UNIVERSIDAD TORCUATO DI TELLA

UNIVERSIDAD TORCUATO DI TELLA, ESCUELA DE NEGOCIOS

MASTER IN MANAGEMENT + ANALYTICS

Modelos predictivos competitivos de morosidad crediticia para entidades argentinas

Análisis descriptivo y predictivo con datos públicos

Resumen

Una importante característica del mercado de créditos en la Argentina es la marcada diferencia que existe en el acceso a la información entre las entidades grandes (mayoritariamente bancos) y las entidades chicas (sociedades anónimas, mutuales y cooperativas), a lo que se suma una menor capacidad analítica de estas últimas (generalmente por no disponer de equipos internos plenamente desarrollados y abocados a la tarea). Esto lleva a que en su operatoria sea común que entidades pequeñas deban recurrir a costosos servicios externos, lo que no sólo impacta en su rentabilidad, sino también en los clientes que efectivamente pueden atender. El objetivo de esta tesis es desarrollar y evaluar una herramienta que, utilizando algoritmos de aprendizaje automático y datos enteramente públicos, prediga morosidad futura en personas que hasta el momento tienen todas sus deudas en situación regular. Una herramienta de estas características permitiría, principalmente a entidades pequeñas, aumentar sus ingresos, reducir sus costos operativos y proyectar mejor sus flujos de fondos. Los resultados obtenidos sugieren que, tomando como insumo datos de la Central de Deudores del Banco Central de la República Argentina y haciendo uso de metodologías modernas de aprendizaje automático, se pueden desarrollar modelos predictivos de detección de mora, los cuales alcanzan resultados competitivos cuando se los compara con la literatura previa. En este trabajo se detalla las diferencias entre ambos tipos de entidades, se presenta en detalle las decisiones metodológicas detrás de los modelos desarrollados, se analiza el efecto marginal que genera la incorporación de variables de tendencias, se evalúa la performance de los mismos utilizando datos reales, y se lleva adelante un ejercicio de interpretación de modelos; finalmente, se discute cómo estos modelos pueden ser aplicados para generar valor en una entidad crediticia.

Alumno: Soules, Lucas M.

Director: Gálvez, Ramiro H. - Departamento de Computación, FCEyN, UBA

Fecha de entrega: 13 de Julio de 2020

UNIVERSIDAD TORCUATO DI TELLA, ESCUELA DE NEGOCIOS

MASTER IN MANAGEMENT + ANALYTICS

Competitive Predictive Credit Default Models for Argentine Entities

Descriptive and Predictive Analysis with Public Data

Abstract

One of the most important characteristics of the Argentine credit market is the strong difference between large entities (mostly banks) and small entities (limited companies, mutuals and cooperatives) in their capability to obtain information. In addition to this, the smaller ones usually have fewer resources to analyse data, mostly because of their lack of internal analytical skills. The result is such that small entities are forced to incur in costly external services, affecting not only their earnings, but also the type and amount of customers they can serve. The purpose of this thesis project is to develop and test a tool, using machine learning algorithms with public data, in order to predict future credit loans default in people that, at the moment, have met all their debt obligations. This tool would allow both types of entities, but mostly smaller ones, to raise their revenue, reduce operating costs and project more accurately future cash flows. The final results suggest it is possible to create competitive and marketable default predictive models using modern machine learning techniques and public data from the Central Bank of Argentina. In this thesis, differences between both types of entities are studied. Moreover, the methodological decisions and the performance behind the created models are exhibited. Also, the marginal effects of using tendency variables in the models are calculated. Finally, a discussion on model interpretation and on how this tool can create value to a company are included.

Student: Soules, Lucas M.

Thesis advisor: Gálvez, Ramiro H. - Departamento de Computación, FCEyN, UBA

Date: July 13th, 2020

1 - Introducción

Reducir la mora es uno de los ejes centrales en una empresa dedicada al mercado crediticio. En Argentina el mercado de créditos se encuentra todavía muy polarizado, habiendo, en términos relativos, pocas entidades grandes y muchas entidades chicas. Esto conlleva a que también exista una asimetría en la información y herramientas que cada una dispone en el diagnóstico de sus clientes. Es decir, empresas chicas encuentran difícil el acceso a información valiosa para poder generar modelos predictivos o de clasificación de deudores. La mayoría de ellas intentan resolver esta cuestión utilizando servicios pagos como Nosis[1] o Veraz[2]; ambos costosos e insuficientes. Incluso, entidades grandes, como bancos nacionales, recurren a estos servicios, ya que son los únicos en el mercado. A lo anterior se le suma el hecho de que muchas de estas entidades chicas no tienen departamentos de riesgo muy desarrollados, y por ende tampoco herramientas avanzadas de análisis de datos enfocados en la toma de decisiones.

El objetivo de este trabajo es poder aprovechar información pública para poder entender cómo se compone el mercado de créditos, y generar modelos de machine learning que predigan la probabilidad de que un individuo, que tiene todas sus deudas en situación normal, pase a ser moroso. Esto tiene gran utilidad en la industria, principalmente por dos motivos. Por un lado permite anticipar qué ingresos tendrá la empresa el siguiente mes, y así adaptar su flujo de fondos. Segundo, pero no menor, permite detectar quiénes serán los candidatos más factibles a defaultear su deuda y entonces dirigir e intensificar la comunicación a este grupo, tratando de minimizar dicho número. 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. Dada la nueva reglamentación del Banco Central de la República Argentina (BCRA),[3] esto tomó incluso

[1] https://www.nosis.com/es
[2] https://www.veraz.com.ar/ECOMMERCE/inicio.ecom
[3] http://www.bcra.gov.ar/

mayor relevancia en los últimos meses. Con la Comunicación "A"6.909 se les quitó a las entidades intermedias (cooperativas, mutuales y sociedades anónimas) la posibilidad de cobrar sus cuotas a través de un débito automático de la cuenta bancaria del cliente (Wende, 2020).[4] Por consecuencia, estas entidades deben hacer un seguimiento más intensivo de los deudores para poder garantizarse el cobro.

La utilización de machine learning en la industria financiera, y puntualmente, en el mundo crediticio no es nueva. Hay literatura que muestra cómo se han desarrollado modelos de ensamble que combinan distintas técnicas como XGBoost, deep learning y regresiones logísticas en mercados fuera de la Argentina (Chen, Ding, Li, Yang, 2018). También, en el exterior, se han desarrollado modelos que además de incluir información privada, utilizan datos de las redes sociales (Yu, 2017). Más aún, se han hecho estudios donde implementan algoritmos predictivos que utilizan datos obtenidos de aplicaciones móviles denominadas *super-apps*, como las de servicio de delivery de comida, que permiten extraer información más precisa y personal del usuario, como su ubicación, monto gastado, cantidad de compras, etc. (Bravo et.al, 2020).

Sin embargo, este trabajo busca mostrar cómo, a partir de datos **enteramente públicos**, se pueden generar modelos predictivos competitivos para el mercado crediticio Argentino. Concretamente, los modelos serán entrenados para predecir si una persona física dada, que tiene deudas y todas ellas se encuentran en situación normal, pasará o no a ser un deudor moroso en el siguiente mes. Además, se estudiará cómo impacta en dichos modelos la incorporación de variables de tendencias. Es decir, se tratará de mostrar y cuantificar cómo mejora la performance predictiva cuando se tiene en consideración información de los meses pasados más cercanos. En definitiva, estos modelos son una herramienta útil para todo tipo de entidad; pero considerablemente especial para las chicas, las cuales tienen muy limitado el acceso a mayor información.

El resto de este documento se encuentra estructurado de la siguiente manera: en la Sección 2 se presentarán los materiales y métodos, en la Sección 3 se hará un análisis

[4] El término 'entidades intermedias' está haciendo referencia a entidades no bancarias. No está directamente relacionado con su tamaño; aunque, en este trabajo, la mayoría entran en la categoría de entidades chicas.

descriptivo a nivel general del mercado de créditos y también a nivel individual, en la Sección 4 se mostrarán los resultados de los modelos con sus *performance* e interpretación, por último, en la Sección 5 se presentarán las conclusiones del trabajo y las posibles futuras mejoras.

2 - Materiales y métodos

2.1 - Datos

2.1.1 - Estructura de los datos originales

Este trabajo fue realizado únicamente con datos públicos de la Central de Deudores del BCRA,[5] disponibles en la página web de la Administración Federal de Ingresos Públicos (AFIP).[6]

Esta base de datos posee información histórica detallada sobre todos los créditos formales accedidos por cada individuo o empresa de la Argentina. Puntualmente, para cada crédito, detalla información, por 24 meses, sobre la entidad prestadora, situación, monto y condición judicial. Cada fila representa un crédito a una determinada persona. Esto significa que pueden haber más de una fila pertenecientes a la misma persona.

Cuando se habla de 'créditos', se está haciendo referencia a un listado amplio de contratos y operaciones financieras. Entre ellos se encuentran:

1. Adelantos
2. Hipotecarios sobre la vivienda
3. Con otras garantías hipotecarias
4. Prendarios sobre automotores
5. Con otras garantías prendarias
6. Personales
7. Tarjetas de Crédito
8. Otros

[5] https://www.bcra.gob.ar/bcrayvos/Situacion_Crediticia.asp
[6] http://www.afip.gob.ar/sitio/externos/

Sin embargo, en los datos no se detalla el tipo de deuda en cuestión.

A continuación se muestra puntualmente qué variables se incluyen en dicha base:

Tabla 1: Variables | Datos Central de Deudores BCRA

Variable	Tipo	Observaciones
Código de identidad	Numérico	Código identificatorio para cada entidad, tanto bancaria como intermedia
Tipo de identificación	Numérico	Determina si la identificación se trata de CUIT, CUIL, Clave de Identificación (CDI) o deudores residentes en el extranjero
Número de identificación	Caracter	Número único para cada persona física / jurídica
Situación *	Numérico	Número del 1 al 6, donde 1 es 'situación normal' y 6 es 'irrecuperable por disposición técnica'. Valores intermedios reflejan, de menor a mayor, distinta gravedad en los días de mora.
Monto *	Numérico	Monto en miles de pesos
Proceso Judicial / Revisión *	Numérico	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 repetirán para cada uno de los 24 meses de información hasta llegar al campo 75.

En este trabajo se utilizará el dataset correspondiente al mes de noviembre del 2019. Es decir, desde esa fecha inclusive, y hasta 23 meses hacia atrás (diciembre del 2017 inclusive). Éste incluye 44.737.761 filas / deudas.

Sólo se tendrán en cuenta las personas físicas. Es decir, las empresas serán dejadas de lado. Esto es así dado que el análisis de una persona física es sustancialmente diferente al de una jurídica. En la investigación de estas últimas se debe incluir el estudio de balances, pronósticos de cash flows y costo del capital, entre otros (Rikkers, Thibeault, 2015). Qué registros se corresponden a una persona física o jurídica se identifica utilizando los valores de la variable 'tipo de identificación'. Además, en este dataset, y en este trabajo, se incluyen los inmigrantes. Particularmente diferenciados por su número de identificación.

En caso de tratarse de una deuda que inició luego del primer mes de información (diciembre 2017), entonces desde ese mes hasta el inicio de la deuda, la situación de morosidad es igual a 0.[7] Es igual si fuese el caso opuesto, el cual la deuda ya se pagó por completo antes de llegar al último mes de información (noviembre 2019).

Como es de esperarse, en la Tabla 2 se puede observar que la gran mayoría de los créditos se encuentran en situación 1, mientras que los que están en una situación superior son la minoría. A problemas con esta característica se los suele llamar 'desbalanceados' .

Tabla 2 : Situación crediticia y cantidad de créditos. Mes: Diciembre 2017

Situación	**0**	**1**	**2**	**3**	**4**	**5**	**6**
Cantidad Absoluta	15.530.491	25.066.564	895.347	640.895	899.515	1.703.287	1.662
Cantidad Porcentual	34,71%	56,03%	2,00%	1,43%	2,01%	3,81%	0,004%

Este desbalance se analizará más en detalle luego, puntualmente con el mes para el cual se estudiarán los datos y generarán variables para el modelo predictivo.

2.1.2 - Clasificación e intuición de los tipos de entidades

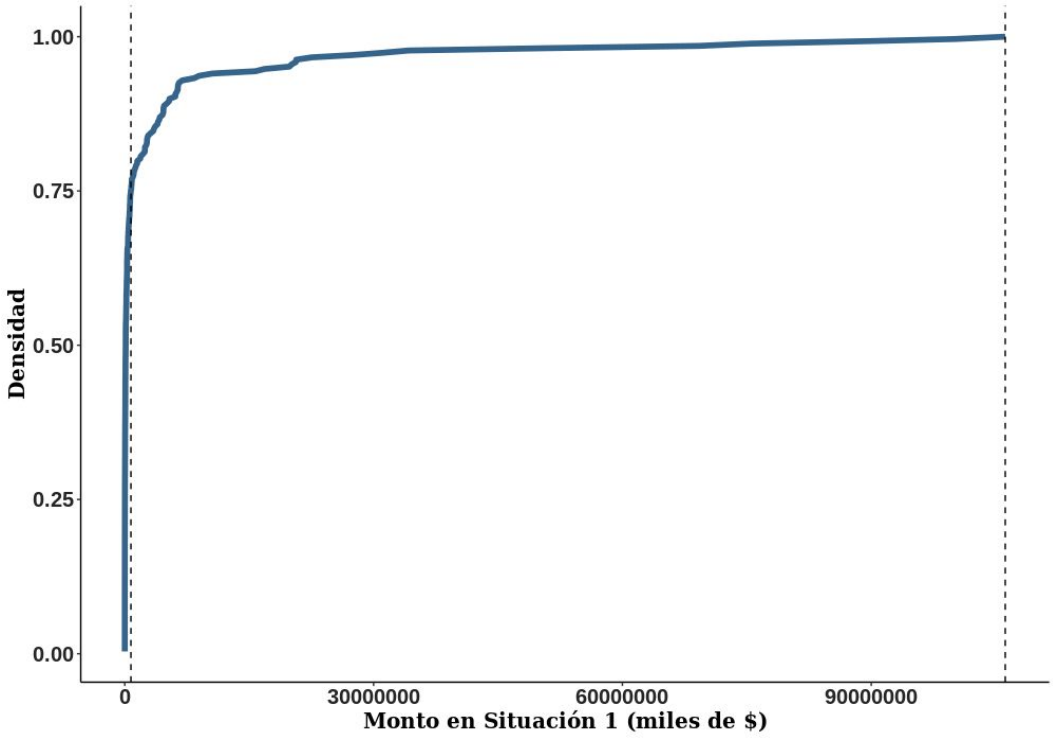
[7] Estas filas no serán tomadas en cuenta para el análisis y armado de los modelos, dado que son consideradas información futura que no debería de ser conocida de antemano por los algoritmos (data leakage).

Por las características propias de este mercado, tiene sentido diferenciar entre entidades grandes y chicas, ya que, tal como se mencionó más arriba, son distintas las herramientas que tienen para cobrar sus préstamos, y también las consecuencias que implican para las personas el dejar de pagarlos. El criterio de división fue enteramente en base a los datos del mes de diciembre del 2017 (primer mes de datos disponible). Concretamente, se calculó el volumen prestado para cada entidad y luego se tomó un punto de corte. Intuitivamente, se deberían tomar todos los datos disponibles; pero, dada una inconsistencia en la carga de datos por parte de las entidades, se tomó solamente los créditos en situación 1. Esta inconsistencia se debe a que no hay una regla establecida de cuándo dar de baja una mora por parte de los acreedores. Hay empresas que en cuanto establecen como incobrable un determinado crédito (o sobrepasa cierta situación crediticia particular, determinada por cada entidad), lo eliminan del sistema. En este caso, si una persona sólo tenía esta deuda, ya no figuraría 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. Por lo tanto, tener en cuenta situaciones distintas de 1 para calcular el volumen de las entidades genera distorsiones en los datos, y por ende en las conclusiones que se puedan llegar a tomar.

Por otro lado, puede 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, hay empresas como bancos tradicionales que generalmente sólo prestan en situación 1, y hay entidades intermedias que se especializan en prestar a individuos incluso en situación 5. Por lo tanto, estas empresas estarían subvaloradas.

De todas formas, se decidió no tener en cuenta situaciones mayores a 1, dado que este segundo efecto es considerado secundario frente al problema mayor de tener empresas que informan deudas 'viejas' y otras que no. Muchas empresas chicas parecen sustancialmente más grandes de lo que son si tenemos en cuenta dichas deudas. Esto es así porque, aunque no hayan informado sobre la baja de la deuda, sí contabilizaron en sus balances las pérdidas y por ende no tienen tal patrimonio. La Figura 1 muestra cómo se distribuyen las entidades en función del volumen prestado en situación igual a 1.

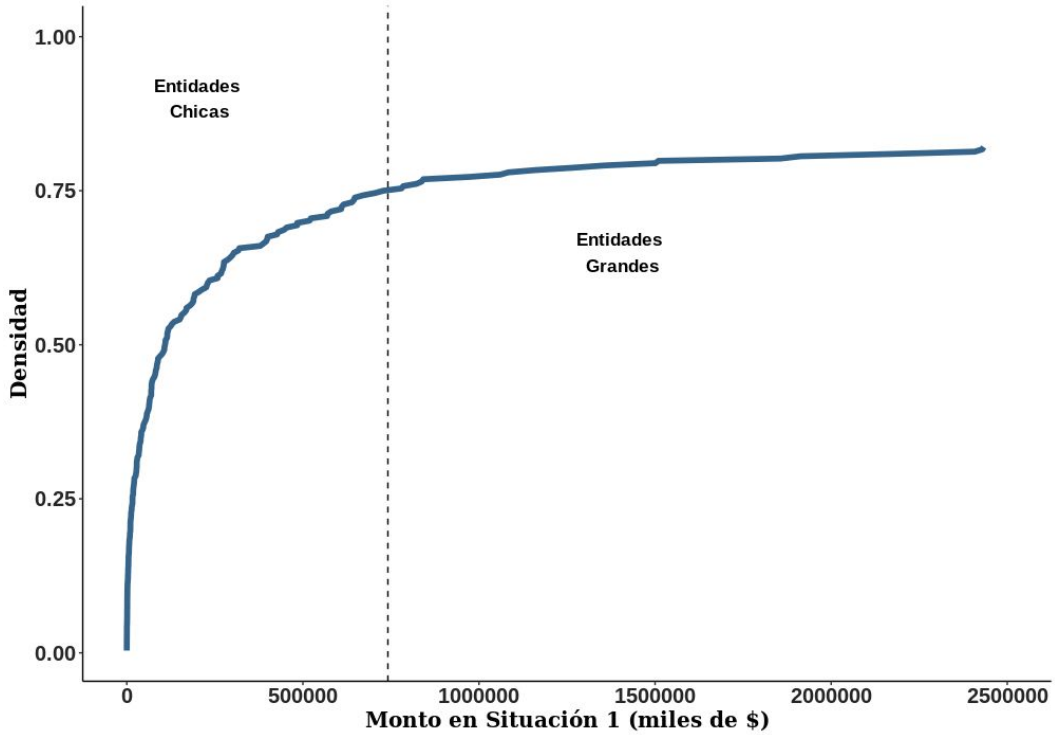
Figura 1: Función de Densidad Empírica del Volumen de las Entidades



Nota: Las líneas punteadas representan el 75% y el 100% de la cantidad de entidades, respectivamente.

En la Figura 1 se puede observar que aproximadamente el 75% de las entidades presentan montos similarmente bajos. Recién empiezan a diferenciarse en el 25% restante. Por lo observado en los datos, resulta coherente la decisión de separar las entidades en dos grupos según su tamaño; y, puntualmente, utilizar el 75% como punto de corte. La Figura 2 muestra exactamente ésto, clasificando hasta el tercer cuartil en 'entidades chicas', y al 25% restante como 'entidades grandes'. Entonces, de un total de 395 entidades que existen en el mercado, a 297 se las clasificaron como chicas, y a 98 como grandes.

Figura 2: Clasificación de Entidades



Nota: La línea punteada representa el tercer cuartil.

2.1.3 - Estructura de los datos a utilizar en los modelos

A partir del dataset original se construyeron 23 datasets adicionales, donde cada uno de éstos representa un mes particular. La estructura es completamente diferente: aquí las filas ya no representan una deuda, sino una persona. Es decir, ya no hay filas repetidas para un mismo individuo.

Como se explicará en la Sección 2.4 - Tendencias, en este trabajo se intentará mostrar el efecto marginal que genera la inclusión de variables de tendencias en los modelos, utilizando un rezago de 3 meses. Para esto se construyeron los mismos datasets anteriores, pero incluyendo dichas variables. Por lo tanto, se hicieron, en total, 43 datasets nuevos.[8]

[8] 23 datasets sin tendencias y 20 con tendencias. Notar que para los primeros 3 meses disponibles no pueden crearse datasets con tendencias, dado que no hay 3 meses anteriores para utilizar de rezagos.

Cada una de estas bases de datos tienen variables con información de las deudas de un sólo mes, a excepción de las variables a predecir que incluyen información futura (del mes siguiente), y de las variables de tendencias, que incluyen información pasada. Por ejemplo, si el dataset es el correspondiente al de marzo del 2018 con tendencias, todas las variables serán sobre los montos y cantidades de deudas de ese mes, salvo las variables a predecir que tendrán información de abril del 2018, y las variables de tendencia que tendrán información de 3 meses hacia atrás . Las variables a predecir son tres:

1. 'y_grandes': Toma valor 1 si para el mes en cuestión la persona se encuentra en situación 1 y, además, se convierte en morosa en el mes siguiente; considerando exclusivamente entidades grandes. Toma valor 0 en caso contrario.
2. 'y_chicas': Toma valor 1 si para el mes en cuestión la persona se encuentra en situación 1 y, además, se convierte en morosa en el mes siguiente; considerando exclusivamente entidades chicas. Toma valor 0 en caso contrario.
3. 'y_total': Toma valor 1 si para el mes en cuestión la persona se encuentra en situación 1 en ambos tipos de entidades y, además, se convierte en morosa en el mes siguiente en al menos uno de los dos tipos de entidades. Toma valor 0 en caso contrario.

La forma de clasificar a una persona en morosa es subjetiva. No hay una regla general que determine la demora necesaria en el pago para poder clasificarla como tal. Este límite lo determina cada empresa de forma independiente. En este trabajo se clasifica a un individuo como moroso siempre que tenga **al menos una de sus deudas en situación mayor a 1**. A modo de ejemplo, si una persona tiene en el mes actual 3 créditos en situación 1, y uno de ellos, otorgado por una entidad chica, pasa a situación 2, entonces 'y_chicas' será igual a 1, 'y_grandes' igual a 0, e 'y_total' igual a 1.

El resto de las variables que integran cada uno de los datasets mensuales serán explicadas más adelante en la Sección 2.3 Ingeniería de Atributos.

2.2 - Metodología

2.2.1 - Modelos

Como se mostrará a continuación, los datos se encuentran muy desbalanceados. Debido a esto, y a su gran poder predictivo, se utilizará el algoritmo Extreme Gradient Boosting (XGBoost) para predecir si un individuo, que ya posee deudas, será moroso en al menos una de ellas el próximo mes (Kuhn, Johnson, 2016). Para esto se tendrá en cuenta la clasificación de entidades grandes y chicas. Es decir, se armarán modelos independientes que buscarán predecir, por un lado, si una persona se convertirá en morosa en al menos un préstamo con entidades chicas, y por otro, si se convertirá en morosa en al menos un préstamo con entidades grandes. Además se construirá un tercer tipo de modelo que buscará predecir si una persona se convertirá en morosa en al menos uno de los dos tipos de entidades.

Por otro lado, como se explicará en detalle luego, se analizará el efecto de incluir variables de tendencia en los modelos. Estos modelos con tendencias tendrán como objetivo predecir lo mismo que los otros 3 modelos anteriores, y entonces así poder compararlos.

De esta forma, se evaluarán 6 modelos:

1. Predecir si un deudor se convertirá en moroso en alguna entidad el mes siguiente. (y_total)
2. Predecir si un deudor se convertirá en moroso puntualmente en entidades grandes el mes siguiente. (y_grandes)
3. Predecir si un deudor se convertirá en moroso puntualmente en entidades chicas el mes siguiente. (y_chicas)
4. Predecir si un deudor se convertirá en moroso en alguna entidad el mes siguiente, utilizando en el modelo variables de tendencia. (y_total)
5. Predecir si un deudor se convertirá en moroso puntualmente en entidades grandes el mes siguiente, utilizando en el modelo variables de tendencia. (y_grandes)

6. Predecir si un deudor se convertirá en moroso puntualmente en entidades chicas el mes siguiente, utilizando en el modelo variables de tendencia. (y_chicas)

En la Tabla 3 se puede observar que las clases están muy desbalanceadas (se omitieron los NA dado que no son tenidos en cuenta para construir el modelo).[9]

Tabla 3: Proporción de cada clase para dataset de diciembre 2017

Variable	0	1
y_grandes	97,08%	2,92%
y_chicas	95,80%	4,20%
y_total	96,80%	3,20%

2.2.2 - XGBoost

XGBoost es un algoritmo de la familia de los 'boosting algorithms' muy eficiente en la reducción del sesgo y de la varianza de un modelo. Estos algoritmos implican crear muchos árboles de decisión a partir de los datos de entrenamiento. La particularidad más importante es que los árboles son creados de manera secuencial. Es decir, cada árbol utiliza información del árbol anterior. Puntualmente, cada nuevo árbol tiene el objetivo de explicar los residuos del modelo previo. Es por esto que se suele decir que son algoritmos que 'aprenden despacio' (James, Witten, Hastie, Tibshirani, 2017). Entrenar todos los árboles de forma simultánea es un problema muy complejo e imposible computacionalmente. Por lo tanto, se opta por entrenar sucesivamente uno por vez (Chen, Guestrin, 2016).

[9] Notar que si una persona, para un tipo de entidad (grande o chica), no tiene créditos, entonces la variable a predecir para ese grupo toma valor NA.

Los típicos hiperparámetros a optimizar en XGBoost son:

Tabla 4: Hiperparámetros de XGBoost a optimizar

Hiperparámetro	Descripción	Rango
max_depth	Máxima profundidad de los árboles	$(0; \infty]$
eta	Proporción que aprende de cada árbol	$[0; 1]$
gamma	Mínima reducción del error necesaria en una hoja para generar una nueva partición	$[0; \infty]$
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	$[0; \infty]$
nrounds	Cantidad de árboles a construir	$(0; \infty]$

2.2.3 - Métrica de evaluación de modelos

Para evaluar los modelos y poder compararlos se utilizará el área bajo la curva ROC (AUC), métrica que mide el grado de separación de las clases; especialmente útil cuando se trata de problemas con clases muy desbalanceadas (Tan, Steinbac, Kumar, 2006). Puntualmente relaciona la tasa de verdaderos positivos ($TPR = \frac{TP}{P}$) con la de falsos positivos ($FPR = \frac{FP}{N}$) (Davis, Goadrich, 2006). Esta métrica toma valores en un rango de $[0\ ;\ 1]$, donde mayor AUC implica mejor *performance* del modelo. Un valor igual a 0,5 es sinónimo de predecir de forma

azarosa, igual a 1 representa una separación perfecta, y menor a 0,5 sugiere una *performance* peor que el azar.

2.2.4 - Conjuntos de entrenamiento, validación y testeo

El armado de los conjuntos de entrenamiento, validación y testeo se hace teniendo en cuenta la temporalidad de los datos. Concretamente, se entrenarán todos los modelos con el dataset correspondiente al mes de agosto del 2019. Se utilizará el del mes siguiente, septiembre del 2019, como validación. Una vez que se haya elegido el mejor modelo para cada variable a predecir, se los testeará en el dataset correspondiente al mes de octubre del 2019. Allí se obtendrá el rendimiento final de cada modelo. Luego, utilizando estos datos (octubre del 2019), se re-entrenarán los modelos elegidos (utilizando la mejor combinación de hiperparámetros encontrada para los datos de septiembre 2019); para luego, a partir de los datos de noviembre del 2019, predecir los morosos de diciembre 2019.

En la Tabla 5 se muestra la cantidad de filas que posee cada dataset. Como se mencionó anteriormente, cada una de éstas representa una persona. En la Sección 2.3 - Ingeniería de atributos se detallará la cantidad de columnas y explicará cada una de ellas.

Tabla 5: Cantidad de individuos (filas) en los datasets utilizados para el armado de los modelos

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

2.2.5 - Modelos *Benchmark*

Con el objetivo de tener una medida base para poder comparar los modelos construidos, se creó un modelo *benchmark* para cada uno de los modelos desarrollados. Éstos son considerados 'ingenuos' ya que no buscan ser complejos, ni tener sus hiperparámetros optimizados. El fin de esta implementación no es más que el de poder mostrar cómo, utilizando

los mismos datos, un modelo complejo se diferencia de uno sencillo en su capacidad predictiva; y de esta forma justificar su relevancia. Para esto se construyeron otros 6 modelos, que buscan predecir las mismas variables que los 6 explicados anteriormente. Dado que en esta etapa el objetivo no es desarrollar modelos sofisticados, se decidió utilizar árboles de decisión, con sus hiperparámetros por *default*.[10] Los resultados son presentados a continuación en la Tabla 6.

Tabla 6: Modelos benchmark

Modelo	AUC en validación	AUC en testeo
Y_grandes	0,67	0,70
y_chicas	0,67	0,66
y_total	0,72	0,71
y_grandes con tendencias	0,74	0,73
y_chicas con tendencias	0,70	0,68
y_total con tendencias	0,73	0,73

A lo largo de este documento se buscará crear modelos que logren mejorar este desempeño. Puntualmente, en la Sección 4 - Resultados se presentan los resultados alcanzados.

2.2.6 - Optimización de hiperparámetros

La estrategia utilizada para encontrar los mejores hiperparámetros fue realizando un 'random search'. Alternativamente, se podrían buscar a través de un 'grid search' que pruebe

[10] Vale la pena mencionar que como los datos presentan valores con NA, es imposible utilizar métodos como regresión lineal o regresión logística. Para poder implementar árboles de decisión, teniendo en cuenta la presencia de este tipo de valores, se utilizó XGBoost con *nrounds* igual a 1 (véase Sección 2.2.2 - XGBoost). Dado esto y que los hiperparámetros del estilo *colsample_by* por default toman valor 1, el resultado es similar al de un árbol de clasificación o regresión (CART, por sus siglas en inglés), pero capaz de trabajar con NA.

cada combinación posible con cada valor prefijado, pero esto es muy costoso computacionalmente y no trae mayores beneficios. Se demostró que random search es una forma mucho más eficiente y de igual (y a veces mayor) eficacia para encontrar los mejores valores (Bengio, Bergstra, 2012).[11] La metodología consiste en definir un rango de posibles valores para cada hiperparámetro y luego, aleatoriamente, seleccionar uno para cada uno. De esta forma quedan seleccionados los hiperparámetros correspondientes para un posible modelo. Repitiendo esto varias veces, quedan armados distintos modelos a evaluar.

Puntualmente, en este trabajo, se realizó en dos etapas:

1. Se probaron 15 configuraciones de hiperparámetros distintas, tomando un rango muy amplio como posibles valores . Luego se seleccionó la mejor de ellas.

   Los rangos seleccionados para cada hiperparámetro son:

Tabla 7: 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% para ambos sentidos. Es

[11] Existen otras técnicas eficientes, como métodos bayesianos de optimización de hiperparámetros. Sin embargo, se eligió 'random search' dada su sencillez y su uso intensivo en la industria.

decir el rango de cada hiperparámetro resultó ser $[0,85 \times valor_{etapa\ 1}; valor_{etapa\ 1} \times 1,15]$

Como resultado de esto, para cada modelo, se probaron 22 configuraciones de hiperparámetros distintas, definiendo como óptima a la que mejor predice en los datos de validación.

2.3 - Ingeniería de atributos

A partir del dataset público original se construyeron 23 dataset adicionales. Como se explicó anteriormente, cada uno de éstos tiene información para un único mes determinado, a excepción de las variables a predecir que poseen información del mes siguiente, y de las variables de tendencia que poseen información pasada. Cada fila representa una persona física (13.325.397 filas para el caso del mes de agosto 2019), y cada columna es una variable explicativa de los modelos a construir. Se crearon 839 columnas:[12]

1. 3 variables a predecir
2. 18 variables generales
3. 12 ratios
4. 395 variables donde cada una representa una entidad
5. 395 variables donde cada una indica la situación en cada entidad
6. 16 variables de tendencia

[12] Esto es así para los datasets con tendencias. Aquellos que no las incluyen tienen 823 columnas.

1. Variables a predecir

Tabla 8: Variables a predecir

Variable	Descripción
y_chicas	Toma valor 1 cuando el individuo actualmente no es moroso en entidades chicas pero lo será en el mes siguiente. 0 en caso contrario.
y_grandes	Toma valor 1 cuando el individuo actualmente no es moroso en entidades grandes pero lo será en el mes siguiente. 0 en caso contrario.
y_total	Toma valor 1 cuando el individuo actualmente no es moroso en ningún grupo de entidad, pero lo será en al menos un grupo en el mes siguiente. 0 en caso contrario.

2. Variables generales

Tabla 9: Variables generales

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

3. Ratios

Tabla 10: 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

4. Entidades

Tabla 11: Variables sobre entidades

Variable	Descripción
entidad_<ent>	Toma valor 1 si el individuo tiene al menos una deuda 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.

5. Situaciones

Tabla 12: Variables sobre situaciones en cada entidad

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

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.

6. Variables de tendencia

Estas variables serán explicadas en la Sección siguiente 2.4 - Tendencias

2.4 - Tendencias

Uno de los desafíos más grandes de utilizar XGBoost en problemas donde el paso del tiempo importa es justamente poder reflejar cómo cambian las variables predictivas a lo largo del tiempo. Por otro lado, muy probablemente, no utilizar la información pasada para intentar predecir resultará en un modelo más limitado. Es por esto que se crearon variables de tendencias que tratan de captar el comportamiento temporal de las deudas para cada individuo. Más adelante se analizará el efecto marginal de incorporar estas tendencias, al comparar los modelos que las incluyen y los que no (detallados en la Sección 2.2 - Metodología)

Puntualmente se armaron 16 variables de tendencia. Cada dataset incluye 8 nuevas variables creadas sobre la variable 'suma_monto_total_grandes', y 8 sobre 'suma_monto_total_chicas'. Estas variables consisten en tomar las diferencias mes a mes en términos absolutos y en términos porcentuales del monto total adeudado, utilizando el mes actual y 3 meses anteriores; y así calcular 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

Se eligió utilizar 3 meses de 'lag' dado que tener una ventana temporal más amplia implica, en cierto sentido, atenuar la información de los sucesos más recientes. Por otro lado, otras variables pueden ser tenidas en cuenta, como por ejemplo, la suma total adeudada en situación mayor a 1, pero por una cuestión de poder de cómputo, sólo se pudo utilizar la suma total adeudada para cada tipo de entidad.

De esta forma, las variables de tendencias creadas son:

Tabla 13: Variables de tendencia

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'

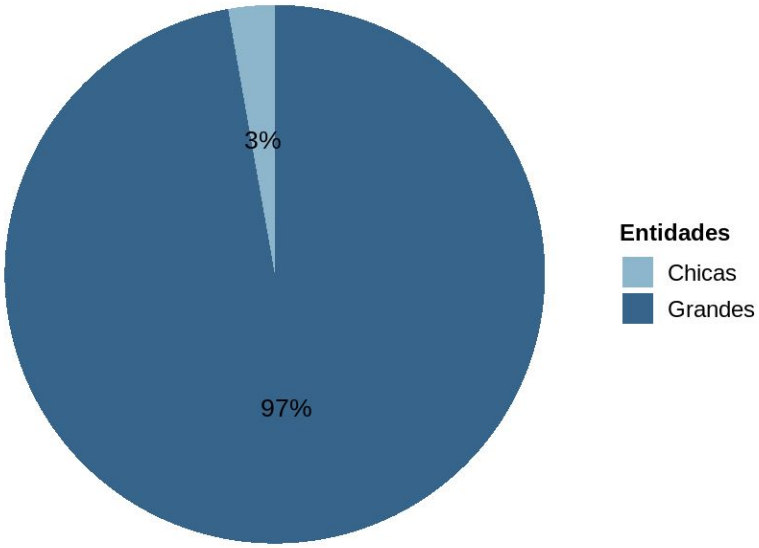
3 - Análisis descriptivo

En esta sección se presentará un análisis exploratorio de los datos analizados. El objetivo detrás de esto es doble: por un lado, poder obtener información de los datos para intentar de entender cómo se comporta el mercado y poder tomar decisiones mejor informados; y, por otro, descubrir las variables relevantes para predecir y así incluirlas en los modelos. Para esto se estudiará el mercado de forma agregada, pasando luego a un análisis diferenciado por individuos.

3.1 - Análisis del mercado

El mercado de créditos en la Argentina se encuentra muy concentrado en manos de unos pocos jugadores. Como se mostró anteriormente, las entidades que se categorizaron como grandes representan el 25% del total. Sin embargo, como se ve en la Figura 3, tienen el 97% del volumen total prestado.

Figura 3: Proporción del mercado según monto total



El motivo de esto, entre otros factores, podría ser el hecho de que los créditos de las entidades grandes (puntualmente bancos) tienen incluidos los saldos de las tarjetas de crédito. Confirmar esta idea resulta imposible con estos datos, ya que no hay información de qué tipo de crédito es cada deuda. Sin embargo, parece razonable pensar que las tarjetas son un factor importante y diferencial a la hora de explicar el volumen prestado.

Con esta información, pareciera que se podría llegar a la conclusión de que el mercado de entidades chicas es irrelevante para el análisis. Sin embargo, no hay que olvidar que estas empresas representan el 75% del total de entidades. Además, como se verá a continuación, atacan a mercados muy distintos, los cuales quedarían afuera del sistema si estas entidades no existieran. Muchas veces, incluso, son empresas que tienen como clientes personas del interior del país, los cuales no pueden acceder al sistema bancario tradicional porque directamente éstos no operan allí (Anastasi et. al, 2010). Es por todo esto que tiene sentido esta clasificación y que sea parte del análisis.

En las Figuras 4 y 5 se pueden observar la evolución de la cantidad de créditos según cada tipo de entidad a lo largo del tiempo.

Figura 4: Cantidad de créditos a lo largo del tiempo en entidades grandes

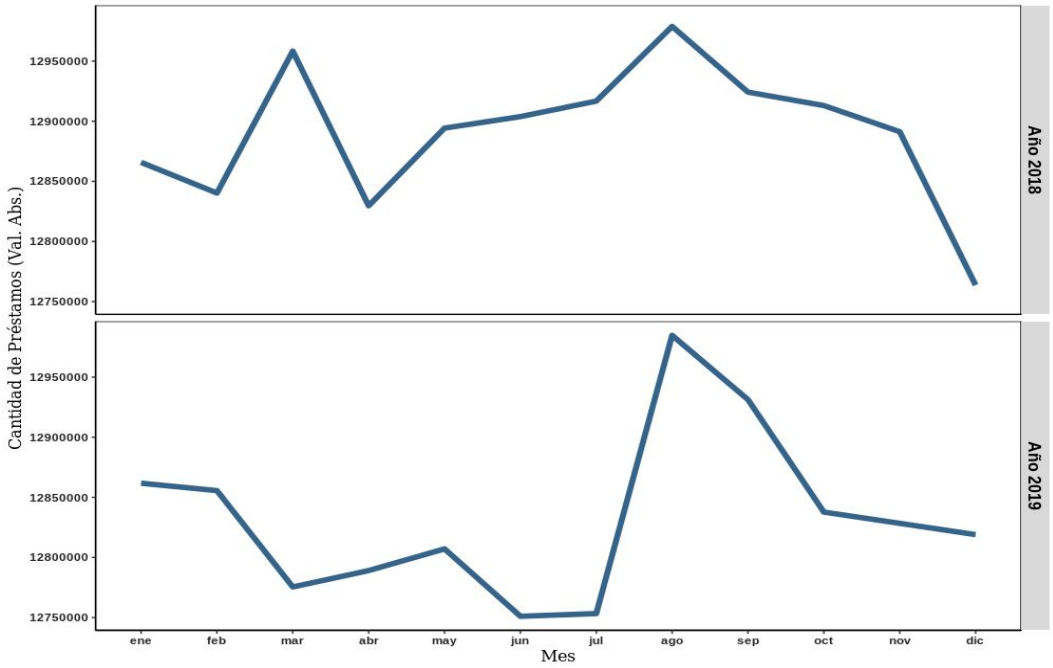
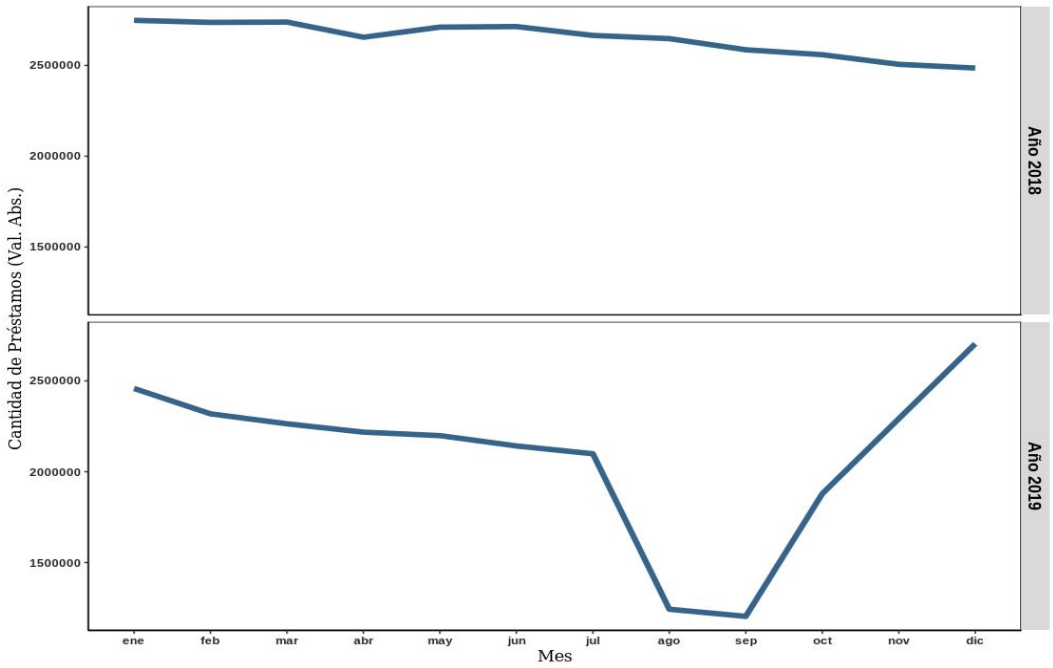


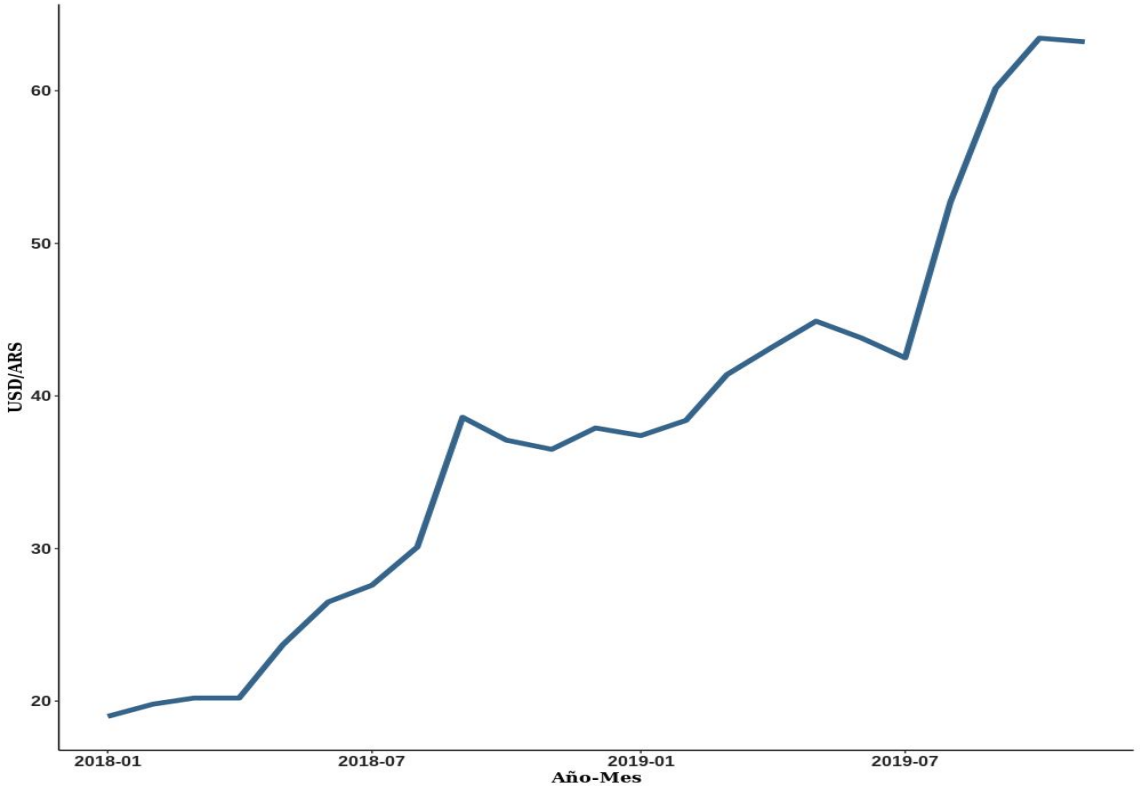
Figura 5: Cantidad de créditos a lo largo del tiempo en entidades chicas



Es notorio lo que sucede en los meses de agosto y septiembre del año 2019. Las entidades chicas tuvieron un marcado descenso en la cantidad de créditos otorgados, mientras

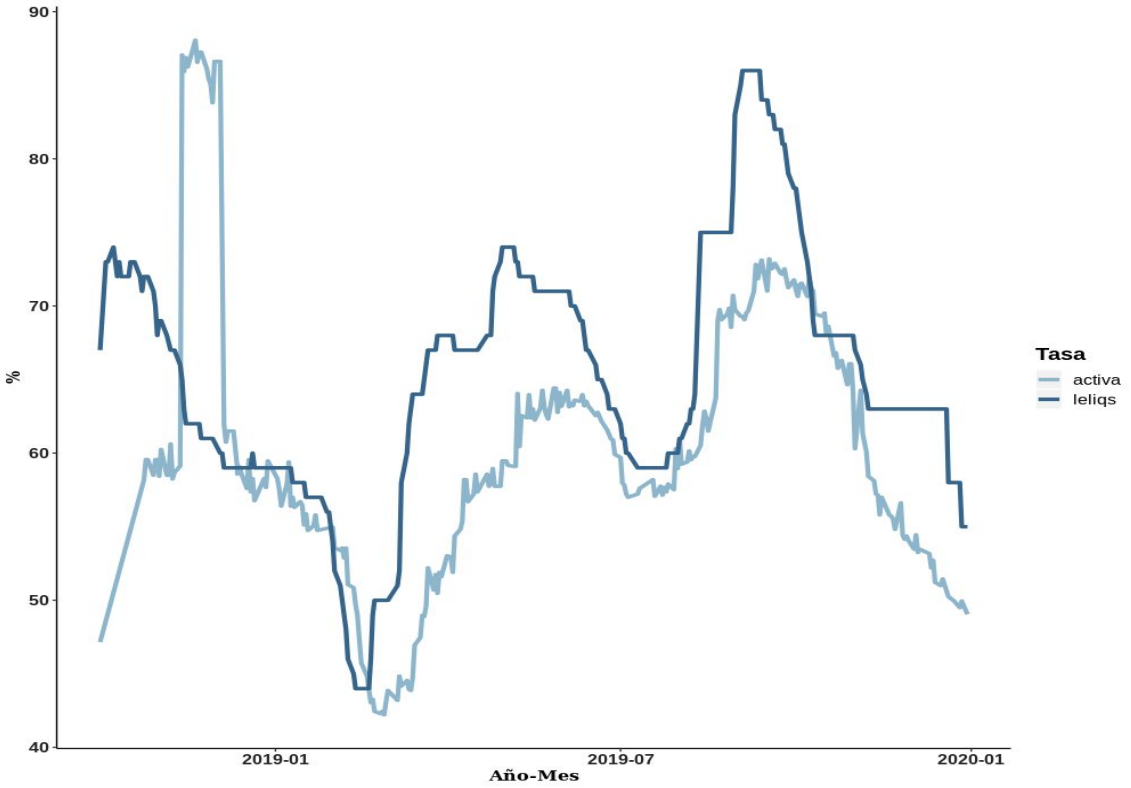
las entidades grandes tuvieron una suba abrupta. Esto puede ser explicado por las características propias de estos dos segmentos. Los bancos prestan los depósitos de sus clientes. Es decir, no están exponiendo su propio capital, ni fondeándose a tasas altas. Por otro lado, las entidades intermedias (mayoritariamente entidades chicas) prestan su patrimonio o se fondean a través de fondos de inversión, los cuales exigen un rendimiento en dólares. Por lo tanto, evaluar lo que sucede con el tipo de cambio es fundamental para entender este comportamiento. Expectativas de una devaluación del peso generan una contracción en la cantidad de créditos dados en las entidades chicas. La Figura 6 representa el valor del dólar para el mismo período que las Figuras 4 y 5. En esas fechas la presión en el tipo de cambio era un tema diario de discusión y las medidas tomadas por los organismos públicos estaban enfocadas en reducirla. Exactamente a mediados de agosto de 2019 el BCRA reduce la tenencia permitida de dólares por parte de los bancos (Barbería, 2019). Esto les quita la única alternativa que tenían para cubrirse de la devaluación y son 'incentivados' a aumentar la cantidad de créditos para así poder obtener algún rendimiento.

Figura 6: Tipo de cambio



Otra forma de calmar las presiones en el tipo de cambio es a través de la tasa de interés. A mayor tasa, menor incentivo a comprar dólares. La Figura 7 muestra la tasa de interés de las Letras de Liquidez (Leliqs), instrumento utilizado por los bancos para colocar sus excedentes de liquidez (y por el BCRA para esterilizar los pesos del sistema), y la tasa activa del Banco de la Nación Argentina (BNA). Con relación a esto, se ve cómo aumentan ambas tasas en los meses de agosto y septiembre del 2019. La tasa activa es aquella que cobran los bancos a sus clientes y puntualmente la que pagan las entidades intermedias cuando venden su cartera con responsabilidad a las entidades grandes. Esta suba tiene un efecto doble en estas entidades. Por un lado implica mayores costos para obtener fondos, y por ende mayor tasa a la cual colocan; y por otro lado, implica mayor morosidad esperada. Una mayor mora esperada genera que las entidades intermedias cobren tasas incluso más altas. Todo esto termina generando tasas extremadamente caras que imposibilitan la operatoria, y por consecuencia una baja en la cantidad prestada por parte de estas entidades.

Figura 7: Tasa de Leliq y Tasa Activa del BNA



Como conclusión, dada la situación macroeconómica de los meses de agosto y septiembre del 2019, era esperable el comportamiento observado en cada tipo de entidad. Puntualmente, la suba en la cantidad de créditos por parte de las entidades grandes, y la baja por parte de las chicas. Sin embargo, es válido preguntarse qué pasó con los montos totales prestados. Es decir, ya se vio que se dieron menos cantidad de créditos, pero ¿cambió el volumen total en pesos de cada tipo de entidad?

Figura 8: Monto total prestado en entidades grandes

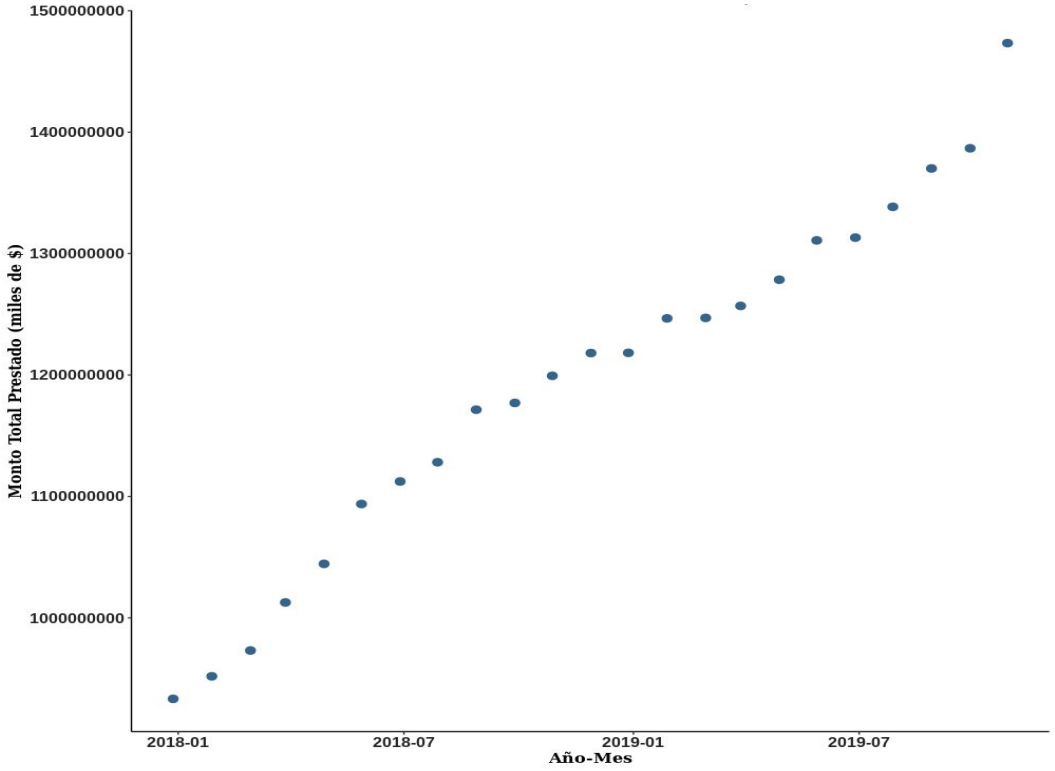
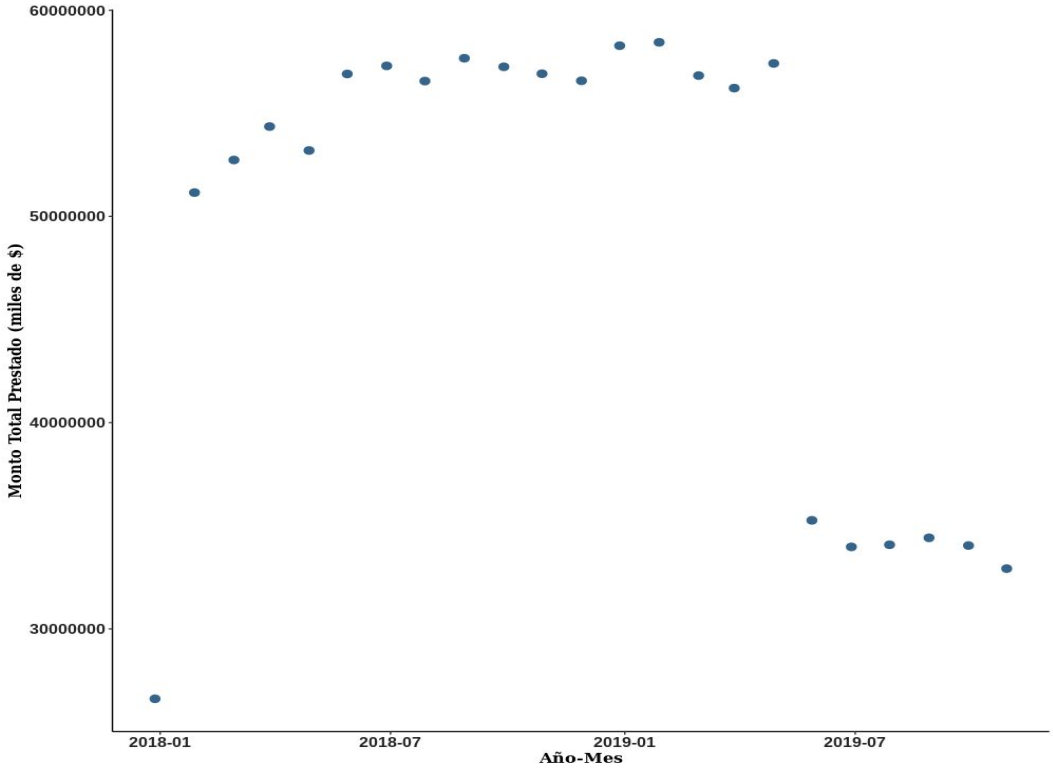
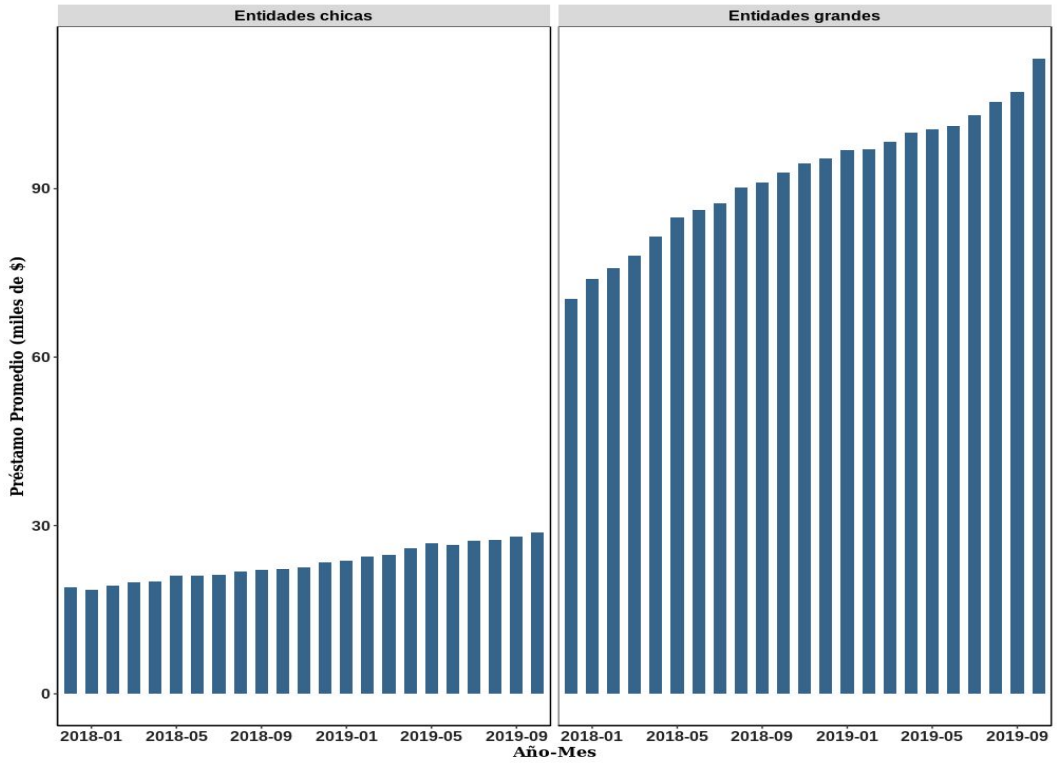


Figura 9: Monto total prestado en entidades chicas



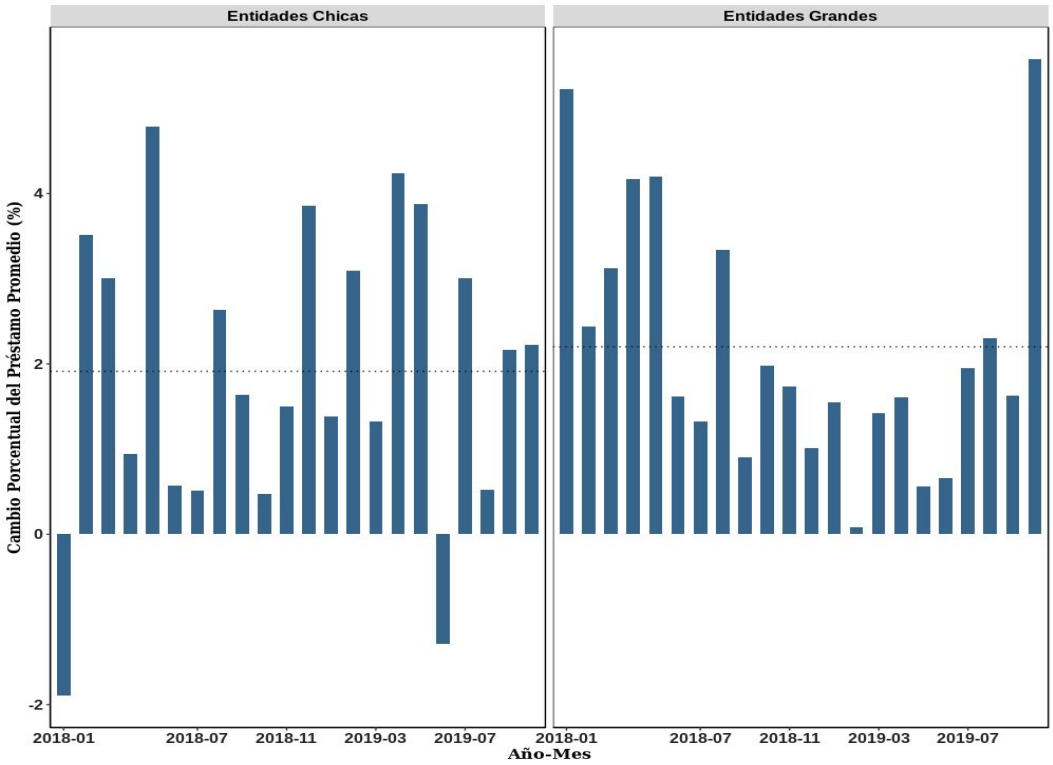
En las Figuras 8 y 9 se puede ver que el comportamiento va en la misma línea que lo planteado anteriormente: las entidades chicas achicaron su posición en pesos en los meses de agosto y septiembre del 2019, mientras que las entidades grandes lo aumentaron. Incluso se pude ver que estas entidades prestaron cada vez más, cuando las entidades chicas estuvieron relativamente estables, salvo los últimos meses. De los gráficos anteriores se desprende la Figura 10.

Figura 10: Préstamo promedio



Allí se puede observar la diferencia en el tipo de préstamos. Las entidades chicas claramente dan créditos más chicos que las entidades grandes. Esto puede deberse a varios factores. Por un lado, y como se verá a continuación en detalle, es explicado por el público al cual atienden. Estas firmas prestan a personas con mayor morosidad, y por ende tiene sentido que sigan una estrategia de microcréditos que diversifique el riesgo. Además, los solicitantes de estas deudas no suelen querer / poder endeudarse en grandes montos, dado las altas tasas que estas entidades cobran. Por otro lado, los bancos suelen prestarles a personas en situación 1, y por ende más confiables. Además, como se mencionó anteriormente, estos montos tienen incluidos los gastos con tarjeta de crédito.

Figura 11: Cambio porcentual del préstamo promedio

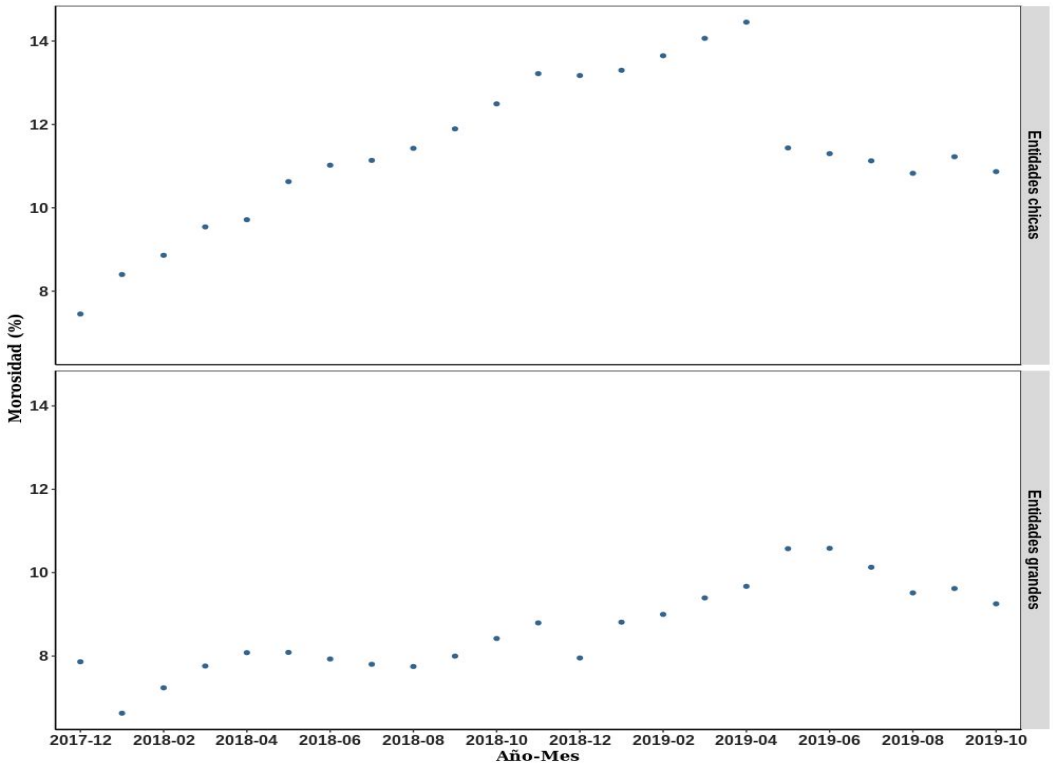


Nota: Las líneas punteadas representan la media para cada grupo

En la Figura 11 se puede ver incluso que las entidades grandes, en promedio, aumentaron proporcionalmente más el tamaño de su préstamo promedio que las chicas (líneas punteadas).

La contracara de lo que se ve en la Figura 10 es lo que se puede observar en la Figura 12.

Figura 12: Porcentaje de mora en función del tiempo y tipo de entidad



Es evidente que cada tipo de entidad trabaja con mercados distintos. La morosidad es sustancialmente mayor en las entidades chicas, donde el mínimo fue del 8% y el máximo del 16%; mientras que en las entidades grandes el mínimo es cercano al 1% y el máximo ronda el 10% (cercano a lo que era el mínimo de las entidades chicas). Además de lo explicado anteriormente sobre el tipo de clientes de cada entidad, hay dos cuestiones importantes a remarcar que pueden explicar este comportamiento.

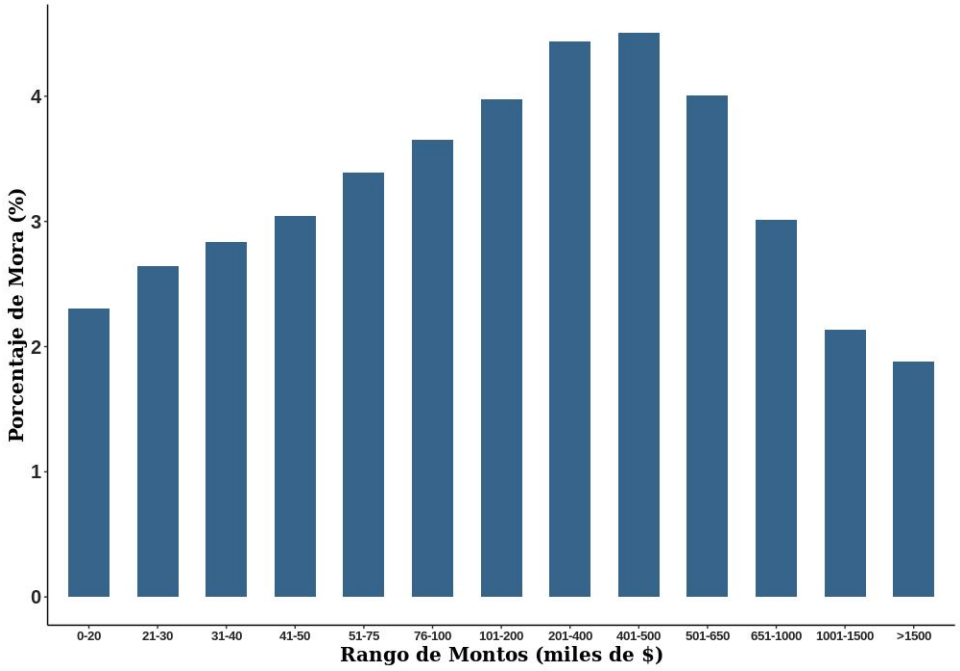
Primero, las herramientas de cobranza tienen un rol fundamental. Es decir, las entidades grandes (puntualmente los bancos) tienen mayor 'poder de fuego' para cobrar las cuotas. La gente cobra sus sueldos en sus cajas de ahorro y el banco automáticamente puede hacer uso de cierta cantidad de esos fondos para pagarse lo que le corresponde. Además, las consecuencias que implican ser moroso en un banco son más perjudiciales que las que implican serlo en una entidad intermedia. Estas últimas, pueden tener convenios o herramientas de cobranzas con los bancos, empresas grandes, sindicatos, etc. para trabajar de la misma forma. Claramente, disponer de este tipo de contratos baja sustancialmente el riesgo y el costo

operativo del proceso de cobranza. Esto es algo que tomó mucha más relevancia a fines de febrero del 2020 cuando el BCRA prohibió el cobro a través del débito de la cuenta del cliente de todo tipo de cuotas provenientes de préstamos. Esto sin dudas fue un revés para todas las empresas no bancarias dedicadas al sector crediticio, y que impactará negativamente en la oferta de créditos, y generando también un aumento en la tasa de interés. Anterior a esa normativa, era posible utilizar la Cámara Electrónica de Compensación de Medios de Pago Minorista de la República Argentina (COELSA) para debitar las cuotas,[13] pero por lo general, eran mucho más eficientes las herramientas de cobranzas.

Segundo, y por último, es notoria la tendencia que se puede observar en el último gráfico. La morosidad estaba aumentando de forma muy pronunciada a medida que pasaba el tiempo. Es incluso mayor para las entidades chicas. Esta tendencia puede ser explicada por los fenómenos macroeconómicos explicados anteriormente.

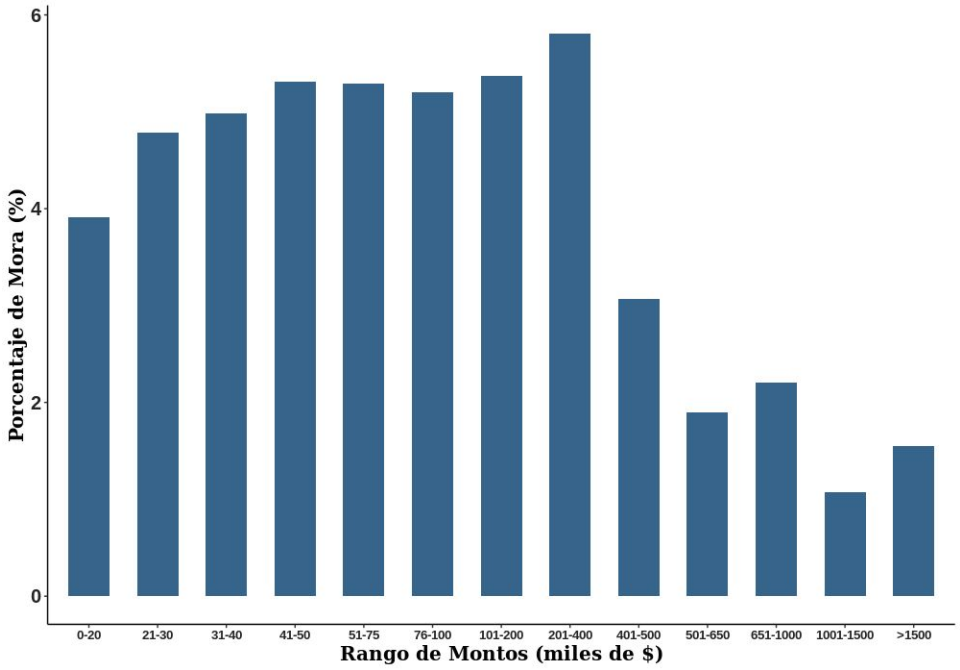
Siguiendo el análisis del comportamiento en la mora, en la Figura 13 y Figura 14 se muestran la proporción de deudores atrasados en sus pagos en función de los rangos de montos.

Figura 13: Proporción de mora según monto en entidades grandes



[13] http://www.coelsa.com.ar/

Figura 14: Proporción de mora según monto en entidades chicas



Intuitivamente uno creería que a mayor monto, mayor tasa de morosidad. Sin embargo, en ambos tipos de entidades sucede algo llamativo. En un primer tramo, se cumple que la mora aumenta a medida que crece la cantidad prestada; pero luego, pasado cierto umbral la mora cae drásticamente. Esto se explica porque pasado este límite (ambiguo y completamente empírico), los préstamos son tomados por personas con alto poder adquisitivo, los cuales tienen buena condición de repago. Además, es lógico pensar que una empresa (del tamaño que sea) analizará con mayor cuidado cuando los préstamos sean relativamente altos.

También se puede ver cómo se compone la cartera de morosos en ambos tipos de entidades. Las chicas concentran la mayor cantidad de mora y de forma bastante homogénea en los préstamos más bajos, mientras que las entidades grandes tienen una subida más escalonada y su máximo en montos mayores que las entidades anteriores. Esto último quizá explicado porque estudian con mayor cuidado a quiénes prestar a partir de montos más elevados que lo que lo hacen las entidades chicas. Puntualmente la Figura 15 muestra esta correlación.

En la Figura 16 se puede notar más claramente la correlación negativa que hay entre el tamaño de la entidad y su proporción de mora. Esto tiene sentido tras lo ya explicado sobre la capacidad que tienen las entidades bancarias de asegurarse su pago, frente a las entidades intermedias que no todas tienen garantizada una buena herramienta de cobranza. Además, también se explica por el hecho de que los bancos prestan a un sector distinto que las demás entidades, generalmente sólo prestan a individuos en situación igual a 1. Por lo tanto, es bastante de esperar el comportamiento observado.

Figura 15: Proporción de mora en función del tamaño del préstamo

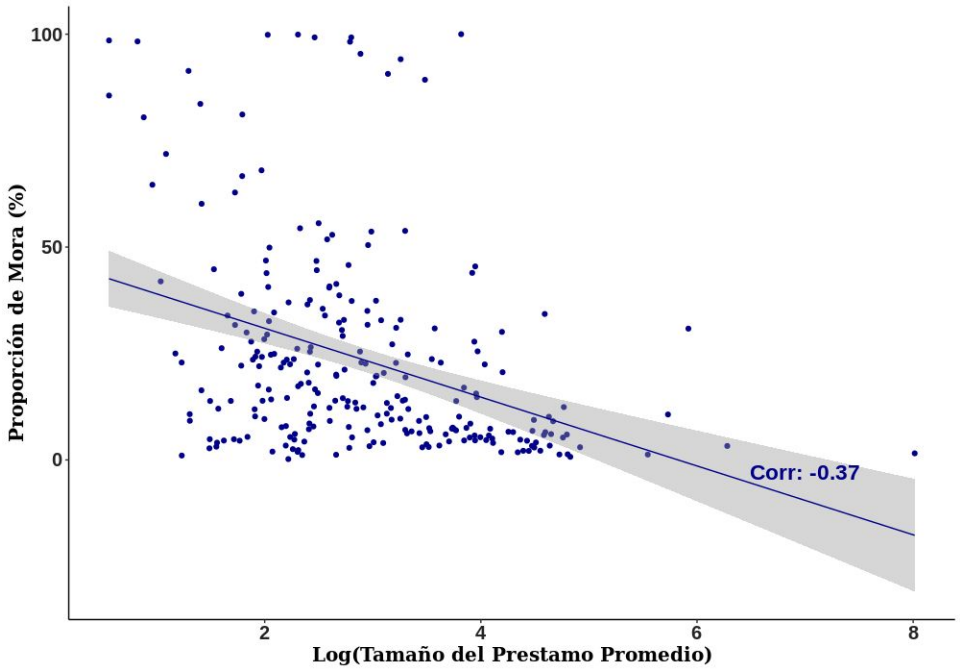
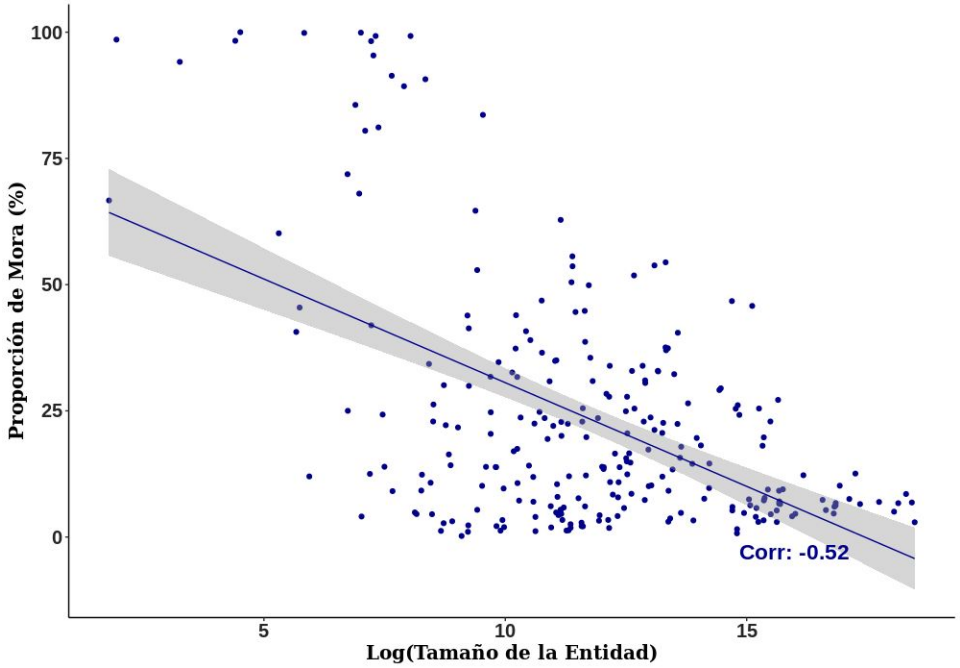
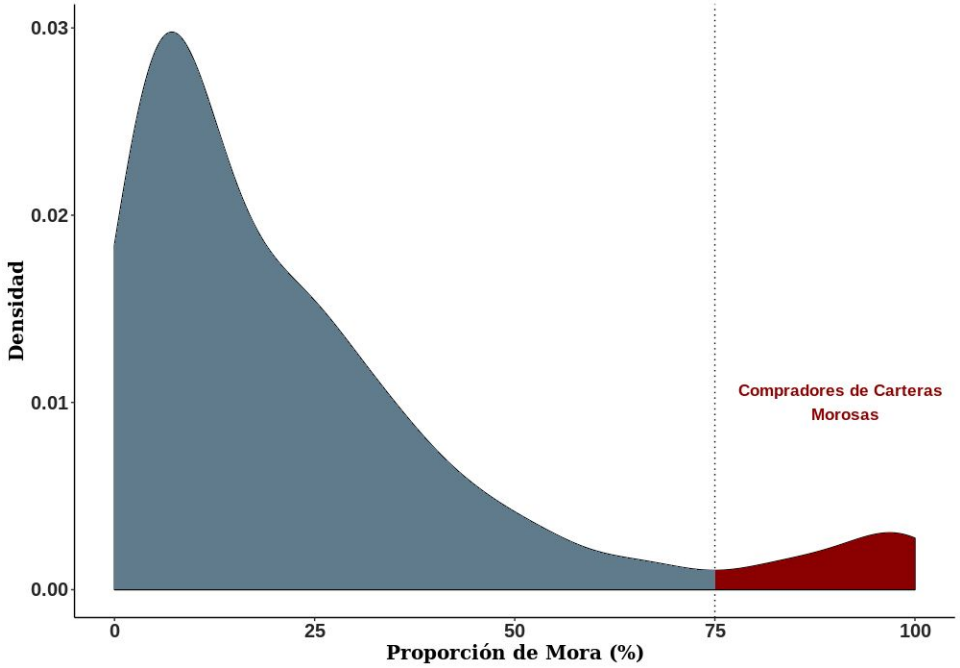


Figura 16: Proporción de mora en función del tamaño de las entidad



Algo curioso que llama la atención de estos últimos dos gráficos son las entidades que tienen 100% de morosidad, o valores muy cercanos a eso. Esto es un indicador de que hay otro tipo de empresas que trabajan de forma distinta a las analizadas en este trabajo, dado que claramente no pueden existir empresas de estas características que posean dichos valores y se mantengan en el largo plazo.

Figura 17: Distribución de las entidades en función de su proporción de mora



En la Figura 17 se puede observar que efectivamente hay una disminución de entidades a medida que aumenta la mora, pero que luego, en valores extremadamente altos de incumplimiento, aumenta nuevamente la cantidad de empresas. Es este grupo puntual el que llama la atención e indica que hay otro tipo de negocio en este mercado. Analizando las empresas que formaban parte de este grupo, se descubrió que eran firmas dedicadas a la compra a muy bajo precio de deudas morosas para llevarlas a juicio y esperar obtener un rendimiento una vez finalizado el litigio.

Por último, y relacionado con lo dicho anteriormente a lo largo de varios gráficos, se puede ver cómo está conformada la cartera de préstamos para cada tipo de entidad. La Figura 18 muestra que la gran concentración de créditos otorgados por entidades grandes son de 0-20 mil pesos, mientras que la cantidad para préstamos más altos es sustancialmente menor y pareja. Lo mismo sucede con las entidades chicas, visible en la Figura 19, con la diferencia de que la cantidad de préstamos dados en el tramo 0-20 mil es incluso mayor que en las grandes (75% vs. 40% aprox), y por consiguiente es menor la cantidad prestada en montos mayores.

Figura 18: Proporción de créditos en entidades grandes

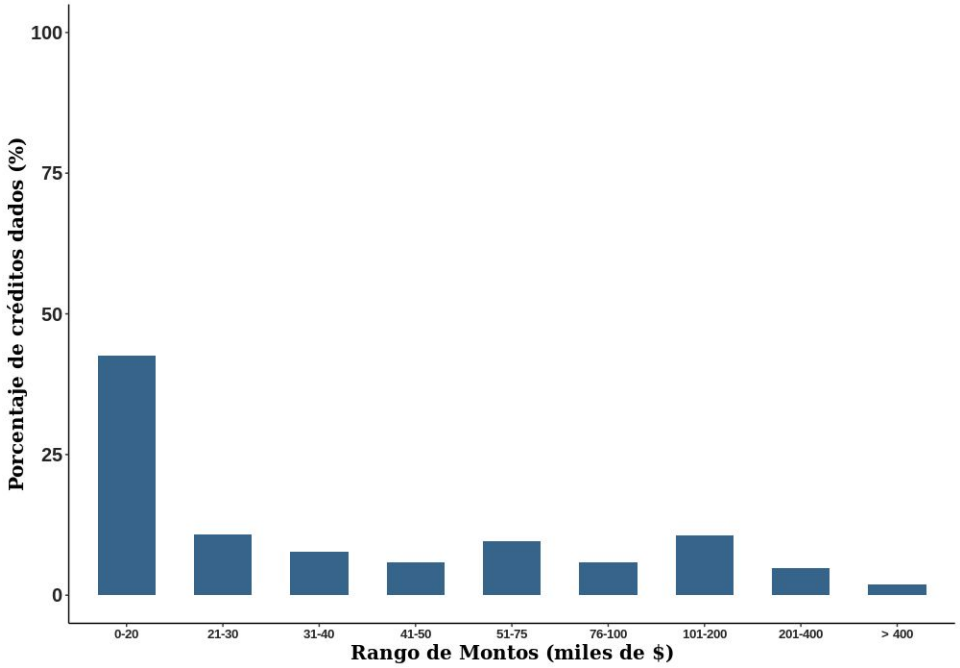
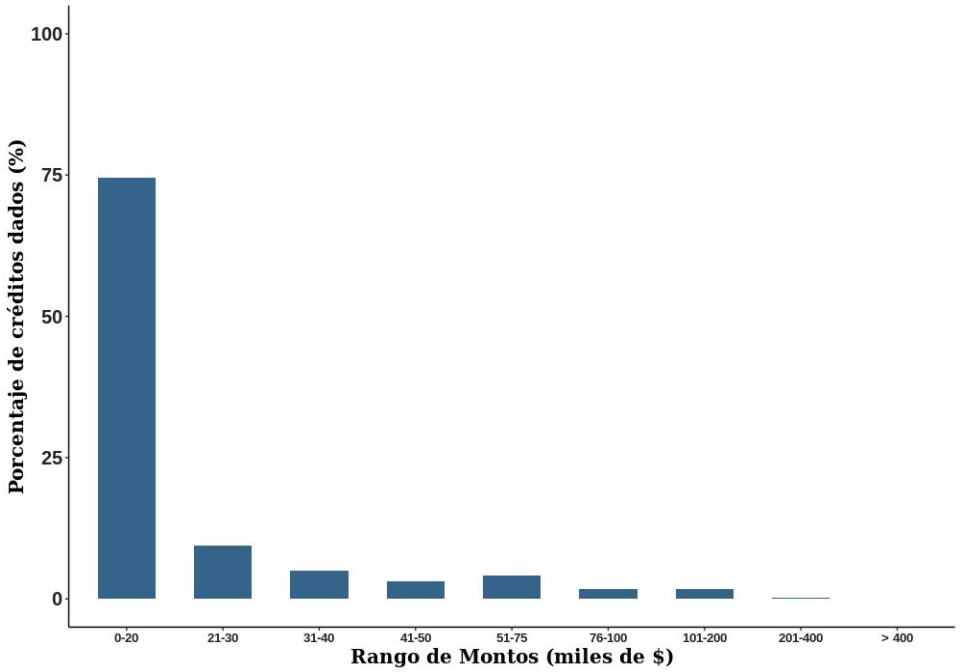


Figura 19: Proporción de créditos en entidades chicas



A modo de resumen de lo explicado en esta sección y como puntos importantes para el armado de los modelos se pueden enumerar los siguientes conceptos:

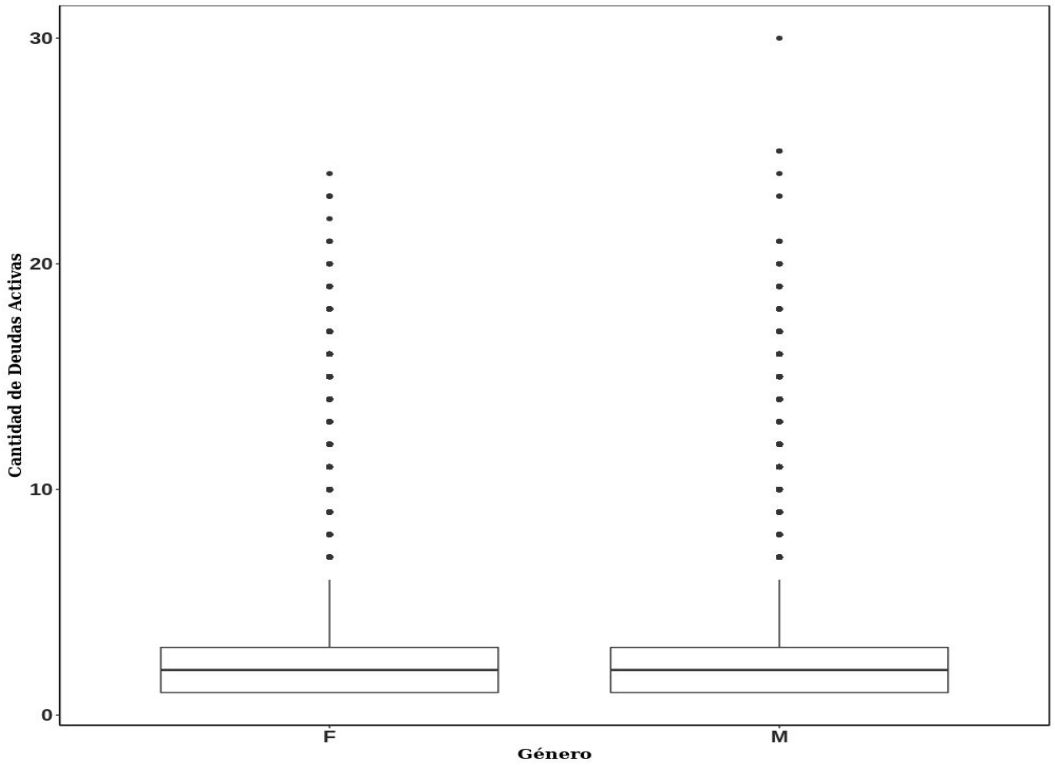
- Las entidades chicas dan créditos más chicos que las entidades grandes.
- Las entidades chicas tienen tasas de morosidad mayores que las entidades grandes.
- El tipo de entidad importa, dado que ser moroso en entidades grandes puede implicar mayores penalidades para el individuo con respecto a las entidades chicas. Por esto y por los puntos anteriores también, tiene sentido diferenciar según el tipo de entidad en los modelos predictivos.
- Incluir en los modelos todas las entidades donde un individuo es moroso es relevante dado que cada una tiene distintas herramientas de cobranzas y, por ende, distintas eficiencias operativas en el cobro de sus cuotas.
- Debajo de cierto umbral, a mayor monto adeudado, mayor tasa de morosidad. Por encima de dicho umbral, la relación es contraria.
- Existe un tercer tipo de entidades que se dedica a comprar carteras de deudas morosas.

3.2 - Análisis de los individuos

En esta sección se estudiará más en detalle cómo se comportan los individuos y se tratará de encontrar correlaciones entre sus características y comportamientos.

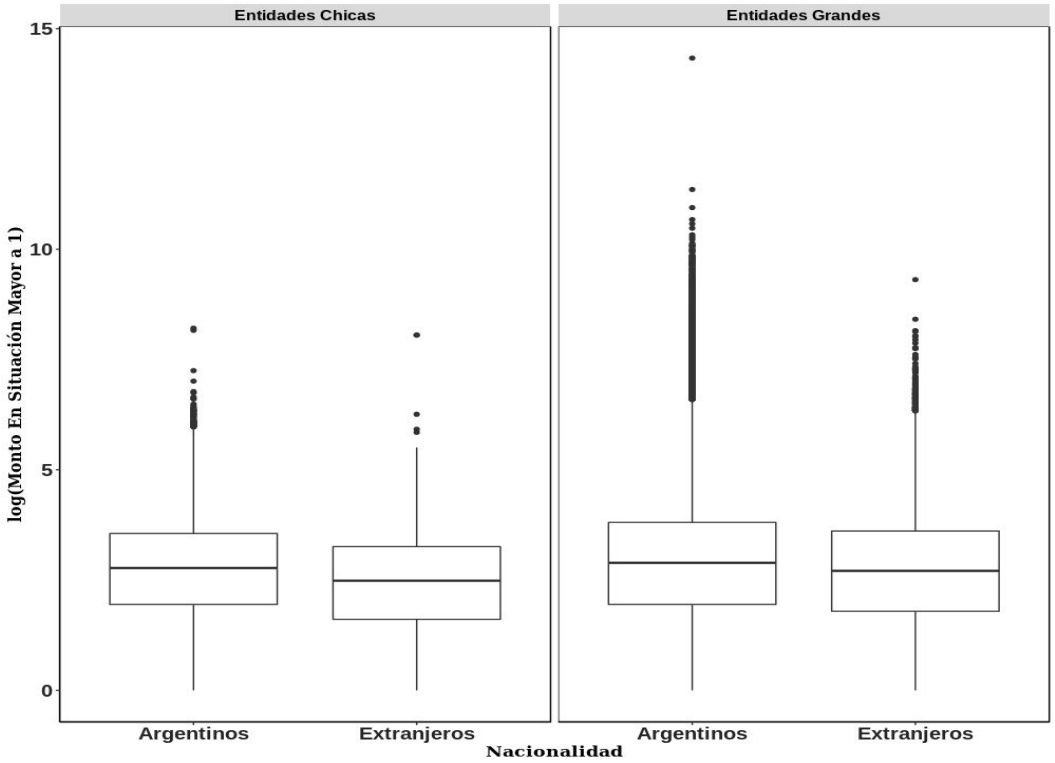
En la Figura 20 se ve que no habría mucha diferencia en las cantidad de deudas entre hombres y mujeres. Sin embargo, se decidió correr un test de hipótesis, donde la hipótesis nula es que el promedio de la cantidad de deudas activas de las mujeres es igual al de los hombres. Con un $p - valor < 10^{-15}$ se puede sostener que hay evidencia estadística significativa para rechazar la hipótesis nula.

Figura 20: Relación entre el género y la cantidad de deudas



Por otro lado, sí pareciera haber alguna diferencia entre extranjeros y ciudadanos argentinos. La Figura 21 muestra la relación entre ser extranjero y el logaritmo de la cantidad de de deudas en situación mayor a 1. Se observa que en ambos tipos de entidades los extranjeros presentan valores más bajos, aspecto que, en un principio, llama la atención. Además, también presentan menor dispersión. El sentido de esto puede venir por el hecho de que estos individuos tengan menos acceso al crédito o se cuiden más a la hora de tener moras, quizá por los riesgos que esto implica.

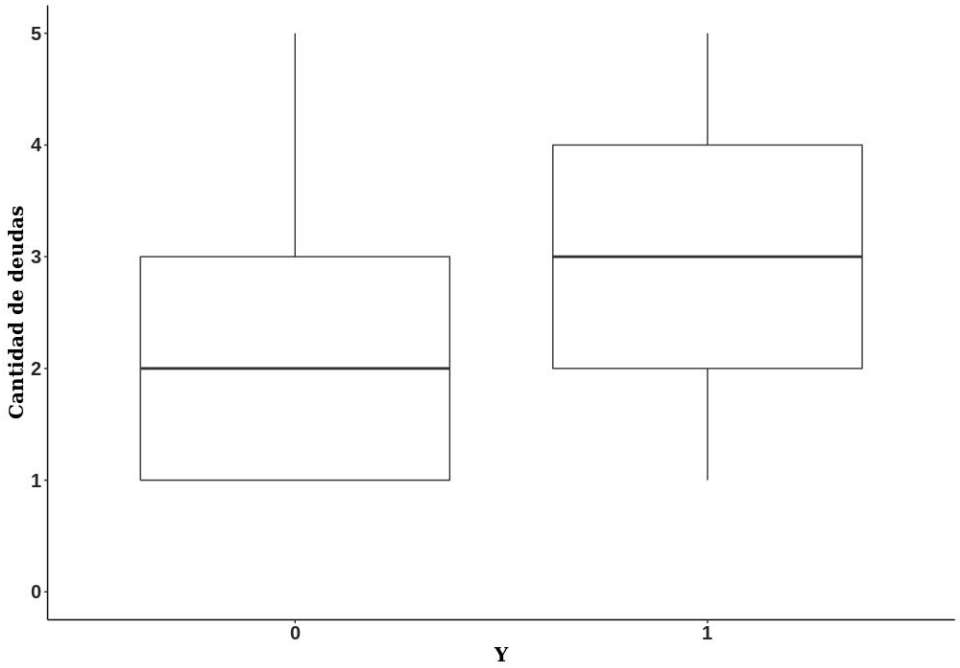
Figura 21: Relación entre ser extranjero y la cantidad de deudas en situación mayor a 1



Otro aspecto fundamental del análisis es tener en cuenta la variable a predecir para diferenciar ambos grupos (futuros morosos de no morosos) y estudiar por dónde pasan sus diferencias, o qué correlaciones existen entre ciertas variables y el hecho de convertirse efectivamente en moroso el mes próximo. Para esto se utilizó únicamente el mes de diciembre del 2017, primer mes de datos disponible. La idea detrás de esto es no tomar datos 'futuros' para predecir, ya que los modelos de machine learning empleados se podrían entrenar con meses intermedios, y por ende, en esos casos, se estarían usando datos que debieran ser desconocidos para definir variables.

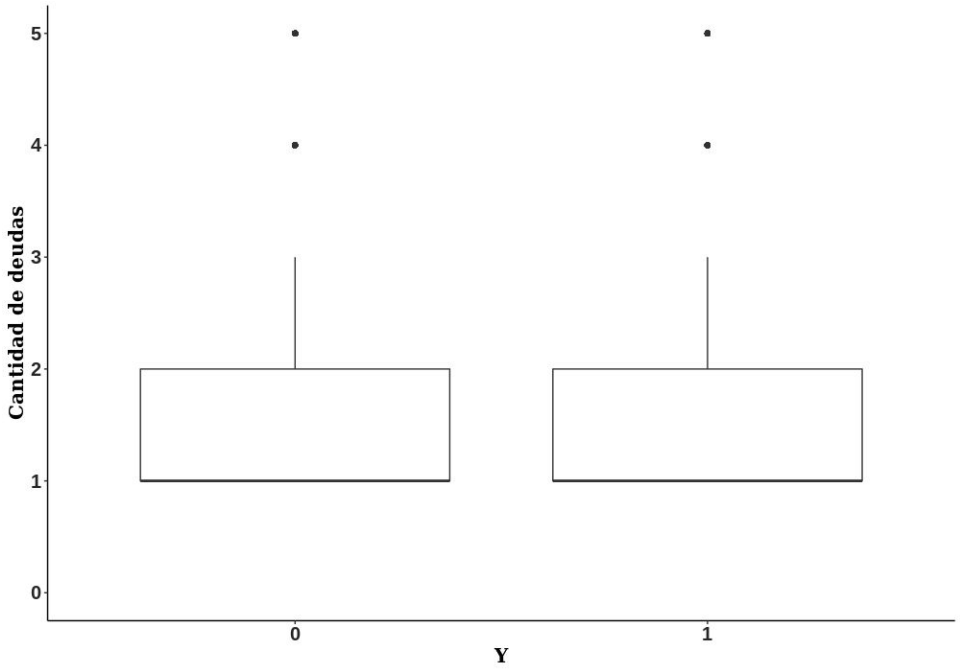
En la Figura 22 se puede ver cómo se diferencian ambos grupos con relación a la cantidad de deudas en las entidades grandes. La mediana para la gente que efectivamente se convertirá en morosa es mayor que para los que no lo serán. Por lo tanto la variable cantidad de deudas parece ser una buena variable para explicar la variable a predecir.

Figura 22: Mediana de la cantidad de deudas según Y en entidades grandes



Por otro lado, como se observa en la Figura 23, no pareciera comportarse de la misma manera en las entidades chicas, donde ambos gráficos de caja son idénticos.

Figura 23: Mediana de la cantidad de deudas según Y en entidades chicas



Otra variable interesante para estudiar es la la cantidad total de deuda de una persona. La Figura 24 es la función de distribución empírica del monto total teniendo en cuenta sólo las entidades grandes; mientras que la Figura 25, las chicas.

Figura 24: Distribución empírica del monto total en entidades grandes

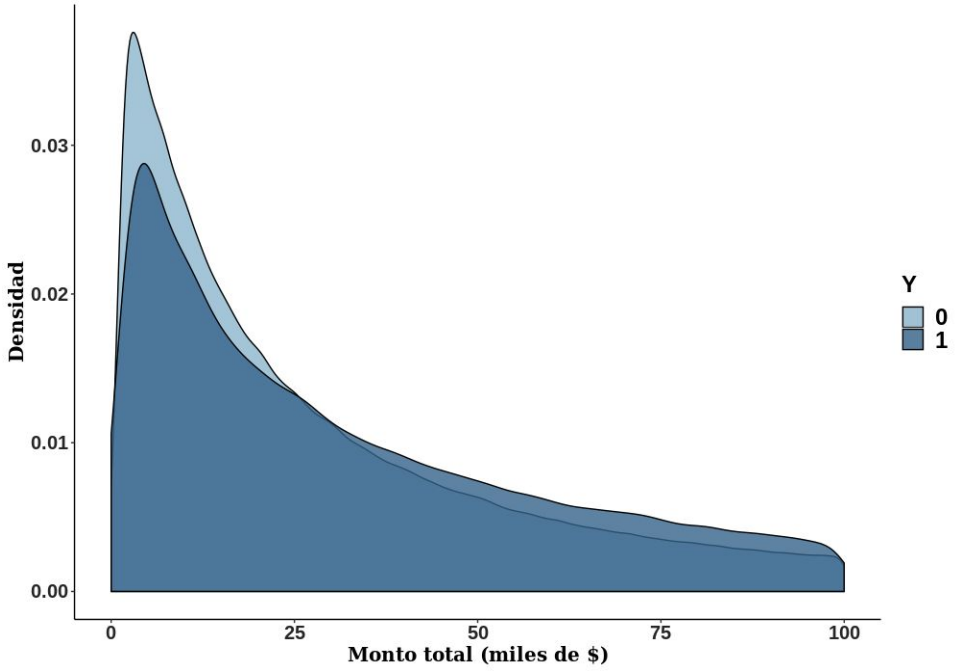
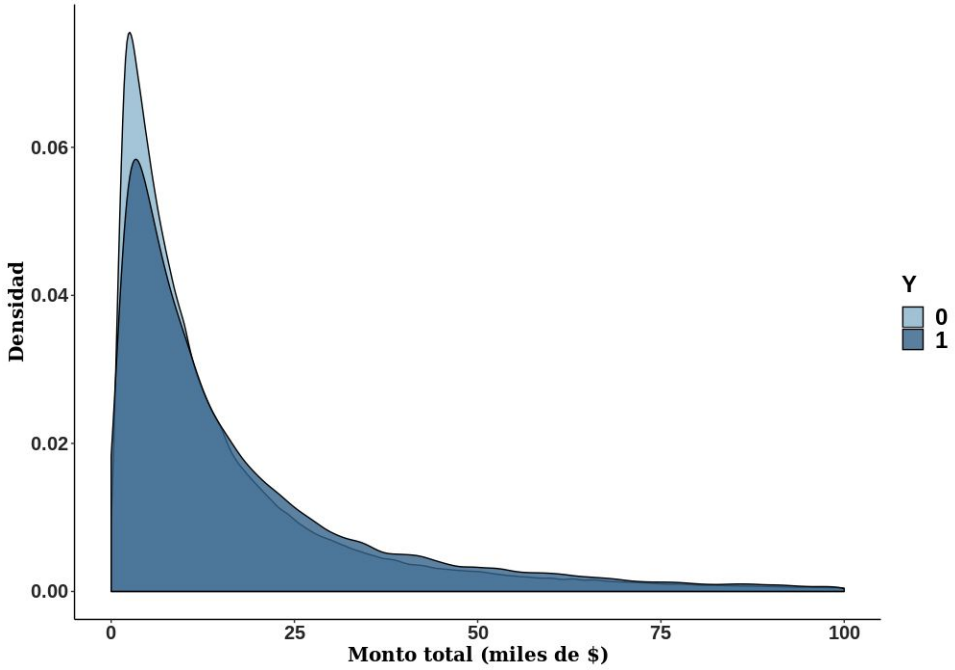


Figura 25: Distribución del monto total en entidades chicas



En ambos casos el comportamiento es similar: las personas que efectivamente se convertirán en morosas están, en promedio, más endeudadas que personas que no lo serán. La diferencia entre ambos tipos de entidades es la densidad de personas que hay en los distintos montos. Las entidades chicas, en comparación con las grandes, tienen mayor densidad de personas en montos relativamente bajos, y menos en montos altos, consistente con lo mencionado anteriormente.

A modo de resumen de lo explicado en esta sección y como puntos importantes para el armado de los modelos se pueden enumerar los siguientes conceptos:

- Hay evidencia estadística significativa para afirmar que la cantidad de deudas activas de los hombres es distinta a la de las mujeres.
- Extranjeros presentan valores más bajos en la cantidad de deudas activas.
- Las personas que se convertirán en morosas en el próximo período tienen una mediana mayor en la cantidad de deudas activas en entidades grandes.
- En ambos tipos de entidades la distribución de las personas que se convertirán en morosas presenta mayor individuos en valores altos de monto total adeudado, y menores en valores bajos, con respecto a los que no se convertirán en morosos.

4 - Resultados

4.1 - Performance

En esta sección se expondrán los resultados de los 6 modelos explicados en la Sección 2.2.1 - Modelos. Se dará detalles acerca de los hiperparámetros elegidos y se mostrará el efecto que genera la inclusión de variables de tendencia.

4.1.1 - Modelos sin tendencias

Se construyeron los siguientes 3 modelos:

1. Predecir si un deudor se convertirá en moroso en alguna entidad el mes siguiente. (y_total)
2. Predecir si un deudor se convertirá en moroso puntualmente en entidades grandes el mes siguiente. (y_grandes)
3. Predecir si un deudor se convertirá en moroso puntualmente en entidades chicas el mes siguiente. (y_chicas)

De las 22 combinaciones de hiperparámetros que se probaron para cada modelo, las 5 que más se destacaron son:

Tabla 14: Modelos para predecir y_total

Modelo	nrounds	max_depth	eta	gamma	colsample_bytree	min_child_weight	subsample	AUC entrenamiento	AUC validación
1	254	12	0,27	16,03	0,96	1,02	0,82	0,80	0,7942
2	277	13	0,26	18,35	0,91	1,14	0,93	0,80	0,7941
3	330	11	0,27	16,79	0,86	1,06	0,77	0,80	0,7941
4	269	11	0,27	19,24	0,90	1,16	0,77	0,80	0,7935
5	257	10	0,29	17,56	0,91	1,14	0,94	0,80	0,7935

Tabla 15: Modelos para predecir y_grandes

Modelo	nrounds	max_depth	eta	gamma	colsample_bytree	min_child_weight	subsample	AUC entrenamiento	AUC validación
1	277	13	0,26	18,35	0,92	1,14	0,93	0,79	0,7880
2	254	12	0,27	16,03	0,96	1,02	0,82	0,79	0,7879
3	330	11	0,27	16,79	0,86	1,06	0,76	0,79	0,7878
4	257	10	0,29	17,56	0,91	1,14	0,94	0,79	0,7877
5	298	11	0,28	18,23	0,99	1,12	0,89	0,79	0,7874

Tabla 16: Modelos para predecir y_chicas

Modelo	nrounds	max_depth	eta	gamma	colsample_bytree	min_child_weight	subsample	AUC entrenamiento	AUC validación
1	256	11	0,26	19,24	0,98	1,20	0,97	0,83	0,7702
2	254	12	0,27	16,03	0,96	1,02	0,82	0,84	0,7698
3	298	11	0,28	18,23	0,99	1,12	0,89	0,83	0,7698
4	330	11	0,27	16,79	0,86	1,06	0,76	0,83	0,7695
5	277	13	0,26	18,35	0,92	1,14	0,93	0,83	0,7693

4.1.2 - Modelos con tendencias

Se construyeron los siguientes 3 modelos:

1. Predecir si un deudor se convertirá en moroso en alguna entidad el mes siguiente, utilizando en el modelo variables de tendencia. (y_total)
2. Predecir si un deudor se convertirá en moroso puntualmente en entidades grandes el mes siguiente, utilizando en el modelo variables de tendencia. (y_grandes)
3. Predecir si un deudor se convertirá en moroso puntualmente en entidades chicas el mes siguiente, utilizando en el modelo variables de tendencia. (y_chicas)

De las 22 combinaciones de hiperparámetros que se probaron para cada modelo, las 5 que más se destacaron son:

Tabla 17: Modelos para predecir y_total utilizando tendencias

Modelo	nrounds	max_depth	eta	gamma	colsample_bytree	min_child_weight	subsample	AUC entrenamiento	AUC validación
1	330	11	0,27	16,79	0,86	1,06	0,76	0,84	0,8252
2	277	13	0,26	18,35	0,92	1,14	0,93	0,84	0,8251
3	298	11	0,28	18,23	0,99	1,12	0,89	0,84	0,8250
4	254	12	0,27	16,03	0,96	1,02	0,82	0,84	0,8250
5	256	11	0,26	19,24	0,98	1,20	0,97	0,84	0,8245

Tabla 18: Modelos para predecir y_grandes utilizando tendencias

Modelo	nrounds	max_depth	eta	gamma	colsample_bytree	min_child_weight	subsample	AUC entrenamiento	AUC validación
1	330	11	0,27	16,79	0,86	1,06	0,76	0,83	0,8207
2	298	11	0,28	18,23	0,99	1,12	0,89	0,83	0,8204
3	277	13	0,26	18,35	0,92	1,14	0,93	0,84	0,8203
4	269	11	0,27	19,24	0,90	1,16	0,76	0,83	0,8201
5	254	12	0,27	16,03	0,96	1,02	0,82	0,84	0,8199

Tabla 19: Modelos para predecir y_chicas utilizando tendencias

Modelo	nrounds	max_depth	eta	gamma	colsample_bytree	min_child_weight	subsample	AUC entrenamiento	AUC validación
1	277	13	0,26	18,35	0,92	1,14	0,93	0,89	0,8243
2	298	11	0,28	18,23	0,99	1,12	0,89	0,89	0,8242
3	254	12	0,27	16,03	0,96	1,02	0,82	0,90	0,8233
4	256	11	0,26	19,24	0,98	1,20	0,97	0,88	0,8232
5	257	10	0,29	17,56	0,91	1,15	0,94	0,88	0,8227

Es interesante remarcar cómo impacta en los resultados el incluir variables de tendencia. Todos los modelos construidos mejoran cuando estas variables son utilizadas. Más aún, esta comparación es cierta, no sólo para la mejor combinación de hiperparámetros encontrada, sino también para todas las otras 4.

Para este trabajo se utilizó únicamente la variable 'suma_monto_total' (tanto para entidades grandes como para chicas). Sin embargo, muy probablemente, estos resultados sean incluso mejores al agregar otras variables, como por ejemplo 'suma_monto_mayor_1'. Tal como se mencionó, debido a cuestiones de capacidad y poder de cómputo, no fue posible probar esto último.

Otro punto interesante a remarcar es el hecho de que los modelos con tendencia parecieran aprender más ruido de los datos de entrenamiento (*overfitting*) que aquellos sin tendencia. Lo mismo sucede con los dos modelos que predicen y_chicas con respecto al resto. Esto significa que la diferencia entre la *performance* de los datos de entrenamiento y de validación es mayor para los modelos con tendencia y aquellos que predicen 'y_chicas'. De todas formas, es importante ver qué sucede con los datos de testeo, dado que también se puede estar sobreajustando los datos de validación.

4.1.3 - Performance en testeo

Habiendo seleccionado los modelos finales, se los evaluó en los datos de testeo y así se calculó su performance final:

Tabla 20: Performance de los modelos en datos de testeo

Modelo	Performance en entrenamiento	Performance en testeo	Performance en testeo del *benchmark*
Modelo para y_total sin tendencias	0,80	0,78	0,71
Modelo para y_grandes sin tendencias	0,79	0,78	0,70
Modelo para y_chicas sin tendencias	0,83	0,74	0,66
Modelo para y_total con tendencias	0,84	0,81	0,73
Modelo para y_grandes con tendencias	0,83	0,81	0,73
Modelo para y_chicas con tendencias	0,89	0,80	0,68

Efectivamente los modelos con tendencias predicen mejor que aquellos que no las tienen incluidas. También realizan apenas un poco más de *overfitting*, aunque de forma muy sutil, si no se tienen en cuenta los modelos que predicen y_chicas. En estos últimos es sorprendentemente distinto cómo se comportan en entrenamiento comparado con como lo hacen en testeo. Por último, es importante mencionar que los modelos construidos son sustancialmente más eficientes a la hora de predecir que los creados como *benchmark* en la Sección 2.2.5 Modelos *Benchmark*.

4.2 - Interpretación

Los modelos generados por algoritmos tan potentes como los utilizados en este trabajo tienen la ventaja de ser muy eficaces a la hora de predecir, pero suelen ser una 'caja negra' en

sus resultados, en qué variables se centran y cómo son utilizadas. Tradicionalmente, para mostrar qué variables son las más importantes, se computa cuánto disminuye el error debido a cada una de éstas, y se seleccionan las n variables que mayor lograron reducirlo. Sin embargo, una forma reciente y más moderna de calcular la importancia de las variables es a través del método *'SHapley Additive exPlanations'* (SHAP), el cual calcula el aporte de cada variable en la predicción (Lundberg y Lee, 2017). Más aún, dice cómo se distribuye esta contribución dentro de los valores de cada variable. Los valores SHAP (ϕ_{ij}) son calculados para cada observación i de cada variable j del dataset de entrenamiento. De esta forma, calculando $\sum_i \left|\phi_{ij}\right|$ se obtiene la contribución total de la variable j .

A continuación se muestran las variables que resultaron ser más importantes para cada modelo, junto con sus valores SHAP. Para el cálculo de estos valores se utilizaron los modelos con mejor *performance* en el conjunto de datos de validación.

Figura 26: Valores SHAP y_total sin tendencias

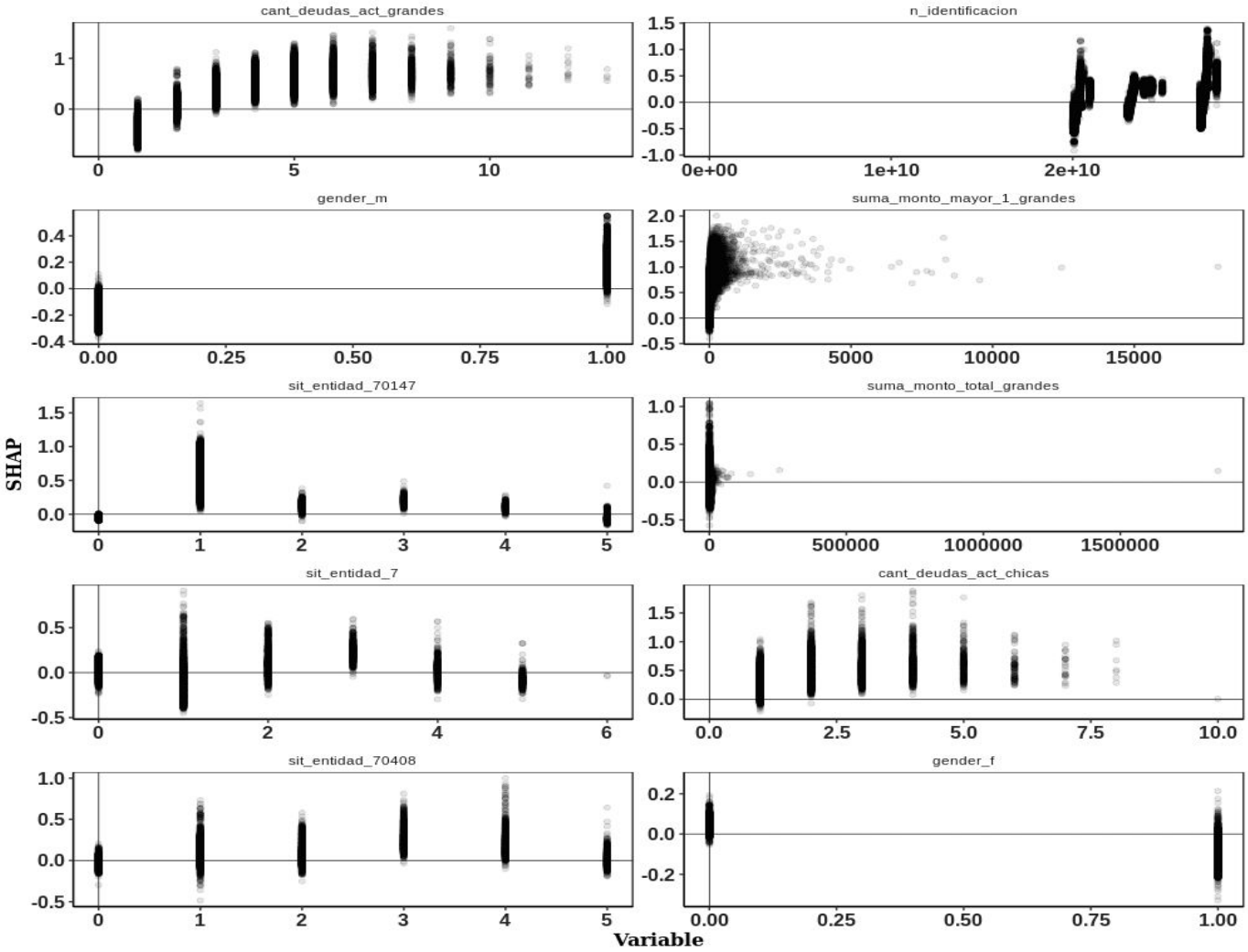


Figura 27: Valores SHAP y_total con tendencias

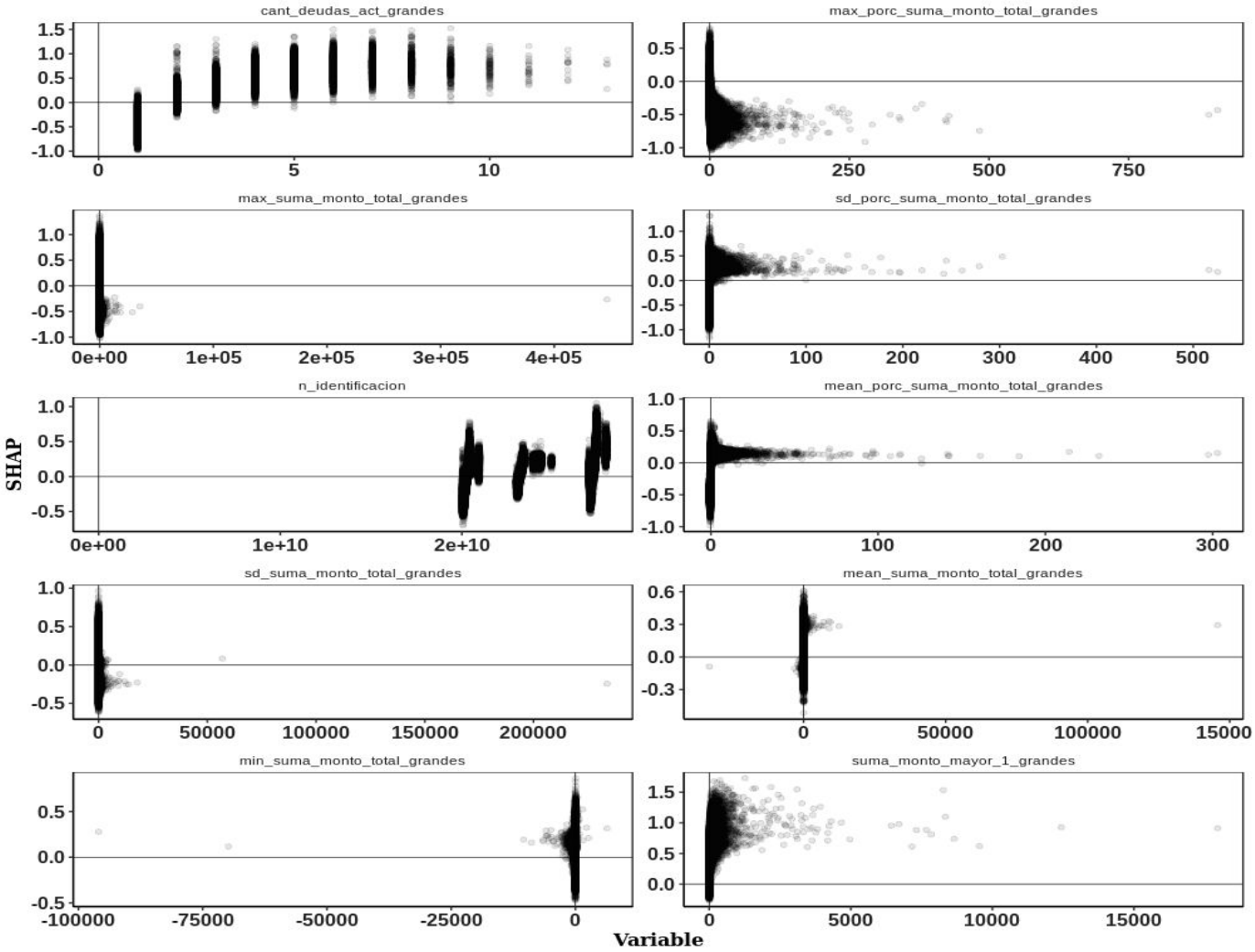


Figura 28: Valores SHAP y_grandes sin tendencias

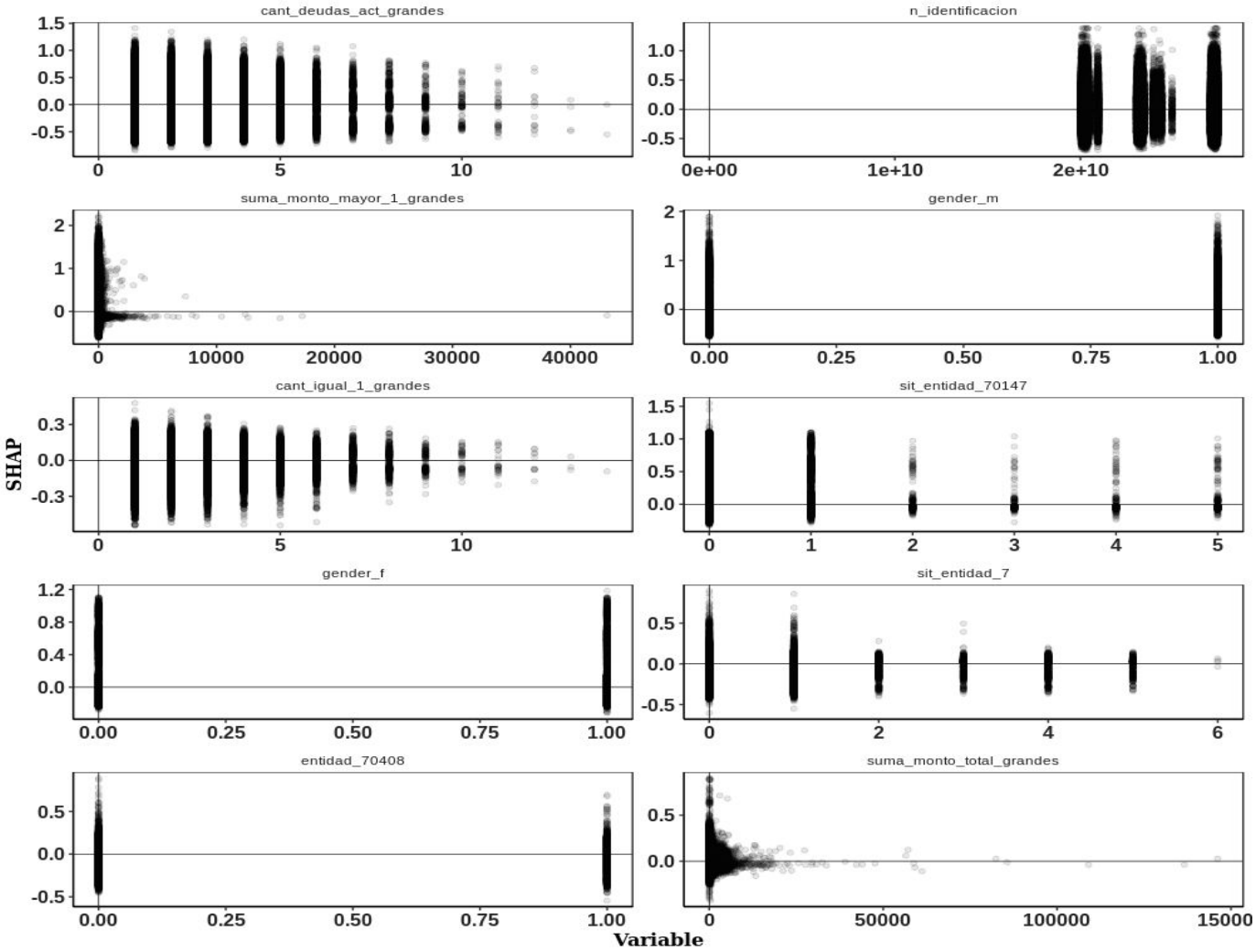


Figura 29: Valores SHAP y_grandes con tendencias

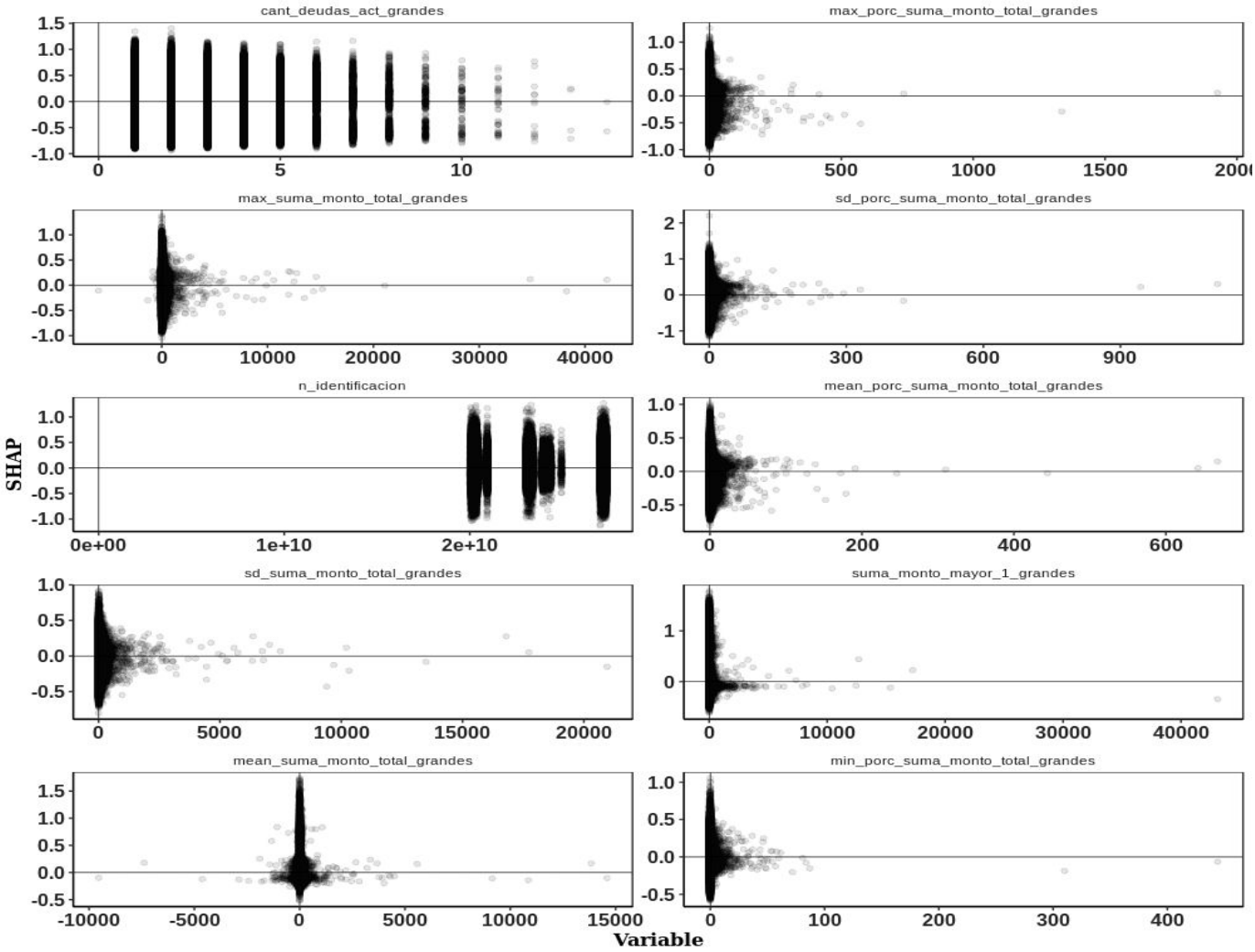


Figura 30: Valores SHAP y_chicas sin tendencias

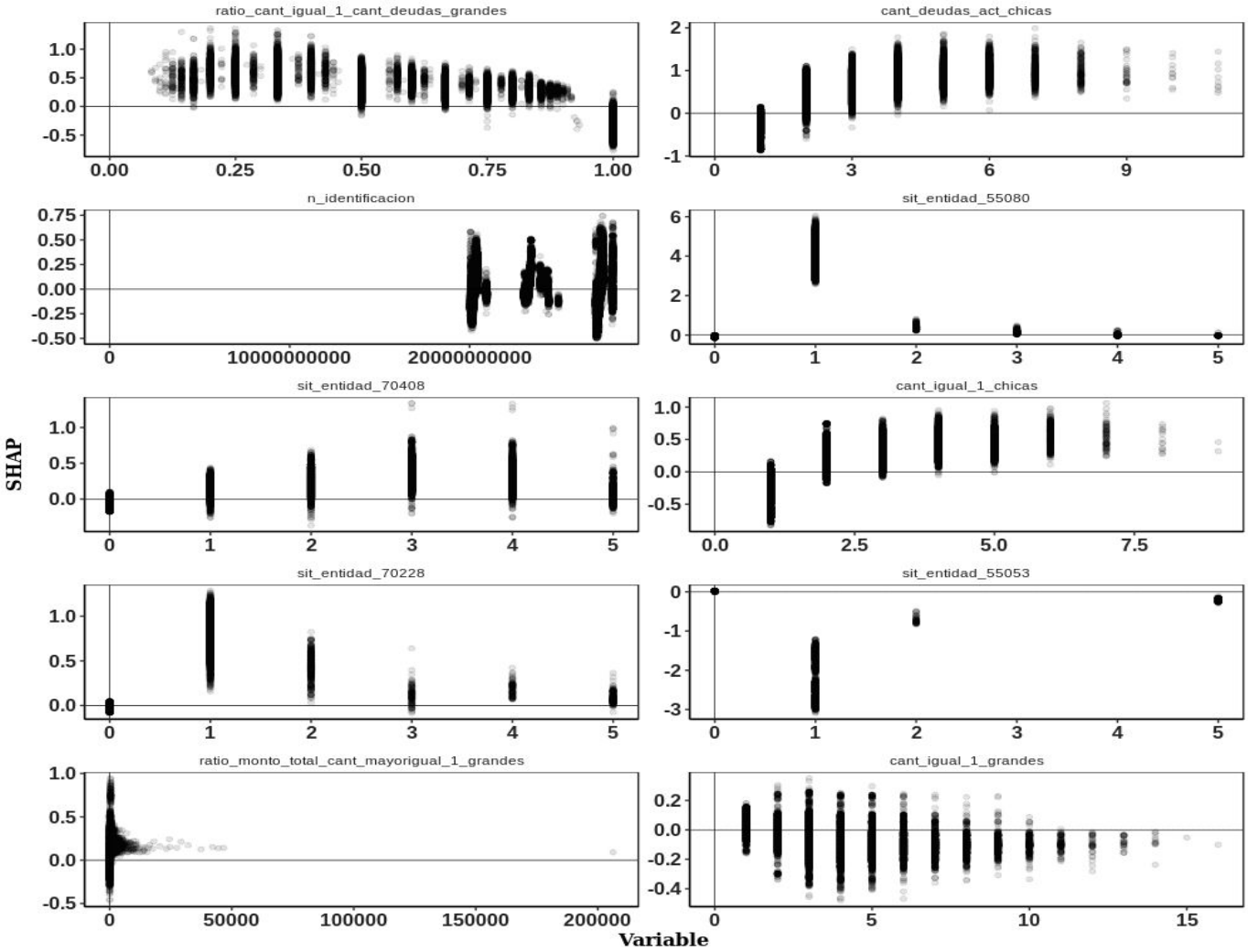
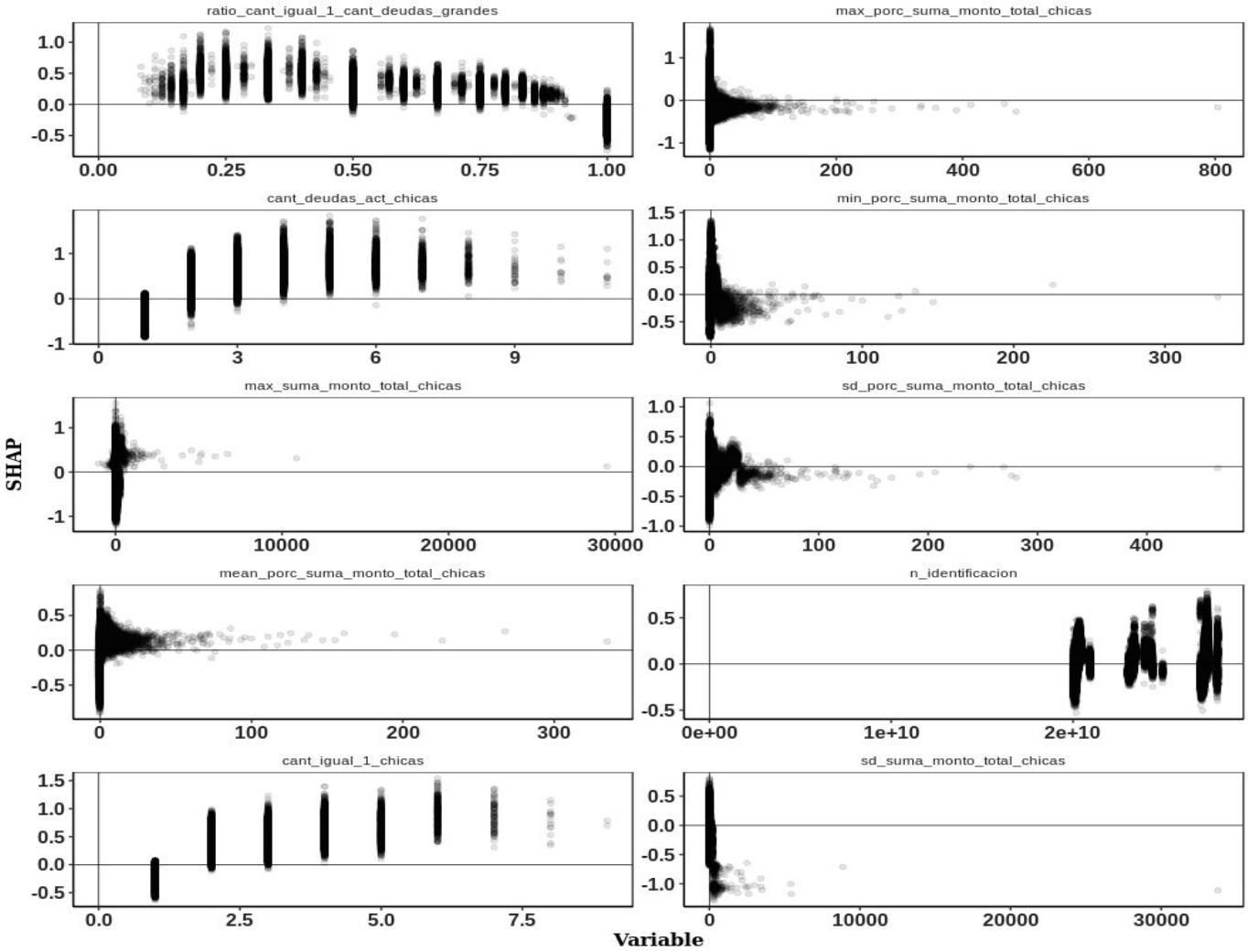


Figura 31: Valores SHAP y_chicas con tendencias



De las figuras anteriores se desprenden algunas conclusiones en línea con lo esperado y señalado anteriormente en las Secciones 3.1 - Análisis del mercado y 3.2 - Análisis de los individuos. Por un lado, como se mostró anteriormente en los resultados de cada modelo, las variables de tendencia importan mucho. En aquellos donde fueron incluidas pasaron a ser parte, en su mayoría, de las variables más importantes. Muchas de ellas se comportaron como era esperado: valores mayores tienden a 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, la variable 'max_porc_suma_monto_total', para los

3 modelos, a mayores valores, menor es la probabilidad que el modelo asigna a que un individuo se convierta en moroso.

Por otro lado, hay una gran relación entre lo que el modelo está prediciendo y el tipo de variables que utiliza. Es decir, en modelos donde se busca predecir la morosidad en entidades chicas, se utilizan, en su mayoría, las variables de este grupo de entidades. Aunque hay excepciones, y es lógico que así sea, donde se utilizan variables del otro grupo. Es razonable que esto sea así dado que una persona podría tener deudas en dos tipos de entidades distintas y ambas estar influyendo en las condiciones de pago de un mismo individuo.

También vale la pena mencionar que aparecen variables que intuitivamente tienen sentido; y de hecho se comportan de la forma esperada. Ejemplos de esto son: cant_deudas_act_grandes, suma_monto_mayor_1_grandes, cant_igual_1_chicas, entre otras. 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 principalmente para que trate de captar algún patrón relacionado a la edad de los individuos. Esa información no está disponible explícitamente en los datos, pero sí tiene una alta correlación con esa secuencia de números. De hecho, se puede observar cómo el algoritmo identifica los 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).[14] En muchos de los modelos esta variable pareciera comportarse de manera parecida dentro de cada grupo: primero toma valores altos y luego, de forma irregular, tiene a disminuir. Esto es un indicador de que, por un lado, logró efectivamente detectar los 3 grupos; y más importante aún, logró captar la edad. Según lo que se observa en estos gráficos, mayor edad contribuye a tener menor probabilidad de volverse moroso.

Por último, hay entidades donde la situación de morosidad pareciera ser muy importante. Estas son:

[14] De hecho la variable 'gender' fue construida a partir de estos números, utilizando sólo a los individuos cuya variable 'n_identificacion' empieza con 20 ó 27. Se asignaron NA en caso contrario. Observar que la variable 'gender' también resulta relevante en ciertos modelos.

Tabla 21: Entidades más importantes según variable a predecir

Modelo	Variable	Entidad
y_total	sit_entidad_70147	Cencosud S.A[15].
	sit_entidad_70408	Tarjeta Naranja S.A.[16]
	sit_entidad_7	Banco De Galicia Y Buenos Aires S.A.U.[17]
y_grandes	sit_entidad_70147	Cencosud S.A.
	sit_entidad_70408	Tarjeta Naranja S.A.
	sit_entidad_7	Banco De Galicia Y Buenos Aires S.A.U.
y_chicas	sit_entidad_70408	Tarjeta Naranja S.A.
	sit_entidad_55080	DAP Cooperativa de Crédito y Consumo Ltda.[18]
	sit_entidad_70228	Santa Mónica S.A.[19]
	sit_entidad_55053	Casa Luis Chemes S.A.[20]

[15] http://www.tarjetacencosud.com.ar/
[16] https://www.naranja.com/
[17] https://www.bancogalicia.com/banca/online/web/Personas
[18] https://bit.ly/2AEXNXl
[19] https://www.dateas.com/es/explore/empresas-emisoras-tarjetas-credito/santa-mnica-sa-69
[20] https://www.chemesweb.com.ar/

5 - Conclusiones

5.1 - Limitaciones y futuras posibles mejoras

Un gran desafío para la resolución de este trabajo fue, en parte, el manejo de una gran base de datos. Poder procesar tanta información y realizar cómputos para la creación de variables no fue una etapa trivial. De hecho, como se mencionó anteriormente, por cuestiones de espacio, no se han podido incorporar variables que muy probablemente sean útiles a la hora de predecir. En versiones futuras de este trabajo, o en caso de buscar perfeccionarlo, se podría, en el mejor de los casos, utilizar computadoras con mayor memoria. En caso de no ser posible, otra alternativa es adaptar el modelo y buscar qué variable de tendencia es la más importante para predecir. Una última alternativa es estudiar con mayor detenimiento los ratios creados, los cuales demostraron no ser muy utilizados por los algoritmos. Así, entonces, se podrían mantener sólo aquellos que sean relevantes. De esta forma, quizá, haya espacio suficiente para incorporar otras variables.

En línea con estudiar nuevas variantes, también se podrían probar otros algoritmos de clasificación. En este trabajo se eligió utilizar XGBoost, principalmente por su gran poder predictivo. Sin embargo, otros tipos también pueden ser tenidos en cuenta, como por ejemplo, Random Forest o Support Vector Machines, que presentan un nivel de complejidad similar al aquí utilizado. En caso de llevarse a la práctica, y tener los recursos suficientes, se podría incluso generar un algoritmo más complejo basado en redes neuronales.

Una tendencia que estuvo surgiendo en el último tiempo, enfocada en los algoritmos que deciden sobre cuestiones humanas, es tratar de evitar que ellos aprendan los sesgos que tienen incorporadas las sociedades. Bajo esta línea de pensamiento, se considera que existen algoritmos sexistas o racistas que básicamente se comportan de esta forma dado que los datos que fueron utilizados para ser entrenados ya tenían estos sesgos incorporados (Zou, Schiebinger, 2018). Una forma para minimizar esto puede ser a través de 'variables reservadas'. Es decir, variables que se dejan de lado a la hora de construir el modelo, ya que de tener poder predictivo, sería básicamente debido a un sesgo social (Bellamy et. al, 2018). En estos datos,

existen variables como 'gender', 'n_identificacion' y 'extranjero' que podrían ser candidatas a ser variables de este estilo. Sin embargo, en este trabajo, no se siguió este enfoque, dado que, debido a la estructura de los datos y de las variables, no es posible resolver el problema omitiendo dichas variables y construyendo el modelo sobre las demás. Por ejemplo, pueden existir 2 variables que no sean reservadas y sin embargo estén correlacionadas entre ellas y con una que sí (como el ser o no extranjero). De esta forma, se eliminaría la variable reservada pero no la correlación entre las otras dos variables y por ende no es cierto decir que el modelo no estaría aprendiendo sesgos. En caso de buscar construir un modelo con un objetivo similar al de este trabajo, pero que no tenga sesgos en sus datos de entrenamiento, habría que hacer un análisis más exhaustivos de los mismos para no caer en estos problemas.

5.2 - Aplicaciones prácticas

El punto central de este trabajo fue poder brindar una herramienta altamente competitiva que le permita a una entidad reducir su mora, alocar mejor sus recursos, reducir sus costos y proyectar mejor sus ingresos y *cashflows*. A su vez, se tenía como objetivo lograr todo esto con datos enteramente públicos para que sea replicable en cualquier entidad.

Pensando en esto último, se construyeron diversos modelos enfocados tanto para entidades grandes, como chicas. Sin embargo, el aporte principal de esta tesis es sobre estas últimas. Como se explicó a lo largo de este trabajo, las entidades intermedias tienen motivos relevantes para construir una herramienta eficiente que les permita aumentar sus beneficios, aprovechando el uso de datos públicos de fácil acceso. Teniendo en cuenta el tipo de problema y el nivel de detalle de los datos, todos los modelos resultaron tener una *performance* más que aceptable. En caso de implementarse, se podrían sumar datos privados de la entidad en cuestión y así mejorar incluso más el comportamiento del modelo.

Sin dudas que reducir la morosidad impacta enormemente en los beneficios de una firma. Sin embargo, optimizar los recursos internos de la empresa puede tener también un impacto sustancial. Esta herramienta debe implementarse para especificar en quiénes hay que enfocar tiempo y dinero, y en quiénes no. De esta forma, no sólo subirían los ingresos, sino también bajarían los costos operativos.

Por último, es una herramienta útil para los departamentos de finanzas, principalmente para los encargados de proyección de flujos, dado que permite estimar cuáles serán los ingresos el próximo mes. Es cierto que es una ventana temporal muy reducida, pero en países como la Argentina, con tanta volatilidad e incertidumbre, puede ser de gran utilidad.

5.3 - Conclusión

En función de los resultados obtenidos en trabajos similares se puede afirmar que los de este trabajo son más que favorables. Por ejemplo, en estudios de predicción de morosidad de tarjetas de crédito, donde sí se incluye información personal de cada individuo (como la edad exacta, su estado civil, nivel de educación, etc.), se ha obtenido una *performance* de AUC similar e incluso menor que en la de este trabajo. En modelos logísticos se obtuvo un área bajo la curva de 0,72; con redes neuronales, una de 0,77; con *support vector machines* (SVM), 0,72; con XGBoost 0,78; entre otros (Yang, Zhang, 2018).

Como conclusión, este trabajo es un ejemplo de cómo utilizando datos enteramente públicos se puede hacer un análisis exploratorio minucioso del cual sacar información, construir modelos competitivos de machine learning, y lo más importante: generar herramientas rigurosas y aplicables que generen valor para una compañía.

6 - Bibliografía

- Anastasi A, Blanco E, Elosegui P, Sangiácomo P. 2010. *La bancarización y los determinantes de la disponibilidad de servicios bancarios en Argentina.* Ensayos Económicos, 60 (Octubre - Diciembre). Banco Central de la República Argentina.

- Banco Central de la República Argentina. 2020. *Comunicación "A"6.909.* *http://www.bcra.gov.ar/Pdfs/comytexord/A6909.pdf*

- Barbería, M. 2019. *El BCRA hace que los bancos vendan dólares para tratar de estabilizar al mercado*. El Cronista. *https://www.cronista.com/finanzasmercados/El-BCRA-hace-que-los-bancos-vendan-dolares-para-tratar-de-estabilizar-al-mercado-20190815-0025.html*

- Bellamy R, Dey K, Hind M, Hoffman S, Houde S, Kannan K, Lohia P, Martino J, Mehta S, Mojsilovic A, Nagar S, Ramamurthy K, Richards J, Saha D, Sattigeri P, Singh M, Varshney K, Zhang Y. 2018. *AI Fairness 360: An Extensible Toolkit for Detecting, Understanding, and Mitigating Unwanted Algorithmic Bias*. IBM Journal of Research and Development.

- Bengio Y, Bergstra J. 2012. *Random Search for Hyper-Parameter Optimization*. Journal of Machine Learning Research.

- Bravo C, Correa-Bahnsen A, Cortés-Tejada F, Luque, Roa M, Suarez G. 2020. *Super-App Behavioral Patterns in Credit Risk Models: Financial, Statistical and Regulatory Implications*. arXiv. *https://arxiv.org/abs/2005.14658*

- Chen Y, Ding S, Li W, Yang S. 2018. *Heterogeneous Ensemble for Default Prediction of Peer-to-Peer Lending in China*. IEEE Access.

- Chen T, Guestrin C. 2016. *Xgboost: A scalable tree boosting system*. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Association for Computing Machinery.

- Davis J, Goadrich M. 2006. *The Relationship Between Precision-Recall and ROC Curves*. Proceedings of the 23rd International Conference on Machine Learning. Association for Computing Machinery.

- James G, Witten D, Hastie T, Tibshirani R. 2017. *An Introduction to Statistical Learning*. 8 edición. Springer.

- Kuhn M, Johnson K. 2016. *Applied Predictive Modeling*. 5 edición. Springer.

- Lundberg S, Lee S. 2017. *A Unified Approach to Interpreting Model Predictions*. NIPS.

- Rikkers F, Thibeault A. 2015. *A Structural form Default Prediction Model for SMEs, Evidence from the Dutch Market*. Multinational Finance Journal.

- Tan P, Steinbac M, Kumar V. 2006. *Introduction to Data Mining*. Pearson.

- Wende P. 2020. *Con una nueva regulación a los créditos, el Banco Central le dio un duro golpe a la industria fintech*. Infobae. https://www.infobae.com/economia/2020/02/20/con-una-nueva-regulacion-a-los-creditos-el-banco-central-le-dio-un-duro-golpe-a-la-industria-fintech/

- Yang S, Zhang H. 2018. *Comparison of Several Data Mining Methods in Credit Card Default Prediction*. Intelligent Information Management Vol.10 No.05.

- Yu X. 2017. *Machine learning application in online lending risk prediction*. arXiv. *https://arxiv.org/abs/1707.04831*

- Zou J, Schiebinger L. 2018. *AI can be sexist and racist — it's time to make it fair*. Nature. https://www.nature.com/articles/d41586-018-05707-8