TRABAJO FIN DE MÁSTER

- MEMORIA-

CREACIÓN DE UN MODELO DE PREDICCIÓN DE DEFAULT PARA MICRO-EMPRESAS Y EMPRENDEDORES

MÁSTER UNIVERSITARIO EN INSTITUCIONES Y MERCADOS FINANCIEROS

MARIANO RUIZ

XABIER VALCÁRCEL

ANTONIO JAVIER VALDERRAMA

MANUEL VERDEJO

Nos gustaría aprovechar este espacio para agradecer a CUNEF la oportunidad que nos ha brindado de sentirnos partícipes de una institución de tan reconocido prestigio. Además de ello, nos ha permitido conocer a un grupo de enorme calidad humana formado por alumnos, profesores y trabajadores del propio centro, gracias a los cuales hemos experimentado un gran desarrollo personal y profesional.

En particular, queremos agradecer a cada uno de los profesores que han puesto su pequeño pero importante granito de arena en nuestra formación, aportando un gran valor añadido con sus experiencias profesionales y personales. Muy especialmente, queremos agradecer a nuestro tutor, Jorge Morán Santor, en representación de todos ellos, por su capacidad de sacrificio, atención y disponibilidad para no solo darnos la oportunidad de elaborar este TFM sino también para guiarnos y darnos la posibilidad de aprender durante el camino.

Finalmente, no podemos olvidarnos de todos nuestros seres queridos que, con su esfuerzo y apoyo incondicional, nos han permitido disfrutar y superar una de las experiencias más enriquecedoras de nuestra vida.

Javier, Manuel, Mariano y Xabier.

Índice de Contenidos

1 INTRODUCCIÓN

A menudo, las pymes y empresas encuentran serias dificultades para obtener acceso a financiación ajena y poder llevar a cabo su actividad. Para muchas de estas empresas, el no tener acceso a esta financiación supone el cese de su actividad.

Para poder mitigar estos problemas de acceso al crédito con los que se encuentran, lo ideal sería contar con un modelo que permitiera evaluar el riesgo de que estas empresas no fueran capaces de hacer frente a sus compromisos de pago en el momento en el que se les conceda una financiación, con el fin de que tuvieran acceso a crédito el mayor número posible de pymes con capacidad de pago y no ser discriminadas por un mal análisis.

Por este motivo el presente documento tiene como principal objetivo la creación de un modelo que logre discernir aquellas variables significativas que permitan evaluar el riesgo de default de una empresa.

Es importante destacar que el trabajo realizado no es un *credit rating*, ya que este engloba aspectos tales como la capacidad de pago o impago de una empresa, la capacidad de esta de hacer frente a obligaciones futuras, su capacidad de endeudamiento o el coste de capital entre otros. Es decir, aspectos tanto cualitativos como cuantitativos, que permiten poder elaborar un rating que permita diferenciar y clasificar las distintas empresas.

En el caso de este trabajo, se basa únicamente en aspectos cuantitativos para lograr el comentado de objetivo de identificar las variables significativas para evaluar el riesgo de que se produzca default. Para la selección de estas variables se ha realizado un análisis de la literatura relacionada al mismo tiempo que se han analizado las distintas variables presentes en la base de datos utilizada.

Con el objetivo definido, consideramos que este documento podría ayudar a las instituciones financieras y avalistas como herramienta de apoyo, teniendo en cuenta que en ningún momento sería un sustituto del analista, sino una herramienta que éste podría utilizar para optimizar su trabajo.

En la primera parte del trabajo se describe con un mayor nivel de detalle los objetivos y propósitos perseguidos, discerniendo entre los principales objetivos que se plantearon al comienzo de la realización de este trabajo y otros con un carácter secundario.

A continuación, y con el objetivo de adquirir una base que nos permita afrontar con garantías la realización de este trabajo realizaremos un análisis de la literatura relacionada con la temática del TFM. En este apartado se muestran los autores más relevantes sobre el comentado tema, y se realizará un análisis con sus principales ideas, metodologías y variables utilizadas. Consideramos importante remarcar que la revisión de la literatura planteada en este TFM puede no ser exhaustiva asumiendo la necesidad de seguir profundizando en la investigación.

Seguidamente, y debido a que nuestro principal objetivo es la creación de un modelo que nos permita distinguir las variables significativas para evaluar el nivel de riesgo de default, realizaremos un marco teórico de los modelos *logit* y *probit*, para explicar detalladamente los motivos y razones que han hecho que finalmente nos decantáramos por uno o por otro a la hora de la realización del trabajo.

Como punto final del trabajo se detallará profundamente la metodología seguida para la obtención final de nuestro modelo, presentando las conclusiones alcanzadas en la elaboración de éste.

2 OBJETIVOS Y PROPOSITO PERSEGUIDO

El principal objetivo de este trabajo es la creación y desarrollo de una metodología que nos permita identificar aquellas variables que resultan significativas a la hora de predecir un posible evento de crédito en la concesión de avales o garantías a empresas y pymes madrileñas.

De acuerdo con la consecución de este objetivo principal, el trabajo plantea otros objetivos secundarios que permitirán afrontar con éxito la realización del trabajo.

Uno de estos objetivos es aprovechar la realización del trabajo para profundizar nuestro conocimiento acerca de modelos de valoración de riesgo de crédito y metodologías de *credit rating*, para lo que realizaremos una lectura y estudio de la literatura relacionada más relevante sobre la evolución del riesgo de crédito que nos permitirá conocer mejor la evolución que han seguido los modelos, las variables más utilizadas o significativas, la capacidad de predicción de los distintos modelos e incluso las críticas recibidas o en qué momento se encuentra la materia en la actualidad.

Este proceso de documentación servirá como paso previo a la realización del modelo, y nos ayudará a tener una sólida base de conocimientos que nos permitirá afrontar con ciertas garantías el trabajo.

Por otro lado, el siguiente objetivo propuesto es la elección, en base a lo anteriormente estudiado, de un modelo que consideremos adecuado para la evaluación del riesgo en empresas con las características anteriormente comentadas.

Éste será uno de los principales pasos del trabajo, ya que todo lo realizado posteriormente se basará en esta elección, por ello deberemos fijarnos en una serie de variables con los conocimientos adquiridos en la etapa previa para seleccionar correctamente el modelo que mejor se adapte a nuestras necesidades.

Otro objetivo pretendido con la realización de este trabajo es el manejo y utilización de información financiera real, en esta ocasión se ha utilizado una base de datos empresarial real[1] sobre la que se realizará un trabajo de depuración de información, que nos permita filtrar aquellas variables que aporten valor al trabajo, y posteriormente seleccionar aquellas variables que consideremos significativas.

Con esta base de datos, se pretende obtener también la efectividad del modelo seleccionado. Mediante métodos econométricos, y con las variables seleccionadas, pondremos a prueba el modelo con datos de empresas reales y podremos obtener el porcentaje de acierto del modelo a la hora de detectar la probabilidad de quiebra de una empresa.

[1] Base de datos cedida por una SGR para uso exclusivamente académico: Empresas con domiciliación en la Comunidad de Madrid

Con los resultados obtenidos, el último objetivo del trabajo supondrá la reflexión y puesta en común de la aplicación del modelo desarrollado, analizando las distintas variables usadas y la importancia de estas y hacer una crítica del modelo.

3 LITERATURA RELACIONADA

Los orígenes del *credit scoring* se atribuyen a David Durand, profesor del MIT que en 1941 publica un estudio sobre buenos y malos préstamos realizado a 37 empresas. En dicho estudio, Durand (1941) subraya la importancia de la calidad de los datos.

Altman (1968) publica un modelo de predicción que determinaba el grado de cercanía de una empresa a la quiebra, basado en la combinación de los estados financieros de una empresa con su valor de mercado. El modelo, mejor conocido como Z-score, emplea una metodología estadística de análisis discriminante.

En la década de 1980 se comienza a usar una metodología de estimación por máxima verosimilitud del modelo logístico condicional, prueba de ello son las obras de Ohlson (1980) y Wiginton (1980). Este último destaca por una mayor proporción de clasificaciones correctas que el propio Z-Score aunque recalca que la capacidad predictiva de cualquier modelo depende de la disponibilidad completa de información.

Ya en la década del 2000, surgen metodologías para estimar los diferentes ratios de riesgo de la cartera minorista de una entidad financiera. El trabajo de Westgaard & Van de Wijst (2001) es un ejemplo en el que se muestra que se han desarrollado en el ámbito empresarial y académico modelos y sistemas que monitorizaban la evolución del riesgo que emergía de sus líneas de negocio.

El incremento de presión procedente de las condiciones del mercado y una clara tendencia de las autoridades a incluir los sistemas de control del riesgo dentro del esquema regulatorio (léase Basilea I, II y III) estimulan adicionalmente este desarrollo de Westgaard y Van de Wijst (2001). Aparecen modelos que relacionan el *credit scoring* con el *behavioral scoring*, como el de Thomas et al. (2001), así como otros que certifican la importancia de incluir "*soft information*" (o información cualitativa) como forma de mejorar la predictibilidad de los modelos Lehmann (2003).

En esta línea, Altman & Sabato (2005) aportan que los requerimientos de capital de las entidades financieras, en el marco de Basilea II, se reducirían si éstas considerasen a las PYMES clientes minoristas y no empresas.

En la mayoría de los estudios publicados y que hemos revisado, se emplean tres técnicas de minería de datos: regresión logística (RL), redes neuronales (NN) y árbol de decisión (DT). Koh et al. (2006) compara estos modelos y concluye que cada uno de los tres, mejora a los demás según el criterio que se establezca y, por tanto, depende de los aspectos que quiera enfatizar el usuario para determinar el modelo a utilizar.

Otros elaboran un modelo de predicción del riesgo para PYMES en base a indicadores de información no financiera, en el caso del modelo de Altman et al. (2010).

Así mismo, Koh et al. (2006) y Abdou & Pointon (2011) proponen el uso de varios modelos a modo de crear sinergias, que permitan obtener mejores resultados predictivos conjuntamente que por separado, aunque mantienen que la elección de dichos modelos dependerá de la aversión al riesgo del usuario.

Chen & Huang (2003) desarrollan un modelo basado en redes neuronales y algoritmos genéticos mientras que Blanco et al. (2013) se centran en una tipología de red neuronal (enfoque de percepción multicapa - MLP) para desarrollar un modelo que supera el rendimiento de las técnicas estadísticas clásicas.

También se han desarrollado modelos que permiten visualizar efectos en el riesgo crediticio ante cambios en las condiciones macroeconómicas, como los de Hwang (2013) y Jacobs & Bag, (2011), y en el ciclo económico como el de Berteloot et al. (2013). En esta línea, numerosos autores cuestionaron el modelo de Altman (1968), concretamente, por su fiabilidad en otros períodos de estudio precisamente debido a variables macroeconómicas como la inflación, el tipo de cambio o la disponibilidad de crédito. Para ello, Salimi (2015) verifica que el Z-Score sí funciona en un período diferente (2000-2005) y con una capacidad de predicción media aceptable (79,4%)

Siguiendo los modelos *logit*, Modina & Pietrovito (2014) establecen un modelo de probabilidad de default concluyendo que la estructura de capital y los gastos financieros son variables muy relevantes para predecir el default de la muestra (PYMES Italianas).

En una línea similar se presenta la revisión de Lessmann et al. (2015) sobre la clasificación realizada por Baesens et al. (2003) o Abdou & Pointon (2011), en la que confirman los mejores resultados obtenidos por modelos que siguen una técnica de red neuronal o bosque aleatorio.

Ello se confirma en Stevenson & Pond (2016), que observan que a la hora de evaluar el riesgo en los bancos alemanes y británicos, éstos últimos confían en el uso del credit scoring para asignar una probabilidad de quiebra a las empresas sin tener apenas en cuenta el uso de datos históricos cualitativos, mientras que los alemanes, por su parte, sí que tienen en cuenta el histórico de la gestión de la empresa. Siendo un hecho visible que se ha reducido la importancia de la información cualitativa para la evaluación de riesgos.

Vistos los resultados positivos de modelos lineales (modelos *logit*) y no lineales (redes neuronales), Raei et al. (2016) desarrollan un modelo híbrido que combine el modelo *logit* y de red neuronal para estimar la probabilidad de default de empresas desde el punto de vista de un banco comercial, mejorando los resultados que dichos modelos obtenían de forma individual.

Por su parte, Bequé & Lessmann (2017) o Bathia et al. (2017) desarrollan así mismo, modelos predictivos basados en el Machine Learning.

A continuación, resumimos los principales modelos, variables, capacidades de predicción y base de datos utilizados en la literatura previa revisada en el siguiente cuadro resumen:

Autor	Modelo usado	BB.DD usada	Capacidad de predicción	Variables usadas en el modelo*
(Altman, 1968)	Z-score		80-90% (1 año)	WC/ AT; RE/AT; Ebit/AT; Valor de Mercado Equity/ PT; Ventas/AT
(Iazzolino, et al., 2013)	Data Envelopment Analysis (DEA)		El modelo solo mide la eficiencia de la empresa.	Pasivo; PMC; EBITDA; Flujos de caja
(Koh, et al., 2006)	Modelo logit, Redes Neuronales y Árbol de Decisión	Base de datos de créditos alemanes	Logit: 77,0% Redes Neuronales: 73,4% Árbol de Decisión: 71,4%	Estado cuenta corriente, duración, propósito, préstamo, historial crediticio, deudores, garantes, tasa pago sobre ingreso disponible, cuenta ahorro, plan de ahorro, años de empleo, tipo propiedad del bien objeto del préstamo
(Lessmann, et al., 2015)	Redes neuronales, modelo logit, bosque aleatorio y HCES con muestra de bootstrap	Base de datos de crédito Australiano y Alemán (UCI Library). Procedentes también de instituciones en UK y Benelux.	HCES con muestra de bootstrap tiene la mejor capacidad predictiva en base a clasificadores individuales y en conjunto.	
(Abdou & Pointon, 2011)	Modelo lineal Discriminante, modelo probit y logit, árboles de decisión, redes neuronales y programación genética		Modelos complejos (destacan híbridos) tienen mejor capacidad predictiva que modelos simples	
(Michael Jacobs, 2010)				TTD, ORR, tamaño empresa (BVTA), medidas liquidez (CR), rentabilidad del PM, medida intangibilidad, el PERCBNK, PERCSEC y el MSG12MTDR
(Ohlson, 1980)	Modelo logit	2.163 empresas de USA entre 1970 y 1976		SIZE, TLTA, WCTA, WCTA, CLCA, OENEG, NITA, FUTL, INTWO, CHIN

Autor	Modelo usado	BB.DD usada	Capacidad de predicción	Variables usadas en el modelo*
(Wiginton, 1980)	Modelo logit y Modelo Lineal Discriminante.	1.908 individuos	Modelo Logit mejor que discriminante/ 62% y 58% en las dos sub-muestras	Nº personas dependientes, estado vida, mudanza últ. año, Uso automóvil, ocupación e industria trabajo, años empleo actual
(Westgaard & Van de Wijst, 2001)	Modelo logit	Base de datos Dun & Bradstreet y Compañías AS noruegas entre 1995 y 1999		Deuda, liquidez, solidez, antigüedad, tamaño
(Thomas, et al., 2001)	Modelos matemáticos mediante credit scoring y behavioral scoring			
(Lehmann, 2003)	Modelo logit	20.000 observaciones de PYMES Alemanas	Información cualitativa mejora el modelo	Variables cuantitativas ("soft") y cualitativas ("hard")
(Chen & Huang, 2003)	Redes neuronales y algoritmos genéticos.	Repositorio UCI de bases de datos de Machine Learning	Alrededor del 87%	15 variables independientes
(Altman & Sabato, 2005)		PYMES de USA, Italia y Australia	30% mayor que el modelo genérico de empresas	
(Altman & Sabato, 2005)	Modelo logit	20.193 empresas italianas entre 2000-2003.		Deuda/Equity, Deuda bancaria, Pasivos LP, Valor económico agregado, Activos tangibles, Cuentas a pagar, Deuda bancaria LP
(Altman & Sabato, 2005)	Modelo logit	3.073 empresas USA entre 2000-2003.		WC/AT, RE/AT; Ebit/AT, Valor en libros/Valor en libros PT,
(Altman & Sabato, 2005)	ZETA-Score model	Proporcionada por el Corporate Scorecard Group para empresas australianas		EBIT/AT, Estabilidad ganancias, EBIT/Total Intereses Pagados, RE/AT, Current ratio, Equity/CT, AT
(Altman, et al., 2010)		5,8 millones de empresas de UK entre 2000-2007	78% con información cualitativa (71% sin).	Capital empleado/PT; AD/AC; AC/PC; PT/AD; ACO/DCO; ACO/PT; DCO/AT; inventario/WC; efectivo/AT; efectivo neto/PN; RE/AT; y deuda a corto/ PN corto plazo
(Bathia, et al., 2017)	(Modelo lineal discriminante,			

Autor	Modelo usado	BB.DD usada	Capacidad de predicción	Variables usadas en el modelo*
	algoritmo de bosque aleatorio, modelo logit y XGBoost)			
(Raei, et al., 2016)	Modelo híbrido: Logit con redes neuronales	Compañías cotizadas en Tehran Stock Exchange entre 2008 y 2014	95,33%	Beneficio bruto/venta; RE/AT; Activo fijo/AT; Intereses/deuda total; Beneficio bruto/activo; Beneficio operativo/venta; EBIT/venta.
(Modina & Pietrovito, 2014)	Modelo logit	9.208 PYMES Italianas provistas por el Centrale Rischi Finanziari (CRIF) entre 2006 y 2010	81%	Composición Fuentes de financiación; rentabilidad; Eficiencia; Relación activo-pasivo; Habilidad de generar liquidez.
(Berteloot, et al., 2013)	Modelo logit	+ 6.200 compañías USA entre 1984 y 2011		Balance presupuesto; Ratio utilización capacidad; S & P/IP casas Case-Shiller; CNCF; CNCF con iva y ccadj; Índice confianza consumidor; Activos banco comercial; Crédito consumidor; BDI con iva & ccadj; IP al consumidor; Saldo cuenta corriente; Ingreso personal disponible; PIB; PNB; IP viviendas; Producción industrial; Stock de dinero M1; Precio petróleo; Tasa préstamo principal; BDI; BDI no financieros; BAI; BAI no financieros; Deuda pública; Préstamos inmobiliarios en bancos comerciales; S & P 500 - IP; Tasa ahorro; Balanza comercial; TT 1 año; TT 3 meses; TT 10 años; Tasa de desempleo;
(Blanco, et al., 2013)	Enfoque de percepción multicapa (Tipología de red neuronal)	5.541 prestadores en Perú proporcionados por Edpyme Proempresa entre 2003 y 2008	Mejor MLP -> AUC: 93,22%	Carac. personales; ratios económicos y financieros; carac. operación financiera actual; variables macroeconómicas; demora pago crédito

*Definición de las abreviaturas utilizadas en el campo "*variables usadas en el modelo*" definidas en Anexo 8.1

4 MARCO TEÓRICO DE MODELOS DE ELECCIÓN DISCRETA

4.1 INTRODUCCIÓN

Retomando el objetivo principal de este trabajo, que como comentábamos anteriormente no es otro que la creación y desarrollo de un modelo capaz de identificar las variables significativas para predecir si una empresa cumplirá o no con sus obligaciones de pago, es conveniente comenzar estableciendo un marco teórico sobre el que edificar dicho modelo.

La probabilidad de *default* hace referencia a la probabilidad de que el receptor o peticionario de un aval - en el caso de este trabajo serán PYMES y empresas registradas en la Comunidad de Madrid - incurra en un evento de crédito, es decir, que incurra en el incumplimiento total o parcial, en su importe total o en su debido momento de sus obligaciones contractuales. Es necesario señalar que la probabilidad de default comprende desde la situación en la que se incumple un pago en fecha y forma hasta la situación de bancarrota o imposibilidad de hacer frente a los pagos que se deben atender.

El evento de crédito es un fenómeno dicotómico, es decir, puede ocurrir o no y ésta es la principal diferencia entre el riesgo de crédito y el de mercado en el que la fuente de riesgo son variables de mercado, generalmente precios, que pueden tomar un continuo de valores, generalmente entre cero e infinito.

Es por ello, que este trabajo se basa en el desarrollo de un modelo dicotómico de elección discreta cuya variable endógena es una variable dependiente binaria o dicotómica y las variables exógenas son linealmente independientes. Es decir, se trata de un modelo que se utiliza para explicar fenómenos en los cuales la variable de relevancia sólo puede tomar dos valores.

Los modelos de respuesta binaria o dicotómica son aquellos cuya variable de respuesta Y puede tomar dos valores, generalmente 0 o 1. La distribución de Y es una Bernoulli cuya esperanza es:

$$E\,[Y\,] \;=\; P\,[Y \;=\; 1] \;=\; p\,(0 \;<\; p \;<\; 1)$$

Dada una variable X, posible predictora de la variable Y, entonces la distribución condicional de Y sobre un valor de X = x, también sigue una distribución de Bernoulli de forma que la esperanza condicionada de Y sobre X = x es:

$$E\,[Y\,|X \;=\; x] \;=\; P\,[Y \;=\; 1\,|\,X \;=\; x] \;=\; p(x)$$

Y la varianza condicionada.

$$Var[Y\,|\,X \;=\; x] \;=\; p(x)\;\cdot\;(1 \;-\; p(x))$$

Un modelo para la variable Y en función de X sería de la forma

$$Y \;=\; f\;(parámetros, x, error)$$

Una primera aproximación al problema sería aplicar un modelo de regresión lineal clásico para estimar Y en función de X. Si X es continua el modelo sería:

$$Y = \alpha + \beta x + \varepsilon(x)$$

Dónde los errores son variables aleatorias independientes con esperanza 0, y cuya distribución es una Bernoulli. El modelo de regresión lineal sería:

$$E[Y \mid X = x] = p(x) = \alpha + \beta x$$

Es decir, un modelo lineal para estimar la probabilidad condicionada. Este modelo adolece de varios problemas, tales como:

- Ausencia de normalidad de la variable *Y,* y de los errores ya que estos se distribuyen según una Bernoulli
- Presencia de heterocedasticidad, es decir, la varianza de la variable respuesta no es constante sobre los valores de *x,* sino que depende de la esperanza condicionada.
- No acota los valores de *p(x).* A pesar de que la probabilidad está acotada entre 0 y 1, el modelo puede predecir probabilidades fuera de ese intervalo.
- Relación lineal entre *X* y *p(x)* lo que llevaría a que variaciones iguales en X producen variaciones iguales en *p(x)*. Por lo tanto, la probabilidad en los extremos varía más lentamente.

Debido a los problemas del modelo de probabilidad lineal, se han buscado modelos alternativos de la forma

$$Y = F(\alpha + \beta x) + \varepsilon(x)$$

con $\varepsilon(x)$ variables aleatorias independientes con esperanza 0, con lo que el modelo sobre la probabilidad condicionada se puede escribir como:

$$p(x) = F(\alpha + \beta x)$$

con F función monótona creciente. También se puede expresar como sigue:

$$F^{-1}(p(x)) = \alpha + \beta x$$

es decir, se busca una función F cuya inversa transforme las probabilidades condicionadas p(x) y posteriormente, modelar linealmente esta transformación.

Según se elija una determinada función F se tienen distintas formulaciones.

En resumen, un modelo lineal de probabilidad no satisface las condiciones de homocedasticidad, se subestima el coeficiente de estimación y no acota las estimaciones realizadas entre 0 y 1, (Alamilla-López & Arauco Comargo, 2009), hechos suficientes como para decidirnos a utilizar en nuestras estimaciones un modelo no lineal como veremos a continuación: *logit* o *probit*.

Transformación *logit*. La transformación *logit* es:

$$\boxed{logit(p(x)) = ln\frac{p(x)}{1 - p(x)}}$$

Con lo que un modelo para la transformación *logit* sobre p(x) sería:

$$logit(p(x)) = ln \frac{p(x)}{1 - p(x)} = \alpha + \beta x$$

o en términos de p(x):

$$p(x) = \frac{exp\,(\alpha + \beta x)}{1 + exp\,(\alpha + \beta x)} = \frac{1}{1 + exp\,[-(\alpha + \beta x)]}$$

De esta forma, la expresión para p(x) está acotada entre 0 y 1, siendo deseable al tratarse de una probabilidad. Otra ventaja es la sencillez de la interpretación, puesto que $\frac{p(x)}{1-p(x)}$ se corresponde con la ventaja de la respuesta Y = 1 para el valor x.

Transformación *probit*. La transformación *probit* consiste en considerar como función de transformación la inversa de la función de distribución de una normal estándar N (0, 1).

$$F(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{1}{2}t^2} dt$$

y la expresión del modelo sería

$$F^{-1}\,(p(x)) = \alpha + \beta x$$

Esta transformación también acota p(x) entre 0 y 1. La función *probit* se acerca más rápidamente a probabilidades de 0 y 1 que la función *logit.*

El tipo de modelos que se obtienen mediante las transformaciones descritas, pueden considerarse un caso particular de los modelos lineales generalizados (GLM).

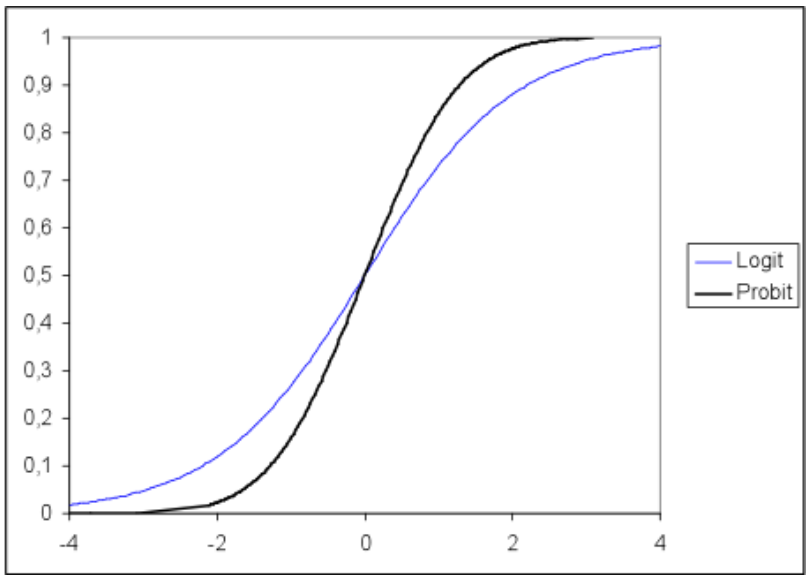
En los GLM, la varianza $Var(Y \mid X1 = x1, \ldots, Xk = xk)$ viene dada por un parámetro de escala positivo, φ y por una función de la media condicionada de y dados los valores de las variables predictoras.

$$Var(Y \mid X1 = x1, \ldots, Xk = xk) = \varphi \times f\,[\mu(x)]$$

Para las distribuciones binomial o *poisson*, $\varphi = 1$, y la varianza depende sólo de μ . Para la distribución normal (gaussiana), la varianza depende sólo del parámetro de dispersión φ que en ese caso es σ^2.

Dada la similitud existente entre las curvas de la normal estándar y de la logística, los resultados estimados por ambos modelos no difieren mucho entre sí. La menor complejidad de manejo que caracteriza al modelo *logit* es lo que ha potenciado su aplicación en la mayoría de los estudios empíricos. La función de densidad de la distribución normal no es tan fácil de integrar, de manera que los modelos *probit* generalmente requieren de la simulación. Otra diferencia entre ambos modelos es que la función logística tiene colas más anchas, por lo que la probabilidad de éxito será mayor en los extremos cuando se use el modelo *logit*. En el gráfico siguiente se aprecia cómo la función *logit* cuenta con colas más anchas que la función del modelo *probit*.

Gráfico 1. Comparativa distribución logística con distribución normal



Fuente: (Cifuentes, 2015).

La matriz de datos muestrales estará formada por n observaciones pudiendo ser el valor de la variable endógena para cada una de ellas 1 ó 0. La naturaleza dicotómica de la variable dependiente en este tipo de modelos impide la utilización de los métodos tradicionales en la estimación de los parámetros, al no poderse calcular la inversa de la varianza utilizada como ponderación del modelo. Para la estimación de los parámetros (coeficientes β_i) se utiliza el método de Máxima Verosimilitud.

Dada una variable aleatoria, caracterizada por unos parámetros, y dada una muestra poblacional, se consideran estimadores de Máxima-Verosimilitud de los parámetros de una población determinada, aquellos valores de los parámetros que generarían con mayor probabilidad la muestra observada. Es decir, son aquellos valores para los cuales la función de densidad conjunta (o función de verosimilitud) alcanza un máximo (Moral, 2003).

Suponiendo que las observaciones son independientes, la función de densidad conjunta de la variable dicotómica Yi queda como:

$$Prob\ (Y_1 Y_2 \dots\ Y_i \dots Y_n) = \prod_{i=1}^{n} M_i^{Y_i}(1 - M_i)^{1-Y_i}$$

donde M_i recoge la probabilidad de que $Y_i = 1$. Por simplicidad se trabaja con la función de densidad conjunta en logaritmos, cuya expresión es:

$$\zeta = Ln\ L = \sum_{i=1}^{i} Y_i Ln\ M_i + \sum_{i=1+i}^{n-i} (1 - Y_i) Ln\ (1 - M_i)$$
$$= \sum Y_i Ln\ M_i + \sum (1 - Y_i)\ Ln\ (1 - M_i)$$

El método de estimación de máxima verosimilitud elige el estimador del parámetro que maximiza la función de verosimilitud ($\zeta = Ln\ L$), por lo que el

procedimiento a seguir será calcular las derivadas de primer orden de esta función con respecto a los parámetros que queremos estimar, igualarlas a 0 y resolver el sistema de ecuaciones resultante. Se trata de un sistema de ecuaciones no lineales por lo que es necesario aplicar un método iterativo o algoritmo de optimización que permita la convergencia en los estimadores.

4.2 ELECCIÓN DEL TIPO DE MODELO

Una vez establecido el marco teórico de los modelos de elección discreta – modelo *logit* y *probit* – es necesario realizar una explicación de los motivos que nos han llevado a la elección de un modelo *logit* en este trabajo.

En primer lugar, como se observa en la revisión del estado del arte, la gran mayoría de autores destacados en el desarrollo de modelos de predicción se decantan por el modelo de regresión logística. Entre ellos, destacan Altman, Sabato, Wiginton, Ohlson, Martens y Baesens, entre otros.

Por otro lado, se ha pretendido demostrar empíricamente que el modelo *logit* tiene una capacidad predictiva mayor que el modelo *probit*. Para ello, se han tomado diferentes medidas de bondad de ajuste de los modelos.

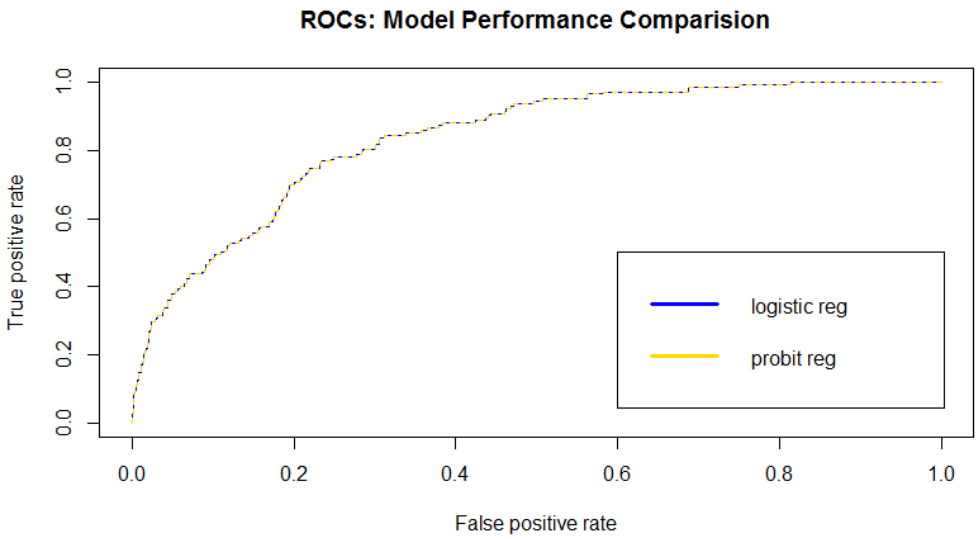
1. La Curva ROC (acrónimo de *Receiver Operating Characteristic* o Característica Operativa del Receptor) se define como una representación gráfica de la sensibilidad frente a la especificidad para un sistema clasificador binario según se varía el umbral de discriminación (valor a partir del cual decidimos que un caso es un positivo).

   La sensibilidad se define como la probabilidad de que el modelo de predicción prediga una observación como "positiva" y realmente lo sea, es decir, la proporción de observaciones verdaderamente positivas que el modelo clasifica como tales. Por el contrario, la especificidad es la probabilidad de que el modelo prediga "negativo" una observación que lo sea. El siguiente es un gráfico ROC, el cual es un gráfico bidimensional en el que el “true positive rate” (sensibilidad) se dibuja en el eje vertical y el “false positive rate” (1-especificidad) en el horizontal.

   De modo que un gráfico ROC representa el equilibrio relativo entre beneficios (verdaderos positivos) y costes (falsos negativos).

   Donde la diagonal Y=X representa la decisión de predecir aleatoriamente si la empresa cumplirá o no con sus obligaciones de pago y el punto (0,1) (vértice superior izquierdo) representa la clasificación perfecta. Por ello son preferibles los modelos más cercanos a la zona de arriba a la izquierda. En el siguiente gráfico se puede observar dicha curva ROC de los modelos logit y probit, respectivamente, así como el AUC (área under the curve), que facilita la interpretación de la curva ROC.

Gráfico 2. Comparaciones Curvas ROC de los modelos logit y probit.



Fuente: Elaboración propia a través de R.

Tabla 1. Comparación Área AUC de los modelos logit y probit.

	LOGIT	PROBIT
AUC	0,836	0,834

Fuente: Elaboración propia a través de R.

2. Como se acaba de explicar, se sabe que el modelo cuya curva ROC se acerque más a la esquina superior izquierda tendrán más precisión. Sin embargo, existe cierta complicación a la hora de observar qué curva se aproxima más a dicha esquina. Es por ello que se analiza el AUC (*Area Under Curve*) o área bajo la curva. Si el modelo con mayor capacidad de predicción es aquel que tiene una curva ROC más cercana a la esquina superior izquierda de la gráfica, este coincidirá con el modelo que deje una mayor área entre la curva y la diagonal que supone elegir aleatoriamente.

   Por lo tanto, el área bajo la curva tomará valores entre 0 y 1, correspondiendo el valor 0.5 a la elección aleatoria. En este caso, el modelo *logit* cuenta con un AUC de 0,836 y el modelo *probit* con un AUC de 0,834. A pesar de contar con un AUC más elevado el modelo *logit*, la diferencia es tan mínima que no se puede considerar esta medida como suficiente para decantarse por un modelo u otro.

3. Como ya se ha explicado anteriormente, en los modelos *logit* y *probit*, se utiliza el método de estimación de máxima verosimilitud, en lugar de mínimos cuadrados ordinarios. Es por ello que no se puede utilizar el coeficiente de determinación clásico R^2 para medir la bondad del ajuste. En su lugar, se utiliza el pseudo R^2 de McFadden:

$$R^2_{McFadden} = 1 - \frac{Ln\ L}{Ln\ L_r}$$

donde $Ln\ L$ es el logaritmo neperiano de la función de verosimilitud del modelo sin restricciones (el modelo con todas las variables explicativas) y $Ln\ L_r$ es el logaritmo neperiano de la función de verosimilitud del modelo restringido (solo incluye el término independiente del modelo).

Es habitual que valores superiores a 0,2 sean considerados como un ajuste robusto y estadísticamente significativo.

Ilustración 1. Pseudo R^2 McFadden modelos *logit* y *probit*, respectivamente

```
> pR2 <- pR2(model)
> pR2 [4]
 McFadden
0.2092844
```

```
> pR2 <- pR2(model)
> pR2 [4]
 McFadden
0.2041427
```

Fuente: Elaboración propia a través de R.

En este caso, el pseudo R^2 de McFadden de ambos modelos son superiores a 0,2 y por tanto, se pueden considerar como un ajuste robusto y estadísticamente significativo. Sin embargo, se observa que el modelo *logit* se ajusta mejor al contar con una medida mayor que la del modelo *probit*.

4. El método Stepwise (paso a paso) es un proceso que permite elegir el mejor modelo en forma secuencial, incluyendo o excluyendo una sola variable predictora en cada paso de acuerdo a ciertos criterios. El proceso secuencial termina cuando se satisface una regla de parada establecida. En este caso, será conseguir el modelo con menor AIC posible.

   Por su parte, el criterio de información de Akaike (AIC) es una medida de la calidad relativa de un modelo estadístico, por la que se valora positivamente la bondad de ajuste y se penaliza la complejidad del mismo, basada en sobreajustes.

   En general, el AIC es:

   $$AIC\ =\ 2k\ -\ 2\ Ln\ L$$

   donde k es el número de parámetros en el modelo estadístico, y L es el máximo valor de la función de verosimilitud para el modelo estimado.

   Por lo tanto, se elegiría el modelo con menor AIC.

Ilustración 2. AIC modelos logit y probit, respectivamente

```
    Null deviance: 1198.54  on 3625  degrees of freedom
Residual deviance:  964.94  on 3613  degrees of freedom
AIC: 990.94

Number of Fisher Scoring iterations: 10
```

```
    Null deviance: 1198.54  on 3625  degrees of freedom
Residual deviance:  966.13  on 3612  degrees of freedom
AIC: 994.13

Number of Fisher Scoring iterations: 10
```

Fuente: Elaboración propia a través de R.

En este caso, se observa que el modelo *logit* cuenta con un menor AIC. Sin embargo, siguen siendo diferencias mínimas.

5. El test de Kolmogorov-Smirnov se utiliza para contrastar si un conjunto de datos se ajusta o no a una distribución normal. Es similar en este caso al test de Shapiro Wilk, pero la principal diferencia con éste radica en el número de muestras. Mientras que el test de Shapiro Wilk se puede utilizar con hasta 50 datos, el test de Kolmogorov Smirnov es recomendable utilizarlo con más de 50 observaciones. Antes de realizar el test de Kolmogorov-Smirnov en R, es necesario conocer cuál es el contraste de hipótesis que se va a realizar.

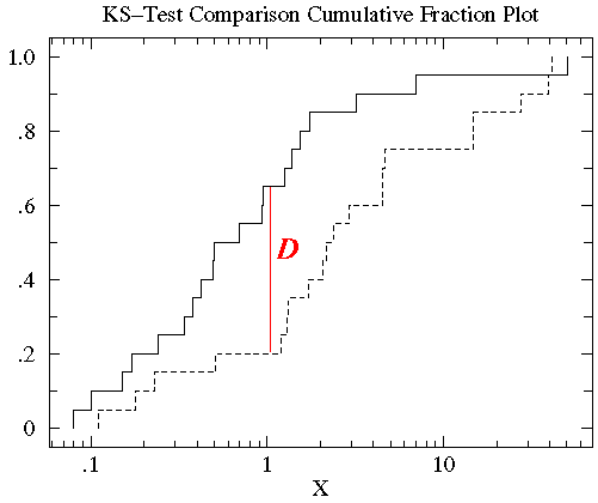
   H0: los datos proceden de una distribución normal
   H1: los datos no proceden de una distribución normal

   El test K-S está basado en la función de distribución empírica (ECDF). Dado N puntos de datos ordenados $Y_1, Y_2, \dots, Y_N$, la ECDF se define como:

$$E_N = \frac{n(i)}{N}$$

   Donde n(i) es el número de puntos menor que Y_i y el Y_i se ordenan del menor al mayor valor. Esta es una función de paso que aumenta en $\frac{1}{N}$ en el valor de cada punto de datos ordenados.
   Cómo alternativa al contraste del test de Kolgomorov-Smirnov, (Nist/Sematech, 2018) grafica la curva de la función de distribución empírica (ECDF) con la función de distribución acumulativa normal. En este gráfico, el test K-S se basa en la máxima distancia entre estas dos curvas, suponiendo un mayor poder predictivo, cuanto mayor es la diferencia entre estas dos curvas (Mondal, 2016).

Gráfico 3. Test K-S. Curva de la función de distribución empírica (ECDF) con la función de distribución acumulativa normal



Fuente: (Kirkman, 2018).

6. El Test de Hosmer y Lemeshow es otro método para estudiar la bondad de ajuste del modelo que consiste en comparar los valores previstos por el modelo con los valores realmente observados. Ambas distribuciones, esperada y observada, se contrastan mediante una prueba de chi cuadrado. La hipótesis nula del test de Hosmer-Lemeshow es que no hay diferencias entre los valores observados y los valores pronosticados. Por lo tanto, valores elevados del p-valor (por encima de los niveles estándar d significación del 1% y 5%) no permitirían rechazar la hipótesis nula de que el modelo está bien ajustado.

   El estadístico C y el estadístico H que aparecen en las tablas siguientes, son dos tipos de ajustes diferentes. El estadístico C se basa en la agrupación de las probabilidades estimadas, por lo que los puntos de corte varían. Sin embargo, el estadístico H basa su ajuste en una agrupación predeterminada con puntos de corte fijos. Los propios autores, Hosmer y Lemeshow recomiendan el uso de ambos estadísticos, si bien no existe consenso sobre la preponderancia de uno en concreto.

Ilustración 3. Test de Hosmer-Lemeshow modelos *logit* y *probit*, respectivamente

```
> hosmerlem(entrenamiento$Fallido, fitted.values(model))
          Hosmer-Lemeshow C statistic Hosmer-Lemeshow H statistic
X-squared                   12.929055                   6.5370859
p.value                      0.114314                   0.5872971

> hosmerlem(entrenamiento$Fallido, fitted.values(model))
          Hosmer-Lemeshow C statistic Hosmer-Lemeshow H statistic
X-squared                   6.3952562                  14.22111907
p.value                     0.6030478                   0.07618099
```

Fuente: Elaboración propia a través de R.

En ambos casos, el p-valor es superior al 5%, por lo que no se rechaza la hipótesis nula y se podría decir que el modelo tanto *logit* como *probit* estaría bien ajustado. Concretamente, para ambos estadísticos y tanto *probit* como *logit* se establece un p-valor superior al 5%.

5 ESTIMACIÓN DEL MODELO DE PROBABILIDAD DE DEFAULT

El siguiente apartado es la fase final del trabajo, tras analizar la literatura relacionada y estudiar y decidir el modelo estadístico utilizado nos disponemos a abordar el principal objetivo de este trabajo. Un modelo de probabilidad que nos ayude a identificar que variables (cuantitativas o cualitativas, económicas, demográficas o sectoriales y macroeconómicas o microeconómicas) son influyentes a la hora de determinar si una empresa será capaz de hacer frente a los compromisos de pago que surgirán en el momento que le concedan financiación y qué peso tiene dicha influencia.

Para una mejor comprensión del apartado y con el objetivo de tener un modelo lo más ajustado a la realidad posible, aplicamos la siguiente metodología:

- Depuración de la base de datos y selección de la muestra.

  De las bases de datos disponibles, que presentaremos y detallaremos más a fondo a continuación, verificaremos que todos los campos que se presentan están calculados y significan lo mismo en cada una de ellas, transformaremos aquellas variables que no se expresan en el formato apropiado para nuestro análisis y obtendremos otras variables que consideramos que podrían ser significativas en nuestro modelo en base a relacionar variables que tenemos con información externa.

  Haremos también una partición en dos de la base de datos (una vez hayamos depurado toda la muestra) con el objetivo de estimar el modelo con una parte y verificar el grado de predicción del modelo estimado con la segunda. Esta partición no se realiza en base a ningún método probabilístico o estadístico.

- Selección de las variables.
  Una vez la depuración sea efectiva, procederemos a seleccionar las variables que incluiremos en la estimación para nuestro modelo final. En esta etapa, realizaremos una estimación de cada uno de los modelos revisados en la literatura previa con nuestra base de datos (y las variables que dispongamos en la misma que los autores utilicen) y con aquellas variables que disponemos en nuestra base de datos y no se hayan considerado en la literatura previa revisada.

  En base a ello, procederemos a crear una tabla con todas las variables significativas resultantes de estas estimaciones, que conformarán el inicio de nuestro modelo de estimación de la probabilidad de default.

- Ajuste del modelo estimado y su validación y bondad.

  Obtenidas las variables significativas en modelos de la literatura previa y de nuestra propia base de datos, crearemos un modelo que procederemos a estimar para calcular los coeficientes βs y su nivel de significación respecto a la variable dependiente.

  Seguidamente, validaremos el modelo y analizaremos su bondad a través de la capacidad de predicción del mismo, el análisis de los residuos, y los valores atípicos relevantes.

- Interpretación del modelo.

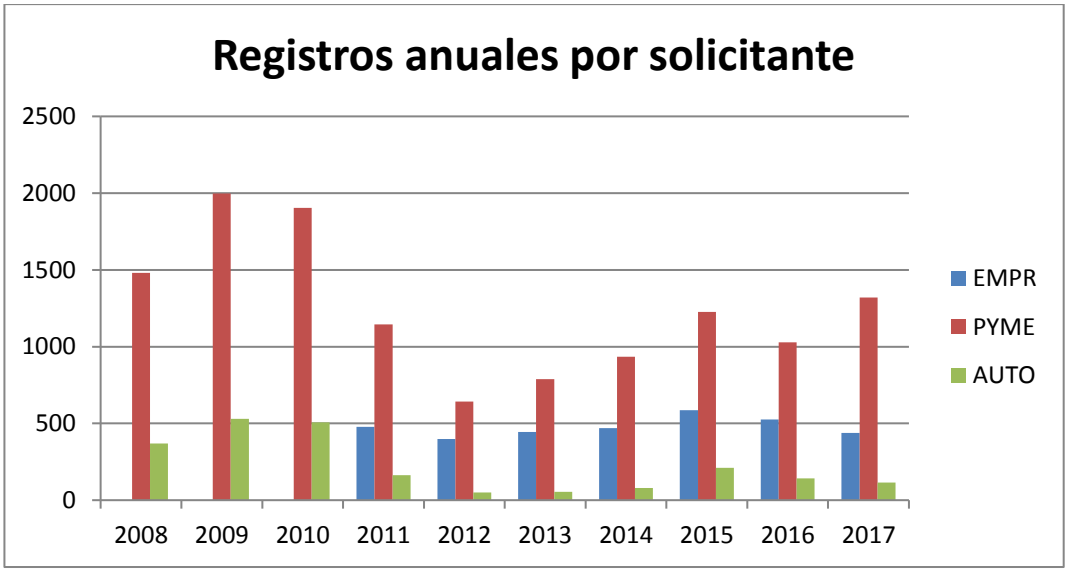
  Una vez ajustemos el modelo, extrapolamos los resultados al resto de la muestra y establecemos las conclusiones obtenidas durante el proceso y sus resultados.

5.1 DEPURACIÓN DE LA BASE DE DATOS Y SELECCIÓN DE LA MUESTRA

La base de datos que disponemos ha sido facilitada por una sociedad de garantía recíproca madrileña.

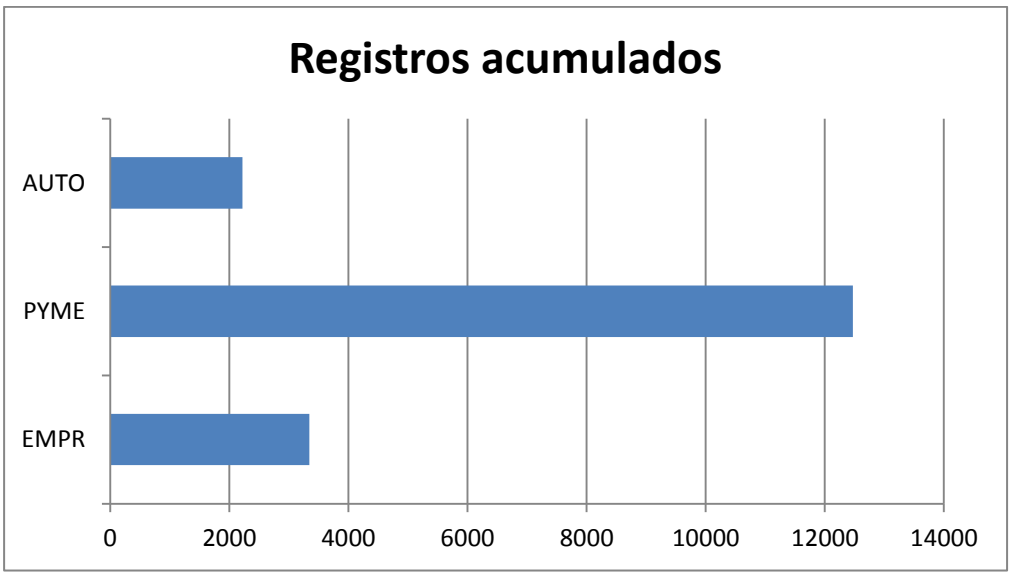
La base de datos contiene inicialmente 18.030 registros de avales concedidos a Empresas, PYMES y autónomos, de acuerdo a la clasificación que hace la propia SGR en el período comprendido entre 2008-2017.

Gráfico 4. Registros anuales por tipo de solicitante disponibles en nuestra base de datos



Fuente: Elaboración propia a partir de la base de datos proporcionada por la SGR.

Gráfico 5. Registros acumulados por tipo de solicitante disponibles en nuestra base de datos



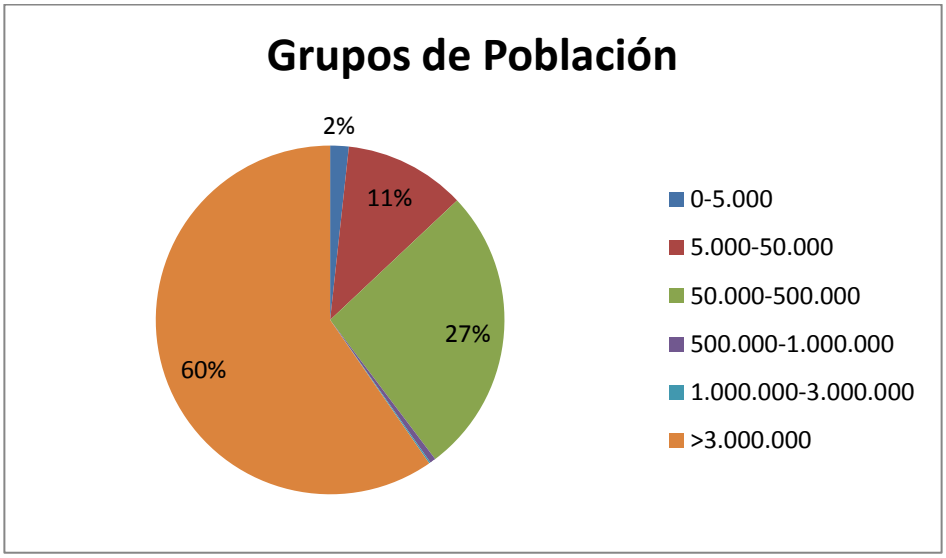
Fuente: Elaboración propia a partir de la base de datos proporcionada por la SGR.

A continuación, los pasos que hemos seguido para llegar a tener la base de datos depurada son:

1. La base de datos contiene la muestra en diferentes archivos según se traten de una solicitud de aval procedente de Empresas, PYMES o Autónomos. El primer paso será verificar que los campos contenidos en estas bases de datos son los mismos en cada una de ellas a fin de poder trabajarlas conjuntamente en un mismo archivo
2. Transformamos el formato de algunas variables que al descargarse se han convertido en valores de texto en vez de valores numéricos, como el tipo de interés, el ROA u otros ratios.
3. Desagregamos también las diferentes fechas disponibles (fecha de firma, fecha de creación de la empresa o fecha de fallido) con el fin de poder obtener variables como la antigüedad de la compañía en el momento de la solicitud del aval, variable significativa en investigaciones como (Westgaard & van de Wijst, 2001).
4. Eliminamos cualquier registro de la base de datos en que el solicitante sea un autónomo. Esta decisión viene motivada porque el objetivo del TFM es la creación de un modelo que determine el default empresas con personalidad jurídica (PYMES y empresas) y no con personalidad física (Autónomos). La aplicación de un modelo de probabilidad de default de estos últimos se asemejaría más a un credit scoring que a un *credit rating*.
5. Añadimos nuevas variables que consideramos que podrían ser significativas en nuestro modelo en base a variables que disponemos actualmente en nuestra base de datos y a información externa. Es el caso

de la variable Población, que presentamos a continuación, y que obtenemos a través del INE, 2018. Esta variable nos dice el número de habitantes de las que dispone el municipio donde tiene sede la PYME solicitante del aval. Consideramos que la variable es relevante por el hecho de que el grado de digitalización de este tipo de empresas no es muy alto y además, la mayoría de ellas pertenecen al sector servicios, lo cual, implica que su ámbito geográfico de actuación queda reducido, principalmente, a su municipio.

Gráfico 6. Distribución del número de registros disponibles en la base de datos según el número de habitantes del municipio del domicilio social



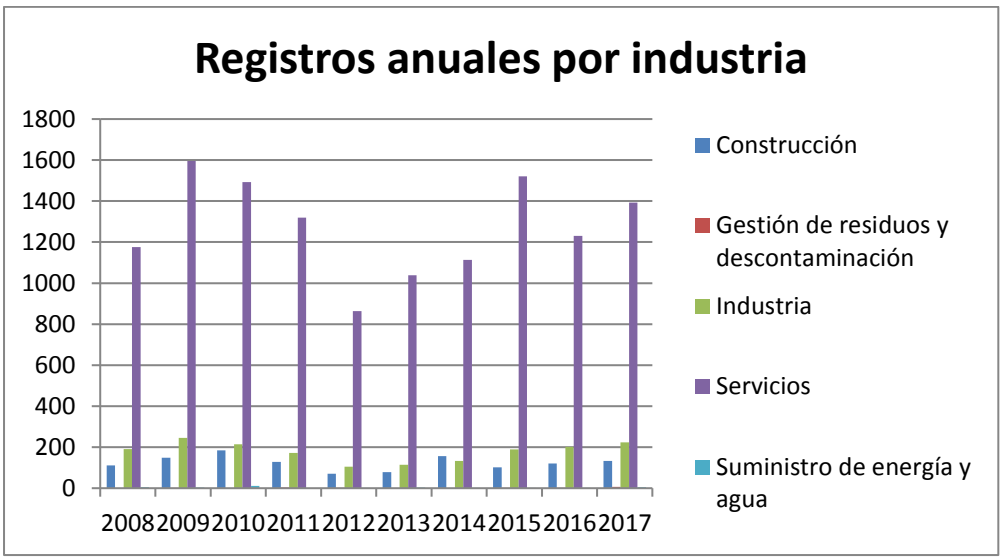
Fuente: (INE, 2018)

En la misma línea, necesitamos añadir algunas variables que se utilizan en modelos de la literatura previa y de las que no disponemos en nuestra base de datos. Un ejemplo es el Producto Bruto Nacional Level Index, necesario para incluir una de las variables que menciona (Ohlson, 1980) en su modelo para el cálculo de la SIZE[2].

Finalmente, creamos una variable en base al CNAE que dividida en 5 grupos atendiendo a la industria a la que pertenece cada solicitante y depuramos la variable Destino de la operación con el objetivo de agruparla en 8 categorías que nos permita trabajar con ella e incluirla en el análisis econométrico.

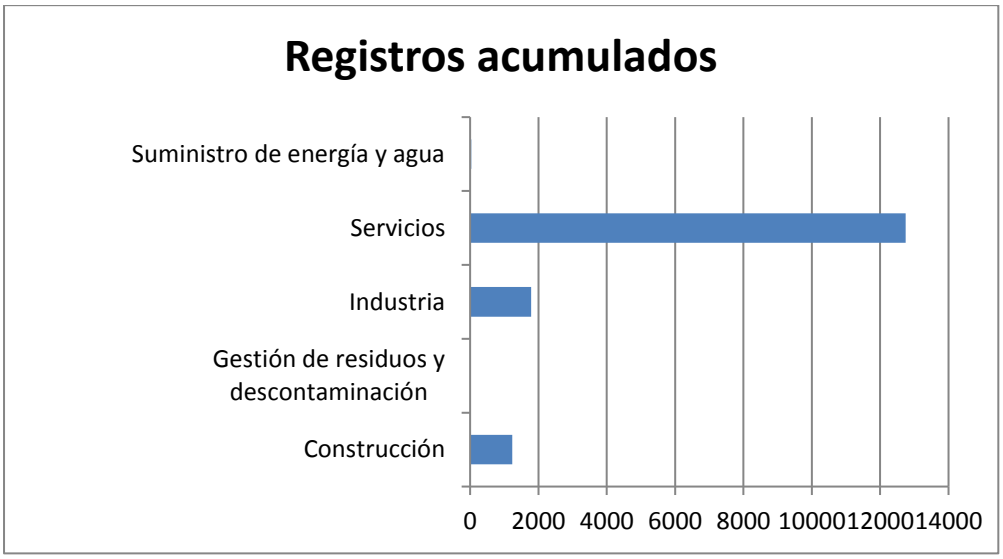
[2] SIZE = log(Activo total / Índice del Producto Bruto Nacional).

Gráfico 6. Registros anuales por tipo de industria del solicitante disponibles en nuestra base de datos



Fuente: Elaboración propia a partir de la base de datos proporcionada por la SGR.

Gráfico 7. Registros acumulados por tipo de industria del solicitante disponibles en nuestra base de datos



Fuente: Elaboración propia a partir de la base de datos proporcionada por la SGR.

6. Disponemos de la posibilidad de analizar la probabilidad de default en seis horizontes diferentes (A 2, 3, 4, 5 y más de 5 años o sin horizonte definido, es decir, todos juntos).

   En nuestro caso, hemos decidido no escoger ningún horizonte temporal y tratar de obtener un modelo lo más generalista posible que nos permita identificar un mayor número de variables significativas para la probabilidad de default.

Tabla 2. Número de registros disponibles en la base de datos que han resultado impagados y su desglose según el año que se ha producido el default

	TOTAL	%	% sobre Muestra total
Fallido 1 año	0	0%	0,00%
Fallido 2 año	102	13%	0,65%
Fallido 3 año	211	28%	1,33%
Fallido 4 año	208	27%	1,32%
Fallido 5 año	131	17%	0,83%
Fallido +5 año	108	14%	0,68%
Fallido TOTAL	760		
Muestra	15.812		

Fuente: Elaboración propia a partir de la base de datos proporcionada por la SGR.

7. Realizados todos los pasos previos anteriores, la muestra final permanece en nuestra base de datos para su estimación y validación es de 5.179 registros

5.2 SELECCIÓN DE LAS VARIABLES

Una vez depurada la base de datos, seleccionaremos las variables que formarán parte de nuestro modelo final a estimar en base a los modelos de la literatura previa probados con nuestra base de datos (siempre y cuando dispongamos de las variables) y a un modelo propio en el que se incluyen aquellas variables de nuestra base de datos que no se han tratado en los modelos de la literatura previa considerados.

Se presenta, a continuación, aquellas variables que han resultado significativas en cada una de las estimaciones realizadas. Estas variables compondrán nuestro modelo de probabilidad de default (que estimaremos, validaremos y ajustaremos en los siguientes puntos):

- **Total Activos**
  Se define activo como todo aquel bien o derecho perteneciente a una empresa y del que se espera un beneficio o que aporte un rendimiento económico a la compañía.
  (Altman & Sabato, 2005)

- **Activos Fijos/ Activos**
  Dentro de los activos, los activos fijos son aquellos de los cuales no se espera obtener un beneficio en el plazo de un año. Por lo tanto, este ratio nos indica el porcentaje de inversiones que esperan realizarse a largo plazo del total de las poseídas por la compañía.
  (Altman & Sabato, 2005) y (Raei, et al., 2016)

- **Beneficio Bruto/ Activos**
  Este ratio nos permite, comparando el beneficio bruto con los activos, conocer si los activos de una compañía son rentables o no.
  (Altman & Sabato, 2005)

- **Pasivo No Corriente/ Activos**
  Ratio que determina el porcentaje de activos financiado con financiación a largo plazo.
  (Altman & Sabato, 2005)

- **Ebit/Activos**
  El rendimiento de los activos totales, conocido como ROTA es un ratio que mide las ganancias de una compañía antes de intereses e impuestos (EBIT) en relación con sus activos netos totales. Este ratio sirve para medir la eficacia con la que la empresa hace uso de sus activos para generar ganancias antes del pago de otras obligaciones.
  (Altman & Sabato, 2005)

- **Caja/ Activos**
  La relación de efectivo a activos nos dice qué porción del activo está constituida por los activos más líquidos de la empresa: efectivo y equivalentes de efectivo y valores negociables.
  (Altman, et al., 2010)

- **Activos Corrientes/ Pasivos Corrientes**
  Índice de liquidez que sirve para medir la capacidad de una empresa de hacer frente a sus obligaciones corto plazo.
  (Altman, et al., 2010)

- **Inventario/ Capital Circulante**
  Este ratio se define como una forma de qué parte de los inventarios de una empresa se financia con su efectivo disponible. Esto es esencial para las empresas que mantienen inventario y sobreviven con suministros de efectivo. En general, cuanto menor es la relación, mayor es la liquidez de una empresa.
  (Altman, et al., 2010)

- **"Quick Assets" / Activos Corrientes**
  Los activos rápidos son activos que se pueden convertir en efectivo rápidamente. Por lo general, incluyen efectivo, cuentas por cobrar,

valores negociables e inventario. Por otro lado, se e entiende por activos corrientes aquellos activos que son susceptibles de convertirse en dinero en efectivo en un periodo inferior a un año, por lo que este ratio mide la capacidad de convertir los activos más líquidos de la empresa en liquidez (Altman, et al., 2010)

- **INONE**
  Variable binomial que toma el valor 1 en caso de que la empresa solicitante de la operación ha obtenido un beneficio positivo en su último ejercicio o 0 en caso contrario.
  Proxy de la variable INTWO de (Ohlson, 1980) que tomaba el valor 1 en caso de que los dos últimos ejercicios de la empresa arrojasen resultados positivos o 0 en caso contrario.

- **OENEG**
  Variable binomial que es uno si el pasivo total excede los activos totales, cero de lo contrario.
  (Ohlson, 1980)

- **SIZE**
  Log (activos totales / índice de precio de GNP). El índice supone un valor base de 100 para 2003. El año del índice es a partir del año antes del año de la fecha del balance
  (Ohlson, 1980)

- **Antigüedad**
  Años de actividad de la empresa en el momento de solicitud del aval
  (Westgaard & Van de Wijst, 2001)

- **Empleos Fijos**
  Empleados caracterizados por carecer límite de tiempo en la prestación de los servicios. Este tipo de contrato refleja estabilidad tanto a nivel del contratado como de la empresa.
  (Westgaard & Van de Wijst, 2001)

- **Empleos temporales**
  Todo aquel contrato que se haga a través de una empresa de trabajo temporal ETT. Son siempre de duración determinada.
  (Westgaard & Van de Wijst, 2001)

- **Empleos Indirectos**
  Se define empleo indirecto como aquellos puestos de trabajo generados por la actividad de la empresa pero que no son propiamente de la empresa.
  (Westgaard & Van de Wijst, 2001)

- **Financov**
  Está relacionada con la cobertura financiera, y se mide como el resultado neto antes de los costes financieros dividido entre los costes financieros. (Westgaard & Van de Wijst, 2001)

- **Industria**
  Sector de actividad a la que se dedica la empresa analizada
  (Westgaard & Van de Wijst, 2001)

- **Población**
  Localidad o municipio donde se encuentra la empresa.
  (Westgaard & Van de Wijst, 2001)

- **Avales financieros**
  Número de avales financieros que dispone la empresa. Se define aval financiero como una garantía prestada por un banco por la que se compromete a responder en el cumplimiento de una obligación ante un tercero.

- **Avales técnicos**
  Número de avales técnicos de la empresa. Estos avales garantizan el incumplimiento de compromisos no económicos asumidos por su empresa.

- **Destino de la operación**
  Esta variable se refiere al uso que tendrán los fondos adquiridos por la empresa en la operación.

- **Duración**
  Número de períodos durante los cuales el prestatario asume la obligación de pagar las cuotas.

- **Garantías hipotecarias**
  La garantía hipotecaria es el derecho que se concede sobre un inmueble a una persona o entidad con la que se contrae una deuda o compromiso, para que en caso de dicha deuda no sea satisfecha o el compromiso incumplido, tenga la posibilidad de convertirse en propietario y vender dicho inmueble para recuperar el dinero prestado.

- **Garantías personales**
  Aquellas garantías que asumen directamente personas físicas teniendo como respaldo su patrimonio.

- **Garantías pignoración**
  Número de préstamos con garantía pignorada, es decir, aquel en el que el cliente presenta como garantía una prenda. Estos suelen ser activos financieros o artículos de valor como joyas.

- **Margen sobre ventas**
  Ratio calculado a partir del margen comercial entre el precio de venta. El margen es la diferencia entre el precio de venta, sin incluir los impuestos y el coste del producto, sin impuestos también.

- **Tipo de interés de la financiación.**
  Tipo de interés exigido por la institución financiera como cargo para aprobar la concesión de un crédito.

5.3 AJUSTE DEL MODELO ESTIMADO Y SU VALIDACIÓN Y BONDAD

Trataremos en este apartado de obtener un modelo de probabilidad de default a través de la regresión logística (ya explicada anteriormente). Finalmente, como hemos remarcado en el punto 7 del apartado 4.1 Depuración de la base de datos y selección de la muestra, la base de datos ya depurada contiene 5.179 registros solicitados a la SGR entre el 2008 y 2017.

Realizamos a continuación la primera estimación del modelo:

Ilustración 4. Primera estimación del modelo con todas las variables disponibles

```
> summary(model) #model o sec.model

Call:
glm(formula = Fallido ~ ., family = binomial(link = "logit"),
    data = entrenamiento)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.3433  -0.3052  -0.1557  -0.0300   4.0822

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -9.876e-01  1.153e+00  -0.856 0.391850
AT          -3.708e-08  2.684e-08  -1.381 0.167159
AF_AT       -1.274e+00  3.951e-01  -3.225 0.001259 **
MB_AT       -2.929e-01  2.233e-01  -1.312 0.189601
Dlp_AT       1.537e+00  4.876e-01   3.152 0.001623 **
EBIT_AT     -5.581e-01  6.807e-01  -0.820 0.412253
Caja_AT     -5.323e+00  1.853e+00  -2.873 0.004067 **
D.Fian      -1.595e+00  4.725e-01  -3.375 0.000738 ***
D.Circ      -5.297e-01  2.150e-01  -2.464 0.013747 *
OENEG        2.499e-01  3.056e-01   0.818 0.413633
Poblac      -6.206e-08  6.298e-08  -0.985 0.324430
AC_PC        3.703e-02  5.589e-02   0.663 0.507621
Inv_Circ    -1.073e-08  2.392e-08  -0.449 0.653746
QA_AC       -6.837e-01  3.752e-01  -1.822 0.068457 .
IN1         -9.660e-01  2.396e-01  -4.031 5.55e-05 ***
Tama         6.152e-01  2.440e-01   2.522 0.011684 *
Antg        -1.447e-02  1.000e-02  -1.447 0.147963
E.Fij       -2.339e-04  1.837e-03  -0.127 0.898667
E.Ind        8.251e-03  1.227e-02   0.672 0.501340
Fincov      -1.618e-02  7.948e-03  -2.035 0.041835 *
I.Indus      2.136e-01  2.466e-01   0.866 0.386392
I.Energ     -1.365e+01  4.378e+02  -0.031 0.975119
Liq         -1.764e-02  1.154e-02  -1.529 0.126344
A.Fin       -1.329e-01  4.915e-02  -2.703 0.006865 **
A.Tec       -1.073e-01  4.056e-02  -2.646 0.008152 **
Dur         -1.937e-03  1.060e-03  -1.827 0.067688 .
G.Hip       -3.943e-01  1.960e-01  -2.012 0.044270 *
G.Per       -9.468e-02  6.105e-02  -1.551 0.120943
G.Pig        1.278e-01  4.209e-01   0.304 0.761406
MB_Ventas   -1.187e+00  4.901e-01  -2.421 0.015466 *
Int         -7.433e+00  1.815e+00  -4.095 4.23e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1198.5  on 3625  degrees of freedom
Residual deviance:  935.7  on 3595  degrees of freedom
AIC: 997.7

Number of Fisher Scoring iterations: 14
```

Fuente: Elaboración propia a través de R.

El modelo estimado resultante es de 15 variables significativas sobre 30 introducidas. Es un ajuste que creemos que se puede mejorar, tratando de simplificar el modelo con la premisa básica de no perder información.

Por ello, a continuación, realizaremos varios contrastes para comprobar la significación individual y conjunta de las variables del modelo, la auto-correlación existente entre ellas y diversos estadísticos sobre la bondad del ajuste (test de Hosmer-Lemeshov o Kolgomorov-Smirnov).

5.3.1 Contraste de significación

Una vez hemos estimado el modelo, es necesario verificar que los parámetros βn son significativamente diferentes de 0, tanto de forma individual como de forma conjunta.

Para el contraste de significación individual, utilizaremos diversos métodos que nos permitan confirmar la hipótesis de que βn ≠ 0: el contraste de Wald, los intervalos de confianza y el contraste condicional de razón de verosimilitud. Para el contraste de significación conjunta utilizaremos el contraste condicional de razón de verosimilitud en el que confrontaremos a nuestro modelo estimado respectivo a un modelo constante (formado por nuestra variable dependiente y una constante).

5.3.1.1 Contraste de significación individual

5.3.1.1.1 Wald

Test de Wald trata de ver la coherencia de afirmar un valor concreto de un parámetro de un modelo probabilístico una vez tenemos ya un modelo previamente seleccionado y ajustado (Pérez, 2018). Esto es, siempre que sea cierta la hipótesis nula, valoramos si lo que vemos es o no muy distante de lo esperado.

El contraste planteado es el siguiente (Hauck & Donner, 1977):

$$H_0 : \beta_n = 0$$

$$H_1 : \beta_n \neq 0$$

A través del siguiente estadístico de contraste:

$$W_r = \frac{\beta_n}{SE(\beta_r)}$$

Y su interpretación, suponiendo un nivel de significación por defecto del 5% (0,05):

- Si el p-valor es menor que 0,05, se rechaza la hipótesis nula que afirma que ese coeficiente es 0 y que por lo tanto el modelo es útil para representar una relación determinada.
- Si el p-valor es mayor que 0,05, significa que el valor del coeficiente podría ser igual a 0 y, por tanto, esa variable no influiría a la hora de determinar la variable dependiente del modelo de regresión.

Una de las formas de aplicar el test de Wald, es utilizar el "z value" (valor en la tercera columna de la salida de un modelo). En este caso, la variable independiente influirá sobre la variable dependiente si el valor absoluto de su "z value" es superior a 1,96.

Ilustración 5. "Z value" de cada una de las variables introducidas en la primera estimación del modelo

```
> summary(model)$coefficients[,3]
(Intercept)          AT       AF_AT       MB_AT      Dlp_AT     EBIT_AT     Caja_AT      D.Fian      D.Circ
 -0.8562677  -1.3813895  -3.2251370  -1.3117614   3.1517980  -0.8199355  -2.8728834  -3.3748936  -2.4637955
      OENEG      Poblac       AC_PC    Inv_Circ       QA_AC         IN1       Tama        Antg       E.Fij
  0.8175179  -0.9853952   0.6625464  -0.4485644  -1.8219888  -4.0310590   2.5215530  -1.4467642  -0.1273450
      E.Ind      Fincov     I.Indus     I.Energ         Liq       A.Fin       A.Tec         Dur       G.Hip
  0.6723828  -2.0351581   0.8661795  -0.0311882  -1.5286801  -2.7033035  -2.6457147  -1.8270795  -2.0115258
      G.Per      G.Pig   MB_Ventas         Int
 -1.5508267   0.3036350  -2.4212744  -4.0947078
```

Fuente: Elaboración propia a través de R.

Como vemos en la tabla anterior, solamente las variables que no son significativas del modelo estimado contienen un "z value" inferior a 1,96 en valor absoluto (el ya mencionado punto crítico e_(α/2)=1,96), que agrupamos a continuación:

Ilustración 6. "Z value" de las variables introducidas en la primera estimación que no cumplen con la hipótesis de significación del contraste de Wald

```
> no.sign.var.zvalue
          AT     MB_AT    EBIT_AT     OENEG     Poblac     AC_PC   Inv_Circ     QA_AC      Antg     E.Fij
[1,] -1.38139 -1.311761 -0.8199355 0.8175179 -0.9853952 0.6625464 -0.4485644 -1.821989 -1.446764 -0.127345
       E.Ind     I.Ind    I.Energ       Liq       Dur     G.Per     G.Pig
[1,] 0.6723828 0.8661795 -0.0311882 -1.52868 -1.827079 -1.550827 0.303635
```

Fuente: Elaboración propia a través de R.

5.3.1.1.2 Intervalo de confianza

Una forma alternativa al test de Wald, es el uso de los intervalos de confianza.

Siguiendo a (Roche, 2013), debido a la interpretación de los parámetros en los modelos de regresión logística, se suelen calcular los intervalos de confianza para los exponenciales de los parámetros, que se corresponden con los cocientes de ventajas. Este contraste se define como:

$$H_0: e^{\beta_r} = 1$$

$$H_1: e^{\beta_r} \neq 1$$

Para este método, sabemos que los parámetros βn son significativamente distintos de 1 si no contienen dicho valor en su intervalo de confianza (que en nuestro caso está elaborado al 95%, es decir, a nivel de significación de α = 5%).

Ilustración 7. Intervalos de confianza de los exponenciales de los parámetros de las variables introducidas en la primera estimación

```
> exp(confint.default(model))
                    2.5 %      97.5 %
(Intercept) 3.885098e-02  3.57125540
AT          9.999999e-01  1.00000002
AF_AT       1.288786e-01  0.60657245
MB_AT       4.816027e-01  1.15575684
Dlp_AT      1.788007e+00 12.08951890
EBIT_AT     1.507207e-01  2.17284559
Caja_AT     1.291208e-04  0.18423477
D.Fian      8.041746e-02  0.51247559
D.Circ      3.863015e-01  0.89733407
OENEG       7.052736e-01  2.33705311
Poblac      9.999998e-01  1.00000006
AC_PC       9.300495e-01  1.15786960
Inv_Circ    9.999999e-01  1.00000004
QA_AC       2.419357e-01  1.05313535
IN1         2.379546e-01  0.60877129
Tama        1.146858e+00  2.98462484
Antg        9.664964e-01  1.00514676
E.Fij       9.961730e-01  1.00337217
E.Ind       9.843234e-01  1.03283066
Fincov      9.687452e-01  0.99940253
I.Indus     7.636248e-01  2.00732184
I.Energ     0.000000e+00         Inf
Liq         9.605377e-01  1.00498981
A.Fin       7.951610e-01  0.96412251
A.Tec       8.296228e-01  0.97257241
Dur         9.959939e-01  1.00014086
G.Hip       4.591356e-01  0.98994481
G.Per       8.070770e-01  1.02529268
G.Pig       4.980219e-01  2.59269357
MB_Ventas   1.167817e-01  0.79763058
Int         1.686531e-05  0.02075599
```

Fuente: Elaboración propia a través de R.

En línea con el contraste realizado en el test de Wald, el intervalo de confianza para los exponenciales de los parámetros arroja resultados similares que determinan que las variables de la ilustración 5. 6. anterior no son significativas de individualmente en el modelo, es decir, que cada una de ellas no tiene influencia en la variable dependiente.

5.3.1.2 Contraste de significación conjunta

El contraste condicional de razón de verosimilitudes nos sirve para contrastar si un subconjunto de parámetros es iguales a 0, frente a que alguno de ellos no lo sea.

El contraste de significación conjunto será válido para nuestros intereses en caso de que rechacemos la hipótesis nula.

Ilustración 8. Contraste de significación conjunto

```
> anova(modelo.cte, model, test = "Chisq")
Analysis of Deviance Table

Model 1: Fallido ~ 1
Model 2: Fallido ~ AT + AF_AT + MB_AT + Dlp_AT + EBIT_AT + Caja_AT + D.Fian +
    D.Circ + OENEG + Poblac + AC_PC + Inv_Circ + QA_AC + IN1 +
    Tama + Antg + E.Fij + E.Ind + Fincov + I.Indus + I.Energ +
    Liq + A.Fin + A.Tec + Dur + G.Hip + G.Per + G.Pig + MB_Ventas +
    Int
  Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
1      3625     1198.5
2      3595      935.7 30   262.84 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Fuente: Elaboración propia a través de R.

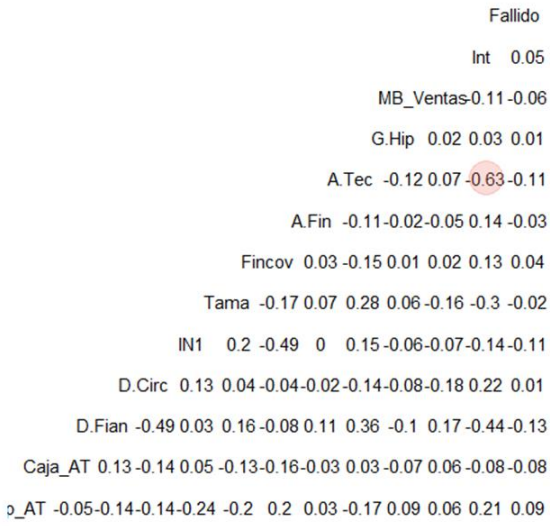
Como vemos en la tabla anterior, se rechazaría la hipótesis nula de que los parámetros del siguiente modelo son 0 a un nivel de significación de $\alpha = 0{,}1\%$:

5.3.2 Ausencia de autocorrelación

La multicolinealidad es un problema que surge cuando las variables explicativas del modelo están altamente correlacionadas entre sí o cuando, directamente, alguna variable explicativa es combinación lineal exacta de otras.

Cuando existe colinealidad, no se puede estimar de forma precisa el efecto de cada una de las variables independientes en la variable dependiente del modelo porque se reduce la precisión de los coeficientes estimados, es decir, se incrementa su varianza (Roche, 2013).

Ilustración 9. Matriz de autocorrelación



Fuente: Elaboración propia a través de R.

Es, por tanto, una condición indispensable que deben cumplir las variables explicativas del modelo.

Vista la matriz de autocorrelación, no observamos ningún indicio de colinealidad entre variables a excepción de un posible indicio de colinealidad entre el tipo de interés y el número de avales técnicos. Para confirmar la hipótesis de inexistencia de colinealidad, calculamos a continuación el Factor de Inflación de la Varianza (VIF).

El VIF proporciona un índice que mide hasta qué punto la varianza (el cuadrado de la desviación estándar estimada) de un coeficiente de regresión estimado se incrementa a causa de la colinealidad.

Para su interpretación acudimos a los siguientes límites (PennState, 2018):

- VIF = 1 → Ausencia de colinealidad.
- 1 < VIF < 4 → Posible indicio de colinealidad.
- VIF > 10 → Serios indicios de colinealidad

Ilustración 10. Factor de Influencia de la Varianza

```
> vif(model)
      AT    AF_AT    MB_AT   Dlp_AT  EBIT_AT  Caja_AT   D.Fian   D.Circ    OENEG
2.574542 1.382943 2.111655 1.576048 2.407345 1.210092 1.238716 1.277495 1.683011
  Poblac    AC_PC Inv_Circ    QA_AC      IN1     Tama     Antg    E.Fij    E.Ind
1.112746 1.276292 1.203868 1.386553 1.739476 3.345803 1.507328 1.302128 1.105038
  Fincov I.Indus I.Energ      Liq    A.Fin    A.Tec      Dur    G.Hip    G.Per
1.751615 1.273157 1.000002 1.242907 1.154688 1.264958 1.270282 1.133140 1.146173
   G.Pig MB_Ventas      Int
1.048801 2.062932 1.140629
```

Fuente: Elaboración propia a través de R.

5.4 STEPWISE

Una vez comprobamos que el modelo se ajusta correctamente a los contrastes planteados, le aplicamos a nuestro modelo el proceso de Stepwise reciente descrito con

el objetivo de simplificar el modelo al máximo sin perder información. Esta restricción la manejaremos a través del criterio de información de Akaike (AIC) eligiendo el modelo que tenga un menor valor de dicho criterio AIC:

Ilustración 11. Resultado del proceso de Stepwise aplicado a la primera estimación tras la validación del modelo

```
Step:  AIC=990.94
Fallido ~ A.Tec + Caja_AT + D.Fian + IN1 + Int + A.Fin + MB_Ventas +
    Dlp_AT + AF_AT + G.Hip + Tama + D.Circ

            Df Deviance     AIC
<none>           964.94  990.94
+ Fincov     1   963.05  991.05
- MB_Ventas  1   969.32  993.32
- D.Circ     1   969.65  993.65
- Tama       1   971.22  995.22
- G.Hip      1   972.29  996.29
- A.Fin      1   974.72  998.72
- AF_AT      1   975.90  999.90
- Dlp_AT     1   978.48 1002.48
- Int        1   980.04 1004.04
- Caja_AT    1   986.34 1010.34
- D.Fian     1   989.38 1013.38
- IN1        1   990.74 1014.74
- A.Tec      1  1024.58 1048.58
```

Fuente: Elaboración propia a través de R.

El modelo resultante tras el proceso Stepwise, determina que debemos tener en cuenta las siguientes variables independientes:

- Avales técnicos.
- Caja / Activo Total.
- Destino: Fianza.
- Beneficio neto positivo en el año anterior.
- Tipo de interés.
- Avales financieros.
- Margen bruto / Ventas.
- Deuda a largo plazo / Activos totales.
- Activos fijos / Activos totales.
- Garantías hipotecarias.
- Tamaño.
- Destino: Financiación de circulante.

La estimación de este modelo que consideramos el modelo final es la siguiente:

Ilustración 12. Modelo estimado resultante del proceso Stepwise

```
Call:
glm(formula = Fallido ~ A.Tec + Caja_AT + D.Fian + IN1 + Int +
    A.Fin + MB_Ventas + Dlp_AT + AF_AT + G.Hip + Tama + D.Circ,
    family = binomial(link = "logit"), data = snd.model)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.1217  -0.3139  -0.1663  -0.0438   3.9902

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.33879    0.69885  -1.916 0.055401 .
A.Tec       -0.12876    0.04131  -3.117 0.001829 **
Caja_AT     -6.67876    1.77719  -3.758 0.000171 ***
D.Fian      -1.89520    0.46055  -4.115 3.87e-05 ***
IN1         -0.97603    0.18986  -5.141 2.74e-07 ***
Int         -6.78070    1.73243  -3.914 9.08e-05 ***
A.Fin       -0.12298    0.04624  -2.660 0.007820 **
MB_Ventas   -0.73285    0.35237  -2.080 0.037545 *
Dlp_AT       1.64656    0.44359   3.712 0.000206 ***
AF_AT       -1.24728    0.38143  -3.270 0.001075 **
G.Hip       -0.44700    0.19611  -2.279 0.022643 *
Tama         0.34966    0.13944   2.508 0.012154 *
D.Circ      -0.44486    0.20698  -2.149 0.031612 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1198.54  on 3625  degrees of freedom
Residual deviance:  964.94  on 3613  degrees of freedom
AIC: 990.94

Number of Fisher Scoring iterations: 10
```

Fuente: Elaboración propia a través de R.

5.5 VALIDACIÓN Y BONDAD DEL MODELO

Ajustado el modelo, el siguiente y último paso anterior a la interpretación de los datos obtenidos, es la bondad del ajuste y el análisis de los residuos del modelo estimado. Estudiaremos además los valores atípicos.

5.5.1 Análisis de residuos

El análisis de los residuos es fundamental para evaluar la adecuación del modelo y detectar valores anómalos e influyentes.

Aunque existen varios tipos de errores que se pueden utilizar según el objetivo perseguido, utilizaremos el método de Pearson y los residuos de la devianza.

5.5.1.1 Residuos de Pearson

Serían los q-ésimos componentes del estadístico X2 de Pearson de bondad del ajuste global, que veremos en la siguiente sección. Su expresión es:

$$e_{p_q} = \frac{y_q - n_q \hat{p}_q}{\sqrt{n_q \hat{p}_q \, (1 - \hat{p}_q)}}$$

Una vez obtenidos los residuos, se realiza el siguiente contraste:

$$H_0\colon\ e_{p_q} = 0$$

$$H_1\colon\ e_{p_q} \neq 0$$

Bajo la hipótesis nula, e_{Pq}, tiene distribución asintótica normal con media 0 y varianza estimada menor que 1. No obstante, se suele considerar la distribución normal estándar, y se consideran significativos aquellos residuos cuyo valor absoluto sea mayor que 2.

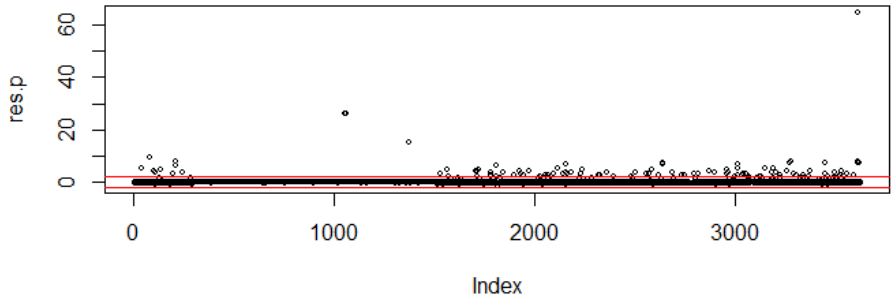
Si analizamos los residuos de Pearson aplicados a nuestro modelo y consideramos solo aquellos cuyo valor absoluto es superior a 2, obtenemos que solamente 113 cumplen con esta restricción, lo que representa que tan solo un 3,11% de los errores corresponden con valores significativos.

Ilustración 13. Residuos de Pearson del modelo estimado restringidos a un valor absoluto mayor a 2

```
> table(res.p.std.sig)
res.p.std.sig
FALSE  TRUE
 3513   113
```

Fuente: Elaboración propia a través de R.

Gráfico 8. Residuos de Pearson del modelo estimado



Fuente: Elaboración propia a través de R.

5.5.1.2 Residuos de la devianza

Considerando que la devianza del modelo es:

$$G^2 = \sum_{q=1}^{Q} d_q^2 = \sum_{q=1}^{Q} 2\left[y_q \log\frac{y_q}{\hat{\mu}_q} + (n_q - y_q)\log\frac{n_q - y_q}{n_q - \hat{\mu}_q}\right]$$

Los residuos de la devianza son los elementos d_q, tomando como signo el signo de $y_q - \mu_q$con lo que serían:

$$e_{Dq} = d_q = signo\left(y_q - \hat{\mu}_q\right)\left(2\left[y_q \log\frac{y_q}{\hat{\mu}_q} + (n_q - y_q)\log\frac{n_q - y_q}{n_q - \hat{\mu}_q}\right]\right)^{\frac{1}{2}}$$

Estos residuos, al igual que los de Pearson, bajo la hipótesis nula que son igual a 0, tienen distribución asintótica de media 0 y varianza estimada menor que 1.

Ilustración 14. Residuos de la devianza del modelo estimado restringidos a un valor absoluto mayor a 2

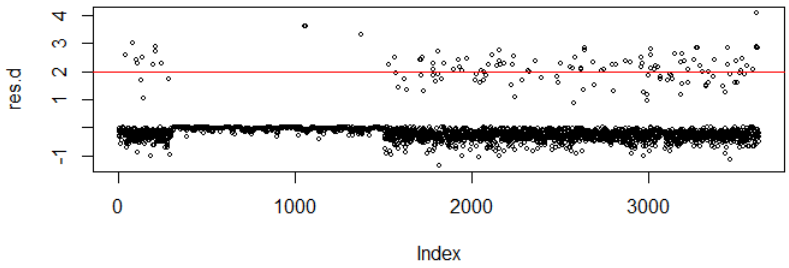
```
> res.dev.std <- rstandard(model, type = "deviance")
> table(abs(res.dev.std) > 2)

FALSE  TRUE
 3535    91
```

Fuente: Elaboración propia a través de R.

Si analizamos los residuos de la devianza aplicados a nuestro modelo y consideramos solo aquellos cuyo valor absoluto es superior a 2, obtenemos que solamente 91 cumplen con esta restricción, lo que representa que tan solo un 2,50% de los errores corresponden con valores significativos.

Gráfico 9. Residuos de la Devianza del modelo estimado



Fuente: Elaboración propia a través de R.

5.5.2 Valores influyentes

Independientemente de si el modelo se ha aceptado o no, es importante identificar si existe algún valor que está condicionando el modelo, es por ello, por lo que utilizamos los Hat Values y la Distancia de Cook como indicador para la identificación de estos posibles valores influyentes.

5.5.2.1 Distancia de Cook

Este indicador tiene como objetivo medir aquellas observaciones que suelen tener un peso específico mayor en la estimación de los coeficientes de regresión, esto es, mide el efecto provocado al eliminar una observación determinada.

La distancia de Cook (D_i) de una observación i ($\forall_i = 1, \dots, n$) está definida como la suma de todos los cambios en la regresión del modelo cuando la observación i está eliminada:

$$D_i = \frac{\sum_{j=1}^{n} \left(\hat{y}_j - \hat{y}_{j(i)}\right)^2}{ps^2}$$

Donde $\hat{y}_{j(i)}$ es el valor de respuesta ajustado obtenido al excluir i, y $s^2 = (n-p)^{-1} * e^T * e$ es el error cuadrático medio de la regresión del modelo. De manera equivalente, se puede expresar utilizando el apalancamiento (MathWorks, 2018):

$$D_i = \frac{e_i^2}{s^2 p}\left[\frac{h_i}{(1-h_i)^2}\right]$$

Siguiendo con nuestro modelo, y tras calcular cada distancia de Cook para cada uno de los valores respuesta obtenidos, comprobamos que no tenemos ningún valor atípico que supere una distancia de Cook de 1, valor que la literatura fija como límite de referencia para la realización de otros análisis.

Ilustración 15. Distancia de Cook del modelo estimado restringido a un valor absoluto mayor a 1

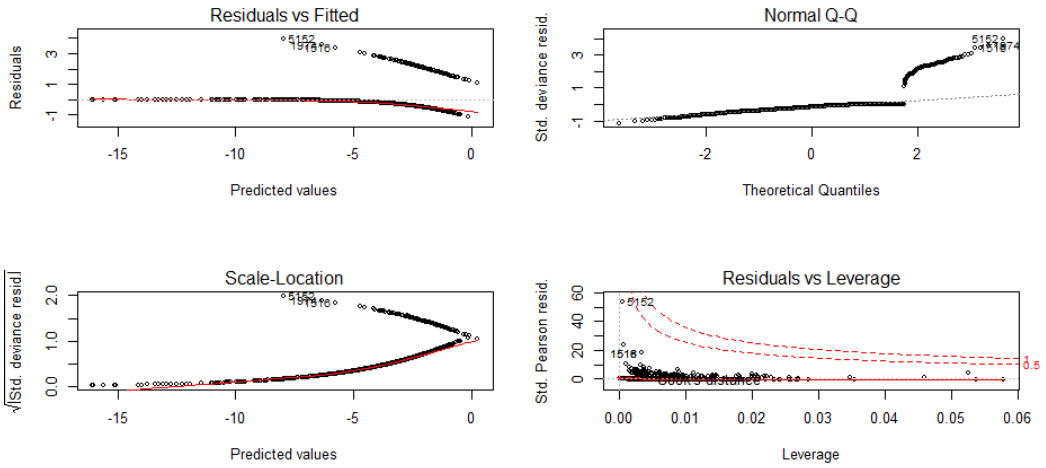
```
> table(distancias.cook > 1)

FALSE
 3626
```

Fuente: Elaboración propia a través de R.

En relación a los gráficos de a continuación, presentamos una serie de gráficos que muestran la relación entre los valores atípicos y sus efectos en la variable dependiente.

Gráfico 10. Varios gráficos que presentan los residuos de la devianza frente a diferentes valores del modelo estimado



Fuente: Elaboración propia a través de R.

En la esquina superior izquierda presentamos los residuos de la devianza frente al predictor lineal mientras que en la esquina inferior se muestra la raíz cuadrada de los residuos de la devianza estandarizados. En la esquina superior derecha, presentamos los residuos de la deviación estándar y en su esquina inferior el gráfico de más interés en nuestra opinión ya que es donde mejor se reflejan los resultados de comparar los residuos estandarizados con los hat values y la distancia de Cook.

Vemos en este "4º gráfico" como se comparan los residuos estandarizados de Pearson con los hat values, mostrando además líneas de contorno para las distancias de Cook. Como comprobamos con los hat values y la distancia de Cook, nuestro modelo no presenta valores influyentes que pudiese afectar a la variable dependiente.

5.5.3 Test de Hosmer-Lemeshov

El test de Hosmer-Lemeshov es un contraste apropiado para datos no agrupados como nuestra base de datos dónde se propone crear grupos de la variable respuesta en base a las probabilidades estimadas por el modelo, y comparar las frecuencias de éxito

observadas con las estimadas, mediante el estadístico usual X2 de Pearson (Roche, 2013).

Si estimamos los dos estadísticos (C y H) del modelo, obtenemos los siguientes salidos:

Ilustración 16. Test de Hosmer-Lemeshow. Estadísticos C y H

```
> hosmerlem(entrenamiento$Fallido, fitted.values(model))
          Hosmer-Lemeshow C statistic Hosmer-Lemeshow H statistic
X-squared                   12.929055                   6.5370859
p.value                      0.114314                   0.5872971
```

Fuente: Elaboración propia a través de R.

Y aceptaríamos la hipótesis nula de que el modelo se ajusta globalmente a los datos.

5.5.4 Poder de clasificación del modelo

Para calcular el poder de clasificación del modelo, procedemos a realizar la estimación para cada uno de los registros de la probabilidad de default, para posteriormente poder elaborar una tabla de clasificación del modelo con el siguiente formato:

Tabla 3. Tabla teórica de clasificación de valores resultantes

	Clasificación: Éxito	**Clasificación: Fracaso**
Éxito	Verdaderos positivos (VP)	Falsos Negativos (FN)
Fracaso	Falsos positivos (FP)	Verdaderos negativos (VN)

Fuente: (Mondal, 2016)

Estimamos un punto de corte a partir del cual transformaremos la probabilidad estimada por el modelo en variables 0 para el caso de pago y 1 para el caso de impagos. En nuestro trabajo, hemos establecido como punto de corte 0,5. Entonces:

- Si PD>0,5 el registro se clasifica como 1
- Si PD<0,5 el registro se clasifica como 0

A raíz de esto, podemos determinar dos tipos distintos de errores:

- **Tipo 1**: Aquel en el que se clasifica una pyme o empresa como que no va a entrar en default, cuando sí que produce default.
- **Tipo 2**: Aquel en el que se clasifica una pyme o empresa como que va a entrar en default cuando en realidad no es así.

Por tanto, una vez estimado la probabilidad de default de cada uno de los registros y comparados con su clasificación real, podemos completar la tabla definida anteriormente y estimar el poder de predicción del modelo.

Tabla 4. Tabla de clasificación de valores resultantes

	Clasificación: Éxito	**Clasificación: Fracaso**
Éxito	3.484	0
Fracaso	141	1

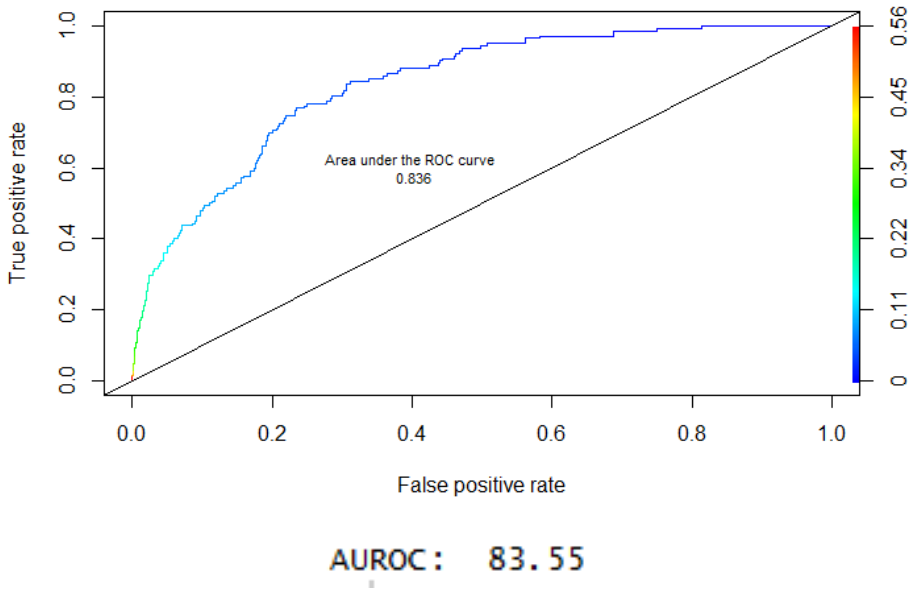
Fuente: (Mondal, 2016)

Con los que tendríamos una capacidad de predicción en nuestro modelo del 96,11%.

Como hemos indicado anteriormente en este trabajo, hemos aplicado además otros indicadores para conocer el grado de bondad del ajuste del modelo. Todos ellos han sido ya desarrollados en el marco teórico por lo que a continuación presentamos la interpretación de los resultados obtenidos en el ajuste:

5.5.4.1 AUROC

Gráfico 11. Curva ROC y AUROC del modelo estimado



Fuente: Elaboración propia a través de R.

Con un valor de 83,55%, de la imagen anterior, y dada la tabla posterior donde nos da una referencia del grado de ajuste del modelo de acuerdo a su AUROC, consideramos que el ajuste del modelo es muy bueno.

Tabla 5. Interpretación de la curva AUROC

AUROCC	Interpretation
1.0 (100%)	Perfect model
0.9-0.99 (90-99%)	Excellent Model
0.8-0.89 (80-89%)	Very Good Model
0.7-0.79 (70-79%)	Fair Model
0.51-0.69(51-69%)	Poor Model
<0.5(50%)	Worthless Model

Fuente: (Mondal, 2016)

5.5.4.2 K-S

Como hemos remarcado en el marco teórico, el test de Kolmogorov-Smirnov (K-S test) tiene como objeto determinar si dos clases difieren significativamente (Mondal, 2016), es decir, es un indicador de la distancia máxima entre las distribuciones de dos clases, que implica un mejor poder predictivo cuanto más cerca está su valor de 1.

En nuestro caso, intentamos medir si el modelo es capaz de clasificar de una forma muy fiable como "bueno" o "malo". Si estas dos categorías están completamente separadas (esta situación supondría un AUROCC = 100%), el valor del test de K-S sería 1 mientras que este valor sería 0 cuando ambas curvas toman el mismo valor.

Realizado el test K-S, observamos cómo dado un p-valor: 0,9299, aceptaríamos la hipótesis nula de distribución de normalidad a cualquier nivel de significación dado.

Ilustración 17. Test de Kolgomorov-Smirnov del modelo estimado

```
> ks.test(y1,"pnorm")

        One-sample Kolmogorov-Smirnov test

data:  y1
D = 0.0090136, p-value = 0.9299
alternative hypothesis: two-sided
```

Fuente: Elaboración propia a través de R.

Como confirmación de la bondad del ajuste, calculamos la diferencia entre las curvas de la función de distribución empírica (ECDF) y de la función de distribución acumulativa normal, con un valor cercano a 80, lo que nos indica un buen ajuste de la bondad del modelo.

Ilustración 18. Coeficiente del test de Kolgomorov-Smirnov del modelo estimado

```
KS:   79.83
```

Fuente: Elaboración propia a través de R.

5.5.4.3 McFadden

Como hemos descrito anteriormente, McFadden es un indicador de bondad del ajuste para el modelo *logit* muy similar al coeficiente de determinación R^2 utilizado en los métodos de mínimos cuadrados ordinarios.

El coeficiente de McFadden tiene un valor en nuestro modelo estimado de 0,2092 (en la imagen posterior) con lo que consideramos que el modelo presenta un ajuste robusto y estadísticamente significativo.

Ilustración 19. Coeficiente McFadden del modelo estimado

```
> pR2 <- pR2(model)
> pR2 [4]
 McFadden
0.2092844
```

Fuente: Elaboración propia a través de R.

5.6 INTERPRETACIÓN DEL MODELO

Partiendo de un correcto ajuste del modelo, un porcentaje reducido de errores correspondientes a valores significativos (< 3,10%) y sin valores influyentes en los indicadores realizados, consideramos que el modelo estimado responde muy correctamente (a la vista de una tasa de clasificaciones correctas que supera el 96%) a la siguiente fórmula:

$$Fallido \sim e^{-1{,}33879} * e^{-0{,}12876*Avales\ técnicos} * e^{-6{,}67876*Caja/Activo\ total} * e^{-1{,}8952*Destino:Fianza} * e^{-0{,}97603*\ Beneficio\ neto\ positivo} * e^{-6{,}78070*\ Tipo\ interés} * e^{-0{,}12298*Avales\ financieros} * e^{-0{,}73285*Margen\ bruto.Ventas} * e^{+1{,}64656*Deuda\ a\ largo\ plazo/Activo\ total} * e^{-1{,}24728*\ Activo\ fijo/Activo\ total} * e^{-0{,}447*Garantías\ hipotecarias} * e^{+0{,}34966*Tamaño} * e^{-0{,}44486*Destino:Circulante}$$

6 CONCLUSIONES

En múltiples ocasiones, las pymes y empresas se encuentran en serias dificultades para obtener acceso a financiación ajena o avales y poder llevar a cabo su actividad. Para muchas de estas pequeñas empresas y pymes, el no tener acceso a esta financiación supone el cese de su actividad y por consiguiente su cierre.

Para poder mitigar estos problemas de acceso al crédito con los que se encuentran las pymes, lo ideal sería contar con un modelo que permita evaluar el riesgo de que estas sean capaces de hacer frente a sus compromisos de pago en el momento en el que se les conceda una financiación, con el fin de que tengan acceso a un crédito el mayor número posible de pymes y no queden discriminadas por un mal análisis.

Por este motivo el presente trabajo tiene como principal objetivo crear un modelo que consiga diferenciar de una base de datos de empresas y pymes madrileñas las variables significativas que nos permitan evaluar el riesgo de que se produzca default.

Para cumplir con nuestro objetivo, hemos trabajado con una base de datos proporcionada por una Sociedad de Garantía Recíproca madrileña con registros desde el 2.008 al 2.017 y el lenguaje de programación R.

No encontramos apenas diferencia de resultados en los distintos indicadores utilizados para poder discernir entre si debiéramos de usar un modelo de transformación *logit* o pobit. Por tanto, en base a los ligeramente superiores valores del modelo *logit* sobre el *probit* (AUROCC, McFadden, AIC y el contraste de Hosmer-Lemeshow) y, sobre todo, a la múltiple literatura relacionada que utiliza en su mayoría un modelo *logit* sobre *probit*, realizaremos nuestra estimación del modelo siguiendo la transformación logarítmica.

Seleccionamos para nuestro modelo inicial aquellas variables de la literatura relacionada revisada que resultaran ser significativas en una estimación con nuestra base de datos (Ver anexo 1) y procedemos a realizar el proceso de estimación en el que se han tanto contrastes de validación (contrastes de significación individual y conjunta, contrastes de colinealidad y comprobación de valores influyentes) como de bondad del modelo (AUROCC, Hosmer-Lemeshow, grado de predicción, Kolgomorov-Smirnov o McFadden).

El resultado final del modelo estimado incluye las siguientes variables significativas: Avales técnicos, Caja / Activo Total, Destino: Fianza, Beneficio neto positivo en el año anterior, Tipo de interés, Avales financieros, Margen bruto / Ventas, Deuda a largo plazo / Activos totales, Activos fijos / Activos totales, Garantías hipotecarias, Tamaño, Destino: Financiación de circulante.

De las variables que acabamos de mencionar, el coeficiente estimado del ratio Caja / Activo Total, estudiado en (Altman, 2010), sigue el mismo sentido que dicho paper, por lo que compañías con un alto ratio de caja sobre activo total exhibirán una menor proporción al default. Al contrario que en (Ohlson, 1980), las posibilidades de default en una empresa aumentan cuanto mayor sea el logaritmo del cociente entre sus

activos totales y el índice de precios de Producto Nacional Bruto, mientras que dichas posibilidades disminuyen si el beneficio neto en el año previo ha sido positivo. Respecto a los ratios introducidos en nuestro modelo procedentes de evaluaciones previas en (Altman & Sabato 2007), tanto el Margen bruto sobre Ventas, como el ratio Activo fijo sobre Activo total y como la deuda a largo plazo sobre el activo total siguen la lógica financiera por la que los dos primeros reducen las probabilidades de default cuanto mayor sea su valor, mientras que una mayor cantidad a largo plazo de deuda aumenta las probabilidades de default de una compañía. Finalmente, hemos introducido una serie de variables no comentadas en la literatura relacionada pero que creíamos relevantes para el modelo. En los cinco casos (Avales técnicos, Destino: Fianza, , Tipo de interés, Avales financieros, Garantías hipotecarias, Destino: Financiación de circulante) el sentido de los coeficiente en el que nos indica la lógica financiera, por la que las posibilidades de default disminuyen cuántos más avales o garantías se aporten en la operación mientras que también disminuyen si la operación se destina a la financiación de circulante o aporte de fianzas respecto a otras categorías (léase adquisición de CAPEX, expansión nacional o internacional o de inicio de la actividad).

Como punto final de este trabajo, consideramos muy interesantes futuras líneas de actuación enfocadas en tres principales direcciones: por un lado, verificar que los resultados obtenidos son de aplicación a empresas de otras comunidades autónomas españolas (recordemos que nuestra base de datos solo contiene registros de la Comunidad Autónoma de Madrid); consideramos que las empresas analizadas en este estudio (PYMES en su mayoría) tienen una exposición mucho mayor a una crisis económico-financieras como la acaecida en el 2008 que cualquier multinacional (o empresa ya diversificada) por lo que creemos necesario la inclusión de variables macroeconómicas como así lo hacen Berteloot et al. (2013); y, por otro lado, dado el avance significativo que ha tenido en los últimos años las Fintech en general, y las plataformas de *crowdlending* en particular, consideramos de especial interés un estudio comparativo sobre los diferentes métodos de estimación de estas instituciones con otras más “tradicionales” como instituciones financieras o sociedades de garantía recíproca.

Índice de Ilustraciones

Índice de Tablas

Índice de Gráficos

7 BIBLIOGRAFÍA

Abdou, H. & Pointon, J., 2011. Credit scoring, statistical techniques and evaluation criteria: A review of the literature. *Intelligent System in Accounting, Finance and Management,* pp. 59-88.

Alamilla-López, N. E. & Arauco Comargo, S., 2009. Limitaciones del modelo lineal de probabilidad y alternativas de modelación microeconómica. En: *Temas de Ciencia y Tecnología.* s.l.:s.n., pp. 3-12.

Altman, E. I., 1968. Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy. *The Journal of Finance,* pp. 589-609.

Altman, E. I. & Sabato, G., 2005. Effects of the New Basel Capital Accord on Bank Capital Requirements for SMEs. *Journal of Financial Services Research,* pp. 15-42.

Altman, E. I., Sabato, G. & Wilson, N., 2010. The value of non-financial information in small and medium-sized enterprise risk management. *The Journal of Credit Risk,* pp. 1-33.

Angulo, J. C. C., s.f. *Modelo microeconómico para el análisis de la diferenciación de productos.* s.l.:s.n.

Baesens, B. y otros, 2003. Benchmarking state-of-the-art classification algorithms for credit scoring. *Journal of Operational Research Society,* pp. 627-635.

Bartlett, J., 2014. *The Stats Geek.* [En línea]
Available at: http://thestatsgeek.com/2014/05/05/area-under-the-roc-curve-assessing-discrimination-in-logistic-regression/

Bathia, S. y otros, 2017. Credit Scoring using Machine Learning Techniques. *International Journal of Computer Applications.*

Benavente, J. M., 2003. *Microeconometría,* s.l.: Universidad de Chile.

Bequé, A. & Lessmann, S., 2017. Extreme learning machines for credit scoring: An empirical evaluation. *Ecpert Systems With Applications,* pp. 42-53.

Berteloot, K. y otros, 2013. A Novel Credit Rating Migration Modeling Approach Using Macroeconomic Indicators. *Journal of Forecasting,* pp. 654-672.

Blanco, A., Pino-Mejías, R., Lara, J. & Rayo, S., 2013. Credit scoring models for the microfinance industry using neural networks: Evidence from Peru. *Expert Systems with Applications,* pp. 356-364.

Cardone-Riportella, C., Trujillo-Ponce, A. & Briozzo, A., 2013. Analyzing the role of mutual guarantee societies on bank capital requirements for small and medium-sized enterprises. *Journal of Economic Policy Reform,* pp. 142-159.

Chakravarti, I. M., Laha, R. G. & Roy, J., 1967. Handbook of Methods of Applied Statistics. En: Hoboken: John Wiley & Sons, pp. 392-394.

Chen, M.-C. & Huang, S.-H., 2003. Credit scoring and rejected instances reassigning through. *Expert Systems with Applications,* pp. 433-441.

Cifuentes, C. C., 2015. *Modelos Logit y Probit,* s.l.: Facultad Ciencias Sociales. Estadística III.

Durand, D., 1941. Risk Elements in Consumer Instalment Financing.

Garcia-Tabuenca, A. & Crespo-Espert, J. L., 2008. Credit guarantees and SME efficiency. *Small Business Economics,* pp. 113-128.

Hauck, W. W. & Donner, A., 1977. Wald's Test as Appied to Hyputheses in Logit Analysis. *Journal of the American Statiscal Association,* pp. 851-853.

Hwang, R.-C., 2013. Forecasting credit rating with the varying-coefficient model. *Journal of Quantitative Finance,* pp. 1947-1965.

Iazzolino, G., Bruni, M. E. & Beraldi, P., 2013. Using DEA and Financial ratings for credit risk evaluation. *Applied Economics Letters,* 20(14), pp. 1310-1317.

INE, 2018. *Demografia y poblacion.* [En línea]
Available at:
http://www.ine.es/dyngs/INEbase/es/categoria.htm?c=Estadistica_P&cid=1254734710984

Investopedia, 2018. *Investopedia.* [En línea]
Available at: https://www.investopedia.com/terms/g/gini-index.asp

Jacobs, M. J. & Bag, P., 2011. What do we know about exposure at default on contigent credit lines? - A survey of the literature, empirical analysis and models. *Journal of Advanced Studies in Finance,* pp. 26-46.

Kirkman, T., 2018. *Statistics to Use.* [En línea]
Available at: http://www.physics.csbsju.edu/stats/

Koh, H. C., Goh, C. . P. & Tan, W. C., 2006. A Two-step Method to Construct Credit Scoring Models with Data Mining Techniques. *International Journal of Business and Information,* 1(1), pp. 96-118.

Lehmann, B., 2003. *Is It Worth the While? The Relevance of Qualitative Information in Credit Rating.* Helsinki, s.n.

Lessmann, S., Baesens, B., Seow, H.-V. & Thomas, L. C., 2015. Benchmarking state-of-the-art classification algorithms for credit scoring: An update or research. *European Journal of Operations Research,* pp. 1-32.

Marín, J., s.f. *Tema 4: Modelos avanzados con Regresión Logística,* s.l.: Universidad Carlos III Madrid.

MathWorks, 2018. *Documentación MathWorks.* [En línea]
Available at: http://se.mathworks.com/help/stats/cooks-distance.html

Michael Jacobs, J., 2010. *What do we know about exposure at default on contingent credit lines? - A survey of the literature, empirical analysis and models,* s.l.: s.n.

Modina, M. & Pietrovito, F., 2014. A default prediction model for Italian SMEs: the relevance of the capital structure. *Journal of Applied Financial Economics,* pp. 1537-1554.

Mondal, A., 2016. *R Studio pubs.* [En línea]
Available at: https://rstudio-pubs-static.s3.amazonaws.com/225209_df0130c5a0614790b6365676b9372c07.html#41_receiver_operating_characteristic(roc)_curve

Moral, E. M., 2003. *Modelos de elección discreta,* Madrid: Universidad Autónoma de Madrid.

Nist/Sematech, 2018. *e-Handbook of Statistical Methods.* [En línea]
Available at: http://www.itl.nist.gov/div898/handbook/

Ohlson, J. A., 1980. Financial ratios and the probabilistic prediction of bankruptcy. *Journal of Accounting Research,* pp. 109-131.

PennState, 2018. *STAT 501 | Regression Methods.* [En línea]
Available at: https://onlinecourses.science.psu.edu/stat501/node/347/

Pérez, J. L., 2018. *La estadística: Una orquesta hecha instrumento.* [En línea]
Available at: https://estadisticaorquestainstrumento.wordpress.com/2013/04/30/test-de-wald/

Raei, R., Kousha, M. S., Fallahpour, S. & Fadaeinejad, M., 2016. A hybrid model for estimating the probability of default of corporate customers. *Iranian Journal of Management,* pp. 651-673.

Reche, J. L. C., 2013. *Regresión logística. Tratamiento computacional con R,* s.l.: Universidad de Granada.

Roche, J. L. C., 2013. *Regresión lgística. Tratamiento computacional con R,* Granada: s.n.

Rudden, R., s.f. *Caribbean Information & Credit Rating Services Limited.* [En línea]
Available at: http://www.caricris.com/images/pdfs/article/evolutionpart1.pdf
[Último acceso: 26 03 2018].

S&P Global, 2018. *Definiciones de Calificaciones de S&P Global Ratings,* s.l.: s.n.

Salimi, A. Y., 2015. Validity of Altmans Z-Score model in predicting brankruptcy in recent years. *Academy of Accounting and Financial Studies Journal,* pp. 233-238.

Silva, V. H. U., 2013. *Comparación de los modelos logit y probit del análisis multinivel, en el estudio del rendimiento escolar",* s.l.: Universidad Nacional Mayor de San Marcos.

Stefan Lessmann, B. B. ,. H.-V. S. ,. L. C. T., s.f. *Benchmarking state-of-the-art classification algorithms for credit scoring: A ten-year update,* s.l.: s.n.

Stevenson, T. & Pond, K., 2016. SME lending decisions – the caseof UK and German banks. An international comparison. *Studies in Economics and Finance,* pp. 501-508.

Sylla, R., 2002. An Historical Primer on the Business of Credit Rating. En: *Ratings Agencies and the Global Financial System.* Boston: Spinger, pp. 19-40.

Thomas, L. C., Edelman, D. B. & Crook, J. N., 2001. *Credit Scoring and Its Applications.* s.l.:SIAM.

Universidad de Granada, s.f. *Modelos de eleeción discreta,* s.l.: s.n.

Vásquez, F. P., 2002. *Los modelos logit y probit en la investigación social,* Lima: Centro de Investigación y Desarrollo (CIDE).

Westgaard, S. & Van de Wijst, N., 2001. Default probabilities in a corporate bank portfolio: A logistic model approach. *European Journal of Operations Research,* pp. 338-349.

White, L. J., 2013. Credit Rating Agencies: An Overview. *Annual Review of Financial Economics,* pp. 93-122.

Wiginton, J. C., 1980. A Note on the Comparison of Logit and Discriminant Models of Consumer Credit Behaviour. *The Journal of Financial and Quantitative Analysis,* pp. 757-770.

8 ANEXOS

8.1 DEFINICIÓN DE LAS ABREVIATURAS DEL CUADRO RESUMEN DE LA LITERATURA RELACIONADA

Abreviatura	Definición
WC	Working Capital
RE	Retained Earnings
TTD	Time To Default
ORR	Obligor Risk Rating
PERCBNK	Proportion of bank debt in the capital structure
PERSEC	Debt on the obligor capital structure in the data set
MSG12MTDR	Moody's Speculation 12 Month Trailing Default Rate
TLTA	Total de pasivo entre activos totales
WCTA	Fondo de maniobra entre activos totales
CLCA	Pasivo corriente entre activo corriente
NITA	Beneficio neto entre activos totales
FUTL	Beneficio de explotación entre pasivo total
INTWO	Uno si el beneficio neto de los dos últimos años fue negativo, cero en caso contrario
OENEG	Uno si el pasivo total excede del activo, cero en caso contrario
AT	Activo total
PT	Pasivo total
PMC	Periodo medio de cobro
LP	Largo plazo
CT	Capital total
AD	Activos disponibles
AC	Activos corrientes
PC	Pasivos corrientes
ACO	Acreedores comerciales
DCO	Deudores comerciales
PN	Patrimonio Neto
BDI	Beneficio después de impuestos
CNCF	Corporate Net Cash Flow
PIB	Producto interior bruto
PNB	Producto nacional bruto
IP	Índice precios
BAI	Beneficio antes de impuestos
TT	Tasa del tesoro

8.2 CUADRO RESUMEN DE LAS VARIABLES SIGNIFICATIVAS RESULTANTES DE LAS DIFERENTES ESTIMACIONES

Modelo	Variable independiente	Variable dependiente	Estimate	Std. Error	z value	Pr(>|z|)	Significativa
Propio (29 variables)	Población	Fallido	0,00	0,00	-3,49	0,00	***
Propio (29 variables)	Destino Operación. Fianza	Fallido	-2,90	0,73	-3,40	0,00	***
Propio (29 variables)	ROA	Fallido	3,70	1,56	2,38	0,02	*
Propio (29 variables)	Periodo Medio de Cobro	Fallido	0,00	0,00	2,11	0,03	*
Propio (29 variables)	Activo fijo sobre Activo total	Fallido	-0,83	0,39	-2,10	0,04	*
Propio (29 variables)	Activo total	Fallido	0,00	0,00	-2,97	0,00	**
Propio (29 variables)	Deuda bancaria l/p sobre Deuda bancaria total	Fallido	0,68	0,39	1,77	0,08	.
Propio (29 variables)	Fondos propios / Pasivo (valor en libros)	Fallido	-0,28	0,13	-2,21	0,03	*
Propio (29 variables)	Caja / Activo total	Fallido	-5,41	1,31	-4,13	0,00	***
Propio (29 variables)	Margen bruto / Activo total	Fallido	-0,36	0,17	-21,05	0,04	*
Propio (29 variables)	OENEG	Fallido	-0,65	0,23	-2,83	0,00	**
Propio (29 variables)	Indicador. McFadden	Fallido	0,25				
Propio (29 variables)	Indicador. Predicción	Fallido	0,93				
Propio (29 variables)	Indicador. AUC	Fallido	0,67				

Propio (29 variables)	Población	Fallido2	0,00	0,00	-1,82	0,07	.
Propio (29 variables)	Indicador. McFadden	Fallido2	0,27				
Propio (29 variables)	Indicador. Predicción	Fallido2	No disponible				
Propio (29 variables)	Indicador. AUC	Fallido2	No disponible				
Propio (29 variables)	Activo fijo sobre Activo total	Fallido3	-2,02	0,82	-2,45	0,01	*
Propio (29 variables)	Deuda bancaria l/p sobre Deuda bancaria total	Fallido3	2,77	0,82	3,37	0,00	***
Propio (29 variables)	Fondos propios / Pasivo (valor en libros)	Fallido3	-0,36	0,17	-2,19	0,03	*
Propio (29 variables)	Caja / Activo total	Fallido3	-8,12	3,89	-2,09	0,04	*
Propio (29 variables)	Caja / Deuda bancaria	Fallido3	-0,47	0,21	-2,28	0,02	*
Propio (29 variables)	Margen bruto / Activo total	Fallido3	-0,67	0,40	-1,68	0,09	.
Propio (29 variables)	OENEG	Fallido3	-0,97	0,48	-2,04	0,04	*
Propio (29 variables)	Indicador. McFadden	Fallido3	0,27				
Propio (29 variables)	Indicador. Predicción	Fallido3	0,99				
Propio (29 variables)	Indicador. AUC	Fallido3	0,41				
Propio (29 variables)	ROA	Fallido4	5,84	2,93	1,99	0,05	*
Propio (29 variables)	Activo total / Deuda bancaria	Fallido4	-1,44	0,84	-1,72	0,09	.

Propio (29 variables)	Fondos propios / Pasivo (valor en libros)	Fallido4	-0,54	0,26	-2,07	0,04	*
Propio (29 variables)	Caja / Activo total	Fallido4	-3,90	1,67	-2,34	0,02	*
Propio (29 variables)	Caja / Deuda bancaria	Fallido4	0,82	0,35	2,38	0,02	*
Propio (29 variables)	Margen bruto / Activo total	Fallido4	-0,91	0,40	-2,31	0,02	*
Propio (29 variables)	OENEG	Fallido4	-1,29	0,47	-2,73	0,01	**
Propio (29 variables)	Pasivo / Activo Total	Fallido4	-2,12	0,87	-2,44	0,01	*
Propio (29 variables)	Indicador. McFadden	Fallido4	0,24				
Propio (29 variables)	Indicador. Predicción	Fallido4	0,98				
Propio (29 variables)	Indicador. AUC	Fallido4	0,64				
Propio (29 variables)	Población	Fallido5	0,00	0,00	-2,23	0,03	*
Propio (29 variables)	Deuda bancaria l/p sobre Deuda bancaria total	Fallido5	2,00	0,96	2,07	0,04	*
Propio (29 variables)	Caja / Activo total	Fallido5	-5,57	3,28	-1,70	0,09	.
Propio (29 variables)	Indicador. McFadden	Fallido5	0,24				
Propio (29 variables)	Indicador. Predicción	Fallido5	0,98				
Propio (29 variables)	Indicador. AUC	Fallido5	0,66				
Propio (29 variables)	Población	Fallido+5	-0,37	0,00	-2,75	0,01	**
Propio (29 variables)	Empleos temporales	Fallido+5	0,01	0,01	2,02	0,04	*
Propio (29 variables)	Destino Operación. Fianza	Fallido+5	-2,45	1,32	-1,86	0,06	.

Propio (29 variables)	Destino Operación. Ampliación de actividad	Fallido+5	2,16	1,23	1,76	0,08	.
Propio (29 variables)	Activo Corriente / Pasivo Corriente	Fallido+5	0,42	2,2293-01	1,89	0,06	.
Propio (29 variables)	Fondos propios / Pasivo (valor en libros)	Fallido+5	-9,74	4,1163-01	-2,12	0,03	*
Propio (29 variables)	Caja / Activo total	Fallido+5	-0,20	5,16	-2,33	0,02	*
Propio (29 variables)	Deuda bancaria / Fondos propios	Fallido+5	0,00	0,00	-1,82	0,07	.
Propio (29 variables)	Capital Circulante / Activo total	Fallido+5	-53770,00	265000,00	-2,03	0,04	
Propio (29 variables)	Indicador. McFadden	Fallido+5	0,28				
Propio (29 variables)	Indicador. Predicción	Fallido+5	0,98				
Propio (29 variables)	Indicador. AUC	Fallido+5	0,71				
(Altman & Sabato, 2005) (Australia)	Activo total	Fallido	0,00	0,00	-1,72	0,09	.
(Altman & Sabato, 2005) (Australia)	Indicador. McFadden	Fallido	0,03				
(Altman & Sabato, 2005) (Australia)	Indicador. Predicción	Fallido	1,00				
(Altman & Sabato, 2005) (Australia)	Indicador. AUC	Fallido	0,56				
(Altman & Sabato,	Activo total	Fallido2	0,00	0,00	-1,72	0,09	.

2005) (Australia)							
(Altman & Sabato, 2005) (Australia)	Indicador. McFadden	Fallido2	0,03				
(Altman & Sabato, 2005) (Australia)	Indicador. Predicción	Fallido2	1,00				
(Altman & Sabato, 2005) (Australia)	Indicador. AUC	Fallido2	0,56				
(Altman & Sabato, 2005) (Australia)	Activo total	Fallido3	0,00	0,00	-2,61	0,01	**
(Altman & Sabato, 2005) (Australia)	Indicador. McFadden	Fallido3	39699,65				
(Altman & Sabato, 2005) (Australia)	Indicador. Predicción	Fallido3	0,99				
(Altman & Sabato, 2005) (Australia)	Indicador. AUC	Fallido3	0,49				
(Altman & Sabato, 2005) (Australia)	EBIT / Activo total	Fallido4	-0,28	0,15	-1,85	0,06	.
(Altman & Sabato, 2005) (Australia)	Indicador. McFadden	Fallido4	4864,57				
(Altman & Sabato, 2005) (Australia)	Indicador. Predicción	Fallido4	0,99				
(Altman & Sabato, 2005) (Australia)	Indicador. AUC	Fallido4	0,65				

(Altman & Sabato, 2005) (Australia)	Activo total	Fallido5	0,00	0,00	-2,15	0,03	*
(Altman & Sabato, 2005) (Australia)	Indicador. McFadden	Fallido5	38960,60				
(Altman & Sabato, 2005) (Australia)	Indicador. Predicción	Fallido5	0,98				
(Altman & Sabato, 2005) (Australia)	Indicador. AUC	Fallido5	0,60				
(Altman & Sabato, 2005) (Australia)	Activo Corriente / Pasivo Corriente	Fallido+5	-6,20	0,03	-2,11	0,03	*
(Altman & Sabato, 2005) (Australia)	Activo total	Fallido+5	0,00	0,00	-1,75	0,08	.
(Altman & Sabato, 2005) (Australia)	Indicador. McFadden	Fallido+5	26712,43				
(Altman & Sabato, 2005) (Australia)	Indicador. Predicción	Fallido+5	0,98				
(Altman & Sabato, 2005) (Australia)	Indicador. AUC	Fallido+5	0,50				
(Altman & Sabato, 2005) (Italia)	Pasivo l/p / Activo total	Fallido	0,95	0,25	3,88	0,00	***
(Altman & Sabato, 2005) (Italia)	Activo fijo sobre Activo total	Fallido	1,11	0,28	4,00	0,00	***
(Altman & Sabato, 2005) (Italia)	Margen bruto / Activo total	Fallido	-0,18	0,11	-1,65	0,10	.

(Altman & Sabato, 2005) (Italia)	Indicador. McFadden	Fallido	33729,67				
(Altman & Sabato, 2005) (Italia)	Indicador. Predicción	Fallido	0,93				
(Altman & Sabato, 2005) (Italia)	Indicador. AUC	Fallido	0,55				
(Altman & Sabato, 2005) (Italia)	Indicador. McFadden	Fallido2	5989,92				
(Altman & Sabato, 2005) (Italia)	Indicador. Predicción	Fallido2	No disponible				
(Altman & Sabato, 2005) (Italia)	Indicador. AUC	Fallido2	No disponible				
(Altman & Sabato, 2005) (Italia)	Pasivo l/p / Activo total	Fallido3	1,41	0,28	5,08	0,00	***
(Altman & Sabato, 2005) (Italia)	Activo total / Deuda bancaria	Fallido3	-0,40	0,13	-3,03	0,00	**
(Altman & Sabato, 2005) (Italia)	Margen bruto / Activo total	Fallido3	-0,19	0,08	-2,22	0,03	*
(Altman & Sabato, 2005) (Italia)	Indicador. McFadden	Fallido3	52579,38				
(Altman & Sabato, 2005) (Italia)	Indicador. Predicción	Fallido3	0,99				
(Altman & Sabato, 2005) (Italia)	Indicador. AUC	Fallido3	0,41				
(Altman & Sabato, 2005) (Italia)	Pasivo l/p / Activo total	Fallido4	0,73	0,37	1,97	0,05	*
(Altman & Sabato, 2005) (Italia)	Margen bruto / Activo total	Fallido4	0,60	0,15	-4,00	0,00	***
(Altman & Sabato, 2005) (Italia)	Indicador. McFadden	Fallido4	28570,84				

(Altman & Sabato, 2005) (Italia)	Indicador. Predicción	Fallido4	0,99				
(Altman & Sabato, 2005) (Italia)	Indicador. AUC	Fallido4	0,55				
(Altman & Sabato, 2005) (Italia)	Pasivo l/p / Activo total	Fallido5	1,13	0,37	3,02	0,00	**
(Altman & Sabato, 2005) (Italia)	Activo fijo sobre Activo total	Fallido5	1,22	0,60	2,01	0,04	*
(Altman & Sabato, 2005) (Italia)	Activo total / Deuda bancaria	Fallido5	-0,33	0,18	-1802,00	0,07	.
(Altman & Sabato, 2005) (Italia)	Indicador. McFadden	Fallido5	33869,62				
(Altman & Sabato, 2005) (Italia)	Indicador. Predicción	Fallido5	0,98				
(Altman & Sabato, 2005) (Italia)	Indicador. AUC	Fallido5	0,57				
(Altman & Sabato, 2005) (Italia)	Activo fijo sobre Activo total	Fallido+5	3,29	0,61	5,41	0,00	***
(Altman & Sabato, 2005) (Italia)	Pasivo l/p / Activo total	Fallido+5	-1,20	0,54	-2,25	0,02	*
(Altman & Sabato, 2005) (Italia)	Deuda bancaria / Fondos propios	Fallido+5	-0,01	0,00	-1,94	0,05	.
(Altman & Sabato, 2005) (Italia)	Indicador. McFadden	Fallido+5	33869,62				
(Altman & Sabato, 2005) (Italia)	Indicador. Predicción	Fallido+5	0,98				
(Altman & Sabato, 2005) (Italia)	Indicador. AUC	Fallido+5	0,57				
(Altman & Sabato, 2005) (USA)	EBIT / Activo total	Fallido	-0,41	0,10	-4,07	0,00	***

(Altman & Sabato, 2005) (USA)	Indicador. McFadden	Fallido	15159,34				
(Altman & Sabato, 2005) (USA)	Indicador. Predicción	Fallido	0,93				
(Altman & Sabato, 2005) (USA)	Indicador. AUC	Fallido	0,60				
(Altman & Sabato, 2005) (USA)	EBIT / Activo total	Fallido2	-0,30	0,15	-1,95	0,05	.
(Altman & Sabato, 2005) (USA)	Indicador. McFadden	Fallido2	10802,54				
(Altman & Sabato, 2005) (USA)	Indicador. Predicción	Fallido2	1,00				
(Altman & Sabato, 2005) (USA)	Indicador. AUC	Fallido2	0,54				
(Altman & Sabato, 2005) (USA)	Indicador. McFadden	Fallido3	8964,30				
(Altman & Sabato, 2005) (USA)	Indicador. Predicción	Fallido3	No disponible				
(Altman & Sabato, 2005) (USA)	Indicador. AUC	Fallido3	No disponible				
(Altman & Sabato, 2005) (USA)	Indicador. McFadden	Fallido4	1164,64				
(Altman & Sabato, 2005) (USA)	Indicador. Predicción	Fallido4	No disponible				
(Altman & Sabato, 2005) (USA)	Indicador. AUC	Fallido4	No disponible				
(Altman & Sabato, 2005) (USA)	EBIT / Activo total	Fallido5	-0,32	0,12	-2,75	0,01	**
(Altman & Sabato, 2005) (USA)	Indicador. McFadden	Fallido5	12719,93				

(Altman & Sabato, 2005) (USA)	Indicador. Predicción	Fallido5	0,98				
(Altman & Sabato, 2005) (USA)	Indicador. AUC	Fallido5	0,62				
(Altman & Sabato, 2005) (USA)	Indicador. McFadden	Fallido+5	8318,05				
(Altman & Sabato, 2005) (USA)	Indicador. Predicción	Fallido+5	No disponible				
(Altman & Sabato, 2005) (USA)	Indicador. AUC	Fallido+5	No disponible				
(Altman et al., 2010)	Quick assets / Activo Circulante	Fallido	0,09	0,04	2,56	0,01	*
(Altman et al., 2010)	Inventario / Capital Circulante	Fallido	0,00	0,00	-1,98	0,05	*
(Altman et al., 2010)	Activo Corriente / Pasivo Corriente;	Fallido	0,01	0,01	2,33	0,02	*
(Altman et al., 2010)	Caja / Activo total	Fallido	-7,63	1,24	-6,17	0,00	***
(Altman et al., 2010)	Indicador. McFadden	Fallido	53062,35				
(Altman et al., 2010)	Indicador. Predicción	Fallido	0,94				
(Altman et al., 2010)	Indicador. AUC	Fallido	0,63				
(Altman et al., 2010)	Caja / Activo total	Fallido2	-9,99	4,77	-2,10	0,04	*
(Altman et al., 2010)	Indicador. McFadden	Fallido2	0,05				
(Altman et al., 2010)	Indicador. Predicción	Fallido2	0,99				
(Altman et al., 2010)	Indicador. AUC	Fallido2	0,54				
(Altman et al., 2010)	Quick assets/ Activo Circulante	Fallido3	0,10	0,04	2,49	0,01	*
(Altman et al., 2010)	Activo Corriente / Pasivo Corriente;	Fallido3	0,02	0,01	3,83	0,00	***
(Altman et al., 2010)	Caja / Activo total	Fallido3	-21,66	5,96	-3,64	0,00	***
(Altman et al., 2010)	Indicador. McFadden	Fallido3	112887,70				

(Altman et al., 2010)	Indicador. Predicción	Fallido3	0,99				
(Altman et al., 2010)	Indicador. AUC	Fallido3	0,49				
(Altman et al., 2010)	Caja / Activo total	Fallido4	-6,71	2,11	-3,18	0,00	**
(Altman et al., 2010)	Indicador. McFadden	Fallido4	36945,47				
(Altman et al., 2010)	Indicador. Predicción	Fallido4	0,99				
(Altman et al., 2010)	Indicador. AUC	Fallido4	0,57				
(Altman et al., 2010)	Acreedores / Deudores	Fallido5	0,00	0,00	-1,96	0,05	.
(Altman et al., 2010)	Acreedores / Pasivo total	Fallido5	0,43	0,20	2,22	0,03	*
(Altman et al., 2010)	Activo Corriente / Pasivo Corriente;	Fallido5	-0,06	0,03	-1,83	0,07	.
(Altman et al., 2010)	Indicador. McFadden	Fallido5	19139,36				
(Altman et al., 2010)	Indicador. Predicción	Fallido5	0,98				
(Altman et al., 2010)	Indicador. AUC	Fallido5	0,62				
(Altman et al., 2010)	Activo Corriente / Pasivo Corriente;	Fallido+5	-0,15	0,07	-2,07	0,04	*
(Altman et al., 2010)	Caja / Activo total	Fallido+5	-23,54	6,60	-3,57	0,00	***
(Altman et al., 2010)	Indicador. McFadden	Fallido+5	0,12				
(Altman et al., 2010)	Indicador. Predicción	Fallido+5	0,98				
(Altman et al., 2010)	Indicador. AUC	Fallido+5	0,62				
(Ohlson, 1980)	Tamaño	Fallido	-0,47	0,10	-4,73	0,00	***
(Ohlson, 1980)	OENEG	Fallido	1,04	0,18	5,90	0,00	***
(Ohlson, 1980)	INONE	Fallido	0,56	0,16	3,42	0,00	***
(Ohlson, 1980)	Indicador. McFadden	Fallido	0,07				
(Ohlson, 1980)	Indicador. Predicción	Fallido	0,93				

(Ohlson, 1980)	Indicador. AUC	Fallido	0,62				
(Ohlson, 1980)	Indicador. McFadden	Fallido2	22627,70				
(Ohlson, 1980)	Indicador. Predicción	Fallido2	No disponible				
(Ohlson, 1980)	Indicador. AUC	Fallido2	No disponible				
(Ohlson, 1980)	Tamaño	Fallido3	-0,66	0,19	-3,51	0,00	***
(Ohlson, 1980)	OENEG	Fallido3	1,13	0,33	3,44	0,00	***
(Ohlson, 1980)	INONE	Fallido3	0,61	0,32	1,90	0,06	.
(Ohlson, 1980)	Indicador. McFadden	Fallido3	82133,52				
(Ohlson, 1980)	Indicador. Predicción	Fallido3	0,99				
(Ohlson, 1980)	Indicador. AUC	Fallido3	0,50				
(Ohlson, 1980)	Tamaño	Fallido4	-0,57	0,20	-2,90	0,00	**
(Ohlson, 1980)	OENEG	Fallido4	0,77	0,35	2,20	0,03	*
(Ohlson, 1980)	INONE	Fallido4	0,57	0,31	1,83	0,07	.
(Ohlson, 1980)	Indicador. McFadden	Fallido4	49198,03				
(Ohlson, 1980)	Indicador. Predicción	Fallido4	0,99				
(Ohlson, 1980)	Indicador. AUC	Fallido4	0,58				
(Ohlson, 1980)	Tamaño	Fallido5	-0,72	0,24	-3,24	0,00	**
(Ohlson, 1980)	Indicador. McFadden	Fallido5	44555,69				
(Ohlson, 1980)	Indicador. Predicción	Fallido5	0,98				
(Ohlson, 1980)	Indicador. AUC	Fallido5	0,59				
(Ohlson, 1980)	OENEG	Fallido+5	2,05	0,41	5,02	0,00	***

(Ohlson, 1980)	INONE	Fallido+5	0,94	0,41	2,30	0,02	*
(Ohlson, 1980)	Indicador. McFadden	Fallido+5	103807,10				
(Ohlson, 1980)	Indicador. Predicción	Fallido+5	0,98				
(Ohlson, 1980)	Indicador. AUC	Fallido+5	0,67				
(Raei et al., 2016)	Activo fijo sobre Activo total	Fallido	1,64	0,30	5,37	0,00	***
(Raei et al., 2016)	Indicador. McFadden	Fallido	24985,05				
(Raei et al., 2016)	Indicador. Predicción	Fallido	0,94				
(Raei et al., 2016)	Indicador. AUC	Fallido	0,54				
(Raei et al., 2016)	Indicador. McFadden	Fallido2	4335,78				
(Raei et al., 2016)	Indicador. Predicción	Fallido2	No disponible				
(Raei et al., 2016)	Indicador. AUC	Fallido2	No disponible				
(Raei et al., 2016)	Activo fijo sobre Activo total	Fallido3	1,52	0,59	2,58	0,01	**
(Raei et al., 2016)	Indicador. McFadden	Fallido3	16159,52				
(Raei et al., 2016)	Indicador. Predicción	Fallido3	0,99				
(Raei et al., 2016)	Indicador. AUC	Fallido3	0,38				
(Raei et al., 2016)	Margen bruto / Activo total	Fallido4	-1,01	0,26	-3,91	0,00	***
(Raei et al., 2016)	Indicador. McFadden	Fallido4	33047,99				
(Raei et al., 2016)	Indicador. Predicción	Fallido4	0,99				
(Raei et al., 2016)	Indicador. AUC	Fallido4	0,57				
(Raei et al., 2016)	Activo fijo sobre Activo total	Fallido5	1,40	0,70	2,00	0,05	*

(Raei et al., 2016)	Indicador. McFadden	Fallido5	13975,19				
(Raei et al., 2016)	Indicador. Predicción	Fallido5	0,98				
(Raei et al., 2016)	Indicador. AUC	Fallido5	0,56				
(Raei et al., 2016)	Activo fijo sobre Activo total	Fallido+5	3,56	6,54 4-1	5,44	0,00	***
(Raei et al., 2016)	Indicador. McFadden	Fallido+5	73148,58				
(Raei et al., 2016)	Indicador. Predicción	Fallido+5	0,98				
(Raei et al., 2016)	Indicador. AUC	Fallido+5	0,58				
(Westgaard & Van de Wijst, 2001)	Industria. Industria	Fallido	0,66	0,38	1,76	0,08	.
(Westgaard & Van de Wijst, 2001)	Industria. Suministro de energia y agua	Fallido	2,27	1,15	1,97	0,05	*
(Westgaard & Van de Wijst, 2001)	Población	Fallido	0,00	0,00	-4,68	0,00	***
(Westgaard & Van de Wijst, 2001)	Empleos fijos	Fallido	-0,01	0,00	-3,18	0,00	**
(Westgaard & Van de Wijst, 2001)	Empleos indirectos	Fallido	0,13	0,03	3,83	0,00	***
(Westgaard & Van de Wijst, 2001)	Financov	Fallido	0,00	0,00	-1,76	0,08	.
(Westgaard & Van de Wijst, 2001)	Antigüedad	Fallido	-0,02	0,01	-1,88	0,06	.
(Westgaard & Van de Wijst, 2001)	Liquidez	Fallido	-0,05	0,01	-3,79	0,00	***
(Westgaard & Van de Wijst, 2001)	Indicador. McFadden	Fallido	0,10				

(Westgaard & Van de Wijst, 2001)	Indicador. Predicción	Fallido	0,93				
(Westgaard & Van de Wijst, 2001)	Indicador. AUC	Fallido	0,56				
(Westgaard & Van de Wijst, 2001)	Indicador. McFadden	Fallido2	-4,00				
(Westgaard & Van de Wijst, 2001)	Indicador. Predicción	Fallido2	No disponible				
(Westgaard & Van de Wijst, 2001)	Indicador. AUC	Fallido2	No disponible				
(Westgaard & Van de Wijst, 2001)	Población	Fallido3	0,00	0,00	1,66	0,10	.
(Westgaard & Van de Wijst, 2001)	Antigüedad	Fallido3	-0,03	0,02	-1,69	0,09	.
(Westgaard & Van de Wijst, 2001)	Liquidez	Fallido3	-0,07	0,04	-2,02	0,04	*
(Westgaard & Van de Wijst, 2001)	Indicador. McFadden	Fallido3	83275,47				
(Westgaard & Van de Wijst, 2001)	Indicador. Predicción	Fallido3	0,99				
(Westgaard & Van de Wijst, 2001)	Indicador. AUC	Fallido3	0,54				
(Westgaard & Van de Wijst, 2001)	Industria. Suministro de energía y agua	Fallido4	3,03	1,27	2,39	0,02	*
(Westgaard & Van de Wijst, 2001)	Empleos fijos	Fallido4	-0,03	0,01	-2,36	0,02	*
(Westgaard & Van de Wijst, 2001)	Antigüedad	Fallido4	-0,03	0,02	-1,71	0,09	.
(Westgaard & Van de Wijst, 2001)	Indicador. McFadden	Fallido4	96675,75				

(Westgaard & Van de Wijst, 2001)	Indicador. Predicción	Fallido4	0,99				
(Westgaard & Van de Wijst, 2001)	Indicador. AUC	Fallido4	0,52				
(Westgaard & Van de Wijst, 2001)	Población	Fallido5	0,00	0,00	-3,07	0,00	**
(Westgaard & Van de Wijst, 2001)	Empleos fijos	Fallido5	-0,01	0,01	-1,67	0,10	.
(Westgaard & Van de Wijst, 2001)	Financov	Fallido5	0,00	0,00	-1,68	0,09	.
(Westgaard & Van de Wijst, 2001)	Indicador. McFadden	Fallido5	113885,30				
(Westgaard & Van de Wijst, 2001)	Indicador. Predicción	Fallido5	0,98				
(Westgaard & Van de Wijst, 2001)	Indicador. AUC	Fallido5	0,57				
(Westgaard & Van de Wijst, 2001)	Población	Fallido+5	0,00	0,00	-3,05	0,00	**
(Westgaard & Van de Wijst, 2001)	Liquidez	Fallido+5	-0,21	0,09	-2,33	0,02	*
(Westgaard & Van de Wijst, 2001)	Indicador. McFadden	Fallido+5	158559,60				
(Westgaard & Van de Wijst, 2001)	Indicador. Predicción	Fallido+5	0,98				
(Westgaard & Van de Wijst, 2001)	Indicador. AUC	Fallido+5	0,54				
Propio (20 variables)	Duración	Fallido	0,00	0,00	-4,20	0,00	***
Propio (20 variables)	Población	Fallido	0,00	0,00	-2,11	0,03	*
Propio (20 variables)	Empleos temporales	Fallido	0,02	0,01	3,60	0,00	***

Propio (20 variables)	Empleos indirectos	Fallido	0,08	0,03	2,35	0,02	*
Propio (20 variables)	Avales financieros	Fallido	-0,07	0,02	-3,09	0,00	**
Propio (20 variables)	Avales técnicos	Fallido	-0,04	0,01	-4,83	0,00	***
Propio (20 variables)	Destino de la operación. Fianza	Fallido	-1,97	0,44	-4,47	0,00	***
Propio (20 variables)	Destino de la operación. Financiación de circulante	Fallido	-0,74	0,40	-1,83	0,07	.
Propio (20 variables)	Margen bruto / Ventas	Fallido	0,29	0,07	-3,90	0,00	***
Propio (20 variables)	Garantías hipotecarias	Fallido	-0,35	0,16	-2,20	0,03	*
Propio (20 variables)	Garantías pignoración	Fallido	-2,70	0,42	-6,47	0,00	***
Propio (20 variables)	Garantías personales	Fallido	-0,15	0,04	-3,98	0,00	***
Propio (20 variables)	Tipo interés	Fallido	-3,26	0,93	-3,51	0,00	***
Propio (20 variables)	Indicador. McFadden	Fallido	225358,50				
Propio (20 variables)	Indicador. Predicción	Fallido	0,90				
Propio (20 variables)	Indicador. AUC	Fallido	0,73				
Propio (20 variables)	Ventas / Activo total	Fallido2	-0,35	0,14	-2,45	0,01	*
Propio (20 variables)	Garantías pignoracion	Fallido2	-1,97	0,78	-2,46	0,01	*
Propio (20 variables)	Garantías personales	Fallido2	-0,51	0,14	-3,78	0,00	***
Propio (20 variables)	Tipo interés	Fallido2	8,06	2,33	3,45	0,00	***
Propio (20 variables)	Indicador. McFadden	Fallido2	0,25				
Propio (20 variables)	Indicador. Predicción	Fallido2	0,99				
Propio (20 variables)	Indicador. AUC	Fallido2	0,78				

Propio (20 variables)	Duración	Fallido3	-0,02	0,00	-3,26	0,00	**
Propio (20 variables)	Población	Fallido3	0,00	0,00	-1,91	0,06	.
Propio (20 variables)	Empleos fijos	Fallido3	0,01	0,00	2,07	0,04	*
Propio (20 variables)	Empleos temporales	Fallido3	-0,25	0,13	-1,93	0,05	.
Propio (20 variables)	Empleos indirectos	Fallido3	0,16	0,06	2,70	0,01	**
Propio (20 variables)	Empleos hace un año	Fallido3	-0,06	0,03	-1,88	0,06	.
Propio (20 variables)	Avales financieros	Fallido3	-0,08	0,05	-1,70	0,09	.
Propio (20 variables)	Margen bruto / Ventas	Fallido3	-0,18	0,11	-1,67	0,10	.
Propio (20 variables)	Garantías pignoración	Fallido3	-1,92	0,61	-3,13	0,00	**
Propio (20 variables)	Garantías personales	Fallido3	-0,21	0,07	-2,93	0,00	**
Propio (20 variables)	Indicador. McFadden	Fallido3	0,22				
Propio (20 variables)	Indicador. Predicción	Fallido3	0,98				
Propio (20 variables)	Indicador. AUC	Fallido3	0,63				
Propio (20 variables)	Duración	Fallido4	0,00	0,00	-2,35	0,02	*
Propio (20 variables)	Empleos indirectos	Fallido4	0,09	0,04	2,07	0,04	*
Propio (20 variables)	Avales técnicos	Fallido4	-0,03	0,01	-2,98	0,00	**
Propio (20 variables)	Destino de la operación. Fianza	Fallido4	-3,00	0,58	-5,19	0,00	***
Propio (20 variables)	Destino de la operación. CAPEX	Fallido4	-1,87	0,54	-3,49	0,00	***
Propio (20 variables)	Destino de la operación. Financiación de circulante	Fallido4	-1,34	0,48	-2,80	0,01	**

Propio (20 variables)	Destino de la operación. Financiación de inicio de actividad	Fallido4	-1,15	0,46	-2,50	0,01	*
Propio (20 variables)	Margen bruto / Ventas	Fallido4	-3,45	0,11	-2,21	0,03	*
Propio (20 variables)	Garantías pignoración	Fallido4	-2,93	1,01	-2,89	0,00	**
Propio (20 variables)	Tipo interés	Fallido4	-6,35	1,65	-3,84	0,00	***
Propio (20 variables)	Indicador. McFadden	Fallido4	0,18				
Propio (20 variables)	Indicador. Predicción	Fallido4	0,98				
Propio (20 variables)	Indicador. AUC	Fallido4	0,69				
Propio (20 variables)	Garantías hipotecarias	Fallido5	-1,21	0,61	-1,99	0,05	*
Propio (20 variables)	Tipo interés	Fallido5	-14,73	2,47	-5,97	0,00	***
Propio (20 variables)	Indicador. McFadden	Fallido5	0,23				
Propio (20 variables)	Indicador. Predicción	Fallido5	0,97				
Propio (20 variables)	Indicador. AUC	Fallido5	0,68				
Propio (20 variables)	Población	Fallido+5	0,00	0,00	-2,96	0,00	**
Propio (20 variables)	Empleos temporales	Fallido+5	0,03	0,01	3,29	0,00	**
Propio (20 variables)	Avales técnicos	Fallido+5	-0,50	0,25	-1,98	0,05	*
Propio (20 variables)	Garantías hipotecarias	Fallido+5	-1,81	0,74	-2,46	0,01	*
Propio (20 variables)	Tipo interés	Fallido+5	-13,55	2,53	-5,35	0,00	***
Propio (20 variables)	Indicador. McFadden	Fallido+5	0,25				
Propio (20 variables)	Indicador. Predicción	Fallido+5	0,98				
Propio (20 variables)	Indicador. AUC	Fallido+5	0,75				

8.3 CÓDIGO R

```
#Le cambiamos el nombre al archivo y cubrimos las celdas en blanco con NAs
training.data.raw <-
read.csv2('entrenamiento_Variables.Significativas.Literatura_fallido_v4.csv',hea
der=T, sep=";", dec = ",", na.strings=c(""))
#Cuántos NAs tiene cada variable
sapply(training.data.raw,function(x) sum(is.na(x)))
#Nos proporciona los diferentes valores que toma cada variable
sapply(training.data.raw, function(x) length(unique(x)))
#Activamos el paquete Amelia
library(Amelia)
#Seleccionamos la variable que vamos a utilizar la estimación del modelo
data <- subset(training.data.raw, select= c(2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14,
15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32))
#Separar la muestra - Stratified Random Sampling (R Pubs)
library(knitr)
# Function 1: Create function to calculate percent distribution for factors
pct <- function(x){
  tbl <- table(x)
  tbl_pct <- cbind(tbl,round(prop.table(tbl)*100,2))
  colnames(tbl_pct) <- c('Count','Percentage')
  kable(tbl_pct)
  }
library(caret)
pct(data$Fallido)
div_part_1 <- createDataPartition(y = data$Fallido, p = 0.7, list = F)
# Parte para correr el modelo -> entrenamiento
entrenamiento <- data[div_part_1,] # 70% here
pct(entrenamiento$Fallido)
# Parte para comprobar el modelo estimado -> test
test <- data[-div_part_1,] # rest of the 30% data goes here
pct(test$Fallido)
save(entrenamiento, file="entrenamiento.RData")
```

```
save(test, file="test.RData")
str(entrenamiento)
str(test)
#Estimación del modelo
model <- glm(Fallido ~.,family=binomial(link='logit'),data=entrenamiento)
warnings(model)
summary(model)

model <- glm(Fallido ~.,family=binomial(link='probit'),data=entrenamiento)
warnings(model)
summary(model)
#Estimación con toda la base de datos
pred.model <- glm(Fallido ~.,family=binomial(link = 'logit'),data=data)
summary(pred.model)

pred.model <- glm(Fallido ~.,family=binomial(link = 'probit'),data=data)
summary(pred.model)
#Contraste de significación - Wald (forma 1)
summary(model)$coefficients[,3]
#Variables no significativas por Wald: AT, MB_AT, EBIT_AT, OENEG, Poblac,
AC_PC, Inv_Circ, QA_AC, Antg, E.Fij, E.Ind, I.Ind, I.Energ, Liq, DUR, G.Per, G.Pig
#Excluimos variables no significativas por Wald.
snd.model.logit <- subset(entrenamiento, select = c(2, 4, 6, 7, 8, 14, 15, 19, 23,
24, 26, 29, 30, 31))

snd.model.probit <- subset(entrenamiento, select = c(2, 4, 6, 7, 8, 13, 14, 15,
19, 23, 24, 25, 29, 30, 31))
#Seleccionamos solo las variables cuyo |z value| no es superior a 1,96
no.sign.var.zvalue <- cbind(summary(model)$coefficients[2,3],
summary(model)$coefficients[4,3], summary(model)$coefficients[6,3],
summary(model)$coefficients[10,3], summary(model)$coefficients[11,3],
summary(model)$coefficients[12,3], summary(model)$coefficients[13,3],
summary(model)$coefficients[14,3], summary(model)$coefficients[17,3],
summary(model)$coefficients[18,3], summary(model)$coefficients[19,3],
```

```
summary(model)$coefficients[21,3], summary(model)$coefficients[22,3],
summary(model)$coefficients[23,3], summary(model)$coefficients[26,3],
summary(model)$coefficients[28,3], summary(model)$coefficients[29,3])

colnames(no.sign.var.zvalue) <- c("AT", "MB_AT", "EBIT_AT", "OENEG",
"Poblac", "AC_PC", "Inv_Circ", "QA_AC", "Antg", "E.Fij", "E.Ind", "I.Ind",
"I.Energ", "Liq", "Dur", "G.Per", "G.Pig")

no.sign.var.zvalue

#Contraste de significación conjunta

modelo.cte <- glm(Fallido ~ 1, data = entrenamiento, family = binomial (link =
'probit'))

anova(modelo.cte, model, test = "Chisq")

#SI da significativo 2 es que se rechaza la hipótesis nula de que ambos
parámetros son 0.

#Contraste de signficación - IC exponencial Wald

exp(confint.default(model))

#El contraste es que sea = 1 por lo tanto, si contiene 1 el IC tenemos un
problema

#Contraste de colinealidad

#Matriz de autocorrelación

library(MASS)

library(ggplot2)

library(ggcorrplot)

library(GGally)

ggcorr(snd.model[,-1], geom = "blank", label = TRUE, hjust =
0.75,label_round=2) +  geom_point(size = 10, aes(color = coefficient > 0, alpha =
abs(coefficient) > 0.6)) +   scale_alpha_manual(values = c("TRUE" = 0.25,
"FALSE" = 0)) +   guides(color = FALSE, alpha = FALSE)

#VIF

library(car)

vif(model)

#Variables no significativas por Wald: AT, MB_AT, EBIT_AT, OENEG, Poblac,
AC_PC, Inv_Circ,

#QA_AC, Antg, E.Fij, E.Ind, I.Ind, I.Energ, Liq, DUR, G.Per, G.Pig

#Stepwise

modelo.full <- glm(Fallido~., data = snd.model, family = binomial (link = 'logit'))
```

```
modelo.inicial <- glm(Fallido ~ 1, data = snd.model, family = binomial(link = 'logit'))


modelo.full <- glm(Fallido~., data = snd.model, family = binomial (link = 'probit'))
modelo.inicial <- glm(Fallido ~ 1, data = snd.model, family = binomial(link = 'probit'))


model <- stepAIC(modelo.inicial, scope = list(upper = modelo.full), direction = "both")
summary(model)


stepwisedata <- subset(snd.model, select= c(1, 2, 3, 4, 5, 6, 7, 9, 10, 11, 12, 13, 14))
#Análisis de los residuos - Pearson
res.p <- residuals(model, type = "pearson")
head(res.p)
summary(res.p)
plot(res.p, cex = 0.5)
abline(h = c(-2, 2), col = "red")
#Serán significativos aquellos cuyo valor absoluto sea mayor de 2.
res.p.sig <- abs(res.p) > 2
table(res.p.sig)
plot(res.p.sig, cex = 0.9)
abline(h = c(-2, 2), col = "red")
#Análisis de los residuos - Devianza
res.d <- residuals(model, type = "deviance")
summary(res.d)
plot(res.d, cex = 0.5)
abline(h = c(-2, 2), col = "red")
res.d.sig <- abs(res.d) > 2
table(res.d.sig)
#Sólo cogemos los que superan +2 o -2
```

```
#Y lo representamos por gráfico
signif <- which(abs(res.d) > 2)
plot(res.d[signif], type = "n")
text(1:length(signif), res.d[signif], label = signif, cex = 0.4)
#O histograma
hist(res.d, breaks = 40)
abline(v = quantile(res.d, probs = c(0.05/2, 1 - 0.05/2)), lty = 2)
plot(fitted.values(model), res.d, xlab = "Prob.predichas", ylab = "Residuos")
plot(data$Fallido, fitted.values(pred.model), xlab = "Valores observados",
    ylab = "Valores predichos")
#Representa los residuos (por defecto los de Pearson, pero se pueden cambiar) frente a las variables predictoras, y también frente a las transformaciones logit
residualPlots(model, type = "deviance", cex = 0.6)#Un p-valor alto que indica que no hay falta de ajuste
#Valores atípicos - Distancia de Cook y HatValores
distancias.cook <- cooks.distance(model)
hat.valores <- hatvalues(model)
table(distancias.cook > 1)


plot(distancias.cook)


par(mfrow = c(2, 2))
plot(model, cex = 0.6) #Interesante el cuarto gráfico es el más útil para detectar valores influyentes, ya que compara los residuos estandarizados de Pearson con los hat values y además muestra líneas de contorno para las distancias de cook.
#Contraste basado en el estadístico de Hosmer-Lemeshow
#Function
hosmerlem <- function(y, yhat, g = 10) {
   cutyhat1 = cut(yhat, breaks = quantile(yhat, probs = seq(0, 1, 1/g)),
 include.lowest = TRUE)
   obs = xtabs(cbind(1 - y, y) ~ cutyhat1)
   expect = xtabs(cbind(1 - yhat, yhat) ~ cutyhat1)
```

```
  chisq.C = sum((obs - expect)^2/expect)
  P.C = 1 - pchisq(chisq.C, g - 2)
  cutyhat2 = cut(yhat, breaks = g, include.lowest = TRUE)
  obs = xtabs(cbind(1 - y, y) ~ cutyhat2)
  expect = xtabs(cbind(1 - yhat, yhat) ~ cutyhat2)
  chisq.H = sum((obs - expect)^2/expect)
  P.H = 1 - pchisq(chisq.H, g - 2)
  res <- data.frame(c(chisq.C, P.C), c(chisq.H, P.H))
  colnames(res) <- c("Hosmer-Lemeshow C statistic", "Hosmer-Lemeshow H
statistic")
  rownames(res) <- c("X-squared", "p.value")
  return(res)
}

yhat <- fitted.values(model)
y <- snd.model$Fallido

hosmerlem(entrenamiento$Fallido, fitted.values(model))
#Error tipo I / Error tipo II
table(entrenamiento$Fallido)

prediccion <- ifelse(fitted.values(model) >= 0.5,1,0)
table(prediccion)

table(entrenamiento$Fallido, prediccion)
#Tasa de clasificaciones correctas
tabla.clasif <- table(entrenamiento$Fallido, prediccion)
tcc <- 100 * sum(diag(tabla.clasif))/sum(tabla.clasif)
tcc

library(ROCR)
```

```
pred.logit <- prediction(fitted.values(model), stepwisedata$Fallido)
pred.probit <- prediction(fitted.values(model), stepwisedata$Fallido)

perf1.probit <- performance(pred.probit, measure = "acc")
perf1.logit <- performance(pred.logit, measure = "acc")
# el punto de corte que maximiza 'acc' es
(posicion.max <- sapply(perf1@y.values, which.max))
(punto.corte <- sapply(perf1@x.values, "[", posicion.max))
plot(perf1, col = "darkred")
# Añadimos una línea horizontal al valor de 0.8
abline(h = 0.8, lty = 2)
# Añadimos recta con el punto de corte que maximiza la tasa de
# clasificaciones correctas
abline(v = punto.corte, lty = 2)
# auc : Area under curve
AUC.logit <- performance(pred.logit, "auc")
AUC.logit@y.name
AUC.logit@y.values

AUC.probit <- performance(pred.probit, "auc")
AUC.probit@y.name
AUC.probit@y.values
# con performance se selecciona tpr (true positive rate) y fpr (false
# positive rate)
perf2.logit <- performance(pred.logit, "tpr", "fpr")
plot(perf2.logit, colorize = TRUE) # mostramos colores según el punto de corte

perf2.probit <- performance(pred.probit, "tpr", "fpr")
plot(perf2.probit, colorize = TRUE) # mostramos colores según el punto de
corte
# Añadimos la recta y=x que sería la correspondiente al peor clasificador
abline(a = 0, b = 1)
```

```
# añadimos el valor del área bajo la curva
text(0.4, 0.6, paste(AUC.logit@y.name, "\n", round(unlist(AUC.logit@y.values), 3)), cex = 0.7)
#Comparar gráficos ROC logit-probit
plot(perf2.logit, col='blue', lty=1, main='ROCs: Model Performance Comparision') # logistic regression
plot(perf2.probit, col='gold',lty=2, add=TRUE); # probit regression
 legend(0.6,0.5,
       c('logistic reg','probit reg'),
       col=c('blue','gold'),
       lwd=3)
#Kolgomorov-Smirnov test.
y1 = rnorm(stepwisedata$Fallido, mean = 0, sd = 1)
ks.test(y1,"pnorm")
#AUROCC
AUROC <- round(performance(pred, measure = "auc")@y.values[[1]]*100, 2)
#McFadden
library(pscl)
pR2 <- pR2(model)
pR2 [4]
```