Un modelo de aprendizaje automático orientado a predecir mora crediticia sobre la base de datos públicos, abiertos y masivos: desarrollo, evaluación e implicancias prácticas para el mercado crediticio argentino

Lucas María Soules

Segundo Premio / Categoría Jóvenes Profesionales

12° Premio de Investigación Económica

"Dr. Raúl Prebisch" 2020

BANCO CENTRAL
DE LA REPÚBLICA ARGENTINA

Un modelo de aprendizaje automático orientado a predecir mora crediticia sobre la base de datos públicos, abiertos y masivos: desarrollo, evaluación e implicancias prácticas para el mercado crediticio argentino

Concurso Anual de Investigación "Dr. Raúl Prebisch"
Año 2020

Resumen

En Argentina, como en muchos otros países, el mercado de créditos está compuesto por pocas entidades de gran tamaño — que representan casi la totalidad del dinero prestado — y por muchas entidades pequeñas que buscan atender mercados no satisfechos por las anteriores. El rol que cumplen las entidades pequeñas es fundamental a la hora de democratizar el acceso al crédito, representando una importante fuente de financiamiento para individuos que no cumplen con los requisitos que las entidades grandes, como bancos comerciales, imponen. Sin embargo, son justamente las entidades pequeñas las que disponen de menos recursos a la hora de evaluar sus clientes o potenciales clientes utilizando técnicas analíticas. Los motivos detrás de esto son principalmente dos. En primer lugar, por ser más pequeñas, disponen de un menor volumen de datos y por ende no cuentan con, o se encuentran en desventaja en lo que se refiera a, la materia prima a analizar. En segundo lugar, muchas de estas entidades no cuentan con departamentos de riesgo capacitados en lo que se refiere a desarrollar herramientas avanzadas de análisis de datos orientadas a la toma de decisiones. De este modo, es común que entidades pequeñas deban recurrir a servicios pagos, costosos e, incluso, insuficientes para administrar sus carteras de créditos. En este trabajo se estudia cómo combinando el uso de datos públicos abiertos como los brindados por la Central de Deudores del Banco Central de la República Argentina (los cuales son accesibles a cualquier tipo de entidad) con algoritmos de aprendizaje automático (*machine learning*) avanzados se pueden desarrollar modelos que permiten predecir morosidad futura — tanto en entidades grandes como pequeñas — de manera competitiva cuando se los compara con la literatura previa (la cual comúnmente utiliza datos propietarios). Un sistema como el propuesto permitiría a ambos tipos de entidades, aunque fundamentalmente a las pequeñas, aumentar sus ingresos, reducir sus costos operativos y proyectar mejor sus flujos de fondos. Como consecuencia de lo anterior, estas empresas tendrían una herramienta eficiente que les permitiría ofrecer préstamos con tasas más competitivas, generando un aumento en las posibilidades de acceso al crédito para muchos argentinos.

1. Introducción

El mercado de créditos se caracteriza por el inherente riesgo asociado a la propia actividad (de no repago, sectoriales, shocks agregados, entre otros) (Bank of International Settlements, 2000). Este riesgo no sólo impacta de manera directa en el rendimiento de las entidades prestamistas, sino que también, al dar lugar a racionamiento crediticio (Stiglitz & Weiss, 1981; Walsh, 2003), juega en contra de promover la inclusión financiera (The World Bank, 2013). Dentro de los factores de riesgo asociados a esta industria, el de incumplimiento (*default*) o mora se destaca por ser considerado uno de los más relevantes a los fines operativos y económicos. A modo ilustrativo, en Ali y Iraj (2006) se reporta que la reducción de dicho riesgo se destaca entre las actividades orientadas a reducir el riesgo de la industria y que, ya desde 2006, las entidades prestamistas ponderaban el uso de modelos predictivos que predijesen la probabilidad de repago de créditos.

La importancia que se le da a los modelos predictivos no es exclusiva de la industria financiera. En los último años se ha vivido una explosión en lo que se refiere al uso o potenciales usos de modelos predictivos en dominios tan disímiles como la salud (Dash y col., 2019), las políticas públicas (Kleinberg y col., 2015), la gestión de relaciones con clientes (Anshari y col., 2019), el planeamiento urbano (Rathore y col., 2016), entre otros. Aun así, se puede ver a la industria financiera como pionera en lo que se refiere a utilizar modelos predictivos a los fines de tomar decisiones basadas en datos (Provost & Fawcett, 2013).[1] Sin embargo, resulta oportuno mencionar que el desarrollo y uso de estos modelos no es algo que sea asequible a toda entidad financiera, pues, entre otras cosas, requiere personal capacitado, infraestructura idónea, e, igual o más importante, datos de calidad que permitan efectivamente detectar patrones útiles para la toma de decisiones (véase Iqbal y col., 2018).

En Argentina, al igual que en muchos países, el mercado de créditos se caracteriza por estar conformado por unas pocas entidades grandes que otorgan la mayor cantidad de créditos del sistema y por muchas entidades pequeñas que, en términos relativos, otorgan pocos créditos cada una.[2] Esto da lugar a que exista una fuerte asimetría en cuanto a la capacidad que tienen las distintas entidades pequeñas de hacer uso de herramientas analíticas avanzadas en su operatoria. Los motivos detrás de estos hechos son principalmente dos. En primer lugar, por ser más pequeñas, disponen de un menor volumen de datos y por ende no cuentan con, o se encuentran en desventaja en lo que se refiera a, la materia prima a analizar. En segundo lugar, muchas de estas entidades pequeñas (la mayoría cooperativas, mutuales y sociedades anónimas pequeñas) no cuentan con departamentos de riesgo sofisticados, y por ende tampoco cuentan con la capacidad de desarrollar herramientas avanzadas de análisis de datos orientadas a la toma de decisiones. De este modo, es común que entidades pequeñas deban recurrir a servicios pagos, costosos e,

[1] A modo de ejemplo, FICO, el primer modelo de *credit scoring* comercial, fue desarrollado en 1958 (véase Demyanyk y col., 2008).

[2] En la Sección 2.2 se detallará sobre esta característica sectorial.

incluso, insuficientes para cubrir esta necesidad.[3]

En este trabajo se lleva adelante un ejercicio en donde se combina el uso de técnicas modernas de aprendizaje automático (*machine learning*) con información pública provista por el Banco Central de la República Argentina (BCRA) a los fines de desarrollar modelos predictivos competitivos que permitan, tanto a entidades grandes como a entidades pequeñas, reducir el riesgo asociado al manejo de carteras de crédito. Puntualmente, sobre la base de información obtenida a partir de la Central de Deudores del Sistema Financiero del BCRA,[4] la cual detalla la evolución de más de aproximadamente 45 millones de créditos otorgados por entidades crediticias argentinas,[5] se hace uso de técnicas modernas de aprendizaje supervisado (Hastie, Tibshirani & Friedman, 2001) para desarrollar y evaluar modelos que predigan la probabilidad de que una persona física, que al día de la fecha tiene todas sus deudas en situación "normal" (situación 1 de acuerdo a la nomenclatura utilizada por el BCRA), pase a ser moroso en al menos una de sus deudas.[6] A estos modelos se los conoce como *behavioral credit scoring models* y su importancia en el manejo de carteras crediticias es ampliamente reconocida tanto por la literatura como por la industria crediticia (véase, a modo de ejemplo, Ceylan & Elif, 2018; Hsieh, Lee & Lee, 2010).[7] De manera complementaria, además de desarrollar y evaluar estos modelos, se lleva adelante un ejercicio de interpretación de los mismos, en donde, también utilizando técnicas provenientes del campo del aprendizaje automático, se detectan cuáles son los patrones/características que hacen que el modelo prediga que un prestatario tienen alta o baja probabilidad de convertirse en moroso.

El presente trabajo se circunscribe dentro de una amplia literatura en la que se proponen, desarrollan y evalúan modelos orientados a predecir incumplimiento crediticio. Tal como se mencionó arriba, el uso de este tipo de modelos en esta industria dista de ser un fenómeno novedoso, lo cual queda claramente evidenciado en Hand y Henley (1997). En dicho artículo los autores llevan adelante una revisión de la literatura publicada hasta 1997, encontrando que ya en esa época estos modelos eran considerados importantes por la industria crediticia, que el volumen de publicaciones se encontraba limitado por el hecho de que los datos utilizados en las mismas fueran propietarios, que las variables utilizadas por los mismos solían centrarse en características sociodemográficas (e.g., edad, sexo, nivel educativo, situación laboral, estado civil, entre otras), y que, cuando se utilizaban técnicas estadísticas/analíticas, solían ser en su mayoría técnicas "tradicionales" tales como: análisis discriminante (Durand, 1941; Eisenbeis, 1977), regresión lineal (Orgler, 1970), regresión logística (Wiginton, 1980) o algoritmos de programación lineal

[3] De hecho, resulta relevante mencionar que, aun cuando su dependencia es menor, grandes bancos también recurren a estos servicios.

[4] https://www.bcra.gob.ar/bcrayvos/Situacion_Crediticia.asp

[5] Nótese que este valor hace referencia al universo de créditos otorgados al momento en que los datos fueron obtenidos, noviembre de 2019.

[6] Esto mismo se podría adaptar fácilmente para predecir sobre cada crédito puntual.

[7] Los otros dos grandes problemas asociados a *credit scoring* son el de *application scoring* (predecir si prestatario al que se evalúa otorgar un crédito será o no buen pagador) y el de *collection scoring* (predecir si un prestatario con un crédito ya otorgado cometerá o no fraude). Para mayores detalles véase Ceylan y Elif (2018).

(Kolesar & Showers, 1985). Aun así, resulta sumamente interesante mencionar que ya en este artículo se menciona el uso, quizás más experimental, de técnicas que hoy son asociadas a aplicaciones modernas de aprendizaje automático o inteligencia artificial, tales como: árboles de clasificación (Carter & Catlett, 1987), métodos no paramétricos (en particular k-vecinos más cercanos, véase Chatterjee & Barcun, 1970), redes neuronales (Rosenberg & Gleit, 1994), sistemas expertos (Leonard, 1993) y modelos bayesianos (Srinivasan & Kim, 1987). Esto último no hace más que reforzar el hecho que históricamente la industria financiera ha sido innovadora al momento de incorporar descubrimientos en lo que se refiere a analítica (*analytics*). Tal como se describe en Abdou y Pointon (2011), con el pasar de los años el foco de una rama de la literatura fue centrándose en mayor medida en evaluar modelos de aprendizaje automático de mayor complejidad. Notoriamente, estos estudios rara vez hacen uso de conjuntos de datos de mayor riqueza o volumen que los usados por la literatura que les precedía. A modo de ejemplo, en West (2000) se evalúa el desempeño predictivo de múltiples modelos de redes neuronales utilizando dos conjuntos de datos, uno compuesto únicamente por 1.000 registros y otro por 590 registros. Teniendo esto en cuenta, otra línea de investigación, quizás más novedosa (y en línea con la idea de asociar *Big Data* a variedad en los datos), se centra en incorporar fuentes alternativas de información como insumo de modelos de predicción de incumplimiento crediticio. A modo ilustrativo, en este línea de estudios se encuentran Óskarsdóttir y col. (2019), quienes estudian cómo el análisis de registros de llamadas de telefonía móvil combinado con el análisis de redes puede mejorar el desempeño de estos modelos; Roa y col. (2020), quienes muestran cómo a partir de datos recolectados por *super-apps* (aplicaciones de teléfonos móviles que incorporan múltiples aplicaciones más pequeñas o que ofrecen una amplia gama de productos) se pueden generar modelos de credit scoring competitivos al ser comparados con los resultados que obtienen modelos generados a partir de fuentes tradicionales; y Wei y col. (2016), quienes estudian cómo datos obtenidos a partir de la actividad de usuarios en redes sociales (e.g., Facebook, LinkedIn, Twitter) podrían ser utilizados a los fines de desarrollar este tipo de modelos. Sin embargo, aun cuando esta última línea de investigación es interesante tanto desde el punto académico como del comercial, no puede dejar de mencionarse que estas nuevas fuentes de datos exploradas mantienen la característica de ser propietarias y que de hecho, cuando se las compara con fuentes tradicionalmente utilizadas, quizás sea más complejo que entidades prestamistas puedan acceder y procesar las mismas. De este modo, resulta difícil afirmar que los avances que esta última línea de trabajos ofrece generan un impacto positivo en el agregado de las entidades financieras del sistema crediticio argentino.

Teniendo en cuenta el estado de la literatura, este trabajo cubre un punto que aun no ha sido plenamente abordado por la misma: ¿Puede desarrollarse un modelo de predicción de incumplimiento crediticio competitivo enteramente en base a datos de acceso público? Nótese que, más allá del interés académico que acarrea responder esta pregunta,[8] un sistema de estas

[8]Resulta oportuno mencionar que los resultados obtenidos también aportan al debate en torno a lo que se refiere a herramientas desarrolladas sobre la base de datos abiertos (Janssen, Charalabidis & Zuiderwijk, 2012) y con

características sería de gran utilidad práctica para la industria crediticia. Primero, permitiría anticipar los ingresos que tendría una entidad prestamista el siguiente mes, pudiendo la misma adaptar sus proyecciones de flujos de fondos en base a las predicciones. Segundo, pero no menos importante, permitiría detectar quiénes serían los candidatos más factibles a incumplir en el pago de su deuda y, de esta manera, dirigir e intensificar la comunicación a este grupo.[9] Nótese también que un sistema como el propuesto, sería de particular utilidad para entidades pequeñas, ayudando a nivelar en gran medida la capacidad que las mismas tienen en lo referido a incorporar técnicas analíticas en sus decisiones operativas y facilitándoles de esta manera el brindar créditos más baratos (lo cual podría ampliar el acceso al crédito a más argentinos de una manera más competitiva).

Los resultados obtenidos son sumamente alentadores, pues como se detallará más adelante, los sistemas propuestos en este trabajo obtienen resultados altamente competitivos cuando se los compara con la literatura previa. A su vez, y de manera complementaria, tras realizar un ejercicio de interpretación de los modelos desarrollados se detectan patrones que podrían ser de utilidad no sólo a las entidades prestamistas en su operatoria diaria, sino también a los fines de entender y caracterizar el mercado crediticio argentino en sus conjunto.

Lo que resta de este documento se encuentra estructurado de la siguiente manera. En la Sección 2 se detalla sobre los datos y la metodología utilizada. En la Sección 3 se presentan los principales resultados obtenidos. En la Sección 4 se provee discusión y se concluye.

2. Materiales y métodos

En esta sección se presentan los materiales y métodos utilizados en este trabajo. En primer lugar, se presenta en detalle el conjunto de datos utilizado a lo largo del presente trabajo (Sección 2.1), luego se presenta la metodología adoptada para clasificar a una entidad como pequeña o grande (Sección 2.2), después se provee detalles referidos a cómo el conjunto de datos original fue procesado a los fines de ser utilizados por los modelos propuestos (Sección 2.3). Finalmente, en la Sección 2.4 se presentan y justifican todas las decisiones metodológicas referidas al modelado estadístico adoptadas en este trabajo.

2.1. Datos de la Central de Deudores del BCBA

Tal como se mencionó con anterioridad, a los fines de facilitar el acceso de técnicas de analytics a entidades pequeñas, resulta indispensable que los insumos que dichas técnicas utilizan sean de fácil acceso a las mismas (ya sea por ser no propietarios y/o por ser simples de

potencial de generar impactos sociales positivos.

[9] Esto es importante teniendo en cuenta que el seguimiento al cliente por parte de los departamentos de cobranzas es un factor clave y altamente costoso para esta industria. Un costo asociado a esto es, por ejemplo, el destinado al *call center*. Detectar los candidatos que cesarán su pago implicaría poder enfocar los llamados, *mailings*, mensajes, etc. a estas personas, y no a otras, de forma tal de alocar los recursos de forma eficiente y no desperdiciarlos en aquellos que tienen baja probabilidad de mora.

procesar). De esta manera el conjunto de datos provisto por la Central de Deudores del BCRA se destaca por tener un alto potencial para los fines de este trabajo. Este conjunto de datos contiene información estructurada e histórica en donde, para cada crédito formal otorgado por una entidad crediticia argentina a un individuo o empresa, se detalla, para los últimos 24 meses, información referida al monto adeudado mes a mes, a la situación del mismo mes a mes (más detalles abajo), a la persona (ya sea física o jurídica) que lo solicitó, a la entidad que lo otorgó, entre otras cosas. Algo de suma importancia a los fines de este trabajo, es que los datos pueden ser obtenidos de manera simple y agregada a través de la página web de la Administración Federal de Ingresos Públicos (AFIP)[10] sin más requerimiento que poseer una clave fiscal válida nivel 3 o superior.

En lo que se refiera a la estructura de este conjunto de datos se tiene que, tal cual es provisto, se trata de un gran archivo de texto plano con estructura de tabla en donde cada fila representa un crédito otorgado y en donde cada columna representa una característica de cada crédito (montos, prestamista, prestatario, etc. — más detalles abajo —). Nótese que, de acuerdo a esta estructura, un mismo individuo puede estar asociado a más de un registro en esta tabla (es decir, puede tener más de un crédito otorgado).

En el contexto de este conjunto de datos, cuando se hace referencia a "créditos", se está haciendo referencia a un listado amplio de contratos y operaciones financieras, entre los cuales se encuentran: adelantos, créditos hipotecarios sobre la vivienda, créditos hipotecarios con otras garantías hipotecarias, créditos prendarios sobre automotores, créditos prendarios con otras garantías prendarias, créditos personales, saldos de tarjetas de crédito, entre otros. Sin embargo, en los datos obtenidos no se proveen detalles referidos al tipo de deuda de cada registro.

A los fines de presentar en detalle la estructura de los datos utilizados, en la Tabla 1 se detalla qué variables componen los datos provistos por la Central de Deudores del BCRA, el tipo de cada variable y se provee una pequeña descripción que indica qué refleja cada una.

Teniendo en cuenta el detalle provisto en la Tabla 1 se tiene que para cada registro se dispone de 75 mediciones. En este trabajo se utilizaron datos descargados al mes de diciembre de 2019. De este manera se tiene que para cada crédito se dispone de información mes a mes desde diciembre 2017 hasta noviembre 2019 (inclusive). En total, el conjunto de datos obtenido consta de 45.694.595 deudas/filas y tiene un tamaño en disco de 17,4 gigabytes.[11]

La estructura del conjunto de datos analizado presenta algunos desafíos. Nótese que si un crédito fue otorgado después de diciembre 2017 o fue cancelado antes de noviembre 2019, igualmente se dispondrá de información referida a los meses que no estuvo activo. La manera de identificar estos casos es través del valor que asume la variable situación para dichos meses. Concretamente, se tiene que para los casos de créditos no activos en un mes dado su situación en dicho mes es igual a 0. Nótese que la información referida a los meses en que un crédito no

[10] http://www.afip.gob.ar/sitio/externos/

[11] Este volumen de datos puede ser trabajado, utilizando las herramientas y técnicas idóneas, en equipos de gama media en lo que se refiere a memoria RAM. De este modo, no debería ser un limitante para ninguna entidad crediticia el disponer del equipamiento adecuado para trabajarlos. Concretamente, en este trabajo se utilizó un ordenador portátil con 32 gigabytes de memoria RAM.

Tabla 1: Variables contenidas en los datos provistos por la Central de Deudores BCRA

Variable	Tipo	Descripción
Código de identidad	Numérica	Código identificatorio para cada **entidad prestataria**, tanto bancaria como intermedia
Número de identificación	Caracter	Número único para cada **prestatario**, ya sea persona física o jurídica
Tipo de identificación	Numérica	Determina si la identificación del prestatario es el CUIT, el CUIL, la Clave de Identificación (CDI) o si se trata deudores residentes en el extranjero
Situación*	Numérica	Número del 0 al 6, donde 1 es "situación normal" y 6 es "irrecuperable por disposición técnica". Valores intermedios reflejan, de menor a mayor, distinta demora en los días de mora. Asume el valor 0 si en un mes dado el crédito aun no hubiese sido otorgado o ya hubiera sido cancelado.
Monto*	Numérica	Monto adeudado (en miles de pesos)
Proceso Judicial/Revisión*	Numérica	Valor 0 si no se observa el dato, 1 si se encuentra en proceso judicial, y 2 si se encuentra en revisión

* Estos campos se repiten para los 24 meses anteriores al momento de obtener los datos hasta llegar al campo 75.

estuvo activo deben ser ser tratados con cautela al momento de modelar los datos. En la Sección 2.3 detallaremos cómo se hará uso de esta información a los fines de evitar no caer en errores metodológicos de filtrado de información (*data leakage*).

Adelantando una decisión metodológica tomada, y en línea con el objetivo de crear modelos de predicción de mora para personas físicas, a partir de este punto, en este trabajo sólo se tendrán en cuenta los créditos otorgados a personas físicas (44.737.761 créditos). De esta manera todos los créditos otorgados a personas jurídicas serán dejados de lado. Esto es así dado que el análisis realizado al momento de otorgar créditos a personas físicas es sustancialmente diferente al realizado para personas jurídicas. En estas últimas se suelen incluir el estudio de balances, pronósticos de flujos de fondos (*cash flows*), costo del capital, entre otros (véase Rikkers & Thibeault, 2009). Qué registros corresponden a una persona física o jurídica se identifica sobre la base del campo "tipo de identificación".

A los fines de entender cómo se distribuye la situación de los créditos del sistema, la Tabla 2 presenta la frecuencia absoluta y relativa por situación considerando únicamente los créditos activos a diciembre de 2017.[12] Nótese que esta tabla deja de lado 15.530.491 créditos que estarán activos en al menos uno de los meses comprendidos en el conjunto de datos pero que no lo estuvieron en diciembre 2017.

La Tabla 2 permite ver que aproximadamente el 86 % de los créditos activos a diciembre de 2017 se encontraban en situación 1, mientras que los que están en una situación superior son la minoría. Teniendo en cuenta esto, ya se puede prever que el problema de clasificación al

[12]Con el fin de evitar cualquier posible filtrado de información el análisis exploratorio de los datos se realizó sobre los créditos activos este mes, el más alejado respecto a la fecha de descarga del conjunto de datos.

Tabla 2: Distribución de la situación crediticia para crédito activos en diciembre 2017

Frecuencia	**Situación**					
	1	**2**	**3**	**4**	**5**	**6**
Absoluta	25.066.564	895.347	640.895	899.515	1.703.287	1.662
Relativa	85,82 %	3,07 %	2,19 %	3,08 %	5,83 %	0,01 %

que se enfrentarán los modelos a desarrollar será "desbalanceado", en donde una categoría será ampliamente mayoritaria y la otra estará subrepresentada. Este desbalance se analizará en detalle más adelante y afectará en las decisiones metodológicas adoptadas.

2.2. Clasificación de entidades en pequeñas y grandes

Como se mencionó con anterioridad, el mercado argentino de crédito está conformado por un pequeño conjunto de grandes entidades (en término de volumen prestado) y por un gran conjunto de entidades pequeñas. Tal como se mencionó más arriba, tiene sentido diferenciar entre entidades pequeñas y grandes, ya que son distintas las herramientas que las mismas tienen disponibles para hacer frente al riesgo de sus carteras de créditos. Sin embargo, no existe una nomenclatura establecida para definir a una entidad como "chica" o "grande". En este trabajo optamos por seguir un criterio de división enteramente basados en datos. Concretamente, sobre la base de los prestamos a persona físicas activos en diciembre del 2017 (primer mes de datos disponible), se calculó el volumen prestado por cada entidad, se analizó la distribución de este valor y se eligió un umbral de volumen prestado a partir del cual considerar a una entidad como chica o grande.

Aun cuando a los fines de llevar adelante esta clasificación se deberían considerar todos los créditos activos a personas físicas, dada una inconsistencia en la carga de datos por parte de las entidades prestamistas, en este cálculo se consideró únicamente los créditos en situación 1. Esta inconsistencia ocurre porque no existe una regla establecida en lo referido a cuándo dar de baja una mora por parte de los acreedores.[13] Por lo tanto, tener en cuenta situaciones distintas de 1 para calcular el volumen de las entidades podría generar sesgos en las distribuciones y por ende también en las conclusiones arribadas. Por otro lado, también podría ser un problema no considerar situaciones mayores a 1 dado que las entidades del mercado de créditos no son todas iguales y prestan a distintos segmentos. Es decir, existen entidades que concentran sus préstamos en prestatarios seguros, mientras que otras entidades se especializan en prestar a prestatarios riesgosos. Por lo tanto, estas últimas empresas estarían subrepresentadas. Ponderando estos dos efectos, optamos por no tener en cuenta créditos en situación mayores a 1 en este cálculo.

[13] Hay entidades que en cuanto establecen como incobrable un determinado crédito (o sobrepasa cierta situación crediticia particular, determinada por cada entidad), eliminan este crédito del sistema. En este caso, si una persona sólo tenía esta deuda, ya no figurará como morosa, sin embargo la deuda incobrable existe y ésta debería ser tenida en cuenta en el historial crediticio de la persona. En cambio, hay entidades que no eliminan nunca la información de incobrables.

La Figura 1 muestra cómo se distribuyen las entidades en función del volumen prestado a personas físicas durante diciembre 2017, considerando únicamente préstamos en situación igual a 1 en dicho mes.

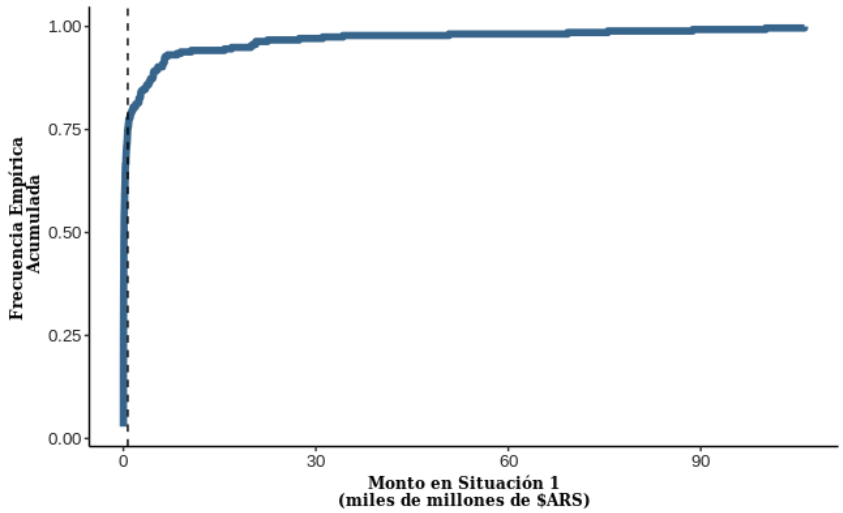

Figura 1: Frecuencia Empírica Acumulada del Volumen Prestado por Entidad. *Nota*: La línea punteada representa el 75 % de la distribución (Q_3=728,6 millones de $ARS).

La Figura 1 es muy ilustrativa, a partir de la misma se observa que aproximadamente el 75 % de las entidades crediticias tenían un cartera de créditos menor o igual a 728.6 millones de $ARS en situación 1 durante diciembre 2017 y que recién a partir de este monto las mismas empiezan a diferenciarse. De esta manera, se tomó la decisión de separar las entidades en dos grupos según el volumen de sus montos prestados. La Figura 2 muestra exactamente esto, clasificando a todas aquellas entidades hasta el tercer cuartil como “chicas” (297 entidades) y al 25 % restante como “grandes” (98 entidades).

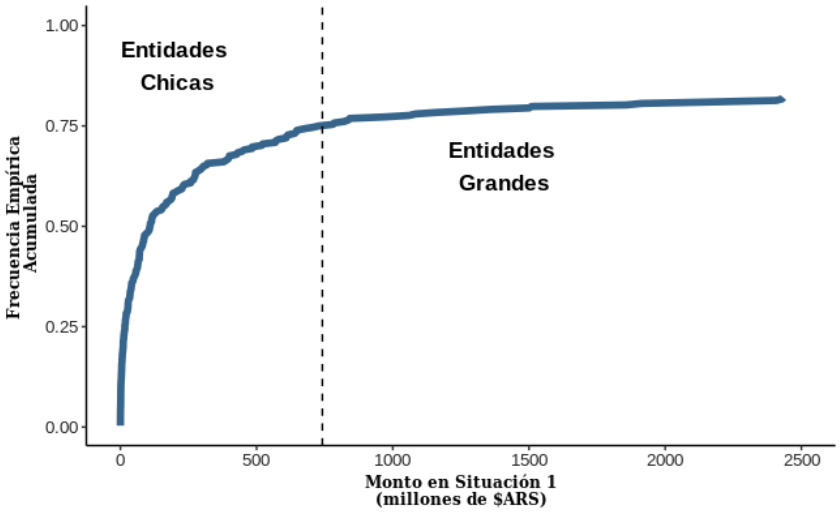

Figura 2: Clasificación de Entidades. *Nota*: La línea punteada representa el tercer cuartil (Q_3=728,6 millones de $ARS).

2.3. Agregación a nivel de individuo e ingeniería de atributos

Tal como se describió en la Sección 2.1, el conjunto de datos obtenido de la Central de Deudores del BCRA detalla información a nivel de créditos otorgados para un total de 24 meses. Sin embargo, como se mencionó más arriba, los modelos que aquí se desarrollan tienen como fin realizar predicciones a nivel de individuo (el cual puede tener o haber tenido más de un crédito activo en el periodo). De esta manera, antes de proceder a ser utilizados por los modelos, se deben agregar a nivel de individuo los datos originales. En esta sección se detalla cómo se hizo esto.

Concretamente, para un mes particular a analizar (e.g., agosto 2019) se genera un conjunto de datos en donde cada registro representa un individuo que durante el mes en cuestión tiene todas sus deudas activas en situación 1 y en donde cada variable representa mediciones referidas a 1) si alguna de sus deudas en el mes actual pasará a tener una situación mayor a 1 en el mes siguiente (septiembre 2019 en caso de analizar agosto 2019) — la variable que se quiere predecir —, 2) características de sus deudas en el mes a analizar (por ejemplo, total de deudas activas en el mes analizado, monto total adeudado en el mes analizado), 3) características referidas a la **evolución** de sus deudas en el mes en cuestión y los tres meses anteriores (agosto 2019, julio 2019, junio 2019 y mayo 2019 en caso de analizar agosto 2019) — las cuales en este trabajo se denominan **variables de tendencia** —, 4) características sociodemográficas propias del individuo (e.g., género, si es o no extranjero, etc., atributos que se pueden inferir a partir de la variable "número de identificación"). En total, teniendo en cuenta estos 4 grupos de variables, se construyeron 839 variables. Es importante mencionar que a los fines de evitar el filtrado de información, al generar estas variables se dejaron de lado todos los créditos aun no otorgados (lo que cuales son simples de identificar analizando el código de situación).

Cada uno de los grupos de variables mencionados conlleva múltiples decisiones de modelado. Al proceso de crear variables predictoras que puedan ser aprovechadas por los modelos predictivos se lo conoce como "ingeniaría de atributos" y es ampliamente reconocido como de vital importancia al momento de desarrollar modelos con buen desempeño predictivo (véase Zheng & Casari, 2018). Por esto motivo a continuación se detallan las decisiones que se tomaron para cada grupo por separado.

Variable a predecir Tal como se mencionó con anterioridad, el objetivo de este trabajo es desarrollar modelos que predigan la probabilidad de que una persona física, que al día de la fecha tiene todas sus deudas en situación "normal" (situación 1 de acuerdo a la nomenclatura utilizada por el BCRA), pase a ser moroso en al menos una de sus deudas el mes siguiente. En vez de desarrollar un único modelo, y en línea con el objetivo de priorizar el uso de este tipo de modelos por parte de las entidades pequeñas, se desarrollarán tres distintos modelos: uno que prediga mora únicamente en entidades pequeñas, otro que prediga mora únicamente en entidades grandes y otro que prediga mora en cualquier entidad. Nótese que a los fines de este trabajo, el más relevante es aquel que predice mora en entidades pequeñas.

De este manera se crearán tres variables a predecir (una por modelo):

1. y_{chicas}: Toma valor igual 1 si para el mes en cuestión la persona tiene todas sus deudas en entidades pequeñas en situación igual a 1 y al mes siguiente una de sus deudas en entidades pequeñas pasa a una situación mayor a 1. Toma valor 0 en caso que ninguna de sus deudas en entidades pequeñas cambie de situación. En caso que la persona no tenga al menos una deuda en una entidad pequeña, asume valor faltante (*missing*) y, de esta manera, la observación es ignorada por los modelos que predicen mora en entidades pequeñas.

2. $y_{grandes}$: Toma valor igual 1 si para el mes en cuestión la persona tiene todas sus deudas en entidades grandes en situación igual a 1 y al mes siguiente una de sus deudas en entidades grandes pasa a una situación mayor a 1. Toma valor 0 en caso que ninguna de sus deudas en entidades grandes cambie de situación. En caso que la persona no tenga al menos una deuda en una entidad grande, asume valor faltante (*missing*) y, de esta manera, la observación es ignorada por los modelos que predicen mora en entidades grandes.

3. y_{total}: Toma valor igual 1 si para el mes en cuestión la persona tiene todas sus deudas en situación igual a 1 y al mes siguiente una de sus deudas pasa a una situación mayor a 1. Toma valor 0 en caso que ninguna de sus deudas cambie de situación.

Es importante mencionar que cada modelo utilizará sólo una de estas variables. A modo de ejemplo, el modelo que se enfoca en predecir deudas en entidades pequeñas sólo utilizará como insumo y_{chicas} y no considerará en lo absoluto y_{grande} e y_{total}. La Tabla 3 muestra la distribución de estas tres variables considerando los datos de diciembre 2017.

Tabla 3: Distribución de la variables a predecirs para diciembre 2017

Variable	0	1
y_{chicas}	95,80 %	4,20 %
$y_{grandes}$	97,08 %	2,92 %
y_{total}	96,80 %	3,20 %

Al analizar los números contenidos en la Tabla 3 se observa claramente que el desbalance ya adelantado en la Tabla 2 es efectivamente pronunciado: son relativamente pocas las personas que, teniendo todas sus deudas en situación igual a 1, pasan a ser morosas en al menos una. Otro patrón sumamente interesante de mencionar es que los datos sugieren que es más común que una persona pase a ser morosa en sus deudas en entidades pequeñas que en sus deudas en entidades grandes. Esto no hace más que reforzar la importancia que tiene para las entidades pequeñas una herramienta como la que en este trabajo se propone.

Variables referidas a características de las deudas en el mes a analizar Se crearon una serie de variables que reflejan el estado de la situación crediticia del individuo en el mes a analizar. Éstas también se subdividen en distintos grupos: variables generales relacionadas a la cantidad de deudas y montos totales, ratios construidos a partir de las anteriores, especificación del endeudamiento en cada una de las entidades y situaciones en cada una de las entidades donde un individuo posee un crédito activo.

1. Variables generales[14]

Tabla 4: Variables generales

Variable	Descripción
cant_deudas_act_grandes	Cantidad de deudas activas (situaciones distintas de 0) en entidades grandes
cant_mayor_1_grandes	Cantidad de deudas en situación mayor a 1 en entidades grandes
cant_igual_1_grandes	Cantidad de deudas en situación igual a 1 en entidades grandes
cant_igual_0_grandes	Cantidad de deudas en situación igual a 0 en entidades grandes, teniendo en cuenta sólo deudas pasadas y no las que todavía no sucedieron
cant_deudas_act_chicas	Cantidad de deudas activas (situaciones distintas de 0) en entidades chicas
cant_mayor_1_chicas	Cantidad de deudas en situación mayor a 1 en entidades chicas
cant_igual_1_chicas	Cantidad de deudas en situación igual a 1 en entidades chicas
cant_igual_0_chicas	Cantidad de deudas en situación igual a 0 en entidades chicas, teniendo en cuenta sólo deudas pasadas y no las que todavía no sucedieron
suma_monto_total_grandes	Suma del monto total adeudado en entidades grandes
suma_monto_igual_1_grandes	Suma del monto total adeudado en situación igual a 1 en entidades grandes
suma_monto_mayor_1_grandes	Suma del monto total adeudado en situación mayor a 1 en entidades grandes
suma_monto_total_chicas	Suma del monto total adeudado en entidades chicas
suma_monto_igual_1_chicas	Suma del monto total adeudado en situación igual a 1 en entidades chicas
suma_monto_mayor_1_chicas	Suma del monto total adeudado en situación mayor a 1 en entidades chicas

[14]Vale la pena aclarar que, dentro de este grupo, también se incluyen algunas variables creadas que tienen información de meses pasados. Puntualmente, todas aquellas variables que consideran préstamos en situación igual a 0 están obteniendo información de los préstamos vencidos a esa fecha. Recuérdese que los préstamos que tienen situación igual a 0, debido a que aún no ocurrieron, no fueron tenidos en cuenta en el armado de ninguna variable.

2. Ratios

Tabla 5: Ratios

Variable	Descripción
ratio_monto_total_cant_mayorigual_1_grandes	Monto total adeudado / Cantidad de deudas en sit. mayor o igual a 1; en entidades grandes
ratio_cant_igual_1_cant_deudas_grandes	Cantidad de deudas en sit. igual a 1 / Cantidad total de deudas; en entidades grandes
ratio_cant_mayor_1_cant_deudas_grandes	Cantidad de deudas en sit. mayor a 1 / Cantidad total de deudas; en entidades grandes
ratio_cant_igual_0_cant_deudas_grandes	Cantidad de deudas en sit. igual a 0 / Cantidad total de deudas; en entidades grandes
ratio_monto_1_cant_igual_1_grandes	Monto adeudado en sit igual a 1 / Cantidad de deudas en sit. igual a 1; en entidades grandes
ratio_cant_mayor_1_cant_mayorigual_1_grandes	Cantidad de deudas en sit. mayor a 1 / Cantidad de deudas en sit. mayor o igual a 1; en entidades grandes
ratio_monto_total_cant_mayorigual_1_chicas	Monto total adeudado / Cantidad de deudas en sit. mayor o igual a 1; en entidades chicas
ratio_cant_igual_1_cant_deudas_chicas	Cantidad de deudas en sit. igual a 1 / Cantidad total de deudas; en entidades chicas
ratio_cant_mayor_1_cant_deudas_chicas	Cantidad de deudas en sit. mayor a 1 / Cantidad total de deudas; en entidades chicas
ratio_cant_igual_0_cant_deudas_chicas	Cantidad de deudas en sit. igual a 0 / Cantidad total de deudas; en entidades chicas
ratio_monto_1_cant_igual_1_chicas	Monto adeudado en sit igual a 1 / Cantidad de deudas en sit. igual a 1; en entidades chicas
ratio_cant_mayor_1_cant_mayorigual_1_chicas	Cantidad de deudas en sit. mayor a 1 / Cantidad de deudas en sit. mayor o igual a 1; en entidades chicas

3. Entidades

Tabla 6: Entidades

Variable	Descripción
entidad_<ent>	Toma valor 1 si el individuo tiene al menos una deuda activa con la entidad "<ent>", 0 en caso contrario

Esta variable se repite para todas las entidades. Por lo tanto, hay 395 variables de este estilo, donde en vez de "<ent>" figura el número identificatorio de la entidad.

4. Situaciones en entidades

Tabla 7: Situaciones

Variable	Descripción
sit_entidad_<ent>	Toma el valor de la situación máxima teniendo en cuenta las deudas activas que tenga el individuo en la entidad "<ent>", de 1 a 6 inclusive. Toma valor 0 si no tiene deudas en dicha entidad.

Variables de tendencias Puesto que, como se verá en la Sección 2.4.1, el modelo de aprendizaje automático supervisado elegido no aprovecha de manera directa la estructura temporal de los datos, se crearon variables de tendencias que tratan de captar la evolución de las deudas de cada individuo. Durante el análisis de los resultados haremos hincapié en analizar el efecto marginal de incorporar estas variables de tendencias, esto lo haremos comparando modelos que las incluyen con modelos que no lo hacen.

Puntualmente se crearon 16 variables de tendencia. De las cuales 8 fueron creadas sobre la variable *suma_monto_total_grandes* y las restantes 8 sobre *suma_monto_total_chicas*. Estas variables se generaron a partir de tomar, para múltiples meses, las diferencias mes a mes en términos absolutos y en términos porcentuales y calculando, en base a los valores obtenidos, los siguientes indicadores:

- El máximo cambio en términos absolutos
- El mínimo cambio en términos absolutos
- El máximo cambio en términos porcentuales
- El mínimo cambio en términos porcentuales
- El promedio de los cambios en términos absolutos
- El promedio de los cambios en términos porcentuales
- El desvío estándar de los montos totales adeudados
- El desvío estándar de los cambios en términos porcentuales

Las diferencias se calculan mes a mes para el mes analizado y los tres anteriores. Se eligió considerar hasta tres meses hacia atrás dado que de considerar una ventana temporal más amplia, se atenuaría la información referida a lo que ocurre en los meses cercanos al analizado.[15]

De esta forma, las variables a incorporar en los modelos son:

[15] Otras variables de tendencia podrían haber sido tenidas en cuenta, como por ejemplo, tendencias de la suma total adeudada en situación mayor a 1, pero por una cuestión de poder de cómputo, se optó por utilizar únicamente la suma total adeudada para cada tipo de entidad.

Tabla 8: Variables de tendencias

Variable	Descripción
max_suma_monto_total_grandes	El máximo cambio absoluto de la variable "suma_monto_total_grandes"
min_suma_monto_total_grandes	El mínimo cambio absoluto de la variable "suma_monto_total_grandes"
mean_suma_monto_total_grandes	El promedio de los cambios absolutos de la variable "suma_monto_total_grandes"
sd_suma_monto_total_grandes	El desvío estándar de la variable "suma_monto_total_grandes"
max_porc_suma_monto_total_grandes	El máximo cambio porcentual de la variable "suma_monto_total_grandes"
min_porc_suma_monto_total_grandes	El mínimo cambio porcentual de la variable "suma_monto_total_grandes"
mean_porc_suma_monto_total_grandes	El promedio de los cambios porcentuales de la variable "suma_monto_total_grandes"
sd_porc_suma_monto_total_grandes	El desvío estándar de los cambios porcentuales de la variable "suma_monto_total_grandes"
max_suma_monto_total_chicas	El máximo cambio absoluto de la variable "suma_monto_total_chicas"
min_suma_monto_total_chicas	El mínimo cambio absoluto de la variable "suma_monto_total_chicas"
mean_suma_monto_total_chicas	El promedio de los cambios absolutos de la variable "suma_monto_total_chicas"
sd_suma_monto_total_chicas	El desvío estándar de la variable "suma_monto_total_chicas"
max_porc_suma_monto_total_chicas	El máximo cambio porcentual de la variable "suma_monto_total_chicas"
min_porc_suma_monto_total_chicas	El mínimo cambio porcentual de la variable "suma_monto_total_chicas"
mean_porc_suma_monto_total_chicas	El promedio de los cambios porcentuales de la variable "suma_monto_total_chicas"
sd_porc_suma_monto_total_chicas	El desvío estándar de los cambios porcentuales de la variable "suma_monto_total_chicas"

Variables referidas a características sociodemográficas Por último, se crearon variables sociodemográficas propias de cada individuo. La Central de Deudores del BCRA no provee de manera directa datos sociodemográficos de los deudores. Sin embargo, a partir de la variable *n_identificacion* (que representa el CUIL de las personas), se pueden determinar algunas características sociodemográficas. Las variables *gender_f* y *gender_m* fueron construidas a partir de los primeros dos números de este indicador (aquellos iniciados en 20 representan hombres; en 27, mujeres; y en 24, tanto hombres como mujeres).[16] Por otro lado, los 8 números siguientes representan el Documento Nacional de Identidad (DNI), único para cada individuo. El número de DNI correlaciona con la edad de las personas y a partir del primer número del mismo se puede determinar si la persona es o no extranjera.[17]

Tabla 9: Variables de tendencias

Variable	Descripción
n_identificacion	CUIL del individuo
extranjero	Toma valor 1 si el individuo es de nacionalidad extranjera y 0 en caso contrario
gender_f	Toma valor 1 si el individuo es mujer y 0 en caso contrario
gender_m	Toma valor 1 si el individuo es hombre y 0 en caso contrario

2.4. Modelado predictivo

En esta sección se presentan todas las decisiones metodológicas referidas al modelado predictivo. Concretamente en la Sección 2.4.1 se presenta el modelo de aprendizaje supervisado elegido, en la Sección 2.4.2 se presenta la métrica de evaluación elegida, en la Sección 2.4.3 se presenta cómo se eligieron los conjuntos de validación y testeo, en la Sección 2.4.4 se presenta la estrategia de búsqueda de hiperparámetros adoptada, y finalmente, en la Sección 2.4.5 se presenta la metodología seguida para llevar adelante el ejercicio de interpretación de modelos.[18]

2.4.1. Modelo de aprendizaje supervisado elegido (XGBoost)

Se eligió *Extreme Gradient Boosting* (XGBoost) (Chen & Guestrin, 2016) como algoritmo de aprendizaje supervisado a utilizar para predecir si un individuo, que al día de la fecha tiene todas sus deudas en situación 1, será moroso en al menos una de ellas el próximo mes. XGBoost es un algoritmo de la familia de los *boosting algorithms*, reconocido por su capacidad de reducir tanto el sesgo como la varianza de las predicciones del modelo (véase Hastie, Tibshirani & Friedman, 2001). A grandes rasgos XGBoost construye de manera secuencial árboles de decisión a partir de

[16] Las variables *gender_f* y *gender_m* para los individuos cuyo CUIL inicia en 24 toman valor faltante (*missing*).

[17] Se optó por incluir directamente el CUIL de los individuos y no del DNI, pues, como se verá, los modelos utilizados tienen la capacidad de detectar esta correlación entre la edad y el CUIL.

[18] Resulta oportuno mencionar que el objetivo de esta sección es motivar las decisiones tomadas y no explicar en detalle cada una de las técnicas expuestas, para el lector que quiera ahondar en este temas se referencia literatura relevante.

los datos de entrenamiento, en donde cada nuevo árbol tiene como objetivo predecir los residuos que resultan del modelo conformado únicamente por los árboles que le preceden. XGBoost es ampliamente reconocido tanto en la literatura académica, la industria y en competencias de aprendizaje automático como un algoritmo con excelente desempeño predictivo y notoriamente bien implementado (lo que facilita experimentar con distintas configuraciones del mismo en poco tiempo). Los típicos hiperparámetros a optimizar en XGBoost se muestran en la Tabla 10.

Tabla 10: Hiperparámetros comúnmente optimizados al utilizar XGBoost

Hiperparámetro	**Descripción**	**Rango**
max_depth	Máxima profundidad de los árboles	(0, ∞]
eta	Proporción que aprende de cada árbol	[0, 1]
gamma	Mínima reducción del costo necesaria en una hoja para generar una nueva partición	[0, ∞]
colsample_bytree	Porcentaje de columnas elegidas (al azar) para construir un árbol	(0, 1]
subsample	Porcentaje de observaciones elegidas (al azar) para construir un árbol	(0, 1]
min_child_weight	Cantidad mínima exigida de observaciones por hoja	[0, ∞]
nrounds	Cantidad de árboles a construir	(0, ∞]

2.4.2. Métrica de evaluación

Para evaluar el desempeño predictivo de los modelos se utilizará el área bajo la curva ROC (AUC, por su siglas en inglés). Para un modelo dado, esta métrica AUC mide el grado en que las predicciones logran "separar las clases". Es decir, qué tanto ocurre que para observaciones que efectivamente pertenecen a la clase positiva (nuevos morosos en el problema atacado en este trabajo) se predicen probabilidades mayores a las predichas para la clase negativa (no morosos en el problema atacado en este trabajo) (para mayores detalles véase Tan, Steinbach & Kumar, 2006).

AUC es una métrica ampliamente utilizada en la literatura de aprendizaje automático y es reconocida, particularmente, por ser robusta al momento de evaluar el desempeño predictivo en problemas con clases desbalanceadas. Esta métrica toma valores en el intervalo [0, 1], donde mayor AUC implica mejor desempeño predictivo (un valor igual a 1 representa una separación perfecta). A su vez, un valor igual a 0,5 indica que el modelo predice acorde a lo que predecería un modelo totalmente azaroso, un valor menor a 0,5 sugiere un desempeño peor que el azar.

2.4.3. Esquema de validación y testeo elegido

A los fines de desarrollar modelos con capacidad de predecir de manera acertada en datos desconocidos, la literatura de aprendizaje automático reconoce como indispensable contar con conjuntos de validación y testeo adecuados (véase Hastie, Tibshirani & Friedman, 2001). El conjunto de validación es un conjunto de datos que no es utilizado directamente por el modelo

para aprender los patrones predictivos, pero que es utilizado para medir el desempeño predictivo del mismo. Este conjunto de datos es de suma utilidad, pues sobre la base de los valores de la métrica de desempeño que se obtienen del mismo, se elige el modelo que se considera mejor a los fines de predecir en datos cuya variable a predecir se desconoce. Finalmente, una vez elegido el modelo considerado como mejor, se utiliza el conjunto de testeo para tener una estimación final de su desempeño en datos desconocidos. Los datos con los que se entrenan los distintos modelos candidatos (es decir, de donde se aprenden de manera directa los patrones predictivos) son conocidos como conjunto de entrenamiento. Nótese que tanto para el conjunto de entrenamiento, el de validación y el de testeo se deben conocer tantos las variables predictoras como las variables a predecir.

Al momento de definir qué datos utilizar como conjunto de entrenamiento, de validación y de testeo en un problema como el abordado en este trabajo se debe tener en cuenta la temporalidad de los datos.[19] En este trabajo se optó por el siguiente esquema: se entrenan todos los modelos utilizando el conjunto de datos correspondiente al mes de agosto del 2019 y se utiliza el conjunto de datos correspondiente a septiembre del 2019, el mes siguiente, como conjunto de validación. Una vez hallado el modelo con mejor desempeño en el conjunto de validación, se evalúa su desempeño predictivo sobre los datos de octubre 2019 (el conjunto de testeo). Nótese que, de esta manera, se tendrán tres valores de AUC para el modelo final elegido: el obtenido sobre el conjunto de entrenamiento, el obtenido sobre el conjunto de validación y el obtenido sobre el conjunto de testeo (siendo este último el más relevante).

En la Tabla 11 se muestra la cantidad de filas que contiene cada uno de estos conjuntos de datos. Como se mencionó anteriormente, cada una de éstas representa una persona.

Tabla 11: Cantidad de individuos (filas) en los conjuntos de datos utilizados para el desarrollo de los modelos

Agosto 2019	**Septiembre 2019**	**Octubre 2019**
13.325.397 individuos	13.262.230 individuos	13.314.295 individuos

2.4.4. Optimización de hiperparámetros

Tal como se mencionó con anterioridad, al utilizar XGBoost se deben definir los valores de múltiples hiperparámetros, los cuales suelen afectar dramáticamente el desempeño predictivo del algoritmo. Dado que la búsqueda exhaustiva de valores óptimos de estos hiperparámetros es sumamente costosa en términos computacionales, se suele optar por estrategias predefinidas de búsqueda. En este trabajo se optó por utilizar *random search* (Bergstra & Bengio, 2012). La misma consiste en definir un rango de posibles valores para cada hiperparámetro y luego seleccionar de manera aleatoria un valor para cada uno. De esta forma quedan seleccionados los hiperparámetros correspondientes para un posible modelo. A su vez el proceso de muestrear los

[19] En caso de no tener esto en cuenta se estaría corriendo el riesgo de llevar adelante filtrado de información (*data leakage*).

valores de cada hiperparámetros es repetido múltiples veces, eligiendo de esta manera múltiples combinaciones de hiperparámetros a validar.[20]

Este proceso en este trabajo se realizó en dos etapas:

1. Se probaron 15 configuraciones de hiperparámetros distintas, tomando un rango muy amplio como posibles valores (véase la Tabla 12) y se seleccionó la mejor de ellas.

Tabla 12: Rango de cada hiperparámetro

Hiperparámetro	**Rango**
max_depth	[5, 14]
eta	[0.2, 0.6]
gamma	[0.2, 20]
colsample_bytree	[0.2, 1]
subsample	[0.4, 1]
min_child_weight	[0, 5]
nrounds	[50, 300]

2. Se corrieron 7 configuraciones aleatorias de hiperparámetros distintas, pero esta vez reduciendo el rango. Puntualmente, se tomaron los valores de la mejor opción encontrada en el punto anterior y se le dio un margen del 15 % en ambos sentidos. Para cada hiperparámetro se muestreo en el intervalo $[0{,}85 \cdot valor_{etapa1}, valor_{etapa1} \cdot 1{,}15]$.

De este modo, para cada modelo, se probaron 22 configuraciones distintas de hiperparámetros, eligiendo como la mejor aquella que mejor predice en los datos de validación.

2.4.5. Intepretación de modelos de aprendizaje supervisado

Modelos de aprendizaje supervisado como los utilizados en este trabajo tienen la ventaja de poseer una gran capacidad predictiva, pero la misma suele darse a costa de sacrificar la interpretabilidad de los mismos.[21] Por este motivo muchas veces estos modelos suelen tratarse como "cajas negras" en lo que se refiere a las variables en que se centran y cómo las utilizan. Sin embargo, cada vez se reconoce en mayor medida la importancia de entender cómo es que estos modelos hacen uso de las variables predictoras al momento de predecir, tarea a la que se la conoce como "interpretación de modelos" (véase Molnar, 2019).

Una técnica recientemente propuesta para interpretar modelos complejos es el método conocido como *SHapley Additive exPlanations* (SHAP) (Lundberg & Lee, 2017). SHAP toma como input un modelo entrenado ($\hat{f}$) y una matriz de datos (X) (que sólo debe contener las

[20] Nótese que cada una de estas combinaciones es utilizada para que el modelo aprenda los patrones predictivos utilizando el conjunto de entrenamiento, y que luego el desempeño predictivo de cada uno de estos modelos se evalúa utilizando el conjunto de validación.

[21] En la literatura esto se conoce como "*accuracy interpretability trade-off*" (véase, a modo de ejemplo, Hastie, Tibshirani & Friedman, 2001; Molnar, 2019)

variables predictoras, siendo comúnmente la utilizada para entrenar $\hat{f}$), y devuelve como output una matriz Φ de las mismas dimensiones que X. Cada elemento $\Phi_{i,j}$ indica cómo la variable j "empuja" la predicción que $\hat{f}$ hace para i respecto al promedio de las predicciones de $\hat{f}$ en X.[22] De este manera se tiene que valores absolutos altos de $\Phi_{i,j}$ sugieren que la variable j impacta en la predicción de i (ya sea aumentándola o disminuyéndola), mientras que valores cercanos a 0 sugieren que la variable j no impactó en dicha predicción.

Teniendo en cuenta lo previamente mencionado, se tiene que una manera de medir la importancia de las distintas variables predictoras es calculando $\sum_i |\Phi_{i,j}|$ para las distintas variables j. Adicionalmente, si para una variable j^* se grafican los valores de Φ_{i,j^*} contra los valores de X_{i,j^*} se obtiene lo que se conoce como gráficos de dependencias SHAP (*SHAP dependence plots*), los cuales permiten visualizar cómo distintos valores de una variable impactan en las predicciones de los modelos. En este trabajo se hará uso de ambas estrategias para interpretar cómo los modelos utilizan las variables predictoras.[23]

3. Resultados

En esta sección se presentan los resultados obtenidos por los modelos propuestos. La sección se divide en dos. En primer lugar, en la Sección 3.1, se reporta el desempeño predictivo obtenido por los distintos modelos desarrollados. Luego, en la Sección 3.2, se presentan los resultados obtenidos tras llevar adelante un ejercicio de interpretación de modelos. Dado que es el que más interesa a los fines de este trabajo, este ejercicio se llevará adelante únicamente para el modelo que predice incumplimiento en entidades pequeñas.

3.1. Resultado de desempeño predictivo

Antes de presentar los resultados obtenidos en el conjunto de testeo, en la Tabla 13 se reportan las mejores configuraciones de hiperparámetros encontradas para los diferentes modelos evaluados. Adicionalmente, también se presenta el desempeño predictivo que los modelos tuvieron tanto en el conjunto de entrenamiento como en el de validación. Tal como se mencionó con anterioridad, a los fines de entender qué rol cumplen las variables de tendencia, los resultados se presentan tanto para modelos que no incluyen variables de tendencia como para modelos que sí lo hacen.

Los resultados presentados en la Tabla 13 indican que, en primer lugar, los modelos logran obtener valores de AUC nada despreciables (recuérdese que valores iguales a 0,5 sugieren predicciones equiparables a las que se obtendrían por puro azar). Adicionalmente, es destacable

[22] En el caso de XGBoost aplicado a problemas de clasificación binaria, las predicciones no son directamente las probabilidades predichas, sino la razón de oportunidades (*odds-ratio*).

[23] Resulta oportuno mencionar que toda conclusión a la que se arribe a partir de este análisis no sugiere ninguna relación de causalidad entre la morosidad y una variable predictora; simplemente son estimaciones de cómo los valores de la variable predictora impactan en las predicciones del modelo.

Tabla 13: Hiperparámetros óptimos y desempeño en entrenamiento y validación

	Modelo					
	Sin tendencias			**Con tendencias**		
	y_{chicas}	$y_{grandes}$	y_{total}	y_{chicas}	$y_{grandes}$	y_{total}
Hiperparámetros						
nrounds	256	277	254	277	330	330
max_depth	11	13	12	13	11	11
eta	0,26	0,26	0,27	0,26	0,27	0,27
gamma	19,24	18,35	16,03	18,35	16,79	16,79
colsample_bytree	0,98	0,92	0,96	0,92	0,86	0,86
min_child_weight	1,2	1,14	1,02	1,14	1,06	1,06
subsample	0,97	0,93	0,82	0,93	0,76	0,76
Performance (AUC)						
Entrenamiento	0,8266	0,7939	0,8018	0,8914	0,8343	0,8398
Validación	0,7702	0,7881	0,7942	0,8243	0,8207	0,8252

la manera positiva en que impacta en los resultados el incluir las variables de tendencia. Nótese que, tanto en entrenamiento como en validación, en todos los modelos se evidencia un aumento en su desempeño al incorporar dichas variables.

Aun cuando los resultados reportados ya sugieren que los modelos desarrollados efectivamente podrían predecir de una manera razonable la probabilidad de que una persona, que al día de la fecha tiene todas sus deudas en situación 1, pase a tener al menos una deuda en situación mayor el siguiente mes, lo cierto es que los valores de AUC reportados en la Tabla 13, al ser los que se obtienen en el conjunto de validación, son excesivamente optimistas (véase Hastie, Tibshirani & Friedman, 2001). Por este motivo en la Tabla 14, además de volver a reportar los resultados obtenidos en el conjunto de validación, se presentan los resultados obtenidos en el conjunto de testeo. Recuérdese que los mismos se obtienen haciendo que el modelo reportado en la Tabla 13 prediga en datos de octubre 2019 (los cuales nunca fueron utilizados para su construcción).

Tabla 14: Desempeño de los modelos en validación y testeo

Modelo	**Performance (AUC)**	
	Validación	**Testeo**
Sin tendencias		
y_{chicas}	0,7702	0,7375
$y_{grandes}$	0,7880	0,7788
y_{total}	0,7942	0,7833
Con tendencias		
y_{chicas}	0,8243	0,7988
$y_{grandes}$	0,8207	0,8116
y_{total}	0,8252	0,8148

Los resultados reportados en Tabla 14 una vez más sugieren que los modelos desarrollados

tienen poder predictivo en lo que se refiere a predecir mora crediticia. En línea con la idea de no subestimar la importancia que tiene la ingeniería de atributos en este tipo de modelos, se evidencia el gran impacto de incluir variables de tendencia en los mismos. Resulta importante mencionar que en todos los casos se observa una caída en el desempeño predictivo al comparar los resultados de validación con los de testeo. Notoriamente, esta caída parece ser mayor en los modelos que predicen morosidad en entidades pequeñas.

Cuando se comparan los resultados obtenidos con la literatura previa, se observa que los mismos son claramente competitivos, incluso superando en muchas ocasiones los reportados por estudios previos. A modo de ejemplo, en Yang, Zhang y col. (2018) se lleva adelante un estudio de predicción de morosidad de tarjetas de crédito, en donde utilizando un conjunto de datos que, a diferencia del aquí utilizado, sí incluye información personal de cada individuo (edad exacta, su estado civil, nivel de educación, etc.), se evalúa el desempeño de múltiples modelos de aprendizaje supervisado (i.e., regresiones logísticas, ensambles XGBoost, redes neuronales, entre otros). En dicho trabajo los autores obtienen valores de AUC menores a los que en este trabajo se reportan. De igual manera, en Hsu y col. (2019) se propone un complejo modelo conformado por redes neuronales recurrentes para predecir mora en pago de tarjetas de crédito, obteniendo un valor de AUC en el orden de 0.78. Finalmente, en Wang, Yu y Ji (2018) se lleva adelante un ejercicio de predicción de mora crediticia utilizando un conjunto de datos de 1.000 observaciones que incorpora atributos referidos tanto a saldos bancarios como a características sociodemográficas; los autores reportan un AUC de 0,83 utilizando un esquema de validación cruzada (concretamente, 10-*fold cross validation*).

3.2. Resultados del ejercicio de interpretación

Por cuestiones de espacio y teniendo en cuenta que el modelo más relevante a los fines de este trabajo es el que predice mora en entidades pequeñas, en esta sección sólo se llevará adelante el ejercicio de interpretación de este modelo. En la Tabla 15 se presentan las variables que, utilizando SHAP, se estiman como más importantes.[24]

La Tabla 15 muestra patrones sumamente interesantes. En primer lugar, es notorio el rol que cumple la variable *ratio_cant_igual_1_cant_deudas_grandes* (tanto en el modelo con tendencias como en el que no las incorpora). Pareciera que el ratio de deudas en situación igual a 1 respecto al total de deudas en entidades grandes es un buen predictor de si un individuo se convertirá en moroso en entidades chicas, se utilicen o no tendencias. Por otro lado, cuando se incluyen tendencias en el análisis, casi todas las variables más relevantes pasan a estar representadas por estas últimas. Más aún, sólo figuran variables de tendencias relacionadas a entidades chicas y ninguna a entidades grandes. Esto, sumado a la presencia de variables como *cant_deudas_act_chicas* y *cant_igual_1_chicas*, explicaría cómo este tipo de entidades atiende un

[24]Para estimar estos valores SHAP tomó como input el modelo con mejor desempeño en el conjunto de datos de validación y los datos de entrenamiento.

Tabla 15: Importancia de atributos de acuerdo a SHAP

y_{chicas} **- sin tendencias**	
Variable	**Importancia**
ratio_cant_igual_1_cant_deudas_grandes	100,0
cant_deudas_act_chicas	69,0
n_identificacion	39,9
sit_entidad_55080	25,6
sit_entidad_70408	24,2
cant_igual_1_chicas	20,7
sit_entidad_70228	18,8
sit_entidad_55053	17,6
ratio_monto_total_cant_mayorigual_1_grandes	15,5
cant_igual_1_grandes	13,2

y_{chicas} **- con tendencias**	
Variable	**Importancia**
ratio_cant_igual_1_cant_deudas_grandes	100,0
max_porc_suma_monto_total_chicas	88,9
cant_deudas_act_chicas	83,1
min_porc_suma_monto_total_chicas	79,4
max_suma_monto_total_chicas	66,4
sd_porc_suma_monto_total_chicas	48,1
mean_porc_suma_monto_total_chicas	46,1
n_identificacion	43,4
cant_igual_1_chicas	40,2
sd_suma_monto_total_chicas	33,7

Nota: los valores están escalados para que la variable más importante tome un valor igual a 100.

mercado con tendencias distintas respecto al gran mercado que representa el de las entidades grandes. También es importante remarcar que existen entidades que son más importantes que otras a la hora de predecir morosidad. Estas entidades son tanto grandes como pequeñas.[25]

A los fines de caracterizar mejor cómo el modelo hace uso de las variables predictoras, en la Figura 3 y en la Figura 4 se presentan gráficos de dependencias SHAP para cada una de las 10 variables identificadas como más relevantes para el modelo sin tendencias y el modelo con tendencias respectivamente.

[25]Las identificadas como importantes resultaron ser: DAP Cooperativa de Crédito y Consumo Ltda (cód. 55080), Tarjeta Naranja S.A. (cód. 70408), Santa Mónica S.A. (cód. 70228) y Casa Luis Chemes S.A. (cód. 55053).

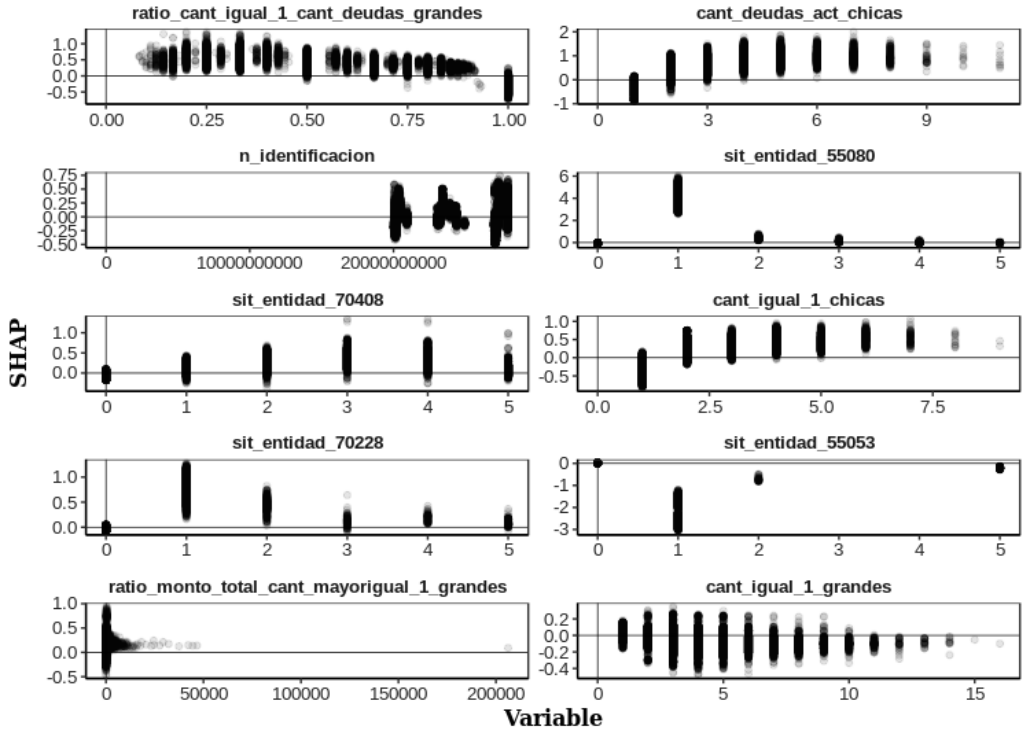

Figura 3: Valores SHAP para y_{chicas} sin tendencias

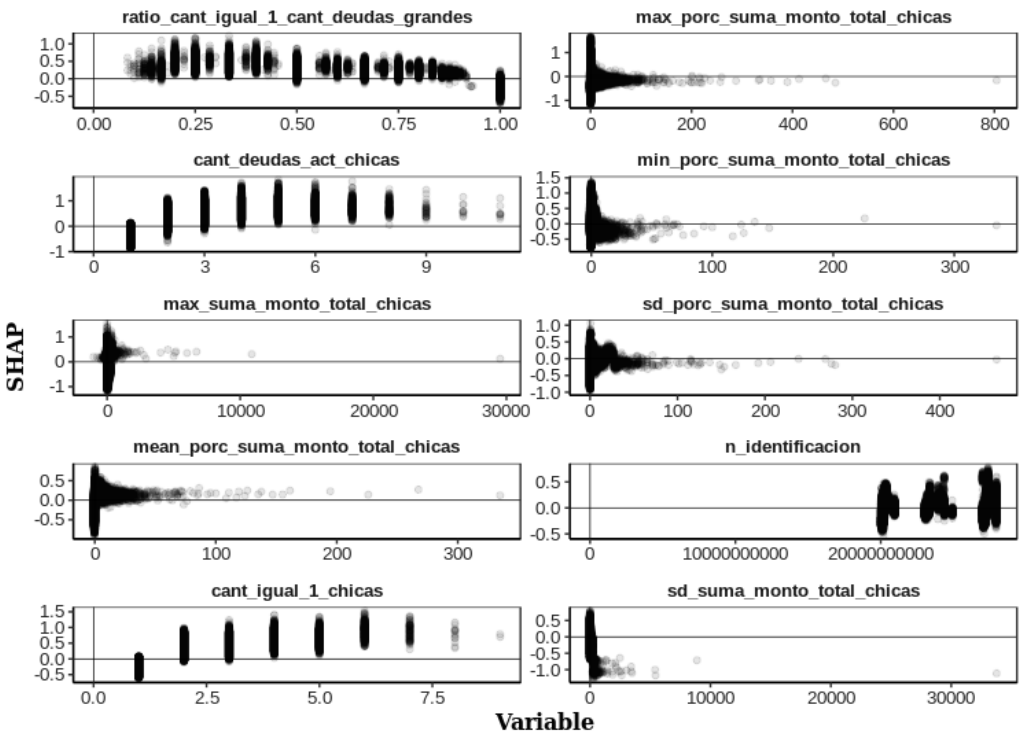

Figura 4: Valores SHAP para y_{chicas} con tendencias

De las figuras anteriores se desprenden algunas conclusiones que permiten entender mejor el comportamiento del mercado crediticio de las entidades pequeñas. Por un lado, como se mostró anteriormente en los resultados de cada modelo, las variables de tendencia importan mucho. Cuando fueron incluidas pasaron a ser parte, en su mayoría, de las variables más importantes. Muchas de ellas se comportaron como se esperaría: valores mayores tienden a hacer que el modelo prediga que es más probable ser moroso. Sin embargo, hay ciertas variables de tendencia en donde esto no se cumple, de hecho se comporta de forma inversa; por ejemplo, para la variable *max_porc_suma_monto_total_chicas* se observa que a mayores valores, menor es la probabilidad de mora que predice el modelo.

Por otro lado, también vale la pena mencionar que las variables estimadas como importantes por SHAP intuitivamente tienen sentido; y de hecho se comportan de la forma esperada. Ejemplos de esto son: *cant_deudas_act_chicas*, *cant_igual_1_chicas*, *cant_igual_1_grandes*, entre otras; donde a mayores valores de las mismas el modelo tiende a predecir un aumento en la probabilidad de mora.[26] Pero, por otro lado, no aparecen (o aparecen poco) otras que sí se esperarían, como por ejemplo los ratios construidos en base a las cantidades y montos totales.

Otra variable interesante de analizar es *n_identificacion*. Ésta fue incluida en el modelo con el fin de captar características sociodemográficas de los individuos. Tal información no está disponible explícitamente en los datos, pero, como se menciono más arriba, tiene una fuerte

[26]Una observación interesante es que, para estas variables, a partir de valores muy elevados, la tendencia cambia de sentido, y mayores valores implican menor probabilidad de ser moroso. Esto tiene sentido financiero, ya que se puede estar tratando de personas con alto poder adquisitivo (raramente morosos) o de préstamos para los cuales las entidades tienen extremados recaudos a la hora de asignar el crédito.

correlación con esta secuencia de números. Cuando se analizan las figuras presentadas, se observa cómo el algoritmo identifica tres grupos: hombres (aquellos que empiezan con 20), mujeres (aquellos que empiezan con 27) y el grupo que incluye ambos géneros de forma indistinta (aquellos que empiezan con 24). En los modelos esta variable pareciera comportarse de manera parecida dentro de cada grupo: primero toma valores altos y luego, de forma irregular, tiende a disminuir. Este patrón sugiere que, por un lado, logró efectivamente detectar los 3 grupos, y, más importante aún, que pareciera haber logrado captar la edad de los prestatarios. Según lo que se observa en estos gráficos, mayor edad contribuye a que se prediga menor probabilidad de volverse moroso.

4. Discusión y conclusiones

En este trabajo se desarrollaron y evaluaron modelos predictivos que, tomando como único insumo información pública contenida en la Central de Deudores del BCRA, predicen, de manera competitiva, si una persona física que hoy en día tiene todas sus deudas en situación regular, pasará a ser morosa en al menos una de ellas. Adicionalmente, se mostró que sobre la base de este conjunto de datos es posible desarrollar modelos que mantengan este desempeño competitivo y que se enfoquen únicamente en predecir mora de créditos otorgados por entidades pequeñas. Finalmente, se llevó adelante un ejercicio de interpretación de modelos el cual permite caracterizar qué tipos de prestatarios se asocian a altas o bajas probabilidades predichas de mora.

En lo que resta se presentarán limitaciones y posibles mejoras (Sección 4.1), luego se cierra el trabajo discutiendo posibles aplicaciones prácticas de los modelos desarrollados (Sección 4.2)

4.1. Limitaciones y futuras posibles mejoras

Aun cuando a lo largo de este trabajo no se ahondó en el tema, un desafío al momento de desarrollar los modelos presentados fue, en parte, el manejo de un gran volumen de datos (como el contenido en la Central de Deudores del BCRA). Poder procesar tanta información y realizar cómputos para la creación de variables no fue una etapa trivial. De hecho, como se mencionó anteriormente, por limitaciones en la capacidad técnica de los equipos utilizados, no se pudo incorporar variables que probablemente fuesen útiles a la hora de predecir (por ejemplo, tendencias calculadas sobre otros atributos). En versiones futuras de este trabajo se podría avanzar en la línea de expandir el conjunto de variables predictoras utilizando equipamiento más potente. En caso de no ser posible, otra alternativa es adaptar el modelo y llevar adelante estrategias de búsqueda de variables importantes de tendencia para predecir.[27] Una última alternativa podría ser estudiar con mayor detenimiento los ratios creados, muchos de los cuales no demostraron ser utilizados por los modelos propuestos, de manera de mantener sólo aquellos que fuesen

[27] Por ejemplo siguiendo metodologías del tipo *forward feature selection* (véase Hastie, Tibshirani & Friedman, 2001).

relevantes.

En este trabajo, por diseño, se planteó como objetivo desarrollar modelos orientados a predecir mora a nivel de persona física y considerando como horizonte temporal un mes a futuro. Nuevas iteraciones podrían enfocarse tanto en desarrollar modelos que consideren distintas poblaciones y/o distintos horizontes temporales. A modo de ejemplo, se podrían incorporar al modelo los créditos otorgados a personas jurídicas, se podrían desarrollar modelos que predigan mora a nivel de crédito y no de individuo, se podría evaluar el desempeño de los mismos al predecir mora en plazos más largos, entre otros. Los resultados a los que se arribó en este trabajo sugieren que modelos alternativos como los mencionados tienen alto potencial de arribar a buenos resultados.

En este trabajo se utilizaron técnicas de modelado predictivo que al día de la fecha son consideradas competitivas. Sin embargo, es posible que existan otras que pudieran tener aun mejor desempeño que el propuesto. Trabajo futuro podría centrarse en explorar caminos alternativos de modelado. Aun así vale la pena repetir que los resultados arribados son competitivos cuando se los compara con la literatura previa y que, teniendo en cuenta que pareciera existir margen para mejorar tanto lo referido a ingeniería de atributos como lo referido a modelado predictivo, los resultados que aquí se presentan deben ser entendidos como una cota inferior.

Finalmente, dentro de la literatura de aprendizaje automático, una tendencia que cobra cada vez mayor relevancia es tratar de evitar que los algoritmos propuestos aprendan los sesgos que tienen incorporadas las sociedades (véase, a modo de ejemplo, Chang, Prabhakaran & Ordonez, 2019). Por ejemplo, un modelo de aprendizaje podría incorporar recomendaciones sexistas o racistas simplemente por haber sido desarrollados utilizando datos que contenían estos sesgos incorporados (véase Zou & Schiebinger, 2018). Una forma para minimizar estos efectos puede ser a través del uso de "variables reservadas". Es decir, variables que se dejan de lado a la hora de construir el modelo, ya que de tener poder predictivo, éste surgiría básicamente debido sesgos sociales (Bellamy y col., 2018). En estos datos, existen variables como "gender_f", "gender_m", "n_identificacion" y "extranjero" que podrían ser candidatas a ser variables de este estilo. Futuras iteraciones de este trabajo podrían estudiar el impacto de omitir este tipo de información. Sin embargo, vale la pena aclararse que la tarea dista de ser simple, pues es posible que terceras variables estén correlacionadas con estas variables sociodemográficas y que de manera indirecta igualmente se capten estos sesgos.

4.2. Implicancias prácticas

El objetivo central de este trabajo fue desarrollar una herramienta que, sobre la base de datos públicos, le permita a una entidad reducir su mora, alocar mejor sus recursos, reducir sus costos y proyectar mejor sus ingresos y flujos de fondos. Teniendo esto en cuenta, se construyeron diversos modelos enfocados tanto para entidades grandes como pequeñas. Sin embargo, el aporte principal está sobre estas últimas. Como se explicó a largo de este trabajo, las entidades pequeñas

serían las mayores beneficiadas si pudieran hacer uso de una herramienta como la propuesta, que les permita aumentar sus beneficios y reducir el riesgo de su cartera de créditos. Teniendo en cuenta el tipo de problema y el nivel de detalle de los datos, todos los modelos resultaron tener un desempeño más que aceptable. Incluso, en caso de implementarse, se podrían sumar datos privados de la entidad en cuestión y así mejorar incluso más el comportamiento del modelo.

Tal como ya se mencionó, sería de esperar que una herramienta como la propuesta no sólo impacte de manera positiva en los beneficios de las entidades crediticias, sino que también, al disminuir el riesgo asociado a otorgar y administrar sus créditos, tenga un impacto positivo en términos de reducir el costo de acceso al crédito y de esta manera fomentar la inclusión financiera. De esta manera, el presente trabajo pone en manifiesto la viabilidad de desarrollar herramientas de analytics que, a contramano de la tendencia actual de utilizar cada vez en mayor medida información propietaria, utilicen enteramente información pública con el objetivo de impactar de manera positiva tanto en los beneficios privados como en los sociales.

Referencias

Abdou, H. A. & Pointon, J. (2011). Credit scoring, statistical techniques and evaluation criteria: a review of the literature. *Intelligent Systems in Accounting, Finance and Management*, *18*(2-3), https://onlinelibrary.wiley.com/doi/pdf/10.1002/isaf.325, 59-88. https://doi.org/10.1002/isaf.325

Ali, F. & Iraj, F. (2006). Credit risk management: a survey of practices. *Managerial Finance*, *32*(3), 227-233. https://doi.org/10.1108/03074350610646735

Anshari, M., Almunawar, M. N., Lim, S. A. & Al-Mudimigh, A. (2019). Customer relationship management and big data enabled: Personalization & customization of services. *Applied Computing and Informatics*, *15*(2), 94-101. https://doi.org/https://doi.org/10.1016/j.aci.2018.05.004

Bank of International Settlements. (2000). *Principles for the Management of Credit Risk* (B. C. on Banking Supervision, Ed.; inf. téc.). https://www.bis.org/publ/bcbs75.htm

Bellamy, R. K. E., Dey, K., Hind, M., Hoffman, S. C., Houde, S., Kannan, K., Lohia, P., Martino, J., Mehta, S., Mojsilovic, A., Nagar, S., Ramamurthy, K. N., Richards, J., Saha, D., Sattigeri, P., Singh, M., Varshney, K. R. & Zhang, Y. (2018). AI Fairness 360: An Extensible Toolkit for Detecting, Understanding, and Mitigating Unwanted Algorithmic Bias.

Bergstra, J. & Bengio, Y. (2012). Random Search for Hyper-Parameter Optimization. *J. Mach. Learn. Res.*, *13*(null), 281-305.

Carter, C. & Catlett, J. (1987). Assessing Credit Card Applications Using Machine Learning. *IEEE Expert*, *2*(3), 71-79.

Ceylan, O. & Elif, Ö. (2018). A review of credit scoring research in the age of Big Data. *Journal of Financial Regulation and Compliance*, *26*(3), 382-405. https://doi.org/10.1108/JFRC-06-2017-0054

Chang, K.-W., Prabhakaran, V. & Ordonez, V. (2019). Bias and Fairness in Natural Language Processing, En *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): Tutorial Abstracts*, Hong Kong, China, Association for Computational Linguistics.

Chatterjee, S. & Barcun, S. (1970). A Nonparametric Approach to Credit Screening [Full publication date: Mar., 1970]. *Journal of the American Statistical Association*, *65*(329), 150-154. https://doi.org/10.2307/2283581

Chen, T. & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System, En *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco, California, USA, Association for Computing Machinery. https://doi.org/10.1145/2939672.2939785

Dash, S., Shakyawar, S. K., Sharma, M. & Kaushik, S. (2019). Big data in healthcare: management, analysis and future prospects. *Journal of Big Data*, *6*(1), 54. https://doi.org/10.1186/s40537-019-0217-0

Demyanyk, Y. Y col. (2008). *Did credit scores predict the subprime crisis?* (F. R. B. of St. Louis, Ed.; inf. téc.). https://www.stlouisfed.org/publications/regional-economist/october-2008/did-credit-scores-predict-the-subprime-crisis

Durand, D. (1941). *Risk elements in consumer installment financing*. National Bureau of Economic Research, New York.

Eisenbeis, R. A. (1977). Pitfalls in the Application of Discriminant Analysis in Business, Finance, and Economics [Full publication date: Jun., 1977]. *The Journal of Finance*, *32*(3), 875-900. https://doi.org/10.2307/2326320

Hand, D. J. & Henley, W. E. (1997). Statistical Classification Methods in Consumer Credit Scoring: A Review [Full publication date: 1997]. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, *160*(3), 523-541. www.jstor.org/stable/2983268

Hastie, T., Tibshirani, R. & Friedman, J. (2001). *The Elements of Statistical Learning*. New York, NY, USA, Springer New York Inc.

Hsieh, H., Lee, T. & Lee, T. (2010). Data Mining in Building Behavioral Scoring Models, En *2010 International Conference on Computational Intelligence and Software Engineering*.

Hsu, T., Liou, S., Wang, Y., Huang, Y. & Che-Lin. (2019). Enhanced Recurrent Neural Network for Combining Static and Dynamic Features for Credit Card Default Prediction, En *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Iqbal, M., Kazmi, S. H. A., Manzoor, A., Soomrani, A. R., Butt, S. H. & Shaikh, K. A. (2018). A study of big data for business growth in SMEs: Opportunities challenges, En *2018*

International Conference on Computing, Mathematics and Engineering Technologies (iCoMET).

Janssen, M., Charalabidis, Y. & Zuiderwijk, A. (2012). Benefits, Adoption Barriers and Myths of Open Data and Open Government. *Information Systems Management*, *29*(4), 258-268. https://doi.org/10.1080/10580530.2012.716740

Kleinberg, J., Ludwig, J., Mullainathan, S. & Obermeyer, Z. (2015). Prediction Policy Problems. *American Economic Review*, *105*(5), 491-95. https://doi.org/10.1257/aer.p20151023

Kolesar, P. & Showers, J. L. (1985). A Robust Credit Screening Model Using Categorical Data [Full publication date: Feb., 1985]. *Management Science*, *31*(2), 123-133. www.jstor.org/stable/2631510

Leonard, K. J. (1993). Detecting credit card fraud using expert systems. *Computers & Industrial Engineering*, *25*(1), 103-106. https://doi.org/https://doi.org/10.1016/0360-8352(93)90231-L

Lundberg, S. M. & Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions (I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan & R. Garnett, Eds.). En I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan & R. Garnett (Eds.), *Advances in Neural Information Processing Systems 30*. Curran Associates, Inc. http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf

Molnar, C. (2019). *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable* [https://christophm.github.io/interpretable-ml-book/].

Orgler, Y. E. (1970). A Credit Scoring Model for Commercial Loans [Full publication date: Nov., 1970]. *Journal of Money, Credit and Banking*, *2*(4), 435-445. https://doi.org/10.2307/1991095

Óskarsdóttir, M., Bravo, C., Sarraute, C., Vanthienen, J. & Baesens, B. (2019). The value of big data for credit scoring: Enhancing financial inclusion using mobile phone data and social network analytics. *Applied Soft Computing*, *74*, 26-39. https://doi.org/https://doi.org/10.1016/j.asoc.2018.10.004

Provost, F. & Fawcett, T. (2013). Data science and its relationship to big data and data-driven decision making. *Big data*, *1*(1), 51-59.

Rathore, M. M., Ahmad, A., Paul, A. & Rho, S. (2016). Urban planning and building smart cities based on the Internet of Things using Big Data analytics [Industrial Technologies and Applications for the Internet of Things]. *Computer Networks*, *101*, 63-80. https://doi.org/https://doi.org/10.1016/j.comnet.2015.12.023

Rikkers, F. & Thibeault, A. E. (2009). A Structural form Default Prediction Model for SMEs, Evidence from the Dutch Market. *Multinational Finance Journal*, *13*(3-4), 229-264. https://EconPapers.repec.org/RePEc:mfj:journl:v:13:y:2009:i:3-4:p:229-264

Roa, L., Correa-Bahnsen, A., Suarez, G., Cortés-Tejada, F., Luque, M. A. & Bravo, C. (2020). Super-App Behavioral Patterns in Credit Risk Models: Financial, Statistical and Regulatory Implications.

Rosenberg, E. & Gleit, A. (1994). Quantitative Methods in Credit Management: A Survey [Full publication date: Jul. - Aug., 1994]. *Operations Research*, *42*(4), 589-613. www.jstor.org/stable/171615

Srinivasan, V. & Kim, Y. H. (1987). Note-The Bierman-Hausman Credit Granting Model: A Note. *Manage. Sci.*, *33*(10), 1361-1362.

Stiglitz, J. E. & Weiss, A. (1981). Credit Rationing in Markets with Imperfect Information. *The American Economic Review*, *71*(3), 393-410. www.jstor.org/stable/1802787

Tan, P.-N., Steinbach, M. & Kumar, V. (2006). *Introduction to Data Mining*. Pearson Education.

The World Bank. (2013). *Global financial development report 2014 : financial inclusion* (T. W. Bank, Ed.; inf. téc.). https://www.bis.org/publ/bcbs75.htm

Walsh, C. E. (2003). *Monetary Theory and Policy, 2nd Edition* (Vol. 1). The MIT Press.

Wang, M., Yu, J. & Ji, Z. (2018). Credit Fraud Risk Detection Based on XGBoost-LR Hybrid Model. *ICEB 2018 Proceedings*, *68*. https://aisel.aisnet.org/iceb2018/68

Wei, Y., Yildirim, P., Van den Bulte, C. & Dellarocas, C. (2016). Credit Scoring with Social Network Data. *Marketing Science*, *35*(2), 234-258. https://doi.org/10.1287/mksc.2015.0949

West, D. (2000). Neural network credit scoring models. *Computers & Operations Research*, *27*(11), 1131-1152. http://www.sciencedirect.com/science/article/pii/S0305054899001495

Wiginton, J. C. (1980). A Note on the Comparison of Logit and Discriminant Models of Consumer Credit Behavior [Full publication date: Sep., 1980]. *The Journal of Financial and Quantitative Analysis*, *15*(3), 757-770. https://doi.org/10.2307/2330408

Yang, S., Zhang, H. Y col. (2018). Comparison of several data mining methods in credit card default prediction. *Intelligent Information Management*, *10*(05), 115. https://doi.org/10.4236/iim.2018.105010

Zheng, A. & Casari, A. (2018). *Feature Engineering for Machine Learning: Principles and Techniques for Data Scientists* (1st). O'Reilly Media, Inc.

Zou, J. & Schiebinger, L. (2018). AI can be sexist and racist—it's time to make it fair. Nature Publishing Group. https://doi.org/10.1038/d41586-018-05707-8