Clasificación de créditos utilizando máquinas de soporte vectorial sobre la base de datos de LendingClub

Credit classification using support vector machines on the LendingClub database

Karen Estefanía Guevara-Díaz*

* Magíster en Finanzas. Trader. Banco Davivienda S.A. [keguevar@davivienda.com], [0000-0001-6360-7377].

Artículo recibido el 01 de febrero de 2020.
Aceptado el 01 de marzo de 2020.

Para citar este artículo:

Guevara-Díaz, K. E. (2020). Clasificación de créditos utilizando máquinas de soporte vectorial sobre la base de datos de LendingClub. ODEON, 18, 59-98.
DOI: https://doi.org/10.18601/17941113.n18.03

Resumen

Se presenta la teoría de máquinas de soporte vectorial (Support Vector Machine – SVM) aplicada a la clasificación de créditos otorgados por la banca *fintech* (*financial technology*) de Estados Unidos LendingClub. Se estudiará la clasificación que da el método SVM a los créditos contra la ya otorgada por la entidad. Se analizan las variables más importantes que maneja LendingClub para el otorgamiento de créditos comparándolas con la clasificación de pago o impago que arroja el experimento de SVM.

Palabras clave: riesgo de crédito; máquinas de soporte vectorial; *fintech*.

Clasificación JEL: C63, G21.

Abstract

The theory of support vector machines applied to the classification of credits granted by the United States fintech banking LendingClub is presented. The classification given by the SVM method to credits will be analyzed against what has already been granted by the entity. The most important variables that LendingClub manages for the granting of credits are analyzed, comparing it with the classification of payment or non-payment that the SVM experiment gives.

Key words: Credit risk; vector support machines; fintech.

JEL classification: C63, G21.

Introducción

En la práctica financiera, la clasificación de clientes entre buenos y malos es importante para el buen funcionamiento de los establecimientos de crédito. El riesgo de crédito se asocia a la "probabilidad de pago de un agente al que se le otorga el crédito" (Moreno y Melo, 2011, p. 2). Así, el problema de riesgo de crédito puede acotarse a un problema de pago o impago de los agentes/clientes.

Analizar el riesgo de crédito es significativo porque es el primer factor para proteger la solvencia de las entidades financieras. La importancia de las entidades financieras se funda en su papel de proveedor de liquidez y transformador de plazos en la sociedad. Al buscar modelos que ayuden a mejorar la clasificación de aquellos agentes que pagan y los que no pagan, mejoramos la solvencia de los bancos y, a su vez, protegemos la economía de un colapso de liquidez.

Esta búsqueda de calidad crediticia no es ajena a las nuevas entidades financieras que enfocan su colocación de crédito en el análisis de variables novedosas que se extraen netamente a través de sistemas digitales. Sobre la novedad de variables para el análisis de crédito se hablará más adelante en este trabajo.

En este artículo se analiza la base de datos de clientes de la compañía estadounidense de préstamos entre particulares llamada LendingClub, aplicando la metodología de máquina de soporte vectorial (Support Vector Machine – SVM) para clasificar entre "buenos y malos" los clientes y asignarles una clase crediticia.

El interés de analizar a los clientes de LendingClub se debe a que es la primera compañía de préstamos *peer to peer* (P2P, préstamos entre particulares) en listarse en la Securities and Exchange Comission en el año 2013. Su modelo de negocios de desintermediación, acerca a inversionistas y prestamistas a un mismo objetivo de eficiencia en los recursos. Los mecanismos P2P aproximan a quienes tienen excedentes de liquidez a aquellos que necesitan de esos excedentes. Así, a través de una plataforma tecnológica (el lugar de encuentro) y sin las barreras de la banca tradicional, se unen dos grupos de interés para darle solución a su problema. Por un lado, las bajas tasas de remuneración en inversiones como cuentas de ahorro o depósitos a término y, por otro, la necesidad de crédito.

No obstante, el primer grupo de interés de este negocio –quienes tienen el exceso de liquidez– necesita prestar mayor atención a las tasas de *default* de los créditos, tanto como lo hacen los bancos comerciales para evitar el colapso del negocio. Así, al aplicar SVM se quiere resolver si la clasificación con la metodología SVM clasifica mejor los créditos presentados en el *data set* original de la compañía LendingClub.

El presente documento consta de cinco partes. Primero, se abordará la importancia del riesgo de crédito y se hará una breve descripción de metodologías en esta área de uso habitual. Segundo, se describirá la metodología SVM. Tercero, se explicarán las particularidades de los datos de clientes que publica LendingClub en su página web. Cuarto, se aplicará la metodología SVM al problema de clasificación de crédito. Por último, se presentarán las conclusiones de este experimento.

1. Importancia del riesgo de crédito

El riesgo de crédito es el riesgo de impago de dinero como resultado de la imposibilidad de honrar la obligación de crédito por parte del acreedor (Bessis,

2015, p. 213). El problema de la posibilidad de que se materialice el impago es el riesgo que se deriva para el prestador de no recuperar el capital prestado, los intereses que honraban la deuda y el aumento de costos asociados a recuperarla.

De manera tradicional, el riesgo de crédito ha sido tratado como la evaluación del agente acreedor y su capacidad de pago medido por el flujo de caja. Sin embargo, a lo largo de la historia financiera, son los apretones de liquidez derivados de los impagos de créditos lo que ha llevado a la extinción de entidades financieras sólidas. El caso más renombrado de la era reciente data de 2008, con la quiebra del banco de inversión Lehman Brothers. No obstante, su nombre se toma en este documento solo como referencia y no hace parte del análisis.

El desarrollo de la tecnología en la última década y los avances computacionales generaron la expansión de los mecanismos de crédito. Ante esto, el análisis crediticio entró en una era de cuantificación y modelación computacional. Entre los desafíos que enfrenta el riesgo de crédito tenemos:

- Existe un problema: no hay un patrón único por seguir en la evaluación de factores de crédito. Por eso, se desarrolla la robótica y se da la oportunidad de aprendizaje a las máquinas para analizar estos patrones.
- Nos enfrentamos a volúmenes masivos de información. Así que, al existir muchos datos, analizar esa información se vuelve difícil.
- Las máquinas de aprendizaje –una de las cuales hace parte del objeto de este artículo, la SVM– pueden aplicarse para dar solución a este problema de volumen de datos y patrones no únicos.

En este artículo abordaremos las características y el análisis de la clasificación de riesgo de crédito de LendingClub. Esta entidad es originaria de Estados Unidos y hace préstamos basados en el análisis crediticio de múltiples patrones no convencionales a los usados de manera tradicional por la banca de personas.

El análisis de crédito tradicional se basa en medir la probabilidad de pérdida como un factor de incumplimiento basado en el estudio de un balance[1] estático del agente al que se le prestan los recursos. Este punto de vista se ha transformado en una investigación que cambia el supuesto anterior, donde el prestatario y el prestamista permanecen en el mismo estado crediticio durante toda la vida de la

1 Balance: activos = pasivos + patrimonio en una fecha dada.

deuda. Somos conscientes de que, por el contrario, a medida que pasa el tiempo, las condiciones de préstamo y pago de las obligaciones pueden también variar de forma más segura o más riesgosa. Un estado crediticio se caracteriza por la probabilidad de pago o incumplimiento, y esta probabilidad cambia a medida que el prestatario migra hacia otros estados de crédito.

El riesgo de crédito es medido de forma tradicional como una *scoring* de crédito, en donde de acuerdo con las variables analizadas, en su mayoría de casos el flujo de caja en un momento del tiempo, se da una calificación más o menos positiva para otorgar crédito. Los estados de crédito están ampliamente documentados por las transiciones observadas de las calificaciones crediticias de un determinado periodo a medida que pasa el tiempo. Pero en el modelo tradicional de *scoring*, el cambio de los patrones no se analiza y esto lleva a que la medición de riesgo no sea en ocasiones eficiente.

Con el avance de la tecnología podemos hacer uso de herramientas de análisis de datos de grandes volúmenes. El desarrollo y avance de las máquinas de autoaprendizaje da la posibilidad de evaluar grandes volúmenes de información de forma rápida y eficiente para entregar soluciones puntuales. En este caso se estudiará el método de SVM para el análisis de los datos.

2. Explicación de la metodología de máquinas de soporte vectorial

La SVM es una herramienta desarrollada dentro del campo de *machine lerning* (máquinas de aprendizaje), creada por Vladimir Vapnik en 1995. En la última década, el campo de *machine learning* ha llegado a usarse de manera masiva dado el avance computacional que nos permite, en menor tiempo, aumentar la cantidad de datos por analizar e implementar en cada experimento.

Entre las principales ventajas de esta metodología están (Moreno y Melo, 2011, p. 8):

- Es uno de los modelos más exitosos en *machine lerning* ya que requiere menos supuestos sobre los datos. Por ejemplo, no asume normalidad o continuidad.
- El desempeño del modelo no está ligado al volumen de datos que debe incluir.
- Al igual que las teorías generales de solución única, como por ejemplo Black and Scholes, la metodología plateada por Vapnik se resuelve mediante programación cuadrática, lo que genera una solución única.

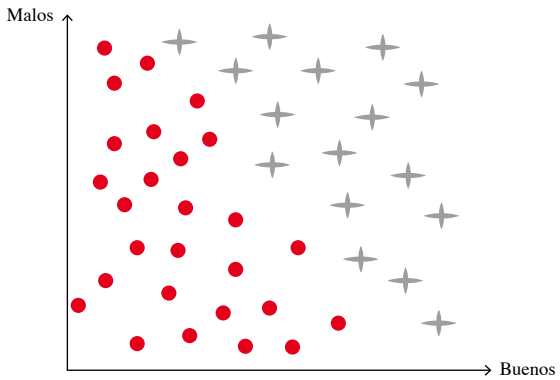
- El algoritmo puede programarse para soluciones tanto lineales como no lineales a través de transformaciones de Kernel.

Una SVM es transformar un grupo de datos de cualquier dimensión, en un plano lineal, para poder clasificarlos en grupos que nos arrojarán conclusiones sobre su comportamiento. Su ventaja es que sin importar la dimensión de la base de datos, ni la función que siga, estos datos pueden transformarse en un espacio lineal para resolver el problema de esa misma forma, arrojando una solución única.

Asumamos que tenemos un espacio de dos predictores, x_1 y x_2 que representan un problema de clasificación como puede ser pago (buenos, estrellas negras) o impago (malos, puntos rojos). Los predictores son tomados como dos grandes grupos de los datos por analizar (figura 1).

La SVM fue creada como una máquina de aprendizaje para problemas de clasificación de dos grupos. Sin embargo, a través del concepto de Hyperplano, que se describe más adelante, el problema de clasificación de dos grupos se extiende a más grupos. La idea central de esta teoría es hacer un mapeo de los predictores de algún espacio de características de alta dimensión (gran número de datos), para luego arrojar una solución lineal con propiedades especiales para ese mapeo que se realizó, a fin de asegurar una alta generalización de los datos analizados.

Figura 1: Representación de predictores buenos y malos



Fuente: elaboración propia.

No obstante, la generalización de datos a través de una predicción lineal crea el problema de la separación lineal, dado que los datos pocas veces pueden separarse de esta forma. Asimismo, el análisis de grandes volúmenes de datos y su separación en varias dimensiones enfrenta a esta teoría ante el problema de construir ecuaciones de polinomios tan amplios que podrían llevar a formar millones de dimensiones.

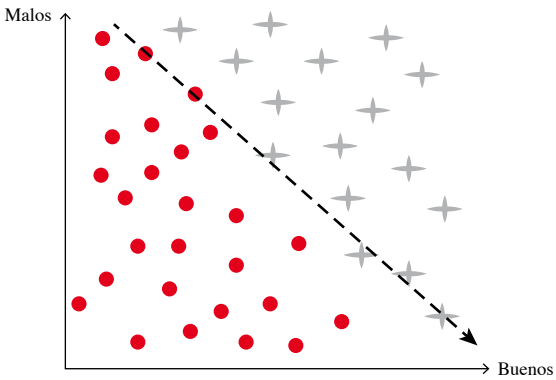
Acá es donde el autor del método, Vapnik, propone separar los predictores mencionados anteriormente, en hiperplanos. Un hiperplano es definido por el autor como "la función de decisión lineal con máximo margen entre los vectores que clasifican los tipos de predictores".

$$\beta_0 + \beta_1 X_1 + \beta_2 X_2 = 0$$

Se observó que, para construir tales hiperplanos, solo se debe tener en cuenta una pequeña cantidad de los datos de entrenamiento, llamados vectores de soporte, que determinan este margen (Vapnik y Cortés, 1995).

Si el hiperplano de separación de los datos es óptimo, entonces el vector de separación será la mejor frontera para clasificar entre dos o más clases estos datos. Asimismo, se reduce la probabilidad de cometer un error de clasificación de los datos (figura 2).

Figura 2: Margen de separación óptimo ilustrativo



Fuente: elaboración propia

Hasta este momento parece sencillo el concepto de las SVM. Sin embargo, ¿qué sucede cuando los predictores o grupo de datos son más de dos? Se debe producir una clasificación de varios grupos en varias dimensiones que hace engorrosos los cálculos. La novedad del modelo SVM es su eficiencia en transformar múltiples dimensiones en un espacio lineal para arrojar una solución.

2.1. Definición del hiperplano

Un hiperplano es una función que maximiza el margen de separación de un determinado conjunto de datos.

En un espacio de dimensión *p* (p-dimensional), el hiperplano será el plano más afín a la dimensión de los datos de esa dimensión *p*-1. En una sola dimensión, el hiperplano será una línea; en dos dimensiones será un plano separador de los datos del espacio; en tres dimensiones, el hiperplano será un subespacio de dos dimensiones. Más allá de tres dimensiones es difícil de imaginar, pero los computadores ayudarán en ese ámbito.

La definición matemática del hiperplano es:

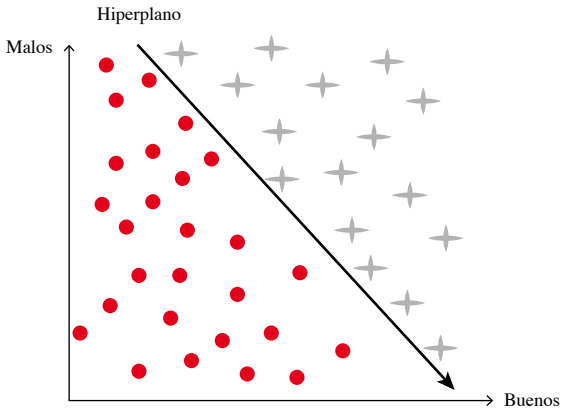
$$\beta_0 + \beta_1 X_1 + \ldots \beta_p X_p = 0$$

Los parámetros $\beta_0, \beta_1 \ldots$ definen el espacio del hiperplano. El resultado de las betas serán los coeficientes que definan la separación de los datos en el modelo. Esto significa que para todo $f(x) = (x_1, x_2)$ evaluado en la función del hiperplano, su resultado corresponderá a un punto en este. La primera definición arroja como resultado una línea recta dado que el primer ejemplo de hiperplano solo considera una dimensión (figura 3).

El hiperplano en el ejemplo más sencillo será una función lineal que divide el espacio en dos.

Ahora, ¿qué pasa si el conjunto de datos no puede ser separado por una línea y, por tanto, el modelo clasificador debe extenderse a más dimensiones para arrojar el mejor resultado posible?

Figura 3: Hiperplano separador



Fuente: elaboración propia.

En el caso de más dimensiones, el hiperplano deberá expandir sus betas para ilustrar ese mayor número de conjuntos. En más de dos dimensiones será definido como:

$$\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots \beta_p X_p > 0$$

En un gráfico donde los datos tengan una ubicación con más de dos dimensiones (por ejemplo, un cubo) la función debe extenderse para hacer la división óptima como es el caso de la ecuación anterior.

Todas las betas, $\beta_0, \beta_1, \beta_2 \ldots$ definen el espacio del hiperplano. Lo que significa que para cada observación $f(x) = (x_1, x_2)$ evaluada en la anterior función. El resultado se interpreta como un punto en el hiperplano. En este modelo, el interés está en evaluar si los datos están por encima del hiperplano, en el hiperplano o por debajo de él.

2.2. Clasificador de máximo margen

Si definitivamente los datos no pueden ser separados por un hiperplano de manera perfecta, usamos la idea de separarlos, pero desde el punto más lejano de la observación con el hiperplano. Con esto, se busca la distancia más pequeña del margen, o la división óptima, a los datos.

Si el clasificador es óptimo, entonces tendremos una función de este tipo:

$$f(x) = \left\{ \begin{array}{c} > 0, clase\,1. Clase\ superior \\ = 0 \\ < 0, clase - 1. Clase\ inferior \end{array} \right\}$$

El problema será encontrar el mejor hiperplano que clasifique los datos con más de dos dimensiones. Al evaluar con ese hiperplano las observaciones, dadas las características anteriores, el que logre maximizar la distancia a todos los puntos será el mejor hiperplano.

Asimismo, significa que el hiperplano que mejor separa todos los grupos de datos es el que tenga la mayor distancia acumulada a todos los puntos. A esto llamamos un clasificador de máximo margen.

Dado $\beta_0, \beta_1, \ldots \beta_p$
Maximizar el margen (M) sujeto a:

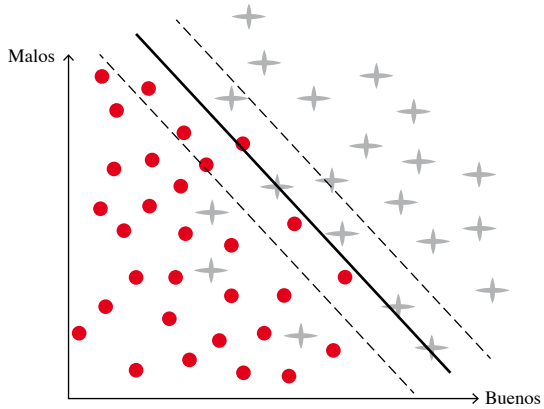
$$\sum_{j=1}^{p} B_j^2 = 1$$

$$Y_i(\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots \beta_p X_{ip} \geq M \textit{ para todo } i = 1, \ldots, n$$
(todos los β deben sumar 1)

Esto asegura que la solución sea cuadrática y el vector sea unitario y garantice la máxima distancia del vector que clasifica los datos contra el plano. En este caso, los puntos más importantes serán los que estén más cerca del hiperplano; estos los llamaremos puntos de soporte (figura 4).

Esto quiere decir que cuando se evalúa la función para cada punto y se tiene en cuenta que ese resultado puede dar positivo o negativo, la segunda restricción garantiza que la observación evaluada va a estar en el lado correcto del hiperplano.

Figura 4: Margen máximo de clasificación

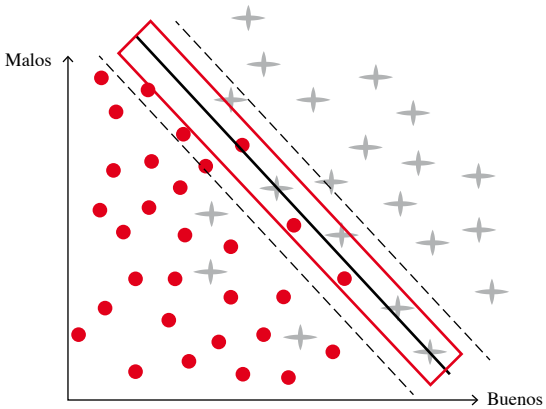


Fuente: elaboración propia.

Las condiciones de la maximización del hiperplano juntas aseguran que $f(x) = (x_1, x_2)$ esté correctamente clasificado en cada lado del hiperplano y que el conjunto de esos puntos tenga la mínima distancia de M desde el hiperplano.

En caso de que no haya solución de un solo hiperplano para la clasificación de los datos, entonces no podremos separar en dos exclusivas clases los datos por analizar (como se representa en la figura 4 con los datos dentro del margen). La solución a este problema se llama clasificador de vector de soporte y suaviza el concepto de hiperplano al dar un margen suave de clasificación a los datos (figura 5).

Figura 5: Clasificador de vector de soporte



Fuente: elaboración propia.

El modelo SVM tiene una baja variabilidad y es robusto al cambio de los datos; solo importa si los datos cerca del plano cambian. Entre más grande $f(x)$ más seguro está el modelo de que ese valor es de esa clase. Sin embargo, el problema del clasificador del máximo margen es que solo funciona si los datos son exactamente separables.

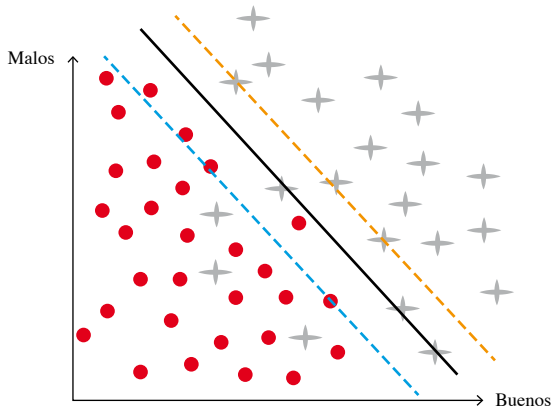
Para avanzar hacia la clasificación se busca ubicar la línea en la mejor separación posible. Esta línea o margen es la que maximiza la distancia a la observación más cercana. Se permitirá un margen de error con el clasificador de vector de soporte.

2.3. Clasificador de vectores de soporte

Cuando una selección de datos no es clasificable por un único hiperplano, en vez de restringir la clasificación de los datos a solo un vector, podemos suavizar el concepto introduciendo un margen suave a la clasificación que dará como resultado permitir algunos datos dentro de ese margen que no son correctamente clasificables (figura 6).

Azul: $h(x_1) = x * w + b = 1$
Rojo: $h(x_1) = x * w + b = -1$
Línea negra: representa el hiperplano.

Figura 6: Margen máximo de clasificación y puntos de soporte



Fuente: elaboración propia.

Puede suceder que una observación no solo esté en el lado equivocado del margen, sino que esté completamente equivocada frente a hiperplano. Este margen, al igual que el máximo clasificador de margen, tiene solución en la maximización de M, así:

Sea M el factor por maximizar dado $\beta_0, \beta_1, \ldots, \beta_p$
Sujeto a

$$\sum_{j=1}^{p} B_j^2 = 1$$

$$Y_i(\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_p X_{ip}) \geq M(1 - j)$$

- $i \geq 0, \sum_{i=1}^{n} \in i \leq C,$

- Variables flojas que permiten que las observaciones individuales estén en el lado incorrecto del margen del hiperplano.

Donde C es el número óptimo de errores permitidos en el margen que admiten los errores. El factor C se calcula a través de experimentación y es un factor de regularización.

En la medida que el parámetro C crece o disminuye en el ejercicio de maximización, el modelo permitirá más violaciones de la clasificación o las restringirá del todo llevando C a cero. Esto a su vez logrará que el modelo tenga un intercambio óptimo entre sesgo y varianza, lo que aumenta una de las dos medidas estadísticas dependientes de la flexibilidad de C.

Cuando C aumenta, el modelo tendrá alto sesgo, pero poca varianza; al contrario, si C es cercano a cero los vectores de soporte que clasifican los datos tendrán bajo sesgo, pero alta varianza.

Al igual que en el margen máximo de clasificación, lo que arroja la disposición del hiperplano debe ser más grande o igual al margen. El objetivo del problema de clasificación es encontrar la línea que hace que el margen se maximice. Se toma la distancia más pequeña al grupo de evaluaciones para que la evaluación contra el margen sea la máxima posible. El margen, que se interpreta como la distancia contra el hiperplano, tiene una última condición donde todas las betas deben sumar 1.

2.4. Máquina de soporte vectorial

Cuando las clasificaciones lineales se desempeñan de manera pobre ante un *data set* evidentemente no lineal, se debe considerar ampliar las características del espacio de los datos usando las funciones de los predictores como funciones cuadráticas o cúbicas para poder solucionar problemas de no linealidad.

Así, los predictores de margen de clasificación y el vector de soporte de clasificación que se describe como: $x_1, x_2, \ldots x_p$ pueden migrar a una forma con la siguiente característica: $x_1, x_1^2, x_2, x_2^2 \ldots x_p, x_p^n$. Esto es, transformar el espacio de datos de tal manera que el resultado de la transformación sea lineal.

Al transformar los datos en un espacio lineal, se puede aplicar el método de SVM. Si el cambio del espacio de datos es exitoso, se puede resolver el problema de clasificación aplicando el método de soporte de vector de clasificación. Este es el punto crítico del asunto de la transformación de los datos en ese espacio lineal.

Para deformar el espacio de los datos en las dimensiones que se necesitan, podemos usar una expansión del espacio de variables. El ejemplo sería el siguiente:

Asumiendo que se tiene un modelo con dos predictores y tres observaciones $f(x) = (x_1, x_2, x_3)$, tenemos que[2]:

$$\begin{gathered}\beta_0 + \beta_1 x_{11} + \beta_2 x_{12} \\ \beta_0 + \beta_1 x_{21} + \beta_2 x_{22} \\ \beta_0 + \beta_1 x_{31} + \beta_3 x_{32}\end{gathered}$$

Expandimos el espacio con tres observaciones, así: las funciones que antes teníamos vamos a expresarlas en función de una cantidad de términos que va a depender del número de observaciones. A esta operación se le llama producto punto:

$$\begin{gathered}\beta_0 + \propto_1 (x_{11}x_{11} + x_{12}x_{12}) + \propto_2 (x_{11}x_{21} + x_{12}x_{22}) + \propto_3 (x_{11}x_{31} + x_{12}x_{32}) \\ \beta_0 + \propto_1 (x_{21}x_{11} + x_{22}x_{12}) + \propto_2 (x_{21}x_{21} + x_{22}x_{22}) + \propto_3 (x_{31}x_{31} + x_{22}x_{32}) \\ \beta_0 + \propto_1 (x_{31}x_{11} + x_{32}x_{12}) + \propto_2 (x_{31}x_{21} + x_{32}x_{22}) + \propto_3 (+ x_{32}x_{32})\end{gathered}$$

Con la anterior transformación de producto punto hemos ampliado el espacio de las funciones que determinan los parámetros a un espacio lineal para poder

2 Predictores: función del hiperplano evaluada en tres puntos.

entender y solucionar el problema de la base de datos. Lo importante es que las betas y las alfas son lo mismo y al hacer expansión lineal se llega a esa equivalencia. Las betas serán una ponderación por alfa de cada observación.

Para los predictores que no son soporte, es decir que no están dentro de la banda o en la banda margen de las figuras anteriores, alfa es igual a cero. Al ser igual a cero, su función (la que evalúa si se está arriba o abajo del margen) no afectará el desempeño de la clasificación.

Las betas serán uno por cada predictor o función y las alfas serán una por cada observación. La conexión de ambos términos es la transformación lineal de los datos. Y si se quiere conectar a los dos términos, las betas serán una ponderación por alfa de cada observación.

Para los predictores que no son soporte, que no están dentro de la banda, el alfa es igual a cero. Eso significa que su función, la que evalúa si se está arriba o abajo, será igual a beta cero más la suma sobre las observaciones que pertenecen al soporte (la banda). Sin embargo, a medida que aumentan los predictores, los cálculos se vuelven más complejos. Al complicarse las operaciones matemáticas, llegamos a la expansión lineal de la *data set* a través de la aplicación de las expansiones de Kernel.

2.5. Expansión de Kernel (K)

La expansión de Kernel (K) es una función que cumple la propiedad de expandir un espacio de datos de varias dimensiones en un solo espacio lineal, ahorrando los cálculos de un factor punto a punto.

Es una generalización para calcular el producto punto entre dos vectores teniendo en cuenta una expansión. Permite calcular de manera rápida la operación sin necesidad de hacer todo el proceso algebraico.

Resumiendo lo anterior, las SVM son el clasificador de máximo margen el cual utiliza una transformación que convierte los datos de una base de datos en un espacio lineal ampliado. El éxito de este modelo es encontrar esa función Kernel (K) de expansión lineal que sea óptima y eficiente al problema de clasificación de la *data set* original. El secreto del éxito será encontrar la expansión de Kernel que mejor se adecue a los datos.

Entre las funciones de Kernel (K) más utilizadas están:

- Kernel lineal

$$K(x, x')=(x^* x')$$

El Kernel lineal es la abreviación de las operaciones punto a punto descritas anteriormente. Al aplicar un Kernel lineal a un experimento de SVM, tanto el clasificador SVM como el clasificador de vector de soporte serán el mismo.

- Kernel polinómico

$$K(x, x') = (x^* x' + c)^d$$

En el caso de que $d = 1$, y $c = 0$, el resultado de la operación anterior será igual al Kernel lineal. En caso de que $d > 1$ o $d < 1$, se producirán límites de decisión y el resultado será un margen. La no linealidad aumentará en el modelo a medida que aumenta el factor d. La literatura revisada en este artículo habla de no emplear este Kernel con un factor d mayor a 5 ya que podría causar problemas de sobreajuste.

- Kernel radial

$$K(x, x') = e(-\gamma \, ||x^* x'||^d)$$

Un Kernel radial busca clasificar unos datos con un tipo de formación circular. El valor de γ será el que determine la dimensión del Kernel. Un γ pequeño se asimilará a un Kernel lineal. A medida que γ aumenta hace más flexible el modelo.

3. Descripción de la base de datos de LendingClub

LendingClub fue creado en 2007 según datos de su página web. Está denominado como una banca *fintech* (que hace uso de la tecnología para prestar servicios bancarios) cuya promesa de valor es que los inversionistas y prestatarios tendrán mejores tasas de captación y colocación frente a la banca tradicional. Su negocio consiste en desintermediar la labor bancaria acercando más a los agentes con excedentes de liquidez a aquellos con necesidad de recursos.

La base de datos que puede obtenerse de la página web de LendingClub, y que se analizará, contiene información de personas que aplicaron y les fue aprobado

un préstamo. La compañía genera préstamos personales con montos máximos de 40.000 dólares. Para solicitar un crédito, los prestatarios deben tener unas características específicas como una calificación mínima en el puntaje FICO[3], una relación de solvencia del prestatario del 35 % y varios datos demográficos y personales por cada crédito. Entre estos podemos encontrar que piden las razones de solicitud de préstamos, el plazo en que se planea pagar la deuda, si se tiene bienes raíces o no, si ya se ha tomado una hipoteca, entre otras características.

En su página web[4], LendingClub permite acceso a la base de datos de los créditos aprobados y negados. Tiene en cuenta 120 variables entre cualitativas y cuantitativas para entregar una aprobación de crédito. Una vez el crédito tiene una respuesta es desembolsado en la cuenta que haya decidido el prestatario.

En este artículo, el objetivo es el análisis de 177.610 datos de un horizonte de tiempo de dos años (enero 2012 a diciembre 2013) donde la variable dependiente será la *loan status*. Esta variable es cualitativa y binaria, se presenta en términos de: "C" cumple con el préstamo o "D" incumple el préstamo. Sin embargo, por temas de capacidad de procesamiento de datos, el análisis de los datos se hará con toda la base, mientras que en el experimento de clasificación solo se tomarán 44.400 datos.

Asimismo, de las 121 variables que analiza LendingClub para la aprobación de crédito, solo se tomarán 22 que son las seleccionadas por Malekipirbazari (2015) en su estudio previo de clasificación supervisada[5].

En el grupo de 22 variable que se van a analizar, nuestra variable dependiente será *lending status*, las 21 variables restantes se dividen en 14 cuantitativas y 7 cualitativas. En principio, la base original es una base desbalanceada donde el 84,41 % de los datos de *loan status* pertenece a la clase de "C" cumple y tan solo el 15,59 % de los datos corresponde a la clase "D" incumple (tabla 1).

3 FICO score fue creada en 1989 por la compañía Fair, Isaac, and Company. Definen su producto como una solución analítica al *scoring* de crédito. Es una medida de riesgo para medir el riesgo de incumplimiento de una obligación financiera. Fuente: https://www.fico.com/en/about-us#our-company

4 www.lendingclub.com

5 Para mayor información sobre la selección de variables favor remitirse al autor.

Tabla 1: Porcentaje de datos "C" y "D" en la base de datos

	Loan Status	Participación (%)
C	149.923	84,41
D	27.687	15,59
Total	177.610	100

Fuente: elaboración propia

Debido a que las variables cualitativas deben medirse para analizar los datos en conjunto, con la base de datos se realizó un proceso de binarización de las variables cualitativas. Las variables cualitativas se transformarán en variables *dummies* en términos de 0 y 1 para su análisis, donde 0 será un dato negativo o "NO" y 1 será un dato positivo nombrado para temas de análisis como "SÍ".

La base con la que trabajamos corresponde a una base obtenida de un trabajo previo sobre LendingClub, analizando la misma variable dependiente *loan status* bajo el modelo de Gradient Boosting. Este experimento previo nos entregó dos bases, una primera de datos limpios y con el respectivo proceso de binarización de las variables cualitativas pero desbalanceada entre los créditos clasificados con "C" o con "D". De igual manera, otra base de datos balanceada al 50 % con datos de "C" cumple y 50 % con datos de "D" incumple. Con estas dos bases se realizará el experimento de clasificación supervisada para entregar conclusiones (tabla 2).

Tabla 2: Descripción y tipo de variables por analizar en el experimento

Variables	Descripción	Tipo de variable
loan_amnt	Monto del préstamo solicitado	Cuantitativa
Term	Número de pagos del crédito medido en meses	Cuantitativa
Income_To_payment	Porcentaje del ingreso que se destina al pago de la deuda	Cuantitativa
Grade	Calificación de crédito asignada por LendingClub: 1 mejor grado, 7 peor grado	Cuantitativa
emp_length	Antigüedad laboral	Cuantitativa
annual_inc	Ingreso anual reportado por el prestatario durante su registro	Cuantitativa

Variables	Descripción	Tipo de variable
dti	Razón calculada entre el total del pago mensual sobre el total de las obligaciones	Cuantitativa
delinq_2yrs	El número de moras mayor a 30 días en los últimos dos años	Cuantitativa
earliest_cr_line	El mes en que se abrió la primera línea de crédito informada del prestatario	Cuantitativa
inq_last_6mths	Número de solicitudes de los últimos 6 meses excluyendo hipotecas y crédito de vehículo	Cuantitativa
open_acc	El número de créditos abiertos por el prestamista en su historial	Cuantitativa
total_acc	El número de créditos actuales en el historial del prestamista	Cuantitativa
revol_util	La cantidad de crédito usado por el prestatario en relación con todo el crédito otorgado	Cuantitativa
revol_inco	La porción restante de crédito que el prestamista no está usando	Cuantitativa
home_ownership_MORTGAGE	El prestamista posee o no una hipoteca	Cualitativa
home_ownership_OWN	El prestamista posee vivienda propia	Cualitativa
home_ownership_RENT	El prestamista tiene vivienda arrendada	Cualitativa
purpose_credit_card	Propósito de la tarjeta de crédito	Cualitativa
purpose_debt_consolidation	Propósito de la consolidación de la deuda	Cualitativa
purpose_home_improvement	Propósito de las mejoras locativas	Cualitativa
purpose_other	Otro tipo de propósito del préstamo	Cualitativa
loan_status	Estado actual del préstamo	Cualitativa

Fuente: elaboración propia

El proceso de binarización de variables se realizó con 7 variables de las 22 por tomar en el experimento. Las variables antes descritas se presentan en la tabla 3 y tienen la siguiente estructura para su análisis y modelación:

Tabla 3: Estructura interna de los objetos en R

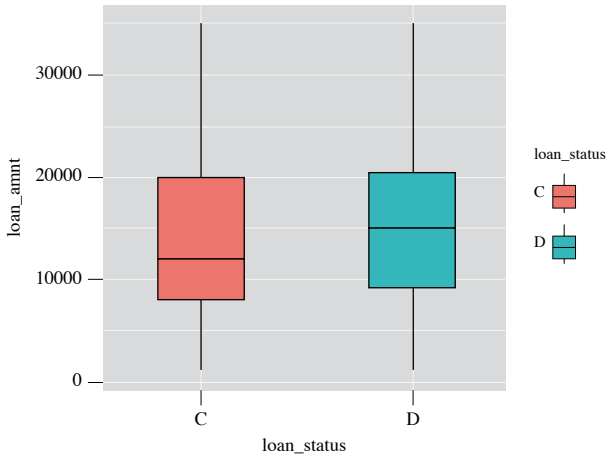
```
> str(DATOS)
Classes 'tbl_df', 'tbl' and 'data.frame':       177610 obs. of  22 variables:
 $ loan_amnt                 : num  12000 27050 12000 28000 27600 ...
 $ term                      : num  36 36 36 36 60 36 36 36 36 36 ...
 $ Income_To_payment         : num  0.0465 0.1932 0.0368 0.0322 0.1201 ...
 $ grade                     : num  7 6 6 7 4 6 3 6 3 6 ...
 $ emp_length                : num  3 10 10 5 6 4 10 10 1 2 ...
 $ annual_inc                : num  96500 55000 130000 325000 73000 60000 90000 40000 26000 39600 ...
 $ dti                       : num  12.6 22.9 13 18.6 23.1 ...
 $ delinq_2yrs               : num  0 0 0 0 1 0 1 0 0 0 ...
 $ earliest_cr_line          : num  37865 31686 35735 34639 32660 ...
 $ inq_last_6mths            : num  0 0 1 1 1 1 0 0 0 2 ...
 $ open_acc                  : num  17 14 9 15 10 15 9 7 12 3 ...
 $ total_acc                 : num  30 27 19 31 24 18 12 32 28 8 ...
 $ revol_util                : num  0.557 0.612 0.67 0.546 0.828 0.24 0.662 0.688 0.528 0.161 ...
 $ revol_inco                : num  1.647 7.994 0.997 1.092 4.439 ...
 $ home_ownership_MORTGAGE   : Factor w/ 2 levels "0","1": 2 1 2 2 2 1 2 1 1 2 ...
 $ home_ownership_OWN        : Factor w/ 2 levels "0","1": 1 2 1 1 1 1 1 1 1 1 ...
 $ home_ownership_RENT       : Factor w/ 2 levels "0","1": 1 1 1 1 1 2 1 2 2 1 ...
 $ purpose_credit_card       : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
 $ purpose_debt_consolidation: Factor w/ 2 levels "0","1": 2 2 2 2 2 2 1 2 2 1 ...
 $ purpose_home_improvement  : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 2 ...
 $ purpose_other             : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 2 1 1 1 ...
 $ loan_status               : Factor w/ 2 levels "C","D": 1 1 1 1 2 1 1 1 1 1 ...
```

Fuente: elaboración propia.

Como se mencionó, 14 variables tienen característica cualitativa (numérica) y 8 son cualitativas y binarizadas. En total tenemos 177.610 filas con 22 columnas.

Para el análisis de las variables cualitativas usaremos los diagramas de cajas, con el objetivo de identificar valores atípicos y comparar distribuciones de los datos. Para las variables cualitativas usaremos gráficos de barras (figura 7).

Figura 7: Estatus del crédito frente a monto del crédito

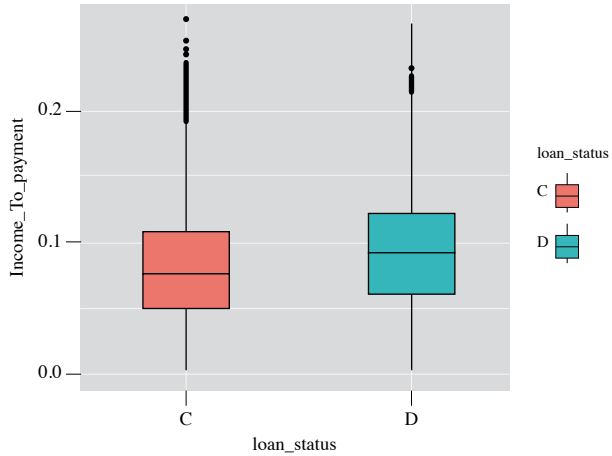


Monto del préstamo	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Cumple (C)	1.000	8.000	12.000	14.261	20.000	35.000
Incumple (D)	1.000	9.325	15.000	15.589	20.500	35.000

Fuente: elaboración propia.

La relación entre *loan status* y el monto del crédito nos permite observar que un mayor monto en el préstamo no es determinante para cumplir o incumplir la obligación. La mediana y el promedio del monto del crédito en los parámetros C y D son muy similares (figura 8).

Figura 8: Estatus del crédito frente a porcentaje del ingreso destinado al pago

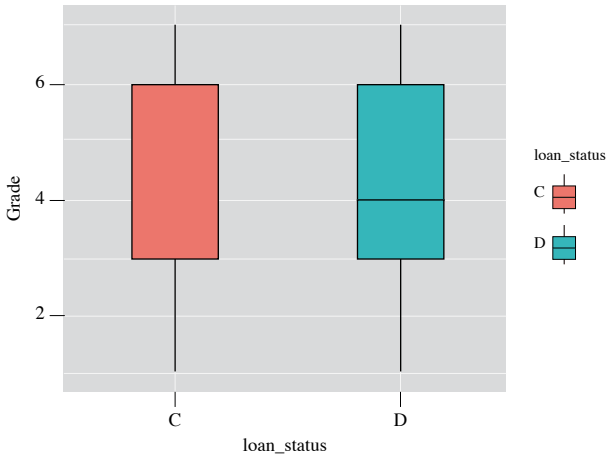


Porcentaje del ingreso destinado al pago	Min. (%)	1st Qu. (%)	Median (%)	Mean (%)	3rd Qu. (%)	Max. (%)
Cumple (C)	0,06	4,99	7,59	8,05	10,72	26,60
Incumple (D)	0,17	6,16	9,05	9,24	12,26	23,16

Fuente: elaboración propia.

En la figura 8 podemos observar que el porcentaje de ingreso destinado para pago tampoco determina de manera significativa el estado del crédito. Tanto la mediana como la media de los datos son similares. Ahora bien, se puede observar que quienes tienen un mayor porcentaje de ingreso destinado al pago son los créditos clasificados como "C" cumple, que se evidencian en la concentración de puntos del diagrama de cajas en su tercer quintil.

Figura 9: Estatus del crédito frente a calificación



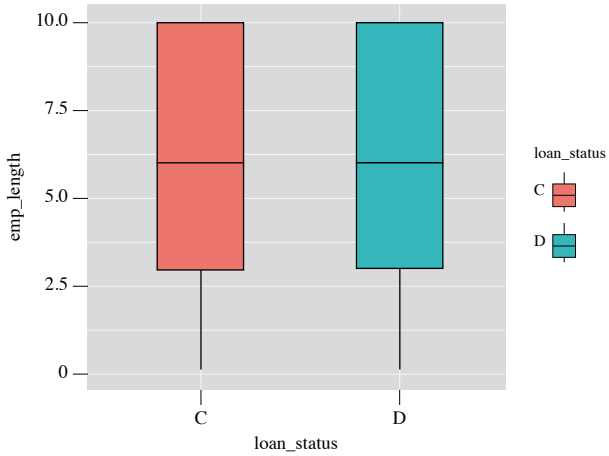
Calificación	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Cumple (C)	1	3	6	4.947	6	7
Incumple (D)	1	3	4	4.264	6	7

Fuente: elaboración propia.

La calificación de crédito que asigna LendingClub a los préstamos sí tiene una incidencia directa en el comportamiento de la variable dependiente (figura 9). Podemos observar que el grupo "C" cumple tiene como media una calificación de 6, dos grados mejor que la calificación del grupo "D" incumple, que solo llega a una calificación de 4. Al analizar el promedio, el mayor número de datos en la variable "C" cumple debido al desbalance disminuye un grado de calificación los datos de "C". Sin embargo, ratifica que los créditos con menor calidad crediticia tienen menor probabilidad de impago.

La antigüedad laboral tampoco es significativa en la condición de incumplimiento de los créditos. La media de los datos para ambas características es la misma y el promedio de las observaciones no tiene diferencia significativa. Tampoco se encuentran datos atípicos en esta relación (figura 10).

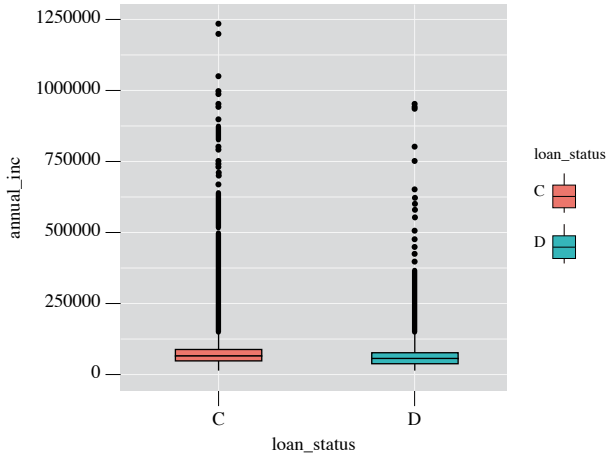
Figura 10: Estatus del crédito frente a antigüedad laboral



Antigüedad laboral	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Cumple (C)	0	3	6	6.092	10	10
Incumple (D)	0	3	6	6.097	10	10

Fuente: elaboración propia.

Figura 11: Estatus del crédito frente a ingreso anual

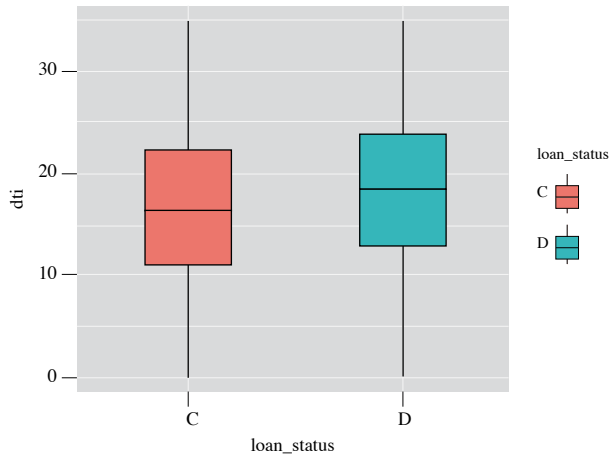


Ingreso anual	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Cumple (C)	5.000	46.000	65.000	74.131	90.000	1.233.000
Incumple (D)	7.600	43.000	59.500	67.171	80.000	950.000

Fuente: elaboración propia.

El ingreso anual es una variable con muchos datos atípicos que se concentran en los bigotes de la caja. Los créditos con la condición "C" tienden a tener mayores ingresos que aquellos clasificados como "D". Sin embargo, el diagrama de cajas y sus valores no nos muestra una alta influencia del ingreso anual frente a la condición de incumplimiento (figura 11).

Figura 12: Estatus del crédito frente a DTI



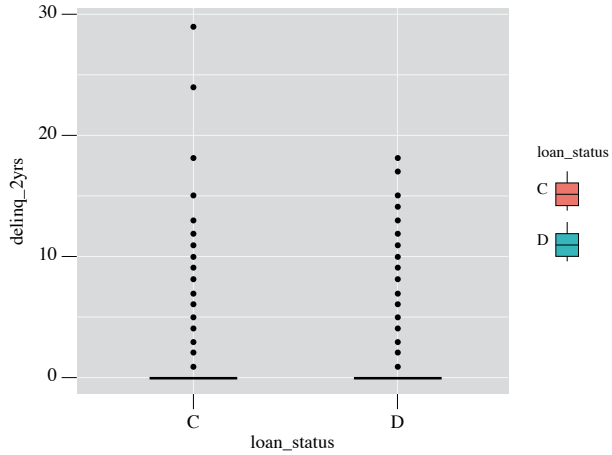
DTI	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Cumple (C)	0,00	11,06	16,41	16,77	22,21	34,99
Incumple (D)	0,00	12,82	18,42	18,41	23,95	34,99

Fuente: elaboración propia.

La relación calculada entre el pago mensual total de las deudas del prestatario en el total de sus obligaciones "DTI", nos muestra que una mayor relación de esta variable aumenta el incumplimiento del crédito. Tanto la mediana como la media son más altas en los créditos "D" frente a los créditos "C". Esta variable no presenta datos atípicos (figura 12).

Las moras mayores a 30 días nos muestran que la calidad de la cartera es buena ya que en promedio los prestatarios tienen una media y una mediana de 0 moras en los últimos dos años. Sin embargo, esta característica es debatible ya que el estado "C" y el estado "D" presentan datos atípicos en los bigotes de las cajas. Asimismo, al tener una base desbalanceada, esto puede incidir en mostrar este comportamiento positivo de los clientes que puede generar un dato positivo falso (figura 13).

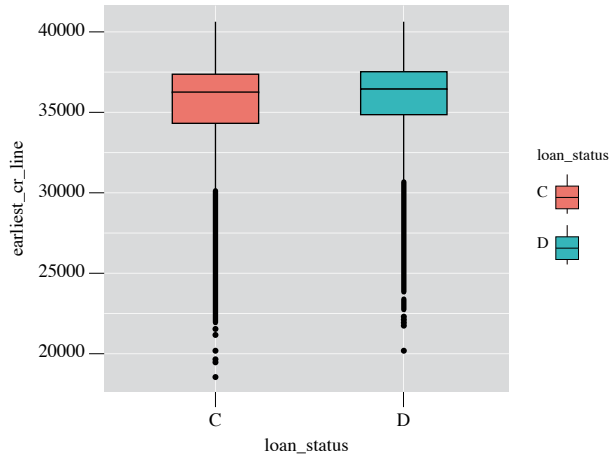
Figura 13: Estatus del crédito frente a mora mayor a los 30 días en los últimos dos años



Mora mayor a 30 días en dos años	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Cumple (C)	0,00	0,00	0,00	0,24	0,00	29,00
Incumple (D)	0,00	0,00	0,00	0,25	0,00	18,00

Fuente: elaboración propia.

Figura 14: Estatus del crédito frente a mes en que se solicitó

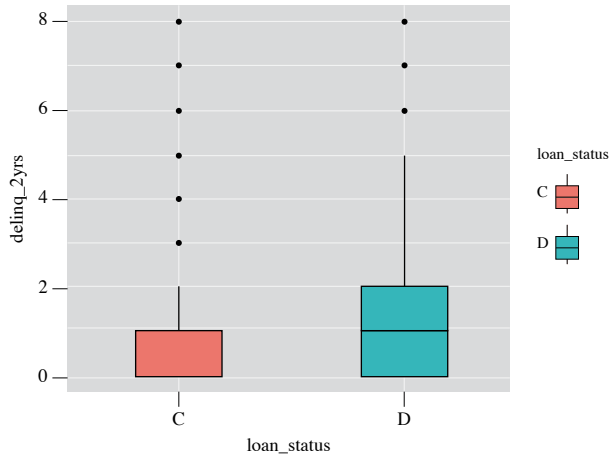


Mes en el que el cliente abrió el primer crédito en su vida	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Cumple (C)	nov-50	may-94	feb-99	oct-97	jun-02	nov-10
Incumple (D)	abr-55	abr-95	sep-99	jun-98	oct-02	oct-10

Fuente: elaboración propia.

El análisis de cajas nos muestra que, de la base de datos analizada, las personas pidieron su primer crédito hace más de 20 años, lo que nos enfrenta ante una población de prestatarios de unos 40 años (figura 14).

Figura 15: Estatus del crédito frente a antigüedad laboral



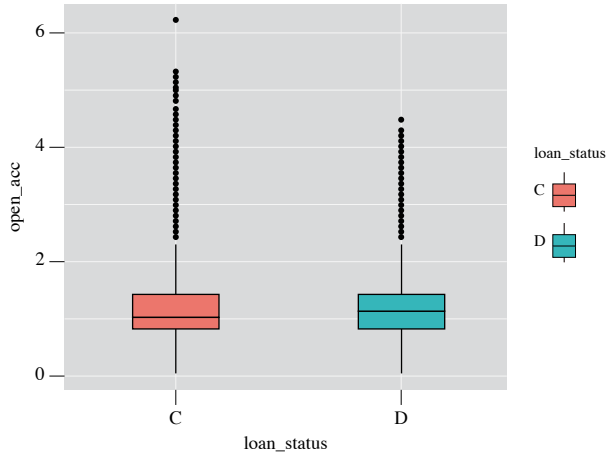
Número de solicitudes en los últimos 6 meses	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Cumple (C)	0,00	0,00	0,00	0,78	1,00	8,00
Incumple (D)	0,00	0,00	1,00	0,95	2,00	8,00

Fuente: elaboración propia.

En esta variable sí se puede observar que, a mayor número de solicitudes de crédito en el corto plazo, aumenta el número de incumplimiento del crédito. No obstante, ambas variables tienen datos atípicos de hasta 8 créditos solicitados que se observan en los bigotes de las cajas (figura 15).

En promedio, la población de la base de datos tiene alrededor de 40 años y ha pedido en su vida 11 créditos. Sin embargo, el diagrama de cajas no nos muestra un efecto significativo entre "C" y "D" con esta variable. Los bigotes de las cajas presentan una alta acumulación de datos atípicos en la parte superior, lo que refleja una alta dispersión de los datos (figura 16).

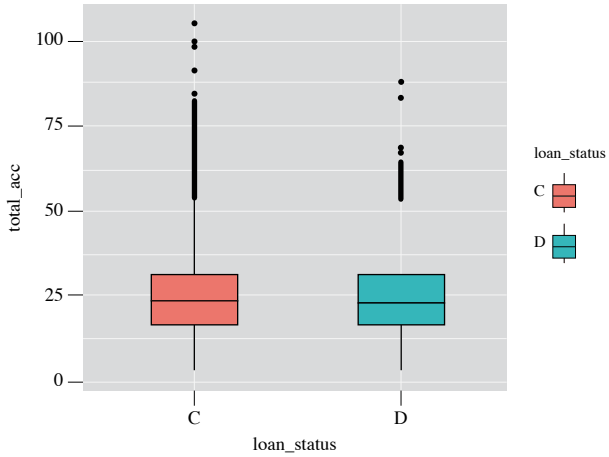
Figura 16: Estatus del crédito frente a total de créditos abiertos en la vida del prestatario



Total de créditos abiertos en la vida del prestatario	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Cumple (C)	0	8	10	11,01	14	62
Incumple (D)	0	8	11	11,26	14	45

Fuente: elaboración propia.

Figura 17: Estatus del crédito frente a líneas de crédito actualmente abiertas

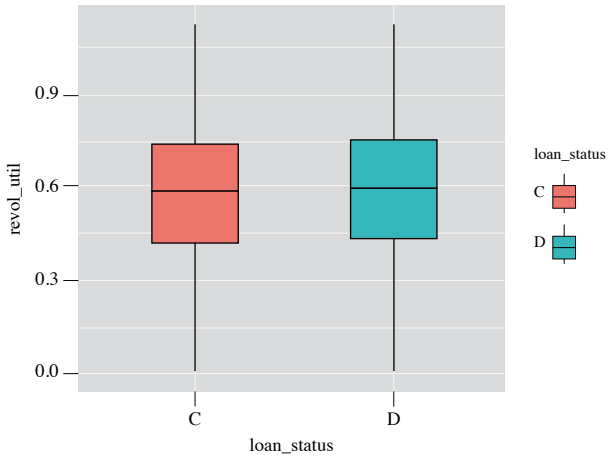


Líneas de crédito actualmente abiertas	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Cumple (C)	2	16	23	25	31	105
Incumple (D)	2	16	23	24,34	31	88

Fuente: elaboración propia.

En estos datos podemos observar que los créditos actuales superan a los pedidos de forma histórica. Este fenómeno puede explicarse por el horizonte de tiempo de los datos tomados. En ese periodo, en Estados Unidos las tasas de la Reserva Federal estuvieron entre 0,07 % hasta 0,17 %[6], lo que generó un alto acceso a crédito barato para las personas. Ahora bien, el aumento del número de créditos no incrementa la media o el promedio de las características "C" y "D" (figura 17).

Figura 18: Estatus del crédito frente a razón de crédito rotativo usado



Razón de crédito rotativo usado	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Cumple (C)	0,0000	0,4240	0,5860	0,5772	0,7420	11,270
Incumple (D)	0,0000	0,4360	0,5950	0,5849	0,7490	11,270

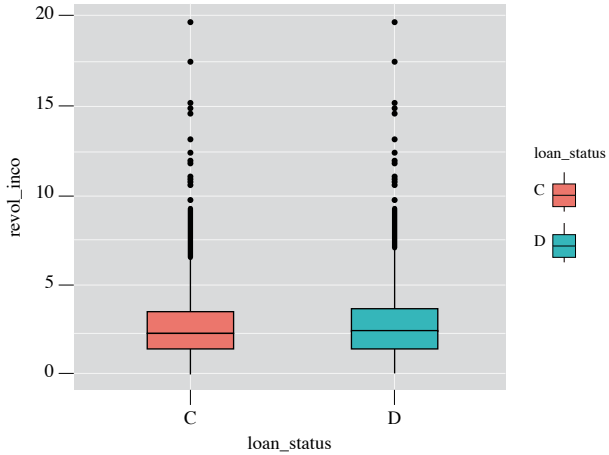
Fuente: elaboración propia.

En esta relación, la variable independiente tanto en los créditos calificados como "C" como los calificados como "D" se comportan de manera similar en su media y su promedio, por lo cual no es determinante la relación en el incumplimiento del préstamo (figura 18).

El gráfico de cajas nos muestra que el no uso de los créditos rotativos tampoco determina el incumplimiento de los préstamos. Sin embargo, esta variable sí tiene una alta acumulación de datos atípicos en los bigotes superiores de las cajas, lo que genera una alta dispersión de los datos de la variable independiente (figura 19).

6 Tasas históricas de Federal Fund Rate publicadas en: https://www.federalreserve.gov/

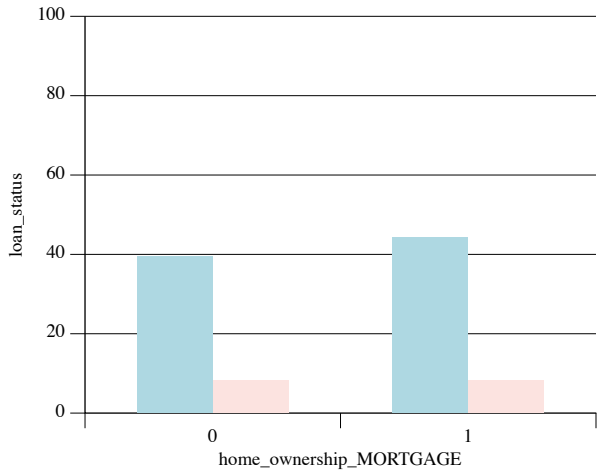
Figura 19: Estatus del crédito frente a razón de crédito rotativo no usado



Razón de crédito rotativo no usado	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Cumple (C)	0,000	1,351	2,256	2,618	3,482	19,751
Incumple (D)	0,000	1,454	2,394	2,774	3,713	19,751

Fuente: elaboración propia.

Figura 20: Estatus del crédito frente a prestatario con hipoteca sobre su vivienda

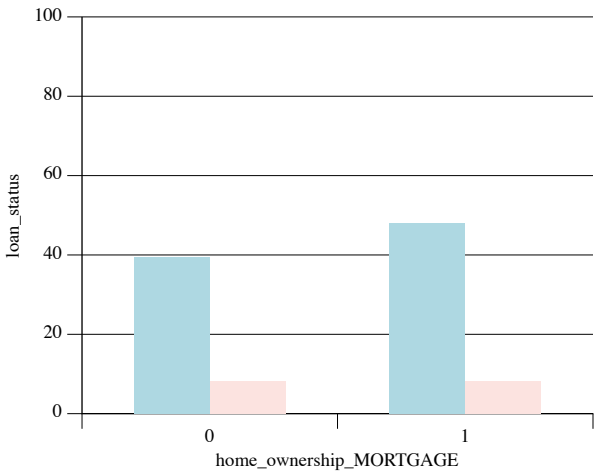


El prestatario posee vivienda con hipoteca	NO (%)	SÍ (%)
Cumple (C)	40,00	44,42
Incumple (D)	8,24	7,35

Fuente: elaboración propia.

Del 100 % de la población analizada, el 44 % posee vivienda con hipoteca; similar a la población que representa la mayoría de créditos con calificación "C". Asimismo, el análisis de datos continúa presentando el desbalance de la base hacia la clasificación "C", donde el 84,42 % de la población analizada tiene clasificación "C" (figura 20).

Figura 21: Estatus del crédito frente a vivienda propia



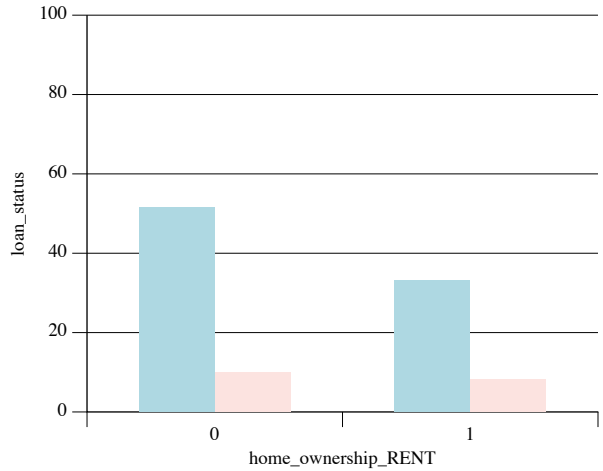
Vivienda propia	NO	SI
Cumple (C)	77,80%	6,61%
Incumple (D)	14,34%	1,25%

Fuente: elaboración propia.

Los datos muestran que la mayoría de los usuarios analizados no tienen vivienda propia. Sin embargo, el 77,80 % de la población cumple con el crédito (figura 21).

Los clientes con la variable "vivienda alquilada" continúan mostrando el buen comportamiento de sus patrones de pagos, al ser el mayor grupo frente a toda la base analizada (figura 22).

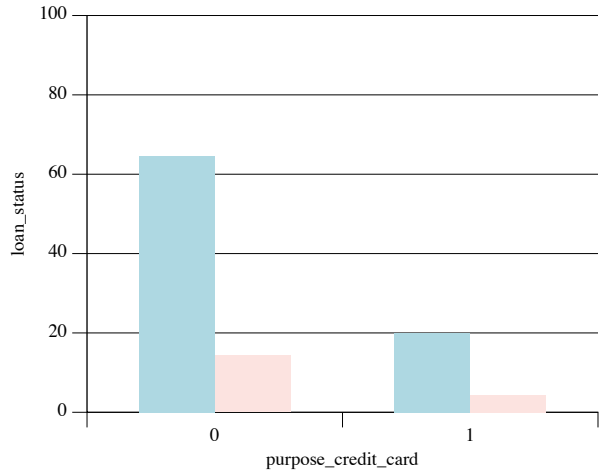
Figura 22: Estatus del crédito frente a vivienda en arriendo



Vivienda alquilada	NO (%)	SÍ (%)
Cumple (C)	51,03	33,39
Incumple (D)	8,60	6,99

Fuente: elaboración propia.

Figura 23: Estatus del crédito frente a propósito tarjeta de crédito

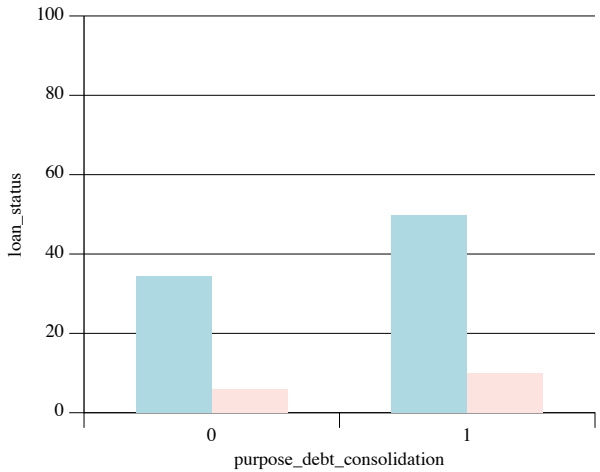


Propósito tarjeta de crédito	NO (%)	SÍ (%)
Cumple (C)	64,52	19,89
Incumple (D)	12,64	2,95

Fuente: elaboración propia.

Los datos continúan presentando el desbalance de la base de datos hacia la clasificación “C”. Una mayoría significativa de la base tiene propósitos clasificados como negativos por LendingClub sobre el uso de la tarjeta de crédito. Sin embargo, la población se clasifica como “C” (figura 23).

Figura 24: Estatus del crédito frente a propósito consolidación de la deuda



Propósito consolidación de la deuda	NO (%)	SÍ (%)
Cumple (C)	34,69	49,72
Incumple (D)	5,76	9,83

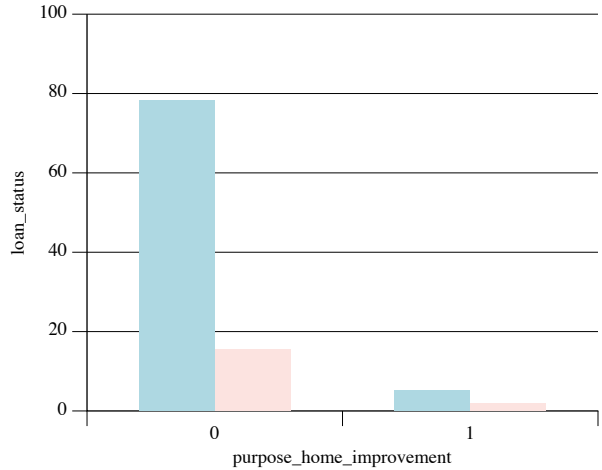
Fuente: elaboración propia.

En su mayoría, los propósitos de consolidación de la deuda son tomados como positivos por LendingClub. Asimismo, esta variable continua muestra el desbalance de la base de datos al observarse que el 49,72 % de los datos son clasificados como “C” (figura 24).

Contrario a los datos anteriores, la variable “propósito arreglos locativos” sí tiene una mayoría negativa asignada por LendingClub. A pesar de esto, el desbalance de los datos continúa mostrando que, en esta comparación, la mayoría de la población se clasifica con “C” (figura 25).

La variable Otro propósito tiene un comportamiento similar a arreglos locativos, donde a pesar de que la clasificación del propósito es negativa, la variable binaria “NO” posee la mayor población de clasificados como “C” del universo de datos (figura 26).

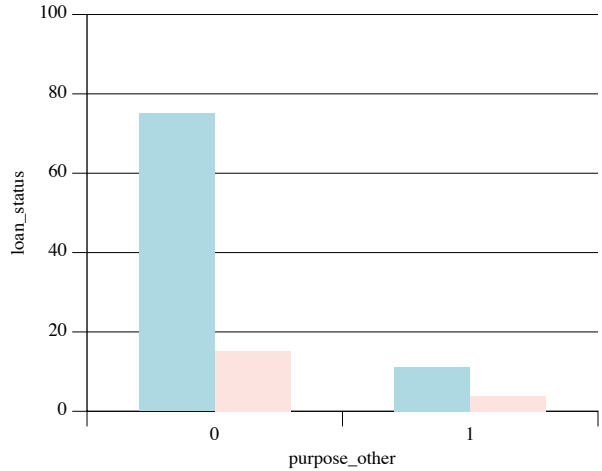
Figura 25: Estatus del crédito frente a propósito arreglos locativos



Propósito arreglos locativos	NO (%)	SÍ (%)
Cumple (C)	79,76	4,65
Incumple (D)	14,87	0,721

Fuente: elaboración propia.

Figura 26: Estatus del crédito frente a otro propósito



Otro propósito	NO (%)	SI (%)
Cumple (C)	74,262	10,149
Incumple (D)	13,494	2,094

Fuente: elaboración propia.

4. Experimento de SVM aplicado a la base de datos de LendingClub

El objetivo del experimento es clasificar de manera supervisada la variable *loan status* de la base de datos. Esto con el objetivo de evaluar si la SVM mejora la clasificación original dada por LendingClub.

En este experimento utilizamos programación en R para análisis de datos. En una primera etapa cargamos las bases de datos desbalanceadas (datos originales) y balanceadas (variable *loan status* clasificada como "C" y "D" en participaciones iguales).

Dada la cantidad de datos, su separación no puede ser planteada únicamente a través del hiperplano. Por esto utilizamos la expansión Kernel para transformar la base en espacio lineal. Utilizamos tres kernels: lineal, radial y polinómico.

Debido a la importancia de la transformación de los datos para el experimento, y su directa incidencia en la eficiencia de la predicción, se implementaron los tres tipos de Kernel descritos anteriormente y su calibración se hizo a través de búsqueda directa.

Debido a la cantidad de datos, en los tres tipos de kernels se generó una partición de datos experimental para mejorar los tiempos de respuesta del experimento. Tanto en el Kernel lineal como en el polinómico, la partición menor al 50 % de los datos no generaba resultados en la máquina. El único Kernel que nos permitió análisis en tiempos menores a 24 horas fue el radial, con una partición de datos al 50 %.

4.1. Experimento con Kernel lineal

Se procesaron los 177.610 datos con esta transformación. No obstante, al ser 22 variables diferentes, la transformación en un solo espacio lineal no fue posible en tiempos óptimos, incluso aumentando el parámetro C del modelo para mayor flexibilidad. Se decidió realizar una partición de los datos al 50 %. Luego de 24 horas de análisis en la máquina que procesó los datos, se decidió interrumpir el experimento y avanzar a un Kernel radial.

El experimento con Kernel radial se realizó tanto en la base desbalanceada como en la balanceada y sobre estos dos resultados se arrojan conclusiones.

4.2. Experimento con Kernel polinómico

Se procesaron los 177.610 datos con esta transformación. En principio, la clasificación de los datos no arrojaba conclusiones pasadas las 24 horas de análisis de la máquina utilizada. Se tomó la misma decisión que el Kernel lineal. Pasadas 24 horas de experimento con una partición de la base al 50 %, la misma no generaba conclusiones. Se decidió interrumpir el experimento y realizarlo con un Kernel radial.

4.3. Experimento con Kernel radial

Se realizó una partición de los datos al 50 %, así:

```
index <- sample(177610,177610*0,5, replace = TRUE)
datossample = DATOS[index,]
```

Esto arrojó una base de 88.805 datos con los cuales seguimos el experimento. Luego de esto generamos una partición de los datos: de experimentación y de prueba. Los datos de experimentación correspondían al 80 % de la base llamada *datossample* y el 20 % restante se dejó para realizar el testeo del modelo.

```
in.train= createDataPartition(as.factor(datossample$loan_status),
p=0,8, list=FALSE)
datos.train=datossample[in.train,]
datos.test=datossample[-in.train,]
```

Datos.train arrojó una nueva base de análisis de 71.044 datos, mientras que la base de test, llamada datos.test, arrojó 17.761 datos para la prueba del modelo.

Finalmente, el modelo SVM con Kernel radial fue programado con costo 1, gamma 1 como parámetros:

```
svm.fit=svm(loan_status ~.,data=datos.train,kernel='radial',
cost=1,gamma=1,probability=TRUE)
summary(svm.fit)
set.seed(1)
```

En la base desbalanceada el resultado fue el siguiente:

Parámetros:
SVM-Type: C-classification
SVM-Kernel: radial
cost: 1

Number of Support Vectors: 68400

(57267 11133)

Number of Classes: 2

Levels:
C D

El resultado de la clasificación de SVM con Kernel radial nos muestra que el 96 % de la base creada a partir de la partición son soportes del modelo o puntos sobre el hiperplano de separación. De este 96 %, el 83 % de los vectores soporte corresponde a la clasificación "C", lo que evidencia que el desbalance de la base de datos inclina la clasificación hacia este resultado.

En el caso de la base balanceada, el experimento arrojó exactamente los mismos resultados.

Para validar los resultados del SVM con Kernel radial se programó una validación cruzada que arrojó una matriz de confusión (tabla 4). La matriz muestra de forma sencilla los resultados de la clasificación supervisada. En este resultado, las columnas representan las predicciones de cada clase y las filas representan las instancias de la clase real.

Tabla 4: Matriz de confusión base desbalanceada

	Predicted	
Truth	C	D
C	14967	6
D	2633	150

La intersección (C,C) representa los verdaderos positivos, la intersección (C,D) los falsos positivos, la intersección (D,C) los falsos negativos y, finalmente, la intersección (D;D) los verdaderos negativos (tabla 4).

Tabla 5: Matriz de confusión base balanceada

	Predicted	
Truth	C	D
C	7194	0
D	6556	456

La intersección (C,C) representa los verdaderos positivos, la intersección (C,D) los falsos positivos, la intersección (D,C) los falsos negativos y, finalmente, la intersección (D;D) los verdaderos negativos (tabla 5).

Tanto en la base balanceada como en la base desbalanceada, la tasa de verdaderos positivos es superior a los falsos positivos; al revisar la curva ROC (Receiver Operating Characteristic), esta arroja un resultado por encima de la diagonal central, lo que nos indica un resultado positivo en el modelo entre verdaderos positivos y falsos positivos (figuras 27a y 27b).

Figura 27a: ROC base desbalanceada

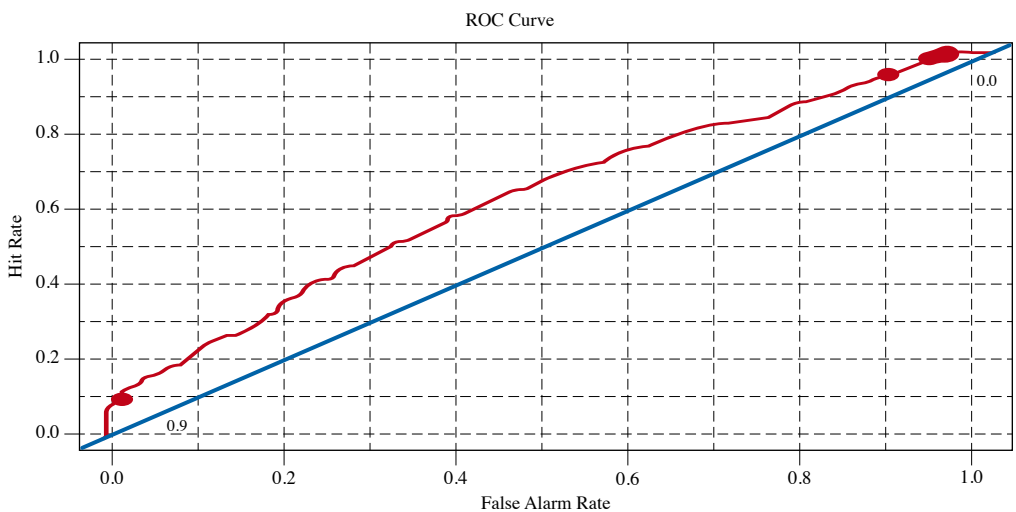
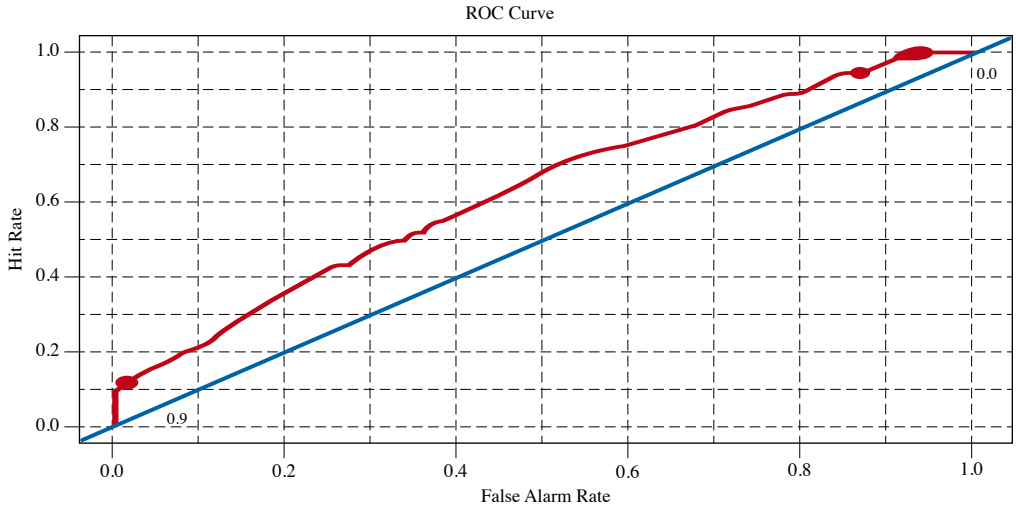


Figura 27b: ROC base balanceada



Conclusiones

La metodología SVM permitió analizar la clasificación de la variable *loan status* de la base de datos de LendingClub, esta dio como resultado que la mayoría de los datos tanto en la base balanceada como en la desbalanceada están bien clasificados, con una tasa de verdaderos positivos superior al 50 %.

Las curvas ROC nos ratificaron esta mejor clasificación al arrojar un resultado por encima de la diagonal central.

El error es menor en la base original, con solo 0,1486233, frente al error en la base balanceada que fue de 0,4614783.

Referencias

Bessis, J. (2015). *Risk Management in Banking* (e ed.). New York: Wiley.

Fernández-Sainz, A. (2010). ¿Bancos con problemas? Un sistema de alerta temprana para la prevención de crisis bancarias. *Cuadernos de Gestión*, *11*(2), 149-168.

Gareth, J. *et al*. (2017). *An Introduction to Statistical Learning with Applications in R*. Berlin: Springer.

Malekipirbazari, M. y Aksakalli, V. (2015). Risk assessment in social lending via random forests. *Expert Systems With Applications*, *42*(10), 4621-4631.

Moreno Gutiérrez, J. y Melo Velandia, L. (2011). Pronóstico de incumplimiento de pago mediante máquinas de vectores de soporte: una aproximación inicial a la gestión del riesgo de crédito. *Borradores de Economía*, *677*.

Vapnik, V. y Cortés, C. (1995). Support Vector Networks. *Machine Learning*, *20*, 273-297.

Xi, Y. *et al*. (2019). Risk control of online P2P lending in China based on health investment. *Ekoloji*, *107*, 2013-2022.

Yu, L. (2014). Credit risk evaluation with a least squares fuzzy support vector machines classifier. *Arcif*, 1-9.

Anexo 1
Resumen de las variables utilizadas en el experimento de la base de datos de LendingClub

	loan_amnt	term	Income_To_payment	grade	emp_length	annual_inc	dti
Min. :	1000	36	0,0005916	1	0	5000	0,00
1st Qu.:	8000	36	0,0514171	3	3	45000	11,31
Median :	12400	36	0,0780347	5	6	63000	16,73
Mean :	14468	41,76	0,0823144	4,84	6	73046	17,02
3rd Qu.:	20000	36	0,1098869	6	10	88000	22,51
Max. :	35000	60	0,266004	7	10	1233000	34,99

	delinq_2yrs	earliest_ cr_line	inq_last_ 6mths	open_acc	total_acc	revol_util	revol_inco
Min. :	0	18568	0	0	2	0	0,00
1st Qu.:	0	34516	0	8	16	0,426	1,37
Median :	0	36220	0	10	23	0,587	2,28
Mean :	0,2421	35763	0,8054	11,05	25	0,5784	2,64
3rd Qu.:	0	37438	1	14	31	0,744	3,52
Max. :	29	40483	8	62	105	1,127	19,75

	home_ ownership_ MORTGAGE	home_ ownership_ OWN	home_ ownership_ RENT	purpose_ credit_ card	purpose_ debt_ consolidation	purpose_ home_ improvement	purpose_ other
Min. :	0	0	0	0	0	0	0,00
1st Qu.:	0	0	0	0	0	0	0,00
Median :	1	0	0	0	1	0	0,00
Mean :	0,5177	0,07861	0,4037	0,2283	1	0,05372	0,12
3rd Qu.:	1	0	1	0	1	0	0,00
Max. :	1	1	1	1	1	1	1,00

	loan_status
C	149923
D	27687