Ingeniería de Sistemas y Computación
Escuela de Posgrado
MINE-4101: Ciencia de Datos Aplicada

Integrantes:

- Álvaro Andrés Castiblanco López
- Camilo Andrés Morrillo Cervantes
- Lorraine Jazlady Rojas Parra
- Vladimir Emil Rueda Gómez

ENTREGA 2– PROYECTO FINAL

Definición de la problemática y entendimiento del negocio:

Para tener una comprensión adecuada de la problemática a solucionar, es fundamental comprender el funcionamiento de Sika, por eso, a continuación, se hará una breve descripción de su modelo de negocio.

Sika es una empresa global de productos químicos tanto para la construcción como para la manufactura. **Sika desarrolla y comercializa especialidades químicas para impermeabilizar, adherir, amortiguar, reforzar y proteger estructuras.**

Fundada en Zurich, Suiza en 1910, abre su primera sede en Bogotá en el año 1951 y hoy en día cuenta con más de 400 empleados en siete oficinas regionales: **Bogotá, Barranquilla, Bucaramanga, Cali, Medellín y Pereira**, con las cuales suple las necesidades del mercado de la construcción en el territorio colombiano.

Sika posee dos líneas de negocio principales: productos comerciales y productos técnicos. En la figura 1 se observa su distribución.

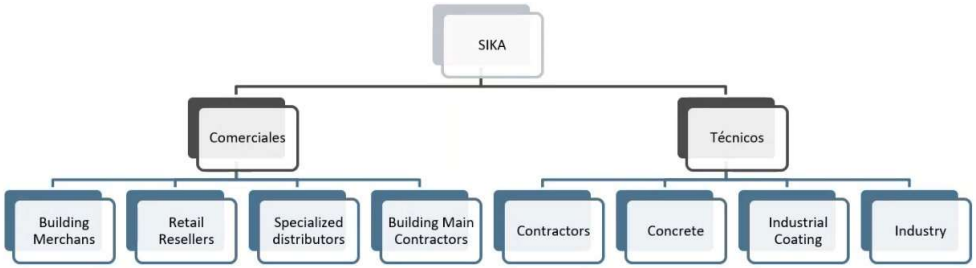


Figura 1. Línea de negocio de Sika.

Los productos comerciales son aquellos que no necesitan de ningún nivel experticia para su aplicación, en contra posición, los técnicos deben ser aplicados por personal certificado o se

corre el riesgo de que el producto se estropee y en el peor de los casos que se generen daños irreparables a la estructura.

La figura 2 ilustra como se distribuyen estos productos entre sus múltiples sub-líneas de negocio.

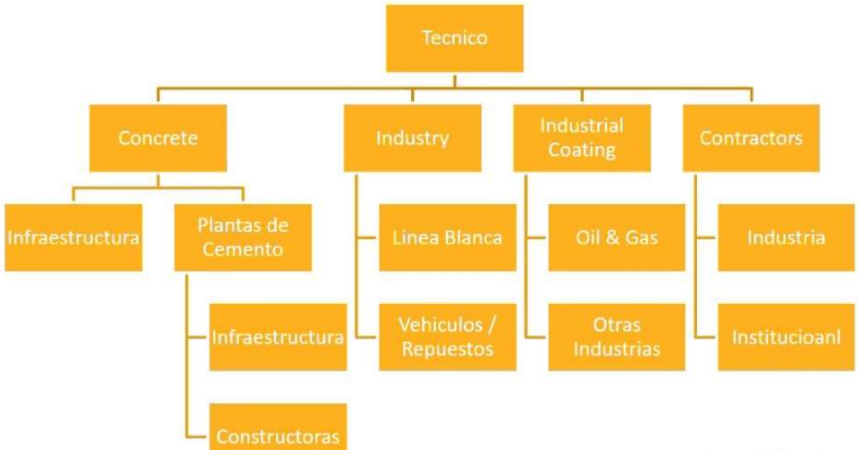


Figura 2. Sub-líneas de negocio técnicas de Sika.

Con todas las aclaraciones anteriores, podemos enfocarnos en el problema a solucionar, este es el de la planificación de demanda de la línea técnica, lo que comprende tanto insumos de fabricación de productos como concreto, barnices o esmaltes y productos listos para su venta.

El objetivo principal de la solución será:

- **Realizar un modelo de planificación de demanda a corto (un mes) y mediano plazo (cuatro meses). Con el fin de generar una estimación correcta para evitar tanto subestimación como sobrestimación**

Los periodos de tiempo del primer objetivo se deben a que Sika debe planificar su inventario con un mes de antelación para los productos nacionales y con al menos cuatro meses para los que deben importar.

Sika cuanta actualmente un proceso de planificación de demanda, pero debido a su simpleza y altos porcentajes de error no satisface las demandas de calidad que una empresa como Sika necesita.

El proceso actual tiene rangos de error bastante grandes debido a sobrestimaciones, lo que genera exceso de inventario y puede llevar a la caducidad de los productos y peor aún, subestimación que puede generar déficit de inventario, lo que afecta la credibilidad en Sika en el mercado.

La planificación de demanda es el input mediante el cual los gerentes regionales planifican la compra de insumos y finalmente se ponen de acuerdo con el área comercial. La solución por desarrollar debe precisamente mejorar ese input que reciben los gerentes regionales. Una subestimación aumentará los costos de almacenaje de producto y generará posibles vencimientos, lo que finalmente se traduce en pérdidas. **Una subestimación generará escasez de producto y pérdida de credibilidad en la empresa, por lo tanto, es fundamental que la solución pueda abordar de manera adecuada el problema.**

En la figura 3 se observa un breve diagrama del proceso.

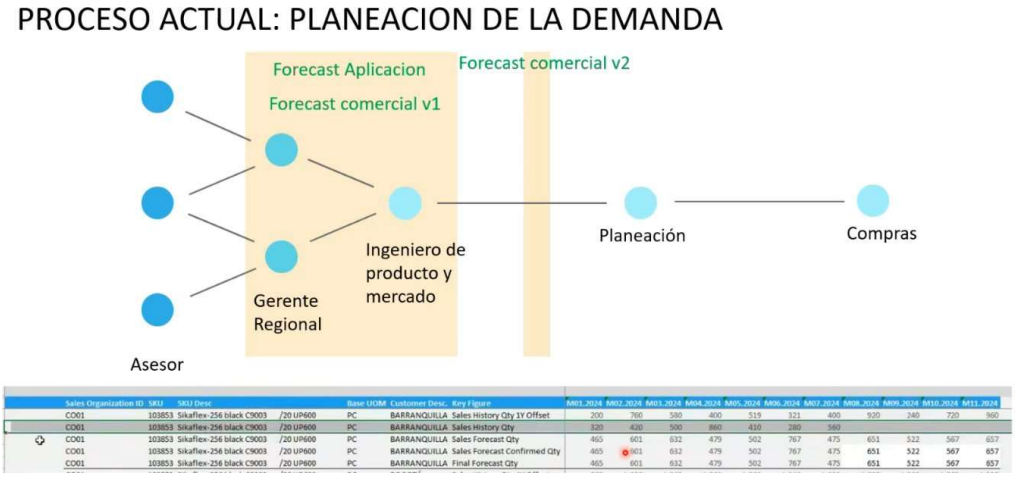
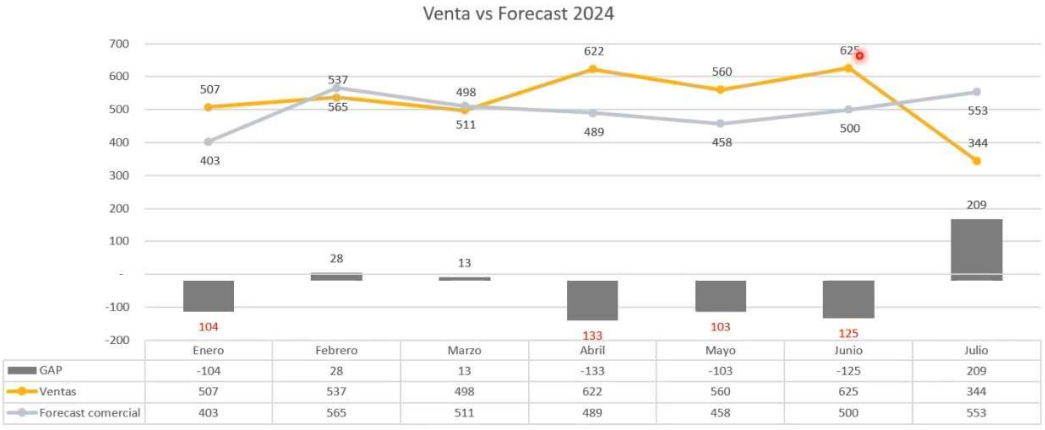


Figura 3. Proceso actual de demanda.

En la figura 4 se puede apreciar la serie de tiempo de ventas y el gap generado entre estas y modelo de planificación de demanda actual.



	Enero	Febrero	Marzo	Abril	Mayo	Junio	Julio
GAP	-104	28	13	-133	-103	-125	209
Ventas	507	537	498	622	560	625	344
Forecast comercial	403	565	511	489	458	500	553

TM	Enero	Febrero	Marzo	Abril	Mayo	Junio	Total
Building Finishing	-10.0%	5.4%	11.0%	-16.2%	-8.6%	-6.6%	-4.6%
Flooring	-15.6%	32.6%	2.0%	-48.2%	-48.8%	-58.1%	-33.9%
Roofing	-12.0%	20.8%	42.6%	-53.7%	-51.8%	-47.1%	-29.6%
Sealing & Bonding	-35.2%	1.8%	-10.4%	-14.4%	-16.9%	-24.9%	-16.6%
Total	**-20.6%**	**5.2%**	**2.6%**	**-21.4%**	**-18.4%**	**-20.1%**	**-12.7%**

Figura 4. Gap generado debido al error del modelo de planificación.

Ideación

El diseño de un producto de datos requiere un enfoque integral que combine el entendimiento profundo de las necesidades de los usuarios, los procesos actuales y las oportunidades de mejora identificadas en el manejo de la información.

- Procesos actuales:
  - La empresa utiliza un modelo de forecast basado en estadísticas básicas, pero este se ve complementado por los ajustes manuales de los gerentes regionales, lo que puede generar ineficiencias
  - Se utilizan datos históricos de ventas, pero carecen de integración adecuada con variables externas que pueden influir en la demanda (p.ej., inversión extranjera, incertidumbre política)
- Usuarios potenciales:
  - Gerentes regionales: Son los responsables de hacer el primer ajuste del forecast generado por la aplicación con base en su conocimiento del mercado.
  - Ingenieros de producto: Validan y ajustan los pronósticos a nivel técnico.
  - Área de planeación: Encargados de consolidar los pronósticos finales para generar órdenes de compra.
- Problemas del proceso actual:
  - Sobreinventarios o agotados
  - Falta de precisión en los pronósticos debido a la simplicidad del modelo estadístico utilizado y la falta de integración de variables externas.
  - Retrasos en la cadena de suministro por pronósticos incorrectos, especialmente con productos importados.

Requerimientos del producto de datos:

- Funcionales:
  - Carga de Datos: Permitir la carga de archivos históricos de ventas en formato CSV. Validar el formato y estructura de los datos cargados para asegurar consistencia y evitar errores.
  - Filtrado de Información: Ofrecer filtros de selección por producto y región para visualizar información específica según las necesidades del usuario.
  - Generación de Pronósticos: Generar pronósticos de demanda basados en datos históricos y variables externas como clima o factores económicos.
  - Exportación de Resultados: Permitir la exportación de los datos analizados y pronósticos en formatos como CSV o PDF para que puedan ser utilizados en otros sistemas o reportes
- No funcionales:
  - Escalabilidad: La aplicación debe ser capaz de manejar grandes volúmenes de datos, especialmente si se integran datos externos o históricos extensos.
  - Desempeño: La generación de pronósticos y visualización de datos debe ser rápida, con tiempos de respuesta mínimos para garantizar una experiencia de usuario fluida.
  - Interfaz Intuitiva: La interfaz debe ser fácil de usar para personas no técnicas, con una disposición lógica y visualización clara de los datos. Streamlit se usará para crear una experiencia amigable.

Mockup del producto:

Figura 5. Mockup inicial de la solución

La figura 5 ilustra la interfaz inicial del producto de datos diseñado para optimizar el proceso de análisis de ventas históricas y proyección en Sika. La pantalla se divide en dos secciones principales: a la izquierda, un área destinada a la carga de archivos históricos en formato CSV, permitiendo que el usuario suba los datos necesarios para el análisis. A la derecha, se encuentran dos filtros desplegables para seleccionar el producto y la región, lo que facilita la visualización específica de las ventas. Debajo de estos filtros, un gráfico de barras presenta los datos de ventas por mes, destacando visualmente los picos o cambios significativos en el rendimiento, lo que permite a los usuarios identificar patrones y tendencias de manera rápida y efectiva; además, la proyección de ventas de los

meses siguientes. Este diseño inicial busca simplificar la interacción del usuario con el sistema, asegurando que pueda cargar y analizar información de manera intuitiva y segmentada según sus necesidades.

Posibles fuentes tecnológicas:

- Kedro (tentativa): Kedro es una herramienta de desarrollo de pipelines de datos que facilita la organización y escalabilidad de los proyectos de ciencia de datos. Su estructura modular ayuda a gestionar el flujo de datos de manera ordenada, lo cual sería útil para mantener un código limpio y reutilizable en el procesamiento y transformación de datos históricos.
- PySpark: Es la API de Python para Apache Spark, se utilizaría para manejar grandes volúmenes de datos, como el historial de ventas o datos externos. Es ideal para el procesamiento distribuido, lo cual mejora el rendimiento y permite analizar grandes datasets en menor tiempo, optimizando el proceso de predicción de demanda.
- Databricks (tentativa): Databricks podría utilizarse como plataforma de colaboración para manejar notebooks y gestionar pipelines de datos en la nube. Esta herramienta permite integrar y escalar PySpark fácilmente, además de ofrecer un ambiente colaborativo para los miembros del equipo de ciencia de datos.
- Streamlit.io: Se utilizaría para desarrollar la interfaz de usuario de la aplicación de forma rápida y sencilla. Es ideal para construir aplicaciones de datos interactivas en Python, como el dashboard donde los usuarios podrán cargar datos, aplicar filtros y visualizar pronósticos y tendencias de ventas de manera amigable.
- GitHub: Fundamental para la gestión del código y la colaboración en el equipo. Con este sistema de control de versiones, se podrá rastrear cambios en el código, gestionar ramas y colaborar eficientemente en el desarrollo del producto de datos, asegurando la integridad y el orden del proyecto.

Estos componentes combinados permiten una solución robusta, escalable y fácil de usar para el desarrollo, manejo y visualización de datos, facilitando un flujo de trabajo colaborativo y una interacción ágil con el usuario final.

Responsabilidad:

La gestión responsable por parte de Sika se basa en la privacidad de los datos utilizados para el desarrollo del forecast y datos futuros que sean adicionados al mismo, esto se debe a que se trabaja con información sensible de volumen de ventas, clientes, productos, datos geográficos los cuales deben cumplir con las normativas de protección de datos como lo es la ley 1581 de 2012 y el decreto 1074 de 2015. Por esta razón los datos utilizados para el proyecto fueron anonimizados para cumplir con estas políticas y garantizar la información de la empresa.

La transparencia y responsabilidad del uso del forecast es fundamental para que los gerentes regionales como la parte comercial y el departamento de compras tenga un entendimiento pleno de los datos obtenidos y de las predicciones que se puedan recibir, pese a que el forecast desarrollado

para el proyecto tenga una alta fiabilidad es claro tener en cuenta que este es solo una herramienta de apoyo en la toma de decisiones, las decisiones finales deben reposar sobre personal ideo de la compañía.

Uno de los grandes desafíos éticos es garantizar la equidad y evitar sesgos que pudieran heredarse de los datos históricos, en el caso de la compañía Sika un modelo de forecast pude verse afectado por sesgos regionales o de alguno de sus productos los cuales pueden llevar a generar un dato que no coincida con las tendencias del mercado actual, la inclusión de variables externas como factores económicos o políticos permitirá que el modelo se ajuste con una perspectiva mas completa minimizando el riesgo de sesgos

Enfoque analítico:

Para abordar la solución, se propone desarrollar tres modelos diferentes a los que se les evaluará su RMSE y MAE para seleccionar el que mejor predicción genere en cada periodo para cada regional con el fin de tomarlo como la predicción definitiva.

La hipótesis de modelado en este caso será la siguiente: Los valores se encuentran auto correlacionados y por lo tanto podrían ser modelados como series de tiempo.

Los tres modelos que se proponen inicialmente son:

- Modelo de series de tiempo ARIMA (Aproximación clásica confiable)
- Modelo de regresión RandomForest(Regresor robusto y de alto rendimiento)
- Red neuronal recurrente LSTM (Su alta capacidad de capturar patrones complejos lo convierte en el modelo que posiblemente retornará los mejores resultados)

Recolección de datos:

- Fuentes de datos**:**
    - La fuente principal de datos es un conjunto de archivos históricos proporcionados por la empresa en formato Excel. Estos datos provienen del ERP SAP que la empresa utiliza para gestionar sus operaciones de ventas.
    - Los datos abarcan todas las ventas realizadas en distintos puntos de venta distribuidos por todo el país. Esto permite tener una visión completa y geográficamente distribuida del comportamiento de las ventas.
    - Adicionalmente, el ERP SAP captura otras variables de interés como la información sobre productos, clientes, y datos operativos que podrían ser útiles para el análisis.
- Estructura de los datos: Se tienen tres campos principales:

1. SGAN (str): Código de identificación único de cada producto
2. Suma de QTY (float): Ventas en Unidad de medida
3. FirstDayOfMonth (date): Hace referencia al mes en el cual se calculó la facturación.
4. Target_Market_Name (str): Linea de producto Nivel I (Recordemos que se tomarán únicamente los relaciones con productos técnicos como se describe en la figura 2)

5. Sales_Office_Name (str): Oficina de ventas bajo el cual se comercializo los productos y que obedecen a una region o zona geografica.

- Utilidad de los datos: Los datos proporcionados son fundamentales para generar un modelo predictivo de ventas debido a la diversidad de variables y su relevancia en la dinámica de ventas. A continuación, se destacan las principales utilidades de los datos:

    - Identificación de tendencias y estacionalidad: El campo FirstDayOfMonth (fecha de facturación) es clave para analizar la evolución temporal de las ventas, lo que permite identificar patrones de estacionalidad o tendencias a lo largo del tiempo.
    - Segmentación por oficina de ventas y mercado objetivo: Variables como Sales_Office_Name y Target_Market_Name proporcionan una visión detallada de las oficinas y mercados donde se realizan las ventas, lo cual es útil para identificar áreas con mayor o menor desempeño y ajustar estrategias a nivel local o por segmento.
    - Análisis por línea y subcategoría de producto: Campos como Application_Field y Target_Market_Name permiten desglosar las ventas por tipo de producto o aplicación, lo que ayudará a identificar los productos más demandados en diferentes categorías.
    - Análisis de canales de venta: El campo BU_Name_CO ofrece información sobre los canales de comercialización utilizados, lo que facilita evaluar el desempeño por canal y determinar cuál genera más ingresos o necesita ajustes.

Estos datos, combinados en un modelo predictivo, permiten capturar patrones y realizar proyecciones más precisas, maximizando la capacidad de la empresa para planificar sus operaciones de ventas

Entendimiento de los datos

Ventas mensuales totales

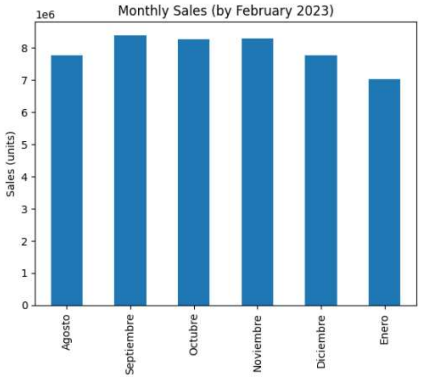


Figura 6. Gráfico de barras ventas de agosto a enero del 2023

En primer lugar, se tiene un gráfico de barras con la información de ventas de los pasados 6 meses al mes de febrero. Dado que los datos están anonimizados el valor del eje Y no es muy relevante, en su lugar es relevante la tendencia que tienen las ventas a lo largo de los meses. Aquí se nota como

en los meses de septiembre, octubre y noviembre se tienen las ventas más altas para posteriormente ir disminuyendo gradualmente.

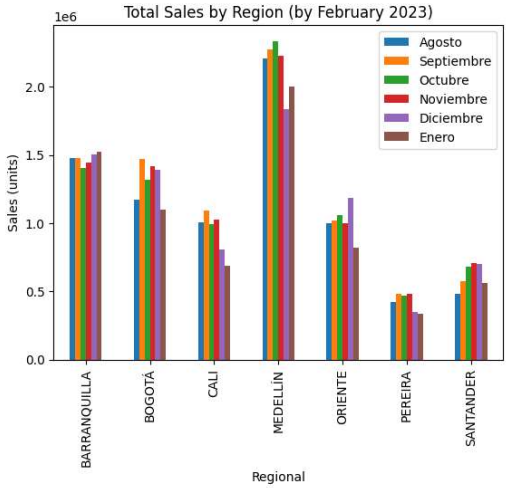


Figura 7. Gráfico de barras ventas Febrero del 2023 por regional

Al observar las ventas por región se puede notar que hay una diferencia notable entre cada región, pues la zona que pertenece a Medellín es la que tiene más ventas mientras que la de Pereira es la que tiene menos ventas. Aquí se puede notar que como es de esperar las regiones más pobladas y desarrolladas del país presentan una cantidad de ventas mayor a comparación de las demás regiones.

Información de los materiales

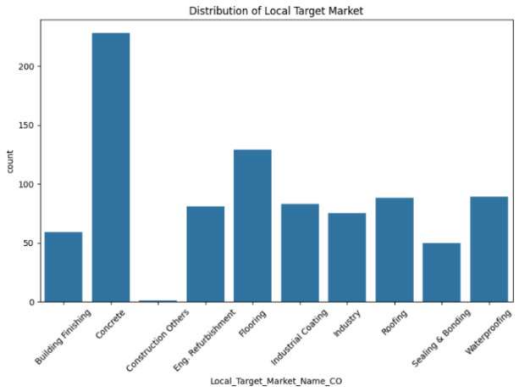


Figura 8. Cantidad de materiales por segmento de mercado

Los materiales que ofrece la empresa se pueden categorizar en distintos segmentos de mercado tal como se presenta en la siguiente grafica. Aquí se nota que la mayor cantidad de materiales entran en la categoría de concreto y la categoría de otros de construcción es la que tiene la menor cantidad de materiales. Las demás categorías tienen aproximadamente la misma cantidad de materiales siendo

así bastante consistentes entre sí, presentando así una gama consistente de materiales a lo largo de todas las categorías.

Graficas por regional de los productos top en ventas.

Realizando el análisis exploratorio de los daros es importante resaltar el impacto que pueden ocasionar los productos de mayores ventas en cada región en el proceso de manufactura al igual que en el proceso de aprovisionamiento de materia prima que requiere ser importada.

Mapa de calor productos

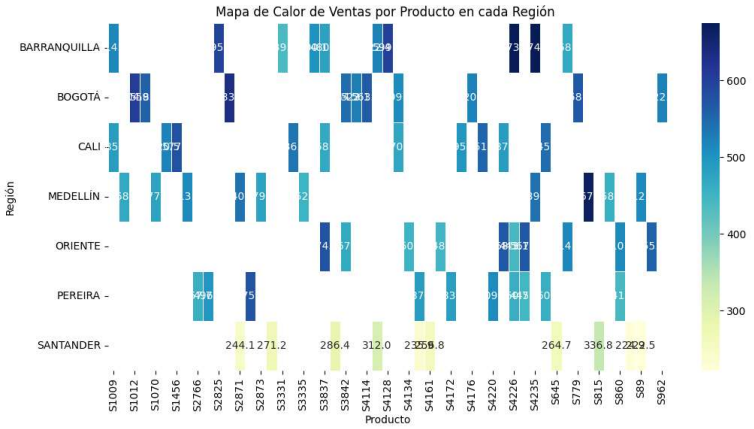


Figura 9. Mapa de calor top 10 productos más vendidos por región

De la gráfica anterior podemos evidenciar que los productos top en ventas presentan una variación dependiendo de la regional a la que se está realizando el análisis, de igual manera se observa que regionales como Medellín y Cali mantiene un promedio de ventas en productos equilibrado a diferencia de Barranquilla y Bogotá que presenta productos con demanda sobresaliente frente a otros, la regional de Santander es quien presenta promedios más bajos en sus productos top.
Lo anterior nos lleva a la **conclusión que la mejor manera de abordar el problema sería mediante un modelo para cada SGAN por cada Regional**, esto permitiría calcular patrones específicos de zona por cada SGAN.

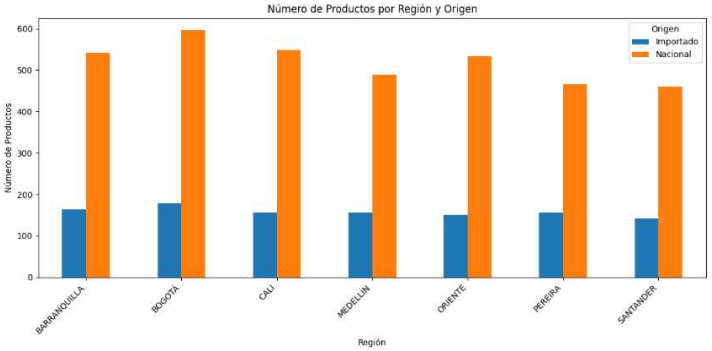


Figura 10. Cantidad de productos importados y nacionales por región

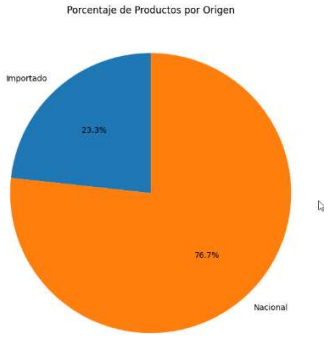


Figura 11. Cantidad de productos importados y nacionales totales

Podemos observar que la mayoría de los productos tiene una base de insumos nacional siendo en cada regional muy superiores las ventas de productos que requieren materia prima importada. Teniendo una relación de 23,3% de productos que requieren materia prima importada y 76.7% productos que requieren materia prima nacional.

```
Chi-cuadrado: 1067440.6369913141
Valor p: 0.0
Grados de libertad: 5396
Frecuencias esperadas:
 [[9.58509603e-02 8.38695903e-02 1.19813700e-02 ... 6.07455461e+00
  7.68005819e+00 3.24695128e+00]
 [1.76627143e-01 1.54548750e-01 2.20783928e-02 ... 1.11937452e+01
  1.41522498e+01 5.98324445e+00]
 [5.68215853e+00 4.97188872e+00 7.10269817e-01 ... 3.60106797e+02
  4.55282952e+02 1.92483120e+02]
 [1.02484657e-02 8.96740750e-03 1.28105821e-03 ... 6.49496514e-01
  8.21158315e-01 3.47166776e-01]
 [2.03511490e+00 1.78072554e+00 2.54389362e-01 ... 1.28975407e+02
  1.63063581e+02 6.89395172e+01]]
Rechazamos la hipótesis nula: hay una asociación significativa entre el Prmedio Mensual de Ventas y el Clima (invierno).
```

Se realiza la prueba estadística de chi-cuadrado para verificar si existe una asociación significativa entre el promedio mensual de ventas y el periodo de invierno (incremento de lluvias), encontrando un comportamiento en las ventas influenciado por los factores climáticos. Información que podrá ser útil para realizar el ajuste del forecast, según proyecciones climáticas.

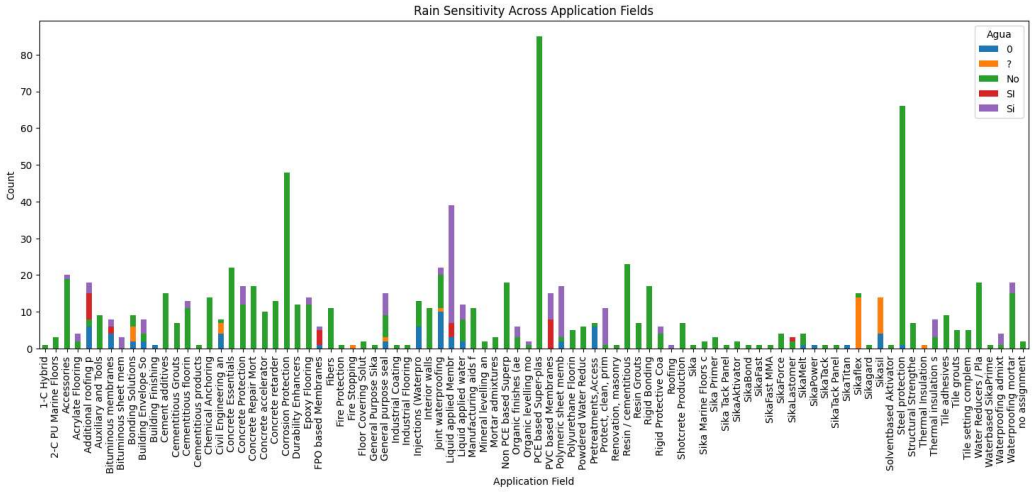


Figura 12. Sensibilidad de los materiales a la lluvia

Una de las variables presentadas por la empresa es la afectación de la demanda del material de acuerdo con la lluvia. Hay materiales cuya demanda aumenta dependiendo de si está lloviendo o no siendo así un factor que la empresa considera importante a la hora de realizar una predicción de la demanda correspondiente. Por ello en la siguiente grafica se presenta la sensibilidad de las categorías específicas de los materiales al hecho de que llueva o no. Aquí se puede notar como las categorías cuya demanda es más sensible a la lluvia son Liquid Applied Membrane y PVC based Membranes siendo ambas categorias membranas que las personas utilizan en techos para arreglar posibles goteras.

Nota: Todos los cálculos, limpieza de datos, procedimientos y gráficas vistas anteriormente se pueden encontrar en el siguiente Notebook (eda_proy_cda.ipynb)

Procesamiento de datos y resultados

Teniendo en cuenta la figura 9, es posible concluir que el mejor enfoque para la solución del problema sería el de tomar un top 5 de códigos SGAN por cada regional.

Una vez sleccionado el top 5, se analizará si las series de tiempo son estaconarias y por lo tanto es posible aplicar un modelamiento mediante la familia Arima, utilizando el método auto_arima de la librería pmdarima de Python. En caso de que la serie no sea estacionaria, es posible diferenciarla para convertirla en una serie hasta convertirla en una serie estacionaria, para rectificar si la serie es estacionaria, se emplará una prueba estaística de Dickey-Fuller. Sin embargo, diferenciar la serie implica pérdida de información y por lo tanto no puede ser un proceso que se repita un número de veces muy elevado. Incluso, puede ocurrir que al diferenciarla se llegue a una serie estacionaria de ruido blanco, lo que tampoco puede ser modelo mediante la familia Arima. Para reconocer series de este tipo, se utilizará una prueba estadística del tipo Ljuang-Box. Si finalmente, la serie de tiempo es ruido blanco o es estacionaria, se utilizarán los regresores de soporte como el Random Forest o la red neuronal LSTM.

Para las series estacionarias que no son ruido blanco, se utilizarán los modelos de la familia ARIMA mencionados. Para analizar su dependencia de la componente autorregresiva y de media móvil, se utilizarán sus funciones de autocorrelación y autocorrelación parcial. También se calculará para estas los dos regresores de soporte, esto con el fin de elegir el modelo que mejores resultados genere. A continuación, se agrega in diagrama de cajas que ilustra el proceso a seguir.

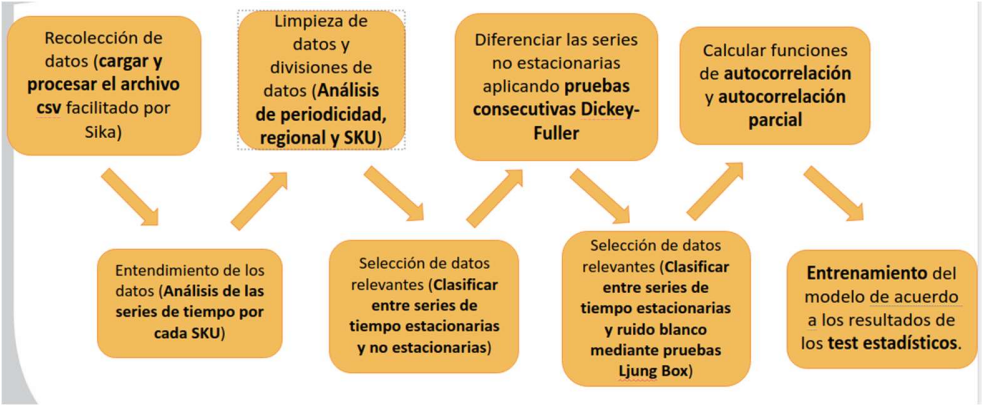


Figura 13. Diagrama de cajas del procedimiento para preprocesar los datos antes de entrenar los modelos

Veamos como ejemplo de los resultados obtenidos por este proceso para el SGAN 3019 para la regional Bogotá:

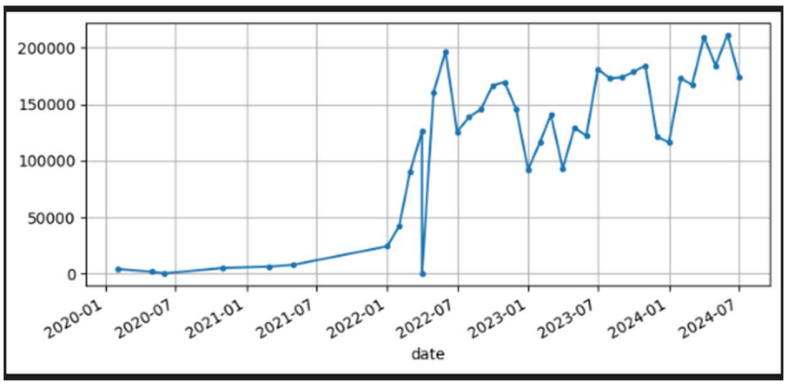


Figura 14. Serie de tiempo SGAN 3019 para la regional Bogotá

La serie de tiempo parece bien comportada, no se aprecian a simple vista estacionalidad aunque sí se aprecia una tendencia creciente. Para observar de mejor manera esta tendencia, calculemos un promedio anual de la serie.

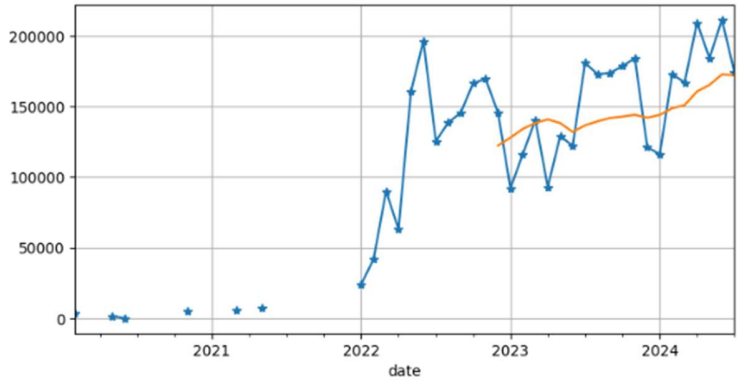


Figura 15 Serie de tiempo SGAN 3019 para la regional Bogotá (curva azul) y serie diferenciada (curva naranja)

Ahora sí se aprecia con total claridad la tendencia de la serie. Esta tendencia indica claramente que la serie es no estacionaria y por lo tanto debe ser diferenciada para convertirla en una serie de este tipo, sin embargo, para cerciorarse, se aplicó la prueba de Dickey-Fuller, y se obtuvo que no es posible rechazar la hipótesis nula (no hay evidencia estadísticamente significativa de que la serie sea estacionaria) lo que confirma lo que ya se había detectado de manera visual.

Veamos ahora su función de autocorrelación y de autocorrelación parcial.

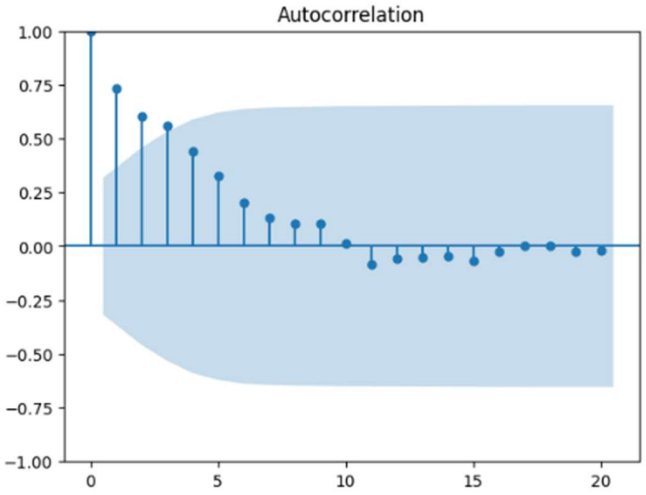


Figura 16. Función de autocorrelación de la serie de tiempo SGAN 3019 para la regional Bogotá.

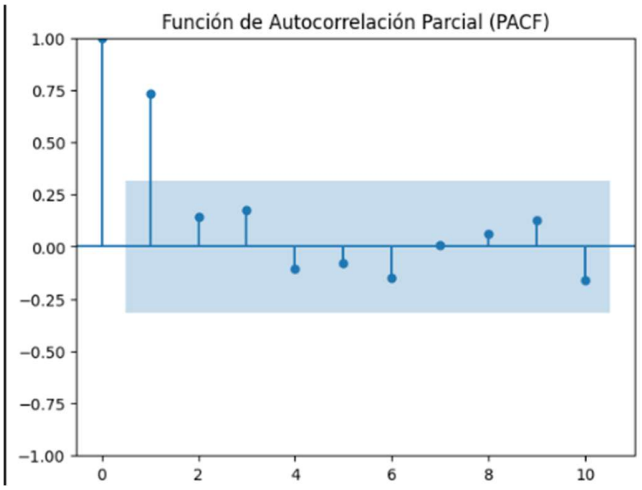


Figura 16. Función de autocorrelación parcial de la serie de tiempo SGAN 3019 para la regional Bogotá.

En la figura 15 se aprecia una dependencia de orden superior a 3 para la media móvil y en la figura 16 de orden 2 para la componente autorregresiva.

Sin embargo, para verificar si podemos modelar la serie mediante la familia de modelos ARIMA, es necesario diferenciarla serie para verificar si la serie diferenciada es estacionaria.

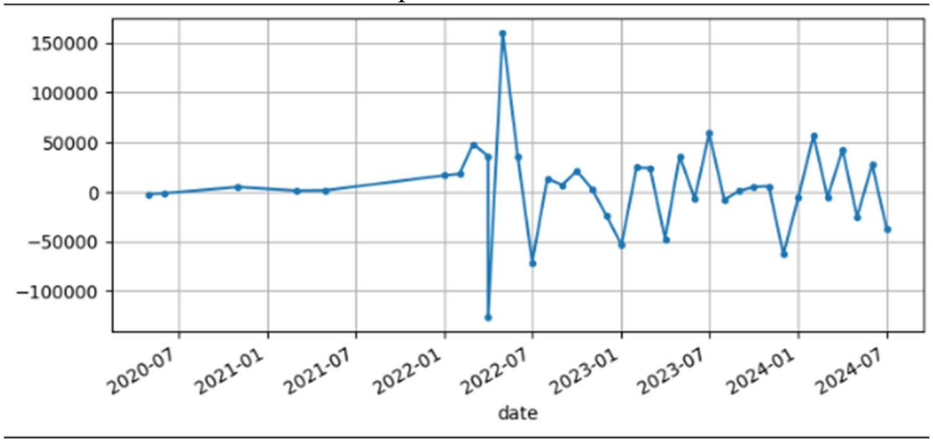


Figura 16. Serie de tiempo SGAN 3019 para la regional Bogotá diferenciada.

En la figura 16 se aprecia la serie diferenciada. A simple vista parece una serie estacionaria. Al aplicar la prueba estadística se confirma que la serie diferenciada es estacionaria y por lo tanto puede ser modelada mediante la familia ARIMA. Además, sus funciones de autocorrelación y autocorrelación parcial muestran dependencias de orden 2 para su componente de media móvil y autorregresiva, lo que permite descartar que la serie sea ruido blanco.

Modelando mediante cada uno de los tres modelos y eligiendo el modelo que minimiza las métricas de error de interés se obtiene la siguiente curva de predicción.

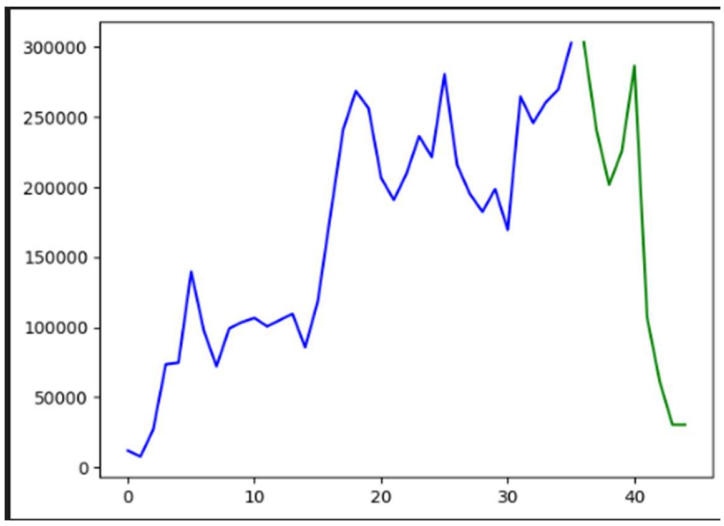


Figura 16. Serie de tiempo SGAN 3019 para la regional Bogotá (curva azul) y predicción (curva verde).

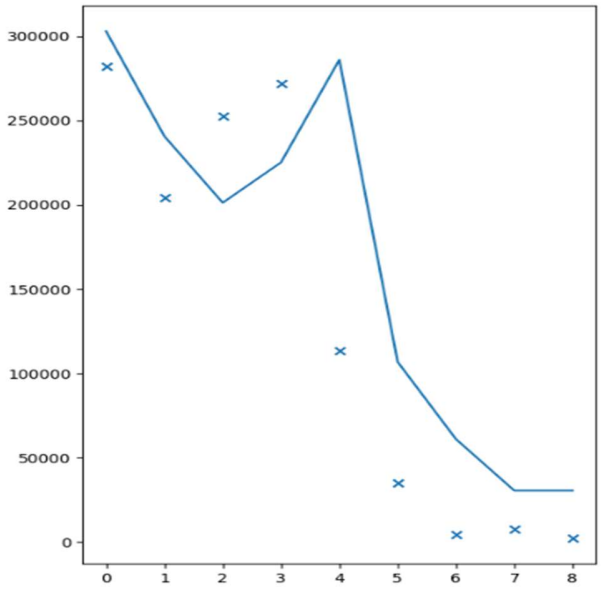


Figura 17. Serie de tiempo SGAN 3019 para la regional Bogotá (marcadores xl) y predicción (curva azul).

Se aprecia que la predicción es bastante buena y logra capturar un patrón descendiente de la serie, son justo este tipo de patrones los que desean capturar, pues el comportamiento inicial de la serie invitaría a pensar que esta mantendrá su tendencia a seguir aumentando, sin embargo, lo que realmente sucede es una disminución de la demanda. Si Sika logra capturar este tipo de patrones con clariadad con antelación, podrá ver impactado de manera positiva su planificación de demanda, lo que finalmente generará un fuerte impacto en sus finanzas. Como ejemplo puede ponerse el caso anterior, lo que pudo terminar en una sobrestimación y un aumento de costos de almacenaje de producto e incluso un posible riesgo de vencimiento, **pudo ser evitado mediante un modelo con una buena capacidad de predicción**, **ahorrando dinero y disminuyendo pérdidas para la empresa.**

A continuación, se ilustran algunos resultados de las métricas de error de los modelos generando predicciones de los 3 SGAN más vendidos para la regional de Bogotá.

	Modelo MINE		Base Line	
S3019	RMSE	748288.19	RMSE	229635,58
	MAE	43019.25	MAE	229250,0
	MAPE	0.072	MAPE	0,5094
S276	RMSE	80304.16	RMSE	41442,69
	MAE	70381.65	MAE	38247,5
	MAPE	0.302	MAPE	0,19
S681	RMSE	41200.55	RMSE	53060,74
	MAE	40333.44	MAE	49375,0
	MAPE	0.049	MAPE	0,054

Tabla 1 Comparativo de métricas para productos en modelo usado actualmente y modelo generado.

El modelo generado durante la actividad es más efectivo y reduce significantemente los errores relativos MAPE pasando de un valor original 0.5094 a un valor del nuevo modelo de 0.072, esto destaca de manera positiva su capacidad de reducción de errores que es esencial en donde la variabilidad de los datos es alta. Este modelo sugiere un enfoque prometedor y potencial para su futura implementación, con iteraciones adicionales y refinamiento el modelo podría convertirse en una herramienta de predicción altamente eficiente.

La solución propuesta será entonces, aplicar este procedimiento descrito anteriormente con el top 5 de SGAN para todas las regionales. Todo el procedimiento puede observarse en el notebook llamado models.ipynb que se encuentra en el repositorio del proyecto.

Arquitectura de la solución

1i

Figura 18. Arquitectura de la solución propuesta

La arquitectura de la solución propuesta presenta el esquema general de la aplicación. Aquí se observa cómo se tiene una fuente de datos que son los archivos CSV que Sika genera cada mes. Dichos archivos se consumen cada mes y se procede a realizar las predicciones de las ventas a partir del modelo implementado. Los datos de las predicciones de ventas producidos por el modelo son posteriormente almacenados por un datalake donde son consumidos por el dashboard de la aplicación para mostrarlos al usuario. En dicho dashboard se podrá observar la gráfica de las predicciones de ventas para las diferentes sucursales y los diferentes SKU para los cuales se realizaron las predicciones. Dicha solución tiene un despliegue tipo Batch Prediction mediante shadow deployment (con A/B test) y Backtesting.

Conclusiones:

- El desarrollo de un *forecast* con alta eficiencia en la proyección generará un mejor aprovechamiento de los recursos de la compañía al igual que la optimización de adquisición de insumos provenientes del extranjero lo cual dará un mejor flujo de caja de la compañía y la optimización de la cadena de abastecimiento.
- A pesar de que un enfoque inicial teniendo en cuenta un modelo únicamente autorregresivo puede generar buenos resultados, es fundamental explorar la inclusión de variables externas que puedan mejorar las métricas de éxito de los modelos.
- Puede pensarse a futuro el desarrollo de un modelo híbrido que minimice las métricas de error (RMSE).
- El desarrollo de tres modelos independientes nos permite evaluar en cada periodo de tiempo el que genere predicciones con la menor métrica de error para ser utilizado.

1

-
- El modelo generó en al gunos casos predicciones sobresalientes a corto (un mes) y mediano plazo (cuatro meses), debido al funcionamiento del negocio. Sin embargo, sería un gran deseable generarlas también a largo plazo (mayor a cuatro meses).
- Fue fundamental tener retroalimentación del personal de Sika para poder desarrollar un producto que sea amigable para los potenciales usuarios de la compañía.
- Utilizar tres modelos de predicción permite elegir entre el modelo con las mejores métricas.
- En algunos casos, no hubo un preprocesamiento adecuado de los datos, lo que pudo haber impactado en los errores de predicción de algunos SGANs.
- Algunos SGANs no podrán ser modelados como series de tiempo debido a no cumplen con la hipótesis inicial (La serie es autorregresiva), por lo tanto, ningún modelo de serie de tiempo va a poder modelarla. Solo los modelos alternativos podrán hacerlo.
- Es fundamental finalizar con la inclusión de variables externas con el fin de mejorar las métricas de éxito de los modelos.

Trabajo a futuro:

A futuro, se recomienda incorporar variables externas como indicadores económicos, condiciones climáticas y tendencias de mercado para mejorar la precisión de los modelos y capturar factores externos que influyen en la demanda. Asimismo, se sugiere desarrollar modelos híbridos que combinen enfoques clásicos y avanzados de machine learning para abordar comportamientos complejos en las series de tiempo. El trabajo de realizar un tratamiento de los valores atípicos presente en los datos de entrenamiento puede ser una mejora significativa a futuro. Además, la implementación de un sistema de retroalimentación continua permitirá ajustar y validar pronósticos con aportes de los usuarios finales. También se puede presentar el trabajo de cuantificar cuanto podría ahorrase Sika en caso de que se prevenga una sobrestimación o subestimación de la demanda.

Se plantea rediseñar el dashboard para incluir funcionalidades avanzadas como alertas predictivas y análisis de sensibilidad, mientras se exploran opciones para migrar la solución a la nube, garantizando escalabilidad y soporte a largo plazo. Finalmente, se recomienda realizar estudios de estacionalidad y validar los modelos en entornos reales, lo que permitirá adaptarse mejor a las dinámicas del negocio y maximizar el impacto financiero positivo de las predicciones. Estas acciones consolidarán la herramienta como un pilar estratégico para Sika en la optimización de su planificación de demanda.

---

i