UNIVERSIDAD DE CHILE
FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS
DEPARTAMENTO DE INGENIERÍA INDUSTRIAL

ESTIMACIÓN DEL PRECIO DE PRENDAS DE SEGUNDA MANO CON ALGORITMOS DE MACHINE LEARNING

TESIS PARA OPTAR AL GRADO DE MAGÍSTER EN GESTIÓN DE OPERACIONES
MEMORIA PARA OPTAR AL TÍTULO DE INGENIERO CIVIL INDUSTRIAL

GERMÁN NICOLÁS SILVA ARANDA

PROFESOR GUÍA:
CHARLES THRAVES CORTÉS-MONROY

MIEMBROS DE LA COMISIÓN:
MARCEL GOIC FIGUEROA
ANDRÉS MUSALEM SAID

SANTIAGO DE CHILE
2024

ESTIMACIÓN DEL PRECIO DE PRENDAS DE SEGUNDA MANO CON ALGORITMOS DE MACHINE LEARNING

Este trabajo propone una metodología para estimar el precio de venta de una tienda dedicada a vender prendas de lujo de segunda mano, quienes actualmente determinan de manera manual los precios para cada nueva prenda que ponen a la venta. El objetivo es proponer un modelo de estimación de precios de venta para prendas de vestir de segunda mano utilizando algoritmos de *Machine Learning*. Con este fin, se utilizan caracterísicas de la prenda: marca, categoría, sub categoría, estado, composición y si pertenece al catálogo de productos más accesibles. Debido a que algunas prendas no cuentan el atributo *precio de referencia*, se hace la distinción entre dos modelos predictivos: aquellos que usan esta variable, y aquellos que no. Luego, para cuantificar los errores, se compara el precio de venta estimado con el precio de venta usado por la empresa.

Los modelos que no utilizan el *precio de referencia* como input tienen un *MAPE* entre el 20 % y el 30 % en el set de testeo para cada una de las categorías de prenda, donde los métodos con mejor desempeño son *Linear Regression* y *Support Vector Machine*. Asimismo, se obtiene un *MAPE* entre 10 % y el 18 % para los modelos que sí utilizan el *precio de referencia*, donde los métodos destacados son *Random Forest* y *XGBoost*. Debido a la alta variabilidad de los datos, también se encontraron intervalos de confianza de los precios predichos para generar recomendaciones de niveles de precio para las nuevas prendas. Además, se abordó el impacto en las utilidades de la empresa al utilizar estos modelos para la estimación de precios: cuando se utiliza *precio de referencia* las utilidades bajan alrededor de 5 % mientras que cuando no, bajan alrededor de un 17 %. Esto viene a reforzar el hecho de que contar con este precio de referencia al ayudar enormemente a las predicciones también tiene un impacto en las utilidades que derivan de una buena puesta de precios.

Se concluye que los modelos propuestos son capaces de estimar los *precios de venta* de las prendas de segunda mano con un error aceptable, y que el uso de *precio de referencia* como input es una variable que mejora de forma significativa el desempeño de los modelos.

Para María Isabel, Olga, Germán y Juan.

Agradecimientos

A mi familia, que siempre me ha apoyado y ha sido un pilar fundamental en mi vida. En especial a mi madre que ha hecho todo lo posible para que mis hermanos y yo tengamos la mejor vida posible. A mis cuatro abuelos, María Isabel (Tati) que me ha hecho sentir su cariño y amor incondicional desde pequeño, a Olga que siempre quiere lo mejor para sus nietos, a Germán que me dio valores que van conmigo a todas partes y a Juan que es una inspiración y ejemplo para mí.

Quiero agradecer de corazón a personas que me ha regalado la vida dentro de la Universidad. En primer lugar, a mi novia Sofía que me anima a seguir adelante y me apoya en todo momento, con ella he compartido momentos que han servido de motivación para seguir adelante. A mis amigos: Eduardo, Diego, Félix, Francisca, Martín, Nicolás, Pedro y Vicente(s), de los cuales siempre estaré agradecido.

Mi gratitud hacia el profesor Charles Thraves por su paciencia y generosidad al responder cada una de mis inquietudes, siendo un apoyo muy importante en el desarrollo de la tesis. Fue un gusto trabajar con él.

Agradezco también a la Universidad de Chile por darme la oportunidad de estudiar en la mejor de Chile, y a la Facultad de Ciencias Físicas y Matemáticas por la calidad de la formación que recibí. Por último a una persona cuyo recuerdo me acompañará toda la vida: mi tío Felipe Álvarez, que esta era su casa y le hubiera encantado verme estudiar en la Escuela y recibir este título. Hoy lo ve mi tía Claudia, mi madrina, de la que estaré eternamente agradecido por su cariño y apoyo.

Tabla de Contenido

Índice de Tablas

Índice de Ilustraciones

Capítulo 1

Introducción

En los mercados de productos de segunda mano no es fácil poner precios cuando se tiene una gran variedad de ítems donde practicamente cada unidad es diferente al resto. Se debe procurar ser competitivo con los productos nuevos y realmente ofrecer descuentos atractivos para que compradores prefieran una prenda de segunda mano, en comparación a sus alternativas en el comercio tradicional.

Este es el caso de Market People, empresa que nace bajo el alero de las socias Stephanie Truan y María Ignacia Cartoni, cuyo negocio consiste en recibir prendas en consignación de personas dispuestas a desocupar parte de su closet y en convertir estas prendas -mediante una cuidadosa selección, toma de fotografias y una posterior publicación- en productos para que otras personas tengan la opción de comprarlas ya sea mediante su sitio web (https://www.emarketpeople.com) o visitando su tienda ubicada en Santiago de Chile. A cambio, por generar toda esta plataforma que permite que este nexo exista, la compañía gana una porción de la venta como comisión.
Market People busca transmitir una identidad clara a sus clientes y es la de vender lujo accesible. De esta forma, la compañía tiene un ticket promedio alto al vender por lo general ropa muy costosa, aunque no por eso dejan de aceptar prendas de precios más bajos siempre que vayan alineadas con el estándar de la empresa.

El proceso entre que se recibe en consignación la prenda y se pone a la venta no es menor e intervienen en este diversas personas dentro de la empresa. A grandes rasgos primero se

recibe un formulario con las prendas que el cliente quiere consignar, luego se realiza una selección de las prendas a recibir, posteriormente estas prendas se revisan en las oficinas de Market People, se les toman fotografías con modelos y finalmente se llega a una etapa crítica del negocio: fijarles un precio. Después de publicada, la prenda en muchos casos posee dos precios asociados: el *precio de venta*, naturalmente, y un *precio de referencia* o *precio de retail* (ambos nombres, que representan lo mismo, se utilizarán indistintamente a lo largo de la tesis), que es el valor del producto nuevo en alguna tienda del retail. Actualmente, la búsqueda del *precio de referencia* la hace parte del equipo de la empresa y deben buscar prenda por prenda en múltiples sitios de internet y ver cuál es el precio que tiene cada prenda nueva si es que este precio se encuentra, lo que no siempre ocurre: puede que sea una prenda de diseñador o de una producción muy acotada, no logrando asignarle así este precio del producto 'nuevo'. Por otra parte, poner el *precio de venta* es un proceso más sutil, pues requiere juicio experto y por lo mismo es María Ignacia, una de las cofundadores, la que debe analizar cada una de las prendas y determinar el *precio de venta* que le corresponderá de cara al público. El costo de oportunidad incurrido al tener a una de las cofundadores poniendo precios de manera manual es enorme y no es escalable, teniendo en cuenta que una de las ambiciones de Market People es constantemente ir creciendo en tamaño, volumen de ventas y variedad de productos que ofrecen.

De esta manera, se propone hacer más eficiente la puesta de *precios de venta* para Market People y generar sugerencias de precio que sirvan de orientación para la decisión final mediante la utilización de datos históricos de ventas de la compañía. Estos *precios de venta* se estimarán utilizando métodos de *Machine Learning*, los cuales recibirán como *input* características (desde ahora en adelante *atributos*) de la prenda en cuestión. Así, los modelos serán entrenados utilizando datos que fueron inputados en base al juicio experto de María Ignacia y tomarán en consideración factores como la marca, la sub categoría, el estado en el que se encuentra la prenda, entre otros atributos. La utilización de este tipo de métodos posee una serie de ventajas para la compañía, dentro de las que destaca capturar tendencias de precios históricos que pueden no saltar a la vista a un ser humano, procesar de manera rápida y eficiente altos volúmenes de datos que de forma manual tomarían muchas horas hombre, poner precios de manera consistente a lo largo del tiempo y así no generar desajustes en las expectativas de precio de clientes recurrentes, además de generar escalabilidad de cara a un futuro donde la

tienda aumente tanto su número de prendas como su variedad de tipos de prendas.

Alternativamente, se podría pensar en un enfoque donde lo que se busque sea orientar a la empresa a poner precios "óptimos", donde tradicionalmente lo que se realiza es estimar una curva de demanda y luego maximizar el precio. Este enfoque si bien surge de manera natural al considerar un problema de fijación de precios, no es el objetivo de esta tesis por una serie de razones. Primero, la empresa considera que la fijación de precios no ha sido un problema. Se podría argumentar que por más que la empresa considere que está haciendo un buen trabajo al fijar los precios no necesariamente está en el óptimo. Esto puede ser cierto, sin embargo lo que el enfoque alternativo omite es el hecho de que cada producto como combinación de atributos es dificilmente comparable a otro producto, aún cuando este producto pueda tener la misma combinación de atributos, debido a que su precio varía enormemente por diversos factores difícilmente capturables que además no se encuentran en la base de datos utilizada, como puede ser el diseño. Por lo anterior, este enfoque alternativo es un problema completamente distinto al que busca realizar esta tesis, para el habría que tomar una serie de supuestos no menores para resolver un problema que desde la compañía no consideran un problema.

1.1. Estado del Arte

La estimación de precios es un tema que ha sido estudiado ampliamente en la literatura. Hasta hace unas décadas los trabajos relacionados eran principalmente teóricos, con modelos que se basaban en la teoría económica para explicar el comportamiento de los precios de los bienes. Sin embargo, con el avance de la tecnología y la masificación de internet, se ha hecho posible el uso de grandes cantidades de datos para la estimación de precios. Principalmente en los últimos 20 años es que se encuentra un número importante de trabajos relacionados a la estimación de precios utilizando diversos métodos de *Machine Learning*.

La masificación del uso de métodos de *Machine Learning* ha trascendido áreas del conocimiento, siendo utilizados ampliamente en campos sumamente variados como las finanzas, la medicina, la agricultura, el deporte, la robótica, entre otros. Esto no es casual, y es que las ventajas de emplear *Machine Learning* son numerosas y sustanciales. En primer lugar, la capacidad

de procesar grandes cantidades de datos de manera eficiente permite identificar patrones y tendencias que podrían pasar desapercibidos para los métodos tradicionales. Esto se traduce en una toma de decisiones más informada y precisa en diversas disciplinas. Además, el *Machine Learning* posibilita la automatización de tareas repetitivas y laboriosas, liberando recursos humanos para tareas más estratégicas y creativas. Otra ventaja clave es la capacidad de adaptación y mejora continua del modelo. A medida que se obtienen más datos y se refina el modelo, el rendimiento del *Machine Learning* tiende a mejorar, lo que no siempre es posible con enfoques estáticos o reglas predefinidas.

En el caso de esta tesis, se utilizarán métodos de *Machine Learning* para predecir *precios de venta* de productos de segunda mano en base a caracterísicas o atributos de estos. De esta manera, lo que se busca es poder generar una serie de precios sugeridos que sirvan de orientación para la fijación de precios de la empresa. Así, se lograría la *automatización* de un proceso que en la actualidad es realizado en su totalidad por humanos. Lo anterior conlleva una serie de beneficios, sobre todo si se considera que como se mostró en Hammond (1996) [1] el error humano es inevitable: la persona que actualmente pone los precios podría sufrir síntomas de fatiga mental al tener que tomar decisiones de manera constante, lo que podría llevar a errores en la fijación de precios, impactando de esta forma su entendimiento y en última instancia poniendo en riesgo la calidad de los precios que se fijan (Breton et. al (2003) [2]).

En lo que a predicción de precio en base a atributos se refiere, la literatura es amplia y variada, centrándose los trabajos principalmente en el mercado de segunda mano de los automóviles usados y en el de viviendas.
En el mercado de viviendas encontramos los trabajos de Phan (2018) [3], Park y Bae (2015) [4] que utilizando datos históricos de la ciudad de Melbourne, Australia y del condado de Fairfax en Virginia, Estados Unidos, respectivamente, utilizan atributos tanto de la propiedad en sí misma como de la localización de esta como atributos para predecir el *precio de venta* de la propiedad. Similar es el caso del estudio realizado por Tchuente y Nyawa (2021) [5] donde utilizan datos aún más granulares, ya que poseen datos proporcionados por el gobierno de Francia que contiene 5 años históricos de transacciones en el mercado de viviendas de dicho país, que utilizan para comparar múltiples métodos utilizando la serie de atributos que posee

la base.
En el caso de los automóviles usandos, encontramos los trabajos de Listiani (2009) [6], Wu et al. (2009) [7], Peerun et al. (2015) [8] y Gegic et al. (2019) [9], donde cada trabajo utiliza una serie de métodos para predecir *precios de venta* en base a los atributos que posee la base de datos. Por lo general poseen atributos bastante similares tales como el año del vehículo, su kilometraje, marca, número de dueños, y donde difieren es en el procesamiento de datos, dependiendo de qué métodos pretende utilizar cada estudio.

Sin embargo, en el caso de ropa la literatura es más escasa en lo que a predicción de precio en base a atributos se refiere. En este rubro los trabajos se centran fuertementente en la predicción de demanda, como es el caso de Frank et al. (2003) [10], Yu et al. (2019) [11], y Li et al. (2021) [12], donde los trabajos emplean herramientas de tanto de *Machine Learning* como de *Deep Learning* pero en el contexto de la predicción de demanda.

Un trabajo que toca el tema de manera general es el de Fathalla et al. [13], estudio que utilizando datos de plataformas del estilo de *eBay* y *OLX*, propone combinar procesamiento de texto e imágenes para obtener un indicador aproximado de la calidad de cada artículo listado, el cual combinan con los rangos de precio que poseen productos de su categoría en otros anuncios dentro de la plataforma para así generar un precio sugerido.
Más específicos son los trabajos de Munaweera (2019) [14] y Katai et al. (2019) [15], los cuales se centran en la predicción de precio en ropa basada en los atributos de las prendas. El primero de estos estudios es algo más acotado ya que únicamente se centra en una categoría de productos, ropa interior femenina, y sus datos poseen una única marca que es *Victoria's Secret*; además, esta es ropa nueva. El segundo estudio es bastante más amplio y similar a lo que se pretende realizar en el estudio, ya que utiliza datos provenientes de un importante *e-commerce* dedicado a vender ropa de segunda mano que por motivos de confidencialidad llaman *sitio M*, pero del que se infiere que posee multiplicidad de marcas y tipos de prendas. Este estudio si bien es bastante más amplio que el primero, se queda corto en lo que se pretende en el estudio, ya que su análisis solo prueba dos métodos: *Random Forest* y *Multiple Linear Regression* y no realiza un análisis centrado en la predicción del precio como tal, si no que termina enfocándose en la diferencia en el error al usar el *precio de retail* como input.

La principal diferencia entre estos trabajos y el objetivo que busca esta tesis es poder realizar un estudio más completo que incorpore además de una comparativa de los métodos utilizados, también un análisis de los atributos que se utilizarán como input para los modelos, buscando así no solo mejorar las predicciones, si no que también generar un mejor entendimiento de qué atributos son los que finalmente impactan de mayor forma esta predicción del *precio de venta*. Además, se buscará crear una estructura de análisis generalizable a variadas categorías de productos, ya que muchos de estos trabajos se centran únicamente en un tipo de prenda cuando en este caso el objetivo de la compañía es seguir escalando el negocio y abarcar cada vez más categorías de productos, donde el foco está en generar un *marketplace* de buenos productos de segunda mano.

Capítulo 2

Datos

Los datos utilizados provienen de la base de datos de Consignación con la que cuenta la empresa. Esta incluye el histórico de todas las prendas que han recibido para posteriormente ponerlas a la venta. Los datos que contiene van desde fines de Enero del 2022 hasta comienzos de Abril del 2023, es decir, abarca 14 meses. A continuación se presentará la estructura de la base de datos, donde cada fila corresponde a una prenda con los siguientes atributos. Se omitirán las columnas que no son de interés para el estudio.

2.1. Descripción de variables a utilizar

2.1.1. Variables independientes

- ***Id***: Identificador único de la prenda en cuestión.
- ***Marca***: Indica a qué marca pertenece la prenda.
- ***Categoría***: Indica qué tipo de prenda o producto es a nivel grueso.
- ***Sub Categoría***: Es un nivel más abajo que la Categoría, ya que indica la pertenencia a cierto sub tipo de prendas dentro de la Categoría, como pueden ser la Sub Categoría de 'Manga corta' dentro de la Categoría de 'Poleras'
- ***Precio de referencia***: El precio del mismo producto nuevo en el retail.
- ***Estado de prenda***: Indica cómo se encuentra la prenda a nivel de desgaste. Los niveles que presenta la base de datos son los siguientes: *new_with_label, new_without_label,mild_signs,*

very_good, no_signs, good.

- ***Composición de prenda***: Indica los materiales (en porcentaje) de los que está compuesta la prenda.

- ***Market Ganga***: Binaria, indica si es que dicho producto pertenece al apartado 'Market Ganga' que tiene Market People utilizado para promocionar prendas de valores bajos.

2.1.2. Variable dependiente

- ***precio de venta***: El precio en el que Market People trata de vender (o ya vendió) la prenda. Este en la base de datos se almacena como un valor estático, además de no poseer estacionalidad de acuerdo a su fecha de publicación, por lo que se puede considerar como un valor fijo en el tiempo.

2.2. Categorías a utilizar

La empresa trabaja con un número importante de categorías, por lo que se decidió centrar el análisis en un número acotado de estas. Lógicamente, la pregunta de cuáles elegir surge de inmediato. Para esto, se realizó un análisis tanto de la cantidad de prendas presentes en la base de datos (q) como del ingreso que se espera percibir por estos ($p \cdot q$) para desde ahí seleccionar las categorías que tendrán cabida en el análisis.

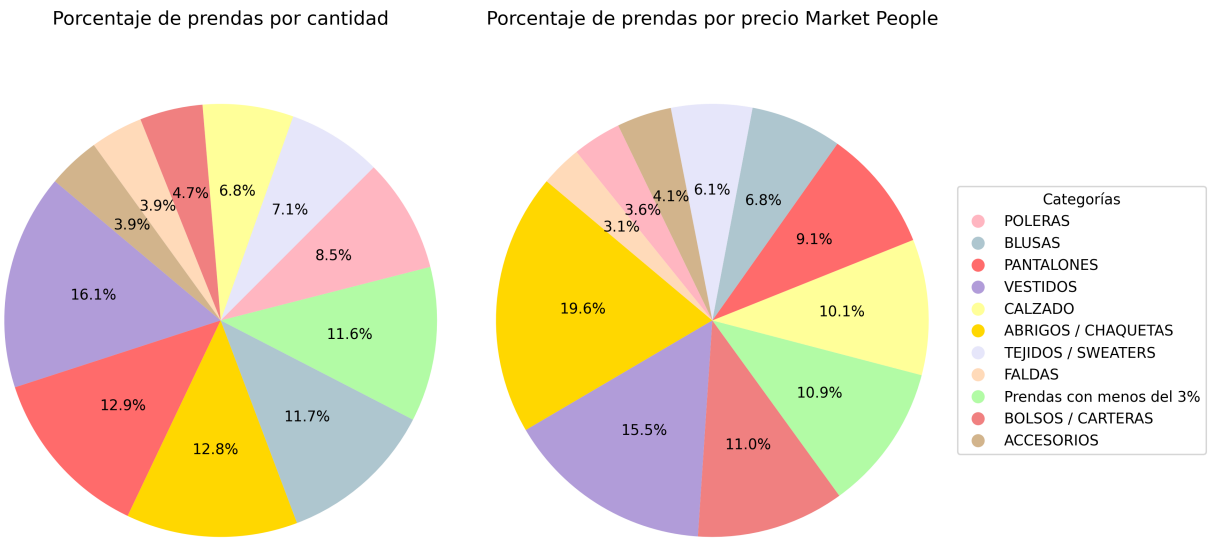

Figura 2.1: Presencia porcentual de prendas en la base de datos. Por cantidad (izquierda) y por precios (derecha).

De los gráficos de *pie* presentados saltan a la vista unas 10 categorías con un número considerable de datos (definiendo el corte en un 3 % del total de ítems). Dentro de estas hay algunas que destacan por tener mucha cantidad relativa (q) como es el caso de los pantalones o las blusas pero no tanto potencial ingreso ($p * q$); otras son el caso contrario: bolsos y carteras junto con calzado si bien no muestran gran cantidad al ser comparadas con otras categorías, sí son productos de ticket alto cuya venta genera grandes ingresos. Vestidos con abrigos y chaquetas son dos categorías que tienen simultáneamente bastante cantidad junto con ingresos potenciales altos.

Sin embargo, hay dos categorías que serán excluidas del análisis posterior y son 'accesorios' y 'bolsos y carteras'. Esto pues en realidad no son prendas, son accesorios y por ende escapan el propósito del estudio.

En consecuencia, se trabajará con un total de 8 categorías, siendo estas: Abrigos / Chaquetas, Blusas, Calzado, Faldas, Pantalones, Poleras, Tejidos / Sweaters y Vestidos.

Será de interés cuantificar el porcentaje de prendas vendidas por categoría, que denotaremos de aquí en más como s. Para esto, hay que tener en consideración que Market People tiene, en teoría, cada prenda a la venta por un máximo de 9 meses. Con esto en mente, de los 14 meses que abarca la base de datos, para el cálculo de estos porcentajes se considerarán

únicamente los primeros 5 meses, puesto que de esta forma nos aseguramos que dichas prendas cumplieron el ciclo completo: o se vendieron antes de los 9 meses, o ya pasaron sus 9 meses y no fueron vendidas. Dicho de otra manera, si utilizáramos todos los datos para el cálculo de este porcentaje, podría ocurrir que consideremos prendas que lleven menos de 9 meses publicadas y, por ende, aún no hayan sido vendidas pero que eventualmente podrían serlo, por lo que no se puede afirmar que dichos productos forman parte de los no vendidos. Esto lógicamente hace que perdamos información, pero es un *trade-off* que asumiremos para ser consistentes según lo mencionado previamente.

El cálculo de estos porcentajes resulta ser el de la Tabla 2.1.

Tabla 2.1: Porcentaje de ventas según categoría

Categoría	**Probabilidad de ser vendido (s)**
POLERAS	72,0 %
BLUSAS	64,3 %
PANTALONES	62,5 %
VESTIDOS	69,5 %
CALZADO	61,0 %
FALDAS	63,2 %
TEJIDOS / SWEATERS	74,2 %
ABRIGOS / CHAQUETAS	73,6 %

2.3. Preprocesamiento de los datos

A la base de datos con las columnas que se utilizarán en el estudio se le aplica un preprocesamiento antes de comenzar el análisis.

Se comienza eliminando las filas (prendas) que tengan valores faltantes en alguna de las variables previamente mencionadas.

Para el caso de ***Estado de prenda***, esta es una variable con muchos niveles pero muchos de ellos tienen una diferencia sumamente sutil que en realidad para efectos del estudio no presenta ventaja alguna, lo cual se valida en la sección 3.2, especificamente en la Tabla 3.1. Se consultó al respecto a la empresa quienes confimaron la hipótesis, argumentando que a grandes rasgos hay dos estados de prenda: las prendas en un estado óptimo (ropa nueva,

con o sin etiqueta) y el resto de las prendas en buen estado. Esto se debe a que Market People es muy cuidadoso en la selección de las prendas que reciben, por lo que prendas en un estado menor a 'Bueno' son devueltas a sus dueños. Uno de los problemas que podrían tener los modelos predictivos al usar todos los niveles de estado de prenda es que al ser tan similares entre sí algunos de los niveles de esta variable, se corre el riesgo de que los modelos sean miopes al no ver que otros estados representan básicamente lo mismo, pudiendo, los niveles con pocos datos, ser especialmente propensos a ser sensibles a outliers, además de ser computacionalmente costoso. De esta forma, se aplicará un agrupamiento de los niveles actuales de la columna a solo dos niveles: *Óptimo* y *Bueno*.

new_with_label	Óptimo
new_without_label	
mild_signs	Bueno
very_good	
no_signs	
good	

La ***Composición de prenda*** es una variable que sin ser numérica podría decirse que es relativamente continua, teniendo en cuenta que hay muchos materiales disponibles y una infinidad de combinaciones porcentuales posibles para cada prenda. Con esto en mente se decidió realizar una agrupación que tome en cuenta si -fijada la categoría, ya que las composiciones a priori no son transversales a todos los tipos de prendas- la ropa en cuestión supera cierto umbral porcentual (>50 %) de ciertos materiales definidos con la ayuda de la empresa como *premium*. El detalle de estos diferenciado por categoría se encuentra en la Tabla A.1 del Anexo. Esto ya que por el solo hecho de contenerlos en gran porcentaje, son prendas de mayor valor. Así, las prendas que contengan al menos en la mitad de su composición un material *premium* pasarán a tener un 1 en la columna y las que no un 0, convirtiendo así esta columna de una columna con características continuas en una columna binaria donde los 1's representan prendas con composiciones *premium* y los 0's prendas que no tienen ese tipo de composición.

Vale la pena mencionar que en estos casos al ser una variable que se transforma en binaria, se pierde información, pero se asume que el beneficio de tener una variable binaria que capture la presencia de materiales *premium* es mayor que el costo de perder información que sería un

continuo sumamente complejo de analizar.

Distribución de *precios de venta* por categoría

Finalizado el preprocesamiento de datos se realiza un *boxplot* para visualizar la dispersión de precios que presenta cada una de las 8 categorías que serán objeto de estudio. Importante

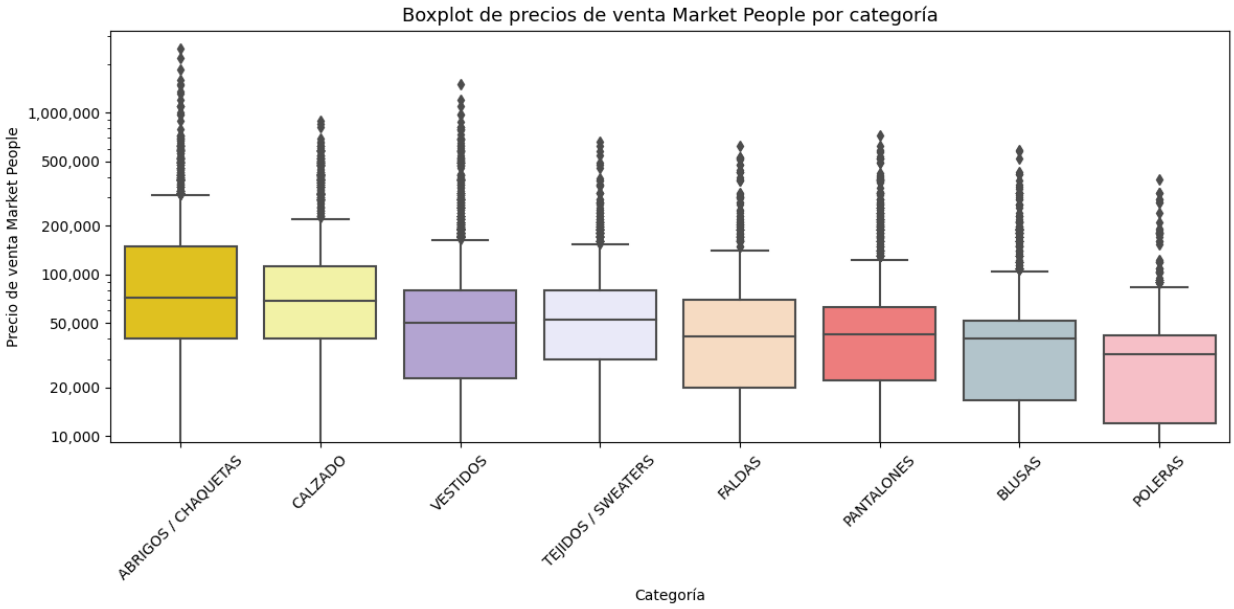

Figura 2.2: Distribución *precios de venta* por categoría

notar que el eje de las ordendas presenta una transformación a logarítmo para una mejor visualización.

El detalle del número de prendas por categoría tanto antes como después del preprocesamiento se encuentra en la Tabla A.2 del Anexo.

Capítulo 3

Desarrollo

3.1. Metodología

3.1.1. Separación en entrenamiento y testeo

Con la finalidad de obtener estimaciones certeras y que sirvan para datos nuevos en el futuro, es preciso que los métodos sean capces de aprender de los datos con los que contamos y, usando ese aprendizaje, predecir de buena forma sobre datos que no conoce. Esta es la lógica que subyace un procedimiento estándar en la ciencia de datos que es el de separar la totalidad de la data en dos: por una parte, datos que servirán para entrenar, por el otro, datos que servirán para probar el desempeño de los métodos.
Entonces, para realizar la estimaciones en datos nuevos y que sean *desconocidos* para los modelos y así evitar caer en problemas ligados a overfitting es que se separa la data en datos que serán utilizados para entrenar los modelos (desde ahora: *entrenamiento*) y datos que serán utilizados para testear los modelos (desde ahora: *testeo*).

Esta separación en *entrenamiento* y *testeo* será realizada de forma aleatoria, dejando el 80 % de filas (prendas) en la data utilizada para entrenar los modelos y el 20 % restante en la data utilizada para testear el desempeño de estos métodos ya previamente calibrados en prendas que no ha visto antes.
Sin embargo en este procedimiento es necesario hacer una salvedad, ya que hay casos que excluiremos del set de testeo por razones prácticas. Estos son los casos donde se tiene un único ítem por marca, los que se dejarán únicamente en el set de entrenamiento, de lo contrario

habrían modelos que no podrían calibrar los parámetros adecuados para la marca respectiva, y por ende no se podría hacer la predicción.

Así, en caso de haber dejado en esta separación aleatoria de 80-20 % a alguna Marca sin representación en *entrenamiento*, lo que se realiza es tomar alguna prenda aleatoria de dicha Marca (que lógicamente estará en *testeo*) e intercambiarla por alguna prenda aleatoria de los datos de *entrenamiento* siempre y cuando esta prenda aleatoria no sea una tal que su marca quedará sin representación en *entrenamiento*.

3.1.2. Transformación logarítmica del *precio de venta*

Para efectos de una mejor precisión se decidió realizarle una transformación logarítmica al *precio de venta*, la variable a estimar. Esta es una transformación sumamente estándar a la hora de trabajar con precios y ha sido ampliamente aplicada en diversos estudios [16]. ¿Pero por qué utilizarla en el caso de precios? Los precios son por lo general variables que tienen valores con diferentes órdenes de magnitud, por lo que si los datos se mantienen con esas diferencias los errores podrían ser capturados por los ítems de mayor magnitud. Para mayor certeza con respecto al uso de esta transformación se probaron estimaciones de los métodos que se mencionarán a continuación con y sin la transformación aludida, testeando utilizando *cross-validation* en el set de *entrenamiento* y para todos los métodos mostraba mejores resultados el uso de la transformación $log(precio\_venta)$.

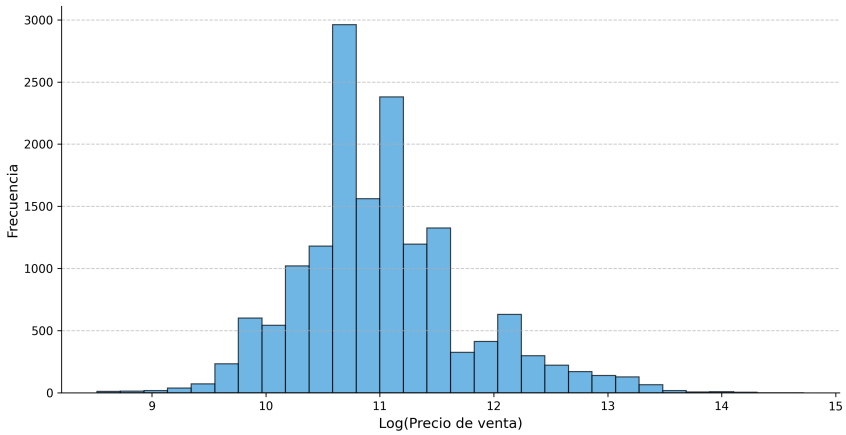

Figura 3.1: Distribución de *precio de venta* y su transformación logarítmica

En el Apéndice, Figura B.1 del Anexo B se muestra la distribución de *precio de venta* sin

la transformación logarítmica.

3.1.3. Variables a utilizar

Las variables que se utilizarán de input para los diferentes métodos serán Marca, Sub Categoría, Is ganga, Estado de prenda y Composición de prenda. Hay que recordar que en el tratamiento realizado a las variables, la naturaleza de algunas de estas se vio modificada en beneficio de mejores predicciones y una mayor interpretabilidad de los métodos. El caso de la variable *precio de retail* o *precio de referencia* merece especial atención, ya que es una variable particular por los siguientes motivos: no todas las prendas tienen un *precio de referencia* puesto que el hecho de obtener este precio es costoso en términos de recursos, implicando utilizar tiempo de trabajadores en buscar las mismas prendas que la tienda recibe en consignación pero en diferentes tiendas. Este trabajo que ya es costoso de por sí no siempre logra su objetivo, ya que existen prendas que no tienen su precio publicado en internet o derechamente no aparecen en las búsquedas por la web.

En base a lo anterior se podría pensar que una simplificación tanto del análisis como del mismo proceso de la compañía sería obviar este *precio de referencia*. Sin embargo, como se verá a continuación, este atributo puede ser muy valioso para la predicción de *precios de venta*.

En la siguiente tabla se presentan las variables a utilizar junto a la naturaleza de cada una de estas

Variable	**Naturaleza de la variable**
Marca	Categórica
Sub Categoría	Categórica
Is ganga	Binaria
Estado de prenda	Binaria
Composición de prenda	Binaria
Precio de Retail*	Entera

Donde encontramos que *precio de retail* cuenta con un asterisco ($*$) pues es una variable que irá variando en el análisis, ya que se probará el efecto de utilizarla o no en los modelos.

Ambos casos, tanto al utilizar todos los atributos junto con *precio de retail* como sin este, tienen los mismos sets de entrenamiento y testeo, con la finalidad de así comparar las predicciones sobre exactamente los mismos ítems.

3.1.4. Métrica de error escogida

La métrica de error escogida es el *Promedio del Error Porcentual Absoluto* (desde ahora *MAPE* por sus siglas en inglés). Como se mencionó anteriormente, la variable dependiente a predecir es *precio de venta* con su transformación logarítmica ya aplicada.

3.1.5. Métodos

Se presentarán los 8 métodos utilizados para predecir el *precio de venta*. Todos estos métodos, exceptuando la *Linear Regression* -que no posee hiperparámetros como tal- fueron probados con diferentes hiperparámetros, los que fueron testeados utilizando *cross-validation* dentro del set de *entrenamiento* para finalmente utilizar los mejores hiperparámetros que cada uno de los métodos arrojó. El detalle del *pool* probado -incluyendo los que resultaron escogidos- se encuentra en la tabla A.3 del Apéndice A.

Los métodos son los siguientes:

- eXtreme Gradient Boosting (XGBoost): Este método se basa en el principio del boosting, en el que se entrena una serie de modelos secuencialmente, cada uno aprendiendo de los errores del anterior. En cada iteración, se agrega un nuevo árbol de decisión al modelo, ajustando sus parámetros para minimizar el error residual del modelo combinado.
  El algoritmo utiliza una técnica llamada 'gradient boosting' para optimizar los parámetros de los árboles de decisión. Esta técnica calcula la derivada del error con respecto a los parámetros del árbol, lo que permite ajustar los parámetros de manera más eficiente.

- K-Nearest-Neighbor (KNN): Es un algoritmo de aprendizaje automático bastante antiguo que se utiliza para resolver problemas de clasificación y regresión. El método se basa en el principio de que los puntos de datos similares tienden a pertenecer a la misma clase o categoría. Así, KNN clasifica un nuevo punto de datos asignando la clase más común entre sus k vecinos más cercanos en el espacio de características, siendo k un parámetro

que recibe el modelo.

- LASSO Regression (LASSO): El Método Lasso (Least Absolute Shrinkage and Selection Operator), conocido desde la década de los 90's, es un método que combina un modelo de regresiónn linea con un procedimiento de contracción de algunos parámetros hacia cero y selección de variables, imponiendo penalizaciones sobre los coeficientes de la regresión.

- LightGBM: El método es una implementación de Gradient Boosting Framework que destaca por su eficiencia y alta precisión. El algoritmo utiliza una estrategia de crecimiento de árboles que se centra en dividir los nodos con las características que tienen el mayor impacto en la reducción del error.

- Linear Regression (LR): La regresión lineal es quizá el método más conocido de los mencionados. Este se basa en tratar de establecer una relación entre la variable y a predecir con una serie de variables explicativas, buscando ajustar los parámetros que minimicen su error al ajustarse a los datos existentes. A pesar de ser simple y fácil de interpretar, es ampliamente utilizada y puede capturar muy bien relaciones entre variables.

- Neural Network (NN): Las redes neuronales son modelos inspirtados en la forma en que el cerebro humano procesa la información. Consiste en nodos conectados entre sí que se organizan en diferentes capas capas. Estos nodos o neuronas por sí solos tienen una lógica bastante sencilla, pero es en su conexión con el resto donde son capaces de generar un conocimiento mayor que permite reconocer patrones, predecir, clasificar, entre otras cosas.

- Random Forest (RF): Random Forest es un algoritmo de aprendizaje supervisado que se basa en la creación de múltiples árboles de decisión. Los árboles se crean de forma aleatoria, utilizando subconjuntos tanto de los datos de entrenamiento como de las variables. Esto ayuda a reducir el riesgo de *overfitting* y a mejorar la precisión de las predicciones. Los resultados de los árboles se combinan mediante votación para clasificar un nuevo punto de datos.

- Support Vector Machine (SVM): Este método se basa en un problema de optimización que lo que hace es encontrar el hiperplano óptimo que mejor separa los datos. Una

vez que se ha encontrado el hiperplano óptimo, se puede utilizar para clasificar nuevos puntos de datos. Para ello, se calcula la distancia entre el nuevo punto y el hiperplano. Si la distancia es mayor que 0, se clasifica el punto en la clase positiva. Si la distancia es menor o igual que 0, se clasifica el punto en la clase negativa.

3.2. Resultados

3.2.1. Estimación del *precio de venta*

A continuación se presentan los resultados obtenidos a partir de la utilización de los métodos previamente mencionados para la estimación del *precio de venta*. Como se mencionó en la sección Metodología, se probaron los métodos utilizando por una parte, todos los atributos *sin precio de retail* y, por otra, todos los atributos *incluyendo precio de retail*.

En la Figura 3.2 se muestra el *MAPE* en el set de testeo para diferentes categorías con los 8 métodos predictivos usados al utilizar todos los atributos exceptuando el *precio de retail* e incluyendo este atributo.

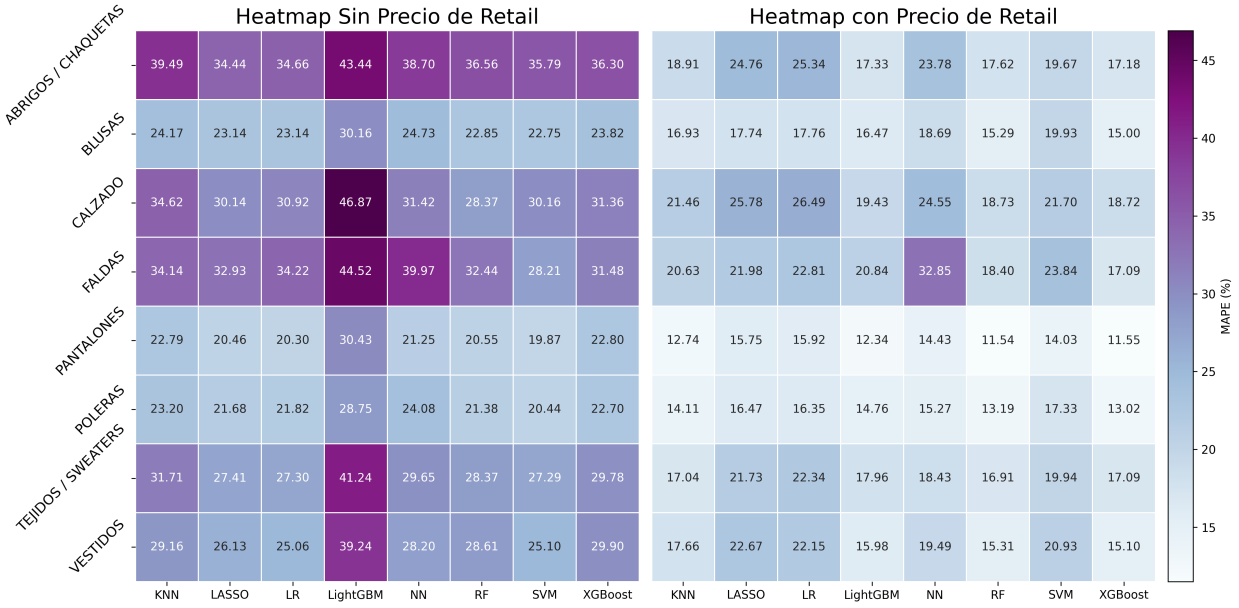

Figura 3.2: Heatmap que muestra los *MAPE* para todas las categorías y métodos utilizados

Es interesante notar que cuando no está *precio de retail* dentro de los atributos, los

métodos que tienen un mejor desempeño predictivo son *Linear Regression* (LR) y *Support Vector Machine* (SVM). Esto se evidencia de manera transversal a las categorías, teniendo un desempeño sobre el promedio de los métodos para cada una de estas.

Cuando *precio de retail* está dentro de los atributos, los resultados son diferentes al caso anterior. Ahora los métodos *Random Forest* (RF) y *eXtreme Gradient Boosting* (XGBoost) son quienes tienen un menor *MAPE*, llegando a errores porcentuales que disminuyeron en más de 10 puntos porcentuales para ambos métodos en la comparativa de su desempeño al no tener el atributo *precio de retail*.

En el Apéndice C, Figura C.2 se presentan los *MAPE* de utilizar todos los niveles de *Estado de prenda* en vez de la simplificación realizada en la Sección 2.3, donde se evidencia que la simplificación cumple el objetivo al obtener resultados mejores (ver Tabla 3.1) en el caso sin *precio de referencia* y muy similares en el caso que incluye la variable.

Otra posible variación que surge de manera natural es el utilizar como variable *log*(*precio de retail*) en vez de *precio de retail*, ya que estamos prediciendo el logaritmo del *precio de venta*. De esta forma, se presentan en la Figura C.1 los *MAPE* de utilizar dicha transformación, donde se observa que dicha transformación no genera mejoras sustanciales en relación con el caso base (ver Tabla 3.1).

A modo de resumen, se presentan en la Tabla 3.1 los *MAPE* promedio para el caso base (el mismo que se presentó en la Figura 3.2) y las diferentes variaciones recién mencionadas (exceptuando la última columna, la cual se abordará al final de la sección), donde se calcula el promedio de los *MAPE* de todos los métodos para cada una de las categorías, diferenciando si se utilizó o no *precio de retail* en los atributos.

Tabla 3.1: Promedio del *MAPE* para cada caso

	Caso base	**Estado sin agrupar**	**log(*precio de retail*)**	**Variable dependiente es el % de descuento.**
Sin *precio de retail*	29,26 %	29,41 %	29,26 %	No aplica
Con *precio de retail*	18,55 %	18,48 %	23,23 %	17,28 %

3.2.1.1. Importancia de variables en la predicción

Con el objetivo de comprender con mayor profundidad la importancia que asignan a cada una de las variables estos métodos que en su mayoría son *black-boxes* se utilizará la librería *SHAP* de *python*, donde su nombre proviene de SHapley Additive exPlanations y permite dar un enfoque en base a teoría de juegos, ocupando los Valores de Shapley para poder explicar comó afecta cada variable en los modelo de aprendizaje automático (SHapley Additive exPlanations, 2018, s.p.).

Este análisis será efectuado para *Random Forest*, ya que es el método que mostró un mejor desempeño en el caso de utilizar todos los atributos con *precio de retail*. Para no sobrecargar el informe, por momentos se fijará una categoría arbitraria para realizar el análisis, mientras que los gráficos asociados a las categorías restantes se encontrarán en los Anexos.

3.2.1.2. Aporte de variables

A continuación se presentan gráficos para visualizar el aporte promedio en valor absoluto de cada una de las variables del modelo, diferenciando este aporte en cada una de las 8 categorías del estudio.

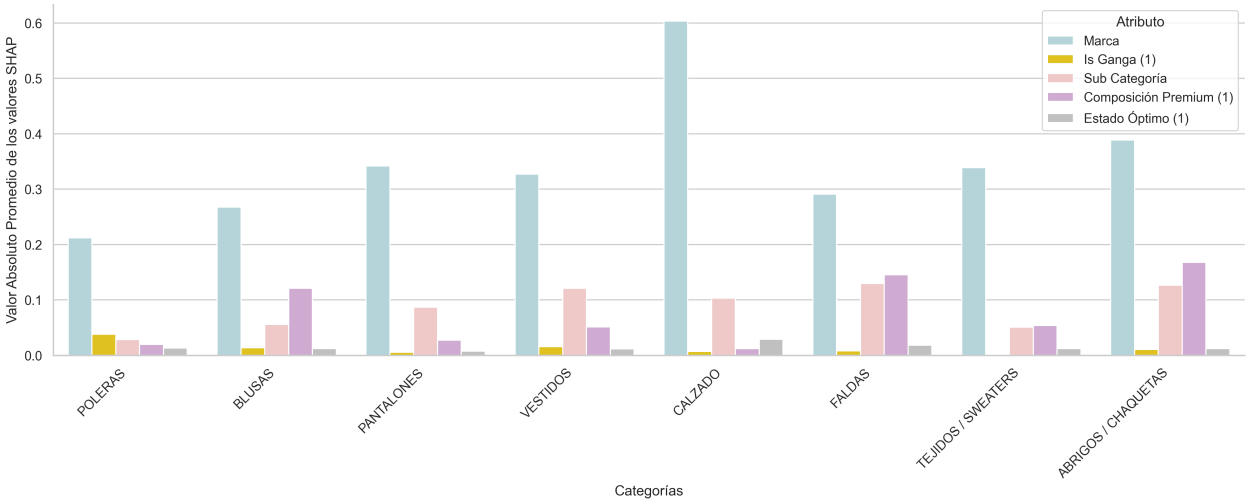

Figura 3.3: Promedio absoluto de SHAP values *sin precio de retail*

Se observa de la Figura 3.3 que ante la ausencia del *precio de retail* se cumple con lo que

indica la lógica (más cuando se trata de productos de lujo), y es que la variable asociada a la Marca es el atributo con una mayor influencia a la hora de generar la predición de precio.

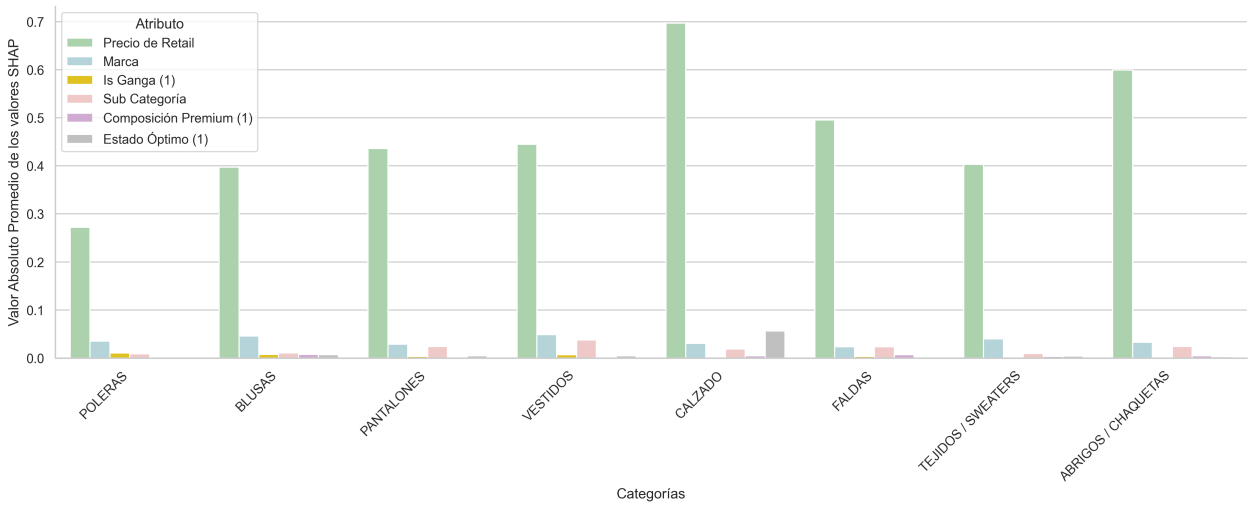

Figura 3.4: Promedio absoluto de SHAP values *con precio de retail*

En la Figura 3.4 se evidencia que cuando se le agrega el *precio de retail* al *pool* de variables se hace más que notoria la preponderancia que tiene este nuevo atributo a la hora de generar un impacto en el método. Su aporte es órdenes de magnitud mayor que el resto de los atributos, lo que vendría a explicar la gran mejora que *Random Forest* experimenta una vez se incluye el *precio de retail* como variable.

De ahora en adelante, en algunas secciones el análisis se realizará para la categoría *Vestidos* para no sobrecargar el informe. La elección de esta categoría obedece a que posee un número importante de ítems (ver Figura 2.1), lo que permite realizar un análisis más robusto; además, es una categoría que según la Figura 2.2 no es de las más caras ni de las más baratas, por lo que se espera que el análisis sea representativo de lo que ocurre en el resto de las categorías.

Ya habiendo analizado las importancias de variables en términos de su valor absoluto, es válido cuestionarse cómo es que los valores de dichas variables afectan a la predicción final en su signo: ¿tienden a generar *precios de venta* más altos o más bajos?.
Para esto se presentarán gráficos de tipo *beeswarm* que permiten visualizar el aporte de las variables en la predicción de *precio de venta* para los ítems de una categoría en particular, la cual será *Vestidos*. En esta visualización se puede observar cómo es que cada variable afecta

a la predicción en relación a su valor relativo, si es que aumenta o disminuye el *precio de venta*.

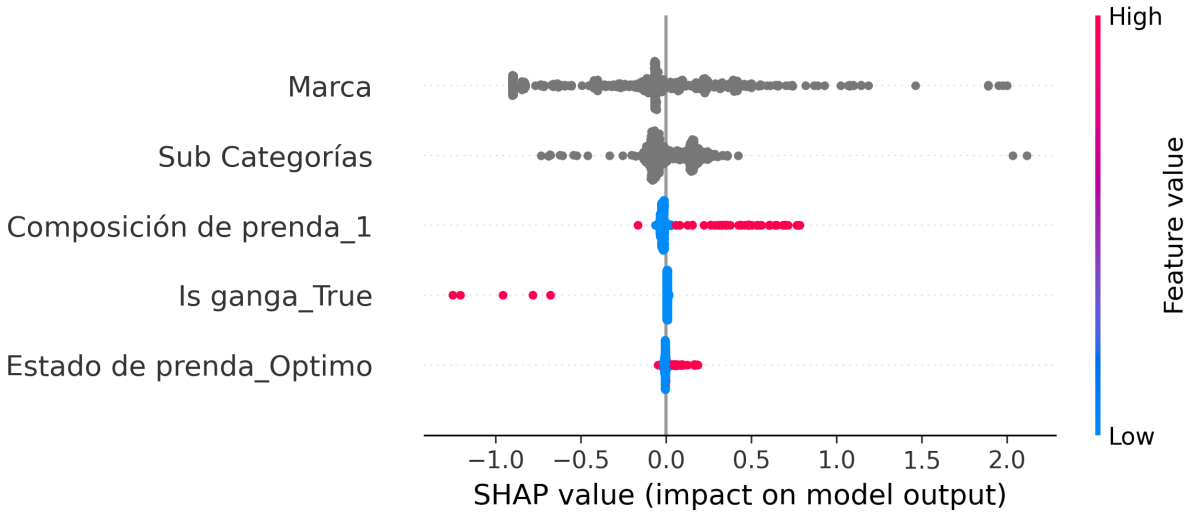

Figura 3.5: *beeswarm sin precio de retail*

En la Figura 3.5 se observa que al analizar los signos estos son concordantes con lo esperado, ya que el hecho de que una variable sea *premium* (*i.e.* que "Composición de prenda" se active) tiene un aporte positivo a la predicción, mientras que al activarse "Is ganga" esta tiene un aporte negativo. Asimismo ocurre con el "Estado óptimo", ya que cuando esta variable es 1 quiere decir que está casi nueva y esto refleja que dicha caracterísica hace que la predicción del *precio de venta* sea más positiva.

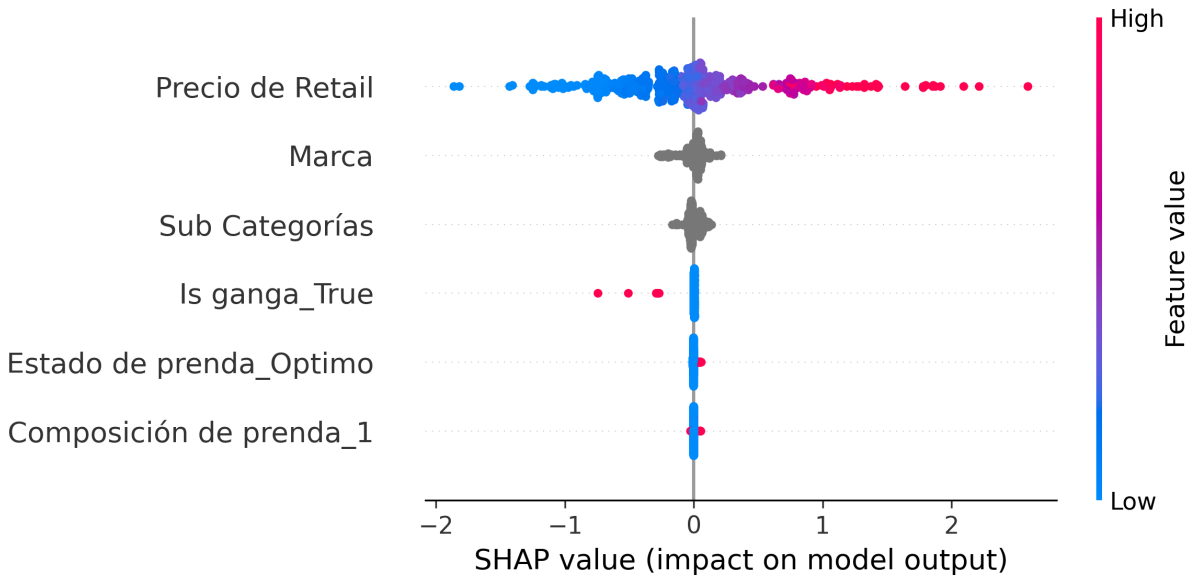

Figura 3.6: *beeswarm con precio de retail*

La Figura 3.6, cuando se incluye *precio de retail*, sigue la misma lógica con la salvedad de que tal como se vio en la Figura 3.4 el *precio de retail* es el atributo que más aporta a la predicción. Como la variable *precio de retail* es un continuo, a diferencia de las binarias (donde el valor 1 es el rojo y el 0 es el azul), se evidencian más cambios en la escala de color que lo que muestran es el valor relativo dentro de esa variable en particular. Lo que se evidencia de esta variable es que su aporte está estrechamente ligado al valor relativo que tiene el *precio de retail* de la prenda a analizar: si es bajo, su constribución tenderá a ser negativa, mientras que si es alto, su constribución será positiva a la hora de predecir el *precio de venta*.

3.2.1.3. Reconstrucción de predicción vía Shapley values

Los Shapley values que calcula la librería son para cada ítem, por lo que podemos *reconstruir* la lógica que siguió el método aludido para entregar la predicción final asociada a cada ítem en particular. La visualización sigue un formato de cascada, siendo natural leerla desde abajo hacia arriba, donde la base comienza en $E[f(x)]$ que es el estimado promedio para ítems de la categoría fijada. Luego, se comienzan a añadir de forma lineal las contribuciones de cada variable, que van aumentando o decreciendo este valor inicial hasta llegar a la parte superior, donde junto a la última barra de aporte está $f(x)$ que viene a ser la predicción final para

el ítem. Es necesario recordar que se está trabajando con precios logarítmicos, por lo que la escala está distorsionada con respecto al peso chileno, pero al ser logarítmo una función monótona los signos permanecen inalterables, siendo una buena herramienta para ver qué variables terminaron por aumentar o disminuir el valor predicho de esta prenda en particular.

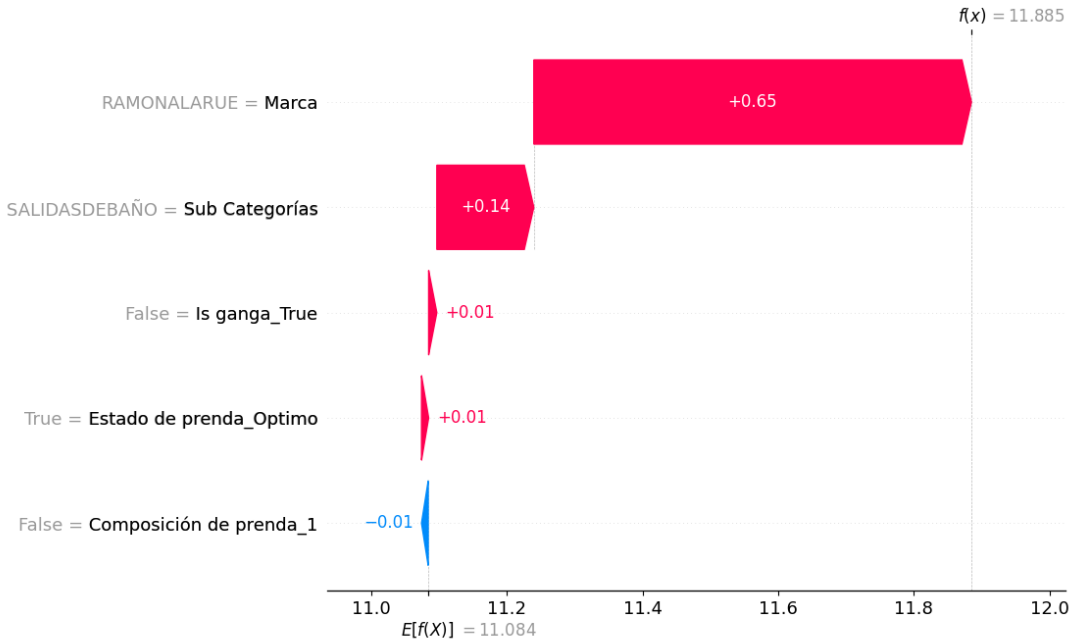

Figura 3.7: Ejemplo: Gráfico de cascada para un ítem de la categoría Vestidos

En la Figura 3.7 se presenta un ejemplo de este gráfico de cascada para un ítem de la categoría Vestidos. Se puede ver que el valor inicial de la predicción es de $11,084$, que es el promedio de las predicciones de *precios de venta* de los ítems de la categoría Vestidos (en su set de testeo lógicamente). Luego, siguiendo la lógica mencionada, notamos que el hecho de que como este ítem no posee una composición *premium*, tiene un aporte negativo a la predicción. En cambio, tiene un aporte positivo -aunque también de magnitud leve- el hecho de que su estado de prenda es *óptimo* y que no pertenece al catálogo *Is ganga*, donde la empresa lista prendas de precios más bajos. Es interesante notar que la Sub Categoría que posee, *Salidas de baño*, sí le entrega un aporte positivo que hace que la predicción para este ítem en particular sea por sobre el promedio de los datos, con una magnitud bastante mayor que los atributos previamente mencionados. Por último está el atributo marca, que es por mucho el que aporta más a la predicción, generando un aumento positivo de $0,65$. La marca

de esta prenda es *Ramona LaRue*, la cual es

> *(...) una tienda de moda femenina bohemia de alta gama en la que cada diseño comienza con una pintura de seda, acuarela o boceto único en su tipo, y está firmado por la diseñadora Arianne [17]*

por lo que es concordante con la lógica de que prendas de marcas de lujo, en este caso de diseñador, tengan un *precio de venta* mayor.
Finalmente, esta prenda en particular termina con un precio estimado -en logaritmo- de $11,885$, que al ser transformado de vuelta a su escala original es de $145.074, sumamente similar al precio predicho por *Random Forest* que es $145.084 y con un error de menos del $10\,\%$ con respecto al precio real, que es $159.990.

3.2.1.4. Estimación del descuento a aplicar a la variable *precio de retail*

Si el *precio de referencia* fuese una variable que se conociera con certeza, se podría haber predicho el descuento que se le aplicará a este para obtener el *precio de venta*. Sin embargo, como se mencionó anteriormente, la empresa no tiene un descuento fijo que se le aplique a todas las prendas, sino que este varía dependiendo de la prenda, su estado, su composición, entre otros factores. Por esta razón el trabajo en esta tesis se centró en objetivo de la empresa es poder predecir el *precio de venta* de la prenda, no el descuento que se le aplicará al *precio de retail*. Sin embargo, en la Figura 3.8 se muestra en una estructura similar a la presentada anteriormente el error en las predicciones cuando la variable a predecir (y) no es el *precio de venta*, si no que el *descuento* que se le aplicará al *precio de retail* para obtener el *precio de venta*. Los métodos utilizan los mismos hiperparámetros que en el caso de la estimación del *precio de venta* y para el calculo del *precio de venta* predicho para cada prenda lo que se debe hacer es multiplicar el *precio de retail* de dicha prenda con $1 - \hat{y}$, donde $\hat{y}$ es predicción asociada al *descuento* que se le aplica. Luego, se computa la diferencia entre el *precio de venta* verdadero con el predicho y se obtiene el *MAPE* para cuantificar los errores en la predicción.

Figura 3.8: Heatmap que muestra los *MAPE* para todas las categorías y métodos utilizados cuando la variable dependiente no es el precio de venta, si no el descuento aplicado al precio de retail

En la Figura 3.8 se observan los *MAPE* para cada método y categoría, donde se puede ver que el método que tiene un mejor desempeño es *Random Forest*, similar al caso base. Estos presentan un error promedio similar al caso base con *precio de retail*.

El *MAPE* resumen de este método se encuentra en la Tabla 3.1 junto con las distintas variantes de predecir el *precio de venta*. Notar que en este caso no aplica el caso sin *precio de retail* pues sin este precio sería imposible generar una predicción en este caso.

3.2.2. Estimación de intervalos de precio

Hasta ahora se han centrado los esfuerzos en estimar el *precio de venta* para cada ítem, sin embargo, como se observa en la Tabla 3.2 un aspecto que no se puede obviar es que existe una considerable heterogeneidad en precios para ítems que son idénticos en los atributos que se tienen, no obstante, son productos distintos con *precios de venta* diferentes. Esto hace surgir la pregunta: además de entregar el estimado del *precio de venta*, ¿es posible entregar intervalos de confianza que ayuden a orientar posibles *precios de ventas*?

Tabla 3.2: Heterogeneidad en *precios de venta* para ítems idénticos

Marca	Sub Categoría	Is ganga	Estado de prenda	Composición de prenda (Premium)	Precio de Retail	Precio de Venta
FREEPEOPLE	MINI	FALSO	Óptimo	0	$70.100	$42.990
FREEPEOPLE	MINI	FALSO	Óptimo	0	$59.700	$35.990
FREEPEOPLE	MINI	FALSO	Óptimo	0	$299.400	$192.990
FREEPEOPLE	MINI	FALSO	Óptimo	0	$299.400	$192.990
FREEPEOPLE	MINI	FALSO	Óptimo	0	$83.500	$65.990
FREEPEOPLE	MINI	FALSO	Óptimo	0	$63.700	$49.990
FREEPEOPLE	MINI	FALSO	Óptimo	0	$86.000	$66.000

Como se vio anteriormente, un método sólido a la hora de estimar el *precio de venta* era *Linear Regression*. Este método posee una serie de características que ayudan a interpretar de manera directa los resultados del modelo, generando conocimiento de cómo opera el método al utilizar las variables que obtiene como *input*. Más importante aún para lo que motiva la sección, *Linear Regression* puede generar de forma cerrada ecuaciones que entreguen intervalos de confianza para el $\hat{y}$. Por lo anterior, se utilizará este método para generar los intervalos de predicción.

A continuación se presentará el resultado de la derivación de estos intervalos de confianza para la estimación.

Expresión para intervalos de confianza

Sea $se(\hat{Y}_i) = \sqrt{\widehat{Var}(\hat{Y}_i)}$ la raíz cuadrada del elemento correspondiente a la iésima diagonal de $\hat{Var}(\hat{Y})$. Entonces, el intervalo de confianza al $100 \cdot (1-\alpha)\,\%$ para la estimación media se puede calcular como:

$$\hat{Y}_i \pm t_{\left(1-\frac{\alpha}{2}, N-2\right)} \cdot se(\hat{Y}_i)$$

3.2.3. Intervalos obtenidos

Para la obtención de los intervalos deseados se utilizó la expresión para aplicarla en los datos de testeo. Se fijó una confianza del 80 % y la categoría *Vestidos* para el análisis. Además, como ha sido habitual, se hizo la distinción entre las estimaciones que incluían (o no) el *precio de retail* como atributo.

Las Figuras 3.9 y 3.10 muestran en el eje-x los *precios de venta* verdaderos de cada una de las prendas de la categoría, ordenados de manera creciente; en el eje-y se encuentran los

intervalos de confianza que surgen a partir de las predicciones para cada uno de estos ítems, teniendo en los extremos líneas horizontales que representan los límites del intervalo. Además, para cada intervalo asociado a un ítem, este se encuentra coloreado verde o rojo, dependiendo si este contiene (o no) el *precio de venta* verdadero de dicha prenda.

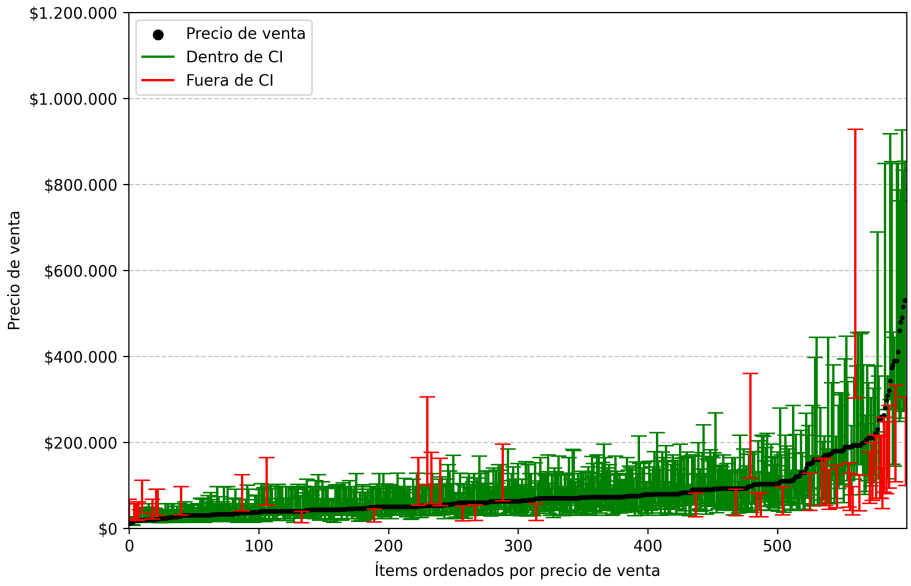

Figura 3.9: Intervalos de confianza para cada punto: *Linear Regression* sin *precio de retail*

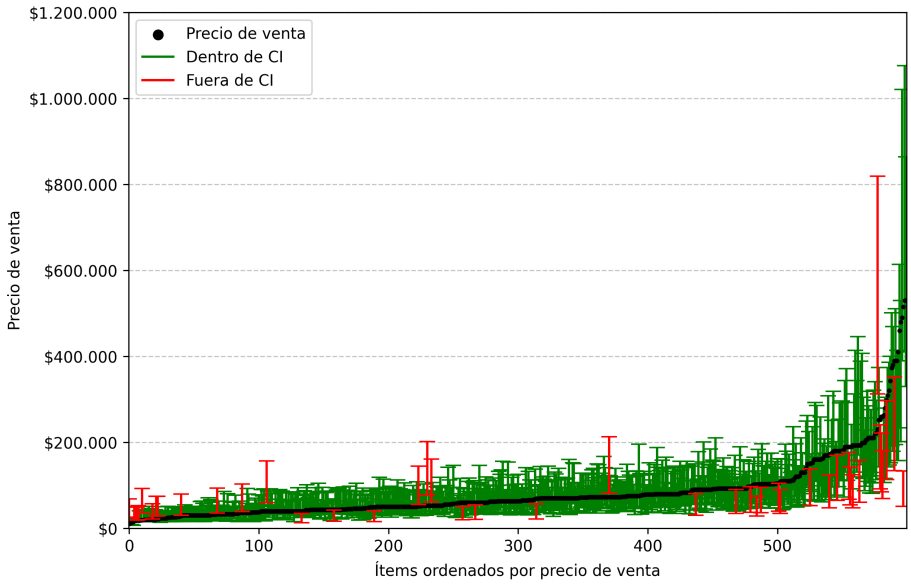

Figura 3.10: Intervalos de confianza para cada punto: *Linear Regression*con *precio de retail*

Para el caso en el que no está *precio de retail* dentro de los atributos, el 78 % de los ítems tienen su *precio de venta* dentro de su intervalo de confianza respectivo. Para cuando sí está *precio de retail*, esta cifra es un 80 %. Estos resultados muestran ser consistentes con el nivel de confianza buscado (80 %) para ambos casos. Es importante destacar que, a pesar de que porcentualmente ambos gráficos son consistentes con el nivel de confianza deseado, se evidencia una clara mejora a la hora de incluir el *precio de retail*, lo que se ve reflejado en que para cada uno de los puntos decrece la amplitud de su intervalo de confianza, mostrando que para un mismo nivel de confianza se puede entregar un intervalo mucho más preciso para la predicción del ítem. En particular se evidencia una mejora a la hora de indentificar los intervalos para los productos caros cuando sí se utiliza la variable *precio de retail* en relación a cuando no.

Las Figura 3.11 muestra -para cada una de las categorías del análisis- cómo se comporta la pertenencia del *precio de venta* verdadero de los ítems a los intervalos creados a partir de la *Linear Regression* para niveles de confianza desde un 5 % hasta un 95 %. Ambas figuras muestran que los datos se ajustan de buena forma a la confianza con la que son creados estos intervalos, donde la línea negra punteada es la referencia de una correspondencia perfecta

(1:1) entre Nivel de confianza y el Porcentaje de ítems dentro de este intervalo.

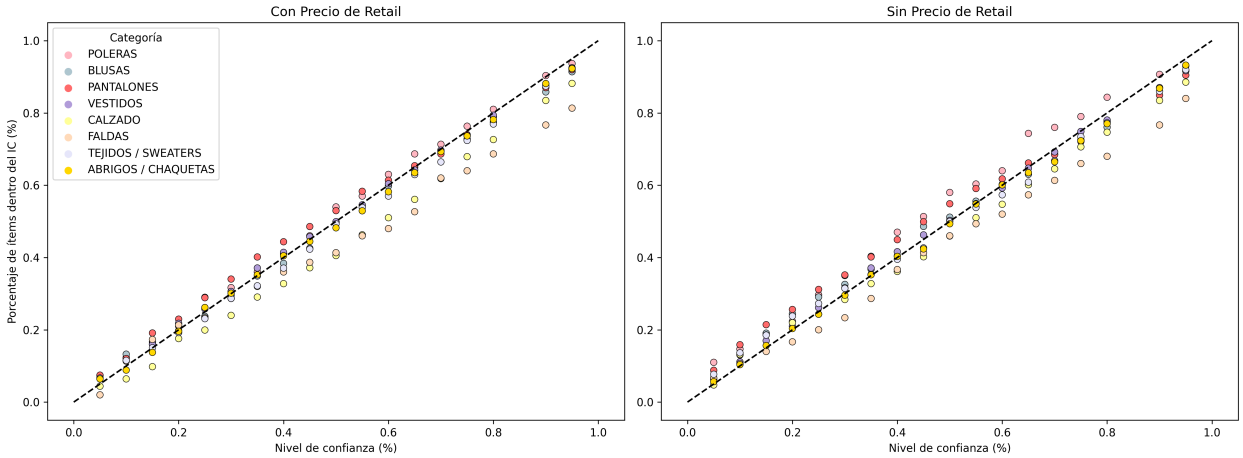

Figura 3.11: Scatterplot: niveles de confianza (por categoría)

En vista de los buenos resultados obtenidos con los intervalos a partir de la *Linear Regression*, se intentó generalizar el procedimiento a otros métodos como *Random Forest* utilizando el módulo de Python *forest-confidence-interval* [18]. Sin embargo, no se logró obtener resultados satifactorios a la hora de generar intervalos de precio (ver resultados en Apéndice D). Esto se debe principalmente al hecho de que, a diferencia de la *Linear Regression*, estos métodos *black-boxes* son calibrados internamente. En particular, *Random Forest* lo que hace es combinar una serie de árboles diferentes, con discrepancias en las variables utilizadas, en los datos considerados y con una importante componente aleatoria, sin poseer tampoco fórmulas cerradas que permitan una aproximación matemática como fue el caso de la *Linear Regression*, donde en el desarrollo de la derivación de la expresión se utilizaron una serie de supuestos matemáticos en los que se basa este método y que permiten obtener los intervalos de confianza deseados.

3.2.4. Curva de demanda

La motivación de esta subsección es poder cuantificar el impacto en las utilidades al utilizar las predicciones generadas a partir de los modelos predictivos como *precios de venta* en vez de los precios escogidos por Market People. Adicionalmente, nos interesa evaluar el desempeño, tanto de métodos predictivos que usan el *precio de retail* en la predicción, como en el caso en que este no se usa. Los datos utilizados para este análisis serán las predicciones para la categoría *Vestidos*, donde se utilizarán los modelos *Linear Regression* para el caso sin *precio*

de retail y *Random Forest* para el caso con *precio de retail*.

Para esto naturalmente es necesario generar una curva de demanda que permita relacionar el precio de una prenda y la probabilidad de que esta se venda. No obstante, el problema de usar el precio es que es relativo a cada ítem, por ende, precios de ítems distintos no son comparables. Para abordar esto, lo que se hará es normalizar posibles precios con respecto al *precio de venta* que efectivamente puso Market People para cada prenda. De esta forma, un ratio (de ahora en más, r) de 1.5 implica que el precio de venta de la prenda es un 50 % mayor que el precio que efectivamente puso Market People.

Queremos obtener una forma funcional para $q(r)$ que represente la probabilidad de venta en términos porcentuales respecto al total de la categoría (q) en función de r, que es el porcentaje total de un precio arbitrario con respecto al *precio de venta* real. Junto con eso, deseamos que la función cumpla que $q(1) = s$, donde s es el porcentaje de venta real correspondiente a dicha categoría, es decir, es la probabilidad empírica que una prenda de una categoría en específico se venda, como se mencionó en la Tabla 2.1. En otras palabras, si es que el ratio es 1, esto implica que el precio a usar es igual al *precio de venta* usado por Market People. En este caso, la probabilidad de venta de una prenda sería exactamente la probabilidad empírica, s.

Para esto impondremos una forma sinusoidal para las probabilidades de venta, donde se asume que la probabilidad de venta es máxima cuando el precio es el más bajo posible, y que esta disminuye a medida que el precio aumenta. La forma funcional que se utilizará es la siguiente:

$$q(r) = \frac{e^{\beta_0 + \beta_1 \cdot \ln(r)}}{1 + e^{\beta_0 + \beta_1 \cdot \ln(r)}} \tag{3.1}$$

Dado que se busca que $q(1) = s$, se tiene que $\frac{e^{\beta_0}}{1+e^{\beta_0}} = s$, de donde se obtiene que $\beta_0 = \ln\left(\frac{s}{1-s}\right)$.

Se necesita asignarle un valor a β_1 de alguna forma puesto que no se tiene una forma cerrada para este parámetro. Para esto se utilizará el supuesto de que β_1 es tal que es el que proviene cuando el ratio r es el que maximiza las utilidades esperadas, es decir r^*. Por ende, para obtener el parámetro β_1 se busca resolver el problema de optimización que se presenta

en la Ecuación (3.2) de donde se obtiene el r^* que maximiza las utilidades esperadas.

$$\max r \cdot q(r) = \max r \cdot \frac{e^{\beta_0+\beta_1 \cdot \ln(r)}}{1+e^{\beta_0+\beta_1 \cdot \ln(r)}} \tag{3.2}$$

De resolver este problema mediante condiciones de primer orden (CPO) se obtiene la siguiente relación en el óptimo:

$$1 + \frac{s}{1-s} \cdot r \cdot \beta_1 = -\beta_1 \tag{3.3}$$

Donde β_1 es el parámetro a estimar para diferentes r.

Utilizando métodos numéricos para resolver el problema, pues no cuenta con una solución analítica, se obtendrá $\beta_1(r^*)$. Acá podemos notar que para el caso particular en donde $r^* = 1$, esto es, el precio óptimo que maximiza las utilizades esperadas es el usado por Market People, entonces sí hay fórmula cerrada para β_1 en la Ecuación (3.3) donde $\beta_1 = \frac{-1}{1-s}$ donde s, como se mencionó anteriormente, es la probabilidad empírica de que una prenda de la misma categoría sea vendida.

En la Figura 3.12 se observa el comportamiento de $q(r)$ para diferentes valores de r en el caso particular de la categoría *Vestidos*, donde $s = 0.695$ y asumiendo tres posibles r^* como se muestra en la leyenda. El punto rojo que se marca muestra que cuando $r = 1$, la probabilidad de venta es el s que se observa de manera empírica, lo que es consistente con lo que se mencionó anteriormente.

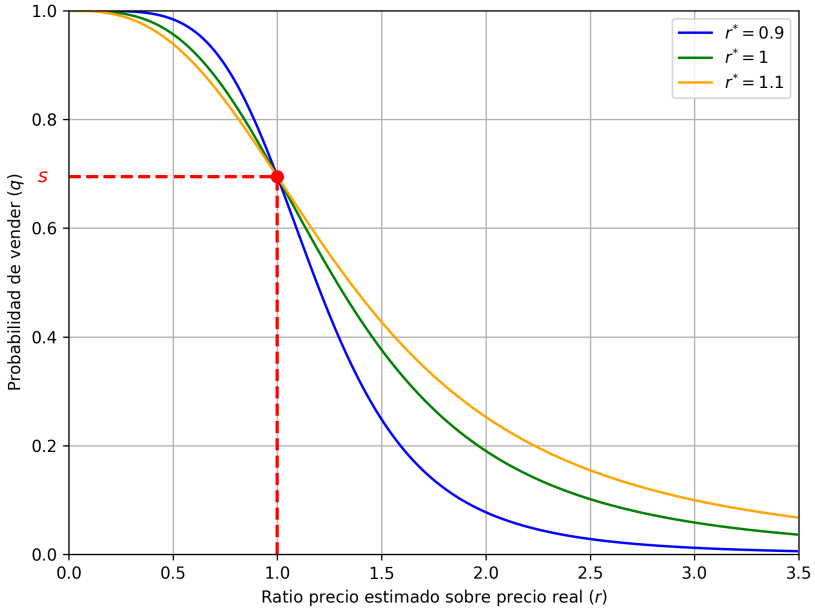

Figura 3.12: Curva de demanda como probabilidad de venta en función del ratio para la categoría *Vestidos*

Para finalizar, en el set de testeo se computará la diferencia porcentual de utilidades al usar los precios estimados en vez de los precios Market People para cada una de las categorías del estudio, donde el subíndice i hace referencia a cada uno de los ítems presentes en el set de testeo en la categoría bajo estudio:

$$\Delta U = \left(\frac{\sum_i \hat{p}_i \cdot q(\frac{\hat{p}_i}{p_i})}{\sum_i p_i \cdot \underbrace{q\,(1)}_{s}} - 1 \right) \cdot 100 \tag{3.4}$$

Donde la diferencia de utilidades ΔU dependerá de la categoría de prenda (ya que de esto dependerá el s) y el r^* que viene a ser el ratio entre el nivel de precios óptimo y el nivel de precios actual de Market People, por lo que se probarán diversos r^* considerando que puede que el nivel de precios óptimo esté sobre o por debajo del nivel de precios actual, viendo así cómo se alteran las utilidades en función de esto.

Utilizando la Expresión (3.4) en la categoría *Vestidos* y variando el r^* entre 0.9 y 1.2, se obtiene la Figura 3.13, que representa en el eje-x el r^* (ratio óptimo) y en el eje-y la diferencia porcentual de utilidades al usar los precios estimados en vez de los precios de Market People.

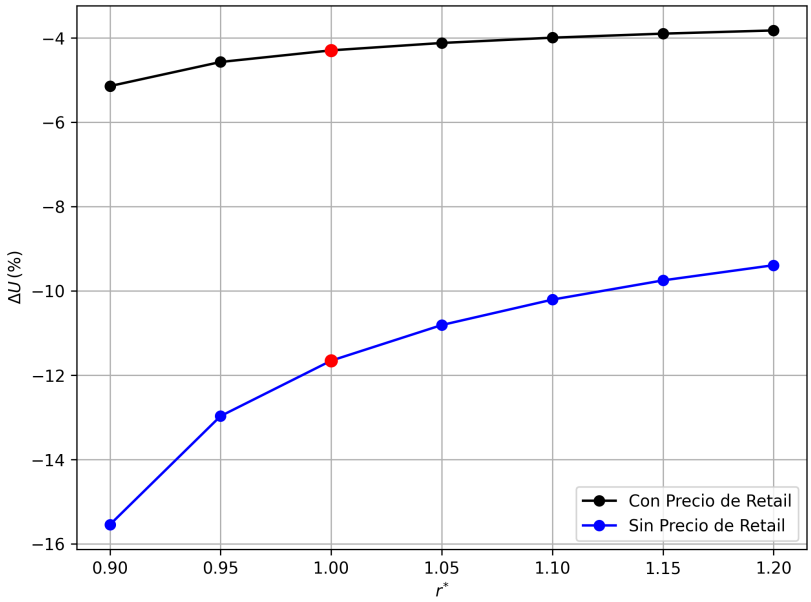

Figura 3.13: Diferencia de utilidades

Los puntos rojos de la Figura 3.13 corresponden a puntos donde se está asumiendo que el nivel de precios actual de Market People es, en efecto, el nivel óptimo.

Es interesante notar cómo decrece en magnitud esta diferencia porcentual de utilidades al incorporar el *precio de retail* como atributo. Esto viene a reforzar el hecho de que contar con este *precio de referencia* al ayudar enormemente a las predicciones también tiene un impacto en las utilidades que derivan de una buena puesta de precios. Notar que esto es obvio para el caso $r^* = 1$, no obstante, también se cumple numéricamente para otros valores de r^*.

Esto implica que usar *precios de retail* para generar estimaciones de los *precios de venta* tiene un impacto positivo en las utilidades de la empresa. Además, aún cuando la compañía no esté poniendo en la actualidad los precios que maximizan utilidades (línea vertical sobre la cual se ubican los puntos rojos), usar predicciones con *precio de retail* no genera tanta perdida al ser menos sensible al óptimo como el caso donde no se utiliza esta variable (que tiene una pendiente bastante más pronunciada), lo que le añade robustez al método y motiva la utilización de esta variable en la predicción.

En el Apéndice E, Figuras E.1 y E.2 se presentan tres gráficos para cada caso (con y sin *precio de referencia*) donde lo que varía es el valor de r^*. En función de esto se analiza brevemente el comportamiento de la probabilidad de venta y la utilidad esperada para diferentes valores de r^*, diferenciando por color la probabilidad de venta y el ingreso esperado, respectivamente.

Trabajo futuro: incidencia de la presencia de *precio de retail* en ventas

Una posible extensión de este trabajo es estudiar la incidencia de la presencia de *precio de retail* en las probabilidades de venta de las prendas. Para esto, se podría realizar un A/B testing donde para dos grupos de personas separadas de forma aleatoria se les muestren las mismas prendas, pero a un grupo se le muestra el *precio de retail* y al otro no. Luego, se podría medir la probabilidad de venta de las prendas para cada uno de los grupos y comparar si es que hay una diferencia significativa en las probabilidades de venta entre los dos grupos. Con esto, se podría cuantificar el impacto que tiene el *precio de retail* en las probabilidades de venta de las prendas, lo que sería de gran utilidad para la empresa, ya que le permitiría saber si es que vale la pena el esfuerzo de obtener este dato para todas las prendas, o si es que este esfuerzo no se ve reflejado en un aumento significativo en las probabilidades de venta de las prendas.

Se buscó realizar este análisis estudiando las ventas de prendas que poseen *precio de retail* y las que no, no obstante, se encontró que la gran mayoría de las prendas que poseen *precio de retail* no son consideradas *Is ganga*, lo que hace que sea imposible separar los efectos de la presencia de *precio de retail* y el hecho de que una prenda sea *Is ganga* en la probabilidad de venta del productos, pues como se mencionó, la gran mayoría de las prendas que poseen *precio de retail* no son consideradas *Is ganga* como se evidencia en la Figura 3.14.

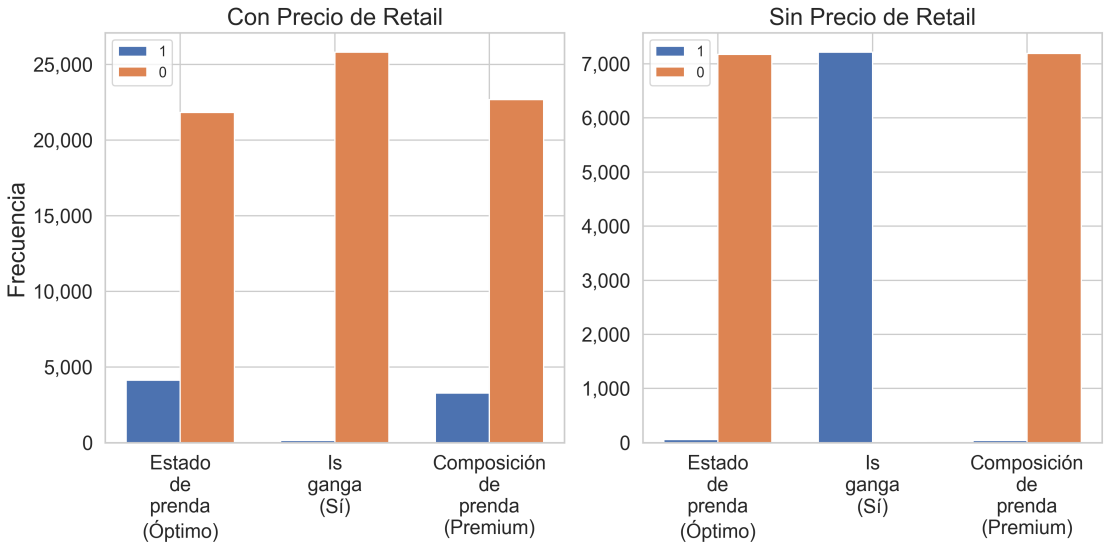

Figura 3.14: Características de los datos a utilizar

Este análisis podría ser de gran utilidad para la empresa, ya que se podría analizar si es que valdría la pena utilizar métodos avanzados de *web scraping* o inteligencia artificial, por ejemplo, para obtener el *precio de retail* de las prendas si es que se observa que este dato tiene un impacto significativo en las probabilidades de venta de las prendas.

Capítulo 4

Conclusiones

El trabajo realizado en esta tesis buscaba, principalmente, entregar mecanismos que permitan predecir precios de ropa de segunda mano en base a las características de estas. Esto, con la finalidad de que los vendedores de ropa de segunda mano puedan tener una referencia de precio que les permita setear un precio de forma más rápida y en consecuencia lograr así que publicar -y posteriormente vender- sus prendas sea un proceso más veloz, escalable y eficiente, logrando además ser consistente con los precios fijados anteriormente.

Para lograr esto se utilizó la base de datos de la compañía Market People, empresa dedicada principalmente a la venta de ropa de lujo de segunda mano, la cual contenía atributos de las prendas como su categoría, marca, composición, estado, entre otros, además de su *precio de venta* y en algunos casos su *precio de referencia*, también conocido como *precio de retail*.
Se decidió centrar el análisis en 8 categorías principales, por dos principales motivos: primero, porque estas categorías eran las que tenían mayor cantidad de datos y segundo, porque estas categorías son precisamente prendas, ya que la empresa vende también accesorios como bolsos, carteras, cinturones, artículos para el hogar como alfombras, entre otros. Luego, se realizó una separación de los datos en entrenamiento y testeo, se aplicó una transformación logarítmica a la variable *precio de venta*, para posteriormente utilizar una serie de diferentes métodos de *Machine Learning* para predecir el *precio de venta* de las prendas. Cada uno de estos métodos fue calibrado utilizando validación cruzada y se utilizó como métrica de error el *MAPE*, puesto que es una métrica que permite comparar errores relativos entre distintos modelos.

Los modelos, con sus respectivos hiperparámetros ajustados, fueron entrenados utilizando únicamente los datos de entrenamiento. Después, con los datos de testeo, cada método tenía que estimar el *precio de venta* de cada una de las prendas presentes en dicha base. Con esto, se calculó el *MAPE* de cada modelo y se compararon los resultados obtenidos. Los modelos utilizaron dos posibles bases de datos: una que contenía el atributo *precio de retail* y otra que no lo contenía. De esta forma, se buscaba simular la realidad, donde en algunos casos se tiene el *precio de retail* y en otros no, ya sea por un tema de asignación de recursos (pues toma tiempo obtener el dato) o por el hecho de que puede ser prácticamente imposible asignarle a prendas muy únicas.

Los resultados de las predicciones se separaron según su categoría, donde para la base de datos sin *precio de retail* los modelos que mejor predijeron fueron *Linear Regression* y *Support Vector Machine (SVM)* mientras que para la base de datos con *precio de retail* los modelos que mejor predijeron fueron *Random Forest* y *eXtreme Gradient Boosting (XGBoost)*, de manera transversal a las categorías, lo que le da consistencia a los resultados obtenidos. Además, la diferencia entre los MAPEs de los modelos con mejor capacidad predictiva al comparar el caso con *precio de retail* y sin *precio de retail* es de aproximadamente 10 %, lo que es un resultado que da cuenta de una importancia bastante alta del *precio de retail* a la hora de predecir el *precio de venta*. Para darle mayor soporte a esta idea, se utilizó el análisis de importancia de variables de los SHAP values para el caso del *Random Forest*, donde se pudo observar que el atributo *precio de retail* es el atributo más importante para predecir el *precio de venta*, lo que confirma que es por lejos el atributo con mayor importancia en la predicción.

A continuación se realizaron predicciones a nivel de intervalo, esto teniendo en cuenta la gran heterogeneidad que presentaban los datos, donde para un producto con los mismos atributos se podían tener múltiples *precios de ventas* muy diferentes entre sí. Para esto, se utilizó el modelo *Linear Regression*, puesto que es un modelo que permite predecir intervalos de confianza de manera cerrada. Los resultados obtenidos fueron bastante buenos, puesto que con un nivel de confianza buscado de un 80 % en promedio el 80 % de los *precios de venta* reales se encontraban dentro del intervalo de confianza predicho, el cual si bien era bastante amplio, es un resultado que sirve para darle una referencia del nivel de precio a los vendedores.

Junto con lo anterior, se cuanficó el impacto en las utilidades al utilizar las predicciones generadas a partir de los modelos en vez de los precios escogidos por la empresa. En este apartado se generó una curva de demanda como probabilidad de venta en función del ratio del precio estimado sobre el precio real. Utilizando los datos reales de la compañía se mostró que la diferencia de utilidad porcentual al utilizar modelos con *precio de retail* versus modelos sin *precio de retail* es bastante importante no únicamente en el caso de suponer que la empresa está poniendo los precios que maximizan su utilidad, si no que este resultado se extiende aún cuando los *precios de venta* actuales de la empresa no sean los óptimos. Para el caso con *precio de retail* la diferencia porcentual de utilidades oscila entre un -5 % y un -4 % y para el caso sin *precio de retail* va desde el -15 % hasta el -10 % aproximadamente.

Por último, se realizó una subsección para proponer un posible trabajo futuro, que consiste en estudiar la incidencia de la presencia de *precio de retail* en las probabilidades de venta de las prendas y, por consiguiente, en las utilidades de la empresa. Para esto existen diversas metodologías que se podrían utilizar, como por ejemplo un A/B testing, para estudiar si es que el *precio de retail* tiene un impacto significativo en las probabilidades de venta de las prendas y en consecuencia en la utilidad esperada de la empresa, lo cual con los datos actuales no es posible concluir debido a la imposibilidad de separar los efectos.

En resumen, este estudio ha abordado con datos reales la tarea de predecir precios de ropa de segunda mano mediante el uso de modelos de *Machine Learning*, destacando la importancia del atributo *precio de retail* en la mejora de la capacidad predictiva. Aunque los resultados son buenos, existen áreas de mejora y posibles direcciones para futuras investigaciones. Se podría considerar la incorporación de más atributos relacionados con tendencias de moda, estacionalidad o eventos especiales. Además, se podría considerar el procesamiento de imágenes de las prendas para incorporarlas como atributos en los modelos, o también la utilización de modelos de deep learning que incorporen embeddings de texto para atributos como la descripción de la prenda o el nombre de la marca podría ser una línea de investigación interesante para trabajos futuros.

Lo realizado en esta tesis tiene consecuencias directas en la gestión de la empresa. La capacidad de predecir precios de venta de manera más precisa y rápida, y la posibilidad de generar intervalos de confianza para estas predicciones, permitirá a la empresa ser más

eficiente en la fijación de precios, lo que se traduce en una mayor eficiencia en la publicación de prendas, liberando de esta forma recursos que podrían ser utilizados en otras áreas de la empresa que requieran factores donde la creatividad humana sea más relevante. Esto no quiere decir que de ahora en más la tarea de fijación de precios quede relevada a los modelos predictivos, sino que estos modelos pueden ser utilizados como una herramienta que permita a los encargados de la tarea dedicarse más bien a supervisar y ajustar los precios que los modelos generen, en vez de tener que generarlos desde cero. Esta forma de operar se ha estudiado y ha mostrado ser eficiente y escalable [19], permitiendo a la empresa optimizar de mejor manera sus recursos y ser más competitiva en el mercado.

Bibliografía

[1] Hammond, K. R., Human Judgement and Social Policy: Irreducible Uncertainty, Inevitable Error, Unavoidable Injustice. Oxford University Press USA, 1996.

[2] Breton, R. y Bosse, E., "The cognitive costs and benefits of automation," p. 13, 2003.

[3] Phan, T. D., "Housing price prediction using machine learning algorithms: The case of melbourne city, australia," en 2018 International Conference on Machine Learning and Data Engineering (iCMLDE), (Sydney, Australia), Macquarie University, 2018.

[4] Park, B. y Bae, J. K., "Using machine learning algorithms for housing price prediction: The case of fairfax county, virginia housing data," Expert Systems with Applications, vol. 42, no. 6, pp. 2928–2934, 2015, doi:10.1016/j.eswa.2014.11.040.

[5] Tchuente, D. y Nyawa, S., "Real estate price estimation in french cities using geocoding and machine learning," Annals of Operations Research, vol. 308, pp. 571–608, 2022, doi:10.1007/s10479-021-03932-5.

[6] Listiani, M., "Support vector regression analysis for price prediction in a car leasing application," 2009.

[7] Wu, J. D., Hsu, C. C., y Chen, H. C., "An expert system of price forecasting for used cars using adaptive neuro-fuzzy inference," Expert Systems with Applications, vol. 36, no. 4, pp. 7809–7817, 2009.

[8] Peerun, S., Chummun, N. H., y Pudaruth, S., "Predicting the price of second-hand cars using artificial neural networks," en The Second International Conference on Data Mining Internet Computing and Big Data, pp. 17–21, 2015.

[9] Gegic, E., Isakovic, B., Keco, D., Masetic, Z., y Kevric, J., "Car price prediction using machine learning techniques," TEM Journal, vol. 8, pp. 113–118, 2019, doi:10.18421/T

EM81-16.

[10] Frank, C., Garg, A., Sztandera, L., y Raheja, A., "Forecasting women's apparel sales using mathematical modeling," International Journal of Clothing Science and Technology, vol. 15, no. 2, pp. 107–125, 2003, doi:10.1108/09556220310470097.

[11] Yu, S., Dong, H., Chen, Y., He, Z., y Shi, X., "Clothing sales forecast based on arima-bp neural network combination model," en 2019 IEEE International Conference on Power, Intelligent Computing and Systems (ICPICS), (Shenyang, China), pp. 367–372, 2019, doi:10.1109/ICPICS47731.2019.8942427.

[12] Li, Y., Yang, Y., Zhu, K., y Zhang, J., "Clothing sale forecasting by a composite gru–prophet model with an attention mechanism," IEEE Transactions on Industrial Informatics, vol. 17, pp. 8335–8344, 2021, doi:10.1109/TII.2021.3057922.

[13] Fathalla, A., Salah, A., Li, K., Li, K., y Francesco, P., "Deep end-to-end learning for price prediction of second-hand items," Knowledge and Information Systems, vol. 62, no. 4, pp. 4541–4568, 2020, doi:10.1007/s10115-020-01495-8.

[14] Munaweera, R. P., "Sales price prediction for women's innerwear," 2019.

[15] Katai, Y. y Hasuike, T., "Sales price prediction of used clothes considering the market price of flea market application," en 2019 8th International Congress on Advanced Applied Informatics (IIAI-AAI), 2019, doi:10.1109/iiai-aai.2019.00154.

[16] Benoit, K., "Linear regression models with logarithmic transformations," vol. 22, no. 1, pp. 23–36, 2011.

[17] Miami, M., "Ramona larue." = https://www.myguidemiami.com/es/compras/ramona-larue,, 2023.

[18] Polimis, K., Rokem, A., y Hazelton, B., "Confidence intervals for random forests in python," Journal of Open Source Software (JOSS), 2019, https://doi.org/10.21105/joss.00124. eScience Institute, University of Washington.

[19] Haight, J. M. y Kecojevic, V., "Automation vs. human intervention: What is the best fit for the best performance?," Process Safety Progress, vol. 24, no. 1, pp. 45–51, 2005.

Anexos

Anexo A. Tablas Anexas

Tabla A.1: Sub Categorías por Categorías

Categorías	**Sub Categorías**	**Composiciones Premium**
Abrigos / Chaquetas	Chaquetas, Blazer, Abrigos, Cuero, Parcas, Denim, Kimonos, Bordadas, Impermeables, Trench, Sobrecamisas, Tapados, Bomber, Tweed, Peludos, Gilet, Montgomery, Capas, Ligeras	Seda, Cuero, Lana, Cashmere, Poliamida, Piel
Blusas	Manga larga, Sin mangas, Manga corta, Crop tops	Seda, Cuero
Calzado	Zapatillas, Botines, Sandalias, Zapatos, Mules, Mocasines, Botas cowboy, Alpargatas, Botas, Bototos, Botas largas, Botas de agua, Terraplen, Flats, Salón, Babuchas, Suecos, Niño/a	Cuero
Faldas	Mini, Maxi, Midi	Seda, Cuero, Lana
Pantalones	Denim, De tela, Leggins, Buzos, Cuero, Culotte, De vestir, Palazzos, Cotelé, Bombachos, Maternales, Jeans	Seda, Cuero
Poleras	Manga corta, Manga larga, Sin mangas, Crop tops, Deportivas	Seda, Cuero
Tejidos / Sweaters	Cerrados, Abiertos, Ponchos, Sweater	Alpaca, Lana
Vestidos	Mini, Maxi, Midi, Sin Mangas, Manga corta, Salidas de baño, Manga larga, Novias, Fiesta, Con mangas, Kimonos	Seda, Cuero, Lana

Tabla A.2: Tabla descriptiva por categoría

Prenda	Pre procesamiento	Post procesamiento	Is ganga (Sí)	Estado de prenda (Óptimo)	Composición de prenda (Premium)
Blusas	2.864	2.150	1 %	13 %	10 %
Poleras	2.089	1.493	1 %	15 %	1 %
Pantalones	3.169	2.609	1 %	15 %	3 %
Vestidos	3.950	3.007	1 %	16 %	7 %
Calzado	1.663	1.485	1 %	33 %	31 %
Faldas	962	751	1 %	15 %	11 %
Tejidos / Sweaters	1.731	1.433	1 %	13 %	12 %
Abrigos / Chaquetas	3.142	2.656	1 %	11 %	23 %

Tabla A.3: Hiperparámetros testeados por método

Modelo	Hiperparámetro	Opciones testeadas
XGBoost	*n_estimators*	100, 200, **300**
	max_depth	3, 4, **5**
	learning_rate	**0.1**, 0.01, 0.001
KNN	*n_neighbors*	3, 5, 7, **9**
	weights	uniform, **distance**
LASSO	*alpha*	np.logspace(-4, 4, 100): **0.0001**
LightGBM	*n_estimators*	**100**, 200, 300
	max_depth	**-1**, 3, 4, 5
	learning_rate	**0.1**, 0.01, 0.001
Neural Networks	*hidden_layer_sizes*	(64, 32), (128, 64), **(64, 64, 32)**
	activation	relu, **tanh**
	solver	adam, **lbfgs**
	learning_rate	**adaptive**, invscaling
RF	*n_estimators*	50, **100**, 150
	max_depth	**None**, 5, 10, 20
	min_samples_split	2, 5, **10**
	min_samples_leaf	**1**, 2, 4
SVM	*C*	0.1, **1**, 10
	kernel	**linear**, poly, rbf
	degree	**2**, 3, 4
	gamma	**scale**, auto

*Destacados en **verde** se encuentran los hiperparámetros escogidos.*

Anexo B. Histograma de precios sin aplicar transformación logarítmica

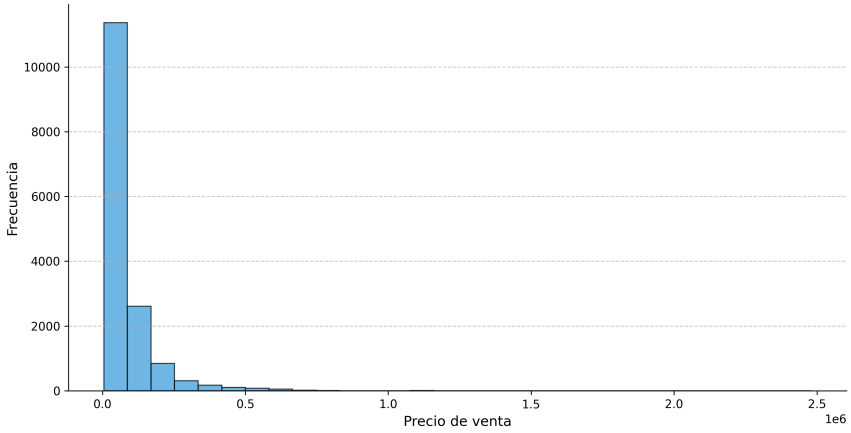

Figura B.1: Distribución de *precio de venta* sin transformación logarítmica

Anexo C. Resultados del MAPE al utilizar diferentes modelos

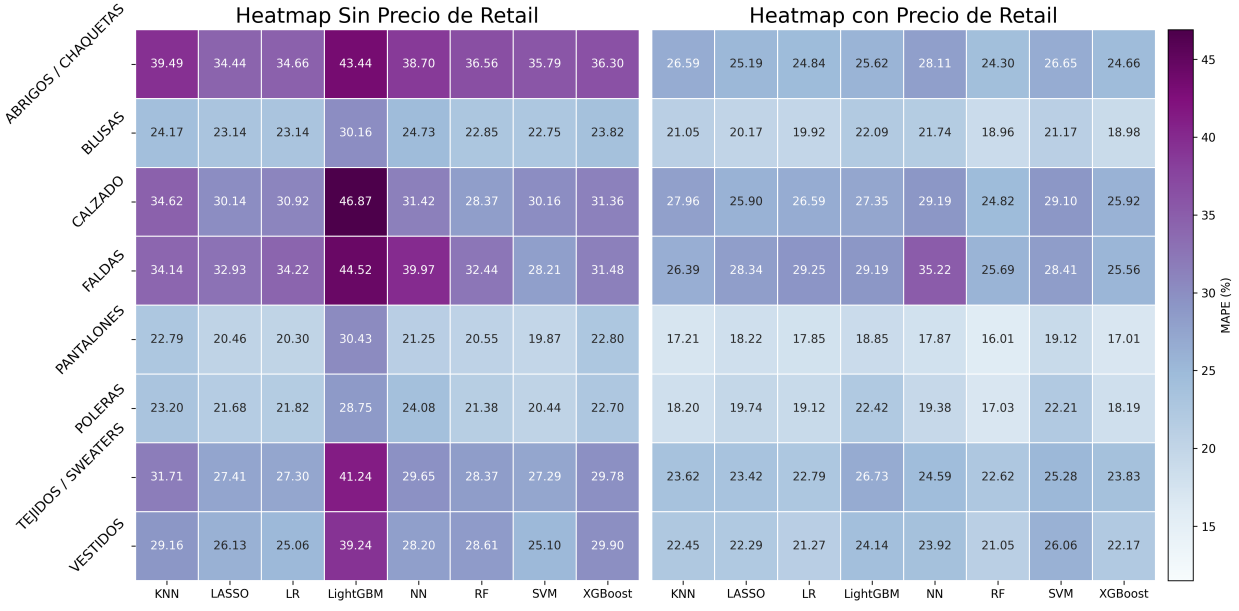

Figura C.1: Heatmap que muestra los *MAPE* para todas las categorías y métodos utilizados cuando la variable Precio de Retail es logarítmica

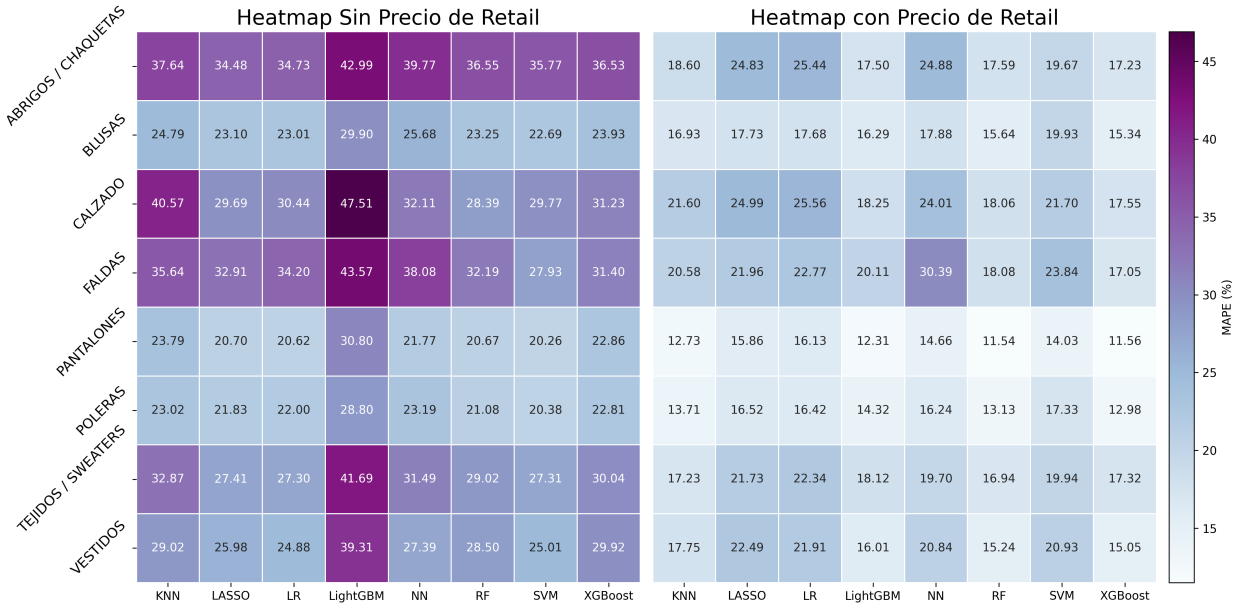

Figura C.2: Heatmap que muestra los *MAPE* para todas las categorías y métodos utilizados cuando la variable Estado no es agrupada en niveles

Anexo D. Resultados del uso de *Random Forest* para intervalos de precio

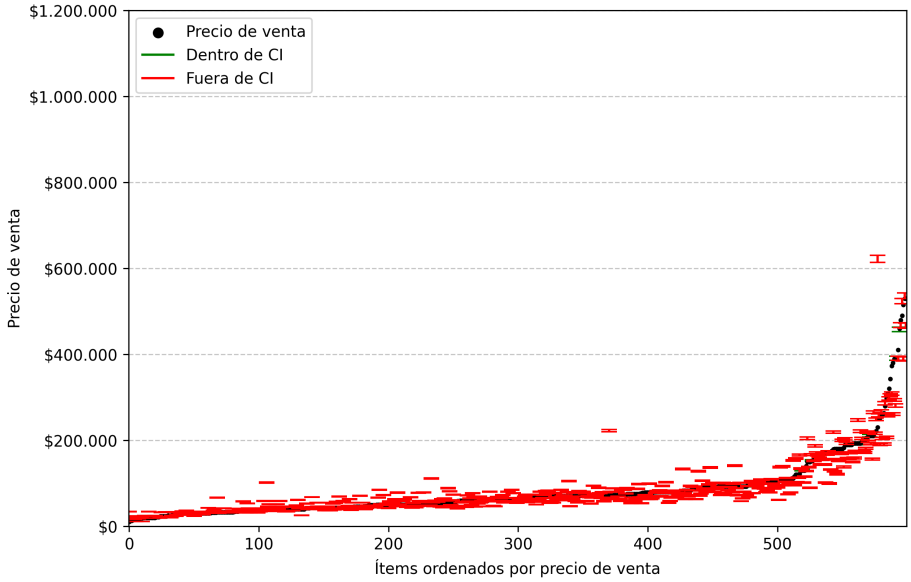

Figura D.1: Intervalos de confianza generados por Random Forest para la categoría *Vestidos*, utilizando *precio de retail*

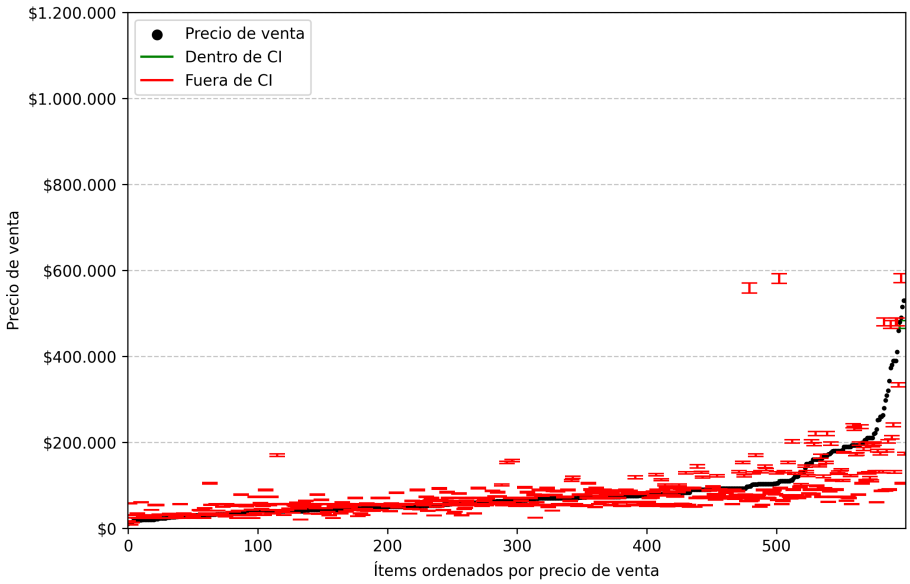

Figura D.2: Intervalos de confianza generados por Random Forest para la categoría *Vestidos*, sin utilizar *precio de retail*

Anexo E. Gráficos de probabilidad y utilidad esperada

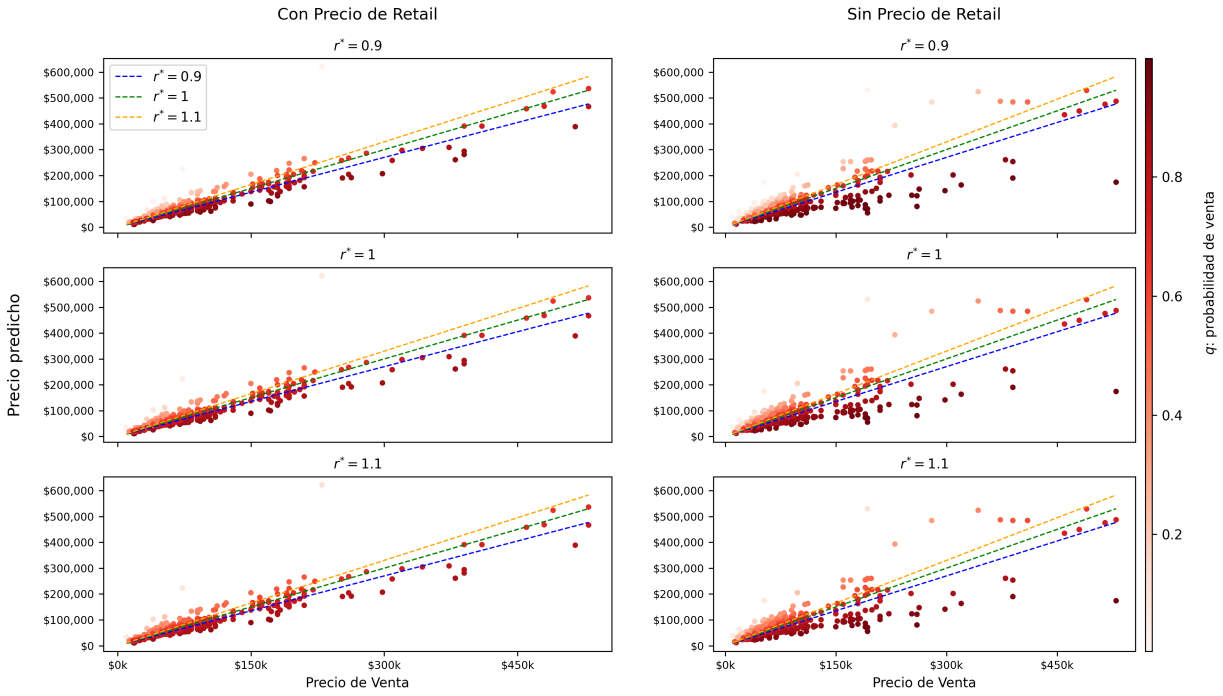

Figura E.1: Precio predicho vs *precio de venta*. Color indica probabilidad de venta

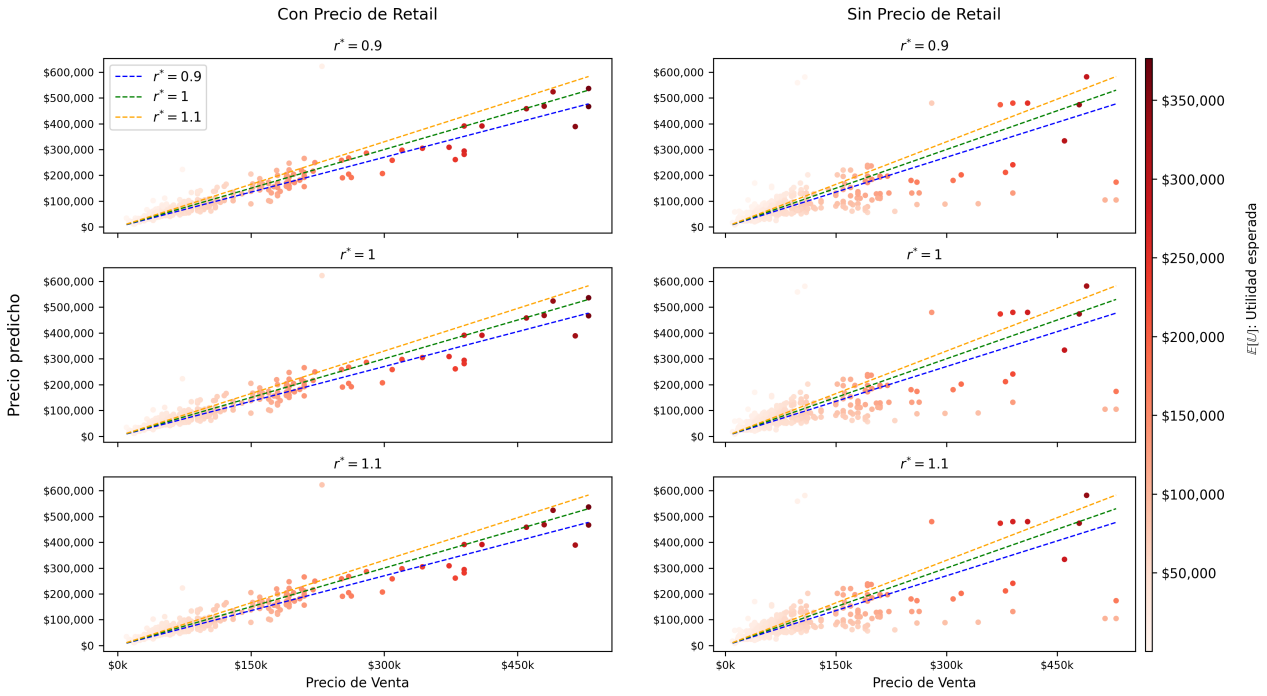

Figura E.2: Precio predicho vs *precio de venta*. Color indica ingreso esperado

En la Figura E.1 se ve que la probabilidad de venta mejora al ser mayor el r^*. Esto se explica ya que, como se puede ver en la Figura 3.12, un r^* mayor permite que la curva que relaciona el ratio con la probabilidad de venta se desplace hacia la derecha, lo que implica que, pasado el $r = 1$, para un mismo ratio, la probabilidad de venta sea mayor para las curvas con r^* más alto. A diferencia de lo que ocurre entre el 0 y $r = 1$, donde se observa una diferencia sutil en la probabilidad de venta de las tres líneas, para $r > 1$ se observa una diferencia mucho más marcada que termina afectando de manera importante la utilidad esperada, en particular a ítems que poseen un precio predicho considerablemente mayor al *precio de venta* de Market People, los que verán cómo a medida que crece el r^* su probabilidad de venta asignada (y en consecuencia su utilidad esperada) será mayor.

De la Figura E.2 se desprende que para cuando se incluye *precio de retail* casi no se observa diferencia en la esperanza de la utilidad entre los diferentes niveles de r^* (lo que hace sentido con la curva negra de la Figura 3.13, cuya pendiente es muy poco pronunciada). Para el caso sin *precio de retail* sí se observa de manera clara para algunos puntos (como los tres que están sobre las líneas puntedas, alrededor de los $400.000) que la utilidad esperada mejora

cuando el r^* crece.