CAPÍTULO III

Aprendizaje automático en modelos de concesión de crédito: oportunidades y riesgos

Andres Alonso-Robisco
Jose Manuel Carbo

En el presente capítulo analizamos las oportunidades y riesgos que supone la aplicación de modelos de aprendizaje automático (*ML,* por sus siglas en inglés) en la concesión de crédito. Para ello realizamos una revisión guiada de la literatura, y acompañamos la discusión con la realización de un ejercicio empírico con una base de datos de libre acceso. Por un lado, la mayor capacidad de predicción de *ML* puede suponer un impacto económico, tanto por la provisión de nuevo negocio, como por el ahorro en capital regulatorio que una entidad financiera podría obtener si adoptara *ML* en sus modelos internos. El mayor rendimiento predictivo de ***ML*** también puede conllevar efectos positivos en la inclusión financiera, aunque estos resultados deben ser valorados junto con nuevos riesgos que vendrían de la mano de esta tecnología, como son especialmente la discriminación (sesgos) entre clases sociales, o la explicabilidad de los resultados, la cual analizaremos usando nuevas técnicas de interpretabilidad. Concluimos con una reflexión sobre la necesidad de lograr una satisfactoria explicabilidad de los modelos como condición necesaria para la puesta en producción de esta tecnología financiera, dando confianza a los usuarios de sistemas automatizados de concesión de crédito.

Palabras clave: machine learning, aprendizaje automático, *credit scoring,* explicabilidad.

1. INTRODUCCIÓN

El uso de modelos de aprendizaje automático o *machine learning (ML)* está ganando terreno en finanzas (Alonso-Robisco & Carbo, 2022c; Fernández, 2019; Gimeno y Sevillano, 2022) impulsado en gran medida por la suma de dos factores. En primer lugar, por el acceso a grandes conjuntos de datos (caracterizados por volumen, variedad, y variabilidad), los cuales requieren una tecnología escalable para su almacenamiento eficiente, tratamiento y posterior análisis. En segundo lugar, por los avances en computación en la nube y el uso de nuevas técnicas de modelización estadística como la inteligencia artificial (IA), de la que podemos considerar el *ML* como un subconjunto. En concreto, es a partir de la década de los ochenta cuando se populariza este tipo de modelos, permitiendo que los ordenadores aprendan automáticamente de los datos sin instrucciones directas de un ser humano, simulando un comportamiento inteligente (Estévez, 2022). Al igual que los modelos de econometría tradicional, el objetivo principal del *ML* es extraer información de los datos y hacer predicciones, si bien la gran diferencia entre ambos paradigmas de modelización estadística es que el uso de *ML* permite tratar el proceso que gobierna los datos como desconocido (Breiman, 2001). El énfasis, por tanto, no reside en modelos teóricos que describen el comportamiento supuesto de las variables, dejándose dicha tarea a complejos algoritmos que aprenden iterativamente el patrón que describen los datos.

Este cambio en el paradigma presenta oportunidades y riesgos a la hora de su implementación en un entorno financiero. Existe una extensa literatura académica que respalda el potencial de esta tecnología para resolver un elevado número de problemas dentro de la gestión del riesgo de crédito (Goodell *et al.,* 2001; Königstorfer y Thalmann, 2020), como por ejemplo la predicción de impagos, el establecimiento de sistemas de alerta temprana, o la prevención del fraude e identificación de comportamientos anómalos. En concreto, recientes encuestas identifican que una de las áreas donde más interés existe por parte de las entidades financieras para el uso de *ML* es en la concesión de préstamos (Institute of International Finance, 2019; Jung *et al.,* 2019). Precisamente, la posibilidad de sumar registros comportamentales de los clientes (por ejemplo, a través de la huella digital) hace de esta actividad un entorno ideal para un buen rendimiento de los modelos de ML: alta dimensionalidad de los datos y relaciones no lineales entre las variables que describen el comportamiento de los individuos (Berg *et al.,* 2020).

En este contexto, las autoridades y supervisores financieros han identificado el uso del *ML* como un área con gran potencial para la innovación financiera, incluyéndolo dentro de su ámbito de seguimiento (Alonso-Robisco & Carbo, 2022c; Gimeno y Sevillano, 2022). De hecho, podemos destacar iniciativas de experimentación desarrolladas en el BIS Innovation Hub como el proyecto Ellipse, el cual trata de combinar fuentes de información estructurada y no estructurada con datos regulatorios existentes para mejorar la información sobre los riesgos sistémicos.

Sin embargo, el elevado potencial de esta tecnología no está exento de riesgos. Son numerosas las referencias académicas (Blattner, Nelson y Spiess, 2021; Tyagi, 2022) y super-

visoras (BAFIN, 2022; EBA, 2021; Dupont, Fliche y Yang, 2020; Institute of International Finance, 2018) que describen nuevos factores de riesgo asociados a esta tecnología que suscitan preocupación. Los modelos de *ML* son cada vez más complejos, y su naturaleza de "caja negra" les confiere un riesgo de modelo inherentemente nuevo (Alonso-Robisco & Carbo, 2022a) en particular en el campo de la concesión de crédito, debido a las necesidades de explicabilidad de los resultados, y al control de los posibles sesgos o discriminación. En este sentido, el enfoque regulador trata de evaluar el riesgo asociado al uso de sistemas basados en IA a través de la evaluación del potencial impacto negativo para los usuarios de los mismos. Así, bajo la pionera Directiva de Inteligencia Artificial propuesta por la Comisión Europa (legislación conocida como *AI Act*) un recomendador automático de productos de inversión o un *chatbot* para la interacción con los clientes no tendría tanto impacto potencial negativo como un sistema de evaluación de riesgo crediticio (*credit scoring*), cuyas potenciales decisiones de denegación de un préstamo pueden dañar la solvencia de empresas e individuos. En este último caso se establece como prioritario la gestión del riesgo de modelo, y el buen gobierno de la tecnología, existiendo el requerimiento de cumplir con un test de conformidad regulatorio (Floridi *et al.,* 2022).

Dado el elevado potencial de esta tecnología para mejorar la capacidad de predicción, así como su incipiente uso por parte de la industria financiera, y atención captada por los supervisores financieros, en este capítulo repasamos el estado actual de la investigación sobre el diseño e implementación de sistemas automáticos de evaluación y concesión de crédito basados en *ML*, realizando una revisión guiada de la literatura académica, unido a un ejercicio empírico con una base de datos de crédito de libre acceso en https://www.kaggle.com. La sección 2 comienza analizando las principales ventajas derivadas de su uso, a saber la mejor capacidad de predicción de impagos, y su potencial para una mayor inclusión financiera. Posteriormente, en la sección 3 se revisan las consecuencias no intencionadas del uso de *ML*, en particular centrándose en los problemas de discriminación o sesgos entre clases sociales protegidas que se pueden ver perjudicadas por la ausencia de juicio humano en la concesión de crédito, y en la explicabilidad de sus resultados, donde se expone el problema sobre la discrepancia en las explicaciones que actualmente existe en la literatura académica, ilustrándolo con un ejercicio empírico. Por último, en la sección 4 se concluye con una serie de recomendaciones para economistas, expertos en *ML*, reguladores y supervisores financieros.

2. OPORTUNIDADES

En esta sección, vamos a analizar las oportunidades que ofrece el uso de *ML* en concesión de crédito. En primer lugar, analizamos las ganancias en capacidad de predicción con respecto a modelos tradicionales, y lo hacemos con una revisión de la literatura reciente, y con la realización de un ejercicio empírico. En segundo lugar, traducimos estas ganancias predictivas a impacto económico. Y finalmente, analizamos el posible efecto de *ML* en la inclusión financiera.

2.1. Mejor capacidad de predicción

2.1.1. Revisión de la literatura

Existen numerosos artículos académicos que analizan el uso de algoritmos de *ML* para la predicción de impagos en comparación con técnicas estadísticas tradicionales. Este apartado comienza repasando algunos de estos artículos a raíz de un resumen realizado en Alonso-Robisco y Carbo (2022a). En el cuadro 1 se muestra un extracto de dichos artículos, junto con el tipo de préstamo subyacente (corporativos, consumo, o hipotecas), el tamaño de la muestra, el rendimiento estadístico de los modelos de *ML*, y el que obtienen con un *Logit* tradicional. En estos artículos el rendimiento estadístico se mide normalmente usando la métrica denominada *AUC-ROC*[1], que es un método de evaluación de sistemas de clasificación. La curva ROC se representa mediante la tasa de verdaderos positivos (*TPR*, por sus siglas en inglés) y la tasa de falsos positivos (*FPR*, por sus siglas en inglés), para todos los posibles umbrales de discriminación (Fawcett, 2006). En particular:

$$TPR = TP / (TP+FN)$$

$$FPR = FP / (FP+TN)$$

Donde *TP* (*true positives*) son los préstamos que, habiendo hecho impago, están correctamente estimados como tal; *FN* (*false negatives*) son los préstamos que, habiendo hecho impago, son incorrectamente estimados como no impagados; *FP* (*false positives*) son los préstamos que no impagaron, pero fueron estimados como impagos; y *TN* (*true negatives*) son los préstamos que no impagaron y correctamente se estimaron como no impagados. Así, para cada posible umbral, si un préstamo tiene una probabilidad de impago mayor que dicho umbral clasificaremos a ese préstamo como impagado. De este modo, cuanto menor es el umbral, mayor será la tasa *TPR* y menor la *FPR* (parte superior derecha de la curva ROC). Igualmente, a mayor umbral menor será la tasa *TPR* y mayor será la tasa *FPR* (parte inferior izquierda de la curva ROC).

Como alternativa al *AUC-ROC,* se muestra también en ocasiones la precisión, que es la tasa de verdaderos positivos entre el total de verdaderos y falsos positivos. Todas estas métricas, al representar tasas de acierto o fallo, tienen un rango igual a una probabilidad de ocurrencia, es decir, entre 0 y 100, y un mayor valor indica un mejor rendimiento estadístico.

Lo primero que destaca es que en todos los estudios revisados se encuentra una mejora en predicción estadística al usar modelos de *ML* (como *random forest, gradient boosting* o redes neuronales profundas) respecto a *Logit.* Estas ganancias pueden ser de hasta el 20 %. Observamos también que no necesariamente los modelos más complejos, como aquellos basados en redes neuronales, predicen mejor. De hecho, los resultados de estos artículos

[1] Por sus siglas en inglés, *Area Under the Curve of the Receiver Operator Characteristic.*

Cuadro 1.

Revisión de la literatura: precisión *ML* en predicción de impagos

Autor, año, revista	*Activo subyacente*	*Tamaño muestra*	*Predicción* ML	*Predicción* Logit
Jones, Johnstone y Wilson (2015) *Journal of Banking and Finance*	Préstamos corporativos	5.000 empresas en 20 años	*Random forest* 93 % AUC	83 % AUC
Petropoulos *et al.* (2019) *ECB, Working paper*	Préstamos corporativos	200.000 empresas	*Gradient boosting and neural net* 78 % AUC	66 % AUC
Sigrist y Hirnschall (2019) *Journal of Banking and Finance*	Préstamos corporativos	850 préstamos a 141 pymes	*Grabit (Gradient boosting and Tobit)* 83% AUC	66 % AUC
Moscatelli *et al.* (2020) *Expert Systems with Applications*	Préstamos consumo	300.000 empresas	*Random forest* 75,9 % AUC	73,2 % AUC
Butaru *et al.* (2016) *Journal of Banking and Finance*	Préstamos consumo	1 millón de préstamos	*Random forest* 66,6 % Recall	59,2 % Recall
Kvamme *et al.* (2018) *Expert Systems with Applications*	Hipotecas	20.000 hipotecas	Convolutional neural net 91,5 % AUC	86,6 % AUC
Sirignano y Cont (2019) *Quantitative Finance*	Hipotecas	120 millones de hipotecas	*Grabit (Gradient boosting and Tobit)* 83 % AUC	66 % AUC
Albanesi y Vamossy (2019) *NBER, Working paper*	Préstamos consumo	1 millón de préstamos	*Neural net* 90 % Precision	86 % Precision

sugieren que los modelos basados en árboles de decisión como *Random forest* o *XGBoost* tienen mejor rendimiento predictivo.

2.1.2. Ejercicio empírico

Cabe resaltar que los datos usados por estos artículos son de variada naturaleza, difieren en tipo de préstamo subyacente y en el tamaño de muestra (desde unos pocos miles de préstamos a millones de ellos). Por lo tanto, con el objetivo de mejorar la comparabilidad de los resultados observados, en este apartado vamos a utilizar diferentes modelos de *ML* (árbol de decisión[2], *XGBoost,* red neuronal profunda) y un modelo *Logit* para predecir los impagos en una misma base de datos, de libre acceso, de crédito al consumo. En concreto, utilizamos la base de datos *Give me some credit*, accesible en la plataforma Kaggle.com. La base de datos forma parte de uno de los concursos de predicción más famosos de *credit scoring* en Kaggle,

[2] Forzamos al árbol de decisión a tener únicamente tres ramas, para que sea interpretable.

publicado en 2011. En las bases del concurso se menciona que se trata de datos de préstamos al consumo, aunque no se confirma si son datos reales. Los datos se componen de un total de 100.000 préstamos etiquetados con una variable objetivo de impago con una frecuencia del 6 %. Esta variable objetivo, *SeriousDlqin2yrs,* es una variable binaria que determina si el préstamo resultó fallido o no, en función de si el prestatario ha tardado más de 90 días en realizar el pago. El resto de variables cuantitativas, cuyas características se muestran en el cuadro 2, son las siguientes. *Revolving* se refiere al saldo total en tarjetas de crédito y líneas de crédito personales excepto patrimonio y sin deuda a plazos. Las variables *Age* y *Dependents* aluden a la edad del prestatario y al número de dependientes en la familia (cónyuge, hijos, etc.). Las variables *MonthlyIncome, DebRatio* y *CreditLines* referidas a la renta mensual, al pago mensual de deuda, pensión alimenticia y costos de vida divididos por el ingreso bruto mensual, y al número de préstamos abiertos (a plazos, como préstamos para automóviles o hipotecas) y líneas de crédito (por ejemplo, tarjetas de crédito), respectivamente. Las variables *30-59Days* y *90Days* describen las veces que el prestatario se ha retrasado de 30 a 59 días, y de 60 a 89 días, respectivamente. Por último, *RealEstate* se refiere al número de préstamos hipotecarios, incluidas las líneas de crédito con garantía hipotecaria.

Cuadro 2.

Descripción de la *Give me some credit*

	Default	*Revolving*	*Age*	*30-59Days*	*DebtRatio*	*Monthly Income*	*Credit Lines*	90Days	*RealEstate*	*Dependents*
mean	0.068	5.952	51.246	0.385	26.921	6635	8.752	0.216	1.054	0.852
std	0.253	266.619	14.410	3.558	394.013	13615	5.163	3.525	1.155	1.149
min	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
25 %	0.000	0.035	40.000	0.000	0.143	3400	5.000	0.000	0.000	0.000
50 %	0.000	0.178	51.000	0.000	0.296	5400	8.000	0.000	1.000	0.0000
75 %	0.000	0.581	61.000	0.000	0.482	8226	11.000	0.000	2.000	2.000
max	1.000	50708.000	103.000	98.000	61106.500	3008750	58.000	98.000	54.000	20.000

Para realizar este ejercicio, se realiza una partición de 80 % para entrenar[3], y 20 % para testear, utilizando los modelos anteriormente mencionados y evaluando su rendimiento en la muestra de test a través de diferentes métricas estadísticas. En concreto, se usa el anteriormente mencionado *AUC-ROC,* el *recall* y el *F1 score* (otra métrica con rango entre 0 y 100, que combina las métricas de *recall* y precisión). En el cuadro 3 se muestran los resultados. Lo primero que llama la atención es que el *Logit,* el modelo estadístico tradicional, es el que peor rendimiento tiene en las tres medidas. Por ello, se propone utizar un *Logit* al cuadrado, en el que seleccionamos las once variables explicativas e incluimos las variables al cuadrado para

[3] De la muestra de entrenamiento, se selecciona un 20 % para validar. En esa muestra de validación, se eligen entre diferentes arquitecturas o hyperparámetros de los modelos de *ML* en función del *AUC* que consiguen. En el caso del árbol de decisión, el número de ramas se mantiene fijo en tres para que se pueda interpretar. En *XGBoost,* se elige la combinación de número de árboles y máximo de ramificaciones que obtiene un mayor *AUC* en la muestra de validación, y en el caso de las redes neuronales profundas, el número de capas internas y cantidad de nodos por capa.

ayudar al mismo a captar posibles no linealidades entre las variables. Su rendimiento mejora así considerablemente (medido en cualesquiera de las tres métricas), pero como veremos más adelante, la inclusión de variables al cuadrado empeorará la interpretabilidad del modelo, lo cual en principio era una de sus ventajas. En cualquier caso, los tres modelos de *ML* propuestos obtienen una mejor predicción estadística que los modelos *Logit,* siendo el modelo con mejor *AUC-ROC* y *F1 score* el *XGBoost*. Concluimos de esta manera que los resultados están en línea con la literatura académica antes mencionada, sumando evidencia de que los modelos más complejos de *ML* no necesariamente se comportan mejor en la clasificación de impagos.

Cuadro 3.

***Give me some credit:* métricas de rendimiento**

	Logit	*Logit cuadrado*	*Árbol decisión*	*XGBoost*	*Red neuronal profunda*
AUC-ROC	0,64	0,78	0,81	0,84	0,82
TPR / Recall	0,61	0,61	0,71	0,74	0,75
F1	0,17	0,31	0,31	0,32	0,20

2.2. Impacto económico

Llegado este punto nos preguntamos, ¿cómo traducir esta mejora estadística en impacto económico? Existe un reducido número de estudios que tratan de monetizar el impacto de usar modelos de *ML* utilizados para tomar decisiones crediticias. Una de las propuestas pioneras se recoge en Khandani, Kim y Lo (2010) quienes calculan ahorros en costes de entre el 6 % y el 25 % medidos a través de las pérdidas totales por impagos usando datos de tarjetas de crédito, o Albanesi y Vamossy (2019) quienes usan redes neuronales en un modelo de evaluación crediticia obteniendo mejor rendimiento estadístico, y ahorros de hasta el 9 % medido a través de una métrica llamada "Valor Añadido", siguiendo la siguiente ecuación, que usan para calcular el resultado neto derivado de los aciertos y errores del modelo utilizado para tomar las decisiones de concesión de créditos:

$$VA = TN \cdot B_d - FN \cdot B_r \cdot P_m \qquad [1]$$

B_r: saldo pagadores; P_m: margen beneficios; B_d: saldo impagado; *TN*: verdaderos negativos; *FN*: falsos negativos.

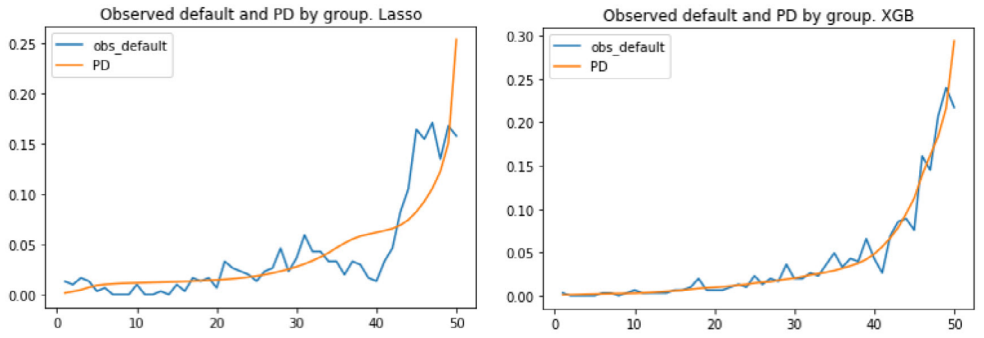
No obstante, es importante destacar que estas métricas de ahorro son retrospectivas, ya que sobre una cartera ya concedida se calcula el impacto en términos de ahorro entre un escenario donde se utiliza un modelo de *ML,* y un escenario en el que no se usa *ML*. Por lo tanto, no se trata de un importe económico que una entidad financiera pudiera materializar para valorar si implementar esta tecnología o no. Otra posible forma de calcular el impacto de predecir mejor los impagos es el posible ahorro en capital regulatorio. Un reciente estudio (Fraisse y Laporte, 2022) usa *ML* para predecir la probabilidad de impago de una cartera de bonos corporativos de una entidad financiera francesa y analiza los ahorros en capital regu-

latorio que dicha entidad podría obtener si usara *ML* bajo un enfoque de modelos internos (*Internal Rating Based* o *IRB*), los cuales alcanzarían hasta un 25 %. De manera similar, en nuestro estudio (Alonso-Robisco & Carbó, 2022b) se utiliza una base de datos de crédito al consumo de un banco español para estimar igualmente el ahorro de capital regulatorio, obteniendo hasta un 17 % en comparación con un cálculo del riesgo de esa exposición crediticia usando *Logit*.

Para ilustrar el porqué de estos ahorros, vamos a fijarnos en la figura 1 de nuestro artículo (Alonso-Robisco & Carbo, 2022b). En esta figura se dividen las predicciones obtenidas con regresión logística penalizada vía *Lasso*[4] (a la izquierda) y con *XGBoost* (a la derecha). Las predicciones se dividen en 50 grupos de igual número de préstamos, de forma ascendente en función de la probabilidad estimada de impago. Es decir, en el primer grupo están las predicciones con menor probabilidad de impago, y en el último grupo están las predicciones con mayor probabilidad de impago. La línea naranja indica la media de probabilidad estimada de cada grupo, y la línea azul la media de impagos observados en cada grupo. Cuanto más cercanas estén estas líneas, mejor será la predicción. Se puede observar a simple vista que las predicciones de *Lasso* infraestiman y sobreestiman la frecuencia observada de impagos. También, que la probabilidad de impago presenta una distribución donde en muchos grupos apenas cambia, no se diferencia bien entre préstamos. De este modo, teniendo en cuenta que el capital regulatorio es una función cóncava y creciente de la probabilidad de impago estimada, con el modelo tipo *Lasso* el capital regulatorio es mayor porque (i) tiende a sobreestimar para probabilidades de impago bajas y, (ii) no se pueden organizar los préstamos en grupos tan pequeños como se podría hacer con *XGBoost,* y dada la concavidad de la fórmula de capital regulatorio, cuantos más grupos se crean (mayor granularidad), más bajo será el capital regulatorio.

Figura 1.

***Logit:* variables más importantes según coeficiente**



[4] Un modelo paramétrico lineal que, a diferencia de *Logit,* incluye mecanismos para eliminar variables con poco poder predictivo.

2.3. Mayor inclusión financiera

La automatización de las decisiones de crédito ofrece la posibilidad de democratizar el acceso a los servicios financieros. Por inclusión financiera la literatura económica distingue habitualmente tres aspectos: el acceso, la disponibilidad y el uso de servicios financieros (Sarma, 2008). Además, la exclusión financiera suele tener carácter involuntario y estar motivada por la carencia de ahorros, el coste de los servicios, la distancia o la ausencia de documentación, o la falta de educación financiera, entre otros. No obstante, también existen causas culturales y religiosas, o motivos como la desconfianza en el sistema financiero (Demirguc-Kunt *et al.,* 2018).

Una de las vías más importantes por las que la tecnología financiera permite influir en la inclusión financiera es principalmente a través de la recopilación granular de datos. Por ejemplo, el uso de la información con origen en la huella digital abriría posibilidades de financiación a individuos residentes en zonas no bancarizadas (Berg *et al.,* 2020) , pero con recurso a otros proveedores financieros como plataformas de préstamos *online* (*crowdlending*), donde el uso de *ML* es cada vez más extendido (véase la base de datos de crédito para competiciones entre científicos de datos de Lending Club, en EE. UU.), o neobancos como el brasileño Nubank. En este sentido, Philippon (2019) apunta a que el *big data* tiene el potencial de reducir los prejuicios negativos en la concesión de crédito, pero, a su vez, podría perjudicar la protección de determinadas minorías sociales. Asimismo, (Barruetabeña, 2020) concluye que la nueva generación de servicios financieros accesibles a través de los móviles e internet está contribuyendo al progreso en el sector financiero.

Dicha transformación se está dando, además, tanto por actores tradicionales como por nuevos agentes que están irrumpiendo en el mercado, ya sean grandes empresas tecnológicas (*bigtech*) o pequeñas compañías especializadas (*fintech*). La evidencia empírica muestra que una mayor competencia en la provisión de servicios financieros está reduciendo los costes de información y transformación asociados a la actividad crediticia (Philippon, 2019). Por ejemplo, un reciente trabajo (Fuster *et al.,* 2022) estudia la diferencia en la concesión de crédito hipotecario en EE. UU. entre *fintech* y bancos tradicionales, encontrando que los primeros son capaces de procesar la información hasta un 20 % más rápido sin aumentar la exposición a riesgo de impago. En esta línea, otros investigadores (Buchak *et al.,* 2018) encuentran que las empresas *fintechs* suelen atender a clientes con mayor nivel de renta, cobrando un diferencial de tipo de interés de entre 14-16 puntos básicos, concluyendo que los consumidores están dispuestos a pagar por decisiones rápidas y mejor experiencia de usuario. Así, esta parte de la literatura académica apunta a que las empresas *fintech* han mejorado la eficiencia en la provisión de crédito hipotecario, mostrándose como una vía prometedora para reducir la desigualdad en el acceso a financiación.

Es importante resaltar que el origen de esta mejora en la inclusión financiera no vendría únicamente a partir de la captura y uso de nuevos datos, sino también por el recurso a nuevas técnicas y modelos de *ML* que permiten captar nueva información de los mismos datos. Así queda de relieve en diversos estudios como Huang *et al.* (2020) , donde se concluye que usar *big data* junto con modelos de *ML* ofrece ventajas significativas en la predicción de impagos de pequeñas y medianas empresas (pymes) en China, tanto por el uso de nuevos

datos alternativos (ventaja de información), como por la mayor flexibilidad de los nuevos métodos de predicción (ventaja de modelo). De manera similar, otros autores (Gambacorta *et al.,* 2019) sugieren que los algoritmos de *ML* pueden capturar mejor las relaciones no lineales entre variables en tiempos de estrés. Esto es especialmente relevante para determinadas clases sociales como los más jóvenes, quienes suelen carecer de historial crediticio, lo cual les imposibilita acceder a oportunidades de inversión (por ejemplo, en formación) a través de intermediarios financieros tradicionales. Así se pone de relieve en Agarwal *et al.* (2020) que describe como el uso de algoritmos de *ML* y los datos alternativos (en particular, el historial de navegación en el móvil), está facilitando en EE. UU. la predicción de impagos y el consiguiente acceso a financiación de la generación *millenial.* En esta misma línea, Qi y Xiao (2018) describen cómo la tecnología de *ML* aplicada para la concesión de microcréditos en China (de nuevo, donde los datos financieros son escasos o nulos), está mejorando la calidad de vida de las personas. En definitiva, existe evidencia que apunta hacia el potencial del *ML* para contribuir a un incremento responsable de inclusión financiera, no obstante, existiendo riesgos asociados a la adopción de esta tecnología (Bazarbash, 2019).

3. RIESGOS

Existe una extensa literatura que alerta de nuevos riesgos a la hora de aplicar *ML* en riesgo de crédito (BAFIN, 2022; Dupont, Fliche y Yang, 2020), desde la estabilidad de las predicciones, pasando por la transparencia de los algoritmos, o la privacidad de los datos. Para una correcta clasificación de estos factores de riesgo, y comprender cuáles son nuevos o genuinamente debidos al uso de esta tecnología financiera, podemos acudir a nuestro trabajo previo (Alonso-Robisco & Carbo, 2022a), donde se revisan y clasifican los factores asociados al uso de *ML* en tres categorías: estadística, tecnología y conducta de mercado, tal y como se muestra en el cuadro 4. Estos factores de riesgo están asociados a los requisitos que se exigen a los modelos cuantitativos usados en los esquemas de validación de modelos internos (*IRB*). Por ejemplo, en la validación de modelos de *IRB*, se exige que los modelos cuantitativos usados no sobreajusten, y sean interpretables, lo que se puede traducir en factores de riesgo a la hora de aplicar modelos de *ML*. De esta forma, encontramos hasta 13 potenciales factores de riesgo derivados del uso de *ML*.

Cuadro 4.
Factores de riesgo de modelo

Estadística	*Tecnología*	*Conducta de mercado*
Estabilidad	Transparencia	Privacidad de datos
Hiperparámetros	Huella de carbono	Auditabilidad
Sobreajuste	Dependencia de proveedores	Interpretabilidad
Calibración dinámica	Cyberataques	Sesgos
Preprocesamiento de datos		

Fuente: Alonso-Robisco y Carbó (2022a).

En concreto, dos de estos factores destacan por encima del resto (Institute of International Finance, 2018 y 2019), como son la interpretabilidad de los resultados y la discriminación financiera (sesgos), de los cuales hablaremos en las siguientes secciones.

3.1. Discriminación financiera (sesgos)

Recientes estudios (Fuster *et al.,* 2022) apuntan a que las últimas innovaciones en tecnología estadística (asociadas a *ML*) han provocado un aumento en el riesgo de impactos redistributivos en clases sociales protegidas como religión, género o raza. La pregunta sobre si las decisiones automatizadas de concesión de crédito promueven o mitigan la discriminación es especialmente relevante en el mercado de crédito al consumo, dada su repercusión sobre el bienestar de las familias. Estos impactos pueden venir por la mayor flexibilidad para descubrir relaciones estructurales o por triangulación de otras características previamente excluidas. Por ejemplo, a través del código postal se podría capturar la pertenencia a una determinada etnia en determinadas ciudades. De este modo, a pesar de eliminar variables sensibles de nuestra base de datos, de forma no intencionada podríamos incumplir políticas sociales de igualdad.

En este sentido, en Fuster *et al.* (2022), los autores se preguntan si son más laxas las condiciones de los préstamos que conceden las empresas *fintech* o tienen mejores sistemas de evaluación crediticia usando *ML,* encontrando que los prestamistas negros e hispanos están desproporcionadamente más desfavorecidos por la introducción de *ML.* En esta misma línea otros autores (Bartlett *et al.,* 2022) encuentran que aunque las *fintech* pueden reducir la discriminación, no la eliminan, y observan que los prestamistas negros e hispanos pagan un diferencial positivo de interés en préstamos hipotecarios, cuyo importe ascendería según sus estimaciones a 450 millones de dólares anuales que tendrían que sobrepagar; y por último, en otro estudio (Dobbie *et al.,* 2021) se encuentra que modelos de *ML* guiados por objetivos de largo plazo pueden aumentar el beneficio de las entidades financieras y reducir los sesgos, pero guiados por objetivos cortoplacistas penalizan a minorías como ancianos o inmigrantes.

Como conclusión de estos estudios podríamos destacar que la mejor capacidad predictiva de los algoritmos de *ML* proviene sensiblemente por un mejor uso de la información disponible. Sin embargo, una razón importante para evaluar su impacto a la hora de usarlo en la provisión de crédito es que el reparto entre ganadores y perdedores por el uso de esta tecnología puede estar desigualmente distribuido en la sociedad, especialmente a la hora de distinguir entre raza, edad o género de los consumidores o clientes. Para mitigar este riesgo existen nuevas herramientas, como las "explicaciones contrafactuales" (Wachter, Mittelstadt y Russel, 2017) que nos ayudan a averiguar qué factores o variables deberían cambiar para modificar una decisión de crédito tomada por un algoritmo.

3.2. Interpretabilidad de los resultados

3.2.1. Necesidad de interpretabilidad de las decisiones de crédito

¿Por qué queremos explicaciones de las decisiones que nos afectan? Los individuos tienden a buscar explicaciones ante eventos que no entienden, ya que una buena explicación puede facilitar el aprendizaje y crear un sentimiento de confianza (Miller, 2019). En el entorno de decisiones de crédito, la necesidad de las explicaciones cobra una especial importancia,

dado que son decisiones que pueden tener un impacto adverso en la vida de los usuarios. De hecho, como mencionamos anteriormente, la Directiva de Inteligencia Artificial (*AI Act*) incluye como actividades de alto riesgo en la aplicación de IA aquellas relacionadas con el riesgo de crédito, en concreto mencionando: "sistemas de IA [...] en la evaluación de la calidad crediticia o establecimiento de prioridad en el acceso a dichos servicios". De modo similar, en EE. UU. existen regulaciones destinadas a eludir la discriminación en crédito, como son el *Equal Credit Opportunity Act* (*ECOA*) y *Fair Housing Act* (*FHA*). Ambas normativas están centradas en evitar el *disparate treatment* y el *disparate impact,* dos conceptos de equidad crediticia. El primero se enfoca en si los prestamistas han tratado a los solicitantes de crédito de manera diferente según variables protegidas, como la raza o el género. El segundo aborda el uso por parte de los prestamistas de prácticas que puedan tener un efecto negativo desproporcionado en ciertos segmentos de la población. Para comprobar que no se incumplen estos dos conceptos, se exigen pruebas estadísticas y análisis de datos que pueden ser más difíciles de implementar para modelos complejos de *ML* que para modelos estadísticos sencillos. Además, se requiere a los proveedores financieros responder ante una solicitud de los clientes por decisión desfavorable (conocida como *adverse action notice*). No obstante, no existe todavía una obligación legal de "abrir las cajas negras", puesto que no se reconoce el derecho de los usuarios a acceder al código de programación de los algoritmos (Wachter, Mittelstadt y Russel, 2017). Pero sí que existe un derecho a incluir el juicio humano en las decisiones de concesión de crédito, tal y como se establece en el artículo 22 del Reglamento General de Protección de Datos (*GDPR,* por sus siglas en inglés): "Todo interesado tendrá derecho a no ser objeto de una decisión basada únicamente en el tratamiento automatizado, incluida la elaboración de perfiles, que produzca efectos jurídicos en él o le afecte significativamente de modo similar.

Por lo tanto, queda patente que la normativa vigente exige cierto nivel de interpretabilidad en los modelos cuantitativos en riesgo de crédito. Esto se traduce en un riesgo que debe ser gestionado, porque mientras que los modelos econométricos tradicionales son inherentemente explicables, interpretar el resultado de los modelos de *ML* es más complicado. Es por ello que en los últimos años está floreciendo la disciplina de *Explainable AI (xAI)*, que se centra en interpretar los modelos de ML. En este campo, cuando hablamos de interpretar un modelo de ML, nos referimos a indicar qué variables son las más importantes en la salida de un modelo de *ML* para un individuo en particular (interpretabilidad global) y qué variables son las más relevantes para un conjunto de individuos (interpretabilidad global).

3.2.2. Técnicas de interpretabilidad de ML

Explainable AI engloba un conjunto amplio de herramientas (Hoffman *et al.,* 2018). En este artículo destacamos las conocidas como técnicas de interpretabilidad *post hoc,* las cuales se pueden aplicar a cualquier modelo de *ML* una vez entrenado, por ello también se les conoce como agnósticas. Algunas de las más populares son *SHAP – Shapley Additive exPlanations* (Lundberg, 2017), *LIME* (Ribeiro, Singh y Guestrin, 2016)[5] y *permutation Feature*

[5] Recientemente se ha propuesto una modificación más ambiciosa denominada Anchors (Ribeiro, Singh y Guestrin, 2018).

Importance o *FI* (Breiman, 2001). En general, todas ellas se basan en permutar de diferentes maneras las variables de entrada de los modelos para averiguar cómo cambia el resultado o predicción. En las siguientes secciones explicamos en más detalle cómo funcionan cada una de ellas.

Actualmente el uso de estas técnicas de interpretabilidad comienza a extenderse en la investigación académica de concesión de crédito (Tyagi, 2022), y en la propia industria financiera (Blattner *et al.*, 2021). Hoy entendemos además que diferentes usuarios de las decisiones de crédito pueden requerir diferentes tipos de explicaciones (Davis *et al.,* 2022). En concreto, las entidades financieras se beneficiarían de tener capacidad de explicación global (*SHAP, FI*) y local (*SHAP, LIME*) de los resultados, mientras que los clientes obtendrán información más valiosa a partir de ejemplos contrafactuales (Wachter, Mittelstadt y Russel, 2017) que les ayudarán a comprender qué cambios en determinadas variables podrían significar un cambio en la concesión o denegación del crédito. Por otro lado, los reguladores se podrían enfocarse en estresar el resultado de estas técnicas de interpretabilidad, tratando de asegurar su robustez. Por último, los ingenieros de datos deberán tratar de valorar la incertidumbre alrededor de sus estimaciones (*conformal ML*) y la posibilidad de explicar sus resultados con modelos más fácilmente comprensibles por otros usuarios (*surrogate models*).

Cabe recordar que, en última instancia, el uso de *ML* en finanzas suscita cuestiones éticas que deben ser examinadas con cautela, tales como integridad y justicia. Si bien tanto humanos como máquinas pueden errar en sus decisiones de concesión de créditos, lo importante es asegurar que se está usando el mejor juicio posible desde un punto de vista ético (Rizinski *et al.,* 2022), para lo cual no debemos nunca olvidar la necesidad de colaboración entre humanos y máquinas para lograr un proceso de decisión justo y eficiente.

Permutation Feature Importance

Permutation Feature Importance (también referida en este artículo como *feature importance,* o *FI*) es una técnica de evaluación *post hoc* que mide la importancia de una variable en la predicción del modelo en función del aumento del error en la predicción cuando cambiamos el valor de dicha variable. Se introdujo inicialmente para el modelo *random forest* (Breiman, 2001) aunque ya existen versiones agnósticas del mismo (Fisher *et al.,* 2019). Siguiendo con el ejemplo de datos de *Give me some credit,* imaginemos que hemos entrenado un modelo *XGBoost* para predecir impagos, y que estamos interesados en la importancia de la variable *Revolving.* En primer lugar, medimos el error en predicción obtenido con el modelo *XGBoost.* Hay diferentes medidas de error, pero dado que estamos en un problema de clasificación donde la variable a predecir es binaria, proponemos usar como métrica (1 – AUC-ROC), de forma que cuanto más alto es el *AUC-ROC,* más bajo es el error. A continuación, cambiamos el valor de *Revolving* de forma aleatoria, y estimamos de nuevo el error en la predicción. Repetimos el proceso *n* veces, para asegurarnos que los resultados no dependen de un cambio de valor en concreto. Después de *n* iteraciones, comparamos el error medio de esas repeticiones con el error del modelo original. Si la variable *Revolving* es importante, entonces el error medio obtenido debería ser mucho mayor que con el modelo original. Si la variable no es importante, entonces ambos errores deberían ser parecidos.

El método es simple, pero tiene un gran coste computacional. Implica repetir el proceso *n* veces para cada variable, por lo que cuanto mayor sea el número de variables, más tiempo llevará calcular sus importancias. Además, tiene dos inconvenientes adicionales. En primer lugar, se asume la independencia de variables entre ellas, algo que puede resultar poco realista en las decisiones de crédito. Por ejemplo, al cambiar de forma aleatoria el valor de las variables *Revolving,* el valor del resto de variables no cambia, lo que podría generar situaciones poco realistas. Por ejemplo, si suponemos que los préstamos *Revolving* se asocian a vencimientos cortos, nosotros no estaríamos restringiendo esta interacción entre ambas variables, permitiendo que existieran observaciones en nuestra simulación con vencimiento alto y estructura *Revolving,* en contra de lo que observamos en la realidad. Además, esta técnica no puede indicarnos la dirección del efecto de una determinada variable. Podemos saber si *Revolving* es importante, pero no si los aumentos o disminuciones en el valor de la variable están relacionados con aumentos o disminuciones en la probabilidad de impago.

SHAP

SHAP es una técnica que se basa en medir la contribución de una variable a la predicción de la probabilidad de impago, para un prestatario en particular, en comparación con la predicción promedio. Estas contribuciones se denominan "valores de Shapley" y nos darían información sobre la interpretabilidad local de nuestro modelo. Una vez que tenemos los valores de Shapley para cada variable y para cada prestatario, estos se pueden sumar para obtener la importancia global de la variable (*SHAP*). El proceso por el cual se calculan los valores de Shapley puede explicarse desde la perspectiva de la teoría de juegos. El juego sería reproducir el resultado del modelo (en nuestro caso, la probabilidad de impago). Los jugadores serían todas las posibles coaliciones de variables explicativas. Finalmente, la recompensa sería la contribución de cada coalición al resultado final del modelo. Veamos cómo funciona con un ejemplo. Supongamos que disponemos de una base de datos parecida a la de *Give me some credit,* pero donde únicamente tenemos tres variables explicativas: *Revolving*, "Edad" e "Ingresos", además de la variable binaria a predecir, "Impago". Imaginemos que estamos interesados en calcular la importancia de la variable *Revolving* en la probabilidad de impago de un individuo dado. Primero consideremos las cuatro posibles coaliciones de variables sin *Revolving*: Conjunto vacío, Edad, Ingresos, Edad e Ingresos.

Para las cuatro coaliciones, calculamos la probabilidad de impago del prestatario *i* con y sin la variable *Revolving*. El valor de Shapley de *Revolving* en relación con la probabilidad de impago para el prestatario *i* es el promedio ponderado de esas contribuciones marginales. Para obtener la importancia global de *Revolving* para la predicción de impago en toda la muestra, repetimos el proceso para todos los prestatarios de la base de datos, y calculamos el promedio de los valores de Shapley. Las variables con valores Shapley absolutos grandes se consideran variables locales importantes (y, en consecuencia, igualmente con valores agregados *SHAP*, para una explicación global). La técnica tiene dos problemas que también afectan a *permutation FI*. *SHAP* hace la suposición de que las variables no están correlacionadas, por lo que se generan coaliciones poco realistas. Y computacionalmente puede ser muy costoso también. Una ventaja con respecto a *permutation FI* es que aporta interpretabilidad local además

de la global, y además puede indicar la dirección del efecto de la variable en la probabilidad de impago. O sea, nos dice si la probabilidad aumenta o disminuye cuando el valor de la variable explicativa de interés es alto o bajo.

LIME

LIME o *Local Interpretable Model-Agnostic Explanations* es un modelo *post hoc* de interpretabilidad local y agnóstico. Como su nombre indica, *LIME* explica el resultado de modelos de *ML* aproximando el modelo subyacente por uno interpretable. Los modelos interpretables pueden ser modelos lineales con o sin regularización (por ejemplo, penalizando el tamaño de los coeficiente a través de técnicas como *Lasso* o *Ridge*), o árboles de decisión sencillos. Estos modelos interpretables se entrenan en pequeñas perturbaciones de los datos de entrada originales para predecir la predicción del modelo original subyacente. Imaginemos que hemos usado redes neuronales para predecir la probabilidad de impago de un préstamo concreto de nuestra base de datos. Si queremos usar *LIME* para explicar la predicción de ese préstamo, lo primero que deberíamos hacer es perturbar las variables explicativas de este préstamo agregando ruido a las variables continuas o eliminando algunas de ellas, y obtener la nueva predicción con nuestro modelo usando estos datos generados. Repetimos este proceso varias veces, otorgando un peso a estas nuevas muestras según su proximidad a los datos originales. Finalmente, usamos un modelo interpretable en estos nuevos datos, y explicamos las predicciones interpretando el modelo local. Una de sus ventajas es la relatividad facilidad para llevar a cabo el proceso, pero *LIME* también tiene sus puntos débiles. El mayor problema es la creación de las perturbaciones de datos, sobre todo para variables categóricas. No hay un consenso sobre cómo generar estos datos, y se pueden crear datos poco realistas. A su vez, las explicaciones pueden ser poco estables al depender de la perturbación de los datos y del modelo interpretable que se elige. Otro problema es que, a diferencia de los modelos anteriormente mencionados, no aporta interpretabilidad global.

3.2.3. El problema de la discrepancia en las interpretaciones

Como vemos, existe una amplia variedad de técnicas para explicar las razones por las que un modelo de *ML* ha tomado una decisión, lo cual a su vez puede crear un problema si existe discrepancia entre ellas, Krishna *et al.* (2022). Es decir, según la técnica utilizada, la conclusión sobre qué variables han sido más relevantes para un modelo puede variar. En Krishna *et al.* (2022) se establecen una serie de entrevistas con ingenieros de datos para entender cuándo existe una discrepancia entre explicaciones. Siguiendo este artículo, hay tres dimensiones a partir de las cuales podemos medir esta discrepancia. En primer lugar, podemos medir la discrepancia entre explicaciones en función del distinto *ranking* de variables importantes que ofrecen las diferentes técnicas. En segundo lugar, la discrepancia se puede medir en función de la magnitud de importancia asignada por las explicaciones a las diferentes variables (las técnicas como *SHAP* y *permutation FI* no solo indican el orden de importancia, también indican su magnitud). Por último, se puede medir la discrepancia en función de las explicaciones en diferentes momentos del tiempo. La preocupante realidad es que actualmente no existe un

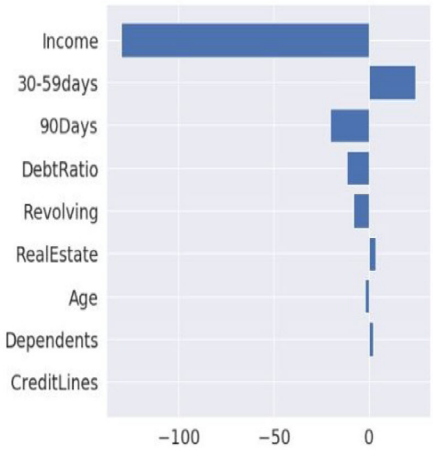
consenso sobre qué es una discrepancia ni sobre cómo resolverla. Para ilustrar el problema de la discrepancia de explicaciones proponemos de nuevo utilizar la base de datos *Give me some credit*. Además, añadiremos una dimensión a esta discrepancia, la cual descompondremos en dos componentes. Primero, la discrepancia entre modelos, bajo la cual un mismo individuo que se enfrente a una decisión de concesión de crédito otorgada por un modelo basado en árboles o basado en redes neuronales, puede diferir. Para ilustrarlo, intentaremos explicar los resultados de los cinco modelos usados en el apartado 2.1 (*Logit, Logit* al cuadrado, árbol de decisión, *XGBoost,* y redes neuronales profundas), y veremos que en función del modelo de *ML* utilizado, encontraremos explicaciones muy diferentes. En segundo lugar, analizaremos la discrepancia dentro del mismo modelo, es decir, usando distintas técnicas de interpretabilidad sobre el mismo modelo de ML veremos cómo obtenemos resultados discrepantes también.

Discrepancia entre modelos

Partimos del modelo *Logit*, que es el más sencillo de interpretar, aunque su rendimiento en clasificación era el más bajo como vimos en el apartado 2.1. Dado que hemos estandarizado las magnitudes de las variables, podemos comparar directamente los coeficientes estimados de las variables para entender qué variables son las más importantes. Es por ello que *Logit* es un modelo inherentemente interpretable, sin necesidad de aplicar ninguna técnica de interpretabilidad *post hoc*. En la figura 2 podemos ver las magnitudes de los coeficientes del *Logit* estimado, donde se ve claramente que la variable "Ingresos" es la más importante, con signo negativo, lo cual es de esperar, pues mayor nivel de "Ingresos" se asocia a menor probabilidad de impago.

Figura 2.

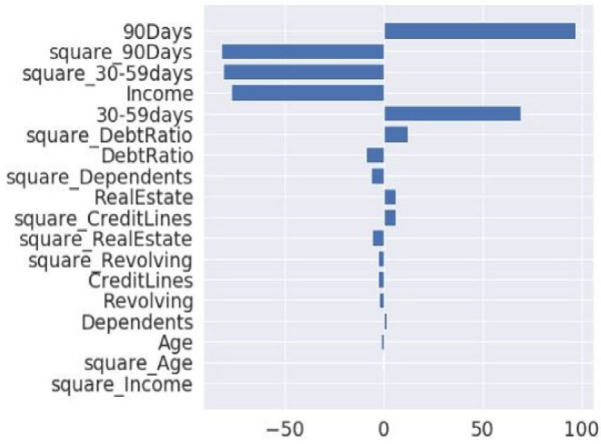
***Logit:* variables más importantes según coeficiente**



En definitiva, el *Logit* es fácilmente interpretable pero su predicción es muy baja. Añadiendo variables al cuadrado podemos mejorar su poder predictivo como vimos anteriormente. Sin embargo, aunque todavía sería inherentemente interpretable, no es tan sencillo explicar qué variables son las más importante puesto que hay que tener en cuenta los términos al cuadrado. En la figura 3 mostramos la estimación de los coeficientes de *Logit* al cuadrado, y vemos que esta vez la variable *90Days* es la más importante, con signo positivo. Sin embargo, la segunda variable más importante es la variable *90Days* al cuadrado, con signo opuesto. Esto sugiere que cuantas más veces se retrase 90 días en el reembolso de sus obligaciones, más alta la probabilidad de impago del prestatario. Pero a partir de un número de días, mayor número es el retraso más disminuye la probabilidad de impago. Esto pone de manifiesto las limitaciones del *Logit:* a pesar de que incluir términos polinómicos para incluir no linealidades puede ayudar en la predicción, esto elimina la facilidad de interpretar el resultado del modelo.

Figura 3.

***Logit²:* variables más importantes según coeficiente**

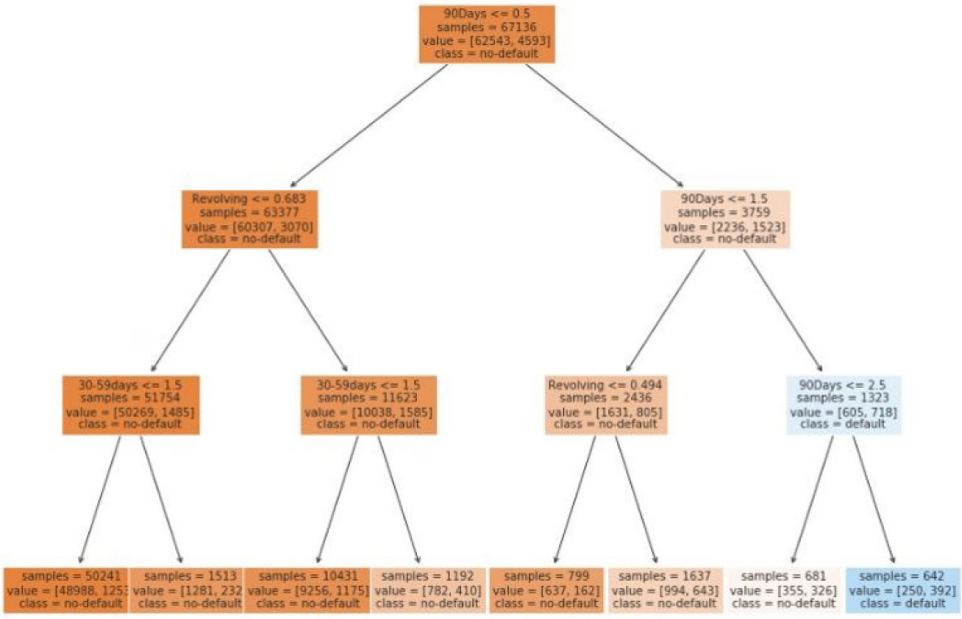


A continuación, pasamos a explicar las variables más importantes para el árbol de decisión. Al tratarse de un árbol sencillo, de solo tres ramas, podemos considerarlo un modelo inherentemente interpretable. En la figura 4 mostramos el resultado del árbol que estimamos en el apartado 2.1. La forma para interpretar el árbol es la siguiente: en cada ramificación, se muestra la variable que el modelo ha usado para ramificar la muestra en dos, de forma que al final se creen grupos lo más homogéneos posibles (los impagos con los impagos, y los no impagos con los no impagos). Por ejemplo, siguiendo la figura 4, la variable más importante es *90Days,* que es la que el modelo usa para ramificar la muestra en dos al comienzo del árbol, dejando a la izquierda aquellas observaciones con *90Days* igual a cero, y a la derecha las observaciones con *90Days* mayor a cero. En las siguientes ramificaciones se sigue usando principalmente la variable *90Days*, manteniendo a la derecha en cada ramificación primero las que tienen *90Days* mayor a uno, y finalmente las que tienen *90Days* mayor a dos. Finalmente encontramos una ramificación donde hay más impagos (392) que no impago (250). El

hecho de que el árbol considere *90Days* como la variable más relevante encaja con la explicación de *Logit* al cuadrado. Sin embargo, como se aprecia en la figura 4, la variable *Revolving* también se usa en muchas ramificaciones, y sin embargo no parece una de las más relevantes en *Logit* al cuadrado. Por lo tanto, aquí tenemos ya dos modelos inherentemente interpretables de *ML, Logit* al cuadrado y árbol, con un rendimiento de predicción parecido, pero con explicaciones diferentes.

Figura 4.

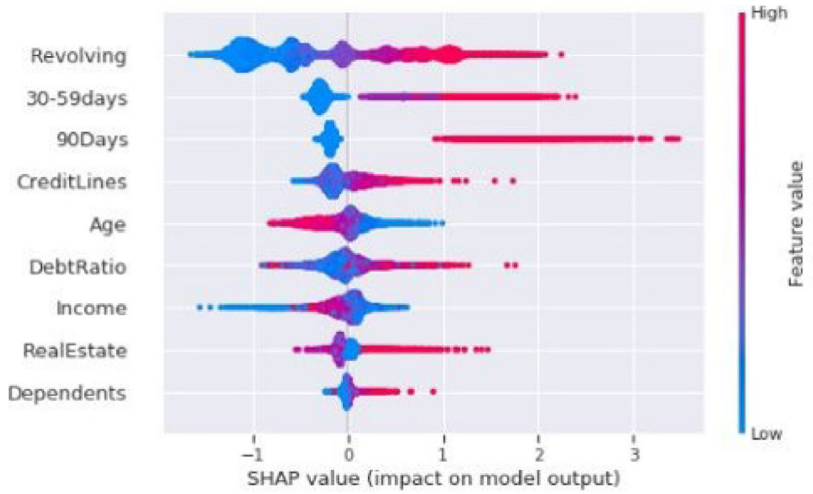
Árbol de decisión: variables más importantes



Por último, vamos a interpretar las variables más importantes para los modelos *XGBoost* y redes neuronales profundas. A diferencia de los tres modelos inherentemente interpretables vistos antes, para interpretar *XGBoost* y la red neuronal profunda es necesario usar una técnica de XAI. En este caso, vamos a usar *SHAP,* cuyo funcionamiento explicamos en la sección anterior. Mostramos en las figuras 5 y 6 el resultado de *SHAP* para *XGBoost* y redes neuronales profundas, respectivamente. La imagen de *SHAP* corresponde con la salida de la aplicación de *Python tree.shap* para *XGBoost* y de *deep.shap* para redes neuronales profundas. Las variables aparecen ordenadas de mayor (arriba) a menor (abajo) importancia. En el eje horizontal se indica el impacto de la variable explicativa sobre nuestro objetivo a predecir. Los puntos azules corresponden con valores bajos de la variable, y los puntos rojos corresponden con valores altos. Por ejemplo, en 5, la variable más relevante es *Revolving,* y para valores bajos de *Revolving* (puntos azules), menor es la probabilidad de impago, mientras que para

Figura 5.

XGBoost*: variables más importantes segun *SHAP



valores altos de *Revolving* (puntos rojos) mayor es la probabilidad de impago. En el caso de "Edad" (la quinta variable más importante), valores altos de la misma reducen la probabilidad de impago, y valores bajos de la variable aumentan la probabilidad de impago. Como queda patente viendo las figuras 5 y 6, la interpretación de *XGBoost* y de la red neuronal profunda

Figura 6.

Red neuronal profunda: variables más importantes según *SHAP*

a través de *SHAP* es completamente diferente. Por ejemplo, la variable más relevante para *XGBoost* es la menos importante para la red neuronal. Y la tercera variable más importante para la red neuronal es la sexta para *XGBoost*. Esto es así a pesar de haber usado la misma técnica de interpretabilidad *post hoc*.

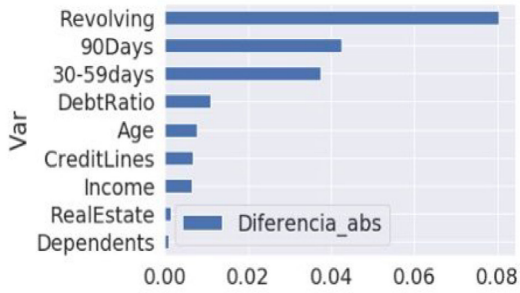
Como conclusión, esta evidencia muestra cómo las explicaciones de los modelos interpretables (*Logit* y árbol de decisión) no coinciden con las de los modelos no interpretables.

Discrepancias dentro del mismo modelo

Además de las fuertes discrepancias entre diferentes modelos de *ML*, en esta sección mostraremos que también existen discrepancias dentro de un mismo modelo de *ML*. Vamos a comparar la explicación de *XGBoost* con *SHAP 5*, con la explicación que nos brinda otra técnica de interpretabilidad global, *permutation FI* (ver figura 7). En este caso, ambas técnicas identifican como variable más relevante *Revolving,* pero existen diferencias en el resto del *ranking,* particularmente en cuanto a *DebRatio,* "Edad", "Ingresos", etc. Mostrando que, por lo tanto, también puede haber discrepancias dentro del mismo modelo.

Figura 7.

XGBoost:* variables más importantes según *permutation FI



Hasta ahora hemos visto técnicas de interpretabilidad global. Pero, de modo similar, podemos observar discrepancias también si tratamos de explicar las predicciones préstamo a préstamo, o sea interpretabilidad local. Por ejemplo, para un mismo préstamo de la muestra, podemos explicar qué variables son las más importantes con valores de Shapley (ver figura 8) o con *LIME* (ver figura 9). Este prestamista tiene 65 años, una renta de 6.000 dólares, 19 "líneas de crédito", valor de *Revolving* igual a 0,67, ha pagado entre 30 y 59 días tarde 3 veces, y ha pagado 90 días tarde 8 veces. Ambas técnicas coinciden con que *90Days* es la variable más relevante en la probabilidad de impago de este prestamista (cuanto mayor *90Days,* mayor probabilidad de impago), y que mayor "Edad" reduce la probabilidad de impago. Sin embargo, la importancia otorgada a *Revolving* difiere mucho en ambas explicaciones. Mientras que para *LIME* es la segunda más importante, para *SHAP* es prácticamente la menos

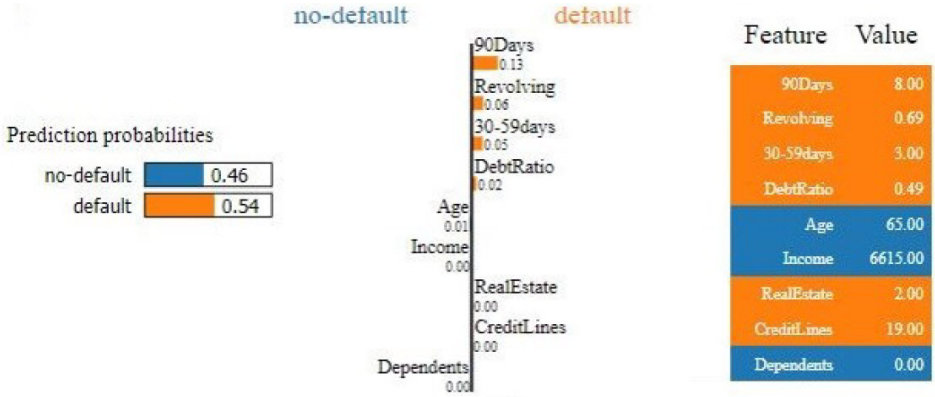
importante. De nuevo, un ejemplo de discrepancia entre *rankings,* esta vez buscando interpretabilidad local.

Figura 8.

***XGBoost:* variables más importantes según *SHAP*, a nivel local**

Figura 9.

***XGBoost:* variables más importantes según *LIME*, a nivel local**



A pesar de que pueden ser una herramienta fundamental para aplicar *ML* en decisiones de crédito, existen todavía muchas dudas sobre las técnicas de interpretabilidad *post hoc* ¿Cómo de preocupantes son las discrepancias entre técnicas? ¿Cómo de fiables y robustas son? ¿Son suficientes como explicación? (Miller, 2019). Motivados por estas preguntas, existe una literatura incipiente que trata de usar datos sintéticos para responder a algunas de estas preguntas. Generar datos sintéticos permite controlar y conocer la verdadera naturaleza de la relación entre variables, pudiendo comprobar hasta qué punto las explicaciones de las técnicas de intepretabilidad coinciden con las relaciones verdaderas entre los datos. Si bien, no existe un procedimiento estandarizado sobre cómo crear estos conjuntos de datos sintéticos. Por ejemplo hay estudios que (Barr *et al.,* 2020) usan cópulas gaussianas y demuestran que la correlación de variables redundantes puede afectar a las explicaciones dadas por *SHAP*. En esta línea, otros estudios (Aas, Jullum y Løland, 2021) modifican la implementación de *SHAP* para poder aproximar mejor las explicaciones cuando las variables explicativas tienen cierto nivel de dependencia. También usando datos sintéticos, estudios como Hall *et al.* (2021) y

Zhang *et al.* (2019) demuestran que existen varias fuentes de incertidumbre e inestabilidad para *LIME*. Por último mencionar que en nuestro trabajo (Alonso-Robisco & Carbo, 2022d) proponemos la generación de datos sintéticos para crear un entorno controlado donde realizar un *stress test* de estas técnicas. Nuestros datos generados representan escenarios similares a las condiciones normalmente observadas en entornos de crédito, y nuestros resultados sugieren que *SHAP* es mejor que *permutation FI* capturando características relevantes, aunque las explicaciones pueden variar significativamente según las características del conjunto de datos y el modelo *ML* utilizado, por lo tanto, se requiere cautela para generalizar unas conclusiones sobre la precisión de estas técnicas.

4. CONCLUSIONES

Los modelos de *ML* pueden revolucionar el campo de la concesión de crédito, creando sistemas automáticos de decisión que mejoren tanto la experiencia del usuario (por ejemplo, la velocidad de concesión del crédito), sin empeorar la exposición de riesgo de los proveedores de crédito (Gambacorta *et al.,* 2019). Estas mejoras se deben tanto a la capacidad de esta tecnología financiera para gestionar grandes cantidades de datos (*big data*), como a la mayor flexibilidad de estos modelos predictivos para captar relaciones no lineales entre las variables (Fuster *et al.,* 2022; Gimeno y Sevillano, 2022; Hoffman *et al.,* 2018). De hecho, como hemos visto, su mayor rendimiento estadístico se puede llegar a traducir en clasificaciones más granulares del activo ponderado por riesgo, derivando en un menor consumo potencial de capital regulatorio (Alonso-Robisco & Carbo, 2022d); Fraisse y Laporte, 2022) e inclusive, a gran escala, en mayor inclusión financiera (Berg *et al.,* 2020; Buchak, 2018; Barruetabeña, 2020).

No obstante, el uso de *ML* trae consigo una serie de consecuencias no intencionadas, en la forma de nuevos factores de riesgo (Alonso-Robisco & Carbo, 2022a). Son de especial relevancia para los supervisores financieros la discriminación y la interpretabilidad de los resultados (BAFIN, 2022; Blattner, Nelson y Spiess, 2021; Dupont, Fliche y Yang, 2020). En cuanto a la discriminación, el uso de *ML* puede dificultar el cumplimiento de determinadas políticas de equidad en la concesión de crédito (Philippon, 2019). Esta preocupación se sostiene en numerosos estudios que demuestran que los sistemas de concesión de crédito basados en *ML* perjudican especialmente a minorías sociales o a la población de elevada edad (Bartlett *et al.,* 2022; Dobbie *et al.,* 2021). Por ello, nuevos actos legislativos dedicados a la regulación de los sistemas basados en inteligencia artificial, como por ejemplo, la Directiva Europea de Inteligencia Artificial (*AI Act*) clasifican el uso de esta tecnología para la concesión de crédito como un área de alto riesgo, debido al potencial impacto negativo en la solvencia de las empresas y el bienestar de las personas que se enfrentan a este tipo de decisiones automatizadas. En cuanto a la interpretabilidad, la reciente normativa sobre privacidad de los datos (*GDPR,* por sus siglas en inglés), establece la necesidad de transparencia e inclusión del juicio humano en decisiones donde se hace un perfilado del riesgo de los clientes. Esto puede ser un problema en la aplicación de *ML* en crédito debido a la dificultad para explicar las decisiones de estos modelos. Existen herramientas novedosas que podrían ayudar a mitigar estos

problemas de explicabilidad, conformando un campo conocido como *Explainable AI* (*xAI*). Dentro de este campo, pueden ser de especial interés para el uso en sistemas de evaluación del riesgo crediticio las conocidas como técnicas de interpretabilidad *post hoc,* las cuales pueden ser utilizadas para explicar los resultados de cualquier modelo de *ML* previamente entrenado.

En este capítulo, para ilustrar tanto las oportunidades derivadas de la mayor capacidad predictiva de estos modelos, como los nuevos riesgos asociados a la explicabilidad de sus resultados, realizamos un ejercicio empírico con una base de datos de libre acceso denominada *Give Me Some Credit.* Estimamos cinco modelos de *ML* comúnmente utilizados para la predicción de impagos, como son *Logit,* árboles de decisión, r*andom forest, XGBoost* y redes neuronales profundas. Tras evidenciar su mejor rendimiento estadístico tratamos de interpretar sus resultados a través de tres técnicas: *SHAP, permutation FI* y *LIME,* poniendo de relieve la existencia de un problema por la discrepancia de explicaciones (Krishna *et al.,* 2022), tanto dentro de cada modelo, comparando distintas técnicas de interpretabilidad; como entre modelos, usando la misma técnica, pero aplicada a distintos modelos predictivos.

Sin duda, para lograr una implementación de esta tecnología a mayor escala en el sistema financiero será necesario lograr robustas herramientas que logren satisfacer las necesidades de explicación de todos los agentes que intervienen en un proceso de concesión de crédito: entidades financieras, consumidores, reguladores y científicos de datos (Davis *et al.,* 2022). La solución a los problemas derivados del uso de sistemas de *ML* posiblemente no radique en usar exclusivamente más *ML* (*xAI*), por lo que debemos prestar especial atención al uso responsable de esta tecnología, siguiendo unos principios éticos (Rizinski *et al.,* 2022), sin olvidarnos de que esta tecnología no es sustitutiva de la actual econometría, sino complementaria (Alonso-Robisco & Carbo, 2022c); Kaji *et al.,* 2020). Por ello, concluimos con un llamamiento hacia la colaboración interdisciplinar, entre economistas e ingenieros, haciendo especial énfasis en la formación en técnicas de *ML* a los profesionales del sector financiero, gestores de riesgos, y economistas dedicados a la investigación académica (Athey e Imbens, 2019).

Referencias

Aas, K., Jullum, M. y Løland, A. (2021). Explaining individual predictions when features are dependent: More accurate approximations to Shapley values. *Artificial Intelligence,* 298, 103502.

Agarwal, S., Alok, S., Ghosh, P. y Gupta, S. (2020). Financial inclusion and alternate credit scoring for the millennials: Role of big data and machine learning in fintech. *Business School, National University of Singapore Working Paper,* SSRN, 3507827.

Albanesi, S. y Vamossy, D. F. (2019). Predicting consumer default: A deep learning approach (No. w26165). National Bureau of Economic Research.

Alonso-Robisco, A. y Carbo Martinez, J. M. (2022a). Measuring the model risk-adjusted performance of machine learning algorithms in credit default prediction. *Financial Innovation, 8*(1), pp. 1-35.

Alonso-Robisco, A. y Carbo Martinez, J. M. (2022b). Can machine learning models save capital for banks? Evidence from a Spanish credit portfolio. *International Review of Financial Analysis,* 84, 102372.

Alonso-Robisco, A. y Carbó, J. M. (2022c). Inteligencia Artificial y Finanzas: Una Alianza Estratégica (Artificial Intelligence and Finance: A Strategic Alliance). *Banco de España Occasional Paper,* 2222.

Alonso-Robisco, A. y Carbó, J. M. (2022d). "Accuracy of explanations of machine learning models for credit decisions," *Working Papers 2222*, Banco de España.

Athey, S. e Imbens, G. W. (2019). Machine learning methods that economists should know about. *Annual Review of Economics,* 11, pp 685-725.

Bafin (2022). Machine learning in risk models – characteristics and supervisory priorities. responses to the consultation paper. Rep. Germany: Federal Financial Supervisory Authority.

Bartlett, R., Morse, A., Stanton, R. y Wallace, N. (2022). Consumer-lending discrimination in the FinTech era. *Journal of Financial Economics, 143*(1), pp. 30-56.

Barr, B., Xu, K., Silva, C., Bertini, E., Reilly, R., Bruss, C. B. y Wittenbach, J. D. (2020). Towards ground truth explainability on tabular data. *arXiv preprint arXiv*:2007.10532.

Barruetabeña Lorente, E. (2020). La influencia de las nuevas tecnologías en la inclusión financiera. *Boletín económico/Banco de España [Artículos],* n. 1.

Bazarbash, M. (2019). Fintech in financial inclusion: machine learning applications in assessing credit risk. International Monetary Fund.

Berg, T., Burg, V., Gombović, A. y Puri, M. (2020). On the rise of fintechs: Credit scoring using digital footprints. *The Review of Financial Studies, 33*(7), pp. 2845-2897.

Blattner, L., Nelson, S. y Spiess, J. (2021). Unpacking the black box: Regulating algorithmic decisions. *arXiv preprint arXiv*:2110.03443.

Blattner, L., Stark, PR., Spiess, J., McElfresh, D., Yazdi, S. y Kalashnov, G. (2021). Machine Learning Explainability & Fairness: Insights from Consumer Lending. *finreglab.org.*

Breiman, L. (2001). Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical science, 16*(3), pp. 199-231.

Breiman, L. (2001). Random forests. *Machine learning,* 1, pp. 5-32

Buchak, G., Matvos, G., Piskorski, T. y Seru, A. (2018). Fintech, regulatory arbitrage, and the rise of shadow banks. *Journal of Financial Economics 130*(3), pp. 453– 483.

Butaru, F., Chen, Q., Clark, B., Das, S., Lo, A. W. y Siddique, A. (2016). Risk and risk management in the credit card industry. *Journal of Banking & Finance,* 72, pp. 218-239.

Davis, R., Lo, A. W., Mishra, S., Nourian, A., Singh, M., Wu, N. y Zhang, R. (2022). *Explainable Machine Learning Models of Consumer Credit Risk.* Available at SSRN.

Demirguc-Kunt, A., Klapper, L., Singer, D. y Ansar, S. (2018). *The Global Findex Database 2017: Measuring financial inclusion and the fintech revolution.* World Bank Publications.

Dobbie, W., Liberman, A., Paravisini, D. y Pathania, V. (2021). Measuring bias in consumer lending. *The Review of Economic Studies, 88*(6), pp. 2799-2832.

Dupont, L., Fliche, O. y Yang, S. (2020). *Governance of Artificial Intelligence in Finance.* Banque De France.

EBA. (2021). Discussion paper on machine learning for IRB models. EBA/DP/2021/04. November 2021.

Estévez Almenzar, M., Fernández Llorca, D., Gómez, E. y Martinez Plumed, F. (2022). Glossary of human-centric artificial intelligence (No. JRC129614). Joint Research Centre (Seville site).

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern recognition letters, 27*(8), pp. 861-874.

Fernández, A. (2019). Inteligencia artificial en los servicios financieros. *Boletín Económico* 2/2019. Artículos Analíticos. Banco de España.

Fisher, A., Rudin, C. y Dominici, F. (2019). All Models are Wrong, but Many are Useful: Learning a Variable's Importance by Studying an Entire Class of Prediction Models Simultaneously. *J. Mach. Learn. Res., 20*, 177, pp-1–81.

Fraisse, H. y Laporte, M. (2022). Return on investment on artificial intelligence: The case of bank capital requirement. *Journal of Banking and Finance,* 138, 106401.

Floridi, L., Holweg, M., Taddeo, M., Amaya Silva, J., Mökander, J. y Wen, Y. (2022). capAI-A Procedure for Conducting Conformity Assessment of AI Systems in Line with the EU Artificial Intelligence Act. Available at SSRN 4064091.

Fuster, A., Goldsmith-Pinkham, P., Ramadorai, T. y Walther, A. (2022). Predictably unequal? The effects of machine learning on credit markets. *The Journal of Finance, 77*(1), pp. 5-47.

Gambacorta, L., Huang, Y., Qiu, H. y Wang, J. (2019). *How do machine learning and non-traditional data affect credit scoring? New evidence from a Chinese fintech firm.*

Gimeno, R. y Sevillano, J. M. M. (2022). Tradición e inteligencia artificial: oportunidades y retos del machine learning para los servicios financieros. Información Comercial Española, ICE: *Revista de economía,* (926), pp. 109-118.

Goodell, J. W., Kumar, S., Lim, W. M. y Pattnaik, D. (2021). Artificial intelligence and machine learning in finance: Identifying foundations, themes, and research clusters from bibliometric analysis. *Journal of Behavioraland Experimental Finance,* 32,100577.

Hall, P., Cox, B., Dickerson, S., Kannan, R., Arjun, Kulkarni, R. y Schmidt, N. (2021). A United States fair lending perspective on machine learning. *Frontiers in Artificial Intelligence,* 4.

Hoffman, R. R., Mueller, S. T., Klein, G. y Litman, J. (2018). Metrics for explainable AI: Challenges and prospects. *arXiv preprint arXiv*:1812.04608.

Huang, Y., Zhang, L., Li, Z., Qiu, H., Sun, T. y Wang, X. (2020). *Fintech credit risk assessment for SMEs: Evidence from China.*

Institute of International Finance. (2018). *Explainability in predictive modelling.*

Institute of International Finance. (2019). *Machine learning in credit risk.*

Institute of International Finance. (2019). *Bias and Ethical Implications in Machine Learning.*

Jones, S., Johnstone, D. y Wilson, R. (2015). An empirical evaluation of the performance of binary classifiers in the prediction of credit ratings changes. *Journal of Banking and Finance,* 56, pp. 72-85.

Jung, C., Mueller, H., Pedemonte, S., Plances, S. y Thew, O. (2019). *Machine learning in UK financial services.* Bank of England and Financial Conduct Authority.

Kaji, T., Manresa, E. y Pouliot, G. (2020). An adversarial approach to structural estimation. *arXiv preprint arXiv*:2007.06169.

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

Königstorfer, F. y Thalmann, S. (2020). Applications of Artificial Intelligence in commercial banks–A research agenda for behavioral finance. *Journal of behavioral and experimental finance,* 27, 100352.

Krishna, S., Han, T., Gu, A., Pombra, J. Jabbari, S., Wu, S. y Lakkaraju, H. (2022). The Disagreement Problem in Explainable Machine Learning: A Practitioner's Perspective. *arXiv preprint arXiv*:2202.01602.

Kvamme, H., Sellereite, N., Aas, K. y Sjursen, S. (2018). Predicting mortgage default using convolutional neural networks. *Expert Systems with Applications,* 102, pp. 207-217.

Lundberg, S. y Lee, S. (2017). A unified approach to interpreting model predictions. *Advances in neural information processing systems,* 30.

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence,* 267, pp. 1-38.

Moscatelli, M., Parlapiano, F., Narizzano, S. y Viggiano, G. (2020). Corporate default forecasting with machine learning. *Expert Systems with Applications,* 161, 113567.

Petropoulos, A., Siakoulis, V., Stavroulakis, E. y Klamargias, A. (2019). A robust machine learning approach for credit risk analysis of large loan level datasets using deep learning and extreme gradient boosting. *IFC Bulletins chapters,* 49.

Philippon, T. (2019). On fintech and financial inclusion (No. w26330). National Bureau of Economic Research.

Petralia, K., Philippon, T., Rice, T. y Véron, N. (2019). Banking disrupted? financial intermediation in an era of transformational technology. Technical Report 22, Geneva Reports on the World Economy, ICMB and CEPR.

Qi, Y. y Xiao, J. (2018). Fintech: AI powers financial services to improve people's lives. *Communications of the ACM, 61*(11), pp. 65-69.

Ribeiro, M., Singh, S. y Guestrin, C. (2016). Model-agnostic interpretability of machine learning. *arXiv preprint arXiv*:1606.05386.

Ribeiro, M. T., Singh, S. y Guestrin, C. (2018, April). Anchors: High-precision model-agnostic explanations. In *Proceedings of the AAAI conference on artificial intelligence* (*Vol. 32,* No. 1).

Rizinski, M., Peshov, H., Mishev, K. y Chitkushev, L. T., Vodenska, I. y Trajanov, D. (2022). Ethically Responsible Machine Learning in Fintech. *IEEE Access,* 10, pp-97531–97554

Sarma, M. (2008). Index of financial inclusion (No. 215). *Working paper.*

Sigrist, F. y Hirnschall, C. (2019). Grabit: Gradient tree-boosted Tobit models for default prediction. *Journal of Banking and Finance,* 102, pp. 177-192.

Sirignano, J. y Cont, R. (2019). Universal features of price formation in financial markets: perspectives from deep learning. *Quantitative Finance, 19*(9), pp. 1449-1459.

Tyagi, S. (2022). Analyzing Machine Learning Models for Credit Scoring with Explainable AI and Optimizing Investment Decisions. *arXiv preprint arXiv*:2209.09362.

Wachter, S. Mittelstadt, B. y Russel, C. (2017). Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harv. JL & Tech.,* 31, pp. 841-.

Zhang, Y., Song, K., Sun, Y., Tan, S. y Udell, M. (2019). "Why Should You Trust My Explanation?" Understanding Uncertainty in LIME Explanations. *arXiv preprint arXiv*:1904.12991.