DATATHON

1. Introducción

1.1 Objetivo

El objetivo de este trabajo es la creación de un modelo de predicción de demanda de vinos y espumantes para una reconocida cadena de supermercados de nuestro país. El fin principal es proporcionarles una herramienta robusta y precisa que permita anticipar la demanda futura de los productos mencionados.

2. Descripción del Problema

2.1 Contexto

La empresa busca obtener información más precisa sobre la demanda de los productos 15 días después. Dicha información es utilizada como insumo para generar las ordenes de pedidos de reabastecimiento de los locales, por lo que ajustar la misma les permitiría optimizar la gestión del inventario, mejorar la planificación de compras, reducir tanto el desabastecimiento como el exceso de stock y aumentar la eficiencia en el uso de los recursos tanto físicos como humanos.

2.2 Desafíos

El modelo de predicción de demanda de vinos y espumantes enfrenta múltiples desafíos, incluyendo la estacionalidad que provoca variaciones significativas durante festividades; el impacto de promociones y descuentos, que distorsionan los patrones de demanda habituales y los cambios en las preferencias del consumidor. Además, la diversidad de productos añade complejidad, ya que diferentes tipos de vinos tienen patrones de demanda distintos. Factores externos como las condiciones climáticas también afectan la demanda. Abordar estos desafíos es esencial para proporcionar pronósticos precisos y útiles que apoyen la toma de decisiones estratégicas de la empresa.

3. Datos

3.1 Fuentes de Datos

Para abordar la problemática planteada, la empresa proporcionó tres sets de datos con información histórica sobre ventas, productos y promociones. Además de esto, se decidió incorporar datos relativos al clima extraídos de INUMET.

3.2 Descripción de los Datos

Las variables comprendidas en las bases proporcionadas por la empresa son las que se detallan a continuación y abarcan el periodo comprendido entre 02.05.2021 y 30.04.2024.

Ventas

Variable	Descripción
LOCCOD	Número del local
MOVFEC	Fecha de venta
PRDCODEXT	Código de producto vendido
ACPRCANTVEND	Cantidad vendida
ACPRIMPVEN	Facturación

Producto

Variable	Descripción
Cod	Código de identifica al producto
Status Hoy	Situación del producto: 1=activo, 2=congelado, 3=congelado compra
Prov	Código que identifica al proveedor
Cat	216=Vinos Finos, 217=Espumantes
Tipo Prov	Tipo de proveedor: 1=Proveedor Externo, 3=Logistico

Productos en Promoción

Variable	Descripción
PRDCODEXT	Código que identifica al producto
PROMDEFFECDES	Fecha desde cuando entró en promoción
PROMDEFFECHAS	Fecha hasta cuando duró la promoción

La información que se decidió incorporar con datos relativos al clima, se resume en el siguiente cuadro:

Variable	Descripción
Fecha	Fecha del día
HumRelativa	Máximo valor de humedad registrado en el día
TempAire	Máximo temperatura registrada en el día

4. Exploración y Preprocesamiento de Datos

4.1 Limpieza de Datos

Las tablas proporcionadas y mencionadas en el punto anterior no contenían valores faltantes, sin embargo, la unión de la tabla Ventas con la tabla Producto arrojó 239 SKU vendidos sin los datos correspondientes al producto. Al considerarse una cantidad poco significativa de registros, se decidió eliminar las filas correspondientes.

Adicionalmente se eliminaron las filas con valores negativos o valor 0 tanto para cantidades vendidas como para monto de facturación, en el entendido de que ese tipo de transacciones

corresponden a devoluciones de mercadería en el primer caso y errores en el segundo. Al tratarse de una cantidad pequeña de registros se entiende que su incidencia en el modelo sería poco significativa.

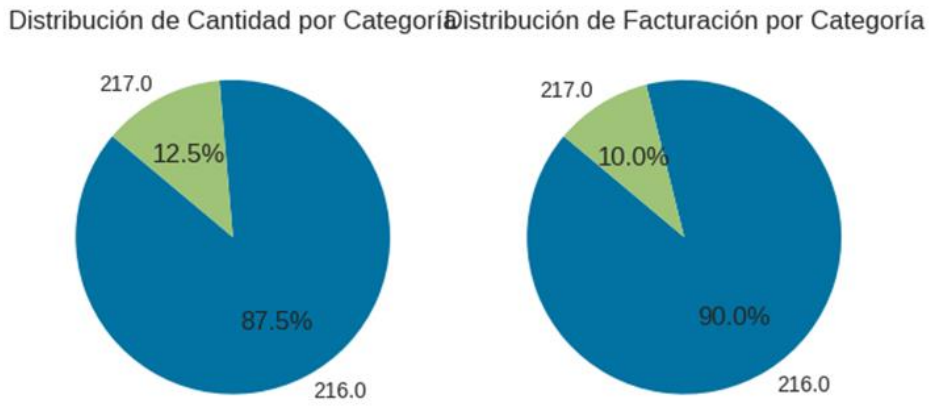
4.2 Análisis Exploratorio de Datos

Esta cadena de supermercados trabaja con 79 proveedores y comercializa 1316 vinos distintos y 163 espumantes.

Las estadísticas descriptivas de los datos se resumen en el siguiente cuadro:

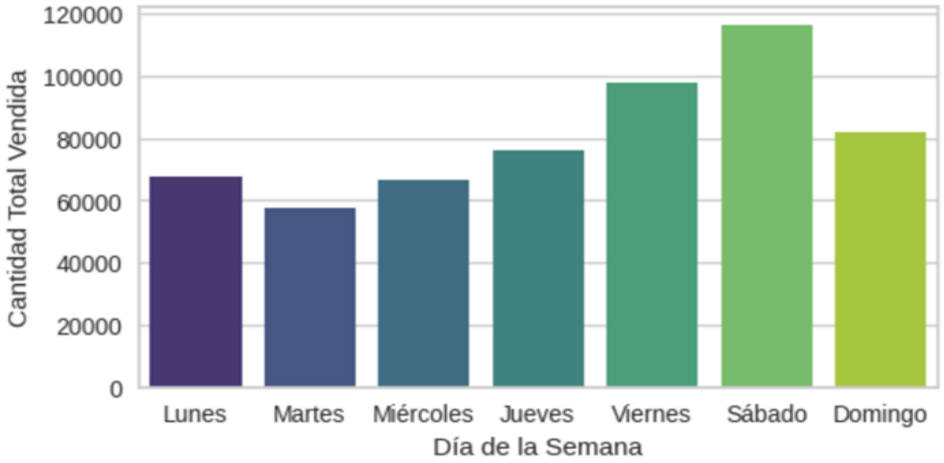
	LOCCOD	MOVFEC	SKU	CANTIDAD	Status Hoy	Prov	Cat	Tipo Prov
count	145276.0	145276	145276.000000	145276.000000	145037.000000	145037.000000	145037.000000	145037.0
mean	5201.0	2022-10-07 05:48:55.686555136	571156.444809	3.884702	1.693202	28273.125575	216.101795	1.0
min	5201.0	2021-05-02 00:00:00	306252.000000	-152.000000	1.000000	10067.000000	216.000000	1.0
25%	5201.0	2021-12-24 00:00:00	561244.000000	1.000000	1.000000	10214.000000	216.000000	1.0
50%	5201.0	2022-09-24 00:00:00	562230.000000	1.000000	1.000000	22010.000000	216.000000	1.0
75%	5201.0	2023-06-18 00:00:00	562799.000000	3.000000	1.000000	31066.000000	216.000000	1.0
max	5201.0	2024-04-30 00:00:00	660998.000000	678.000000	5.000000	85300.000000	217.000000	1.0
std	0.0	NaN	18674.468358	11.035535	1.514033	21871.308226	0.302379	0.0

Del análisis exploratorio surge que el dataset contiene mayor proporción de ventas de vinos que de espumantes:

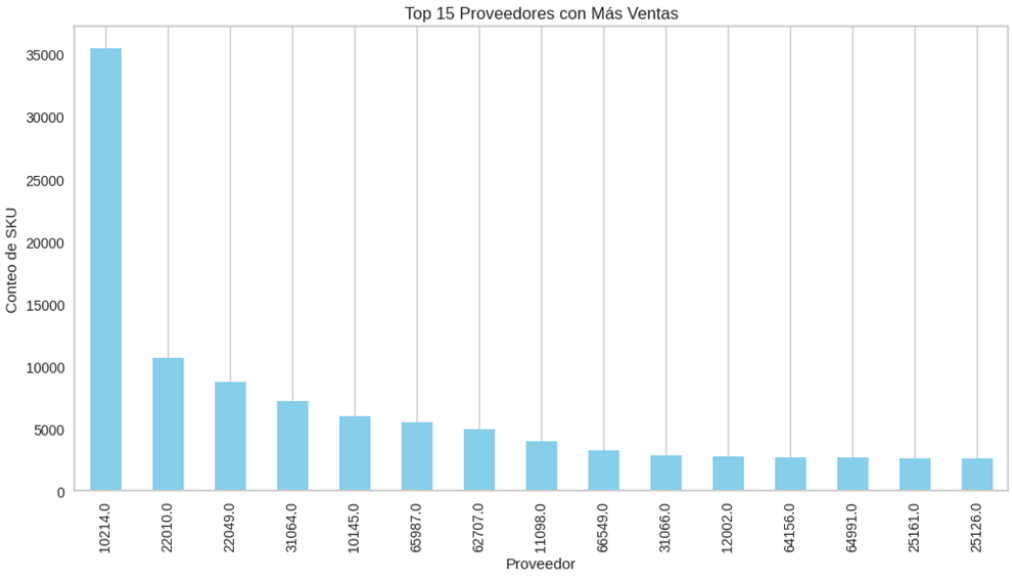


Respecto a su distribución en el tiempo se visualiza claramente la estacionalidad de las ventas durante los periodos de promoción y los días próximos a las festividades de fin de año.

La distribución por día de la semana muestra claramente una tendencia al alza a partir de los días miércoles, alcanzando su máximo los sábados. Luego los domingos se observa una gran caída, que continúa, aunque en menor medida, hasta el martes donde alcanza el mínimo.



Al analizar los proveedores, se observa que el 24% de los productos vendidos pertenecen al mismo proveedor:



5. Metodología

5.1 Selección de Variables

La selección de variables se realizó teniendo en cuenta su relevancia para el modelo. Los datos contenidos en las tablas proporcionadas por la empresa fueron incluidos en su totalidad con excepción de la variable que indica el número del local del cual se obtuvieron los datos, y las variables que indican desde y hasta cuando el producto se encontraba en promoción que fue transformada en una dummy indicando, para cada día de venta y cada SKU, 1 si el producto estaba en promoción y 0 si no lo estaba.

Adicionalmente y con el objetivo de proporcionar mayor información al modelo se crearon las variables categoría_precio y demanda. La primera divide la totalidad de los precios en tres categorías: bajo (1), medio (2) y alto (3), la segunda divide los días de la semana entre días de demanda baja (0) o alta (1).

Por otro lado, se crearon variables que permitieran al modelo poder captar la estacionalidad de los datos. Para esto se calcularon las ventas relativas a cada uno de los días de la semana inmediata anterior, de los 14 días anteriores y del año anterior. Respecto a la fecha, se crearon las variables temporales día del mes, mes, día de la semana (ordenado), y seno y coseno de cada una de ellas.

Finalmente, y teniendo en cuenta el objetivo perseguido por la empresa, se creó la variable dependiente, la cual refleja la cantidad vendida de cada SKU 15 días después del día de venta.

5.2 Técnicas de Modelado

Si bien la literatura referida a predicción de demanda menciona principalmente modelos como ARIMA, SARIMA y Prophet por su capacidad para trabajar con series temporales, su utilización en el caso planteado sería insuficiente ya que permiten predecir únicamente el valor de la variable dependiente por fecha, lo que implica perder la predicción por producto vendido que es el principal objetivo de la empresa en el marco de intentar aumentar la eficiencia en la realización de órdenes de pedidos.

Ante dicha realidad y teniendo en consideración la gran variedad de vinos y espumantes ofrecida por la cadena de supermecados, asi como también el amplio margen de variación que puede haber entre el precio de dichos productos, se decidió aplicar en primera instancia una técnica de cluster de modo de agrupar datos con similares características respecto a su comportamiento a lo largo del tiempo. Se seleccionó para tal fin timeserires K-means el cual utiliza DTW(Dynamic Time Warping) como métrica de la similitud entre dos series temporales que pueden variar en velocidad.

Luego de esto y habiendo obtenido 4 clusters se procedió con la utilización de algoritmos de predicción, utilizando para ello la biblioteca Pycaret la cual compara y evalúa distintas métricas sobre los siguientes modelos de regresión:

	Model
et	Extra Trees Regressor
lightgbm	Light Gradient Boosting Machine
rf	Random Forest Regressor
xgboost	Extreme Gradient Boosting
gbr	Gradient Boosting Regressor
knn	K Neighbors Regressor
br	Bayesian Ridge
lr	Linear Regression
ridge	Ridge Regression
lar	Least Angle Regression
en	Elastic Net
omp	Orthogonal Matching Pursuit
lasso	Lasso Regression
llar	Lasso Least Angle Regression
dummy	Dummy Regressor
huber	Huber Regressor
dt	Decision Tree Regressor
par	Passive Aggressive Regressor
ada	AdaBoost Regressor

5.3 Evaluación de los Modelos

La evaluación del modelo se realizó teniendo en cuenta principalmente las métricas MAE (Error Absoluto Medio) y RMSE (Raíz del Error Cuadrático Medio), ya que ambas permiten evaluar la precisión de las predicciones tanto en train como en test y cuya interpretación se realiza en las mismas unidades que la variable objetivo, permitiendo mayor comprensión de los resultados arrojados por el modelo, lo que entendemos de suma relevancia a la hora de transmitir los resultados obtenidos.

6. Resultados

6.1 Rendimiento de los Modelos

Tal como se mencionaba anteriormente se aplicó un modelo de predicción para cada uno de los Clusters detectados, siendo los algoritmos Light Gradient Boosting Machine y Extra Trees Regressor los que arrojan mejores resultados, tal como se puede visualizar en el siguiente cuadro:

	Cluster 0	Cluster 1	Cluster 2	Cluster 3
Modelo seleccionado	Light Gradient Boosting Machine	Extra Trees Regressor	Light Gradient Boosting Machine	Extra Trees Regressor
MAE	1.7054	1.6864	1.8538	1.9833
RMSE	5.5185	5.3833	4.6720	11.7004

6.2 Interpretación de Resultados

Al analizar los resultados obtenidos en test se puede observar que los distintos modelos y Clusters alcanzan un error absoluto promedio que se sitúa entre 1,6 y 1,9, es decir, que en promedio, la predicción en términos absolutos se desvía en menos de dos unidades de la demanda real.

7. Conclusiones, Recomendaciones y Próximos pasos

7.1 Conclusiones

La aplicación de técnicas de cluster con anterioridad a la utilización de algoritmos de predicción permitió desarrollar modelos con mayor nivel de precisión, especialmente si consideramos que las ventas de productos individuales pueden ser bastante volátiles.
Dicha performance ayudará a la empresa a optimizar el inventario, reduciendo costos por exceso o falta de stock y mejorando la planificación y eficiencia operativa.

7.2 Recomendaciones

Como posible mejora se plantea la inclusión de más variables en el modelo, agregar datos de otros locales, o bien aplicar algún algoritmo de clusterización al set de datos previamente segmentado en función de variables temporales con el fin de identificar subgrupos dentro de los clusters iniciales.
También se podría mejorar la información meteorológica incluida, especificando por ejemplo la magnitud de las precipitaciones diarias, así como también los días en que hubo alerta meteorológica.
Finalmente, otra posible mejora consistiría en identificar el tipo de error que la empresa desea priorizar. No es lo mismo incurrir en un error por falta de stock que por exceso. Si la empresa especifica cuál de estos errores es más crítico evitar, se podría ajustar el modelo con mayor precisión para reducir el impacto del tipo de error prioritario.

7.3 Próximos pasos

Para avanzar con la implementación del modelo desarrollado en primer lugar sería necesario evaluar la viabilidad de su integración con los sistemas de la empresa, tanto en aspectos técnicos como operativos. En segunda instancia deberían hacerse los ajustes correspondientes para su puesta en producción, y una vez sucedido esto avanzar a la etapa de evaluación del modelo, esto es evaluar su capacidad de predicción una vez en funcionamiento y monitorear su desempeño.

Además de esto sería necesario definir el procedimiento para mantener el modelo funcionando correctamente, así como también la estrategia de actualización del mismo con los nuevos datos que se vayan generando.

Por último, se podría evaluar la implementación de modelos con similar metodología para la predicción de cualquier otro producto comercializado por la empresa.