UNIVERSIDAD DE CONCEPCIÓN
FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS

Predecir la probabilidad de incumplimiento a través de un modelo de regresión sobre el intervalo unitario

Por: Camila Andrea Silva Calabrano

Tesis presentada a la Facultad de Ciencias Físicas y Matemáticas de la Universidad de Concepción para optar al titulo profesional de Ingeniería Civil Matemática

Enero de 2024
Concepción, Chile

Profesores: Guillermo Ferreira, Pamela Meléndez

Resumen

En un entorno económico y financiero cada vez más complejo, la gestión y evaluación de riesgos es extremadamente fundamental para la toma de decisiones informada y sostenible en instituciones financieras. La capacidad de anticiparse y proyectar eventos adversos y comprender la probabilidad de incumplimiento, así como las pérdidas asociadas, se convierte en un componente crucial para garantizar la estabilidad y solidez de dichas instituciones. Aunque comúnmente se adopta la técnica estadística de regresión múltiple para proyectar tanto la probabilidad de incumplimiento como la pérdida dado el incumplimiento forward-looking, muchos autores sugieren la existencia de varios modelos estadísticos capaces de realizar esta proyección. Es por esto que el objetivo de la presente memoria de titulo consiste en comparar la capacidad predictiva de 5 modelos de regresión y/o series de tiempo para proyectar la probabilidad de incumplimiento y la pérdida dado el incumplimiento forward-looking. Especificamente los modelos estudiados fueron la Regresión Cox, la Regresión Beta, el Modelo Aditivo Generalizado, el Modelo Autorregresivo con Variables Exógenas y el Modelo Autorregresivo Integrado de Media Móvil con Variables Exógenas.

Es importante destacar que, si bien el objetivo de la presente memoria de título es comparar la capacidad predictiva de cinco modelos estadísticos en la proyección de la probabilidad de incumplimiento y la pérdida dado el incumplimiento forward-looking, no todos los modelos considerados incorporan explícitamente la dependencia temporal en sus estructuras. Siendo conscientes de esta distinción, se realiza una evaluación exhaustiva de cada modelo en función de su capacidad predictiva y su idoneidad para el análisis de riesgos financieros.

Índice

Índice de cuadros

Índice de figuras

1. Introducción

La probabilidad de incumplimiento es un término financiero que refleja la posibilidad de que un cliente no cumpla con sus obligaciones de deuda en un horizonte de tiempo específico. Por su parte, la pérdida dado el incumplimiento es una métrica que permite cuantificar la pérdida esperada en caso de que un cliente incumpla con sus obligaciones de pago, es decir, la pérdida esperada representa el porcentaje que se espera perder en caso de incumplimiento.

De acuerdo con la normativa IFRS 9 (Board, 2014), se considera que un cliente entra en incumplimiento al presentar un retraso en sus pagos igual o superior a 90 días, o mostrar una capacidad de pago deteriorada o nula, o bien requerir la renegociación de sus deudas.

En la misma normativa se define el componente forward-looking como un ajuste esencial, pues permite no solo incorporar información de morosidad, sino también considerar toda la información crediticia pertinente, incluida la información macroeconómica con una perspectiva hacia el futuro. El objetivo principal es anticipar las pérdidas crediticias a lo largo del tiempo de vida del activo.

Así se definen la probabilidad de incumplimiento forward-looking y la pérdida dado el incumplimiento forward-looking, la cuales se refieren a las proyecciones de la probabilidad de que una cuenta incurra en incumplimiento y la pérdida esperada dado esos incumplimientos, basándose en las variables macroeconómicas proyectadas, respectivamente.

En este contexto, se emplean métodos como la regresión y análisis de series temporales para proyectar esta probabilidad de incumplimiento. Las regresiones permiten entender la relación entre variables macroeconómicas y la probabilidad de incumplimiento, mientras que los modelos de series temporales incorporan la dimensión temporal, capturando patrones y tendencias en los datos a lo largo del tiempo.

Para las instituciones financieras, resulta fundamental contar con procedimientos efectivos para proyectar la probabilidad de incumplimiento de sus carteras. Una decisión equivocada podría aumentar el riesgo de sufrir pérdidas al no poder recuperar el monto prestado. Esto se debe a que el nivel de provisiones podría no ser suficiente para cubrir dichas pérdidas.

1.1. Revisión Bibliográfica

El modelo de probabilidad de incumplimiento forward-looking, según la norma IFRS 9 (Board, 2014), tiene como requisito contar con la influencia de las condiciones macroeconómicas actuales y previstas sobre las tasas de incumplimiento. A continuación, se presentan algunas referencias relevantes:

En el trabajo de Bellini (Bellini, 2019), se enumeran los modelos tradicionales VAR (vectorial autorregresivo) y VEC (corrección de errores vectoriales) como herramientas esenciales para abordar series temporales macroeconómicas y predecir la probabilidad de incumplimiento. En el trabajo de Jacobs (Jacobs Jr, 2019), se utilizaron modelos vectoriales autorregresivos, especialmente VARMAX, que incorpora variables exógenas a los modelos VAR más convencionales.

Por otro lado, Tasche (Tasche, 2015) emplea modelos de regresión para realizar una regresión de la tasa de incumplimiento con respecto a las variables macroeconómicas, aunque destaca que este enfoque tiene la desventaja de requerir series temporales largas de observaciones.

1.2. Objetivos

Como se observa, no existe una única metodología utilizada para predecir la probabilidad de incumplimiento y la pérdida dado el incumplimiento, siendo las más comunes los modelos de series temporales y los modelos de regresión. Por lo tanto, no es posible seleccionar un mejor modelo de manera global.

Es por esto que, en colaboración con una institución financiera, se define esta memoria de título con el objetivo de comparar metodologías para predecir la probabilidad de incumplimiento forward-looking y la pérdida dado el incumplimiento utilizando el software estadístico R. Se proponen diferentes metodologías para la proyección, entre las cuales se incluyen: Regresión Cox, Regresión Beta, Modelo Aditivo Generalizado, Modelo Autorregresivo con Variables Exógenas y Modelo Autorregresivo Integrado de Media Móvil con Variables Exógenas.

Así se definen los siguientes objetivos:

1. Estudiar e implementar los diferentes modelos a datos reales en la predicción de la probabilidad de incumplimiento y pérdida dado el incumplimiento

forward-looking para una institución financiera.

2. Comparar los resultados de los diferentes modelos, con el fin de determinar la metodología con mejor rendimiento y ajuste.

2. Marco Teórico

En este capítulo se describirán los conceptos básicos asociados a los diferentes algoritmos y/o modelos a utilizar.

2.1. Conceptos claves

2.1.1. Definición de incumplimiento

Se detalla en la norma IFRS 9 Instrumentos Financieros (Board, 2014), apartado B 5.5.37 lo siguiente respecto a la marca de incumplimiento a utilizar:

Al definir incumplimiento a efectos de determinar el riesgo de que ocurra un incumplimiento, una entidad aplicará una definición de incumplimiento que sea congruente con la definición utilizada a efectos de gestión del riesgo crediticio interno para el instrumento financiero relevante y considerará indicadores cualitativos (por ejemplo, pactos financieros) cuando sea apropiado. Sin embargo, hay una presunción refutable de que un incumplimiento no ocurrirá después de que un activo financiero esté en mora 90 días, a menos que una entidad tenga información razonable y sustentable que un criterio de incumplimiento más aislado es más apropiado. La definición de incumplimiento utilizada a estos efectos deberá aplicarse de forma congruente a todos los instrumentos financieros, a menos que la información pase a estar disponible lo que demuestra que otra definición de incumplimiento es más adecuada para un instrumento financiero concreto.

2.1.2. Fuentes de información

Se detalla en la norma IFRS 9 Instrumentos Financieros (Board, 2014), apartado B 5.5.52 lo siguiente respecto a las fuentes de información a utilizar:

La información histórica es un sostén o base importante desde la cual medir las pérdidas crediticias esperadas. Sin embargo, una entidad ajustará la información histórica, tal como la experiencia de pérdidas crediticias, sobre la base de la información observable actual para reflejar los efectos de las condiciones actuales y su pronóstico de condiciones futuras que no afecten al periodo sobre el cual se basa la información histórica, y eliminar los efectos de las condiciones en el periodo histórico que no son relevantes para los flujos de efectivo contractuales futuros.

2.1.3. Componente Forward-Looking (FL)

Se detalla en/ la norma IFRS 9 Instrumentos Financieros (Board, 2014), apartado B5.5.4 y B5.5.41 lo siguiente respecto a la definición de Forward-Looking y Probabilidad de incumplimiento Forward-looking, respectivamente.

Probabilidad de incumplimiento Forward-Looking: *La interpretación de la Junta de Normas Internacionales de Contabilidad (IASB, por sus siglas en inglés), en el estándar IFRS 9, respecto a mirada prospectiva o forward-looking, hace alusión a las proyecciones de largo plazo del ciclo económico, con el fin de capturar tanto escenarios favorables como desfavorables que pudieran llegar a suceder e impactar en la cartera. Dado lo anterior, el ajuste forward-looking, corresponde a proyectar la probabilidad de que una cuenta incurra en un incumplimiento, con base en variables macroeconómicas proyectadas.*

Forward-looking: *Esta información sobre el riesgo crediticio integral debe incorporar no solo información sobre morosidad, sino también toda la información crediticia relevante, incluida la información macroeconómica con vistas al futuro, para aproximarse al resultado de reconocer las pérdidas crediticias durante el tiempo de vida del activo cuando había habido un incremento significativo en el riesgo crediticio desde el reconocimiento inicial a un nivel de instrumento individual.*

2.2. Métodos de Selección de Variables

La selección de variables es fundamental en el análisis de datos y el modelado estadístico, pues permite contar con una interpretación simplificada y precisa de los modelos estadísticos, además permite ahorrar en los costes de almacenamiento y eliminar los ruidos o variables no relacionadas con los modelos.

El problema de selección de variables consiste principalmente en identificar y elegir las características o atributos más relevantes y significativas de un conjunto de datos.

Por esta razón se utilizarán los métodos de selección de variables Boruta (Kursa and Rudnicki, 2010) y el método de regularización de Lasso (Tibshirani, 1996), según sea conveniente. Ambos métodos serán definidos a continuación.

2.2.1. Boruta

Boruta (Kursa and Rudnicki, 2010) es una técnica de selección de variables diseñada para identificar las variables más relevantes en un conjunto de datos, basada principalmente en el algoritmo Random Forest. Boruta proporciona una estimación numérica de la importancia de las variables la cual es calculada por separado para todos los árboles del bosque que utilizan una variable determinada. Teniendo la importancia entregada por Random Forest es posible calcular los puntajes estándar como sigue:

$$Z_i = \frac{x_i - \mu}{\sigma},$$

donde μ es la media y σ es la desviación estándar de las importancias de las variables, $i \in V$, con V el conjunto de variables del modelo. Este puntaje mide qué tan lejos está la importancia de la i-ésima variable con respecto a la desviación estándar de todas las importancias. Este puntaje es de utilidad en el algoritmo Boruta pues es utilizado para discernir entre las variables importantes y no importantes.

El algoritmo Boruta evalúa la importancia de las variables mediante el cálculo de puntajes estándar basados en la desviación de las importancias obtenidas a través del algoritmo Random Forest. Estos puntajes indican qué tan significativa es cada variable en relación con el conjunto de datos. Durante el proceso, se comparan estas puntuaciones con un umbral máximo derivado de las variables 'sombra' aleatorizadas, lo que permite discernir entre variables importantes y no importantes. Aquellas variables con importancia significativamente mayor o menor que este umbral son consideradas respectivamente como importantes o no importantes y son ajustadas en consecuencia.

El algoritmo Boruta funciona de la siguiente forma:

1. Se expande la tabla de datos agregando copias de todas las variables explicativas, a estas copias se le llaman variables sombra.
2. A las variables sombra se les aleatoriza para eliminar sus correlaciones con la variable de respuesta.
3. Se ejecuta un clasificador de Random Forest en la tabla de datos extendida

y se calculan los puntajes Z.

4. Se encuentra el máximo puntaje Z entre los variables sombra (que definiremos como MZSA), y luego se asigna un conteo a cada variable i que obtuvo un puntaje mejor que el MZSA ($Z_i > MZSA$).
5. Para cada variable explicativa con importancia no determinada, se realiza una prueba bilateral de igualdad con respecto al MZSA.
6. Se considerarán las variables que tienen importancia significativamente menor que el MZSA como no importantes y se eliminan de forma permanente de la tabla de datos.
7. Se considerarán las variables que tienen importancia significativamente mayor que el MZSA como importantes.
8. Se eliminan todas las variables sombra.
9. Se repite el procedimiento hasta que se asigne la importancia a todas las variables.

Si bien el algoritmo Boruta está construido para trabajar sobre un modelo Random Forest, esto no limita su uso a algoritmos de clasificación o problemas de regresión.

2.2.2. Regularización Lasso

La Regularización Lasso (Tibshirani, 1996) es una técnica utilizada para mejorar la capacidad de generalización de un modelo y seleccionar automáticamente las variables más importantes.

Uno de los problemas más comunes al construir modelos es el sobre-ajuste (overfitting). Esto ocurre cuando el modelo es muy complejo y se ajusta demasiado a los datos de entrenamiento. Como resultado, el modelo no se generaliza bien a nuevos datos y tiene un mal rendimiento en la predicción.

La regularización Lasso aborda el problema de sobre-ajuste (overfitting) agregando una penalización (la cual llamaremos L_1) a la función de costo. La penalización se define de la siguiente manera:

$$L_1(\beta) = \lambda \sum_{j=1}^{p} |(\beta_j)|,$$

donde $L_1(\beta)$ es la penalización, λ es el hiperparámetro de regularización que controla el grado de penalización y $|\beta_j|$ es el valor absoluto del coeficiente β_j correspondiente a la variable predictora x_j.

Esta penalización tiene como efecto forzar a algunos coeficientes a ser exactamente iguales a cero, lo que significa que algunas variables se excluyen completamente del modelo, esto realiza automáticamente la selección de variables, identificando las variables menos importantes. En otras palabras, algunas variables se excluyen completamente del modelo, lo que conduce a una selección automática de características al identificar las menos importantes. Además, la penalización Lasso reduce la magnitud de los coeficientes no nulos, lo que ayuda a prevenir el sobreajuste al reducir la complejidad del modelo. El hiperparámetro de regularización λ controla el grado de penalización: a medida que λ aumenta, la penalización también aumenta, lo que resulta en una mayor exclusión de variables y una reducción de coeficientes.

2.3. Modelos de Regresión

2.3.1. Regresión Cox

La regresión de Cox, también conocida como el modelo de riesgos proporcionales de Cox, es un modelo estadístico utilizado en el análisis de datos de supervivencia. Fue desarrollado por el estadístico británico Sir David Cox en 1972 (Cox, 1972).

Los modelos de riesgo tienen la función de dar cuenta del tiempo, lo que los convierte en modelos dinámicos. Tienen la capacidad de incorporar variables que varían en el tiempo, es decir, variables explicativas que pueden asumir diferentes valores para diferentes tiempos.

El modelo de Cox asume que la relación entre las variables explicativas (o covariables) y la función de riesgo es multiplicativa y constante en el tiempo. Esto se conoce como la "proporcionalidad de riesgos". Matemáticamente, la función de riesgo en el modelo de Cox se expresa como:

$$h_i\left(t\right) = h_0\left(t\right) \cdot \ exp\left(\sum_{j=1}^{p} \beta_j x_j^i\left(t\right)\right),$$

donde $h(t)$ es la función de riesgo que describe la tasa de cambio en la probabilidad

de falla en el tiempo t, $h_0(t)$ es la función de riesgo de referencia, $x_j^i(t)$ denota el valor de la j-ésima variable explicativa de i en el momento t y β_j describe los respectivos coeficientes de la regresión.

Los coeficientes obtenidos en la Regresión Cox se pueden interpretar de la siguiente forma:

- Si $\exp(\beta_j) > 1$: Esto indica que un aumento en el valor de la variable explicativa está asociado con un aumento en la tasa de riesgo relativo.
- Si $\exp(\beta_j) < 1$: Esto indica que un aumento en el valor de la variable explicativa está asociado con una disminución en la tasa de riesgo relativo.
- Si $\exp(\beta_j) = 1$: Esto indica que la variable explicativa no tiene ningún efecto sobre la tasa de riesgo relativo. En otras palabras, la variable no afecta el riesgo de que ocurra el evento.

El modelo de Regresión de Cox estima los coeficientes de regresión a través de un método llamado máxima verosimilitud parcial. Los coeficientes estimados proporcionan información sobre la dirección y la magnitud de la asociación entre las variables explicativas y el riesgo relativo de experimentar el evento de interés.

2.3.1.1. Supuestos

Entre los supuestos que debe cumplir la Regresión de Cox se encuentran (Grambsch and Therneau, 1994):

- Proporcionalidad de riesgos: El efecto de las variables predictoras sobre la función de riesgo instantáneo debe ser constante a lo largo del tiempo.
- Independencia: Los individuos deben ser independientes entre sí, es decir, los eventos que ocurren en un individuo no están relacionados con los eventos que ocurren en otros individuos.
- Linealidad en el logaritmo de la función de riesgo: debe existir una relación lineal entre las variables predictoras y el logaritmo de la función de riesgo.

2.3.1.2. Ventajas y desventajas

Algunas ventajas de la Regresión Cox es que permite incluir múltiples variables explicativas, lo que facilita la exploración de múltiples factores que pueden influir

en el tiempo hasta el evento, además los coeficientes del modelo son interpretables y proporcionan información sobre cómo las variables explicativas afectan la tasa de riesgo relativo. Esto permite identificar factores que aumentan o disminuyen el riesgo de que ocurra el evento.

Algunas desventajas de la Regresión Cox es que como cualquier modelo estadístico puede ser sensible a valores atípicos en los datos, lo que puede influir en la estimación de los coeficientes, el modelo no es capaz de capturar posibles efectos temporales de los datos, además la calidad de los resultados del modelo depende en gran medida de la elección adecuada de las variables explicativas.

2.3.2. Regresión Beta

Cuando se desea hacer una análisis a una variable dependiente que toma valores en el intervalo de unitario estándar (0, 1) se transforman los datos de manera que la respuesta transformada, llamémosle $\hat{y}$, asumiera valores en la línea real y luego aplicar un análisis de regresión lineal estándar. Una transformación comúnmente utilizada es la logit, $\hat{y} = \log\left(\frac{y}{1-y}\right)$.

Sin embargo, ese enfoque tiene deficiencias, primero, los parámetros de la regresión son interpretados en términos de la media de la variable transformada, y no en términos de la media de la variable requerida. Segundo, las regresiones que involucran datos del intervalo unitario suelen ser heterocedásticas y finalmente, las distribuciones de tasas y proporciones suelen ser asimétricas y, por lo tanto, las aproximaciones basadas en Gauss para la estimación de intervalos y la prueba de hipótesis pueden ser bastantes inexactas en muestras pequeñas.

Por esto Ferrari and Cribari-Neto (Ferrari SLP, 2004) propusieron un modelo de regresión para variables continuas que asumen valores en el intervalo unitario estándar (0,1). Donde el modelo se basa en la suposición de que la respuesta tiene una distribución beta, de allí el nombre del modelo de Regresión Beta, en este modelo los parámetros de la regresión son interpretables en términos de la media de la variable de interés, el modelo es heterocedástico por naturaleza y se adapta fácilmente a las asimetrías.

Sea $y_1, \ldots, y_n$ una muestra aleatoria tal que $y_i \sim \mathcal{B}(\mu, \phi), i = 1, \ldots, n$. El modelo

de regresión beta está definido por:

$$g\left(\mu_i\right) = x_i^T \beta = \eta_i,$$

donde $\beta = \left(\beta_1, \ldots, \beta_k\right)^T$ es un vector de $k \times 1$ parámetros de regresión desconocidos $(k < n)$, $x_i = \left(x_{i1}, \ldots, x_{ik}\right)^T$ es el vector de k regresores (o variables independientes) y η_i es un predictor lineal (es decir, $\eta_i = \beta_1 x_{i1} + \ldots + \beta_k x_{ik}$, donde usualmente $x_{i1} = 1$ para todo i, para que el modelo tenga un intercepto.

La interpretación del signo del coeficiente (β_j) en un modelo de regresión beta indica la dirección de la relación entre la variable predictora (x_j) y la variable de respuesta (Y).

- Si β_j es positivo, un aumento en x_j se asocia con un aumento en la probabilidad de que Y esté más cerca de 1.
- Si β_j es negativo, un aumento en x_j se asocia con un aumento en la probabilidad de que Y esté más cerca de 0.

2.3.2.1. Supuestos

Entre los supuestos que debe cumplir la regresión Beta se encuentran (Geissinger, 2022):

- Adecuación del Enlace: Este supuesto se refiere a la elección apropiada de la función de enlace que relaciona el predictor lineal con la media de la variable de respuesta. En la regresión beta, se utilizan funciones de enlace como el enlace logit o el enlace categórico.
- Residuos Homogéneos: Este supuesto se relaciona con la homogeneidad de los residuos en función del predictor lineal. En otras palabras, se espera que la variabilidad de los residuos sea constante en todos los niveles del predictor.
- Normalidad: Este supuesto se refiere a la normalidad de los residuos. Se espera que estos residuos sigan una distribución normal. Aunque la regresión beta no asume directamente la normalidad de los residuos, la verificación de este supuesto es importante para garantizar la validez de las inferencias realizadas con el modelo.

- Ausencia de Valores Atípicos: Este supuesto implica que no debe haber valores atípicos significativos en los datos que afecten de manera desproporcionada el modelo.

2.3.2.2. Ventajas y desventajas

Algunas ventajas de utilizar la Regresión Beta es que a diferencia de la transformación logit utilizada en el análisis de regresión logística, el modelo de regresión beta permite una interpretación directa de los coeficientes en términos de la media de la variable de interés (en el intervalo 0-1). Esto facilita la comprensión de cómo las variables predictoras afectan la variable de , además el modelo de regresión beta es especialmente adecuado cuando la variable de respuesta está restringida al intervalo (0,1), como tasas y proporciones. En lugar de forzar una transformación logit, el modelo trabaja directamente con estos tipos de datos.

Una desventaja de utilizar la Regresión Beta es que al igual que otros modelos, el modelo de Regresión Beta puede ser sensible a valores atípicos en los datos, lo que puede influir en la estimación de los coeficientes.

2.3.3. Modelos Aditivos Generalizados

El Modelo Aditivo Generalizado fue introducido por Trevor Hastie y Robert Tibshirani (Hastie and Tibshirani, 1990). Un modelo aditivo generalizado es un modelo lineal generalizado en el que la variable de respuesta depende linealmente de funciones suaves de algunas variables predictoras.

La característica principal de los Modelos Aditivos Generalizados es que permiten incluir términos suaves no lineales para capturar relaciones complejas entre las variables predictoras y la variable respuesta. Los términos suaves son funciones suaves de las variables predictoras y se pueden ajustar utilizando técnicas de suavizado, como splines o funciones de base penalizadas. Los Modelos Aditivos Generalizados también pueden incorporar términos lineales para variables predictoras de forma similar a los modelos lineales generalizados.

El Modelo Aditivo Generalizado relaciona una variable univariante Y, con algunas variables predictoras, x_i, de la siguiente forma:

$$Y = \beta_0 + f_1\left(x_{1i}\right) + f_2\left(x_{2i}\right) + \ldots + f_n\left(x_{ni}\right) + \epsilon,\ i = 1, \ldots, n\ ,$$

donde Y es la variable respuesta que se desea predecir, $f_j, j = 1, 2, 3, \ldots$ son las funciones suaves de las variables predictoras $x_{1i}, \ldots, x_{ni}$, y β_0 es el intercepto del modelo y ϵ es el término de error.

Los coeficientes en los términos suaves de un modelo aditivo generalizado no se interpretan de la misma manera que los coeficientes en un modelo lineal. En lugar de representar cambios lineales en la variable de respuesta por unidad de cambio en el predictor, representan cómo la contribución del término suave afecta a la predicción. Es decir, un coeficiente positivo o negativo indica si el término suave contribuye positiva o negativamente a la predicción de la variable de respuesta, respectivamente.

2.3.3.1. Supuestos

Entre los supuestos que debe cumplir el modelo aditivo generalizado se encuentran (Augustin et al., 2012), (Wood, 2017):

- Elección adecuada de la dimensión base k: La dimensión de base k establece el número de funciones básicas utilizadas para crear una función suave. Cada función suave en un modelo aditivo generalizado es esencialmente la suma ponderada de muchas funciones más pequeñas, llamadas funciones básicas. Cuantas más funciones básicas se utilicen para construir una función suave, más ondulante será la función suave. Los grados efectivos de libertad o edf son una medida que indica cuántos grados de libertad se están utilizando efectivamente en un término suave para modelar la relación entre una variable predictora y la variable de respuesta. Si el edf está muy cerca de k, esto significa que el movimiento del modelo está demasiado restringido, y por lo tanto el modelo esta sobre-ajustando.
- Distribución de los residuos del modelo: Se debe verificar la suposición de que los residuos se distribuyen de manera aleatoria alrededor de cero y sin mostrar patrones evidentes, esto significa que no muestran ningún patrón discernible en relación con las variables independientes y tienen una media cercana a cero. Además se debe verificar que los residuos se ajustan bien a una distribución normal.

2.3.3.2. Ventajas y desventajas

Algunas ventajas de los Modelos Aditivos Generalizados es que son altamente flexibles y pueden capturar relaciones no lineales y complejas entre las variables predictoras y la variable de respuesta. Esto los hace adecuados para abordar problemas en los que los modelos lineales simples no son suficientes, apesar de su flexibilidad, los Modelos Aditivos Generalizados permiten una interpretación intuitiva de los efectos de las variables predictoras.

Una desventaja de los Modelos Aditivos Generalizados es que pueden ser computacionalmente intensivos, especialmente cuando se ajustan términos suaves con un gran número de funciones básicas. Esto puede requerir más tiempo y recursos computacionales.

2.4. Modelo de series temporales

2.4.1. Modelo Autorregresivo Integrado de Media Móvil

El Modelo Autorregresivo Integrado de Media Móvil o ARIMA es una técnica estadística utilizada para analizar y predecir series de tiempo que fue desarrollado por el estadístico británico George E.P. Box y su colega Gwilym M. Jenkins (Box and Jenkins, 1970). El modelo ARIMA se compone de tres componentes claves: el componente autorregresivo (AR), el componente de integración (I) y el componente de media móvil (MA).

El componente autorregresivo (AR) (Yule, 1927) utiliza la idea de que los valores pasados de la serie de tiempo tienen un efecto sobre los valores futuros. La notación AR(p) representa el componente autorregresivo de orden p, lo que significa que se utilizan p valores anteriores para predecir el siguiente valor en la serie de tiempo. La ecuación AR(p) viene dada por:

$$X_t = c + \phi_1 X_{t-1} + \phi_2 X_{t-2} + \cdots + \phi_p X_{t-p} + \epsilon_t,$$

donde X_t es el valor en el tiempo t, c es una constante, $\phi_1, \ldots, \phi_p$ son coeficientes autorregresivos y ϵ_t es un término de error en el tiempo t.

El componente de integración (I) se refiere al número de diferenciaciones necesarias para hacer que la serie de tiempo sea estacionaria, es decir, que tenga una media y una varianza constantes a lo largo del tiempo. La notación I(d) representa el

componente de integración de orden d, donde d es el número de diferenciaciones requeridas.

El componente de media móvil (MA) (Slutzky, 1927) se basa en la idea de que los errores pasados de predicción afectan los valores futuros de la serie de tiempo. La notación MA(q) representa el componente de media móvil de orden q, lo que significa que se utilizan q términos de error anteriores para predecir el siguiente valor en la serie de tiempo. La ecuación MA(q) viene dada por:

$$X_t = \mu + \theta_1\epsilon_{t-1} + \theta_2\epsilon_{t-2} + \cdots + \theta_q\epsilon_{t-q} + \epsilon_t,$$

donde X_t es el valor en el tiempo t, μ es la media de la serie de tiempo, $\theta_1, \ldots, \theta_q$ son coeficientes de medias móviles y ϵ_t es un término de error en el tiempo t.

Así un proceso estocástico (X_t) es integrado de orden d ($d > 0$ entero) si y sólo si (X_t) sigue un modelo autorregresivo integrado media móvil de orden (p, d, q), o $ARIMA(p, d, q)$, donde la ecuación general del modelo viene dada por:

$$Y_t = c + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \cdots + \phi_p Y_{t-p} - \theta_1\epsilon_{t-1} - \theta_2\epsilon_{t-2} - \cdots - \theta_q\epsilon_{t-q} + \epsilon_t,$$

donde Y_t es el valor en el tiempo t, c es una constante, $\phi_1, \ldots, \phi_p$ son los coeficientes autoregresivos de orden p que capturan la relación lineal entre los valores anteriores de la serie, $\theta_1, \ldots, \theta_q$ son los coeficientes de media móvil de orden q que modelan la relación lineal entre los términos de error anteriores y ϵ_t es un término de error en el tiempo t.

2.4.1.1. ARIMAX

El Modelo Autorregresivo Integrado de Media Móvil con Variables Exógenas o ARIMAX es una extensión del modelo ARIMA que permite incorporar variables exógenas en la predicción de una serie temporal. Las variables exógenas son variables que no son parte de la serie temporal principal, pero se cree que tienen un impacto en la variable que se está pronosticando.

La forma general de un modelo ARIMAX viene dada por:

$$Y_t = c + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \cdots + \phi_p Y_{t-p} - \theta_1 \epsilon_{t-1} - \theta_2 \epsilon_{t-2} - \cdots - \theta_q \epsilon_{t-q} + \beta_1 X_{1,t} + \beta_2 X_{2,t} + \cdots + \beta_k X_{k,t} + \epsilon_t,$$

donde X_t es el valor en el tiempo t, c es una constante, $\phi_1, \ldots, \phi_p$ son los coeficientes autoregresivos de orden p que capturan la relación lineal entre los valores anteriores de la serie, $\theta_1, \ldots, \theta_q$ son los coeficientes de media móvil de orden q que modelan la relación lineal entre los términos de error anteriores, $\beta_1, \ldots, \beta_k$ son los coeficientes de las variables exógenas $X_{1,t}, \ldots, X_{k,t}$ y ϵ_t es un término de error en el tiempo t.

Los modelos ARIMAX son útiles en situaciones donde se cree que factores externos influyen en la serie temporal que se está estudiando. Estos modelos permiten capturar y tener en cuenta estas influencias adicionales en las predicciones, lo que puede mejorar la precisión de los pronósticos en comparación con un modelo ARIMA estándar.

2.4.1.2. Supuestos

El modelo autorregresivo integrado de media móvil o ARIMA debe cumplir los siguientes supuestos (Robert et al., 2006):

- Estacionariedad en la serie: La serie temporal debe ser estacionaria o transformable en estacionaria a través de diferenciación. Esto /significa que las propiedades estadísticas, como la media y la varianza, deben ser constantes a lo largo del tiempo.
- Residuos como ruido blanco: Se espera que los residuos del modelo ARIMA sean ruido blanco, lo que significa que cumplen con las siguientes condiciones: La media de los residuos debe ser cercana a cero, la varianza de los residuos debe ser constante a lo largo del tiempo y los residuos no deben mostrar patrones significativos de autocorrelación.

2.4.2. Modelo Autorregresivo

El Modelo Autorregresivo o AR, es una técnica estadística utilizada para analizar y predecir series de tiempo (Yule, 1927). Este modelo se basa en la idea de que los valores pasados de la serie de tiempo tienen un efecto sobre los valores futuros.

La notación AR(p) significa que se utilizan p valores anteriores para predecir el siguiente valor en la serie de tiempo. La ecuación AR(p) viene dada por:

$$X_t = c + \phi_1 X_{t-1} + \phi_2 X_{t-2} + \cdots + \phi_p X_{t-p} + \epsilon_t,$$

donde X_t es el valor en el tiempo t, c es una constante, $\phi_1, \ldots, \phi_p$ son coeficientes autorregresivos y ϵ_t es un término de error en el tiempo t.

2.4.2.1. ARX

El Modelo Autorregresivo con Variables Exógenas, abreviado como ARX, es una extensión del modelo AR que al igual que el modelo ARIMAX permite incorporar variables exógenas en la predicción de una serie temporal.

La forma general de un modelo ARX se expresa de la siguiente manera:

$$X_t = c + \phi_1 X_{t-1} + \phi_2 X_{t-2} + \cdots + \phi_p X_{t-p} + \beta_1 X_{1,t} + \beta_2 X_{2,t} + \cdots + \beta_k X_{k,t} + \epsilon_t,$$

donde X_t es el valor en el tiempo t, c es una constante, $\phi_1, \phi_2, \ldots, \phi_p$ son los coeficientes autoregresivos, $\beta_1, \beta_2, \ldots, \beta_k$ son los coeficientes de las variables exógenas $X_{1,t}, X_{2,t}, \ldots, X_{k,t}$ y ϵ_t es un término de error en el tiempo t.

2.4.2.2. Supuestos

El modelo autorregresivo o AR debe cumplir los siguientes supuestos (Diversi et al., 2008):

- Estacionariedad en la serie: La serie temporal debe ser estacionaria. Esto significa que las propiedades estadísticas, como la media y la varianza, deben ser constantes a lo largo del tiempo.
- Residuos como ruido blanco: Se espera que los residuos del modelo sean ruido blanco, lo que significa que cumplen con las siguientes condiciones: La media de los residuos debe ser cercana a cero, la varianza de los residuos debe ser constante a lo largo del tiempo y los residuos no deben mostrar patrones significativos de autocorrelación.

2.5. Criterios de evaluación de modelos

2.5.1. Error Cuadrático Medio

La Raíz del Error Cuadrático Medio (RECM) o Root Mean Square Error (RMSE) en inglés es una métrica utilizada para la precisión de modelos de predicción o regresión. El RECM mide la diferencia entre los valores predichos por un modelo y los valores reales. El RECM permite comparar los errores de predicción de diferentes modelos para un conjunto de datos en particular (Hyndman and Koehler, 2006).

La fórmula del RECM es la siguiente:

$$RECM = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2},$$

donde n es el número de observaciones en el conjunto de datos, y_i es el valor real u observado e $\hat{y}_i$ es el valor predicho por el modelo. El RECM es siempre no negativo, mientras más bajo sea su valor es mejor, pues indica un mejor ajuste de los datos.

2.5.2. Coeficiente de Determinación

El coeficiente de determinación (R^2) es una medida del grado de relación existente entre la variable dependiente y las variables independientes. Mide cuánta variabilidad en la variable dependiente es explicada por el modelo de regresión (Steel and Torrie, 1960). El R^2 varía entre 0 y 1, y su fórmula viene dada por:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2},$$

donde y_i es el valor real u observado, $\hat{y}_i$ el valor predicho por el modelo, $\bar{y}_i$ la media de los valores reales y n es el número de observaciones en el conjunto de datos. Mientras más alto sea el valor del R^2 mejor es el ajuste del modelo.

2.5.3. Criterio de Información de Akaike

El criterio de Información de Akaike (AIC) es una métrica utilizada para comparar modelos estadísticos, el AIC toma en cuenta la calidad del ajuste del modelo y la complejidad del modelo (número de parámetros) (Sakamoto et al., 1986). El objetivo es seleccionar el modelo que minimice el AIC. La fórmula del AIC es:

$$AIC = 2log(L) + 2k,$$

donde L es la verosimilitud del modelo (una medida de cuán bien el modelo se ajusta a los datos) y k es el número de parámetros del modelo. Un valor de AIC más bajo indica un mejor equilibrio entre el ajuste y la complejidad del modelo.

3. Ajuste a datos reales

En este capítulo, se llevará a cabo el ajuste a datos reales de los diferentes modelos y algoritmos presentados en el capítulo anterior. Para ello, se iniciará con una breve explicación de la base de datos original. Posteriormente, se mostrarán los resultados asociados con la aplicación tanto del algoritmo Boruta como de la regularización Lasso. A continuación, se realizará un análisis exploratorio de las variables marcadas como importantes por ambos algoritmos de selección de variables. Finalmente, se ajustarán los modelos de regresión y series temporales definidos en el capítulo anterior, describiendo sus resultados y realizando un análisis detallado.

El análisis de datos y el posterior ajuste y validación de modelos se llevaron a cabo en el software R Studio versión 4.2.2.

3.1. Descripción de la base de datos

La base de datos original consta de 84 registros de 228 variables macroeconómicas (84 observaciones para cada variable). Los primeros 48 registros corresponden a las variables macroeconómicas reales, tanto brutas como derivadas, mientras que los siguientes 36 registros restantes corresponden a proyecciones de las 228 variables macroeconómicas, indicando valores estimados para períodos futuros.

Es importante destacar que 6 de las 228 variables corresponden a variables crisis covid, variables que describen el comportamiento de la pandemia covid-19 a través de variables dummys, por ejemplo, 1 si fue un mes afectado por la pandemia covid-19 y 0 en caso contrario. Estas variables fueron creadas para medir el efecto de la pandemia sobre la tasa de incumplimiento.

Por su parte, las variables macroeconómicas brutas utilizadas son:

- Producto Interno Bruto (PIB)
- Tasa de Desempleo (%)
- Índice de Precios al Consumidor Nacional (IPC)
- Tipo de Cambio Promedio del Período (S/ por UM) - Dólar Americano (US$) (Paridad)

- Tasas de Interés del Banco Central de Reserva. Tasa de Referencia de la Política Monetaria (TPM)

A partir de estas variables macroeconómicas brutas, se crearon las variables derivadas (por ejemplo, la variación mensual del producto interno bruto), variables que podrían ser de utilidad para predecir el incumplimiento.

En la tabla 3.1, se muestra un resumen de la distribución de las 228 variables macroeconómicas, estás variables macroeconómicas y sus respectivas proyecciones fueron creadas por la institución financiera.

Variable	**Cantidad de variables**
Producto Interno Bruto	54
Tasa de Desempleo (%)	54
Índice de Precios al Consumidor Nacional	54
Tipo de Cambio Promedio del Período	30
Tasa de Referencia de la Política Monetaria	30
Crisis Covid	6

Cuadro 3.1: Distribución de variables.

(Fuente: Elaboración propia)

Por razones de confidencialidad, los nombres originales de las variables serán cambiados. Los nombres de las variables *Producto Interno Bruto* serán cambiados a PIB, . . . , PIB 54; los nombres de las variables *Tasa de Desempleo* serán cambiados a Desempleo 1, . . . , Desempleo 54; los nombres de las variables *Índice de Precios al Consumidor Nacional* serán cambiados a IPC 1, . . . , IPC 54; los nombres de las variables *Tipo de Cambio Promedio del Período* serán cambiados a Paridad 1, . . . , Paridad 30; los nombres de las variables Tasa de Referencia de la Política Monetaria serán cambiados a TPM 1, . . . , TPM 30; y los nombres de las variables Crisis Covid serán cambiados a crisis 1, . . . , crisis 6.

Además de las 228 variables macroeconómicas, la base de datos original cuenta con una variable objetivo llamada *tasa de incumplimiento*, de la cual se tienen 66 registros. Los primeros 48 registros se utilizan para la construcción de los diferentes modelos, mientras que los 18 registros restantes se utilizan para realizar un backtest de los modelos, con el fin de determinar cuál modelo proyecta mejor la tasa de incumplimiento para diferentes periodos.

La *tasa de incumplimiento* es una variable numérica que varía entre 0 y 1, la cual representa los incumplimientos históricos en cada periodo. Esta medida refleja la proporción de operaciones que han incurrido en incumplimiento con respecto al total de operaciones.

En la figura 3.1, se muestra la tasa de incumplimiento para los 66 registros disponibles.

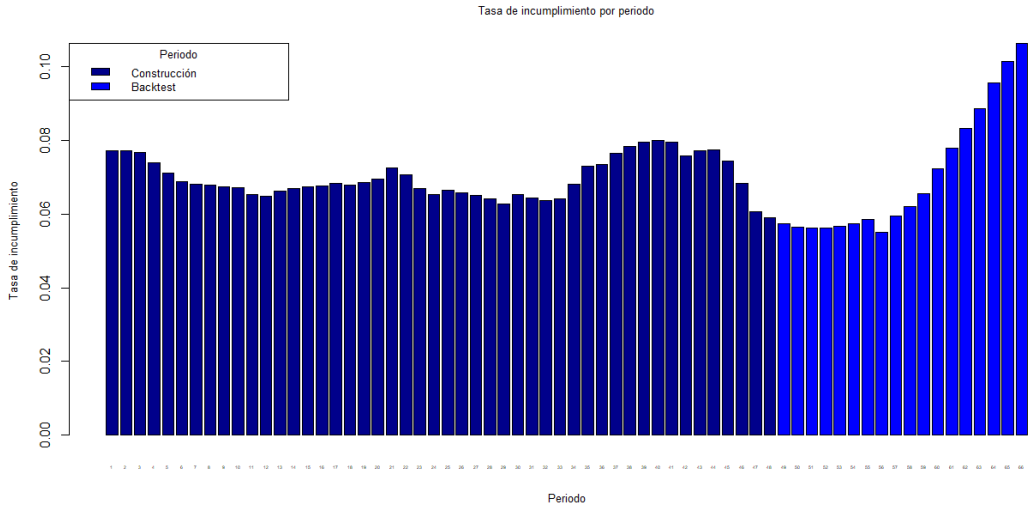

Figura 3.1: Tasa de incumplimiento por periodo.

También la base de datos cuenta con una variable objetivo llamada *Pérdida dado el incumplimiento*, de la cual se tiene 48 registros. Los primeros 24 registros se utilizan para la construcción de los diferentes modelos, mientras que los 24 registros restantes se utilizan para realizar un backtest de los modelos.

La *Pérdida dado el incumplimiento* es una variable numérica que varía entre el 0 al 100 %, y representa la pérdida esperada en caso de que un cliente incumpla con sus obligaciones de pago. Esta medida se calcula como:

$$LGD = 1 - \frac{\text{Valor Recuperable}}{\text{Exposición Total}}$$

donde la exposición total es el monto total en riesgo, como el valor del préstamo o la inversión y el valor recuperable es el valor estimado que se puede recuperar después de un incumplimiento.

En la figura 3.2, se muestra la pérdida dado el incumplimiento para los 48 registros

disponibles.

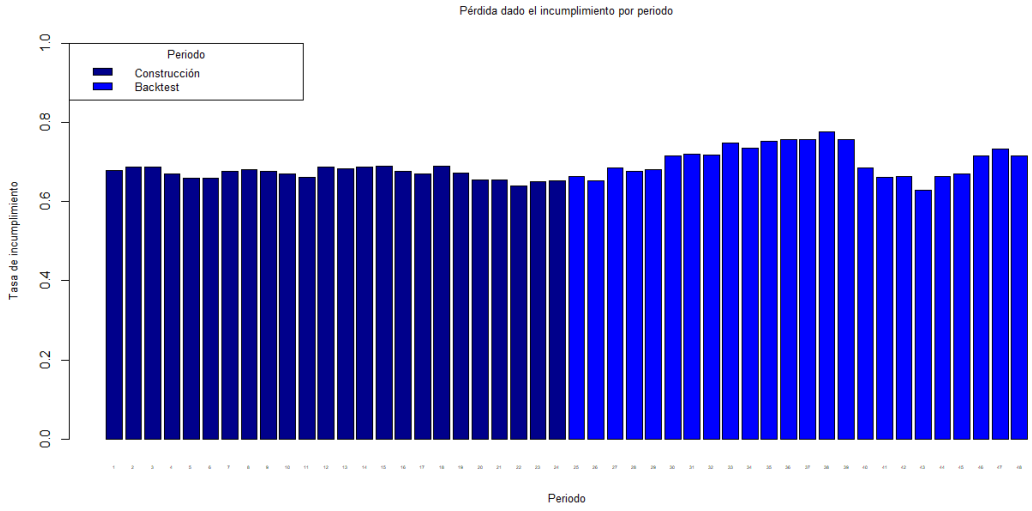

Figura 3.2: Pérdida dado el incumplimiento por periodo.

3.2. Resultados selección de variables

3.2.1. Algoritmo Boruta

Al aplicar Boruta sobre la base de datos original y la variable objetivo tasa de incumplimiento, se obtuvieron 23 variables seleccionadas de las 228 variables macroeconómicas originales. Estas variables se distribuyen de la siguiente manera según las distintas variables macroeconómicas (Figura 3.3).

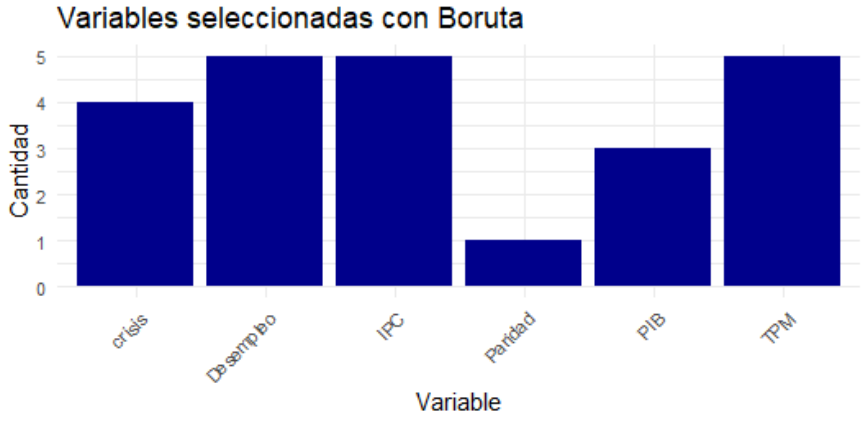

Figura 3.3: Distribución de variables seleccionadas por Boruta para tasa de incumplimiento.

Dado que cada conjunto de variables proviene de la misma variable bruta (por

ejemplo, todas las variables del conjunto Desempleo provienen de la Tasa de Desempleo histórica bruta para cada periodo), se decide seleccionar solo una variable de cada conjunto. De esta manera, dado que Boruta encontró variables importantes para los 6 conjuntos de variables, nos quedaremos con 6 variables macroeconómicas. Estas son: TPM 21, Crisis 4, IPC 36, Desempleo 5, PIB 15 y Paridad 1.

Cabe recalcar que la variable escogida de cada tipo es escogida de la siguiente forma: se ordenan las variables confirmadas que entrega el algoritmo y nos quedamos con la variable más importante de cada tipo, este método de elección será utilizado tanto para el algoritmo Boruta, como la Regularización Lasso.

Posteriormente, al aplicar Boruta sobre la base de datos original y la variable objetivo pérdida dado el incumplimiento, se obtuvieron 5 variables seleccionadas de las 228 variables originales. Las variables se distribuyen de la siguiente manera según las distintas variables macroeconómicas (Figura 3.4).

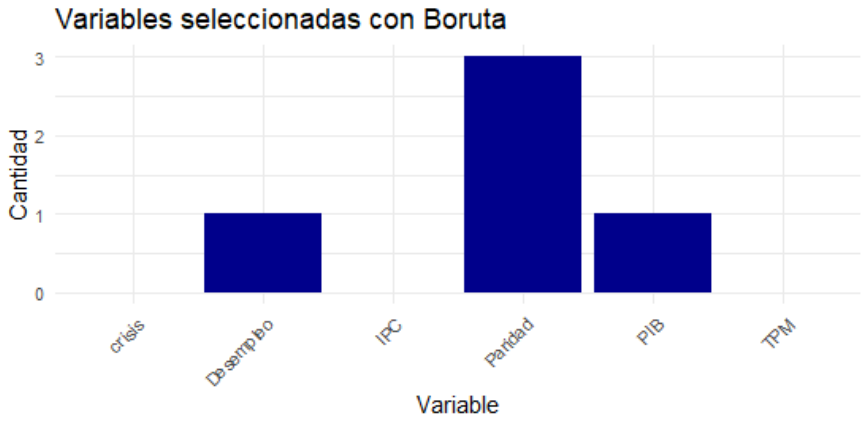

Figura 3.4: Distribución de variables seleccionadas por Boruta para pérdida dado el incumplimiento.

Tal como se explicó, dado que cada conjunto de variables proviene de la misma variable bruta, se decide seleccionar solo una variable de cada conjunto. De esta manera, dado que Boruta encontró variables importantes para 3 conjuntos de variables, nos quedaremos con 3 variables macroeconómicas. Estas son: Paridad 2, Desempleo 29 y PIB 12.

3.2.2. Regularización Lasso

Al aplicar la regularización Lasso sobre la base de datos original y la variable objetivo tasa de incumplimiento, se obtuvieron 31 variables seleccionadas de las 228 variables macroeconómicas originales. Estas variables se distribuyen de la siguiente manera según las distintas variables macroeconómicas (Figura 3.5).

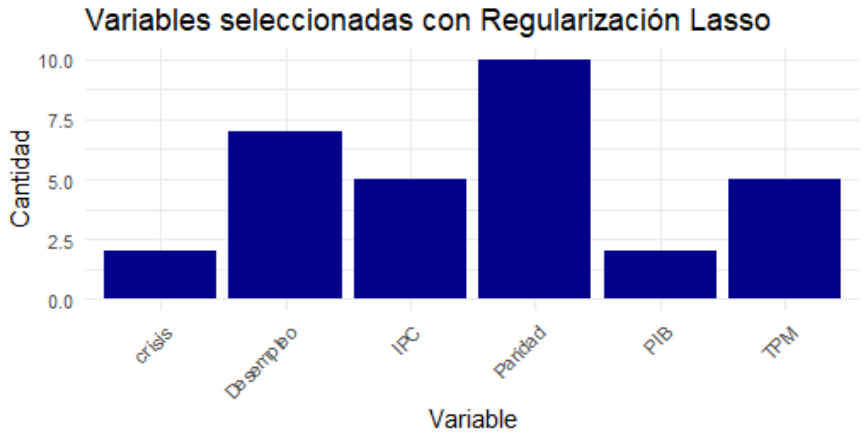

Figura 3.5: Distribución de variables seleccionadas por Regularización Lasso.

Al igual que en Boruta, se decide seleccionar una variable de cada conjunto, dado que la regularización Lasso encontró variables importantes para los 6 conjuntos de variables, nos quedaremos con 6 variables macroeconómicas. Estas son: TPM 10, Crisis 4, IPC 16, Desempleo 11, PIB 19 y Paridad 23.

Luego, al aplicar Regularización Lasso sobre la base de datos original y la variable objetivo pérdida dado el incumplimiento, se obtuvieron 9 variables seleccionadas de las 228 variables originales. Las variables se distribuyen de la siguiente manera según las distintas variables macroeconómicas (Figura 3.6).

Como en el caso anterior, se decide seleccionar solo una variable de cada conjunto. De esta manera, dado que Regularización Lasso encontró variables importantes para 2 conjuntos de variables, nos quedaremos con 2 variables macroeconómicas. Estas son: Paridad 14 y Desempleo 17.

3.3. Análisis exploratorio de los datos

A continuación se analizarán las variables seleccionadas por el algoritmo Boruta y la regularización Lasso, con el fin de identificar patrones, tendencias, valores atípicos y/o influyentes.

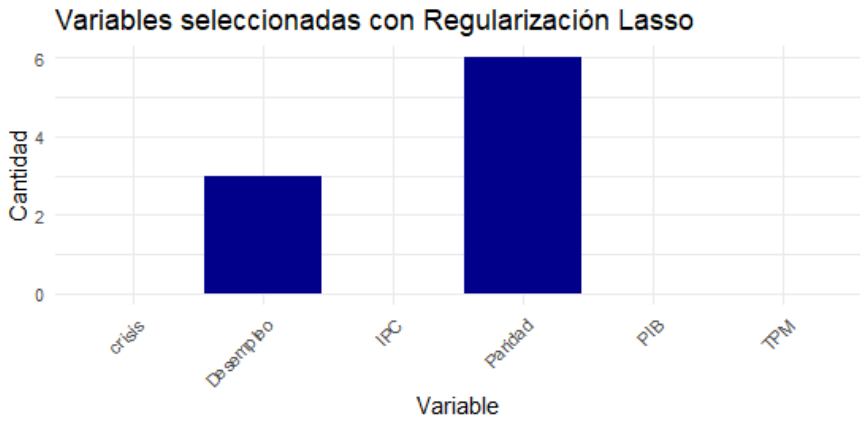

Figura 3.6: Distribución de variables seleccionadas por Regularización Lasso para pérdida dado el incumplimiento.

3.3.1. Variables seleccionadas mediante Boruta

Al aplicar el algoritmo Boruta para la variable objetivo tasa de incumplimiento, se seleccionaron 6 variables macroeconómicas, todas ellas de naturaleza numérica. Estas variables son examinadas visualmente en las Figuras 3.7, 3.8, y 3.9, con el objetivo de identificar tendencias y posibles datos influyentes.

La Figura 3.7 revela que las variables TPM y crisis exhiben una clara tendencia a la baja, mientras que las variables restantes muestran una clara inclinación al alza.

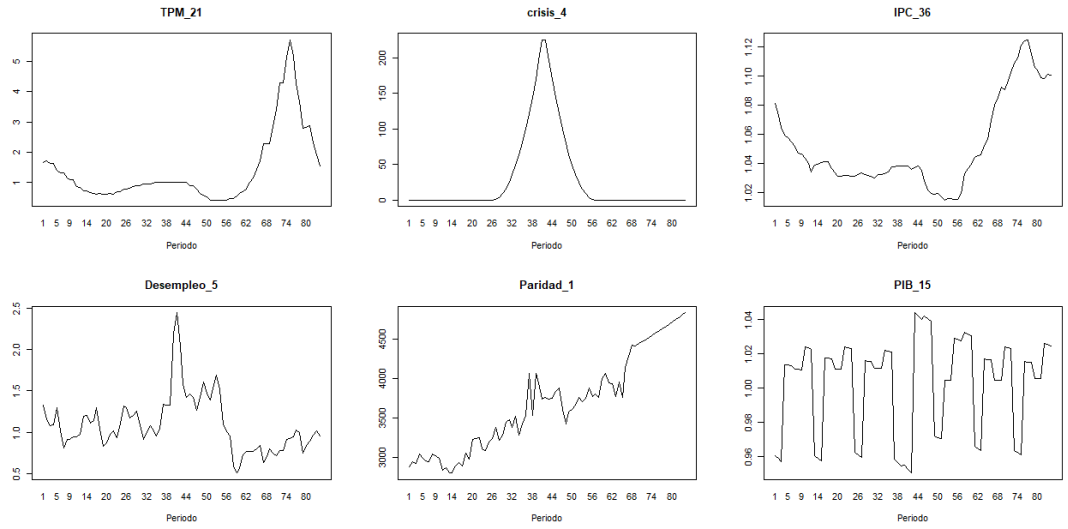

Figura 3.7: Gráficas variables seleccionadas por Boruta para la variable objetivo tasa de incumplimiento.

La Figura 3.8 muestra que las variables TPM, crisis, IPC y Desempleo contienen posibles datos influyentes dado el comportamiento de los histogramas en comparación a su eje x, pues se observa que existen frecuencias alrededor del cero alrededor del eje x.

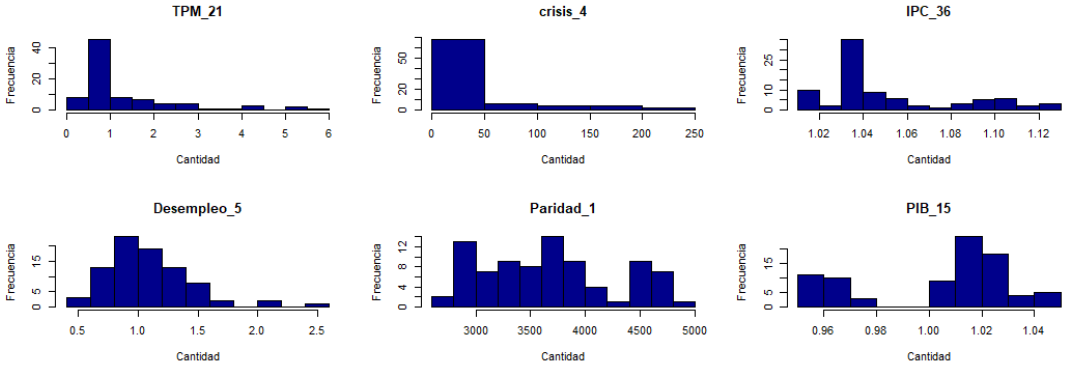

Figura 3.8: Histogramas variables seleccionadas por Boruta para la variable objetivo tasa de incumplimiento.

La Figura 3.9 muestra que las variables TPM, crisis, IPC y Desempleo salen del rango del diagrama de caja, es decir, poseen datos atípicos, datos que podrían influir significativamente en la proyección de la tasa de incumplimiento.

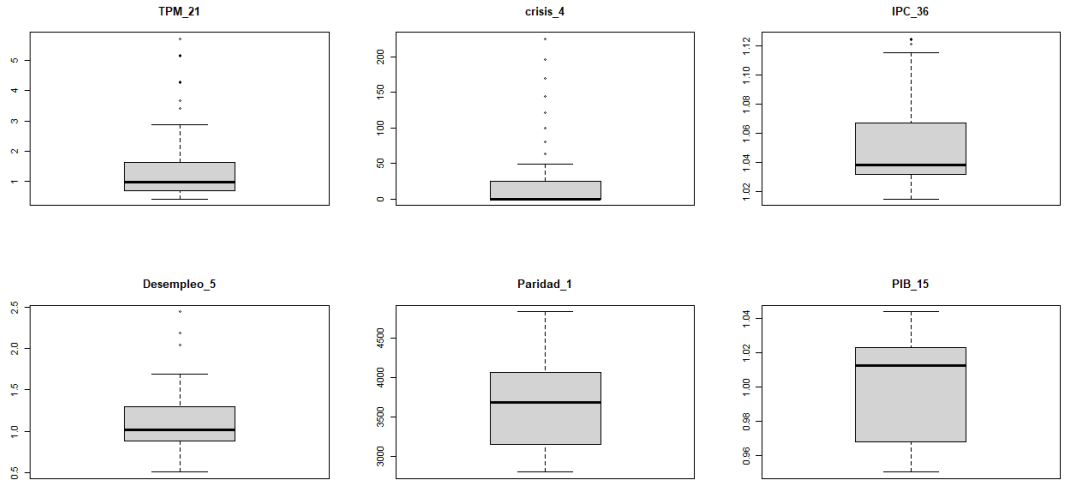

Figura 3.9: Diagramas de caja variables seleccionadas por Boruta para la variable objetivo tasa de incumplimiento.

Por su parte, al aplicar el algoritmo Boruta para la variable objetivo pérdida dado el incumplimiento, se seleccionaron 3 variables macroeconómicas, todas ellas de naturaleza numérica. Estas variables son examinadas visualmente en las Figuras

3.10, 3.11, y 3.12, con el objetivo de identificar tendencias y posibles puntos de datos influyentes.

La Figura 3.10 revela que las variables exhiben una clara tendencia decreciente o a la baja.

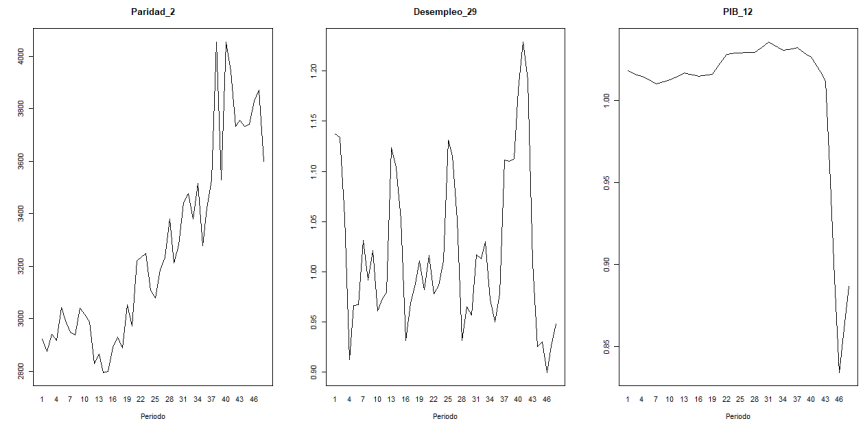

Figura 3.10: Gráficas variables seleccionadas por Boruta para la variable objetivo pérdida dado el incumplimiento.

De la Figura 3.11 se observa que la variable PIB contiene posibles datos influyentes dado el comportamiento de los histogramas en comparación a su eje x, pues se observa que existen frecuencias alrededor del cero alrededor del eje x.

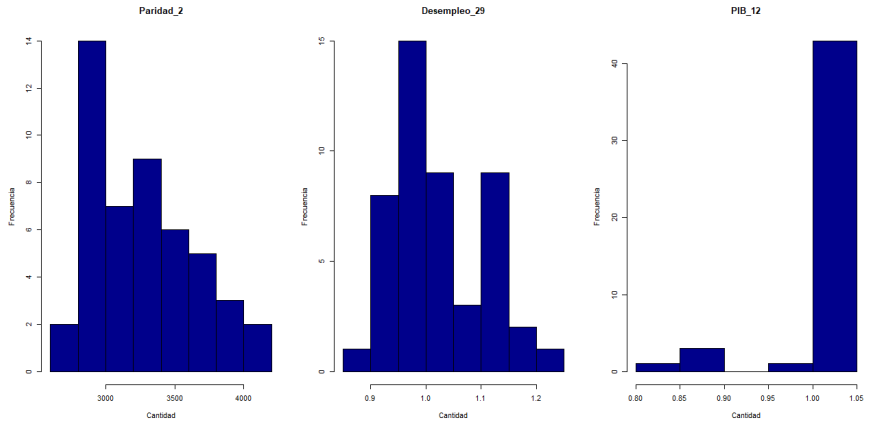

Figura 3.11: Histogramas variables seleccionadas por Boruta para la variable objetivo pérdida dado el incumplimiento.

Por su parte, en la Figura 3.12 se muestra que la variable PIB sale del rango del diagrama de caja, es decir, posee datos atípicos, datos que podrían influir significativamente en la proyección de la pérdida dado el incumplimiento.

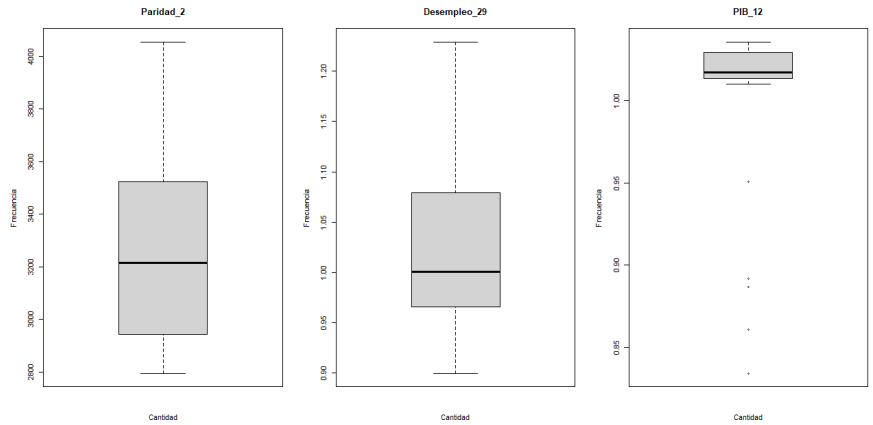

Figura 3.12: Diagramas de caja variables seleccionadas por Boruta para la variable objetivo pérdida dado el incumplimiento.

3.3.2. Variables Seleccionadas mediante Regularización Lasso

Al aplicar la técnica de regularización Lasso a la variable objetivo tasa de incumplimiento, al igual que con Boruta, se opta por seleccionar 6 variables numéricas. Estas variables se analizarán visualmente en las Figuras 3.13, 3.14 y 3.15 con el objetivo de identificar tendencias y posibles datos influyentes.

La Figura 3.13 revela que las variables TPM, crisis y Desempleo muestran una tendencia a la baja, mientras que las demás variables presentan una tendencia al aumento.

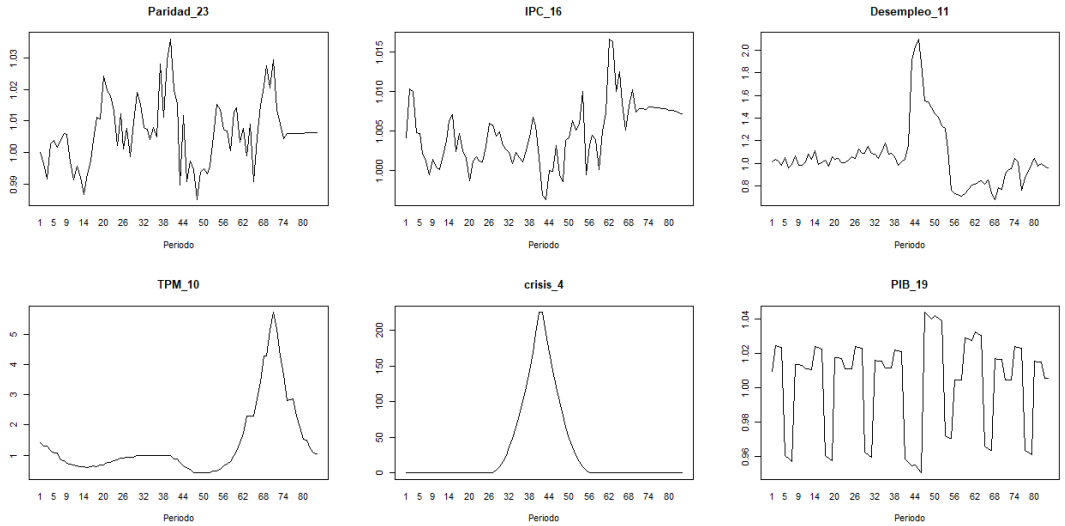

Figura 3.13: Gráficos de las variables seleccionadas mediante Regularización Lasso para la variable objetivo tasa de incumplimiento.

.

En la Figura 3.14, se observa que las variables TPM, crisis y Desempleo podrían contener datos influyentes, dado el comportamiento de los histogramas en comparación con su eje x. Se evidencian frecuencias alrededor del cero, indicando posibles valores atípicos.

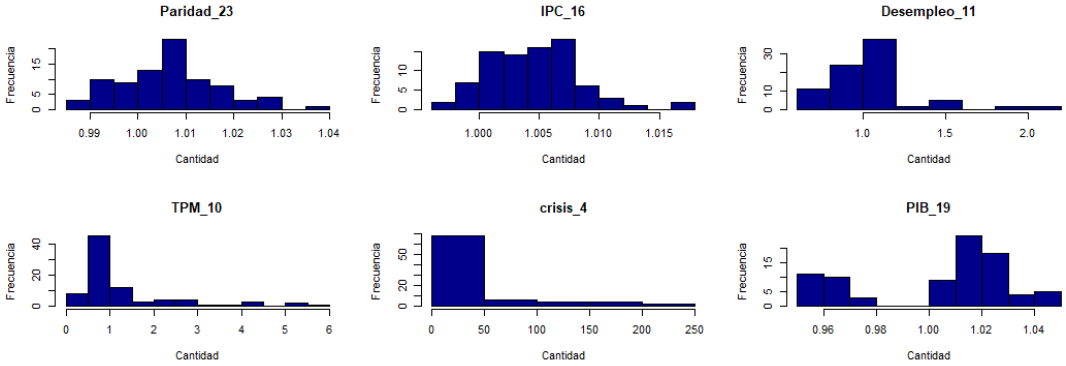

Figura 3.14: Histogramas de las variables seleccionadas mediante Regularización Lasso para la variable objetivo tasa de incumplimiento.

La Figura 3.15 destaca que las variables TPM, crisis y Desempleo se encuentran fuera del rango del diagrama de caja, lo que sugiere la presencia de datos atípicos. Estos valores atípicos podrían tener una influencia significativa en la proyección de la tasa de incumplimiento.

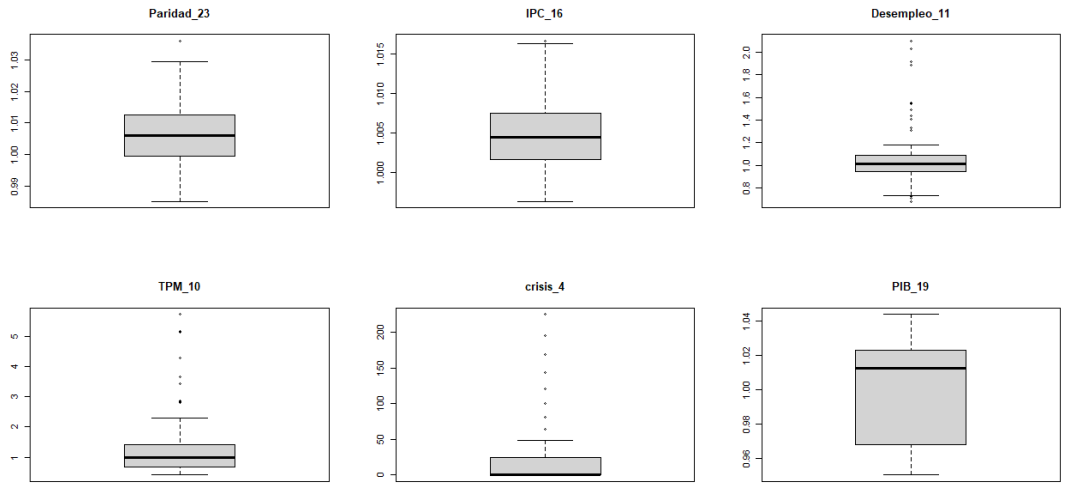

Figura 3.15: Diagramas de caja de las variables seleccionadas mediante Regularización Lasso para la variable objetivo tasa de incumplimiento.

Por otra parte, al aplicar la Regularización Lasso a la variable objetivo pérdida dado el incumplimiento, seleccionamos 2 variables macroeconómicas, todas ellas

de naturaleza numérica. Las variables serán visualizadas en las Figuras 3.16, 3.17 y 3.18 con el objetivo de identificar tendencias y posibles datos influyentes.

La Figura 3.16 revela que las variables muestran claras tendencias al alza y la baja.

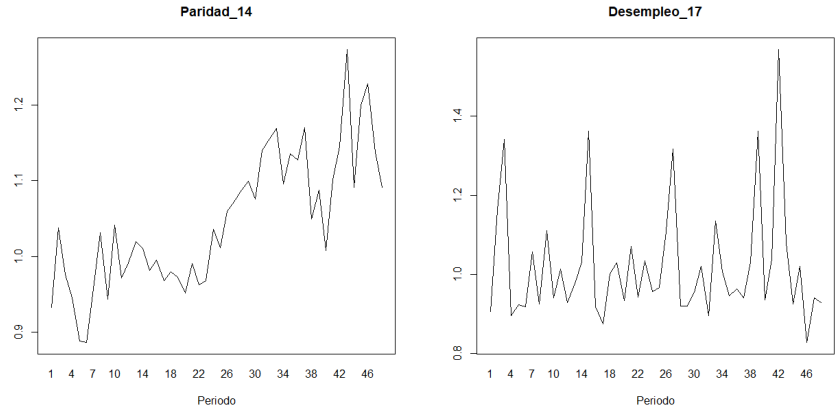

Figura 3.16: Gráficos de las variables seleccionadas mediante Regularización Lasso para la variable objetivo pérdida dado el incumplimiento.

.

En la Figura 3.17, se observa que la variables Desempleo podría contener datos influyentes, dado el comportamiento de los histogramas en comparación con su eje x. Se evidencian frecuencias alrededor del cero, indicando posibles valores atípicos.

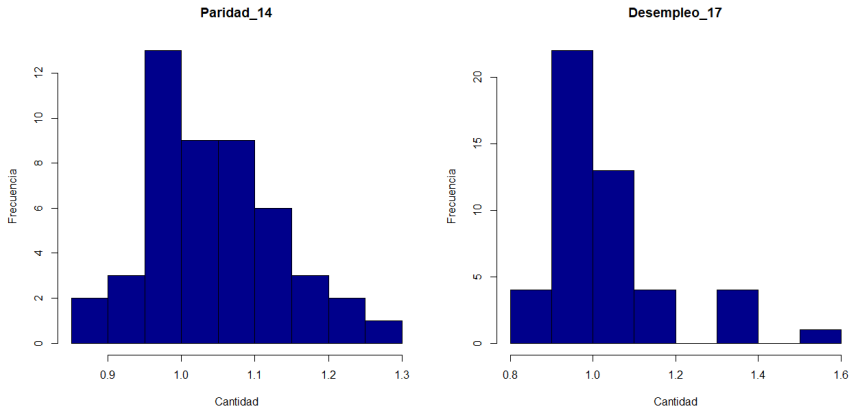

Figura 3.17: Histogramas de las variables seleccionadas mediante Regularización Lasso para la variable objetivo pérdida dado el incumplimiento.

Finalmente, de la Figura 3.18 se destaca que las variable Desempleo se encuentran fuera del rango del diagrama de caja, lo que sugiere la presencia de datos atípicos.

Estos valores atípicos podrían tener una influencia significativa en la proyección de la pérdida dado el incumplimiento.

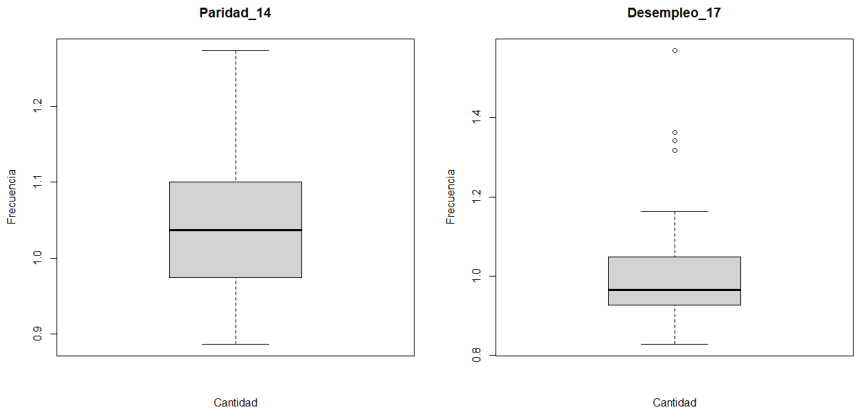

Figura 3.18: Diagramas de caja de las variables seleccionadas mediante Regularización Lasso para la variable objetivo pérdida dado el incumplimiento.

3.4. Modelos de regresión

Al aplicar y seleccionar los modelos de regresión (salvo para el modelo aditivo generalizado), se analizará el sentido negocio, asegurándose de que la relación entre la variable objetivo, en este caso, la tasa de incumplimiento y/o la pérdida dado el incumplimiento, y las variables predictoras tenga sentido coherente.

Se espera que un aumento en el producto interno bruto conduzca a una disminución en la tasa de incumplimiento y/o la pérdida dado el incumplimiento; por el contrario, se anticipa que ante un incremento en el desempleo, el índice de precios al consumidor, el tipo de cambio y la tasa de referencia de la política monetaria la tasa de incumplimiento y/o la pérdida dado el incumplimiento aumente.

3.4.1. Regresión Cox

Los estimadores de la exponencial de los coeficientes β obtenidos durante la construcción del modelo de Regresión de Cox para la variable objetivo probabilidad de incumplimiento, junto con sus valores p asociados, se presentan en la Tabla 3.2 para las variables seleccionadas mediante los métodos de Boruta y Regularización Lasso.

Variables Boruta			**Variables Regularización Lasso**		
Variable	**exp(Coef.)**	**Valor p**	**Variable**	**exp(Coef.)**	**Valor p**
TPM 21	$5{,}115 \times 10^{2}$	0.0001	Paridad 23	$2{,}795 \times 10^{-21}$	0.008
crisis 4	0.975	0.0003	IPC 16	$1{,}802 \times 10^{-69}$	0.017
IPC 36	$3{,}831 \times 10^{-127}$	$9{,}82 \times 10^{-7}$	Desempleo 11	0.496	0.455
Desempleo 5	6.074	0.040	TPM 10	0.047	0.002
Paridad 1	0.997	0.016	crisis 4	0.9859	0.0009
PIB 15	$9{,}324 \times 10^{5}$	0.061	PIB 19	61.82	0.543

Cuadro 3.2: Coeficientes del modelo de Regresión Cox para la variable objetivo probabilidad de incumplimiento.

De la Tabla 3.2, se observa que el modelo con las variables Boruta no tienen sentido negocio. Las variables crisis 4, IPC 36, Paridad 1 y PIB 15 poseen coeficientes diferentes a los esperados según la teoría del modelo y la definición de sentido de negocio. Se esperaría que la exponencial del coeficiente fuera mayor a 1 para todas las variables, excepto la variable PIB. Por lo tanto, este modelo se descarta.

Por su parte, la Tabla 3.2 muestra que el modelo con las variables de regularización

Lasso tampoco cuenta con sentido negocio. Se observa que la exponencial de los coeficientes de todas las variables es contraria a lo esperado.

Dado que ningún modelo cumple con lo esperado, se probó una nueva combinación de variables, se probó con una combinación de variables IPC, Paridad e IPC, la cual se presenta en la Tabla 3.3.

Variable	**exp(Coef.)**	**Valor p**
IPC 34	$1{,}293 \times 10^{60}$	0.022
Paridad 27	$3{,}216 \times 10^{01}$	0.069
PIB 7	$9{,}999 \times 10^{-01}$	0.0001

Cuadro 3.3: Coeficientes del modelo de Regresión Cox.

Desde la Tabla 3.3, se observa que la exponencial del coeficiente de la variable PIB es menor que 1, y para las demás variables, la exponencial del coeficiente es mayor a 1, es decir, este modelo cuenta con sentido negocio. Por lo tanto, este es el posible modelo escogido, nos queda probar que el modelo es adecuado.

Para confirmar que el modelo es adecuado, debemos asegurarnos de que cumpla con los supuestos de la Regresión Cox, es decir, se deben cumplir los supuestos de proporcionalidad de riesgos y linealidad en el logaritmo de la función de riesgo.

Variable	**chisq**	**df**	**p**
IPC 34	0.300	1	0.58
Paridad 27	0.418	1	0.52
PIB 7	0.432	1	0.51
GLOBAL	0.988	3	0.80

Cuadro 3.4: Supuesto de proporcionalidad de riesgo.

Según la Tabla 3.4, la prueba no es estadísticamente significativa para cada una de las covariables, pues se observa que todos los valores p son mayores a 0.05, y la prueba global tampoco es estadísticamente significativa. Por lo tanto, se cumple el supuesto de riesgos proporcionales.

Desde la Figura 3.19, se observa que las líneas trazadas en negro son casi lineales, lo que significa que si trazamos líneas rectas a través de las curvas, esta debería seguir una tendencia general similar a la curva. Por lo tanto, se cumple la linealidad de la función de riesgo.

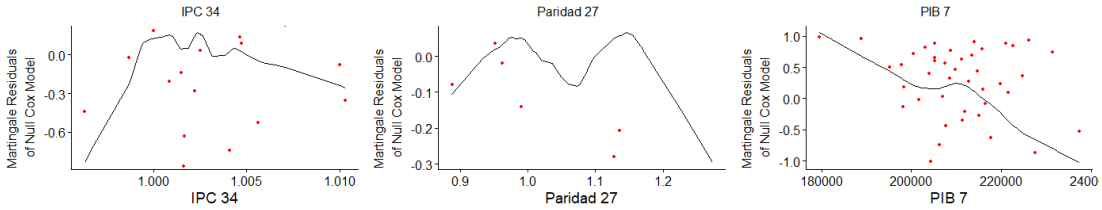

Figura 3.19: Supuesto de linealidad en el logaritmo de la función de riesgo.

Así, el modelo cumple con todos los supuestos esperados de la Regresión de Cox. Con esto, es posible crear proyecciones para la tasa de incumplimiento. En la Figura 3.20, se observa un gráfico del ajuste (antes de la línea roja punteada) y las proyecciones (después de la línea roja punteada) entregadas por el modelo de Regresión de Cox, junto con las tasas de incumplimiento reales.

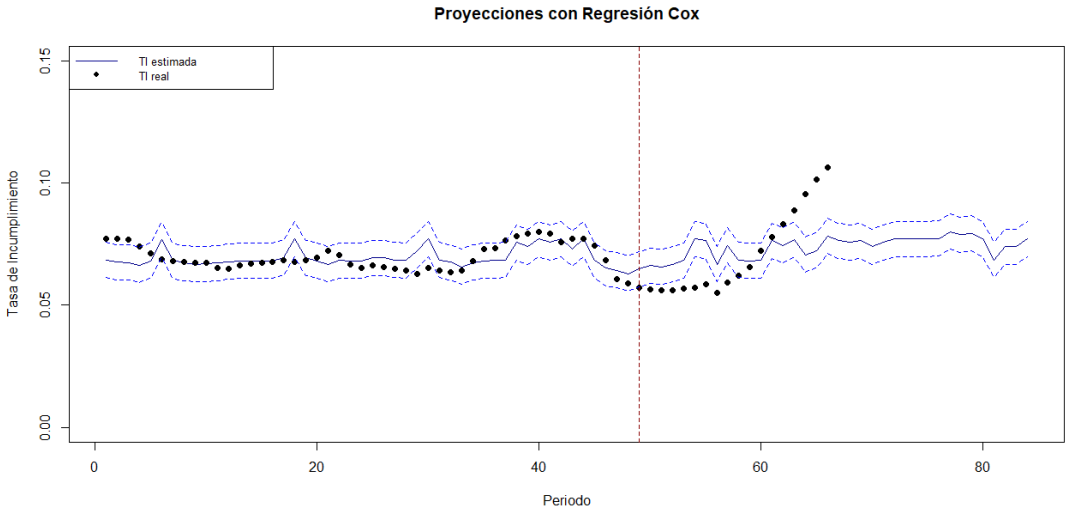

Figura 3.20: Ajuste y predicciones con el modelo de regresión de Cox.

De la Figura 3.20, se observa a priori que el modelo tiene un mal ajuste para los periodos de construcción y una mala predicción para los periodos de backtest. Por esta razón, este modelo no será utilizado para crear proyecciones de la pérdida dado el incumplimiento.

3.4.2. Regresión Beta

Los estimadores de los coeficientes β obtenidos durante la construcción del modelo de Regresión Beta para la variable objetivo probabilidad de incumplimiento, junto con sus valores p asociados con las variables seleccionadas por Boruta y

Regularización Lasso, se presentan en la Tabla 3.5.

Variables Boruta			**Variables Regularización Lasso**		
Variable	**Coef.**	**Valor p**	**Variable**	**Coef.**	**Valor p**
TPM 21	$-1{,}624 \times 10^{-1}$	0.000	Paridad 23	0,575	0.504
crisis 4	$8{,}76 \times 10^{-4}$	$5{,}18 \times 10^{-6}$	IPC 16	5,83	0.071
IPC 36	8,53	$2{,}37 \times 10^{-12}$	Desempleo 11	0,00271	0.950
Desempleo 5	$-0{,}0664$	0.0211	TPM 10	0,141	0.00116
Paridad 1	$8{,}04 \times 10^{-5}$	0.0195	crisis 4	0,000702	$2{,}72 \times 10^{-6}$
PIB 15	$-0{,}367$	0.068	PIB 19	$-0{,}226$	0.435

Cuadro 3.5: Coeficientes del modelo de Regresión Beta para la variable objetivo probabilidad de incumplimiento.

En la Tabla 3.5, se observa que las variables TPM 21 y Desempleo 5 (variables regularización lasso) cuentan con coeficientes negativos, lo que no tiene sentido negocio, pues se espera, por ejemplo, que a mayor desempleo aumente la probabilidad de incumplimiento, es decir, que la probabilidad de incumplimiento esté más cercana a 1. Por lo tanto, el modelo de Regresión Beta con las variables marcadas como importantes por el algoritmo Boruta queda descartado.

De la Tabla 3.5, se observa que la variable PIB cuenta con coeficiente negativo y las demás variables cuentan con coeficiente positivo, teniendo esto sentido negocio. Por lo tanto, este es el posible modelo escogido.

Para confirmar que el modelo de Regresión Beta con variables de la regularización Lasso es adecuado, debemos asegurarnos de que cumpla con los supuestos de la Regresión Beta, es decir, el enlace debe ser adecuado, los residuos deben ser homogéneos y deben distribuir normal, además debe existir ausencia de valores atípicos.

De la Figura 3.21, en el gráfico de la esquina superior izquierda, se verifica si el enlace logit es apropiado para el modelo. Como los residuos se dispersan de manera uniforme alrededor de cero en el gráfico, se sugiere que el enlace logit es apropiado. En la misma Figura, en el gráfico de la esquina superior derecha, se verifica que los residuos sean homogéneos. Se observa que los residuos están razonablemente distribuidos de manera homogénea alrededor de cero con respecto a las predicciones del modelo, lo que sugiere homogeneidad.

De la Figura 3.21, en el gráfico de la esquina inferior izquierda, se verifica

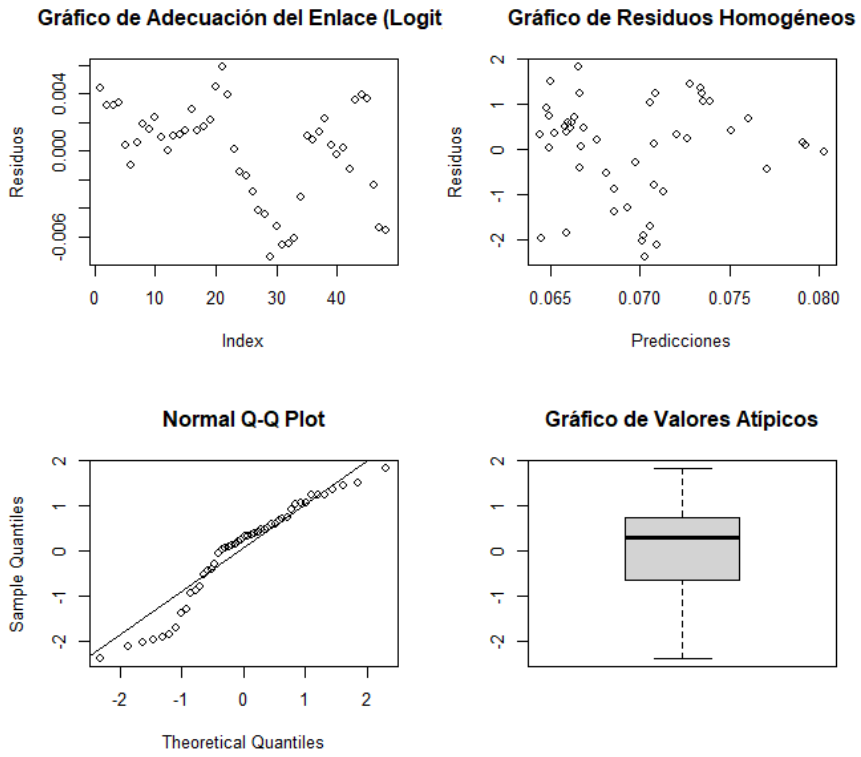

Figura 3.21: Gráficos supuestos del modelo de regresión beta.

el supuesto de normalidad de los residuos. Si los puntos en el gráfico se ajustan aproximadamente a la línea diagonal, indica que los residuos siguen una distribución normal. En este caso, los residuos parecen seguir una distribución aproximadamente normal, ya que se ajustan a la línea diagonal. Finalmente, en la misma Figura, en el gráfico de la esquina inferior derecha, se verifica que no existan valores atípicos en los datos. En este caso, no parece haber valores atípicos notables, ya que no se observan puntos fuera de los límites del boxplot.

Así, el modelo con las variables seleccionadas con la regularización Lasso cumple todos los supuestos esperados de la Regresión Beta. Con esto, es posible crear proyecciones para la tasa de incumplimiento. En la Figura 3.22, se observa un gráfico del ajuste (antes de la línea roja punteada) y las proyecciones (después de la línea roja punteada) del modelo de regresión beta, junto con las tasas de incumplimientos reales.

De la Figura 3.22, se observa a priori que el modelo tiene un mal ajuste para los periodos de construcción, en comparación al ajuste presentado por otros modelos y que se verá en mayor detalle en el capitulo de comparación de modelos. Por lo tanto,

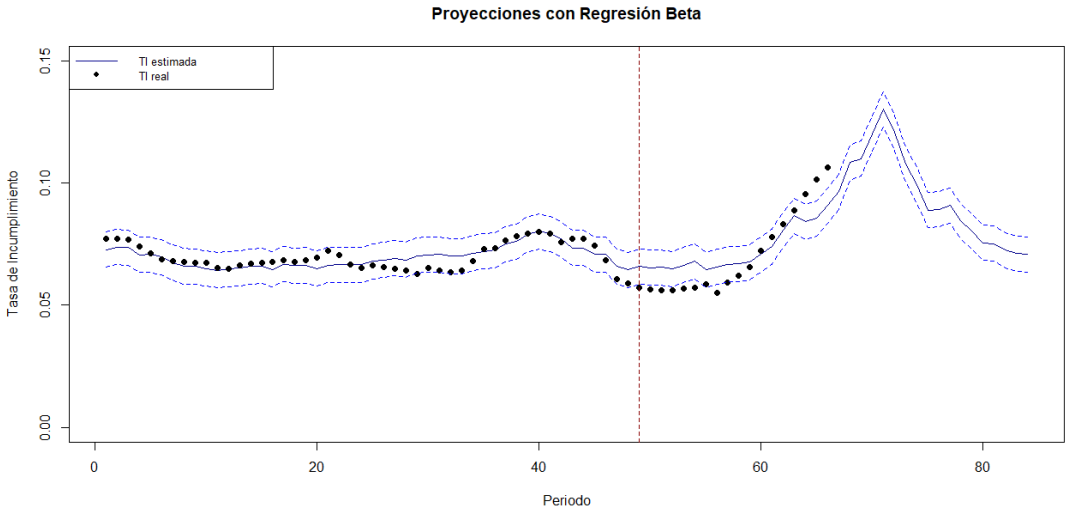

Figura 3.22: Ajuste y predicciones con el modelo de regresión beta.

este modelo no será utilizado para proyectar la pérdida dado el incumplimiento.

3.4.3. Modelo Aditivo Generalizado

Los grados de libertad utilizados (o edf) obtenidos al construir el modelo aditivo generalizado para la variable objetivo probabilidad de incumplimiento (utilizando la función suave spline cúbica penalizada). y sus valores p asociados con las variables seleccionadas por Boruta y regularización Lasso se presentan en la Tabla 3.6.

Variables Boruta			**Variables Regularización Lasso**		
Variable	**Edf**	**Valor p**	**Variable**	**Edf**	**Valor p**
TPM 21	4.836	0.000591	Paridad 23	5.242	0.00138
crisis 4	5.736	1.31e-05	IPC 16	1.000	0.04933
IPC 36	5.477	$6{,}05 \times 10^{-7}$	Desempleo 11	7.138	$< 2 \times 10^{-16}$
Desempleo 5	2.197	0.066308	TPM 10	1.000	0.41082
Paridad 1	4.300	0.025773	crisis 4	7.754	$< 2 \times 10^{-16}$

Cuadro 3.6: Edf del modelo aditivo generalizado para la variable objetivo probabilidad de incumplimiento.

Dado que, a diferencia de otros modelos de regresión, acá no existen coeficientes β estimados, para escoger entre ambos modelos se utilizará el coeficiente de determinación o R^2.

En la Tabla 3.7 se muestra el coeficiente de determinación para ambos modelos, se

Modelo	$\mathbf{R^2}$
Modelo Boruta	0.978
Modelo Lasso	0.982

Cuadro 3.7: Coeficientes de determinación para modelos aditivos generalizado para la variable objetivo probabilidad de incumplimientos.

observa que el modelo aditivo generalizado ajustado con las variables seleccionadas por Regularización Lasso posee un mejor ajuste, pues su valor es más cercano a 1 que el coeficiente de determinación del otro modelo. Es por esto que este es el posible modelo escogido.

Para confirmar que el modelo aditivo generalizado con variables de la regularización Lasso es adecuado debemos asegurarnos que cumpla con los supuestos del modelo aditivo generalizado, es decir, que la elección de la base k' sea adecuada y que los residuos distribuyan normal.

Recordemos que la base k' hace referencia al número de funciones de suavizado o funciones base que se utilizan para modelar la relación no lineal entre las variables predictoras y la variable de respuesta. La elección de $k´$ es crucial porque determina la flexibilidad del modelo y, por lo tanto, puede afectar la capacidad del modelo para capturar patrones complejos en los datos.

Variable	k'	Edf	k-index	Valor p
s(Paridad 23)	9.000	5.242	1.020	0.500
s(IPC 16)	9.000	1.000	0.971	0.350
s(TPM 10)	9.000	7.138	1.342	0.973
s(Desempleo 11)	9.000	1.000	0.989	0.408
s(crisis 4)	9.000	7.754	0.994	0.418

Cuadro 3.8: Elección adecuada de la dimensión base k para el modelo aditivo generalizado para la variable objetivo probabilidad de incumplimiento.

De la Tabla 3.8 se observa que los edf son mucho más pequeños que k',es decir, dado que los edf son mucho más pequeños que k' puede interpretarse como que el modelo está evitando el sobreajuste, así la elección de la base k' es adecuada.

En la Figura 3.23, en la parte superior izquierda hay un gráfico QQ, que compara los residuos del modelo con una distribución normal, dado que los residuos se encuentran cerca de la linea recta el modelo ajusta bien. Vemos también que el

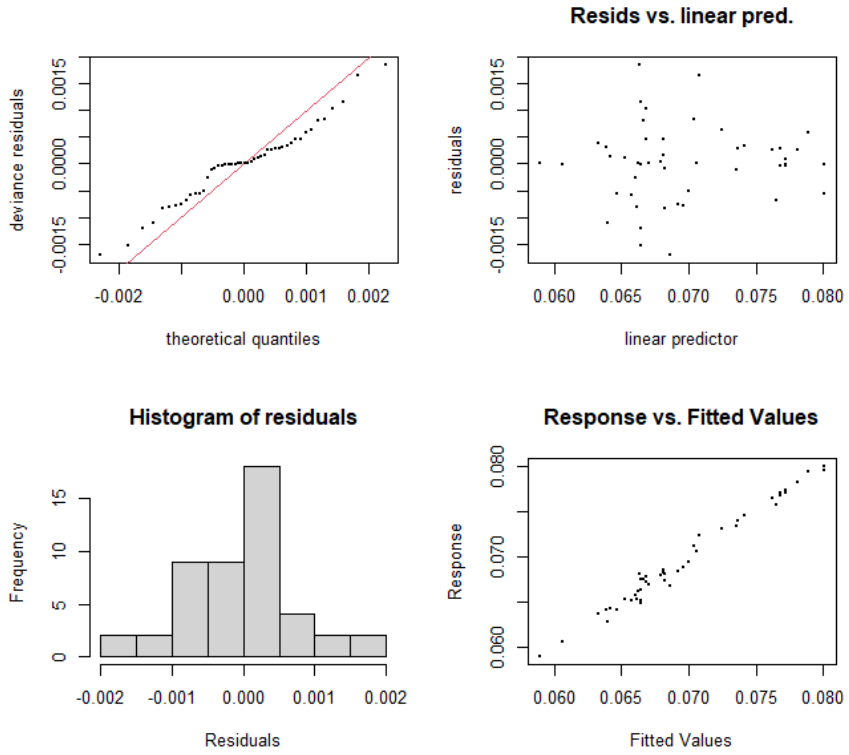

Figura 3.23: Supuesto distribución de los residuos para el modelo aditivo generalizado para la variable objetivo probabilidad de incumplimiento.

histograma en la esquina inferior izquierda tiene forma de campana tal como se esperaría. En la misma Figura, en la parte superior derecha hay un gráfico de valores residuales los cuales se distribuyen uniformemente alrededor de cero, es decir, los residuos distribuyen normal.

Finalmente, en la parte inferior derecha está el gráfico de respuesta frente a los valores ajustados. En un modelo perfecto formaría una línea recta. De la gráfica podemos observar que el modelo forma una línea muy cercana a una recta. Todo esto indica un buen ajuste del modelo, es decir, confirmamos que los residuos efectivamente distribuyen normal.

Por lo tanto, el modelo con las variables seleccionadas con la Regularización Lasso cumple todos los supuestos del modelo aditivo generalizado. Con esto es posible crear proyecciones para la tasa de incumplimiento. En la Figura 3.24 se observa un gráfico del ajuste (antes de la linea roja punteada) y las proyecciones (después de la linea roja punteada) del modelo aditivo generalizado, junto con las tasas de incumplimientos reales.

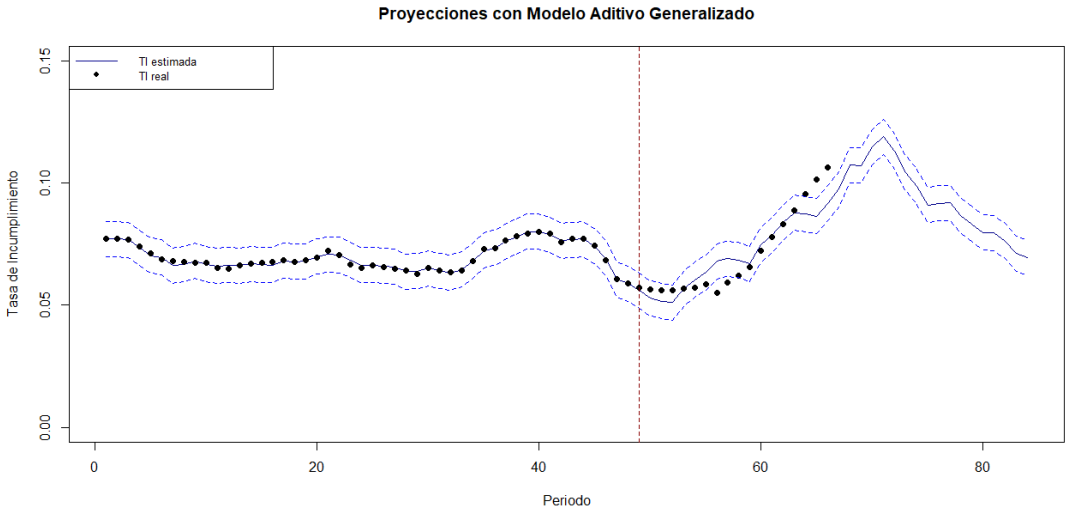

Figura 3.24: Ajuste y predicciones con el modelo aditivo generalizado para la variable objetivo probabilidad de incumplimiento.

De la Figura 3.24 se observa que el modelo cuenta con buen ajuste para los periodos de construcción y para los periodos de backtest, por lo tanto este modelo será utilizado para proyectar la pérdida dado el incumplimiento.

Los grados de libertad (edf) obtenidos al construir el modelo aditivo generalizado para la variable objetivo pérdida dado el incumplimiento (utilizando la función suave spline cúbica penalizada), junto con sus valores p asociados con las variables seleccionadas por Boruta y regularización Lasso se presentan en la Tabla 3.9.

Variables Boruta			**Variables Regularización Lasso**		
Variable	**Edf**	**Valor p**	**Variable**	**Edf**	**Valor p**
Paridad 2	8.708	$2{,}50 \times 10^{-5}$	Paridad 14	3.936	0.396
Desempleo 29	3.284	0.00337	Desempleo 17	1.000	0.080

Cuadro 3.9: Edf del modelo aditivo generalizado para la variable objetivo pérdida dado el incumplimiento.

Como se ha mencionado, para escoger entre ambos modelos se utilizará el coeficiente de determinación o R^2.

En la Tabla 3.10 se muestra el coeficiente de determinación para ambos modelos, se observa que el modelo aditivo generalizado ajustado con las variables seleccionadas por Boruta posee un mejor ajuste, pues su valor es más cercano a 1 que el coeficiente de determinación del otro modelo. Es por esto que este es el posible

Modelo	R^2
Modelo Boruta	0.957
Modelo Lasso	0.394

Cuadro 3.10: Coeficientes de determinación para modelos aditivos generalizados para la variable objetivo pérdida dado el incumplimiento.

modelo escogido.

Para confirmar que el modelo aditivo generalizado con variables Boruta es adecuado debemos asegurarnos que cumpla con los supuestos del modelo aditivo generalizado, es decir, que la elección de la base k sea adecuada y los residuos distribuyan normal.

Variable	k'	edf	k-index	Valor p
s(Paridad 2)	9	8.708	1.255	0.863
s(Desempleo 29)	9	3.284	1.406	0.965

Cuadro 3.11: Elección adecuada de la dimensión base k para el modelo aditivo generalizado para la variable objetivo pérdida dado el incumplimiento.

De la Tabla 3.11 se observa que los edf son mucho más pequeños que k',es decir, dado que los edf son mucho más pequeños que k' puede interpretarse como que el modelo está evitando el sobreajuste y por lo tanto, la elección de la base k fue adecuada.

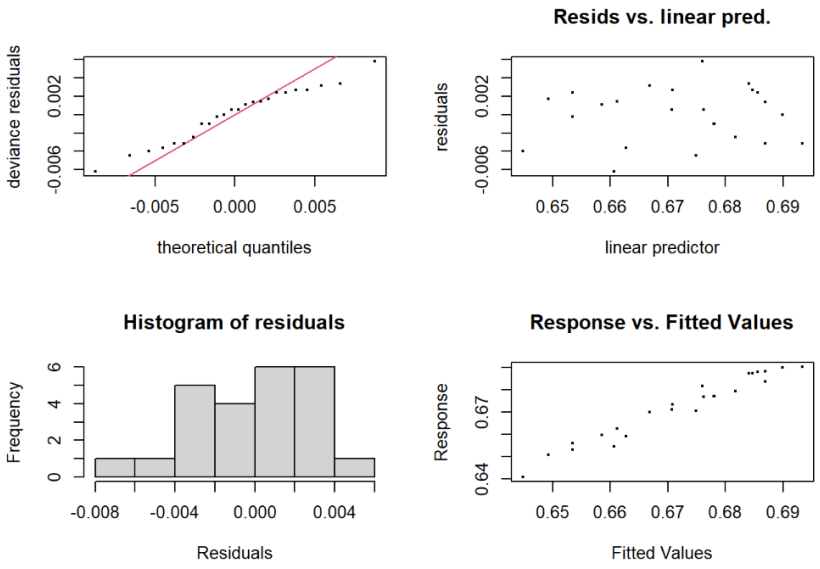

Figura 3.25: Supuesto distribución de los residuos para el modelo aditivo generalizado.

En la Figura 3.25, en la parte superior izquierda hay un gráfico QQ, que compara los residuos del modelo con una distribución normal, dado que los residuos se encuentran cerca de la linea recta el modelo ajusta bien. El histograma en la esquina inferior izquierda de la misma Figura tiene forma de campana tal como se esperaría. En la parte superior derecha de la misma Figura se encuentra un gráfico de valores residuales los cuales se distribuyen uniformemente alrededor de cero, por lo tanto, los residuos se distribuyen normal.

Finalmente de la Figura 3.25, en la parte inferior derecha está el gráfico de respuesta frente a los valores ajustados. En un modelo perfecto formaría una línea recta. De la gráfica podemos observar que el modelo forma una línea muy cercana a una recta. Todo esto indica un buen ajuste del modelo y por lo tanto, confirmamos que los residuos distribuyen normal.

Por lo tanto, el modelo con las variables seleccionadas con Boruta cumple todos los supuestos del modelo aditivo generalizado. Con esto es posible crear proyecciones para la pérdida dado el incumplimiento. En la Figura 3.26 se observa un gráfico del ajuste (antes de la linea roja punteada) y las proyecciones (después de la linea roja punteada) del modelo aditivo generalizado, junto con las pérdidas dado el incumplimiento reales.

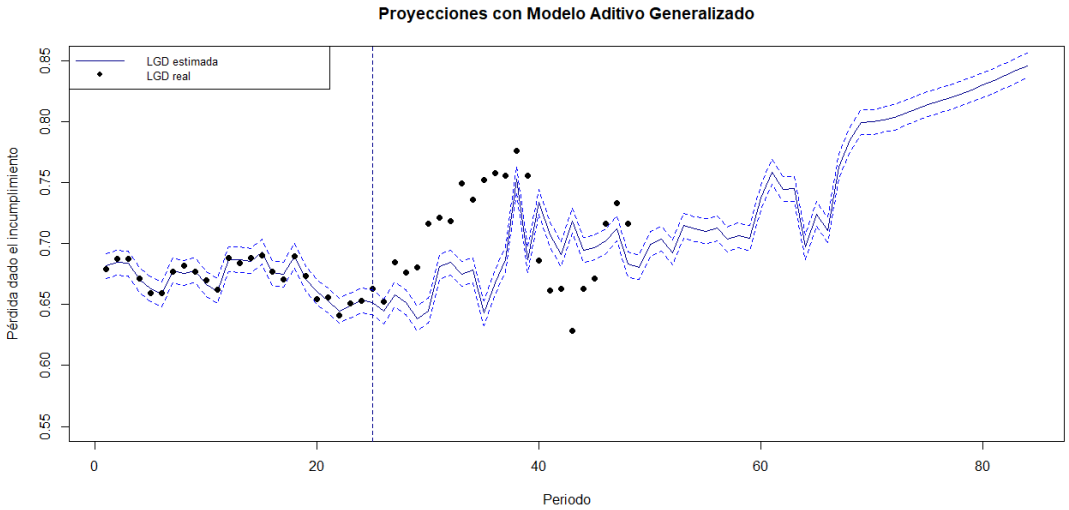

Figura 3.26: Ajuste y predicciones con el modelo aditivo generalizado para la variable objetivo pérdida dado el incumplimiento.

3.5. Modelos de series de tiempo

Al igual que en el caso de los modelos de regresión (a excepción del modelo aditivo generalizado), al abordar los modelos de series temporales generados, se focalizará en el sentido lógico de negocio. Esto implica que la relación entre la tasa de incumplimiento y/o la pérdida dado el incumplimiento con las variables macroeconómicas utilizadas debe tener coherencia. Es decir, todas las variables deben exhibir una relación positiva con la tasa de incumplimiento y/o la pérdida dado el incumplimiento, excepto las variables del Producto Interno Bruto (PIB), que deben mostrar una relación negativa con la tasa de incumplimiento y/o la pérdida dado el incumplimiento.

Adicionalmente, tanto para el modelo Autorregresivo Integrado de Media Móvil con Variables Exógenas (ARIMAX) como para el modelo Autorregresivo con Variables Exógenas (ARX), es necesario que ambas series de tiempos, sean estacionarias. Para evaluar la estacionariedad, se utiliza el test de Phillips-Perron. La hipótesis nula de este test plantea la existencia de una raíz unitaria, es decir, la serie no es estacionaria.

```
    Phillips-Perron Unit Root Test


 data:  serie
 Dickey-Fuller Z(alpha) = -7.7198, Truncation lag parameter = 3, p-value
 = 0.652
 alternative hypothesis: stationary
```

Al aplicar el test de Phillips-Perron a la serie de tasas de incumplimiento sin diferenciar, se concluye que la serie de tiempo no es estacionaria, ya que el valor-p es mayor a 0.05.

```
    Phillips-Perron Unit Root Test


data:  seriedif
Dickey-Fuller Z(alpha) = -21.239, Truncation lag parameter = 3, p-value
= 0.0326
alternative hypothesis: stationary
```

Posteriormente, al aplicar el test a la serie de tasas de incumplimiento diferenciada

se observa que la serie es estacionaria.

```
    Phillips-Perron Unit Root Test

data:  serie
Dickey-Fuller Z(alpha) = -10.125, Truncation lag parameter = 2, p-value
= 0.4669
alternative hypothesis: stationary
```

Al aplicar el test de Phillips-Perron a la serie pérdida dado el incumplimiento sin diferenciar, se concluye que la serie de tiempo no es estacionaria, ya que el valor-p es mayor a 0,05.

```
    Phillips-Perron Unit Root Test

data:  seriedif
Dickey-Fuller Z(alpha) = -21.505, Truncation lag parameter = 2, p-value
= 0.01574
alternative hypothesis: stationary
```

Posteriormente, al aplicar el test a la serie pérdida dado el incumplimiento diferenciada se observa que la serie es estacionaria. Teniendo en cuenta esto, se generaron los modelos ARIMAX y ARX.

3.5.1. Modelo Autorregresivo Integrado de Media Móvil con Variables Exógenas

Para escoger los parámetros p y q se utilizará el criterio de información de Akaike (AIC). El modelo que minimice el valor de AIC se considera el más adecuado para los datos observados.

Se observa así que los parámetros escogidos son. $p = 1, d = 1$ y $q = 1$, pues el modelo ARIMAX generado con esos parametros es el que menor AIC posee (con un valor de $-195{,}5077$).

Los coeficientes obtenidos al construir el modelo autorregresivo integrado de media móvil con variables exógenas (o ARIMAX) para la variable objetivo probabilidad

p/q	0	1	2	3	4
0	-190.7428	-192.5058	-194.1890	-190.9320	-194.0392
1	-190.6461	-195.5077	-191.8400	-189.9567	-192.0607
2	-193.4539	-191.9957	-194.0025	-192.6547	-193.1084
3	-192.1211	-190.1223	-192.5296	-195.7434	-192.6620
4	-190.1250	-193.2952	-190.3736	-194.0053	-192.4748

Cuadro 3.12: Valores AIC para la serie de tiempo probabilidad de incumplimiento.

de incumplimiento con las variables seleccionadas por Boruta y regularización Lasso se presentan en la Tabla 3.13.

Variables Boruta		**Variables Regularización Lasso**	
Variable	**Coeficiente**	**Variable**	**Coeficiente**
TPM 21	0.055	Paridad 23	0.421
crisis 4	0.001	IPC 16	3.782
IPC 36	3.866	Desempleo 11	0.135
Desempleo 5	0.0002	TPM 10	0.083
Paridad 1	0	crisis 4	0.002
PIB 15	-0.073	PIB 19	0.100

Cuadro 3.13: Coeficientes del modelo ARIMAX para la variable objetivo probabilidad de incumplimiento.

De la Tabla 3.13 se observa que las variables cuentan con sentido negocio, pues los coeficientes encontrados para las diferentes variables son los esperados.

Por su parte de la Tabla 3.13 se observa que la variable PIB cuenta con un coeficiente positivo lo cual no tiene sentido, pues a mayor producto interno bruto (PIB) se espera que los incumplimientos disminuyan y por tanto la tasa de incumplimiento disminuya.

Así el modelo modelo autorregresivo integrado de media móvil con variables exógenas generado con las variables Boruta es el posible modelo escogido. Teniendo el modelo se realizaron diferentes test para probar que el modelo es adecuado.

El test de Ljung-Box permite comprobar si una serie de observaciones en un período de tiempo específico son aleatorias e independientes. Así, si se obtiene que las observaciones no son independientes, existirá autocorrelación, lo cual puede disminuir la exactitud del modelo predictivo. De la Tabla 3.14 se ve que el

Prueba	Estadístico	Grados de libertad	Valor p
Box-Ljung test	0.0047891	1	0.9448
Box-Pierce test	0.0045017	1	0.9465
Jarque Bera Test	0.98258	2	0.6118
Shapiro-Wilk normality test	0.9661	NA	0.1778

Cuadro 3.14: Supuestos modelo ARIMAX para la variable objetivo probabilidad de incumplimiento.

resultado del test es de 0.9448, así se acepta la hipótesis nula, la cual dice que los datos del modelo se distribuyen de forma independiente, cumpliéndose así el supuesto de independencia.

El text de Box-Pierce sirve para probar la independencia de las observaciones en un serie, donde la hipótesis nula es que existe independencia. De la Tabla 3.14 se observa que resultado del test es de 0.9465, por lo que se acepta la hipótesis nula, cumpliéndose así el supuesto de independencia de las observaciones.

EL test de Jarque-Bera es una prueba de bondad de ajuste que permite comprobar si una muestra tiene asimetría y la curtosis de una distribución normal. La hipótesis nula es que existe normalidad, de la Tabla 3.14 se observa que el resultado del test fue de 0.6118, con lo que se acepta la hipótesis nula, así el modelo distribuye normal.

Finalmente el test de Shapiro-Wilk permite comprobar la normalidad de la serie; siendo la hipótesis nula que la serie distribuye normal. De la Tabla 3.14 se observa que el resultado del test es de 0.1778, es decir, se acepta la hipótesis nula, de normalidad de la serie de tiempo.

Así, se concluye que la serie de tiempo generada tiene comportamiento de ruido blanco gaussiano, es decir, independiente y normal.

Dicho esto, el modelo cumple con todos los supuestos necesarios, con esto es posible crear proyecciones para la tasa de incumplimiento. En la Figura 3.27 se observa un gráfico del ajuste (antes de la linea roja punteada) y las proyecciones (después de la linea roja punteada) del modelo autorregresivo integrado de media móvil con variables exógenas, junto con las tasas de incumplimientos reales.

De la Figura 3.27, se observa que el modelo cuenta con buen ajuste para el periodo de construcción y parece tener buen ajuste para los primeros periodos de

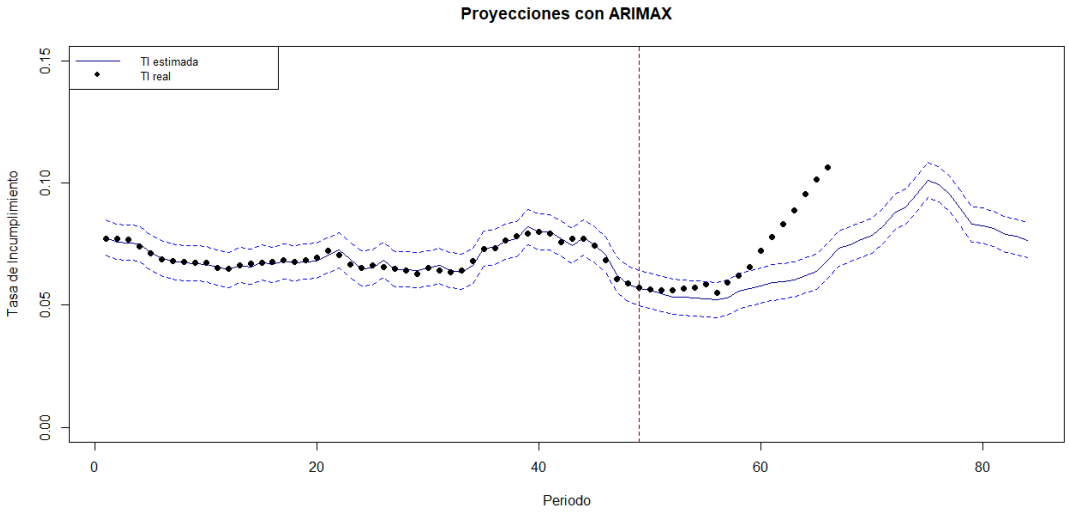

Figura 3.27: Ajuste y predicciones con el modelo autorregresivo integrado de media móvil con variables exógenas.

backtest, es por esto que el modelo será utilizado para proyectar la pérdida dado el incumplimiento.

Para escoger los parámetros p y q se utilizará el criterio de información de Akaike (AIC). El modelo que minimice el valor de AIC se considera el más adecuado para los datos observados.

p/q	0	1	2	3	4
0	-208.7575	-211.6394	-209.9125	-210.0973	-207.4006
1	-209.9175	-209.9837	-210.3261	-209.5343	-205.8677
2	-208.8614	-210.3670	-209.0555	-207.5549	-208.9233
3	-209.1954	-207.2523	-205.2716	-203.9052	-207.1695
4	-207.2337	-205.2558	-205.7304	-205.7781	-204.1378

Cuadro 3.15: Valores AIC para la serie de tiempo pérdida dado el incumplimiento.

Se observa así que los parámetros escogidos son. $p = 0, d = 1$ y $q = 1$, pues el modelo ARIMAX generado con esos parametros es el que menor AIC posee (con un valor de $-211{,}6394$).

Los coeficientes obtenidos al entrenar el modelo autorregresivo integrado de media móvil con variables exógenas (o ARIMAX) para la variable objetivo pérdida dado el incumplimiento con las variables seleccionadas por Boruta y regularización Lasso se presentan en las Tablas 3.16.

Variables Boruta		Variables Regularización Lasso	
Variable	**Coeficiente**	**Variable**	**Coeficiente**
Paridad 2	-0.0001	Paridad 14	0.098
Desempleo 29	0.052	Desempleo 17	0.030
PIB 12	-0.274		

Cuadro 3.16: Coeficientes del modelo ARIMAX para la variable objetivo pérdida dado el incumplimiento.

De la Tabla 3.16 se observa que la variable Paridad 2 cuenta con un coeficiente negativo lo cual no cuenta con sentido negocio, por lo que este modelo queda descartado.

Por su parte, de la Tabla 3.16 se observa que las variables cuentan con sentido negocio, pues los coeficientes encontrados para las diferentes variables son los esperados.

Así el modelo modelo autorregresivo integrado de media móvil con variables exógenas generado con las variables Regularización Lasso es el posible modelo escogido. Teniendo el modelo se realizaron diferentes test para probar que tan adecuado es el modelo.

Prueba	**Estadístico**	**Grados de libertad**	**Valor p**
Box-Ljung test	0.28364	1	0.5943
Box-Pierce test	0.25091	1	0.6164
Jarque Bera Test	2.3191	2	0.3136
Shapiro-Wilk normality test	0.95983	NA	0.435

Cuadro 3.17: Supuestos modelo ARIMAX para la variable objetivo pérdida dado el incumplimiento.

El test de Box-Ljung comprueba si la serie de observaciones son aleatorias e independientes, de la Tabla 3.17 se observa que el resultado del test es de 0.5943, así se acepta la hipótesis nula, la cual dice que los datos del modelo se distribuyen de forma independiente y aleatoria, por lo tanto los datos del modelo se distribuyen de forma independiente, es decir, no existe autocorrelación, así se cumple el supuesto de independencia.

El text de Box-Pierce sirve para probar la independencia de las observaciones en un serie, donde la hipótesis nula es que existe independencia. De la Tabla 3.17 se

observa que resultado del test es de 0.6164, por lo que se acepta la hipótesis nula. es decir existe independencia de las observaciones.

La hipótesis nula del test de Jarque Bera es que la serie sigue una distribución normal, de la Tabla 3.17 se observa que el resultado del test fue de 0.3136, con lo que se acepta la hipótesis nula de normalidad, por lo que el modelo distribuye normal.

La hipótesis nula del test de Shapiro-Wilk es que la serie distribuye normal, de la Tabla 3.17 se observa que el resultado del test es de 0.435, es decir, se acepta la hipótesis nula de normalidad de la serie.

Así , se concluye que la serie de tiempo generada tiene comportamiento de ruido blanco gaussiano, es decir, independiente y normal.

Desde el análisis previo, se deduce que el modelo cumple con todos los supuestos necesarios, con esto es posible crear proyecciones para la pérdida dado el incumplimiento. En la Figura 3.28 se observa un gráfico del ajuste (antes de la linea roja punteada) y las proyecciones para la pérdida dado el incumplimiento (después de la linea roja punteada) del modelo autorregresivo integrado de media móvil con variables exógenas, junto con las pérdidas dado el incumplimiento reales.

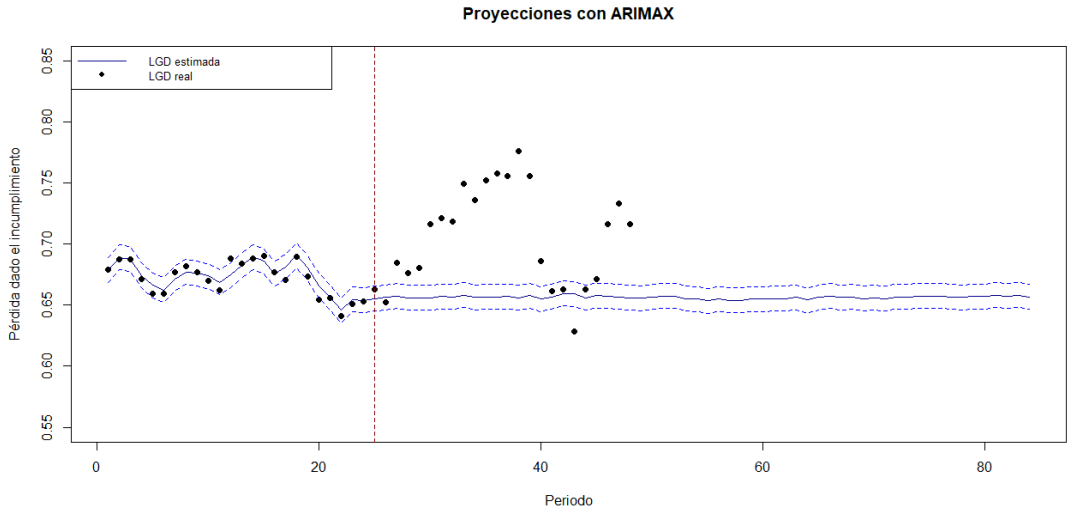

Figura 3.28: Ajuste y predicciones con el modelo autorregresivo integrado de media móvil con variables exógenas para la pérdida dado el incumplimiento.

3.5.2. Modelo Autorregresivo con Variables Exógenas

Para escoger el parámetro p se utilizará el criterio de información de Akaike (AIC). El modelo que minimice el valor de AIC se considera el más adecuado para los datos observados.

p	AR
0	-190.7428
1	-190.6461
2	-193.4539
3	-192.1211
4	-190.1250
5	-189.2292

Cuadro 3.18: Valores AIC para la serie de tiempo probabilidad de incumplimiento.

Se observa así que los parámetros escogidos son. $p = 2, d = 1$, pues el modelo ARX generado con esos parámetros es el que menor AIC posee (con un valor de $-193{,}4539$).

Los coeficientes obtenidos al entrenar el modelo autorregresivo (o ARX) para la variable objetivo probabilidad de incumplimiento con las variables seleccionadas por Boruta y regularización Lasso se presentan en la Tabla 3.19.

Variables Boruta		**Variables Regularización Lasso**	
Variable	**Coeficiente**	**Variable**	**Coeficiente**
TPM 21	0.050	Paridad 23	0.521
crisis 4	0.001	IPC 16	3.903
IPC 36	3.127	Desempleo 11	0.139
Desempleo 5	0.006	TPM 10	0.104
Paridad 1	0	crisis 4	0.001
PIB 15	-0.097	PIB 19	0.085

Cuadro 3.19: Coeficientes del modelo ARX para la variable objetivo probabilidad de incumplimiento.

De la Tabla 3.19 se observa que las variables cuentan con sentido negocio, pues los coeficientes encontrados para las diferentes variables son los esperados.

Por su parte de la Tabla 3.19 se observa que la variable PIB cuenta con un coeficiente positivo lo cual contradice lo esperado, por lo que el modelo queda descartado.

Así el modelo modelo autorregresivo con variables exógenas generado con las variables Boruta es el posible modelo escogido. Teniendo el modelo se realizaron diferentes test para probar que tan adecuado es el modelo.

Prueba	**Estadístico**	**Grados de libertad**	**Valor p**
Box-Ljung test	0.070021	1	0.7913
Box-Pierce test	0.065819	1	0.7975
Jarque Bera Test	0.60902	2	0.7375
Shapiro-Wilk normality test	0.96489	NA	0.159

Cuadro 3.20: Supuestos modelo ARX para la variable objetivo probabilidad de incumplimiento.

De la Tabla 3.20 se obtiene que el resultado para el test de Ljung-Box es de 0.7913, así se acepta la hipótesis nula, la cual dice que los datos del modelo se distribuyen de forma aleatoria e independiente. Por su parte del test de Box-Pierce se obtiene que el resultado del test es de 0.7975, por lo que se acepta la hipótesis nula, la cual nos dice que la serie de observaciones es independiente.

De la Tabla 3.20 del test de Jarque-Bera se tiene que el resultado del test fue de 0.7375, con lo que se acepta la hipótesis nula, la cual dice que la serie distribuye normal. Finalmente el resultado del test de Shapiro-Wilk es de 0.159, es decir, se acepta la hipótesis nula de normalidad de la serie.

Así , se concluye que la serie de tiempo generada tiene comportamiento de ruido blanco gaussiano, es decir, independiente y normal.

Dicho esto, el modelo cumple con todos los supuestos necesarios, con esto es posible crear proyecciones para la tasa de incumplimiento. En la Figura 3.29 se observa un gráfico del ajuste (antes de la linea roja punteada) y las proyecciones (después de la linea roja punteada) del modelo autorregresivo con variables exógenas, junto con las tasas de incumplimientos reales.

De la Figura 3.29, se observa que el modelo cuenta con buen ajuste para el periodo de construcción y parece tener buen ajuste para los primeros periodos de

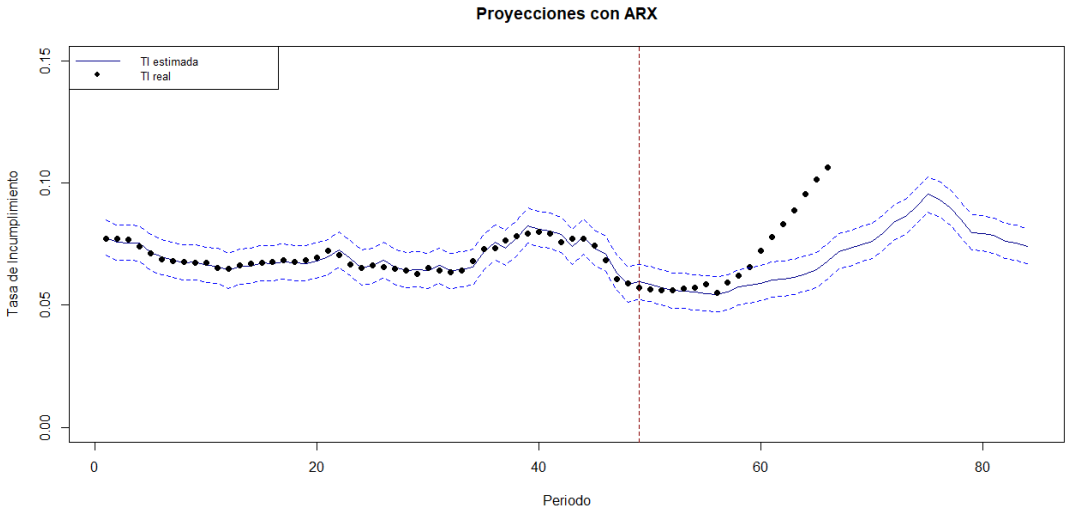

Figura 3.29: Ajuste y predicciones con el modelo autorregresivo con variables exógenas.

backtest, es por esto que el modelo será utilizado para proyectar la pérdida dado el incumplimiento.

Para escoger el parámetro p se utilizará el criterio de información de Akaike (AIC). El modelo que minimice el valor de AIC se considera el más adecuado para los datos observados.

p	AR
0	-208.7575
1	-209.9175
2	-208.8614
3	-209.1954
4	-207.2337
5	-205.3628
5	-189.2292

Cuadro 3.21: Valores AIC para la serie de tiempo pérdida dado el incumplimiento.

Se observa así que los parámetros escogidos son. $p = 1d = 1$, pues el modelo ARX generado con esos parámetros es el que menor AIC posee (con un valor de $-209{,}9175$).

Los coeficientes obtenidos al entrenar el modelo autorregresivo con variables exógenas (o ARX) para la variable objetivo pérdida dado el incumplimiento con

las variables seleccionadas por Boruta y regularización Lasso se presentan en la Tabla 3.22.

Variables Boruta		**Variables Regularización Lasso**	
Variable	**Coeficiente**	**Variable**	**Coeficiente**
Paridad 2	-0.000	Paridad 14	0.100
Desempleo 29	0.063	Desempleo 17	0.030
PIB 12	-0.807		

Cuadro 3.22: Coeficientes del modelo ARX para la variable objetivo pérdida dado el incumplimiento.

De la Tabla 3.22 se observa que la variable Paridad cuenta con un coeficiente negativo lo cual no tiene sentido, por lo que este modelo queda descartado.

Por su parte de la Tabla 3.22 se observa que las variables cuentan con sentido negocio, pues los coeficientes encontrados para las diferentes variables son los esperados.

Así el modelo modelo autorregresivo integrado de media móvil con variables exógenas generado con las variables Regularización Lasso es el posible modelo escogido. Teniendo el modelo se realizaron diferentes test para probar que tan adecuado es el modelo.

Prueba	**Estadístico**	**Grados de libertad**	**Valor p**
Box-Ljung test	0.11233	1	0.7375
Box-Pierce test	0.099365	1	0.7526
Jarque Bera Test	5.5784	2	0.06147
Shapiro-Wilk normality test	0.92716	NA	0.08422

Cuadro 3.23: Supuestos modelo ARX para la variable objetivo pérdida dado el incumplimiento.

De la Tabla 3.23 se obtiene que el resultado para el test de Ljung-Box es de 0.7375, así se acepta la hipótesis nula, la cual dice que los datos del modelo se distribuyen de forma aleatoria e independiente. Por su parte del test de Box-Pierce se obtiene que el resultado del test es de 0.7526, por lo que se acepta la hipótesis nula, la cual nos dice que la serie de observaciones es independiente.

De la Tabla 3.23 del test de Jarque-Bera se tiene que el resultado del test fue de 0.06147, con lo que se acepta la hipótesis nula, la cual dice que la serie distribuye

normal. Finalmente el resultado del test de Shapiro-Wilk es de 0.08422, es decir, se acepta la hipótesis nula de normalidad de la serie.

Así , se concluye que la serie de tiempo generada tiene comportamiento de ruido blanco gaussiano, es decir, independiente y normal.

Desde el análisis previo, se deduce que el modelo cumple con todos los supuestos necesarios, con esto es posible crear proyecciones para la pérdida dado el incumplimiento. En la Figura 3.30 se observa un gráfico del ajuste (antes de la linea roja punteada) y las proyecciones (después de la linea roja punteada) del modelo autorregresivo integrado de media móvil con variables exógenas, junto con las tasas de incumplimientos reales.

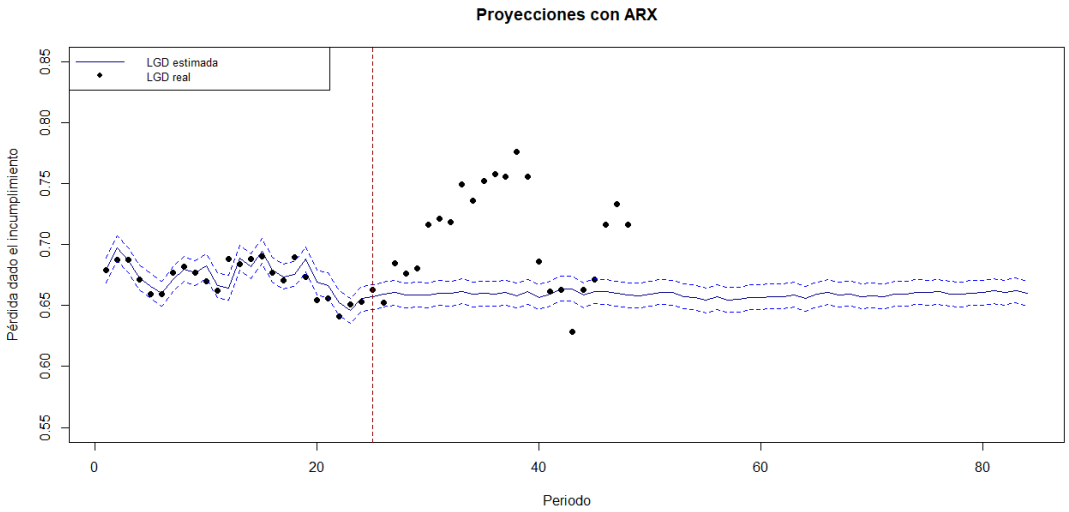

Figura 3.30: Ajuste y predicciones con el modelo autorregresivo con variables exógenas para la pérdida dado el incumplimiento.

3.6. Comparación de modelos

A continuación, se presentan las métricas de bondad de ajuste correspondientes a los diferentes modelos durante el periodo de construcción. El propósito es comparar estos modelos en función de su capacidad para predecir la probabilidad de incumplimiento y la pérdida dado el incumplimiento, evaluando aspectos como el Error Cuadrático Medio de la Raíz (RECM), el coeficiente de determinación (R^2), y el Criterio de Información de Akaike (AIC).

De las Tablas 3.24 y 3.25, se desprende que el modelo aditivo generalizado exhibe

Modelo	RECM	R^2	AIC
Regresión Beta	0.003	0.609	-393.652
Modelo Aditivo Generalizado	0.001	0.982	-511.340
Regresión Cox	0.005	0.182	272.708
ARIMAX	0.001	0.948	-190.627
ARX	0.002	0.915	-182.158
Modelo Institución	0.002	1.000	-161.547

Cuadro 3.24: Métricas de bondad de ajuste para el periodo de construcción para la variable objetivo tasa de incumplimiento.

Modelo	RECM	R^2	AIC
Modelo Aditivo Generalizado	0.003	0.957	-184.187
ARIMAX	0.006	0.845	-141.790
ARX	0.009	0.596	-140.609
Modelo Institución	0.015	0.084	NA

Cuadro 3.25: Métricas de bondad de ajuste para el periodo de construcción para la variable objetivo pérdida dado el incumplimiento.

el menor RECM y AIC, siendo uno de los modelos con el mejor coeficiente de determinación (R^2), solo superado por el modelo actual de la institución financiera.

Adicionalmente, se observa que, tanto en el periodo de construcción como en el de backtest, los modelos de regresión beta y regresión cox presentan los peores ajustes, evidenciados por sus altos valores de RECM y bajos coeficientes de determinación.

Modelo	RECM	R^2
Regresión Beta	0.009	0.730
Modelo Aditivo Generalizado	0.007	0.824
Regresión Cox	0.015	0.204
ARIMAX	0.018	-0.186
ARX	0.018	-0.113
Modelo Institución	0.013	0.385

Cuadro 3.26: Métricas de bondad de ajuste para el periodo de backtest para la variable objetivo tasa de incumplimiento.

En las Tablas 3.26 y 3.27, se destaca que el modelo aditivo generalizado continúa mostrando el RMSE más bajo en comparación con los demás modelos y posee el mejor coeficiente de determinación para los periodos de backtest. Asimismo, se observa que, a pesar de un ajuste deficiente durante el periodo de construcción, la

Modelo	RECM	R^2
Modelo Aditivo Generalizado	0.058	-1.702
ARIMAX	0.063	-2.201
ARX	0.060	-1.952
Modelo Institución	0.052	-1.168

Cuadro 3.27: Métricas de bondad de ajuste para el periodo de backtest para la variable objetivo pérdida dado el incumplimiento.

regresión beta mejora significativamente al proyectar su ajuste durante el periodo de backtest para la variable objetivo tasa de incumplimiento.

Dado el análisis de las métricas de bondad de ajuste en los periodos de construcción y backtest, se llega a la conclusión de que el modelo aditivo generalizado es la opción que mejor proyecta la probabilidad de incumplimiento y la pérdida dado el incumplimiento.

4. Conclusión

El objetivo principal de esta tesis era realizar una comparación entre modelos que predicen la probabilidad de incumplimiento y la pérdida dado el incumplimiento con una perspectiva forward-looking. Para lograr esto, se llevó a cabo una exhaustiva revisión bibliográfica que abarcó los algoritmos de selección de variables Boruta y Regularización Lasso y los modelos de regresión y series temporales utilizados en la predicción de la probabilidad de incumplimiento y la pérdida dado el incumplimiento, revelando una amplia gama de enfoques estudiados. Seguidamente, se realizó una revisión teórica detallada de los algoritmos y modelos propuestos. Estos modelos fueron luego implementados utilizando los datos detallados en las secciones del Capitulo 3: Descripción de la base de datos, Resultados selección de variables y Análisis exploratorio de los datos.

A través del análisis de los datos y los resultados obtenidos, se destacó la fuerte dependencia de la calidad de los datos en el rendimiento de los modelos. Además, al comparar los modelos tanto de regresión como de series temporales, se concluyó que la metodología más adecuada para predecir la probabilidad de incumplimiento y la pérdida dado el incumplimiento forward-looking es el modelo aditivo generalizado. Como se explicó en el capitulo de Implementación este modelo no solo es el que mejor ajusta y predice en comparación a los demás modelos de regresión y/o series temporales, sino que además es capaz de capturar y la tendencia al alza y/o a la baja tanto de la probabilidad de incumplimiento como de la pérdida dado el incumplimiento lo que proporciona una predicción más cercana a la realidad.

Es así, que en base al análisis estadístico, se puede afirmar que el modelo aditivo generalizado ofrece un mejor ajuste y predicción de la probabilidad de incumplimiento y la pérdida dado el incumplimiento, pues este modelo refleja de manera conjunta varios aspectos fundamentales:

1. El modelo es capaz de capturar tendencias que los demás modelos presentados en esta tesis no son capaces de capturar.
2. Sus métricas de construcción y backtest superan a las obtenidas por otras metodologías mencionadas en esta tesis.
3. No presenta sobre-ajuste al conjunto de construcción.
4. Cumple con todos los supuestos teóricos del modelo.

Estos hallazgos respaldan la conclusión de que el modelo aditivo generalizado es la elección más sólida y confiable para abordar la predicción de la probabilidad de incumplimiento y la pérdida dado el incumplimiento en una perspectiva forward-looking.

4.1. Limitaciones del estudio

Entre las limitaciones encontradas para llevar a cabo el estudio se encuentran la escasa disponibilidad de datos para un análisis más extenso y detallado. La escasez de datos limitó la aplicación de modelos de Machine Learning (los modelos de machine learning son menos sensibles a los supuestos sobre la distribución de los datos, lo que puede ser beneficioso en situaciones donde los datos no cumplen con los supuestos de los modelos de regresión tradicionales), los cuales suelen utilizarse con conjuntos de datos más grandes para lograr una mayor precisión y generalización.

Además, la falta de registros suficientes pudo haber tenido un efecto en la selección de modelos y en la capacidad para describir posibles relaciones temporales en los datos. Específicamente, el modelo propuesto no fue capaz de capturar posibles dependencias temporales debido a la limitación en la cantidad de datos disponibles.

En términos metodológicos, es importante reconocer que cualquier elección específica, como por ejemplo, en la selección de variables, puede tener limitaciones inherentes que podrían afectar la robustez de los resultados, debido a que los algoritmos de selección de variables dependen fuertemente de los parámetros o hiper-parámetros que se ingresan en el algoritmo, como por ejemplo, en la Regularización Lasso, dependiendo del λ o parámetro de regularización se tienen más o menos variables con coeficientes igual a cero, pues a medida que λ aumenta, la regularización Lasso penaliza más fuertemente los coeficientes, lo que puede llevar a más coeficientes exactamente iguales a cero.

4.2. Trabajos Futuros

A lo largo de este trabajo, se ha observado que las métricas de bondad de ajuste, especialmente el coeficiente de determinación, para el periodo de backtest no superan el 83 % en el caso de la probabilidad de incumplimiento, e incluso muestran valores negativos para la pérdida dado el incumplimiento. Aunque

estos resultados son significativamente mejores que las métricas obtenidas por la Institución Financiera, existe la oportunidad de mejorar aún más mediante el desarrollo de metodologías basadas en diferentes algoritmos y modelos, como los modelos de Machine Learning. Es evidente que la capacidad predictiva de los modelos podría beneficiarse de la exploración de algoritmos más avanzados o de la combinación de enfoques diversos.

Por otro lado, la aplicación de esta metodología a otros modelos de provisión podría ser de gran beneficio. La necesidad de prever la probabilidad de incumplimiento y la pérdida dado el incumplimiento no se limita únicamente al marco de IFRS9, sino que es una necesidad compartida en diversos modelos de provisiones utilizados en el ámbito financiero. Por ende, se propone la aplicación de esta metodología en diferentes contextos, con el objetivo de evaluar su robustez y eficacia en escenarios diversos.

Referencias

Augustin, N., Sauleaub, E.-A., and Wood, S. (2012). *On quantile quantile plots for generalized linear models.* Computational Statistics Data Analysis.

Bellini, T. (2019). *IFRS 9 and CECL Credit Risk Modelling and Validation: A Practical Guide with Examples Worked in R and SAS.* Academic Press.

Board, I. A. S. (2014). *NIIF 9 Instrumentos Financieros.*

Box, G. E. and Jenkins, G. M. (1970). *Time Series Analysis: Forecasting and Control.* Holden-Day.

Cox, D. R. (1972). *Regression Models and Life-Tables.* Journal of the Royal Statistical Society, Series B.

Diversi, R., Guidorzi, R., and Soverini, U. (2008). Identification of autoregressive models in the presence of additive noise. *International Journal of Adaptive Control and Signal Processing*, 22(5):465–481.

Ferrari SLP, C.-N. F. (2004). *Beta Regression for Modelling Rates and Proportions.* Journal of Applied Statistics.

Geissinger, e. a. (2022). *A case for beta regression in the natural sciences.* Ecosphere.

Grambsch, P. M. and Therneau, T. M. (1994). *Proportional Hazards Tests and Diagnostics Based on Weighted Residuals.* Biometrika.

Hastie, T. J. and Tibshirani, R. J. (1990). *Generalized Additive Models.* Chapman Hall/CRC.

Hyndman, R. J. and Koehler, A. B. (2006). Another look at measures of forecast accuracy. *International Journal of Forecasting.*

Jacobs Jr, M. (2019). An analysis of the impact of modeling assumptions in the current expected credit loss (cecl) framework on the provisioning for credit loss. *The Journal of Risk and Control*, 6(1):65–114.

Kursa, M. B. and Rudnicki, W. R. (2010). *Feature selection with the boruta package.* Journal of Statistical Software.

Robert, H. et al. (2006). Time series analysis and its applications with r examples second edition.

Sakamoto, Y., Ishiguro, M., and Kitagawa, G. (1986). *Akaike Information Criterion Statistics*. D. Reidel Publishing Company.

Slutzky, E. E. (1927). *The Summation of Random Causes as the Source of Cyclic Processes*. Econometrica.

Steel, R. and Torrie, J. (1960). *Principles and Procedures of Statistics with Special Reference to the Biological Sciences*. McGraw Hill.

Tasche, D. (2015). *Forecasting Portfolio Credit Default Rates*.

Tibshirani, R. (1996). *Regression Shrinkage and Selection via the lasso*. Journal of the Royal Statistical Society. Series B (methodological).

Wood, S. (2017). *Generalized Additive Models: An Introduction with R (2nd edition)*. Chapman and Hall/CRC Press.

Yule, G. U. (1927). *On a Method of Investigating Periodicities in Disturbed Series, with Special Reference to Wolfer's Sunspot Numbers*. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences.

A. Anexo

A.1. Librerías Utilizadas

En el cuadro a continuación se presentan las librerías utilizadas en el software R para cada algoritmo y modelo estadístico utilizado.

Metodología	**Función**	**Libreria**
Boruta	Boruta	Boruta
Regularización Lasso	cv.glmnet	glmnet
Regresión Beta	betareg	betareg
Modelo Aditivo Generalizado	gam	mgcv
Regresión Cox	coxph	survival
ARIMAX	arima	stats
ARX	arima	stats

Cuadro A.1: Librerías utilizadas para la construcción de los algoritmos y modelos.

A continuación se especifican los parámetros utilizados para construir cada uno de los algoritmos y modelos.

A.1.1. Boruta

Los principales argumentos utilizados para construir el algoritmo Boruta son los siguientes:

- **fórmula:** Expresión de la forma $y \sim x_1 + \cdots + x_n$, donde y es la variable de respuesta y $x_1, \ldots, x_n$ son las posibles variables explicativas.
- **data:** Representa el conjunto de datos del que se tomarán las variables especificadas en fórmula.

A.1.2. Regularización Lasso

Los principales argumentos utilizados para construir el algoritmo Regularización Lasso son los siguientes:

- **X:** Representa el conjunto de datos en donde se encuentran las posibles variables explicativas.
- **Y:** Representa la variable de respuesta.

- **lambda:** Representa el hiperparámetro de regularización que controla el grado de penalización, en este caso, tiene un valor de 1 (el mayor valor de alpha posible).

A.1.3. Regresión Cox

Los principales argumentos utilizados para construir el modelo de Regresión Cox son los siguientes:

- **fórmula:** Expresión de la forma $y \sim x_1 + \cdots + x_n$, donde y es la variable de respuesta y $x_1, \ldots, x_n$ son las posibles variables explicativas.
- **data:** Representa el conjunto de datos del que se tomarán las variables especificadas en fórmula.

A.1.4. Regresión Beta

Los principales argumentos utilizados para construir el modelo de Regresión Beta son los siguientes:

- **fórmula:** Expresión de la forma $y \sim x_1 + \cdots + x_n$, donde y es la variable de respuesta y $x_1, \ldots, x_n$ son las posibles variables explicativas.
- **data:** Representa el conjunto de datos del que se tomarán las variables especificadas en fórmula.
- **link:** Representa la función de enlace en el modelo que se utilizan en diferentes situaciones dependiendo de la naturaleza de los datos y la relación que se espera entre las variables, en este caso, se utiliza la función de enlace logit.

A.1.5. Modelo Aditivo Generalizado

Los principales argumentos utilizados para construir el Modelo Aditivo Generalizado son los siguientes:

- **fórmula:** Expresión de la forma $y \sim s(x_1) + \cdots + s(x_n)$, donde y es la variable de respuesta y $x_1, \ldots, x_n$ son las posibles variables explicativas.
- **data:** Representa el conjunto de datos del que se tomarán las variables especificadas en fórmula.

- **family:** Representa a la familia de distribución de probabilidad que se asume para el término de error en el modelo, en este caso, se escoge la familia gaussiana.

A.1.6. Modelo Autorregresivo Integrado de Media Móvil con Variables Exógenas

Los principales argumentos utilizados para construir el Modelo Autorregresivo Integrado de Media Móvil con Variables Exógenas son los siguientes:

- **serie:** Representa la serie temporal que estás modelando.
- **order:** Representa los órdenes autorregresivos, diferenciales e integradores del modelo.
- **xreg:** Es una matriz que contiene las variables exógenas que se incorporan al modelo.

A.1.7. Modelo Autorregresivo con Variables Exógenas

Los principales argumentos utilizados para construir el Modelo Autorregresivo con Variables Exógenas son los siguientes:

- **serie:** Representa la serie temporal que estás modelando.
- **order:** Representa los órdenes autorregresivos y diferenciales del modelo.
- **xreg:** Es una matriz que contiene las variables exógenas que se incorporan al modelo.