COMILLAS
UNIVERSIDAD PONTIFICIA

Facultad de Ciencias Económicas y Empresariales

Análisis de interpretabilidad de modelos de Machine Learning aplicado a la gestión de riesgos financieros

Autor: María Sanmartín González
Director: Jenny Alexandra Cifuentes Quintero

MADRID| Junio 2023

Resumen

En el campo del Machine Learning (ML), ha habido una notable evolución en la complejidad de los modelos, lo que ha mejorado su capacidad predictiva. Sin embargo, esta alta complejidad ha dado lugar a un desafío en términos de interpretación, ya que los modelos se vuelven menos comprensibles para los usuarios. En este contexto, el presente trabajo se centra en abordar el problema de la interpretabilidad local en modelos de ML, específicamente en el contexto del riesgo financiero. En el ámbito del riesgo financiero, la interpretabilidad de los modelos de ML es de vital importancia para promover la transparencia en las decisiones tomadas por dichos modelos.La capacidad de comprender y explicar las predicciones o clasificaciones es esencial para garantizar decisiones éticas y cumplir con los crecientes requisitos regulatorios en el sector financiero. Por lo tanto, el objetivo de este trabajo es identificar y aplicar las técnicas de interpretabilidad más relevantes en el ámbito del riesgo de crédito.

El enfoque se centrará en técnicas de interpretabilidad local, que permiten analizar y comprender las decisiones de los modelos a nivel de cada observación individual. Al aplicar estas técnicas a un caso concreto de riesgo de crédito, se buscará obtener una visión detallada de cómo el modelo clasifica el riesgo y qué variables influyen en dicha clasificación. Para lograr este objetivo, se presenta una variedad de técnicas de interpretabilidad local ampliamente utilizadas en la literatura. Se emplean 5 técnicas específicas: LIME, SHAP, Explicaciones contrafácticas, anclas y los gráficos de ICE. Se aplican sobre un modelo de ML específico, en este caso una red neuronal entrenada para clasificar solicitudes de préstamos como “buenas” o “malas”. Cada una de estas técnicas generará explicaciones que permitirán comprender el proceso de toma de decisiones del modelo, brindando información valiosa sobre las variables más importantes que influyen en la clasificación de las solicitudes de préstamos.

Se ha realizado un análisis de resultados en 3 observaciones seleccionadas aleatoriamente, teniendo en cuenta la frecuencia con la que las variables son consideradas relevantes y la coherencia de los resultados con el contexto financiero. Estos análisis muestran la importancia de variables como la edad, la cantidad solicitada, la duración del préstamo y la presencia de una gran cantidad de dinero en la cuenta de ahorros en la clasificación del riesgo crediticio para ciertas observaciones. Sin embargo, no se llega a un consenso claro para todas las variables, lo que subraya la complejidad y la necesidad de un enfoque local robusto en la interpretabilidad de estos modelos.

Palabras clave: interpretabilidad, riesgo de crédito, Machine Learning, SHAP, LIME, ICE, Explicaciones contrafácticas, anclas.

Abstract

In the field of Machine Learning (ML), there has been a remarkable evolution in the complexity of models, which has improved their predictive capability. However, this high complexity has led to a challenge in terms of interpretability, as models become less understandable to users. In this context, this paper focuses on addressing the problem of local interpretability in ML models, specifically in the context of financial risk. In the field of financial risk, the interpretability of ML models is of vital importance to promote transparency in the decisions made by such models. The ability to understand and explain predictions or ratings is essential to ensure ethical decisions and to comply with increasing regulatory requirements in the financial sector. Therefore, the aim of this paper is to identify and apply the most relevant interpretability techniques in the field of credit risk.

The focus will be on local interpretability techniques, which allow to analyse and understand model decisions at the level of each individual observation. By applying these techniques to a specific credit risk case, the aim is to obtain a detailed view of how the model classifies risk and which variables influence this classification. To achieve this goal, a variety of local interpretability techniques widely used in the literature are presented. Five specific techniques are employed: LIME, SHAP, Counterfactual explanations, anchors and the ICE graphs. They are applied on a specific ML model, in this case a neural network trained to classify loan applications as “good” or “bad”. Each of these techniques will generate explanations that will provide insight into the decision-making process of the model, providing valuable information on the most important variables that influence the classification of loan applications.

An analysis of results has been carried out on 3 randomly selected observations, considering the frequency with which the variables are considered relevant and the consistency of the results with the financial context. These analyses show the importance of variables such as age, the amount requested, the duration of the loan and the presence of a large amount of money in the savings account in the classification of credit risk for certain observations. However, no clear consensus is reached for all variables, which underlines the complexity and the need for a robust local approach in the interpretability of these models.

Keywords: interpretability, credit risk, Machine Learning, SHAP, LIME, ICE, counterfactual explanations, anchors.

Agradecimientos

En primer lugar, me gustaría expresar mi agradecimiento a tutora, Jenny Alexandra Cifuentes Quintero, por su apoyo constante y valiosos comentarios a lo largo de este trabajo. Su conocimiento y experiencia han sido fundamentales para el desarrollo y la finalización exitosa de este Trabajo de Fin de Grado.

También quiero agradecer a mi familia y amigos por su apoyo y acompañamiento durante toda la carrera.

Índice general

Índice de figuras

Índice de tablas

Acrónimos

CFPB	Consumer Financial Protection Bureau
ECOA	Equal Credit Opportunity Act
IA	Inteligencia Artificial
ICE	Individual Conditional Expectations
LIME	Local Interpretable Model-Agnostic Explanations
ML	Machine Learning
OECD	Organización para la Cooperación y el Desarrollo Económicos
PDP	Partial Dependence Plot
RGPD	Reglamento General de Protección de Datos
SHAP	SHapley Additive exPlanations
UE	Unión Europea
XAI	Explainable Artificial Intelligence

Capítulo 1

Introducción

El uso de los modelos de Machine Learning (ML) durante los procesos de toma de decisiones ha experimentado una rápida y extensa expansión. La implementación de estas técnicas ha revolucionado diversos sectores y áreas del conocimiento gracias a su excepcional capacidad predictiva (Angelov, Soares, Jiang, Arnold, y Atkinson, 2021). De hecho, el gran desarrollo tecnológico que ha tenido lugar en este campo ha llevado a la generación y perfeccionamiento de nuevos métodos que han reportado métricas de desempeño aún más precisas a lo largo del tiempo. Estos nuevos métodos se tratan de modelos más complejos, conocidos como modelos de *caja negra*, que han incluido la caracterización de relaciones no lineales, diversas interacciones entre variables y un elevado número de parámetros. Sin embargo, aunque sus métricas de desempeño son muy precisas, el carácter interpretable de estos modelos ha disminuido considerablemente.

En este punto, es importante destacar que la *interpretabilidad* se define como el grado en que los humanos pueden entender la causa de una decisión tomada por un modelo. Considerando esta definición, los modelos llamados de caja negra no son interpretables debido a que su funcionamiento y razonamiento no son fácilmente comprensibles. En la figura 1.1 se puede observar la relación entre la complejidad de los modelos de ML y su interpretabilidad. Los modelos más complejos ofrecen mejores predicciones pero su interpretabilidad disminuye considerablemente, mientras que los modelos más interpretables son más sencillos computacionalmente y tienen una capacidad predictiva menor (Miller, 2019).

El creciente incremento del uso de estos modelos de gran complejidad en los procesos de toma de decisiones ha generado una necesidad de transparencia para asegurar su uso apropiado y ético (Arrieta et al., 2020). Aunque en algunos casos es suficiente conocer el resultado predicho, en otros es importante entender e interpretar las decisiones tomadas por el modelo (Molnar, 2020). Para ello, se han desarrollado diversas estrategias de interpretabilidad en diferentes sectores de aplicación. Por ejemplo, en el sector de la salud, estos análisis han sido implementados en áreas como inmunología, neurología y oncología (Bloch y Friedrich, 2021; Yu, Wei, Deng, Peng, y Hu, 2021; Yuan et al., 2021). Asimismo, en el sector finan-

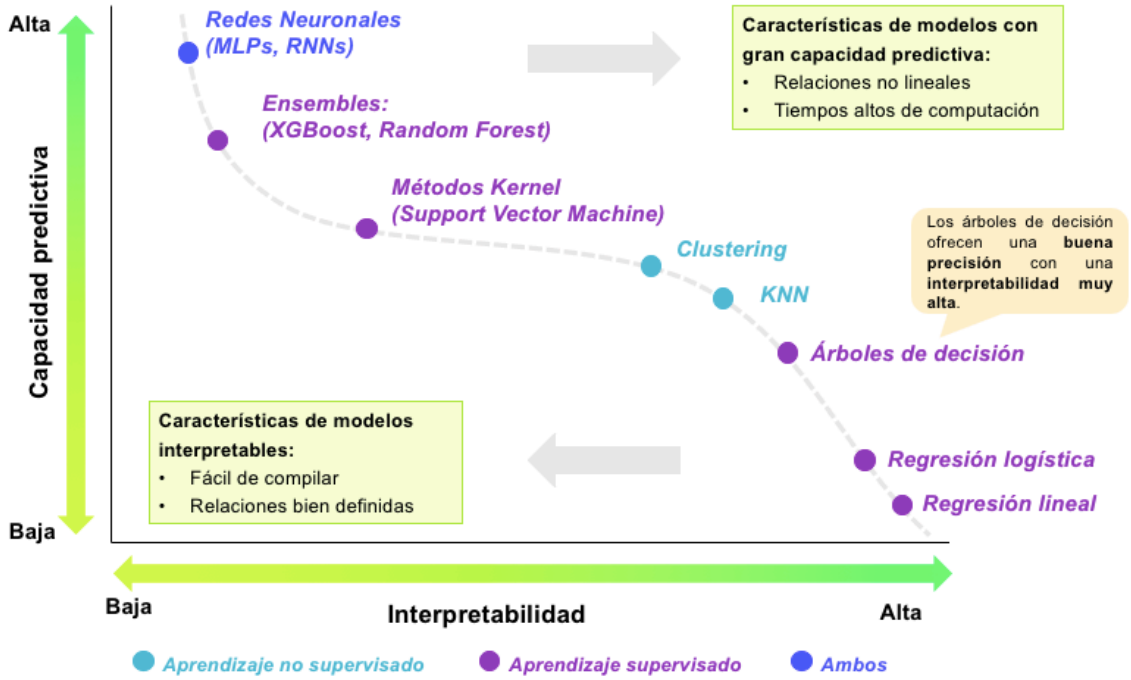


Figura 1.1: Compromiso entre Interpretabilidad y Desempeño en modelos de ML

Elaboración Propia, adaptada de: Srushti Dhamangaonkar (2020)

ciero se han incorporado para mejorar la experiencia del usuario, personalizar las ofertas de producto y servicios, detectar posibles causas de fraude, diseñar estrategias de trading, entre otras (Chen et al., 2022; Jiang, Wang, y Zhao, 2019). Es importante destacar que, aunque la complejidad de los modelos más precisos puede disminuir su interpretabilidad (Miller, 2019), la transparencia en estos modelos es esencial para su correcta implementación y aceptación.

En este contexto, cabe mencionar que la necesidad de transparencia y explicabilidad en los modelos de ML no solo es una preocupación ética y moral, sino que también se ha convertido en una necesidad legal en muchos lugares del mundo. En Europa, por ejemplo, el Reglamento General de Protección de Datos (RGPD) de la Unión Europea reconoce el derecho de las personas a ser informadas del razonamiento detrás de las decisiones tomadas por algoritmos de ML (European Parlament and the Council, 2016). De manera similar, en Estados Unidos, el Consumer Financial Protection Bureau (CFPB) establece que es necesario proveer explicaciones con los motivos del rechazo cuando se deniega un crédito (Consumer Financial Protection Bureau, 2018). Estas regulaciones buscan garantizar la transparencia y la rendición de cuentas en los procesos de toma de decisiones automatizados.

Además, es importante destacar que la necesidad de interpretabilidad en ML surge en gran parte del concepto de justicia ("Fairness"), en el que estos análisis buscan asegurar la equidad para evitar cualquier tipo de discriminación social, como la discriminación por raza, religión o sexo (Cornacchia, Narducci, y Ragone, 2021). La principal razón de que se generen

decisiones discriminatorias en los modelos de ML es la presencia de sesgos en los datos de entrenamiento. Por lo tanto, lograr un carácter interpretable en el modelo permite la detección de sesgos y predicciones erróneas realizadas de manera sistemática.

De hecho, en Estados Unidos, la Ley de Igualdad de Oportunidades de Crédito (Equal Credit Opportunity Act, ECOA) prohíbe a los prestamistas cualquier tipo de discriminación hacia los solicitantes del crédito por raza, color, religión, origen, sexo, estado civil o edad (The United States Department of Justice, 2022). Además, el grupo G20, los miembros de la Organización para la Cooperación y el Desarrollo Económicos (OECD) y algunos países adheridos se han comprometido a unos principios éticos de Inteligencia Artificial (IA) por los cuales deben velar por la igualdad, asegurándose que los modelos deben tomar decisiones justas y que los actores de la IA deben comprometerse a ofrecer transparencia e información pertinente para entender la correspondiente toma de decisiones (OECD, 2019). Enlazando la regulación con el concepto de fairness, estos principios éticos buscan garantizar la equidad en los modelos de IA y evitar cualquier tipo de discriminación social en la toma de decisiones automatizadas.

Es así como la interpretabilidad de los modelos de ML es una característica clave que permite conocer las variables más importantes en su funcionamiento y, de esta manera, dar trazabilidad a la justificación de las decisiones tomadas. También es posible caracterizar las relaciones entre las variables, lo cual facilita la detección de sesgos. Por ejemplo, si se quiere saber si la raza étnica ha incrementado o no la probabilidad de que se conceda un crédito, la interpretabilidad permitiría detectar este tipo de sesgos y evitar la discriminación. Por último, otra ventaja importante de la interpretabilidad en los modelos de ML es que facilita la detección de errores, también conocido como "debugging". Cuando un modelo tiene un rendimiento bajo, si se entiende por qué y cómo el modelo toma las decisiones, será más fácil identificar por qué las predicciones fallan y corregir los errores. En resumen, la interpretabilidad en los modelos de ML es una herramienta valiosa que ayuda a garantizar la equidad y la transparencia en la toma de decisiones automatizadas.

En los últimos años, ha habido un aumento significativo en el interés por el desarrollo de técnicas de interpretabilidad. Este creciente interés ha dado lugar al surgimiento de *Explainable AI* (XAI), un campo de estudio que se centra en comprender los modelos de ML. Para comprender a fondo esta área de investigación, es fundamental familiarizarse con ciertos conceptos clave. La interpretabilidad en el contexto del aprendizaje automático puede referirse tanto a la capacidad de entrenar modelos que sean intrínsecamente interpretables como a la tarea de explicar un modelo una vez que ha sido entrenado. Estos enfoques se conocen como explicabilidad *ante-hoc* y *post-hoc*, respectivamente. Dentro de los métodos de post-hoc, abordados en este trabajo de grado, existen dos categorías principales de técnicas de interpretabilidad. En primer lugar, están los métodos específicos al modelo, los cuales se aplican exclusivamente a ciertos tipos de modelos. Estos métodos aprovechan las características y estructuras particulares de los modelos específicos para generar explicaciones comprensibles.

Por ejemplo, en el caso de los árboles de decisión, se pueden visualizar las reglas de división y las rutas de decisión.

En segundo lugar, encontramos los métodos agnósticos al modelo, que son más flexibles, ya que pueden aplicarse a cualquier tipo de modelo, independientemente de su complejidad o estructura. Estos métodos se basan en técnicas generales de interpretabilidad que no dependen de las particularidades de un modelo específico. Algunos ejemplos de métodos agnósticos al modelo incluyen la importancia de características, las aproximaciones lineales locales y las técnicas de descomposición de la contribución de características. El uso de métodos agnósticos al modelo facilita las comparaciones entre diferentes modelos, lo cual es especialmente útil cuando se desarrollan múltiples modelos para abordar un mismo problema (Molnar, 2020). En el contexto de este trabajo de fin de grado, se dará un enfoque específico a los métodos agnósticos al modelo. El objetivo es explorar su aplicabilidad en el ámbito financiero y su potencial para brindar una mayor transparencia en el proceso de toma de decisiones.

Dentro de los métodos agnósticos, es importante distinguir dos tipos de interpretabilidad: la interpretabilidad global y la interpretabilidad local. La **interpretabilidad global** se enfoca en comprender el modelo en su conjunto, considerando cómo las variables de entrada son utilizadas por el modelo para tomar decisiones. Este enfoque es especialmente útil cuando se busca analizar el funcionamiento general del modelo, identificar sesgos o detectar posibles problemas. Sin embargo, obtener una interpretabilidad global completa puede resultar desafiante en la práctica debido a la complejidad y la opacidad de muchos modelos de ML. Por otro lado, la **interpretabilidad local** se centra en explicar cómo las diferentes variables del modelo han influido en una observación específica. Permite comprender el impacto de cada variable en la predicción realizada para esa observación en particular. Además, proporciona la capacidad de simular observaciones cercanas a la observada, lo que permite analizar cómo cambios en las variables de entrada afectarían a la predicción. La interpretabilidad local es especialmente útil en casos donde las decisiones individuales pueden tener consecuencias significativas, como en el ámbito del diagnóstico médico o la evaluación crediticia. La Figura 1.2 proporciona una representación visual que ilustra la diferencia entre los dos tipos de interpretabilidad mencionados anteriormente, resaltando la perspectiva global y local en la comprensión de los modelos de ML.

Este trabajo de fin de grado se enfocará en el análisis de técnicas de interpretabilidad local. El objetivo es explorar cómo estas técnicas pueden proporcionar una mayor comprensión de los modelos de ML en el contexto financiero y mejorar la transparencia en la toma de decisiones de riesgo financiero. Al utilizar métodos de interpretabilidad local, se podrán examinar de manera más detallada las influencias específicas de las variables en las predicciones del modelo y se podrán realizar análisis de sensibilidad para evaluar el impacto de cambios en las variables de entrada.

Teniendo en cuenta las consideraciones descritas anteriormente, el objetivo de este trabajo

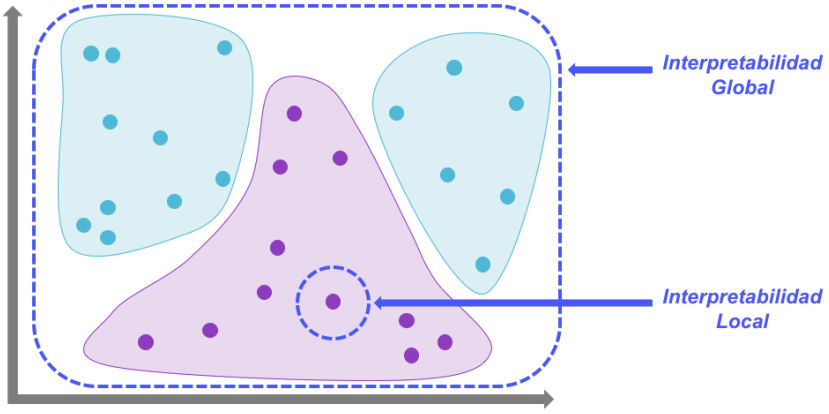


Figura 1.2: Interpretabilidad global y local

Elaboración Propia

es aplicar las técnicas de interpretabilidad local disponibles al área de riesgo financiero. En el contexto de la importancia de la transparencia en el ámbito financiero, donde las decisiones tomadas tienen un gran impacto en la vida de las personas, es esencial garantizar que los modelos de ML utilizados sean éticos y justos. Además, dada la regulación y supervisión estricta en el uso de modelos de ML en este ámbito, la interpretabilidad es un factor crítico para su adopción y aplicación efectiva. En la siguiente sección, se presentarán los objetivos detallados de este trabajo.

1.1. Objetivos

1.1.1. Objetivo General

El objetivo de este trabajo es evaluar las técnicas de interpretabilidad local agnósticas para modelos de Machine Learning, en un contexto de aplicación a la gestión de riesgos financieros.

1.1.2. Objetivos Específicos

- Contextualizar la definición de interpretabilidad en el área de Machine Learning e identificar los factores de importancia en su aplicación.
- Identificar y exponer las diferentes técnicas de interpretabilidad local agnósticas para modelos de Machine Learning, resaltando sus principales ventajas y desventajas.
- Entrenar un modelo no interpretable en un caso de gestión de riesgos financieros.

- Evaluar las técnicas más usadas en interpretabilidad local para el caso de aplicación con el fin de brindar un análisis de importancia de variables, a nivel local.

1.2. Organización de la Memoria

El presente documento se estructura en cinco capítulos que abordan diversos aspectos relacionados con la interpretabilidad en modelos de ML. En el primer capítulo, se ha establecido el contexto de la necesidad de interpretabilidad, brindando la motivación para el desarrollo del trabajo. Además, se han presentado los conceptos clave y se han establecido el objetivo general y los objetivos específicos del estudio. En el capítulo 2, se realizará una exhaustiva revisión sobre el desarrollo y la aplicación de las técnicas de interpretabilidad local agnósticas en modelos de ML. Se explorarán los avances más relevantes en este campo y se analizarán las investigaciones más destacadas. Posteriormente, en el capítulo 3, se llevará a cabo una revisión crítica de las técnicas de interpretabilidad identificadas. Se evaluará su eficacia, limitaciones y aplicabilidad en el contexto de la interpretabilidad local para modelos de ML.

En el capítulo 4, se desarrollará un caso de uso específico. En este capítulo, se presentarán los datos utilizados, la metodología empleada para entrenar el modelo y la aplicación de las técnicas de interpretabilidad local. Además, se analizarán y discutirán los resultados obtenidos a partir de la aplicación de dichas técnicas. Finalmente, en el capítulo 5, se presentarán las conclusiones derivadas del trabajo realizado. Se resumirán los hallazgos más relevantes, se discutirán las implicaciones de los resultados y se ofrecerán recomendaciones para futuras líneas de investigación en el campo de la interpretabilidad en modelos de ML.

Capítulo 2

Interpretabilidad Local en Riesgos Financieros: Una revisión de métodos y aplicaciones

En este capítulo se aborda el progreso y desarrollo de las técnicas de interpretabilidad local agnóstica al modelo, centrándose específicamente en su aplicación en el ámbito de la gestión de riesgo financiero. El objetivo principal de implementar estas técnicas en este sector es brindar una comprensión más profunda de los modelos de Machine Learning utilizados, con el fin de aumentar la confianza en las predicciones realizadas. La relevancia de este capítulo se deriva del notable crecimiento en la investigación en este campo durante los últimos años. Es importante destacar que el propósito de este capítulo es presentar las técnicas, analizar el estado actual de su investigación y examinar sus aplicaciones en el sector financiero. Se profundizará en cada una de estas técnicas y su funcionamiento en el Capítulo 3.

La primera técnica de interpretabilidad local que se abordará tiene su origen en una técnica de interpretabilidad global conocida como Gráfico de Dependencia Parcial (PDP, Partial Dependence Plot), introducida por (Friedman, 2001). Las PDPs proporcionan una representación gráfica de la relación entre las variables predictoras y la variable de interés en general. A partir de esta técnica, (Goldstein, Kapelner, Bleich, y Pitkin, 2015) desarrollaron los gráficos de Expectativa Condicional Individual (ICE, Individual Conditional Expectation), los cuales funcionan de manera similar al PDP, pero se enfocan en una observación específica. En el ámbito del riesgo financiero, la técnica ICE ha sido implementada para interpretar modelos que predicen la solvencia bancaria (Fernández, 2020), la tasa porcentual anual de préstamos hipotecarios (Gill, Hall, Montgomery, y Schmidt, 2020) y la morosidad de estos préstamos (Farzad, 2019). Además, en muchas ocasiones, se utiliza la técnica ICE para evaluar la fiabilidad y consistencia de la PDP, lo que la convierte, en realidad, en una mejora del enfoque PDP.

En cuanto a las técnicas más ampliamente implementadas en el análisis de interpretabili-

dad local en modelos de caja negra, se destacan dos en particular: Local Interpretable Model-Agnostic Explanations (LIME) y SHapley Additive exPlanations (SHAP) (Cornacchia et al., 2021; Misheva, Osterrieder, Hirsa, Kulkarni, y Lin, 2021). Estas técnicas se centran en el análisis de la importancia de las características, es decir, en determinar qué características son más relevantes para el modelo y cómo afectan a sus predicciones, lo que permite a los usuarios comprender el impacto de cada característica en los resultados. En 2016, (Ribeiro, Singh, y Guestrin, 2016) presentaron LIME con el objetivo de lograr que los seres humanos puedan confiar en los modelos de ML, asegurándose de que las explicaciones proporcionadas por esta técnica sean fácilmente comprensibles. Un año después, en 2017, (Lundberg y Lee, 2017) propusieron la técnica de interpretabilidad SHAP basada en los valores de Shapley de la teoría de juegos. Desde su presentación, tanto LIME como SHAP se han convertido en dos de las técnicas de interpretabilidad más populares. Como resultado, han sido utilizadas en diversas aplicaciones en el área de riesgo crediticio. Por ejemplo, (Misheva et al., 2021) entrenaron modelos de ML para predecir el incumplimiento de pagos en préstamos entre particulares (*P2P lending*) y aplicaron tanto SHAP como LIME para interpretar los modelos. Estos análisis permitieron concluir que las explicaciones proporcionadas por ambas técnicas eran consistentes y contaban con lógica financiera. Hallazgos similares se han obtenido en el ámbito de la detección de fraude crediticio, donde la aplicación de SHAP y LIME ha demostrado aumentar la confianza de los usuarios en el algoritmo predictivo (Ji, 2021).

Sin embargo, se han identificado aspectos a mejorar. Como es el caso de (Cornacchia et al., 2021), quienes ponen en manifiesto que ambas técnicas están centradas en los propios desarrolladores de los modelos, ya que para poder comprender las explicaciones que estas técnicas generan se necesita cierto conocimiento sobre el funcionamiento del algoritmo. Ante esto, proponen un paso adicional tras la aplicación de las técnicas de interpretabilidad que consiste en la generación de explicaciones en lenguaje natural, para que estas sean más claras y comprensibles para el usuario final. Este problema ha sido también reconocido por (Z. Zhang, Wu, Qu, y Chen, 2022) que destaca como, a día de hoy, el estudio de la interpretabilidad en los modelos de caja negra se ha centrado mayoritariamente en proveer a los ingenieros de ML de técnicas para realizar el debugging de sus modelos. Ante esto, se propone un enfoque nuevo a la hora de interpretar los modelos de caja negra que consiste en identificar las necesidades de los stakeholders y responder directamente a ellas. Concretamente, este modelo propuesto se centra en predecir si una empresa acabará en bancarrota o no. Al enfocarse en resolver las necesidades de los stakeholders del problema, no solo se entiende como afectan las variables, sino que facilita el encontrar métodos para que estas empresas eviten la bancarrota. Estas investigaciones resaltan la importancia de ir más allá de la mera generación de explicaciones técnicas y enfocarse en proporcionar explicaciones más comprensibles y relevantes para los usuarios finales.

Con este enfoque centrado en el usuario final, el uso de explicaciones contrafácticas emerge como una herramienta valiosa para ofrecer soluciones que puedan cambiar el resultado de

una decisión. Las explicaciones contrafácticas (*Counterfactual Explanations*), una técnica de interpretabilidad local agnóstica al modelo, se basan en identificar las características que, de haber sido diferentes, habrían llevado a un resultado o situación alternativa. Estas explicaciones se originan en el campo de la psicología, ya que reflejan un razonamiento inherente al ser humano (Byrne, 2016). En el contexto del riesgo crediticio, las explicaciones contrafácticas pueden ser especialmente útiles. Por ejemplo, si se deniega un crédito a un solicitante, las explicaciones contrafácticas pueden proporcionar información sobre qué variables específicas deberían haber sido modificadas para aumentar la probabilidad de aprobación en el futuro. Esta capacidad de ofrecer soluciones concretas y accionables no solo ayuda a comprender el impacto de las variables en la decisión, sino que también brinda a los usuarios la oportunidad de tomar medidas concretas para mejorar sus posibilidades de obtener crédito.

La última técnica de interpretabilidad local agnóstica al modelo en surgir son los "anchors" o anclas, presentados por (Ribeiro, Singh, y Guestrin, 2018), los mismos investigadores que introdujeron la técnica LIME. Estas anclas se centran en la búsqueda de reglas simples que expliquen las predicciones del modelo de una manera comprensible para los usuarios. En el campo de la evaluación crediticia, los "anchors" han sido implementados con resultados satisfactorios. Las explicaciones generadas por esta técnica han generado resultados de gran utilidad y suficientemente detallados y comprensibles para los usuarios involucrados en el proceso de toma de decisiones (Demajo, Vella, y Dingli, 2020). Además, se ha realizado su implementación en el ámbito de la auditoría, con el propósito de evaluar el riesgo de error material del cliente a nivel de los estados financieros. En este contexto, se ha observado que estas técnicas son capaces de generar explicaciones fácilmente comprensibles, lo cual es especialmente relevante en el campo de la auditoría, donde se requiere una interpretación clara y precisa de los resultados obtenidos (C. A. Zhang, Cho, y Vasarhelyi, 2022). El enfoque basado en "anchors" ofrece una forma intuitiva y transparente de comprender las decisiones tomadas por los modelos de caja negra. Al identificar reglas simples que explican las predicciones, los usuarios pueden tener una mayor confianza en los resultados y comprender el razonamiento detrás de las decisiones.

En resumen, se han identificado cinco técnicas de interpretabilidad local, agnósticas al modelo, ampliamente utilizadas, las cuales han demostrado ser efectivas en la comprensión de los modelos de ML en el ámbito financiero: los gráficos de Expectativa Condicional Individual (ICE), Local Interpretable Model-Agnostic Explanations (LIME), SHapley Additive exPlanations (SHAP), explicaciones contrafácticas y Anchors. La Tabla 2.1 presenta las referencias de las aplicaciones en el ámbito financiero mencionadas previamente, junto con el objetivo del artículo en el que se presentaron y el resultado concreto de la técnica de interpretabilidad implementada. Estas técnicas ofrecen explicaciones interpretables a nivel local, lo que resulta invaluable para los profesionales financieros al permitirles comprender las razones subyacentes a las predicciones realizadas por los modelos. Esto a su vez les brinda la capacidad de tomar decisiones más fundamentadas y confiables en la gestión del riesgo

financiero. Aunque cada una de estas técnicas aborda la interpretabilidad desde diferentes enfoques, todas comparten el objetivo de proporcionar una comprensión clara y precisa de cómo las variables influyen en las predicciones del modelo.

En el siguiente capítulo se proporcionará una explicación detallada del funcionamiento de cada una de estas técnicas de interpretabilidad local agnósticas al modelo. Además, se llevará a cabo un análisis crítico de sus fortalezas y limitaciones, considerando su aplicabilidad en el contexto financiero. Esto permitirá una evaluación más completa de estas técnicas y facilitará la selección adecuada de la herramienta interpretable más apropiada para abordar los desafíos específicos de la gestión del riesgo financiero.

Tabla 2.1: Resumen de las aplicaciones de las técnicas de interpretabilidad en el ámbito del riesgo financiero.

Referencia	Estrategia de Interpretabilidad	Objetivo	Resultados
(Fernández, 2020)	ICE	Analizar la predicción de la estabilidad bancaria en Estados Unidos de 1990 a 2017, utilizando variables de solvencia, morosidad y un indicador de estabilidad bancaria ad hoc	En el índice de estabilidad bancaria formado, la variable más importante en su predicción es e10 (Rendimiento Corporativo BBB). Para valores superiores al 8.5 %, el índice toma valores muy bajos, lo que indica una mayor inestabilidad bancaria.
(Gill et al., 2020)	ICE	Ofrecer un marco de referencia viable para las aplicaciones de aprendizaje automático que requieren una gran precisión e interpretabilidad y que mitiguen los riesgos de discriminación.	Se diseña un modelo de caja negra utilizando datos de préstamos hipotecarios para predecir la tasa porcentual interanual. Las fluctuaciones en las gráficas ICE muestran un posible sobreajuste o filtración de señales no monótonas de características importantes.
(Farzad, 2019)	ICE	Determinar los factores que influyen en los impagos de préstamos hipotecarios.	Durante el período de 2013Q1 a 2017Q4, se descubrió que los factores más importantes que predicen una morosidad de 90 días en préstamos hipotecarios son aquellos que afectan los términos del contrato, especialmente la tasa de interés.
(Misheva et al., 2021)	SHAP y LIME	Contribuir a la literatura existente explorando la utilidad de las técnicas SHAP y LIME en el contexto de la gestión del riesgo de crédito.	Se analizó un modelo que predecía dificultades a la hora de pagar un crédito. El análisis de los resultados de LIME indicó que una persona que no tiene recuperaciones en su cuenta, no tiene cargos por pagos atrasados, una tasa de interés más baja y pagos totales más altos, puede considerarse como un cliente sin riesgo. El análisis con SHAP también mostró que una cantidad de préstamo más baja está asociada con una probabilidad más baja de incumplimiento de pago.
(Ji, 2021)	SHAP y LIME	Evaluar los métodos XAI para la detección de fraudes con tarjetas de crédito.	El principal indicador de una transacción fraudulenta es el método de uso de la tarjeta. Esto tiene sentido ya que la mayoría de transacciones fraudulentas ocurren online.
(Z. Zhang et al., 2022)	Explicaciones contrafácticas	Establecer un nuevo enfoque de XAI centrado en las necesidades de los stakeholders externos.	Para disminuir la probabilidad de bancarrota de una empresa estudiada, se recomienda incrementar la tasa de crecimiento del beneficio operativo junto con la tasa de crecimiento del flujos de efectivo por actividades de explotación.
(Demajo et al., 2020)	Anchors	Proponer un modelo de evaluación crediticia preciso e interpretable mediante el uso de algoritmos de ML y técnicas de explicabilidad.	Tras entrenar un modelo de ML y generar explicaciones locales con anchors, expertos en el campo de préstamos y riesgos fueron entrevistados y todos estuvieron de acuerdo con la predicción del modelo al evaluar las solicitudes de préstamo, además determinaron que las explicaciones proporcionadas por los anchors eran correctas y fáciles de entender.
(C. A. Zhang et al., 2022)	Anchors	Presentar a los profesionales e investigadores de la auditoría las técnicas de XAI, y mostrar ejemplos de técnicas populares de interpretabilidad aplicadas específicamente a la auditoría.	El anchor generado explica que el modelo predice que habrá un error material en las cuentas anuales porque su Deuda a Largo Plazo, se encuentra entre 19.75 y 695.29, y que su ratio Ventas/Facturación Neta, es superior a 13.25.

Capítulo 3

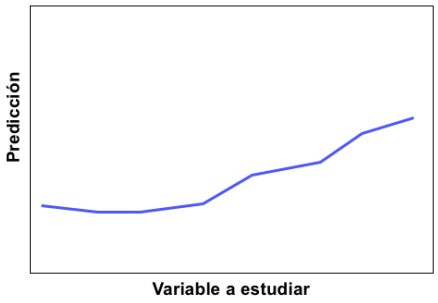
Estrategias de Interpretabilidad Local Agnósticas al Modelo

Todas las técnicas revisadas en el Capítulo 2 ofrecen enfoques distintos para lograr la interpretación local de modelos de aprendizaje automático. Aunque comparten el objetivo de brindar explicaciones interpretables, cada técnica tiene su propio funcionamiento y enfoque característico. En este capítulo, se realizará una explicación conceptual detallada de cada técnica, se analizarán cuidadosamente sus ventajas y desventajas, y se expondrán las limitaciones específicas de cada una.

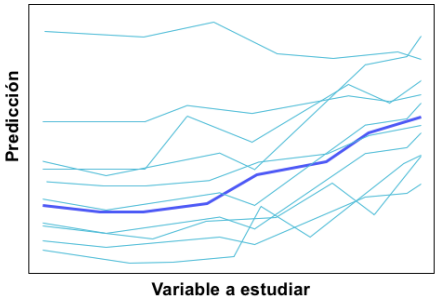
3.1. Gráficos de Expectativa Condicional Individual

Los gráficos de Expectativa Condicional Individual (ICE) consisten en una representación visual que permite examinar el efecto que tienen las variaciones en las variables predictoras sobre las predicciones del modelo. Cada línea del gráfico representa una observación individual, lo que permite una representación visual clara y concisa de cómo cambian las predicciones a medida que se modifican características específicas. Este enfoque visual ofrece una perspectiva intuitiva y detallada de la relación entre las características y las predicciones del modelo, brindando una mayor comprensión de los factores que influyen en los resultados. A nivel global, al presentar todas las líneas en un solo gráfico, se pueden identificar patrones de comportamiento en diferentes subconjuntos de observaciones y detectar aquellas que se desvían del comportamiento general. Esto brinda una visión más completa y valiosa de cómo las variables afectan las predicciones en diferentes contextos, mejorando así la interpretación y comprensión del modelo en su conjunto.

La Figura 3.1a muestra el gráfico ICE correspondiente a una sola observación, caracterizado por su simplicidad y facilidad de comprensión cuando se desea analizar el comportamiento específico de dicha observación. Por otro lado, la Figura 3.1b representa tanto la observación en cuestión (resaltada en un color diferente) como las demás observaciones del



(a) Ejemplo de gráfico ICE para una única observación



(b) Ejemplo de gráfico ICE para varias observaciones

Figura 3.1: Gráficos de ICE

Elaboración Propia: adaptada de: (Wai On, 2020)

conjunto de datos. Esta visualización permite evaluar si la observación en estudio presenta un comportamiento similar al de las demás observaciones o si se trata de un valor atípico (*outlier*). Sin embargo, en conjuntos de datos extensos, esta visualización puede resultar compleja y dificultar la identificación de patrones o la distinción entre las observaciones.

El proceso para obtener los datos necesarios para construir el gráfico es el siguiente (Goldstein et al., 2015). En primer lugar, se elige la observación específica que se desea estudiar y la variable cuya variación se quiere analizar. Se extraen los valores únicos que toma dicha variable dentro del conjunto de datos. Para cada uno de estos valores, se obtienen las predicciones del modelo manteniendo los valores constantes de las demás variables correspondientes a la observación seleccionada. Al obtener múltiples predicciones para diferentes valores de la variable, se construye un gráfico que muestra la relación entre la variable seleccionada y la respuesta del modelo. Es importante destacar que este proceso puede repetirse para diferentes observaciones y variables, lo que permite explorar y comparar el comportamiento de diferentes casos.

Como se puede observar, el gráfico ICE es una técnica de interpretación simple y fácil de implementar. Sin embargo, es fundamental comprender sus limitaciones. En primer lugar, en conjuntos de datos extensos, las visualizaciones pueden volverse complejas y sobrecargadas, especialmente cuando se busca una interpretabilidad global. Sin embargo, en este trabajo nos centramos en la interpretabilidad local, lo que nos permite generar gráficos de ICE para observaciones individuales o subconjuntos específicos de datos seleccionados, superando así la complejidad visual asociada a conjuntos de datos grandes. En segundo lugar, los gráficos de ICE asumen la independencia entre las variables predictoras, es decir, que las variables no están correlacionadas entre sí. Sin embargo, en la realidad, es común que existan correlaciones entre las variables, lo que puede afectar la precisión de los resultados. Además, al construir los gráficos de ICE, los puntos de datos pueden representar combinaciones improbables o incluso imposibles de existir en la realidad cuando hay correlación entre las variables.

Por último, es importante destacar que los gráficos ICE se enfocan en mostrar el efecto de una sola característica en la predicción. Sin embargo, se ha considerado la posibilidad de incorporar una segunda variable predictora al gráfico, representando ambas variables en los ejes y visualizando el efecto del cambio a través de una escala de colores (Z. Zhang et al., 2022). No obstante, es importante tener en cuenta que al añadir una segunda variable, la interpretación de los resultados puede volverse más compleja. Estas limitaciones deben considerarse al utilizar los gráficos ICE, lo que resalta la importancia de realizar una cuidadosa selección de las variables sobre las cuales se aplicará esta técnica.

3.2. Local Interpretable Model-Agnostic Explanations

La técnica de interpretabilidad LIME (Local Interpretable Model-Agnostic Explanations) es ampliamente reconocida por su capacidad de proporcionar explicaciones locales precisas para modelos de ML. LIME se basa en el entrenamiento de modelos sustitutos locales, los cuales son construidos utilizando algoritmos interpretables como árboles de decisión o regresión lineal. Estos modelos sustitutos se entrenan mediante la generación de una muestra de datos modificando las variables de la observación de interés, lo que permite crear un conjunto de datos "cercano" a la observación estudiada.Una vez obtenido este conjunto de datos, se introducen en el modelo de ML original para obtener las predicciones correspondientes. Posteriormente, se entrena un modelo interpretable utilizando este nuevo conjunto de datos, el cual busca aproximar las predicciones del modelo de caja negra a nivel local. De esta manera, el modelo sustituto proporciona una explicación detallada el efecto de cada variable en la predicción de la observación estudiada.

Es importante destacar que LIME considera la distancia entre las nuevas instancias generadas y la observación estudiada durante el entrenamiento del modelo sustituto. Esto implica asignar un mayor peso a aquellas instancias que se encuentran más cercanas, ya que se considera que tienen una mayor influencia en la construcción del modelo interpretable (Ribeiro et al., 2016). Esta estrategia asegura que las interpretaciones locales sean más precisas y relevantes. La combinación de modelos interpretables que generan explicaciones locales precisas y el enfoque agnóstico al modelo ha posicionado a LIME como una de las técnicas más utilizadas en el campo de la interpretabilidad de modelos de ML.

En la Figura 3.2, se puede observar de manera visual la aplicación de la técnica LIME, brindando una comprensión intuitiva de su funcionamiento. En esta representación, la cruz más gruesa representa la observación específica que se está tratando de explicar, mientras que las demás cruces y puntos corresponden a las observaciones generadas por LIME. La ubicación y tamaño de estas observaciones generadas por LIME representan el peso asignado a cada una de ellas durante el proceso de entrenamiento del modelo sustituto. Este peso se basa en la distancia entre las observaciones generadas y la observación de interés. Cuanto más

cercana esté una observación generada a la observación estudiada, mayor será su influencia en la construcción del modelo sustituto. Utilizando estas observaciones generadas y ponderadas, se genera una explicación local que se representa visualmente en la Figura 3.2 mediante una línea discontinua. Es importante destacar que la Figura 3.2 es solo un ejemplo ilustrativo, y en aplicaciones reales, el número de observaciones generadas por LIME puede ser mucho mayor.

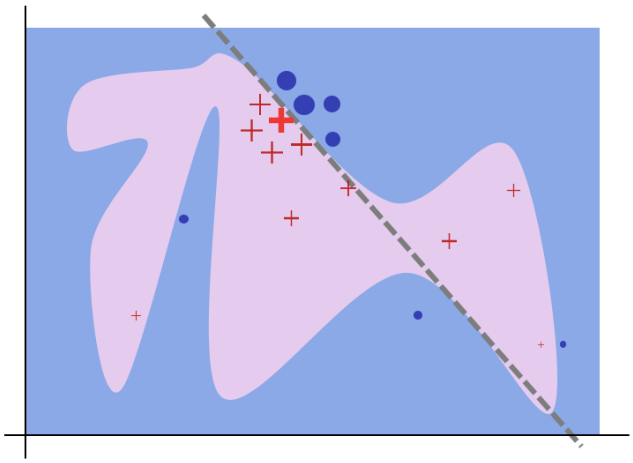


Figura 3.2: Representación gráfica de la lógica de LIME.

Elaboración Propia: adaptado de (Ribeiro et al., 2016)

La implementación de la técnica LIME ha revelado ciertos desafíos que es importante tener en cuenta. En primer lugar, se ha observado una falta de estabilidad en las explicaciones generadas por LIME. Debido a que los datos utilizados para entrenar el modelo sustituto se generan de manera aleatoria, es posible obtener explicaciones diferentes si se repite el proceso de muestreo. Esta variabilidad en las explicaciones puede reducir la confianza en los resultados y dificultar la interpretación precisa de las observaciones. En segundo lugar, LIME emplea un kernel como función para asignar pesos a las nuevas observaciones en función de su distancia a la observación de interés. El ancho del kernel desempeña un papel crucial, ya que determina cómo se distribuyen los pesos. Un kernel con un ancho pequeño asignará pesos significativos únicamente a las observaciones muy cercanas, mientras que un ancho más amplio otorgará mayor peso a las observaciones más alejadas. Sin embargo, encontrar el valor óptimo del ancho de kernel es una dificultad en sí misma, ya que no existe un método definido para determinarlo. Esto puede llevar a dificultades al interpretar el grado de influencia de las diferentes observaciones en la explicación generada.

Además, al igual que los gráficos de ICE y las explicaciones contrafácticas (Ver sección 3.4), LIME no tiene en cuenta la correlación entre las variables. Debido a que el muestreo utilizado para generar los puntos de entrenamiento es aleatorio, existe la posibilidad de que se generen puntos de datos que sean imposibles o poco probables de existir en la realidad. Esta limitación puede afectar la precisión de las explicaciones y su capacidad para reflejar

el comportamiento real del modelo. Estos desafíos resaltan la importancia de considerar las limitaciones y posibles sesgos al utilizar la técnica de LIME. Es esencial comprender la variabilidad en las explicaciones, explorar diferentes valores de ancho de kernel y ser consciente de la falta de consideración de la correlación entre variables.

3.3. Shapley Additive exPlanations (SHAP)

SHAP (Shapley Additive Explanations) es una técnica de interpretación de modelos de aprendizaje automático que se enfoca en comprender la influencia de cada variable predictora en las predicciones realizadas por el modelo. La principal característica de SHAP radica en el cálculo de los **valores de Shapley**, que representan la contribución media de una variable teniendo en cuenta todas las posibles combinaciones de variables y considerando todas las permutaciones posibles. Esta consideración del orden de las combinaciones es fundamental, ya que tiene un impacto en la contribución de cada variable. Al calcular los valores de Shapley, se obtiene una medida precisa de la importancia relativa de cada variable en el proceso de toma de decisiones del modelo (Lundberg y Lee, 2017). Estos valores de Shapley se utilizan para construir explicaciones individuales (locales), donde se detalla la influencia de cada variable en la predicción de una observación específica. Mediante esta metodología de descomposición aditiva, se pueden obtener información detallada y cuantitativa sobre la contribución de cada variable en la predicción final del modelo.

SHAP ofrece una serie de ventajas significativas, entre las cuales destaca su capacidad para proporcionar explicaciones tanto a nivel global como local. A nivel global, SHAP evalúa la importancia relativa de cada variable al considerar su impacto en todas las predicciones realizadas por el modelo. Esto permite identificar las variables que tienen un efecto más significativo y más limitado en las predicciones generales del modelo. Por otro lado, a nivel local, SHAP ofrece explicaciones específicas y detalladas para una observación particular. Esto significa que no solo se obtiene una comprensión general de la influencia de las variables en las predicciones, sino que también se pueden analizar esas influencias en un caso específico.

Dentro de la técnica SHAP, existen dos variantes principales: KernelSHAP y TreeSHAP (Mayer, Meier, y Wuthrich, 2023). KernelSHAP utiliza un enfoque basado en un kernel para ponderar las observaciones generadas en cada permutación. Este kernel asigna pesos a las diferencias entre las instancias perturbadas y las instancias originales, lo que permite determinar la importancia relativa de cada característica en la predicción del modelo. En este sentido, KernelSHAP comparte similitudes con LIME, ya que ambos asignan pesos a las observaciones generadas. Sin embargo, mientras que LIME asigna los pesos según la distancia a la observación que se está explicando, SHAP se enfoca en la presencia o ausencia de variables en las permutaciones. Esto se debe a que se obtiene más información sobre las

contribuciones de cada variable cuando se trata de la única característica presente o de la única característica ausente. Estos pesos se utilizan luego para calcular las contribuciones individuales de cada característica en el resultado final.

Por otro lado, TreeSHAP es una variante de SHAP específicamente diseñada para modelos basados en árboles de decisión. A diferencia de KernelSHAP, que es agnóstica al modelo y puede aplicarse a cualquier tipo de modelo de aprendizaje automático, TreeSHAP se enfoca en modelos basados en árboles de decisión, como árboles de clasificación o regresión, bosques aleatorios o gradient boosting. Una de las ventajas clave de TreeSHAP radica en su eficiencia computacional. Es menos costosa y más rápida en comparación con KernelSHAP, especialmente cuando se trabaja con conjuntos de datos grandes o modelos complejos que contienen numerosos árboles. Esto se debe a que TreeSHAP aprovecha las estructuras inherentes de los árboles de decisión para calcular las contribuciones individuales de manera eficiente, lo que reduce significativamente los tiempos de computación y permite un análisis rápido y efectivo de la importancia de las variables.

3.4. Explicaciones contrafácticas

Las explicaciones contrafácticas se enfocan en la interpretabilidad del modelo desde la perspectiva de alterar las predicciones. Su objetivo es identificar las variables que deberían haber sido diferentes para obtener un resultado alternativo. Para generar estas explicaciones, se crean instancias ficticias que se asemejan a una instancia dada, pero con características modificadas. Estas modificaciones se realizan con el propósito de explorar los efectos de esos cambios en las características en las predicciones del modelo. En resumen, la explicación contrafáctica analiza el cambio mínimo necesario en una variable de entrada para lograr un cambio en la predicción (Molnar, 2020).

La principal ventaja que ofrecen las explicaciones contrafácticas es un gran nivel de claridad al usuario final ya que se proporciona una explicación precisa de las variables que requieren modificación. De esta manera, se destacan las características que deben alterarse en una instancia dada para obtener un resultado diferente. Esto facilita la comprensión y la toma de decisiones por parte del usuario, ya que se le brinda una guía clara sobre los aspectos relevantes y su influencia en los resultados deseados. Además, esta claridad permite que incluso el usuario que no disponga de grandes conocimientos sobre ML sea capaz de entender las explicaciones.

Las explicaciones contrafácticas enfrentan una limitación compartida con las técnicas ICE y LIME cuando las variables presentan una alta correlación entre sí, lo que implica que modificar una variable puede tener un impacto en las demás. En estas situaciones, las explicaciones contrafácticas pueden perder su utilidad y resultar difíciles de implementar en la práctica (Taly, Ankur and Shanbhag, Aalok, 2020). Cuando las variables están altamente

correlacionadas, las explicaciones contrafácticas pueden generar resultados poco realistas o inviables en términos prácticos. Esto se debe a que modificar una variable sin considerar su relación con las demás puede llevar a escenarios poco probables o incluso imposibles de alcanzar en la realidad. Por lo tanto, es necesario tener en cuenta esta consideración adicional al interpretar y utilizar las explicaciones contrafácticas en tales escenarios. En casos de alta correlación entre variables, es fundamental reconocer las restricciones impuestas por estas relaciones y tener precaución al extraer conclusiones de las explicaciones contrafácticas. Es importante evaluar cuidadosamente si las modificaciones propuestas son factibles y coherentes dentro del contexto del problema en cuestión. Además, se deben considerar enfoques alternativos o complementarios para abordar la influencia de las variables correlacionadas, como análisis de sensibilidad o técnicas de selección de características.

Además, es importante tener en cuenta que las explicaciones contrafácticas pueden generar resultados contradictorios entre sí, aunque todos puedan ser considerados como válidos y correctos. Esto se debe a que existen diferentes formas o caminos posibles para modificar las variables de una observación y así cambiar la predicción. Esta característica puede plantear un desafío en la interpretación de las explicaciones contrafácticas, ya que puede generar cierta ambigüedad o incertidumbre al momento de seleccionar la mejor explicación o tomar decisiones basadas en estas explicaciones. Por lo tanto, al utilizar explicaciones contrafácticas, es fundamental comprender que pueden existir diferentes soluciones posibles y que la elección de una explicación en particular dependerá del contexto específico.

3.5. *Anchors* o anclas

Los anchors son una técnica de interpretabilidad que busca identificar reglas simples que expliquen las predicciones realizadas. Los anchors se centran en encontrar condiciones específicas en los datos que, si se cumplen, conducen a una determinada predicción, aunque cambien el resto de las características. En otras palabras, buscan variables que 'anclan' la predicción, de allí, su nombre. Al igual que LIME, evalúa puntos de datos cercanos a la observación estudiada. Estas reglas son fáciles de entender, ya que el formato de expresión es 'SI-ENTONCES'. Es decir, si la variable 'A' tomara el valor X entonces la predicción tomaría el valor Y (Mi, Li, y Zhou, 2020). En la Figura 3.3 puede observarse de manera visual el funcionamiento de los anchors. Los círculos que rodean las observaciones estudiadas representa el alcance de la muestra utilizada para establecer las reglas, y los cuadrados con línea discontinua representan las reglas locales identificadas.

Esta estructura condicional brinda una gran ventaja, ya que no se requiere un conocimiento profundo sobre el aprendizaje automático para comprender las explicaciones generadas. Al igual que la técnica LIME, los anchors evalúan puntos de datos cercanos a la observación de interés para obtener explicaciones locales. Esto implica que los anchors se enfocan en

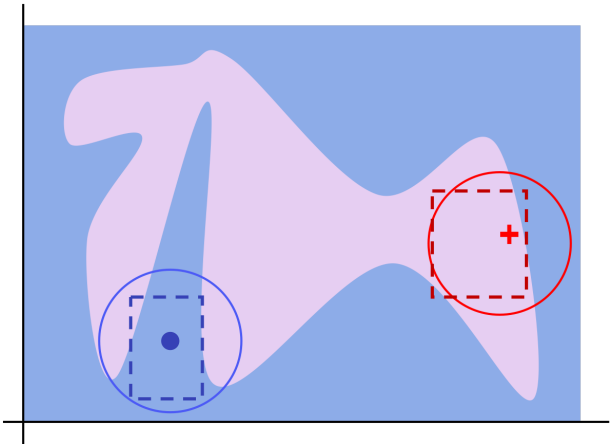


Figura 3.3: Representación gráfica de la lógica de los anchors.

Elaboración Propia: adaptado de (Mi et al., 2020)

encontrar reglas específicas que sean aplicables a la observación analizada y que pueden no generalizarse a otras instancias del conjunto de datos.

El proceso para generar los anchors implica la generación de explicaciones candidatas y la selección de las mejores entre ellas. Comienza generando explicaciones individuales considerando cada variable de entrada para la observación seleccionada. Estas explicaciones iniciales se evalúan y se seleccionan las que proporcionan las mejores interpretaciones de las predicciones del modelo. A partir de las explicaciones seleccionadas, se procede a generar explicaciones candidatas adicionales mediante la gradual incorporación de otras variables de entrada a las explicaciones existentes. Esto implica la exploración de diferentes combinaciones de variables para ampliar el alcance de las reglas identificadas. Cada nueva explicación candidata se evalúa en términos de su precisión. El proceso de selección de los mejores anchors se basa en encontrar las explicaciones que logren un equilibrio entre la simplicidad y la precisión. Se busca identificar las reglas más concisas y fácilmente comprensibles que aún sean capaces de capturar las relaciones relevantes entre las variables y las predicciones del modelo (Molnar, 2020).

Sin embargo, una desventaja importante de los anchors radica en su rendimiento computacional en conjuntos de datos muy grandes o con un gran número de variables. A medida que aumenta la complejidad del conjunto de datos, el proceso de búsqueda y selección de reglas puede volverse computacionalmente costoso y requerir un tiempo significativo para obtener resultados. A pesar de esta limitación, los anchors siguen siendo una técnica valiosa para generar explicaciones interpretables en diversos contextos. Su enfoque en reglas simples y fáciles de entender permite que los usuarios no especializados en aprendizaje automático comprendan y confíen en las explicaciones proporcionadas por los modelos. Además, el proceso de selección de los mejores anchors permite obtener explicaciones más precisas y significativas, brindando una mayor confianza en la interpretación de las predicciones del modelo.

Capítulo 4

Caso de Uso

El objetivo de este capítulo es aplicar las técnicas de interpretabilidad previamente mencionadas a un caso concreto de riesgo de crédito, a fin de obtener una comprensión más clara y detallada de las decisiones tomadas por el modelo. Para lograrlo, se seguirá un enfoque metodológico estructurado, que abarcará las siguientes etapas (Ver Figura 4.1):

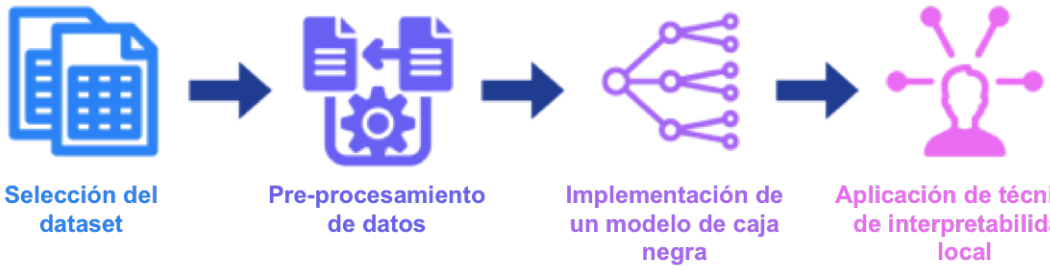


Figura 4.1: Flujo de actividades

Elaboración Propia

1. **Selección del dataset:** Se realizará la selección cuidadosa de un conjunto de datos apropiado para el análisis de riesgos financieros en el ámbito de crédito. Este dataset deberá contener información relevante sobre variables financieras, características del cliente y del prestamo, entre otros aspectos relevantes.

2. **Pre-procesamiento de datos:** Se llevará a cabo un proceso de pre-procesamiento de los datos seleccionados, que incluirá tareas como transformación y normalización de variables, así como la codificación adecuada de variables categóricas. Este paso es fundamental para asegurar la calidad y coherencia de los datos antes de su análisis.

3. **Implementación de un modelo de caja negra:** Se construirá y entrenará un modelo de aprendizaje automático de caja negra o de baja interpretabilidad específico para el problema de clasificación o regresión en cuestión. Este modelo será el punto de partida para la aplicación de las técnicas de interpretabilidad local.

4. **Aplicación de técnicas de interpretabilidad local:** Una vez que se haya entrenado el modelo, se aplicarán las técnicas de interpretabilidad local explicadas en el Capítulo 3. Estas técnicas permitirán obtener explicaciones detalladas a nivel de instancia, revelando la influencia de cada variable en las predicciones del modelo. Se explorarán técnicas como ICE, LIME, SHAP, anchors o explicaciones contrafácticas para analizar el comportamiento del modelo y generar información interpretable sobre las decisiones del modelo. Durante el análisis de interpretabilidad, se han utilizado diversas librerías para implementar las diferentes técnicas, con el objetivo de obtener explicaciones más claras y comprensibles. A continuación, se detallan las librerías utilizadas para cada técnica:

    - LIME: librería 'lime'.
    - SHAP: librería 'shap'.
    - Explicaciones contrafácticas: librería 'dice_ml'. Además, para cada observación se han generado 4 explicaciones para aumentar la probabilidad de que alguna de ellas sea viable y fácil de implementar.
    - Anchors: librería 'anchor-exp'. Durante la generación de los anchors se ha establecido un umbral de precisión del 80 % para asegurar una suficiente confianza en las explicaciones generadas.
    - Gráficos de ICE: librería pycebox.

A lo largo de todo el proceso, se ha empleado Python como lenguaje de programación principal. Python ofrece una ventaja significativa frente a otros lenguajes debido a su ecosistema robusto y diverso. PyPI (Python Package Index) es un repositorio que alberga una amplia variedad de librerías y paquetes desarrollados por la comunidad de Python. Esta característica enriquece considerablemente el entorno de Python, proporcionando numerosas opciones y herramientas para el desarrollo de programas. De hecho, varias de las librerías utilizadas en el análisis de interpretabilidad se encuentran disponibles en PyPI, lo que facilitó su instalación y utilización en el proyecto. Además, Python cuenta con una comunidad activa y comprometida, lo que se traduce en una abundancia de recursos y soluciones disponibles. El código desarrollado durante este trabajo se encuentra disponible en: `https://github.com/mariasanmartin/analisis_interpretabilidad/`.

4.1. Descripción y tratamiento de los datos

La predicción del riesgo de crédito es de suma importancia para las entidades bancarias y financieras, ya que les permite evaluar el nivel de riesgo asociado a cada solicitud de préstamo. Esta evaluación tiene un impacto directo en la toma de decisiones sobre la aprobación

del préstamo, así como en las condiciones y tasas de interés aplicadas. Además de su relevancia para las entidades financieras, la predicción del riesgo de crédito está influenciada por regulaciones y requisitos legales, como se ha explicado en el Capítulo 1. La transparencia en estos procesos es fundamental para garantizar la equidad, evitar la discriminación y proteger los derechos de los consumidores. En este sentido, la implementación de técnicas de interpretabilidad contribuye a dicha transparencia al proporcionar explicaciones claras y comprensibles sobre los factores que influyen en la evaluación del riesgo de crédito. Esto permite que los consumidores comprendan mejor las razones detrás de las decisiones tomadas por las instituciones financieras y, además, facilita la presentación de sugerencias y opciones centradas en el cliente final para aumentar sus posibilidades de que su solicitud sea aceptada en el futuro.

Para abordar un tema de aplicación en este campo, se utilizará un conjunto de datos disponible en Kaggle (`https://www.kaggle.com/datasets/kabure/german-credit-data-with-risk`). Este conjunto de datos consta de 1000 observaciones y 10 variables. Cada observación representa a un individuo que ha solicitado un préstamo a una entidad bancaria. El objetivo será entrenar un modelo de aprendizaje automático con el fin de predecir qué préstamos se pagarán sin problemas y, por lo tanto, deben ser concedidos, y cuáles podrían enfrentar dificultades y requerir una evaluación más exhaustiva. Es un problema de clasificación binaria, ya que la variable dependiente (‘Risk’) es binaria, tomando el valor ‘good’ en caso de que el préstamo pueda ser concedido sin problemas, y el valor ’bad’ en caso de que presente un mayor riesgo y requiera una evaluación más detallada.

Una vez seleccionado el conjunto de datos con el que se va a trabajar, se procedió al tratamiento de los datos, comenzando por la identificación de los valores faltantes en dicho conjunto. Durante este proceso, se detectaron dos variables con valores ausentes: ‘Checking account’, con un 39,4 % de valores faltantes, y ‘Savings account’, con un 18,3 %. Con el objetivo de establecer una gestión adecuada de los valores faltantes, se definió un umbral máximo del 30 %. En consecuencia, se optó por eliminar la variable ‘Checking account’ debido a su alta proporción de valores ausentes. Por otro lado, en el caso de la variable ‘Savings account’, que presentaba un porcentaje de valores faltantes del 18,3 %, por debajo del umbral establecido, se decidió crear una nueva categoría denominada ‘no_inf’ para todas las observaciones que carecían de información en dicha variable. Esta estrategia permite preservar la integridad de los datos disponibles y evitar la pérdida de observaciones relevantes en el análisis.

En este punto del estudio, es de vital importancia adquirir un profundo entendimiento de cada variable y su significado, con el objetivo de llevar a cabo un procesamiento de datos adecuado y, posteriormente, lograr un análisis de interpretabilidad efectivo. Cada variable contenida en el conjunto de datos posee información única y relevante que influye en la predicción del riesgo crediticio. Por lo tanto, es esencial examinar en detalle la tabla 4.1, la cual proporciona las definiciones de cada una de las variables presentes en el conjunto de datos.

Variable	Tipo	Significado
Age	Cuantitativa discreta	Edad del solicitante en años
Sex	Cualitativa binaria	Sexo del solicitante
Job	Cualitativa ordinal	Experiencia en el ámbito laboral (0 - unskilled and non-resident, 1 - unskilled and resident, 2 - skilled, 3 - highly skilled).
Housing	Cualitativa nominal	Situación de propiedad sobre la vivienda (own, rent, or free)
Saving accounts	Cualitativa nominal	(little, moderate, quite rich, rich)
Credit amount	Cuantitativa discreta	Cantidad solicitada (en marcos alemanes)
Duration	Cuantitativa discreta	Duración del préstamo en meses
Purpose	Cualitativa nominal	La finalidad que se le va a dar al dinero (car, furniture/equipment, radio/TV, domestic appliances, repairs, education, business, vacation/others)
Risk	Cualitativa binaria	Evaluación del préstamo (good, bad)

Tabla 4.1: Explicación de variables.
Elaboración Propia

Dado que la variable 'Job' representa categorías de trabajo que están ordenadas de menor a mayor experiencia, se puede tratar como una variable cuantitativa discreta, utilizando los valores 0, 1, 2 y 3 para representar diferentes niveles de experiencia. En esta codificación, se asigna el valor 0 a los individuos con menos experiencia y que no son residentes, el valor 1 a los individuos con menos experiencia pero que sí son residentes, el valor 2 a los individuos con experiencia y el valor 3 a los individuos con alta experiencia. Esta codificación permite representar la información ordinal de la variable 'Job' en una forma numérica. Por su parte, para incorporar variables cualitativas al modelo, se ha adoptado el enfoque del One-Hot encoding. Este método implica crear variables binarias para cada valor único presente en la variable categórica que se está codificando. De esta manera, si una observación cumple con una determinada característica, la variable correspondiente toma el valor 1; de lo contrario, toma el valor 0. Para evitar la colinealidad, se elimina una de las variables creadas mediante One-Hot encoding de cada una de las variables originales, es decir, si hay N valores únicos en una variable cualitativa se crean $N - 1$ variables binarias, de esta manera se garantiza la independencia lineal entre estas nuevas variables.

Finalmente, se realizaron análisis de correlación entre las variables, y no se encontraron correlaciones fuertes entre ellas. Debido a esto, se procedió al último paso del preprocesamiento de los datos, que consiste en dividirlos en conjuntos de entrenamiento y prueba. En este caso, se decidió asignar el 70 % de los datos disponibles (700 observaciones) al conjunto de entrenamiento, mientras que el 30 % restante (300 observaciones) se utilizará más adelante para evaluar el rendimiento del modelo. Al finalizar esta etapa, se obtuvieron cuatro conjuntos de datos, cada uno con las siguientes características:

- Las variables predictoras de entrenamiento.
- La variable objetivo de entrenamiento.
- Las variables predictoras de prueba.
- La variable objetivo de prueba.

Esta división de los datos es esencial para garantizar una evaluación imparcial y precisa del modelo, ya que los conjuntos de entrenamiento y prueba permiten simular el desempeño del modelo en datos nuevos, o no vistos previamente.

4.2. Entrenamiento de la red neuronal

El enfoque principal de este trabajo se centra en la interpretabilidad, priorizando la comprensión del razonamiento subyacente del modelo sobre la maximización de su rendimiento predictivo. Con este objetivo en mente, se decidió utilizar una red neuronal sencilla con solo dos capas ocultas. Aunque este tipo de modelo puede ser menos complejo en términos de arquitectura y número de parámetros, ha demostrado una precisión (*accuracy*) del 74,3 %, lo cual se considera satisfactorio para los propósitos de este estudio. Por lo tanto, las predicciones generadas por este modelo serán utilizadas en las etapas posteriores del análisis.

La elección de una red neuronal como modelo de machine learning se basa en su capacidad demostrada para ofrecer una excelente capacidad predictiva. Sin embargo, también se reconoce que las redes neuronales pueden presentar una alta complejidad en cuanto a su interpretación. Estas redes consisten en capas compuestas por “neuronas” que realizan operaciones matemáticas en los datos de entrada y transmiten los resultados a las neuronas de la siguiente capa (Figura 4.2). Durante el entrenamiento, los pesos de las conexiones entre las neuronas se ajustan para permitir que la red aprenda a reconocer patrones en los datos y realice predicciones (Gurney, 1997).

Para el entrenamiento de la red neuronal, se ha utilizado la librería ‘scikit-learn’, una potente librería de Python especializada en machine learning que ofrece una amplia gama de técnicas y herramientas para su desarrollo (Pedregosa et al., 2011). La elección de esta librería se basa en su facilidad de uso, versatilidad y la comunidad activa que la respalda, lo

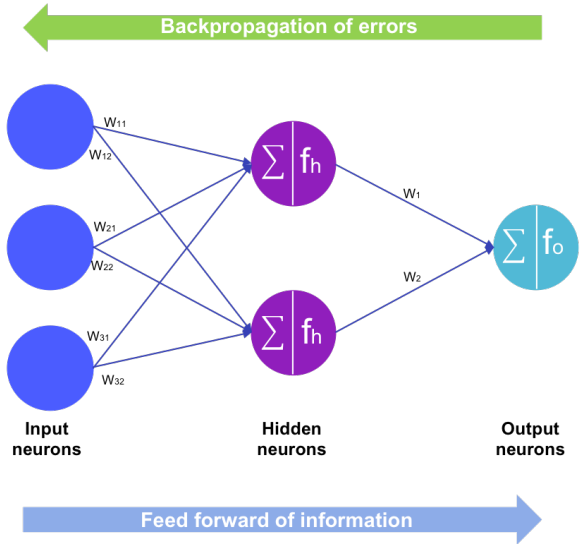


Figura 4.2: Diagrama de una red neuronal

Elaboración Propia, adaptada de: (Dieterle, 2003)

que permite aprovechar las capacidades de machine learning de manera eficiente y efectiva en este estudio.

4.3. Análisis de interpretabilidad

Con respecto al análisis de la interpretabilidad, en este estudio, se ha optado por utilizar métodos de interpretabilidad local. Para ello, se han seleccionado tres observaciones de forma aleatoria, que serán sometidas a un análisis detallado. Estas observaciones representan casos individuales dentro del conjunto de datos y proporcionan una visión del funcionamiento del modelo en situaciones concretas. En la Tabla 4.2, se pueden apreciar las características particulares de estas observaciones.

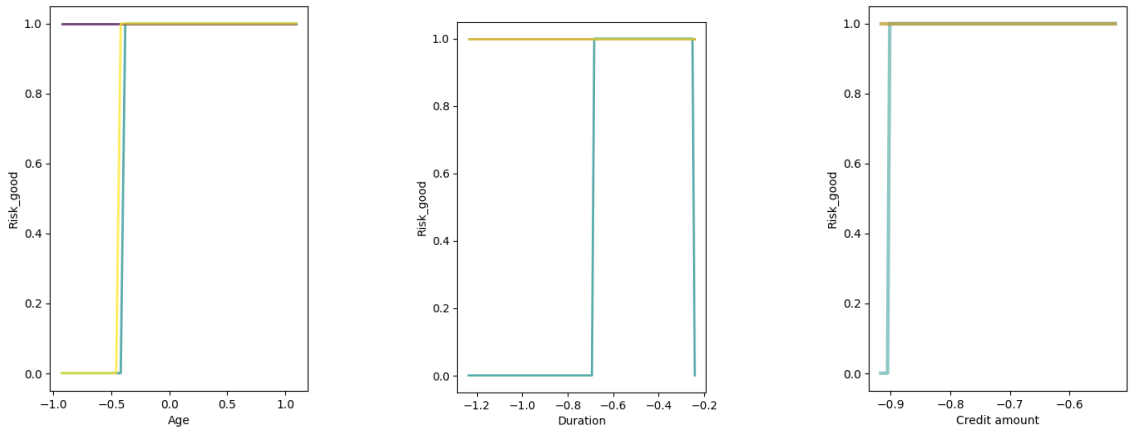
ID	Age	Sex	Job	Housing	Savings account	Credit amount	Duration	Purpose	Risk
313	25	male	1	own	little	685	12	car	bad
485	47	male	3	own	little	1209	6	car	bad
801	48	female	1	rent	little	1795	18	radio/TV	good

Tabla 4.2: Observaciones seleccionadas para el análisis y sus características.
Elaboración Propia

Es importante mencionar que tanto la observación 313 como la 801 son clasificadas correctamente por el modelo, mientras que, la observación 485 es clasificada como de riesgo bueno por el modelo cuando en realidad es de riesgo malo.

4.3.1. Gráficos de Expectativa Condicional Individual

Para comenzar, en la Figura 4.3 se presentan los gráficos de ICE para las tres variables cuantitativas del conjunto de datos: Edad, Duración del préstamo y Cantidad solicitada. En estos gráficos, el eje 'x' representa la variable independiente, mientras que el eje 'y' muestra la predicción del riesgo. Las observaciones 313, 485 y 801 se representan en color azul, morado y amarillo respectivamente. En relación a la variable de edad, se observa que tanto la observación 313 como la observación 801 presentan una disminución del riesgo a medida que aumenta la edad, mientras que para la observación 485, la edad no parece afectar la calificación del riesgo. La duración según los gráficos de ICE solo afecta a la observación 313, donde existe un rango de duración que disminuye el riesgo. En cuanto a la duración del préstamo, se observa que en la observación 313 existe un rango de duración que conlleva a una disminución del riesgo. Esto indica que la duración del préstamo puede desempeñar un papel relevante en la clasificación de riesgo para esta observación en particular, mientras que para las otras dos observaciones (485 y 801), no se aprecia una influencia significativa de la duración. Por último, en relación a la cantidad solicitada, se observa que solo la observación 313 muestra una relación aparente con el riesgo, sin evidencias claras de influencia en las observaciones 485 y 801.



(a) Gráfico de ICE según la edad

(b) Gráfico de ICE según la duración

(c) Gráfico de ICE según la cantidad solicitada

Figura 4.3: Gráficos de ICE generados

Elaboración Propia

4.3.2. Local Interpretable Model-Agnostic Explanations (LIME)

Para continuar, se ha implementado la técnica de interpretabilida LIME. En la figura 4.4 se muestra el resultado de aplicar LIME, donde se pueden apreciar las explicaciones generadas. La simplicidad y eficiencia con la que se ha implementado LIME, gracias a la ayuda de

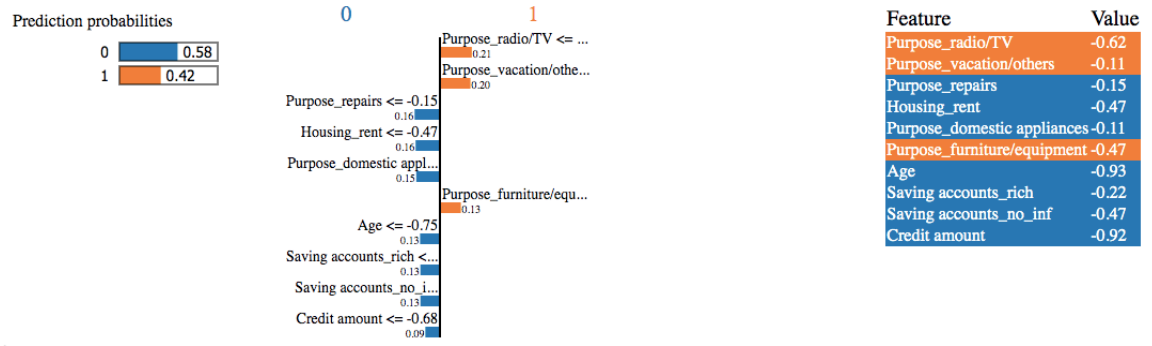
la librería *lime*, han sido aspectos destacados en este proceso de interpretabilidad. Cada subfigura se compone de tres secciones distintas que proporcionan información clave. En primer lugar, la parte izquierda refleja la estimación de probabilidad realizada por el modelo para determinar si la observación adopta un valor de 1 (la solicitud del crédito se considera de bajo riesgo y por tanto se clasifica como buena) o 0 (la solicitud del crédito se considera de alto riesgo y por tanto se clasifica como mala). Seguidamente, en la sección central se presentan las 10 variables más relevantes que han influido significativamente en el cálculo de dicha probabilidad. Por último, en la parte derecha se muestran los valores reales correspondientes a la observación en consideración

Al examinar detenidamente los resultados obtenidos mediante LIME, se puede identificar un desafío en cuanto a la interpretabilidad del modelo. Este obstáculo se relaciona con la escala de los datos utilizados durante el entrenamiento de la red neuronal, lo cual se refleja en las explicaciones proporcionadas por LIME en forma de valores escalados. Esta situación puede dificultar la comprensión y la intuición al interpretar los efectos calculados. Es importante tener presente esta limitación al analizar los resultados de LIME y considerar la necesidad de aplicar una transformación inversa a los valores escalados. Esta transformación permitirá obtener una interpretación más directa y comprensible de los efectos generados por LIME, brindando una visión más clara y significativa de la influencia de las variables en las predicciones del modelo.

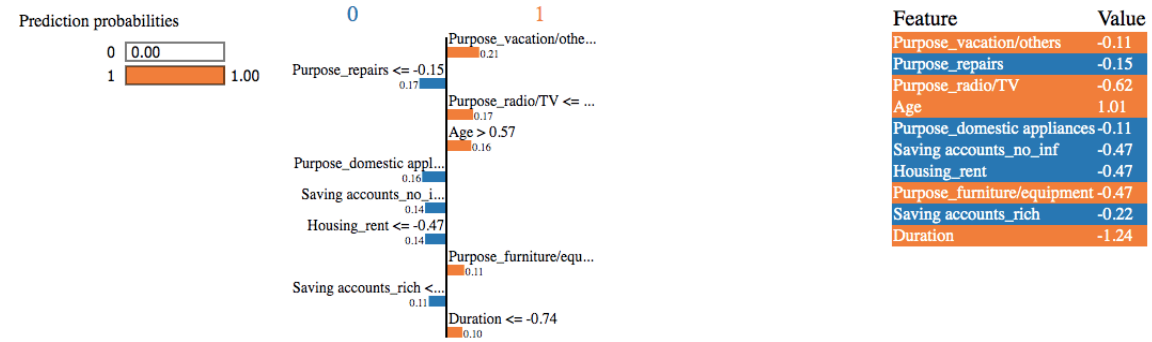
La pérdida de interpretabilidad debido al escalado de las variables afecta a todas ellas, aunque el impacto puede ser más significativo en aquellas que no son binarias. En el caso de las variables binarias, se puede aplicar una transformación sencilla en la que un valor escalado negativo se interpreta como 0, mientras que un valor positivo se interpreta como 1. Por ejemplo, en el gráfico de LIME correspondiente a la observación 801 (Ver Figura 4.4c), la variable 'Purpose_radio_TV' tiene un valor de 1.6, que en los datos desescalados equivaldría a 1. Por otro lado, la variable 'Purpose_repairs' tiene un valor de -0.15, lo que en los datos desescalados representa un 0. Además, la explicación de LIME asociada a esta variable indica que si 'Purpose_repairs' es menor que -0.15, aumenta la probabilidad de un mal crédito. Esto implica que, según el modelo, si el propósito del crédito no es realizar reparaciones, se considera más arriesgado. De esta manera, en el caso de las variables binarias, es posible preservar una interpretación intuitiva, donde un valor negativo representa 0 o la ausencia de la característica, y un valor positivo representa 1 o la presencia de la misma. Por lo tanto, el escalado de las variables binarias no compromete significativamente la interpretación de sus valores.

Por otro lado, en el caso de la variable 'Age', la explicación proporcionada por LIME indica que cuando la edad supera 0.57, el modelo la considera un indicador de buen riesgo. Si bien sabemos que el valor escalado de 'Age' es 1.1 para la observación 801, lo que corresponde a 48 años, desconocemos a qué edad se refiere el valor 0.57 y a partir de qué edad se produciría una disminución en el riesgo crediticio. Ante esta barrera, hemos logrado deses-

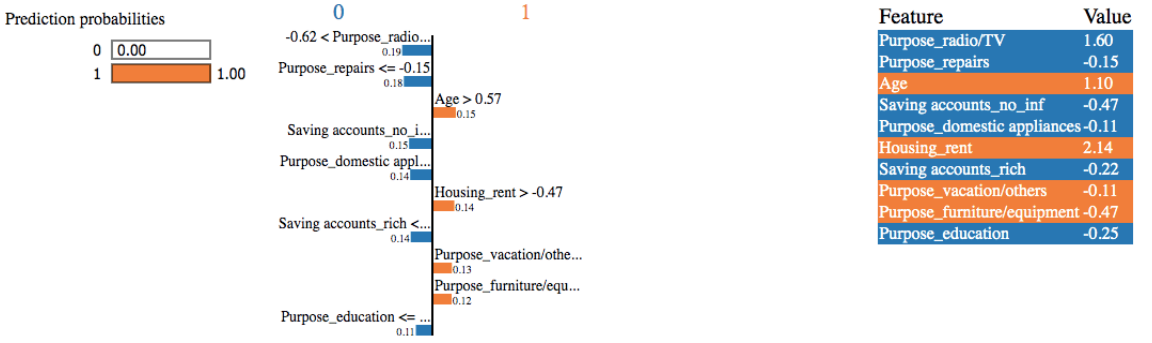
calar las variables y se presentarán a continuación los valores desescalados. Sin embargo, es importante destacar que actualmente, con las librerías disponibles en Python, no es posible visualizar directamente los valores desescalados en los gráficos de un modelo entrenado con valores escalados.



(a) Observación 313



(b) Observación 485



(c) Observación 801

Figura 4.4: Explicaciones de LIME para las observaciones objeto de estudio
Elaboración Propia

A continuación, se procederá al análisis de las explicaciones proporcionadas por LIME. Para la observación 313 (Figura 4.4a), el modelo calcula que hay un 58 % de probabilidad de que el riesgo sea malo (valor 0), y por tanto, lo clasifica como tal. Las variables que han tenido mayor peso en esta clasificación son la edad del solicitante, el hecho de que la cuenta de ahorros no sea calificada como “rich”, y que el propósito del préstamo no sea ni para electrodomésticos domésticos ni para reparaciones. Por otro lado, el hecho de que el propósito

del préstamo tampoco sea para una radio o un televisor, ni para muebles/equipamiento, ni para vacaciones, parece aumentar la calidad percibida del préstamo.

Este análisis permite identificar las variables más influyentes en la clasificación de riesgo del préstamo. La edad del solicitante emerge como un factor relevante, sugiriendo que a medida que la edad aumenta, la probabilidad de que el préstamo se considere de bajo riesgo también se incrementa. Además, la no disponibilidad de una cuenta de ahorros con un saldo elevado (denominada como "rich") se muestra como un aspecto desfavorable, ya que está asociada con una mayor probabilidad de clasificar el préstamo como de alto riesgo. Esto sugiere que cuanto mas dinero tengas ahorrado más fácil sea obtener un crédito. Asimismo, el propósito del préstamo desempeña un papel importante en la evaluación del riesgo, ya que es posible que propósitos específicos como la compra de electrodomésticos o reparaciones, se asocien con una mayor probabilidad de ser clasificados como préstamos de alto riesgo.

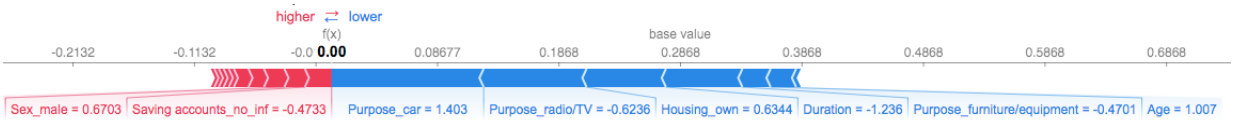
La observación 485 (Figura 4.4b) ha sido clasificada por el modelo como de riesgo "bueno". Según las explicaciones proporcionadas por LIME, el propósito del préstamo ha tenido un efecto similar al observado en la observación 313. Es decir, el hecho de que el propósito no esté relacionado con la adquisición de una radio, un televisor, muebles/equipamiento ni vacaciones ha contribuido a considerar el préstamo como de buena calidad. Además, otra variable que ha influido en la clasificación positiva del préstamo es la edad del solicitante. Según los valores proporcionados por LIME, una edad mayor a 42 años ha tenido un impacto favorable en la evaluación crediticia específica de esta observación. Es interesante destacar que mientras en la observación 313 la edad (25 años) tuvo un efecto negativo en la clasificación de riesgo, en esta observación (con una edad de 47 años) la edad ha tenido un impacto positivo. Este resultado resalta la importancia de la edad como factor determinante en la evaluación crediticia.

Por último, en el caso del préstamo de la observación 801 (Ver Figura 4.4c), también ha sido clasificado como "bueno". En esta instancia, se han identificado tres variables clave que han influido en dicha clasificación. En primer lugar, la edad del solicitante ha tenido un impacto positivo en la evaluación del préstamo, lo cual indica que una edad más avanzada se considera favorable en términos de riesgo crediticio. Además, la presencia de un contrato de alquiler en el momento actual ha contribuido a clasificar el préstamo como "bueno". Por último, se ha observado que el propósito del crédito no estar relacionado ni con vacaciones ni con muebles/equipamiento ha sido un factor determinante en la clasificación positiva del préstamo. Por otro lado, se ha identificado que la variable más influyente en la clasificación negativa ha sido el propósito del préstamo para la adquisición de una televisión o una radio. Este resultado sugiere que, en este contexto, solicitar un préstamo con el propósito de adquirir estos productos electrónicos específicos se considera un indicador de mayor riesgo.

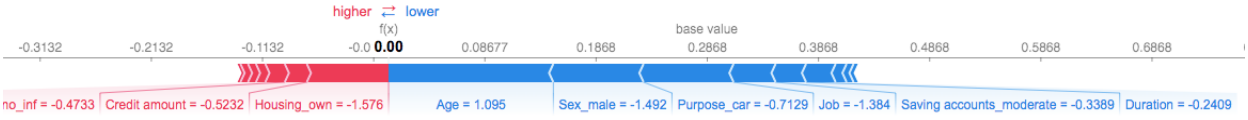
4.3.3. Shapley Additive exPlanations (SHAP)

Se ha procedido a implementar la técnica de SHAP utilizando la biblioteca ‘shap’, y específicamente se ha aplicado el método *KernelSHAP* para obtener los valores de SHAP. Con el fin de visualizar estos valores, se han generado los gráficos ’shap.plots.force’ que se presentan en la Figura 4.5. Estos gráficos representan de manera intuitiva la influencia de cada variable en la probabilidad de que la variable dependiente (riesgo del crédito) tome el valor 0. Mediante esta representación visual, es posible apreciar la naturaleza aditiva de SHAP, donde la suma de los valores de Shapley de cada variable, junto con la media de las predicciones, se iguala al resultado final. Esta característica aditiva de SHAP ofrece una comprensión clara de la contribución de cada variable al riesgo crediticio y de la influencia de su combinación en el resultado final.

(a) Observación 313



(b) Observación 485



(c) Observación 801

Figura 4.5: Valores de SHAP para las observaciones objeto de estudio
Elaboración Propia

En los gráficos, la etiqueta ‘base_value’ representa la media de la probabilidad predicha por el modelo de que la variable objetivo de la observación tome el valor 0, lo que indica un riesgo ’malo’. Los valores resaltados en rojo indican las variables que incrementan dicha probabilidad en el modelo, mientras que los valores en azul representan las variables que la disminuyen. Al igual que en LIME, los valores se presentan escalados.

En cuanto a la interpretación de la observación 313 (Figura 4.5a), se ha observado que las variables que más han influido en la clasificación del crédito como “malo” son: la edad del solicitante, la cantidad solicitada y el propósito de adquirir un coche. Estas variables han tenido una contribución significativa en la predicción del riesgo crediticio en esta observación específica. Por otro lado, se ha identificado que las variables que han contrarrestado en cierta medida la influencia negativa de las variables mencionadas anteriormente y han contribuido parcialmente a la consideración de un crédito “bueno” son: la duración del préstamo y el hecho de que el propósito del crédito no esté relacionado con la compra de una televisión o

una radio. Estas variables han desempeñado un papel importante en la mitigación del riesgo crediticio en esta observación en particular.

En la observación 485, que ha sido clasificada como de riesgo "bueno" por el modelo, se han identificado las variables que más han contribuido a esta clasificación, según los valores de SHAP (Figura 4.5b). Las principales variables que han influido positivamente en la calificación de riesgo son: el propósito del crédito ser adquirir un coche, que el propósito no esté relacionado con una TV o una radio, y si el solicitante es propietario de una vivienda. Por otro lado, se han encontrado características que han tenido un efecto contrario, es decir, una influencia negativa en la calificación de riesgo. Estas características incluyen la falta de información sobre la cuenta de ahorros y que el solicitante sea de sexo masculino. Sin embargo, es importante destacar que estas características han tenido una contribución mucho menor en comparación con aquellas que han influido positivamente en la calificación de riesgo según el modelo.

Por último, la observación 801 (Figura 4.5c) también ha sido clasificada como de riesgo "bueno" por el modelo. En este caso, se han identificado las variables que han tenido mayor influencia en esta decisión. Las características más relevantes incluyen la edad del solicitante, el género femenino y el propósito del crédito, especialmente si no es para adquirir un coche. Por otro lado, se ha observado que el hecho de que el solicitante no sea propietario de la vivienda en la que reside ha contribuido hacia la clasificación del crédito como "malo".

4.3.4. Explicaciones Contrafácticas

Otra técnica de interpretabilidad local implementada en este trabajo de fin de grado son las explicaciones contrafácticas. El individuo que representa la observación 313 será el más beneficiado o interesado en esta técnica, ya que le proporcionará información para cambiar la clasificación del modelo de riesgo "malo" a "bueno". Esto no quiere decir que para las observaciones 485 y 801 no sean de interés las explicaciones generadas, ya que les proporcionarán información sobre aquellos aspectos a los que tienen que prestar especial atención y evitar realizar cambios en ellos para no cambiar la clasificación del modelo de riesgo de "bueno" a "malo".

Las explicaciones generadas para la observación 313 están detalladas en la Tabla 4.3. En la primera explicación, se sugiere que el individuo agregue el propósito de educación en la solicitud, pero al mismo tiempo mantenga el propósito de coche. Sin embargo, esto no es factible, ya que normalmente solo se puede seleccionar un propósito en una solicitud, por lo que esta regla es descartada. La segunda explicación resalta que modificando únicamente la edad se podría influir en la evaluación del crédito. Si el individuo espera hasta alcanzar los 35 años (valor desescalado), existe una alta probabilidad de que el crédito le sea concedido. Esta información sugiere que la edad es un factor crítico en la clasificación del riesgo crediticio en este caso particular. La tercera regla identifica que otra forma de obtener la aprobación del

crédito es aumentando la cantidad solicitada de 685 a 9986 y alcanzando una clasificación en la cuenta de ahorros de 'rich'. Esta modificación en los valores del préstamo y el estado de la cuenta de ahorros incrementa las posibilidades de obtener la aprobación crediticia. Por último, la última explicación generada también resulta inviable, ya que presenta una situación similar a la primera regla mencionada. En este caso, se seleccionan dos propósitos, coche y 'domestic appliances', lo cual no es compatible con las reglas establecidas.

Por su parte, las explicaciones contrafácticas generadas para la observación 485 se detallan en la Tabla 4.4. En la primera explicación, se indica que al aumentar la cantidad solicitada de 1209 a 16108, habría una alta probabilidad de que el préstamo fuera denegado. Este incremento representa un aumento del 1292 % en la cantidad solicitada, lo cual resulta poco probable en la realidad. La segunda explicación muestra el impacto de no proporcionar información sobre la cuenta de ahorros. En este escenario contrafáctico, es probable que la solicitud sea clasificada como riesgo 'malo'. Esto resalta la relevancia de la información sobre la cuenta de ahorros en esta observación específica. La tercera explicación, al igual que la primera, enfatiza la importancia de considerar la cantidad solicitada en la evaluación del crédito. En este caso, se muestra que si la cantidad solicitada se incrementara a 16475 y la duración del préstamo se extendiera de 6 meses a 19 meses, el préstamo habría sido clasificado como riesgo 'malo'. Finalmente, la última explicación generada no es factible, ya que establece que se deben considerar tanto los propósitos de reparaciones como de educación en la solicitud, cuando en realidad solo se puede seleccionar uno de ellos.

Por último, las explicaciones generadas para la observación 801 se presentan en la Tabla 4.5. En este caso, ninguna de las explicaciones generadas resulta viable. La primera explicación propone reducir la duración del crédito de 18 meses a 15 meses, pero requiere seleccionar tanto reparaciones como radio/TV como propósitos del crédito. Sin embargo, sabemos que no es posible elegir múltiples propósitos, lo que invalida esta explicación. La segunda explicación plantea una situación similar al proponer la adquisición de un coche y una radio/TV como propósitos del crédito, lo cual también excede las restricciones establecidas. Estas dos primeras explicaciones demuestran la importancia de considerar las restricciones y limitaciones prácticas al generar explicaciones. La tercera explicación indica que si el solicitante tuviera 28 años en lugar de su edad actual, es probable que el crédito fuera clasificado como riesgo 'malo'. Aunque esta explicación no es posible, ya que la edad no puede cambiarse, sí proporciona información sobre la relevancia de la edad en la evaluación del crédito. La última explicación generada tampoco es realista, ya que implica cambiar el sexo del solicitante y seleccionar múltiples propósitos: radio/TV y reparaciones.

Tabla 4.3: Explicaciones contrafácticas generadas - Observación 313.

313											
							Variables actuales				
Age	Job	Credit Amount	Duration	Sex_male	Housing_own	Housing_rent	Saving accounts_-moderate	Saving accounts_no_-inf	Saving accounts_quite rich	Saving accounts_rich	Purpose_car
-0.927	-1.38	-0.917	-0.739	0.67	0.634	-0.467	-0.339	-0.473	-0.259	-0.224	1.403
							Explicaciones	**contrafácticas**			
-0.927	-1.38	-0.917	-0.739	0.67	0.634	-0.467	-0.339	-0.473	-0.259	-0.224	1.403
-0.039	-1.38	-0.917	-0.739	0.67	0.634	-0.467	-0.339	-0.473	-0.259	-0.224	1.403
-0.927	-1.38	2.38	-0.739	0.67	0.634	-0.467	-0.339	-0.473	3.85	-0.224	1.403
-0.927	-1.38	-0.917	-0.739	0.67	0.634	-0.467	-0.339	-0.473	-0.259	-0.224	1.403

313 (Continuación)						
			Variables actuales			
Purpose_domestic appliances	Purpose_education	Purpose_furniture/equipment	Purpose_radio/TV	Purpose_repairs	Purpose_vacation/others	Risk_good
-0.11	-0.25	-0.47	-0.623	-0.149	-0.11	0
			Explicaciones	**contrafácticas**		
-0.11	3.99	-0.47	-0.623	-0.149	-0.11	1
-0.11	-0.25	-0.47	-0.623	-0.149	-0.11	1
-0.11	-0.25	-0.47	-0.623	-0.149	-0.11	1
9.07	-0.25	-0.47	-0.623	-0.149	-0.11	1

Tabla 4.4: Explicaciones contrafácticas generadas - Observación 485.

485											
							Variables actuales				
Age	Job	Credit Amount	Duration	Sex_male	Housing_own	Housing_rent	Saving accounts_-moderate	Saving accounts_no_-inf	Saving accounts_quite rich	Saving accounts_rich	Purpose_car
1.01	1.678	-0.731	-1.236	0.67	0.634	-0.467	-0.339	-0.473	-0.259	-0.224	1.403
							Explicaciones	**contrafácticas**			
1.01	1.678	4.55	-1.236	0.67	0.634	-0.467	-0.339	-0.473	-0.259	-0.224	1.403

1.01	1.678	-0.731	-1.236	0.67	0.634	-0.467	-0.339	2.11	-0.259	-0.224	1.403
1.01	1.678	4.68	-0.15	0.67	0.634	-0.467	-0.339	-0.473	-0.259	-0.224	1.403
1.01	1.678	-0.731	-1.236	0.67	0.634	-0.467	-0.339	-0.473	-0.259	-0.224	-0.71

485 (Continuación)

			Variables actuales			
Purpose_domestic appliances	Purpose_education	Purpose_furniture/equipment	Purpose_radio/TV	Purpose_repairs	Purpose_vacation/others	Risk_good
-0.11	-0.25	-0.47	-0.623	-0.149	-0.11	1
			Explicaciones	**contrafácticas**		
-0.11	-0.25	-0.47	-0.623	-0.149	-0.11	0
-0.11	-0.25	-0.47	-0.623	6.66	-0.11	0
-0.11	-0.25	-0.47	-0.623	-0.149	-0.11	0
-0.11	3.99	-0.47	-0.623	-0.149	-0.11	0

Tabla 4.5: Explicaciones contrafácticas generadas - Observación 801.

801

							Variables actuales				
Age	Job	Credit Amount	Duration	Sex_male	Housing_own	Housing_rent	Saving accounts_-moderate	Saving accounts_no_-inf	Saving accounts_quite rich	Saving accounts_rich	Purpose_car
1.09	-1.38	-0.52	-0.24	-1.49	-1.57	2.14	-0.339	-0.473	-0.259	-0.224	-0.71
							Explicaciones	**contrafácticas**			
1.09	-1.38	-0.52	-0.468	-1.49	-1.57	2.14	-0.339	-0.473	-0.259	-0.224	-0.71
1.09	-1.38	-0.52	-0.24	-1.49	-1.57	2.14	-0.339	-0.473	-0.259	-0.224	1.40
-0.649	-1.38	-0.52	-0.24	-1.49	-1.57	2.14	-0.339	-0.473	-0.259	-0.224	-0.71
1.09	-1.38	-0.52	-0.24	0.67	-1.57	2.14	-0.339	-0.473	-0.259	-0.224	-0.71

801 (Continuación)

			Variables actuales			
Purpose_domestic appliances	Purpose_education	Purpose_furniture/equipment	Purpose_radio/TV	Purpose_repairs	Purpose_vacation/others	Risk_good

-0.11	-0.25	-0.47	1.60	-0.149	-0.11	1
			Explicaciones	**contrafácticas**		
-0.11	-0.25	-0.47	1.60	6.667	-0.11	0
-0.11	-0.25	-0.47	1.60	-0.149	-0.11	0
-0.11	-0.25	-0.47	1.60	-0.149	-0.11	0
-0.11	-0.25	-0.47	1.60	6.667	-0.11	0

4.3.5. *Anchors* o anclas

En este estudio, se ha utilizado la técnica de interpretabilidad de los "anchors" para obtener explicaciones adicionales sobre las predicciones del modelo. En la Tabla 4.7 se presentan las explicaciones generadas mediante esta técnica, las cuales revelan las condiciones específicas que, si se cumplen, mantendrían la predicción del modelo sin cambios. Estas explicaciones brindan una mayor comprensión del funcionamiento del modelo en diferentes escenarios.

Antes de analizar estas explicaciones, es importante comprender dos conceptos clave: cobertura y precisión. La cobertura se refiere a la proporción de observaciones en el conjunto de datos que pueden ser explicadas por el "anchor". En otras palabras, indica cuántas observaciones cumplen con las condiciones establecidas por el "anchor". Una cobertura alta implica que el "anchor" es aplicable a un gran número de observaciones, lo que aumenta su utilidad y generalización. Por otro lado, la precisión es un indicador de la capacidad del "anchor" para mantener la predicción del modelo sin cambios en los casos que cumplen con sus condiciones. Representa la probabilidad de que el modelo mantenga la misma clasificación cuando se cumplen las reglas generadas por el "anchor". Una alta precisión significa que las condiciones establecidas por el "anchor" son altamente confiables y respaldadas por el modelo.

En primer lugar, la observación 313 (Ver Tabla 4.7a) proporciona un ejemplo interesante para analizar los resultados obtenidos mediante la técnica de los "anchors". En este caso, se observa que el "anchor" establece un número considerablemente alto de condiciones, lo que se refleja en una cobertura baja. Esto significa que solo un pequeño porcentaje de observaciones en el conjunto de entrenamiento cumple con todas esas condiciones específicas. En concreto, la cobertura es del 2 %, lo que implica que solo 14 observaciones del conjunto de entrenamiento se ajustan a estas características establecidas por el "anchor". Es importante destacar que, si se mantienen las condiciones establecidas por el "anchor" en estas 14 observaciones, existe una probabilidad del 79 % de que la predicción del modelo se mantenga sin cambios. Sin embargo, dado que la observación 313 fue clasificada como de riesgo 'malo' por el modelo, es posible que el solicitante del préstamo no tenga un gran incentivo para seguir las condiciones establecidas por el "anchor". Su objetivo principal sería evitar la clasificación de riesgo 'malo' y buscar una evaluación más favorable.

En contraste, para las observaciones 485 y 801, las cuales fueron clasificadas como de riesgo 'bueno' por el modelo, es más probable que el solicitante esté interesado en seguir las condiciones establecidas por el "anchor". Esto se debe a que cumplir con estas condiciones puede proporcionar cierta garantía de mantener la clasificación de riesgo 'bueno' y, por lo tanto, obtener la aprobación del préstamo o incluso condiciones más favorables. En ambos casos, se observa que la duración del préstamo es un factor determinante para mantener la clasificación de riesgo 'bueno'. Para la observación 485 (Ver Tabla 4.7b), el "anchor" esta-

Anchor
Purpose_car >-0.71
Age <= -0.75
Saving accounts_no_inf <= -0.47
Housing_own >-1.58
Saving accounts_rich <= -0.22
Purpose_education <= -0.25
Purpose_furniture/equipment <= -0.47
Purpose_repairs <= -0.15
Saving accounts_moderate <= -0.34
Job <= 0.15
Housing_rent <= -0.47
Purpose_domestic appliances <= -0.11
Saving accounts_quite rich <= -0.26
Purpose_vacation/others <= -0.11
Precision: 0.79
Coverage: 0.02

(a) Observación 313

Anchor
Duration <= -0.74
Precision: 0.82
Coverage: 0.37

(b) Observación 485

Anchor
Duration <= -0.24
Purpose_radio/TV >-0.62
Precision: 0.85
Coverage: 0.16

(c) Observación 801

Tabla 4.7: Anchors generadas para las observaciones objeto de estudio
Elaboración Propia

blece una regla que indica que la duración del préstamo debe ser inferior a un valor escalado de -0.74, es decir, menos de 12 meses. Esta regla se aplica al 37 % de las observaciones en el conjunto de datos de entrenamiento, con una precisión del 82 %. En el caso de la observación 801 (Ver Tabla 4.7c), el "anchor" también establece un límite para la duración del crédito, aunque es menos estricto. En este caso, se requiere que la duración del préstamo sea inferior a un valor escalado de -0.24, es decir, menos de 18 meses. Además, se agrega la condición de que el propósito del crédito sea la compra de un televisor o una radio. Al incluir esta nueva condición, la cobertura se reduce al 16 %, pero se obtiene una precisión del 85 %.

4.3.6. Resumen de resultados

Después de llevar a cabo la aplicación de cinco técnicas de interpretabilidad local sobre un conjunto de observaciones, se procede a evaluar los resultados obtenidos. Es importante destacar que el enfoque adoptado se centra en la interpretación a nivel local, en lugar de buscar una visión global de la interpretabilidad. Para evaluar estos resultados y la importancia de las diferentes características, se ha considerado la frecuencia en la que estas han aparecido como variables relevantes y la coherencia que existe en los resultados con el contexto financiero. Este análisis proporcionará una comprensión más completa y confiable de cómo estas características influyen en la clasificación del riesgo crediticio.

En primer lugar, se ha observado que para la observación 313 la variable Edad ha desempeñado un papel determinante. Todas las técnicas establecen que la edad joven del solicitante ha sido uno de los factores principales que ha contribuido a la clasificación negativa del crédito. Esto tiene sentido financiero, ya que una menor edad suele conllevar menos experiencia laboral y por tanto menor sueldo, también suele conllevar a una menor capacidad a la hora de gestionar adecuadamente los asuntos financieros. Además, es probable que para los solicitantes más jóvenes se disponga de menos información detallada sobre su historial crediticio o incluso no tengan historial crediticio, lo cual también puede influir en la evaluación del riesgo crediticio. En la observación 801 se aprecia igualmente la relevancia de la edad sin embargo en este caso todas las técnicas, excepto los anchors indican la edad del solicitante ha afectado positivamente a la clasificación del riesgo. Además, se ha identificado otra variable relevante: la cantidad solicitada. En el caso de la observación 313, se observa que ha solicitado un valor por debajo de la media, lo cual parece afectar negativamente a la calificación del crédito, según las técnicas de LIME y SHAP. Además, al realizar una explicación contrafáctica, se ha encontrado que aumentar la cantidad solicitada tiene un efecto positivo en la clasificación del crédito. Esto sugiere que solicitar una cantidad más alta podría influir de manera favorable en la evaluación del riesgo crediticio para esta observación en particular.

Por otro lado, para la observación 485 observamos que tanto LIME como SHAP indican que la duración del préstamo ha afectado positivamente a que el crédito se clasificara como bueno. Además la regla generada por los anchors indica que mientras el préstamo se man-

tenga por debajo de 12 meses, el crédito probablemente sea clasificado como bueno. Esto tiene sentido financiero ya que cuanto menor sea el plazo, hay menos probabilidad de que se produzcan contratiempos y por tanto conlleva menos riesgo.

Otra variable la cual LIME, para las tres observaciones, revela su importancia es el hecho de que ninguno tenga una gran cantidad de dinero en la cuenta de ahorros ("Savings_rich"), que afecta negativamente a la clasificación del riesgo. Esto tiene sentido financiero ya que tener una gran cantidad de dinero en la cuenta de ahorros puede considerarse como un indicador de estabilidad financiera y capacidad para afrontar posibles riesgos.

Estos resultados son de gran utilidad para comprender el modelo y brindan herramientas que promueven la toma de decisiones transparente. Sin embargo, para otras muchas ocasiones y variables no se ha alcanzado un consenso entre las diferentes técnicas de interpretabilidad utilizadas, es decir, ninguna de las variables aparece de manera clara como la más importante. Por ejemplo, las explicaciones contrafácticas indican que de no proveer información sobre la cuenta de ahorros el crédito probablemente se rechazara. Pero LIME y SHAP indican que el hecho de que si que haya información sobre la cuenta de ahorros ha afectado negativamente. Otro ejemplo es la variable Cantidad solicitada donde encontramos explicaciones contradictorias para la observación 313, LIME indica que el valor actual de esta observación ha afectado negativamente a la clasificación mientras que SHAP indica lo contrario, que ha afectado positivamente y las explicaciones contrafácticas establecen que solicitar una cantidad mucho más elevada también podría favorecer a la clasificación del riesgo.

Capítulo 5

Conclusiones y trabajo futuro

La interpretación y comprensión de los modelos de Machine Learning (ML) se han vuelto cada vez más relevantes en diversos campos, y el contexto del riesgo financiero no es una excepción. La motivación principal de este trabajo ha sido abordar la interpretabilidad local de los modelos de ML en el ámbito del riesgo financiero. Se busca promover la transparencia en las decisiones tomadas por los modelos, garantizar una toma de decisiones ética y cumplir con los crecientes requisitos regulatorios en el sector financiero. Los objetivos de este trabajo se han enfocado en aplicar las técnicas de interpretabilidad local más relevantes y populares en un modelo de ML específico para clasificar solicitudes de préstamos como "buenas" o "malas". Para ello, se ha utilizado una variedad de técnicas, incluyendo LIME, SHAP, explicaciones contrafácticas, anclas y gráficos de ICE. Estas técnicas generan conjuntos de explicaciones que ayudan a comprender el proceso de toma de decisiones del modelo y proporcionan información valiosa sobre las variables más importantes que influyen en la clasificación de las solicitudes de préstamos.

Para llevar a cabo el análisis, se seleccionó un conjunto de datos asociado a solicitudes de préstamo en una entidad bancaria. A continuación, se modeló una red neuronal con el propósito de abordar el problema de clasificación binaria, permitiendo predecir qué préstamos se pagarán sin problemas y, por lo tanto, deben ser concedidos, y cuáles podrían enfrentar dificultades y requerir una evaluación más exhaustiva. Una vez entrenada la red neuronal, se implementaron técnicas de interpretabilidad local sobre tres observaciones específicas. Posteriormente, se analizaron los resultados considerando la frecuencia y coherencia de las variables relevantes presentes en las explicaciones generadas por cada técnica. Además, se tuvo en cuenta el contexto financiero para validar la relevancia de las variables identificadas en el análisis.

La aplicación de técnicas de interpretabilidad local ha sido fundamental en este estudio, ya que ha permitido generar explicaciones comprensibles y específicas para cada observación analizada. En la mayoría de los casos, estas técnicas han logrado proporcionar resultados relevantes en el contexto financiero. Entre los resultados más destacados se encuentra la im-

portancia de la edad en la calificación crediticia, donde se ha observado que las personas más jóvenes tienden a tener una posición más desfavorable en este aspecto. Además, se ha evidenciado la importancia de la duración del crédito, donde se ha encontrado que los plazos más cortos son los más deseables para minimizar el riesgo. Sin embargo, es importante mencionar que se ha identificado una limitación en el análisis realizado: no todas las explicaciones generadas por estas técnicas concuerdan entre sí. Esto sugiere que existen múltiples factores y complejidades que influyen en la evaluación del riesgo crediticio, y que algunas variables pueden tener un impacto diferente según el contexto.

Este resultado resalta la necesidad de una evaluación crítica de las explicaciones generadas por diferentes técnicas de interpretabilidad local. No todas las técnicas proporcionan la misma perspectiva o énfasis en las variables relevantes, lo que puede generar discrepancias en la interpretación de los resultados. Esto no implica que la interpretación de los resultados por estas técnicas sea incorrecta, sino que el modelo es tan complejo que pueden ocurrir interacciones entre variables que no se pueden capturar de manera simplificada. Esta falta de consenso entre las explicaciones resalta la complejidad de la interpretabilidad de los modelos de ML y destaca la importancia de considerar múltiples enfoques y técnicas en futuras investigaciones, así como la necesidad de mantener una actitud crítica coherente con el contexto. A pesar de esta limitación, los resultados obtenidos a través de las técnicas de interpretabilidad local ofrecen un análisis exploratorio sobre los determinantes clave de la calificación crediticia. Estos resultados podrían servir como punto de partida para las entidades financieras y bancarias, ya que les permitiría tomar decisiones más informadas en la evaluación de solicitudes de préstamos,

Durante el desarrollo de este estudio, se han identificado limitaciones adicionales que pueden dificultar el proceso de interpretabilidad de los modelos. En primer lugar, se encontró que el tratamiento de variables categóricas, como las variables ‘Purpose’ (propósito del préstamo) y ‘Savings account’ (cuenta de ahorros), ha incrementado la dimensionalidad del conjunto de datos. Esta mayor dimensionalidad presenta desafíos para comprender la estructura de las variables y las relaciones entre ellas, lo que a su vez agrega complejidad al modelo. La presencia de variables categóricas requiere enfoques adicionales, como la codificación adecuada de las categorías o el uso de técnicas específicas de manejo de variables categóricas, para garantizar una interpretabilidad efectiva. En segundo lugar, se observó que si los datos utilizados en el modelo están escalados, las explicaciones generadas también se presentarán de manera escalada. Esto significa que la interpretación de las contribuciones de las variables puede volverse más complicada, ya que se pierde la escala original de los datos. La interpretación de las relaciones y la importancia de las variables se ve influenciada por este escalamiento, lo que puede dificultar la comprensión de los resultados y la toma de decisiones basada en las explicaciones, si no se tiene en cuenta esta salvedad.

Por último, es importante destacar que algunas de las técnicas presentadas en este estudio pueden presentar dificultades cuando se enfrentan a altas correlaciones entre variables,

como se ha mencionado anteriormente para ICE, LIME y las explicaciones contrafácticas, ya que no alcanzan a modelar la presencia de interacciones. Sin embargo, sería valioso en trabajos futuros abordar el desarrollo de modelos más sofisticados capaces de identificar y capturar adecuadamente estas interacciones entre variables en casos donde las correlaciones sean particularmente significativas.

Referencias

Angelov, P. P., Soares, E. A., Jiang, R., Arnold, N. I., y Atkinson, P. M. (2021). Explainable artificial intelligence: an analytical review. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *11*(5), e1424.

Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., … others (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information fusion*, *58*, 82–115.

Bloch, L., y Friedrich, C. M. (2021). Data analysis with Shapley values for automatic subject selection in Alzheimer's disease data sets using interpretable machine learning. *Alzheimer's Research & Therapy*, *13*(1), 1–30.

Byrne, R. M. (2016). Counterfactual thought. *Annual review of psychology*, *67*, 135–157.

Chen, C., Lin, K., Rudin, C., Shaposhnik, Y., Wang, S., y Wang, T. (2022). A holistic approach to interpretability in financial lending: Models, visualizations, and summary-explanations. *Decision Support Systems*, *152*, 113647.

Consumer Financial Protection Bureau. (2018). *Comment for 1002.9 - Notifications.* (acceso Enero 15, 2023) `https://www.consumerfinance.gov/rules-policy/regulations/1002/interp-9/`.

Cornacchia, G., Narducci, F., y Ragone, A. (2021). Improving the user experience and the trustworthiness of financial services. En *Human-computer interaction–interact 2021: 18th ifip tc 13 international conference, bari, italy, august 30–september 3, 2021, proceedings, part v 18* (pp. 264–269).

Demajo, L. M., Vella, V., y Dingli, A. (2020). Explainable ai for interpretable credit scoring. *arXiv preprint arXiv:2012.03749*.

Dieterle, F. J. (2003). *Multianalyte quantifications by means of integration of artificial neural networks, genetic algorithms and chemometrics for time-resolved analytical data* (Tesis Doctoral no publicada). Universität Tübingen.

European Parlament and the Council. (2016). *REGULATION (EU) 2016/679 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation).* (acceso Enero 15, 2023) `https://eur-lex.europa.eu/eli/reg/2016/`

679/oj".

Farzad, T. (2019). *Determinants of mortgage loan delinquency: Application of interpretable machine learning.*

Fernández, J. A. F. (2020). United states banking stability: an explanation through machine learning. *Banks and Bank Systems*, *15*(4), 137.

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of statistics*, 1189–1232.

Gill, N., Hall, P., Montgomery, K., y Schmidt, N. (2020). A responsible machine learning workflow with focus on interpretable models, post-hoc explanation, and discrimination testing. *Information*, *11*(3), 137.

Goldstein, A., Kapelner, A., Bleich, J., y Pitkin, E. (2015). Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation. *journal of Computational and Graphical Statistics*, *24*(1), 44–65.

Gurney, K. (1997). *An introduction to neural networks*. CRC press.

Ji, Y. (2021). *Explainable ai methods for credit card fraud detection: Evaluation of lime and shap through a user study.*

Jiang, C., Wang, Z., y Zhao, H. (2019). A prediction-driven mixture cure model and its application in credit scoring. *European Journal of Operational Research*, *277*(1), 20–31.

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

Mayer, M., Meier, D., y Wuthrich, M. V. (2023). Shap for actuaries: Explain any model. *Available at SSRN*.

Mi, J.-X., Li, A.-D., y Zhou, L.-F. (2020). Review study of interpretation methods for future interpretable machine learning. *IEEE Access*, *8*, 191969–191985.

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

Misheva, B. H., Osterrieder, J., Hirsa, A., Kulkarni, O., y Lin, S. F. (2021). Explainable AI in credit risk management. *arXiv preprint arXiv:2103.00949*.

Molnar, C. (2020). *Interpretable machine learning*. Lulu. com.

OECD. (2019). *Recommendation of the Council on Artificial Intelligence.* (acceso Enero 15, 2023) https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., … others (2011). Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, *12*, 2825–2830.

Ribeiro, M. T., Singh, S., y Guestrin, C. (2016). "Why should I trust you?" Explaining the predictions of any classifier. En *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (pp. 1135–1144).

Ribeiro, M. T., Singh, S., y Guestrin, C. (2018). Anchors: High-precision model-agnostic explanations. En *Proceedings of the aaai conference on artificial intelligence* (Vol. 32).

Srushti Dhamangaonkar. (2020). *ML: Model Interpretability Methods PDP, ICE ELI5, LIME, SHAP implementation on Tabular Dataset.* (acceso Mayo 10, 2023) `https://dhamangaonkar-s.medium.com/ml-model-interpretability-methods-7c2fc02f51b6`.

Taly, Ankur and Shanbhag, Aalok. (2020). *'Counterfactual Explanations vs. Attribution Based Explanations.* (acceso Junio 10, 2022) `https://www.fiddler.ai/blog/counterfactual-explainable-vs-attribution-based-explanations`.

The United States Department of Justice. (2022). *The Equal Credit Opportunity Act.* (acceso Enero 14, 2023) `https://www.justice.gov/crt/equal-credit-opportunity-act-3`.

Wai On. (2020). *Model-Agnostic Local Explanations using Individual Conditional Expectation (ICE) Plots.* (acceso Mayo 21, 2023) `https://towardsdatascience.com/how-to-explain-and-affect-individual-decisions-with-ice-curves-1-2-f39fd751546f`.

Yu, F., Wei, C., Deng, P., Peng, T., y Hu, X. (2021). Deep exploration of random forest model boosts the interpretability of machine learning studies of complicated immune responses and lung burden of nanoparticles. *Science Advances*, *7*(22), eabf4130.

Yuan, B., Shen, C., Luna, A., Korkut, A., Marks, D. S., Ingraham, J., y Sander, C. (2021). Cellbox: interpretable machine learning for perturbation biology with application to the design of cancer combination therapy. *Cell systems*, *12*(2), 128–140.

Zhang, C. A., Cho, S., y Vasarhelyi, M. (2022). Explainable artificial intelligence (xai) in auditing. *International Journal of Accounting Information Systems*, *46*, 100572.

Zhang, Z., Wu, C., Qu, S., y Chen, X. (2022). An explainable artificial intelligence approach for financial distress prediction. *Information Processing & Management*, *59*(4), 102988.