
  10.0   1.00  0.002100090  0.6026625  0.001630380
  10.0  10.00  0.002518940  0.4124115  0.002114863

RMSE was used to select the optimal model using the smallest value.
The final values used for the model were sigma = 0.1 and C = 10.
```

Figura 5.18: Modelo óptimo SVM

Observamos que el modelo que minimiza el RMSE es aquel que tiene un valor de $C = 10$ y $\sigma = 0{,}1$ correspondiéndole un RMSE $= 0.001877$ y un $R^2 = 0.6695$.

Con estos resultados, junto con el modelo ajustado, podemos probar el modelo sobre la muestra de test, obteniendo los resultados que se observan en el cuadro 5.8.

Cuadro 5.8: Métricas de rendimiento o desempeño para el modelo SVM.

Modelo	ME	RMSE	Rsquared
SVM	-0.0010047	0.0025912	0.4124331

El modelo actual muestra un RMSE cercano a cero, indicando una baja discrepancia entre las predicciones y los valores reales. Sin embargo, su coeficiente de determinación R^2 es considerablemente bajo, apenas alcanzando el 41 %. Este valor sugiere que el modelo explica menos de la mitad de la variabilidad en la variable de respuesta. A pesar de ajustar diferentes kernels y modificar sus hiperparámetros, los resultados no han mejorado significativamente. Además, para descartar la posibilidad de que estos resultados fueran producto de la aleatoriedad, se llevaron a cabo 100 iteraciones en la segunda fase del estudio, las cuales confirmaron la persistencia de este bajo rendimiento.

En vista de esta situación, se considera la posibilidad de enriquecer el modelo con más información. Se propone incorporar una nueva variable, específicamente la variable COSECHA, con el objetivo de mejorar la capacidad explicativa y predictiva del modelo. Seguidamente, se vuelve a ajustar el modelo teniendo en cuenta los mismos hiperparámetros que en el primer modelo y se chequean los resultados que aparecen en la Figura 5.19

```
Support Vector Machines with Radial Basis Function Kernel

43 samples
 3 predictor

Pre-processing: centered (3), scaled (3)
Resampling: Cross-Validated (5 fold)
Summary of sample sizes: 35, 34, 35, 34, 34
Resampling results across tuning parameters:

  C     sigma  RMSE         Rsquared   MAE
   0.5   0.01  0.002611259  0.6315740  0.002103857
   0.5   0.10  0.001848354  0.7365127  0.001439921
   0.5   1.00  0.001521975  0.8402366  0.001093206
   0.5  10.00  0.002715560  0.6853617  0.002127536
   1.0   0.01  0.002304495  0.6434015  0.001834566
   1.0   0.10  0.001674725  0.7659527  0.001289992
   1.0   1.00  0.001358431  0.8566912  0.000971540
   1.0  10.00  0.002362465  0.7041506  0.001874630
   5.0   0.01  0.002034673  0.6657308  0.001639669
   5.0   0.10  0.001482282  0.8166986  0.001138330
   5.0   1.00  0.001395779  0.8356654  0.001012971
   5.0  10.00  0.002260220  0.7167402  0.001799971
  10.0   0.01  0.002001952  0.6803182  0.001591328
  10.0   0.10  0.001478168  0.8183367  0.001143137
  10.0   1.00  0.001404102  0.8328656  0.001023697
  10.0  10.00  0.002260220  0.7167402  0.001799971

RMSE was used to select the optimal model using the smallest value.
The final values used for the model were sigma = 1 and C = 1.
```

Figura 5.19: Modelo óptimo SVM, incluyendo previamente la variable COSECHA.

Para este modelo se escoge nuevamente el que minimiza el RMSE, siendo este el que tiene un valor de $C = 1$ y $\sigma = 1$, arrojando una RMSE=0.0013584 y un $R^2 = 0.85669$.

Respecto a los resultados que se representan en el cuadro 5.9 relativo a la muestra test se aprecia que este modelo presenta un error medio negativo y muy próximo a cero, lo cual indica que el sesgo es casi inexistente. El RMSE también es muy pequeño, medido en términos de porcentaje, debido a que como se ha mencionado en apartados anteriores esta métrica depende de la medida en que se encuentre la variable respuesta, es de 0.10 %, lo cual es un resultado muy positivo pues es un error reducido. En esa misma línea se observa que el R^2 es mucho más alto que en el primer modelo que se había ajustado, arrojando un valor del 88.75 %, lo cual indica que aproximadamente el 88 % de la variabilidad en la variable respuesta puede ser explicada por las variables explicativas del modelo. Los resultados obtenidos del modelo de regresión SVM añadiendo una variable más, mejoran considerablemente, por lo que es el modelo óptimo para llevar a la segunda etapa de este procedimiento.

Modelo	ME	RMSE	Rsquared
SVM	-0.0001223	0.0010263	0.8875892

Cuadro 5.9: Métricas de rendimiento o desempeño para el modelo SVM.

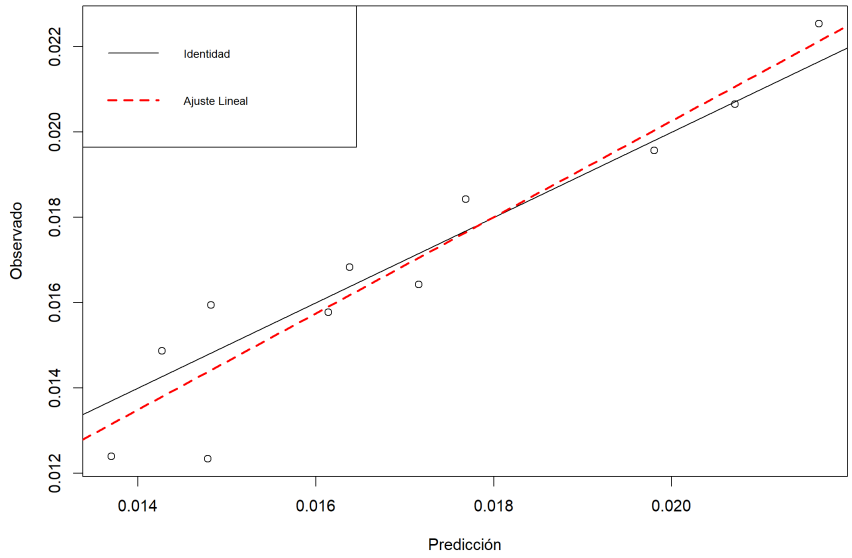

Figura 5.20: Gráfico de dispersión de observaciones vs predicciones.

En el gráfico 5.20 observamos que efectivamente el modelo tiende a sobreestimar para valores pequeños de tasas de cancelación pero que también parece que infraajusta para valores grandes de ER. En todo caso, la recta de regresión obtenida a partir de los pares (observado, predicción) está muy próxima a la bisectriz del primer cuadrante, lo que es indicativo del buen compartamiento predictivo del modelo.

COMPARATIVA DE MODELOS

Con todos estos resultados se procede a hacer una comparativa entre los cuatro modelos para determinar cuál de todos es el mejor modelo a la hora de predecir la tasa de cancelación anticipada. Los resultados pueden verse en el Cuadro 5.10.

Cuadro 5.10: Comparativa de métricas de rendimiento o desempeño en la muestra test para los modelos LM, GAM, BOOSTING y SVM.

Modelo	ME	RMSE	Rsquared
LM	-0.0007949	0.0019743	0.5840509
GAM	-0.0007881	0.0018321	0.6418026
BOOSTING	-0.0009276	0.0012011	0.8460480
SVM	-0.0001223	0.0010263	0.8875892

A la vista de estos resultados se concluye que el mejor modelo tanto en términos de error medio, RMSE y R^2 es el modelo SVM. Le sigue el modelo Boosting con un R^2 del 84 % y un RMSE de 0.12 %, situando a cuntinuación el modelo GAM y, por último, el LM, aunque es importante tener en cuenta que el modelo SVM tiene tres variables predictoras por lo que el modelo es más complejo y difícil

de interpretar que el moodelo Boosting, al que le corresponde un R^2 tan solo 4 puntos porcentuales menor, por lo que, tendiendo en cuenta criterios combinados de simplicidad y bondad de ajuste, será interesante trabajar con el modelo Boosting, más sencillo e interpretable.

5.2. Segunda etapa

Una vez se han evaluado los modelos para una única muestra de entrenamiento y test, se repite este procedimiento 100 veces con el objetivo de determinar si los resultados que se apreciaban en el apartado anterior son fruto de la aleatoriedad o si por el contrario efectivamente son resultados reproducibles independientemente de la muestra a partir de la cual se tomen.

Los resultados son los que se pueden apreciar en los Cuadros 5.11y 5.12, en concreto se ha calculado el valor medio y la desviación típica (a partir de 100 muestras) correspondientes a cada una de las medidas de bondad de ajuste.

Cuadro 5.11: Resumen de métricas de precisión para muestras de entrenamiento.

Métrica / Modelo	LM	GAM	Boosting	SVM
ME (media)	1.67e-18	-2.00e-18	1.59e-06	3.45e-05
RMSE (media)	0.0017	0.0016	0.0008	0.0006
R^2 (media)	0.6810	0.8617	0.9297	0.9523
ME (Std)	4.83e-18	2.41e-17	7.02e-06	7.69e-05
RMSE (Std)	0.0001	9.09e-05	7.88e-05	0.0003
R^2 (Std)	0.0417	0.0217	0.0130	0.0456

Cuadro 5.12: Resumen de métricas de precisión para muestras de test.

Métrica / Modelo	LM	GAM	GBM	SVM
ME (media)	$-2{,}84 \times 10^{-5}$	$-1{,}10 \times 10^{-6}$	$5{,}10 \times 10^{-5}$	$6{,}06e-05 \times 10^{-5}$
RMSE (media)	0.0018	0.0016	0.0014	0.0013
R^2 (media)	0.6058	0.6966	0.7198	0.7796
ME (Std)	0.0006	0.0005	0.0005	0.0005
RMSE (Std)	0.0004	0.0003	0.0003	0.0003
R^2 (Std)	0.1949	0.1568	0.1689	0.1402

A la vista de los resultados, atendiendo al valor del R^2 correspondiente al modelo de SVM, se concluye que éste presenta el mejor desempeño, ya sea en la muestra de test como en la de entrenamiento.

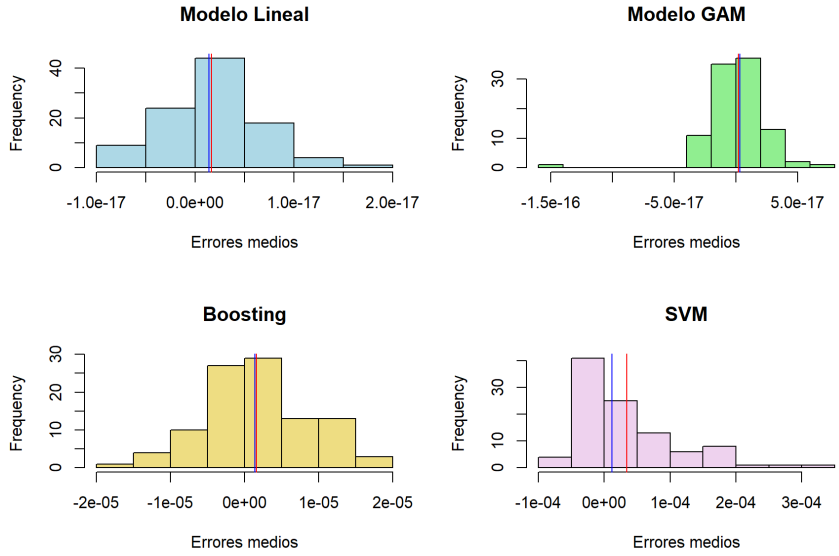

Figura 5.21: Histogramas del error medio para los distintos modelos en la muestra de entrenamiento.

Para la muestra de entrenamiento, SVM tiene el menor RMSE (0.0006) y el mayor R^2 (0.9523). En la muestra de test, también presenta el menor RMSE (0.0013) y un R^2 más alto (0.7796). Debido a esto, en primera instancia podríamos determinar que el mejor modelo para predecir la tasa de cancelación anticipada es el SVM, pero debemos recordar que fue necesario incluir una variable adicional para que el modelo presentara resultados óptimos, lo cual indica añadir una mayor complejidad al modelo. En base a esto, y viendo los valores de tanto el coeficiente de determinación como del RMSE son similares a los correspondientes al modelo boosting, deberemos considerar si la ligera pérdida en capacidad predictiva se justifica por la mayor simplicidad y mejor interpretabilidad de un modelo más sencillo.

Se han representado también para una mejor comprensión de los resultados los histogramas, tanto de la muestra de entrenamiento como de la muestra test, de los errores medios, del RMSE y del R^2, en los que la línea roja vertical indica la media de los datos representados en cada histograma y la línea azul la mediana. Los histogramas que recogen los resultados de las muestras de entrenamiento son los que se representan en las Figuras 5.21, 5.22 y 5.23 mientras que los histogramas para la muestra test se muestran en las Figuras 5.24, 5.25 y 5.26. Estos resultados nos ayudan a tener una comprensión más fidedigna de los resultados, pudiendo hacernos una idea de cómo están distribuidos los valores de las diferentes métricas y observando que en líneas generales la media no siempre es un indicador fiable de rendimiento para estos índices, dada su susceptibilidad a ser distorsionada por valores extremos. Es por esto por lo que sería interesante considerar no solo la media si no otras medidas descriptivas que puedan reflejar con más fidelidad la distribución de los datos. Por ello se representan los boxplot de cada una de las métricas para cada uno de los modelos, tanto para la muestra de entrenamiento como para la muestra test, con el objetivo de tener una perspectiva aún más completa del comportamiento de nuestras predicciones ya que esta herramienta gráfica ofrece una visión más detallada de la distribución, incluyendo la mediana, los cuartiles y los valores atípicos, que son fundamentales para entender la dispersión y la asimetría de los datos. Con toda esta información podremos tener y configurar una idea clara y precisa de cuál es el mejor modelo atendiendo a todas las casuísticas planteadas en este apartado.

Los boxplots para las muestras de entrenamiento están representados en las Figuras 5.27, 5.28 y 5.29 y de las cuales podemos extraer las conclusiones que se muestran a continuación. Específicamente,

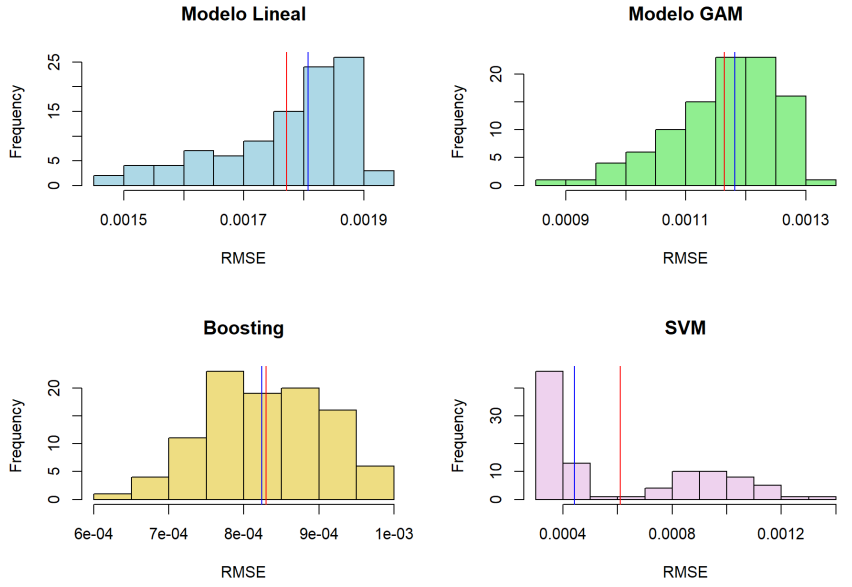

Figura 5.22: Histogramas de RMSE para los distintos modelos en la muestra de entrenamiento.

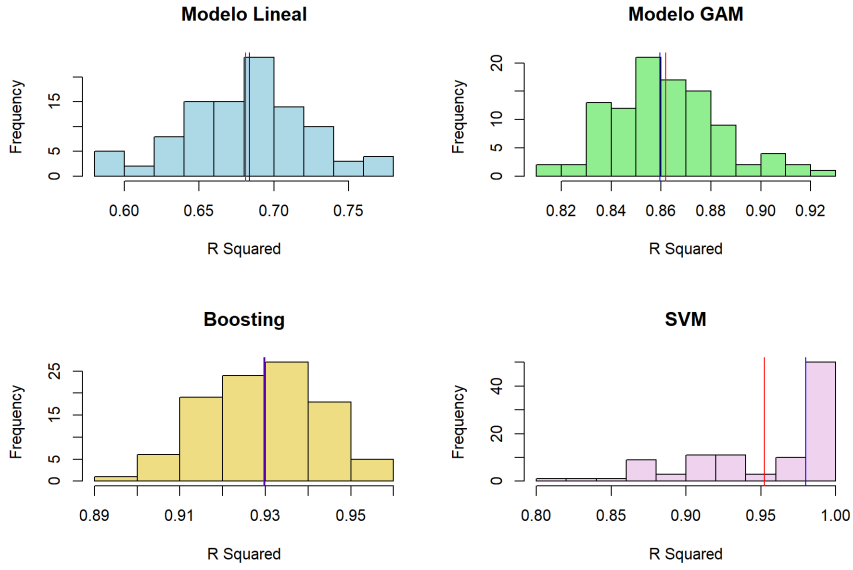

Figura 5.23: Histogramas de R Squared para los distintos modelos en la muestra de entrenamiento.

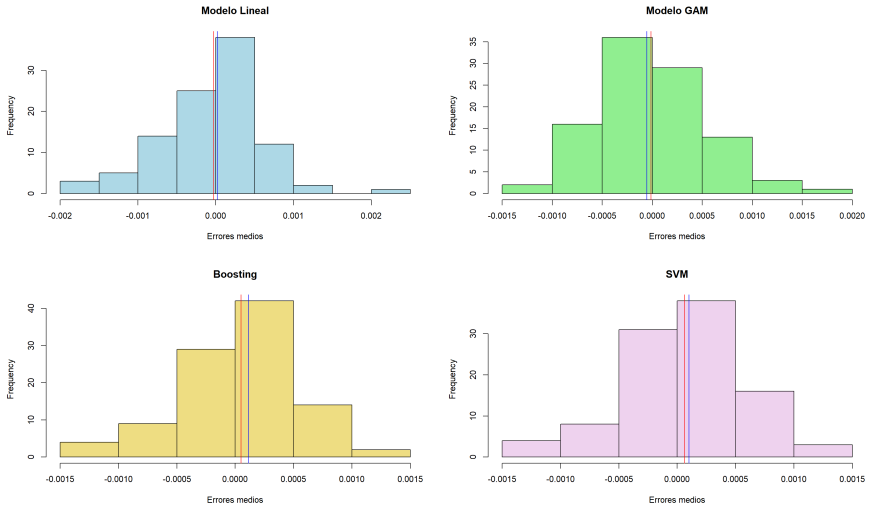

Figura 5.24: Histogramas del error medio para los distintos modelos.

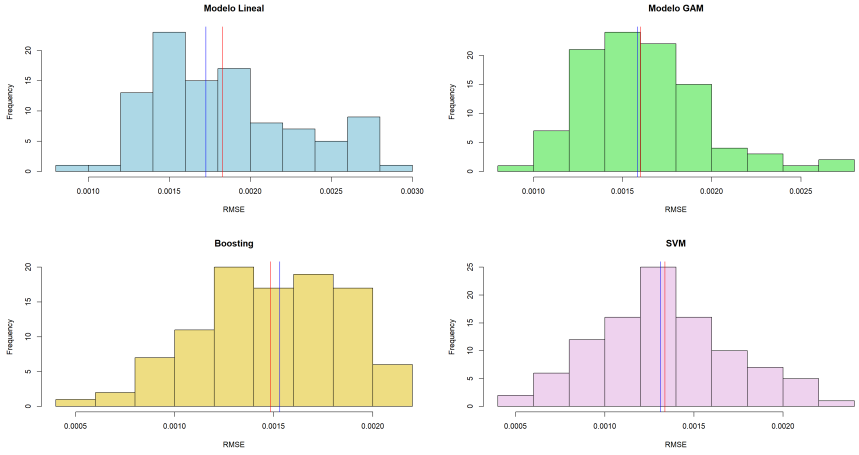

Figura 5.25: Histogramas de RMSE para los distintos modelos.

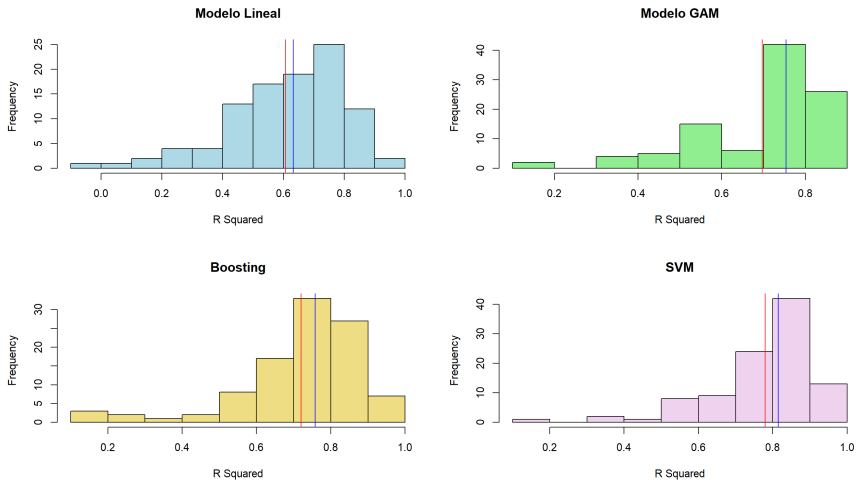

Figura 5.26: Histogramas del coeficiente de determinación para los distintos modelos.

el SVM muestra el error medio más alto y menos consistente presentando una mayor variabilidad, mientras que el GAM, GBM y LM muestran un sesgo bajo y con una variabilidad mínima.

Por otro lado, el modelo SVM tiene la mediana de RMSE más baja, lo que indica que tiende a tener errores de predicción menores en más de la mitad de los casos en comparación con los otros modelos. Por otro lado, el modelo LM muestra la mediana de RMSE más alta, lo que sugiere que tiene errores de predicción generalmente mayores. Los modelos GBM y GAM tienen medianas de RMSE que están entre las del SVM y LM, indicando un rendimiento intermedio en términos de errores de predicción. Además el modelo SVM, aunque tiene la mediana de RMSE más baja, también muestra una variabilidad mayor en los errores, como se refleja en la longitud de su caja. Esto sugiere que mientras sus errores pueden ser bajos en muchos casos, también hay casos donde los errores son significativamente más altos.

Además, el modelo SVM tiene la mediana de R^2 más alta, lo que indica que es el que mejor se ajusta a los datos del conjunto de entrenamiento. Por otro lado, el modelo LM muestra la mediana de R^2 más baja, lo que sugiere que es el que tiene el peor ajuste de los cuatro modelos. Por otra parte, parece que el modelo GAM tiene el mayor rango intercuartítico seguido del modelo SVM. Por su parte el modelo boosting tiene un rango intercuartílico muy pequeño, presentando por lo tanto una variabilidad pequeña en sus resultados.

Las Figuras 5.30, 5.31 y 5.32 muestran los boxplots de la distribución de las métricas de la muestra test donde podemos apreciar que la mediana de los errores medios está en todos los modelos en torno a cero presentando todos una variabilidad similar. Asimismo, en términos de RSME nuevamente el modelo SVM presenta los mejores resultados, presentando la menor mediana aunque en términos de variabilidad, el modelo más estable es el modelo GAM. En lo que respecta al R^2, la mediana del modelo SVM es con diferencia superior al resto de modelos y presenta la menor variabilidad, por ello podemos concluir que es un modelo bastante estable en términos de coeficiente de determinación presentando valores bastante consistentes.

En este caso, en términos de mediana, los modelos GAM y boosting se encuentran equiparados en términos de coeficiente de determinación aunque el modelo GAM presenta claramente una mayor variabilidad en sus resultados que el modelo boosting. En lo que respecta a la comparativa entre el modelo GAM y el SVM si bien en términos de media podíamos plantearnos escoger el modelo boosting en detrimento del SVM por ser un modelo más sencillo y presentar en media tan solo una diferencia aproximada del 6 %, si atendemos al valor mediano, el cual podemos ver con más precisión en las tablas

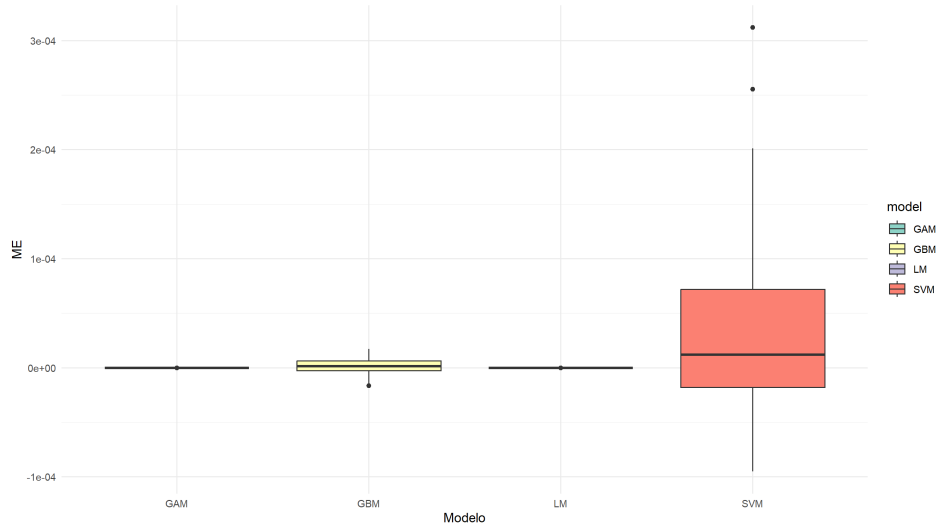

Figura 5.27: Boxplots del error medio de los modelos para las muestras de entrenamiento.

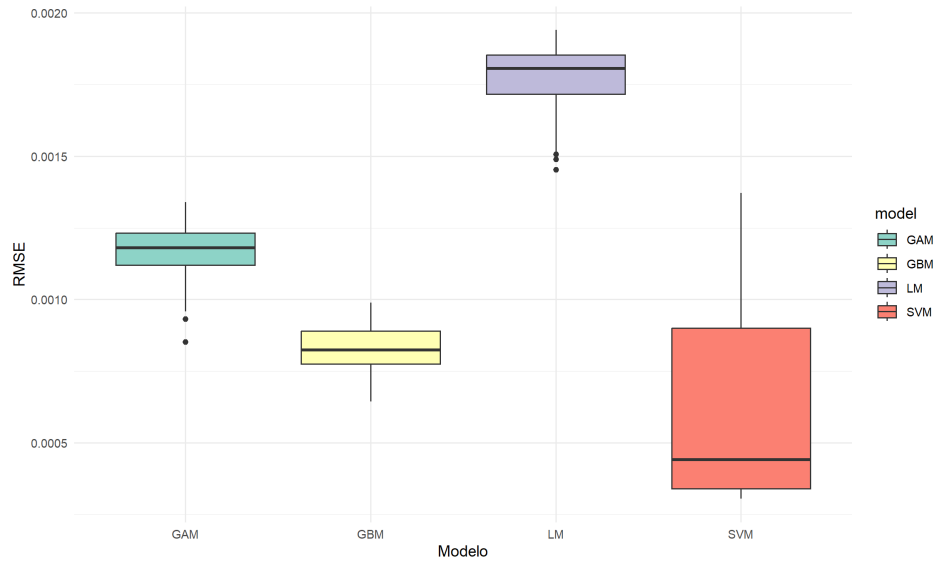

Figura 5.28: Boxplots del RMSE de los modelos para las muestras de entrenamiento.

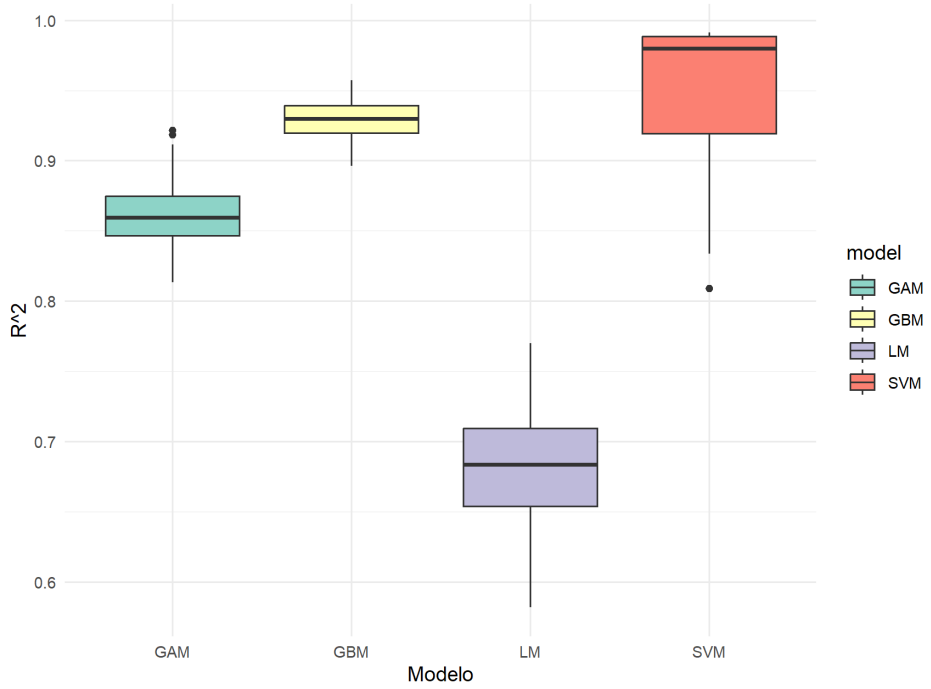

Figura 5.29: Boxplots del R^2 de los modelos para las muestras de entrenamiento.

5.15,5.16 la diferencia con el modelo SVM sigue siendo de aproximadamente del 6 %, pero es un modelo más variable que el SVM, lo cual hace replantearnos la posibilidad de escoger el modelo boosting como el mejor modelo para predecir las cancelaciones anticipadas, ya que el SVM parece arrojar predicciones más precisas.

Por último presenta en los Cuadros 5.13, 5.14, 5.15,5.16 se muestran los resúmenes numéricos correspondientes a las métricas obtenidas para los distintos modelos, cuyos valores son coherentes con respecto a las conclusiones obtenidas a partir de la información de las Figuras 5.21 - 5.32.

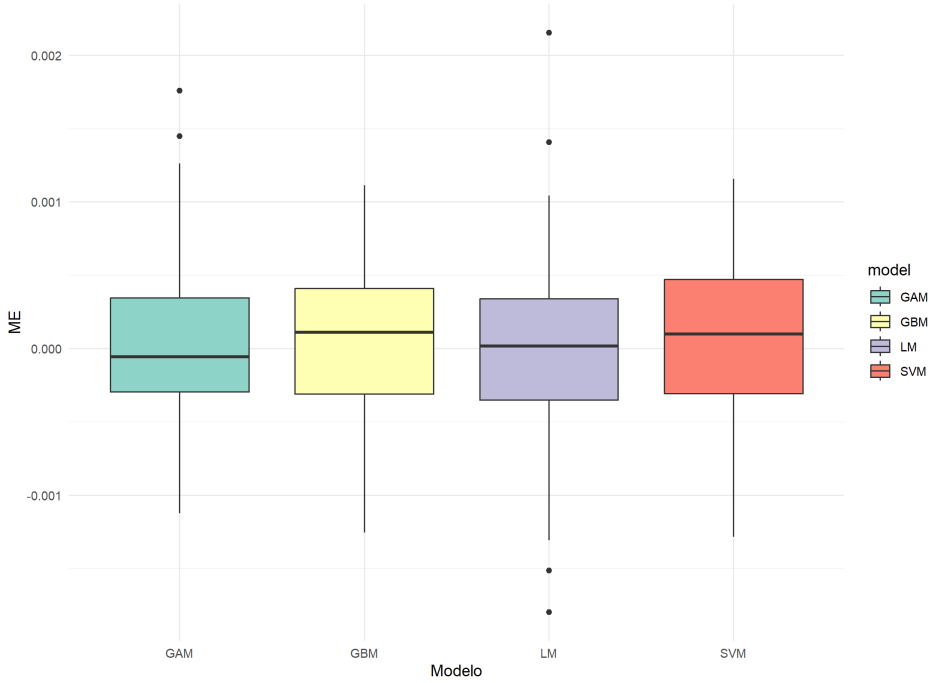

Figura 5.30: Boxplots del error medio de los modelos para las muestras test.

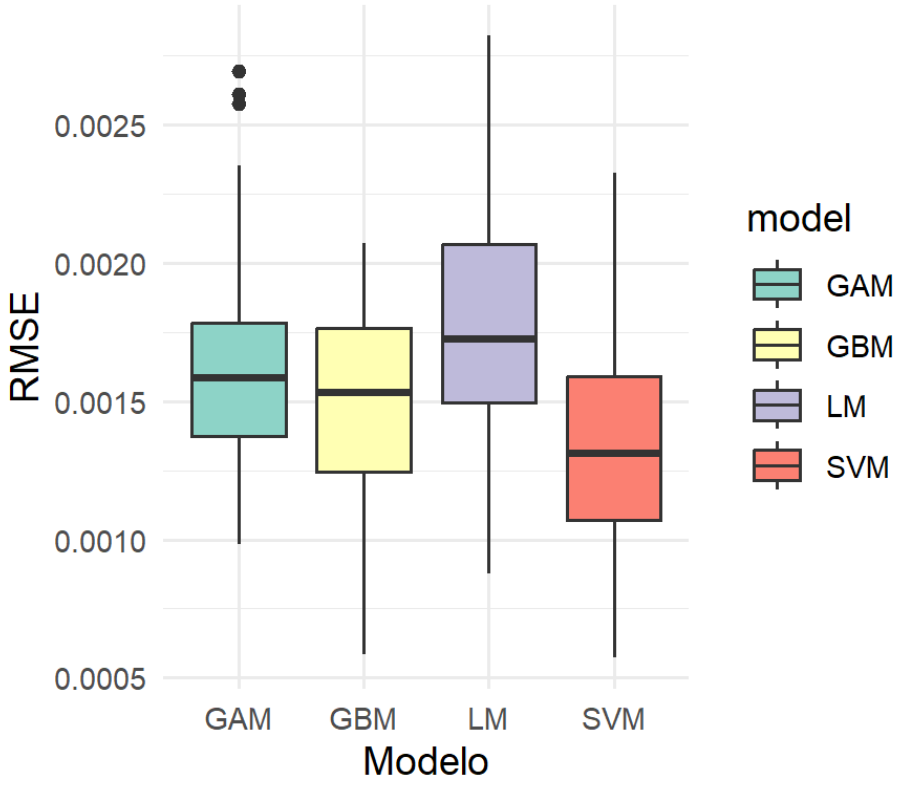

Figura 5.31: Boxplots del RMSE de los modelos para las muestras test.

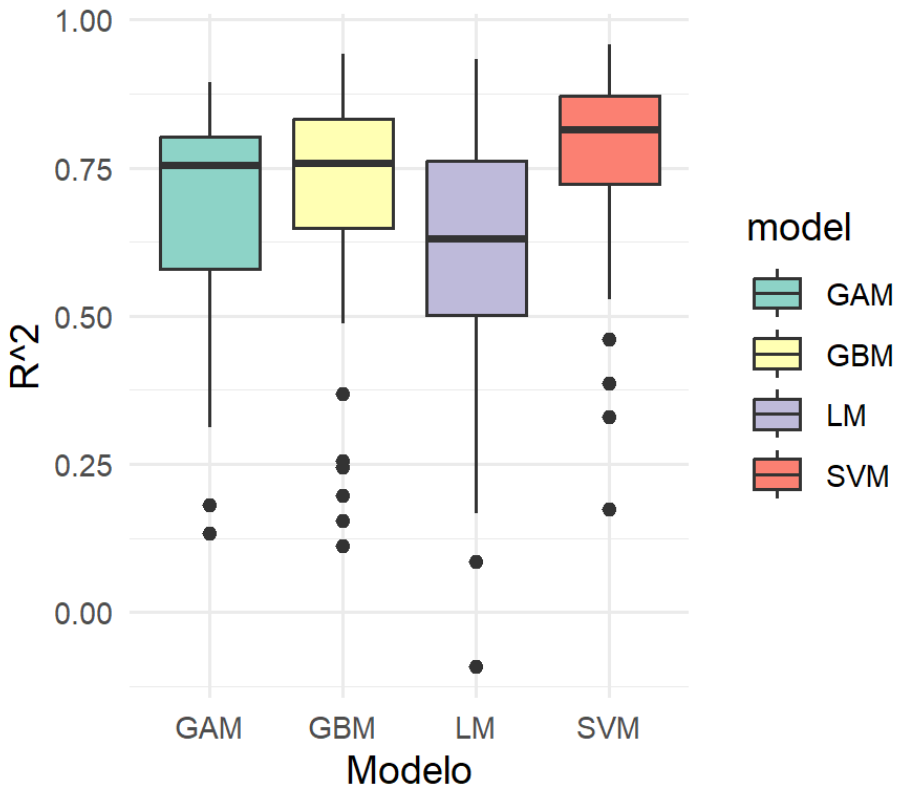

Figura 5.32: Boxplots del R^2 de los modelos para las muestras test.

Cuadro 5.13: Resumen de las métricas del modelo lineal para los conjuntos de test y entrenamiento.

Métrica	Mínimo	1er Cuartil	Mediana	Media	3er Cuartil	Máximo
RMSE (Train)	0.001453	0.001716	0.001807	0.001771	0.001853	0.001941
ME (Train)	-8.674e-18	-1.926e-18	1.432e-18	1.673e-18	4.569e-18	1.529e-17
R Squared (Train)	0.5819	0.6538	0.6837	0.6810	0.7093	0.7701
RMSE (Test)	0.000876	0.001495	0.001726	0.001830	0.002066	0.002819
ME (Test)	-0.001795	-0.0003485	0.00001849	-0.00002845	0.0003401	0.002153
R Squared (Test)	-0.09194	0.50111	0.63153	0.60583	0.76224	0.93395

Cuadro 5.14: Resumen de las métricas del modelo GAM para los conjuntos de test y entrenamiento.

Métrica	Mínimo	1er Cuartil	Mediana	Media	3er Cuartil	Máximo
RMSE (Train)	0.000852	0.001120	0.001181	0.001164	0.001231	0.001340
ME (Train)	-1.563e-16	- 9.077e-18	3.389e-18	2.001e-18	1.538e-17	6.277e-17
R Cuadrado (Train)	0.8134	0.8465	0.8595	0.8618	0.8746	0.9218
RMSE (Test)	0.0009829	0.0013723	0.0015856	0.0016005	0.0017822	0.0026918
ME (Test)	-1.122e-03	- 2.935e-04	-5.436e-05	-1.107e-05	3.463e-04	1.758e-03
R Cuadrado (Test)	0.1327	0.5797	0.7543	0.6966	0.8028	0.8953

Cuadro 5.15: Resumen de las métricas del modelo GBM para los conjuntos de test y entrenamiento.

Métrica	Mínimo	1er Cuartil	Mediana	Media	3er Cuartil	Máximo
RMSE (Train)	0.0006439	0.0007740	0.0008237	0.0008297	0.0008901	0.0009889
ME (Train)	-1.637e-05	-2.569e-06	1.432e-06	1.592e-06	6.209e-06	1.706e-05
R Cuadrado (Train)	0.8961	0.9196	0.9299	0.9297	0.9393	0.9571
RMSE (Test)	0.0005838	0.0012442	0.0015311	0.0014832	0.0017648	0.0020699
ME (Test)	-0.001254	-0.0003102	0.0001126	0.00005106	0.0004116	0.001114
R Cuadrado (Test)	0.1123	0.6482	0.7580	0.7198	0.8327	0.9422

Cuadro 5.16: Resumen de las métricas del modelo SVM para los conjuntos de test y entrenamiento.

Métrica	Mínimo	1er Cuartil	Mediana	Media	3er Cuartil	Máximo
RMSE (Train)	0.0003039	0.0003392	0.0004420	0.0006096	0.0008996	0.0013722
ME (Train)	-9.507e-05	-1.798e-05	1.221e-05	3.460e-05	7.176e-05	3.119e-04
R Cuadrado (Train)	0.8089	0.9193	0.9800	0.9524	0.9884	0.9912
RMSE (Test)	0.0005737	0.0010700	0.0013130	0.0013377	0.0015891	0.0023243
ME (Test)	-1.284e-03	-3.055e-04	1.015e-04	6.061e-053	4.714e-04	1.156e-03
R Cuadrado (Test)	0.1731	0.7226	0.8149	0.7796	0.8717	0.9576

Capítulo 6

Conclusiones y líneas futuras de investigación.

La gestión correcta del riesgo de tipo de interés es un problema que debe abordarse con suma cautela desde las entidades bancarias pues es algo que afecta a la gran mayoría de los productos financieros de estas entidades. Una mala gestión y previsión por parte de la entidad del comportamiento de los tipos de interés y de cómo este puede afectar a los distintos productos financieros que ofertan, puede conllevar a serios problemas y consecuencias nada deseables para la empresa. Es por esto que, debido al aumento de los tipos de interés que se han producido en los últimos años, la entidad financiera ABANCA detectó la necesidad de implementar mecanismos para predecir las cancelaciones en los depósitos a plazo como parte de la gestión del riesgo de tipo de interés. Hasta el momento no disponían de una herramienta que controlara este tipo de riesgo en esta clase de productos, debido a que durante un amplio periodo de tiempo, los tipos habían permanecido cercanos a cero, e incluso en negativo. Es por este motivo por el cual, hasta que no se ha producido este repunte en los tipos de interés, estos productos no eran demasiado populares, ya que tienen condiciones más restrictivas que una cuenta corriente y a muchos usuarios no les compensa mantener su dinero inactivo durante el periodo que dure el contrato del depósito a cambio de una remuneración nula.

Una vez detectado por parte de la entidad esta necesidad se propone crear un "Early Redemption Model"para poder predecir y detectar las cancelaciones anticipadas de este tipo de producto financiero. Para ello:

- Se realizó un preprocesamiento de los datos, incluyendo la limpieza y depuración de la muestra eliminando datos atípicos, datos faltantes, delimitación del perímetro, así como la generación de nuevas variables tras la detección de posibles factores de riesgo como puede ser la estacionalidad o el tiempo de vida del depósito.
- Se realiza un análisis exploratorio de los mismos para poder tener una mejor comprensión de estos y se procede a realizar un estudio de correlación entre variables con el objetivo de identificar las variables que a posteriori se utilizarán para la creación de los modelos de regresión.
- Partición de la muestra tras sospechar que las cancelaciones anticipadas podrían comportarse de maneras diferentes atendiendo a diferentes categorías. Para ello, se realizaron test estadísticos no paramétricos, con el objetivo de identificar diferencias estadísticamente significativas entre grupos, determinando que, efectivamente, se apreciaban diferencias estadísticamente significativas atendiendo a la variable categórica PLAZO. Una vez confirmado esto, se procede a particionar la muestra en función del plazo y a analizar las cancelaciones anticipadas para la submuestra PLAZO = 12 meses.

- Agregación de variables en base a las necesidades de la empresa, para tener una mejor comprensión de las cancelaciones anticipadas y su evolución mensual. Concretamente, se calculó el promedio de las cancelaciones anticipadas y del SPREAD en función de las variables SEASONING y COSECHA.
- Creación de los distintos modelos de regresión (lineal multivariante, aditivos generalizados, support vector machines y boosting) para finalmente extraer las siguientes conclusiones:
  - Se confirma que el modelo estándar que se suele emplear en la industria para tratar los modelos de riesgo, a pesar de arrojar resultados positivos, presenta un amplio margen de mejora ya que se ha postulado como el peor modelo de los cuatro que se han evaluado.
  - Que los modelos de aprendizaje estadístico SVM y boosting, son las mejores alternativas para modelar el riesgo de tipo de interés aplicado a los depósitos a plazo, arrojando predicciones precisas y fiables.
  - El modelo boosting a pesar de presentar una bondad de ajuste ligeramente inferior al SVM, se presenta como un modelo más sencillo e interpretable que este último y que por ello, se debe valorar la posibilidad de elegirlo como la mejor alternativa para predecir las cancelaciones anticipadas.
  - Que ambos modelos (SVM y boosting) muestran un buen desempeño cuando el experimento se repite 100 veces, dejando patente que esos resultados no son fruto de la aleatoriedad, sino que son resultados reproducibles con independencia de la muestra a partir de la cual se tomen.

A la vista de estos resultados se abre una vía de investigación en lo que se refiere al uso de modelos de aprendizaje estadístico en el área de los modelos de cancelación anticipada y de riesgo de tipo de interés, los cuales se pueden ampliar no solo a los depósitos a plazo sino, también, a las cancelaciones anticipadas de hipotecas, cancelaciones anticipadas de créditos o cancelaciones anticipadas de otros tipos de instrumentos financieros en los cuales el tipo de interés juegue un papel fundamental. Además se cumplen los objetivos del trabajo, ya que parece que se han encontrado métodos alternativos al normalmente implementado en el sector, siendo capaces de dar predicciones confiables de las cancelaciones anticipadas.

Atendiendo a problemas no resueltos, encontramos como limitación de estos modelos su implementación en los softwares de gestión de riesgo de tipo de interés. Existe ámbito de mejora en las herramientas utilizadas por la entidad para permitir la implementación de modelos de aprendizaje estadístico.

También sería interesante volver a realizar este estudio contando con una ventana temporal mayor de la que se ha trabajado en esta memoria, con el objetivo de tener una muestra más grande de datos y poder evaluar de nuevo el desempeño de los distintos modelos con un volumen mucho mayor de información, así como evaluar el desempeño de estos modelos para las submuestras PLAZO = 3 meses y PLAZO = 6 meses con el objetivo de determinar si los resultados obtenidos para la submuestra estudiada en la presente memoria, son extrapolables a diferentes plazos.

Bibliografía

[1] Benigno, P., Canofari, P., Di Bartolomeo, G., & Messori, M. (2023). The ECB?s new inflation target from a short-and long-term perspective. Journal of Policy Modeling, 45(2), 286-304.

[2] Bernhard E. Boser, Isabelle M. Guyon, and Vladimir N. Vapnik. 1992. A training algorithm for optimal margin classifiers. In Proceedings of the fifth annual workshop on Computational learning theory (COLT '92). Association for Computing Machinery, New York, NY, USA, 144?152. https://doi.org/10.1145/130385.130401

[3] Bissiri, Matteo and Cogo, Riccardo, Modeling Behavioral Risk (November 12, 2014). Available at SSRN: https://ssrn.com/abstract=2523349 or http://dx.doi.org/10.2139/ssrn.2523349

[4] Breiman, L. (1996). Bagging predictors. Machine learning, 24, 123-140. https://doi.org/10.1145/130385.130401

[5] Drucker, H., Burges, C. J., Kaufman, L., Smola, A., y Vapnik, V. (1997). Support Vector Regression Machines. Advances in Neural Information Processing Systems, 9, 155-161.

[6] Fernández, R., Costa, J., & Oviedo, M. (2021). Aprendizaje estadístico. https://rubenfcasal.github.io/ aprendizaje estadistico

[7] Freund, Y. (1995). Boosting a weak learning algorithm by majority. Information and Computation, 121(2), 256-285.

[8] Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. The Annals of Statistics, 1189-1232. https://doi.org/10.1214/aos/1013203451

[9] Grömping U (2006) Relative importance for linear regression in R: the package relaimpo. J Stat Softw 17(1):1

[10] Hastie, T., & Tibshirani, R. (1986). Generalized Additive Models. Statistical Science, 1(3), 297-310. https://doi.org/10.1214/ss/1177013604

[11] H. B. Mann. D. R. Whitney. .ºn a Test of Whether one of Two Random Variables is Stochastically Larger than the Other..ªnn. Math. Statist. 18 (1) 50 - 60, March, 1947. https://doi.org/10.1214/aoms/1177730491

[12] James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). An Introduction to Statistical Learning: With Applications in R (2nd ed.). Springer.

[13] Kearns, M. J., & Valiant, L. G. (1989). Cryptographic limitations on learning Boolean formulae and finite automata. Journal of the ACM (JACM), 41(1), 67-95.

[14] Kruskal, W. H., y Wallis, W. A. (1952). Use of Ranks in One-Criterion Variance Analysis. Journal of the American Statistical Association, 47, 583-621. http://dx.doi.org/10.1080/01621459.1952.10483441

[15] Kuhn, M., & Johnson, K. (2016). Applied Predictive Modeling. (5ª impresión). [Springer]. https://doi.org/10.1007/978-1-4614-6849-3

[16] Kuhn, M. (2008). Building Predictive Models in R Using the caret Package. Journal of Statistical Software, 28(5), 1?26. https://doi.org/10.18637/jss.v028.i05

[17] Lans, R. F. V. D. (1992). An SQL Guide to Oracle. Addison-Wesley Longman Publishing Co., Inc..

[18] Lindeman RH, Merenda P, Gold R (1980) Introduction to bivariate and multivariate analysis. Scott Foresman, Glenview

[19] Maggi, F., Natale, A., Pepanides, T., Risso, E., & Schröck, G. (2017). IFRS 9: A silent revolution in banks? business models. New York: McKinsey & Company.

[20] Martín Guareño, J. J. (2016). Support Vector Regression: Propiedades y aplicaciones [Trabajo de fin de grado]. Universidad de Sevilla.

[21] Nelder, J. A., & Wedderburn, R. W. M. (1972). Generalized linear models. Journal of the Royal Statistical Society: Series A (General), 135(3), 370-384. Oxford University Press.

[22] Pardo.F, Juan Carlos. (2023). Curso de Contrastes de Especificación. Máster Universitario en Técnicas Estadísticas.

[23] Pearson K (1896) VII. Mathematical Contributions to the Theory of Evolution.-III. Regression, Heredity, and Panmixia. Philosophical Transactions of the Royal Society A, 187, 253-318. https://doi.org/10.1098/rsta.1896.0007

[24] Rao, C. R., Toutenburg, H., Shalabh, & Heumann, C. (2007). Linear Models and Generalizations: Least Squares and Alternatives (3rd ed.). Springer.

[25] Richardson, Alice. (2015). Nonparametric Statistics: A Step-by-Step Approach. International Statistical Review.

[26] Schapire, R. E. (1990). The strength of weak learnability. Machine Learning, 5(2), 197-227.

[27] Valiant, L. G. (1984). A theory of the learnable. Communications of the ACM, 27(11), 1134-1142.

[28] Vapnik, V., Guyon, I., & Hastie, T. (1995). Support vector machines. Mach. Learn, 20(3), 273-297.

[29] Vilar, Jose A, (2022). Métodos no paramétricos. Máster Universitario en Técnicas Estadísticas.

[30] Walsh, C. E. (2022). Inflation surges and monetary policy. Bank of Japan, Institute for Monetary and Economic Studies.

[31] Wilson, L. (2022). Toxic asset subsidies and the early redemption of TALF loans. International Journal of Financial Studies, 10(2), 23.

[32] Wood, S. N. (2007). Generalized Additive Models: An Introduction with R. Chapman and Hall/CRC.

Apéndice

A continuación se muestra el nombre y breve descripción de cada una de las variables estudiadas en el presente estudio.

NUC: Identificador del contrato.

SUBNUC: Subidentificador del contrato.

ID: Identificador conjunto del contrato (NUC-SUBNUC).

FECHA CONSTITUCION: Fecha de constitución del contrato.

FECHA VENCIMIENTO: Fecha de vencimiento del contrato.

FECHA MOVIMIENTO: Fecha de movimiento de saldo (retirada / ampliación / renovación / constitución / ...).

FECHA NACIMIENTO: Fecha de nacimiento del cliente.

PRODUCTO ID: Identificador del tipo producto.

INTERÉS: Tipo de interés medio al que se remunera el depósito.

TIPO MOVIMIENTO: Clase de movimiento efectuado:

- NUEVA IMPOSICIÓN = 0-INUE

- IMPOSICIÓN RENOVACIÓN = 0-IREN

- NO RENOVACIONES = 1-NREN

- REINTEGROS = 1-REIN

- REINTEGRO RENOVACIÓN =1-RREN

- CANCELACIONES = 2-CANC

TIPO PERSONA: Indica el tipo de persona que ha contratado el depósito:

- 0 = Jurídica

- 1 = Física

- 2 = Otros

SECTOR ID: Indica los distintos tipos de sectores al que pertence el cliente.

IMPORTE: Importe en euros del movimiento.

SALDO: Saldo / nominal tras la entrada / salida de efectivo por movimiento.

CANCELACIÓN: Retiradas de dinero con cancelación del producto.

REINTEGRO: Retiradas de dinero antes de la fecha de vencimiento.

TOTAL: CANCELACION + REINTEGRO.

SALDO ANT: Saldo / nominal tras la entrada / salida de efectivo por movimiento anterior.

ER: Tasa porcentual de precancelación.