UNIVERSIDAD TORCUATO DI TELLA

MASTER IN MANAGEMENT & ANALYTICS

Modelos predictivos del nivel de demanda en la industria láctea: un abordaje para reducir los costos asociados a la gestión de devoluciones y evitar quiebres de góndola

Resumen

El problema en torno a la predicción de la demanda resulta fundamental en la industria láctea debido principalmente a las características propias de sus productos como la fecha corta de vencimiento, la imposibilidad de stockearlos y la alta rotación de los mismos. No contar con niveles de inventarios adecuados en las tiendas se traduce en quiebres de góndola o devoluciones – ambos con sus costos financieros y logísticos aparejados. El objetivo del presente trabajo fue entonces modelar, a través de diferentes técnicas de aprendizaje supervisado, la demanda en tres cadenas de supermercados. Para ello, se utilizaron datos de sell in, sell out y variables que ofrecían información característica de los productos y las tiendas. Los resultados encontrados en las tres cadenas fueron muy similares. El modelo con el que se alcanzó la mejor performance predictiva sobre un conjunto de validación fue el XGBoost en los tres casos. Dentro de los outputs encontrados más destacables de los modelos se encuentra la importante influencia que ejercen las variables históricas en la demanda ya que, tanto los precios como las demandas de las semanas anteriores resultaron ser de las variables más significativas a la hora de predecir la variable objetivo. Este fenómeno contribuye a dar mayor visibilidad y tomar decisiones en torno a los niveles de inventarios que deberían tener las tiendas de cara a la implementación de acciones comerciales, aumentos de precio o lanzamientos de nuevos productos – entre otros - para evitar quiebres de góndola, así como también, excesos de inventario que se traduzcan luego en devoluciones.

Alumna: María Belén Santangelo

Director: Gabriel Martos Venturini

Fecha de entrega: 1 de Junio 2021

UNIVERSIDAD TORCUATO DI TELLA

MASTER IN MANAGEMENT & ANALYTICS

Predictive Demand Models in the Dairy Industry: An Approach to Reduce Costs Associated with Return Management and Prevent Out of Stock in Stores

Abstract

The demand´s prediction problem turns out to be essential in the dairy industry mainly because of certain of its product´s characteristics such as their short expiration date, the inability to store them for many days and the highly rotation around them. Inadequate inventory levels at stores might turn into out of stock as well as returns – both of them involving financial and logistical costs. The purpose of the present work was to model, using different supervised learning techniques, the demand in three supermarket chains. In order to achieve this, information regarding the sell in and sell out was used as well as different variables characterizing the products and the stores. The results that were obtained in the three cases were very similar. The model that was able to achieve the best performance within a validation dataset was XGBoost in every case. One of the main interesting output of the models was the highly importance that historical variables have around the demand, taking into account that not only the past weeks´ prices but also the historical demands turned out to be one of the most significant attributes to predict future demand. This phenomenon contributes to make decisions regarding the inventory levels the stores should keep in order to face discounts policies, price increases or the launching of a new product – among others – to avoid running out of stocks or keeping excess of inventory that eventually might turn into returns.

Student: María Belén Santangelo

Thesis advisor: Gabriel Martos Venturini

Date: June 1st, 2021

INDICE

GRÁFICOS

TABLAS

1. INTRODUCCIÓN

1.1 Marco teórico

La estimación de demanda es una parte fundamental en la gestión de la cadena de suministro; sobre todo teniendo en cuenta que la demanda futura de un producto será la base para definir las políticas de reposición de los mismos. Hoy en día, las cadenas de supermercados cuentan con una gran cantidad de información respecto a las decisiones que toman los consumidores a la hora de realizar una compra: qué productos eligen llevar, qué marcas, a qué precios, qué cantidades, entre otros. De esta manera, es importante poder aprovechar toda esta información y capitalizarla en mejorar los procesos de abastecimiento y garantizar niveles de inventario adecuados, tanto para las cadenas como para los proveedores.

Según Ozhegov y Teterina (2018), el interés por predecir demanda en el mundo del retail nació a fines de la década del 90 cuando Nilson y el IRI Marketing Research comenzaron a relevar información de scanning (datos que se obtienen a partir de todos los productos facturados por línea de caja en los supermercados). Esta herramienta no sólo permitió obtener información acerca de los SKU (stock keeping unit) que llevaban los clientes en cada compra, sino también empezar a tener mayor visibilidad en lo que respecta a otros aspectos de las preferencias del consumidor como el precio, momento del día / horario en que se realiza la compra y cantidades entre otros. El hecho de que las cadenas de supermercados empezaran a tener disponible información más detallada y desagregada sobre lo que pasaba en sus tiendas, aumentó el interés por poder modelar el comportamiento de los consumidores y utilizar estos análisis para mejorar la toma de decisiones (Bajari, Nekipelov, Ryan y Yang, 2015).

El trabajo colaborativo entre las grandes cadenas de supermercados – de ahora en adelante, Grandes Cuentas - y los proveedores es fundamental para mejorar la coordinación en la toma de decisiones y lograr de esta manera una mayor precisión en las predicciones de demanda. Esto no sólo implicará menores costos asociados a la gestión de las devoluciones (productos que no fueron demandados) sino también un manejo más eficiente de los inventarios (Carbonneau, Laframboise y Vahidov, 2007). A su vez, contar con buenos forecasts de demanda permite a los proveedores tener una noción más precisa respecto de los niveles de inventario que deberían tener las cadenas para no quedarse sin stock en las góndolas. Este aspecto es central para evitar una pérdida potencial de ventas para ambas partes y una mala imagen de cara al consumidor final.

Una parte esencial del trabajo colaborativo es la información que se comparte y fluye entre las cadenas y sus proveedores ya que implica numerosos beneficios para ambas partes. Por un lado, compartir información sobre el sell out permite conocer mejor las preferencias de los consumidores, ayuda a entender cómo responde el mercado frente a distintos estímulos y es un input fundamental a la hora de elaborar forecasts más precisos. Por otro lado,

compartir información sobre los inventarios en las tiendas es de vital importancia para evitar los quiebres de stock y los costos asociados a tener que almacenar inventarios en exceso (Lotfi, Mukhtar, Sahran y Zadeh, 2013).

En las últimas décadas, la industria del retail se vio atravesada por la introducción de sistemas de gestión en la cadena de suministro (SCM: Supply Chain Management). Estos avances en materia de sistemas de abastecimiento fueron esenciales para no perder competitividad dentro del ambiente dinámico que caracteriza al sector. Aburto y Weber (2007) definen Supply Chain Management (SCM) como la práctica de coordinar el flujo de bienes, servicios e información a lo largo de toda la cadena de suministro (desde las materias primas hasta que el producto llega al consumidor final).

Por su parte, The Global Supply Chain Forum identificó 8 procesos claves que conforman la base del SCM, entre los cuales se encuentran la gestión de la demanda y la gestión de las devoluciones. La gestión de la demanda se trata de poder predecir la demanda y sincronizarla con las capacidades de abastecimiento, producción y distribución de los fabricantes. Para poder generar estas predicciones es fundamental contar con información histórica de ventas, con los market shares del negocio, con los planes de precio y promoción y con los objetivos estratégicos de la compañía (Croxton, García-Dastugue, Lambert y Rogers, 2001).

La mejora no sólo del nivel de servicio y satisfacción del cliente, como también la de los niveles de competitividad, la reducción de los costos involucrados a la hora de producir y la búsqueda permanente de eficiencia, se encuentran dentro de los objetivos perseguidos por el SCM. A su vez, SCM apunta a mantener en niveles aceptables los niveles de inventario y los costos asociados a su gestión para aumentar las ganancias del negocio (Lotfi, et.al, 2013).

La eficiente gestión en el tratamiento de las devoluciones juega un rol fundamental en el CSM a la hora de aumentar la rentabilidad del negocio si se tienen en cuenta todos los costos asociados a la gestión de productos que son devueltos. No sólo por los costos de transporte que implican sino también por los costos de almacenar esos productos que fueron devueltos hasta que se defina su tratamiento final (destrucción o reciclado). Desde un punto de vista más estratégico, una gestión eficiente de las devoluciones puede impactar de manera positiva en las relaciones con los clientes y mejorar la reputación con el consumidor final; contribuyendo de esta manera a crear valor no sólo para los proveedores sino también para los clientes (Mollenkopf, Russo y Frankel, 2007).

La problemática en torno a la gestión de las devoluciones también fue objeto de estudio de Cui, Rajagopalan y Ward (2019) quienes reconocen los desafíos desde el punto de vista logístico y operativo que la misma implica. El hecho de que las empresas tengan que destinar recursos (humanos y físicos) para el tratamiento de los productos que son devueltos impacta negativamente sobre las finanzas de la compañía. A su vez, en el caso de

que las devoluciones impliquen un reparamiento, podría afectar los ciclos y tiempos de producción.

Dentro de la industria del consumo masivo, y especialmente cuando se trata de productos perecederos y con una corta vida útil como los lácteos, contar con buenos modelos predictivos de demanda resulta fundamental para poder manejar de forma eficiente los inventarios, evitar quiebres en el punto de venta y aumentar la rentabilidad. Según Tarallo, Akabane, Shimabukuro, Mello y Amancio (2019) la corta aptitud que caracteriza a los productos lácteos, así como también, la necesidad de mantenerlos frescos en la etapa de distribución como de almacenado, aumentan la relevancia de contar con buenos modelos predictivos de demanda. Esto contribuiría a minimizar las pérdidas por falta de stock, reducir las devoluciones por tener una fecha cercana a la de vencimiento y mejorar la disponibilidad del producto en góndola.

1.2 Modelos de Machine Learning para predecir demanda

Históricamente los modelos econométricos eran los más utilizados a la hora de predecir demanda. Sin embargo, el acceso a grandes volúmenes de información, que fue posible gracias a los sistemas de scanning, dio lugar a que los modelos de Machine Learning comenzaran a tomar protagonismo. Estos modelos no sólo generaban estimaciones más precisas sobre los conjuntos de validación, sino que también permitieron trabajar con bases de datos más robustas y con un mayor nivel de detalle; pudiendo capturar de esta manera información más desagregada sobre los consumidores y su comportamiento en las góndolas (Ozhegov y Teterina, 2018).

La superioridad de los modelos de Machine Learning frente a los modelos estadísticos a la hora de predecir demanda fue demostrada en varios trabajos de investigación. Puntualmente dentro de la industria del retail, Ozhegov y Teterina (2018) utilizaron modelos de Machine Learning para predecir la demanda en el negocio de las pastas. A fines del presente trabajo, considero interesante los aportes de su investigación teniendo en cuenta las similitudes que presentan las pastas con los alimentos lácteos ya que ambos forman parte de la canasta básica de alimentos y se consumen con relativa frecuencia. Si bien hay diferencias estructurales como el hecho que las pastas pueden almacenarse por más tiempo, mientras que los lácteos tienen fecha corta de vencimiento y no pueden stockearse; ambos productos tienen una gran variedad de SKU y variación de precios entre los mismos. Para entrenar los modelos utilizaron como variables distintas características de las pastas como el precio, el sabor, el peso, el tipo de paquete, el tipo de tienda y el mes en que se realizó la compra, entre otros. Los resultados de su investigación fue que los modelos de Machine Learning que utilizaron tuvieron una notable mejora en la performance respecto a modelos estadísticos más sencillos.

Por su parte, Carbonneau, Laframboise y Vahidov (2007) encontraron que las predicciones de demanda que obtuvieron fueron más precisas al utilizar modelos de Machine Learning que cuando replicaron el análisis usando modelos lineales de regresión.

Ferreira, Lee y Simchi-Levi (2015) utilizaron modelos de Machine Learning para predecir la demanda online de un retailer de ropa (y luego, utilizar esta información como input para optimizar precios y maximizar ganancias). Su objetivo principal era poder anticipar la demanda futura de productos nuevos que todavía no habían sido lanzados al mercado y de los cuales no contaban con información histórica de ventas. Dentro de las variables que utilizaron para entrenar los modelos, incluyeron distintos aspectos de los SKU como el precio, la fecha de inicio y finalización de las promociones, el inventario inicial, el color, el tamaño y la marca. Los autores encontraron que los resultados obtenidos utilizando modelos de bagging eran muy superiores a los de otras técnicas de regresión estadísticas.

La utilización de modelos de Machine Learning para predecir ventas también fue analizado por Pavlyshenko (2019). El autor menciona la importancia de estos algoritmos a la hora de encontrar patrones en series de tiempo y reconoce que la implementación de modelos de aprendizaje supervisado permite identificar patrones más complejos en los datos. Pavlyshenko utiliza información histórica de ventas de "Rossmann Store Sales" para sus predicciones.

Por su parte, Cui, Rajagopalan y Ward (2019) recurrieron a la utilización de modelos de Machine Learning para predecir el nivel de devoluciones en la industria de accesorios para autos que eran vendidos de manera online a través de distribuidores. El principal objetivo era poder anticipar qué productos eran devueltos, en qué momento del año y cuál era la motivación de la devolución; para poder luego focalizar las políticas de descuento (en lugar de dinamizar todos los productos, enfocar el presupuesto en aquellos con mayores niveles de devolución, y en los meses del año más críticos). En particular, los autores trabajaron con una base de datos que incluía información sobre los productos que fueron vendidos y devueltos, la fecha de compra y de devolución, el distribuidor a través del cual se realizó la operación, entre otros.

1.3 Objetivo del trabajo

El objetivo del presente trabajo de investigación será predecir la demanda de los productos lácteos en tres cadenas de supermercados utilizando distintos modelos de Machine Learning; y elegir en última instancia aquel que tenga mejor performance sobre un conjunto de validación. Teniendo en cuenta las particularidades de los productos que serán analizados en cuanto a su rápida pérdida de aptitud, a su corta fecha de vencimiento y a su dificultad para poder ser stockeados; se espera que los resultados de este trabajo contribuyan a darle mayor visibilidad tanto a la empresa productora de lácteos como a las Grandes Cuentas acerca de los niveles de inventario que deberían estar disponibles en las

tiendas de manera de evitar quiebres de góndola pero también reducir al máximo posible los niveles de devoluciones.

A su vez, se espera que los resultados de este trabajo sirvan de input a la hora de tomar decisiones en lo que respecta a la gama de productos que es conveniente mantener en cada formato de tienda en cada cadena (tanto para la discontinuación de SKU´s que no estén performando correctamente como para el alta de productos nuevos). En vistas de evitar pérdidas de ventas e ingresos, tanto para la empresa proveedora como para las Grandes Cuentas, se espera que el output de los modelos contribuya también a poder anticipar a tiempo potenciales quiebres de góndola y asegurarse que las cadenas planifiquen a tiempo los stocks para evitar este escenario (sobre todo en época de descuentos y promociones).

Dentro de los objetivos se encuentra a su vez el de construir modelos que contribuyan a dar mayor visibilidad en lo que respecta a la correcta asignación de productos en las tiendas. Es decir, el foco del trabajo no estará puesto en lograr entrenar modelos de planificación de demanda agregada de largo plazo sino en la asignación a nivel de cada tienda enfocado en el corto plazo. Por este motivo, se espera que las predicciones de los modelos encontrados permitan pronosticar la demanda de la próxima semana.

Por último, se espera que las predicciones de demanda resultantes del presente trabajo generen un impacto considerable en la reducción de los costos logísticos asociados a la gestión de las devoluciones. El hecho de contar con modelos predictivos ayudará a estimar con mayor precisión los niveles de demanda de las semanas siguientes, ajustar los pedidos de las cadenas y evitar de esta manera abastecer a los mercados con producto que luego no será demandado en las tiendas.

Cabe aclarar que, si bien hay diversos motivos por los cuales un producto puede ser devuelto (vencimiento o pérdida de aptitud, pérdida de la cadena de frío, roturas generadas en el trayecto de la planta a las tiendas, entre otros), los modelos destinados a la predicción de demanda aspirarían a reducir el impacto de las devoluciones generadas por excesos de stock en las góndolas que luego no son demandados. Las devoluciones provocadas por razones distintas al exceso de oferta quedan por fuera del alcance de estos modelos ya que, al momento de ingresar en el sistema las devoluciones, no se especifica el motivo de las mismas y, por ende, no se cuenta con este tipo de información para poder sumar a los modelos.

2. MATERIALES Y MÉTODOS

2.1 Datos

Para realizar este trabajo se utilizaron datos provenientes de una empresa productora de lácteos. En vistas de preservar la confidencialidad de los mismos, se enmascararon los datos y se anonimizaron los valores de las variables. El período de análisis que se contempló fue

de enero 2020 a diciembre 2020; y la unidad temporal utilizada fue semanal con el fin de poder capturar potenciales efectos estacionales. La base de datos se construyó con el mayor nivel de granulación posible con el fin de poder realizar las predicciones de demanda para cada SKU, en cada tienda, en cada semana de cada mes.

Cabe aclarar que, si bien el período de análisis considerado es el más reciente temporalmente, el mismo se encuentra atravesado por el impacto económico y social generado por la pandemia. Es por ello, que las conclusiones que se desprenden de este análisis quedan circunscriptas al corto plazo. Una vez que las condiciones económicas y sanitarias se normalicen, sería importante volver a modelar la demanda de los productos lácteos usando datos que representen de forma más fiel los patrones de consumo de los individuos.

A la hora de consolidar y armar la base se utilizaron datos provenientes de diferentes fuentes. En primer lugar, toda la información de sell in (son las ventas que la empresa proveedora de lácteos realizó a las cadenas), devoluciones y nivel de entregas fue provista directamente por la empresa. Por otro lado, la información de sell out (ventas que las cadenas realizan al consumidor final en sus tiendas), de inventarios y de precios se obtuvo a partir del proceso de scanning que tiene lugar en la línea de cajas de los supermercados. Por último, toda la información que contiene características de los SKU como de las tiendas también fue provista por el proveedor de lácteos.

El conjunto de variables que conformaban la base de datos inicial (es decir, previo a la ingeniería de atributos) puede separarse en 4 grandes grupos: atributos del cliente, atributos del SKU, atributos temporales y variables numéricas. El detalle y descripción de las variables iniciales puede observarse en la Tabla 1:

Tabla 1: Descripción de variables iniciales

Variable	Tipo	Descripción	Grupo de variables
MES	Categórica	Mes del año	Atributos temporales
SEMANA	Categórica	Número de semana del mes	
CADENA	Categórica	Código de identificación de cada cadena	Atributos del cliente
FORMATO	Categórica	Tipo de formato de la tienda	
PROVINCIA	Categórica	Provincia donde está ubicada la tienda	
REGION	Categórica	Región donde está ubicada la tienda	
TIENDA	Categórica	Código de identificación de cada tienda	
ENVASE	Categórica	Tipo de envase contenedor del SKU	Atributos del SKU
FAMILIA	Categórica	Familia de productos a la que pertenece el SKU	
GRM	Numérica	Gramaje del SKU	
MARCA	Categórica	Marca del SKU	
PCB	Numérica	Cantidad de unidades por bulto del SKU	
PRECIO	Numérica	Precio vigente esa semana	
PRECIO_T_1	Numérica	Precio vigente la semana anterior	
PRECIO_T_2	Numérica	Precio vigente 2 semanas atrás	
PRECIO_T_3	Numérica	Precio vigente 3 semanas atrás	
PRECIO_T_4	Numérica	Precio vigente 4 semanas atrás	
REEMPLAZO	Categórica	Dummy para capturar efecto de reemplazos de SKU	
SABOR	Categórica	Sabor del SKU	
SKU	Categórica	Código de identificación de cada SKU	
VIDA_UTIL	Numérica	Vida útil del SKU	
CSL	Numérica	Customer Service Level / Nivel de Servicio	Atributos numéricos
DEMANDA	Numérica	Demanda de esa semana (en unidades)	
DEM_T_1	Numérica	Demanda de la semana anterior (en unidades)	
DEM_T_2	Numérica	Demanda de 2 semanas atrás (en unidades)	
DEM_T_3	Numérica	Demanda de 3 semanas atrás (en unidades)	
DEM_T_4	Numérica	Demanda de 4 semanas atrás (en unidades)	
DEV	Numérica	Devolución de la tienda al proveedor (en toneladas)	
DEV_T_1	Numérica	Devolución de la semana anterior (en toneladas)	
DEV_T_2	Numérica	Devolución de 2 semanas atrás (en toneladas)	
DEV_T_3	Numérica	Devolución de 3 semanas atrás (en toneladas)	
DEV_T_4	Numérica	Devolución de 4 semanas atrás (en toneladas)	
%DEV	Numérica	Devolución de esa semana (en %)	
INV_INI	Numérica	Inventario al iniciar la semana en las tiendas (en unidades)	
INVENTARIO_SEMANAL	Numérica	Inventario al finalizar la semana en las tiendas (en unidades)	
SELL_IN	Numérica	Ventas del proveedor a las cadenas (en toneladas)	
UNID_ENT	Numérica	Unidades entregadas en esa tienda	
UNID_PED	Numérica	Unidades pedidas por la cadena en esa tienda	
VENTA_SEMANAL	Numérica	Sell out de la tienda (en unidades)	

Cabe aclarar que algunas de estas variables fueron calculadas a partir de otras:

- CSL →el Customer Service Level se obtuvo a partir del cociente entre unidades entregadas y las unidades pedidas por el cliente.

- Demanda → esta variable es central ya que será la que se buscará predecir con los distintos modelos. La demanda se construyó como el mínimo entre la venta semanal (sell out) y el inventario disponible al inicio de la semana. En aquellos casos en los cuales la venta semanal y el stock inicial era el mismo se asignó un valor vacío a la demanda ya que probablemente se trate de un caso de quiebre de góndola. Es decir, no era posible determinar cuál sería el valor de demanda porque se agotó el stock en las tiendas de ese SKU. Para estos casos puntuales, como paso previo al

entrenamiento de modelos, se computó la demanda faltante utilizando regresiones lineales (este punto será detallado en la sección 2.4.1)

- Inv_ini → el inventario disponible en las tiendas al iniciar la semana se calculó como la suma del stock remanente al finalizar la semana y la venta semanal.

- % dev → el porcentaje de devoluciones se calculó como el cociente entre las devoluciones (en toneladas) y el sell in (en toneladas).

- Reemplazo → si bien no es una variable calculada, se incluyó esta dummy para contemplar pequeñas modificaciones que tuvo un SKU (como cambio de gramaje o una mejora en la fórmula), que por cuestiones administrativas se les asigna un código de identificación nuevo pero es muy similar a otro SKU existente que se discontinuó. El objetivo al incluir esta dummy es generar una suerte de continuidad entre el SKU viejo y el nuevo para no perder la información histórica de precios, demanda y devoluciones. Se asignó un 1 a los SKU que fueron reemplazos de otros y un 0 a aquellos que no lo fueron.

Otro aspecto importante a tener en cuenta es la inclusión de variables con información histórica de precios, demanda y devoluciones (puntualmente de las últimas 4, 3 y 2 semanas, así como también de la semana anterior a la que se está analizando). La incorporación de información rezagada sobre las ventas, los precios y las devoluciones de cada sucursal permiten modelar patrones autorregresivos que pudieran existir en el comportamiento de las mismas y potencialmente tener un aporte significativo (en términos predictivos) a la hora de predecir la demanda de la semana actual.

2.2 Análisis exploratorio de los datos

Para comenzar con el análisis descriptivo, la Tabla 2 resume los principales indicadores que se verán a lo largo de la sección para las 3 cadenas. Las Cadenas 2 y 3 presentan valores muy similares de quiebres de góndola en torno al 3,5% mientras que el porcentaje correspondiente a la Cadena 1 es de 1,5%. En lo que respecta a variedad de portfolio, la Cadena 1 es la que presenta una gama de productos más amplia con 232 SKU´s mientras que las Cadenas 2 y 3 tienen 206 y 96 respectivamente.

La Cadena 2 es la más grande en términos de cantidad de tiendas, pero las mismas se encuentran todas concentradas en 4 provincias. Por su parte, la Cadena 3 si bien es la más chica con 90 tiendas, es la que presenta una distribución más equitativa a lo largo del país teniendo presencia en 22 provincias. La Cadena 1 tiene una posición intermedia en relación a sus pares con 161 tiendas repartidas en 10 provincias.

Con respecto a la distribución del sell out, son las mismas 5 marcas las que concentran el grueso de las ventas tanto en la Cadena 1 como la 2. En la Cadena 3, en cambio, la Marca 5 forma parte de este grupo en lugar de la Marca 9. A pesar de que las marcas sean las mismas, la participación dentro del sell out en cada cadena es diferente. La Marca 15, por ejemplo, si bien es la que concentra la mayor parte de las ventas en las tres cadenas, en la tercera tiene un peso mayor que en las otras dos. En el caso de la Marca 10, en la Cadena 2 tiene una participación de más del doble en relación al resto. La Marca 4, por su parte, concentra alrededor del 26,7% del sell out en la Cadena 3 mientras que en la Cadena 1 es de tan sólo un 10,5% y de un 18,9% en la Cadena 2. La participación de la Marca 6 para las dos primeras cadenas es muy similar; situándose levemente por encima para la tercera cadena. Por último, la Marca 9 tiene mucha presencia en la primer cadena (alcanzando casi un 32% de las ventas) mientras que el Cadena 2 su participación es menos significativa siendo de un 12% nomás.

Tabla 2: Comparación entre cadenas análisis descriptivo

		CADENA 1	CADENA 2	CADENA 3
Atributos de los SKU´s	% quiebres	1,5%	3,4%	3,5%
	# SKU´s	232	206	96
Atributos de las tiendas	# tiendas	161	205	90
	# regiones	5	5	6
	# provincias	10	4	22
	# formatos	3	3	2
Marcas con mayor sell out (% del sell out)	Marca 4	10,5%	18,9%	26,7%
	Marca 5	-	-	4,7%
	Marca 6	3,3%	3,7%	5,0%
	Marca 9	31,9%	12,1%	-
	Marca 10	6,6%	16,8%	8,1%
	Marca 15	35,4%	30,9%	42,3%

2.2.1 Cadena 1

La base inicial con los datos de la Cadena 1 tenía 39 variables y 844.514 observaciones. De las 39 variables iniciales, 12 eran categóricas, 9 integers y 18 numéricas. Se analizaron datos de 161 tiendas (pertenecientes a 3 tipos de formatos), ubicadas en 5 regiones y 10 provincias. Con respecto a los productos, se incluyeron 232 SKU´s, pertenecientes a 17 tipos de familias diferentes, 27 sabores, 16 marcas y 5 tipos de envase.

Tabla 3: Variables categóricas Cadena 1

Variable	# Categorías
Tienda	161
SKU	232
Región	5
Provincia	10
Formato	3
Sabor	27
Marca	16
Envase	<
Familia	17

Si se pone el foco en la variable "demanda" (que, en definitiva, será la variable que buscaremos predecir con los distintos modelos), hay un 1,5% de los registros del conjunto de datos que se corresponde con situaciones de quiebres de góndola. Es decir, a la hora de construir la variable demanda, hubieron 13.158 registros de los 844.514 en los cuales la venta semanal en unidades igualó el inventario inicial en las tiendas y por ende, no fue posible conocer el verdadero valor de la demanda. A efectos de modelar la demanda, estos casos se considerarán como casos perdidos y utilizando un modelo de regresión lineal se imputarán los mismos a fines de la estimación. La Tabla 4 muestra, para las variables consideradas, las que tuvieron los mayores niveles de quiebres:

Tabla 4: Análisis quiebres Cadena 1

Variable	% de quiebres	% del total de tiendas	% del total de SKU	% del sell out
Provincia: Río Negro	18,4%	18,6%	-	18,2%
Formato: Formato_2	46,2%	53,4%	-	34,6%
Marca: Marca_15	34,0%	-	34,1%	35,4%
Familia: Familia_5	36,0%	-	37,1%	61,5%
Sabor: Sabor_12	22,3%	-	22,4%	36,4%

Es decir, Río Negro fue la provincia que mayor concentración de quiebres tuvo con un 18,4%; pero a su vez, es la que mayor porcentaje de tiendas tiene. El formato 2 de tiendas fue el que concentró el 46,2% de los quiebres, pero más de la mitad de las tiendas pertenecen a este formato. Dentro de los atributos de los SKU´s, los mayores porcentajes de quiebres se corresponden con valores muy similares al porcentaje del total de SKU´s que están incluidos dentro de cada categoría.

Sin tener en cuenta los casos de quiebres, la demanda (en unidades) por local y por semana presenta una distribución asimétrica positiva, concentrada principalmente en valores menores a 250. Esto es razonable si tenemos en cuenta que estamos analizando demanda semanal a nivel tienda y SKU.

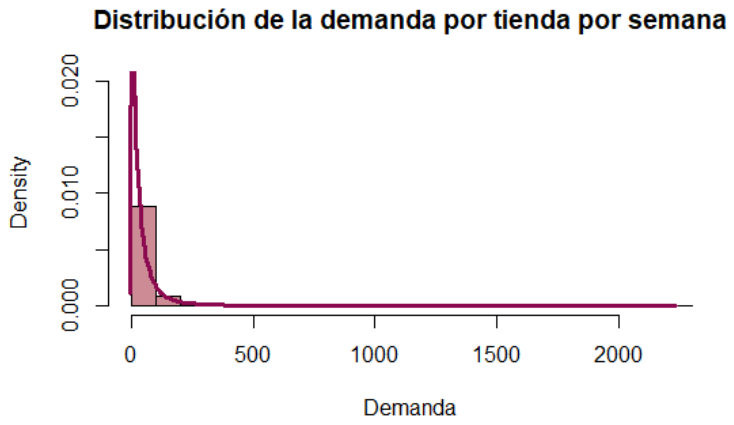


Gráfico 1: Distribución de la demanda en unidades de la Cadena 1

A nivel región, la distribución de la demanda no presenta grandes diferencias, a excepción de la región Sur donde se observan algunos outliers por encima de la media (tomando valores mayores a 1500).

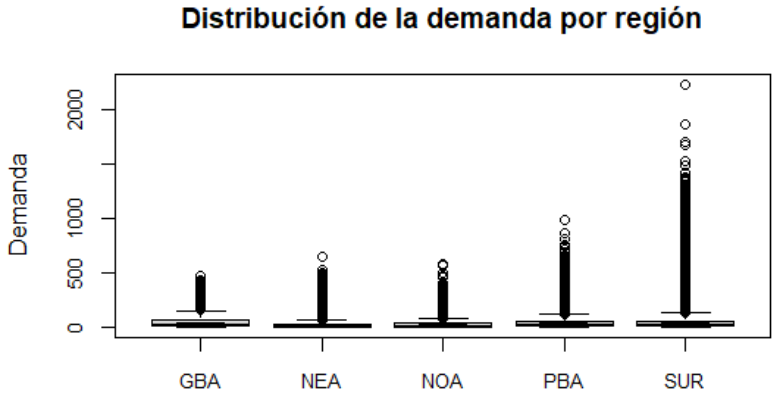


Gráfico 2: Distribución de la demanda en unidades abierto por región en la Cadena 1

Resultan interesantes también los resultados que se desprenden de analizar la evolución a lo largo de los meses de los niveles de demanda junto con los niveles de inventario. El Gráfico 3 muestra que, para todos los meses, el stock en unidades se mantuvo en niveles

superiores a la demanda experimentada por las tiendas. Es decir, en términos generales, la Cadena 1 tuvo a disposición niveles de inventario que permitieron abastecer la demanda sin quebrar stock. A su vez, tanto los niveles de inventario como los de demanda presentaron tendencias similares a lo largo del año considerado.

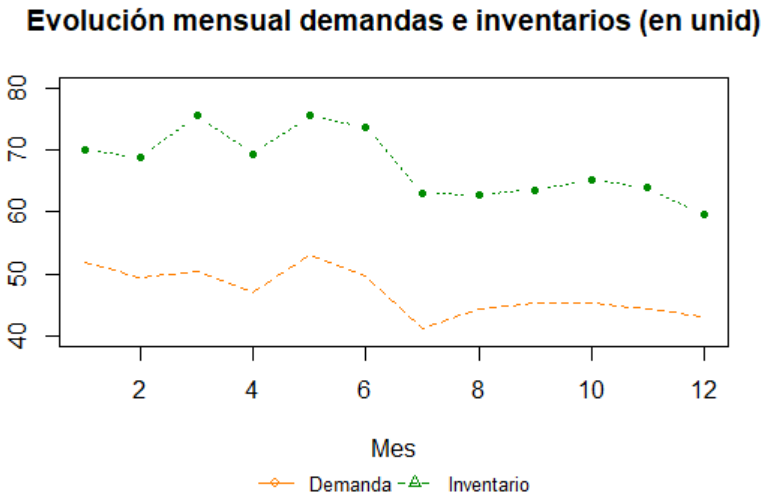


Gráfico 3: Evolución mensual de la demanda e inventario en unidades en la Cadena 1

Al desagregar el análisis por marca, encontramos que las 5 marcas que concentran el mayor porcentaje del sell out son: la Marca_15 (con el 35,4%), la Marca_9 (con el 31,86%), la Marca_4 (con el 10,46%), la Marca_10 (con el 6,56%) y la Marca_6 (con el 3,26%). Es decir, aproximadamente el 87% de las ventas al consumidor final por parte de la Cadena 1 se concentra en estas 5 marcas mencionadas. Teniendo esto en cuenta, podemos observar en el Gráfico 4 cómo se distribuye la demanda en estas 5 marcas:

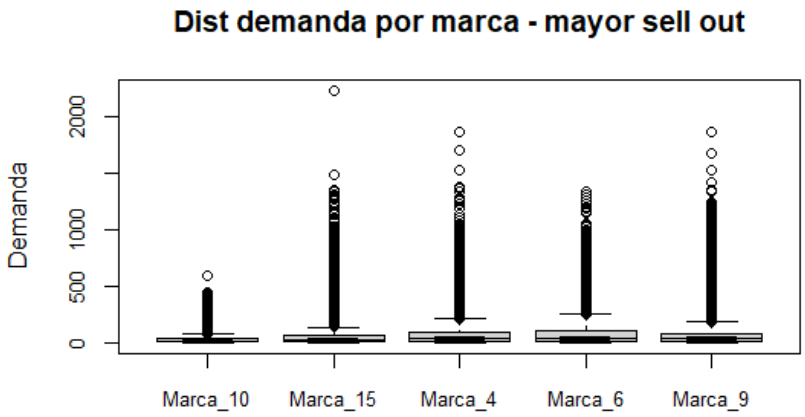


Gráfico 4: Distribución de la demanda en las 5 marcas con mayor sell out en la Cadena 1

Otra variable que probablemente tendrá un rol fundamental en los modelos predictivos de demanda es el precio. Y es en esta instancia donde considero importante tener en cuenta no sólo el precio por unidad que el consumidor termina pagando por el producto, sino también el precio por kilo ($/kg) ya que éste último es el que define qué formatos de SKU son los más rentables para la empresa. El Gráfico 5 muestra que si bien el precio unitario presenta una distribución asimétrica positiva donde el grueso de los valores se sitúa entre 25 y 100; la variable precio por kilo ($/kg) tiene una distribución más centrada en torno a la media de 335.

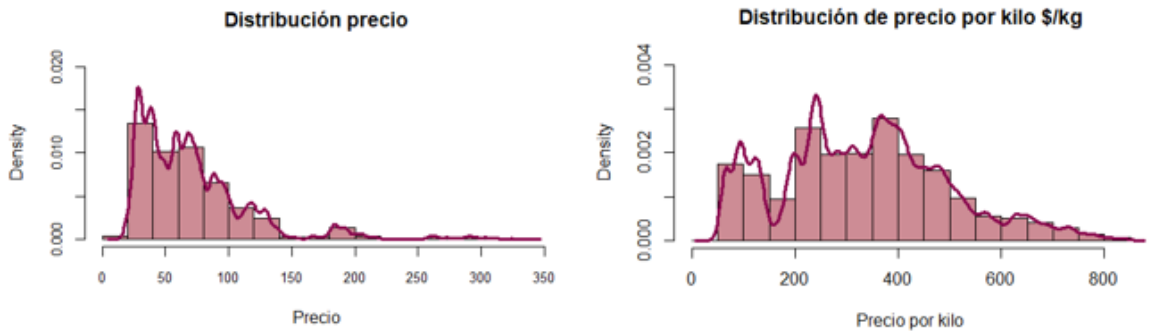


Gráfico 5: Distribución del precio y precio por kilo en la Cadena 1

Así como la distribución de la demanda no presentaba grandes diferencias entre las distintas regiones, algo similar ocurre al analizar la distribución del precio y del precio por kilo ($/kg). La región del Sur presenta valores levemente mayores que en el resto de las regiones, con los valores máximos perteneciendo también a esta región.

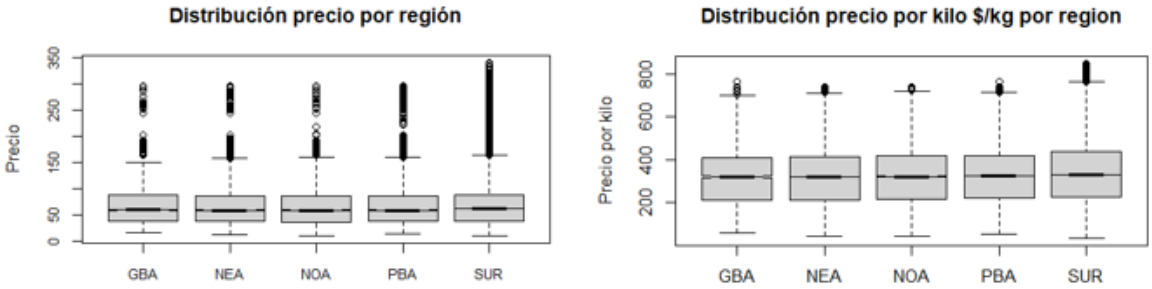


Gráfico 6: Distribución del precio y precio por kilo por región en la Cadena 1

Resulta interesante también entender cómo es la distribución del precio unitario y del precio por kilo en las 5 marcas identificadas previamente que concentran alrededor del 87% del sell out de la Cadena 1. En el Gráfico 7 se puede observar que la Marca 4 no sólo es la

más cara en términos de “out of pocket” sino que también es la que presenta la mayor dispersión de precio (con valores medios de 130). La Marca 9, por su parte, pareciera ser la más económica con un precio medio de $39. Por otro lado, si consideramos al precio por kilo como variable, notamos que es la Marca 6 la que presenta una mayor dispersión en sus valores, así como también, la que tiene los valores máximos. Este fenómeno es interesante ya que, en términos de precio, la Marca 6 no presentaba grandes dispersiones y era una de las marcas que tenía más bajos los precios en góndola. Por su parte, la Marca 4 es la que presenta la menor dispersión de los precios por kilo de sus productos; esto demuestra que si bien los SKU que engloba esta marca son más caros (en términos absolutos) para el consumidor cuando se enfrenta a la góndola, si se tiene en cuenta el gramaje, no hay grandes diferencias entre los mismos.

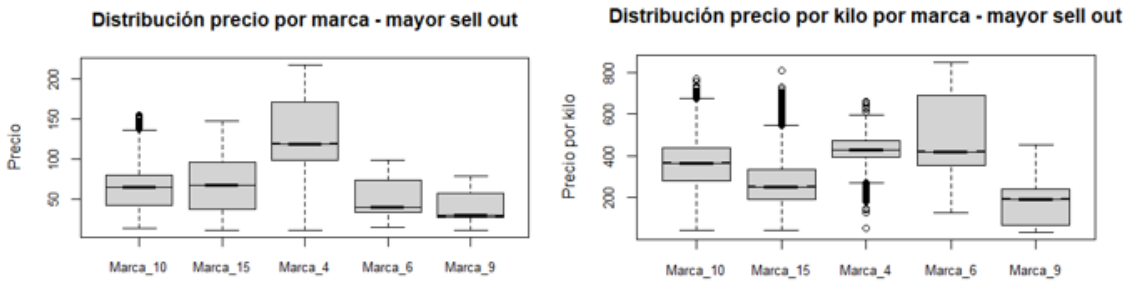


Gráfico 7: Distribución del precio y precio por kilo por marca con mayor sell out en la Cadena 1

Para finalizar el análisis de la variable precio para la Cadena 1, considero interesante entender cómo fueron sus variaciones a lo largo de los meses y compararlas con las variaciones que experimentaron la demanda y los niveles de inventario. El Gráfico 8 muestra que, de enero a agosto, los niveles de demanda y stocks presentaron tendencias similares. Por otro lado, las variaciones de precio se movieron en direcciones opuestas a las de precio e inventarios (con mucha lógica ya que en los meses en donde hubo aumentos de precios, la demanda y los stocks disminuyeron). Esta relación inversa se muestra de manera bastante clara en el mes de julio, donde un aumento del 3% de precio se vio acompañado de una caída del 15% aproximadamente de la demanda y de los niveles de inventario.

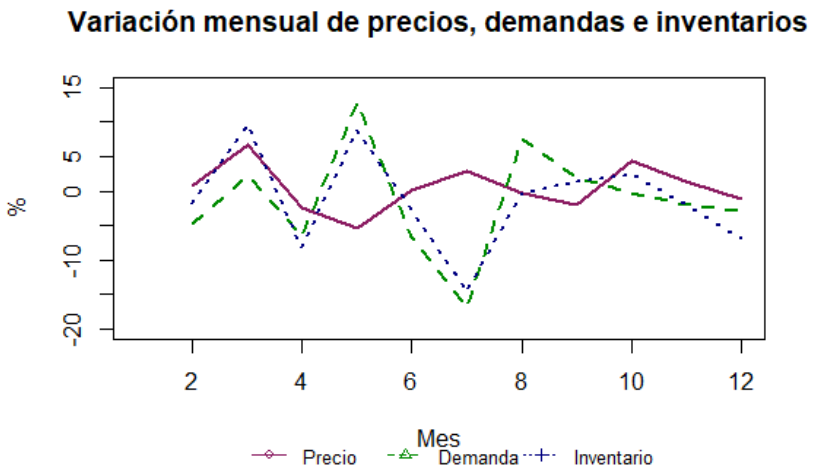
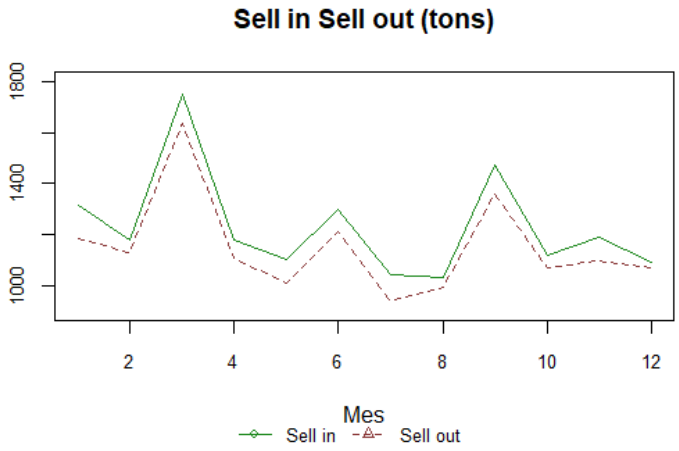


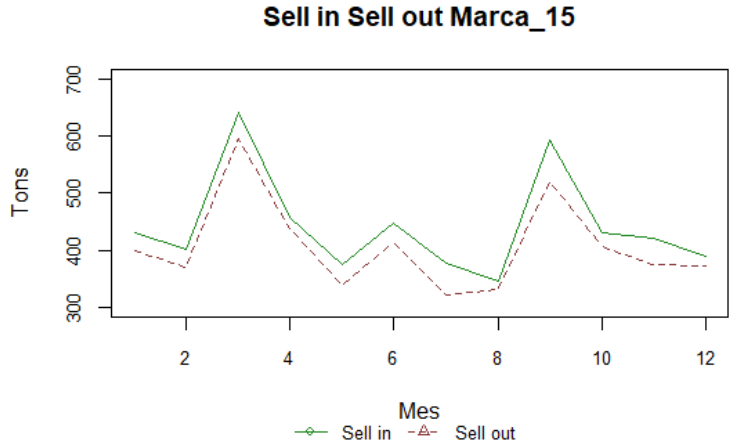
Gráfico 8: Variación mensual del precio, la demanda y del inventario en la Cadena 1

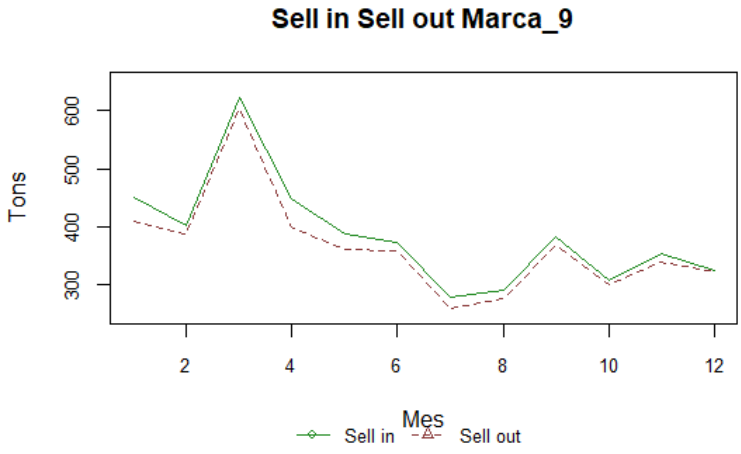
Otro aspecto interesante a tener en cuenta a la hora de estudiar la demanda en consumo masivo, es cómo es la relación entre el sell in y el sell out. Entender cómo se mueven ambas curvas sirve como termómetro para saber si el volumen de sell in que experimenta el proveedor es producto de demanda genuina por parte del consumidor final o si las Grandes Cuentas están stockeandose por algún motivo.

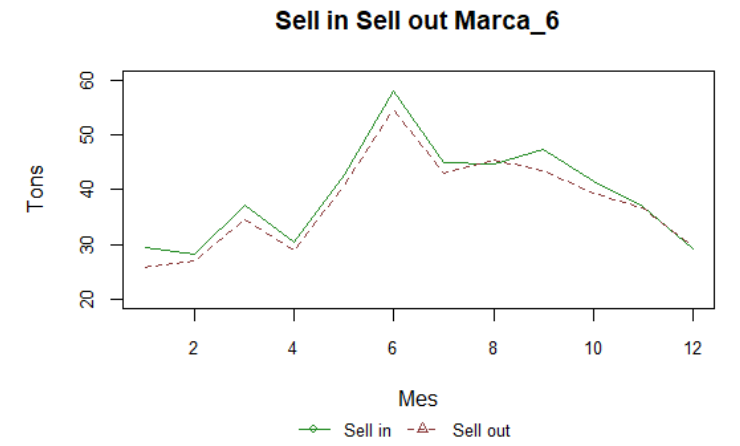
El Gráfico 9 muestra la evolución de sell in y sell out a nivel total para la Cadena 1 y para cada una de las 5 marcas identificadas previamente con mayor sell out. Se puede observar que en general, ambas curvas se mueven en la misma dirección, donde la curva que muestra los volúmenes del sell in se ubica por encima de las de sell out. Esto implica que las ventas en toneladas de la empresa proveedora de lácteos a las cadenas de supermercados es mayor que las ventas que las Grandes Cuentas realizan en sus tiendas.

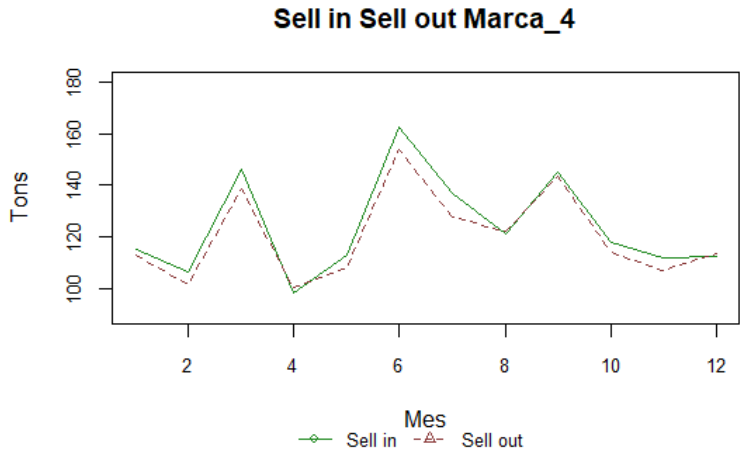
Un aspecto interesante a destacar es que las Marcas 10, 15 y 9 muestran una tendencia muy similar a la que se observa en el primer gráfico (que describe la situación total de la Cadena 1). Sin embargo, la Marca 6 pareciera tener cierta estacionalidad durante el invierno ya que el pico de ventas lo experimenta de junio a septiembre y luego empieza a caer durante los meses del verano.


Sell in Sell out (tons)
1800
1400
1000
2
4
6
8
10
12
Mes
Sell in
Sell out


Sell in Sell out Marca_15
Tons
700
600
500
400
300
2
4
6
8
10
12
Mes
Sell in
Sell out


Sell in Sell out Marca_9
Tons
600
500
400
300
2
4
6
8
10
12
Mes
Sell in
Sell out





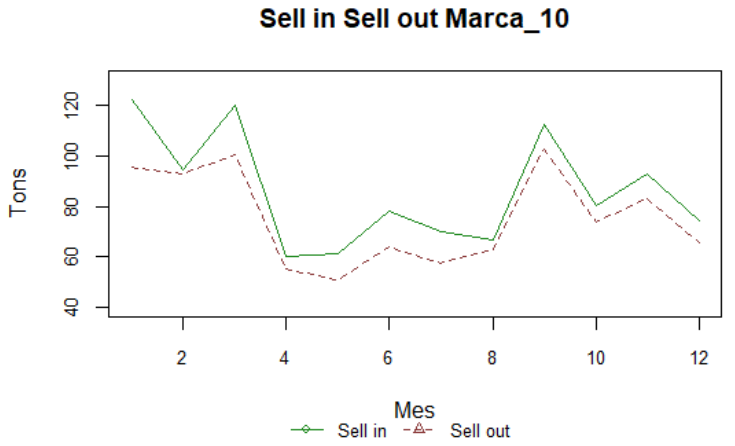


Gráfico 9: Variación mensual del sell in sell out total cadena y marcas con mayor sell out en la Cadena 1

2.2.2 Comparación entre cadenas

El mismo análisis planteado en la sección 2.2.1 se realizó para las Cadenas 2 y 3. Los resultados obtenidos fueron bastante similares a los de la Cadena 1 y el detalle se puede encontrar en las Tablas 6.1.1, 6.1.2 y 6.1.3 dispuestas en el Anexo. A continuación, se presentan los aspectos más destacables encontrados para las Cadenas 2 y 3.

En relación a la Cadena 2, resulta interesante entender cómo fue la evolución mensual de los niveles de inventario y demanda a lo largo del año. El Gráfico 10 muestra que, a excepción del mes de marzo, para el resto del 2020 los stocks se mantuvieron por encima de la demanda. Lo que se observa en marzo resulta interesante ya que la demanda y los inventarios tienen el mismo valor; esto implica que en ese mes probablemente se hayan generado quiebres de góndola producto de que la gente salió a stockearse de cara a la pandemia. Es decir, la Cadena 2 no estaba lo suficientemente abastecida de productos para satisfacer el pico de demanda que se generó en marzo. Sin embargo, de abril en adelante se puede notar cómo la cadena levantó sus inventarios de manera de mantener siempre stocks por encima de la demanda.

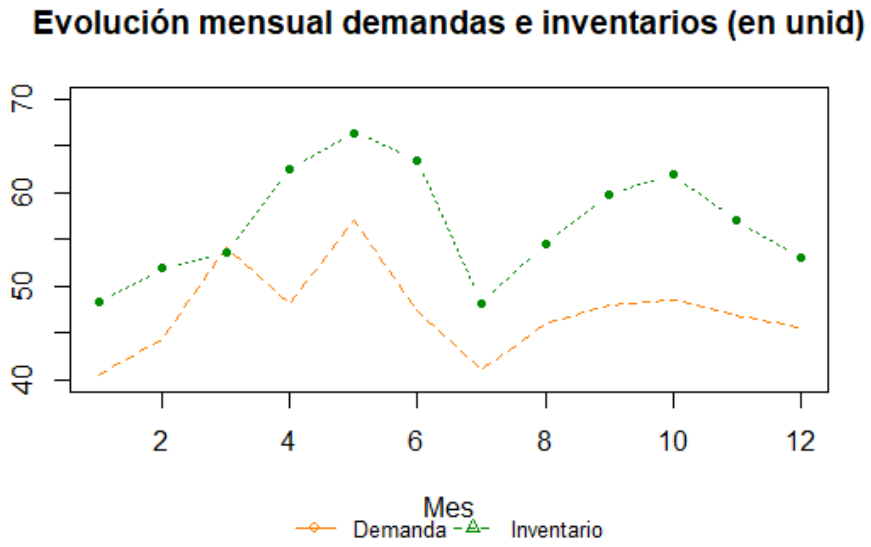


Gráfico 10: Evolución mensual de la demanda y del inventario en la Cadena 2

Otro aspecto relevante a la hora de describir los datos que caracterizan a la Cadena 2 es entender cómo fue la evolución mensual de la variación del precio, de la demanda y de los inventarios. En el Gráfico 11 se puede ver que la demanda experimentó diversos picos de aumento y de caída a lo largo del año. Si miramos específicamente el mes de marzo, que fue el inicio de la pandemia, la demanda aumentó un 20% aproximadamente respecto al mes anterior, mientras que pareciera no haber habido variaciones del precio en ese mes. El pico de demanda de marzo tuvo como contraparte una caída de los niveles de inventario.

En abril se puede observar una recomposición en los stocks (con un aumento por encima del 10%) mientras que la demanda experimentó una caída del 10% aproximadamente. En el mes de julio, se observa un aumento de precio que se vio acompañado de una disminución en los niveles de demanda y stocks. Acercándonos a fin de año, pareciera haber otro aumento de precio relativamente pequeño, junto con una caída de la demanda e inventarios.

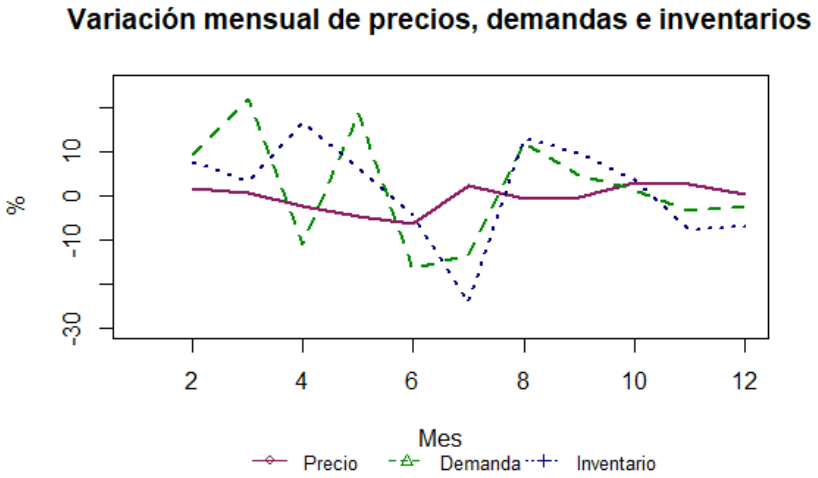
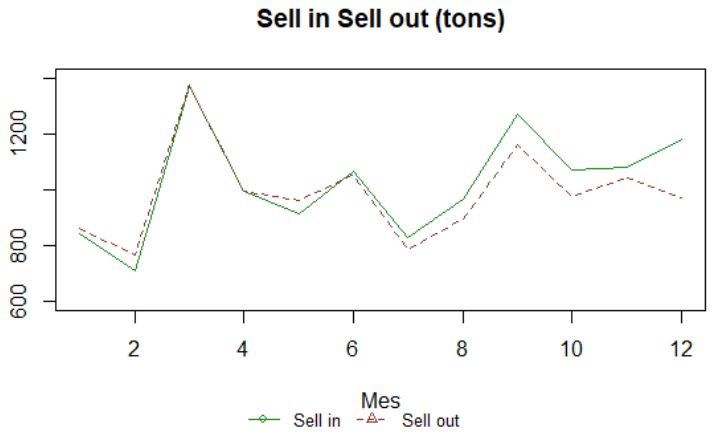
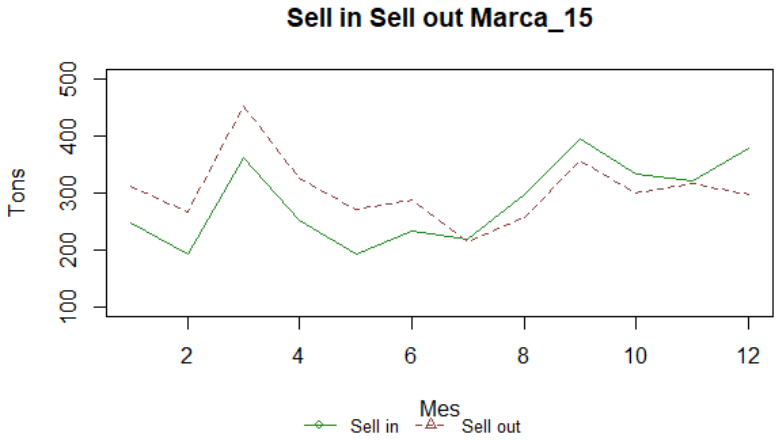
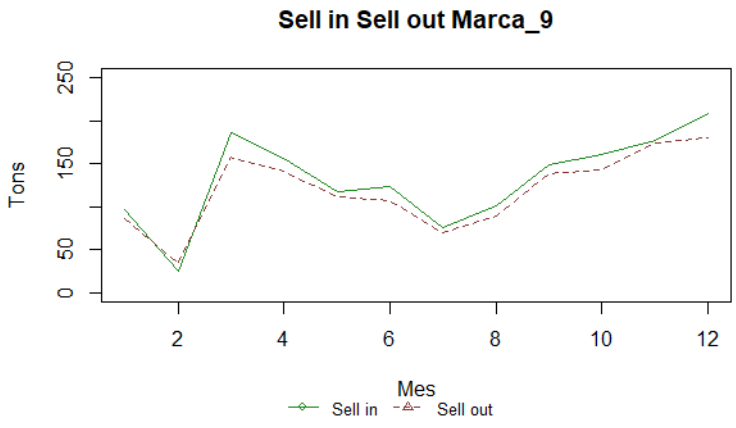


Gráfico 11: Variación mensual del precio, la demanda y del inventario en la Cadena 2

Con respecto a la evolución del sell in y sell out, el Gráfico 12 muestra que en términos generales ambas métricas se movieron de manera conjunta. En marzo tanto el sell in como el sell out experimentaron un pico producto del inicio de la pandemia. Sin embargo, a diferencia de lo que ocurría en la Cadena 1, no se observa que el sell in se ubique notoriamente en niveles superiores que el sell out. Al abrir el análisis para las 5 marcas con mayor sell out, se puede observar que todas ellas, a excepción de la Marca 6, presentan tendencias similares a las del total de la Cadena 2. La Marca 6, por su parte, pareciera tener cierta estacionalidad positiva de marzo a septiembre. Un aspecto a destacar es que la Marca 15 tiene niveles de sell out mayores que el sell in de enero a julio, pero esta tendencia se revierte a partir de agosto. Esto quiere decir que las ventas de la Cadena 2 al consumidor final en toneladas fue mayor que las compras que la Cadena 2 realizó a su proveedor.


Sell in Sell out (tons)
600
800
1200
2
4
6
8
10
12
Mes
Sell in
Sell out


Sell in Sell out Marca_15
Tons
100
200
300
400
500
2
4
6
8
10
12
Mes
Sell in
Sell out


Sell in Sell out Marca_9
Tons
0
50
150
250
2
4
6
8
10
12
Mes
Sell in
Sell out

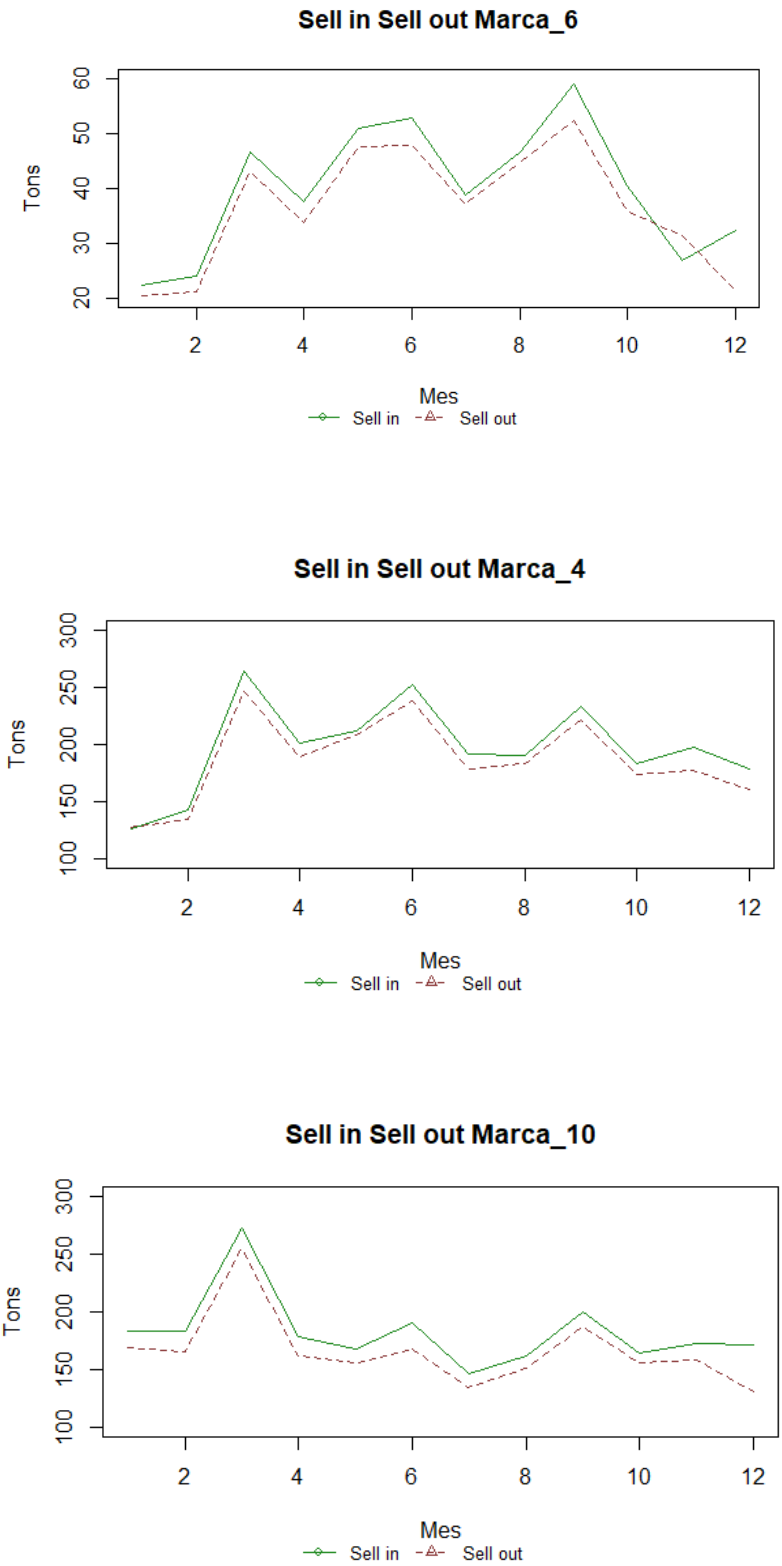


Gráfico 12: Variación mensual del sell in sell out total cadena y marcas con mayor sell out en la Cadena 2

La Cadena 3, por su parte, en términos de evolución de niveles de inventario y demanda, presenta una tendencia similar a la de la Cadena 1 en el sentido que los niveles de inventario se mantuvieron en todo momento por encima de la demanda. En el mes de julio la brecha entre ambas se redujo significativamente (pasando de un inventario de 100 unidades a 65 aproximadamente sin que este hecho se viera acompañado de un crecimiento de la demanda). Si bien en el segundo semestre la Cadena 3 logró recomponer sus stocks, los mismos se mantuvieron en niveles inferiores a los vigentes durante la primera mitad del año.

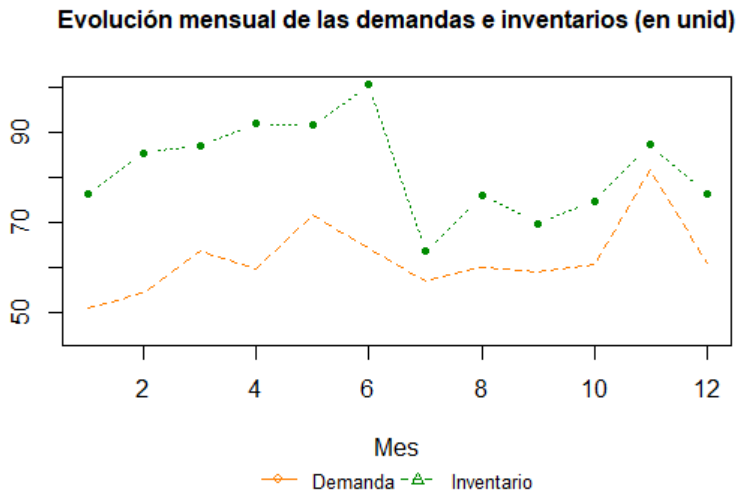


Gráfico 13: Evolución mensual de la demanda y del inventario en la Cadena 3

Otro aspecto interesante que surgió del análisis descriptivo de la Cadena 3, fue la evolución de la variación porcentual de la demanda, el precio y los inventarios a lo largo del año. El Gráfico 14 muestra que si bien las variaciones de precio experimentadas por la Cadena 3 no parecieran haber sido muy grandes (menores al 5%), la demanda y los stocks sí presentan variaciones más pronunciadas. En el caso de los inventarios, los mismos presentan una caída del 30% en el mes de julio en relación al mes anterior; para luego volver a recomponerse a través de un incremento del 20%. La demanda, por su parte, también experimentó picos de crecimiento (en noviembre por ejemplo fue del 34%) y caídas, pero no tan agudas como el inventario (la caída máxima fue de un 25% en diciembre).

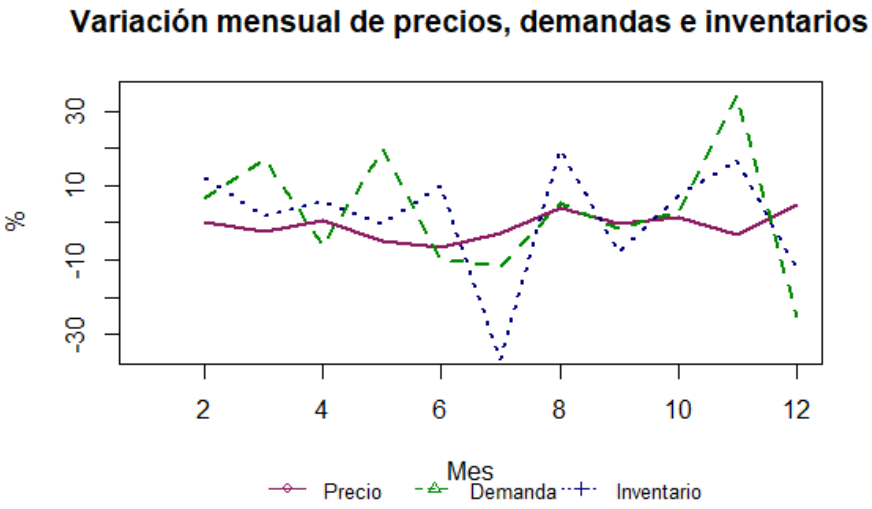
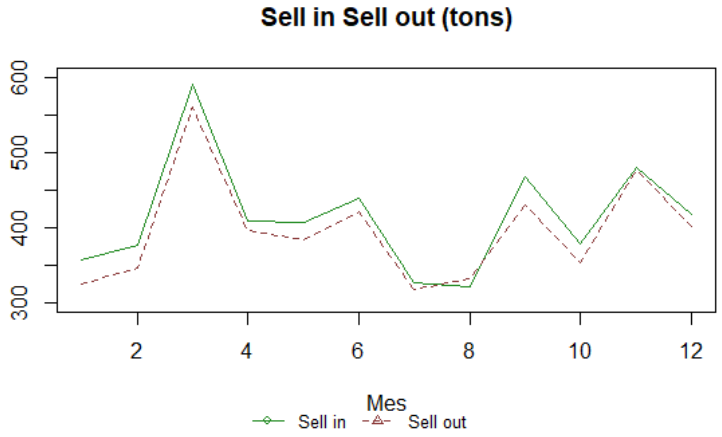
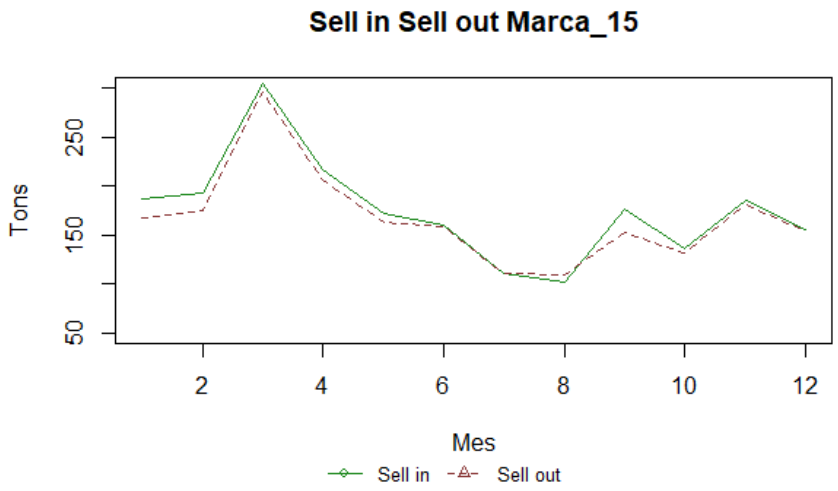
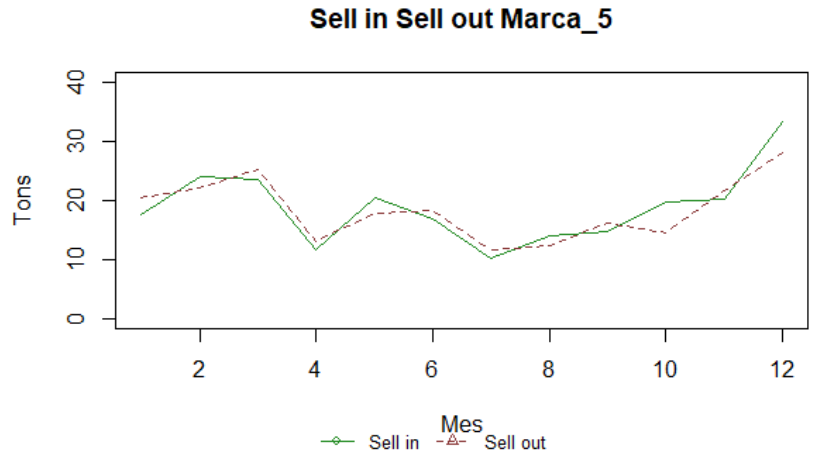


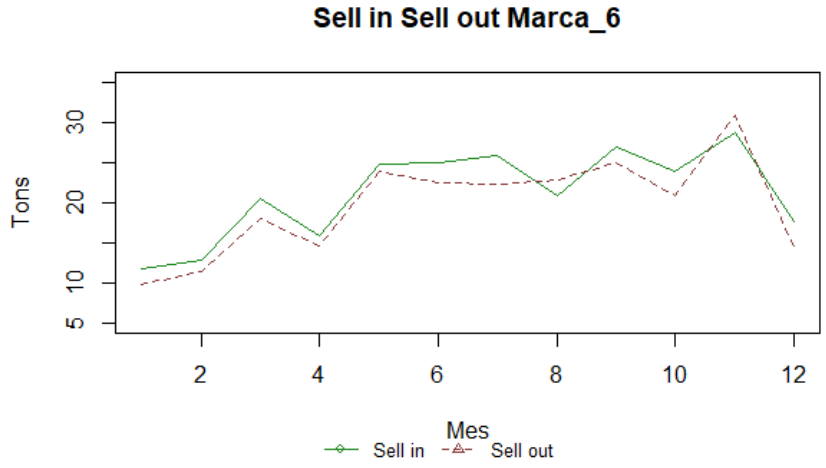
Gráfico 14: Variación mensual del precio, la demanda y del inventario en la Cadena 3

Para finalizar el análisis descriptivo de la Cadena 3, resulta interesante entender cómo fue la evolución del sell in y sell out a nivel general y en las 5 marcas con mayor sell out. El Gráfico 15 muestra que la Cadena 3 experimentó un pico de ventas en marzo (asociado con el inicio de la pandemia) y que el sell in se mantuvo levemente por encima del sell out a lo largo de todo el año. Tanto la Marca 15, como la 10 y la 4 presentan tendencias similares a la del total de la Cadena. La Marca 5, por su parte, pareciera no sólo tener cierta estacionalidad durante el verano ya que a partir de septiembre comienzan a crecer las ventas sino también que en algunos meses el sell out se ubica por encima del sell in. Finalmente, la Marca 6 presenta cierta estacionalidad durante el invierno ya que las ventas alcanzan valores mayores de mayo a noviembre para luego caer durante el verano.




Sell in Sell out Marca_15
Tons
50
150
250
2
4
6
8
10
12
Mes
Sell in
Sell out


Sell in Sell out Marca_5
Tons
0
10
20
30
40
2
4
6
8
10
12
Mes
Sell in
Sell out


Sell in Sell out Marca_6
Tons
5
10
20
30
2
4
6
8
10
12
Mes
Sell in
Sell out

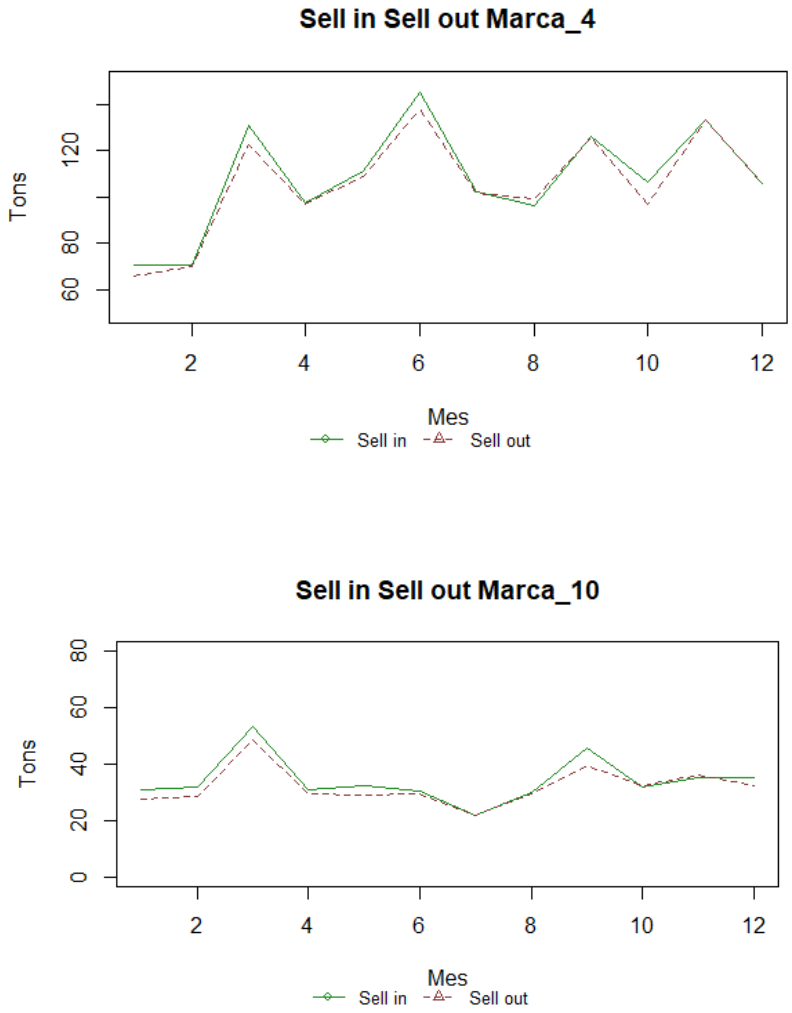


Gráfico 15: Variación mensual del sell in y sell out total cadena y marcas con mayor sell out en la Cadena 3

Las diferencias sistemáticas entre el sell in y sell out presente en las tres cadenas analizadas, refuerza la necesidad y la importancia de poder contar con modelos predictivos que contribuyan a mejorar las estimaciones de la demanda. De esta manera, se podría reducir la brecha existente entre la venta realizada en las tiendas y las ventas del proveedor de lácteos a las Grandes Cuentas.

2.3 Ingeniería de atributos

Se denomina “ingeniería de atributos” al proceso de formular y transformar las variables teniendo en cuenta los datos, el modelo y la tarea a realizar (Zheng y Casari, 2018). Este proceso tiene lugar una vez que ya se cuenta con los datos, pero es un paso previo al entrenamiento de los modelos.

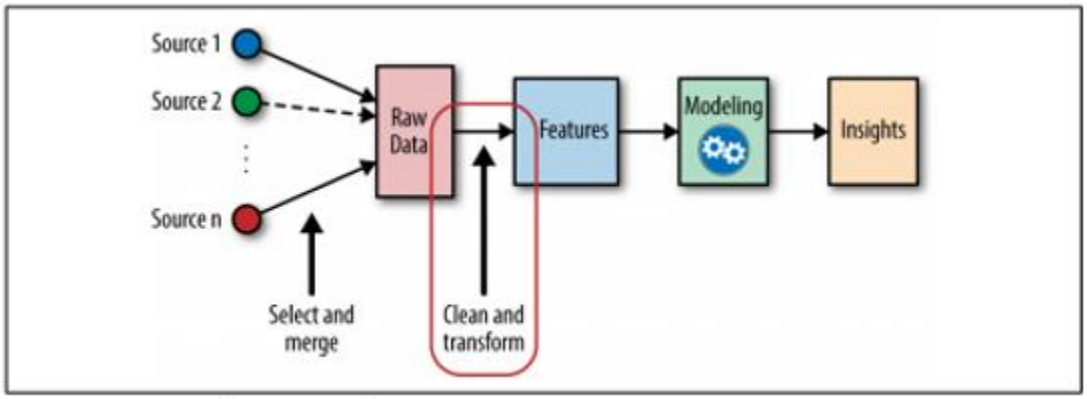


Fig.1: Ubicación de la ingeniería de atributos en el proceso de Machine Learning (Zheng y Casari, 2018)

Entre las múltiples variables que se construyeron para predecir demanda se destaca, en primer lugar, el sell out en toneladas definido como:

Sell out (tons) = (GRM/1.000.000) * sell out (unid)

En segundo lugar, a partir de la variable precio, se generó la variable precio/kg ya que se consideró que el mismo es un atributo fundamental a la hora de predecir demanda y que sería importante sumarla al modelo. La diferencia entre ambas medidas es que el precio refleja el "out of pocket", es decir, la cantidad de dinero que el consumidor tiene que desembolsar por esa unidad de producto. En cambio, el precio/kg es una manera de homogeneizar a todos los SKU y permite evaluar efectivamente cuales son más caros.

Precio/kg ($) = (Precio*1.000)/GRM

Otro aspecto que se esperaba (previo al entrenamiento de los modelos) que fuera fundamental a la hora de predecir demanda son las variaciones de precio. Para poder capturar este efecto, se crearon tres variables:

Var_precio_porc → refleja la variación porcentual del precio entre la semana actual y la anterior para cada SKU en cada tienda

Descuento → es una dummy que tiene el valor de 1 si la baja de precio entre la semana actual y la anterior es mayor al 5%

Aumento → es una dummy que tiene el valor de 1 si el aumento de precio entre la semana actual y la anterior es mayor al 5%

La técnica de ingeniería de atributos que está por detrás de la generación de las variables "descuento" y "aumento" se llama "binarización" y consiste en crear dummys en base a algún valor (en este caso, a valores mayores al 5% para la variable aumento y menores al 5% para la de descuento).

En tercer lugar, se crearon 4 variables dummys para generar agrupaciones por categoría de productos:

Yogures → una dummy que tiene el valor de 1 si el SKU en cuestión forma parte de la categoría de yogures, y un 0 en caso contrario.

Postres → una dummy que tiene el valor de 1 si el SKU en cuestión forma parte de la categoría de postres, y un 0 en caso contrario.

Quesos → una dummy que tiene el valor de 1 si el SKU en cuestión forma parte de la categoría de quesos, y un 0 en caso contrario.

Leches → una dummy que tiene el valor de 1 si el SKU en cuestión forma parte de la categoría de leches, y un 0 en caso contrario.

En cuarto lugar, se crearon 8 variables de interacción utilizando cada una de las dummys de categorías de productos y las variables de aumento y descuento generadas previamente. Según Zheng y Casari (2018) las variables de interacción son el producto entre otras dos variables y su utilización ayuda a capturar cualquier tipo de interacción que pueda llegar a existir entre ellas. En el caso de que los dos atributos que participen de la interacción sean dummys, la interacción resultante tendrá el valor de 1 si esa combinación cumple con los dos requisitos, y tendrá el valor de 0 en caso de que al menos una de las dos condiciones no se cumpla. Puntualmente, las variables de interacción que se crearon fueron las siguientes:

Yogures_dinamica → interacción entre la dummy "yogures" y la dummy "dinámica". Ej: tendrá el valor de 1 si ese SKU pertenece a la categoría de yogur y la baja de precio respecto a la semana anterior fue mayor al 5%

Postres_dinamica → interacción entre la dummy "postres" y la dummy "dinámica". Ej: tendrá el valor de 1 si ese SKU pertenece a la categoría de postres y la baja de precio respecto a la semana anterior fue mayor al 5%

Quesos_dinamica → interacción entre la dummy "quesos" y la dummy "dinámica". Ej: tendrá el valor de 1 si ese SKU pertenece a la categoría de quesos y la baja de precio respecto a la semana anterior fue mayor al 5%

Leches_dinamica → interacción entre la dummy "leches" y la dummy "dinámica". Ej: tendrá el valor de 1 si ese SKU pertenece a la categoría de leches y la baja de precio respecto a la semana anterior fue mayor al 5%

Yogures_aumento → interacción entre la dummy "yogures" y la dummy "aumento". Ej: tendrá el valor de 1 si ese SKU pertenece a la categoría de yogur y la suba de precio respecto a la semana anterior fue mayor al 5%

Postres_aumento → interacción entre la dummy "postres" y la dummy "aumento". Ej: tendrá el valor de 1 si ese SKU pertenece a la categoría de postres y la suba de precio respecto a la semana anterior fue mayor al 5%

Quesos_aumento → interacción entre la dummy "quesos" y la dummy "aumento". Ej: tendrá el valor de 1 si ese SKU pertenece a la categoría de quesos y la suba de precio respecto a la semana anterior fue mayor al 5%

Leches_aumento → interacción entre la dummy "leches" y la dummy "aumento". Ej: tendrá el valor de 1 si ese SKU pertenece a la categoría de leches y la suba de precio respecto a la semana anterior fue mayor al 5%

Por último, se eliminaron ciertas variables con demasiadas categorías que no aportaban mucho al análisis y cuyas características ya estaban incluidas dentro de otros atributos. Este tratamiento se hizo para las variables de SKU, número de tienda y cadena, mes y número de la semana. Sin embargo, en vistas de poder contar con predicciones individualizadas, en la sección Anexo se puede encontrar un análisis realizado para los dos SKU con mayor participación en el sell out de cada cadena; donde se aplica el mejor modelo encontrado a cada uno de ellos y se compara la performance respecto al modelo benchmark.

2.4 Modelos y Métricas de Evaluación

En primer lugar, todos los modelos utilizados a lo largo del presente trabajo son modelos de aprendizaje supervisado. Lo que caracteriza a este tipo de modelos es que para cada observación se tiene un conjunto de atributos, que de ahora en adelante denoto como X_1, X_2, X_3 … X_p, así como también una variable respuesta (Y_i) asociada que se buscará predecir utilizando los atributos antes mencionados. El objetivo será entonces poder entrenar un modelo que relacione las X con la Y, poniendo el foco principalmente en lograr generar predicciones lo más precisas posibles de cara a nuevas y futuras observaciones (James, Witten, Hastie y Tibshirani, 2013).

Un aspecto central previo a la modelización de la demanda, consiste en definir un conjunto de métricas que se usarán para medir la calidad de las predicciones. Siguiendo las prácticas habituales cuando se trata de problemas de regresión, en esta tesis se utilizaron el Error Cuadrático Medio (ECM) y el Coeficiente de Determinación (R^2) definidos a continuación:

$$\mathrm{ECM} = \frac{1}{n}\sum_{i=1}^{n}(\hat{Y}_i - Y_i)^2$$

$$\hat{R}^2 = 1 - \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}$$

El Error Cuadrático Medio (ECM) mide qué tan lejos se encuentran las predicciones realizadas por el modelo de los verdaderos valores observados; donde Yˆ denota las predicciones de demanda obtenidas por los modelos estimados e Y es la variable que indica los valores reales observados de demanda. Cuanto menor sea esta medida, mejor será el modelo en cuestión. Es fundamental tener en cuenta que lo que verdaderamente importa a la hora de definir la calidad predictiva de un modelo es cómo es su performance por fuera de los datos de entrenamiento. Es decir, un modelo puede tener un ECM bajo sobre los datos de train pero ser mucho más elevado sobre los datos de validación. Es por esta razón, que, a la hora de optar por un modelo en lugar de otro, hay que inclinarse por aquel que tenga el menor ECM sobre un conjunto de validación.

El R^2 por su parte, indica qué proporción de la variabilidad total presente en la variable respuesta está siendo explicada por el modelo. En este caso, cuanto mayor sea el valor del R^2, mejor será la capacidad predictiva del mismo. Al igual que lo que ocurría con el ECM, lo que verdaderamente va a definir si el modelo propuesto es superior a otros en términos de estimaciones precisas será el valor que adopte el R^2 sobre datos que no hayan sido utilizados para entrenar el modelo.

2.4.1 Modelo de Regresión Lineal y regularización LASSO

El modelo lineal asume que la relación entre las variables predictoras y la que se quiere estimar es lineal:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon,$$

Donde Ɛ es un componente aleatorio no modelable de la demanda.

En contextos de alta dimensión, como en el caso del presente trabajo en donde disponemos de 120 covariables, resulta sumamente necesario implementar técnicas de regularización y/o selección de variables a fin de controlar la variabilidad del modelo y con ello obtener estimaciones de la demanda más precisas. Con estas técnicas se pretende eliminar o reducir aquellas variables que no sean del todo significativas a la hora de explicar la variable dependiente del modelo, ya que incluir variables poco relevantes aumenta la complejidad del modelo, así como también, los tiempos de procesamiento (James et.al, 2013).

En vistas de poder hacer selección de variables, en el presente trabajo se utilizó la técnica de regularización LASSO cada vez que se plantearon regresiones lineales, con lo cual, dada una muestra de entrenamiento, los parámetros del modelo de regresión se aprenden minimizando la siguiente expresión:

$$\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2+\lambda\sum_{j=1}^{p}|\beta_j| \;=\mathrm{RSS}+\lambda\sum_{j=1}^{p}|\beta_j|.$$

Donde λ es un hiperparámetro que controla el trade off entre el sesgo y la variabilidad del modelo; y su valor se elige utilizando técnicas de validación cruzada (James et.al, 2013).

2.4.2 Modelo Random Forest

Los modelos de Random Forest se encuentran dentro del grupo de técnicas de Machine Learning basadas en la construcción de árboles de decisión. Estos últimos modelos aprenden la variable de respuesta particionando el espacio de atributos en distintas regiones sin que haya superposición entre ellas. A su vez, a todas las observaciones que caigan dentro de la misma región, se les realizará la misma predicción que será el valor medio de la variable respuesta de todas las observaciones del conjunto de entrenamiento que formen parte también de esa región. El objetivo entonces de los árboles de decisión será encontrar aquella configuración de partición del espacio que logre minimizar el ECM en datos desconocidos. Es importante destacar que la construcción de los árboles se realizará en base a minimizar el ECM sobre los datos de entrenamiento a pesar de que éste no sea el principal foco de interés (James et.al, 2013).

Dentro de las principales ventajas de los árboles de decisión se encuentra su capacidad de poder lidiar con facilidad con variables categóricas y el hecho de que permiten distinguir cuáles son las variables más importantes del modelo (las que más contribuyeron a reducir el error). Sin embargo, el poder predictivo de los árboles en sí mismos no es tan potente debido a la elevada varianza que los caracteriza. A pesar de ello, existe la posibilidad de ir ensamblándolos y así poder alcanzar mejores predicciones; dando origen a los modelos de ensamble como Random Forest y XGBoost, entre otros.

Los modelos de ensamble, a diferencia de los árboles de decisión, se generan a partir del promedio de muchos modelos de árbol que se construyen utilizando técnicas de remuestreo con el fin de reducir la varianza del modelo predictivo. En particular, los modelos Random Forest consisten en crear múltiples muestras Bootstrap sobre los datos de entrenamiento y para cada una de ellas se entrena por separado un árbol de decisión. La predicción final del modelo en un problema de regresión será entonces el promedio de las predicciones de cada uno de los árboles que fueron entrenados por separado.

Los modelos Random Forest tienen varios hiperparámetros que es importante calibrar en vistas de lograr mejores predicciones. Zheng (2015) sostiene que la principal diferencia entre los parámetros y los hiperparámetros de un modelo es que los primeros son

aprendidos en la etapa de entrenamiento mientras que los segundos son valores que deben establecerse de antemano, como paso previo a entrenar los modelos. A su vez, los hiperparámetros controlan no sólo la flexibilidad del modelo, sino que también definen la capacidad predictiva del mismo.

Entre los hiperparámetros que podemos considerar sensibles del modelo Random Forest se encuentran:

- **Ntree** → es la cantidad de árboles o remuestras Bootstrap del modelo (en el caso puntual de Random Forest, una mayor cantidad de árboles no implica overfitting[1]).

- **Mtry** → es la "m" cantidad de variables que se tendrán en cuenta a la hora de realizar cada una de las particiones. Este hiperparámetro da cuenta del doble remuestreo presente en Random Forest que permite decorrelar los árboles.

- **Sample** → es el tamaño, o cantidad de observaciones, que tendrán las remuestras Bootstrap (pueden ser del mismo tamaño que la base inicial o de un tamaño menor)

- **Maxnodes** → es la cantidad máxima de nodos terminales que pueden tener cada uno de los árboles que se entrenen (cuanto mayor sea este valor, mayores probabilidades habrá de overfitting)

- **Nodsize** → es la cantidad mínima de observaciones que debe tener un nodo para que se genere una partición (cuanto menor sea este valor, mayores chances de overfitting).

Todos estos hiperparámetros se eligen minimizando el error out-of-bag (OOB); que es una manera eficiente de estimar el error del modelo similar a las técnicas de validación cruzada (James, et.al, 2013).

2.4.3 XGBoost

Los modelos de boosting también forman parte de los modelos de ensamble. Sin embargo, a diferencia de Random Forest, en boosting los árboles se construyen de forma secuencial usando información proveniente de árboles anteriores. A su vez, otro aspecto que diferencia a los modelos de boosting de los de bagging es que los primeros no realizan un remuestreo Bootstrap; sino que cada árbol utiliza de input una versión modificada del dataset original (James et.al, 2013).

[1] Overfitting es una situación que se genera cuando se tiene un modelo que se ajusta demasiado a las particularidades de los datos utilizados para entrenar el modelo. Estos casos se caracterizan por tener un bajo ECM sobre train pero un alto ECM sobre los datos de validación.

En el presente trabajo se optó por utilizar el modelo XGBoost teniendo en cuenta que es de los modelos más potentes dentro de boosting. Los hiperparámetros del modelo son los siguientes:

- **Nrounds** → es la cantidad de árboles que se van a entrenar (en este caso, a diferencia de Random Forest, un mayor número de árboles aumentará las probabilidad de cometer overfitting).

- **Max_depth** → es la profundidad máxima de cada árbol (nuevamente, cuanto más profundos sean los árboles, más probabilidades de overfitting habrá).

- **Eta** → es el learning rate λ que controla la velocidad con la que aprende el algoritmo.

- **Gamma** → es la mínima reducción del error necesaria para aceptar un nuevo corte (cumple el mismo rol que α en los árboles de decisión).

- **Colsample_bytree** → indica qué proporción del total de las variables se van a considerar en el entrenamiento de cada árbol.

- **Subsample** → indica qué proporción de las observaciones se van a considerar en cada árbol.

- **Min_child_weight** → es la cantidad mínima de observaciones que debería tener un nodo para poder realizar un corte.

2.5 Optimización de hiperparámetros

Los hiperparámetros de un modelo juegan un rol fundamental a la hora de lograr buenas predicciones. Una correcta calibración de los mismos es importante para evitar hacer overfitting, tener un modelo que sea demasiado flexible y se ajuste demasiado a los datos del conjunto de entrenamiento. Un aspecto a tener en cuenta es que la configuración óptima de hiperparámetros es propia de cada base de datos; con lo cual, los mismos deberán calibrarse adhoc en cada una de ellas (Zheng, 2015).

En el caso de los modelos lineales, se utilizó la técnica de k-fold cross validation para encontrar el mejor valor de λ. Para los modelos de Random Forest y XGBoost, la búsqueda de hiperparámetros se llevó adelante a través de una grilla. Es decir, se fueron conformando distintas grillas con combinaciones diferentes de hiperparámetros y se terminó optando por aquella que arrojara el menor ECM.

Tabla 5: Técnicas de selección de hiperparámetros por modelo

Modelo	Hiperparámetros	Método de selección
Regresión lineal	λ	k fold CV (con 5 folds)
Random Forest	Mtry	Grid Search
	Ntree	
	Maxnodes	
	Nodsize	
	Sample	
XGBoost	Nrounds	Grid Search
	Max_depth	
	Eta	
	Gamma	
	Colsample_bytree	
	Min_child_weigth	
	Subsample	

3. RESULTADOS

Como paso previo a obtener el modelo benchmark, se estimó un primer modelo de regresión para cada una de las cadenas con el fin de estimar la demanda faltante en aquellos casos identificados como "quiebres de góndola". Estos últimos casos se corresponden con tiendas y SKU´s donde las ventas semanales y el inventario semanal inicial eran iguales; y, por lo tanto, el valor observado de la demanda se encontraba censurado. Es decir, que la demanda que efectivamente se hubiera observado esa semana en esa tienda para ese SKU hubiera sido al menos igual o superior a las ventas realizadas. En otras palabras, se trataron a estas instancias como si fuesen datos faltantes y se utilizó primero un modelo de regresión lineal para imputar los valores perdidos. En la sección Anexo se puede encontrar el detalle del ejercicio realizado para cada una de las cadenas para validar la calidad predictiva de este primer modelo.

Una vez computados los valores faltantes de demanda, se dividió de forma aleatoria la totalidad de la base en dos grupos: el 80% se lo apartó para entrenar el modelo y el 20% restante se lo utilizó como conjunto de validación.

3.1 Modelo de Regresión Lineal y regularización LASSO

Los resultados obtenidos en las 3 cadenas analizadas fueron muy similares en lo que respecta a este primer modelo benchmark. En las Tablas 6.1.4 y 6.1.5 (Cadena 1), 6.1.6. y 6.1.7. (Cadena 2) y 6.1.8 y 6.1.9 (Cadena 3) dispuestas en el Anexo, se puede encontrar el detalle de los coeficientes estimados para cada una de las cadenas, así como también, cuáles fueron los atributos que el modelo terminó dejando de lado por no considerarlos relevantes. A su vez, los gráficos 6.2.13 y 6.2.14 (Cadena 1), 6.2.15 y 6.2.16 (Cadena 2) y

6.2.17 y 6.2.18 (Cadena 3) de la sección Anexo muestran cómo va seleccionando variables la técnica LASSO y cómo se van achicando los coeficientes de la regresión a medida que aumenta el valor de λ.

Dentro de las variables con un coeficiente estimado positivo (es decir, que un aumento en una unidad de ese atributo implica un aumento también en la demanda), en las tres cadenas se destacan el sell in, la covariable dummy que da cuenta si se trata de un producto incluido dentro de la categoría de quesos y a su vez, tuvo un descuento de precio superior al 5% respecto a la semana anterior (quesos_dinamica), la covariable dummy que da cuenta si se trata de un producto incluido dentro de la categoría de postres y a su vez, tuvo un descuento de precio superior al 5% respecto a la semana anterior (postres_dinamica) y descuento. Estos resultados, estarían demostrando que el hecho de aplicar una baja de precio en los productos, y en especial en ciertas categorías como quesos y postres, tiene un efecto positivo sobre la demanda. Por su parte, el hecho de que el sell in esté relacionado positivamente con los niveles de demanda tiene sentido, ya que, si las cadenas de supermercados no se abastecen de productos, la demanda no podrá tener lugar en las tiendas. Por otro lado, dentro de las variables con un coeficiente estimado negativo, se destacan las devoluciones (en toneladas), la covariable dummy que da cuenta si se trata de un producto incluido dentro de la categoría de quesos y a su vez, tuvo un aumento de precio superior al 5% respecto a la semana anterior (quesos_aumento) y la covariable dummy que da cuenta si se trata de un producto incluido dentro de la categoría de postres y a su vez, tuvo un aumento de precio superior al 5% respecto a la semana anterior (postres_aumento). Los resultados arribados por el modelo parecen razonables ya que un incremento en el nivel de devoluciones estaría reflejando una caída en el nivel de demanda. Nuevamente, las categorías de quesos y postres parecerían ser las más sensibles a variaciones en el precio (tanto para la suba como para la baja).

El valor óptimo de λ en cada una de las cadenas se buscó a través de validación cruzada con 5 folds. Los resultados obtenidos sobre el conjunto de validación en este primer modelo se resumen en la Tabla 6:

Tabla 6: Performance out of sample del modelo Regresión Lineal con LASSO

CADENA	ECM	R2	LAMBDA
Cadena 1	943,16	81,20	0,015
Cadena 2	1025,55	86,40	0,054
Cadena 3	1468,95	85,03	0,022

3.2 Modelo Random Forest

A partir del modelo benchmark, se buscó mejorar la performance sobre el conjunto de validación a través de los modelos basados en la construcción de árboles de decisión. Tanto

para Random Forest como para XGBoost la metodología fue la misma: entrenar primero un modelo con los parámetros que vienen seteados por default y luego ir calibrando hiperparámetros para lograr mejores predicciones. La Tabla 7 muestra los resultados encontrados para las tres cadenas sobre un conjunto de validación para esta primera versión previa a la calibración de hiperparámetros:

Tabla 7: Performance out of sample del modelo Random Forest sin optimizar

CADENA	ECM	R2
Cadena 1	808,75	83,98
Cadena 2	938,82	86,65
Cadena 3	1145,55	88,33

Las conclusiones más interesantes de este primer modelo de Random Forest son las que se desprenden de analizar las variables que el mismo identificó como las más significativas a la hora de explicar la demanda. El Gráfico 16 muestra que los atributos que más contribuyeron a reducir el error fueron, dentro de todo, muy similares en las tres cadenas analizadas. En los tres casos, la variable identificada por los modelos como la más relevante a la hora de estimar demanda fue la demanda de la semana anterior a la que se quiere estimar. En una segunda instancia, tanto las unidades entregadas como pedidas por las cadenas fueron destacadas por el algoritmo a la hora de predecir la demanda. A su vez, las variables relacionadas al precio (como el precio por kilo, los precios históricos y la variación porcentual del mismo) resultaron relevantes en la predicción de la demanda. Por último, en la Cadena 1, el modelo destacó algunos atributos característicos del SKU en sí, como la cantidad de unidades por bandeja (PCB) y el gramaje (GRM), como importantes.

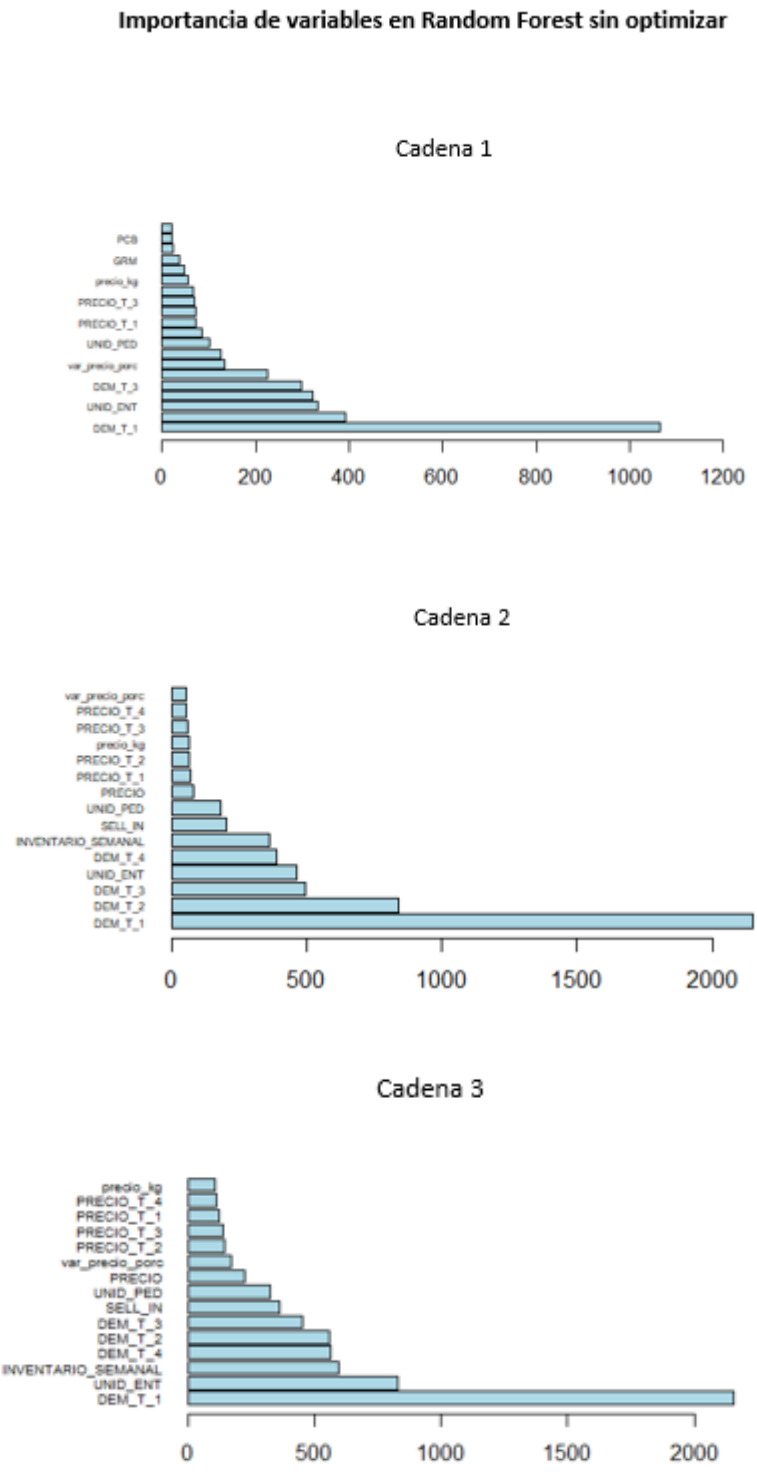


Gráfico 16: Variables más importantes del modelo Random Forest sin optimizar

En vistas de poder mejorar la calidad de las predicciones, se probaron distintas combinaciones de hiperparámetros con el fin de lograr reducir el ECM respecto del modelo

de Random Forest benchmark. No se encontraron diferencias significativas entre los modelos benchmark y aquellos cuyos hiperparámetros fueron optimizados para las tres cadenas analizadas. Resulta importante destacar que la exploración de dichos hiperparámetros fue acotada debido a la escasez de recursos computacionales disponibles – cada iteración con los modelos insumió una cantidad enorme de tiempo. Es parte del trabajo futuro explorar esta familia de modelos disponiendo de recursos computacionales adecuados con el fin de analizar en profundidad las potencialidades de Random Forest en la predicción de demanda de los productos lácteos.

La performance sobre un conjunto de validación de los modelos de Random Forest optimizados se detallan a continuación:

Tabla 8: Performance out of sample del modelo Random Forest optimizado

CADENA	ECM	R2
Cadena 1	812,33	83,90
Cadena 2	954,31	86,30
Cadena 3	1148,50	88,30

La combinación de hiperparámetros con la que se logró la mejor performance sobre los datos de validación en cada una de las cadenas fue la siguiente. Para las tres cadenas, se presenta la configuración de hiperparámetros que logró una performance lo más cercana posible a la del modelo de Random Forest sin optimizar:

Tabla 9: Hiperparámetros óptimos en Random Forest en cada cadena

CADENA	Mtry	Ntree	Maxnodes	Nodsize
Cadena 1	16	10000	7000	14
Cadena 2	16	15500	1225	10
Cadena 3	16	2000	22000	11

Por último, resulta interesante analizar cuáles fueron las variables que los modelos de Random Forest optimizados identificaron como las más relevantes a la hora de estimar la demanda. Las mismas se pueden observar en el Gráfico 17:

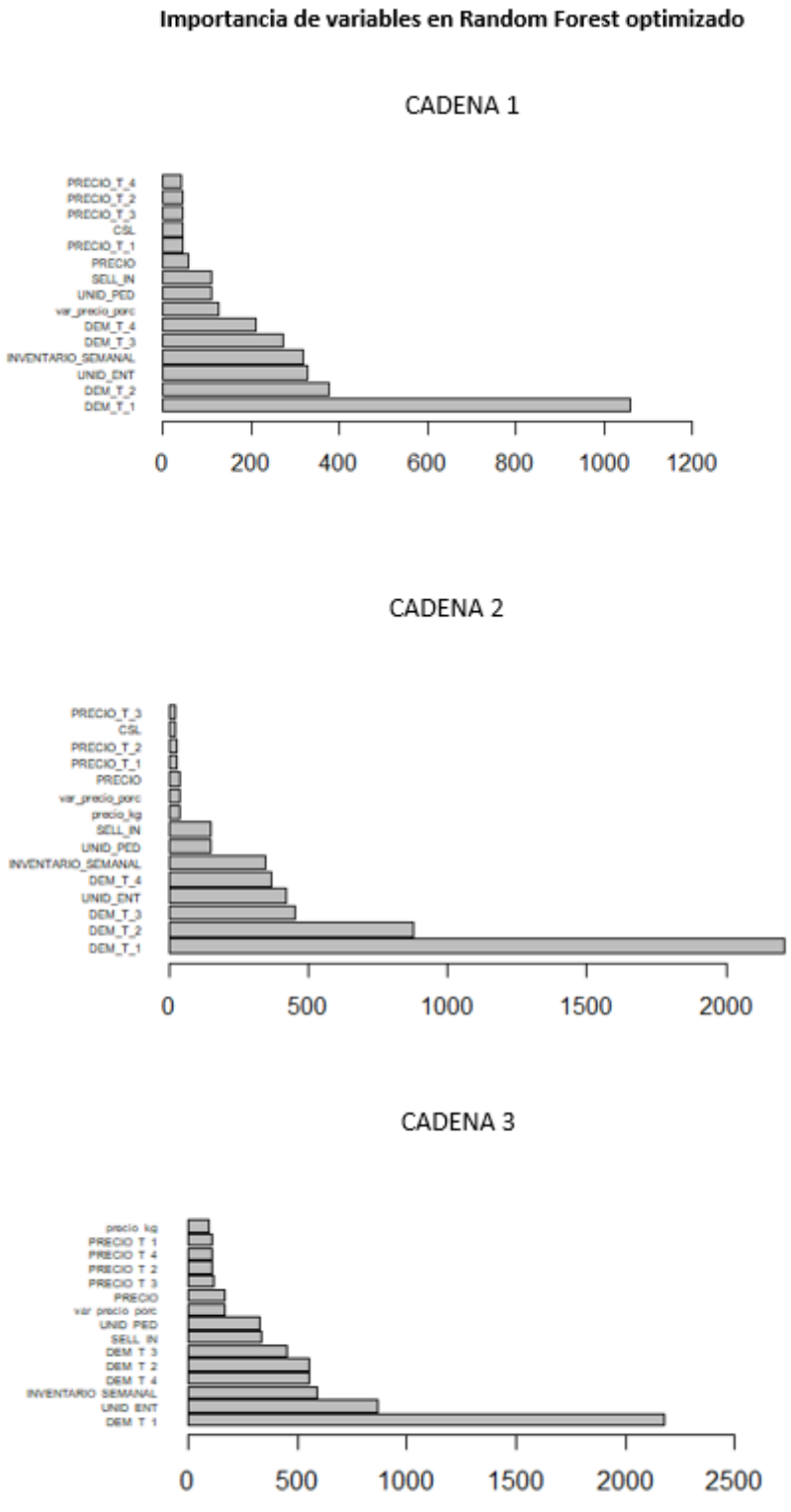


Gráfico 17: Variables más importantes del modelo Random Forest optimizado

3.3 Modelo XGBoost

Por último, se corrió un modelo de XGBoost en vistas de seguir mejorando la calidad de las predicciones. Al igual que la metodología implementada en Random Forest, primero se entrenó un modelo benchmark con los hiperparámetros seteados por default y luego, se fueron calibrando los mismos.

Los resultados sobre un conjunto de validación obtenidos en las tres cadenas en el primer modelo benchmark XGBoost se encuentran resumidos en la Tabla 10:

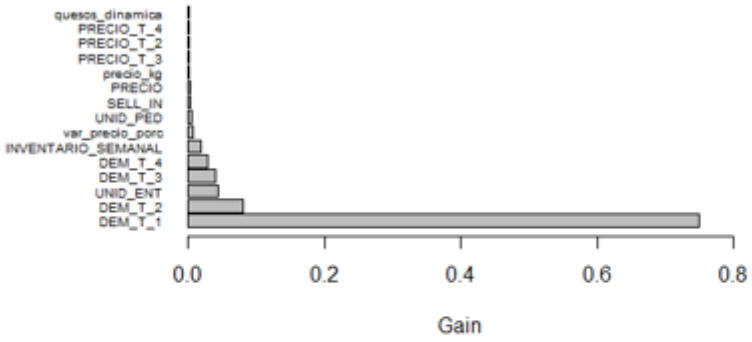
Tabla 10: Performance out of sample del modelo XGBoost sin optimizar

CADENA	ECM	R2
Cadena 1	761,34	84,83
Cadena 2	811,94	88,87
Cadena 3	1132,56	88,46

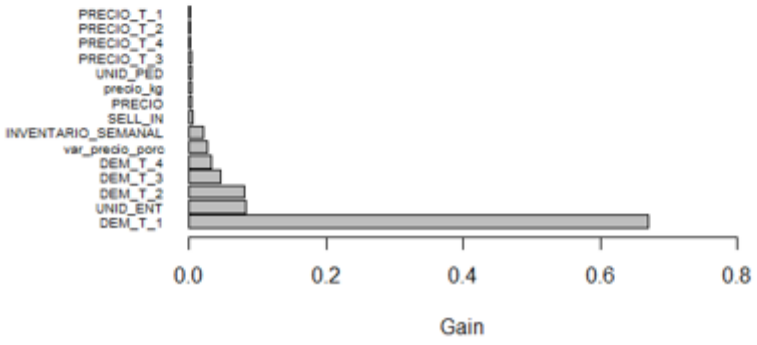
Al igual que lo que ocurrió con Random Forest, los atributos que el modelo identificó como los más significativos fueron muy similares para las tres cadenas. El Gráfico 18 muestra los 15 atributos que más contribuyeron a reducir el error en este primer modelo XGBoost sin optimizar hiperparámetros. Se puede observar que, al igual que lo que ocurría en Random Forest, la variable que más contribuye a reducir el error es la demanda de la semana anterior seguida por las unidades entregadas, y, en tercer lugar, por las demandas de las semanas pasadas. A su vez, pareciera ser importante el impacto que tienen las variables históricas tanto de precio como de demanda sobre la demanda. Cabe destacar también que el precio no sólo aparece en su versión histórica sino también en términos unitarios, en precio por kilo y en la variación porcentual entre semanas. Otro aspecto importante a tener en cuenta es el rol protagónico que juega la demanda de la semana anterior, en las tres cadenas analizadas, a la hora de explicar la demanda actual lo cual estaría mostrando que la variable en cuestión se encuentra fuertemente influenciada por sus rezagos. Por último, en el caso de la Cadena 1, resulta interesante la aparición de la variable quesos_dinamica (que es una dummy que da cuenta si el SKU está dentro de la categoría de quesos y a su vez, tuvo un descuento de precio superior al 5% respecto a la semana anterior) dentro de las más significativas sobre todo si se tiene en cuenta la fuerte influencia sobre la demanda que había identificado el modelo de regresión lineal inicial.

Importancia de variables en XGBoost sin optimizar

CADENA 1



CADENA 2



CADENA 3

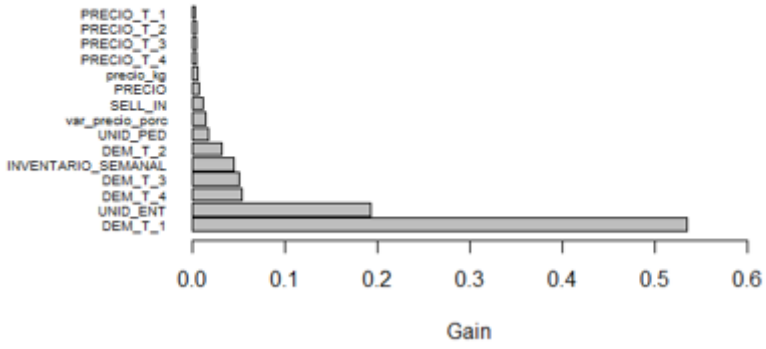


Gráfico 18: Variables más importantes en XGBoost sin optimizar

Sin embargo, en vistas de poder mejorar aún más la calidad de las predicciones, se procedió a calibrar hiperparámetros. Como se mencionó en la sección 2.5 la técnica empleada fue la de ir probando una grilla de valores y terminar optando por aquella configuración que arrojara el mejor ECM. Finalmente, la combinación con la que se logró la mejor performance sobre los datos de validación en cada una de las cadenas fue la siguiente:

Tabla 11: Hiperparámetros óptimos en XGBoost en cada cadena

CADENA	Nrounds	Max_depth	Eta	Gamma	Colsample_bytree	Min_child_weight	Subsample
Cadena 1	595	10	0,04	0,18	0,63	1,67	0,50
Cadena 2	304	13	0,02	0,18	0,66	1,36	0,71
Cadena 3	199	12	0,06	0,24	0,54	1,95	0,87

La Tabla 12 muestra los resultados sobre un conjunto de validación que se obtuvieron a partir de optimizar los hiperparámetros de XGBoost:

Tabla 12: Performance out of sample del modelo XGBoost optimizado

CADENA	ECM	R2
Cadena 1	724,76	85,56
Cadena 2	811,94	89,23
Cadena 3	1095,06	88,84

Las variables más importantes identificadas por el modelo XGBoost optimizado se presentan en el Gráfico 19. Se puede observar que los resultados no son tan distintos respecto a los del XGBoost sin optimizar. Nuevamente las demandas de los períodos anteriores son protagonistas a la hora de explicar la demanda. Por su parte, las unidades entregadas también se destacan en este rol. En una segunda instancia, las variables vinculadas al precio también fueron seleccionadas por este modelo como relevantes para explicar la demanda.

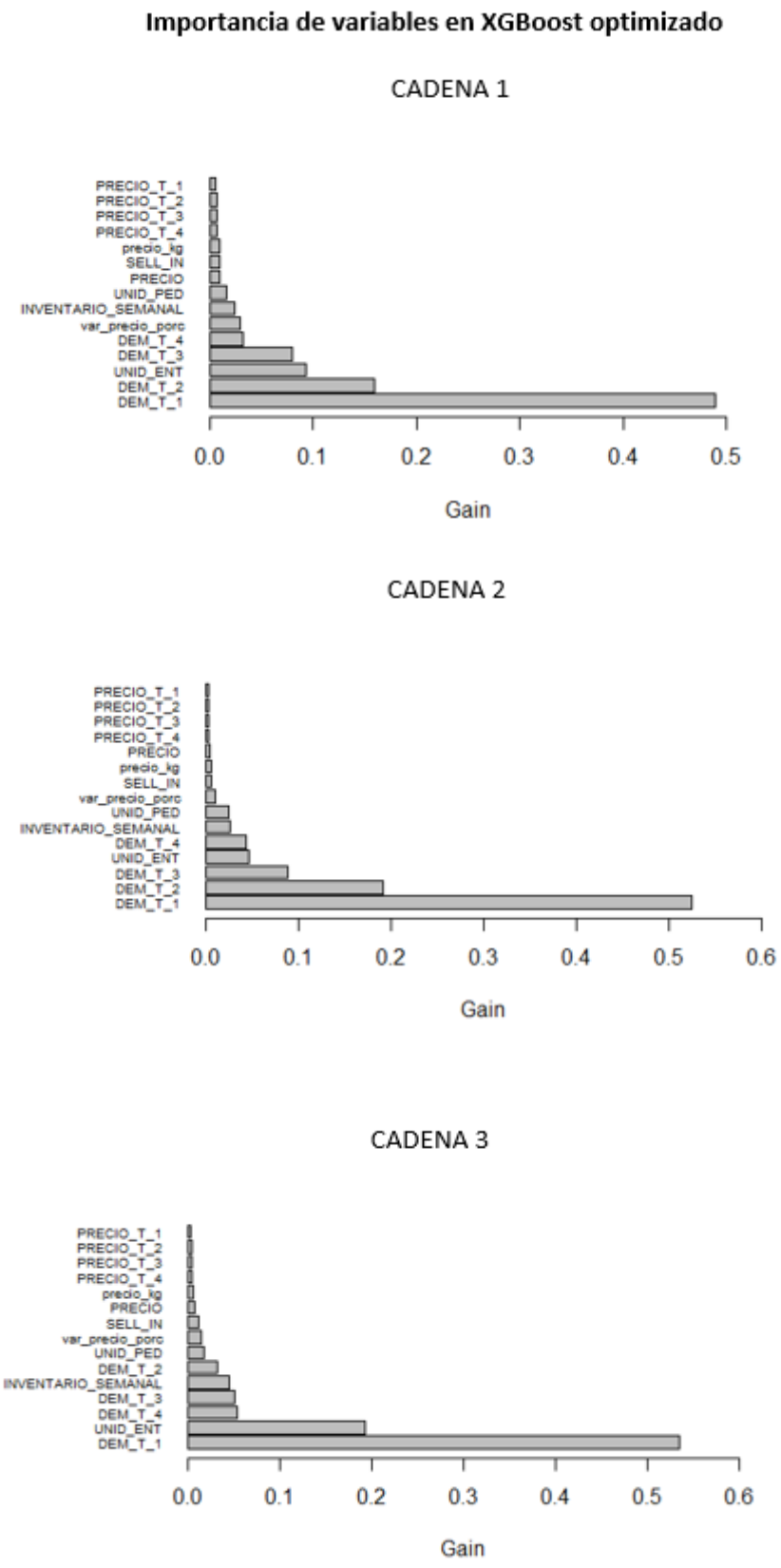


Gráfico 19: Variables más importantes en XGBoost optimizado

3.4 Comparación de resultados entre cadenas

Luego de haber entrenado distintos modelos de aprendizaje supervisado en vistas de predecir la demanda en tres cadenas de supermercados, en los tres casos se encontró que el modelo de XGBoost optimizado fue el que mejores predicciones alcanzó. Los resultados sobre un conjunto de validación se resumen en la Tabla 13:

Tabla 13: Comparación de resultados obtenidos en las tres cadenas

CADENA	LASSO		RANDOM FOREST			RANDOM FOREST OPT			XG BOOST			XG BOOST OPT		
	ECM	R2	ECM	R2	vs LASSO	ECM	R2	vs LASSO	ECM	R2	vs LASSO	ECM	R2	vs LASSO
Cadena 1	943,2	81,2	808,8	84,0	-14,3%	812,3	83,9	-13,9%	761,3	84,8	-19,3%	724,8	85,6	-23,2%
Cadena 2	1025,5	86,4	938,8	86,7	-8,5%	954,3	86,3	-6,9%	839,4	88,9	-18,2%	811,9	89,2	-20,8%
Cadena 3	1469,0	85,0	1145,6	88,3	-22,0%	1148,5	88,3	-21,8%	1132,6	88,5	-22,9%	1095,1	88,8	-25,5%

En el caso de la Cadena 1, el modelo XGBoost optimizado fue el que logró la mayor reducción del ECM (un 23,2%) mientras que una situación similar ocurre para las Cadenas 2 y 3, siendo la reducción del ECM del 20,8% y 25,5% respectivamente. Si comparamos la performance alcanzada entre las tres, se puede observar que la Cadena 3 fue la logró la máxima reducción del ECM. En términos de R^2, el valor máximo se logró en la Cadena 2 siendo del 89,2 en el modelo XGBoost optimizado.

Si consideramos la situación inicial del modelo LASSO, nuevamente fue la Cadena 2 la que alcanzó el máximo valor entre las tres cadenas en lo que respecta al R^2 (siendo del 86,4). Sin embargo, si tenemos en cuenta la eficiencia a la hora de reducir el ECM respecto del modelo LASSO, la Cadena 3 fue la que logró los mejores resultados en todos los modelos entrenados.

3.5 Aplicaciones al negocio

Luego de haber entrenado distintos modelos de aprendizaje supervisado para estimar la demanda se encontraron resultados interesantes de cara a contribuir en la toma de decisiones del negocio. Teniendo en cuenta las variables más importantes encontradas por el modelo XGBoost optimizado a la hora de predecir demanda, uno de los outputs del mismo sería enfocar acciones a garantizar en primer lugar las entregas de los productos demandados por las cadenas (principalmente como una manera de evitar los quiebres de stock). La variable unidades entregadas (UNID_ENT) no sólo resultó ser de las más significativas sobre la demanda en el mejor modelo encontrado, sino que el coeficiente estimado asociado en el modelo de regresión lineal fue positivo en las tres cadenas. Por su parte, el inventario semanal también resultó ser otra de las covariables identificadas por los modelos como relevante en la estimación de la demanda. Es decir, el hecho de que el producto esté disponible en la góndola para la compra es relevante para la variable demanda. Este fenómeno refuerza la importancia de que el proveedor de lácteos se asegure de poder entregar lo mejor posible los pedidos realizados por las cadenas para que

los productos puedan ser exhibidos y se garanticen niveles de inventarios adecuados en las tiendas, de manera de evitar quiebres de góndola y pérdidas financieras por ventas no realizadas.

Por su parte, la importante influencia de los rezagos sobre la demanda fue un tema muy presente en las tres cadenas en los modelos basados en la construcción de árboles de decisión; ya que las demandas de las semanas anteriores fueron seleccionadas por los modelos como las más significativas a la hora de reducir el error. En vistas de evitar futuras devoluciones, sería importante ir monitoreando los niveles de demanda de las semanas previas (sobre todo de la semana anterior).

El hecho de que la demanda de las próximas semanas esté en gran parte influenciada por los precios y las demandas pasadas es un buen insight a tener en cuenta a la hora de evaluar y tomar decisiones en torno al lanzamiento de un nuevo producto. Una buena forma de estimar la conveniencia o no de incluir dentro de la gama de productos de cierta tienda un lanzamiento podría ser analizar cómo fueron las ventas de algún SKU con características similares. El hecho de ya contar con esta información preliminar, contribuiría a evitar colocar productos que por sus características y las de los consumidores de la tienda no funcionen correctamente y terminen con altos niveles de devoluciones.

Por otro lado, todas las variables relacionadas con el precio (tanto en términos absolutos como en precio por kilo y variación porcentual) mostraron tener un papel influyente en la demanda. El hecho de que los precios vigentes en las semanas anteriores hayan sido destacados por el mejor modelo encontrado para estimar demanda, refuerza la importante influencia de los rezagos sobre la variable a predecir. Por su parte, la variación porcentual de los precios respecto a la semana anterior también resultó ser un factor importante a la hora de estimar demanda. De cara al lanzamiento de un producto, o a la hora de estimar qué nivel de stock habría que mantener en las tiendas cuando se va a implementar alguna dinámica de precio, sería altamente recomendable que las cadenas revisaran en qué niveles se ubicaron las demandas y los precios en las semanas anteriores. Por su parte, es importante que la empresa proveedora de lácteos tenga en cuenta la influencia del precio sobre la demanda a la hora de diseñar políticas de precios (tanto de aumentos como de planificar acciones comerciales) para evitar quiebres de stock en las góndolas y devoluciones.

En resumen, en las tres cadenas analizadas se observó un fenómeno similar caracterizado por el hecho de que la demanda resultó estar muy influenciada por todas las variables vinculadas al precio; lo cual es importante tener en cuenta a la hora de realizar aumentos de precio o acciones comerciales. En el primer caso para evitar devoluciones y en el segundo, para evitar quedarse sin stock. A su vez, en las tres cadenas analizadas la demanda resultó estar fuertemente influenciada no sólo por los precios sino también por las demandas de las semanas anteriores. Por último, las covariables relacionadas con los

niveles de inventario que hay en las góndolas, así como también la cantidad de unidades entregadas resultaron tener un papel importante a la hora de estimar demanda en las tres cadenas analizadas.

4. CONCLUSIONES

A partir del entrenamiento de diversos modelos de aprendizaje supervisado para estimar la demanda en tres cadenas de supermercados, se encontraron varios resultados interesantes para la toma de decisiones en la industria de los productos lácteos. En primer lugar, teniendo en cuenta la importancia del trabajo colaborativo entre la empresa proveedora de lácteos y las Grandes Cuentas, los resultados encontrados resultan fundamentales para poder darle visibilidad a las cadenas sobre los niveles de inventario que deberían mantener para poder afrontar correctamente la demanda, así como también, asesorarlas a la hora de generar los pedidos. A su vez, los resultados encontrados por el mejor modelo en cada caso, ayudarán a dar visibilidad respecto a qué surtido de productos es conveniente o no mantener en cada tienda/región.

Las variables que resultaron ser las más significativas a la hora de estimar demanda por los modelos entrenados fueron muy similares para las tres cadenas. Tanto las demandas como los precios de las semanas anteriores fueron las variables que los modelos identificaron como las más relevantes en la estimación de la variable objetivo. El hecho de que la demanda de las próximas semanas esté influenciada por las demandas de las semanas anteriores es un buen indicador que ayudaría en la decisión en torno a la discontinuación o lanzamiento de productos. En el caso de que algún SKU no esté performando correctamente y se esté considerando quitarlo del surtido, como paso previo a tomar la decisión, sería interesante analizar cómo fueron las demandas de las semanas anteriores (para poder entender si los bajos niveles de demanda es algo que viene ocurriendo recurrentemente y se replicará en las demandas futuras; y recién en ese caso, optar por la discontinuación). A su vez, tener presente la fuerte influencia de las demandas pasadas sobre la demanda futura puede ser de suma utilidad para que las cadenas de supermercados ajusten los pedidos de aquellos productos que vengan creciendo en volumen para evitar desabastecimiento.

Otro de los outputs más relevantes de los modelos entrenados fue la importancia que tiene el precio a la hora de influir en la demanda. Tanto el precio en las góndolas como el precio por kilo, sus variaciones porcentuales y los precios históricos fueron seleccionados por los modelos dentro de las variables más importantes. De cara al negocio, es sumamente útil tener esto en cuenta sobre todo a la hora de planificar políticas de precios – ya sea aumentos o acciones comerciales – para asegurarse de abastecer correctamente a las tiendas las semanas previas a las dinámicas y evitar quiebres de góndola. En el caso de un aumento de precio, sería recomendable que las cadenas ajusten los pedidos de sell in antes

de su implementación para evitar acumular excesos de inventarios que luego se traducirán en devoluciones (teniendo en cuenta sobre todo que las variaciones porcentuales de precio resultaron ser muy significativas en la estimación de demanda).

Otro resultado interesante que se desprende de los modelos de regresión lineal es la alta sensibilidad de las categorías de postres y quesos a las variaciones de precios. Si bien estos modelos no fueron los de mejor performance, considero que es un aspecto sumamente relevante a la hora de decidir sobre políticas de precios que involucren a estas dos categorías. Llegado el momento de implementar un aumento o baja de precio, es importante que tanto las Grandes Cuentas como el proveedor tengan en cuenta la sensibilidad de ambas categorías para asegurar una adecuada disponibilidad de productos en las góndolas – tanto para evitar quiebres cuando se realice una dinámica y evitar devoluciones cuando se realicen aumentos de precios.

En lo que respecta al abastecimiento de las tiendas, una variable fundamental para asegurar buenos niveles de inventario, y que de hecho también fue destacada por los modelos como relevante a la hora de estimar demanda, son las unidades entregadas. En vistas de evitar quiebres de stock, es sumamente importante que la empresa proveedora de lácteos asegure la correcta distribución de sus productos y tratar de mantener niveles de servicio lo más cercanos al 100% (ya que, si el producto solicitado por las Grandes Cuentas no es entregado, el mismo no se encontrará en las góndolas y, por ende, no podrá ser vendido).

Desde el punto de vista más técnico del trabajo, el modelo XGBoost luego de calibrar hiperparámetros fue el que mejor performance alcanzó en las tres cadenas analizadas. En particular, la Cadena 3 fue la que logró la máxima reducción del ECM sobre un conjunto de validación respecto al modelo benchmark (con un -25%). Sin embargo, en términos de R^2, fue la Cadena 1 la que alcanzó la máxima mejora (con un +4,4 puntos) respecto al modelo de Regresión Lineal con LASSO.

En resumen, la importancia de contar con modelos que ayuden a predecir la demanda es fundamental en vistas de asegurar buenos niveles de inventarios en las tiendas. El problema en torno a la estimación de demanda resulta más crítico aún en la industria láctea analizada debido a las características propias de sus productos. Desde el punto de vista del negocio, excesos de stock en las tiendas se traducen en devoluciones que traen aparejadas costos logísticos y financieros y deteriora la relación con los clientes. Por otro lado, la insuficiencia de stocks implica quiebres de góndola que se traduce en una mala imagen frente al consumidor final y en pérdida de ventas e ingresos tanto para el proveedor como para las cadenas de supermercados. Los modelos de aprendizaje supervisado entrenados en el presente trabajo encontraron que las demandas de las semanas anteriores son fundamentales a la hora de explicar la demanda futura. A su vez, la variable precio – en todas sus formas – también fue de las más relevantes en la estimación de la demanda. De cara al negocio los outputs encontrados por los modelos serán de utilidad a la hora de tomar

decisiones en torno a diversos aspectos que afectarán a la rentabilidad del mismo como, por ejemplo: qué niveles de inventario tratar de asegurar en las tiendas cuando hay modificaciones de precio, la conveniencia o no de mantener cierto surtido de productos según la performance de los mismos y la posibilidad de reducir costos asociados a la gestión de las devoluciones.

Dentro de los próximos pasos a seguir se encuentran seguir profundizando el trabajo colaborativo entre la empresa proveedora de lácteos y las cadenas de supermercados para poder seguir encontrando formas de mejorar los niveles de inventarios en las tiendas. A su vez, volver a entrenar estos modelos una vez que la situación pandémica se regularice en base a información que refleje de manera más fiel los patrones de consumo de los clientes. Otro aspecto a considerar, tomando como referencia el comentario de un referee anónimo, es la posibilidad de considerar como bechmark modelos ARIMA; en particular para predecir no sólo a nivel de tienda la demanda, sino como marco de referencia para hacer predicciones de largo plazo sobre las ventas a nivel de cadenas (ya que por hipótesis se espera que éstas se comporten de forma más estable que la demanda de cada tienda en particular). Estos modelos agregados pueden también servir de apoyo para tomar decisiones fundamentales sobre cómo planificar la producción a mediano y largo plazo.

Por otra parte, queda incluido también dentro de futuras acciones a realizar, seguir explorando potenciales mejoras predictivas en los modelos Random Forest que surjan de una posible mejora en la optimización de los hiperparámetros. A su vez, en caso de ser posible, complementaría muy bien al análisis introducir productos de la competencia de manera de introducir efectos de sustitución (de precios principalmente) y evaluar su impacto en la demanda. Por último, sería interesante poder extender el trabajo realizado para las Cadenas 1,2 y 3 para otros canales como distribuidores, autoservicios y mayoristas.

5. BIBLIOGRAFÍA

- Aburto, L. y Weber, R. (2007) "Improved supply chain management based on hybrid demand forecasts", *Applied Soft Computing*, N° 7, Enero 2007, pp. 134-144. Disponible en: (PDF) Improved supply chain management based on hybrid demand forecasts (researchgate.net) Última consulta: 12/5/2021.

- Bajari, P., Nekipelov, D., Ryan, S.P. y Yang, M. (2015) "Machine Learning Methods for Demand Estimation", *American Economic Review,* Vol. 105, N° 5, Mayo 2015, pp. 481-485. Disponible en: https://faculty.washington.edu/bajari/published/RYAN_manuscript.pdf Última consulta: 19/1/2021.

- Carbonneau, R., Laframboise, K. y Vahidov, R. (2007) "Application of machine learning techniques for supply chain forecasting", *European Journal of Operational Research*, Vol. 184, N° 3, Febrero 2008, pp. 1140 – 1154. Disponible en: https://www.sciencedirect.com/science/article/abs/pii/S0377221706012057 Última consulta: 12/5/2021.

- Croxton, K.L., García-Dastugue, S.J., Lambert, D.M. y Rogers, D.S. (2001) "The Supply Chain Management Processes", *The International Journal of Logistics Management*, Vol. 12, N° 2, pp. 13 – 36. Disponible en: http://ecsocman.hse.ru/data/474/089/1217/article4.pdf Última consulta: 12/5/2021.

- Cui, H., Rajagopalan, S. y Ward, A.R. (2019) "Predicting product return volume using machine learning methods", *European Journal of Operational Research*, Vol. 281, N° 3, Julio 2019, pp. 612 – 627. Disponible en: Predicting product return volume using machine learning methods - ScienceDirect Última consulta: 12/5/2021.

- Ferreira, K.J., Lee, B.H.A., Simchi-Levi,D. (2015) "Analytics for an Online Retailer: Demand Forecasting and Price Optimization", *Manifacturing & Service Operations Management*, Vol. 18, N° 1, Noviembre 2015, pp. 69-88. Disponible en: (PDF) Analytics for an Online Retailer: Demand Forecasting and Price Optimization (researchgate.net) Última consulta: 12/5/2021.

- James, G., Witten, D., Hastie, T. y Tibshirani, R. (2013) *An Introduction to Statistical Learning with Applications in R*, New York: Springer.

- Lotfi, Z., Mukhtar, M., Sahran, S. y Zadeh, A.T. (2013) "Information Sharing in Supply Chain Management", *The 4th International Conference on Electrical Engineering and*

Informatics, 2013, Vol.11, pp.298-304. Disponible en: https://www.sciencedirect.com/science/article/pii/S2212017313003484 Última consulta: 12/5/2021.

- Mollenkopf, D., Russo, I. y Frankel, R. (2007) "The returns management process in supply chain strategy", *International Journal of Physical Distribution & Logistics Management*, Agosto 2007, Vol.37, N° 7, p. 568 – 592. Disponible en: (PDF) The Returns Management Process in Supply Chain Strategy (researchgate.net) Última consulta: 12/5/2021.

- Ozhegov, E.M. y Teterina D. (2018) "Ensemble Method for Censored Demand Prediction", *International Journal of Physical Distribution & Logistics Management*, 23 Octubre 2018. Disponible en https://www.researchgate.net/publication/328445501_Ensemble_Method_for_Censored_Demand_Prediction Última consulta: 12/5/2021.

- Pavlyshenko, B.M. (2019) "Machine – Learning Models for Sales Time Series Forecasting", *Data*, 18 de Enero 2019, Vol. 4, N° 15. Disponible en https://www.researchgate.net/publication/330484523_Machine-Learning_Models_for_Sales_Time_Series_Forecasting/link/5c420d5aa6fdccd6b5b6e4a9/download Última consulta: 12/5/2021.

- Tarallo, E., Akabane, G.K., Shimabukuro, C.I., Mello, J. y Amancio, D. (2019) "Machine Learning in Predicting Demand for Fast-Moving Consumer Goods: An Exploratory Research", *IFAC – PapersOnlIne,* Enero 2019, Vol. 52, N° 13, pp. 737-742. Disponible en https://www.researchgate.net/publication/338172563_Machine_Learning_in_Predicting_Demand_for_Fast-Moving_Consumer_Goods_An_Exploratory_Research Última consulta: 12/5/2021.

- Zheng, A. (2015) *Evaluating Machine Learning Models. A Beginner´s Guide to Key Concepts and Pitfalls*, Sebastopol: O' Reilly.

- Zheng, A. y Casari, A. (2018) *Feature Engineering for Machine Learning. Principles and Techniques for Data Scientists*, Sebastopol: O' Reilly.

6. ANEXOS

6.1 Tablas

6.1.1 Variables categóricas en las bases de las Cadenas 2 y 3

Tabla 6.1.1: Variables categóricas Cadenas 2 y 3

CADENA 2		CADENA 3	
Variable	**# Categorías**	**Variable**	**# Categorías**
Tienda	205	Tienda	90
SKU	206	SKU	96
Región	5	Región	6
Provincia	4	Provincia	22
Formato	3	Formato	2
Sabor	27	Sabor	23
Marca	16	Marca	16
Envase	5	Envase	5
Familia	17	Familia	17

6.1.2 Análisis de los quiebres de góndola en la Cadena 2

Tabla 6.1.2: Análisis quiebres Cadena 2

Variable	% de quiebres	% del total de tiendas	% del total de SKU	% del sell out
Provincia: Buenos Aires	42,5%	37,1%	-	46,2%
Formato: Formato_6	50,0%	51,2%	-	34,0%
Marca: Marca_15	26,7%	-	27,7%	30,9%
Familia: Familia_5	32,5%	-	31,1%	38,9%
Sabor: Sabor_27	27,4%	-	23,8%	36,0%

6.1.3 Análisis de los quiebres de góndola en la Cadena 3

Tabla 6.1.3: Análisis quiebres Cadena 3

Variable	% de quiebres	% del total de tiendas	% del total de SKU	% del sell out
Provincia: Buenos Aires	51,7%	33,3%	-	46,9%
Formato: Formato_7	65,5%	34,4%	-	66,1%
Marca: Marca_15	40,6%	-	28,1%	42,3%
Familia: Familia_5	39,9%	-	27,1%	42,0%
Sabor: Sabor_27	26,6%	-	41,7%	27,4%

6.1.4 Coeficientes estimados del modelo de Regresión Lineal con LASSO en la Cadena 1

Variable	Coef est (+)
Sell_in	31,1291
quesos_dinamica	17,3480
Intercepto	16,1521
postres_dinamica	11,2103
MARCAMarca_13	7,6725
FAMILIAFlia_14	5,9161
descuento	5,8130
SABORSabor_18	4,9075
SABORSabor_14	4,5540
FAMILIAFlia_3	3,1109
SABORSabor_13	2,7940
MARCAMarca_6	2,6310
Dev_t_1	2,5354
MARCAMarca_14	2,2122
REGIONGBA	1,8641
yogures_aumento	1,8560
FAMILIAFlia_5	1,4570
PROVINCIANqn	1,4321
SABORSabor_19	1,2711
SABORSabor_26	1,1457
SABORSabor_17	0,9942
SABORSabor_7	0,9877
PROVINCIARioNeg	0,7720
SABORSabor_20	0,6484
MARCAMarca_8	0,5705
MARCAMarca_7	0,4987
SABORSabor_25	0,4677
REGIONSUR	0,3922
quesos	0,3440
Dem_t_1	0,3164
SABORSabor_5	0,3069
Unid_ent	0,2110
Dev_t_4	0,2177
SABORSabor_22	0,1588
Dem_t_2	0,1306
Dem_t_3	0,1200
Precio_t_2	0,1189
Dem_t_4	0,1150

Variable	Coef est (-)
precio	-0,0003
porc_dev	-0,0008
vida_util	-0,0074
SABORSabor_27	-0,0174
precio_kg	-0,0210
CSL	-0,0358
UNID_PED	-0,0377
FAMILIAFlia_16	-0,0548
PROVINCIACba	-0,0916
PROVINCIAChu	-0,1241
REGIONNOA	-0,1444
MARCAMarca_16	-0,1455
FAMILIAFlia_2	-0,1483
REGIONNEA	-0,2066
SABORSabor_15	-0,2233
Precio_t_1	-0,3057
SABORSabor_2	-0,3399
var_precio_porc	-0,3633
leches_aumento	-0,4172
PROVINCIAStaFe	-0,5234
FAMILIAFlia_15	-0,5240
REEMPLAZO	-0,6973
Dev_t_3	-0,7309
FORMATOFormato_3	-0,7553
PROVINCIASaCruz	-0,8000
PROVINCIATdelF	-0,8829
SABORSabor_21	-0,9263
FAMILIAFlia_6	-0,9275
PROVINCIALaPam	-0,9605
Dev_t_2	-1,2559
SABORSabor_9	-1,2777
MARCAMarca_2	-1,4474
SABORSabor_10	-1,4474
SABORSabor_16	-1,5656
yogures	-1,8159
SABORSabor_8	-2,0023
FAMILIAFlia_4	-2,3584
MARCAMarca_15	-2,4662

Precio_t_3	0,0974
Precio_t_4	0,0857
SABORSabor_11	0,0856
PCB	0,0554
INVENTARIO_SEMANAL	0,0479
SABORSabor_12	0,0286
GRM	0,0005

ENVASEEnvase_5	-2,7128
SABORSabor_6	-2,7737
SABORSabor_24	-2,8210
ENVASEEnvase_4	-2,8496
FORMATOFormato_2	-3,0106
MARCAMarca_5	-3,1260
FAMILIAFlia_7	-3,2810
leches_dinamica	-3,3900
FAMILIAFlia_12	-3,5520
ENVASEEnvase_2	3,6803
postres_aumento	-3,7740
SABORSabor_4	-4,6096
ENVASEEnvase_3	-7,0723
FAMILIAFlia_17	-7,8126
SABORSabor_3	-8,4268
quesos_aumento	-9,8800
DEV	-44,4028

6.1.5 Variables no seleccionadas por el modelo de Regresión Lineal con LASSO en la Cadena 1

Variables no seleccionadas por el modelo
aumento
FAMILIAFlia_10
FAMILIAFlia_11
FAMILIAFlia_13
FAMILIAFlia_8
FAMILIAFlia_9
leches
MARCAMarca_10
MARCAMarca_11
MARCAMarca_12
MARCAMarca_3
MARCAMarca_4
MARCAMarca_9
postres
PROVINCIACntes
REGIONPBA
SABORSabor_23
yogures_dinamica

6.1.6 Coeficientes estimados del modelo de Regresión Lineal con LASSO en la Cadena 2

Variable	Coef est (+)
quesos_dinamica	15,9602
Intercepto	6,8730
SABORSabor_22	5,2620
FAMILIAFlia_14	4,4410
descuento	3,5170
MARCAMarca_3	3,3953
postres_dinamica	2,5770
SABORSabor_18	1,7830
FAMILIAFlia_3	1,7109
FORMATOFormato_5	1,1672
SABORSabor_13	1,1521
SABORSabor_7	0,8290
MARCAMarca_2	0,7765
FAMILIAFlia_5	0,5379
MARCAMarca_10	0,5208
SABORSabor_27	0,5036
SABORSabor_23	0,4680
SABORSabor_26	0,4616
REGIONCYO	0,4261
Dem_t_1	0,3474
MARCAMarca_8	0,2818
FAMILIAFlia_16	0,1940
UNID_ENT	0,1863
Dem_t_2	0,1324
SABORSabor_20	0,1073
INVENTARIO_SEMANAL	0,1052
Dem_t_3	0,0994
Dem_t_4	0,0981
MARCAMarca_6	0,0304
UNID_PED	0,0091
FAMILIAFlia_11	0,0027
Precio_t_3	0,0015

Variable	Coef est (-)
precio	-0,0001
precio_t_1	-0,0001
dev_t_4	-0,0002
dev_t_2	-0,0003
dev_t_1	-0,0003
porc_dev	-0,0007
PROVINCIACba	-0,0012
precio_kg	-0,0055
PCB	-0,0196
CSL	-0,0200
VIDA_UTIL	-0,0273
MARCAMarca_14	-0,0935
var_precio_porc	-0,1766
FAMILIAFlia_4	-0,2815
SABORSabor_2	-0,3685
ENVASEEnvase_5	-0,4358
MARCAMarca_7	-0,4903
FAMILIAFlia_6	-0,5046
PROVINCIACaba	-0,5585
REGIONPBA	-0,7924
REEMPLAZO	-0,8679
SABORSabor_12	-0,8828
SABORSabor_15	-0,9126
REGIONNOA	-0,9176
FAMILIAFlia_7	-0,9345
ENVASEEnvase_4	-1,0534
FORMATOFormato_6	-1,1378
SABORSabor_6	-1,1972
MARCAMarca_11	-1,2923
postres_aumento	-1,2929
SABORSabor_21	-1,3835
SABORSabor_9	-1,5010
SABORSabor_10	-1,5382
ENVASEEnvase_3	-1,5586
MARCAMarca_16	-1,7171
SABORSabor_3	-1,8147
leches_dinamica	-1,8869
FAMILIAFlia_2	-2,0369

SABORSabor_4	-2,0876
FAMILIAFlia_12	-2,6244
SABORSabor_16	-2,9340
SABORSabor_24	-2,9784
SABORSabor_8	-4,6619
FAMILIAFlia_17	-5,0952
MARCAMarca_5	-8,3680
quesos_aumento	-9,5699
DEV	-186,8030

6.1.7 Variables no seleccionadas por el modelo de Regresión Lineal con LASSO en la Cadena 2

Variables no seleccionadas por el modelo
aumento
Dev_t_3
ENVASEEnvase_2
FAMILIAFlia_10
FAMILIAFlia_13
FAMILIAFlia_15
FAMILIAFlia_8
FAMILIAFlia_9
GRM
leches
leches_aumento
MARCAMarca_12
MARCAMarca_13
MARCAMarca_15
MARCAMarca_4
MARCAMarca_9
postres
Precio_t_2
Precio_t_4
PROVINCIAMza
quesos
REGIONGBA
REGIONSUR
SABORSabor_11
SABORSabor_14
SABORSabor_17
SABORSabor_19
SABORSabor_25

SABORSabor_5
SELL_IN
yogures
yogures_aumento
yogures_dinamica

6.1.8 Coeficientes estimados del modelo de Regresión Lineal con LASSO en la Cadena 3

Variable	Coef est (+)
quesos_dinamica	26,8484
SELL_IN	9,4299
Intercepto	8,9770
postres_dinamica	8,4998
descuento	6,9426
FAMIILIAFlia_14	4,1745
FAMILIAFlia_3	3,1890
MARCAMarca_4	3,0617
PROVINCIACatam	3,0450
leches_aumento	2,3713
SABORSabor_22	2,3080
SABORSabor_23	2,2410
PROVINCIALaRioja	2,0328
SABORSabor_17	1,7387
postres_aumento	1,5935
PROVINCIACaba	1,5470
PROVINCIAJuj	1,5429
PROVINCIANqn	1,3580
quesos	1,1506
REGIONGBA	1,0986
SABORSabor_5	0,9022
PROVINCIARioNeg	0,7901
PROVINCIASalta	0,7248
MARCAMarca_7	0,6800
SABORSabor_27	0,5900
SABORSabor_13	0,5831
REGIONPBA	0,5817
Dem_t_1	0,3410
UNID_ENT	0,2471
MARCAMarca_11	0,1966
Precio_t_3	0,1690
PROVINCIAMza	0,1637

Variable	Coef est (-)
Dev_t_4	-0,0003
porc_dev	-0,0004
PROVINCIASanJuan	-0,0086
VIDA_UTIL	-0,0086
precio_kg	-0,0148
CSL	-0,0320
REGIONSUR	-0,0345
FAMILIAFlia_4	-0,0570
Precio_t_1	-0,1089
Precio	-0,1933
FAMILIAFlia_11	-0,2130
var_precio_porc	-0,2652
SABORSabor_20	-0,4275
PROVINCIAMnes	-0,4575
PROVINCIATuc	-0,4687
SABORSabor_19	-0,4714
SABORSabor_26	-0,6965
REGIONNEA	-0,7030
PROVINCIAChu	-0,7055
PROVINCIACba	-0,7637
yogures	-0,7918
SABORSabor_24	-0,8633
PROVINCIASanLuis	-0,9107
MARCAMarca_3	-0,9532
PROVINCIACntes	-1,0034
PROVINCIALaPam	-1,0932
PROVINCIAStaFe	-1,1475
MARCAMarca_16	-1,3360
FAMILIAFlia_7	-1,3420
SABORSabor_9	-1,5410
FAMILIAFlia_2	-1,6130
MARCAMarca_15	-1,6260

Precio_t_2	0,1442
Dem_t_4	0,1201
REEMPLAZO	0,1156
INVENTARIO_SEMANAL	0,1104
Dem_t_3	0,0764
Dem_t_2	0,0530
PCB	0,0503
PROVINCIAEntRios	0,0040
SABORSabor_12	0,0044
GRM	0,0015
Dev_t_2	0,0004
Dev_t_3	0,0002
UNID_PED	0,0001
Precio_t_4	0,0001
Dev_t_1	0,0000

SABORSabor_16	-1,6670
SABORSabor_6	-1,7948
FAMILIAFlia_12	-1,8707
MARCAMarca_2	-1,8795
PROVINCIAChaco	-1,9767
MARCAMarca_9	-2,1542
SABORSabor_15	-2,1547
ENVASEEnvase_2	-2,6368
ENVASEEnvase_5	-2,8235
FORMATOFormato_8	-3,7239
SABORSabor_10	-3,7691
SABORSabor_8	-3,9375
FAMILIAFlia_17	-8,0434
ENVASEEnvase_3	-8,2770
SABORSabor_3	-8,9330
leches_dinamica	-9,0220
SABORSabor_25	-19,9302
quesos_aumento	-30,1667
DEV	-270,9317

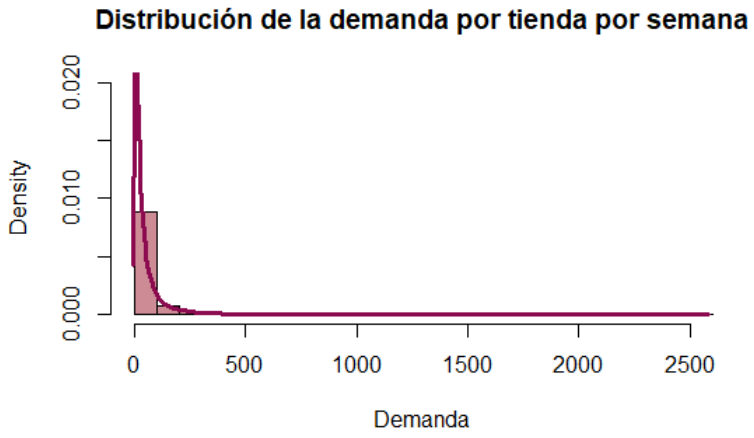
6.1.9 Variables no seleccionadas por el modelo de Regresión Lineal con LASSO en la Cadena 3

Variables no seleccionadas por el modelo
aumento
ENVASEEnvase_4
FAMILIAFlia_10
FAMILIAFlia_13
FAMILIAFlia_15
FAMILIAFlia_16
FAMILIAFlia_5
FAMILIAFlia_6
FAMILIAFlia_8
FAMILIAFlia_9
leches
MARCAMarca_10
MARCAMarca_12
MARCAMarca_13
MARCAMarca_14
MARCAMarca_5
MARCAMarca_6
MARCAMarca_8

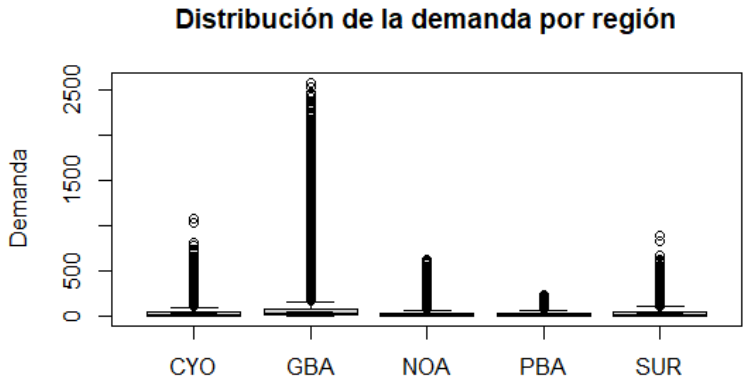
postres
PROVINCIAForm
PROVINCIASdelE
REGIONCYO
REGIONNOA
SABORSabor_11
SABORSabor_2
SABORSabor_21
yogures_aumento
yogures_dinamica

6.2 Gráficos

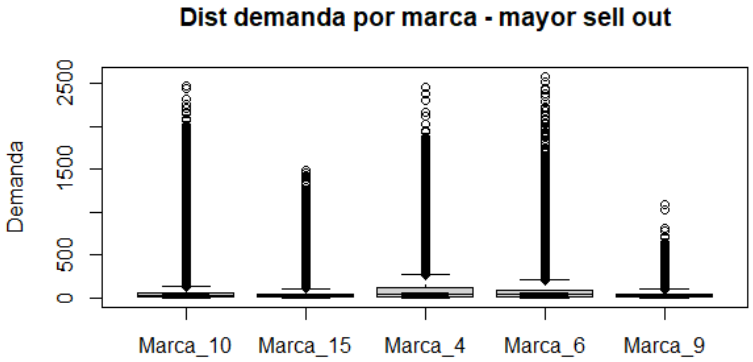
6.2.1 Distribución de la demanda por semana en la Cadena 2



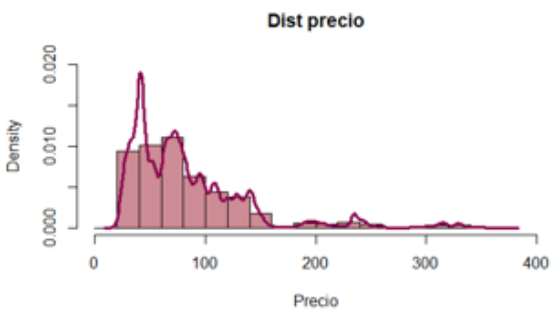
6.2.2 Distribución de la demanda por región en la Cadena 2

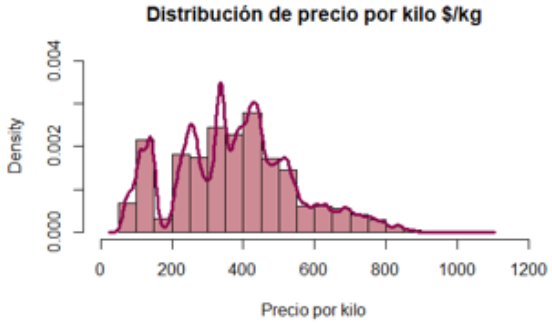


6.2.3 Distribución de la demanda por marca con mayor sell out de la Cadena 2

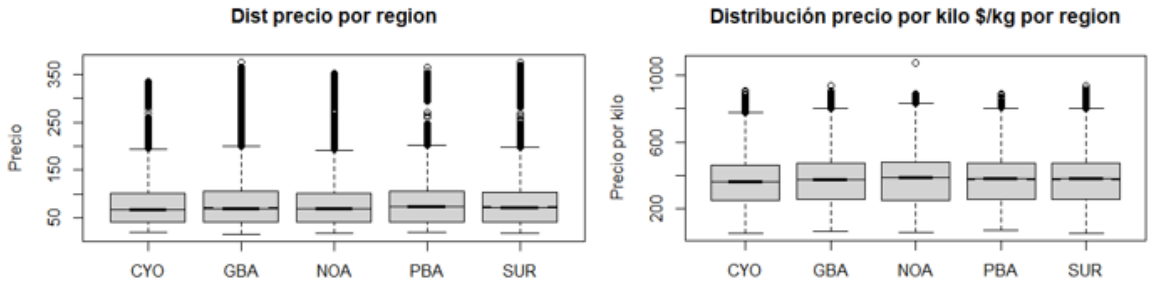


6.2.4 Distribución del precio y del precio por kilo en la Cadena 2

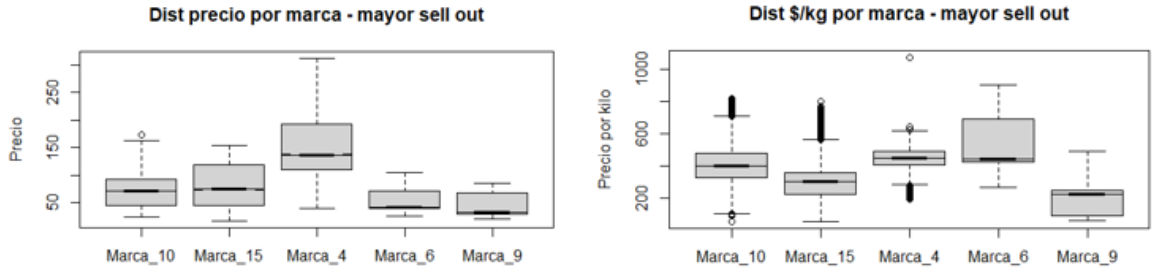




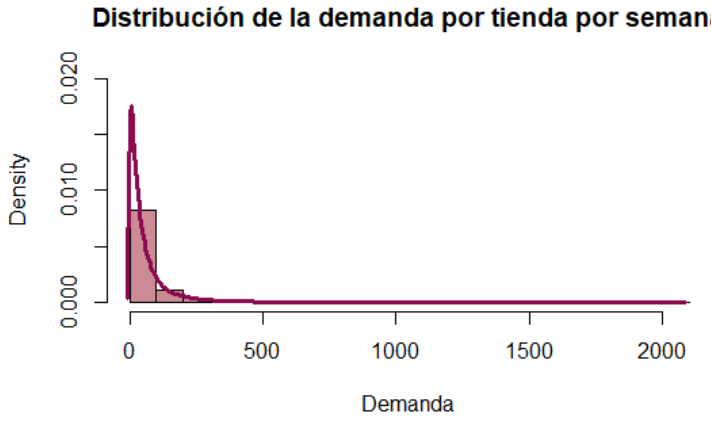
6.2.5 Distribución del precio y precio por kilo según la región en la Cadena 2



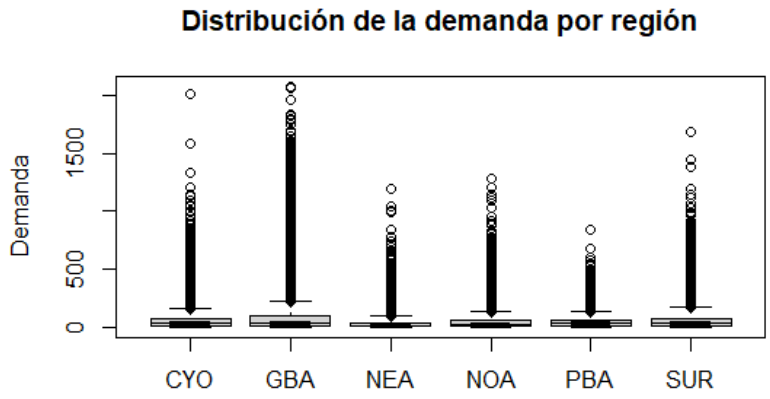
6.2.6 Distribución del precio y del precio por kilo en las 5 marcas con mayor sell out en la Cadena 2



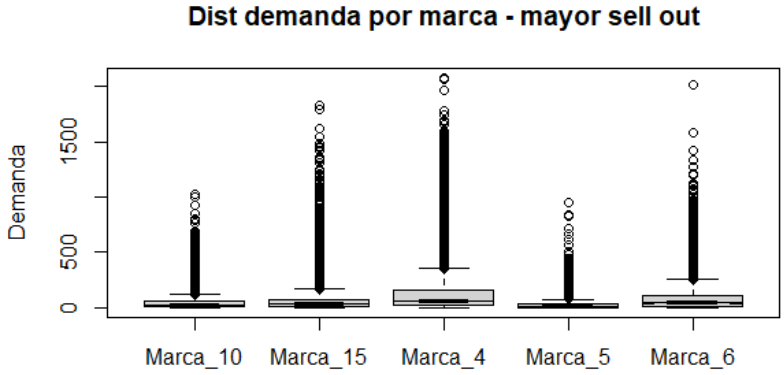
6.2.7 Distribución de la demanda semanal en la Cadena 3



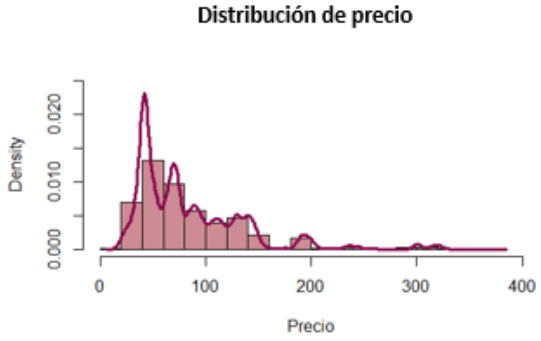
6.2.8 Distribución de la demanda por región en la Cadena 3

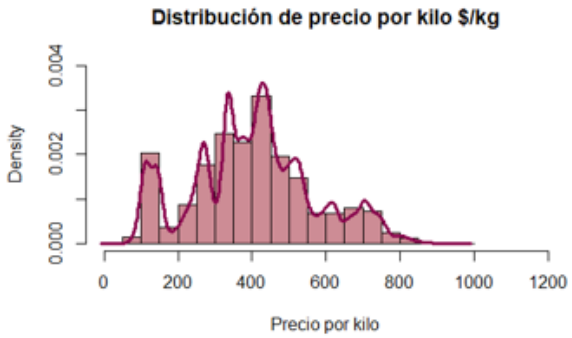


6.2.9 Distribución de la demanda en las 5 marcas con mayor sell out en la Cadena 3

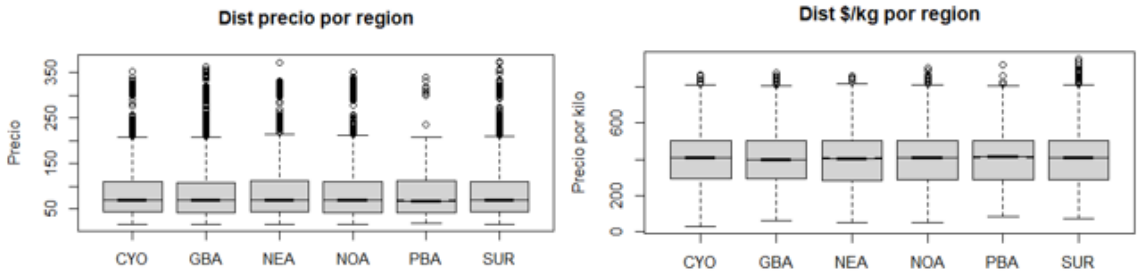


6.2.10 Distribución del precio y del precio por kilo en la Cadena 3

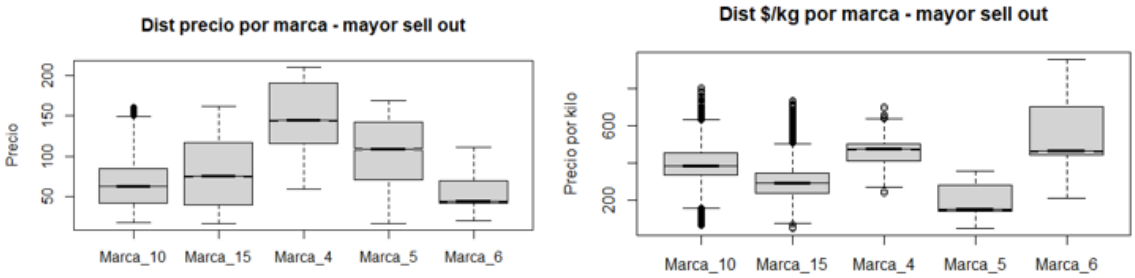




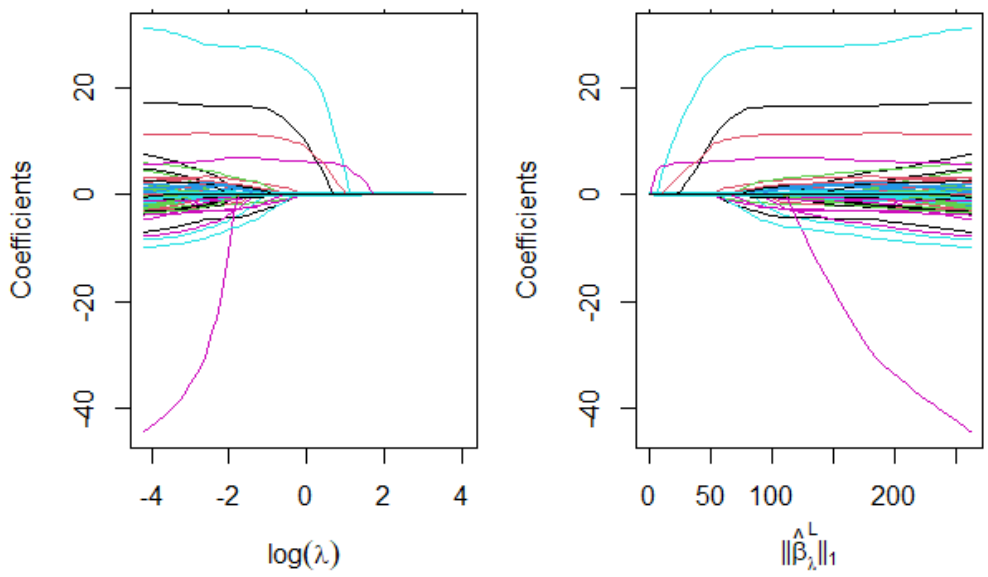
6.2.11. Distribución del precio y del precio por kilo por región en la Cadena 3



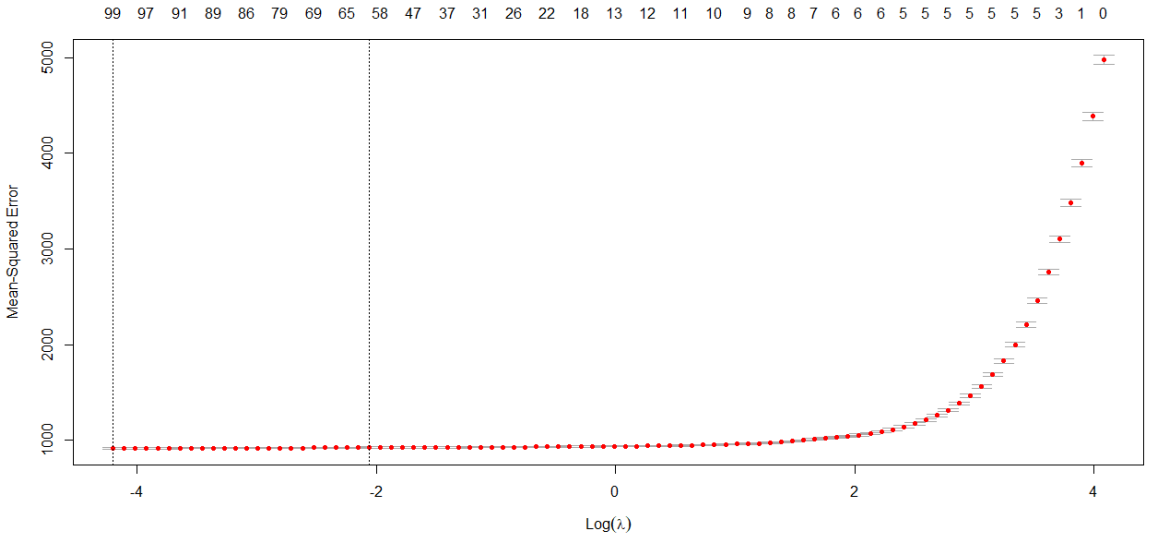
6.2.12 Distribución del precio y del precio por kilo en las 5 marcas con mayor sell out en la Cadena 3



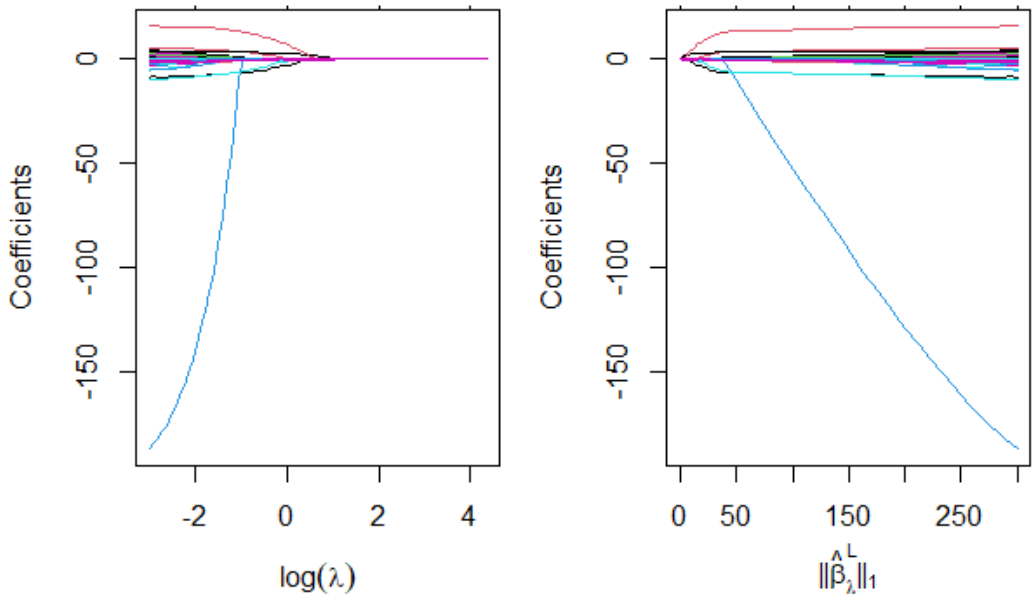
6.2.13 Evolución de los coeficientes de la Regresión Lineal con regularización LASSO en función del valor de λ en la Cadena 1



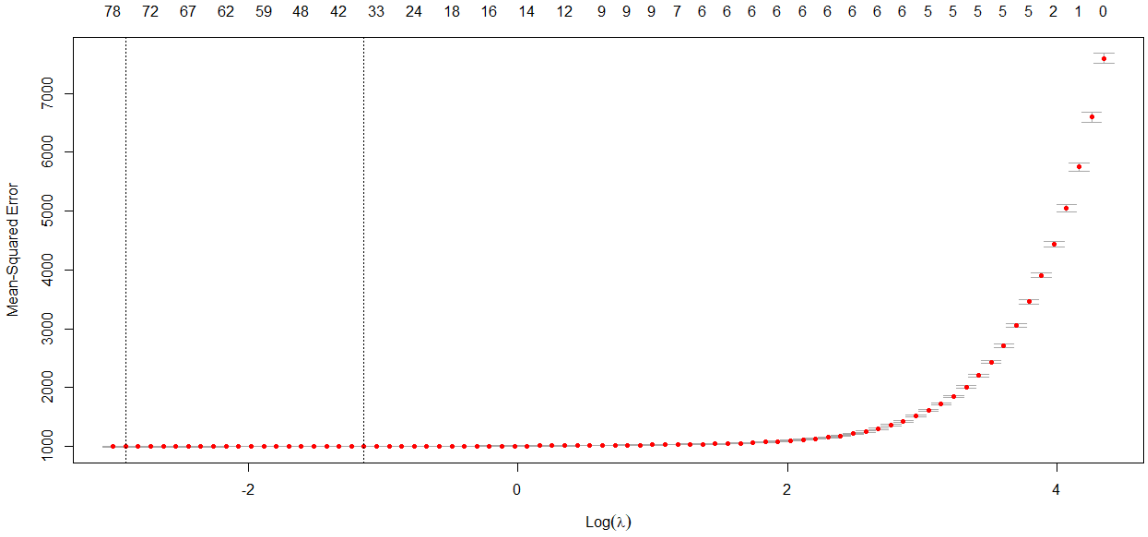
6.2.14 Selección de variables por validación cruzada en el modelo de Regresión Lineal con regularización LASSO en la Cadena 1



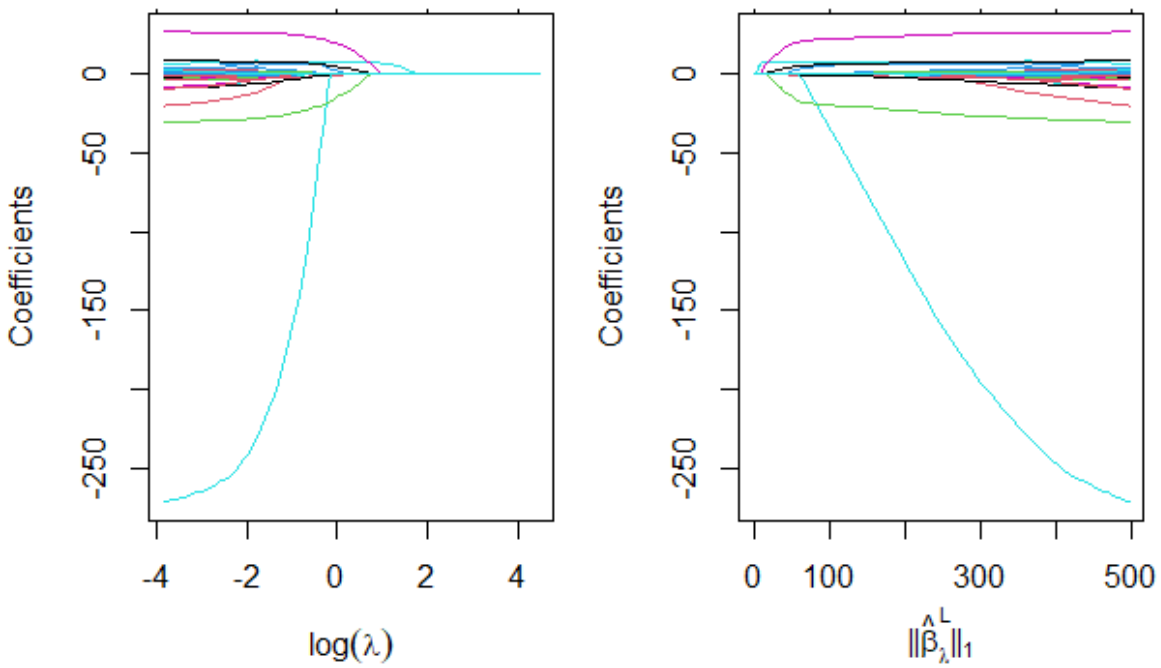
6.2.15 Evolución de los coeficientes de la Regresión Lineal con regularización LASSO en función del valor de λ en la Cadena 2



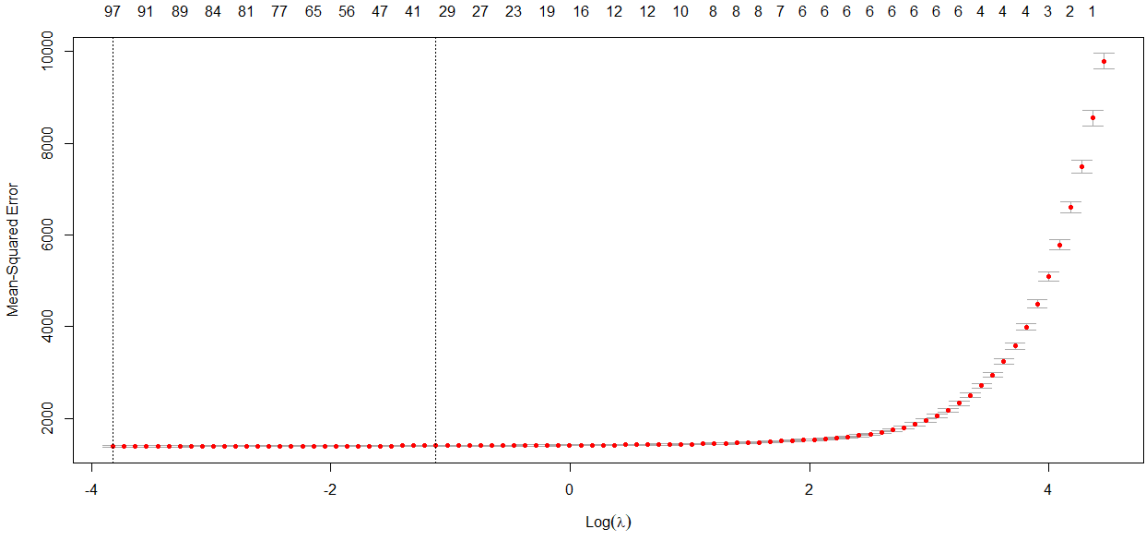
6.2.16 Selección de variables por validación cruzada en el modelo de Regresión Lineal con regularización LASSO en la Cadena 2



6.2.17 Evolución de los coeficientes de la Regresión Lineal con regularización LASSO en función del valor de λ en la Cadena 3



6.2.18 Selección de variables por validación cruzada en el modelo de Regresión Lineal con regularización LASSO en la Cadena 3



6.3 Validación del modelo que computa la demanda faltante

En vistas de poder evaluar la calidad predictiva del modelo utilizado para computar la demanda faltante, se dividió al conjunto de entrenamiento en dos grupos: el 80% se lo apartó para entrenar el modelo y el 20% restante se lo utilizó como conjunto de validación. En las tres cadenas, se empleó el mismo valor de λ que se utilizó para computar la demanda faltante.

A continuación, se detallan los resultados encontrados para las tres cadenas:

Tabla 6.3.1: Performance del modelo empleado para computar demanda faltante

CADENA	ECM	R2	LAMBDA
Cadena 1	941,88	81,27	0,015
Cadena 2	1077,37	86,15	0,050
Cadena 3	1412,47	85,41	0,024

6.4 Aplicación de los modelos en SKU´s con mayor sell out

El objetivo de esta sección es poder realizar predicciones individualizadas. En particular, para cada cadena se seleccionaron los dos SKU´s con mayor participación en el sell out. Siguiendo la misma metodología empleada para la realización de esta tesis, se realizó una primera estimación con un modelo de Regresión Lineal como punto de partida. Luego, se aplicó el modelo XBoost optimizado encontrado y se compararon las métricas para evaluar la calidad predictiva de los modelos:

Tabla 6.4.1: Predicciones a nivel SKU en la Cadena 1

CADENA 1	LASSO		XG BOOST OPT		
	ECM	R2	ECM	R2	vs LASSO
13082	10.809	73,2	9.731	75,9	-10,0%
13801	7.225	70,2	6.056	75,1	-16,2%
TOTAL CADENA 1	**943,2**	**81,2**	**724,8**	**85,6**	**-23,2%**

Tabla 6.4.2: Predicciones a nivel SKU en la Cadena 2

CADENA 2	LASSO		XG BOOST OPT		
	ECM	R2	ECM	R2	vs LASSO
11340	2.884	94,0	2.516	94,7	-12,8%
11320	3.228	93,0	2.802	93,9	-13,2%
TOTAL CADENA 2	**1.025,5**	**86,4**	**811,9**	**89,2**	**-20,8%**

Tabla 6.4.3: Predicciones a nivel SKU en la Cadena 3

CADENA 3	LASSO		XG BOOST OPT		
	ECM	R2	ECM	R2	vs LASSO
6508	4.355	85,1	4.185	85,7	-3,9%
6509	945	84,9	921	85,3	-2,6%
TOTAL CADENA 3	**1.469,0**	**85,0**	**1.095,1**	**88,8**	**-25,5%**

El hecho de que, a nivel individual, las ganancias en términos de reducción del error no resultaran tan significativas como a nivel agregado de la cadena, se deba a que, para poder realizar predicciones individualizadas, los modelos se entrenaron con una menor cantidad de observaciones y esto influye en la calidad de las predicciones. Ya que, cuanto mayor sea la cantidad de datos con las que el modelo cuente para poder aprender, mejores serán las estimaciones.