UNIVERSIDAD NACIONAL DE ROSARIO
FACULTAD DE CIENCIAS ECONÓMICAS Y ESTADÍSTICA

CARRERA DE POSGRADO

MAESTRÍA EN ESTADÍSTICA APLICADA

Tema: Predicción del precio de acciones utilizando los estados contables mediante un modelo SARIMAX.

Autor: Almada, Agustín

Director: Mg. Del Rosso, Rodrigo

Fecha: 02/12/2024

Tribunal examinador

- Mg. Sergio Buzzi
- Dr. Martín E. Masci
- Mg. Fernanda Méndez

Resumen

La presente tesis aborda el desafío de predecir el precio de cierre de una acción el día siguiente a la publicación de sus estados contables, integrando la información contenida en dichos documentos y las estimaciones de los analistas de mercado. Para ello, se desarrollaron y ajustaron modelos SARIMAX que capturan la dinámica estacional y autorregresiva de la serie histórica de precios, al tiempo que incorporan variables exógenas para mejorar la precisión predictiva.

Se seleccionaron dos empresas representativas del mercado argentino (YPF y Grupo Financiero Galicia) sobre cuyos datos trimestrales se aplicó un exhaustivo proceso de ingeniería de características. Este incluyó el cálculo de diferencias entre los valores reales y esperados de los principales campos contables, la utilización de ratios financieras e indicadores macroeconónicos, la imputación de valores atípicos mediante rangos intercuartílicos y técnicas de *clustering*, la generación de componentes principales para reducir la dimensionalidad y la incorporación de variables de interacción.

Se aplicaron técnicas de selección de variables (*forward/backward selection*) y explicabilidad mediante SHAP para identificar los indicadores con mayor impacto en la estimación: flujo de caja, EBITDA y ventas. Además, se evaluó la exactitud de las predicciones de los analistas sobre los estados contables mediante la prueba de signos de Wilcoxon, aportando evidencia robusta sobre la precisión de sus estimaciones en muestras pequeñas.

Los resultados muestran un desempeño sobresaliente de los modelos, con errores cuadráticos medios reducidos y una clara identificación de las variables clave, lo que demuestra la eficacia de la metodología SARIMAX en combinación con análisis de características y pruebas no paramétricas para la predicción de precios en contextos de alta generación de información financiera.

Palabras Clave: Predicción de precios, modelos SARIMAX, series temporales, variables exógenas, estados contables, análisis financiero, mercado bursátil argentino.

Índice

1. Introducción

La empresa, institución predominante en el siglo XXI, constituye el eje fundamental de la actividad económica al ser la principal fuente de empleo, innovación y crecimiento. Como sistema abierto, está en constante interacción con la sociedad en la que opera: utiliza recursos que esta provee y, a su vez, genera trabajo para las familias. Esta dinámica la hace especialmente sensible a los cambios en el contexto económico, tanto a nivel local como internacional. Factores como la etapa del ciclo económico del país, los movimientos de la economía global, eventos políticos y la evolución de variables económicas clave -como la inflación, el tipo de cambio y la tasa de interés- tienen un impacto directo en su desempeño.

Cuando las empresas se hacen públicas mediante una oferta pública inicial (IPO), incorporan capital accionario a través del mercado bursátil, abriendo así su propiedad a inversores externos. Este proceso obliga a las empresas a cumplir con normativas de transparencia financiera, como la comunicación periódica de su información financiera a través de la presentación trimestral de estados contables. Según expresa Damodaran (2012), estos informes muestran la situación económica y financiera de la empresa en un momento específico, consistiendo en una representación estructurada de los activos, pasivos, patrimonio neto, flujo de efectivo y resultados, proporcionando así una imagen global de la solvencia, liquidez y estabilidad financiera de la empresa.

La publicación de los estados financieros brinda a accionistas, inversores y otras partes interesadas acceso a información objetiva y verificable sobre la salud financiera de la empresa. Estos informes detallan cómo se gestionan los recursos y si se generan beneficios, permitiendo evaluar la eficiencia y la eficacia de la administración. Además, proporcionan datos clave para analizar el desempeño financiero, lo cual influye directamente en la valoración y el precio de cotización de la acción.

El análisis financiero es un proceso que ayuda a determinar la salud financiera de una empresa y su capacidad para generar ganancias y crecimiento futuro, lo cual se logra a través de una revisión detallada de los estados financieros y el uso de ratios contables. Este análisis desempeña un papel crucial en la predicción de retornos futuros de acciones, un campo de investigación central debido a su impacto en la toma de decisiones estratégicas tanto a nivel corporativo como gubernamental. No solo los precios de las acciones pueden anticipar cambios económicos significativos, como sugiere Pearce (1983), sino que también la correcta interpretación de los estados contables y las ratios financieras puede

mejorar la precisión de estas predicciones, fortaleciendo la estabilidad financiera y fomentando un crecimiento económico sostenible (Fama, 1970). Estudios más recientes, como el de Pettenuzzo y Timmermann (2017), establecen una relación entre las recesiones, los ciclos económicos y el rendimiento del mercado.

Entender qué campos de los estados contables tienen un mayor impacto en la variación de los precios de las acciones es crucial para la gestión empresarial. Para los directivos de las empresas, identificar estas variables clave permite enfocar sus estrategias y recursos en mejorar los aspectos de la empresa que más influyen en la percepción del mercado, optimizando así el valor de las acciones y la confianza de los inversores. Este enfoque en las métricas financieras críticas no solo puede mejorar el rendimiento empresarial a corto plazo, sino que también puede contribuir a una estrategia de crecimiento sostenible y alineada con las expectativas del mercado (Kaplan y Norton, 1996). Además, al priorizar los indicadores financieros más relevantes, la dirección de la empresa puede tomar decisiones más informadas y eficaces, mejorando la transparencia y la comunicación con los accionistas (Black, Wright y Davies, 2001).

La predicción del precio de las acciones es una temática ampliamente abordada desde diversos enfoques, como modelos lineales, series de tiempo y, desde los últimos avances computacionales, técnicas de *machine learning*. Sin embargo, parece no haber un método que destaque sobre los demás, especialmente en empresas de Argentina. Particularmente, los modelos de series temporales ofrecen un marco teórico sólido para el análisis de datos, permitiendo a los investigadores utilizarlos como una herramienta robusta para analizar cómo las variaciones en los resultados financieros afectan el valor de mercado de una empresa. En particular, los modelos SARIMAX (*Seasonal AutoRegressive Integrated Moving Average with eXogenous factors*), una extensión de los modelos ARIMA (*AutoRegressive Integrated Moving Average*), son capaces de capturar no solo la estacionalidad de una serie temporal, sino también el impacto de variables exógenas, como indicadores económicos o métricas de rendimiento financiero, que influyen directamente en su comportamiento.

En el presente trabajo, se comparan los valores reales de los principales campos de los estados contables y sus ratios con las expectativas de los analistas de mercado, utilizando esta comparación para estimar el precio de la acción al día siguiente de la presentación de los estados financieros. Para este propósito, se aplican modelos SARIMAX, que permiten predecir no solo la evolución histórica del precio de la acción, sino también incorporar factores externos relevantes. La inclusión de ratios

financieras es esencial, ya que estas normalizan la información contable, facilitando comparaciones entre empresas de diferentes tamaños o sectores y proporcionando una evaluación temporal más precisa (Penman, 2010). Además, dado que la percepción de los inversores sobre el valor futuro de una acción influye significativamente en los mercados (Derakhshan, 2019), se incorporan como variables las estimaciones de los principales analistas de inversión, enriqueciendo así el modelo predictivo.

En este contexto, surgen varios interrogantes fundamentales que guiarán el desarrollo de la investigación: ¿Cuáles son los campos de los estados contables o ratios que tienen mayor impacto en la estimación del precio de las acciones? ¿Los campos de los estados contables influyen de la misma manera en la estimación del precio en las dos empresas seleccionadas? ¿Qué tan precisos son los analistas al predecir los valores de los campos de los estados contables?

Esta tesis introduce un enfoque novedoso al integrar variables que hasta ahora no han sido combinadas en un modelo predictivo de precios de acciones. A diferencia de estudios anteriores, que se abordarán en detalle en el análisis del estado del arte, este trabajo propone ajustar un modelo SARIMAX que no solo incorpora la comparación entre los valores reales y estimados de los principales campos de los estados contables, sino que también integra factores macroeconómicos claves como la tasa de interés de los bonos a 10 años del gobierno de Estados Unidos y el valor del principal índice accionario del mundo.

El análisis se centrará en empresas argentinas que cotizan en Estados Unidos, abarcando el período comprendido entre 2014 y 2024. La atención se dirigirá exclusivamente a aquellas empresas con mayor volumen de operaciones diarias, dado que sus estimaciones son las más sólidas y fiables, ya que son objeto de un seguimiento exhaustivo por parte de un mayor número de analistas especializados.

Esta tesis está organizada en 9 secciones. En el marco teórico se definirán conceptos que se utilizarán a lo largo de todo el trabajo. Se abordarán allí las definiciones y características del mercado de capitales de Argentina y Estados Unidos, la hipótesis del mercado eficiente y como se conforma la predicción del precio de una acción mediante la utilización de los estados contables, método conocido como análisis fundamental. Además, se efectúa una descripción precisa sobre el estado de arte en cuanto a metodologías para predecir el precio de una acción, abordando las principales investigaciones sobre estados financieros y series de tiempo.

En la sección 3 se describen la hipótesis y objetivos del trabajo, el cual se centra en

estimar a través de un modelo SARIMAX cuál va a ser el precio de una acción el día de la presentación de los estados contables, utilizando la información proporcionada en el mismo documento y las estimaciones de analistas del mercado.

En la Sección 4, se presenta la metodología empleada para abordar el estudio. Se define la población y la muestra utilizada para cada uno de los modelos predictivos, describiendo el origen de los datos y los criterios de selección. Además, se detalla el proceso de recolección de información financiera y bursátil, junto con las variables de interés, tanto las correspondientes a los estados contables como las relacionadas con precios y tasas de interés.

En la sección 5 se analizan los modelos aplicados a las empresas Galicia e YPF. Allí se realiza un análisis descriptivo de las variables, se verifica la estacionariedad de la variable objetivo y se ajustan distintos modelos SARIMA y SARIMAX, evaluando componentes principales y posibles interacciones entre variables. Finalmente, se analiza el ajuste del modelo propuesto y sus residuos.

En la Sección 6, se presentan y discuten los resultados obtenidos tras el ajuste de los modelos. Se evalúa la importancia de las variables seleccionadas y se comparan los desempeños de los modelos en términos de métricas de error. Además, se visualizan los resultados en gráficos que permiten interpretar los patrones observados.

La Sección 7 se dedica a las conclusiones, donde se sintetizan los hallazgos más relevantes del estudio. Se revisan las implicancias de los resultados tanto para la predicción del precio de acciones como para el análisis de los estados contables. Asimismo, se discuten posibles limitaciones del enfoque utilizado y se sugieren líneas futuras de investigación para mejorar el rendimiento predictivo de los modelos.

En la sección 8 se listan todas las referencias bibliográficas utilizadas para la elaboración de la tesis, incluyendo citas de *papers*, libros y links a páginas de internet y en la sección 9 se incluyen los anexos de la tesis.

2. Marco Teórico

El mercado de capitales, al proporcionar un espacio para la transferencia, negociación y distribución de acciones, ha sido ampliamente reconocido como un medio crucial para que las grandes empresas obtengan financiamiento de los accionistas. La emisión de acciones conlleva un flujo significativo de capital hacia dicho mercado, lo que mejora la configuración orgánica del capital corporativo al promover la concentración de capital (Jin, 2019). Asimismo, los mercados, al facilitar las operaciones a bajos costos de transacción, desempeñan un papel vital en la economía al dirigir de manera eficiente el flujo de ahorros e inversión, lo que fomenta la acumulación de capital y estimula la producción de bienes y servicios, brindando beneficios tanto a los prestatarios como a los inversores y contribuyendo al óptimo funcionamiento económico.

En un entorno caracterizado por la presencia de ruido y volatilidad inherente a los mercados, la tarea de predecir el precio de las acciones resulta un desafío al menos complejo tanto para inversores como para accionistas. La multitud de variables ocultas detrás de las fluctuaciones, los riesgos de burbujas económicas y los inevitables altibajos, añaden una dosis de incertidumbre a esta compleja labor.

El proceso de predicción de los cambios en el precio de las acciones generalmente se considera un proceso difícil y complejo. Requiere la mezcla de varios factores y el comportamiento especial de factores individuales, incluidos factores políticos, económicos y de mercado, así como tecnología y comportamiento del inversor (Jin, 2019).

2.1 Hipótesis del Mercado Eficiente

La Hipótesis del Mercado Eficiente (EMH), desarrollada por Fama (1965), sugiere que es imposible obtener consistentemente rendimientos superiores a los del mercado de acciones en su conjunto, ya que este es un mercado eficiente donde los precios reflejan toda la información disponible y relevante. Cada vez que surge nueva información, esta se difunde inmediatamente a través de las noticias y el precio de la acción se ajusta al instante. Por lo tanto, las acciones siempre se negocian a su valor justo, y no es posible hacer predicciones sobre las tendencias del mercado o identificar acciones infravaloradas.

El supuesto fundamental de esta teoría es que asume que el mercado de acciones es eficiente. Es decir:

- Todos los inversores deben tener acceso a sistemas avanzados y de alta

velocidad para el análisis de precios de acciones.

- Debe existir un método de análisis de precios de acciones universalmente aceptado como correcto.
- Todos los inversores en el mercado deben ser tomadores de decisiones racionales, cuyas decisiones no estén influenciadas por sus emociones.

Las implicancias de EMH son que el mercado reacciona instantáneamente a las noticias y que nadie puede superar al rendimiento del mercado de manera sistemática. Sin embargo, el grado de eficiencia del mercado es controvertido debido al no cumplimiento de las condiciones mencionadas (Yen, 2008).

Bajo el marco de la hipótesis del mercado eficiente, se asume que el precio de la acción refleja toda la información disponible hasta ese momento. En este contexto, la publicación de los estados contables introduce información actualizada y detallada que impacta en las expectativas del mercado. El nuevo precio de la acción, por lo tanto, se ajustará en función de la oferta y la demanda impulsadas por la interpretación que hacen los inversores de esta información financiera reciente.

Existen numerosos métodos desarrollados que intentan predecir los precios de las acciones. Sin embargo, aunque se han llevado a cabo un gran número de investigaciones relacionadas al pronóstico de precios (Lin et al, 2009; Adebiyi et al, 2016; Hu et al, 2021; Gupta et al, 2022; Payal Soni *et al,* 2022; Cakra et al, 2023), aún existen muchos desafíos por resolver. Hasta la fecha, no se ha encontrado ningún modelo completo, preciso y exhaustivo para predecir el rendimiento del mercado de capitales (Agrawal, 2013).

2.2 Análisis fundamental

Ball y Brown (1968) indican que el análisis fundamental se basa en la suposición de que existe una brecha entre el valor de mercado y el valor intrínseco, y que el precio de la acción converge hacia el valor intrínseco. El análisis fundamental es un método a través del cual se analizan los precios de las acciones de una empresa mediante datos históricos contables y financieros.

El análisis fundamental considera tanto factores macroeconómicos, como el estado general de la economía, y factores microeconómicos, como los ingresos o ganancias de una empresa; para determinar el potencial de crecimiento futuro de una empresa e intentar estimar el valor real de sus acciones, si están sobrevaloradas o subvaloradas y obtener rendimientos a largo plazo.

El objetivo principal de la presentación de informes financieros es proporcionar información sobre la posición financiera y el rendimiento de las empresas. Los inversores utilizan la información como por ejemplo ganancias y gastos de la empresa, activos, pasivos, experiencia de la dirección, beneficios y dinámica de la industria; para predecir los rendimientos futuros de las acciones (Cantemir, 2013). De esta manera, se utilizan estadísticas de informes financieros para estudiar los valores intrínsecos de las empresas (Graham, 2004).

Si bien hay autores que sostienen que la información de los estados contables es difícil de formalizar y estandarizar; y la interpretación de ese conocimiento puede ser subjetiva (Agrawal, 2013); famosas literaturas financieras, como Basu (1983) y Fama y French (1988, 1992, 2017), han sugerido que los estados financieros y contables son herramientas muy importantes para analizar el rendimiento inminente del mercado de valores. Apoyando esta postura, Arkan (2016) señala que los usuarios de la información contable, con el fin de evaluar y predecir la rentabilidad, el crecimiento del patrimonio, el flujo de efectivo y los dividendos de las empresas, utilizan estos documentos como una guía para sus decisiones.

Para realizar el análisis fundamental de una empresa o sector público, los inversores y analistas suelen analizar las métricas en los estados financieros de una empresa. Los cuales incluyen el balance general, el estado de resultados, el estado de flujos de efectivo y el estado de situación patrimonial. Estos contienen información útil e ideas sobre la empresa, que pueden explotarse con el uso de ratios financieras.

2.3 Mercado bursátil argentino

El Sistema Bursátil Argentino es un conjunto de instituciones que proveen el marco operativo para la realización de las diferentes operaciones bursátiles. Está compuesto por un organismo de control (Comisión Nacional de Valores), un Agente Depositario Central de Valores Negociables (Caja de Valores S.A.), un Mercado que liquida y compensa las operaciones de compra y venta de valores negociables (BYMA) y los Agentes a través de los cuales estas operaciones son concertadas (ALyC).

La Bolsa de Comercio de Buenos Aires fue fundada en el año 1854 donde se efectuaban transacciones en onzas de oro, y en donde hoy cotizan las empresas más importantes del país. Con la premisa de brindar un marco de transparencia y eficiencia para la unión de oferta y demanda, cotidianamente la Bolsa favorece el flujo de capitales. Es una entidad que se autorregula y su organismo de control es

la Comisión Nacional de Valores (CNV).

La Comisión Nacional de Valores es un organismo autárquico con jurisdicción en toda la República Argentina, que funciona bajo la órbita del Ministerio de Economía y es el encargado de la promoción, supervisión y control del mercado de capitales. Está orientada a proteger a los inversores y crear un marco normativo capaz de contribuir al fomento del desarrollo de un mercado de capitales federal, transparente, inclusivo y sustentable que propenda a canalizar el ahorro hacia la inversión, contribuyendo al fortalecimiento de las diversas actividades económicas del país.

Los mercados son sociedades anónimas autorizadas por la CNV con el objeto principal de organizar las operaciones con valores negociables que cuenten con oferta pública. Es el ámbito donde se ofrecen públicamente valores negociables con el objetivo de canalizar el ahorro hacia la inversión productiva.

En BYMA, principal mercado de Argentina, cotizan empresas de diversos rubros, entre los que se destacan el sector Financiero, Energético, Petrolero, Campo, Materias Primas, Consumo Básico, Consumo Discrecional, Telecomunicaciones, Construcción y Telecomunicaciones.

La Comisión Nacional de Valores, en su carta orgánica reglamentada por las leyes 27.440 y 26.831, establece que los estados contables constituyen un tipo de informe a través del cual la sociedad que lo emite da a conocer públicamente su situación patrimonial, económica y financiera durante un determinado período de tiempo; resultado de vital importancia para accionistas, acreedores, propietarios, miembros de los órganos de administración y fiscalización e inversores en general. Todos los sujetos autorizados, controlados, regulados y fiscalizados por la CNV deben presentar sus estados contables con periodicidad trimestral y anual. Los documentos que componen los estados contables son:

Balance General: Es un estado financiero que informa la situación patrimonial sobre los activos, pasivos y el patrimonio neto de una empresa en una fecha específica. A menudo se describe como una foto de la condición financiera de una empresa, que incluye lo que la empresa posee y debe, así como la cantidad invertida por los accionistas.

Estado de Resultados: Mientras que el balance general proporciona una visión general financiera de una empresa en un punto específico del tiempo, el estado de resultados informa los ingresos durante un período de tiempo específico, generalmente a lo largo de un año o un trimestre del año. Proporciona los estados financieros centrales de una empresa, mostrando sus ganancias y gastos, al

considerar todos los ingresos, gastos, ganancias, pérdidas e ingresos netos durante el período especificado.

Estado de Flujos de Efectivo o el estado de flujo de caja: Es un estado financiero que contiene datos agregados sobre todos los flujos de efectivo entrantes y salientes de una empresa. Estos incluyen el efectivo recibido de sus actividades en curso y fuentes externas de inversión, así como el efectivo pagado por las operaciones comerciales y la financiación de nuevas inversiones. Ayuda a los inversores a comprender cómo una empresa administra su posición de efectivo, es decir, de dónde proviene su dinero y cuán bien genera efectivo para pagar sus deudas y gastos operativos.

Al igual que BYMA en Argentina, el *New York Stock Exchange* (NYSE) ofrece un marco transparente y eficiente para la negociación de valores en Estados Unidos. En él cotizan tanto empresas locales como aquellas de otros países, a través de instrumentos conocidos como *American Depositary Receipts* (ADRs). Los ADRs son certificados emitidos por bancos estadounidenses que representan la propiedad de acciones de una empresa extranjera, permitiendo a inversores en EE. UU. adquirir participaciones en compañías internacionales sin necesidad de que estas estén listadas directamente en el NYSE. Dado que en esta tesis se utilizan precios de ADRs de YPF y Galicia, es útil comprender cómo estos certificados facilitan el acceso de empresas argentinas al mercado estadounidense: por un lado, brindan a los inversores locales una forma sencilla de diversificar sus portafolios con acciones internacionales; y, por otro, otorgan a los emisores extranjeros una mayor visibilidad y liquidez en un mercado de gran profundidad, sin verse obligados a cumplir con todos los requisitos regulatorios y operativos que implica una cotización directa en Estados Unidos.

2.4 Estado del arte

2.4.1 Predicción del precio de las acciones mediante estados financieros

En cuanto a predicción del precio de una acción mediante los estados financieros, Ou y Penman fueron pioneros. Encontraron que, a través del análisis fundamental, los inversores pueden obtener rendimientos superiores al retorno promedio del mercado.

Afirman que el análisis de los estados financieros publicados puede descubrir valores que no se reflejan en los precios de las acciones. En lugar de tomar los precios como puntos de referencia de valor, los 'valores intrínsecos' descubiertos a

partir de los estados financieros sirven como puntos de referencia con los cuales se comparan los precios para identificar acciones sobrevaloradas y subvaloradas (Ou y Penman, 1989). Su labor consistió en analizar los estados financieros combinando un gran conjunto de sus campos, y a partir de ellos elaborar una medida indicador de la dirección de las ganancias futuras. Este indicador captura valores de capital que no se reflejan en los precios de las acciones.

Siguiendo en la línea de Ou y Penman, Holthausen investigó la capacidad de modelos puramente estadísticos. Basado exclusivamente en información contable de costos históricos y utilizando las ratios financieras como variables independientes, predijo rendimientos excesivos en un período posterior al estudiado. Holthausen (1992) arribó a la misma conclusión que Ou y Penman, los elementos de los estados financieros se pueden combinar en una medida resumida para obtener información sobre el movimiento posterior de los precios de las acciones.

Comenzando desde los mismos objetivos, Lev y Thiagarajan (1993) utilizaron 12 factores fundamentales que fueron considerados útiles para la evaluación de valores por los investigadores y encontraron que las señales financieras tenían poder predictivo. Abarbanell y Bushee (1998) también trabajaron en el mismo fenómeno y encontraron los mismos resultados.

Clubb (2007) presenta evidencia sólida de que un modelo lineal simple que combina la relación valor-libro con el retorno sobre el patrimonio neto futuro explica una parte significativa de la variación cruzada en los rendimientos futuros de las acciones. Este artículo presenta un modelo log-lineal que incluye el valor libro actual (BM), las expectativas del BM, y el rendimiento sobre el patrimonio neto (ROE) futuro como variables explicativas de los rendimientos futuros de las acciones. Demuestran que estas tres variables explican una parte significativa de los rendimientos cruzados de las acciones en el Reino Unido y que siguen siendo estadísticamente significativas incluso después de incluir variables adicionales que representan el riesgo. Esto respalda la relevancia del análisis fundamental para explicar los rendimientos de las acciones e indica su utilidad potencial para predecir los rendimientos futuros.

Imran (2008) utilizó regresión logística para predecir cómo será la tendencia o performance de una acción trabajando con información de los estados financieros. La investigación examinó el crecimiento de las ventas, la relación deuda sobre patrimonio, la relación precio – valor libro, las ganancias por acción, el retorno sobre el patrimonio y la ratio corriente para la predicción del rendimiento de las acciones.

El estudio de Arkan (2016) sobre la correlación entre la información contable y el precio de las acciones también indica la importancia del análisis financiero y el uso de ratios para predecir y evaluar el rendimiento de las empresas y sus acciones en el mercado. El estudio analiza varias ratios financieras y busca determinar su correlación con las tendencias de los precios de las acciones utilizando un análisis de regresión múltiple para encontrar una ecuación que explique la relación entre las ratios financieras (variables dependientes) y el precio de las acciones (variable independiente). El estudio concluyó que se puede confiar en un conjunto de ratios financieras para cada sector para predecir el precio de las acciones, pues estas ratios pueden tener una alta correlación con la rentabilidad y la predictibilidad de las acciones, lo que las convierte en herramientas valiosas para los inversores y tomadores de decisiones financieras.

Boozer (2017) trabajó con un modelo de regresión lineal múltiple en el cual el precio de las acciones de la empresa es la variable dependiente y las variables independientes son efectivo de actividades operativas, efectivo de actividades financieras, ventas netas, poder de ganancia básico, pasivos corrientes totales, y capital de trabajo neto. El modelo concluye que las variables de los estados financieros analizadas tienen capacidad predictiva.

2.4.2 Predicción del precio de las acciones mediante series de tiempo

Otro enfoque ampliamente utilizado para la predicción del precio de las acciones es el análisis de series temporales, particularmente los modelos ARIMA, los cuales son conocidos por ser robustos y eficientes en la predicción de series financieras.

Los procedimientos de modelado tradicionales, introducidos por Box y Jenkins (Box, 1970) en la década de 1970, combinan la Auto-Regresión lineal (AR) y el Promedio Móvil (MA), como el popular modelo ARIMA (Box, 1994). El mismo es un referente en la literatura de pronóstico de series temporales.

Pese a que han sido extensamente utilizados en el campo de finanzas y economía, el método presenta algunas limitaciones, entre ellas:

- Los modelos ARIMA generalmente asumen que los residuos del modelo tienen un valor promedio cero y la misma varianza, siendo que en realidad las series temporales de precios de acciones tienen varianzas que varían a lo largo del tiempo (Zhang, 2009).
- Para considerar todos los rezagos significativos es necesario formular y experimentar con un número muy grande de modelos, lo que es un proceso lento y engorroso (Kumar, 2021).

- Predicen los precios futuros teniendo en cuenta únicamente los precios pasados, sin considerar todos los factores complejos que afectan a la empresa. (Zhang, 2009).

Como paliativo a estas limitaciones, Jarret (2011) ha planteado un modelo ARIMA con intervenciones, este agregado le resultó útil para explicar la dinámica del impacto de interrupciones graves en una economía y los cambios en la serie de tiempo de un índice de precios de manera precisa y detallada.

El modelo ARIMAX (ARIMA con variables exógenas) añade variables exógenas con un rezago de cada una. La literatura ofrece diferentes conclusiones sobre su efectividad (Peter y Silvia, 2012; Kongcharoen y Kruangpradit, 2013), encontrando en ambos casos que su precisión predictiva es menor o mayor en comparación con la de un modelo ARIMA estándar.

Green (2011), utilizó el método de Box-Jenkins para ajustar modelos ARIMA a los precios de cierre de las acciones AAPL, MSFT, COKE, KR, WINN, ASML, AATI y PEP. Tras el análisis, casi todos sus modelos fueron modelos AR(1) en forma diferenciada o no diferenciada. Se encontró que las acciones de la misma industria no se comportaban de manera similar.

Otro estudio, realizado por Yu (2012), intentó combinar técnicas tradicionales de análisis de series temporales con información del sitio web de tendencias de Google y del sitio web de Yahoo Finance para predecir cambios semanales en los precios de las acciones. Recolectaron noticias importantes relacionadas con una acción particular durante un período de cinco años y utilizaron los valores del índice de tendencias de Google de esta acción para medir la magnitud de estos eventos. Encontraron una correlación significativa entre los valores de las noticias/eventos importantes y los precios semanales de las acciones. Para analizar los precios históricos de acciones, realizaron un análisis de series temporales ARIMA tras una diferenciación de primer grado de la raíz cuadrada de los datos sin procesar. Se encontró, al graficar la función de autocorrelación y la función de autocorrelación parcial, que los precios de las acciones transformados seguían esencialmente un proceso ARIMA(0,1,0).

Por su parte, Adebiyi, Adewumi y Ayo (2014) utilizaron datos de precios de cierre del Índice de Acciones de Nokia y del Índice de Acciones del Banco Zenith para construir modelos ARIMA separados para las dos empresas. Encontraron que sus modelos construidos proporcionaban satisfactoriamente predicciones a corto plazo.

Mondal (2014) condujo un estudio sobre la efectividad de los modelos ARIMA para predecir precios futuros de cincuenta y seis acciones de siete sectores de la India.

Todos sus modelos construidos fueron capaces de predecir los precios de las acciones con una precisión superior al 85%.

Yetginer (2017) se propuso pronosticar el Índice de Precios BIST-100 utilizando sus determinantes macroeconómicos y financieros más significativos. El algoritmo, que está construido en forma de modelos ARIMAX lineales, explota cada posible combinación de variables explicativas para capturar el comportamiento del índice durante el período de tiempo de 2002 a 2013 utilizando datos mensuales. Las variables consideradas fueron:

Indicadores Macroeconómicos: Índice de Precios al Consumidor, Índice de Producción Industrial (ajustado estacionalmente), Tipo de Cambio USD/TRY, Oferta Monetaria M1 y M2 y Balanza Comercial.

Indicadores Financieros: Tasa de Interés de Depósitos; Tasa de Interés de Bonos de EE. UU a 1 y 10 Años, Índice de Precios Industriales Dow Jones, Bovespa y DAX, Precio del Petróleo Brent y de la Onza de Oro; Interés del Bono de Turquía a 2 Años.

3. Hipótesis y Objetivos

3.1 Hipótesis

El precio de cierre de la acción al día siguiente de la presentación de los estados contables puede estimarse con precisión mediante un modelo SARIMAX que incorpore las diferencias entre los valores reales y las estimaciones de los analistas para los principales campos contables y ratios financieras derivados de dichos campos, y variables macroeconómicas exógenas.

3.2 Objetivo General

Predecir cuál va a ser el precio de una acción el día posterior al de la presentación de los estados contables, utilizando la información proporcionada en el mismo documento y las estimaciones de analistas del mercado.

3.3 Objetivos Específicos

- Determinar las variables explicativas que aportan mayor relevancia al ajuste del modelo predictivo, evaluando su impacto individual en la estimación de la variable objetivo. Asimismo, se compararán los conjuntos de variables seleccionadas para las dos empresas de la muestra (YPF y Galicia) con el fin de analizar la consistencia de los indicadores financieros más influyentes en distintos contextos empresariales.
- Analizar la contribución de las variables explicativas derivadas de los estados contables en la predicción de la dirección del precio de la acción tras la presentación de los estados contables.
- Evaluar la precisión de las estimaciones realizadas por los analistas de mercado sobre los principales campos de los estados contables empleando pruebas no paramétricas.

4. Metodología

4.1 Población y muestra

En este estudio se seleccionarán los ADRs de origen argentino con mayor volumen operado en el mercado estadounidense, de entre los 13 ADRs disponibles. La elección de estos instrumentos se fundamenta en el hecho de que las variables explicativas del modelo se construyen a partir de las estimaciones realizadas por analistas financieros. Tal como afirman Bhushan (1989) y Brennan y Subrahmanyam (1995), las empresas con mayor volumen de operaciones suelen ser objeto de análisis más exhaustivos y profundos por parte de los analistas, lo que proporciona estimaciones más robustas y confiables para la modelización. Esta relación entre el volumen de operaciones y la profundidad del análisis ha sido documentada en la literatura económica, donde se destaca que los analistas tienden a concentrarse en empresas con mayor liquidez y visibilidad en el mercado, ya que estas presentan mayor interés para los inversores.

En la tabla 4.1 se listan las empresas que se utilizarán.

Empresa	Ticker
Yacimientos Petrolíferos Fiscales, S. A.	YPF
Grupo Financiero Galicia	GGAL

Tabla 4.1

4.1.1 YPF

YPF S.A. (Yacimientos Petrolíferos Fiscales) es la empresa de energía más grande de Argentina, dedicada a la exploración, explotación y producción de petróleo, gas y energías renovables como eólica y solar. Según su sitio institucional, su composición es mixta, con un 51% de las acciones en manos del Estado argentino y el resto cotizando en bolsa. Con más de 100.000 empleados y presencia en todo el país, YPF es líder en producción de recursos no convencionales, destacándose la formación Vaca Muerta, uno de los mayores del mundo en gas de lutita y petróleo de esquisto.

4.1.2 GGAL

Grupo Galicia (GGAL) es un holding financiero argentino fundado en 1999 que opera principalmente a través de sus subsidiarias en Argentina. Sus principales empresas incluyen Banco Galicia, uno de los tres bancos privados más grandes del país, Tarjetas Regionales, Sudamericana Holding, y Galicia Administradora de

Fondos. Según su sitio institucional, ofrece una amplia gama de servicios financieros como ahorro, crédito, seguros, depósitos, préstamos comerciales y personales, hipotecas, banca de inversión, corretaje de valores, y gestión de fondos. Con más de 9.500 empleados, 550 sucursales y cerca de 8 millones de clientes, GGAL es un referente en el sector financiero privado argentino.

4.2 Origen de la información

Se trabajó con información de carácter público brindada en los estados financieros trimestrales presentados a la Comisión Nacional de Valores. Dichos estados contables se presentan aplicando el régimen las Normas Internacionales de Información Financiera (NIIF). La obligatoriedad de aplicar esta normativa entró en vigencia en el 2011 para mejorar la transparencia y compatibilidad de la información financiera a nivel global, mediante un lenguaje común utilizable por los distintos sectores económicos e industrias.

Los valores de las variables correspondientes a los campos de los estados contables se obtendrán a partir de los documentos presentados por las empresas ante la Comisión Nacional de Valores (CNV). Para los datos relativos al precio de la acción antes y después de la presentación de los estados contables, así como para la información del índice SPY, se recurrirá a la información pública proporcionada por el mercado NYSE. Los valores estimados para cada uno de los campos de las empresas serán extraídos de la base de datos de Bloomberg, reconocida por ser una fuente confiable de datos financieros y bursátiles.

4.3 Modelo a utilizar

En el presente estudio se empleará el modelo SARIMAX *(Seasonal AutoRegressive Integrated Moving Average with eXogenous factors*) como herramienta principal para la predicción del precio de las acciones tras la presentación de los estados contables. Este modelo es una extensión del modelo ARIMA, que permite no solo capturar patrones de tendencia y estacionalidad en series temporales, sino también incorporar la influencia de variables exógenas que pueden tener un impacto significativo en la evolución de la serie, como son los indicadores financieros y macroeconómicos relevantes. La elección del modelo SARIMAX está fundamentada en su capacidad para manejar tanto la complejidad inherente a las series temporales financieras, caracterizadas por su volatilidad y dependencia temporal, como la necesidad de incluir factores externos, como las ratios financieras y las condiciones económicas globales, que afectan al comportamiento

del precio de las acciones. Este enfoque permitirá capturar de manera más precisa la dinámica del mercado y ofrecer una herramienta robusta para la toma de decisiones en el ámbito financiero. Para asegurar la validez del modelo y su capacidad predictiva, se realizarán análisis exhaustivos de los residuos y pruebas de diagnóstico, garantizando que los supuestos subyacentes se cumplen y que el modelo proporciona estimaciones fiables.

4.4 Variable respuesta

Para cada observación, la variable respuesta del modelo es el precio de cierre posterior a la presentación de los estados contables. Si la presentación se realiza antes de la apertura del mercado, el precio objetivo es el precio de cierre de ese mismo día; en cambio, si la presentación ocurre después del cierre del mercado, el precio objetivo es el precio de cierre del día siguiente.

4.5 Variables explicativas

4.5.1 Variables correspondientes a los campos de los estados contables

Estas variables pertenecen a los estados financieros de la empresa, los cuales son el balance general, el estado de resultados y el estado de flujos de efectivo. Cada uno tiene información sensible y juntos narran la performance que la empresa ha tenido durante el periodo analizado. El alcance, composición, partidas contables que incluye y excluye cada variable considerada se define mediante las Normas Internacionales de Información Financiera (NIIF).

Para este conjunto de variables, los valores se calculan como:

$$\text{Valor} = \frac{Valor\_REAL}{Valor\_ESPERADO} - 1 \qquad (4.1)$$

Se interpreta como una medida relativa del rendimiento o desviación del valor real presentado en los estados contables respecto al valor esperado por los analistas para dicho campo o ratio. Esta medida es útil para normalizar y comparar diferentes variables en términos de sus desviaciones relativas respecto a sus valores esperados.

Interpretación:

- Valor igual a 0: Indica que el valor real es igual al valor esperado. No hay desviación.

- Valor mayor a 0: Indica que el valor real es mayor que el valor esperado. La magnitud del valor positivo representa el porcentaje por el cual el valor real supera al valor esperado. Si la variable es 0.2, esto significa que el valor real es un 20% mayor que el valor esperado.
- Valor menor a 0: Indica que el valor real es menor que el valor esperado. La magnitud del valor negativo representa el porcentaje por el cual el valor real está por debajo del valor esperado. Por ejemplo, si la variable es -0.1, esto significa que el valor real es un 10% menor que el valor esperado.

A continuación, se presentan los campos y ratios de los estados financieros que se utilizan como variables.

- **EBITDA_TO_REVENUE** (EBITDA entre Ingresos)

Este índice se obtiene dividiendo el EBITDA (Beneficio Antes de Intereses, Impuestos, Depreciación y Amortización) entre los Ingresos Totales. Mide la proporción de ingresos convertidos en EBITDA, proporcionando una visión de la eficiencia operativa excluyendo gastos no operativos.

- **PX_TO_EBITDA**: (Precio entre EBITDA)

Mide la relación entre el precio de mercado de una empresa y su EBITDA. Esta ratio es utilizada para evaluar si una empresa está sobrevalorada o infravalorada en relación con su capacidad de generar beneficios operativos. Un valor bajo de "Precio/EBITDA" puede indicar que la empresa está infravalorada o es una buena oportunidad de inversión, mientras que un valor alto puede sugerir una sobrevaloración.

- **SALES_REV_TURN** (Ingresos por ventas)

Calculado como las ventas netas divididas por los activos totales, esta ratio evalúa la eficiencia con la que la empresa utiliza sus activos para generar ventas. Su análisis proporciona información relevante sobre la eficiencia operativa, que impacta las proyecciones de ingresos futuros y el valor de la acción.

- **EARN_FOR_COM_TO_TOT_REV** (Ganancia entre Ingresos)

Mide qué porcentaje de los ingresos totales de una empresa se traduce en ganancias netas disponibles para los accionistas comunes. Se calcula como las ganancias atribuidas a los accionistas comunes dividido por los ingresos totales, expresado como porcentaje. Esta métrica es útil para evaluar la eficiencia con la que una empresa convierte sus ingresos en beneficios para sus accionistas, reflejando la rentabilidad después de cubrir los costos operativos y financieros, así como los impuestos.

- **CF_FREE_CASH_FLOW** (Flujo de caja libre)

Determinado como el flujo de caja operativo menos los gastos de capital, este indicador representa el efectivo disponible tras cubrir los gastos de capital necesarios para mantener o expandir la base de activos. Es fundamental para evaluar la capacidad de la empresa para invertir en el negocio o devolver valor a los accionistas, siendo una medida clave de la salud financiera.

- **CF_NET_INC** (Ingreso neto entre flujo de caja operativo)

Se calcula dividiendo la utilidad neta por el flujo de caja operativo. Esta ratio compara la utilidad neta con el flujo de caja operativo, proporcionando una perspectiva sobre la eficiencia en la conversión de la utilidad neta en flujo de caja. Un valor bajo puede sugerir problemas en la calidad de las ganancias.

- **IS_COMP_NET_INCOME** (Ingreso neto ajustado)

Este valor ajusta la utilidad neta para eliminar elementos no recurrentes o no operativos, reflejando una utilidad neta más representativa de las operaciones normales de la empresa. Su propósito es proporcionar una visión más precisa del desempeño operativo ajustado.

- **IS_COMP_SALES** (Ventas)

Calcula las ventas ajustadas para excluir efectos de adquisiciones, desinversiones o cambios significativos en la estructura de la empresa. Ofrece una visión más clara del crecimiento orgánico de las ventas, permitiendo ajustar las expectativas basadas en un análisis más puro de ventas comparables.

- **IS_COMPARABLE_EBITDA** (EBITDA)

Ajusta el EBITDA para eliminar efectos de eventos no recurrentes o cambios significativos en la estructura de la empresa. Mide la rentabilidad operativa ajustada y permite una evaluación más precisa del desempeño operativo,

- **IS_COMPARABLE_EBIT** (EBIT)

Este valor ajusta el EBIT para eliminar elementos no recurrentes o cambios importantes en la estructura de la empresa, proporcionando una medida del beneficio operativo ajustado. Permite ajustar la estimación de beneficios operativos para prever con mayor precisión el desempeño futuro.

- **IS_OPER_INC** (Ingreso operativo)

Es equivalente al EBIT, representando el ingreso operativo de la empresa, es decir, el beneficio antes de intereses e impuestos. Este índice proporciona una medida fundamental de la rentabilidad operativa de la empresa.

4.5.2 Variables correspondientes a precios y tasa de interés

- **PX_PRE_BALANCE**

Precio de la acción antes de presentar balance. Último precio que tenía la acción antes de la presentación del balance.

- **PX_SPY**

El precio del SPY, que es el ETF del principal índice accionario del mundo, es crucial para ajustar un modelo de predicción de precios de acciones porque refleja variables globales y el nivel general de precios del mercado. Captura la salud y tendencias del mercado bursátil global, las expectativas de crecimiento económico y el sentimiento de los inversores, lo que puede influir significativamente en las valoraciones individuales de las acciones.

- **YIELD GOVT 10**

La tasa de interés de referencia de los bonos del gobierno de Estados Unidos a 10 años refleja la percepción del mercado sobre la salud económica a largo plazo, influye en el costo del capital y las tasas de interés a largo plazo, sirve como indicador de inflación y afecta las decisiones de los inversores al compararla con las rentabilidades potenciales de otros activos.

5. Modelos ajustados

5.1 Modelos sobre GGAL

El conjunto de datos contiene 43 observaciones que cubren el periodo entre el 1 de enero de 2014 y el 31 de diciembre de 2024, correspondientes a los estados contables trimestrales de la empresa GGAL. Cada observación incluye 11 variables: una variable objetivo, GGAL_post_eecc, y 10 variables explicativas (tabla 5.1), todas detalladas previamente en la sección 4.1 de este trabajo.

Para proceder con el ajuste de los modelos SARIMAX, el conjunto de datos fue dividido en dos subconjuntos: el 80% de los datos fue reservado para el entrenamiento del modelo, mientras que el 20% restante se utilizó para evaluar su capacidad de predicción sobre datos no vistos. Esta partición asegura una evaluación más confiable del rendimiento de los modelos propuestos, evitando el sobreajuste y maximizando su capacidad de generalización a futuros períodos.

Variable	Detalle
PX_TO_EBITDA	Precio / EBITDA
EBITDA_TO_REVENUE	EBITDA / Ingresos Totales
CF_NET_INC	Utilidad Neta / Flujo de Caja Operativo
IS_COMP_NET_INCOME	Utilidad neta
SALES_REV_TURN	Ventas Netas / Activos Totales
IS_OPER_INC	Ingreso operativo
EARN_FOR_COM_TO_TOT_REV	(Resultado neto / Patrimonio neto) / Ingresos
YIELD GOVT 10 PRE BALANCE	Tasa de interés de bonos de US a 10 años.
SPY_pre_eecc	Precio del SPY el día anterior al balance.
GGAL_pre_eecc	Precio de GGAL el día anterior al balance.

Tabla 5.1

5.1.1 Análisis descriptivo

En esta sección, se presenta un análisis detallado de las características principales del conjunto de datos. En la imagen 5.1, se observa la evolución del precio de la acción GGAL después de la presentación de los estados contables. A lo largo del periodo, se pueden notar varios picos y valles, destacándose el fuerte incremento hacia el año 2018, seguido de una marcada caída. Estos ciclos se asocian a los periodos políticos vividos en Argentina.

Imagen 5.1

En el Anexo 9.3 se presenta el análisis exploratorio del conjunto de datos, donde se examinan la distribución y la dispersión de cada variable. Para ello se utilizó una matriz de gráficos de dispersión y diagramas de densidad correspondientes a las variables derivadas de los estados contables. Esta exploración preliminar no solo permite visualizar posibles relaciones o asociaciones entre las variables, sino también identificar valores atípicos y verificar supuestos de normalidad. De este modo, se sienta una base sólida para las etapas posteriores del análisis estadístico, facilitando la selección de variables y la aplicación de transformaciones que mejoren el ajuste de los modelos.

El análisis de estacionariedad es fundamental en modelos de series temporales, ya que estos modelos asumen que la serie es estacionaria, lo que implica que sus propiedades estadísticas, como la media y la varianza, se mantengan constantes a lo largo del tiempo.

Para verificar la estacionariedad de la serie temporal, ejecutamos la prueba de Dickey-Fuller aumentada (ADF), la cual contrasta la hipótesis nula de que la serie presenta una raíz unitaria, es decir, no es estacionaria. El resultado arrojó un p-value de 0.48, lo cual no permite rechazar la hipótesis nula al nivel de significancia de 0.05. Esto indica que la serie no es estacionaria, y, por lo tanto, es necesario aplicar transformaciones, como la diferenciación, para estabilizar las propiedades estadísticas y permitir el ajuste adecuado del modelo. Esta falta de estacionariedad se evidencia también en el gráfico 5.1.

Realizamos la prueba de Dickey-Fuller aumentada (ADF) sobre la serie diferenciada, obteniendo un p-value de 0.0005, lo que es significativamente menor a 0.05. Este resultado nos permite rechazar la hipótesis nula, confirmando que la

serie diferenciada es estacionaria. De esta forma, la serie queda preparada para ser modelada con un modelo SARIMA. Los efectos de la diferenciación se pueden observar en el gráfico 5.2.

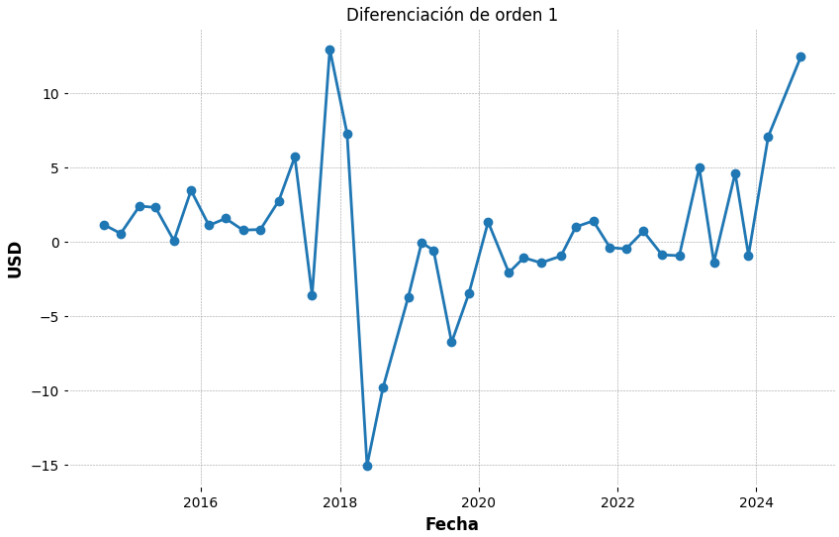

Imagen 5.2

Una vez confirmada la estacionariedad de la serie diferenciada, realizamos un análisis de autocorrelación (ACF) y autocorrelación parcial (PACF) para identificar patrones de dependencia temporal en la estructura de la serie. La función de autocorrelación examina cómo una observación está correlacionada con valores en distintos rezagos temporales, mientras que la autocorrelación parcial permite aislar la relación entre una observación y un rezago, eliminando el efecto de los rezagos intermedios.

En los gráficos 5.3 y 5.4 vemos los gráficos de autocorrelación y autocorrelación parcial respectivamente, antes y después de diferenciar. Se puede apreciar como las series diferenciadas muestran un comportamiento que refleja la ausencia de tendencia o patrones repetitivos a largo plazo. En el ACF se observa que las correlaciones se acercan rápidamente a cero después de unos pocos rezagos, lo que indica que las observaciones pasadas tienen una influencia limitada en los valores futuros. En el PACF, los coeficientes significativos se reducen rápidamente, mostrando un corte abrupto después del primer rezago.

El comportamiento de la serie indica que la misma no tiene una estructura temporal fuerte que se extienda más allá de un corto periodo, lo cual es típico en series estacionarias, avalando que la podamos modelar mediante procesos ARIMA.

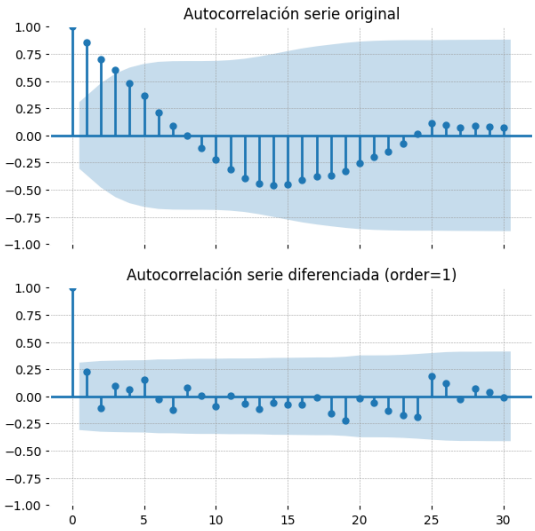

Imagen 5.3

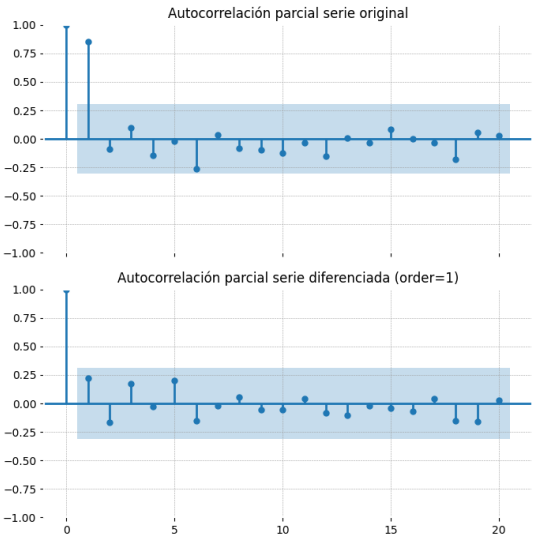

Imagen 5.4

5.1.2 Ingeniería de características

Con el fin de preparar el conjunto de datos para el ajuste de los modelos, se llevaron a cabo las modificaciones necesarias para tratar los valores atípicos. Si bien no existen datos faltantes, sí se detectaron observaciones extremas en algunas variables, las cuales pueden distorsionar el ajuste de los modelos SARIMAX y comprometer su capacidad para predecir el comportamiento del precio de la acción. Estos valores atípicos fueron analizados y corregidos mediante imputación basada en rangos intercuartílicos, garantizando que el conjunto de datos mantenga una distribución adecuada para el modelado. En el Anexo 9.3 se proporciona un detalle pormenorizado de este proceso, incluyendo los umbrales utilizados y el criterio de sustitución aplicado.

A su vez, dado que algunas variables tienen fuertes asociaciones, y motivados por la poca cantidad de observaciones con relación al número de variables, es interesante plantear un modelo con una reducción de la dimensionalidad de las variables. Para ello, el Análisis de Componentes Principales (ACP) resulta un método adecuado, ya que permite capturar la mayor parte de la variabilidad de los datos en un conjunto reducido de nuevas variables.

El detalle de su aplicación se presenta en el anexo 9.4. Para determinar cuántas componentes principales se utilizarán, se analiza la varianza explicada por cada una de ellas, tal como se observa en la imagen 5.5. El análisis de la varianza permite identificar cuántas componentes son necesarias para capturar un porcentaje significativo de la variabilidad total del conjunto de datos. En este caso, para explicar el 90% de la varianza, se requieren cuatro componentes principales.

Esto indica que estas cuatro dimensiones contienen la mayor parte de la información original, permitiendo así una reducción dimensional de 7 a 4 variables sin perder información crucial para el análisis. El umbral del 90% surge a partir de la necesidad de equilibrar la cantidad de información retenida con la simplicidad del modelo, evitando el uso de demasiadas componentes que podrían introducir ruido o redundancia. En este sentido, las cuatro componentes seleccionadas proporcionan una representación compacta y eficiente del comportamiento de las variables financieras.

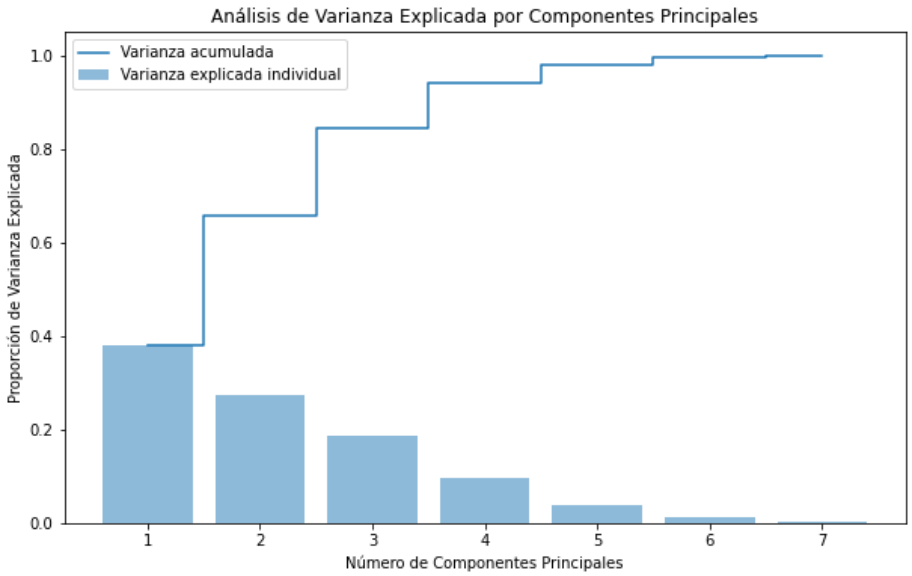

Imagen 5.5

Al detectar un patrón de dependencia conjunta entre el precio del índice SPY y la tasa de interés, se infiere que su interacción puede incidir de manera significativa sobre el valor de Galicia. En consecuencia, se decidió incorporar una nueva variable que capture este efecto de interacción en los modelos predictivos, con el fin de mejorar la capacidad de ajuste y de captar dinámicas no lineales que de otro modo quedarían ocultas. En el Anexo 9.5 se describe en detalle el procedimiento utilizado para generar dicha variable de interacción y su inclusión en los modelos SARIMAX, así como el impacto observado en las métricas de ajuste.

5.1.3 Modelos ajustados sobre Galicia

Para ajustar cada uno de los modelos propuestos se utiliza un procedimiento de búsqueda por grilla (conocido como *grid search*) para encontrar los mejores parámetros p,q,P,D,Q,S; donde:

- p,q: órdenes del modelo autorregresivo y de media móvil, respectivamente.
- P,Q,D,S: órdenes correspondientes a la componente estacional del modelo.

Los parámetros se seleccionan optimizando el criterio de Akaike (AIC), una métrica que evalúa la calidad de los modelos estimados penalizando su complejidad. El AIC mide el balance entre el ajuste de los datos y el número de parámetros, donde un menor valor de AIC indica un mejor modelo ajustado.

Para cada modelo, se entrenaron 1.250 combinaciones de parámetros, correspondientes a las variaciones de parámetros p, q, P, D, Q de 0 a 3; el parámetro S en los valores 4 y 12; y d = 1.

5.1.3.1 SARIMA

Como Modelo 1, se ajustó un modelo univariado SARIMA (*Seasonal AutoRegressive Integrated Moving Average),* que es una extensión del modelo ARIMA que incluye una componente estacional. El modelo SARIMA es útil cuando los datos presentan patrones estacionales, permitiendo modelar la tendencia, estacionalidad y componentes de ruido de una serie temporal.

El modelo con el menor AIC fue el que presenta los parámetros: (3,1,1)×(3,1,2,4), lo que indica que el componente autorregresivo y estacional tienen un orden de 3, con una diferenciación de primer orden, y una estacionalidad de 4 periodos. Matemáticamente, el modelo se expresa como:

$$(1 - \phi_1 L - \phi_2 L^2 - \phi_3 L^3)(1 - \phi_1 L^S - \phi_2 L^{2S} - \phi_3 L^{3S})(1 - L)(1 - L^S) y_t = (1 + \theta_1 L)(1 + \Theta_1 L^S + \Theta_2 L^{2S})\varepsilon_t \quad (5.4)$$

Este modelo arrojó un RMSE de 43.9, lo que implica que el error promedio entre los valores reales y predichos es de 43.9 dólares. Considerando que el precio promedio de la acción es de 20 dólares, este error es significativo, lo que sugiere la necesidad de mejorar el ajuste o incorporar variables exógenas al modelo para capturar mejor la dinámica del mercado.

5.6.2 SARIMAX

En el modelo 2 se agregaron todas las variables explicativas (tanto las provenientes de los estados contables como aquellas que proporcionan información sobre el mercado) al modelo anterior, este mejoró considerablemente, con un RMSE de 0.79. Este resultado indica que la predicción del precio de la acción se volvió mucho más precisa al incluir estas variables, lo que resalta la importancia de las variables explicativas en la mejora del modelo. La inclusión de información adicional sobre el mercado y los estados contables permitió capturar mejor las dinámicas que afectan el precio.

El modelo SARIMAX que mejor ajusta es el que tiene los parámetros: (1, 1, 1) x (2, 0, 1, 4).

5.6.3 SARIMAX Con componentes principales

El Modelo 3 reemplaza las siete variables provenientes de los estados contables por cuatro componentes principales que explican el 90% de la varianza, lo que reduce el número de variables explicativas de 10 a 7. Estas 7 variables incluyen las cuatro componentes principales, el precio de la acción y del SPY antes de la presentación de los estados contables, y la tasa de interés de los bonos del gobierno de Estados Unidos a 10 años. Este modelo obtuvo un RMSE de 0.76, lo que indica un rendimiento superior al modelo con todas las variables explicativas originales. Señal de que trabajar con componentes principales en vez de las variables originales, no sólo ayuda a la parsimonia del modelo, sino que disminuye el sobre ajuste al no capturar el ruido del conjunto de datos de entrenamiento.

El modelo SARIMAX que mejor ajusta es el que tiene los parámetros: (1, 1, 1) x (2, 0, 2, 4).

5.6.4 SARIMAX Con componentes principales e interacción de variables

Debido a la asociación entre la tasa de interés y el precio del índice SPY, se decidió crear una nueva variable, denominada YIELD_i_SPY, que captura la interacción entre ambas mediante su producto. Esta variable fue añadida al conjunto de datos del Modelo 3, que ya incluía las cuatro componentes principales de las variables contables, los precios de GGAL y SPY previos a la presentación de los estados contables, y la tasa de interés de los bonos a 10 años. Con esta nueva variable, el modelo ajustado resultó en un RMSE de 1.25, lo cual señala que la nueva variable interacción no aporta información relevante o introduce multicolinealidad.

El modelo SARIMAX que mejor ajusta es el que tiene los parámetros: (1, 1, 1) x (2, 0, 1, 4).

5.1.4 Modelo propuesto para Galicia

El modelo que mejor ajustó a los datos de testeo es el modelo 3, el cual utiliza componentes principales de las variables derivadas de los estados contables. Para este modelo proseguimos a analizar los residuos debido a que, si los residuos muestran patrones sistemáticos, podrían indicar que el modelo no captura correctamente la estructura subyacente de los datos o que hay factores no modelados que afectan el desempeño predictivo.

En gráfico 5.6 se muestran los residuos a lo largo del tiempo, observamos una tendencia general a oscilar alrededor de cero, lo que sugiere que no hay un sesgo significativo en las predicciones. Además, se observa que los residuos no presentan patrones de correlación ni se modifica su rango a lo largo del tiempo.

En el gráfico 5.7 se observa la serie temporal a predecir junto a las estimaciones del modelo. Ambas curvas se superponen en la mayoría de los puntos, lo que sugiere un buen ajuste general. Sin embargo, en la última observación se aprecia una discrepancia más pronunciada. Este desajuste podría estar relacionado con el contexto de volatilidad económica que vivió Argentina en la fase previa a las elecciones Primarias Abiertas Simultáneas y Obligatorias (PASO), celebradas en octubre de 2023, lo cual habría afectado los precios de las acciones.

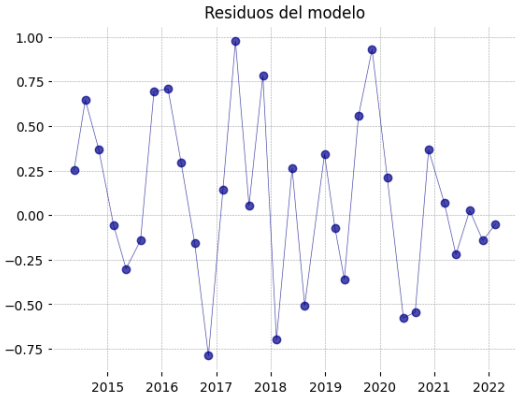

Imagen 5.6

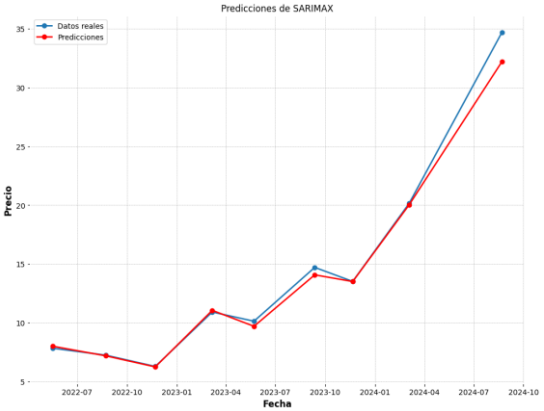

Imagen 5.7

En el histograma de la imagen 5.8, se observa que los residuos no siguen una distribución claramente normal, ya que hay acumulación de valores cercanos a 0.25 y -0.25, y los extremos presentan mayor frecuencia de lo esperado, lo que sugiere posibles problemas de ajuste en ciertos rangos del modelo. Para validar esto, se realizaron dos pruebas estadísticas: la prueba de normalidad de Shapiro-Wilk, que evalúa si una muestra proviene de una distribución normal (H_0: los datos siguen una distribución normal; H_1: no siguen una distribución normal), y la prueba de Jarque-Bera, que contrasta si los residuos tienen una curtosis y asimetrías compatibles con una distribución normal (H_0: los datos son normales; H_1: no lo son). Los *p-values* obtenidos fueron de 0.8 y 0.7, respectivamente, lo que indica que no se puede rechazar la hipótesis nula en ambos casos. Así, los errores se distribuyen normalmente según estas pruebas.

Visualmente se complementa con el gráfico Q-Q (imagen 5.9). Este gráfico muestra cómo los residuos se alinean con la distribución normal teórica. Si bien hay ligeras desviaciones en los extremos, en su mayoría los puntos se alinean bien con la línea diagonal roja, lo que indica que los residuos siguen razonablemente una distribución normal. Las desviaciones en los cuantiles más altos y bajos pueden

indicar que el modelo no captura perfectamente los valores extremos, pero la distribución es aceptablemente cercana a la normalidad.

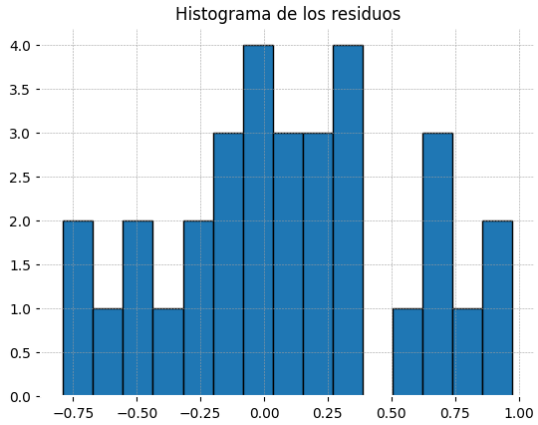

Imagen 5.8

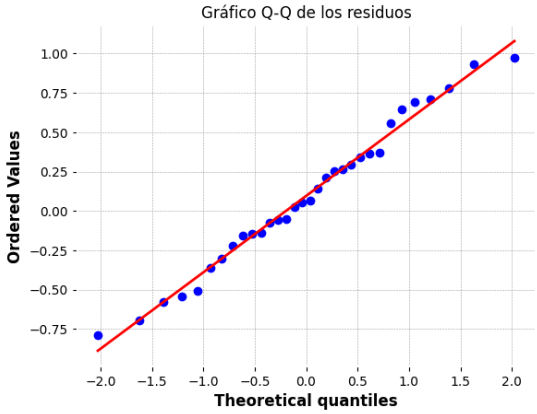

Imagen 5.9

Las imágenes 5.10 y 5.11 muestran los gráficos de correlación y correlación parcial respectivamente. En el ACF se observa que las correlaciones se acercan rápidamente a cero después de unos pocos rezagos, lo que indica que las observaciones pasadas tienen una influencia limitada en los valores futuros. En el PACF, los coeficientes significativos se reducen rápidamente, mostrando un corte abrupto después del primer rezago.

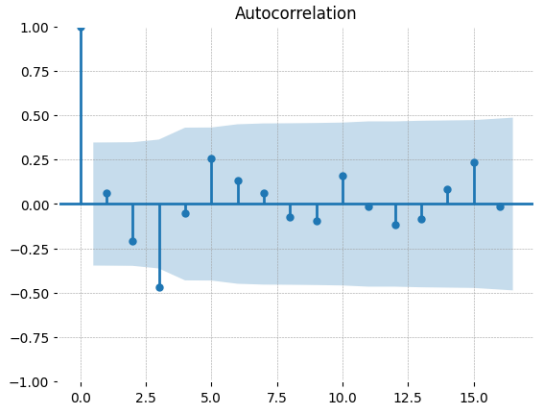

Imagen 5.10

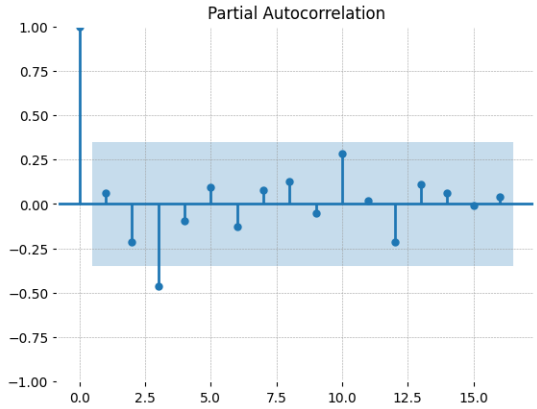

Imagen 5.11

El análisis de los residuos permitió verificar que el modelo cumple con los supuestos de homocedasticidad, ausencia de autocorrelación y normalidad.

5.2. Modelos sobre YPF

El conjunto de datos utilizado para analizar la empresa YPF comprende 43 observaciones que abarcan el período del 1 de enero de 2014 al 31 de diciembre de 2024, correspondientes a los estados contables trimestrales. En cada observación, se incluyen 16 variables: una variable objetivo, YPF_post_eecc, y 12 variables explicativas (tabla 5.2), las cuales han sido previamente detalladas en la sección 4.1. Entre las variables explicativas, destacamos que EBITDA_TO_REVENUE, CF_NET_INC, IS_COMP_NET_INCOME, SALES_REV_TURN, e IS_OPER_INC también fueron utilizadas en el análisis de la empresa Galicia.

Variable	Detalle
EBITDA_TO_REVENUE	EBITDA / Ingresos Totales
CF_NET_INC	Utilidad Neta / Flujo de Caja Operativo
IS_COMP_NET_INCOME	Utilidad neta
SALES_REV_TURN	Ventas Netas / Activos Totales
IS_OPER_INC	Ingreso operativo
CF_FREE_CASH_FLOW	Flujo de Caja Operativo menos los Gastos de Capital
IS_COMP_SALES	Ventas ajustadas
IS_COMPARABLE_EBITDA	EBITDA Ajustado
IS_COMPARABLE_EBIT	EBIT Ajustado
YIELD GOVT 10 PRE BALANCE	Tasa de interés de bonos de US a 10 años.
SPY_pre_eecc	Precio del SPY el día anterior al balance.
GGAL_pre_eecc	Precio de GGAL el día anterior al balance.

Tabla 5.2

5.2.1 Análisis descriptivo

En la imagen 5.12, se muestra la evolución del precio de la acción YPF posterior a la presentación de los estados contables. El gráfico está segmentado en dos partes: el conjunto de entrenamiento y el conjunto de prueba. A lo largo del período analizado, se puede observar una tendencia negativa predominante en el precio de la acción, que experimenta una notable reversión después de las elecciones en Argentina en 2023.

Imagen 5.12

En el Anexo 9.6 se presenta el análisis exploratorio del conjunto de datos, donde se examina la distribución de cada variable y se identifican valores atípicos. Estos últimos fueron tratados de la misma manera que en el conjunto de datos de Galicia, mediante imputación basada en rangos intercuartílicos para garantizar un comportamiento más homogéneo de las series.

En lo que respecta a la interacción de variables, el Anexo 9.7 muestra que la combinación más sobresaliente corresponde a YPF_pre_eecc y SPY_pre_eecc, lo cual sugiere una relación entre el valor de la empresa y el índice SPY, tradicionalmente considerado un indicador del "sentimiento" de los inversores. Esta interacción se incorporó como variable adicional en los modelos con el objetivo de capturar efectos conjuntos que podrían escapar a un análisis univariante.

Dado que las variables provenientes de los estados contables están intrínsecamente correlacionadas, pues varias reflejan aspectos similares de rentabilidad, eficiencia operativa y liquidez, se detectó una redundancia que podría afectar la estabilidad del modelo. Por ello, en el Anexo 9.8 se aplicó el Análisis de Componentes Principales (ACP), que sintetiza la información en un número reducido de componentes no correlacionados. En este caso, se seleccionaron cinco componentes principales que explican el 90 % de la varianza total, optimizando así la modelización sin sacrificar la información relevante.

Tal como se observa en la figura 5.12, la serie temporal original no es estacionaria, ya que presenta tendencia y heterocedasticidad. Este comportamiento es corroborado por la prueba de Dickey-Fuller aumentado, el cual arroja un *p value* de 0.48, lo que sugiere no rechazar la hipótesis nula de no estacionariedad. Para

solucionar esto, se aplica una diferenciación de orden 1, generando la serie que se presenta en la figura 5.13. En esta nueva serie, el *p value* de la prueba es 6.9 × 10^-11, indicando que ahora sí es estacionaria, lo que permite ajustar un modelo ARIMA.

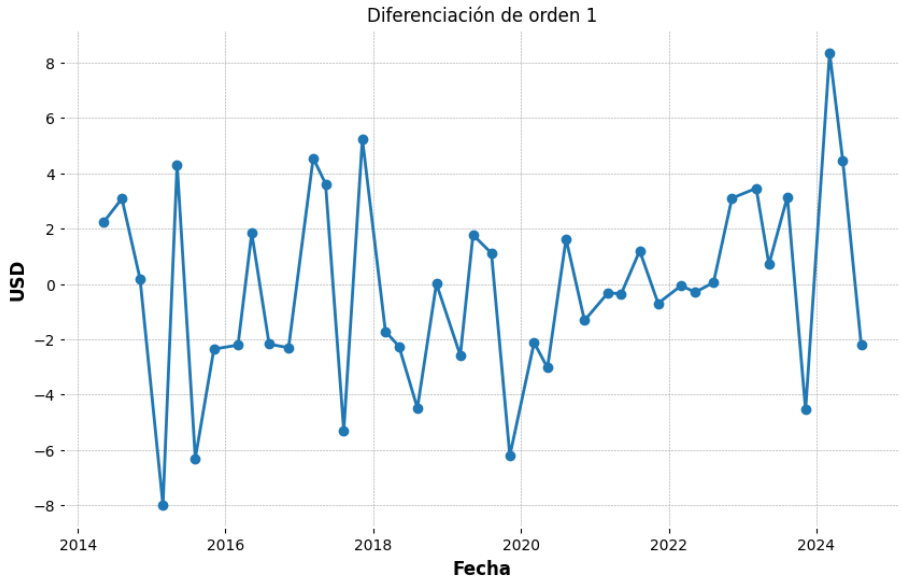

Imagen 5.13

Esto se puede ver también en los gráficos de autocorrelación (ACF) y autocorrelación parcial (PACF) de las figuras 5.14 y 5.15, respectivamente. Al ser una serie que es estacionaria, el ACF muestra una rápida disminución de la correlación después de los primeros rezagos, indicando que no hay una correlación significativa a largo plazo. En cuanto a la PACF, presenta solo un rezago significativo, lo que sugiere un modelo ARIMA adecuado para capturar la estructura de la serie.

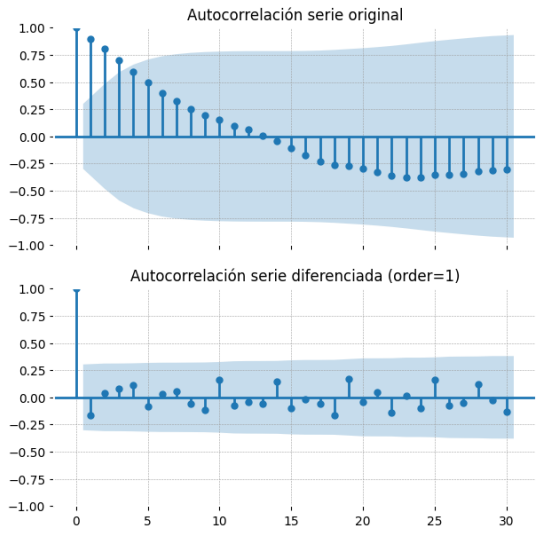

Imagen 5.14

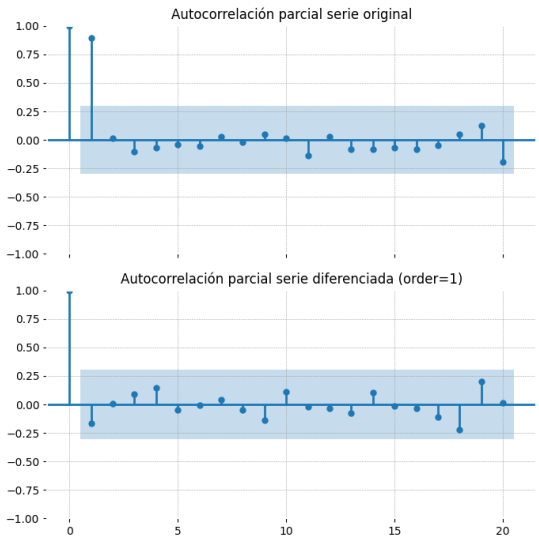

Imagen 5.15

5.2.2 Modelos ajustados sobre YPF

Para ajustar cada uno de los modelos propuestos se utilizó el procedimiento de búsqueda por grilla para encontrar los mejores parámetros p,d,q,P,D,Q,S; donde:

- p,d,q: órdenes del modelo autorregresivo y de media móvil, respectivamente.
- P,Q,D,S: órdenes correspondientes a la componente estacional del modelo.

Se entrenaron 1.250 combinaciones de modelos, correspondientes a las variaciones de parámetros p, q, P, D, Q de 0 a 3; el parámetro S en los valores 4 y 12; y d = 1 y 2. Los parámetros se seleccionan optimizando el criterio de Akaike.

5.2.2.1 SARIMA

Como primer modelo, se ajustó un SARIMA univariado, que solo considera la información proporcionada por la variable objetivo. El modelo óptimo fue ajustado con los parámetros (1, 2, 1)(1, 2, 0, 4), lo que implica una diferenciación de orden 2 tanto en el componente regular como en el estacional. Tras el ajuste, el modelo arrojó un RMSE de 7.1.

5.2.2.2 SARIMAX

Para aprovechar la información de las variables explicativas, se ajustó un modelo SARIMAX, cuya parametrización óptima fue (1, 1, 2)(2, 0, 1, 4). Este modelo redujo considerablemente el error, arrojando un RMSE de 0.6. La marcada diferencia con respecto al SARIMA univariado (RMSE 7.1) evidencia la importancia de incluir variables explicativas, ya que estas aportan información adicional que captura mejor las relaciones subyacentes en los datos, mejorando la precisión predictiva del modelo y su capacidad para ajustarse a las fluctuaciones del entorno financiero.

5.2.2.3 SARIMAX Con componentes principales

El modelo número tres fue ajustado utilizando componentes principales para las variables provenientes de los estados contables. Esta técnica permitió reducir la dimensionalidad, lo que ayuda a evitar la captura de ruido durante el entrenamiento del modelo. El modelo óptimo fue parametrizado con (1, 1, 1) y un orden estacional de (2, 0, 2, 4), logrando un RMSE de 0.53. Este resultado refuerza la ventaja de la reducción de dimensionalidad, permitiendo que el modelo se enfoque en la información más relevante para mejorar la predicción.

5.2.2.4. SARIMAX Con componentes principales e interacción de variables

Para capturar las relaciones entre las variables que no provienen de los estados

contables, se ajustó un modelo que incluye la interacción entre las variables YPF_pre_eecc y SPY_pre_eecc. Este enfoque permitió tener en cuenta cómo los movimientos del índice SPY influyen en la acción de YPF antes de la presentación de los estados contables. El modelo optimizado resultante, con parámetros (1, 1, 2) (2, 1, 2, 4), arrojó un RMSE de 0.47, lo que indica una mejora notable en comparación con modelos sin interacción.

5.2.3 Modelos propuesto.

El modelo que mejor ajustó a los datos de *testing* es el modelo 4, el cual utiliza componentes principales de las variables derivadas de los estados contables e interacción de las variables que no provienen de los estados contables. Para este modelo se prosigue a analizar los residuos.

En los gráficos 5.16 y 5.17 se puede observar un análisis complementario del ajuste del modelo. En el primero, los residuos se distribuyen de manera uniforme alrededor de cero, lo que indica ausencia de sesgo y de patrones de correlación, sugiriendo un buen comportamiento del modelo a lo largo del tiempo. En el segundo gráfico, la serie temporal a predecir se compara con las estimaciones, observándose una superposición significativa en la mayoría de los puntos, lo que refuerza la idea de un ajuste adecuado.

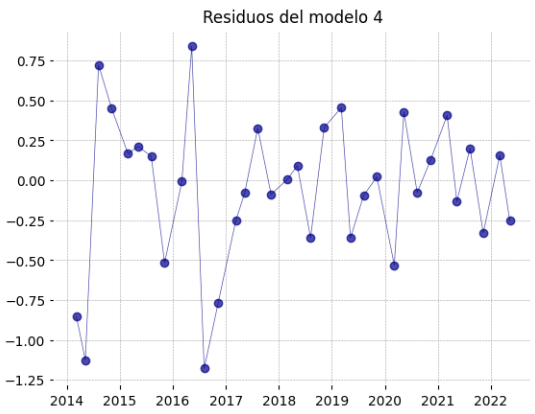

Imagen 5.16

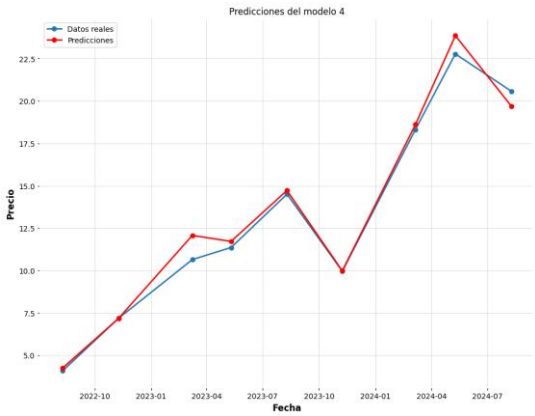

Imagen 5.17

Para evaluar la normalidad de los residuos, se realizaron dos pruebas estadísticas: la prueba de Jarque-Bera y la prueba de Shapiro-Wilk. El *p value* de la prueba de Jarque-Bera es 0.39, y el de la prueba de Shapiro-Wilk es 0.33, ambos superiores al umbral típico de 0.05. Esto indica que no se puede rechazar la hipótesis nula de que los residuos siguen una distribución normal, lo que sugiere que los residuos del modelo presentan una distribución cercana a la normalidad. Estos resultados se

complementan con el gráfico QQ, en donde las observaciones se acoplan sobre la recta identidad (gráfico 5.18) y sobre el histograma de los residuos, que se aproximan a un comportamiento normal (imagen 5.19).

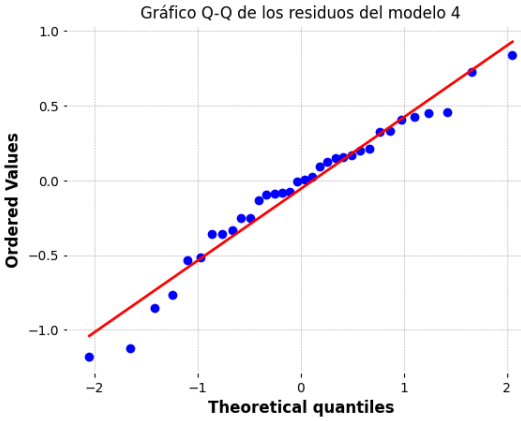

Imagen 5.18

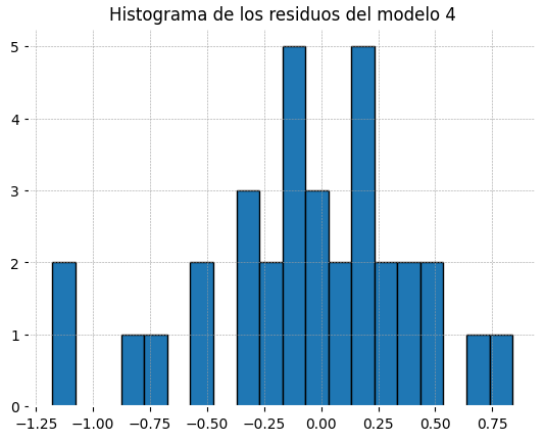

Imagen 5.19

Las imágenes 5.20 y 5.21 presentan los gráficos de autocorrelación (ACF) y autocorrelación parcial (PACF) de los residuos, respectivamente. En el ACF, se observa que las correlaciones caen rápidamente a cero después de un rezago, sugiriendo una influencia limitada de las observaciones pasadas en los valores futuros. El PACF también muestra un corte abrupto después del primer rezago, con pocos coeficientes significativos. Esto se refuerza con los resultados de la prueba de Ljung-Box, donde los *p values* (0.971, 0.609, 0.489) indican la ausencia de autocorrelación significativa en los residuos.

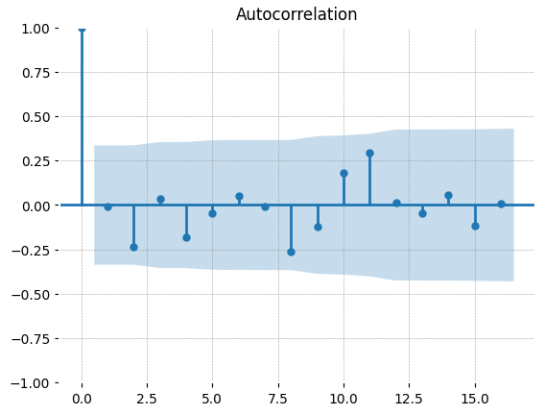

Imagen 5.20

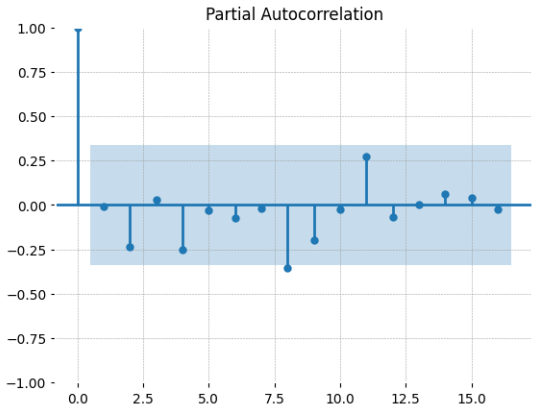

Imagen 5.21

El análisis de los residuos permitió verificar que el modelo cumple con los supuestos de homocedasticidad, ausencia de autocorrelación y normalidad.

6. Resultados

En la tabla 6.1 se presenta un resumen de todos los modelos propuestos, donde se comparan sus desempeños utilizando la métrica raíz del error cuadrático medio (RMSE). Esta métrica proporciona una medida de la diferencia promedio entre los valores observados y los predichos, penalizando los errores grandes de manera más significativa, lo cual nos permite identificar no solo qué modelo se ajusta mejor a los datos, sino también su capacidad para generalizar en predicciones futuras.

Modelo	GGAL	YPFD
1 (SARIMA)	43.9	7.1
2 (SARIMAX)	0.79	0.6
3 (SARIMAX con CP)	0.76	0.53
4 SARIMAX con CP, e interacción de variables	1.25	0.47

Tabla 6.1

Los gráficos de velas japonesas son una herramienta visual para representar el movimiento del precio de una acción en un periodo determinado. Cada "vela" muestra cuatro puntos clave: el precio de apertura, el precio de cierre, y los precios más altos y bajos alcanzados durante ese periodo (imagen 6.1). En los días en donde el precio de la acción subió, las velas se colorean de verde, y en estos casos el precio de cierre (que es el que predicen los modelos ajustados) se puede observar en el extremo superior de la caja que conforma el cuerpo de la vela. Cuando el precio de la acción baja, la vela se colorea de rojo. Y en este caso, el precio de cierre se ubica en el extremo inferior de la caja.

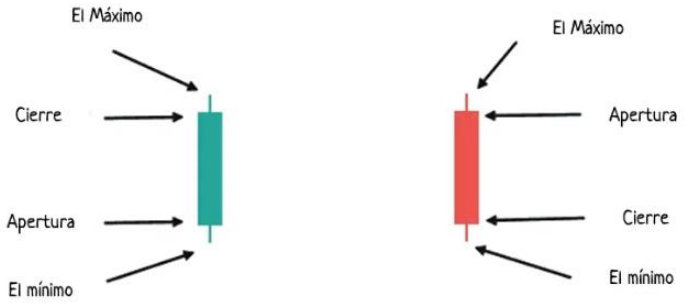

Imagen 6.1

Se ha superpuesto a los gráficos de velas japonesas para ambas empresas las estimaciones de los modelos propuestos, donde los puntos azules representan las predicciones. Esta visualización permite comparar de manera clara cómo las estimaciones se alinean con el comportamiento real del mercado, ofreciendo una representación tangible de la efectividad del modelo en la predicción de los precios de ambos activos, lo que contribuye a la validación de la metodología aplicada.

En la imagen 6.2 se observan las estimaciones sobre Galicia, en donde se puede apreciar que las estimaciones se ubican sobre el precio de cierre en todas las observaciones, excepto en la última observación. Por su parte, en la imagen 6.3 se observan las estimaciones sobre YPF. En este caso, aunque la variabilidad es levemente mayor, se observa como las estimaciones se encuentran en el rango de cotización del precio en todas las observaciones.

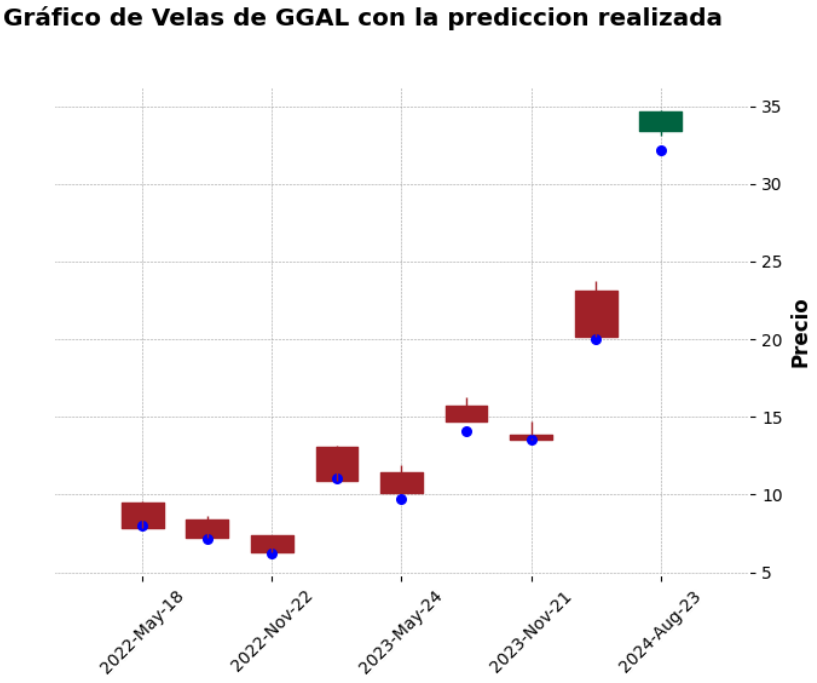

Imagen 6.2

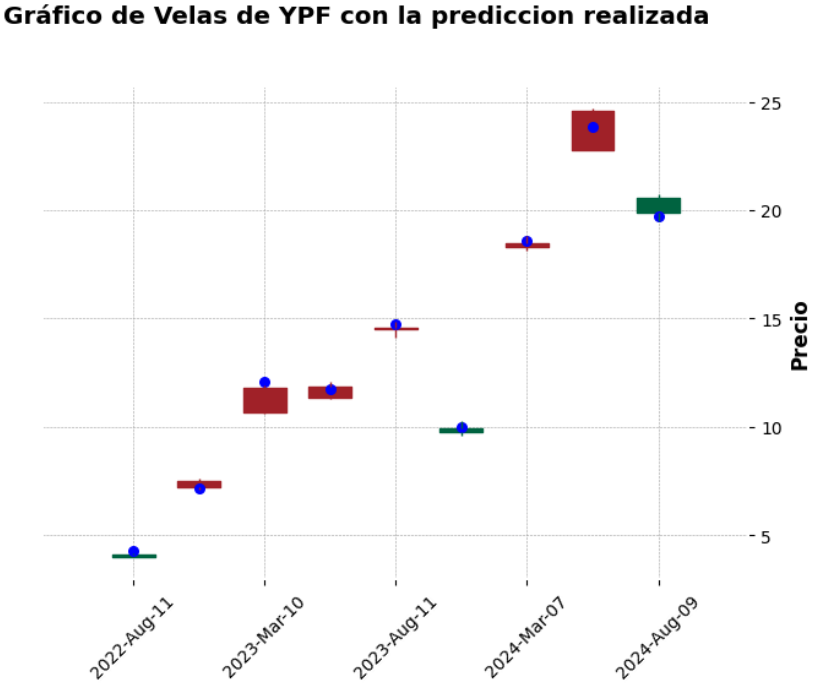

Imagen 6.3

6.1 Selección de variables

Para evaluar la contribución de las variables exógenas en la predicción de la variable objetivo se emplearon técnicas de selección de variables. La selección hacia adelante (*forward selection*) permite incorporar progresivamente variables según su impacto en la mejora del ajuste del modelo, evaluado con métricas como criterio de información de Akaike o criterio de información Bayesiano (BIC). De esta forma, se comienza con un modelo sin variables y, en cada iteración, se agrega aquella que mejore más el modelo. Este proceso continúa hasta que no se observe una mejora significativa.

De manera complementaria, la eliminación hacia atrás (*backward elimination*) parte de un modelo completo, e iterativamente elimina las variables menos significativas una a una, evaluando en cada paso si el ajuste del modelo mejora.

Ambas técnicas se aplicaron en los modelos propuestos para identificar las variables exógenas más relevantes, utilizando los parámetros según lo especificado en las secciones de modelos propuestos para Galicia e YPF.

En la tabla 6.1 se detallan las variables seleccionadas para Galicia a través de las dos técnicas de selección, junto con las variables que no fueron seleccionadas por ninguna de ellas. Las variables seleccionadas se presentan en orden decreciente de importancia, determinado por el impacto que tienen en el ajuste del modelo. De manera similar, la tabla 6.2 ofrece la misma información para YPF.

Selección hacia adelante	Selección hacia atrás	No seleccionadas
PX_TO_EBITDA	PX_TO_EBITDA	EBITDA_TO_REVENUE
IS_COMP_NET_INCOME	CF_NET_INC	IS_OPER_INC
SALES_REV_TURN	EARN_FOR_COM_TO_TOT_REV	

Tabla 6.1

Selección hacia adelante	Selección hacia atrás	No seleccionadas
IS_COMPARABLE_EBITDA	CF_NET_INC	SALES_REV_TURN
EBITDA_TO_REVENUE	IS_COMP_NET_INCOME	IS_COMP_SALES
	IS_OPER_INC	
	CF_FREE_CASH_FLOW	
	IS_COMPARABLE_EBIT	

Tabla 6.2

Los resultados muestran que las variables contables más importantes para predecir los resultados de Galicia e YPF son diferentes, reflejando la naturaleza única de

cada empresa, pero también presentan algunas similitudes.

En Galicia, las variables PX_TO_EBITDA, IS_COMP_NET_INCOME y SALES_REV_TURN se destacan en la selección hacia adelante, mientras que la selección hacia atrás introduce CF_NET_INC y EARN_FOR_COM_TO_TOT_REV. La repetición de PX_TO_EBITDA en ambas selecciones resalta su relevancia crítica como predictor para Galicia, alineada con su enfoque en la eficiencia y la rentabilidad operativa.

Para YPF, se observa una mayor dispersión de variables entre los métodos de selección, siendo más compleja su estructura contable. En la selección hacia adelante, solo dos variables, IS_COMPARABLE_EBITDA y EBITDA_TO_REVENUE, se destacaron, mientras que el método hacia atrás revela una mayor dependencia de variables relacionadas con los flujos de caja, tales como CF_NET_INC y CF_FREE_CASH_FLOW, además de incorporar la variable IS_OPER_INC.

El hecho de que en Galicia e YPF los indicadores relacionados con el flujo de caja sean cruciales en la selección hacia atrás sugiere la importancia de la liquidez y la generación de efectivo en la estabilidad financiera y predictibilidad de estas empresas, lo que es coherente con la necesidad de mantener una sólida posición financiera, sobre todo en economías fluctuantes.

6.2 Contribución de las variables.

Se desarrolló un modelo de clasificación con el objetivo de analizar la contribución de las variables explicativas que provienen de los estados contables en la predicción de la variable objetivo, la cual se definió en función del comportamiento del precio de la acción el día en que la empresa presentó los estados contables. En este contexto, la variable toma el valor de 1 si el precio subió y 0 si bajó. Los conjuntos de datos se estructuraron utilizando la variable dicotómica "variación" como objetivo, mientras que las variables explicativas provienen de los estados contables de la empresa.

La regresión logística es una técnica estadística ampliamente utilizada en contextos donde la variable dependiente es categórica, particularmente cuando se trata de una variable dicotómica (que toma solo dos posibles valores). A diferencia de la regresión lineal, que busca predecir un valor continuo, la regresión logística modela la probabilidad de que un evento ocurra (en este caso, si el precio de una acción sube o baja) en función de las variables explicativas.

Este modelo se basa en la función logística, que garantiza que la probabilidad estimada se encuentre entre 0 y 1. Específicamente, la regresión logística asume una relación logarítmica entre la razón de probabilidades (*odds*) y las variables explicativas. Esta relación se expresa mediante la siguiente ecuación:

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 * X_1 + .. + \beta_n * X_n \qquad (6.1)$$

donde p es la probabilidad de que ocurra el evento (en este caso, que el precio de la acción suba), X_1, X_2, … , X_n son las variables explicativas y β_0, β_1, … , β_n son los coeficientes que representan la magnitud y dirección del efecto de cada variable sobre la probabilidad del evento.

Este enfoque es particularmente útil para evaluar el impacto de cada uno de los campos de los estados contables en la probabilidad de que el precio de una acción suba o baje luego de la presentación de estados contables. Al analizar los coeficientes estimados, podemos identificar qué variables tienen un mayor peso en la predicción, y así evaluar si todas las variables explicativas aportan de manera similar al modelo.

Además, la regresión logística tiene la ventaja de ser robusta a la naturaleza categórica de la variable dependiente, lo que la convierte en una opción ideal para problemas de clasificación binaria en series temporales financieras, como el presente análisis (Hosmer et al., 2013; Agresti, 2015).

Al ajustar un modelo de clasificación bivariado utilizando regresión logística, es

interesante analizar cuál es el aporte de cada variable explicativa en el modelo. En el análisis de la explicabilidad del modelo de regresión logística ajustado, se ha empleado la librería *SHapley Additive exPlanations* (SHAP), una herramienta que permite interpretar de manera clara el impacto de cada variable explicativa en la predicción del modelo.

La motivación detrás de utilizar SHAP es entender no solo qué variables contribuyen al modelo, sino también cómo y en qué magnitud lo hacen. Según su documentación, SHAP se basa en los valores de Shapley, un concepto proveniente de la teoría de juegos cooperativos. En este marco, los valores de Shapley miden la contribución de cada jugador (en este caso, cada variable explicativa) al resultado del juego (la predicción del modelo). En otras palabras, SHAP descompone la predicción del modelo en contribuciones individuales de cada variable, asignando una "cuota justa" del resultado total a cada una de ellas. Este enfoque tiene la ventaja de ser consistente y aditivo, lo que significa que las contribuciones de todas las variables se suman para igualar la predicción total del modelo.

Los resultados se presentan en dos gráficos. El Gráfico de barras SHAP resume el impacto medio absoluto de cada variable en el modelo. Muestra cuáles son las variables más importantes en términos de su contribución promedio a la predicción. Las variables se ordenan de mayor a menor impacto, lo que permite identificar cuáles tienen mayor relevancia en la estimación de la variable objetivo. En este caso, permite determinar qué campos de los estados contables influyen más en la clasificación de si el precio de la acción subió o bajó. Por su parte, el gráfico de beeswarm proporciona una representación más detallada. Para cada variable, se visualiza cómo sus diferentes valores (representados por puntos) afectan las predicciones del modelo. Los puntos de colores indican si el valor de la variable es alto o bajo en cada observación, y la posición horizontal de los puntos muestra cómo ese valor específico impacta en la predicción. Si el punto se encuentra hacia la derecha (valores positivos de SHAP), significa que ese valor específico está aumentando la probabilidad de que el precio de la acción suba (variable objetivo = 1), mientras que, si está hacia la izquierda, disminuye dicha probabilidad.

6.2.1 Variables del modelo Galicia.

En el grafico 6.4 se observa el gráfico beeswarm para la empresa Galicia. En este gráfico, cada punto representa una observación y está posicionado a lo largo del eje horizontal según su valor SHAP, que indica el impacto de esa variable en la predicción del modelo. A partir de aquí es interesante interpretar las variables que

sobresalen.

La variable que más aporta en el modelo es IS_COMP_NET_INCOME. Se observa que los valores altos de esta variable (puntos rojos) tienden a estar a la derecha del gráfico (valores SHAP positivos), lo que significa que cuando el ingreso neto consolidado es alto, aumenta la probabilidad de que el precio de la acción suba.

Por el contrario, los valores bajos (puntos azules) se distribuyen hacia la izquierda, indicando que valores bajos de esta variable están asociados con una mayor probabilidad de que el precio de la acción baje.

Los valores bajos de PX_TO_EBITDA (azul) tienden a impactar negativamente en la predicción (hacia la izquierda), mientras que los valores altos (rojo) no tienen un efecto tan claro en una dirección particular. Esto sugiere que un bajo PX/EBITDA puede ser más decisivo para predecir una caída en el precio de la acción.

Con respecto a CF_NET_INC, los valores bajos (azules) de esta variable también están más distribuidos hacia la izquierda, indicando que cuando los flujos de caja netos son bajos, la predicción del modelo favorece una caída del precio de la acción.

La variable EARN_FOR_COM_TO_TOT_REV tiene una dispersión equilibrada en ambos lados, lo que sugiere que esta variable tiene una influencia mixta en las predicciones, sin un claro sesgo hacia el aumento o la caída del precio.

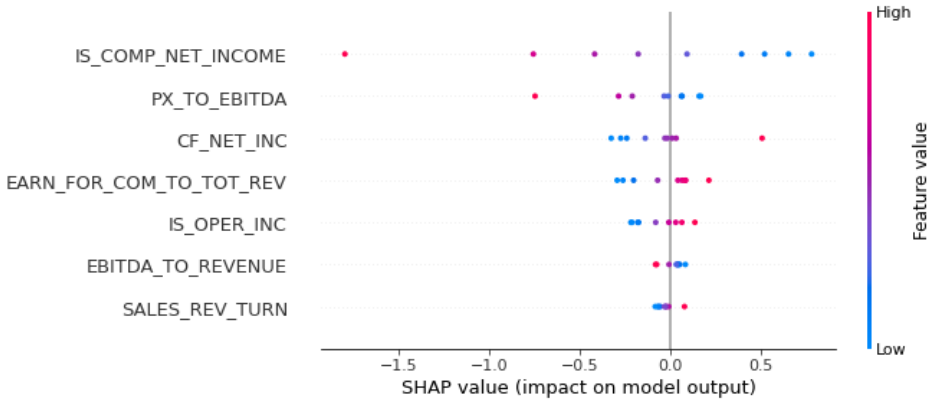

Imagen 6.4

El gráfico de barras SHAP (imagen 6.5) visualiza el impacto promedio de cada variable en las predicciones del modelo de regresión logística. El eje horizontal muestra el valor promedio absoluto de los valores SHAP para cada variable. Este valor refleja la magnitud del impacto que cada variable tiene en las predicciones, sin tener en cuenta si el efecto es positivo o negativo. Las variables que tienen un valor promedio SHAP más alto son las que tienen un mayor efecto sobre las predicciones del modelo. Esto indica qué variables explicativas contribuyen más a la estimación de si el precio de la acción subirá o bajará.

La variable que tiene mayor impacto en el modelo es IS_COMP_NET_INCOME, con un valor promedio SHAP de 0.62, lo que indica que es la variable más influyente en la predicción de si el precio de la acción sube o baja. Esto es coherente con el gráfico beeswarm, donde IS_COMP_NET_INCOME mostraba una dispersión clara de los valores SHAP, con una tendencia fuerte a influir en las predicciones positivas (subida del precio).

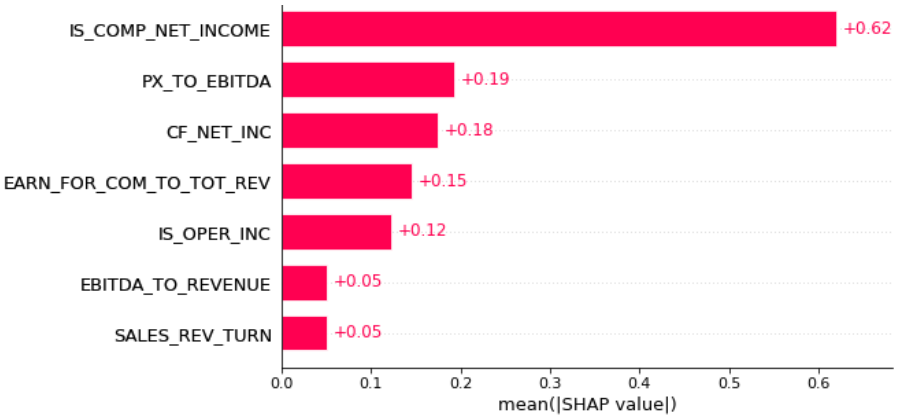

Imagen 6.5

6.2.2 Variables del modelo YPF.

El gráfico de barras SHAP revela las características que, en promedio, más influyen en las predicciones, proporcionando una visión general de la importancia de las variables. Para el caso de YPF (imagen 6.6), IS_COMP_SALES es la variable con mayor impacto, con un valor medio SHAP de +0.45. Esto indica que es la característica más relevante para explicar las predicciones del modelo. Le sigue la variable EBITDA_TO_REVENUE (+0.29), siendo también una característica importante. Y luego, CF_NET_INC y IS_COMP_NET_INCOME, las cuales tienen impactos similares, ambos alrededor de +0.19 y +0.15 respectivamente.

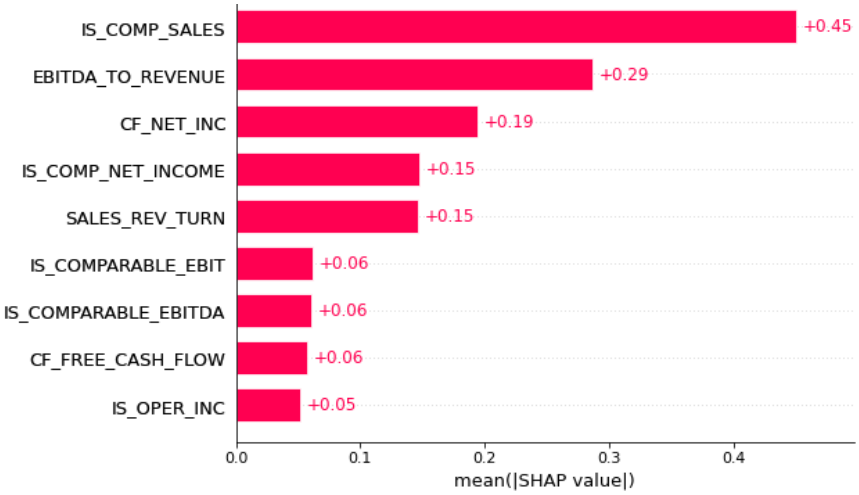

Imagen 6.6

El gráfico beeswarm (imagen 6.7) muestra el impacto individual de cada observación para las variables en el modelo. Se observa que IS_COMP_SALES tiene un gran impacto en ambos extremos del espectro: observaciones con valores altos (rojos) tienden a aumentar la probabilidad de que el precio de la acción suba, mientras que los valores bajos (azules) disminuyen esta probabilidad. EBITDA_TO_REVENUE sigue un patrón similar, donde un mayor EBITDA relativo a los ingresos tiende a asociarse con subas en el precio de la acción. Para variables como CF_NET_INC y IS_COMP_NET_INCOME, observamos que tanto valores altos como bajos influyen de manera más equilibrada en las predicciones, aunque los valores altos tienden a estar más asociados con aumentos en el precio de la acción.

Ambos gráficos coinciden en señalar que IS_COMP_SALES y EBITDA_TO_REVENUE son las dos variables más importantes para las predicciones del modelo, pero el gráfico beeswarm revela que incluso dentro de esas variables, su impacto puede variar significativamente según los valores observados.

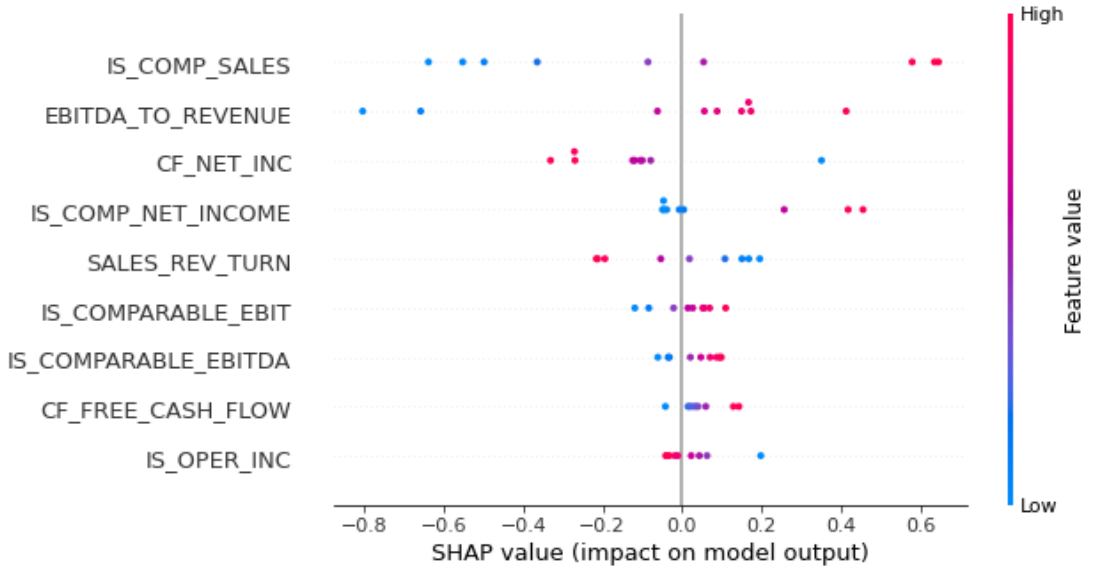

Imagen 6.7

6.3 Precisión de las estimaciones de los analistas.

Para determinar si los analistas aciertan en sus estimaciones de los resultados de los distintos campos de los estados contables, se realiza un análisis basado en la mediana. Se opta por este procedimiento robusto debido a la poca cantidad de observaciones. Siguiendo esta línea, se aplica la prueba de signo de Wilcoxon con el objetivo de comparar los valores observados en los estados contables con las predicciones de los analistas.

Esta prueba no paramétrica resulta adecuada cuando el tamaño de la muestra es pequeño y no se puede asumir normalidad en la distribución de los datos, como ocurre en este caso, donde contamos con 43 observaciones. La hipótesis nula de la prueba es que la mediana de las diferencias entre los valores reales y los estimados es cero, lo cual indicaría que los analistas aciertan en sus predicciones de manera consistente.

Según se expresó en la sección metodología donde se especifica en detalle cada una de las variables utilizadas, aquellas que provienen de los estados contables se definen a partir de la relación entre el valor real y el estimado.

Sea la desviación relativa para la observación i:

$$D_i = \frac{Valor\_real_i}{Valor\_estimado_i} - 1 \qquad (6.2)$$

Siguiendo el procedimiento propuesto por Corso y Jimmy (2005), la hipótesis nula es que la mediana de las diferencias es cero. Esto implica que no hay diferencias sistemáticas entre los valores reales y los estimados. Y la hipótesis alternativa es que la mediana de las diferencias es distinta de cero, lo que indicaría un sesgo en las predicciones de los analistas.

Se ordenan las diferencias Di en valor absoluto y se asignan rangos a las mismas. Los signos positivos y negativos de las diferencias originales se mantienen para el cálculo de los rangos con signo. Luego, el estadístico de la prueba de Wilcoxon se obtiene sumando los rangos correspondientes a los valores positivos y negativos de D_i por separado. El estadístico W es el menor de las dos sumas de rangos:

$$W = \min\left(\sum_{Di>0} R_i \ , \sum_{Di<0} R_i\right) \qquad (6.3)$$

donde R_i es el rango asignado a la observación i.

El *p value* se calcula comparando el estadístico W observado con su distribución bajo la hipótesis nula. Un *p value* menor a 0.05 indicaría que se rechaza H_0 a favor de H_A, lo cual sugeriría que las diferencias entre los valores reales y estimados no son debidas al azar.

En la tabla 6.3 y 6.4 se muestran los resultados para la empresa Galicia e YPF

respectivamente.

Variables	p_value
PX_TO_EBITDA	9.09E-13
EBITDA_TO_REVENUE	0.004
CF_NET_INC	6.45E-07
IS_COMP_NET_INCOME	1.15E-09
SALES_REV_TURN	9.09E-13
IS_OPER_INC	1.53E-08
EARN_FOR_COM_TO_TOT_REV	9.09E-13

Tabla 6.3

En el caso de Banco Galicia (tabla 6.3), los resultados muestran *p values* extremadamente pequeños (menores a 0.05) por lo que se rechaza la hipótesis nula para todas las variables, indicando que la mediana es significativamente diferente de 0. Esto sugiere que los analistas no están acertando con sus predicciones para ninguna variable de forma consistente. Los resultados son estadísticamente significativos, lo que implica que existen diferencias sistemáticas entre las predicciones de los analistas y los valores reales en estas variables.

Variables	p_value
EBITDA_TO_REVENUE	0.0274
CF_NET_INC	0.1245
IS_COMP_NET_INCOME	0.0005
SALES_REV_TURN	0.0000
IS_OPER_INC	0.2939
CF_FREE_CASH_FLOW	0.0197
IS_COMP_SALES	0.0001
IS_COMPARABLE_EBITDA	0.0001
IS_COMPARABLE_EBIT	0.9001

Tabla 6.4

Para la empresa YPF (tabla 6.4), los resultados son más variados. Las variables EBITDA_TO_REVENUE, IS_COMP_NET_INCOME, SALES_REV_TURN, CF_FREE_CASH_FLOW, IS_COMP_SALES, IS_COMPARABLE_EBITDA, y

IS_COMPARABLE_EBIT presentan *p values* muy bajos, indicando nuevamente que la mediana es significativamente diferente de 0. Esto sugiere que los analistas no aciertan en sus predicciones para estas variables de manera precisa y constante. En cambio, para las variables CF_NET_INC, IS_OPER_INC, y IS_COMPARABLE_EBIT, los *p values* son mayores a 0.05, lo que implica que no se rechaza la hipótesis nula. Esto sugiere que, sólo para estas variables, no se puede concluir que las predicciones de los analistas difieran significativamente de los valores reales.

7. Conclusiones

El objetivo principal de esta tesis fue predecir cuál iba a ser el precio de una acción el día posterior al de la presentación de los estados contables. Para ello, se emplearon modelos estadísticos combinando variables financieras provenientes de los estados contables con indicadores macroeconómicos relevantes para el mercado bursátil; y se aplicaron a las empresas Galicia e YPF.

Para ambas empresas se ajustaron múltiples modelos SARIMAX. Se implementó un exhaustivo proceso de ingeniería de características, lo que permitió mejorar significativamente el rendimiento de los modelos. Este proceso incluyó la imputación de valores atípicos mediante técnicas de *clustering*, la aplicación de análisis de componentes principales, y la incorporación de variables de interacción. Los modelos finales lograron una performance destacada. Para Banco Galicia, la raíz del error cuadrático medio (RMSE) fue de 0,76, mientras que para YPF fue de 0,47. Esto último se traduce en que el error promedio para YPF equivale a 47 centavos, o aproximadamente un 2% de su precio, lo que refleja una capacidad predictiva notable en ambos casos.

Los modelos SARIMAX ajustados para cada empresa difieren en la selección de variables, lo que refleja las particularidades de cada sector. No obstante, tanto para Banco Galicia como para YPF, el flujo de caja neto y el EBITDA emergen como variables clave en la predicción del precio de las acciones. Estas métricas, fundamentales para evaluar la rentabilidad y la capacidad operativa de una empresa, están estrechamente ligadas con la salud financiera y la eficiencia en la gestión operativa. A pesar de las diferencias en la naturaleza de sus operaciones, estos resultados sugieren que existen indicadores financieros universales que son esenciales para analizar el desempeño corporativo, independientemente del sector al que pertenezcan las empresas, ya sea financiero o energético.

El análisis con SHAP proporcionó una interpretación detallada del modelo de regresión logística ajustado, permitiendo identificar qué variables de los estados contables tienen mayor impacto en la predicción de si el precio de la acción sube o baja. Esta evaluación no solo facilita la identificación de las variables explicativas más relevantes, sino que también ofrece una mayor comprensión del comportamiento del modelo y su relación con el fenómeno estudiado. Este conocimiento es fundamental para mejorar la toma de decisiones basada en modelos predictivos y para optimizar las estrategias de inversión.

En el caso de Banco Galicia, las variables más influyentes en el modelo resultaron ser IS_COMP_NET_INCOME, PX_TO_EBITDA y CF_NET_INC, todas ellas

directamente relacionadas con el desempeño financiero. Por su parte, en el modelo ajustado para YPF, las variables clave fueron IS_COMP_SALES y EBITDA_TO_REVENUE, lo que subraya la relevancia de los ingresos y la eficiencia operativa en la predicción del comportamiento del precio de la acción tras la presentación de los estados contables. Esto evidencia la importancia de ciertos indicadores financieros en la evaluación del desempeño corporativo, adaptados al contexto de cada empresa.

Un aspecto destacado de los resultados es que, en la mayoría de las variables analizadas, los analistas no logran acertar de manera consistente en sus estimaciones, ya que las medianas de las diferencias de las variables construidas a partir de los estados contables no son iguales a 0. Esto es particularmente evidente en el caso de Banco Galicia, donde todas las variables presentan diferencias estadísticamente significativas. En contraste, para YPF, los analistas parecen realizar leves mejores predicciones en ciertos aspectos contables, lo que sugiere una mayor precisión en sus estimaciones. Esta diferencia podría explicarse por el mayor volumen de operaciones de YPF, lo que podría influir en la cantidad de recursos y atención que los analistas dedican a seguir esta empresa.

Algunas de las limitaciones del modelo ajustado en esta tesis residen en que los modelos SARIMAX, si bien permiten incorporar factores externos, asumen relaciones lineales entre las variables. Esta suposición podría restringir su capacidad para capturar dinámicas más complejas en los datos financieros. Además, el tamaño reducido del conjunto de observaciones (43 estados contables) puede afectar la robustez del modelo y su capacidad para generalizar a otros períodos o empresas.

Si bien se incluyeron variables macroeconómicas clave, como la tasa de interés de los bonos del Tesoro de EE. UU. y el valor del S&P 500, existen otros factores externos que podrían influir en los precios de las acciones y no fueron considerados en el modelo. Incorporar más indicadores económicos o variables sectoriales podría mejorar la capacidad predictiva. Además, el análisis se centró solo en dos empresas (YPF y Grupo Financiero Galicia); replicar este estudio en un conjunto más amplio de compañías permitiría obtener resultados más robustos y generalizables, fortaleciendo la validez de las conclusiones.

En futuras investigaciones, la implementación de métodos de aprendizaje automático, como redes neuronales, podría abordar este desafío al modelar dinámicas no lineales y patrones más complejos en los datos financieros. Estas técnicas, al requerir grandes volúmenes de información, serían más efectivas si se

amplía el conjunto de datos, lo que permitiría mejorar la capacidad predictiva del modelo y captar mejor las interacciones entre variables exógenas y el comportamiento del mercado.

Así mismo, se podrían explorar otras técnicas estadísticas, como modelos GARCH. Estos modelos son especialmente útiles para capturar la heterocedasticidad presente en las series temporales financieras, donde la varianza no es constante a lo largo del tiempo. Al aplicar un modelo GARCH, se puede modelar y predecir no solo el valor esperado del precio de la acción, sino también su volatilidad. Esto permitiría obtener estimaciones más precisas del riesgo asociado a las predicciones, proporcionando una visión más completa para la toma de decisiones y complementando la capacidad de los modelos SARIMAX en el análisis de dinámicas temporales y factores exógenos.

El abordar el problema desde múltiples enfoques y utilizando técnicas de estadística tradicional y moderna permitiría compararlas entre sí para identificar fortalezas y debilidades relativas de cada enfoque, determinando cuál se adapta mejor a las particularidades del mercado financiero analizado. La combinación de diferentes métodos también podría dar lugar a modelos híbridos, que integren las ventajas de varias técnicas para mejorar la precisión en la predicción de los precios de las acciones.

Una línea prometedora de investigación futura consiste en integrar datos no estructurados provenientes de redes sociales y portales de noticias financieros. Estos recursos pueden proporcionar *insights* valiosos sobre las expectativas del mercado respecto a los estados contables antes de su publicación. El análisis de sentimiento de publicaciones en redes como Twitter o foros especializados podría capturar percepciones y reacciones anticipadas de los inversores. Al incluir esta información en los modelos predictivos, se podrían mejorar las estimaciones, ya que se considerarían no solo los datos históricos y contables, sino también las expectativas del mercado, ofreciendo un enfoque más integral y adaptado a la dinámica actual de la información financiera.

8. Referencias.

8.1 Libros.

Agresti, A. 2015. *Foundations of Linear and Generalized Linear Models*. John Wiley & Sons.

Alpaydın, E. 2004. *Introduction to Machine Learning*. MIT Press.

Graham, B. y Dodd, D. 2004. *Security Analysis.* McGraw-Hill.

Hosmer, D. W., Lemeshow, S., y Sturdivant, R. X. 2013. *Applied Logistic Regression*. John Wiley & Sons.

MacKay, D. 2003. *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press.

8.2 Papers.

Adebiyi, A. O., Adewumi, A. A., y Ayo, C. K. 2014. Stock price prediction using the arima model. *AMSS 16th International Conference on Computer Modeling and Simulation*: 105–111.

Agrawal, J., Chourasia, V., y Mittra, V. 2013. State-of-the-Art in Stock Prediction Techniques. *International Journal of Advanced Research in Electrical, Electronics and Instrumentation Energy*.

Ariyo, A., Aderemi, O., y Ayo, C. 2014. Stock Price Prediction Using the ARIMA Model. *16th International Conference on Computer Modelling and Simulation.*

Arkan, T. 2016. The Importance of Financial Ratios in Predicting Stock Price Trends: A Case Study in Emerging Markets. *Zeszyty Naukowe Uniwersytetu Szczecińskiego Finanse Rynki Finansowe Ubezpieczenia.*

Atsalakis, G., y Valavanis, K. 2009. Surveying stock market forecasting techniques – Part II: Soft computing methods. *Expert Systems with Applications*.

Black, A., Wright, P., y Davies, J. 2001. In Search of Shareholder Value: Managing

the Drivers of Performance. *Financial Times/Prentice Hall*.

Bhushan, R. 1989. Firm characteristics and analyst following, *Journal of Accounting and Economics*: 11(2-3), 255-274.

Boozer, B., Rainwater, L., y Lowe, S. 2017. Using financial statement variables to predict stock prices: Lessons from the 2007-2009 financial crisis. *Journal of Finance and Accountancy:* Vol. 21, 1-10.

Brennan, M. J., y Subrahmanyam, A., 1995. Investment analysis and price formation in securities markets- *Journal of Financial Economics:* 38(3), 361-381.

Cakra, Y., y Distiawan, T. 2015. Stock price prediction using linear regression based on sentiment analysis. *2015 International Conference on Advanced Computer Science and Information Systems (ICACSIS)*.

Cakra, Y., Trisedya, B. 2023. Stock price prediction using linear regression based on sentiment analysis. *11th International Conference on Cyber and IT Service Management (CITSM).*

Cantemir, D. 2013. Does Fundamental Analysis Predict Stock Returns? Evidence from Non-Financial Companies Listed on KSE. *Knowledge Horizons*: Vol 5

Clubb, C., y Naffi, M. 2007. The Usefulness of Book-to-Market and ROE Expectations for Explaining UK Stock Returns. *Journal of Business Finance and Accounting*.

Corso S. y Jimmy, A. 2005. Estadística no paramétrica: métodos basados en rangos. *Universidad Nacional de Colombia - Facultad de Ciencias - Departamento de Estadística.*

Damodaran, A. 2012. *Investment Valuation: Tools and Techniques for Determining the Value of Any Asset.* 3rd ed. New York: Wiley.

Derakhshan A. y Beigy H. 2019. Sentiment analysis on stock social media for stock price movement prediction. *Engineering Applications of Artificial Intelligence. Vol 85*: 569-578.

Dzikevičius, A., y Šaranda, S. 2011. Can financial ratios help to forecast stock

prices?, *Journal of Security and Sustainability Issues*.

Fama, E., 1965. The behavior of stock market prices. *Journal of Finance.* 34–105.

Fama, E., 1970. Efficient Capital Markets: A Review of Theory and Empirical Work. *Journal of Finance.* 383-417

Gong, J., y Sun, S. 2009. A New Approach of Stock Price Prediction Based on Logistic Regression Model. *International Conference on New Trends in Information and Service Science (NISS)*

Green, S., 2011. Time series analysis of stock prices using the box-jenkins approach. *Electronic Theses and Dissertations*.

Gupta, R., Chen, M. 2022. Sentiment Analysis for stock price prediction. *2020 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR).*

Heo, J., y Yang, J. 2016. Stock Price Prediction Based on Financial Statements Using SVM. *International Journal of Hybrid Information Technology*.

Holthausen, R., y Larcker, D. 1992. The prediction of stock returns using financial statement information. *Journal of Accounting and Economics 15*.

Hu, Z., Zhao, Y. 2020. A Survey of Forex and Stock Price Prediction Using Deep Learning. *Aplied system innovation.*

Imran, K. 2018. Prediction of stock performance by using logistic regression model: evidence from Pakistan Stock Exchange (PSX). *Asian Journal of Empirical Research.*

Islam, M., y Nguyen, N. 2020. Comparison of Financial Models for Stock Price Prediction. *Journal of Risk and Financial Management*.

Jarrett, J., y Kyper, E. 2011. ARIMA Modeling with Intervention to Forecast and Analyze Chinese Stock Prices. *International Journal of Engineering Business Management.*

Jin Z., Yang, Y. y Liu, Y. 2019. Stock closing price prediction based on sentiment

analysis and LSTM. *Neural Comput Appl:*1-17.

Kaplan, R. S., y Norton, D. P. 1996. *The Balanced Scorecard: Translating Strategy into Action*. Harvard Business School Press.

Kumar, M. 2021. Forecasting stock market prices using mixed ARIMA model: a case study of Indian pharmaceutical companies. *Investment Management and Financial Innovations*.

Lin, X. Yang, Z., Song, Y. 2009, Short-term stock price prediction based on echo state networks. *Expert Systems with Applications*: Vol 36, pag. 7313-7317.

Lu, W., Li, J. y Qui, L. 2020. A CNN-BiLSTM-AM method for stock price prediction. *Neural computing and applications*. Springer.

Min, J. 2020. Financial Market Trend Forecasting and Performance Analysis Using LSTM. *ArXiv*.

Mondal, P., Shit, L. y Goswami, S. 2014. Study of effectiveness of time series modeling (arima) in forecasting stock prices. *International Journal of Computer Science, Engineering and Applications (IJCSEA)*. 4:13–29.

Ou, J. y Penman, S. 1989. Financial statement analysis and the prediction of stock returns. *Journal of Accounting and Economics*.

Pearce, D. K. 1983. Stock prices and the economy. *Federal Reserve Bank of Kansas City Economic Review*. pages 7–22.

Penman, Stephen H. 2010. *Financial Statement Analysis and Security Valuation.* 5th ed. New York: McGraw-Hill Education.

Pettenuzzo, D., y Timmermann, A. 2017. *Predicting Recessions and Stock Market Performance: International Evidence*. Review of Economics and Statistics, 99(4), 720-733.

Sharaf, M. 2021. StockPred: a framework for stock Price prediction. *Multimedia Tools and Applications.*

Siew, H., y Nordin, J. 2012. Regression techniques for the prediction of stock price trend. *International Conference on Statistics in Science, Business and Engineering*

(ICSSBE2012).

Sotiris, F. 2021. Financial Forecasting based on Fundamental Analysis with Machine Learning. *Aristotle University of Thessaloniki*.

Xu, S. Y. 2012. Stock price foreasting using information from yahoo finance and google trend. *Berkeley University*.

Yen, G. y Lee, C. 2008. Efficient Market Hypothesis (EMH): Past, Present and Future, Review of Pacific Basin Financial Markets and Policies. vol 11, issue 2.

Yetginer, B. 2017. Forecasting BIST-100 price index. *Middle East Techincal University*.

Zhang, J., Shan, R., y Su, W. 2009. *Applying Time Series Analysis Builds Stock Price Forecast Model.* Modern Applied Science.

8.3 Páginas web.

Sitio de SHAP. https://shap.readthedocs.io/en/latest. Acceso: 23/10/2024

Sitio oficial de Bolsas y Mercados de Argentina. https://www.byma.com.ar. Acceso: 03/09/2024.

Sitio oficial de Galicia SA. https://www.gfgsa.com. Acceso: 04/10/2024.

Sitio oficial de la Comisión Nacional de Valores. https://cnv.gov.ar. Acceso: 03/09/2024.

Sitio oficial de YPF SA. https://www.ypf.com. Acceso: 04/10/2024.

Sitio oficial del Banco Central de la República Argentina. https://www.bcra.gob.ar. Acceso: 03/09/2024.

Texto ordenado de las normas de la CNV basado en la reglamentación de las leyes 27.440 y 26.831.
https://servicios.infoleg.gob.ar/infolegInternet/verNorma.do;jsessionid=AD418DF9B8AAED8DEC9BA0861C048E71?id=219405. Acceso: 05/09/2024.

9. Anexos.

9.1 Dataset de GGAL

fecha	YIELD10	SPY_pre_eecc	GGAL_pre_eecc	GGAL_post_eecc	PX_TO_EBITDA	EBITDA_TO_REVENUE	CF_NET_INC	IS_COMP_NET_INCOME	SALES_REV_TURN	IS_OPER_INC	EARN_FOR_COM_TO_TOT_REV
23/514	2.53	190.35	12.64	12.26	-0.78	1.42	0.63	0.79	1.68	0.55	-0.39
7/814	2.41	192.07	14.15	14.68	-0.79	1.48	0.38	0.54	1.48	0.57	-0.44
5/1114	2.34	201.07	14.88	15.26	-0.79	1.63	0.21	0.36	1.17	0.20	-0.44
12/215	1.99	206.93	18.00	18.05	-0.79	1.66	0.04	0.18	0.93	-0.03	-0.46
7/515	2.18	208.04	20.99	20.80	-0.77	1.52	0.65	0.76	1.98	0.00	-0.45
12/815	2.15	208.67	21.01	20.74	-0.75	1.34	0.65	0.70	1.88	0.15	-0.43
10/1115	2.34	208.08	25.47	26.12	-0.74	1.33	0.13	0.16	1.72	-0.08	-0.35
12/216	1.75	182.86	26.90	27.19	-0.75	1.29	0.64	0.63	1.70	-0.01	-0.39
10/516	1.76	205.89	28.93	29.19	-0.76	1.43	0.48	0.47	1.63	0.29	-0.39
9/816	1.55	218.18	29.86	29.64	-0.78	1.46	0.33	0.32	1.34	-0.09	-0.43
9/1116	2.06	214.11	30.93	29.91	-0.78	1.50	0.14	0.17	1.13	-0.14	-0.47
14/217	2.47	232.77	34.49	34.70	-0.78	1.56	0.11	0.09	0.84	-0.29	-0.41
9/517	2.40	239.44	41.83	42.05	-0.70	0.94	-0.02	0.03	0.76	-0.06	-0.45
8/817	2.26	247.26	37.13	36.53	-0.63	0.58	0.19	0.18	0.97	0.05	-0.39
9/1117	2.34	259.11	53.71	54.32	-0.56	0.39	0.16	0.11	0.79	0.30	-0.35
8/218	2.83	267.67	63.03	60.29	-0.54	0.36	0.21	0.22	0.89	0.24	-0.36
24/518	2.98	272.80	43.49	43.66	-0.43	0.11	0.45	0.33	1.23	1.64	-0.30
16/818	2.87	284.06	31.04	29.29	-0.35	-0.06	0.66	0.57	1.60	1.13	-0.36
27/1218	2.77	248.07	26.34	26.92	-0.41	-0.05	0.82	0.82	1.94	0.51	-0.39
7/319	2.64	277.33	26.27	25.60	-0.42	-0.02	0.52	0.59	1.47	0.26	-0.41
9/519	2.44	286.66	25.55	24.72	-0.62	0.10	2.38	0.66	2.27	2.02	0.06
12/819	1.65	288.07	16.75	17.23	-0.69	0.22	1.55	0.61	2.05	2.28	-0.16
11/1119	1.94	308.94	12.41	12.72	-0.72	0.27	3.98	0.92	3.07	2.28	0.23
20/220	1.52	338.34	14.09	14.31	-0.77	0.37	1.28	0.51	2.99	2.71	-0.43
8/620	0.88	323.20	11.50	10.77	-0.69	0.08	1.94	0.33	2.75	2.40	-0.22
26/820	0.69	344.12	10.17	10.04	-0.61	0.00	0.59	0.11	3.01	1.42	-0.60
24/1120	0.88	357.46	8.31	8.82	-0.57	-0.07	0.44	-0.02	2.51	0.99	-0.58
9/321	1.53	381.72	7.13	7.19	-0.58	-0.11	0.42	0.30	2.52	0.91	-0.59
26/521	1.58	419.07	8.31	8.57	-0.61	-0.02	-0.12	-0.15	2.61	0.85	-0.76
26/821	1.35	446.26	10.05	10.23	-0.69	-0.17	-0.33	-0.01	3.53	0.93	-0.85
22/1121	1.63	468.89	9.58	9.15	-0.75	0.26	-0.32	0.01	2.80	1.03	-0.81
15/222	2.05	439.02	9.00	9.17	-0.84	0.38	0.04	0.28	3.98	1.65	-0.79
17/522	2.99	408.32	9.49	9.25	-0.74	0.04	0.12	-0.01	6.28	2.46	-0.74
23/8/22	3.05	412.35	8.46	8.55	-0.69	-0.08	0.44	-0.05	5.77	1.91	-0.68
22/11/22	3.76	394.59	7.28	7.33	-0.73	-0.51	0.40	0.21	9.13	1.95	-0.87
7/3/23	3.97	404.47	13.04	12.67	-0.85	-0.14	0.47	1.11	11.22	3.69	-0.88

Continuación.

fecha	YIELD10	SPY_pre_eecc	GGAL_pre_eecc	GGAL_post_eecc	PX_TO_EBITDA	EBITDA_TO_REVENUE	CF_NET_INC	IS_COMP_NET_INCOME	SALES_REV_TURN	IS_OPER_INC	EARN_FOR_COM_TO_TOT_REV
23/5/23	3.70	414.09	11.46	11.76	-0.67	-0.79	0.88	0.80	19.36	9.78	-0.92
12/9/23	4.28	448.45	15.74	15.97	-0.60	-0.79	1.22	0.43	22.85	7.00	-0.88
21/11/23	4.39	454.26	14.65	14.56	0.63	-0.55	-0.48	0.37	4.33	1.09	-0.76
4/3/24	4.21	512.85	22.29	21.71	-0.06	-0.56	-0.19	1.27	6.71	1.66	-0.89
22/8/24	3.85	556.22	33.12	34.68	-0.48	-0.09	-0.12	0.02	2.90	1.83	-0.77

9.2 Dataset de YPF

fecha	YIELD10	SPY_pre_eecc	YPF_pre_eecc	YPF_post_eecc	EBITDA_TO_REVENUE	CF_NET_INC	IS_COMP_NET_INCOME	SALES_REV_TURN	IS_OPER_INC	CF_FREE_CASH_FLOW	IS_COMP_SALES	IS_COMPARABL_EBITDA	IS_COMPARABLE_EBIT
09/08/24	3.94	530.65	19.78	20.58	-0.35	-2.61	-0.32	0.66	-1.85	0.50	0.66	-0.38	-2.65
10/05/24	4.50	520.17	24.38	22.78	-0.81	-2.62	-0.86	1.12	-4.07	0.48	1.12	-0.72	-3.81
06/03/24	4.10	509.75	18.67	18.32	-1.38	-6.09	-2.69	0.21	-16.58	-0.97	0.21	-1.45	-16.58
08/11/23	4.49	437.25	10.03	9.96	-0.52	-0.32	-0.06	0.51	-1.18	-0.21	0.51	0.51	-0.77
10/08/23	4.11	445.91	14.39	14.51	0.01	1.14	1.05	0.53	0.08	-3.10	0.53	0.39	0.44
11/05/23	3.39	412.13	11.75	11.36	0.09	2.45	1.35	0.52	0.68	3.38	0.52	0.37	0.68
09/03/23	3.91	391.56	11.56	10.65	-0.05	0.18	0.18	0.12	-0.00	-0.44	0.12	0.55	-0.00
09/11/22	4.10	374.13	7.12	7.19	-0.13	0.45	1.50	0.48	0.59	-0.11	0.43	0.13	0.49
10/08/22	2.79	419.99	3.91	4.09	-0.10	1.79	2.10	0.32	0.59	-0.01	0.28	0.45	0.45
11/05/22	2.93	392.75	4.06	4.04	0.09	-0.51	-0.41	0.15	-0.03	0.31	0.14	0.32	-0.17
03/03/22	1.84	435.71	4.55	4.34	-0.10	-1.03	-1.04	0.10	-0.41	0.01	0.06	-0.05	-0.41
09/11/21	1.44	467.38	4.56	4.40	0.28	-1.30	1.05	0.12	-5.26	0.20	0.12	0.84	-5.26
10/08/21	1.35	442.68	4.74	5.10	0.03	-0.51	1.36	0.12	-1.76	0.39	0.12	0.67	-1.76
11/05/21	1.62	417.94	4.07	3.90	-0.16	-0.22	0.73	0.16	-0.41	0.13	0.16	0.66	-0.41
05/03/21	1.57	376.70	4.19	4.25	-0.08	-0.51	-0.26	0.07	-0.33	-0.45	0.07	0.19	-0.33
10/11/20	0.96	354.04	4.75	4.58	-0.37	1.71	1.72	0.20	1.53	-0.51	0.20	0.02	1.53
11/08/20	0.64	335.57	6.24	5.90	-0.56	4.36	2.71	0.21	14.32	-0.62	0.21	0.09	14.32
11/05/20	0.71	292.50	4.11	4.27	-0.12	-5.03	2.27	0.23	-1.77	-0.85	0.23	0.27	-1.77
06/03/20	0.76	302.46	7.75	7.30	-0.18	1.78	-0.04	0.06	-2.34	-0.69	0.06	0.07	-2.34
07/11/19	1.92	308.18	9.50	9.42	-0.22	-1.42	0.25	0.36	-1.02	-0.82	0.36	0.38	-1.02
08/08/19	1.72	287.97	15.85	15.63	-0.00	0.19	0.15	0.13	-0.04	-1.08	0.13	0.14	-0.04
09/05/19	2.44	286.66	15.00	14.63	0.03	1.91	1.13	0.29	0.13	1.72	0.29	0.34	0.13
07/03/19	2.64	275.01	12.74	12.84	0.00	0.09	1.47	0.23	0.59	-3.21	0.23	0.13	0.59
09/11/18	3.18	277.76	15.44	15.54	0.24	3.49	3.13	0.56	1.07	-13.73	0.56	0.68	0.40
07/08/18	2.97	285.58	16.01	15.53	0.25	3.19	1.98	0.36	0.52	-1.51	0.36	0.42	-0.15
08/05/18	2.98	266.92	19.31	20.08	0.25	5.93	2.86	0.28	0.69	-2.66	0.28	0.21	-0.01
02/03/18	2.87	269.08	22.52	22.37	0.05	8.80	4.67	0.09	0.19	-1.29	0.09	0.08	-2.79
08/11/17	2.34	259.11	24.81	24.11	-0.02	-5.97	-3.12	0.12	0.03	-1.79	0.12	0.12	-0.06
08/08/17	2.26	247.26	18.85	18.80	-0.62	-23.36	-5.64	0.17	-2.64	-1.47	0.17	0.12	-0.10
10/05/17	2.42	239.44	23.69	24.30	-0.61	-16.48	-3.32	0.03	-2.33	-2.12	0.03	-0.01	-0.09
10/03/17	2.58	236.86	21.02	20.61	-0.61	-0.03	0.13	0.01	-6.26	-2.73	0.01	0.08	1.32
08/11/16	1.86	214.11	16.63	15.97	-0.71	-7.99	-3.17	0.13	-2.55	-4.74	0.13	0.14	-0.39
04/08/16	1.50	216.41	19.20	18.32	-0.02	-0.89	-0.91	0.27	-0.05	-19.09	0.27	0.35	-0.25
10/05/16	1.76	208.45	20.15	20.70	-0.20	-1.49	-0.15	-0.04	-0.59	5.26	0.37	0.44	-0.02
03/03/16	1.84	199.78	18.78	18.81	-0.05	-0.08	-0.05	0.49	0.40	2.92	0.49	0.54	0.40
05/11/15	2.23	210.15	21.42	21.08	-0.05	-0.55	0.02	0.02	0.36	2.74	0.07	0.05	0.36
05/08/15	2.27	210.07	23.02	23.50	-0.02	0.91	0.16	0.09	0.53	2.26	0.09	0.09	0.33
07/05/15	2.18	208.87	30.28	30.15	-0.00	0.79	-0.02	0.10	0.19	-0.27	0.10	0.11	0.23
26/02/15	2.03	211.38	25.24	25.69	-0.06	0.13	-0.01	0.04	-0.20	0.63	0.04	-0.02	-0.20

Continuación.

fecha	YIELD10	SPY_pre_eecc	YPF_pre_eecc	YPF_post_eecc	EBITDA_TO_REVENUE	CF_NET_INC	IS_COMP_NET_INCOME	SALES_REV_TURN	IS_OPER_INC	CF_FREE_CASH_FLOW	IS_COMP_SALES	IS_COMPARABLE_EBITDA	IS_COMPARABLE_EBIT
05/11/14	2.34	202.34	33.58	33.94	0.08	0.32	0.18	0.23	0.64	-0.21	0.23	0.44	0.64
07/08/14	2.41	191.03	33.43	33.76	0.07	0.19	0.20	0.31	0.66	-0.24	0.31	0.54	0.66
08/05/14	2.62	187.68	31.76	30.70	-0.02	0.15	0.72	0.42	0.51	0.15	0.42	0.57	0.51
07/03/14	2.79	188.26	28.36	28.39	0.01	-0.14	0.47	0.20	0.11	-0.27	0.20	0.58	0.79

9.3 Análisis exploratorio de Galicia.

En la imagen 9.1, se presenta una matriz de gráficos de dispersión y distribuciones de las variables provenientes de los estados contables. Este análisis visual permite identificar relaciones interesantes entre algunas de las variables, así como su distribución univariada. Por ejemplo, la relación entre IS_OPER_INC y SALES_REV_TURN es notablemente positiva, lo que refleja la lógica económica detrás de la relación entre ingresos operativos y ventas totales. Caso contrario, en las variables IS_OPER_INC y IS_COMP_NET_INCOME no se observa relaciones aparentes.

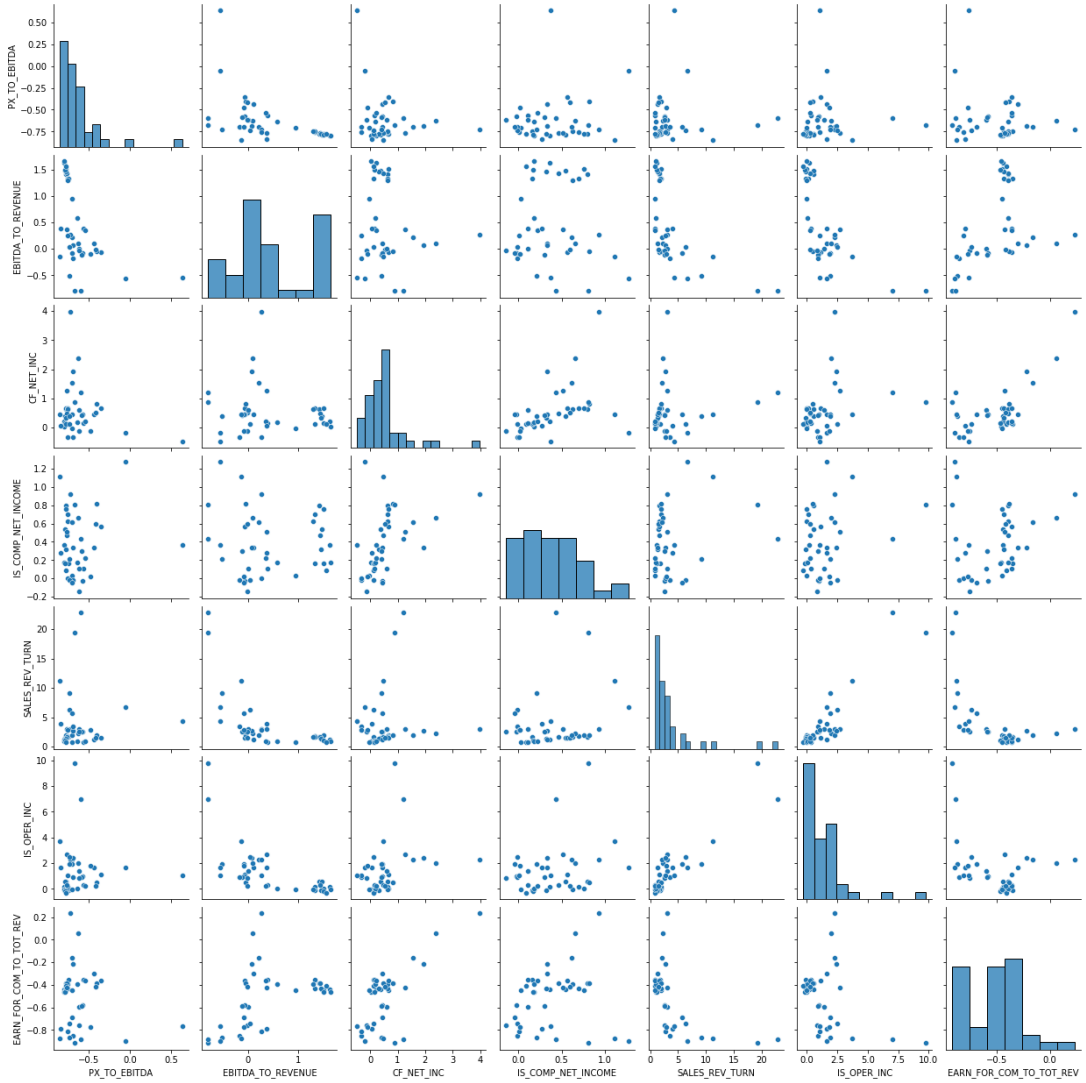

Imagen 9.1

Complementariamente, en la imagen 9.2, se muestra la matriz de correlación de Pearson entre las variables contables. Aquí se destacan relaciones más fuertes, como la correlación entre SALES_REV_TURN y IS_OPER_INC (0.90), lo cual refuerza la asociación entre los ingresos por ventas y los ingresos operativos. Otra

correlación significativa es entre CF_NET_INC y EARN_FOR_COM_TO_TOT_REV (0.67), lo que indica que las ganancias netas de la empresa están estrechamente relacionadas con los ingresos totales reportados para los accionistas. Estos patrones son consistentes con las expectativas sobre la dinámica operativa de las empresas y brindan una base sólida para las etapas posteriores del análisis estadístico.

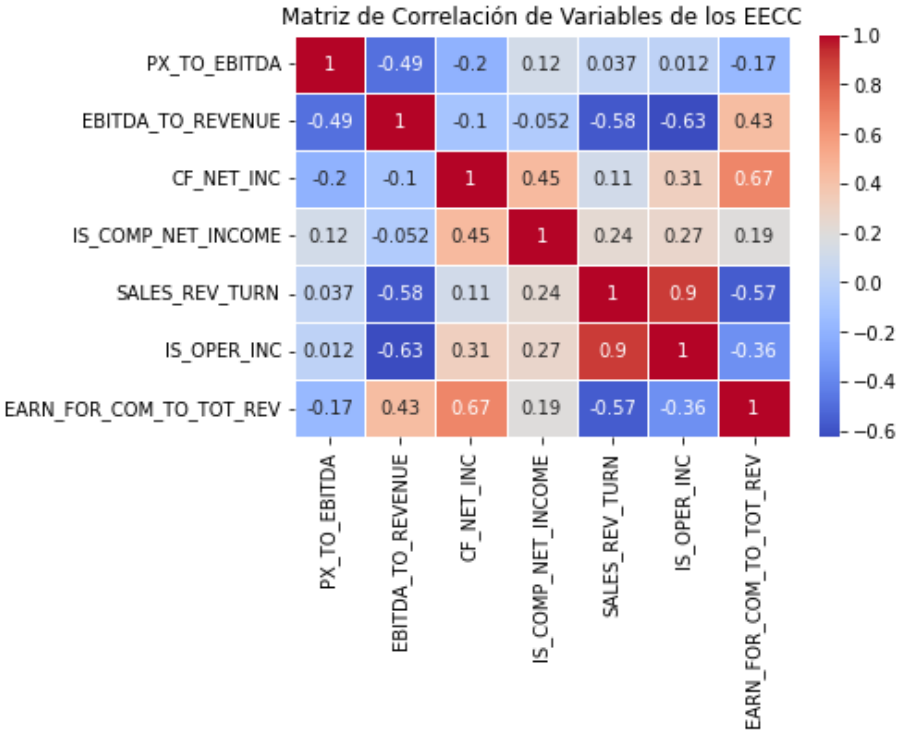

Imagen 9.2

En la imagen 9.3, se muestra un gráfico boxplot de las variables relacionadas con los estados contables. Este gráfico es útil para identificar la distribución y la dispersión de cada variable. Se observa que la mayoría de las variables tienen medianas cercanas a 0, lo que sugiere que, en promedio, las diferencias entre los valores reportados por la empresa y los valores esperados podrían no ser significativas. Sin embargo, se destacan valores atípicos en variables como SALES_REV_TURN, IS_OPER_INC y CF_NET_INC.

Estas variables son relaciones entre un valor real y uno esperado. Por lo tanto, la presencia de valores atípicos puede surgir de discrepancias en dos direcciones: un valor real que difiere sustancialmente del estimado o un valor esperado que, por alguna razón, no capturó adecuadamente el comportamiento de la empresa. Esta doble componente amplía el margen de explicación para los valores atípicos, que pueden ser producto tanto de eventos internos, como cambios operativos, o errores

en las expectativas del mercado.

Estos valores extremos deben ser analizados cuidadosamente en las etapas de modelado, ya que podrían afectar el ajuste de los modelos SARIMAX y su capacidad para predecir el comportamiento del precio de la acción.

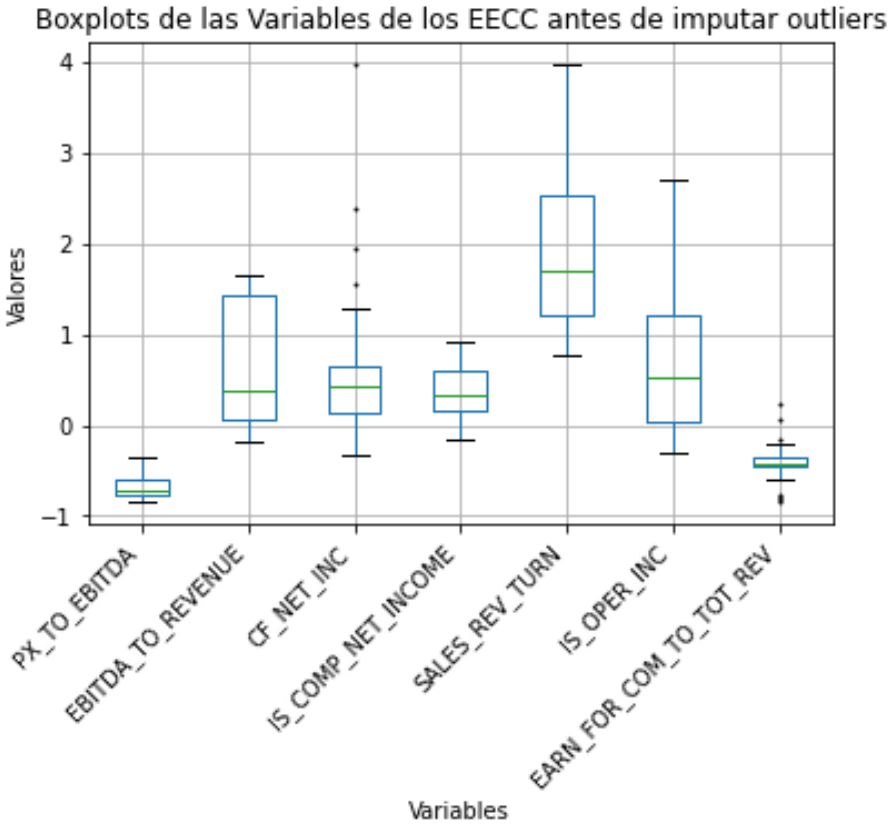

Imagen 9.3

Para la detección de valores atípicos se siguió un procedimiento basado en los cuartiles. Primero se calcularon los cuartiles Q_1 (percentil 25) y Q_3 (percentil 75), lo que nos permite determinar el rango intercuartil (IQR) definido como:

$$IQR = Q_3 - Q_1 \qquad (9.1)$$

Con esto, establecimos un umbral de detección de valores atípicos usando una constante multiplicativa de 2. Si los valores de las variables se encuentran por debajo de Q1−2×IQR o por encima de Q3+2×IQR, se consideran valores atípicos. Estos valores serán corregidos mediante imputación por *K-Nearest Neighbors* (KNN).

La imputación de valores atípicos usando el algoritmo KNN es una técnica de análisis multivariado que se basa en la proximidad de observaciones similares en el espacio de características. Formalmente, el KNN asume que los valores faltantes o

atípicos en una variable pueden estimarse tomando como referencia las k observaciones más cercanas según alguna medida de distancia, típicamente la distancia euclidiana en el espacio de n dimensiones.

Dado un punto x_i que contiene un valor atípico o valor faltante, el KNN busca los k puntos más cercanos $xj_1, x_{j2}, \ldots, x_{jk}$ para realizar la imputación. Luego, el valor de x_i se estima tomando el promedio de los valores de las k observaciones cercanas.

Matemáticamente, si $X \in R^{n \times p}$ es el conjunto de datos con n observaciones y p características, para cada observación x_i el valor imputado para una variable $x_{i,j}$ es:

$$x_{ij} = \frac{1}{k} \sum_{m=1}^{k} x_{jm} \qquad (9.2)$$

donde x_{jm} son los valores de los k vecinos más cercanos de x_i.

En la imagen 9.4 se observa el gráfico boxplot de las variables de los estados contables luego de imputar los valores atípicos.

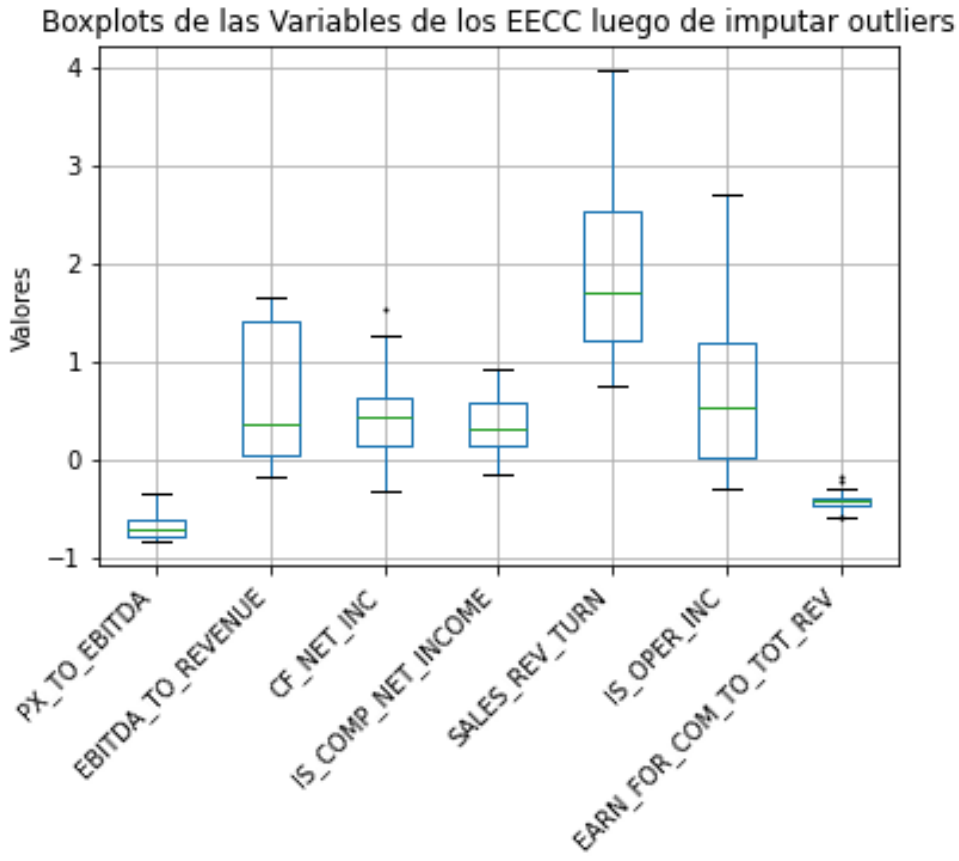

Imagen 9.4

9.4 Componentes principales de Galicia

El ACP es una técnica estadística que transforma un conjunto de variables correlacionadas en un nuevo conjunto de variables no correlacionadas llamadas componentes principales. La principal idea detrás del ACP es identificar direcciones en el espacio de los datos donde la varianza es máxima, es decir, aquellas combinaciones lineales de las variables originales que explican la mayor cantidad de información en los datos.

Formalmente, dado un conjunto de datos X con n observaciones y p variables, el ACP busca encontrar una nueva base de componentes principales. Cada componente principal z_j se define como una combinación lineal de las variables originales $x_1,x_2,...,x_p$ de la forma:

$$z_j=a_{j1}\,x_1+a_{j2}\,x_2+\cdots+a_{jp}\,x_p \qquad (9.3)$$

donde los coeficientes a_{jk} son tales que maximizan la varianza de z_j sujeta a la restricción de que los componentes sean ortogonales entre sí.

El ACP se basa en la descomposición en valores de la matriz de covarianza de los datos. A través de esta descomposición, se obtienen los autovalores (λ_1, $\lambda_2,...,\lambda_p$) y autovectores ($v_1,v_2,...,v_p$) de la matriz de covarianza, donde los autovalores indican la cantidad de varianza explicada por cada componente principal, y los autovectores proporcionan las direcciones de máxima variabilidad en el espacio de las variables originales.

Una de las principales ventajas del ACP es la reducción de la dimensionalidad, lo cual es crucial cuando trabajamos con conjuntos de datos que tienen un número elevado de variables en comparación con las observaciones disponibles, como ocurre en el presente caso. Al proyectar los datos sobre los primeros componentes principales, se logra un resumen eficaz de la información contenida en las variables originales, lo que facilita el análisis y la interpretación de los resultados, además de reducir el riesgo de sobreajuste en modelos predictivos.

El ACP permite entender cómo las variables originales se combinan para formar nuevas dimensiones ortogonales, conocidas como componentes principales, que capturan la varianza total de los datos. Los pesos, o contribuciones, de cada variable en las componentes principales se pueden observar en la tabla 9.1.

En la Componente Principal 1 (CP1) se observa que IS_OPER_INC tiene el mayor peso positivo (0.536), lo que sugiere que las ganancias operativas son el principal factor que define esta componente. CF_NET_INC (0.445) y SALES_REV_TURN (0.329) también tienen una fuerte influencia en esta dimensión, lo que resalta que los ingresos netos por operaciones y el rendimiento de las ventas juegan un rol

importante. Por su parte, EBITDA_TO_REVENUE presenta un peso negativo (-0.433), lo que implica que un aumento en esta métrica contribuye de manera opuesta a la variabilidad explicada por CP1. Esto puede señalar una relación inversa entre la rentabilidad medida por EBITDA y otros indicadores de rendimiento en esta dimensión.

En conjunto, CP1 parece reflejar principalmente una combinación de métricas relacionadas con la eficiencia operativa y la rentabilidad de la empresa, con algunas variables financieras mostrando direcciones opuestas.

	PC1	PC2	PC3	PC4
PX_TO_EBITDA	0.267	0.144	-0.731	-0.274
EBITDA_TO_REVENUE	-0.433	-0.437	0.288	-0.024
CF_NET_INC	0.445	-0.393	0.071	-0.227
IS_COMP_NET_INCOME	0.296	-0.500	0.137	-0.506
SALES_REV_TURN	0.329	0.398	0.509	-0.064
IS_OPER_INC	0.536	0.140	0.243	0.292
EARN_FOR_COM_TO_TOT_REV	0.245	-0.452	-0.202	0.726

Tabla 9.1

En los gráficos 9.5 y 9.6 se observan los gráficos de carga de las componentes principales 1 y 2, y de las componentes 3 y 4, respectivamente. La interpretación de estos gráficos permite identificar qué variables tienen mayor peso en la explicación de las componentes y sus correlaciones subyacentes, lo que contribuye a la reducción de dimensionalidad y la simplificación del análisis de las relaciones entre variables. En estos gráficos, los vectores representan las variables originales y su proyección en los ejes de las componentes indica su contribución a las mismas.

En el gráfico de las componentes 1 y 2, las variables SALES_REV_TURN e CF_NET_INC tienen una gran influencia en la CP1, dado que sus vectores son largos y están alejados del origen. Las variables EARN_FOR_COM_TO_TOT_REV y IS_COMP_NET_INCOME están positivamente correlacionadas, ya que el ángulo entre ambos vectores es pequeño. Por otro lado, la variable EBITDA_TO_REVENUE está más correlacionada con la CP2, debido a que su vector está alineado con dicho eje. Esta variable tiene una contribución negativa importante, al encontrarse en la dirección opuesta de otras como PX_TO_EBITDA.

En el gráfico de las componentes 3 y 4, los vectores más largos como PX_TO_EBITDA y EARN_FOR_COM_TO_TOT_REV indican que estas variables tienen una gran influencia en las respectivas componentes.

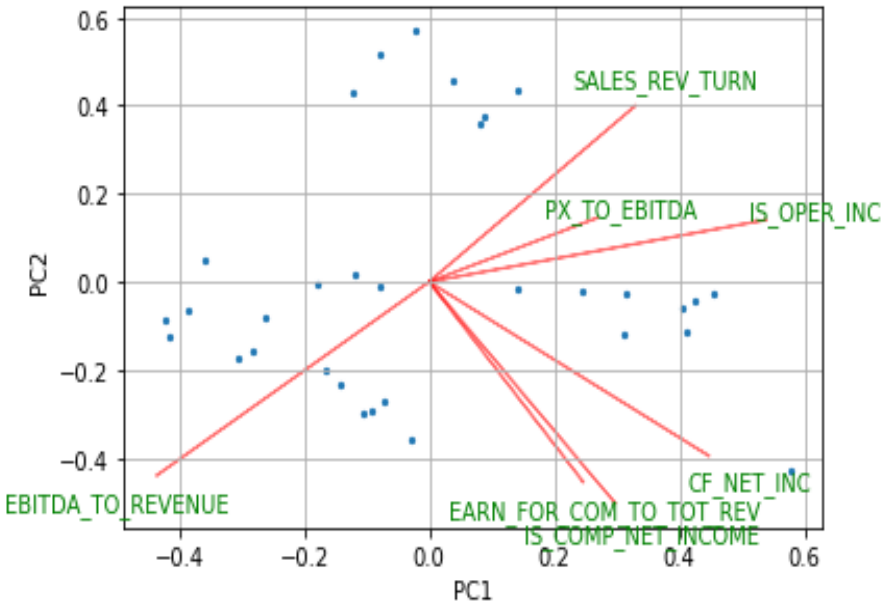

Imagen 9.5

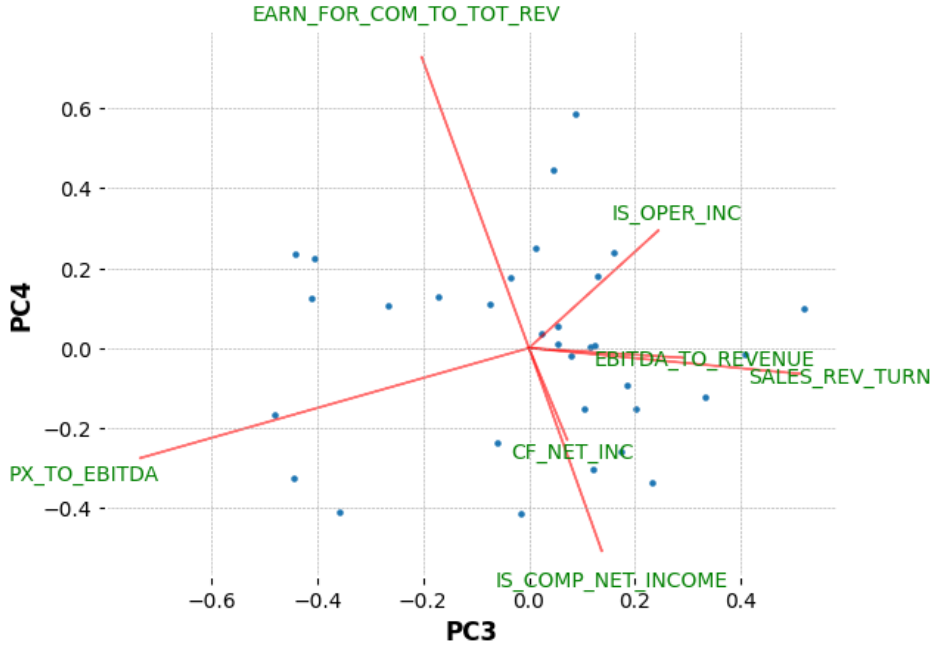

Imagen 9.6

9.5 Interacción de variables de Galicia

Utilizar interacción de variables es una técnica que permite analizar cómo dos o más variables afectan de manera conjunta a una variable dependiente, superando la simple suma de efectos individuales. La inclusión de esta nueva variable permitiría capturar estos efectos combinados, que de otro modo quedarían ocultos en los análisis que no consideran estas relaciones no lineales. Esto también podría mejorar el ajuste del modelo y su capacidad predictiva.

El análisis visual proporcionado por el gráfico 9.7 sugiere que la relación entre la variable SPY_pre_eecc y el resultado GGAL_post_eecc puede variar según los niveles de otra variable, como YIELD_GOVT_10_PRE_BALANCE. Analizando cómo cambian los patrones de color en el gráfico, se observa que distintos niveles del marcador de colores muestran distribuciones diferentes para los mismos valores de SPY_pre_eecc. Esto sugiere que hay una posible interacción entre estas dos variables.

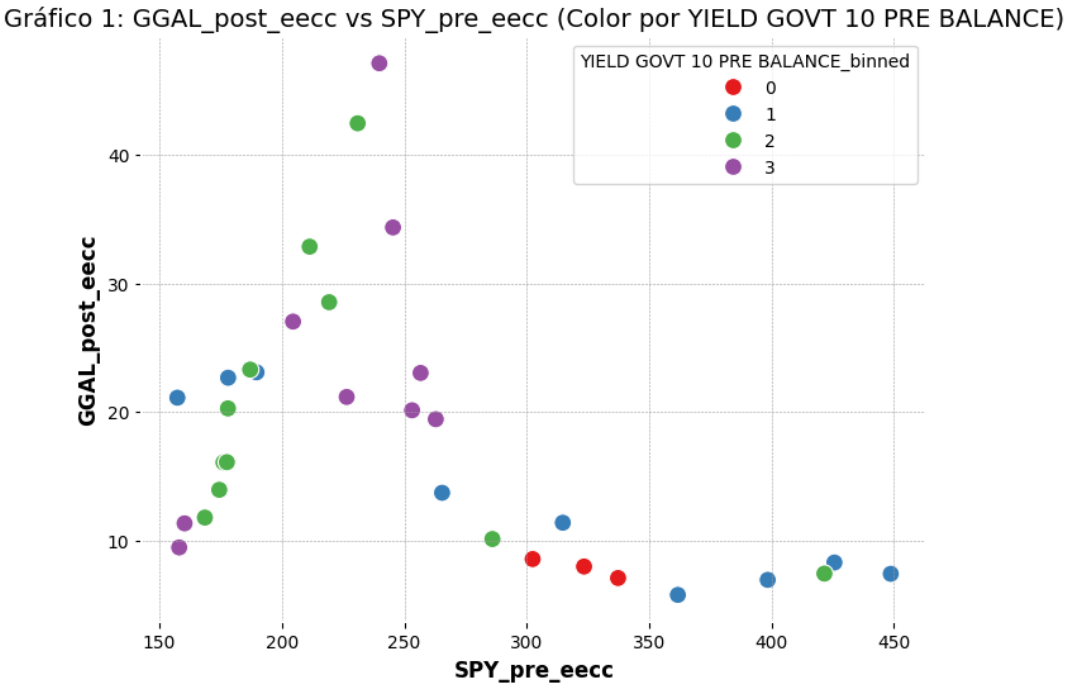

Imagen 9.7

9.6 Análisis exploratorio de YPF.

En las imágenes 9.8 y 9.9 se puede observar una clara asociación entre algunas de las variables de los estados contables. Destacan las correlaciones entre IS_COMP_NET_INCOME y CF_NET_INCOME, lo cual es razonable dado que ambas variables reflejan los ingresos netos, pero provenientes de distintos estados financieros: el estado de resultados y el flujo de caja, respectivamente. Asimismo, existe una fuerte correlación entre IS_OPER_INC y IS_COMPARABLE_EBIT, lo cual es lógico ya que los ingresos operativos, tras la deducción de costos fijos y variables, resultan en el EBIT (ganancia antes de intereses e impuestos).

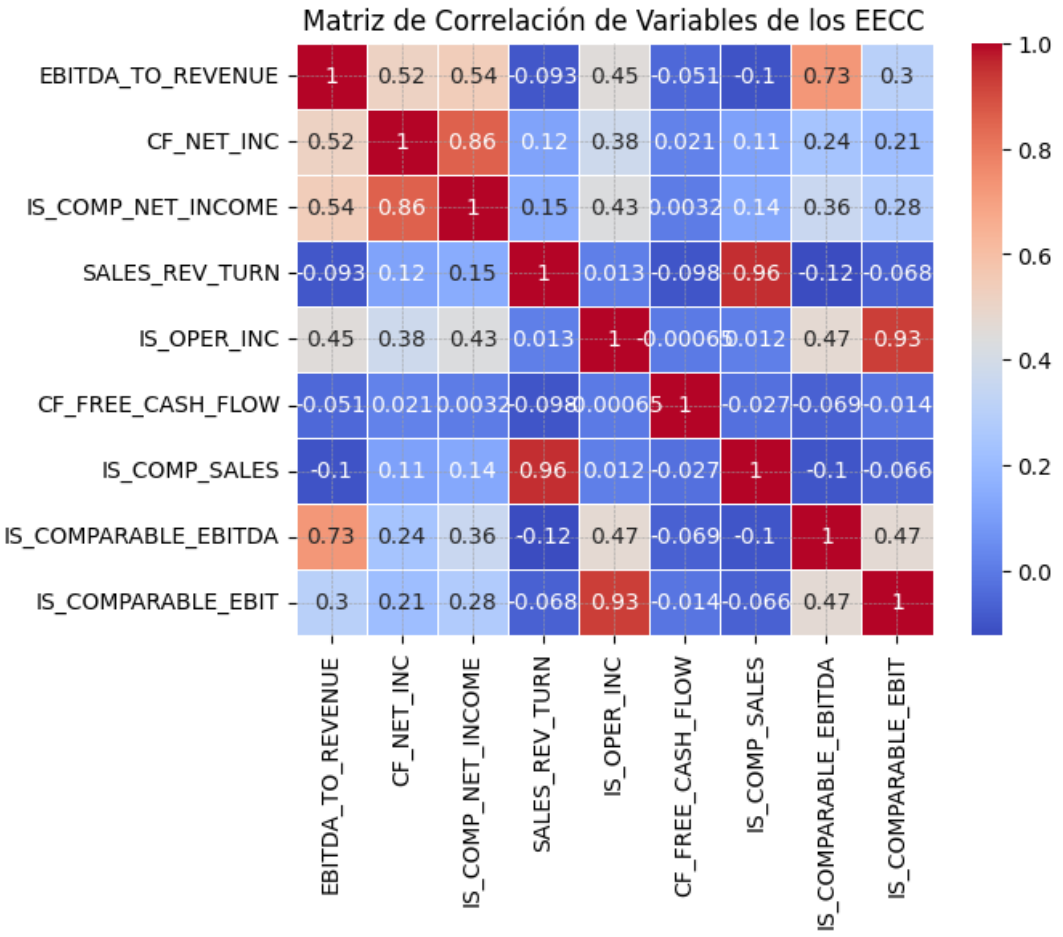

Imagen 9.8

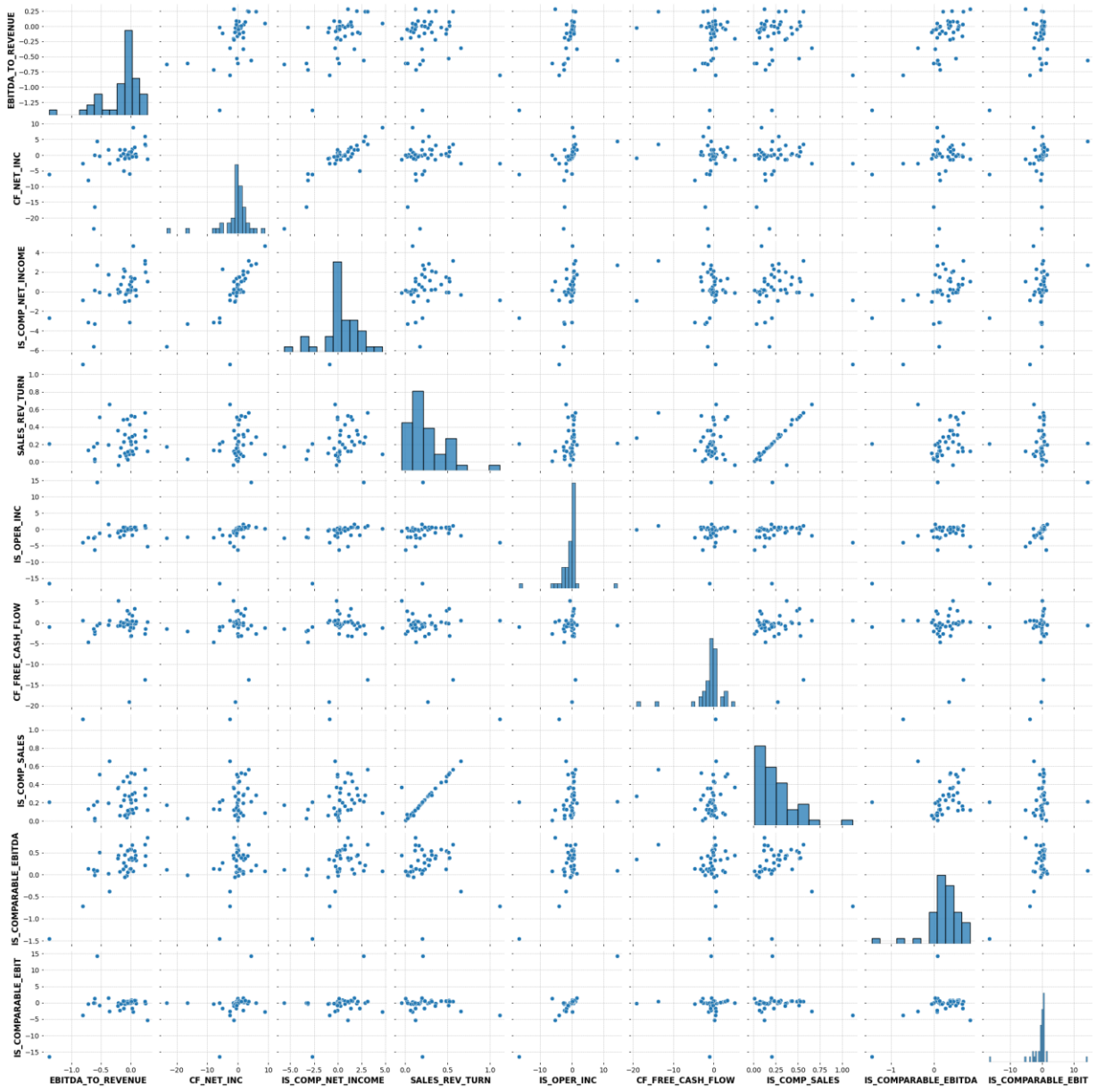

Imagen 9.9

Al igual que en el caso del conjunto de datos de Galicia, los datos de YPF no presentan valores faltantes, pero sí se identifican valores atípicos en algunas variables, como se observa en la imagen 6.4. Los valores atípicos más notables se encuentran en CF_NET_INCOME y CF_FREE_CASH_FLOW. Esto es posiblemente atribuible a la naturaleza del flujo de caja, que sigue el criterio de lo devengado, lo que lo hace más difícil de predecir. Para manejar estos valores atípicos, se volvió a utilizar el método KNN, imputando los valores mediante los vecinos más cercanos. Los datos transformados se muestran en la figura 9.11.

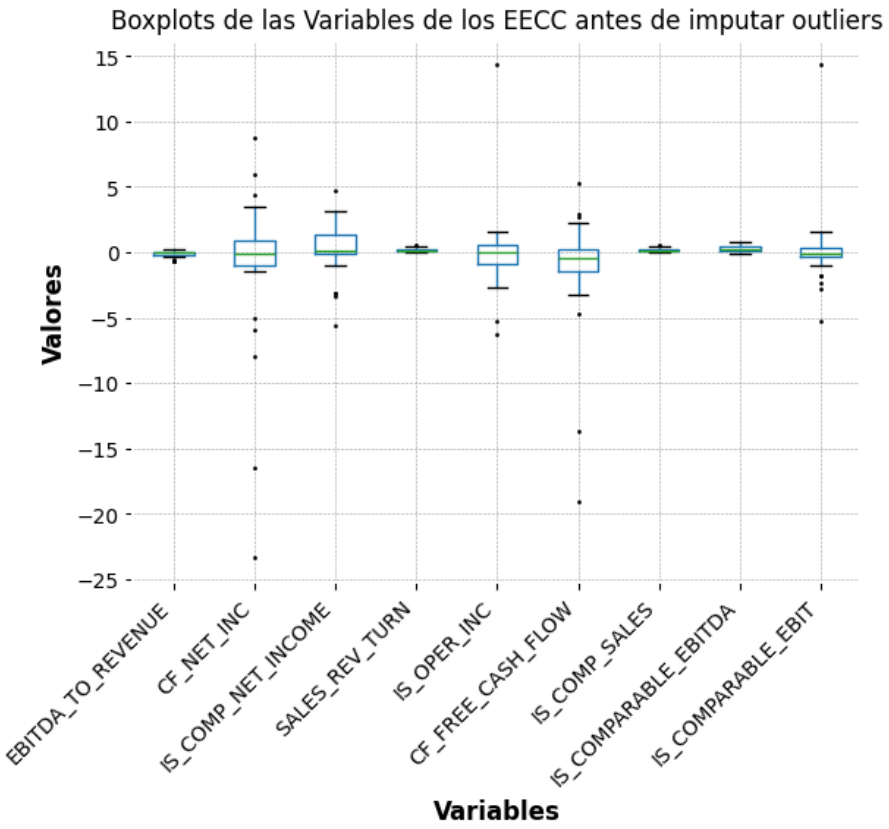

Imagen 9.10

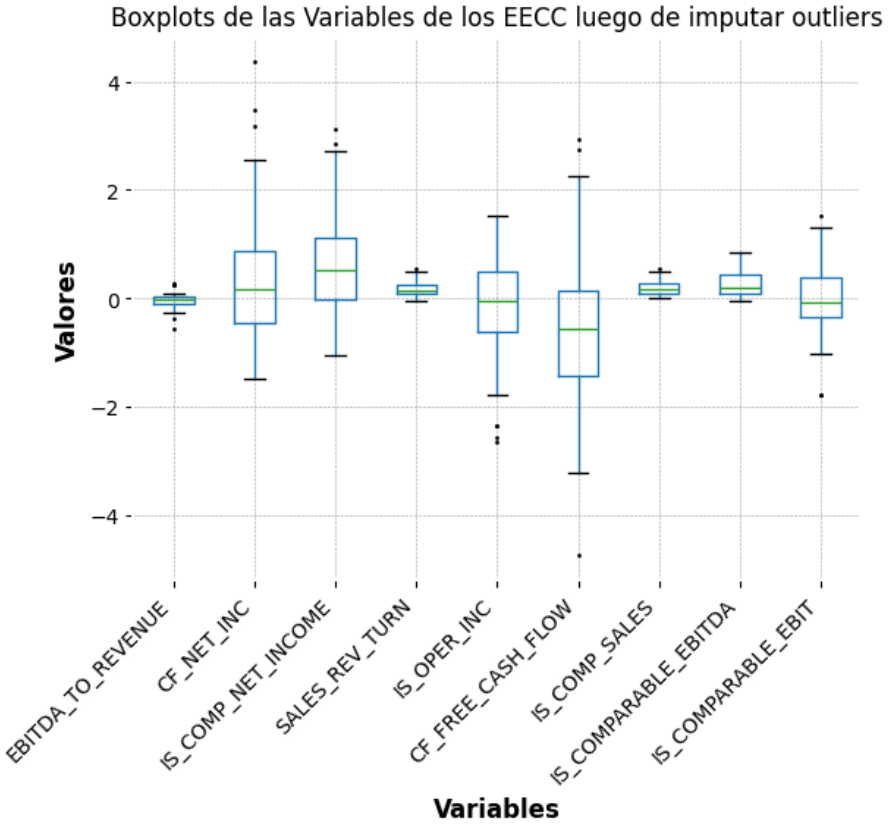

Imagen 9.11

9.7 Interacción de variables de YPF.

Para analizar la interacción entre las variables que no corresponden a los estados contables, como el precio del SPY, el precio de la empresa previo a la presentación de los estados contables, y la tasa de interés, se implementó una inspección visual. Esta se llevó a cabo mediante una función que genera gráficos de dispersión, permitiendo observar cómo varía la variable dependiente (precio de la empresa después de la presentación) en función de dos variables independientes. Para ello, se aplican *bins* a una de las variables, lo que facilita identificar patrones según los rangos categorizados.

La interacción más destacada corresponde a las variables YPF_pre_eecc y SPY_pre_eecc, lo que podría sugerir una relación entre el valor de la empresa y el índice SPY, comúnmente utilizado como indicador del "sentimiento" de los inversores. En la imagen 9.12 se observan agrupamientos significativos, lo cual sugiere que estas variables podrían tener un rol relevante en el modelo ajustado, al reflejar la influencia del entorno económico general sobre el precio de la empresa.

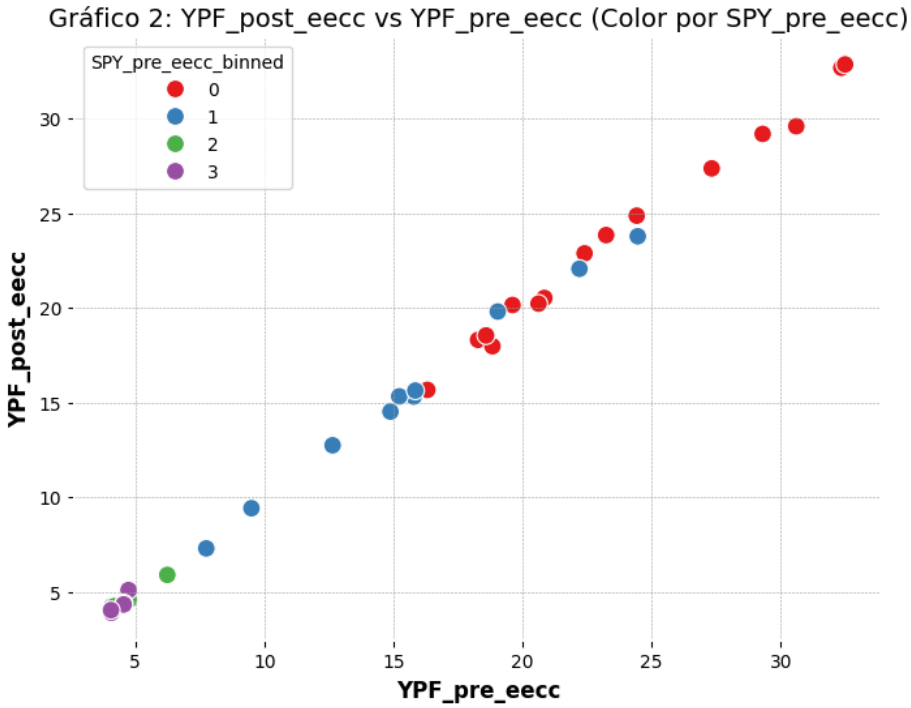

Imagen 9.12

9.8 Componentes principales de YPF.

En la imagen 9.13 se puede observar el aporte individual de cada una de estas componentes principales en la varianza explicada del modelo.

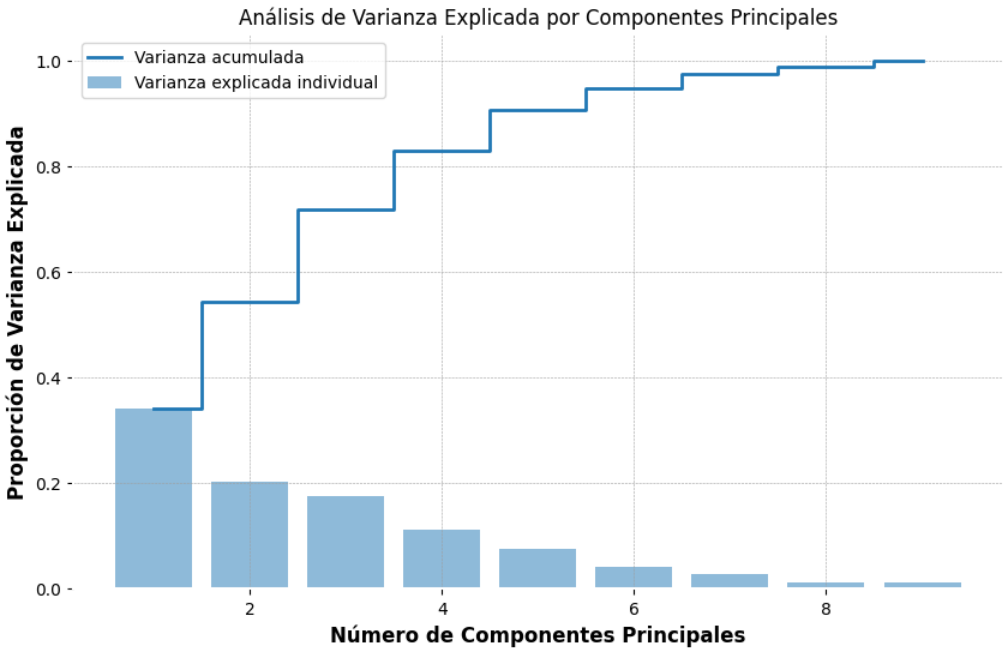

Imagen 9.13

Como se puede observar en la tabla 9.2, la primera componente principal (PC1) es la que más varianza explica, lo que indica que captura la mayor parte de la información contenida en el conjunto de datos. Vemos que las variables SALES_REV_TURN (0.510), IS_COMP_SALES (0.494) y IS_COMPARABLE_EBITDA (0.349) tienen las mayores cargas positivas, lo que sugiere que esta componente está fuertemente asociada a las métricas de ingresos, ventas y EBITDA, que son indicadores clave del desempeño financiero de la empresa.

En cuanto a las otras componentes principales, PC2 presenta una alta carga positiva en CF_NET_INC (0.531) e IS_COMP_NET_INCOME (0.515), lo que refleja una relación fuerte con los flujos de caja y el ingreso neto. PC3 tiene una alta carga positiva en IS_COMPARABLE_EBIT (0.679), lo que sugiere que está relacionada principalmente con las ganancias antes de intereses e impuestos (EBIT), mientras que PC4 y PC5 capturan variabilidad asociada a EBITDA_TO_REVENUE y CF_FREE_CASH_FLOW, señalando su relación con la eficiencia operativa y el flujo de caja.

	PC1	PC2	PC3	PC4	PC5
EBITDA_TO_REVENUE	0.255	-0.228	-0.130	-0.717	-0.420
CF_NET_INC	0.266	0.531	0.228	0.148	-0.358
IS_COMP_NET_INCOME	0.334	0.515	-0.089	0.008	-0.261
SALES_REV_TURN	0.510	-0.005	-0.100	0.162	0.225
IS_OPER_INC	0.328	-0.250	0.509	-0.072	-0.168
CF_FREE_CASH_FLOW	0.035	-0.476	0.080	0.599	-0.536
IS_COMP_SALES	0.494	-0.062	-0.127	0.204	0.401
IS_COMPARABLE_EBITDA	0.349	-0.309	-0.414	-0.050	0.022
IS_COMPARABLE_EBIT	0.134	-0.112	0.679	-0.171	0.316

Tabla 9.2

En la imagen 9.14, que analiza las componentes principales 1 (PC1) y 2 (PC2), se observa que variables SALES_REV_TURN, IS_COMP_SALES y IS_COMP_NET_INCOME tienen una fuerte influencia en PC1, mostrando relaciones positivas con esta componente. Estas variables están relacionadas con los ingresos y el desempeño neto de la empresa, lo que indica que PC1 captura principalmente la varianza explicada por la rentabilidad y las ventas. En cambio, CF_FREE_CASH_FLOW tiene una fuerte correlación negativa, lo que sugiere que esta métrica de flujo de caja juega un papel opuesto en PC1.

En la imagen 9.15, que muestra las componentes principales 3 (PC3) y 4 (PC4), se destaca que CF_NET_INC y CF_FREE_CASH_FLOW tienen una influencia significativa sobre PC4, mientras que EBITDA_TO_REVENUE y IS_OPER_INC impactan fuertemente a PC3. Esto indica que PC3 está más asociada con la eficiencia operativa y la generación de ingresos, mientras que PC4 parece capturar aspectos relacionados con el flujo de caja y el rendimiento neto, lo que sugiere que esta componente representa la estabilidad financiera de la empresa.

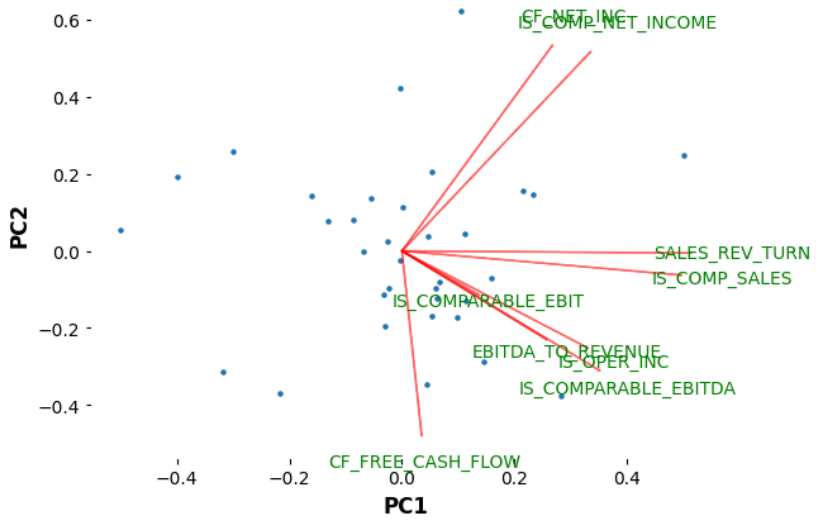

Imagen 9.14

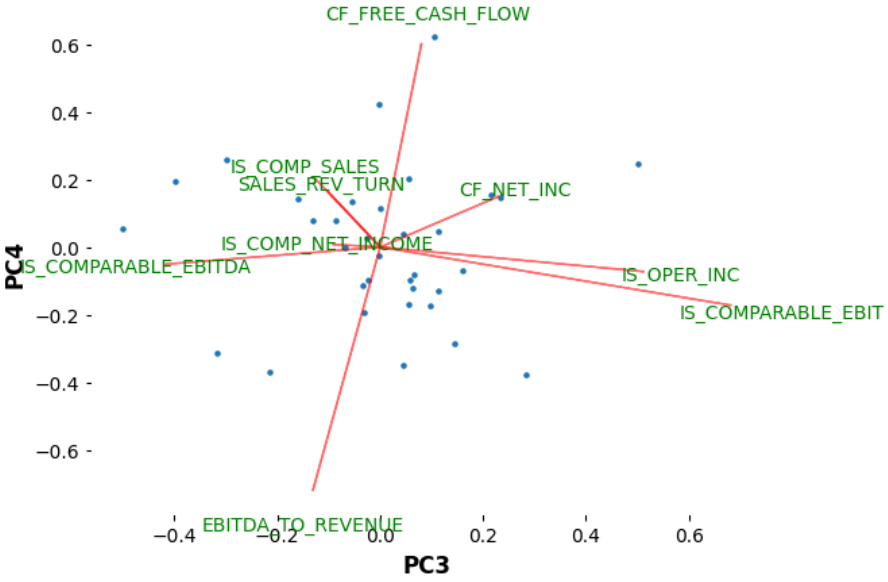
CF_FREE_CASH_FLOW
IS_COMP_SALES
SALES_REV_TURN
CF_NET_INC
IS_COMP_NET_INCOME
IS_COMPARABLE_EBITDA
IS_OPER_INC
IS_COMPARABLE_EBIT
EBITDA_TO_REVENUE
PC4
PC3
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6
−0.4
−0.2
0.0
0.2
0.4
0.6

Imagen 9.15

9.9 Enlace a repositorio

En el siguiente enlace se puede acceder al repositorio que contiene los códigos y *datasets* utilizados en esta tesis. El repositorio incluye dos archivos en formato Python y dos archivos CSV, correspondientes a cada una de las empresas analizadas. Los archivos Python documentan detalladamente todas las etapas del proceso, desde la carga de los datos, el análisis descriptivo y la ingeniería de características, hasta el ajuste de los modelos y el análisis de residuos. Por su parte, los archivos CSV contienen los *datasets* utilizados en el estudio.

Link: https://github.com/almadaagus/MaestriaEstadistica-UNR-Tesis