Universidad de Buenos Aires
Facultad de Ciencias Económicas
Escuela de Estudios de Posgrado

MAESTRÍA EN GESTIÓN ECONÓMICA Y FINANCIERA DE RIESGOS

TRABAJO FINAL DE MAESTRÍA

Implementación de *Text Mining* en R para predecir el riesgo de *default* corporativo

AUTOR: FLAVIA MUNAFO

DIRECTOR: MAURO SPERANZA

AGOSTO 2019

Resumen

El tema de la presente tesis se encuentra centrado en torno al riesgo de crédito, el cual constituye el principal riesgo del sector bancario y es uno de los temas claves para determinar la estabilidad financiera de una empresa. El riesgo de crédito se divide en tres tipos: el spread de crédito, la cesación de pagos o probabilidad de *default* y el descenso en la calificación crediticia. El trabajo se centrará en estos dos últimos por ser los de mayor relevancia para las instituciones bancarias en la decisión de otorgamiento de créditos.

Tradicionalmente los bancos se basan en modelos econométricos de calificación crediticia para tomar decisiones vinculadas al otorgamiento de préstamos. Sin embargo, cabe destacar que la deficiencia primordial que presentan dichos modelos de riesgo de crédito utilizados en la práctica radica en que están basados en información contable y financiera que se emite de manera trimestral o semestral. En este contexto, se plantea la pregunta problematizante de la tesis: ¿podría cubrirse esta deficiencia incorporando a los modelos de *credit scoring*[1] convencionales un coeficiente adicional que mida la información diaria contenida en la red social Twitter sobre el desempeño de la empresa? ¿La incorporación de este coeficiente, mejoraría la estimación de la probabilidad de *default*?

Para responder al objetivo propuesto, en el presente trabajo se calcula la probabilidad de *default* de empresas que cotizan en bolsa a través de un modelo de regresión logística que incorpora fuentes alternativas de datos (*text mining*[2]). Para ello, en primer lugar, se calibra un modelo de regresión logística convencional con el objetivo de estimar la probabilidad de *default*. Luego se desarrolla un índice de sentimiento que mide cuantitativamente la información contenida en Twitter y se vuelve a calibrar el modelo anterior incorporando el nuevo coeficiente. Finalmente se realiza una comparación entre ambos modelos y se describe la importancia de la utilización de *big data*[3] en los modelos de riesgo de crédito.

Palabras claves: Riesgo de crédito, *Default, Big data*, *Text mining*, Análisis de sentimiento, R Studio, Twitter

[1] Puntuación de crédito, según su traducción al español.

[2] Minería de texto, según su traducción al español.

[3] Grandes volúmenes de datos, según su traducción al español.

Índice

Introducción

El riesgo es entendido como la probabilidad de ocurrencia de un evento adverso y sus consecuencias. El riesgo financiero se refiere a la probabilidad de ocurrencia de que un evento tenga consecuencias negativas para una organización, originado dentro de las actividades que genera para su normal funcionamiento. En particular, el riesgo de crédito es definido como la pérdida potencial provocada por el incumplimiento de la contraparte en una operación que incluye un compromiso de pago. El mismo constituye el principal riesgo del sector bancario y es uno de los temas claves para determinar la estabilidad financiera de una empresa o de una persona física.

Una definición alternativa plantea dos formas de manifestación del riesgo de crédito y lo describe como la incertidumbre derivada de la probabilidad de sufrir quebrantos por el incumplimiento de alguna o de todas las obligaciones contractuales de la contraparte en una operación financiera, ya sea por la entrada en mora del deudor, provocada por el retraso en el cumplimiento, o por el impago definitivo de las obligaciones, lo que deviene en la insolvencia del mismo.

Los enfoques actuales de las instituciones financieras para un modelo de riesgo de crédito se centran en la estimación de parámetros claves requeridos por el Segundo Acuerdo de Basilea (Medina, 2007), estos son; la probabilidad de *default*, definida como la probabilidad de que una empresa entre en *default* en el período de un año, la pérdida dada por el *default* o *loss given default*[4] entendido como la cantidad de dinero que un banco u otra institución financiera pierde cuando un prestatario deja de pagar un préstamo y la exposición en el momento del incumplimiento o *exposure at default*[5] definida como el importe de deuda pendiente de pago en el momento de incumplimiento del cliente. Los bancos pueden calcular su carga de capital regulatorio para el riesgo de crédito sobre la base de estas estimaciones.

Un elemento importante en el riesgo de crédito es el evento de *default*. El mismo es definido por el Banco Central de la República Argentina (BCRA) como un evento que ocurre cuando el deudor tarda más de 90 días en realizar sus pagos o es poco probable que pague una obligación. Mientras que la primera parte de la definición puede o no implicar pérdidas, ya que el deudor puede pagar todas sus deudas transcurridos los 90 días, la segunda parte implica un juicio subjetivo que puede resultar correcto o no.

Los métodos o modelos de *credit scoring*, a veces denominados *score-cards* [6] o *classifiers*[7], son algoritmos que de manera automática evalúan el riesgo de crédito de un solicitante de financiamiento o alguien que ya es cliente de una entidad evaluadora entre las clases de riesgo “bueno” y “malo” en base a su probabilidad de *default* (Hand & Henley, 1997). Tienen una dimensión individual, ya que se enfocan en el riesgo de

[4] Pérdida crediticia esperada, según su traducción al español.
[5] Exposición al default, según su traducción al español.
[6] Tarjetas de calificación, según su traducción al español
[7] Clasificadores, según su traducción al español.

incumplimiento del individuo o empresa, independientemente de lo que ocurra con el resto de la cartera de préstamos.

La utilización de modelos de *credit scoring* para la evaluación del riesgo de crédito, es decir, para estimar probabilidades de *default* y ordenar a los deudores y solicitantes de financiamiento en función de su riesgo de incumplimiento, comenzó en los años 70 pero se generalizó a partir de los 90. Esto se ha debido tanto al desarrollo de mejores recursos estadísticos y computacionales, como por la creciente necesidad por parte de la industria bancaria de hacer más eficaz y eficiente la originación de financiaciones y de tener una mejor evaluación del riesgo de su portafolio.

Para modelar la probabilidad de *default* de las empresas existe una gran cantidad de modelos, que pueden clasificarse en dos grandes categorías: los modelos estructurales y los modelos de forma reducida. A su vez estos modelos se clasifican en paramétricos y no paramétricos. La diferencia radica en que los modelos paramétricos parten de una función de distribución conocida y reducen el problema a estimar los parámetros que mejor la definen, mientras que, por el contrario, los modelos no paramétricos no se encuentran sujetos a ninguna forma funcional por lo que el problema consiste en calcular los parámetros de una función estimada y no de una función conocida.

El modelo de Merton como en Black & Scholes (1973, 1974) constituye el modelo estructural más representativo. En 1995, la agencia de calificación crediticia, Moody´s, lo incorporó a su trabajo alterando algunos conceptos y lo recategorizó como modelo Merton-KMV. En este modelo, la línea de crédito se considera como un pasivo contingente en el valor de los activos de la firma y se valúa de acuerdo con la teoría de las opciones financieras. Se establece así que la empresa alcanzará el *default* cuando el valor de mercado de los activos de la compañía sea menor que el valor de sus pasivos. Por este motivo, el modelo también fue llamado modelo basado en el valor de la empresa (Scandizzo, 2016).

Vasicek (1977) y Shimko (1993) utilizaron tasas de interés estocásticas para evaluar el precio de los bonos en dicho modelo. Longstaff and Schwartz (1995) y Hui et al. (2003) realizaron ciertas modificaciones al modelo original de Merton. Sin embargo, además de los factores internos de la empresa, hay muchos otros factores de distinta índole que podrían causar el incumplimiento corporativo. Los factores externos del medio ambiente han hecho que gradualmente los modelos estructurales se hicieran menos populares. En este contexto, los modelos de forma reducida también conocidos como modelos de intensidad, se encargan de explorar el vínculo entre el incumplimiento corporativo y las variables explicativas.

Existe una cantidad infinita de combinaciones de factores de riesgo y metodologías de puntuación que pueden utilizarse para calcular la probabilidad de *default* en los modelos de forma reducida, pero la mayoría se basa en los mismos tipos de factores, financieros (como los ratios de las hojas de balance e indicadores financieros), no financieros (como la capacidad de gestión y la flexibilidad financiera) así como factores de comportamiento (como el estado de morosidad y la utilización del crédito).

En este contexto, cada modelo de forma reducida incorpora en su estructura distintos ratios para predecir la probabilidad de *default* corporativo. Altman (1968), Ohlson (1980) y Zmijewski (1984) utilizaron de tres a nueve ratios financieros. Shumway (2001) incluyó dos ratios financieros y tres variables de mercado. Chava y Jarrow (2004) agregaron variables industriales a los ratios de Altman (E. I. Altman, 1968) y Zmijewski (1984). Lee y Yeh (2004) se centró en la relación entre el gobierno corporativo y la dificultad financiera. Duffie et al. (2007) agregaron variables macroeconómicas al modelo de intensidad dinámica. Campbell, Hilscher, & Szilagyi (2008) agregaron dos ratios financieros específicos de la empresa y el retorno de las acciones a la lista de variables compiladas por Shumway. (2001). Finalmente, Standard & Poor considera dieciocho variables que explican liquidez, términos de rentabilidad, estructura de capital, flujo de caja y capacidad de pago de interés, etc. en la calificación crediticia de la empresa.

El primer modelo de forma reducida fue propuesto por Jarrow & Turnbull (1995) donde el *default* se modeló como el momento en el que ocurre el primer salto en un proceso de Poisson con una intensidad *random walk*[8]. Luego, se desarrollaron una gran cantidad de modelos relacionados, basados en técnicas estadísticas, matemáticas, econométricas y de inteligencia artificial. En este contexto existe una gran cantidad de técnicas disponibles incluido el análisis regresión múltiple (R. West, 1970), regresión lineal, el análisis discriminante multivariante, el modelo *Z score*[9] (E. I. Altman, 1968), modelos de regresión logística (Ohlson, 1980), modelos *probit* (Zmijewski, 1984), modelos de orden de probabilidad (Blume, Lim, & MacKinlay, 1998; Gentry, Newbold, & Whitford, 1985; Güttler & Wahrenburg, 2007), el modelo fijo de riesgos proporcionales (Bharath & Shumway, 2008; Cox, 1972; Lane, Looney, & Wansley, 1986), modelos de riesgo de tiempo discreto (Chava & Jarrow, 2004; Shumway, 2001), modelos de matrices de transición (Lando & Skodeberg, 2002), modelos de intensidad de *default* dinámica (Duffie et al., 2007), algoritmos de particionamiento recursivo como los árboles de decisión (E. I. Altman, 1968; Beaver, 1966), algoritmos genéticos, redes neuronales y finalmente el juicio humano es decir, la decisión de un analista acerca de otorgar un crédito (Mester Loretta, 1997; Srinivasan & Kim, 1987).

A pesar de la proliferación de los modelos de *credit scoring*, el juicio del analista continúa siendo utilizado en la originación de créditos, en algunos casos expresado como un conjunto de reglas que la entidad aplica de manera sistemática para filtrar solicitudes o deudores. De hecho, en la práctica ambas metodologías muchas veces coexisten y se complementan entre sí, definiendo sistemas híbridos.

Aunque esta última presenta la ventaja de ser más eficaz, los métodos de *credit scoring* son más eficientes a la vez que sus predicciones más objetivas y consistentes, por lo que pueden analizar y tomar decisiones sobre una gran cantidad de solicitudes de crédito en poco tiempo y a un bajo costo. La literatura sugiere que todos los métodos de *credit scoring* arrojan resultados similares, por lo que la conveniencia de usar uno u otro depende de las características particulares del caso de estudio. Dentro de los enfoques econométricos, los modelos de probabilidad lineal han caído en desuso por sus

[8] Paseo aleatorio, según su traducción al español.
[9] Puntuación Z, según su traducción al español.

desventajas técnicas, en tanto que los modelos de variables econométricas discretas como los *probit*, *logit* y la regresión logística son superiores al análisis discriminante ya que proveen para cada deudor una probabilidad de *default*, en tanto que este sólo clasifica a los deudores en grupos de riesgo.

Los modelos no paramétricos y los de inteligencia artificial, como por ejemplo los árboles de clasificación o decisión, las redes neuronales y los algoritmos genéticos, son superiores a los modelos estadísticos cuando se desconoce la forma probable de la relación funcional y se presume que esta no es lineal. Los árboles tienen la ventaja de no requerir la formulación de supuestos estadísticos sobre distribuciones estadísticas o formas funcionales. A su vez, presentan la relación entre las variables, los grupos y el riesgo de manera visual, con lo cual, si el conjunto de variables en el análisis es reducido, facilita entender cómo funciona el *scoring*. Las redes neuronales y los algoritmos genéticos, a pesar de las ventajas mencionadas, son poco intuitivos y de difícil implementación. Srinivasan y Kim (1987) compararon diversas técnicas y encontraron que los árboles de decisión superan a las regresiones logísticas, mientras que éstas arrojan mejores resultados que el análisis discriminante.

Entre todas las metodologías disponibles, la revisión de la literatura muestra que entre los métodos más usados en la industria para la confección de estos modelos predominan los enfoques econométricos, tales como los modelos *probit*, junto con las regresiones lineales y logísticas. Los motivos para su predominio son básicamente dos: en general las metodologías relevadas muestran resultados similares, por lo que tienden a emplearse aquellas cuyo funcionamiento e interpretación son más sencillos, en contraposición a enfoques más sofisticados y de difícil interpretación, como ser las redes neuronales (Gutiérrez Girault, 2007).

La mayoría de los bancos se basan en algunos de los modelos econométricos de calificación crediticia mencionados anteriormente para tomar decisiones vinculadas al otorgamiento de préstamos. Sin embargo, cabe destacar que estos modelos presentan ciertas deficiencias, ya que se basan en informes financieros formales de los prestatarios. Por su parte, la información sobre las ganancias de la corporación y otra información vinculada a la contabilidad o las finanzas de la empresa se publican de manera trimestral o mensualmente, pero el mercado bursátil funciona de manera diaria, y a menudo se encuentra fuertemente influenciado por las noticias del día a día, donde el precio de cierre no solo refleja las condiciones operativas de la corporación sino también la información diaria del mercado.

En este contexto, la información de carácter no convencional – como la textual – puede ayudar a los bancos a superar algunos de estos desafíos y mejorar su evaluación del riesgo crediticio, en particular su enfoque de evaluación cualitativa. Esta información incluye contenido producido profesionalmente, como informes de analistas y periodismo empresarial, así como textos informales, como blogs y publicaciones en redes sociales. Los artículos de noticias describen los últimos desarrollos de las empresas; los informes de los analistas ofrecen análisis profundos sobre las estrategias de las empresas, el posicionamiento competitivo y las perspectivas; las clasificaciones de productos en los sitios de compras en línea ofrecen vistas sin filtro de la satisfacción

del cliente; y redes sociales como *Twitter* distribuyen las últimas noticias y comentarios de clientes en tiempo real.

Diversos autores (Braun, Nelson, & Sunier, 1995; K. C. Brown, Harlow, & Tinic, 1988; Coval & Shumway, 2001; Pandher & Currie, 2013) han abordado esta problemática desde diferentes ángulos. Tetlock (2007) estudió el impacto que tienen los medios como el *Wall Street Journal* en los inversores y encontraron impactos significativos de las noticias negativas en el volumen de negociación de acciones. Tetlock et al. (2008) muestra que la información negativa afectará los ingresos corporativos y se puede utilizar como un predictor importante de las devoluciones de acciones y los ingresos corporativos. Antweiler y Frank (2004) estudiaron el impacto de las noticias web en el mercado de valores. Sin embargo, resulta difícil evaluar los impactos compuestos de noticias de diferentes fuentes ya que sus características básicas pueden ser diferentes unas de otras.

Para resolver este inconveniente en los modelos de calificación, muchos autores han incorporado a los modelos de calificación una técnica denominada análisis de sentimiento como un factor de calificación adicional donde a la información obtenida de las búsquedas de texto se agrega trimestralmente un índice de sentimiento que representa un tipo y grado de opinión expresada por el escritor como optimismo o pesimismo. Después del análisis estadístico, el índice se integra en el sistema de calificación con un peso adecuado. Esto puede ser particularmente valioso en la evaluación de nuevos clientes corporativos para los cuales los bancos suelen tener solo información limitada. Una proyección sistemática de información pública puede revelar información adicional importante que puede tener un peso significativo en la calificación.

En este contexto el objetivo de la tesis es incorporar en un modelo de regresión logística, fuentes alternativas de datos (*text mining*) para calcular la probabilidad de *default* de empresas que cotizan en la bolsa. Para el logro de dicho objetivo en primer lugar, en el Capítulo 1, se realizará un breve recorrido por los modelos más representativos de riesgo de crédito utilizados en el sector corporativo. Luego en el Capítulo 2, se introducirá el concepto de *text mining* en la era del *big data*, en particular aplicado a los modelos de riesgo de crédito y su importancia para el análisis de sentimiento. El concepto será introducido bajo el paraguas de la responsabilidad social, incluyendo sus potenciales riesgos y beneficios. Finalmente, en el Capítulo 3, se desarrollarán dos modelos utilizando la herramienta *R Studio*. En una primera instancia se calibrará un modelo de forma reducida utilizando una regresión logística y en una segunda instancia se incorporará una variable adicional al modelo que mida el análisis de sentimiento de la red social Twitter y se sacarán conclusiones al respecto. Finalmente, se expondrán en las conclusiones, los resultados a los que se ha arribado a lo largo del desarrollo de este trabajo, junto con las futuras líneas de investigación que se han abierto durante el mismo.

1.1 Planteamiento del problema

Para modelar la probabilidad de *default* de las empresas existe una gran cantidad de modelos y se pueden dividir en dos grandes categorías: por un lado, los modelos estructurales y por otro los modelos de forma reducida. Entre todas las metodologías disponibles, la revisión de la literatura muestra que entre los métodos más usados en la industria para la confección de estos modelos predominan los enfoques econométricos de forma reducida, tales como los modelos *probit*, junto con las regresiones lineal y logística, el análisis discriminante y los árboles de decisión. Según el análisis de la literatura, los motivos de su predominio se deben a que tienden a utilizarse los modelos cuya interpretación y funcionamiento son más sencillos en contraposición a enfoques más sofisticados. Por este motivo se seleccionaron los modelos de regresión logística.

La mayoría de los bancos se basan en alguno de los modelos econométricos de calificación crediticia mencionados anteriormente para tomar decisiones vinculadas al otorgamiento de préstamos. Sin embargo, cabe destacar que la deficiencia primordial que presentan los modelos de riesgo de crédito utilizados radica en que están basados en información contable y financiera que se emite de manera trimestral o semestral. En este contexto, se plantea la pregunta problematizante ¿podría cubrirse esta deficiencia incorporando a los modelos de *credit scoring* convencionales un coeficiente adicional que mida la información diaria contenida en la red social Twitter sobre el desempeño de la empresa? ¿La incorporación de este coeficiente, mejoraría la estimación de la probabilidad de *defaul*t?

1.2 Objetivos

Objetivo general: Incorporar en un modelo de regresión logística fuentes alternativas de datos (*text mining*) para calcular la probabilidad de *default* de empresas que cotizan en la bolsa.

Objetivo específico 1: Calibrar un modelo de regresión logística para calcular la probabilidad de *default* de empresas que cotizan en la bolsa.

Objetivo específico 2: Desarrollar un índice que mida cuantitativamente con una probabilidad de 0-1 la información contenida en *Twitter* y volver a calibrar el modelo incorporando el nuevo coeficiente.

Objetivo específico 3: Realizar una comparación entre ambos modelos y describir los aportes de utilizar *text mining* en modelos de riesgo de crédito.

1.3 Hipótesis

La incorporación de un índice de sentimiento cualitativo en los modelos de riesgo de crédito tradicionales (cuantitativos) mejoraría la estimación de la probabilidad de *default* ya que incorporaría información no cuantificable sobre el funcionamiento del mercado bursátil.

CAPÍTULO 1: Modelos de riesgo de crédito

1.1 Introducción a los modelos de riesgo de crédito

Los modelos de riesgo de crédito se clasifican en dos tipos principales: los modelos estructurales y los de forma reducida, y a su vez estos se clasifican en paramétricos y no paramétricos. La diferencia entre esta última clasificación radica en que los modelos paramétricos parten de una función de distribución conocida y reducen el problema a estimar los parámetros que mejor la definen, por el contrario, los modelos no paramétricos no se encuentran sujetos a ninguna forma funcional por lo que el problema consiste en calcular los parámetros de una función estimada y no de una función conocida.

Mayormente, los modelos estructurales son también conocidos como modelos de valor de activos o modelos de opción teórica. Suponen que se dispone de información de mercado de las firmas, específicamente de información bursátil y consideran al riesgo predeterminado como una opción de venta europea sobre el valor del activo de la firma (Black and Scholes, 1973; Merton, 1974). El *default* es entendido en el momento en el cual los activos de la firma caen por debajo del nivel de las obligaciones al vencimiento. Se denominan modelos estructurales ya que el riesgo de crédito y sus componentes principales como la probabilidad de *default* y el *loss given default* son una función de las características estructurales de la empresa como el apalancamiento o la volatilidad de los activos. Entre los principales exponentes de los modelos estructurales se encuentra Merton (1974) y en segundo lugar Credit Metrics de JP Morgan (1997) y el modelo Credit Portafolio Manager de KMV Moody's (Crouhy, Galai y Mark, 2000).

Por otro lado, los modelos de forma reducida no intentan establecer una explicación causal del *default*, sino que más bien lo consideran como parte de un proceso aleatorio, mientras que la pérdida dada por el *loss given default* se especifica exógenamente. Los modelos de forma reducida están basados en datos y modelan la probabilidad de un evento (el *default*) por separado de la pérdida correspondiente y luego producen una medida agregada de riesgo mediante métodos numéricos. Los eventos de *default* se supone que ocurren inesperadamente debido a uno o más eventos exógenos, independientemente del valor del activo del prestatario.

Los factores de riesgo observables que se producen exógenamente incluyen cambios en los factores macroeconómicos, tales como el PBI, la tasa de interés, los tipos de cambio y la inflación mientras que los factores de riesgo inobservables pueden ser específicos para una empresa, industria o país.

Entre los modelos de forma reducida, se incluyen el modelo tradicional de (Altman, 1968), los modelos que se basan en *logit, probit* y lineales de probabilidad (Greene, 2000; Gujarati, 2003); los modelos que utilizan simulación (Dunkel & Weber, 2007; Ramaswamy, 2005); y los modelos no paramétricos de redes neuronales artificiales (Atiya, 2001) y metodología borrosa (Wei, 2008). Estos modelos suponen que se tiene información a priori de las empresas o de los créditos clasificados como cumplidos o

incumplidos en torno a las obligaciones que implica el crédito y que este incumplimiento se puede relacionar con variables internas y externas de sus clientes.

Siendo el principal motor de este trabajo, se observa que, en su forma general, los modelos de forma reducida suponen que existe una variable observable (Y_i^*) que representa la utilidad que obtiene un individuo *i* al incumplir un crédito. De esta forma, Y_i^* puede ser explicada a través de un conjunto de factores o variables de la siguiente forma:

$$Y_i^* = \beta' X_i + u_i \quad (1.1)$$

Donde: X_i es la matriz de variables explicativas asociadas, β es el vector de parámetros a estimar y u_i es el término de perturbación del modelo o el error. Y_i^* es una variable inobservable dicotómica, es decir que puede tomar el valor 1 en caso de que se haya incurrido en *default* y 0 en caso contrario. A los fines del modelo interesa estimar su valor esperado $E(Y_i^*/X_i) = \beta_i$ pero solo podemos calcular $E(Y_i/X_i)$, por lo tanto:

$$P_i = E(Y_i/X_i) = F(\beta'/X_i) \quad (1.2)$$

Siendo F una función de distribución acumulada de u_i.

En los próximos apartados se procederá a explicar con más detalle los modelos de *credit scoring* más representativos presentes en la literatura académica. En primer lugar, se introducen los modelos univariados. Luego se presentan los modelos de forma reducida, entre los que se destaca el modelo de Altman o Z Score y los modelos que utilizan regresiones logísticas. A continuación, se detallan los modelos estructurales, entre los que se destaca el modelo de Merton, el modelo de *Credit Portfolio Manager*[10] de *KMV*[11] *Moody's* y el modelo de *Credit Metrics*[12] *de JP Morgan*. Finalmente, se concluye el capítulo presentando los modelos no paramétricos como el Modelo de redes neuronales, *Support Vector Machine* y los árboles de decisión.

1.2 Análisis crediticio tradicional o modelos univariados

El modelo tradicional más conocido es el de las cinco "C" del crédito (capacidad de pago, el capital, el colateral, el carácter y las condiciones) también llamado modelo de experto, en el cual la decisión se deja en manos de un analista de crédito (experto) que analiza cinco factores claves para la toma de decisiones. Implícitamente, la experiencia de dicha persona, su juicio subjetivo y la evaluación de dichos factores constituyen elementos determinantes a la hora de otorgar o rechazar un crédito.

Aunque este modelo se ha visto influenciado por condiciones cambiantes en el entorno financiero y se ha llegado a sustituir por técnicas probabilísticas y estadísticas más

[10] Gestión de cartera de riesgo, según su traducción al español.
[11] Modelo de varianza, covarianza según su traducción al español.
[12] Metricas de riesgo, según su traducción al español.

sofisticadas, a continuación, se explicará en qué consiste ya que el criterio del analista continúa siendo utilizado como complemento de modelos más sofisticados y constituye la base de los mismos. A continuación, se procederá a explicar más en detalle las 5 C del crédito: la capacidad de pago, el capital, el colateral, el carácter y las condiciones.

La capacidad de pago del acreditado es el factor más importante en la decisión de un banco. Consiste en evaluar la habilidad y la experiencia en los negocios que tenga la persona o empresa, su administración y sus resultados prácticos. Para realizar esta valuación se toma en cuenta la antigüedad, el crecimiento de la empresa, sus canales de distribución, las actividades que desarrolla, la zona de influencia, el número de empleados, las sucursales, etc., ya que se requiere saber cómo pagará el préstamo y para ello se necesita determinar el flujo de efectivo del negocio; incluso necesitan el historial del crédito del dueño y sus deudas pasadas y presentes (tanto las personales como las comerciales).

El capital hace referencia a los valores invertidos en el negocio del acreditado, así como las obligaciones, es decir, implica realizar un análisis de la situación financiera de la empresa. Un estudio financiero detallado permite conocer completamente las posibilidades de pago, el flujo de ingresos y egresos, así como la capacidad de endeudamiento. El flujo de liquidez, la rotación de inventario, el tiempo promedio que tarde en pagar, etc. son algunas razones financieras importantes para este análisis.

El colateral son todos aquellos elementos con los que dispone el acreditado para garantizar el cumplimiento del pago en el crédito, es decir, las garantías o apoyos colaterales. Se evalúa a través de sus activos fijos, el valor económico y la calidad de estos, ya que en el análisis del crédito se establece que no deberá otorgarse un crédito sin tener prevista una segunda fuente de pago.

El carácter son las cualidades vinculadas a la honorabilidad y la solvencia moral que tiene el deudor para responder al crédito. Se busca información sobre sus hábitos de pago y comportamiento en operaciones crediticias pasadas y presentes, en relación con sus pagos. La valuación del carácter o solvencia moral de un cliente debe hacerse a partir de elementos contundentes, cuantificables y verificables, como: referencias comerciales de otros proveedores con los que tenga un crédito, reportes de buró de crédito, referencias bancarias y calificación crediticia entre otros.

Las condiciones son los factores exógenos que pueden afectar la marcha del negocio del acreditado, como las condiciones económicas y del sector o la situación política y económica de la región. Aunque dichos factores no están bajo el control del acreditado, se consideran en el análisis de créditos para prever sus posibles efectos.

El empleo de modelos de riesgo de crédito basados en el juicio de expertos no debe ser entendido de forma inequívoca para determinar la futura solvencia de una empresa ya que presenta ciertos inconvenientes y riesgos. Los dos problemas principales que se destacan son la consistencia y la subjetividad. Los factores subjetivos aplicados las cinco “C” por un experto pueden variar de acreditado a acreditado. Este hecho dificulta la comparación de rangos y la toma de decisiones, por lo que se aplican estándares

diferenciados por parte de analistas de crédito dentro de una misma institución a distintos tipos de acreedores (Saavedra García & Saavedra García, 2010). Debido a la subjetividad que presentan estos modelos, han sido desplazados por otras metodologías y técnicas más sofisticadas que incorporan el criterio del analista como un complemento en sus análisis.

Los modelos que se presentan a continuación se clasifican en modelos de forma reducida y modelos estructurales y no paramétricos y utilizan de manera implícita en su metodología el criterio del analista mencionado anteriormente.

1.3 Modelos de forma reducida

Los modelos de forma reducida consideran al evento de *default* como parte de un proceso aleatorio que ocurre inesperadamente debido a uno o más eventos exógenos, independientemente del valor del activo del prestatario. Los factores de riesgo observables que se producen exógenamente incluyen cambios en los factores macroeconómicos, tales como el PBI, la tasa de interés, los tipos de cambio y la inflación mientras que los factores de riesgo inobservables pueden ser específicos para una empresa, industria o país. A continuación, se expondrán los dos modelos de forma reducida más representativos, tal es el caso del modelo de Altman o *Z Score* y los modelos de regresión logística.

1.3.1 Modelo de Altman o Z score

El primero de los exponentes vinculado al riesgo de crédito es Altman (1968), cuyo modelo se basa en identificar una serie de variables definidas en función de ratios financieros de la compañía y a los cuales se les asigna diferentes pesos estadísticos para generar un *score*. Cabe destacar que las empresas que no obtienen el score necesario no son rechazadas directamente si no que se someten a un análisis más detallado.

Para la creación del modelo se construyen dos muestras, unas que hayan dado buenos resultados y otras que hayan dado malos para la compañía para desarrollar después una escala para cada variable relevante. Las variables deben maximizar la varianza entre los dos grupos (empresas con resultados favorables y desfavorables) y minimizar la varianza dentro de cada grupo. Luego, se seleccionan aquellas variables que tengan mayor poder explicativo de la probabilidad de impago. De esta manera se construye el puntaje score como una combinación lineal de cinco relaciones comerciales comunes ponderadas por sus coeficientes. Para formalizar su modelo, Altman (1968) propuso el siguiente postulado lineal:

$$Z = 1{,}2\,X_1 + 1{,}4\,X_2 + 3{,}3\,X_3 + 0{,}6\,X_4 + 1{,}0\,X_5 \qquad (1.3)$$

Siendo:

X_1 = capital de trabajo / activos totales. Mide los activos líquidos en relación con el tamaño de la empresa.

X_2 = ganancias retenidas / activos totales. Mide la rentabilidad que refleja la antigüedad y el poder adquisitivo de la empresa.

X_3 = ganancias antes de intereses e impuestos / activos totales. Mide la eficiencia operativa además de los factores fiscales y de apalancamiento. Reconoce que las ganancias operativas son importantes para la viabilidad a largo plazo.

X_4 = valor de mercado del patrimonio / valor contable de los pasivos totales. Agrega una dimensión de mercado que puede mostrar la fluctuación del precio de seguridad como una posible bandera roja.

X_4 = ventas / activos totales. Medida estándar para la rotación total de activos (varía mucho de una industria a otra).

Z= Índice o valor discriminante

Una vez obtenido el *score*, se hace un contraste de hipótesis para comprobar su validez. Se realiza bajo la hipótesis nula de que la empresa no quiebra y la alternativa de que sí lo haga. La probabilidad de quiebra de una empresa dependerá del resultado obtenido con la fórmula de Altman *Z Score*. Según el resultado, la empresa podrá encontrarse en la zona segura, zona gris o en la zona de peligro. Si el *Z Score* resulta superior a 2,99 se trata de la zona segura, si se encuentra entre 1,81 y 2,99 se trata de la zona gris y es probable que la empresa pueda quebrar en los próximos dos años y si el Z-score es inferior a 1,81 se trata de la zona de peligro por lo que la quiebra es inminente.

Sin embargo, se puede incurrir en el error de tipo I al considerar impaga una empresa que no es y el error de tipo II al considerar una empresa solvente como fallida. El primer error resulta mucho más importante para las empresas, por las pérdidas reales que acarrea, mientras que el segundo, sólo tiene un costo de oportunidad de lo que se pudo dejar de ganar.

En su prueba inicial, se encontró que el puntaje de Altman tenía una precisión del 72% en la predicción de quiebra con una antelación de dos años del evento, con un error de tipo II (falsos negativos) del 6% (Altman, 1968). En una serie de pruebas posteriores que cubrieron tres períodos durante los siguientes 31 años, hasta 1999, se encontró que el modelo tenía aproximadamente un 80%-90% de precisión en la predicción de quiebra un año antes del evento, con un error de tipo II de aproximadamente 15% - 20% (Altman, 2000).

Desde alrededor de 1985 en adelante, los puntajes Z obtuvieron una amplia aceptación por para ser utilizados en la evaluación de préstamos. El enfoque de la fórmula se ha utilizado en una amplia variedad de contextos y países. Las variaciones posteriores de Altman fueron diseñadas para ser aplicables a compañías privadas (Altman Z'-Score) y compañías no manufactureras (Altman Z-Score). Este modelo fue utilizado por *Standard & Poors* para el cálculo del rating de bonos. Más adelante, Altman (1993) extendió el modelo para empresas no cotizadas y para la calificación de deuda de países emergentes.

A partir de las observaciones de Johnson (1970), Altman llegó a comprender algunas limitaciones y defectos en su modelo; en particular en lo referente a su capacidad predictiva. Según Johnson existía poca capacidad de los modelos para llevar predicciones *ex ante*. En cambio, cuando las predicciones eran *ex post* el modelo de Altman tendía a ser más exacto y se podían identificar con claridad las causas del fracaso a partir de los estados financieros.

A su vez los ratios financieros del modelo predictivo de Altman tenían poca capacidad para captar y describir la dinámica del proceso de *default* empresarial. Tradicionalmente los ratios solo habían servido para realizar análisis comparativos estáticos, fue el mismo Altman quien reconoció el problema del dinamismo. A modo de conclusión, con el surgimiento de modelos no lineales el modelo de Altman entró en desuso, sin embargo, sus ratios financieros fueron incluidos en modelos no estructurales como los modelos de regresión logística que se detallan a continuación.

1.3.2 Modelos de regresión logística

En esta sección, se mostrará cómo especificar un modelo de puntuación utilizando una técnica estadística denominada regresión logística o simplemente *logit*. El mismo equivale a codificar información en un valor específico (por ejemplo, medir el apalancamiento como deuda /activos) y luego encontrar la combinación de factores que hacen el mejor trabajo para explicar el comportamiento predeterminado histórico.

Introduciendo la regresión logística, un score resume la información contenida en los factores que afectan a la probabilidad de *default*. Un modelo estándar de regresión logística es la aproximación de la combinación lineal de las variables exógenas. El vector X denota los factores o variables explicativas (hasta el k-ésimo) y el vector b constituye el peso de los coeficientes unidos a ellos. De esta forma se puede representar el score que se obtiene en la instancia i como:

$$Score_i = b_1 x_{i1} + b_2 x_{i2} + \cdots + b_k x_{ik} = b' x_i \quad (1.4)$$

Esta expresión puede simplificarse colocando los b_s y x_s en un vector columna para b y el vector columna para x. El mismo se describe de la siguiente manera:

$$x_i = \begin{bmatrix} x_{i1} \\ x_{i2} \\ \dots \\ x_{iK} \end{bmatrix} \; b = \begin{bmatrix} b_1 \\ b_2 \\ \dots \\ b_K \end{bmatrix} \quad (1.5)$$

Si el modelo incluye a la constante b_1 entonces $x_{i1} = 1$ para ese i.

Asumimos por simplicidad que se elegirán los factores X por lo que queda determinar el peso del vector b asociado. Usualmente, el mismo se estima basándose en el comportamiento del *default* observado. Para ello se supone que se recoleta la información anual de empresas con factores y un comportamiento de *default* determinado. La información referida al default se almacena en la variable y_i. La misma

tomará el valor 1 si la empresa *defaultea* en el año siguiente para el que se han recopilado los valores del factor y 0 en caso contrario. El número total de observaciones se denota con la letra N.

El modelo debe predecir una alta probabilidad de impago para aquellas observaciones que *defaultean* y una baja probabilidad de impago para aquellas que no lo están. Para elegir los pesos apropiados de los coeficientes b se deben vincular los scores con las probabilidades de incumplimiento. Esto puede hacerse representando las probabilidades de incumplimiento como una función *F* de scores:

$$Prob\ (Default_i) = F(Score_i) \qquad (1.6)$$

Al igual que las probabilidades de *default*, la función *F* debe estar restringida en el intervalo 0 a 1, esto crea un campo de probabilidad de impago para cada score. Este hecho podría ser cumplido con una función de distribución de probabilidad acumulada. La función de distribución considerada para cubrir este propósito es la distribución logística.

La función de distribución logística $A(z)$ se define como $A(z) = \exp\ (z)/(1 + \exp(z))$ Aplicando esta ecuación a la condición (6) obtenemos:

$$Prob(Default_i) = A\ (Score_i) = \frac{\exp\ (b'x_i)}{1 + \exp\ (b'x_i)} = \frac{1}{1 + \exp\ (-b'x_i)} \quad (1.7)$$

Los modelos que vinculan información de probabilidad usando la función de distribución logística se llaman modelos *logit*. Habiendo recolectado los valores de las variables X, una forma de estimar los coeficientes *b* es utilizando el método de máxima verosimilitud. De acuerdo a este principio, los pesos de los coeficientes *b* son elegidos tal que la probabilidad de observar el comportamiento del *default* sea máximo.

El primer paso en la estimación de los coeficientes de máxima verosimilitud es establecer la función de verosimilitud. Para el caso de un prestatario que se encuentra impago (Yi=1), la probabilidad de observar el *default* es:

$$Prob\ (Default_i) = A(b'x_i) \quad (1.8)$$

De forma análoga, para un prestatario que no entró en *default* (Yi=0) la probabilidad es:

$$Prob\ (No\ Default_i) = 1 - A(b'x_i) \quad (1.9)$$

Podemos combinar las dos fórmulas en una sola que automáticamente brinde la probabilidad correcta de *defaultear* o no. Como cualquier número elevado a la potencia de 0 da como resultado 1, la probabilidad de la observación se puede escribir como:

$$L_i = A(b'x_i)^{yi}(1 - A(b'x_i))^{1-yi} \quad (1.10)$$

Asumiendo que los *defaults* son independientes, la probabilidad de un conjunto de observaciones resulta del producto de las probabilidades individuales.

$$L = \prod_{i=1}^{N} Li = \prod_{i=1}^{N} A(b'x_i)^{yi}(1 - A(b'x_i))^{1-yi} \quad (1.11)$$

Para los fines de la maximización resulta más conveniente tomar el ln de L, de esta manera el logaritmo de la probabilidad queda expresado como:

$$\ln L = \sum_{i=1}^{N} y_i \ln\left(A(b'_{xi})\right) + (1 - y_i)\ln\left(1 - A(b'_{xi})\right) \quad (1.12)$$

La función se puede maximizar haciendo la derivada primera con respecto a b.

$$\frac{\partial \ln L}{\partial b} = \sum_{i=1}^{N} (y_i - A(b'_{xi}))x_i \quad (1.13)$$

Para resolver la ecuación se puede aplicar el método de Newton que consiste en derivar la ecuación dos veces, de esta manera resulta:

$$\frac{\partial^2 \ln L}{\partial b\, \partial b'} = -\sum_{i=1}^{N} A(b'x_i)(1 - A(b'x_i))x_i x'_i \quad (1.14)$$

Entre todas las metodologías disponibles, la revisión de la literatura muestra que entre los métodos más usados en la industria para la confección de los modelos de riesgo de crédito predominan los enfoques econométricos, tales como los modelos *probit*, junto con las regresiones lineal y logística, el análisis discriminante y los árboles de decisión. Los motivos para su predominio son básicamente dos: en general las metodologías relevadas muestran resultados similares, por lo que tienden a emplearse aquellas cuyo funcionamiento e interpretación son más sencillos, en contraposición a enfoques más sofisticados y de difícil interpretación, como ser las redes neuronales (Gutiérrez Girault, 2007).

Cabe destacar que la mayoría de los bancos se basan en este tipo de modelos econométricos de calificación crediticia para tomar decisiones vinculadas al otorgamiento de préstamos de la banca de consumo minorista. Sin embargo, para los fines del trabajo, el modelo se aplicará al sector corporativo, lo cual no es lo más habitual.

A continuación, se procederá a explicar con más detalle los modelos estructurales más representativos en la literatura del riesgo de crédito, tal es el caso del modelo de Merton y el KMV que se deriva del anterior.

1.4. Modelos estructurales

Los modelos estructurales suponen que se dispone de información de mercado de las firmas, específicamente de información bursátil y consideran al riesgo predeterminado como una opción de venta europea sobre el valor del activo de la firma (Black and Scholes, 1973; Merton, 1974). Se denominan modelos estructurales debido a que la probabilidad de *default* depende exclusivamente de características estructurales de la empresa como el apalancamiento o la volatilidad de los activos. En estos modelos el evento de *default* ocurre cuando los activos caen por debajo del nivel de los pasivos de la empresa al vencimiento. Entre los principales exponentes se encuentran: el modelo de Merton (1974), el modelo *Credit Portafolio Manager* de KMV Moody's (Crouhy, Galai

y Mark, 2000) y el modelo de *Credit Metrics de JP Morgan* (1997) que se expondrán a continuación.

1.4.1 Modelo estructural de Merton

El Modelo de Merton (1974) forma parte de los modelos estructurales de valuación de riesgo de crédito. Relaciona el riesgo de *default* con la teoría de la valuación de las opciones financieras y la estructura de capital de las empresas. Supone que las empresas tienen dos formas de financiación, a través de acciones y de deuda. Bajo los supuestos del modelo establece que la empresa entrará en *default* cuando sus pasivos sean superiores a sus activos.

Analiza las acciones de la compañía como una opción *call*[13] sobre el valor de los activos y establece que se incurrirá en *default* cuando los activos de la empresa sean inferiores a sus pasivos. En consecuencia, la probabilidad de incumplimiento es la probabilidad de que, en el momento T, el valor de los activos este por debajo del valor de los pasivos. Para ello, utiliza una formulación matemática basada en la fórmula de Black-Scholes que permite medir el número de desviaciones estándar entre el valor esperado del activo y el valor de la deuda (punto de *default*), lo que se conoce como distancia de *default* (DD).

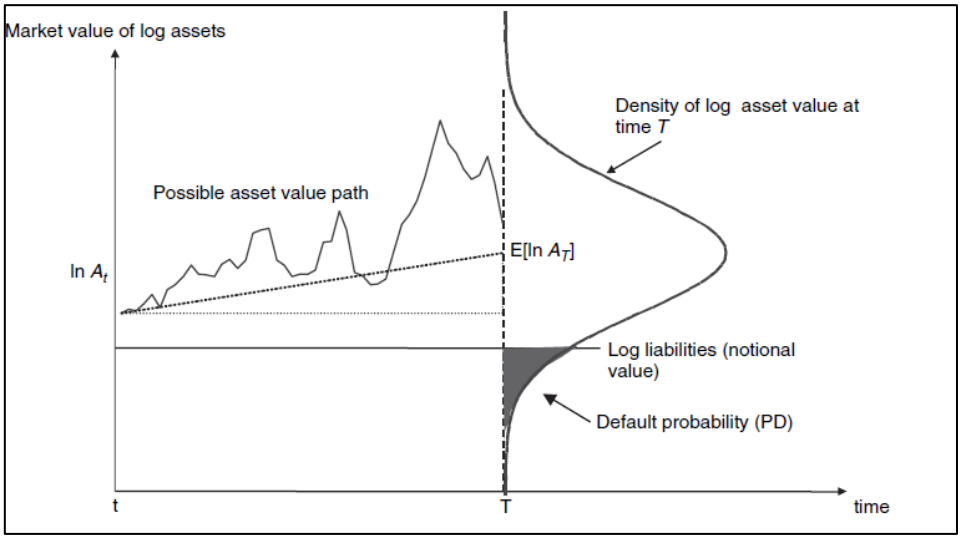
El modelo de Merton establece que los pasivos de la empresa consisten en un bono cupón cero con valor teórico L y con vencimiento en T. Supone que no hay pagos hasta T, y los titulares de acciones esperarán hasta T antes de que decidan si se declaran en *default* o no. Si ellos *defaultean* antes de T renunciarían a la posibilidad de beneficiarse de un aumento del valor del activo. En consecuencia, la probabilidad de *default* es entonces la probabilidad de que, en el momento T, el valor de los activos se encuentre por debajo del valor de los pasivos. Para determinar esta probabilidad se requiere información sobre los pasivos de la empresa que se obtiene de la hoja de balance.

Como se observa en el gráfico 1, que a continuación se presenta, para estimar la probabilidad de *default* se calcula el valor de los pasivos y el valor de los activos. El valor de los pasivos se considera como un bono cupón cero y además se supone fijo en el corto plazo, hasta T, por lo que su valor se puede extraer de los estados financieros de la empresa. Para estimar el valor de los activos en el corto plazo, el cual no se supone fijo, se necesita especificar la distribución de probabilidad en T. Una suposición común establece que el valor de los activos financieros (A_t) sigue una distribución log normal. El cambio anual esperado del valor logarítmico de los activos se denota por $\mu - \delta^2/2$. Simbolizamos a t como el período actual, donde μ es la media de los retornos y δ la volatilidad de los activos. Siendo t el período actual, el valor logarítmico del activo en T sigue una distribución logarítmica normal con los siguientes parámetros (1.15):

$$\ln A_t \sim N\left(\ln A_t + \left(\mu - \frac{\delta^2}{2}\right) * (T-t)\,,\ \delta^2\,(T-t)\right) \quad (1.15)$$

[13] Opción de compra, según su traducción al español.

Gráfico 1: Representación gráfica del modelo de Merton



Fuente: Löeffler, G., & Posch, M. P. N. (2011). Credit risk modeling using Excel and VBA. John Wiley & Sons. P. 28

En general, la probabilidad de que una variable X que se distribuye en forma normal caiga por debajo de Z, está dado por $\emptyset[(z - E[x])/\sigma(x)]$ donde $\emptyset$ denota la probabilidad acumulada de una distribución normal estándar. Aplicando este resultado a nuestro caso obtenemos la probabilidad de que el valor de los activos (A_t), caiga por debajo del valor de los pasivos (L). Esta probabilidad conocida como la probabilidad de *default* se obtiene como (16):

$$Prob\ (Default) = \frac{\emptyset[\ln L - ln\ A_t - (\mu - \delta^2/2)(T-t)]}{\delta\sqrt{T-t}}$$
$$= \frac{\emptyset[\ln(L/A_t) - (\mu - \delta^2/2)(T-t)]}{\delta\sqrt{T-t}} \qquad (1.16)$$

Usualmente se aplica el término distancia de *default* (DD) para medir el número de desviaciones estándar donde el valor esperado del activo At se separa del punto de *default*. De este modo, la ecuación anterior la podemos reescribir como (17):

$$DD = \frac{\ln A_t + (\mu - \delta^2/2)(T-t) - \ln L}{\delta\sqrt{T-t}}$$
$$= Pob\ (Default) = \emptyset(-DD) \qquad (1.17)$$

Si se conociera el valor de las variables y de los parámetros, la estimación de probabilidad de *default* se podría realizar fácilmente. Sin embargo, para una empresa típica no es posible observar el valor actual de mercado de los activos A_t lo que podemos observar es el valor libros de los activos que difiere del valor de mercado. Si no se puede observar el valor de los activos A_t, tampoco se conoce su volatilidad δ, por lo que no se puede aplicar la fórmula (1.16) y (1.17) para determinar la probabilidad de *default*.

En este contexto, Merton aplicó la teoría de *pricing*[14] de las opciones para resolver el problema de la valuación de los pasivos de una empresa en presencia de *default*. Esta teoría permite establecer una relación entre los valores inobservables de (At, δ^2) y las variables observables. De esta manera, el modelo estima el valor de los activos a partir del valor de mercado de las acciones que surge de la multiplicación el precio de la acción por el número de acciones en circulación y el valor de los pasivos a partir de su valor libros. Utilizando el valor de los activos y de los pasivos se construye una medida que representa el número de desviaciones estándar entre el activo y el pasivo.

Merton establece una relación entre la teoría de valuación de las opciones financieras y la estructura de capital de las empresas. Esto permite tratar el valor de mercado de las acciones y el valor nominal de los pasivos como opciones financieras sobre el valor de los activos de la empresa. La empresa emite dos tipos de responsabilidades sobre el valor de sus activos: acciones y deuda. Merton representa la acción como una opción *call*, comprada por los accionistas y la deuda como una opción *put*[15] vendida por los acreedores, siendo L el precio de ejercicio de los dos casos.

Al ser representada la estructura de capital de la empresa como una opción *call*, el valor de la opción es el valor de mercado de las acciones y el precio spot es el valor de mercado de los activos (At). En el periodo T, se establece la siguiente relación para el valor esperado del *call*:

$$E_T = \max(0, A_t - L) \quad (1.18)$$

Si al vencimiento el valor de los activos supera la deuda la empresa pagará y su capital sería la diferencia entre los activos y los pasivos. En términos de derivados, la empresa estaría *In The Money*[16]. Sin embargo, si la deuda supera a los activos, no se ejercería la opción y se declararía en quiebra, por tanto, estaría *Out the Money*[17]. Si el valor de los activos supera al de la deuda, la opción estará *In the Money* y se pagará la deuda. En caso contrario, la empresa se declarará en quiebra.

En el periodo T se puede establecer una relación entre el valor del capital y el de los activos. Mientras que el valor de los activos se encuentre por debajo del valor de los pasivos, el valor del *equity*[18] será nulo, no se ejercerá la opción de compra, la empresa entrará en quiebra y los activos pasarán a manos de los acreedores. Si, por el contrario, el valor de los activos supera el valor nominal de los pasivos, las acciones tomarán un valor positivo y su valor se incrementará linealmente con los incrementos del activo.
En la fecha T el pago de la deuda es:

$$B_T = A_t - E_t = A_t - Max\ A_t - L, 0 = Min\ A_t, L \quad (1.19)$$

[14] Fijación de precios, según su traducción al español.
[15] Opción de venta, según su traducción al español.
[16] Opción que ha superado su precio de ejercicio, según su traducción al español.
[17] Opción que no ha superado su precio de ejercicio, según su traducción al español.
[18] Capital, según su traducción al español.

El pago del bono está determinado por el valor de los pasivos menos la opción de venta (opción *put*) sobre el valor de la compañía. Se pueden determinar dos escenarios, primero el titular de la deuda recibe en T la deuda pactada, es decir L, ya que en T la empresa no ejerce el derecho de venta dado que el valor de los activos supera a los pasivos. En el segundo escenario, el titular de la deuda recibe un valor inferior a la deuda pactada ya que la empresa ejerce la opción de venta sobre el valor de la compañía dado que los pasivos superan el valor de los activos en T.

La acción se puede considerar como un *call* sobre el valor de los activos, este valor de la acción está dado por la fórmula de Black y Scholes. De esta manera el *pay-off*[19] de los tenedores de bonos es un portafolio compuesto por un bono cupón cero con un valor nocional de L y un corto en un *put* en los activos de la firma, con un *strike* de L. Si la firma no paga dividendos, el valor del capital se puede determinar con la fórmula estándar de opciones de Black-Scholes de la siguiente manera:

$$E_t = A_t * \emptyset(d_1) - L\, e^{-r(T-t)} * \emptyset(d_2) \quad (1.20)$$

Donde,

$$d_1 = \frac{\ln(A_t/L) + (r + \delta^2)(T-t)}{\delta\sqrt{T-t}} \qquad d_2 = d_1 - \delta\sqrt{T-t} \qquad (1.21)$$

Siendo r la tasa libre de riesgo de los retornos logarítmicos. Gracias a la teoría de *pricing* de las opciones Merton obtuvo una ecuación que vincula los valores observados (valor del *equity*) con los dos parámetros desconocidos (At, δ^2) a partir de la ecuación (1.20) y (1.21). De esta manera se obtiene una ecuación con dos variables desconocidas. El modelo puede resolverse mediante un proceso iterativo o planteando un sistema de ecuaciones.

A modo de conclusión cabe destacar que, aunque el modelo de Merton presenta ventajas debido a su sencilla representación y calibración, su evaluación empírica no resulta del todo absoluta ya que se han encontrado sesgos a la hora de realizar su contrastación y se ha detectado a través de contrastaciones empíricas que el modelo subestima el riesgo de *default* de las empresas.

Una de las limitaciones del modelo de Merton radica en que asume que el pasivo de una empresa está compuesto únicamente por la emisión de bonos y que la insolvencia o probabilidad de *default* puede producirse al vencimiento de tal obligación, que se asume de un año. Ese hecho impide determinar la probabilidad de impago para un periodo inferior a la vencimiento del bono (Munafo, 2018). A su vez, el modelo parte del supuesto de que el valor de los activos y de la acción sigue un Movimiento Browniano Geométrico. Este supuesto es utilizado cuando se supone que el valor del activo puede ser calculado a partir del valor de mercado de las acciones que surge de multiplicar su

[19] Ganancia, según su traducción al español.

precio por las cantidades de las mismas, y al calcular la probabilidad de *default* en términos de la distancia al *default*.

El supuesto de Movimiento Browniano Geométrico incluye suposiciones adicionales que restringen al modelo. El cambio de los precios es independiente, es decir el precio del periodo actual no guarda relación con el precio del período anterior. El cambio en los precios sigue una distribución normal, por lo tanto, el proceso está dominado por eventos ordinarios, mientras que los eventos extremos ocurren con poca frecuencia. El cambio en los precios es continuo, es decir sin saltos. A pesar de las limitaciones del modelo inicial de Merton de 1974, la literatura sugiere que el modelo presenta buenos resultados en su aplicación práctica cuando el mismo se contrasta con modelos más avanzados y por eso se continúa siendo el padre de los modelos estructurales utilizado en el área de riego de crédito. Además, constituye la base para los cálculos de las probabilidades de incumplimiento en el modelo de KMV Moody's que se detalla a continuación.

1.4.2 Modelo de *Credit Portfolio Manager* de KMV *Moody's*

En 1995, la empresa americana KMV, buscó cómo estimar la función de distribución de las pérdidas de una cartera de créditos, teniendo en cuenta cambios en la calidad crediticia y el umbral de quiebra. Considera que la probabilidad de quiebra no es constante para cada rating, sino que por el contrario sigue un proceso continuo. En este modelo, cada prestatario tiene una frecuencia esperada de incumplimiento, que depende de que el valor de sus activos disminuya por debajo del punto de quiebra.

Moody´s buscó calcular el valor esperado de una empresa y su volatilidad a través del valor de sus activos, utilizando para la estimación de éstos el valor de mercado las acciones ya que no se conoce directamente el valor de los activos de la compañía. Lo mismo ocurre con la volatilidad de los activos, que resulta en principio desconocida. Con estos datos se busca calcular el número de desviaciones típicas desde el valor esperado de la empresa y desde el punto de quiebra.

Para la construcción del modelo, al igual que el modelo de Merton se asume que las acciones de la compañía equivalen a una opción de compra o *call* sobre el activo de la compañía. El precio de ejercicio sería el valor de la deuda y el tiempo hasta vencimiento el de la deuda en cuestión. Así, si el valor de la deuda es superior al de los activos, los accionistas ejercerán la opción. El valor de capital de la empresa, el cual es la diferencia entre activos y deuda depende del valor de mercado de los activos (denotado por A), de la volatilidad de los activos, del punto de inflexión o *break even*, a partir del cual la empresa se declara en quiebra, del tiempo hasta el vencimiento, de los tipos de interés libre de riesgo y de los cupones que se pagarán.

Bajo los supuestos antes mencionados la distancia al punto de quiebra puede expresarse como:

$$Dist.punto\ de\ quiebra = \frac{E(A) - K}{\sigma_A} \quad (1.22)$$

Siendo $E(A)$ el valor estimado de los activos de la empresa, K el punto de quiebra y σ_A la desviación típica normalizada del valor estimado de los activos.

Según el número de desviaciones típicas obtenidas como distancia al punto de quiebra se puede calcular según una normal la probabilidad de que el valor de los activos descienda sobre el punto de quiebra y estimar, por tanto, la probabilidad de quiebra de la empresa. No obstante, cabe destacar que el modelo no asume la hipótesis de normalidad. Para estimar la probabilidad entonces, este modelo utiliza el dato obtenido de distancia al punto de quiebra y lo corrobora con datos históricos de empresas con el mismo dato que entraron en default, sobre el total de empresas con ese dato.

En el modelo de Merton el incumplimiento se presenta cuando el valor de los activos es inferior al monto total de la deuda financiera, mientras que en el modelo de *KMV Moody's* el incumplimiento se presenta cuando el valor de los activos está por debajo de un umbral definido entre los pasivos totales y la deuda de corto plazo. Esta diferencia en el establecimiento del umbral afecta directamente la probabilidad de incumplimiento, manteniendo los demás factores constantes, mientras más alejado se ubique el umbral del monto total de la deuda menor será la probabilidad de incumplimiento.

Una vez calculada la distancia al punto de quiebra en el modelo *KMV* sólo hay que hacer divisiones en categorías de las distintos datos que se obtienen con la distancia al punto de quiebra comparándolos con una escala de rating como la de *Moody's* o *S&P*. Luego, es posible calcular la matriz de transición de unas categorías de rating a otras. Este modelo detecta los cambios en la calidad crediticia de un deudor con antelación a lo que lo hacen las agencias de rating. Sin embargo, la desventaja principal del modelo es que la estructura de deuda que asumió la empresa va a permanecer inalterada durante la vida de la compañía, lo cual no resulta muy realista.

1.4.3 Modelo de *Credit Metrics* de JP Morgan

El modelo de *Credit Metrics* fue desarrollado en 1997 por un grupo de instituciones financieras encabezadas por JP Morgan (Morgan, 1997). Este modelo tiene como propósito estimar el valor a riesgo (VaR)[20] de crédito. Supone que el riesgo de crédito depende de los cambios en la calificación crediticia y en la tasa de incumplimiento de los deudores. Desarrolla el modelo de riesgo de crédito a través de etapas; primero específica un sistema de calificaciones y una matriz de transición utilizando la información de las agencias calificadoras (*Moody's o Standard & Poor's*); después establece un horizonte de tiempo que por lo general es de un año, luego desarrolla un modelo de valoración, así logra analizar los cambios en el valor de la cartera de créditos, y por último define el incumplimiento como el momento en el cual el valor de los activos se encuentra por debajo del valor nominal de los créditos.

El componente principal del modelo es la matriz de transición que está relacionada con un sistema de clasificación, el cual modela la migración de la calidad de los créditos. Con esto se determina las perdidas resultantes de los incumplimientos del deudor y los

[20] Se entiende por valor a riesgo como una medida de riesgo de mercado en una cartera de inversiones de activos. Resume la máxima pérdida esperada para un horizonte de tiempo y un determinado nivel de confianza.

cambios en el valor de mercado de los créditos de la cartera. El modelo calcula el riesgo de crédito basándose en la probabilidad de cambiar de *rating*, así como en el impacto asociado a la correlación para un período dado. Modeliza la distribución futura de los valores de un bono o de una cartera, donde los cambios se relacionan solo con la migración en el rating. Los tipos de interés se suponen deterministas.

Dado que el modelo *Credit Metrics* utiliza información de las agencias calificadoras, esto genera un inconveniente el cual radica en que las probabilidades de incumplimiento actuales sean iguales al promedio de las probabilidades de incumplimiento de la compañía calculada con datos históricos. Otro inconveniente que presenta el modelo *Credit Metrics* es que supone que todas las compañías calificadas en una misma categoría presentan la misma probabilidad de incumplimiento. Por lo que los cambios en la calidad de la deuda son idénticos y el rating y la probabilidad de impago son sinónimos.

1.5 Modelos no paramétricos

Los modelos no paramétricos y los de inteligencia artificial, como por ejemplo los árboles de clasificación o decisión, las redes neuronales y los algoritmos genéticos, son superiores a los modelos estadísticos que se mencionaron anteriormente cuando se desconoce la probable forma de la relación funcional y se presume que esta no es lineal. Los árboles tienen la ventaja de no requerir la formulación de supuestos estadísticos sobre distribuciones estadísticas o formas funcionales. A su vez, presentan la relación entre las variables, los grupos y el riesgo de manera visual, con lo cual, si el conjunto de variables en el análisis es reducido, facilita entender cómo funciona el *scoring*.

A continuación, se procederá a explicar con más detalle el modelo de redes neuronales, el *Support Vector Machine* y los árboles de decisión.

1.5.1 Modelo de redes neuronales

Es una metodología catalogada dentro de las técnicas no paramétricas y no estadísticas de *credit scoring*. Las redes neuronales utilizan procesos de aprendizaje para buscar solución a diferentes problemas; son un conjunto de algoritmos matemáticos que encuentran relaciones no lineales entre conjuntos de datos; suelen ser utilizadas como herramientas para la predicción de tendencias y como clasificadoras de conjuntos de datos (Perez Ramirez & Fernandez Castaño, 2007). Tratan de imitar al sistema nervioso, construyendo sistemas con cierto grado de inteligencia. La red está formada por una serie de procesadores simples, denominados nodos, que se encuentran interconectados entre sí. Como nodos de entrada se consideran las características o variables de la operación de crédito. El nodo de salida sería la variable respuesta definida como la probabilidad de no pago. La finalidad de cada nodo consiste en dar respuesta a una determinada señal de entrada. El proceso de c*redit scoring* mediante el uso de esta técnica resulta complicado, pues el proceso interno de aprendizaje funciona como una "caja negra", donde la comprensión de lo que ocurre dentro requiere de conocimientos de especialistas.

Desai, Crook & Overstreet (1996); West(2000) y Soydaner & Kocadağlı (2015) buscaron explicar el uso de redes neuronal que utilizan un sistema artificial que se asemeja al cerebro humano y es capaz de arribar a una predicción eficiente. Por otro lado, Ripley (1994) y Rosenberg y Gleit (1994) describieron algunas de las aplicaciones de las redes neuronales empleadas en las decisiones gerenciales sobre el crédito y sobre la detección del fraude.

Las redes neuronales resultan superiores a otras técnicas de *credit scoring* ya que se entrenan, auto organizan, aprenden y olvidan; son robustas y tolerantes a fallas, la falla de una o varias neuronas no implica un fallo total en la red; son flexibles por lo que pueden adaptarse fácilmente a nuevos ambientes; funcionan como sistemas independientes; presentan una gran velocidad de respuesta; son hábiles en el proceso de asociar, evaluar o reconocer patrones. La principal desventaja de utilizar redes neuronales para el otorgamiento de crédito radica en que funciona como una caja negra, donde se toma una decisión, pero desconociendo los fundamentos en los que se basa. En este contexto podría ocurrir que un director de riesgo niegue un crédito solo porque las variables de la salida de la caja negra así lo determinan, sin que pueda argumentar esta decisión por desconocer el funcionamiento de la red neuronal, por este motivo resulta necesario acompañar el uso de esta técnica con el juicio del analista.

1.5.2 *Support Vector Machine* (SVM)

Nacido del modelo presentado en la sección anterior, surge el modelo de *Support Vector Machine* (SMV) para el análisis de riesgo de crédito. Entre los principales exponentes se encuentran Martens et al. (2010); J. F. Moreno & Melo (2011) y Zhou, Lai & Yen (2009) que identificaron el proceso que usa geometría euclidiana para discriminar correctamente los datos; la operación matemática trata de identificar los espacios que existen entre los datos, llamados “hiperplanos”, que muestran la distancia discriminante que hay en respuestas de tipo binomial [0,1]. El factor de entrenamiento de los datos puede ser representado por la siguiente fórmula:

$$y(x) = sign(w^T x + b) \qquad (1.23)$$

Dicha expresión identifica dos posibles relaciones de discriminación:

$$w^T x + b \geq +1, si\ y_k = +1 \qquad (1.24)$$

$$w^T x + b \leq +1, si\ y_k = -1 \qquad (1.25)$$

Siendo y_k= es un valor -1 o 1 que determina la clase a la que pertenece x, siendo este un vector real de carácter p dimensional, y w, un es un vector normal en el hiperplano

Siendo esta una función lineal clásica que tomará el nombre de hiperplano, el objetivo es analizar las distancias de cada punto, se bifurca en posiciones de [-1, +1] mostrando la ecuación óptima que sirve para entrenar y discriminar la información, la cual proporciona una predicción óptima.

1.5.3 Árboles de Decisión

Al igual que el modelo de redes neuronales, la principal ventaja que presentan los árboles de decisión radica en que no están sujetos a supuestos estadísticos referentes a distribuciones o formas funcionales. Aunque conllevan una comprensión interna difícil sobre su funcionamiento, presentan relaciones visuales entre las variables, los grupos de la variable respuesta y el riesgo; por ello, este método es muy utilizado en *credit scoring.*

Los aportes de Breiman, Friedman, Olshen y Stone (1984) fueron determinantes para el desarrollo de otros trabajos utilizando esta técnica. Entre ellos, Makowski (1985)**,** Coffman (1986) y Carter y Catlett (1987) aplicaron modelos de árboles de decisión para la clasificación de clientes en términos de *credit scoring*.

1.6 Consideraciones

En el capítulo, fueron presentadas diferentes metodologías utilizadas para la calibración de modelos de *credit scoring*. En una primera instancia, se presentaron los modelos de forma reducida, entre los que se encuentran como principales exponentes, el modelo de Altman o Z Score y los modelos de regresión logística. En una segunda instancia, fueron presentados los modelos estructurales, entre los que se destaca el modelo de Merton, el modelo de *Credit Portfolio Manager* de KMV Moody's y el modelo de *Credit Metrics* de JP Morgan. Finalmente, se expusieron los modelos, menos utilizados, pero no por ello menos importantes, entre los que se encuentran los modelos no paramétricos, tales como el modelo de redes neuronales, *Support Vector Machine* y los árboles de decisión.

Entre todas las metodologías disponibles, para la calibración de los modelos de modelos de riesgo de crédito, la revisión de la literatura sugiere que los modelos más utilizados en el mercado son los modelos *probit* y *logit*. Su predominio se debe a su sencillo funcionamiento e interpretación a la vez que se implementación permite arribar a resultados confiables a través de una metodología entendida por diversas áreas de una institución financiera.

Sin embargo, a modo de conclusión, cabe destacar que, a pesar de la proliferación de las numerosas metodologías utilizadas para la estimación de la probabilidad de *default* en los modelos de riesgo de crédito, el juicio del analista continúa siendo utilizado en la originación de créditos, en algunos casos expresado como un conjunto de reglas que la entidad aplica de manera sistemática para filtrar solicitudes o deudores. En la práctica, conviven de forma conjunta el juicio del analista con el modelo elegido por la Entidad.
En el próximo Capítulo se introducirán el concepto de *text mining* y su importancia para los modelos de riesgo de crédito en la era del *big data*.

CAPÍTULO 2: *Text Mining* aplicado a modelos de *Credit Scoring*

2.1 Introducción

Los conceptos relacionados con Big Data se presentan como una tecnología disruptiva que está revolucionando la forma en que funciona nuestro mundo siendo capaz de impactar de manera estratégica en toda la sociedad, por su capacidad de generar transformaciones productivas, económicas y sociales de gran envergadura.

En particular, el *big data* hace referencia a la construcción, organización y utilización de enormes cantidades de datos, particularmente no estructurados que superan la capacidad de un software convencional para ser capturados, administrados y procesados en un tiempo razonable, por lo que se hace necesario extraer relaciones o crear nuevas formas de valor. Una vez almacenados los datos no estructurados resulta necesario recurrir a su análisis para ello se utilizan diversas técnicas como la asociación, la minería de datos o *data mining*, la agrupación o *clustering* y el análisis de texto o *text mining*.

En particular, el *text mining* surge como una ciencia capaz de convertir el texto no estructurado en datos estructurados, extraer índices numéricos y, por lo tanto, hacer que la información contenida en el texto sea accesible a los diversos algoritmos de minería de datos. En términos más generales, la minería de texto permite convertir información textual en información numérica, que luego se podrán incorporar en otro tipo de análisis como proyectos de minería de datos predictivos la aplicación de métodos de aprendizaje no supervisado como el análisis de sentimiento.

En este contexto, la información textual, como el uso de datos de las redes sociales permite ofrecer un enfoque alternativo de puntuación de crédito y mejorar su evaluación del riesgo crediticio, en particular su enfoque de evaluación cualitativa. Esta información incluye contenido producido profesionalmente, como informes de analistas y periodismo de negocios, así como textos informales como blogs y publicaciones en redes sociales, tales como *Twitter*. En comparación con la información financiera disponible sobre diversas corporaciones, la cantidad de contenido textual es inmenso y proporciona un gran volumen de información útil para su análisis y posterior procesamiento.

Si bien el *big data* promete mejoras tecnológicas que posibiliten un mejor conocimiento del mercado, descubriendo y potenciando las necesidades de una compañía, su utilización puede generar ciertos riesgos y peligros por el exceso y el uso de la información. Cabe destacar que, más allá de los riesgos potenciales en los que se encuentra inmerso el *big data*, el mismo tiene el potencial de generar grandes ventajas y beneficios para la sociedad en su conjunto, siempre y cuando se utilice bajo el paraguas de la responsabilidad social.

El presente capítulo se encuentra estructurado en seis secciones. En primer lugar, se introducirá el concepto de *big data*, luego se desarrollará el concepto de *text mining* en la era del *big data* para luego vincularlo al área de riesgo de crédito. Posteriormente, del concepto de text mining se deriva el análisis de sentimiento como una herramienta que permite convertir información textual en información cuantitativa para su posterior análisis y procesamiento. En este concreto, se introduce la red social *Twitter* para el análisis de sentimiento y se concluye el capítulo problematizando la importancia de la utilización del *big data* en un contexto de responsabilidad social, considerando sus principales riesgos y beneficios para la sociedad en su conjunto.

2.2 *Big data*

El *big data* se presenta como una tecnología disruptiva que está revolucionando la forma en que funciona nuestro mundo siendo capaz de impactar de manera estratégica en toda la sociedad, por su capacidad de generar transformaciones productivas, económicas y sociales de gran envergadura (Schmarzo, 2013). Sin embargo, a diferencia de otras revoluciones tecnológicas como la Revolución industrial donde el impulso estaba en la energía, las tecnologías de información y comunicación (TIC's) donde el centro era el procesamiento y transmisión de la información, en el caso de la era del *big data* en la que vivimos el impulso se centra en la transformación, análisis, uso y almacenamiento de enormes volúmenes de información automatizada (McAfee, Brynjolfsson, Davenport, Patil, & Barton, 2012).

El concepto de *big data* se aplica a todo el conjunto de información que no puede ser procesado o analizado utilizando herramientas o procesos tradicionales. Hace referencia a la construcción, organización y utilización de enormes cantidades de datos que superan la capacidad de un software convencional para ser capturados, administrados y procesados en un tiempo razonable, por lo que se hace necesario extraer relaciones o crear nuevas formas de valor (John Walker, 2014). De esta manera, a través del análisis y procesamiento de la información se busca encontrar patrones repetitivos que permitan crear relaciones para el fácil acceso a la información. En este contexto, el objetivo del *big data*, al igual que los sistemas analíticos convencionales, es convertir el dato en información útil que facilite la toma de decisiones, en tiempo real (Schmarzo, 2013).

Si bien las definiciones de *big data* no son uniformes entre sí, todas presentan como común denominador el análisis de grandes volúmenes de información (B. Brown, Chui, & Manyika, 2011). Una definición provista en el 2001 por Douglas considera al término *big data* en relación a sus características principales, lo que se conoce como el desafío de las 5 Vs. Las mismas son el Volumen, la Velocidad, la Variedad, la Veracidad y el Valor (Douglas, 2011). A continuación, se procederá a explicar en detalle cada una de ellas. En primer lugar, el volumen hace referencia a los datos y metadatos que debe ser capaz de recolectar, almacenar y tratar. El crecimiento exponencial de estos datos dificulta que sea analizado mediante herramientas y procesos tradicionales como se hacía anteriormente tales como MS Excel o SQL, para ello es necesario utilizar nuevos sistemas como NoSQL o el software Apache Hadoo, que permiten trabajar millones de bytes de información y organizarlos en miles de nodos.

La Velocidad con la que se deben procesar los datos se encuentra en continuo aumento, el *big data* permite analizar tanto datos estáticos como lo hacían las tecnologías tradicionales, como los dinámicos que se van creando en tiempo real, permitiendo la realización de predicciones. La Variedad de formas que pueden tomar los datos que se recolectan, pueden ser de tres tipos: estructurados, semi estructurados y no estructurados. Los primeros son aquellos en los que la longitud y el formato se encuentran bien definidos, pudiendo ser almacenados en tablas. Los datos semi-estructurados son aquellos que no residen de bases de datos relacionales, pero presentan una organización interna que facilita su tratamiento, tales como documentos XML y datos almacenados en bases de datos NoSQL. Y finalmente, los datos no estructurados que carecen de un formato específico y por lo tanto no se encuentran almacenados en una base de datos tradicional o predefinida.

La Veracidad de las bases de datos hace referencia al nivel de fiabilidad o calidad de los datos que se recolectan de los grandes volúmenes de información. Este hecho resulta más difícil cuando se trata de datos no estructurados. A su vez, algunos datos son inciertos, como los creados en las redes sociales, por lo que resulta importante el concepto de incertidumbre en estas áreas. Finalmente, el Valor que se obtiene por la información extraída de los datos resulta el fin último de la implementación de *big data*. El valor puede ser entendido como las oportunidades económicas que se obtienen de los grandes volúmenes de información.

La revolución del *big data* radica en que años atrás, con otras revoluciones tecnológicas, las 5 Vs de las bases de datos, el Volumen, la Velocidad, la Variedad, la Veracidad y el Valor resultaban incompatibles entre sí, creando una tensión que obligaba a elegir entre algunas de ellas. Por ejemplo, se podían analizar grandes volúmenes de información, pero estos debían ser sencillos como datos estructurados; es decir que había que sacrificar la variedad de los datos en post de un mayor volumen. Del mismo modo, se podían utilizar grandes volúmenes de datos, pero a un ritmo de trabajo lento, en este caso se sacrificaba la velocidad. O podían analizarse datos a gran velocidad, pero se carecía de veracidad en la información o en el peor de los casos se sacrificaba la generación de valor. Con el surgimiento del *big data* las 5 Vs dejaron de actuar de manera aislada para ser complementarias en la generación de valor.

Los grandes volúmenes de datos están estrechamente relacionados con una necesidad constantemente y creciente de análisis veloces para generar conocimientos en tiempo real, desde una perspectiva del uso de *big data* para el crecimiento. Ambos, el volumen y la velocidad tienen un fuerte impacto en la veracidad, ya que analizan una gran cantidad de datos en diferentes formatos (estructurados, no estructurados y semiestructurados), procesándose a gran velocidad no tendrían valor si esos datos fueran incorrectos. Los datos incorrectos, tienen el potencial de generar problemas cuando se utilizan en los procesos de toma de decisiones de los gobiernos, las empresas y en última instancia termina afectando a los consumidores. Es en este sentido que el Volumen, la Velocidad, la Veracidad y la Variedad se encuentran estrechamente interconectados para la generación de valor.

Una de las preguntas fundamentales que surgen cuando se habla del término *big data* es de dónde provienen las grandes masas de información. Solo basta ver a nuestro alrededor para dar cuenta de toda la información que se genera por segundo en las

redes. En un día, millones de personas envían correos electrónicos, mensajes por *Whatsapp*, publican estados en *Facebook*, en *Instagram*, responden encuestas, generando una enorme cantidad de datos y metadatos que necesitan ser almacenados en alguna parte del universo. En este contexto, cabe destacar que la información no estructurada que se genera diariamente no solo es formada por personas, sino que intervienen en el proceso otro tipo de operaciones como las transacciones bancarias y la información que se genera maquina a máquina. Estas últimas forman parte de la tecnología que comparte datos con dispositivos, medidores, sensores de temperatura, de luz, de altura, de presión, de sonido, que transforman las magnitudes físicas o químicas y las convierten en datos.

Una vez almacenados los datos no estructurados resulta necesario recurrir a su análisis para ello se utilizan diversas técnicas como la asociación, la minería de datos o *data mining*, la agrupación o *clustering* y el análisis de texto o *text mining*. La asociación permite relacionar diferentes variables con el fin de encontrar una predicción en el comportamiento de otras variables; la minería de datos trata de descubrir patrones en grandes cantidades de datos englobando los métodos estadísticos y el aprendizaje automático. La agrupación o *claustering* intenta buscar similitudes entre grupos y el descubrimiento de nuevos a través de las cualidades que los definen. Finalmente, la minería de texto permite extraer información de datos y así modelar patrones o predecir palabras. El trabajo se encuentra centrado en las técnicas de análisis de texto que procederá a explicarse con mayor detalle en el apartado que sigue.

2.3 *Text Mining* en la era del *Big Data*

Antes de definir lo que el término *text mining* es capaz de abarcar, resulta necesario definir sus orígenes en el *data mining*. La minería de datos puede definirse como un proceso de descubrimiento de nuevas y significativas relaciones, patrones y tendencias al examinar grandes volúmenes de información (López, 2007). También puede ser considerada como una combinación de técnicas semiautomáticas de inteligencia artificial, análisis estadístico, bases de datos y visualización gráfica, para la obtención de información que no se encuentra representada explícitamente en los datos (Martínez, E., 2000).

De esta manera, a través de la minería de datos pueden deducirse patrones y tendencias, que no podrían detectarse mediante una exploración tradicional porque las relaciones resultan demasiado complejas o por el volumen de datos que se maneja. En este contexto, surge la minería de datos como aquella parte de la estadística no paramétrica que se utiliza para solventar problemas que se presentan en el análisis de datos.

Introducido el concepto de *data mining* es posible entender que implica el *text mining*, el cual tiene como objetivo la búsqueda del conocimiento en grandes colecciones de documentos. La diferencia con el *data mining* radica en que se obtiene información nueva a partir de grandes cantidades de texto, en la que la información suele estar no estructurada.

En este contexto, el *text mining* surge como una ciencia capaz de convertir el texto no estructurado en datos estructurados, extraer índices numéricos significativos del texto y, por lo tanto, hacer que la información contenida en el texto sea accesible a los diversos algoritmos de minería de datos (estadística y aprendizaje automático). En términos más generales, la minería de texto "convertirá el texto en números" (índices significativos), que luego se podrán incorporar en otro tipo de análisis como proyectos de minería de datos predictivos, la aplicación de métodos de aprendizaje no supervisado.

Para realizar minería de textos existen cinco pasos fundamentales involucrados en el proceso. En primer lugar, se realiza la recopilación de datos no estructurados que proviene de múltiples fuentes como páginas web, redes sociales, archivos pdf, correos electrónicos, blogs, etc. En una segunda instancia se identifica la información relevante, eliminando anomalías en los datos mediante operaciones de preprocesamiento y limpieza de datos, lo que permite extraer y retener información valiosa oculta en el texto y ayudar a identificar palabras específicas. Luego se convierte la información relevante extraída de datos no estructurados en formatos estructurados a través de la elaboración de índices y diversos señalizadores. A continuación, se analizan los patrones dentro de los datos a través de sistemas de gestión de la información y finalmente se almacena la información valiosa en una base de datos segura para realizar análisis de tendencias y mejorar el proceso de toma de decisiones en una empresa.

El *text mining* permite identificar el “quién", "qué", "cuándo", "dónde", "por qué" y el sentimiento en un texto, de esta manera puede ser utilizado para desarrollar una mejor comprensión de los gustos, aversiones y motivaciones del cliente. Las personas se expresan a través de las redes sociales en el momento en que tienen una experiencia e interactúan con una marca. Las empresas pueden tomar este hecho como un indicador por adelantado de la actitud del cliente y detectar con anterioridad como este hecho repercutirá en sus ventas en un futuro. Con este ejemplo, se puede notar que la información que surge del procesamiento de textos resulta valiosa ya que puede utilizarse como un predictor del comportamiento, adelantándose ante cualquier efecto adverso y brindando resultados efectivos para la toma de decisiones y maximización de las ganancias. En los próximos apartados, se procederá a explicar con más detalle como el *text mining* puede ser utilizado en el sector financiero, en particular como una herramienta con potenciales beneficios para predecir con mayor exactitud el riesgo de *default* corporativo.

2.4 *Text Mining* aplicado al riesgo de crédito

Tal como se explicó en el capítulo anterior los modelos de *credit* scoring constituyen una herramienta de apoyo a la decisión utilizada para identificar el nivel de riesgo asociado con los solicitantes para un determinado servicio. Se basa en la aplicación de un conjunto de técnicas estadísticas para predecir el comportamiento de los aspirantes de crédito y asignar puntuaciones que reflejen que tan bueno o malo se espera que sean. Los modelos de *credit scoring* son ampliamente utilizados en la gestión del riesgo de los bancos, compañías de seguros y otras instituciones financieras y pretenden identificar

la calidad o el riesgo de los clientes. El propósito del modelo de puntuación de crédito es aumentar la eficiencia, y la fiabilidad del proceso de juicio convencional (Ghailan, Mokhtar, & Hegazy, 2016).

La mayoría de los bancos utilizan modelos de calificación crediticia para poder tomar decisiones sobre préstamos a empresas. Tales modelos son, de hecho, un requisito para los bancos que utilizan el enfoque basado en calificaciones internas de Basilea II. Estos modelos de riesgo de crédito adoptaron principalmente dos tipos de variables de entrada (Cao, Guan, & Jingqing, 2010; Chan, 2003; Huang, Chen, Hsu, Chen, & Wu, 2004). El primer tipo de variables son los números contables publicados en reportes financieros. Se cree que el rendimiento informado en los balances puede realmente reflejar el empeoramiento de la calidad crediticia en empresas vulnerables (E. I. Altman, 1968; Beaver, 1966; Ohlson, 1980). El otro tipo de variables se obtiene de los mercados financieros. Ejemplos como este incluyen retornos de las acciones, precios de la deuda y actividades en derivados relacionados. Las variables del mercado financiero pueden complementar variables contables proporcionando información actualizada a los informes que se elaboran de forma trimestral o anual.

Sin embargo, cabe destacar que a menudo estos modelos presentan deficiencias significativas. En primer lugar, frecuentemente son retrógrados. En segundo lugar, para su calibración utilizan datos históricos, es decir se basan en la información financiera formal de los prestatarios, lo que significa que los datos siempre tienen al menos 6 meses de antigüedad y hacia el final del año fiscal, los datos tienen casi 18 meses de antigüedad. En tercer lugar, las evaluaciones cualitativas de los prestatarios son simplistas. Finalmente, muchos bancos confían en sus modelos de calificación crediticia para proporcionar una visión instantánea y de largo plazo, con lo cual el resultado no resulta del todo correcto.

Se cree que la información cualitativa pública puede mejorar los modelos de calificación crediticia por varios motivos. En primer lugar, las noticias sobre una empresa pueden proporcionar una alerta temprana o pistas sobre el deterioro de su situación crediticia antes que se refleje en los estados contables y financieros. Para las empresas privadas, las noticias son aún más valiosas porque este tipo de empresas carecen de información de mercado. Las noticias pueden proporcionar información adicional para empresas con mala calidad contable causada por manipulaciones de sus estados. Sin embargo, cabe destacar que los informes exagerados de los medios pueden generar efectos negativos induciendo a los depositantes o prestamistas a retirar sus fondos e incluso terminar provocando la quiebra de las empresas insolventes.

En este contexto, la información textual, como el uso de datos de las redes sociales puede ofrecer un enfoque alternativo de puntuación de crédito y ayudar a los bancos a superar algunos de estos desafíos y mejorar su evaluación del riesgo crediticio, en particular su enfoque de evaluación cualitativa. Esta información incluye contenido producido profesionalmente, como informes de analistas y periodismo de negocios, así como textos informales como blogs y publicaciones en redes sociales. Los artículos periodísticos describen los últimos desarrollos de las empresas; los informes de los analistas proporcionan análisis sobre las diversas estrategias de negocio, el

posicionamiento competitivo y la perspectiva; las calificaciones de los productos en los sitios de compras en línea brindan vistas sin filtro de la satisfacción del cliente; y los microblogs como *Twitter* distribuyen las últimas noticias con una velocidad sin precedentes. En comparación con la información financiera disponible sobre pequeñas y medianas empresas (PYME) o corporaciones, la cantidad de contenido textual sobre las empresas es inmensa y proporciona un gran volumen de información para su análisis y posterior procesamiento.

Estudios financieros recientes en mercados bursátiles encontraron que la cobertura de los medios y su sentimiento están relacionados con el rendimiento de las acciones. Por ejemplo, Chan (2003) encontró que las acciones experimentaron una fuerte variación después de las malas noticias. Tetlock (2007) y Tetlock (2008) descubrió que la fracción de palabras negativas en noticias específicas de la empresa predicen bajas ganancias y bajos retornos de acciones. Los resultados en Fang y Peress (2009) demostraron que la cobertura de los medios constituye un factor clave para explicar los retornos de acciones esperados. Sin embargo, los efectos de las noticias sobre el riesgo de crédito no se han investigado con profundidad, por lo que constituye un tema novedoso. Las relaciones entre los medios de comunicación y la calificación crediticia de las empresas representan una investigación valiosa por este motivo.

Enormes cantidades de información textual están disponibles; esta información ofrece a las empresas una visión profunda de su salud financiera y rendimiento. En este contexto, el objetivo expuesto consiste en desarrollar un modelo de calificación crediticia que pueda identificar y cuantificar el sentimiento dentro de una cantidad dada de información textual. Si los bancos o compañías calificadoras pudieran utilizar incluso una parte de esta información en sus sistemas, se espera una mejoraría en la precisión, la puntualidad y el carácter prospectivo de sus sistemas de evaluación de riesgo crediticio. De esta manera, el análisis de sentimiento podría ayudar a los bancos a mejorar sus análisis tradicionales de industrias y sectores. Por ello, a continuación, se presenta el análisis de sentimiento como una técnica derivada del *text mining*, la cual permitirá llevar a cabo el objetivo propuesto.

2.5 Análisis de sentimiento como medida de riesgo de crédito

Los bancos y compañías financieras han comenzado a emplear una nueva técnica denominada análisis de sentimiento realizado por programas específicos. Dicha técnica utiliza el procesamiento del lenguaje, análisis de texto y herramientas computacionales para clasificar comentarios subjetivos de diferentes usuarios. De esta forma, a la información textual expresada en cualquier formato (palabras, oraciones, párrafos, artículos o libros) se les asigna un "índice de sentimiento", es decir, un número que representa un tipo y grado de opinión expresado por el escritor, como optimismo, confianza, escepticismo, desconfianza, pesimismo, etc. Medir el sentimiento con un índice, hace posible que las máquinas analicen los grandes volúmenes de información textual disponible. De esta manera, la información cualitativa puede ser procesada, convertida y comparada. Finalmente, el índice puede acabar siendo utilizado para realizar análisis estadísticos y construir modelos de predicción.

En la literatura existen, esencialmente dos enfoques para abordar el análisis de sentimiento (Liu, 2012): las técnicas de aprendizaje computacional (Pang, Lee, & Vaithyanathan, 2002) y las aproximaciones semánticas (Turney, 2002). Los enfoques semánticos se caracterizan por el uso de diccionarios de términos con orientación semántica de polaridad u opinión. Típicamente los sistemas pre procesan el texto y lo dividen en palabras, eliminan las *stopwords* o palabras de parada, realizan una normalización lingüística por *stemming* o *lematización* y luego comprueban la aparición de los términos para asignar un valor al texto (positivo, negativo o neutral). En el núcleo del proceso se encuentra un léxico que enumera palabras o frases que representan un cierto tipo de sentimiento. El léxico se debe contextualizar de manera apropiada para tener el tipo de precisión que necesitan los bancos. Tan importante como definir correctamente el léxico es seleccionar y filtrar las fuentes de datos. Una búsqueda amplia dará como resultado más artículos potencialmente relevantes para ser analizados, pero también atraerá mucho material irrelevante, por lo que se genera un *trade-off* difícil de resolver.

Por otra parte, los enfoques basados en aprendizaje computacional consisten en entrenar un clasificador usando un algoritmo de aprendizaje supervisado a partir de una colección de textos, donde cada texto habitualmente se representa con un vector de palabras (*bag of words*), *n-gramas* o *skip-grams*[21], en combinación con otro tipo de características semánticas que intentan modelar la estructura sintáctica de las frases, la intensificación, la negación, la subjetividad o la ironía. Los sistemas utilizan diversas técnicas, aunque las más populares son los clasificadores basados en *Support Vector Machines* (SVM), *Naive Bayes* y *K-Nearest Neighbor* (KNN). En investigaciones más recientes se han empezado a utilizar otras técnicas más avanzadas, como *Latent Semantic Analisis* (LSA) e incluso *Deep Learning*.

El análisis del sentimiento y la información que la misma produce puede mejorar los modelos de calificación crediticia de los bancos, y contribuir con otras dos tareas importantes. En primer lugar, en los modelos de calificación, los bancos pueden usar el índice de sentimiento como un factor de calificación adicional. La información obtenida de las búsquedas de texto puede agregarse trimestralmente a un índice de sentimiento para cada compañía. Después del análisis estadístico, el índice se integra en el sistema de calificación con un peso apropiado. Esto resulta ser particularmente valioso en la evaluación de nuevos clientes corporativos para los cuales los bancos generalmente solo tienen información limitada, la mayoría de la cual es proporcionada por el mismo cliente. En los mercados emergentes, donde los datos confiables de los clientes son escasos, el análisis de la información textual también puede proporcionar información valiosa. Sin embargo, cabe destacar que analizar información textual presenta ciertos desafíos que deben ser tenidos en consideración al implementar esta herramienta.

Los desafíos de extraer información y separar la señal del ruido resultan sustanciales para el análisis de sentimiento. Para usar datos textuales, los bancos deben enfrentar un desafío práctico fundamental: la capacidad computacional. La cantidad de información disponible basada en texto es enorme y está creciendo a pasos agigantados.

[21] Algoritmo de palabras, según su traducción al español.

Las herramientas de programación deben tener la capacidad de poder leer, procesar y analizar los grandes volúmenes de información. Por otro lado, se adiciona un nuevo inconveniente ya que los datos textuales no se encuentran estructurados, es decir no están almacenados en una estructura tradicional para su análisis. Si bien es relativamente fácil analizar los datos financieros de una manera estadística, las cifras se vuelven significativas en tamaño, además los textos a priori no tienen significado para una computadora.

A su vez, no existen procedimientos estándar o estadísticos para que una máquina analice e interprete textos. La tarea de clasificar automáticamente un texto escrito en un lenguaje natural en un sentimiento positivo o negativo, opinión o subjetividad (Pang and Lee, 2008), es a veces tan complicada que incluso es difícil para los expertos llegar a un común acuerdo a la hora de asignar un determinado sentimiento a un texto, ya que se ve afectada por el juicio del analista, su cultura y sus vivencias. Esta tarea resulta aún más difícil cuanto más corto sea el texto, y si a su vez se encuentra escrito en un lenguaje coloquial, como suele ser el caso de mensajes en redes sociales como *Twitter* o Facebook. En particular, el significado de los mensajes cortos en redes sociales resulta difícil de interpretar por los métodos convencionales. Si bien las estructuras de oraciones complicadas se pueden enseñar a las herramientas de programación como *R* o *Python*, el concepto de metáfora, sarcasmo o ironía resulta extremadamente difícil de procesar y entender para una computadora. De hecho, casi todas las dificultades semánticas del lenguaje escrito presentan enormes problemas para su análisis posterior.

Teniendo en cuenta los desafíos antes mencionados a continuación se procederá a explicar con mayor detalle una alternativa utilizada para realizar análisis de sentimiento a la información contenida en la red social Twitter.

2.6 *Twitter* para el análisis de sentimiento

El microblogging, entendido como el sitio que permite a los usuarios enviar y publicar mensajes breves se ha convertido en una herramienta de comunicación muy popular entre los usuarios de Internet, tal es el caso *Twitter*, *Facebook* e *Instagram*. Los datos de las redes sociales son uno de los indicadores más efectivos y precisos de la opinión pública. Millones de usuarios comparten opiniones sobre diferentes aspectos de la vida cotidiana, opinan sobre una gran variedad de tópicos y comentan sobre asuntos actuales. Por lo tanto, los sitios web de microblogging constituyen fuentes ricas de información para la extracción de opiniones y el análisis de sentimiento (Pak & Paroubek, 2010).

A medida que más usuarios publican sobre productos y servicios, los sitos se convierten en fuentes valiosas de opinión, tales datos pueden ser utilizados en minería de datos y tareas de análisis de sentimiento como marketing o para estudios sociales diversos. Por lo tanto, extraer opiniones sobre diversos temas de las redes sociales constituye un enfoque innovador para el análisis de mercado. Por ejemplo, las empresas pueden estar interesadas en las siguiente preguntas: qué piensa la gente de sus productos (servicio o

compañía), qué tan positivo o negativo es la opinión del público en general con respecto a un producto y cómo preferiría que sea dicho producto, entre otras cuestiones.

En este contexto, *Twitter* es el líder destacado de los sistemas de *microbloging* para realizar análisis de sentimientos y minería de opinión. Por este motivo, en la presente tesis se utilizará el *microbloging* y en particular *Twitter*, dicha elección se debe a las razones que se enuncian a continuación: las plataformas de *microblogging* son utilizadas por diferentes personas para expresar su opinión sobre diversos temas, por lo que resulta una fuente valiosa de opinión; *Twitter* contiene una enorme cantidad de mensajes de texto y el corpus recogido puede ser arbitrariamente grande; la audiencia de *Twitter* varía de usuarios regulares a celebridades, representantes de la empresa, políticos, e incluso presidentes de países por lo tanto, es posible recolectar publicaciones de texto de usuarios de diferentes intereses y grupos sociales.

El acceso a los datos de *Twitter* puede realizarse a través de tres APIs: *Rest* API, *Streaming* API, y *Search API*. La *Rest* API ofrece a los desarrolladores el acceso al *core* de los datos de twitter, permite leer tweets y escribirlos, permite consultar el *timeline* de un usuario en particular, permite buscar tweets de usuarios específicos, pero presenta ciertas limitaciones vinculadas al límite de tweets que pueden extraerse diariamente. Por otro lado, API *Streaming* proporciona un conjunto de tweets en tiempo real. Se establece una conexión permanente por usuario con los servidores de Twitter y mediante una petición *http* se recibe un flujo continuo de tweets en formato *json*. Entre las opciones que presenta, se puede obtener todo el caudal de tweets, sólo los que tienen enlaces, una muestra aleatoria, un filtrado por palabras claves o por usuarios o sólo los tweets con retweets por poner un ejemplo. Sin embargo, la desventaja de esta opción es que no es de acceso gratuito. Finalmente, el Search API suministra los tweets con una profundidad en el tiempo de 7 días. Permite filtrar por cliente utilizando lenguaje y localización. No requiere autenticación, es de acceso gratuito y los tweets se obtienen en formato *json* o *atom*. La desventaja es que ofrece una información más limitada del tweet, en concreto sobre los datos del autor en el que sólo indica el Id, el *screen_name* y la url de su avatar. Los otros dos APIs si ofrecen el perfil completo del autor en el momento de la escritura del tweet. En el trabajo se extraerán los tweets a través de la API REST de Twitter. En el siguiente capítulo se retomará esta idea y se explicará en detalle el proceso de extracción, procesamiento y clasificación de los tweets.

La sociedad de la información en la que nos encontramos inmensos experimenta los primeros pasos en la utilización del *big data*, lo que conlleva ciertos beneficios y a su vez los primeros errores y peligros por el exceso de información. Cada rastro digital que dejamos plasmado en las redes puede ser utilizado para recrear nuestra vida cotidiana y nuestros comportamientos, individuales y colectivos (Casanovas, De Koker, Mendelson, & Watts, 2017). Mientras más rastros digitales dejamos, perdemos espacios de privacidad, este hecho puede evidenciarse en las búsquedas que se realizan por internet, el uso de teléfonos celulares, hasta pagos con tarjeta de crédito. Incluso, la información pública que circula en las redes puede ser analizada para conceder o negar un crédito a un solicitante. Más allá de los riesgos potenciales a los que se encuentra inmerso el *big data*, el mismo tiene el potencial de generar grandes ventajas y beneficios

para la sociedad en su conjunto siempre y cuando se utilice bajo el paraguas de la responsabilidad social (Kitchin, 2014).

En el siguiente apartado se explica con más detalle este último punto con el objetivo de enfatizar sobre los aspectos que convierten al *big data* en una herramienta con potenciales beneficios y a su vez riesgos asociados a la privacidad en el uso de los datos.

2.7 *Big data* y su vinculación con la responsabilidad social: potenciales riesgos y beneficios

Entre los beneficios potenciales del *big data* se encuentra la implementación de mejoras tecnológicas que posibilitan un mejor conocimiento del mercado, la adquisición de datos que permiten descubrir las necesidades y puntos de mejora de una compañía. Al mismo tiempo, el análisis de datos puede mejorar sustancialmente la toma de decisiones dentro de la empresa, reduciendo al mínimo los riesgos. Facilita que las compañías puedan crear productos nuevos y rediseñar los ya existentes. Suministra la segmentación de los clientes lo que permite direccionar y personalizar la oferta a la satisfacción de sus necesidades específicas. Y por último mejora la accesibilidad y la fluidez de la información dentro de la propia compañía generando una lógica de trabajo más rápida y eficaz.

De esta manera, las organizaciones en sus tratamientos de *big data* obtienen valor ofreciendo un mejor servicio a sus clientes. Es de esperar que con el correr del tiempo el *big data* ayude a crear nuevas oportunidades de negocio, nuevos mercados y nuevas categorías de empresas. En este contexto, los beneficios que se esperan obtener con el acceso al universo de la información apuntan a ser mayores a nivel general en comparación a los riesgos potenciales que supone el fácil acceso a la información. Por este motivo, es de esperar, que permanecer al margen de esta realidad suponga un costo para las empresas en su crecimiento y una pérdida de su posicionamiento competitivo.

Con la introducción del *big data* comenzó a tratarse el tema de la responsabilidad social en este tipo de innovaciones. La adquisición, el análisis y el uso del *big data* generan un impacto potencial en la sociedad tanto en el sector público como privado. A su vez, la falta de una definición conceptual de lo que el término es capaz de abarcar plantea el mayor inconveniente ya que no se pueden contabilizar los riesgos al no saber exactamente qué es lo que se está evaluando o midiendo.

En este contexto, varios estudios se han realizado recientemente para reflexionar sobre el uso de grandes datos con el fin de analizar el cambio social (Manovich, 2012), analizar las necesidades de privacidad de los datos (Cranor, Rabin, Shmatikov, Vadhan, & Weitzner, 2016), aumentar la preocupación sobre cuestiones sociales y éticas (Boyd & Crawford, 2012) e identificar aspectos éticos, sociales y desafíos de política (Metcalf, Keller, & Boyd, 2016). La mayoría de los enfoques sostienen que los riesgos fundamentales y potenciales del *big data* radican en la violación del límite de la

privacidad, el uso indebido de datos, la confianza mal depositada en la tecnología y la seguridad de los datos.

El proyecto de la Comisión Europea, (2014), *The Big data Roadmap for cross-disciplinar and community for addressing societal Externalities,* se centró en caracterizar los aspectos positivos y negativos del *big data* para el desarrollo de una economía del *big data* socialmente responsable en Europa. En este artículo, los aspectos negativos del *big data* son entendidos por las dificultades para adaptar los mecanismos reguladores a las nuevas características de las interacciones y a asegurar su efectividad.

Por su parte, entre los aspectos positivos, destacan los efectos de datos en red como un incentivo para la generación de modelos de negocios más intensivos en datos y más instructivo y focalizado en las peculiaridades de los individuos. De esta manera, el uso de la información se vuelve más eficiente cuantos más datos se pueden recopilar, convirtiendo los datos personales en un activo, mejorando la eficiencia y la innovación, tanto desde el punto de vista social, económico y ético. Sin embargo, la toma de decisiones puede ser más opaca cuando está basada en una amplia gama de fuentes de datos y por lo tanto menos responsable (Casanovas et al., 2017).

En este contexto cobra especial relevancia el marco de gobernanza en *big data* (Soares, 2012; Tallon, 2013), la cual es entendida como un conjunto de metodologías que proporcionan un marco para establecer políticas de protección de los datos personales e implementar controles para asegurar la calidad, consistencia, accesibilidad y explotación de la información. Sin embargo, el cumplimiento de la normativa sobre esta materia ha sufrido un cambio radical con el desarrollo de las tecnologías *big data*, ya que se deben categorizar, modelar y mapear los datos a medida que estos son capturados y almacenados en tiempo real, con el inconveniente de que puede tratarse de datos no estructurados o semiestructurados que dificultan el análisis.

La rapidez con la que circula la información hace muy difícil la implementación de procesos de verificación de calidad, por lo que se hace necesario que se desarrollen nuevas metodologías y herramientas adecuadas para el análisis. La recolección de grandes volúmenes de datos y su posterior análisis generan preocupaciones sobre la privacidad y la confidencialidad de la información. Por este motivo, el principal desafío que enfrenta el *big data* es la generación de valor garantizando la seguridad de los datos y la privacidad de las personas a través de políticas preventivas y mitigadoras del riesgo. Tradicionalmente, las organizaciones utilizaban métodos de desidentificación, tales como la anonimización, la seudonimización, el cifrado, la codificación de claves y la fragmentación de datos (Tene & Polonetsky, 2012) para distanciar los datos de las identidades reales y permitir su análisis. Sin embargo, en los últimos años se ha demostrado que, en ocasiones, los datos anónimos pueden volver a identificarse y atribuirse a individuos, por lo que la privacidad de la información sigue siendo un inconveniente latente.

En este contexto, Paul Ohm (2009) observó que la ciencia de la identificación interrumpe el panorama de la política de privacidad al socavar la fe que hemos puesto en la anonimización. Bajo este concepto, todos los datos deben ser tratados como

personalmente reidentificables y por lo tanto sujetos al marco regulatorio. Por su parte, Betsy Masiello y Alma Whitten (2009) señalan que la información anónima siempre acarrea algún riesgo de reidentificación. Los riesgos de privacidad más urgentes existen solo si hay certeza en la reidentificación, es decir, si la información puede ser fácilmente autentificada. Cuando existe incertidumbre en la reidentificación, no se puede conocer con exactitud si la información realmente corresponde a un individuo en particular, lo que dificulta la implementación de un marco regulatorio.

Las leyes de privacidad y protección de datos se basan en el control individual sobre la información y sobre principios como la minimización de los datos. Sin embargo, no está claro que la minimización de la recopilación de información sea siempre una práctica del enfoque a la privacidad en la era de los grandes datos. Los principios de privacidad y la protección de datos deben equilibrarse con valores sociales adicionales, como la seguridad nacional y aplicación de la ley, la protección del medio ambiente y la eficiencia económica. Se podría pensar en una matriz de riesgos, teniendo en cuenta el valor de los diferentes usos de los datos frente a los riesgos potenciales a la autonomía individual y la privacidad. Donde los beneficios del uso prospectivo de datos superen los riesgos de privacidad, asumiendo la legitimidad del procesamiento, incluso si las personas se niegan a dar su consentimiento (Tene & Polonetsky, 2012).

Los formuladores de políticas también deben abordar la función del consentimiento en el marco de la privacidad. Actualmente, las actividades de procesamiento y análisis de información se basan en el consentimiento de los usuarios. Sin embargo, las personas no están en condiciones de tomar decisiones responsables sobre sus datos personales, por una parte, por los sesgos cognitivos en los potenciales usos del *big data* y por otro lado por la complejidad de los grandes volúmenes de información. Es imposible predecir las formas en que un conjunto de información puede ser analizada para entender un problema hasta que verdaderamente exista dicho problema. Este hecho aplica tanto para los que brindan la información como para quienes la analizan. Un ejemplo, lo constituye cuando se construyó el motor de búsqueda Google, ningún ingeniero pensó en los usos que tiene hoy en día y tampoco ningún usuario pensó que la información iba a ser capaz de predecir epidemias y enfermedades como los brotes de gripe (Masiello & Whitten, 2009).

Turow, Hoofnagle, Mulligan, & Good (2007) han demostrado que cuando los usuarios de un servicio o consumidores ven el término “políticas de privacidad” creen que su información personal estará protegida de maneras especificas; en particular suponen que un sitio web que publicita una política de privacidad no compartirá públicamente su información personal, cuando en realidad esto no siempre es así. Las políticas de privacidad sirven más a menudo como descargos de responsabilidad para las empresas que como garantía de privacidad para los usuarios. Un enfoque de consentimiento expreso y minimización de datos, con poca agregación de valor de los usos de datos, podría poner en peligro la innovación y los avances sociales beneficiosos.

Los defensores de la privacidad y los reguladores de datos menosprecian la era de los macrodatos a medida que observan la facilidad con la que pueden recopilarse a través de procesadores y almacenare de forma ilimitada. Sin embargo, los beneficios del *big*

data apuntan a ser mayores a los riesgos potenciales que puede generar, por este motivo permanecer al margen de esta revolución no constituye una solución viable.

En los últimos años, existe una clara tendencia de las compañías financieras a vincular el otorgamiento de crédito a técnicas algorítmicas de aprendizaje automático proveniente de información no estructurada como la provista en las redes sociales. En particular, dicha tendencia puede verse en mayor volumen en las nuevas *Fintech* como es el caso de Mercado Libre, que otorga préstamos a individuos que no tienen historial crediticio, basado en técnicas de *Machine Learnining*. Si bien los beneficios que pueden obtenerse resultan elevados, ya que personas y Pymes sin historial crediticio se les dificulta o incluso se le niega el acceso al primer crédito por las estructuras bancarias tradicionales, la incorporación de datos no estructurados ofrece alternativas prometedoras, que nunca antes fueron imaginadas en el sector financiero. Dicha política resulta un avance fundamental en la gobernanza financiera, pero, a su vez, es necesario tener en cuenta que existe una fina línea que separa el beneficio potencial del riesgo que implica utilizar información "privada" para la toma de decisiones crediticias.

La irrupción del *big data* está revolucionando las estructuras financieras tradicionales. A la vez que se incorpora nueva información no estructurada en la toma de decisiones, resulta necesario que las políticas de regulación y protección de datos personales acompañen el proceso de mejora y que no queden aisladas del mismo. Si bien es cierto que la normativa llega después del hecho a regular, las políticas de regulación financiera se encuentran atrasadas en materia normativa.

Estamos viviendo una tercera revolución industrial y con ella la restructuración de una sistema financiero más trasparente, más justo e inclusivo, que beneficiará a individuos y a Pymes que necesiten financiamiento. Lo que determinará el grado de expansión del mismo será la capacidad y voluntad de los gobiernos por adoptar regulaciones que permitan a la tecnología evolucionar para el bien común.

En el siguiente capítulo se desarrollará un modelo de regresión logística convencional con el objetivo de predecir la probabilidad de *default* de una cartera de empresas del índice S&P 500. Luego se volverá a calibrar el modelo incorporando información no estructural, en particular través de la incorporación de coeficiente adicional que mida el sentimiento que perciben los usuarios de la red social *Twitter* sobre el desempeño de la empresa. El trabajo será desarrollado en un marco responsable de privacidad de la información, manteniendo el anonimato de los tweets y eliminado todo contenido personal que pueda permitir una reidentificación con el usuario.

Resulta necesario destacar que el trabajo está focalizado en empresas que cotizan en bolsa debido a que la información financiera es pública. Dichas empresas tienen historial crediticio y a su vez cuentan con fácil acceso al crédito. Sin embargo, la muestra será utilizada como caso ilustrativo para explicar como la información no estructurada como es el desempeño reputacional de la empresa en la red social Twitter puede impactar en la probabilidad de *default*.

CAPÍTULO 3: El modelo

3.1 Introducción

En el siguiente capítulo se implementará un modelo de forma reducida utilizando la técnica de regresión logística para calcular la probabilidad de *default* de una cartera de empresas que cotizan en el índice *S&P 500*. En un segundo apartado, se volverá a calibrar el modelo, pero incorporando una variable adicional que realice un análisis de sentimiento de la información contenida en Twitter. Al finalizar, se analizarán los dos modelos a través de diversas técnicas estadísticas y se sacarán conclusiones al respecto.

3.2. Modelo 1: Modelo de regresión logística sin análisis de sentimiento

Tal como se explicó anteriormente en el capítulo uno, la regresión logística constituye un método para ajustar la curva de regresión, $y = f(x)$, cuando la variable y es una variable categórica. Es decir que constituye un conjunto de procesos estadísticos que miden la relación entre la variable dependiente categórica y una o más variables independientes mediante la estimación de probabilidades a través de una función logística. El modelo se utiliza para predecir el valor de la variable y, dado un conjunto de predictores x. Los predictores tienen la característica que pueden ser continuos, categóricos o una combinación de ambos.

A continuación, se implementará un modelo de regresión logística para calcular la probabilidad de *default* de empresas que cotizan en bolsa. La variable categórica y, en general puede asumir diferentes valores. En este caso se supondrá que asume el valor 1 o 0, siendo 1 el caso de *default* y 0 el caso de no *default.* En el presente trabajo, el *default* es definido como la baja en la calificación crediticia de las empresas. Para determinar dicha caída crediticia se toma en consideración el precio de las acciones en el mercado. Dicha suposición parte de la hipótesis de los mercados eficientes que establece que el precio de mercado de la acción incorpora toda la información que tiene el mercado sobre el desempeño de la empresa (Kealhofer, S, 2003). Por lo tanto, si el precio resume toda la información de mercado, una caída en el precio puede asociarse un descenso en la calificación crediticia y por lo tato permite ser un aproximador bastante certero del *default*.

De esta manera, entendiendo al *default* como una caída en la calificación crediticia generada por la caída en el precios de mercado se estableció como umbral de corte el 1,8%. Es decir, para los casos en los que las empresas registren una caída en el precio superior al 1,8% trimestral, la variable y asumirá el valor de 1, que corresponde al *default*; y en los casos en los que la variación trimestral de los precios de las acciones se mantenga o registre una caída inferior al 1,8%, la variable y asumirá el valor de 0, correspondiente al caso de no *default*.

Para la construcción de la variable y se realizó una búsqueda de los precios de cierres trimestrales de las acciones de las 505 empresas, desde el último trimestre 2017 hasta el primer trimestre 2019, es decir de 6 trimestres consecutivos. Luego, se realizaron las variaciones logarítmicas y con dichos valores se construyó un histograma de frecuencias,

seleccionando una variación de corte superior o igual -1,8% para la determinación del evento de *default*. Es decir, a la variable y se le asignó el valor de 1 en el caso en que la empresa tuviera una caída en el precio superior al 1,8% y un valor de 0 para el caso contrario. El análisis descriptivo de los datos arrojó los siguientes resultados:

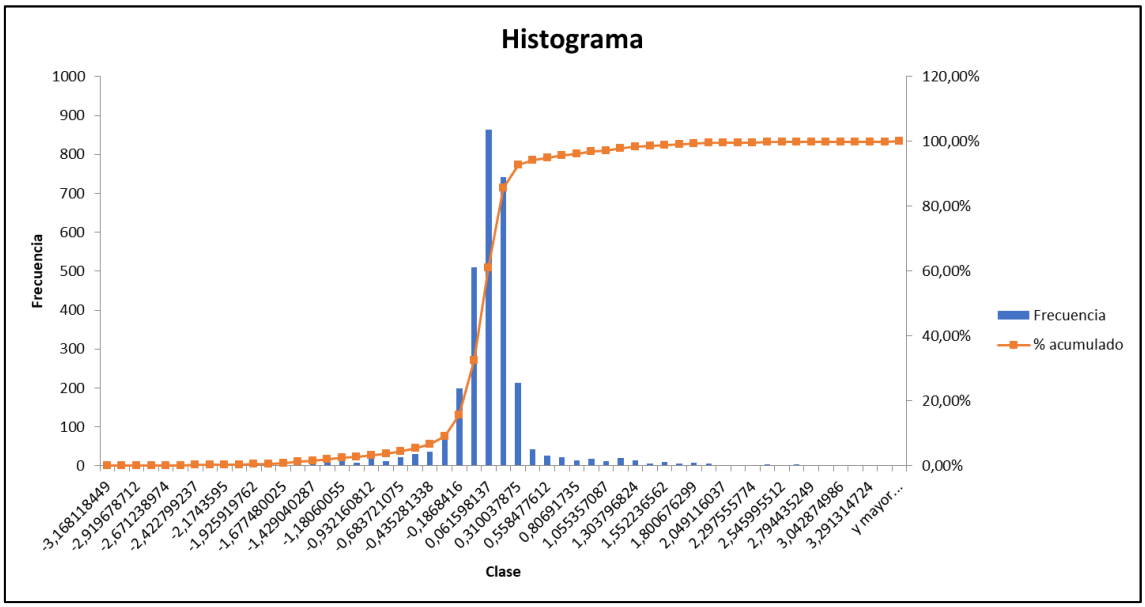
Tabla 1: Análisis descriptivo

Estadística descriptiva	
Media	0,001826908
Error típico	0,008268414
Mediana	0,017053715
Moda	0,181127618
Desviación estándar	0,454236306
Varianza de la muestra	0,206330622
Curtosis	12,1841892
Coeficiente de asimetría	0,037430347
Rango	6,707872911
Mínimo	-3,168118449
Máximo	3,539754461
Suma	5,513609362
Cuenta	3018

Fuente: Elaboración propia

El Gráfico 2 muestra el histograma de frecuencias de las variaciones logarítmicas de los precios:

Gráfico 2: Histograma de frecuencias



Fuente: Elaboración propia

Para determinar la probabilidad de *default* de las empresas se consideró el precio de las acciones en el mercado, partiendo de la hipótesis de mercados eficientes que establece que el precio de mercado de la acción incorpora toda la información que tiene el mercado sobre el desempeño de la empresa, lo que permite calcular el riesgo de *default* (Kealhofer, S, 2003).

Para la elaboración de la base de datos se seleccionaron 505 empresas del índice S&P 500. Las mismas fueron seleccionadas de forma aleatoria. De cada una de ellas se extrajo un total de siete índices y ratios de Thomson Reuters de apalancamiento, de rentabilidad y de liquidez. Los mismos son:

Ratio de rentabilidad:

Flujos de fondo/Ventas: Fondos de Operaciones / Ventas Netas o Ingresos * 100
Margen neto: Ingreso neto - resultado final/ Ventas o ingresos netos *100
Para el caso de las compañías de seguros si el ingreso neto – resultado final no está disponible, se sustituye el excedente del titular de la póliza.

Ratio de apalancamiento:

Deuda total/Capital de accionistas comunes: (Deuda de largo plazo + Deuda a corto plazo y porción actual de la deuda a largo plazo) / Capital de accionistas comunes*100.
Capital de accionistas comunes/Activos totales
Activos totales/ Ratio de capital común
Deuda neta: La deuda neta representa la deuda total menos el efectivo. El efectivo no solo representa el dinero líquido, sino también el vencimiento de los bancos y para los bancos, el efectivo para compañías de seguros y las inversiones a corto plazo para todas las demás industrias.

Ratio de liquidez ratio:

Deuda de largo plazo: Representa todas las obligaciones financieras que devengan intereses, excluyendo los montos adeudados dentro de un año. Se muestra neto de prima o descuento. Incluye, pero no está restringido a: Las hipotecas, bonos, debentures, deuda convertible, bono de fondos hundidos, sobregiros bancarios a largo plazo, notas a largo plazo, cuentas a largo plazo, préstamos a mediano plazo y regalías a largo plazo.

Las variables seleccionadas fueron elegidas teniendo en cuenta dos criterios principales, en primer lugar, se consideraron los ratios que mayor influencia tienen el riesgo de crédito, como es el caso de los ratios de apalancamiento, liquidez y rentabilidad. En segundo lugar, para su elección, se tuvo en cuenta que fueran similares a los ratios presentados en el modelo *Z Score* de Altman, pero sujetos a la disponibilidad de la plataforma Thomson Reuters.

3.2.1 Derivación de la ecuación de regresión logística del modelo

La regresión logística forma parte de una clase más grande de algoritmos conocidos como Modelo Lineal Generalizado (GLM). El GLM generaliza la regresión lineal al permitir que el modelo lineal esté relacionado con la variable de respuesta a través de una función de enlace y al permitir que la magnitud de la varianza de cada medición sea una función de su valor predicho. Nelder y Wedderburn (1972) propusieron este modelo como una forma de unificar otros modelos estadísticos, como la regresión lineal, la regresión logística y la regresión de Poisson. Formularon un método de mínimos cuadrados iterativamente ponderados para la estimación de máxima verosimilitud de los parámetros del modelo. La estimación de máxima probabilidad sigue siendo popular y es el método predeterminado en muchos paquetes de computación estadística como *R Studio y Python*.

La ecuación fundamental del modelo lineal generalizado puede determinarse de la siguiente forma:

$$g(E(Y))) = \beta_0 + \beta_1 X_1 + \cdots + \beta_i X_i \qquad (3.1)$$

Siendo $g()$ la función que vincula la esperanza de y con los predictores lineales $x_1, \dots, x_n$. $E(y)$ es la esperanza de la variable Y, y $\beta_0 + \beta_1 X_1 + \cdots + \beta_i X_i$ es el predictor lineal, siendo β_0 β_1, β_i los estimadores que se busca predecir.

Cabe destacar que GLM no asume una relación lineal entre variables dependientes e independientes. Sin embargo, si asume una relación lineal entre la función de enlace y las variables independientes en el modelo *logit.* La variable dependiente no necesita ser distribuida normalmente, a la vez que los errores deben ser independientes, pero no distribuidos normalmente.

En la regresión logística interesa calcular la probabilidad de la variable dependiente y. Como se describió anteriormente $g()$ es la función de enlace. Esta función se establece utilizando dos componentes: la probabilidad de éxito (p) y probabilidad de fracaso ($1 - p$). Como se trata de probabilidades, p debe ser mayor o igual a cero y menor o igual que uno.

Como la probabilidad debe ser positiva, se puede expresar la ecuación lineal en forma exponencial. Para cualquier valor, el exponente de la ecuación nunca será negativo.

$$p = \exp(\beta_0 + \beta_1 X_1) = e^{(\beta_0 + \beta_1 X_1)} \quad (3.2)$$

Para lograr que la probabilidad sea menor que uno debemos dividir p por un número mayor que p. Esto puede realizarse de la siguiente manera:

$$p = \exp(\beta_0 + \beta_1 X_1))/\exp(\beta_0 + \beta_1 X_1) + 1 = \frac{e^{\beta_0 + \beta_1 X_1}}{e^{\beta_0 + \beta_1 X_1}} + 1 \qquad (3.3)$$

Utilizando las ecuaciones (3.1), (3.2) y (3.3) podemos redefinir la probabilidad como:

$$p = e^y/1 + e^y \qquad (3.4)$$

Siendo ésta la función logística y p la probabilidad de éxito.

De forma análoga, $1 - p$ es la probabilidad de fracaso y puede escribirse como:

$$q = 1 - p = 1 - \left(\frac{e^y}{1} + e^y\right) \quad (3.5)$$

Al dividir entre sí la ecuación (3.4) y (3.5) obtenemos:

$$log\left(\frac{p}{1} - p\right) = y \quad (3.6)$$

Siendo esta última la función de enlace. La transformación logarítmica de la variable de resultado nos permite modelar una asociación no lineal de manera lineal.
Sustituyendo el valor de la variable y en la ecuación (3.1) obtenemos:

$$logit\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 \quad (3.7)$$

Para el caso en el que se tenga más de un regresor, la ecuación anterior puede expresarse como:

$$logit\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 X_{1,i} + \cdots + \beta_k X_{k,i} \quad (3.8)$$

Al cociente $\frac{p_i}{1-p_i}$ se lo conoce como *odds-ratio*. Por tanto, los coeficientes del modelo logit se interpretan como el logaritmo del *odds-ratio*.

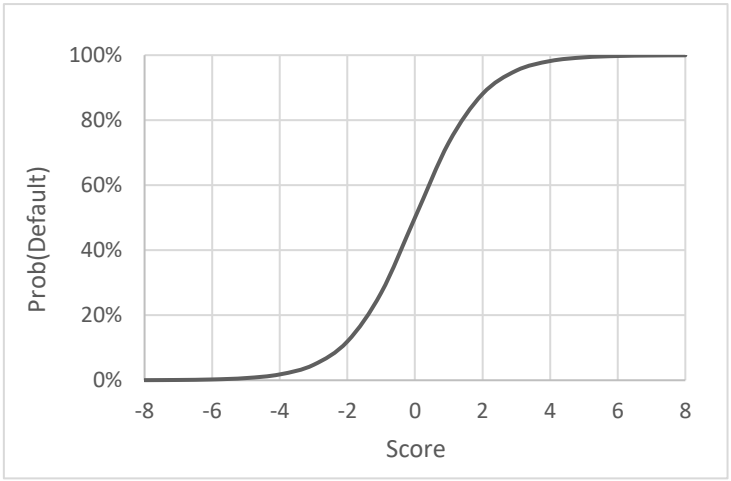
Cada elemento particular de X_i, puede ser ajustado para todo *i* obteniéndose una variable independiente en el modelo. Los parámetros desconocidos β_k son usualmente estimados a través de máxima verosimilitud.

La interpretación de los estimadores del parámetro β_i indica los efectos aditivos en el logaritmo de la variable y (*odds-ratio*) para una unidad de cambio en la *j-esima* variable explicativa. En el caso de una variable explicativa dicotómica, el modelo tiene una formulación equivalente dada por:

$$pi = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_{1,i} + \cdots + \beta_k x_{k,i})}} \quad (3.9)$$

Esta es la ecuación utilizada en la regresión logística. A continuación, en el Gráfico 3 se muestra la visualización de un modelo logístico típico, donde tal como puede observarse la probabilidad nunca se encuentra por debajo de 0 ni por encima de 1.

Gráfico 3: Representación gráfica de una Regresión Logística



Fuente: Elaboración propia

Ecuación del modelo:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 + \beta_7 x_7 \quad (3.10)$$

Siendo:

β_0=$Intercepto\ del\ modelo$

$x_1 = Flujo\ de\ fondos/Ventas$

$x_2 = Deuda\ de\ largo\ plazo$

$x_3 = Margen\ neto$

$x_4 = Deuda\ neta$

$x_5 = Deuda\ total/Capital\ de\ los\ accionistas\ comunes$

$x_6 = Capital\ de\ los\ aciconistas\ comunes/Activos\ totales$

x_7=$Activos\ totales/Ratio\ de\ capital\ común$

$\beta_1, \beta_2, \beta_3, \beta_4, \beta_5, \beta_6, \beta_7$
$= Coeficientes\ asociados\ a\ las\ respectivas\ variables\ indep.$

La variable y fue definida de la siguiente manera:

Siendo:

$$y = \ln(P_1/P_0) \begin{cases} \leq -1{,}8\% & \therefore \quad y = 1 \\ > -1{,}8\% & \therefore \quad y = 0 \end{cases} \quad (3.11)$$

Donde:

$y = 1 \sim Default$

$y = 0 \sim No\ Default$

3.2.2 Implementación de la regresión logística binaria en R

La herramienta R posibilita ajustar con mayor facilidad un modelo de regresión logística. La función utilizada para dicho fin, se denomina glm(). A continuación, se procederá a explicar con mayor detalle el proceso de implementación de la regresión logística utilizando esta herramienta de programación.

El proceso de limpieza y análisis de datos

Cuando se extrae una base de datos, es necesario tener en cuenta que muchas veces la información puede encontrarse incompleta, lo que resulta un inconveniente si el objetivo es realizar una regresión logística. Para el caso particular de estudio, cuando se extrajeron los ratios financieros de la plataforma de Thomson Reuters, muchos de ellos no se encontraban calculados para todos los períodos bajo análisis. Por este motivo, como primer paso se procedió a completar los índices y ratios faltantes calculando un promedio de los datos de los trimestres del año.

Luego se cargó la base de datos completa utilizando la función **read.csv()**. Se utilizó la función **sapply()** para verificar que no haya quedado ninguna fila vacía de contenido. Luego se eliminaron la columna id y años de la planilla. A continuación, se le indicó a R que interprete a la variable *default* como una variable categórica y por lo tanto como predictor del modelo.

Modelo de ajuste

Para hacer la regresión logística los datos fueron divididos en dos partes, un conjunto de entrenamiento *(train)* y un conjunto de prueba *(test).* El conjunto de entrenamiento se utilizó para ajustar al modelo que luego se probó durante el conjunto de pruebas.

Luego se corrió el modelo con la función **glm()**, especificando **family=binomial**. Se realizó la regresión sobre todos sus regresores, para determinar cuáles de ellos resultaban significativos. Al utilizar la función **summary()** se obtienen los resultados del modelo, devuelve los coeficientes beta, el error estándar, el valor Z y el *p-value*[22].

También se obtiene el criterio de información de Akaike (AIC). La misma es una medida de la calidad relativa de un modelo estadístico para un conjunto dado de datos. Esta medida de ajuste penaliza el modelo por el número de coeficientes. Por lo tanto, siempre se preferirá un modelo con un AIC mínimo.

El criterio de información de AIC, a diferencia de otros criterios de decisión estadísticos, como el R cuadrado, no solo recompensa por la bondad de ajuste del modelo, sino que a su vez penaliza la sobre parametrización del mismo, por este motivo, se selecciona dicho criterio de decisión.

[22] P-valor, según su traducción al español.

El criterio de información de Akaike puede expresarse a través de la siguiente ecuación:

$$AIC = 2k - 2\ln(L) \quad (3.12)$$

Siendo:

$k = Número\ de\ parámetros\ del\ modelo$
$L = Máximo\ valor\ de\ la\ función\ de\ verosimilitud\ para\ el\ modelo\ estimado$

Interpretación de resultados del modelo de regresión logística

Para obtener la tabla de desviación nula y residual se aplicó **anova()**. La desviación nula indica la respuesta predicha por el modelo que tiene como regresor solo el intercepto β_0. Cuanto más bajo sea el valor, mejor será el modelo. El desvío residual indica la variable respuesta predicha por el modelo cuando se agregan variables independientes. Cuanto más bajo sea el valor, mejor será el modelo. De forma análoga, la diferencia entre la desviación nula y la desviación residual muestra como se está desempeñando nuestro modelo en comparación con el modelo nulo. Cuanto más amplia sea la brecha mejor es el modelo.

Para la evaluación de la capacidad predictiva del modelo, se utiliza el parámetro **type='response'**. Con esta función R genera probabilidades de forma $P(y = 1|x)$. Nuestro límite de decisión será de 0,5. Si $P(y = 1|x) > 0{,}5$ entonces $y = 1$, de lo contrario $y = 0$. También podrían implantarse diferentes límites de decisión.

Con los valores calculados anteriormente se procedió a realizar la construcción de la matriz de confusión. La misma es una representación tabular de los valores actuales Vs los predichos por el modelo. Esta matriz resulta útil para encontrar el nivel de predicción del modelo. Se representa de la siguiente manera:

Tabla 2: Matriz de Confusión

		Valores Actuales	
		Positivos	**Negativos**
Valores predichos	**Positivos**	Verdaderos positivos (D)	Falsos negativos (C)
	Negativos	Falsos positivos (B)	Verdaderos negativos (A)

Fuente: Elaboración propia

Para calcular la precisión del modelo:

$$\frac{Verdaderos\ positivos + verdaderos\ negativos}{Verdaderos\ positivos + Verdaderos\ negativos + Falsos\ positivos + Falsos\ negativos} \quad (3.13)$$

$$= \frac{D + A}{A + B + C + D} \quad (3.14)$$

A partir de la matriz de confusión, la especificidad y la sensibilidad se pueden derivar de la siguiente manera:

Ratio de Verdaderos negativos (Especificidad)= $\frac{A}{A+B}$

Ratio de Falsos positivos (1-Especificidad) = $\frac{B}{A+B}$

Ratio de Verdaderos positivos (Sensibilidad)=$\frac{D}{C+D}$

Ratio de Falsos negativos= $\frac{C}{C+D}$

Como último paso se realizó el gráfico de la curva ROC Y se calculó el AUC (área bajo la curva) que constituyen las medidas típicas que se utilizan cuando se realiza una clasificación binaria. La curva ROC se genera al trazar la tasa de verdaderos positivos (TPR) contra la tasa de falsos positivos (FPR), en varias configuraciones del umbral. Es decir, resume el desempeño del modelo mediante la evaluación de las compensaciones entre la tasa de verdaderos positivos (sensibilidad) y la tasa de falsos positivos (1-especificiadad) Para trazar la curva ROC se asumió que $p > 0{,}5$ ya que es más importante para el análisis los casos de éxito. Bajo este análisis la curva ROC resume el poder predictivo para todos los valores posibles de $p > 0{,}5$.

El ROC de un modelo predictivo perfecto tiene un TPR igual a 1 y un FPR igual a 0. Una curva que presente estas características tocará la esquina superior izquierda de la gráfica Por otro lado el área bajo la curva (AUC), también conocida como el índice de precisión (A) o índice de concordancia, es el área bajo la curva ROC. Como regla general un modelo con una buena capacidad predictiva debe tener un AUC más cercano a 1 que a 0,5. Siendo 1 el número de perfecta predicción.

3.2.3 Análisis de resultados

A continuación, se realiza un análisis de los resultados obtenidos tras realizar la corrida del primer modelo logístico en R. En primer lugar, se procede a realizar una visualización del *data set*[23] con el objetivo de que quede ningún *missing value*[24] (Iustración 1, Anexo B) y luego se le pide a R que identifique a la variable *default* como variable categórica, de esta manera el *default* pasa de ser de una variable numérica a una variable expresada en niveles 0 y 1 (Ilustración 2, Anexo B).

En una segunda instancia se procede a calcular el desvío estándar para cada uno de los ratios. Los mismas miden la precisión con la que son estimados los parámetro, es decir, indican el "grado de confianza" de los estimadores. Tal como puede verse en el gráfico 4, se presenta la mayor desviación en el ratio 1, el cual representa el cociente entre el flujo de fondos y las ventas y la menor desviación en el ratio 5 asociado al cociente deuda total/capital de los accionistas comunes.

[23] Base de datos, según su traducción al español.
[24] Valor perdido, según su traducción al español.

Gráfico 4: Desvío estándar de las variables

```
> #Calcula el desvio estandar
> sapply(tabla, sd)
     Ratio 1      Ratio 2      Ratio 3      Ratio 4      Ratio 5      Ratio 6      Ratio 7
7.564465e+04 2.994367e+07 3.895534e+04 4.195273e+07 1.796938e+03 2.325386e+01 3.764544e+01
>
```

Fuente: Salida de R

Una vez realizado el análisis de descriptivo de los datos, se realiza la corrida de la regresión logística (Ilustración 3, Anexo B) obteniéndose los valores del intercepto β_0 y de los coeficientes $\beta_i,\ i = 1, \ldots, 7.$, tal como puede observarse en la columna una del gráfico 5.

Como puede observarse en el Gráfico 5, el intercepto β_0 resulta negativo, lo que implica que la regresión presenta pendiente negativa. Los signos de los coeficientes $\beta_1,\ \beta_4,\ \beta_5,\ \beta_6$ resultan negativos lo que implica que incrementos en $x_1\,,\ x_4\,,\ x_5, x_6\,,$ disminuyen la probabilidad de permanecer en *default* $(y = 1)$. Por el contrario, los signos de los coeficientes $\beta_2,\ \beta_3,\ \beta_7,$ resultan positivos, lo que implica que incrementos en $x_2,\ x_3,\ x_7,$ incrementan la probabilidad de permanecer en *default*.

Acorde al modelo, cabe destacar, que los coeficientes de los ratios en cuestión resultan inferiores al p valor asociado, por lo que existe una relación significativa entre las variables seleccionadas en el modelos y su capacidad de predicción.

Gráfico 5: Regresión logística

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.844e+00  2.224e-01 -12.786   <2e-16 ***
`Ratio 1`   -9.078e-03  5.280e-03  -1.719   0.0855 .
`Ratio 2`    1.070e-08  6.587e-09   1.624   0.1043
`Ratio 3`    4.432e-03  4.340e-03   1.021   0.3072
`Ratio 4`   -1.164e-08  5.860e-09  -1.987   0.0469 *
`Ratio 5`   -2.434e-04  1.017e-04  -2.393   0.0167 *
`Ratio 6`   -1.813e-03  4.634e-03  -0.391   0.6956
`Ratio 7`    1.187e-02  4.933e-03   2.407   0.0161 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 734.90  on 1882  degrees of freedom
Residual deviance: 723.23  on 1875  degrees of freedom
AIC: 739.23

Number of Fisher Scoring iterations: 14
```

Fuente: Salida de R

A su vez, en el gráfico 5, puede observarse el desvío nulo y el desvío residual del modelo. El desvío nulo muestra la respuesta predicha por el modelo, solo considerando el intercepto. Cuanto más bajo sea el valor, mejor será el modelo. Por el contrario, el desvío residual muestra la respuesta predicha por el modelo cuando se incluyen los predictores. De forma análoga, la diferencia entre la desviación nula y la desviación residual muestra como se está desempeñando nuestro modelo en comparación con el modelo nulo. Cuanto más amplia sea la brecha mejor es el modelo.

El desvío nulo para el modelo es de 7334,90 en 1882 grados de libertad y el desvío residual de 723,23 en 1875 grados de libertad, la diferencia entre ambos valores o también conocida como diferencia de residuos es de 11,6651. Los valores obtenidos resultan razonables. Por su parte, el criterio de información de Akaike (AIC), entendido como la medida de calidad relativa del modelo o bondad de ajuste resultó de 739,23.

En el Grafico 6, se muestra la matriz de confusión del modelo. La misma es utilizada para evaluar la capacidad predictiva del modelo. Para generar la matriz, R genera probabilidades de la forma $P(y = 1|x)$. El límite de decisión será 0,5. Si $P(y = 1|x) > 0{,}5$ entonces $y = 1$, de lo contrario $y = 0$.

Con los valores calculados anteriormente se procedió a realizar la construcción de la matriz de confusión. La misma es una representación tabular de los valores actuales Vs los predichos por el modelo. Esta matriz resulta útil para encontrar el nivel de predicción del modelo. A continuación, se muestra la matriz de confusión obtenida para el modelo 1.

Para la construcción de la matriz de confusión fue necesario separar la información en *data train* (información que se a entrenar) que es de un 80% y el *data test* (información que se va a testear) que es de un 20%. Se obtiene así, que el modelo es capaz de clasificar correctamente 0,9516 (95,16%) de los casos.

Gráfico 6: Matriz de confusión

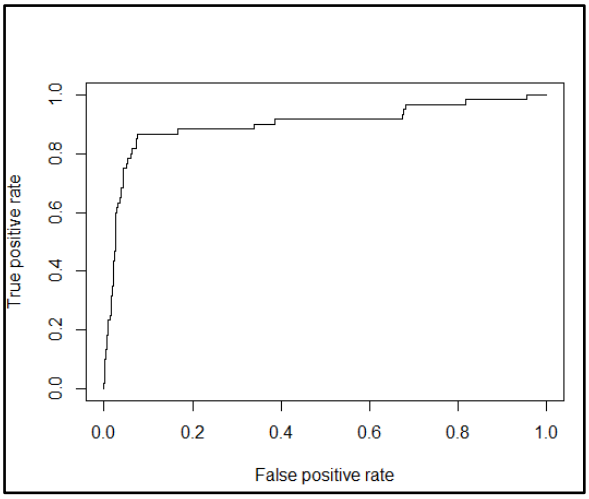
```
> #Para clasificacr el error matriz de confusion
> pred1<-ifelse(p1>0.5,1,0)
> tab1<-table(Predicted = pred1, Actual = train$Default)
> tab1
         Actual
Predicted    0    1
        0 1791   91
        1    0    1
>
```

Fuente: Elaboración propia

De esta manera utilizando la información de la matriz de confusión pueden obtenerse los diferentes ratios explicados en el apartado anterior. Entre los más destacados se encuentra, el ratio verdaderos positivos del modelo que es de 0,9516 y los falsos negativos 0,0483 o también conocida como probabilidad de error del modelo. Los resultados arrojados indican que el modelo predice de forma bastante certera los resultados.

A su vez se realizó el Gráfico 7 de la curva ROC, tal como puede observarse el área bajo la curva ROC, también conocida como AUC, arrojó un valor de 0,8161, siendo la perfecta predictibilidad de 1, podría decirse que el modelo explica con una precisión de aproximadamente el 80%. Los resultado de las salida del código de R pueden verse en la Ilustración 4 del anexo B.

Gráfico 7: Representación gráfica de la curva ROC



Fuente: Salida de R

Según los resultados obtenidos en el análisis anterior puede concluirse que el modelo explica con bastante precisión los resultados. A continuación, se implementará el modelo de riesgo crediticio utilizando la metodología de regresión logística, descripto anteriormente, pero en este caso incorporando una variable adicional que mida el análisis de sentimiento.

3.3 Modelo 2: Modelo de regresión logística con análisis de sentimiento

A continuación, se implementará el modelo de riesgo crediticio utilizando la metodología de regresión logística, descripto anteriormente, pero, en este caso, incorporando una variable cualitativa, que mida el análisis de sentimiento de la empresa en la red social Twitter, al modelo logístico de variables cuantitativas desarrollado en el apartado anterior.

3.3.1 Metodología para el análisis de sentimiento en R

Tal como se mencionó en el Capítulo 2, el análisis de sentimientos es definido como el proceso de identificar y categorizar las opiniones expresadas en un texto, especialmente para determinar si la actitud hacia un tema o producto en general es positiva, negativa o neutral. Para hacer un análisis de sentimiento, se atraviesa por cuatro etapas fundamentales. En primer lugar, es necesario realizar una recolección de datos, que en este caso se llevará a cabo a través de la API de Twitter. La segunda parte de la metodología es el preprocesamiento de los datos, en el cual se limpia la información y se transforma en el formato necesario para su análisis. La tercera etapa es el análisis de sentimiento que consiste en realizar una clasificación de los tweets de acuerdo al sentimiento con el que fueron expresados por el usuario. La última etapa es la predicción, a través de la construcción de un indicador. A continuación, se procederá a explicar con mayor detalle cada una de las etapas antes mencionadas.

Para la recolección de los tweets fue necesario, en primer lugar, crear una aplicación de Twitter. Luego, se ingresó en el siguiente link: http://apps.twitter.com/, se seleccionó la opción “Crear una nueva app” y se completó el formulario de solicitud propuesto por Twitter. Allí se seleccionó el nombre de la aplicación, luego se accedió a la sección *“Keys and Access Tokens”* donde se encuentran todas las claves que se usaron luego en el R script y que autorizan a R a acceder a la API de Twitter.

Con la ayuda de la API de Twitter es posible extraer tweets a gran escala y de forma automática. En este caso se utilizó la versión *"TwitterAouth"* de la API pública. Twitter proporciona una gran cantidad de parámetros de filtrado para que se pueda obtener un conjunto bien definido de tweets. Dentro de los principales filtros de la opción **searchTwitter()** se incluye el idioma, la fecha de inicio y fin de búsqueda, palabras claves del tweet, mensajes de un usuario específico y geolocalización. En este último aspecto cabe destacar que inicialmente la API de Twitter permitía encontrar tweets en forma de latitud y longitud donde el usuario había hecho pública su ubicación. Pero debido a problemas de seguridad y privacidad de la información de los usuarios, esta función se detuvo en el año 2012. Esto implica que la ubicación geográfica desde donde se creó el tweet ya no está disponible con el tweet. Lo que Twitter permite, por otro lado, es el uso de la ubicación como parámetro de filtrado, que se incorpora en el código de R, para la extracción de tweets. La latitud, la longitud y el radio son todos valores asignados al parámetro "geolocalización" en la construcción de consulta.

Para los fines del trabajo, se realizó una búsqueda indicando como palabra clave de búsqueda, el nombre de la empresa. Esta función se repitió para cada una de las empresas bajo análisis. Cabe destacar que la API gratuita de Twitter devuelve los tweets exclusivamente de la semana anterior a la fecha de búsqueda. Luego, a través de la función **write.csv**, los tweets fueron exportados desde R a una planilla de Excel.

Una vez obtenidos los tweets y exportados a una planilla de Excel, se procedió a realizar un preprocesamiento, manipulación, limpieza, formato y filtrado de información. En primer lugar, los tweets originales obtenidos desde la API de Twitter son limpiados de todo tipo de información irrelevante. Los elementos no gramaticales más habituales en el entorno web que deben eliminarse son:

Id de los usuarios (‘@’): En la red social, cada usuario dispone de un alias, precedido por el símbolo ‘@’ (ej: @alex). Este tipo de símbolos puede generar problemas para el análisis de sentimiento, dado que es un elemento no gramatical. Por este motivo a través de la plataforma R Studio fueron eliminados los nombres de los usuarios.

Hashtags #: Son términos que los usuarios incluyen en sus tweets precedido por el símbolo ‘#’, con el objetivo de etiquetar los mensajes. Al hacer *click* sobre el hashtag el usuario es redireccionado a un conjunto de tweets que contienen la misma etiqueta. Es habitual que el período de vida de los hashtag sea corto, dado que hacen referencia a eventos específicos y son situados al principio o al final del tweet.

Stopwords: Son palabras que no agregan información adicional al mensaje, suelen ser conectores, proposiciones y artículos como *“a”*, *“an”* y *“the”*.

Links URL (http://): que redireccionan a otros sitios web fuera de Twitter

Signos de puntuación y otros signos: Para poder realizar un correcto análisis de sentimiento es necesario eliminar de los tweets los signos de puntuación que resultan irrelevantes para el análisis (ej: , ; . & ($ " ¡! > % = * + - / ¿? [ ] {}).

Una vez limpiado el texto de contenido irrelevante se procedió a realizar el análisis de sentimiento para ello se utilizó el paquete "*Sentiment*"[25] de R. Este paquete a su vez se subdivide en dos paquetes principales. En primer lugar, el "Paquete *Sentiment*", el cual para poder utilizarlo es necesario instalar la versión *´devtools'*, este paquete utiliza un clasificado *Naive Bayes* entrenado sobre un léxico emociones y otro para la subjetividad de los textos. En segundo lugar, el paquete "*Sentiment Score"* Esta función ayuda a analizar un texto y clasificarlo en diferentes tipos de emociones, tales como: ira, asco, miedo, alegría, tristeza y sorpresa. La clasificación se puede realizar utilizando dos algoritmos: uno es el clasificador *Naive Bayes* entrenado por Carlo Strapparava y Alessandro Valitutti; el otro es sólo un algoritmo de votantes (*Simple Voter Algorithm*).

Con los tweets recolectados y limpiados de contenido irrelevante, se inicia la etapa de análisis de sentimiento, para ello se realizó un análisis lingüístico para el corpus y se construyó un clasificador de sentimiento que utiliza la recolección del corpus como datos de entrenamiento. En el trabajo se presentó un método para recolectar un corpus con diversos sentimientos, sin la necesidad de realizar un esfuerzo humano para la clasificación de los tweets. Con la herramienta de análisis de sentimiento de R los tweets fueron clasificados en 10 categorías diferentes: enojo (*anger*), esperanza (*anticipation*), disgusto (*disgust*), temor (*fear*), alegría (*joy*), tristeza (*sadness*), asombro (*surprise*), confianza (trust), negativo (*negative*) y positivo (*positive*). Para cada una de las categorías R asigna un Score.

Como los tweets recolectados no pueden exceder los 140 caracteres por reglas de la plataforma de R Studio, se analizó exclusivamente el sentimiento de los primeros 140 caracteres sin considerar los caracteres restantes. En este sentido, se asume que el sentimiento aplicado a los primeros 140 caracteres es equivalente a los caracteres restantes. En la investigación se utilizó el lenguaje inglés para la búsqueda de los tweets. Cabe destacar que pueden obtenerse tweets en otros lenguajes, pero el paquete para análisis de sentimiento de la plataforma R Studio se puede utilizar exclusivamente en inglés, por lo que si se selecciona otro lenguaje debería realizarse una adaptación al código original.

Para los fines de la tesis, se utilizó exclusivamente el score positivo y negativo, para luego evaluar el efecto que tienen en el riesgo de *default* corporativo. De este modo, se elaboró un ratio calculado como el efecto neto del score positivo menos el efecto del score negativo sobre la sumatoria del score positivo y negativo. En líneas generales se obtuvieron tres tipos de resultados: ratios positivos implican un predominio del sentimiento positivo, ratios negativos un predominio del sentimiento negativo y ratios

[25] Sentimiento, según su traducción al español.

nulos, implican igual magnitud de sentimiento positivo y negativo, por lo que el efecto final se considera neutro.

Índice de sentimiento:

$$S = \frac{Score\ positivo - Score\ negativo}{Total\ score\ positivo\ y\ negativo} \quad (3.15)$$

Resultados posibles:

$S > 0 = Predominio\ de\ sentimiento\ positivo$
$S < 0 = Predeominio\ de\ sentieminto\ negativo$
$S = 0 = Sentimiento\ neutro$

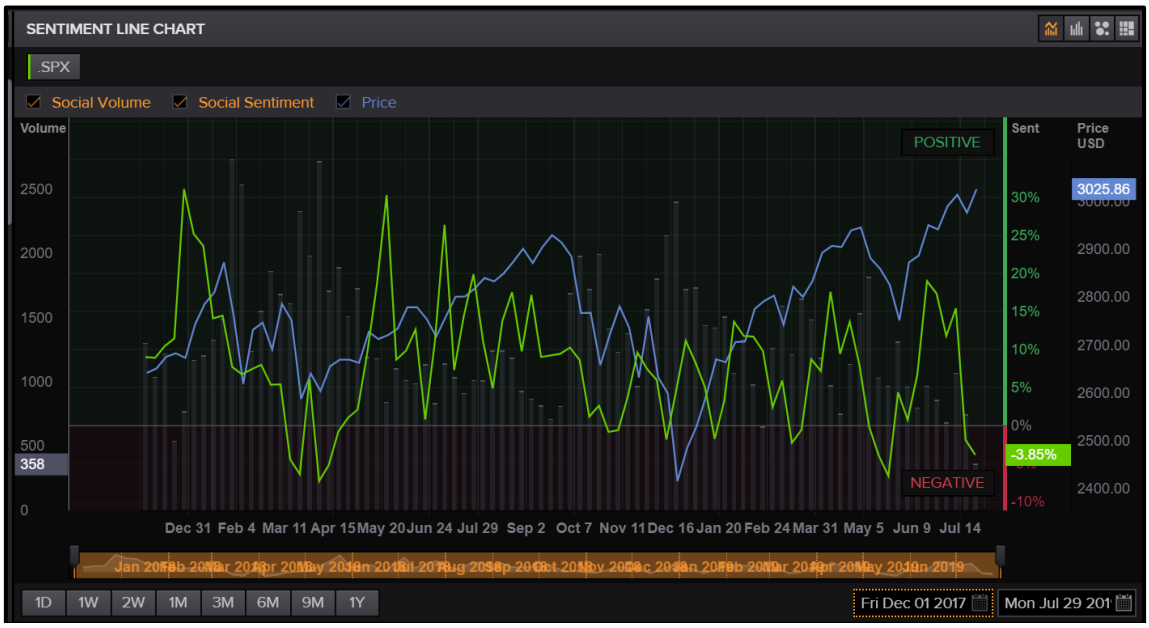
Para la incorporación del coeficiente adicional de análisis de sentimiento S a la regresión logística adicional se partió de la premisa de que existe una relación positiva entre el precio de las acciones, con su sentimiento o reputación existente en la red social Twitter. Dicha relación se fundamenta en la premisa de la economía del comportamiento que establece que las emociones y los estados de ánimo de los individuos afectan su proceso de toma de decisiones (Deaton & Muellbauer, 1980), lo que lleva a una correlación directa entre el sentimiento público y el sentimiento de mercado, entendiéndose el sentimiento público como aquel que se expresa a través de la red social Twitter y el sentimiento de mercado como la fluctuaciones del precio en el mercado de valores (Nguyen, Shirai, & Velcin, 2015).

La predicción del mercado de valores sobre la base de los sentimientos públicos expresados en Twitter constituye un campo de investigación reciente (Asur & Huberman, 2010). La hipótesis de mercados eficientes afirma que los precios del mercado de valores son impulsados en gran medida por información nueva y que siguen un proceso aleatorio *random walk*. Aunque esta hipótesis resulta ampliamente aceptada por la comunidad como un paradigma que gobierna los mercados en general, mucha gente ha tratado de extraer patrones para determinar la forma en la que los mercados se comportan y responden a estímulos externos.

Diversos estudios y análisis llegaron a la conclusión de que el estado del ánimo del público recopilado de Twitter puede estar correlacionado con los principales índices de mercado como Dow Jones Industrial Average Index (DJIA) y Standard&Poor 500 (SPX). En este contexto, si bien el precio de las acciones no puede predecirse de forma total a través del análisis de sentimiento, este último resulta un proxy cercano para que en conjunto con el precio de las acciones se pueda explicar con mayor exactitud la probabilidad de *default* corporativo.

En el siguiente gráfico se muestra la relación existente entre el precio del índice de *S&P 500* y el sentimiento derivado del análisis de los *Tweets*, extraídos de la plataforma Thomson Reuters. Tal como puede observarse, existe una correlación bastante perfecta entre ambas variables, los incrementos y decrementos del precio del índice con las opiniones púbicas expresadas en la red social *Twitter*.

Gráfico 8: Relación entre el Precio del índice y el Sentimiento de Twitter para S&P 500



Fuente: Elaboración propia, construido con la plataforma Thomson Reuters

A continuación, a modo de ejemplo, se presentan la salida del código en R que arrojó el procedimiento de análisis de sentimiento para el caso de la empresa Amazon.

En primer lugar, se realizó una búsqueda de tweets de la última semana colocando como palabra clave "Amazon" en la función **searchTwitter()**. Luego se realizó el análisis de sentimiento a los tweets y la salida de R arrojó el siguiente resultado:

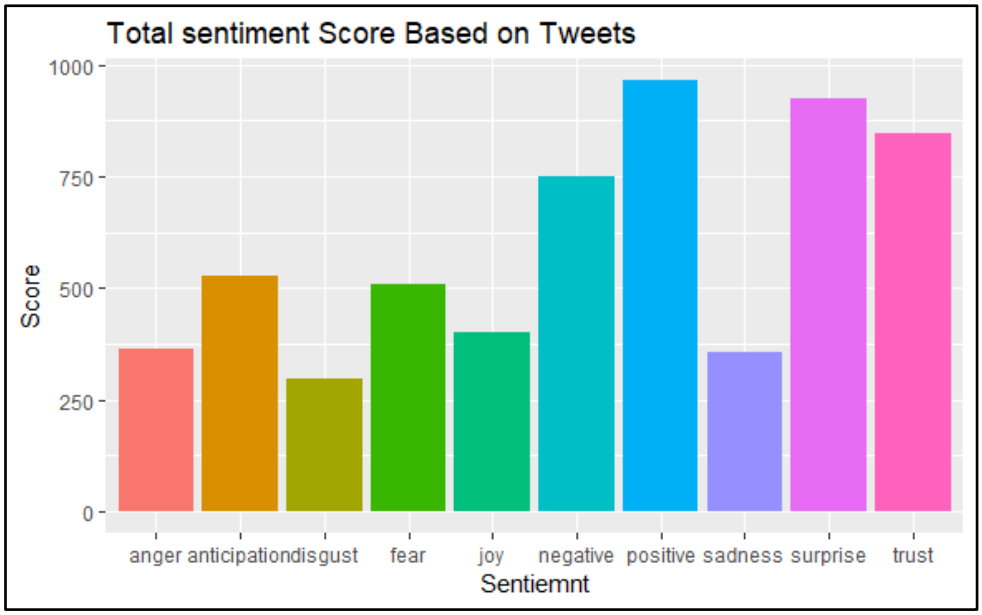
Gráfico 9: Score de sentimiento

	sentiment	Score
1	anger	363
2	anticipation	528
3	disgust	296
4	fear	509
5	joy	400
6	sadness	356
7	surprise	926
8	trust	845
9	negative	751
10	positive	965

Fuente: Salida de R

Por lo que puede observarse en Gráfico 9 precedente, el score positivo es de 965 y el score negativo de 751, por lo tanto, el índice de sentimiento es de 0,12 positivo (el mismo surge de (965-751)/1716). El mismo resultado puede expresarse de forma tabular a través de la función ggplot de R. La misma arrojó el siguiente grafico para la empresa Amazon:

Gráfico 10: Representación gráfica del Score de sentimiento



Fuente: Elaboración propia, construido con la plataforma R Studio

También puede crearse una nube de palabras a través de la opción **wordcloud()** para poder visualizar con mayor claridad aquellas palabras que más se utilizaron en los Tweets. Esta función fue solo utilizada de forma ilustrativa, pero no para los fines del trabajo.

Gráfico 11: Nube de palabras

Fuente: Elaboración propia, construido con la plataforma R Studio

En la siguiente sección se procederá analizar los resultados obtenidos de la incorporación del análisis de sentimiento a la regresión logística y se sacarán conclusiones al respecto.

3.3.2 Análisis de resultados

A continuación, se muestra el modelo de regresión logística incorporando el análisis de sentimiento. Para ello se modificó la variable y incorporando la variación logarítmica de la sumatoria de los precios y del score de sentimiento, expresado por la ecuación (3.11). Los ratios incorporados en el modelo 1 explicado anteriormente se mantuvieron sin modificaciones.

De este modo la variable y fue redefinida:

Siendo:

$$y = \ln\left(\frac{P_{t+1}+S_{t+1}}{P_t+S_t}\right) \begin{cases} \leq -1{,}8\% & \therefore \quad y = 1 \\ > -1{,}8\% & \therefore \quad y = 0 \end{cases} \quad (3.16)$$

$1 \leq S \leq -1$

Donde:

$y = 1 \sim Default$

$y = 0 \sim No\ Default$

Es decir que a la variación logarítmica del precio se le incorporó un coeficiente adicional y se volvió a determinar el histograma de frecuencias y el corte de *default* en -1,8.

A continuación, se realiza un análisis de los resultados obtenidos tras realizar la corrida del modelo logístico en R. En primer lugar, al igual que el modelo 1, presentado anteriormente, se procede a realizar una visualización del *data set* con el objetivo de que no quedara ningún *missing value* (Iustración 1, Anexo B) se le pide a R identifique a la variable *default* como variable categórica, de esta manera el *default* pasa de ser una variable numérica a una variable expresada en niveles 0 y 1 (Ilustración 2, Anexo B).

En una segunda instancia se procede a calcular el desvío estándar para cada uno de los ratios. Los mismos miden la precisión con la que son estimados los parámetro, es decir, indican el "grado de confianza" de los estimadores. Tal como puede verse en gráfico que sigue, se presenta la mayor desviación en el ratio 1, el cual representa el cociente entre el flujo de fondos y las ventas, y la menor desviación en el ratio 5, el cual representa el cociente entre la deuda total y el capital de los accionistas comunes. Este resultado resulta similar al modelo 1 debido a que las variables explicativas no fueron modificadas.

Gráfico 12: Desvío estándar de las variables explicativas

```
> #Calcula el desvio estandar
> sapply(tabla4, sd)
    Default      Ratio 1      Ratio 2      Ratio 3      Ratio 4      Ratio 5      Ratio 6
         NA 7.371122e+04 2.935094e+07 3.795938e+04 4.096960e+07 1.751539e+03 2.322966e+01
    Ratio 7
3.669177e+01
>
```

Fuente: Salida de R

Una vez realizado el análisis de descriptivo de los datos, se realizó la corrida de la regresión logística, obteniéndose los valores del intercepto β_0 y de los coeficientes β_i, $i = 1, \dots, 7.$, tal como puede observarse en la columna una del Gráfico 13.

Como puede observarse en el Gráfico 13, el intercepto β_0 resulta negativo, lo que implica que la regresión presenta pendiente negativa. Los signos de los coeficientes β_1, β_4, β_5, β_6 resultan negativos, lo que implica que incrementos en x_1, x_4, x_5, x_6, disminuyen la probabilidad de permanecer en *default* $(y = 1)$. Por otro lado, los signos de los coeficientes β_2, β_3, β_7, resultan positivos, lo que implica que incrementos en x_2, x_3, x_7, incrementan la probabilidad de permanecer en *default*. El resultado arrojado es igual que el modelo 1.

Acorde al modelo, cabe destacar, que los coeficientes de los ratios en cuestión resultan inferiores al p valor asociado, por lo que existe una relación significativa entre las variables seleccionadas en el modelos y su capacidad de predicción, tal como puede verse en el grafico que sigue.

Gráfico 13: Regresión logística

```
Call:
glm(formula = Default ~ ., family = binomial(link = "logit"),
    data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.9997  -0.2155  -0.1978  -0.1764   3.0759

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.438e+00  2.985e-01 -11.516  < 2e-16 ***
`Ratio 1`   -1.599e-02  6.168e-03  -2.592  0.00953 **
`Ratio 2`    4.751e-09  1.181e-08   0.402  0.68736
`Ratio 3`    2.147e-03  5.170e-03   0.415  0.67801
`Ratio 4`   -1.041e-08  1.043e-08  -0.999  0.31788
`Ratio 5`   -3.128e-04  1.110e-04  -2.819  0.00482 **
`Ratio 6`   -4.962e-03  6.481e-03  -0.766  0.44397
`Ratio 7`    1.597e-02  5.203e-03   3.069  0.00215 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 387.28  on 1882  degrees of freedom
Residual deviance: 371.37  on 1875  degrees of freedom
AIC: 387.37

Number of Fisher Scoring iterations: 15
```

Fuente: Salida de R

A su vez, en el gráfico 13, puede observarse el desvío nulo y el desvío residual del modelo. El desvío nulo muestra la respuesta predicha por el modelo, solo considerando el intercepto. Cuanto más bajo sea el valor, mejor será el modelo. Por el contrario, el desvío residual indica muestra la respuesta predicha por el modelo cuando se incluyen los predictores. De forma análoga, la diferencia entre la desviación nula y la desviación

residual muestra como se está desempeñando nuestro modelo en comparación con el modelo nulo. Cuanto más amplia sea la brecha mejor es el modelo.

El desvío nulo para el modelo es de 387,28 en 1882 grados de libertad y el desvío residual de 371,37 en 1875 grados de libertad, la diferencia entre ambos valores o también conocida como diferencia de residuos es de 15,91. Los valores obtenidos resultan razonables. Por su parte, el criterio de información de Akaike (AIC), entendido como la medida de calidad relativa del modelo o bondad de ajuste resultó de 387,37.

En el gráfico 14 se muestra la matriz de confusión del modelo. La misma es utilizada para evaluar la capacidad predictiva del modelo. Para generar la matriz, R genera probabilidades de la forma $P(y = 1|x)$. El límite de decisión será 0,5. Si $P(y = 1|x) > 0{,}5$ entonces $y = 1$, de lo contrario $y = 0$.

Con los valores calculados anteriormente se procedió a realizar la construcción de la matriz de confusión. La misma es una representación tabular de los valores actuales Vs los predichos por el modelo. Esta matriz resulta útil para encontrar el nivel de predicción del modelo. A continuación, se muestra la matriz de confusión obtenida para el modelo 2.

Gráfico 14: Matriz de confusión

```
> #Para clasificacr el error matriz de confusion
> pred1<-ifelse(p1>0.5,1,0)
> tab1<-table(Predicted = pred1, Actual = train$Default)
> tab1
         Actual
Predicted    0    1
        0 1843   39
        1    0    1
```

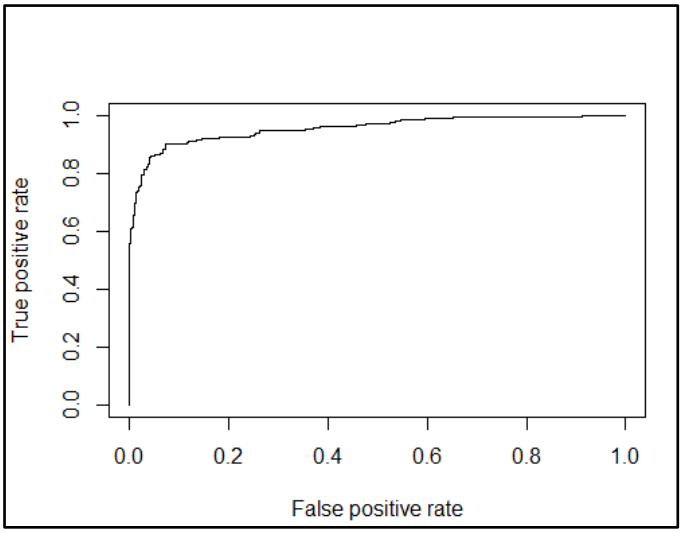
Fuente: Salida de R

Para la construcción de la matriz de confusión fue necesario separar la información en *data train* (información que se a entrenar) que es de un 80% y el *data test* (información que se va a testear) que es de un 20%. El modelo es capaz de clasificar correctamente 0,9792 (97,92%) de los casos, lo cual resulta superior al porcentaje de predicción del modelo 1, que era de un 95,16%.

De esta manera utilizando la información de la matriz de confusión puede obtenerse los diferentes ratios explicados en el apartado anterior. Entre los más destacados se encuentra, el ratio de especificidad del modelo que resulta ser de 1, el ratio de verdaderos positivos que es de 0,9792 y los falsos negativos 0,0207 o también conocida como probabilidad de error del modelo. Los resultados arrojados indican que el modelo predice de forma bastante certera los resultados.

A su vez se realizó el Gráfico 15 de la curva ROC, tal como puede observarse el área bajo la curva ROC, también conocida como AUC, arrojó un valor de 0,9554, siendo la perfecta predictibilidad de 1, podría decirse que el modelo explica con una precisión de aproximadamente el 90%. Los resultado de las salida del código de R pueden verse en la Ilustración 5 del anexo B.

Gráfico 15: Representación gráfica de la curva ROC



Fuente: Salida de R

Finalmente, en la última sección del capítulo se compararán los resultados obtenidos del modelo de la regresión logística sin análisis de sentimiento y la regresión logística incorporando el índice de sentimiento y se sacarán conclusiones al respecto.

3.4 Comparación de resultados

Comparando los resultados de ambos modelos, puede concluirse que el segundo modelo, es decir el que incorpora el índice de sentimiento en la variable explicada, arroja mejores resultados en términos estadísticos que el modelo 1. A continuación se muestra el resumen de los resultados obtenidos en ambos modelos.

Tabla 3. Comparación de resultados

Concepto	Modelo 1	Modelo 2
Desvío nulo	734,9	387,28
Desvío residual	723,23	371,37
Dif. de desvíos	11,6651	15,9278
Akaike	739,23	387,37
Prob. de error	0,0483	0,0207
P valor	0,1121	0,0207
AUC	0,8161	0,9554
Precisión del modelo	0,9516	0,9792

Fuente: Elaboración propia

El modelo 2 presenta un desvío nulo y residual más bajo que el modelo 1, en 347,62 puntos y 351,86 respectivamente. La desviación nula es preferible que sea la más baja posible e indica la respuesta predicha por el modelo que tiene como regresor solo el intercepto β_0. De forma análoga el desvío residual explica la variable predicha del modelo cuando se incorporan más variables, por lo que será conveniente que también sea lo más bajo posible. La diferencia entre la desviación nula y el desvío residual muestra cómo se está desempeñando nuestro modelo en comparación con el modelo

nulo, cuanto más amplia sea la brecha mejor será el modelo. En este caso resulta ser de 11,66 en el modelo 1 y de 15,92 en el modelo 2, resultando mayor en el segundo modelo, en 4,26 puntos.

Por otro lado, se calculó el criterio de información de Akaike (AIC), el cual proporciona un medio para la elección del modelo. El criterio establece que dado un conjunto de modelos, el modelo preferido será el que tenga el AIC más chico. En este caso el modelo 2 presenta un AIC de 387,37, mientras que el modelo 1 de 739,23. Por lo que el modelo 2 resulta preferible al modelo 1.

A su vez, la probabilidad de error del modelo 1 es de 0,04 y mientras que la del modelo 2 de 0,02 por lo que se comporta mejor. Finalmente, el área bajo la curva ROC, también conocida como el índice de precisión o índice de concordancia, resulta de 0,0816 en el modelo 1 y de 0,9554 en el modelo 2. Como regla general un modelo con una buena capacidad predictiva debe tener un AUC más cercano a 1 que a 0,5. Siendo 1 el número de perfecta predicción. Cabe destacar que a través de la realización de la matriz de confusión pudo concluirse que la precisión del modelo 1 es de 0,95 (95%), mientras que la precisión del modelo 2 es de 0,9702 (97,02%).

Por los motivos antes mencionados, puede concluirse que el modelo 2 resulta mejor en términos estadísticos que el modelo 1. De esta manera pude evidenciarse como la incorporación de un índice de sentimiento en la variable explicada, otorga mayor precisión al cálculo de la regresión logística del modelo para la predicción de la probabilidad de *default*.

Todo lo anterior parece indicar que el modelo que incorpora información no estructurada de textos obtenida en tiempo real desde Twitter parece reproducir mejores resultados que el modelo que no incorpora el análisis de sentimiento en la regresión.

4. Conclusiones

A modo de conclusión se considera relevante realizar una síntesis de las temáticas desarrolladas a lo largo del trabajo final de tesis, así como de los resultados obtenidos.

En primer lugar, en el primer capítulo, se desarrolló un recorrido por los modelos de riesgo de crédito más representativos tanto en la literatura académica como en el ámbito profesional. En primer lugar, se intrdujeron los modelos de univariados. Luego, se presentaron los modelos de forma reducida, tales como, el modelo de Altman o *Z Score* y los modelos de regresión logística. A continuación, se explicaron los modelos estructurales, entre los que se destacó el modelo de Merton como principal exponente y luego los modelo derivados tales como, el modelo de *Credit Portfolio Manager* de *KMV* Moody's y el modelo de *Credit Metrics* de JP Morgan. Finalmente, fueron presentados los modelos no paramétricos, en particular, el modelo de redes neuronales, el Support Vector Machine y los árboles de decisión.

Se concluyó el capítulo estableciendo que, entre todas las metodologías disponibles, para la calibración de los modelos de *credit scoring*, la revisión de la literatura sugiere que los modelos más utilizados en el mercado son los modelos *probit* y *logit*. Su predominio se debe a su sencillo funcionamiento e interpretación a la vez que su implementación permite arribar a resultados confiables a través de una metodología entendida por diversas áreas de una institución financiera. Sin embargo, cabe destacar que, a pesar de la proliferación de las numerosas metodologías utilizadas para la estimación de la probabilidad de *default* en los modelos de riesgo de crédito, el juicio del analista continúa siendo utilizado en la originación de créditos, en algunos casos expresado como un conjunto de reglas que la entidad aplica de manera sistemática para filtrar solicitudes o deudores. En la práctica, conviven de forma conjunta el juicio del analista con el modelo elegido por la Entidad.

Una vez introducidos los modelos de *credit scoring*, se procedió en el Capítulo 2 a introducir el concepto de *big data* y luego se desarrolló el concepto de *text mining* en la era del *big data* para finalmente vincularlo al área de riesgo de crédito. Del concepto de *text mining* se derivó el análisis de sentimiento como una herramienta que permite convertir información textual en información cuantitativa para su posterior análisis y procesamiento. En concreto, se introdujo la red social *Twitter* para el análisis de sentimiento y se concluyó el capítulo problematizando la importancia de la utilización del *big data* en un contexto de responsabilidad social, considerando sus principales riesgos y beneficios para la sociedad en su conjunto.

Para finalizar, se desarrolló un modelo de regresión logística convencional con el objetivo de calcular la probabilidad de *default* de una cartera de empresas que cotizan en bolsa. Luego se volvió a calibrar el modelo incorporando un coeficiente adicional que mide el sentimiento que perciben los usuarios de la red social Twitter sobre el desempeño de la empresa. Este modelo fue desarrollado en un marco responsable de privacidad de la información, manteniendo el anonimato de los Tweets y eliminado todo contenido personal que pueda permitir una reidentificación con el usuario.

Para el desarrollo del primer modelo se planteó una regresión logística sobre la base de una muestra de empresas del índice *S&P 500*. Para cada una de ellas se extrajo un total

de siete índices y ratios de Thomson Reuters de apalancamiento, de rentabilidad y de liquidez, en el cual la variable y fue explicada por las variaciones logarítmicas del precio de las acciones, partiendo de la hipótesis de mercados eficientes que establece que el precio de mercado de la acción incorpora toda la información que tiene el mercado sobre el desempeño de la empresa, lo que permite calcular el riesgo de *default*.

Luego se desarrolló el modelo dos incorporando en la variable explicada y un índice de sentimiento. Para ello, se partió de la premisa de que existe una relación positiva entre el precio de las acciones, con su sentimiento o reputación existente en la red social Twitter. Dicha relación fue fundada en la premisa de la economía del comportamiento que establece que las emociones y los estados de ánimo de los individuos afectan su proceso de toma de decisiones, lo que lleva a una correlación directa entre el sentimiento público y el sentimiento de mercado.

Como resultado, se concluyó que el modelo dos arrojó mejores resultados en términos estadísticos, corroborando así la hipótesis planteada como puntapié inicial de la tesis, que establecía que la incorporación de un índice de sentimiento cualitativo en los modelos de crédito tradicionales mejoraría la estimación de la probabilidad de *default* ya que incorporaría información no cuantificable sobre el funcionamiento del mercado bursátil.

De esta manera, el trabajo final de tesis deja planteada la posibilidad de la incorporación de técnicas de aprendizaje automático proveniente de información no estructurada, como la provista por las redes sociales en el sector financiero, particularmente en lo concerniente al otorgamiento de crédito. La generalización de esta tendencia que aún es incipiente ofrece beneficios prometedores. Resulta un avance fundamental en la gobernanza financiera que debería acompañar el proceso de mejora de la mano de normativas que regulen y al tiempo que potencien la expansión del mismo a nivel mundial.

Como futuras líneas de investigación se deja planteada la necesidad de realizar un análisis profundo de la mejora en los procesos regulatorios para la generalización de las buenas prácticas del *big data* en el sector financiero. A su vez, se plantea la posibilidad de la utilización de *big data* no solo como una práctica particular pensada en el otorgamiento de crédito, sino más bien, como tecnología incipiente que permitirá la restructuración del sistema financiero tradicional.

A la vez que se incorpora nueva información no estructurada en la toma de decisiones, resulta necesario que las políticas de regulación y protección de datos personales acompañen el proceso de mejora y que no queden aisladas del mismo, en este contexto se deja abierto el análisis de los potenciales riesgos que puede generar el uso del *big data* en un contexto no responsable y desregulado.

Actualmente, el sistema financiero se encuentra experimentando un cambio importante en la mejora y automatización de procesos. En este contexto el trabajo plantea una herramienta innovadora que puede implementarse en el sector financiero para el otorgamiento de crédito, incorporando información pública no confidencial. Dicha información permitiría la agregación de valor y la mejora en los modelos de scoring crediticio tradicionales utilizados en la práctica.

5.1 Anexo A

5.1.1 Código en R: Regresión logística

1) Limpieza y análisis de datos

```
#Regresión Logísitca
library(readxl)
prueba2 <- read_excel("C:/Users/Flavia/Desktop/mas para la tesis/empresa.xlsm")
View(Modelo1)
str(Modelo1)

summary(Modelo1)

Modelo1 <- Modelo1 [,-1]
View(Modelo1)

limpios$Default=as.factor(limpios$Default)
str(Modelo1)
```

2) El modelo de ajuste

```
set.seed(1234)
ind <- sample(2,nrow(Modelo1), replace = T,prob = c(0.8,0.2))
train<- Modelo1 [ind==1,]
test<- Modelo1 [ind==2,]
```

3) Implementación de la regresión logísitca 1

```
#Modelo 1
reg <- glm(Default ~.,family=binomial(link='logit'),data=train)
summary(reg)

p1<-predict(reg, train, type="response")
head(p1)
head(train)

pred1<-ifelse(p1>0.5,1,0)
tab1<-table(Predicted = pred1, Actual = train$Default)
tab1

table(train$Default)
table(test$Default)

1-sum(diag(tab1)/sum(tab1))

with(reg,pchisq(null.deviance-deviance, df.null-df.residual,lower.tail = F))

install.packages("gplots")
install.packages("ROCR")
```

```
library(gplots)
library(ROCR)
pred=prediction(p1,train$Default)

as.numeric(performance(pred,"auc")@y.values)

predictTrain=predict(reg,type="response")
ROCRpred=prediction(predictTrain, train$Default)
ROCRperf=performance(ROCRpred, "tpr","fpr")
plot(ROCRperf, colorize=FALSE, text.adj= c (-0.2,1.7))

dif_residuos <- reg$null.deviance - reg$deviance
paste("Diferencia de residuos:", round(dif_residuos, 4))

df <- reg$df.null - reg$df.residual
paste("Grados de libertad:", df)

p_value <- pchisq(q = dif_residuos,df = df, lower.tail = FALSE)
paste("p-value:", round(p_value, 4))

anova(reg, test = "Chisq")
```

El mismo código fue utilizado para el realizar la corría del modelo 2, para ello en donde dice “Modelo1”, fue reemplazado por “Modelo2”.

5.1.2 Código en R: *Text Mining* y análisis de sentimiento

```
install.packages("twitteR")
install.packages("ROAuth")
install.packages("plyr")
install.packages("stringr")
install.packages("ggplot2")
install.packages("wordcloud")
install.packages("RCurl")
install.packages("syuzhet")
install.packages("devtools")
install.packages("httr")
install.packages("sentimentr")
install.packages("httpuv")
install.packages("base64enc")
install.packages("tm")

library(twitteR)
library(ROAuth)
```

```
library(plyr)
library(dplyr)
library(stringr)
library(ggplot2)
library(httr)
library(wordcloud)
library(RCurl)
library(syuzhet)
library(devtools)
library(sentimentr)
library(openssl)
library(httpuv)
library(base64enc)
library(tm)

download.file (url='http://curl.haxx.se/ca/cacert.pem',destfile='cacert.pem')
reqURL <- 'https://api.twitter.com/oauth/request_token'
accessURL <- 'https://api.twitter.com/oauth/access_token'
authURL <- 'https://api.twitter.com/oauth/authorize'

consumerKey <- '__________'
consumerSecret <- '__________'
accesstoken <- '__________'
accesssecret <- '__________'
Cred <- OAuthFactory$new(consumerKey=consumerKey,
                         consumerSecret=consumerSecret,
                         requestURL=reqURL,
                         accessURL=accessURL,
                         authURL=authURL)

save (Cred, file='twitter authentication.Rdata')
load ('twitter authentication.Rdata')
setup_twitter_oauth (consumer_key = consumerKey, consumer_secret =
consumerSecret, access_token = accesstoken, access_secret = accesssecret)
some_tweets = searchTwitter("Amazon", n=1000, since = "2017-03-03", lang = "es")

length.some_tweets <- length(some_tweets)
length.some_tweets

some_tweets.df <- ldply(some_tweets, function(t) t$toDataFrame())

write.csv (some_tweets.df, "trump.csv")

some_txt = sapply(some_tweets, function(x) x$getText())

some_txt1 = gsub("RT|via)((?:\\b\\w*@\\w+)+)", " ", some_txt)
some_txt2 = gsub("http[^[:blank:]]+", " ", some_txt1)
```

```
some_txt3 = gsub("@\\w+", " ", some_txt2)
some_txt4 = gsub("[[:punct:]]", " ", some_txt3)
some_txt5 = gsub("[^[:alnum:]]", " ", some_txt4)

write.csv (some_txt5, "trump1.csv")

some_txt6 <- Corpus(VectorSource(some_txt5))

some_txt6 <- tm_map(some_txt6, removePunctuation)
some_txt6 <- tm_map (some_txt6, content_transformer(tolower))
some_txt6 <- tm_map(some_txt6, removeWords, stopwords("english"))

some_txt6 <- tm_map(some_txt6, stripWhitespace)

pal <- brewer.pal(8, "Dark2")

wordcloud(some_txt6, min.freq = 5, max.words = Inf, width= 1000, height=1000,
random.order = FALSE, color=pal)

my_sentiment <- get_nrc_sentiment(some_txt5)
Sentiment_Score <- data.frame (colSums(my_sentiment[,]))
names(Sentiment_Score) <- "Score"
Sentiment_Score <- cbind("sentiment"= rownames(Sentiment_Score),
Sentiment_Score)
rownames(Sentiment_Score) <- NULL
Sentiment_Score

ggplot(data = Sentiment_Score, aes(x = sentiment, y = Score )) +
  geom_bar(aes(fill = sentiment), stat = "identity") +
  theme(legend.position = "none") + xlab("Sentiemnt") + ylab("Score") +
  ggtitle("Total sentiment Score Based on Tweets")
```

5.2 Anexo B

5.2.1 Modelo 1: Modelo de regresión logística sin análisis de sentimiento

Ilustración 1: Visualización de la información

```
> #Descripcion del dataset y ver si hay missing values
> summary(tabla)
   Default          Ratio 1           Ratio 2             Ratio 3             Ratio 4
Min.   :0.0000   Min.   :   -199   Min.   :       -5   Min.   :-410000.0   Min.   :-105641000
1st Qu.:0.0000   1st Qu.:     12   1st Qu.:  2179000   1st Qu.:      5.2   1st Qu.:    876713
Median :0.0000   Median :     20   Median :  5304000   Median :     11.2   Median :   4049662
Mean   :0.0479   Mean   :   1709   Mean   : 13204818   Mean   :    602.3   Mean   :  11215633
3rd Qu.:0.0000   3rd Qu.:     32   3rd Qu.: 12108750   3rd Qu.:     20.4   3rd Qu.:  10327000
Max.   :1.0000   Max.   :3541500   Max.   :282031000   Max.   :1788900.0   Max.   : 511303000
   Ratio 5            Ratio 6           Ratio 7
Min.   :-56985.97   Min.   :-201.42   Min.   :-957.665
1st Qu.:    39.99   1st Qu.:  20.14   1st Qu.:   1.957
Median :    76.19   Median :  34.75   Median :   2.727
Mean   :    53.15   Mean   :  33.81   Mean   :   3.054
3rd Qu.:   132.21   3rd Qu.:  48.25   3rd Qu.:   4.362
Max.   : 51000.00   Max.   : 105.25   Max.   :1213.857
```

Fuente: Salida de R

Ilustración 2: Convertir el default en variable categórica

```
Classes 'tbl_df', 'tbl' and 'data.frame':       2338 obs. of  8 variables:
 $ Default: Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
 $ Ratio 1: num  2.91 33.96 20.37 27.16 31.81 ...
 $ Ratio 2: num  3969000 3973000 3981000 3986000 4012000 ...
 $ Ratio 3: num  -9.34 30.18 9.78 27.24 22.78 ...
 $ Ratio 4: num  -9.80e+07 -9.78e+07 -9.87e+07 -1.03e+08 -1.06e+08 ...
 $ Ratio 5: num  2.6 3.3 2.46 2.35 2.26 ...
 $ Ratio 6: num  77.6 78 76.8 76.9 76.5 ...
 $ Ratio 7: num  1.29 1.28 1.3 1.3 1.31 ...
```

Fuente: Salida de R

Ilustración 3: Regresión logísitica

```
Call:
glm(formula = Default ~ ., family = binomial(link = "logit"),
    data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.9906  -0.3218  -0.3087  -0.2921   2.7146
```

Fuente: Salida de R

Ilustración 4: Salida del código R: Resultados relevantes

```
#Indica la probabilidad de error del modelo
> 1-sum(diag(tab1)/sum(tab1))
[1] 0.04832714

Nivel de significatividad del modelo
with(reg,pchisq(null.deviance-deviance, df.null-df.residual,lower
.tail = F))
[1] 0.1121252

> paste("Diferencia de residuos:", round(dif_residuos, 4))
[1] "Diferencia de residuos: 11.6651"


paste("Grados de libertad:", df)
[1] "Grados de libertad: 7"


#Mide la precisión, el área bajo la curva
> as.numeric(performance(pred,"auc")@y.values)
[1] 0,8161924
```

Fuente: Salida de R

5.2.2 Modelo 2: Regresión logística con análisis de sentimiento

Ilustración 5: Salida del código R: Resultados relevantes

```
> #Indica la probabilidad de error del modelo
> 1-sum(diag(tab1)/sum(tab1))
[1] 0.02071163

> #Para obtener el p valor 1 (como el p valor es
chico , el nivel de significatividad del modelo e
s alto)
> with(reg,pchisq(null.deviance-deviance, df.null
-df.residual,lower.tail = F))
[1] 0.02587807

#Mide la precisión, el área bajo la curva
> as.numeric(performance(pred,"auc")@y.values)
[1] 0.9554488

> paste("Diferencia de residuos:", round(dif_resi
duos, 4))
[1] "Diferencia de residuos: 15.9178"
>
> #Grados de libertad
> paste("Grados de libertad:", df)
[1] "Grados de libertad: 7"
```

Fuente: Salida de R

6. Referencias bibliográficas

Altman, E. I. (1968). Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *The Journal of Finance*, *23*(23), 589–609.

Altman, Edward I. (1993). Corporate Bond and Commercial Loan Portfolio. *New York University, Salomon Brothers Center, Nueva York*.

Altman, Edward I. (2000). Predicting financial distress of companies: Revisiting the Z-score and ZETA models. *Stern School of Business, New York University*, 9-12.

Antweiler, W., & Frank, M. Z. (2004). Is all that talk just noise? The information content of internet stock message boards. *The Journal of Finance*, *59*, 1259–1294.

Asur, S., & Huberman, B. A. (2010). *Predicting the future with social media.*

Atiya, A. F. (2001). Bankruptcy prediction for credit risk using neural networks: A survey and new results. *IEEE Transactions on Neural Networks*, *12*(4), 929-935.

Beaver, W. H. (1966). Financial ratios as predictors of failure. Empirical Research in Accounting: Selected Studies. *Supplement to Journal of Accounting Research*, 71–111.

Bharath, S. T., & Shumway, T. (2008). Forecasting default with the merton distance to default model. *Review of Financial Studies*, *21*, 1339–1369.

Black, F., & Scholes, M. (1973). The pricing of options and corporate liabilities. *The Journal of Political Economy*, *81*, 637–654.

Blume, M. E., Lim, F., & MacKinlay, A. C. (1998). The declining credit quality of U.S. corporate debt: Myth or reality? *Journal of Finance*, *53*, 1389-1414.

Boyd, D., & Crawford, K. (2012). *Critical questions for big data. Information, Commun*.

Braun, P. A., Nelson, D. B., & Sunier, A. M. (1995). Good news, bad news, volatility, and betas. *The Journal of Finance*, *50*, 1575–1603.

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). Classification and Regression Trees. *Monterey, CA: Wadsworth, Inc*.

Brown, B., Chui, M., & Manyika, J. (2011). Are you ready for the era of 'big data'. McKinsey Quarterly,. *McKinsey Quarterly*, *4*(1).

Brown, K. C., Harlow, W. V., & Tinic, S. M. (1988). Risk aversion, uncertain information, and market efficiency. *Journal of Financial Economics*, *22*, 355–385.

Campbell, J. Y., Hilscher, J., & Szilagyi, J. (2008). In search of distress risk. *Journal of Finance*, *63*, 2899–2939.

Cao, L., Guan, L. K., & Jingqing, Z. (2010). Bond rating using support vector machine. Intelligent Data Analysis. *Intelligent Data Analysis*, *10*, 285-296.

Carter, C., & Carlett, J. (1987). Assessing Credit Card Applications Using Machine Learning. *IEEE Expert*, *2*(3), 71-79.

Casanovas, P., De Koker, L., Mendelson, D., & Watts, D. (2017). Regulation of Big Data: Perspectives on strategy, policy, law and privacy. *Health and Technology*.

Chan, W. S. (2003). Stock price reaction to news and no-news: Drift and reversal after headlines. *Journal of Financial Economics*, *70*, 223-260.

Chava, S., & Jarrow, R. A. (2004). Bankruptcy prediction with industry effects. *Review of Finance, 8*, 537-569.

Coffman, J. Y. (1986). The Proper Role of Tree Analysis in Forecasting the Risk Behaviour of Borrowers. *Management Decision Systems , MDS Reports 3, 4, 7 & 9.*

Coval, J. D., & Shumway, T. (2001). Is sound just noise? *The Journal of Finance*, *56*, 1887-1910.

Cox, D. R. (1972). *Regression models and life-tables. Journal of the Royal Statistical Society Series*. *34*, 187–220.

Cranor, L., Rabin, T., Shmatikov, V., Vadhan, S., & Weitzner, D. (2016). *Towards a Privacy Research Roadmap for the Computing Community*.

Deaton, A., & Muellbauer, J. (1980). Economics and consumer behavior. *Cambridge university press*.

Desai, V. S., Crook, J. N., & Overstreet, G. A. (1996). A comparison of neural networks and linear scoring models in the credit union environment. *European Journal of Operational Research*, *95*(1), 24-37.

Douglas, L. (2011). 3d data management: Controlling data volume, velocity and variety. *Gartner. Retrieved*.

Duffie, D., Saita, L., & Wang, K. (2007). Multi-period corporate default prediction with stochastic covariates. *Journal of Financial Economics*, (83), 635–665.

Dunkel, J., & Weber, S. (2007). Efficient Monte Carlo methods for convex risk measures in portfolio credit risk models. *Proceedings of the 2007 Winter Simulation Conference*.

Fang, L., & Peress, J. (2009). Media coverage and the cross-section of stock returns. *The Journal of Finance*, *64*, 2023-2052.

Gentry, J. A., Newbold, P., & Whitford, D. T. (1985). Predicting bankruptcy: If cash flow's not the bottom line, what is? *Financial Analyst's Journal*, *41*, 47-56.

Ghailan, O., Mokhtar, H. M. O., & Hegazy, O. (2016). *Improving Credit Scorecard Modeling Through Applying Text Analysis*. *7*(4).

Greene, W. H. (2000). *Econometric analysis* (2a ed.). Nueva York: Prentice Hall Internacional Editions.

Gujarati, D. (2003). *Econometría. México, D. F.* McGraw-Hill.

Gutiérrez Girault, M. A. (2007). *Modelos de Credit Scoring. Qué, Cómo, Cuándo y Para Qué*. Banco Central de la República Argentina.

Güttler, A., & Wahrenburg, M. (2007). The adjustment of credit ratings in advance of defaults. *Journal of Banking and Finance*, *31*, 751–767.

Hand, D. J., & Henley, W. E. (1997). Statistical classification methods in consumer credit scoring: A review. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *160*(3), 523-541.

Huang, Z., Chen, H. C., Hsu, C. J., Chen, W. H., & Wu, S. S. (2004). Credit rating analysis with support vector machines and neural networks: A market comparative study. *Decision Support Systems*, *37*, 543-558.

Hui, C. H., & Lo, C. F. (2003). Pricing corporate bonds with dynamic default barriers. *Journal of Risk*, *5*(3), 17-37.

Jarrow, R. A., & Turnbull, S. M. (1995). Pricing derivatives on financial securities subject to credit risk. *The journal of finance*, *50*(1), 53–85.

John Walker, S. (2014). *Big data: A revolution that will transform how we live, work, and think.*

Johnson, C. G. (1970). Ratio analysis and the prediction of firm failure. *The Journal of Finance*, *25*(5), 1166-1168.

Kitchin, R. (2014). *Big Data, new epistemologies and paradigm shifts*.

Lando, D., & Skodeberg, T. (2002). Analyzing ratings transitions and rating drift with continuous observations. *Journal of Banking and Finance*, *26*, 423–444.

Lane, W. R., Looney, S. W., & Wansley, J. W. (1986). An application of the cox proportional hazards model to bank failure. *Journal of Banking and Finance*, *10*, 511–531.

Lee, T. S., & Yeh, Y. H. (2004). Corporate governance and financial distress: Evidence from Taiwan. *Corporate Governance*, *12*(3).

Liu, B. (2012). Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies*, *5*(1), 1-167.

Longstaff, F. A., & Schwartz, E. S. (1995). A Simple Approach to Valuing Risky Fixed and Floating Rate Debt. *The Journal of Finance*, *50*(3), 789–819.

López, C. P. (2007). *Minería de datos: Técnicas y herramientas*. Recuperado de https://books.google.com.ar/books?hl=es&lr=&id=wz-D_8uPFCEC&oi=fnd&pg=PR4&dq=Miner%C3%ADa+de+datos:+t%C3%A9cnicas+y+herramientas&ots=ThZ0yn7w6H&sig=_O_gajJYb6mX7Fq2MSt5cTdusaU

Makowski, P. (1985). Credit Scoring Branches Out: Decision Tree-Recent Technology. *Credit World*, *75*.

Manovich, L. (2012). Trending: The promises and the challenges of big social data. *Gold, M.K. Debates in the Digital Humanities. University of Minnesota Press*.

Masiello, B., & Whitten, A. (2009). Engineering Privacy in an Age of Information Abundance. *Google, Inc*.

McAfee, A., Brynjolfsson, E., Davenport, T. H., Patil, D. J., & Barton, D. (2012). *Big data: The management revolution. Harvard business*.

Medina, R. S. (2007). El riesgo de crédito en el marco del acuerdo de Basilea II. *Delta Publicaciones*.

Mester Loretta, J. (1997). What´s the Point of Credit Scoring? *Federal Reserve Bank of Philadelphia*, 3-16.

Metcalf, J., Keller, E. F., & Boyd, D. (2016). *Perspectives on Big Data, Ethics, and Society.*

Morgan, J. (1997). Creditmetrics-technical document. *JP Morgan, New York.*

Munafo, F. (2018). Aplicación del modelo de Merton utilizando VBA. *Revista de investigación en Modelos Financieros*, *07*(01), 109-123.

Nelder, J. A., & Wedderburn, R. (1972). Generalized linear models. Journal of the Royal Statistical Society. *Journal of the Royal Statistical Society*, *135*(3), 370-384.

Nguyen, T. H., Shirai, K., & Velcin, J. (2015). Sentiment analysis on social media for stock movement prediction. Expert Systems with Applications. *Expert Systems with Applications*, *42*(24), 9603-9611.

Ohlson, J. A. (1980). Financial ratios and the probabilistic prediction of bankruptcy. *Journal of Accounting Research*, *18*, 109-131.

Ohm, P. (2009). *Broken promises of privacy: Responding to the surprising failure of anonymization*.

Pak, A., & Paroubek, P. (2010). Twitter as a corpus for sentiment analysis and opinion mining. In LREc. *LREc, Universit´e de Paris-Sud, Laboratoire LIMSI-CNRS*, *10*, 1320-1326.

Pandher, G., & Currie, R. (2013). CEO compensation: A resource advantage and stakeholderbargaining perspective. *Strategic Management Journal*, *34*, 22-41.

Pang, B., Lee, L., & Vaithyanathan, S. (2002). Thumbs up?: Sentiment classification using machine learning techniques. *In Proceedings of the ACL-02 conference on Empirical methods in natural language processing. Association for Computational Linguistics.*, *10*, 79-86.

Perez Ramirez, F. O., & Fernandez Castaño, H. (2007). Las redes neuronales y la evaluación del riesgo de crédito. *Revista Ingenierías Universidad de Medellín*, *6*(10), 77-91.

Ramaswamy, S. (2005). *Simulated credit loss distribution. Journal of Portfolio Management*. *31*, 91-99.

Saavedra García, M. L., & Saavedra García, M. J. (2010). Modelos para medir el riesgo de crédito de la banca. *Cuadernos de administración*.

Scandizzo, S. (2016). *The validation of risk models*.

Schmarzo, B. (2013). *Big Data: Understanding how data powers big business. John Wiley & Sons* (Wiley).

Shimko, D. C. (1993). Bounds of probability. *Risk Magazine*, *6*(4), 33–37.

Shumway, T. (2001). Forecasting bankruptcy more accurately: A simple hazardmodel. *The Journal of Business*, *74*, 101-124.

Soares, S. (2012). *Big data governance: An emerging imperative. Mc Press.* Mc Press.

Soydaner, D., & Kocadağlı, O. (2015). Artificial Neural Networks with Gradient Learning Algorithm for Credit Scoring. *Journal of the School of Bussiness Administration*, *44*(2), 3-12.

Srinivasan, V., & Kim, Y. H. (1987). Credit Granting: A Comparative Analysis of Classification Procedures. *The Journal of Finance*, *XLII*(3).

Tallon, P. P. (2013). Corporate governance of big data: Perspectives on value, risk, and cost. *Computer*.

Tene, O., & Polonetsky, J. (2012). Privacy in the age of big data: A time for big decisions. *HeinOnline*.

Tetlock, P. C. (2007). Giving content to investor sentiment: The role of media in the stock market. *The Journal of Finance*, *62*, 1139–1168.

Tetlock, P. C., Saar-Tsechansky, M., & Macskassy, S. (2008). More than words: Quantifying language to measure firms' fundamentals. *The Journal of Finance*, *63*, 1437–1467.

Turney, P. D. (2002). Thumbs up or thumbs down?: Semantic orientation applied to unsupervised classification of reviews. *In Proceedings of the 40th annual meeting on association for computational linguistics . Association for Computational Linguistics.*, 417-424.

Turow, J., Hoofnagle, C. J., Mulligan, D. K., & Good, N. (2007). *The federal trade commission and consumer privacy in the coming decade*.

Vasicek, O. (1977). An equilibrium characterization of the term structure. *Journal of Financial Economics*, *5*, 177–188.

Wei, R. (2008). Development of credit risk model based on Fuzzy theory and its application for credit risk management of commercial banks in China. *4th International Conference on Wireless Communications, Networking and Mobile Computing*, 1-31 , 10339-10342.

West, D. (2000). Neural network credit scoring models. *Computers and Operations Research*, *27*(11-12), 1131–1152.

West, R. (1970). An alternative approach to predicting corporate bond ratings. *Journal of Accounting Research*, *7*(118-127).

Zmijewski, M. E. (1984). Methodological issues related to the estimation of financial distress prediction models. *Journal of Accounting Research*, *22*, 59-82.