IA Explicable en administración de riesgo de crédito

Presentado para obtener el título de

Maestría en Matemáticas Aplicadas y Ciencias de la Computación

Cristhian Camilo Zamora Mahecha

Dirección:

Oscar Samuel Fernández Barreto

Universidad del Rosario

Escuela de Ingeniería, Ciencia y Tecnología

Maestría en Matemáticas Aplicadas y Ciencias de la Computación

ABSTRACT

Español

La adopción de métodos de inteligencia artificial (IA) en el sector financiero puede conducir a mejoras significativas en temas de experiencia de cliente, bancarización de poblaciones remotas, lucha contra el lavado de capitales, administración del riesgo de crédito, entre otros. Particularmente en el campo de riego de crédito, el principal objetivo es estimar probabilidades de incumplimiento lo más cercanas al incumplimiento observado en la realidad, este objetivo puede alcanzarse mediante la aplicación de algoritmos nuevos y potentes que logran mejorar las medidas de precisión con respecto a métodos más tradicionales. Sin embargo, estos algoritmos pierden transparencia y explicabilidad, por lo que generalmente son denominados como "cajas negras", lo que significa que se conocen las entradas y salidas del algoritmo, pero es difícil de entender y explicar lo que hace el algoritmo en su interior. Dicha falta de inteligibilidad de los métodos es contraria a los requerimientos de los reguladores financieros, llevando a que exista un rezago en el campo respecto al estado del arte en IA.

El propósito de este proyecto es motivar la adopción de métodos de IA en el campo de riego de crédito, mediante la aplicación de un modelo de XGBoost a un conjunto de datos, y la aplicación de un conjunto de metodologías que incluyen la aplicación de Shapeley Values, expectativa condicional individual, diagramas de dependencia parcial, extracción de reglas, entre otras. Estas metodologías se enmarcan en un esquema de preguntas correctamente formuladas que permiten explicar el funcionamiento del modelo a las partes interesadas.

English

The adoption of artificial intelligence (AI) methods in the financial sector can lead to significant improvements in customer experience, banking of remote populations, the fight against money laundering, and credit risk management, among others. Particularly in the field of credit risk, the main objective is to estimate probabilities of default as close

to the default observed in reality, this objective can be achieved through the application of new and powerful algorithms that manage to improve the precision measures with respect to traditional methods. However, these algorithms lose transparency and explainability, so they are generally referred to as “black boxes”, which means that the inputs and outputs of the algorithm are known, but it is difficult to understand and explain what the algorithm does inside. The lack of intelligibility of the methods is contrary to the requirements of financial regulators, leading to a lag in the field concerning the state of the art in AI.

The purpose of this project is to motivate the adoption of AI methods in the field of credit risk, by applying an XGBoost model to a data set, and applying a set of methodologies that include the application of Shapeley Values, individual conditional expectation, partial dependency plots, rule extraction, among others. These methodologies are framed in a scheme of correctly formulated questions that allow explaining the operation of the model to the stakeholders.

TABLA DE CONTENIDO

LISTA DE TABLAS

LISTA DE ILUSTRACIONES

1 JUSTIFICACIÓN Y DESCRIPCIÓN DEL PROBLEMA

A medida que la Inteligencia Artificial (IA) se posiciona como principal propulsor de la cuarta revolución industrial, se especializa para solucionar problemas específicos como la conducción autónoma, la localización de recursos estatales, el diagnóstico médico, la administración de portafolios de inversión, o el desarrollo de modelos de riesgo crediticio. Es previsible que en el futuro cercano gran parte de las decisiones de nuestra vida estén mediadas por una IA especializada. Estas IA, comúnmente se basan en modelos complejos de aprendizaje automático que son muy efectivos para proporcionar un alto rendimiento en la tarea especifica en la que se entrenan. Sin embargo, estas técnicas nuevas y avanzadas carecen de la transparencia necesaria para comprender cómo y por qué se obtiene una determinada salida. Es decir, que en tanto estos modelos son muy buenos realizando la tarea en la que se especializan es muy difícil entender los pasos que el algoritmo tomó para llegar a su decisión.

En este contexto, naturalmente, la confianza que se tiene sobre las predicciones y el modelo en sí mismo es clave. ¿Podemos confiar en los algoritmos que conducirán nuestros automóviles, tomarán decisiones médicas y financieras por nosotros? La confianza es un ingrediente clave requerido para la adopción de IA a gran escala [1]. De acuerdo con Ribeiro [2], es importante diferenciar entre dos definiciones distintas pero relacionadas de confianza: (1) confiar en una predicción, es decir, si un usuario confía lo suficiente en una predicción individual para tomar alguna acción basada en ella, y (2) confiar en un modelo, es decir, si el usuario confía en que un modelo se comportará de manera razonable si se implementa. Comprender las razones detrás de las predicciones es bastante importante para evaluar la confianza, lo cual es fundamental si uno planea tomar medidas basadas en una predicción, o al elegir implementar un nuevo modelo.

Por otra parte, el incremento de la difusión de estas IA especializadas no ha pasado desapercibido por legisladores y autoridades regulatorias, la transparencia en la forma de un “derecho a una explicación” se ha convertido en una opción convincentemente atractiva, ya que intuitivamente promete abrir la “caja negra” algorítmica [3]. Por

ejemplo, para los países miembros de la Unión Europea el Reglamento General de Protección de Datos (GDPR por sus siglas en inglés) especifica que cuando una persona es afectada por decisión basada únicamente en un tratamiento automatizado de datos, se garantiza el derecho a una explicación significativa sobre la lógica subyacente a decisión tomada. Este tipo de normatividad es particularmente evidente en sectores económicos altamente regulados, como la salud y las finanzas, por ejemplo, un informe de la Autoridad Federal de Supervisión Financiera de Alemania (BaFin por sus siglas en inglés) establece que "No hay excusas de caja negra: es necesaria la explicabilidad y trazabilidad de los modelos" [4].

Con este fin, surgió el concepto de inteligencia artificial explicable (XAI). XAI es un área emergente de investigación en el campo de la Inteligencia Artificial (IA). XAI puede explicar cómo la IA obtuvo una solución particular (p. ej., clasificación o detección de objetos) y también puede responder otras preguntas relevantes para entender el funcionamiento de los modelos de IA [5]. A pesar de la prevalencia de la investigación sobre la explicabilidad, las definiciones exactas en torno a la IA explicable aún no están consolidadas [6], para poner solamente un par de ejemplos, para Miller [7] esta explicabilidad se puede definir como el grado en que un ser humano puede comprender la causa de una decisión tomada por una IA especializada, mientras para Turri [6] XAI se define como conjunto de procesos y métodos que permite a los usuarios humanos comprender y confiar en los resultados y resultados creados por los algoritmos de aprendizaje automático. Por su parte Misheva [8], parece tomar elementos de las dos definiciones anteriores proponiendo que XAI introduce un conjunto de técnicas de aprendizaje automático (ML) que buscan producir modelos que ofrezcan un balance aceptable entre explicabilidad y potencia predictiva, y permitan a los humanos comprender, confiar y gestionar modelos complejos de IA .

Específicamente en el sector financiero, en el campo de la administración de riesgo de crédito, el elemento más importante es la estimación de una probabilidad de incumplimiento (PI) lo más cercana posible al incumplimiento observado. Son varios los artículos académicos que aplican algoritmos nuevos y potentes a la estimación de la PI, tales como redes neuronales [9], gradient boosting y otros métodos de aprendizaje

conjunto [10], por mencionar solamente un par de ejemplos. Estos métodos de aprendizaje automático, considerados de caja negra, han demostrado ser capaces de obtener métricas de rendimiento y ajuste más destacadas que los modelos tradicionales, sin embargo, su adopción en la administración de riesgo de crédito ha sido un reto. Lo anterior se explica en gran medida dado que carecen de transparencia algorítmica y explicabilidad, dos elementos de vital importancia en este campo en donde la confianza en las predicciones y en el modelo en general, además las regulaciones gubernamentales exigen contar con estos factores.

Para superar este rezago, desde la academia ha examinado los métodos existentes y propuesto nuevos caminos que permitan avanzar en una mayor adopción de metodologías recientes en el área de riesgo de crédito. En este sentido [11] comparara los marcos SHAP y LIME evaluando su capacidad para definir distintos grupos de observaciones. Por su parte, [12] aplica redes de correlación a los valores de Shapley para que las predicciones de la Inteligencia Artificial se agrupen según la similitud en las explicaciones subyacentes. Giudici [13] propone un nuevo un método de IA explicable global que se basa en descomposiciones de Lorenz, extendiendo así las contribuciones anteriores basadas en descomposiciones de varianza lo que permite que la descomposición de Shapley-Lorenz resultante sea de aplicación más general y proporciona un criterio de importancia variable único. A pesar de estos métodos y avances, además de muchos otros, no existe un esquema de trabajo que permita responder preguntas correctamente formuladas que conduzcan a explicar el funcionamiento del modelo a las partes interesadas a partir de la aplicación de un conjunto de metodologías disponibles en el campo de XAI.

2 OBJETIVOS

2.1 Objetivo general

Motivar la adopción de métodos de IA en el campo de la administración de riesgo de crédito, demostrando como la aplicación de metodologías de aprendizaje automático explicado puede responder preguntas relevantes, acerca del modelo y sus predicciones, para que las partes implicadas ganen confianza e inteligibilidad sobre modelos denominados de caja negra.

2.2 Objetivos específicos

1. Construir un modelo XGBoost sobre un conjunto de datos de riesgo de crédito.
2. Demostrar como la aplicación de metodologías de aprendizaje automático explicado (SHAP, LIME, PDP, ICE, etc.) puede explicar diversos aspectos de modelo opacos.
3. Proponer un marco de trabajo basado en un conjunto de preguntas relevantes relacionadas con el método apropiado para responder a cada pregunta.

3 MARCO TEÓRICO Y ESTADO DEL ARTE

3.1 Criterios de Transparencia

En algunos casos, no importa por qué se tomó una decisión, basta con saber que el rendimiento predictivo en un conjunto de datos de prueba fue bueno. Pero en otros casos, saber el "por qué" puede ayudar a aprender más sobre el problema, los datos y la razón por la cual un modelo podría fallar [14]. Lo anterior depende del nivel de riesgo inherente al objetivo con el que se usa el modelo, por ejemplo, el nivel de riesgo de un sistema antispam es significativamente menor que el riesgo de un modelo cuyo objetivo es otorgar o negar solicitudes de crédito. Como ya se vio en el planteamiento del problema, dados dichos niveles de riesgo, en muchas áreas de aplicación industrial es necesario comprender la razón por la cual se llega a una predicción para generar comprensibilidad sobre el modelo y confianza sobre su funcionamiento.

La comprensibilidad que el ser humano tiene sobre el método de modelado depende en gran medida de la su transparencia. Una IA especializada o modelo de aprendizaje automático se puede considerar transparente si para cualquier entrada dada, no solo proporciona cadenas de razonamiento como funciones lineales, gráficos, listas de reglas, sino que también proporciona la justificación de las decisiones en términos de precisión, consistencia, confiabilidad y seguridad [15].

Según Lipton [16] es posible evaluar la transparencia tomando en cuenta tres criterios:

- Simulabilidad: Este criterio hace referencia a si es posible para el ser humano replicar el modelo, es decir, si teniendo los datos de entrada y los parámetros del modelo, un ser humano podría ser capaz de aplicar los cálculos necesarios para obtener una predicción. La cantidad de tiempo requerida para lograr una predicción debe ser razonable, es decir que un modelo compuesto de una gran cantidad de operaciones simples puede considerarse no simulable por el ser humano.
- Descomponibilidad: Este criterio considera si el modelo es susceptible de ser dividido en partes (Entradas, parámetros, y cálculos), y que cada una de estas partes pueden ser explicadas de forma intuitiva. Cabe destacar que aquellas variables de

entrada obtenidas mediante ingeniería de características o técnicas de reducción de dimensionalidad y que no puedan ser explicadas de forma intuitiva no cumplen con este criterio.

- Transparencia algorítmica: El ultimo criterio de transparencia se relaciona con el algoritmo de aprendizaje en sí mismo. El único requisito para cumplir con este criterio es que el usuario pueda probar el modelo mediante análisis matemático. Por ejemplo, en el caso de un árbol de decisión poco profundo, un ser humano puede ser capaz de seguir los pasos y cálculos requeridos para probar que un nodo está tomando el punto de corte correcto. Este tipo de validaciones analíticas no son aplicables a métodos como las redes neuronales.

Aun teniendo en cuenta los anteriores criterios, la clasificación de un modelo entre transparente u opaco puede llegar a no ser trivial. Por ejemplo, el método k vecinos más cercanos (KNN por sus siglas en ingles), cumple con el criterio de transparencia algorítmica en el sentido que se pude probar matemáticamente, además se puede considerar que dicho modelo cumple con descomponibilidad ya que es posible explicar sus partes de forma independiente e intuitiva. Sin embargo, este método requiere calcular gran cantidad de distancias por lo que puede decirse que no cumple con el criterio de simulabilidad. En la práctica, a pesar de los matices, es convencional ver los árboles de decisión, la regresión lineal, entre otros, como modelos más simples y transparentes, y los bosques aleatorios, el aprendizaje profundo, entre otros, como modelos opacos [17].

3.2 Tipos de explicaciones

De acuerdo con los criterios de transparencia expuestos en la sección anterior es posible determinar el nivel de transparencia de un sistema de IA especializado. Para aquellos sistemas de IA basados en modelos de aprendizaje automático considerados como opacos XAI ofrece un conjunto de métodos post-hoc que buscan explicar aspectos importantes en su funcionamiento. Es importante resaltar que, aunque estas técnicas están enfocadas en modelos opacos, su aplicación no se limita únicamente a estos, es decir, es posible aplicar estas técnicas sobre modelos transparentes. Según Arrieta [18], se podría

considerar que el conjunto de técnicas ofrecidas por XAI se pueden agrupar en los siguientes tipos:

- Explicaciones visuales: Tienen como objetivo generar visualizaciones que logren explicar funcionamiento de algún aspecto de interés del modelo, por ejemplo, el límite de decisión o la interacción de un conjunto de variables con la predicción realizada.
- Explicaciones locales: Intentan explicar el funcionamiento del modelo en un determinado subespacio de decisión. Esto significa que las explicaciones obtenidas no necesariamente pueden generalizarse para crear un orden global que represente el comportamiento total del modelo. En su lugar, normalmente aproximan el funcionamiento del modelo sobre una única observación que el usuario desea describir, para crear una descripción de cómo debe comportarse el modelo cuando se enfrenta observaciones similares.
- Explicaciones de importancia de variables buscan explicar la salida de un modelo cuantificando la influencia de cada variable de entrada sobre dicha salida. Esto da como resultado un valor discreto, puntaje de importancia, que determina el impacto (sensibilidad) que tiene cada una de las variables sobre la salida del modelo.
- Explicaciones por simplificación se refieren a las técnicas que aproximan un modelo opaco a otro más simple, más fácil de interpretar. El principal desafío proviene del hecho de que el modelo simple tiene que ser lo suficientemente flexible para que pueda aproximarse con precisión al modelo complejo. En la mayoría de los casos, esto se mide comparando la precisión (para problemas de clasificación) de estos dos modelos [17].

Dado el constante desarrollo de técnicas novedosas de XAI y la diversidad de estas, no siempre es sencillo situar una técnica en una o varias categorías de las expuestas. Adicional a lo anterior, también hay que considerar una distinción más, y es que algunas técnicas solamente son aplicables a un tipo modelo en específico (i. e. InTrees solamente puede aplicarse a modelos basados en arboles), mientras que otras son agnósticas al tipo de modelo sobre el que se aplican.

3.2.1 Explicaciones Visuales

3.2.1.1 Gráfica de Dependencia Parcial

El gráfico de dependencia parcial visualiza el efecto marginal que tienen una o dos variables sobre el resultado esperado de un modelo de aprendizaje automático [19]. Este método apunta a representar la frontera de decisión del estimador en función de una o dos variables de interés mientras se promedian las variables restantes, es decir, busca explicar el efecto marginal promedio de las variables de interés sobre la salida del modelo. Un gráfico de dependencia parcial puede mostrar si la relación entre el objetivo y una variable es lineal, monótona o más compleja [14]. Este método nos ayuda a comprender cómo los diferentes valores de una variable impactan las predicciones del modelo, sin embargo, dadas las restricciones de la percepción humana y de la visualización en más de tres dimensiones, la cantidad de variables de interés generalmente se limita a una o dos, para identificar esas variables es recomendable aplicar este método junto con otros que permitan medir la relevancia global de las variables en el modelo.

3.2.1.2 Expectativa Condicional Individual

Funciona a nivel de observación, es decir este es un método de interpretación local, expresando los límites de decisión del modelo en términos de una variable y manteniendo el resto constante. Los gráficos ICE refinan el gráfico de dependencia parcial al representar la relación funcional entre la respuesta predicha y una variable a nivel de observaciones individuales [20]. Así, la influencia de esta variable puede estudiarse en las decisiones del modelo en un contexto específico determinado por el resto de las variables. Las curvas de expectativa condicional individual son más intuitivas de entender que las gráficas de dependencia parcial. Una línea representa las predicciones para una instancia si variamos la característica de interés [14].

3.2.2 Explicaciones locales

3.2.2.1 Explicaciones agnósticas del modelo localmente interpretables (LIME)

El objetivo general de LIME es identificar un modelo transparente sobre la representación interpretable que sea localmente fiel al estimador original [2], en otras

palabras, el método busca representar un modelo de aprendizaje automático de caja negra con un modelo interno transparente para explicar cada predicción, es decir este es un método de interpretación local. LIME trata de ajustar un modelo local utilizando puntos de datos de muestra que son similares a la observación a explicar. El modelo local puede ser de la clase de modelos interpretables como modelos lineales, árboles de decisión, etc. [11].

3.2.2.2 SHapley Additive exPlanations (SHAP)

El objetivo de SHAP es explicar la predicción de una observación calculando la contribución marginal de cada variable a la predicción. El método de explicación SHAP calcula los valores de Shapley a partir de la teoría de juegos de coalición [14]. Desde una perspectiva computacional, SHAP (abreviatura de SHpley Additive exPlanation) devuelve valores de Shapley que expresan las predicciones del modelo como combinaciones lineales de variables binarias que describen si cada covariable está presente o no en el modelo [11]. El valor SHAP para una observación se debe interpretar como la contribución marginal de una variable a la diferencia entre la predicción realizada y la predicción media. La importancia de la variable a nivel global se da sumando el valor absoluto de los valores SHAP para cada punto de datos individual [21].

3.2.3 Explicaciones de importancia de variables

3.2.3.1 Importancia por permutación de variables.

La idea subyacente a este método es bastante intuitiva, se trata de monitorear la medida de rendimiento del modelo a medida que los valores de una variable se van volviendo progresivamente aleatorios, de tal manera que aquellas variables con mayor impacto sobre la medida de rendimiento serán las más importantes [22]. Este enfoque tiene la ventaja de ser altamente interpretable ya que ya que la importancia de cada variable se representa en términos del deterioro de la métrica de rendimiento del modelo.

3.2.3.2 Análisis contrafactual

El enfoque de este método consiste en tomar una observación con su clasificación predicha y simular una nueva observación hipotética en la cual el valor de una variable se

modifica hasta que la clasificación predicha cambie de categoría [23]. Adicional al nuevo punto de datos contrafactual, es posible evaluar en que magnitud cambio la observación, de acuerdo con una métrica de similitud que puede ser defina de acuerdo con criterio experto, lo que sirve como base para estimar la importancia de la característica modificada.

3.2.4 Explicaciones por simplificación

3.2.4.1 Reglas de ámbito (anclajes)

La técnica de anclajes tiene como objetivo hallar un conjunto de reglas "Si-Entonces" que logre anclar la predicción del modelo a nivel de observación, se dice que un conjunto de reglas ancla la decisión del modelo si ante cambios en los valores de las variables que no están contenidas en la regla no se logra cambiar la predicción. Los anclajes resaltan la parte de la entrada que es suficiente para que el clasificador haga la predicción, haciéndolos intuitivos y fáciles de entender [24]. La técnica de anclajes se apoya en técnicas de aprendizaje por refuerzo en combinación con un algoritmo de búsqueda en grafos para lograr optimizar la cantidad de llamados al modelo que el método debe hacer.

3.2.4.2 Sustituto global

Lo que busca este enfoque es emular con la mayor fidelidad posible la función de predicción de caja negra f con la función de predicción de modelo sustituto g, bajo la restricción de que g es interpretable [14]. Es decir, el modelo sustituto se entrena tomando como etiqueta de los datos las predicciones realizadas por el modelo de caja negra. Esta técnica puede ser aplicada usando cualquier tipo de modelo sustituto siempre y cuando este cumpla con los criterios de transparencia mencionados en la sección 3.1. Es muy importante destacar que este método es útil para lograr entender mejor el modelo de caja negra, ya que en esta técnica el modelo sustituto nunca tiene como entrada las etiquetas reales de los datos este no debe ser usado para obtener predicciones.

4 PROPUESTA METODOLOGICA

4.1 Flujo de trabajo

4.1.1 Recopilación de los datos

Después de una búsqueda de un conjunto de datos de uso público en la que se consideraron una docena de posibles opciones, se decidió usar el conjunto de datos de proporcionado por American Express para su competencia de predicción de default alojada en la plataforma Kaggle. El sitio describe la competencia como:

"El objetivo de esta competencia es predecir la probabilidad de que un cliente no pague el monto del saldo de su tarjeta de crédito en el futuro en función de su perfil de cliente mensual. La variable binaria objetivo se calcula observando la ventana de rendimiento de 18 meses después del último extracto de la tarjeta de crédito, y si el cliente no paga el monto adeudado 120 días después de la fecha del último extracto, se considera un evento de incumplimiento." [25].

El conjunto de datos tiene un total de 191 variables relacionadas con el comportamiento de los clientes en aspectos clave como morosidad, pago, balance y riesgo. Se considera que las dimensiones de los datos, la cantidad de variables disponibles, la recencia de la información y la pertinencia de esta misma hacen que este conjunto de datos sea ideal para el proyecto.

4.1.2 Preprocesamiento de los datos

En este punto se realizará un análisis de valores nulos y del tratamiento adecuado que debería darse a estos valores faltantes. Así mismo se realizará un análisis específico para las columnas que contienen datos de tipo categórico para determinar cuál es la mejor estrategia de encoding para incluir estas variables al modelo. De manera similar, para las variables de tipo numérico, es necesario identificar si se requiere aplicar escalamiento o algún tipo de tratamiento para valores atípicos.

4.1.3 Análisis exploratorio de los datos

En este punto se busca familiarizarse con la estructura de los datos para identificar aspectos importantes y útiles para la etapa de modelado. En primer lugar, se aplicará un análisis descriptivo sobre la variable objetivo para identificar que tan balanceada se encuentra esta y si es necesario aplicar técnicas de aprendizaje automático imbalanceado. También se explorará la distribución de las variables predictivas y las relaciones que existen entre ellas mediante un análisis de correlación.

4.1.4 Modelado y evaluación

Para el modelado se escogió el algoritmo XGBoost ya que este es un modelo de caja negra que en la literatura académica ha demostrado superar el rendimiento de las técnicas más tradicionales en el campo de la administración de riesgo de crédito. Se planea ajustar un modelo de inicial y sobre este aplicar técnicas de ajuste de hiper-parametros como random search y grid search. Una de las ventajas que ofrece tomar un conjunto de datos que hace parte de una competencia en la plataforma Kaggle es que la evaluación del modelo puede ser realizada de forma automática subiendo una muestra de evaluación a la plataforma que automáticamente y de forma independiente evalúa el rendimiento del modelo, agregando que es posible comparar el rendimiento del modelo versus los resultados de los más de 4800 participantes de esta competición.

4.1.5 Aplicación de técnicas de inteligencia artificial explicable

Al obtener un modelo satisfactorio, se aplicarán cada una de las técnicas de XAI expuestas en el marco teórico con el fin de obtener información relevante acerca del funcionamiento del modelo. En este punto se busca demostrar como dichas técnicas ayudan a responder preguntas relevantes acerca de aspectos clave del modelo obtenido de modo que se muestre de forma práctica como estas técnicas pueden motivar la adopción de algoritmos que han demostrado mayor poder de discriminación en problemas de riesgo de crédito, aun a pesar de que estos algoritmos sean considerados de caja negra.

4.2 Herramientas

El desarrollo del flujo de trabajo propuesto se realizará en lenguaje Python, apoyándose en librerías como Pandas, Numpy, SciKit-Learn, Matplotlib, XGBoost, entre otras. Como entorno de desarrollo se usará Jupyter Notebook dadas las ventajas que este presenta para documentar el proceso y los resultados obtenidos. Además, como ya se mencionó, se usará la plataforma Kaggle para evaluar el rendimiento del modelo de forma automática e independiente.

5 RESULTADOS

5.1 Preprocesamiento y análisis exploratorio

5.1.1 Recopilación de los datos

Como se comentó en la propuesta metodológica, para el desarrollo de los análisis propuestos se usó el conjunto de datos proporcionado por American Express para su competencia de predicción de eventos de impago alojada en la plataforma Kaggle. El conjunto de datos contiene variables del perfil financiero de un conjunto de clientes en distintas fechas de estado de cuenta, la identificación de los clientes se encuentra anonimizada, además, las variables se encuentran normalizadas y clasificadas en los siguientes grupos identificables por el primer carácter de su nombre:

- D_* = Variables de morosidad
- S_* = Variables de gasto
- P_* = Variables de pago
- B_* = Variables de balance
- R_* = Variables de riesgo

Dentro del grupo de variables hay 11 variables de tipo categórico, que junto con las variables de tipo numérico conforman un conjunto de 188 disponibles para modelado.

El archivo de la población de entrenamiento tiene 5.531.451 registros, correspondientes a 458.913 clientes únicos en distintas fechas de estado de cuenta. Consecuentemente, hay un archivo que contiene las etiquetas de entrenamiento para los 458.913 clientes.

El archivo de la población de evaluación tiene 11.363.762 registros, correspondientes a 924.621 clientes únicos en distintas fechas de estado de cuenta.

5.1.2 Preprocesamiento

5.1.2.1 Ingeniería de características

Dado que la población de entrenamiento contiene 458.913 clientes únicos, es necesario realizar el preprocesamiento de la información disponible, para ello se sigue un proceso

de ingeniería de características en donde para cada cliente se derivan variables agregadas basándose en las distintas fechas de estados de cuenta que hay disponibles. En el caso de las variables continuas, se agrupa por cliente calculando el promedio, la desviación estándar, el valor mínimo, el valor máximo y el ultimo valor que toma cada variable. Para las variables de tipo categórico, se decidió agregarlas a nivel de cliente usando un encoding por conteo de casos, conteo de casos únicos, y el ultimo valor que toma cada variable.

De esta manera se obtiene un conjunto de datos de entrenamiento que cuenta con 458.913 registros y 920 campos, es decir 918 variables predictoras, el campo que identifica al cliente y la variable objetivo.

Para los datos de evaluación, en donde hay 924.621 clientes únicos, se realizó el mismo proceso de ingeniería de características descrito antes, sin embargo, dada la dimensión de los datos el costo computacional superó la capacidad de los recursos disponibles, por lo que dicho proceso se realizó después de obtener el conjunto de variables finales en la etapa de modelado, de tal manera que se acotó al conjunto de variables a aquellas requeridas para la evaluación y aplicación del modelo final.

5.1.2.2 Datos nulos y valores atípicos

A partir del conjunto de datos resultado de la etapa de ingeniería de características, se calculó el porcentaje de datos nulos presentado por cada una de las variables disponibles para modelado, en donde se encontró que algunas variables presentan una gran cantidad de valores nulos, sin embargo, es difícil determinar el método más adecuado para el tratamiento de los nulos dado que desconocemos el significado y el proceso generador de cada una de las variables. Por lo tanto, se decidió descartar aquellas variables que presentaron un porcentaje de valores nulos superior a cierto limite (97% para este ejercicio). Se considera que, en este caso en particular, esta forma de manejar los valores nulos es la que mejor asegura la confiabilidad e interpretabilidad de los resultados. Adicionalmente, cabe destacar que se tiene en cuenta un criterio de descarte adicional que se denomina como variabilidad, en este punto se considera que una variable no presenta

suficiente variabilidad si el valor de la variable en el un percentil dado es igual a su valor en uno menos el percentil, en este ejercicio, si el valor de la variable en su percentil 2.5% es igual a su valor en el percentil 97.5%, se considera que la característica no presenta la suficiente variabilidad y se descarta del análisis.

Como ya se mencionó, los datos vienen normalizados desde la fuente, y dado que el algoritmo que se aplicó en la etapa de modelado se basa en arboles de decisión, no existe la necesidad de aplicar otras transformaciones a los datos. Por otra parte, dado que los modelos basados en arboles de decisión son robustos frente a valores atípicos debido a la naturaleza de su construcción, no es necesario realizar tratamiento de datos atípicos.

5.1.3 Análisis exploratorio de los datos

5.1.3.1 Variable objetivo

En primer lugar, se analiza la distribución de la variable objetivo, al conocer su distribución es posible tomar decisiones informadas respecto a la estrategia de modelado. En este caso, se observa que la clase positiva (impago) representa aproximadamente el 26% de del total de la población de entrenamiento, esto significaría que no es necesario aplicar técnicas de aprendizaje automático imbalanceado, sin embargo, se debe tener en cuenta que desde la fuente de los datos la clase negativa se submuestreó al 5%, por lo tanto, recibe una ponderación de 20x en la métrica de rendimiento.

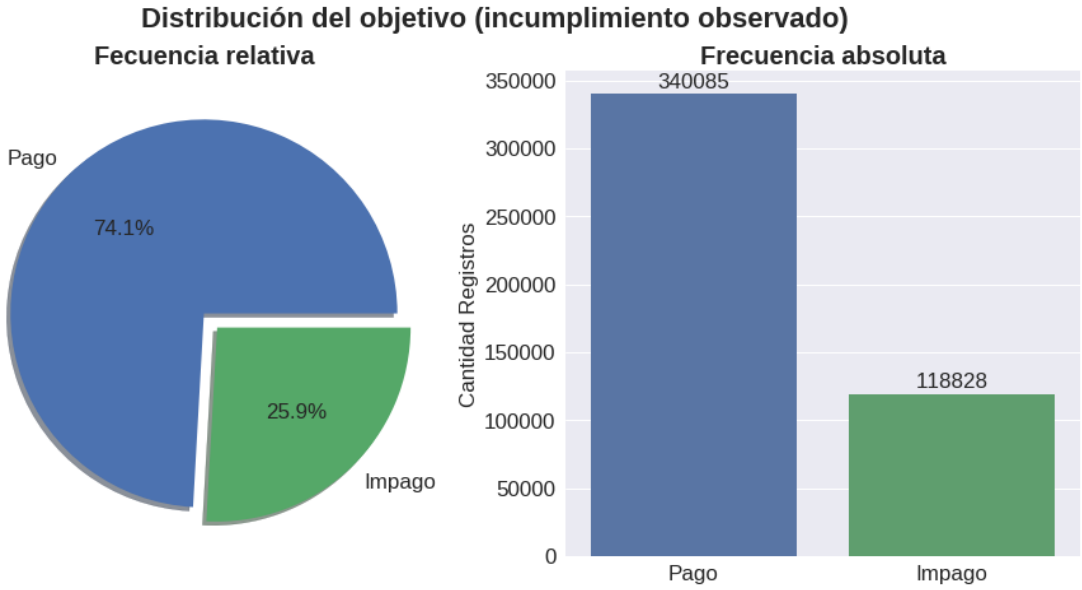

Ilustración 1- Distribución variable objetivo

Dado dicho escenario, es indispensable que la métrica de rendimiento usada en la etapa de modelado sea correcta para la casuística descrita, de tal manera que el modelo desarrollado proporcione estimaciones precisas y confiables.

5.1.3.2 Variables explicativas

En esta etapa del análisis se buscó identificar aquellas relaciones entre las variables que podrían ser importantes para la etapa de modelado. Dado el gran número de variables disponibles y que algunas variables derivadas son subgrupos de características de una misma variable original, es necesario realizar un análisis de correlaciones exhaustivo que garantice que solamente aquellas variables que aporten información valiosa para ingresen a la etapa de modelado.

Por construcción los modelos basados en arboles son robustos ante la multicolinealidad, pero teniendo en cuenta el costo computacional de la etapa de modelamiento para datos de grandes dimensiones, es deseable iniciar la etapa de modelado con un conjunto depurado de variables.

Para lograr, lo anterior se desarrolló un proceso que calcula el índice de correlación de Pearson para cada posible combinación de parejas de variables, en el caso de encontrar

una pareja cuyo índice de correlación es superior al cierto limite (80% para este ejercicio) se evaluó la correlación de cada una con la variable objetivo, la característica con menor correlación con la indicadora se descarta del análisis.

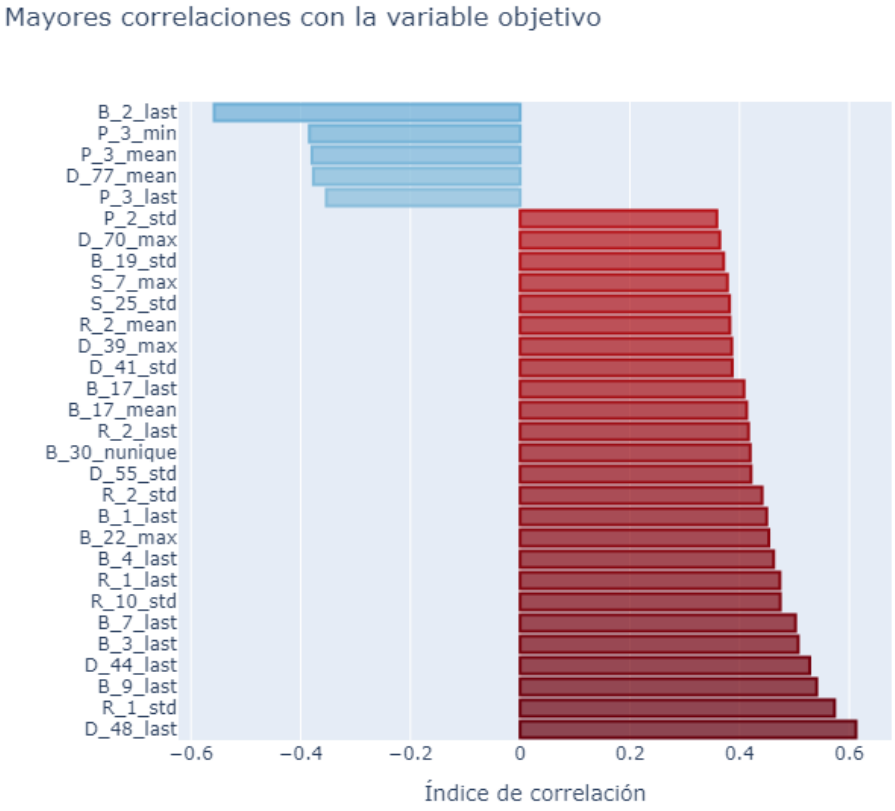

Ilustración 2-Mayores correlaciones con la variable objetivo

Se observa que entre las primeras 30 correlaciones más altas en términos absolutos la mayoría se correlaciona en positivamente con la variable objetivo, mientras el grupo que presenta correlaciones negativas es más reducido.

5.2 Modelado

5.2.1 Conjunto de datos inicial

Como se mencionó anteriormente, antes de iniciar la etapa de modelado se llevó a cabo un descarte de variables por completitud y variabilidad, que descarta variables que contienen nulos en su mayoría y que no presentan suficiente variabilidad en su contenido.

Así como también un descarte por correlación, que descarta variables altamente coleccionadas tomando como criterio de decisión la correlación de cada variable con la variable objetivo. Como resultado de este descarte de variables inicial, la cantidad de variables explicativas disponibles para el entrenamiento del modelo es de 364.

5.2.2 Partición de los datos

A partir del conjunto de datos descrito en la sección previa, se realizó una partición de los datos separándolos en conjuntos de entrenamiento y validación, con una proporción de 30% para validación. Cabe destacar que la partición de validación tiene el objetivo de garantizar la correcta generalización del modelo, servir para detectar posibles problemas y para evaluar el ajuste de los hiperparámetros, el rendimiento del modelo se evalúa sobre los datos proporcionados para tal fin, y se realiza de manera independiente y automática mediante la plataforma Kaggle.

5.2.3 Ajuste inicial de hiperparámetros

El primer paso de modelado consistió en hallar una malla de hiperparámetros aproximada que sirvió de insumo para las posteriores etapas de modelado. El objetivo en este punto es encontrar un conjunto de hiperparámetros que se aproxime a los óptimos y que permita hacerse una idea preliminar de los valores que debería tomar cada uno de estos. Para tal fin se llevó a cabo un ajuste de hiperparámetros por medio del razonamiento bayesiano, u optimización bayesiana. Este método puede reducir el tiempo necesario para llegar al conjunto óptimo de parámetros y comúnmente brinda un mejor rendimiento de generalización en el conjunto de prueba comparado con los métodos tradicionales de búsqueda aleatoria o búsqueda por rejilla. Al igual que los métodos tradicionales, este itera sobre un conjunto de posibles valores definidos ajustando un modelo y evaluando su rendimiento, sin embargo, lo hace teniendo en cuenta la información sobre las combinaciones de hiperparámetros que ha visto, es decir, en cada nueva iteración, se centra en áreas específicas alrededor de los parámetros óptimos recuperados de la ejecución anterior.

Cabe destacar que en cada interacción el modelo se ajusta sobre los datos de entrenamiento mediante una validación cruzada de cinco capas, y se evalúa usando la métrica AUC.

5.2.4 Selección de variables mediante múltiples modelos y SHAP Values

Dado que se tienen más de 364 variables explicativas es necesario establecer un criterio para el descarte de aquellas que no aporten o aporten muy poca información al modelo. Para lograrlo, se desarrolló un proceso que, haciendo uso de la malla de hiperparámetros encontrada en el paso anterior, ajustó iterativamente 100 veces un modelo XGBoost sobre particiones aleatorias del total de los datos disponibles para entrenamiento, en cada una de estas 70 iteraciones el proceso calculó y almacenó la importancia de las variables medida mediante SHAP Values, el proceso retornó un informe detallado en donde para cada variable se tiene la importancia que obtuvo en cada muestra aleatoria y modelo. Se considera que una variable es importante y estable si su participación se mantiene constante sin importar la muestra.

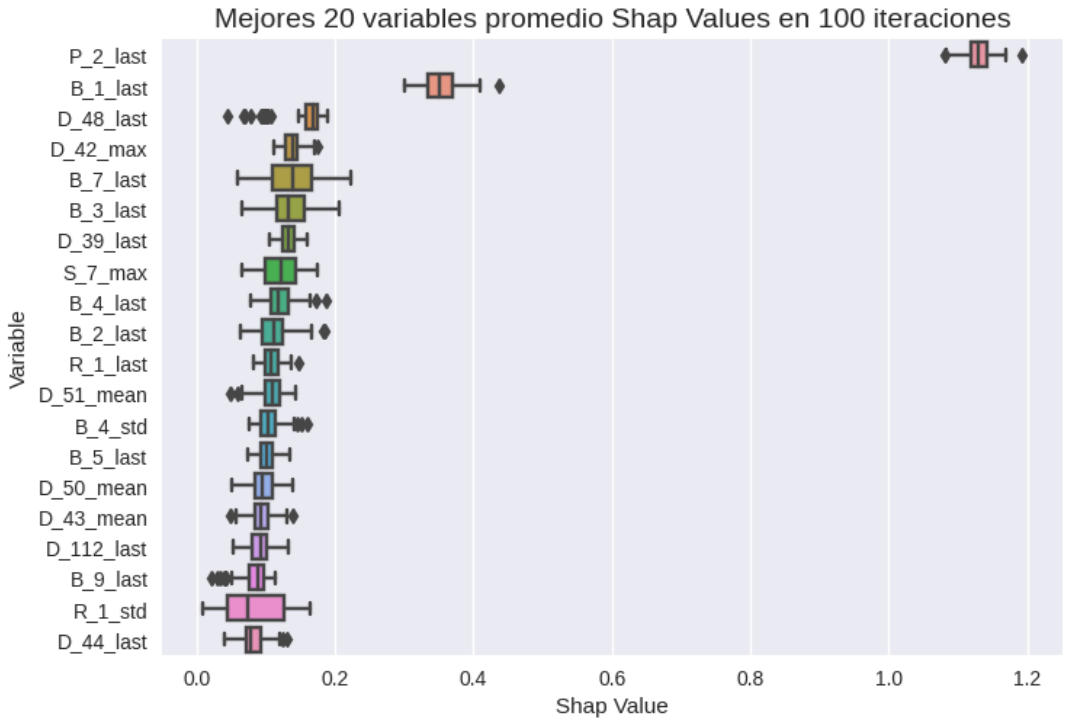

Ilustración 3-Mejores 20 variables promedio SHAP values

Se observó que, en general el Shap value obtenido por cada variable en cada una de las 100 iteraciones no presenta una gran cantidad de valores atípicos, y también la que varianza de estos no es grande, también se verificó que las variables fueran significativas en cada una de las 100 muestras aleatorias del análisis, en donde se encontró que incluso hasta la décimo tercera variable ordenadas por Shap promedio se cumplió esta condición. Se destacó que las dos primeras variables parecen tener un Shap Value promedio muy superior a todas las demás, por lo tanto, se espera que estas sean las más importantes del modelo final.

Con este resultado fue posible determinar las mejores variables para la etapa final de modelado. Para mantener la explicabilidad del modelo, es importante que el grupo de variables explicativas finales sea lo más parsimonioso posible, de tal manera que el conjunto de relaciones capturado intuitivamente por las variables sea razonable y no exceda la retentiva humana. Por esta razón se tomó la determinación de seleccionar las primeras 15 variables con mayor Shap Value promedio en esta etapa para seguir en la etapa de modelamiento final.

5.2.5 Ajuste final de hiperparámetros y modelo definitivo.

Se procedió a realizar una etapa final de ajuste de hiperparámentros mediante optimización bayesiana, en este punto el objetivo fue llegar al conjunto de valores de hiperparámetros que garanticen el mejor ajuste del modelo usando las 15 variables seleccionadas en la etapa anterior. Para ello, en esta etapa se tuvo en cuenta los resultados obtenidos en la primera ronda de ajuste mediante optimización bayesiana y se acoto el espacio de búsqueda de cada hiperparámetro a un rango más pequeño alrededor del valor obtenido en dicha primera etapa. De esta manera, manteniendo el número de iteraciones constante, se realizó una búsqueda más exhaustiva dado que el espacio de búsqueda fue de menor dimensión.

Por último, se ajustó el modelo XGBoost final usando el conjunto de variables obtenido en la etapa de selección de características mediante Shap values, y el conjunto de hiperparámetros obtenidos anteriormente.

Hiperparámetro	Valor
learning_rate	59.7%
colsample_bytree	98.6%
subsample	91.0%
reg_lambda	0.79
reg_alpha	47
min_child_weight	25

Rendimiento en Validación	
AUC	95.2%
Competencia	75.3%

Tabla 1-Hiperpámetros y rendimiento del modelo

Al evaluar el modelo sobre la partición de validación, se obtiene un valor satisfactorio de AUC que refleja que el modelo es capaz de discriminar correctamente la variable objetivo. Por otra parte, cabe destacar que en el caso de la competencia de alojada de en Kaggle la evaluación automática se realiza mediante una métrica especifica determinada por American Expres, la métrica de evaluación, M, para esta competencia es la media de dos medidas de orden de clasificación: Coeficiente de Gini normalizado, G, y tasa de incumplimiento capturada en el 4% superior de la probabilidad de incumplimiento predicha, D, es decir:

$$M = 0.5 * (G + D)$$

Al evaluar el rendimiento del modelo bajo esta última métrica de rendimiento se obtiene que el modelo generaliza correctamente y tiene un rendimiento aceptable. Conclusión que se refuerza al analizar la matriz de confusión, en donde se encuentra que, del total de casos de impago observados, el modelo es capaz de clasificar correctamente el 78.5%, con una precisión total de 89%.

	precision	recall	f1-score	support
Pago	92%	93%	93%	1,018
Impago	79%	79%	79%	35,830
accuracy			89%	137,674
macro avg	86%	86%	86%	137,674
weighted avg	89%	89%	89%	137,674

Tabla 2-Reporte de clasificación

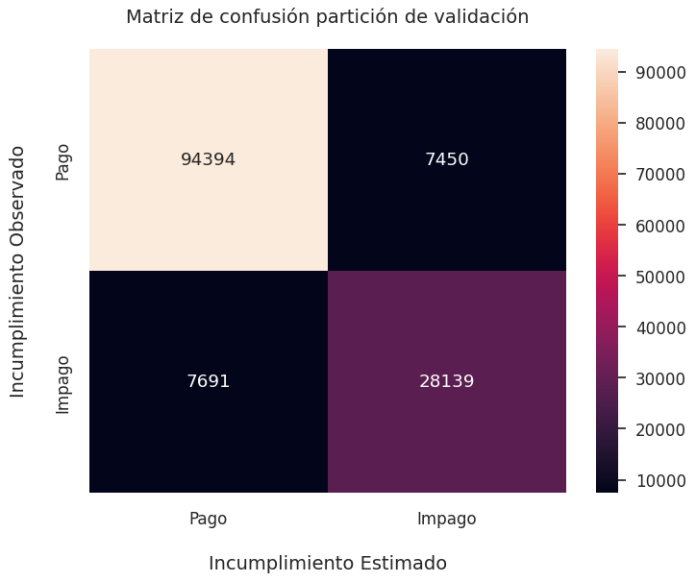

Ilustración 4-Matriz de confusión

Así mismo, para los datos disponibles para evaluación, en los cuales no se cuenta con el incumplimiento observado y por lo tanto solamente pueden ser evaluados bajo la métrica de rendimiento determinada para la competencia y que se realizó de forma independiente y automática por la plataforma Kaggle, se obtuvo:

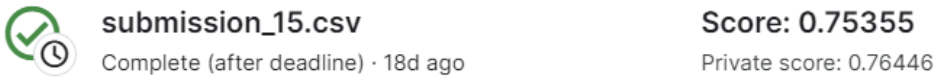

Ilustración 5-Métrica de rendimiento determinada para la competencia

Se observa que la métrica de rendimiento sobre los datos de prueba es muy similar a la obtenida en los datos de validación, lo que refleja que el modelo cuenta con una buena capacidad de generalización en diferentes poblaciones sobre las cuales no fue entrenado.

5.3 Técnicas de explicabilidad

Se aplicaron diversas técnicas desarrolladas en el campo de la inteligencia artificial explicable, algunas de ellas con el fin de entender aspectos de interés global sobre el modelo, otras con el objetivo de explicar los resultados del modelo para un caso en

particular, es decir, a nivel local.

5.3.1 Importancia de las variables mediante Shap values

Aunque en el marco teórico la técnica SHapley Additive exPlanations se presenta como un método de explicabilidad local, como su nombre lo indica este método cuenta con la propiedad de ser aditivo, por lo que es posible obtener la importancia global de una variable promediando los valores Shap para cada una de sus observaciones individuales. El siguiente grafico resume de forma combinada la importancia de cada una de las variables y su respectivo efecto sobre la probabilidad de incumplimiento.

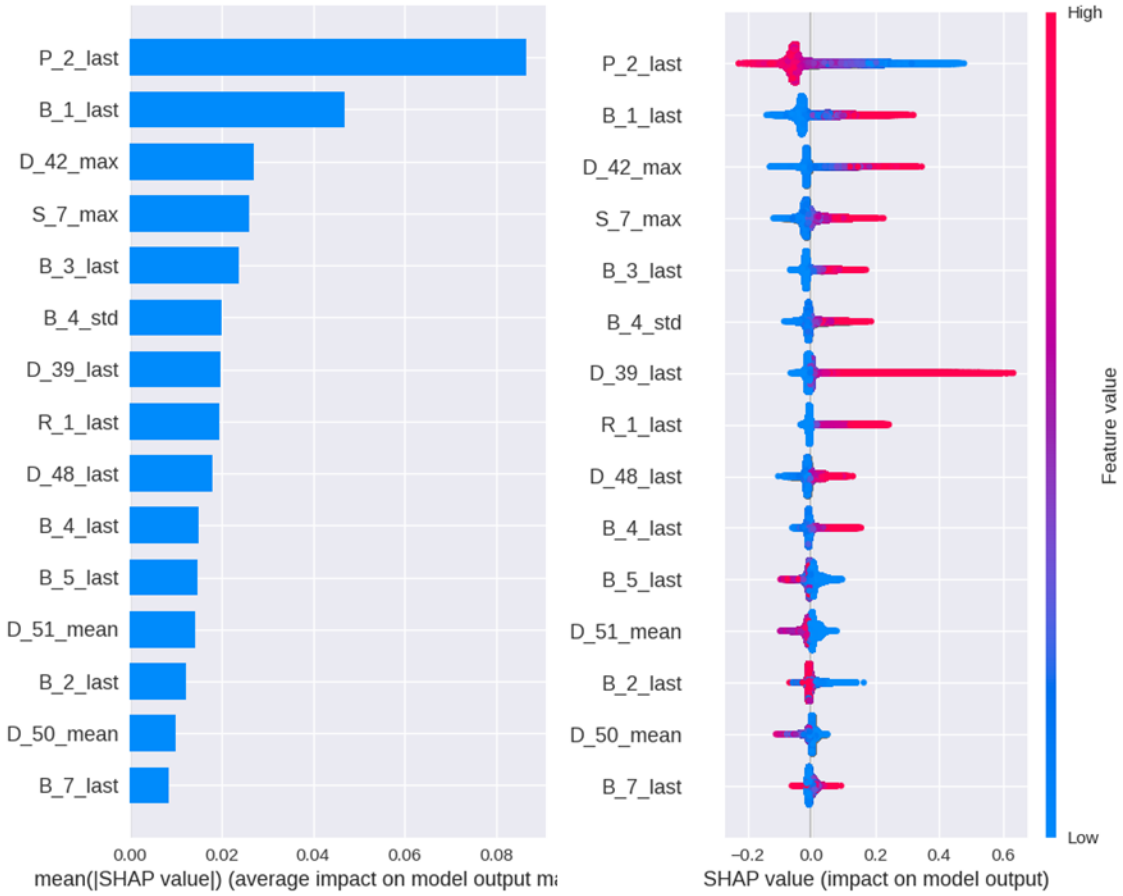

Ilustración 6-Importancia de las variables mediante Shap values

En la parte derecha se presenta la importancia promedio de las variables, es decir el impacto promedio que la variable tiene sobre la probabilidad de incumplimiento. En la

parte izquierda se visualiza cada uno de los puntos correspondientes al valor SHAP para una variable y observación, la posición del eje y determina la variable y el eje x representa el valor SHAP. La super posición de los puntos para cada variable permite hacerse una idea de su distribución. El Color permite representar el valor que toma la variable. Las variables se presentan ordenadas según su importancia.

De esta manera podemos observar que la variable P_2_last es la más importante, y esta tiene un impacto promedio de 8.6% sobre la probabilidad de incumplimiento estimada. Así mismo se concluye que valores altos de esta variable se relacionan con menores probabilidades de incumplimiento, mientras que valores bajos de esta variable se relacionan con altas probabilidades estimadas de incumplimiento. Análogamente, es posible extender el análisis para las demás variables involucradas en el modelo, de tal manera que permite entender las relaciones entre las variables predictoras y la salida del modelo.

5.3.2 Importancia por permutación de variables

Como ya se había comentado el método de importancia por permutación de variables tiene la ventaja de ser altamente interpretable dado que se basa en un concepto simple, medir la magnitud del incremento del error de las predicciones a medida que la variable estudiada tiende a hacerse aleatoria. Para este ejercicio se llevaron a cabo 10 rondas de permutación de variables, esto dado que la permutación de una variable es un proceso aleatorio que puede resultar en valores distintos de perdida de rendimiento, al realizar varias rondas es posible evaluar la media y la deviación de los resultados.

En el grafico se presenta a la derecha la importancia de las variables medida mediante el concepto de gain, con fines de comparación. En el gráfico de la izquierda se visualiza para cada variable un boxplot construido con los valores resultado de las 10 rondas de permutación para cada variable, se observa que la característica con mayor importancia fue P_2_last asociada con un aumento de error en la métrica de rendimiento (AUC) de aproximadamente 0.05.

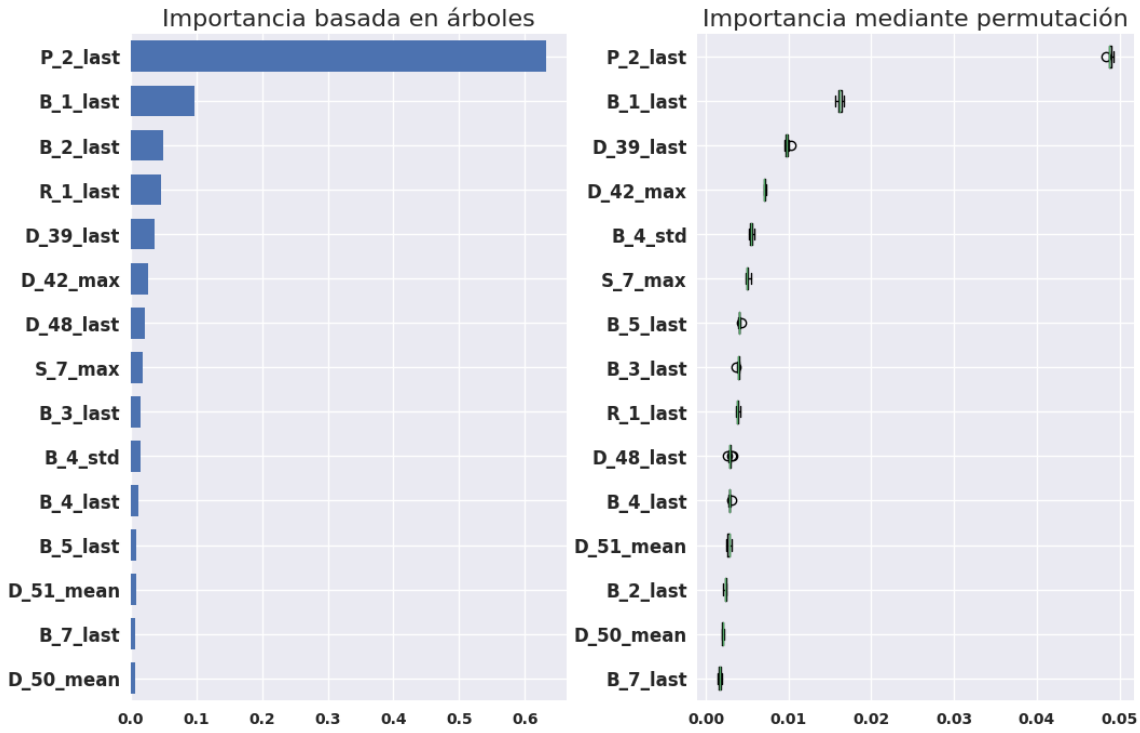

Ilustración 7-Importancia por permutación de variables

Se desataca que la variable B_2_last que a la luz de la importancia basada en arboles es la tercera más relevante, evaluada mediante permutación ocupa el puesto 13, lo que es consecuente con los resultados presentados anteriormente mediante SHAP values, este caso permite ejemplificar como la medida de importancia basada en arboles puede ser no tan adecuada para evaluar la importancia de una variable.

5.3.3 Graficas de dependencia parcial y expectativa condicional individual

Los gráficos PDP e ICE representan la forma funcional de una variable con respecto a la variable objetivo. En el eje y se puede observar el impacto de la variable sobre la probabilidad de incumplimiento estimada, en el eje x se representan los valores de la variable. Las líneas azules representan la expectativa condicional individual, es decir, el impacto de una observación sobre la probabilidad de incumplimiento. La línea naranja representa la dependencia parcial de la variable, es decir, el efecto promedio de cada una de las observaciones sobre la probabilidad de incumplimiento. Los gráficos PDP y ICE

utilizan el mismo enfoque para modelar cambios de predicción, pero mientras que los gráficos PDP muestran las tendencias generales, los gráficos ICE proporcionan una visualización más granular de la variación de la predicción a lo largo de los valores de la variable estudiada para cada observación.

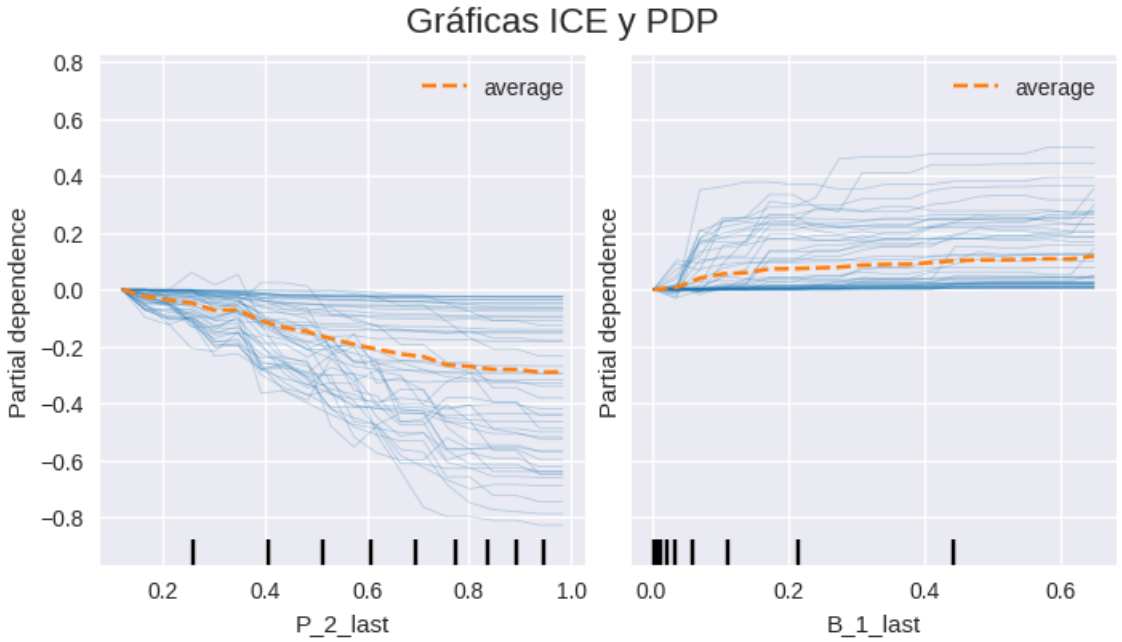

Ilustración 8-Gráficos PDP e ICE

En el grafico se presentan los gráficos ICE y PDP para las dos variables de mayor importancia según los resultados de la sección anterior. Se observa que la forma funcional de la relación entre la variable P_2_last y la probabilidad de impago estimada es monótona decreciente, es decir, valores altos de esta variable se relacionan con probabilidades bajas de impago. Por su parte, la relación funcional de B_1_last y el pronóstico del modelo es monótona creciente, pero se destaca que afecta la probabilidad de impago estimada en una magnitud menor que P_2_last. Cabe destacar que estos resultados consecuentes con lo hallado en la sección anterior.

5.3.4 Explicaciones agnósticas del modelo localmente interpretables

En primer lugar, se debe definir la función que estima la probabilidad que usará el marco de trabajo LIME, esta función no es más que el modelo conseguido. Con esto es posible

obtener la explicación que nos entrega este marco de trabajo para una observación en particular.

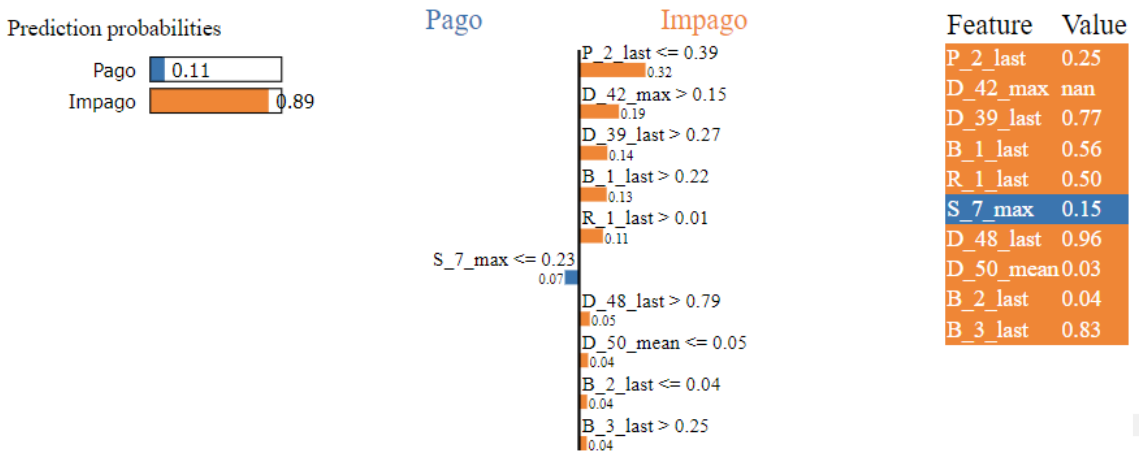

Ilustración 9-Explicabilidad local mediante LIME

El grafico anterior está conformado por tres elementos que proporcionan información relevante para entender la decisión del modelo para una observación. La coloración indica el sentido de la relación entre el valor de la variable y su contribución a la probabilidad estimada, en naranja aquellas variables que contribuyen a aumentar la probabilidad y en azul aquellas que su contribución disminuye la estimación. A la izquierda se muestra la categoría que el modelo estima para la observación, en este caso impago con una probabilidad estimada de 89%. En el centro podemos observar un gráfico que nos muestra una estimación de la importancia de la variable sobre la decisión individual de esta observación, P_2_last tiene la mayor importancia con 32%, seguida de D_42_max con un 19% de importancia. Finalmente, a la derecha se presentan los valores de cada variable para la observación estudiada.

5.3.5 Explicabilidad local mediate SHAP Values

El marco de trabajo SHAP Values cuenta con la ventaja de ser bastante versátil, en este caso es posible aplicarlo para logar explicabilidad local del modelo. El grafico mediante la coloración indica el sentido de la relación entre el valor de la variable y su contribución

a la probabilidad estimada, en aquellas variables que contribuyen a aumentar la probabilidad se muestran rosado oscuro y en azul aquellas que su contribución disminuye la estimación. La magnitud de la barra para cada variable indica su importancia, en la parte inferior se muestra el valor que toma la variable para la observación estudiada. En la parte superior el grafico indica el valor base, es decir la probabilidad promedio estimada por el modelo, también se puede ver el valor de la función de estimación f(x), en este caso 89% de probabilidad de impago.

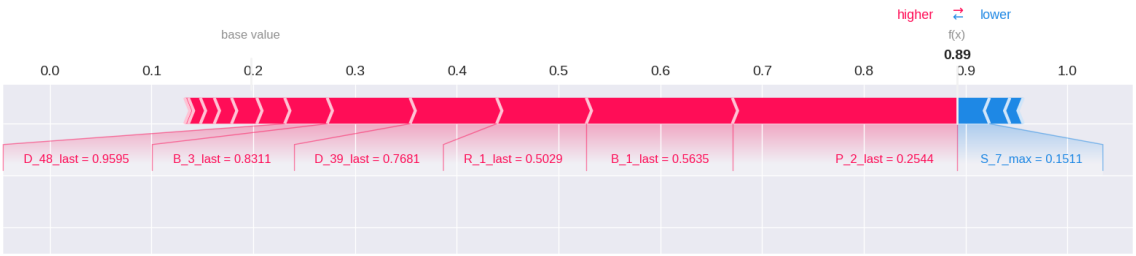

Ilustración 10-Explicabilidad local mediante SHAP values

Se observa, nuevamente, que la variable P_2_last es la que mayor aporte absoluto tiene sobre la decisión del modelo, seguida de B_1_last y R_1_last. Cabe destacar que la observación tomada es la misma estudiada en la sección anterior, salta a la vista que la variable D_42_max, segunda en importancia según LIME en este caso ni siquiera aparece entre las más relevantes.

5.3.6 Análisis contrafactual

Las explicaciones contrafactuales permiten obtener información del funcionamiento del modelo mediante la perturbación de las variables y su impacto en la salida del modelo, es decir, proporciona explicaciones de la forma “que pasaría si” se simula que una o varias variables toman valores diferentes hasta que el resultado de la predicción del modelo cambie. Se busca entender para una observación cuáles son los mínimos cambios en las variables de entrada que conducen a que la predicción del modelo cambie, en el ámbito de la administración de riesgo de crédito en personas naturales, esto significa aprobar o negar un préstamo.

Variable	Original	Contrafactual 1	Contrafactual 2
P_2_last	- 0.016	0.629	0.980
B_1_last	0.100	-	-
D_48_last	0.979	-	-
D_42_max	0.305	-	-
B_7_last	0.665	-	-
B_3_last	0.193	-	-
D_39_last	0.037	-	-
S_7_max	0.627	-	-
B_4_last	0.648	-	-
B_2_last	0.144	-	-
R_1_last	1.506	-	-
D_51_mean	0.004	-	-
B_4_std	0.097	0.097	0.097
B_5_last	0.004	7.559	1.803
D_50_mean	0.058	-	-
target	1	0	0

Tabla 3-Resultados análisis contrafactual

En el caso de ejemplo podemos observar como la observación en particular los valores de sus variables llevan a que su categoría estimada sea 1 (impago), al aplicar análisis contrafactual es posible llegar a que si el cliente desea cambiar el resultado de su estimación lo mínimo que debería hacer es intentar cambiar los valores que presenta en las variables P_2_last, B_4_std y B_5_last según se indica en la tabla anterior.

5.3.7 Reglas de ámbito (anclajes)

De cierta forma, las reglas de ámbito se pueden considerar como la contraparte del análisis contrafactual, en este caso el método busca encontrar valores de variables que hacen que la predicción del modelo se mantenga constante independientemente de los valores que puedan tomar otras variables, es decir, el objetivo es encontrar aquella combinación de valores de un conjunto de variables que anclan la predicción del modelo. Este tipo de reglas que anclan la predicción del modelo pueden ser muy útiles para entender la forma como el modelo toma decisiones y su respectiva lógica subyacente.

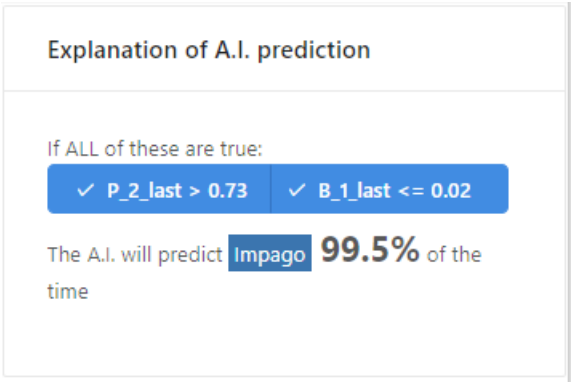

Ilustración 11-Resultado reglas de ámbito

La interpretación de esta regla de ámbito es la siguiente, si P_2_last > 0.73 y B_1_last <=0.02 entonces la predicción del modelo será impago en un 99.5% sin importar los valores que puedan llegar a tomar las demás variables del modelo.

5.3.8 Sustituto global

El ajuste de un modelo sustituto global cuenta con la ventaja de que es fácilmente aplicable independiente del modelo de caja negra que se quiere explicar, solamente se requiere tener acceso a los datos y a la predicción del modelo. Cabe destacar nuevamente, que el modelo sustituto al no ser entrenado sobre las etiquetas reales de los datos no debe ser usado para estimación.

En este caso específico se usó un árbol de decisión con una profundidad igual a 5, se ajustó el modelo de manera tal que la variable objetivo fue la estimación del modelo de caja negra. Como resultado se obtiene un modelo que presenta una métrica ROC-AUC de 97,6%, es decir que captura muy bien el comportamiento del modelo XGBoost que se busca explicar.

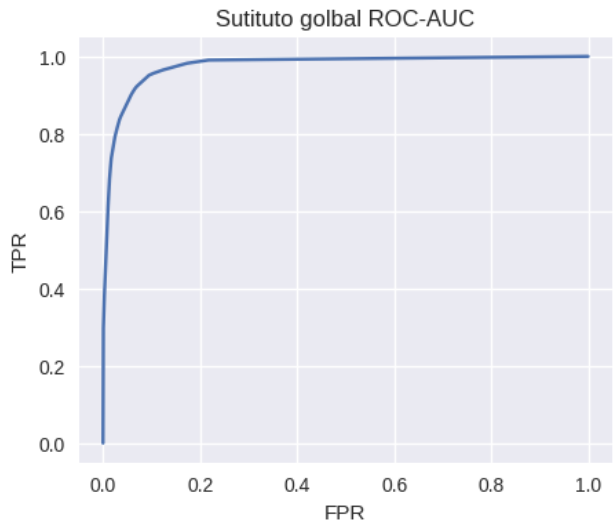

Ilustración 12-Métrica ROC-AUC para sustituto global

A continuación, se presenta el modelo sustituto global en sus dos primeros niveles, podemos observar que la condición del primer nodo involucra la variable P_2_last, y los nodos de la primer a profundidad del árbol involucran la variable B_1_last, lo que concuerda con resultados obtenidos anteriormente. Este resultado es mucho más interpretable y se puede tomar como una aproximación del funcionamiento lógico del modelo XGBoost.

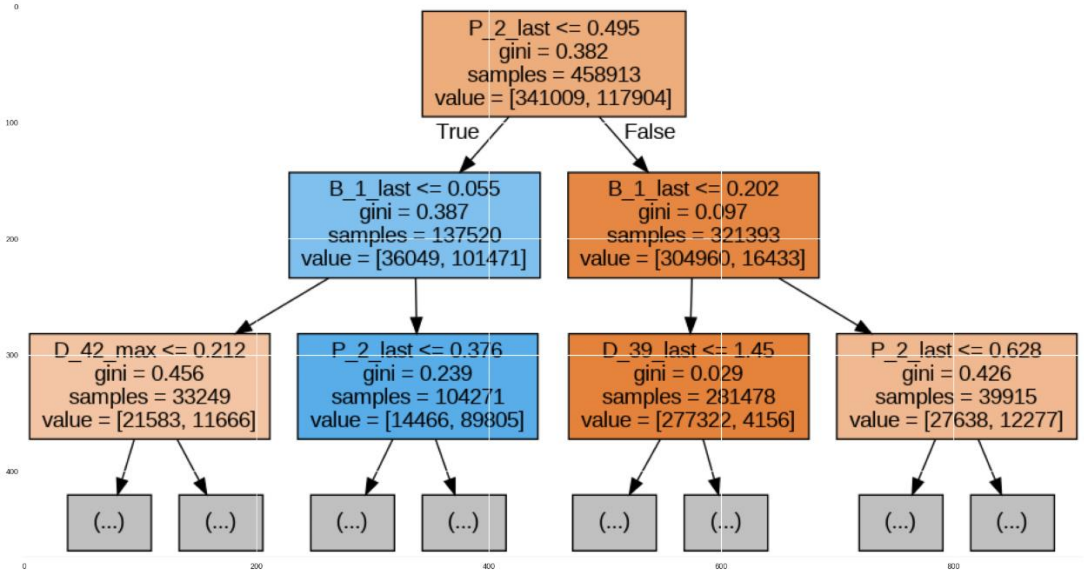

Ilustración 13-Primeros niveles árbol sustituto global

5.4 Marco de trabajo para la inteligencia artificial explicable en la administración de riesgo de crédito

En esta sección se pretende proponer un marco de trabajo que integre los métodos de inteligencia artificial explicable en el proceso de desarrollo de modelos de estimación de probabilidad de incumplimiento de pago en créditos a personas naturales. Aunque el dominio de aplicación de este trabajo es especifico, es fácilmente extrapolable a otros dominios que requieren entendimiento y explicabilidad de los modelos predictivos.

Como se mencionó anteriormente, en el campo de riego de crédito, el principal objetivo es estimar probabilidades de incumplimiento lo más cercanas al incumplimiento observado en la realidad, con base en esto es posible estimar los niveles de perdida esperados, calcular los niveles de provisiones de acuerdo con las perdidas esperadas, fijar políticas de pricing (tasas de interés crediticio), ajustar los niveles de aprobación, entre otras estrategias de gestión de riesgo. Por lo tanto, existe un costo de oportunidad significativo en la precisión que los modelos de estimación de probabilidad de incumplimiento logren obtener.

Sin embargo, en este campo la precisión no lo es todo, como también se ha mencionado anteriormente, por normatividad y por ética empresarial es importante mantener un entendimiento profundo de los modelos y las lógicas subyacentes a los pronósticos generados. En otras palabras, la transparencia algorítmica y explicabilidad, son dos elementos de vital importancia en este campo, en donde la confianza en las predicciones y en el modelo en general, además las regulaciones gubernamentales exigen contar con estos factores.

Las partes interesadas en el desarrollo del modelo deben contar con la confianza y el entendimiento suficiente en la herramienta. Los científicos de datos se benefician de esta explicabilidad dado que les permite comprender cómo funcionan los modelos que han creado, identificar posibles mejoras o errores en el proceso de desarrollo y validar la integridad y robustez de los resultados. Por su parte, al nivel directivo de las organizaciones, la explicabilidad les ayuda a evaluar la confiabilidad y la validez de las estimaciones del modelo y entender el impacto de su despliegue. Así mismo, los

reguladores y organismos de control tienen la responsabilidad de garantizar la equidad, la ética y la legalidad en el uso de los modelos predictivos. Para cumplir con esta tarea, necesitan tener una comprensión clara de cómo funciona el modelo y si hay algún sesgo o discriminación involucrada en las decisiones tomadas por el mismo. A su vez, los equipos de ingeniería responsables del despliegue a producción de los modelos y su integración con el sistema analítico de la organización deben tener un cierto entendimiento de cómo funciona el modelo. Por último, los usuarios evaluados por el modelo pueden querer entender la razón por la cual el modelo toma determinada decisión, la explicabilidad les brinda la confianza y transparencia requerida.

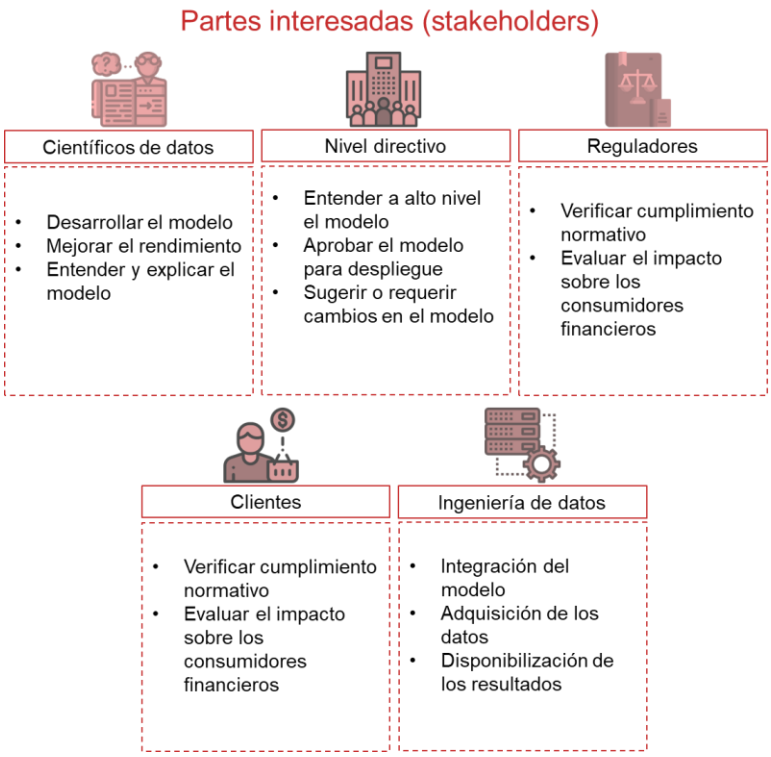

Ilustración 14-Partes interesadas (stakeholders)

En este contexto, el primer paso en el desarrollo de un modelo de estimación de probabilidad de incumplimiento es evaluar detenidamente que tipo de modelo emplear, una decisión puede ser desarrollar un modelo transparente, en este dominio el ejemplo clásico es la regresión logística, con la cual es totalmente explicable. Por otra parte, es

posible desarrollar un modelo algorítmicamente más complejo y por lo tanto más opaco en sus lógicas, como ya se ha dicho, este tipo de métodos más sofisticados por lo general logran un rendimiento mayor y mejor capacidad de generalización que los métodos tradicionales a costa de explicabilidad.

Lo más recomendable es recorrer ambos caminos, es decir, desarrollar por lo menos un par de modelos, uno transparente y uno opaco, y comparar el rendimiento de estos. De frente a los resultados obtenidos evaluar en cual caso se logra una relación rendimiento-transparencia más acorde con los requerimientos del caso. En el caso de que la decisión resultante sea aplicar el modelo de caja negra, es necesario afrontar la desventaja que se pierde transparencia en la función de predicción, por lo que se vuelve necesario aplicar técnicas de inteligencia artificial explicable para ganar entendimiento sobre el modelo.

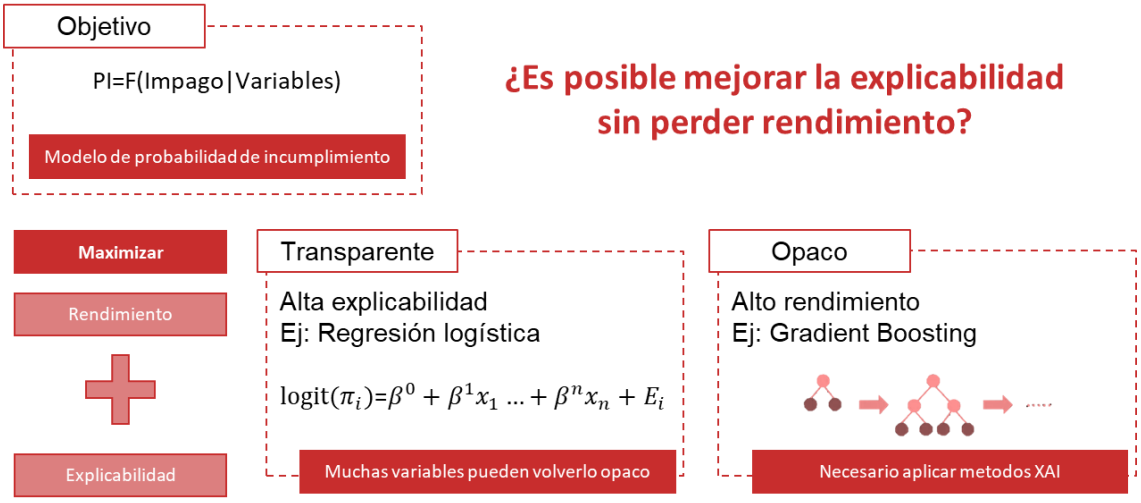

Ilustración 15-Objetivos y retos en la estimación de probabilidad de incumplimiento

En caso de encontrarse en dicho escenario, es aconsejable seguir un enfoque que busque ganar entendimiento del modelo desde lo general a lo especifico, es decir, que en primer lugar se deben aplicar métodos de explicabilidad global para extraer las lógicas más representativas y generales del modelo. Al adoptar un enfoque desde lo general, se busca obtener una visión panorámica de las variables y patrones generales que el modelo utiliza para tomar decisiones.

Es importante comprender las relaciones y el tipo de razonamiento que le modelo aplica sobre las variables de entrada, lo cual implica analizar su importancia e influencia sobre las estimaciones, la manera en que se relaciona cada variable con los resultados de modelo, y cualquier otro factor que pueda influir en la toma de decisiones. Al explorar estas lógicas representativas, se obtiene una visión más clara de cómo el modelo procesa y utiliza la información.

En este punto se pueden usar técnicas descritas y aplicadas anteriormente como la importancia por permutación o los Shap values aplicados a la importancia de variables. Hoy por hoy, el marco Shap es bastante difundido en la industria, ya que adicional a la importancia global de las variables permite entender la magnitud y la dirección del efecto de cada variable sobre las estimaciones resultado del modelo, esto dado que cumplen con la propiedad de descomposición aditiva, lo que significa que la importancia total de un modelo se puede explicar como la suma de las importancias individuales de las variables para cada instancia. Esto facilita el análisis de la contribución de cada variable y la comprensión de las interacciones entre ellas.

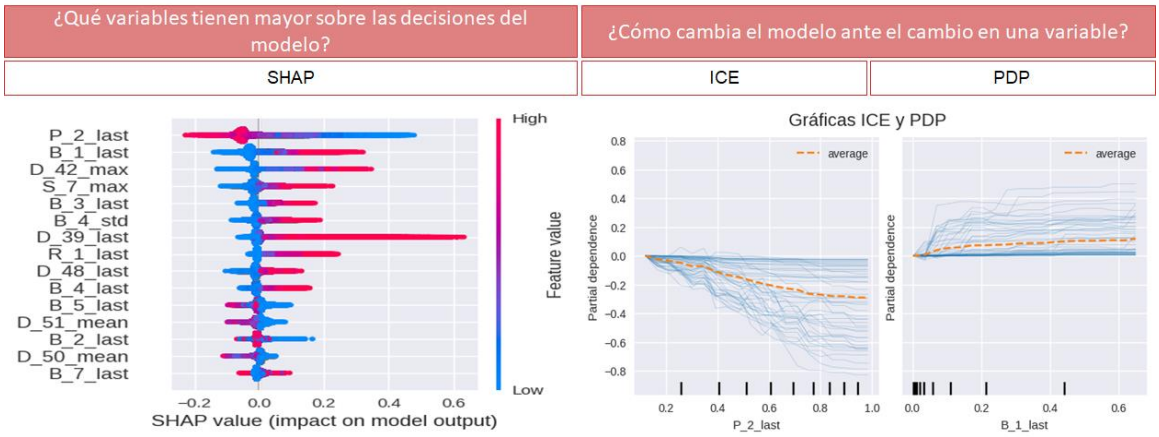

Ilustración 16-preguntas y explicaciones a nivel global

A medida que se gana entendimiento sobre el funcionamiento del modelo es posible explorar lógicas más específicas, por ejemplo, se puede tener interés en entender cómo funciona en modelo cuando una variable de interés es relativamente alta o baja. En dicho escenario es posible apoyarse en un método como las gráficas de expectativa condicional

individual para evaluar el comportamiento del modelo sobre un conjunto de observaciones en donde los valores de la variable de interés varían, mientras que los valores de las demás variables se mantienen constantes en sus valores observados. Adicional a lo anterior, es posible aplicar el método de diagramas de dependencia parcial, que es el complemento natural de ICE, en este caso se puede explorar la frontera de decisión del como una función de la variable de interés, mientras que el valor del resto de variables es distorsionado hacia sus medias respectivamente.

Puede darse la situación en que la explicabilidad sea requerida en el nivel más granular de los datos, es decir a nivel de observación. En el caso de que alguna de las partes interesadas requiera entender el razonamiento que llevo al modelo, por ejemplo, a rechazar una solicitud de crédito es posible dar una respuesta apoyándose en técnicas como LIME o Shap values a nivel de instancia.

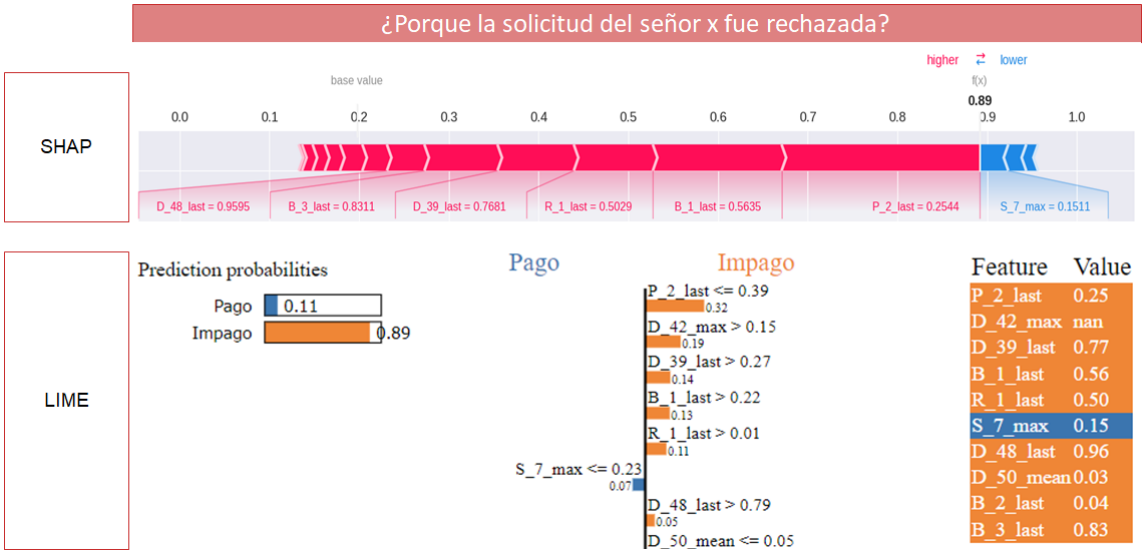

Ilustración 17-preguntas y explicaciones a nivel local

Tanto LIME como Shap values ofrecen herramientas que permiten explicar a nivel de instancia las decisiones tomadas por un modelo. Estas técnicas pueden ayudar a las partes interesadas a comprender y razonar sobre el proceso de toma de decisiones del modelo, especialmente en situaciones donde la explicabilidad detallada es necesaria. Además, estas técnicas permiten mostrar sus resultados de manera gráfica, lo que facilita su comunicación y comprensión. Adicional a lo anterior, es posible apoyar la explicación

mediante la aplicación de análisis contrafactual, de tal manera que el cliente o la parte interesada obtenga la información acerca de cuáles son los mínimos cambios en las variables de entrada que conducen a que la predicción del modelo cambie.

En caso de necesitar explicaciones adicionales acerca del funcionamiento y las lógicas del modelo es posible aplicar otras técnicas de las que se disponen en el campo de la inteligencia artificial explicable, por ejemplo, la aplicación de un modelo sustituto global puede ayudar a representar y resumir de manera sencilla y comprensible las reglas que sigue el modelo de caja negra, o en el mismo sentido, un análisis de reglas de ámbito puede dar razón de aquellas reglas más representativas.

En resumen, el marco de trabajo propuesto para integrar los métodos de inteligencia artificial explicable en el proceso de desarrollo de modelos de estimación de riesgo de crédito sigue los siguientes pasos:

1. Evaluar la pertinencia del uso de un modelo de caja negra comparado con un modelo transparente, es posible que para un conjunto de datos en específico un modelo transparente presente mejor rendimiento que uno de mayor complejidad algorítmica.
2. En caso de que el modelo de caja negra tenga un rendimiento superior, se debe iniciar a explicar el funcionamiento de lo general a lo especifico, los primeros esfuerzos deben enfocarse en entender la importancia de las variables, su impacto y su magnitud sobre las estimaciones del modelo.
3. A continuación, es aconsejable comprender el efecto de las variables en subdominios específicos de los datos, es decir, entender con mayor profundidad el efecto de las variables cuando se fijan los valores de otras variables en valores relativamente altos o bajos, por ejemplo, cómo funciona el modelo en un segmento de clientes con salarios relativamente bajos.
4. El caso más granular de explicabilidad se da al nivel de observación en donde es posible mostrar la lógica que el modelo aplica sobre las variables para generar una estimación, así mismo, se puede informar cuales deberían ser los cambios en los valores observados en las variables si se desea obtener un resultado distinto del modelo.

Cabe destacar que en cada uno de los puntos del marco de trabajo propuesto es posible aplicar un conjunto distinto de técnicas, lo más recomendable es aplicar varias de estas y siempre apoyar los resultados obtenidos con los obtenidos en otros análisis, de manera que se genere un entendimiento robusto del funcionamiento del modelo. En general, intentar usar un conjunto más amplio de métodos de explicabilidad permite una comprensión más profunda.

6 CONCLUSIONES

En este trabajo se mostró cómo con la aplicación de metodologías de inteligencia artificial explicable (XAI) es posible explicar cómo un modelo de aprendizaje automático algorítmicamente complejo (caja negra) obtuvo una solución particular, y también como estas metodologías pueden responder otras preguntas relevantes para entender el funcionamiento del modelo, lo anterior aplicado específicamente al campo de la administración del riesgo de crédito en el otorgamiento de préstamos personales.

Se propone un marco de trabajo que busca generar respuestas precisas, comprensibles y oportunas a las posibles preguntas que puedan tener las partes interesadas en el proceso de modelamiento, despliegue y aplicación de un modelo considerado de caja negra para la estimación de la probabilidad de incumplimiento crediticio.

Lo anterior pretende motivar la adopción de métodos de IA y modelos de aprendizaje automáticos más sofisticados en el campo de la administración de riesgo de crédito, demostrando cómo la aplicación de metodologías de aprendizaje automático explicado puede responder preguntas relevantes, acerca del modelo y sus predicciones, para que las partes implicadas ganen confianza e inteligibilidad sobre modelos denominados de caja negra.

A lo largo de este proyecto se expone la forma en que la aplicación de diversas metodologías XAI permiten romper la limitante de explicabilidad que ha llevado al rezago en la adopción de métodos predictivos más recientes y precisos en el campo de la administración de riesgo. Con una estimación más precisa de las probabilidades de incumplimiento, las entidades financieras pueden tomar decisiones más informadas sobre la concesión de crédito y la gestión del riesgo. Esto les permitiría identificar de manera más efectiva a los prestatarios de alto riesgo y establecer políticas adecuadas de asignación de recursos y provisiones para cubrir posibles pérdidas.

Por último, el campo de la inteligencia artificial explicable es de reciente aparición y constante desarrollo, por lo que desarrollos novedosos son propuestos desde la academia y la industria frecuentemente, algunas de las líneas más prometedoras y destacables son las investigaciones en ética y sesgo en la inteligencia artificial que se centra en el

desarrollo de métodos que garanticen la equidad y la imparcialidad en la toma de decisiones, evitando la discriminación basada en características sensibles. Otra línea destacable es el desarrollo de modelos híbridos interpretables este tipo de enfoques combinan métodos de inteligencia artificial complejos con modelos más interpretables, como regresiones lineales o árboles de decisión, para obtener un equilibrio entre rendimiento y explicabilidad. Con el mismo objetivo, la línea de visualización de datos y resultados busca técnicas para representar visualmente información compleja sobre el modelo, lo que facilita la comprensión y la toma de decisiones por parte de los analistas y los usuarios finales.

7 REFERENCIAS

[1] B. S. Caffo, F. A. D'Asaro, A. Garcez, y E. Raffinetti, «Editorial: Explainable artificial intelligence models and methods in finance and healthcare», *Front. Artif. Intell.*, vol. 5, p. 970246, ago. 2022, doi: 10.3389/frai.2022.970246.

[2] M. T. Ribeiro, S. Singh, y C. Guestrin, «"Why Should I Trust You?": Explaining the Predictions of Any Classifier». arXiv, 9 de agosto de 2016. Accedido: 4 de septiembre de 2022. [En línea]. Disponible en: http://arxiv.org/abs/1602.04938

[3] L. Edwards y M. Veale, «Slave to the Algorithm? Why a "right to an explanation" is probably not the remedy you are looking for», LawArXiv, preprint, nov. 2017. doi: 10.31228/osf.io/97upg.

[4] «Study: Big Data meets artificial intelligence», *BaFin*. https://www.bafin.de/SharedDocs/Downloads/EN/dl_bdai_studie_en.html (accedido 19 de noviembre de 2022).

[5] P. Gohel, P. Singh, y M. Mohanty, «Explainable AI: current status and future directions». arXiv, 12 de julio de 2021. Accedido: 10 de septiembre de 2022. [En línea]. Disponible en: http://arxiv.org/abs/2107.07045

[6] V. Turri, «What is Explainable AI?», *SEI Blog*. https://insights.sei.cmu.edu/blog/what-is-explainable-ai/ (accedido 19 de noviembre de 2022).

[7] T. Miller, «Explanation in Artificial Intelligence: Insights from the Social Sciences». arXiv, 14 de agosto de 2018. Accedido: 4 de septiembre de 2022. [En línea]. Disponible en: http://arxiv.org/abs/1706.07269

[8] B. H. Misheva, J. Osterrieder, A. Hirsa, O. Kulkarni, y S. F. Lin, «Explainable AI in Credit Risk Management». arXiv, 1 de marzo de 2021. Accedido: 8 de agosto de 2022. [En línea]. Disponible en: http://arxiv.org/abs/2103.00949

[9] H. Kvamme, N. Sellereite, K. Aas, y S. Sjursen, «Predicting mortgage default using convolutional neural networks», *Expert Syst. Appl.*, vol. 102, pp. 207-217, jul. 2018, doi: 10.1016/j.eswa.2018.02.029.

[10] T. Chen y C. Guestrin, «XGBoost: A Scalable Tree Boosting System», en *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ago. 2016, pp. 785-794. doi: 10.1145/2939672.2939785.

[11] A. Gramegna y P. Giudici, «SHAP and LIME: An Evaluation of Discriminative Power in Credit Risk», *Front. Artif. Intell.*, vol. 4, p. 752558, sep. 2021, doi: 10.3389/frai.2021.752558.

[12] N. Bussmann, P. Giudici, D. Marinelli, y J. Papenbrock, «Explainable Machine Learning in Credit Risk Management», *Comput. Econ.*, vol. 57, n.º 1, pp. 203-216, ene. 2021, doi: 10.1007/s10614-020-10042-0.

[13] P. Giudici y E. Raffinetti, «Shapley-Lorenz eXplainable Artificial Intelligence», *Expert Syst. Appl.*, vol. 167, p. 114104, abr. 2021, doi: 10.1016/j.eswa.2020.114104.

[14] C. Molnar, *Interpretable Machine Learning*. Accedido: 19 de noviembre de 2022. [En línea]. Disponible en: https://christophm.github.io/interpretable-ml-book/index.html
[15] Y. Zhou y M. Kantarcioglu, «On Transparency of Machine Learning Models: A Position Paper», p. 5.
[16] Z. C. Lipton, «The Mythos of Model Interpretability». arXiv, 6 de marzo de 2017. Accedido: 2 de octubre de 2022. [En línea]. Disponible en: http://arxiv.org/abs/1606.03490
[17] V. Belle y I. Papantonis, «Principles and Practice of Explainable Machine Learning», *Front. Big Data*, vol. 4, p. 688969, jul. 2021, doi: 10.3389/fdata.2021.688969.
[18] A. Barredo Arrieta *et al.*, «Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI», *Inf. Fusion*, vol. 58, pp. 82-115, jun. 2020, doi: 10.1016/j.inffus.2019.12.012.
[19] J. H. Friedman, «Greedy function approximation: A gradient boosting machine.», *Ann. Stat.*, vol. 29, n.º 5, oct. 2001, doi: 10.1214/aos/1013203451.
[20] A. Goldstein, A. Kapelner, J. Bleich, y E. Pitkin, «Peeking Inside the Black Box: Visualizing Statistical Learning With Plots of Individual Conditional Expectation», *J. Comput. Graph. Stat.*, vol. 24, n.º 1, pp. 44-65, ene. 2015, doi: 10.1080/10618600.2014.907095.
[21] ashutosh nayak, «Idea Behind LIME and SHAP», *Medium*, 22 de diciembre de 2019. https://towardsdatascience.com/idea-behind-lime-and-shap-b603d35d34eb (accedido 3 de noviembre de 2022).
[22] L. Breiman, «Random Forests», *Mach. Learn.*, vol. 45, n.º 1, pp. 5-32, oct. 2001, doi: 10.1023/A:1010933404324.
[23] G. Tolomei, F. Silvestri, A. Haines, y M. Lalmas, «Interpretable Predictions of Tree-based Ensembles via Actionable Feature Tweaking», en *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ago. 2017, pp. 465-474. doi: 10.1145/3097983.3098039.
[24] M. T. Ribeiro, S. Singh, y C. Guestrin, «Anchors: High-Precision Model-Agnostic Explanations», *Proc. AAAI Conf. Artif. Intell.*, vol. 32, n.º 1, abr. 2018, doi: 10.1609/aaai.v32i1.11491.
[25] «American Express - Default Prediction». https://kaggle.com/competitions/amex-default-prediction (accedido 6 de noviembre de 2022).