Bogotá, 3 de mayo del 2024

Buenos días, mi nombre es Jorge Jaramillo (www.linkedin.com/in/jorgeivanjh).

A continuación encontrarán mi respuesta al reto "Predicción de demanda", cómo parte del proceso de selección para unirme del equipo de ciencia de datos de Sumz. Espero de corazón cumplir con sus expectivas

AMBIENTE DE TRABAJO

Inicié creando un ambiente virtual en Conda exclusivo para este proyecto, llevando juicioso seguimiento del verisonamiento de las librerías requeridas en el proceso. Creé también un .git y un repositorio PRIVADO en GitHub, asegurándome de incluir en el .gitignore las direcciones de los datos y el pdf con la información confidencial de la prueba, esto para llevar una facilidad en el seguimiento de los avances en el proceso de la prueba.

Si clonó el repositorio, y desea recrear el ambiente virtual utilizado para desarrollar este proyecto, deberá correr los siguientes comandos sobre la raíz del proyecto, dónde encontrará el documento environment.yml

```
conda env create -f environment.yml
conda activate stock_prediction
```

si requiere añador o actualizar version de librerías, hacerlo sobre el environment.ynl y correr

```
conda env update -n stock_prediction -f environment.yml --prune
```

PREPROCESAMIENTO

La analitica inició importando la información disponible y convirtiéndola a DataFrames mediante Pandas. Después empecé un proceso de verificación y limpieza de los datos. Después de verificar que no hubieran outliers determiné la cantidad de datos vacíos en las tablas, encontrando huecos principalmente ubicados en la información de catalogo_productos.csv, dónde se encontraron:

- 32 valores faltantes en la variable "subcategoria"
- 5 valores faltantes en la variable "premium"
- 3 valores faltantes en la variable "tamañö
- 1 valore faltantes en la variable "nit_proveedor"

Después de descartar el acceso a la fuente directa de los datos y métodos de imputación que hubieran sido apropiados para variables numéricas, se llevó a cabo un análisis para imputar los valores más apropiados. Dicho análisis fue el siguiente:

```
"Para cada categoría y tamaño de los registros faltantes, se hizo una
comparación de la subcategoría, estado premium y tamaño con los
registros no faltantes de categorías similares, llenando de forma
acorde los vacíos en esta tabla"
Por ejemplo:
Categoría "jabones","Antibacterial" contrastada con valores en
"premium" de categorías "shampoos" y "aseo" -> La mayoría con valor
premium == 0.0, por lo tanto en categoría "jabones", valor premium =
0
```

Pasado esto, se revisó el tipo y formato de los datos, haciéndose las siguientes modificaciones:

- categoria cambió de string a categoría sin valor ordinal
- subcategoria cambió de string a categoría sin valor ordinal, el formato de texto fue además cambiado a snake casing para estar en concordancia con el formato en categoria

- tamaño cambió de string a categoría con valor ordinal 'pequeño' < 'mediano' < 'grande'
- premium cambió de float a booleano
- marca_exclusiva cambió de int a booleano
- estacional cambió de int a booleano

Luego se agregó una variable booleana llamada"after_competition", la cual expresa el los registros de fechas después de que llegara la competancia el 2 de julio del 2021. Asimismo se eliminó la columna"nit_proveedor" puesto que contenía el mismo valor para todos los registros y no aportaba ningún valor significativo.

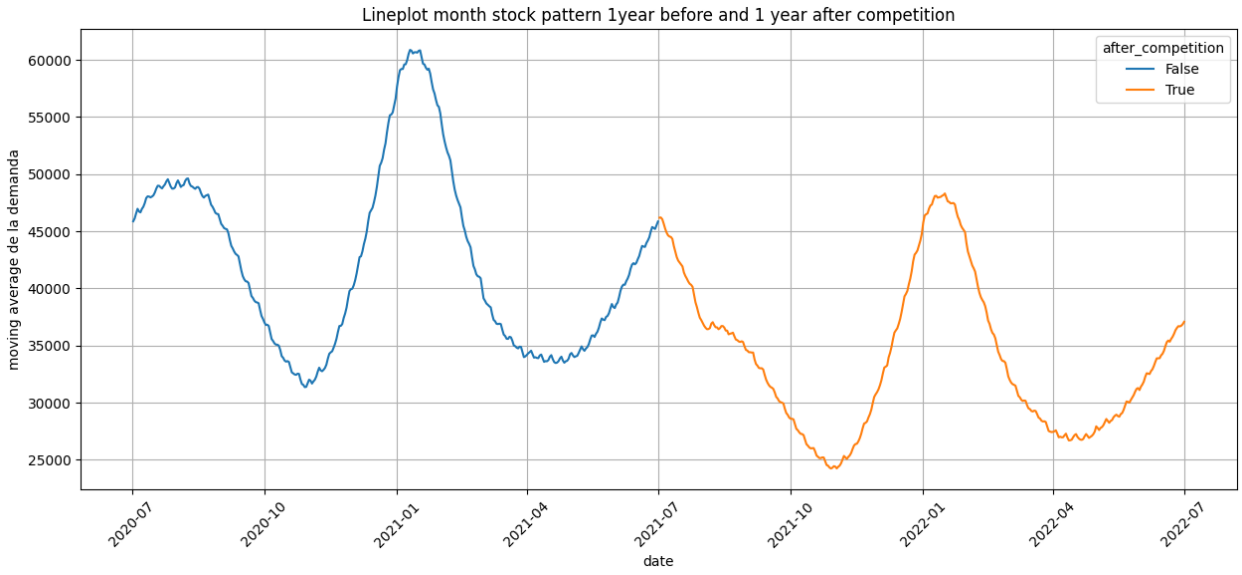
Habiendo varificado los datos hasta acá, se paso a hacer un left join de la información de los productos en"catalogo_productos.csv" con los registros temporales en "demanda.csv" y "demanda_test.csv".

EDA

Se inició el análisis exploratorio de datos haciendo una comparación generalizada de la demanda 1 año antes (azul) y 1 año después (naranja) de la entrada de la competencia el 2 de julio del 2021 (rojo). Para esto se utilizó una gráfica de linea que mostrara la media móvil a 30 días de las demandas diarias para todos los productos

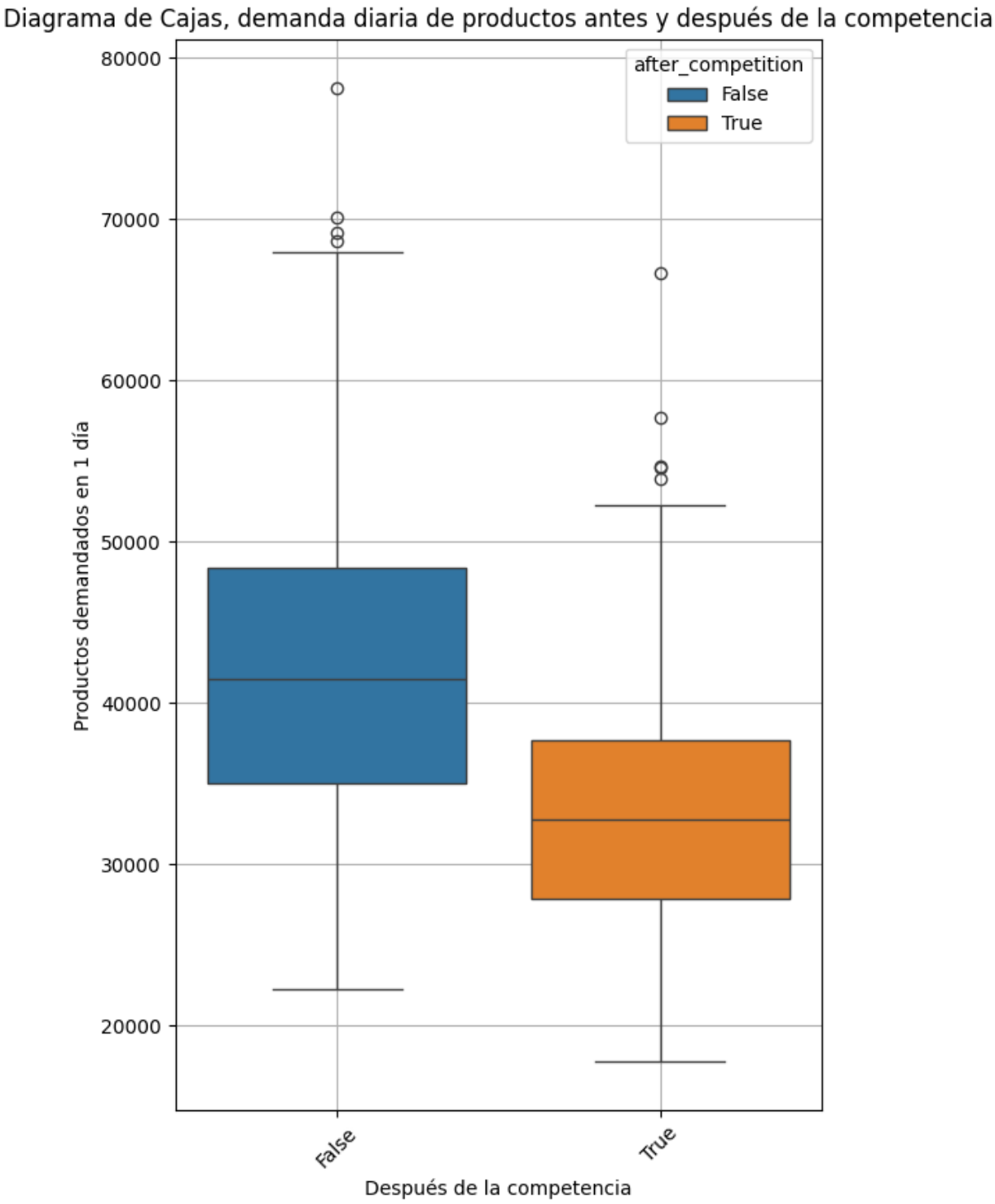
A simple vista se objserva una tendencia a la baja de la demanda de los productos desde que llegó la competencia.

nota: se graficó la media móvil a 30 días para visualizar con mayor facilidad el patrón de la demanda en el tiempo



alt text

Para evidenciar mejor dicha diferencia se calcularon los cuartiles 1,2 y 3, y se relizó un diagrama de cajas, mediente el cual se observa con mayor claridad una importante disminución de las ventas desde que llegó aquel competidor.



alt text

Para comprobar que esta diferencia es estadísticamente significativa, se procedió a realizar un t-test sobre la demanda de productos antes y después del 2 de julio del 2021, tomando cómo:

- Hipótesis nula (H0): no existe diferencia en la demanda entre antes y después de la llegada de la

competencia

- Hipótesis alternativa (Ha): sí existe una diferencia significativa entre la demanda antes y después de la llegada de la competencia

```
TtestResult(statistic=13.918088276062297,
pvalue=3.237331780181677e-39, df=728.0)
```

Se obtuvo como resultado un p-value mucho menor que 0.05, por lo que se rechaza la hipótesis nula, y se concluye que dicha diferencia sí es significativa, y que corresponde con alrededor de una pérdida de 41504.0-32768.0 = 8736 productos diarios en la mediana

Se agregó un segundo diagrama de cajas, que expresa la cantidad de ventas antes y después de la llegada de la competancia.

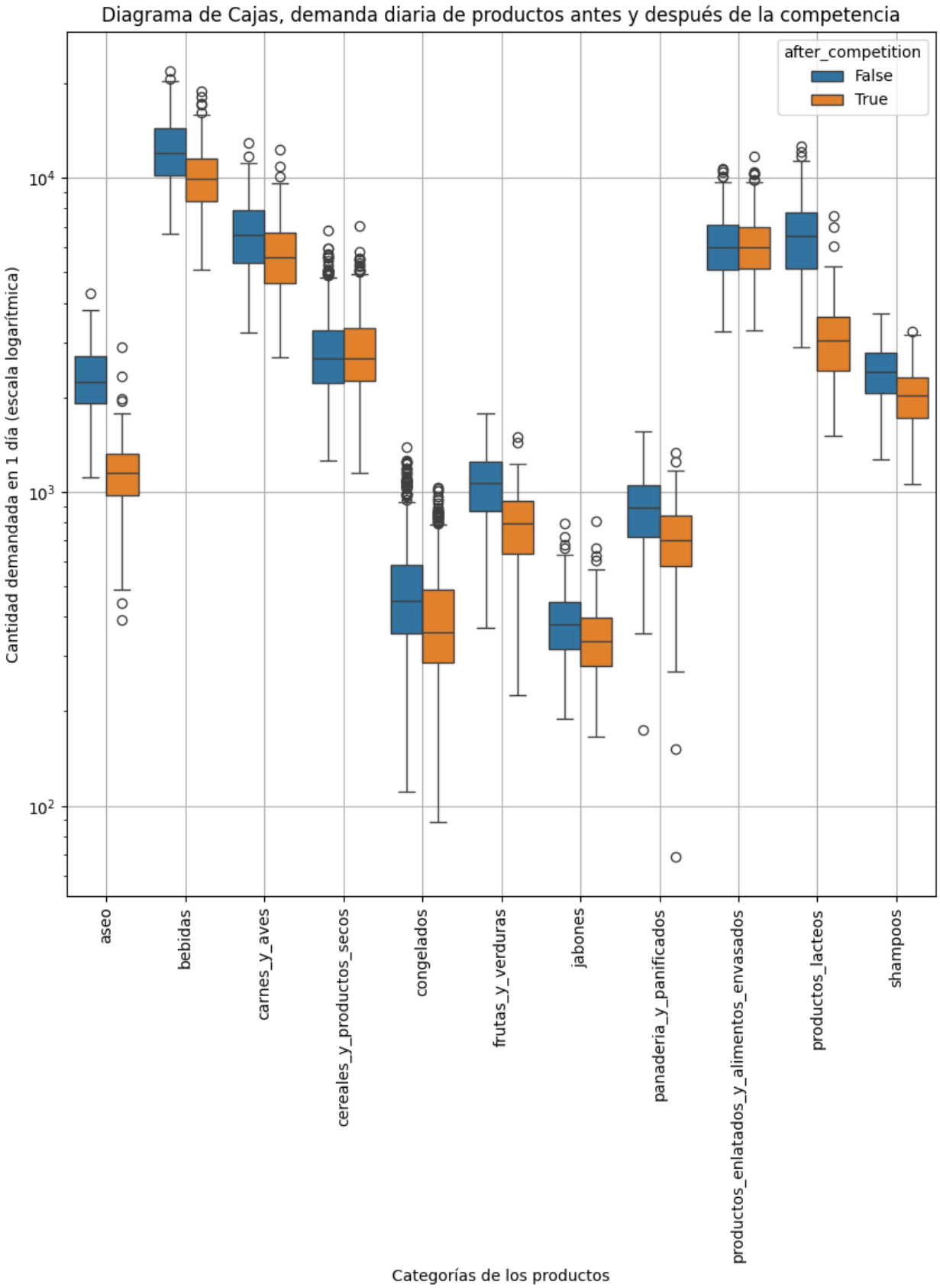


Diagrama de Cajas, demanda diaria de productos antes y después de la competencia
after_competition
False
True
Cantidad demandada en 1 día (escala logarítmica)
10^4
10^3
10^2
aseo
bebidas
carnes_y_aves
cereales_y_productos_secos
congelados
frutas_y_verduras
jabones
panaderia_y_panificados
productos_enlatados_y_alimentos_envasados
productos_lacteos
shampoos
Categorías de los productos

alt text

Junto con el cálculo de la diferencia en la mediana de ventas antes y después de la llegada de la competencia

```
categoria
productos_lacteos                          3472.0
bebidas                                    2042.0
aseo                                       1098.0
carnes_y_aves                              1016.0
shampoos                                    382.0
frutas_y_verduras                           272.0
panaderia_y_panificados                     185.0
congelados                                   93.0
jabones                                      44.0
cereales_y_productos_secos                    4.0
productos_enlatados_y_alimentos_envasados   -36.0
dtype: float64
```

Los productos que mayor cantidad de demanda perdió fueron:

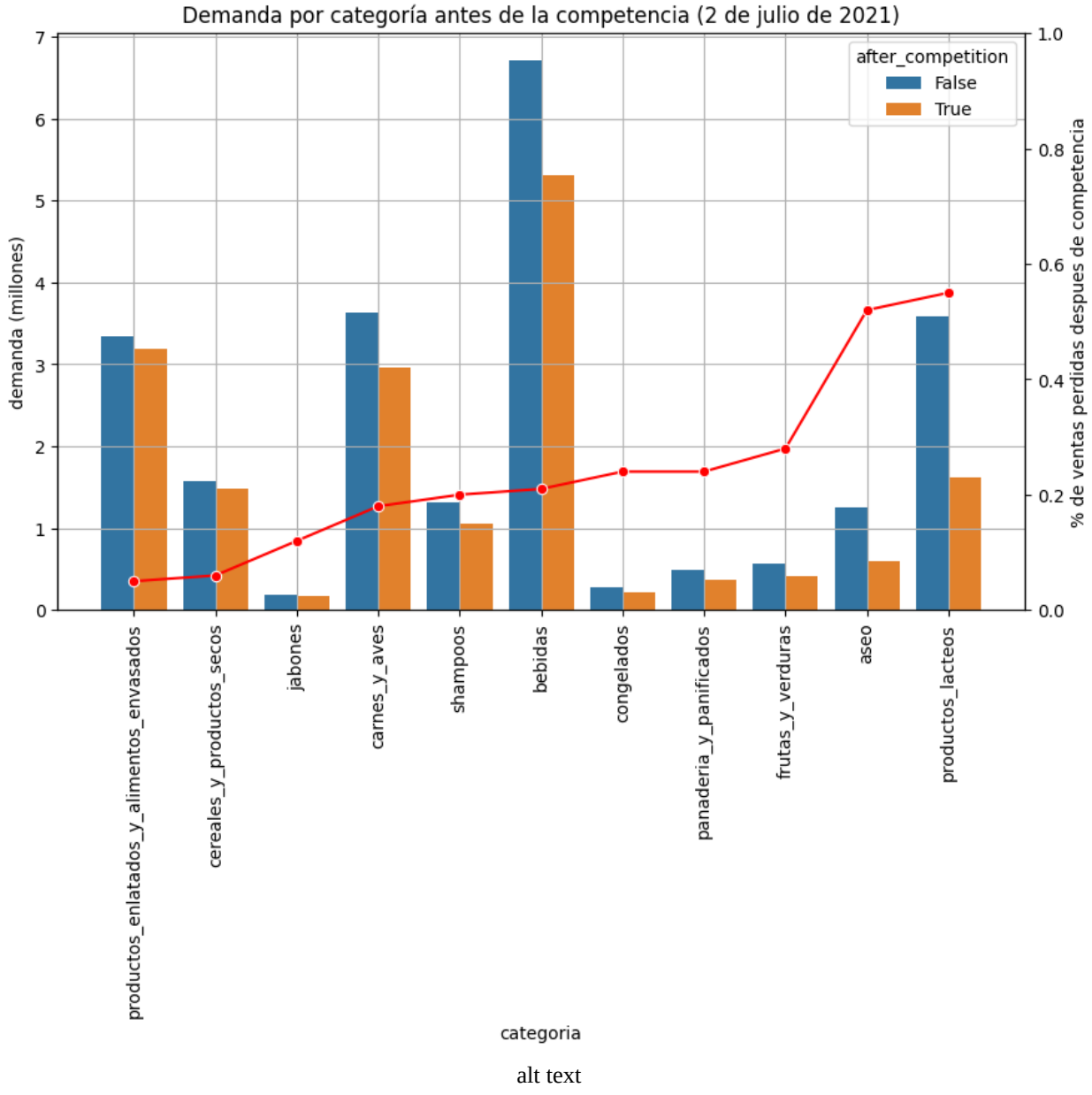
1° productos lacteos, con una diferencia de 3472 unidades

2° bebidas, con una diferencia de 2042 unidades

3° aseo, con una diferencia de 1097 unidades

4° carnes y aves, con una diferencia de 1016 unidades

En un gráfico de barras se organizaron de forma ascendente las categorías según el porcentaje de ventas perdidas después del 2 de julio del 2021. En rojo se superpuso dicho porcentaje.



alt text

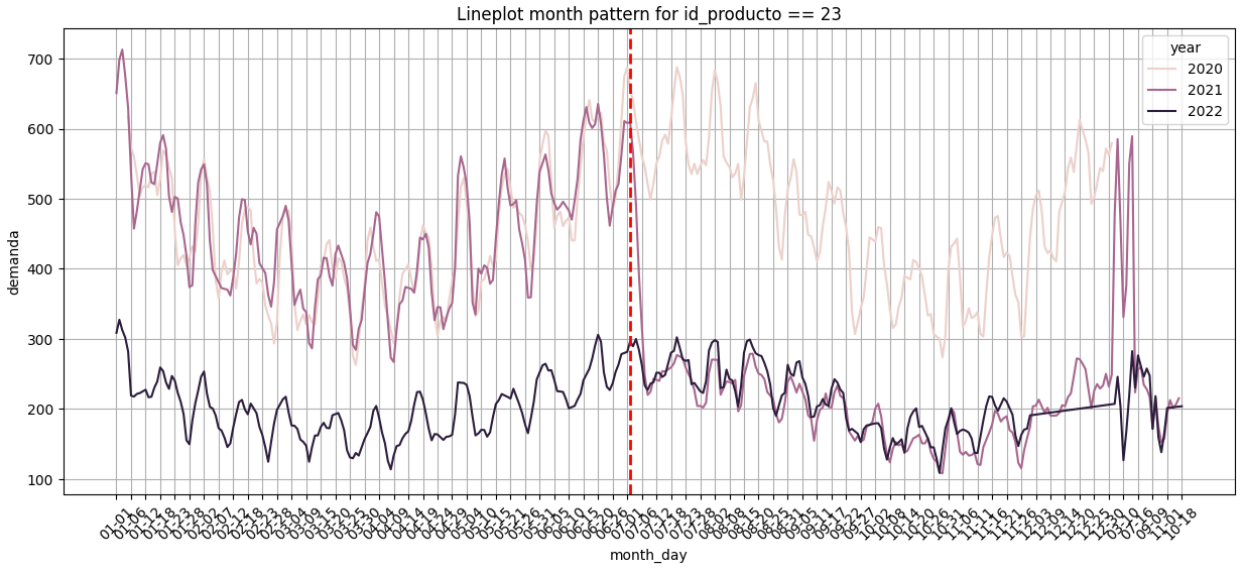
El porcentaje de pérdidas para cada una de las categorías fueron:

categoria	percent_dec_after
productos_enlatados_y_alimentos_envasados	0.04
cereales_y_productos_secos	0.05
jabones	0.11
carnes_y_aves	0.17
shampoos	0.19
bebidas	0.20
congelados	0.23
panaderia_y_panificados	0.23
frutas_y_verduras	0.27

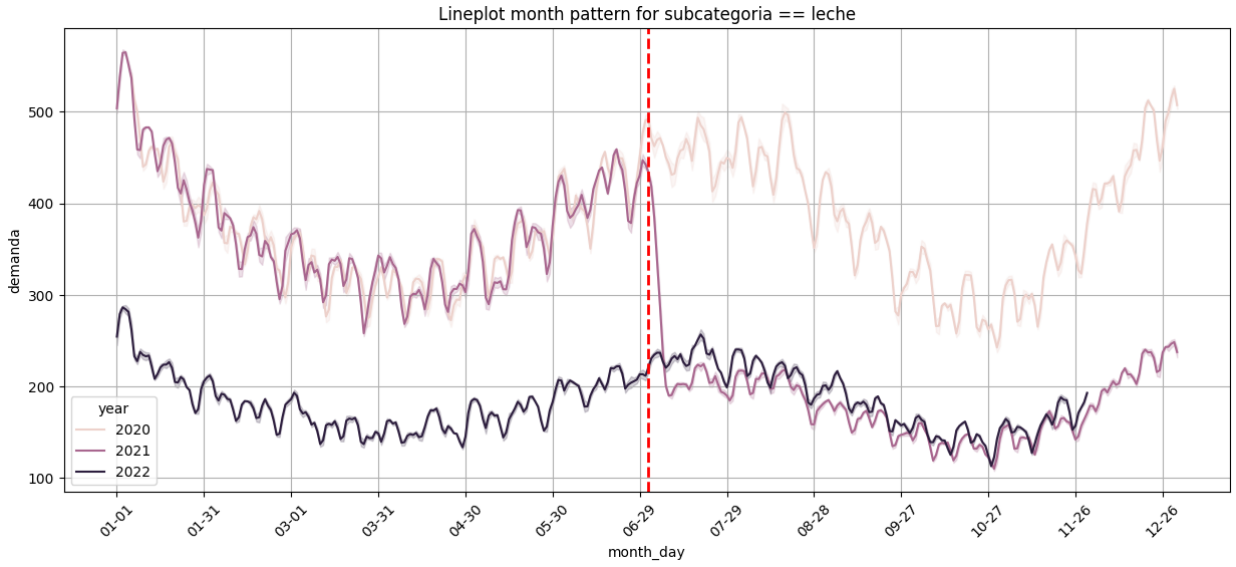
| aseo | 0.51 |
| productos_lacteos | 0.54 |

Con la intención de realizar una exploración detallada de los patrones de ventas, se construyeron las funciones lineplot _movingaverage_pattern junto, barplot_month_pattern y lineplot_weekday_pattern. Las cuales permiten analizar los patrones de demanda de cualquier producto que se desee dentro de los datos.

Usando dichas funciones se crearon gráficos que expresan unos patrones de gran interés respecto del impacto que tuvo la competencia en la demanda por productos lácteos. Tanto para el producto con id==23 (yogurt), cómo para la subcategoría entera dé'leche", se observa el efecto en la reducción de ventas desde la llegada de la competenicia el 2 de julio del 2021.



alt text

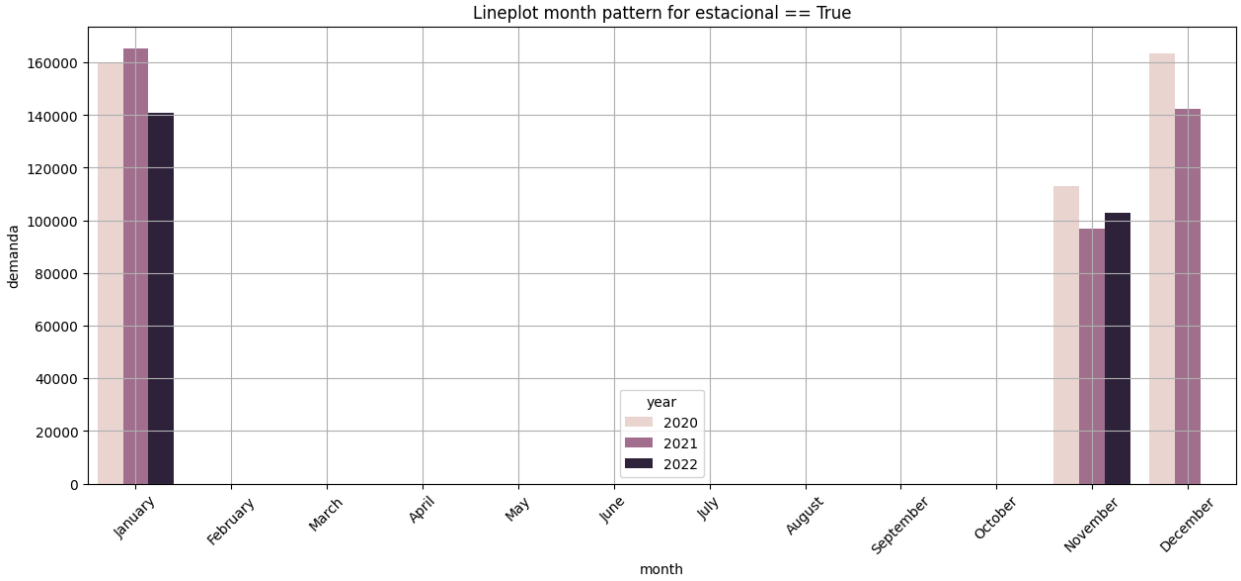


alt text

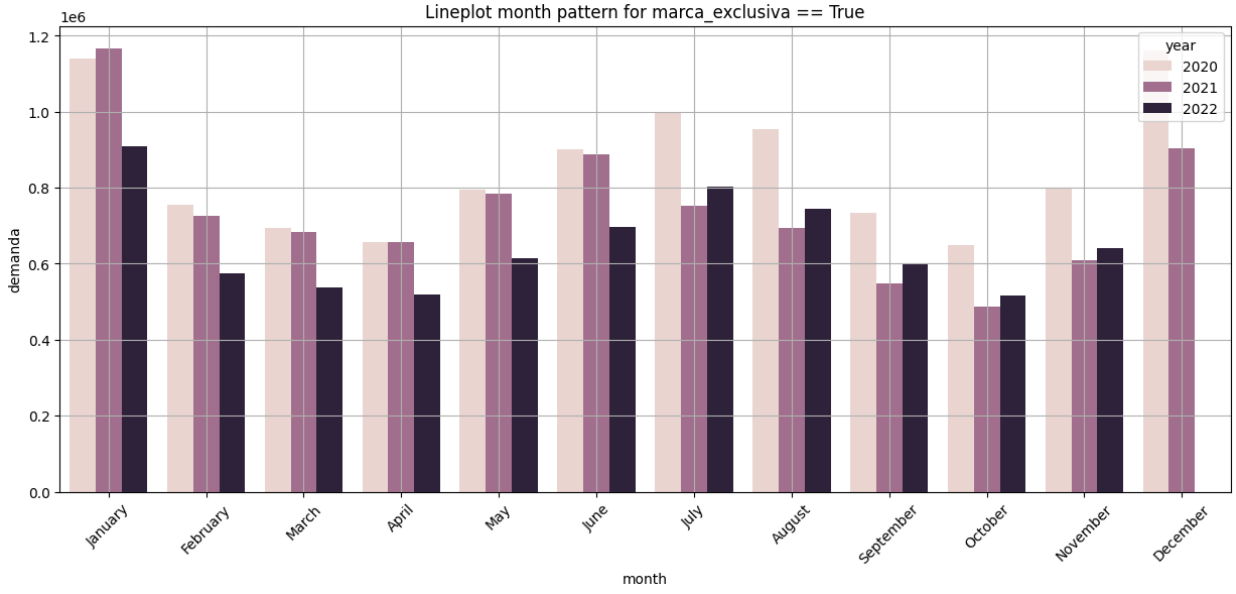
Mediante barplot_month_pattern se puede ver para los meses de los años disponibles en la base de datos, el total de

las ventas conseguidas por mes, para el prodcto, categoría o subcategoría que sean especificados.

En los gráficos de abajo se objervan los patrones de ventas para productos estacionales y de marcas exclusivas. Podemos ver que hay una reducción importante de ventas para las marcas exclusivas, lo que puede significar que la competencia ofrece algún producto competitivo para estos, o incluso que puo haber llegado a tener accesibilidad de dichos productos para su venta.



alt text

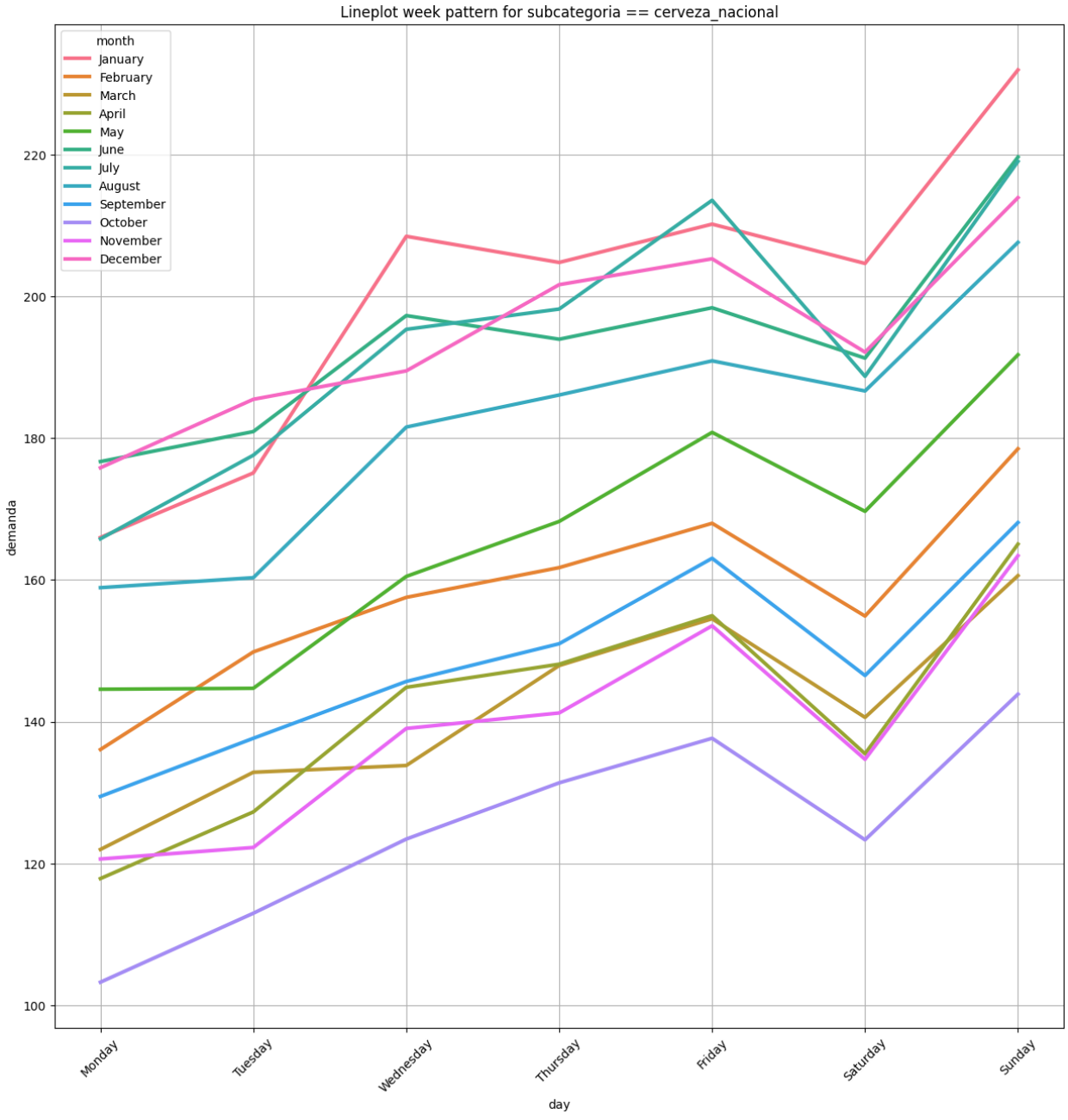


alt text

Mediante lineplot_weekday_pattern se puede observar el patrón de promedio de ventas para los productos indicados a lo largo de los días de la semana para cada mes.

En el gráfico se observa la tendencia en el consumo de cerveza nacional a lo largo de la semana, teniendo su mínimo el lunes, y el pico en el día domingo. Asimismo su consumo es mayor en los primeros y últimos meses del año, y

menor durante los meses de mitad de año.



alt text

Modelo

Se dividieron los datos de entrenamiento (demanda.csv) en entrenamiento (primer 80% de las fechas) y validación (20% restante de fechas). Sobre la sección de entrenamiento se corrieron diferentes modelos de regresión, cuyos resultados de predicción fueron combinados en una sola predicción robusta usando la media.

demanda.csv contiene datos en un rango de tiempo total de 1064 días. Se tomará el primer 80% (851 días) de esos

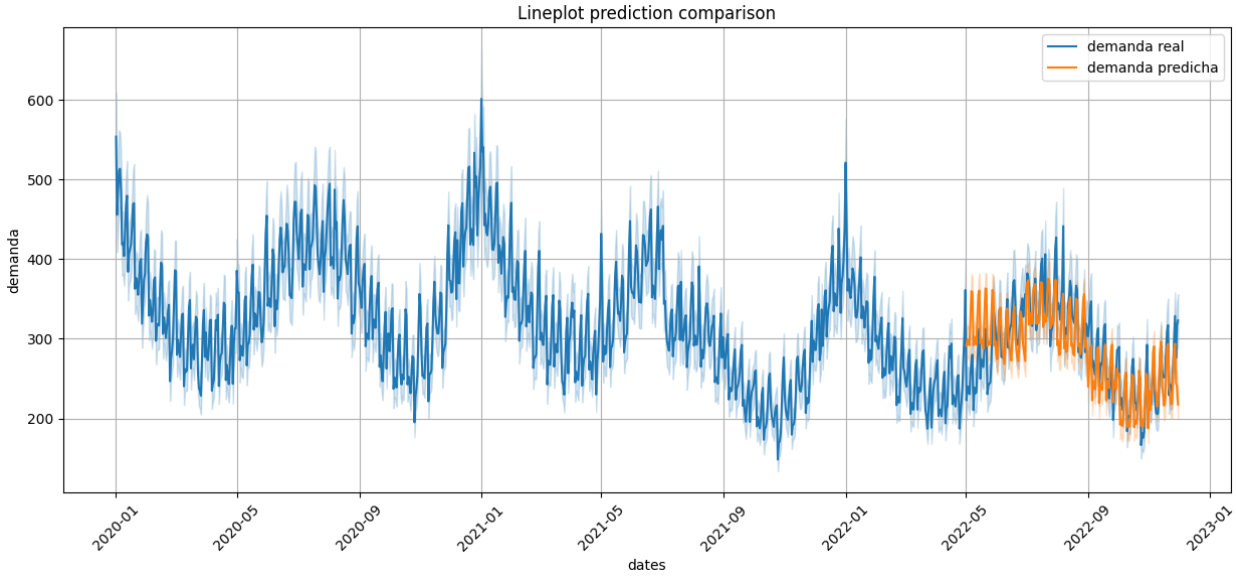
datos para entrenar el modelo, y el 20% (213 días) restante para validación

Regresión Lineal

El modelo de regresión lineal que logró el R-squared más alto llegó a un coeficiente de determinación de 0.506. El cual considera la fecha, todas las variables categóricas disponibles, y las relaciones que pueda haber entre ellas. De este modelo se puede adicionalmente determinar información adicional de la relación de las variables con la demanda.

De esta regresión se puede determinar que, manteniendo todas las variables contantes, mismo día del año, mismo producto y en su misma presentación, por cada año que pase hay una tendencia de una reducción en su demanda de 14 unidades (con un 95% de confiabilidad de que varía entre -19 y -10 unidades) diarias. Este resultado en el resumen de la regresión demuestra además que es un valor significativo (pvalue< 0.05)

```
        coef    std_err  t   P>|t| [0.025      0.975]
year -14.8473 2.493 -5.955 0.000  -19.734 -9.960
```



alt text

Este modelo consiguió un RMSE = 138.66 sobre la sección de datos de prueba

LSTM

Se entrenó un segundo modelo, de redes reuronales recurrentes tipo"Long Short Term Memory" usando tensorflow, al cual se le pasaron los datos de entrenamiento codificadon con one-hot encoding. El modelo Creado cuenta con 4 capas:

- 512 neuronas
- 256 neuronas
- 128 neuronas
- 64 neuronas
- 1 neurona de salida con activación lineal

Se agregaron parámetros de reducción de learning rate a la mitad cada 2 épocas sin mejora en el error, y una finalización automática del entrenamiento después de 5 épocas sin mejoras en la métrica de evaluación.

Este modelo consiguió un RMSE de 157.56

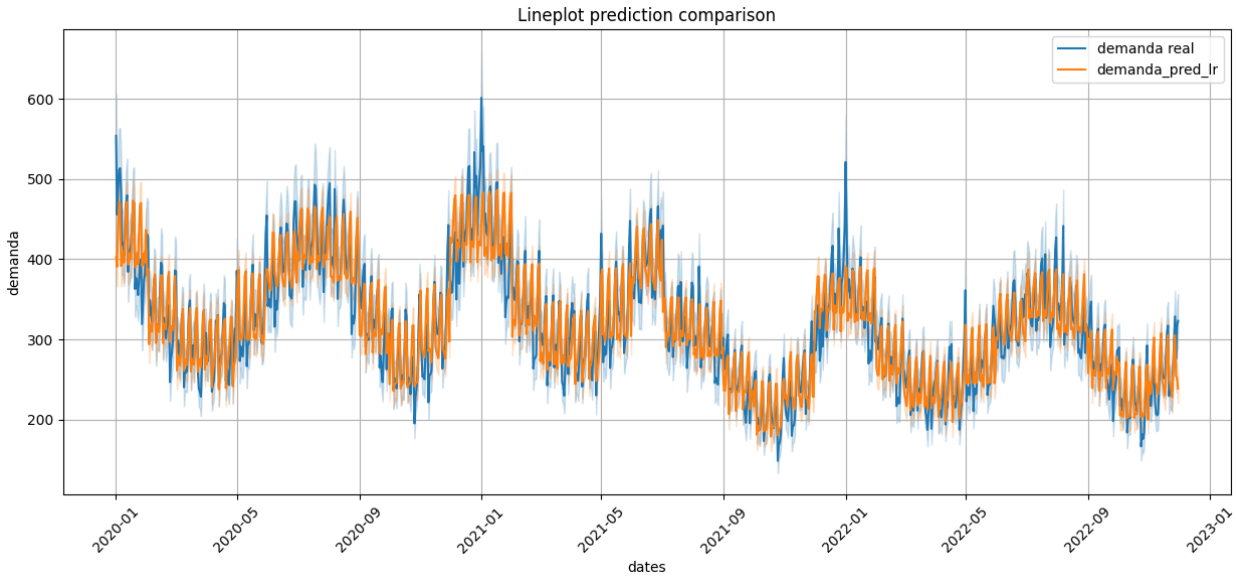
Predicción Final

Después de ajustar los hiperparámetros de ambos modelos usando el test set y el cross validation set, se procedió a entrenarlos nuevamente usando la totalidad de los datos en demanda_test.csv. El resultado de este entrenamiento será usado para hacer las predicciones finales sobre los datos de demanda.csv, de los cuales no se tiene información sobre demanda en sus registros.

La predicción final se realizó considerando las predicciones tanto de la regresión lineal, cómo de la red LSTM, calculando la media entre la demanda predicha por ambos modelos.

Ajuste Linear Regression

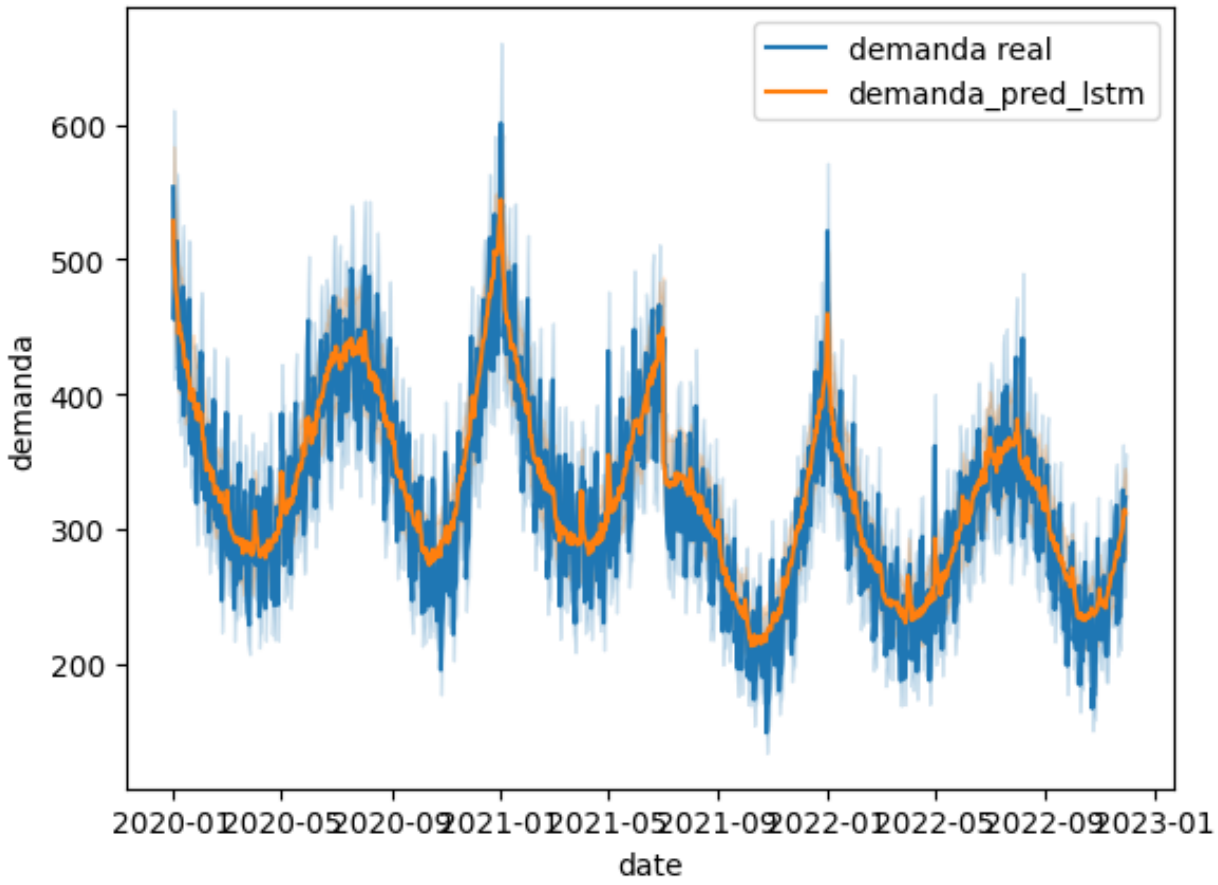
El ajuste de la regresión lineal a los datos de prueba (rmse==138.66) se ve de la siguiente manera



alt text

Ajuste LSTM

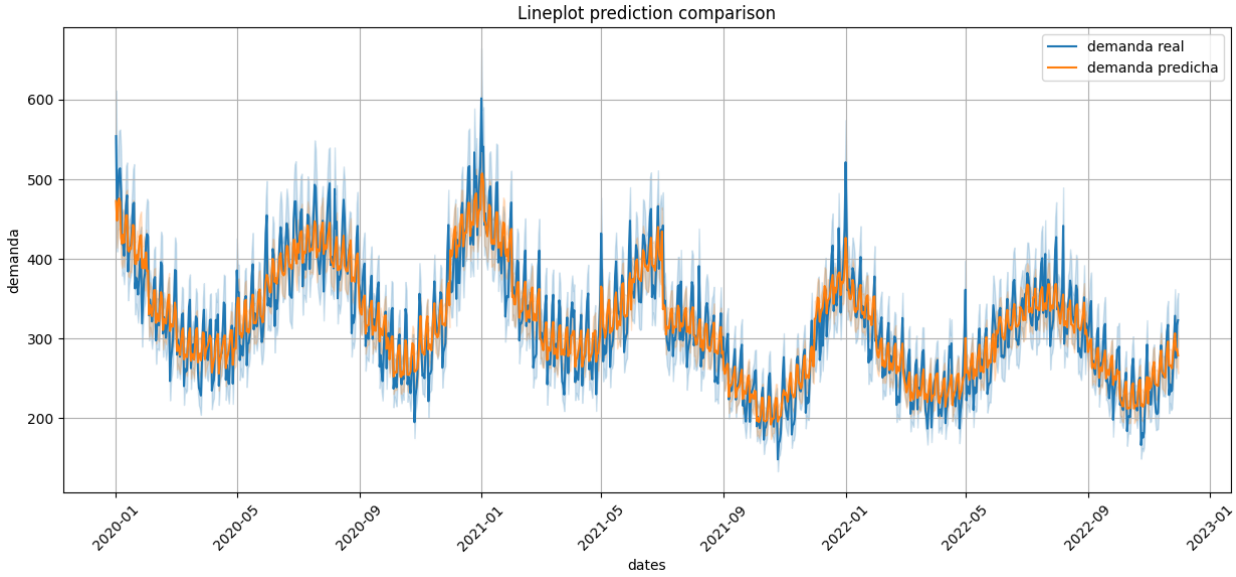
El ajuste de la LSTM a los datos de prueba (rmse==133.07) se ve de la siguiente manera



alt text

Ajuste Model Ensemble

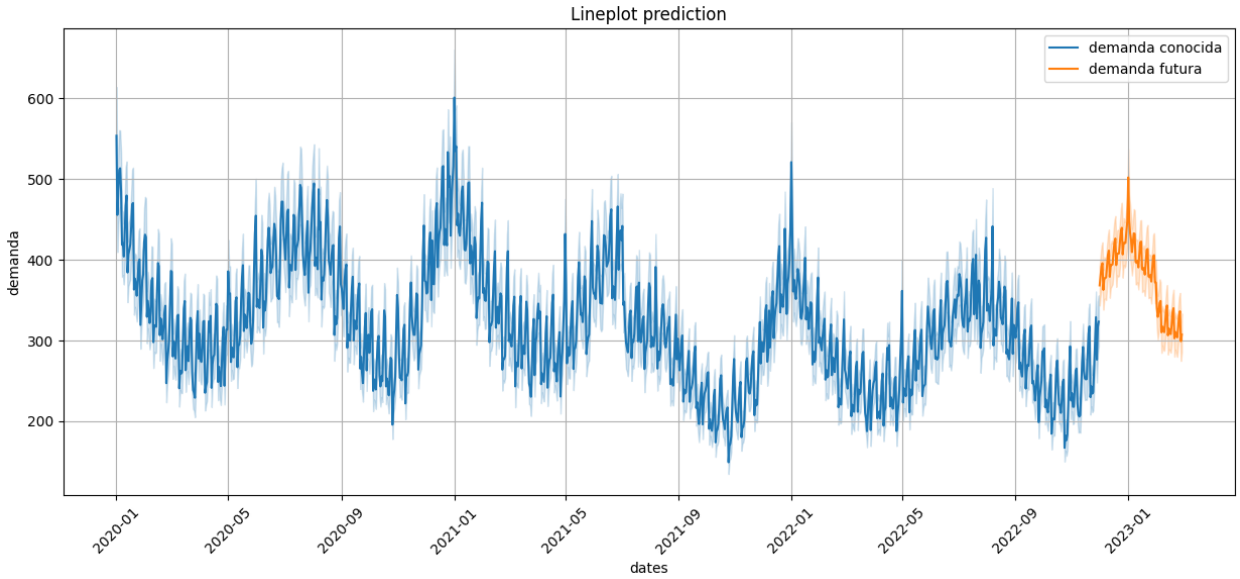
El ajuste de **ambos modelos** a los datos de prueba (rmse==81.14) se ve de la siguiente manera



alt text

Predicción Final

Se usó el ajuste de ambos modelos para realizar las predicciones de los datos en demanda_test.csv, viéndose la continuidad de la semanda en los productos de la siguiente manera:



alt text

En el archivo resultado_prueba.csv se encontrarán los valores de las predicción es realizadas para los datos futuros (date, id_producto y demanda)

Muchas gracias.