Crisis empresarial en Colombia
Probabilidad de entrar en proceso de insolvencia: 2016-2019*

Laura Vanessa Hernández-Cruz[1]

Candidata a Mágister en Economía Aplicada
Universidad de los Andes, Colombia
lv.hernandezc@uniandes.edu.co

Abstract. La detección temprana de la probabilidad de que una empresa entre en insolvencia empresarial puede servir como insumo a los diseñadores de política pública para mitigar este fenómeno, así como sus posibles efectos sobre el empleo y el bienestar social. Si bien en Colombia en épocas de crisis económica este fenómeno tiende a exacerbarse, el grado de afectación empresarial va a depender de la naturaleza de la crisis, y puede incidir de forma distinta a las empresas de acuerdo con sus características intrínsecas como su desempeño operativo y financiero, su tamaño o el sector empresarial al que pertenece. Mediante la comparación de técnicas de pronóstico que contemplan diferentes tipos de relacionamiento de variables financieras y económicas, este documento identifica que, en el problema de insolvencia empresarial en Colombia, los predictores se relacionan de formas complejas que no pueden ser detectadas por técnicas tradicionales de pronóstico, por lo cual los modelos de aprendizaje de máquinas, como los árboles de decisión estimados mediante técnicas de *boosting*, generan un mejor desempeño de pronóstico que los modelos lineales.

Keywords: Contraste de modelos · Predicción · Técnicas computacionales · Insolvencia empresarial · *Machine Learning* · Modelos de regularización · *Boosting*.
Códigos JEL: C52, C53, C63, G33.

1 Introducción

En las últimas décadas, Colombia y el mundo han experimentado crisis económicas que han tenido repercusiones en diferentes sectores productivos de la economía. Las empresas, como unidades productivas, se ven directamente afectadas por los choques económicos que desencadenan fluctuaciones de demanda, a punto tal que, bajo diversas circunstancias, pueden quedar en riesgo inminente de quiebra. Una empresa que entra en insolvencia puede llegar a desencadenar pérdidas económicas que deben ser asumidas no sólo por la empresa en crisis, sino por sus acreedores. Además, por una parte, genera pérdidas en capital que en ocasiones no llega a ser reincorporado con rapidez en la economía, y por otra, produce presiones en el mercado laboral, afectando el bienestar social.

En este sentido, se considera importante que los diseñadores de política pública, y en particular la Superintendencia de Sociedades como institución con injerencia en los procesos de

*Tesis presentada a la Facultad de Economía de la Universidad de los Andes, en cumplimiento parcial de los requerimientos para la obtención del título de Maestría en Economía Aplicada.

Se extiende un agradecimiento especial a David Ibáñez Parra, asesor de esta tesis, por sus comentarios y sugerencias en la realización de la misma. dibanezp@uchicago.edu, dibanez@javeriana.edu.co

insolvencia empresarial en el país, cuente con insumos que le permitan identificar las empresas que se encuentran en riesgo de entrar en bancarrota para tomar medidas que anticipen y prevengan su crisis, minimizando los impactos económicos que la insolvencia empresarial genera en la sociedad. Este trabajo busca proporcionar, mediante el uso de diferentes metodologías, un modelo que permita determinar la probabilidad de que una empresa entre en proceso de insolvencia. Para esto, se realiza un análisis comparativo en términos de calidad de pronóstico tanto de las metodologías que tradicionalmente se utilizan en la literatura, basadas en modelos lineales, como de metodologías de *machine learning*, que se basan en relaciones complejas y no lineales. Lo anterior, se realiza teniendo como base variables financieras, económicas y de estructura, de empresas insolventes y no insolventes de Colombia, con datos entre 2016 y 2019.

Este documento se encuentra estructurado en ocho secciones incluyendo esta introducción. En la segunda sección, se desarrolla brevemente el estado del arte en el que se menciona parte de la literatura que ha abordado la temática de insolvencia empresarial. En la tercera sección se realiza una breve caracterización del fenómeno de insolvencia en el país de acuerdo con las características de las empresas, en particular, con el tamaño y sector empresarial al que pertenecen. En la cuarta sección se hace una descripción de los datos que sirvieron de insumo para desarrollar el estudio, así como de su preprocesamiento. La quinta sección explica brevemente las diferentes metodologías de pronóstico. Los resultados son comparados en la sexta sección, en la cual también se hace un análisis de su robustez, mediante la comparación de la densidad de probabilidades generadas por los modelos estimados. Posteriormente, en la séptima sección se realizan dos ejercicios que muestran algunas implicaciones de política pública derivadas de la ejecución de los modelos de pronóstico planteados; estos ejercicios muestran las ventajas que tiene utilizar modelos que identifiquen las relaciones no lineales entre las variables predictoras en los modelos de pronóstico, tanto en términos de costos de estimación como de eficiencia, así como la capacidad de pronóstico de estos modelos segmentados por tamaño y por sector económico. Finalmente se presentan las conclusiones.

2 Estado del arte

El papel de las empresas en la sociedad parte de su concepción como unidad productiva y como fuente generadora de empleo, rol que de acuerdo con Freeman (1984), se ve ampliado a la importancia que las decisiones empresariales tienen sobre los grupos de interés que pueden influir en las actividades desarrolladas por la misma, como los accionistas, empleados, organizaciones sindicales, proveedores, clientes, instituciones reguladoras, entre otros [Sánchez Jiménez, 2015]. De esta manera, las empresas se constituyen como un pilar determinante en el crecimiento económico y el desarrollo social.

Por lo anterior, existen diversas consecuencias que son asumidas tanto por la sociedad como por las firmas cuando éstas entran en insolvencia empresarial. Ésta es entendida como la incapacidad que adquieren las empresas de pagar sus deudas, y que resulta en la imposibilidad de continuar ejecutando su actividad económica.

En la actualidad existe una amplia serie de investigaciones que buscan establecer modelos que predigan el riesgo que tienen las empresas de entrar en una crisis financiera. Uno de los estudios pioneros con mayor acogida en la literatura es el de Altman [1968], quien utilizó un modelo de análisis discriminante para, basado en variables financieras, determinar si una empresa estaba en riesgo de convertirse en una "empresa fallida". En particular, este estudio se basó en el análisis de 66 empresas de Estados Unidos, del sector manufacturero, cotizantes en bolsa, y estableció un puntaje denominado Z-Score, de acuerdo con el cual y dado un

umbral determinado, se puede inferir si la empresa está o no en riesgo de insolvencia. Este modelo ha sido ampliado y recalibrado a lo largo de los años para poder ser aplicado en empresas no cotizantes en bolsa o de sectores no manufactureros [Altman, 2000].

Cabe señalar que a pesar de que los coeficientes y los umbrales del modelo Z-Score de Altman son estimados principalmente con base en empresas estadounidenses [1], su acogida en la literatura se evidencia en que estos mismos coeficientes se aplican incluso para otros países como Jordania [Alareeni and Branson, 2013], Bangladesh [Hamid et al., 2016] y Grecia [Gerantonis et al., 2009], por citar algunos ejemplos, arrojando según los diferentes autores, resultados con buenos ajustes.

La segunda técnica más utilizada en la determinación de insolvencia empresarial corresponde a modelos de regresión logística, implementados por primera vez por Ohlson [1980]. El uso de este tipo de modelos surgió bajo la crítica al modelo de Altman, según la cual el puntaje del análisis discriminante carecía de interpretación intuitiva. En este sentido, este modelo busca capturar los factores que determinan el riesgo de que una empresa caiga en estado de insolvencia, permitiendo que, contrario al modelo de Altman, los parámetros de estimación de los umbrales de insolvencia no sean estáticos [Ibánez, 2020]. De los dos modelos previamente enunciados, han partido gran número de estudios que han contribuido a su sofisticación, por ejemplo, incluyendo otro tipo de variables (no financieras), o ampliando el número y las características de las empresas analizadas.

De otro lado, gracias al avance en los sistemas de recolección de datos, en la actualidad la disponibilidad y calidad de información a nivel de empresa ha mejorado considerablemente, lo que ha impulsado el uso de herramientas estadísticas más avanzadas como lo son las técnicas de *machine learning*, en el análisis de riesgo crediticio. Siguiendo esta idea, Barboza et al. [2017] desarrollaron un estudio en el que hacen una comparación entre los métodos estadísticos tradicionales (análisis discriminante y modelos de regresión logística) con las técnicas de *machine learning*, en la predicción de insolvencia empresarial. En particular, estos autores utilizan una base de más de 10 mil empresas públicas de Estados Unidos, con información entre 1985 y 2013. Entendiendo que el contexto colombiano tiene particularidades que afectan en diferente medida a las empresas de acuerdo con factores como el tamaño y el sector económico al que pertenece, y siguiendo el trabajo de los autores antecitados, este trabajo busca comparar el desempeño predictivo de modelos tradicionales, con el de estimaciones derivadas del uso de técnicas de *machine learning*. De esta manera, se analizará el desempeño de modelos que basan sus estimaciones en funciones lineales entre los predictores, con modelos que consideran relaciones más complejas e incluso no lineales entre los mismos. Estos análisis se desarrollarán de forma tal que evalúen el desempeño no sólo de variables financieras, sino también de variables macroeconómicas y de variables que capturen otras características intrínsecas de las firmas.

3 Contexto

En Colombia, los procesos de insolvencia empresarial son competencia de la Superintendencia de Sociedades, entidad que actúa como juez del proceso de insolvencia para todas empresas que adelanten dicha solicitud. El régimen judicial de insolvencia es regulado mediante la Ley 1116 de 2006 y los decretos 560 y 772 de 2020 [2]. En particular, la Ley 1116 de 2006 tiene

[1] Altman también ha hecho extensiones calibrando el modelo para el caso de México, como guía para la calibración del modelo en países en desarrollo [Altman, 2000].

[2] Estos son decretos transitorios, y surgieron a raíz de la crisis sanitaria originada por la pandemia del COVID-19. El decreto 560 de 2020 es aquel "por el cual se adoptan medidas transitorias especiales

como objeto "la protección del crédito y la recuperación y conservación de la empresa como unidad de explotación económica y como fuente generadora de empleo, a través de los procesos de reorganización y de liquidación judicial, siempre bajo el criterio de agregación de valor" [Congreso de la República, 2006].

Dicho régimen consta de dos tipos de procesos, los cuales pueden ser solicitados bajo una persona natural o jurídica [Ibánez, 2020]:

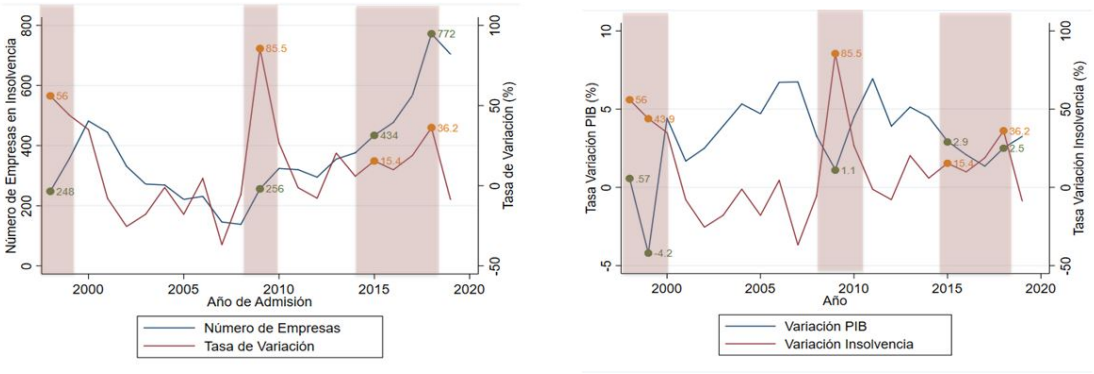
- **Proceso de reorganización**, se establece cuando la persona o empresa solicitante se encuentra en dificultad para cumplir sus obligaciones o está a punto de cesar pagos a sus proveedores. Los objetivos de este proceso son:
  - La recuperación y conservación de la empresa como unidad de explotación económica y fuente generadora de empleo.
  - La preservación de empresas viables.
  - La normalización de las relaciones comerciales y crediticias de la empresa solicitante mediante su reestructuración operacional, administrativa, de activos o pasivos.
  - La protección del crédito y de los proveedores.
- **Proceso de liquidación**, se establece cuando la persona o empresa solicitante incumple los acuerdos de reorganización del proceso anterior, o incumple alguna de las causales de liquidación judicial inmediata previstas en la Ley 1116 de 2006. Los objetivos de este proceso son:
  - La liquidación pronta y ordenada de las empresas solicitantes.
  - El aprovechamiento de los bienes del deudor.

Lo que lleva a que una empresa pueda llegar a entrar en estado de insolvencia son típicamente los problemas de flujo de efectivo, los cuales pueden ser causados por un bajo desempeño en sus ventas o por un alto nivel de sus gastos operativos. Ante esta situación, la empresa puede llegar a aumentar sus préstamos a corto plazo, y en la medida en que esta situación no mejore, se hace más probable que la empresa no pueda pagar sus deudas y que entre en insolvencia.

De acuerdo con Ibánez [2020], como hecho estilizado, se ha observado que las crisis económicas aumentan la probabilidad de que una empresa entre en bancarrota. Al respecto, el autor evidencia que en las últimas dos décadas, Colombia ha experimentado cuatro momentos importantes de moderación de crecimiento económico: en 1999, causado tanto por presiones internas como por presiones internacionales; en 2008-2010 y 2015-2018, causados principalmente por un entorno internacional inestable; y la más reciente desde 2020, producto de la pandemia del COVID-19. Como se observa en la Figura 1a, durante los momentos de crisis se presentan repuntes en la tasa de variación de empresas solicitantes de procesos de insolvencia; así mismo, en la Figura 1b se observa que esos momentos coinciden con la moderación en la tasa de variación del PIB. De esta manera se evidencia la relación inversa que existe entre la tasa de variación de empresas en insolvencia y la tasa de variación de la actividad económica del país.

---

en materias de procesos de insolvencia, en el marco del Estado de Emergencia, Social y Ecológica". Por su parte, el decreto 772 de 2020 es aquel "por el cual se dictan medidas especiales en materia de procesos de insolvencia, con el fin de mitigar los efectos de la emergencia social, económica y ecológica en el sector empresarial".



(a) Evolución histórica (b) Insolvencia vs PIB real

Fig. 1: Serie histórica empresas en insolvencia en Colombia

Nota: Las áreas sombreadas corresponden a los tres momentos de moderación del crecimiento económico: 1999, 2008-2010, 2015-2018.

Fuente: Superintendencia de Sociedades

El impacto de las crisis ha sido heterogéneo entre las firmas y, por ende, en su probabilidad de caer en insolvencia. En Colombia, la mayor proporción de empresas se configuran como pequeñas y medianas; en efecto, de acuerdo con Confecámaras [2021], la tasa de creación de empresas en el país se conforma principalmente por microempresas (99,53%), seguido por pequeñas empresas (0,44%). La estructura del aparato empresarial explica que, como se evidencia en la Tabla 1, la mayor proporción de empresas que entran en estado de insolvencia ante fuertes cambios de la economía sean las micro y pequeñas empresas, las que a su vez son las más vulnerables ante choques negativos de demanda al no contar con un músculo financiero lo suficientemente fuerte para afrontar los periodos de crisis.

Año admisión	**Total empresas**	**% por tamaño de empresas**		
		Grandes	**Medianas**	**Pequeñas y Micro**
1999	357	12,6%	17,6%	69,7%
2009	256	7,8%	18,4%	73,8%
2015	434	9,2%	18,2%	72,6%
2020*	997	5,9%	14,3%	79,7%

Tabla 1: Empresas en insolvencia, por tamaño

Nota: * Datos preliminares para 2020

Fuente: Superintendencia de Sociedades

Por otra parte, vale la pena resaltar que el efecto de la moderación del crecimiento económico varía dependiendo de la naturaleza del evento que lo genera. Así, la crisis del año 1999, relacionada con el sector vivienda, generó un mayor impacto en la tasa de insolvencia en el sector de la construcción; la crisis de 2008-2010, cuyo efecto en el país se manifestó en presiones a la demanda, afectó principalmente a los sectores de manufacturas y comercio; la moderación económica de 2015-2018, que estuvo relacionada con los efectos desencadenados por la crisis previa de la deuda en Europa, afectando los niveles de demanda mundial, y que adicionalmente, se presentó a la par con la fuerte caída de los precios del petróleo a nivel internacional, tuvo una mayor afectación en el sector minero y el sector de manufactura;

finalmente, la crisis relacionada con el COVID-19, ha tenido un efecto mayor en los sectores de comercio y servicios (ver Tabla 2).

Año admisión	Total empresas	Macrosector					
		Servicios	Comercio	Agricultura	Construcción	Manufactura	Minero
Totales y participación %							
1999	357	48,2%	22,4%	6,4%	12,3%	10,6%	0,0%
2009	256	55,5%	28,5%	5,1%	8,6%	2,3%	0,0%
2015	434	36,6%	26,0%	7,1%	12,4%	14,1%	3,7%
2020*	997	43,0%	25,4%	4,9%	9,9%	15,2%	1,5%
Variación frente al año anterior							
1998-99	44,0%	39,8%	12,7%	64,3%	131,6%	90,0%	-100,0%
2008-09	85,5%	94,5%	121,2%	-31,6%	83,8%	500,0%	-
2014-15	15,4%	-4,8%	5,6%	34,8%	25,6%	79,4%	700,0%
2019-20*	41,6%	81,8%	63,2%	-12,5%	7,6%	-1,9%	50,0%

Tabla 2: Empresas en insolvencia por sector (participación y variación)

Nota: * Datos preliminares para 2020
Fuente: Superintendencia de Sociedades

Los hechos anteriores sugieren que en el análisis para determinar la probabilidad de que una empresa entre en estado de insolvencia puede ser pertinente incluir, además de las variables financieras ampliamente aceptadas en la literatura, factores como el sector económico y el tamaño de la empresa. El propósito de las siguientes secciones es el de identificar un modelo que permita llevar a cabo pronósticos precisos que puedan servir como insumos de política pública para detectar, de manera temprana, la probabilidad de que una empresa entre en insolvencia en Colombia y de ser el caso, aplicar las acciones de mitigación que se consideren pertinentes.

4 Datos

Los datos que se utilizan para el desarrollo de este estudio se basan principalmente en los registros contables [3] de las empresas que son supervisadas y requeridas [4] por la Superintendencia de Sociedades, los cuales cuentan con información tanto de empresas admitidas en proceso de insolvencia, como de empresas que no están en dicho proceso [5]. Cabe resaltar que, debido a la calidad de la información original, la base utilizada en este estudio corresponde a una muestra de la totalidad de registros con los que cuenta la Superintendencia; sin embargo, se considera que la base tiene una cantidad de registros suficiente para desarrollar el análisis. En suma, la base utilizada cuenta con información de 82.450 empresas, de las cuales 1.853

[3] Los registros en los que se basa este estudio incluyen la información contable de los Estados de Situación Financiera, Estado de Resultados y Estado de Flujos de Efectivo, principalmente.

[4] Actualmente, la base de dichas empresas cuenta con información de aproximadamente 30.000 sociedades. Sin embargo, es una muestra variable, es decir, no cuenta con información de todas las empresas para todos los periodos de tiempo.

[5] Es importante mencionar que una vez las empresas solicitan el proceso, éstas pueden entrar en insolvencia, sin que ello implique que entren en un proceso de insolvencia empresarial jurídico. Esto implica que la solicitud del proceso no siempre es exitosa, pues puede verse afectada ante el no cumplimiento o reporte de la totalidad de los requisitos solicitados. Por lo anterior, este estudio se centra en las empresas que efectivamente entran en proceso de insolvencia empresarial jurídico.

entraron en algún momento en proceso de insolvencia [6], es decir, sólo el 2,2% de los registros de la base.

La Tabla 3 muestra la caracterización de las empresas no insolventes e insolventes, en términos del año observado, el sector económico al que pertenecen y el tamaño. Por una parte, puede verse que el desbalance de la base se mantiene durante todos los años de observación. En cuanto a la caracterización por sectores, ambos grupos de empresas tienen una distribución similar, salvo por los sectores de servicios y manufacturas, pues en el primero, la proporción es 24 p.p.[7] mayor en el caso de las empresas no insolventes, mientras que para el segundo caso, la proporción es inferior en este grupo por 17,8 p.p. Finalmente, en cuanto al tamaño de las empresas, tanto en el caso de insolventes como de no insolventes, se tienen una mayor representatividad de las empresas medianas.

		No insolventes		**Insolventes**	
Variable	**Valores**	**Frecuencia**	**Proporción**	**Frecuencia**	**Proporción**
Año	2016	21.391		471	
	2017	18.305		445	
	2018	17.815		437	
	2019	23.086		500	
	Total	**80.597**		**1.853**	
Variable	**Valores**	**Frecuencia**	**Proporción**	**Frecuencia**	**Proporción**
Sector	Agropecuario	4.970	6,2%	150	8,1%
	Comercio	20.540	25,5%	479	25,8%
	Construcción	9.748	12,1%	277	14,9%
	Manufactura	12.181	15,1%	609	32,9%
	Minero-Hidrocarburos	1.768	2,2%	61	3,3%
	Servicios	31.383	38,9%	277	14,9%
	No informa	7	0,0%	-	0,0%
	Total	**80.597**	**100,0%**	**1.853**	**100,0%**
Variable	**Valores**	**Frecuencia**	**Proporción**	**Frecuencia**	**Proporción**
Tamaño	Grande	16.510	20,5%	364	19,6%
	Mediana	34.096	42,3%	830	44,8%
	Pequeña	29.991	37,2%	659	35,6%
	Total	**80.597**	**100,0%**	**1.853**	**100,0%**

Tabla 3: Caracterización de empresas no insolventes e insolventes

Fuente: Elaboración propia

Con base en los indicadores considerados por Altman [2000] y Ohlson [1980], se calculó una batería de 27 variables financieras, las cuales se clasifican en cinco categorías: liquidez, apalancamiento, operacionales, rentabilidad, solvencia y efectividad; estos indicadores se encuentran detallados en el Anexo A. . Cabe resaltar que a partir de 2016, en Colombia, las empresas se acogieron a la Norma Internacional de Información Financiera (NIIF), lo que no permite que los indicadores financieros que se calculen antes de dicho periodo sean comparables con los de los últimos años. Por esta razón, este estudio se centra en el análisis del periodo 2016-2019.

Adicionalmente, y con el fin de poder identificar si las variables de contexto económico tienen incidencia en la probabilidad de que una empresa entre en proceso de insolvencia,

[6] Se considera que una empresa entra en estado de insolvencia cuando ha sido admitida por la Superintendencia de Sociedades, tanto para proceso de reorganización como para proceso de liquidación.

[7] p.p.: puntos porcentuales

se consideraron nueve variables adicionales relacionadas con el valor agregado del sector económico al que pertenece la empresa, la incidencia que el sector tiene en el empleo departamental, la tasa de desempleo del departamento, y variables relacionadas con el comercio exterior. Estas variables se encuentran detalladas en el Anexo B. Dentro del análisis también se consideró el tamaño de la empresa y su sector económico. De esta manera, la base considera un total de 38 variables.

4.1 Preprocesamiento de los datos

Debido a la ausencia de algunos datos en los registros contables, esta base tuvo 486.487 datos faltantes, a lo largo de las diferentes variables. Para poder contar con la mayor información posible en los modelos, estos datos se imputaron mediante el método de imputación multivariada de árboles aleatorios *(random forest)*, el cual es uno de los métodos más adecuados cuando se cuenta con datos de alta dimensión y que contemplan relaciones complejas entre variables [Song, 2018]. Cabe aclarar que en el entrenamiento de los árboles aleatorios a partir de los cuales se predijeron los datos faltantes, no se utilizó como predictor la variable de insolvencia.

Tras dicho proceso se logró imputar el 93,4% de los datos faltantes, pasando a tener 32.214 datos no observados. Esto a raíz de la no imputación de una de las variables de rentabilidad [8], y de otros datos que no fueron calculados por el algoritmo. Así, luego de descartar la variable no imputada, la base presentó 1.160 datos que no fueron calculados, distribuidos en cuatro variables, evidenciando una reducción importante de datos faltantes en la base. Las estadísticas descriptivas de las variables ya imputadas se encuentran detalladas en el Anexo C.

Cabe mencionar que en el desarrollo de los modelos de clasificación no se consideraron las empresas que tuvieran algún dato faltante. De esta manera, la base utilizada para los análisis pasó a tener información de 79.656 empresas no insolventes y 1.853 empresas insolventes.

4.2 Partición de la muestra y técnicas de remuestreo

Es importante considerar que típicamente, al ejecutar modelos de pronóstico, se suele hacer una partición aleatoria de los datos, generando una submuestra de entrenamiento con base en la cual se estiman los parámetros de interés, y otra submuestra de prueba, cuyas observaciones no son consideradas en los métodos de estimación. De esta manera, al conocer los valores de interés en la submuestra de prueba, éstos pueden compararse con los valores que son pronosticados a raíz del modelo estimado, lo que permite establecer su calidad de pronóstico.

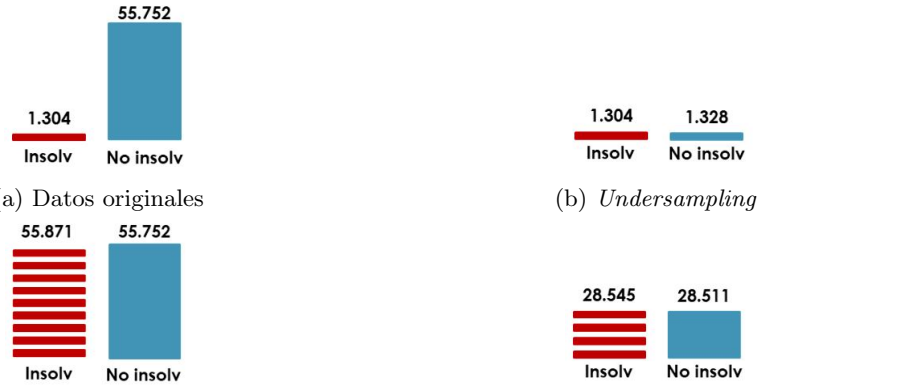
Así, la base de datos fue particionada aleatoriamente de forma tal que la submuestra de entrenamiento contara con el 70% de los datos, y la submuestra de prueba con el 30% restante [9]. Ambas submuestras cuentan con la misma proporción de empresas no insolventes e insolventes (aproximadamente 98% y 2%). De esta forma, la submuestra de entrenamiento cuenta con 57.056 observaciones, de las cuales 55.752 corresponde a empresas no insolventes y 1.304 a empresas insolventes.

[8] La variable no imputada fue la correspondiente a la rentabilidad acumulada del año anterior (R5 en los Anexos).

[9] La partición de los datos en proporción de 70%-30% se realizó buscando tener una submuestra de prueba lo suficientemente grande que permita validar de forma más precisa los resultados de los modelos de pronóstico. Cabe resaltar que esta proporción puede variar de acuerdo con el criterio del investigador.

Ahora, como ya se ha mencionado previamente, con una distribución de aproximadamente 98% de empresas no insolventes y 2% insolventes, la base de datos de este estudio se encuentra altamente desbalanceada. Varios algoritmos, entre los que se encuentran los métodos de clasificación, son altamente sensibles a este tipo de datos, pues la tasa de precisión *(accuracy)* del modelo se ve demasiado influenciada por los datos de los que se tiene más información. Es importante, por lo tanto, utilizar métodos que permitan equilibrar los datos antes de introducirlos en los modelos.

En este sentido, unas de las técnicas más usadas en la literatura para tratar el problema de datos desbalanceados son las técnicas de remuestreo. Estas técnicas consisten en transformar la base de datos en una base más balanceada, bien sea añadiendo observaciones a la clase minoritaria *(oversampling)*, eliminando observaciones de la clase mayoritaria *(undersampling)*, o combinando estos dos métodos [Charfaoui, 2019]. Para los datos de este estudio, las técnicas de remuestreo se ejecutaron mediante métodos aleatorios, cuya representación gráfica se muestra en la Figura 2.



(a) Datos originales (b) *Undersampling*

(c) *Oversampling* (d) *Both: Undersampling - Oversampling*

Fig. 2: Representación gráfica de la distribución de los datos por las técnicas de remuestreo, en la submuestra de entrenamiento

Fuente: Elaboración propia

Así, los modelos de pronóstico se estimaron a partir de las submuestras derivadas de las tres técnicas de remuestreo para determinar con cuál de ellas se obtiene el mejor desempeño o calidad de pronóstico.

5 Metodología

El desarrollo metodológico de este estudio contempla la comparación del desempeño del pronóstico que se obtiene tras la aplicación de los coeficientes del modelo Z-Score de Altman a las variables financieras de las empresas colombianas, siguiendo la práctica mencionada en la sección 2, según la cual estos coeficientes han sido utilizados para el cálculo de riesgo de quiebra empresarial en diferentes países. Los modelos posteriores corresponden a un modelo logarítmico, dos modelos de regularización *(lasso y elastic net)* y un algoritmo de *boosting* aplicado a árbones de decisión *(XGBoost)*, los cuales se calibran considerando las variables enunciadas en la sección anterior.

Es importante señalar que, tanto el modelo logístico como el modelo de análisis discriminante (en el cual se basa el modelo Z-Score) basan sus estimaciones en espacios lineales; por

su parte, los modelos de regularización, que también consideran relaciones lineales entre los predictores, permiten a su vez incorporar relaciones más complejas entre estos; mientras que los modelos de árboles de decisión por *boosting* incorporan relaciones complejas y no lineales entre predictores. En este sentido, el desempeño de los modelos va a estar supeditado al tipo de relaciones que en efecto presenten las variables en el mercado colombiano.

Por otra parte, las calidades de pronóstico de dichos modelos se evaluarán principalmente mediante las medidas de especificidad (tasa de verdaderos negativos), sensibilidad (tasa de verdaderos positivos) y *accuracy* (tasa de empresas correctamente clasificadas), métricas que se obtienen a partir de la matriz de confusión y que pueden ser analizadas gráficamente mediante la curva ROC *(Receiver Operating Characteristic)*. En particular, la clasificación pronosticada de insolvencia se asignará, para todos los modelos, cuando la probabilidad estimada sea superior a 0,5; por otra parte, se realizará una comparación de las probabilidades, para determinar el grado de sensibilidad de los modelos ante diferentes puntos de corte al inicialmente planteado. Todos estos resultados serán analizados y comparados en la séptima sección, previo a lo cual, a continuación, se explican brevemente las técnicas de estimación que se implementaron como parte del desarrollo metodológico y que fueron mencionados al principio de esta sección.

5.1 Modelo Z-Score de Altman

Como se mencionó en la sección 2, el modelo Z-Score de Altman es uno de los más acogidos en la literatura, lo que se evidencia en la estimación del riesgo de insolvencia empresarial mediante estos coeficientes, en diferentes países y contextos socioeconómicos. Es interesante, por lo tanto, calcular las estadísticas de especificidad, sensibilidad y *accuracy* que se obtienen tras aplicar los coeficientes del modelo sobre los datos de las empresas de la submuestra de prueba, y compararlos con los eventos observados de insolvencia empresarial.

Dando un contexto metodológico, cabe mencionar que el modelo de Altman parte de la estimación de un modelo de análisis discriminante [10]. Dicha metodología parte de modelar de forma separada la distribución de las variables independientes X, para cada una de las clasificaciones de la variable Y, y luego, basado en el teorema de Bayes, estimar la probabilidad condicional [James et al., 2013]:

$$P(Y = k \mid X = x) \tag{1}$$

La metodología de Análisis Discriminante Lineal (LDA por sus siglas en inglés) seguida por Altman [1968] [11], tomó como base un conjunto de empresas estadounidenses manufactureras. El modelo resultante, el cual puede aplicarse para determinar el "puntaje" de cualquier empresa $\rangle$, se estructura de la siguiente manera:

$$Z - Score_i = 1,2_i c_1 + 1,4_i c_2 + 3,3_i c_3 + 0,6_i c_4 + 0,99_i c_5 \tag{2}$$

Siendo:

[10] El análisis discriminante lineal asume que las observaciones, dentro de cada grupo de clasificación, se comportan bajo una distribución normal multivariada, con un vector de medias específico por clase y una matriz de covarianza común a todas las clases. De forma alternativa, se encuentra el análisis discriminante cuadrático (QDA), el cual asume que cada grupo tiene su propia matriz de covarianza.

[11] En su paper, Altman [2000] hace referencia a un modelo alternativo llamado modelo ZETA, en el que se testean tanto modelos LDA como QDA. Los coeficientes de dicho modelo son de propiedad privada y por lo tanto no se encuentran publicados en dicho documento.

- $c_1 = \frac{Capital\ circulante}{Activos\ totales}$.
- $c_2 = \frac{Beneficios\ no\ distribuídos}{Activos\ totales}$.
- $c_3 = \frac{EBITDA}{Activos\ totales}$.
- $c_4 = \frac{Capitalización\ bursátil}{Deuda\ total}$.
- $c_5 = \frac{Ventas\ netas}{Ventas\ totales}$.

De acuerdo con el puntaje obtenido por las empresas, Altman estableció los siguientes umbrales de clasificación para determinar el "riesgo" de que la empresa entre en situación de insolvencia:

$$Zonas = \begin{cases} segura & \text{, si } Z - Score_i > 2,99, \\ gris & \text{, si } 1,81 < Z - Score_i < 2,99, \\ riesgo & \text{, si } Z - Score_i < 1,81, \end{cases} \tag{3}$$

Entendiendo las limitadas características de las empresas tomadas originalmente, en años posteriores Altman planteó modificaciones a su modelo, generando modelos alternativos. Algunos de los más conocidos son los que se enuncian en las Ecuaciones 4 y 5. El primero de ellos se calibró para determinar la probabilidad de insolvencia de empresas no manufactureras, eliminando el último componente de la ecuación original (c_5); el segundo modelo se calibró para incluir empresas que no cotizaran en bolsa, modificando el cuarto componente de la ecuación original (c_4) por ($\hat{c_4} = \frac{Patrimonio}{Pasivo}$) [Altman, 2000].

$$Z - Score'_i = 6,5_i c_1 + 3,26_i c_2 + 6,72_i c_3 + 1,05_i c_4 \tag{4}$$

$$Z - Score''_i = 0,717_i c_1 + 0,847_i c_2 + 3,107_i c_3 + 0,420_i \hat{c_4} + 0,998_i c_5 \tag{5}$$

Los puntajes obtenidos de los diferentes modelos se contrastan bajo los mismos umbrales de clasificación enunciados en la Ecuación 3.

5.2 Modelo logístico

El análisis de regresión logística, utilizado por primera vez en la predicción de quiebra empresarial por Ohlson [1980], busca capturar los determinantes de insolvencia. Este modelo parte de una probabilidad condicional, donde la variable dependiente corresponde a una variable dicotómica que toma valores de 0 cuando la empresa es solvente, y valores de 1 si es insolvente; las variables dependientes suelen ser variables que reflejen el desempeño financiero de la empresa. Esta probabilidad condicional es expresada mediante:

$$P(Y = 1 \mid X) = f(X'\beta) \tag{6}$$

Siendo:

$$P(Y = 1 \mid X) = \frac{e^{X'\beta}}{1 + e^{X'\beta}} = \frac{exp\{\beta_0 + \beta_1 x_1 + ... + \beta_k x_k\}}{1 + exp\{\beta_0 + \beta_1 x_1 + ... + \beta_k x_k\}}$$

Cuando esta función es utilizada con fines de clasificación, se puede determinar un punto de corte p, el cual corresponde a la probabilidad p a partir de la cual los valores de pronóstico se interpretan como se enuncia en la Ecuación 7. Por lo general este punto de corte o umbral se toma en 0,5.

$$\hat{y_i} = \begin{cases} 0 \text{ (no insolvente)} & \text{, si } p_i < p, \\ 1 \text{ (insolvente)} & \text{, si } p_i > p, \end{cases} \tag{7}$$

Como técnica para seleccionar las variables que mejor determinan la insolvencia empresarial estimada a partir de los modelos *logit*, se utilizó un método híbrido que combina la selección progresiva de variables *(forward stepwise selection)* y la eliminación progresiva de variables *(backward stepwise selection)*. Este método estima múltiples modelos añadiendo progresivamente variables, comprobando que por cada nueva variable exista una mejora en el ajuste del modelo; así mismo, va removiendo las variables que dejen de proporcionar una mejora en el ajuste. Este es un proceso iterativo que permite escoger las variables que más información aportan a la estimación, de acuerdo con su desempeño dentro de muestra [James et al., 2013].

5.3 Métodos de regularización: Lasso y Elastic Net

Los métodos de regularización se consideran como otra aproximación bajo la cual se puede hacer una selección de variables (similar a los métodos de *forward* y *backward selection*). Esta técnica consiste en estimar el modelo con todas las variables predictoras, regularizando los coeficientes de forma tal que tiendan a cero; esta regularización hace que la varianza de los coeficientes se reduzca significativamente [James et al., 2013].

Dentro de los métodos de regularización más utilizados se encuentran las regresiones *Lasso, Elastic Net* y *Ridge*; en este estudio se utilizaron las dos primeras. Las técnicas de regularización consisten en estimar el modelo (en este caso, logístico), añadiendo un término de penalización al problema de optimización de la función de máxima verosimilitud. . Siguiendo a Awati [2017], el problema de regularización para las regresiones *Lasso* y *Elastic Net* pueden escribirse como se muestra en las Ecuaciones 8 y 9, siendo L el negativo de la función de máxima verosimilitud del modelo; β_s el vector de coeficientes relacionados con las $p-1$ variables (no se penaliza el intercepto); α un hiperparámetro que determina la magnitud del tipo de penalización que se aplica en la regresión por *Elastic Net*; y λ el parámetro que determina el peso de la penalidad aplicada a los coeficientes.

$$\min_{\beta} L(\beta) = L + \lambda \sum_{s=2}^{p} |\beta_s| \tag{8}$$

$$\min_{\beta} L(\beta) = L + \lambda \sum_{s=2}^{p} (\alpha \beta_s^2 + |\beta_s|) \tag{9}$$

Como se observa en las ecuaciones anteriores, la diferencia en los modelos se centra en el tipo de penalidad impuesta. Así, la regresión por *Lasso* impone una penalidad que fuerza a algunos coeficientes a que tomen valores de cero a medida que λ aumenta, mientras que la regresión por *Elastic Net* Net impone una penalidad un poco más flexible al incluir un término que penaliza menos los coeficientes que originalmente están menos alejados de cero. Cabe resaltar que si λ es igualado a 0, las ecuaciones expresarían el problema optimización de la función de máxima verosimilitud, el cual en este caso sería exactamente el mismo que se resuelve en el modelo logístico.

Siguiendo a James et al. [2013], la introducción del término de penalización hace que, a medida que λ aumente, los coeficientes estimados disminuyan su varianza (aumentando su

sesgo). Cabe mencionar que una mayor varianza de los coeficientes implica que un cambio en el set de datos que se utilicen como entrenamiento pueda llegar a causar un cambio significativo en los coeficientes estimados, por lo que los modelos de regularización implican coeficientes más estables. Por otra parte, este tipo de métodos pueden estimar modelos con gran número de variables, modelando inclusive datos donde se cuente con más variables que observaciones; esto lo hacen logrando disminuir considerablemente la varianza, a costo de pequeños incrementos en el sesgo de los coeficientes.

En estas metodologías la selección del mejor modelo se centra en la selección del λ. Para esto, los algoritmos siguen una "ruta de regularización" en donde se estiman diferentes modelos, empezando con un λ tan grande tal que todos los coeficientes sean iguales a cero, y siguiendo con λ cada vez más pequeños. Como mecanismo de selección se pueden seguir dos caminos, bien sea utilizando los criterios de información [12] o mediante métodos de validación cruzada [13]. En este caso se hizo la comparación de pronósticos para ambos casos, y se seleccionó el de mejor desempeño.

5.4 Métodos de boosting - XGBoost

Los métodos predictivos convencionales, como el logarítmico, generan modelos que resultan en una única ecuación que se estima mediante funciones lineales sobre las variables predictoras. Por el contrario, los modelos de árboles de decisión se construyen de forma tal que son capaces de identificar interacciones complejas y no lineales entre dichas variables. Así, los árboles de decisión son técnicas no paramétricas que se forman por reglas binarias (si/no), a partir de las que distribuye las observaciones en función de sus atributos, prediciendo así el valor de la variable respuesta [Amat, 2017]. Por si solos, los árboles de decisión no tienen una buena capacidad predictiva, pues tienen una alta tendencia a la sobreestimación *(overfitting)* y a presentar alta varianza [James et al., 2013].

Por su parte, los métodos de *boosting* son estrategias que se emplean para complementar diferentes métodos de *machine learning*, tanto de regresión como de clasificación, con el fin de lograr un equilibrio entre sesgo y varianza. Estas estrategias buscan lograr mejores predicciones de las que se logran de forma previa a su implementación.

Siguiendo a Amat [2017], la idea detrás del *boosting* aplicado a árboles de clasificación consiste en ajustar de forma secuencial múltiples árboles sencillos, con pocas ramificaciones, que generen una predicción ligeramente mejor a lo esperado por azar *(weak learners)*, de forma tal que cada nuevo árbol emplee la información del modelo anterior para aprender de sus errores e ir mejorando de iteración a iteración. En particular, el *XGBoost (Extreme Gradient Boosting)* es uno de los algoritmos predictivos más usados en la actualidad, debido a que obtiene buenos resultados de predicción con relativamente poco esfuerzo computacional [Mendoza, 2020]. Este tipo de modelos implica la estimación de una cantidad importante de hiperparámetros, cuyo valor óptimo se identifica mediante validación cruzada. Cabe mencionar que, al tratarse de un problema de clasificación, en este estudio se utilizó como función objetivo de las iteraciones la maximización del área bajo la Curva ROC.

[12] Los criterios de información se refieren al criterio de Akaike (corregido por los grados de libertad del modelo ajustado) y al criterio Bayesiano. Se conocen bajo las siglas AICc y BIC, respectivamente.

[13] Se utilizaron dos criterios: seleccionar el λ que minimiza la desviación promedio entre los datos observados y estimados, fuera de muestra *(min deviance)* o seleccionar el λ más grande que se encuentre a una distancia de una desviación estándar del *min deviance (1se)*. El criterio por *min deviance* está enfocado principalmente en desempeño predictivo, mientras que el criterio por *1se* busca balancear la predicción disminuyendo el riesgo de falsos descubrimientos.

6 Resultados

Las medidas de sensibilidad, especificidad y *accuracy* son algunas de las métricas que pueden obtenerse a partir de la matriz de confusión, la cual compara las predicciones de cada modelo con los datos observados, de acuerdo con un umbral de predicción determinado. Como se muestra en la Tabla 4 , la sensibilidad hace referencia a la proporción de empresas insolventes correctamente identificadas en el modelo (tasa de verdaderos positivos); la especificidad se refiere a la proporción de empresas no insolventes que fueron correctamente estimadas (tasa de verdaderos negativos); y el *accuracy* indica el total de empresas que fueron bien identificadas en el modelo.

		Clasificación estimada	
		NEGATIVOS	**POSITIVOS**
Clasificación Observada	**NEGATIVOS (N)**	Verdaderos Negativos (VN)	Falsos Positivos (FP)
	POSITIVOS (P)	Falsos Negativos (FN)	Verdaderos Positivos (VP)

Tabla 4: Esquema de la matriz de confusión

Fuente: Elaboración propia

$$sensibilidad = \frac{VP}{P} \tag{10}$$

$$especificidad = \frac{VN}{N} \tag{11}$$

$$accuracy = \frac{VP + VN}{P + N} \tag{12}$$

Por su parte, la Curva ROC es una representación gráfica que muestra las medidas de sensibilidad y especificidad y su variación frente al umbral de clasificación. En conjunto, estas métricas proporcionan una herramienta para seleccionar los modelos que tengan una mejor capacidad de pronóstico.

En línea con lo enunciado en secciones anteriores, a continuación se muestran los resultados obtenidos tras aplicar las metodologías ya explicadas, partiendo desde la capacidad de pronóstico resultante de aplicar los coeficientes del modelo Z-Score de Altman, y continuando con las distintas estimaciones que surgen tras ejecutar los modelos mediante las diferentes técnicas de remuestreo. Cabe resaltar que, como se mencionó previamente, el umbral de predicción a partir del cual las empresas se clasifican como insolventes, se tomó a partir de una probabilidad estimada superior al 0,5, en los modelos logísticos, de regularización y de árboles de decisión. No obstante, en esta sección también se realiza una comparación de las densidades de probabilidad, la cual muestra el grado de sensibilidad que tienen los modelos estimados ante diferentes umbrales de predicción.

6.1 Modelo Z-Score de Altman

Como se mostró en la Sección 5.1, el modelo de Altman tuvo diferentes modificaciones; en particular, considerando que los datos de las empresas colombianas corresponden en su gran mayoría a empresas no cotizantes en bolsa, el modelo utilizado corresponde al enunciado en

la Ecuación 5, tomando el equivalente a las variables calculadas y enunciadas en el Anexo A (ver Ecuación 13) [14].

$$Z - Score''_i = 0,717_i l_4 + 0,847_i r *_3 + 3,107_i r_6 + 0,420_i a_6 + 0,998_i o_1 \quad (13)$$

Aplicando el modelo a la submuestra de prueba, y de acuerdo con el criterio de decisión expresado en la 3, se obtiene la matriz de confusión y las métricas de predicción que se muestran en la Tabla 5.

		Datos estimados		
		No insolv	**Insolv**	**Observados**
Obs	**No insolv**	13.587	10.317	23.904
	Insolv	95	454	549
	Estimados	13.682	10.771	24.453

(a) Matriz de confusión

Métrica	**Tasa (%)**
Sensibilidad	82,70
Especificidad	56,84
Accuracy	57,42

(b) Métricas de predicción

Tabla 5: Desempeño de pronóstico modelo Altman Z-Score

Fuente: Elaboración propia

Este modelo pronostica 13.682 empresas como no insolventes y 10.771 empresas insolventes. Ahora, de las 23.904 empresas que fueron observadas como no insolventes, el modelo detecta correctamente 56,84% *(tasa de especificidad)*, y de las 549 empresas no insolventes, el modelo detecta 82,7% *(tasa de sensibilidad)*. Sin embargo, a pesar de la alta tasa de sensibilidad, en la predicción global el modelo hace una correcta clasificación para el 57,42% de las empresas. Este bajo *accuracy* puede generar altos costos en la implementación de políticas públicas pues, en este caso en particular, a partir de estas estimaciones se podrían tomar acciones de mitigación de riesgos en las más de 10.700 empresas estimadas como insolventes, y el diseñador de política no tendría cómo discernir o focalizar los recursos para detectar de forma más acertada la fracción de estas empresas que están en un riesgo real de ser insolventes (454 empresas). Lo anterior se refleja en que este modelo genera una tasa de falsos positivos del 43,2% [15].

Este desempeño del modelo puede deberse a que los coeficientes no fueron tomados con base en empresas colombianas, por lo que es muy probable que no capturen factores que en el contexto nacional estén siendo relevantes para determinar la probabilidad de insolvencia empresarial. Por otra parte, es importante mencionar que la estimación de este modelo no genera probabilidades, característica que dificulta la manipulación de los umbrales de clasificación, y por lo tanto, la posibilidad de generar criterios que permitan hacer una mayor focalización de políticas de mitigación, con una interpretación clara para los diseñadores de política pública. En parte, como se mencionó en la Sección 2, este factor fue lo que en la literatura motivó al desarrollo de modelos como los que se estiman a continuación.

[14] Cabe mencionar que el segundo factor del modelo, el cual originalmente corresponde al indicador de beneficios no distribuidos sobre activos totales fue reemplazado por el factor ROA (r3 en la fórmula). Lo anterior porque, como se explicó en la metodología, el indicador original, que correspondía a la variable r5 que no fue imputada por el algoritmo y por lo tanto fue excluida de la base. Al contrastar los datos originales sin imputar se encontró que las variables r5 y r3 tenían una correlación de 0,999 por lo cual se infiere que esta variable es una buena proxy para determinar el efecto que busca capturar el modelo.

[15] Calculado como 100%-tasa de especificidad

6.2 Comparación de modelos

Los primeros modelos que se estimaron a partir de los datos recopilados corresponden a modelos logísticos, siguiendo la Ecuación 7, y siendo X el conjunto de variables financieras y de contexto económico y empresarial mencionados en la Sección 4.

Si bien se utilizaron las diferentes técnicas de remuestreo para calibrar los modelos, con fines ilustrativos también se hizo la estimación utilizando la submuestra de entreno sin remuestreo. Los resultados de esta última estimación se observan en la Tabla 6 y reflejan que, efectivamente, al no haber un balance en la muestra el modelo pierde toda capacidad de pronóstico de insolvencia empresarial.

		Datos estimados		
		No insolv	**Insolv**	**Observados**
Obs	**No insolv**	23.904	-	23.904
	Insolv	549	-	549
	Estimados	24.453	-	24.453

(a) Matriz de confusión

Métrica	**Tasa (%)**
Sensibilidad	0,00
Especificidad	100,00
Accuracy	97,75

(b) Métricas de predicción

Tabla 6: Desempeño de pronóstico modelo Logístico – Datos sin hacer remuestreo

Fuente: Elaboración propia

Esto evidencia la importancia de tratar el problema de desbalance en las muestras para ejecutar modelos de pronóstico, así como la relevancia de analizar en conjunto diferentes métricas de predicción, pues como se observa en este caso, éstas pueden llegar a verse sesgadas por el comportamiento de los datos; a manera de ejemplo, en este caso en particular se obtuvo una especificidad del 100% y un *accuracy* del 97,75%, sin que esto implique que el modelo tenga un buen desempeño; en efecto, el modelo obtiene una sensibilidad del 0%, reflejando que no es capaz de identificar ninguna empresa insolvente. Precisamente este problema es el que se minimiza al utilizar métodos como el de remuestreo, explicado en la Sección 4.

Continuando con las estimaciones realizadas a partir de las diferentes técnicas de remuestreo, se presenta la Tabla 7, la cual resume las métricas de predicción para cada uno de los modelos estimados. Antes de entrar en el detalle de estos resultados, en primera instancia se observa que con todos los modelos existe una mejora sustancial de las métricas en relación con el modelo Z-Score, con tasas de *accuracy* que superan el 60%, las cuales, al mirarlas en conjunto con las otras métricas, nos dan indicios de las mejoras en los pronósticos frente a los modelos previamente descritos.

En cuanto a los modelos logísticos, cuyos coeficientes estimados están registrados en el Anexo D, se observa que si bien alcanzan una sensibilidad superior al 81%, la identificación correcta de empresas no insolventes es baja en relación con otros modelos. En particular, la estimación que logra un mejor balance entre sensibilidad y especificidad bajo la regresión logística se logra con la técnica de remuestreo de *undersampling*, la cual se optimiza con 18 variables.

En este caso, el modelo clasifica 9.128 empresas como insolventes y 15.325 como no insolventes. De las 24.453 empresas de la base, el modelo tuvo una tasa de *accuracy* del 63,68% (15.668 empresas), superando la tasa lograda por el modelo de Altman. La mejora se logra principalmente por una mejor detección de empresas no insolventes (reflejada en la tasa de especificidad), lo cual a su vez se traduce en una reducción en la tasa de falsos positivos (36,32%). De las 18 variables que optimizan el modelo, las que tienen mayor significancia estadística son el ROA, el EBITDA sobre Activos, y la solvencia, e identifica como relevante el sector económico al que pertenece la empresa (ver Anexo D).

Métrica	Under	Over	Both
Modelo: Logístico			
N. Coeficientes	18	18	19
Sensibilidad	81,24	81,42	81,24
Especificidad	63,68	62,37	62,07
Accuracy	64,07	62,80	62,50
Modelo: Lasso (selección AICc)			
N. Coeficientes	250	186	139
Sensibilidad	74,68	71,04	73,59
Especificidad	73,81	75,01	72,10
Accuracy	73,83	74,92	72,13
Modelo: Elastic Net (selección AICc)			
N. Coeficientes	252	180	128
Sensibilidad	74,86	70,86	73,41
Especificidad	74,08	75,12	72,56
Accuracy	74,10	75,03	72,58
Modelo: XGBoost			
Sensibilidad	83,61	15,48	29,87
Especificidad	79,16	99,62	98,85
Accuracy	79,26	97,73	97,30

Tabla 7: Resumen de resultados de pronóstico por modelo, para cada tipo de remuestreo

Fuente: Elaboración propia

En cuanto a los modelos de regularización vale la pena mencionar que, si bien éstos parten de funciones lineales sobre los predictores, también permiten indagar sobre la existencia de interacciones un poco más complejas entre estos. En este caso en particular, los modelos se estimaron considerando las interacciones simples entre todas las variables, lo que explica que cuenten con más de 100 predictores.

Como se mencionó en la Sección 5.3, existen varios criterios de selección de variables aplicables a los modelos de regularización, los cuales se obtienen de estimaciones por validación cruzada o directamente por los criterios de AICc y BIC. Los resultados de las métricas de predicción que se obtuvieron con los modelos resultantes de dichos criterios se detallan en el Anexo E, en donde se observa que los mejores balances entre sensibilidad y especificidad se obtienen al utilizar el criterio de Akaike [16].

Como se observa en la 7, la aplicación de estos métodos genera una disminución en la tasa de sensibilidad con respecto a los modelos logísticos, sin embargo, también generan una mejora de alrededor de 10 p.p. en la tasa de especificidad, lo que conlleva a mejores tasas de *accuracy* de los modelos. El mejor desempeño de los modelos de regularización evidencia que existen interacciones entre las variables que el modelo logístico no está capturando. Por otra parte, similar al caso logístico, el mejor balance entre sensibilidad y especificidad se obtiene con la técnica de *undersampling*, con un desempeño ligeramente superior en el modelo de *Elastic Net*.

Finalmente, los modelos de *XGBoost* se estimaron con base en el cálculo de unos hiperparámetros, cuya selección se realizó a partir de procesos de validación cruzada, utilizando como métrica de desempeño el área bajo la Curva ROC. Similar a los casos anteriores, las mejores métricas de pronóstico se obtienen con la técnica de *undersampling*, con la cual se obtiene una identificación correcta de 83,6% de las empresas insolventes y de 79,2% de empresas no insolventes, mejorando considerablemente la calidad de pronóstico de los modelos

[16] Corregido por los grados de libertad del modelo (AICc).

anteriores. Consecuentemente, la tasa de falsos positivos se reduce a más de la mitad que en el caso del modelo de Altman (20,84%), haciendo de este una mejor herramienta de focalización en términos de política pública. Este último aspecto puede inclusive fortalecerse, tomando umbrales de predicción más altos (por ejemplo 0,6), para detectar las empresas con probabilidades más altas de caer en estado de insolvencia, característica que no posee el modelo Z-Score.

El mejor desempeño del *XGBoost* puede verse gráficamente en la Figura 3, en donde se observa que la Curva ROC de este modelo se encuentra siempre por encima de las estimaciones realizadas con los otros modelos. Esto evidencia que las relaciones entre las variables que generan una mejor predicción de la insolvencia empresarial tienen niveles de complejidad y contemplan interacciones no lineales que no pueden ser capturadas por los modelos convencionales.

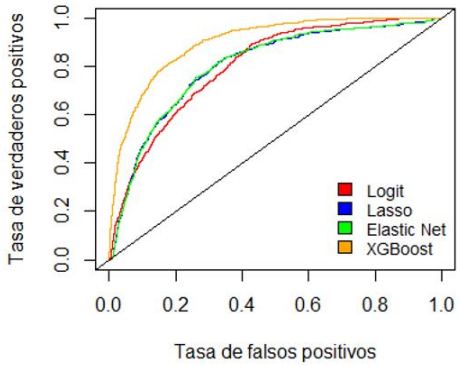


Fig. 3: Curva ROC - estimación sobre submuestra de prueba

Modelo	AUC
Logístico	0,8034
Lasso	0,8034
Elastic Net	0,8044
XGBoost	0,8138

Tabla 8: Área bajo la curva

Fuente: Elaboración propia

Ahora, si bien bajo este método no se obtiene una especificación de un modelo global que se pueda aplicar a todos los datos, si es posible determinar cuáles son los predictores más importantes. Estos se muestran en la Figura 4, donde se destacan las variables correspondientes al margen bruto ($r2$), al factor de endeudamiento ($a3$) y al autofinanciamiento ($a2$), seguidos por la solvencia de la empresa ($ot1$) y el margen neto ($r1$).

Vale la pena mencionar que, de acuerdo con este modelo, las variables adicionales de tamaño, sector, y las demás variables de contexto económico, no se encuentran entre las más importantes en la determinación de insolvencia empresarial en el país. No obstante, no se puede excluir la idea de que la estructura del claustro industrial y el tamaño de la empresa si pueden llegar a influir en la probabilidad de insolvencia. El análisis de estos aspectos podría profundizarse en futuras investigaciones, por ejemplo, mediante estudios que indaguen sobre relaciones causales entre las variables, más allá de la utilidad en la predicción en el fenómeno de estudio.

6.3 Robustez de los resultados: Comparación de densidades de probabilidad

Como se mencionó previamente, las matrices de confusión a partir de las cuales se calcularon las métricas de pronóstico de los diferentes modelos se generaron considerando un punto de

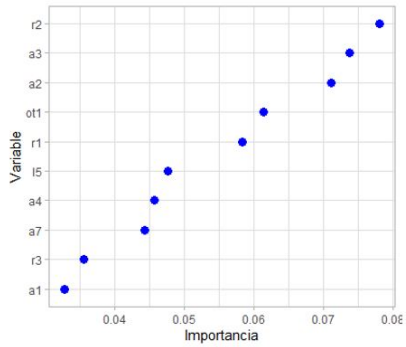


Fig. 4: Predictores más importantes *XGBoost-Undersampling*
Nota: De acuerdo con el Anexo A, las variables corresponden a: $r2$: margen bruto; $a3$: factor de endeudamiento; $a2$: autofinanciamiento; $ot1$: solvencia; $r1$: margen neto; $l5$: capital de trabajo neto sobre deuda de largo plazo; $a4$: nivel de cobertura I; $a7$: cobertura de costos de intereses; $r3$: ROA; $a1$: apalancamiento financiero.
Fuente: Elaboración propia

corte de 0,5, de acuerdo con el cual, si la probabilidad estimada superaba dicho valor, la empresa era clasificada como insolvente. No obstante, es pertinente analizar la sensibilidad del pronóstico ante diferentes puntos o umbrales de corte.

La Figura 5 muestra las densidades de probabilidad de los diferentes modelos estimados, bajo la técnica de remuestreo de *undersampling*. En esta, puede observarse que el modelo logístico tiene, contrario a los otros, una acumulación de empresas con probabilidades asignadas que van entre 0,4 y 0,6, lo que implica que el modelo logístico tiene una mayor sensibilidad en el punto de corte asignado. Así, por ejemplo, si el punto de clasificación de insolvencia se moviera ligeramente hacia la izquierda, bajo este modelo habría una mayor cantidad de nuevas empresas que caerían en dicha clasificación, en relación de las nuevas que serían asignadas bajo otros modelos.

Por el contrario, los modelos de regularización y, en mayor medida el modelo por *XGBoost*, tienen una mayor acumulación de empresas hacia los extremos, lo que evidencia que este modelo predice probabilidades altas tanto de insolvencia como de no insolvencia. Estos resultados muestran que, ante variaciones del punto de corte de clasificación, el modelo por *XGBoost* tiene una mayor robustez que los otros modelos evaluados, y que puede ser de mayor utilidad si se manipula el umbral de corte, por ejemplo, para implementar acciones más focalizadas hacia empresas con mayores probabilidades de entrar en bancarrota.

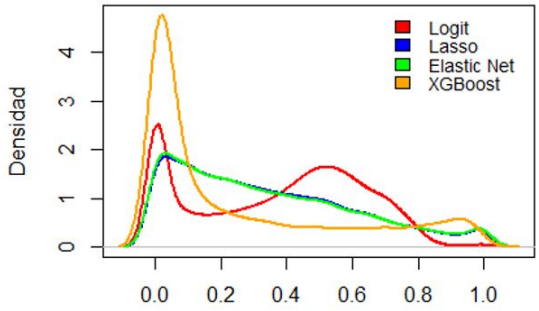


Fig. 5: Densidad de las probabilidades estimadas
Fuente: Elaboración propia

7 Herramientas de Política Pública

7.1 Estimación con menos predictores

Teniendo en cuenta el costo potencial en el que puede llegar a incurrir el diseñador de política pública en la estimación de las 38 variables consideradas en el modelo para la estimación del *XGBoost*, es pertinente analizar el cambio en el desempeño del pronóstico si se utilizan únicamente las variables que tienen mayor importancia en la identificación de la insolvencia empresarial.

De esta manera, la Tabla 9 muestra las métricas de predicción obtenidas tras utilizar todas las variables, y las compara con las métricas obtenidas tras utilizar los diez y los cinco predictores más importantes, identificados previamente en la Figura. A su vez, estas métricas son comparadas con las obtenidas bajo el modelo logístico (obtenido con la técnica de remuestreo de *undersamplig*) con el fin de determinar si existe una verdadera ganancia al utilizar estos modelos frente a utilizar el modelo más simple de pronóstico analizado en este documento, el cual, vale la pena mencionar, se estimó considerando 18 predictores.

Modelo	**Variables**	**Sensibilidad**	**Especificidad**	**Accuracy**
XGBoost (Undersampling)	Todas (38)	83,61	79,16	79,26
	10 Principales	81,97	77,20	77,31
	5 Principales	80,69	73,45	73,61
Logit (Undersampling)	18	81,24	63,68	64,07

Tabla 9: Métricas de predicción del *XGBoost* con las principales variables, y del modelo logístico

Fuente: Elaboración propia

En primer lugar, se observa que inclusive al hacer la estimación únicamente con las cinco variables principales, el modelo genera buenos desempeños de pronóstico, haciendo una correcta identificación, entre empresas insolventes y no insolventes, superior al 73%. Sin embargo, la ganancia en términos de pronóstico se obtiene principalmente por la correcta identificación de empresas no insolventes. Si el objetivo principal del diseñador de políticas públicas se centra en la identificación de empresas insolventes, puede ser más conveniente hacer la estimación mediante el modelo logístico que utilizar solamente las cinco variables principales, mediante el *XGBoost*.

Sin embargo, si el diseñador de política pública puede estimar las diez principales variables, el modelo de *XGBoost* va a tener un mejor desempeño tanto en la identificación de empresas insolventes como no insolventes, lo que, como se ha mencionado previamente, puede disminuir la tasa de empresas falsamente identificadas como insolventes, generando una mayor eficiencia en la implementación de posibles acciones de mitigación.

En conclusión, estos resultados evidencian que la estimación mediante la técnica de *XGBoost* genera buenos resultados de pronóstico, incluso si se toman sólo las variables principales. Con el fin de minimizar la pérdida en el desempeño de pronóstico, se sugiere que al menos se consideren los diez predictores más importantes, lo cual en todo caso genera menores costos de estimación frente al modelo logístico propuesto, el cual considera 18 variables en su estimación.

Por otra parte, este resultado también refleja lo ya enunciado previamente, y es que existen relaciones no lineales entre las variables que están determinando la probabilidad de que una empresa entre en estado de insolvencia, y que dichas relaciones no lineales no están siendo capturadas por el modelo logístico.

7.2 Capacidad de pronóstico por sector y tamaño de empresa

De acuerdo con lo evidenciado en las Secciones 3 y 4, la heterogeneidad de las empresas colombianas, al considerar su tamaño y el sector económico al que pertenecen, hacen que las empresas tengan un diferente grado de afectación frente al fenómeno de insolvencia. En ese sentido, es pertinente analizar el desempeño del modelo de pronóstico estimado mediante la metodología *XGBoost* para estas características.

Por una parte, la Tabla 10 muestra las métricas de predicción obtenidas al aplicar el modelo en la submuestra de prueba, para cada uno de los tamaños de empresa. Se observa que la mayor tasa de identificación correcta de empresas insolventes se obtiene en las empresas medianas (85,6%), mientras que la mayor tasa de identificación correcta de empresas no insolventes se da en las pequeñas empresas (81,8%). A pesar de que el menor desempeño en pronóstico se obtiene para las grandes empresas, en general puede afirmarse que el modelo genera un buen desempeño de pronóstico para los tres tamaños, con tasas que superan el 76% de predicción correcta.

Tamaño	**Sensibilidad**	**Especificidad**	**Accuracy**
Pequeña	81,25	81,77	81,76
Mediana	85,60	78,34	78,51
Grande	83,18	76,16	76,31
Total	**83,61**	**79,16**	**79,26**

Tabla 10: Métricas de predicción por tamaño de empresa

Nota: A partir del remuestreo de la base, las submuestra de entrenamiento y de prueba también estuvieron balanceadas en cuanto a tamaño empresarial. La proporción de empresas de ambas bases corresponde aproximadamente a: 20% grandes empresas; 43% empresas medianas; y 37% pequeñas empresas.
Fuente: Elaboración propia

De forma similar, se puede observar que, al analizar los pronósticos por sector económico, la tasa de predicción correcta de empresas insolventes y no insolventes supera siempre el 65% (ver Tabla 11). Por una parte, el desempeño más alto en la identificación correcta de empresas insolventes se obtiene para las empresas del sector de manufactura (90,9%) y el más bajo para las empresas del sector de comercio (79,2%). Por otra parte, el mejor desempeño en la identificación de empresas no insolventes se obtiene para el sector de servicios (90%), mientras que el menor desempeño se obtiene para el sector de construcción (65,2%).

8 Conclusiones

Este documento realiza una comparación de calidad de pronóstico de insolvencia empresarial, partiendo desde la aplicación de los coeficientes del modelo Z-Score de Altman, pasando por estimaciones de modelos logísticos, de regularización, y de *boosting* aplicado a árboles de decisión. Con base en las estimaciones puede concluirse, en primer lugar, que si bien el modelo de Altman aplicado a las empresas colombianas genera una alta tasa de sensibilidad, la predicción global del modelo, medida por medio del *accuracy*, no es óptima, pudiendo generar altos costos en la implementación de políticas públicas derivados de la alta tasa de falsos positivos que se genera en su estimación. Debido a que este modelo no genera probabilidades de pronóstico, flexibilizar los umbrales de clasificación se vuelve difícil, al no existir una interpretación clara de los mismos, lo que dificulta la posibilidad de una mayor focalización de políticas de mitigación que se pretendan aplicar a partir de la estimación de este modelo.

Sector económico	Sensibilidad	Especificidad	Accuracy
Servicios	79,73	89,96	89,88
Comercio	76,19	79,02	78,95
Manufactura	90,86	67,33	68,45
Construcción	87,84	65,29	65,86
Agropecuario	77,08	69,15	69,40
Minero-hidrocarburos	85,00	76,30	76,62
Total	**83,61**	**79,16**	**79,26**

Tabla 11: Métricas de predicción por sector económico

A partir del remuestreo de la base, las submuestra de entrenamiento y de prueba también estuvieron balanceadas en cuanto a sector económico. La proporción de empresas de ambas bases corresponde aproximadamente a: 32% servicios; 26% comercio; 20% manufactura; 13% construcción; 7% agropecuario; y 2% minero-hidrocarburos.
Fuente: Elaboración propia

Por otra parte, se evidenció la necesidad de aplicar técnicas de remuestreo para balancear la base de datos. Lo anterior dado que los modelos de pronóstico y clasificación son altamente sensibles ante la presencia de datos desbalanceados, generando pronósticos sesgados. En este caso en particular, la técnica de remuestreo por *undersampling* es la que genera los mejores desempeños de pronóstico de insolvencia empresarial, al aplicar diferentes metodologías de estimación y generar los respectivos pronósticos sobre la submuestra de prueba.

Entre los modelos estimados, el método de *XGBoost* mejora considerablemente los pronósticos en relación con los modelos que basan sus estimaciones en funciones lineales sobre los predictores. Esto implica que, en Colombia, en el problema de determinación de insolvencia empresarial, las variables presentan relaciones complejas e interacciones no lineales que no son capaces de identificar los modelos convencionales, lo que da una ventaja a las predicciones realizadas mediante las técnicas de *machine learning*.

El buen desempeño de esta metodología se mantiene tanto si se analiza por tamaño como por sector económico, mostrando una mayor tasa de detección de empresas insolventes en pequeñas empresas y en empresas pertenecientes al sector de manufacturas.

Por otra parte, teniendo en consideración el costo potencial de estimación de todas las variables estimadas en el desarrollo del modelo de *XGBoost*, puede afirmarse que aun realizando la estimación con los diez predictores principales, este modelo genera ganancias en términos de desempeño de pronóstico al compararlo con la estimación por un modelo logístico.

Finalmente, vale la pena mencionar que, de acuerdo con el modelo de *XGBoost*, las variables de tamaño empresarial, sector económico, y las variables adicionales de contexto económico, no se encuentran entre las más importantes en la determinación de insolvencia empresarial en el país. Las variables que tienen una mayor incidencia en la explicación de este fenómeno corresponden a las variables financieras de margen bruto, factor de endeudamiento, autofinanciamiento, solvencia y margen neto. No obstante, no se puede descartar la idea de que las variables que dan cuenta de la estructura de la empresa puedan influir, por ejemplo de forma causal, en el fenómeno de insolvencia, tema que se puede profundizar en estudios posteriores.

Referencias

Alareeni, B. and Branson, J. (2013). Predicting listed companies' failure in jordan using altman models: A case study. *International Journal of Business and Management*, 8(1):113–126.

Altman, E. (1968). Financial ratios, discriminant analysis and the prediction of corporate failure. *Journal of Finance*, 23(4):589–609.

Altman, E. (2000). Predicting financial distress of companies: revisting the z-score and zeta models. *Handbook of Research Methods and Applications in Empirical Finance*, 5.

Amat, J. (2017). Árboles de decisión, random forest, gradient boosting y c5.0. *Obtenido de RPubs by RStudio*.

Awati, K. (2017). A gentle introduction to logistic regression and lasso regularisation using r. Technical report, Eight to Late.

Barboza, F., Kimura, H., and Altman, E. (2017). Machine learning models and bankruptcy prediction. *Expert Systems With Applications*, 83:405–417.

Charfaoui, Y. (2019). Resampling to properly handle imbalanced datasets in machine learning. Technical report, Heartbeat.

Confecámaras (2021). Dinámica de creación de empresas en colombia, enero - junio de 2021. Technical report, Análisis Económico.

Congreso de la República (2006). Ley 1116 de 2006. por la cual se establece el régimen de insolvencia empresarial en la república de colombia y se dictan otras disposiciones. Technical report.

Gerantonis, N., Vergos, K., and Christopoulos (2009). Can altman z-score models predict business failures in greece? *Research Journal of International Studies*, pages 21–28.

Hamid, T., Akter, F., and Binte Rab, N. (2016). Prediction of financial distress of non-bank financial institutions of bangladesh using altman's z score model. *International Journal of Business and Management*, 11(12):261–270.

Ibánez, D. (2020). Actualización impacto de la coyuntura del coronavirus en la economía colombiana. Technical report, Colombia. Superintendencia de Sociedades.

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). *An introduction to statistical learning*, volume 112. Springer, New York.

Mendoza, J. (2020). Tutorial: Xgboost en python. Technical report, Medium.

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

Song, X. (2018). Imbalanced classification on bankruptcy prediction. Technical report, RPubs by Rstudio:.

Sánchez Jiménez, V. (2015). La redefinición del papel de la empresa en la sociedad. *BARATARIA: Revista Castellano-Manchega de Ciencias Sociales*, 20:129–145.

Anexos

A Variables financieras

Grupo	Variable	Nombre variable
Liquidez	Razón corriente: Activo corriente/Pasivo corriente	L1
	Prueba ácida: (Activo corriente-Inventarios ctes y nctes)/Pasivo corriente	L2
	Ratio de estabilidad financiera: PPE/(Pasivo no corriente + Patrimonio)	L3
	Capital de trabajo neto sobre activos totales: Capital circulante/Activos	L4
	Capital de trabajo neto sobre deuda a largo plazo: Capital circulante/Pasivo no corriente	L5
Apalancamiento	Apalancamiento financiero: Pasivo/Patrimonio	A1
	Autofinanciamiento: Patrimonio/Activos x100	A2
	Factor de endeudamiento: Pasivo total / (Ingresos por actividades ordinarias + Ajustes por gastos de depreciación y amortización)	A3
	Nivel de cobertura I: Patrimonio/PPE	A4
	Nivel de cobertura II: (Patrimonio+Pasivo no corriente)/PPE	A5
	Patrimonio/Pasivo	A6
	Cobertura de costos de intereses: Ganancias por actividades de operación / Costos financieros	A7
Operacionales	Rotación de activos 1: Ingresos por actividades ordinarias / Activo total promedio*	O1
	Rotación de activos 2: Ingresos por actividades ordinarias/ Activo no corriente promedio*	O2
	Rotación de capital de trabajo: Ingreso por actividades ordinarias/capital circulante	O3
	Rotación de inventario: Ingreso por actividades ordinairas/Inventario promedio* (cte y no cte)	O4
Efectividad	Efectividad ventas: Ingresos de actividades ordinarias / Costo de ventas	E1
	Efectividad financiamiento: Ingresos financieros / Costos financieros	E2
	Efectividad operativa: (Ingresos de actividades ordinarias + Otros ingresos + Participación en las ganancias de asociados y negocios conjuntos que se contabilicen utilizando el método de participación) / (Gastos de ventas+ Gastos de administración + Otros gastos)	E3
Rentabilidad	Márgen neto: (Ganancia neta/Ingresos por act ordinarias)x100	R1
	Márgen bruto: (Ganancia bruta/Ingresos por act ordinarias)x100	R2
	Retorno sobre los activos (ROA): Ganancia antes de impuestos/Activos	R3
	Retorno sobre el patrimonio (ROE): Ganancia antes de impuestos / Patrimonio	R4
	Rentabilidad acumulada año anterior: Beneficios no distribuidos/Activos totales	R5
	EBITDA/Activos	R6
	EBITDA/Ingresos por act ordinarias	R7
Otros ratios	Solvencia: Activo/Pasivo	OT1

Aclaraciones:

PPE=Propiedad, Planta y Equipo
Capital circulante=Activos corrientes-Pasivos corrientes
Ingresos por actividades ordinarias: Proxy de ventas
Beneficios no distribuidos: ganancia(perdida) año anterior - dividendos pagados año actual
EBITDA: ganancia (pérdida) por actividades de operación+ajustes por gastos de depreciación y amortización+ajustes por provisiones

Fuente: Elaboración propia

B Variables de contexto económico

Variable	Descripción	Fuente
VAGcorr_Part	Participación del valor agregado del macro sector dentro del valor agregado departamental. Calculado a partir del valor agregado corriente.	DANE, Cuentas nacionales
VAG_Var	Variación del valor agregado del macrosector en el departamento. Calculado a partir del valor agregado departamental, serie encadenada de volúmen, con año de referencia 2015.	DANE, Cuentas nacionales
Ocup_Part	Participación del número de ocupados por rama, por departamento. El DANE sólo toma información para 23 departamentos y Bogotá, por lo que el valor para el resto (Arauca, Casanare, Putumayo, San Andrés, Amazonas, Guainía, Guaviare y Vichada) se imputó, tomando como referencia por una parte el total nacional de ocupados, y por otro, la participación por ramas del valor agregado en valores constantes.	DANE, Gran Encuesta Integrada de Hogares
TD	Tasa de desempleo departamental. El DANE sólo toma información para 23 departamentos, por lo que el valor para el resto (Arauca, Casanare, Putumayo, San Andrés, Amazonas, Guainía, Guaviare y Vichada) se imputó con base en el Boletín para ciudades de la Amazonía, tomando como referencia la TD promedio reportada entre 2016-2018, y la TD promedio nacional. En cuanto a San Andrés, se tomó como referencia el reporte del DANE para la Gobernación del departamento, TD promedio 2018-2019.	DANE, ECH - GEIH
EX_part	Participación de cada una de las ramas, dentro de las exportaciones del departamento.	DANE, Comercio Internacional
EX_var	Variación de las exportaciones por rama, dentro de cada departamento.	DANE, Comercio Internacional
IM_part	Participación de cada una de las ramas, dentro de las importaciones del departamento (CIF).	DANE, Comercio Internacional
IM_var	Variación de las importaciones por rama, dentro de cada departamento (CIF).	DANE, Comercio Internacional
EX_IM	Razón entre exportaciones FOB e importaciones FOB. Si Importaciones =0 se imputó un valor de 0,01 para que el valor no se indeterminara pero que reflejara la relación comercial.	DANE, Comercio Internacional

Fuente: Elaboración propia

C Estadísticas descriptivas - datos imputados

Var resumida	Descripción	No insolventes				Insolventes			
		Faltantes	Media	Mediana	Desv. Est	Faltantes	Media	Mediana	Desv. Est
Liquidez1	Razón corriente	0	4760	1,71	183000	0	5,23	1,52	60,4
Liquidez2	Prueba ácida	537	366	1,25	45400	0	2,21	0,888	8,65
Liquidez3	Estabilidad financiera	0	0,536	0,322	12,3	0	0,306	0,459	18
Liquidez4	Ktrabajoneto/activos	0	0,144	0,175	5,32	0	0,129	0,151	0,45
Liquidez5	Ktrabajoneto/deuda largo plazo	0	1810	0,958	313000	0	-10,5	0,393	344
Apalancamiento1	Apalancamiento financiero	0	7,2	0,872	453	0	5,21	1,96	229
Apalancamiento2	Autofinanciamiento	0	-11,5	50,2	12000	0	13,4	23,4	81,4
Apalancamiento3	Factor de endeudamiento	0	5210	0,585	585000	0	3,39	1,12	12
Apalancamiento4	Nivel de cobertura I	0	359	2,07	25100	0	-1,73	0,75	439
Apalancamiento5	Nivel de cobertura II	0	571	2,95	56900	0	32,6	1,98	360
Apalancamiento6	Patrimonio/Pasivo	103	1840	1	191000	0	0,465	0,305	0,937
Apalancamiento7	Cobertura de costos de intereses	0	1980	3,64	1620000	0	36,6	1,15	725
Operacionales1	Rotación de activos I	0	1,19	0,673	11,2	0	0,904	0,581	5,68
Operacionales2	Rotación de activos II	0	788	2,01	74500	0	68,1	1,52	2440
Operacionales3	Rotación de k de trabajo	0	20,3	2,05	6480	0	35,8	1,49	1290
Operacionales4	Rotación de inventario	0	6970	6,66	1050000	0	39,5	5,16	331
Rentabilidad1	Márgen neto	0	-414000	3,53	110000000	0	-15,4	-0,653	891
Rentabilidad2	Márgen bruto	0	-1040	34,5	217000	0	23,9	24	40,8
Rentabilidad3	ROA	0	2,55	0,0354	799	0	-0,051	0,00025	0,254
Rentabilidad4	ROE	0	-0,0393	0,0891	33,8	0	1,18	0,0354	59
Rentabilidad5	EBITDA/Activos	512	2,85	0,0632	801	0	0,0141	0,0332	0,234
Rentabilidad6	EBITDA/Ingresos por act ordinarios	0	2700	0,0978	2110000	0	0,0825	0,0527	8,99
Otros ratios	Solvencia	0	27800	2,01	923000	0	1,47	1,31	0,937
Efectividad1	E. ventas	0	865	1,37	6090	0	29	1,32	1030
Efectividad2	E. Financiamiento	0	11,6	0,129	191	0	1,46	0,106	13,7
Efectividad3	E. Operativa	0	14	2,76	303	0	16,2	3,08	339
$Vagg_part$	Valor agregado (Part)	0	27,7	19,8	20,5	0	17,8	15,1	15,3
$Vagg_var$	Valor agregado (var)	0	-0,987	1,69	7,83	0	-1,83	1,08	7,95
TD	Tada de desempleo dpto	0	12,1	10,9	3,47	0	11,8	10,8	3,46
EX_part	Exportaciones (part)	0	13,1	0,305	28,1	0	26,7	0,886	36,7
IM_part	Importaciones (part)	0	14,6	0,192	33,1	0	31	0,259	43,1
EX_IM	Ratio Exp/Imp	0	4360000	0,496	758000000	0	22200	0,496	147000
EX_var	Exportaciones (var)	0	454	0	9700	0	244	0	2210
IM_var	Importaciones (var)	0	2890	0	17400	0	2000	0	14600
$Ocup_Part$	Ocupados sector/dpto	8	29,1	19,9	20,2	0	19	14,5	15,3

Fuente: Elaboración propia

D Modelos logísticos

	Under	**Over**	**Both**
(Intercept)	2.156 ***	2.900 ***	2.870 ***
	(0.237)	(0.047)	(0.066)
L2: Prueba ácida	-0,004		0,000
	(0.003)		(0.000)
L4: K de trabajo neto sobre activos totales	0,179	0.115 ***	0.152 ***
	(0.127)	(0.015)	(0.022)
L5: K de trabajo neto sobre deuda largo plazo	-0.003 **	-0.000 ***	-0.000 ***
	(0.001)	(0.000)	(0.000)
A3: Factor de endeudamiento	-0.007 **	-0.002 ***	-0.002 ***
	(0.002)	(0.000)	(0.000)
O1: Rotación de activos 1	-0,017	-0.012 ***	-0.011 ***
	(0.016)	(0.001)	(0.002)
R2: Márgen bruto	-0,001		0,000
	(0.001)		(0.000)
R3: Retorno sobre Activos (ROA)	-2.515 ***		-0.024 ***
	(0.700)		(0.006)
R6: EBITDA/Activos	2.512 ***	-0.037 **	
	(0.699)	(0.014)	
OT1: Solvencia	-0.809 ***	-0.878 ***	-0.861 ***
	(0.063)	(0.010)	(0.013)
E1: Efectividad ventas	-0.000 *	-0.000 ***	-0.000 ***
	(0.000)	(0.000)	(0.000)
E2: Efectividad financiamiento	-0.007 *	-0.005 ***	-0.005 ***
	(0.003)	(0.000)	(0.001)
R4: Retorno sobre el patrimonio (ROE)		0.001 ***	0.001 **
		(0.000)	(0.000)
Tasa de Desempleo departamental		-0.024 ***	-0.023 ***
		(0.002)	(0.003)
Exportaciones (Part)		-0.005 ***	-0.005 ***
		(0.001)	(0.001)
Ocupados sector a nivel departamental		-0.013 ***	-0.014 ***
		(0.001)	(0.002)
R7: EBITDA/Ingresos por act ordinarias			0.001 ***
			(0.000)
Variables Tamaño	Si	Si, signif.	Si, signif.
Variables Macrosector	Si, signif.	Si, signif.	Si, signif.
N	2632	111623	57056
logLik	-1424,566	-61064,841	-31232,689
AIC	2893,132	122189,683	62533,379

Fuente: Elaboración propia

E Métricas de pronóstico de los modelos de regularización

Modelos lasso

Undersampling

	min deviance	1se deviance	min AICc	BIC
N. Coeficientes	6	5	250	29
Sensibilidad	85,43	86,34	74,68	72,50
Especificidad	41,68	39,29	73,81	67,96
Accuracy	42,66	40,35	73,83	68,07

Oversampling

	min deviance	1se deviance	min AICc	BIC
N. Coeficientes	13	13	186	186
Sensibilidad	74,50	74,50	71,04	71,04
Especificidad	56,67	56,67	75,01	75,01
Accuracy	57,07	57,07	74,92	74,92

Both

	min deviance	1se deviance	min AICc	BIC
N. Coeficientes	123	51	139	139
Sensibilidad	72,86	70,31	73,59	73,59
Especificidad	72,15	71,25	72,10	72,10
Accuracy	72,17	71,23	72,13	72,13

Modelos Elastic Net

Undersampling

	min deviance	1se deviance	min AICc	BIC
N. Coeficientes	5	4	252	29
Sensibilidad	86,52	86,52	74,86	73,95
Especificidad	38,68	38,65	74,08	68,54
Accuracy	39,76	39,73	74,10	68,66

Oversampling

	min deviance	1se deviance	min AICc	BIC
N. Coeficientes	11	11	180	180
Sensibilidad	74,32	74,32	70,86	70,86
Especificidad	59,56	59,56	75,12	75,12
Accuracy	59,89	59,89	75,03	75,03

Both

	min deviance	1se deviance	min AICc	BIC
N. Coeficientes	117	37	128	128
Sensibilidad	73,04	68,49	73,41	73,41
Especificidad	72,58	74,89	72,56	72,56
Accuracy	72,59	74,74	72,58	72,58

Fuente: Elaboración propia