



82.02 - Proyecto Final Analítica

Forecast de una mayorista y minorista de tecnologías

Alumnos:

María Felicitas Baronio - 59355

Ezequiel Shinzato - 61629

Manuel Parodi - 61517

OBJETIVO

Una empresa vendedora de artículos de tecnología (línea blanca y línea marrón) se acerca a nosotros con la necesidad de generar un forecast de ventas que estime las mismas de forma mensual, con el fin de lograr una planificación más óptima de su actividad y sus recursos. El mismo expresa que el método de forecasting actual se basa en números acordados entre los directores comerciales (quienes transmiten la estrategia comercial de la empresa) y los product managers (quienes conocen el mercado y administran el producto). El método actual se basa mayormente en las suposiciones de estos actores y no en datos empíricos.

Nuestro objetivo es generar un forecast de ventas más preciso para los principales grupos de productos de su empresa, utilizando un enfoque basado en datos empíricos, que logre resultados más precisos que los métodos actuales. Este pronóstico de ventas debería ser capaz de ayudar a la planificación estratégica anual de la empresa y tomar decisiones mejor informadas sobre la asignación de recursos, acorde a la previsión de demanda de cada grupo de productos.

MÉTRICAS Y KPIs

Para evaluar si nuestros modelos están brindando estimaciones precisas y efectivas hay diversas métricas y KPIs que podríamos utilizar para medirlo.

Algunas de las métricas que utilizaremos serán las siguientes:

MAPE → compararemos el error porcentual medio absoluto de los mismos con el del forecast vigente en la empresa para los meses transcurridos del corriente año (YTD). La elección de esta métrica se basa en que brinda la posibilidad de comparar no solo entre modelos de un mismo producto y canal sino también entre ellos, ya que su unidad se refleja de forma porcentual. Esta métrica se definirá como:

$$MAPE = \frac{100}{N} \times \sum_{i=1}^N \left| \frac{x_i - \hat{x}_i}{x_i} \right|$$

Costo de inventario → Para observar si el forecast está teniendo un impacto positivo en la empresa, se calculará esta métrica (tanto para el forecast actual como para el que generemos) que representa el costo de los productos no vendidos por sobrestimar la demanda. Esta se calculará como $CI = \sum (C_e + C_o * t)$ en un período de tiempo determinado, donde C_e es el costo efectivo del producto no vendido (valor por el que lo he comprado + costos indirectos), C_o es el costo de oportunidad del mismo y t es el período de tiempo.

Costo de subestimación → Por último, con ayuda de información del mercado provista por el cliente, calcularemos para ambos métodos de forecasting el costo de subestimación de demanda, el cual se basará en una investigación de mercado que completará la fórmula: $CS = \sum (R_{vp} + C_{ae}) + C_p$ donde R_{vp} es la rentabilidad que hubieran significado las ventas perdidas, C_{ae} es el costo de adquisición de emergencia por salir a comprar sin la planificación adecuada el producto que hayamos subestimado y C_p es una valuación de los clientes perdidos como consecuencia de que no hemos podido suplir su demanda.

Además de esta métrica, también es importante tener en cuenta otros factores clave que pueden afectar la precisión del forecast, como la estacionalidad y las tendencias del mercado, entre otros. Al monitorear y evaluar estas métricas durante el desarrollo del forecast, podremos identificar oportunidades para mejorar el pronóstico de ventas y ajustar la estrategia en consecuencia.

ENTREGABLES Y OUTPUTS

Durante el proyecto se documentará todos los procedimientos y herramientas utilizadas, con el fin de registrar la metodología y sus resultados de forma replicable. Se determinará una entrega parcial durante el proyecto donde se comunicarán los resultados del análisis exploratorio de los datos a utilizar.

A su vez, se entregará un informe del forecast con el resultado final del proyecto el cual incluirá todas las predicciones de ventas para los principales grupos de productos que decidimos analizar para la empresa. El informe incluirá información sobre el modelo de forecast utilizado, los supuestos y limitaciones que encontramos, las métricas y los KPIs que evaluamos, y si encontráramos, recomendaciones o metodologías que podrían utilizarse para mejorar el forecast de ventas en un futuro.

Por supuesto, una vez creada la solución, se entregará al cliente la herramienta ya validada y funcional.

Se prevé que a futuro, el informe del forecast de ventas, ayudará en la planificación y presupuestación de la empresa. Si logramos obtener una predicción óptima de las ventas esperadas, la empresa podrá planificar mejor su inventario, sus gastos y su inversión.

METODOLOGÍAS Y PUBLICACIONES ACADÉMICAS RELEVANTES

Utilizar machine learning para modelos de forecasting de ventas viene siendo una tendencia cada vez mayor. Grandes empresas ponen cada vez mayor atención a los datos que producen, con el fin de encontrar patrones e insights que les otorguen ventajas competitivas u optimicen sus procesos. Algunas incluso publican sus datos en línea y otorgan un premio económico al usuario que logre dar la solución más precisa, en lo que llaman “competencias de ciencia de datos” como es el caso de Grupo Bimbo, en agosto de 2016. El ganador del mismo fue Bohdan Pavlyshenko, quién escribió un interesante artículo al respecto.

En el mismo, explica su preferencia por modelos de regresión que por series de tiempo para este tipo de problemas, ya que este no solo se acota a la búsqueda de patrones temporales, sino que puede también mostrar tendencias, correlaciones y efectos de factores externos. A su vez, muestra la utilidad de no aferrarse a solo un modelo, sino que un conjunto de ellos puede ser integrado para encontrar una estimación más precisa. Estas palabras de un experto que pudo implementar un modelo de forecasting en una empresa transnacional y líder de mercado serán tenidas en cuenta a la hora de realizar nuestro trabajo. Se probarán modelos relacionados a regresión y se contemplará usar modelos de ensamble en casos de ser

necesarios.

Otro artículo muy interesante y que refleja casi a la perfección nuestro trabajo, es “How To Apply Machine Learning To Demand Forecasting”, escrito por Anastasiia Molodoria en octubre de 2022. En el mismo, la autora menciona los beneficios de realizar forecasting utilizando metodologías relacionadas al Machine Learning y explica cuáles son las más adecuadas para este tipo de problemas, paso por paso y con un profundo detalle. Este artículo será tomado como referencia y será de vital ayuda por su similaridad a nuestro trabajo.

ABORDAJE DEL PROBLEMA

Inicialmente pondremos el foco en llevar a cabo una limpieza y preparación adecuada de los datos antes de comenzar el análisis. Esto puede incluir la corrección de registros no coherentes, el tratamiento de valores faltantes y el tratamiento de outliers.

En caso de encontrar registros no coherentes (por ejemplo, un rubro “celular” que pertenezca a la clase “línea blanca”), se solicitará información más detallada al respecto al cliente para identificar si se trata de un error que debe ser corregido.

Puesto que el dataset contiene información de las ventas de la empresa y fue creado a partir de esta, se espera que no exista información de aquellos meses en los que no se haya vendido un determinado producto. Por esta razón, en caso de que este producto sea de interés, se añadirá un registro para dicho mes asignándole una cantidad de ventas igual a cero para tener completitud de la información en términos del período de tiempo analizado.

Existen diversas metodologías que pueden emplearse para realizar un forecast de ventas de los principales grupos de productos de la empresa.

Una metodología que se probará utilizar son los análisis de series de tiempo. Con estos podríamos identificar patrones y tendencias de las ventas históricas y utilizar los mismos para predecir las ventas futuras. Este método sería útil si nos encontramos con datos que son

estacionales.

Otra metodología potencial para utilizar son aquellas que utilizan algoritmos de regresión lineal, ya que podrían identificar relaciones lineales y brindar información acerca de la ponderación de cada variable en la predicción, indicando cuáles de estas son más importantes.

Cada uno de los modelos creados se validarán utilizando técnicas de testeo como la comparación con datos reales de ventas para asegurarse de que el modelo esté produciendo pronósticos precisos.

PLAN DE TRABAJO

El proyecto de forecast de ventas para los principales grupos de productos de la empresa los podemos dividir de la siguiente manera:

1. *Definición del problema* → en esta primera etapa establecemos el objetivo general del proyecto y definimos la problemática que queremos resolver, ayudar a la planificación estratégica anual de la empresa y tomar decisiones mejor informadas sobre la asignación de recursos. A su vez, se determinan las variables relevantes, los principales grupos de productos a pronosticar y los datos necesarios para el análisis.
2. *Obtención y limpieza de datos* → en esta segunda etapa recopilamos los datos históricos que vamos a usar para el pronóstico de ventas y también los limpiaremos para que sean de utilidad para el análisis.
3. *Análisis exploratorio de los datos* → en esta etapa vamos a explorar los datos para identificar patrones y tendencias. Se utilizarán técnicas de visualización de datos y estadística descriptiva para obtener información sobre los datos y realizar un análisis preliminar.
4. *Desarrollo del modelo de forecasting* → en esta etapa se construirán varios modelos de forecast utilizando los datos de entrenamiento y se ajustarán los parámetros del modelo para obtener las mejores predicciones posibles. Asimismo, por cada producto se seleccionará el

algoritmo que mejores resultados haya tenido de acuerdo a las métricas de evaluación seleccionadas.

En cuanto a los tiempos y riesgos del proyecto, estos dependerán de diversos factores, como el tamaño y complejidad de los datos, la disponibilidad de recursos y la calidad de la información provista. Es importante realizar una planificación detallada del proyecto y establecer plazos realistas para cada fase. También se deben identificar los posibles riesgos del proyecto, como la falta de datos o la falta de experiencia en la metodología de forecasting. Se deben establecer medidas de contingencia y mitigación de riesgos para minimizar su impacto en el proyecto. El diagrama de Gantt propuesto es el siguiente:

Actividad	Mes 1	Mes 2	Mes 3	Mes 4
Definición del problema				
Obtención y limpieza de datos				
Análisis exploratorio de datos				
Desarrollo de la solución				
Entrega final al cliente				

DATOS A UTILIZAR

Los datos a utilizar consisten principalmente en una planilla Excel con los datos históricos de venta desde el año 2019 hasta el año 2023. Dichas ventas son categorizadas por su unidad de negocio (Celulares, componentes, línea blanca, etc.), luego son divididas por su rubro y finalmente por su subrubro. Es decir, un producto es catalogado como línea blanca, dentro del

rubro de aires acondicionados y del subrubro Split. Para dichas ventas se detalla el canal de venta y la cantidad de productos vendidos. Estos datos serán utilizados para entender tendencias, estacionalidades con el fin de poder crear un modelo de forecasting.

Por otro lado, se utilizarán datos de las compras de los productos y del control de los inventarios con el fin de entender el manejo de la empresa y así poder optimizar los recursos, gastos e inversiones.

Asimismo, se añadirán datos externos que contextualice los períodos de tiempo analizados, con el fin de brindar mayor explicación a nuestra variable a predecir y que las estimaciones sean más certeras. Estos serán en un principio: tasa de inflación de los últimos tres meses, índice de confianza del consumidor de los últimos tres meses, cantidad de días del mes correspondientes a eventos comerciales (HotSale, CyberMonday, etc.) y cantidad de meses restantes para el fin del aislamiento preventivo obligatorio, durante el mismo (la misma se inicializa en 24 al comenzar el aislamiento, presentando ceros antes del mencionado mes).

Al tratar con datos sensibles de la empresa, se trabajará con extremo cuidado y confidencialidad de los datos para no revelar información sensible de la misma.

HERRAMIENTAS A UTILIZAR

Para el desarrollo del Proyecto se utilizarán diferentes herramientas, lenguajes y tecnologías para su seguimiento y colaboración.

En primer lugar, para la obtención y manejo de datos se usarán planillas de Excel y Hojas de Cálculo de Google para la colaboración del equipo.

En segundo lugar, para el análisis exploratorio de datos, se usarán lenguajes de programación de análisis descriptivo, como R en RStudio y librerías de Python como Matplotlib para generar gráficos que ayuden a entender tendencias en los datos. Además, se utilizan herramientas de visualización para crear gráficos más avanzados que representen de mejor manera los datos. Una de estas podría ser el software de visualización Tableau.

Para generar los modelos de forecasting, se utilizarán librerías de Python con el fin de crear modelos de Machine Learning que puedan predecir la demanda. Esto se realizará en Google Colab que permite a todos los miembros del equipo trabajar en simultáneo.

Por último, los resultados y sus estadísticos se analizarán en softwares compatibles y amigables para ejecutar funciones estadísticas como lo es Minitab.

ANÁLISIS DESCRIPTIVO

A fines de comprender la información disponible y obtener una visión completa de los datos se realizó un análisis exploratorio de las variables con las que contará nuestro dataset. Este análisis nos permitirá descubrir patrones, tendencias, relaciones y posibles anomalías en los datos, lo que nos ayudará a tomar decisiones informadas y construir un modelo de pronóstico de ventas más preciso.

El dataset con el que trabajaremos contiene información de las ventas mensuales de grupos de artículos seleccionados que ofrece la empresa. Para poder llevar a cabo el análisis, es necesario entender las variables que contiene el dataset. Respecto a las variables categóricas tenemos las siguientes variables: Unidad de Negocio, Rubro, Canal y Origen. También se cuenta con las variables Subrubro y Marca, pero se omitirán en una primera instancia debido a la suposición de sustituibilidad de las mismas. Luego, cuando se realice un análisis más profundo de cada rubro, se evaluará si se decide reincorporar estas variables. Además de estas, tenemos las variables Id, Mes, Año y Cantidad, las cuales son numéricas. La primera fue eliminada inmediatamente del análisis ya que no brinda ningún tipo de información relevante sobre el negocio.

Para trabajar con las ventas de estos productos hay factores externos que deben ser considerados. Por lo tanto, se agregaron variables como la inflación, el índice de precios al consumidor, meses en aislamiento y días de eventos en los que hubo promociones (Cyber Mondays, Black Fridays, etc.).

El foco del análisis girará en torno a comprender qué variables afectan las ventas de los productos. En primer lugar se inspeccionará el dataset para poder determinar si es necesario hacer limpieza o corregir alguna de sus variables, para así poder comenzar con el análisis correctamente. Luego, se analizarán las variables de manera individual para poder comprender con mayor profundidad los datasets, ya que los mismos cuentan con una gran cantidad de observaciones. En tercer lugar, habiendo estudiado las variables de manera univariada, se continuará con un análisis bivariado. Por último, se observarán las correlaciones entre estas y se observará el comportamiento de ventas por rubro, a lo largo del tiempo.

Limpeza

Luego de algunas primeras vistas del dataset, se observó que, particularmente en el mes inicial, enero de 2019, se encontraba una gran cantidad de ventas de artículos mal clasificados. Al consultar en la empresa, esta fecha coincide con la implementación del sistema ERP de la misma, por lo que se sabe que contienen cierto error. Dado este conocimiento, se decidió eliminar este mes del análisis, en búsqueda de realizar el trabajo con datos “limpios”. Algunos ejemplos de estos errores se pueden evidenciar en la siguiente imagen, donde el subrubro no es consistente con el rubro:

Unidad de M	RUBRO	SUBRUBRO	MARCA	Canal	Origen	Semana	Año	Cantidad
Linea Blanca	Aire Acondicion	Celulares	LG	Distribucion	Local	1	2019	7
Linea Blanca	Aire Acondicion	Celulares	LG	Distribucion	Local	2	2019	3
Linea Blanca	Aire Acondicion	Celulares	LG	Online	Local	2	2019	2
Linea Blanca	Aire Acondicion	Disco Externo	WD	Distribucion	Importado	2	2019	2
Linea Blanca	Aire Acondicion	Insumos Tintas	GENERICA	Distribucion	Local	1	2019	3

Imagen 1: errores de clasificación en el dataset

Los errores encontrados en febrero de 2019 eran menores, por lo que fueron corregidos manualmente. Además, se eliminan también los datos de abril de 2023 debido a que este mes aún no ha terminado, por lo que contiene menos días que los demás.

Otra inconsistencia encontrada, es la existencia de dos unidades de negocio que en su nombre indica “No usar”. Estas mismas fueron reemplazadas por una más representativa, según el caso.

También se observa que no existe completitud en los registros mensuales de cada rubro, sino solo para los que hubieron ventas. Es por esto que se decide añadir los mismos, con un valor de cantidad igual a cero.

La variable target que se determinó es la variable numérica “*Cantidad*”, por lo que es la de nuestro mayor interés. Es por esto que resulta relevante ver si presenta outliers y su distribución.

El boxplot que se presenta a continuación muestra la distribución de la variable en cada uno de los rubros analizados. En este caso, se observan outliers en todos los rubros, lo cual indica que existen valores atípicos o extremos en los datos. Sin embargo, es importante destacar que estos outliers no necesariamente son valores erróneos o incorrectos. En lugar de eso, podrían representar eventos excepcionales o variaciones naturales en los datos. Los outliers pueden surgir debido a factores como eventos inesperados, cambios en las condiciones del mercado, comportamientos atípicos de los clientes, o incluso errores en la medición o registro de los datos. Por lo tanto, en este caso específico, se ha decidido no eliminar los outliers del dataset, ya que no se consideran valores erróneos y podrían contener información relevante.

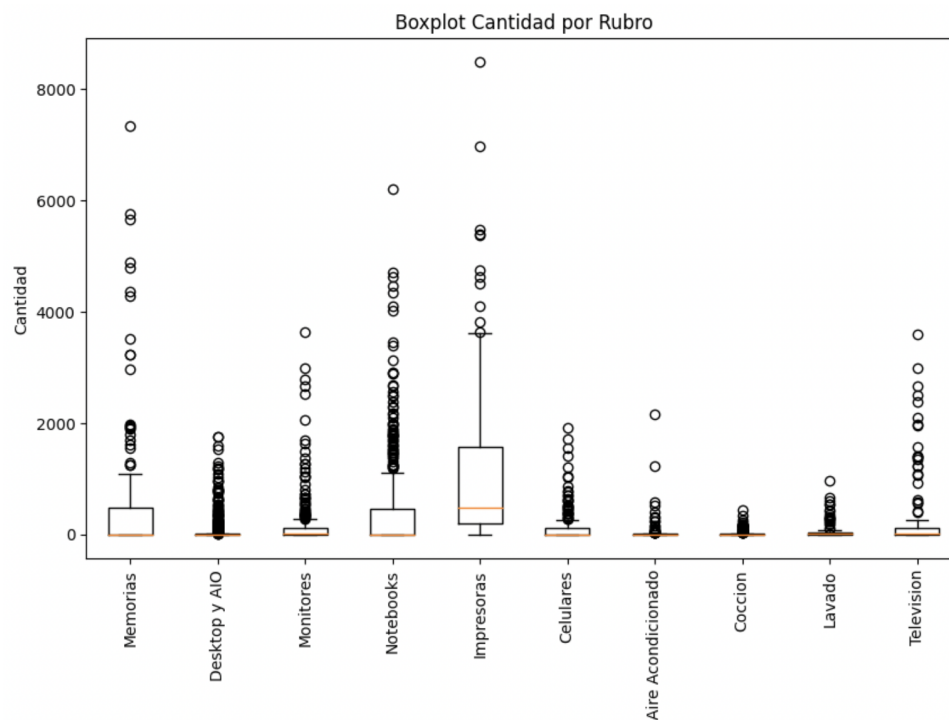


Imagen 2: Boxplot de la variable target por Rubro

También se realizó un análisis de la variable “Rubro” para poder entender la frecuencia de ventas de la misma.

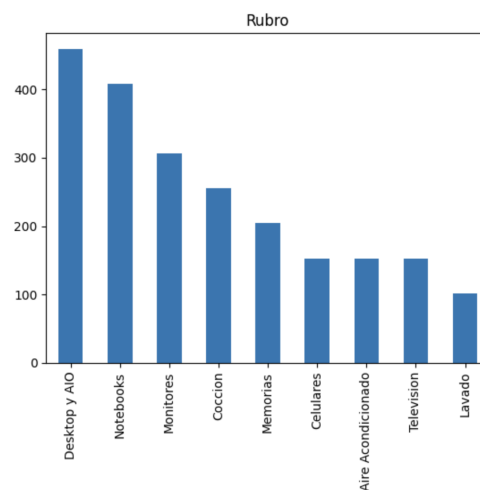


Imagen 3: Cantidad de ventas por rubro

Se puede observar que los rubros Desktop y AIO, Notebook y Monitores representan el 50% de las ventas. Sin embargo, podría suceder que no representen la mayor cantidad de unidades vendidas, por lo tanto, en el análisis bivariado veremos esta relación.

Siguiendo con el análisis para entender la frecuencia de ventas, a continuación se puede ver este análisis para cada una de las unidades de negocio analizadas.

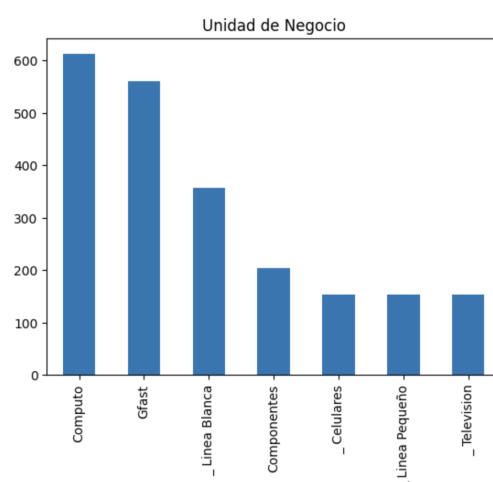


Imagen 4: Cantidad de ventas por unidad de negocio

A partir del gráfico, donde se muestran las frecuencias de las unidades de negocio, podemos observar que la unidad de negocio más frecuente es “Cómputo”, seguido por "Gfast", con una diferencia en la cantidad de ventas poco significativa, 612 contra 561.

En cuanto a la variable categórica que indica el canal a través del cual se realizó la venta, se puede observar en el gráfico a continuación que no hay diferencias significativas. El 51.1% de las ventas se realizan por Distribución y el resto por el canal Online.

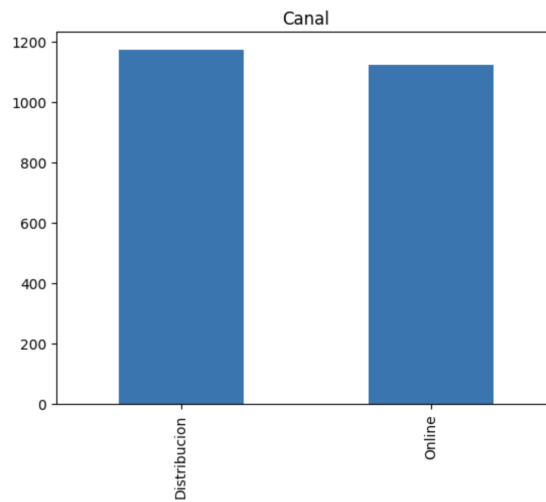


Imagen 5: Cantidad de ventas por canal

Por último, se realizó un análisis univariado de la variable categórica “Origen” para ver la frecuencia de los tres posibles valores: local, importado y fábrica.

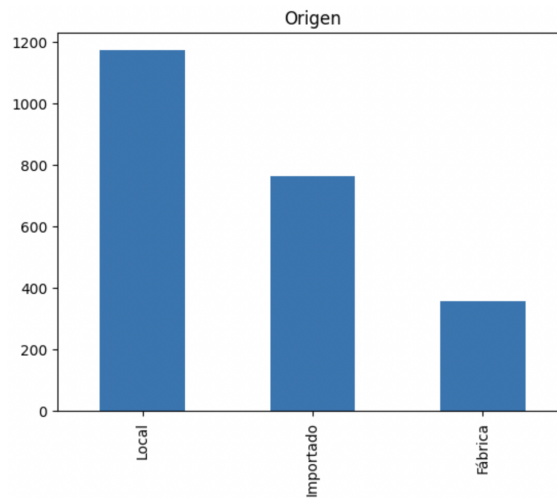


Imagen 6: Cantidad de ventas por origen

En este caso se encontró que los productos de origen local representan el 51.2% de las ventas, seguido por los importados que representan el 33.3% y por último los de fábrica con el 15.5%.

Luego de haber analizado la cantidad de ventas por las distintas variables categóricas, se continuará con un análisis de las unidades vendidas.

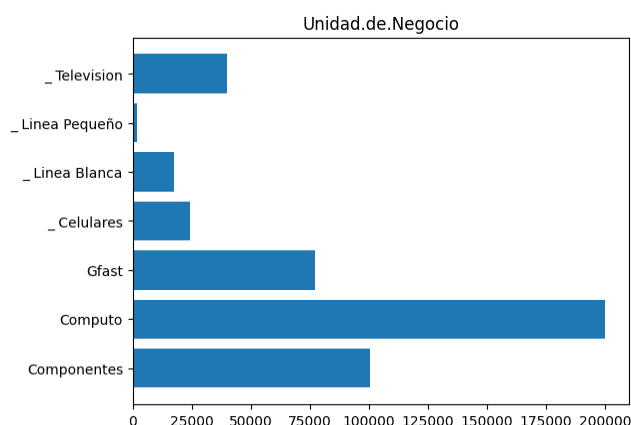


Imagen 7: Unidades vendidas por unidad de negocio

Si bien sería lógico pensar que los componentes, al ser productos más pequeños y de menores precios tendrían el mayor volumen de ventas, observamos que, de las unidades de negocio seleccionadas, Computo es la que mayor unidades vendidas presenta, seguida de la ya mencionada. Una posible causa de esto, es que al segmentar las agrupaciones por rubros, haya más del tipo Computo que Componentes, por lo que no estaríamos viendo la porción completa de estas unidades de negocio.

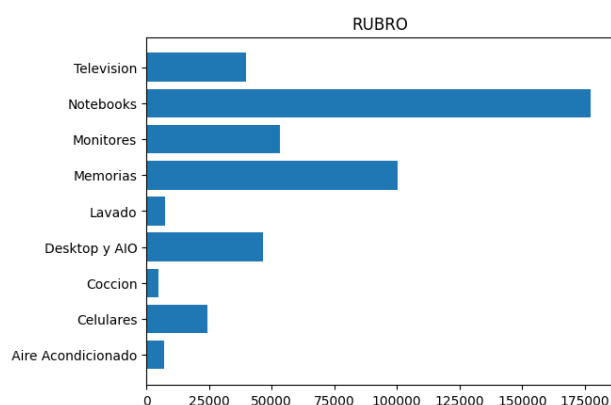


Imagen 8: Unidades vendidas por rubro

Al hacer el mismo análisis por rubro, encontramos que, por mucho, Notebooks es el rubro con más ventas. Lo planteado anteriormente se confirma al ver que, si bien Memorias tiene una alta cantidad de ventas en relación al resto, es el único rubro perteneciente a la unidad de negocio Componentes, mientras que tanto Notebooks como Desktop y AIO se suman formando la unidad de negocio Computo.

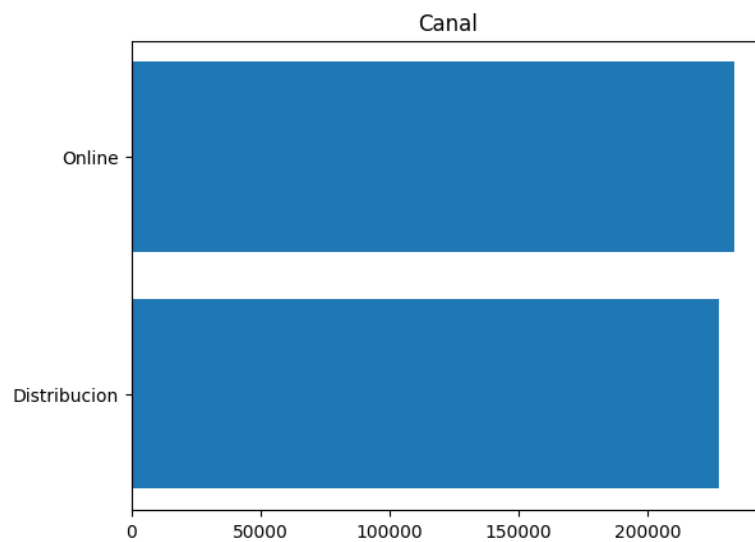


Imagen 9: Unidades vendidas por canal

En cuanto al canal, la paridad se mantiene. No hay diferencias significativas entre ambos, al igual que en el análisis univariado.

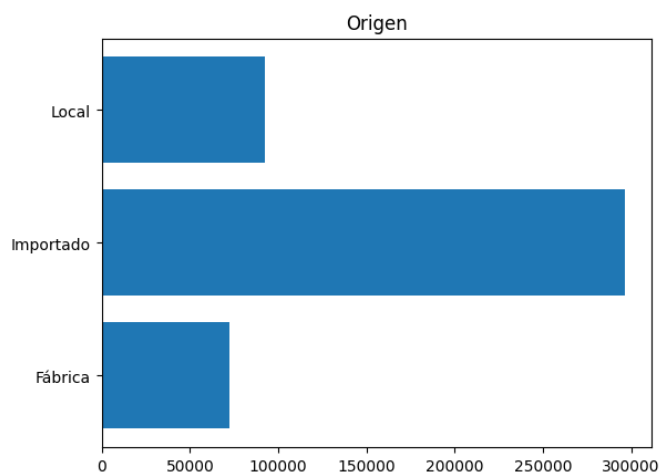


Imagen 10: Unidades vendidas por origen

La variable Origen muestra que aproximadamente dos tercios de las ventas de estos artículos en el período analizado son de productos importados, mientras que los productos locales y de fábrica presentan valores similares.

En cuanto a las ventas anuales, se puede observar a primera vista en el gráfico un alto índice de crecimiento. A saber, el año 2019 presenta ventas correspondientes a once meses (debido a la exclusión de enero) y 2023 presenta apenas tres meses (los transcurridos de forma completa al día de la escritura de este informe). Por lo tanto, para poder realizar una comparación de los años, se decidió tomar el promedio de unidades vendidas mensualmente por cada año. Dicho esto, se puede observar que en cada año se fue creciendo aproximadamente un 60% en el promedio de unidades vendidas por mes.

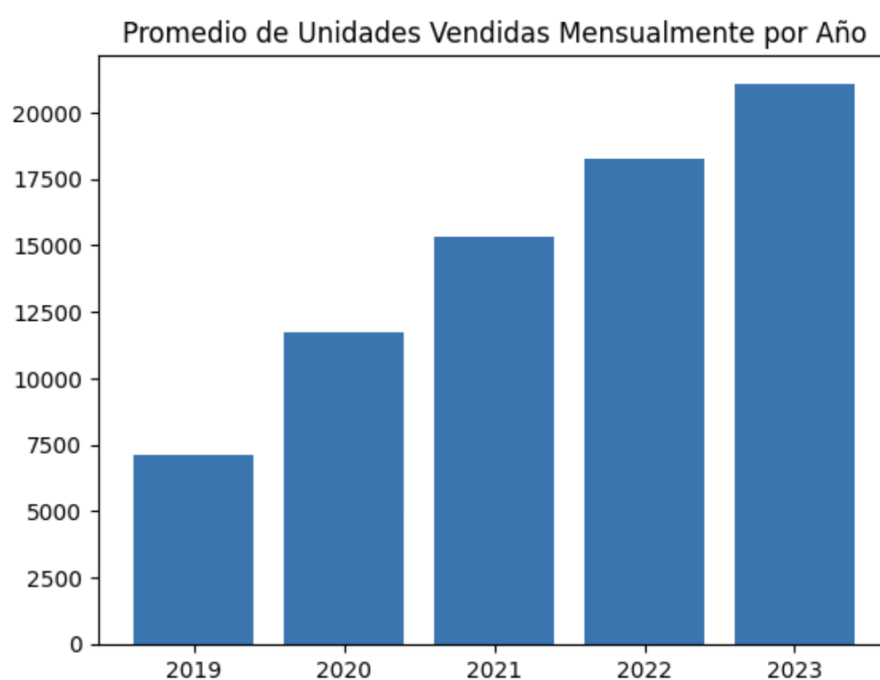


Imagen 11: Unidades vendidas en promedio mensualmente por año

Para ver esta información de forma más comparativa, contamos con la siguiente tabla de ventas promedio mensuales por año:

Año	Ventas mensuales promedio	Crecimiento
2019	3085,727	
2020	6566,75	+112,81%
2021	9625,667	+46,58%
2022	15072	+56,58%
2023	17422,667	+15,6%

Tabla 1: Unidades promedio mensuales por año

Dada una idea de la distribución de las ventas a lo largo de los años, procederemos a hacer un análisis sobre los meses. Para esto, se excluyó 2019 y 2023 puesto que estos años poseen menor cantidad de meses para analizar:

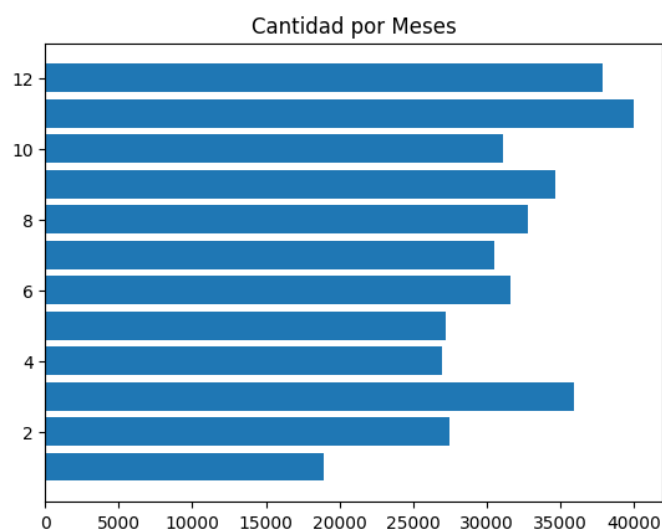


Imagen 12: Unidades vendidas por mes

Observamos que el mes de enero suele ser el que menos ventas presenta. Por el contrario, se ven picos en marzo, noviembre y diciembre. Es importante destacar que en este análisis se encuentra intrínseco el crecimiento interanual.

Correlaciones

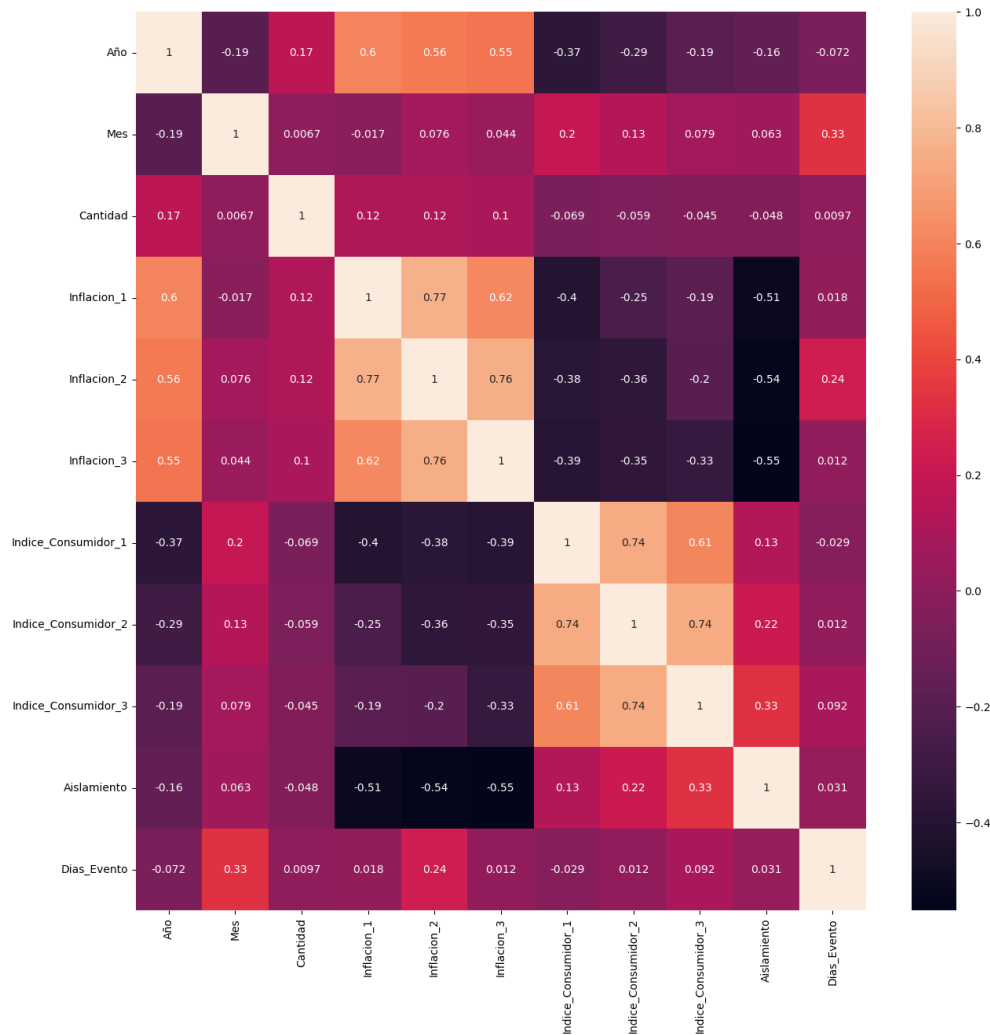


Imagen 13: Matriz de correlación

Se realiza un gráfico de correlaciones para observar qué variables pudieran tener una relación lineal entre sí. Las únicas correlaciones leves se aprecian entre las variables de inflación. Esto nos indica que es probable que los meses suelen encontrar variaciones inflacionarias similares a sus anteriores o consecuentes, no siendo tan comunes los cambios de tendencia abruptos. No se encuentran más correlaciones relevantes entre las variables. Estos indicadores pueden ser evidencia de inexistencia de multicolinealidad.

Variables en el tiempo

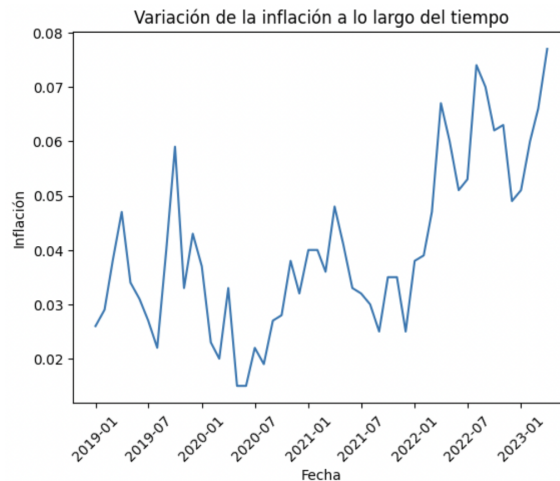


Imagen 14: variación inflacionaria por mes

Si observamos la variable inflacionaria, vemos que tiene una tendencia fluctuante. Se puede observar que en el 2023 se ha disparado alcanzando un máximo de 0,077.

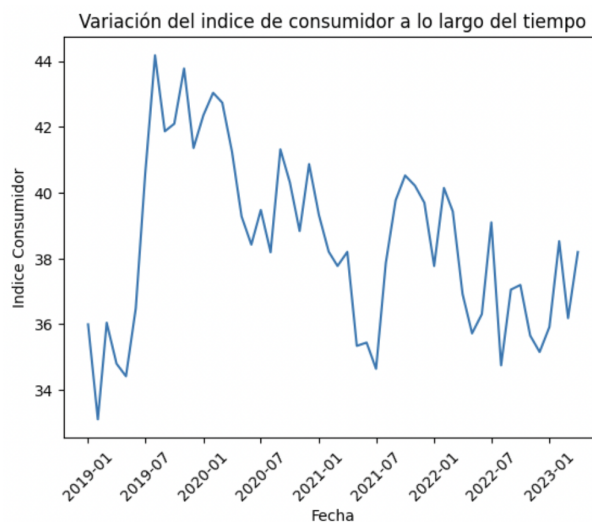


Imagen 15: índice de confianza del consumidor por mes

Al ver el índice de confianza del consumidor, observamos un alto pico en mediados de 2019, seguido de un decrecimiento marcado pero errático, con una nueva suba en finales de 2021 y un nuevo decrecimiento pero más leve.

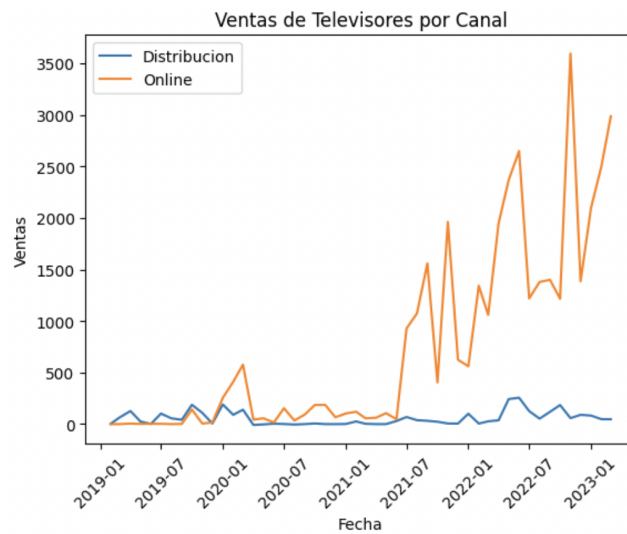


Imagen 16: histórico de ventas de televisores por canal

Al realizar un análisis segmentado de las ventas por rubro, se ha identificado un crecimiento exponencial en la venta de televisores a través del canal online, específicamente a partir de mediados de 2021. Esta observación indica un cambio significativo en el comportamiento de compra de los consumidores en relación con los televisores, favoreciendo cada vez más las compras a través de plataformas digitales.

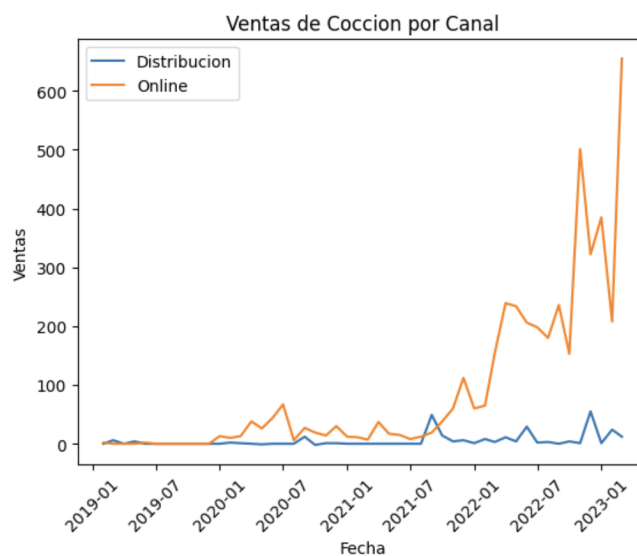


Imagen 17: histórico de ventas de cocción por canal

En cuanto al rubro de cocción, que incluye principalmente cocinas y anafes, se observa una distribución similar a la de las televisiones, con un gran crecimiento en el canal online a partir de mediados de 2019. Sin embargo, este crecimiento no se da de forma tan inmediata como el caso anterior.

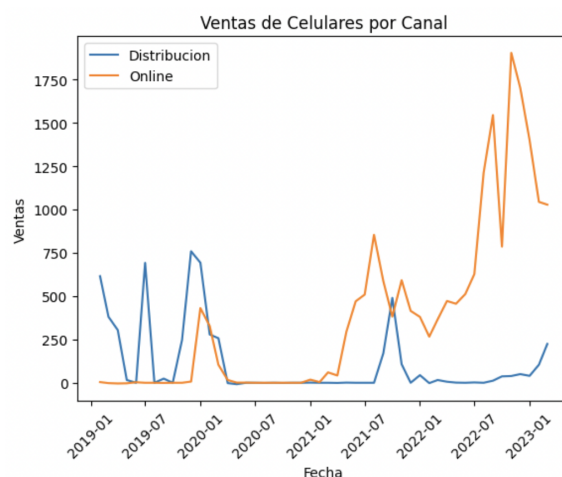


Imagen 18: histórico de ventas de celulares por canal

Las ventas de celulares comienzan su crecimiento en online un poco antes que los otros dos rubros ya mencionados, en principios de 2021. A su vez, existen algunos picos por el canal de distribución en 2019 y principios de 2020.

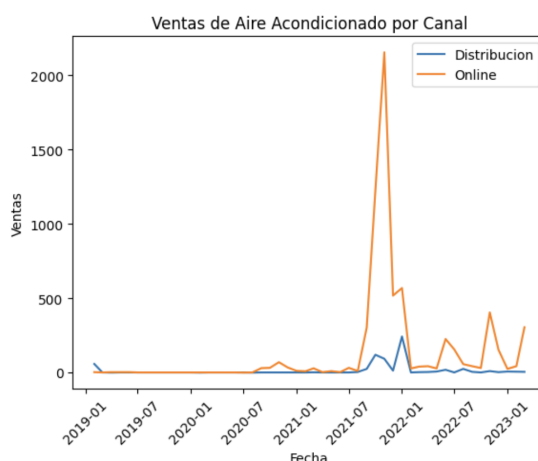


Imagen 19: histórico de ventas de aires acondicionados por canal

Los aires acondicionados parecen tener un comportamiento cíclico, con picos marcados en los segundos semestres a partir de 2021, con especial énfasis en el canal online.

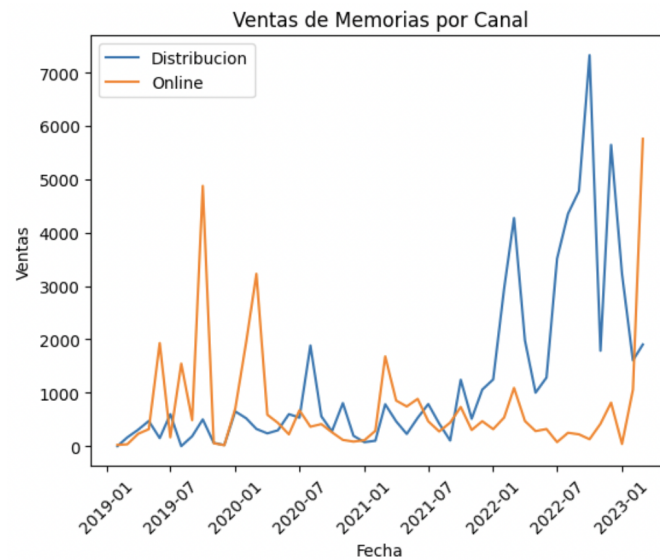


Imagen 20: histórico de ventas de memorias por canal

En cuanto a las memorias, si bien en 2019 y parte de 2020 su fuerte se encontraba en online, a partir de 2022 ha puesto su foco en el canal de distribución.

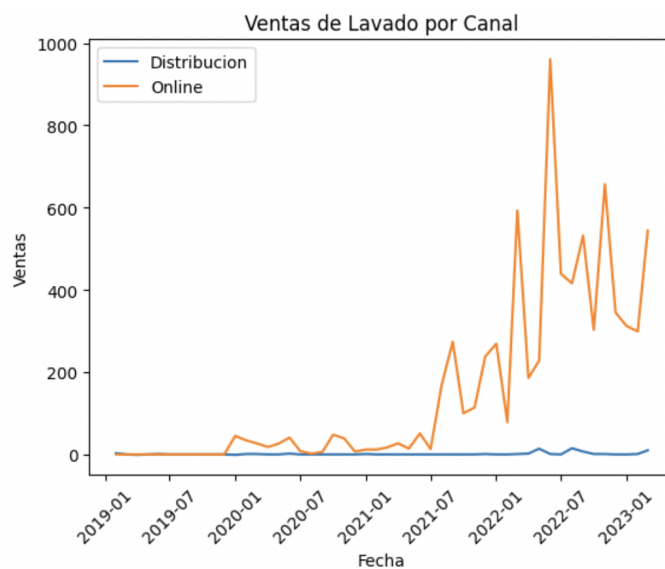


Imagen 21: histórico de ventas de lavado por canal

El rubro de lavado se muestra claramente enfocado hacia el canal online, especialmente a partir de mediados de 2021.

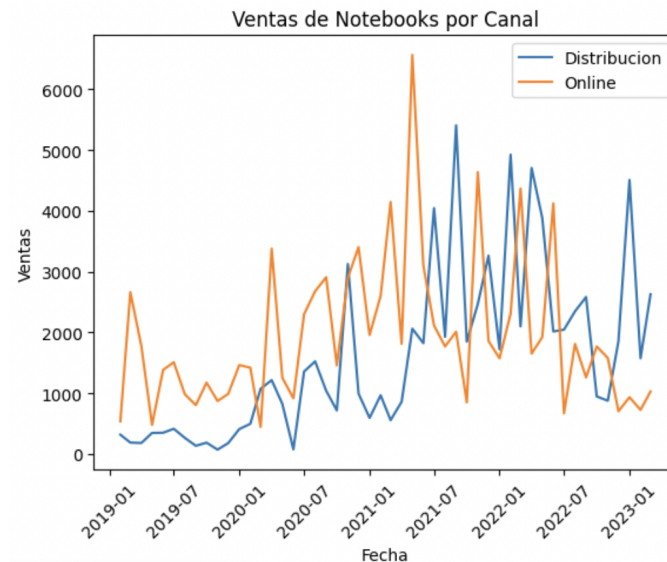


Imagen 22: histórico de ventas de notebooks por canal

El gráfico permite observar que las notebooks, dentro de los diferentes rubros de productos analizados, son un rubro que muestra una presencia sólida y significativa en ambos canales de venta. A su vez, esta fortaleza en ambos canales de venta sugiere que las notebooks son productos altamente demandados y populares entre los consumidores.

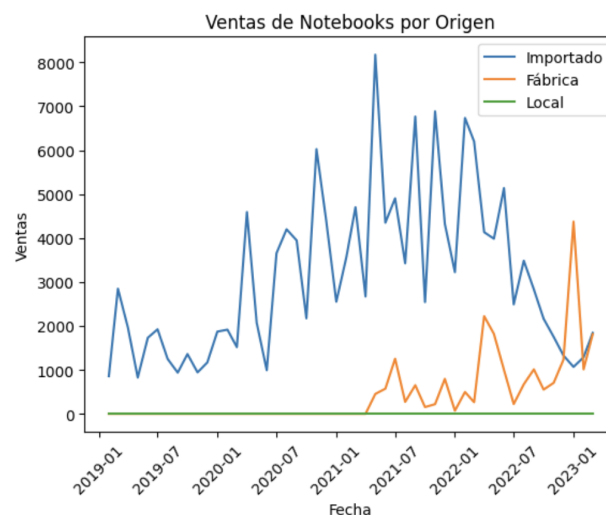


Imagen 23: histórico de ventas de notebooks por origen

En cuanto a su origen, existen ventas de notebooks importadas y fabricadas. Las fabricadas se empiezan a vender en 2021 y a medida que estas se venden más, las ventas de las notebooks importadas decrece.

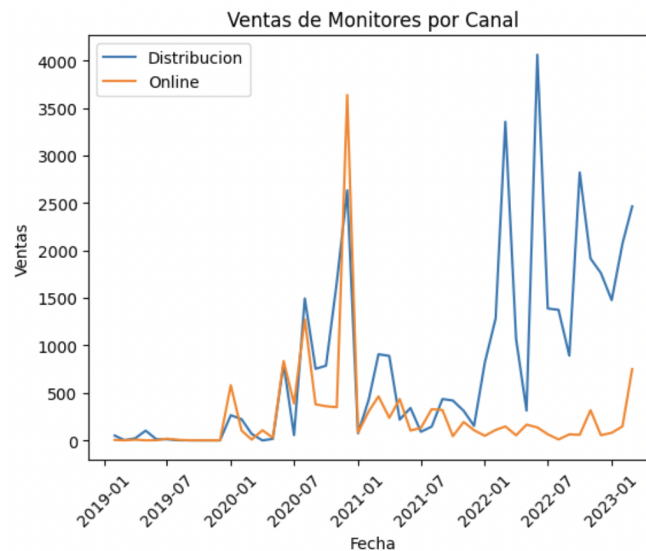


Imagen 24: histórico de ventas de monitores por canal

Los monitores presentan dos picos de ventas muy marcados, uno a mediados de 2021 y otro a lo largo de 2022. El primero se produce en ambos canales, mientras que el segundo sólo en distribución.

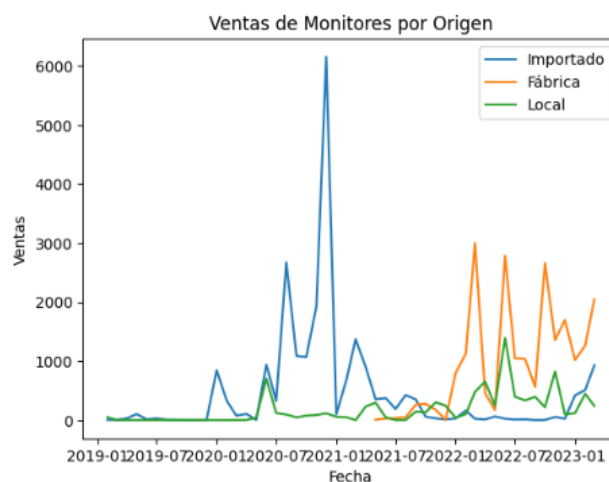


Imagen 25: histórico de ventas de monitores por origen

Este rubro es uno de los pocos de los seleccionados que presenta los tres tipos de origen. El pico de 2021 pertenece claramente a monitores importados, mientras que el de 2022 es una mezcla de productos fabricados y locales, especialmente de los primeros.

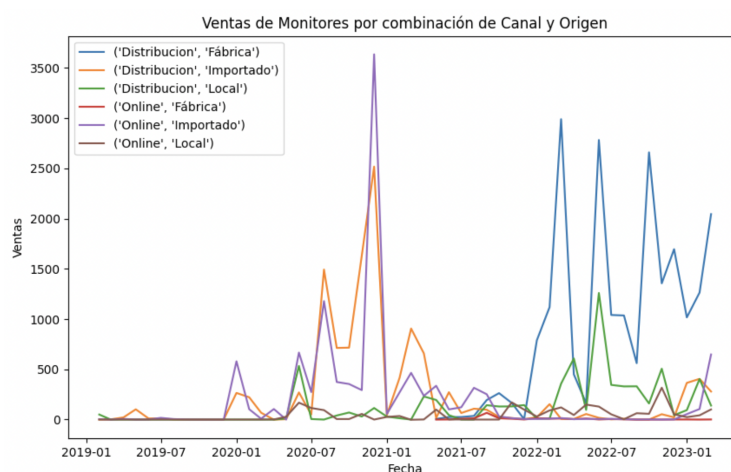


Imagen 26: histórico de ventas de monitores por canal y origen

En este último gráfico se puede observar que el primer pico se realiza sobre monitores importados, en ambos canales, mientras que el segundo es principalmente productos fabricados en el canal de distribución.

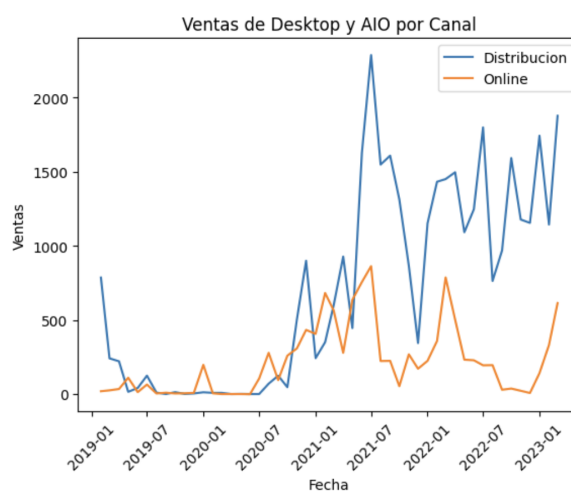


Imagen 27: histórico de ventas de desktop y aio por canal

El rubro de Desktop y AIO se ha vendido de forma constante en el canal online a partir de mediados de 2020, mientras que ha mostrado una tendencia creciente desde ese mismo año.

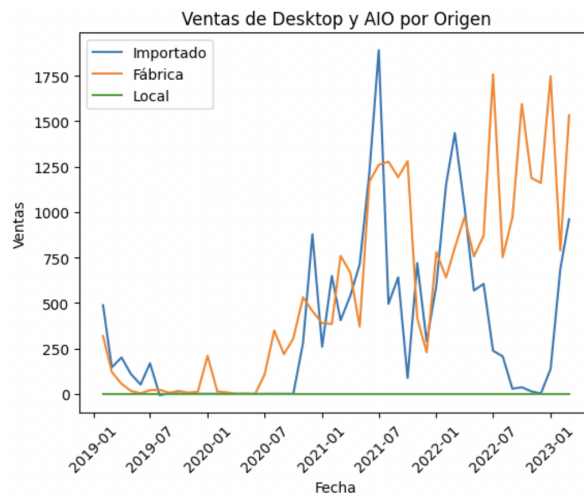


Imagen 28: histórico de ventas de desktop y aio por origen

Este rubro se vende tanto mediante artículos importados como fabricados por la misma empresa.

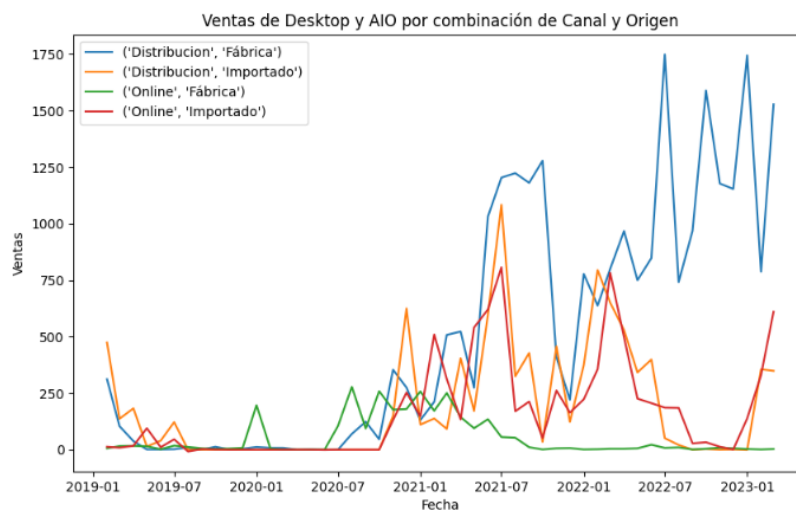


Imagen 29: histórico de ventas de desktop y aio por canal y origen

En este gráfico final se observa que este rubro tiene una mayor importancia a partir de mediados de 2020, principalmente las Desktop y AIOs fabricadas, vendidas mediante el canal distribución, También tienen una importancia significativa, en mediados de 2021 las importadas, en ambos canales.

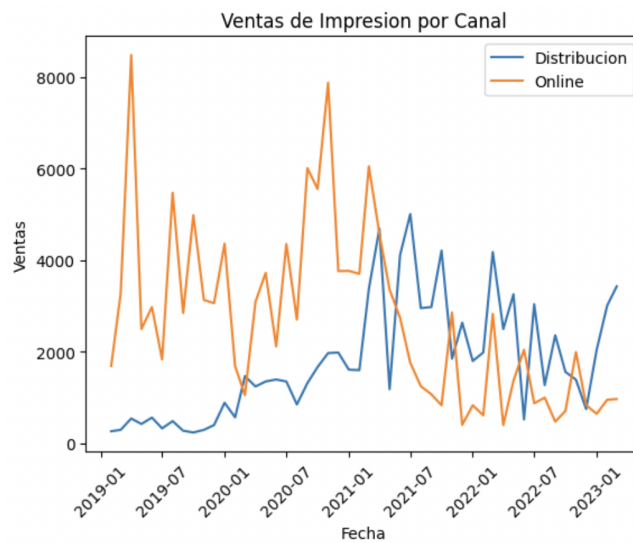


Imagen 30: histórico de ventas de impresión por canal

Las impresoras han observado un claro cambio de tendencia en cuanto a su canal de venta. Históricamente se habían vendido de forma online, pero a partir de 2020 han crecido sus ventas por distribución y disminuido por el mencionado anteriormente.

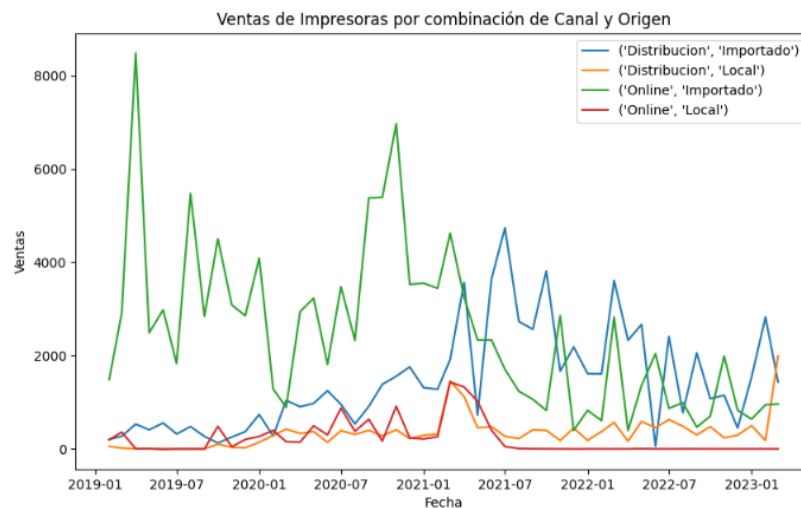


Imagen 31: histórico de ventas de impresión por canal y origen

Si bien hay ventas de impresoras del ámbito local, están muy por debajo de las ventas de productos importados.

CONCLUSIONES DE ANÁLISIS DESCRIPTIVO

A pesar de la buena predisposición del cliente a proveer y facilitar el acceso a sus datos históricos, se perciben errores de carga en los mismos. Si bien hemos intentado tratarlos y limpiar el dataset, es importante ser conscientes de que puede existir un margen de error implícito.

En cuanto a la variable target, tiene una distribución dispar, con muchos valores cercanos a cero y pocos valores mayores a mil. Tampoco pareciera estar correlacionada directamente con ninguna otra variable.

En cuanto a los productos, sus características y distribuciones de venta son muy desiguales entre sí, difiriendo tanto en sus canales principales de venta como en sus orígenes. A partir de esto, se evaluará realizar un modelo para cada uno de ellos, según su canal y su origen, dando prioridad a los de mayor facturación para la empresa.

ENFOQUE DE SOLUCIÓN

Si bien la empresa cuenta con alrededor de treinta rubros catalogados, en este trabajo nos enfocamos en el forecasting de siete de ellos, los cuales fueron seleccionados por el cliente como los de mayor interés para analizar. Estos rubros son notebooks, monitores, impresoras, cocción, memorias, celulares y televisión. Para cada uno de estos rubros, se realizaron sus respectivos datasets filtrando los registros correspondientes del dataset original. En cada uno de los casos se utilizaron de base las variables ya mencionadas, pudiendo agregarse nuevas o descartar las existentes dependiendo del modelo y distribución que se esté trabajando. En cada uno de los rubros, se analizaron las distribuciones de las cantidades a lo largo del tiempo, segmentadas según las variables categóricas. Esto se realizó con el fin de darse una primera idea sobre la superposición de ventas entre distintos valores de una misma variable categórica. Por ejemplo, conocer cuál es el comportamiento de las cantidades vendidas en el canal de distribución cuando crecen las ventas en el canal de online. Según este análisis, se

evaluó la posibilidad de omitir las variables categóricas (canal y origen) en cada caso, buscándose obtener los resultados que muestren más precisión acorde a la realidad.

METODOLOGÍAS A IMPLEMENTAR

Para realizar modelos con los datasets seleccionados (que como dijimos anteriormente, pudiera ser sobre la totalidad de registros de un rubro o sobre un segmento específico de estos), se utilizaron tres tipos de algoritmos: series de tiempo, regresión y algunos modelos de ensemble como xgboost, y random forest. En el caso de estos modelos se optimizaron los hiperparametros con el fin de buscar una mejora en el score.

Hiperparametros para XGBoost:

- nrounds: número de iteraciones que se realizarán en el ajuste del modelo.
- max_depth: especifica la profundidad máxima de cada árbol en la secuencia.
- eta: controla la tasa de aprendizaje del modelo.
- gamma: especifica la reducción mínima requerida en la función de pérdida para realizar una nueva partición en un árbol.
- colsample_bytree: proporcion de características que se utilizan para entrenar cada árbol
- min_child_weight: especifica la suma mínima de los pesos de las muestras,
- subsample: proporción de muestras que se utilizan para entrenar cada arbol.

Hiperparametros para Random Forest:

- max_depth: profundidad máxima de cada árbol en el bosque
- n_estimators: número de árboles de decisión que se construirán en el bosque
- max_features: cantidad máxima de características que se consideran al buscar la mejor división de cada nodo del árbol de decisión

La elección de las series de tiempo se dio por la naturaleza del problema y su capacidad de identificar ciclos y estacionalidad. Para el analisis de serie de tiempo se emplearon dos métodos ampliamente utilizados: Holt-Winters y Naive. El método de Holt-Winters es un enfoque de suavizado exponencial que se utiliza para pronosticar y modelar patrones. Este

método considera tres componentes principales en una serie de tiempo: nivel, tendencia y estacionalidad. El suavizado exponencial se utiliza para estimar y actualizar estos componentes a medida que avanza el tiempo. Holt-Winters es especialmente útil cuando se trata de series de tiempo con patrones de tendencia y estacionalidad, permitiendo realizar pronósticos precisos y capturar cambios en el comportamiento de la serie a lo largo del tiempo. Por otro lado, el método Naive, también conocido como "persistencia", es una técnica de pronóstico simple que asume que el valor futuro de la serie de tiempo será igual al último valor observado. Es decir, supone que no hay cambios ni patrones en la serie y simplemente replica el último valor conocido como predicción para el siguiente período. Aunque es un método muy básico y no captura patrones complejos, puede ser útil como punto de referencia para evaluar la efectividad de otros métodos más sofisticados. La regresión se eligió por su potencial para manejar datos continuos y medir el peso y dirección de las variables. Se optó por los modelos de ensemble como tercera alternativa por su capacidad de combinar y consensuar un resultado mediante subalgoritmos, siendo altamente parametrizables y pudiendo encontrar relaciones no lineales. Los mismos se compararon mediante la métrica MAPE (error porcentual absoluto medio), ya que al devolver resultados en términos porcentuales, los mismos son comparables entre modelos realizados a distintos datasets. De esta forma, pudimos verificar en los casos que haya sido necesario si la aplicación de los algoritmos al dataset entero del rubro fue más óptimo que los resultados obtenidos por datasets subdivididos por variables categóricas. También se obtuvo la capacidad de comparar modelos realizados a diferentes rubros, para observar sobre cuáles obtuvimos mejores resultados.

SETEO DE EXPERIMENTACIÓN

Para la experimentación se decidió separar los modelos en los diez rubros con mayor cantidad de ventas, los cuales son: memorias, desktop y AIO, monitores, notebooks, impresoras, celulares, aire acondicionado, cocción, lavado y televisión. En cada uno de los subdatasets producidos por la división de rubros y (en caso de haber sido considerada útil) la subdivisión

de canales, se aplicó la partición de los datos, en conjunto de entrenamiento y testeo, teniendo el 80% de los datos en el conjunto de entrenamiento y un 20% en el de testeo, en orden de tiempo, el cual corresponde a los últimos 10 meses.

DESARROLLO DE SOLUCIÓN

Modelo Memorias

Debido a las ventas históricas y al análisis previamente hecho sobre el rubro memorias, se decidió realizar un solo modelo para el rubro en donde se unificaban tanto los orígenes como los canales.

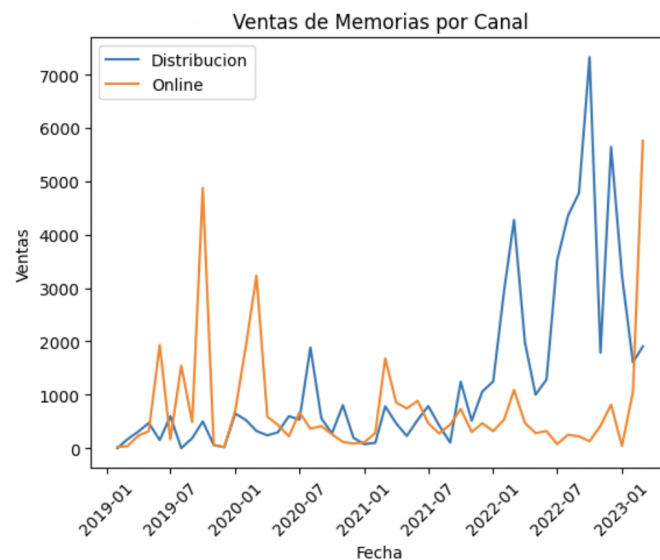
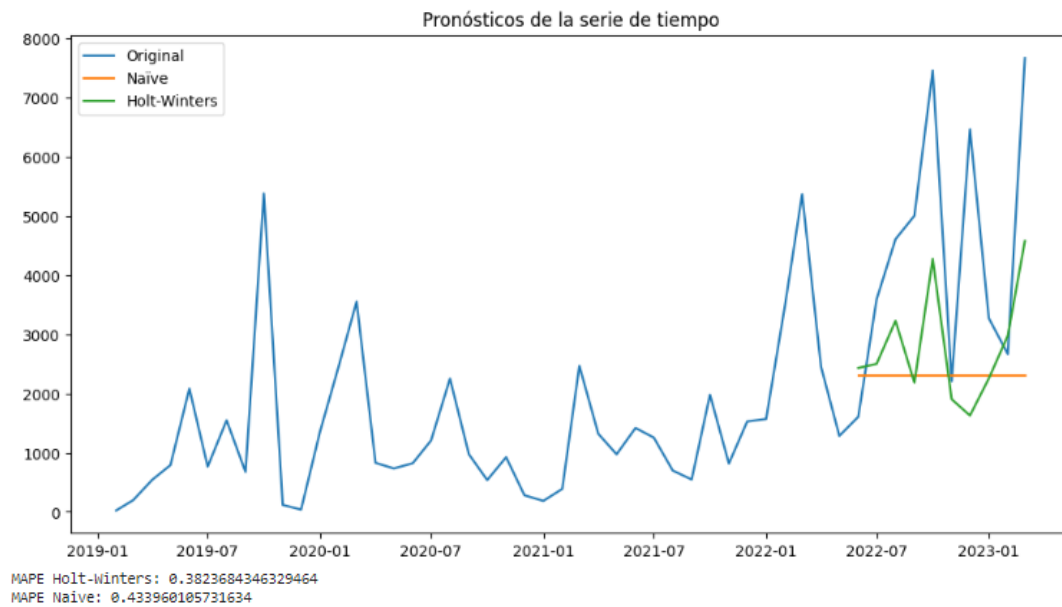


Imagen 33: Memorias por canal

Como se puede ver en el gráfico, las ventas del rubro memorias comienzan siendo principalmente por ventas online pero a partir del año 2022, son “reemplazadas” mediante el canal distribución.

Luego de experimentar con algoritmos como Random Forest, XG Boost y Regresión Lineal Múltiple con diferentes conjuntos de variables, se llegó a la conclusión que el modelo con mejores modelados era mediante un pronóstico de serie de tiempo utilizando el método de

Holt-Winters, en donde se usaban los últimos 10 meses para testear. Para el caso de Random Forest se realizó una optimización de los hiperparametros en donde se analizó max_depth, n_estimators y max_features . Los valores finales fueron: max_depth=10, n_estimators=300, max_features=1.



En el gráfico se puede ver los dos métodos utilizados, el de Holt-Winters y el Naive, en donde el primero logra obtener un MAPE menor.

Algoritmo	MAPE
Regresión Lineal Multiple	3,1
Random Forest	2,1
XGBoost	0,8
Naive	0,43
Holt Winters	0,38

Modelo Celulares

A diferencia del modelo de las memorias, el de celulares se debe dividir por Canal, ya que hay periodos de tiempo en donde se vende el mismo rubro por ambos canales.

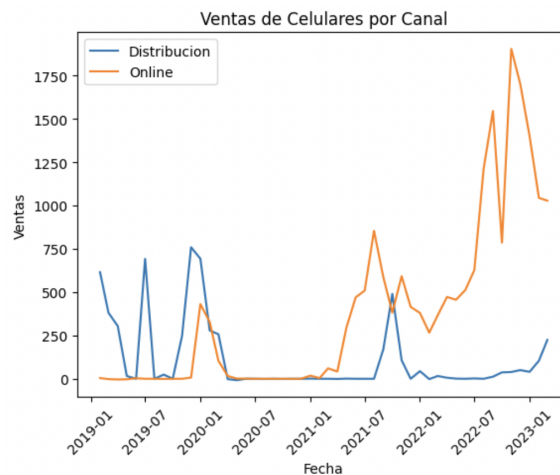


Imagen 34: Celulares por canal

En el gráfico se demuestra que en algunos meses, el rubro Celulares comparte ventas tanto por medio del canal online como del canal distribución; por ende, se debe analizar por separado.

Al igual que el modelo anterior, se aplicaron algoritmos como Random Forest , XG Boost y Regresión Lineal Múltiple, en donde se hizo una selección de variables y ajuste de hiperparametros, pero los mejores scores se obtuvieron usando series de tiempo. Para el ajuste de hiperparametros se analizó la optimización del max_depth, n_estimators y max_features, al igual que los modelos anteriores. Para la Regresión Lineal Múltiple, se realizaron distintos modelos, cambiando las variables en relación al p valor obtenido.

Modelo celulares distribución

En el caso del modelo para celulares por medio del canal de distribución, no se pudo lograr buenos resultados para ninguna de las métricas de los algoritmos. Para solucionar este problema, se decidió también hacer un modelo unificando los canales del rubro.

Modelo celulares online

Para el modelo de los celulares mediante el canal online se muestra que el mejor score se obtiene mediante el método de Holt-Winters con un MAPE de 0,477 en la partición de test, la cual abarca los últimos 10 meses del conjunto de datos.

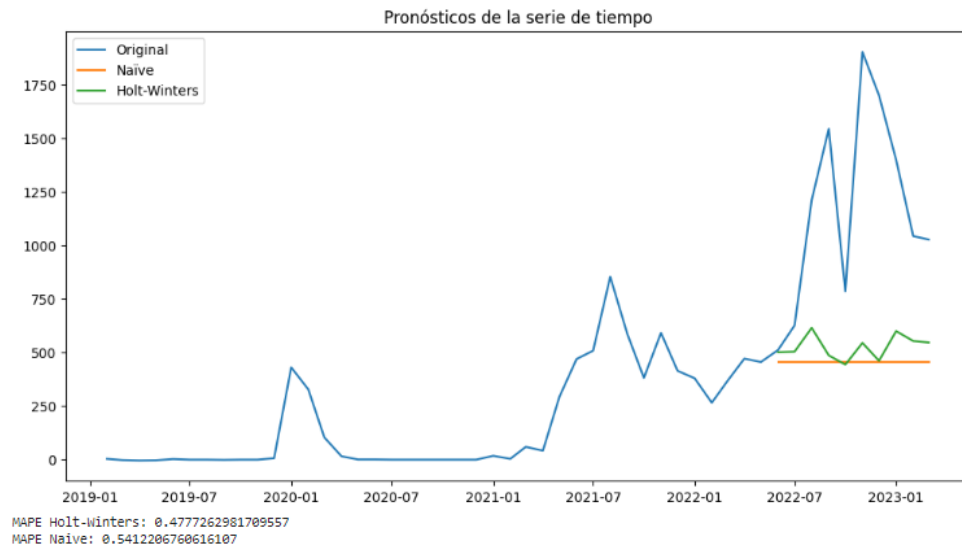


Imagen 35: Predicción de celulares online

Algoritmo	MAPE
Regresión Lineal Multiple	-
Random Forest	-
XGBoost	-
Naive	0,54
Holt Winters	0,48

Modelo celulares unificado

A raíz de los valores obtenidos en el modelo de los celulares mediante el canal de distribución, se decidió probar con un modelo unificando los canales, como se había hecho

previamente con las memorias. Luego de probar con distintos algoritmos, se llegó a la conclusión que el método de Holt-Winters en un modelo unificado era la mejor manera de hacer un pronóstico para el rubro celulares, con un MAPE de 0,353.

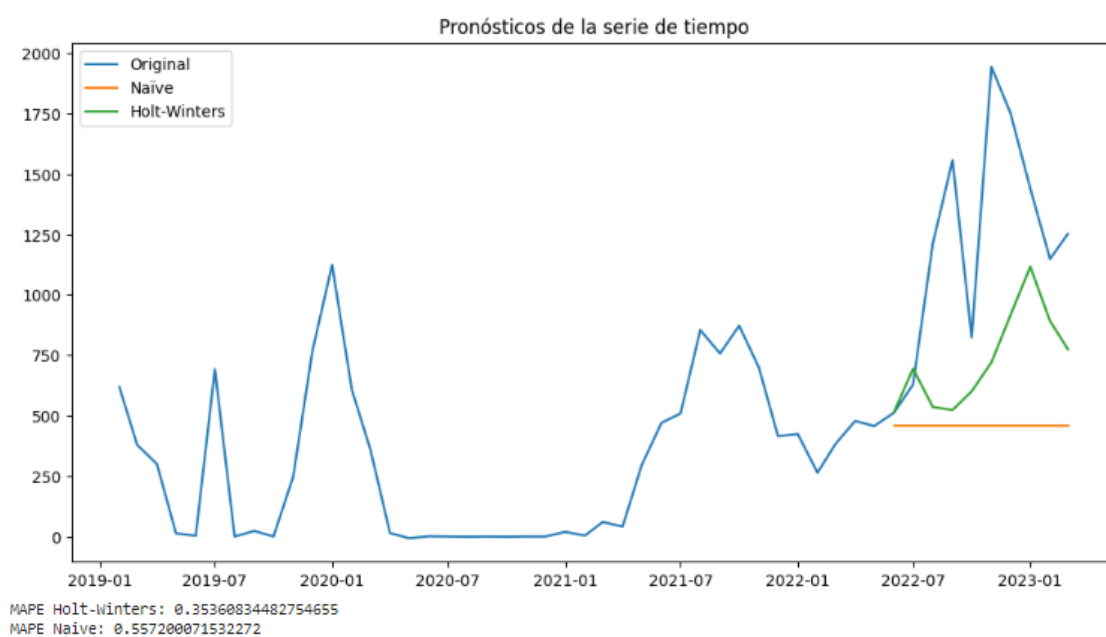


Imagen 36: Predicción de celulares totales

Algoritmo	MAPE
Regresión Lineal Multiple	19,9
Random Forest	13,4
XGBoost	6,6
Naive	0,56
Holt Winters	0,35

Modelo Televisores

En el caso de los televisores ocurre que casi la totalidad de sus ventas se realizan a través del canal online, es decir, casi no hay ventas por medio de la distribución. Es por esto, que al igual que las memorias, se decidió unificar los canales y sumar sus cantidades vendidas.

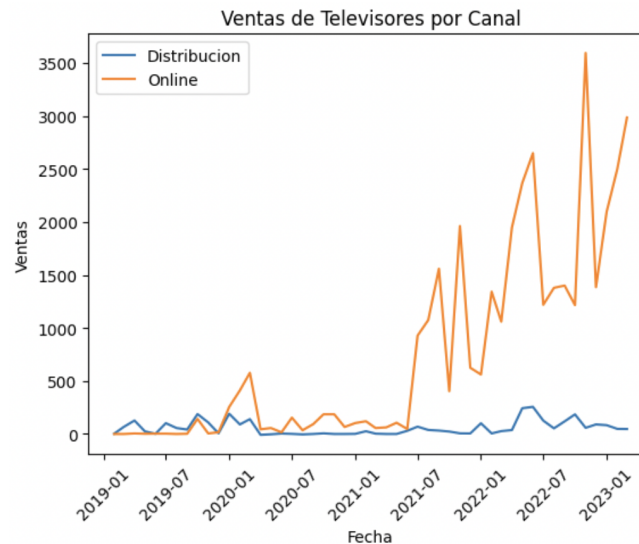


Imagen 37: Televisores por canal

En el gráfico se puede ver la diferencia que hay entre las ventas por distribución y las ventas online.

Para el modelado también se ha experimentado con Random Forest, XG Boost y Regresión Lineal Múltiple. En XG Boost se realizó un ajuste de los hiperparametros y en la Regresión Lineal Múltiple se buscó distintas combinaciones de variables teniendo en cuenta su p valor. Sin embargo, el modelo que presentó mejores resultados fue el de Naive con un MAPE de 0,36.

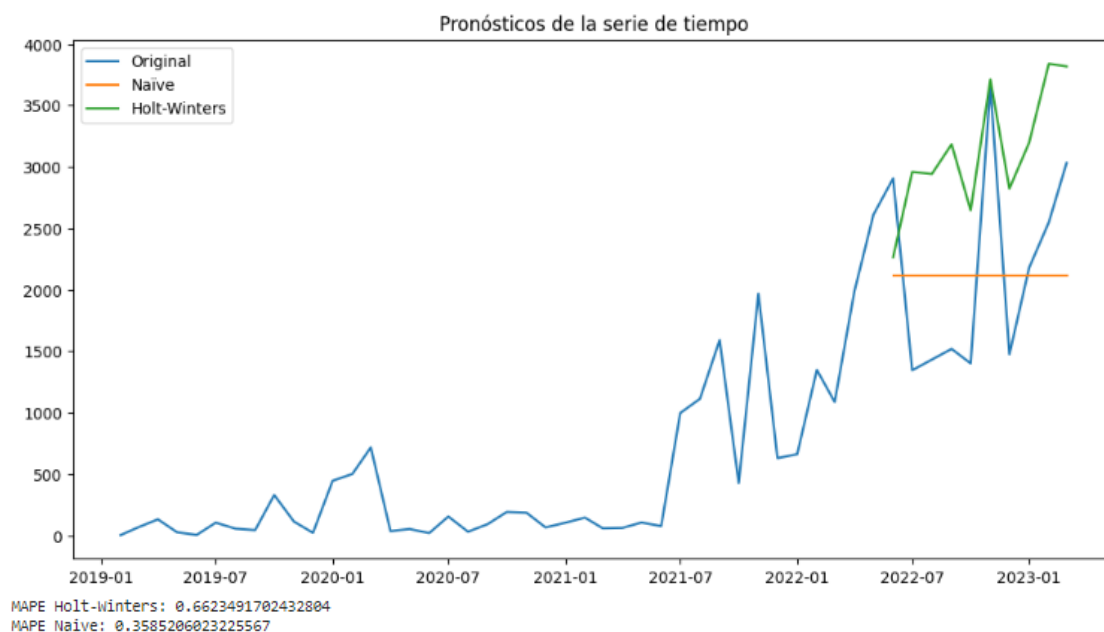


Imagen 38: Predicción de televisores totales

Algoritmo	MAPE
Regresión Lineal Multiple	19,9
Random Forest	12,8
XGBoost	6,66
Naïve	0,38
Holt Winters	0,66

Modelo Notebooks

Al observar la distribución de las cantidades de ventas a lo largo del tiempo según el canal, no se observa que las ventas de un canal “se coman” las ventas del otro. De hecho, a grandes rasgos, pareciera que el crecimiento y decrecimiento de los canales se dan de forma conjunta y no contraria.

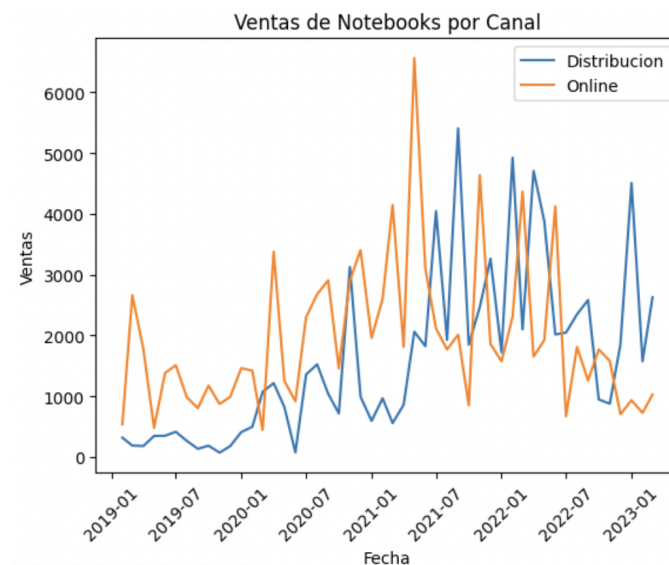


Imagen 39: Notebooks por canal

Tomadas estas conclusiones, se decide empezar por un modelo sobre el dataset completo del rubro.

Antes de empezar a modelizar, cabe destacar que no había registros para enero de 2019. Por este motivo, se agregó dicho valor ficticio, tomándose el promedio de ese año como referencia.

Se probaron algoritmos de regresión lineal (utilizando diferentes combinaciones de variables, dándoles prioridad según su p-valor), diferentes tipos de series de tiempo y xgboost. Los resultados fueron los siguientes:

Algoritmo	MAPE
Naive estacionario	0,77
Drift	1
TSLM	1,43
Naive	0,81
Exponential smoothing	0,81

Holt Winters	1,11
ARIMA	1,23
TBATS	0,79
Regresión lineal	1,06
XGBoost	0,48

De los algoritmos probados, el que mejores resultados arrojó fue el de xgboost, parametrizado mediante la función *train* de la librería caret de R. El resultado de la métrica MAPE fue de 0.6241. Los parámetros utilizados se detallan a continuación:

nrounds	10
max_depth	5
eta	0.35
gamma	0
colsample_bytree	1
min_child_weight	1
subsample	1

Tabla 1: Parametrización del algoritmo xgboost

Este modelo predijo un único valor para todos los meses, y es que al observar la importancia que este le había asignado a las variables se encontró que únicamente “Año” se tomaba como input para arrojar la predicción.



Imagen 39: Predicción de notebooks totales

Debido a la importancia de este rubro en la empresa (es el rubro de mayor facturación), se intentó buscar mejores resultados dividiendo el dataset según el canal de venta. Tanto al subset de online como para el de distribución, se les asignó un valor para enero de 2019 de la misma forma que para el anterior.

Algoritmo	MAPE
Naive estacionaria	0,67
Drift	1
TSLM	0,85
Naive	0,36
Exponential smoothing	0,46
Holt Winters	1,24
ARIMA	0,55

TBATS	0,39
Regresión lineal	0,97
XGBoost	0,9

Al aplicarse los mismos algoritmos, se encontró un mejor resultado en la predicción de las ventas por el canal online, al obtener un MAPE de 0,3651. El mismo se logró mediante la serie de tiempo naive simple.



Imagen 40: Predicción de notebooks online

Luego se procedió a hacer lo mismo para el canal de distribución:

Algoritmo	MAPE
Naive estacionaria	0,79
Drift	0,99
TSLM	1,68

Naive	1,27
Exponential smoothing	1,09
Holt Winters	1,88
ARIMA	1,33
TBATS	1,54
Regresión lineal	1,7
XGBoost	0,67

Para el dataset de distribución, el algoritmo que más se acercó a los valores reales fue el de XGBoost, con un MAPE de 0,59. El mismo fue parametrizado de la siguiente manera:

nrounds	10
max_depth	2
eta	0.25
gamma	0
colsample_bytree	1
min_child_weight	1
subsample	1



Imagen 41: Predicción de notebooks distribución mediante XGBoost

Siendo que los tres modelos mejores modelos resultaron tener la característica de que arrojaron valores constantes en sus predicciones, se indagó sobre cuáles podrían ser las causas de esto y se teorizó con que se debe a cierta particularidad del conjunto de testeo. Si se observa la *imagen 11*, se puede observar un claro cambio de tendencia que se da a partir del primer semestre de 2022. Esta coincide casi de forma idéntica con la partición utilizada en este trabajo. Por esta razón, se cree que los datos con los que el modelo está entrenando no son acordes a la tendencia actual de ventas de la empresa en este rubro. Por cuestiones de tiempos de entrega del trabajo, no se continuó probando con otros métodos de partición. Sin embargo, se tomará en cuenta esto en caso de prorrogarse el proyecto.

Modelo Monitores

En cuanto al rubro monitores se puede observar en el gráfico a continuación que a partir del 2022 el canal con la mayoría de ventas se convirtió en el canal de distribución. Se podría decir que el fin de la cuarentena contribuyó en estancar la cantidad de ventas por canal

online. Por esta razón, para realizar el análisis de ventas de monitores se decidió agrupar las ventas de cada canal en uno único.

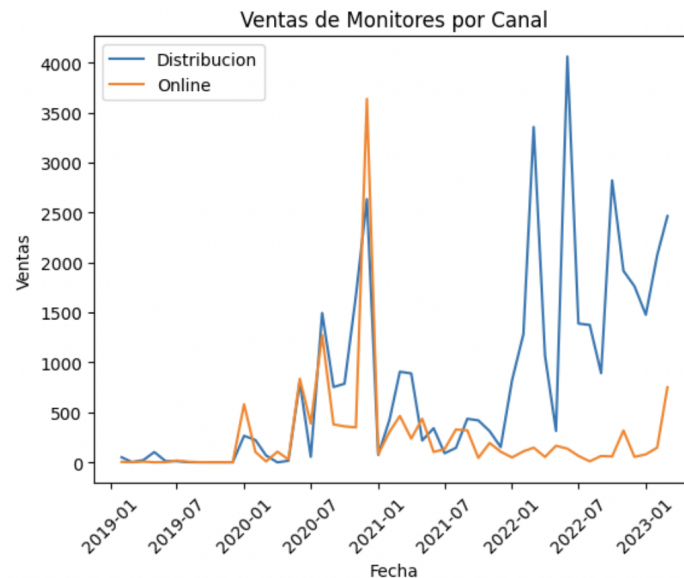


Imagen 43: Monitores por canal

Se llevaron a cabo experimentos con varios algoritmos de modelado, incluyendo Random Forest, XG Boost y Regresión Lineal Múltiple. Se realizó un ajuste de hiperparámetros en XG Boost y se probaron diferentes combinaciones de variables teniendo en cuenta su valor-p en la Regresión Lineal Múltiple. Sin embargo, el algoritmo que mostró los mejores resultados fue el modelo de serie de tiempo. A continuación se presentan los resultados.



Imagen 44: Predicción de monitores totales

Los resultados obtenidos con los distintos modelos que se probaron fueron los siguientes:

Algoritmo	MAPE
Regresión Lineal Múltiple	1,7
Random Forest	0,5
XG Boost	0,17
Naive	0,37
Holt Winters	0,47

El modelo Naive Bayes tuvo un MAPE de 0.3762, lo que indica que, en promedio, las predicciones difirieron del valor real en un 37.62%. Por otro lado, el modelo de Holt-Winters tuvo un MAPE de 0.4778, lo que indica que las predicciones difirieron del valor real en un 47.78%, en promedio. Por lo que, en este caso, el modelo Naive Bayes parece ser más preciso que el modelo de Holt-Winters para pronosticar las ventas de monitores.

Modelo Cocción

Hasta el año 2021, las ventas por ambos canales eran similares en términos de cantidad, lo que puede haberse atribuido a las restricciones de la cuarentena. Sin embargo, a partir de mediados de 2021, las ventas por el canal de distribución experimentaron un crecimiento significativo, mientras que las ventas en línea también aumentaron, pero en menor medida en comparación con el canal de distribución.

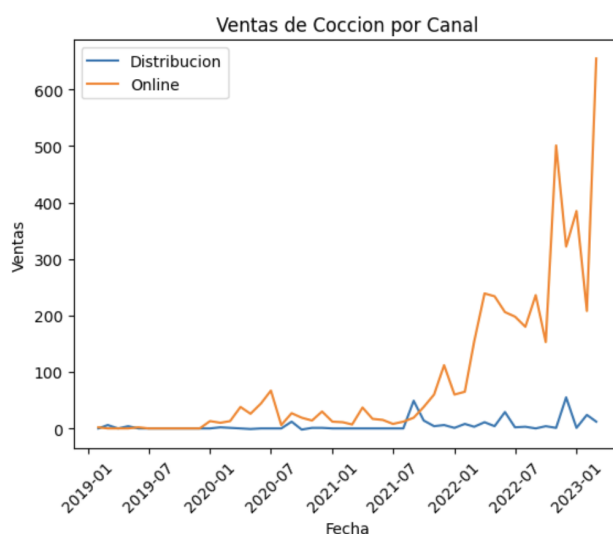


Imagen 45: Cocción por canal

Al igual que los demás modelos, se realizaron experimentos con diversos algoritmos de modelado para predecir las ventas en el rubro de cocción, incluyendo Random Forest, XG Boost y Regresión Lineal Múltiple. Se ajustaron los hiper parámetros en XG Boost y se probaron distintas combinaciones de variables en la Regresión Lineal Múltiple, teniendo en cuenta su valor-p. Sin embargo, luego de realizar todos los modelos mencionados, se llegó a la conclusión que el modelo que presentó los mejores resultados fue el de serie de tiempo. Se pueden observar los resultados obtenidos a continuación.



Imagen 46: Predicción de coccion total

Los resultados obtenidos con los distintos modelos que se probaron fueron los siguientes:

Algoritmo	MAPE
Regresión Lineal Múltiple	3,6
Random Forest	6,1
XG Boost	5,6
Naive	0,29
Holt Winters	0,50

Los resultados indican que el modelo Naive Bayes obtuvo un MAPE de 0.2982, lo que significa que en promedio las predicciones difirieron del valor real en un 29.82%. Por el contrario, el modelo de Holt-Winters tuvo un MAPE de 0.508, indicando que las predicciones difirieron del valor real en un 50.8%, en promedio. En este caso, el modelo Naive Bayes parece ser más preciso que el modelo de Holt-Winters para pronosticar las ventas del rubro cocción, ya que el MAPE obtenido por Naive Bayes es mucho menor que el MAPE obtenido por el modelo de Holt-Winters.

Modelo Impresoras

En lo que respecta al sector de impresoras, se puede apreciar en el gráfico siguiente que a partir de 2021 el canal online dejó de ser el líder de ventas. Es posible que la cuarentena haya contribuido a frenar la cantidad de ventas en línea. En la actualidad, se observa una correlación positiva entre las ventas por canal, lo que indica que cuando uno de los canales experimenta un aumento en las ventas, el otro canal también lo hace, y cuando uno disminuye, el otro también disminuye.

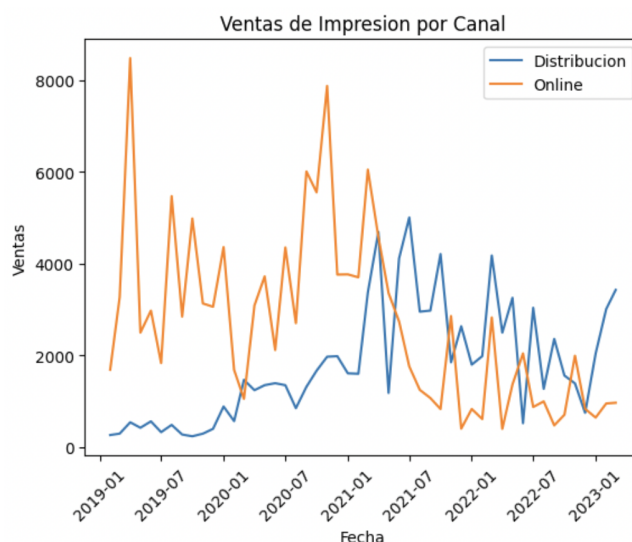


Imagen 47: Impresoras por canal

Al igual que los otros modelos, se realizaron experimentos utilizando diferentes algoritmos de modelado para predecir las ventas en el sector de impresoras. Entre ellos se incluyeron Random Forest, XG Boost y Regresión Lineal Múltiple. En XG Boost se ajustaron los hiperparámetros y en Regresión Lineal Múltiple se probaron varias combinaciones de variables, considerando su valor-p. Si bien con ningún modelo se llegó a la precisión deseada, se puede concluir que si se continúan realizando ajustes en la serie de tiempo se podría llegar a una mayor precisión para modelar. A continuación se muestran los resultados obtenidos.

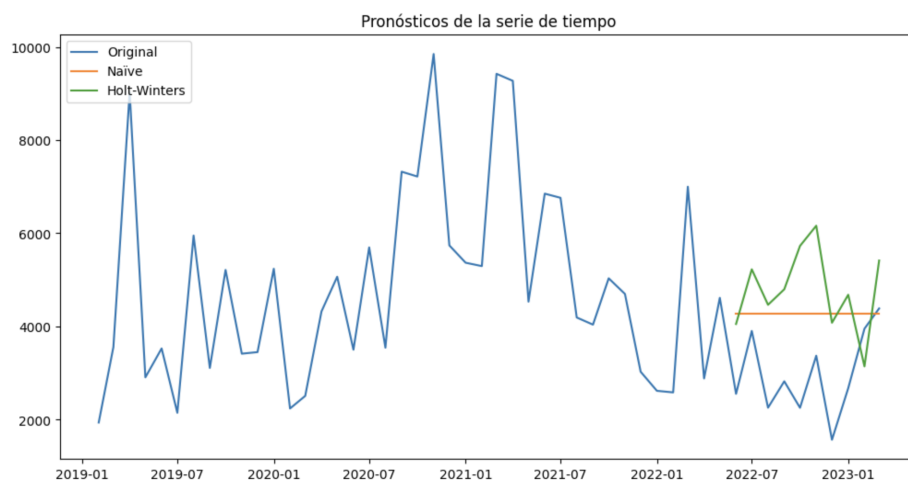


Imagen 48: Predicción de impresoras totales

Los resultados obtenidos con los distintos modelos que se probaron fueron los siguientes:

Algoritmo	MAPE
Regresión Lineal Múltiple	4,1
Random Forest	3,3
XG Boost	1,7
Naive	0,57
Holt Winters	0,75

Como se puede observar, el MAPE de Naive Bayes para las ventas del rubro de impresoras fue del 0.5779, lo que indica que en promedio, las predicciones difirieron del valor real en un 57.79%. Mientras que el MAPE de Holt-Winters fue del 0.7755, lo que indica que las predicciones difirieron del valor real en un 77.55% en promedio.

Estos resultados sugieren que ambos modelos tuvieron dificultades para hacer predicciones precisas sobre las ventas del rubro impresoras. Sin embargo, es importante tener en cuenta que los modelos son tan precisos como la calidad de los datos de entrada que se utilizan para

entrenarlos. Por lo tanto, se podría trabajar en un mayor tratamiento de las variables, lo que podría mejorar la precisión del modelo y reducir el error de predicción.

RESULTADOS Y CONCLUSIONES

En conclusión, después de experimentar con varios algoritmos de modelado, incluyendo Random Forest, XG Boost y Regresión Lineal Múltiple se encontró que los modelos de series de tiempo presentaron los mejores resultados. Se obtuvo un MAPE promedio de 0.30 en la mayoría de los modelos, lo que indica que en promedio, las predicciones difirieron del valor real en un 30%. A pesar de esto, se sigue trabajando en mejorar la precisión de los modelos y reducir los errores en las predicciones, para poder brindar pronósticos más precisos y útiles para la toma de decisiones.

Bibliografía

Competición Grupo Bimbo:

<https://www.kaggle.com/competitions/grupo-bimbo-inventory-demand/overview/description>

Artículo de B. Pavlyshenko: <https://www.mdpi.com/2306-5729/4/1/15#B36-data-04-00015>

Artículo de A.
Modoloria: <https://mobidev.biz/blog/machine-learning-methods-demand-forecasting-retail>

Presentación:

<https://docs.google.com/presentation/d/1MeAWBjEV-Kdx8ZDJi6mxYGh031loGXHF/edit?usp=sharing&ouid=117464510099101037580&rtpof=true&sd=true>