Escuela de Negocios
Tipo de documento: Tesis de maestría

Master in Management + Analytics

Detección de Anomalías de Precio en Comercio Electrónico

Autoría: Komel, Lucas Julián
Año: 2025

UNIVERSIDAD TORCUATO DI TELLA

UNIVERSIDAD TORCUATO DI TELLA, ESCUELA DE NEGOCIOS

MASTER IN MANAGEMENT + ANALYTICS

Detección de Anomalías de Precio en Comercio Electrónico

Resumen

En el dinámico mercado del comercio electrónico en Latinoamérica, la precisión y eficiencia en la fijación de precios son cruciales debido al vasto volumen y la diversidad de productos gestionados. Esta tesis aborda el desafío de detectar anomalías de precio en plataformas de ecommerce, donde los errores en la fijación de precios pueden provocar pérdidas económicas significativas. Es un problema de gran complejidad porque se cuenta con una gran cantidad de datos pero relativamente pocas anomalías de precio etiquetadas como tal. Para ello, se desarrollaron y compararon diversos modelos de detección de anomalías, incluidos un modelo base, regresión logística, Random Forest e Isolation Forest.

El análisis exploratorio se realizó sobre una base de datos con más de dos millones de registros de cambios de precios, identificando patrones y distribuciones en diferentes verticales de negocio. Los modelos fueron entrenados y evaluados utilizando datos históricos, aplicando una ventana móvil de 60 días para detectar anomalías en un período de un mes.

Los resultados mostraron que el modelo Random Forest tuvo el mejor desempeño, con un AUC de hasta 0,93 y un F1 Score de hasta 0,85 dependiendo de la vertical, reduciendo significativamente el número de falsos positivos y, por lo tanto, el tiempo de revisión manual respecto al modelo base.

Además, se implementó un nuevo método llamado Conformal Prediction sobre el modelo Random Forest, el cual permite calibrar el balance entre el nivel de seguridad y las tareas manuales asociadas a la revisión de anomalías.

La tesis concluye con recomendaciones para la implementación práctica de estos modelos en empresas de ecommerce en Latinoamérica, destacando la importancia de considerar eventos promocionales y estacionales en futuros estudios. También se sugiere la evaluación de modelos adicionales y la integración de más datos históricos para mejorar la detección de anomalías de precio.

Alumno: Lucas Julian Komel

Tutor: Fernando Delbianco

Abril 2025

UNIVERSIDAD TORCUATO DI TELLA, ESCUELA DE NEGOCIOS

MASTER IN MANAGEMENT + ANALYTICS

Price Anomaly Detection in Ecommerce

Abstract

In the dynamic ecommerce market in Latin America, precision and efficiency in pricing are crucial due to the vast volume and diversity of products managed. This thesis addresses the challenge of detecting price anomalies on ecommerce platforms, where pricing errors can lead to significant economic losses. It is a highly complex problem because there is a large amount of data but relatively few price anomalies labeled as such. To this end, various anomaly detection models were developed and compared, including a baseline model, logistic regression, Random Forest, and Isolation Forest.

Exploratory analysis was performed on a database of over two million price change records, identifying patterns and distributions across different business verticals. The models were trained and evaluated using historical data, applying a 60-day rolling window to detect anomalies over a one-month period.

The results showed that the Random Forest model performed best, with an AUC of up to 0.93 and an F1 score of up to 0.85 depending on the vertical, significantly reducing the number of false positives and, therefore, manual review time compared to the baseline model.

In addition, a new method called Conformal Prediction was implemented on the Random Forest model, which allows for calibrating the balance between the level of security and the manual tasks associated with anomaly review.

The thesis concludes with recommendations for the practical implementation of these models in e-commerce companies in Latin America, highlighting the importance of considering promotional and seasonal events in future studies. It also suggests evaluating additional models and integrating more historical data to improve price anomaly detection.

Student: Lucas Julian Komel

Thesis Advisor: Fernando Delbianco

April 2025

1. Índice

2. Introducción

2.1. Contexto general

Este trabajo consiste en la aplicación de diversos modelos estadísticos al problema de la detección de anomalías de precio en el comercio electrónico. Es un problema que se destaca por su complejidad, dado que aunque por un lado se cuenta con una gran cantidad de datos de precios, en términos relativos, muy pocos precios fueron revisados por un ser humano y se han etiquetado como incorrectos. Aún más, en algunos de los casos en los que tenemos precios etiquetados como anómalos, esa etiqueta tampoco está libre del error humano, que puede clasificar precios correctos como incorrectos. Por ejemplo, un mismo precio en un momento del tiempo, puede ser juzgado como correcto o incorrecto dependiendo del juicio o los incentivos de cada individuo.

Teniendo en cuenta la naturaleza intrínseca del problema, además de ciertos modelos más comunes como por ejemplo, Random Forest, se buscará aplicar una nueva técnica no cubierta dentro del programa de contenidos del Master in Management + Analytics llamada Conformal Prediction. Esta técnica se destaca por poder construirse sobre cualquier otro modelo y permite cuantificar la incertidumbre, fundamental en este contexto donde el problema es tan difícil de abordar.

Comenzaremos por dar contexto respecto al problema de negocio que se quiere resolver. Según estimaciones de Statista Digital Market Insights, el mercado latinoamericano de ecommerce superó los 117.000 millones de dólares en ventas para 2023. Se prevé que esta cifra se duplique en los próximos 5 años. La aplicación de medidas de aislamiento por la pandemia de COVID-19 llevó a un auge sin precedentes del comercio electrónico, donde se estima que se produjo un salto del 30% anual entre 2020 y 2021. Los principales mercados en esta región son Brasil y México, seguidos por Argentina, Chile, Colombia y Perú. Estos mercados además, tienen una de las mayores tasas de crecimiento previstas en el comercio electrónico de todo el mundo.

En este contexto, cada vez más plataformas profundizan los modelos híbridos, posicionándose como intermediarios entre terceros (third party) y a la vez como un vendedor más en la plataforma (first party). Un ejemplo es el caso de Amazon, que actúa como revendedor en el negocio de First Party, comprando productos a proveedores y vendiéndolos en su plataforma. Esto lo hace con el objetivo de capturar el valor del gran volumen de ventas de su plataforma y su desarrollado sistema logístico, mejorar la calidad del servicio y atraer nuevos usuarios. Otros competidores en este sector incluyen a MercadoLibre, Walmart, Magazine Luiza y Americanas.

Dada la inmensa cantidad volumen y variedad de productos que manejan, estas empresas generalmente utilizan sistemas automáticos para definir los precios de cada uno de los ítems que comercializan. Estos sistemas pueden incluir herramientas para seguir de forma automática los precios de ciertos competidores y definir el precio para liquidar el stock antes de llegar a niveles no saludables, entre otras posibilidades. Estos precios suelen estar contaminados por errores manuales, por lo tanto, de no tener algún tipo de restricción podrían ocasionar grandes pérdidas económicas al vender mercaderías a precios irrisorios.

La contracara de esto es generar un sistema de restricciones tan estricto que la cantidad de falsos positivos a gestionar activamente por los empleados de estas empresas se vuelva inmanejable, requiriendo la contratación de cada vez más personal para lidiar con la gestión de precios de los productos. Estos problemas se vuelven aún más profundos en contextos de alta inflación, con una frecuencia de variación de precios muy alta, dado que ante mayor cantidad de cambios de precios se producen mayor será la probabilidad de que se incurra en un error.

Como se puede ver en la bibliografía citada, Walmart es un claro líder en la implementación de modelos estadísticos para optimizar la respuesta ante este problema. Partiendo del paper "Anomaly Detection for an E-commerce Pricing System" (2019) de Jagdish Ramakrishnan y otros autores de Walmart Labs, buscaremos replicar algunos de los modelos aplicados para la detección de anomalías. Nos enfocaremos en los modelos de Isolation Forest y Random Forest.

Random Forest es un enfoque supervisado que utiliza múltiples árboles de decisión entrenados con diferentes subconjuntos de datos y características. Por otro lado, Isolation Forest es un enfoque no supervisado que utiliza árboles de decisión para identificar anomalías basándose en la facilidad con la que se pueden aislar puntos de datos individuales.

Además, aplicaremos el modelo de Regresión Logística, ya que es un modelo relativamente simple que se puede entrenar y ejecutar rápidamente, lo cual es beneficioso para trabajar con grandes volúmenes de datos.

2.2. El Modelo de Negocio *"First Party"*

En esta sección se desarrolla el modelo de negocios de "First Party" para dar contexto al problema de negocio al que se busca dar respuesta. Los datos que se utilizan posteriormente tienen su origen en los precios definidos por una empresa siguiendo este modelo de negocio. Considero fundamental dar este contexto para poder entender la complejidad intrínseca de la naturaleza de los datos, los cuales son manipulados por diversas personas que no necesariamente todas siguen los mismos criterios a la hora de evaluar si un precio es correcto o incorrecto para un producto en cierto momento del tiempo.

Dentro de los negocios de plataformas como son los sitios de ecommerce se pueden dar tres tipos de modelos:

1) **Modelo puro de plataforma *"enabling mode"* o de mero intermediario.**
   Estas plataformas facilitan o posibilitan la interacción y no controlan la producción del servicio. En general se las conoce como "modelo third party".
2) **Modelo puro de plataforma "controlling mode" o de integración vertical.**
   Controlan la producción del servicio o transacción. Un ejemplo son las empresas de ecommerce que actúan como revendedores o resellers dentro de la misma plataforma. Se conoce comúnmente como "modelo first party".
3) **Modelo de negocio híbrido: combinación de ambos modelos.** Dentro de las motivaciones para adoptar este modelo se encuentran capturar valor, mejorar la calidad de la plataforma y además permite a la plataforma continuar creciendo en un estadio ya maduro.

Un ejemplo de modelo de negocio híbrido de plataformas es el caso de Amazon, que actúa como revendedor en el negocio de first party: compra productos a proveedores y los vende en su plataforma. En el caso de los grandes sitios de ecommerce, existe una enorme oportunidad de negocio para capturar valor, dado el gran volumen de venta en su plataforma y el sistema logístico desarrollado. Además, sirve para atraer nuevos usuarios a la plataforma, especialmente en aquellas plataformas más maduras. Las empresas que cuentan con Marketplaces adoptan el modelo de first party por diversas razones, todas alineadas a cubrir gaps en la experiencia de usuarios. Entre ellas se encuentran:

- Ofrecer productos que la empresa no puede vender directamente sin contacto directo con las marcas.
- Ser competitivos en precio y ofrecer al consumidor el mejor precio del mercado.
- Ofrecer la mejor experiencia para el cliente, por ejemplo, gestionando los envíos en centros de distribución propios.

- Garantizar el flujo de devoluciones de productos, también conocido como logística inversa.

Estos objetivos permiten cubrir de la mejor forma el ecosistema, no desplazando, si no conviviendo en conjunto con vendedores de tipo third party. Ejemplos de empresas que siguen este modelo son:

- En Argentina: Frávega y MercadoLibre (opera en todo latinoamérica).
- En Brasil: Americanas, Casas Bahia, Extra, Magazine Luiza.
- En Chile: Falabella.
- En Colombia: Éxito.
- En México: Elektra, Linio, Liverpool y Walmart.

2.3. Problema de negocio a abordar

El problema que intenta abordar esta tesis es la optimización de los recursos humanos medidos en horas hombre dedicadas a atender la revisión de anomalías de precio, sin comprometer la seguridad del sistema de pricing, manteniendo un nivel de seguridad lo suficientemente alto para evitar grandes pérdidas económicas.

Un precio anómalo es un valor que se desvía significativamente del rango de precios esperados o normales para un producto o servicio en particular. Estos precios anómalos pueden ser causados por diversos factores, como errores de entrada de datos, fluctuaciones inusuales del mercado, estrategias de precios agresivas o eventos extraordinarios.

La capacidad de identificar precios anómalos permite detectar errores en la fijación de precios, fraude interno o externo, pérdida de competitividad externa como motivo para negociar mejores condiciones comerciales con los proveedores y estrategias de precios desleales por parte de los competidores.

2.4. Objetivo

El objetivo de este trabajo será analizar diversos modelos de detección de estas anomalías de precio, buscando llegar a una recomendación de negocio sobre qué modelo a aplicar a este problema para una empresa de comercio electrónico en Latinoamérica. La pregunta central será: ¿Tienen una performance significativamente mejor los modelos de Machine Learning o estadística avanzada aplicados a este problema que un modelo estadístico simple?

El criterio de éxito estará dado por lograr la reducción significativa de la cantidad de falsos positivos partiendo de utilizar un modelo base de alertamiento ante toda baja de precio mayor al 20%. Aplicaremos al mismo período modelos más complejos y compararemos la performance.

La reducción de los falsos positivos implica una menor cantidad de horas hombre asignadas a la revisión manual de cada uno de los cambios de precio, liberando estas horas para dedicarlas a otros aspectos estratégicos de este negocio.

Para poder estimar el tiempo en horas hombre que se le debe dedicar a la gestión de anomalías si se aplicara cada uno de los modelos, supondremos que a cada representante del equipo comercial le toma en promedio 3 minutos revisar y tomar una decisión sobre cada anomalía de precio reportada.

3. Revisión de referencias existentes en bibliografía

Para comenzar, realizaremos una exploración de referencias existentes en la bibliografía, relacionado a la detección de anomalías en series de tiempo, modelos estadísticos que se pueden aplicar a este problema y el método de Conformal Prediction.

3.1. Bibliografía sobre detección de anomalías en series de tiempo

3.1.1. Ramakrishnan, J., Li C., Shaabani, E., Sustik, M. (2019). *Anomaly Detection for an E-commerce Pricing System*

El paper "Anomaly Detection for an E-commerce Pricing System" (2019) de Jagdish Ramakrishnan y otros autores de Walmart Labs, describe el desarrollo y la implementación de un sistema de detección de anomalías de precios en ecommerce. Los autores destacan que la fijación de precios requiere actualizaciones constantes y precisas debido a la gran cantidad de productos y la velocidad de los cambios en el mercado. En este contexto, los errores en la fijación de precios pueden tener graves consecuencias financieras y afectar la confianza de los clientes.

El estudio abarca tanto enfoques supervisados como no supervisados para la detección de anomalías, implementados en contextos de procesamiento por lotes (batch) y transmisión en tiempo real (por eventos).

En el ecommerce la fijación de precios dinámica es crucial para mantenerse competitivo. Empresas como Walmart manejan millones de actualizaciones de precios diariamente, lo que requiere un sistema robusto para detectar y corregir anomalías en tiempo real. La investigación se enfoca en cómo diseñar y desplegar estos modelos en un entorno de producción a gran escala, considerando factores como la arquitectura del sistema, la eficiencia del modelo y la velocidad de procesamiento.

El sistema de detección de anomalías desarrollado para Walmart emplea una combinación de características basadas en precios, datos históricos, y transformaciones de características para mejorar el rendimiento de los modelos. Se utilizaron tanto modelos supervisados, como Random Forest y Gradient Boosting Machine, como no supervisados, como Gaussian Naive Bayes y Autoencoder. Los modelos fueron entrenados y evaluados utilizando datos reales de comercio minorista, lo que permitió ajustar los parámetros y seleccionar los modelos más efectivos para el entorno de producción. La arquitectura del sistema se diseñó para soportar tanto el procesamiento por lotes como la transmisión en tiempo real, asegurando la detección y corrección rápida de anomalías. Los resultados de este texto muestran que los modelos supervisados, específicamente

Random Forest y Gradient Boosting Machine, superan a los modelos no supervisados en términos de precisión y recall. Sin embargo, en el contexto de transmisión en tiempo real, donde la velocidad de predicción es crucial, el modelo Gaussian Naive Bayes fue preferido debido a su rapidez.

3.1.2. Tinawi, I. (2019). *Machine Learning for Time Series Anomaly Detection*

El documento "Machine Learning for Time Series Anomaly Detection" (2019) presentado por Ihssan Tinawi al Department of Electrical Engineering and Computer Science del Massachusetts Institute of Technology (MIT), explora técnicas de aprendizaje automático y métodos estadísticos para la detección de anomalías en series temporales de datos obtenidos de sensores para capturar datos de telemetría de satélites para entrenar modelos que pronostican señales basadas en patrones históricos, las cuales al superar un umbral de error dinámico, los puntos son marcados como anómalos. Los modelos aplicados incluyen Long Short-Term Memory (LSTM), autorregresión, Perceptrón Multicapa (MLP) y LSTM Encoder-Decoder. El autor mejoró el desempeño en la detección de anomalías respecto a un estudio anterior al que hace referencia.

El trabajo se divide en varios capítulos que abordan la introducción al aprendizaje automático y detección de anomalías, trabajos relacionados, detalles del dataset, arquitectura del sistema, análisis y conclusiones. También se realiza una evaluación cualitativa y cuantitativa de varios modelos de detección de anomalías aplicados a los datos de NASA, destacando el uso de umbrales dinámicos no paramétricos que mejoran la detección en datos no estacionarios.

3.2. Bibliografía sobre Conformal Prediction

3.2.1. Lei, J., G'Sell M., Rinaldo A., Tibshirani R., Wasserman L. (2017). *Distribution-Free Predictive Inference For Regression*

El documento "*Distribution-Free Predictive Inference For Regression*" (2017) fue desarrollado por Jing Lei, Max G'Sell, Alessandro Rinaldo, Ryan J. Tibshirani, y Larry Wasserman del Departamento de Estadística de la Universidad Carnegie Mellon. Los autores presentan un marco general para la inferencia predictiva sin supuestos de distribución en regresión, utilizando conformal prediction. Esta metodología permite la construcción de bandas de predicción para la variable de respuesta utilizando cualquier estimador de la función de regresión.

La inferencia predictiva es una técnica estadística que se utiliza para estimar o predecir un valor desconocido en función de un conjunto de datos observados. Es una tarea fundamental en la ciencia de datos, especialmente en contextos de regresión donde el objetivo es predecir valores futuros de una variable de respuesta. Un desafío clave en este ámbito es construir intervalos de predicción que sean válidos incluso cuando no se cumplen supuestos distribucionales específicos.

En este contexto, este método propuesto inicialmente por Vladimir Vovk en 2005, permite la construcción de conjuntos de predicción que garantizan la cobertura de muestra finita, independientemente de la distribución subyacente de los datos. Este enfoque es particularmente útil en escenarios de alta dimensionalidad donde los métodos tradicionales pueden fallar debido a las fuertes suposiciones que requieren.

3.2.2. Vovk, V., Gammerman, A., & Shafer, G. (2022). *Algorithmic Learning in a Random World*

En el libro "Algorithmic Learning in a Random World" (2022) originalmente publicado en 2005, Vladimir Vovk explora la teoría y aplicación de predicciones algorítmicas en ambientes de alta dimensionalidad y bajo supuestos de aleatoriedad. La obra se enfoca principalmente en el método de Conformal Prediction y la predicción Venn, metodologías que permiten realizar predicciones fiables sin asumir una distribución particular para los datos. En el libro se discuten diferentes enfoques y algoritmos de aprendizaje, destacando sus ventajas y limitaciones cuando se enfrentan a datos generados de manera aleatoria.

En las últimas décadas, el avance tecnológico computacional ha revolucionado el campo del aprendizaje automático, permitiendo abordar problemas cada vez más complejos. Sin embargo, uno de los desafíos persistentes es desarrollar algoritmos que no solo aprendan de los datos, sino que también sean capaces de generalizar de manera robusta en presencia de incertidumbre. Conformal Prediction es una técnica que permite asignar intervalos de confianza a las predicciones de un modelo de aprendizaje automático, garantizando un nivel predefinido de validez estadística sin asumir una distribución fija para los datos. Esta metodología se basa en la idea de que, para cualquier nuevo punto de datos, el modelo puede proporcionar un rango dentro del cual la predicción es válida con una cierta probabilidad. Por otro lado, la predicción Venn es una extensión que no solo proporciona intervalos de confianza, sino que también permite manejar la incertidumbre de manera más detallada, dividiendo el espacio de posibles predicciones en subconjuntos mutuamente excluyentes y exhaustivos.

Las técnicas de predicción algorítmica bajo aleatoriedad, como Conformal Prediction y Venn, son particularmente útiles en dominios donde los datos son inherentemente ruidosos o donde la estructura de los datos no se conoce a priori. Por ejemplo, en la detección de anomalías de precios en ecommerce, estas metodologías pueden ser empleadas para identificar precios atípicos con una alta confianza, incluso cuando los datos históricos son escasos o altamente variables. No obstante, implementar estas técnicas en la práctica presenta desafíos significativos, incluyendo la necesidad

de cálculos computacionalmente intensivos y la adaptación de los algoritmos a contextos específicos sin perder su generalidad y robustez.

3.2.3. Tibshirani, R. (2023). *Conformal Prediction, Advanced Topics in Statistical Learning*

En "Conformal Prediction, Advanced Topics in Statistical Learning" (2023), Ryan Tibshirani, Profesor de Estadística en UC Berkeley, describe Conformal Prediction como un marco relativamente nuevo para cuantificar la incertidumbre en las predicciones realizadas por algoritmos arbitrarios de predicción. Fundamentalmente, este enfoque convierte las predicciones de un algoritmo en conjuntos de predicción que poseen propiedades de cobertura en muestras finitas. Esta capacidad es particularmente relevante en el contexto del ecommerce, donde la detección de anomalías de precio requiere no solo identificar valores atípicos sino también estimar la confianza en estas predicciones.

De acuerdo al texto, el concepto de conformal prediction surgió a mediados de la década de 1990 en la Universidad de Londres, a partir de discusiones entre Vladimir Vovk y sus colegas. Desde entonces, ha sido objeto de intenso interés y desarrollo continuo, destacándose como una herramienta poderosa en la estadística y el aprendizaje automático. Larry Wasserman y sus colaboradores en la Universidad Carnegie Mellon han sido fundamentales en la adopción y el desarrollo de este marco teórico dentro de la comunidad estadística, aportando valiosas contribuciones que han mejorado su aplicabilidad.

En el ámbito del ecommerce, este método puede desempeñar un papel crucial en la detección de anomalías de precio. La capacidad de generar conjuntos de predicción con propiedades de cobertura garantizadas permite identificar con mayor precisión precios que no siguen las tendencias esperadas. Esto es particularmente útil en escenarios donde los patrones de precios pueden ser altamente variables y susceptibles a múltiples factores externos. Uno de los principales beneficios de conformal prediction es su capacidad para adaptarse a la dificultad del problema de predicción. A pesar de esto, su implementación también presenta desafíos, especialmente en términos de la complejidad computacional y la necesidad de manejar grandes volúmenes de datos típicos en plataformas de ecommerce.

3.2.4. Molnar, C., (2023). *Introduction To Conformal Prediction With Python: A Short Guide For Quantifying Uncertainty Of Machine Learning Models*

En el texto "Introduction To Conformal Prediction With Python: A Short Guide For Quantifying Uncertainty Of Machine Learning Models" (2023), Christoph Molnar presenta una serie de ejemplos prácticos de Conformal Prediction utilizando Python y la librería MAPIE. Esta librería ofrece una interfaz sencilla y accesible para incorporar conformal prediction en tareas de clasificación y regresión. Un caso de estudio ilustrativo es la clasificación de variedades de frijoles, donde se demuestra cómo generar conjuntos de predicción que garantizan la cobertura del 95% de las clases verdaderas.

La esencia de esta técnica radica en su capacidad para transformar puntajes de no conformidad en intervalos de predicción con cobertura garantizada. A diferencia de otros métodos de cuantificación de incertidumbre, como los intervalos de predicción bayesianos o el bootstrap, la predicción conformal no requiere suponer distribuciones específicas, lo que aumenta su aplicabilidad en diversas situaciones. Este texto proporciona una base teórica sólida junto con explicaciones intuitivas que facilitan la comprensión y la aplicación de este método en diversos contextos de modelado. Conformal prediction no se limita a problemas de clasificación y regresión. Su aplicabilidad se extiende a una variedad de tareas, incluyendo la detección de anomalías, la calibración de probabilidades y la predicción de series temporales.

4. Análisis exploratorio de datos disponibles

Partimos de una base de datos de 2.310.260 líneas de evolución de precios de items de un ecommerce durante tres meses para un país de Latinoamérica.

Originalmente los datos constan de una base de datos que guarda todos los eventos de cambio de precio de un grupo de ítems de dos países a lo largo del tiempo. Es decir, es posible que un ítem que no tenga cambios de precio durante varios días no tenga nuevos registros en esta base de datos. No necesariamente son los cambios de un día al día siguiente. De hecho, un ítem puede cambiar de precio varias veces dentro del mismo día. Adicionalmente, también puede que existan múltiples líneas repetidas para el mismo registro de precio si no hay cambios de un punto del tiempo al otro.

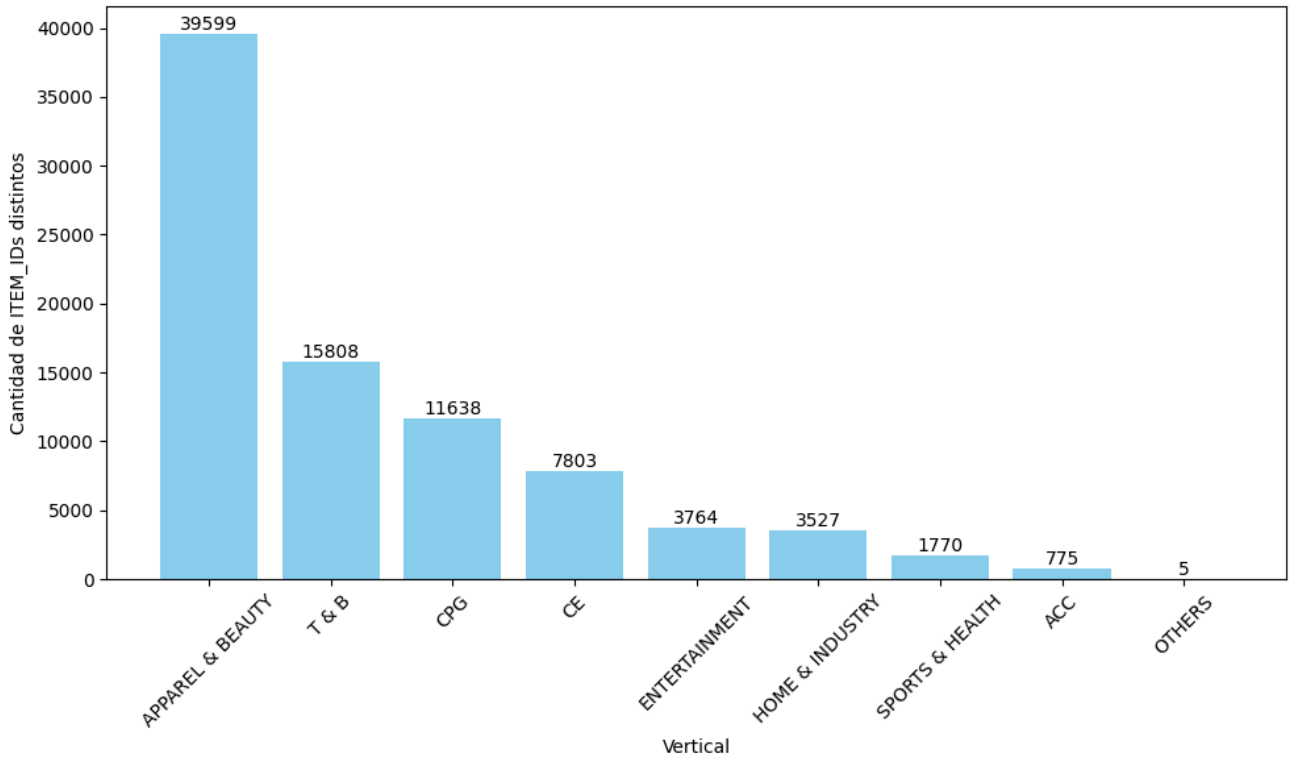
La base de datos original cuenta con las siguientes columnas:

Tabla 1: Columnas de la base de datos

Variable	Tipo	Observaciones
ITEM_ID	Caracter	Identificador único de cada item que se vende.
DATE	Fecha	Fecha de cambio de precio o status.
PRICE	Numérico	Precio de la publicación en cierta fecha.
STATUS	Caracter	Describe si la publicación se encuentra activa o pausada.
KEY_PRICE	Caracter	Concatenado entre DATE, ITEM_ID y PRICE.
FEEEDBACK	Booleano	Es 1 si un precio fue etiquetado como incorrecto por una persona del equipo que se encarga de revisar anomalías de precio, 0 si el precio no fue etiquetado ni como correcto ni como incorrecto.
VERTICAL	Caracter	Vertical de negocio a la que pertenece un ITEM_ID.

Los ítems se distribuyen en verticales de negocio, siendo la vertical de APPAREL & BEAUTY la de mayor cantidad de ítems, seguida por T & B, CPG, CE, ENTERTAINMENT, HOME & INDUSTRY, SPORTS & HEALTH, ACC, y finalmente OTHERS.

Figura 1: Cantidad de ITEM_IDs distintos por Vertical (Elaboración Propia)[1]

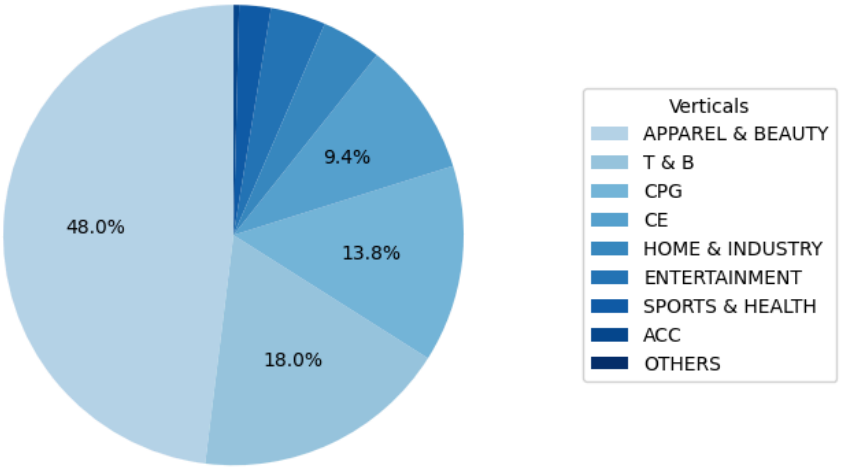


Si revisamos la cantidad de ítems medidos en términos relativos, vemos que la vertical de APPAREL & BEAUTY representa cerca del 50% del total de ítems de los datos con los que contamos.

La siguiente vertical en proporción del total es la de “T & B” con casi el 19% de los ítems, seguida por “CPG” con el 14%, y “CE” con el 9%.

[1] Todas las tablas y figuras de este trabajo son de elaboración propia.

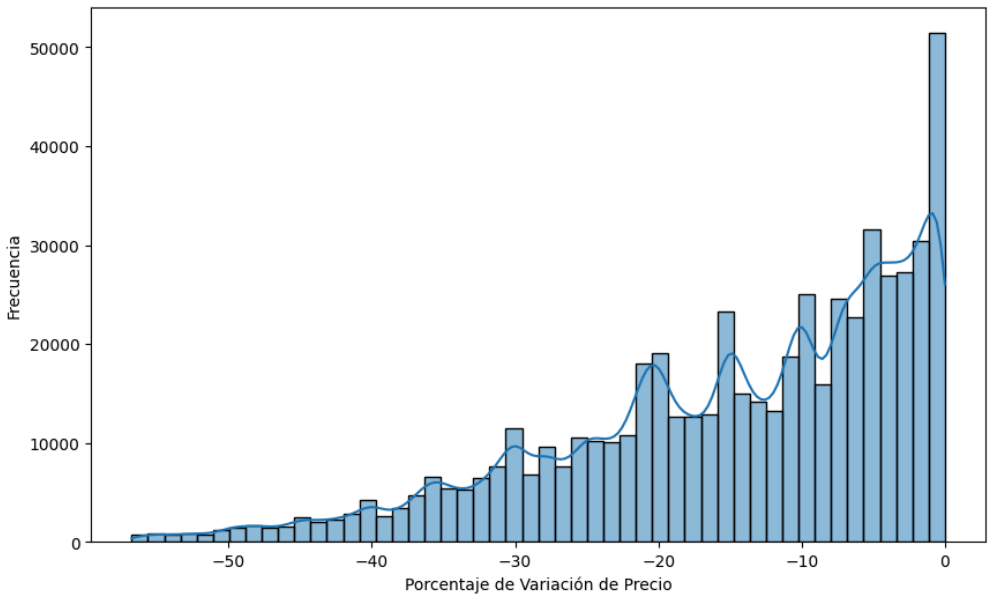
Figura 2: Distribución de ITEM_IDs distintos por VERTICAL



Para comenzar a analizar los datos, nos enfocaremos ahora en las variaciones de precio. Dado que el problema de negocio planteado es el de los errores de precio para no vender items a precios irrisorios, nos enfocaremos únicamente en los cambios de precio a la baja. Vender a precios exageradamente altos también puede ser un problema, por ejemplo perdiendo potenciales ventas, pero se considera como por fuera del foco que se quiere analizar en esta tesis

Para avanzar, es necesario entender cómo es la distribución de los cambios de precio. La distribución de los cambios de precio a la baja de todo el dataset en general es la siguiente:

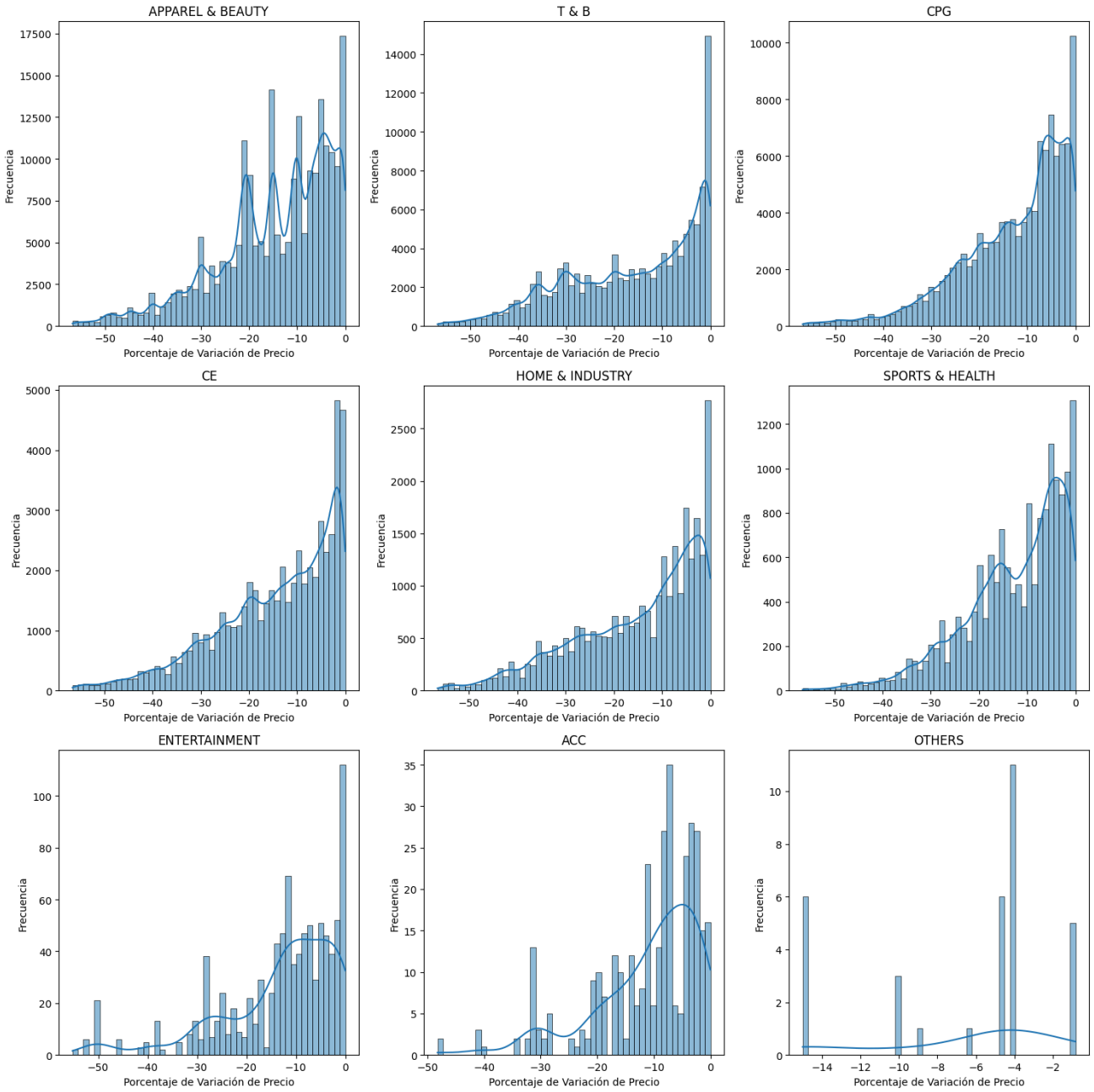
Figura 3: Distribución de Porcentajes de Variaciones de Precio a la Baja - Nivel General



En general se observan picos en números “redondos”, como por ejemplo 5%, 10%, 15%, 20%, lo cual hace sentido con la lógica comercial usual al hacer promociones con esas características.

Ahora, es necesario entender si hay comportamientos definidos para cada vertical. Es esperable que verticales con ítems de mayor valor se comporten de distinta manera que verticales con ítems de tickets más bajos respecto a su política de descuento. Aperturando el gráfico anterior por categorías podemos comenzar a entender mejor los datos.

Figura 4: Distribución de Porcentajes de Variaciones de Precio a la Baja por VERTICAL



- **“APPAREL & BEAUTY”:**

  Esta vertical se compone de ítems de vestimenta, accesorios, artículos de belleza y calzado. Nuevamente observamos el comportamiento de concentración de variaciones de precio “exactas”.

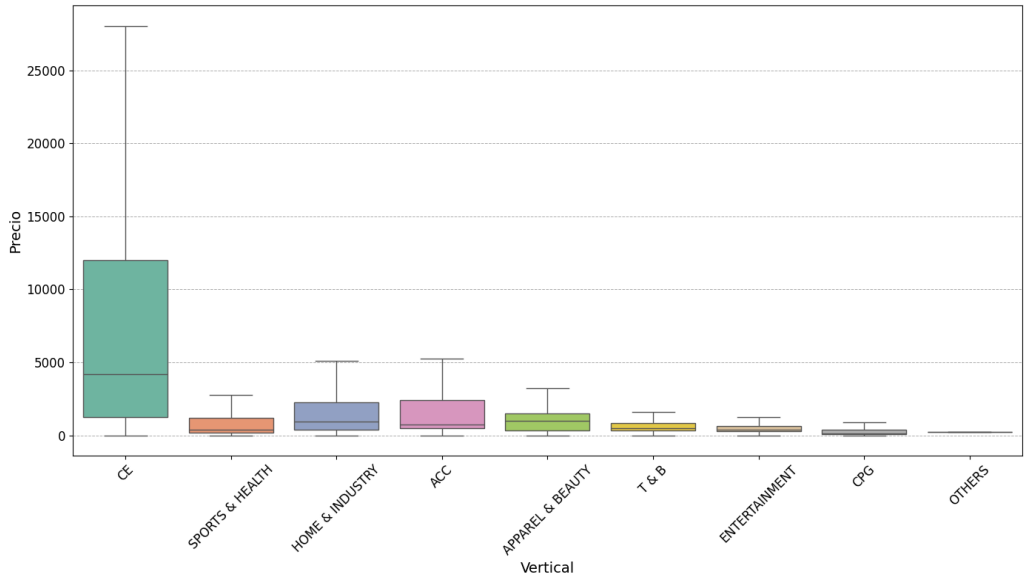
- **“CPG” o “Consumer Packaged Goods”**:

  Incluye bienes de consumo empaquetados, que los consumidores promedio utilizan diariamente y que deben ser reemplazados o reabastecidos periódicamente. Estos pueden incluir bienes como alimentos, bebidas, ropa, maquillaje, papel higiénico y otros productos para el hogar. Si bien la demanda de bienes de consumo envasados por parte de los consumidores se mantiene en gran medida constante, es un sector altamente competitivo. Esto se debe principalmente a la alta saturación del mercado y los bajos costos de cambio de los consumidores, donde los consumidores pueden cambiar fácil y económicamente sus lealtades a la marca dependiendo del precio o la calidad de los mismos.

- **“CE” o “Consumer Electronics”**:

  Incluye cualquier dispositivo electrónico diseñado para ser comprado y utilizado por usuarios finales o consumidores para fines diarios y no comerciales/profesionales. Dentro de esta vertical se incluyen ítems tales como dispositivos de audio, cámaras, drones, computadoras, notebooks, juegos, consolas, electrodomésticos para el hogar, accesorios y periféricos, smartphones, tablets, televisores, relojes inteligentes y wearables, entre otros. Como se puede apreciar, en general son ítems de un valor más alto que el resto de las verticales.

Figura 5: Distribución de Precios por Vertical (Ordenado por Precio Promedio)



- **“T & B” o “Toys and Babies”:**

  Incluye ítems de artículos para bebés, juguetes y juegos.

- **“HOME & INDUSTRY”:**

  Incluye ítems de construcción e iluminación, artículos de decoración, muebles, sábanas, acolchados, cortinas, artículos de jardín, herramientas de construcción entre otros.

- **“ENTERTAINMENT”:**

  Esta vertical incluye ítems tales como libros, artículos multimedia e instrumentos musicales.

- **“SPORTS & HEALTH”:**

  Incluye ítems relacionados a los deportes y el bienestar físico.

- **“ACC”:**

  Corresponde a accesorios de automóviles y motocicletas, neumáticos, aceites, filtros y demás herramientas.

- **“OTHERS”**:

  Incluye ítems que no encajan dentro de las otras categorías. Dado que posee tan pocos ítems, la descartaremos para correr los modelos.

Revisemos ahora la cantidad de filas que tiene el data frame para cada mes. Como se puede observar el mes de febrero es el de mayor cantidad de observaciones. Parece razonable utilizar como datos de entrenamiento los datos de los primeros dos meses y luego utilizar un período móvil de 60 días para aplicar los modelos sobre los cambios de precio del mes de marzo.

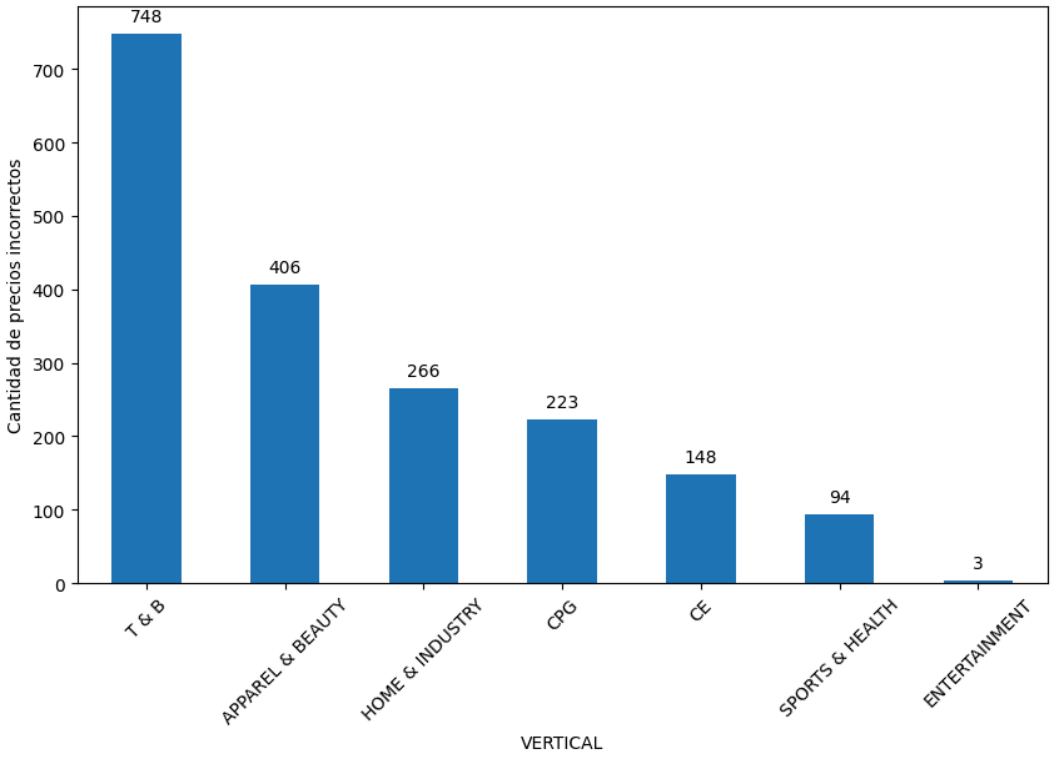
Tabla 2: Cantidad de filas de datos por mes

Mes	Cantidad de Filas
enero 2024	695.898
febrero 2024	885.919
marzo 2024	728.443

4.1 Análisis de datos etiquetados

Del total de observaciones, 1888 líneas han sido etiquetadas como incorrectas. Si revisamos la cantidad de observaciones con precios etiquetados como incorrectos por vertical obtenemos el siguiente gráfico:

Figura 6: Cantidad de precios etiquetados como incorrectos por VERTICAL



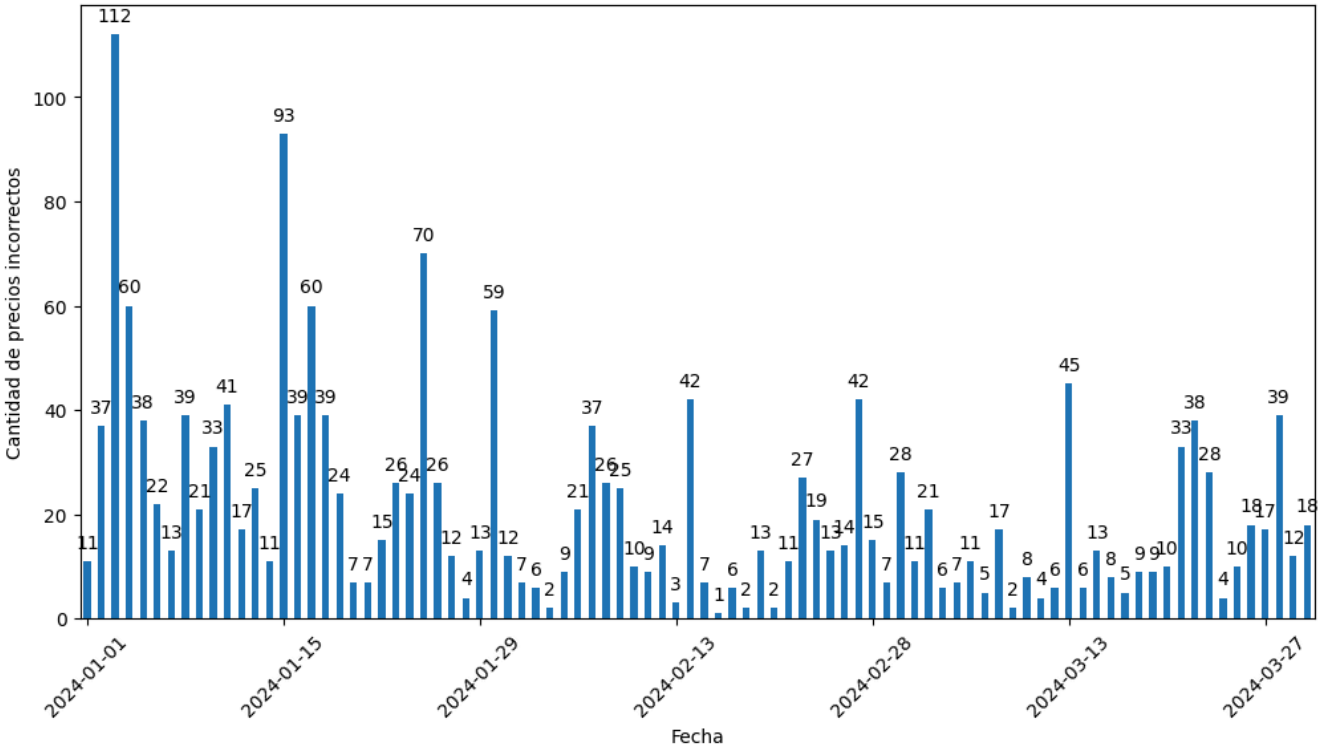
La vertical con mayor cantidad de casos etiquetados como erróneos es “T & B”. Sin embargo si comparamos el ratio de la cantidad de casos erróneos sobre la cantidad de items activos para cada vertical, aquella que obtiene el ratio más alto es “HOME & INDUSTRY” con un 8%.

Tabla 3: Ratio de precios incorrectos por vertical

Vertical	Cantidad de Items	Cantidad de precios etiquetados como incorrectos	Ratio de cantidad precios incorrectos sobre cantidad de items activos
HOME & INDUSTRY	3.527	266	8%
SPORTS & HEALTH	1.770	94	5%
T & B	15.808	748	5%
CPG	11.638	223	2%
CE	7.803	148	2%
APPAREL & BEAUTY	39.599	406	1%
ENTERTAINMENT	3.764	3	0%

Podemos también graficar la cantidad de precios etiquetados como incorrectos para cada fecha a lo largo del tiempo:

Figura 7: Cantidad de precios etiquetados como incorrectos por fecha



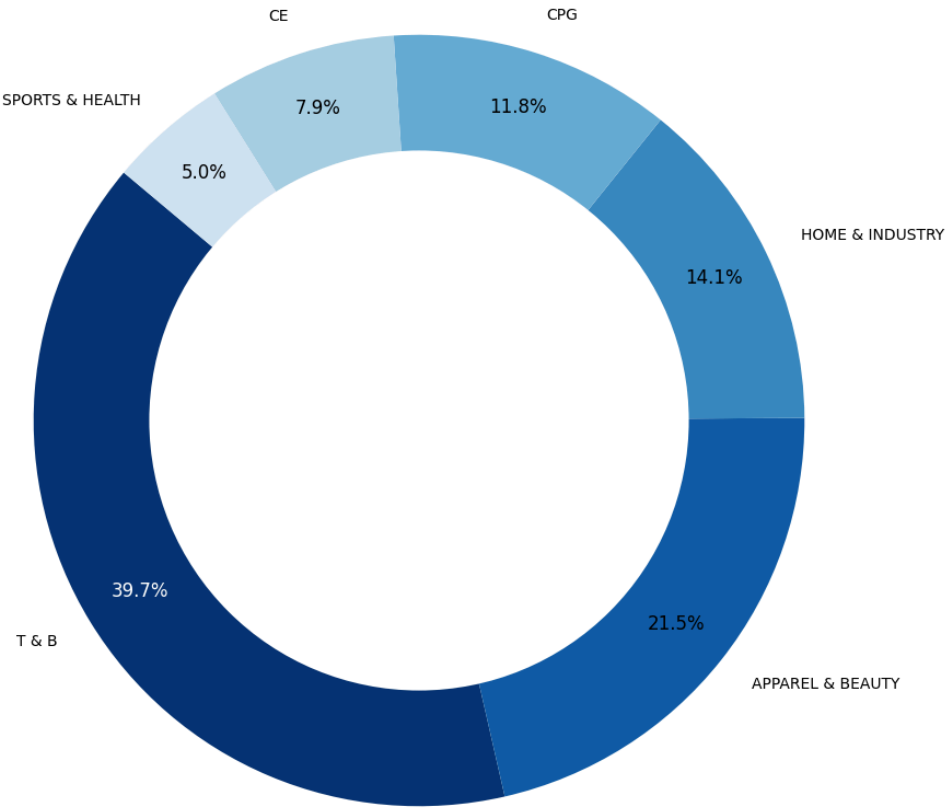
Este gráfico nos permite entender que existen ciertas fechas con mayor concentración de precios etiquetados como incorrectos.

Las 10 fechas con mayor cantidad de precios etiquetados como erróneos se muestran en la siguiente tabla:

Tabla 4: Cantidad de precios etiquetados como incorrectos por fecha, ordenados de mayor a menor cantidad por fecha

Fecha	Cantidad de precios etiquetados como erróneos
2024-01-03	119
2024-01-15	94
2024-01-25	72
2024-01-04	61
2024-01-17	60
2024-01-30	59
2024-01-02	46
2024-03-13	45
2024-01-18	42
2024-02-14	42

Figura 8: Distribución de precios etiquetados como incorrectos por VERTICAL



5. Metodología

5.1. Tipos de modelos

A partir de los modelos desarrollados en libro de James et al (2023)[2] para realizar este análisis partiremos de los siguientes modelos:

- Modelo Base: bajas de precio de más de 20%.
- Modelo Lineal: Regresión Logística.
- Modelo No Lineal: Random Forest.
- Modelo No Supervisado: Isolation Forest.

5.2. Pasos a seguir

Para cada uno de los modelos la metodología será la siguiente:

- Se importan las librerías necesarias para el manejo de datos (pandas, numpy), para la construcción de modelos (sklearn), y para el manejo de fechas (datetime).
- Se filtran los datos para incluir solo aquellos pertenecientes a la categoría 'T & B' de la columna VERTICAL. Esta vertical se seleccionó por ser aquella con mayor cantidad de precios etiquetados como erróneos lo que permite un mejor entrenamiento de los modelos.
- Se convierte la columna DATE al formato datetime para facilitar el manejo de fechas.
- Tomaremos a modo de datos de entrenamiento aquellos datos de entre el 01-01-2024 y el 29–02-2024. De esta forma, contaremos con dos meses de entrenamiento y un mes de testeo.
- Luego, trabajaremos con una ventana móvil de 60 días para ir evaluando los cambios de precio de cada nuevo día en base a los datos históricos de comportamiento de precios de los 60 días anteriores. Esto será para los datos de entre el 01-03-2024 y el 31-03-2024.
- Correremos cada uno de los modelos.
- Compararemos los resultados utilizando métricas de desempeño.
- Seleccionaremos el modelo con mejor performance y realizaremos una recomendación de qué modelo aplicar para una empresa de ecommerce.
- Aplicaremos Conformal Prediction al modelo con mejor performance.
- Evaluaremos si este método mejora la performance del modelo seleccionado.

[2] James, G., Witten, D., Hastie, T., Tibshirani, R., & Taylor, J. (2023). An Introduction to Statistical Learning: with Applications in Python. Springer Nature.

6. Modelo Base: bajas de precio de más de 20%

6.1. Descripción general

Partiremos de un modelo base donde se considerará necesario de revisión manual todo cambio de precio a la baja de más de 20%. Este porcentaje se ha seleccionado en base a la observación de los gráficos de distribución de descuentos a la baja. En general, las empresas de ecommerce, especialmente las que cotizan en el Mercado de Valores, se ven obligadas a establecer controles sobre las variaciones de precio a la baja. Esto es con el objetivo de reducir posibles potenciales pérdidas económicas generadas por la aplicación de precios irrisorios sobre las publicaciones.

Un precio erróneo puede generar en muy pocos minutos, una pérdida muy grande para estas empresas. Existen grupos organizados que están a la espera de estas oportunidades y comparten aquellas publicaciones en grupos multitudinarios para aprovechar los errores de las empresas a la hora de establecer los precios de las publicaciones. Teniendo esto en cuenta, las instituciones que regulan estas empresas suelen definir como mínimo controles basados en porcentajes fijos de descuento. Este modelo base se ha incorporado al análisis para representar este aspecto de la relación del negocio de first party con las instituciones que lo regulan.

6.2. Aplicación del modelo

Para aplicar este modelo, se siguen los siguientes pasos:

- Se define una función llamada ‘detect_price_drop’ que toma los datos de cambio de precio como entrada.
- Se ordenan los cambios de precio por la columna 'DATE' para asegurarse de que los precios estén en orden cronológico.
- Para cada producto identificado por 'ITEM_ID', se agrega una nueva columna 'PREVIOUS_PRICE' que contiene el precio del día anterior. La función ‘shift(1)’ desplaza los valores hacia abajo una fila dentro de cada grupo de 'ITEM_ID'.
- Se agrega una nueva columna 'PRICE_DROP' que calcula la proporción de la caída de precio entre el precio anterior y el precio actual.
- Se agrega una columna 'ANOMALY' que marca True si la caída de precio es mayor al 20%, indicando que hay una baja significativa en el precio.
- Finalmente, la función retorna el DataFrame modificado con las nuevas columnas 'PREVIOUS_PRICE', 'PRICE_DROP' y 'ANOMALY'.
- Se obtiene la cantidad de anomalías de precio detectadas bajo esta definición.

7. Modelo Lineal: Regresión Logística

7.1. Descripción general

La regresión logística es un modelo estadístico utilizado para predecir la probabilidad de un resultado binario basado en una o más variables independientes. Se basa en la función logística, que es una función sigmoide que mapea cualquier valor real en el intervalo (0, 1).

Es un método de clasificación supervisada diseñado para predecir la probabilidad de que una observación pertenezca a una de dos categorías, como por ejemplo “precio normal” y “precio anómalo". A diferencia de la regresión lineal, en la que el objetivo es ajustar una recta, aquí se modela la probabilidad $p(x) = P(Y = 1 \mid X = x)$ mediante la función logística:

$$log\frac{p(x)}{1-p(x)} = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p$$

donde $\beta_0, \ldots, \beta_p$ son parámetros a estimar.

Para la estimación de parámetros se utiliza máxima verosimilitud, buscando los coeficientes $\widehat{\beta}$ que maximicen:

$$L(\beta) = \prod_{i=1}^{n} p(x_i)^{y_i} [1 - p(x_i)]^{1-y_i}$$

Un coeficiente β_j positivo indica que un aumento en x_j eleva el cociente de probabilidades $\frac{p(x)}{1-p(x)}$.

El valor $\widehat{p}(x) \in (0, 1)$ puede interpretarse directamente como el score de pertenecer a la clase 1.

El umbral se ajusta según el tradeoff entre falsos positivos (coste de investigar precios normales) y falsos negativos (anomalías no detectadas).

Para detectar anomalías de precio, el modelo de regresión logística puede ser entrenado con un conjunto de datos históricos de precios de productos. Es decir, se usa para modelar un resultado que puede tomar dos valores posibles, típicamente etiquetados como 0 y 1. En el contexto de la detección de anomalías, estos valores pueden representar "precio normal" (0) y "precio anómalo" (1).

En el contexto de la detección de anomalías de precio en ecommerce, este modelo puede ser una herramienta valiosa para identificar precios que se desvían significativamente de los patrones normales.

7.2 Aplicación del modelo

Para aplicar este modelo en Python se utiliza la clase 'LogisticRegression' de 'sklearn.linear_model'.

- Se selecciona la columna 'PRICE' como característica '(X)' y 'FEEDBACK' como etiqueta '(y)'.
- Se define un transformador de columnas que deja pasar la columna 'PRICE' sin cambios ('passthrough'), ya que no se necesita codificación adicional.
- Se crea un pipeline para la Regresión Logística.
- Se entrena el modelo de Regresión Logística y se calculan y almacenan el AUC de ROC y las predicciones.

8. Modelo No Lineal: Random Forest

8.1. Descripción general

Según James et al (2023)[3] el modelo de Random Forest es un enfoque supervisado que utiliza múltiples árboles de decisión para mejorar la precisión de la detección de anomalías. Se fundamenta en la combinación de varios árboles de decisión independientes, cada uno entrenado sobre una muestra aleatoria del conjunto de datos original (bagging), con reemplazo. Un árbol de decisión es un clasificador secuencial que particiona el espacio de características mediante umbrales definidos en cada nodo, generando hojas que asignan una etiqueta o probabilidad de pertenencia a la clase de interés.

El procedimiento de bagging (Bootstrap Aggregating) consiste en generar subconjuntos de entrenamiento, cada uno obtenido por muestreo aleatorio con reemplazo de la muestra original. Cada árbol se ajusta a uno de estos subconjuntos, lo que reduce la varianza del estimador global sin incrementar el sesgo de manera significativa. Además, en cada división de nodo, Random Forest incorpora un nivel adicional de aleatoriedad: en lugar de considerar todas las variables explicativas, selecciona al azar un subconjunto de características Esta estrategia disminuye la correlación entre árboles y, por ende, mejora la capacidad de generalización del conjunto. La predicción del Random Forest en problemas de clasificación probabilística consiste en promediar las probabilidades estimadas por cada árbol.

A pesar de su alta precisión el modelo puede presentar generalmente dos limitaciones.

[3] James, G., Witten, D., Hastie, T., Tibshirani, R., & Taylor, J. (2023). An Introduction to Statistical Learning: with Applications in Python. Springer Nature.

Su complejidad computacional para el entrenamiento y la inferencia de cientos o miles de árboles pueden resultar costosos en tiempo y memoria, lo que dificulta su aplicación en escenarios con restricciones de recursos o requisitos de respuesta en tiempo real. Sin embargo, no consideramos esto un impedimento en el contexto del problema de negocio que estamos analizando, dado que las grandes empresas de ecommerce cuentan con los recursos necesarios para aplicar estos modelos a gran escala con un rendimiento en línea con la expectativa para este problema.

Por otro lado, su reducida interpretabilidad puede ser un problema para explicar por qué un precio fue considerado como una anomalía o no a los miembros del equipo que se encargan de la revisión de los casos. La naturaleza agregada del modelo impide una fácil comprensión de las razones subyacentes a cada predicción. Sin embargo, consideramos que es conveniente evaluar estos modelos de acuerdo a sus métricas de desempeño y establecer una cultura en la que se confíe en los modelos aplicados a este problema sin necesidad de tener explicabilidad.

El Random Forest se fundamenta en la combinación de varios árboles de decisión independientes, cada uno entrenado sobre una muestra aleatoria del conjunto de datos original (bagging), con reemplazo. Un árbol de decisión es un clasificador secuencial que particiona el espacio de características mediante umbrales definidos en cada nodo, generando hojas que asignan una etiqueta o probabilidad de pertenencia a la clase de interés.

8.2. Aplicación del modelo

Para aplicar este modelo en Python se utiliza la clase ‘RandomForestClassifier’ de ‘sklearn.ensemble’.

- Se selecciona la columna ‘PRICE’ como característica (X) y ‘FEEDBACK’ como etiqueta (y).
- Se entrena el modelo de Random Forest y se calcula el AUC de ROC y las predicciones.
- Se almacenan los resultados de AUC y predicciones para el modelo.

9. Modelo No Supervisado: Isolation Forest

9.1. Descripción general

Isolation Forest es un modelo de aprendizaje automático no supervisado que representa un conjunto de árboles de decisión. Se basa en la idea central de que las anomalías son observaciones raras y, por tanto, más fácilmente separables. Es decir, que los puntos de datos que se dividen anteriormente en el árbol tienen más probabilidades de ser anomalías que aquellos que viajan más abajo en el árbol.

Cada árbol se construye a partir de una submuestra aleatoria de la población original, y en cada nodo se selecciona aleatoriamente una característica y un umbral de corte para dividir el espacio de atributos. Una vez aislado hasta un nodo hoja, se mide la profundidad recorrida: los puntos que requieren pocas particiones reciben un score de anomalía alto, reflejando su rareza.

El método es el siguiente:

1. Se selecciona una submuestra aleatoria de datos para un árbol binario.
2. Dentro de esa submuestra aleatoria, la ramificación comienza a tener lugar en función de una característica y un umbral aleatorios. Si un punto de datos cae por debajo del valor umbral, se asigna al subárbol izquierdo. De lo contrario, se asigna al subárbol derecho.
3. Este flujo de trabajo de ramificación continúa hasta que se alcanza la profundidad máxima del árbol o hasta que cada punto de datos esté completamente aislado. Este proceso se repite para cada árbol del conjunto.
4. Dependiendo de la profundidad requerida para alcanzar un punto de datos, se asignan scores de anomalía. Cuanto más negativo es el score, más anómalo es un dato.

Para evaluar los resultados se utiliza un conjunto de datos etiquetado y métricas como la precisión, la sensibilidad y la especificidad.

9.2. Aplicación del modelo

Para aplicar este modelo en Python se utiliza la clase ‘IsolationForest’ de la librería ‘sklearn.ensemble’.

- Se selecciona la columna ‘PRICE’ como característica ‘(X)’ y ‘FEEDBACK’ como etiqueta ‘(y)’.
- Se entrena el modelo Isolation Forest.
- Se convierten las predicciones del Isolation Forest a formato binario y se calcula el AUC de ROC.
- Se almacenan los resultados de AUC y predicciones para el modelo.

10. Comparación y evaluación de desempeño de los modelos

10.1. Métricas de desempeño de resultados

10.1.1. Área bajo la curva ROC

El área bajo la curva ROC (AUC-ROC) es una métrica fundamental utilizada en la evaluación del rendimiento de un modelo de clasificación. ROC significa "Receiver Operating Characteristic" y es una gráfica que ilustra la capacidad de un clasificador binario a medida que varía su umbral de decisión.

La curva ROC es una representación gráfica que traza la tasa de verdaderos positivos (True Positive Rate, TPR) frente a la tasa de falsos positivos (False Positive Rate, FPR) a diferentes umbrales de clasificación.

- Tasa de verdaderos positivos (TPR) o sensibilidad: = VP / (VP + FN)
- Tasa de falsos positivos (FPR) = FP / (FP + VN)

Donde:

- VP (Verdaderos Positivos) son los casos correctamente identificados como positivos.
- FN (Falsos Negativos) son los casos positivos incorrectamente identificados como negativos.
- FP (Falsos Positivos) son los casos negativos incorrectamente identificados como positivos.
- VN (Verdaderos Negativos) son los casos correctamente identificados como negativos.

El área bajo la curva ROC (AUC) es un valor que cuantifica el desempeño global del modelo de clasificación. Este valor varía entre 0 y 1, donde:

- AUC = 1: Indica un modelo perfecto que clasifica correctamente todas las instancias.
- AUC = 0.5: Indica un modelo que no tiene capacidad de discriminación entre las clases (similar a una clasificación aleatoria).
- AUC < 0.5: Indica un modelo con desempeño peor que la clasificación aleatoria, lo que generalmente sugiere que el modelo está invertido en su capacidad de clasificación.

Un valor más alto de AUC indica un mejor rendimiento del modelo y permite comparar múltiples modelos de clasificación independientemente de sus umbrales de decisión.

10.1.2. F1-Score

El F1-Score es una métrica de desempeño utilizada en problemas de clasificación binaria. Es especialmente útil cuando hay un desbalance significativo entre las clases, lo que significa que una clase es mucho más frecuente que la otra. El F1-Score es la media armónica de la precisión y el recall (sensibilidad o exhaustividad), y proporciona una única métrica que equilibra ambos aspectos.

- La precisión es la proporción de verdaderos positivos entre todas las predicciones positivas. Mide la exactitud de las predicciones positivas del modelo.
- El recall es la proporción de verdaderos positivos entre todos los casos que son realmente positivos. Mide la capacidad del modelo para encontrar todos los casos positivos.

El F1-Score es la media armónica de la precisión y el recall. La media armónica penaliza los valores extremos más que la media aritmética, lo que significa que un F1-Score alto solo se alcanza si tanto la precisión como el recall son altos.

- Valor Máximo (1): Un F1-Score de 1 indica que el modelo tiene precisión y recall perfectos.
- Valor Mínimo (0): Un F1-Score de 0 indica que el modelo no tiene precisión ni recall.

El F1-Score es particularmente útil en contextos donde existe un desbalance de clases. Cuando una clase es mucho más frecuente que la otra, el F1-Score ayuda a medir el rendimiento de la clase minoritaria.

10.2. Análisis de Resultados

10.2.1. Resultados Modelo Base

El Modelo Base detecta un gran número de casos (10.191), dado que cualquier baja de precio superior al 20% es considerada una anomalía. Esto resulta en un tiempo de revisión muy alto, de aproximadamente 509 horas (más de 21 días de trabajo continuo a 24 horas por día), lo cual es considerablemente ineficiente.

Tabla 5: Resultados Modelo Base

Métrica	Resultado
Modelo	Modelo Base
Anomalías Detectadas	10.191
Tiempo Total de Revisión (minutos)	30.573

10.2.2. Resultados Modelo Regresión Logística

El modelo de Regresión Logística no detecta ninguna anomalía, lo que es problemático porque significa que el modelo no está capturando correctamente los eventos anómalos. Además, un AUC de 0,46 indica que el modelo no tiene una buena capacidad para discriminar entre las clases (anomalía y no anomalía).

Tabla 6: Resultados Modelo Logistic Regression

Métrica	Resultado
Modelo	Logistic Regression
AUC	0,4602
Anomalías Detectadas	0
Tiempo Total de Revisión (minutos)	0
Precision	1
Recall	0
F1-Score	0

10.2.3. Resultados Modelo Random Forest

El modelo de Random Forest tiene un AUC significativamente más alto (0,83), lo que indica un mejor rendimiento en la discriminación entre clases. Detecta 11 anomalías, lo que resulta en un tiempo de revisión de solo 33 minutos. Aunque su precisión y recall son bajos, el modelo es más práctico en términos de tiempo de revisión.

Tabla 7: Resultados Modelo Random Forest

Métrica	Resultado
Modelo	Random Forest
AUC	0,8301
Anomalías Detectadas	11
Tiempo Total de Revisión (minutos)	33
Precision	0,45
Recall	0,05
F1-Score	0,09

10.2.4. Resultados Modelo Isolation Forest

El modelo de Isolation Forest detecta 772 anomalías, lo que resulta en un tiempo de revisión de 2316 minutos (aproximadamente 38,6 horas). Sin embargo, con un AUC de 0,465, su capacidad de discriminación es pobre, y las métricas de precisión, recall y F1-Score para la clase de anomalías son 0, indicando que el modelo no está haciendo un buen trabajo en identificar correctamente las las variaciones de precio.

Tabla 8: Resultados Modelo Isolation Forest

Métrica	Resultado
Modelo	Isolation Forest
AUC	0,4651
Anomalías Detectadas	772
Tiempo Total de Revisión (minutos)	2316
Precision	0
Recall	0
F1-Score	0

10.3 . Conclusión y Recomendación

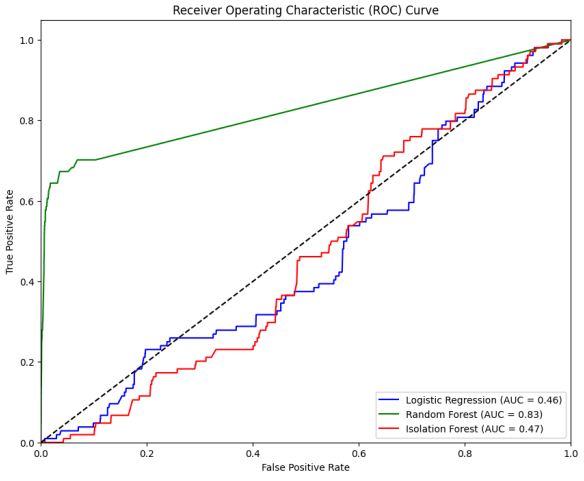
A continuación se hace una recopilación de los resultados de los modelos:

Tabla 9: Comparación de Resultados de Modelos

Métrica	Resultado			
Modelo	**Modelo Base**	**Logistic Regression**	**Random Forest**	**Isolation Forest**
Anomalías Detectadas	10.191	0	11	772
Tiempo Total de Revisión (minutos)	30.573	0	33	2.316
AUC	-	0,4602	0,8301	0,4651
Precision	-	1	0,45	0
Recall	-	0	0,05	0
F1-Score	-	0	0,09	0

La curva ROC es una representación gráfica de la sensibilidad (tasa de verdaderos positivos) frente a especificidad (tasa de falsos positivos):

Figura 9: Curva ROC comparativa entre los modelos



Analicemos ahora cada una de las curvas:

- Random Forest: La curva del Random Forest es la más alejada de la línea diagonal (la línea de azar), lo que indica un buen rendimiento. La AUC de 0,83 sugiere que este modelo es, en términos relativos, el más efectivo para diferenciar entre anomalías y no anomalías.
- Isolation Forest: La curva está cerca de la diagonal, lo que indica un desempeño cercano al azar. La AUC de 0,47 sugiere que el modelo no es eficaz para este problema.
- Logistic Regression: Similar al Isolation Forest, la curva de la Regresión Logística está cerca de la línea diagonal. El AUC de 0,46 también indica un rendimiento pobre, apenas mejor que el azar.

El mejor modelo de acuerdo a estos resultados es el modelo de Random Forest con AUC de 0,83. Respecto al tiempo de revisión detecta 11 anomalías, lo que resulta en un tiempo de revisión de 33 minutos.

Revisemos ahora el F1-Score, que es la media armónica de la precisión y el recall, y proporciona una medida de la exactitud del modelo en términos de la clasificación de las clases positivas.

Analicemos los resultados:

- Random Forest: La puntuación F1 de 0,09, aunque muy baja, es superior en comparación con los otros dos modelos.
- Logistic Regression e Isolation Forest: Ambos modelos tienen una puntuación F1 de 0. Esto indica que no lograron detectar correctamente las anomalías o que clasificaron todos los casos como negativos, resultando en una baja efectividad.

En conclusión para una empresa de ecommerce que enfrenta problemas de detección de anomalías en precios, teniendo en cuenta estos resultados, se vuelve evidente la necesidad de realizar mejoras sobre el preprocesamiento de los datos y la aplicación de los modelos. Por lo tanto, pasamos a una nueva sección donde intentaremos aplicar estas mejoras.

11. Iteración y mejora de los modelos

11.1. Descripción general

Dadas las grandes oportunidades sobre el enfoque inicial aplicado, se pasan a aplicar nuevas capas de pre procesamiento de datos sobre el data frame original y la aplicación de los modelos estadísticos.

11.2. Mejoras sobre el pre procesamiento

Teniendo en cuenta los resultados anteriores, se procede a iterar el análisis trabajando sobre los datos iniciales. Se realizaron los siguientes pasos:

11.2.1. Iteración sobre la base de datos en general

1. Para comenzar, se calculan las variaciones de precio para todos los items.

2. Dado que ensucian el análisis con filas sin cambios de precio, se eliminan los registros donde cambiaba la fecha pero el precio era el mismo para fechas seguidas para un mismo item id.

3. Dado que estamos analizando cambios de precio a la baja, se eliminan del análisis las filas donde se produce un cambio de precio con una variación mayor o igual a cero, quedándonos únicamente con las variaciones de precio con signo negativo.

11.2.2. Iteración de la variable FEEDBACK

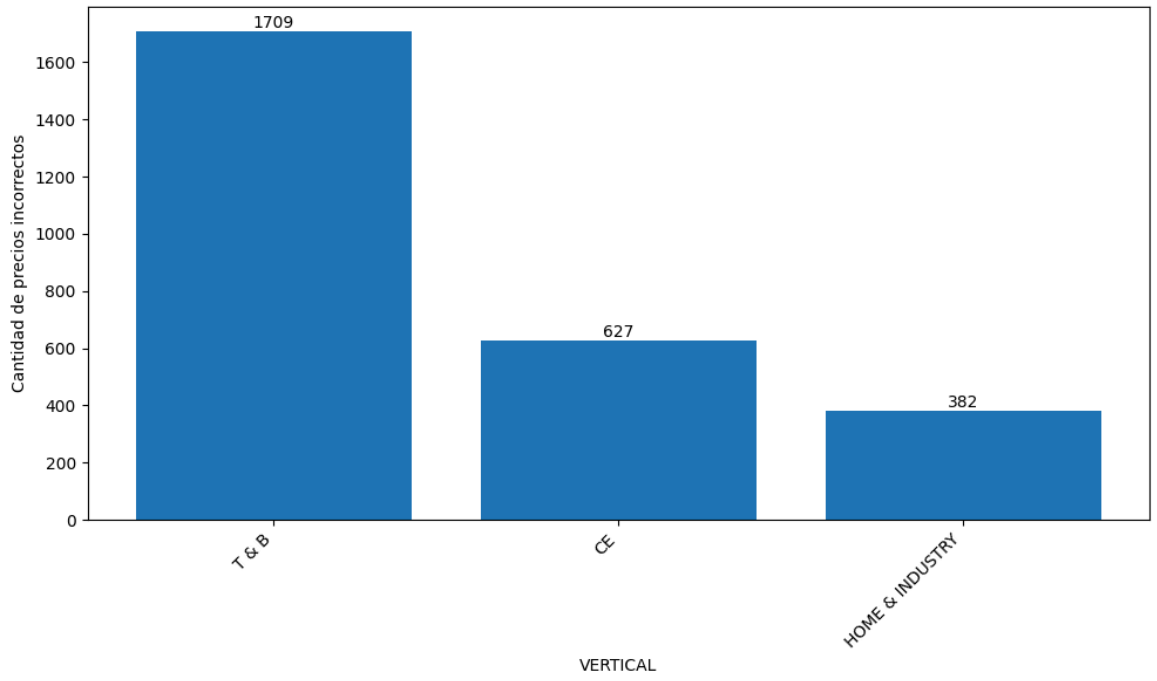
1. Se detectaron casos donde el feedback humano consideraba como incorrecto casos donde el precio había subido respecto a su precio anterior, lo cual es lógicamente inconsistente. Para intentar mejorar las performance de los modelos, se eliminan estos casos de consideración.

2. Adicionalmente, para reflejar cierta lógica que se ve en la práctica comercial y sumar más casos para poder entrenar los modelos, se genera una nueva columna llamada FEEDBACK_V2, en la cual:

    a. Se asigna un valor de “1” en los casos en los que el feedback humano etiquetó un precio surgido de una variación a la baja.

b. Para sumar una mayor cantidad de datos etiquetados como anómalos, consideramos que se pueden tomar medidas adicionales. En línea con esto, en esta columna, todo precio originado de una variación a la baja mayor al 50% se considera como una anomalía y se etiqueta con un “1”. Esta decisión se basa en criterios estándar de la industria del ecommerce, en la cual se presta especial atención a las variaciones de precio de esta magnitud.

Figura 10: Cantidad de precios etiquetados como incorrectos por VERTICAL post procesamiento



Las acciones descritas en estos puntos ayudan a reducir la cantidad de casos no clasificados y aumentar la cantidad de precios etiquetados como incorrectos, con la esperanza de mejorar la performance de los modelos aplicados a la detección de anomalías.

11.2.3. Incorporación de nuevas variables

1. Se calculan como nuevas variables para sumar al análisis el promedio de precio de los últimos “x” días, donde “x” = 2, 5, 7, 14, 21 y 28.

2. Se calcula la diferencia porcentual entre el precio promedio de los últimos “x” días, donde “x” = 2, 5, 7, 14, 21 y 28, y el precio al momento de la evaluación. Esto se hace con el racional de poder comparar el precio en revisión con cada uno de los promedios calculados, dando mayor contexto a los modelos para evaluar si el precio actual es correcto. Precios

con una variación muy grande a la baja respecto al promedio de períodos anteriores deberían indicar una anomalía de precio.

11.2.4. Nueva descripción de la base de datos pre procesados

Como resultado, la base de datos queda constituida con las siguientes columnas:

- ITEM_ID: id identificador único de cada ítem que se vende.
- PRICE: es el precio de un ITEM_ID en cierta fecha.
- DATE: es la fecha de cada cambio de precio (entre 2024-01-01 y 2024-03-31).
- KEY_PRICE: es un campo que concatena DATE, ITEM_ID y PRICE.
- variacion_pct: la variación porcentual de precio respecto al registro anterior para cada ITEM_ID.
- FEEDBACK_V2: 1 si un precio fue etiquetado como incorrecto por una persona del equipo que se encarga de revisar anomalías de precio, 0 si el precio no fue etiquetado ni como correcto ni como incorrecto.
- SIT_SITE_ID: país al que pertenece el ítem.
- VERTICAL: categoría a la que pertenece un ITEM_ID.
- avg_2d: es el promedio de precios de los últimos 2 días para el ITEM_ID.
- avg_5d: es el promedio de precios de los últimos 5 días para el ITEM_ID.
- avg_7d: es el promedio de precios de los últimos 7 días para el ITEM_ID.
- avg_14d: es el promedio de precios de los últimos 14 días para el ITEM_ID.
- avg_21d: es el promedio de precios de los últimos 21 días para el ITEM_ID.
- avg_28d: es el promedio de precios de los últimos 28 días para el ITEM_ID.
- diff_pct_2d: es la diferencia porcentual entre PRICE y avg_2d.
- diff_pct_5d: es la diferencia porcentual entre PRICE y avg_2d.
- diff_pct_7d: es la diferencia porcentual entre PRICE y avg_7d.
- diff_pct_14d: es la diferencia porcentual entre PRICE y avg_14d.
- diff_pct_21d: es la diferencia porcentual entre PRICE y avg_21d.
- diff_pct_28d: es la diferencia porcentual entre PRICE y avg_28d.

11.3. Mejoras en la aplicación de los modelos

Como primera medida, teniendo en cuenta las dificultades que se tuvieron anteriormente para procesar la cantidad de datos necesarios, se ha migrado el código que se utilizó para correr estos modelos de un entorno local utilizando Jupyter Notebooks a un entorno “cloud”, con mayor capacidad de procesamiento, utilizando Google Colab. Esto permitió realizar iteraciones sobre el código exponencialmente más rápido y reducir el tiempo de procesamiento.

Adicionalmente, se han aplicado técnicas para afrontar el problema de tener relativamente pocas anomalías etiquetadas por el feedback humano y manejar el desbalance de clases. Se ha modificado la función de pérdida del clasificador para penalizar más los errores sobre la clase minoritaria, aplicando la técnica conocida como "cost-sensitive learning". De esta forma, el algoritmo otorga mayor peso a cada muestra anómala. Utilizando la librería scikit-learn, se ajusta el parámetro class_wight para que sea igual a "balanced" y aplique sobre el conjunto de entrenamiento. Esta técnica es recomendable en casos como este donde el volumen de datos es muy elevado. Con esta estrategia, los modelos supervisados como Random Forest y Regresión Logística mejoran significativamente su performance.

11.4. Comparación y evaluación de desempeño de los modelos con las mejoras aplicadas

Comencemos ahora por evaluar la performance de los modelos para la vertical "T & B". Obtenemos una mejora en todas las métricas dado que redujimos la base de casos no clasificados y aumentamos la cantidad de precios etiquetados como incorrectos. Recordemos que suponemos que revisar cada anomalía lleva 3 minutos.

Tabla 10: Resultados de los modelos para la Vertical "T & B"

Resultados Vertical: T & B	Precision	Recall	F1	AUC	Cantidad de Anomalías Detectadas	Tiempo de Revisión Total (minutos)	Tiempo de Revisión Total (horas)	Diferencia absoluta sobre el Modelo Base	Diferencia porcentual sobre el Modelo Base
Modelo Base	-	-	-	-	4.894	14.682	244,70	-	-
Logistic Regression	0,5510	0,8367	0,6444	0,9112	679	2.037	33,95	-210,75	-86,13%
Random Forest	0,6914	0,8130	0,7321	0,9024	526	1.578	26,30	-218,40	-89,25%
Isolation Forest	0,0196	1,0000	0,0383	0,5030	21.108	63.324	1.055,40	810,70	331,30%

Puntos destacados para la Vertical "T & B":

- Random Forest resulta de nuevo el más equilibrado, manteniendo una buena combinación de precisión y recall, lo que se refleja en el F1 score de 0.73.

- Logistic Regression es moderado, pero inferior a Random Forest en estabilidad global.

- Isolation Forest logra detectar el 100% de las anomalías (recall perfecto), pero a costa de una precisión prácticamente nula, haciendo que la mayoría de las detecciones sean falsos positivos.

Comparando estos resultados con los anteriores obtenemos una clara mejora en la performance de todos los modelos. Pasemos ahora a revisar la performance de estos modelos en otras verticales:

Tabla 11: Resultados de los modelos para la Vertical "CE"

Resultados Vertical: CE	Precision	Recall	F1	AUC	Cantidad de Anomalías Detectadas	Tiempo de Revisión Total (minutos)	Tiempo de Revisión Total (horas)	Diferencia absoluta sobre el Modelo Base	Diferencia porcentual sobre el Modelo Base
Modelo Base	-	-	-	-	2.061	6.183	103,05	-	-
Logistic Regression	0,6220	0,8703	0,7033	0,9307	198	594	9,90	-93,15	-90,39%
Random Forest	0,8478	0,8659	0,8464	0,9317	135	405	6,75	-96,30	-93,45%
Isolation Forest	0,0886	0,8780	0,1552	0,8338	2.024	6.072	101,20	-1,85	-1,80%

Puntos destacados para la Vertical "CE":

- Random Forest destaca por tener la mayor precisión y un recall alto, lo que se traduce en un F1 score equilibrado de 0.85 y AUC alta, de 0.93.

- Logistic Regression ofrece un rendimiento razonable, pero inferior en precisión y F1 en comparación con Random Forest.

- Isolation Forest logra un recall alto, es decir, casi detecta todas las anomalías, pero su precisión muy baja indica que se generan demasiados falsos positivos, haciendo poco útil la detección.

Tabla 12: Resultados de los modelos para la Vertical "HOME & INDUSTRY"

Resultados Vertical: HOME & INDUSTRY	Precision	Recall	F1	AUC	Cantidad de Anomalías Detectadas	Tiempo de Revisión Total (minutos)	Tiempo de Revisión Total (horas)	Diferencia absoluta sobre el Modelo Base	Diferencia porcentual sobre el Modelo Base
Modelo Base	-	-	-	-	1.288	3.864	64,40	-	-
Logistic Regression	0,5634	0,8259	0,6372	0,9064	143	429	7,15	-57,25	-88,90%
Random Forest	0,6674	0,7746	0,6991	0,8844	100	300	5,00	-59,40	-92,24%
Isolation Forest	0,0312	1,0000	0,0600	0,7108	3.234	9.702	161,70	97,30	151,09%

Puntos destacados para la Vertical "HOME & INDUSTRY":

- En este caso, Random Forest sigue ofreciendo el mejor balance, con un F1 de 0.70 y una precisión y recall aceptables.
- Logistic Regression vuelve a mostrar un comportamiento similar al de verticales anteriores, aunque con métricas algo inferiores.
- Isolation Forest mantiene su patrón de recall perfecto o cercano a 1 a costa de una precisión extremadamente baja.

Tabla 13: Resultados de los modelos para la Vertical "ENTERTAINMENT"

Resultados Vertical: ENTERTAINMENT	Precision	Recall	F1	AUC	Cantidad de Anomalías Detectadas	Tiempo de Revisión Total (minutos)	Tiempo de Revisión Total (horas)	Diferencia absoluta sobre el Modelo Base	Diferencia porcentual sobre el Modelo Base
Modelo Base	-	-	-	-	66	198	3,30	-	-
Logistic Regression	0,6875	0,6875	0,6667	0,8344	16	48	0,80	-2,50	-75,76%
Random Forest	0,6500	0,8750	0,7083	0,9054	26	78	1,30	-2,00	-60,61%
Isolation Forest	0,5810	1,0000	0,6728	0,9236	46	138	2,30	-1,00	-30,30%

Puntos destacados para la Vertical "ENTERTAINMENT":

- En esta vertical, Random Forest muestra un excelente balance con el F1 más alto (0.71) y un AUC de 0.91, lo que lo hace adecuado para este dominio.

- Isolation Forest logra detectar todas las anomalías (recall perfecto) pero su precisión menor puede implicar un número importante de falsos positivos.

- Logistic Regression tiene un desempeño similar a Isolation Forest en f1, pero sin destacar en algún aspecto concreto.

11.5. Conclusión

Revisemos el desempeño de cada uno de los modelos para llegar a una conclusión respecto a cuál es más conveniente utilizar.

El modelo de Isolation Forest, tiende a detectar la mayoría o todas las anomalías con un recall cercano a 100% en todas las verticales. Sin embargo su precisión es muy baja, lo que sugiere que aunque no se pierden anomalías, se incluyen muchos falsos positivos. Esto reduce significativamente el F1 score, lo que puede ser problemático en aplicaciones donde revisar falsos positivos es costoso, como es en este caso.

El modelo de Logistic Regression ofrece resultados aceptables pero consistentemente inferiores a los de RandomForest en términos de precisión y F1 score.

El modelo de Random Forest, muestra un equilibrio óptimo entre precisión y recall en la mayoría de las verticales, lo que se traduce en F1 scores y AUC altos. La detección es moderada en cantidad, lo que sugiere que no se generan alarmas excesivas.

Teniendo en cuenta todo lo anterior, y tomando una decisión en términos relativos con respecto a los otros dos modelos podemos destacar al modelo de Random Forest como una buena opción para aplicar a este problema. Pasaremos ahora a revisar si es posible mejorar estos resultados utilizando el método de Conformal Prediction.

12. Conformal Prediction

12.1. Descripción general

La detección de anomalías de precio es esencial en el comercio electrónico para identificar errores, fraudes o eventos anormales. El método de Conformal Prediction ofrece una metodología robusta para abordar esta tarea mediante el control de la tasa de falsos positivos, permitiendo así la detección de valores atípicos sin necesidad de datos etiquetados.

Conformal Prediction es un método que otorga garantías de cobertura sobre las predicciones que hace un modelo. Es una metodología que transforma las predicciones puntuales de un modelo de aprendizaje automático en regiones de predicción, ofreciendo una garantía probabilística de que estas regiones contienen el valor verdadero del resultado.

12.2. Aplicación del método

Para aplicar este método en Python se utiliza la librería “MAPIE” como se destaca en el texto de Christoph Molnar (2023)[4]. Para aplicar este método, se siguen los siguientes pasos:

1. Preparación de Datos:
    a. Entrenamiento del Modelo: Entrena un modelo de clasificación o regresión (en este caso, Random Forest) con los datos históricos de precios y de precios erróneos etiquetados.
    b. División de Datos: Dividir los datos en conjuntos de entrenamiento, calibración y prueba. Es crucial que el conjunto de calibración sea independiente del conjunto de entrenamiento para evitar sobreajuste y garantizar una evaluación adecuada.

2. Cálculo de Score de No Conformidad:
    a. Utilizar un conjunto de datos de calibración (diferente del conjunto de entrenamiento) para calcular los "non-conformity scores" (‘s_i’) que miden qué tan inusuales son las predicciones del modelo para estos datos. Estos puntajes se ordenan y se utiliza un cuantil para determinar el umbral de predicción.

3. Cálculo del Umbral (‘q’):
    a. Determinar el umbral ‘q’ que cubra el ‘1 - α’ (nivel de confianza) de las puntuaciones de no conformidad. Este umbral se usa para definir la región de predicción,

[4] Molnar, C., (2023). Introduction To Conformal Prediction With Python: A Short Guide For Quantifying Uncertainty Of Machine Learning Models.

asegurando que la cobertura promedio sea al menos ‘1 - α’. Es posible ajustar ‘α’ para controlar el trade-off entre cobertura y tamaño del conjunto de predicción. Valores más bajos de ‘α’ aumentan la cobertura pero pueden resultar en conjuntos de predicción más grandes.

4. Predicción y Evaluación:
    a. Para nuevos datos, se calculan los "non-conformity scores" y se generan las regiones de predicción asegurando que el umbral ‘q’ determinado en la calibración se mantenga, proporcionando así la garantía de cobertura deseada.
    b. Las observaciones fuera del conjunto de predicción se consideran anomalías (precios que bajaron más de lo esperado).

Por otro lado, es importante definir las métricas que se deben observar para evaluar el desempeño del método de Conformal Prediction:

- Cobertura: Porcentaje de predicciones que incluyen el valor verdadero en el conjunto de predicción. Esto indica la fiabilidad del método para cubrir la verdadera etiqueta o valor en su intervalo de predicción.
- Tamaño Medio del Conjunto de Predicción: Indica cuántas posibles etiquetas o valores se incluyen en promedio en los conjuntos de predicción. Menores tamaños indican una mayor precisión del modelo.
- Exactitud: Proporción de predicciones correctas. Aunque el método de Conformal Prediction se enfoca más en la cobertura, la exactitud puede proporcionar información adicional sobre la calidad general del modelo.
- Matriz de Confusión: Proporciona un desglose de verdaderos positivos, falsos positivos, verdaderos negativos y falsos negativos, ayudando a identificar posibles problemas en la detección de anomalías.

12.3. Resultados de aplicar Conformal Prediction

Partiendo del modelo de Random Forest, avancemos a aplicar el método de Conformal Prediction para la vertical de “T & B” con un alpha = 0,005, siguiendo los pasos de la sección anterior. Tras correr el modelo, obtenemos los siguientes resultados que analizaremos en detalle:

- Cobertura: 98,72%
  La cobertura indica que en el 98,72% de los casos, el conjunto de predicción generado por Conformal Prediction incluye la clase verdadera. Este valor sugiere que el modelo es

bastante confiable en incluir la etiqueta correcta dentro de su conjunto de predicción, lo cual es crucial para minimizar los falsos negativos. De las 436 muestras anómalas reales, en el 98,72% de los casos la etiqueta correcta aparece dentro del conjunto de predicción conformal, lo cual está muy cerca del 99,5% esperado por 1-alpha.

- Tamaño medio del conjunto de predicción: 0,99
  Un tamaño medio del conjunto de predicción de 0,99 indica que en promedio el conjunto de predicción incluye menos de una clase, lo cual es bastante preciso. Un valor cercano a 1 indica un alto nivel de confianza en las predicciones.

- Matriz de confusión:

  [[20785; 0]
  [107; 329]]

  - TN = 20785 (Verdaderos Negativos): El número de casos correctamente clasificados como no anomalías es muy alto, lo que indica que el modelo es eficaz en identificar los precios normales.
  - FP = 0: No se han cometido errores al identificar precios normales como anomalías.
  - FN = 107: El modelo no detectó 107 casos que eran anomalías reales de acuerdo con los datos etiquetados
  - TP = 329 (Verdaderos Positivos): Al aplicar Conformal Prediction fue posible detectar 329 de las 436 anomalías.

- Exactitud:
  (TN + TP) / (total) = (20785 + 329) / 21221 = 99,50%. Esta exactitud del 99,50% refleja un alto nivel de precisión general del modelo, indicando que la gran mayoría de las predicciones son correctas.

- Precisión:
  TP / (TP + FP) = 329 / 329 = 100%. Cada vez que CP predice “anómalo”, acierta.

- Recall (tasa de TP):
  TP / (TP + FN) = 329 / (329+107) = 75%.

- Especificidad = 100 %
  TN / (TN + FP) = 20785 / 20785 = 100%. Ningún caso sano es marcado erróneamente.

En conclusión, con alpha = 0,005, aplicar Conformal Prediction, entrega una cobertura muy alta, mantiene un solo label casi siempre, tiene 0 falsos positivos, 100% de precisión y alcanza una exactitud global de 99,50%, ofreciendo un buen equilibrio entre garantía de cobertura y rendimiento operativo.

Iteremos ahora con distintos valores de alpha para calibrar el método, donde alpha = {0,2; 0,15; 0,1; 0,05; 0,03; 0,02; 0,01; 0,005; 0,001}.

Tabla 14: Iteración del parámetro alpha para Conformal Prediction.

Número de Prueba	1	2	3	4	5	6	7	8	9
Alpha	0,2	0,15	0,1	0,05	0,03	0,02	0,01	0,005	0,001
Cobertura (CP)	88.16%	88.16%	88.16%	93.09%	95.09%	96.10%	97.33%	98.72%	99.50%
Tamaño medio (CP)	0.88	0.88	0.88	0.93	0.95	0.96	0.98	0.99	1.00
Matriz de confusión	[[20785; 0] [337; 99]]	[[20785; 0] [337; 99]]	[[20785; 0] [337; 99]]	[[20785; 0] [234; 202]]	[[20785; 0] [178; 258]]	[[20785; 0] [140; 296]]	[[20785; 0] [123; 313]]	[[20785; 0] [107; 329]]	[[20785; 0] [106; 330]]
Verdaderos Negativos (TN)	20785	20785	20785	20785	20785	20785	20785	20785	20785
Falsos Positivos (FP)	0	0	0	0	0	0	0	0	0
Falsos Negativos (FN)	337	337	337	234	178	140	123	107	106
Verdaderos Positivos (TP)	99	99	99	202	258	296	313	329	330
Exactitud global	98.41%	98.41%	98.41%	98.90%	99.16%	99.34%	99.42%	99.50%	99.50%
Precisión	100%	100%	100%	100%	100%	100%	100%	100%	100%
Recall (tasa de TP)	23%	23%	23%	46%	59%	68%	72%	75%	76%
Especificidad (tasa TN)	100%	100%	100%	100%	100%	100%	100%	100%	100%

En esta tabla se puede observar que la cobertura crece monótonamente al disminuir alpha, pasando de 88,16% con alpha mayor o igual a 0,1 hasta 99,5% con alpha igual a 0,001.

Para poner estos números en perspectiva, para una empresa de ecommerce para la cual cada persona del equipo comercial dedica 3 minutos en revisar cada alerta, es conveniente traducir la cantidad de anomalías detectadas de cada configuración del método de Conformal Prediction a minutos de trabajo humano.

Tabla 15: Impacto operativo de la iteración de alpha en Conformal Prediction

Alpha	Anomalías Detectadas	Tiempo total (min)	Tiempo total (hs)	Alertas por día (≈ 31 días)	Tiempo por día (min)	Recall (TP/(TP+FN))
0,2	99	297 min	5,0 hs	3,19	10 min	23 %
0,1	99	297 min	5,0 hs	3,19	10 min	23 %
0,05	202	606 min	10,1 hs	6,52	20 min	46 %
0,03	258	774 min	12,9 hs	8,32	25 min	59 %
0,02	296	888 min	14,8 hs	9,55	29 min	68 %
0,01	313	939 min	15,7 hs	10,10	30 min	72 %
0,005	329	987 min	16,5 hs	10,61	32 min	75 %
0,001	330	990 min	16,5 hs	10,65	32 min	76 %

A partir de la tabla anterior se puede interpretar la relación entre recall y tiempo dedicado por el equipo comercial a la revisión de anomalías:

- A menor alpha se generan más alertas, aumentando el recall, y generando mayor carga en tiempo de revisión.

- A mayor alpha se generan menos alertas, lo que provoca un ahorro de tiempo a riesgo de un menor recall.

De esta forma, es posible adaptar la configuración al tiempo del equipo comercial de una empresa de ecommerce que se quiera dedicar a este problema. Para concluir, podemos decir que la iteración de alpha demuestra que el método de Conformal Prediction permite ajustar el compromiso o trade off entre la garantía de cobertura (incorporar la etiqueta real), la capacidad de detección (recall) y la operacionalidad (número de alertas generadas).

Para un balance óptimo en detección de anomalías de precio, es recomendable un alpha en el rango entre 0,005 y 0,01, que sitúa la cobertura entre 97% y 99% y el recall entre 72% y 75%, sin generar falsos positivos y con conjuntos de predicción casi siempre unitarios, facilitando su adopción en el caso de una empresa de ecommerce que enfrente el problema de calibrar un sistema de detección de anomalías de precio.

13. Conclusiones finales

En este trabajo se evaluaron diversos modelos para detectar anomalías de precio en ecommerce. destacando el modelo Random Forest como la opción más robusta, con un AUC de hasta 0,93 con un F1 Score de hasta 0,85, dependiendo de la vertical en la que se implemente. Los resultados obtenidos demuestran que la aplicación de técnicas avanzadas de machine learning puede reducir significativamente la cantidad de falsos positivos y, por ende, el tiempo de revisión manual en comparación con modelos más simples, lo cual es crucial para las empresas que manejan grandes volúmenes de datos e items.

Un aspecto a destacar de este trabajo fue la implementación del método de Conformal Prediction sobre el modelo de Random Forest. Este enfoque permitió cuantificar la incertidumbre de las predicciones, proporcionando intervalos de predicción con garantías de cobertura. Esto ofreció una herramienta valiosa para manejar la variabilidad en los datos de precios.

A pesar de esto, nos enfrentamos a la dificultad de la naturaleza de este problema, donde contamos con una gran cantidad de datos de cambios de precio, pero relativamente muy pocos casos de precios etiquetados como anómalos a partir de la revisión de miembros del equipo comercial. Este es el punto central en la complejidad de este trabajo. Quedaron por fuera del alcance de esta tesis varios puntos debido a limitaciones en los datos disponibles, ya sea por la extensión en tiempo de los mismos o por la falta de variables adicionales que serían de gran utilidad.

Puede ser de gran valor para obtener una solución más sofisticada de este problema analizar períodos más largos para tener en cuenta el impacto de eventos especiales anuales como por ejemplo Black Friday o Amazon Prime Day, donde se realizan gran cantidad de ofertas de forma generalizada en todos los sitios de ecommerce y durante los cuales se produce la mayor cantidad de ventas de cada año. Se recomienda realizar este análisis para un periodo de varios años, para tener en cuenta en los modelos como se comportan los precios en contextos de alta demanda por eventos promocionales y estacionales para capturar patrones más complejos en la variación de precios.

Otro punto a considerar que pudo haber afectado al análisis es la limitación de no contar en los datos con la capacidad de identificar a la persona que ha etiquetado cada precio como una anomalía en los datos de base. Podría ser posible que la misma persona responda con diferentes respuestas sobre si un precio es correcto o no, o que el criterio entre distintas personas para tomar

esta decisión sea incompatible. Profundizar en esta variable podría ser de gran valor para este problema en análisis futuros.

Un factor fundamental a considerar para seleccionar un modelo para esta tarea es el costo computacional. Aunque el modelo de Random Forest corrió en un tiempo razonable para un grupo reducido de ítems, por fuera del alcance de esta tesis se encuentra evaluar la factibilidad técnica de implementar este modelo para la revisión de anomalías en tiempo real a gran escala. Alternativas a explorar ante esta problemática podrían ser modelos basados en medidas estadísticas más simples como IQR (interquartile range) o Z score detection.

El desafío es encontrar el balance y la sensibilidad, entre el control para asegurar la seguridad del ecosistema de precios y la flexibilidad para no sobrecargar a los equipos de tareas repetitivas e innecesarias. La capacidad de adaptarse a los cambios del mercado y mejorar continuamente los sistemas de detección de anomalías es clave para lograr consolidar la confianza de los clientes en los sitios de ecommerce.

13. Bibliografía

1. Lei, J., G'Sell M., Rinaldo A., Tibshirani R., Wasserman L. (2017). Distribution-Free Predictive Inference For Regression. Department of Statistics, Carnegie Mellon University. *Descripción: Marco general para la inferencia predictiva utilizando Conformal Inference.*

2. Ramakrishnan, J., Li C., Shaabani, E., Sustik, M. (2019). Anomaly Detection for an E-commerce Pricing System. Walmart Labs. *Descripción: Análisis de modelos de detección de anomalías de precios desarrollados e implementados a gran escala para el sistema de ecommerce de Walmart.*

3. Tinawi, I. (2019). Machine Learning for Time Series Anomaly Detection. Massachusetts Institute of Technology. *Descripción: Análisis de diversos modelos para la detección de anomalías de sensores satelitales en base a sus patrones históricos.*

4. Vovk, V., Gammerman, A., & Shafer, G. (2022). Algorithmic Learning in a Random World, Second Edition. Springer International Publishing. *Descripción: Primer libro en desarrollar la metodología de Conformal Prediction.*

5. Molnar, C., (2023). Introduction To Conformal Prediction With Python: A Short Guide For Quantifying Uncertainty Of Machine Learning Models. *Descripción: El libro introduce Conformal Prediction como una técnica esencial para cuantificar la incertidumbre en modelos de aprendizaje automático.*

6. Tibshirani, R. (2023). Conformal Prediction, Advanced Topics in Statistical Learning. *Descripción: Desarrolla Conformal prediction como un marco relativamente nuevo para cuantificar la incertidumbre en las predicciones hechas por algoritmos de predicción.*

7. Downey, A., Elkner J., & Meyers, C. (2009) Introducción a la Programación con Python. *Descripción: Libro usado de referencia para realizar los análisis exploratorios.*

8. Fry B. (2007), Visualizing Data, Exploring and Explaining Data with the Processing Environment, O'Reilly Media. *Descripción: Libro usado de referencia para realizar los gráficos.*

9. Heumann, C., & Schomaker, M. (2016). Introduction to statistics and data analysis. Springer International Publishing Switzerland. *Descripción: Libro usado de referencia respecto a estadística.*

10. Angelopoulos, A., & Bates, S. (2022) A Gentle Introduction to Conformal Prediction and Distribution-Free Uncertainty Quantification. *Descripción: Explicación metodológica y con ejemplos de la aplicación de Conformal Prediction.*

11. James, G., Witten, D., Hastie, T., Tibshirani, R., & Taylor, J. (2023). An Introduction to Statistical Learning: with Applications in Python. Springer Nature. *Descripción: Este libro presenta algunas de las técnicas de modelado y predicción más importantes, junto con sus aplicaciones relevantes. Los temas incluyen regresión lineal, clasificación, métodos de remuestreo, enfoques de contracción, métodos basados en árboles, máquinas de vectores de soporte, agrupación, aprendizaje profundo, análisis de supervivencia, pruebas múltiples y más.*

12. Liu, F.,Ming Ting, K., & Zhou, Z. (2008). Isolation Forest. International Conference on Data Mining (ICDM ’08). *Descripción: Marco general del método de Isolation Forest.*