UNIVERSIDAD NACIONAL DE COLOMBIA

Asignación de cupo de crédito rotativo aplicada a nuevos clientes de una fintech Colombiana

Camilo Ernesto Medina González
Matemático

Universidad Nacional de Colombia
Facultad de Ciencias, Departamento de Matemáticas
Bogotá, Colombia
2023

Asignación de cupo de crédito rotativo aplicada a nuevos clientes de una fintech Colombiana

Camilo Ernesto Medina González

Tesis o trabajo de grado presentada(o) como requisito parcial para optar al título de:
Magíster en Ciencias - Matemática Aplicada

Director(a):
Francisco Albeiro Gómez Jaramillo
Ph.D. en Ingeniería - Sistemas y Computación

Línea de Investigación:
Aplicaciones del aprendizaje de máquina
Grupo de Investigación:
Computational Modeling of Biological Systems Research Group - COMBIOS

Universidad Nacional de Colombia
Facultad de Ciencias, Departamento de Matemáticas
Bogotá, Colombia
2023

Dedicatoria

A mi madre, que sin saber ni un poco que estoy haciendo, me ofreció su ayuda cada dia destinado para realizar este proyecto.

Agradecimientos

Un gran agradecimiento a Francisco Gómez, ya que gracias su orientación y ayuda me fue posible finalizar este proyecto, y quien siempre tuvo la disposición para ayudar y llevar a buen puerto esta idea. A todos los miembros de la organización, que aportaron de múltiples maneras, con conocimiento, con datos, con ideas, con interpretaciones, y un sin fin de herramientas que fueron usadas a lo largo de la construcción y concepción de esta tesis. A mis amigos, que siempre que lo necesiten sacaron el tiempo para tenderme una mano, de manera directa en muchos casos, o a veces siendo el pilar que me mantuvo en pie para hacer esto, un agradecimiento especial a quimiplones, que me dieron cada dia la moral para hacer las cosas. Agradecimiento a mi familia, por todo el amor, en especial a mi madre, que lo merece todo.

Asignación de cupo de crédito rotativo aplicada a nuevos clientes de una fintech Colombiana

Resumen

La asignación de cupos de crédito es uno de los grandes problemas a los que se enfrenta la industria financiera en Colombia actualmente. Esta tesis propone un nuevo modelo de asignación de cupos de crédito mediante la extracción de información financiera relevante de los registros históricos de solicitudes de crédito y estrategias de aprendizaje de máquina. En particular, se propone estimar directamente el segmento de cupo de crédito a partir de información económica, sociodemográfica y de riesgo crediticio de los clientes, en contraste con modelos previos que abordan el problema en dos pasos: 1) estimando la probabilidad de incumplimiento, y seguidamente, 2) el cupo de crédito con optimización. La información para ajustar los modelos de aprendizaje es extraída de históricos de aprobaciones de créditos realizadas por expertos. Seguidamente, modelos de clasificación multiclase son entrenados y comparados para resolver la tarea de asignación directa de cupos de crédito en segmentos. El modelo fue evaluado en datos crediticios de una entidad financiera colombiana en la tarea de asignación de cupos en microcréditos. Los resultados sugieren que el modelo propuesto supera los desempeños en asignación respecto a modelos lineales del estado del arte y presenta bajos niveles de sesgo. Se espera que este modelo de asignación de cupo pueda ser utilizado para la automatización de los procesos de originación de microcréditos, generando un impacto positivo en el acceso a los servicios financieros.

Palabras clave: Asignación de cupo, Aprendizaje de máquina, Finanzas, Crédito.

Allocation of revolving credit limit applied to new customers of a Colombian fintech

Abstract

Currently, the allocation of credit quotas is one of the major problems facing the financial industry in Colombia. Therefore, this dissertation proposes a new credit quota allocation model by extracting relevant financial information from historical credit application records and machine learning strategies. Indeed, it is suggested to directly estimate the credit quota segment from customers' economic, sociodemographic, and credit risk information, in contrast to previous models that tackle the problem in two steps: firstly, 1) by estimating the probability of default, and secondly, 2) by optimizing the credit quota. Thus, the information to fit the learning models is extracted from historical credit approvals by experts. Subsequently, multi-class classification models are trained and compared to solve the task of direct assignment credit quotas in segments. Finally, the model was evaluated in credit data from a Colombian financial institution in assigning microcredit quotas. The results suggest that the proposed model outperforms state-of-the-art linear models and presents low bias levels. Eventually, it is expected that this quota allocation model can be used to automate the microcredit origination processes, generating a positive impact on access to financial services.

Keywords: Credit allocation, Machine Learning, Finance, Credit.

Contenido

Lista de Figuras

1. Introducción

La distribución de la riqueza es una de las mayores preocupaciones del crecimiento económico, y puede ser entendida como un mecanismo de desarrollo. En este sentido, se requieren políticas públicas orientadas a mejorar la distribución de los ingresos (Marrugo Arnedo, 2013). Por esta razón, para los gobiernos alrededor del mundo, en particular en América Latina, la inclusión financiera se ha establecido como uno de los temas más relevantes de las agendas públicas (Franco, 2017).

La inclusión financiera se puede entender como la capacidad de un individuo, o cierto grupo de personas, para obtener acceso a productos financieros formales y poder, eventualmente, hacer uso de los mismos (Cano et al., 2014). Una de las principales estrategias para aumentar la inclusión financiera, comúnmente utilizadas en países en vías de desarrollo, son los microcréditos (Miled and Rejeb, 2015). Un microcrédito es un instrumento financiero consistente en la concesión de cantidades reducidas de capital a los individuos. Los microcréditos permiten, por ejemplo, financiar pequeños capitales de trabajo, la compra de activos fijos y/o cartera con condiciones financieras razonables. Generalmente, estos instrumentos financieros se enfocan en los segmentos más pobres de la sociedad, por ejemplo, pequeños emprendedores (Patiño, 2008). Hay evidencia que sugiere que el uso de microcréditos en países en vías de desarrollo aumenta la inclusión financiera ya que permiten que personas de escasos recursos tengan un fácil acceso al sistema financiero, mejorando su calidad de vida (Lattanzio Carrioni and Pinilla Jaramillo, 2013), no obstante, las ganancias asociadas al uso de estos

instrumentos son marginales (Banerjee et al., 2015). Los microcréditos, a diferencia de los créditos estándar, comúnmente no requieren de historiales de crédito extensos y verificables para su asignación. Por esta razón, en la ausencia de mecanismos de mitigación del riesgo de no pago para la asignación periódica del crédito, la definición del monto de capital a prestar, representa una tarea fundamental en el diseño de estos instrumentos financieros.

Un componente fundamental en la operación de los microcréditos son los sistemas de administración de riesgo crediticio (SARC). Un SARC es un sistema para el manejo de cuentas de crédito, desde la evaluación del riesgo y la determinación de monto de crédito, hasta él envió de facturas para cobrar los pagos. Los SARC comúnmente operan en las instituciones financieras, y su uso es obligatorio en la regulación colombiana (Hernández, 2004). Estos sistemas soportan las tareas relacionadas con la gestión de los procesos de otorgamiento, recuperación, seguimiento y control del crédito. Los SARC son particularmente críticos para la definición de los procesos de otorgamiento o procesos de originación, que corresponden a dos de las tareas más comunes abordadas al inicio del proceso de asignación de cualquier crédito.

1.1. Proceso de asignación de un crédito

Una tarea fundamental a resolver dentro de los procesos de originación es el denominado filtrado de clientes potenciales, es decir, la determinación de potenciales beneficiarios de los créditos. Este filtrado es necesario, ya que no todas las personas cuentan con la capacidad financiera para hacer frente a sus deudas, bien sean las actuales o las deudas que van a adquirir en un futuro. Una estrategia común para abordar esta tarea consiste en la clasificación de los créditos de las personas como potenciales buenos o malos créditos (Ayling, 2018). Dicha clasificación comúnmente se realiza siguiendo el criterio de expertos, y más recientemente, utilizando datos asociados a los clientes. Este proceso de perfilamiento del

cliente como bueno o malo recibe el nombre de scoring crediticio (Ochoa et al., 2010).

El problema del scoring crediticio se ha abordado desde los años setenta, aunque su uso se generalizó a partir de los años noventa. Actualmente, este problema cuenta con una extensa base teórica y práctica (Gutierrez Girault, 2007). Inicialmente, este scoring era realizado por analistas que ayudaban a tomar decisiones sobre a qué clientes se les otorga o no un crédito dado (González Fernández, 2016). Sin embargo, en los últimos años el área ha tenido avances significativos, gracias al uso técnicas de inteligencia artificial, así como a la consolidación de conjuntos de datos cada vez más informativos sobre la capacidad de endeudamiento y de pago de los clientes (Rodríguez Martínez and Serna García, 2020). Gracias a esto, actualmente es posible automatizar el cálculo de este scoring para muchas empresas en la industria financiera (Espin-García and Rodríguez-Caballero, 2013).

1.2. Scoring crediticio y asignación del cupo

El scoring crediticio corresponde a la primera etapa del proceso de originación, y consiste en asignación de un indicador a cada cliente relacionado con la capacidad de pago de un crédito. Este score comúnmente es utilizado para perfilar un cliente como bueno o malo. Sin embargo, una vez el crédito haya sido aprobado, es necesario especificar las condiciones puntuales del crédito, en particular, el monto a prestar o cupo. Es importante anotar que, a diferencia del scoring crediticio, actualmente este último proceso se realiza de forma manual en muchas instituciones financieras. La estrategia más común para la definición del monto crediticio es la definición de un conjunto reglas de negocio. Naturalmente, esta estrategia manual depende críticamente del nivel de experticia del tomador de decisiones y de la estructura de las reglas de negocio, las cuales no necesariamente, se adaptan a las particularidades de los clientes, ni a las necesidades financieras del negocio. Es importante anotar, que la tarea de la asignación de cupos para los créditos, puede impactar negativamente las finanzas de

la organización, incrementando el riesgo financiero en los casos en los que se sobreestime la capacidad de pago por parte del cliente, o afectando los ingresos en los casos en los que esta se subestime (Badel Coronell and Herrera Valdez, 2013).

Naturalmente, el problema de asignación de cupos está relacionado con el problema de scoring crediticio, puesto que al cambiar los cupos de crédito también puede cambiar el riesgo de no pago. Por esta razón, los procesos se vuelven indisociables ya que no se puede hacer referencia al segundo sin tener en consideración el primero. Realizar un análisis de la relación del scoring con la asignación de cupos, sin tener en cuenta el tipo de crédito sobre el cual se hace el análisis, resultará insuficiente y descontextualizado.

1.3. Microcréditos y créditos rotativos

Dentro de los tipos de crédito más populares en el contexto colombiano, destacan los créditos rotativos. Estos créditos se caracterizan por que el monto es previamente aprobado por el acreedor, se pagan comisiones e intereses únicamente por el monto y por el tiempo utilizado. Adicionalmente, los microcréditos permiten desembolsos casi automáticos, mediante la firma de pagarés o con un mecanismo parecido al de una tarjeta de crédito convencional. Estos créditos también son revolventes. Es decir, vuelven a tener vigencia con arreglo a las condiciones estipuladas originalmente después de haber sido amortizado (Vázquez, 2012). Finalmente, el poseedor tiene la misma disponibilidad de cupo o compra mediante abonos al saldo de la cuenta (Granados and Quijano, 2016).

Los créditos rotativos facilitan el acceso a la financiación al no tener que pasar por el proceso de originación del crédito múltiples veces, sino una única vez para poder seguir accediendo de manera periódica a estos préstamos. En este sentido, los créditos rotativos tienen una mayor facilidad de acceso frente a otras alternativas del sector financiero tradicional, las cuales en su mayoria cuentan con procesos de originación largos, tardíos y complicados para

los usuarios.

1.4. Industria financiera

En los últimos años, el uso de tecnologías de la información, y en particular basadas en inteligencia artificial y aprendizaje de máquina, ha impactado positivamente a pequeñas y medianas empresas. Estas nuevas tecnologías han facilitado el desarrollo de nuevas cadenas de valor, plazos de entregas más cortos y una mejor calidad de los productos ofrecidos (Lu et al., 2022). Este crecimiento se potenció por la pandemia del COVID-19 impactando de manera positiva a sectores altamente tecnológicos como el de las fintech, permitiendo el desarrollo de productos financieros más simples, rápidos, cómodos, eficaces para los clientes, y aumentando sus niveles de cobertura, gracias a sus difusión en formatos digitales y remotos (Rodríguez Martínez and Serna García, 2020).

Igualmente, las opciones financieras promocionadas por las fintech, son menos costosas, altamente escalables y actualmente son las formas más prometedoras para llenar los vacíos que dejan las instituciones financieras tradicionales, por ejemplo, al acceso a financiamiento en los hogares. (Yang and Zhang, 2022).

1.5. Políticas de asignación de cupos

Es importante destacar que dependiendo de la organización, la política para la asignación de cupos de crédito puede estar enfocada en generar ganancias o en reducir las pérdidas económicas sobre todos los créditos desembolsados. Dado que los aumentos excesivos del cupo de crédito impactan de manera negativa las tasas de retorno de los créditos, los préstamos ofrecidos a futuro pueden llevar a los clientes a dejar solicitar créditos (Shema, 2022). Por esta razón, es necesario enfocar la asignación de los cupos de crédito a la reducción de las

pérdidas.

El ofrecimiento de microcreditos por parte de las fintech, puede ayudar a mejorar la inclusión financiera, en específico, ofrecer microcréditos rotativos, haciendo una correcta asignación de cupo de crédito de los mismos, y enfocándose en la reducción de pérdidas.

En este contexto, el problema de investigación a abordar será: *la construcción de un modelo matemático de asignación de cupos, enfocado en reducción de pérdidas, para el sector de las fintech en Colombia.*

1.6. Justificación

América Latina presenta en la actualidad un rezago notorio frente a competitividad, productividad y crecimiento, tal como la CEPAL (Comisión Económica para América Latina y el Caribe) expone en su informe (Pollack and García Hurtado, 2004). Una de las causas de este rezago es la falta de acceso de las empresas y personas naturales a financiamientos para sus proyectos y negocios (Pollack and García Hurtado, 2004). Para lograr un crecimiento con equidad es posible mejorar el acceso de empresas y personas naturales a diferentes instrumentos de financiación. Esto se puede lograr, según la CEPAL (Pollack and García Hurtado, 2004), a partir de la modificación de los marcos regulatorios, inyectando liquidez a la banca privada, y asegurando una máxima cobertura y accesibilidad al sistema financiero a través de créditos, en especial a las personas que recién se incorporan a este sistema. En este contexto, el foco de investigación de este proyecto, la asignación de cupos de crédito, es de alto interés para la industria financiera. Una correcta asignación de cupos asegura que las personas que ingresan al sistema financiero puedan acceder bajo las mejores condiciones a los préstamos, aumentando la accesibilidad al sistema. Un modelo que asigne correctamente un cupo de crédito a los usuarios que ingresan a este sistema es fundamental para que este rezago sea cada vez menor.

El proceso de asignación de cupos, para muchas empresas, representan una ventaja diferenciadora (Redroban Bermudez, 2016). Una mejora en la efectividad y capacidad de respuesta en este proceso, permite incrementar ventas y captar un mayor mercado. Por lo tanto, una correcta asignación de cupos puede representar una ventaja competitiva, y puede fortalecer la cadena de valor de estas empresas. Adicionalmente, dentro de los activos en los estados financieros de las pequeñas y medianas empresas, en general, las cuentas por cobrar son el activo más grande y líquido (Gómez González, 2010). Por esta razón, soportar los requerimientos de liquidez será mucho más sencillo si se cuenta con un adecuado manejo de

la cartera de cuentas por cobrar. En este sentido, la correcta asignación de cupos no solo representa una ventaja competitiva, sino también parte fundamental del manejo de cartera de una empresa, y en consecuencia, parte importante de la buena salud financiera de la misma. Adicionalmente, dado que la cartera de una empresa no es más que la transferencia de unos bienes o préstamo de servicios a crédito, la misma constituye el activo más líquido después del capital disponible (Badel Coronell and Herrera Valdez, 2013). Por esta razón, la correcta gestión de la asignación de cupos de créditos y la determinación de los plazos de pago de los mismos, es importante para la eficiente recuperación del capital invertido o prestado, y tiene un gran impacto en la rentabilidad de una empresa.

En conclusión, recuperar el capital invertido se convierte en un problema de suma importancia para la rentabilidad, las finanzas, la ventaja competitiva y la cadena de valor de una empresa. Por estas razones, esta investigación se enfocará en el riesgo crediticio, primando la recuperación del capital invertido, antes que el rendimiento sobre el mismo.

1.7. Objetivos

En este contexto y partiendo del problema de investigación identificado, esta tesis aborda los siguientes objetivos:

1.7.1. Objetivo General

- Desarrollar un modelo de asignación de cupos de crédito enfocado en riesgo y aplicarlo en el proceso de originación de crédito.

Con el fin de cumplir con el objetivo general, se plantean los siguientes objetivos específicos:

1.7.2. Objetivos Específicos

- Formular un modelo de asignación de cupos de crédito enfocado en reducir pérdidas basado en técnicas de aprendizaje de máquina supervisado.
- Implementar un modelo computacional de asignación de cupos de crédito enfocado en reducir perdidas en el proceso de originación de crédito de la empresa.
- Validar un modelo de asignación de cupos de crédito enfocado en reducir pérdidas utilizando datos históricos de clientes de una entidad crediticia.

2. Estado del arte

Para abordar el estado del arte y las secciones posteriores, se asumen ciertos conocimientos básicos sobre probabilidad, finanzas y otros temas. En el anexo A se pueden encontrar distintas definiciones de estos temas en caso de ser necesarias.

El problema de construcción de modelos de asignación de cupos se ha abordado previamente. El estado del arte relacionado puede estructurarse a partir de 5 dimensiones principales:

1. Soporte basado en teoría financiera.
2. Consideración de la probabilidad de incumplimiento.
3. Uso de múltiples clases para la caracterización de los clientes.
4. Interpretabilidad de los resultados.
5. Enfoque en reducción de pérdidas.

A continuación, se hará una revisión de algunas de las aproximaciones más relevantes asociadas al problema. Teniendo en cuenta las dimensiones anteriormente descritas, se propone dividir las propuestas de solución encontradas en la literatura de 2 maneras, las que son basadas en la probabilidad de incumplimiento, y las que no.

2.1. Estrategias basadas en la probabilidad de incumplimiento

Una primera familia de modelos son los que basan su formulación en la probabilidad de incumplimiento, en particular, (Herga et al., 2016) utiliza datos recolectados sobre cientos de empresas Europeas para estimar la probabilidad de incumplimento a través de una regresión logística.

Este modelo surge a partir de ciertas variables que describen a los potenciales clientes, en un vector de características, y buscan predecir en primera instancia, una probabilidad de incumplimiento, para posteriormente usar estas probabilidades para el cálculo del cupo para cada cliente. A partir de estos datos se creó un modelo que estimaba dicha probabilidad. Inicialmente, para cada uno de los elementos del vector de características, se crean n divisiones, escogidas a partir del juicio del experto del dominio, y se contabilizan cuántos datos de cada división fueron clasificados como impagos, y cuantos como pagados. A partir de estos conteos, se genera el WOE (Weight of Evidence) para cada división, el cual no es más que un valor numérico que muestra qué tan relacionada está tanto el elemento del vector de características, como el rango estudiado, respecto a la variable que se pretende predecir. Finalmente, se transforma cada indicador financiero, asignando el WOE para su división correspondiente. A partir de estos datos transformados, se ajusta un modelo de regresión logística, y se crea la probabilidad de incumplimiento para cada uno de los datos en el conjunto.

Para la construcción del cupo, se formula un problema de optimización basado en el Valor en Riesgo (VaR), y el Valor en Riesgo Condicional (CVaR), métodos para cuantificar la exposición al riesgo. Mientras el primero habla de las posibles pérdidas para una probabilidad dada, el segundo se enfoca más en la pérdida para los peores escenarios. En el problema de optimización se busca maximizar el retorno de dinero de estos préstamos, sujetos a ciertas

restricciones creadas a partir del VaR y CVaR, y restricciones propias del contexto del problema de la propuesta.

Dicha aproximación está bien fundamentada desde el punto de vista teórico, haciendo uso del VaR y CVaR, adicionalmente provee probabilidades de incumplimiento calculadas sobre el conjunto de datos. Se obtine una respuesta del modelo sobre un conjunto de datos continuo y es una solución sumamente interpretable al solo hacer uso de la probabilidad de incumplimiento y de una regresión logística.

Dado que se basan en el Var y CVaR, está enfocado en reducir las pérdidas, pero se evidencia que el modelo en algunos casos asigna el menor o mayor cupo posible, y en estos casos solo algunos clientes obtienen parte del mayor cupo posible, por lo cual su nivel de predicción puede llegar a ser pobre. Es necesario validar la solidez de esta propuesta en el contexto propio de nuestro problema.

Por otro lado, So Young Sohn (Sohn et al., 2014), aborda el problema utilizando el beneficio total neto de una empresa, el cual, a su vez, es dependiente del beneficio neto de cada préstamo. Para calcular este beneficio neto, se hace uso del costo y del ingreso generado por cada crédito, y estos mismos dependen del balance del crédito, la probabilidad de incumplimiento y de valores propios de cada línea de crédito. La probabilidad de incumplimiento se calcula utilizando una regresión logística sobre ciertas variables que representan la actividad de cuenta del cliente, una vez calculada la misma para cada cliente, se utiliza un árbol de regresión para clasificar a los clientes según su cupo.

Además, la probabilidad de incumplimiento debe ser reajustada para cada grupo de clientes obtenido en la clasificación, para obtener una probabilidad de incumplimiento más acorde a un caso de uso real. Para este ajuste se hace uso de la probabilidad de incumplimiento del tipo de crédito revisado, el promedio de la probabilidad de incumplimiento del grupo en cuestión y el promedio de la probabilidad de incumplimiento de todo el grupo.

Finalmente, se utiliza un algoritmo genético sobre las probabilidades de incumplimiento ajustadas para cada grupo, y el árbol de regresión creado, para determinar el límite de crédito que maximiza el beneficio neto total, el cual cambia para cada mes.

La propuesta de Sohn se basa mucho en la teoría financiera, al hacer uso de conceptos como el beneficio total neto y el beneficio neto de cada línea de crédito, además de estar basada en la probabilidad de incumplimiento del cliente. Este modelo permite el uso de múltiples clases, ya que se utiliza un árbol de regresión para clasificar a los clientes según un rango de cupo. Pero la propuesta se queda corta respecto a la interpretabilidad, ya que al aplicar un algoritmo genético sobre un árbol regresor, los resultados obtenidos, sin importar lo buenos que sean, no son interpretables (Michalewicz, 1996). Adicionalmente su construcción se basa únicamente en las ganancias de cada crédito, por tanto, su enfoque respecto a las perdidas económicas es nulo, y por su planteamiento no muestra signos de poderse adaptar a un modelo enfocado en reducir perdidas.

Uttiya (Paul and Biswas, 2017) et al. propone utilizar la teoría de la decisión Bayesiana para predecir los cupos de crédito, ya que cuando la misma se aplica conduce a la minimización de la probabilidad de errores de clasificación. A esto se suma el concepto del puntaje Fico, un puntaje continuo, sumamente usado en la industria financiera, y que es considerado como un indicador robusto de riesgo, y el porcentaje del uso de cupo de crédito.

Adicionalmente utiliza lógica difusa, también conocida como lógica borrosa, que permite el uso de valores de verdad distintos de verdadero o falso; la combinación de esto con la teoría de la decisión Bayesiana permite la creación de un modelo de asignación de cupos que se ajusta automáticamente a cambios en los datos para que el error en la clasificación no aumente abruptamente.

Esta aproximación también se basa en la probabilidad de incumplimiento. Pero respecto a los otros 4 aspectos a revisar, se queda corto en todos. El modelo no muestra buenas bases

financieras detrás para su construcción, sino más bien bases estadísticas y lógicas bastante potentes. Además, el modelo solo permite el uso de 2 clases para la clasificación, por lo que se queda bastante atrás en comparación a las otras propuestas ya revisadas, esto sin contar que su enfoque no está en reducir perdidas, sino en aumentar las ganancias. Por último, es importante destacar que el uso de combinar la decisión bayesiana con lógica difusa, 2 ramas bastante desconocidas para los expertos en el área, hace que la interpretación de este modelo sea prácticamente nula.

2.2. Estrategias no basadas en la probabilidad de incumplimiento

Otra manera de abordar el problema es la propuesta por Haimowitz (Haimowitz and Schwarz, 1997). En este caso se hace un acercamiento al problema de asignación de cupos de crédito, desde el problema de scoring crediticio. En particular, se deja en claro que el problema del scoring crediticio es el hecho de solo calificar a los clientes como buenos o malos, además de no tratar el cupo como una variable independiente y endógena, lo cual si se logra en el planteamiento de Haimowitz.

En el modelo Haimowitz se propone la creación de un marco de trabajo para la asignación de cupos, separado en 3 partes. En la primera parte se toman los datos históricos de los clientes de varios meses y se agrupan en K conjuntos, basado en el comportamiento de pago. Además, se introduce el concepto de valor presente neto, relacionado con cuál es el valor de una inversión en el tiempo donde la misma inicia, lo cual permite evaluar que tan rentable es una inversión en cuestión.

Una vez realizado esto, se utiliza un árbol de decisión, más específicamente, un árbol regresor, para calcular la probabilidad de que un cliente esté en un determinado grupo de los

creados en la primera parte de este flujo. Por último, se utilizan estas probabilidades junto a la tasa interna de retorno para calcular el valor presente neto de la inversión.

El modelo propuesto puede adaptarse para un modelo de clustering determinado, y para cualquier modelo de predicción que utilice probabilidades. Es importante destacar que el uso del valor presente neto de la inversión y la tasa interna de retorno le da mucho fundamento financiero a la solución, además de permitir el uso de K clases que se pueden determinar previamente, siendo una salida multiclase.

La interpretabilidad de la propuesta va a depender netamente de la selección del modelo en el marco de trabajo. No obstante, este modelo no se enfoca en riesgo, sino en el aumento de las ganancias. Esta limitación se puede abordar, eliminando el valor presente neto, y convirtiendo el mismo en un valor esperado, lo cual lo convertiría en un modelo enfocado netamente en reducir perdidas financieras.

3. Asignación de cupo en la organización

3.1. Sistema de Administración del Riesgo Crediticio

Los cambios que el mundo sufrió durante el siglo XX, época marcada por grandes acontecimientos históricos, tales como la gran depresión, la guerra fría, la primera guerra mundial, y la segunda guerra mundial, provocaron transformaciones socioeconómicas profundas, en especial, transformaciones que abrieron los mercados financieros acentuados por el fenómeno de la globalización. (Rodríguez and Hernández, 2008)

A mediados de los años 70, se empiezan a ver cambios sustanciales en el manejo de las instituciones financieras, que obligan a que se creen nuevas regulaciones motivadas por estos cambios. Después del colapso en 1974 de Bankhaus Herstatt en Alemania, y del Banco Nacional Franklin en Estados Unidos, el Banco Internacional de Pagos (BIS), crea el Comité de Supervisión Bancaria de Basilea, con el objetivo de formular distinto tipo de recomendaciones para la regulación de instituciones financieras a nivel mundial, y poder contrarrestar las inestabilidades propias del mercado. (Ochoa et al., 2010)

Para 1988 el Comité de Supervisión Bancaria de Basilea ya había formulado un primer pliego de recomendaciones, el Acuerdo de Capitales de Basilea, también conocido como Basilea I. En 1999 el Comité se reúne de nuevo, para la creación del acuerdo Basilea II, que sería publicado en su totalidad sólo hasta el año 2006.

Las recomendaciones del acuerdo Basilea II se resumen en 3 grandes pilares:

- Requisitos de capital mínimo: Cubrimiento del capital en riesgo.
- Proceso de examen supervisor: Debe existir un ente supervisor para las entidades financieras.
- La disciplina del mercado: El acceso a la información de las entidades financieras debe ser transparente.

(Ochoa et al., 2010)

Es así que siguiendo la tendencia mundial, fomentada por los acuerdos de Basilea, en Colombia, la Superintendencia Bancaria reglamenta la creación del Sistema de Administración de Riesgo Crediticio (SARC). Esto se reglamenta a través de 2 documentos, el primero, la Carta Circular 31 de 2002, y la Circular externa 11 de 2002, donde se definen los lineamientos básicos para que se implementen mediciones de riesgo en las entidades financieras colombianas. (Hernández, 2004)

Es importante aclarar que la Carta Circular 31 de 2002, tuvo modificaciones y adiciones a lo largo de los años, a través de las siguientes Circulares:

- Circular Externa 052 de 2004
- Circular Externa 035 de 2005
- Circular Externa 035 de 2006
- Circular Externa 029 de 2007
- Circular Externa 039 de 2007
- Circular Externa 010 de 2008

Por tanto, de ahora en adelante se hará referencia de manera directa a la Circular Externa 010 de 2008, al ser la modificación más reciente de la regulación.

3.2. Elementos que componen un SARC

Tal como indica la Circular Externa 010 de 2008 (financiera de Colombia, 2008), las entidades reguladas por la superintendencia financiera en Colombia, tienen la obligación de hacer la evaluación del riesgo crediticio adoptando un SARC en la organización. El mismo debe contar con 5 componentes básicos:

- Políticas de administración del riesgo crediticio
- Procesos de administración del riesgo crediticio
- Modelos internos o de referencia para la estimación y la cuantificación de las pérdidas esperadas
- Sistema de provisiones para cubrir el riesgo crediticio
- Procesos de control interno

Para el caso de las políticas de administración del riesgo crediticio, debe existir una junta directiva o un consejo administrativo en la institución, y la misma debe definir bajo qué criterios se evaluará, calificará, controlará y asumirá el riesgo crediticio. Estas políticas deben incluir, como mínimo, los siguientes aspectos:

- Estructura organizacional: Se debe desarrollar una estructura organizacional clara, con responsabilidades asignadas a distintas áreas o personas, además de generar reglas internas para evitar conflictos de intereses y mantener la reserva de la información de los clientes.
- Límites de exposición crediticia y de pérdida total: Dentro de las políticas hay que incluir algunas pautas generales para la limitación de la exposición, cupos adjudicados y límites de concentración, para los créditos totales e individuales.
- Otorgamiento del crédito: Se deben definir las características básicas para que un sujeto sea considerado para un crédito, y el nivel de adjudicación de crédito del mismo.

- Garantías: Se deben generar criterios claros para la exigencia y aceptación de garantías para cada crédito.
- Seguimiento y control: Deben existir políticas claras para hacer seguimiento y control de los portafolios, incluyendo clasificar y recalificar las operaciones realizadas en el proceso de otorgamiento.
- Construcción de provisiones: Se debe prever el cubrimiento de las pérdidas mediante provisiones generales e individuales.
- Capital económico: Debe existir un nivel de patrimonio para cubrir las pérdidas no esperadas.
- Recuperación de cartera: Se deben desarrollar políticas que permitan enfrentar incumplimientos en las obligaciones, con objeto de minimizar las pérdidas.
- Políticas de las bases de datos que soportan el SARC: Deben existir políticas claras para estas bases de datos, además de cumplir con ciertas normativas, como una antigüedad mayor a 7 años.

Adicional a todo lo anteriormente mencionado, existen 3 procesos básicos que se deben incluir en el SARC, los cuales corresponden al otorgamiento, seguimiento y control, y recuperación, donde existen criterios de contenido mínimo para cada uno de estos procesos. Dado que el proceso de asignación de cupo hace parte del proceso básico de otorgamiento, sólo se hará énfasis en este mismo.

3.3. Proceso de otorgamiento o originación en Colombia

El otorgamiento de crédito de las entidades debe basarse en el conocimiento del sujeto de crédito o contraparte, de su capacidad de pago y de las características del contrato a celebrar entre las partes. (financiera de Colombia, 2008) Existen algunos parámetros mínimos a tener en cuenta para hacer este otorgamiento La primera es entregar al potencial deudor del crédito información comprensible y legible sobre su crédito, previo a la aceptación del mismo, y debe incluir como mínimo:

- Información completa sobre la tasa de interés, indicando su equivalente en tasa efectiva anual, además de la periodicidad de los pagos, el tipo de pago (fijo o variable) e información explícita de cualquier cambio que pueda existir en la tasa.
- El capital sobre el cual se aplicarán los intereses.
- Comisiones y recargos.
- Plazo de préstamo, con periodos muertos o de gracia descritos.
- Condiciones de prepago.
- Derechos del acreedor en caso de incumplimiento.
- Derechos del deudor.
- Cualquier información relevante adicional sobre el crédito.

Luego, dentro del proceso de otorgamiento, se debe hacer una selección de las variables que permiten, con mayor significancia, separar a los potenciales clientes de la entidad según el perfil de riesgo que asume la entidad. Estas variables deben ser determinantes para todo el proceso de otorgamiento, y adicionalmente, para el seguimiento de los créditos.

Posteriormente, entra el proceso de evaluación de capacidad de pago del deudor. Esta evaluación es primordial para determinar la probabilidad de incumplimiento del crédito, y adicionalmente, para el cupo asignado al mismo. Para hacer el análisis se requiere que por

lo menos, se analice lo siguiente:

- Flujo de ingresos y egresos.
- Solvencia del deudor, a través de distintos indicadores socioeconómicos.
- Información del cumplimento de anteriores y actuales obligaciones.
- Reestructuraciones realizadas en el crédito, y sus respectivas características.
- En el caso de entidades públicas territoriales, se evaluará la capacidad de pago basado en la reglamentación vigente.
- Riesgos financieros a los que se expone el deudor, como cambios de tasas, efectos sobre la moneda, volatilidad de tasas de cambio o riesgos operacionales del mismo.
- En el caso de microcréditos, debe existir una metodología adecuada al riesgo del deudor, y que compensen la deficiencia o falta de información sobre los sujetos.

Por último, deben existir garantías sobre las operaciones realizadas. Estas son necesarias para calcular las pérdidas esperadas y el nivel de cubrimiento sobre estas mismas pérdidas. Estas garantías pueden ser de varios tipos, tales como prendarias, hipotecarias, pignoraciones, entre otras, y en todo caso, se debe evaluar como mínimo la naturaleza de la garantía, la liquidez de la misma, el valor y la cobertura total en caso de un evento de no pago.

3.4. Asignación de cupo dentro del proceso de originación

Tal como indica Gómez González (2010), para la buena administración del riesgo crediticio es necesario que las empresas determinen el nivel máximo de riesgo y exposición para cada cliente, lo que será representado a través de la asignación de un cupo máximo de crédito acorde a su capacidad de pago. Adicionalmente, la pérdida esperada de un crédito, según la superintendencia financiera financiera de Colombia (2008), debe calcularse de la siguiente manera:

$$PE = [PI] \cdot [EA] \cdot [PEA] \tag{3-1}$$

Donde:

- $PE = Pérdida\ esperada$
- $PI = Probabilidad\ de\ incumplimiento$
- $EA = Exposición\ del\ activo$
- $PEA = Pérdida\ esperada\ de\ valor\ del\ activo\ dado\ el\ incumplimiento$

Por tanto, la pérdida esperada de un portafolio estará directamente relacionada con la exposición del activo, o en este caso, el cupo máximo asignado a cada línea de crédito que hace parte de este portafolio. Además la probabilidad de incumplimiento está relacionada a dicha exposición, dado que a mayor exposición, se espera una mayor probabilidad de incumplimiento, al tener que hacer uso los clientes de mayor parte de sus activos para hacer el pago de sus obligaciones, reduciendo su capacidad de pago. Por ende, es normal esperar que tanto la probabilidad de incumplimiento como la asignación del cupo máximo sean calculados por la regulación de la superintendencia financiera, en la fase de otorgamiento.

3.5. Estado actual de la asignación de cupo en la organización

Actualmente la organización cumple con todos los requisitos solicitados por la superintendencia financiera, relacionado con la existencia de un SARC en la organización, aunque en este documento solo se hará énfasis en las políticas y regulaciones relacionadas con la fase de otorgamiento de crédito. En primer lugar al deudor se le hace entrega de toda la información relacionada con su crédito en el momento de realizar la solicitud, de manera clara y legible.

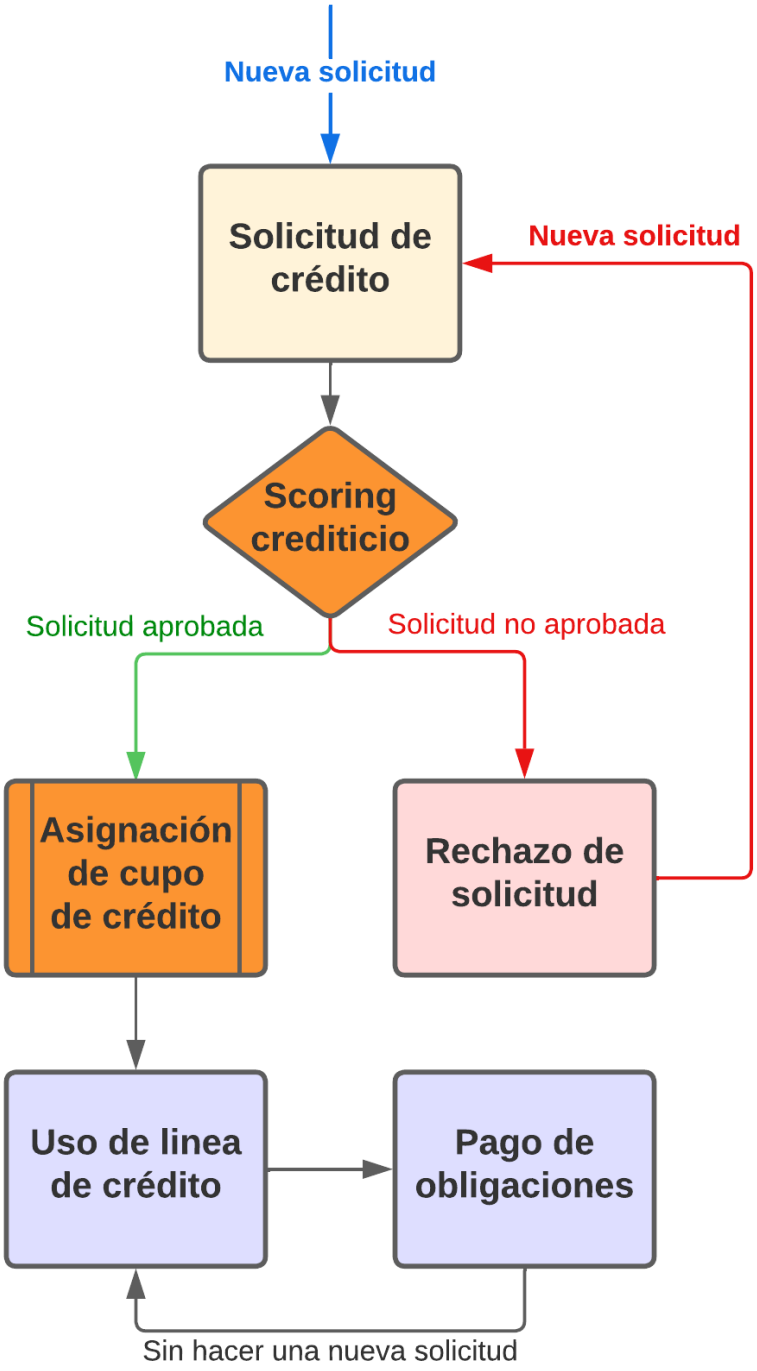
Respecto a las garantías, la organización cuenta con 2 alianzas estratégicas para tal fin, la primera con el FNG (Fondo Nacional de Garantías), y la segunda con el FGA (Fondo de Garantías). Estos prestan el servicio de garantía de deudas, cubriendo parte del capital de un cliente en caso de incumplimiento del mismo con sus obligaciones financieras adquiridas con la organización.

Adicional a la base de datos propia de la organización, se cuenta con el apoyo de 2 organizaciones externas: Datacredito: Es el buró de crédito más grande de Colombia, es una institución que se encarga de recopilar datos de personas naturales y jurídicas, relacionadas con su comportamiento crediticio. Recopila entre otras cosas, datos de comportamiento de pago, deudas, capacidad de endeudamiento, y da un puntaje crediticio basado en estos datos. Mareigua: Es el primer buró de ingresos y empleo de Colombia, al igual que datacredito, recopila datos de personas naturales y jurídicas, pero en este caso relacionados con ingresos y empleo.

Utilizando todos estos datos, se hace el cálculo de un score de riesgo para cada cliente. Al tratarse de microcréditos, tal como la superintendencia financiera lo requiere, el modelo de scoring debe estar adaptado a la falta de información sobre los potenciales clientes, además de hacer uso de una metodología clara y adaptada al riesgo del deudor. A partir de esos scores se genera un decil de riesgo, y con el mismo, se decide si un cliente cumple o no

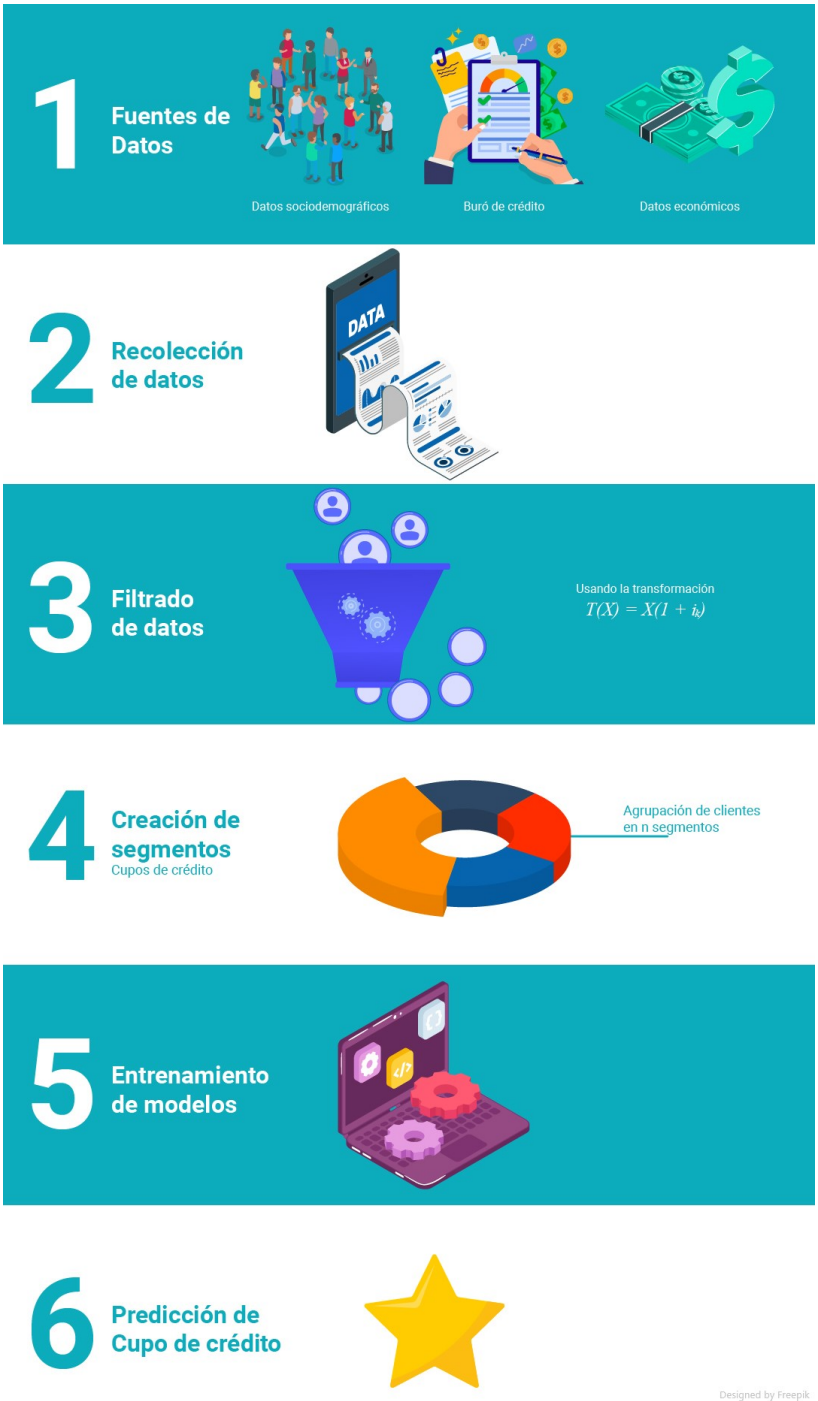
el perfil de riesgo creado en el SARC de la organización. Una vez aprobada la solicitud de crédito, también motivado por la regulación de la superintendencia, se conoce gracias a algunos análisis previamente realizados, cuales son las variables con mayor significancia para la asignación de este decil de riesgo. Utilizando las mismas, y con el apoyo de un experto del dominio, se hace la asignación del cupo al sujeto. Es decir, el proceso de asignación de cupo no es automático, sino que depende de la decisión de algún experto de negocio. Este último paso es el que se busca cambiar, y encontrar un modelo que automáticamente, haciendo uso de todos los datos disponibles sobre el cliente, pueda asignar un cupo máximo de crédito, sin necesidad del apoyo del experto, buscando que este experto únicamente valide, más no cree, la regla o modelo de asignación de cupos.

Figura 3-1. Diagrama de flujo de las solicitudes de crédito en la organización.



4. Materiales y métodos

Figura 4-1. Diagrama de la propuesta realizada en este documento



En la Figura **4-1** se describe la propuesta que se llevará a cabo en este documento. Se cuentan con tres fuentes de datos, correspondientes a datos económicos, sociodemográficos y de buró de crédito, sobre un conjunto de solicitudes, las cuales serán recopiladas, y luego procesadas. En el proceso se harán transformaciones a las variables económicas, tales como el ajuste de los valores por inflación y las creaciones de segmentos de cupo de crédito. Eso último es importante, ya que se convertirá en un problema de clasificación gracias a esta transformación. Posteriormente se entrenarán y escogerá un modelo basado en la métrica de ROC-AUC score. Luego, se hará una predicción del cupo asignado basado en las variables recopiladas y procesadas de las tres fuentes de datos.

4.1. Datos

4.1.1. Descripción de los datos

Dado que la organización cuenta con una experiencia superior a los 10 años ofreciendo microcréditos en Colombia, se ha logrado recopilar datos de más de 7 millones de solicitudes de crédito, para aproximadamente 2.8 millones de clientes a nivel nacional. La antigüedad de esta base de datos es sumamente importante para estar bajo los estándares solicitados por la Superintendencia Financiera, ya que la misma solicita una antigüedad mínima para estas bases de datos propias en 7 años.

Es importante destacar que dado que esta base de datos no es de acceso público, si no que es confidencial y de uso exclusivo de la organización, y para evitar dar a conocer a cualquier competidor en la industria que datos se utilizan de los clientes para el proceso de otorgamiento de los créditos, solo se hará una descripción superficial de los mismos, divididos en 3 grupos.

- Datos sociodemográficos del cliente: Gracias al formulario de inscripción y solicitud de

crédito, de entrada se cuenta con numerosa información sobre el cliente, tal como edad, género, ciudad y departamento de la solicitud, dirección, entre muchos otros.

- Datos económicos del cliente: Dentro del formulario anteriormente mencionado, también se encuentra información económica del cliente, como salario. La información que falte o que el cliente entregue de manera fraudulenta, es contrastada en un buró de ingresos y empleo, validando que el salario, ocupación, entre otros datos suministrados por el cliente, correspondan a la realidad económica del mismo.

  Es importante aclarar que la información suministrada por el buró prima sobre la suministrada por el cliente, en búsqueda de evitar que se falsifiquen datos intentando mejorar la posibilidad de acceder a una línea de crédito.

- Datos de buró de crédito: Tal como la superintendencia financiera lo requiere, se toman datos de buró de crédito para evaluar las solicitudes de crédito. La información entregada por este buro es cuantiosa, pero en algunos casos, resulta ser excesiva, por tanto se toman solo los datos que se consideran necesarios para la evaluación de riesgo del cliente según la organización.

4.1.2. Recolección de los datos

Los datos fueron recolectados desde inicios de 2012 hasta finales del año 2022. Se descartaron los datos de una solicitud en cualquiera de los siguientes casos:

- Cualquier reuso de la línea de crédito, es decir, cualquier desembolso generado después del primero.
- Cualquier solicitud que llegará a ser descartada por el sistema de scoring crediticio.
- Cualquier solicitud que fuera calificada como fraude, incluso si la misma llegó a ser desembolsada.
- Cualquier solicitud que no contara con más de 180 días de haber sido realizada. Se

necesitan al menos 180 días para que un desembolso llegue a su maduración según las reglas de la organización.

Por tanto, al final se recolectaron los datos de aproximadamente 260.000 solicitudes, las cuales corresponden únicamente a primeras solicitudes, a las cuales se les desembolso, que no fueron fraudulentas, y adicionalmente, que tuvieron la cantidad de tiempo suficiente para llegar a su maduración, y por ende, poder evaluar la mora de las mismas.

4.2. Procesamiento de los datos

4.2.1. Preprocesamiento de los datos

Los datos de entrada del modelo son vectores de características de cada solicitud, mientras que los datos de salida son el valor de cupo máximo que debe ofrecerse para la solicitud dada. Inicialmente se hizo una limpieza de los datos, para aquellos valores que por errores de manipulación o por errores a la hora de su obtención, no fueron guardados correctamente, o en su defecto, no existen. La organización ya cuenta con métodos propios para hacer esta tarea, los cuales no serán abordados en este documento dada la confidencialidad de estos métodos.

Una vez los datos son limpiados, se procede a hacer una transformación de las variables de fechas. Las mismas son utilizadas para calcular moras, duración de los créditos, etc. Esta transformación se realiza utilizando como base el día en el cual se realizó la solicitud, y se calcula la diferencia con otras fechas, como por ejemplo, la fecha de nacimiento, para calcular la edad del cliente. Esta diferencia se puede devolver en formato de días, meses o años según se requiera para la transformación requerida. Una vez realizado esto, se procede a eliminar todas las fechas, y por último, se realiza una transformación de las variables cualitativas en el modelo, a variables cuantitativas, con el fin de poder procesar los datos en los modelos propuestos con mayor facilidad.

4.2.2. Transformaciones a las variables relacionadas con la moneda

Dado que se manejan datos relacionados con dinero, tales como salarios, montos de crédito, entre otros, es importante convertir los mismos a valor futuro. Tal como indica Finan Finan (2014), el valor del dinero cambia con el tiempo, y como se tienen datos entre el año 2012 y 2022, es importante comparar valores de dinero en tiempos equivalentes. Para hacer esto se hará uso de 2 elementos principales:

- La inflación de los últimos 10 años, dato proporcionado por el Banco de la República, banco central de Colombia. La inflación es uno de los mejores indicadores para evaluar el cambio del poder adquisitivo de la moneda en un periodo de tiempo dado.

  Con la inflación de cada año se puede calcular una tasa equivalente desde el año de la solicitud al presente.

- Función de acumulación para un único periodo, utilizando la tasa de inflación equivalente desde el año de la solicitud al presente.

Por tanto, una vez calculada la tasa equivalente, se hará la siguiente transformación para cualquier variable de moneda:

$$T(X) = X \cdot (1 + i_k) \tag{4-1}$$

Donde i_k es la tasa de inflación equivalente desde el año de la solicitud al presente.

Adicionalmente, se necesita generar una penalidad para los clientes en mora. No hacer esa transformación implicaría entrenar un modelo que prediga la asignación de cupo anteriormente otorgada, pero que no mejore la misma, teniendo los mismos problemas expuestos anteriormente. Para esto se usará la definición de un buen cliente proporcionada por el área de negocio de la organización, la cual considera que un buen cliente es aquel que paga sus obligaciones en los primeros 180 días de su línea de crédito.

El ajuste se realizará reemplazando el cupo otorgado originalmente, por el capital abonado en la línea de crédito en los primeros 180 días de la misma, penalizando así a quienes entraron en una mora mayor a los 180 y dejando con el mismo cupo aquellos que cumplieron con sus pagos.

Adicionalmente, ya que existe un cupo mínimo y máximo para estos créditos, se realizará el ajuste correspondiente para que los valores estén en los rangos determinados por negocio. Por último, el cupo no se tomará como una variable continua, sino discreta, y se hará la separación en n segmentos, donde la etiqueta asignada será el cupo máximo para una

solicitud dada.

4.3. Modelos de estimación del cupo de crédito

4.3.1. Modelo propuesto

El modelo que se propone en este documento no está basado en la probabilidad de incumplimiento. Se tienen los datos ya procesados, es decir, con las transformaciones de la moneda, y más importante aún, con los segmentos de cupo de crédito calculados, y las solicitudes agrupadas según el cupo de crédito correspondiente. Se procederá a crear una configuración experimental. Con esto, se entrenarán y ajustarán los hiperparámetros de 7 modelos presentados. Finalmente, se tomará el mejor modelo en rendimiento según la métrica de desempeño del ROC-AUC score. Esta propuesta es innovadora al hacer un cálculo directo del segmento de cupo sin necesidad de calcular previamente la probabilidad de incumplimiento tal como es presentado en la literatura.

Conjuntos de entrenamiento, testeo y validación

Se tomará el conjunto completo, y se dividirá el conjunto de la siguiente manera:

- 75 % de los datos serán usados para el conjunto de entrenamiento.
- Con el 25 % restante se construirá el conjunto de testeo y validación. Se dividirán en proporción 25/75, usando un 25 % de estos datos para el conjunto de testeo, y el restante 75 % será usado para el conjunto de validación.

Finalmente se obtendrán 3 conjuntos tales que:

- El 75 % del total serán usados para el conjunto de entrenamiento.
- El 18,75 % del total serán usados para el conjunto de validación.
- El 6,25 % del total serán usados para el conjunto de testeo.

Por último, se hará un balance de clases en el conjunto de entrenamiento, utilizando bootstrapping.

Estandarización

Se hará una estandarización de cada una de las variables,se tomará el conjunto de entrenamiento, y se realizará la siguiente transformación para cada variable.

$$z_i = \frac{x_i - \mu_i}{\sigma_i} \tag{4-2}$$

donde:

- μ_i es la media aritmética de la variable i, sobre el conjunto de entrenamiento.
- σ_i es la desviación estándar de la variable i, sobre el conjunto de entrenamiento.
- x es el vector de características, por tanto, x_i es el elemento i de este vector.

Posteriormente, se hará la misma transformación para el conjunto de validación y testeo, teniendo en cuenta que los valores de μ_i y σ_i, para todo i son los calculados sobre el conjunto de entrenamiento.

Métricas de desempeño

Se introducirán algunas definiciones que serán usadas para el calculo de las métricas de desempeño de esta sección y futuras secciones

Elemento i del conjunto: Se define el elemento i, como x_i

Etiqueta del elemento i del conjunto: Se define la etiqueta del elemento i, como y_i

Predicción del modelo, para el elemento i: Sea $f(x)$ la función de predicción del modelo, se define la predicción del modelo para el elemento i como $f(x_i)$.

Verdadero positivo (True positive): Se define una predicción como verdadero positivo si $f(x_i) = 1$ y $y_i = 1$, o equivalentemente, si el modelo de clasificación clasificó el elemento como positivo, teniendo el mismo una etiqueta positiva Fawcett (2006). La cantidad total de verdaderos positivos se denota como tp.

Falso negativo (False negative): Se define una predicción como falso negativo si $f(x_i) = 0$ y $y_i = 1$, o equivalentemente, si el modelo de clasificación clasificó el elemento como negativo, teniendo el mismo una etiqueta positiva Fawcett (2006). La cantidad total de falsos negativos se denota como fn.

Verdadero negativo (True negative): Se define una predicción como verdadero negativo si $f(x_i) = 0$ y $y_i = 0$, o equivalentemente, si el modelo de clasificación clasificó el elemento como negativo, teniendo el mismo una etiqueta negativa Fawcett (2006). La cantidad total de verdaderos negativos se denota como tn.

Falso positivo (False positive): Se define una predicción como verdadero positivo si $f(x_i) = 1$ y $y_i = 0$, o equivalentemente, si el modelo de clasificación clasificó el elemento como positivo, teniendo el mismo una etiqueta negativa Fawcett (2006). La cantidad total de falsos positivos se denota como fp.

Sensibilidad o Recall: Es la proporción de predicciones correctas, entre todos los elementos pertenecientes a la clase positiva Fawcett (2006). Se calcula como:

$$Recall = \frac{tp}{tp + fn} \tag{4-3}$$

Tasa de falsos positivos: Es la proporción de elementos negativos clasificados erróneamente, en las predicciones de la clase negativa Fawcett (2006). Se calcula como:

$$Tasa\ de\ falsos\ positivos = \frac{fp}{tn + fp} \tag{4-4}$$

Precisión: Es la proporción de elementos positivos clasificados correctamente, entre todos los elementos clasificados como positivos Fawcett (2006). Se calcula como:

$$Precisión = \frac{tp}{tp + fp} \tag{4-5}$$

Especificidad: Es la proporción de elementos negativos clasificados correctamente, entre todos los elementos pertenecientes a la clase negativa Fawcett (2006). Se calcula como:

$$Especificidad = \frac{tn}{tn + fp} \tag{4-6}$$

F1 Score: Se define como la media armónica entre la precisión y la sensibilidad Lipton et al. (2014). Se calcula como:

$$F1 = \frac{2}{\frac{1}{precisión} + \frac{1}{sensibilidad}} \tag{4-7}$$

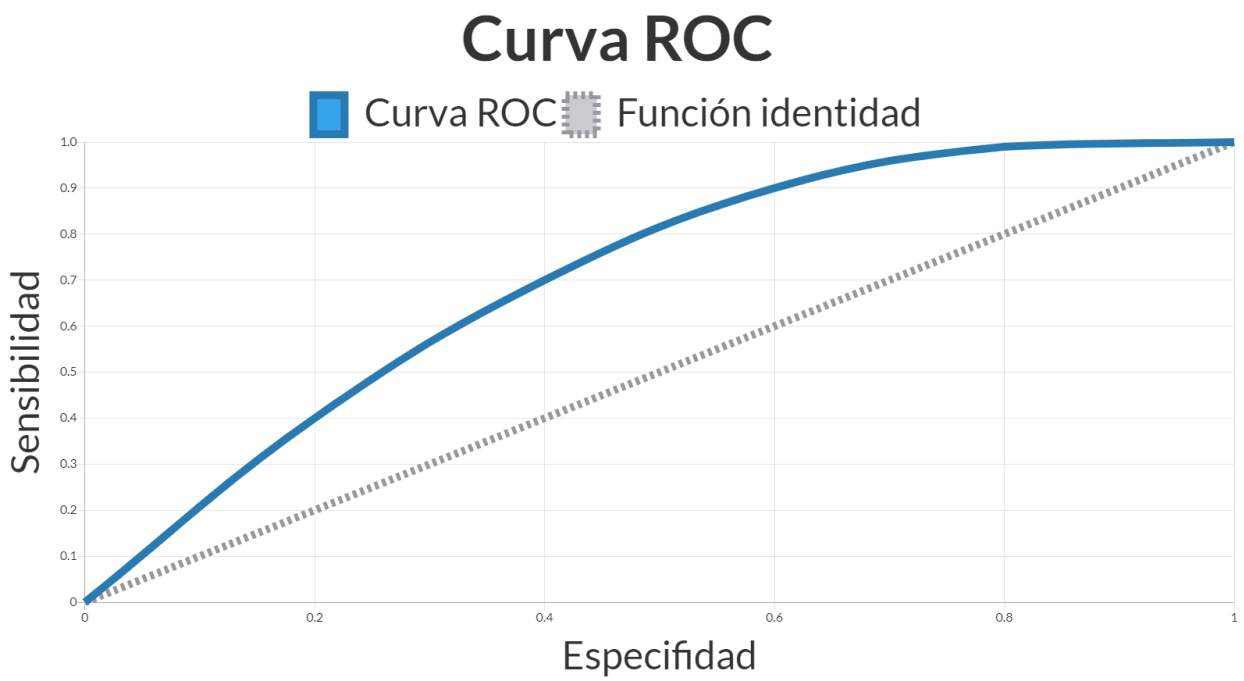
También se puede expresar en función de tp, fp y fn como:

$$F1 = \frac{2 \cdot tp}{2 \cdot tp + fp + fn} \tag{4-8}$$

Curva ROC: Una curva ROC es una gráfica de 2 dimensiones, donde la abscisas corresponden a la tasa de falsos positivos, mientras que las ordenadas corresponden a la sensibilidad Fawcett (2006). Una curva ROC muestra las compensaciones relativas entre los beneficios

(verdaderos positivos) y los costos (falsos positivos).

Figura 4-2. Ejemplo de una curva ROC



Por ejemplo, para la gráfica una especificidad del 0,4 corresponde a una sensibilidad aproximadamente del 0,7. En este tipo de gráficas se espera que la relación entre las variables sea directamente proporcional. Para medir la efectividad de un modelo dado, se requiere más que solo hacer el gráfico de la curva ROC.

Área bajo la curva ROC (ROC AUC): Dado que la curva ROC es una medida bidimensional de desempeño, se busca reducir a una dimensión esta medida para poder comparar modelos de clasificación Fawcett (2006). Calcular el área bajo la curva ROC (AUC) es la medida más habitual. Dada las características de la curva ROC este área está entre 0 y 1.

AUC para una clase: Sea C_i la clase i de un modelo de clasificación multiclase. Sea $C_i{}^c$ el complemento de la clase i. Se define:

$$AUC(C_i, C_i{}^c) \tag{4-9}$$

Como la métrica obtenida al determinar la clase i como la clase objetivo. Intuitivamente se puede entender como el AUC calculado al determinar la clase i como 1 (o clase positiva), y cualquier otra clase como 0 (o clase negativa) Gimeno et al. (2021). Esta métrica es válida únicamente en la aproximación One vs Rest.

AUC One vs Rest: Sea C_i la clase i de un modelo de clasificación multiclase. Sea $C_i{}^c$ el complemento de la clase i y sea a_i la proporción de la clase i frente al total de datos. Se define el AUC One vs Rest como

$$AUC_{ovr} =: \sum_{i=0}^{n-1} a_i \cdot AUC(C_i, C_i{}^c) \tag{4-10}$$

Es decir, es la media ponderada de los AUC score de todas las n clases del modelo de clasificación multiclase.

Para la comparación de los modelos en la propuesta de este documento se hará uso del ROC-AUC score. El mismo, tal como indica Toh Toh et al. (2008) indica una medida relacionada con la capacidad del modelo de discriminar las diferentes clases, por tanto, entre más alto sea el ROC-AUC score, se puede considerar que un modelo puede diferenciar mejor entre las n clases que está clasificando.

Modelos a entrenar y propuesta

La metodología propuesta se basa en la propuesta realizada por Haimowitz Haimowitz and Schwarz (1997). Esta propuesta está compuesta a grandes rasgos por 3 fases:

- Fase de agrupamiento.
- Fase de predicción de probabilidades de cada grupo.
- Fase de predicción de cupo para cada cliente.

Para la fase de agrupamiento, en este caso, se crearán los grupos de manera manual, asignando a cada cliente el segmento correspondiente, tal como fue descrito en la sección 5.2.2 de este documento.

Ahora, la fase de predicción de probabilidades de cada grupo, y la fase de predicción de cupo para cada cliente, se harán en un solo proceso. Se hará un entrenamiento de los modelos descritos posteriormente en esta sección, haciendo uso del conjunto de entrenamiento creado para tal fin, y posteriormente se obtendrá un vector de probabilidades para esta solución. Luego, se hará el cálculo del retorno esperado, para cada vector de características x, de la siguiente manera:

$$C.I(x) = \sum_{i=1}^{n} \text{máx}(B_i) * P(x \in B(i)) \tag{4-11}$$

donde:

- C_i Es el i-ésimo segmento de cupo.
- máx(C_i) Es el cupo máximo para el segmento i de cupo.
- $P(x \in C_i)$ Es la probabilidad de pertenecer al segmento dado, calculada por el modelo.

Posteriormente, se asignará a cada $C.I(x)$ su respectivo segmento, siendo esta la predicción hecha por el modelo. Se entrenarán 7 tipos de modelos diferentes. El ajuste de hiperparámetros se hará usando el conjunto de entrenamiento generado para esta sección, y la métrica de AUC One vs Rest, llamada de ahora en adelante ROC-AUC Score para simplicidad.

Regresión logística: Tal como indica Kleinbaum (Kleinbaum et al., 2002), la regresión logística permite generar una relación entre ciertas variables con una variable dicotómica dependiente de ellas, de hecho, no es la única manera de generar estas relaciones, pero si es la

más popular en algunos sectores, como la epidemiología. Para entender la popularidad de este método, es necesario entender en primer lugar la función en la que se basa su construcción. Esta es la función logística y se define como:

$$\begin{aligned} f &:= \mathbb{R} \to (0,1) \\ f(z) &= \frac{1}{1+e^{-z}} \end{aligned} \tag{4-12}$$

Esta función tiene ciertas propiedades interesantes:

- Su dominio incluye todos los reales.
- Si $z \to \infty$, se tiene que $f(z) \to 0$.
- Si $z \to -\infty$, se tiene que $f(z) \to 1$.
- Su rango es el intervalo (0,1).

Dado el rango que posee, es muy sencillo relacionar esta función con una medida de probabilidad. Luego, el modelo de regresión logística está diseñado justamente para otorgar una probabilidad, y por tanto, el mismo es seleccionado en la mayoría de casos que se requiere conocer la probabilidad de que ocurra o no ocurra un evento, para un vector de características x.

Ahora, es importante poder relacionar ese vector de características con un valor numérico real, para poder evaluarlo en la función logística. Esto se realiza escribiendo z como una combinación lineal de cada uno de los elementos del vector de características, de la siguiente manera:

$$z = \beta_0 + \beta_1 \cdot x_1 + \ldots + \beta_n \cdot x_n \tag{4-13}$$

para ciertos valores $\beta_0, ..., \beta_n$. Es decir, la función logística se puede transformar para que dependa solo del vector de características, obteniendo lo siguiente:

$$f(x) = \frac{1}{1+e^{-(\beta_0+\beta_1 \cdot x_1+\ldots+\beta_n \cdot x_n)}} \tag{4-14}$$

O equivalentemente, se puede calcular la probabilidad de ocurrencia del vector x como:

$$\pi(x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 \cdot x_1 + \ldots + \beta_n \cdot x_n)}} \tag{4-15}$$

Haciendo ciertas transformaciones, se puede calcular la siguiente ecuación equivalente:

$$\beta_0 + \beta_1 \cdot x_1 + \ldots + \beta_n \cdot x_n = \ln\left(\frac{\pi\,(x)}{1 - \pi\,(x)}\right) \tag{4-16}$$

Esta ecuación es importante ya que tal como señala Hosmer (Hosmer Jr et al., 2013), de esta manera se obtiene una ecuación que puede ser optimizada haciendo uso del modelo de regresión lineal. Existen múltiples métodos para ajustar los coeficientes $\beta_0, \beta_1, ..., \beta_n$ para un modelo de regresión lineal. El más famoso es el conocido método de Newton. Consiste en utilizar las derivadas de primer y segundo orden para aproximar iterativamente la solución del modelo de regresión lineal, y es sumamente usado para problemas de optimización convexa (Tibshirani, 2019). Es un algoritmo computacionalmente costoso al tener que hacer uso de las derivadas de segundo orden.

Basado en el método de Newton, se creó el algoritmo L-BFGS. Sigue la misma idea de método de newton, salvo que no calcula las derivadas de segundo orden, sino que aproxima el cálculo de las mismas (Najafabadi et al., 2017). Esto lo hace un algoritmo computacionalmente más eficiente que el método de Newton.

También está la familia de métodos de descenso coordinado. En estos algoritmos se resuelve el problema de optimización realizando sucesivamente búsquedas a lo largo de una dirección de coordenadas, sin utilizar derivadas de primer o mayor orden (Wright, 2015). Estos algoritmos pueden ser ejecutados en paralelo en una máquina, lo que los hace eficientes, pero pueden quedarse atrapados en puntos no estacionarios.

El gradiente en descenso consiste en hacer aproximaciones sucesivas al punto óptimo uti-

lizando el gradiente de la función. Adicionalmente, existe una variación de este algoritmo llamada gradiente en descenso estocástico, donde en cada iteración se agrega una muestra aleatoria, que dificulta el estancamiento en puntos no estacionarios (Schmidt et al., 2017). Pese a esta ventaja, el algoritmo es computacionalmente muy costoso al necesitar más iteraciones para converger. Basado en este algoritmo también se construyó el algoritmo SAGA (Defazio et al., 2014), que suele ser mucho más veloz para procesar grandes cantidades de datos. Los primeros 2 modelos, juntos a los hiperparámetros a ajustar, son respectivamente:

- Regresión logística multiclase.
  - Cantidad máxima de iteraciones para converger.
  - Parámetro de regularización.
  - Tipo de algoritmo a utilizar en el problema de optimización.
- Modelo lineal regularizado con descenso en gradiente estocástico multiclase (SGD).
  - Cantidad máxima de iteraciones para converger.
  - Parámetro de regularización.
  - Método de cálculo de la tasa de aprendizaje.
  - Valor inicial de la tasa de aprendizaje.

Clasificador Navie-Bayes: El clasificador Naive-Bayes es uno de los clasificadores más efectivos según su rendimiento predictivo. Este clasificador aprende sobre los datos de entrenamiento, la probabilidad condicional de cada uno de los elementos del vector de características, para una etiqueta de predicción dada, haciendo uso de la regla de bayes (Friedman et al., 1997). Estos enfoques probabilísticos hacen suposiciones muy fuertes sobre cómo los datos fueron generados, para postular un modelo probabilístico que haga uso de estas suposiciones (McCallum et al., 1998).

El rendimiento de este modelo, a pesar de que en la literatura sea de los que mejor

rendimiento predictivo tenga, se ve limitado por la gran cantidad de suposiciones sobre los datos de entrenamiento (Friedman et al., 1997). Entre estas suposiciones se encuentra asumir que los elementos del vector de características son condicionalmente independientes unos de otros, por tanto, puede presentar un rendimiento pobre en el caso de que esta suposición no sea cierta sobre el conjunto de entrenamiento. Se entrenará un clasificador Navie-Bayes para modelos multivariados de Bernoulli, ajustando los siguientes hiperparámetros.

- Parámetro de regularización.
- Uso de las probabilidades de clase en el conjunto de entrenamiento.

Redes neuronales: Una red neuronal es una máquina diseñada para simular el modelo con el que nuestro cerebro procesa los estímulos. Las redes neuronales hacen uso de múltiples unidades de procesamiento independientes, llamadas neuronas, que están conectadas entre sí a través de una red (Haykin, 2009). También tiene múltiples ventajas como modelos, entre ellas:

- La no linealidad del modelo.
- Adaptabilidad.
- Respuesta basada en evidencia.
- Capacidad de manejar la información contextual.
- Tolerante a los fallos.

Las redes neuronales no solo están formadas por las neuronas, sino adicionalmente por (Haykin, 2009):

- La sinapsis o redes de conexión, que conectan las neuronas entre sí. Cada conexión tiene un peso particular.
- Un combinador lineal, que suma las señales generadas, multiplicadas por el respectivo peso de la conexión.

- Una función de activación, que limita la amplitud de la salida de cada neurona.

Existen múltiples funciones de activación, entre las más conocidas están la función umbral, función logística, función hiperbólica y función RELU (Rectified Lineal Unit). Además se pueden encontrar diversos algoritmos para optimizar los pesos en la sinapsis de estas redes, entre ellos:

- Algoritmo ADAM: Es un algoritmo basado en la optimización de funciones objetivo estocásticas, haciendo uso de derivadas de primer orden. Es un algoritmo fácil de implementar, computacionalmente eficiente, con poco uso de memoria y con gran adaptabilidad (Kingma and Ba, 2014).
- Gradiente en descenso estocástico.
- Algoritmo L-BFGS.

Para este caso, se entrenará un modelo de clasificación de redes neuronales, ajustando los siguientes hiperparámetros:

- Cantidad máxima de iteraciones para converger.
- Función de activación para las neuronas de la red neuronal.
- Método de cálculo de la tasa de aprendizaje.
- Valor inicial de la tasa de aprendizaje.
- Parámetro de regularización.
- Tipo de algoritmo a utilizar en el problema de optimización.

Máquinas de vectores de soporte: Las máquinas de vector de soporte son un conjunto de métodos populares de aprendizaje automático, muy usadas para resolver problemas de regresión y clasificación (Chang and Lin, 2011). Estos métodos implementan la siguiente idea: Los vectores de características son transformados a un espacio de mayor dimensionalidad, y en este espacio se construye una superficie de decisión, es decir, un hiperplano en este

espacio que separa las etiquetas correspondientes a los vectores de características, y dadas algunas propiedades de esta superficie de decisión, se garantiza una excelente generalización (Cortes and Vapnik, 1995). Una característica importante de esta familia de métodos, es que mientras intentan reducir los errores de clasificación, crean representaciones duales, lo que los convierte en métodos computacionalmente eficientes (Cristianini et al., 2000).

En algunos casos los objetivos de los problemas de clasificación no pueden ser descritos como una combinación lineal de los elementos del vector de características, por lo que es necesario tener un espacio de hipótesis mucho más expresivo y general que un espacio lineal (Cristianini et al., 2000). En este sentido, las funciones de kernel son una gran solución, haciendo una transformación, en la mayoría de casos una transformación no lineal, de cada vector de características a un nuevo espacio de características (Valenzuela González et al., 2022). Las funciones de kernel más usadas son las lineales, polinómicas y las funciones gaussianas, que son las más usadas en la práctica (Valenzuela González et al., 2022). En particular, se entrenará una máquina de vector de soporte multiclase, donde se ajustaran estos hiperparámetros:

- Cantidad máxima de iteraciones para converger.
- Tipo de kernel utilizado.

Árboles de decisión: Los árboles de decisión funcionan a grandes rasgos de la siguiente manera (Studer et al., 2011):

- Se agrupan todos los datos de entrenamiento en un conjunto o nodo.
- Se dividen estos datos, haciendo uso de alguna regla aplicada sobre cierto elemento del vector de características y las etiquetas de clasificación. Se busca que los elementos generados en esta división difieran entre sí lo más posible.
- En el paso anterior, cada conjunto generado se considera un nuevo nodo, y se vuelve a

aplicar el mismo procedimiento recursivamente.

- El algoritmo finaliza al alcanzar algún criterio de parada determinado.

Siguiente la analogía de forma de árbol, los árboles de decisión están compuestos por los siguientes elementos (Gareth et al., 2013):

- Nodos terminales o hojas, que son los últimos nodos generados en el algoritmo.
- Nodos internos, que representan las separaciones hechas por el algoritmo, que no son las últimas en ser generadas.
- Los segmentos del árbol que unen a los nodos internos entre sí, son llamados ramas.

Además, los árboles de decisión presentan grandes ventajas sobre otros métodos de clasificación, tales como (Maimon and Rokach, 2014):

- Son autoexplicativos y es sencillo entender las reglas de los nodos.
- Son flexibles para manejar distintos tipos de datos, sean nominales, numéricos, textuales, entre otros.
- Se adaptan fácilmente a conjuntos de datos con errores o valores faltantes.
- Tienen una capacidad de predicción alta, comparado con su relativamente bajos requerimientos computacionales.
- Son útiles para trabajar con conjuntos de datos muy grandes.

Basado en este algoritmo, también se creó el algoritmo de árboles de decisión extremadamente aleatorios. El algoritmo es similar a los árboles de decisión clásicos, salvo que a la hora de dividir los datos para la generación de un nuevo nodo, la selección de la regla se hace de una manera totalmente aleatoria (Geurts et al., 2006). Puede aumentar o disminuir significativamente la varianza y el sesgo sobre las predicciones respecto a los árboles de decisión clásicos, dependiendo de la selección de hiperparamétros del mismo. Los modelos de árboles, junto a los respectivos hiperparámetros a ajustar, son los siguientes:

- Clasificador de árboles de decisión.
  - Estrategia para elegir la división en cada nodo del árbol.
  - Función utilizada para elegir la división en cada nodo del árbol.
  - Mínimo de ejemplos requeridos para hacer una división en un nodo.
- Clasificador de árboles de decisión extremadamente aleatorios.
  - Función utilizada para elegir la división en cada nodo del árbol.
  - Mínimo de ejemplos requeridos para hacer una división en un nodo.

Se seleccionará el modelo que mayor ROC-AUC Score tenga respecto al conjunto de validación. Todo esto se realizará utilizando la librería Sklearn de Python. Puede consultar una guia del codigo en el anexo IV, para replicar los resultados D.

4.3.2. Modelo líneal

El modelo lineal es un modelo que estima el cupo crediticio basado en la probabilidad de incumplimiento, y se utilizará como línea base de comparación para el modelo propuesto. Primero se crea una configuración experimental para entrenar un modelo de regresión logística, para calcular la probabilidad de incumplimiento de cada solicitud. Una vez calculada las probabilidades de incumplimiento, se utilizan las mismas para generar restricciones en un problema de optimización lineal, con el cual se obtienen los cupos asignados para cada cliente.

Probabilidad de incumplimiento

Cálculo de pesos de evidencia (WOE): Para simular los resultados obtenidos por Herga (Herga et al., 2016), se calculará para cada una de las variables su respectivo peso de evidencia (Weight Of Evidence o WOE). Originalmente se creaban n rangos para una variable en particular, asignando un valor máximo y mínimo, y asignando a cada solicitud su rango

correspondiente. Para este caso particular se seleccionaron solo 2 rangos para cada variable, ya que el experto del dominio de negocio asegura que es la manera en la que se separan normalmente a los clientes con buen comportamiento de pago, de los de mal comportamiento. Esta separación se hizo en su totalidad en acompañamiento del experto del dominio. Una vez se separan los 2 rangos para cada variable, se procede a hacer el cálculo del WOE de la siguiente manera:

- Para una variable dada, se selecciona un rango.
- Una vez seleccionado el rango, se procede a hacer el conteo de cuántas solicitudes fueron marcadas como solicitudes con buen comportamiento de pago. Para nuestro caso, se seleccionaron las solicitudes que pagaron el total del capital en los primeros 180 días de realizada la solicitud.
- Para este mismo rango, se realiza un conteo de las solicitudes que se marcaron con un mal comportamiento de pago, es decir, la que no pagaron en su totalidad el capital solicitado.
- Se calcula el WOE para ese rango y variable de la siguiente manera:

$$\log \frac{P(Buen\ pago)}{P(Mal\ pago)} \tag{4-17}$$

  Siendo $P(x)$ el conteo de los casos x.
- Se asigna a cada rango su respectivo WOE.

Estos datos se usarán posteriormente para entrenar un modelo de regresión lineal.

Configuración experimental para la probabilidad de incumplimiento: Se tomará el conjunto completo, y se dividirá el conjunto de la siguiente manera:

- 75 % de los datos serán usados para el conjunto de entrenamiento.
- Con el 25 % restante se construirá el conjunto de testeo y validación. Se dividirán

en proporción 25/75, usando un 25 % de estos datos para el conjunto de testeo, y el restante 75 % será usado para el conjunto de validación.

Finalmente se obtendrán 3 conjuntos tales que:

- El 75 % del total serán usados para el conjunto de entrenamiento.
- El 18,75 % del total serán usados para el conjunto de validación.
- El 6,25 % del total serán usados para el conjunto de testeo.

Se hará uso de una semilla aleatoria fija para poder replicar el experimento con facilidad.

Métricas de desempeño: Dado que Herga Herga et al. (2016) no especifica qué métrica fue utilizada para el cálculo de la probabilidad de incumplimiento, se hará uso del F1-score. Este score permite evaluar tanto la precisión como la sensibilidad, ya que un F1-score alto, se relaciona también con una sensibilidad y precisión altos.

Modelo de probabilidad de incumplimento

Para el cálculo de la probabilidad de incumplimiento se entrenará una regresión logística, ya abordado en la sección anterior. La misma se entrenará utilizando el conjunto de entrenamiento generado para este fin, y se hará el ajuste de los hiperparámetros utilizando el conjunto de validación, y haciendo el cálculo del F1-score para cada configuración de hiperparámetros. Se debe clasificar a los clientes según su comportamiento de pago, es decir, cumple o no con la obligación pactada. Los hiperparámetros que se ajustarán son los siguientes:

- Cantidad máxima de iteraciones para converger.
- El parámetro de regularización.
- Tipo de algoritmo a utilizar en el problema de optimización.

Finalmente se escogerá la configuración de hiperparámetros que tuviera un rendimiento

más alto sobre la métrica escogida. La probabilidad de incumplimiento para el vector de características x se calculará como:

$$PD(x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 \cdot x_1 + \ldots + \beta_n \cdot x_n)}} \tag{4-18}$$

Donde β_i es el i-ésimo coeficiente de regresión lineal calculado por el modelo.

Modelo de optimización lineal

Un problema de optimización, según Boyd Boyd et al. (2004), tiene la forma:

$$\begin{aligned} &\text{minimizar} \quad f_0(x) \\ &\text{sujeto a} \quad f_i(x) \leqslant b_i, \quad i = 1, \ldots, m \end{aligned} \tag{4-19}$$

Donde:

- $x = (x_1, ..., x_n)$ es la variable de optimización del problema, o vector de características.
- $f_0 := \mathbb{R}^n \to \mathbb{R}$ es la función objetivo.
- Las funciones $f_i := \mathbb{R}^n \to \mathbb{R}, \quad i = 1, \ldots, m$ son las funciones de restricción.
- Las constantes $b_1, ..., b_m$ son los límites de las restricciones.

Un vector x es una solución factible si está sujeto a todas las restricciones impuestas. Se dice que un vector x^* es óptimo, si es una solución factible y el valor de la función objetivo en este vector, es el menor posible. Es decir, x^* es óptimo sí $f_0(x) \geqslant f_0(x^*)$, para todo x tal que $f_1(x) \leqslant b_1, ..., f_m(x) \leqslant b_m$.

En particular, un problema de optimización se considera un problema de optimización lineal si tanto la función objetivo como las funciones de restricción son lineales, es decir, si:

$$f_i(\alpha x + \beta y) = \alpha f_i(x) + \beta f_i(y) \tag{4-20}$$

Para todo $x, y \in \mathbb{R}^n$ y todo $\alpha, \beta \in \mathbb{R}$. Tal como indica Boyd Boyd et al. (2004), no existe una fórmula analítica simple para la solución de un problema de optimización lineal, pero existe una gran variedad de métodos para resolver estos problemas. Las 2 más comunes son el algoritmo simplex, creado por George Dantzig , y los métodos de punto interior. El algoritmo simplex es un método iterativo. Para resolver el problema de optimización lineal se construye un poliedro Boyd et al. (2004) y se recorren iterativamente los vértices del mismo buscando minimizar el valor de la función. Dado que estos vértices son finitos, el algoritmo debe hallar una solución óptima en tiempo finito Dantzig (1989).

Los algoritmos de punto interior son una familia de métodos, que recorren iterativamente el interior de la región factible de solución del problema, moviéndose en la dirección que se minimiza la función. Adicionalmente genera transformaciones de la región factible para mejorar la convergencia del algoritmo Andersen et al. (1996).

Para nuestro problema en particular, inicialmente hay que considerar la pérdida esperada para un portafolio de inversión dado. Sea $f(x, y)$ la pérdida asociada al portafolio de inversión x, para un vector aleatorio y. Considere $p(y)$ la función de densidad asociada al vector aleatorio y, el CVaR, denotado como $\phi_\alpha(x)$, tal como indica Herga Herga et al. (2016), se puede calcular como:

$$\phi_\alpha(x) = (1-\alpha)^{-1} \int_{f(x,y)>VaR_\alpha} f(x,y)p(y)dy \tag{4-21}$$

Adicionalmente, $\phi_\alpha(x)$ se puede calcular minimizando la siguiente función:

$$F_\alpha(x,\zeta) = \zeta + (1-\alpha)^{-1} \int_{y\in\mathbb{R}^n} \text{máx}\{f(x,y)-\zeta, 0\}p(y)dy \tag{4-22}$$

Donde ζ es el valor del VaR_α, cuando ζ minimiza la función.

Como se quiere limitar las pérdidas esperadas, se espera que se cumpla la siguiente res-

tricción:

$$F_\alpha(x,\zeta) \leqslant \omega \tag{4-23}$$

Por tanto, encontrar el portafolio de inversión x que maximice el retorno esperado (Denotado como $R(x)$) es equivalente a resolver el siguiente problema de optimización:

$$\begin{aligned} \underset{x\in X,\zeta\in\mathbb{R}}{\text{máx}} \quad & R(x) \\ \text{sujeto a} \quad & F_\alpha(x,\zeta) \leqslant \omega \end{aligned} \tag{4-24}$$

Que es equivalente al problema de optimización lineal:

$$\begin{aligned} \underset{x\in X,\zeta\in\mathbb{R}}{\text{mín}} \quad & -R(x) \\ \text{sujeto a} \quad & F_\alpha(x,\zeta) \leqslant \omega \end{aligned} \tag{4-25}$$

Ahora se pueden generar j escenarios aleatorios diferentes, en donde para cada cliente se simulara que pagó o no pagó la deuda en cuestión. Para cada uno de estos escenarios aleatorios, denotados como ${}^iy$, se puede calcular la probabilidad de ocurrencia del mismo, usando las probabilidades de incumplimiento calculadas en la sección anterior. Por tanto, para cada vector aleatorio ${}^iy$ se conoce la probabilidad de ocurrencia del mismo, denotada como π_i, y se calcula como:

$$\pi_i = \prod_{k=1}^{n} p(x_k = {}^iy_k) \tag{4-26}$$

Siendo n la cantidad total de solicitudes. Se puede aproximar el valor de $F_\alpha(x,\zeta)$ con la siguiente suma:

$$\tilde{F}_\alpha(x,\zeta) = \zeta + (1-\alpha)^{-1} \sum_{i=1}^{j} \pi_i \,\text{máx}\{f(x,{}^iy) - \zeta, 0\} \tag{4-27}$$

Y aplicando lo siguiente:

$$z_i \geqslant f\left(x,{}^iy\right) - \zeta, \quad z_i \geqslant 0, \quad i = 1,\ldots,j, \quad \zeta \in \mathbb{R} \tag{4-28}$$

Se pueden convertir las restricciones del problema de optimización lineal expuesto en 4-25, como lo siguiente:

$$\zeta + (1-\alpha)^{-1} \sum_{i=1}^{j} \pi_i \cdot z_i \leqslant \omega \tag{4-29}$$

$$f(x, {}^{i}y) - \zeta - z_i \leqslant 0, \quad z_i \geqslant 0, \quad i = 1, \ldots, j, \quad \zeta \in \mathbb{R} \tag{4-30}$$

Adicionalmente se generó una limitación sobre los valores que puede tomar cada elemento del portafolio, es decir, el cupo máximo permitido, acotado por unos valores máximos y mínimos para los cupos, permitido por las reglas de la organización. Para este caso en particular, el retorno esperado será:

$$R(x) = \sum_{i=1}^{n} (1 - pd_i) \cdot x_i - pd_i \cdot x_i \cdot margen \tag{4-31}$$

Por tanto, se procede a resolver el siguiente problema de optimización lineal:

$$\begin{array}{ll} \underset{x \in X, \zeta \in \mathbb{R}}{\text{mín}} & -\sum_{i=1}^{n} (1 - pd_i) \cdot x_i - pd_i \cdot x_i \cdot margen \\ \text{sujeto a} & \begin{cases} \zeta + (1-\alpha)^{-1} \sum_{i=1}^{j} \pi_i \cdot z_i \leqslant \omega \\ f\,(x, {}^{i}y) - \zeta - z_i \leqslant 0, \quad i = 1, \ldots, j \\ z_i \geqslant 0, \quad i = 1, \ldots, j \\ lb \leqslant x_i \leqslant ub, \quad i = 1, \ldots n \end{cases} \end{array} \tag{4-32}$$

donde:

- *lb* Es la cota inferior para el cupo.
- *ub* Es la cota superior para el cupo.

Este sera resuelto en Python, haciendo uso de la librería scipy. El algoritmo utilizado será HiGHS, el cual escoge automáticamente entre los algoritmos del método simplex y los métodos de punto interior. La implementación de este algoritmo fue descrita por Huangfu y Hall en 2018. Huangfu and Hall (2018). Se determinó el valor del margen como 0.5, determinado

por un especialista de negocio como el porcentaje de capital que se espera recuperar en caso de que un cliente no pague la totalidad de su obligación, y se determinó el valor de α como 0.95 Por otro lado, se determinaron 3 valores de ω, también con ayuda de un especialista de dominio:

- $\omega_{máx}$, determinado como el valor máximo permitido de pérdida. Es la mayor pérdida económica que la organización está dispuesta a pagar.
- ω_{real}, determinado como el valor esperado de pérdida de la organización.
- $\omega_{mín}$, determinado como el valor ideal de pérdida de la organización, siendo este hasta un orden de magnitud menor que ω_{real} Una vez obtenida una solución, se hará el cálculo del segmento para cada predicción hecha en el modelo de optimización lineal, y se calculará el ROC-AUC Score sobre el conjunto de validación, seleccionando el modelo que tenga mejor rendimiento.

4.3.3. Comparación

Se comparará el rendimiento entre el modelo lineal, respecto al modelo propuesto de este documento. Adicionalmente, se genera un modelo aleatorio para validar que las soluciones tienen un mejor rendimiento que la aleatoriedad misma, todo esto sobre el conjunto de validación. Se compararán:

- ROC-AUC Score sobre el conjunto de validación
- Gráfica de distribución de errores
- Algunos scores adicionales, como F1-Score, precisión y sensibilidad.
- Algunas medidas estadísticas sobre la distribución de errores.

4.4. Evaluación de resultados obtenidos

Se evaluará el modelo seleccionado, sobre el conjunto de testeo, calculando lo siguiente:

- ROC-AUC score
- F1-Score
- Precisión
- Sensibilidad

Además, se incluirá un análisis de la asignación del modelo respecto a 4 clientes aleatorios, junto a un análisis de correlaciones entre las variables del vector de características y la predicción del modelo.

5. Resultados

A continuación, se reportan los resultados del modelo para asignación del cupo de crédito. Primero, se reportan algunos ejemplos de asignaciones de crédito realizadas por el modelo propuesto en comparación con asignaciones realizadas por expertos. Seguidamente, se comparan los resultados del modelo contra modelos de línea base. Posteriormente, se reportan algunos resultados adicionales relacionados con los desempeños detallados del modelo, sesgos de sub o sobreestimación en la asignación realizada por el modelo, y posibles correlaciones entre predictores y segmentos de cupo asignados. Finalmente, se reportan algunos detalles cuantitativos del modelo de línea base lineal.

5.1. Asignaciones individuales de cupos crediticios

La Figura **5-1** ilustra las predicciones de cupo realizadas por el modelo propuesto y su comparación con los cupos propuestos por un experto de negocio para cuatro clientes diferentes. Cada panel ilustra una predicción particular. La cantidad de símbolos de billetes a la izquierda se refiere al nivel socioeconómico. Las casas simbolizan el nivel sociodemográfico. Y los dos símbolos de caras se relacionan con reportes de centrales crédito positivas (verde) y negativas (rojas).

Figura 5-1. Predicciones individuales realizadas por el modelo propuesto frente a las realizadas por un experto del dominio. Cada panel ilustra a la izquierda con símbolos particulares los valores de tres variables predictoras (económicas, sociodemográficas y de crédito) representativas de la caracterización de los clientes, a la derecha el segmento de cupo asignado: bajo, medio-bajo, medio-alto y alto.

"Imagen de Freepik". Esta portada ha sido diseñada usando imágenes de Freepik.

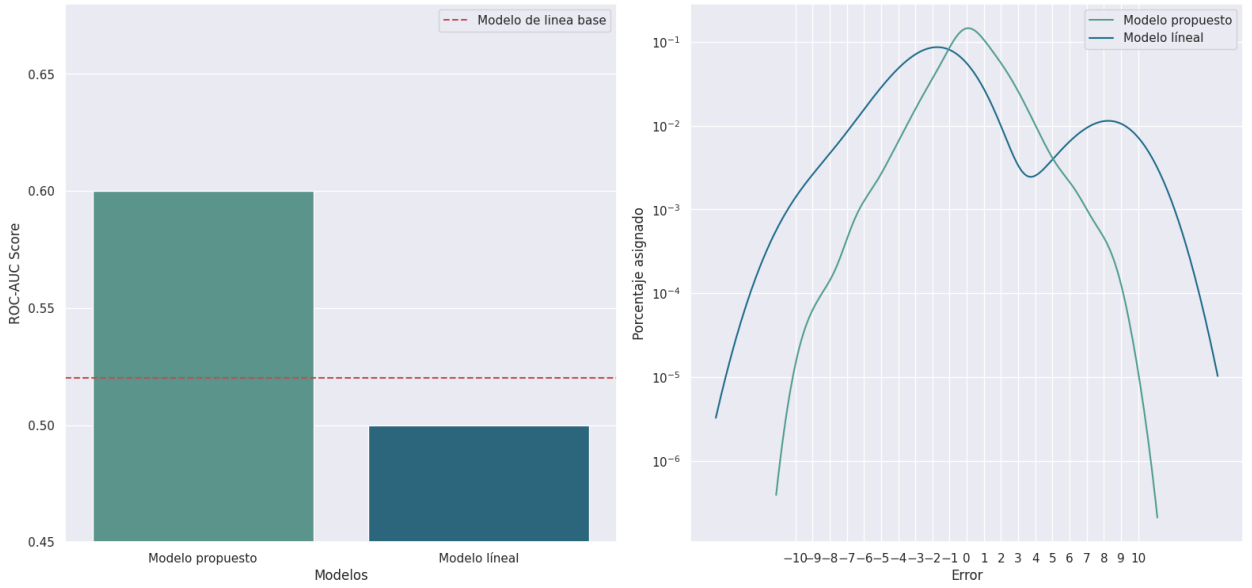
Es importante anotar que, por aspectos de privacidad del origen de los datos, no es posible reportar en este documento datos detallados sobre los clientes, ni la naturaleza específica de las variables predictores, inclusive para clientes anonimizados debido a aspectos comerciales.

El panel A ilustra como para un cliente con perfil socioeconómico, sociodemográfico y de crédito positivo, el modelo asigna un valor alto de cupo de crédito, coincidiendo con el experto. Similarmente, el panel D muestra cómo el modelo asignó exitosamente un cupo de crédito mínimo, que coincide con el sugerido por el experto, cuando el perfil es desfavorable respecto a su nivel económico y de reporte de crédito. El panel B muestra cómo el modelo asigna el cupo correctamente (cupo medio) para un cliente con reporte de crédito positivo, nivel sociodemográfico bajo, y nivel económico medio. El panel C ilustra cómo el método asignó un cupo a personas con perfiles económicos y sociodemográficos medios, pero con un reporte de crédito negativo. En este caso el modelo propuesto sugiere asignar un cupo medio, el cual es muy cercano al cupo bajo sugerido por el experto. En este caso el modelo sobreestimó la asignación del cupo beneficiando al cliente. Es importante anotar que todas las predicciones reportadas y los análisis de resultados subsecuentes se realizaron con el modelo de árboles de decisión, el cual resultó ganador del proceso de selección de modelos (ver Sección 5.4).

5.2. Comparación del modelo propuesto contra modelos de línea base

La Figura **5-2** reporta los resultados de la comparación del modelo contra dos modelos de línea base. En particular, el modelo lineal (vea la Sección **4.3.2**) y un modelo de asignación aleatoria del cupo. A la izquierda se reporta el desempeño del modelo propuesto (verde), cuando se compara contra el modelo lineal (azul), y de selección aleatoria del cupo (línea punteada). Se reportan los valores de ROC-AUC calculados sobre el mismo conjunto de validación para los tres modelos.

Figura 5-2. (Izquierda) Comparación del modelo propuesto (verde) con dos modelos de línea base: lineal (azul) y aleatorio (línea punteada). (Derecha) Sub y sobreestimaciones del cupo realizadas por el modelo sobre datos de validación considerando como referencia a los expertos.



Como puede observarse, el modelo propuesto reporta un rendimiento mayor al de los modelos de línea base. En particular, el ROC-AUC, que mide el desempeño del clasificador variando especificidad y sensibilidad, supera por más de 10 puntos porcentuales al modelo basado en optimización lineal, y por más de 8 puntos porcentuales al modelo aleatorio. Mostrando una capacidad de generalización alta en comparación con el modelo lineal reportado en la literatura. Interesantemente, el modelo de asignación aleatoria superó al modelo lineal de asignación.

5.3. Resultados detallados del modelo propuesto

La Tabla **5-1** reporta los resultados detallados para los tres modelos comparados (propuesto, lineal y aleatorio).

Modelos	*Precisión (%)*	*Sensibilidad (%)*	*F1-score (%)*
Propuesto	40	36	38
Lineal	9	1	2
Aleatorio	20	4	7

Tabla 5-1.: Métricas de desempeño detalladas para el modelo propuesto. En particular se reporta la precisión, sensibilidad y F1-score para el modelo propuesto y los modelos de línea base.

Como puede observarse, el método propuesto tiene una capacidad más alta de asignar cupos correctamente que los modelos de línea base utilizados (precisión). Igualmente, el método propuesto tiene una alta capacidad para asignar correctamente los diferentes segmentos (sensibilidad). Finalmente, el método propuesto muestra un buen balance entre precisión y exhaustividad comparado con los métodos de línea base (F1-score).

5.3.1. Sesgos por sub y sobreestimación del cupo

Con el fin de estudiar de forma cuantitativa posibles sesgos por sub y sobreestimación del cupo generados por el modelo propuesto se estudiaron los momentos estadísticos de la distribución de errores, los cuales dan cuenta de la forma de la distribución. La Tabla **5-2** reporta los cuatro primeros momentos estadísticos (media, varianza-desviación estándar, asimetría y curtosis) para las distribuciones de los errores en asignación de cupo (ver Panel derecho en Figura **5-2**) para el modelo propuesto y el lineal.

Modelo	*Media*	*Desviación estándar*	*Asimetría*	*Curtosis*
Propuesto	0.25	1.64	0.12	1.56
Líneal	-1.06	3.6	1.6	2.49

Tabla 5-2.: Momentos de la distribución de errores de los modelos. Se reportan los valores para los cuatro primeros momentos para el modelo propuesto y el modelo lineal.

Como puede observarse, el modelo propuesto tiene una media cercana a cero (0.25) que indica en promedio los errores se distribuyen cerca a cero. En comparación con el modelo

lineal que comete errores de subestimación por debajo de -1. Respecto a la dispersión de los datos el modelo propuesto tiene una dispersión más pequeña comparada con el modelo lineal. Indicando que los errores se agrupan más alrededor de la media en el modelo propuesto. Los valores de asimetría confirman que la distribución de los errores para el modelo lineal tienen un alto sesgo hacia las subestimaciones. El modelo propuesto tiene una asimetría positiva (0.12), pero pequeña en magnitud. Finalmente, la curtosis positiva y alejada de 0 indica que la distribución de errores del modelo propuesto leptocúrtica, es decir, hay una alta concentración de errores alrededor de la media central, y siendo esta cercana a 0, se concluye que el modelo propuesto suele tener errores pequeños, en comparación con el modelo lineal en el cual los datos se distribuyen de forma más dispersa.

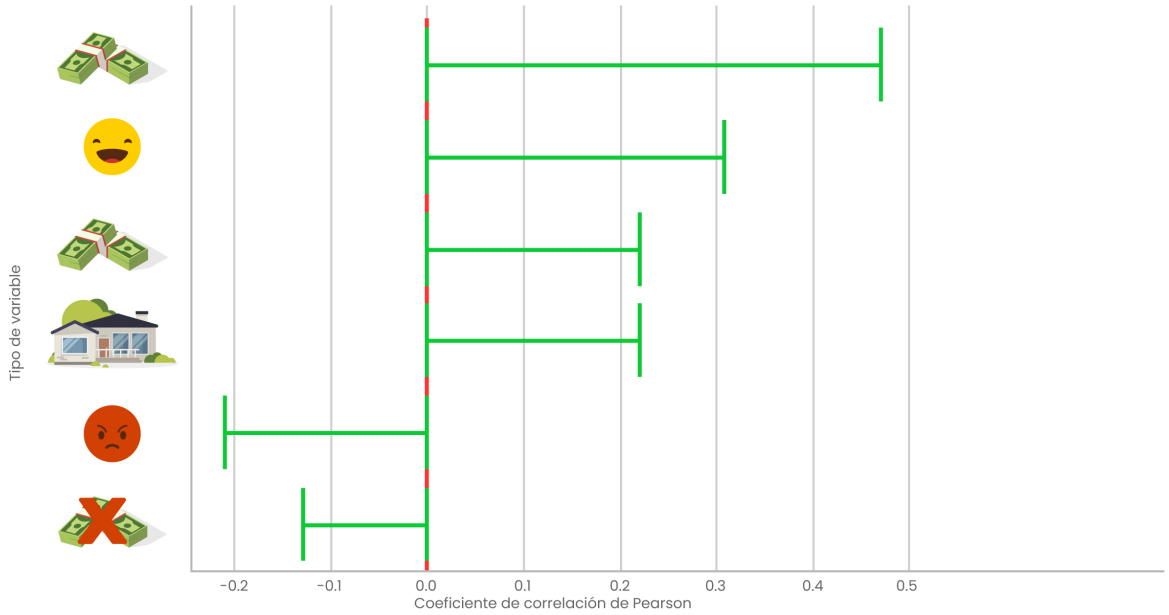
5.3.2. Relación entre variables predictoras y el cupo asignado

En principio, podría pensarse que un experto solamente considera un conjunto pequeño de características de los clientes para determinar el valor de crédito. Por ejemplo, la condición económica del cliente puede determinar directamente el cupo crediticio. En este caso, considerar un modelo complejo basado en aprendizaje de máquina como el propuesto, no tendría sentido. Con el fin de estudiar este posible comportamiento, se calcularon las correlaciones entre un conjunto de variables predictoras representativas y los cupos de crédito generados por el modelo.

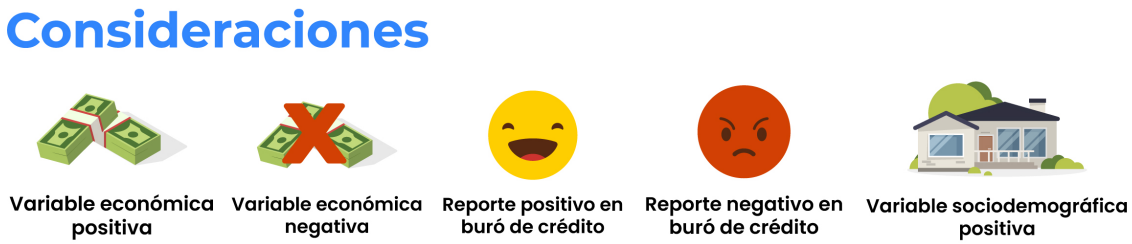
La Figura **5-3** reporta las correlaciones entre un conjunto de variables predictoras representativas de los clientes y las predicciones realizadas por el modelo. Como antes por cuestiones de privacidad solamente se ilustran variables económicas, sociodemográficas y de crédito. La figura muestra a la izquierda las variables predictoras con mayor valor magnitud en el coeficiente de correlación.

En primer lugar, es importante observar cómo las variables económicas y de crédito se

Figura 5-3. Coeficientes de correlación de Pearson para variables económicas, sociodemográficas y de buró de crédito representativas. Se evidencia cómo las variables que se consideran positivas para el riesgo crediticio tienen una correlación positiva frente a la predicción, mientras que las que tienen connotación negativa, tienen un coeficiente de correlación negativo con la predicción.



"Imagen de Freepik". Esta portada ha sido diseñada usando imágenes de Freepik.



correlacionan (positiva y negativamente) con los cupos de crédito asignados (altos y bajos), sugiriendo que el modelo captura información de variables representativas del problema modeladas por los expertos. No obstante, los valores de correlación no son lo suficientemente altos como para que una sola variable explique las predicciones realizadas por el modelo. Este comportamiento, sugiere que es posible que exista una relación no-lineal entre las variables predictoras y el cupo asignado por el modelo.

5.4. Anexo: Selección de modelos

Para la selección del modelo se construyó un espacio de hipótesis con los modelos descritos en la sección **4.3.1**. Estos modelos fueron entrenados con los hiperparametros reportados en anexo B utilizando los datos de entrenamiento y fueron evaluados utilizando el subconjunto de validación.

La Tabla **5-3** reporta los desempeños en ROC-AUC Score (%) obtenidos para la comparación de los modelos de aprendizaje de máquina. Estos desempeños se obtuvieron sobre la misma partición de validación.

Modelos	*ROC-AUC Score (%)*
Modelo aleatorio	52
Regresión logística multiclase	52
Modelo lineal (SGD)	52
Clasificador Navie-Bayes	48
Redes neuronales	58
Máquina de vector de soporte	50
Árboles de decisión	**60**
Árboles de decisión extremadamente aleatorios	58

Tabla 5-3.: Rendimiento de los modelos sobre el conjunto de validación, respecto al modelo de línea base

Como puede observarse el modelo de árboles de decisión obtuvo el mejor desempeño,

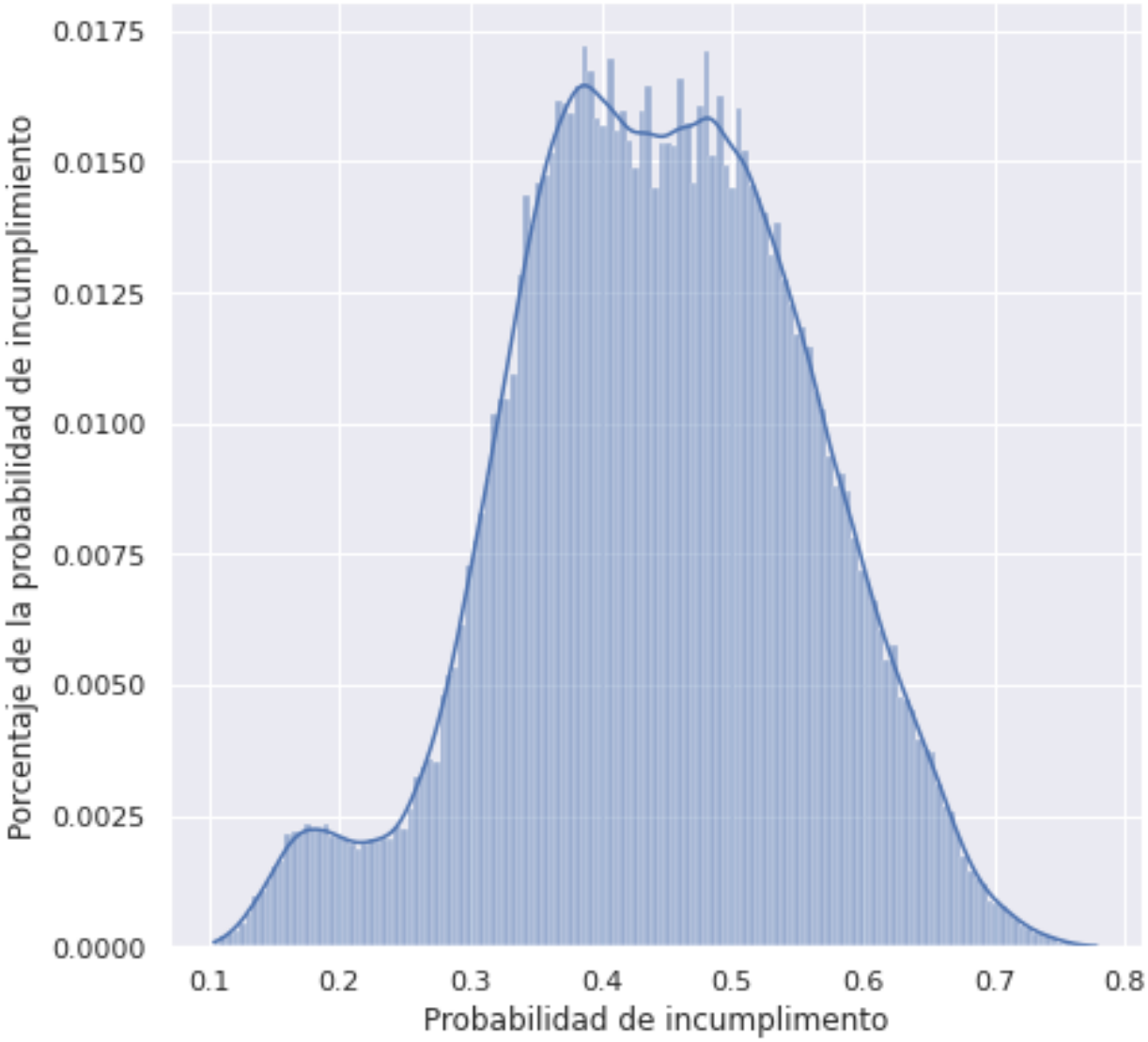
comparado con los otros modelos evaluados. Por esta razón este modelo fue seleccionado para la fase de evaluación.

5.5. Modelo lineal

Los tres modelos lineales planteados (ver Sección **4.3.2**): ω_{min}, ω_{real}, $\omega_{máx}$ fueron abordados utilizando optimización líneal. Los resultados sugieren que el modelo ω_{min} no tiene solución factible. Los valores de ROC-AUC para los modelos con solución factible fueron calculados, transformando los valores obtenidos a los segmentos correspondientes. Y su desempeño se comparó sobre el conjunto de datos de validación. El modelo con mejor desempeño fue ω_{real} con un ROC-AUC de 0.502.

La Figura **5-4** reporta la distribución de las probabilidades de incumplimiento estimadas para el ajuste de modelo lineal. Como puede observarse, no existen solicitudes con probabilidades de incumplimiento menores a un 10 % ni mayores a un 80 %. Adicionalmente, más de un 68 % de las solicitudes tiene menos de un 50 % de probabilidad de incumplimiento. Estas solicitudes corresponden a las que el modelo de regresión logística clasificó solicitudes en las cuales se cumplira posteriormente con las obligaciones pactadas.

Figura 5-4. Distribución de las probabilidades de incumplimiento calculadas por el modelo de regresión logistica. Se evidencia un sesgo marcado a izquierda, con más de un 68 % de las predicciones que no superan el umbral del 50 %.



6. Discusión

En primer lugar es importante analizar los resultados de la probabilidad de incumplimiento calculada. En este caso, dado que para aproximadamente un 68 % de las solicitudes fue calculada una probabilidad de incumplimiento menor al 50 %, estas solicitudes fueron clasificados por el modelo como solicitudes sobre las cuales se cumpliran con sus obligaciones. Lo importante es hacer una comparación frente al modelo de calificación con el que cuenta la organización, que clasificó al 100 % de las solicitudes como solicitudes en las cuales se cumpliran con las obligaciones posteriormente. Este modelo es confidencial y no se tuvo acceso al mismo para la elaboración de esta propuesta.

Esto se debe a que el proceso de scoring crediticio, al menos en esta organización, se hace previo a la asignación del cupo de crédito, y por ende, toda solicitud que recibe una asignación de cupo, previamente fue calificada como una solicitud donde se cumpliran las obligaciones crediticias. Eso puede representar un problema a la hora de calcular la probabilidad de incumplimiento, y dado que no se tuvo acceso al modelo de scoring de la organización por temas de confidencialidad del mismo, no se pueden comparar estas probabilidades de incumplimiento calculadas a posteriori, frente a las calculadas a priori para el modelo de scoring de la organización.

Estas diferencias que no son calculables, pueden introducir un sesgo en las probabilidades de incumplimiento calculadas, lo cual tendría un impacto sustancial en las mismas. Este impacto influirá sustancialmente sobre el modelo de optimización lineal, basado justamente

en las probabilidades de incumplimiento. Tener acceso a las probabilidades de incumplimiento reales podría haber representado una mejora significativa en el pobre rendimiento, aunque no existe manera de evaluar o confirmar esto dada la confidencialidad del modelo de scoring original.

Frente a la propuesta realizada, en primer lugar se puede abordar el pobre rendimiento de los modelos lineales. Si evaluamos los coeficientes de correlación de Pearson entre los elementos del vector de características y las variables de respuesta, no se observan coeficientes con una magnitud mayor a 0.5. Por tanto, se concluye que no existen correlaciones lineales entre las variables del modelo, lo que podría explicar el pobre rendimiento.

Tanto las redes neuronales como los modelos de árboles mostraron un rendimiento bueno. Lo más probable es que este buen rendimiento se deba a la falta de correlaciones lineales entre los elementos del vector de características y la variable de respuesta. Además se evidencia que el modelo propuesto tiene una correlación entre las variables del vector de características y los cupos que se predicen, sugiriendo que se captura información importante de las variables representativas del problema, sin que exista una que sea más representativa y pueda explicar el funcionamiento del mismo.

El modelo presenta pequeños sesgos a la subestimación, y dado que se pretende enfocar el modelo en reducir los riesgos crediticios, esto es positivo, además de concentrar los datos cerca de un error 0, mostrando que el modelo en cuestión tiene errores pequeños en general, lo cual explica sus buenas métricas y capacidad de diferenciación de los segmentos de cupos.

7. Conclusiones y recomendaciones

El modelo propuesto se mostró superior al modelo lineal, previamente reportado en la literatura. Esta superioridad se evidenció tanto en las métricas de desempeño seleccionadas, como en la distribución de los errores y otros aspectos varios. Los resultados sugieren que la predicción directa del cupo de crédito es posible, sin estimar de forma intermedia la probabilidad de incumplimiento, resultando en mejores desempeños en la estimación. demostró pobres desempeños, no obstante, trabajos futuros pueden explorar de forma más profunda posibles fallos en esta aproximación. El rendimiento de los modelos dependerá del caso particular bajo estudio, y es más probable que ante una probabilidad de incumplimiento calculada a priori o un proceso de originación con un orden diferente al presentado en la organización bajo estudio el modelo lineal tenga mayor impacto. En todo caso, se presenta una nueva e innovadora manera de abordar el problema de asignación de cupo únicamente basado en las variables que describen la solicitud del cliente. Esta propuesta puede ser adaptada a casos particulares, distintos tipos de modelos o métricas, dando una gran adaptabilidad para los casos que surjan en una industria financiera mundial constantemente en desarrollo.

A. Anexo I: Definiciones

A.1. Definiciones relacionadas con finanzas

A.1.1. Probabilidad de incumplimiento

Es una medida de clasificación para clientes, la cual se otorga internamente en una organización para determinar la probabilidad de que el mismo incurra en un incumplimiento en sus obligaciones crediticias, en un tiempo determinado.

La probabilidad de incumplimiento para la solicitud de crédito i se puede denotar como:

- $PD(x_i)$
- PD_i
- pd_i

A.1.2. Beneficio Bruto

Es el beneficio que obtiene una sociedad antes de haber deducido impuestos Heriberto and Vicente (2006).

A.1.3. Beneficio neto

Exceso de ingresos sobre costos y gastos efectuados en un periodo de tiempo. Si los gastos exceden a los ingresos, tal diferencia recibe el nombre de pérdida neta Heriberto and Vicente (2006).

A.1.4. FICO Score

FICO es un puntaje crediticio creado por Fair Isaac Corporation para evaluar el riesgo crediticio. El puntaje FICO es ampliamente utilizado en la industria de préstamos. Se percibe como un indicador robusto de riesgo y, por lo tanto, tiene una amplia aceptabilidad en la industria. Al ser un cuadro de mando, el cuadro de mando de Fico es una variable continua Paul and Biswas (2017).

A.1.5. Utilización de línea de crédito

La utilización de línea de crédito es la cantidad de un cupo de crédito utilizado en una línea de crédito dada. Como práctica general se expresa con un porcentaje del crédito disponible utilizado por el cliente, así que en general se puede expresar como una variable continua Paul and Biswas (2017). En algunos casos se puede determinar por niveles o rangos, llevando a que el mismo pueda ser considerado una variable discreta.

A.1.6. Función de descuento

Es la cantidad que debe invertirse hoy a la tasa de interés i por período para generar una cantidad de \$1 al final de t períodos de tiempo Finan (2014).

$$v^t = \frac{1}{(1+i)^t} \tag{A-1}$$

A.1.7. Valor presente

Corresponde a la cantidad que debe invertirse hoy a la tasa de interés i por período para generar una cantidad de $\$X$ al final de t períodos de tiempo Finan (2014). Se puede calcular haciendo uso de la función de descuento.

$$PV = v^t \cdot X \tag{A-2}$$

O equivalentemente:

$$PV = \frac{X}{(1+i)^t} \tag{A-3}$$

A.1.8. Función de acumulación

Representa el valor acumulado de un capital de \$1 invertido a un tiempo $t \geqslant 0$ Finan (2014) . Para el caso particular del interés compuesto, se puede calcular de la siguiente manera:

$$a(t) = (1+i)^t \tag{A-4}$$

A.1.9. Valor futuro

Representa el valor acumulado de un capital de $\$X$ invertido a un tiempo $t \geqslant 0$ Finan (2014). Para el caso particular del interés compuesto, se puede calcular de la siguiente manera, utilizando la función de acumulación:

$$FV = a(t) \cdot X \tag{A-5}$$

O equivalentemente:

$$FV = X \cdot (1+i)^t \tag{A-6}$$

A.1.10. Valor en riesgo

El VaR a un nivel de confianza α denotado $\alpha - VaR$, es la máxima pérdida potencial de un portafolio de inversión, en un horizonte de tiempo t con un nivel de nivel de confianza α Arbeláez and Ceballos (2005).

A.1.11. Valor en riesgo condicional

El CVaR está definido como el valor esperado de las pérdidas que exceden al VaR, para un nivel de confianza α Melo and Granados (2011).

A.1.12. Retorno Esperado

Sea X un portafolio, y sea EX el valor esperado del portafolio. Sea v^t la función de descuento. Se define el Retorno Esperado para el tiempo t como:

$$ER = v^t \cdot EX \tag{A-7}$$

A.2. Definiciones relacionadas con probabilidad

Para una variable aleatoria discreta X, definimos la función de probabilidad $p(x_i)$ de X como Ross (2014):

$$p(x_i) = P(X = x_i) \tag{A-8}$$

A.2.1. Valor esperado

Sea X una variable aleatoria discreta, y sea $p(x)$ su función de probabilidad. Se define el valor esperado EX (esperanza matemática, media) de X como Ross (2014):

$$EX =: \sum_{i=1}^{n} x_i \cdot p(x_i) \tag{A-9}$$

A.2.2. Probabilidad condicional

Sean A,B eventos de probabilidad. La probabilidad condicional de que B ocurra dado que A ha ocurrido se define como Ross (2014):

$$P(B|A) =: \frac{P(A \cap B)}{P(A)} \tag{A-10}$$

A.2.3. Teorema de probabilidad total

Sean $A_1, A_2, .., A_n$ eventos mutuamente excluyentes que forman una partición del espacio muestral, con $P(A_i) > 0$ para todo i, entonces, para todo evento B con $P(B) > 0$ se tiene que Bertsekas and Tsitsiklis (2008):

$$P(B) = P(A_1 \cap B) + ... + P(A_n \cap B) \tag{A-11}$$

$$P(B) = P(B|A_1)P(A_i) + ... + P(B|A_n)P(A_n) \tag{A-12}$$

$$P(B) = \sum_{i=1}^{n} P(B|A_i)P(A_i) \tag{A-13}$$

A.2.4. Regla de Bayes

Sean $A_1, A_2, .., A_n$ eventos mutuamente excluyentes que forman una partición del espacio muestral, con $P(A_i) > 0$ para todo i, entonces, para todo evento B con $P(B) > 0$ se tiene que Bertsekas and Tsitsiklis (2008):

$$P(A_i|B) =: \frac{P(A_i)P(B|A_i)}{P(B)} \tag{A-14}$$

$$P(A_i|B) =: \frac{P(A_i)P(B|A_i)}{P(B|A_1)P(A_1) + ... + P(B|A_n)P(A_n)} \tag{A-15}$$

$$P(A_i|B) =: \frac{P(A_i)P(B|A_i)}{\sum_{j=1}^{n} P(B|A_j)P(A_j)} \tag{A-16}$$

A.2.5. Independencia condicional

Sean A, B, C eventos de probabilidad. Se dice que A es condicionalmente independiente de B, dado un evento C, si Friedman et al. (1997):

$$P(A|B \cap C) = P(A|C) \tag{A-17}$$

A.3. Definiciones varias

Para empezar, considere un problema de clasificación usando únicamente 2 clases, también conocido como problema de clasificación binario.

A.3.1. Vector de características

Es el espacio de características que definen a un elemento a clasificar. El vector de características i será denotado como x_i.

A.3.2. Etiquetas reales

Para cada vector de características x_i se asigna una etiqueta y_i. Para el caso del problema de clasificación binaria, se tiene que $y_i \in \{0, 1\}$, o equivalentemente, cada elemento se define con una etiqueta negativa o positiva respectivamente.

A.3.3. Modelo de clasificación binario

Es una función cuyo dominio es el conjunto de características del problema, y cuya imagen es el conjunto $\{0, 1\}$.

$$f := X \rightarrow \{0, 1\} \tag{A-18}$$

Los valores asignados por el modelo de clasificación también son llamados etiquetas predichas. Estas etiquetas no deben confundirse con las etiquetas reales. Se denota $f(x_i)$ a la predicción hecha para el vector de características i.

A.3.4. Modelo de clasificación multiclase (de n clases)

Es una función cuyo dominio es el conjunto de características del problema, y cuya imagen es el conjunto $\{0, .., n-1\}$.

$$f := X \rightarrow \{0, .., n-1\} \tag{A-19}$$

B. Anexo II: Hiperparámetros de los modelos entrenados

	Hiperparámetros
Regresión logística multiclase	C-value: 1.0, max-iter: 100, Algotirmo: L-BFGS, penalty: L2
Modelo lineal regularizado con descenso en gradiente estocástico multiclase (SGD)	alpha: 0.01, max-iter: 100, eta_0: 1, learning_rate: constant, penalty: L2
Clasificador Navie-Bayes para modelos multivariados de Bernoulli	alpha:0, fit_prior: true
Modelo de clasificación de redes neuronales	alpha: 0.01, max-iter: 100, solver:adam, learning_rate_init: 1, activation: tanh
Máquina de vector de soporte multiclase	max-iter: 10, kernel: rbf
Clasificador de árboles de decisión	criterion: gini, splitter: best, min_samples_split: 2
Clasificador de árboles de decisión extremadamente aleatorios	criterion: gini, min_samples_split: 2

Tabla B-1.: Hiperparámetros para modelos entrenados

C. Anexo III: Hiperparámetros del modelo de regresión logistica

	Hiperparámetros
Modelo de regresión logística	C-value: 1.0, max-iter: 100, Algotirmo: L-BFGS, Penalty: L2

Tabla C-1.: Hiperparámetros para modelo de regresión logística

D. Anexo IV: Código guia

```
import pandas as pd
import numpy as np
import pickle
from sklearn.metrics import roc_auc_score
from sklearn.metrics import make_scorer
from sklearn.MODELFAMILY import MODEL
# Cargar los datos procesados, ya divididos en entrenamiento, prueba y
 validacion,usando pickle.
X, y, X_train, X_test, y_train, y_test, X_validation, y_validation, amount
data.load()
# Definir una funcion para generar la etiqueta y segmento de cada valor,
esta funcion cambiara segun los segmentos de cupo definidos.
def amount_class_generator(x):
    return f(x)
# Definir una funcion para calcular la puntuacion utilizando el ROC-AUC
score ponderado, y las probabilidades de pertenecer a los
segmentos de cupos creados.
def scoring_function(estimator, X, y_true):
    predict = np.dot(estimator.predict_proba(X), estimator.classes_)
```

```
    predict = amount_class_generator(predict)
    score = roc_auc_score(pd.get_dummies(predict, drop_first=False),
            pd.get_dummies(y_true, drop_first=False), average='weighted')
    return score
# Crear un modelo inicial sin hiperparametros ajustados, y entrenarlo.
final_model = MODEL(random_state=0)
final_model.fit(X_train, y_train)
final_score = scoring_function(final_model, X_validation, y_validation)
# Entrenar cada modelo con uno de los hiperparametros en la lista, y
compararlo con el mejor modelo. En caso de tener mejor puntaje, lo reemplaza.
for params in params_list:
    model = MODEL(random_state=0, parametros=params)
    model.fit(X_train, y_train)
    score = scoring_function(model, X_validation, y_validation)
    if score > final_score:
        final_score = score
        final_model = model
# Guardar el mejor modelo usando pickle, y mostrar sus hiperparametros.
final_model.save()
final_model.get_params()
```

Bibliografía

Andersen, E. D., Gondzio, J., Mészáros, C., Xu, X., et al. (1996). *Implementation of interior point methods for large scale linear programming*. HEC/Université de Geneve.

Arbeláez, L. C. F. and Ceballos, L. E. F. (2005). El valor en riesgo condicional cvar como medida coherente de riesgo. *Revista Ingenierías Universidad de Medellín*, 4(6):43–54.

Ayling, N. (2018). Cresk: scoring de riesgo alternativo.

Badel Coronell, P. D. J. and Herrera Valdez, A. E. (2013). Diseño de una guía de procedimientos para lograr la efectiva asignación de cupos de créditos en la empresa ci antillana sa.

Banerjee, A., Karlan, D., and Zinman, J. (2015). Six randomized evaluations of microcredit: Introduction and further steps. *American Economic Journal: Applied Economics*, 7(1):1–21.

Bertsekas, D. and Tsitsiklis, J. N. (2008). *Introduction to probability*, volume 1. Athena Scientific.

Boyd, S., Boyd, S. P., and Vandenberghe, L. (2004). *Convex optimization*. Cambridge university press.

Cano, C. G., Esguerra, M., García, N., Rueda, L., and Velasco, A. (2014). Inclusión financiera en colombia. *Recuperado de: http://www. banrep. gov. co/sites/default/files/eventos/archivos/sem_357. pdf.*

Chang, C.-C. and Lin, C.-J. (2011). Libsvm: a library for support vector machines. *ACM transactions on intelligent systems and technology (TIST)*, 2(3):1–27.

Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine learning*, 20(3):273–297.

Cristianini, N., Shawe-Taylor, J., et al. (2000). *An introduction to support vector machines and other kernel-based learning methods*. Cambridge university press.

Dantzig, G. B. (1989). Making progress during a stall in the simplex algorithm. *Linear Algebra and its Applications*, 114:251–259.

Defazio, A., Bach, F., and Lacoste-Julien, S. (2014). Saga: A fast incremental gradient method with support for non-strongly convex composite objectives. *Advances in neural information processing systems*, 27.

Espin-García, O. and Rodríguez-Caballero, C. V. (2013). Metodología para un scoring de clientes sin referencias crediticias. *Cuadernos de economía*, 32(59):137–162.

Fawcett, T. (2006). An introduction to roc analysis. *Pattern recognition letters*, 27(8):861–874.

Finan, M. B. (2014). A basic course in the theory of interest and derivatives markets: A preparation for the actuarial exam fm/2. *Arkansas Tech University*.

financiera de Colombia, S. (2008). Circular externa 010.

Franco, P. (2017). Tecnología, inclusión financiera y regulación: acercando el financiamiento a las personas. In *21 Conferencia anual de la Asociación Latinoamericana e Ibérica de Derecho y Economía (ALACDE)–Universidad del Pacífico–Perú*, pages 1–38.

Friedman, N., Geiger, D., and Goldszmidt, M. (1997). Bayesian network classifiers. *Machine learning*, 29(2):131–163.

Gareth, J., Daniela, W., Trevor, H., and Robert, T. (2013). *An introduction to statistical learning: with applications in R*. Spinger.

Geurts, P., Ernst, D., and Wehenkel, L. (2006). Extremely randomized trees. *Machine learning*, 63(1):3–42.

Gimeno, P., Mingote, V., Ortega, A., Miguel, A., and Lleida, E. (2021). Generalizing auc optimization to multiclass classification for audio segmentation with limited training data. *IEEE Signal Processing Letters*, 28:1135–1139.

Gómez González, S. (2010). Análisis de riesgo crediticio y asignación de cupos de crédito.

González Fernández, K. L. (2016). Identificación de factores de riesgo en un scoring crediticio mediante técnicas de estadística espacial. *Escuela de Estadística*.

Granados, P. and Quijano, D. (2016). El financiamiento: más que un problema, una solución. *Realidad Empresarial*, (1):21–22.

Gutierrez Girault, M. A. (2007). Modelos de credit scoring: qué, cómo, cuándo y para qué.

Haimowitz, I. J. and Schwarz, H. (1997). Clustering and prediction for credit line optimization. In *Proceedings of AAAI-97 Workshop on AI Approaches to Fraud Detection & Risk Management*, pages 29–33.

Haykin, S. (2009). *Neural networks and learning machines, 3/E*. Pearson Education India.

Herga, Z., Rupnik, J., Škraba, P., and Fortuna, B. (2016). Modeling probability of default and credit limits. In *Conference on Data Mining and Data Warehouses*.

Heriberto, E. G. and Vicente, C. M. (2006). Diccionario económico financiero.

Hernández, P. A. C. (2004). Aplicación de árboles de decisión en modelos de riesgo crediticio. *Revista Colombiana de estadística*, 27(2):139–151.

Hosmer Jr, D. W., Lemeshow, S., and Sturdivant, R. X. (2013). *Applied logistic regression*, volume 398. John Wiley & Sons.

Huangfu, Q. and Hall, J. J. (2018). Parallelizing the dual revised simplex method. *Mathematical Programming Computation*, 10(1):119–142.

Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kleinbaum, D. G., Dietz, K., Gail, M., Klein, M., and Klein, M. (2002). *Logistic regression*. Springer.

Lattanzio Carrioni, S. and Pinilla Jaramillo, J. A. (2013). El microcrédito como herramienta de inclusión financiera para fomentar el desarrollo y crecimiento económico de colombia.

Lipton, Z. C., Elkan, C., and Narayanaswamy, B. (2014). Thresholding classifiers to maximize f1 score. *arXiv preprint arXiv:1402.1892*.

Lu, Y., Yang, L., Shi, B., Li, J., and Abedin, M. Z. (2022). A novel framework of credit risk feature selection for smes during industry 4.0. *Annals of Operations Research*, pages 1–28.

Maimon, O. Z. and Rokach, L. (2014). *Data mining with decision trees: theory and applications*, volume 81. World scientific.

Marrugo Arnedo, V. (2013). Crecimiento económico y desarrollo humano en colombia (2000-2010). *Revista de economía del Caribe*, (11):127–143.

McCallum, A., Nigam, K., et al. (1998). A comparison of event models for naive bayes text classification. In *AAAI-98 workshop on learning for text categorization*, volume 752, pages 41–48. Madison, WI.

Melo, L. F. and Granados, J. C. (2011). Regulación y valor en riesgo. *Ensayos sobre política económica*, 29(SPE64):110–177.

Michalewicz, Z. (1996). *Genetic algorithms + data structures = evolution programs*. Springer Science Business Media, Berlin, Heidelberg, 3rd edition.

Miled, K. B. H. and Rejeb, J.-E. B. (2015). Microfinance and poverty reduction: A review and synthesis of empirical evidence. *Procedia-Social and Behavioral Sciences*, 195:705–712.

Najafabadi, M. M., Khoshgoftaar, T. M., Villanustre, F., and Holt, J. (2017). Large-scale distributed l-bfgs. *Journal of Big Data*, 4(1):1–17.

Ochoa, J. C., Galeano, W., and Agudelo, L. G. (2010). Construcción de un modelo de scoring para el otorgamiento de crédito en una entidad financiera. *Perfil de Coyuntura Económica*, (16):191–222.

Patiño, O. A. (2008). Microcrédito historia y experiencias exitosas de su implementación en américa latina. *Revista Escuela de Administración de Negocios*, (63):41–58.

Paul, U. and Biswas, A. (2017). Consumer credit limit assignment using bayesian decision theory and fuzzy logic–a practical approach. *Uttiya Paul, Angshuman Biswas. Consumer Credit Limit Assignment Using Bayesian Decision Theory and Fuzzy Logic–A Practical Approach. Journal of Management*, 4(2).

Pollack, E. M. and García Hurtado, A. (2004). *Crecimiento, competitividad y equidad: rol del sector financiero*. CEPAL.

Redroban Bermudez, P. V. (2016). Proceso de asignacion de cupos de credito en empresas ferreteras y el enfoque iso/tc 176/sc 2/n 544r3. Master's thesis, Universidad de Guayaquil Facultad de Ciencias Administrativas.

Rodríguez, M. C. and Hernández, F. P. (2008). Gestión del riesgo crediticio: un análisis comparativo entre basilea ii y el sistema de administración del riesgo crediticio colombiano, sarc. *Cuadernos de Contabilidad*, 9(24).

Rodríguez Martínez, V. and Serna García, S. (2020). Las fintech, la nueva opción para el acceso a una banca ágil y económica para personas naturales y pymes.

Ross, S. M. (2014). *A first course in probability*. Pearson.

Schmidt, M., Le Roux, N., and Bach, F. (2017). Minimizing finite sums with the stochastic average gradient. *Mathematical Programming*, 162(1):83–112.

Shema, A. (2022). Effects of increasing credit limit in digital microlending: A study of airtime lending in east africa. *The Electronic Journal of Information Systems in Developing Countries*, 88(3):e12199.

Sohn, S. Y., Lim, K. T., and Ju, Y. (2014). Optimization strategy of credit line management for credit card business. *Computers & operations research*, 48:81–88.

Studer, M., Ritschard, G., Gabadinho, A., and Müller, N. S. (2011). Discrepancy analysis of state sequences. *Sociological Methods & Research*, 40(3):471–510.

Tibshirani, R. (2019). Newton's method. In *Notes for Convex Optimization: Machine Learning 10-725*.

Toh, K.-A., Kim, J., and Lee, S. (2008). Maximizing area under roc curve for biometric scores fusion. *Pattern Recognition*, 41(11):3373–3392.

Valenzuela González, G. et al. (2022). Aprendizaje supervisado: Métodos, propiedades y aplicaciones.

Vázquez, S. (2012). Crédito empresarial simple vs revolvente:¿ cuándo y para qué? *ESTUDIOS ECONÓMICOS CNBV Volumen 1, 2012*.

Wright, S. J. (2015). Coordinate descent algorithms. *Mathematical Programming*, 151(1):3–34.

Yang, T. and Zhang, X. (2022). Fintech adoption and financial inclusion: Evidence from household consumption in china. *Journal of Banking & Finance*, 145:106668.