Universidad Zaragoza
1542

Trabajo Fin de Máster

Como minimizar la tasa de error en la clasificación de los préstamos: el caso *peer to peer lending*

Autor

Víctor Cebollada Mellado

Directora

Begoña Gutiérrez Nieto

Facultad de Economía y Empresa

2015

Resumen

En este trabajo se va a analizar la posibilidad de minimizar la tasa de error en la clasificación de los préstamos sociales, también denominados entre iguales, alternativa online de financiación sin intermediación financiera tradicional y que en los últimos años está obteniendo una relevancia considerable. El procedimiento utilizado consiste en la utilización de varios algoritmos que seleccionan aquellas variables consideradas significativas para minimizar el error en la clasificación dada una muestra de entrenamiento. Los resultados sin embargo son pocos coherentes, ya que muestran que minimizamos el error con una única variable significativa que en la práctica no tendría sentido. El trabajo se estructura de la siguiente manera: en el epígrafe 1 se presenta la literatura sobre los préstamos entre iguales y en la sección 2 se muestra la metodología y los datos utilizados. En la sección 3 se presentan los resultados y en el epígrafe final las conclusiones.

Palabras claves: préstamos entre iguales, préstamos sociales, microfinanzas

Abstract

This work analyzes the possibility of minimizing the error rate in the classification of social lending, also known peer to peer lending, online alternative of financing without traditional financial intermediation, which is getting a considerable relevance in the last few years. The procedure used is based on the use of several algorithms, which selected those variables considered significant to minimize the error rate in the classification given a training sample. However, the results are inconsistent, since they show that we minimize the error rate with only one significant variable that in the practice wouldn´t make sense. The work is structured as follows: section 1 provides a literature review on peer to peer lending and in the section 2 it is shown the methodology and data used. In section 3 it is presented the results and in the finally section the conclusion.

Keywords: peer to peer lending, social lending, microfinance

1. INTRODUCCIÓN

El avance de la tecnología de la información durante los últimos años ha conducido al desarrollo de los mercados electrónicos donde la importancia de los tradicionales intermediarios financieros puede ser menos relevante o incluso redundante para la interacción económica de los participantes del mercado (Benjamin y Wigand 1995; Evans y Wurster 1997). Precisamente esta es la principal propiedad de los prestamos *peer to peer* (en adelante P2P), un nuevo modelo de financiación para los usuarios de internet creado sobre la base de los principios del microcrédito.

Los prestamos P2P se refieren a aquellos sin necesidad de garantía entre prestamistas y prestatarios mediante plataformas online sin la intermediación de ninguna institución financiera (Lin et al. 2009; Collier y Hampshire 2010; Bachman et al. 2011), lo cual ha provocado un punto de crítica ya que la mediación financiera, según Leland y Pyle (1977) existe principalmente para mitigar los problemas de asimetría en la información. Como una revolucionaria aplicación de las tecnologías de la información en los campos financieros (Iyer et al. 2009; Lin et al. 2012) los prestamos P2P permiten facilitar de una forma efectiva la búsqueda de información y ofrecen todas las funciones necesarias para llevar a cabo las transacciones (Brown 2008; Herzenstein et al. 2008), permitiendo el acceso a servicios financieros a determinados segmentos de la sociedad que sufrían cierto grado de exclusión financiera. Las ventajas de este innovador modelo de préstamos incluyen:

1) Se expone que los prestatarios consiguen préstamos con menores tasas de interés y los prestamistas obtienen mayores rentabilidades (Magee 2011). Sin embargo, muchas observaciones empíricas demuestran que no siempre se alcanzan las rentabilidades esperadas, lo que ha conducido a la discusión sobre el futuro de los prestamos P2P (Kauffman y Riggins 2012; Magee 2011; Rose 2007; Wang y Greiner 2011).
2) Prestamistas y prestatarios pueden fácilmente enviar y buscar información, completando las transacciones en las plataformas online con menores costes de transacción (Lin et al. 2009; Lin et al. 2012).

3) Bajos costes de transacción hacen factibles pequeños préstamos, denominados micropréstamos, que pueden ser equipados juntos para financiar proyectos que requieren gran cantidad de financiación mientras se reduce el riesgo del préstamo. (Klafft 2008)

4) Los prestamistas pueden conseguir más información sobre los prestatarios en las redes sociales y mitigar la asimetría de la información que se incrementa al no haber intermediarios tradicionales, contribuyendo a reducir el riesgo y ayudando a expandir el negocio más allá del tradicional círculo conocido. (Freedman y Jin 2008; Lin et al. 2011; Berger y Gleisner 2009; Collier y Hampshire 2010).

Los prestamos online P2P tienen el primordial propósito de realizar prestamos pequeños, los cuales no solamente financian pequeños negocios sino que, como indican Johnson et al. (2010) y Wang et al. (2009), proporcionan liquidez a corto plazo. Incluso las plataformas de préstamos P2P juegan el rol de las instituciones financieras conectando prestamistas y prestatarios, lo cual provoca en muchas ocasiones, beneficios procedentes de comisiones, aun existiendo gran diferencia de concepto entre los depósitos y los préstamos. Debido a su reciente popularidad, los prestamos P2P han recogido la atención de los profesionales de diversas disciplinas y de los medios de comunicación (Brown 2008; Galloway 2009; Lin 2009 y Bachmann et al. 2011).

Desde 2005, año del lanzamiento de la primera plataforma, Zopa, en el Reino Unido los prestamos online P2P han experimentado un rápido crecimiento a través de un amplio número de países, incluyendo Estados Unidos, Canadá, Reino Unido; Japón, Italia y China con algunas diferencias entre ellos. Este auge ha sido favorecido por la recesión económica, conduciendo a estas plataformas a inversionistas desalentados por la banca tradicional. Algunas plataformas tienen objetivos no lucrativos, recogiendo y proporcionando dinero para la población carente de riqueza, mientras que otras tienen fines comerciales, facilitando los prestamos entre prestamistas y prestatarios. Entre las plataformas con más éxito se encuentran la ya mencionada Zopa en el Reino Unido, Prosper, Kiva y Lending Club en Estados Unidos. Por ejemplo, Prosper fundada en 2006, ha financiado préstamos por valor superior a los 4 billones de dólares y Kiva, con carácter no lucrativo, tiene registrados más de 705 millones en préstamos.

Gartner predijo en 2008 que en los próximos años tales plataformas podrían alcanzar una cuota de mercado del 10% del mercado mundial en lo que respecta a los préstamos al por menor y la planificación financiera y en el 2010 vaticinó que en 2013 la industria superaría los 5 billones de dólares. Según Slavin (2007), se espera que dentro de unos años los intercambios de préstamos P2P se conviertan en una plataforma alternativa para el ahorro tradicional y los inversionistas. En los últimos años, los prestamos sociales online han crecido desde una pequeña singularidad a una industria que no solamente abarca pequeños préstamos personales sino que ha empezado a hacer incursiones en los préstamos inmobiliarios y de pequeños negocios. No es difícil imaginarse un futuro donde los prestamos sociales online sean una fuente común de financiación para empezar nuevos negocios. Es un instrumento que tiene el potencial para, de una forma eficiente, reunir y formar fondos de un ilimitado número de pequeños inversores virtuales, por lo que el impacto futuro de una amplia gama de mercados financieros podría ser significativo. Harvard Business Review[1] argumentó que cada banco principal tendrá su propio sistema de préstamos P2P en los próximos años y que esta forma de recibir financiación será considerada una de las innovaciones de servicios financieros más importantes de la última década. De hecho desde 2014, el 80% de los préstamos publicados fueron financiados por grandes empresas, las cuáles las plataformas de préstamos P2P intentaban evitar y la tendencia continúa debido a la creciente demanda y a la insuficiente oferta.

A pesar de su rápido crecimiento, los mercados de préstamos P2P se encuentran aún en etapas débiles ya que solamente unas pocas están capacitadas para alcanzar la eficiencia operacional y la supervivencia ante una posible competencia. Lin (2009); Coller y Hampshire (2010) argumentan que la asimetría en la información y la desconfianza por determinados comportamientos oportunistas de los prestatarios pueden causar ineficiencias tanto a prestamistas como a prestatarios. Las investigaciones muestran que no solamente los factores técnicos pueden afectar al comportamiento de los prestamistas y prestatarios, ya que los factores psicológicos también pueden influir. Así que la realización de un examen del comportamiento tanto de prestamistas como de

[1] Ver Sviokla, J. Breakthrough Ideas: Forget Citibank - Borrow form Bob. *Harvard Business Review*. 2009, Feb.

prestatarios y la identificación de los préstamos exitosos anteriores se considera significativa para el buen desarrollo de las plataformas de préstamos P2P.

Al contrario de lo que ocurre en los bancos considerados tradicionales, en estas nuevas plataformas de préstamos P2P es difícil que los prestamistas dispongan de información comprensiva sobre los prestatarios, causando, tal como se indica en Lin et al. (2009) importantes problemas de asimetría de la información. Por lo que la mayoría de estudios realizados, como los de Klafft (2008), se centran en mitigar esta asimetría con el objetivo de reducir riesgos, ya que la mayoría de los prestamistas no tienen experiencia.

Los prestatarios proporcionan información crediticia a la plataforma que puede ser cuantificada con exactitud y ser transmitida de forma eficiente, denominada *"hard credit information"*. Este tipo de información incluye aspectos como el perfil de crédito, el ratio deuda/ingreso, el número de solicitudes de préstamo en el pasado y el número de tarjetas de crédito del que dispone el prestatario. En una transacción online, el comprador normalmente no puede acceder a la información que requiere con gran detalle y tiene que juzgar la honradez del vendedor por algunos comportamientos o señales (Bacharach y Gambetti 2001). Algo similar ocurre en las plataformas de préstamos P2P: Collier y Hampshire (2010) han revelado que juegan un papel esencial en la decisión de prestar el coste de obtener señales y la dificultad de evaluar las señales. Por ello, la información personal de los prestatarios que se incluye en los *listing*[2] publicados por ellos mismos, son señales de gran importancia para considerar la honradez de los prestatarios, así como el riesgo de impago y la fijación de los tipos de interés (Collier y Hampshire 2010; Lin 2009).

Todas las plataformas sociales confían las decisiones de inversión en el denominado "*credit score*[3]" proporcionado por agencias como Experian TransUnion LLC, Equifax Inc y Schufa Holding AG. Sin embargo trabajos como el de Zachakis y Meyer (2000) muestran grandes tasas de error para los modelos estadísticos, basados en

[2] Ver Everett, C. R. Group membership, relationship banking and loan default risk: the case of online social lending, 4-5. March 2010

[3] Ver Rayo Cantón, S; Lara Rubio, J; y Camino Blasco, D. A credit Scoring Model for Institutions of Microfinance under the Basel II Normative. *Journal of Economics, Finance and Administrative.* Vol 15 nº 28, jun 2010.

el "*credit score*". Iyer et al. (2014) encuentra que los modelos econométricos superan a los mercados en términos de habilidad predictiva, existiendo evidencias empíricas que muestran que los prestatarios con las puntuaciones más altas no son necesariamente los mejores, lo cual indica que las medidas tradicionales financieras no están bien equipadas para capturar la dinámica no convencional prevaleciente en los préstamos sociales.

Ante tal situación, es necesario aportar información subjetiva por parte de los prestatarios, confusa y difícilmente cuantificable conocida como "*soft credit information*". Este tipo de información puede ser conseguida a través de las redes sociales, tal como indican Collier y Hampshire (2010); Iyer et al. (2009); Kattherine y Sergio (2009) y López (2009), ya que la teoría de las microfinanzas sugiere que las redes sociales pueden ayudar a reducir la asimetría en la información, compensando la ausencia de información "*hard*" y de ahí motivar a los prestatarios a devolver los préstamos que se les ha concedido (Katherine y Sergio 2009). Greiner y Wang (2009) encuentran que a mayor capital social por parte de los prestatarios, mayor posibilidad de ser financiados y menores tasas de interés tienen que abonar. Estos resultados son congruentes con los de Lin et al. (2009), en donde la utilización de las redes sociales conlleva mejores resultados, incluyendo mayor probabilidad de que el préstamo sea financiado, menores tasas de impago y menores tasas de interés. De modo que el uso de las redes sociales puede llegar a mitigar la selección adversa. Collier y Hampshire (2010) investigaron el impacto de las redes sociales en el comportamiento a la hora de prestar basándose en la teoría de las señales y encontraron que la frecuencia en la comunicación con otros miembros de las plataformas de préstamos P2P a la hora de realizar transacciones podría ser utilizado por parte de los prestamistas como una señal fuerte de la honradez de los prestamistas. Estos resultados muestran que los prestamistas combinan tanto información objetiva como subjetiva disponible para evaluar el grado de confianza de los potenciales prestatarios.

Otra opción para reducir la asimetría en la información en estas plataformas, es la de ser miembro de lo que se denominan "*trusted group*" o grupos de confianza en los prestamos P2P, con características muy similares a la metodología utilizada por las instituciones del microcrédito para proporcionar servicios microfinancieros denominada grupos solidarios, que proporciona ventajas de economía de escala y externalidades positivas, ya que los costes de otorgamiento y administración de los préstamos se reducen

al conceder un único préstamo a varios solicitantes integrados en el grupo y el coste de la morosidad se reduce como consecuencia de la presión ejercida por el grupo sobre cada miembro para que pague, así como la garantía mutua otorgada por los miembros del grupo antre el incumplimiento de alguno (Gutiérrez, 2009). Estos "*trusted group*" los cuáles, si se crean bajo unos incentivos correctos, mejoran las tasas de éxito a la hora de financiar a los prestatarios conjuntamente y reducen las tasas de interés que tienen que soportar los prestatarios con bajos "*credit score"* eliminando el riesgo moral, ya que existe una responsabilidad conjunta, evaluando y supervisando la capacidad de pago de los miembros individuales, haciendo cumplir las reglas de reembolso o en su defecto imponiendo sanciones sociales y verificando y obteniendo información adicional (López 2009). Berger y Gleisner (2009) proporcionaron apoyo al efecto del capital social en la tasa de impago basada en la afiliación a un grupo. Berger explica que los líderes de estos grupos pueden disponer de información privada sobre los prestatarios que los prestamistas estándar no poseen, por lo que tienen una mayor capacidad de selección en el momento de elegir a los prestatarios correctos y además influyen un mayor poder a la hora de forzar a los prestatarios a reembolsar el préstamo. Sin embargo estos resultados no son consistentes con Freedam y Jin (2008), que divulgaron que la rentabilidad estimada de los préstamos de grupo son menores que los que se tendría si no se formara parte de un grupo. Esto es debido a que los líderes de grupo pueden tener numerosos incentivos inadecuados y financiar prestamos de baja calidad para obtener mayor recompensas. En consonancia con estas conclusiones, Wang y Greiner (2011), encontraron un gradual descenso en la publicación de "*listings*", después de disipar las recompensas a los líderes.

Una petición de préstamo puede ser catalogada en formato cerrado, la cual finaliza una vez alcanzada la financiación del montante solicitado, o en formato abierto, donde los prestamistas pueden seguir ofreciendo su dinero incluso si la cantidad requerida esta ya totalmente financiada, por debajo del tipo de interés establecido. Investigaciones previas sugieren que el formato de la subasta influye en los juicios que realiza los prestamistas: el formato cerrado tiene mayores posibilidades de ser financiado pero con tasas de interés superiores. Sin embargo, no existen diferencias en las tasas de impago entre los dos formatos (Lin et al. 2012; Puro et al. 2010). También se ha probado que el objetivo del préstamo influye de algún modo en la decisión de los prestamistas: menos

éxito y altas tasas de interés para préstamos comerciales que para aquellos relacionados con deudas (Wang et al. 2009, Collier y Hampshire 2010).

Debido a la inexperiencia financiera de los inversionistas individuales que normalmente participan en las plataformas de préstamos P2P, sus decisiones están basadas en lo que hacen los demás y no en sus propias ideas o decisiones, o dicho de otra forma, las decisiones individuales se realizan con la fuerte influencia de las decisiones de otros. Duan et al. (2009) propone dos razones que explicarían este fenómeno: 1) el exceso de información que provoca que los usuarios no la utilicen completamente (Brynjolfsson y Smith 2000) y 2) se puede observar con facilidad las opciones y decisiones en internet. De modo que los participantes realizan el comportamiento conocido como "*herding*", cuando las decisiones se basan en información imperfecta (Lee 2012; Herzenstein et al. 2011; Shen et al. 2010; Zhang y Liu 2012). Según Zhang y Liu (2012), los prestamistas observan el comportamiento del conjunto de participantes y también de las características de los prestatarios y por lo tanto, el comportamiento de "*herding*" es racional. Herzenstein (2011) lo define como la existencia de gran probabilidad de realizar pujas en subastas que ya tienen más ofertas.

Existen varios estudios que proponen directrices para realizar decisiones puramente racionales en la inversión social y que reducen el riesgo de impago (Klafft 2008): 1) invertir solamente en prestatarios sin cuentas morosas, 2) invertir en los que cumplan el primer criterio y además tengan un ratio deuda/ingreso inferior al 20% y 3) invertir en los que cumplan los dos primeros criterios y no tengan investigaciones en sus cuentas en los últimos 6 meses.

Cuando se afrontan múltiples decisiones en un periodo de tiempo limitado, en muchas ocasiones se usa la heurística, señales de decisión simple para simplificar el proceso (Tvesky y Kahneman 1974). En el contexto de los préstamos P2P, la heurística es un potente mecanismo de posible uso ya que los prestamistas tienen normalmente numerosas alternativas para elegir (Payne et al. 1993) y deben evaluar el uso de los prestamos dada una información objetiva limitada (Iyer et al. 2009). Mientras que la heurística añade gran eficacia, conduce a desviaciones de la decisión óptima (Tresky y Kahneman 1974), lo que pude conducir a no alcanzar rentabilidades deseadas.

Dadas las características singulares expuestas hasta el momento de esta modalidad de préstamos, que facilitan una mayor predisposición por el impago de estos, la importancia de tomar correctas decisiones en el momento de invertir en un préstamo se hace indispensable. Es por ello que, junto a la corriente cada vez más numerosa de realizar predicciones mediante la estrategia de buscar patrones dentro de grandes conjuntos de datos con los que entrenarse, provoca que la búsqueda de minimizar el riesgo de crédito sea una tarea en la que se deben realizar todos los esfuerzos necesarios. Las primeras investigaciones de predicciones fallidas fueron las realizadas por Ramser y Foster (1931) y Durand (1941). Desde entonces, la investigación teórica y empírica ha evolucionado considerablemente, enfocándose en la búsqueda de variables que conlleven menores tasas de error en la predicción y en la exploración de métodos estadísticos que mejoren la exactitud en la predicción.

2. METODOLOGÍA Y DATOS

Los datos utilizados en este trabajo corresponden a los préstamos otorgados a través de la plataforma de préstamos P2P Lending Club, la cual ofrece información amplia de cada petición de préstamo así como atributos del prestatario que pueden ser de interés tanto para los prestamistas como para el público en general y que se recoge en el anexo I. Algunas de las variables proporcionadas por la plataforma serán transformadas en ratios, otras en dicotómicas ordinales y algunas serán eliminadas, a fin de un mejor uso de la información disponible y del correcto funcionamiento del *software* utilizado, como se muestra en el anexo II. Nuestro propósito consiste, a través del *software* Tanagra, una herramienta para el tratamiento de datos, la obtención de aquellas variables que nos permitan proporcionar unos menores errores a la hora de predecir si los prestamistas van a cobrar o no los préstamos así como unos mayores valores de precisión y acierto dados unos prestamos definidos como *train* y que deben contener información completa y consistente, aplicando la predicción a una muestra de *test*.

El periodo analizado comprende desde junio 2007 hasta marzo 2015 inclusive, con un número total de 550.564 préstamos y de 9.407.250.600 millones de dólares intercambiados entre esas fechas.

La plataforma permite prestamos tanto a 36 meses como a 60, pero el análisis se va a realizar con los préstamos a 36 meses, eliminando los de 60, lo cual proporciona un abanico más amplio de estudio e investigación, debido a la necesidad de que transcurran esos 36 meses para calificar el préstamo como pagado (*fully paid*) o impagado (*default* o *charged off* [4]), tanto en los utilizados como *train* como a los que vamos a realizar el *test*. Esto nos permitirá conocer en una etapa posterior los siguientes ratios: tasa de error, error tipo I y error tipo II, precisión, acierto y F-score.

		Predicción	
		Solvente	Fallido
Realidad	Solvente	Verdadero positivo *a*	Falso negativo (error tipo II) *b*
	Fallido	Falso positivo (error tipo I) *c*	Verdadero negativo *d*

Tabla 1. Matriz de confusión

Tasa de error ((*c*+*b*)/*n*): mide el porcentaje de error en la clasificación de los préstamos.

Tasa de acierto ((*a*+*d*)/*n*): calcula el porcentaje de acierto en la clasificación de los préstamos.

Error tipo I (*c*/(*c*+*d*)): mide el porcentaje de los préstamos que se habían clasificados como solventes y que en realidad no lo son del total de préstamos fallidos.

Error tipo II (*b*/(*a*+*b*)): representa el porcentaje de préstamos que son pagados y que se habían catalogado como insolventes del total de préstamos pagados.

Precisión positivos (*a*/(*a*+*c*)): porcentaje de los préstamos concedidos a prestatarios catalogados como solventes y que en la realidad lo son del total de prestatarios clasificados como solventes.

Precisión negativos (*d*/(*b*+*d*)): porcentaje de préstamos adjudicados a prestatarios clasificados como insolventes y que al final serán pagados del total de prestatarios catalogados como insolventes.

Acierto pagados (*a*/(*a*+*b*)): % de acierto en los prestatarios catalogados como solventes

Aciertos fallados (*d*/(*c*+*d*)): % de aciertos en los prestatarios clasificados como no solventes.

[4] Un préstamo se convierte en *charged off* cuando no existe una expectativa razonable de pago. Ocurre normalmente cuando el préstamo esta vencido hace 150 días.

F-score ((*2*precisión*acierto)/(precisión + acierto*)): media armónica de precisión y acierto.

Una vez reducida la muestra a aquellos préstamos a 36 meses, obtenemos 394.523 préstamos. Disminuimos la muestra filtrando aquellos préstamos no clasificados como pagados (*fully paid*) o impagados (*default* o *charged off*): prestamos con todos los pagos hasta la fecha (*current*), prestamos en periodo de gracia de 15 días (*in grace period*), prestamos atrasados (*late 16-30 days* o *late 31-120 days*) y préstamos recibidos por parte de prestatarios que no conocían las políticas de crédito. De tal forma que trabajamos con 118.734 préstamos.

Una vez realizados estos filtros, hacemos una clasificación de los préstamos por semestres para poder elegir cuales utilizamos como muestra *train* y a cuales vamos a realizar el *test* y que se detalla en el anexo III. Elegimos como lapso de tiempo al que realizar el *test* el trimestre de 2012 de enero a marzo ya que es el último periodo al que se lo podemos realizar debido a que deben transcurrir como bien se ha comentado antes, 36 meses hasta que se le impone el estatus correspondiente al préstamo, tomando en consideración que existe información disponible hasta marzo de 2015. Se va a utilizar como *train* todo el año de 2008 al no poder solaparse en el tiempo prestamos utilizados como *train* cuando no han transcurrido aún 36 meses y prestamos empleados para el *test*. No puede existir un número distinto de préstamos pagados e impagados en la parte de *train*, lo que conduce a realizar un procedimiento comúnmente denominado parear en la búsqueda del mismo número de préstamos pagados e impagados mensuales con características similares entre cada préstamo pagado e impagado elegido. En lo que respecta a la muestra de test, no existe ninguna restricción que aplicar. Esto determina que obtengamos 484 préstamos que utilizaremos como entrenamiento, de los cuales la mitad serán pagados y la otra mitad impagados, y 6.042 préstamos formaran la muestra *test.*

Tras realizar este procedimiento, y dotándonos del *software* Tanagra, procedemos a la selección de variables relevantes, tanto al 1% como al 5% de significatividad, ya que el objetivo es eliminar variables redundantes, atributos espúreos y todos aquellos atributos que no aporten ningún aumento de la información. que nos permitirá obtener un modelo

más robusto y mucho más interpretable. En el proceso de selección de variables empleamos los siguientes procedimientos: 1) *forward logit*: se inicia por un modelo que no contiene ninguna variable explicativa y se añade como primera de ellas a la que presente un mayor coeficiente de correlación en valor absoluto con la variable dependiente. En los pasos sucesivos se va incorporando al modelo aquellas variables con un mayor coeficiente de correlación parcial con la variable dependiente dadas las independientes ya incluidas en el modelo. El procedimiento se detiene cuando el incremento en el coeficiente de determinación debido a la inclusión de una nueva variable explicativa en el modelo ya no es significativo. 2) *stepdisc*: técnica que se origina como *forward logit* pero en cada etapa se determina si las variables introducidas deben de permanecer. El proceso finaliza cuando ninguna variable se introduce o se elimina de la regresión. El algoritmo es el siguiente: se fijan unos criterios de entrada, F_{IN} y de salida, $F_{OUT,}$ de tal forma que se concretan unas regiones de aceptación de la hipótesis nula, donde la variable regresora no es significativa, y de rechazo, tanto para entrar como para salir del modelo. Se calculan los coeficientes de correlación simple (Y, xi), $_{i=1,2\ldots k}$ y la de mayor valor, por ejemplo i=k, es la candidata a entrar en el modelo. Se obtiene la regresión Y sobre x_k y se calcula el estadístico F_k. Se determina si la variable se introduce en el modelo ($|F_k| > F_{IN}$). Una vez introducida la variable x_k en el modelo, se calculan las correlaciones parciales, eliminando la influencia de $x_{k,}$ (Y, x_i), $_{i=1,2\ldots k-1}$ y la de mayor valor será candidata a entrar en el modelo, en este caso x_{k-1}. Se obtiene la regresión de Y sobre x_k y x_{k-1} y se calculan los estadísticos F_k y F_{k-1}. Se establece si la variable x_{k-1} se introduce en el modelo ($|F_{k-1}| > F_{IN}$) y si la variable x_k debe permanecer en el modelo ($|F_k| > F_{OUT}$). Y así sucesivamente. 3) FCBF[5] (*fast correlation based filter*): consta de un procedimiento en dos pasos conectados, primero, la selección de un subconjunto de características relevantes S_{list}, y seguidamente la selección de aquellas predominantes. Dado un conjunto de datos con N características y clase C, el algoritmo encuentra un conjunto de características principales. En el primer paso, se calcula el valor SU (*symmetrical uncertainty*), que mide la correlación de cada característica con la clase C, también denominado correlación C, se seleccionan las variables relevantes basándose en un delta predefinido y se clasifican en orden descendente según el valor SU. En el segundo paso se procesa la lista ordenada para seleccionar las características predominantes. Una

[5] Ver Yu, L. & Liu, H. Efficient feature selection via analysis of relevance and redundancy. *Journal of machine learning research* 5 (2004)

característica, F_j que ya se ha establecido como predominante, puede ser utilizada para filtrar otras características siguiendo la "*manta de Markov*"[6]. Ya que la característica con el mayor valor SU no tendrá ninguna aproximación con la "*manta de Markov*", entonces debe ser una característica predominante. Para todas las características restantes (desde la inmediatamente posterior a F_j hasta la última de S_{list}) F_j pasa a formar una "*manta de Markov*" con F_i, $_{i=1,2\ldots N}$, F_i saldrá fuera de la lista S_{list}. El algoritmo toma como nueva referencia a la característica restante que más próxima a F_j se encuentre en S_{list}, y se repite el mismo proceso hasta que no se pueden seleccionar más características predominantes. 4) MIFS[7] (*mutual information feature selection*): usado por Battiti y Al (1994), se parte del intento, dado un conjunto inicial de *n* características de encontrar un subconjunto *k*, con $k < n$, que maximiza la información sobre la clase, también denominado *FRn-k*. Sin embargo este procedimiento es limitado dada su dificultad de aplicación práctica. Por ello se hace una aproximación del vector "*mutual information*", MI, usando el "*mutual information*" entre los componentes individuales del vector. A pesar de calcular el "*mutual information*" I (F, C), entre un vector característico y la variable clase C, solamente calculamos I (f, C) e I (f´, C), donde f y f´ son características individuales. Entonces el análisis de todos los posibles subconjuntos es sustituido por un algoritmo "*greedy*" buscando en cada etapa la opción óptima, con la esperanza de alcanzar una solución general óptima. Dado un conjunto de variables seleccionados, el algoritmo elige la variable, tal que se maximiza la información sobre la clase restando una cantidad proporcional a la media MI de las variables seleccionadas.

Tras realizar la selección de variables que se consideran más relevantes en el análisis, dotándonos de los algoritmos que a continuación se exponen, se examinan aquellas variables que son significativas para la predicción logrando minimizar la tasa de error en la clasificación de los préstamos que van a ser utilizados como *test*: 1) KNN (*k-nearest neighbors*): este algoritmo es un método no paramétrico extensamente usado y desarrollado por Fix y Hodges (1951) que se basa en el aprendizaje de casos. Los inputs son los k casos de prueba (*training*) más cercanos en lo que concierne a cierta función de distancia, denominada euclidiana. La clasificación de una nueva muestra está basada

[6] Ver Perez Rubido, R. A review of feature selection algorithms that treat the microarray data redundancy. *Revista Cubana de Ciencias Informaticas*, Vol. 7, nº 4 (2013), 23-24.
[7] Ver Battiti, R. Using Mutual Information for Selecting Features in Supervised Neural Net Learning. *IEEE Transactions on neural Networks*, Vol. 5, nº 4, July 1994.

en el voto mayoritario de los k vecinos más cercanos. Es aconsejable pensar que los vecinos más cercanos tuvieran un mayor peso en la clasificación (Aha y Kibler 1991). 2) Regresión logística binaria: las primeras aplicaciones de esta regresión se encuentran en Martín (1977) y Ohlson (1980). Es un procedimiento similar a un modelo de regresión lineal, que presenta ciertos problemas cuando la variable dependiente es binaria, adaptándose, ya que no presenta condiciones de aplicación restrictivas y es por ello que la regresión logística es más robusta que el análisis discriminante, al requerir menos supuestos (Pérez y Santin 2007). A diferencia del modelo lineal, que suele hacer uso en la estimación de mínimos cuadrados ordinarios, en la estimación de la regresión logística se utiliza el método de máxima verosimilitud en la estimación de los parámetros. 3) Regresión logística multinomial: los primeros usos se encuentran en Hosmer y Lemeshow (1989). Es una regresión muy similar a la regresión logística binaria, con la única diferencia de que la variable dependiente no está restringida a dos categorías. 4) SVM (*support vector machine*): presentado por Boser, Guyon, y Vapnik (1992). Para un conjunto de datos con una variable de respuestas binaria, el apoyo de un vector hace separar los datos en dos regiones, uno para cada clase, en el espacio p-dimensional a través de un hiperplano con el máximo margen entre las distancias de las dos clases (Cortes y Vapnik 1995). La maximización del margen disminuye la complejidad del modelo y el riesgo de errores. Cuando los datos no son separables por un hiperplano, que en la práctica es la mayoría de las veces, un margen suave es usado. Ante tal situación una holgura positiva es añadida en el lado considerado malo del margen. Estas holguras incrementan la distancia del margen. El objetivo es minimizar la suma de estas holguras mientras maximizamos la anchura del margen (Han y Zhao 2013; Schebesch y Stecking 2005; Zhor, Lai y Yu 2010). 5) *Random forest*: técnica que se remonta a Hunt, Marin y Stone (1966). El árbol de decisión es un método de clasificación popular que cultiva una estructura de árbol con probabilidad al final de cada rama del árbol. El árbol de decisión comienza con un nodo raíz y gradualmente se construyen sub-arboles con los nodos internos que están conectados por ramas que emanan y acaban en nodos terminales, denominados hojas. Cada nodo interno corresponde a un test de una característica y las ramas representan una partición binaria del atributo al que se le realiza el test. Sin embargo, existen dos publicaciones críticas con este método: i) como escoger la división de los atributos en cada nodo y ii) cuantos niveles tiene que tener cada rama hasta que deja de dividirse. Las divisiones son realizadas por el índice de Gini y el número de ramas

es controlado por un parámetro d (Breiman 2001). 6) *Naive bayes continuous*: este algoritmo está basado en el teorema de Bayes (1973), de probabilidades condicionales. Se centra en datos históricos y utiliza el teorema de Bayes para calcular la probabilidad de que un suceso ocurra dada la probabilidad de un suceso que ya ha ocurrido. Tiene la ventaja de que no tiene necesidades de las distribuciones de las variables y no tiene requerimientos completos de información de las variables. 7) Análisis discriminatorio lineal: utilizado por Durand (1941), este método está basado en modelos de probabilidad lineal. Supone que si se tienen n préstamos para los que se conocen k variables explicativas, y se observa que n_1 de ellas pertenece a un grupo y n_2 a otro, es posible construir una función lineal de las k variables que puede usarse para predecir si una nueva observación pertenece a un grupo u a otro con una probabilidad determinada.

Tras la selección de aquellas variables significativas que minimizan los errores en la predicción, se calculan los ratios expuestos anteriormente al comienzo de este epígrafe.

3. RESULTADOS

Se presentan las siguientes variables seleccionadas en función del método utilizado:

Si introducimos todas las variables en el modelo con el algoritmo *forward logit*, al 1% de significatividad, la variable seleccionada únicamente es *loan_estatus_fully_paid* para minimizar la tasa de error en la predicción de si un préstamo se reembolsa o no, lo cual es indudable y obvio, de modo que dejamos de introducir esta variable junto al resto de variables que indican el estatus del prestamos además del *% devuelto* al considerarla de índole similar. Esto ocurre principalmente debido a que estos métodos de selección de variables que se presentan en este trabajo reemplazan la búsqueda de un óptimo global por la consideración sucesiva de óptimos locales, y por lo tanto no es seguro que encuentren al final el óptimo global ni que este sea el mismo para cada procedimiento. Además pueden surgir problemas de inestabilidad, ya que pequeños cambios en el conjunto de datos pueden producir grandes modificaciones en las variables seleccionadas (Breiman 1996). Estos problemas pueden ser agravados por predictores fuertemente

correlacionados, ya que una variable que tiene efecto sobre la respuesta puede no ser elegida si alguna variable correlacionada con esta se encuentra ya en el modelo.

Obtenemos entonces, el siguiente resultado:

Variable	Chi-2	p-valor
1. total_rec_prncp	*147,333*	*0,000*
2. total_rec_int	*165,477*	*0,000*
3. grado_g	*99,902*	*0,000*
4. pagosprestamo_annual_inc	*82,664*	*0,000*
5.last_pymnt_amnt	*112,493*	*0,000*

Tabla 2. Selección de variables con método forward

Resultados idénticos se observan al realizar el análisis al 5% de significatividad.

Pero la selección de determinadas variables, como se ha visto anteriormente, debido a su evidencia o al carácter del algoritmo, provoca que no se seleccionen otras variables que bien podrían explicar con cierta calidad el modelo y/o disponer de una tasa de error mínima. De tal forma que a continuación, se exponen otros modelos obtenidos al excluir una serie de variables.

Si no se introduce *pagosprestamo_annual_inc*, las variables seleccionadas son los siguientes:

Variable	Chi-2	p-valor
1. total_rec_prncp	*147,333*	*0,000*
2. total_rec_int	*165,477*	*0,000*
3. grado_g	*99,902*	*0,000*
4. 2_int_rate	*49,110*	*0,000*
5. loan_amnt	*34,961*	*0,000*
6. loan_amnt-funded_amnt	*52,769*	*0,000*

Tabla 3. Selección de variables con método forward

Si se elimina *total_rec_prncp, total_rec_int, int_rate, loan_amnt* y *loan_amnt-funded_amnt* y se vuelve a introducir *pagosprestamo_annual_inc* entre las posibles variables explicativas, logramos:

Variable	Chi-2	p-valor
1.total_pymnt	*114,428*	*0,000*
2. pagosprestamo_annual_inc	*126,744*	*0,000*
3. grado_g	*423,840*	*0,000*
4. 2_int_rate	*59,338*	*0,000*
5. grado_f	*20,682*	*0,000*
6. total_rec_late_fee	*16,221*	*0,0001*
7. last_credit_pull_d	*10,382*	*0,0013*

Tabla 4. Selección de variables con método forward

Al 5% de significatividad obtenemos, al introducir todas las variables en el algoritmo *forward*, la selección de las variables obtenidas al 1% de forma idéntica. Otro modelo obtenido, al eliminar *% devuelto*, las variables de *loan_status, total_rec_prncp, total_rec_int* y *funded_amnt* es:

Variable	Chi-2	p-valor
1. last_credit_pull_d	*112,053*	*0,000*
2. total_pymnt_inv	*63,526*	*0,000*
3. last_pymnt_amnt	*46,802*	*0,000*
4. loan_amnt	*24,264*	*0,000*
5. collection_recovery_fee	*14,961*	*0,0001*
6. issue_d	*16,915*	*0,000*

Tabla 5. Selección de variables con método forward

Al utilizar en la selección de variables el algoritmo *stepdisc*, al 1 % de significatividad y utilizando todas la variables, excepto las de *loan_status* obtenemos como variables seleccionadas: *% devuelto, total_rec_late_fee e int_rate*, lo cual no

proporciona una información mínima que pueda ser útil en los siguientes pasos. Se alcanza otro modelo, excluyendo del análisis diversas variables.

Variable	F	p-valor
1. total_rec_prncp	*210,93*	*0,000*
2. funded_amnt	*295,30*	*0,000*
3. last_credit_pull_d	*31,02*	*0,000*
4. total_rec_int	*31,66*	*0,000*
5. installment	*18,85*	*0,000*
6. last_pymnt_amnt	*12,92*	*0,000*
7. total_rec_late_fee	*7,56*	*0,0062*

Tabla 6. Selección de variables con método stepdisc

Al 5 % de significatividad conseguimos una selección de variables mucho más amplia. La primera, sin la utilización de *% devuelto* y *loan_status*:

Variable	F	p-valor
1. total_rec_prncp	*210,93*	*0,000*
2. funded_amnt	*295,30*	*0,000*
3. last_credit_pull_d	*31,02*	*0,000*
4. total_rec_int	*31,66*	*0,000*
5. installment	*18,85*	*0,000*
6. last_pymnt_amnt	*12,92*	*0,004*
7. total_rec_late_fee	*7,56*	*0,0062*
8. purpose_money	*5,74*	*0,0169*
9. collection_recovery_fee	*5,16*	*0,0236*
10. total_pymnt_inv	*5,06*	*0,0250*

Tabla 7. Selección de variables con método stepdisc

La segunda eliminando *loan_amnt y total_rec_prncp* del conjunto de variables e introduciendo *% devuelto*:

Variables	F	p-valor
1. % devuelto	*1777,13*	*0,000*
2. total_rec_late_fee	*16,01*	*0,0001*
3. int_rate	*8,49*	*0,0037*
4. emp_length_bien	*5,57*	*0,0187*
5. subgrade_2_int_rate	*4,48*	*0,0348*
6. installment	*4,73*	*0,0302*
7. total_rec_int	*39,36*	*0,000*
8. total_pymnt	*8,67*	*0,0034*
9. last_pymnt_amnt	*14,10*	*0,0002*
10. funded_amnt	*8,97*	*0,0029*
11. recoveries	*8,89*	*0,0030*

Tabla 8. Selección de variables con método stepdisc

Y la tercera, eliminando *loan_amnt, total_rec_prncp, % devuelto* y todos los *loan_status*:

Variables	F	p-valor
1. total_pymnt	*149,24*	*0,000*
2. installment	*249,58*	*0,000*
3. recoveries	*46,98*	*0,000*
4. total_rec_int	*52,00*	*0,000*
5. last_credit_pull_d	*39,46*	*0,000*
6. funded_amnt	*26,60*	*0,000*
7. last_pymnt_amnt	*12,34*	*0,0005*
8. total_rec_late_fee	*7,98*	*0,0049*
9. purpose_moving	*5,64*	*0,0180*
10. total_pymnt_inv	*4,87*	*0,0278*

Tabla 9. Selección de variables con método stepdisc

Con el método *MIFS filtering*, las variables introducidas en el modelo deben de ser discretas, de modo que la selección es la siguiente, con una beta de 0,01 y eliminando *loan_status* del grupo de posibles variables a seleccionar:

Variable	I
1. addr_state	*0,138981*
2. sub_grade	*0,032551*
3. purpose	*0,022208*
4. verification_status	*0,002349*
5. grade	*0,004774*
6. home_ownership	*0,002023*

Tabla 10. Selección de variables con método MIFS filtering

Y finalmente con el procedimiento *FCBF filtering*, al igual que con el método anterior las variables deben ser discretas. Con un delta nulo y todas las variables como posibles seleccionadas obtenemos: *addr_state* y *loan_status*. Si eliminamos *loan_status* debido al gran poder explicativo que tiene, solamente se selecciona *addr_state.*

Una vez realizado los cuatro procedimientos para la selección de variables, detectando aquellas consideradas como irrelevantes y/o redundantes, seleccionamos dos modelos que contienen las variables consideradas, donde aplicamos los algoritmos propuestas en el apartado 2, permitiéndonos perfeccionar los modelos clasificatorios con las variables significativas que se obtienen para alcanzar la predicción más exacta y obtener los menores errores posibles. Seleccionamos dos modelos: el tercero del procedimiento *forward* al 1% de significatividad, que corresponde a la *Tabla 3* y el tercero del método *stepdisc* al 5% de significatividad, que concierne a la *Tabla 8*.

Partiendo del modelo *forward* seleccionado, con las variables *pagosprestamo_annual_inc*, *2_int_rate*, *total_rec_late_fee* y *last_credit_pull_d* como base, obviando *grade_g*, *grade_f* y *total_pymnt*, a causa de su carácter que impiden tomarlas como variables finales a la hora de considerar su análisis para alcanzar los menores errores en la clasificación de los préstamos, aunque necesarias a la hora de la selección de las variables debido al funcionamiento que siguen los métodos de selección, debemos ajustar el modelo obteniendo cuál de ellas son significativas y si es posible añadir alguna más que consiga reducir los errores en la predicción, ya que como se ha comentado anteriormente, dos métodos diferentes en la selección de variables obtienen diferentes resultados, por lo tanto no son infalibles.

Utilizando los algoritmos descritos en el epígrafe anterior, el mejor modelo considerado como base obtenido es el que nos proporciona *la regresión logística binaria* y *la regresión logística multinomial*. Para la muestra utilizada como *train* obtenemos los siguientes resultados:

		Predicción		
		solvent	*failed*	*sum*
Realidad	*solvent*	178	64	242
	failed	31	211	242
	sum	209	275	484

Tabla 11. Matriz de confusión train

Ratio	Valor
Tasa de error	*19,63%*
Tasa de acierto	*80,37%*
Acierto pagados	*73,55%*
Acierto fallados	*87,19%*
Precisión pagados	*85,16%*
Precisión fallados	*76,72%*
Error tipo I	*12,80%*
Error tipo II	*26,44%*
F-score	*78,94%*

Tabla 12. Ratios para train

Variable	Coeficiente	Significatividad
constant	*81,9208*	*0,0000*
total_rec_late_fee	*0,01739*	*0,0118*
last_credit_pull_d	*-0,0020*	*0,0000*
pagosprestamo_annual_inc	*-0,000005*	*0,3352*
2_int_rate	*15,5818*	*0,4753*

Tabla 13. Variables y significatividad

Como se puede observar, *pagosprestamo_annual_inc y 2_int_rate* son no significativas para minimizar la tasa de error en el momento de clasificar correctamente a los préstamos otorgados.

Para la predicción del *test* se estiman los siguientes resultados:

		Predicción		
		solvent	*failed*	*sum*
Realidad	*solvent*	*5258*	*0*	*5258*
	failed	*784*	*0*	*784*
	sum	*6042*	*0*	*6042*

Tabla 14. Matriz de confusión test

Como se puede apreciar, la muestra *train* en este caso no proporciona una buena predicción para el *test*, ya que los resultados que se obtienen para los *failed* son más erróneos, como se muestra en la siguiente tabla.

Ratio	Valor
Tasa de error	*12,97%*
Tasa de acierto	*87,03%*
Acierto pagados	*100%*
Acierto fallados	*0%*
Precisión pagados	*87,02%*
Precisión fallados	*0%*
Error tipo I	*100%*
Error tipo II	*0%*
F-score	*93,05%*

Tabla 15. Ratios para test

Tras obtener una tasa de error del *12,97%*, la cual puede ser considerada como un valor aceptable, ya que se establecen valores por debajo del *20%* como adecuados, y un

error tipo I del *100%*, se pretende perfeccionar la predicción, en el intento de minimizar lo máximo la tasa de error con la introducción y/o eliminación de aquellos atributos que sean o no significativos. Pero nos encontramos ante un problema: la introducción de más atributos en el modelo que permiten reducir la tasa de error, como es el caso de la variable *collection_recoveries_fee,* donde minimizamos la tasa de error, con el *15,70%* en el *train* y en la predicción del *test* del *5,11%* provoca que la única variable significativa sea *last_credit_pull_d.* Tal suceso indica que el analista financiero tomará la decisión de otorgar un préstamo en función de esta única variable, la cual considera relevante, ya que es con la que menor error tendrá en la predicción de si el préstamo será reembolsado o en su defecto impagado. Sin embargo, en la práctica es un resultado irracional y poco coherente, ya que este atributo indica que se minimiza el error con la fecha de la última consulta o reporte crediticio.

Los resultados obtenidos son los siguientes, tanto para el *train* como para el *test*:

		Predicción		
		solvent	*failed*	*sum*
Realidad	*solvent*	*223*	*19*	*242*
	failed	*57*	*185*	*242*
	sum	*280*	*204*	*484*

Tabla 16. Matriz de confusión train

Ratio	Valor
Tasa de error	*15,70%*
Tasa de acierto	*84,30 %*
Acierto pagados	*92,14%*
Acierto fallados	*76,44%*
Precisión pagados	*79,64%*
Precisión fallados	*90,68%*
Error tipo I	*23,55%*
Error tipo II	*7,85%*
F-score	*85,44%*

Tabla 17. Ratios para train

Variable	Coeficiente	Significatividad
constant	*70,1874*	*0,0000*
total_rec_late_fee	*0,0062*	*0,4066*
collection_recovery_fee	*48,5589*	*0,7827*
last_credit_pull_d	*-0,0017*	*0,0000*
pagosprestamo_annual_inc	*0,0000*	*0,8457*
2_int_rate	*30,4129*	*0,2369*

Tabla 18. Variables y significatividad

		Predicción		
		solvent	*failed*	*sum*
Realidad	*solvent*	*5258*	*0*	*5258*
	failed	*309*	*475*	*784*
	sum	*5567*	*475*	*6042*

Tabla 19. Matriz de confusión test

Ratio	Valor
Tasa de error	*5,11%*
Tasa de acierto	*94,89%*
Acierto pagados	*100%*
Acierto fallados	*60,58%*
Precisión pagados	*94,44%*
Precisión fallados	*100%*
Error tipo I	*39,41%*
Error tipo II	*0%*
F-score	*97,15%*

Tabla 20. Ratios para test

Como se puede apreciar, todos los ratios han mejorado en el *test* si los comparamos con los mismos del modelo base, en detrimento de un menor número de variables significativas y de una predicción menos coherente. Centrándonos en el modelo *stepdisc* seleccionado, y los atributos *int_rate*, *emp_length_bien* y *funded_amnt/annual_inc* como variables base, desechando el resto seleccionadas por no considerarlas variables correctas ante el análisis. Este modelo, obtenido con el algoritmo *support vector machine (SVM)* proporciona la siguiente información para el *train*:

		Predicción		
		solvent	*failed*	*sum*
Realidad	*solvent*	148	94	242
	failed	131	111	242
	sum	279	205	484

Tabla 21. Matriz de confusión train

Ratio	Valor
Tasa de error	*46,49%*
Tasa de acierto	*53,51%*
Acierto pagados	*61,15%*
Acierto fallados	*45,86%*
Precisión pagados	*53,04%*
Precisión fallados	*54,14%*
Error tipo I	*54,13%*
Error tipo II	*38,84%*
F-score	*56,81%*

Tabla 22. Ratios para train

Variable	Weight	Significatividad
constant	*2,066*	*0,1532*
int_rate	*-8,599*	*0,3218*
emp_length_bien	*-0,009*	*0,5539*
funded_amnt/annual_inc	*-2,854*	*0,1625*

Tabla 23. Variables y significatividad

Con un resultado para la predicción del *test*:

		Predicción		
		solvent	*failed*	*sum*
Realidad	*solvent*	*2850*	*2408*	*5258*
	failed	*314*	*470*	*784*
	sum	*3164*	*2878*	*6042*

Tabla 24. Matriz de confusión test

Ratio	Valor
Tasa de error	*45,05%*
Tasa de acierto	*54,95%*
Acierto pagados	*54,20%*
Acierto fallados	*59,94%*
Precisión pagados	*90,07%*
Precisión fallados	*16,33%*
Error tipo I	*40,05%*
Error tipo II	*45,79%*
F-score	*67,68%*

Tabla 25. Ratios para test

Ninguna de las variables es significativa. En el intento de mejorar el modelo, obteniendo variables significativas y reduciendo la tasa de error, se nos presenta idéntico problema al que surgía en el procedimiento *forward*. El menor ratio de error se obtiene introduciendo las variables *collection_recovery_fee* y *last_credit_pull_d* en el modelo base, obteniendo como única variable significativa, como en el otro caso, al atributo *last_credit_pull_d*, con una tasa de error en el *train* de 15,29 % y en el *test* de 5,11 %, para los algoritmos *regresión logística binaria* y *regresión logística multinomial.*

Los resultados obtenidos para el *train* y el *test* se desarrollan a continuación:

		Predicción		
		solvent	*failed*	*sum*
Realidad	*solvent*	*225*	*17*	*242*
	failed	*57*	*185*	*242*
	sum	*282*	*202*	*484*

Tabla 26. Matriz de confusión train

Ratio	Valor
Tasa de error	*15,29%*
Tasa de acierto	*84,71%*
Acierto pagados	*92,97%*
Acierto fallados	*76,63%*
Precisión pagados	*79,78%*
Precisión fallados	*91,58 %*
Error tipo I	*23,55 %*
Error tipo II	*7,02%*
F-score	*85,88%*

Tabla 27. Ratios para train

Variable	Coeficiente	Significatividad
constant	*68,9684*	*0,0000*
int_rate	*7,4239*	*0,2483*
collection_recovery_fee	*43,5862*	*0,6585*
last_credit_pull_d	*-0,0017*	*0,0000*
emp_length_bien	*0,0421*	*0,2043*
funded_amnt/annual_inc	*1,5840*	*0,1914*

Tabla 28. Variables y significatividad

		Predicción		
		solvent	*failed*	*sum*
Realidad	*solvent*	*5258*	*0*	*5258*
	failed	*309*	*475*	*784*
	sum	*5567*	*475*	*6042*

Tabla 29. Matriz de confusión test

Ratio	Valor
Tasa de error	*5,11%*
Tasa de acierto	*94,89%*
Acierto pagados	*100%*
Acierto fallados	*60,58%*
Precisión pagados	*94,44%*
Precisión fallados	*100%*
Error tipo I	*39,41%*
Error tipo II	*0%*
F-score	*97,15%*

Tabla 30. Ratios para test

Como se puede comprobar, se obtienen idénticos ratios en el *test* para los modelos *forward* y *stepdisc* finales, debido a que la variable única variable significativa es *last_credit_pull_d*.

Ante estos resultados, surge el dilema si es más conveniente disponer de una clasificación con un error mínimo, que si bien es la finalidad u objetivo de los algoritmos, que se basa exclusivamente en una única variable, *last_credit_pull_d,* la cual aunque el análisis la refleja como el atributo en el que se debe basar la decisión de prestar para obtener el menor error, ya que en la práctica no sería sensato tomar una decisión fundamentándose en esta única variable o bien alcanzar una predicción con mayor error, con mayor posibilidad de tener errores al dilucidar los prestamistas que reembolsarán sus préstamos y que habíamos predicho que no lo harían y que prestamistas no lo harán habiendo vaticinado que eran solventes basándose no solamente en un número mayor de variables, sino además más coherentes.

A causa de este motivo, se plantea una predicción más coherente en la práctica con el objetivo de disponer con un mayor número de variables significativas que los resultados obtenidos, si bien tendrá que ser a costa de un mayor error en la probabilidad de equivocarnos en la clasificación. Se seleccionan diez atributos que la experiencia dice que podrían ser significativas en el momento de predecir si un préstamo va a ser devuelto o no. Las distintas variables son las siguientes: *int_rate, home_ownership, delinq_2yrs, grade_ord, purpose_nominal emp_length_bien, days_historial, dti, open_acc* y *pagosprestamo_annual_inc*. Se obtienen los mejores resultados para los algoritmos *regresión logística binaria* y *regresión logística multinomial*.

En el *train* obtenemos:

Variable	Coeficiente	Significatividad
constant	*3,4560*	*0,2235*
int_rate	*-15,7306*	*0,2551*
dti	*-0,0036*	*0,8291*
delinq_2yrs	*-0,4020*	*0,0475*
open_acc	*-0,0202*	*0,3088*
emp_lenght_bien	*-0,0120*	*0,6367*
home_ownership_ord	*0,2223*	*0,1681*
pagosprestamo_annual_inc	*0,0000*	*0,5165*
purpose_nominal	*0,0362*	*0,1909*
days_historial	*0,0000*	*0,9464*
grade_ord	*-0,4343*	*0,0607*

Tabla 31. Variables y significatividad

		Predicción		
		solvent	*failed*	*sum*
Realidad	*solvent*	*134*	*108*	*242*
	failed	*105*	*137*	*242*
	sum	*239*	*245*	*484*

Tabla 32. Matriz de confusión train

Ratio	Valor
Tasa de error	*44,01%*
Tasa de acierto	*55,99%*
Acierto pagados	*55,37%*
Acierto fallados	*56,61%*
Precisión pagados	*56,07%*
Precisión fallados	*55,92%*
Error tipo I	*43,39%*
Error tipo II	*44,63%*
F-score	*55,72%*

Tabla 33. Ratios para train

Para el *test* presentamos los siguientes resultados:

		Predicción		
		solvent	*failed*	*sum*
Realidad	*solvent*	*4394*	*864*	*5258*
	failed	*676*	*108*	*784*
	sum	*5070*	*972*	*6042*

Tabla 34. Matriz de confusión para test

Ratio	Valor
Tasa de error	*25,49%*
Tasa de acierto	*74,51%*
Acierto pagados	*83,57%*
Acierto fallados	*13,78%*
Precisión pagados	*86,67%*
Precisión fallados	*11,11%*
Error tipo I	*86,22%*
Error tipo II	*16,43%*
F-score	*95,09%*

Tabla 35. Ratios para test

Se puede observar como únicamente *delinq_2yrs* y *grade_ord* son significativas aunque se podía pensar que el número fuese mayor. Pero no es un resultado aislado. Existe la evidencia de que las mejores predicciones, con los menores errores para esta plataforma se consiguen con un número mínimo de variables significativas, en concreto con *last_credit_pull_d.* Resultados subóptimos también se obtienen con un número muy reducido de variables significativas. No se puede, por lo tanto, realizar predicciones coherentes y razonables ampliando el número de variables significativas y a la vez obtener valores casi despreciables en el error a la hora de clasificar a los prestamistas dada la muestra con la que hemos realizado el *train* y el *test*.

4. CONCLUSIÓN

Desde su primera aparición en el Reino Unido, los prestamos *peer to peer* han experimentado un gran crecimiento en los últimos años, identificándose como un nuevo método de microfinanzas que proporciona servicios financieros a través de una plataforma online sin la presencia de intermediarios financieros tradicionales. En España, con la existencia de una cultura muy bancarizada, este fenómeno se encuentra aún poco desarrollado, aunque favorecido por la crisis económica, ha experimentado un gran crecimiento en los últimos años. Estas transacciones afrontan una variedad de riesgos provocados por la incertidumbre implícita en el uso de una plataforma online y/o por las características y la conducta de los prestatarios.

Dadas estas características y el funcionamiento del proceso de financiación, la aparición de información asimétrica es un fenómeno muy presente en estos mercados, favoreciendo la selección adversa y el riesgo moral, que hacen fomentar el riesgo de impago. Existen evidencias empíricas donde el análisis tradicional, que se apoya en la base del "*credit score*" de los préstamos otorgados, para la clasificación de los prestatarios, no proporciona señales correctas en la predicción de la probabilidad de impago.

Es por ello que se planeta en este trabajo un procedimiento diferente, que permite calcular la tasa error en la clasificación de los préstamos basándose en las similitudes con aquellos concedidos anteriormente mediante el uso de varios algoritmos, que proporcionan aquellas variables y/o atributos de los prestatarios que son significativos para minimizar la tasa de error en la clasificación. Anteriormente se realiza un mecanismo de selección de variables, a través de cuatro procedimientos diferentes, para eliminar aquellas variables que son redundantes para la predicción y que no aportan información alguna.

Dotándonos de la información proporcionada por Lending Club, que actualmente es la mayor plataforma de préstamos P2P del planeta, tanto en lo que respecta a los préstamos financiados como a las características que de los prestatarios obtenemos los siguientes resultados: de las técnicas utilizadas, los algoritmos *regresión logística binaria*

y *regresión logística multinomial* proporcionan el menor error en la clasificación de los prestatarios, concretamente del 5, 11%, con un error del tipo I del 39,31% y del tipo II del 0%, con unos valores, además de precisión y de acierto bastante aceptables, a través de dos modelos diferentes, uno procedente del método de selección de variables *forward* y otro del *stepdisc*. El problema surge cuando estos resultados están basados en una un único atributo, *last_credit_pull_d,* variable que indica, que la última fecha en la que se realizó una consulta crediticia es la variable significativa que nos establece el menor error en la clasificación de los préstamos. Esta conclusión, aunque si bien es cierto que cumple el propósito del estudio, en la práctica no es un resultado coherente, ya que el uso de única variable, posiblemente por la existencia de correlación de las variables, puede conducir a errores en la realidad.

Estos resultados toman relevancia cuando se plantea un modelo alternativo, obviando los procedimientos de selección de variables, que en principio parece coherente con el objetivo de obtener una predicción basándose en un mayor número de variables significativas en el que alcanzamos mayor tasa de error en la clasificación de los préstamos, en concreto del 25,49% a través de los algoritmos *regresión logística binaria* y *regresión logística multinomial*. Pero nuevamente la variable significativa es únicamente *last_credit_pull_d,* lo cual nos lleva a decir que el objetivo de nuestro modelo planteado alternativo no obtiene los resultados esperados, teniendo que aceptar que las decisiones que tienen el propósito de minimizar la tasa de error en la clasificación se deben basar exclusivamente en *last_credit_pull_d.*

Hay que considerar que estos resultados, provienen únicamente dada una muestra utilizada como *train*, año 2008, y otra de *test*, tres primeros meses de 2012, y que no conlleva que estos resultados se obtengan para otros periodos de tiempo ni para otras plataformas de préstamos *peer to peer*.

REFERENCIAS

Aha, D. & Kibler, D. (1991).Instance – based learnings algorithms. *Machine Learning*, 6, pp.37-66.

Bacharach, M. & Gambetti, D. (2001) Trust in signs, In Trust in society, K. Cook (ed.). New York: *Russell Sage Foundation*.

Bachmann, A; Becker, A; Buerckner, D; Hilker, M; Kock, F; Lehmann, M. & Tiburtius, P. (2011). Online Peer to Peer Lending. A literature Review. *Journal of Internet Banking and Commerce*, Vol.16, nº 2, pp.1-18,

Battiti, R. Using Mutual Information for Selecting Features in Supervised Neural Net Learning. *IEEE Transactions on neural Networks*, Vol. 5, nº. 4, July 1994.

Benjamin, R. & Wigand, R. (1995) Electronic Markets and Virtual Value Chains on the Information Superhighway. *Sloan Management Review* (Winter), 62-72.

Berger, S.C. & Gleisner, F. Emergence of Financial Intermediaries in Electronics Markets: The Case of Online P2P Lending. *BuR Business Research. Verband der Hochschullehrer für Betriebswirtschaft* e.V, Vol. 2, issue 1. May 2009, 39-65.

Boser, B.E; Guyon, I. & Vapnik, V.N. (1992). A training algorithm for optimal margin classifiers. *In Proceedings of the 5th Annual Workshop of Computational Learning Theory*, 5 pp.144-152.

Breiman, L. (1996). Bagging predictors. *Machine Learning* 26(2), pp.123-140

Breiman, L. (2001) Random Forest. *Machine Learning* 45, pp.5-32.

Brown, C.M. (2008). Is peer to peer lending right for you? *Black Enterprise* (39:2), pp. 146.

Brynjolfsson, E. & Smith, M.D. (2000). Frictionless Commerce? A comparison of Internet and Conventional Retailers. *Management Science*, pp. 259-314. *Elsevier Science Publishers B.V.*

Collier, B. & Hampshire, R. (2010). Sending mixed signals: Multilevel reputation effects in peer to peer lending markets, proceeding of the CSCW, Savannah, Georhia, USA, pp.197-206.

Cortes, C. & Vapnik, V. (1995). Support-vector networks. Machine Learnings, 20, pp-273-297.

Duan, W; Gu, B. & Whinston, A (2009). Informational Cascades and Software Adoption on the Internet. An Empirical Investigation. *MIS Quarterly*, 23-48.

Durand, D. Risk Elements in Consumer Installment Financing, *National Bureau of Economic Research*, New York, 1941.

Evans, P. & Wurster, T. Strategy and the New Economics of Information. *Harvard Business Review*, September-October, 1997.

Everett, C.R. Group membership, relationship banking and loan default risk: the case of online social lending. *Social Science Research Network*, March 2010, 4-5.

Fix, E. & Hodges, J.L. (1951). Discriminatory analysis. Nonparametric discrimination: small performance. Report Number 11, pp.280-322.

Freedman, S. & Jin, G.Z. (2008). Do Social Networks Solve Information Problems for Peer to Peer Lending? Evidence from Prosper.com

Galoway, I. Peer to Peer Lending and Community Development Finance. *Community Investments*, Winter 2009/2010, Vol. 21, issue 3.

Gartner. Gartner says social banking platforms threaten traditional banks for control of financial relationship. Egham, UK. January 5, 2010.

Greiner, M.E. & Wang, H. (2009). The role of social capital in people to people lending marketplaces. *Proceedings of the 3rd International Conference on Information Systems*, Phoenix, pp.1-17.

Gutiérrez, M. (2009) Las microfinanzas: El Sistema financiero en Guatemala, CEPAL, Santiago de Chile.

Han, L. & Zhao, H. (2013), Orthogonal support vector machine for credit scoring. Engineering Applications of Artificial Intelligence, 26, pp-848-862.

Herzenstein, M; Andrews, R.L; Dholakia, U. & Lyandres, E. (2008). The democratization of personal consumer loans? Determinants of success in online peer to peer lending communities.

Herzenstein, M; Dholakia, U.M. & Andrews, R. Strategic herding behavior in peer to peer loan auctions. *Journal of Interactive Marketing*, 25, 1, pp.-27-36.

Hosmer, D.W. & Lemeshow, S. (1989). Best subsets logistic regression. *Biometrics*, 45, pp. 1265-1270.

Hunt, E.B; Marin, J. & Stone, P.J. (1966) Experiments in induction. New York: Academic Press.

Iyer, R; Khwaja, A.I. & Luttmer, E.F.P. (2009). Screening in new credit market: Can individual lender infer borrower creditworthiness in peer to peer lending? *Management*. Cambridge, MA.

Iyer, R; Khwaja, A.I; Luttmer, E.F.P. & Shue, K. (2014). Screening peer s softly: Inferring the quality of small borrowers. *Harvard University.*

Johnson, S; Arvind, A. & Assadi, D. (2010). Online or offline? : The rise of peer to peer lending in microfinance. *Journal of Electronic Commerce in Organizations* (8:3), pp.26-37.

Kauffman, R.J. & Riggins, F.R. Information and communication technologies and the sustainability of microfinance. *Electronic Commerce Research and Applications,* 10, 5, 2012.

Klafft, M. (2008). Online peer to peer lending: A lenders´ perspective. *Proceedings of the international conference on E-learnings, E-business, enterprise information systems and E-government, EEE*, pp. 371-375.

Katherine, A.K. & Sergio, H. (2009). Lending behavior and community structure in an online peer to peer economic network, *Proceedings of the 2009 International Conference on Computational Science and Engineering*, Vancouver, Canada, pp.613-618.

Lee, E; Lee, B. & Chae, M. Herding Behavior In Online P2P Lending: An Empirical Investigation. *Electronic Commerce Research and Applications,* Vol. 11, issue 5. October 2012. 495-503.

Leland, H.E. & Pyle, D. (1977). Informational asymmetries, financial structure and financial intermediation, *J. Finance* 32, pp.371-387.

Lin, M.F; Prabhala, N.R. & Viswanathan, S. (2009). Social networks as signaling mechanisms: Evidence from online peer to peer lending.

Lin, M.F; Prabhala, N.R. & Viswanathan, S. (2011). Judging borrowers by the company the keep: Friendship networks and information asymmetry in online peer to peer lending. *Social Science Research Network*, July 1, 2011.

Lin, M.F; Prabhala, N.R. & Viswanathan, S. (2012). Judging borrowers by the company they keep: Social networks and adverse selection in online peer to peer lending, *Management Science*.

Lopez, S.H. (2009). Social interactions in P2P lending. *In: Proceedings of the 3rd Workshop on Social Network Mining and Analysis*, NY, USA.

Magee, J. Peer to peer lending in the United States: surviving after Dodd-Fran. *North Caroline Banking Institute Journal*, 15, 2011, pp.139-174.

Martin, D. (1977). Early warnings of bank failure: A logit regression approach. *Journal of Banking and Finance, 1, 249-276.*

Ohlson, J.A. 1980. Financial ratios and Probabilistic Prediction of Bankruptcy. *Journal of Accounting Research* 18 (1), PP.109-131.

Payne, J; Bettman, J.R. & Johnson, E.J. (1993). The Adaptive Decision Maker, *Cambridge University Press*, New York

Perez Rubido, R. A review of feature selection algorithms that treat the microarray data redundancy. *Revista Cubana de Ciencias Informáticas*, Vol. 7, nº4 (2013), 23-24.

Pérez, S. & Santín, D. Minería de datos. Técnica y herramientas. Thomson, 2007.

Puro, L; Teich, J.E; Wallenius, H. & Wallenius, J. (2010). Borrower decision aid for people to people lending, *Decision Support System* (49), pp.52-60.

Ramser, J. & Foster, L. A demonstration of ratio analysis. Bulletin nº 40, University of Illinois, *Bureau of Business Research*, Urbana.

Rayo Cantón, S; Lara Rubio, J; y Camino Blasco, D. A credit Scoring Model for Institutions of Microfinance under the Basel II Normative. *Journal of Economics, Finance and Administrative.* Vol. 15 nº 28, Jun 2010.

Rose, S. (2007) The Prosper Lender Rebellion and the US credit/borrowing black hole, *P2P foundation*.

Schebesch, K.B. & Stecking, R. (2005). Support vector machine for classifying and describing credit applicants: Detecting typical and critical regions. *Journal of the Operational Research Society*, 56, pp.1082-1088.

Serrano Cinca, C. & Gutiérrez Nieto, B. Partial Least Square Discriminant Analysis for bankruptcy prediction. *Decision Support Systems* 54 (2013) 1245-1255.

Shen, D; Krumme, C. & Lippman, A. Follow the profit or the herd? Exploring social effects in peer to peer lending. *In Proceedings of the 2010 IEEE Second International Conference on Social Computing*. August 2010, pp.137-144

Slavin, B. Peer to peer lending: an industry insight. BradSlavin.com, June 21, 2007.

Sviokla, J. Breakthrough Ideas: Forget Citibank - Borrow form Bob. *Harvard Business Review*. 2009, Feb.

Tversky, A. & Kahneman, D. (1974). Judgment under uncertainty: heuristics and biases. *Science, New Series*, Vol. 185 nº 4157, pp.1124-1131.

Wang, H. & Greiner, M. (2011). Prosper: the eBay for money in lending 2.0. *Communications of the Association for the Information Systems*, 29, pp. 243-258.

Wang, H; Greiner, M. & Aronson, J.E. (2009). People to people lending: The emerging e-commerce transformation of a financial market, *Value Creation in E-Business Management* (36), pp.182-195.

Yu, L. & Liu, H. Efficient feature selection via analysis of relevance and redundancy. *Journal of machine learning research* 5 (2004).

Zacharakis, A. L. & Meyer, G.D. (2000). The potential of actuarial decision models: Can they improve the venture capital investment decision? *Journal of Business Venturing*, 15, 323.346.

Zhang, J. & Liu, P. Relational herding in microloans markets. *Management Science*, 58, 5, 2012, pp.892-912.

Zhou, L; Lai, K. & Yu, L. (2010). Least squares support vector machines ensemble models for credit scoring. *Expert Systems with Applications*, 37, pp.127-133.

ANEXOS

ANEXO I: variables proporcionadas por la plataforma de préstamos P2P Lending Club

Variable	*Escala*	*Descripción*
id	NUM	número asignado por LC para cada préstamo
member_id	NUM	número asignado por LC para cada prestatario
loan_amnt	USD	cantidad del préstamo solicitado por el prestatario
funded_amnt	USD	cantidad total del préstamo comprometida en ese momento de tiempo
funded_amnt_inv	USD	cantidad total del préstamo comprometida por inversores en ese momento de tiempo
term	NUM	número de pagos (en meses) del préstamo. 36 o 60
int_rate	%	tasa de interés del préstamo
installment	USD	pagos mensuales adeudados por el prestatario
grade	A-G	grado del préstamo asignado por LC
sub_grade	A1-G5	subgrado del préstamo asignado por LC
emp_title		nombre de la compañía donde está actualmente trabajando el prestatario
emp_length	<1-+10 years	duración del empleo en años
home_ownership	LETRA	estatus de la propiedad de la vivienda proporcionado por el prestatario. Posibles valores: rent, own, mortage, other
annual_income	USD	ingresos anuales proporcionados por el prestatario
verification_status		indica si los ingresos están verificados por LC
issue_d	DD/MM/AAAA	fecha en la que el préstamo es financiado

loan_status	LETRA	estatus actual del préstamo. Posibles valores: charged off, default, fully paid, current, in grace period, late, does not meet the credit policy
pymnt_plan		indica si un plan de pagos se ha puesto en práctica
url		URL de la página de LC
desc		descripción del préstamo proporcionada por el prestatario.
purpose	LETRA	categoría dada por el prestatario en la petición del préstamo. Posibles valores: debt_consolidation, medical, wedding, home_improvement, renewable_energy, small_business, vacation, moving, house, car, major_purchase, credit_card, educational, other.
tittle		título del préstamo dado por el prestatario
zip_code		3 primeras letras del código postal proporcionado por el prestatario
addr_state		dirección proporcionada por el prestatario durante el proceso
dti	%	ratio deuda/ingreso calculado utilizando los pagos mensuales de la deuda total, excluyendo las obligaciones hipotecarias dividido por los ingresos mensuales dados por el prestatario
delinq_2yrs	NUM	días por encima de los 30 días de morosidad en el historial crediticio del prestatario en los últimos 2 años
earliest_cr_line	DD/MM/AAAA	fecha en la que el prestatario informo de la apertura de una línea de crédito

inq_last_6mths	NUM	número de consultas de los acreedores en los últimos 6 meses
mths_since_last_delinq	NUM	meses desde la última morosidad del prestatario
mths_since_last_record	NUM	meses desde el último registro público
open_acc	NUM	cifra de líneas de crédito abiertas en el historial crediticio del prestatario
pub_rec	NUM	número de registros públicos despectivos
revol_bal	NUM	cantidad total de crédito rotatorio
revol_util	%	tasa de utilización del crédito rotatorio, o cantidad de crédito utilizada por el prestatario en relación con el crédito rotatorio disponible
total_acc	NUM	número de líneas de crédito abiertas actualmente en el historial crediticio del prestatario
initial_list_status	LETRA	estado inicial del listado del préstamo. Posibles valores: w,f
out_prncp	USD	restante principal pendiente de la cantidad financiada total
out_prncp_inv	USD	restante principal pendiente de la cantidad financiada total por inversores
total_pymnt	USD	pagos recibidos hasta la fecha del importe total financiado
total_pymnt_inv	USD	pagos recibidos hasta la fecha de la parte total financiada por inversores
total_rec_prncp	USD	principal recibido hasta la fecha por parte del prestatario
total_rec_int	USD	intereses recibidos hasta la fecha por parte del prestatario
total_rec_late_fee	USD	comisiones por pagos tardíos la fecha
recoveries	USD	recuperación bruta de pagos adeudados y que eran poco probables de ser recuperados

collection_recovery_fee	USD	recaudación de comisiones ante pagos adeudados y poco probables de ser recuperados
last_pymnt_d	DD/MM/AAAA	última fecha en la que el prestatario realizó un pago
last_pymnt_amnt	USD	última cantidad de pago recibida
next_pymnt_d	DD/MM/AAAA	próxima fecha programada de pago
last_credit_pull_d	DD/MM/AAAA	última fecha en la que LC realizó una consulta crediticia para el préstamo del prestatario
collections_12_mths_ex_med	NUM	número de recobros de impagos en los doce últimos meses sin tener en cuenta recobros de impagos médicos
mths_since_last_major_derog	NUM	meses desde 90 días o peor calificación
policy code	NUM	indica si la información está disponible públicamente. Valor 1 si lo esta

ANEXO II: transformaciones realizadas a las variables:

Variable	*Procedimiento*
loan_amnt	a partir de esta variable se crea la variable *loan_amnt-funded_amnt,* como diferencia entre lo solicitado y lo financiado, y la misma diferencia en porcentaje con *%(loan-amnt-funded_amnt)*
funded_amnt_inv	eliminada, ya que son valores idénticos en su mayoría a funded_amnt
int_rate	se añade el cuadrado del tipo de interés, *2_int_rate*
term	eliminada debido a la utilización única de préstamos a 36 meses
grade	se crea la variable *grade_ord:* si el grado es A, se impone un 7, si es B, un 6…hasta 1 si es G. Además se crean las variables dicotómicas *grado_a, grado_b*…hasta *grado_g* que toma valores 1 o 0 dependiendo del grado al que pertenece el préstamo.
sub_grade	se establece la variable *subgrade_ord,* colocando, si el préstamo es A1, un 35, A2, un 34,…2 si es G4 y finalmente 1 si es G5. Se añade además el cuadrado del subgrado, *2_subgrade,* el cuadrado del subgrado por el tipo de interés, *2_subgrade_int_rate* y el cuadrado del tipo de interés por el subgrado, *subgrade_2_int_rate*
emp_title	eliminada al no proporcionar información alguna
emp_length	se crea la variable *emp_lenght_bien*, poniendo un 0 si lleva trabajando menos de un año, 1 si lleva un año…hasta 10 si el periodo es superior a diez, eliminando emp_length
home_ownership	se establecen las variables dicotómicas *home_ownership_own*, *home_ownership_rent*, *home_ownership_mortage* y *home_owership_none* y *home_ownership_other* y la variable *home_ownership_ord* poniendo de 5 a 1, siendo 5 own, 4 mortgage, 3 rent, 2 other y 1 none
annual_income	a partir de esta variable se crean: *pagosprestamo_annual_inc* de tal forma, 12*installment/annual_income y *funded_amnt/annual_inc*
verification_staus	se establecen dos dicotómicas; *inc_v_si* e *inc_v_no,* indicando si los ingresos están verificados o no.

loan_status	se añade la variable *loan_status_ord* con valores de 7 a 1 de la siguiente forma: 7 fully paid, 6 current, 5 does not meet the credit policy, 4 in grace period, 3 late, 2 default y 1 charged off. Además se crean las siguientes variables dicotómicas según el estado del préstamo: *loan_staus_fullypaid, loan_staus_current, loan_status_late, loan_staus_doesnotmeet, loan_status_ingrace, loan_status_default y loan_status_chargedoff*
purpose	se crea a partir de esta variable *purpose_nominal* indicando lo siguiente: 1 si la finalidad es car, 2 credit card, 3 debt consolidation, 4 educactional, 5 home improvement, 6 house, 7 major purchose, 8 medical, 9 moving, 10 other, 11 renewable energy, 12 small business, 13 vacation, 14 wedding y las variables dicotómicas dependiendo de la finalidad del préstamo: *purpose_car, purpose_credit_card, purpose_other, purpose_house, purpose_debt_consolidation purpose_renewable_energy, purpose_educational, purpose wedding, purpose_moving, purpose_vacation, purpose_home_improvement, purpose_majorpurchase, purpose_medical y purpose_small_business*
pymnt_plan	eliminada
url	eliminada debido a que no proporciona ninguna información para el software Tanagra
desc	eliminada debido a que no proporciona ninguna información para el software Tanagra
title	eliminada debido a que no proporciona ninguna información para el software Tanagra
zip_code	eliminada debido a que no proporciona ninguna información para el software Tanagra
earliest_cr_line	a partir de esta variable, añadimos *days_historial,* calculado como la diferencia entre la fecha actual tomada como referencia y earliest_cr_line, la cual eliminamos
mths_since_last_delinq	eliminada al no disponer de una muestra lo suficientemente grande
mths_since_last_record	eliminada al no disponer de una muestra lo suficientemente grande

revol_util	eliminada
total_rec_prncp	a partir de esta variable añadimos otra: *% devuelto*, como el ratio total_rec_prncp/funded_amnt
last_pymnt_d	eliminada
next_pymnt_d	eliminada
mths_since_last_major_derog	eliminada

Las variables indicadas en el *Anexo I* no mencionadas en este segundo, no sufren modificación alguna.

ANEXO III: muestra de todos los préstamos concedidos, pagados e impagados, tanto para los posibles utilizados como *train* como para los del *test*.

Train

fecha	préstamos concedidos	fully paid	charged off or default
julio 2007	*63*	*30*	*0*
agosto 2007	*74*	*26*	*7*
septiembre 2007	*53*	*15*	*3*
octubre 2007	*105*	*37*	*10*
noviembre 2007	*112*	*30*	*7*
diciembre 2007	*171*	*66*	*17*
enero 2008	*169*	*140*	*30*
febrero 2008	*169*	*149*	*24*
marzo 2008	*234*	*196*	*40*
abril 2008	*152*	*128*	*26*
mayo 2008	*71*	*61*	*10*
junio 2008	*65*	*59*	*7*
julio 2008	*141*	*66*	*17*
agosto 2008	*96*	*65*	*5*
septiembre 2008	*54*	*27*	*5*
octubre 2008	*116*	*81*	*14*
noviembre 2008	*202*	*153*	*31*
diciembre 2008	*248*	*190*	*33*

Test

fecha	préstamos concedidos	fully paid	charged off or default
enero 2009	*269*	*211*	*28*
febrero 2009	*302*	*226*	*33*
marzo 2009	*324*	*245*	*30*
abril 2009	*333*	*250*	*39*
mayo 2009	*359*	*277*	*42*
junio 2009	*406*	*313*	*41*

julio 2009	*411*	*327*	*44*
agosto 2009	*446*	*368*	*39*
septiembre 2009	*507*	*392*	*56*
octubre 2009	*604*	*458*	*84*
noviembre 2009	*662*	*519*	*80*
diciembre 2009	*658*	*535*	*61*
enero 2010	*662*	*512*	*76*
febrero 2010	*682*	*563*	*61*
marzo 2010	*828*	*668*	*66*
abril 2010	*912*	*745*	*80*
mayo 2010	*807*	*664*	*87*
junio 2010	*639*	*538*	*63*
julio 2010	*784*	*677*	*61*
agosto 2010	*716*	*607*	*61*
septiembre 2010	*732*	*605*	*62*
octubre 2010	*732*	*596*	*66*
noviembre 2010	*765*	*650*	*51*
diciembre 2010	*806*	*795*	*63*
enero 2011	*949*	*807*	*93*
febrero 2011	*831*	*723*	*63*
marzo 2011	*837*	*721*	*73*
abril 2011	*947*	*828*	*90*
mayo 2011	*957*	*811*	*107*
junio 2011	*1258*	*1078*	*126*
julio 2011	*1288*	*1124*	*129*
agosto 2011	*1295*	*1108*	*135*
septiembre 2011	*1367*	*1161*	*142*
octubre 2011	*1411*	*1213*	*143*
noviembre 2011	*1543*	*1348*	*52*
diciembre 2011	*1415*	*1194*	*194*
enero 2012	*2058*	*1727*	*276*
febrero 2012	*2083*	*1735*	*276*
marzo 2012	*2355*	*1997*	*279*
abril 2012	*2680*	*2258*	*323*
mayo 2012	*2741*	*1984*	*360*
junio 2012	*3048*	*1424*	*442*

julio 2012	*3780*	*1749*	*520*
agosto 2012	*4461*	*2006*	*549*
septiembre 2012	*5080*	*2149*	*576*
octubre 2012	*5202*	*2166*	*570*
noviembre 2012	*5174*	*2079*	*544*
diciembre 2012	*5803*	*1880*	*439*
enero 2013	*5536*	*2172*	*515*
febrero 2013	*6011*	*2180*	*571*
marzo 2013	*6440*	*2367*	*563*
abril 2013	*7237*	*2610*	*588*
mayo 2013	*7770*	*2606*	*648*
junio 2013	*8127*	*2565*	*613*
julio 2013	*8977*	*2573*	*650*
agosto 2013	*9181*	*2618*	*590*
septiembre 2013	*9158*	*2526*	*537*
octubre 2013	*9981*	*2581*	*491*
noviembre 2013	*10864*	*2607*	*515*
diciembre 2013	*11088*	*2588*	*490*
enero 2014	*11493*	*2357*	*442*
febrero 2014	*10942*	*2031*	*391*
marzo 2014	*11634*	*1976*	*388*
abril 2014	*13264*	*2135*	*398*
mayo 2014	*13107*	*1952*	*297*
junio 2014	*11508*	*1528*	*211*
julio 2014	*19938*	*2339*	*279*
agosto 2014	*13281*	*1290*	*165*
septiembre 2014	*7369*	*644*	*66*
octubre 2014	*26570*	*1869*	*154*
noviembre 2014	*16572*	*828*	*37*
diciembre 2014	*6871*	*343*	*5*
enero 2015	*23629*	*811*	*6*
febrero 2015	*16022*	*361*	*0*
marzo 2015	*16914*	*264*	*0*