Estimación de la probabilidad de riesgo de quiebra en las empresas colombianas a partir de un modelo para eventos raros*

*Jorge Iván Pérez García***
*Mauricio Lopera Castaño****
*Fredy Alonso Vásquez Bedoya*****

* doi: 10.11144/Javeriana.cao30-54.eprqe. Este artículo recibió financiamiento de la Universidad de Antioquia. El artículo se recibió el 26/01/2017 y se aprobó 15/05/2017. Sugerencia de citación: Pérez, J., Lopera, M. y Vásquez, F. (2017). Estimación de la probabilidad de riesgo de quiebra en las empresas colombianas a partir de un modelo para eventos raros. *Cuadernos de Administración*, *30*(54), 7-38. http://dx.doi.org/10.11144/Javeriana.cao30-54.eprqe.

** Economista de la Universidad de Antioquia, Medellín, Colombia, 2013. Profesor Catedrático de la Facultad de Ciencias Económicas de la Universidad de Antioquia, Medellín, Colombia.
Correo electrónico: jivan.perez@udea.edu.co

*** Magíster en Estadística de la Universidad Nacional de Colombia, Medellín, Colombia, 2007. Profesor Ocasional de la Facultad de Ciencias Económicas de la Universidad de Antioquia, Medellín, Colombia.
Correo electrónico: mloperacl@gmail.com

**** Magíster en Economía de la Universidad de Antioquia, Medellín, Colombia, 2007. Profesor de la Facultad de Ciencias Económicas de la Universidad de Antioquia, Medellín, Colombia. Profesor Catedrático de la Facultad de Ciencias Humanas y Económicas de la Universidad Nacional de Colombia, Medellín, Colombia.
Correo electrónico: fredy.vasquez@udea.edu.co

Estimación de la probabilidad de riesgo de quiebra en las empresas colombianas a partir de un modelo para eventos raros

Resumen

Para discriminar el riesgo de quiebra y no quiebra de las empresas colombianas que reportaron sus estados financieros a la Superintendencia de Sociedades de Colombia para el periodo 2011-2015, este artículo considera la quiebra como un evento raro y emplea un modelo logístico, un modelo aditivo generalizado, un modelo de valor extremo generalizado y un modelo binario aditivo de valor extremo generalizado (BGEVA). En términos comparativos, el modelo BGEVA presenta mejor desempeño predictivo con respecto a los otros al asumir una distribución de valor extremo en la función link y estructuras semi-paramétricas en las estimaciones, permitiendo así determinar la relación existente entre la probabilidad de default y las variables explicativas.

Palabras clave: quiebra; eventos raros; modelos de predicción.
Clasificación JEL: C14, C35, C53, G33.

Estimation of bankruptcy risk probability in Colombian companies from a model for rare events

Abstract

In order to discriminate bankruptcy risk and no bankruptcy risk of Colombian companies that reported their financial statements to the Superintendency of Corporations for the time period 2011-2015, this paper considers bankruptcy as a rare event and it employs a logistic model, a generalized additive model, a generalized extreme value model, and a binary generalized extreme value additive model (BGEVA). In comparative terms, the BGEVA model presents better predictive performance compared to the other models by assuming an extreme value distribution in the link function and semi-parametric structures in its estimations, thus allowing to establish the existing relation between default probability and explanatory variables.

Keywords: bankruptcy, rare events, prediction models.
JEL Classification: C14, C35, C53, G33

Estimativa da probabilidade de risco de insolvência nas empresas colombianas a partir de um modelo para eventos raros

Resumo

Para discriminar o risco de insolvência e não insolvência das empresas colombianas que apresentaram seus balanços financeiros à Superintendência de Sociedades da Colômbia para o período 2011-2015, este artigo considera a falência como um evento raro e utiliza um modelo logístico, um modelo aditivo generalizado, um modelo de valor extremo generalizado e um modelo binário aditivo de valor extremo generalizado (BGEVA). Em termos comparativos, o modelo BGEVA apresenta melhor desempenho preditivo a respeito dos outros ao assumir uma distribuição de valor extremo na função link e estruturas semiparamétricas nas estimativas, o que permite determinar a relação existente entre a probabilidade de default e as variáveis explicativas.

Palavras-chave: insolvência, eventos raros, modelos de previsão.
Classificação JEL: C14, C35, C53, G33

Introducción

Los modelos de riesgo de quiebra pronostican la probabilidad de que una empresa no pueda cumplir con el pago de sus obligaciones adquiridas (probabilidad de default) y por consiguiente deban cesar sus operaciones. El cálculo de esta probabilidad es un insumo importante para la banca a quienes se les exige, según lo establecido por el Comité de Basilea en lo referente a la regulación bancaria, calcular la probabilidad de default de sus clientes en un horizonte de un año para medir la exposición de sus préstamos y protegerse mediante el cálculo de una provisión.

Una de las metodologías pioneras para el cálculo de la probabilidad de default, es la presentada por Beaver (1966), quien emplea técnicas de análisis univariado para determinar los indicadores financieros más relevantes para discriminar empresas en riesgo de quiebra y no quiebra. Otra metodología pionera es la de Altman (1968), quien propone un modelo estadístico conocido como Z-score, para medir la probabilidad de default mediante técnicas de análisis discriminante.

Hoy en día se utilizan diversas metodologías para medir el riesgo de quiebra tales como la regresión logística (Ohlson, 1980; Laitinen y Laitinen, 2000; Bernhardsen, 2001; Martínez, 2003; Charitou et al., 2004; Brédart, 2014), árboles de decisión (Aoki y Hosonuma, 2004; Santos et al., 2006; Zibanezhad et al., 2011), redes neuronales (O'Leary, 1998; Anandarajan et al., 2001; Santos et al., 2006) entre otras. Sin embargo, ninguna de las anteriores considera la quiebra como un evento raro o de baja ocurrencia, aspecto que de no considerarse, provoca la subestimación en el cálculo de la probabilidad de default.

Una estrategia empleada con frecuencia para que estas metodologías funcionen y no se subestime la probabilidad de default, es sesgar la muestra en las estimaciones de tal forma que la quiebra no aparezca como un evento raro, lo cual conlleva a que la proporción de la muestra con que se estima el modelo no sea la misma proporción con la que se realizan los pronósticos. Sesgar la muestra ayuda a mejorar el cálculo de la probabilidad de default, pero al mismo tiempo conlleva a clasificar como empresas en riesgo de quiebra a algunas que no lo están. Además, emplear en la estimación muestras que no representen bien a la población, puede generar sesgos en la estimación de los parámetros.

A diferencia de las metodologías mencionadas, en Calabrese y Osmetti (2013) las autoras contemplan la quiebra como un evento raro y proponen un modelo de variables

dependientes binarias, cuya función link asimétrica está dada por la densidad acumulada de una distribución de valor extremo generalizada (GEV). Es de anotar, que en esta función link el predictor es de carácter lineal, lo cual no es necesariamente adecuado para capturar las relaciones entre la probabilidad de default y las variables explicativas. Con el fin de superar esta dificultad, Calabrese et al. (2016) proponen un modelo binario aditivo de valor extremo generalizado (BGEVA), donde utilizan funciones spline para modelar la relación entre la probabilidad de default y las variables explicativas, logrando así superar de forma notable los resultados obtenidos con el modelo GEV.

En el presente trabajo se aplica un modelo de regresión logística, un modelo aditivo generalizado (GAM), un modelo de valor extremo generalizado (GEV) y un modelo binario aditivo de valor extremo generalizado (BGEVA), con el objetivo de estimar la probabilidad de default para las empresas que reportaron su información financiera a la Superintendencia de Sociedades de Colombia, para el periodo 2011-2015. Este trabajo será de gran interés para las instituciones financieras, puesto que para estas es más costoso clasificar una empresa como no riesgosa, cuando lo es, que clasificarla como riesgosa cuando no lo es. Además, clasificar una empresa como riesgosa cuando no lo es puede tener un alto costo social, dado que se tendrá menos acceso a financiación de proyectos productivos, de manera que es muy importante hacer una correcta clasificación.

Este trabajo está organizado en cinco secciones, además de esta introducción. En la primera se hace una revisión de la literatura sobre metodologías para estimar la probabilidad de default, que son relevantes con el enfoque del artículo a desarrollar. En la segunda se hace una breve exposición de los modelos estadísticos que se implementan en el trabajo y algunas de sus fallas. En la tercera se describen los datos empleados y las variables que se usan en las etapas de estimación y predicción, junto con un breve análisis estadístico descriptivo. En la cuarta se muestran los resultados obtenidos con los diferentes modelos durante el proceso de estimación y predicción. En la quinta se presentan las conclusiones.

1. Revisión de la literatura

Previo a la determinación legal de la quiebra, no es posible afirmar a priori que una empresa va caer o no en esta, pero sí es posible hacer análisis para calcular la probabilidad de que esta situación pueda ocurrir. Para calcular tal probabilidad se han utilizado diferentes herramientas estadísticas y econométricas, entre ellas metodologías univariadas

y multivariadas, que emplean modelos paramétricos y no paramétricos. La mayoría de estudios que tienen por objeto valorar la salud financiera de una empresa emplean en sus análisis los indicadores financieros, que se obtienen de los estados financieros, información del mercado o de las calificadoras de riesgo (Ravi Kumar y Ravi, 2007).

Uno de los primeros estudios sobre predicción de quiebra empresarial es el de Beaver (1966), donde el autor emplea información financiera del periodo 1954-1964 de 79 empresas no quebradas de un total de 12.000, facilitada por el *Moody's Industrial Manuals* y complementada con un listado de 79 empresas en quiebra proporcionadas por el *Dun and Bradstreer*, pertenecientes a los Estados Unidos. A partir de la información financiera, Beaver construye 30 indicadores que usa con el fin de identificar si existen o no diferencias significativas entre empresas quebradas y no quebradas, aplicando para ello una metodología de análisis univariante, que consiste en la comparación de las medias de los ratios financieros, un test de clasificación dicotómico y un análisis de probabilidad de ratios. Por último, concluye que los indicadores *flujo de efectivo/deuda total* y *utilidad final/activos*, son los que presentan un mejor desempeño para discriminar entre empresas quebradas y no quebradas, inclusive con antelación de 5 años.

Continuando con esta línea de trabajos, Altman (1968) introduce el análisis discriminante múltiple en un estudio realizado con información financiera de dos grupos de 33 empresas pertenecientes a los Estados Unidos, facilitadas del *Moody's Industrial Manuals*. El primero, denominado grupo de quiebra, está compuesto por empresas que presentaron una petición de quiebra bajo el capítulo X de la Ley Nacional de Quiebra para el periodo 1946-1965. El segundo está compuesto por empresas que para el año 1966 continuaban sus operaciones con normalidad. A partir de la información financiera el autor construye 22 ratios que somete a evaluación, teniendo como criterios su significancia estadística, su contribución relativa, la intercorrelación entre los mismos, la observación en la exactitud de predicción de diferentes grupos de ratios y su juicio como analista. Luego de este proceso selecciona 5 ratios como los más relevantes para la estimación del modelo Z-Score (que es una función discriminante que arroja un valor Z para cada empresa evaluada, el cual sirve para saber si la empresa está propensa o no a entrar en quiebra.), clasificando como empresas no quebradas aquellas que presenten un valor $Z \geq 2{,}99$ y como quebradas aquellas que presenten un valor $Z \leq 1{,}81$. El autor denomina "zona de ignorancia" a las empresas que presenten valores Z comprendidos entre 1,81 y 2,99, debido a que en este intervalo existe una alta probabilidad de cometer errores de clasificación.

Por su parte, Altman et al. (1977) desarrollan el modelo ZETA®, que a diferencia del Z-Score, incluye aspectos de mercado, un concepto de varianza del valor de los activos y precios de las acciones en los análisis. En su trabajo emplean información financiera facilitada por el *Moody's Industrial Manuals* para el periodo 1969-1975 de 53 empresas quebradas y 58 no quebradas pertenecientes a Estados Unidos. A partir de los estados financieros construyen 27 indicadores, de los cuales seleccionan 7 para la estimación del modelo después de realizar un proceso iterativo de reducción de variables y encuentran que el modelo ZETA® con una estructura lineal y una estructura cuadrática, clasifica de forma correcta el 92,8% de las empresas un año antes de la quiebra. Además, encuentran que la estructura lineal para el modelo ZETA® tiene un mejor desempeño que la estructura cuadrática para predecir la quiebra con dos, tres, cuatro y cinco años de antelación. Por último comparan el modelo ZETA® con el modelo Z-score de Altman (1968), concluyendo que en el caso de las 7 variables contempladas en su trabajo, el modelo ZETA® presenta un mejor desempeño que el Z-score, mientras que para las 5 variables contempladas en Altman (1968) el ZETA® supera ligeramente las predicciones del Z-score.

Otra metodología para la predicción de la quiebra empresarial es la propuesta por Ohlson (1980), quien introduce por primera vez en este campo el uso de modelos logísticos condicionales. En su estudio emplea datos de 105 empresas quebradas y 2.058 empresas no quebradas, obtenidas del *Compustat File* para el periodo 1970-1978, a partir de los que construye 9 ratios financieros para estimar 3 modelos, uno para predecir la quiebra un año antes, otro para predecirla dos años antes y el restante para predecirla uno o dos años antes. De estos concluye que su porcentaje de clasificación correcta es de 96,12%, 95,55% y 92,84% respectivamente. Además señala que el poder predictivo de cualquier modelo depende de cuando esté disponible la información financiera de la empresa, y que las estimaciones son robustas al utilizar transformaciones lineales en los vectores de ratios financieros y que por ende, una mejora significativa en el poder predictivo de las estimaciones, requiere predictores adicionales.

Berg (2007) propone emplear un modelo aditivo generalizado (GAM) para la predicción de la quiebra empresarial. En su estudio utiliza información de los estados financieros del periodo 1996-2000 de sociedades de responsabilidad limitada, pertenecientes al registro noruego de empresas comerciales, de donde selecciona 13 variables y las primeras diferencias de 10 de las variables que son razones financieras, empleando un total de 23 variables para la estimación de cuatro modelos, a saber, un modelo aditivo generalizado, un modelo de análisis discriminante, un modelo lineal generalizado y un modelo de redes neuronales. Con el fin de comparar el desempeño de los modelos

fuera de muestra y fuera de tiempo, estima la Relación de Precisión[1] (AR) propuesta por Sobehart et al. (2000). Luego, al realizar la validación por fuera de muestra y fuera de tiempo, el autor encuentra que el GAM presenta un AR medio más alto que los demás modelos propuestos y concluye que el GAM tiene un mejor desempeño predictivo. Por último, recomienda hacer remuestreo cuando se realiza validación por fuera de tiempo, pues señala que decidir si un modelo es mejor que otro con un nivel de confianza preestablecido, se logra mejor mediante dicho procedimiento.

Calabrese y Osmetti (2013) consideran la quiebra como un evento raro y proponen un modelo de valor extremo generalizado (GEV) para estimar la probabilidad de quiebra, con el fin de superar algunas de las dificultades que se presentan en el modelo de regresión logística. En su estudio emplean los estados financieros de 210.000 pequeñas y medianas empresas italianas, suministrados por *AIDA-Bureau van Dijk* para el periodo 2005-2011, a partir de los cuales seleccionan las variables más usadas en los trabajos de Altman y Sabato (2006), Ciampi y Gordini (2008) y Vozzella y Gabbi (2010), luego examinan sus relaciones de multicolinealidad y eliminan aquellas cuyo Factor Inflador de la Varianza sea mayor a 5, obteniendo así para estimar el modelo GEV un total de 16 variables, de las cuales 7 resultan ser significativas al 5%. Luego, para medir el poder predictivo de los modelos logístico y GEV, emplean el error cuadrático medio, el error absoluto medio, el índice de área bajo la curva y la medida H con una razón de severidad de 0,01. Por último, concluyen que el modelo GEV supera los inconvenientes que presenta la regresión logística en la subestimación de la probabilidad de default, y que a diferencia del modelo de regresión logística, el modelo GEV es un modelo robusto para diferentes porcentajes de default en la muestra.

Una limitación en el trabajo de Calabrese y Osmetti (2013) es asumir una estructura paramétrica entre la probabilidad de default y las variables explicativas que puede ser muy restrictiva en la práctica. Para superar dicha limitación, Calabrese et al. (2016) proponen un modelo binario aditivo de valor extremo generalizado (BGEVA) que permite mediante el uso de funciones spline, determinar la relación que existe entre la probabilidad de default y las variables explicativas. Para probar el desempeño predictivo del modelo BGEVA respecto a los modelos aditivo logístico y adictivo log-log, los autores utilizan los estados financieros de 50.160 pequeñas y medianas empresas italianas, facilitados

[1] Esta métrica consiste en comparar la curva de potencia del modelo bajo investigación con la del modelo perfecto. Cuanto más cerca está la curva de potencia de la curva de potencia perfecta, mejor se comporta el modelo (Berg, 2007).

por *AIDA-Bureau van Dijk* para el periodo 2006-2011. A partir de estos construyen ratios financieros y dado que los mismos pueden poseer una alta dependencia entre sí, realizan un análisis de multicolinealidad, descartando aquellos cuyo Factor Inflador de la Varianza sea mayor a 5, obteniendo así un total de 21 variables explicativas para la estimación de los modelos. En la etapa de estimación, mediante procesos de selección de variables hacia atrás, los autores obtienen para todos los modelos 12 variables significativas a un nivel del 5%. Luego, con el fin de medir el poder predictivo de estos, analizan el error cuadrático medio, el error absoluto medio, el índice de área bajo la curva y la medida H con una razón de severidad de 0,01, concluyendo que el modelo BGEVA provee un desempeño predictivo superior para la estimación del default para diferentes horizontes de tiempo. Además afirman que entre las principales ventajas del modelo BGEVA está el permitir relajar la hipótesis de linealidad entre la probabilidad de default con las variables explicativas obteniéndose así un mejor ajuste del modelo a los datos.

Para estimar la probabilidad de riesgo de quiebra de las empresas colombianas, algunos autores han empleado metodologías encontradas en la literatura para este tipo de estudios. Entre estos se encuentran los trabajos de Rosillo (2002) y Narváez (2010) quienes aplican la metodología de Análisis Discriminante Múltiple propuesta por Altman (1968); Martínez (2003) quien aplica un modelo de regresión probit; Pérez et al. (2013) quienes además de aplicar un modelo de regresión probit, emplean el método de regresión logística propuesto por Ohlson (1980). Es de anotar que para para estimar la probabilidad de riesgo de quiebra de las empresas colombianas no se han realizado hasta el momento trabajos que contemplen el modelo BGEVA propuesto por Calabrese et al. (2016).

Si bien una gran parte de trabajos empíricos que intentan estimar la probabilidad de riesgo de quiebra emplean como variables dependientes indicadores financieros, es posible identificar en la literatura trabajos que usan como fuentes de información las características individuales de la empresa, entre ellas el análisis del flujo de caja y estrategia corporativa, al igual que variables exógenas tales como indicadores macroeconómicos y el comportamiento del mercado accionario, entre otras. En lo referente a los modelos utilizados para estimar la probabilidad de riesgo de quiebra y siguiendo a Romero (2013) es pertinente anotar que además de métodos paramétricos se han empleado enfoques semiparamétricos y no paramétricos. Jayasekera (2018) señala que en los semiparamétricos las redes neuronales artificiales se han convertido en un enfoque prometedor para la predicción del fracaso empresarial, también indica que en los no paramétricos se han empleado los árboles de decisión y los algoritmos Bagging

y Boosting. Finalmente afirma que es posible encontrar modelos basados en la teoría financiera, así como modelos que incorporan aspectos de las metodologías mencionadas previamente a los que denomina híbridos.

2. Modelos estadísticos

En esta sección se realiza una breve descripción de algunos modelos estadísticos empleados para la predicción del riesgo de quiebra empresarial, a saber, el modelo logístico, el modelo aditivo generalizado (GAM), el modelo de valor extremo generalizado (GEV) y el modelo binario aditivo generalizado de valor extremo (BGEVA). Además se resaltan algunas de las fallas que presentan cuando son empleados para estimar la probabilidad de default.

2.1. Modelo logístico

En modelos lineales generalizados, se conoce como función link a aquella función que establece la relación que hay entre el predictor lineal y la media de la variable respuesta *Y*, es decir

$$g[P_i(\boldsymbol{x}_i)] = \boldsymbol{x}_j'\beta \tag{1}$$

donde $P_i(\boldsymbol{x}_i) = E(Y_i \mid \boldsymbol{x}_i)$ representa la media condicional de Y_i dado $\boldsymbol{x}_i$, $\boldsymbol{x}_i'\beta = \beta_1 + \beta_2 x_{2i} + \beta_3 x_{3i} + \ldots + \beta_k x_{ki}$ es el predictor lineal y g es una función monótona y diferenciable.

Cuando la variable de interés es dicotómica, es decir, la variable respuesta solo puede tomar dos valores, que son a su vez complementarios, se llega al modelo de regresión logístico dado por

$$\ln\left(\frac{P_i(\mathbf{x}_i)}{1-P_i(\mathbf{x}_i)}\right) = \mathbf{x}_i'\boldsymbol{\beta} \tag{2}$$

Este modelo ha sido implementado en diversos estudios, entre ellos, aquellos que intentan estimar la probabilidad de default, pero a pesar de su gran uso, la regresión logística presenta algunas limitaciones, tal como se menciona en Hosmer et al. (2013), donde los autores señalan que en la evaluación del modelo, la clasificación correcta es sensible a los tamaños de muestra en los grupos de entrenamiento, teniendo un mayor porcentaje

de clasificación correcta para el grupo más grande, sin importar el nivel de ajuste del modelo a los datos. Por su parte, Calabrese et al. (2016) señalan como otra dificultad del modelo logístico el suponer una relación lineal entre las variables explicativas y la variable respuesta, omitiendo así la posibilidad de relaciones no lineales que puedan mejorar su poder predictivo.

Otra limitación del modelo de regresión logística se debe a la forma de la matriz de varianzas-covarianzas estimada para el vector de parámetros la cual, bajo un escenario de eventos raros provocará que la probabilidad estimada de default tienda a cero, lo que puede generar grandes errores estándar en la estimación de los parámetros afectando así la precisión de las estimaciones y de las predicciones por fuera de muestra.

Una limitación adicional se da cuando se trabaja en escenarios de eventos raros, pues en estos casos la distribución probabilidad para la muestra es asimétrica, y por tanto una distribución simétrica como la logística, tendrá problemas para generar valores extremos en el predictor lineal que permitan realizar estimaciones cercanas a uno, es decir que bajo el escenario de eventos raros, la regresión logística proporciona una buena tasa de clasificación en las fallas (empresas sin riesgo de quiebra) y una mala tasa de clasificación en los éxitos (empresas en riesgo de quiebra).

2.2. Modelo aditivo generalizado (GAM)

Con el fin de superar la dificultad generada por asumir una relación lineal entre la probabilidad de default y las variables explicativas, se plantea el modelo aditivo generalizado de regresión binaria, el cual reemplaza el predictor lineal $\boldsymbol{x}_j'\beta$ dado en la ecuación (2) por

$$\alpha + \sum_{j=1}^{k} f_j\left(x_{ij}\right) \tag{3}$$

donde $f_j(x_{ji})$ son funciones de suavización spline de las variables explicativas continuas x_{ji}, que tienen la forma

$$f_j\left(x_{ji}\right) = \sum_{l=1}^{q} b_l\left(x_{ji}\right)\beta_l \tag{4}$$

con $b_l(x_{ji})$ una función base. Al combinar el componente de suavización spline con la función link de la regresión logística, se obtienen la función link de respuesta binaria del GAM, definida por

$$\ln\left(\frac{P_i(\mathbf{x}_i)}{1-P_i(\mathbf{x}_i)}\right)=\alpha+\sum_{j=1}^{k} f_j\left(x_{ij}\right) \tag{5}$$

que permite una relación más flexible entre la probabilidad de default y las variables explicativas. Sin embargo, como lo señala Wood (2006), la flexibilización tiene como costo la selección de los términos de suavización y el nivel de suavidad que deben tener.

Siguiendo a Wood (2006), la función $f_j(x_{ji})$ puede ser representada entre otras funciones por un spline cúbico, que es una curva formada por secciones de polinomios cúbicos, conectados de forma tal que exista la primera derivada para que se garantice la continuidad de la función en el punto y que exista la segunda derivada para que no se den cambios de concavidad de un lado al otro del punto, garantizando así la suavidad de la curva alrededor del punto. A los puntos de conexión se les conoce como nudos del spline, que pueden ser puntos igualmente espaciados en el rango de x_{ji} o pueden estar posicionados en sus cuantiles. Wood (2006) define los nudos como x^*_{ji} con $i = 1,2,\ldots,q-2$.

Además como lo señala Wood (2006), existen muchas formas equivalentes de representar la base de las funciones spline cubicas, donde una de ellas es la propuesta por Wahba (1990) y Gu (2002) y que dada por: $b_1(x_{ji}) = 1$, $b_2(x_{ji}) = x_{ji}$ y $b_{i+2}(x_{ji}) = R(x_{ji}, x^*_{ji})$ para $i = 1,2,\ldots,q-2$ donde

$$\begin{aligned} R(x_{ji}, x^*_{ji}) = &([(x^*_{ji} - (1/2))^2 - (1/12)][(x_{ji} - (1/2))^2 - (1/12)]/4) - \\ &([(|x_{ji} - x^*_{ji}| - (1/2))^4 - (1/2)(|x_{ji} - x^*_{ji}| - (1/2))^2 + (7/240)]/24) \end{aligned} \tag{6}$$

El usar esta base spline para $f_j(x_{ji})$ implica que el predictor para el modelo de regresión logística continúe siendo lineal. Sin embargo, cada variable explicativa x_{ji} se reemplaza por el vector

$$\boldsymbol{X_{ji}} = [1, x_{ji}, R(x_{ji}, x^*_{j1}), R(x_{ji}, x^*_{j2}), \ldots, R(x_{ji}, x^*_{j(q-2)})] \tag{7}$$

y ahora cada componente $x'_{ji}\beta_j$, en la ecuación (2) se reemplaza por $\boldsymbol{X}'_{ji}\beta_j$ siendo β_j un vector de orden $1 \times q$. Se tienen entonces que el GAM de respuesta binaria está dado por

$$\ln\left(\frac{P_i}{1-P_i}\right) = \sum_{r=1}^{k} \mathbf{X}'_{ri}\boldsymbol{\beta}_r \qquad (8)$$

Nótese que en el GAM cada variable explicativa entra en forma aumentada por una cantidad de componentes spline que son funciones de ella misma, para ayudar a que se aproxime de forma más adecuada a la verdadera relación con P_i y se mejore el desempeño dentro de muestra. Dado que este modelo intenta recoger la verdadera relación subyacente de P_i con las variables explicativas, se espera un mejor desempeño predictivo del GAM respecto al modelo de regresión logística para datos fuera de muestra.

Dado que el GAM contiene una gran cantidad de términos para cada una de las variables explicativas, la maximización clásica de la función de verosimilitud se hace más compleja y se debe recurrir a un esquema de estimación penalizada, que deje los términos lineales o no lineales en el componente de suavización para cada variable explicativa que en realidad se requiera para describir el comportamiento de P_i.

Sin embargo, a pesar de la posible mejora que presenta el GAM en relación con el ajuste a los datos, este sigue utilizando la distribución de probabilidad logística como base para realizar su proceso de optimización penalizada, por lo que es probable que este adopte algunos de los problemas señalados en el modelo de regresión logística. Por tanto, en un escenario de eventos raros, no se esperan mejoras significativas en términos de desempeño predictivo del GAM respecto a la regresión logística.

2.3. Modelo de valor extremo generalizado (GEV)

La teoría de valor extremo[2], como lo muestra Tsay (2010), es útil para estudiar el comportamiento de cola en las distribuciones de los procesos estocásticos. Tal es el caso de la distribución de valor extremo generalizado (GEV) propuesta por Jenkinson (1955), que a diferencia de distribuciones simétricas como la normal o la logística, es capaz de generar grandes valores en el extremo derecho de la distribución.

[2] Una introducción formal a la teoría de valor extremo puede ser encontrada en Beirlant et al. (2004) y una referencia de un carácter práctico en Tsay (2010).

Partiendo de Beirlant et al. (2004) y de Calabrese y Osmetti (2013), la densidad acumulada de la distribución GEV se puede definir como

$$F(x) = e^{-\left[1+\tau\left(\frac{x-\mu}{\sigma}\right)\right]^{-1/\tau}} \tag{9}$$

para un soporte de densidad dada por $S = \{x: 1 + \tau(x - \mu)/\sigma > 0\}$, con μ, σ y τ los parámetros de localización, escala y forma, respectivamente, donde τ determina que distribución de valor extremo está representada. Si $\tau > 0$ se tiene la distribución Fréchet, si $\tau < 0$ se tiene la distribución Weibull, y si $\tau \to 0$ se tiene la distribución Gumbel.

Calabrese y Osmetti (2013) proponen el modelo GEV con el fin de superar los problemas del modelo de regresión logística en escenarios de eventos raros. Para ello generalizan el modelo log-log complementario utilizando la función cuantil de la distribución GEV como función link. La función link que consideran está representada por

$$\frac{\left[-\ln(P_i)\right]^{-\tau} - 1}{\tau} = \mathbf{x}_i'\boldsymbol{\beta} \tag{10}$$

Además, la relación entre las variables explicativas del modelo GEV con la probabilidad de éxito se halla a partir de la expresión

$$\frac{\partial P_i}{\partial x_{ji}} = \beta_j P_i \left[1 + \tau(\mathbf{x}_i'\boldsymbol{\beta})\right]^{-\frac{(1+\tau)}{\tau}} \tag{11}$$

la cual indica que, cuando los parámetros y demás variables del modelo permanecen constantes, existe una relación directa entre la variable explicativa x_{ji} y la probabilidad de éxito P_i.

Es de anotar que el modelo GEV supera una de las limitaciones del modelo de regresión logística y el GAM, a saber, el no generar grandes valores positivos en el predictor lineal en un escenario de eventos raros, lo que imposibilita llegar a probabilidades estimadas de éxito iguales o superiores a 0,5, pero al igual que la regresión logística, continúa asumiendo una relación lineal entre la probabilidad de éxito y las variables explicativas, hecho que podría ser no cierto a la luz de los datos.

2.4. Modelo binario aditivo de valor extremo generalizado (BGEVA)

Una dificultad del modelo GEV propuesto por Calabrese y Osmetti (2013) es asumir una relación lineal entre la función link y las variables explicativas, lo cual puede ser difícil de cumplirse en la práctica, ya que se asume implícitamente que no importa el nivel de la variable explicativa x_{ji}, pues un incremento unitario de esta última, siempre nos llevará al mismo efecto β_j. Para superar esta dificultad Calabrese et al. (2016), proponen el modelo binario aditivo de valor extremo generalizado (BGEVA), que a diferencia del GEV utiliza funciones spline para flexibilizar el supuesto de linealidad entre las variables explicativas y la probabilidad de default tal como lo hace el GAM, pero sigue utilizando la misma función link asimétrica que se emplea en el modelo GEV, dando cabida a probabilidades positivas superiores a 0,5 en escenarios de eventos raros.

Dado que en la práctica el supuesto de linealidad asumido por Calabrese y Osmetti (2013) en el modelo GEV es muy restrictivo, Calabrese et at. (2016) reemplazan el predictor lineal $\boldsymbol{x}_j'\beta$ por el componente de suavización spline, $\alpha + \sum_{j=1}^{k} f_j\left(x_{ji}\right)$, donde $f_j(\cdot)$ son funciones spline de las covariables x_{ji}. Así, al combinar el componente de suavización spline con la función link asimétrica del GEV, obtienen la función link para el modelo BGEVA que está dada por

$$\frac{\left[-\ln\left(P_i\right)\right]^{-\tau}-1}{\tau} = \alpha + \sum_{j=1}^{k} f_j\left(x_{ij}\right) = \eta_i \qquad (12)$$

donde $\tau \in \Re$ es el parámetro de forma que determina el peso de la cola en la distribución, $\eta_i = \sum_{r=1}^{k} \mathbf{X}_{ri}'\boldsymbol{\beta}_r$ y con un soporte dado por $1 + \tau\eta_i \geq 0$.

Como lo señalan Calabrese et al. (2016), al reemplazar los términos de suavización con las funciones spline, se conduce a un modelo paramétrico cuya matriz de diseño incluye bases spline, lo cual implica que el modelo BGEVA pueda ser estimado por el método de máxima verosimilitud. Sin embargo, dado el alto número de componentes de suavización, la estimación de máxima verosimilitud nos llevará a un resultado con muchos parámetros no significativos, lo cual no es lo adecuado en un modelo con muchas variables. Por ello al igual que en el GAM, este problema se puede superar por estimación de máxima verosimilitud penalizada, donde el uso de matrices de penalización elimina los términos que en realidad no son significativos a la luz de los datos, es decir, cada

componente de suavización tiene un componente de penalidad asociado $\beta_r^T \boldsymbol{S}_j \beta_r$, donde $\beta_r^T = (\beta_{j1}, \beta_{j2}, \ldots, \beta_{jk})$ y $\boldsymbol{S}_j$ es una matriz cuadrada semi definida positiva de coeficientes conocidos midiendo la rudeza del j-ésimo componente de suavización.

Es de anotar que algunos valores de τ, en especial los más grandes, pueden llevar a indeterminaciones lo que se convierte en una debilidad del modelo BGEVA, y por tanto su estimación debe restringirse a valores de τ que no generen indeterminaciones.

3. Datos y variables

El conjunto de datos empleado corresponde a 28.677 empresas colombianas que reportaron sus estados financieros a la Superintendencia de Sociedades de Colombia para el periodo 2011-2015. Este conjunto de datos se reduce a 26.046 empresas luego de eliminar las que presentan valores faltantes, no congruentes o poco realistas en algunas de sus cuentas, dado que son necesarias para obtener las covariables que se usan en la estimación de los modelos.

Las 26.046 empresas se dividen en quebradas y no quebradas, quedando el grupo de quebradas compuesto por 237 empresas que durante el periodo 2011-2015 debieron ser disueltas o liquidadas, según lo establecido por el numeral 7 del artículo 34 de la Ley 1258 de 2008, por tener pérdidas que redujeron el patrimonio neto de la sociedad por debajo del 50% del capital suscrito. Por su parte, el grupo de no quebradas queda conformado por 25.809 empresas que durante el periodo 2011-2015 no tenían su patrimonio neto por debajo del 50% del capital suscrito.

Partiendo de los grupos de empresas quebradas y no quebradas, se realiza un muestreo aleatorio con el fin de seleccionar para la estimación de los cuatro modelos el 80% de las empresas de cada grupo, mientras que el 20% restante se empleará para evaluar su desempeño predictivo, teniéndose así 20.837 empresas para los procesos de estimación y 5.209 empresas para los procesos de validación.

Dado el conjunto de empresas que integran cada grupo, se construyen a partir de sus estados financieros diez indicadores, que si bien no son los únicos que se pueden usar para calcular la probabilidad de riesgo de quiebra, si son algunos de los más utilizados en los modelos donde se busca estimar la fragilidad empresarial (Altman, 1968; Ohlson, 1980; Rosillo, 2002; Martínez, 2003; Berg, 2007). Se emplean dos indicadores de liquidez (razón corriente y razón de tesorería), un indicador de actividad (rotación

activos), tres indicadores de rentabilidad (margen neto, rentabilidad activo y rentabilidad patrimonio), dos indicadores de endeudamiento (endeudamiento financiero y nivel de endeudamiento) y dos indicadores de apalancamiento (apalancamiento corto plazo y apalancamiento largo plazo).

La razón corriente (*raz*) se define como *activo corriente/pasivo corriente* y representa la capacidad que tiene la empresa para pagar sus obligaciones financieras en el corto plazo. La razón de tesorería (*teso*) se define como *(caja+bancos)/pasivo corriente* e indica la capacidad que tiene la empresa para cubrir en el corto plazo sus obligaciones financieras con los activos más líquidos que posee. La rotación activos (*rota*) se define como *ventas/total activos* y da cuenta del grado de eficiencia con que la empresa utiliza sus activos para generar ingresos por ventas. El margen neto (*margen*) se define como *utilidad/ventas* e indica la rentabilidad generada por la empresa por cada unidad monetaria en ventas. La rentabilidad activo (*ractiv*) se define como *utilidad/total activos* y representa la utilidad generada por la empresa tomando como base el total de activos. La rentabilidad patrimonio (*rpatri*) se define como *utilidad/patrimonio* e indica la rentabilidad generada tomando como base el capital que se ha invertido en la empresa por los accionistas. El endeudamiento financiero (*endeu*) se define como *obligaciones financieras/ventas netas* y establece el porcentaje que representan las obligaciones financieras de corto y largo plazo, con respecto a las ventas de un periodo determinado. El nivel de endeudamiento (*niven*) se define como *total pasivos/total activos* e indica el porcentaje de recursos y participación que tienen los acreedores dentro de la empresa. El apalancamiento corto plazo (*apalc*) se define como *pasivo corriente/patrimonio* e indica el grado de compromiso del patrimonio de la empresa con los acreedores en el corto plazo. El apalancamiento largo plazo (*apalar*) se define como *total pasivos/patrimonio* y se interpreta como el nivel de compromiso del patrimonio de la empresa con los acreedores en el largo plazo.

Una vez definidos los indicadores a utilizar en los procesos de estimación y predicción de los modelos, se realiza una comparación entre los indicadores de los grupos de empresas quebradas y no quebradas, con el fin de verificar si existen diferencias significativas que permitan a los modelos discriminar entre grupos. La hipótesis nula considerada está dada por:

$$H_0: \mu_{i,no\ frágil} = \mu_{i,frágil}$$
$$H_1: \mu_{i,no\ frágil} \neq \mu_{i,frágil}$$

Tabla 1
Prueba de diferencia de medias

Variable	Grupo no quebradas		Grupo quebradas		Estadístico t	P-valor
	Media	Desviación Estándar	Media	Desviación Estándar		
raz	4,368	9,985	3,568	10,580	0,870	0,384
teso	0,491	2,062	0,202	0,935	3,516	0,000
rota	1,419	2,485	3,282	6,564	-3,272	0,001
margen	-0,005	1,193	-0,796	2,265	4,021	0,000
ractiv	0,033	0,151	-0,086	0,412	3,329	0,001
rpatri	0,036	0,310	-0,086	1,098	1,278	0,201
endeu	0,492	0,399	1,572	1,602	-7,776	0,000
niven	0,609	2,947	1,866	5,570	-2,601	0,009
apalc	1,467	4,167	0,993	12,809	0,427	0,670
apalar	1,449	4,244	0,512	13,462	0,803	0,422

Fuente: cálculos propios.

Al realizar la prueba de diferencia de medias presentada en la tabla 1, se observa que los indicadores razón corriente, rotación activos, margen neto, rentabilidad activos, endeudamiento financiero y nivel endeudamiento, rechazan la hipótesis nula a un nivel de significancia del 5%, lo que permite concluir que estos indicadores parecen ser los más relevantes para discriminar entre grupos de empresas.

De otro lado, se observa que casi todos los indicadores son consistentes con la lógica financiera, pues el grupo de no quebradas respecto al grupo de quebradas presentan en promedio valores más altos en los indicadores razón corriente, razón de tesorería, margen neto, rentabilidad del activo, rentabilidad del patrimonio, apalancamiento corto plazo y apalancamiento total, y más bajos en los indicadores endeudamiento financiero y nivel de endeudamiento. Es de anotar que el indicador rotación de activos no es consecuente con la lógica financiera, pues el grupo de no quebradas debería presentar en promedio un valor más alto que el grupo de quebradas y esto no se cumple.

4. Estimación y predicción

En esta sección se presentan los resultados encontrados en la estimación y predicción para el modelo de regresión logística, el modelo aditivo generalizado, el modelo de valor extremo generalizado y el modelo binario aditivo de valor extremo generalizado, donde se utiliza para la estimación de los modelos el 80% de las empresas de cada grupo, y para

observar el desempeño predictivo de los mismos se usa el 20% restante. Los resultados se obtuvieron mediante el programa R (R Development Core Team, 2017), empleando el paquete stats para realizar el ajuste del modelo logístico, el paquete mgcv (Wood, 2017) para realizar el ajuste del GAM y el paquete GJRM (Marra y Radice, 2017) para realizar el ajuste de los modelos GEV y BGEVA.

En la etapa predictiva se adopta la estrategia clásica del análisis discriminante, donde se define como éxito (grupo en riesgo de quiebra) a las observaciones que en la predicción obtienen un valor igual o superior a 0,5 y como falla (grupo en riesgo de no quiebra) a aquellas observaciones que obtienen un valor inferior a 0,5. De otro lado, se construye para cada modelo la matriz de confusión binaria definida en Powers (2011) y que se presenta en la tabla 2, en la cual *QQ* hace referencia al número de empresas clasificadas por el modelo en riesgo de quiebra y pertenecientes al grupo de quiebra (clasificación correcta), *QN* representa el número de empresas clasificadas por el modelo en riesgo de quiebra y pertenecientes al grupo de no quiebra (clasificación incorrecta), *NQ* hace referencia al número de empresas clasificadas por el modelo en riesgo de no quiebra y pertenecientes al grupo de quiebra (clasificación incorrecta), *NN* hace referencia al número de empresas clasificadas por el modelo en riesgo de no quiebra y pertenecientes al grupo de no quiebra (clasificación correcta).

Tabla 2
Matriz de confusión

		Valor real		
	Grupos	**Quiebra**	**No quiebra**	**Total predicción**
Predicción	Quiebra	*QQ*	*QN*	*QQ + QN*
	No quiebra	*NQ*	*NN*	*NQ + NN*
	Total real	*QQ + NQ*	*QN + NN*	*QQ + NQ + QN + NN*

Fuente: elaboración propia.

Con el fin de observar las proporciones de clasificación correcta e incorrecta, se calculan algunas de las medidas presentadas en Powers (2011), a saber, recall, miss rate, fall-out, inverse recall, precision, false discovery rate, false omission rate e inverse precisión, y definidas como:

- Recall = $\frac{QQ}{QQ + NQ}$

  Representa la proporción de empresas que en realidad se encuentran en quiebra y que son predichas por el modelo en riesgo de quiebra.

- Miss rate $= \frac{NQ}{QQ + NQ}$ — Representa la proporción de empresas que en realidad se encuentran en quiebra y que son predichas por el modelo en riesgo de no quiebra.
- Fall-out $= \frac{QN}{QN + NN}$ — Representa la proporción de empresas que en realidad se encuentran en no quiebra y que son predichas por el modelo en riesgo de quiebra.
- Inverse recall $= \frac{NN}{QN + NN}$ — Representa la proporción de empresas que en realidad se encuentran en no quiebra y que son predichas por el modelo en riesgo de no quiebra.
- Precision $= \frac{NN}{QN + NN}$ — Representa la proporción de empresas que son predichos por el modelo en riesgo de quiebra y que en realidad se encuentran en quiebra.
- False discovery rate $= \frac{NN}{QN + NN}$ — Representa la proporción de empresas que son predichos por el modelo en riesgo de quiebra y que en realidad se encuentran en no quiebra.
- False omission rate $= \frac{NN}{QN + NN}$ — Representa la proporción de empresas que son predichos por el modelo en riesgo de no quiebra y que en realidad se encuentran en quiebra.
- Inverse precision $= \frac{NN}{QN + NN}$ — Representa la proporción de empresas que son predichos por el modelo como en no quiebra y que en realidad son empresas en no quiebra.

En el análisis de resultados se hace mayor énfasis en las medidas recall e inverse recall, debido a que estas representan la proporción de empresas que fueron clasificadas de forma correcta por el modelo ya sea en riesgo de quiebra o en riesgo de no quiebra, cuando su situación real es respectivamente quiebra o no quiebra, pues el conocer la situación real en que se encuentran las empresas que se están clasificando, permite identificar cuál de los modelos es el que presenta una mayor tasa de aciertos.

También pero en menor medida, se analizan las medidas precision e inverse precision debido a que al realizar una comparación entre modelos, los totales de las empresas predichas en riesgo de quiebra y en riesgo de no quiebra varían de modelo a modelo, lo cual hace que los resultados obtenidos no sean comparables entre modelos. Las medidas miss rate, fall-out, false discovery rate y false omission rate se calculan pero no se analizan, debido a que son respectivamente el complemento de las medidas recall, inverse recall, precision e inverse precision.

Por último, con el fin de hacer comparable el desempeño de los modelos, se calcula el F_1-score el cual está definido entre 0 y 1, donde valores cercanos a 1 indican un buen desempeño predictivo del modelo, mientras que valores cercanos a 0 indican un mal desempeño predictivo. A su vez, se calcula el Matthews correlation coefficient (MCC) el cual se encuentra entre –1 y 1, donde valores cercanos a 1 indican un buen desempeño predictivo del modelo y valores cercanos a –1 indican un mal desempeño predictivo, es decir, se estarían clasificando las empresas quebradas en riesgo de no quiebra y/o a las empresas no quebradas como en riesgo de quiebra. Las dos medidas mencionadas también se presentan en Powers (2011).

4.1. Modelo logístico

En la tabla 3 se observa que las variables rotación activos, margen neto, rentabilidad activos, rentabilidad patrimonio, nivel de endeudamiento, endeudamiento financiero, apalancamiento corto plazo y apalancamiento largo plazo son significativas. Además se aprecia que los signos de los parámetros de algunas de las variables significativas no son consecuentes con la lógica financiera, pues se espera que el efecto que tienen las variables rotación activos, rentabilidad patrimonio, apalancamiento corto plazo sobre la variable dependiente sea negativo, ya que incrementos en estas variables deberían dar como resultado una reducción en la probabilidad de riesgo de quiebra.

Tabla 3
Estimación del modelo logístico

Parámetro	Estimación	Error estándar	Valor Z	Pr(>|z|)
Intercept	-5,522	0,113	-48,563	0,000
raz	0,012	0,008	1,534	0,125
teso	-0,104	0,079	-1,322	0,186
rota	0,079	0,017	4,598	0,000
margen	-0,084	0,039	-2,188	0,029
ractiv	-3,255	0,427	-7,630	0,000
rpatri	0,820	0,158	5,203	0,000
niven	0,613	0,068	9,022	0,000
endeu	0,050	0,013	3,929	0,000
apalc	0,577	0,068	8,475	0,000
apalar	-0,549	0,062	-8,809	0,000

Fuente: elaboración propia.

A partir de la tabla 4, la cual es construida a partir del 20% de la muestra reservada para el proceso de validación, se observa que el modelo de regresión logística clasifica en riesgo de quiebra el 2,13% de las empresas que en realidad se encuentran en quiebra (recall), y en riesgo de no quiebra al 99,92% de las empresas que se encuentran realmente en no quiebra (inverse recall). De otro lado se observa que del total de empresas predichas en riesgo de quiebra, solo el 20,00% en realidad se encuentran quebradas (precision), mientras que del total de empresas predichas en riesgo de no quiebra, el 99,12% no se encuentran realmente quebradas (inverse precision).

Además se observa que las medidas F_1-score y Matthews correlation coefficient (MCC) se encuentran cercanas a 0, lo cual indica que el modelo de regresión logística presenta un bajo desempeño predictivo. Este hecho se puede deber a que en el proceso de estimación la quiebra es considerada como un evento de rara ocurrencia y por tanto no se sesga la muestra, estrategia que si es implementada en estudios similares con el objetivo de obtener un buen desempeño predictivo para el modelo de regresión logística.

Tabla 4
Matriz de confusión del modelo logístico

		Valor real		
	Grupos	**Quiebra**	**No quiebra**	**Total predicción**
Predicción	Quiebra	1	4	5
	No quiebra	46	5158	5204
	Total real	47	5162	5209
	Recall	0,0213	Precision	0,2000
	Inverse recall	0,9992	Inverse precision	0,9912
	F_1-Score	0,0385	MCC	0,0626

Fuente: elaboración propia.

4.2. Modelo aditivo generalizado

En la tabla 5 se aprecia que el componente paramétrico del modelo no considera ninguna variable explicativa y que el componente no paramétrico considera como variables significativas o marginalmente significativas a la razón corriente, la rotación activos, la rentabilidad activos, el nivel de endeudamiento y el endeudamiento financiero. A partir de la figura 1 se observa que de estas cinco variables, la razón corriente, el nivel de endeudamiento y el endeudamiento financiero presentan un efecto positivo sobre la variable dependiente en casi toda su extensión, mientras que la rotación activos y la

rentabilidad activos presentan un efecto negativo sobre la variable dependiente en un tramo muy pequeño. Es de anotar que el efecto de la razón corriente no es consecuente con la lógica financiera, pues se esperaría que un aumento en esta debería disminuir la probabilidad de riesgo de quiebra.

Tabla 5
Estimación del GAM

Coeficientes paramétricos				
Parámetro	**Estimación**	**Error Estándar**	**Valor Z**	**Pr(>|z|)**
(Intercept)	-6,693	0,229	-29,180	-6,693
Importancia aproximada de suavización de términos				
Parámetro	**edf**	**Ref.df**	**Chi. sq**	**Valor-P**
raz	1,735	1,953	10,125	0,008
teso	1,270	1,465	0,541	0,478
rota	3,372	3,797	9,180	0,052
margen	5,227	5,641	7,420	0,252
ractiv	4,875	5,274	24,526	0,000
rpatri	3,347	3,624	3,452	0,426
niven	6,773	7,644	97,715	0,000
endeu	1,005	1,010	4,280	0,039
apalc	3,690	4,015	5,719	0,238
apalar	3,387	3,687	3,704	0,388

Fuente: elaboración propia.

En la tabla 6 se observa que el GAM al igual que el modelo de regresión logística, clasifica como en riesgo de quiebra el 2,13% de las empresas que se encuentran realmente quebradas (recall), y del total de empresas que predice en riesgo de no quiebra, el 99,12% no se encuentra en realidad en quiebra (inverse precision). Por otro lado se observa que el GAM mejora respecto al modelo de regresión logística, pues clasifica en riesgo de no quiebra al 99,96% de las empresas que no se encuentran realmente quebradas (inverse recall) y del total de empresas predichas en riesgo de quiebra, mejora clasificando el 33,33% de aquellas que en realidad se encuentran en quiebra (precision).

Además el GAM, al igual que el modelo de regresión logística, presenta valores cercanos a 0 para las medidas F_1-score y para la Matthews correlation coefficient (MCC), lo cual significa que dicho modelo también tiene un bajo desempeño predictivo. Lo anterior puede deberse a que independiente del uso de funciones spline que permita un mayor

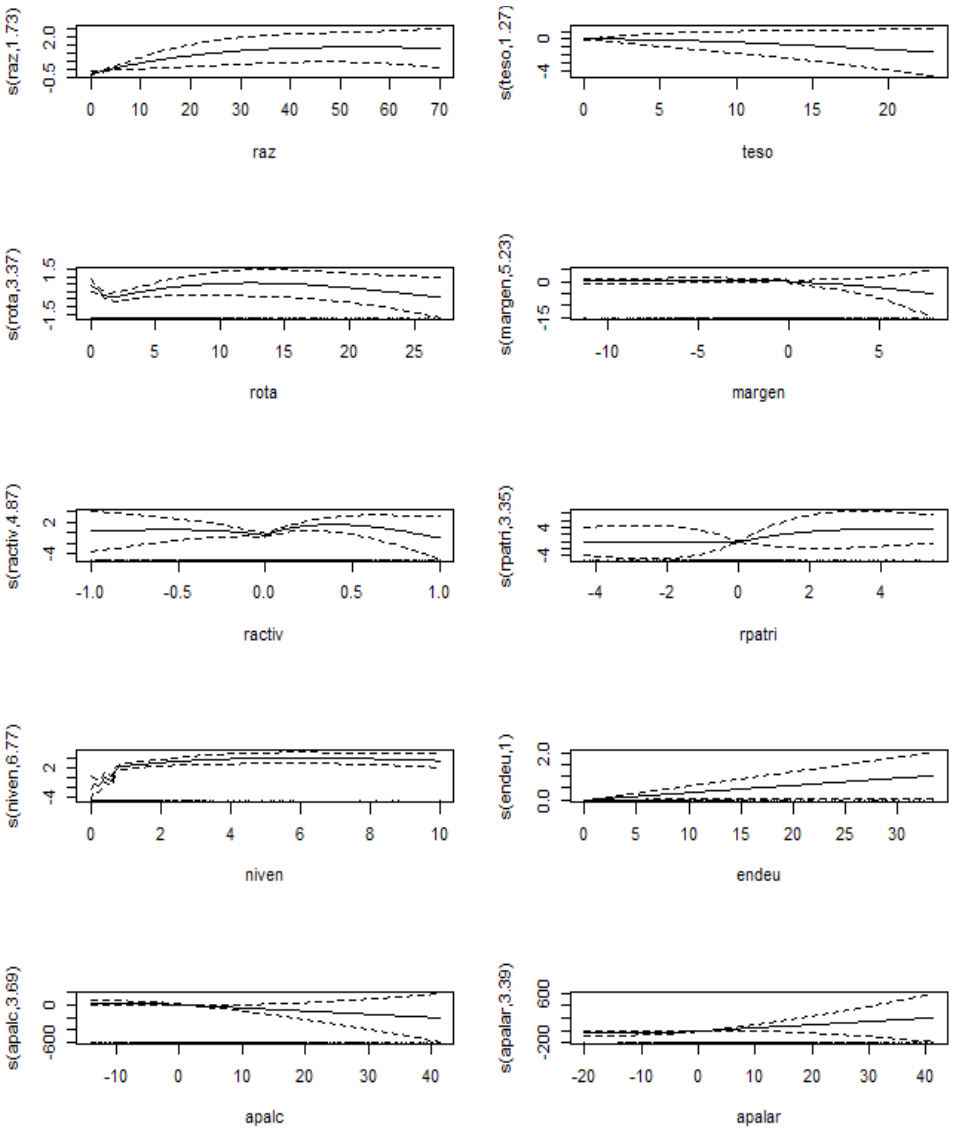

Figura 1. Ajuste variables del GAM

Fuente: elaboración propia.

nivel de ajuste del modelo a los datos, el desempeño predictivo del GAM es sensible a los tamaños muéstrales de los grupos, favoreciendo siempre al grupo más grande (Hosmer et al., 2013).

Tabla 6
Matriz de confusión del GAM

		Valor real		
	Grupos	**Quiebra**	**No quiebra**	**Total predicción**
Predicción	Quiebra	1	2	3
	No quiebra	46	5.160	5.206
	Total real	47	5.162	5.209
	Recall	0,0213	Precision	0,3333
	Inverse recall	0,9996	Inverse precision	0,9912
	F_1-Score	0,0400	MCC	0,0823

Fuente: elaboración propia.

4.3. Modelo de valor extremo generalizado

Como lo señalan Calabrese y Osmetti (2013), en la interpretación de los parámetros estimados en el modelo GEV, se debe suponer que si las demás variables independientes y parámetros permanecen constantes, entonces, un incremento en una unidad en el j-ésimo regresor, provocará una disminución en la estimación de P_i si el parámetro β_j es positivo o provocará un aumento en la estimación de P_i si el parámetro β_j es negativo, es decir, el signo de los parámetros estimados en el modelo GEV se interpretan en sentido contrario a los estimados en la regresión logística. Por tanto, de los resultados presentados en la tabla 7, se infiere que de las variables significativas en la estimación del modelo GEV, el margen neto, la rentabilidad activos, el nivel de endeudamiento y el apalancamiento largo plazo presentan un efecto adecuado sobre la variable dependiente según la lógica financiera, mientras que las variables rotación activos, rentabilidad patrimonio y apalancamiento corto plazo presentan un efecto contrario.

Tabla 7
Estimación del modelo GEV

Parámetro	Estimación	Error estándar	Valor Z	Pr(>|z|)
Intercept	-1,912	0,044	-43,850	0,000
raz	0,003	0,002	1,442	0,149
teso	-0,007	0,014	-0,525	0,599
rota	0,019	0,004	4,195	0,000
margen	-0,040	0,011	-3,698	0,000
ractiv	-0,522	0,112	-4,672	0,000

Continúa →

Parámetro	Estimación	Error estándar	Valor Z	Pr(>|z|)
rpatri	0,157	0,048	3,306	0,001
niven	0,680	0,051	13,418	0,000
endeu	0,005	0,004	1,277	0,201
apalc	0,115	0,024	4,866	0,000
apalar	-0,113	0,023	-4,864	0,000

Fuente: elaboración propia.

A partir de la matriz de confusión presentada en la tabla 8 se infiere que el modelo GEV mejora respecto al modelo de regresión logística y el GAM, pues logra predecir en riesgo de quiebra el 55,32% de las empresas realmente quebradas (recall) y del total de las empresas predichas en riesgo de no quiebra por el modelo, se tiene que en realidad el 99,59% no se encuentran en quiebra (inverse precision). De otro lado, del total de empresas predichas en riesgo de quiebra, solo el 26,53% de estas se encuentran realmente en quiebra (precision), siendo este resultado mayor que el obtenido por el modelo de regresión logística, pero menor que el presentado por el GAM. También se observa que el modelo GEV presenta una menor inverse recall que el modelo logistico y GAM, clasificando en riesgo de no quiebra el 98,61% del total de empresas que en realidad no se encuentran quebradas (inverse recall).

Además se aprecia que las medidas F_1-score y Matthews correlation coefficient (MCC) para el modelo GEV, son mayores que las presentadas en el modelo logístico y en el GAM, siendo ambas medidas superiores a 0,35, hecho que puede ser explicado por la función link asimétrica que se emplea en el modelo GEV, función que permite capturar mejor la información en eventos de rara ocurrencia, como lo es la quiebra.

Tabla 8
Matriz de confusión del modelo GEV

		Valor real		
	Grupos	Quiebra	No quiebra	Total predicción
Predicción	Quiebra	26	72	98
	No quiebra	21	5.090	5.111
	Total real	47	5.162	5.209
	Recall	0,5532	Precision	0,2653
	Inverse recall	0,9861	Inverse precision	0,9959
	F_1-Score	0,3586	MCC	0,3753

Fuente: elaboración propia.

4.4. Modelo binario aditivo de valor extremo generalizado

En la tabla 9 se aprecia que el modelo BGEVA no considera ninguna variable explicativa en su componente paramétrico, tal como en el caso del GAM. Además, permite observar que las variables significativas en el componente no paramétrico del modelo son los indicadores razón corriente, rotación activos, rentabilidad activos y nivel de endeudamiento. En la figura 2 se evidencian los efectos de las variables significativas sobre la variable dependiente, donde para la razón corriente y el nivel de endeudamiento se tiene un efecto positivo en casi toda su extensión, mientras que para la rotación activos y rentabilidad activos un efecto negativo que se da en un tramo muy pequeño. Es de anotar que la razón corriente no presenta el efecto esperado sobre la variable dependiente según la lógica financiera.

Tabla 9
Estimación del modelo BGEVA

Coeficientes paramétricos				
	Estimación	**Error Estándar**	**Valor Z**	**Pr(>|z|)**
Intercept	-1,615	0,029	-56,500	0,000
Significancia aproximada de los componentes suavizados				
	edf	**Est.rank**	**Chi.sq**	**Valor-P**
raz	1,781	1,990	7,554	0,026
teso	1,006	1,012	0,586	0,445
rota	3,315	3,739	11,489	0,027
margen	5,068	5,469	8,013	0,191
ractiv	6,374	6,934	38,54	0,000
rpatri	1,561	1,819	2,052	0,411
niven	2,029	2,214	56,17	0,000
endeu	2,533	2,803	3,253	0,255
apalc	4,387	4,774	6,615	0,228
apalar	4,067	4,421	5,116	0,323

Fuente: elaboración propia.

En la tabla 10 se observa que el modelo BGEVA, respecto a los tres modelos anteriores, presenta mejores medidas para el recall y el inverse precision, pero una medida inferior para el inverse recall. A saber, el modelo BGEVA clasifica de manera correcta el 76,60% de empresas en riesgo de quiebra con relación a las que en realidad se encuentran en quiebra (recall), del total de empresas que clasifica en riesgo de no quiebra, el 99,78%

de ellas no se encuentran en realidad en quiebra (inverse precision), y clasifica de manera correcta el 97,79% de empresas en riesgo de no quiebra, con respecto a las que en realidad no se encuentran en quiebra (inverse recall). En el caso de la precision, el modelo BGEVA presenta una mayor medida que el modelo logístico pero inferior que el GAM y el modelo GEV, clasificando de forma correcta el 24,00% de empresas en riesgo de quiebra, respecto al total de empresas que predice en riesgo de quiebra (precision).

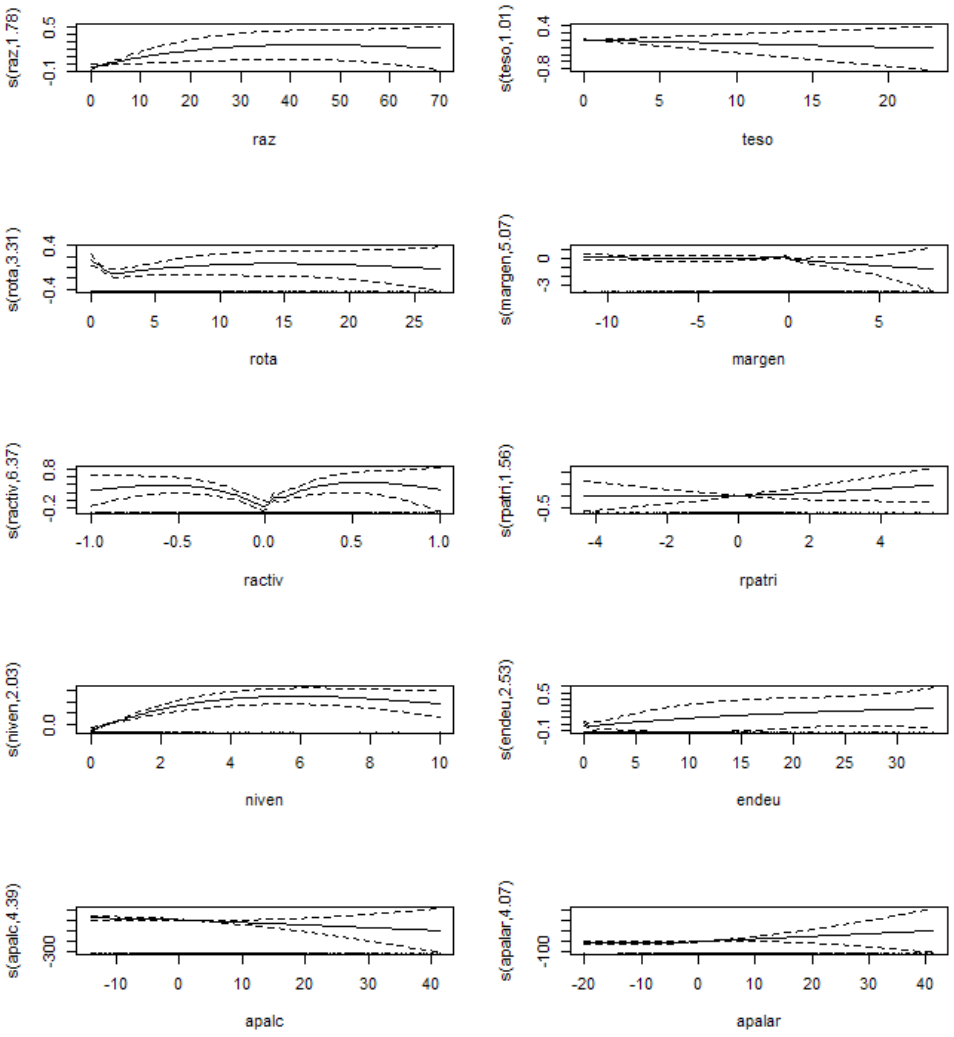

Figura 2. Ajuste variables del modelo BGEVA

Fuente: elaboración propia.

En el caso de las medidas F_1-score y Matthews correlation coefficient (MCC), el modelo BGEVA presenta una mejoría para predecir si una empresa se encuentra o no en riesgo de quiebra respecto a las del modelo logístico, el GAM y el modelo GEV, siendo estas en su orden 0,3655 y 0,4206. Lo anterior puede deberse a que en el modelo BGEVA, al igual que en el modelo GEV, se emplea una función link asimétrica que es apropiada para el estudio de eventos de rara ocurrencia y, que al igual que en el GAM, se usa funciones spline en el proceso de estimación, relajando así el supuesto de linealidad entre las variables independientes y la variable dependiente.

Tabla 10
Matriz de confusión del modelo BGEVA

		Valor real		
	Grupos	**Quiebra**	**No quiebra**	**Total predicción**
Predicción	Quiebra	36	114	150
	No quiebra	11	5.048	5.059
	Total real	47	5.162	5.209
	Recall	0,7660	Precision	0,2400
	Inverse recall	0,9779	Inverse precision	0,9978
	F_1-Score	0,3655	MCC	0,4206

Fuente: elaboración propia.

En la tabla 11 se presentan las medidas calculadas para los cuatro modelos con el fin de facilitar al lector la comparación de los mismos y así poder identificar cuál de ellos es el que presenta mejores resultados sin necesidad de remitirse a la matriz de confusión de cada modelo.

Tabla 11
Resumen medidas por modelos

	Modelos			
Medida	**Logístico**	**GAM**	**GEV**	**BGEVA**
Recall	0,0213	0,0213	0,5532	0,7660
Miss rate	0,9787	0,9787	0,4468	0,2340
Fall-Out	0,0008	0,0004	0,0139	0,0221
Inverse recall	0,9992	0,9996	0,9861	0,9779
Precision	0,2000	0,3333	0,2653	0,2400
False discovery rate	0,8000	0,6667	0,7347	0,7600
False omission rate	0,0088	0,0088	0,0041	0,0022

Continúa →

	Modelos			
Medida	Logístico	GAM	GEV	BGEVA
Inverse precision	0,9912	0,9912	0,9959	0,9978
F_1-Score	0,0385	0,0400	0,3586	0,3655
MCC	0,0626	0,0823	0,3753	0,4206

Fuente: elaboración propia.

5. Conclusiones

En este trabajo se presentó una aplicación del modelo logístico (Ohlson, 1980), el modelo aditivo generalizado (Berg, 2007), el modelo de valor extremo generalizado (Calabrese y Osmetti, 2013) y el modelo binario aditivo de valor extremo generalizado (Calabrese et al., 2016), con el fin de estimar la probabilidad de riesgo de quiebra de las empresas colombianas que reportaron sus estados financieros a la Superintendencia de Sociedades de Colombia para el periodo 2011-2015, utilizando para ello diez de las razones financieras empleadas con frecuencia en trabajos que buscan predecir la fragilidad empresarial, donde se observó que las variables *rentabilidad activos* y *nivel de endeudamiento* resultaron ser significativas en la estimación de los cuatro modelos, siendo el signo en todos los casos consistente con lo establecido en la teoría financiera.

La metodología abordada permitió identificar el efecto que tiene en el poder predictivo del modelo, el asumir en su estructura una función link simétrica o asimétrica, y el asumir una relación lineal o no lineal entre las variables independientes y la variable dependiente. Es de anotar que el emplear una función link asimétrica, contribuye a mejorar el poder predictivo del modelo cuando se trabaja con muestras que contemplan eventos raros, como lo es la quiebra. De otro lado, el no asumir el supuesto de linealidad cuando las variables dependientes no reflejan tal comportamiento, permite un mejor ajuste del modelo al verdadero comportamiento de los datos, y por consiguiente, mejoras en su poder predictivo.

Se observó que los modelos GEV y BGEVA al asumir una distribución de valor extremo en la función link, presentaron un mejor desempeño predictivo respecto a los otros modelos, permitiendo identificar mejor las empresas que se encuentran en riesgo de quiebra. Además, el modelo BGEVA permitió determinar la relación existente entre la probabilidad de default y las variables explicativas al asumir estructuras semi-paramétricas en sus estimaciones, mejorando así su desempeño predictivo respecto al modelo GEV.

Referencias

Altman, E. (1968). Financial Ratios, discriminant analysis and the prediction of corporate bankruptcy. *Journal of Finance*, *23*(4), 589-609.

Altman, E., and Sabato, G. (2006). Modeling credit risk for SMEs: Evidence from the US market, *ABACUS*, *43*(3), 716-723.

Altman, E., Haldeman, R., and Narayanan, P. (1977). ZETA™ Analysis: A new model to identify bankruptcy risk of corporations. *Journal of Banking & Finance*, *1*(1), 29-54.

Anandarajan, M., Lee, P., and Anandarajan, A. (2001). Bankruptcy prediction of financially stressed firms: An examination of the predictive accuracy of artificial neural networks. *International Journal of Intelligent Systems in Accounting, Finance & Management*, *10*(2), 69-81.

Aoki, S., and Hosonuma, Y. (2004). Bankruptcy prediction using decision tree. In *The Application of Econophysics, Proceedings of the Second Nikkei Econophysics Symposium* (pp. 299-302). Tokyo: Springer.

Beaver, W. (1966). Financial ratios as predictors of failure. *Journal of Accounting Research*, 4, 71-111.

Beirlant, J., Goegebeur, Y., Segers, J., and Teugels J. (2004). *Statistics of Extremes: Theory and Applications*. Chichester: John Wiley & Sons, Ltd.

Berg, D. (2007), Bankruptcy prediction by generalized additive models. *Applied Stochastic Models in Business and Industry*, *23*(2), 129-143.

Bernhardsen, E. (2001). *A model of bankruptcy prediction*, Working Paper, Norges Bank. 10.

Brédart, X. (2014). Bankruptcy prediction model: The case of the United State. *International Journal of Economics and Finance*, *6*(3), 1-7.

Calabrese, R., and Osmetti, S. (2013). Modelling small and medium enterprise loan defaults as rare events: The generalized extreme value regression model. *Journal of Applied Statistics*, *40*(6), 1172-1188.

Calabrese, R., Marra, G., and Osmetti, S. (2016). Bankruptcy prediction of small and medium enterprises using a flexible binary generalized extreme value model. *Journal of the Operational Research Society*, *67*(4), 604-615.

Charitou, A., Neophytou, E., and Charalambous, C. (2004) Predicting corporate failure: Empirical evidence for the UK. *European Accounting Review*, *13*(3), 465-497.

Ciampi, F., and Gordini, N. (2008). Using economic-financial ratios for small enterprise default prediction modeling: An empirical analysis. *In 2008 Oxford Business & Economics Conference Proceedings, Association for Business and Economics Research* (pp. 1-21). Oxford: ABER.

Gu, C. (2002). *Smoothing Spline ANOVA Models*. New York: Springer.

Hosmer, D., Lemeshow, S., and Sturdivant, R. (2013). *Applied Logistic Regression*, Hoboken, New Jersey: John Wiley & Sons, Inc.

Jayasekera, R. (2018). Prediction of company failure: Past, present and promising directions for the future. *International Review Of Financial Analysis*, 55, 196-208.

Jenkinson, A. (1955). The frequency distribution of the annual maximum (or minimum) values of meteorological events. *Quarterly Journal of the Royal Meteorological Society*, *81*(348), 158-172.

Laitinen, E., and Laitinen, T. (2000). Bankruptcy prediction: Application of the Taylor's expansion in logistic regression. *International Review of Financial Analysis*, *9*(4), 327-349.

Marra, G. y Radice, R. (2017). *GJRM:* Generalized Joint Regression Modelling, R package version 0.1-1.

Martínez, Ó. (2003). Determinantes de fragilidad en las empresas colombianas. *Borradores de Economía*, 259, Bogotá: Banco de la Republica.

Narváez, L. (2010). *Análisis de la aplicación de los modelos de predicción de quiebras en Colombia*. Tesis pregrado en contaduría pública. Cali: Universidad Autónoma de Occidente.

Ohlson, J. (1980). Financial ratios and the probabilistic prediction of bankruptcy. *Journal of Accounting Research*, *18*(1), 109-131.

O'Leary, D. (1998). Using neural networks to predict corporate failure. *International Journal of Intelligent Systems in Accounting, Finance and Management*, 7, 187-197.

Pérez, J., González, K. y Lopera, M. (2013). Modelos de predicción de la fragilidad empresarial: aplicación al caso colombiano para el año 2011. *Perfil de Coyuntura Económica*, 22, 205-228.

Powers, D. (2011). Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. *Journal of Machine Learning Technologies*, *2*(1), 37-63.

R Core Team (2017). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria.

Ravi Kumar, P., and Ravi, V. (2007). Bankruptcy prediction in banks and firms via statistical and intelligent techniques, a review. *European Journal of Operational Research*, *180*(1), 1-28.

Romero Espinosa, F. (2013). Alcances y limitaciones de los modelos de capacidad predictiva en el análisis del fracaso empresarial, *AD-minister*, (23).

Rosillo, J. (2002). Modelo de predicción de quiebras de las empresas colombianas. *Innovar, revista de ciencias administrativas y sociales*, 19, 109-124.

Santos, M., Cortez, P., Pereira, J., and Quintela, H. (2006). Corporate Bankruptcy Prediction Using Data Mining Techniques, *WIT Transactions on Information and Communication Technologies*, *37*(1), 349-357.

Sobehart, J., Keenan, S., and Stein R. (2000). Validation methodologies for default risk models. *Default Risk*, mayo, 51-56.

Tsay, R. (2010). *Analysis of Financial Time Series*. Hoboken, New Jersey: John Wiley & Sons, Inc.

Vozzella, P., and Gabbi, G. (2010). Default and asset correlation: An empirical study for Italian SMEs. *Working Paper*.

Whaba, G. (1990). *Spline Models or Observational Data*. Philadelphia: Society for industrial and applied mathematics.

Wood, S. (2006). *Generalized Additive Models: An Introduction with R*. Boca Raton: Chapman & Hall/CRC.

Wood, S. (2017). *Mgcv*: Mixed GAM Computation Vehicle with GCV/AIC/REML Smoothness Estimation, R package version, 1.8-17.

Zibanezhad, E., Foroghi, D., and Monadjemi, A. (2011). Applying decision tree to predict bankruptcy. *Proceedings of IEEE International Conference on Computer Science and Automation Engineering*, 4, 165-169.