ESCUELA POLITÉCNICA NACIONAL

FACULTAD DE CIENCIAS

CONTRASTANDO EL MACHINE LEARNING Y LA ECONOMETRÍA EN SERIES TEMPORALES

TRABAJO DE TITULACIÓN PREVIO A LA OBTENCIÓN DEL TÍTULO DE INGENIERO EN CIENCIAS ECONÓMICAS Y FINANCIERAS

PROYECTO DE INVESTIGACIÓN

CHRISTIAN MATEO QUIGUIRI DAQUILEMA
mateo.quiguiri@epn.edu.ec

DIRECTORA: PhD. ANDREA GABRIELA BONILLA BOLAÑOS
andrea.bonilla@epn.edu.ec

Quito, octubre de 2023

CERTIFICACIÓN

Certifico que el presente trabajo fue realizado por **Christian Mateo Quiguiri Daquilema**, bajo mi supervisión.

PhD. Andrea Gabriela Bonilla Bolaños

DECLARACIÓN

Yo, **Christian Mateo Quiguiri Daquilema**, declaro bajo juramento que el trabajo aquí descrito es de mi autoría; que no ha sido previamente presentado para ningún grado o calificación profesional; y, que he consultado las referencias bibliográficas que se incluyen en este documento.

A través de la presente declaración cedo mis derechos de propiedad intelectual correspondientes a este trabajo, a la Escuela Politécnica Nacional, según lo establecido por la Ley de Propiedad Intelectual, por su Reglamento y por la normativa institucional vigente.

Christian Mateo Quiguiri Daquilema

DEDICATORIA

Dedico esta tesis a mamá, una entre las incontables madres que, con coraje y determinación, ha tenido que criar sola a sus hijos. Rindo homenaje a tu esfuerzo sobrehumano, a ese sacrificio incansable para brindarme las oportunidades que nunca tuviste, a tu amor incondicional, y a la protección que siempre me has ofrecido.

Este logro también es un tributo a todas las madres solteras, que, a pesar de las adversidades y los desafíos incesantes, permanecen fuertes. A esas mujeres que, a pesar de las heridas del pasado, siempre encuentran motivos para sonreír. A las que, a pesar de las decepciones, nunca dejan de confiar y de amar.

Reconozco el sacrificio de todas aquellas madres que abandonan sus sueños y metas, que calzan nuestros zapatos, y cuyas vidas se ven reflejadas en nuestros logros y derrotas. Gracias por el amor inmenso que entregan, un amor que sólo puede ser infundido por Dios.

Finalmente, dedico esta tesis también a aquellos a quienes la educación universitaria les fue negada. A aquellos que se vieron forzados a trabajar desde temprano, privados de este derecho.

Este logro es más tuyo que mío, mamá. Gracias por ser la fuerza para escribir cada página.

> ***La ignorancia es la que pone a los pueblos de rodillas y los hombres de rodillas son más rodillas que hombres. – Tirone González, "Canserbero".***

AGRADECIMIENTO

A mi tía Eugenia, cuya paciencia y atención han sido un faro de guía en los momentos más complicados.

A Alex, quien me enseñó la valiosa lección de no rendirme y luchar hasta el final.

A mi papá, quien a pesar de los problemas ha estado presente.

A mi tío Juan y Lorena, siempre dispuestos a extender una mano cuando más lo necesito, gracias por su constante ayuda.

A Ubita, quien sembró en mí la semilla de los valores y me enseñó a soñar en grande.

A Pachito, por enseñarme a sonreír sin importar las circunstancias.

A Alba, por mostrarme que la justicia no se espera si no se lucha, y por no desaprovechar ninguna oportunidad para enseñarme y corregirme.

A Aldo, quien me acompañó en los buenos y malos momentos, sobre todo en los malos.

A Juanita, por siempre animarme, por encontrar lo bueno en mí y confiar en mí.

A Andrea Bonilla, por esa vocación para enseñar, por su apoyo y empuje a mis sueños, por guiarme en el camino para lograrlos.

A Andy, Mela, Daya, Fabri y Xime, quienes siempre encontraron la manera de ayudarme y empujarme.

Finalmente, agradezco a Dios por bendecirme con una familia y amigos tan maravillosos, y por la salud para lograr este importante hito.

ÍNDICE DE CONTENIDO

ÍNDICE DE GRÁFICOS:

ÍNDICE DE TABLAS

RESUMEN

Este estudio busca explorar y comparar la precisión, complejidad e interpretabilidad de los modelos econométricos tradicionales y modelos de Machine Learning (Aprendizaje Automático) en la predicción de un indicador económico clave: el volumen de crédito de los hogares en Ecuador. La elección de esta variable se justifica por su importancia en la literatura económica reciente, donde se destaca su papel como un canal influyente sobre los ciclos económicos. El análisis compara dos modelos muy conocidos en el área de ML, random forest y Xtreme Gradient Boosting (XGB), con los modelos econométricos auto regresivos AR, en particular ARIMA y ARIMA(X). La comparativa no sólo considera una discusión general, sino también una experimentación empírica, y aborda las diferencias entre los modelos en tres fases: la preparación de los datos, el modelado y la evaluación de resultados. Además, se analiza la incorporación de predictores no convencionales, como las tendencias de búsqueda de Google relacionadas con el volumen de crédito. Finalmente, se examina el desempeño de los modelos en los ciclos económicos. Los resultados iniciales demuestran una notable superioridad del modelo random forest en términos del Error Porcentual Absoluto Medio (MAPE), logrando alcanzar un error 1,5 puntos porcentuales menor al mejor modelo econométrico, ARIMA, cuyo MAPE fue de 11,6%. La incorporación de predictores no tradicionales (Google Trends) contribuyó a reducir el MAPE de los modelos multivariados, cerca de 1,3 puntos porcentuales. Asimismo, se evidenció el potencial de las herramientas explicativas disponibles para abrir la "caja negra" de los modelos de ML, permitiendo identificar las de forma visual las relaciones no lineales entre variables, así como describir la importancia de estas últimas para el modelo. De esta forma el presente estudio busca contribuir a la discusión acerca de la aplicabilidad de los modelos de ML en la predicción de variables económicas, tarea común en una ciencia social como es la economía, con una inherente naturaleza no determinística.

ABSTRACT

This study aims to explore and compare the accuracy, complexity, and interpretability of traditional econometric models and Machine Learning (ML) models in predicting a key economic indicator: household credit volume in Ecuador. The choice of this variable is justified by its significance in recent economic literature, highlighting its role as an influential channel on economic cycles. The analysis compares two well-known ML models, namely random forest, and Xtreme Gradient Boosting (XGB), with autoregressive econometric models, particularly ARIMA and ARIMA(X). The comparison encompasses not only theoretical discussion but also empirical experimentation, addressing differences between the models across three phases: data preparation, modeling, and result evaluation. Additionally, the incorporation of unconventional predictors, such as Google search trends related to credit volume, is examined. Finally, the performance of the models across economic cycles is scrutinized. Initial results exhibit a notable superiority of the random forest model in terms of Mean Absolute Percentage Error (MAPE), achieving an error 1.5 percentage points lower than the best econometric model, ARIMA, with a MAPE of 11.6%. The inclusion of non-traditional predictors (Google Trends) helped reduce the MAPE of multivariate models by nearly 1.3 percentage points. Furthermore, the potential of explanatory tools is demonstrated in unveiling the 'black box' of ML models, enabling the visual identification of nonlinear relationships among variables, as well as describing the significance of these variables for the model. In this manner, the present study aims to contribute to the discourse regarding the applicability of ML models in predicting economic variables—a common task in a social science such as economics, characterized by its inherent non-deterministic nature.

INTRODUCCIÓN

Tradicionalmente en economía, los métodos econométricos han sido los pilares del análisis cuantitativo. Estos métodos se basan en la teoría económica y estadística, utilizando pruebas de hipótesis y modelos paramétricos para comprender y predecir fenómenos económicos (Boelaert & Ollion, 2018). No obstante, el auge de datos y los avances en la computación en décadas recientes han dado lugar al surgimiento de campos de estudio de vanguardia, como es el Machine Learning (ML), una rama de la Inteligencia Artificial (AI). En este contexto, Varian (2014) identifica tres factores claves: la necesidad de manejar conjuntos de datos más grandes, la creciente cantidad de predictores que puede requerir técnicas de selección de variables, y la naturaleza flexible de los datos que supera las capacidades de los modelos lineales simples. De esta manera, el ML emerge como una herramienta valiosa para lidiar con estas complejidades y desafíos, permitiendo un análisis más preciso y eficiente de los datos económicos.

A pesar de las ventajas aparentes del ML, la intersección entre el ML y las ciencias sociales, entre ellas la economía, no ha sucedido con la misma rapidez a la vista en otros campos. Esto se debe a factores como limitaciones en la disponibilidad de datos, ya que a menudo los agregados macroeconómicos suelen presentar una cantidad de datos reducida debido a su baja periodicidad (anual, trimestral o mensual). Sin embargo, autores como Medeiros et al. (2021) ilustran cómo el ML puede aplicarse satisfactoriamente en un contexto macroeconómico, en el pronóstico de la inflación. En la misma línea, Chen & Baker (2020) demuestran que el ML supera las limitaciones de modelos tradicionales al predecir índices hipotecarios más precisos a niveles de industria y hogar. A pesar de ello, Boelaert & Ollion (2018) plantean cuestionamientos significativos sobre la aplicabilidad del ML en la economía, destacando temas como la relación entre datos y teoría, la dificultad de demostrar la optimalidad de algoritmos altamente flexibles y la disyuntiva entre complejidad e interpretación de modelos. A esto se suma la afirmación de (Mullainathan & Spiess, 2017) quien menciona que una de las mayores limitantes para la aplicabilidad del ML en las ciencias sociales es la dificultad de interpretación del modelo, también llamada "caja negra".

Frente a esta problemática, este trabajo explora y compara las herramientas de la econometría tradicional y el ML en un contexto macroeconómico, en la predicción de un indicador económico relevante: el volumen de crédito de los hogares en Ecuador.

Agregado que, en la literatura económica contemporánea, ha adquirido gran relevancia, pues se ha ido consolidando como un factor influyente en la dinámica económica, considerándolo un canal clave que incide de manera importante sobre la dinámica económica y por consecuencia sobre los ciclos económicos. Además, se explora algunas de las herramientas disponibles para la interpretación del modelo, lo que es particularmente valioso en la economía, donde además de la predicción, la interpretación del resultado es crucial.

Además, replicando a Stavinova, Timoshina, & Chunaev (2021) se considerará la inclusión de predictores no convencionales, información proveniente de un motor de búsqueda de internet, en este caso las tendencias de búsqueda de Google (Google Trends) relacionadas con el volumen de crédito. La idea detrás de este enfoque es revelar las preferencias y necesidades cambiantes de las personas con respecto al crédito, que antes solo se podían inferir indirectamente. Las búsquedas en línea pueden actuar como un reflejo temprano de las intenciones de las personas, ofreciendo una nueva dimensión de datos que enriquece los modelos predictivos (Choi & Varian, 2009).

La investigación realiza una comparación general y empírica entre los modelos tradicionales de econometría, como ARIMA y ARIMA(X), y dos modelos de ML ampliamente reconocidos: random forest y Xtreme Gradient Boosting (XGB). Se analizan tres fases clave: la preparación de los datos, el modelado y la evaluación e interpretación de resultados, además se explora la inclusión de predictores no tradicionales. Finalmente se evalúa el desempeño de los modelos en los ciclos económicos. El objetivo es comprender si el ML, con su enfoque empírico y no paramétrico, puede superar las técnicas tradicionales de econometría en pronósticos precisos en un contexto donde los datos económicos son limitados. A través de esta investigación, se espera arrojar luz sobre la utilidad relativa y precisión de estos enfoques y así proporcionar información útil para alimentar la discusión acerca de la aplicabilidad del ML en la esfera económica.

Capítulo 1: Marco Teórico

El capítulo de revisión de la literatura de esta investigación se adentra en dos áreas de investigación cruciales que sustentan el propósito de este estudio, a saber, (i) la importancia del crédito de los hogares y los ciclos económicos y (ii) el potencial de las técnicas de Machine Learning (ML) en la previsión económica.

Por un lado, la importancia del crédito de los hogares y los ciclos económicos, según Mian, Sufi, & Verner (2017), es un campo emergente que está ganando relevancia, ya que permite comprender los mecanismos que impulsan las fluctuaciones en la economía, y tiene potencial de predecir y mitigar futuras crisis financieras. En este contexto, Beck et al. (2012) encuentran que el crédito a las empresas es el que impulsa el efecto amortiguador que el desarrollo del sector financiero tiene sobre la desigualdad de ingresos. Así también, Mian et al., (2017) analizan el impacto del endeudamiento de los hogares en las expectativas económicas, donde sugieren que un aumento en la relación entre la deuda de los hogares y el PIB puede conducir a expectativas de crecimiento demasiado optimistas.

Respecto al potencial del Machine Learning (ML) en la previsión económica, la segunda sección profundiza en las afirmaciones como las de Medeiros, Vasconcelos, Veiga, & Zilberman (2021) al referirse al reciente auge del ML y Big Data (grandes cantidades de datos) en economía como una revolución en la previsión económica. Se revisan varios estudios que han explorado el uso de ML en la previsión económica. Además, se discute la utilidad de los predictores no tradicionales en la mejora de la precisión de las predicciones, donde autores como Stavinova et al (2021) y Haselbeck, Killinger, Menrad, Hannus, & Grimm (2022) encuentran que la incorporación de factores externos relevantes puede mejorar significativamente el pronóstico económico.

Finalmente, este capítulo detalla la metodología para la contrastación, reforzando el objetivo de este trabajo de investigación: realizar un análisis comparativo de modelos econométricos y de ML en la previsión del volumen de crédito, abordándolo desde dos perspectivas: desde sus generalidades y de forma empírica, siguiendo la estrategia de Stavinova, Timoshina, & Chunaev (2021).

1.1. El crédito de los hogares y la dinámica económica.

El estudio del crédito privado (crédito de los hogares) y los ciclos económicos es un campo relativamente nuevo que va teniendo cada vez mayor interés. La literatura reciente va demostrando que las dinámicas del crédito de los hogares permiten comprender los mecanismos que impulsan las fluctuaciones en la economía para, potencialmente, predecir y mitigar futuras crisis financieras, lo que sugiere que los cambios en la deuda de los hogares pueden tener un impacto profundo en la actividad económica general (Mian, Sufi, & Verner, 2017). De hecho, Mian & Sufi (2018) señalan que, la relación entre el endeudamiento de los hogares y los ciclos económicos es crucial para entender la economía moderna.

1.1.1. El canal de la demanda de los hogares impulsado por el crédito

En artículos como el de Beck et al. (2012) ya se resaltaba el papel que tiene el crédito sobre el crecimiento económico, sobre todo el crédito de las empresas. Estos autores encuentran que países con ratio deuda empresas/PIB mayor que ratio deuda hogares/ PIB, experimentan mayores crecimientos.

Por otra parte, Mian y Sufi (2018) van más allá en sus conclusiones, al señalar que el volumen de crédito juega un papel fundamental en los ciclos económicos, en particular los que operan a través de la demanda de los hogares. Señalan que la principal regularidad empírica respecto al crédito y el crecimiento económico salió a la luz en La Gran Recesión, ya que cuanto mayor era el aumento del apalancamiento de los hogares antes de la recesión, más severa era la recesión posterior. Esta relación entre la deuda de los hogares y el PIB, el desempleo y el crecimiento del PIB ha sido confirmada por varios estudios (King, 1994; Mian & Sufi, 2010; IMF, 2012; Martin & Phillippon, 2017).

Según Mian & Sufi (2018), los choques de crédito por el lado de la oferta parecen ser absorbidos por el canal de los hogares más que por el de las empresas, argumentando principalmente que, a diferencia de los hogares, las empresas son agentes económicos más racionales y solo piden prestado en tiempos de crisis y no en tiempos de bonanza como sucede en los hogares. De manera que, dada una expansión en la oferta de crédito[1], la demanda de crédito de los hogares incrementa, y por consecuencia gastan más, lo que

[1] Mian & Sufi (2018) llaman una expansión en la oferta de crédito a la situación en donde los prestamistas aumentan la cantidad de crédito o disminuyen la tasa de interés sobre el crédito por razones no relacionadas con los cambios en los ingresos o la productividad de los prestatarios.

a su vez aumenta la demanda agregada de los hogares. Lo cual coincide con lo hallado por Beck et al. (2012) quienes afirman que los préstamos bancarios a las empresas, no a los hogares, son los que impulsan el impacto positivo del desarrollo financiero en el crecimiento económico, señalando al factor crédito empresas-crecimiento económico como la justificación del enfoque que la literatura tradicionalmente ha tenido sobre las finanzas y el crecimiento en el crédito empresarial en contraposición al crédito de los hogares.

En términos de la relación entre la deuda de los hogares y el crecimiento del PIB y el desempleo, Mian & Sufi (2018) encuentran que un mayor aumento en la deuda de los hogares de 1982 a 1989 ve un mayor aumento en el desempleo y una mayor disminución en el crecimiento real del PIB de 1989 a 1992. Además, señalan que la disminución en la oferta de crédito no es un factor externo, sino que es consecuencia del exceso precedente asociado con el incremento inicial de la oferta de crédito, de manera que, ya antes del shock de contracción del crédito, el estado de la economía no es estable.

Mian & Sufi (2018) señalan que detrás del canal de demanda de crédito de los hogares hay tres razonamientos fundamentales, entre ellos:

I. Una expansión en la oferta de crédito, en oposición a los shocks permanentes de ingresos o tecnología, es una fuerza clave que genera expansiones y contracciones de la actividad económica (al menos durante las últimas cinco décadas).
II. Aunque el canal de inversión empresarial también es un factor en cómo las expansiones de crédito pueden afectar los ciclos económicos, el canal de crédito de los hogares parece ser más importante en episodios recientes.
III. La economía se enfrenta a serias dificultades para ajustarse cuando la oferta de crédito se contrae.

El tema en el cual Mian & Sufi (2018) reconocen la mayor incertidumbre es el evento causal de los shocks en la oferta de crédito. Señalan que estos no son factores exógenos, sino que son manifestaciones de eventos como el incremento en la desigualdad en una economía y el rápido incremento en los ahorros en las economías emergentes. Este tipo de eventos generan una acumulación de capital en las economías de origen que luego es inyectada en economías con tipos de interés más altos, generando shocks de crédito.

1.1.2. Relación entre desigualdad, crédito y las expectativas

Beck et al. (2012) encuentran que el crédito a las empresas es el que impulsa el efecto amortiguador que el desarrollo del sector financiero tiene sobre la desigualdad de ingresos, señalando que las finanzas reducen la desigualdad más bien a través de una mejor asignación de capital y una transformación económica que a través de un mayor acceso al crédito.

Por otra parte, Coibion, Gorodnichenko, Kudlyak, & Mondragon (2014) identifican una relación singular, entre la desigualdad y el crédito. En su investigación denotan que la existencia de desigualdad no es sinónimo de mayor deuda de los hogares de bajos ingresos, todo lo contrario. Los hogares con bajos ingresos se encuentran en una zona de baja desigualdad generalmente tiene mayor endeudamiento. La causalidad o la relación entre la desigualdad y el incremento de la deuda se explica por el lado de la oferta, ya que los bancos incrementan la concesión de créditos en zonas de baja desigualdad pues se asocia a un préstamo de menor riesgo, al incrementar las garantías de retorno de dinero. No obstante, conforme la tendencia de desigualdad aumente, la disponibilidad de crédito se reducirá afectando a los hogares a medio y largo plazo.

Un factor importante identificado por Mian et al. (2017) es que, mayores aumentos en la relación entre la deuda hogares/PIB conducen a expectativas de crecimiento demasiado optimistas y errores de pronóstico sugiriendo que los pronosticadores no estén apreciando completamente el impacto del aumento de la deuda de los hogares en el crecimiento económico futuro. Además, mencionan que los acreedores tienden a tener expectativas demasiado optimistas durante los auges, sobreestimando los ingresos futuros y, en consecuencia, el crecimiento económico cuando la expansión en la oferta de crédito termina.

1.1.3. El volumen de crédito y su pronóstico adecuado

Los pronósticos precisos pueden ayudar a los formuladores de políticas a anticipar la gravedad de una próxima recesión. Al comprender y anticipar la dinámica crediticia, los formuladores de políticas pueden medir el riesgo financiero general y el riesgo de crisis financieras. Por ejemplo, si hay un rápido crecimiento en el crédito, podría ser una señal de una economía inestable y puede indicar una próxima crisis financiera, haciendo que contar con un modelo de pronóstico confiable para el volumen de crédito, una herramienta crucial para las intervenciones de política oportunas. Esta previsión no solo es importante

para la política interna, sino también para las instituciones internacionales que supervisan la estabilidad financiera mundial e incluso a nivel de personas y empresas, permitiéndoles ajustar el nivel de endeudamiento de una forma económicamente racional (Levieuge, 2017).

1.2. Potencial del Machine Learning en la previsión económica

A casi 64 años de su nacimiento[2], el ML ha demostrado ser una herramienta valiosa en varios campos de la ciencia, desde la medicina hasta la ingeniería, ofreciendo la capacidad de analizar grandes cantidades de datos y extraer patrones útiles para la toma de decisiones. Hace más de tres décadas, White (1988) publicó un artículo que involucraba una aplicación de redes neuronales (NN) para pronosticar los rendimientos diarios de las acciones de IBM, marcando uno de los primeros intentos de incorporar el aprendizaje automático en el ámbito económico. Desde entonces, la aparición del ML en Economía aumentó constantemente, aunque su aplicación ha estado plagada de desafíos. A pesar de esto, existe un creciente interés en su aplicación. En esta sección se revisarán algunos estudios relevantes para entender el potencial del ML en la previsión económica, así como las limitaciones y desafíos que presenta su aplicación en este campo.

Inicialmente, el ML se aplicó para pronosticar series temporales financieras donde la disponibilidad de conjuntos de datos es extensa. Los modelos de ML de esa época requerían, para un entrenamiento eficiente, amplios conjuntos de datos que no existían en otras áreas de la economía. Además, el entrenamiento requería mucho tiempo debido a la, comparativamente, baja potencia de procesamiento de los ordenadores de la época. Hoy en día, el uso del ML no está necesariamente limitado a conjuntos de datos excesivamente grandes, por lo que es una vía interesante y muy prometedora en la previsión económica. Este es el caso no sólo de los problemas financieros sino también de las aplicaciones macroeconómicas o microeconómicas donde los conjuntos de datos tienen un tamaño inherentemente limitado (Gogas & Papadimitriou, 2021).

En el ámbito de las predicciones económicas y financieras, varios autores han aplicado el ML para pronosticar diferentes aspectos económicos. Por ejemplo, Soybilgen

[2] El término Machine Learning (ML) fue introducido por Arthur Samuel mientras trabajaba para IBM en 1959, principalmente para describir las tareas de reconocimiento de patrones que proporcionaban el componente de "aprendizaje" en los entonces pioneros sistemas de Inteligencia Artificial (IA) (White, 1988).

& Yazgan (2021) y Yoon (2021) se enfocaron en predecir el crecimiento del PIB de países como Estados Unidos y Japón, así como los precios de Bitcoin, utilizando modelos basados en árboles, Random Forests y Gradient Boosting. Bouri, Gkillas, Gupta, & Pierdzioch (2021) investigaron el considerar Bitcoin como una opción de cobertura con técnicas de ML, y Yilmaz & Arabaci (2021) realizaron comparaciones entre modelos de ML para pronosticar tipos de cambio.

En el análisis de texto, se ha explorado cómo las características textuales de las minutas de la FED y las noticias financieras pueden ayudar en la predicción económica. Lima, Godeiro, & Mohsin (2021) investigó las características textuales de las minutas de la FED, mientras que Duarte, Montenegro González, & Cruz (2021) combinó noticias financieras y datos históricos para predecir pérdidas financieras en el mercado de valores brasileño.

En el campo de la ventas hortícolas, Haselbeck, Killinger, Menrad, Hannus, & Grimm (2022) aportan pruebas empíricas de que los modelos de ML superan a los métodos de previsión clásicos. Comparando el rendimiento de nueve modelos de ML y tres algoritmos de pronóstico clásicos para las predicciones de ventas hortícolas. Los métodos de ML fueron superiores en todos los experimentos, teniendo a XGB, como el mejor modelo en 14 de 15 comparaciones.

Un factor importante cuando se habla de ML es la eficiencia de sus modelos cuando se trabaja con datos que tienen muchos predictores. Si bien, hace algunas décadas esto no era un requisito primordial para un modelo debido a la limitada información disponible, el escenario actual es muy distinto y está caracterizado por una enorme cantidad de información, la cual, gracias al ML, se aprovecha de manera efectiva (Varian, 2014).

En este contexto, Haselbeck, Killinger, Menrad, Hannus y Grimm (2022) encontraron que incluir factores externos adicionales en los modelos de ML conduce a una mejora significativa en el pronóstico. Por lo tanto, sugieren que los modelos de ML pueden mejorarse mediante la incorporación de factores externos relevantes, que pueden proporcionar un contexto adicional y mejorar la precisión de las predicciones. Así lo confirma Medeiros et al., (2021) cuando al pronosticar la inflación, encuentran que

algunos modelos, como el modelo random forest, pueden producir pronósticos más precisos que los modelos econométricos tradicionales, sobre todo si el entorno de datos es abundante en número de predictores.

En el ámbito de los estudios de causalidad, Mele & Magazzino (2021) utilizaron métodos de ML para identificar relaciones causales entre la contaminación per cápita y el número de muertes por COVID-19. Utilizando algoritmos de ML junto con enfoques econométricos, obteniendo resultados acerca de la causalidad muy similares en ambos enfoques.

Autores como (Chakraborty, Chakraborty, Biswas, Banerjee, & Bhattacharya, 2021) demuestran que un enfoque híbrido puede tener mejores resultados, los autores combinan ARIMA y redes neuronales para predecir la tasa de desempleo de varios países. Así también en un artículo que analiza la competencia de predicción M4[3], Makridakis, Spiliotis, & Assimakopoulos (2020) indican que los métodos de ML que combinan características estadísticas y de ML tienden a tener un rendimiento superior, con una mejora en la precisión a medida que aumenta la complejidad del modelo.

Es así que, mientras autores como Medeiros, Vasconcelos, Veiga, & Zilberman (2021) quienes sugieren que el reciente auge de ML y Big Data[4] en economía no es una moda pasajera, sino un desarrollo significativo que podría revolucionar la previsión económica, autores como Boelaert & Ollion (2018) señalan que si bien es claro que el ML ofrece una alternativa a los métodos de cuantificación clásicos, un escenario donde el ML reemplace completamente a los métodos estadísticos tradicionales aún es muy lejano y poco probable al señalar que el ML, al ser un tema tan novedoso, ha sido sobreestimado, y ha llegado a convertirse en una moda, así como sucedió con el término "Big Data", donde, los investigadores, con afán de obtener un mayor prestigio en sus investigaciones se han visto tentados a añadir el término "Machine Learning".

[3] Los Concursos Makridakis, también conocidos como concursos M, son una serie de concursos abiertos para evaluar y comparar la precisión de diferentes métodos de pronóstico de series temporales (M Forecasting Competitions, 2018).

[4] Big Data se define como conjuntos de datos extremadamente grandes que se analizan computacionalmente para revelar patrones, tendencias y asociaciones, especialmente en relación con el comportamiento y las interacciones humanas (Varian, 2014).

1.3. Análisis comparativo de modelos econométricos y de Machine Learning en la previsión del volumen de crédito

Para la comparación general, se adoptará el enfoque del artículo desarrollado por Boelaert & Ollion (2018) para describir los métodos. Este enfoque implica una revisión detallada de las características fundamentales de cada método, sus supuestos subyacentes, y las fortalezas y debilidades inherentes a cada uno. Además, se analiza cómo estos métodos pueden ser aplicados en un contexto de datos limitados y las implicaciones que esto puede tener para su interpretación y utilidad. Con este análisis general se busca entender las diferencias fundamentales entre los modelos econométricos y de ML, y cómo estas diferencias pueden influir en su rendimiento en diferentes tareas de predicción.

Desde la perspectiva práctica, se seguirá una estrategia similar a la de Stavinova, Timoshina, & Chunaev (2021). Esta estrategia implica la implementación y evaluación de varios modelos de ML, como random forest y XGB, así como modelos econométricos como ARIMA y ARIMAX. Estos modelos serán aplicados a un conjunto de datos relevante y sus resultados serán comparados al uso de métricas como el MAPE, RMSE y la desviación estándar. Este análisis práctico permitirá una evaluación directa de cómo los modelos utilizados para la contrastación (ver tabla 2.2) se desempeñan en la práctica.

Como parte de la experimentación práctica se evaluará el rendimiento de los modelos al incluir predictores externos adicionales, siguiendo las recomendaciones de Medeiros et al. (2021) y Stavinova et al. (2021). Además, para entender de mejor manera la dinámica del volumen de crédito, se analizará el desempeño de los modelos dentro de los ciclos económicos. A través de este enfoque, se espera obtener aportar pistas para una comprensión más profunda de los modelos de ML y su aplicabilidad en el campo de la economía.

Capítulo 2: Contrastación General

El desarrollo tecnológico de las últimas décadas ha dado lugar a un vertiginoso incremento en la capacidad computacional, lo cual ha sido clave para el desarrollo del ML, pues si bien sus primeros algoritmos aparecen ya hace más de 50 años, no es hasta hace dos décadas que el área explota y se vislumbra su enorme potencial. Desde entonces, los avances tecnológicos se han visto impulsados por algoritmos de ML los cuales se destacan por su habilidad para manejar grandes volúmenes de datos, ya sea en términos de predictores o registros. Estos algoritmos ofrecen una precisión superior en tareas predictivas, además de una gran flexibilidad y la capacidad de identificar relaciones complejas no lineales, entre otras características (Medeiros et al., 2021). Gracias a estos atributos, el ML ha logrado establecerse sin discusión en prácticamente todos los campos científicos. Sin embargo, en áreas como las ciencias sociales el ML aún no ha logrado consolidarse y establecerse.

Los desafíos que han impedido el establecimiento del ML en áreas sociales, de forma específica en la Economía, son varias, entre las más relevantes tenemos que en la Economía, debe existir una coherencia entre la teoría y los datos, por lo que las pruebas puramente empíricas no son suficientes. Además, el propósito de un análisis económico va más allá de predecir con precisión un evento, tarea en la que el ML ha demostrado gran capacidad, sino más bien entender la causalidad. En este sentido, el ML ha hecho esfuerzos por desarrollar herramientas que cumplan con estas tareas, no obstante, no han conseguido desplazar a las herramientas econométricas desarrolladas hace casi un siglo (Boelaert & Ollion, 2018).

Varios autores entre ellos Charpentier, Flachaire, & Ly (2018) y Boelaert & Ollion (2018), han demostrado que dentro de la economía existen ciertas tareas y contextos en los cuales los algoritmos de ML pueden ser buenos sustitutos a los modelos econométricos, sobre todo cuando el conjunto de datos tiene características como gran tamaño, alta dimensionalidad o relaciones no lineales, situaciones en las que los modelos econométricos demuestran limitaciones. Si bien ambos enfoques tienen fortalezas y debilidades, autores como (Varian, 2014) y (Athey & Imbens, 2019) resaltan la necesidad de que el profesional en econometría tenga entre sus herramientas, conocimiento de

modelos de ML por los beneficios que estos puede tener, de forma que la elección de un enfoque u otro sea determinada por el contexto y los datos disponibles y no por la falta de herramientas del analista.

En este capítulo, se explora y se compara la diferencia entre los modelos de ML y los modelos econométricos y se analiza las fortalezas y debilidades de cada enfoque. A través de esta exploración, se busca comprender su terminología, sus premisas, su lógica, los objetivos que persiguen y las principales diferencias metodológicas en su aplicación en el contexto de esta investigación, es decir, en una tarea predictiva de un valor continuo. Así también, se revisa brevemente cómo el contrastar la Econometría con el ML conduce a analizar las culturas del modelado[5] mencionadas por Breiman (2001). Finalmente, se presentarán los modelos utilizados en la comparación empírica realizada en el siguiente capítulo, así como las medidas de evaluación de los modelos.

Este hilo de ideas permitirá una mejor comprensión de cuál de los enfoques contrastados es más apropiado para diferentes contextos y cómo se pueden utilizar de manera efectiva para maximizar la precisión y la utilidad de los modelos predictivos y analíticos en la previsión económica.

Antes de empezar con la comparación es importante recordar que el ML forma parte de la Inteligencia Artificial, que es a su vez un campo de las ciencias de la computación, por lo que su aplicación se ha extendido hacia todo tipo de problemas, desde problemas que involucran la predicción de valores continuos, clasificación de texto, imágenes, procesamiento de lenguaje natural, hasta problemas más complejos como el aprendizaje de refuerzo, por mencionar algunas[6]. Sin embargo, debido a que el objetivo de esta investigación es la predicción de una serie de tiempo, el alcance se limita únicamente a la parte del ML relacionada con la predicción de un valor continuo, categorizado, dentro del ML, como un caso especial de un problema regresivo de aprendizaje supervisado[7], especial porque se toma en cuenta la temporalidad. Es así como,

[5] Se refiere a las dos culturas de modelamiento mencionadas por Breiman (2001b), explicado con mayor profundidad en la subsección 2.2.3

[6] En el Anexo 2 se incluye un diagrama con un breve resumen de cada uno de estos problemas, aplicaciones más comunes y los modelos disponibles para cada tarea. El diagrama fue realizado por (Cotton, 2022)

[7] El diagrama usado para clasificar el problema se encuentra en el Anexo 2, el cual se ha basado en la clasificación realizada por (Scikit-Learn, 2022a).

en adelante, cuando se mencione el término ML se hará referencia únicamente a esta parte del ML, dejando de lado las demás clasificaciones (ver al Anexo A.1).

2.1. Contraste general econometría y Machine Learning

2.1.1. Sobre la terminología

Los contrastes entre el ML y la Econometría empiezan desde su terminología. Muchas de estas diferencias responden a la disciplina de la cual cada área proviene, en el caso de ML, de las Ciencias de la Computación y Estadística, mientras que la Econometría, tiene en sus raíces a la Economía y la Estadística. Su terminología refleja las diferencias fundamentales en el enfoque y método de cada campo, y pueden ser una barrera para la comprensión y colaboración entre ambos campos, por lo que resulta pertinente un breve repaso de los términos más relevantes en el contexto de la presente investigación. La Tabla 2.1 resume varios términos de ML y su equivalente estadístico, con una breve explicación necesaria debido a que en algunos casos la similitud no es precisa sino relativa y depende del contexto en el que se utilice.

Tabla 2.1. Términos equivalentes en estadística y ML

ESTADÍSTICA	Machine Learning	OBSERVACIÓN
dato, registro o fila	ejemplo, instancia	Las dos disciplinas usan el término "observación" para referirse a un valor o un vector de características dependiendo del contexto.
variable respuesta, variable dependiente, variable endógena	label (etiqueta), output variable (variable de salida)	Ambas disciplinas usan también el término "variable objetivo"
variable independiente, variable exógena	vector de características (feature), input, predictor,	El término "variable independiente" se usa tradicionalmente en estadística, a pesar de que generalmente una variable dependa de otras
pronóstico de una variable continua	regresión	
pronóstico de una variable categórica	clasificación	

Estimación	entrenamiento	Ambos términos hacen referencia a la predicción de un “output” a partir de un “input” (Kurtz, 2018).
Hipótesis	hipótesis	En ambas disciplinas una hipótesis es una afirmación científica que debe ser probada. Pero en ML puede hacer referencia a la regla de predicción que se obtiene a partir de un algoritmo de clasificación (Kurtz, 2018).
sesgo	sesgo	Es estadística, el término hace referencia al error o diferencia entre el valor esperado de un estimador y su valor real. En ML el término se utiliza para describir el grado de complejidad y los supuestos que necesita un modelo con respecto a su capacidad de predicción (James, DanielaWitten, Hastie, & Tibshirani, 2013).
Varianza	varianza	En Estadística, se usa el término para referirse a la medida de dispersión como tal, pero en ML se puede utilizar para hacer referencia a la sensibilidad que tiene un modelo cuando se utiliza diferentes datos a los utilizados para entrenar el modelo (Java T Point, 2022).
Muestra	conjunto de entrenamiento o de prueba	El término "muestra" utilizado en Estadística obedece al enfoque inferencial de esta disciplina (Cook, 2010).
Parámetro	parámetro, hiperparámetro, pesos	En econometría, el término parámetro generalmente se utiliza para referirse a los coeficientes estimados por un modelo. En ML existen dos tipos de

		parámetros: los unos que se puede inicializar y actualizar a través del proceso de aprendizaje de datos (p. ej., los pesos de las neuronas en las redes neuronales), llamados parámetros del modelo; mientras que los otros, denominados hiperparámetros, no se pueden estimar directamente a partir del aprendizaje de datos y se deben configurar antes de entrenar un modelo de ML porque definen la arquitectura del modelo. Sino que es definido ex ante mediante un proceso de prueba y error (Yang & Shami, 2020). A este proceso se le conoce como optimización de hiperparámetros (Ver sección 3.5.3) El término pesos o "weights" se asocia a los parámetros desconocidos de una red neuronal. Estos valores se determinan de manera que se ajusten a los datos de entrenamiento.(Hastie, 2009, p. 395).

Realizado por: Christian Mateo Quiguiri Daquilema.

2.1.2. Sobre el objetivo:

El objetivo de la econometría puede ser definido siguiendo a Wooldridge (2016) quien en su libro de Introducción a la Econometría menciona:

> La econometría se define como la ciencia social en la cual las herramientas de la teoría económica, las matemáticas y la inferencia estadística se aplican al análisis de los fenómenos económicos.

Para lograrlo, de acuerdo con Athey & Imbens (2019), se hace uso de un modelo estadístico paramétrico que describe la distribución de un conjunto de variables. Es así como, dada una muestra aleatoria de una población, se estiman los parámetros de interés encontrando los valores que mejor se ajustan a la muestra, utilizando para esto técnicas como maximización de la verosimilitud o minimización de errores de estimación, encontrando el mejor modelo posible dado un conjunto de datos. Una vez obtenidos estos parámetros, la atención se centra en la calidad de estos, los coeficientes y los errores estándar, comúnmente incluyendo también intervalos de confianza.

Por otra parte, el ML resume su objetivo en desarrollar algoritmos para hacer predicciones precisas sobre unas variables dado otras. Para alcanzar este objetivo, el ML hace uso de las propiedades de sus disciplinas fundadoras, por un lado, su base estadística le permite crear modelos que lograr capturar la estructura de un conjunto de datos, mientras que las ciencias de la computación le dan la capacidad de generar algoritmos eficientes que le permiten resolver problemas complejos de optimización. Es así como el ML genera un modelo altamente flexible y complejo que encuentra aproximaciones de la verdadera relación subyacente entre los predictores y la variable objetivo. No obstante, a diferencia de la econometría, los resultados que se reportan son únicamente medidas de precisión y poco o nada acerca de las propiedades que un modelo debe tener para que sus resultados sean válidos (Boelaert & Ollion, 2018).

2.1.3. Las dos culturas de modelamiento de datos

Cuando se trata de usar modelos estadísticos para llegar a conclusiones acerca de los datos existen principalmente dos enfoques, inductivo y deductivo, de los que hacen uso el ML y la econometría respectivamente (Boelaert & Ollion, 2018). Si bien este puede parecer un problema filosófico, tiene gran relevancia cuando de aplicar un modelo estadístico sobre un conjunto de datos se trata, pues se relaciona directamente con la validez de las conclusiones que se pueden obtener de estos.

El estadístico Leo Breiman, en su influyente artículo "Statistical Modelling: The Two Cultures" (Breiman, 2001), subrayó la importancia de estos dos enfoques, advirtiendo sobre los riesgos de una cultura estadística que se concentra más en la estructura de los modelos que en los datos en sí mismos. Breiman anticipó que este enfoque podría dar lugar a conclusiones erróneas, a pesar de que la literatura estadística

predominante a menudo lo presenta como incuestionable. Este artículo reavivó el debate entre los razonamientos inductivo y deductivo en la modelización estadística.

Breiman (2001) dividió los enfoques en dos categorías: la cultura de los Modelos de Datos y la cultura Algorítmica. Los modelos estadísticos tradicionales, utilizados en la Econometría, forman parte de la primera cultura, que se basa en la suposición de que los datos son generados por un modelo de datos definido. Por otro lado, la cultura algorítmica, en la que se enmarca el ML, asume que la relación subyacente en los datos es compleja y completamente desconocida, generalmente no impone relaciones entre predictores y resultados, ni aísla el efecto de una sola variable.

Breiman (2001) argumentó que la cultura de los Modelos de Datos se enfoca en obtener conclusiones sobre el mecanismo del modelo en lugar de entender el mecanismo natural del fenómeno de estudio. Esto implica que, si el modelo resultante es una mala representación de la realidad, las conclusiones serán inevitablemente erróneas. Por lo tanto, Breiman abogó por el uso de la precisión como la medida más directa y confiable para cuantificar cómo se aproxima un modelo a la realidad del fenómeno de estudio. No obstante, autores como Charpentier et al. (2018) mencionan que la frontera entre ambas culturas es cada vez más borrosa, pues en la intersección encontramos la econometría no paramétrica, o métodos que combinan el ML con la estadística.

En las últimas décadas el ML ha demostrado su potencial, revolucionando casi todas las ramas de la ciencia, no obstante, en áreas sociales como la Economía no ha podido sobresalir como en las otras. Este problema se atribuye a la complejidad que un modelo de ML puede llegar a tener, la cual va a estar directamente relacionada con su interpretabilidad. Esto tiene relevancia ya que, en el análisis de un fenómeno, si bien la predicción es importante, lo es aún más el entendimiento del fenómeno. Respecto a este punto Breiman (2001) menciona que, la predicción y la interpretación no se contraponen, sino que son complementos, y al final un modelo más preciso va a aportar información más real acerca de la naturaleza del fenómeno de estudio. Mullainathan & Spiess (2017) toma en cuenta este punto, mencionando que deben encontrarse preguntas dentro de la economía, para las cuales una previsión precisa si sea importante, en esto coinciden autores como Varian (2014) quien llama a que las nuevas generaciones de econometristas o estadistas tengan una formación en Machine Learning.

2.1.4. Fundamentos matemáticos versus eficacia empírica

Uno de los principales contrastes entre el ML y la Econometría tiene que ver con el respaldo teórico que cada una tiene como base. Siguiendo a las definiciones de Gujarati & Porter (2009), la Econometría puede ser entendida como estadística matemática y teoría económica aplicada sobre datos económicos. En consecuencia, hace uso de modelos paramétricos[8] ampliamente conocidos por la estadística, como la regresión lineal y ARIMA, que tienen bases sólidas en la estadística y teoría probabilística. La Teoría Asintótica[9] – que incluye: expansiones de Taylor, Ley de los grandes números, Teorema del Límite Central, entre otros – permite que, una vez estimado un modelo, sea posible obtener estimadores con características ideales, a partir de los cuales se puede hacer inferencias sobre la población de la que proviene la muestra utilizada para entrenar el modelo (Charpentier et al., 2018).

Por el contrario, los artículos y manuales metodológicos desarrollados por el ML se enfocan sobre todo en explicar cómo funciona el algoritmo usado para generar el modelo y no en probar propiedades teóricas para validar sus resultados. Si bien, la matemática y estadística están presente, sus lazos son más débiles. De hecho, muchos de los algoritmos usados en ML son pobremente entendidos desde un punto de vista matemático. Un ejemplo lo encontramos en el modelo random forest, que, a pesar de su gran precisión y facilidad de uso, la dificultad de expresar el algoritmo completo en una expresión matemática manejable hace que no pueda ser entendido del todo (Boelaert & Ollion, 2018). De acuerdo con Mullainathan & Spiess (2017), expresado en manera concisa, el ML pertenece a las herramientas que se preocupan más por el $\hat{y}$ (output) de un modelo, que por el comportamiento de $\hat{\beta}$ (estimador).

2.1.5 Aleatoriedad y optimalidad

Un aspecto importante e inherente a los algoritmos de ML, que dificulta la tarea de matemáticamente encontrar una solución óptima como la Econometría, es el rol de la aleatoriedad pues está presente en casi todas las fases entrenamiento de un modelo de

[8] Hacen referencia a los métodos que asumen que la información acerca de la distribución de la población es conocida y está basada en un conjunto fijo de parámetros(Wackerly, Mendenhall, & Scheaffer, 2008). Además hace referencia a que el número de parámetros del modelo se mantiene fijo con respecto al tamaño de la muestra (Haitao, 2017).

[9] Es un marco bajo el cual se evalúan propiedades de estimadores y pruebas estadísticas. Dentro de este marco se asume que el tamaño de la muestra puede crecer indefinidamente $n \to \infty$ (Wackerly et al., 2008).

ML. Desde el orden en el que las observaciones ingresan al modelo, el proceso de seleccionar el subconjunto de datos que ingresará al algoritmo, la inicialización del algoritmo utilizado, la validación cruzada, en la división del conjunto de datos para evaluar el desempeño del modelo. De esta manera se puede afirmar que el ML en la práctica es estocástico (Brownlee, 2014).

La aleatoriedad implica que sea imposible reproducir los resultados de un modelo de ML a pesar de que se use el mismo algoritmo con los mismos datos. Por consiguiente, se afirma que el ML no ofrece soluciones cerradas u óptimas y su naturaleza estocástica hace que, cuando el algoritmo llega a una solución, esta no sea un óptimo global sino más bien un óptimo local. En la práctica, es usual fijar una semilla[10] a partir de la cual se genera la aleatoriedad, así los resultados de un modelo pueden ser replicados (Mullainathan & Spiess, 2017).

2.1.6 Modelamiento definido y formas flexibles

Para un modelo econométrico, la especificación del modelo, las transformaciones de las variables y sus potenciales interacciones deben ser especificadas ex ante, guiadas por la teoría económica. Un factor sustancial que permite que los modelos ML en general sean más precisos es la flexibilidad y complejidad que estos pueden llegar a alcanzar. El ML tiene un procedimiento que funciona a la inversa, se parte de tratar a las relaciones entre variables como completamente desconocidas y entrena, al uso de un algoritmo un modelo flexible y completo que se ajusta a los datos. Esto no quiere decir que el investigador no debe tomar decisiones al ajustar un modelo de ML, pues ciertos parámetros que definen la complejidad del modelo si son ajustados ex ante mediante la optimización de parámetros (Charpentier et al., 2018).

Esta característica permite al ML soportar la premisa del *Teorema de Aproximación Universal* que afirma que, de existir alguna relación en los datos, entonces un modelo de ML es, al menos en teoría, capaz de encontrarla y reproducirla, dejando de lado restricciones inherentes a los modelos paramétricos, la más relevante, la monotonía (Boelaert & Ollion, 2018).

[10] Establecer una semilla significa inicializar un generador pseudoaleatorio. El prefijo pseudo hace referencia a que a partir de la semilla el generador aleatorio obtendrá la misma salida de números cada vez que desea generar números aleatorios. Si no establecemos una semilla, los números pseudoaleatorios generados son diferentes en cada ejecución (Li, 2022).

2.1.7 Sobre los datos

Uno de los términos que más se ha asociado al ML es "Big Data", que enfatiza al cambio en la escala de los datos, o al incremento exponencial que ha habido en la disponibilidad de datos en las últimas décadas y que el ML ha sabido manejarlo. No obstante, el término esconde el hecho de que, además de la escala, ha existido un cambio en la naturaleza de los datos. Por ejemplo, actualmente se dispone de información satelital, de imágenes, de textos, buscadores, entre otros, todos ellos con una elevada dimensionalidad, que no puede ser manejada por métodos estadísticos tradicionales como los usados en Econometría, pero sí por el ML. Estas nuevas fuentes de datos pueden resultar relevantes para tareas donde la información económica no está disponible o es poco confiable, como sucede en gran parte de los países en vías de desarrollo (Mullainathan & Spiess, 2017).

Si bien, el hecho de que un modelo de ML sea capaz de manejar eficientemente conjuntos enormes de datos supone una ventaja, desde un enfoque pragmático, que para llegar a su máximo desempeño requiera tal cantidad de datos puede suponer una debilidad, sobre todo en el contexto económico, y más específicamente macroeconómico donde los conjuntos de datos llegan difícilmente a sobrepasar los mil registros. Varias investigaciones han demostrado que, mientras los métodos paramétricos usualmente no mejoran su desempeño a partir de unos cuantos cientos de observaciones, los algoritmos de ML requieren gran cantidad de datos para alcanzar el máximo de su capacidad predictiva (Boelaert & Ollion, 2018). No obstante, literatura reciente demuestra beneficios de utilizar métodos de ML aun cuando el número de observaciones es relativamente pequeño. Autores como Medeiros, Vasconcelos, Veiga, & Zilberman (2021) y J. M. Chen & Baker, (2020) evidencian que el aumentar el número de predictores puede jugar un papel importante al mejorar la predicción de un modelo de ML y los segundos evidencian que un número reducido de observaciones no limita la capacidad predictiva de un modelo de ML, demostrando que la naturaleza de los datos y el contexto del fenómeno son muy importante en el desempeño de un modelo, sea este, econométrico o de ML.

2.2 Diferencias Metodológicas

Luego de revisar algunos de los contrastes más marcados entre las dos formas de generar un modelo, queda la interrogante de cómo podemos evaluar la calidad e interpretar la información que podemos obtener de dos tipos de modelos tan diferentes.

2.2.1 Validación y validación cruzada

En la Econometría, la evaluación de un modelo involucra varios aspectos clave. Primero, se analizan las propiedades de los residuos, que incluyen su normalidad, homocedasticidad e independencia. Estos son indicadores importantes de la calidad del modelo, ya que los residuos ideales deben comportarse como ruido blanco.

Después, la significancia estadística es un aspecto importante en la evaluación del modelo, y el P-valor desempeña un papel crucial en esta evaluación. El P-valor se utiliza para realizar contrastes de hipótesis y determinar la significancia de los parámetros estimados. Por lo general, se compara con un nivel de significancia predefinido, que suele establecerse en 0.05. Si el P-valor es menor que este nivel, se considera que el parámetro en cuestión tiene un efecto significativo en el modelo (Boelaert & Ollion, 2018)

Además de las pruebas estadísticas, se realiza una validación de la relevancia económica de las variables. Esto implica examinar los coeficientes estimados en términos de sus signos esperados y magnitud. Se busca asegurar que la significancia estadística sea coherente con la significancia económica y que las relaciones postuladas por la teoría económica sean consistentes con los resultados del modelo (Gujarati & Porter, 2009).

Asimismo, se lleva a cabo una verificación de los supuestos subyacentes al modelo. En el caso de una regresión lineal, se busca confirmar que se cumplan los supuestos de Gauss-Markov, lo que garantiza que los estimadores de Mínimos Cuadrados Ordinarios (OLS) sean insesgados y eficientes. Para modelos de tipo ARMA, se realiza un diagnóstico de residuos formulado por Box-Jenkins para evaluar la adecuación del modelo a los datos (Gujarati & Porter, 2009).

Finalmente, después de realizar estas verificaciones, se puede afirmar que los estimadores son normalmente distribuidos. Esto habilita la construcción de intervalos de confianza para las predicciones, cuyo ancho depende de la desviación estándar de los estimadores (Gujarati & Porter, 2009). En conjunto, estos pasos permiten una evaluación completa de la calidad y la adecuación del modelo econometría."

Esto contrasta de gran manera con el ML ya que carece de medidas similares. En su lugar, el criterio universal de la calidad de un modelo es la predicción, medida

calculada mediante la llamada validación cruzada fuera de la muestra o "out-of-sample cross-validation", una técnica que divide a los datos de entrenamiento en k bloques utilizando $k-1$ bloques para entrenar el modelo y probando su rendimiento en el bloque que queda afuera. De esta forma evita sobreestimar la capacidad de predicción al probar el modelo estimado en el bloque que queda por fuera de los datos de entrenamiento, lo cual asegura obtener comparaciones insesgadas del ajuste. Con esto se puede ver que la calidad de los modelos ML es evaluada sobre el poder de generalización que ofrece el modelo, es decir en la capacidad de predecir correctamente observaciones de otras muestras de la misma población (Athey & Imbens, 2019).

2.2.2 Sobreajuste y regularización

Athey & Imbens (2019) mencionan que el ML está más preocupado por el sobreajuste de un modelo estadístico que la contraparte econométrica, afirmación que, de acuerdo con los autores, responde al objetivo para el que los métodos de ML fueron diseñados, que es la predicción. En lugar de la optimización directa de una función objetivo, como podría ser la minimización de la suma de los residuos al cuadrado (MCO) en una regresión lineal o la maximización del logaritmo de la función de máxima verosimilitud en un modelo ARMA, los modelos de ML añaden un término de regularización a la función objetivo, que penaliza la complejidad del modelo que a su vez está directamente relacionada con el sobreajuste. La intensidad de la penalización, generalmente representada por las letras griegas α o λ, son determinada por los datos, específicamente por el desempeño predictivo fuera de la muestra determinado mediante validación cruzada (Boelaert & Ollion, 2018).

2.2.3 Ajuste de hiperparámetros o "Hyperparameter Tuning"

Cuando un modelo de ML es entrenado, varios parámetros que dependerán del algoritmo utilizado se irán ajustando y variando de acuerdo con los datos que vayan ingresando. Sin embargo, existe otro tipo de parámetros, conocidos como hiperparámetros[11], que no se puede aprender directamente del proceso de entrenamiento regular. Los hiperparámetros son ajustados antes del entrenamiento del modelo como tal. Estos parámetros expresan propiedades importantes del modelo, como su complejidad, la sensibilidad o velocidad de ajuste del modelo a los datos, entre otras.

[11] El significado del término es explicado en la Tabla 2.1 correspondiente a la terminología del ML y la Econometría.

Algunos ejemplos de hiperparámetros del modelo incluyen a la regularización L1[12] o L2[13]. Dependiendo del algoritmo que use, el número de hiperparámetros puede ir de un par a más de 20. Por lo que el objetivo de la optimización de hiperparámetros es encontrar la mejor combinación de estos en términos de una medida de error, como puede ser el RMSE o MSE. Por ejemplo, si queremos ajustar dos hiperparámetros α y λ de un modelo que mide su error por el MSE, la técnica para ajustar estos parámetros consiste en construir muchas versiones del modelo con todas las combinaciones posibles de hiperparámetros y se seleccionará la mejor, es decir la que menor error o MSE (o cualquier métrica similar) presente (Geeks for Geeks, 2022).

2.2.4 Interpretación del modelo estadístico resultante:

Luego de la estimación de un modelo econométrico, la interpretación de este hace parte fundamental del análisis, ya que responde al objetivo mismo de la Econometría, que es entender el fenómeno de estudio, separándose en gran manera con el ML, que se preocupa principalmente por la precisión.

Una vez que se ha ajustado un modelo econométrico, los resultados son interpretados en base a los siguientes puntos:

i. Si un predictor es estadísticamente significante, medido por el *P-valor* de su coeficiente estimado;
ii. El signo o la dirección del efecto del estimador, que permite conocer la relación entre el predictor y la variable objetivo;
iii. La magnitud del efecto, medida por la magnitud del coeficiente estimado.
iv. Finalmente, es posible determinar un intervalo de confianza para la magnitud del coeficiente mediante su desviación estándar.

El análisis de estos resultados permite una interpretación directa acerca del fenómeno que se está estudiando y puede ser usado inmediatamente para hacer inferencias acerca de la población (Gujarati & Porter, 2009).

[12] L1 hace referencia a la regularización de una regresión LASSO el cual agrega el "valor absoluto" del coeficiente como término de penalización a la función de pérdida, funciona como un método de selección de variables al estimar el coeficiente con un valor de 0 para aquellas variables no explicativas para el modelo (Anuja, 2022).
[13] L2 hace referencia a la regularización de una regresión Ridge, que agrega la "magnitud al cuadrado" del coeficiente como término de penalización a la función de pérdida, pero a diferencia de L1 no hace que los coeficientes sean igual a 0, por lo que no es posible hacer selección de variables (Anuja, 2022).

Al contrario que el enfoque econométrico, la interpretación directa de los parámetros en un modelo de ML es muy difícil, desventaja inherente de la enorme flexibilidad que le caracteriza. De nuevo, en el ML el único resultado es la predicción del modelo, característica que ha sido descrita como el mayor impedimento para la plena adecuación del ML en las ciencias sociales. Si se quisiera obtener interpretaciones similares a las de un modelo paramétrico nos encontraríamos con un gran problema que tiene que ver con la naturaleza no paramétrica de los modelos asociada con la ausencia de errores estándar en los coeficientes estimados. De acuerdo con algunos investigadores, el intentar generar estos errores puede ser muy complicado, e incluso imposible, debido a que de entre varios factores, se debería tener en cuenta la selección del modelo por sí misma (la determinación del valor λ del término de regularización de un modelo ML) (Boelaert & Ollion, 2018).

De acuerdo con Mullainathan & Spiess (2017) otro factor que afecta a la interpretación de un modelo de ML es la misma regularización, indicando que promover la selección de un modelo más simple (al usar un término de regularización) para mejorar la generalización del modelo, puede dar lugar a modelos erróneos, ya que el forzar la elección de modelos simples con el afán de mejorar la generalización puede causar que un modelo como por ejemplo; LASSO[14], descarte variables importantes que en realidad puedan explicar el fenómeno de estudio .

2.3 Técnicas y métodos usados en la contrastación empírica

En primer lugar, se revisarán algunos de los modelos econométricos más utilizados para pronosticar series de tiempo. Luego se presentarán dos de los modelos predictivos más utilizados en ML para la predicción de un valor continuo. Estos modelos han sido considerados por algunos autores como algoritmos de vanguardia, conocidos también como "state of the art"[15], por el éxito que han tenido frente a otros algoritmos al aplicarse en determinados problemas. Debemos recordar que, si bien, estos últimos no han sido diseñados para su uso específico en series de tiempo, su flexibilidad y adaptabilidad han permitido su aplicación a series de tiempo, tratándolos como un caso especial de

[14] En Estadística y ML, LASSO hace referencia a un método de regresión que realiza selección de variables y regularización, para mejorar la precisión e interpretabilidad del modelo estadístico resultante (Tibshirani, 1996).
[15] Significa que el método supera a sus competidores en un conjunto particular de problemas.

regresión, donde el muestreo y la validación cruzada que se realizan al ajustar un modelo de ML deben adaptarse a la temporalidad presente en los datos.

2.3.1 Modelos econométricos

2.3.1.1 ARIMA

Para el desarrollo de este modelo se hará uso de la metodología Box Jenkins propuesto por Hyndman & Athanasopoulos (2018).

Para hacer uso del modelo se requiere que la serie o series de tiempo involucradas en el análisis sean estacionarias o que pueda llegar a serlo después de una o más diferenciaciones, ya que el supuesto base para pronosticar al uso de este modelo es la suposición de que las características de la serie son constantes a través del tiempo. El modelo puede ser representado con la siguiente notación:

$$\text{ARIMA}\ (p, d, q)$$

Con p que representa al orden AR no estacional, d el orden de diferenciación no estacional, q el orden MA no estacional. Utilizando el operador de rezagos (L), la representación matemática[16] del modelo es la siguiente:

$$\phi_{\mathrm{p}}(\mathrm{L})(1-\mathrm{L})^{\mathrm{d}}Y_{\mathrm{t}} = \theta_{\mathrm{q}}(\mathrm{L})\varepsilon_{\mathrm{t}} \tag{2.1}$$

Donde los componentes no estacionales son:

$$AR\colon \phi(L) = 1 - \phi_1 L - \cdots - \phi_p L^p \tag{2.2}$$

$$MA\colon \theta(L) = 1 + \theta_1 L + \cdots + \theta_q L^q \tag{2.3}$$

Luego, la predicción de un nuevo punto está dada por:

$$Y_t = \alpha + \phi_1 Y_{t-1} + \cdots + \phi_p Y_{t-p}\epsilon_t + \theta_1 \epsilon_{t-1} + \cdots + \theta_q \epsilon_{t-q} \tag{2.4}$$

[16] La representación matemática del modelo ARIMA se obtuvo del la notación utilizada en el libro de Gujarati & Porter (2009).

El procedimiento general para la aplicación del modelo, de acuerdo con la metodología Box Jenkins sugerida por Hyndman & Athanasopoulos (2018) y Gujarati & Porter (2009) comprende:

1. Exploración: Graficar los datos e identificar observaciones inusuales. Si es necesario, transformar los datos (usando transformaciones Box Cox) para estabilizar la varianza.
2. Identificación: Encontrar los valores apropiados de p, d y q al uso de la función de autocorrelación (FAC), función de autocorrelación parcial (FACP) y los correlogramas resultantes que están en función de los rezagos de la serie. Además, se considera la prueba de Dickey-Fuller para evaluar la estacionariedad de la serie e identificar el orden de diferenciación d.
3. Estimación: Estimar los parámetros de los términos autorregresivos y de promedios móviles incluidos en el modelo.
4. Examen de diagnóstico: Verificar si el modelo seleccionado se ajusta a los datos de forma razonablemente buena contra otras especificaciones, utilizando criterios de información y sobre todo verificar que los residuos del modelo sean ruido blanco[17].
5. Pronóstico: Generación de pronósticos que permitan hacer la evaluación de la capacidad predictiva del modelo.

2.3.1.2 Regresión con Residuos ARIMA (ARIMAX):

A veces la información de la propia serie no es suficiente para explicar su comportamiento, es así como el modelo ARIMA con regresores externos[18] permite la inclusión de variables que pueden ser relevantes para explicar el fenómeno de estudio especificando un modelo similar a una regresión lineal pero expresando los residuos con una representación ARIMA, conservando la facilidad de interpretación de una regresión pero incluyendo las características de un modelo ARIMA Hyndman & Athanasopoulos (2018).

[17] Se refiere a una serie de tiempo que no evidencia autocorrelación o esta es muy cercana a 0, su media es 0 y presentan varianza constante. También es conocido como ruido blanco Gaussiano (Hyndman & Athanasopoulos, 2018).

[18] Regresión con residuos ARIMA, a veces conocido como ARIMAX, es el modelo ajustado cuando se incluyen regresores externos que es la que utiliza los paquetes Forecast y Statsmodels de R y Python respectivamente (Hyndman, 2010).

La representación matemática[19] del modelo es la siguiente:

$$Y_t = \beta_0 + \beta_1 X_{1,t} + \beta_2 X_{2,t} + \cdots + \beta_k X_{k,t} + \eta_t \quad (2.5)$$

Donde $x_{1,t}, x_{2,t}, \ldots x_{k,t}$ son los k regresores externos de la variable y_t; $\beta_0, \beta_1, \ldots \beta_k$ son los coeficientes estimados y η_t es la representación ARIMA de los residuos de la regresión de X sobre Y.

La expresión (2.5) es análoga a la especificación de una regresión lineal común, con la diferencia de que el término de error tiene una representación ARIMA, la cual es expresada como:

$$\eta_t = \phi\eta_{t-1} + \cdots + \phi\eta_{t-p} + \theta_1\varepsilon_{t-1} + \cdots + \theta_q\varepsilon_{t-q} + \varepsilon_t \quad (2.6)$$

Donde ε_t es ruido blanco

A diferencia de modelo ARIMA, donde para el proceso de blanqueamiento de residuos se utiliza la metodología Box Jenkins En un modelo ARIMAX, el proceso de blanqueamiento se refiere a la estimación conjunta de los coeficientes de regresión (β) y los parámetros ARIMA (ϕ, θ), donde se busca garantizar que los residuos del modelo, que incluyen tanto el componente de regresión como el componente ARIMA, sean ruido blanco.

2.3.2 Modelos de Machine Learning

Dado el objetivo y el alcance de esta investigación, los métodos de ML que se han incluido aquí se enfocan en los algoritmos y técnicas para problemas de regresión, es decir, cuando la variable objetivo es numérica y continua. Esta aclaración es necesaria puesto que las técnicas de ML a continuación descritas pueden ser usadas tanto para problemas de clasificación como de regresión.

2.3.2.1 Árboles de Regresión y Clasificación (CART):

De acuerdo con Hastie (2009), un árbol de regresión es un algoritmo que divide el espacio en dos partes de forma recursiva, obteniendo un conjunto de regiones rectangulares, luego, ajusta un modelo simple como una constante, para cada región. Es decir, dado un

[19] Esta representación matemática ha sido obtenida del libro "Forecasting principles and practice" de (Hyndman & Athanasopoulos, 2018). La cuál coincide con el modelo estimado por el paquete Statsmodels utilizado en la contrastación empírica que se detallarán en el capítulo 3.

conjunto x_i de predictores de una variable objetivo y_i con N observaciones: $(x_i, y_i),\ i = 1,2,\dots,N$, con $x_i = (x_{i1}, x_{i2}, \dots, x_{ik})$, el algoritmo decide la variable y el punto para separar los datos. Suponiendo que el espacio se ha divido en M regiones $R_1, R_2, \dots, R_M$, la predicción será una constante c_m en cada región. Es así que se tiene la expresión (2.7)

$$f(x) = \sum_{m=1}^{M} c_m I(x \in R_m) \tag{2.7}$$

Si adoptamos como criterio de minimización de errores o función de pérdida[20] a la suma de los errores al cuadrado, tenemos:

$$\sum (y_i - f(x_i))^2 \tag{2.8}$$

Luego de derivar la expresión (2.8) se determina que el $\hat{c}_m$ óptimo es el promedio de los valores y_i dentro de cada región R_m:

$$\hat{c}_m = promedio\ (y_i \mid x_i \in R_m) \tag{2.9}$$

Ahora, encontrar, en términos de la suma de los errores al cuadrado, la mejor partición binaria o el punto óptimo para dividir el espacio de datos es computacionalmente inviable[21], por lo que se utiliza un algoritmo codicioso, o "greedy algorithm" como se lo conoce comúnmente. Este algoritmo empieza con todo el espacio de datos, considera una variable aleatoria j y empieza considerando un punto de división s, definiendo el siguiente par de hiperplanos:

$$R_1(j,s) = \{X \mid X_j \leq s\} \text{ y } R_2(j,s) = \{X \mid X_j > s\} \tag{2.10}$$

Ahora se busca la variable divisora j y el punto separador s que satisfagan:

$$\min_{j,s} \left[\min_{c_1} \sum_{x_i \in R_1(j,s)} (y_i - c_1)^2 + \min_{c_2} \sum_{x_i \in R_2(j,s)} (y_i - c_2)^2 \right] \tag{2.11}$$

Para cualquier elección de j o s, la minimización interna se consigue por:

[20] Las funciones de pérdida se utilizan para determinar el error (también conocido como "la pérdida") entre la salida de nuestros algoritmos y el valor objetivo dado. En términos sencillos, la función de pérdida expresa qué tan lejos de lo esperado está la predicción calculada por el modelo (Deep AI, 2023).

[21] Esto se debe a que para encontrar el mejor punto de partición habría que probar todos los posibles puntos en todas las variables disponibles, y recordando que el algoritmo es recursivo, se necesitaría, repetir el proceso tomando en cuenta todas las formas posibles de ordenar las variables, lo que hace inviable su cálculo.

$$\hat{c}_1 = \text{promedio}\ (y_i \mid x_i \in R_1(j,s))\ \text{y}\ \hat{c}_2 = \text{promedio}\ (y_i \mid x_i \in R_2(j,s)) \qquad (2.12)$$

Donde $\hat{c}_1$ y $\hat{c}_2$ son los puntos óptimos de cada región. Para cada variable j, la determinación del punto óptimo s se realiza muy rápido, por lo que el escaneo de todas las variables predictoras y la determinación del punto óptimo es muy rápido. Una vez que se tiene los puntos j y s para el nivel o nodo, el proceso se repite para las M regiones creadas en cada nivel. Al revisar el proceso se puede observar que los puntos óptimos que se consiguen son óptimos locales mas no globales.

El repetir este proceso hasta que ya no se pueda dividir el espacio puede dar como resultado un sobreajuste del modelo, implicando un pobre desempeño predictivo, razón por la que, generalmente se utilizan varias estrategias para limitar el crecimiento del árbol, entre estas, limitar el número de observaciones en cada hoja, la profundidad del árbol, construir el árbol con un subconjunto de variables, entre otras. Algunos algoritmos basados en árboles aplican una o varias de estas técnicas para evitar el sobreajuste y así reducir la varianza de la predicción.

2.3.2.2 Random forest

El algoritmo random forest fue introducido por (Breiman, 2001a) y desde entonces se ha consolidado como uno de los algoritmos predictivos más conocidos de su tiempo, siendo aclamado por su adaptabilidad y facilidad de uso. Se trata de un modelo de ensamble[22], donde su output se obtiene al construir paralelamente varios árboles de regresión a partir de muestras Bootstrap[23], de modo que cada árbol depende de los valores de un vector aleatorio muestreado de forma independiente y con la misma distribución para todos los árboles del bosque.

Cada vez que se genera un nuevo nodo en un árbol, una muestra aleatoria de los predictores es usada para determinar el punto de partición del espacio de datos. Es así que para cada nivel del árbol se usa un subconjunto de predictores distinto (James et al., 2013).

[22] Se refiere a la construcción de un modelo predictivo combinando múltiples modelos con el fin de mejorar una tarea predictiva (Hastie, 2009).

[23] La idea básica es extraer al azar conjuntos de datos con reemplazo, de los datos de entrenamiento, cada muestra del mismo tamaño que el conjunto de entrenamiento original. Es decir, si se hace B veces, digamos B = 100, se producirían B conjuntos diferentes de datos generados a partir del conjunto de datos de entrenamiento original (Hastie, 2009).

El algoritmo random forest consta de los siguientes pasos:

1. Para $b = 1\ a\ B$: donde B es el número de árboles que se construirán. Este valor es un hiperparámetro y es definido mediante hiperparameter tuning[24]
   a) Obtener una muestra bootstrap $\boldsymbol{Z}^*$ de tamaño N del conjunto de datos de entrenamiento.
   b) Construir un árbol random forest T_b para la muestra bootstrap, repitiendo recursivamente los pasos siguientes en cada nodo del árbol, hasta que el nodo máximo n_{max} es alcanzado.
      i. Seleccionar aleatoriamente m variables de las p variables.
      ii. Seleccionar la mejor variable/punto de bifurcación de las m variables para el nodo (usando el método descrito para un árbol de regresión).
      iii. Dividir el nodo en dos nodos hijos.
2. Salida del ensamble de árboles o bosque $\{T_b\}_1^B$.

De manera que la predicción de un nuevo punto x está dado por:

$$\hat{f}_{\mathrm{rf}}^B(x) = \frac{1}{B}\sum_{b=1}^{B} T(x;\Theta_b) \tag{2.13}$$

Donde Θ_b representa la estructura del b-iésimo árbol random forest en términos de las m variables de división, puntos de corte en cada nodo y valores del nodo terminal.

2.3.2.3 Extreme Gradient Boosting (XGB)

XGB, al igual que random forest se ha consolidado como uno de los referentes en lo que a algoritmos predictivos se refiere, siendo el preferido de muchos de los ganadores de varias competiciones de ML. El algoritmo ha sido aclamado por su eficiencia, precisión y diversidad de aplicaciones. Las características que lo diferencian de otros algoritmos basados en árboles incluyen la penalización del crecimiento de los árboles, parámetros de aleatoriedad extra, computación distribuida y selección automática de variables (Synced, 2017).

De acuerdo a Alim et al. (2020), la idea básica es aprovechar la predicción resultante de cientos de modelos simples con baja capacidad y combinarlos para generar un modelo complejo con elevada capacidad predictiva. El algoritmo fue propuesto por T.

[24] El procedimiento es explicado en la subsección 2.3.3

Chen & Guestrin (2016) y ha sido optimizado y perfeccionado a través del tiempo. Se caracteriza por ser una variación del modelo Stochastic Gradient Boosting Machine (SGBM) desarrollado por Friedman (2002), con modificaciones enfocadas sobre todo en mejorar su escalabilidad y eficiencia.

Este modelo al igual que random forest, es un modelo de ensamble basado en árboles, y de igual forma, minimiza una función de pérdida, pero a diferencia de random forest, los árboles generados no son independientes, sino que se construyen a partir del output del árbol precedente[25], además, la construcción de los árboles se basa en una medida de calidad o similitud[26]. Es así como para un conjunto de datos con n observaciones y m predictores: $\mathcal{D} = \{(\mathrm{x}_i, y_i)\}(|\mathcal{D}| = n, \mathbf{x}_i \in \mathbb{R}^m, y_i \in \mathbb{R})$, el output del modelo depende de K funciones aditivas consecutivas, las cuales se expresan en la ecuación (2.14):

$$\hat{y}_i = \phi(\mathbf{x}_i) = \sum_{k=1}^{K} f_k(\mathbf{x}_i), f_k \in \mathcal{F} \tag{2.14}$$

Donde $\phi(\mathbf{x}_i)$ representa la predicción para la observación $\mathbf{x}_i$ y que se calcula como la suma de las predicciones individuales $f_k(\mathbf{x}_i)$ de todos los árboles en el ensamble.

De la expresión (2.14) tenemos que $\mathcal{F} = \{f(\mathbf{x}) = w_{q(\mathbf{x})}\}(q: \mathbb{R}^m \to T, w \in \mathbb{R}^T)$ representa al espacio de árboles utilizados en XGB, que se distinguen por ser árboles potencialmente más complejos y optimizados de manera diferente en comparación con los árboles de regresión tradicionales (CART)[27]. Aquí q representa la estructura de cada árbol que asigna una observación al índice de la hoja correspondiente. T es el número de hojas en el árbol. Cada $f(x)$ corresponde a una estructura de árbol q independiente y a una hoja w. Cada hoja de un árbol de regresión contiene un valor continuo que representa la predicción en esa hoja específica, de manera que w_i representa el valor de predicción en la i-ésima hoja. Para cada observación se usan las reglas de decisión en el q árbol y se

[25] Conocido también como Boosting, donde un modelo es construido para mejorar la predicción del modelo precedente. (Geeks For Geeks, 2022)

[26] A diferencia de random forest, que determina el punto de bifurcación del espacio de datos en base al promedio de los puntos y pertenecientes a cada región, XGB al generar arboles sucesivos, usa los residuos de las predicciones de cada árbol generado, es así como la estrategia para determinar el punto de división óptimo del espacio de datos y generar las regiones rectangulares está dado por la ecuación 2.

[27] El término CART son las iniciales de los métodos Classification and Regression Trees, propuestos por Breiman en 1984, explicados previamente en la subsección 2.4.2.1

calcula la predicción final al sumar las predicciones individuales en las hojas correspondientes.

Una vez se ha definido $\hat{y}$ en la expresión (2.14) es necesaria una función objetivo a optimizar para entrenar el modelo XGB.

$$\mathcal{L}(\phi) = \sum_i \, l(\hat{y}_i, y_i) + \sum_k \, \Omega(f_k) \tag{2.15}$$

La ecuación (2.15) representa la función objetivo que se minimiza (optimiza) durante el entrenamiento de XGB. El primer término $\sum_i \, l(\hat{y}_i, y_i)$ es una función de pérdida, convexa y derivable que mide la diferencia entre las predicciones $\hat{y}_i$ (definida en la ecuación (2.14)) y los valores observados y_i. En el segundo término $\sum_k \, \Omega(f_k)$ se incluye un término de regularización donde $\Omega(f) = \gamma T + \frac{1}{2}\lambda \parallel w \parallel^2 + \frac{1}{2}\alpha|w|$, el cuál es utilizado para evitar el sobreajuste del modelo, en el que α y λ representan a la regularización L1y L2 respectivamente y γT es otro término de regularización que limita la profundidad del árbol.

Ahora bien, dado que el ensamble basado en árboles $\hat{y}_i$ de la expresión (2.15) incluye funciones $f_k(\mathbf{x}_i)$ (recordando de la expresión (2.14) que $\hat{y}_i = \sum_{k=1}^{K} \, f_k(\mathbf{x}_i)$) como parámetros, la optimización de la expresión (2.15) no es posible mediante métodos tradicionales como Gradient Descent[28]. Es así como el modelo se entrena u optimiza de manera aditiva, entonces, sea $\hat{y}_i^{(t)}$ la predicción de la i-ésima observación en la t-ésima iteración, si se añade f_t, se podría expresar $\hat{y}_i^{(t)}$en función de la predicción precedente y la función objetivo a minimizar (2.15) se convertiría en (2.16):

$$\mathcal{L}^{(t)} = \sum_{i=1}^{n} \, l\left(y_i, \hat{y}_i^{(t-1)} + f_t(\mathbf{x}_i)\right) + \Omega(f_t), \tag{2.16}$$

[28] Gradient Descent es un método de optimización de los más utilizados. Dada una función $J(\theta)$ parametrizada por parámetros $\theta \in \mathbb{R}^d$, funciona actualizando el valor de los parámetros en la dirección opuesta del gradiente de la función objetivo $\nabla_\theta J(\theta)$ con respecto a los parámetros. La presencia de funciones en la expresión (2.14) es el motivo por el cual no es posible optimizar la expresión utilizando este método (Ruder, 2017).

Al uso de aproximaciones de series de Taylor de segundo orden[29], se transforma la expresión (2.16) en la expresión (2.17) que representa a la función objetivo para la t-ésima iteración o árbol sucesivo, que sí puede ser optimizada:

$$\tilde{\mathcal{L}}^{(t)} = \sum_{i=1}^{n} \left[g_i f_t(\mathbf{x}_i) + \frac{1}{2} h_i f_t^2(\mathbf{x}_i) \right] + \Omega(f_t) \tag{2.17}$$

Donde $g_i = \partial_{\hat{y}^{(t-1)}} l\left(y_i, \hat{y}^{(t-1)}\right)$ representa al gradiente y $h_i = \partial^2_{\hat{y}^{(t-1)}} l\left(y_i, \hat{y}^{(t-1)}\right)$ representa a la hessiana.

Ahora ya se tiene una función convexa que se puede optimizar, por lo que resta derivar la función objetivo $\mathcal{L}$ con respecto a f_t y se obtiene la siguiente expresión:

$$f_t optimo = -\frac{(g_1 + g_2 + \cdots + g_n)}{(h_1 + h_2 + \cdots + h_n + \lambda)} \tag{2.18}$$

Donde los términos g y h se refieren al gradiente y la hessiana, respectivamente, y se utilizan en el proceso de optimización para encontrar el valor óptimo de f_t.

Ahora, recordando de (2.15) que $l(y_i, \hat{y})$ hace referencia a una función de pérdida cualquiera. Por ejemplo, se puede definir a esta como los errores promedio al cuadrado: $l(y_i, \hat{y}) = \frac{1}{2}(y_i - \hat{y}_i)^2$

Entonces la expresión (2.18) se convierte en:

$$f_t optimo = \widehat{f_t} = -\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)}{\sum_{i=1}^{n} k + \lambda} \; donde \; k = 1 \tag{2.19}$$

Donde el numerador es la suma de los residuos y el denominador es el número de residuos. Ahora, el valor óptimo de la función de pérdida está dado por:

$$\tilde{\mathcal{L}}^{(t)}(q) = -\frac{1}{2} \sum_{j=1}^{T} \frac{\left(\sum_{i \in I_j} g_i\right)^2}{\sum_{i \in I_j} h_i + \lambda} + \gamma T \tag{2.20}$$

[29] Se refiere a la técnica de aplicar aproximaciones de segundo orden de Taylor a una función de pérdida $f(x)$ alrededor de su punto óptimo, de manera que se logra obtener una función convexa que es diferenciable (University of Toronto, 2021).

La ecuación (2.20) suele ser usada para determinar la calidad de la estructura de un árbol q, con esta expresión se identifica en qué variable y en qué punto se debe dividir los datos para construir un árbol.

2.3.3 Tabla resumen de los modelos utilizados

Para facilitar la lectura, la Tabla 2.2 resume los modelos utilizados en la comparación empírica realizada en el Capítulo 3:

Tabla 2.2: Resumen de modelos utilizados y sus características

MODELO	TIPO DE MODELO	VENTAJAS	DESVENTAJAS	FUNCIÓN OBJETIVO
ARIMA	Econométrico – Paramétrico	Modelo univariado, fácil de ajustar, buena capacidad de predicción en el corto plazo si la serie es estacionaria, captura tendencias.	Requiere una serie estacionaria, no admite variables externas, sensibles a cambios estructurales, computacionalmente costoso, limitado a relaciones lineales en los datos.	$Y_t = \alpha + \phi_1 Y_{t-1} + \cdots + \phi_p Y_{t-p} \epsilon_t + \theta_1 \epsilon_{t-1} + \cdots + \theta_q \epsilon_{t-q}$
ARIMAX	Econométrico - Paramétrico	Incluye predictores externos, mayor precisión que el modelo ARIMA, fácil interpretación (como el de una regresión	Requiere que las series sean estacionarias, computacionalmente costoso, limitado a relaciones lineales en los datos	$Y_t = \beta_1 X_{1,t} + \beta_2 X_{2,t} + \cdots + \beta_k X_{k,t} + \eta_t$

		lineal), captura tendencias		
Random forest (RF)	ML - No Paramétrico	Facilidad de uso, buena precisión, no requiere normalización de los predictores, controla el sobreajuste, puede lidiar con datos perdidos, captura relaciones no monótonas y no lineales en los datos.	Requiere gran capacidad computacional, dificultad de interpretación, no ofrece significancia estadística de una variable, si no se ajusta correctamente puede sufrir de sobreajuste	$\hat{y} = \hat{f}_{\mathrm{rf}}^{B}(x) = \frac{1}{B}\sum_{b=1}^{B} T(x;\Theta_b)$
Extreme Gradient Boostig (XGB)	ML - No Paramétrico	Alta escalabilidad, excelente precisión, no requiere normalización de los predictores, controla el sobreajuste, captura relaciones no monótonas y no lineales en los datos.	El ajuste es más complejo dado que tiene más parámetros para ajustar, dificultad de interpretación, no ofrece significancia estadística de una variable, si no se ajusta correctamente o los datos son muy limitados puede sufrir de sobreajuste.	$\hat{y}_i = \phi(\mathbf{x}_i) = \sum_{k=1}^{K} f_k(\mathbf{x}_i)$

Realizado por: Christian Mateo Quiguiri Daquilema.

2.3.4 Medidas de evaluación y comparación:

La diferencia metodológica entre un modelo econométrico y de ML dificulta la comparación de los modelos desde su metodología, no obstante, una vez estimado o entrenado el modelo, es posible utilizar medidas comunes para medir el error de la predicción de los modelos, entre las cuales se encuentra; la Raíz del Promedio de los Errores al Cuadrado (RMSE) y el Porcentaje Absoluto Promedio de los Errores (MAPE), cuyas fórmulas se presentan a continuación:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - x_i)^2}$$

$$\text{MAPE} = \frac{1}{n}\sum_{\text{i}=1}^{\text{n}}\left|\frac{\text{x}_\text{i} - \text{x}_\text{i}}{\text{x}_\text{i}}\right| \times 100\%$$

En este capítulo se ha revisado brevemente las principales diferencias entre los modelos econométricos y de ML, analizando sus fortalezas y debilidades. Para el caso del ML sus fortalezas se resumen en obtener métodos poderosos y flexibles de hacer predicciones de calidad y mientras que sus debilidades se manifiestan principalmente por la ausencia de suposiciones sólidas y en su mayoría no verificables. La Econometría por otro lado sintetiza sus fortalezas en su solidez teórica, que permite hacer inferencias sobre la población, y que una vez estimado el modelo permite una interpretación directa de sus resultados, mientras que parte de sus debilidades vienen de la solidez teórica que la hace tan confiables, las asunciones acerca del modelo generador de datos y la poca flexibilidad de sus modelos que pueden resultar en modelos alejados de la realidad (Mullainathan & Spiess, 2017).

En el siguiente capítulo se analiza las diferencias expuestas en el presente capítulo desde un punto de vista experimental, contrastando sus fortalezas y debilidades de forma empírica en la previsión del volumen de crédito.

Capítulo 3: Contrastación Empírica

Este capítulo presenta un análisis comparativo y metodológico entre un modelo de ML y un modelo econométrico en un problema de predicción de una serie de tiempo – la serie corresponde al volumen de crédito de consumo, mensual, de los hogares ecuatorianos durante el periodo comprendido entre enero de 2005 a marzo de 2023. El ejercicio comparativo sigue las cuatro fases estándar de un análisis cuantitativo, a saber: (i) preparación de los datos, (ii) construcción del modelo, (iii) validación del modelo, (iv) interpretación de los resultados. El ejercicio comparativo se realiza en los últimos doce meses de observaciones históricos disponibles. Además, se dará respuesta a varias preguntas que pueden surgir al comparar los modelos en cada fase, entre estos: el problema de la multicolinealidad, datos faltantes, entre otros.

En particular, el pronóstico se realizó para los últimos doce meses de la serie *consumo,* es decir, el periodo comprendido entre abril de 2022 y marzo de 2023. La técnica que se realizó es la de un paso adelante o Walking Forward Forecasting, que consiste en pronosticar $k = 1$ periodos fuera de la muestra, donde el modelo es entrenado recursivamente en cada nueva observación T y los $T + k$ pronósticos adelante, que son computados cada vez basándose en la información disponible hasta T (Brownlee, 2019).

Para la implementación de los modelos se hizo uso de la plataforma integral de analítica Databricks[30]. En específico, para los modelos de ML se usa el paquete Scikit Learn pues ofrece un amplio rango de los más avanzados algoritmos de ML, a través de una interfaz sencilla y lenguaje de alto nivel de propósito general (Pedragosa et al., 2011). Para los modelos econométricos se hace uso del paquete *Statsmodels*, que provee herramientas para el análisis econométrico y estadístico (Seabold & Perktold, 2010). La versión de Python y de las diferentes librerías utilizadas en este trabajo se encuentra en la carpeta del proyecto en Github[31] creada por el autor, la carpeta contiene todos los archivos de código y de datos necesarios para replicar esta investigación.

3.1. Descripción de variables

Los predictores, al igual que la variable objetivo, se componen de series de tiempo con periodicidad mensual, que se han obtenido de fuentes gubernamentales como la

[30] Databricks es una plataforma de análisis e ingeniería de datos basada en la nube.
[31] El Proyecto el cual contiene los notebooks y los datos utilizados. Se puede acceder desde el siguiente enlace: https://github.com/mateoquiguiri1995/Tesis-Pronostico-Volumen-de-Credito.git

Superintendencia de Bancos y el Banco Central del Ecuador (BCE). La Tabla 3.1 resume las variables identificadas. El criterio para la selección de variables ha sido guiado por la literatura, identificando tres factores clave que inciden sobre el volumen de crédito, factores relacionados con la demanda y oferta del crédito, así como factores del entorno macroeconómico.

Con el afán de incluir factores potencialmente útiles para la predicción del volumen de crédito, en la Sección 3.8 se realizó una experimentación considerando predictores no tradicionales, replicando lo realizado por Stavinova, Timoshina, & Chunaev (2021) quienes utilizan el buscador Yandex (plataforma similar a Google Search, en Rusia) para obtener las búsquedas relacionadas con el volumen de crédito en San Petersburgo y luego incluirlos como predictores del mismo. Para el caso de Ecuador, se utiliza la información disponible en la plataforma Google Trends, la cual analiza las búsquedas de los usuarios de Google Search, Youtube, Google News y Google Shopping. Esta plataforma ofrece estadísticas diarias, semanales, mensuales y anuales acerca de los temas más relevantes para sus usuarios. De esta manera se busca sacar provecho de la información disponibles en las nuevas fuentes de datos y evaluar su relevancia en el pronóstico de un agregado económico como es el volumen de crédito.

Tabla 3.1: Descripción y detalles de las variables utilizadas en la comparación de modelos

VARIABLE	DESCRIPCION	FUENTE	TIPO
Crédito Hogares (*consumo*)	Volumen mensual de transacciones crediticias originales de los segmentos hipotecario y consumo en Ecuador	Superintendencia de Bancos	Variable Objetivo
Inflación *(inflacion)*	Riesgo proveniente del mercado real	Banco Central del Ecuador (BCE)	Entorno macroeconómico (Maldonado & Vera, 2011)
Tasa Interés activa *(tasa_activa)*	Riesgo del sector financiero	Banco Central del Ecuador (BCE)	Entorno macroeconómico (Maldonado & Vera, 2011)
Tasa Interés pasiva *(tasa_pasiva)*	Riesgo del sector financiero	Banco Central del Ecuador (BCE)	Entorno macroeconómico (Maldonado & Vera, 2011)

Precio WTI ***(precio_wti)***	Precio del petróleo WTI	U.S. Energy Information Administration	Entorno macroeconómico (Maldonado & Vera, 2011)
Iaec ***(iaec)***	Índice de actividad económica coyuntural	Banco Central del Ecuador (BCE)	Factores de la demanda de crédito (Gattin-turkalj, Ljubaj, Martinis, & Mrkalj, 2007)
Icc ***(icc)***	Índice de confianza al consumidor	Banco Central del Ecuador (BCE)	Factores de la demanda de crédito
Roe SF ***(roe_sf)***	Rentabilidad Patrimonial Sistema Financiero	Superintendencia de Bancos	Factores de la oferta de crédito (Maldonado & Vera, 2011)
GQ: crédito banco Guayaquil ***(gq_credito_bco_guayaquil)***	Búsqueda Google: Índice de interés relativo (0-100) de búsqueda de "crédito banco Guayaquil" en Ecuador	Google Trends	Factores de la demanda de crédito (Stavinova et al., 2021)
GQ: crédito banco Pacífico ***(gq_credito_bco_pacífico)***	Búsqueda Google: Índice de interés relativo (0-100) de búsqueda de "crédito banco Pacífico" en Ecuador	Google Trends	Factores de la demanda de crédito (Stavinova et al., 2021)
GQ: crédito banco Pichincha ***(gq_credito_bco_pichincha)***	Búsqueda Google: Índice de interés relativo (0-100) de búsqueda de "crédito banco Pichincha" en Ecuador	Google Trends	Factores de la demanda de crédito (Stavinova et al., 2021)
GQ: crédito banco Produbanco ***(gq_credito_bco_produbancol)***	Búsqueda Google: Índice de interés relativo (0-100) de búsqueda de "crédito banco Produbanco" en Ecuador	Google Trends	Factores de la demanda de crédito (Stavinova et al., 2021)
GQ: crédito quirografario ***(gq_credito_quirografario))***	Búsqueda Google: Índice de interés relativo (0-100) de búsqueda de "crédito quirografario" en Ecuador	Google Trends	Factores de la demanda de crédito (Stavinova et al., 2021)
GQ: préstamo quirografario ***(gq_prestamo_quirografario))***	Búsqueda Google: Índice de interés relativo (0-100) de búsqueda de "préstamo quirografario" en Ecuador	Google Trends	Factores de la demanda de crédito (Stavinova et al., 2021)
GQ: crédito ***(gq_credito)***	Búsqueda Google: Índice de interés relativo (0-100) de búsqueda de "crédito" en Ecuador	Google Trends	Factores de la demanda de crédito (Stavinova et al., 2021)
GQ: préstamo ***(gq_prestamo))***	Búsqueda Google: Índice de interés relativo (0-100) de	Google Trends	Factores de la demanda de crédito (Stavinova et al., 2021)

	búsqueda de “préstamo" en Ecuador		
GQ: simulador de crédito *(gq_simulador_credito)*	Búsqueda Google: Índice de interés relativo (0-100) de búsqueda de “simulador de crédito" en Ecuador	Google Trends	Factores de la demanda de crédito (Stavinova et al., 2021)

Fuente: Banco Central del Ecuador, Superintendencia de Bancos y Seguros, Google Trends y US. Energy Information
Realizado por: Christian Mateo Quiguiri Daquilema.

Se debe mencionar que la serie objetivo *consumo* es una variable proxy del volumen de crédito total de los hogares de Ecuador, ya que por motivos de disponibilidad de los datos en la fuente (BCE), la variable contabiliza únicamente el crédito concedido por los bancos y no el crédito concedido por las entidades pertenecientes al sector financiero popular y solidario como las Cooperativas de Ahorro y Crédito, por lo que se debe considerar este punto al interpretar los resultados.

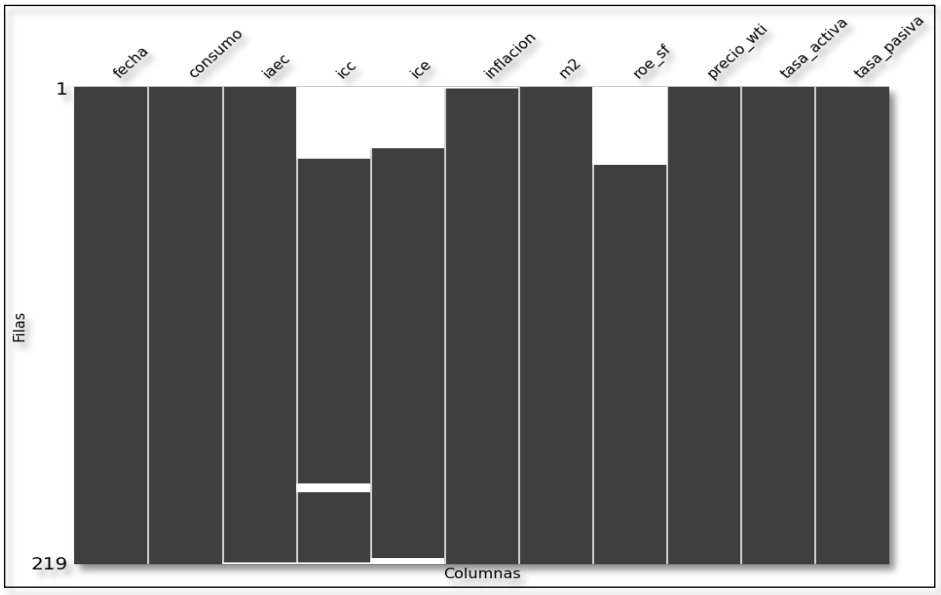
3.2. Exploración de datos

Al revisar la calidad de los datos se evidenciaron dos problemas, datos faltantes y variables altamente correlacionadas. A continuación, se revisa brevemente su implicación dentro de cada enfoque, ya sea econométrico o de ML.

3.2.1. Datos faltantes

El problema de los datos faltantes se puede abordar utilizando herramientas similares, ya sea que se utilice un modelo econométrico o uno de ML. Estas herramientas pueden ser la eliminación de la variable, la eliminación de la observación o la imputación de datos. Sin embargo, a diferencia de los modelos econométricos, los modelos ML de árboles pueden manejar los datos faltantes de forma nativa, es decir, directamente desde el algoritmo que se usa para entrenar el modelo. En el Gráfico 3.1, se visualiza la estructura y proporción de los datos faltantes para cada variable revelando que las variables '*icc'*, '*roe_sf'* e '*ice'* presentan un porcentaje significativo de datos faltantes, 20%, 21% y 17% respectivamente. Un aspecto relevante es que la falta de datos no se distribuye aleatoriamente, sino que se concentra al inicio de la serie, lo cual se debe a una limitación por la disponibilidad de datos. Esta no aleatoriedad de los datos perdidos puede introducir sesgos en el análisis y resultar en una estimación imprecisa de los parámetros del modelo, afectando la fiabilidad de los modelos predictivos que se construyan, ya sea ARIMAX o de árboles de decisión.

Gráfico 3.1: Proporción y ubicación de datos perdidos de las variables utilizadas para los modelos econométricos y de ML.



Fuente: Banco Central del Ecuador y Superintendencia de Bancos y Seguros.
Realizado por: Christian Mateo Quiguiri Daquilema.

Para evitar cualquier sesgo, se considera la implementación de estrategias para manejar estos datos perdidos, a saber: imputación o remoción de la observación. En el caso de los datos faltantes en series temporales, es importante considerar que su comportamiento puede ser diferente al de los datos de sección transversal. En nuestro estudio, observamos que existen datos perdidos que parecen seguir un patrón no aleatorio, especialmente al inicio de las series. Debido a esta falta de aleatoriedad, decidimos eliminar todas las observaciones con datos perdidos al inicio de las series, ya que la imputación no sería apropiada en este contexto.

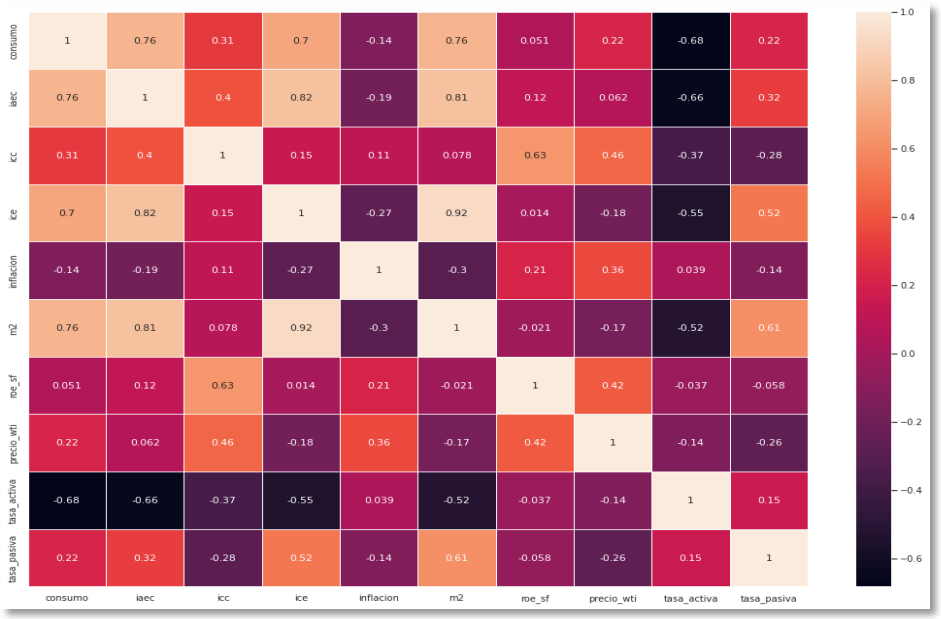
Por otro lado, también identificamos datos faltantes cerca del final de las series 'icc' e 'ice' (ver Gráfico 3.1). En este caso, optamos por realizar la imputación utilizando el último valor observado, la decisión considera que la cantidad de observaciones con datos perdidos en esta parte de las series es baja, representando menos del 2% del total y los valores son estables, lo que respalda la aplicación de esta técnica en esta situación particular (datacamp, 2023).

3.2.2. Correlación

Para visualizar patrones de correlación se confía en la matriz de correlación de Spearman de los predictores ilustrada en el Gráfico 3.2 mediante un mapa de calor. Se hace uso de

la correlación de Spearman sobre la de Pearson, ya que la primera puede indicar relaciones lineales en caso de que existan, pero, además refleja indicios acerca de la monotonía de las variables (Frost, 2020). Al calcular la correlación de Pearson, se evidencia una correlación muy similar a la de Spearman, por lo que se presentan la correlación de Spearman únicamente, sin embargo, la correlación de Pearson se presenta en el Anexo B.3.

Gráfico 3.2: Matriz de correlación de Spearman de la variable objetivo 'consumo' y predictores



Fuente: Banco Central del Ecuador y Superintendencia de Bancos y Seguro.
Realizado por: Christian Mateo Quiguiri Daquilema.

Con base en las medidas de correlaciones calculadas, es evidente que las variables *'ice'*, *'iaec'* y *'m2'* están fuertemente correlacionadas entre sí, todas con una correlación mayor a 0.7. Esta fuerte correlación sugiere que estas variables pueden moverse conjuntamente. En contraste, las variables *'tasa_activa'*, *'iaec'*, *'ice'* y *'m2'* exhiben una correlación negativa, lo que sugiere que pueden moverse en direcciones opuestas. Es necesario mencionar que la elevada correlación en algunos predictores puede dar lugar a un problema de multicolinealidad al ajustar el modelo.

3.3. Preparación de los datos

3.3.1. Modelos econométricos

En la preparación de los datos para los modelos econométricos, tanto univariados como multivariados, hay un requisito clave: los predictores deben ser estacionarios. Esta estacionariedad es una suposición fundamental que permite el desarrollo de la metodología correspondiente para los modelos ARIMA y ARIMAX. Sin embargo, el modelo ARIMAX, a diferencia del ARIMA univariado, requiere que las variables exógenas, además de ser estacionarias, tengan una relación significativa con la variable objetivo. Al añadir estas variables, se mejora la capacidad predictiva del modelo más allá de lo que se podría lograr con un modelo univariado (Hyndman & Athanasopoulos, 2018).

Para determinar la estacionariedad de los variables, en el análisis se ha utilizado la prueba Dickey-Fuller (DF). Después de aplicar esta prueba, se ha encontrado que todos los predictores y la variable objetivo alcanzan la estacionariedad después de una sola diferenciación. Estos resultados se muestran en la Tabla 3.2.

Tabla 3.2: Resultados de la prueba ADF de estacionaridad con y sin diferencia[32]

ADF TEST – P-VALOR		
Variable	**Variable original**	**1 diferencia**
Consumo	0.406	0.000
Iaec	0.476	0.001
Icc	0.463	0.087
Ice	0.774	0.069
Inflación	0.065	0.000
m2	1.000	0.055
roe_sf	0.034	0.000
precio_wti	0.016	0.000
tasa_activa	0.075	0.000
tasa_pasiva	0.480	0.000

Fuente: Banco Central del Ecuador y Superintendencia de Bancos y Seguros
Realizado por: Christian Mateo Quiguiri Daquilema.

[32] En la Tabla 3.2 se presentan los resultados de la prueba ADF de estacionaridad. Si el P-valor es menor que un nivel de significancia predefinido (en este caso, 0.1), se rechaza la hipótesis nula y se concluye que la serie es estacionaria. Un P-valor alto indica la falta de evidencia suficiente para afirmar la estacionaridad, lo que puede requerir análisis o transformaciones adicionales, como la diferenciación. (Gujarati & Porter, 2010).

Ahora bien, dado que cuando se utiliza un modelo de regresión con errores ARIMA (ARIMAX), se está ajustando simultáneamente la parte regresiva y ARIMA, los predictores (*X)* pueden ser introducidos directamente en el modelo sin necesidad de diferenciarlos previamente, ya que el modelo se enfoca en ajustar el patrón temporal en los residuos de la regresión, no en los predictores.

3.3.2. Modelos Machine Learning

3.3.2.1. ¿Cómo se transforma un problema de predicción de series de tiempo en un problema de ML?

La ingeniería de variables para modelos de ML implica la generación de predictores basados en información temporal como el año, mes y día. Una vez creados estos predictores, se desplaza la matriz de predictores (X), n periodos con respecto a la variable objetivo (y), permitiendo realizar pronósticos fuera de la muestra, este paso es al que se hace referencia cuando se habla de transformar un problema de series de tiempo a problema de regresión común en ML. Dado que el pronóstico se realizará un periodo hacia adelante, los predictores deben desplazarse al menos $n = 1$ periodos (Brownlee, 2019).

3.4. Ingeniería de variables

3.4.1. ¿Cómo determinar el número de rezagos en los predictores?

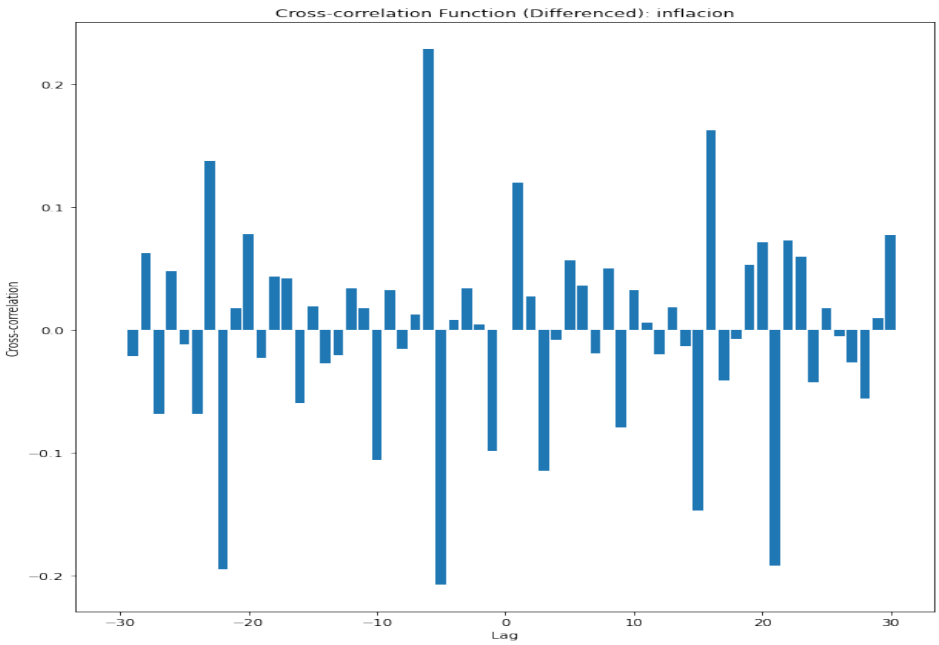
Ahora bien, para determinar cuántos rezagos (p) de los predictores se deben incluir, se utiliza la función de correlación cruzada (CCF), estableciendo p=3 como el número de rezagos que se incluirán de todas las variables. También se incluyen predictores con más rezagos, como es el caso de la inflación, dado que presenta una correlación significativa al 5to y 6to rezago, algo similar sucede con el ROE de las instituciones financieras (*roe_sf*), que tiene una correlación significativa con el consumo en el tercer rezago.

La selección de los rezagos se basa en varias consideraciones. En primer lugar, se ha tenido en cuenta la significatividad estadística, ya que una correlación significativa no implica causalidad, y las correlaciones pueden ser relaciones espurias. También se ha considerado la teoría económica, ya que ayuda a identificar si una relación es espuria o no. Además, se evita el sobreajuste del modelo al incluir demasiados rezagos de los predictores, ya que esto puede hacer que el modelo se ajuste demasiado a los datos. Por último, se ha tenido en cuenta la consistencia de las variables, ya que determinar un

número específico de rezagos para todas las variables puede evitar inconsistencias y facilitar la interpretación (Hyndman & Athanasopoulos, 2018).

Es necesario tomar en cuenta algunas consideraciones acerca de variables como la inflación, que muestra una correlación más fuerte en el quinto y sexto rezago, lo que indica un efecto retardado en el volumen de crédito. Esta relación podría tener una base teórica en la hipótesis de Fisher, que asegura que cuando la inflación incrementa, las tasas nominales aumentan en proporciones iguales, lo cual tendría un efecto de aversión a endeudarse en las personas pues la deuda sería más cara, no obstante, en el Gráfico CCF de la inflación (Gráfico 3.3), se puede evidenciar que en el quinto rezago la correlación es negativa y en el sexto rezago la correlación es positiva por lo indica que el crédito en lugar de reducirse se incrementa, teniendo una interpretación ambigua, siendo lo más probable que se trate de relación espuria (Blanchard, 2006).

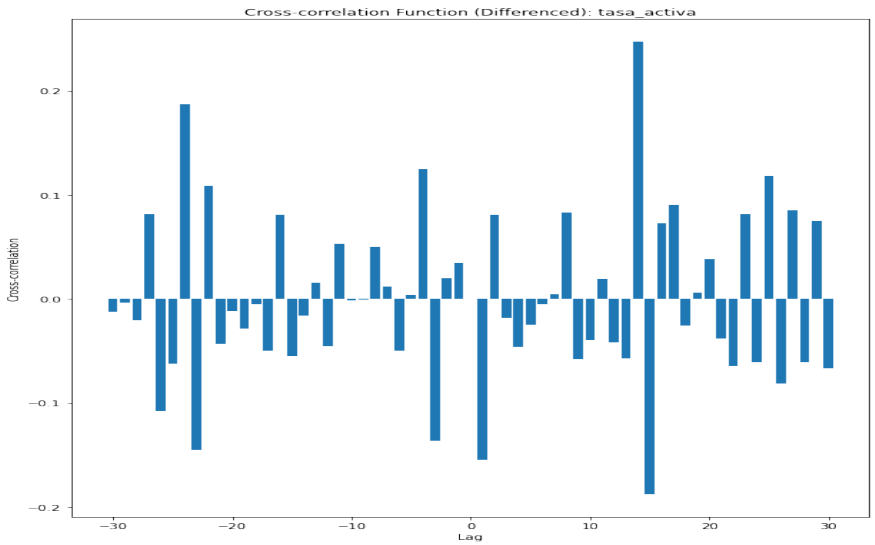
Gráfico 3.3: Función de correlación cruzada (CCF) de *'inflacion'* diferenciada



Fuente: Banco Central del Ecuador.
Realizado por: Christian Mateo Quiguiri Daquilema.

La CCF de la tasa activa (ver Gráfico 3.4) muestra una correlación mayor a 0.12 en el cuarto rezago y dado que uno de los determinantes de la demanda de crédito es la tasa de interés, se lo incluye dentro de los predictores relevantes para el modelo.

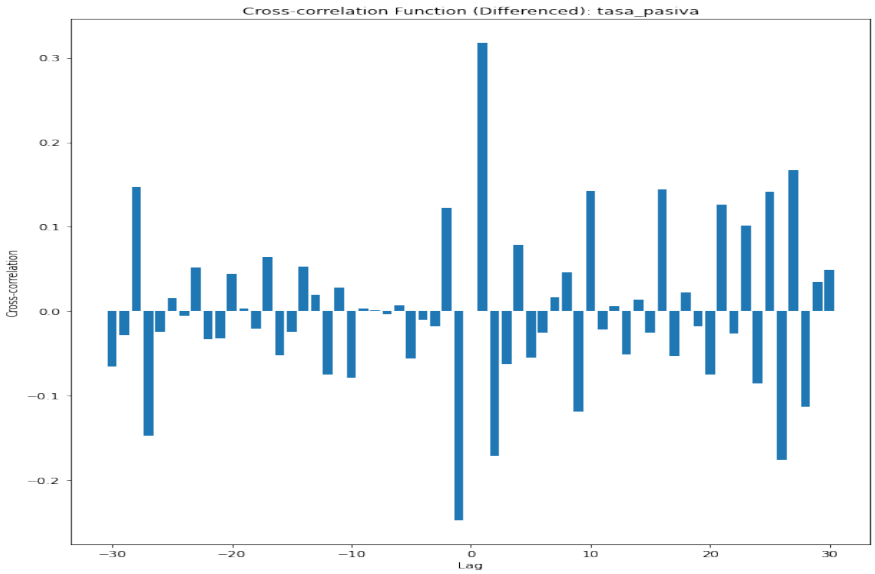
Gráfico 3.4: Función de correlación cruzada (CCF) de *'tasa_activa'* diferenciada



Fuente: Banco Central del Ecuador.
Realizado por: Christian Mateo Quiguiri Daquilema.

Finalmente, la tasa de interés pasiva (Gráfico 3.5), muestra una correlación mayor a 0.25 en el primer rezago, lo que la convierte en un predictor relevante, ya que influye directamente en las decisiones de préstamo y endeudamiento.

Gráfico 3.5: Función de correlación cruzada (CCF) de *'tasa_pasiva'* diferenciada



Fuente: Banco Central del Ecuador.
Realizado por: Christian Mateo Quiguiri Daquilema.

Es importante tener en cuenta que los modelos de ML utilizados en la experimentación pueden capturar relaciones no lineales (Hastie, 2009). Por lo tanto, aunque algunos predictores no presenten relaciones lineales significativas según la función CCF, es posible que se puedan obtener relaciones no identificadas por esta

función. Por esta razón, se mantienen 3 rezagos en todos los predictores, excepto en los casos mencionados anteriormente.

Además, se menciona la posibilidad de causalidad inversa en algunas variables, como la "*tasa_activa"*, "*m2*", "*ice*" e "*iaec*", que muestran una mayor correlación con rezagos hacia adelante (ver en Anexo C.1 los gráficos de las tres últimas). Esto podría sugerir que los cambios en el volumen de crédito podrían estar causando cambios en el volumen de crédito. Sin embargo, debido al alcance de esta investigación, solo se plantea esta inquietud y se invita al lector a profundizar en el tema.

3.4.2. Características de ventanas móviles

Además de las variables antes descritas, se incluyeron estadísticas móviles como la media y a desviación estándar, también conocidos como rolling window features[33]. Para la investigación se ha seleccionado un tamaño de 12 pues la frecuencia de los datos es mensual (Baxter & King, 1995).

También se añadieron estadísticos de ventana incrementales conocidos como expanding window features[34](Brownlee, 2019). Finalmente se obtiene una matriz de 40 predictores a partir de los cuales se realizará la selección de variables para cada tipo de modelo.

3.4.3. Selección de variables

3.4.3.1. Modelos econométricos

En econometría, se busca que un modelo sea parsimonioso debido a la importancia de la simplicidad y la eficiencia en el proceso de modelado. Por esta razón, antes de ingresar todos los predictores descritos anteriormente se realizará una selección de variables utilizando la correlación lineal entre los predictores y la variable objetivo, de manera que se mantendrá únicamente los predictores más relevantes. Para esto se utilizará la función $SelectKbest$ de la librería Scikit-Learn (Pedragosa et al., 2011), conjuntamente con la teoría económica. Es así como se seleccionaron de las variables iniciales aquellas

[33] Es una variable que contiene estadísticos móviles de ventana como la media, el valor mínimo o máximo, media, mediana o desviación estándar, calculado para los valores dentro de la ventana. Para determinar el tamaño lo más recomendable es realizar pruebas y seleccionar el que mejor predicción ofrece (Luka, 2020).

[34] La idea es similar a los estadísticos de ventanas móviles, pero a diferencia de estas que tienen tamaño fijo, las ventanas desplegables o incrementales tienen un punto de inicio fijo e incorporan nuevos datos a medida que están disponibles (Luka, 2020).

más relevantes, antes de aplicar la función $SelectKbest$. Posteriormente se determinó $K = 8$ variables, las cuales están más relacionadas con la variable objetivo.

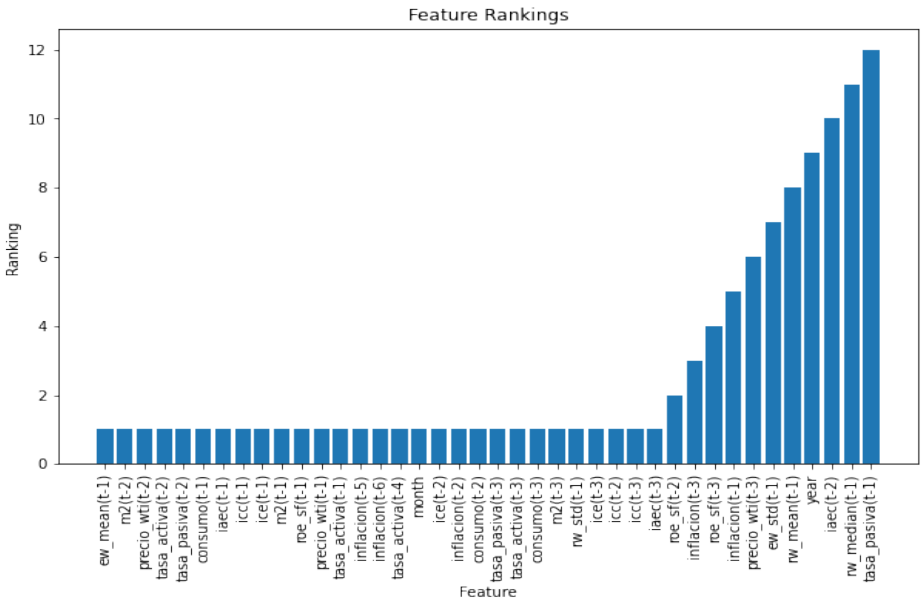
Luego, de estos 8 predictores se ha elegido la mejor combinación de los mismo utilizando el criterio de información AIC y la coherencia económica. En el anexo C.2 se puede observar los distintos modelos y correspondiente AIC y BIC. Finalmente, bajo este criterio, el modelo que ofrece el AIC más bajo es el modelo entrenado con solo cuatro predictores; "*icc(t-3)", "inflación(t-3)", "icc(t-2)"," inflación(t-2)".*

3.4.3.2. Modelos de Machine Learning

Dado el tamaño reducido de la base de datos (cerca de 200 observaciones), agregar todos los predictores generados previamente puede llevar a un sobreajuste del modelo a los datos, alta varianza y baja precisión en la predicción. Para mitigar este problema, es crucial seleccionar solo las variables relevantes.

Para este proceso, se emplea una técnica llamada RFECV[35] (Recursive Feature Elimination with Cross-Validation) que establece un ranking de las variables basándose en su relación (tanto lineal como no lineal) con la variable objetivo. Este ranking se ilustra en el Gráfico 3.6.

Gráfico 3.6: Selección de variables, ranking RFECV

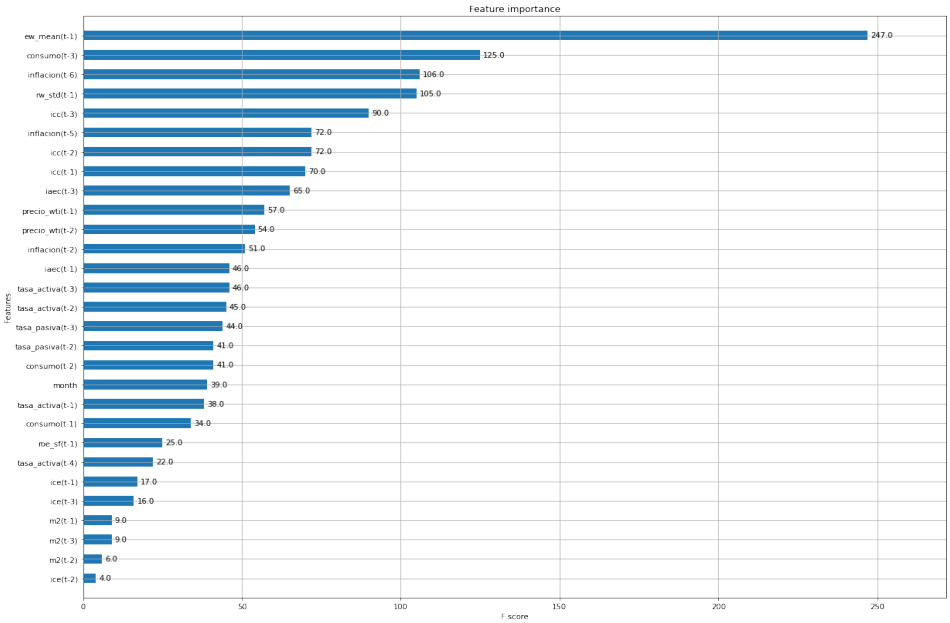


Fuente: Banco Central del Ecuador y Superintendencia de Bancos y Seguro.
Realizado por: Christian Mateo Quiguiri Daquilema.

[35] Dado un estimador externo que asigna pesos a las características (por ejemplo, los coeficientes de un modelo lineal), el objetivo de la eliminación recursiva de variables con validación cruzada (RFECV) es seleccionar variables considerando de forma recursiva conjuntos de variables cada vez más pequeños.

En este ranking, las variables con un valor más cercano a 1 se consideran de mayor importancia. En base a esta metodología, se selecciona 29 variables, que poseen un valor de 1. No obstante, de estas 29 variables, realizamos una segunda selección para reducir la cantidad a 20[36], para lo cual se aplica otro método de selección de variables, pero ahora se utiliza la importancia de las variables del modelo XGB, bajo el criterio de mayor ganancia o "gain"[37] en inglés (ver Gráfico 3.7).

Gráfico 3.7: Importancia de variables bajo el criterio de ganancia "gain" de un modelo XGB



Fuente: Banco Central del Ecuador y Superintendencia de Bancos y Seguro.
Realizado por: Christian Mateo Quiguiri Daquilema.

Es relevante mencionar que, aunque los modelos de ML pueden manejar un gran número de variables, el reducido número de observaciones y la gran cantidad de predictores puede comprometer la estabilidad y la precisión del modelo (alta varianza y baja precisión). Este riesgo se incrementa al añadir predictores poco relacionados con la variable objetivo. Por ende, el realizar una selección de variables adecuada se minimiza el ruido y mejora la estabilidad del modelo (Hyndman & Athanasopoulos, 2018). Es así como la matriz de predictores finales está compuesta por 20 variables, las cuales se utilizaron tanto para el modelo XGB como para el modelo random forest.

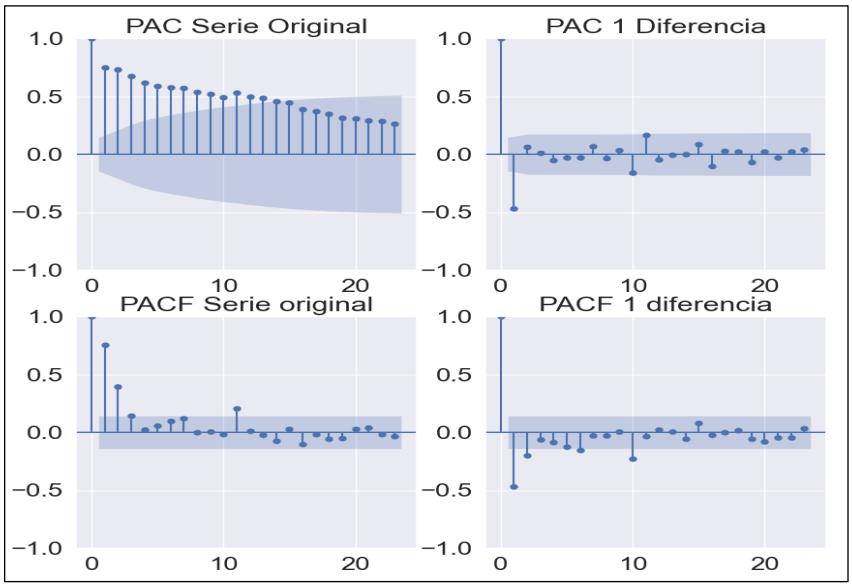
[36] Este número es puramente experimental, se ha determinado este número con el propósito mejorar la estabilidad del modelo debido al alto número de predictores y bajo tamaño de la muestra.
[37] Esta métrica muestra la ganancia promedio de una variable al modelo, calculado al tomar la contribución de una variable para cada árbol del modelo. Un valor más alto implica una mayor importancia de la variable (Abu-Rmileh, 2019).

3.5. Construcción del Modelo

Las principales diferencias entre el enfoque econométrico y de ML se hacen evidentes durante la fase de entrenamiento de los modelos (identificación y estimación para el enfoque econométrico). En primer lugar, el modelo ARIMA sigue la metodología de Box-Jenkins y se basa en pruebas de estacionariedad junto con el análisis de las funciones de Autocorrelación (PAC) y Autocorrelación Parcial (PACF). Los resultados de estos análisis se encuentran en el Gráfico 3.8. Al examinarlos, se identifica que el modelo ARIMA (1,1,1) es adecuado. Sin embargo, se opta por utilizar la especificación ARIMA (2,1,1). La razón se debe a que en el estudio se busca la especificación que ofrezca el mejor pronóstico, dejando que los P-valores de los coeficientes pasen a un segundo plano. A pesar de esto, no se descuida la verificación de los residuos (Hyndman, 2018).

En el caso del modelo ARIMAX, se aplican pruebas de estacionariedad y se analizan las funciones de Autocorrelación (PAC) y Autocorrelación Parcial (PACF) para evaluar la serie temporal en términos de su componente ARIMA. Sin embargo, en este enfoque, también se considera la inclusión de regresores externos o términos de residuos en la especificación del modelo. La elección de estos regresores se basa tanto en la mejora del pronóstico como en su significancia estadística. Los P-valores de los coeficientes de los regresores externos se evalúan junto con la verificación de los residuos, lo que asegura que estos, que incluyen el componente ARIMA y los regresores externos, se comporten como ruido blanco. El modelo ARIMAX sigue la misma especificación que se mencionó anteriormente (2,1,1).

Gráfico 3.8: PAC y PACF de la serie volumen de crédito[38]



Fuente: Banco Central del Ecuador y Superintendencia de Bancos y Seguro.
Realizado por: Christian Mateo Quiguiri Daquilema.

Una vez especificado el modelo ARIMA y ARIMAX los parámetros son estimados mediante Máxima Verosimilitud, y técnicas de estimación conjunta respectivamente. Asumiendo una distribución condicional concreta para la serie objetivo (Wooldridge, 2016).

3.5.1. ¿Cómo se ajusta un modelo predictivo basados en árboles?

En la comparación teórica realizada en el capítulo 2 se destacó que, a pesar de que tanto el modelo random forest como XGB se basan en árboles, su proceso de construcción difiere entre sí. A continuación, se proporcionará una breve descripción de cómo se construye un árbol random forest y se contrastará con un árbol XGB, con el objetivo de mejorar la comprensión.

Antes de revisar como se ajusta un árbol random forest, un primer aspecto a recordar es que esta técnica utiliza muestreo Bootstrap. Para este caso en particular el

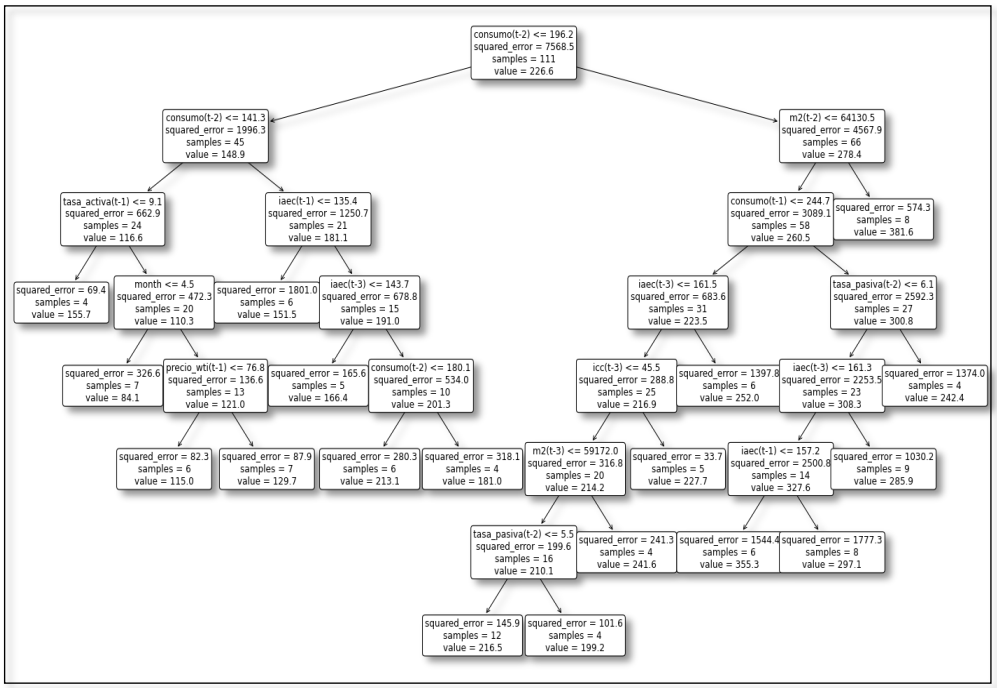
[38] Para identificar el orden p y q de un modelo ARIMA a partir de los gráficos PAC (Partial Autocorrelation Function) y PACF (Partial Autocorrelation Function), es necesario observar los picos significativos en estos gráficos. En el gráfico PAC, los palos que se extienden más allá del intervalo de confianza indican la presencia de una correlación significativa entre las observaciones pasadas y presentes. Por otro lado, en el gráfico PACF, los palos que caen fuera del intervalo de confianza sugieren una correlación significativa entre las observaciones pasadas y presentes, controlando los efectos de las observaciones intermedias. Por ejemplo, si en el gráfico PAC el primer pico significativo se encuentra en el rezago 2 y en el gráfico PACF el primer pico significativo se encuentra en el rezago 1, entonces el orden p del modelo ARIMA sería 1 y el orden q sería 2 (Gujarati & Porter, 2009).

algoritmo utiliza 4/5 de las variables para generar nuevos nodos y 1/2 de los datos para la muestra Bootstrap implicada en la construcción de cada árbol.

El algoritmo empieza seleccionando una de las $4/5 * 20$ variables. Inmediatamente se ordena la muestra Bootstrap de acuerdo con los valores de la variable seleccionada, luego divide la muestra en dos regiones, utilizando como punto divisorio el promedio de los dos primeros valores ordenados de la variable, después calcula el MSE de ambas regiones. Se repite este proceso sucesivamente para los siguientes dos valores hasta el final de la muestra. Una vez calculados los errores promedio al cuadrado de todas las posibles divisiones en esa variable, se toma aquella con el menor MSE. El proceso se repite iterativamente para las siguientes variables. Finalmente se elige aquella variable con el menor MSE, la cual genera el primer nodo del árbol o también llamado raíz.

En el Gráfico 3.9 se observa que, para el primer nodo, la variable que ofrece el menor MSE es la misma variable objetivo $Consumo(t-2)$ con un rezago de dos periodos y el punto de división es 196.2 lo que significa que todas aquellas observaciones con valores menores a 196.2 se agrupan en el lado izquierdo y en el lado derecho las que no cumplen la condición. La forma en la que los siguientes nodos son generados es idéntica, con la diferencia que el MSE es calculado sobre las observaciones que se encuentran en cada región.

Gráfico 3.9: Primer árbol de modelo random forest: estructura de un árbol de decisión[39]



Fuente: Banco Central del Ecuador y Superintendencia de Bancos y Seguro.
Realizado por: Christian Mateo Quiguiri Daquilema.

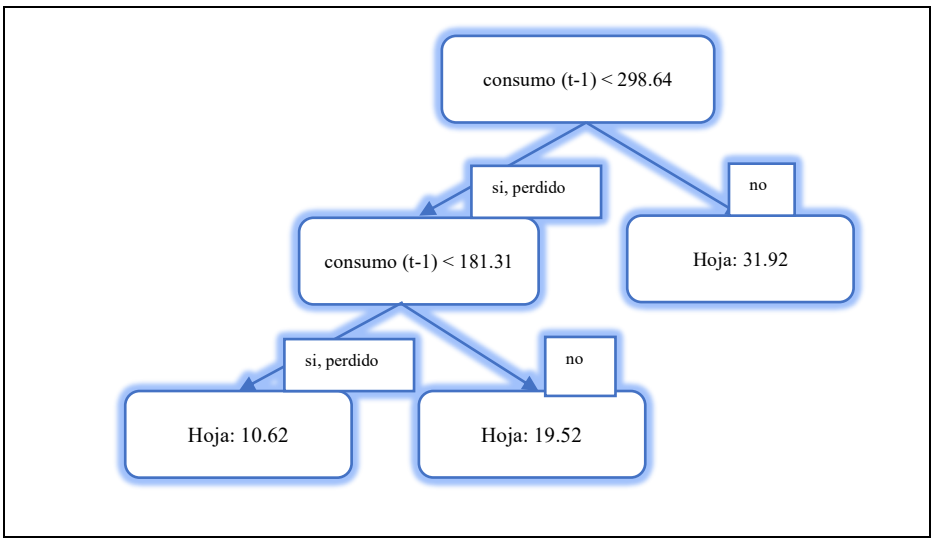
Si se da libertad al algoritmo de iterar indefinidamente, el árbol construido va a generar hojas o nodos que no se pueden dividir ya que cada nodo tendrá solo una observación, en consecuencia, el MSE de este árbol será de 0, en otras palabras, se ajustará perfectamente al conjunto de datos de entrenamiento, no obstante, esto no es deseable, pues es una clara señal de sobreajuste del modelo. De ahí que, generalmente se limite o se pode el crecimiento de los árboles. Por esta razón el árbol del gráfico 3.9 solo tiene 8 niveles.

Algo similar ocurre para el caso de un árbol XGB, pero a diferencia de random forest donde los nodos son construidos al calcular el MSE de la variable objetivo, aquí los árboles se construyen basándose en una medida de similitud y ganancia. Por ejemplo, en el Gráfico 3.10, para determinar la raíz del árbol se divide la muestra de acuerdo con

[39] El gráfico muestra el primer árbol de un modelo random forest, que utiliza una técnica de ensamble de árboles de decisión. Cada nodo representa una división en la variable de interés y muestra información sobre el punto de división, el error al cuadrado, el número de observaciones que se dirigen hacia la izquierda del nodo y el valor promedio del nodo. Por ejemplo: la variable de división es 'consumo(t-2)' <= 196.2. El error al cuadrado asociado es de 7568.5, indicando la discrepancia entre las observaciones y las predicciones en este nodo. Hay 111 observaciones que se dirigen hacia la izquierda del nodo, y el valor promedio en este nodo es de 226.6. Estos valores proporcionan información sobre cómo se realiza la división inicial en el árbol de decisión y cómo se agrupan las observaciones en función de esta división.

el promedio de dos puntos de una determinada variable al igual que random forest, luego en lugar del MSE se calcula la medida de similitud $\frac{(\sum \hat{y}_i - y_i)^2}{n+\lambda}$ para todos los puntos del nodo raíz, pero también para los nodos o regiones resultantes del nodo raíz. Una vez calculadas las similitudes se calcula la Ganancia, definida como $Ganancia = Similitud\ izquierda + Similitud\ derecha - Similitud\ raíz$, de manera que el punto divisorio de cada nodo es aquel punto de entre todos los puntos y variables cuya ganancia es la más alta.

Gráfico 3.10: Primer árbol de modelo XGB: Estructura de un árbol de decisión[40]



Fuente: Banco Central del Ecuador y Superintendencia de Bancos y Seguro.
Realizado por: Christian Mateo Quiguiri Daquilema.

Otra diferencia importante respecto a random forest es que en XGB los cálculos de la similitud y la ganancia son sobre los residuos de la predicción del árbol anterior y no los valores originales de la variable objetivo, como era el caso de random forest. Es precisamente este punto al que se hace referencia cuando se habla del contraste entre un método de Bagging (random forest) y de Boosting (XGB), por una parte, random forest construye árboles de manera independiente, mientras que en XGB los árboles son construidos a partir de los residuos de la predicción del árbol precedente.

[40] El gráfico muestra el primer árbol de un modelo random forest, que utiliza una técnica de ensamble de árboles de decisión. Cada nodo representa una división en la variable de interés y muestra información sobre el punto de división. A diferencia del modelo XGB los valores en los nodos y las hojas no son los valores de predicción final, sino que representan a la predicción del residuo del árbol precedente puesto que se trata de un modelo de Boostig.

La diferencia en el número de niveles entre el árbol random forest y el árbol XGB se debe a la forma en que se construyen cada uno de ellos. El árbol XGB tiene solo 3 niveles debido a que cada árbol se genera a partir de los residuos del árbol anterior. Si se agregaran más niveles a XGB, se produciría un sobreajuste en el modelo. Por otro lado, el árbol random forest construye árboles independientes, por lo que puede tener 7 niveles o más, sin que necesariamente se sobreajuste, ya que cada árbol se crea de forma independiente.

Se debe mencionar que los niveles del árbol, la proporción de las variables para cada nodo, así como la proporción de la muestra para cada árbol y demás parámetros se han obtenido al realizar el proceso de optimizar los hiperparámetros. En la sección 3.5.3 se detalla más a fondo esta herramienta.

3.5.2. ¿Cómo se evita el sobreajuste de un modelo de Machine Learning?

El efecto de la flexibilidad de los modelos de ML les permite ajustarse a estructuras variadas de datos, pero a su vez los hace proclives al sobreajuste. Es aquí cuando entra el juego la regularización, un término que hace referencia a dos aspectos, el primero se refiere a técnicas para limitar el modelo base, en el caso de árboles; el número de niveles, numero de nodos, número mínimo de observaciones mínimo para cada nodo y demás características de un árbol. El otro aspecto hace referencia a un término que es agregado en la función de pérdida a optimizar, que al igual que el anterior busca reducir la sensibilidad del modelo a una observación en particular (Mullainathan & Spiess, 2017).

3.5.3. ¿Cómo se determina la especificación óptima de un modelo de Machine Learning?

A diferencia del enfoque econométrico, los algoritmos utilizados para el entrenamiento de los modelos de ML contienen parámetros que no son ajustados en el entrenamiento del modelo, estos parámetros especiales conocidos como hiperparámetros determinan la complejidad del modelo y la flexibilidad por la que estos métodos son conocidos. La metodología usada para determinar estos parámetros de denomina Hyperparameter Tuning que significa optimización de hiperparámetros. Esta técnica consiste en que dado una rejilla o “grid” de parámetros y una medida del error, evalúa, a través validación cruzada la calidad de la predicción de todas las combinaciones posibles de parámetros de la rejilla, obteniendo así los mejores parámetros para el modelo. Para elegír el mejor modelo se hace uso de MAPE, RMSE o cualquier métrica de interés. A esta forma de

evaluación, donde todos los parámetros de la rejilla son evaluados se la conoce como Grid Search (Hastie, 2009).

Primero se define al MAPE como medida de evaluación del error y para la validación cruzada se utiliza el enfoque de validación un paso adelante sobre los últimos doce meses del conjunto de datos de entrenamiento. El proceso empieza definiendo un conjunto de hiperparámetros de la rejilla de parámetros a evaluar, luego, se realiza la validación cruzada del desempeño del modelo con los hiperparámetros dados. Es decir, se entrena el modelo con los datos disponibles hasta X_t, después se realiza el pronóstico del mes siguiente, $\hat{y}_{t+1}$ y se mide el error de predicción mediante el MAPE, inmediatamente después se reajusta el modelo con los datos de entrenamiento añadido la observación del mes siguiente X_{t+1} y se vuelve a realiza el pronóstico y evaluación de $\hat{y}_{t+2}$, repitiendo el proceso sucesivamente para los siguientes meses, finalmente se calcula el error promedio de las predicciones y se cuantifica el error del modelo para un conjunto de hiperparámetros. Este proceso se repite iterativamente para cada posible combinación de hiperparámetros, razón por la cual su optimización requiere suficientes recursos computacionales. Esta limitación es el principal motivo por el que en la presente investigación se restringe el número de hiperparámetros a optimizar. A pesar de esto, el número de modelos ajustados bordearon los 80.000, tanto para XGB como para random forest, tomándose alrededor de 16 horas en concluir el proceso.

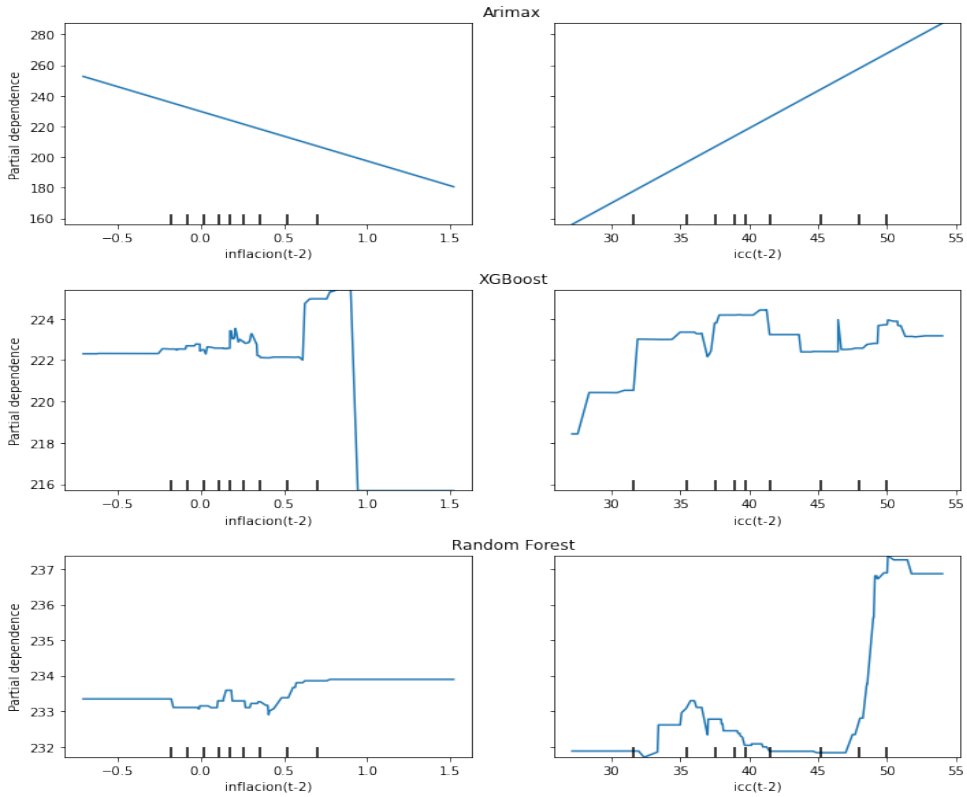
3.5.4. ¿Por qué se dice que los modelos de Machine Learning son muy flexibles?

De nuevo, una diferencia notable se manifiesta al analizar las relaciones que cada modelo captura entre los predictores (input) y la variable objetivo (output). En el Gráfico 3.11 se muestra la dependencia parcial[41] de dos variables comunes para todos los modelos multivariados, donde es posible evidenciar la capacidad del enfoque ML de obtener relaciones no lineales sin requerir ninguna especificación ex ante, como es el caso de los métodos econométricos. Así también, es evidente la relación no monótona que se observa entre las variables predictoras con el volumen de crédito, lo cual da pistas acerca de la razón por la que no fue posible identificar predictores útiles para el modelo ARIMAX puesto que la relación entre los predictores y la variable objetivo no era linear de suerte

[41] Un gráfico de dependencia parcial muestra la dependencia entre la variable objetivo y un conjunto de predictores, es decir evalúa el rendimiento del modelo al alterar los valores de un predictor, manteniendo los otros constante (Scikit-Learn, 2022b).

que ningún predictor mejoraba de forma significativa la capacidad de pronóstico del modelo.

Gráfico 3.11: Dependencia parcial[42] de variables inflacion(t-2) e icc(t-2)



Fuente: Banco Central del Ecuador y Superintendencia de Bancos y Seguro.
Realizado por: Christian Mateo Quiguiri Daquilema.

Al observar con detalle los modelos de ML se puede notar la forma escalonada en la que relacionan los predictores con la variable objetivo. Tanto la inflación como el índice de confianza al consumidor "*icc"*, ilustrando la manera cómo los métodos basados en árboles relacionan las variables los predictores, ya que es posible identificar los nodos o puntos donde se está dividiendo a la variable. Por ejemplo, en el caso random forest se puede ver que, si el índice de confianza al consumidor "*icc(t-2)*" es mayor a 46, en promedio el valor del volumen de crédito incrementará cerca de 5 millones, siendo entonces el punto $icc(t-2) < 46$ un nodo del árbol. Para el caso del modelo XGB es

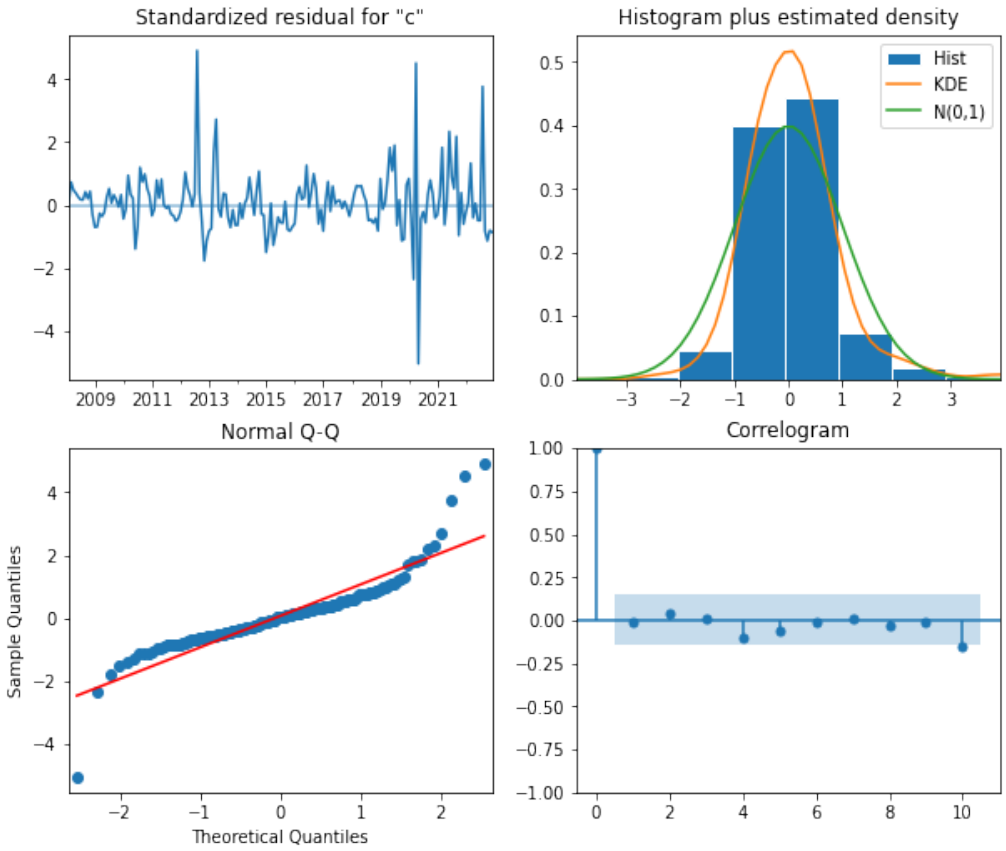
[42] La línea de tendencia muestra cómo varía la variable objetivo *'consumo'* a medida que *'inflacion(t-2)'* cambia, manteniendo constantes todas las demás variables. Para el modelo ARIMAX, los valores más altos de *'inflacion(t-2)'* están asociados con menores valores de la variable objetivo, mientras que en los modelos de árboles se evidencia una relacion no lineal. Es importante tener en cuenta que esta representación se basa en un modelo predictivo y puede no reflejar necesariamente relaciones de causalidad

más complicado definir el nodo puesto que la generación de cada árbol se usa los residuos del modelo anterior y no los valores reales como es el caso de random forest. Sin embargo, se puede evidencia que existe un patrón común que sucede si la inflación es mayor a 1, ya que, en promedio el volumen de crédito disminuirá cerca de 10 millones de dólares.

3.6. Evaluación del modelo

Cuando un modelo econométrico es ajustado, el siguiente paso es la verificación del modelo. Para el caso de ARIMA y ARIMAX se revisa que los residuos sean ruido blanco, pero en este último, debido a que es una regresión se revisa también los supuestos Gauss Markov, a saber, que el modelo sea lineal en los parámetros, que no exista multicolinealidad perfecta y que la media condicional del error sea cero. Así los estimadores de las variables exógenas son insesgados lineales óptimos (ELIO). No obstante, para no extender innecesariamente el análisis y dado el objetivo del estudio, en esta investigación solamente se revisa que los errores sean ruido blanco.

Gráfico 3.12: Análisis gráfico residuos modelo ARIMAX (2,1,1)



Fuente: Banco Central del Ecuador y Superintendencia de Bancos y Seguro.
Realizado por: Christian Mateo Quiguiri Daquilema.

El Gráfico 3.12 ayuda a identificar si los residuos cumplen los supuestos de ruido blanco. En primero lugar podemos ver que los errores, en media se encuentran alrededor de 0. Luego al revisar la tabla de resultados del modelo (ver Anexo D.1 y D.2), la prueba de Ljung-Box y el diagrama de autocorrelación evidencian que no existe correlación serial. El supuesto que no se cumple es la homocedasticidad ya que la prueba ARCH-LM[43] y el gráfico indican que la varianza no es constante. No obstante, dado que lo que se busca en la investigación es la comparación y no la inferencia, no se modela la varianza de los residuos con un modelo ARCH-GARCH y se procede con la predicción (Hardy, 2019).

Los modelos de ML, por otra parte, no necesitan cumplir ninguna validación sobre sus propiedades teóricas (evaluación de residuos u asunciones subyacentes del modelo, medidas de ajuste de bondad) para que los resultados sean válidos, pues la discriminación entre un buen o un mal modelo se realiza únicamente en base a su capacidad predictiva.

3.6.1. ¿Existen ventajas de pronosticar al uso de técnicas de Machine Learning en un contexto de datos limitados?

Para abordar la pregunta es necesario recordar que los modelos de ML han sido diseñados para trabajar con grandes volúmenes de datos y en este terreno han mostrado superioridad sobre los modelos paramétricos, no obstante, como se evidencia a lo largo de este capítulo, cuando se trabaja con bases de datos pequeñas el problema no está del todo claro. De ahí que la contrastación de la calidad de predicción en entornos con datos limitados es relevante sobre todo para áreas como la economía y más aún para la macroeconomía, que, por limitaciones inherentes a su naturaleza, acostumbra a trabajar con bases de datos reducidas.

Así pues, en la Tabla 3.3 se observa que todos los modelos ofrecen una predicción razonable con un MAPE cercano al 10%, pero son los modelos ML los más precisos, siendo el modelo random forest el que ofrece el error más bajo, con un MAPE de 9.92%, 1.5 puntos porcentuales menos que el mejor modelo econométrico. Se observa también un dato interesante respecto al MAPE y al RMSE, y es que el modelo random forest tiene el MAPE (error de porcentaje absoluto medio) más bajo, mientras que el modelo XGB tiene el RMSE (raíz del error cuadrático medio) más bajo. Eso se explica porque mientras

[43] Esta prueba se puede revisar en los anexos D.1 y D.2

el MAPE mide la diferencia porcentual promedio entre las predicciones y los valores reales, RMSE mide la diferencia cuadrática promedio.

Tabla 3.3: Evaluación de los modelos mediante MAPE[44], RMSE[45] y Desviación Estándar[46]

Modelo	MAPE	RMSE	DESV. STD MAPE
ARIMA	11.62	73.72	9.71
ARIMAX	11.78	76.06	10.12
XGB	11.50	**64.55**	9.91
RANDOM FOREST	**9.92**	66.00	**9.0**
PROMEDIO	10.13	66.01	9.18

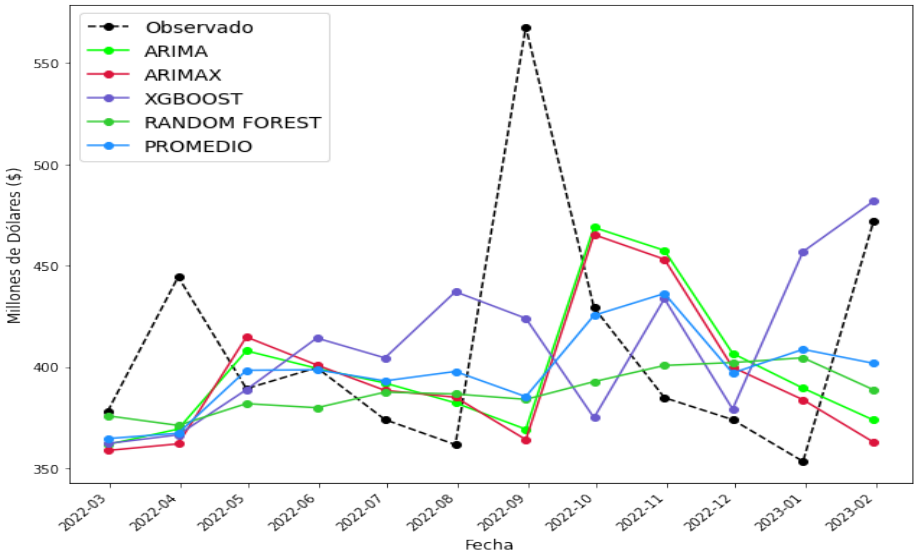
Fuente: Banco Central del Ecuador y Superintendencia de Bancos y Seguro.
Realizado por: Christian Mateo Quiguiri Daquilema.

Las dos métricas capturan diferentes aspectos de precisión. Un modelo con MAPE bajo es relativamente preciso en términos de diferencia porcentual, mientras que un modelo con RMSE bajo tiene desviaciones absolutas más pequeñas. Esto quiere decir que en promedio cuando la predicción del modelo XGB se aleja del valor real, en términos absolutos, es decir en millones de dólares, la diferencia es menor que la predicción del modelo random forest.

[44] El MAPE (Mean Absolute Percentage Error) es una métrica de evaluación de modelos de predicción que calcula el promedio del porcentaje absoluto de error entre las predicciones y los valores reales. Esta métrica proporciona una medida relativa del error promedio en relación con el tamaño de los valores reales y se expresa como un porcentaje.
[45] El RMSE (Root Mean Square Error) es otra métrica comúnmente utilizada para evaluar la precisión de un modelo de predicción. Calcula la raíz cuadrada del promedio de los errores al cuadrado entre las predicciones y los valores reales. El RMSE proporciona una medida de la desviación estándar de los errores y se expresa en la misma unidad que la variable objetivo.
[46] La desviación estándar del MAPE (Mean Absolute Percentage Error) es una medida complementaria que proporciona información adicional sobre la variabilidad de los errores en términos porcentuales. Calculando la desviación estándar del MAPE, se puede obtener una medida de cuánto varían los porcentajes de errores absolutos en relación con el promedio del MAPE. Esto permite evaluar la consistencia o estabilidad de la precisión del modelo a lo largo del tiempo.

Estos resultados evidencian que, a pesar de sus limitaciones, los modelos de ML son herramientas útiles para pronosticar en un contexto de series macroeconómicas, sumando evidencia a lo encontrado por Medeiros et al. (2021) y Chen & Baker (2020).

Gráfico 3.13: Comparación de predicciones de modelos econométricos y de ML para el volumen de crédito en Ecuador



Fuente: Banco Central del Ecuador y Superintendencia de Bancos y Seguro.
Realizado por: Christian Mateo Quiguiri Daquilema.

En el Gráfico 3.13 se observa el desempeño de los modelos. Se evidencia la gran inestabilidad de la serie, y se observa que el modelo random forest es el que mejor captura esta variabilidad.

3.6.2. ¿Cómo se obtienen los intervalos de confianza de los modelos Machine Learning?

La confianza, desde un punto de vista estadístico, de que la predicción se encontrará dentro de un intervalo definido es una de las características que más se echa de menos en los modelos ML. Debido a su naturaleza no paramétrica, no es posible obtener desviaciones estándar de los parámetros, por consiguiente, los intervalos de confianza tradicionales no son posibles de determinar. Existen métodos para obtener intervalos de confianza que involucran cambiar el modelo base y la función de pérdida del algoritmo. Otra técnica involucra la utilización de re-muestreo Bootstrap, y demás, no obstante, las técnicas para realizar pseudo intervalos de confianza están fuera del alcance de esta investigación por lo que queda al interés del lector su profundización.

3.6.3. ¿Representa la multicolinealidad un problema para el Machine Learning?

La multicolinealidad es un problema que puede tener implicaciones[47] en la estimación de modelos lineales sobre todo en la interpretación del modelo, al punto de imposibilitar la estimación del modelo si la correlación es perfecta. No obstante, Kutner, Nachtsheim, Neter, & Li (2005) mencionan que la calidad de predicción de estos modelos no se ve afectada cuando no se trata de una correlación perfecta.

En el caso de los modelos de ML sucede algo similar pues, de acuerdo con T. Chen, He, Michaël, & Tang (2022), creadores del algoritmo XGB, en los modelos basados en árboles, la multicolinealidad es superflua si se evalúa su efecto en la precisión del modelo, debido a que cada vez que un árbol decide dividir los datos para crear un nodo, solo toma en cuenta una variable a la vez y siempre selecciona a la que mejor relación tiene con la variable objetivo. A pesar de esto, se evidenció que realizar una selección cuidadosa de variables impacta en la precisión del modelo (ver Sección 3.4.3). Este resultado puede ser consecuencia del limitado número de observaciones, ya que solo se trabajó con 194 meses. Es posible que las variaciones en la precisión reflejen una leve inestabilidad del modelo. En el caso de multicolinealidad perfecta, el algoritmo seleccionará aleatoriamente solo una de las variables correlacionadas para separar los datos. Esta forma en la que se separa los datos hace que la interpretación se dificulte, haciendo que, de igual forma como en el caso lineal, la multicolinealidad represente un problema de interpretación.

3.7. Interpretación

3.7.1. ¿Es posible hacer inferencias a partir de un modelo de Machine Learning?

Cuando se habla acerca de la interpretación de un modelo de ML por lo general se menciona la palabra “black-box” o caja negra, que hace referencia a la dificultad de extraer información interpretable además de la predicción. A menudo este tema es irrelevante para la mayoría de los problemas en los que el ML es aplicado, no obstante,

[47]De acuerdo con (Wooldridge, 2016), la multicolinealidad no perfecta reduce la precisión de los estimadores al incrementar su varianza, además reduce la significancia estadística que ofrece los P-valores y dificulta la interpretación ceteris paribus.

en las ciencias sociales y más aún en la economía, la sola predicción no es suficiente. Es por tanto que se han desarrollado varias herramientas para esta tarea (Molnar, 2022).

Si hablamos de un modelo econométrico, la interpretación se enfoca en los coeficientes estimados y los *P-valores.* Por ejemplo, de la tabla 3.4, se puede saber que los predictores $icc(t-3)$, $inflacion(t-3)$ y $inflacion(t-2)$ no son estadísticamente significativos. Luego se observa que el predictor con mayor importancia (por la magnitud de su coeficiente) es la variable $inflacion(t-3)$. Así también, es posible una interpretación ceteris paribus para conocer los efectos marginales de la variable, es decir, se puede saber que un incremento porcentual unitario en la inflación significará en promedio una reducción del Volumen de Crédito en 15.2 millones de dólares manteniendo las demás variables constantes.

Tabla 3.4: Resultados modelo SARIMAX

SARIMAX (2,1,1)

Variable Dep:	Consumo			N. Observaciones:		180
Modelo:	SARIMAX (2,1,1)			AIC		1922.338
	Coef.	**Std. Err.**	**Z**	**P>|z|**	**[0.025**	**0.975]**
icc(t-3)	1.87	2.59	0.72	0.47	-3.20	6.96
inflacion(t-3)	-15.20	12.67	1.19	0.23	-9.64	40.05
icc(t-2)	4.35	1.85	2.35	0.019	0.73	7.98
inflacion(t-2)	2.99	14.91	0.20	0.84	-26.24	32.23
ar. L1	0.35	0.118	3.00	0.00	0.12	0.58
ar. L2	0.19	0.095	2.09	0.036	0.01	0.38
ma. L1	-0.88	0.083	-10.56	0.00	-1.04	-0.71
sigma2	2462.54	173.384	18.32	0.00	2122.719	2802.37

Fuente: Banco Central del Ecuador y Superintendencia de Bancos y Seguro.
Realizado por: Christian Mateo Quiguiri Daquilema.

Por otra parte, la elevada complejidad que puede alcanzar un modelo de ML hace imposible una interpretación análoga al caso anterior. Por lo que no es posible una comparación directa en este sentido. No obstante, existen algunas herramientas que permiten hacer una comparación entre métodos, analizando directamente la relación entre los predictores y la variable objetivo.

3.7.2. ¿Cómo se puede comparar la importancia de variable entre un modelo de Machine Learning y un modelo econométrico?

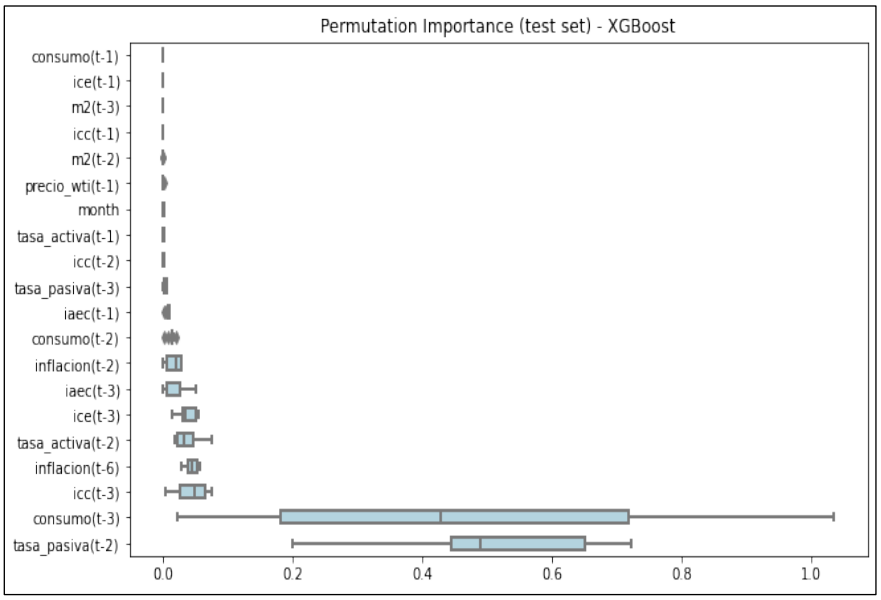
3.7.2.1. Importancia de permutación

La importancia de la permutación sirve como un método apropiado para esta comparación debido a su naturaleza agnóstica del modelo, lo que nos permite evaluar de manera consistente la importancia de las variables en ambos tipos de modelos. Esta metodología mantiene su solidez independientemente del modelo subyacente o la distribución de datos, lo que la convierte en una opción adecuada para la comparación (Breiman, 2001a). Además, la aplicación de la importancia de la permutación en ambos campos (ML y econometría) contribuye a comprender cómo los diferentes modelos asignan importancia a variables específicas, informando así la selección del modelo y mejorando la interpretabilidad.

La importancia mediante permutación rompe la relación entre los predictores y la variable objetivo reordenando aleatoriamente uno o más predictores, de manera que la perdida en la capacidad predictiva es un indicador de la importancia de la variable para el modelo (Scikit-Learn, 2022b).

Una forma sencilla de representar este método es mediante un diagrama de cajas y bigotes. En este diagrama cada cuadro corresponde a una variable o predictor. La posición de la caja a lo largo del eje x indica la importancia media de la variable. Una variable se considera más importante si su caja está más a la derecha en el eje. Dentro de cada cuadro, la línea representa la mediana de las puntuaciones de importancia de permutación para esa variable. Si la mediana del cuadro de una variable está más a la derecha que la de otra, sugiere que esta tiene una puntuación de importancia más alta y es más crucial para las predicciones del modelo. El ancho de la caja representa la variabilidad o estabilidad de la importancia del predictor. Un recuadro más ancho indica una mayor variabilidad en las puntuaciones de importancia en diferentes permutaciones, lo que sugiere una menor estabilidad en su contribución a las predicciones del modelo.

Gráfico 3.14: Importancia de permutación[48] modelo XGB (últimos doce meses)



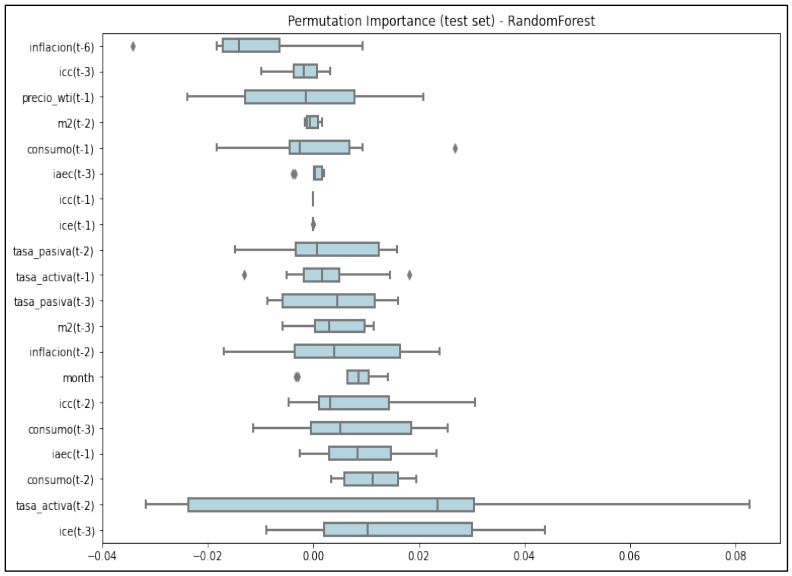
Fuente: Banco Central del Ecuador y Superintendencia de Bancos y Seguro.
Realizado por: Christian Mateo Quiguiri Daquilema.

Para el caso del modelo XGB (Gráfico 3.14), se evidencia que la mayoría de los predictores tienen una importancia relativa en la predicción de los doce últimos meses, sin embargo, es el volumen de crédito rezagado "*consumo(t-3)"* y la "*tasa_pasiva(t-2)*" las que tienen un enorme impacto en el pronóstico.

Por otra parte, para el modelo random forest (ver Gráfico 3.15) los predictores que más influyen en el pronóstico son la "*tasa_activa(t-2)"* y el volumen de crédito rezagado "*consumo(t-2)"*. Sin embargo, la importancia relativa los predictores es mucho menor a la importancia vista en el modelo XGB, denotando las diferencias de los modelos a pesar de estar basados en árboles.

[48] El gráfico de caja y bigotes "boxplot" de la importancia de permutación muestra la relevancia de las variables en un modelo. En el eje Y están las variables y en el eje X la importancia. Los valores más altos en el eje X indican una mayor influencia en el rendimiento predictivo. Además, se muestra la mediana, el rango inter-cuartil y posibles valores atípicos.

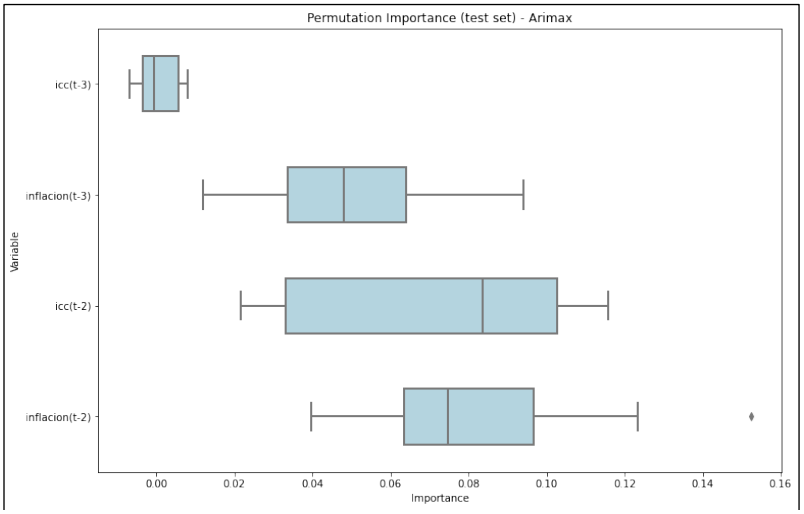
Gráfico 3.15: Importancia de permutación modelo random forest (últimos doce meses)



Fuente: Banco Central del Ecuador y Superintendencia de Bancos y Seguro.
Realizado por: Christian Mateo Quiguiri Daquilema.

Debido a la imposibilidad técnica[49] de calcular la importancia de permutación para el modelo ARIMAX, se ha utilizado una regresión lineal como una aproximación, con el fin de observar la importancia de permutación de en un modelo linear símil a ARIMAX.

Gráfico 3.16: Importancia de permutación[50] modelo ARIMAX (últimos doce meses)



Fuente: Banco Central del Ecuador y Superintendencia de Bancos y Seguro.
Realizado por: Christian Mateo Quiguiri Daquilema.

[49] El cálculo de la importancia de permutación usando la librería Scikit-learn requiere un modelo compatible con la misma, no obstante, el modelo ARIMAX no lo es, por lo tanto, no es posible calcularlo de esta manera.

[50] El gráfico de boxplot de la importancia de permutación muestra la relevancia de las variables en un modelo. En el eje Y están las variables y en el eje X la importancia. Los valores más altos en el eje X indican una mayor influencia en el rendimiento predictivo. El boxplot muestra la mediana, el rango inter-cuartil y posibles valores atípicos. Es útil para identificar las variables más importantes y comprender su impacto en el modelo.

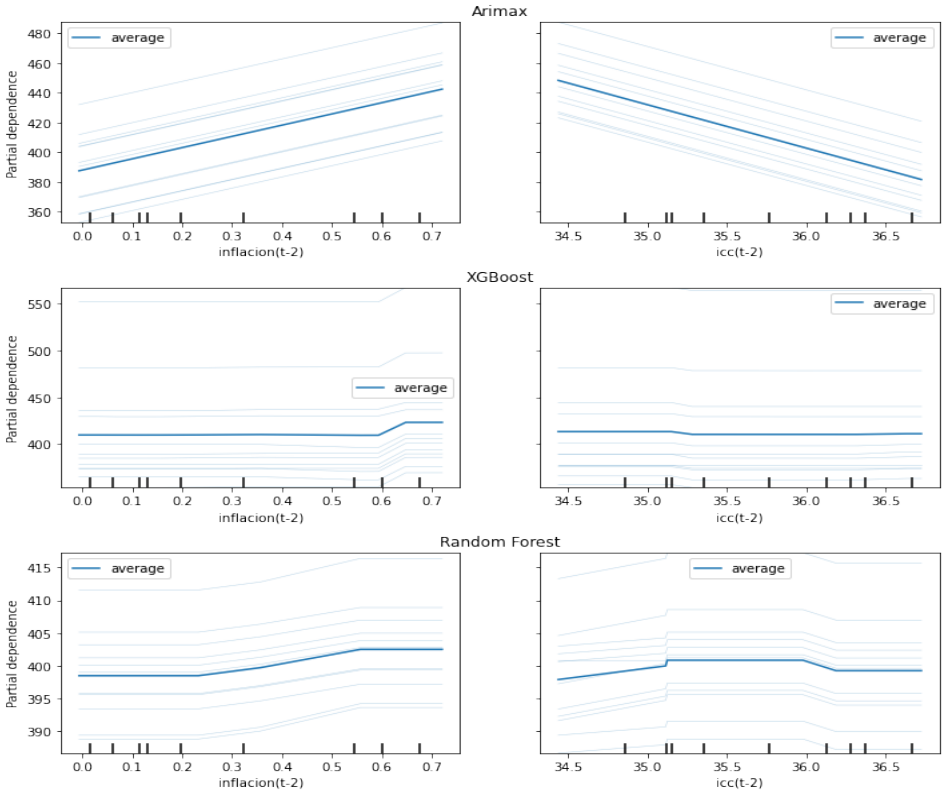
El Gráfico 3.16 muestra que, para el modelo lineal, los cuatro predictores muestran una importancia negativa, que como se mencionó en la fase de preparación de los datos, se debe a que los predictores no presentan una relación lineal significativa con la variable objetivo. Cabe recalcar que también se evidencia este comportamiento en algunas variables de los modelos de ML. La importancia negativa puede darse si al permutar los valores de la variable, en lugar de empeorar el rendimiento del modelo, lo mejora. Esto parece contrario a la intuición, pero puede ocurrir cuando una variable no es relevante para la tarea de predicción o está fuertemente correlacionada con otra variable que también usa el modelo.

Para el primer caso, si una variable no es útil para predecir la variable de destino, permutar sus valores no emporará las predicciones del modelo, sino que podría conducir a ligeras mejoras debido a la aleatoriedad, especialmente cuando el modelo está sobre ajustado al ruido asociado con esa variable. Para el segundo caso, si dos características están fuertemente correlacionadas, contienen información similar, y permutar una de ellas puede hacer que el modelo dependa más de la otra, lo que posiblemente conduzca a una mejora en el rendimiento.

3.7.2.2. Gráficos de dependencia parcial (PDP) y expectativa condicional individual (ICE)

Otra de las herramientas que permite extraer información interpretable de los modelos son los gráficos de Dependencia Parcial (PDP) y la Expectativa Condicional Individual (ICE), el primero que se revisó en la Sección 3.5.4, y que de manera similar al PDP, muestra la dependencia entre la variable objetivo y un predictor de interés. Sin embargo, a diferencia de las gráficas de dependencia parcial, que muestran el efecto promedio de las características de interés, las gráficas ICE visualizan la dependencia de la predicción en una característica para cada observación (mes), con una línea por mes (Scikit-Learn, 2022b).

Gráfico 3.17: Dependencia parcial (PDP) y Expectativa Condicional Individual (ICE) para inflación(t-2) e icc(t-2) sobre conjunto de prueba (últimos doce meses)



Fuente: Banco Central del Ecuador y Superintendencia de Bancos y Seguro.
Realizado por: Christian Mateo Quiguiri Daquilema.

En el Gráfico 3.17, la línea azul más marcada representa a la dependencia parcial (PDP), y las líneas azules delgadas la expectativa condicional individual (ICE). La dependencia parcial indica la relación entre los predictores y la variable objetivo, que para los modelos de ML son claramente ni lineales y ni monótonas, sobre todo para el modelo random forest. Las líneas de la expectativa condicional nos permiten observar la dependencia local, es decir dado una observación o mes, el cambio en la predicción al alterar los valores de una variable individual para este mes, en este caso las variables "inflación*(t-2)"* e *"icc(t-2)"*, manteniendo iguales los valores de las otras variables para los 12 meses o individuos que conforman la serie. Así se observa que, para todos los meses, si la inflación es más alta, el valor de crédito también va a incrementar, aunque se evidencian ligeras diferencias, es decir, no todos los meses incrementan o reaccionan al cambio de la inflación en la misma proporción. Una lectura similar podemos hacer para los otros modelos.

Los resultados anteriores demuestran los beneficios de los modelos de ML, además de la predicción, ya que si bien, no son herramientas de inferencia estadística, las relaciones no lineales y no monótonas que estos algoritmos logran capturan, facilitan la interpretación de las verdaderas relaciones en los datos, por lo que el ML puede brindar información útil no solo cuando el problema es predictivo sino también cuando de entender el fenómeno se trata (Mullainathan & Spiess, 2017).

3.7.3. ¿Es justificada la afirmación de la aproximación universal del Machine Learning?

Si bien el teorema de aproximación universal se aplica sobre todo a modelos de redes neuronales, luego de evaluar dos de los algoritmos más importantes en ML (ver tabla 3.3), los resultados no justifican plenamente esta afirmación. En primer lugar, a pesar de que se evidencia que los modelos de ML utilizados son capaces de capturar relaciones impensables para los modelos econométricos y ofrecer un desempeño predictivo superior, se evidencia la insuficiencia de datos de entrenamiento para generar un modelo lo suficientemente complejo como para que aproxime fielmente la relación entre los predictores y la variable objetivo. Además, se debe recordar que la calidad de la predicción no solo depende de la elección de un modelo, sino de la calidad de los datos, característica que depende de varios factores como errores en la recolección y mantenimiento de datos, variables relevantes omitidas y demás. Luego, es necesario recordar que cuando se hacen predicciones acerca de fenómenos sociales jamás se alcanzará una predicción perfecta, sin importar el modelo que se use, la razón tiene que ver con la inherente aleatoriedad en el comportamiento individual y social que resulta imposible modelar (Boelaert & Ollion, 2018).

3.8. Nuevas Fuentes de Datos: Google Trends

Ahora se profundizará en la aplicación específica de técnicas de ML para mejorar el pronóstico del volumen de crédito mediante la incorporación de nuevos predictores basados en Internet, como los datos de Google Trends.

En el Capítulo 1 se examinaron investigaciones llevadas a cabo en diferentes sectores industriales que resaltan el potencial del ML en la predicción de variables económicas agregadas. La flexibilidad de estos modelos permite incluir gran cantidad de predictores y capturar relaciones no lineales, que en economía sucede a menudo. Un ejemplo de ello es el estudio realizado por Haselbeck, Killinger, Menrad, Hannus, &

Grimm (2022), en el cual se demuestra el beneficio de incorporar factores externos en los modelos de ML, como información meteorológica y días festivos, con el fin de mejorar el rendimiento predictivo. Medeiros et al. (2021) de igual forma menciona la ventaja de pronosticar un agregado económico tan complejo como es la inflación en un entorno de datos abundante, ya que introduce cerca de 900 predictores en su modelo. Hallazgo que suma a las evidencias del potencial de los modelos ML en la previsión económica. Finalmente, (Stavinova et al., 2021) utiliza datos de consultas de búsqueda, tasas bancarias clave y tipos de cambio de divisas para optimizar las predicciones de préstamos hipotecarios.

A continuación, se demuestra el poder predictivo de incorporar datos de Google Trends en el pronóstico de volumen de crédito. A la luz de los resultados, es importante recalcar que aún se necesita realizar más estudios para fundamentar de manera definitiva las capacidades de los métodos de ML en la predicción económica.

Para la comparación se replicó la misma metodología explicada a detalle en las secciones anteriores (Secciones 3.3-3.6), por lo que la preparación de datos, ingeniería de variables, así como la construcción y evaluación de los modelos, es la misma. Cabe recalcar que las series de Google Trends están disponibles a partir de 2015 por lo que el número total de observaciones se redujo de aproximadamente 200 a 120 para para modelos con los nuevos predictores. La introducción de los datos de Google Trends estuvo también motivada por la creciente digitalización de los sistemas financieros, donde las búsquedas en línea de préstamos se han vuelto más frecuentes.

3.8.1. ¿Existen beneficios en el pronóstico al incorporar predictores externos?

Al evaluar los modelos, se encontró que el modelo random forest superó a los demás, arrojando un error porcentual absoluto medio (MAPE) de 8,73 (ver Tabla 3.5) con la inclusión de consultas de Google.

Tabla 3.5: Evaluación de los modelos mediante MAPE[51], RMSE[52] y Desviación Estándar[53], incluyendo predictores novedosos Google Trends

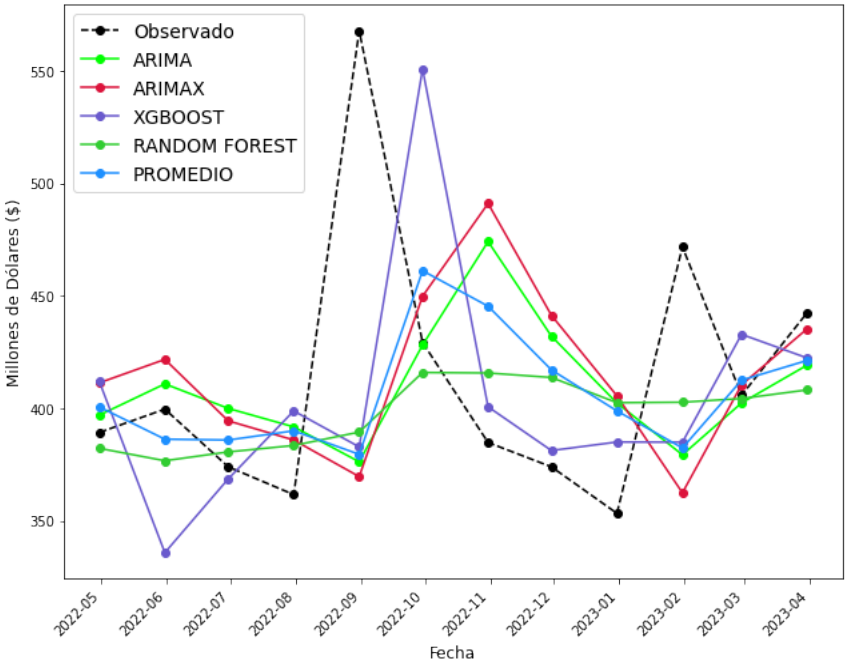
Modelo	MAPE	RMSE	DESV. STD MAPE
ARIMA	11.02	71.45	9.71
ARIMAX	12.44	77.55	10.12
XGB	11.57	**73.53**	9.91
RANDOM FOREST	**8.76**	60.51	**9.0**
ENSAMBLE (PROMEDIO)	10.27	66.92	9.18

Fuente: Banco Central del Ecuador y Superintendencia de Bancos y Seguro.
Realizado por: Christian Mateo Quiguiri Daquilema.

Sin los nuevos predictores, el MAPE fue superior a 9,93, ahora el MAPE se redujo hasta 8.76 lo que indica la contribución significativa de los datos de Google Trends para mejorar el rendimiento de la previsión. De igual forma, en el Gráfico 3.18 se puede observar las predicciones de los modelos.

[51] El MAPE (Mean Absolute Percentage Error) es una métrica de evaluación de modelos de predicción que calcula el promedio del porcentaje absoluto de error entre las predicciones y los valores reales. Esta métrica proporciona una medida relativa del error promedio en relación con el tamaño de los valores reales y se expresa como un porcentaje.
[52] El RMSE (Root Mean Square Error) es otra métrica comúnmente utilizada para evaluar la precisión de un modelo de predicción. Calcula la raíz cuadrada del promedio de los errores al cuadrado entre las predicciones y los valores reales. El RMSE proporciona una medida de la desviación estándar de los errores y se expresa en la misma unidad que la variable objetivo.
[53] La desviación estándar del MAPE (Mean Absolute Percentage Error) es una medida complementaria que proporciona información adicional sobre la variabilidad de los errores en términos porcentuales. Calculando la desviación estándar del MAPE, se puede obtener una medida de cuánto varían los porcentajes de errores absolutos en relación con el promedio del MAPE. Esto permite evaluar la consistencia o estabilidad de la precisión del modelo a lo largo del tiempo.

Gráfico 3.18: Comparación de predicciones de modelos econométricos y de ML para el volumen de crédito en Ecuador incluyendo predictores Google Trends



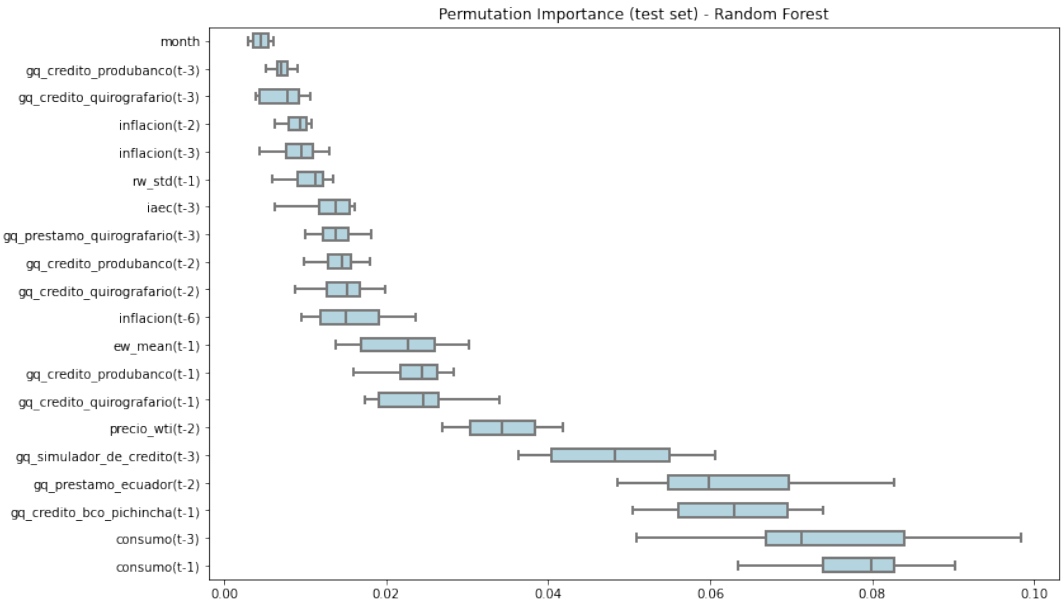
Fuente: Banco Central del Ecuador y Superintendencia de Bancos y Seguro.
Realizado por: Christian Mateo Quiguiri Daquilema.

Estos hallazgos se alinean con los resultados presentados por Medeiros et al. (2021) quienes encontraron que el modelo random forest es efectivo en un entorno de datos enriquecido con predictores externos. El resultado es aún más sorprendente ya que a pesar de que la muestra utilizada para entrenar los modelos con los nuevos predictores fue alrededor de 80 observaciones (meses) más corta, la precisión de la predicción mejoró. Por lo que se puede afirmar que los predictores novedosos de las búsquedas de Google Trends contribuyeron a la riqueza del conjunto de datos, mejorando el poder predictivo del modelo random forest.

3.8.2. ¿Qué tan importantes son los nuevos predictores en el pronóstico?

El análisis de la importancia de las variables reveló que tres de las cinco variables más importantes eran consultas de Google (ver Gráfico 3.19), lo que enfatiza aún más el valor de estos predictores.

Gráfico 3.19: Importancia de Permutación[54] modelo random forest



Fuente: Banco Central del Ecuador y Superintendencia de Bancos y Seguro.
Realizado por: Christian Mateo Quiguiri Daquilema.

Dada la tendencia hacia la digitalización de los sistemas financieros, no sorprende que las consultas de Google hayan surgido como poderosos predictores. A medida que más personas buscan préstamos en línea, estas consultas de búsqueda brindan información en tiempo real sobre el comportamiento del consumidor y la demanda de crédito. Esto último lo evidencia al ver en el número de rezagos de los predictores más relevantes de Google Trends, la mayoría tienen solo un rezago (ver Gráfico 3.16).

3.8.3. ¿Se puede asegurar que añadir predictores no tradicionales como Google Trends mejorará la predicción de un agregado económico?

Si bien este estudio logró resultados prometedores, es importante recordar que el ML y su aplicación en el pronóstico de agregados económicos tiene sus limitaciones. Futuras investigaciones podrían explorar el uso de otras fuentes de datos basadas en Internet, aplicar diferentes modelos de ML al mismo conjunto de predictores, o este conjunto de predictores para otros agregados económicos. Por lo tanto, no se puede asegurar que añadir predictores externos puede conducir a una mejora en la precisión del pronóstico, sino que aún se requiere más investigación en torno al tema.

[54] El gráfico de boxplot de la importancia de permutación muestra la relevancia de las variables en un modelo. En el eje Y están las variables y en el eje X la importancia. Los valores más altos en el eje X indican una mayor influencia en el rendimiento predictivo. El boxplot muestra la mediana, el rango inter-cuartil y posibles valores atípicos. Es útil para identificar las variables más importantes y comprender su impacto en el modelo.

Capítulo 4: Pronóstico en los Ciclos Económicos

Las variables económicas a menudo exhiben un comportamiento diferente durante las expansiones y las recesiones de los ciclos económicos. Generalmente los modelos que exhiben un buen desempeño al pronosticar la dinámica promedio de la variable objetivo pueden mostrar un bajo rendimiento en ciertos escenarios como una aceleración económica, una recesión económica, crisis bancarias entre otros. Una particularidad encontrada en varios estudios es que con frecuencia los mayores errores en el pronóstico se dan en los puntos de inflexión entre ciclos económicos, sobre todo en las recesiones (Klein, 2013).

El filtro Baxter-King (BK) es una herramienta ampliamente utilizada en el análisis de ciclos económicos para identificar las distintas fases de dichos ciclos. Su función principal consiste en aislar las fluctuaciones del ciclo económico en los datos, lo que facilita la diferenciación entre períodos de expansión económica y recesión. En este sentido, se aplica el filtro BK a los datos de la serie trimestral del Producto Interno Bruto (PIB) real, para el periodo comprendido entre el primer trimestre del año 2000 al cuarto trimestre del 2022. De esta manera, el filtro BK permite discernir y analizar de forma más precisa los momentos de crecimiento y contracción económica (Drehmann, Borio, & Tsatsaronis, 2012).

La finalidad de este capítulo es evaluar la capacidad de los modelos para pronosticar las dinámicas del volumen de crédito de los hogares en diferentes fases de los ciclos económicos. De esta manera se busca tener una mejor comprensión de la adaptabilidad y eficacia del modelo durante los períodos de expansión económica y recesión. Además, se analizará un periodo inusual, como lo fue la pandemia de COVID-19.

Adicionalmente, se examina el rendimiento del modelo durante un período atípico: la pandemia de COVID-19. Este período se caracteriza por importantes perturbaciones y un profundo alejamiento de los patrones económicos regulares. Al probar el modelo en este escenario, se busca evaluar su resiliencia frente a un estrés económico extremo y una alta volatilidad, tema que ha tomado notoriedad pues ha relevado la enorme incertidumbre económica a la que un evento como este puede conducir (Baker, Bloom, Davis, & Terry, 2020).

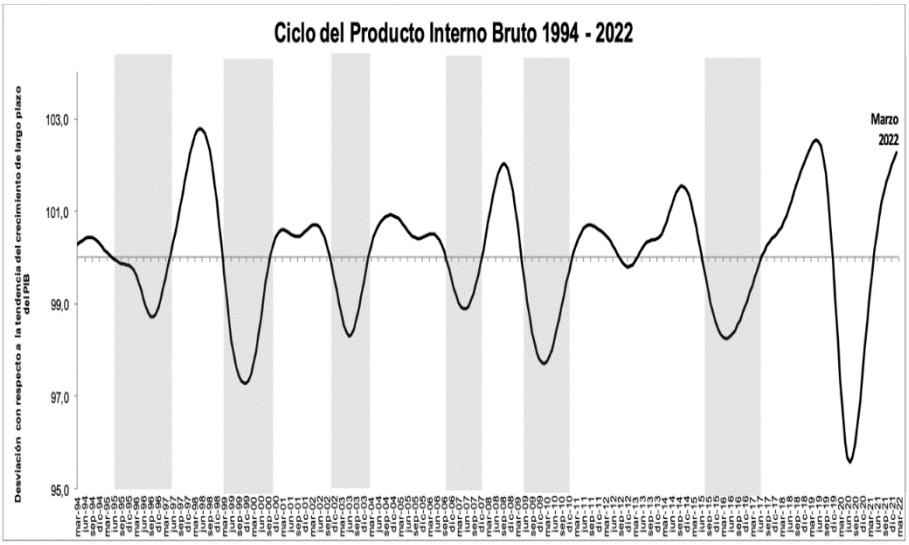
4.1. El crédito y los ciclos económicos en Ecuador

Entre los indicadores económicos importantes para el ciclo económico, el volumen de crédito al sector privado ocupa un lugar destacado. Esto se debe a que este es un indicador clave de la actividad económica. Cuando la economía está en expansión, generalmente hay un aumento en el crédito a medida que las empresas toman prestado para invertir y expandirse, y los consumidores toman prestado para el consumo. Por el contrario, en una recesión, el crédito tiende a contraerse. Este comportamiento coincidente con el ciclo económico hace que el crédito forme parte de los indicadores coincidentes, a diferencia de series como la producción de petróleo o los depósitos a la vista, que por su naturaleza, suelen adelantarse al ciclo económico (Erráez, 2014).

En el país, la institución que reporta la información correspondiente a los ciclos es el Banco Central del Ecuador (BCE) y como lo detalla en sus informes, la metodología que se aplica para determinar los ciclos es la recomendada por la Organización para la Cooperación y el Desarrollo Económico (OCDE) (Banco Central del Ecuador, 2021).

El filtro BK, seleccionado aquí para identificar los ciclos económicos, ha sido diseñado para aislar los componentes cíclicos de una serie de tiempo eliminando los componentes irregulares y de tendencia, enfocándose esencialmente en las fluctuaciones que ocurren dentro de un rango de frecuencia específico (Baxter & King, 1995). Este rango a menudo se elige para que corresponda con las frecuencias típicas del ciclo económico del objetivo en cuestión. En el caso de Ecuador, esto suele corresponder a ciclos que duran de 5 a 18 trimestres (Banco Central del Ecuador, 2022). Estos rangos a su vez son consistentes con los recomendados por los autores del método – Baxter & King (1995), y en este ejercicio han logrado una coincidencia exacta a los ciclos reportados por el BCE (Ver Gráfico 4.1 y 4.2), tanto en ciclos de expansión como de recesión.

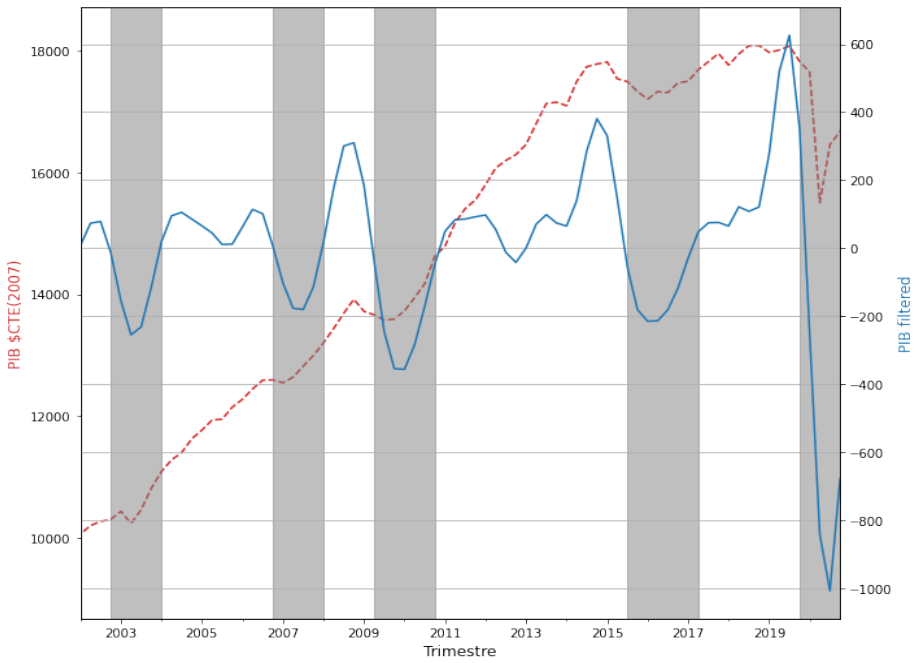
Gráfico 4.1: Gráfico de los ciclos del PIB reportados por el Banco Central del Ecuador



Fuente: Ilustración obtenida de (Banco Central del Ecuador, 2022)
Realizado por: Christian Mateo Quiguiri Daquilema.

Los ciclos económicos identificados, así como la serie original PIB en términos corrientes y la misma serie luego de aplicar el filtro (BK) se pueden observar en el Gráfico 4.2.

Gráfico 4.2: Ciclos económicos, PIB en términos corrientes y PIB filtrado



Fuente: Banco Central del Ecuador.
Realizado por: Christian Mateo Quiguiri Daquilema.

4.2. Pronóstico en los ciclos económicos

La comparación en los ciclos económicos se realiza en el muy corto plazo, específicamente para un periodo de un mes hacia delante (en tiempo real). Este enfoque tiene dos ventajas: por una parte, emula los métodos de pronóstico en tiempo real de los agentes económicos y los Bancos Centrales; y por otra, reduce la sensibilidad de los errores a rápidos cambios entre las fases de los ciclos económicos (Klein, 2013).

Con la intención de garantizar una cantidad suficiente de datos para el entrenamiento de los modelos, el pronóstico se limita a un ciclo económico completo, el cual incluye una recesión y una expansión. Además, se examina el periodo de recesión ocasionado por el SARS-Cov2. El ciclo se inicia con una recesión en mayo de 2015 y concluye con una fase de expansión que se extiende hasta diciembre de 2019. La recesión atribuida a la pandemia comienza en el primer trimestre de 2020 y perdura hasta el primer trimestre de 2021. Para llevar a cabo la comparación, se recurre a los modelos empleados en el Capítulo 3, siguiendo el mismo proceso de preparación de datos, ingeniería de variables y construcción de modelos. Asimismo, para evaluar la estabilidad de los modelos, se emplean el MAPE, el RMSE, y la varianza.

4.2.1. Resultados del pronóstico en ciclos económicos:

En la Tabla 4.1 se presenta, a modo de resumen, los resultados de la experimentación bajo los distintos escenarios considerados. A partir de estos resultados se ha identificado varios puntos importantes que evidencia las ventajas y limitaciones de los modelos. Así, como se evidencia en el resumen de la Tabla 4.1, en los ciclos económicos que comprenden tanto fases de expansión como de recesión, el modelo random forest exhibe el mejor desempeño con el MAPE más bajo. Esto sugiere que ofrece las predicciones más precisas cuando se aplica a condiciones económicas cíclicas típicas. Sin embargo, es importante tener en cuenta que el modelo BASE, que simplemente usa un dato anterior para hacer predicciones, también demuestra un MAPE relativamente bajo, especialmente durante la fase de expansión. Esto podría indicar un posible problema de estabilidad en los modelos, indicando que estos no llegan a converger debido al conjunto de datos limitado. Si los modelos más sofisticados no pueden superar significativamente a un modelo de serie retrasada simple, plantea dudas sobre su capacidad para generalizar a partir de una cantidad tan limitada de datos (Hyndman & Athanasopoulos, 2018).

Tabla 4.1: MAPE, RMSE, y Desviación Estándar de la predicción un mes hacia delante en los ciclos económicos.

	MAPE				RMSE				DESVIACIÓN ESTÁNDAR (MAPE)			
Modelo	**Periodo Completo**	**Expansión**	**Recesión**	**COV-19**	**Periodo Completo**	**Expansión**	**Recesión**	**COV-19**	**Periodo Completo**	**Expansión**	**Recesión**	**COV-19**
BASE	17.92	9.27	15.22	48.00	35.98	24.84	25.68	89.62	40.23	7.70	**8.13**	91.23
ARIMA	17.51	**9.48**	16.50	40.65	35.26	**25.67**	26.70	77.62	33.91	6.83	9.61	76.96
ARIMAX	19.76	11.78	19.24	42.36	39.79	30.93	31.61	80.72	35.98	7.96	10.72	82.04
XGB	18.56	13.07	16.20	38.66	39.91	35.11	27.78	78.98	24.53	10.72	10.67	50.77
RANDOM FOREST	**16.43**	10.59	16.83	**27.89**	**35.21**	29.39	27.27	**60.97**	**20.35**	7.95	11.74	**42.07**
ENSAMBLE	17.20	10.91	**15.63**	37.16	35.98	29.45	**25.97**	74.07	28.05	**7.39**	9.45	62.38

Fuente: Banco Central del Ecuador, Superintendencia de Bancos.
Realizado por: Christian Mateo Quiguiri Daquilema.

Asimismo, se evidencia que los mayores errores de pronóstico se dan durante las recesiones. En las expansiones, el error de predicción medido por el MAPE es la mitad del error de todo el periodo, indicándonos que los modelos se van a desempeñar en la fase expansiva del ciclo.

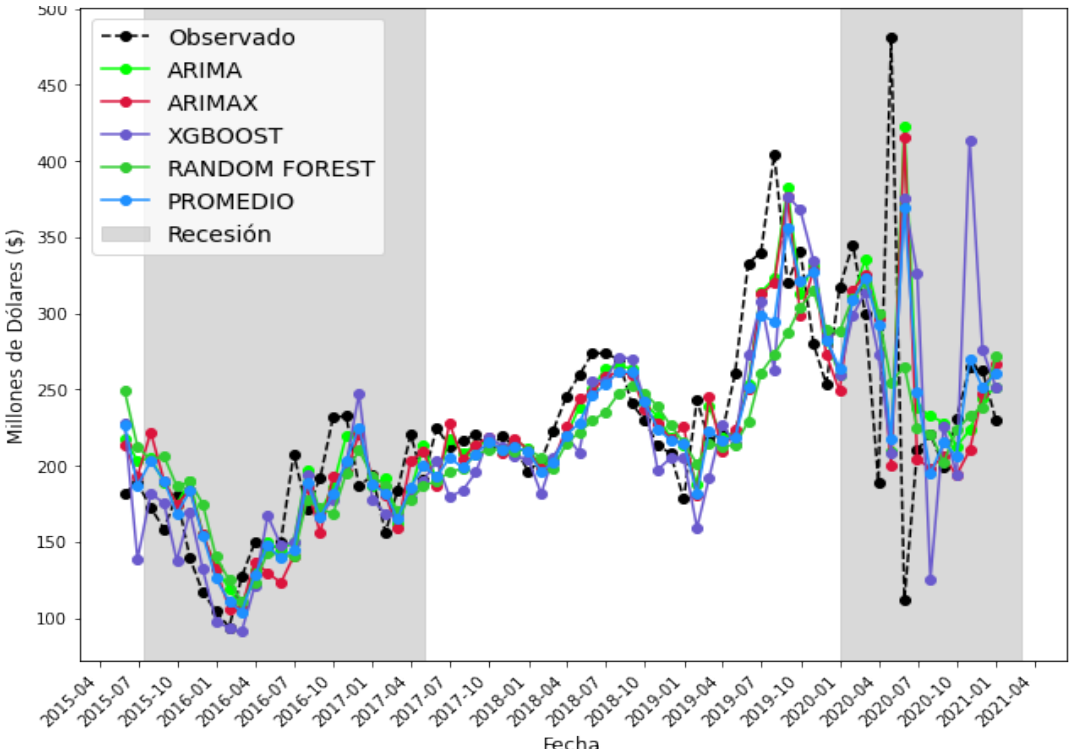
4.2.2. Período COVID-19:

La pandemia de COVID-19 ha sido un evento inusual y de alto riesgo con repercusiones a nivel mundial, cuyas secuelas aún se siguen sintiendo en todos los ámbitos, especialmente en el ámbito económico. Este suceso planteó un desafío sin precedentes para la predicción económica, generando cambios drásticos e inesperados en numerosos indicadores económicos y dando lugar a un nivel de volatilidad e incertidumbre que rara vez se ha experimentado en la historia moderna (Baker et al., 2020).

Además, según un informe publicado por el Fondo Monetario Internacional (FMI) en 2021, la pandemia de COVID-19 ha tenido un impacto significativo en la economía mundial, provocando una contracción económica generalizada y afectando negativamente el crecimiento, el empleo y el comercio en la mayoría de los países. El informe destaca la necesidad de adaptar los modelos de previsión económica para tener en cuenta los efectos duraderos de la pandemia y la incertidumbre en curso (Agarwal & Gopinath, 2021).

Al observar el Gráfico 4.2 es evidente que los meses durante el pico de la pandemia de COVID-19 muestran los patrones más irregulares para el volumen de crédito. En línea con las tendencias económicas mundiales, hay una caída abrupta en el volumen de crédito en marzo 2020, lo que refleja el impacto causado por el inicio de la pandemia. A las restricciones adoptadas por el gobierno se debe sumar una incertidumbre económica generalizada que provocó una contracción en las actividades crediticias.

Gráfico 4.3: Pronóstico modelos en ciclos económicos y periodo SARS-COV2



Fuente: Banco Central del Ecuador.
Realizado por: Christian Mateo Quiguiri Daquilema.

Después de esto, sin embargo, se observa una recuperación casi inmediata en mayo 2020. Este fuerte repunte puede ser el reflejo de varios factores. Por un lado, puede señalar la inyección de liquidez por parte de instituciones financieras y organismos gubernamentales para estimular la actividad económica (El Universo, 2020). En segundo lugar, esto también podría deberse a la adaptación de empresas y consumidores al nuevo entorno económico, con un repunte de la demanda de crédito.

Teniendo en cuenta que los modelos de pronóstico convencionales se basan en gran medida en tendencias pasadas y fluctuaciones cíclicas regulares, ante un evento tan atípico se muestran deficientes. En este contexto, la previsión se convirtió menos en la proyección de tendencias pasadas hacia el futuro y más en medir el impacto de una crisis nueva y multifacética (Baker et al., 2020).

A pesar de estos desafíos, el modelo random forest demostró su solidez al lograr el MAPE más bajo entre los modelos probados durante el período de COVID-19, demostrando la adaptabilidad de un modelo relativamente simple frente a un escenario sin precedente. Esto se alinea con la literatura existente que subraya la adaptabilidad y resiliencia de random forest bajo condiciones complejas e inciertas (Hastie, 2009)

4.2.3. Equilibrio Sesgo/Varianza

Al examinar el rendimiento del modelo, es fundamental tener en cuenta el equilibrio sesgo/varianza, o la compensación entre el sesgo y la varianza. El sesgo de un modelo (representado por MAPE) muestra qué tan cerca están las predicciones, en promedio, de los valores reales, mientras que la varianza (evaluada aquí por la desviación estándar de MAPE) muestra cuánto varían las predicciones para diferentes conjuntos de datos. Un modelo con bajo sesgo y baja varianza es ideal. El modelo random forest exhibe un sesgo bajo y una varianza relativamente baja, y se desempeña de manera consistente en diferentes escenarios económicos.

Según el MAPE y los resultados de la desviación estándar, el modelo random forest generalmente funciona mejor en diferentes ciclos económicos y el período inusual de COVID-19. Sin embargo, vale la pena señalar el desempeño comparable del modelo BASE, signo de que los modelos utilizados no son mucho mejores que la simple predicción a partir del valor pasado. Esto podría sugerir un problema de estabilidad en los

modelos más complejos debido a los datos limitados, por lo que se debe tener cuidado al interpretar los resultados.

El ejercicio de pronosticar los indicadores económicos en ciclos económicos, como el volumen de crédito, tanto en situaciones normales como durante eventos sin precedentes como el COVID-19, presenta desafíos y oportunidades. En este escenario, los modelos de ML prometen ser una solución. Sin embargo, el análisis realizado demuestra que su efectividad está intrínsecamente ligada a los datos en los que son entrenados. Además, se evidenció que en ocasiones un "modelo" simple, como utilizar la variable rezagada de un periodo, puede tener un mejor desempeño que un modelo complejo como XGB.

Se confirma lo mencionado en la literatura, que indica que ningún modelo está preparado para funcionar correctamente en un evento de tal magnitud (Baker et al., 2020). A pesar de que se han demostrado las limitaciones de los modelos, tanto de ML como econométricos, comprender estas limitaciones en mayor profundidad puede arrojar luz sobre el camino hacia el cual deben dirigirse la investigación futura.

Capítulo 5: Conclusiones

A pesar de sus limitaciones, el modelo ARIMA muestra una sólida capacidad de pronóstico a corto plazo, con un error de predicción muy cercano al mejor modelo de ML en los tres casos de comparación. Es importante destacar que este modelo, aunque univariado y sensible a cambios estructurales, presenta una buena capacidad de predicción, pero su limitación radica en la falta de inclusión de variables externas y su restricción a la linealidad en los datos.

El modelo Random Forest sobresale en nuestro estudio, logrando un equilibrio óptimo entre precisión y varianza, incluso en condiciones de datos limitados. Sin embargo, es relevante destacar que el contexto de datos limitados influye significativamente en la capacidad de pronóstico de los modelos de Machine Learning, especialmente el modelo XGBoost, que enfrenta dificultades para converger debido al número limitado de datos disponibles.

Nuestros resultados sugieren que los datos no tradicionales, como las tendencias de búsqueda en Google Trends, pueden mejorar los modelos de predicción económica al reflejar las preferencias cambiantes de las personas. Esto es sorprendente, considerando que la muestra utilizada para entrenar los modelos con estos nuevos predictores fue aproximadamente 80 observaciones (meses) más corta, pero la precisión de la predicción mejoró. Sin embargo, se necesita más investigación en diferentes usos y contextos para confirmar su utilidad en la predicción económica.

Nuestro estudio respalda la literatura existente sobre la predicción de ciclos económicos al demostrar que los mayores errores se producen durante las recesiones en todos los modelos, mientras que los errores son significativamente menores durante las expansiones. Esto subraya la importancia de desarrollar modelos precisos que puedan pronosticar con éxito tanto el crecimiento como la contracción económica. El enriquecimiento de los datos con predictores no tradicionales puede ser un camino prometedor para lograrlo.

Aunque tanto el Machine Learning como los modelos econométricos tienen limitaciones, queda claro que ningún modelo está completamente preparado para funcionar

correctamente en eventos como los ciclos económicos recesivos o eventos inusuales como el surgimiento del SARS-CoV-2. Comprender estas limitaciones a fondo puede orientar la dirección de futuras investigaciones. Por lo tanto, es necesario llevar a cabo más investigaciones para establecer definitivamente las capacidades del Machine Learning, así como del uso de predictores no tradicionales en la predicción económica, ya que, aunque ha mostrado promesas, se requiere evidencia empírica adicional para validar su eficacia en diversos contextos y aplicaciones

BIBLIOGRAFÍA:

Abu-Rmileh, A. (2019). The Multiple faces of 'Feature importance' in XGBoost. Retrieved July 6, 2022, from https://towardsdatascience.com/be-careful-when-interpreting-your-features-importance-in-xgboost-6e16132588e7#:~:text="The Gain implies the relative,important for generating a prediction.

Agarwal, R., & Gopinath, G. (2021). Pandemic Economics: A broad-based economic recovery requires an end to the pandemic. *Finance and Development*, *58*(4), 10–11.

Alim, M., Ye, G. H., Guan, P., Huang, D. S., Zhou, B. Sen, & Wu, W. (2020). Comparison of ARIMA model and XGBoost model for prediction of human brucellosis in mainland China: A time-series study. *BMJ Open*, *10*(12), 1–8. https://doi.org/10.1136/bmjopen-2020-039676

Anuja, N. (2022). L1 and L2 Regularization Methods, Explained. Retrieved February 20, 2022, from https://builtin.com/data-science/l2-regularization

Athey, S., & Imbens, G. W. (2019). Machine Learning Methods That Economists Should Know about. *Annual Review of Economics*, *11*(March), 685–725. https://doi.org/10.1146/annurev-economics-080217-053433

Baker, S., Bloom, N., Davis, S., & Terry, S. (2020). COVID-Induced Economic Uncertainty. *National Bureau of Economic Research*, 17. Retrieved from http://www.nber.org/papers/w26983

Banco Central del Ecuador. (2021). *Ciclo Económico del Ecuador: Resultados al primer trimestre de 2021*. Retrieved from https://contenido.bce.fin.ec/documentos/Estadisticas/SectorReal/Previsiones/IDEAC/CicloEconIT2021.pdf

Banco Central del Ecuador. (2022). *Ciclo Económico del Ecuador: Resultados al Primer Trimestre 2022*. Retrieved from https://contenido.bce.fin.ec/documentos/Estadisticas/SectorReal/Previsiones/IDEAC/CicloEconIT2022.pdf

Baxter, M., & King, R. G. (1995). Approximate band-pass filters for economic time series. *NBER Working Paper Series*, *5022*, 1–53.

Beck, T., Büyükkarabacak, B., Rioja, F. K., & Valev, N. T. (2012). Who Gets the Credit? And Does It Matter? Household vs. Firm Lending Across Countries. *The B.E. Journal of Macroeconomics*, *12*(1). https://doi.org/doi:10.1515/1935-1690.2262

Blanchard, O. (2006). *Macroeconomics 4th Edition. Pearson Prentice Hall, New Jersey*.

Boelaert, J., & Ollion, É. (2018). The great regression machine learning, econometrics, and the future of quantitative social sciences. *Revue Francaise de Sociologie*, *59*(3), 475–506. https://doi.org/10.3917/rfs.593.0475

Bouri, E., Gkillas, K., Gupta, R., & Pierdzioch, C. (2021). Forecasting Realized Volatility of Bitcoin: The Role of the Trade War. *Computational Economics*, *57*(1), 29–53. https://doi.org/10.1007/s10614-020-10022-4

Breiman, L. (2001a). Random Forest. *Machine Learning*, *45*, 5–32. https://doi.org/https://doi.org/10.1023/A:1010933404324

Breiman, L. (2001b). Statistical modeling: The two cultures. *Statistical Science*, *16*(3), 199–215. https://doi.org/10.1214/ss/1009213726

Brownlee, J. (2014). 14 Different Types of Learning in Machine Learning. Retrieved July 27, 2021, from https://machinelearningmastery.com/types-of-learning-in-machine-learning/

Brownlee, J. (2019). How to Convert a Time Series to a Supervised Learning Problem

in Python. Retrieved July 7, 2021, from https://machinelearningmastery.com/convert-time-series-supervised-learning-problem-python/
Chakraborty, T., Chakraborty, A. K., Biswas, M., Banerjee, S., & Bhattacharya, S. (2021). Unemployment Rate Forecasting: A Hybrid Approach. *Computational Economics*, *57*(1), 183–201. https://doi.org/10.1007/s10614-020-10040-2
Charpentier, A., Flachaire, E., & Ly, A. (2018). Econometrics and machine learning. *Economie et Statistique*, *2018*(505–506), 147–169. https://doi.org/10.24187/ecostat.2018.505d.1970
Chen, J. M., & Baker, A. (2020). Forecasting Mortgage Demand: An Application of Traditional Methods, Machine Learning, and Neural Networks. *SSRN Electronic Journal*, 1–111. https://doi.org/10.2139/ssrn.3656924
Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, *13-17-Augu*, 785–794. https://doi.org/10.1145/2939672.2939785
Chen, T., He, T., Michaël, B., & Tang, Y. (2022). Understand your dataset with XGBoost. Retrieved July 1, 2021, from https://cran.r-project.org/web/packages/xgboost/vignettes/discoverYourData.html#numeric-v.s.-categorical-variables
Choi, H., & Varian, H. R. (2009). Predicting the present with Google Trends. Retrieved from https://static.googleusercontent.com/media/www.google.com/es//googleblogs/pdfs/google_predicting_the_present.pdf
Cook, J. (2010). What is the difference between machine learning and statistics? Retrieved July 15, 2022, from https://stackoverflow.com/questions/4205105/whats-the-difference-between-machine-learning-and-statistics#:~:text=Main difference%3A,relying on rules based programming.
Cotton, R. (2022). Top Machine Learning Algorithms. Retrieved March 5, 2021, from https://www.datacamp.com/cheat-sheet/machine-learning-cheat-sheet
datacamp. (2023). Top Techniques to Handle Missing Values Every Data Scientist Should Know.
Deep AI. (2023). Loss Function. Retrieved February 26, 2021, from https://deepai.org/machine-learning-glossary-and-terms/loss-function
Drehmann, M., Borio, C., & Tsatsaronis, K. (2012). Characterising the Financial Cycle: Don't Lose Sight of the Medium Term! *BIS Working Papers*, (380), 1–38. Retrieved from http://ideas.repec.org/p/bis/biswps/380.html
Duarte, J. J., Montenegro González, S., & Cruz, J. C. (2021). Predicting Stock Price Falls Using News Data: Evidence from the Brazilian Market. *Computational Economics*, *57*(1), 311–340. https://doi.org/10.1007/s10614-020-10060-y
El Universo. (2020). Lenín Moreno anuncia medidas económicas y llama a un gran acuerdo nacional ante la triple emergencia que vive Ecuador. Retrieved June 10, 2023, from https://www.eluniverso.com/noticias/2020/04/10/nota/7810659/lenin-moreno-acuerdo-emergencia-ecuador-renegociar-deuda-reforma/
Erráez, J. (2014). Sistema de Indicadores del Ciclo de Crecimiento Económico. *Banco Central Del Ecuador*, (77), 37. Retrieved from https://contenido.bce.fin.ec/documentos/PublicacionesNotas/Catalogo/NotasTecnicas/nota77.pdf
Friedman, J. H. (2002). Stochastic gradient boosting. *Computational Statistics & Data Analysis*, *38*(4), 367–378. https://doi.org/https://doi.org/10.1016/S0167-

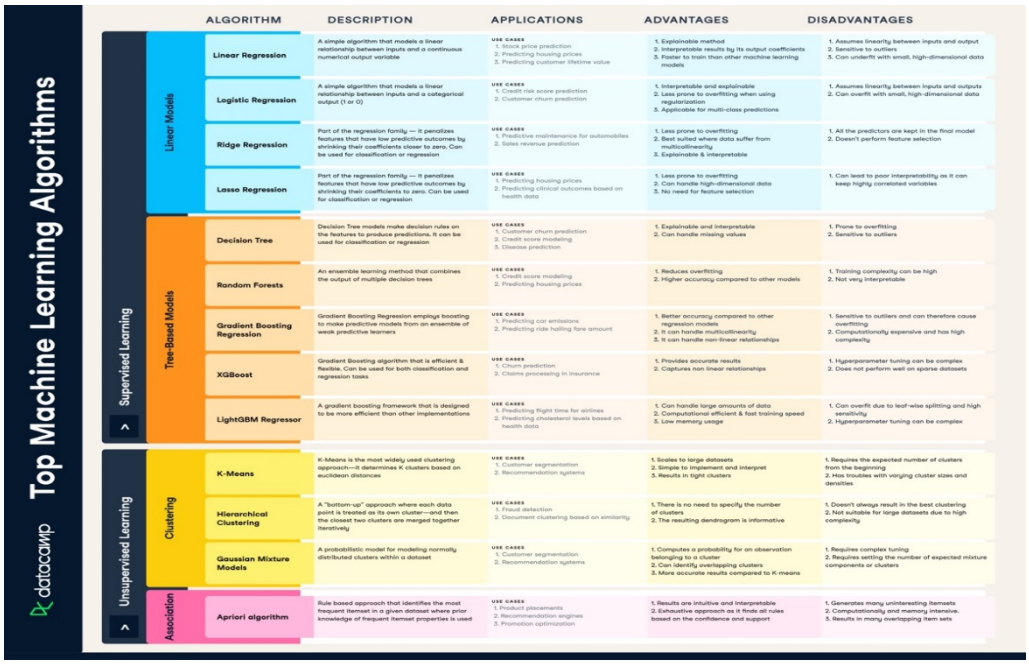
9473(01)00065-2
Frost, J. (2020). Choosing Between Spearman's and Pearson's Correlation. Retrieved August 4, 2021, from https://statisticsbyjim.com/jim_frost/
Gattin-turkalj, K., Ljubaj, I., Martinis, A., & Mrkalj, M. (2007). Estimating Credit Demand in Croatia Draft version. *Croatian National Bank Paper*, (April), 1–36.
Geeks for Geeks. (2022). Hyperparameter tuning. Retrieved March 11, 2021, from https://www.geeksforgeeks.org/hyperparameter-tuning/
Geeks For Geeks. (2022). Bagging vs Boosting in Machine Learning. Retrieved June 6, 2021, from https://www.geeksforgeeks.org/bagging-vs-boosting-in-machine-learning/
Gogas, P., & Papadimitriou, T. (2021). Machine Learning in Economics and Finance. *Computational Economics*, *57*(1), 1–4. https://doi.org/10.1007/s10614-021-10094-w
Gujarati, D. N., & Porter, D. C. (2010). *Econometría Quinta Edición*. (J. M. Chacón, Ed.) (5th ed.). México: McGRAW-HILL/INTERAMERICANA EDITORES, S.A. DE C.V.
Haitao, D. (2017). What exactly is the difference between a parametric and non-parametric model? Retrieved November 1, 2022, from https://stats.stackexchange.com/questions/268638/what-exactly-is-the-difference-between-a-parametric-and-non-parametric-model
Hardy, R. (2019). Can I continue ARIMA model despite my time time series has heterodasticity?
Haselbeck, F., Killinger, J., Menrad, K., Hannus, T., & Grimm, D. G. (2022). Machine Learning Outperforms Classical Forecasting on Horticultural Sales Predictions. *Machine Learning with Applications*, *7*, 100239. https://doi.org/https://doi.org/10.1016/j.mlwa.2021.100239
Hastie, T. et. all. (2009). The Elements of Statistical Learning. *The Mathematical Intelligencer*, *27*(2), 83–85. Retrieved from http://www.springerlink.com/index/D7X7KX6772HQ2135.pdf
Hyndman, R. J. (2010). The ARIMAX model muddle. Retrieved September 1, 2022, from https://robjhyndman.com/hyndsight/arimax/
Hyndman, R. J., & Athanasopoulos, G. (2018). *Forecasting: principles and practice* (2nd ed.). Melbourne: OTexts. Retrieved from OTexts.com/fpp2
James, G., DanielaWitten, Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning with application in R*.
Java T Point. (2022). Bias and Variance in Machine Learning. Retrieved October 30, 2021, from https://www.javatpoint.com/bias-and-variance-in-machine-learning
Klein, L. R. (2013). *Economic Forecasting*. *Kyklos* (Vol. 12). https://doi.org/10.1111/j.1467-6435.1959.tb01823.x
Kurtz, Z. (2018). Translating Between Statistics and Machine Learning. Retrieved October 31, 2021, from https://insights.sei.cmu.edu/blog/translating-between-statistics-and-machine-learning/
Kutner, Mi. H., Nachtsheim, C., Neter, J., & Li, W. (2005). *Applied Linear Statistical Models 5th Edition*. *Journal of Quality Technology* (5th ed., Vol. 29). New York: McGraw-Hill Irwin. https://doi.org/10.1080/00224065.1997.11979760
Levieuge, G. (2017). Explaining and forecasting bank loans. Good times and crisis. *Applied Economics*, *49*(8), 823–843. https://doi.org/10.1080/00036846.2016.1208350
Li, J. (2022). What is set.seed function in R. Retrieved February 1, 2021, from https://www.projectpro.io/recipes/what-is-set-seed-function#:~:text=,
Lima, L. R., Godeiro, L. L., & Mohsin, M. (2021). *Time-Varying Dictionary and the*

Predictive Power of FED Minutes. *Computational Economics* (Vol. 57). Springer US. https://doi.org/10.1007/s10614-020-10039-9
Luka, A. (2020). Rolling and Expanding Windows For Dummies. Retrieved May 16, 2023, from https://robotwealth.com/rolling-and-expanding-windows-for-dummies/
Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2020). The M4 Competition: 100,000 time series and 61 forecasting methods. *International Journal of Forecasting*, *36*(1), 54–74. https://doi.org/10.1016/j.ijforecast.2019.04.014
Maldonado, L., & Vera, L. (2011). Los determinantes de la demanda de crédito de los hogares: un modelo de vectores de corrección de errores para Venezuela. *Nueva Economía*, *19*(November), 13–45. Retrieved from https://www.researchgate.net/publication/317786081_Los_determinantes_de_la_demanda_de_credito_de_los_hogares_un_modelo_de_vectores_de_correccion_de_errores_para_Venezuela/citation/download
Medeiros, M. C., Vasconcelos, G. F. R., Veiga, Á., & Zilberman, E. (2021). Forecasting Inflation in a Data-Rich Environment: The Benefits of Machine Learning Methods. *Journal of Business and Economic Statistics*, *39*(1), 98–119. https://doi.org/10.1080/07350015.2019.1637745
Mele, M., & Magazzino, C. (2021). Pollution, economic growth, and COVID-19 deaths in India: a machine learning evidence. *Environmental Science and Pollution Research*, *28*(3), 2669–2677. https://doi.org/10.1007/s11356-020-10689-0
Mian, A., & Sufi, A. (2018). Finance and business cycles: The credit-driven household demand channel. *Journal of Economic Perspectives*, *32*(3), 31–58. https://doi.org/10.1257/jep.32.3.31
Mian, A., Sufi, A., & Verner, E. (2017). Household Debt and Business Cycles Worldwide*. *The Quarterly Journal of Economics*, *132*(4), 1755–1817. https://doi.org/10.1093/qje/qjx017
Molnar, C. (2022). *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable (2nd ed.)*. Retrieved from christophm.github.io/interpretable-ml-book/
Mullainathan, S., & Spiess, J. (2017). Machine learning: An applied econometric approach. *Journal of Economic Perspectives*, *31*(2), 87–106. https://doi.org/10.1257/jep.31.2.87
Pedragosa, F., Varoquaux, G., Gramfort, A., Thirion, B., Grisel, O., Blondel, M., … Duchesnay, É. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, *12*(9), 2825–2830. https://doi.org/10.1289/EHP4713
Ruder, S. (2017). An overview of gradient descent optimization algorithms, 1–14. Retrieved from http://arxiv.org/abs/1609.04747
Scikit-Learn. (2022a). *Choosing the right estimator*. Retrieved from https://scikit-learn.org/stable/tutorial/machine_learning_map/index.html
Scikit-Learn. (2022b). Partial Dependence and Individual Conditional Expectation Plots. Retrieved July 7, 2021, from https://scikit-learn.org/stable/auto_examples/inspection/plot_partial_dependence.html
Scikit-Learn. (2022c). Permutation Importance. Retrieved July 7, 2021, from https://scikit-learn.org/stable/modules/permutation_importance.html
Seabold, S., & Perktold, J. (2010). Statsmodels: Econometric and statistical modeling with python. *Python in Science Conference.*, *9*.
Soybilgen, B., & Yazgan, E. (2021). *Nowcasting US GDP Using Tree-Based Ensemble Models and Dynamic Factors*. *Computational Economics* (Vol. 57). https://doi.org/10.1007/s10614-020-10083-5
Stavinova, E., Timoshina, A., & Chunaev, P. (2021). Forecasting the volume of mortgage loans with open Internet data in the period of noticeable changes in the

Russian mortgage market. *Procedia Computer Science*, *193*, 266–275. https://doi.org/https://doi.org/10.1016/j.procs.2021.10.027
Synced. (2017). Tree Boosting With XGBoost – Why Does XGBoost Win "Every" Machine Learning Competition? Retrieved February 26, 2021, from https://syncedreview.com/2017/10/22/tree-boosting-with-xgboost-why-does-xgboost-win-every-machine-learning-competition/
Tibshirani, R. (1996). Regression Shrinkage and Selection Via the Lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, *58*(1), 267–288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x
University of Toronto. (2021). CSC 2541: Neural Net Training Dynamics - Lecture 2 - Taylor Approximations. Retrieved February 28, 2021, from https://www.cs.toronto.edu/~rgrosse/courses/csc2541_2021/slides/lec02.pdf
Varian, H. R. (2014). Big data: New tricks for econometrics. *Journal of Economic Perspectives*, *28*(2), 3–28. https://doi.org/10.1257/jep.28.2.3
Wackerly, D. D., Mendenhall, W., & Scheaffer, R. L. (2008). *Mathematical Statistics with Applications*. (C. Crockett, Ed.), *Mathematical Statistics with Applications* (7th ed.). Belmont: The Thomson Corporation. https://doi.org/10.1201/9781315275864
White, H. (1988). Economic prediction using neural networks: The case of IBM daily stock returns, 451–458. https://doi.org/10.1109/icnn.1988.23959
Wooldridge, J. M. (2016). *Introductory Econometrics: A Modern Approach, Fifth Edition* (Fifth). Mason: Cengage Learning.
Yang, L., & Shami, A. (2020). On hyperparameter optimization of machine learning algorithms: Theory and practice. *Neurocomputing*, *415*, 295–316. https://doi.org/10.1016/j.neucom.2020.07.061
Yilmaz, F. M., & Arabaci, O. (2021). Should Deep Learning Models be in High Demand, or Should They Simply be a Very Hot Topic? A Comprehensive Study for Exchange Rate Forecasting. *Computational Economics*, *57*(1), 217–245. https://doi.org/10.1007/s10614-020-10047-9
Yoon, J. (2021). Forecasting of Real GDP Growth Using Machine Learning Models: Gradient Boosting and Random Forest Approach. *Computational Economics*, *57*(1), 247–265. https://doi.org/10.1007/s10614-020-10054-w

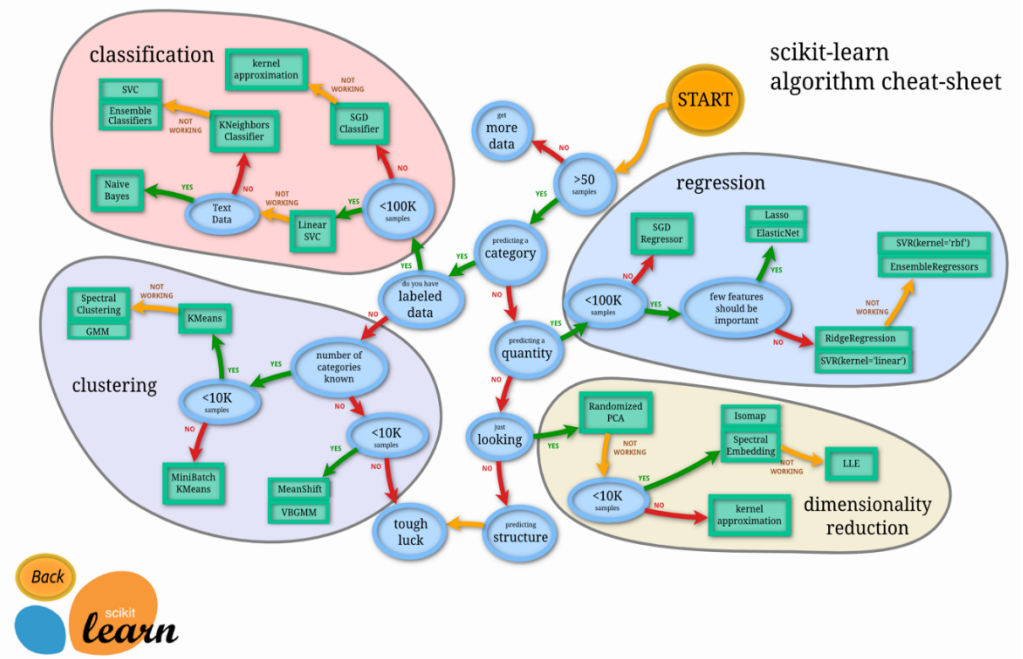
ANEXOS

Anexo A: Conceptos generales del ML
Anexo A.1: Aplicaciones y modelos comunes del ML:



Fuente: Cuadro obtenido de Cotton, (2022)

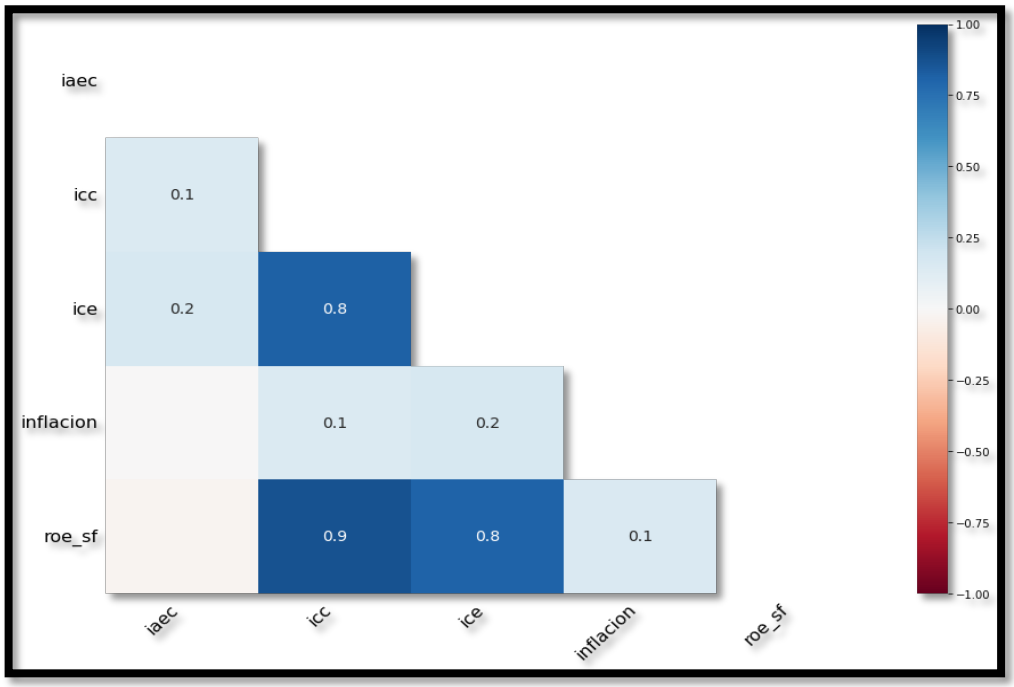
Anexo A.2: Diagrama de resumen de métodos y aplicaciones:



Fuente: Diagrama obtenido de (Scikit-Learn, 2022a)

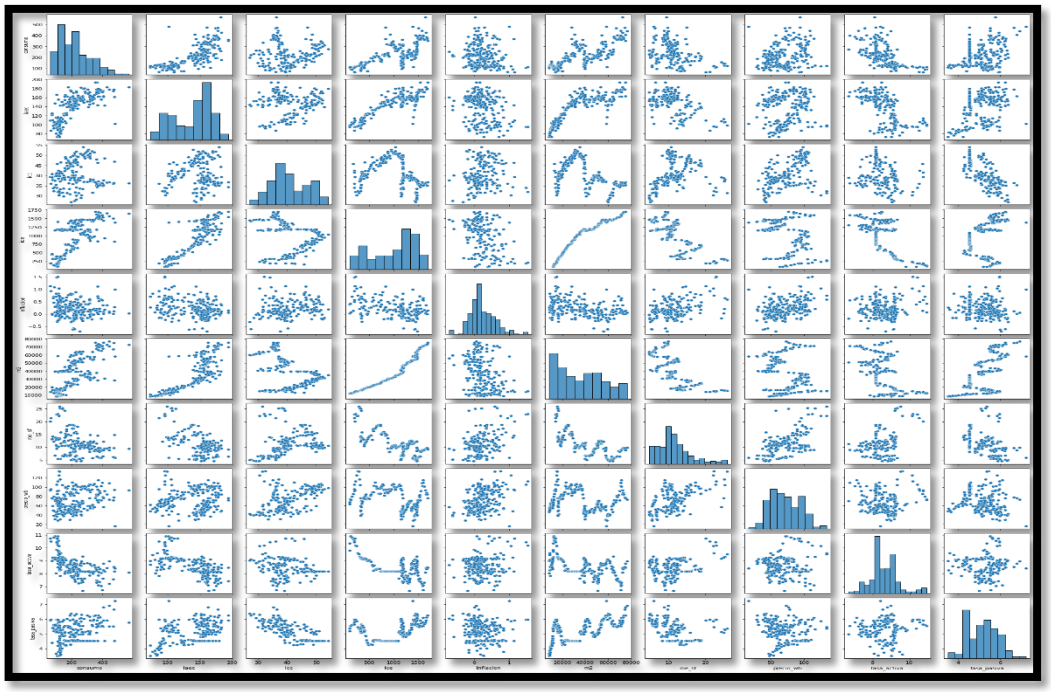
Anexo B: Exploración de datos

Anexo B.1: Gráfico correlación de Datos faltantes



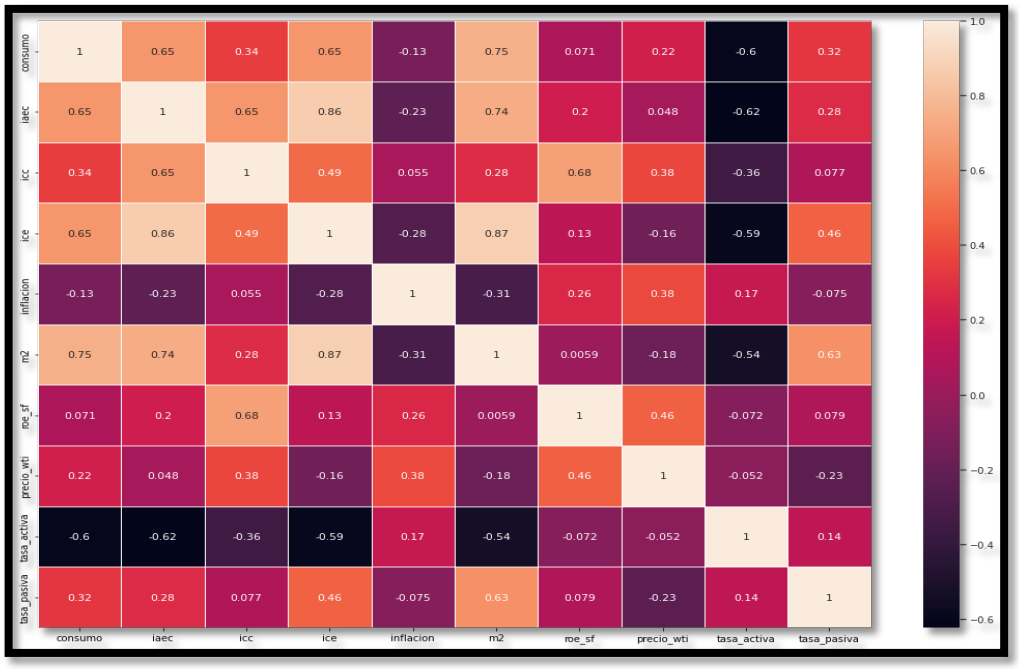
Fuente: Realizado por Autor

Anexo B.2: Gráfico de distribución y dispersión de los datos



Fuente: Realizado por el autor

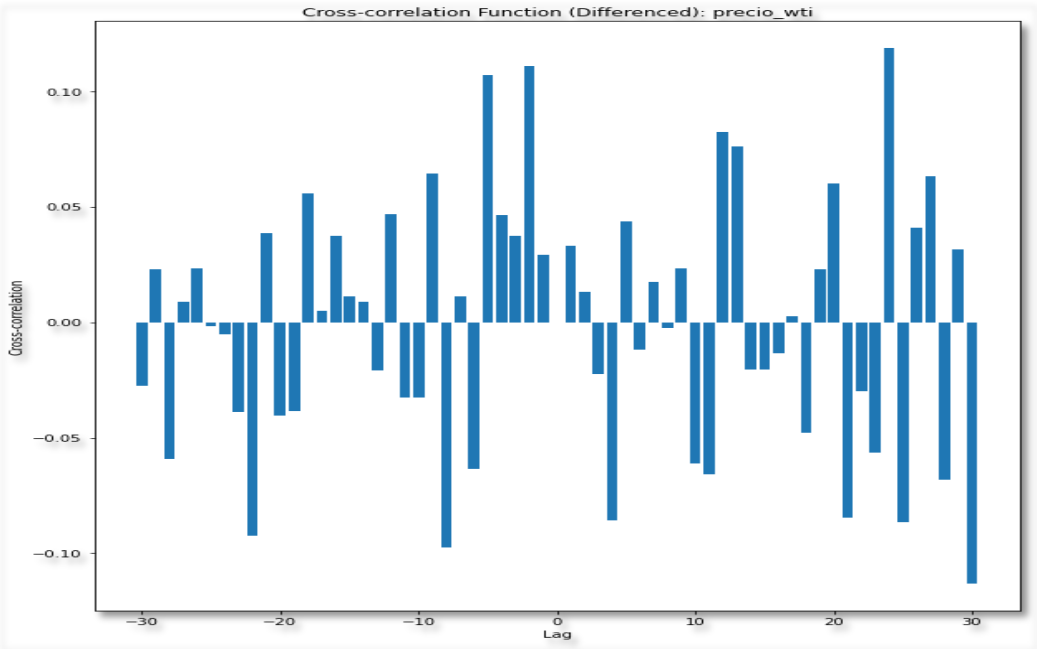
Anexo B.3: Correlación de Spearman de variables

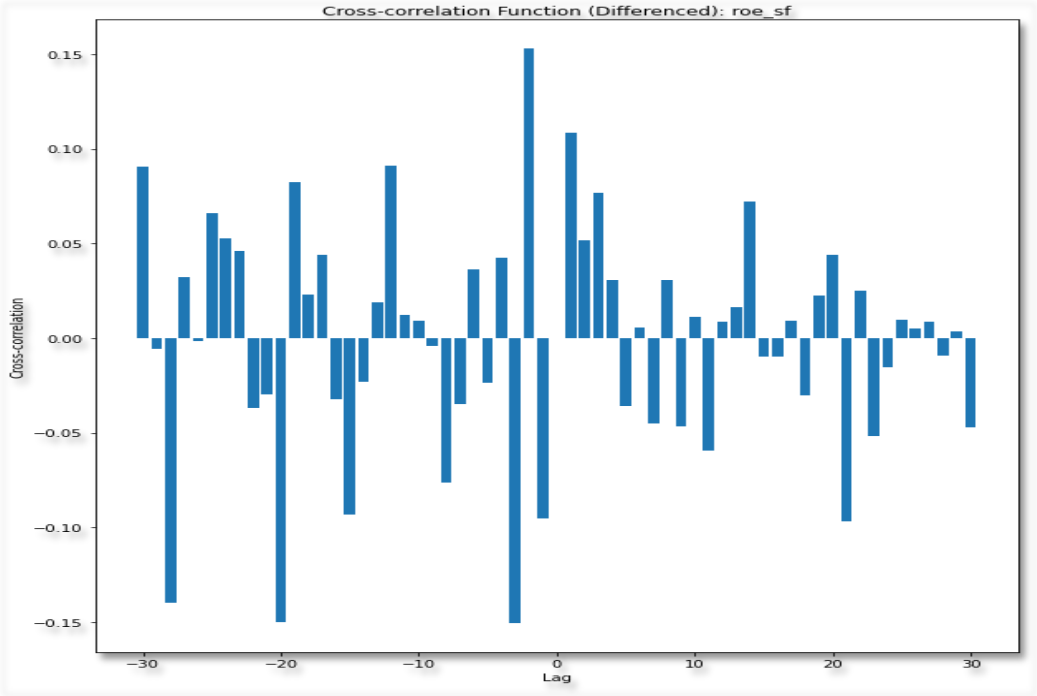


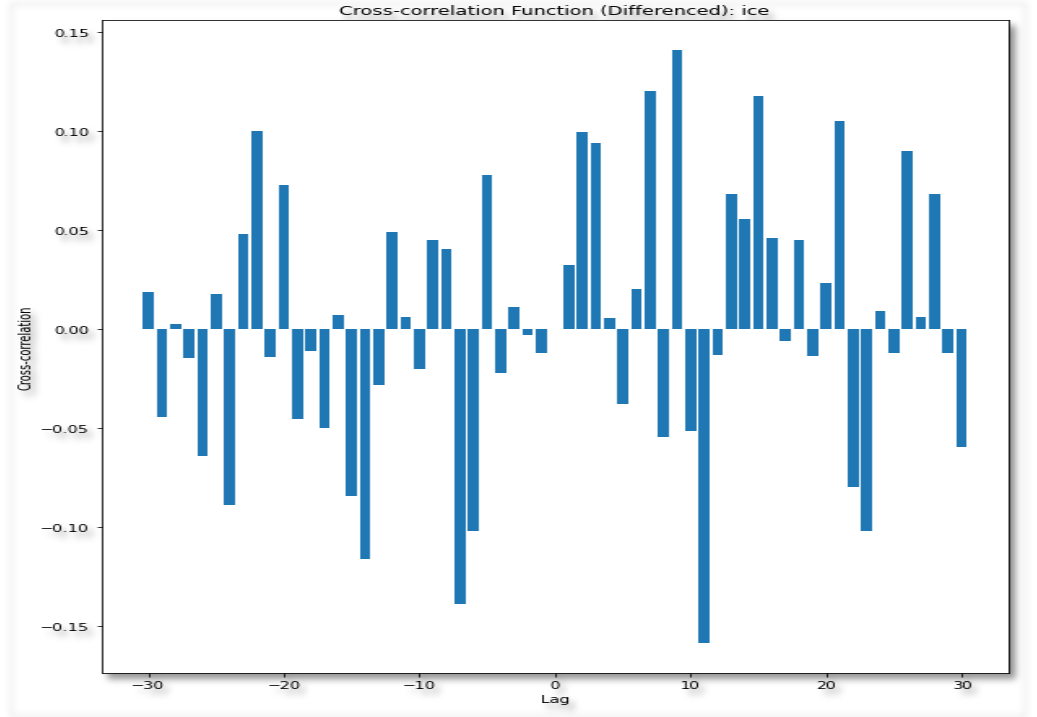
Fuente: Realizado por el autor

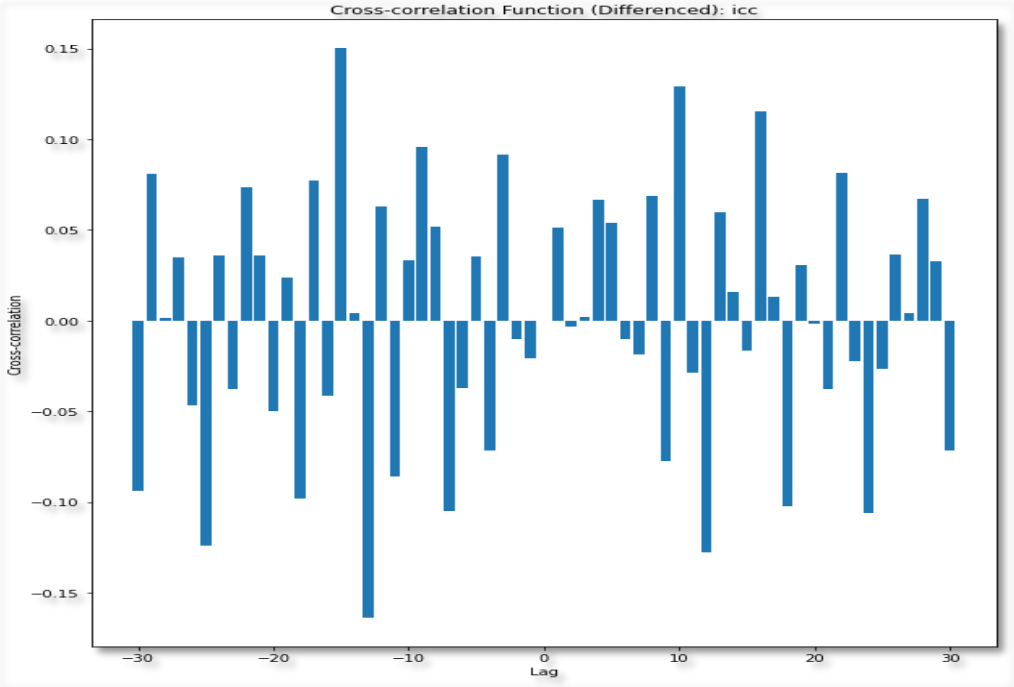
Anexo C: Preparación y transformación de datos

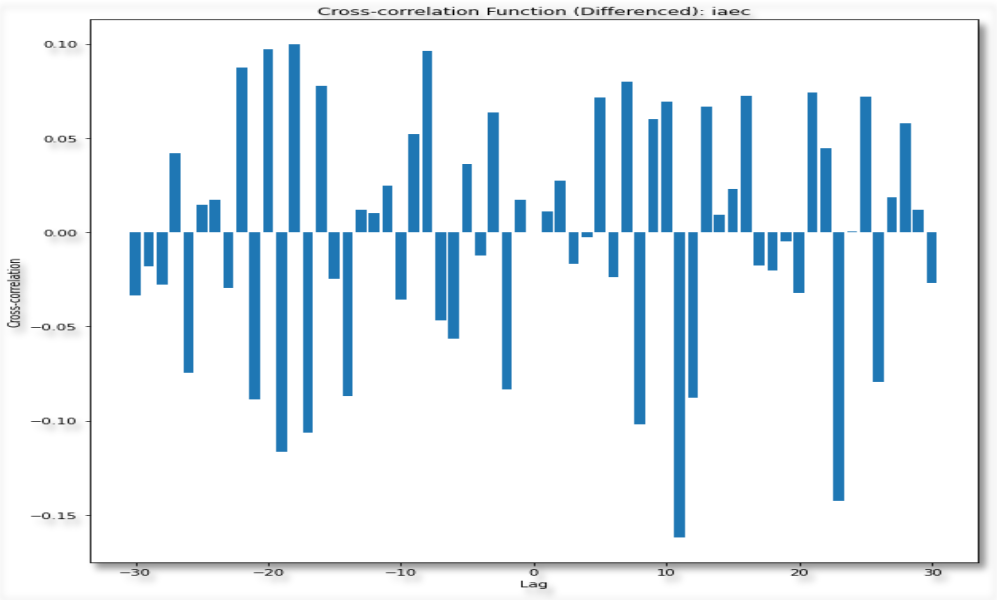
Anexo C.1: Gráficos de correlación cruzada entre predictores y variable objetivo.




Cross-correlation Function (Differenced): roe_sf
Cross-correlation
Lag
0.15
0.10
0.05
0.00
−0.05
−0.10
−0.15
−30
−20
−10
0
10
20
30


Cross-correlation Function (Differenced): ice
Cross-correlation
Lag
0.15
0.10
0.05
0.00
−0.05
−0.10
−0.15
−30
−20
−10
0
10
20
30





Fuente: Realizado por el autor

Anexo C.2: Tabla con el AIC de las diferentes combinaciones de los 8 mejores predictores para el modelo ARIMAX.

Duration	Name	AIC	BIC
4.1s	treasured-finch-928	1941.43	1963.79
4.1s	blushing-crab-103	1941.50	1960.66
3.8s	enchanting-slug-223	1941.66	1960.82
3.9s	flawless-fowl-770	1941.68	1960.84
3.9s	chill-frog-679	1941.91	1964.26
4.0s	honorable-ram-460	1941.93	1964.28
3.7s	carefree-swan-548	1941.95	1964.30
3.7s	righteous-loon-643	1942.00	1964.35
3.8s	polite-snake-709	1942.02	1961.17
3.9s	nosy-bird-429	1942.05	1961.21
3.5s	trusting-ox-251	1942.06	1961.21
3.7s	aged-lamb-501	1942.10	1961.26
4.2s	lyrical-turtle-458	1942.11	1964.46
4.0s	defiant-bug-842	1942.11	1961.27
3.8s	invincible-colt-246	1942.12	1961.27
4.0s	placid-trout-805	1942.13	1964.48
4.0s	burly-asp-834	1942.14	1961.30
4.0s	indecisive-loon-348	1942.15	1964.50
4.1s	fortunate-eel-559	1942.18	1961.34
3.9s	industrious-ray-47	1942.20	1964.55
3.8s	gentle-shrike-729	1942.23	1961.39
3.8s	classy-rat-602	1942.23	1961.39
4.1s	caring-wren-710	1942.27	1964.62
4.7s	rare-kite-666	1942.29	1964.64
3.8s	gifted-mare-820	1942.31	1964.66
4.2s	omniscient-sow-847	1942.46	1964.81
3.9s	powerful-stoat-849	1942.49	1964.84
4.1s	adorable-ray-314	1942.50	1964.86
5.2s	melodic-squid-71	1942.51	1964.86
3.6s	dapper-flea-749	1942.55	1964.90
3.9s	aged-kit-878	1942.60	1964.95
4.2s	glamorous-lynx-996	1942.60	1964.95
5.1s	wistful-shark-22	1942.63	1964.98
4.0s	upset-rook-289	1942.64	1964.99

4.8s	treasured-snake-865	1942.71	1965.06
4.2s	nimble-shrew-28	1942.74	1965.09
4.0s	persistent-rook-955	1942.75	1965.10
3.7s	glamorous-toad-375	1942.75	1965.10
4.1s	smiling-donkey-830	1942.82	1965.17
4.3s	handsome-rat-757	1942.86	1968.40
4.2s	adaptable-crab-69	1943.03	1965.38
4.0s	fortunate-conch-648	1943.04	1968.59
3.9s	sneaky-skink-608	1943.05	1965.40
4.1s	caring-fly-683	1943.09	1965.44
4.0s	crawling-lamb-946	1943.09	1965.44
4.0s	unleashed-ray-199	1943.13	1965.48
4.1s	caring-grouse-47	1943.16	1965.51
3.8s	resilient-dove-367	1943.19	1965.54
4.2s	sneaky-gnat-995	1943.21	1965.56
3.9s	monumental-colt-120	1943.21	1965.56
3.8s	bustling-panda-424	1943.25	1965.61
4.0s	bustling-shrimp-131	1943.26	1965.61
3.8s	funny-loon-783	1943.31	1965.66
4.2s	adventurous-dog-861	1943.31	1965.66

Fuente: Realizado por el autor,

Anexo D: Evaluación de los modelos

Anexo D.1: Tabla resumen modelo ARIMA (2,1,1)

SARIMAX Results

Dep. Variable:	consumo	No. Observations:	180
Model:	SARIMAX(1, 1, 1)	Log Likelihood	-958.619
Date:	Sun, 11 Jun 2023	AIC	1923.239
Time:	09:06:21	BIC	1932.801
Sample:	01-31-2008 - 12-31-2022	HQIC	1927.116
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	0.0805	0.109	0.738	0.460	-0.133	0.294
ma.L1	-0.5867	0.117	-5.010	0.000	-0.816	-0.357
sigma2	2620.4234	151.486	17.298	0.000	2323.516	2917.331

Ljung-Box (L1) (Q):	0.03	Jarque-Bera (JB):	546.17
Prob(Q):	0.87	Prob(JB):	0.00
Heteroskedasticity (H):	2.43	Skew:	0.77
Prob(H) (two-sided):	0.00	Kurtosis:	11.42

Fuente: Realizado por el autor

Anexo D.2: Tabla resumen modelo ARIMAX (2,1,1)

SARIMAX Results

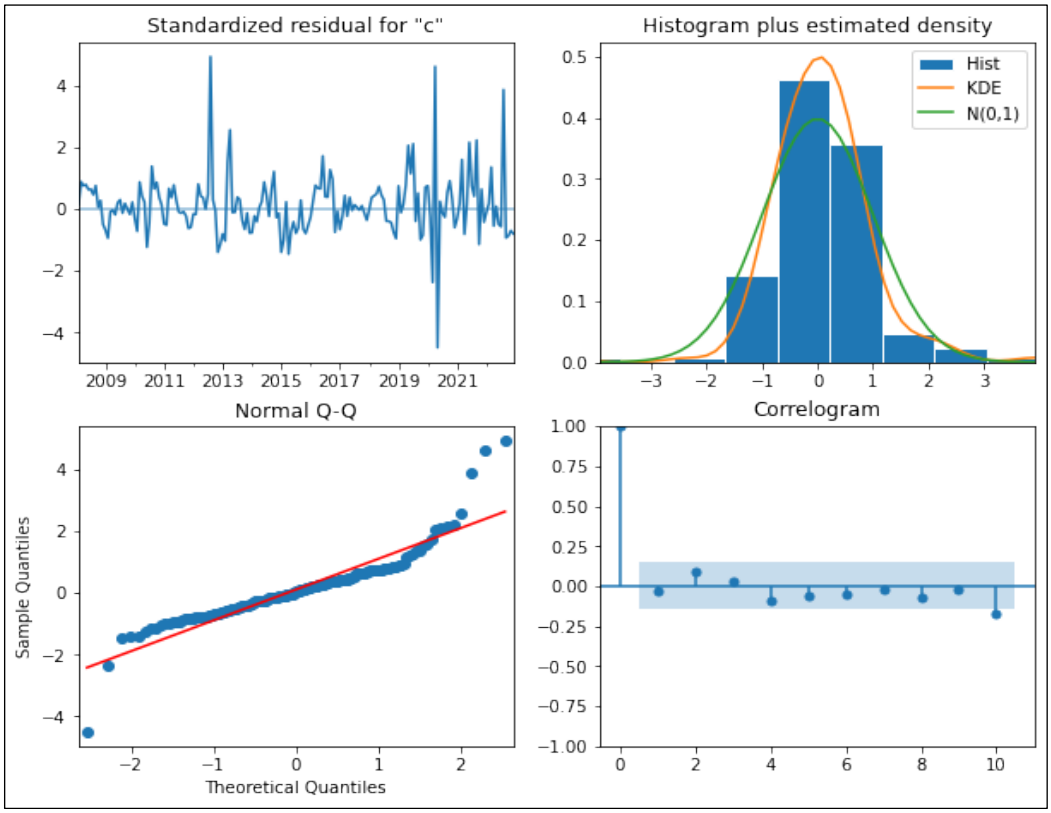
Dep. Variable:	consumo	No. Observations:	180
Model:	SARIMAX(1, 1, 1)	Log Likelihood	-955.358
Date:	Sun, 11 Jun 2023	AIC	1924.717
Time:	06:14:14	BIC	1947.028
Sample:	01-31-2008 - 12-31-2022	HQIC	1933.764
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
icc(t-3)	1.0959	2.482	0.442	0.659	-3.768	5.960
inflacion(t-3)	10.0038	12.465	0.803	0.422	-14.426	34.434
icc(t-2)	4.1234	1.552	2.656	0.008	1.081	7.166
inflacion(t-2)	3.7944	15.352	0.247	0.805	-26.296	33.885
ar.L1	0.2187	0.126	1.729	0.084	-0.029	0.467
ma.L1	-0.7165	0.106	-6.733	0.000	-0.925	-0.508
sigma2	2525.4056	160.125	15.771	0.000	2211.566	2839.245

Ljung-Box (L1) (Q):	0.23	Jarque-Bera (JB):	433.51

Fuente: Realizado por el autor

Anexo D.3: Gráfico de los residuos del modelo Arima (2,1,1)



Fuente: Realizado por el autor