Facultad de Ciencias Económicas y Empresariales

MACHINE LEARNING Y RIESGO DE CRÉDITO

Autor: Jaime Grau Álvarez
Director: José Portela González

MADRID | Abril 2020

Resumen

El objetivo de este trabajo es analizar como podemos medir o predecir el riesgo de crédito hipotecario mediante técnicas de Machine Learning.

Para ello se realizará una aproximación a lo que conocemos como técnicas Machine Learning. Este término hace referencia al aprendizaje automático por el cual, mediante algoritmos, se pueden analizar gran cantidad de datos y obtener información relevante para la toma de decisiones en cualquier ámbito.

Por otro lado, se buscará realizar una aproximación al riesgo de crédito, analizando las causas de la todavía reciente crisis económica de 2007 y la influencia que tuvo el riesgo de crédito hipotecario en la consecución de Defaults crediticios en esa época.

Se realizará un análisis de los principales métodos de cálculo de riesgo de crédito, desde los más simples hasta aquellos que involucren técnicas más avanzadas.

Por último, se realizará un caso de estudio, en el que se seleccionaron un conjunto de créditos hipotecarios a los que se les aplicó una regresión logística, un árbol de decisión y un Gradient Boosting, para ver cuál de ellas nos proporcionaba un mejor análisis desde el punto de vista empresarial.

Palabras clave

Machine Learning, riesgo de crédito, modelo Logit, regresión logística, árbol de decisión, Gradient Boosting.

Abstract

The objective of this paper is to analyze how we can measure or predict mortgage credit risk using Machine Learning techniques.
For this purpose, we will make an approach to what we know as Machine Learning techniques. This term refers to automatic learning by which, through algorithms, a large amount of data can be analyzed, and relevant information obtained, for decision making in any area.
On the other hand, an approach to credit risk is made, analyzing the causes of the still recent economic crisis of 2007 and the influence that mortgage credit risk had on the consecution of credit Defaults at that time.
An analysis of the main credit risk calculation methods will be carried out, from the simplest to those involving more advanced techniques.
Finally, a case study will be conducted, in which a set of mortgage loans were selected to which a logistic regression, a decision tree and a Gradient Boosting were applied, to see which of them provided us with a better analysis from a business point of view.

Key words

Machine Learning, credit risk, Logit model, logistic regression, decision tree, Gradient Boosting.

ÍNDICE

Índice de figuras

Índice de tablas

Índice de ecuaciones

1. INTRODUCCIÓN

1.1. Justificación.

El ser humano ha ido evolucionando a lo largo de la historia gracias a pequeños hechos concretos que han supuesto grandes pasos. El fuego, la rueda, el comercio. Sin embargo, en el último siglo, los avances en el computación y en todo lo que la rodea, proporcionalmente, han revolucionado nuestro modo de vida. Desde finales del siglo XVIII, cuando autores como Gauss o Legendre, introdujeron el concepto de regresiones lineares, hasta la actualidad en donde se han desarrollado sistemas artificiales capaces de imitar comportamientos humanos e incluso mejorarlos.

Por otro lado, hoy 24 de abril, nos encontramos ante una situación sin precedentes, en la que el mundo se encuentra en confinamiento por culpa del Covid-19. Hace un mes, el índice español Ibex35, cayó un 14.06% (Expansión), la mayor caída de la historia de la bolsa. Hace diez años, el mundo entero sufrió una de las peores recesiones económicas que se recuerdan, dando lugar a una situación sin precedentes que no solo destrozó la economía social, si no que su impacto a nivel social fue desastroso. El detonante de esta fue el estallido de la burbuja inmobiliaria y la concesión de hipotecas basura denominadas Subprime, elevando el riesgo de crédito a niveles nunca vistos y aumentando los números de Default en los créditos hipotecarios hasta números estratosféricos.

Este trabajo recoge los aspectos más relevantes de estos dos mundos. Por un lado, se describe todo el proceso de transformación de la Inteligencia Artificial hasta llegar a lo que conocemos hoy como Machine Learning y algunas de las técnicas empleadas para el caso de estudio. Por otro lado, se describe lo acontecido en la recesión del 2007 además de los principales métodos de cálculo del riesgo de crédito. Finalmente, se pone en común la teoría de los dos campos para ver como las técnicas de Machine Learning nos pueden ayudar para calcular ese riesgo de crédito hipotecario y predecir las situaciones de Default de los prestatarios.

1.2. Objetivos

El objetivo de este trabajo es, en primer lugar, analizar la evolución de la Inteligencia Artificial hasta los métodos actuales de cálculo Por otro lado, entender el concepto de riesgo de crédito bajo la descripción de los hechos sucedidos en la crisis de 2007. Además, se estudiarán algunos modelos de cálculo y algunas técnicas que serán empleadas en este trabajo. Por último, se realizará un caso de estudio en donde se aplicarán algunos modelos de Machine Learning para predecir los Defaults de los prestatarios de una cartera de créditos hipotecarios, datada antes de la crisis.

1.3. Metodología

Este trabajo se abordó de dos formas. La primera involucró el método de investigación, por el cual se buscó información acerca de la historia de la Inteligencia Artificial y del Machine Learning, para entender como ha ido evolucionando hasta nuestros días. Se investigó acerca del riesgo de crédito y de lo acontecido en el año 2007, estudiando las causas que produjeron la recesión, así como las consecuencias derivadas del gran número de Defaults en los créditos hipotecarios. A su vez también se investigó acerca de los modelos del cálculo del riesgo de crédito y algunas técnicas como el árboles de decisión o Gradient Boosting que son empleadas en el presente caso de estudio, sin entrar en el desarrollo matemático. Por otro lado, la metodología escogida para el último apartado del trabajo fue el estudio de caso. Este método de investigación se define como una investigación empírica que investiga un fenómeno contemporáneo dentro de su contexto de la vida real; cuando los límites entre el fenómeno y el contexto no son evidentes; y en el que se usan múltiples fuentes de datos (Yin, 1989). La principal aportación de este método es que permite dar respuestas a situaciones contemporáneas y que pueden facilitar el entendimiento de realidades complejas como puede ser el Machine Learning. Se analizó un conjunto de créditos hipotecarios con fecha anterior a la recesión. El conjunto de datos fue analizado mediante el programa Gretl y el lenguaje de programación R, cuyo código se nos fue proporcionado por el Profesor Portela.

1.4. Estructura

El trabajo se divide en cuatro grandes bloques. El primero hace referencia al concepto de Inteligencia Artificial y a su evolución, diferenciándose de las técnicas de Machine Learning y resaltando sus aplicaciones en actividades habituales en el día a día.

El segundo bloque lo conforma el análisis histórico de la recesión de 2007, el proceso de securitización, los Acuerdos de Basilea y algunos tipos de riesgo de créditos relacionados con el objeto de estudio. Se trata de dar a conocer como influyó el riesgo de crédito en la crisis de 2007 y como después de los Acuerdos de Basilea se ha venido regulando de forma más estricta. Es necesario definir qué tipos de riesgo de crédito podemos encontrarnos ya que el objeto de estudio de este trabajo se centra en el riesgo de crédito hipotecario.

El tercer bloque lo compone algunos métodos de cálculo del resigo de crédito. Por otra parte, se describen las técnicas de regresión logística o modelo Logit, los árboles de decisión y el Gradient Boosting, sin explicar su desarrollo matemático.

El último bloque y piedra angular del caso lo conforma el caso de estudio. Se analizará un conjunto de créditos para ver la aplicación de las técnicas expuestas en el ámbito empresarial. Se indicará que técnica nos ofrece mejores resultados y cuál puede ser más útil a la hora conocer la probabilidad de Default en función de los valores de las características del crédito.

2. MARCO CONCEPTUAL: BIG DATA, MACHINE LEARNING Y APLICACIONES

Procedemos a analizar primeramente el marco global del Machine Learning. A lo largo del trabajo nos referiremos a diversas técnicas que se recogen en este, sin embargo, es necesario resaltar las diferencias con lo que se conoce como Inteligencia Artificial. Si bien el Machine Learning puede integrarse como un campo del primero, el funcionamiento es completamente distinto y merece un análisis histórico para ver en qué momento se configuró. Se analizará además algunas de las aplicaciones de estos campos, remarcando la importancia del análisis de los datos en actividades de nuestro día a día.

2.1. Contexto histórico

Toda la terminología que conocemos actualmente en lo que respecta al Big Data o al Machine Learning es prácticamente reciente. La entrada de la digitalización de procesos ha supuesto una revolución en el tratamiento de los datos, optimizando procesos, ahorrando costes e incluso generando beneficios para las empresas y para la sociedad. Como veremos más adelante, el uso de este tipo de herramientas permite por ejemplo a los bancos saber si uno de sus clientes va a incurrir en Default o la probabilidad de que devuelvan el préstamo en los términos acordados. Sin embargo, ambas ciencias tienen un origen común, la Inteligencia Artificial (IA el resto del trabajo) [1].

Para poder entender cómo surge la IA es necesario entender cuáles fueron los autores matemáticos que sirvieron como base para los posteriores estudios. El origen como tal de la IA y posteriormente el del Statistical Learning o Machine Learning proviene de las regresiones lineales. De los primeros estudios, destacan los de Legendre y Gauss: El origen del método de los mínimos cuadrados se empelaba por entonces para cálculos astronómicos como demuestra el estudio *Nuevos métodos para la determinación de las órbitas de los*

[1] A lo largo del presente trabajo nos referiremos como Default a la situación de impago por parte del hipotecado/deudor/prestatario del préstamo concedido.

cometas (Legendre, 1805) [2], siendo Gauss quién introdujo conceptos de Estadística para la aplicación de dicho método en su estudio *Theoria motus corporum coelestium* (Gauss, 1809).

Un siglo más tarde, adentrándonos en la neurociencia y separándonos de la rama científica, en 1943, Pitts y McCulloch, matemático y neurólogo respectivamente, publicaron su trabajo *A logical calculus of the ideas immanent in nervous activity*. Por primera vez se describía el concepto de Inteligencia Artificial bajo la descripción del comportamiento de redes neuronales. En palabras de los propios autores:

> *"El comportamiento de sucesivas redes neuronales se puede describir por medio de la lógica de proposición, con la adición de medios lógicos más complicados para las redes que contienen círculos; y que para cualquier expresión lógica que satisface ciertas condiciones, se puede encontrar una red de comportamiento en la manera que describe. Para cada comportamiento neto bajo una suposición, existe otra red que se comporta en función de la otra y da los mismos resultados, aunque quizás no al mismo tiempo*. (Pitts & McCulloch, 1943).

Se definió, por lo tanto, el concepto de Red Neuronal, y como se relacionaban ciertos parámetros de comportamientos de unas Redes con otras. Este estudió nos permitió conocer que, bajo la determinación de ciertos parámetros y circunstancias, las Redes neuronales eran capaces de analizar los datos y tomar decisiones, en otras palabras, las Redes Neuronales eran capaces de pensar, analizar y racionalizar como seres humanos (Pitts & McCulloch, 1943).

Este fundamento explica el razonamiento del presente trabajo, cómo una empresa, mediante sus algoritmos o redes neuronales artificiales analiza las características de un sujeto o en nuestro caso un crédito y toma una decisión o predice un resultado en base a ellas.

[2] Traducción de *Nouvelles méthodes pour la détermination des orbites des comètes*.

Años más tarde, Turing o Samuel consiguieron trasladar esta toma de decisiones neurales a nivel humano, a máquinas, definiendo que las máquinas eran capaces de tomar propias decisiones y que eran en sí mismas inteligentes. Destacan del primer autor, la elaboración del Test de Turing, en el que se planteaba si una máquina era capaz pensar por sí misma bajo la resolución de unas determinadas preguntas[3] (Turing, 1964).

Del segundo autor destaca la creación de un software que podía jugar a las damas, mejorando cada vez que jugaba una partida gracias a la memorización y al análisis de datos (Samuel, 1959), que posteriormente sirvieron de base para las grandes IA como DeepBlue que eran capaces de ganar a los mejores jugadores de ajedrez del mundo.

El lugar de referencia donde finalmente se acuñó el término Inteligencia Artificial fue en la Conferencia de Darmouth, cuya hipótesis de estudio fue: "*las bases de todos los aspectos de aprendizaje o cualquier otro atributo de la inteligencia son tan descriptibles que una máquina es capaz de realizarlos o simularlos"*. Se intentó estudiar como las máquinas eran capaces de emplear el lenguaje desde conceptos más básicos, hasta algunos más abstractos para la resolución de problemas y su propia mejora respecto a su inicial respuesta (McCarthy et al. 1955).

En relación con lo anterior y regresando al área de la computación y al razonamiento estadístico-econométrico, a comienzos de los años 70, Nelder y Wedderburn establecieron la generalización de los modelos lineares, que incluían tanto modelos de regresión lineal como la regresión logística, y como poco a poco fueron avanzando y evolucionando a lo que conocemos hoy en día como modelización, siendo esto el pilar fundamental teórico del Statiscal Learning (Nelder & Wedderburn, 1972).

Posteriormente, lo que conocemos como Ciencia de la Computación (Computer Science) evolucionó de tal forma que todos los tipos de regresiones tanto las lineares como las no lineares eran computacionalmente factibles. Destacan numerosos estudios acerca de la validación cruzada para la selección de modelos (Breiman et al. 1984) o sobre la

[3] En estudios posteriores, Steven Harnad corrigió en varios puntos el artículo publicado por Turing en 1950, llegando a la conclusión de que el verdadero objeto de estudio de Turing no era si las máquinas podían pensar por sí solas, sino si podían pensar como seres humanos. Para ver más sobre el experimento de Turing visitar: Turing, A. (1964). Computer machinery and intelligence. *Minds and Machines*.

implementación práctica de softwares para la generalización de modelos lineares. (Hastie & Tibshirani, 1987).

Tras todos estos estudios, el Statiscal Learning se ha convertido una rama más de la estadística, concentrándose en la modelización y análisis de modelos. Es innegable por otro lado que el diseño de nuevos lenguajes de programación como el R o Python han sido de gran utilidad y han tenido gran culpa en el progreso de esta rama. (James et al. 2013).

2.2. Diferencias entre Big Data y Machine Learning

Una vez establecido el nexo común o el origen de la IA y de lo que conocemos como Statistical Learning, es necesario realizar una distinción y comparación entre lo que definimos como Big Data y Machine Learning. Está claro que ambos hacen referencia al manejo de grandes cantidades de datos, tanto para su almacenamiento como para su análisis y su aplicación en problemas cotidianos. El número de datos que abarcamos hoy en día es inmanejable (McAfee et al. 2012), (Bughin et al. 2010). Estos autores en numerosos estudios proporcionaron una serie de datos sorprendentes acerca de la información que somos capaces de manejar.

- El 90% del total de datos ha sido creado en los dos últimos años
- En 2010, el número total de teléfonos móviles ascendía a 5.000 millones
- Treinta mil millones de contenidos son compartidos en Facebook en un mes.
- La biblioteca del Congreso de Estados Unidos alberga unos 235 terabytes de información.[4]

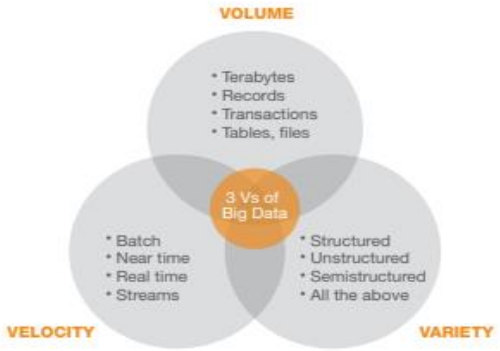
Sin embargo, la popularidad que ha ido adquiriendo el término Big Data en los últimos años nos conduce a una definición difusa y compleja. Y más aún en relación con los diferentes sistemas de análisis de datos que se han ido empleando recientemente entre los que destacan el ya mencionado Machine Learning, el BlockChain, el Data Mining, entre otros

A pesar de lo anterior, las definiciones de los autores ponen en manifiesto un punto común acerca de su definición. La mejor forma de definir al Big Data es como un conjunto de datos

[4] 1 terabyte = 1,000,000,000,000 bytes. En términos comparativos y aproximados 10 bytes = una o dos palabras.

masivos. Estos datos son empleados a gran escala para extraer nuevas percepciones o crear nuevas formas de valor (Mayer-Schonberger & Cukier, 2013). Los datos suponen por otro lado, la confluencia de una multitud de tendencias tecnologías que han irrumpido con fuerza en la sociedad generando una mayor movilidad, aumento de redes sociales, geolocalización, aumento de banda ancha, reducción en los costes de conexión y computación en la nube (Aguilar, 2016). Otros autores indagan más en las características de la información almacenada o contenida en ese Big Data, por ejemplo, Russom incluye factores como la variedad y velocidad de los datos. (Russom, 2011).

Figura 1: Características del Big Data



Fuente: Big Data analytics, Russom (2011).

Nos acercamos pues a una definición mucho más concreta que el simple almacenamiento de datos. Existe cierta tendencia a incluir el análisis de datos dentro del marco común del Big Data, sin embargo, acorde a las referencias previamente consultadas, esta asociación no sería del todo correcta, pues valoran e incluyen el análisis de datos dentro del marco común del Big Data. Tal y como se viene demostrando, una cosa es el almacenamiento de datos masivos y otra su interpretación con técnicas estadísticas o matemáticas avanzadas. Sucede lo mismo con la IA y el Big Data. Mientras que la primera se relaciona con la forma de pensar de las máquinas y su imitación a comportamientos, pensamientos y toma de decisiones humanas, lo segundo pertenece al proceso de almacenamiento de datos sobre el que acabamos de hablar (Russom, 2011).

Una vez esclarecido esto, es necesario desmarcar el concepto de Machine Learning de la definición de Big Data. El primero se conoce como la construcción y el estudio de modelos

que pueden aprender a base de datos. Los modelos construidos aprenden a base de datos objetivos y una vez analizado estos realizan predicciones, estimaciones y toman decisiones. (Provost & Kohavi, 1998). Otros autores estipulan que, en la toma de decisiones los algoritmos actúan como verdaderos expertos en lugar de seguir de manera explícita sus órdenes de programación (Bishop, 2006).

En referencia con lo anterior, el Machine Learning es un método de análisis más que de almacenamiento y entran en juego variables estadísticas y econométricas de alta complejidad matemática. Existen varios sistemas, algunos de los cuales serán introducidos más adelante. Por un lado, existen modelos complejos que eliminan cualquier posible fase exterior de interpretación externa, mientras que otros sirven de ayuda para un colaborador externo. Estos últimos son de gran utilidad pues el sujeto externo es el experto en cuestión, mientras que el algoritmo o la computadora actúa como interacción. Es la persona humana la encargada de la representación de los datos, las formas de manipulación de estos y el análisis final (Monleón-Gentino, 2015).

Tabla 1: Tipos de aprendizaje

Tipo de aprendizaje	**Descripción**	**Ejemplos**
Aprendizaje supervisado	El modelo utilizado produce una función que establece una correspondencia entre las entradas y las salidas deseadas del sistema. Este tipo de aprendizaje puede llegar a ser muy útil en problemas de investigación biológica, biología computacional y bioinformática.	Programa informático que clasifica el mail como “spam” o “no spam” Este es un problema de clasificación, donde el sistema de aprendizaje trata de etiquetar (clasificar) una serie de vectores utilizando una entre varias categorías (clases). La base de conocimiento del sistema está formada por ejemplos de etiquetados realizados anteriormente por el usuario
Aprendizaje no supervisado...	Todo el proceso de modelado se lleva a cabo sobre un conjunto de	Un robot que detecte si hay algún problema en su operación en función

	ejemplos formado tan sólo por entradas al sistema. No se tiene información sobre la variable de salida. El algoritmo tiene que ser capaz de reconocer patrones para poder etiquetar las nuevas entradas	de lo que indican los sensores que posee (temperatura, estado de la batería, etc)
Aprendizaje por refuerzo	El algoritmo aprende observando el mundo que le rodea. Su información de entrada es la retroalimentación que obtiene del mundo exterior como respuesta a sus acciones. Por lo tanto, el sistema aprende a base de ensayo-error. Hay un supervisor que da información al agente sobre si lo está haciendo bien o mal, pero no exactamente lo que debe hacer.	Robot experto que aprende del mundo exterior en base a ensayo-error.
Transducción.	Similar al aprendizaje supervisado, pero el algoritmo no construye de forma explícita una función, ya que los datos no tienen etiqueta, están sin clasificar. Se pretende pronosticar las categorías de los futuros ejemplos basándose en los ejemplos de entrada, sus respectivas categorías y los ejemplos nuevos al sistema.	Análisis automático de texto, aplicaciones de la bioinformática

Fuente: El impacto del Big-data en la Sociedad de la Información, Monleón-Getino (2015)

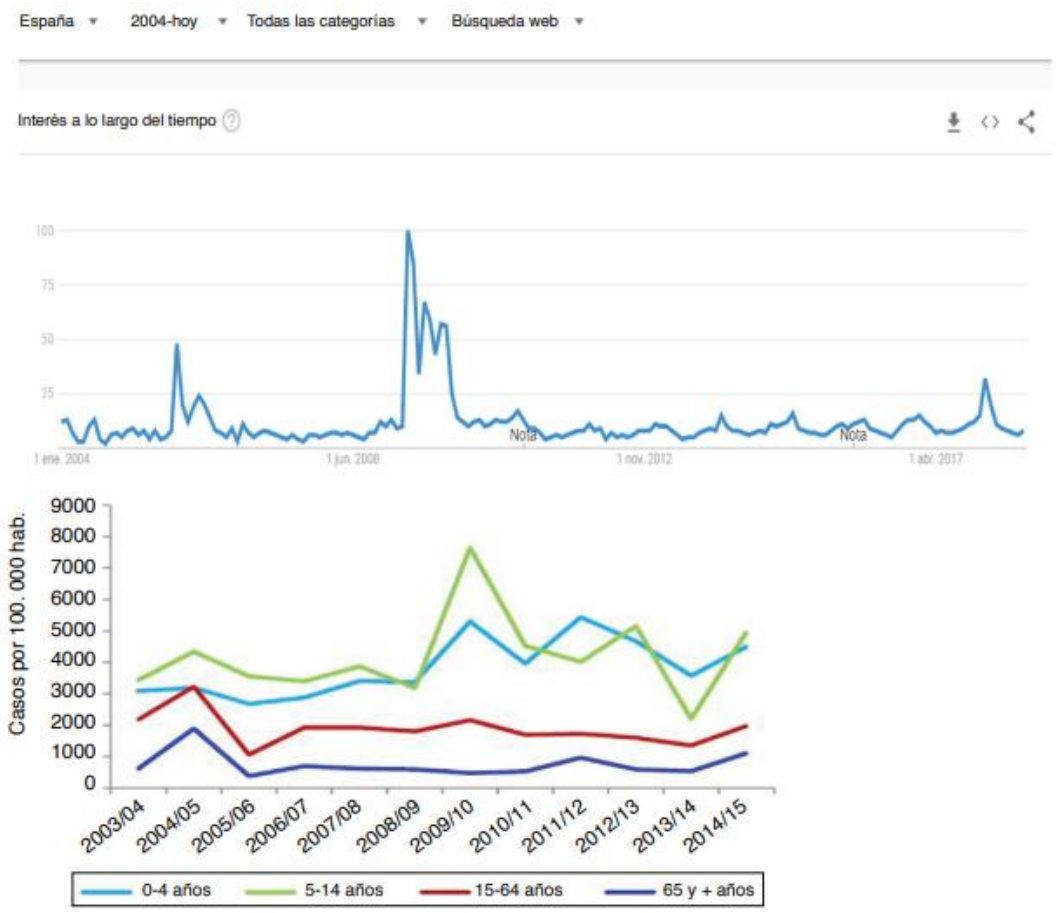
2.3. Aplicaciones prácticas: Empresas y sectores

Como ya hemos podido observar, existen números sistemas de Machine Learning, lo que nos permite adaptarlos a diferentes sectores de la sociedad. Por otro lado, existen numerosas empresas, generalmente relacionados con estos sectores que maximizan sus beneficios y optimizan sus procesos empresariales con el objetivo de reducir costes.

La mecánica y robótica es sin duda una de las ramas que más se puede beneficiar de los avances del Machine Learning. Sobre todo, en ámbitos laborales donde el trabajo de máquinas optimiza y es mucho más eficaz que el de un humano. Véase fábricas de alimentación (en procesos como el envasado, el etiquetado) o en fábricas automovilísticas. (Hinestroza, 2018). Otros campos interesantes son el reconocimiento de objetos basados en imágenes cuya aplicación es muy interesante para sectores como las cámaras, micrófonos o sistemas de geolocalización. (Contreras et al. 2016).

La medicina es posiblemente otro de los sectores más beneficiados del análisis de datos, proporcionando un beneficio directo a la sociedad. Por un lado, podemos observar en base a unos parámetros si un paciente posee una enfermedad o no. Por otro lado, todo el sistema de gestión de citas y seguimiento de los pacientes puede realizarse de forma más rápida. Se puede observar, además, donde es más necesario oftalmólogos o traumatólogos en función de las necesidades de la población de cada zona. (Hinestroza, 2018). Otro ejemplo interesante fue la investigación clínica en casos de gripe. En 2004, se predijo un aumento de casos de gripe N1H1 en función delo número de búsquedas en Google del término “gripe” (Reiz et al. 2013).

Figura 2: Relación entre los casos de Gripe con búsquedas del término en Internet (2004).



Fuente: Big Data Analysis y Machine Learning en medicina intensive, Reiz, A et al. (2013).

La Figura 2 nos muestra como a partir de la búsqueda por internet del término gripe, se podía predecir un patrón común de casos reales en la población. En la tabla de arriba se observó como el número de búsquedas para el año 2009 se disparó, coincidiendo con el repunte de casos de ese año para poblaciones jóvenes.

Las técnicas de Machine Learning se pueden aplicar prácticamente en cualquier ámbito de nuestra vida. Se han descrito dos campos esenciales, sin embargo, existen otras aplicaciones muy interesantes como son el sector público (Plazas, 2017), para la educación (Nájera & de la Calleja Mora, 2017) o para las finanzas y los negocios, como es el caso del presente trabajo.

3. RIESGO DE CRÉDITO

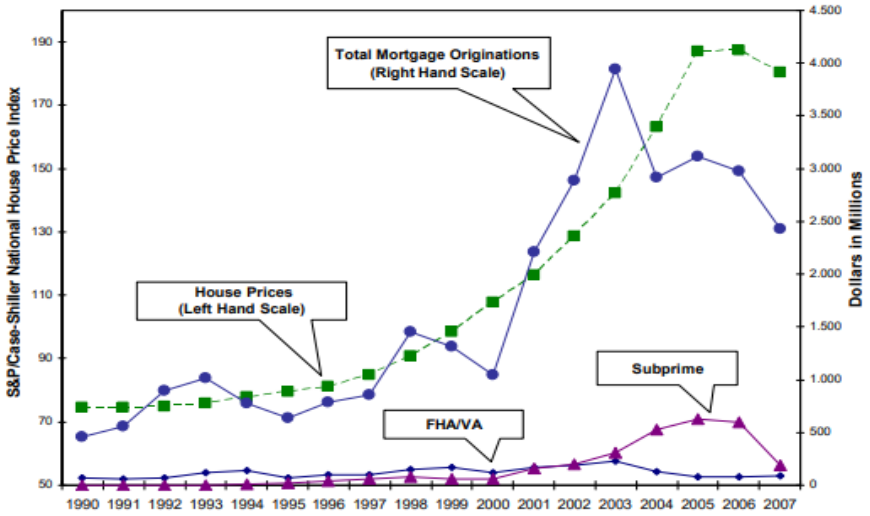
El eje del presente trabajo es el cálculo del riesgo de crédito hipotecario mediante la aplicación de técnicas de Machine Learning. Para ello es fundamental entender qué es el concepto de riesgo de crédito y riesgo de crédito hipotecario, y cómo ha ido evolucionando su importancia y método de cálculo. Es imprescindible entender lo acontecido durante la crisis del 2007 y como a partir de esta, los sistemas de medición de riesgo a nivel mundial han ido transformándose. Se analizará por lo tanto lo sucedido entre los años 2000 y 2007 y posteriormente las regulaciones impuestas por los Acuerdos de Basilea II.5 y Basilea III.

3.1. Introduccón al riesgo de crédito. Crisis financiera del 2007

El riesgo de crédito es un término cuya importancia surge a partir de la crisis de 2007. Pese a que su definición será analizada en el apartado siguiente, es necesario realizar una introducción sobre los acontecimientos ocurridos a principios del siglo XXI, pues el riesgo del que hablamos juega un papel fundamental y descriptivo de lo sucedido. Todos los sistemas modernos dedicados a la gestión del riesgo surgen como consecuencia de la de la aprobación del Acuerdo de Basilea II. Dichas disposiciones no se llegaron a implementar el año del estallido de la recesión y como consecuencia el resultado fue incluso peor de lo esperado. De este modo, es innegable el impacto que tuvo la crisis financiera de 2007 en la economía en general. Por este motivo carece de sentido hablar del riesgo de crédito o de la gestión de cualquier tipo de riesgo sin hacer especial alusión a lo acontecido en esas fechas.

Es difícil establecer una única causa objetiva como desencadenante de la crisis financiera. El punto de partida podría considerarse el mercado inmobiliario estadounidense. A comienzos del siglo XXI, los precios de los inmuebles comenzaron a crecer a un ritmo superior a la década anterior mientras que los tipos de interés se desplomaban durante los años 2002 y 2005. A fecha de 2 de enero de 2001, la Reserva Federal mantenía los tipos de interés al 6.50%, descendiendo hasta cuatro puntos porcentuales ese mismo año, alcanzando el 11 de diciembre de 2001 un 1.75% (Expansión).

Figura 3: Precio de la vivienda y número de hipotecas en Estados Unidos (1990-2007)



Fuente: S&P Case-Schiller, Inside Mortgage Finance

Sin embargo, no se pudo entender la masiva subida de precio de los inmuebles sin analizar la indiscriminada concesión de créditos hipotecarios (Figura 3). A comienzos del año 2000, las condiciones de concesión de hipotecas se ablandaron, haciendo posible la adquisición de hipotecas a familias que anteriormente jamás se les habría sido concedidas. A estas se les denominó Hipotecas Subprime, cuyas características eran más que factibles para los hipotecados pues se les permitía la devolución del principal con unos intereses muy bajos, siendo la capacidad financiera muy pobre en comparación con el importe que se les concedía. (Hull, 2012). Como consecuencia los productos financieros involucraban un riesgo muy elevado pues pese a ser factible la devolución del crédito, muchas no contaban con las suficientes fuentes de financiación para hacer frente al pago y por lo tanto el riesgo de que el deudor incurriese en Default o impago era elevado.

La facilidad de obtención crediticia impulsó la compra de viviendas generando una demanda excesivamente alta tal y como muestra la Figura 3. El aumento de la demanda inmobiliaria generó un consecuente aumento de precios para equilibrar el mercado, alcanzando en 2005 el máximo de la década. (Figura 3).

Los inversores y propietarios aprovecharon la oportunidad que les ofrecía el mercado. A mayor número de hipotecas otorgadas mayor eran los beneficios para ellos. Como a su vez los precios de la vivienda continuaban aumentando, si los hipotecados incurrían en Default,

los inversores podían ejecutar la hipoteca y hacerse con el control del inmueble de tal forma que las pérdidas sufridas se cubrían con el *collateral* o producto subsidiario[5]. El problema no residía del todo en la concesión ilimitada de hipotecas a familias sin los suficientes recursos para pagarlas si no en la contraparte y en la especulación bancaria. (Zimmerman, 2007).

A medida que subieron los precios, los primeros compradores de inmuebles no pudieron hacer frente al precio solicitado[6]. Con el objetivo de atraer más compradores, los agentes inmobiliarios, los prestamistas y los inversores continuaron con la flexibilización de las condiciones de concesión ignorando y no comprobando factores como el salario anual del hipotecado o la capacidad de devolución. Se popularizó la concesión de las ARMs (Adjustable Rate Mortgages) con tipos de interés ajustables (Pennington-Cross, 2010)[7].

Un estudio muy interesante acerca de la indiscriminada concesión de hipotecas a familias que en años anteriores jamás habrían cumplido los criterios establecidos por los bancos fue el elaborado por Mian y Sufi. De acuerdo con el estudio, antes de la crisis de 2007 los códigos Zip de las Subprime experimentaron un crecimiento en comparación con la caída de los salarios en los mismos códigos Zip (Mian & Sufi, 2009). Los códigos Zip eran o son los códigos postales definidos por zonas o estados en Estados Unidos. De esta forma demostraron que:

- La expansión del crédito hipotecario asociado a códigos Zip Subprime tenía una estrecha relación con los Defaults producidos en todo el país.
- Para entender la crisis de crédito de 2007 es necesario analizar la expansión del crédito asociado a Hipotecas Subprime en todos los barrios de todo Estados Unidos

[5] El collateral o bien subsidiario lo constituía el activo que respaldaba el crédito concedido. En caso de hipotecas el bien subsidiario suele ser el objetivo gravado por la hipoteca, para nuestro caso, la vivienda. Sin embargo, se pueden hipotecar otros productos como acciones de una compañía, o bienes del deudor que no forman parte del mismo contrato.

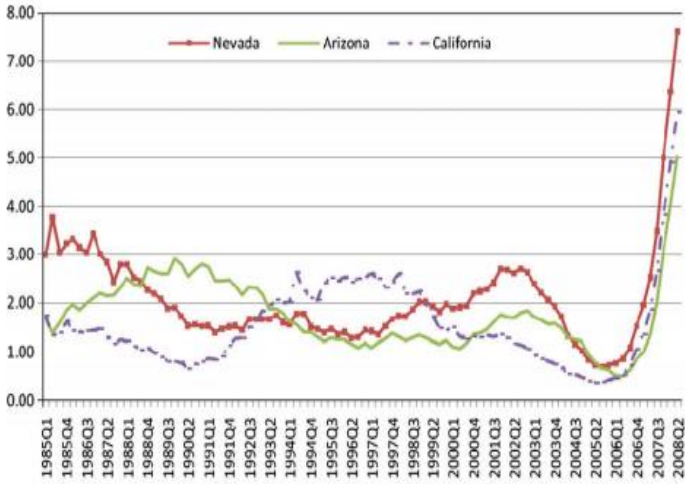
[6] A los prestatarios, hipotecados o deudores se les conocía como *liar loans* pues muchas veces mentían acerca de sus condiciones económicas. Como sabían que el crédito se les iba a conceder igualmente y sus cuentas no iban a ser comprobadas directamente mentían. Otro nombre por los que se les conocía era NINJA (no income, no job, no assets), sin ingresos, sin trabajo, sin activos (Hull, 2012).

[7] En concreto se popularizó las ARM híbridas 2/28. El número 2 hacía referencia a los dos primeros años del préstamo que se concedían a un tipo de interés fijo, mientras que el 28 indicaba que los restantes veintiocho años que le restaban al crédito tendrían un tipo de interés ajustable.

- Ciertos códigos Zip Subprime sufrieron un elevado crecimiento a pesar del nulo crecimiento de los salarios en las mismas zonas o incluso una disminución de estos[8]

En 2005 la demanda inmobiliaria y el precio de la vivienda alcanzaron los niveles más altos de la década.

Figura 4: Variación del precio de bienes inmuebles en Nevada, Arizona y California. (1985-2007)

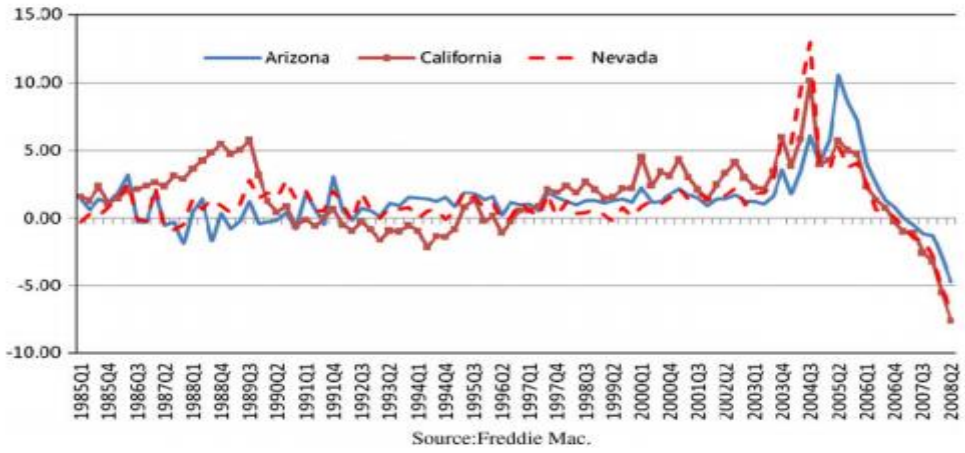


Fuente: Mortgage Bankers Association.

Una vez comenzado el 2006 y por los motivos expuestos anteriormente, la demanda inmobiliaria disminuyó, sin embargo, el precio de la vivienda continuó ascendiendo hasta 2007 (Figura 4).

[8] Para ver más conclusiones del estudio visitar: Mian, A., & Sufi, A. (2009).

Figura 5: Variación de los tipos de interés de créditos hipotecarios en Nevada, Arizona y California. (1985-2007)



Fuente: Freddie Mac.

Las familias hipotecadas habían asumido costes del 100% o más de la vivienda. Al realizarse el ajuste del tipo de interés los hipotecados no podían hacer frente a la devolución del crédito concedido, por lo que la mayoría incurrieron o se posicionaron en Default (Figura 5).

La situación era la siguiente, por un lado, la demanda inmobiliaria disminuyó debido a los altos precios de las viviendas. Por otro lado, la oferta de vivienda aumentó descontroladamente debido a las ejecuciones por parte de los propietarios de las viviendas hipotecadas. El problema del mercado inmobiliario estadounidense es que funcionaba como una opción bursátil, en concreto como una Put[9]. En este sentido, cuando el prestatario incurría en Default, el prestamista tenía la posibilidad de adquirir el inmueble, sin embargo, el resto de los bienes del hipotecado eran intocables (Hull, 2012). De este modo el hipotecado en cualquier momento del Default, podía vender al prestamista la vivienda por el valor actual amortizado. Con la ejecución hipotecaria, la hipoteca sobre el inmueble se cancelaba y esta era adquirida por el prestamista del crédito. A medida que fueron sucediendo las numerosas ejecuciones, los prestamistas se vieron con un portfolio inundado de viviendas vacías, cuyo valor y precio descendía drásticamente (Hull, 2012).

[9] Una Put u opción de venta es un instrumento financiero o contrato que otorga a su propietario el derecho, pero no obligación de vender una cantidad determinada del título subyacente a un precio determinado por un precio determinado

Las consecuencias fueron demoledoras para la sociedad. En el aspecto humano, las familias que no podían hacer frente al pago de la hipoteca se veían obligados a abandonar su casa pues el prestamista ejercía su derecho de adquisición de la vivienda por incurrir en Default. Por otro lado, aumentó la especulación inmobiliaria pues la mayoría de los inmuebles eran propiedad de fondos de inversión que jugaban con la variación de precios para beneficiarse de esto (Zinnerman, 2012).

En una misma calle existían dos casas ambas en posesión de dos familias. Las dos viviendas alcanzaban un valor de 200.000$ y ambas familias solicitaron hipotecas que ascendían a las 250.000$. En ambos contratos, si el hipotecado no cumplía con sus obligaciones crediticias, el prestamista podría ejecutar la hipoteca, hacerse con el control de la vivienda y venderla a un precio estimado de 170.000$. Lo que sucedía en este caso es que ambas familias podrían declararse en posición de Default y seguramente adquirir la vivienda de al lado por un precio mucho menor, es decir, como los bienes personales no eran ejecutables, ambos ejercitaban la opción Put para adquirir por un valor mucho menor la casa de la otra familia. En condiciones normales, en caso de Default del hipotecado, los prestamistas eran capaces de recuperar hasta el 75% del valor de la ejecución. Tras las crisis de 2007, ese mismo año apenas la cifra alcanzaba el 25% del valor del activo (Hull, 2012).

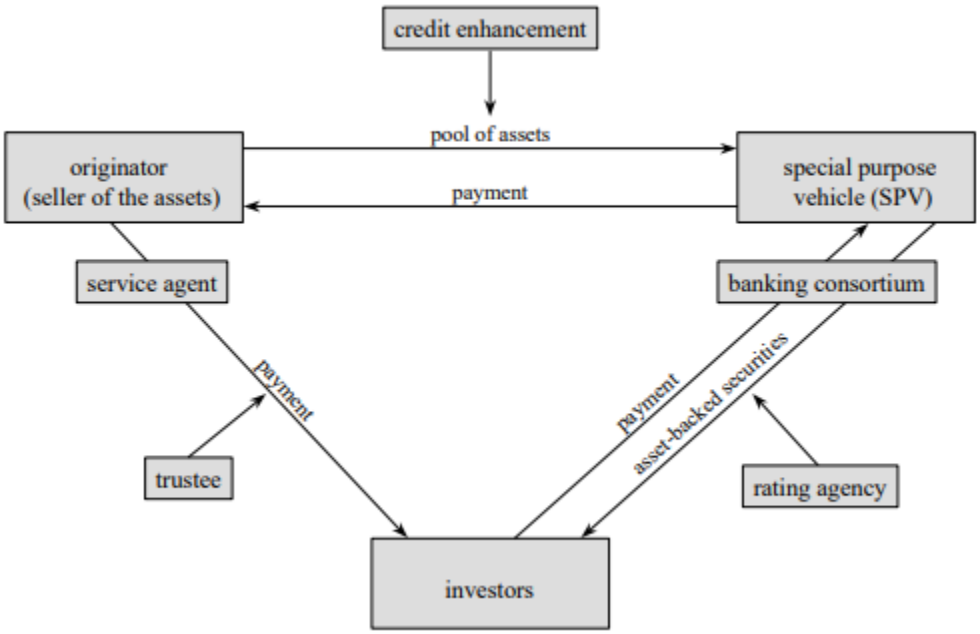
3.2. Proceso de securitización

Hemos visto las causas y las consecuencias de esta crisis a nivel individual, sin embargo, el papel de los inversores, los bancos y las agencias de calificación tuvieron un papel fundamental. Como es lógico, los bancos concedentes de hipotecas vendían productos o sus portfolios a inversores para, bien cubrirse del riesgo de Default de los hipotecados o para cubrir las pérdidas que producían las ejecuciones de hipotecas sobre viviendas cuyo precio empezaba a caer. Los bancos crearon por lo tanto porfolios de hipotecas, la mayoría Subprime, muy arriesgados y atractivos para inversores. Al proceso de creación de dichos portfolios se conoce como securitización o titulación. Consiste en la transformación de crédito no líquido y activos no comercializados en líquidos y comerciales como son los títulos-valor (Berlin, 1994).

El primer proceso de securitización consiste en seleccionar una serie de activos con características homogéneas, en este caso hipotecas. Una vez agrupados todos los activos, el vendedor los enajena a través de SPVs, compañías trusts cuya función es el mantenimiento de estos activos. Ya en posesión de los SPVs, con ayuda de agentes privados de bolsa, se emiten los títulos valor tanto en el mercado público como en el privado (Keys, 2010). Como muchos inversores no estaban dispuestos a correr todo el riesgo asociado al portfolio que estaban adquiriendo, se cubrían o mejoraban el crédito con la sobrecolateralización de los activos o con la firma de seguros. (Henke, 1998).

En la Figura 6 se muestra cómo funcionaba el proceso de securitización para los ABS (asset-backed security), un tipo de título que contenía los packs de préstamos inmobiliarios. El término *backed* hacía referencia a que dicho producto estaban respaldados por los activos que cubrían los préstamos, en este caso las viviendas de los hipotecados.

Figura 6: Proceso de securitización de títulos.



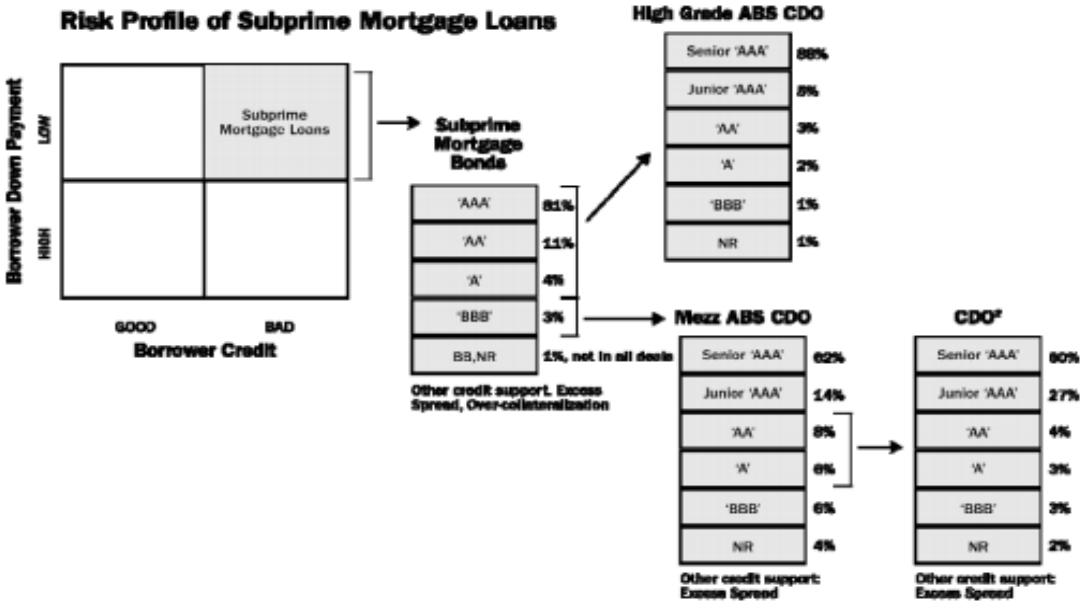
Fuente: Risk Management and Financial Institutions, Hull (2012)

Una vez vendidos, los flujos de caja de los activos se colocaban en ramas. Pese a que en la realidad existían muchos más tipos de ramas, las más populares eran tres, la rama Senior la Mezzanine y la rama Equity. En cada rama se colocaban los flujos de caja de los activos por el orden el expuesto de modo que hasta que no se cubría la rama Senior, la Mezzanine no

dispondría de los flujos de caja restantes (Gorton, 2007). Es por eso por lo que, si nuestro portfolio adquiriese un valor de 100 millones, 75 se destinarían a la rama Senior, una vez cubiertos se destinarían 20 a la rama Mezzanine y, por último, los cinco restantes a la rama Equity. (Hull, 2012).

Como es lógico cada rama poseía una tasa de retorno diferente pues el riesgo de cada una era mayor. Mientras que la rama Senior ofrecía rentabilidades del 5%, las ramas Equity podrían llegar a ofrecer un 30% (Hull, 2012). Sin embargo, las tres ramas se concentraban en un mismo título, el ABS. A la rama Senior se le calificaba como crédito AAA, a la Mezzanine BBB, y generalmente la rama Equity carecía de calificación crediticia. Como era lógico encontrar inversores para la rama Senior, no involucraba gran riesgo ya que su valor crediticio era elevado. La rama Equity o no se vendía o se entregaba a fondos de cobertura capaces de cubrir el riesgo (Gorton, 2007). El problema surgió con la rama Mezzanine, sobre la cual se realizaban sucesivas ABS. El producto resultante era un ABS con sus propias ramas sobre una rama Mezzanine al que se denominó ABS CDO (collateral debt obligation). Sobre un producto cuyo valor era BBB, se realizaban ABS con las mismas ramas anteriores, solo que la rama Senior del ABS CDO poseía una calificación AAA cuando la rama Mezzanine no alcanzaba ni si quiera la calificación de A- (Gorton, 2007).

Figura 7: Proceso de creación de los CDO´s con su valoración crediticia.



Fuente: UBS, Market Commentary (2007)

3.3. Acuerdos de Basilea

Debemos hacer una breve referencia a la aprobación de los Acuerdos de Basilea, ya que han introducido nuevas fórmulas y restricciones para la concesión de créditos hipotecarios

Dentro de los numerosos acuerdos de Basilea los más pertinentes o los que nos conciernen son Basilea II y Basilea III. Bien es cierto que Basilea II fue aprobado con anterioridad al estallido de la crisis de 2007, sin embargo, ciertos economistas achacan a la inaplicación de las disposiciones de dicho Acuerdo a la recesión económica.

Tras la gran recesión económica, el 31 de diciembre de 2011 se firmó la implementación de Basilea II.5. No parece justo achacar las consecuencias de la crisis a la laxitud de Basilea II, sobre todo, porque tal y como se ha demostrado, las consecuencias de la burbuja inmobiliaria estadounidense surgen por causas achacables al año 2000. Por otro lado, tampoco parece que Estados Unidos implementase correcta o íntegramente las disposiciones del Acuerdo Basilea II, por lo que, si hubiese procedido como otros países, posiblemente los efectos de la crisis hubiesen sido más bajos (Hul, 2012).

En los primeros acuerdos de Basilea, en 1998 se establecieron cuatro ponderaciones para medir el riesgo de crédito, que se aplicaban de acuerdo con las categorías sobre las que se clasificaban los créditos por tipo de emisor de la deuda o del prestatario: 0% para las obligaciones a cargo de gobiernos miembros de la OCDE, 20% para títulos emitidos por bancos de países miembro de la OCDE, 50% para los créditos hipotecarios y 100% para las obligaciones de países menos desarrollados o títulos de bancos de países no miembros de la OCDE (Ong, 1999).

Por no incluir todos remedios no pertinentes para el presente estudio, destaca la mayor innovación que introdujo esta modificación, la medición comprensiva del riesgo (CRM). El CRM se diseñó con el objetivo de estudiar la correlación entre los activos ABS y los CDO. Ambos productos en momentos donde la correlación se incrementaba como en 2007, en situaciones de Default sufrían pérdidas conjuntas. Basilea II.5 permite a los bancos calcular su propio CRM, pero para poder ser válido debía incluir diversos factores como la volatilidad de las correlaciones, (Lopez, 2004).

En diciembre de 2009, se firmó por último Basilea III. El objetivo de este fue hacer frente al riesgo de liquidez provocado por el impacto de la crisis de 2007. El caso que mejor ilustra este riesgo fue el caso de Northern Rock y Lehman Brothers. Para financiar operaciones a largo plazo los bancos empleaban pagarés no garantizados cuyos vencimientos rondaban los 90 o 270 días (Hull, 2012). Para seguir manteniendo la financiación, seguían emitiendo estos instrumentos financieros en el momento de vencimiento del anterior. Sin embargo, cuando los bancos sufrían dificultades económicas no podían emitir más pagarés. El resultado era una falta de liquidez preocupante e incapacidad de hacer frente a las deudas de corto plazo (Lopez, 2004).

Los Acuerdos de Basilea han supuesto un cambio en el paradigma del cálculo del riesgo de crédito. Su estudio en este apartado es meramente informativo y merece sin duda un análisis intensivo fuera de este trabajo.

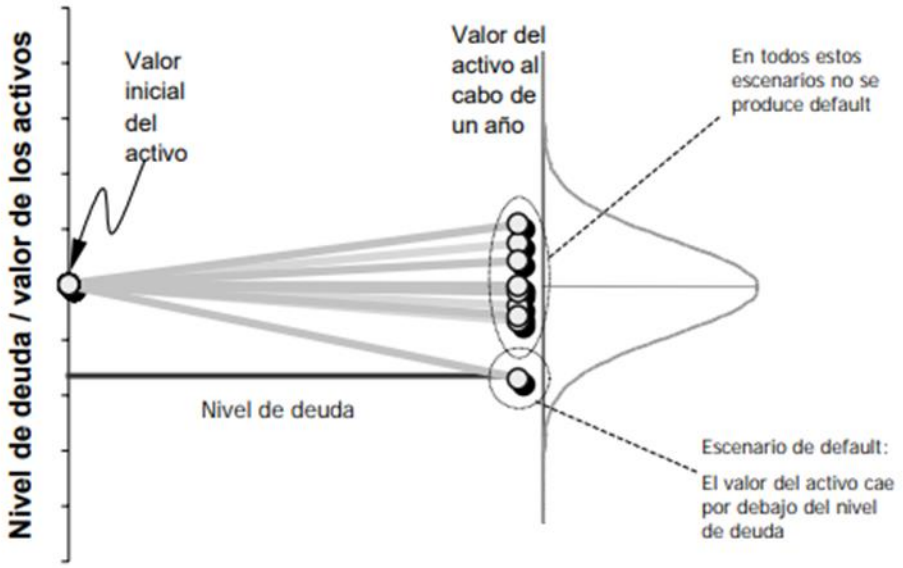
3.4. Perspectiva empresarial

Finalmente, podemos definir el término riesgo de crédito como la probabilidad de que, a su vencimiento, una entidad o persona no haga frente en parte o en su totalidad a su obligación de devolver una deuda o rendimiento acordado sobre un instrumento financiero, debido a la quiebra, iliquidez u otra razón (Chorafas, 2000). Cuando el prestatario o deudor no puede hacer frente a la obligación incurre en lo que hemos denominado como Default, argot financiero empleado para referirse al incumplimiento de una deuda. Para nuestro objeto de estudio, el riesgo de crédito hipotecario será la probabilidad de que, a vencimiento, el deudor no haga frente a la cuota del crédito hipotecario al final del mes.

Para el caso de las empresas, el riesgo de crédito funciona de manera similar. Sin embargo, las empresas suelen tener mayor capacidad de financiación o endeudamiento que por ejemplo una familia de personas. Las empresas al poseer generalmente un abanico o portfolio de activos tienen mejor capacidad de hacer frente a sus deudas (Chorafas, 2000).

Tal y como se muestra en la Figura 8, las empresas también pueden incurrir en Default. En este sentido y tal y como se explica en la imagen, cuando el valor de los activos totales cae por debajo de la deuda total de la empresa, nos encontramos ante una situación de Default.

Figura 8: Escenario de Default para una empresa**.**



Fuente: De Basilea: Recordatorio, BASILEA II (2010).

Como se comentaba en el experimento de Mian y Sufi, el mercado inmobiliario estadounidense funcionaba como una opción Put. Es decir, en caso de Default el deudor (sea una persona física o una jurídica) tenía un derecho de venta del activo inmobiliario. El caso para los bancos cuyo portfolio lo componían viviendas hipotecadas con las Subprime era mucho más grave. Debido a las numerosas ejecuciones hipotecarias, muchos de los pisos quedaron vacíos, sumándose eso a la caída de los precios, el valor de los activos inmobiliarios se desplomó agravando y convirtiendo la situación en un círculo vicioso. (Mian & Sufi, 2009).

Existe por otro lado un problema con la correlación de activos que una empresa puede poseer. Es decir, a mayor correlación entre los activos, las probabilidades de incurrir en Default son mucho mayores. Con todo el sentido del mundo, si el 80% del portfolio de un banco son bienes inmuebles o son hipotecas Subprimes, en el estallido de la crisis si al menos el 50% de esos créditos concedidos, sus deudores incurrían en Default, era muy probable que la entidad bancaria perdiese mucho valor en sus activos y por ende incurriría en Default también

(Hull, 2012). Cuando hablamos de un portfolio de créditos existen ciertos aspectos a tener en cuenta. No estamos hablando de créditos individualizados si no que nos encontramos con un portfolio diversificado compuestos por varios créditos a la vez

Tal y como se comentó en la introducción las principales empresas que emplean métodos para el cálculo de la posibilidad de impago, son los bancos. Cada entidad financiera posee su propio sistema de medición de riesgo de tal forma que forman incluso parte de la propiedad intelectual del banco (Elizondo & Lopez, 1999).

3.5. Tipos de riesgo de crédito

Por último, en este apartado se hará referencia a determinados riesgos de créditos, en función de su origen.

En primer lugar, es necesario analizar el riesgo de crédito hipotecario. Pese que ya ha sido comentado en algunos puntos anteriores, se puede definir como el riesgo de que el prestatario no pague su cuota correspondiente del crédito concedido por el que se adquiría una vivienda (Sirignano et al. 2015). El riesgo de crédito depende no solo de las características del sujeto hipotecado sino en las condiciones en las que se contrajo el crédito o incluso de las características globales del ciclo económico en la que se contrajo el crédito. De esta forma factores como la elevada edad del préstamo, variable que se analizará en el caso de estudio, son una de las principales causas de Default previas a la crisis del 2007 (Bagherpour, 2017).

En segundo lugar, podemos destacar el riesgo de crédito del consumidor. Este hace referencia a las índices de morosidad e incumplimiento de los titulares de la tarjeta de crédito de los clientes de los bancos. (Khandani et al. 2010). Se recogen dentro de este riesgo, los de tarjetas, cuentas corrientes y otros instrumentos de pago o de giro en posesión del consumidor.

Terceramente podemos destacar el riesgo de crédito del préstamo. Las diferencias con el riesgo objeto de estudio son escasas pues en el fondo, el crédito hipotecario actúa generalmente bajo las mismas condiciones que un préstamo normal. La diferencia reside en que otras características o variables tienen mayor peso (Bagherpour, 2017). Por ejemplo, para los préstamos simples, el tipo de interés puede tener un mayor peso que si el prestatario

pertenece a una familia unifamiliar o no. En el caso del riesgo de crédito hipotecario, tal y como se analizará en el caso de estudio, observaremos que variables como el LTV o el Balance en el momento de observación del crédito tienen un impacto mayor en la probabilidad de Default del prestatario que por ejemplo la edad del crédito concedido.

Otro tipo de riesgo común es el riesgo de crédito corporativo. La diferencia con los anteriores reside en que el sujeto pasivo en vez de ser un particular, lo conforma una empresa. En este caso, para su cálculo, además de analizar las características del crédito concedido es necesario hacer referencia a variables empresariales. Las más importantes son las principales cuentas contables, los comportamientos de pago de la empresa prestataria frente a créditos anteriores y toda aquella información relacionada con los préstamos y la posición macroeconómica de la empresa, en comparación con el mercado en el que se encuentra (Carling et al. 2013)

Por último y con el ánimo de no extendernos más, debemos resaltar el riesgo de calificación del crédito, que consiste en el riesgo generado por sobreestimar la calificación de un crédito (Dumitrescu et al. 2018). Este último adquiere una gran importancia en las agencias de rating actuales como Moody´s o Fitch, y han desarrollado modelos predictivos capaces de obtener el riesgo generado por la sobre calificación de créditos.

4. MODELOS DE CÁLCULO DEL RIESGO DE CRÉDITO

Una vez definida la perspectiva histórica de la Inteligencia Artificial y analizada la evolución del riesgo de crédito en la recesión económica se deben de abordar las principales técnicas que van a ser empleadas para el estudio del caso. En este apartado haremos referencia a los métodos más relevantes que han sido empleados para el cálculo del riesgo de crédito entre los que distinguiremos los más tradicionales hasta los más modernos que involucran técnicas estadísticas más avanzadas. Se hará referencia también al modelo Logit o regresión logística que ha sido empleada durante varios años como piedra angular a la hora del cálculo de este tipo de riesgo.

Se ha realizado una división de los principales modelos de medición del riesgo de crédito siguiendo la propuesta de clasificación de García (Tabla 2). Estos sistemas tienen como objetivo identificar los determinantes del riesgo de crédito de las carteras de cada institución, con el propósito de prevenir perdidas potenciales en las que podría incurrir. Es necesario destacar que, a diferencia del riesgo de mercado, el desarrollo para medir el riesgo de crédito ha sido menos cuantiosos, pues cada institución contempla generalmente un modelo diferente de cálculo de este riesgo, por lo que existen numerosas aproximaciones a su cálculo (García, 2010). Siguiendo la clasificación del autor, el análisis del riesgo de crédito debe considerar dos tipos de riesgo, el individual del crédito y el del portfolio que los recoge.

Tabla 2:Clasificación de los tipos de riesgo según García.

Riesgos individuales	**Riesgos del portfolio**
• Probabilidad de incumplimiento • Tasa de recuperación • Migración de crédito	• Incumplimiento y calidad crediticia correlacionada • Concentración del riesgo y concentración crediticia

Fuente: Managing credit risk. Chorafas (2000).

4.1. Métodos Tradicionales

El primer grupo de modelos a analizar son los denominados modelos tradicionales. Se estudiarán dos tipos, los sistemas expertos y los sistemas de calificación. La característica principal de estos modelos son la aplicación de porcentajes para determinar el riesgo que se va a cubrir. De esta forma se realizan proyecciones de las variables económicas y financieras en el tiempo dependiendo del desempeño de la empresa (Marquez, 2006). Amos sistemas por otro lado consideran estático e independiente el comportamiento de las variables e involucra el criterio subjetivo de cada analista basándose en valoraciones de acuerdo con la experiencia adquirida en la asignación de créditos (García, 2010).

4.1.1. Sistemas expertos

Los sistemas expertos podrían componer la base de los sistemas de clasificación actuales basados en técnicas de Machine Learning o Inteligencia Artificial. Un sistema experto puede definirse como un sistema informático (hardware y software) que simula a los expertos humanos en un área de especialización dada (Castillo, 1997). Sin embargo, quedan limitados a la etapa de clasificación ya que no pueden establecer un vínculo entre la probabilidad de Default con la gravedad de la pérdida (García, 2010). Es necesario remarcar que los Sistemas expertos se emplean exclusivamente para el cálculo del riesgo de crédito corporativo ya que como se verá más adelante, se fundamentan en los ratios principales de la empresa deudora. Los Sistemas expertos se fundamentan en los siguientes factores:

- Capacidad: Capacidad del pago del deudor. Valoración acerca de la antigüedad de la empresa, crecimiento a lo largo del tiempo, principales actividades, flujos de caja del negocio, deudas pasadas y presentes y otros factores decisivos para hacer frente a la devolución del crédito. Si bien no es mencionado por García, para personas físicas no empresas, sería necesario otros factores como la renta, deudas pendientes, gastos corrientes.
- Capital: Análisis de la situación financiera del deudor, capacidad de endeudamiento, liquidez, tiempo de rotación de proveedores.

- Colateral: Activos en posesión del deudor para constituirlo como garantía de pago. En el caso de créditos hipotecarios como los mencionados en el apartado de la crisis, el colateral lo conformaban las viviendas hipotecadas, de ahí el nombre de ABS.
- Carácter: Información de hábitos de pago del deudor. Referencias comerciales de otros proveedores, posibles verificaciones de demandas judiciales, referencias bancarias.
- Condiciones: Factores externos que afectan a la situación patrimonial del deudor. Condiciones económicas del país, situación del sector del negocio.

Como se puede observar, estos sistemas contemplan características simples de del prestatario para así conformar un modelo posterior, son por así decirlo las bases de modelos posteriores, los cuales a través de estos datos realizaban predicciones sobre la posibilidad de Default. Una vez se implementaron las técnicas de aprendizaje automático permitieron a los sistemas almacenar la información para de esta forma evitar la complementación del evaluador en la valoración del crédito (Andrés, 2000).

4.1.2. Sistemas de calificación

Por otro lado, destacan los sistemas de calificación comúnmente empleados por ejemplo por la Oficina de Control de Moneada de Estados Unidos para la evaluación de su deuda emitida y adecuación de las reservas para posibles casos de Default. Tras la aprobación de los primeros Acuerdos de Basilea, se decidió transformar los sistemas de medición de crédito que se venían utilizando (Ong, 1999). Se consideraron dos nuevas aproximaciones a la medición del riesgo de crédito. Una estandarizada en la que entraban en consideración la valoración de los créditos por entidades externas y una aproximación basada en calificaciones internas de riesgo, cuyo fundamento eran las pérdidas no esperadas. La metodología más empleada para esta aproximación fue el modelo Altman Z-score en el que se aplicaba un análisis discriminante a un conjunto de indicadores financieros cuyo propósito fue clasificar las empresas en dos grupos, bancarrota y no bancarrota (De la Fuente, 2006)[10]. Debemos destacar que este método se emplea también para el cálculo del riesgo de crédito corporativo.

[10] Para ver más acerca de la aplicación de este modelo ver Anjum, S. (2012). Business bankruptcy prediction models: A significant study of the Altman's Z-score model. *Available at SSRN 2128475*. Y Altman, E. I.,

La función discriminante para empresas públicas estimada por Altman se configuraba de la siguiente forma (Altman, 2017)[11].

$$Z = 1.2X_1 + 1.4X_2 + 3.3X_3 + 0.6X_4 + 1.05X_5$$ [12]

***Ecuación 1**: Función discriminante de Altman.*

Donde:

- X1= Capital circulante / Activos totales.
- X2= Beneficios no distribuidos / Activos totales
- X3= EBITDA / Activos totales
- X4= Capitalización bursátil / Deuda Total
- X5= Ventas Netas / Activos Totales
- Z= Índice de estudio

En función de los resultados:

- Para Z >3: No se encontraban indicios para prever una posible situación de quiebra para la empresa.
- Para 1.8 > Z > 2.7: La empresa se encontraba en zona de alerta. Si las condiciones no eran modificadas se podría prever una situación de quiebra en dos años.
- Para Z < 1.8: La empresa incurriría en quiebra inminente.

Como podemos observar, se fueron introduciendo medios más avanzados basados en regresiones econométricas o aplicaciones estadísticas. Sin embargo, estos se quedan bastante cortos respecto a la aplicación de otros modelos más modernos que incluyen un mayor número de variables o incluso de modelos estadísticos que aplican técnicas de Machine Learning

Iwanicz-Drozdowska, M., Laitinen, E. K., & Suvas, A. (2014). distressed firm and bankruptcy prediction in an international context: A review and empirical analysis of Altman's Z-score model. *Available at SSRN 2536340*.

[11] El propio autor configuró una función diferente para las empresas privadas que disponía de la siguiente forma: Z'' = 3.25 + 6.56·X1 + 3.26·X2 + 6.72·X3 + 1.05·X4.

[12] Las cuatro variables se encuentran expresadas en decimales.

4.2. Métodos Modernos

En contraposición con los modelos anteriores, destacan modelos de medición más modernos donde intervienen un mayor grado de sofisticación.

El modelo KMV de monitoreo de crédito fue implementado por la agencia Moody´s y consistía en estimar la probabilidad de incumplimiento entre activos y pasivos. Definía la probabilidad de incumplimiento como una función de la estructura del capital de la firma, la volatilidad de los activos y su valor actual (García, 2010).

En una primera fase, se estima el valor de los activos y la volatilidad del rendimiento en función del valor de mercado, nunca de los libros. El valor de los activos en el mercado representa el valor efectivo libre y futuro producidos por los activos de la empresa, descontados a una tasa correspondiente. Esta fase emplea bases del conocido modelo de Merton. De acuerdo a este modelo, para una empresa no hay reestructura o renegociación, de manera que existen solo dos posibilidades para hacer frente al riesgo de crédito (Saunders, 2000). La primera es la liquidación total del montante pactado en plazo. La segunda, es la declaración de insolvencia y entrega de los activos sobre los que recaían garantías por ese crédito, al banco prestamista. A fecha de vencimiento, tendremos los créditos o bien como parte del activo de la empresa si se pagan, o bien en el pasivo si se declara insolvente. (Merton, 1977)

En una segunda fase, se calcula el riesgo de los activos en el que se incluyen el riesgo del negocio y del sector en el que trabaja la empresa, medidos por la volatilidad de los activos y condicionada por la propia volatilidad de las acciones. Llegados a este punto, se debe calcular como punto de incumplimiento (DPT), que es el momento en el que el valor de los activos alcanza un nivel medio entre el valor total de las obligaciones y el valor de la deuda a corto plazo. (García, 2010). Por otro lado, la distancia al incumplimiento nos indica el número de desviaciones estándar entre la media y la distribución del valor de los activos en el punto de incumplimiento:

$$DD = \frac{E(V_1) - DPT}{S}$$

***Ecuación 2**: Distancia de incumplimiento para el modelo KMV.*

Donde:

- E(V1) = valor esperado de los activos en un año
- DPT= Punto de incumplimiento= Deuda a corto plazo + 0.5 (deuda a largo plazo)
- S= volatilidad de los rendimientos esperados de los activos
- DD= distancia al incumplimiento

Por último, se calcula la derivación de la probabilidad de incumplimiento, cuyo objetivo es convertir la distancia anteriormente hallada en un estimación de la frecuencia de incumplimiento[13]. El modelo KMV utiliza un modelo de valuación de riesgo neutral para derivar precios descontados al valor esperado de los futuros flujos de caja (García, 2010). La valuación de dichos flujos consiste en:

$$PV = \frac{FV * (1 + LGD) + FV * LGD(1 - Q)}{(1 + i)}$$

***Ecuación 3**: Derivación de la probabilidad de incumplimiento para el modelo KMV*

Donde:

- PV= valor presente de los flujos de caja
- FV= valor futuro
- LGD= influencia de la pérdida expresada en porcentaje
- I= tasa libre de riesgo
- Q= probabilidad de que el emisor incumpla en un año

Concluyendo, lo que nos permite este modelo es encontrar una relación entre la distancia al incumplimiento y la probabilidad de que se produzca el Default. Para esto, se establece una tabla que relaciona la probabilidad de incumplimiento con los niveles de distancias de Default. Así, es constante esta relación entre distancia en incumplimiento y probabilidad de

[13] El modelo KMV toma como referencia el modelo EDF o estimated Default frecuency que transformaba la información contenida en el precio de la acción en una medida del riesgo de incumplimiento de pago.

que ocurra, independientemente del sector, el tamaño o el tiempo en el que se puedan encontrar dos empresas, ya que las diferencias existentes debidas a todos los factores señalados se recogen ya la distancia en incumplimiento. Estos tres métodos que hemos analizado hasta ahora nos sirven para identificar como se ha venido calculando el riesgo de crédito corporativo, de las empresas. Para nuestro caso de estudio, debemos de profundizar más pues lo que nos interesa es hallar el riesgo de crédito hipotecario, donde los sujetos deudores o prestatarios son personas solicitantes de una hipoteca, no empresas.

4.3. Métodos Estadísticos

Por último, parece necesario destacar modelos actuales empleados por empresas. Se citarán a continuación algunos de estos:

1) Modelos de previsión no lineales y no paramétricos del riesgo crediticio de los consumidores. (Khandani et al. 2010). Se calcula el riesgo de crédito del consumidor.
   a) Estudio: Combinación de las transacciones de los clientes y los datos de las oficinas de crédito de enero de 2005 a abril de 2009 para una muestra de los clientes de un importante banco comercial. Elaboración de pronósticos fuera de la muestra que mejoran significativamente la calificación del crédito y las tasas de impago impagos de los titulares de tarjetas de crédito, con una regresión lineal R2 de morosidad pronosticada/realizada del 85% (Khandani et al. 2010).
   b) Conclusión: Los patrones de series temporales de las tasas de morosidad estimadas a partir de este modelo en el curso de la reciente crisis de financiera sugieren que el análisis de riesgo de crédito al consumidor agregado puede tener importantes aplicaciones en el pronóstico del riesgo sistémico (Khandani et al. 2010).

2) Desarrollo de un modelo de aprendizaje profundo de riesgo hipotecario de múltiples períodos. Para analizar un conjunto de datos cuyo origen y registros parten del rendimiento mensual para más de 120 millones de hipotecas originadas en todo EE.UU. entre 1995 y 2014. (Sirignano et al. 2016). Se calcula el riesgo de crédito hipotecario por lo que es el estudio más interesante y que más puntos en común comparte con nuestro caso de estudio.

a) Estudio: Las estimaciones de estructuras de plazos de probabilidades condicionales de pago anticipado, ejecución hipotecaria y varios estados de morosidad, incorporan la dinámica de un gran número de préstamo, así como variables macroeconómicas hasta el nivel de código postal. Las estimaciones descubren la naturaleza altamente no lineal de la relación entre las variables y el comportamiento del prestatario, especialmente el prepago. También destacan el efecto de las condiciones económicas locales sobre el comportamiento de los prestatarios. El desempleo estatal tiene el mayor poder explicativo entre todas las variables, ofreciendo fuerte evidencia de la estrecha conexión entre los mercados de la vivienda y la macroeconomía (Sirignano et al. 2016).
b) Conclusión: La sensibilidad de un prestatario a los cambios en el desempleo depende en gran medida del desempleo actual. También varía significativamente en toda la población de prestatarios, lo que pone de relieve la interacción del desempleo y muchas otras variables. Estas conclusiones tienen importantes repercusiones para los inversores en títulos respaldados por hipotecas, los organismos de calificación y los encargados de la formulación de políticas de vivienda (Sirignano et al. 2016).

5. TÉCNICAS DE MACHINE LEARNING

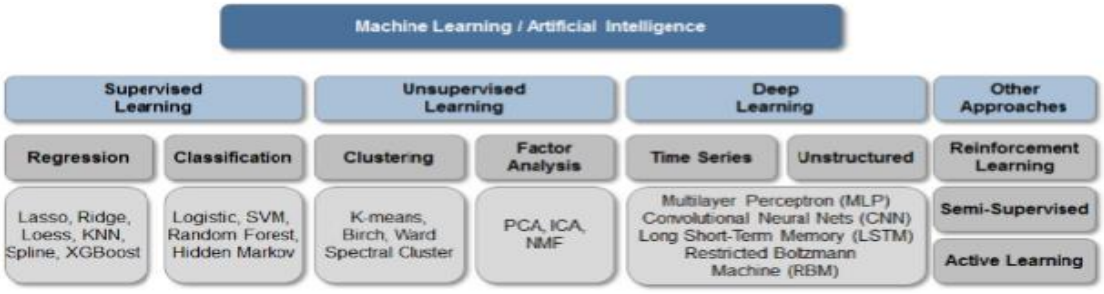
Se desglosará en este apartado las técnicas de Machine Learning empeladas para el posterior caso práctico. Primero se hará referencia al marco conceptual. En segundo lugar, se analizará las bases del modelo Logit o regresión logística, en la que se apoyan las técnicas posteriores. Por último, se estudiará el concepto de árbol de decisión y Gradient Boosting

1.1. Marco conceptual. Problema de clasificación.

El concepto de Statiscal Learning se refiere a una serie de herramientas para la modelización y la comprensión de bases de datos muy complejas. Dividido en supervisado y no supervisado, es el primero el que nos permite construir modelos estadísticos con el objetivo de predecir y estimar una variable en función de otras (Figura 9). En este sentido disciplinas como los negocios, la medicina o las políticas públicas emplean estas herramientas con el objetivo de sacar el mayor partido a una gran base de datos.

Antes de adentrarnos en el problema de clasificación es necesario comprender que razonamiento de estas técnicas es bastante más compleja de lo planteado en este caso. El objetivo es entender cómo aplicar dichas técnicas a problemas de negocios actuales, no hacer un análisis exhaustivo de todas las técnicas posibles de Machine Learning.

Tabla 3: Tipos de aprendizaje**.**



Fuente: JP Morgan Macro QDS

Tal y como muestra la Figura 9, todos los problemas relacionados con el Statiscal Learning se pueden dividir hasta en cuatro categorías. De estas cuatro, nos centraremos en las dos

primeras, el aprendizaje supervisado y los no supervisado. Es el primero el que nos interesa para este trabajo pues son los que nos permiten estudiar el comportamiento de una variable en función de otras. Para cada observación medida por el sujeto predictor, x_i; i $= 1, \ldots$ n. existe una respuesta asociada y_i. Lo que se busca con estos modelos es buscar la relación ente la respuesta con el predictor, para así predecir el comportamiento de futuras observaciones (predicción) o comprender la relación establecida entre la respuesta y el predictor (inferencia). (Kolanovic et al. 2017).

Las principales técnicas estadísticas que se emplean como parte del Machine Learning estudiadas en el presente trabajo pueden resumirse en las siguientes (Figura 9).

- Clasificación: Asignación de una clase a un determinado objeto o individuo. Busca la categorización generalmente en dos etiquetas o respuesta binomial. Un ejemplo práctico es la definición de "Guapo" o "Feo" de una persona, o para el presente objeto de estudio, si el cliente incurre en "Default" o "No Default".
- Regresión: Generalización del problema de clasificación. La respuesta toma forma de un número real.

1.2. Modelo Logit

La diversidad de modelos para predecir la capacidad de pago nos impide establecer un criterio claro y común para todos ellos. Sin embargo, gran parte de las técnicas aplicadas en nuestro caso de estudio parten de las premisas de la regresión logística.

Primeramente, debemos analizar qué tipo de variable es nuestra variable endógena o dependiente. A la hora de analizar variables, estas pueden ser categorizadas en cuantitativas o cualitativas (categóricas), (Martínez de Ibarreta et al. 2017). Las primeras suelen ser valores numéricos tales como el salario de una casa o el número de horas de espera en una consulta médica. Las variables cualitativas toman valores de diferentes clases o categorías. Ejemplos de estas últimas pueden ser el género de una persona (masculino o femenino), el color de ojos (marrón, azulo verde), o para el caso que nuestro caso, respuestas simples como Default o no Default .

Las variables cuantitativas son más fáciles de analizar con los métodos de regresión linear, cuya función es:

$$y_i = \beta_1 + \beta_2 x_{2i} + \cdots \beta_k x_{ki} + \mu_i$$

***Ecuación 4**: Modelo de Regresión lineal clásico*.

Cuando queremos analizar la probabilidad de variables dependientes binarias se puede emplear el modelo de probabilidad lineal cuya función es:

$$y = \beta_1 + \beta_2 x_2 + \cdots \beta_k x_k + u$$

***Ecuación 5**: Modelo de probabilidad lineal.*

Sin embargo, para nuestro caso en el que queremos analizar una variable cualitativa que toma dos sucesos, el modelo de probabilidad lineal no permite analizar correctamente el valor de la variable dependiente. La regresión logística o modelo Logit por el contrario, sí que nos permite obtener correctamente las probabilidades de que ocurran los sucesos contemplados para la variable binomial, sin incurrir en los errores del modelo de probabilidad lineal (Martínez de Ibarreta et al. 2017). Las funciones del modelo Logit son las siguientes:

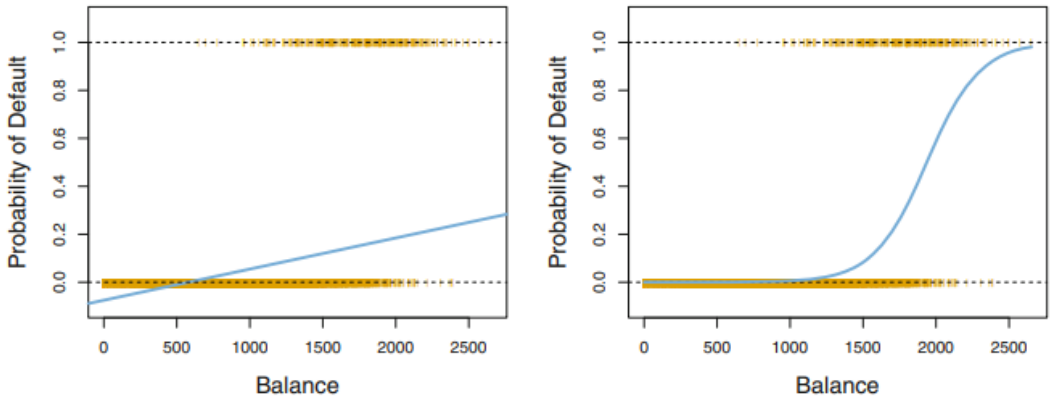
$$Prob\ (y_i = 1) = \frac{e^{z_i}}{1 + e^{z_i}}$$

$$z_i = \beta_1 + \beta_2 x_{2i} + \cdots \beta_k x_{ki}$$

***Ecuación 6**: Modelo Logit*

Existen varias limitaciones al modelo de probabilidad lineal, pero especialmente nos centraremos en el problema de no acotamiento de la probabilidad entre 0 y 1. Si queremos mantener que el valor de y_i sea una medida de la probabilidad, es necesario que la variable y_i no tome valores inferiores a 0 o superiores a 1, es decir que se encuentre dentro del intervalo [0,1]. No podemos obtener una probabilidad de -5.40% o una probabilidad del 120%. Para ello se emplea como relación funcional entre x_i e y_i la función de probabilidad acumulada o función de distribución (Martínez de Ibarreta et al. 2017).

Figura 9: Recta obtenida con el modelo de probabilidad lineal (izq). Función de distribución obtenida con el modelo Logit (dcha).



Fuente: An Introduction to Statistical Learning, James et al. (2017)

Como vemos en la Figura 9, la función de distribución no toma valores inferiores a 0 superiores a 1, por lo tanto, en vez de obtener la recta de regresión que se obtiene con la regresión lineal, obtenemos una función de probabilidad acumulada que podremos utilizar para predecir el comportamiento de la variable endógena.

Por último, debemos destacar la interpretación de los resultados la variable dependiente obtenidos con el modelo Logit. El valor que toma la variable y es la probabilidad de que esa variable tome el valor "1". es decir, si consideramos como suceso 1 incurrir en Default, si y = 0.8, estamos considerando que la empresa tiene un 80% de probabilidad de incurrir en Default. Si se quiere clasificar al sujeto como 0 o 1 y no tener meramente una probabilidad, a partir de cierto resultado se le considera o un suceso u otro. Pongamos un ejemplo. Siendo y = 1 Default o y = 0 No Default, si la variable $y < 0.5$, entonces consideraremos que la empresa no incurrirá en Default. Por el contrario, si $y > 0{,}5$, consideraremos que la empresa sí incurrirá en Default. Si la probabilidad estimada supera un cierto umbral (en nuestro caso 0,5) se calificará como 1. A este umbral le denominamos *cut-off*. (Martínez de Ibarreta et al. 2017).

Concluimos por lo tanto que el Modelo Logit nos permite analizar nuestra variable dependiente y binaria en comparación con el modelo de probabilidad lineal. Nos permite identificar la probabilidad de que el suceso 1 ocurra sin incurrir en los errores provocados por el método de probabilidad línea anteriormente descritos.

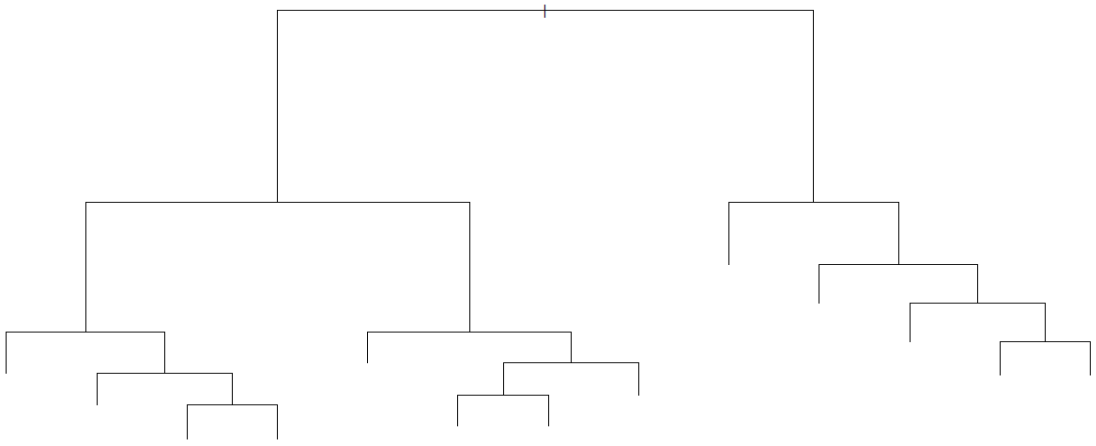
1.3. Árboles de decisión y Gradient Boosting

Una vez descrito el modelo Logit, debemos hacer referencia a las otras dos técnicas que se emplearán en el caso de estudio. Se analizarán en primer lugar, el fundamento de los árboles de decisión, y a continuación la técnica de Gradient Boosting. Bien es cierto que existen otras técnicas como el Bagging, Random Forest, entre otras cuyo punto de partida también son los árboles de decisión y comparten cierta similitud con estas dos. Sin embargo, es imposible abarcar todas las técnicas que se pueden emplear para el análisis del conjunto de datos, por eso nos centraremos en estas dos, porque son las que aplicaremos en el caso de estudio.

1.3.1. Árboles de decisión.

Los árboles de decisión pueden ser aplicados para resolver problemas tanto de regresión como de clasificación. En la siguiente figura se muestra un esquema de un árbol de decisión sin incorporar las variables de los cortes:

Figura 10: Esquema de un árbol de decisión sin variables en los cortes.



Fuente: elaboración propia mediante R.

En los árboles de decisión existen varias regiones, para el caso de la Figura 10 se muestran hasta 13. A estas regiones se le denomina a cada una como R y se les conoce como nodos terminales u hojas de los árboles. (James et al. 2013). A la hora de construir un árbol de

decisión debemos centrarnos en dos pasos. Primero, dividir el espacio de predicción en función de una serie de valores x, que son fundamentalmente las variables explicativas de la variable endógena que queremos analizar. Cada variable en función de su importancia dividirá el espacio en dos regiones para un valor determinado. A esta división, se le denomina corte para la variable x. Dependiendo de la importancia de la variable explicativa es posible que nos encontremos con cortes para una misma variable, como sucederá en nuestro caso de estudio. Una vez analizadas las variables, se obtienen todos los cortes y divisiones, finalizando el árbol en los nodos terminales

Segundo, para cada observación que se realiza dentro de la región se realiza la misma predicción, que es simplemente la media de los valores de respuesta para las observaciones dentro de esa región (James et al. 2013). Es decir, dentro de cada nodo terminal, se realiza la media de los valores de respuesta. El valor con mayor media será catalogado como clase mayoritaria dentro de ese nodo. Si la media de Defaults es mayor que la de no Defaults, la clase mayoritaria que aparecerá en el nodo terminal será Default.

El árbol comienza con todas las observaciones pertenecientes a una sola región y luego continúa dividiéndose sucesivamente hasta llegar a los nodos terminales. Cada división implica la creación de dos nuevas ramas. A este sistema se le denomina *greedy* [14]. (Kolanovic et al. 2017). En cada paso de construcción del árbol la mejor división se realiza en ese paso particular, en vez de mirar características futuras y escogiendo una división que nos favorezca la construcción de un árbol futuro. A este sistema se le conoce como partición binaria del espacio (James et al. 2013).

Para realizar dicha partición binaria, en primer lugar, seleccionamos la predicción Xj. Estableceremos el punto de corte como S y se dividirá el espacio en las regiones:

$$\{X|Xj < S\}\ ;\ \{X|Xj \geq S\}$$

***Ecuación 7**: División del espacio en función de la predicción Xj en un árbol de decisión*

La primera de estas anotaciones nos indica la región del espacio en el que Xj toma un valor menor que S. La segunda anotación indica la región en el que Xj toma un valor mayor que S. Consideramos las predicciones Xj y todos los posibles valores de los puntos de corte S

[14] Traducido al español vendría a denominarse codicioso.

para cada una de las predicciones (Kolanovic et al. 2017). Para cualquier Xj y S definimos las regiones (R) como:

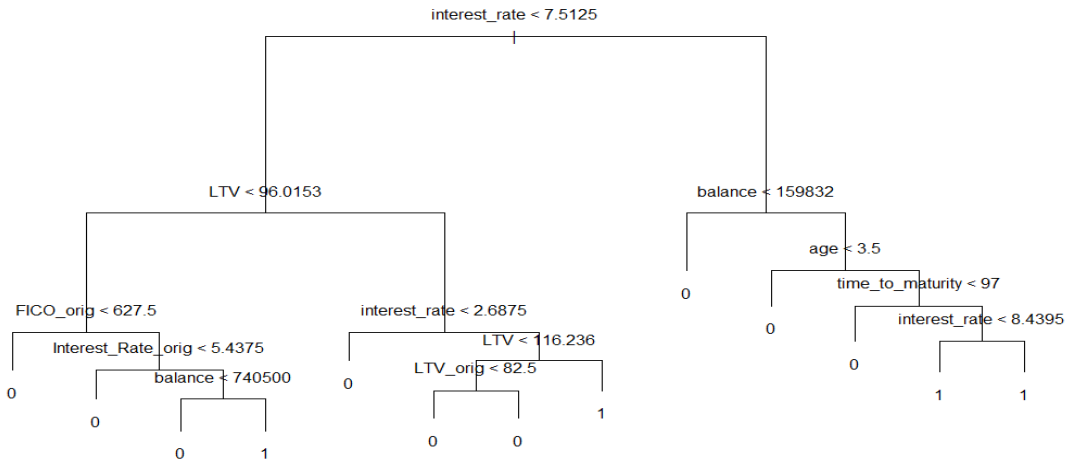
$$R_1(j,S) = \{X|Xj < S\}\ ;\ R_2(j,S)\ \{X|Xj \geq S\}$$

***Ecuación 8**: Obtención de las regiones en función de Xj y S en árbol de decision*.

Encontrar los valores de Xj y S depende del número de características o variables explicativas que queramos observar en el espacio de predicción. Continuando con la fórmula anterior y aplicado un nuevo corte para otra variable explicativa, dividimos no el espacio de predicción total sino una de las regiones obtenidas tras la división de este con la primera característica. El proceso continúa hasta que descartas todas las características que deseas estudiar (James et al. 2013).

Sin embargo, este método de construcción de árboles de decisión, a medida que vamos añadiendo variables, cada vez se hace más difícil, dando resultados confusos y difíciles de interpretar. Un árbol más pequeño con menos divisiones y ramas nos permite una mejor interpretación reduciendo el sesgo. A este proceso se le conoce como poda[15]. Consiste en construir un árbol muy grande, podándolo de tal forma que obtenemos un subárbol. Nuestro objetivo es, por lo tanto, seleccionar un subárbol que nos conduce a un error de testeo menor.

Figura 11: Árbol de decisión con varios subárboles.



Fuente: elaboración propia mediante R.

[15] Traducción literal de pruning

Estimar y validar cada subárbol es un proceso demasiado largo ya que pueden existir numerosos subárboles. Para eso se emplea el método *Cost Complexity Ppruning*, que consiste en considerar una secuencia de subárboles indexados por un parámetro no negativo (α), en vez de considerar cada subárbol posible (Kolanovic et al. 2107)[16]. El proceso, sin embargo, es ciertamente complejo y merece un análisis más profundo y matemático que el que se puede proporcionar en este trabajo.

Los árboles de decisión tienen muchas ventajas respecto a otras técnicas. La primera es su sencillez a la hora de operar con él ya que se asemejan a la toma de decisiones humanas, a diferencia de regresiones lineares o logarísticas. Por otro lado, la disposición gráfica permite un mayor entendimiento e interpretación, sobre todo si no se analizan muchas variables de la muestra de sujetos. El problema con los árboles reside, sin embargo, en que la falta de dificultad no permite grandes niveles de predicción que otros modelos de regresión. Por otro lado, un cambio en algunas de las variables del árbol rompe con las ramas originarias a partir de la división de la variable, por eso se considera un método poco robusto (James et al. 2013). Algunas de estas desventajas serán analizadas en el caso de estudio.

1.3.2. *Gradient Boosting*.

Tal y como se ha mencionado al comienzo de este apartado, el Gradient Boosting es otro de las varias alternativas que pueden ser aplicados a modelos de Statiscal learning para problemas de regresión y clasificación. Este método implica crear múltiples copias del conjunto de datos, ajustando los árboles de decisión para cada copia y después combinándolos para crear un único modelo predictivo (James et al. 2017). A diferencia del Bagging donde cada árbol se construye independientemente de los otros, en el Gradient Boosting los árboles son creados de forma secuencial en el que cada árbol se desarrolla en función de la información aportada por árboles anteriores.

En comparación con otros métodos, el Gradient Boosting implica un proceso de aprendizaje más lento. En vez de crear un único gran árbol de decisión respecto a todos los datos, se van

[16] Para ver profundizar más este concepto véase Bradford, J. P., Kunz, C., Kohavi, R., Brunk, C., & Brodley, C. E. (1998, April). Pruning decision trees with misclassification costs. In *European Conference on Machine Learning* (pp. 131-136).

ajustando los árboles creados a los residuos del modelo, en lugar de al resultado de la variable Y. Este árbol de decisión se añade posteriormente a la función ajustada para actualizar los residuos, de forma que cada uno de estos árboles puede ser bastante pequeño con solo unos pocos nodos determinados por el parámetro del algoritmo. Mientras vamos ajustando los árboles pequeños a los residuos, vamos mejorado la función f en áreas donde no funciona correctamente (Kolanovic et al. 2017). Las expresiones matemáticas del Boosting son mucho más complejas que para cualquier regresión o árbol de decisión. El objetivo de este trabajo no es realizar un análisis matemático de los métodos si no ver su aplicación al ámbito empresarial, por lo que no se incluyen el desarrollo matemático dentro de este trabajo[17].

Por último, es necesario plantearse como podemos aplicar el Boosting al ámbito empresarial que es el asunto que nos concierne. Por ejemplo, para la predicción del fracaso empresarial se puede construir un clasificador capaz de distinguir las empresas sanas de aquellas que van a fracasar. El planteamiento del aprendizaje automático ante este problema sería el siguiente: Empieza recopilando tantos ejemplos como sea posible de empresas que fracasaron, así como de empresas sanas. A continuación, con estas observaciones y las etiquetas con sus clases se entrena al sistema de aprendizaje elegido que produce un clasificador o regla de predicción. Posteriormente, ante una nueva empresa de clase desconocida, dicho clasificador intenta predecir si fracasará o no. El objetivo, por supuesto, es generar un clasificador que realice las predicciones más precisas posibles en nuevos ejemplos de test. Este mismo proceso se realizará a continuación con el objeto de estudio formado por créditos hipotecarios.

[17] Si se desea conocer el razonamiento matemático, se puede acudir a algunas de las fuentes ya citadas, donde se realiza un análisis exhaustivo acerca del desarrollo de construcción de un Gradient Boosting o de árboles de decisión más complejos como James (2013) o Kolanovic (2017).

6. CASO DE ESTUDIO

En este último apartado se va a realizar el caso de estudio. Una vez definidas las técnicas que se van a emplear y el concepto de riesgo de crédito hipotecario, se busca poner en común lo explicado para observar qué técnica nos permite un mejor análisis del conjunto de datos desde el punto de vista empresarial.

6.1. Datos utilizados

Los datos empleados para el análisis del caso corresponden a datos de panel. Utilizamos los datos observados en el mes en cuestión para predecir la situación de no Default o de Default del prestatario. Por otro lado, el conjunto de datos pertenece a una selección aleatoria de datos de carteras de préstamos hipotecarios recopilados de carteras subyacentes de la titularización de títulos[18]. Los datos fueron extraídos del International Financial Research. Por último, el conjunto contiene observaciones de rendimiento para 1045 prestatarios de hipotecas residenciales de Estados Unidos previas a la crisis del 2007. Es necesario destacar que originalmente la base de datos contenía las características de hasta 622.000 prestatarios. Se decidió realizar un balance de clase y obtener una muestra más pequeña con la que poder trabajar. Una de las razones de esto, fue la cantidad de No Defaults que se obtenían de la muestra más grande. Como es lógico, antes de la recesión, los prestatarios pagaban sus respectivas cuotas, siendo el riesgo de Default mucho menor que por ejemplo dos años más tarde. Con la muestra que vamos a trabajar obtenemos respuestas dispares, que nos permiten identificar dentro de los Defaults que variables eran más significativas y tenían más impacto a la hora de determinar cuando el prestatario incurría en esta posición. Se analizan a continuación las variables estudiadas.

Variable endógena y dependiente.

Como se comentó en el apartado anterior, la variable endógena o dependiente que se desea analizar es el Default. Dicha variable corresponde a una variable categórica y binomial, es decir, toma dos valores. 0 si el prestatario ha pagado la parte correspondiente de su crédito

[18] Es la traducción del título RMBS: residential mortgage backed securities.

ese mes. 1 si el prestatario ha incurrido en Default, es decir, no ha pagado la parte correspondiente de su crédito ese mes.

Variables explicativas

- *Age* (*Edad*): Variable explicativa que mide el tiempo transcurrido desde el origen de la concesión del crédito hipotecario hasta el día de la observación. No muestra por lo tanto la edad del prestatario sino el tiempo de transcurso de la obligación crediticia en meses.
- *Time_to_maturity* (*Madurez*): Variable explicativa que mide el tiempo restante en meses para el vencimiento del crédito hipotecario.
- *Balance* (*Balance*): Variable explicativa que muestra el balance restante en el tiempo de observación. Corresponde a la cantidad de dinero que le queda por pagar al prestatario.
- *LTV*: Variable explicativa que muestra el valor de la hipoteca respecto al valor del bien hipotecado en el momento de la observación, expresada en porcentaje. Es decir, un LTV de un 50% indica que el valor de la hipoteca corresponde a la mitad del valor del bien inmueble. A diferencia de la variable LTV en el origen, esta puede ser distinta ya que según se van realizando los pagos del crédito el LTV del crédito disminuirá.
- *Interest_rate* (*Tipo de interés*): Variable explicativa que indica el tipo de interés del crédito hipotecario en el momento de observación.
- *SingleFamily* (*Familia*): Variable explicativa dicotómica que indica si el prestatario es unifamiliar o no. El valor de la variable será 0 cuando no sea unifamiliar. El valor de la variable será 1 cuando el prestatario sea unifamiliar.
- *Balance_origin* (*Balance en el origen*): Variable explicativa que muestra el balance restante en el tiempo de concesión del crédito. Coincidirá con el valor del crédito que el prestatario tiene que devolver.
- *FICO_origin* (*Puntuación FICO*): Variable explicativa que indica la calificación que otorgada por la agencia Fair Isaac Corporation en el momento de concesión del crédito, para determinar su riesgo y la posibilidad de extender su periodo de vencimiento.

- *LTV_origin* (*LTV en el origen*): Variable explicativa que muestra el valor de la hipoteca respecto al valor del bien hipotecado en el momento de la concesión del crédito, expresado en porcentaje.
- *Interest_Rate_orig* (*Tipo de interés en el origen*): Variable explicativa que indica el tipo de interés del crédito en el momento de la concesión del crédito.

6.2.Metodología

El objetivo de este caso es comprender desde un punto de vista empresarial, cuáles de los métodos descritos en las técnicas de Machine Learning puede ser más útil para una empresa. Existen varios escenarios por la que una empresa puede decantarse por un método u otro. Por ejemplo, si se tienen una cartera de créditos hipotecarios y se desea conocer el riesgo de Default de esa cartera para así cubrirse con otros instrumentos financieros. O si por ejemplo se desea conceder un crédito a un nuevo prestatario, en qué condiciones se debe otorgar para asegurar que todos los meses pagará su cuota y no incurrirá en Default.

Se expondrá a continuación un análisis de la muestra de datos comentada en el apartado anterior mediante la aplicación Gretl y el lenguaje de programación R. Con el conjunto de datos se obtuvo una regresión logística, un árbol de decisión y un gradient Boosting. Si bien es cierto que este último suele ser más preciso, como por ejemplo para la muestra original más grande, las otras técnicas también nos ofrecen resultados interesantes para la muestra de los 1045 prestatarios.

6.3.Resultados

Se introdujo el conjunto de datos en el programa Gretl para la obtención de los principales estadísticos de las variables explicativas, así como la matriz de correlación para observar posibles problemas de multicolinealidad entre ellas.

Tabla 4: Estadísticos principales de las variables explicativas.

	Media	Mediana	Mínimo	Máximo
age	14.581	11.000	0.0000	69.000
time_to_maturity	103.95	108.00	9.0000	158.00
balance	2.5205e+005	1.8155e+005	1551.4	2.9816e+006
LTV	88.153	88.782	0.17858	147.74
interest_rate	7.0084	6.9500	1.8750	14.700
SingleFamily	0.61942	1.0000	0.0000	1.0000
balance_orig	2.5977e+005	1.8800e+005	15600	3.0000e+006
FICO_orig	667.16	668.00	441.00	814.00
LTV_orig	79.432	80.000	50.100	100.50
Interest_Rate_or~	5.7797	6.5000	0.0000	14.700
Default	0.32330	0.0000	0.0000	1.0000

	Desv. Típica.	C.V.	AsimetríaExc. de curtosis	
age	10.860	0.74484	1.0779	1.0015
time_to_maturity	18.341	0.17644	-1.4428	5.9810
balance	2.2699e+005	0.90056	3.6315	27.257
LTV	23.700	0.26885	-0.36436	0.12838
interest_rate	2.0793	0.29669	-0.030569	0.54816
SingleFamily	0.48577	0.78423	-0.49191	-1.7580
balance_orig	2.2881e+005	0.88081	3.5332	26.243
FICO_orig	70.313	0.10539	-0.26251	-0.38300
LTV_orig	9.7295	0.12249	-0.36130	0.82054
Interest_Rate_or~	3.0326	0.52469	-0.72593	-0.19604
Default	0.46796	1.4475	0.75555	-1.4291

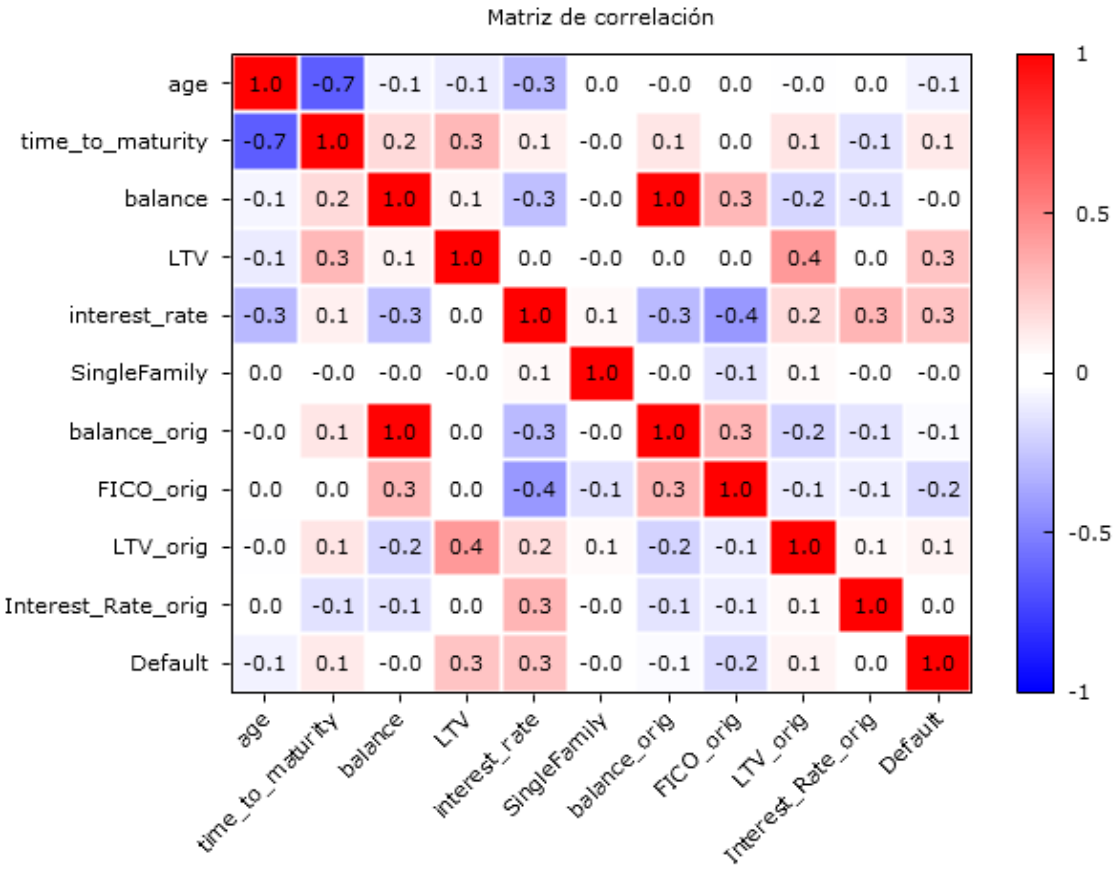
Fuente: Elaboración propia mediante Gretl.

Debemos destacar la media de la variable dependiente, cuyo valor es 0.323, es decir, en media, hubo más prestatarios ese mes que pagaron sus respectivas cuotas. Por otro lado, parece interesante analizar algunos estadísticos de las variables explicativas. La media de la variable *Familia*, 0.619, nos indica que, en la muestra observada, la mayoría de los prestatarios pertenecían a familias monoparentales. La media de la variable *FICO* nos indica que las puntuaciones de los créditos, en media, no eran muy elevadas, justificando en algunas situaciones altas probabilidades de Default, tal y como veremos en el árbol de decisión. En los casos de las variables que son comparadas en el origen y en el momento de observación (*Interés*, *Balance* y *LTV*), el valor de la media es distinto. Para el caso de las variables *Balance* y *Balance en el origen* el valor es casi el mismo. Para los tipos de interés, la media en el momento de observación es mayor que en el origen. Esto puede deberse un aumento de tipos del país o a un ajuste realizado por la entidad prestamista para aumentar la edad del crédito. Respecto a las variables *LTV* y *LTV en el origen*, la media de la primera debería ser menor pues muchos prestatarios han ido pagando sus respectivas cuotas. Esta diferencia puede explicarse bien por el aumento de los tipos de interés, que provocaban que el valor del montante a devolver aumentara, o bien porque en los casos de Default, ambas partes podían

acordar una prórroga de las condiciones del crédito para evitar futuros Defaults. Otro error puede ser el tamaño muestral observado. Tal vez si se analizasen el conjunto original, obtendríamos valores diferentes.

Por otro lado, con el objetivo de identificar problemas de multicolineanlidad en las variables, se examinó la matriz de correlaciones de las variables explicativas. El problema de colinealidad surgiría si alguno de los coefciientes obtenidos superase el 80%.

Figura 12:Matriz de correlación de las variables explicativas.

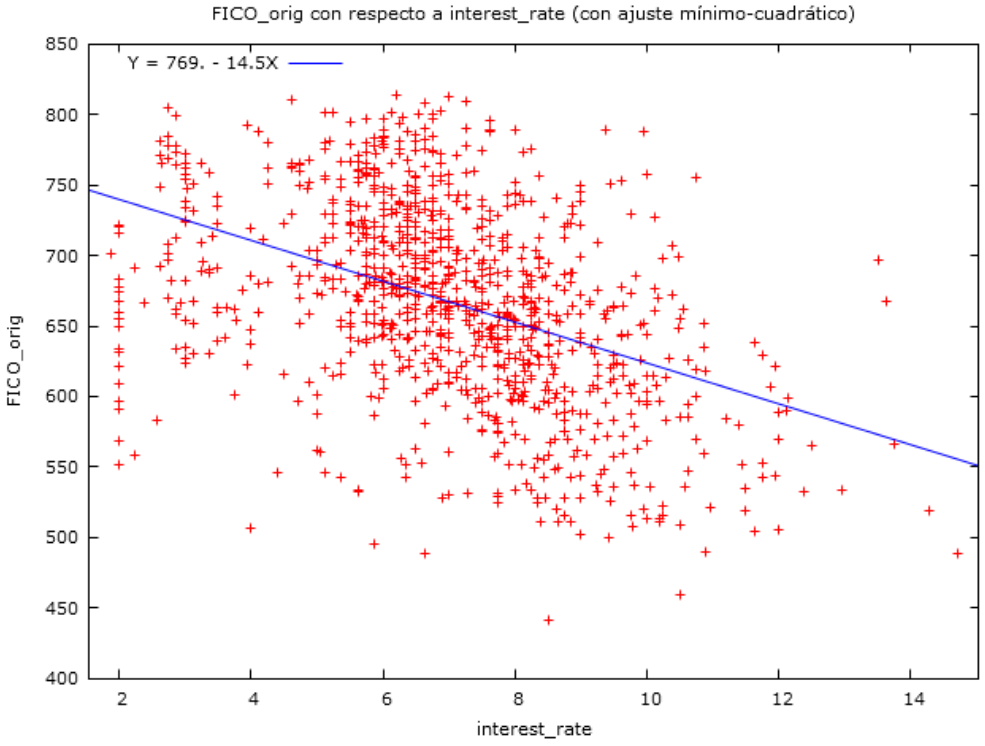


Fuente: Elaboración propia mediante Gretl

A la vista de resultados, solo hay dos variables que supongan grande problemas de multicolinealidad y son fáciles de comprender. Las variables *Balance* y *Balance en el origen* están totalmente relacionadas pues su valor depende del valor de la otra. Una solución a este problema es eliminar alguna de las dos variables, sin embargo, para nuetsro estudio no es necesario. Por otro lado, los valores más altos que alcanzan hasta un 70% de correlación son la edad y el vencimiento del préstamo. Esta relación tiene sentido pues a mayor edad del

crédito, menor será el tiempo que le resta para su vencimiento, pues recordemos que la edad es el número de meses en los que el crédito se ha mantenido activo. Otras correlaciones interesantes a explicar pero que a priori no nos resulta perjudicial para realizar un análisis, se produce entre el LTV y el LTV en el origen (misma explicación que para las variables *Balance* y *Balance en el origen*) y entre el tipo de interés en el momento de la observación con la variable *FICO*. Esto explica la influencia que tiene el tipo de interés del crédito en relación a la puntuación FICO. Como se puede observar en la Figura 13, a mayor tipo de intrés, menor es la calificación FICO:

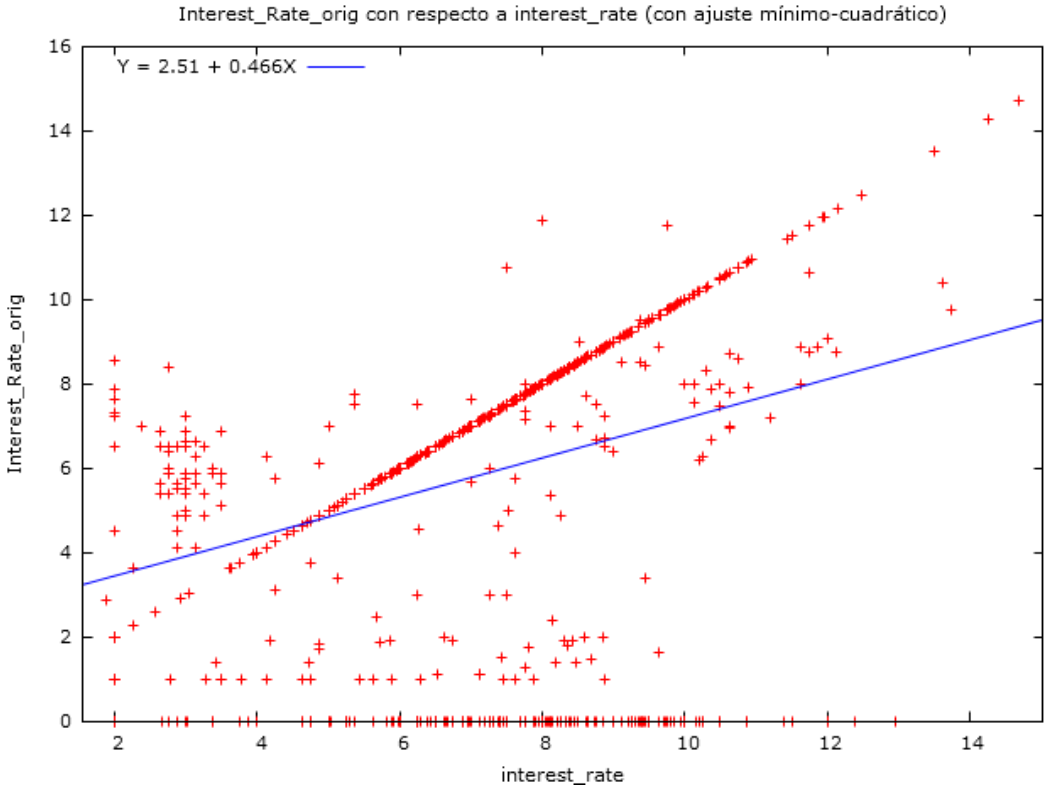
Figura 13:Grafico X-Y Scatter de las variables *FICO* y *Tipo de interés*.



Fuente: Elaboración propia mediante Gretl.

Otra correlación a destacar es el tipo de interés en el momento de observación con los tipos de interés en el momento de concesión. Los valores obtenidos en la Figura 14 nos indican una correlción positiva ya que en la mayoría de las observaciones de la muestra, para tipos deinterés mayoreen el orgien, mayor es el tipo de interés en el momento de la observación.

Figura 14: Gráfico X-Y Scatter de las variables *Tipo de interés y Tipo de interés en el origen*.



Fuente: Elaboración propia mediante Gretl

La variable *Tipo de interés* presenta numerosos análisis y como se verá posteriormente juega un papel fundamental en la determinación de la probabilidad de Default. Otras correlaciones interesantes de esta variable se producen con las variables *Edad* y *Balance*. Estas son negativas por lo que un aumento de los tipos de interés produciría efectos inversos en estas variables.

Por último, puede ser interesante analizar la variable endógena en función de otras variables explicativas. Una relación para estudiar puede ser la implicación que tiene pertenecer a una familia unifamiliar o no, con el Default del crédito:

Tabla 5: Tabulación cruzada de Default (filas) contra *SingleFamily* (columnas).

	[0]	[1]	TOT.
[0]	265	432	697
[1]	127	206	333
TOTAL	392	638	1030

Fuente: Elaboración propia mediante Gretl

A la vista de la figura anterior, observamos que el número total de Defaults en la muestra ascendió a 333, siendo 206 de estos unifamilaires, un 61.8%. Por otro lado, de todos los 697 prestatarios que pagaron su cuota, 432 pertenecían a familias unifamiliares, un 61.97%. Los resultados no nos indican una gran relación pues para los casos de Default y no Default, siempre había un mayor de prestatarios unifamilares.

Una vez obtenidos los estadísticos anteriores con el programa Gretl, se procedió al análisis del conjunto de observaciones con el programa R. Este programa ofrece ciertas ventajas respecto a la aplicación Gretl. Principalmente porque nos permite ver cómo se comporta el modelo en el entrenamiento, para luego ver si su aplicación en el test resulta significativa. El objetivo es, por lo tanto, obtener una predicción de los datos lo mejor posible. Un modelo que obtiene buenos resultados y predice correctamente en el set de entrenamiento, pero en el test los resultados no lo son, no puede ser utilizado como modelo predictivo. Para mejorar la validez del trabajo, se empleó el lenguaje de programación R para mantener constante el conjunto de entrenamiento y de test. Por otra parte, el objeto de estudio del caso es qué modelo de Machine Learning resultaba más favorable desde el punto de vista empresarial, por lo que la gestión y comparación de los modelos fue más sencilla con este código.

Continuando con en análisis, para el set de entrenamiento se seleccionó aleatoriamente de la muestra un 80% de los datos, un total de 824 observaciones. Para el test se empleó el 20% restante, un total de 206 observaciones. Una vez definidos el conjunto de entrenamiento y de test, se procedió a calcular la regresión logística, el árbol de decisión y el gradient Boosting.

Para la regresión logística, se obtuvieron los siguientes resultados:

Tabla 6: Resultado del modelo de regresión logística.

```
Call:
glm(formula = Default ~ ., family = binomial, data = fdata, subset = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9420  -0.8306  -0.5350   1.0136   2.5435

Coefficients:
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)       -1.733e+00  1.542e+00  -1.124  0.26116
age                1.799e-02  1.226e-02   1.467  0.14228
time_to_maturity   5.101e-03  7.626e-03   0.669  0.50354
balance            2.310e-05  7.940e-06   2.910  0.00362 **
LTV                2.709e-02  4.619e-03   5.865 4.49e-09 ***
interest_rate      3.203e-01  5.335e-02   6.004 1.92e-09 ***
SingleFamily      -2.007e-01  1.724e-01  -1.164  0.24430
balance_orig      -2.259e-05  7.921e-06  -2.852  0.00434 **
FICO_orig         -4.472e-03  1.369e-03  -3.266  0.00109 **
LTV_orig          -1.706e-02  1.020e-02  -1.673  0.09431 .
Interest_Rate_orig -3.110e-02  2.829e-02  -1.099  0.27158
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1029.00  on 823  degrees of freedom
Residual deviance:  871.84  on 813  degrees of freedom
AIC: 893.84

Number of Fisher Scoring iterations: 6
```

Fuente: Elaboración propia mediante R

La primera columna obtenida, corresponde al valor de los estimadores. Las caracterísitcas más significativas respecto a la variable endógena de acuerdo con la Tabla 5, son el balance en el momento de observación y en el origen, el *LTV* en el momento de observación, el tipo de interés en el momento de observación y la puntuación FICO. De las cinco variables significativas, los coeficientes de las variables *Balance*, *Tipos de interés* y *LTV* poseen una relación directa con la probabilidad de Default. Si el balance del crédito aumenta, implica que la cantidad a pagar por el prestatario es mayor, por lo tanto, la probabilidad de Default aumentará. Si el LTV del crédito aumenta, esto indica que el porcentaje del valor del bien cubierto por la hipoteca también aumenta, por lo que el prestatario tendrá más posibilidades de incurrir en Default ya que le quedará por pagar más parte del bien. Si el tipo de interés aumentan, las probabilidades de Default también ascienden ya que el prestatario tendrá que pagar unas cuotas mayores ajustadas al tipo de interés aumentado. Las otras dos variables, *Balance* en el origen y *puntuación FICO* implican una relación indirecta con la probabilidad

de Defaul, es decir que un aumento de estas variable supone una disminución de la probabilidad de Default. Si nos fijamos bien todas las variables ex ante, las analizadas en el origen, tienen una relación inversa con la variable dependiente. Una explicación de esto puede ser que como el prestatario no había pagado todavía ninguna parte del crédito y se encontraba en la posición inicial de pago. Por el contrario, las variables ex post, las analizadas en el momento de observación, todas implican relaciones positivas y directas con la variable dependiente.

A continuación, se obtuvieron las matrices de confusión para el conjunto de entrenamiento y test. Las matrices de confusión nos permiten ver la capacidad de predicción del modelo. Se observa por lo tanto las veces que el modelo predijo que el prestatario iba a pagar y pagó o no pagó, y las veces que el modelo predijo que el prestatario iba a incurrir en Default y este incurrió o no. Como es lógico, cuanto más acierte el modelo, mejor será su uso para predecir el comportamiento de las observaciones. Para el conjunto de entrenamiento se obtuvo la siguiente matriz:

Tabla 7: Matriz de confusión del conjunto de entrenamiento para la regresión logística.

```
> #Train set
> Yest.TR.LR=predict(LR ,newdata=fdata[train ,],type = "response")
> Yest.TR.LR <- ifelse(Yest.TR.LR>=0.5,1,0)
> table(Y.TR ,Yest.TR.LR)
    Yest.TR.LR
Y.TR   0   1
   0 496  67
   1 165  96
```

Fuente: Elaboración propia mediante R.

Para el conjunto de entrenamiento en la regresión logística, el modelo predijo que 563 prestatarios pagarían ese mes la cuota del crédito y no incurrirían en Default. De esos 563, 496 realmente pagaron las cuotas de sus créditos, mientras que 67 no. El modelo acertó para los casos de no Default el 88% de los casos. Por otro lado, el modelo predijo que 261 prestatarios incurrirían en Default ese mes, siendo el resultado verdadero 96. Para el caso de Default, el modelo acertó en el 36.7% de los casos. En su totalidad, el modelo acertó 592 predicciones de un total de 824, un 71.84%

Para el conjunto del test se obtuvo la siguiente matriz de confusión:

Tabla 8: Matriz de confusión del conjunto del test para la regresión logística.

```
> #Test set
> Yest.TS.LR=predict(LR ,newdata=fdata[-train ,], type = "response")
> Yest.TS.LR <- ifelse(Yest.TS.LR>=0.5,1,0)
> table(Y.TS ,Yest.TS.LR)
     Yest.TS.LR
Y.TS   0   1
   0 116  18
   1  42  30
```

Fuente: Elaboración propia mediante R

Para el conjunto del test en la regresión logística, el modelo predijo que 134 prestatarios pagarían ese mes su cuota del crédito y no incurrirían en Default. De esos 134, 116 pagaron realmente sus cuotas mientras que 18 no. El modelo acertó para los casos de no Default en el 86.56% de los casos. Por otra parte, el modelo predijo que 72 prestatarios incurrirían en Default ese mes, de los cuales 30 verdaderamente no pagaron las cuotas de sus créditos, mientras que 42 sí lo hicieron. El modelo acertó para los casos de Default un 41.66%. En su totalidad el modelo acertó 146 predicciones de un total de 206, un 70.87%.

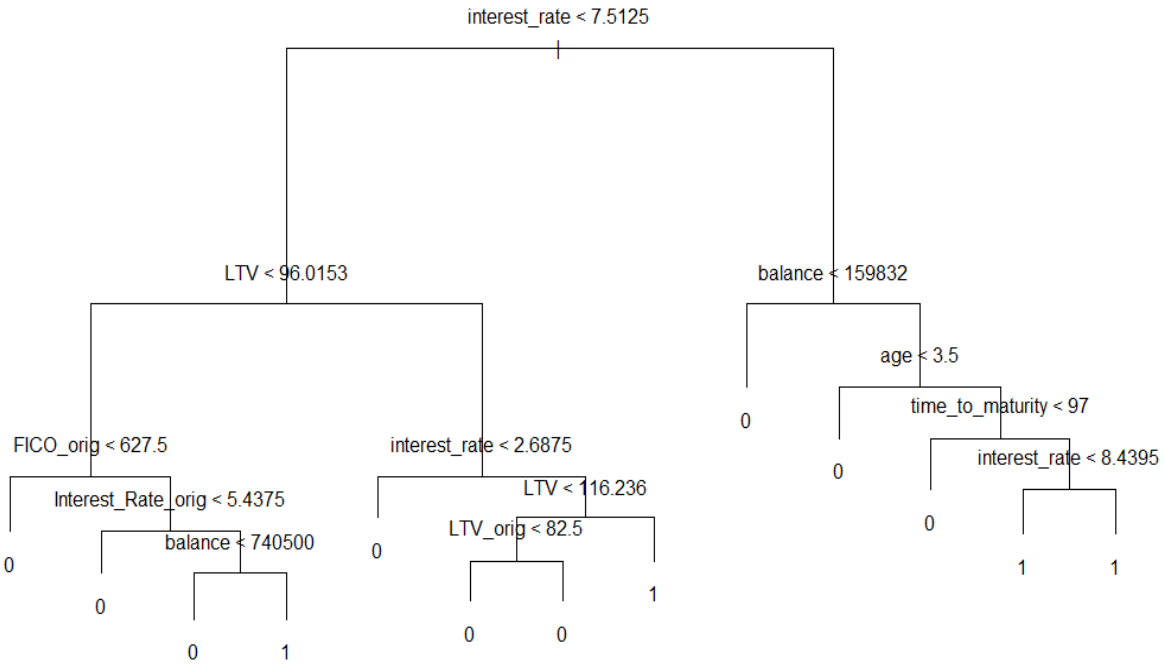
Podemos considerar entonces que, empleando una regresión logística, el modelo obtuvo unos buenos resultados. Tal y como ofrecen los datos, el modelo predice mejor los casos de no Default que de Default. Los resultados son sin duda positivos pues, aunque las predicciones parecen similares, en el conjunto del test, el modelo predijo con mayor exactitud los casos de Default, hasta un 4% más. Sí bien es verdad que el cómputo total, el acierto es un punto porcentual mayor, lo interesante a analizar en este caso es la variable endógena, el Default. Desde el punto de vista empresarial, resulta más interesante conocer cómo se comportará tu modelo respecto a predecir cuantos prestatarios incurrirán verdaderamente en Default que cuántos no lo harán. E incluso para este caso, y siendo la predicción sobre el total del conjunto tan similar para el conjunto de entrenamiento como para el test, el modelo en el test predice mejor los verdaderos casos de Default.

Por último, debemos destacar qué problema presenta emplear la regresión logística para el análisis del conjunto de datos. Si bien las predicciones realizadas para el conjunto de entrenamiento y de test son altas, a diferencia del árbol de decisión, no sabemos qué valores deben tomar las variables explicativas para predecir el Default o no Default. Es decir, los

coeficientes obtenidos nos indican qué variables explicativas son las más significativas y que relación tienen con la variable endógena, sin embargo, no sabemos para que valores de las variables, la probabilidad de Default es mayor. Desde el punto de vista empresarial nos interesa saber el valor de esas variables para analizar posibles escenarios y conceder créditos con unas características determinadas para evitar el Default del prestatario.

El segundo método empleado fue un árbol de decisión. Manteniendo los mismos conjuntos se obtuvo el siguiente árbol de decisión para el conjunto de entrenamiento:

Figura 15: Árbol de decisión



Fuente: Elaboración propia mediante R.

El árbol de decisión ofrece sin duda datos muy interesantes a analizar. Se obtuvieron 13 nodos terminales de los cuales 4, recogían como clase mayoritaria la situación de Default, mientras que en los otros 9, la clase mayoritaria fue no Default. El árbol de decisión nos indicó que, para este conjunto de datos la clase mayoritaria era no Default, ya que para el 68.32% de las observaciones la variable endógena tomaba el valor 0, es decir, no Default.

Por otro lado, se realizaron varios cortes para una misma variable, dos para el tipo de interés y dos para el LTV. La primera división o corte se realizó en función del tipo de interés,

distinguiendo aquellas observaciones en las que el tipo de interés en el momento de observación era menor o mayor que 7.512%. Para los créditos de las observaciones, si el tipo de interés en el momento de la observación fue menor que 7.512%, la probabilidad de no incurrir en Default ascendía a un 79.05%. Para los créditos cuyos tipos de interés en el momento de la observación fuesen mayores de 7.512%, la probabilidad de no incurrir en Default ascendía a un 51.25%. Esto nos indica que el tipo de interés es sin duda, la variable más significativa a tener en cuenta.

Los segundos cortes se produjeron en función de las variables *LTV* y *Balance*. Si el tipo de interés era menor al 7.512%, entonces el segundo corte se producía en función de la variable LTV. Si el valor de la variable en el momento de observación se encontraba por debajo de un 96.01%, la probabilidad de que el prestatario no incurriera en Default era de 88.73%. Por el contrario, si el LTV del crédito era mayor que un 96.01%, la probabilidad de no incurrir en Default ascendía a 62.01%.

Si el tipo de interés era mayor de 7.512% el segundo corte se producía en función de la variable *Balance*. Si el valor de la variable en el momento de observación era menor que 159.832 la probabilidad de que el prestatario no incurriese en Default ascendía a un 62.23%. Sin embargo, si la variable *Balance* alcanzaba valores superiores a 159.832, la situación cambiaba drásticamente ya que entonces la probabilidad de que el prestatario incurriese en Default alcanzaba un 64.61%.

Desde el punto de vista empresarial, es interesante analizar en qué situaciones fue más probable que el prestatario no incurriese en Default. De este modo las empresas pueden establecer las condiciones de los créditos que conceden para evitar en mayor medida el impago de las cuotas y reducir el riesgo de Default.

De acuerdo con el árbol de decisión, en escala de menor a mayor probabilidad de no Default, se obtuvieron los siguientes escenarios:

En el primero, la probabilidad de no Default fue de 95%. Para ello fue necesario que:

- 2.68% < Interés < 7.512%.
- 96.01% < LTV < 116.23%.
- LTV en el origen > 82.5%.

En el segundo, la probabilidad de no Default fue de 98.79%. Para ello fue necesario que:

- Interés < 7.512%.
- LTV < 96.01%.
- FICO > 627.5
- Interés en el origen < 5.437%.

En el tercero, la probabilidad de que el prestatario no incurra en Default fue del 100%. Desde el punto de vista empresarial, si se desea reducir o eliminar el riesgo de Default los créditos deberán cumplir las siguientes características:

- 2.68% < Interés < 7.512%.
- LTV > 96.01%.

Del mismo modo, se obtuvo que, en cuatro nodos terminales la clase mayoritaria fue el Default del prestatario. En escala de menor a mayor probabilidad de Default se obtuvieron los siguientes escenarios:

El primero, para una probabilidad de Default del 62.50% fue necesario que:

- Interés < 7.512%.
- LTV < 96.01%.
- FICO > 627.5.
- Interés en el origen > 5.437%.
- Balance > 740,500.

El segundo, para una probabilidad de Default de 63.41% fue necesario que:

- 2.68% < Interés < 7.512%.
- LTV > 116.23%.

En el tercer escenario se estudió la última rama del árbol, aquella que cumplía las siguientes características:

- Interés > 7.512%.
- Balance > 159,832.
- Edad > 3.5 meses.

- Madurez > 97 meses.

En esta rama la clase mayoritaria obtenida fue Default, con un 77.32% de probabilidad. Dentro de esta misma rama, se obtuvieron dos nodos terminales, uno de los cuales nos indicaba la situación para una probabilidad de Default más alta. Mientras que para un interés menor que 8.43%, la probabilidad de Default fue de 63.26%, para un tipo de interés superior al 8.43%, la probabilidad de Default obtenida fue de un 91.66%.

Desde el punto de vista empresarial, el análisis proporcionado por el árbol de decisión es de gran utilidad. Sabiendo que, de los 1045 casos estudiado, la clase mayoritaria es no Default con un 68.32% de probabilidad, deducimos que la probabilidad de Default para ese conjunto es de 31.67%. El empresario por ejemplo sabe que, si concede un crédito con un interés menor al 7.512%, es muy probable que el prestatario no incurra en Default. Si por otro lado desea obtener la mayor rentabilidad de ese crédito y acuerda con el deudor un interés mayor a ese umbral, debe saber que para reducir el riesgo de Default debe reducir la edad del crédito, su vencimiento, u observar si el balance a fecha de observación es mayor o menor que 740,500 euros. Si en cambio desea concretar un LTV mayor con el prestatario sin que este último incurra en Default, deberá fijarse en que para LTV menores de 96.01%, el riesgo de Default se reduce drásticamente si los tipos de interés son bajos y si el crédito tiene una puntuación FICO mayor que 627.5.

Una vez analizado esto, se obtuvieron las matrices de confusión para los dos conjuntos, con el objetivo de comprar si el modelo predecía mejor las observaciones empleando una regresión logística, o bien un árbol de decisión.

Respecto al conjunto de entrenamiento, se obtuvo la siguiente matriz de confusión:

Tabla 9: Matriz de confusión del conjunto de entrenamiento para el árbol de decisión

```
> #Train set
> Yest.TR.tree=predict(tree ,newdata=fdata[train ,],type ="class")
> table(Y.TR ,Yest.TR.tree)
    Yest.TR.tree
Y.TR   0   1
   0 523  40
   1 155 106
```

Fuente: Elaboración propia mediante R.

Para el conjunto de entrenamiento del árbol de decisión, el modelo predijo que 563 prestatarios pagarían ese mes la cuota del crédito y no incurrirían en Default. De esos 563, 523 realmente pagaron las cuotas de sus créditos, mientras que 40 no. El modelo acertó para los casos de no Default el 92.89% de los casos. Por otro lado, el modelo predijo que 261 prestatarios incurrirían en Default ese mes, siendo el resultado verdadero 106. Para el caso de Default, el modelo acertó en el 40.61% de los casos. En su totalidad, el modelo acertó 629 predicciones de un total de 824, un 76.33%.

Respecto al test, se obtuvo la siguiente matriz de confusión:

Tabla 10: Matriz de confusión del conjunto del test para el árbol de decisión.

```
> #Test set
> Yest.TS.tree=predict(tree ,newdata=fdata[-train ,],type ="class")
> table(Y.TS ,Yest.TS.tree)
    Yest.TS.tree
Y.TS   0   1
   0 115  19
   1  50  22
```

Fuente: Elaboración propia mediante R.

Para el conjunto del test del árbol de decisión, el modelo predijo que 134 prestatarios pagarían ese mes su cuota del crédito y no incurrirían en Default. De esos 134, 115 pagaron realmente sus cuotas mientras que 19 no. El modelo acertó para los casos de no Default un 85.82% de los casos. Por otra parte, el modelo predijo que 72 prestatarios incurrirían en Default ese mes, de los cuales 22 verdaderamente no pagaron las cuotas de sus créditos, mientras que 50 sí lo hicieron. El modelo acertó para los casos de Default un 30.55%. En su totalidad el modelo acertó 137 predicciones de un total de 206, un 66.50%.

El árbol de decisión nos permite identificar que variables pueden ser modificadas para reducir el riesgo de Default del crédito. Desde un punto de vista empresarial, parece más interesante el árbol que la regresión logística pues la obtención de las probabilidades para las situaciones anteriormente descritas es de gran utilidad a la hora de valorar las características con las que se concede el crédito. El problema que hemos obtenido en este análisis fue las diferencias obtenidas entre el conjunto de entrenamiento y en el test. Las diferencias obtenidas en los conjuntos aplicando la regresión logística, eran menores comparadas con las obtenidas empleando un árbol de decisión, obteniendo una diferencia de predicción total de casi un 10%. El modelo sigue prediciendo muy positivamente los casos de no Default, tanto en el

conjunto de entrenamiento como en el test, sin embargo, la predicción de no Default en el test se encuentra por debajo de un 31%. Aplicando el mismo razonamiento de la regresión logística, si bien es fundamental que el modelo prediga correctamente los no Defaults, lo que se busca desde un punto de vista empresarial, es conocer con mayor exactitud los casos de Default que vas a tener para así cubrir el riesgo, y esto se predice mejor aplicando la regresión logística. A pesar de lo anterior, la información que nos ofrece el árbol a través de los cortes con las variables y los nodos es de mayor utilidad que las estimaciones obtenidas por la regresión.

El tercer método empleado para analizar el conjunto de datos que se empleó fue un Gradient Boosting. Manteniendo el mismo conjunto de entrenamiento y de test, se obtuvo la siguiente información acerca de las variables explicativas:

Tabla 11: Resultados del Gradient Boosting.

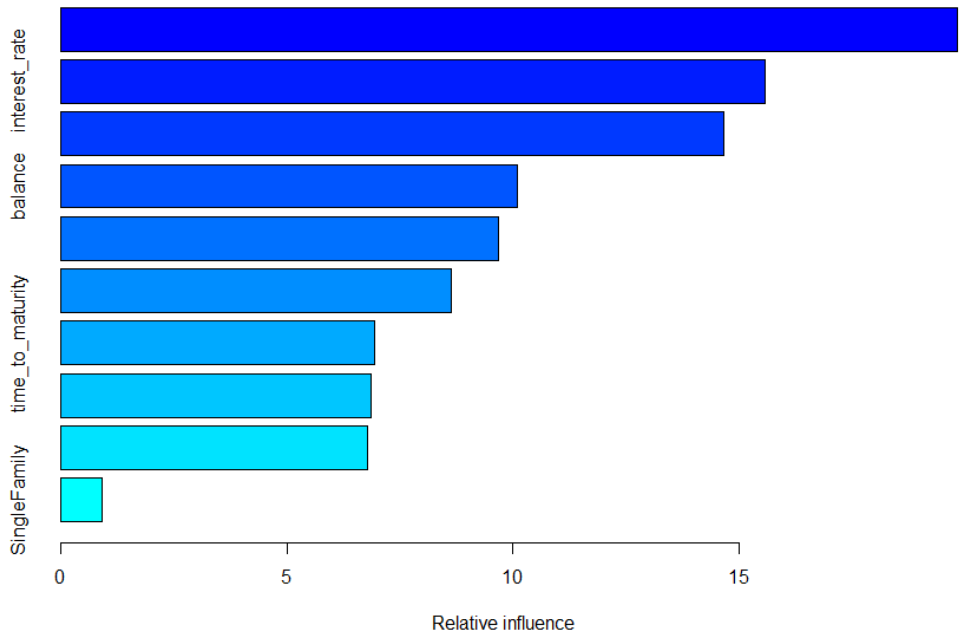
```
> library (gbm)
> set.seed(1)
> fdata$Default <- as.integer(fdata$Default)
> boost = gbm(Default ~ ., data=fdata[train ,],
+             distribution="bernoulli",n.trees =500 , interaction.depth =4)
> #Variable importance
> summary(boost)
                                  var    rel.inf
LTV                               LTV 19.8306366
interest_rate           interest_rate 15.5840405
FICO_orig                   FICO_orig 14.6571725
balance                       balance 10.1035657
balance_orig             balance_orig  9.6831083
Interest_Rate_orig Interest_Rate_orig  8.6332065
time_to_maturity     time_to_maturity  6.9472212
LTV_orig                     LTV_orig  6.8637464
age                               age  6.7822818
SingleFamily             SingleFamily  0.9150207
```

Fuente: Elaboración propia mediante R.

El Boosting nos permite conocer cuáles son las variables con mayor influencia o que aportan mayor información en la predicción de la probabilidad de Default. Las cuatro variables que más peso tienen son, de menor a mayor importancia, el balance en el momento de observación, la puntuación FICO, el tipo de interés en el momento de observación y el LTV en la observación del crédito. No es casualidad que tres de las variables que más información aportan a la probabilidad obtenidas mediante el Boosting, coincidan con las tres características sobre las que se realizaban los primeros cortes en el árbol de decisión.

Si comparamos los resultados obtenidos en la Figura 16 con los obtenidos en el árbol de decisión se repiten ciertos comportamientos de algunas variables.

Figura 16: Información relativa de las variables explicativas proporcionada por el Gradient Boosting.



Fuente: Elaboración propia mediante R.

El caso más evidente es la relevancia de la variable familia. El Boosting nos indica que es la variable que aporta menos información sobre el Default del prestatario. Si observamos a su vez el árbol de decisión, la variable familia no produce ninguna división del árbol y no parece ser determinante a la hora de determinar el mayor caso posible. Lo mismo sucede con el resto de las variables. La importancia que nos indica los coeficientes hallados en el Boosting coincide con la posición de las variables en el árbol de decisión, siendo las más relevantes, aquellas que producen las primeras divisiones de las ramas.

Por lo tanto, tanto el árbol de decisión como el Gradient Boosting nos permiten conocer de forma más precisa la influencia de las variables explicativas sobre la variable endógena, en comparación con la regresión logística.

Una vez obtenida la información relativa de las variables, se obtuvieron las matrices de confusión para el conjunto de entrenamiento y el conjunto del test, para observar si mediante

esta técnica, las predicciones del modelo resultaban más precisas que mediante la regresión logística y el árbol de decisión.

Respecto al conjunto de entrenamiento, se obtuvo la siguiente matriz de confusión:

Tabla 12: Matriz de confusión del conjunto de entrenamiento para el Gradient Boosting.

```
> #Train set
> Yest.TR.boost=predict(boost ,newdata=fdata[train ,],n.trees =500, type = "response")
> Yest.TR.boost <- ifelse(Yest.TR.boost>=0.5,1,0)
> table(Y.TR ,Yest.TR.boost)
    Yest.TR.boost
Y.TR   0   1
   0 556   2
   1  10 256
```

Fuente: Elaboración propia mediante R.

Para el conjunto de entrenamiento del Gradient Boosting, el modelo predijo que 558 prestatarios pagarían ese mes la cuota del crédito y no incurrirían en Default. De esos 558, 556 realmente pagaron las cuotas de sus créditos, mientras que 2 no. El modelo acertó para los casos de no Default el 99.64% de los casos. Por otro lado, el modelo predijo que 266 prestatarios incurrirían en Default ese mes, siendo el resultado verdadero 256. Para el caso de Default, el modelo acertó en el 96.24% de los casos. En su totalidad, el modelo acertó 812 predicciones de un total de 824, un 98.54%.

Respecto al conjunto del test, se obtuvo la siguiente matriz de confusión:

Tabla 13: Matriz de confusión del conjunto del test para el Gradient Boosting.

```
> #Test set
> Yest.TS.boost=predict(boost ,newdata=fdata[-train ,],n.trees =500, type = "response")
> Yest.TS.boost <- ifelse(Yest.TS.boost>=0.5,1,0)
> table(Y.TS ,Yest.TS.boost)
    Yest.TS.boost
Y.TS   0   1
   0 107  32
   1  44  23
```

Fuente: Elaboración propia mediante R.

Para el conjunto del test del Gradient Boosting, el modelo predijo que 139 prestatarios pagarían ese mes su cuota del crédito y no incurrirían en Default. De esos 134, 107 pagaron realmente sus cuotas mientras que 32 no. El modelo acertó para los casos de no Default un

79.85% de los casos. Por otra parte, el modelo predijo que 67 prestatarios incurrirían en Default ese mes, de los cuales 23 verdaderamente no pagaron las cuotas de sus créditos, mientras que 44 sí lo hicieron. El modelo acertó para los casos de Default un 34.32%. En su totalidad el modelo acertó 130 predicciones de un total de 206, un 63.10%.

El Boosting por lo tanto nos permite identificar que variables explicativas nos aportan más información sobre la probabilidad de Default. Sin embargo, no nos ofrece la precisión de valores que obtuvimos con el árbol de decisión. Por otro lado, los resultados del modelo en el conjunto de entrenamiento son casi perfectos, alcanzo porcentajes de acierto cercanos al 100%, sin embargo, esta precisión no se traslada al conjunto de test, que es donde verdaderamente podemos observar la capacidad predictiva del modelo. Las diferencias en los porcentajes nos obligan a pensar que el Boosting está sobreentrenado. Esto implica que el modelo se ha aprendido los datos analizados y que al introducirle nuevos datos no será capaz de analizarlos correctamente. Si los datos fuesen un examen tipo test, el modelo en vez de aprenderse el contenido del examen se aprendió el patrón de respuesta de las preguntas por lo que, al introducirle un examen nuevo, contestaría del mismo modo que el examen anterior. Generalmente, y a pesar de los resultados obtenidos en el presente caso de estudio, el Boosting suele realizar mejores predicciones que las técnicas anteriores. Para resolver este problema se debería ajustar los parámetros del modelo o tal vez aumentar el tamaño muestral para que analizase una muestra más grande.

6.4. Conclusiones del caso de estudio

Se expone en la Tabla 13, el resumen de los datos obtenidos para las matrices de confusión de los tres métodos empleados:

Tabla 14: Resumen de resultados de las matrices de confusión

	Regresión logísitca		Árbol de decisión		Gradient Boosting	
	Entrenamiento	Test	Entrenamiento	Test	Entrenamiento	Test
No Default acertados	88,10%	86,57%	92,90%	85,82%	99,64%	79,85%
Default acertados	36,78%	41,67%	40,61%	30,56%	96,24%	34,33%
Total	71,84%	70,87%	76,33%	66,50%	98,54%	63,11%

Fuente: Elaboración propia mediante Excel.

Como se puede observar, la técnica más eficaz a la hora de predecir los casos es la regresión logística, seguido por el árbol de decisión. Para nuestro caso de estudio debemos descartar el Gradient Boosting como método de referencia. En primer lugar, porque pese a que los datos obtenidos para el conjunto de entrenamiento son sin duda los mejores, estos no se trasladan al test. La diferencia en el total de predicciones respecto al árbol de decisión es de un 3%, pero el árbol ofrece un mejor análisis de las variables explicativas. En segundo lugar, debemos descartar el Boosting ya que, pese a ofrecer qué variables presentan una mayor importancia relativa, la regresión logística también nos indica qué variables son más significativas y el árbol de decisión indica el valor de las variables en cada corte.

Debemos plantearnos entonces qué técnica nos ofrece más ventajas para este modelo y este conjunto de datos. A la vista de los resultados y las predicciones concluimos que el árbol de decisión es la mejor técnica con la que proceder. En primer lugar, porque a pesar de que las predicciones de las matrices de confusión son peores (diferencias de hasta un 11%), no existe gran diferencia en el porcentaje de predicción total (un 4%). En segundo lugar, y la razón con más peso, es el análisis ya comentado que proporciona el árbol de decisión, pues esclarece situaciones que permiten al prestamista modificar las características del crédito que concede.

Debemos recordar que el análisis de este caso se ha realizado constantemente bajo un punto de vista empresarial, interpretándose en todo momento su aplicación para el negocio. Si el análisis se hubiese centrado más en la predicción exquisita de los valores de la variable dependiente, sin duda se hubiese concluido que la técnica más acorde sería la regresión

logística. Sin embargo, desde el punto de vista empresarial, los valores y las situaciones proporcionadas por el árbol de decisión permiten al empresario jugar con diferentes supuestos y gestionar el riesgo de crédito de una mejor forma.

7. CONCLUSIONES

Tras haber realizado la investigación sobre las técnicas de Machine Learning y el riesgo de crédito y haber realizado el caso de estudio, podemos obtener las siguientes conclsuiones:

1) La evolución de la Inteligencia Artificial hasta nuestros ha seguido un crecimiento exponencial. Desde las primeras regresiones lineares descritas por Gauss, hasta la generación de programas de computación y creación de algoritmos capaces de "aprender" por sí solos.
2) El concepto de Inteligencia Artificial abarca muchos campos, en los que se incluye el aprendizaje automático o Machine Learning. Es un error afirmar que estos dos campos de estudio son lo mismo.
3) El riesgo de crédito hipotecario jugó un papel fundamental en la crisis de 2007. El número de Defaults en créditos concedidos bajo la denominación de hipotecas Subprime alcanzó niveles nunca vistos, provocando el estallido de la burbuja inmobiliaria y de la recesión económica
4) A partir de los Acuerdos de Basilea, la concesión de créditos se regularizó evitando prácticas abusivas o *soft lending*, en las que se concedían créditos hipotecarios sin ni siquiera analizar las situaciones financieras de los prestatarios.
5) El riesgo de crédito no solo hace referencia al impago de la cuota por parte del prestatario. Existen otros tipos de riesgo relacionados con este como es el hipotecario, el corporativo o el del consumidor.
6) No existe un modelo del cálculo del riesgo de crédito unificado. Cada empresa, en función de los algoritmos o técnicas que utilice decide qué variables son las que se debe analizar para ver si el deudor incurrirá en Default o no.
7) La evolución de los modelos de cálculo del riesgo de crédito ha sido significativa, desde simples fórmulas matemáticas hasta técnicas de Machine Learning complejas donde algoritmos son capaces de predecir correctamente el número posible de Defaults dentro de una cartera de créditos.

8) La regresión logística o modelo Logit es mucho más interesante y útil para analizar variables cualitativas binomiales que el modelo de regresión lineal y que el modelo de probabilidad lineal.
9) El modelo Logit ha servido de base para el cálculo de riesgo de crédito en métodos modernos.
10) Los árboles de decisión son técnicas fáciles de manejar que nos proporcionan diferentes escenarios en función de las variables explicativas.
11) El Gradient Boosting es de las técnicas analizadas, sin duda la más compleja y permite una mejor predicción de Default que los otros métodos.
12) Para nuestro caso de estudio se obtuvieron las siguientes conclusiones
    a. Las tres técnicas obtuvieron muy buenos resultados para el conjunto de entrenamiento, destacando por encima el Gradient Boosting, indicándonos posiblemente que el modelo estuviese sobreentrenado.
    b. Los resultados para el conjunto del test fueron peores que para el conjunto de entrenamiento.
    c. Lastres técnicas obtienen mejores resultados de predicción para el número de no Defaults que para el número de Defaults.
    d. La técnica que mejores predicciones ofrecía fue la regresión logística, sin embargo, el análisis de las variables explicativas no proporcionaba gran utilidad
    e. El árbol de decisión fue la segunda técnica que mejores predicciones ofreció mejorando en análisis de las variables explicativas obtenidos para la regresión logística y para el Gradient Boosting
    f. El Gradient Boosting obtuvo muy buenos resultados en el conjunto de entrenamiento, sin embargo, los resultados en el test fueron decepcionantes.
13) Sin duda desde un punto de vista empresarial, la técnica más recomendable es el árbol de decisión por sus buenos resultados de predicción y por el análisis de las variables explicativas y cómo influyen a la hora de establecer la probabilidad del suceso de la variable endógena.

8. BIBLIOGRAFÍA

Aguilar, L. J. (2016). *Big Data, Análisis de grandes volúmenes de datos en organizaciones*. Mexico D.F: Alfaomega Grupo Editor. 09/03/2020

Altman, E. I., Iwanicz-Drozdowska, M., Laitinen, E. K., & Suvas, A. (2017). Financial distress prediction in an international context: A review and empirical analysis of Altman's Z-score model. *Journal of International Financial Management & Accounting*, *28*(2), 131-171.

Andrés Suárez, J. (2000). Técnicas de Inteligencia Artificial aplicadas al análisis de la solvencia empresarial. *Documentos de trabajo (Universidad de Oviedo. Facultad de Ciencias Económicas)*, 206, 1-31.

Bagherpour, A. (2017). Predicting mortgage loan Default with machine learning methods. *University of California/Riverside*. 1-30.

Berlin, M. (1992): Securitization. En: Newman, P., Milgate, M., & Eatwell, J. (eds). *The New Palgrave Dictionary of Money and Finance,* 3, (pp. 433-435). Basingstoke: Macmillan 13/03/2020.

Bishop, C. M. (2006). *Pattern recognition and Machine Learning*. Nueva York: Springer. 09/03/2020

Breiman, L., Friedman, J., Stone, C. J., & Olshen, R. A. (1984). *Classification and regression trees*. Londres: Chapman & Hall/CRC.

Bughin, J., Chui, M., & Manyika, J. (2010). Clouds, big data, and smart assets: Ten tech-enabled business trends to watch. *McKinsey quarterly*, *56*(1), 75-86.

Carling, K., Jacobson, T., Lindé, J., & Roszbach, K. (2007). Corporate credit risk modeling and the macroeconomy. *Journal of banking & Finance, 31*(3), 845-868.

Castaño, H. F., & Ramírez, F. O. P. (2005). El modelo logístico: una herramienta estadística para evaluar el riesgo de crédito. *Revista Ingenierías Universidad de Medellín*, *4*(6), 55-75. 06/03/2020

Castillo, E., Gutiérrez, J. M., & Hadi, A. S. (1997). *Sistemas expertos y modelos de redes probabilísticas.* Madrid: Academia Española de Ingenierıa.

Chorafas, D. N. (2000). *Managing credit risk*. Londres: Euromoney Books.

Contreras, S., & De la Rosa, F. (2016, September). Aplicación de Deep Learning en robótica móvil para exploración y reconocimiento de objetos basados en imágenes. *Colombian Computing Conference (CCC)*, 11, 1-8.

De la Fuente, M. D. L. (1999). *La administración integral de riesgos financieros*. Universidad Iberoamericana, Departamento de Economía.

Dumitrescu, E., Hue, S., Hurlin, C., & Tokpavi, S. (2018). Machine Learning for Credit Scoring: Improving Logistic Regression with Non-Linear Decision *Tree Effects. Doctoral dissertation, PhD thesis. Paris Nanterre University, University of Orleans.* 1-29.

Elizondo, A., & Lopez, C. (1999). El riesgo de crédito en México: una evaluación de modelos recientes para cuantificarlo. *Gaceta de economia*, *4*(8), 51-74.

Elul, R. (2005). The economics of asset securitization. *Business Review*, 3, 16-25. 13/03/2020

García, M. L. S., & García, M. J. S. (2010). Modelos para medir el riesgo de crédito de la banca. *Cuadernos de administración*, *23*(40), 295-319.

Garzón, G. R. (2003). Los orígenes del método de mínimos cuadrados. *Suma: Revista sobre Enseñanza y Aprendizaje de las Matemáticas*, (43), 31-38. 07/03/2020

Gorton, G. (2009). The subprime panic. *European Financial Management*, *15*(1), 10-46. 13/03/2020

Hastie, T., & Tibshirani, R. (1987). Generalized additive models: some applications. *Journal of the American Statistical Association*, *82*(398), 371-386.

Henke, S., Burghof, H. P., & Rudolph, B. (1998). Credit securitization and credit derivatives: financial instruments and the credit risk management of middle market commercial loan portfolios. *Center for Finicial Studies Working*. 7, 1-30. 13/03/2020

Hinestroza Ramírez, D. (2018). *El Machine Learning a través de los tiempos, y los aportes a la humanidad.* Risaralda: Universidad Libre Seccional Pereira. 09/03/2020

Hull, J. (2012). *Risk management and financial institutions*. Nueva Jersey*:* John Wiley & Sons. 12/03/202 .

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning*. New York: Springer.

Khandani, A. E., Kim, A. J., & Lo, A. W. (2010). Consumer credit-risk models via machine-learning algorithms. *Journal of Banking & Finance*, *34*(11), 2767-2787.

Keys, J., Mukherjee, T., Seru, A., Vig, V. Did securitization lead to lax screening? Evidende for the Subprimer Loans. *Quaterly Hournal of Economics. 125*(1), 307-362.

Kolanovic, M., & Krishnamachari, R. T. (2017). Big data and AI strategies: Machine learning and alternative data approach to investing. *JP Morgan Global Quantitative & Derivatives Strategy Report*. 1-280.

Legendre, A. M. (1805). *Nouvelles méthodes pour la détermination des orbites des comètes*. Lausana: F. Didot.

Lopez, J.A. (2004). The empirical relationship between Averrage Asset Correlation, Firm Probability of Default and Asset Size. *Journal of Financial Intermediation. 13*(2), 265-283.

Martínez de Ibarreta, C., Álvarez Fernandez, C., Budría Rodriguez, S., Curto Gonzalez, T., Escobar Torres, L., Borrás Pala, F. (2017). *Modelos cuantitativos para la Economía y la Empresa en 101 ejemplos.* Madrid: EV Services.

Mayer-Schönberger, V., & Cukier, K. (2013). *Big data: la revolución de los datos masivos*. Madrid: Turner. 09/03/2020

McAfee, A., Brynjolfsson, E., Davenport, T. H., Patil, D. J., & Barton, D. (2012). Big data: the management revolution. *Harvard business review*, *90*(10), 60-68.

McCarthy, J.; Minsky, M.; Rochester, N.; Shannon, C. (1955). A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence. *AI magazine*, *27*(4), 12-12. *09/03/2020*

McCulloch, W. S. & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, *5*(4), 115-133. 09/03/2020.

Mian, A. & Sufi, A. (2009). The consequences of mortgage credit expansion: Evidence from the US mortgage Default crisis. *The Quarterly Journal of Economics, 124*(4), 1449-1496. 12/03/2020

Monleón-Getino, A. (2015). El impacto del Big-data en la Sociedad de la Información. Significado y utilidad. *Historia y comunicación social*, *20*(2), 427. 09/032020

Nájera, A. B. U., & de la Calleja Mora, J. (2017). Brief review of educational applications using data mining and machine learning. *Redie. Revista Electrónica de Investigación Educativa*, *19*(4), 84-96.

Nelder, J. A., & Wedderburn, R. W. (1972). Generalized linear models. *Journal of the Royal Statistical Society: Series A (General), 135*(3), 370-384.

Ong, M. K. (Ed.). (1999). *Internal credit risk models: Capital allocation and performance measurement*. Londres: Risk publications.

Pennington-Cross, A., & Ho, G. (2010). The Termination of Subprime Hybrid and Fixed-Rate Mortgages. *Real Estate Economics*, *38*(3), 399-426. 12/03/2020

Plazas Wadynski, M. A. (2017). Machine Learning Para la predicción de series temporales en indicadores de Desarrollo Mundial. *Universidad Industrial de Santander, Escuela De Ingeniería de Sistemas*, 1-61.

Provost, F., & Kohavi, R. (1998). Glossary of terms. *Journal of Machine Learning*, *30*(2-3), 271-274. 09/03/2020

Reiz, A. N., de la Hoz, M. A., & García, M. S. (2019). Big Data Analysis y Machine Learning en medicina intensiva. *Medicina Intensiva*, *43*(7), 416-426.

Russom, P. (2011). Big data analytics. *TDWI best practices report, fourth quarter*, *19*(4), 1-34. 09/03/2020

Samuel, A. L. (1959). Some studies in Machine Learning using the game of checkers. *IBM Journal of research and development*, *3*(3), 210-229. 12/03/2020

Sanders, A. (2007). The subprime crisis and its role in the financial crisis. *Journal of Housing Economics*, *17*(4), 254-261.

Sirignano, J., Sadhwani, A., & Giesecke, K. (2016). Deep learning for mortgage risk. *Stanford University*, 1-83.

Turing, A. (1964). Computer machinery and intelligence. En: Epstein, R; Roberts, G & Beber, G. (dirs.): *Parsing the Turing Test: Philosophical and Methodological Issues in the Quest for the Thinking Computer*. Nueva York: Springer. 12/03/2020.

Zimmerman, T. The Great Subprime Meltdown. *Journal of structured Finance*, 7, 7-20. 12/03/2020.

9. ANEXOS

9.1. Anexo 1: Descripción del Código empleado en R.

```
fdata <- read.csv("MortgageClass.csv", header=TRUE, sep = ";")



#-------------------------------- REMOVE OUTLIERS  ----------------------------------------

# Create a new dataset removing negative numbers of balance_origin (row).
fdata <- fdata[-fdata$balance_orig<0,]
str(fdata)
summary(fdata)

#-------------------------------- MISSING VALUES ----------------------------------------

# In addition, the function na.omit() directly obtains the dataset without NA in all variables.
fdata <- na.omit(fdata)
summary(fdata)



#-------------------------------- TRAINING MODELS ----------------------------------


#Take 80% of data for training
train <- sample(dim(fdata)[1] , 0.8*dim(fdata))

#Output Variable in train and test
Y.TR <- fdata$Default[train]
Y.TS <- fdata$Default[-train]
```

```
## Logistic Regression ----------
LR =glm(Default ~ ., data=fdata, subset = train, family =binomial)
summary (LR)
#Train set
Yest.TR.LR=predict(LR ,newdata=fdata[train ,],type = "response")
Yest.TR.LR <- ifelse(Yest.TR.LR>=0.5,1,0)
table(Y.TR ,Yest.TR.LR)
#Test set
Yest.TS.LR=predict(LR ,newdata=fdata[-train ,], type = "response")
Yest.TS.LR <- ifelse(Yest.TS.LR>=0.5,1,0)
table(Y.TS ,Yest.TS.LR)



## Decision tree ----------
library(tree)
fdata$Default <- as.factor(fdata$Default)
tree =tree(Default ~ ., data=fdata, subset = train)
#tree plot
plot(tree)
text(tree,pretty =0)
#Cuts
tree

#Train set
Yest.TR.tree=predict(tree ,newdata=fdata[train ,],type ="class")
table(Y.TR ,Yest.TR.tree)
```

```
#Test set
Yest.TS.tree=predict(tree ,newdata=fdata[-train ,],type ="class")
table(Y.TS ,Yest.TS.tree)
```

```
## Gradient boosting ----------
library (gbm)
set.seed(1)
fdata$Default <- as.integer(fdata$Default)
boost = gbm(Default ~ ., data=fdata[train ,],
        distribution="bernoulli",n.trees =500 , interaction.depth =4)
#Variable importance
summary(boost)

#Train set
Yest.TR.boost=predict(boost ,newdata=fdata[train ,],n.trees =500, type = "response")
Yest.TR.boost <- ifelse(Yest.TR.boost>=0.5,1,0)
table(Y.TR ,Yest.TR.boost)
#Test set
Yest.TS.boost=predict(boost ,newdata=fdata[-train ,],n.trees =500, type = "response")
Yest.TS.boost <- ifelse(Yest.TS.boost>=0.5,1,0)
table(Y.TS ,Yest.TS.boost)
```

9.2. Anexo 1: Descripción del árbol de decisión obtenido con el código en R.

```
> tree
node), split, n, deviance, yval, (yprob)
      * denotes terminal node

 1) root 824 1029.000 0 ( 0.68325 0.31675 )
   2) interest_rate < 7.5125 506  519.400 0 ( 0.79051 0.20949 )
     4) LTV < 96.0153 327  235.000 0 ( 0.88379 0.11621 )
       8) FICO_orig < 627.5 40   50.450 0 ( 0.67500 0.32500 ) *
       9) FICO_orig > 627.5 287  169.800 0 ( 0.91289 0.08711 )
        18) Interest_Rate_orig < 5.4375 83   10.830 0 ( 0.98795 0.01205 ) *
        19) Interest_Rate_orig > 5.4375 204  147.800 0 ( 0.88235 0.11765 )
          38) balance < 740500 196  124.800 0 ( 0.90306 0.09694 ) *
          39) balance > 740500 8   10.590 1 ( 0.37500 0.62500 ) *
     5) LTV > 96.0153 179  237.700 0 ( 0.62011 0.37989 )
      10) interest_rate < 2.6875 11    0.000 0 ( 1.00000 0.00000 ) *
      11) interest_rate > 2.6875 168  226.800 0 ( 0.59524 0.40476 )
        22) LTV < 116.236 127  161.200 0 ( 0.66929 0.33071 )
          44) LTV_orig < 82.5 107  142.400 0 ( 0.61682 0.38318 ) *
          45) LTV_orig > 82.5 20    7.941 0 ( 0.95000 0.05000 ) *
        23) LTV > 116.236 41   53.850 1 ( 0.36585 0.63415 ) *
   3) interest_rate > 7.5125 318  440.600 0 ( 0.51258 0.48742 )
     6) balance < 159832 188  249.300 0 ( 0.62234 0.37766 ) *
     7) balance > 159832 130  168.900 1 ( 0.35385 0.64615 )
      14) age < 3.5 25   31.340 0 ( 0.68000 0.32000 ) *
      15) age > 3.5 105  123.800 1 ( 0.27619 0.72381 )
        30) time_to_maturity < 97 8    6.028 0 ( 0.87500 0.12500 ) *
        31) time_to_maturity > 97 97  103.900 1 ( 0.22680 0.77320 )
          62) interest_rate < 8.4395 49   64.440 1 ( 0.36735 0.63265 ) *
          63) interest_rate > 8.4395 48   27.540 1 ( 0.08333 0.91667 ) *
```