Universidad de San Andrés

Escuela de Administración y Negocios

Magister en Finanzas

Desarrollo de *score* crediticio a través de redes neuronales y regresión logística

Autor: Fida, Alex Eduardo

DNI: 37.248.769

Director: Basaluzzo, Gabriel

Ciudad de Buenos Aires, 20 de Diciembre de 2023

Índice General

Resumen

En este trabajo se realiza una comparación de dos modelos para estimar la probabilidad de no pago de clientes que sacaron un préstamo en una institución financiera. En el primer modelo se utiliza una regresión logística y en el segundo una red neuronal. Se comprueba una mejor *performance* en la red neuronal logrando una mejor discriminación en ambas colas de la distribución del modelo. Para lograr este objetivo se construye una base de datos de originación de préstamos y se divide de forma aleatoria en una muestra de entrenamiento y otra de testeo. Este procedimiento se repite en varios pasos para así lograr obtener distribuciones de las distintas métricas de *performance* de los modelos y realizar el estudio de ellas.

1. Introducción

En las instituciones financieras es importante tener criterios confiables para determinar a quiénes deben otorgar el crédito y en qué medida hacerlo. Por lo dicho, es de vital importancia cuantificar el riesgo que se corre cuando se otorga un crédito y es de sumo interés para los bancos u otra entidad financiera disminuir lo más posible este riesgo cuando se adquieren nuevos clientes, de esta forma se reducirá la pérdida económica por consecuencia de una mala venta. Dentro de este contexto de incertidumbre, los bancos tienen un gran dilema, confiar en cada uno de sus clientes ya que dan dinero a través de préstamos y créditos, y en algunos casos sin solicitar garantía a cambio.

No solamente la incertidumbre ocurre al inicio de la relación, sino también en cada momento de la vida del crédito hasta que el cliente cubra toda su deuda. Por todo lo expuesto, es necesario contar con una segmentación de clientes para fidelizar aquellos que se supone que son buenos pagadores, la forma de lograrlo es ofrecerles un *cross sell* de productos, ampliar sus líneas de crédito, otorgarles una tasa preferencial, etc. Por ende, es fundamental que la política de riesgo de una entidad financiera esté dirigida a cuantificar e identificar a los clientes confiables.

En este trabajo se aplica una metodología de segmentación llamada *scoring*, que busca asignar una puntuación a los clientes que solicitan por primera vez un crédito en una entidad financiera (*Origination score),* con el fin de identificar su nivel de riesgo. Estos tipos de modelos buscan estimar una probabilidad de incumplimiento que logre discriminar entre clientes que son buenos pagadores y los que incumplen en sus pagos.

Los modelos de regresión o aquellos que utilizan alguna técnica de *machine learning*, para diferenciar a los buenos clientes del total de la población, consisten en una función que devuelve un valor que se toma como probabilidad, estas funciones y modelos se pueden estimar con los datos de la institución financiera, es decir, basándose en la experiencia propia.

El uso de modelos de aprendizaje automático se ha vuelto común en diversos campos, incluyendo la regresión logística y las redes neuronales. Sin embargo, elegir entre una y

otra puede ser difícil y depende de factores como la complejidad del problema, el tamaño del conjunto de datos y la necesidad de interpretación. En casos donde se requiere un modelo interpretable y se dispone de un conjunto de datos pequeño, la regresión logística puede ser la mejor opción, mientras que las redes neuronales pueden ser más apropiadas para problemas complejos con grandes conjuntos de datos. Aunque las redes neuronales pueden modelar relaciones complejas, esto no siempre mejora significativamente el rendimiento del modelo, especialmente con datos pequeños o relaciones lineales. En conclusión, la elección debe basarse en una evaluación cuidadosa de los factores relevantes para el problema, no en suposiciones sobre la superioridad de una técnica sobre la otra.

2. Revisión literaria y antecedentes

2.1 Basilea I

Desde los años setenta se comienza a ver cambios en el manejo de los bancos debido tanto a la volatilidad de variables exógenas y endógenas explicadas por la globalización de mercados financieros como a la innovación de productos financieros, entre otros, que obligan a que se introduzcan cambios en materia de regulación. Es así como el colapso en 1974 de Bankhaus Herstatt en Alemania y del Banco Nacional Franklin en los EE.UU, obliga a que en el Banco Internacional de Pagos (BIS), con sede en Suiza, se cree el Comité de Supervisión Bancaria de Basilea con los presidentes de los bancos centrales del grupo G102, con el objetivo de formular recomendaciones para la regulación de instituciones financieras y enfrentar de manera más eficaz las inestabilidades producidas por un mercado financiero mundial. Este organismo, consciente de que las instituciones se enfrentan al riesgo de crédito, hace público el Acuerdo de Capitales de Basilea en 1988, el cual es conocido como Basilea I, en dónde se hacen las recomendaciones necesarias dada la importancia de asegurar la estabilidad del sistema y mantener un capital mínimo con el que se cubran los capitales sujetos al riesgo de posibilidad de impago, denominado capital mínimo regulatorio. Conforme a Basilea I, el capital mínimo era una suma ponderada de activos distribuidos en cuatro categorías de riesgo.

2.2 Basilea II

En 1999 el Comité de Basilea se reúne nuevamente y se crea un nuevo acuerdo: Basilea II, se hace público en 2004 y una versión más completa en 2006, donde se amplía el tratamiento de los riesgos, teniendo en cuenta además del riesgo de crédito, los riesgos operacionales y los de mercado. El acuerdo Basilea II recomienda la gestión del sistema financiero a través de tres pilares:

- *Requisitos de capital mínimo*: cubrimiento de capital en riesgo.

- *Proceso de examen supervisor*: donde el ente supervisor cumple un papel primordial en la vigilancia y supervisión de la administración por parte de las entidades financieras.

- *La disciplina de mercado*: acceso y transparencia de la información suministrada por las entidades financieras. El tercer pilar busca la transparencia de la información. Al existir la posibilidad de crear metodologías de medición y gestión de riesgos para cada banco, este pilar toma mucha importancia. Para ello, intenta fomentar la disciplina de mercado mediante el desarrollo de una serie de requisitos de divulgación que permite a los agentes del mercado evaluar información esencial referida al ámbito de aplicación, el capital, las exposiciones al riesgo, los procesos de evaluación del riesgo y, con todo ello, a la suficiencia del capital de la institución

En Basilea II se emiten recomendaciones para gestionar tres tipos de riesgos:
- Riesgo de Crédito
- Riesgo de Mercado
- Riesgo Operacional

Basilea 1	Basilea 2
Estructura basada en un Pilar	Se establecen 3 Pilares
Medición del Riesgo Crediticio: aplicación de ponderaciones dadas por el regulado	Medición del riesgo crediticio: aplicación de ponderaciones externas (calificadoras) o por métodos internos
Cálculo del Riesgo Crediticio: por medio del enfoque estandarizado	Cálculo del Riesgo Crediticio: mediante 3 métodos: a. Estandarizado b. FIRB: IRB básico (*Internal Ratings Based*) c. AIRB: IRB avanzado
Incorpora la medición del Riesgo de Mercado desde 1996	Permanece igual
No incorpora la medición del Riesgo Operativo	Incorpora la medición del Riesgo Operativo
No incluye posibilidad de requerimiento adicional por otros riesgos	El Pilar 2 da la posibilidad al ente supervisor de requerir mayor capital por otros riesgos (ej. Concentración de mercado)

Tabla 1: principales diferencias entre Basilea I y Basilea II

a. *Método estandarizado*: establece ponderadores en función del riesgo de cada tipo de exposición y la clasificación externa de la contraparte.
b. *Método de modelos internos o IRB (Internal Ratings Based):* incorpora los modelos internos, antes sólo admitidos para riesgo de mercado (enmienda del acuerdo de capital de 1996). Tiene 3 parámetros claves: *PD, LGD y EAD*

2.3 Basilea III

En diciembre de 2010 se aprueba el acuerdo de Basilea III con un paquete de reformas a los estándares de capital y liquidez como respuesta a la crisis financiera internacional iniciada en Estados Unidos por los créditos hipotecarios. El objetivo principal de Basilea III es mejorar la capacidad del sistema bancario de absorber perturbaciones en situaciones de *stress* reduciendo el contagio al sector real y reforzar los requerimientos de capital por el riesgo de contraparte en operaciones con derivados. Además de gestionar políticas anti cíclicas para generar colchones de capital durante los buenos tiempos que permitan hacer frente al cambio del ciclo económico.

2.4 Previsión: antecedentes regulatorios en la Argentina

A. Circular de previsiones

El Banco Central establece un modelo de cálculo de previsiones cuyos objetivos son una correcta valoración de las pérdidas inherentes a las carteras de crédito, reforzar la estabilidad financiera y la competitividad del sector, a partir de previsiones genéricas y específicas.

B. NIC 39 (Norma Internacional de Contabilidad)

Es un modelo denominado de pérdidas incurridas. Las pérdidas esperadas como resultados de hechos futuros "no se reconocen, independientemente de lo alta que pueda ser la probabilidad de que puedan llegar a producirse", aunque también reconoce una pérdida "incurrida pero no reflejada".

C. *IFRS* 9 (*International Financial Reporting Standard*)

E capítulo 1 de la norma determina: "*El objetivo de esta Norma es establecer los principios para la información financiera sobre activos financieros y pasivos financieros,*

de forma que se presente información útil y relevante para los usuarios de los estados financieros para la evaluación de los importes, calendario e incertidumbre de los flujos de efectivo futuros de la entidad". Propone un modelo de pérdidas esperadas basado en valoración de descuento de flujos.

La definición de la norma, se ha desarrollado en 3 bloques

I. Clasificación y valoración:

Se clasifica a los instrumentos financieros según el modelo de negocio que tenga la entidad para el ingreso de flujos de efectivos.

II. Metodología del deterioro de valor:

Modifica de forma significativa el mecanismo de reconocimiento de provisiones ante potenciales pérdidas por incumplimiento. Introduce el concepto de pérdida "*lifetime*" que a diferencia de Basilea III la pérdida se basaba en estimaciones futuras a 12 meses.

Propone un modelo basado en pérdidas esperadas segmentando los instrumentos en 3 grupos:

a. Previsiones basadas en las pérdidas esperadas de los próximos doce meses: para aquellos instrumentos financieros que no se han deteriorado significativamente en calidad de crédito
b. Previsiones basadas en las pérdidas esperadas durante toda la vida del activo financiero: para los activos que se han deteriorado de manera significativa desde su reconocimiento inicial pero que no presenten evidencia objetiva de un evento de pérdida de crédito.
c. Previsiones sobre instrumentos ya en *default*, incorporando todas las pérdidas esperadas futuras: para los activos financieros que tengan evidencia objetiva de deterioro al final del periodo sobre el que se informa.

La determinación de la evidencia objetiva en general se apalanca en la de *default* BIS pudiendo ser ajustada: impago de más de 90 días (default objetivo)

III. Contabilidad de coberturas

Modificación de la contabilidad de micro coberturas con un doble objetivo: aproximar el tratamiento contable a la gestión de riesgos y simplificar el modo de contabilizar las coberturas

2.5 *Credit scoring*

Los modelos de *credit scoring* asignan al evaluado un puntaje o *score* (Bessis, J., 2010). A su vez, permiten relacionar ese valor con estimaciones más concretas del riesgo, en general se busca adquirir alguna estimación de la probabilidad de incumplimiento del deudor (PD, por probabilidad de *default*) asociada a su *score*.

La creación de los modelos matemáticos para la construcción de un *score* crediticio, es decir para estimar probabilidades de *default*, comienza en los años setenta pero se generaliza a partir de la década del noventa. Esto surge debido al desarrollo de mejores recursos estadísticos, computacionales y por la creciente necesidad por parte de las instituciones financieras, principalmente bancos, de hacer más eficaz y eficiente la originación de financiaciones, además de tener una mejor evaluación del riesgo de su portafolio. Estos modelos generalmente se asocian a lo que se ha dado en llamar *data mining* (minería de datos), que son todos aquellos procedimientos que permiten extraer información útil y encontrar patrones de comportamiento de los datos. A pesar de la generación de métodos automáticos para gestionar el riesgo y su uso en los negocios, el juicio humano continúa siendo utilizado en la originación de créditos. De hecho, en la práctica ambas metodologías muchas veces coexisten y se complementan, definiendo sistemas híbridos.

Los *credit scoring*, según Ciby Joseph (2013), son procedimientos estadísticos que se usan para clasificar a aquellos que solicitan crédito, inclusive a los que ya son clientes de la entidad crediticia, en clientes "buenos" y "malos". En su comienzo, el análisis discriminante era la técnica más utilizada, luego fue evolucionando para utilizar herramientas matemáticas y de inteligencia artificial.

Tipos de Score

Dependiendo en que parte del ciclo se esté trabajando se calcula uno de los siguientes *scores*:

- *Score de originación (Aquisition Score)*: en las áreas de políticas de crédito se utiliza este *score* para la aceptación o rechazo de las solicitudes de crédito, en otras palabras, en el comienzo de la relación con un cliente. Los tipos de variables utilizadas son demográficas y de *bureau* de créditos. Este tipo *score* estima la probabilidad de incumplimiento de pago de un posible cliente y de esta manera se decide si se acepta o rechaza como posible consumidor de crédito.

- *Score de comportamiento (Behavior score):* es utilizado en la etapa de administración del ciclo de riesgo. Predice la probabilidad de incumplimiento de los clientes que ya pertenecen a la institución y tienen un comportamiento financiero en ella. Por lo dicho, se utilizan las variables de comportamiento de los productos crediticios con la propia institución.

- *Score de cobranza (Collection score)*: este tipo de *score* es utilizado por las áreas de cobranzas, se calcula en la parte de recuperación de cuentas para estimar la probabilidad de recuperar a un cliente y hacer más eficientes la gestión de cobranza. Las variables utilizadas resultan de la combinación de variables de comportamiento y *bureau* de crédito.

Modelos de Score

Modelo externo: el modelo se construye con información de otras instituciones, esto significa que está listo para usar. Se trata de modelos de crédito genéricos que se compran a consultores externos.

Modelos estadísticos: modelos que se construyen con información propia, son conocidos como modelos *in-house.* Tienen como beneficio que se pueden construir modelos específicos para distintos segmentos de la población.

3. Metodología

En esta sección se detallan los teoremas o herramientas utilizadas en este estudio que hace referencia a la teoría de la probabilidad, estadística y álgebra matricial. (Canavos, G. C., 1984 y Ross, S., 2007)

3.1 Estadística y teoría de la probabilidad

Definición clásica de probabilidad

Si un experimento que está sujeto al azar, resulta de n formas igualmente probables y mutuamente excluyentes, y si n_A de estos resultados tienen un atributo A, la probabilidad de A es la proporción de n_A con respecto a n.

$$P(A) = \frac{n_A}{n}.$$

En muchas situaciones prácticas, los posibles resultados de un experimento no son igualmente probables. En estos casos, no es correcto estimar la probabilidad mediante la definición clásica. En lugar de esta, en muchas ocasiones, se emplea la interpretación de probabilidad como una frecuencia relativa. La interpretación de una frecuencia relativa como una medida de probabilidad se debe a que un experimento se efectúa y se repite muchas veces. A medida que se aumenten las repeticiones, la frecuencia relativa se aproxima al verdadero valor de la probabilidad.

Probabilidad marginal y conjunta

Si lo que se desea es determinar la probabilidad de que ocurra un evento A, dado que ya sucedió el evento B, se habla de una probabilidad condicionada, de que ocurra A dado B. Entonces, sean A y B cualquiera dos eventos en un espacio muestral S de manera tal que $P(B) > 0$. La probabilidad condicional de A al ocurrir el evento B, es el cociente de la probabilidad conjunta de A y B con respecto a la probabilidad marginal de B:

$$P(A/B) = \frac{P(A \cap B)}{P(B)}.$$

Esta relación puede escribirse como un producto, lo que da como resultado la regla de multiplicación de probabilidades dada por:

$$P(A \cap B) = P(B) * P(A/B),$$

donde $P(A \cap B)$ es la probabilidad conjunta de los eventos A y B.

Probabilidad marginal

Si lo que se desea es la probabilidad de cada evento por separado, se deben ignorar una o más características del espacio muestral. Si se quiere obtener la probabilidad marginal del evento A, se debe efectuar la suma de las probabilidades conjuntas de los eventos A y B, sobre todos los eventos de B. Se puede expresar como:

$$P(A) = \sum_1^n P(A/B_n),$$

donde

B_n: son todas las posibles realizaciones del evento B.

En otras palabras, la probabilidad marginal de un evento A, es igual a la suma de las probabilidades conjuntas de A y B, donde la suma se efectua sobre todos los eventos B_j.

Coeficiente de correlación de Pearson

El coeficiente de correlación de Pearson, pensado para variables cuantitativas (escala mínima de intervalo), es un índice que mide el grado de covariancia entre distintas variables relacionadas linealmente. Esto significa que puede haber variables fuertemente relacionadas, pero no de forma lineal. Este coeficiente, cuyo valor no depende de las unidades de medida de las variables porque es adimensional, está acotado entre -1 y +1; su signo indica la dirección, positiva o negativa, de la asociación lineal y su valor absoluto la intensidad de la misma.

El coeficiente de Pearson se calcula de la siguiente manera:

$$r_{xy} = \frac{\sigma_{xy}}{\sigma_x * \sigma_y} = \frac{E[(X - \mu_x) * (Y - \mu_Y)]}{\sigma_x * \sigma_y},$$

Fórmula 1

donde

σ_{xy}: $es\ la\ covarianza\ de\ (X, Y)$.

σ_x: $es\ el\ desvío\ estandar\ de\ X$.

σ_Y: $es\ el\ desvío\ estandar\ de\ Y$.

Interpretación del coeficiente de Pearson

- Si r = 1, existe una correlación positiva perfecta. El índice indica una dependencia total entre las dos variables denominada relación directa: cuando una de ellas aumenta, la otra también lo hace en proporción constante.

- Si 0 < r < 1, existe una correlación positiva.

- Si r = 0, no existe relación lineal. Pero esto no necesariamente implica que las variables son independientes: pueden existir todavía relaciones no lineales entre las dos variables.

- Si -1 < r < 0, existe una correlación negativa.

- Si r = -1, existe una correlación negativa perfecta. El índice indica una dependencia total entre las dos variables llamada relación inversa: cuando una de ellas aumenta, la otra disminuye en proporción constante.

Como se ha indicado el coeficiente de correlación de Pearson es un índice cuyos valores absolutos oscilan entre 0 y 1. Cuanto más cerca de 1 mayor es la correlación, y menor cuanto más cerca de cero. La interpretación de un coeficiente de Pearson depende en gran parte del estudio.

Para nuestro caso, se clasifica al coeficiente de la siguiente manera:

|r|> 0,7: correlación fuerte;

0,3 <|r|<0,7: correlación media;

0 <= |r| <= 0,3: correlación débil,

3.2 Inferencia estadística: prueba de *performance* de los modelos

En este apartado se muestran las técnicas que se utilizan para medir la *performance* de los modelos. Para la medición se hace uso de inferencia estadística como se detalla a continuación.

Test Kolmogorov – Smirnov

El test Kolmogorov-Smirnov (KS) es un test de hipótesis que se utiliza para determinar si hay divergencia entre las distribuciones de probabilidad de dos muestras independientes y para contrastar la distribución empírica de una muestra contra una distribución teórica. El estadístico de prueba se calcula como la máxima diferencia absoluta entre sus distribuciones empíricas (Figura 1), entonces se busca detectar las discrepancias existentes entre las frecuencias relativas acumuladas de las dos muestras de estudio. Estas diferencias están determinadas no solo por las medias sino también por la dispersión. La prueba se construye sobre las hipótesis nula y alternativa como sigue:

$$H_0: Las\ distribuciones\ poblaciones\ son\ iguales$$

Vs

$$H_1: Las\ distribuciones\ poblaciones\ no\ son\ iguales$$

Para esta prueba se requiere tener dos muestras de una variable aleatoria continua, o al menos de escala ordinal. Con los datos agrupados en *k* categorías o intervalos se calculan las frecuencias relativas acumuladas $\hat{F}_i$ y $\hat{G}_i\ con\ i = 1, 2, 3 \dots, k$ que corresponden a las dos muestras de tamaño n_1 y n_2 respectivamente. Se calculan entonces las diferencias de las frecuencias relativas acumuladas. El estadístico está dado como la máxima diferencia de las distribuciones de frecuencias relativas acumuladas.

.

$$KS = \max_{i \le i \le k} |\hat{F}_i - \hat{G}_i|$$

Fórmula 2

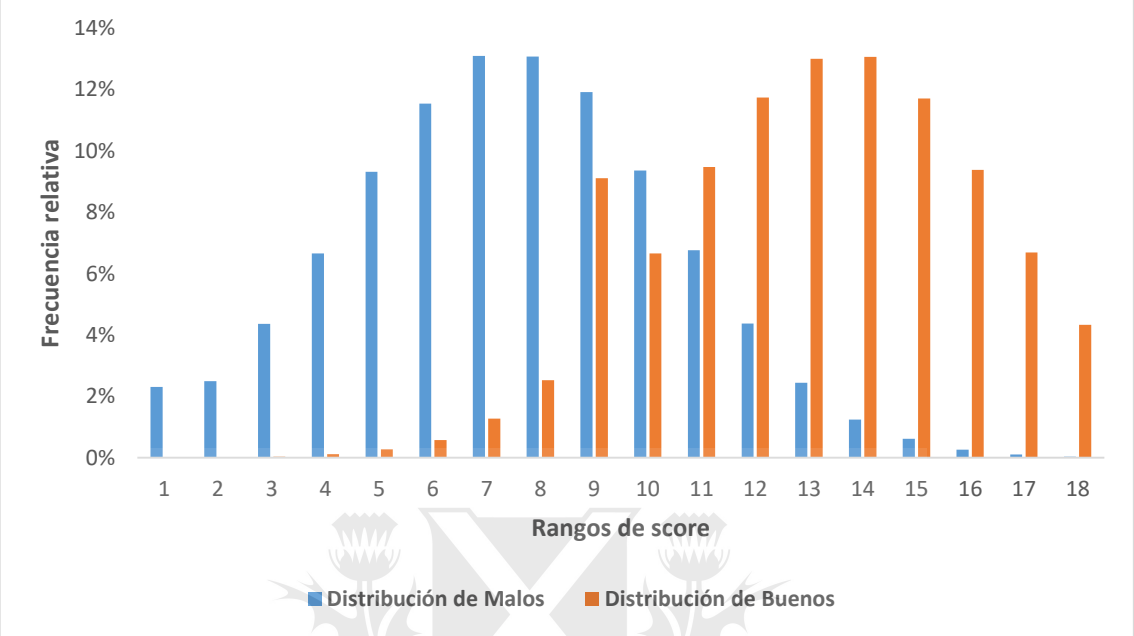


Figura 1: distribución de Buenos y Malos

Curva ROC (*Receiver operating characteristic*)

Se construye en base a la distribución de errores de clasificación de un modelo de caracterización de comportamiento. Es una técnica gráfica para visualizar, organizar y seleccionar clasificadores basados en su capacidad predictiva.

Se define una variable binaria *C* para la cual se consideran como buenos a los clientes con un cierto valor para esta variable, y malos a los clientes con el segundo valor definido. Los errores de clasificación cometidos reciben nombres en estadística, dado que corresponden a distintas variaciones de la hipótesis realizada:

- Error tipo I: clientes buenos son clasificados como posibles malos.
- Error tipo II: clientes malos son clasificados como posibles buenos.

Por lo tanto, se puede generar una matriz de confusión entre la clasificación otorgada por el modelo vs el resultado verdadero.

	Clientes verdaderamente buenos	cliente verdaderamente malos
Clientes clasificados buenos	Verdadero positivo (VP)	Falso positivo (FP)
Clientes clasificados malos	Falso negativo (FN)	Verdadero negativo (VN)
TOTAL	Total buenos (Tb)	Total Malos (Tm)

Tabla 2: matriz de confusión

Siendo:

- Verdadero positivo (VP): la instancia es positiva y es clasificada como positiva.
- Verdadero negativo (VN): la instancia es negativa y es clasificada como negativa.
- Falso positivo (FP): la instancia es negativa y es clasificada como positiva.
- Falso negativo (FN): la instancia es positiva y es clasificada como negativa.

Evaluación de la matriz de confusión

Exactitud: elementos de la diagonal principal de la matriz de confusión; tasa de acierto global del clasificador.

$$Exactitud = \frac{\mathrm{VP} + \mathrm{VN}}{Tb + Tm}$$

Sensibilidad: instancias positivas correctamente clasificadas como positivas.

$$Sensibilidad = \frac{\mathrm{VP}}{Tb}$$

Especificidad: instancias negativas correctamente clasificadas como negativas.

$$Especificidad = \frac{\mathrm{Vn}}{Tn}$$

El gráfico de una curva ROC (Figura 2) consigna en el eje Y la medida de sensibilidad (tasa de verdaderos positivos) y en el eje X el complemento de la especificidad (tasa de falsos positivos). De esta manera, la curva intenta representar el *trade-off* existente entre beneficios y costos.

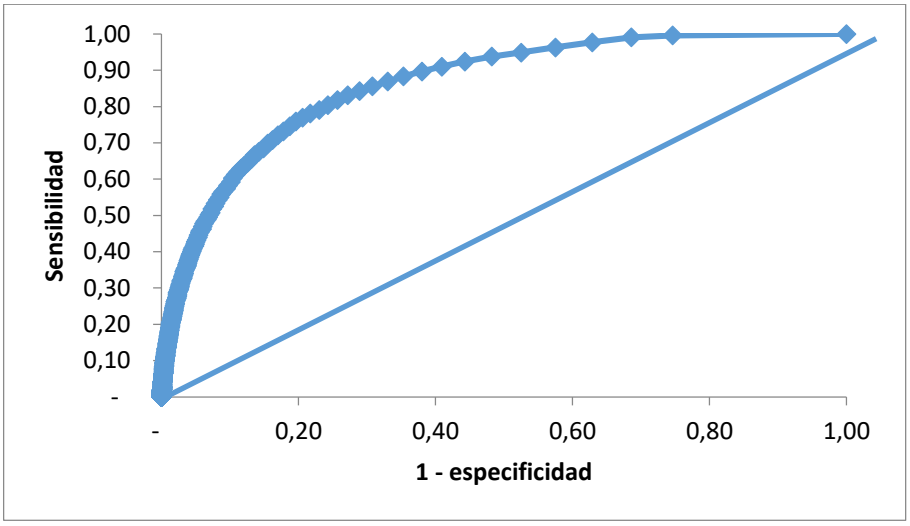


Figura 2: curva Roc

Algunos puntos importantes en el espacio de representación de la curva ROC son:

- Coordenadas (0,0): escenario de clasificación nula.
- Coordenadas (1,1): escenario de clasificación perfecta.
- Diagonal principal: escenario de clasificación aleatoria.

Donde el área bajo esta curva corresponde al resultado del test, este valor se mueve entre 0,5 y 1. Un valor cercano al 0,5 significa que el clasificador no es mejor que el azar; valores superiores a 0,5 y lo más cercanos a 1 son deseables.

3.3 Autovalores y autovectores

Los autovalores y autovectores corresponden a números y vectores asociados a matrices cuadradas. Dada una matriz A de $n * n$, su autovector $\vec{v}$ es una matriz $n * 1$ tal que

$$A * \vec{v} = \lambda * \vec{v}$$

donde el número λ es el autovalor, un valor escalar real asociado con el autovector. Se resta en ambos miembros $\lambda * \vec{v}$

$$A * \vec{v} - \lambda * \vec{v} = \vec{0},$$

luego se multiplica a $\vec{v}$ por I (matriz identidad)

$$A * \vec{v} - \lambda * I * \vec{v} = \vec{0},$$

agrupando:

$$(A - \lambda * I) * \vec{v} = \vec{0}.$$

Resulta un sistema homogéneo siendo $(A - \lambda * I)$ la matriz de coeficientes. Al ser un sistema homogéneo, siempre tiene la solución trivial y por lo tanto siempre es un sistema compatible. Se necesitan vectores que cambien su dirección, por lo tanto el objetivo es que el determinante de la matriz de coeficiente sea igual a 0.

$$det(A - \lambda * I) = 0$$

La ecuación escrita recibe el nombre de ecuación característica de la matriz A. Entonces,

$$P(\lambda) = det(A - \lambda * I)$$

Fórmula 3

es el polinomio característico. Para hallar los autovalores es necesario encontrar las raíces del polinomio característico. Luego para cada autovalor se determina todas las soluciones no triviales para el siguiente sistema

$$(A - \lambda * I) * \vec{v} = \vec{0}.$$

Fórmula 4

Una de las propiedades interesantes de esta herramienta, y por eso utilizada en este estudio, es que todos los autovectores de una matriz son perpendiculares. Esto significa que se pueden expresar los datos respecto a estos autovectores.

4. Crédito y riesgo de crédito

4.1 Crédito

Se entiende como crédito a la obligación pactada entre una persona o institución que otorga capacidad de compra por adelantado al deudor. El crédito permite satisfacer a la necesidad de compra de los consumidores, en función de su capacidad de pago. Los posibles clientes son aquellos que reúnen requisitos para que se les otorgue un crédito. Las instituciones que se dedican a otorgar créditos son los bancos y las instituciones financieras, así como también empresas comerciales, industriales, de servicios, etc. No importa la empresa o la institución que otorga el crédito, en todos los casos se cuantifica, gestiona y mitiga el riesgo de crédito.

Tipos de créditos

Créditos de Consumo o Créditos comerciales: son los otorgados por empresas para la adquisición de bienes o servicios de uso personal en plazos determinados.

Créditos Empresariales: cuando una empresa requiere materia prima, insumos, servicios, etc. solicita el bien o servicio a otras empresas a crédito para continuar su actividad empresarial, realizando convenios para cubrir su deuda en un futuro.

Crédito hipotecario: es el dinero que entrega el banco para adquirir una propiedad ya construida, un terreno, la construcción de viviendas, oficinas y otros bienes raíces, con la garantía de la hipoteca sobre el bien adquirido o construido; normalmente es pactado para ser pagado en el mediano o largo plazo.

Crédito prendario: es el dinero que le entrega el banco a una persona física para efectuar la compra de un bien mueble, generalmente, debe de ser aprobado por el banco o entidad financiera, y puesto que este bien mueble a comprar quedará con una prenda, hasta una vez saldada la deuda con la entidad financiera o Bancaria.

4.2 Definición de *default*

Por definición, el *default* ocurre cuando un prestatario no puede cumplir con sus obligaciones de pago de una deuda en el plazo y forma acordados. Podemos decir, existe un riesgo de *default* cuando existe la posibilidad de que una entidad o persona que haya obtenido un préstamo no pueda pagar el monto adeudado.

Las áreas de riesgos, en las instituciones financieras, tienen una tarea muy importante, estimar la ocurrencia del evento previo a que suceda. Para ello, existen distintos indicadores y métricas de morosidad que se calculan, que son utilizados para poder estimar la probabilidad de incumplimiento. Por lo tanto, es necesario definir dónde se genera tal incumplimiento. Es importante conocer cuándo se realiza ese evento, ya que es nuestra variable objetivo a predecir en nuestros modelos, es decir, estimar la probabilidad de que ocurra o no el *default*.

Métodos para seleccionar la definición de *default*:

- Normativo: Basilea recomienda como *default* clientes que tienen 90 días de atraso.
- Sistemas de información gerencial: a través de la información interna generada por la entidad, utilizando los *roll rates*. Dicha métrica muestra el movimiento neto de las cuentas de un tramo de mora (ej. 60-89 dpd) en un mes determinado, a un tramo de mora más severo (ej. 90-119 dpd) en el siguiente mes.
- Distribución interna de días de atraso: calcular la distribución acumulada de máximos días de atraso en una ventana de observación para aquellos clientes que en el primer mes de observación tengan 0 días de atraso.
- El banco considera probable que el deudor no abone la totalidad de sus obligaciones crediticias frente al grupo bancario, sin recurso por parte del banco a acciones tales como la realización de protecciones (si existieran).

No necesariamente el evento de *default* óptimo para la entidad coincide con la definición Basilea.

Este estudio se basa en el análisis que proveen los sistemas de información gerencial. Por lo tanto, se define como evento de *default* a los clientes con 60 días de atraso en una ventana temporal de 6 meses a partir de la fecha de origen del crédito.

4.3 Definición de riesgo de crédito

El riesgo de crédito es el riesgo de una pérdida económica como consecuencia de la falta de cumplimiento de las obligaciones contractuales por una de las partes. Este efecto es medido como el costo de restituir los fondos si la contraparte incumple sus obligaciones. El motivo para crear modelos de riesgo de crédito radica en la necesidad de calcular cuánto capital económico es necesario para sustentar las actividades de riesgo de un banco o una institución que presta dinero.

4.4 Medición del riesgo de crédito

Para cuantificar el riesgo de crédito de manera precisa, es necesario calcular la distribución de pérdida correspondiente. Esta distribución (ver Figura 3) suele no estar acotada a la derecha, lo que implica una dispersión denominada *Unexpected Loss* (*UL*). Para hacer frente a posibles pérdidas imprevistas que pudieran afectar el capital de la Institución, se requiere una cantidad adicional de capital destinada a absorber dichos riesgos.

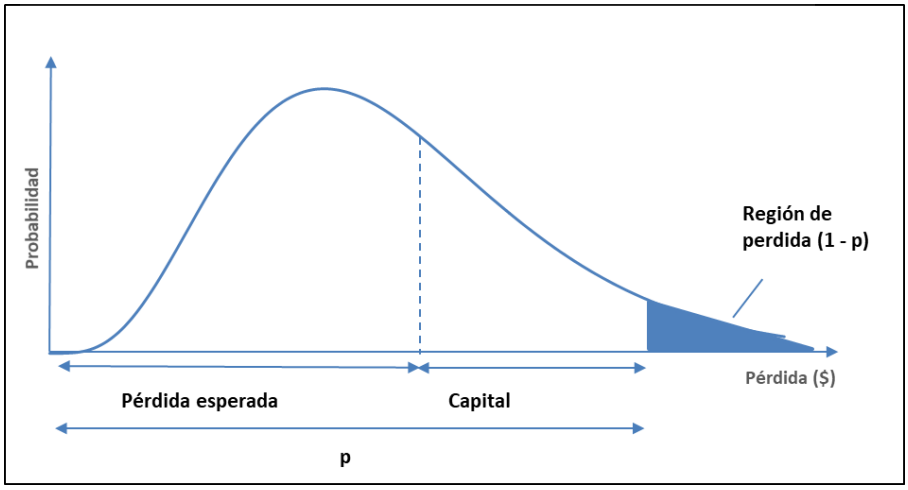


Figura 3: distribución de pérdida

donde:

- p: valor de pérdida correspondiente al percentil p%

- Pérdida esperada: es el saldo de pérdida esperado. Se cubre con provisiones
- Capital: es el saldo que cubre las perdidas inesperadas. Se calcula como var(p) − pérdida esperada.

De la distribución de pérdida es posible calcular la pérdida esperada. Los factores que se usan para el cálculo son:

- *Probabilidad de Default (PD, Probability of Default):* probabilidad de que una contraparte pueda no hacer frente a sus obligaciones en un determinado plazo temporal.

- *Exposición al Riesgo de Crédito (EAD, Exposure at Default):* volumen de riesgo expuesto en el momento de incumplimiento.

- *Pérdida dado el Incumplimiento (LGD, Loss given default)*: porcentaje final que se pierde en caso de incumplimiento, es decir, el porcentaje no recuperado.

Una vez obtenido los parámetros de riesgo, la pérdida esperada se calcula de la siguiente forma:

$$Perdida\ esperada\ (EL,\ expected\ loss\ ) = PD * LGD * EAD.$$

Fórmula 5

Las entidades no conocen "a priori" el número exacto de contrapartes que no pueden cumplir con sus obligaciones, ni su exposición ni su severidad. Por lo tanto, las tres variables son estocásticas.

4.5 Ciclo de riesgo

Se puede describir desde el punto de vista de la administración de riesgo de la siguiente forma:

- *Originación*: la intención en este punto es otorgar crédito a un cliente por primera vez en la institución.
- *Administración*: la intención en esta parte del ciclo es premiar a los clientes que se están "portando bien" (incrementos de límite de crédito) y castigar a los que se están "portando mal" (decrementos de límite de crédito). Aquí se busca la detección temprana de cuentas de alto riesgo y poder realizar acciones tempranas de corrección.
- *Recuperación*: en esta parte del ciclo de riesgo se pretende recuperar a todos aquellos clientes que dejaron de pagar.

4.6. Aplicaciones de los parámetros de riesgo de créditos en la gestión y negocios de una entidad crediticia

En este apartado se dan a conocer algunas utilidades y mejoras en lo que se refiere a la gestión de riesgo de crédito y el uso de los parámetros de dicho riesgo.

Riesgo de crédito – proceso

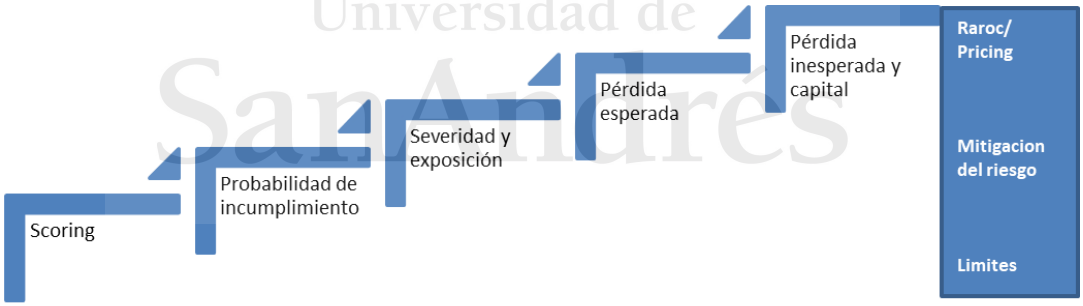


Figura 4: proceso de Riesgo de Crédito

El riesgo de crédito se mide por la pérdida esperada, dicho valor se conforma por 3 parámetros, entre ellos la probabilidad de *default*. A continuación, se nombran algunas aplicaciones en la gestión del riesgo:

- Posibilidad de segmentar la cartera entre distintos niveles de riesgo. Lograr fidelizar a los mejores clientes a través de *cross sell* con otros productos de la entidad.
- Administración de límites: *up grade* de ofertas a clientes con mejor *score* y decremento de límite para clientes malos. Es importante para la entidad la administración de los límites ya que controla la exposición al momento de calcular la pérdida.
- Mecanismo de precio *(Pricing)*: construcción de la rentabilidad ajustada al riesgo (*RAROC*), para la cual se utiliza para la renovación con tasa diferencial, o mismo en el alta de producto.

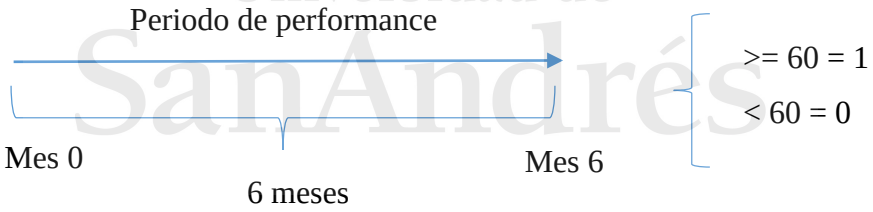
5. Construcción de la base de datos

Los datos utilizados en este estudio corresponden a un Fintech con operatoria en Europa, más específicamente de créditos originados en España.

5.1 Construcción de la variable dependiente

La variable que se predice en este estudio es la probabilidad de que los clientes superen los 60 días atraso en alguno de los siguientes 6 meses desde la fecha de originación del préstamo. Para la construcción de la variable dependiente es necesario entender que existen dos estados cuyo punto crítico se genera a los 60 días de morosidad. Los clientes que se clasifican como buenos son los que no superen este umbral en ningún mes posterior; los clientes clasificados como malos son los que tengan morosidad mayor o igual a 60 días en cualquiera de los siguientes meses durante los próximos 6 meses.

Una vez generada la marca de *default*, esta no puede ser removida, independiente que se logre "curar", la marca es para siempre. Lo anterior se representa en el siguiente diagrama



	Frecuencia	Porcentaje
Malo	10.125	64,8%
Bueno	5.500	35,2%
Total	15.625	100,00%

Tabla 3: distribución de clientes buenos y malos de la muestra total

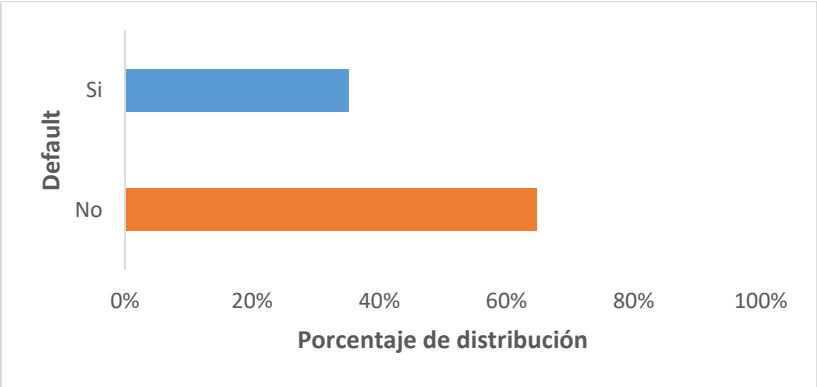


Figura 5: distribución de clientes buenos y malos de la muestra total

5.2 Periodo de observación y *performance*

El periodo de observación refiere al intervalo de tiempo en la ventana temporal, en donde se deja ver el comportamiento de los datos estudiados. Con esta ventana se realizan los cálculos y estimaciones necesarias para llegar al resultado final. En este caso, la ventana de observación corresponde a los periodos comprendidos entre septiembre 2020 y febrero 2021.

El periodo de *performance* refiere al tiempo en el que se hace seguimiento al cliente, con el fin de explicar en qué momento este cae en incumplimiento o *default*.

5.3 Utilización de los datos: Proceso *CRISP-DM (Cross Industry Standard Process for Data Mining*)

CRISP-DM proporciona una descripción normalizada del ciclo de vida de un proyecto estándar de análisis de datos y cubre las fases de un proyecto, sus tareas respectivas, y las relaciones entre estas tareas. (Shearer, C., 2000)
El modelo está estructurado en seis fases:

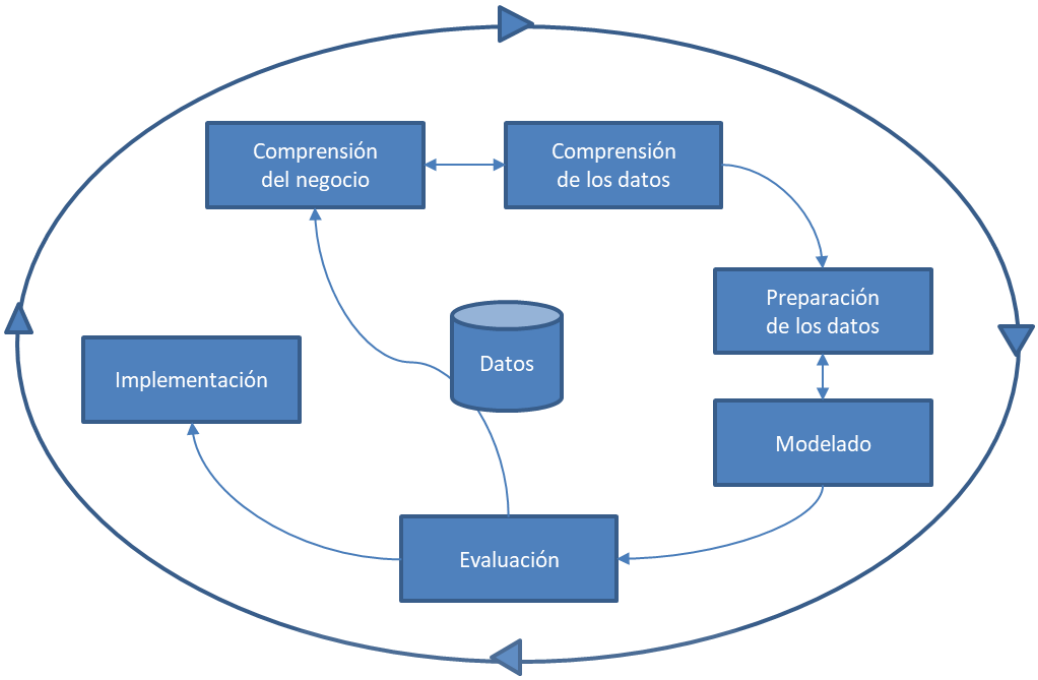


Figura 6: proceso CRISP-DM

1. Fase de comprensión del negocio o problema: se establecen los objetivos para convertirlo en un plan de proyecto.

2. *Fase de comprensión de los datos:* comprende la recolección inicial de datos, identificar su calidad y establecer las relaciones que permitan definir las primeras hipótesis.

3. *Fase de preparación de los datos*: en esta fase se procede a la preparación de los datos. Esto requiere tareas generales de selección de datos a los que se va a aplicar una determinada técnica de modelado, limpieza de datos, generación de variables adicionales, integración de diferentes orígenes de datos y cambios de formato.

4. *Fase de modelado*: en esta fase de CRISP-DM, se seleccionan las técnicas de modelado más apropiadas para el proyecto.

5. *Fase de evaluación*: en esta fase se evalúa el modelo, teniendo en cuenta el cumplimiento de los criterios de éxito del problema.

6. *Fase de implementación:* Una vez obtenido el modelo, se transforma el conocimiento adquirido en acciones dentro del proceso de negocio.

6. Construcción de un modelo de *score*

Para la creación y diseño del modelo es necesario estimar una probabilidad de incumplimiento la cual nos indique qué clientes son buenos y malos pagadores. El paso siguiente es crear un puntaje con el fin de dar un ordenamiento, y a través de éste, poder separar entre distintos grupos o perfiles de riesgo. De esta manera, se puede asignar a cada perfil una tasa de incumplimiento o *bad rate* que ayudará a distinguir entre distintas categorías de riesgo. Luego se tiene que analizar las variables y crear nuevas a partir de ellas para predecir de manera óptima. En una primera instancia se realiza el análisis bivariado para analizar el poder predictivo de cada variable con el *default* y luego un análisis multivariante, con todas ellas, para elegir las variables que luego son utilizadas como *inputs* en los modelos.

Análisis bivariado

En este apartado se detalla el análisis de las variables más explicativas para el evento de *default*. El análisis bivariado consiste en analizar cada variable de modo independiente, relacionándola con la variable dependiente. Para las variables numéricas se analizan su distribución e histograma, así como sus estadísticos descriptivos: media, mediana y percentiles. Para las variables categóricas, se analizan la frecuencia y la tasa de mora de cada clase. Posteriormente, se agrupan las clases tratando de formar grupos lo más homogéneos posibles y con suficiente población.

A continuación, se detallan las métricas utilizadas en este análisis.

1) Weight of Evidence

El peso de la evidencia o WOE indica el poder predictivo de cada categoría. Mide la diferencia entre la proporción de buenos y malos en cada grupo.

La fórmula se define de la siguiente manera:

$$WOE_i = \ln(\frac{\%\ buenos_i}{\%malos_i}),$$

Fórmula 6

donde

- i: $cada\ categoria\ o\ buckets\ que\ fue\ agrupada\ una\ variable$,
- $\%\ buenos_i$: $\frac{buenos_i}{total\ buenos}$,
- $\%\ malos_i$: $\frac{malos_i}{total\ malos}$.

Por definición, no puede haber una categoría que esté formada únicamente por buenos o malos.

2) ***Information Value (IV)***

El *Information Value* es una medida de entropía muy popular en la construcción de *scorecards*. Es un buen indicador a la hora de seleccionar variables para un modelo de regresión logística binario, como es el caso de un modelo de *scoring*.

Se calcula de la siguiente manera, para un número *k* de categorías:

$$Information\ value\ (IV)\ = \sum_{i=1}^{k}(\%\ buenos_i - \%\ malos_i) * WOE.$$

Fórmula 7

En este trabajo, el indicador es utilizado para conocer el poder predictivo de las variables. Generalmente, se establecen los siguientes criterios en cuanto al IV (Brotherton, 2013)

- Menor a 0,02: La variable no es predictiva.
- Entre 0,02 y 0,1: La variable es débilmente predictiva.
- Entre 0,1 y 0,3: La variable es predictiva media.
- Más de 0,3: La variable es fuertemente predictiva.

Ejemplo ilustrativo:

Rangos	# Buenos	# Malos	Total	% Buenos	% Malos	WOE	IV
1	998	2	1,000	10%	1%	2.06	18%
2	997	3	1,000	10%	2%	1.54	12%
3	994	6	1,000	10%	4%	0.97	6%
4	991	9	1,000	10%	6%	0.54	2%
5	989	11	1,000	10%	7%	0.30	1%
6	985	16	1,000	10%	10%	-0.03	0%
7	980	20	1,000	10%	13%	-0.30	1%
8	978	22	1,000	10%	15%	-0.41	2%
9	972	28	1,000	10%	18%	-0.63	5%
10	967	33	1,000	10%	22%	-0.82	10%

Tabla 4: Ejemplo cálculo WOE e IV

Information value (IV): 58% (sumatorio de la columna *IV* de la tabla 4)

3) Test Kolmogorov – Smirnov

Se aplica el test estadístico detallado en la sección *"2.3.1 Test Kolmogorov – Smirnov", Fórmula (2).*

Aplicando las fórmulas de este apartado es que se realiza la selección de las variables explicativas del modelo de *scoring*[1].

Ejemplo ilustrativo:

Rangos	# Buenos	# Malos	Total	% Buenos	% Malos	% Buenos acum	% Malos acum	KS
1	998	2	1,000	10%	1%	0.10	1%	9%
2	997	3	1,000	10%	2%	0.20	3%	17%
3	994	6	1,000	10%	4%	0.30	7%	23%
4	991	9	1,000	10%	6%	0.40	13%	27%
5	989	11	1,000	10%	7%	0.50	21%	30%
6	985	16	1,000	10%	10%	0.60	31%	29%
7	980	20	1,000	10%	13%	0.70	44%	26%
8	978	22	1,000	10%	15%	0.80	59%	21%
9	972	28	1,000	10%	18%	0.90	78%	12%
10	967	33	1,000	10%	22%	1.00	100%	0%

Tabla 5: ejemplo cálculo de KS

KS: 30% (máximo valor de la columna *KS* de la tabla 5)

Análisis multivariante

Una vez que la adecuación de incluir cada variable en el modelo ha sido analizada por separado, es necesario analizar todas las variables en conjunto. Esto se hace a través de

[1] Apéndice A

la matriz de correlaciones utilizando la fórmula (1). El análisis de correlaciones permite eliminar variables que para el modelo pueden ser redundantes o presenten multicolinealidad, es decir, que dos variables diferentes expliquen lo mismo. Esto permite disminuir el conjunto de variables sin una pérdida significativa de información.

Luego se clasifican las variables según el grado de correlación de la siguiente manera:

- Mayor a 70%: correlación fuerte
- Entre 30% y 70%: correlación media
- Menor a 30%: correlación débil

Si hay variables correlacionadas fuertemente se elimina la que tiene menor *IV,* fórmula (7) o KS, fórmula (2) frente a la variable dependiente. Las variables con correlación media se analizan subjetivamente si son necesarias mantenerlas en el modelo[2].

Muestra de *training* y *testing*

Es una práctica común partir la muestra en dos submuestras, la primera es conocida como muestra de entrenamiento (*training*) y se utiliza para estimar el modelo, la segunda se conoce como muestra de validación (*testing*). Para este estudio y creación de un *score* crediticio, se elige para la muestra de entrenamiento un 80% de la muestra original, dejando un 20% para testear los resultados y validar que ambas poblaciones tengan la misma predicción. Con el fin de garantizar que ambas muestras tengan la misma proporción de casos morosos, se hace uso de un muestreo aleatorio simple sin sustitución.

Este procedimiento se realiza *"n"* veces para lograr tener una distribución de los parámetros en distintas muestras.

Elección de variables

De forma general, se realiza una selección de variables significativas, como se muestra en el apéndice B[3], intentando cumplir el principio de parsimonia (es decir, que aporten la mayor capacidad discriminante con el menor número de variables). Durante el proceso

[2] Apéndice B y C
[3] Apéndice B

de selección de variables, se procede a la realización de diferentes modelos tentativos, de manera que se compara el poder predictivo global de cada modelo, el correcto signo de los coeficientes de las variables así como el equilibrio en términos de pesos relativos de cada variable y perfil. Con los datos obtenidos se lleva a cabo, cuando se ha considerado necesario, un tratamiento de las variables, ya sea agrupando sus categorías (re categorización) o relacionando las variables existentes para construir otras nuevas.

6.1 Modelo de regresión logística

La regresión logística es un instrumento estadístico de análisis bivariado o multivariado, de uso tanto explicativo como predictivo. (Hosmer Jr., D. W., Lemeshow, S., & Sturdivant, R. X., 2013)

Si se busca discriminar a los solicitantes de crédito en buenos y malos según su probabilidad de pago, la regresión logística es un buen método de clasificación que se utiliza comúnmente en el *credit scoring*. La regresión logística, se aplica tanto a datos que siguen una distribución normal, como a datos que no están normalmente distribuidos.

La capacidad predictiva del modelo logístico se valora mediante la comparación entre el grupo de pertenencia observado y el pronosticado por el modelo. El modelo debe ser capaz de clasificar a los individuos en cada uno de los dos grupos: buenos o malos, basado en las variables o características de los individuos.

Para aplicar este modelo se crea una variable binaria ficticia cuya estructura es:

$$Y_i = \begin{cases} 1\ Cuando\ ocurre\ el\ fenomeno \\ 0\ Cuando\ no\ ocurre\ el\ fenomeno, \end{cases}$$

donde *i* representa cada observación.

La variable de respuesta *Y* tiene los valores cero y uno, se puede definir *Y = 1* si se trata de un mal cliente e *Y = 0* si se trata de un buen cliente. Por lo tanto, la regresión logística es un modelo estadístico de clasificación binaria que entrega la probabilidad de pertenencia a uno de los dos grupos definidos, utilizando para ello un vector $\dot{X}$ que contiene un conjunto de variables predictoras $x_i \in R^n$ con $i = \{1 \ldots N\}$, siendo *N* el

número de observaciones. Con la regresión logística se modela la probabilidad de que Y sea igual a cero dados los valores observados de las variables predictoras contenidas en *Ẋ*. Se formula como:

$$P\ (Y = 0/\dot{\mathrm{X}}) = \mathrm{P}(\ \mathrm{Y} = 0/x_1, x_2, x_3, \dots, x_N)$$

Sea $P\ (Y = 0/\dot{\mathrm{X}}) = p_i$ la probabilidad que *Y= 0*, dado el vector *Ẋ*, se define una relación entre p_i y un modelo lineal mediante una función monótona y creciente *G(x)*.

$$G\ (p_i) = \beta_0 + \beta_i^T * \ x_i$$

Donde β^T es un vector de parámetros $\beta_i^T = [\beta_1, \beta_2, \beta_3, \dots, \beta_N]$

La función que relaciona la probabilidad con la función lineal se llama función logística (figura 6) y es el logaritmo del cociente de la probabilidad y el complemento.

$$G\ (p_i) = \ln(\frac{p_i}{1 - p_i}) \ = \beta_0 + \beta_i^T * \ x_i$$

El argumento de la función logística se conoce como *odds ratio* y es un parámetro de cuantificación de riesgo. El *odds* asociado a un evento es el cociente entre la probabilidad de que ocurra y la probabilidad de que no ocurra.

$$odds = \frac{p_i}{1 - p_i}$$

Por lo tanto, derivando de la función $G\ (p_i)$ la probabilidad de pertenencia se obtiene mediante:

$$p_i = \frac{1}{1 + e^{-(\beta_0 + \beta_i^T * x_i)}}$$

Fórmula 8

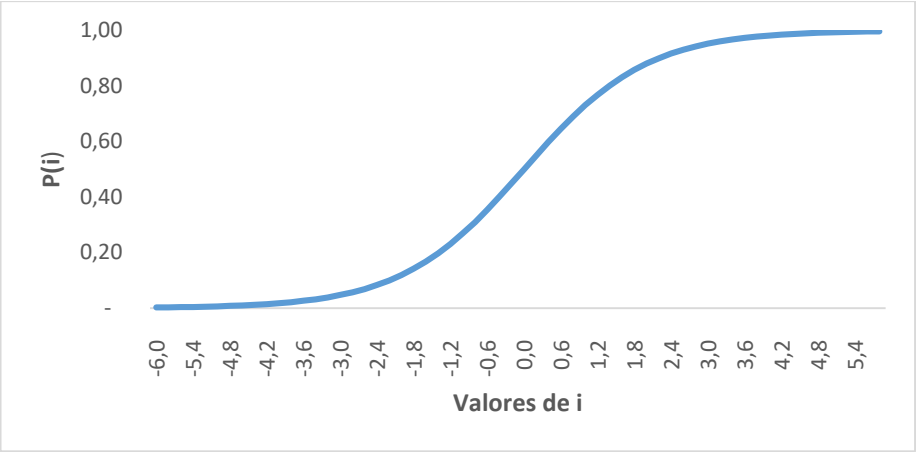


Figura 7: función logística simple

Estadístico de Wald

El Test de Wald es un contraste de hipótesis donde se trata de ver la coherencia de afirmar un valor concreto de un parámetro de un modelo probabilístico. La prueba resulta de contrastar la hipótesis nula:

$$H_0: \beta_i = 0$$

Vs

$$H_1: \beta_i \neq 0.$$

Con un estadístico definido como:

$$w_i = \frac{\hat{\beta}_i}{s(\hat{\beta}_i)},$$

que bajo el supuesto de que H_0 es cierto, sigue una distribución *t* con *n - p - 1* grados de libertad y para muestras grandes se distribuye como una normal estándar. Se entiende que si w_i es un valor alejado de cero se tiene evidencia de que H_0 es falsa, por lo tanto la región crítica de la prueba es de la forma $|w_i| > t_\alpha/2$, para un nivel de significancia α adecuado. Se entiende que si el valor de $\hat{\beta}_i$ es cero, la variable x_i se debe excluir. Otra manera equivalente de escribir la región crítica es usando el *p-value* donde $p = P(t > |w_i|)$. Donde la región crítica es de la forma $p < \alpha$.

6.2 Modelo de red neuronal

Una red neuronal está inspirada en la célula fundamental del sistema nervioso humano: “la neurona" (Jeff Heaton, 2012). Las neuronas (Figura 8) son componentes relativamente simple pero conectadas de a miles forman un poderoso sistema. De hecho, una red de una sola neurona puede ser igual a una regresión logística.

Modelo de una neurona

- w_i: son los parámetros. Son los pesos de la red.
- $U(w_i)$: sumatoria de las señales de entrada ponderadas por los pesos.
- $F(U)$: función de salida. Función de activación en una señal de salida.

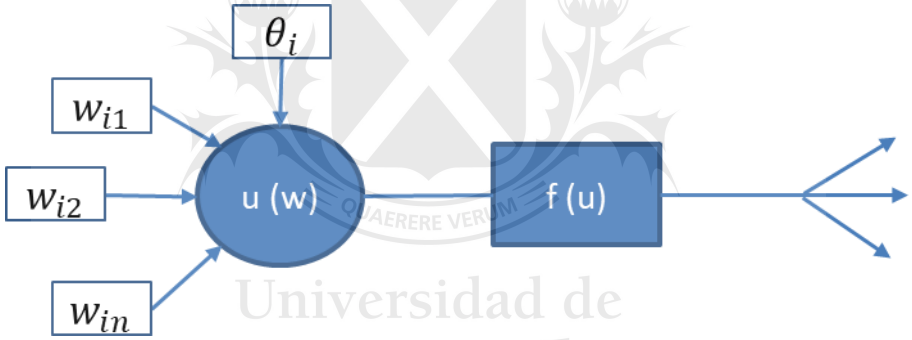


Figura 8: modelo de una neurona

Se interconectan neuronas en tres tipos de capas:

- De entrada: reciben estímulos externos.
- Oculta: elementos internos de procesamiento (se pueden estructurar en varias capas).
- De salida: reciben la información procesada y retornan la respuesta del sistema al exterior.

Red Neuronal

Las redes (Figura 9) son representadas matemáticamente por la siguiente función:

$$U_i(w_i, x) = \sum_{j=1}^{n} w_{ij} * x_j,$$

Fórmula 9

donde

- *i: indica en que neurona se encuentra,*
- *j: indica la variable que se utilizó de input.*

Luego tienen una función de activación que puede ser una simple función escalón. Para este estudio se utiliza la función de activación sigmoidal:

$$f(U_i) = \frac{1}{1 + e^{-U_i}}.$$

Fórmula 10

Por dicha función de activación, si la red se entrena con una sola neurona, se asemeja a una regresión logística.

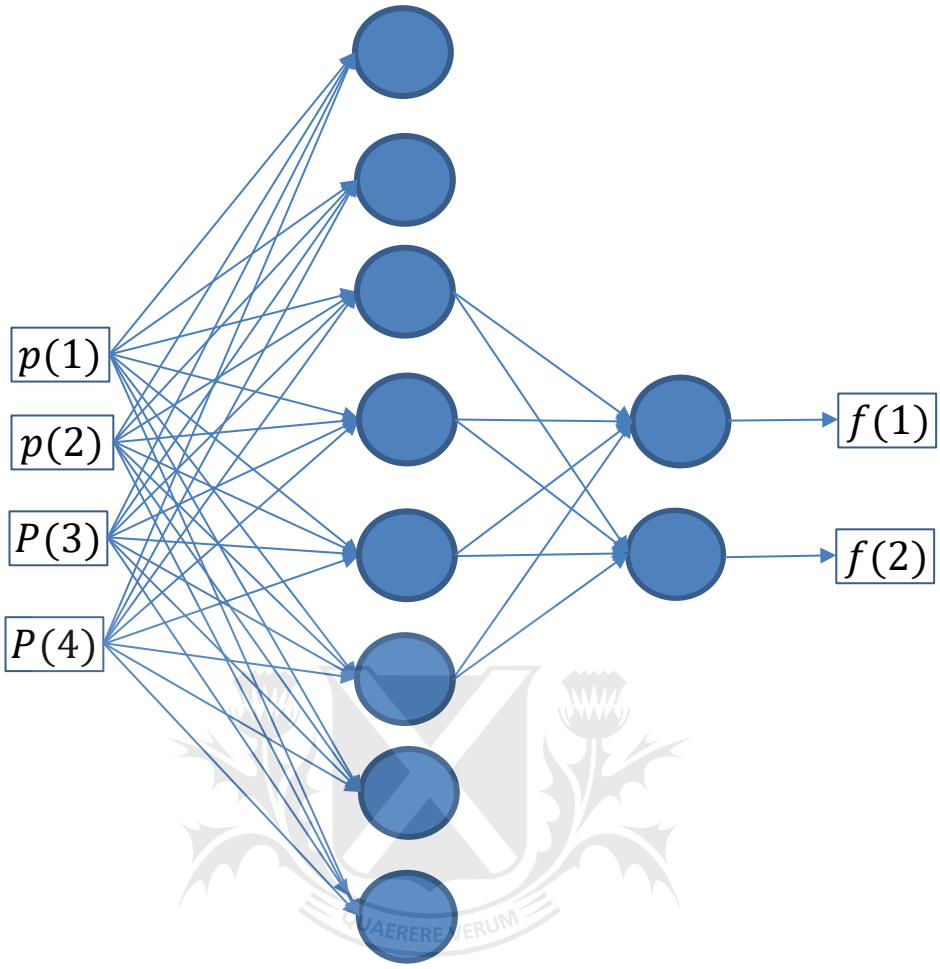


Figura 9: ilustración de red neuronal

Diseño de una red neuronal

Característica de una red neuronal: se tienen que definir las siguientes características:

- Número de capas.
- Número de neuronas por capa.
- Tipo de conexiones. Normalmente, todas las neuronas de una capa reciben señales de la capa anterior (más cercana a la entrada) y envían su salida a las neuronas de la capa posterior (más cercana a la salida de la red).

Diseño de una red neuronal para estimar la probabilidad de *default*

En el presente texto, se describe, a modo de ejemplo, cómo utilizar una red neuronal para evaluar la probabilidad de *default* asociada a cada préstamo en un conjunto de datos de créditos otorgados a diferentes clientes. Para ello, se utiliza la siguiente estructura de red neuronal.

- Capa de entrada: esta capa recibe información sobre el cliente y el préstamo, como su historial crediticio, entre otros. Se pueden utilizar una serie de variables y características relevantes para predecir el riesgo de crédito. Por ejemplo, se utilizan 11 variables diferentes.

- Capa oculta: esta capa se encarga de procesar la información recibida en la capa de entrada y extraer características relevantes para predecir el riesgo de crédito. Por ejemplo, se utiliza una capa oculta con 11 neuronas y una función de activación Sigmoide.

- Capa de salida: esta capa se encarga de predecir el riesgo de crédito asociado a cada préstamo. Se emplea una capa de salida con una sola neurona, cuyo valor representa la probabilidad de que el préstamo no se pague. Se puede utilizar la función de activación Sigmoide para que la salida esté entre 0 y 1.

La red neuronal se entrena utilizando un conjunto de datos de préstamos previamente otorgados, donde se conoce si el préstamo fue pagado en su totalidad o no. Se usa la información de estos préstamos para ajustar los pesos de las conexiones entre las neuronas y para minimizar el error en la predicción de la probabilidad de incumplimiento.

Tipo de aprendizaje en redes *feedforward* (*backpropagation*)

Según Alexander I. Galushkin (2007), podemos identificar los siguientes pasos:

Paso 1: asignar valores aleatorios a los pesos de la red.

Paso 2: mientras no se alcance un nivel deseado de predicción, se toma un dato de entrenamiento “x” y obtener una salida “y”. Si “y” no es el valor deseado para “x” entonces se actualizan los pesos.

El algoritmo actualiza los pesos según la siguiente función:

$$w' = w + \Delta w.$$

Lo importante es calcular Δw. Por lo tanto, se define una medida de error para cada patrón de entrenamiento “p”:

$$E^p = \frac{1}{2} * (d^p - y^p)^2.$$

Se deben encontrar los w_i que minimicen el error cuadrático

$$E = \sum E^p = \frac{1}{2} * \sum (d^p - y^p)^2,$$

para la siguiente regla de aprendizaje:

$$\Delta w_i = -\delta * \frac{\sigma E}{\sigma w_i}.$$

Fórmula 11

Por lo tanto, la idea es definir un vector gradiente:

$$\Delta E = < \frac{\sigma E}{\sigma w_1}, \frac{\sigma E}{\sigma w_2}, \dots, \frac{\sigma E}{\sigma w_i} >.$$

Se debe encontrar la dirección en la que este vector hace menor al error E.

Algoritmo

- Inicialización aleatoria de pesos.
- Aplicar patrón de entrada.

- Propagación de la entrada a través de todas las capas.
- La RNA genera salidas y se calcula el error para cada neurona de salida.
- Los errores se transmiten hacia atrás, partiendo de la capa de salida hacia las neuronas de la capa intermedia.
- Este proceso se repite capa por capa.
- Se reajustan los pesos de conexión de cada neurona en base al error recibido.

Construcción de la red neuronal

El número de capas y el número de neuronas por capa se estiman por prueba y error. En este trabajo se decide realizar una red de una sola capa oculta. El número de neuronas se eligió en base a un estudio de *performance* en algunas simulaciones[4].

Con respecto al tipo de conexión, se utilizó la función de activación sigmoide la cual regresa un valor entre 0 y 1, al igual que la variable objetivo a predecir.

[4] Apéndice D

7. Resultados

Para evaluar la significatividad estadística de una variable concreta dentro del modelo, se observa el valor de chi cuadrado (estadístico de Wald) correspondiente al coeficiente de la variable y su nivel de probabilidad (p-valor) para la regresión logística.
Finalmente, para verificar la bondad del modelo final y considerarlo adecuado comprueban los siguientes aspectos:

- La **significatividad** de cada una de las variables y categorías que influyen en el modelo, mediante el análisis de efectos (p valor inferior a 0.05).
- La **tendencia** que presenta cada una de las variables, de modo que sea intuitiva, asignando a las "peores categorías" una puntuación menor. En el caso de las variables continuas, se estudia si el crecimiento o decrecimiento de la puntuación conforme crece la variable tiene sentido intuitivo.
- El **KS, ROC e *Information Value***, que proporcionan una representación del poder explicativo de la calificación dada por el modelo y una medida de comparación entre distintos modelos.

7.1 Regresión logística

A partir del análisis descriptivo y bivariante de las variables, se descartan todas aquellas con bajo poder predictivo. Luego, mediante el análisis multivariante, se descartan las variables correlacionadas, seleccionando aquella con mayor predictibilidad. Por último, se construye una regresión logística.

Se realizan 1000 simulaciones para seleccionar la muestra de *training* y *testing*. Se utiliza la fórmula (8) en la muestra de *training* para estimar los parámetros y luego se valida en la muestra de *testing*. A continuación, se muestran la distribución (Figuras 10 y 11) de los parámetros de *performance* y la tabla de la distribución de promedio de la tasa de malos en deciles para entender si el modelo discrimina la variable objetivo.

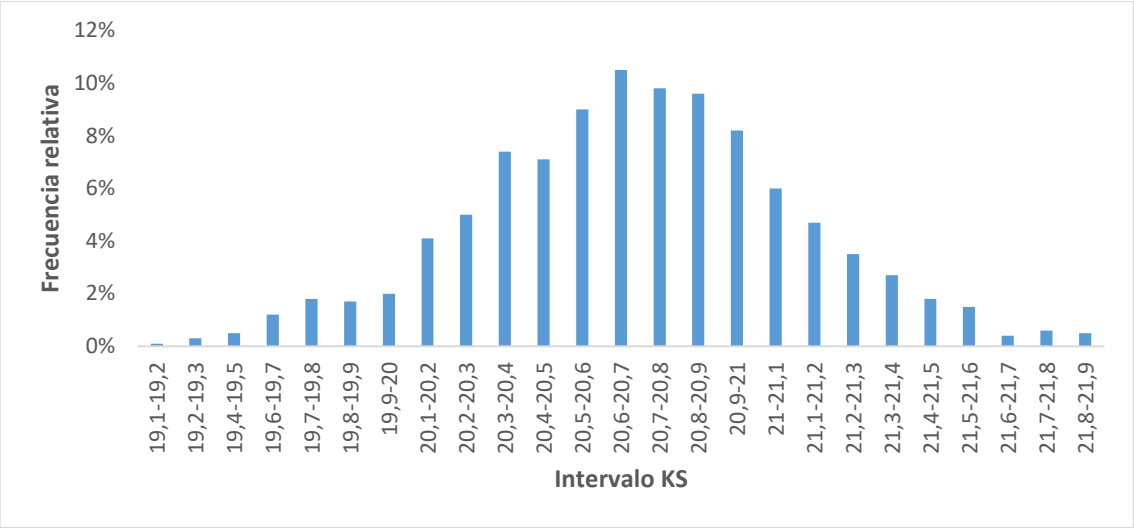


Figura 10: distribución de K-S en muestra de training

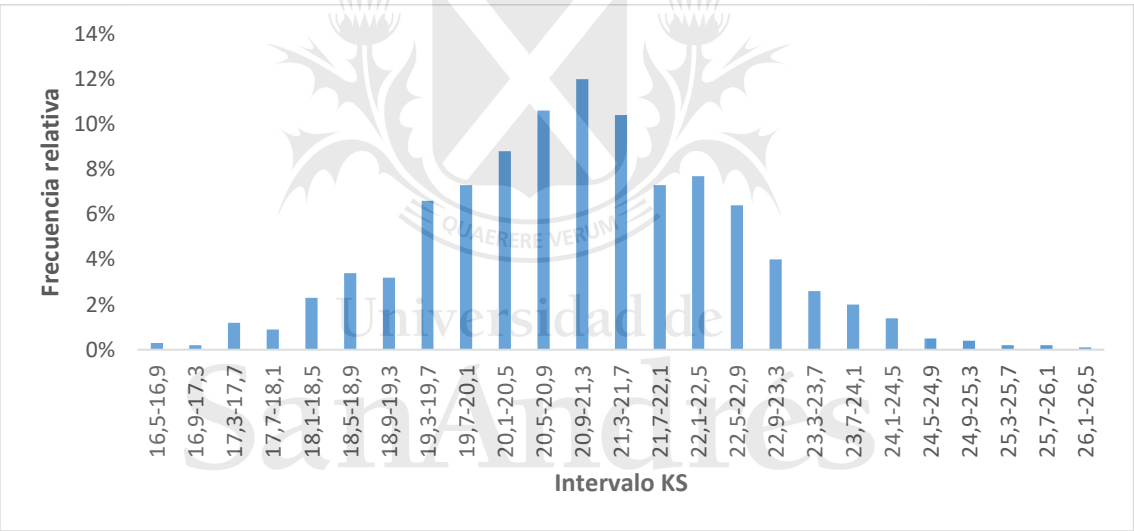


Figura 11: distribución de K-S en muestra de test

Métricas	**KS *Train***	**KS *Test***	**AUC *Train***	**AUC Test**
Promedio	20,6	21,0	64,6	64,3
Desvío	0,4	1,5	0,5	1,0
Percentil 25%	20,4	20,0	64	64
Percentil 50%	20,6	21,0	65	64
Percentil 75%	20,9	22,1	65	65
Min	19,1	16,5	64	62
Max	21,9	26,2	65	68

Tabla 6: resumen estadístico modelo reg. Logística

Distribución promedio de *bad rate* por decil en cada simulación

Decil	*Bad rate*
1	16,05%
2	23,05%
3	28,39%
4	31,42%
5	31,47%
6	34,60%
7	38,16%
8	41,30%
9	48,13%
10	59,39%

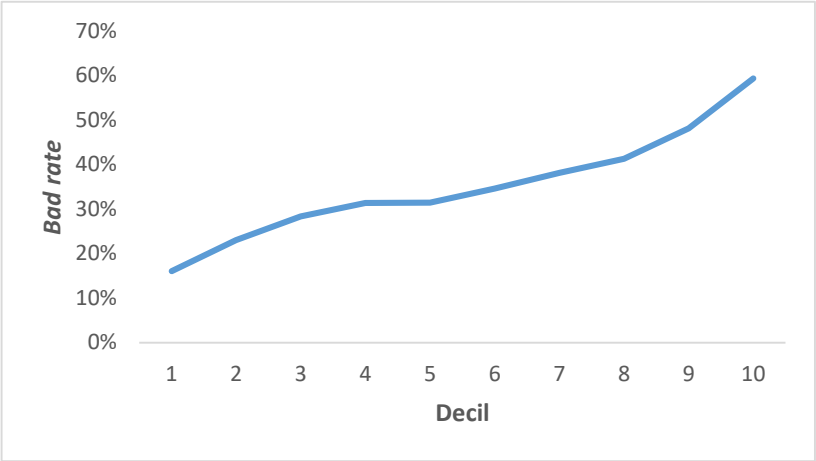


Figura 12: bad rate promedio por decil (Reg logística)

Parámetros estimados

El código realiza 1000 simulaciones independientes para estimar los parámetros de un modelo de regresión logística. En cada simulación, se genera una muestra de entrenamiento aleatoria a partir de un conjunto de datos dado. Esta muestra se utiliza para ajustar un modelo de regresión logística, obteniendo así los coeficientes para cada variable independiente. Este proceso se repite 1000 veces, generando 1000 conjuntos de coeficientes a partir de las muestras de entrenamiento. Luego, se calculan el promedio y la desviación estándar de estos coeficientes a lo largo de las 1000 simulaciones. Estos valores promedio y de desviación estándar proporcionan una idea de la tendencia central y la dispersión de los parámetros estimados a partir de las diferentes muestras, lo que ayuda a comprender mejor la estabilidad y la variabilidad de los coeficientes del modelo.

Variable	Betas	
	Promedio	Desvio
low_balance_days_mod	0.00880	0.00048
exigencia_mensual	- 0.00065	0.00004
ingresos_2	- 0.00004	0.00002
atm_usage_ratio_mod	0.06903	0.04623
account_age	- 0.00173	0.00009
flag_snp_0	0.16727	0.07999
sum_0_3_1	0.11069	0.03391
sum_0_3_2	0.08213	0.04050
sum_0_3_3	0.03390	0.02307
flag_jornalero_0	0.08264	0.03512
flag_ppimoneyman_ult_3m_0	- 0.33404	0.03171

Tabla 7: resumen estadístico coeficientes de la Reg. Logística

En la regresión logística, los coeficientes representan el cambio logarítmico en la odds ratio (la relación de probabilidades) asociada con un cambio de una unidad en la variable independiente, manteniendo constantes todas las demás variables.

Si el coeficiente es positivo para una variable, se espera que un aumento en esa variable independiente esté asociado con un aumento en la probabilidad de que la variable dependiente sea 1. Por otro lado, si el coeficiente es negativo, se espera que un aumento en esa variable esté asociado con una disminución en la probabilidad de que la variable dependiente sea 1.

De los resultados obtenidos podemos ver que el signo asociado a cada coeficiente es el correcto.

7.2 Red neuronal

Se utilizan las mismas variables que fueron utilizadas en la regresión logística para poder comparar la performance. De la misma forma que en el modelo de regresión logística, se realizan 1000 simulaciones para seleccionar la muestra de *training* y *testing*. Se utilizan la fórmula (9) y la fórmula (10) en la muestra de *training* para estimar los parámetros y luego se valida en la muestra de *testing* A continuación, se muestran la distribución de

los parámetros de performance (Figuras 13 y 14) y la tabla de la distribución promedio de la tasa de malos en deciles, para entender si el modelo discrimina la variable objetivo.

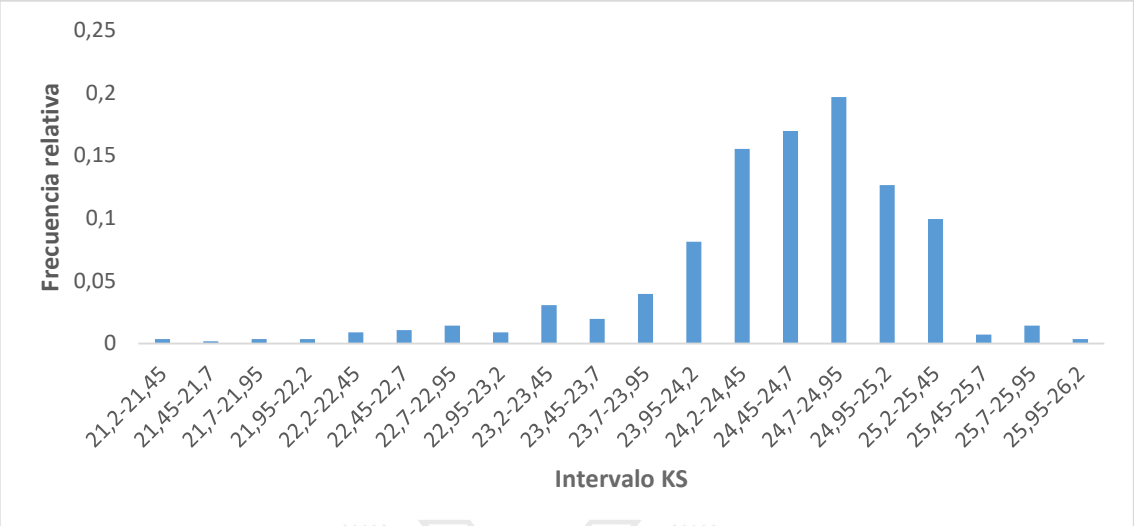


Figura 13: distribución de K-S en muestra de training

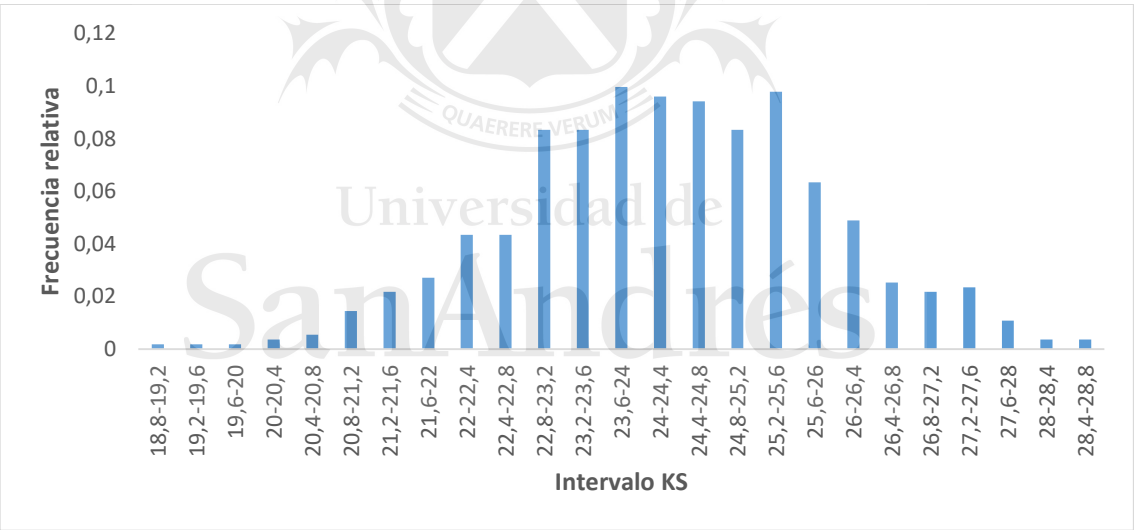


Figura 14: distribución de KS en muestra de test

Métricas	***KS Train***	***KS Test***	***AUC Train***	***AUC Test***
Promedio	24,5	24,2	67,3	66,7
Desvío	0,7	1,6	0,6	1,0
Percentil 25%	24,2	23,2	67.	66
Percentil 50%	24,6	24,2	67	67
Percentil 75%	25,0	25,3	68	67
Min	21,2	18,8	65	63
Max	26,2	28,7	68	69

Tabla 8: resumen estadístico modelo red neuronal

Distribución promedio de *bad rate* por decil en cada simulación

Decil	*Bad rate*
1	13,45%
2	21,44%
3	24,80%
4	29,92%
5	31,44%
6	35,36%
7	39,84%
8	41,60%
9	52,16%
10	64,59%

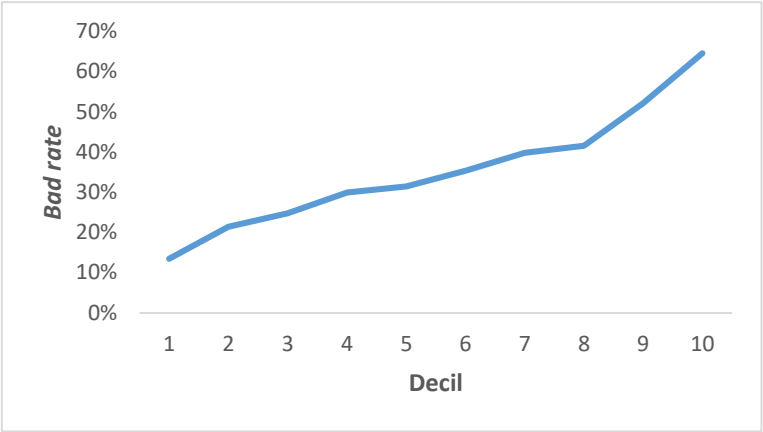


Figura 15: bad rate promedio por decil (red neuronal)

7.3 Análisis de componentes principales

El análisis de componentes principales (*principal component analysis*) o *PCA* es una de las técnicas de aprendizaje no supervisado, según Jolliffe, I. T. (2002). Una de las aplicaciones de *PCA* es la reducción de dimensionalidad (variables), perdiendo la menor cantidad de información (varianza) posible o también sirve como herramienta para la visualización de datos. Para esto último se utiliza esta técnica en este trabajo. Dado los resultados similares entre ambos modelos presentados.

Los componentes principales son una combinación lineal normalizada de las variables originales de un set de datos. Para lograrlo primero se estandariza la matriz de variables con la siguiente formula:

$$\frac{x_i - \ media(x)}{desvio(x)}$$

Una vez realizado ese procedimiento, se calculan los autovalores, fórmula (3), y autovectores, fórmula (4), de la matriz de covarianza. La primera componente principal es aquella cuya dirección refleja o contiene la mayor variabilidad en los datos (por lo que esta componente será la que más información contenga). Este vector define la línea lo más próxima posible a los datos y que minimiza la suma de las distancias perpendiculares entre cada dato y la línea representada por la componente. Esto se da en el componente que más alto valor tiene el autovalor.

Si se suman todos los autovalores, se tiene la varianza total de los componentes y si se quiere obtener la varianza explicada por los primeros *m* componentes teniendo un total de *n* componentes se calcula de la siguiente manera:

$$\sum_{i}^{m} Var(\lambda_i) = \frac{\sum_{i}^{m} \lambda_i}{\sum_{i}^{n} \lambda_i}$$

Fórmula 12

A continuación se muestran los resultados obtenidos para nuestro estudio:

Autovalores en orden descendiente	Varianza explicada individual	Varianza explicada acumulada descendiente
2,08	18,9%	18,9%
1,63	14,8%	33,7%
1,23	11,2%	44,9%
1,17	10,6%	55,5%
1,11	10,1%	65,6%
0,93	8,4%	74,0%
0,87	7,9%	81,9%
0,80	7,2%	89,2%
0,52	4,7%	93,9%
0,46	4,2%	98,1%
0,21	1,9%	100,0%

Tabla 9: Autovalores y varianza explicada

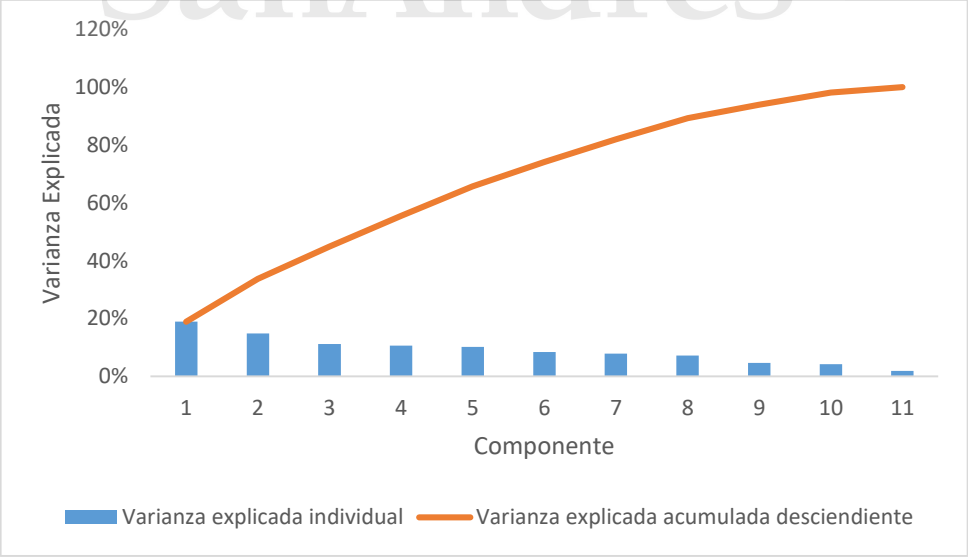


Figura 16: Grafico de varianza explicada

No existe una regla para seleccionar el número de componentes principales que son suficientes para un análisis. Como resaltamos al principio de este estudio, el fin de realizar PCA es para la visualización de los datos. Por lo tanto, lo deseable es poder capturar con los primeros 2 o 3 componentes una variabilidad suficiente para sacar conclusiones.

Si se quiere graficar en dos dimensiones, se captura el 33,7% de la varianza y 44,9% con las primeras 3 componentes. A continuación se muestran ambos gráficos.

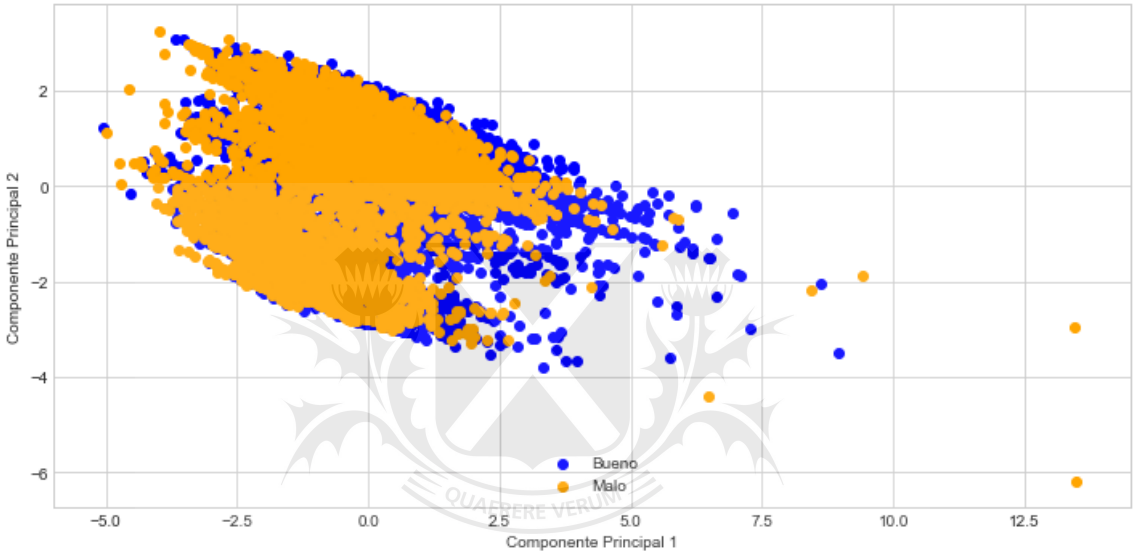


Figura 17: Gráfico en Python sobre el eje de los dos primeros componentes

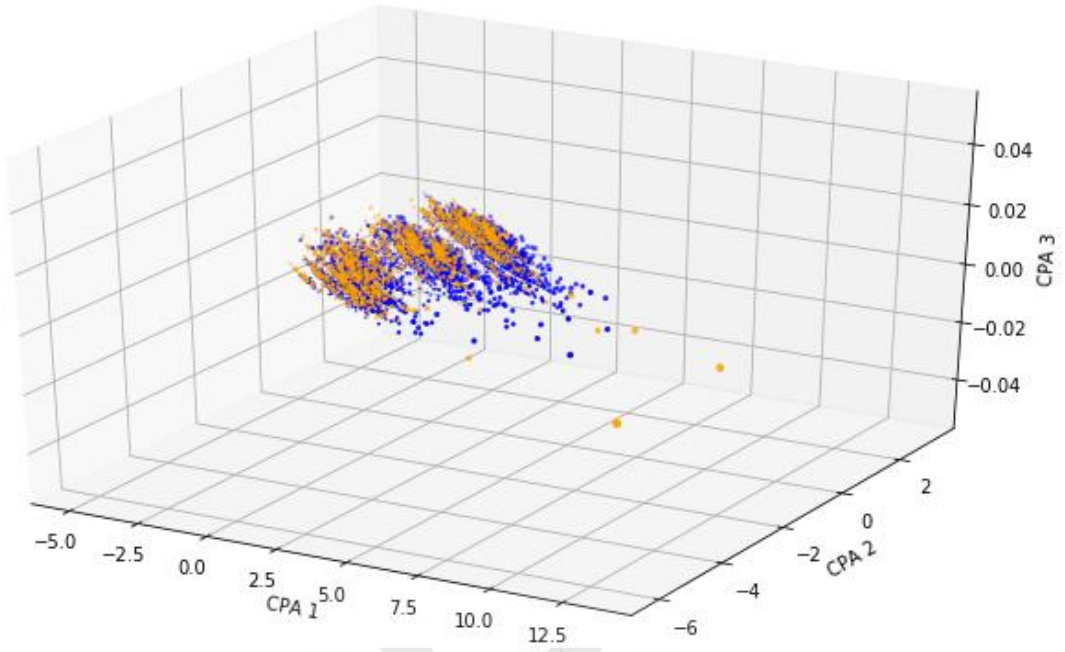


Figura 18: Gráfico en Python sobre el eje de los tres primeros componentes

8. Conclusiones

La investigación realizada diseña un modelo de calificación estadística capaz de predecir con una potencia deseable la probabilidad de *default* de los clientes A este respecto, las medidas de valoración del modelo globalmente indican un ajuste aceptable en regresión logística.

Respecto al modelo realizado mediante una red neuronal, se utilizan las mismas variables que en la regresión logística. Al introducir varias neuronas en la capa oculta para intentar capturar relaciones no lineales entre las variables, se logró una mejor distinción del *bad rate* entre el primer y el último decil, lo que resultó en una predicción más precisa del impago. Sin embargo, no se observa una mejora significativa en el rendimiento del modelo

Los resultados de realizar *PCA* arrojan que son necesarias la mayoría de las variables utilizadas para explicar la varianza total. Ya que para obtener como mínimo el 80% de la varianza, es necesario contar con 7 de los 11 componentes. Esto deja una señal que el tratamiento, realizado previamente a la construcción de los modelos (análisis bivariado y análisis de correlación), es de gran utilidad y sirve para capturar la mayor relación con la variable dependiente u objetivo. Por ello, no se observa una mejora significativa en la *performance* del modelo de red neuronal respecto al modelo de regresión logística.

Diferencias entre los modelos planteados

Un modelo de red neuronal y una regresión logística son dos técnicas de aprendizaje automático que se pueden utilizar para predecir la probabilidad de un evento binomial como el *default*. A continuación se detallan las principales diferencias entre ambos:

- Capacidad de modelado: Una red neuronal puede modelar relaciones más complejas y no lineales entre las variables de entrada y la variable de salida, mientras que una regresión logística se limita a relaciones lineales.

- Interpretabilidad: La regresión logística proporciona una interpretación más directa de los coeficientes, lo que permite identificar qué variables tienen un

mayor impacto en la probabilidad de *default*. En cambio, la interpretación de una red neuronal es más difícil, ya que puede haber múltiples capas ocultas que no tienen una interpretación clara.

- Tamaño del conjunto de datos: Las redes neuronales generalmente requieren un conjunto de datos más grande para obtener resultados precisos, ya que tienen muchos más parámetros que ajustar que una regresión logística. La regresión logística se puede utilizar con conjuntos de datos más pequeños.

- Tiempo de entrenamiento: Entrenar una red neuronal puede llevar más tiempo que entrenar una regresión logística, especialmente si se trata de una red neuronal profunda.

Bibliografía

1. Alexander I. Galushkin (2007). *Neural Networks Theory*
2. Bessis, J. (2010). *Risk management in banking. New York: Wiley.*
3. Canavos, G. C. (1984). *Applied probability and statistical methods. Boston: Little, Brown and company.*
4. Ciby Joseph (auth.). *Advanced Credit Risk Analysis and Management-Wiley* (2013).
5. Hosmer Jr., D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression.*
6. Jeff Heaton. *Introduction to the Math of Neural Networks*. 2012
7. Jolliffe, I. T. (2002). *Principal Component Analysis*
8. Ross, S. (2007). *Introduction to Probability Models (9ª ed.). San Diego: Academic Press.*
9. Shearer, C. (2000). *The CRISP-DM model: The new blueprint for data mining. Journal of Data Warehousing, 5(4), 13-22.*

Apéndices

A. Análisis bivariado

A continuación, se detallan los cálculos del apartado 6 (Construcción de un modelo de *score*) para las variables estudiadas.

B. Análisis predictivo de las variables

A continuación, se muestran la distribución de cada variable, la *bad rate* y los parámetros de poder predictivo *KS* e *IV*.

Low_balance_day: Días en donde la cuenta es negativa.

Low_balance_day	buenos	malos	total	Bad_rate	KS	IV
<=11	2.437	678	3.115	22%	12%	7,86%
<=33	2.504	1.181	3.685	32%	15%	0,46%
<=55	2.259	1.345	3.604	37%	13%	0,20%
>55	2.925	2.296	5.221	44%	0%	4,73%
					15%	13,25%

Exigencia mensual: monto en Euros de pagos realizados los últimos 3 meses

exigencia_mensual	buenos	malos	total	Bad_rate	KS	IV
<=0	1.704	1.158	2.862	41%	4%	0,95%
<=75	1.269	1.119	2.388	47%	12%	3,78%
<=207	1.815	1.066	2.881	37%	13%	0,11%
<=625	3.092	1.462	4.554	32%	10%	0,55%
<=890	905	344	1.249	28%	7%	0,96%
>890	1.340	351	1.691	21%	0%	5,00%
					13%	11,35%

Flag_snp: marca de buen pagador en buro de crédito.

flag_snp	buenos	malos	total	Bad_rate	KS	IV
0	5.748	3.808	9.556	40%	12%	2,47%
1	4.377	1.692	6.069	28%	0%	4,24%
					12%	6,72%

Ingresos: ingresos máximos de los últimos 2 meses

flag_ingresos	buenos	malos	total	Bad_rate	KS	IV
<=815	1.455	1.175	2.630	45%	7%	2,77%
<=1170	2.599	1.631	4.230	39%	11%	0,58%
<=1875	3.383	1.694	5.077	33%	8%	0,21%
>1875	2.688	1.000	3.688	27%	0%	3,17%
					11%	6,73%

Atm_usage_ratio:porcentaje de utilización de cajero automático

r_atm_usage_ratio	buenos	malos	total	Bad_rate	KS	IV
<=0.15	6.599	3.052	9.651	32%	10%	1,56%
<=0.2	1.125	649	1.774	37%	9%	0,04%
>0.2	2.401	1.799	4.200	43%	0%	2,89%
					10%	4,49%

Sum_0_3: suma de créditos a favor en los últimos 3 meses

sum_0_3	buenos	malos	total	Bad_rate	KS	IV
<=2	2.442	1.809	4.251	43%	9%	2,72%
>2	7.683	3.691	11.374	33%	0%	1,08%
					9%	3,80%

Account age: antigüedad de la cuenta

account_age	buenos	malos	total	Bad_rate	KS	IV
<=86	1.306	1.063	2.369	45%	6%	2,60%
>86	8.819	4.437	13.256	34%	0%	0,49%
					6%	3,09%

r_flag_ppimoneyman_ult_3m: saco préstamo en una empresa similar en los últimos 3 meses

r_flag_ppimoneyman_ult_3m	buenos	malos	total	Bad_rate	KS	IV
0	6.441	3.195	9.636	33%	6%	0,50%
1	3.684	2.305	5.989	39%	0%	0,78%
					6%	1,28%

flag_jornalero: cobra el sueldo un día viernes.

flag_jornalero	buenos	malos	total	Bad_rate	KS	IV
0	1.269	1.155	2.424	48%	8%	4,37%
1	8.856	4.345	13.201	33%	0%	0,86%
					8%	5,23%

C. Análisis multivariante

El análisis multivariante tiene como objeto la identificación de las relaciones entre las diferentes variables analizadas. En esta línea, se estudia las correlaciones de las variables con el fin de evitar la inclusión de información redundante, mejorando el poder discriminante del modelo. A continuación, se muestra la matriz de correlaciones de las variables seleccionadas. Para calcular la correlación entre cada posible variable explicativa se utiliza la correlación lineal de Pearson.

Variables	low_balance_days_mod	exigencia_mensual	flag_snp	ingresos_2	atm_usage_ratio_mod	sum_0_3	flag_jornalero	flag_ingresos_sup	account_age	flag_ppimoneyman_ult_3m	flag_bad
low_balance_days_mod	100%	-19%	-10%	-12%	24%	3%	-36%	-11%	32%	-3%	16%
exigencia_mensual	-19%	100%	34%	38%	-16%	11%	13%	36%	-12%	18%	-14%
flag_snp	-10%	34%	100%	15%	-16%	13%	11%	15%	6%	0%	-12%
ingresos_2	-12%	38%	15%	100%	-9%	14%	12%	66%	0%	0%	-8%
atm_usage_ratio_mod	24%	-16%	-16%	-9%	100%	-3%	-14%	-9%	-1%	-4%	10%
sum_0_3	3%	11%	13%	14%	-3%	100%	12%	12%	25%	-9%	-9%
flag_jornalero	-36%	13%	11%	12%	-14%	12%	100%	11%	5%	-2%	-11%
account_age	32%	-12%	6%	0%	-1%	25%	5%	1%	100%	-12%	-4%
flag_ppimoneyman_ult_3m	-3%	18%	0%	0%	-4%	-9%	-2%	0%	-12%	100%	5%
flag_bad	16%	-14%	-12%	-8%	10%	-9%	-11%	-8%	-4%	5%	100%

D. Selección de neuronas en la capa oculta de la red neuronal

Si bien, no hay una formula o ecuación para construir la red. En este estudio se realizan diferentes pruebas con distinta cantidad de neuronas. Se comparan los resultados realizando 10 simulaciones en cada corrida, agregando una neurona a la capa oculta, tal como se muestra a continuación:

Neuronas	*ks_train*	*ks_test*	*auc_train*	*auc_test*
2	21,35	21,07	64,8	64,4
3	21,12	21,77	64,9	64,7
4	21,15	21,61	64,6	64,7
5	21,39	21,81	64,8	64,7
6	21,2	21,65	64,8	64,8
7	21,29	22,28	64,9	64,9
8	21,22	21,93	64,9	64,7
9	21,39	21,65	64,9	64,4
10	21,45	22,07	65	65
11	21,1	22,19	65	65,3
12	21,07	22,33	64,9	65,2
13	21,14	21,87	65	64,7
14	21,4	21,49	65	64,8
15	21,29	22,12	65	64,8
16	21,63	20,97	65	64,7
17	21,3	22,06	64,9	65,2
18	21,13	22,38	65	64,8
19	21,25	22,24	65	65,1
20	21,41	21,72	65	64,5
21	21,55	22,36	65	65,2
22	21,41	22,4	65	65,1

Tabla 10: resultado de performance utilizando distintos números de neuronas en la capa oculta

Al no mirar solamente un punto en la distribución, se utiliza la métrica *AUC* en muestra de Test para decidir cuantas neuronas se ponen en la capa oculta, para luego entrenar una red con más simulaciones y encontrar los parámetros del modelo. Por lo tanto, se utiliza una red de una capa oculta con 11 neuronas, la cual cada una recibe de input las 11 variables del modelo (más el *bias*) y luego hay una capa de salida de una neurona para predecir el impago.