Proyecto Final
Ciencia de Datos Aplicada
Universidad de los Andes

Jairo Vladimir Chaparro Rodríguez 201531080
Oscar Duvan Giraldo Romero 202324939
Andrés Mauricio Martínez Celis 202322624
Danilo Andrés Alfonso Bohórquez 201611827

Proyecto Final de Estrategia enfocado a Comercial Química

1. Definición de la problemática y entendimiento del negocio:

Selección de la Organización y Problemática

Comercial química es una empresa fundada en el año 2001, creada con el fin de responder a las necesidades puntuales de materias primas en la industria farmacéutica en Colombia, buscando el estricto cumplimiento de los más altos estándares de calidad en dichos insumos. Actualmente, la organización opera bajo el modelo de negocio business to business, siendo un distribuidor o intermediario para las grandes empresas de estos sectores.

Actualmente, la empresa está atravesando una problemática relacionada con la gestión de inventarios. En este momento el proceso se realiza a través del uso de plantillas en Excel. Sin embargo, este método ha causado problemas debido a su baja eficiencia y su dificultad para visualizar la información. Por lo tanto, nuestra intención como grupo es crear un dashboard que permita al cliente observar el estado de su inventario, es decir que pueda crear alertas cuando alguna materia prima esté en sobre stock o out of stock al igual que identificar los índices de rotación para dichos productos. De esta manera, una vez el dashboard logre tener el control sobre el inventario, este debe realizar una predicción que a final del año permita calcular el presupuesto esperado para el siguiente año calendario.

Documentación de Información Clave del Negocio

En Colombia, el sector de la industria farmacéutica genera gran movimiento a nivel de importaciones, dado que en el país no se producen la mayoría de las materias primas que demanda esta industria. Según la base de datos LegisComex, entre enero de 2022 y diciembre de 2022 Colombia importó aproximadamente 32,886,997 kg de material con partida arancelaria 2918, la cual contiene 30 subpartidas, donde cada una representa diferentes materias primas necesarias para la producción de fármacos y otros productos.

De acuerdo a lo anterior, en este momento la empresa cuenta con información detallada de su operación desde 2020. Actualmente, la plantilla de Excel que manejan con la información contiene para cada mes del año la cantidad disponible de cada materia prima, los saldos, el precio unitario y el precio de venta total de cada transacción para cada uno de los productos. Adicionalmente, en esta misma plantilla también se manejan los índices de rotación por producto para cada cuatrimestre del año.

Definir Objetivos del Proyecto y Métricas de Negocio

Objetivos:
Proveer una solución visual que permita a la empresa tener un control en tiempo real de su inventario con el fin de establecer estrategias de negocio enfocadas en mejores prácticas de almacenamiento, al igual que eficiencia en la rotación de los productos con el fin de maximizar su ingreso.

Minimizar los problemas existentes en términos de eficiencia operativa, ya que mucha de esta información se construye de manera manual y no existe un estándar óptimo que permita conservar esquemas deseados en términos de calidad y gobierno de datos de dicha información.

Desarrollar un modelo para anticipar de la manera más precisa el presupuesto requerido para el próximo año, con un énfasis en la proyección de los niveles de inventario y las necesidades de inversión.

Métricas de Negocio:
Una de las métricas que nos permitirán observar eficiencia en el proceso de clasificación de inventarios, es el índice de rotación en niveles óptimos, ya que esto permite optimizar costos de almacenamiento y a la vez realizar un flujo adecuado de los insumos.Asimismo, la optimización de las cantidades importadas de acuerdo a la demanda de productos tanto en la compañía como en el sector de materias primas.

Por otro lado, utilizando datos de 2020 a 2023, se pueden realizar iteraciones que permitan desarrollar un modelo de predicción de presupuesto que minimice el error, brindando apoyo al área financiera en la proyección del presupuesto del próximo año así como la determinación de precios.

2. Ideación:
Los potenciales usuarios del producto de datos detallados a continuación son los empleados de la empresa de distribución farmacéutica que están involucrados en la gestión del inventario, la producción o la satisfacción del cliente.

1. Auxiliar de Gerencia de Inventarios y Compras:
    - Responsabilidades: Supervisar el inventario y las compras, realizar un seguimiento de los niveles de stock y gestionar las adquisiciones de productos.
    - Dolores Actuales: Dificultad para prever las necesidades de inventario y tomar decisiones informadas sobre las compras.
2. Compras Internacionales:
    - Responsabilidades: Encargado de adquirir materias primas según los requerimientos de la empresa, gestionar proveedores internacionales y asegurar la disponibilidad oportuna de insumos.
    - Dolores Actuales: Falta de datos precisos para optimizar las compras internacionales de manera predictiva y reducir costos.

3. Líder de Aseguramiento de Calidad y Cumplimiento de Indicadores:
    - Responsabilidades: Supervisar y garantizar el cumplimiento de los estándares de calidad de los productos y el seguimiento de los indicadores clave de rendimiento.
    - Dolores Actuales: Dificultad para acceder a información relevante para el control de calidad y el seguimiento de índices de inventario.

Producto final:
Dashboard que integra la Proyección de Presupuesto y la detección de alertas en los índices de rotación, además de un Modelo de series de tiempo de ventanas móviles para prever el presupuesto ideal de adquisición de materia prima para el año 2024.

Requerimientos:

- El sistema deberá mostrar de manera uniforme los datos relacionados con las ventas y su impacto en el inventario de la compañía con el propósito de respaldar la toma de decisiones, se debe desarrollar un modelo (Dashboard) que refleje los cambios en el inventario derivados

de las ventas, ofreciendo la capacidad de filtrar por material (producto), periodo y proveedor. Este recurso será esencial para una gestión eficiente del inventario.

- El sistema debe tener la capacidad de actualizar la información y cargar datos en los diversos escenarios previstos, que incluyen el cargue inicial y la actualización de información. Cuando estos cambios se presenten, se recalculará y actualizará las gráficas en el Dashboard. Esto según lo dictaminado ya sean cargas manuales o conexiones a recursos externos de la compañía en línea, dependiendo del sistema seleccionado.
- El sistema deberá ser capaz de recopilar y procesar los datos de inventario y ventas proporcionados por el administrador del mismo. Todas las reglas de negocio deben estar claramente definidas.
- El sistema debe generar alertas visuales según las reglas de rotación del inventario, esto con el fin de informar cuando haya un uso indebido de los recursos que puedan generar gastos adicionales en almacenamiento y demás causas por el sobre stock o insuficiencia del mismo.
- El sistema debe estar alojado en la red y debe permitir el acceso a los usuarios descritos en la definición de roles y usuarios.
- Utilizando los datos recopilados, se debe crear una proyección del presupuesto de la empresa con el fin de evaluar el cumplimiento de los objetivos establecidos para inventarios y ventas.

3. **Responsible:**

Implicaciones Éticas: En este proyecto no es significativo el riesgo de sesgo en los datos que puedan resultar en decisiones injustas o discriminatorias. No obstante, se manejará un esquema de responsabilidad y rendición de cuentas con el cliente, ya que este es responsable de las decisiones tomadas por el modelo. Se creará una bitácora donde se lleve registro de los avances del proyecto para poder tener un rastreo y justificación en el proceso de toma de decisiones del modelo.

Privacidad y confidencialidad: Se obtuvo el consentimiento previo por parte del cliente para utilizar sus datos a través de una carta. En ese documento se le explicó al cliente la intención del proyecto y también se aclara que los datos recolectados serán utilizados únicamente para el proceso de investigación. Además, se asegura que los datos no se compartirán con terceros o con cualquier agente externo a la empresa que pueda afectar a la organización, económica y socialmente.

Transparencia: En cuanto a la divulgación de métodos y datos, al cliente se le proporcionará información detallada de los métodos a utilizar y los conjuntos de datos que se emplearán. Adicionalmente, se compartirán con el cliente los resultados y hallazgos que se obtengan durante el desarrollo del proyecto.

Aspectos regulatorios: Existen una serie de regulaciones que pueden afectar al uso de datos y técnicas de IA en el contexto de la problemática abordada. Estas regulaciones pueden variar de un país a otro. En Colombia, algunas de las regulaciones relevantes incluyen la Ley 1581 de 2012, sobre protección de datos personales, el decreto 1377 de 2013, que reglamenta la Ley 1581 de 2012 y el reglamento General de Protección de Datos (RGPD), de la Unión Europea,sin embargo, en el contexto utilizado no se manejarán datos de usuarios, únicamente se trabajarán con información suministrada por la organización y hace referencia a productos comerciales adquiridos por la compañía.

Seguridad y protección contra ataques: Para el desarrollo de este proyecto, nos acogemos a los protocolos de seguridad y protección contra ataques que tenga el cliente. Por otro lado, para mitigar las posibles vulnerabilidades que podrían ser explotadas por terceros malintencionados, se ha

decidido solo compartir información con el cliente a través de medios oficiales como el correo electrónico institucional.

Monitoreo y evaluación continua: Una vez se haga entrega del proyecto es responsabilidad del cliente tomar las decisiones necesarias para poder llevar a cabo el proceso de monitoreo y evaluación continua.

4. **Enfoque analítico:**

Pregunta de Negocio 1: ¿Cómo podemos aportar información sobre el stock disponible y las ventas para minimizar los costos de almacenamiento sin comprometer la disponibilidad de productos especialmente aquellos que tienen índices de rotación menores a 1? 25 y mayores a 0.75?

Pregunta de Negocio 2: ¿Cómo lograr una identificación a través del modelo de predicción del presupuesto de los 10 productos con mayor potencial de oferta y a la vez demanda en la empresa?

- Análisis exploratorio de datos (EDA): Se podría utilizar un análisis descriptivo para identificar que la demanda de algunos productos es estacional, mientras que la demanda de otros productos es más constante y de esta manera definir
- Modelos de aprendizaje automático mediante ventana deslizante: Se empleará un modelo de aprendizaje automático que utilice datos históricos de los presupuestos de los últimos tres años para predecir de manera más precisa el presupuesto del próximo año.

El índice de rotación en los diferentes niveles mostrará si la solución visual (dashboard) permite tomar mejores decisiones en terminos de compra y venta de manterias primas, al igual que el ajuste de presupuesto para insumos con alta demanda. Por otro lado, el seguimiento de ventas por proveedor permitirá observar aumento de la demanda o disminución de la misma. Finalmente, el margen entre el presupuesto calculado por el modelo y el presupuesto asignado a principio del año.

5. **Recolección de datos**

Dada la delicadeza de los datos y su potencial impacto en las operaciones comerciales, nuestra única fuente de información es la proporcionada por la compañía. Estos datos, recopilados desde el año 2020, representan un recurso invaluable que nos permitirá llevar a cabo las proyecciones necesarias para tomar decisiones estratégicas.
Para comenzar a comprender los datos, nos enfocaremos en la siguiente información que, según nuestros resultados esperados, se proyecta como crucial para definir las estrategias propuestas en este informe:

- Años de operación (2020, 2021, 2022, 2023): Esta información se divide en tres secciones relevantes:
    - La primera sección define las ventas por producto y el mes en que se realizaron.
    - La segunda sección ofrece detalles sobre las compras realizadas durante el año, desglosadas por meses.
    - La última sección calcula el inventario total para cada mes mediante una operación aritmética que consiste en restar las ventas a las compras. Esta información es

fundamental para calcular el índice de rotación, que es esencial para los stakeholders. Todos estos datos son de naturaleza cuantitativa.

- Datos de ventas: Esta hoja de cálculo contiene información relevante, como el año, el mes, la fecha de contabilización, la descripción del artículo, el nombre del comprador, la cantidad y atributos específicos del producto. Estos datos, tanto cuantitativos como categóricos, desempeñarán un papel crucial en la predicción del comportamiento y en la formulación de recomendaciones, incluyendo porcentaje de crecimiento sugerido y presupuesto para el próximo año.

6. **Entendimiento de los datos:** Se adjunta el notebook del análisis exploratorio y calidad de datos en el repositorio.

7. Conclusiones/Insights:

Esta primera entrega nos ha permitido como grupo identificar el modelo de negocio y las áreas de mejora que comercial química tiene en su proceso de inventarios y presupuesto. Por otro lado, también se ha logrado hacer un análisis exploratorio de los datos, logrando estructurar la estrategia a seguir y la idea inicial del producto que se quiere obtener. De esta manera se logra cumplir con los primeros pasos del modelo ASUM-DM, los cuales son entendimiento del negocio y aproximación analítica.

En ese sentido,en el análisis exploratorio se pudo observar que la mayor parte de los datos se encuentran completos, no obstante, se normalizaron aquellos que podrían ser considerados como valores atípicos como lo son los precios finales por transacción, lo cual nos permite evidenciar que el precio medio por cada una de las transacciones oscila en 12.915.144 pesos colombianos, asimismo, las regiones en las cuales se ubican los proveedores que tienen mayor demanda son los departamentos de Cundinamarca y Antioquía, ya que en Bogotá se ubica el 56% de estos clientes.

Por otro lado, el nivel de transacciones se ubica entre los 10 y los 8 millones de pesos, lo cual se relaciona fuertemente con las cantidades distribuidas demostrando una relación lineal fuerte entre estas dos variables. Asimismo, los medicamentos con mayor cantidad de productos comprados al igual que aquellos productos con el precio unitario más alto se identifican para lograr determinar potenciales aumentos de presupuesto y rotaciones eficientes. De igual forma, frente a los atributos adicionales, como la presentación y la administración del documento se observa que las inyecciones y las cremas son los más costosos y que aquellos que tienen una vía de administración oral son los más preferidos por los clientes.

Frente a los próximos pasos a implementar en esta estrategia se tiene la selección de datos relevantes en la cual se pretende identificar aquellos que son clave para el problema. Seguido de esto, realizar la limpieza y preparación de esta información con el fin de lograr definir la estrategia de modelado la cual puede incluir el uso de algoritmos y métricas y finalmente crear los primeros modelos de regresión de series de tiempo de ventanas móviles con el fin de tener aproximaciones previas al producto final.

8. Preparación de datos

Para la preparación y acondicionamiento de los datos se realizaron los siguientes pasos:

- Cambio de etiquetas
- Eliminación de datos innecesarios
- Tratamiento de ciudades faltantes
- Cargar información adicional del inventario
- Tratamiento de datos separados (lotes informados)
- Eliminación de datos
- Unión de información
- Asignación de One Hot Encoding
- Verificación del campo “Total líneas”

Al final se exporta un archivo único que tiene toda la información necesaria para realizar el entrenamiento de los diferentes modelos. No obstante, el procesamiento detallado de cada uno de los pasos se puede verificar en el diagrama de flujo que se encuentra adjunto en el repositorio del proyecto en GitHub.

9. Estrategia de validación y selección de modelo:

Selección del modelo

Como se ha mencionado anteriormente, la empresa necesita de herramientas que le permitan mejorar la toma de decisiones en cuanto a la definición del presupuesto. Por lo tanto, el objetivo del modelo de machine learning que se va a diseñar, es poder predecir con cierta precisión las ventas de Comercial Química para el siguiente año basándose en el comportamiento de las ventas de años anteriores. De esta manera la empresa puede tomar mejores decisiones a la hora de establecer un presupuesto.

Teniendo en cuenta que el modelo debe realizar pronósticos basados en datos históricos consideramos diferentes opciones, las cuales fueron un modelo de regresión lineal y modelos de series de tiempo. Como primer escenario o modelo baseline utilizaremos la regresión lineal. Si bien no es el modelo ideal para este caso, sí nos sirve para establecer un punto de partida que permita establecer las métricas necesarias. Una vez se establece dicho modelo baseline, aplicaremos modelos de series de tiempo para poder realizar predicciones mejor ajustadas a nuestro escenario.

Dichas predicciones se realizarán utilizando 3 modelos diferentes. El primer modelo es ARMA (Modelo Autorregresivo de Media Móvil), que combina modelos autorregresivos (AR) y de media móvil (MA). Los modelos autorregresivos modelan la relación entre una observación y un número de observaciones retrasadas. Por otro lado, el modelo de media móvil (MA) modela la relación entre una observación y un error blanco residual retrasado.

El segundo modelo que se usará es ARIMA (Modelo Autorregresivo Integrado de Media Móvil). Este modelo complementa ARMA ya que incluye la diferenciación para hacer que la serie temporal sea estacionaria. Esta diferenciación implica sustraer la observación anterior de la actual para eliminar la tendencia y/o la estacionalidad. Es adecuado para series temporales que tienen una tendencia y no tienen componentes estacionales.

El tercer modelo es VAR (Modelo Autorregresivo Vectorial), el cual es un modelo multivariante que extiende el enfoque autorregresivo a múltiples series temporales. Modela las relaciones lineales

entre múltiples series temporales, permitiendo que cada una sea una función lineal de sus propios valores retrasados y los valores retrasados de las otras series. Es muy útil para analizar la interacción dinámica entre múltiples series temporales y es especialmente útil en sistemas económicos y financieros donde las variables están interrelacionadas.

Finalmente, se tendrán en cuenta modelos más complejos como las redes neuronales recurrentes, prophet y Xgboost con el fin de validar alternativas complementarias a las series de tiempo.

Preparación de los Datos para el modelo:

En los modelos que vamos a realizar dado que se va a trabajar con series de tiempo, es clave que se tenga estacionariedad en los datos, ya que varios de estos asumen dicha propiedad en la serie. Cuando se habla de estacionariedad significa que en la serie temporal sus propiedades estadísticas como la media, la varianza y la autocorrelación deben ser constantes a lo largo de tiempo.

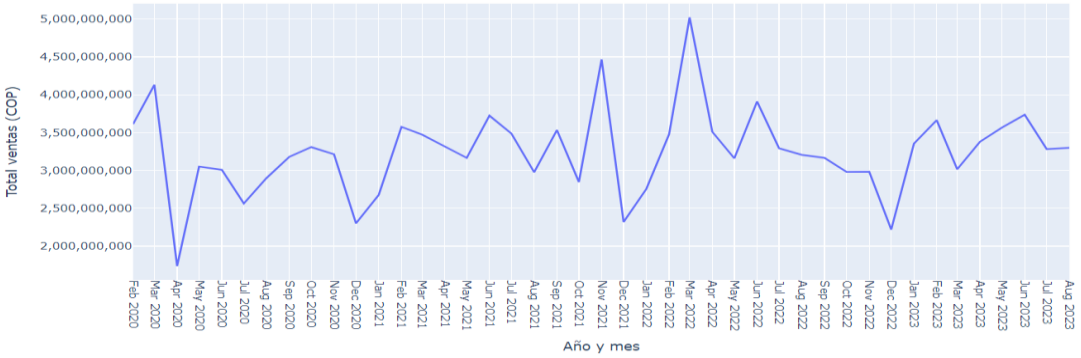
Para asegurarnos que esta propiedad se diera, el data set se reorganizó de diferentes maneras. En el caso de la regresión lineal se sumaron todas las ventas de cada mes, permitiendo que desde el 2020 hasta el 2023 se tenga un valor asignado y así cumplir con la estacionariedad. Luego se ordena el dataset de manera ascendente por fecha, para que los primeros datos sean los más antiguos y los últimos sean los más recientes.

Por otro lado, en el caso de las series de tiempo del cuaderno número 1, se llevaron a cabo pruebas con diversas combinaciones para la partición de los datos de prueba y entrenamiento. Estas decisiones se tomaron en función del tipo de modelo que se estaba validando.

Para una serie temporal de frecuencia diaria, se optó por dividir la muestra en un 85% para entrenamiento y un 15% para prueba, debido a la naturaleza de los datos. En cambio, para modelos como Red LSTM y SARIMAX, se eligió utilizar datos comprendidos entre el 3 de febrero de 2020 y el 31 de diciembre de 2022 para el conjunto de entrenamiento. Para el conjunto de prueba, se tomaron en consideración los datos entre el 1 de enero de 2023 y el 31 de agosto de 2023.

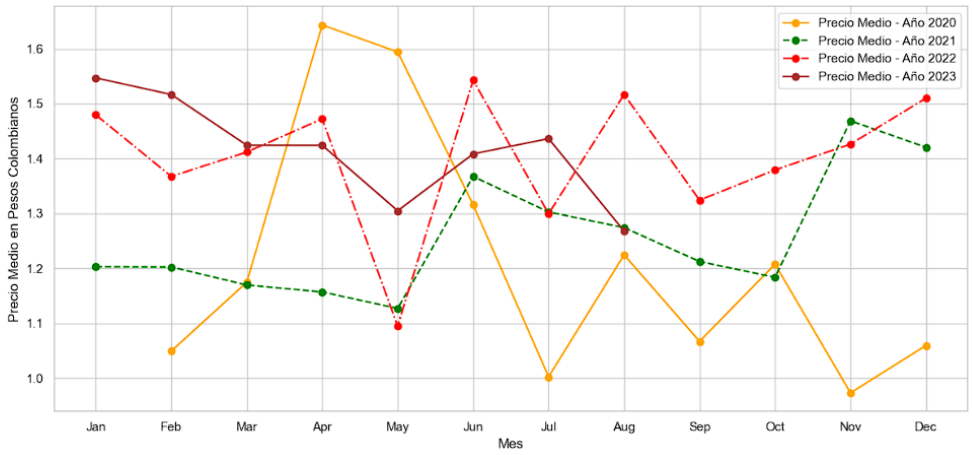
Por otro lado, en la división de datos de entrenamiento y prueba para series temporales mensuales usamos train_test_split, pero especificando en los argumentos “shuffle=false”. De esta manera aseguramos que no se utilicen datos aleatorios en cada partición, sino que se respete la temporalidad, utilizando el 80% de los meses para entrenamiento y el resto para pruebas.

Ilustración 1, distribución mensual de la muestra



La preparación de datos se realizó respetando la muestra, solo se agregaron las ciudades del proveedor con base al dataset compartido de ventas y el otro de inventarios, el lote y la descripción son claves de búsqueda para generar la unión con otros dataset que complementan la información, la perdida de datos fue baja, se quitan los saldos iniciales que son 891 datos y también los datos que no poseen cruce entre descripciones que son 38, los cuales fueron eliminados en la cláusula exclusiva inner join lo que deja 929 datos eliminados por reglas de negocio; esto es el 9,2% de la muestra inicial de 10.092 datos, a continuación se muestra una imagen (ilustración 2) de los datos antes de ser procesados que se contrastan con la ilustración 1.

Ilustración 2, distribución mensual de la muestra antes de ser trasformada



10. Construcción del modelo:

1) Regresión Lineal (Modelo Baseline)

En este modelo se realizaron dos escenarios diferentes que permitieran establecer una base para los demás modelos. En el primero se realizó una predicción sin usar ventanas deslizantes y en el otro se realizaron predicciones utilizando 2 ventanas deslizantes. El resultado en ambos casos fue el esperado, un error cuadrático medio sumamente alto. Esto se debe primero a la falta de datos suficientes y también que la naturaleza del modelo de regresión lineal no se acomoda a este escenario ya que no incorpora el tiempo como variable de manera intrínseca y no es capaz de captar las relaciones dinámicas que existen con el cambio de tiempo ya que asume una relación lineal constante entre las variables independientes y la variable dependiente.

2) Modelo de Vectores Autorregresivos (VAR), Autorregresivo con medias Móviles (ARMA) y Modelo Exponencial de Series de tiempo para series de tiempo mensuales

Para analizar series de tiempo se comienza con la carga de datos desde un archivo Excel y la preparación de los datos mediante la agregación de transacciones mensuales. Luego, se exploran diferentes configuraciones de parámetros AR y MA para seleccionar el modelo óptimo en función de métricas como MSE, RMSE y AIC. Se realizan pruebas para validar supuestos de estacionariedad y distribución de residuos. A pesar de la selección del mejor modelo, se enfatiza que otros factores y métricas deben evaluarse para garantizar un rendimiento óptimo en aplicaciones comerciales.

3) Modelo ARIMA, SARIMA y Modelos Adicionales como Prophet y Red LSTM para series de tiempo diarias

Se valido la muestra de la serie de tiempo y se decidió optar por una frecuencia diaria, con el fin de tener más datos para mejorar la robustez de los modelos, en primer lugar, se ejecutaron series de tiempo ARIMA con diferentes combinaciones de vectores autorregresivos, cointegración y medias móviles, con el fin de minimizar los errores y optimizar el rendimiento esperado. No obstante, a pesar de utilizar varias diferenciaciones de la serie las métricas de validación no cumplen las expectativas, en ese sentido, se probaron alternativas con variables exógenas, sin embargo, se pudo evidenciar que el rendimiento era similar a los modelos expuestos anteriormente por lo que esto no representa una diferencia significativa a la hora de determinar que tipo de modelo usar.

Por otra parte, modelos ARMA y XGBoost no logran capturar de manera eficiente la estructura de la serie temporal ya que el primero de estos, evidencia residuos que evidencian heterocedasticidad y el segundo solo es capaz de predecir algunos días en específico con métricas de error reducidas, pero no está en la capacidad de predecir con la muestra planteada. Finalmente, modelos más robustos como Prophet y LSTM evidencian mejores métricas, sin embargo, son modelos pensados para trabajar con grandes cantidades de datos ya que el primero de estos trabaja con datos históricos que contienen estacionalidad y el segundo es una red neuronal recurrente.

11. Evaluación del mejor modelo

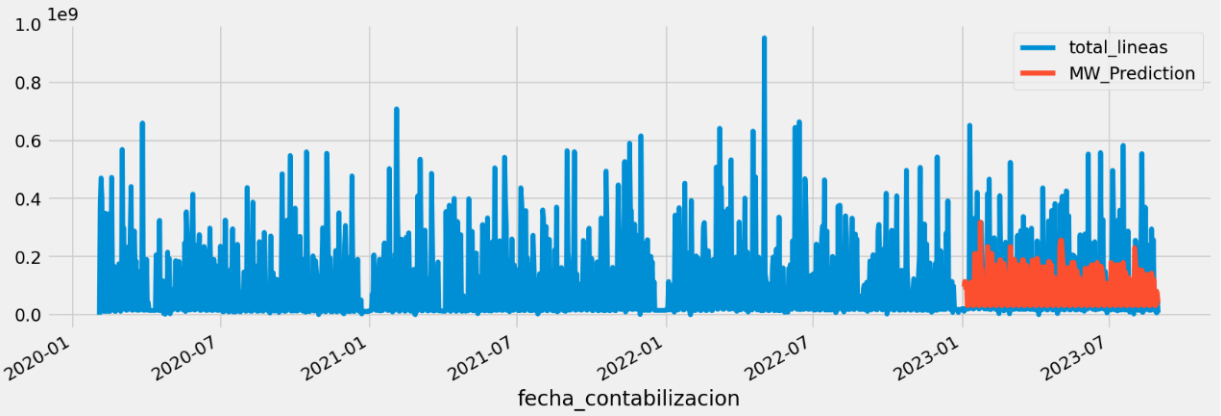
Para la evaluación del mejor modelo aplicado se elegirá uno de una serie temporal mensual y uno diario, con el fin de comparar dichos rendimientos y elegir aquel que mejor se ajuste a los valores actuales.

1) Serie Temporal Diaria

El modelo que muestra un mejor rendimiento es el de la serie temporal de XGboost el cual se entrenó con los datos comprendidos entre el 03 de febrero de 2020 y el 31 de diciembre de 2022 y se probó con los comprendidos entre el 01 de enero de 2023 y el 31 de agosto del mismo año. En ese sentido, se utilizaron 1000 estimadores y se desagrego la fecha en diferentes variables con el fin de determinar la importancia de estas en la contribución de la predicción, donde se determinó que el día de la semana y el día del año son los variables de mayor importancia. Luego de esto, el modelo predijo los datos para las fechas del año 2023, donde la peor predicción obtuvo un error de 213623615,87 que corresponde al 01 de mayo de 2023 mientras que la mejor predicción obtuvo un error de 1350476 correspondiente al 30 de marzo de 2023. Asimismo, a nivel agregado las métricas de predicción se comportan de la siguiente manera:

Métrica	Resultado
RMSE	1.4242416806635096e+16
MAE	82,843,241.06

Se puede deducir que a nivel general el error es muy alto, no obstante, cuando se validan los días predichos podemos observar que el error más alto es de 213 millones, por lo cual podríamos ajustar este modelo para generar recomendaciones enfocadas en aquellos días en los cuales se tienen mejores métricas. La distribución de la serie de tiempo con los valores actuales y predichos se puede observar en el siguiente esquema:

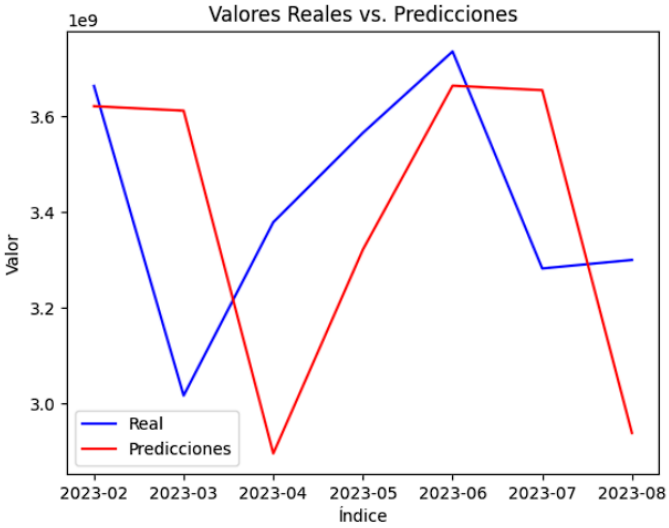


2) Serie Temporal Mensual

El mejor modelo de series de tiempo, en temporalidad mensual seria ExponentialSmoothing, el modelo demostró que aunque tiene un error mayor que el VAR(Vector Autorregresivo), se acopla mejor a la serie temporal a lo largo del tiempo mostrando varianzas en periodos de hasta 35 meses, el VAR en cambio mostró que después de la predicción 12, tiende a quedarse en el mismo momento de predicción, por esto ExponentialSmoothing mostro que cada 4 meses el periodo cambia y genera una mejor predicción, en la siguiente tabla se muestra las métricas de predicción:

Métrica	Resultado
RMSE	363,807,793.10$
MAE	310,510,803.03$

De acuerdo con lo anterior se puede deducir que el error que se genera es de 363 millones al mes en todas las ventas de la compañía un error cerca de 9.47% de las ventas totales, a continuación, se muestran las predicciones del modelo a los primeros meses del año 2023, los anteriores fueron datos de entrenamiento:



12. Conclusiones:

En conclusión, la falta de datos afecta seriamente el rendimiento de los modelos ya que, en el caso de la predicción de series de tiempo a nivel mensual, luego de realizar varias iteraciones dichos modelos no muestran variaciones significativas y no conservan linealidad, lo que no evidencia una estructura óptima con la línea temporal. Asimismo, en las series diarias, los días que no tienen datos generan problemas de consistencia en los modelos ejecutados, por lo cual las predicciones no resultan eficientes, lo que conlleva a determinar que para este tipo de análisis los datos históricos son relevantes y se debe garantizar estacionariedad en las muestras.

Por otra parte, en los datos compartidos por el cliente, se pudieron evidenciar valores atípicos, los cuales corresponden a errores de digitación de los diferentes registros, en ese sentido, esto representa una dificultad a nivel de calidad de datos lo cual genera sesgo y ruido en el resultado esperado, por ende, es importante que la limpieza y calidad de datos sea un paso fundamental antes de la validación de los modelos planteados.

Frente a las dificultades del proyecto, como ya se mencionó anteriormente la escasez de datos es un factor que representa inconvenientes a la hora de ejecutar los diferentes modelos predictivos, sumado a esto registros que corresponde a ajustes de software de seguimiento de inventario generaban categorías con saldos iniciales que debían ser eliminados ya que no corresponden a materias primas las cuales son el objeto del análisis. De igual manera, la falta de completitud de datos relevantes para el análisis como el inventario, representa dificultades a la hora de estandarizar las muestras y contar con más información que permita aplicar alternativas de modelado diferentes a las ejecutadas.

Las estrategias óptimas para solventar este tipo de problemáticas están orientadas hacia la recolección de más datos, los cuales pueden darse a través de “over sampling” o en dado caso, que el cliente provea datos adicionales a los ya compartidos, asimismo, la eliminación de los registros que no son considerados necesarios para el análisis por la naturaleza de estos.

De acuerdo con lo anterior, las condiciones necesarias que deberían cumplir los datos son la estacionariedad, el cual es un factor determinante pero que a su vez no está presente en esta muestra lo que dificulta el análisis como se pudo evidenciar en las métricas de evaluación de los modelos respectivos. Adicional a esto, se debería contar con variables adicionales que complementen la información actual con el fin de poder ejecutar modelos más robustos como por ejemplo vectores autorregresivos o modelos ARIMA con diferentes tipos de parámetros.

Frente a la solución dada por los modelos, estos no logran capturar la sensibilidad de los factores asociados al mercado, como la competencia, la inflación, la demanda y la oferta, entre otros, debido a que, en cierto punto, estos no logran predecir de manera optima los ingresos esperados por las ventas. Tanto, en las series temporales diarias como mensuales, las predicciones con errores relativamente bajos, solo se dan en pequeños periodos de tiempo, sugiriendo que la falta de información definitivamente impide pronosticar mayores periodos de tiempo, conservando la estructura de la tendencia temporal.