Análisis económico
ISSN: 0185-3937
ISSN: 2448-6655
Universidad Autónoma Metropolitana, Unidad Azcapotzalco, División de Ciencias Sociales y Humanidades

Mota Aragón, Martha Beatriz; Moncayo Mejía, Pamela
Un análisis del perfil de riesgo en la dinámica de préstamos persona a persona mediante clústeres de K-medias
Análisis económico, vol. XXXVIII, núm. 99, 2023, Septiembre-Diciembre, pp. 119-144
Universidad Autónoma Metropolitana, Unidad Azcapotzalco, División de Ciencias Sociales y Humanidades

DOI: https://doi.org/10.24275/uam/azc/dcsh/ae/2023v38n99/Mota

Disponible en: https://www.redalyc.org/articulo.oa?id=41376280007

Análisis Económico, vol. XXXVIII, núm. 99, septiembre-diciembre de 2023, pp. 119-144, ISSN: 0185-3937, e-ISSN: 2448-6655

Un análisis del perfil de riesgo en la dinámica de préstamos persona a persona mediante clústeres de K-medias

An analysis of risk profiles in peer-to-peer lending through K-Means Clustering

Recibido: 25/octubre/2022; aceptado: 26/abril/2023; publicado: 20/septiembre/2023

Martha Beatriz Mota Aragón[*]
Pamela Moncayo Mejía[**]

doi.org/10.24275/uam/azc/dcsh/ae/2023v38n99/Mota

RESUMEN

Se analiza de manera empírica la base de datos de la Fintech estadounidense LendingClub, la empresa precursora del mercado de préstamos digitales persona a persona. Se busca contextualizar al modelo de negocio a través de su base de datos y las variables implicadas en el perfil de riesgo de los participantes. Se toman cuatro ventanas de tiempo para capturar ciclos económicos y su impacto en el perfil crediticio de los prestatarios. Para identificar los perfiles de riesgo de manera analítica, implementamos la técnica de clústeres de K-medias, definiendo un patrón de segmentación de participantes en función de la tasa de interés de los préstamos y la calificación FICO (Fair Isaac Company). Los patrones de agrupación son constantes a lo largo de los periodos estudiados y permiten determinar niveles promedio de ingreso, tipos de préstamos adquiridos y proveniencia geográfica según perfil de riesgo. Pocos estudios contemplan toda la información disponible de la operación de LendingClub (2007-20203T); por lo tanto, este estudio también identifica la transición a la madurez de este modelo de negocio. Este trabajo logra dos objetivos: demostrar la evolución de los clientes y la plataforma a través del tiempo en cuanto se refiere a los perfiles de riesgo de los participantes prestatarios, y destacar el uso de clústeres de K-medias como una herramienta apropiada para el perfilamiento frente a otras alternativas analíticas.

Palabras Clave: clústeres de K-medias, LendingClub, Fintech, préstamos persona a persona, riesgo de crédito.
Clasificación JEL: C24; G23; O16.

ABSTRACT

We empirically analyze the LendingClub dataset, the pioneer of the peer-to-peer digital lending market. We aim to contextualize the business model through its database and the variables involved in the risk profile of the participants. We consider four different periods to capture economic cycles and their impact on the

[*] Universidad Autónoma Metropolitana-Iztapalapa. CDMX, México. Correo electrónico: beatrizmota4@gmail.com
[**] Escuela de Negocios del Tecnológico de Monterrey, EGADE Santa Fe, CDMX, México. Correo electrónico: pamemoncayom@gmail.com

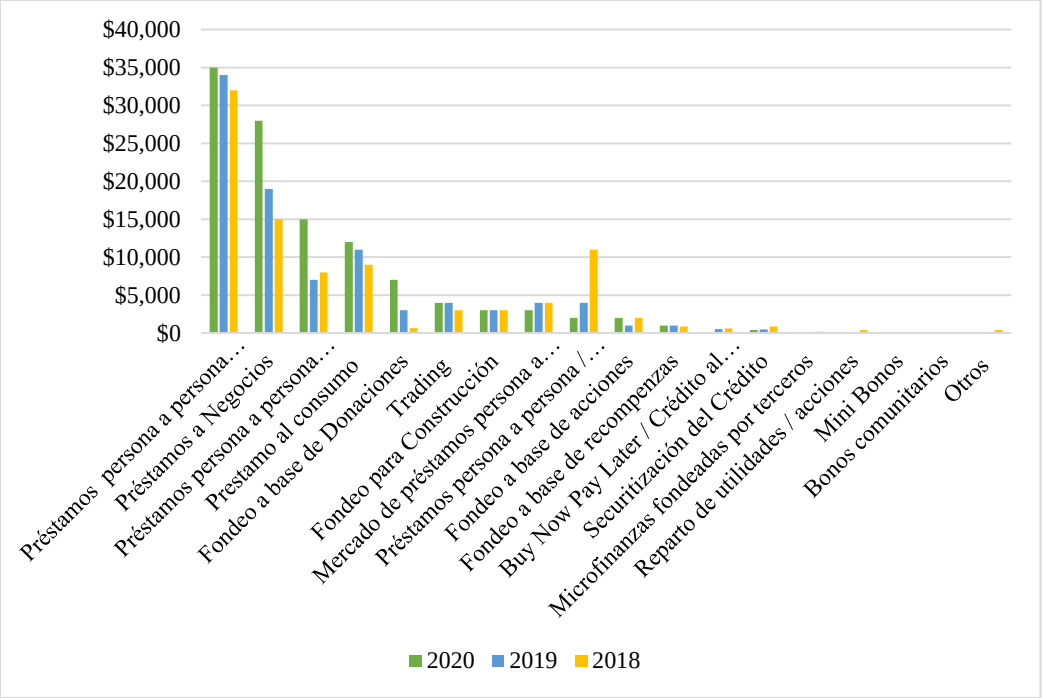
credit profile of borrowers. We implement K-Means Clustering Analysis to analytically define a segmentation pattern for participants based on the interest rate of the loans and the FICO (Fair Isaac Company) score. The clusters are constant throughout the periods studied and allow to detect average income levels, issued loans destination and borrower geographic origin according to risk profile. Few studies contemplate all the information available for the LendingClub operation (2007-2020Q3), therefore, this study illustrates how this business model has transitioned to maturity. This work achieves two objectives: to demonstrate the evolution of clients and the platform over time in terms of the risk profiles of participating borrowers, and to highlight the use of K-Means Clustering as an appropriate tool for profiling against to other analytical alternatives.
Keywords: K-Means Clustering, LendingClub, Fintech, Peer-to-peer lending, Credit Risk.
JEL Classification: C24; G23; O16.

INTRODUCCIÓN

La innovación financiera es un tema que se ha abordado constantemente y a partir de diferentes enfoques, tales como la inclusión financiera, la diferenciación de productos financieros y la competitividad de los mercados (Arner *et al*., 2015, 2016; Funk *et al*., 2011; Lee y Shin, 2018). Más conocida como Fintech, la innovación en los servicios financieros es un tema que ha causado gran curiosidad a nivel industrial y académico, ya que ha revolucionado la forma en la que los usuarios acceden a diferentes servicios por medio de internet (Nüesch *et al*., 2015). La ubicuidad y la facilidad de acceso es lo que ha motivado a los usuarios la preferencia por estas nuevas formas de acceder al sistema financiero, por lo que observamos un desarrollo intensivo de las aplicaciones de banca en línea y la proliferación de *startups* enfocadas en proveer servicios Fintech, como medios de pago y acceso al crédito digital, tal como se observa en el caso mexicano, con la introducción del medio de transacción CoDi. (Bowley, 2011; Herrera-Arizmendi y Amezcua-Núñez, 2020; Mia *et al*., 2007)

El ecosistema Fintech se conforma de varios modelos de negocios, donde predomina la presencia de los mercados de préstamos de consumo en la modalidad "persona a persona". De acuerdo con el reporte de *Cambridge Center for Alternative Finance* (Ziegler *et al*., 2021), el volumen de operaciones de este mercado alcanzó los 3.5 billones de dólares americanos en el año 2020, encabezando la lista de modelos de negocio de finanzas alternativas durante tres años consecutivos. Este reporte muestra la actividad y el volumen de operaciones en distintas regiones como Latinoamérica y el Caribe, Norteamérica (Estados Unidos y Canadá) y Asia Pacífico, donde se encuentra que el crédito de consumo, especialmente bajo la modalidad P2P (persona a persona) es predominante frente a otros modelos de negocio. En la figura 1 se presenta una gráfica del volumen global por modelo de negocio. Por este motivo, encontramos trascendente identificar las características que componen a los participantes de esta dinámica y las áreas de oportunidad presentes en el mercado de préstamos digitales.

Figura 1
Volumen de negocios en Fintech (2018 – 2020)



Fuente: Cambridge Center for Alternative Finance (Ziegler *et al*., 2021)

En la literatura científica observamos un amplio rango de estudios referentes a los préstamos digitales persona a persona, donde se busca determinar los factores que dirigen a este mercado y a sus participantes, así como las características de los productos que compiten con los servicios financieros ofrecidos por la banca tradicional. (Berger y Gleisner, 2014; Mundet y Gutiérrez, 2015). Algunas de las ventajas que presentan estos mercados de préstamos son los bajos costos de transacción y de intermediación financiera por una reducción en monitoreo, evaluación y procedimientos de recuperación de préstamos (Mills y McCarthy, 2014; Wang y Hua, 2014; Wardrop *et al*., 2015). De igual manera la transferencia de costos de operación hacia los participantes de la dinámica también permite ofrecer tasas competitivas para inversionistas y prestatarios en comparación con la banca tradicional (Demirgüç-Kunt y Huizinga, 1999; Maudos y de Guevara, 2004). Finalmente, el éxito de las plataformas de préstamos digitales es atribuible a una evidente contracción del crédito y el endurecimiento de los requisitos aprobatorios. Estas plataformas proveen acceso al financiamiento al segmento de mercado que no resulta de interés para la banca tradicional, así como lo hacen las microfinancieras y las uniones de crédito. De acuerdo con la eficiencia de Pareto, el ochenta por ciento del segmento bancarizado estará atendido por la banca de consumo tradicional, dejando el veinte por ciento restantes para instituciones financieras más pequeñas y para las plataformas de finanzas alternativas como las Fintech, y en este caso específico, aquellas que proveen un mercado para el acceso a préstamos persona a persona. (Dehejia *et al*., 2012; Hales, 1995; Koch, 2011; Peppard, 2000).

En este artículo se propone la revisión de la base de datos de LendingClub, plataforma de préstamos persona a persona localizada en Estados Unidos. Posterior al análisis empírico de toda la información de esta base de datos, se emplea el análisis de clústeres de K-medias, para identificar segmentaciones de los participantes de la dinámica, en función de dos variables relevantes a lo largo de cuatro periodos distintos. Se identifica que existe una relación directa entre la tasa de interés y la

calificación FICO. La calificación FICO es un reconocido modelo de *scoring* desarrollado por la empresa Fair Isaac Company que contempla distintos factores de la información crediticia para asignar una calificación numérica entre 300 y 850. A partir de tres clústeres definidos en el análisis se encuentra que los participantes con menor calificación FICO reciben la asignación de la tasa de interés más alta. Así mismo, se identifican valores promedio respecto a ingreso del prestatario, tasa de interés promedio y calificación FICO promedio por cada segmento identificado. El resultado de este análisis muestra que la metodología de clústeres de K-medias permite identificar la evolución de los perfiles de los participantes de la dinámica de préstamos en la fintech LendingClub a lo largo de su operación, generando agrupaciones congruentes con las variables estudiadas, sean los ingresos anuales, la tasa de interés asignada para el préstamo y el motivo del préstamo asociado a la solicitud. El resto del artículo se divide en I. Revisión de Literatura, II. Metodología, III, Descripción de la base de datos, IV. Resultados del análisis clústeres de K-medias y al final se presentan las conclusiones.

I. Revisión de Literatura

Dado el contexto del acceso al financiamiento, y las restricciones que imponen las instituciones financieras tradicionales, es importante identificar los niveles de riesgo presentes en las dinámicas de préstamos digitales persona a persona, ya que usualmente los participantes poseen de forma personal un nivel de riesgo que no resulta atractivo para los bancos, ya sea porque carecen de un historial crediticio formal, o porque no han sabido llevar adecuadamente sus productos de crédito. La investigación académica señala ciertas cuestiones referentes al riesgo del mercado de préstamos digitales, una de ellas es la asimetría de información existente entre los participantes de la dinámica (Stiglitz y Weiss, 1981). Los inversionistas no conocen con certeza el nivel de riesgo actual de los individuos que solicitan financiamiento. Si bien la plataforma está encargada de seleccionar a los mejores candidatos para participar en el mercado de préstamos persona a persona, existe un nivel de riesgo inherente en estos participantes. Se ha determinado que el rendimiento no compensa el riesgo en el mercado de créditos digitales y que usualmente los inversionistas son propensos a fondear prestatarios con un nivel de riesgo que ellos no estipulaban, atraídos por tasas de retorno más altas (Emekter *et al*., 2015).

Se ha evidenciado que la información referente a cada prestatario presentada por la plataforma incrementa la probabilidad de una mejor selección por parte del inversionista. Esto no solo beneficia el retorno esperado del inversionista, además incrementa el éxito en recuperación del crédito, de modo que se enfatiza en la necesidad de presentar información financiera relevante y hasta cierto punto sofisticada, como aquella que ofrecen las calificadoras de riesgo, por ejemplo, Fair Isaacs Company y la calificación FICO. La contraparte estaría en el nivel de instrucción y de educación financiera de cada inversionista, y qué esfuerzo realiza la plataforma para incrementar el nivel de conocimiento técnico de los participantes de la dinámica (Cumming y Hornuf, 2020)

Otra cuestión discutida respecto a la dinámica del crédito digital involucra la relación riesgo rendimiento, donde se determina que los perfiles de riesgo de los prestatarios están inversamente relacionados al rendimiento ofrecido por la plataforma. (Adhami *et al*., 2019) Esto contrasta con los principios financieros, donde se establece que el rendimiento está positivamente relacionado al nivel de riesgo, de modo que las plataformas de crédito digital no establecen un precio por el riesgo y los inversionistas están aceptando tomar riesgos por retornos menores o iguales. Desde el punto de vista del prestatario y bajo la premisa del préstamo persona a persona, se han estudiado los determinantes de éxito de fondeo de los préstamos solicitados, entendiendo que en este mercado los inversionistas deciden a quién financiar. La investigación académica existente menciona que algunos factores determinantes son el perfil del prestatario, su foto de perfil y en ciertas ocasiones su red de conocidos (redes sociales) y se atribuyen estos factores al nivel de confianza que genera un usuario. Por otro lado, se ha estudiado el

comportamiento de manada[1] como otro factor determinante del fondeo exitoso, los inversionistas fondearán más rápido aquellas solicitudes que ya cuenten con un nivel considerable de fondeo previo. (Gonzalez y Loureiro, 2014; E. Lee y Lee, 2012; Yum *et al*., 2012; Zhang y Liu, 2012).

Un factor que intercepta las necesidades del inversionista y del prestatario es el nivel de confianza en la plataforma. Dado que estos desconocen el riesgo asumido por participar en la dinámica de préstamo persona a persona a nivel técnico, existe una tendencia a confiar en el proceso de selección de prestatarios, así como la generación de oportunidades de fondeo (Moreno-Moreno *et al*., 2018; Moreno-Moreno *et al*., 2019), de modo que un factor de éxito sustancial dentro de estos mercados de crédito digital es el nivel de confianza y la reputación que cada plataforma logra construir.

Si bien la perspectiva de los usuarios y participantes de la dinámica de los préstamos persona a persona es interesante y relevante dentro de la investigación académica, es importante determinar los factores de riesgo que se deben manejar a nivel plataforma. A pesar de fungir como un mercado, la plataforma debe asegurar su rentabilidad y una operación eficiente, lo que derivará en un nivel de confianza mayor por parte de sus usuarios. Como se mencionó, la plataforma estará encargada de propiciar el mejor ambiente para inversión, lo que se traduce en seleccionar prestatarios que no representen niveles de incumplimiento elevados. De igual forma, están obligados a una gestión eficiente del precio del riesgo y de asesorar adecuadamente a sus inversionistas; todo esto con el fin de generar una relación ganar-ganar para todos los participantes. Dentro de la literatura existente se identifican actividades que involucran el análisis del riesgo de crédito, como la asignación de calificaciones de riesgo y la creación de modelos de predicción de *default*. En estas actividades se identifica la necesidad del uso de información alternativa e historiales crediticios en caso de ser posible, para una correcta asignación de calificaciones de riesgo de los prestatarios y la identificación de las probabilidades de incumplimiento. (Serrano-Cinca *et al*., 2015)

Tanto la definición de las calificaciones de riesgo de los prestatarios como los modelos de riesgo de crédito dependen de características del prestatario en sí y de las características del producto de crédito otorgado; por ejemplo, los créditos otorgados para ciertos propósitos presentan más riesgos de incumplimiento, así como los factores socioeconómicos también llegan a influir en la intención de pago de los usuarios. Los criterios empleados de forma común a la hora de determinar el nivel de riesgo de un prestatario son los niveles de endeudamiento, la situación de empleo y vivienda, escolaridad y patrimonio (Agarwal *et al*., 2007; Altman *et al*., 2004; Baesens *et al*., 2005; Komrattanapanya y Suntraruk, 2013). En el caso de usuarios no bancarizados (usualmente jóvenes adquiriendo sus primeros productos financieros), se propone el uso de información alternativa proveniente de redes sociales y otros factores que pudiesen fungir como predictores del poder de adquisición y la rentabilidad de estos usuarios (Serrano-Cinca y Gutiérrez-Nieto, 2016)

A medida que las plataformas recopilan esta información, pueden formar bases de datos para modelar el riesgo de crédito, además se pueden nutrir de fuentes de información alternativa (Jagtiani y Lemieux, 2019) provenientes del *Big Data* para enriquecer sus modelos y adaptarlos mejor a las circunstancias actuales. Sin embargo, la problemática radica en la ausencia de información histórica para el análisis de riesgo pertinente, lo que suele suceder en los inicios de operación de las plataformas Fintech. Ante esta situación, las plataformas deberán basarse en aproximaciones a sus datos y a sus clientes mediante *benchmark* e información pública disponible.

Existen muy pocas fuentes de información pública relacionada al mercado de préstamos digitales persona a persona, el secretismo está muy presente en esta industria puesto que es nueva, y sumamente competitiva. Una fuente que se puede considerar confiable es la base de datos de créditos de *Lending*

[1] Las decisiones de los inversionistas se ven influenciadas por otros inversionistas. El riesgo radica en que la estrategia sea errónea y "en manada" se tomen decisiones perjudiciales. (Bikhchandani & Sharma, 2000).

Club, ya que es una plataforma que opera desde el año 2007 en el mercado más grande de préstamos persona a persona a nivel global (dejando China de lado). Esta plataforma ha sido sujeto de estudio académico en diversas temáticas que incluyen el riesgo de crédito bajo diversas perspectivas de análisis.

Usualmente, la temática recurrente involucra el riesgo de incumplimiento y la predicción de posibles *defaults* en función de las características del prestatario mediante diversas alternativas como regresiones, árboles de decisión, modelación de funciones de distribución de las probabilidades de incumplimiento, o mediante inteligencia artificial (redes neuronales, entrenamiento máquina o *machine learning*) (Calabrese *et al*., 2019; Calabrese y Zanin, 2022; Chengeta y Mabika, 2021; Kim y Cho, 2019; Kriebel y Stitz, 2022) Estos estudios han definido patrones que determinan cómo se atribuye la recuperación de los préstamos a las características de cada usuario. Generalmente se observa que los niveles de endeudamiento aunados a la situación laboral son indicadores determinantes del éxito o fracaso de la recuperación.

Como antecedente respecto a las metodologías utilizadas para el perfilamiento de los prestatarios, a razón de su capacidad de pago y la expectativa de recuperación de los préstamos otorgados encontramos los clásicos del análisis de riesgo de crédito: *CreditMetrics*, que identifica la probabilidad de migración de un cliente entre distintos perfiles de riesgo en un horizonte de tiempo; el modelo de Merton (Merton, 1974) que determina la probabilidad de *default* en función de la calidad de los activos de un agente económico; *CreditRisk+* desarrollado por Credit Suisse, cuyo modelo define al *default* como un proceso exógeno que sigue una distribución de Poisson; finalmente, *CreditPortfolioView* define la probabilidad de default bajo la condicional del estado de variables macroeconómicas (Crouhy *et al*., 2000). Estas metodologías de análisis de riesgo son ampliamente utilizadas por instituciones financieras, que, en su mayoría otorgan préstamos a largo plazo de montos considerables, es por esto que la regulación de Basilea exige que los bancos pertenecientes al G-10 utilicen y reporten sus hallazgos considerando estas metodologías mencionadas previamente, ya que son aquellas con mayor respaldo académico y metodológico, considerando que la estabilidad financiera de estas instituciones tienen un impacto en la economía global, por ende, es necesario que sus niveles de riesgo sean controlados metódicamente.

Desde la perspectiva de los créditos de consumo a escala menor, como es el caso de muchas Fintech en la actualidad, es complejo llegar a implementar los modelos clásicos de análisis de riesgo de crédito, ya que la información requerida para alimentar dichos modelos está limitada tanto a la información histórica disponible como al cumplimiento de los supuestos que se observan en las metodologías clásicas. Un claro ejemplo es el siguiente: El modelo de JP Morgan, *CreditMetric*s (Morgan, 1997) requiere una matriz de transición de las migraciones de estatus o de calidad crediticia para su portafolio de crédito. Grandes instituciones financieras adquieren esta información por parte de Moody's o S&P's, quienes han identificado las probabilidades de migración en esta matriz utilizando información histórica registrada del comportamiento crediticio y los efectos de los cambios del mercado sobre los portafolios de crédito. En otro ejemplo, el modelo de KMV considera la distribución del rendimiento de los activos del deudor. Una Fintech tiene acceso limitado a esta información, ya que está apostando a un nicho de mercado y bancarizado por definición. Por lo tanto, es evidente que estos modelos clásicos presuponen muchas dificultades a la hora de cumplir con todos los supuestos.

Por este motivo, en la literatura académica podemos encontrar una gama amplia de propuestas que divergen de los modelos clásicos, parcialmente o en su totalidad. La naturaleza de la información a la que tienen acceso las Fintech al momento de originar los créditos propicia que los modelos de análisis de crédito y de calificación sean "enfocados en los datos" o "*data driven*" como se conoce en inglés. Por ejemplo, observamos el uso de árboles de decisión para identificar posibles fraudes crediticios (Sahin y Duman, 2010), así como la calidad de los prestatarios en función de las variables a las que se tenga acceso (Zhang *et al*., 2010). Los modelos basados en árboles de decisión permiten obtener resultados de fácil interpretación que permiten identificar las características que generan ciertos eventos, como la

presencia de posibles impagos en función de características sociodemográficas de un prestatario o la probabilidad de estar frente a un evento de fraude.

Otra temática observada en la literatura es la implementación de análisis de clústeres para identificar cierta estructura en los mercados de préstamos persona, a persona o segmentar la población perteneciente a esta dinámica, a fin de identificar prestatarios similares en función de alguna característica propia de la información presente. (Li, 2016) Así mismo, Pujun *et al*., (2016) proponen el uso de análisis de clústeres como una aproximación visual para identificar estructura en la población de la base de datos de LendingClub. Los autores afirman que, dada la complejidad y dimensionalidad de la base de datos, no es posible identificar de forma visual alguna estructura; sin embargo, mediante la selección de variables específicas es posible encontrar algunos patrones de agrupación, por ejemplo, tomando en cuenta el propósito del préstamo. Lim y Sohn (2007) utilizan clústeres para definir un modelo dinámico de calificación crediticia, donde se evalúan los cambios en las características del prestatario posterior a otorgar el préstamo. Identifican que esta perspectiva dinámica permite una identificación temprana de posibles incumplimientos al agrupar prestatarios que muestran deterioro de sus capacidades de pago. El análisis de clústeres es ampliamente utilizado en distintas disciplinas dada la flexibilidad del modelo y su interpretabilidad.

II. METODOLOGÍA

El objetivo de este artículo es caracterizar de manera empírica a la población perteneciente a la base de datos de LendingClub en función del estatus del crédito, tomando todas las observaciones y haciendo un análisis estadístico en tres ventanas de tiempo que fungen como representación de distintos ciclos económicos, tomando en cuenta que: LendingClub emergió en la época de la Gran Recesión producto de la crisis hipotecaria (*subprime)* en Estados Unidos y la contracción del crédito; tuvo actividad durante la época post recesión, cuando el Banco Central redujo las tasas de interés para incentivar la economía; y durante la pandemia por COVID-19 en el año 2020.

Mediante estas ventanas temporales se puede extraer información relevante respecto al perfil socioeconómico y la colocación de productos de crédito digital en el mercado de préstamos persona a persona más grande de la región. Bajo esta premisa, se evaluará la estadística descriptiva en los periodos indicados, así como las variables que caracterizan el contexto socioeconómico de los participantes de la dinámica, así como su historial crediticio al momento de suscribir el préstamo. De igual forma, se empleará la técnica de clústeres de K-*medias*, con la finalidad de identificar subconjuntos homogéneos no observables dentro de la base de datos, que permitan identificar rasgos determinantes en la población participante del préstamo digital.

Clústeres de K-medias

La técnica de Clústeres de K-medias permite identificar K subconjuntos no superpuestos dentro de una base de datos. Es necesario definir la cantidad de clústeres para efectuar la partición de la información; para determinar esta cantidad se emplea la técnica del codo (Syakur *et al*., 2018). El algoritmo de clústeres de k-medias asignará cada observación a exactamente uno de los K clústeres. El procedimiento de este algoritmo resulta de un problema matemático bastante intuitivo. Se tiene $C_1, \dots, C_K$ que representan los conjuntos que contienen los índices de las observaciones en cada clúster. Estos conjuntos satisfacen las siguientes propiedades:

1. $C_1 \cup C_2 \cup \dots \cup C_k = \{1, \dots, n\}$. En otras palabras, cada observación pertenece a al menos uno de los K clústeres.

2. $C_k \cap C_{k'} = \emptyset$ para cada $k \neq k'$. En otras palabras, los clústers no se superponen. Una observación no pertenece a más de un clúster.

Por lo tanto, si la i-ésima observación pertenece al k-ésimo clúster, entonces $i \in C_k$. El objetivo de esta metodología es que la variación dentro de cada clúster sea lo más pequeña posible. Esta medida $W(C_k)$ para el clúster C_k representa la cantidad por la cual la observación dentro de un clúster difiere entre clústeres. Por lo tanto, se busca minimizar el siguiente problema:

$$\min_{C_1,\ldots,C_k} \left\{\sum_{k=1}^{K} W(C_k\right\}. \quad (1)$$

Esta fórmula indica que se busca particionar las observaciones en K clústeres de forma que la variación total entre clústeres, sumada entre los K clústeres determinados, sea la más pequeña posible. Para resolver la ecuación 1, es necesario definir la variación entre clústeres. Una alternativa común es la implementación de la distancia euclidiana, de forma que podemos definir:

$$W(C_k) = \frac{1}{|C_k|}\sum_{i,i' \in C_k}\sum_{j=1}^{p}\left(x_{ij} - x_{i'j}\right)^2, \quad (2)$$

Donde $|C_k|$ representa el número de observaciones en el k-ésimo clúster. En otras palabras, la variación entre clústeres para el k-ésimo clúster es la suma de todas las distancias euclidianas al cuadrado entre las observaciones del k-ésimo clúster, dividido por el número total de observaciones del k-ésimo clúster. Combinando las ecuaciones (1) y (2) se obtiene el problema de optimización que define la técnica de Clústeres de K-medias.

$$\min_{C_1,\ldots,C_K} \left\{\sum_{k=1}^{K} \frac{1}{|C_k|} \sum_{i,i' \in C_k}\sum_{j=1}^{p}\left(x_{ij} - x_{i'j}\right)^2\right\}. \quad (3)$$

El algoritmo de clústeres de K-medias asigna números aleatorios de 1 a K para cada observación de modo que se genera una partición inicial de los datos. Posteriormente, se realiza un proceso iterativo para minimizar la función objetivo descrita en (3). Para cada K clústeres, se identifica el centroide del clúster. El centroide es el vector de las p medias para las variables o características de las observaciones pertenecientes al k-ésimo clúster. Finalmente se asigna cada observación al clúster cuyo centroide sea más cercano usando la distancia euclidiana. (James *et al*., 2013)

Método del codo

El método del codo es una aproximación visual a la cantidad de clústeres óptimos para resolver el problema de minimización establecido en (3). El método consiste en graficar la variación explicada en función del número de clústeres, tomando "el codo" de la curva para determinar el número ideal de clústeres necesarios para particionar un conjunto de datos. (Dangeti, 2017).

III. DESCRIPCIÓN DE LOS DATOS

La base de datos de LendingClub (años 2007-2020 Tercer trimestre) cuenta con 142 variables y 2,916,374 observaciones. Esta base de datos es un corte transversal del estado de los préstamos a lo largo de la vida operacional de LendingClub, por lo que observamos préstamos que han sido pagados, impagos, atrasados, en mora, etc. Así como diversas variables que caracterizan el contexto socioeconómico de los individuos pertenecientes a esta base de datos. No es posible identificar si existe un individuo con operaciones repetidas a lo largo del tiempo porque no existe la variable de identificación disponible.

Antes de proceder con la implementación de cualquier modelo econométrico o de *Machine Learning* es necesario hacer un filtrado inicial de la información. Esta base de datos presenta grandes cantidades de valores faltantes para algunas variables que representan el historial crediticio de los prestatarios, en este estudio descartamos todas las variables que presenten un porcentaje de valores faltantes superior al treinta por ciento. Este argumento fundamenta la decisión de evaluar la base de datos en distintas temporalidades, de modo que sea posible aprender de la información de una forma más homogénea respecto a la influencia de las condiciones económicas sobre los usuarios de esta plataforma en cada ventana de tiempo estudiada. A continuación, en la tabla 1, se presenta un resumen de la cantidad de observaciones y variables disponibles por periodo:

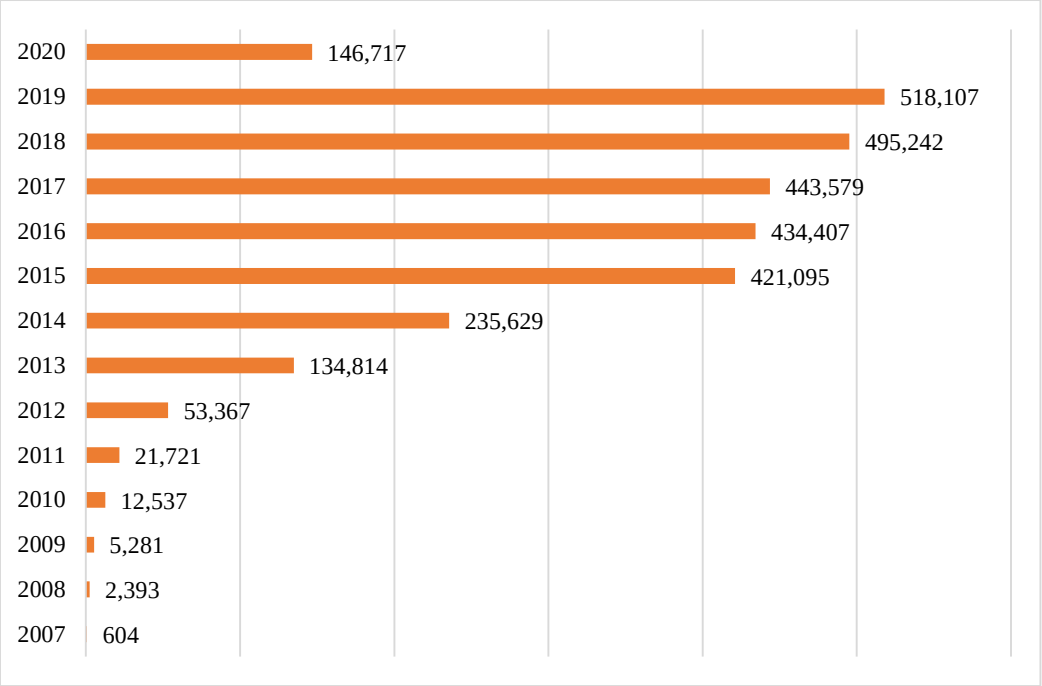
Tabla 1
Préstamos colocados por año

PERIODO	# OBSERVACIONES	# VARIABLES
TOTAL	2,925,493	99
2007-2011	42,536	53**
2012-2016	1,279,312	87**
2017-2019	1,456,928	99
2020	146,717	99

**La diferencia en el número de variables por periodo es el resultado de la presencia de variables con un alto porcentaje de valores faltantes que se eliminaron en el preprocesamiento. El periodo con mayor cantidad de variables con valores faltantes es 2007-2011. Se asume que esto es producto de transiciones en las políticas de recolección de información crediticia para los participantes de la dinámica.
Fuente: Elaboración propia.

El gráfico de barras en la Figura 2, ilustra la distribución de los préstamos colocados a lo largo de la operación de LendingClub como plataforma de préstamos persona a persona. Es interesante observar el crecimiento cuasi exponencial de la colocación hasta la llegada del año 2020. Se debe recordar que la plataforma retiró sus “Notas” de inversión (el instrumento mediante el cual se efectuaban los préstamos persona a persona) en el último trimestre del año 2020. La colocación en este año iguala al año 2013, cifra que genera la inquietud del impacto de la pandemia por COVID-19 sobre la actividad económica que se llevaba a cabo en esta plataforma.

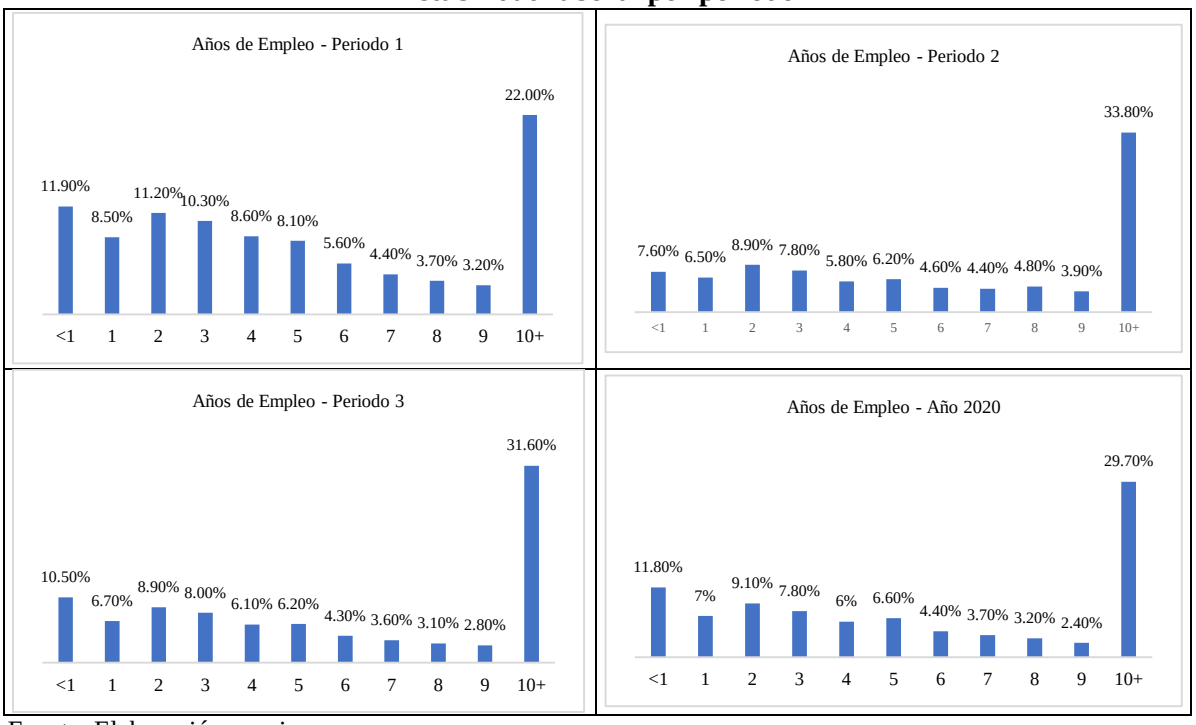
Figura 2
Gráfico de barras. Préstamos por año de colocación



Fuente: Elaboración propia

A continuación, se revisarán algunas variables de interés que representan las características de los participantes de la dinámica y de los préstamos que se les otorgaron. Este proceso permite identificar a grandes rasgos el comportamiento de variables que se consideran relevantes para el análisis de crédito, como el monto del préstamo, la tasa de interés asignada, el nivel de endeudamiento del cliente y su estabilidad laboral, entre otros. De forma inicial se revisan las características del prestatario para finalizar con las variables que representan al préstamo.

Figura 3
Estabilidad laboral por periodo



Fuente: Elaboración propia.

Para la variable que determina la estabilidad laboral (que hace referencia al tiempo que el prestatario ha conservado un puesto) observamos que existe una predominancia de personas que mantienen su posición con un mismo empleador por más de diez años (Figura 3). La distribución de las demás categorías de empleo es más homogénea, sin embargo, es posible identificar que, en el primer periodo, el tercer periodo y el año 2020 cuenta con un once por ciento de prestatarios que tienen menos de un año de antigüedad en su posición laboral. Respecto al riesgo, existe preferencia por clientes con mayor estabilidad laboral, y esto se ve reflejado en la base de datos de LendingClub. Se podría pensar que la cantidad de prestatarios con antigüedad mayor podría equiparar el riesgo de aquellos que presentan antigüedades menores.

Respecto a la variable de ingreso anual "annual_inc", se identifican valores extremos para el nivel superior. Esta es una variable que los prestatarios ingresan como campo de texto en sus solicitudes. La plataforma puede hacer una verificación de la fuente de ingreso, así como el monto, pero existe un porcentaje considerable de ingresos anuales no verificados. Es posible que los valores extremos que presenta esta variable se deban a un error por parte de los prestatarios al momento de generar su solicitud. En la Tabla 2, se presenta un resumen de la estadística descriptiva de esta variable para los periodos establecidos.

Tabla 2
Estadística descriptiva – Ingreso Anual (en USD)

Ingreso Anual	Periodo 1	Periodo 2	Periodo 3	Año 2020
Media	69,136.6	76,740.6	82,009.8	90,361.2
Desv. Estándar	64,096.3	69,259	139,399.9	112,868.2

Límite Inferior	1,896	0	0	0
25%	40,000	46,131.5	47,880	52,000
50%	59,000	65,000	68,000	75,000
75%	82,500	91,000	98,040	106,000
Límite Superior	6,000,000	9,573,072	110,000,000	9,999,999

Fuente: Elaboración propia

Ante esta situación se podría optar por eliminar los valores extremos de los límites superior e inferior, o utilizar otra variable que pueda contener la información del ingreso anual. En este caso, la base de datos de LendingClub presenta la variable Ratio deuda ingreso "DTI – debt to income". Esta variable podría suplir tanto el ingreso como otras variables que representan el nivel de endeudamiento, y, por lo tanto, la calidad crediticia de cada prestatario. Sin embargo, dado que esta variable se construye utilizando el valor de ingreso reportado por cada prestatario, es necesario eliminar observaciones que presenten valores extremos en el ingreso para evitar sesgar cualquier modelo. Otra alternativa es utilizar solamente las observaciones cuyo ingreso anual haya sido verificado. Para el Periodo 1, el 50% de los ingresos son verificados, 70% en el Periodo 2, 60% en el periodo 3, y 50% en el año 2020. La eliminación del 40% de la base por la variable de ingreso anual podría representar un costo muy alto respecto a las características que generen eventos de *default* en un análisis de riesgo de crédito, por lo tanto, la opción más conservadora sería eliminar solamente las observaciones que sobrepasen un umbral predefinido en función de las necesidades de análisis.

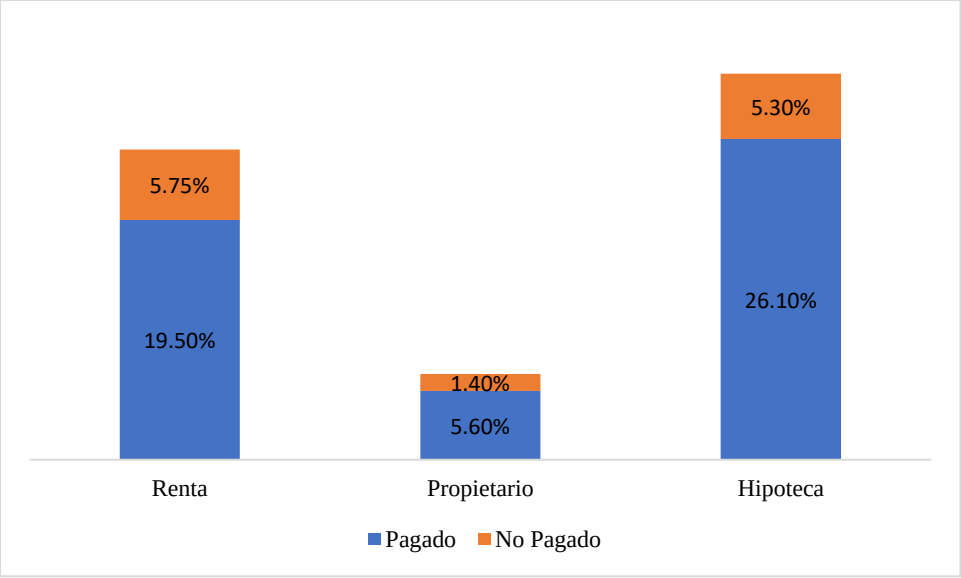
Otra variable que representa el estado socioeconómico de los participantes de la dinámica es su situación de vivienda que se divide en tres categorías. Casa propia, Hipoteca y Casa rentada. En la Tabla 3, se presenta la distribución de observaciones de acuerdo con esta variable para los periodos definidos. Se observa que existe una predominancia de prestatarios que adquirieron una hipoteca y que rentan su vivienda. Por una parte, una persona con un crédito hipotecario tiene ya una relación bancaria y, por lo tanto, información disponible respecto a su historial crediticio. Desde la perspectiva de riesgo de crédito, puede ser interesante saber cómo fue el comportamiento de pago del prestatario en función de su situación de vivienda. En la Figura 4, se presenta el porcentaje de impagos en función de la situación de vivienda para toda la base de datos.

Tabla 3
Porcentajes en Situación de Vivienda

Situación de vivienda	Periodo 1	Periodo 2	Periodo 3	Periodo 4
Casa Propia	8%	10%	12%	12%
Hipoteca	45%	50%	49%	49%
Renta	48%	40%	39%	39%

Fuente: Elaboración propia.

Figura 4
Situación de vivienda vs Default en préstamos (Base de datos completa)



Fuente: Elaboración propia

Para representar los préstamos pagados, utilizamos el color azul, para los impagos, el color rojo. Por lo que se observa, existen pocas observaciones cuyo estado del préstamo representa default. A simple vista, no se podría determinar que un prestatario que rente su vivienda o que mantenga un crédito hipotecario, presente una propensión al default. La Tabla 4, muestra el porcentaje de defaults en función de la situación de vivienda. En el Periodo 4 observamos que existen porcentajes sumamente bajos porque la mayor parte de esos préstamos siguen vigentes.

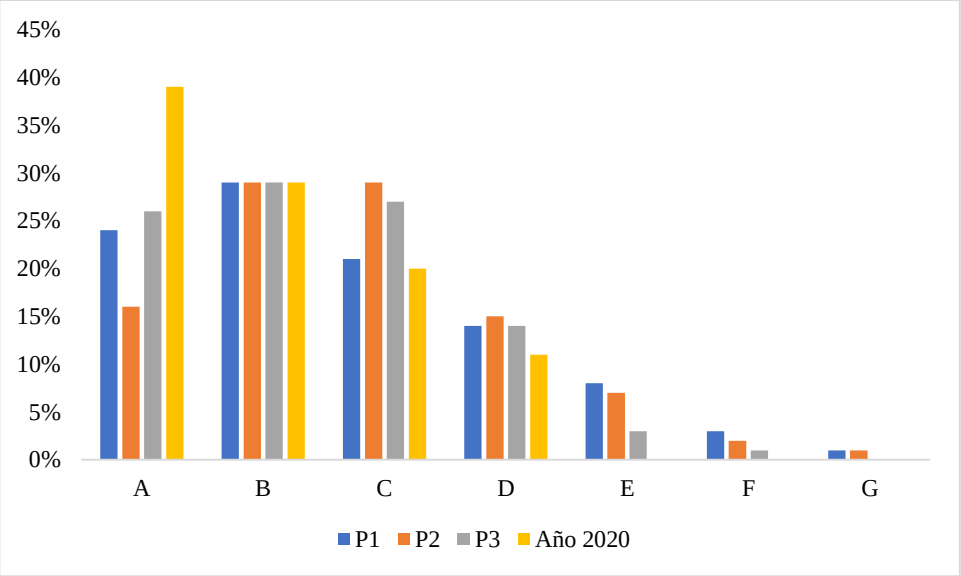
Tabla 4
El Default en función de la situación de vivienda

Default	Periodo 1	Periodo 2	Periodo 3	Periodo 4
Casa Propia	1%	2%	1.1%	0.05%
Hipoteca	5.5%	7.8%	3.6%	0.02%
Renta	6.8%	8%	4%	0.01%

Fuente: Elaboración propia

Dentro de las variables que representan la calidad o el historial crediticios de los prestatarios, se destacan dos; la calificación FICO y la calificación de riesgo asignada por LendingClub (“fico_range” y “grade” respectivamente). La variable “grade” se representa por categorías de la “A” a la “G”, donde A significa que el prestatario tiene bajo riesgo, y G indica mayor riesgo. La asignación de la tasa de interés para cada prestatario, y por lo tanto la tasa de rendimiento de los inversionistas, se realiza en función de esta calificación de riesgo más una comisión por servicio. Adicionalmente, existen cinco subcategorías por cada clasificación, representadas por la variable “sub_grade”. Únicamente el año 2020 solo presenta calificaciones de riesgo de la “A” a la “D”. En la Figura 5, se muestra la distribución de las subcategorías para la variable de calificación de riesgo, para todos los años disponibles.

Figura 5
Distribución calificación de riesgo



Fuente: Elaboración propia

Tabla 5
Calificación de riesgo por periodos

Periodo	A	B	C	D	E	F	G
P1	24%	29%	21%	14%	8%	3%	1%
P2	16%	29%	29%	15%	7%	2%	1%
P3	26%	29%	27%	14%	3%	1%	-
Año 2020	39%	29%	20%	11%	-	-	-

Fuente: Elaboración propia

En la base de datos existen cuatro calificaciones FICO disponibles; el primer par son las calificaciones más altas y bajas del historial crediticio del prestatario al momento del inicio del préstamo, el segundo par representan la calificación más alta y baja en un periodo de consulta de la calificación FICO durante el periodo de existencia del préstamo con LendingClub. Esta calificación está desarrollada por "Fair Isaac Corporation", y consiste en un algoritmo que define una calificación numérica dentro de un rango de 300 a 850 en función del historial crediticio disponible en burós de crédito. En este estudio se utilizará la información perteneciente a la variable "fico_range_high" considerando que esta variable es la que se evalúa al momento del inicio del préstamo. En la Tabla 5 se muestra la distribución de las calificaciones de riesgo por periodo de la base de datos de LendingClub, para la calificación FICO. En la Tabla 6, se despliega la media y desviación estándar para cada periodo de la variable mencionada.

Tabla 6
Calificaciones FICO

Calificación FICO	Periodo 1	Periodo 2	Periodo 3	Año 2020
Media	717.1	689.5	708.6	713
Desv. Estándar	36.2	30.4	35.3	36.3

Fuente: Elaboración propia

Adicionalmente, la base de datos de LendingClub presenta otras variables representativas del historial crediticio de los prestatarios, como las cuentas abiertas, el crédito revolvente utilizado, cantidad de veces que se ha presentado como moroso en los últimos dos años, cantidad de tarjetas de crédito, cantidad de cuentas en mora, bancarrotas registradas, entre otras. En la Tabla 7, se despliegan los valores medios de variables que se consideran relevantes e interpretables para identificar las características de los prestatarios.

Tabla 7
Valores medios para variables de historial crediticio.

Variable	Periodo 1	Periodo 2	Periodo 3	Año 2020
Cuentas Abiertas "open_acc" (cantidad)	9.3	11.7	11.6	12.3
Default en 2 años "delinq_2yrs" (expresado en años)	0.2	0.3	0.3	0.2
Ratio deuda ingreso "dti"	13.4	18.5	19.9	21.4
Balance Crédito Revolvente "revol_bal" (en dólares)	$14,297.9	$17,015.5	$16,804.4	$18,890.2
Límite de crédito total "total_rev_hi_lim" (en dólares)	NA	$32,764.7	$38,526.7	$45,098.3
Balance de crédito excluyendo hipotecas "total_bal_ex_mort" (en dólares)	NA	$50,299.3	$53,805.2	$61,707.0

Fuente: Elaboración propia.

Existen más variables relacionadas con el historial crediticio, pero se han seleccionado aquellas de las que se puede obtener una explicación intuitiva y apegada a la teoría de riesgo de crédito, tomando en cuenta que los niveles de endeudamiento y la liquidez del prestatario definen su calidad crediticia. (Jiménez y Saurina, 2004). Adicionalmente, es muy posible que la variable FICO represente de manera eficaz todas las demás variables observadas en la base de datos. Si bien el algoritmo de la calificación FICO es secreto, se reporta que la calificación está compuesta por información de cuentas pendientes, el historial de pago, nuevos créditos, el tiempo del historial crediticio y los componentes de productos de crédito de cada prestatario[2]. Por lo tanto, se considera a esta variable como una suerte de resumen para caracterizar la información crediticia que se presenta en la base de datos de LendingClub.

Ahora, con respecto a las variables que representan al préstamo, tenemos el monto del préstamo ("loan_amnt"), tasa de interés ("int_rate"), plazo ("term"), pago mensual ("installment"), motivo del préstamo ("purpose"), estatus del préstamo ("loan_status"), los montos pendientes del préstamo y de los intereses ("total_rec_prncp", "total_rec_int") y recuperación de préstamos vencidos ("recoveries"). En la Tabla 8, se presentan los valores medios para las variables numéricas que representan el préstamo.

[2] https://www.myfico.com/credit-education/whats-in-your-credit-score

Tabla 8
Promedios para variables del préstamo

Variable	Periodo 1	Periodo 2	Periodo 3	Año 2020**
Monto préstamo (en dólares)	$11,098.7	$14,869.9	$15,824.2	$16,238.2
Tasa interés (%)	12.1	13.1	12.9	12.9
Pago mensual (en dólares)	$322.6	$443.0	$462.4	$472.4**
Recuperación préstamo (en dólares)	$9,675.7	$12,820.0	$8,236.3	$1,379.8**
Recuperación intereses (en dólares)	$2,240.0	$3.069.0	$2,182.1	$467.6**
Recuperación préstamos vencidos (en dólares)	$103.3	$261.0	$109.5	$0.4**

** La mayor parte de los préstamos del año 2020 siguen vigentes.
Fuente: Elaboración propia

Los montos de préstamo fluctúan en un rango de $500 a $35,000 para el periodo 1, y de $1,000 a $40.000 para los periodos 2, 3 y 4. Para la tasa de interés, el rango se encuentra entre 5% y 30% para todos los periodos.

La variable que representa el motivo del préstamo se divide en doce categorías que los prestatarios eligen al momento de la solicitud. Estas categorías son: consolidación de deuda, tarjeta de crédito, mejoras del hogar, consumo, atención médica, negocios pequeños, automóvil, mudanza, vivienda, vacaciones, energía renovable, boda, crédito educativo, y otros. Los motivos más comunes son el refinanciamiento de deuda y el pago de tarjetas de crédito. En la Tabla 9, se presentan los porcentajes para las categorías mencionadas. La identificación de los propósitos del préstamo es importante para identificar el estado del préstamo (pago o impago) según el propósito. Se observa que para los propósitos más comunes (consolidación de deuda y tarjeta de crédito) el porcentaje de *defaults* respectivo es de 1% y 6.5% para el periodo 1, 11% y 3.5% para el periodo 2, 5% y 1.7% para el periodo 3, y finalmente, en el año 2020, 0.02% y 0.01%, considerando que la mayor parte de los préstamos colocados en este año siguen vigentes.

Tabla 9
Porcentajes para las categorías de propósito del préstamo

Propósito	Periodo 1	Periodo 2	Periodo 3	Año 2020
Consolidación deuda	46.5%	58.9%	54%	53.4%
Tarjeta crédito	12.9%	22.9%	24.6%	26.9%
Consumo	5.4%	2%	2.2%	2.1%
Mejora hogar	7.5%	6.2%	6.8%	7.2%
Atención médica	1.8%	1%	1.3%	1.2%
Negocios	4.7%	1%	0.9%	0.8%
Automóvil	3.8%	0.9%	1%	0.8%

Mudanza	1.5%	0.6%	0.7%	0.6%
Vivienda	1%	0.4%	0.9%	0.6%
Boda	2.4%	0.1%	0.01%	-
Crédito educativo	1%	0.01%	0%	-
Energía renovable	0.2%	0.1%	0.1%	0.001%
Vacaciones	0.9%	0.6%	0.8%	0.6%
Otro	10.4%	5.2%	6.7%	5.7%

Fuente: Elaboración propia.

IV. ANÁLISIS DE RESULTADOS – CLÚSTERES DE K-MEDIAS

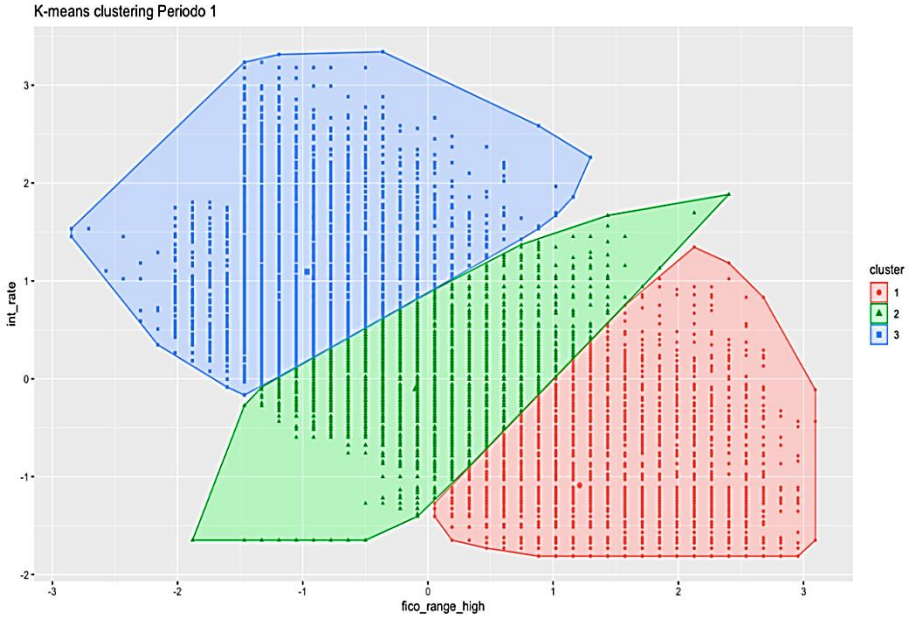
Bajo el análisis realizado, se puede recabar que las variables más importantes dentro de la base de datos de LendingClub, en una forma directa, son la tasa de interés y la calificación FICO. Con base en estas dos variables, es posible analizar a la población desde una perspectiva segmentada, mediante un análisis de clústeres. Se ha decidido utilizar estas variables porque, primeramente, la tasa de interés es una representación del riesgo inherente al crédito, en cuanto a propósito, plazo y monto. Segundo, la calificación FICO es una representación global de la calidad crediticia del prestatario, puesto que está conformada por varios elementos provenientes del historial crediticio. Mediante un análisis de clústeres tomando en cuenta estas dos variables, es posible identificar características sobre otras variables relevantes de la base de datos, como el monto del préstamo, el propósito del préstamo, y la ubicación geográfica desde donde fueron originadas. Se realizó la estandarización de los datos por la presencia de heterogeneidad entre las magnitudes de las variables estudiadas.

Aplicando la regla del codo definida por Dangeti, (2017) se identificó que el número óptimo de clústeres es tres para todos los periodos estudiados en este documento. De esta forma se identifica cómo se comportan las observaciones en función de las dos variables seleccionadas. Es posible observar una formación definida de las agrupaciones sin que ninguna se traslape en ningún periodo estudiado. A continuación, se presentan los resultados gráficos de este análisis junto con su interpretación.

1. **Periodo 1**

- De las 42,535 observaciones
- NA: 0
- K= 3

Figura 6
Clúster Periodo 1



Fuente: Elaboración propia.

- Observaciones por clúster

Clúster	1	2	3	Total
Observaciones	12,322	16,415	13,798	42,535

  - 4 NA's pertenecientes al Clúster 1 de ingresos

Tabla 10
Resumen para Clúster Periodo 1

Variable	Media	Desviación estándar	Error estándar	Clúster
FICO	760.88	22.30	0.39	1
	713.52	18.05	0.28	2
	682.12	15.95	0.27	3
Tasa de interés	8.08	1.88	0.03	1
	11.73	1.64	0.03	2
	16.17	2.20	0.04	3
Ingreso (en dólares)	$70,432.87	$65,736.29	1,160.89	1
	$67,394.31	$69,405.52	1,061.77	2
	$70,051.98	$55,404.73	924.47	3

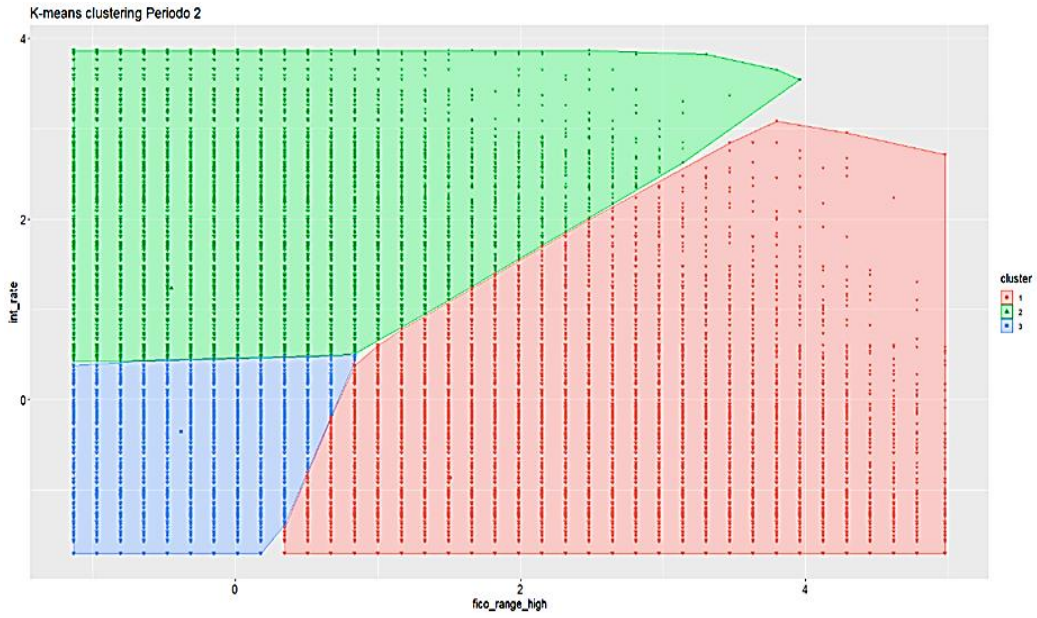
Fuente: Elaboración propia.

Para el primer periodo estudiado (Figura 6), en la Tabla 10 se identifica la relación inversa entre la tasa de interés y la calificación FICO. Para los prestatarios con mayor puntaje FICO, existe una asignación de tasa de interés menor; evidentemente esto representa que existe menor riesgo de crédito y por ende menor retorno para los inversionistas que seleccionen este tipo de prestatarios.

2. <u>Periodo 2</u>

- De las 1,279,312 observaciones
- NA: 0
- K= 3

Figura 7
Clúster Periodo 2



Fuente: Elaboración propia.

- Observaciones por clúster

Clúster	1	2	3	Total
Observaciones	271,767	373,453	634,092	1,279,312

Tabla 11
Resumen para Clúster Periodo 2

Variable	Media	Desviación estándar	Error estándar	Clúster
FICO	744.25	26.44	0.10	1
	684.81	17.87	0.06	2
	686.93	15.44	0.04	3
Tasa de interés	9.19	2.93	0.01	1

	18.80	3.01	0.01	2
	11.51	2.30	0.01	3
Ingreso (en dólares)	$86,591.27	$83,360.17	313.41	1
	$70,736.70	$56,018.84	179.67	2
	$76,054.73	$ 69,195.85	170.32	3

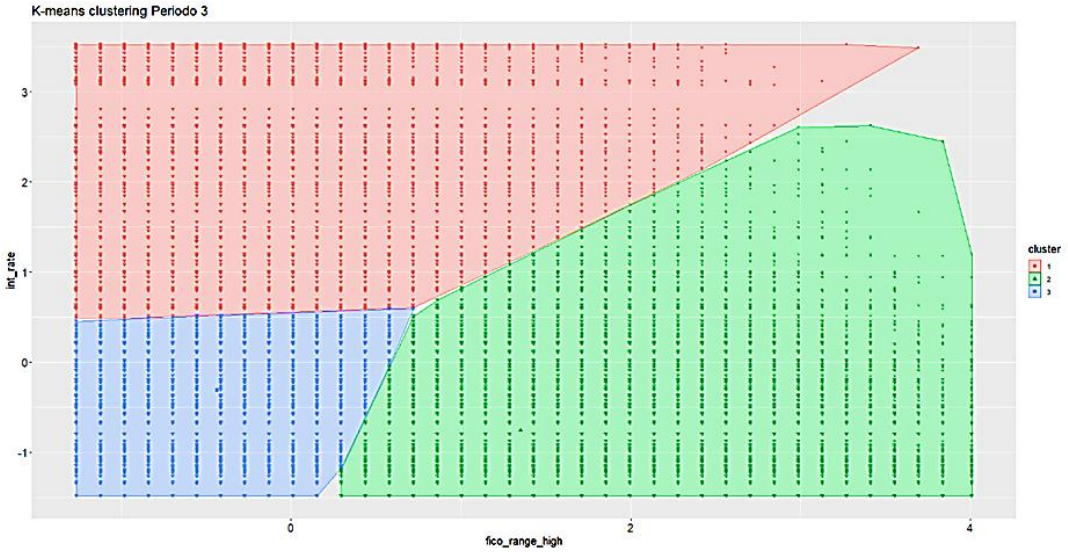
Fuente: Elaboración propia.

En el segundo periodo estudiado (Figura 7), el clúster número dos (color verde) recibe la tasa de interés más alta y evidentemente presenta la calificación FICO más baja, como se presenta en la tabla 11. Gráficamente, es posible identificar que existen prestatarios que cuentan con una calificación FICO baja, que no han recibido una tasa de interés más alta como se observó en los clústeres del primer periodo, como lo demuestra el clúster número tres (color azul).

3. <u>Periodo 3</u>

- De las 1,456,928 observaciones
- NA: 0
- K= 3

Figura 8
Clúster Periodo 3



Fuente: Elaboración propia.

- Observaciones por clúster

Clúster	1	2	3	Total
Observaciones	375,705	381,181	700,042	1,456,928

Tabla 12
Resumen para Clúster Periodo 3

Variable	Media	Desviación estándar	Error estándar	Clúster
FICO	688.89	20.08	0.06	1
	756.27	26.14	0.08	2
	693.17	17.89	0.04	3
Tasa de interés	19.76	3.73	0.01	1
	9.03	2.77	0.01	2
	11.31	2.58	0.01	3
Ingreso (en dólares)	$75,121.81	$90,174.55	288.35	1
	$85,180.88	$91,708.29	291.14	2
	$83,979.83	$177,381.39	415.53	3

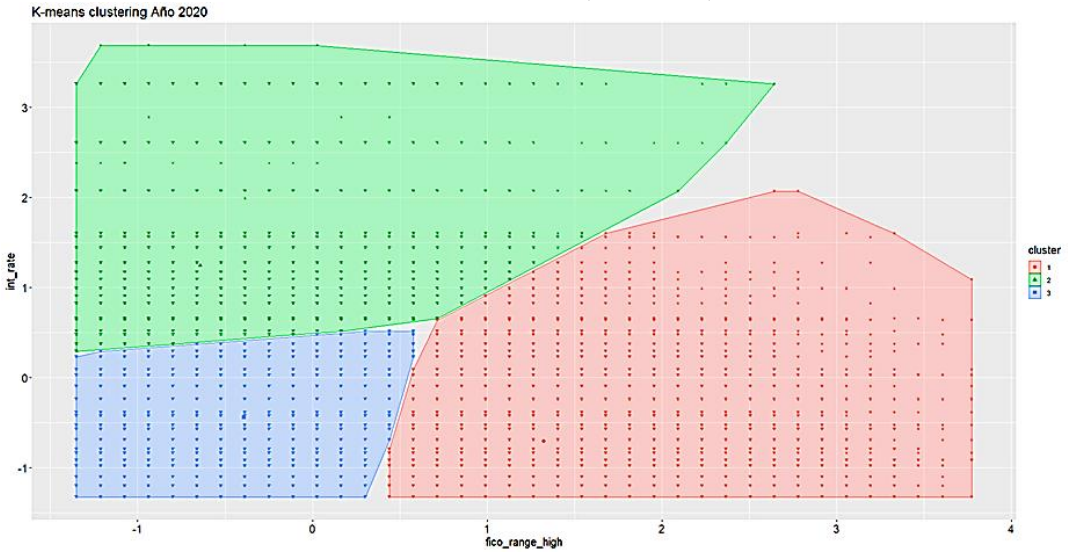
Fuente: Elaboración propia.

Para el periodo 3 (Figura 8), los clústeres con tasa de interés más alta y calificación FICO más baja son 1 (color rojo) y 3 (color azul). Respecto a los anteriores periodos, se observa que la calificación FICO promedio por clúster no cambia radicalmente, de todas formas, en el primer periodo los participantes tenían mejores calificaciones. Las tasas de interés en este periodo, comparada con los dos periodos anteriores, ha incrementado en menos de un punto porcentual en promedio, tal como se presenta en la Tabla 12.

4. <u>Periodo 4 – Año 2020</u>

- De las 1,456,928 observaciones
- NA: 0
- K= 3

Figura 9
Clúster Periodo 4 (Año 2020)



Fuente: Elaboración propia.

- Observaciones por clúster

Clúster	1	2	3	Total
Observaciones	40,083	44,713	61,921	146,717

Tabla 13
Resumen para Clúster Periodo 4 (Año 2020)

Variable	Media	Desviación estándar	Error estándar	Clúster
FICO	761.04	24.55	0.24	1
	689.7	19.88	0.18	2
	698.78	18.97	0.15	3
Tasa de interés	9.45	2.54	0.02	1
	18.96	3.32	0.03	2
	10.73	2.35	0.02	3
Ingreso (en dólares)	$92,502.96	$132,084.97	1,293.09	1
	$84,175.37	$91,599.46	849.05	2
	$93,441.59	$113,104.9	890.88	3

Fuente: Elaboración propia.

Para el año 2020 (Figura 9), se presenta al clúster dos (color verde) como el que recibe mayor tasa de interés, seguido del clúster tres (azul). Para este año se observa que el ingreso promedio ha aumentado respecto de los periodos anteriores y que para el clúster tres, el ingreso promedio es el más

bajo respecto a los demás clústeres. Como se presenta en la Tabla 13, la calificación FICO no muestra cambios importantes ya que se mantiene dentro de un rango similar al de los periodos anteriores.

CONCLUSIONES

En este estudio se ha capturado la evolución de las características de los prestatarios participantes de la dinámica de préstamos persona a persona en la plataforma LendingClub. A partir de los cuatro periodos de estudio seleccionados, se identificaron diferencias entre las variables que se consideran predominantes a la hora de asignar un nivel de riesgo para estos prestatarios. La operación de la dinámica de préstamos se llevó a cabo desde el año 2007 al tercer trimestre del año 2020, por lo tanto, la información captura el contexto socioeconómico de distintos ciclos y el efecto de este contexto sobre las características crediticias de los prestatarios participantes. Mediante el análisis empírico de la base de datos se identificaron variables relevantes respecto a las características de los prestatarios y la naturaleza de los préstamos a los que estaban accediendo. En función de la dinámica del negocio, se definió la implementación del análisis de clústeres de K-medias para ratificar la coherencia entre la asignación de la tasa de interés y la calificación FICO, una variable que conglomera diferentes datos del historial crediticio de los prestatarios.

El conjunto de resultados permite identificar la evolución de los participantes de la dinámica en cuanto a perfiles de riesgo, así como la evolución del manejo de riesgo de crédito de la plataforma, lo que resulta atractivo para nuevos participantes del ecosistema Fintech interesados en el mercado de préstamos digitales. Si bien este estudio se realiza bajo una perspectiva *data-driven* utilizando información de una fintech americana, permite la identificación de variables interesantes dentro del contexto del préstamo en las nuevas formas de acceso al financiamiento, como la tasa de interés asignada por la fintech a cada participante, y la calificación que la empresa Fair Isaac Company (FICO) les asigna como representación de su historial crediticio.

Los resultados presentados brindan un panorama general sobre qué debería incluir un análisis de riesgo de crédito para evaluar prestatarios, y a grandes rasgos, cómo son los perfiles crediticios de los participantes más riesgosos y menos riesgosos. De igual manera, ratifica la utilidad de los modelos de clústeres para llevar a cabo una segmentación de población en función de variables de interés. Como se observa en el desarrollo de este estudio, los resultados de los clústeres muestran de forma consistente cómo los participantes riesgosos se agrupaban por propósito de préstamo e incluso por su ubicación geográfica. El procedimiento y la metodología pueden ser replicados utilizando información de cualquier otro negocio relacionado con el riesgo de crédito, por lo que es replicable para nuevos emprendimientos latinoamericanos que busquen replicar el modelo de negocios que desarrolló LendingClub.

REFERENCIAS

Adhami, S., Gianfrate, G., & Johan, S. (2019). Risks and returns in crowdlending. Available at SSRN 3345874. https://dx.doi.org/10.2139/ssrn.3345874

Agarwal, S., Ambrose, B. W., & Chomsisengphet, S. (2007). Asymmetric information and the automobile loan market. In Agarwal, S., Ambrose, B. W. (eds) Household Credit Usage. (pp. 93–116). Springer. https://doi.org/10.1057/9780230608917_6

Altman, E., Resti, A., & Sironi, A. (2004). Default recovery rates in credit risk modelling: a review of the literature and empirical evidence. *Economic Notes*, 33(2), 183–208. https://doi.org/10.1111/j.0391-5026.2004.00129.x

Arner, D. W., Barberis, J. & Buckley, R. P. (2015). The evolution of Fintech: A new post-crisis paradigm? University of Hong Kong Faculty of Law Research Paper No. 2015/047, *UNSW Law Research Paper No. 2016-62*. http://dx.doi.org/10.2139/ssrn.2676553

Arner, D. W., Barberis, J., & Buckley, R. P. (2016). 150 years of Fintech: An evolutionary analysis. *Jassa-the Finsia Journal of Applied Finance*, 3, 22–29.

Funk, B., Alex, Bachmann, E., Becker, E., Buerckner, D., Hilker, M., Kock, F., Lehmann, M. & Tiburtius, P. (2011). Online peer-to-peer lending-a literature review. *Journal of Internet Banking and Commerce*, 16(2), 1-18.

Baesens, B., van Gestel, T., Stepanova, M., van den Poel, D. & Vanthienen, J. (2005). Neural network survival analysis for personal loan data. *Journal of the Operational Research Society*, 56(9), 1089–1098. https://doi.org/10.1057/palgrave.jors.2601990

Berger, S. C. & Gleisner, F. (2014). Emergence of financial intermediaries in electronic markets: The case of online P2P lending. *BuR Business Research*, 2(1), 39-65. https://doi.org/10.1007/BF03343528

Bikhchandani, S. & Sharma, S. (2000*).* Herd behavior in financial markets. *IMF Staff Papers*, 47(3), 279–310. https://doi.org/10.2307/3867650

Bowley, G. (2011). The new speed of money, reshaping markets. *New York Times*, January 3.

Calabrese, R., Osmetti, S. A. & Zanin, L. (2019). A joint scoring model for peer-to-peer and traditional lending: a bivariate model with copula dependence. *Journal of the Royal Statistical Society*. Series A: Statistics in Society, 182(4), 1163–1188. https://doi.org/10.1111/rssa.12523

Calabrese, R. & Zanin, L. (2022). Modelling spatial dependence for Loss Given Default in peer-to-peer lending. *Expert Systems with Applications*, 192, April. https://doi.org/10.1016/j.eswa.2021.116295

Chengeta, K. & Mabika, E. R. (2021). Peer to Peer Social Lending Default Prediction with Convolutional Neural Networks. IcABCD 2021 - 4th International Conference on Artificial Intelligence, Big Data, Computing and Data Communication Systems, Proceedings. https://doi.org/10.1109/icABCD51485.2021.9519309

Crouhy, M., Galai, D. & Mark, R. (2000). A comparative analysis of current credit risk models. *Journal of Banking & Finance*, 24(1–2), 59–117. http://dx.doi.org/10.1016/S0378-4266(99)00053-9

Cumming, D. J. & Hornuf, L. (2020). Marketplace Lending of SMEs. Center for Economic Studies and Ifo Institute (CESifo), 8100. http://hdl.handle.net/10419/215102

Dangeti, P. (2017). *Statistics for machine learning*. Packt Publishing Ltd.

Dehejia, R., Montgomery, H. & Morduch, J. (2012). Do interest rates matter? Credit demand in the Dhaka slums. *Journal of Development Economics*, 97(2), 437–449. https://doi.org/10.1016/j.jdeveco.2011.05.002

Demirgüç-Kunt, A. & Huizinga, H. (1999). Determinants of commercial bank interest margins and profitability: some international evidence. *The World Bank Economic Review*, 13(2), 379–408. https://doi.org/10.1093/wber/13.2.379

Emekter, R., Tu, Y., Jirasakuldech, B. & Lu, M. (2015). Evaluating credit risk and loan performance in online Peer-to-Peer (P2P) lending. *Applied Economics*, 47(1), 54–70. https://doi.org/10.1080/00036846.2014.962222

Gonzalez, L. & Loureiro, Y. K. (2014). When can a photo increase credit? The impact of lender and borrower profiles on online peer-to-peer loans. *Journal of Behavioral and Experimental Finance*, *2* (C), 44–58. https://doi.org/10.1016/j.jbef.2014.04.002

Hales, M. G. (1995). Focusing on 15% of the pie. *Bank Marketing*, 27(4), 29–33.

Herrera-Arizmendi, P. J. & Amezcua-Núñez, J. B. (2020). El uso de pagos electrónicos, con CoDi en México. *Vincula Tégica Efan.* Año 6, Vol. 2, julio-diciembre, 1111-1119. http://www.web.facpya.uanl.mx/vinculategica/Vinculategica6_2/9_Herrera_Amezcua.pdf

Jagtiani, J. & Lemieux, C. (2019). The roles of alternative data and machine learning in fintech lending: Evidence from the LendingClub consumer platform. *Financial Management*, 48(4), 1009–1029. https://doi.org/10.1111/fima.12295

James, G., Witten, D., Hastie, T. & Tibshirani, R. (2013). *An introduction to statistical learning* (Vol. 112). Springer. https://link.springer.com/book/10.1007/978-1-0716-1418-1

Jiménez, G. & Saurina, J. (2004*).* Collateral, type of lender and relationship banking as determinants of credit risk. *Journal of Banking & Finance*, 28(9), 2191–2212. https://doi.org/10.1016/j.jbankfin.2004.06.010

Kim, J.-Y. & Cho, S.-B. (2019). Predicting repayment of borrows in peer-to-peer social lending with deep dense convolutional network. *Expert Systems*, 36(3). https://doi.org/10.1111/exsy.12403

Koch, R. (2011). *The 80/20 Principle: The Secret of Achieving More with Less: Updated 20th anniversary edition of the productivity and business classic*. Hachette UK.

Komrattanapanya, P. & Suntraruk, P. (2013). Factors influencing dividend payout in Thailand: A tobit regression analysis. *International Journal of Accounting and Financial Reporting*, 3(2), 255. http://dx.doi.org/10.5296/ijafr.v3i2.4443

Kriebel, J. & Stitz, L. (2022). Credit default prediction from user-generated text in peer-to-peer lending using deep learning. *European Journal of Operational Research*. Vol. 302 (1), 309-323. https://doi.org/10.1016/j.ejor.2021.12.024

Lee, E. & Lee, B. (2012). Herding behavior in online P2P lending: An empirical investigation. *Electronic Commerce Research and Applications*, 11(5), 495–503. https://doi.org/10.1016/j.elerap.2012.03.001

Lee, I. & Shin, Y. J. (2018). Fintech: Ecosystem, business models, investment decisions, and challenges. *Business Horizons*, 61(1), 35–46. https://doi.org/10.1016/j.bushor.2017.09.004

Li, H. (2016). Clustering similar lenders in P2P lending. Proceedings of 2016 3rd International Conference on Education, Management and Computing Technology (ICEMCT 2016), 1045–1048. https://doi.org/10.2991/icemct-16.2016.219

Lim, M. K. & Sohn, S. Y. (2007). Cluster-based dynamic scoring model. *Expert Systems with Applications*, 32(2), 427–431. https://doi.org/10.1016/j.eswa.2006.01.034

Maudos, J. & de Guevara, J. F. (2004). Factors explaining the interest margin in the banking sectors of the European Union. *Journal of Banking & Finance*, 28(9), 2259–2281. https://doi.org/10.1016/j.eswa.2006.01.034

Merton, R. C. (1974). On the pricing of corporate debt: The risk structure of interest rates. *The Journal of Finance*, 29(2), 449–470. https://doi.org/10.2307/2978814

Mia, M. A. H., Rahman, M. A. & Uddin, M. (2007). E-Banking: Evolution, Status and Prospect. *The Cost and Management*, 35(1), 36-48. https://ssrn.com/abstract=2371134

Mills, K. & McCarthy, B. (2014). The state of small business lending: Credit access during the recovery and how technology may change the game. Harvard Business School General Management Unit Working Paper, 15–004. https://dx.doi.org/10.2139/ssrn.2470523

Moreno Moreno, A. M., Berenguer, E. & Sanchís Pedregosa, C. (2018). A model proposal to determine a crowd-credit-scoring. *Economics & Sociology*, 11(4), 69-79. https://doi.org/10.36008/eands.11.4.06

Moreno-Moreno, A. M., Sanchis-Pedregosa, C. & Berenguer, E. (2019). Success Factors in Peer-to-Business (P2B) Crowdlending: A Predictive Approach. IEEE Access, 7, 148586–148593. https://doi.org/10.1109/ACCESS.2019.2946858

Morgan, J. P. (1997). Creditmetrics-technical document. JP Morgan, New York. https://dx.doi.org/10.2139/ssrn.3278236

Mundet, H. B. & Gutiérrez, D. M. (2015). La financiación colectiva y su papel en el mundo de la empresa. *Análisis Financiero*, 129,68-78.

Nüesch, R., Alt, R. & Puschmann, T. (2015). Hybrid Customer Interaction. *Business & Information Systems Engineering*. Vol. 57, No. 1. 73-78). https://doi.org/10.1007/s12599-014-0366-9

Peppard, J. (2000). Customer relationship management (CRM) in financial services. *European Management Journal*, 18(3), 312–327. https://doi.org/10.1016/S0263-2373(00)00013-X

Pujun, B., Nick, C. & Max, L. (2016). *Demystifying the workings of Lending Club*. CS229 Stanford. http://cs229.stanford.edu/proj2016spr/report/039.pdf

Sahin, Y. & Duman, E. (2010). Detecting credit card fraud by decision trees and support vector machines. World Congress on Engineering 2012. July 4-6, 2012. London, UK., 2188, 442–447.

Serrano-Cinca, C. & Gutiérrez-Nieto, B. (2016). The use of profit scoring as an alternative to credit scoring systems in peer-to-peer (P2P) lending. *Decision Support Systems*, 89, 113–122. https://doi.org/10.1016/j.dss.2016.06.014

Serrano-Cinca, C., Gutiérrez-Nieto, B., & López-palacios, L. (2015). Determinants of Default in P2P Lending. *PLOS One*, 1–22. https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0139427

Stiglitz, J. E. & Weiss, A. (1981). Credit rationing in markets with imperfect information. *The American Economic Review*, 71(3), 393–410. https://www.jstor.org/stable/1802787

Syakur, M. A., Khotimah, B. K., Rochman, E. M. S. & Satoto, B. D. (2018). Integration k-means clustering method and elbow method for identification of the best customer profile cluster. IOP Conference Series: Materials Science and Engineering, 336(1), 012017. http://dx.doi.org/10.1088/1757-899X/336/1/012017

Wang, Y. & Hua, R. (2014). Guiding the healthy development of the P2P industry and promoting SME financing. *2014 International Conference on Management of E-Commerce and e-Government*, 318–322. https://doi.org/10.1109/ICMeCG.2014.53

Wardrop, R., Zhang, B., Rau, R. & Gray, M. (2015). Moving mainstream. *The European Alternative Finance Benchmarking Report*, 1, 43. https://ideas.repec.org/a/ris/jofipe/0087.html

Yum, H., Lee, B. & Chae, M. (2012). From the wisdom of crowds to my own judgment in microfinance through online peer-to-peer lending platforms. *Electronic Commerce Research and Applications*, 11(5), 469–483. https://doi.org/10.1016/j.elerap.2012.03.001

Zhang, D., Zhou, X., Leung, S. C. H. & Zheng, J. (2010). Vertical bagging decision trees model for credit scoring. *Expert Systems with Applications*, 37(12), 7838–7843. https://doi.org/10.1016/j.eswa.2010.04.054

Zhang, J. & Liu, P. (2012). Rational herding in microloan markets. *Management Science*, 58(5), 892–912. https://www.jstor.org/stable/41499528

Ziegler, T., Shneor, R., Wenzlaff, K., Suresh, K., Ferri, F., Paes, C., Mammadova, L., Wanga, C., Kekre, N., Mutinda, S., Wang, B. W., Closs, C. L., Zhang, B., Forbes, H., Soki, E., Alam, N. & Knaup, C. (2021). Global Alternative Finance Market Benchmarking *The 2nd Global Alternative Finance Market Benchmarking Report*. https://ssrn.com/abstract=3878065