CMA IADCOM - UBA CENTRO DE INVESTIGACIÓN EN METODOS CUANTITATIVOS APLICADOS A LA ECONOMÍA Y LA GESTIÓN

Universidad de Buenos Aires Facultad de Ciencias Económicas

LATINDEX
http://www.latindex.org/latindex/ficha?folio=2751

http://www.economicas.uba.ar/institutos_y_centros/revista-de-investigacion-en-modelos-financieros/

LA IMPORTANCIA DE LA GESTIÓN DE DATOS Y SU IMPACTO EN EL RIESGO DE CRÉDITO DE INSTITUCIONES FINANCIERAS[1]

Flavia Munafo

Analista de riesgo financiero en Banco BBVA. Av. Córdoba 111 -C1054AAA- Ciudad Autónoma de Buenos Aires, República Argentina.

flaviamu_19@hotmail.com

Resumen

Recibido: 02/2019

Aceptado: 07/2019

Palabras clave

Gestión de datos.

Big data.

Text mining.

Riesgo de crédito.

Instituciones financieras.

El *big data* se presenta como una nueva tecnología disruptiva que está revolucionando la forma en la que se gestionan los datos. Tiene la capacidad de impactar de manera estratégica en toda la sociedad, por su capacidad de generar transformaciones productivas, económicas y sociales de gran envergadura. A diferencia de otras revoluciones tecnológicas como la Revolución industrial donde el impulso estaba en la energía, las 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.

El riesgo es entendido como la probabilidad de ocurrencia de un evento adverso y sus consecuencias. 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 financiero y es uno de los temas claves para determinar la estabilidad financiera de una empresa o de una persona física.

El presente artículo tiene por objetivo entender como la gestión de datos, en particular la era del *big data* ha revolucionado la forma en la que se gestiona el riesgo de crédito en las instituciones financieras.

ISSN: 2250-687X - ISSN (En línea): 2250-6861

[1] Los conceptos y opiniones contenidos en este trabajo son de exclusiva responsabilidad de la autora.

THE IMPORTANCE OF DATA MANAGEMENT AND ITS IMPACT ON THE CREDIT RISK OF FINANCIAL INSTITUTIONS

Flavia Munafo

Analista de riesgo financiero en Banco BBVA. Av. Córdoba 111 -C1054AAA- Ciudad Autónoma de Buenos Aires, República Argentina.

flaviamu_19@hotmail.com

Abstract

KEYWORDS

Data management,

Big data.

Text mining.

Credit risk.

Financial institutions.

Big data is presented as a new disruptive technology that is revolutionizing the way data is managed. It has the ability to impact strategically throughout society, for its ability to generate large-scale productive, economic and social transformations. Unlike other technological revolutions such as the Industrial Revolution where the impulse was in energy, ICT where the center was the processing and transmission of information, in the case of the era of big data in which we live the momentum focuses on the transformation, analysis, use and storage of automated information variations.

Risk is understood as the probability of occurrence of an adverse event and its consequences. In particular, credit risk is defined as the potential loss caused by the default of the counterparty in an operation that includes a payment commitment. It constitutes the main risk of the financial sector and is one of the key issues in determining the financial stability of a company or an individual.

This article aims to understand how data management, in particular the era of big data has revolutionized the way in which credit risk is managed in financial organizations.

ISSN: 2250-687X - ISSN (En línea): 2250-6861

INTRODUCCIÓN

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).

El riesgo de crédito puede ser entendido como la incertidumbre derivada de la probabilidad de sufrir quebranto 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, 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* 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*, definida como el importe de deuda pendiente de pago en el momento de incumplimiento del cliente.

Un elemento importante en el riesgo de crédito es el evento de *default*, el mismo es definido por el 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*, 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 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´s pero se generalizó a partir de los 90's. 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 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, y 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 el Black & Scholes (1973, 1974) constituye el modelo estructural más representativo. En 1995, la agencia de calificación crediticia, Moody´s, 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.

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 se encargan de explorar el vínculo entre el incumplimiento corporativo y diversas 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 de ellos se basa en los mismos tipos de factores: financieros (como los ratios de las hojas de balance e indicadores) y 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 (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 en 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*. 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 (West, 1970), regresión lineal, el análisis discriminante multivariante, el modelo *Z-score* (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 (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.

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 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 que se genera diariamente en el mercado.

En este contexto, la información 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 como (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.

El presente artículo se encuentra estructurado en tres secciones, en la primera sección se explicará con mayor detalle en qué consiste la era del *big data*, en la segunda sección se especificará el desarrollo del *text mining* en la era del *big data* y su vinculación al riesgo de crédito para la gestión del riesgo en las instituciones financieras. Finalmente, se expondrán los beneficios y principales desafíos que presenta dicha tecnología disruptiva y se sacarán conclusiones al respecto.

1 LA NUEVA REVOLUCIÓN EN LA GESTIÓN DE DATOS: EL *BIG DATA*

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 podrá 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. En comparación con la información financiera disponible sobre diversas corporaciones, la cantidad de contenido textual es inmensa 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 en el uso de la información. Sin embargo, a pesar de los potenciales riesgos en los que se encuentra inmerso, 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.

Si bien las definiciones de *big data* no son uniformes entre sí, en el entorno de gestión de datos, 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). 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. 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 técnicas de *big data*. El valor puede ser entendido como las oportunidades económicas que se obtienen de los grandes volúmenes de información.

La revolución en la gestión de datos impuesta de manera disruptiva por el *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.

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, en Twitter, 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 de máquina 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, la agrupación y el análisis de

texto. 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 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. A continuación, se expondrán con mayor detalle las técnicas de análisis de texto.

2 *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 qué 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 la minería de datos, por el contrario, el conocimiento se obtiene de bases de datos donde la información está 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 y la aplicación de métodos de aprendizaje no supervisados.

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 la marca. Las empresas pueden tomar este hecho como un indicador por adelantado de la actitud del cliente y detectar con anterioridad cómo 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.

A continuación se procederá a explicar con más detalle cómo el *text mining* puede ser utilizado en el sector financiero en particular como una herramienta para predecir con mayor exactitud el riesgo de default corporativo.

3 *TEXT MINING* APLICADO AL RIESGO DE CRÉDITO

El *credit scoring* constituye 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.

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. 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 (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 insalubres.

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 publicaciones en redes sociales como Twitter distribuyen las últimas noticias con una velocidad sin precedentes.

Estudios financieros recientes en mercados bursátiles encontraron que la información 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 reciente de análisis.

Enormes cantidades de información textual están disponibles; esta información ofrece a las empresas una visión profunda de su salud financiera y rendimiento corporativo. En este contexto, 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.

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. 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. 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, el hecho de que los datos no se encuentren almacenados en una estructura tradicional para su análisis.

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. 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.

La sociedad de la información en la que nos encontramos inmersos 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. 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.

CONCLUSIONES

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*, que otorgan 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 a personas y Pymes sin historial crediticio se les dificulta o incluso se le niega el acceso al primer crédito por las estructuras financieras 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 hacia el bien común.

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 del *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.

REFERENCIAS BIBLIOGRAFICAS

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

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.

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.

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

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.

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.

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.

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.

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.

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.

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).

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

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

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

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

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.

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

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.

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

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.