Modelos Predictivos de Defaults con Machine Learning: Una Comparación Basada en Métricas de Desempeño

Departamento de Economía

Licenciatura en Economía

Autoras: Valentina Lucini y María Mercedes García Fagalde

Mentores: Juan Cruz López Del Valle y Belén Michel Torino

Victoria, Provincia de Buenos Aires, Julio de 2022

Resumen

El siguiente trabajo propone una actividad de predicción de defaults de una selección de cinco países quienes resultaron ser aquellos con mayor cantidad de incumplimientos de deuda soberana desde 1970 hasta la actualidad. Para ello, se utilizarán herramientas y modelos de predicción de Machine Learning. La propuesta consta de una comparación o "carrera de caballos", de diversos modelos (entre ellos regresión logit, KNN y árboles de regresión), y a través del cálculo de diversas métricas de desempeño poder seleccionar el que mejor predice los defaults de los países seleccionados dentro del conjunto de datos de entrenamiento.. (...)

Índice

1. Introducción

Son innumerables y obvias las razones que justifican la profunda búsqueda de la existencia de modelos predictivos sobre deuda soberana. Para quienes escribimos este trabajo el interés por este estudio nace por la simple razón de que nos encontramos en un estado de necesidad para poder entender e intentar prevenir situaciones económicas recurrentes en el país en el que vivimos. Por esto es que, para los argentinos es un tema recurrente pero no menor. Pasados los doscientos años del

primer default de nuestro país, la historia pareciera no tener fin. No son una ni dos las ocasiones donde el gobierno no pudo pagar su deuda externa, sino que nos encontramos frente a un escenario conocido, repetitivo contando nueve defaults oficiales a lo largo de la historia. En un país como Argentina, ya es algo habitual para sus ciudadanos, donde no solo sufren las consecuencias económicas que este hecho conlleva sino también el continuo desánimo de vivir en un país que no progresa. El último préstamo solicitado al Fondo Monetario Internacional fue en 2018, durante la gestión del Presidente Mauricio Macri donde se acordó un monto de 57 millones de dólares (si bien terminó siendo cerca de 44 millones debido al triunfo de Alberto Fernandez en las elecciones Primarias Abiertas, Simultáneas y Obligatorias y el posterior salto del dólar), el más alto de los préstamos acordados entre el FMI y el Estado argentino. Actualmente, como era de esperar, el gobierno argentino negoció con el fondo para reestructurar su deuda, (...)

Esta misma situación se vivió en otros países en simultáneo, siendo España el país que más defaults tuvo en la historia (14), seguido por Venezuela (12) y Ecuador (11). Si bien nuestro objetivo principal era realizar el trabajo para Argentina, nos encontramos con dificultades al predecir con pocas variables y pocos datos disponibles. Es por ello que decidimos extender el análisis para aquellos países que más defaults tuvieron desde 1970 (datos disponibles). Además de Argentina, se encuentran Ecuador con 5 defaults, Paraguay con 2 , Bolivia 3 y por último Nigeria con 5.

Se entiende por default soberano al incumplimiento de pago por parte de un Estado al pagar su deuda en su totalidad. Diversas son las consecuencias que recaen en un país al dejar en deuda a sus acreedores, siendo la más inmediata el impedimento del acceso a los mercados internacionales de crédito. Los períodos de reestructuración de la deuda suelen ser largos, y mientras no se llega a un acuerdo entre el gobierno y sus acreedores, el país pierde el acceso a los mercados internacionales; como es el ejemplo de Argentina, que luego de su default en 2002 estuvo catorce años sin acceso al financiamiento externo. No solo el gobierno se ve perjudicado sino también los municipios y empresas, que no son indemnes de las consecuencias que supone el no acceso al crédito; si los acreedores no prestan al gobierno, tampoco lo hará a las empresas que operan en el país. Ante esta situación el gobierno se ve obligado a elegir una de las siguientes alternativas: recortar el gasto público, aumentar los impuestos o financiarse a través de la emisión de pesos, todas con fuertes consecuencias en la economía. Por otro lado, la confianza de los agentes en una economía es vital para que esta funcione adecuadamente y luego de una cesación de pagos el aumento de la incertidumbre en la economía es un problema no menor. El ambiente de incertidumbre que se genera luego que un país entre en default, hace que los consumidores e inversores restrinjan sus ganas de consumir e invertir, afectando la actividad económica y consecuentemente generando presión en el mercado cambiario.

Antes del siglo XIX los default eran resultado de eventos extraordinarios pero a partir de entonces se han vinculado en mayor medida a desmanejos financieros. Las décadas del 70 y 80 fueron protagonistas de tres grandes crisis; crisis mundial del petróleo, crisis del dólar y la de deuda externa latinoamericana. Las dos primeras crearon una elevada inflación en los países en desarrollo mientras que la última dejó a los países latinoamericanos en una posición difícil para hacer frente a los retos que impondría la economía mundial a fines del siglo 20. Agosto de 1982 marca convencionalmente el inicio de la denominada "década perdida" para el desarrollo latinoamericano, pero esta crisis reconoce sus raíces en sus desequilibrios macroeconómicos y choques externos que se desarrollaron en la década previa, en casi todos los casos el gasto total se expandió por encima del producto. Los países estudiados son los que incumplieron su deuda soberana más veces desde la década del 70 hasta la actualidad.

Por un lado, los países latinoamericanos estudiados comparten características similares en cuanto a las crisis de deuda que enfrentaron a comienzos de la década de los 80, pues ninguno queda exento de la "década perdida" de la región. Tanto Argentina como Paraguay, Ecuador y Bolivia se enfrentaron a gobiernos cívico militares en la década del 70, acompañado con alza de precios, déficit en sus balanzas de pagos y consecuentemente creciente deuda externa para financiarlos. Para el caso argentino, y para el período estudiado, el país sufrió cinco de sus nueve defaults; en 1982 declaró su quinto default, seguido por aquel del 1989, luego tras la crisis de convertibilidad en 2001 se declaró el séptimo y en el 2014 luego de dos canjes (en 2005 y 2010) con los bonistas afectados por el default del 2001 el país entró en default por no aceptar pagarle a los fondos litigantes. Por último en 2020, el gobierno de Alberto Fernandez, entró en lo que llaman "default selectivo" por no pagar ni cerrar un acuerdo de reestructuración o canje de deuda con sus bonistas. Se lo llama así dado que una parte de la deuda externa del gobierno, esos títulos que quedaron impagos, entraron en cesación de pagos, pero no el resto de su deuda soberana (que en total supera los 320 millones de dólares). Por otro lado, Bolivia vivió tres defaults durante el período estudiado, todos durante la década del 80; en 1980, 1986 y 1989. Para mayor visibilidad, la deuda soberana en mora aumentó considerablemente durante este período, de USD 6,6 millones en 1978 a USD 2.236 millones en 1986. Al igual que los países mencionados anteriormente, ni Ecuador ni Paraguay presentaron una situación favorable para el período estudiado. Ecuador incurrió en cinco defaults; dos de ellos en la década del 80 tras los desequilibrios económicos que el país enfrentó a raíz de las deudas desmesuradas que el gobierno adquirió durante la crisis del petróleo; y las tres consiguientes en 1999, 2008 y finalmente en 2020. Al igual que Bolivia, Paraguay incurre en incumplimiento de su deuda externa en 1986 tras el escenario latinoamericano y nuevamente en el 2003. Si bien Nigeria no forma parte de la crisis de deuda latinoamericana, no se escapa de las consecuencias de la crisis mundial del petróleo y del dólar de esos tiempos. Para el período en cuestión, dejó de pagar sus deudas en cinco ocasiones; 1982, 1986, 1992, 2001 y 2004. (...)

Como se mencionó anteriormente, no son pocas las consecuencias que recaen sobre la economía luego de un default. En este sentido, es importante poder entender no solo cuales son sus determinantes sino también poder predecirlo. Proponemos utilizar las técnicas de Machine Learning para predecir la probabilidad de que un Estado soberano, incumpla con sus préstamos. La explosión de nuevas técnicas de aprendizaje automático (ML) cumplen un papel importante y pueden utilizarse para gestionar complejos modelos económicos equipados con enorme cantidad de datos. Por otro lado, ML representa una técnica clave que ajusta formas funcionales complejas y muy flexibles a los datos sin sobre ajustar, es decir, encuentra funciones que predicen bien fuera de la muestra.

En este trabajo se propone analizar una suerte de carrera de caballos entre modelos predictivos (CART, SVM, Random Forest, LASSO, XG Boosting). A partir de estos, se intentará evaluar cuál de ellos ha tenido un mejor desempeño respecto de las métricas y así poder predecir un default. Para ello se propone analizar cuales son los determinantes de defaults soberanos según la literatura previa, y más aún cuales son significativos para predecir. En cuánto los modelos hayan sido evaluados, comparándolos e identificado distintos parámetros relevantes en cada uno, se llegará a la conclusión de cuál es el modelo que mejor se adecúa a los eventos ya acontecidos. La elección del modelo se hará mediante la técnica de validación cruzada (CV) que tiene en cuenta el poder predictivo, más que la estimación de un parámetro estructural o causal concreto, y además el método utiliza comparaciones fuera de la muestra, en lugar de medidas de bondad de ajuste dentro de la muestra. Esto garantiza que obtenemos comparaciones no sesgadas del ajuste.

Se observó que (...)

Es importante estudiar este tipo de cuestiones y más aún en países como los estudiados, donde el sector económico está continuamente analizando, evaluando y actuando en base a ello. A su vez, con esta investigación se propone entender las razones por las que un país no paga sus deudas, y ver qué factores son los que más influyen en dicha cesación de pagos. Por estos motivos se decidió tomar Argentina, Ecuador, Paraguay, Bolivia y Nigeria como casos de estudio, siendo los países que más veces entraron en cesación de pagos desde 1970 hasta la actualidad. Este trabajo es relevante (...)

El trabajo se divide de la siguiente manera…

2. Literatura relacionada

La previsión de impagos soberanos como campo de investigación surgió en la década de 1970, cuando los niveles de deuda externa de los países en desarrollo aumentaron significativamente, lo que llevó a un creciente volumen de reestructuraciones soberanas en la década de 1980. A raíz de la caída del comunismo, se produjo un importante volumen de inversiones extranjeras en Europa del Este, Asia,

América Latina y África, y los inversores internacionales comenzaron a alarmarse de los riesgos que podría implicar la globalización del comercio mundial y de los mercados financieros. Frente al ataque terrorista en 2001 y los posteriores acontecimientos turbulentos pusieron en manifiesto los riesgos asociados con las relaciones internacionales que ya eran cada vez más difíciles de analizar y predecir. Tras la crisis financiera mundial de 2008-2010, el mismo cambio en la dependencia internacional impulsó a varios países a entrar en crisis de deuda, principalmente Grecia, España, Italia, Portugal e Irlanda. De esta forma, se volvió cada vez más relevante poder predecir este tipo de acontecimientos para que las consecuencias que conlleva sean por lo menos más leves.

La literatura sobre crisis de deuda se divide en dos grandes categorías: modelos teóricos de deuda soberana y el estudio de los determinantes de la deuda soberana. La mayoría de los estudios se centran en un aspecto concreto de la crisis de la deuda o en ciertos determinantes. En conjunto, la literatura sugiere una serie de factores macroeconómicos que influyen en la probabilidad de impago de las deudas soberanas.

La literatura teórica destaca una variedad de factores que pueden desencadenar el impago de la deuda soberana y crisis de deuda. El riesgo país es un término recurrente en este tipo de estudios que incorpora tanto riesgos soberanos como políticos y es elemental para entender los defaults soberanos. Expresa la posibilidad de que un país, como deudor/emisor soberano, no pueda o no tenga intención de cumplir sus obligaciones contractuales de pago con los acreedores o inversores extranjeros (Krayenbühl 1985). La calificación de impago otorgada por las agencias de calificación se considera, por tanto, como default soberano. Dado que los países afectados por el impago soberano suelen estar sujetos a las condiciones establecidas por el Fondo Monetario Internacional (FMI), varias publicaciones consideran un evento de este tipo como el período en el que un país supera el límite de los préstamos no concesionales del FMI. Manasse et al. (2003) realizaron un trabajo para el FMI, donde desarrollan un modelo de "Sistema de Alerta Temprana" de crisis de deuda soberana. Mediante un análisis logit y binario recursivo, identificaron variables macroeconómicas que reflejan factores de solvencia y liquidez que predicen un episodio de crisis de deuda con un año de antelación. El modelo logit predice el 74% de todas las entradas en crisis y envía pocas falsas alarmas, mientras que el árbol recursivo 89% pero envía más falsas alarmas.

A su vez, es extensa la literatura que examina cuales son los determinantes de los impagos soberanos. En un capítulo del documento del Banco Mundial, Primo Braga y Vincelette (2011) utilizan una base de datos que abarca 25 años para 46 países emergentes utilizando técnicas de promedios de modelos Bayesianos para determinar el conjunto de determinantes de defaults. Un primer examen de los datos indica que para toda la muestra, la probabilidad de impago se asocia sólo con el nivel de endeudamiento. Las variables que representan los costes de la deuda y el riesgo de refinanciación no parecen robustas como predictores del impago de la deuda, por lo que este

resultado confirma la opinión de la literatura de que sólo unas pocas variables macroeconómicas y de calidad institucional son necesarias para predecir los impagos (Kraay y Nehru 2006). Estos autores examinan empíricamente los determinantes del "debt distress". Utilizando regresiones probit, encuentran que los defaults soberanos dependen de un pequeño conjunto de factores: la carga de la deuda, la calidad de las políticas e instituciones, y las perturbaciones. Se demuestra que estos resultados son robustos y muestran que sus especificaciones básicas tienen un importante poder predictivo fuera de la muestra.

En los últimos años, las técnicas de Machine Learning fueron poco a poco predominando en los estudios empíricos por su mayor precisión al momento de predecir en comparación a los enfoques econométricos tradicionales. Si bien las técnicas de aprendizaje automático han sido extensamente utilizadas en estudios de otras disciplinas, esto no ha sido así en el caso de la economía, en donde la literatura y los avances son relativamente recientes y escasos. En los enfoques econométricos tradicionales, los problemas predictivos o clasificatorios involucran algún tipo de modelo de regresión lineal o logístico en los cuales la prioridad está marcada por el ajuste dentro de la muestra. En cambio, la tecnología de Machine Learning se enfoca en mejorar la capacidad predictiva fuera de la muestra y se destaca por capturar relaciones no lineales entre los predictores y por identificar patrones generales entre los datos (Varian, 2014). En este sentido, se detallan a continuación algunos trabajos que han aplicado estas técnicas para estudiar la problemática de la pobreza en distintos contextos.

Por un lado, Savona y Vezzoli (2013) utilizan un nuevo enfoque basado en un árbol de regresión para alcanzar el mejor compromiso entre la bondad de ajuste dentro de la muestra y la predictibilidad fuera de la muestra de los impagos soberanos. Utilizan datos de países emergentes como Grecia, Irlanda, Portugal y España (GIPS) durante el período 1975-2010. Los resultados muestran que la iliquidez y el historial de impagos junto con el crecimiento del PIB real y los tipos de interés de EE.UU, son los principales determinantes de la reciente crisis de la deuda soberana europea. Dado que este análisis implica una suerte de ejercicio de pronóstico comparando distintos modelos, se puede incluir el modelo que utilizan los autores para predecir los defaults de Argentina. x

Por otro lado, Manasse y Roubini (2009) desarrollaron un análisis utilizando el modelo CART para examinar factores macroeconómicos, financieros y políticos que explican las crisis de la deuda soberana. Las 50 variables iniciales se redujeron a 10 utilizando árboles de decisión en los que se desarrollaron reglas para reconocer las características de los países en situación de impago. Se llegó a la conclusión de que no todas las crisis eran similares y podrían diferenciarse en términos de solvencia, liquidez y riesgos macroeconómicos.

Alaminos et al. (2019) aplicaron la metodología de árboles de decisión difusos para predecir las crisis de la deuda soberana utilizando datos entre 1970 y 2017, y aplicando 30 variables y una validación cruzada de diez veces. El área bajo la curva ROC (AUROC) del modelo global fue del 94%, lo que

indica un gran poder de predicción. A su vez, Huang y Sethi (2017) desarrollaron modelos de SVM, RF y logit utilizando una base de datos del FMI que contiene 1200 observaciones. Las variables se redujeron mediante el análisis de componentes principales, y los resultados se sometieron a pruebas retrospectivas mediante un método de validación cruzada de diez veces. El RF resultó ser el mejor modelo de predicción, con una precisión de clasificación del 91%, seguido de los métodos SVM (89%), KNN (88%) y logit (87%).

3. Set de Datos

En esta sección se propone generar una descripción y explicación de la elección de variables que conforman el set de datos de la actividad económica real de Argentina, Bolivia, Ecuador, Paraguay y Nigeria. Para la construcción de esta base, se comenzó definiendo una búsqueda basada en los determinantes e indicadores principales que mueven, en términos reales, las 5 economías mencionadas. Para la elección de dichos países se distinguieron aquellos en los que se pueden notar importantes y numerosos períodos de default de deuda soberana en los últimos 50 años. Es por esto que el período en consideración se extiende entre 1970 a 2020. Formalmente, el indicador de default se define a partir de una composición de información proveniente del Fondo Monetario Internacional y de la calificadora Standard & Poor´s (S&P 500). Los datos de deuda externa pública y privada fueron extraídos de las estadísticas de deuda internacional del Banco Mundial y los datos macroeconómicos para cada país, de los indicadores internacionales del Banco Mundial.

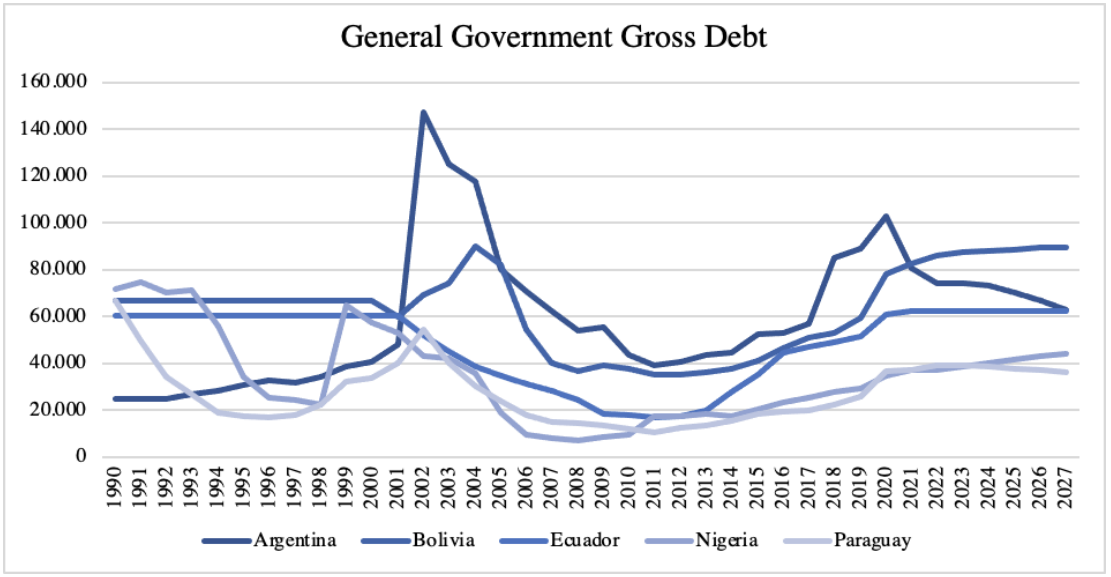
Para poder llevar a cabo una investigación con alta precisión, es importante establecer *a priori*, cuáles son los determinantes de un default de deuda soberana, e identificarlo para cada uno de los cinco países que se abordan en este trabajo. La evidencia empírica sugiere que la probabilidad de una crisis de deuda está positivamente correlacionada con niveles más altos de deuda total (McFadden et al., 1985) y deuda total a corto plazo (Detragiache y Spilimbergo, 2001), negativamente correlacionada con el crecimiento del PBI (Sturzenegger, 2004), y el nivel de reservas internacionales (Dooley, 2000). Además, también están relacionados con fluctuaciones de la producción más volátiles y persistentes (Catao y Sutton, 2002), una menor apertura comercial (Cavallo y Frankel, 2008), responde fuertemente a las condiciones políticas (Manasse et al., 2003), el historial previo de deuda (Reinhart et al., 2003) y el contagio (Eichengreen et al., 1996). En conjunto, estas investigaciones previas contribuyen a nuestra comprensión integral de algunos predictores de defaults soberanos.

Gran parte de la investigación previa a la predicción es explorar y entender en profundidad los datos que se disponen para poder usarlos de la manera más eficiente posible. Es por esto que, en primer lugar, se define que un país se encuentra en crisis de deuda si está clasificado como en default por el índice de Standard & Poor´s o si recibe un importante préstamo de parte del Fondo Monetario Internacional excediendo el 100 por ciento de la cuota asignada disponible. Investigaciones de la

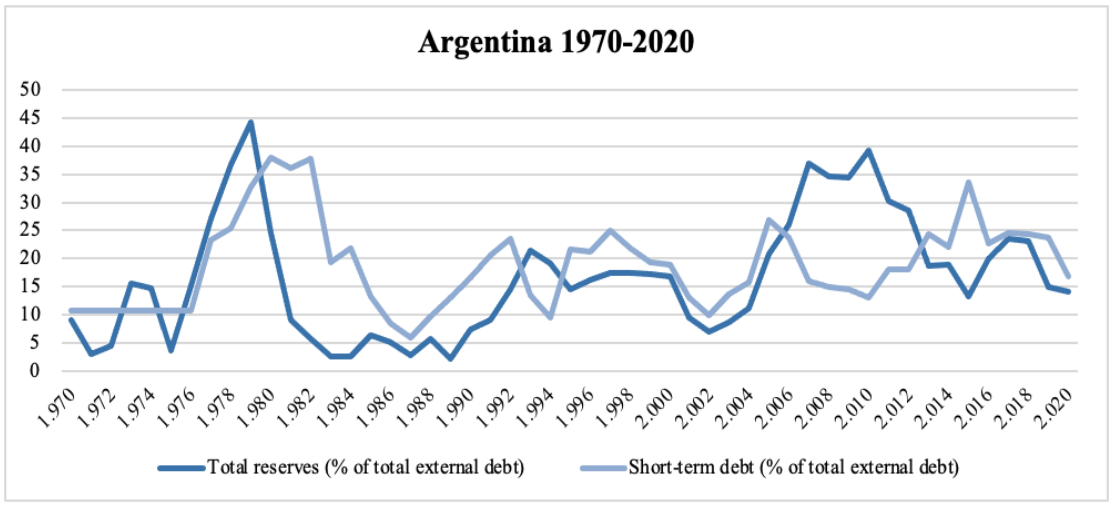
calificadora evidencian que las mayores vulnerabilidades del perfil crediticio de los países con historial de incumplimiento de pagos de deuda provienen del entorno financiero que se deteriora rápidamente, la falta de confianza en los mercados financieros acerca de las iniciativas políticas bajo próximas administraciones, y la incapacidad del Tesoro de financiarse a corto plazo con el sector privado.

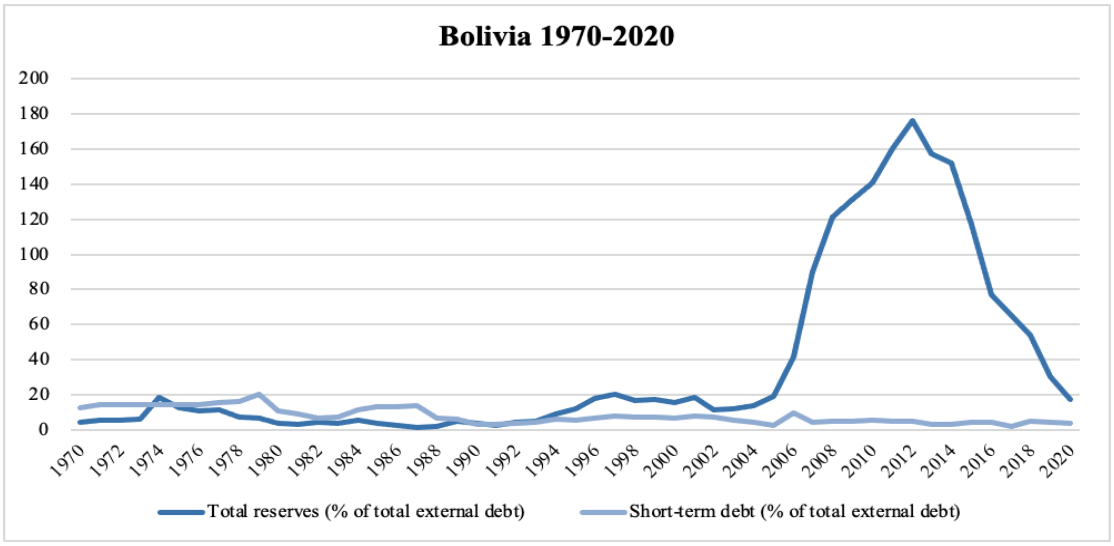
Según S&P un emisor es calificado como en incumplimiento de pagos o default si no ha cumplido con alguna o algunas de sus obligaciones financieras y lo considera como incumplimiento general, cuando el deudor no pagará específicamente ninguna o casi ninguna de sus obligaciones dentro de los plazos establecidos en las condiciones del préstamo. Como se menciona en el paper de Manasse et al. (2003), existe un potencial problema con esta definición ya que no captura los cuasi defaults que fueron prevenidos con un programa de ajuste y un importante paquete financiero proveniente del FMI. Es por esto mismo, que se decide añadir nuevas condiciones al significado de default de S&P para poder así abarcar estos potenciales problemas, en este caso, considerando los datos sobre préstamos y desembolsos incurridos a los 5 países considerados de parte del FMI. La institución del FMI se conforma de cuotas que son los componentes principales de la estructura financiera y de gobierno del FMI. La cuota de cada país miembro del fondo refleja su disposición económica relativa en la economía mundial. Estas se definen como Derechos especiales de giro (DEG), y funciona como unidad de cuenta del FMI. Basado en esta última definición, se clasifica a un país en crisis de deuda, si recibe un préstamo significativo que excede el porcentaje de cuota asignado y que un desembolso es realizado sobre dicho préstamo durante el primer año.

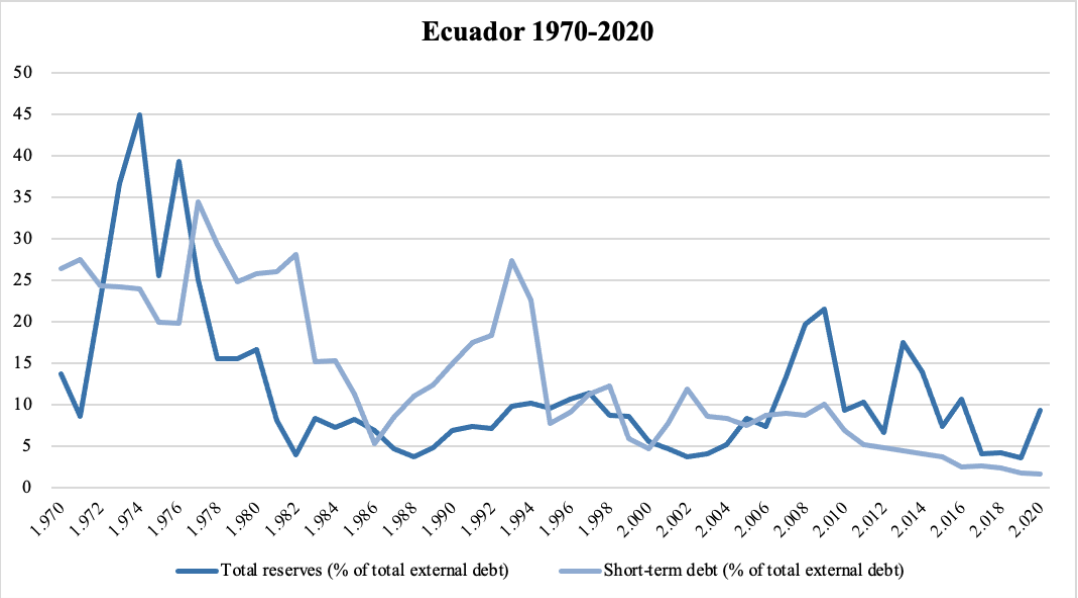
En segundo lugar, es de suma importancia estudiar la correlación entre las variables del dataset y distinguir los movimientos o tendencias a través del tiempo que puedan identificar un determinado patrón en los países investigados. También es relevante hacer este estudio para poder identificar las variables que determinan, en mayor medida, el default en un país. Para esto, a continuación se realizan distintos gráficos comparativos (Comentario: estos gráficos los haremos devuelta en Stata para mejorar el formato).

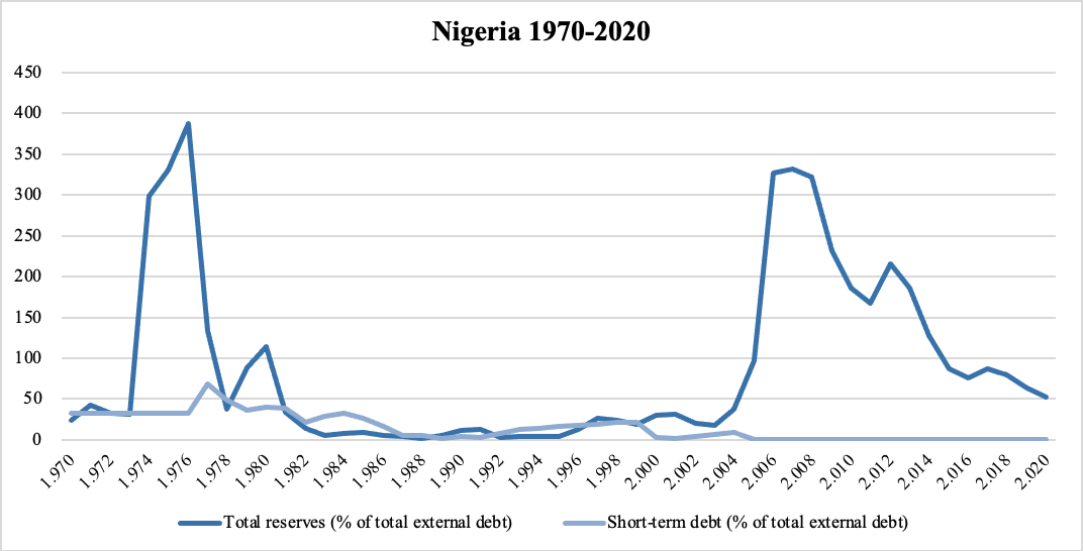


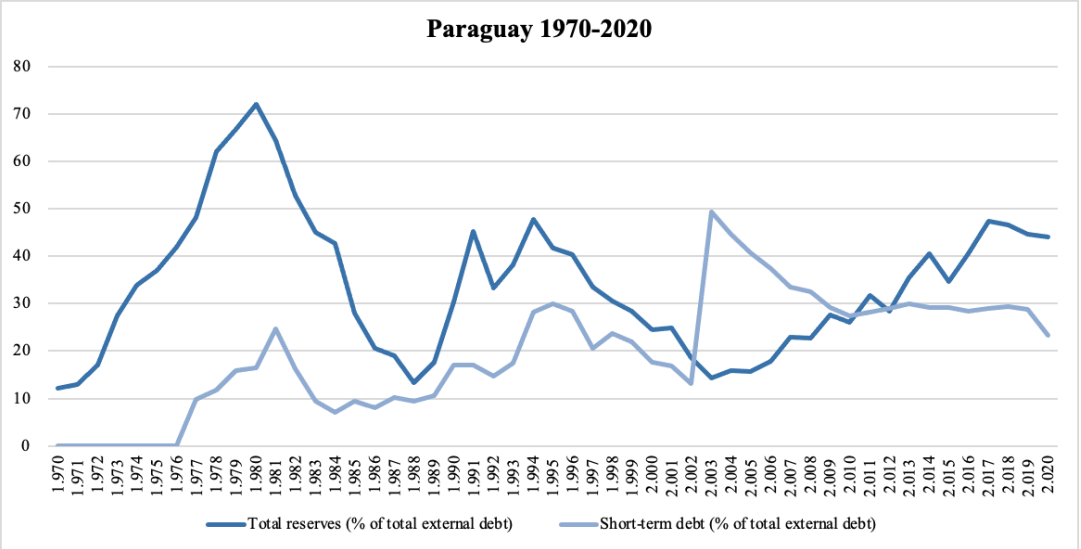
Short-term debt (% of total external debt) to total reserves (% of total external debt)



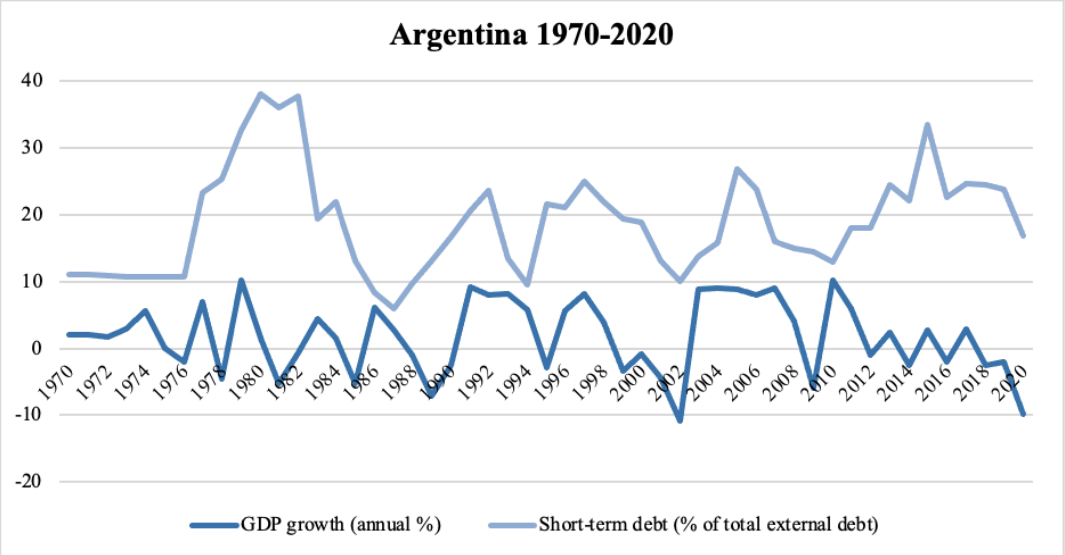


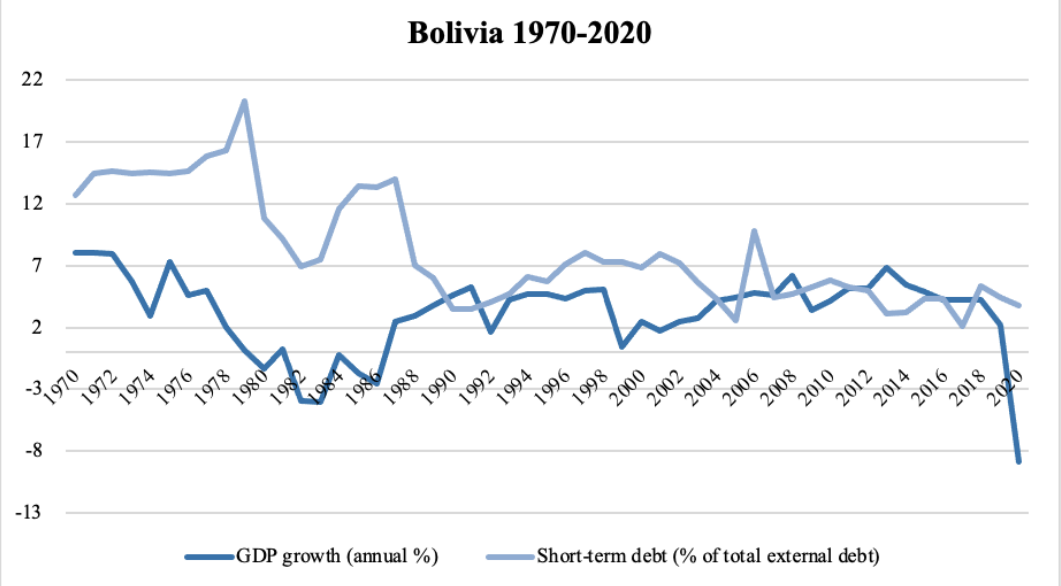

Ecuador 1970-2020
50
45
40
35
30
25
20
15
10
5
0
1.970 1.972 1.974 1.976 1.978 1.980 1.982 1.984 1.986 1.988 1.990 1.992 1.994 1.996 1.998 2.000 2.002 2.004 2.006 2.008 2.010 2.012 2.014 2.016 2.018 2.020
Total reserves (% of total external debt)
Short-term debt (% of total external debt)

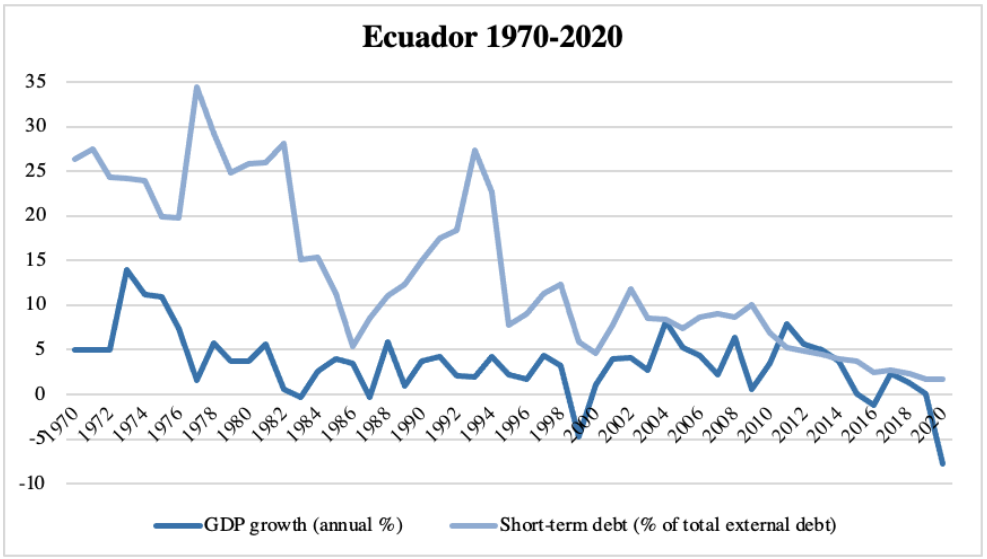

Nigeria 1970-2020
450
400
350
300
250
200
150
100
50
0
1.970 1.972 1.974 1.976 1.978 1.980 1.982 1.984 1.986 1.988 1.990 1.992 1.994 1.996 1.998 2.000 2.002 2.004 2.006 2.008 2.010 2.012 2.014 2.016 2.018 2.020
Total reserves (% of total external debt)
Short-term debt (% of total external debt)

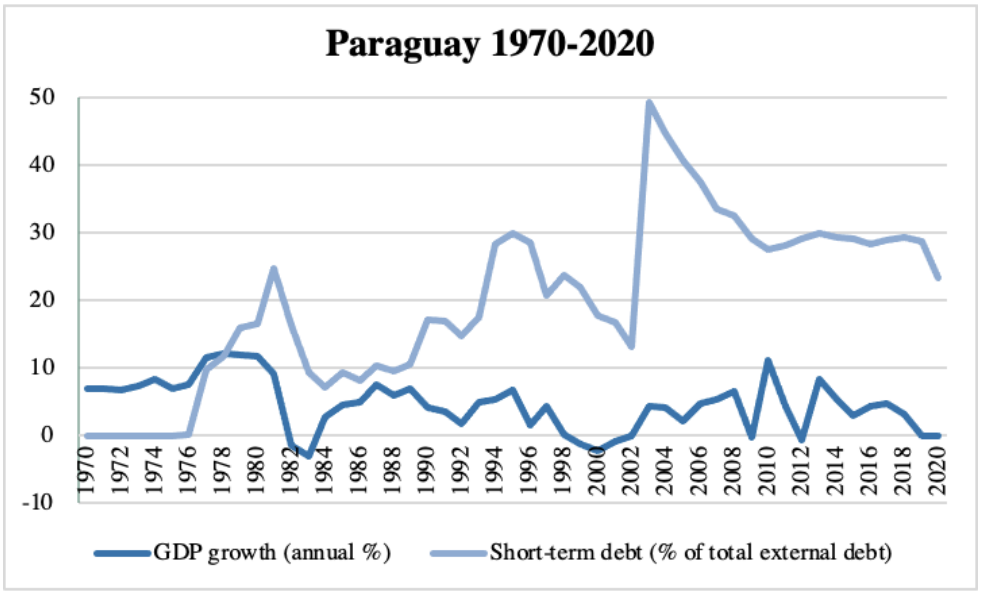

Paraguay 1970-2020
80
70
60
50
40
30
20
10
0
1.970 1.971 1.972 1.973 1.974 1.975 1.976 1.977 1.978 1.979 1.980 1.981 1.982 1.983 1.984 1.985 1.986 1.987 1.988 1.989 1.990 1.991 1.992 1.993 1.994 1.995 1.996 1.997 1.998 1.999 2.000 2.001 2.002 2.003 2.004 2.005 2.006 2.007 2.008 2.009 2.010 2.011 2.012 2.013 2.014 2.015 2.016 2.017 2.018 2.019 2.020
Total reserves (% of total external debt)
Short-term debt (% of total external debt)

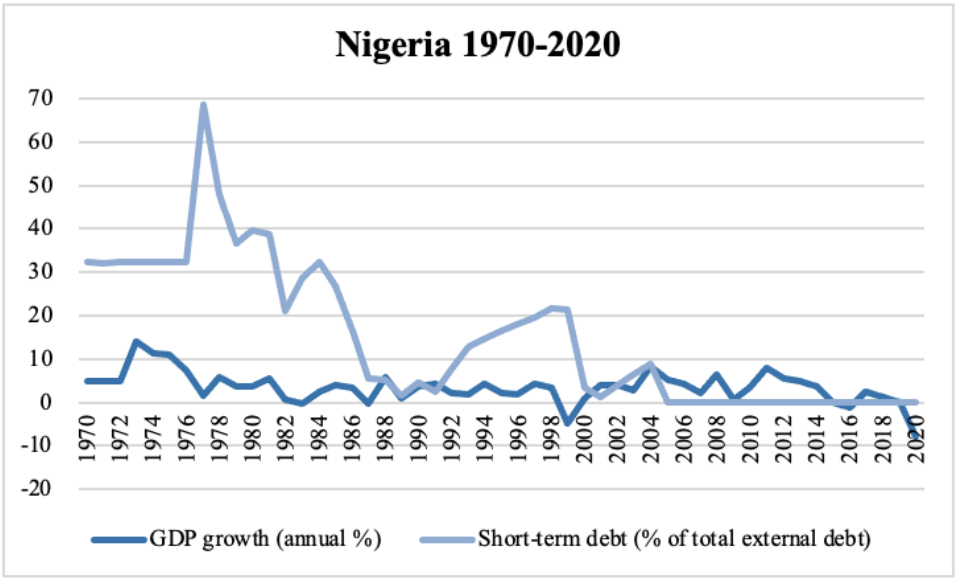
GDP Growth (% annually) to Short-Term Debt (% of Total External Debt)






Ecuador 1970-2020
35
30
25
20
15
10
5
0
-5
-10
1970 1972 1974 1976 1978 1980 1982 1984 1986 1988 1990 1992 1994 1996 1998 2000 2002 2004 2006 2008 2010 2012 2014 2016 2018 2020
GDP growth (annual %)
Short-term debt (% of total external debt)


Paraguay 1970-2020
50
40
30
20
10
0
-10
1970 1972 1974 1976 1978 1980 1982 1984 1986 1988 1990 1992 1994 1996 1998 2000 2002 2004 2006 2008 2010 2012 2014 2016 2018 2020
GDP growth (annual %)
Short-term debt (% of total external debt)


Nigeria 1970-2020
70
60
50
40
30
20
10
0
-10
-20
1970 1972 1974 1976 1978 1980 1982 1984 1986 1988 1990 1992 1994 1996 1998 2000 2002 2004 2006 2008 2010 2012 2014 2016 2018 2020
GDP growth (annual %)
Short-term debt (% of total external debt)

Luego de identificar las tendencias existentes entre las variables, es relevante notar la limitación presente en este trabajo de que la base de datos implementada se encuentra desbalanceada. Para esto, se deben encontrar los algoritmos de aprendizaje más adecuados, ya que la gran mayoría asumen una distribución relativamente equilibrada. Como se explica en la investigación de Miravet (2021), la presencia de clases desbalanceadas en bases de datos supone, en muchos casos, un problema en los modelos predictivos ya que estos tienden a centrar su atención sobre los casos de la clase mayoritaria. Es decir, en la mayoría de los años en estudio de cada país no hubo default, es una fracción menor, pero no menos significativa la de los años en default. En la muestra aproximadamente el % de los años hubo default. En consecuencia, se obtienen resultados que aparentan ser buenos, pero en definitiva, la predicción se realiza sólo en base a la clase mayoritaria. Es por esto que, el análisis de datos desbalanceados presenta una serie de características particulares que necesitan ser tratados de manera diferente tanto a la hora de entrenar como a la hora de evaluar en relación a los datos con distribución equilibrada. Un algoritmo que clasifique a todos los años como no default calificaría correctamente casi al % de las observaciones. Es por esto que mientras mayor sea el desbalance en la muestra, aumenta de manera artificial la precisión de las estimaciones. Según Somasundaram y Reddy (2016), existen dos formas principales de lidiar con este fenómeno.

La primera implica modificar los algoritmos de manera que le den una ponderación mayor a las observaciones de la categoría minoritaria, mientras que la segunda consiste en métodos de remuestreo aleatorio que equilibren la proporción de observaciones de las categorías de la variable de respuesta. Las técnicas más utilizadas dentro de este segundo grupo son: Sobremuestreo: consiste en incrementar aleatoriamente el número de observaciones

4. Los modelos

Siguiendo la literatura de Athey & W. Imbens (2019), el procedimiento tradicional en econometría, como se ha explicado en los conocidos textos de Angrist & Pischke (2008) y Wooldrige (2010), es especificar un objetivo que es funcional de una distribución conjunta de datos. Este objetivo es generalmente un parámetro de un modelo estadístico que describe la distribución de un conjunto de variables, condicionales a otras variables, en términos de un set de parámetros, que puede ser tanto finito como infinito. Dada una muestra aleatoria de la población de interés, los parámetros son estimados buscando los valores que mejor se ajusten a los datos, y esto se lleva a cabo, usando una función objetivo como la suma de los errores al cuadrado o con una función de probabilidad. Dicho esto, se puede identificar que en econometría tradicional, el foco se encuentra en la calidad de los estimadores del objetivo, que por lo general se mide en base a la eficiencia de muestras grandes. Los autores también mencionan el interés por construir intervalos de confianza, y que típicamente, se reportan errores estándares y estimaciones puntuales. En otras palabras, lo que importa es conocer la forma de la función de donde salen los datos. Este enfoque se concentra en la estimación de un

modelo, típicamente representado por $y_i = x_i'\beta + \mu_i$, exógeno en donde la relación entre la variable y_i dependiente y el vector de regresores x_i' está determinada por una teoría o una "estructura" (Sosa Escudero, 2018). Esto significa que lo que se busca es estimar de la mejor manera posible a los coeficientes β en donde la calidad que posee el estimador suele asociarse con ciertas propiedades deseables. Existen preferencias lexicográficas por la insesgadez en la mayoría de los casos, dejando a la eficiencia del estimador relegado en un segundo plano.

Sin embargo, en contraste con la literatura de la econometría tradicional, en la literatura de Machine Learning, los autores argumentan que el foco principal se encuentra en el desarrollo de algoritmos. El objetivo de estos algoritmos es generalmente, hacer predicciones sobre algunas variables dadas otras variables, o clasificar unidades sobre la base de información limitada. La diferencia con técnicas anteriores está en su capacidad para adaptarse a los cambios en los datos a medida que van entrando en el sistema y aprender de las propias acciones del modelo. Es por esto que, el aspecto iterativo es importante porque a medida que los modelos son expuestos a nuevos conjuntos de datos, estos pueden adaptarse independientemente. La forma del modelo, se aprende en base a los datos, y no se realiza inferencia, sino una predicción puntual, ya que se requiere obtener un poder predictivo fuera de la muestra. En otras palabras, el aprendizaje automático busca predecir y en base a x, donde el modelo en sí, no tiene un rol destacado. Particularmente, lo que se quiere es predecir correctamente fuera de la muestra, esto es, evaluar la capacidad de predicción en observaciones que no se utilizaron para la construcción del modelo. Existe una preferencia por tolerar métodos más sesgados a cambio de una significativa disminución en la varianza, en comparación a los métodos de econometría frecuentista (Sosa Escudero, 2018). Los modelos con mayor grado de complejidad tienden a ser menos sesgados, pero a tener una mayor varianza y ser más erráticos, mientras que un menor grado de complejidad en el modelo permite disminuir la varianza a costas de un sesgo mayor. Para controlar dicha complejidad existen los hiperparámetros que maximizan la precisión de la predicción en base a una función de pérdida. Otra gran ventaja del enfoque de Machine Learning es a la hora de entender o encontrar patrones generales en los datos. Dentro del aprendizaje automático existen diferentes algoritmos o tipos de modelo que difieren en el tipo de datos ya sea de entrada como de salida, la estructura y la complejidad computacional. Existen dos tipos de algoritmos de aprendizaje estadístico que son comúnmente empleados en este tipo de estudio. Como señala Plukikova (2016), los algoritmos de aprendizaje no supervisado buscan elaborar e identificar alguna estructura entre los datos sin distinguirla de antemano. Funcionan a partir de un espacio de características sin la necesidad de una variable de respuesta. Los algoritmos de aprendizaje supervisado distinguen un patrón inicial en los datos y por lo general, funcionan con variables predictoras y con variables de respuesta. Estos se utilizan para tareas predictivas como lo es el problema de clasificación binaria de este trabajo.

En este caso fueron entrenados 8 modelos en el set de entrenamiento: regresión logística, análisis discriminante lineal, KNN, árbol de decisión, support vector machine (SVM), bagging, random forest y boosting. Para poder llevarlo a cabo, se eligió usar SKLearn para implementar, entrenar y testear los modelos (Pedregosa et al., 2011). Para el ajuste y la validación de hiperparámetros se utiliza k-fold validación cruzada, con un k=10. Para ello, se siguen los siguientes pasos generales para cada algoritmo utilizado.

A continuación abordaremos, con detalle, la explicación del funcionamiento de cada uno de los modelos propuestos. Es importante remarcar que la selección de los modelos no es exhaustiva a todos los que se podrían tener en cuenta en dicha carrera de caballos, pero incluimos aquellos que identificamos como los más relevantes para el caso dado.

Regresión Logística

La regresión logística es utilizada para estimar la probabilidad de que una variable pertenezca a una clase particular. Dicho esto, si la probabilidad estimada es mayor al 50%, el modelo predecirá que la variable pertenece a la clase positiva, caso contrario el modelo predice que pertenece a la clase negativa. Como se indica en el paper de Muñoz Jaramillo (2021), la probabilidad estimada se puede identificar con la siguiente ecuación:

$$\widehat{p} = \sigma(X^T\theta) \qquad \widehat{y} = \begin{cases} 0 & \text{if } \widehat{p} < 0{,}5 \\ 1 & \text{if } \widehat{p} \geq 0{,}5 \end{cases} \tag{2-1}$$

En donde $\widehat{p}$ es la probabilidad estimada, σ es la función sigmoide, X es la matriz de características, θ es el vector de parámetros del modelo y $\widehat{y}$ corresponde a la predicción.

Análisis Discriminante Lineal

KNN

Árbol de Decisión

SVM

Este modelo de SVM, por sus siglas en inglés, o Máquina de Soporte Vectorial, forma parte de los métodos de aprendizaje supervisado para la regresión, clasificación o detección de extremos. En este modelo se identifican funciones base que están centradas en los puntos de datos de entrenamiento de la muestra y, seguidamente, se selecciona un subconjunto de dichos puntos los cuales se los nombra vectores de soporte. Este tipo de modelos predictivos son empleados para la clasificación de conjuntos

de datos complejos, pero donde el tamaño es mediano o pequeño, y además pueden realizar tareas de clasificación lineales y no lineales (Burges, 1998).

Una propiedad importante de este tipo de modelos es que los parámetros se determinan como la solución a un problema de optimización con función de costo convexa, por lo que, a pesar de que se involucre un problema no lineal, la solución es relativamente sencilla.

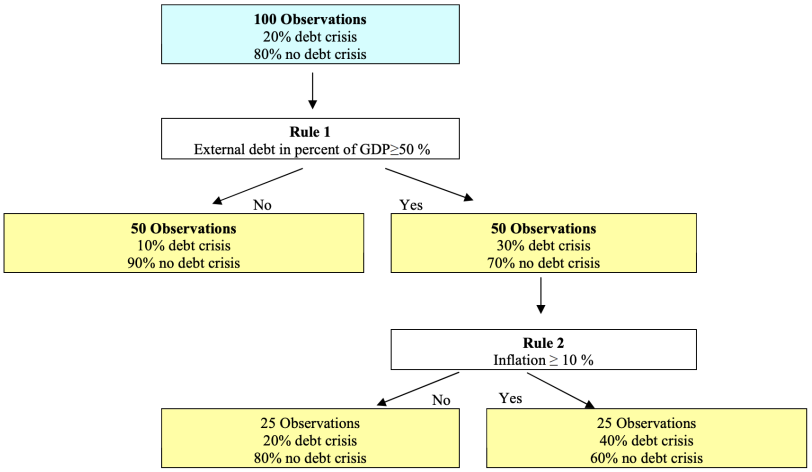
Bagging

Random Forest

Decidimos emplear el modelo de random forest (Liaw et al., 2002) porque este obtuvo buenos resultados en estudios previos, y además tiende a no estar tan afectado por los datos no balanceados como otros modelos.

Boosting

En primer lugar, propondremos utilizar la metodología de CART, que realiza un análisis en árbol de regresión a través de una secuencia de reglas para predecir un outcome binario, descrito en la literatura de Manasse, P., Roubini, N. and Schimmelpfennig, A. (2003). CART realiza una regresión no paramétrica, es decir que no se estiman parámetros, sino que directamente los datos nos indican la distribución. Por este motivo, CART puede detectar relaciones complejas entre las variables explicativas o independientes (X) y las variables dependientes (Y). En esencia, lo que hace este método es partir los espacios de atributos (de las X), y ajustar un modelo simple para Y dentro de cada región, donde propone como predicción la media muestral de Y. Este procedimiento es una partición recursiva binaria, y se debe encontrar el mejor ajuste global, es decir, la variable y el punto de partición óptimos. Este método funciona muy bien para estructuras no-lineales. Para un mejor entendimiento de la metodología que utiliza CART, a continuación dejamos como ejemplo el árbol creado en el paper académico de Manasse et al. (2003) y que es la base del árbol de regresión que proponemos realizar para nuestra investigación.



El siguiente es el modelo Logit. Logit es una técnica estadístico de predicción bastante simple en su lógica de metodología, dado que se utilizan conceptos estadísticos conocidos, como máxima verosimilitud, para estimar los parámetros de interés y poder predecir la probabilidad condicional de que ocurra un suceso (en nuestro caso que suceda un default) dado que se cumple un set determinado de características.

Vecinos cercanos, por otro lado, es un modelo relativamente simple, pero que podría brindarnos mucha información acerca de las características de los países que entran en default, y cuáles de ellas son relevantes para poder realizar una predicción correcta. Aunque no creemos que termine siendo el modelo que mejor prediga, lo que se puede hacer es explotar la facilidad con la que este modelo se extiende a múltiples categorías para poder tener en cuenta distintas características cualitativas o cuantitativas de los países (cómo orientación política del presidente, disturbios sociales, inestabilidad económica, nivel de confianza en el gobierno, nivel de confianza en sistema financiero, distancia (en días) respecto al último default, nivel de deuda respecto a gasto público, entre otros) y conseguir un modelo, por falta de mejor palabra, de país propenso a caer en defaults

5. Resultados

En esta sección abordaremos en primer lugar las diferentes métricas de desempeño y luego las analizaremos para cada algoritmo en particular.

6. Conclusiones

Referencias

Alaminos, David, Sergio M. Fernández, Paulo Magalhães Neves, and José António C Santos. (2019). "Predicting sovereign debt crises with fuzzy decision trees". Journal of Scientific & Industrial Research 78: 733–37

Catao, L., Sutton, B. (2002). 'Sovereign Defaults: The Role of Volatility', IMF Working Paper, 02/149.

Cavallo, E. A. and Frankel, J. A. (2008). 'Does Openness to Trade Make Countries More Vulnerable to Sudden Stops, or Less? Using Gravity to Establish Causality', Journal of International Money and Finance, Vol. 27, pp. 1430-1452.

Detragiache, E., Spilimbergo, A,. (2001). "Crises and Liquidity: Evidence and Interpretation". IMF Working Paper No. 01/2.

Dooley, M. (2000). 'A Model of Crises in Emerging Markets', The Economic Journal, Vol. 110, pp. 256-272.

Eichengreen, B., Rose, A. and Wyplosz, C. (1996). "Contagious Currency Crises: First Tests", Scandinavian Journal of Economics, Vol. 98, pp. 463-84.

Hébert, B., Schreger, J. 2017. "The costs of sovereign default: Evidence from Argentina". American Economic Review, 107(10), 3119-45

Huang, Andrew, and Taresh Sethi. (2017). "Predicting sovereign default". In Proceedings of the 34th International Conference on Machine Learning. Sydney: PMLR 70

Krayenbuehl, Thomas E. 1985. "Country Risk: Assessment and Monitoring". Cambridge: Woodhead-Faulkner. [CrossRef]

Kraay, A., and V. Nehru. 2006. "When Is External Debt Sustainable?" World Bank Economic Review 20 (3): 341–65.

Manasse, P. and Roubini, N. (2009). 'Rules of Thumb for Sovereign Debt Crises', Journal of International Economics, Vol. 78, pp. 192-205.

Manasse, P., Roubini, N. and Schimmelpfennig, A. (2003). 'Predicting Sovereign Debt Crises', IMF Working Paper, 03/221.

McFadden, D., Gershon, F., Vassilis, H., O'Connell, S,. (1985). "Is There Life After Debt? An Econometric Analysis of the Creditworthiness of Developing Countries", in Gordon Smith and John Cuddington, eds., International Debt and the Developing Countries. World Bank: Washington, DC.

Reinhart, C., Rogoff, K. and Savastano, M. (2003). 'Debt Intolerance', Brookings Papers on Economic Activity, Vol. 1, pp. 1-74

Savona, R., & Vezzoli, M. 2015. "Fitting and Forecasting Sovereign Defaults using Multiple Risk Signals". Oxford Bulletin of Economics and Statistics, 77(1), 66–92.

Sturzenegger, F. (2004). 'Toolkit for the Analysis of Debt Problems', Journal of Restructuring Finance, Vol. 1, pp. 201-203

Miravet, Blanca Abella (2021). "Mejora de las predicciones en muestras desbalanceadas", Universidad Autónoma de Madrid de Escuela Politécnica Superior.

Muñoz Jaramillo, Victor Daniel (2021). "Evaluación de Modelos de Machine Learning para la Predicción de Crímenes en la Ciudad de Medellín", Universidad Nacional de Colombia.

Athey, Susan and W.Imbens, Guido (2019). "Machine Learning Methods That Economists Should Know About", Annual Review of Economics.

Angrist JD, Pischke JS. (2008). "Mostly Harmless Econometrics: An Empiricist's Companion", Princeton, NJ: Princeton Univ. Press.

Wooldridge JM. (2010). “Econometric Analysis of Cross Section and Panel Data”, Cambridge, MA: MIT Press

Burges C. J. C. , (1998). “A tutorial on support vector machines for pattern recognition,” Data Mining and Knowledge Discovery, vol. 2, pp. 121–167.

Plulikova, N. (2016). Poverty analysis using machine learning methods. Bachelor’s in Mathematics Thesis, Comenius University in Bratislava.

Sosa Escudero, W. (2018). Big data y aprendizaje automático: Ideas y desafíos para economistas, en una nueva econometría: Automatización, big data, econometría espacial y estructural. Universidad Nacional del Sur.