MÁSTER UNIVERSITARIO EN INGENIERÍA INDUSTRIAL

TRABAJO FIN DE MÁSTER

PREDICCIÓN DE DEVOLUCIONES EN EL SECTOR TEXTIL (VENTAS E-COMMERCE)

Autor: Laura González Legall
Director: Álvaro de la Cruz Sánchez de Rojas

Madrid

Julio de 2020

Declaro, bajo mi responsabilidad, que el Proyecto presentado con el título

Predicción de devoluciones en el sector textil (ventas e-commerce)

en la ETS de Ingeniería - ICAI de la Universidad Pontificia Comillas en el

curso académico 2019/2020 es de mi autoría, original e inédito y

no ha sido presentado con anterioridad a otros efectos.

El Proyecto no es plagio de otro, ni total ni parcialmente y la información que ha sido tomada de otros documentos está debidamente referenciada.

Fdo.: Laura González Legall Fecha: 15/07/2020

Autorizada la entrega del proyecto

EL DIRECTOR DEL PROYECTO

Fdo.: Álvaro de la Cruz Fecha: 15/07/2020

AUTORIZACIÓN PARA LA DIGITALIZACIÓN, DEPÓSITO Y DIVULGACIÓN EN RED DE PROYECTOS FIN DE GRADO, FIN DE MÁSTER, TESINAS O MEMORIAS DE BACHILLERATO

1º. Declaración de la autoría y acreditación de la misma.

El autor D. **Laura González Legall**

DECLARA ser el titular de los derechos de propiedad intelectual de la obra: **"Predicción de devoluciones en el sector textil (ventas e-commerce)"**, que ésta es una obra original, y que ostenta la condición de autor en el sentido que otorga la Ley de Propiedad Intelectual.

2º. Objeto y fines de la cesión.

Con el fin de dar la máxima difusión a la obra citada a través del Repositorio institucional de la Universidad, el autor **CEDE** a la Universidad Pontificia Comillas, de forma gratuita y no exclusiva, por el máximo plazo legal y con ámbito universal, los derechos de digitalización, de archivo, de reproducción, de distribución y de comunicación pública, incluido el derecho de puesta a disposición electrónica, tal y como se describen en la Ley de Propiedad Intelectual. El derecho de transformación se cede a los únicos efectos de lo dispuesto en la letra a) del apartado siguiente.

3º. Condiciones de la cesión y acceso

Sin perjuicio de la titularidad de la obra, que sigue correspondiendo a su autor, la cesión de derechos contemplada en esta licencia habilita para:

a) Transformarla con el fin de adaptarla a cualquier tecnología que permita incorporarla a internet y hacerla accesible; incorporar metadatos para realizar el registro de la obra e incorporar "marcas de agua" o cualquier otro sistema de seguridad o de protección.
b) Reproducirla en un soporte digital para su incorporación a una base de datos electrónica, incluyendo el derecho de reproducir y almacenar la obra en servidores, a los efectos de garantizar su seguridad, conservación y preservar el formato.
c) Comunicarla, por defecto, a través de un archivo institucional abierto, accesible de modo libre y gratuito a través de internet.
d) Cualquier otra forma de acceso (restringido, embargado, cerrado) deberá solicitarse expresamente y obedecer a causas justificadas.
e) Asignar por defecto a estos trabajos una licencia Creative Commons.
f) Asignar por defecto a estos trabajos un HANDLE (URL *persistente)*.

4º. Derechos del autor.

El autor, en tanto que titular de una obra tiene derecho a:

a) Que la Universidad identifique claramente su nombre como autor de la misma
b) Comunicar y dar publicidad a la obra en la versión que ceda y en otras posteriores a través de cualquier medio.
c) Solicitar la retirada de la obra del repositorio por causa justificada.
d) Recibir notificación fehaciente de cualquier reclamación que puedan formular terceras personas en relación con la obra y, en particular, de reclamaciones relativas a los derechos de propiedad intelectual sobre ella.

5º. Deberes del autor.

El autor se compromete a:

a) Garantizar que el compromiso que adquiere mediante el presente escrito no infringe ningún derecho de terceros, ya sean de propiedad industrial, intelectual o cualquier otro.
b) Garantizar que el contenido de las obras no atenta contra los derechos al honor, a la intimidad y a la imagen de terceros.
c) Asumir toda reclamación o responsabilidad, incluyendo las indemnizaciones por daños, que pudieran ejercitarse contra la Universidad por terceros que vieran infringidos sus derechos e intereses a causa de la cesión.

d) Asumir la responsabilidad en el caso de que las instituciones fueran condenadas por infracción de derechos derivada de las obras objeto de la cesión.

6º. Fines y funcionamiento del Repositorio Institucional.
La obra se pondrá a disposición de los usuarios para que hagan de ella un uso justo y respetuoso con los derechos del autor, según lo permitido por la legislación aplicable, y con fines de estudio, investigación, o cualquier otro fin lícito. Con dicha finalidad, la Universidad asume los siguientes deberes y se reserva las siguientes facultades:

- La Universidad informará a los usuarios del archivo sobre los usos permitidos, y no garantiza ni asume responsabilidad alguna por otras formas en que los usuarios hagan un uso posterior de las obras no conforme con la legislación vigente. El uso posterior, más allá de la copia privada, requerirá que se cite la fuente y se reconozca la autoría, que no se obtenga beneficio comercial, y que no se realicen obras derivadas.
- La Universidad no revisará el contenido de las obras, que en todo caso permanecerá bajo la responsabilidad exclusive del autor y no estará obligada a ejercitar acciones legales en nombre del autor en el supuesto de infracciones a derechos de propiedad intelectual derivados del depósito y archivo de las obras. El autor renuncia a cualquier reclamación frente a la Universidad por las formas no ajustadas a la legislación vigente en que los usuarios hagan uso de las obras.
- La Universidad adoptará las medidas necesarias para la preservación de la obra en un futuro.
- La Universidad se reserva la facultad de retirar la obra, previa notificación al autor, en supuestos suficientemente justificados, o en caso de reclamaciones de terceros.

Madrid, a 15 de Julio de 2020

ACEPTA

Fdo 

Motivos para solicitar el acceso restringido, cerrado o embargado del trabajo en el Repositorio Institucional:

MÁSTER UNIVERSITARIO EN INGENIERÍA INDUSTRIAL

TRABAJO FIN DE MÁSTER

PREDICCIÓN DE DEVOLUCIONES EN EL SECTOR TEXTIL (VENTAS E-COMMERCE)

Autor: Laura González Legall
Director: Álvaro de la Cruz Sánchez de Rojas

Madrid

Julio de 2020

PREDICCIÓN DE DEVOLUCIONES EN EL SECTOR TEXTIL (VENTAS E-COMMERCE)

Autor: González Legall, Laura.

Director: Cruz Sánchez de Rojas, Álvaro de la.

Entidad Colaboradora: PricewaterhouseCoopers.

RESUMEN DEL PROYECTO

Palabras clave: Devoluciones, Predicción, E-commerce.

Introducción

Planteamiento del problema

La aparición del Internet revolucionó sin duda alguna todas las industrias, siendo el sector retail uno de los más transformados. Este no solo ha aprovechado el auge del e-commerce para ir más allá de su oferta física a través de la personalización de catálogos de ropa, artículos y ofertas disponibles solamente online… si no también como fuente de datos sobre los consumidores, útiles para ser analizados y empleados posteriormente mediante diferentes iniciativas que aumenten la ventaja competitiva.

Se espera que este impacto siga creciendo. De hecho, según el último reporte de Forrester, para 2022 las ventas online van a suponer el 36% del total [1]. Sin embargo, los costes de logística inversa suponen un obstáculo para las compañías, pues se calcula que suponen de un 2% a un 4% de su facturación total [2]. Es por eso que, con el fin de aumentar el margen de la empresa, una de sus prioridades es optimizarla sin deteriorar la satisfacción del cliente que acompañan las devoluciones bien gestionadas.

La mejora del desempeño financiero de la compañía, el aprovechamiento más eficiente de recursos, la reducción de inventarios, la gestión de stock… son algunos de los beneficios de la optimización de la logística inversa. Una de las herramientas que puede contribuir potencialmente es la identificación de patrones y la predicción de devoluciones, para luego poder implantar barreras que las reduzcan.

Es importante mencionar que la inteligencia artificial (IA) ha contribuido notoriamente en el avance del sector retail mediante el desarrollo de distintas iniciativas, como por ejemplo:

- Programas basados en la conversación.
- Búsqueda de artículos similares.
- Personalización sobre la marcha de ofertas de compra.
- Optimización de la cadena de suministros.
- Predicción de demanda y posterior definición de los niveles de stock.
- Establecimiento automático de precios basado en datos internos y externos.

Objetivos

Enfocándonos en la situación previamente explicada, este proyecto tiene como objetivo principal el desarrollo de un modelo que sea capaz de predecir las devoluciones en el mundo del retail, en concreto en el sector textil y con respecto a las ventas e-commerce, con el fin de utilizar esta información para reducir los costes que la logística inversa tiene en la compañía.

Para conseguir este objetivo, se llevará a cabo un análisis de los consumidores, de los productos y de las transacciones, con el fin de agruparlas y determinar patrones que posibiliten identificar qué productos tienen más probabilidad de ser devueltos.

Una vez obtenidos los resultados, se planteará un caso de negocio. En él se estudiarán diferentes estrategias, y se analizará el impacto económico que la implantación de cada una de ellas tiene en una empresa, con el fin de posteriormente proponer las más factibles.

Metodología

Para cumplir con los objetivos de este trabajo, se investigará sobre las políticas de devolución de las grandes empresas de retail de hoy en día, implementaciones de machine learning para optimizarlas, el impacto de la logística inversa en los márgenes comerciales… Esta información se abstraerá de diferentes fuentes como Google Scholar, EBSCO, Academia…

A su vez, se analizará una base de datos que simulan a los de una empresa del sector moda utilizando Python. Se analizarán variables con el fin de identificar patrones de devolución, estudiar el comportamiento del consumidor según diferentes parámetros…

Tras analizar los datos se elaborará el modelo predictivo con el mismo lenguaje de programación, Python, utilizando el módulo de inteligencia artificial h2o. En concreto, la tipología del modelo es GBM. Se seleccionó este porque es capaz de obtener resultados cada vez mejores gracias a aproximaciones progresivamente refinadas mediante árboles de decisión.

Es importante mencionar que los datos se dividen en tres conjuntos distintos: training, validation y test. El modelo observa y aprende del primero de ellos, ajusta sus parámetros según los resultados con el segundo, y finalmente comprueba su eficacia con el último.

Por último, como ya se ha mencionado se llevará a cabo un análisis económico de diferentes estrategias que se propondrán para reducir el impacto que la logística inversa tiene en la compañía.

Resultados

A la hora de evaluar la validez del modelo propuesto, se seleccionaron diferentes métricas teniendo en cuenta la naturaleza de los datos y de los resultados, y el objetivo final del proyecto. Estas son misclassification, precision y recall. Los valores obtenidos para cada conjunto son los siguientes:

	Entrenamiento	**Validación**	**Test**
Misclassification	0,07%	0,11%	0,09%
Precision	99,99%	99,98%	99,95%
Recall	99,87%	99,81%	99,86%

En base a estos resultados, se puede afirmar:

- Los resultados de '*misclassification'* son muy inferiores en los tres casos. Esto significa que el error general que el modelo comete es bajo.
- La métrica '*precision'* toma valores muy elevados, lo que indica que un porcentaje muy alto de productos estimados que se van a devolver, en realidad serán devueltos, es decir, la calidad del modelo en cuanto a clasificación es excelente.
- Por último, tal y como se puede observar los valores de '*recall'* son también elevados. Esto indica no solo que el modelo es capaz de identificar los relativamente pocos casos positivos, si no también que lo hace de manera precisa.

Por otro lado, se proponen diferentes iniciativas a corto y a largo plazo para reducir el impacto que tienen los costes de logística inversa en el margen de las empresas. Sin embargo, el estudio de la viabilidad económica se centró en las estrategias que la empresa podría implementar a corto plazo, que son las que se indican a continuación.

- Estrategia 1: Que el cliente pague una cierta cantidad de dinero por producto si el modelo predice que éste será devuelto.

	3,00 €	**5,00 €**	**6,00 €**	**7,00 €**	**9,20 €**
Ingreso extra al año	52.980,32 €	88.300,53 €	105.960,64 €	123.620,75 €	162.499,77 €
Ingreso extra al mes	4.075,41 €	6.792,35 €	8.150,82 €	9.509,29 €	12.499,98 €
Payback (meses)	36,81	22,08	18,40	15,77	12,00
Payback (años)	3,1	1,8	1,5	1,3	1,0

Tabla 1. Análisis económico de la estrategia 1

- Estrategia 2: Que el cliente pague un porcentaje de lo que cuesta el producto (según su categoría) si se estima que éste será devuelto.

Categoría	5%	10%	15%	20%
Ingreso extra al año	36.697,59 €	73.395,19 €	110.092,78 €	146.790,38 €
Ingreso extra al mes	2.822,89 €	5.645,78 €	8.468,68 €	11.291,57 €
Payback (meses)	53,1	26,6	17,7	13,3
Payback (años)	4,4	2,2	1,5	1,1

Tabla 2. Análisis económico de la estrategia 2

- Estrategia 3: Que el cliente pague un porcentaje de lo que cuesta la devolución del producto (según su categoría) si se estima que éste será devuelto.

Categoría	10%	20%	30%	40%
Ingreso extra al año	42.156,70 €	84.313,40 €	126.470,10 €	168.626,80 €
Ingreso extra al mes	3.242,82 €	6.485,65 €	9.728,47 €	12.971,29 €
Payback (meses)	46,3	23,1	15,4	11,6
Payback (años)	3,9	1,9	1,3	1,0

Tabla 3. Análisis económico de la estrategia 3

Conclusiones

Para determinar qué estrategias de todas las propuestas se proponen como iniciativas para que la empresa reduzca el impacto de las devoluciones, se ha tenido en cuenta:

- El ingreso adicional anual que supondrían.
- El coste de las devoluciones.
- El tiempo de retorno de la inversión inicial.
- La posibilidad y disposición de los clientes a pagar dinero extra que contrarreste los costes.

Tras analizar la viabilidad económica de cada una de ellas, se ha determinado que la más factible, y por lo tanto se propone como solución a los costes de logística inversa, es la estrategia 3. Esta consiste en que el cliente pague una penalización del **40%** del coste de devolución según la categoría a la que pertenezca el producto. El retorno de inversión (ROI) sería un **12,42%.** Además, los flujos de caja mensuales que se obtendrían se representan en la Tabla 4.

Diciembre 2017	14.304,64 €
Enero 2018	14.671,07 €
Febrero 2018	10.279,22 €
Marzo 2018	12.231,80 €
Abril 2018	10.187,10 €
Mayo 2018	5.454,28 €
Junio 2018	7.926,65 €
Julio 2018	24.987,84 €
Agosto 2018	11.302,37 €
Septiembre 2018	11.742,36 €
Octubre 2018	6.641,11 €
Noviembre 2018	11.459,02 €
Diciembre 2018	21.366,25 €
TOTAL	162.553,72 €

Tabla 4. Flujos de caja mensuales de la estrategia propuesta

Además, para proporcionar una visión más clara del impacto que tendría esta estrategia en los clientes, se calculó la penalización a la que estaría sujeto cada uno de ellos. Esta sería de **3,10 €** al año. Esto quiere decir que bastaría con añadir alrededor de 0,50 € por transacción como coste de envío, por ejemplo, si se supone que cada cliente realiza 6 compras al año.

Referencias

[1] Catchoom, "eCommerce In The Fashion Industry. Industry Changes, Stats & Trends. Image Recognition, AR and Artificial Intelligence Solutions," 2018. [Online]. Disponible: https://catchoom.com/blog/how-is-ecommerce-changing-fashion-industry-stats-trends-predictions/?cn-reloaded=1.

[2] W. Kofler, "Artificial Intelligence in Retail – What to expect and how to act," PwC. [Online]. [Accessed 2019].

[3] F. Ma, "The Study on Reverse Logistics for E-Commerce," IEEE Conference Publication, 2010. [Online]. Available: https://ieeexplore.ieee.org/document/5575577 .

PREDICTION OF RETURNS IN THE TEXTILE SECTOR (E-COMMERCE SALES)

Author: González Legall, Laura.

Director: Cruz Sánchez de Rojas, Álvaro de la.

Collaborating Entity: PricewaterhouseCoopers.

EXECUTIVE SUMMARY

Keywords: Returns, Predict, E-commerce.

Introduction

Problem Statement

The emergence of the Internet revolutionized all industries, being the retail sector one of the most transformed. This industry has not only taken advantage of the e-commerce boom to go beyond its physical offer through the personalization of clothing catalogues, articles and offers available only online... but also as a source of data about consumers, which are analyzed and used later through different initiatives that increase competitive advantage.

This impact is expected to continue to grow. In fact, according to Forrester's latest report, by 2022 online sales will account for 36% of the total [1]. However, reverse logistics costs are a burden for companies. They account for 2% to 4% of their total turnover [2]. That is why, in order to increase the company's margin, one of its priorities is to optimize it without deteriorating the customer satisfaction that entails well-managed returns.

As it has been already mentioned, the improvement of the company's financial performance, more efficient use of resources, reduction of inventories and stock management... are some of the benefits of the optimization of reverse logistics. One of the tools that can potentially contribute to this is the identification of patterns and the prediction of returns, so barriers can be later implemented to reduce them.

It is important to mention that artificial intelligence (AI) has contributed significantly to the progress of the retail sector by developing different initiatives, such as:

- Conversation-based programs.
- Search for similar garments.
- On-the-fly personalization of purchase offers.
- Optimization of the supply chain.
- Demand forecasting and subsequent definition of stock levels.
- Automatic pricing based on internal and external data.

Objectives

Focusing on the situation previously explained, this project has as its main objective the development of a model that is capable of predicting returns in the world of retail, specifically in the textile sector and with respect to e-commerce sales. This information will be used to reduce the costs that reverse logistics cause in the company.

To achieve this objective, an analysis of consumers, products and transactions will be carried out in order to group them together and determine patterns that make it possible to identify which products are most likely to be returned.

Once the results are obtained, a business case will be presented. Different strategies will be studied, and the economic impact that the implementation of each one of them has on a company will be analyzed, in order to propose the most feasible ones as solutions.

Methodology

To meet the objectives of this project, research will be conducted on the return policies of large retail companies today, machine learning implementations to optimize them, the impact of reverse logistics on the finances of a company... This information will be abstracted from different sources such as Google Scholar, EBSCO, Academia...

At the same time, a database simulating those of a company in the fashion sector will be studied using the programming language Python. Variables will be analyzed in order to identify return patterns, study consumer behavior according to different parameters...

After analyzing the data, the predictive model will be developed with the same programming language, Python, using the h2o artificial intelligence module. Specifically, the model typology is GBM. This was selected because it is capable of obtaining increasingly better results thanks to progressively refined approaches using decision trees.

It is important to mention that the data is divided into three different sets: training, validation and test. The model observes and learns from the first one, adjusts its parameters according to the results with the second one, and finally checks its effectiveness with the last one.

Finally, as already mentioned, an economic analysis of different strategies that will be proposed to reduce the impact that reverse logistics has on the company will be carried out.

Results

When evaluating the reliability of the proposed model, different metrics were selected taking into account the nature of the data and results and the final objective of the project. These are misclassification, precision and recall. The values obtained for each set are as follows:

	Training	Validation	Test
Misclassification	0,07%	0,11%	0,09%
Precision	99,99%	99,98%	99,95%
Recall	99,87%	99,81%	99,86%

Based on these results, it can be stated:

- The '*misclassification'* results are very low in all three cases. This means that the overall error that the model makes is low.
- The '*precision'* metric takes very high values, indicating that a very high percentage of estimated returned products will actually be returned, i.e. the quality of the model in terms of classification is excellent.
- Finally, '*recall'* values are also high. This indicates not only that the model is able to identify the relatively few positive cases, but also that it does so accurately.

On the other hand, the different strategies whose economic feasibility was analyzed are as follows:

- Strategy 1: That the client pays a certain amount of money per product if the model predicts that the product will be returned.

	3,00 €	5,00 €	6,00 €	7,00 €	9,20 €
Extra income per year	52.980,32 €	88.300,53 €	105.960,64 €	123.620,75 €	162.499,77 €
Extra income per month	4.075,41 €	6.792,35 €	8.150,82 €	9.509,29 €	12.499,98 €
Payback (months)	36,81	22,08	18,40	15,77	12,00
Payback (years)	3,1	1,8	1,5	1,3	1,0

Table 1. Economic analysis of strategy 1

- Strategy 2: Have the customer pay a percentage of the cost of the product (according to its category) if it is estimated that it will be returned.

Category	5%	10%	15%	20%
Extra income per year	36.697,59 €	73.395,19 €	110.092,78 €	146.790,38 €
Extra income per month	2.822,89 €	5.645,78 €	8.468,68 €	11.291,57 €
Payback (months)	53,1	26,6	17,7	13,3
Payback (years)	4,4	2,2	1,5	1,1

Table 2. Economic analysis of strategy 2

- Strategy 3: Have the customer pay a percentage of the cost of returning the product (according to its category) if it is estimated that it will be returned.

Category	10%	20%	30%	40%
Extra income per year	42.156,70 €	84.313,40 €	126.470,10 €	168.626,80 €
Extra income per month	3.242,82 €	6.485,65 €	9.728,47 €	12.971,29 €
Payback (months)	46,3	23,1	15,4	11,6
Payback (years)	3,9	1,9	1,3	1,0

Table 3. Economic analysis of strategy 3

Conclusions

In order to determine which strategies are proposed as initiatives for the company to reduce the impact of returns, the following has been taken into account:

- The additional annual income they would bring.
- The cost of returns.
- The payback of the initial investment.
- The possibility and willingness of customers to pay extra money to offset the costs.

After analyzing the economic viability of each one of them, it has been determined which strategies are feasible, and therefore proposed as solutions, and which are not. This is summarized in Table 9.

After analyzing the economic viability of each of them, it has been determined that the most feasible, and therefore proposed as a solution to the costs of reverse logistics, is strategy 3. This consists of the client paying a penalty of 40% of the cost of return depending on the category to which the product belongs. The return on investment (ROI) would be **12.42%.** In addition, the monthly cash flows that would be obtained are represented in Table 4.

December 2017	14.304,64 €
January 2018	14.671,07 €
February 2018	10.279,22 €
March 2018	12.231,80 €
April 2018	10.187,10 €
May 2018	5.454,28 €
June 2018	7.926,65 €
July 2018	24.987,84 €
August 2018	11.302,37 €
September 2018	11.742,36 €
October 2018	6.641,11 €
November 2018	11.459,02 €
December 2018	21.366,25 €
TOTAL	162.553,72 €

Table 4. Monthly cash flows of the proposed strategy

In addition, in order to provide a clearer picture of the impact this strategy would have on clients, the penalty to which each client would be subjected was calculated. This would be **3.10 €** per year. This means that it would be sufficient to add around 0.50 € per transaction as a shipping cost, for example, if we assume that each customer makes 6 purchases per year.

References

[1] Catchoom, "eCommerce In The Fashion Industry. Industry Changes, Stats & Trends. Image Recognition, AR and Artificial Intelligence Solutions," 2018. [Online]. Disponible: https://catchoom.com/blog/how-is-ecommerce-changing-fashion-industry-stats-trends-predictions/?cn-reloaded=1.

[2] W. Kofler, "Artificial Intelligence in Retail – What to expect and how to act," PwC. [Online]. [Accessed 2019].

[3] F. Ma, "The Study on Reverse Logistics for E-Commerce," IEEE Conference Publication, 2010. [Online]. Available: https://ieeexplore.ieee.org/document/5575577

Índice de la memoria

Índice de tablas

Índice de figuras

CAPÍTULO 1. INTRODUCCIÓN Y PLANTEAMIENTO DEL PROYECTO

1.1 INTRODUCCIÓN

La aparición del Internet ha cambiado la vida de las personas en más ámbitos y más rápidamente que el desarrollo de cualquier otra tecnología. Su acelerado crecimiento y el de las tecnologías de la información ha generado la evolución del comercio al ámbito online. El e-commerce se define como el comportamiento de comprar y vender bienes o servicios, así como comunicaciones comerciales y transacciones, a través de redes informáticas y computadores individuales vinculadas al World Wide Web (WWW).

Los avances tecnológicos traen consigo oportunidades de transformación, y el sector retail ha sido quizás el primero en comprenderlo. El internet ha afectado significativamente a la industria textil. Motivada a proporcionar comodidad al consumidor, la mayoría de las industrias de este sector han creado versiones virtuales de algún aspecto de su ambiente físico. Algunos han utilizado el internet para ir más allá de su oferta actual, brindando al consumidor la experiencia en internet como valor añadido: catálogos personalizados de ropa, artículos disponibles solamente online…

Las empresas ven el salto al e-commerce como una oportunidad de negocio para aumentar la productividad e incentivar la comunicación. Además, es una fuente de información extremadamente valiosa para la empresa, pues es un medio ideal para la recolección de datos.

Asimismo, la industria textil ha aumentado significativamente su presencia en el e-commerce en los últimos años. E-tailing es un subconjunto del e-commerce que engloba las actividades de venta de bienes retail. Los expertos auguran un gran crecimiento de su cuota de mercado en los próximos años. Según el último reporte de Forrester, para 2022 va a alcanzar el 36% del total de las ventas de retail. Esto se debe a que cada vez más personas prefieren comprar online. De hecho, el 65% utiliza sus dispositivos cuando necesitan comprar algo [1].

Mejores promociones, mayor variedad y más comodidad son algunas razones por las cuales los consumidores prefieren comprar online. Que el 67% de los millennials lo prefieran por encima de tiendas físicas lo demuestra [1].

Sin embargo, las compañías que se han aventurado en el e-commerce no han tenido demasiado en cuenta todo aquello que conllevan las devoluciones de productos. Es por eso que cada vez es más importante para las compañías optimizar su logística inversa. Muchas de ellas la están implementando para obtener ventajas competitivas. Las devoluciones bien gestionadas tienen un impacto muy potente en los costes, ingresos y en la satisfacción del cliente. Algunos de los beneficios de un planteamiento efectivo de la logística inversa son la mejora del desempeño financiero de la compañía, reducción de inventarios, utilización de recursos más eficiente y el aumento de la satisfacción del cliente y mantenimiento de su lealtad [2].

A pesar de esto, hay compañías que siguen percibiendo la logística inversa como un problema, pues ésta supone de un 2% a un 4% de su facturación total. Con el fin de reducir su impacto, la predicción de devoluciones y la identificación de patrones juegan un papel cada vez más interesante para las empresas, para luego poder implantar barreras que las reduzcan.

El impacto que tienen las devoluciones en el margen comercial de las empresas depende mucho de su política. Para determinar las diferentes estrategias se utilizan modelos analíticos. Es más probable que posean políticas más generosas si [8]:

- Los productos no se pueden consumir en un período corto de tiempo.
- Su línea de productos ofrece oportunidades de venta cruzada.
- Los productos devueltos tienen un alto valor de recuperación.

Además, el impacto ambiental es uno de los aspectos más destacados en la sociedad hoy en día. Los consumidores están muy al corriente de la huella de la industria textil en el medioambiente, y son conscientes de que sus compras juegan un papel importante. Esto significa que las marcas tendrán que empezar a integrar esto en su cultura de operaciones y administración, cosa que muchas ya están haciendo. Una logística inversa más eficiente puede

ayudar a reducir las emisiones al aire y al agua, a reducir el consumo de energía y la utilización de recursos naturales, y a evitar el almacenamiento de desechos.

Por otra parte, estamos avanzando hacia un mundo donde las tareas tediosas y repetitivas serán llevadas a cabo por máquinas. 2019 fue definitivamente el auge de la inteligencia artificial (IA), y se espera que en el 2020 se consolide y se siga desarrollando. De hecho, se espera que el impacto de la IA sea masivo a largo plazo. Un estudio llevado a cabo por PwC estima que el impacto en el PIB global sea de 15.7 trillones de dólares para 2030 [3], tal y como se ilustra en la Ilustración 1.

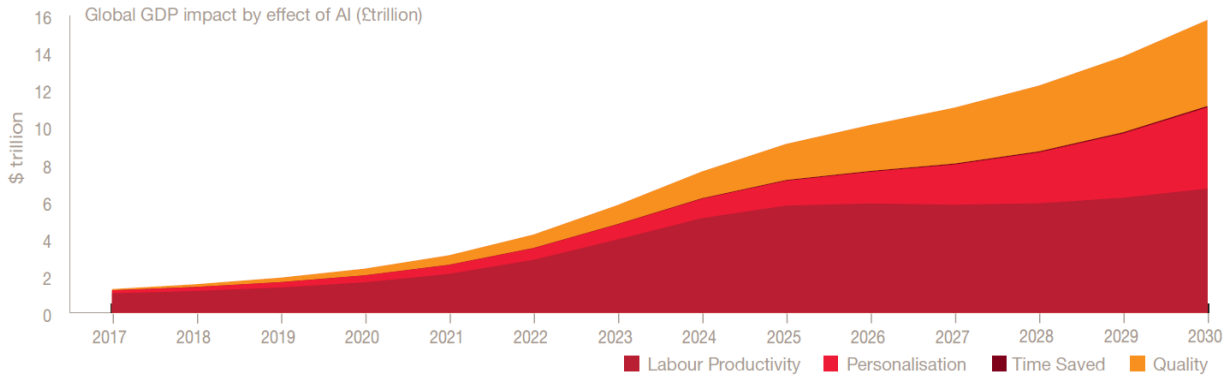


Ilustración 1. Impacto de la IA en el PIB global

Con el tiempo, el aumento previsto para 2030 de 15,7 billones de dólares del PIB mundial será el resultado del impacto de la IA en la productividad y en el comportamiento de los consumidores [3]. Estos se sentirán atraídos sobre todo por productos y servicios de mayor calidad y más personalizados.

Algunos de los principales usos de la IA en el sector retail son la optimización automática de precios basada en datos internos y externos, la optimización de la cadena de suministros desde la predicción de demanda hasta la definición de los niveles de stock, la personalización sobre la marcha de ofertas de compra, programas basados en la conversación a través de métodos auditivos o textuales, la búsqueda basada en imágenes de artículos similares... Compañías como Amazon y Privalia ya están implementando esta tecnología en diferentes áreas de su cadena de valor.

A pesar de todos estos avances, las compañías pertenecientes a la industria de la moda tendrán que seguir adaptándose para poder sobresalir. Los más exitosos serán aquellos que apuesten por las nuevas tecnologías y que aboguen por causas sociales: potenciación de las interacciones con sus clientes mediante voz, incorporación de contenido visual, previsión de devoluciones… La inteligencia artificial tendrá cada vez más protagonismo en este contexto.

1.2 MOTIVACIÓN

Las compras online se han normalizado como resultado de muchos factores: el auge de las redes sociales, el desarrollo de sistemas de logística sofisticados, promesa de envíos y devoluciones gratis… pero sobretodo, como resultado del avance de la tecnología. La mejora de las conexiones a internet, el desarrollo veloz de dispositivos móviles, la penetración de smartphones y tablets en las actividades del día a día, la creación de aplicaciones… han ocasionado que los consumidores prefieran comprar online cada vez más.

Esto supone que el cliente no tenga la oportunidad de analizar el producto al detalle, como lo haría si estuviese en una tienda física. Por este motivo y por las políticas de envío y devolución gratuitas han surgido distintos tipos de clientes que se han acostumbrado a comprar online para luego devolver parte o la totalidad de su pedido.

A pesar de que las empresas han buscado soluciones como la implementación de sistemas de inteligencia artificial y realidad virtual para reducir el impacto en el margen comercial, el número de devoluciones ha aumentado con el auge del e-commerce, impulsando así los costes de la logística inversa. Esta se ha convertido en el campo de más rápido crecimiento de la logística de una empresa [4].

Es por esta razón que las principales motivaciones de este proyecto son las siguientes:

- La implantación de un sistema eficaz que permita predecir las devoluciones, y así poder reducirlas y con ello su impacto en el margen comercial.

- En un contexto donde el e-tailing entre fronteras está creciendo cada vez más, perjudicando así al comercio nacional, las consecuencias medioambientales van a aumentar debido a las distancias de transporte más largas, cambios en la configuración de almacenes… [4]. Proporcionar información sobre el impacto ambiental que tienen las devoluciones. Un planteamiento efectivo puede reducir los residuos aumentando la reutilización de productos.
- La posibilidad de idear estrategias para grandes empresas del mundo del retail como Zalando, ASOS o Inditex, con el objetivo de reducir el impacto de las crecientes devoluciones en su beneficio.
- A pesar de que el e-commerce ha revolucionado de manera fundamentalmente positiva la manera en que las personas realizan sus compras, sería interesante proponer medidas para resolver uno de los asuntos más negativos que ha traído consigo: las devoluciones descontroladas.
- Optimizar la política de devoluciones con el fin de reforzar la lealtad del cliente a la marca y aumentar su satisfacción.
- Ofrecer la posibilidad de gestionar el stock de manera más eficiente. La capacidad de predecir devoluciones permitirá conocer con mayor certeza cuándo hacer pedidos, su tamaño…

1.3. OBJETIVOS

Enfocándonos en esta situación, este proyecto tiene como objetivo principal el desarrollo de un modelo que sea capaz de predecir las devoluciones en el mundo del retail, en concreto en el sector textil, con el fin de utilizar esta información para reducir los costes que la logística inversa tiene en la compañía.

Para ello, se llevará a cabo un proceso de definición de tipos de consumidor, analizando su comportamiento según sus características: tipo, edad, género… para poder determinar y agrupar las causas principales por las cuales un cliente decide devolver un producto. Así, se conseguirá identificar diferentes patrones de devolución.

Será necesario tener en cuenta las políticas de devolución de las empresas más importantes del sector textil como Inditex, Nike, Adidas, H&M… Se identificarán qué estrategias existen actualmente y cuáles se pueden implantar para optimizar la situación.

Una vez identificados y analizados los resultados, se planteará un caso de negocio para estudiar el impacto económico que ocasionan las devoluciones en el margen de las empresas pertenecientes al sector textil. Además, se incluirán una serie de propuestas que tendrán como objetivo reducir las devoluciones, para que afecten en la menor medida posible al funcionamiento de la empresa.

Para todo esto, se utilizarán los datos proporcionados de un marketplace del sector que ha ofrecido su uso para tal fin.

1.4. METODOLOGÍA

Los pasos a seguir para la elaboración de este trabajo son:

1. **Formación de Python + Búsqueda de información**

Las primeras semanas se seguirán cursos de Python con el fin de adquirir los conocimientos necesarios para analizar los datos. Al mismo tiempo se investigará sobre las políticas de devolución de las grandes empresas de retail de hoy en día, implementaciones de machine learning para optimizarlas, el impacto de la logística inversa en los márgenes comerciales…

2. **Recolección de datos**

Se simularán datos de un marketplace del sector textil, desarrollando una lógica para añadir todas las variables que se consideran necesarias para la predicción. Las proporciones de las variables se obtendrán de bases de datos ya existentes, y de datos de reportes de empresas del sector.

3. **Análisis de variables**

Se analizarán los datos proporcionados con el fin de identificar patrones de devolución, estudiar el comportamiento del consumidor según su tipo o región en la que vivan, se considerará la posibilidad de abrir una tienda física…

4. **Utilización de modelos predictivos**

Tras analizar los datos, se utilizarán modelos predictivos para predecir las devoluciones con el fin de reducir su impacto en el margen de operación de las empresas.

5. **Caso de negocio**

Se analizará el impacto económico de las devoluciones y se propondrán medidas para reducirlas.

6. **Redacción de la memoria**

A lo largo de todo el proyecto se redactarán tanto los anexos a entregar como la memoria del trabajo de fin de máster.

1.5. RECURSOS

Para poder llevar a cabo este proyecto se emplearán tres recursos principales:

- Los datos a analizar serán **datos** simulados de un Marketplace del sector moda. Se desarrollará una lógica que simule datos específicos del sector.
- Para analizar los datos se utilizará **Python**, que es un lenguaje de programación avanzado que permitirá identificar patrones y comportamientos de devolución para luego poder proponer medidas de mejora. El modelo predictivo se desarrollará en este lenguaje.
- Con el fin de estudiar la viabilidad económica de todas las estrategias propuestas, se utilizará **Microsoft Excel.**

CAPÍTULO 2. DESCRIPCIÓN DE LAS TECNOLOGÍAS: ESTADO DEL ARTE

Como ya se ha explicado, las devoluciones son las principales impulsoras de la logística inversa. Es por eso por lo que en los últimos años se han intentado desarrollar medidas para prevenirlas, con el fin de reducir su impacto en el margen comercial. La realidad virtual se ha convertido en uno de los recursos más populares para intentar dar con soluciones.

2.1 REALIDAD VIRTUAL

La tecnología y el internet han cambiado nuestra experiencia espacial y han creado un nuevo espacio: el espacio virtual [5]. Las compras físicas permiten al cliente interactuar con los productos a través de todos los sentidos antes de realizar una compra. Sin embargo, en las compras online los consumidores sólo pueden utilizar palabras claves o jugar con la clasificación de una página web para encontrar lo que están buscando, por no hablar de que solo cuentan con una pantalla 2D para considerar la compra del producto. Sin embargo, este último comportamiento satisface necesidades diferentes. Es por eso que la creación de *una plataforma de compra de realidad virtual* se plantea como una herramienta para incorporar los dos tipos de experiencias de compra.

Esta plataforma incluiría las siguientes prestaciones [5]:

- Los usuarios podrán utilizar sus manos para coger los productos, rotarlos, acercárselos… con el fin de analizar todos los detalles que deseen.
- Implantación de un asistente de compras virtual controlado por voz que ayude al cliente en lo que necesite. Por ejemplo, que lo guie a la sección donde están los productos que le interesarían o que suele adquirir.
- La plataforma estaría conectada a las redes sociales, con el objetivo de que los usuarios puedan enviar solicitudes a sus amigos para ‘ir de compras juntos’.
- Al imitar el ambiente de una tienda virtualmente, la experiencia de compra online se acercará aún más a la experiencia física.

Se espera que la simulación de un centro comercial virtual sea capaz de reducir las devoluciones en el sector textil.

2.2 MACHINE LEARNING

Por otro lado, la cantidad de datos generados hoy en día es tan inmensa que los humanos no podemos procesarla sin ayuda. El machine learning es capaz de generalizar información a partir de bases de datos extensas, además de ofrecer técnicas de aprendizaje automáticas para extraer patrones. Es por eso que esta tecnología se utiliza cada vez más, y el sector textil no es una excepción. Introduciendo datos como el intervalo de tiempo de compra, número de compras y el género, edad y región del consumidor se pueden obtener conclusiones sobre qué productos son más probables de ser devueltos [6].

Además, el data mining ya ha sido utilizado de manera exitosa en el sector textil para tareas como la identificación de perfiles de usuarios con el fin de crear, innovar y mejorar productos textiles, para entender mejor los requerimientos de los clientes, reconocer y clasificar defectos para el control de calidad, descubrimiento de hábitos de compra del consumidor para aumentar las ventas [6]…

En concreto, ya se ha llevado a cabo la creación de un modelo analítico con el fin de estimar el comportamiento de los consumidores e identificar patrones. Se segmentó a los usuarios en diferentes categorías [7]:

- Consumidores egoístas, que tienen altas tendencias de devolver los productos que compran.
- Consumidores honestos, que tienen muy claro qué necesitan comprar, y solo devolverán el producto si tiene defectos de calidad.
- Consumidores estafadores, que compran el producto para probarlo gratis intencionalmente, o para reclamar compensaciones al devolver el producto.
- Consumidores irrelevantes, que no tienen tendencias de devolución.

Se espera que cada uno de ellos presente un comportamiento diferente en cuanto a las devoluciones que realizan. Se considera que hay cuatro problemas principales por los cuales un cliente devuelve un producto [7]:

- El producto no cumple con las expectativas del cliente.
- Problemas de calidad e inventario.
- Devoluciones fraudulentas (productos usados, en mal estado…)
- Errores accidentales del cliente, es decir, compró el artículo equivocado.

Este estudio ('trust-aware random walk model') se apoyó en la segmentación de los clientes para estudiar sus preferencias al realizar un pedido, y así poder aplicar la idea de la red de confianza en datos reales, específicamente en la etapa de pedido de los clientes, fase crucial para poder detectar el pedido de artículos posibles de devolución.

2.3 INTELIGENCIA ARTIFICIAL

A partir de estos estudios, se puede decir que el uso de la inteligencia artificial en retail está en aumento. Cada vez más empresas invierten en esta tecnología para controlar su inventario, proporcionar recomendaciones personalizadas de productos a los clientes, predecir sus gustos para luego diseñar ropa que quieran comprar, para facilitar la búsqueda de productos… Las aplicaciones más destacadas en diferentes aspectos de la industria retail se indican en la Ilustración 2.

Según un estudio llevado a cabo por PwC [3], es probable que los retailers se adapten a diferentes tipos de aplicaciones. La IA permitirá tener acceso a información que antes no lo era. Estos conocimientos nutrirán el engagement de los consumidores, además de permitir la aceleración empresarial.

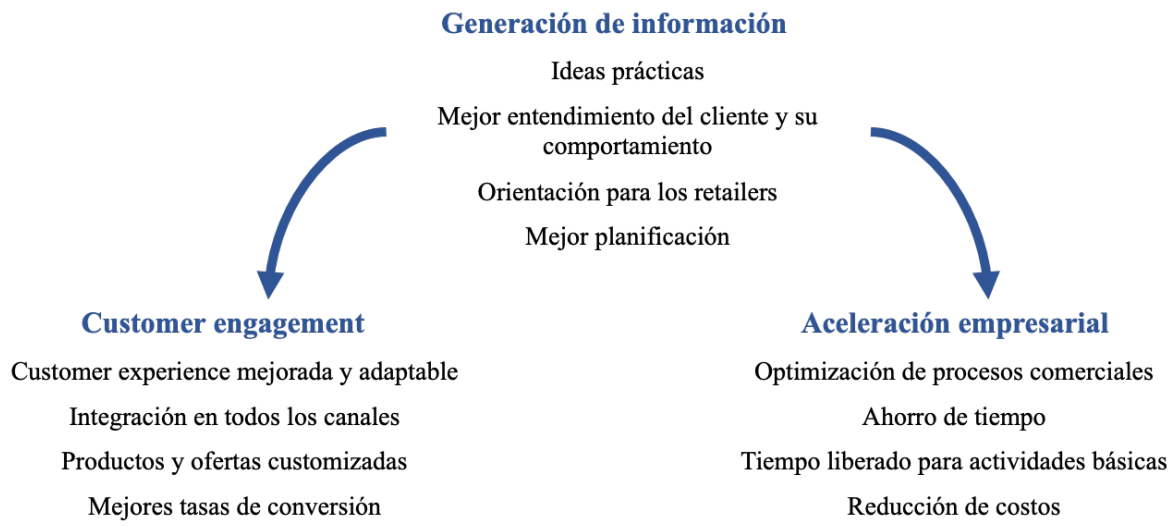


Ilustración 2. Aplicaciones de la IA en la industria retail

Los tipos de aplicaciones comentadas en la Ilustración 2 impactarán toda la cadena de valor de un retailer. En el Apéndice B: Impacto de la Inteligencia Artificial en la cadena de valor de un retailer se detalla este impacto. A continuación, se detallan tres iniciativas interesantes que se están implementando hoy en día, o están en fase de desarrollo [3].

a) Habilitar búsqueda visual

Esta herramienta consiste en la posibilidad de buscar un ítem específico utilizando la cámara de un smartphone con la ayuda de IA. Empresas como Target, Asos y Zalando ya lo están implementando.

El usuario cargaría desde su smartphone una foto del producto, un anuncio… a la app del retailer. El siguiente paso es que el cliente marque sus áreas de interés. El algoritmo examina la oferta de productos para encontrar este o uno similar. Si la búsqueda es exitosa, el producto es ofrecido a los clientes. También se les enseña otros productos parecidos disponibles.

Uno de los beneficios principales es el aumento de la tasa de conversión, lo que conlleva un incremento de las ventas dentro del grupo de usuarios.

b) Escanear estanterías vacías con robots

Consiste en el análisis de estanterías y gestión de inventarios a través del reconocimiento visual del espacio de las estanterías. Empresas como Target, Walmart y Loewe ya lo utilizan.

Se lleva a cabo mediante robots que se movilizan por la tienda escaneando las estanterías, intentando encontrar problemas de stock, errores en el precio… El robot compara la imagen que consigue de la estantería con imágenes de la situación ideal, cargadas en su memoria anteriormente. La información que recolecta se alimenta en tiempo real a un dispositivo, para que el staff de la tienda pueda resolver el problema.

Se puede decir que esta herramienta incentiva las ventas debido a la mejor disponibilidad de productos en las estanterías. También contribuye a la aceleración empresarial, pues reduce el esfuerzo de revisión de los estantes, lo que lleva a la reducción de los costos operativos de la tienda.

c) Servicio al cliente apoyado por IA en desarrollo

Esta iniciativa, aunque aún está en desarrollo, consiste en conseguir un servicio al cliente más eficiente con la ayuda de la IA. Starbucks, y la empresa de supermercados online británica Ocado ya lo está implementando.

El software impulsado por IA escanea los correos electrónicos entrantes de los clientes y analiza su contenido. Posteriormente, dicho software contextualiza, categoriza y prioriza el correo electrónico. El software puede proponer una respuesta para el agente de atención al cliente o enviar automáticamente una respuesta cuando se supera un grado preestablecido de certeza de que la respuesta es correcta.

Los beneficios de esta herramienta incluyen la disminución del costo por consulta, lo que conlleva una reducción significativa de los costos de operación. Además, los ingresos son potencialmente más altos debido a un mejor y más rápido servicio al cliente y, por lo tanto. a una mayor satifacción.

2.4 OTROS ESTUDIOS

La empresa J. Walter Thompson Intelligence publicó un report sobre los retos que todas las empresas del sector retail tendrán que afrontar. En este estudió descubrió que el 72% de los Millennials en los Estados Unidos cree que la IA será capaz de predecir lo que quieren [8]. Información sobre las preferencias de los clientes, la frecuencia de búsqueda de ciertos colores o categorías… permite a una empresa conocer mejor a su clientela.

Este report también incluye un estudio de consumidores de Estados Unidos y de Gran Bretaña, en el que responden a la afirmación "*es más probable que compre en una tienda física si esta tuviese…*".

Los resultados se presentan en la Ilustración 3. Se puede observar cómo las generaciones más jóvenes son más susceptibles a estas nuevas tecnologías. Entre otras cosas, si las tiendas físicas contasen con demostraciones de realidad virtual y aumentada, el 80% de los jóvenes pertenecientes a la generación Z en EEUU y el 67% en UK se inclinarían más a comprar en una tienda física.

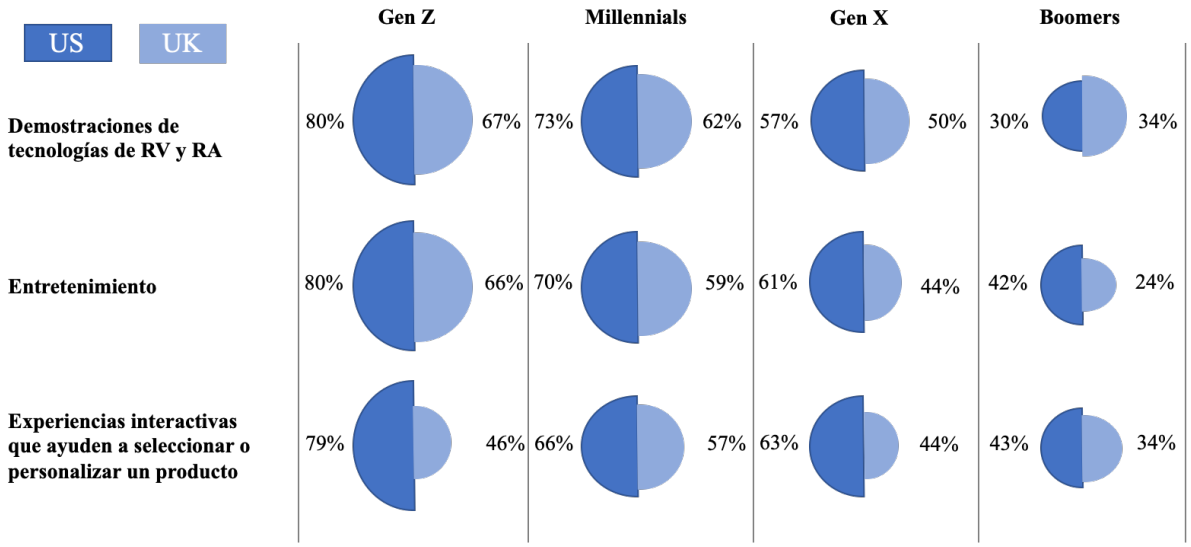


Ilustración 3. Respuestas del estudio de consumidores elaborado por J. Walter Thompson

Por otra parte, según una investigación llevada a cabo por IHL Group [9], los retailers pierden alrededor de 643 billones de dólares al año en devoluciones que se pueden prevenir. En conclusión, es por eso por lo que las compañías en la industria de la moda están cada vez más interesadas en el uso de iniciativas relacionadas con la realidad virtual y la inteligencia artificial como respuesta a sus complicaciones relacionadas con la logística inversa.

CAPÍTULO 3. DESCRIPCIÓN DEL MODELO DESARROLLADO

3.1. OBJETIVOS Y ESPECIFICACIÓN

Como ya se ha explicado anteriormente, el objetivo principal de este proyecto es la creación de un modelo analítico que sea capaz de predecir si un producto va a ser devuelto posteriormente a su compra o no. Para ello, se compararán diferentes parámetros tanto sobre el cliente como sobre el producto comprado.

El modelo se creará utilizando el lenguaje de programación Python, en concreto el módulo h2o. Esta una herramienta para recopilar datos y crear diferentes tipos de modelos de manera rápida y escalable.

3.2. DATOS

Algunas características sobre los datos desarrollados para crear el modelo son:

- Simulan pertenecer a un *marketplace español*, concretamente del sector retail y de la industria textil.
- Como el objetivo de este proyecto es la predicción de devoluciones sobre compras online, estos datos se corresponden únicamente a *ventas online*.
- El período de tiempo incluye *desde diciembre de 2017 hasta mediados de diciembre de 2018*. Es interesante que este período tenga una duración de aproximadamente un año para poder estudiar los picos de demanda en fechas clave como Navidad, las rebajas en distintas épocas del año (rebajas de verano, rebajas de invierno y Black Friday). Además, también es importante porque se puede analizar el comportamiento de los diferentes perfiles de cliente durante un año entero: número de compras año, devoluciones, categorías más compradas...
- Los datos son sobre el *producto*, el *consumidor y la transacción*. Los parámetros y sus posibles valores se representan en la Tabla 1, Tabla 2, y Tabla 3.

PRODUCTO						
Descripción	**Color**	**Categoría**	**Subcategoría**	**Talla**	**Precio (€)**	
Nombre específico del producto	Amarillo	Abrigos	Abrigos	**Partes de arriba**		
	Azul		Anorak	Chica XS-XL	**Bajo**	7,95-55,95
	Beige		Blazer	Chico S-XXL	**Medio**	56-105,95
					Premium	106-149,95
	Blanco		Camisas	**Partes de abajo**		
	Burdeos		Cazadoras y Bombers	Chica 32-42		
	Camel		Chalecos	Chico 38-46		
	Denim claro		Chaquetas	**Zapatos**		
	Denim oscuro		Parkas y Trench	Chica 37-41		
	Dorado		Sobrecamisas	Chico 41-45		
	Gris claro	Accesorios	Corbatas			
	Gris oscuro		Gorros			
	Lila		Guantes			
	Marrón		Pañuelos			
	Negro	Bisutería				
	Plateado	Bolsos				
	Rojo	Camisas y tops	Blusas			
	Rosa		Camisas			
	Verde		Tops			
		Camisetas	Camisetas			
Proveedor			Polos			
1-50		Cinturones				
		Faldas				
		Jerséis				
		Monos				
		Pantalones	Leggings			

			Pantalones Pantalones de vestir Vaqueros		
		Sastrería	Abrigos sastrería Camisas sastrería Chalecos sastrería Pantalones sastrería		
		Sudaderas			
		Trajes	Americanas Chalecos traje Faldas traje Pantalones traje		
		Vaqueros			
		Vestidos	Vestido corto Vestido largo Vestido midi		
		Zapatos	Bailarinas Botas y botines Deportivas Mocasín Zapato cerrado		

Tabla 1. Datos proporcionados sobre el producto

CONSUMIDOR			
Género	**Edad**		**Customer ID**
Mujer Hombre	Joven Adulto Persona mayor	16-27 28-60 >60	

Tabla 2. Datos proporcionados sobre el consumidor

TRANSACCIÓN		
Número de recibo	**Fecha**	**Naturaleza**
		Devolución Compra

Tabla 3. Datos proporcionados sobre la transacción

3.3. ANÁLISIS DE VARIABLES

Para analizar las características que los productos devueltos tienen en común y así posteriormente identificar patrones, se estudiaron todas las transacciones y se llegó a las conclusiones que se presentan a continuación. Cabe destacar que se utilizaron tanto Excel como Python para llegar a los resultados.

Para un mejor entendimiento de las conclusiones, se han dividido según el parámetro examinado.

- **Según la categoría**

Separando las diferentes categorías, se obtuvieron tanto los ingresos que cada una trae al Marketplace, como el dinero total que constituyen las devoluciones de los productos pertenecientes a dicha categoría. Con estos datos se obtuvo el porcentaje que representa dicha categoría sobre el total de las devoluciones, tal y como se presenta en la Tabla 4. Con el fin de obtener una visión más gráfica de los resultados, se muestra la Ilustración 4.

Es importante mencionar que los porcentajes se calculan según la cantidad de dinero y no sobre el número de transacciones porque al final de cuentas el objetivo de este proyecto es reducir los costes que las devoluciones causan a las empresas. Por ello, parece más lógico tratar los datos según el impacto que tienen en el capital.

% de devoluciones sobre el total	
Abrigos	26,19%
Jerséis	11,11%
Zapatos	8,68%
Camisas y tops	7,46%
Pantalones	6,90%
Faldas	6,43%
Vestidos	5,76%
Accesorios	4,78%
Sudaderas	4,14%
Bolsos	3,01%
Trajes	2,99%
Cinturones	2,89%
Camisetas	2,13%
Vaqueros	2,05%
Botas	1,58%
Monos	1,52%
Bisutería	1,24%
Sastrería	1,14%

Tabla 4. Porcentaje que representa cada categoría en el total de devoluciones

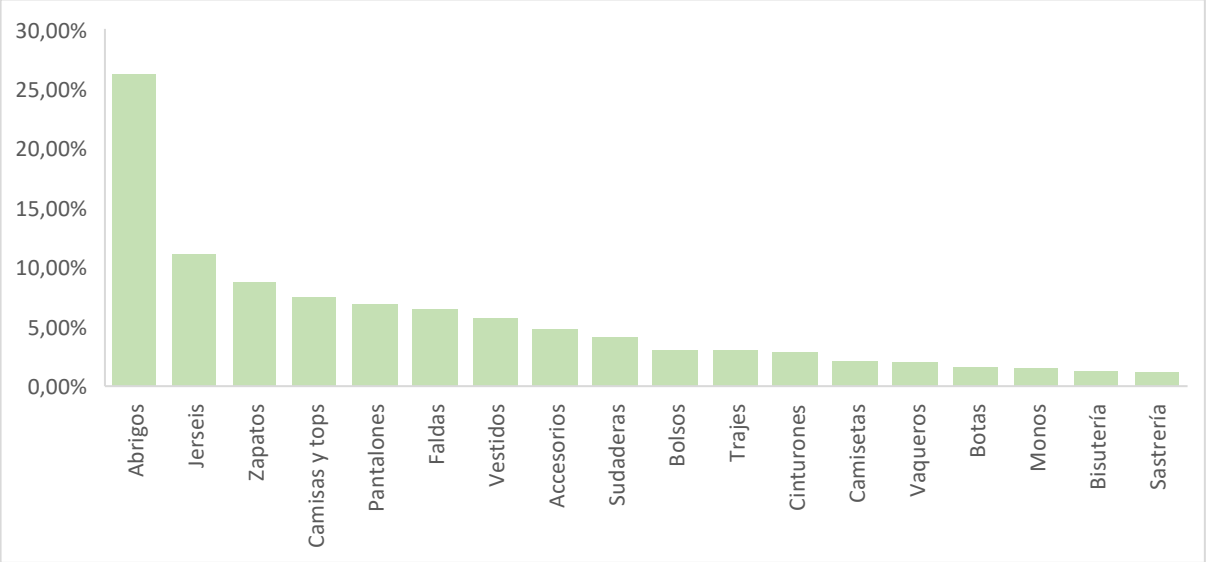


Ilustración 4. Categorías más devueltas

Como se puede observar en la Ilustración 4, la categoría que representa el mayor porcentaje de las devoluciones totales son los abrigos, seguida por los jerséis y los zapatos. Esto se puede explicar porque, como ya se ha comentado, estos porcentajes están calculados sobre cantidad de dinero y no sobre número de transacciones. Los abrigos, al ser productos normalmente más

caros frente a otros pertenecientes a categorías menos costosas, suponen una cantidad de dinero superior al ser devueltos, en contraste con otras como bisutería que representa uno de los porcentajes más pequeños.

- **Según el color**

Se agruparon todas las transacciones según el color del producto para obtener los ingresos y el dinero totales de las devoluciones por colores. Se calculó qué porcentaje supone cada color en las devoluciones totales, tal y como se enseña en la Tabla 5.

Colores más devueltos	
Denim oscuro	3,65%
Beige	3,28%
Gris oscuro	2,54%
Marrón	2,48%
Amarillo	2,44%
Denim claro	2,36%
Camel	2,35%
Verde	2,32%
Burdeos	2,26%
Rosa	2,24%
Rojo	2,10%
Plateado	1,98%
Blanco	1,92%
Gris claro	1,92%
Negro	1,87%
Dorado	1,86%
Lila	1,85%
Azul	1,64%

Tabla 5. Porcentaje sobre las devoluciones totales por colores

Como se puede observar en la Ilustración 5, el color más devuelto es el denim oscuro, que se asocia a ropa de tela vaquera (chaquetas, pantalones vaqueros…). Le siguen colores neutros: beige, gris oscuro, marrón… Esto se debe a que una gran parte de los productos analizados están disponibles en estos colores.

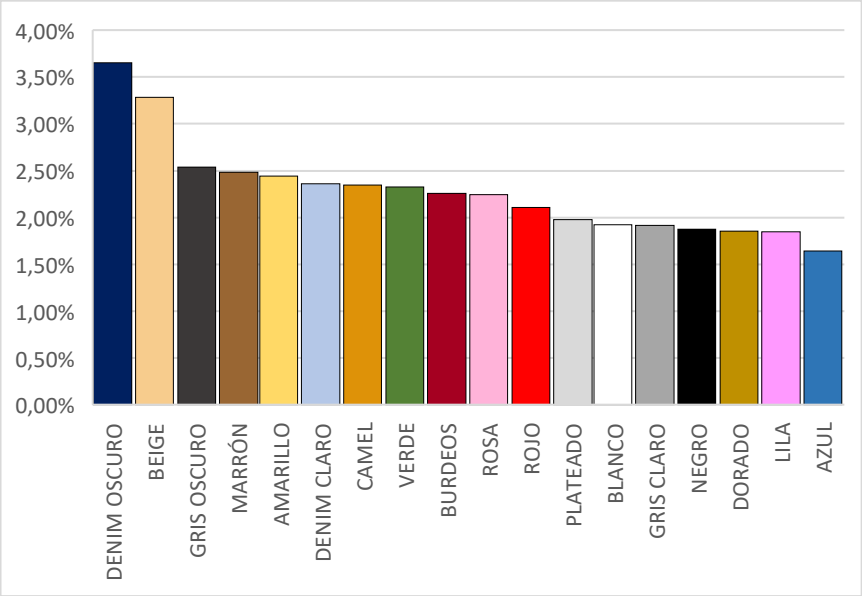


Ilustración 5. Colores más devueltos

Para un mejor entendimiento de los colores más devueltos, se estudió cuáles son los colores más vendidos, presentado en la Ilustración 6. No es sorprendente que el color más vendido sea el negro, seguido otra vez por colores neutros como el marrón y el verde.

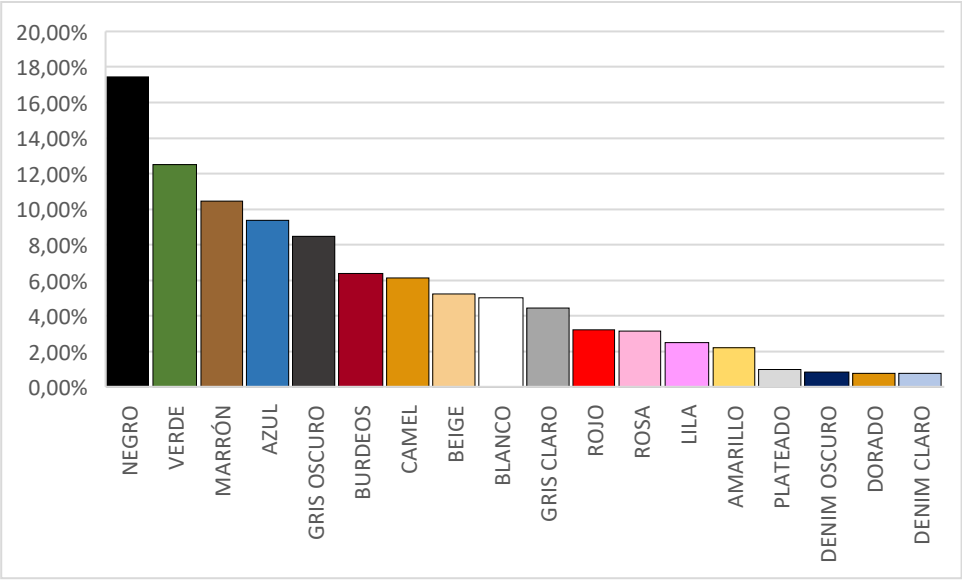


Ilustración 6. Colores más vendidos

- **Según el género**

Por otro lado, se estudió el género del comprador de todas las transacciones para obtener conclusiones sobre cómo afecta esto a la posibilidad que tiene un producto de ser devuelto.

En primer lugar, se comparó el número de transacciones realizadas por cada género. Como se puede observar en la Ilustración 7, los hombres realizan más transacciones que las mujeres. Esto justifica el hecho de que las mujeres hacen menos transacciones, pero más grandes, es decir, más productos por compra.

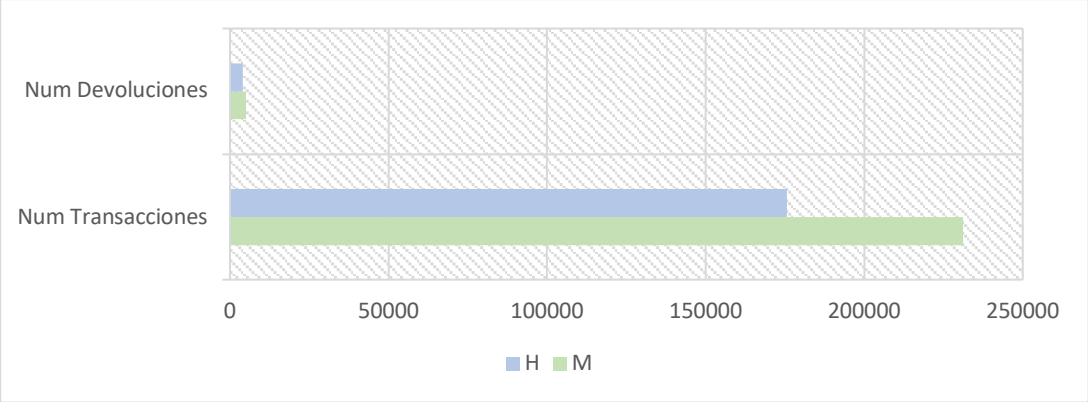


Ilustración 7. Número de transacciones y devoluciones por género

En cuanto al número de devoluciones, están más o menos igualadas, tal y como se enseña en la Ilustración 8.

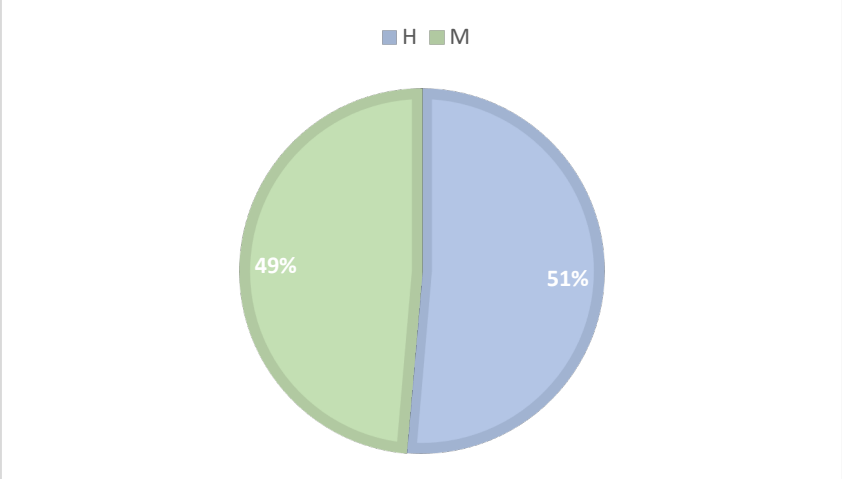


Ilustración 8. Porcentaje de devoluciones realizadas por género

- **Según la edad**

En cuarto lugar, se estudió el factor edad. Para ello se dividió a los consumidores en tres categorías, jóvenes, adultos y personas mayores. Los rangos de edad se explican en la Tabla 6.

Joven	16-27
Adulto	28-60
Persona mayor	>60

Tabla 6. Definición de rango de edades de los consumidores

Una vez más se calcularon los porcentajes que suponen las devoluciones de cada rango de edad sobre el total de devoluciones, y se obtuvieron los resultados enseñados en la Ilustración 9.

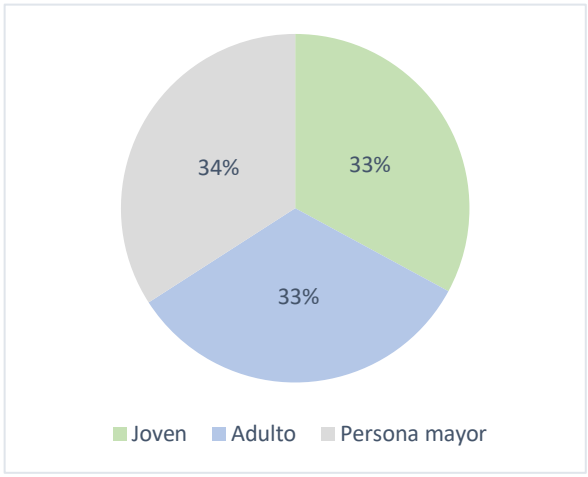


Ilustración 9. Porcentaje devuelto por rango de edad

Sin embargo, para poder poner esta información en perspectiva es importante entender los porcentajes de venta por rango de edad. Este estudió también se realizó, y se enseña en la Ilustración 10.

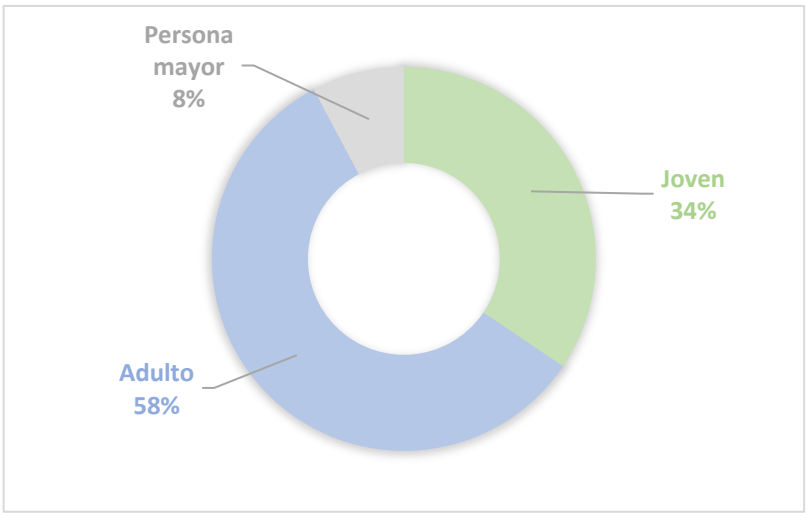


Ilustración 10. Porcentaje de ventas por rango de edad

Analizando ambos casos, se puede observar como era de esperarse los adultos son los clientes que más realizan compras y las personas mayores los que menos. No obstante, en cuanto a las devoluciones, todos tienen porcentajes parecidos.

Es posible pensar que el estudio por edad sería más interesante si se separase por géneros. Sin embargo, al hacerlo los resultados son los mismos, es decir, los porcentajes de ventas y de devoluciones por edad se mantienen. Para evitar la redundancia, se ha decidido no incluirlos en este reporte.

- **Según la fecha de la transacción**

Los datos proporcionados por el Marketplace incluyen información de aproximadamente un año natural. Para estudiar las variaciones de demanda en épocas punta, se realizó un análisis de la fecha de las transacciones.

En la Ilustración 11, se puede observar cómo tanto en diciembre de 2017 como en diciembre de 2018 las ventas se disparan. Esto se debe a que estas fechas se corresponden con la época navideña. En concreto, se ilustra como las ventas empiezan a aumentar en noviembre debido al Black Friday.

Por otro lado, también se observa un pico en julio, que se corresponde con el inicio de las rebajas de verano. En enero las ventas también son más elevadas por las rebajas de invierno.

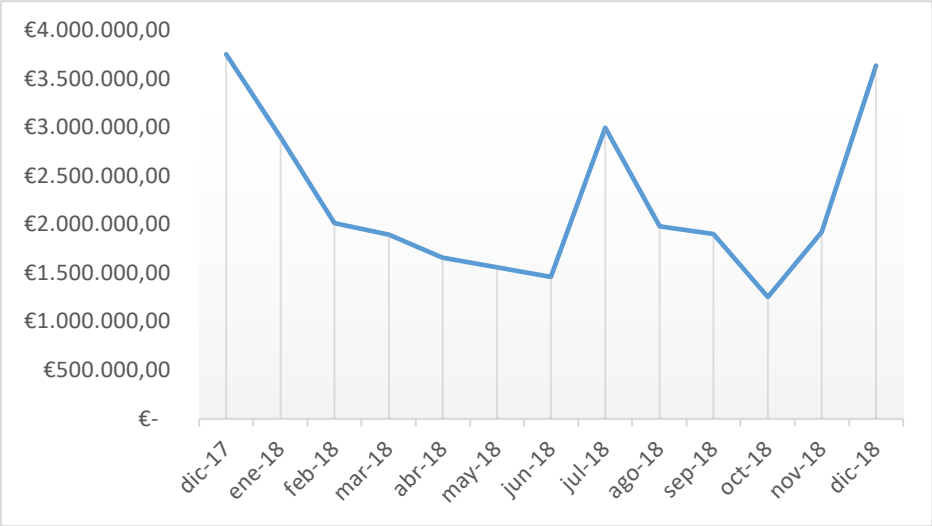


Ilustración 11. Ventas totales por mes

A pesar de que las devoluciones no son tan intuitivas, en la Ilustración 12 se observa cómo son más elevadas en julio, enero y noviembre y diciembre, que se corresponden respectivamente con las rebajas de verano, de invierno, Black Friday y navidades.

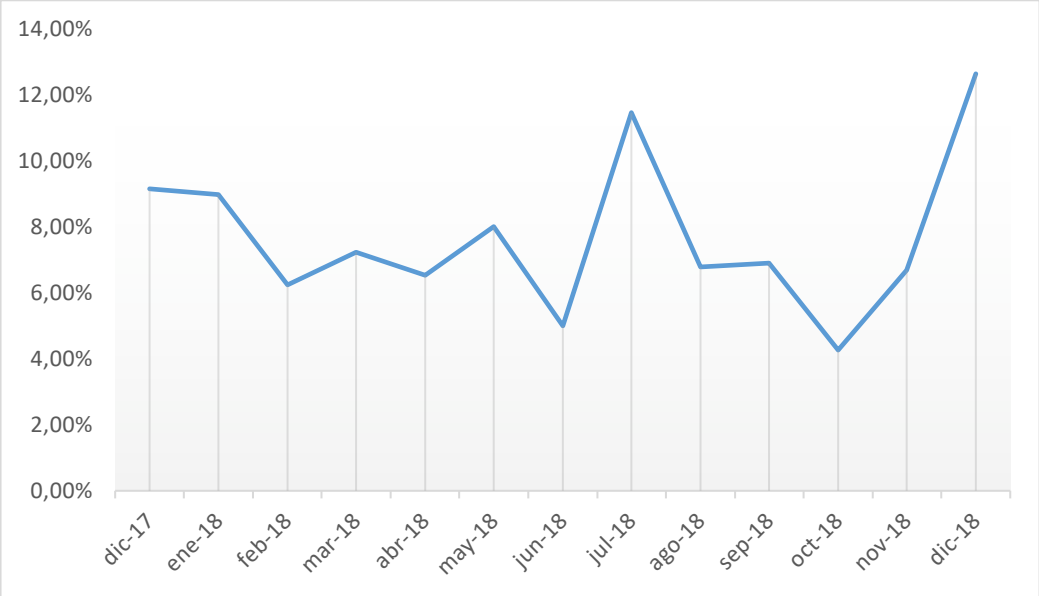


Ilustración 12. Porcentaje de devoluciones por mes

- **Según el rango de precios**

Para estudiar el impacto que tiene el precio por unidad del producto en sus probabilidades de ser devuelto, se creó una clasificación según éste que se explica en la Tabla 7.

BAJO	7,95 €	55,95 €
MEDIO	59,95 €	105,95 €
PREMIUM	110,00 €	149,95 €

Tabla 7. Definición de rango de precios

Una vez hecho esto, se estudiaron las ventas totales. Se calculó que porcentaje de estas se corresponde con cada rango de precio (bajo, medio o premium) y se obtuvo que los productos más comprados son los más baratos, y los menos comprados los más caros, tal y como se presenta en la Ilustración 13.

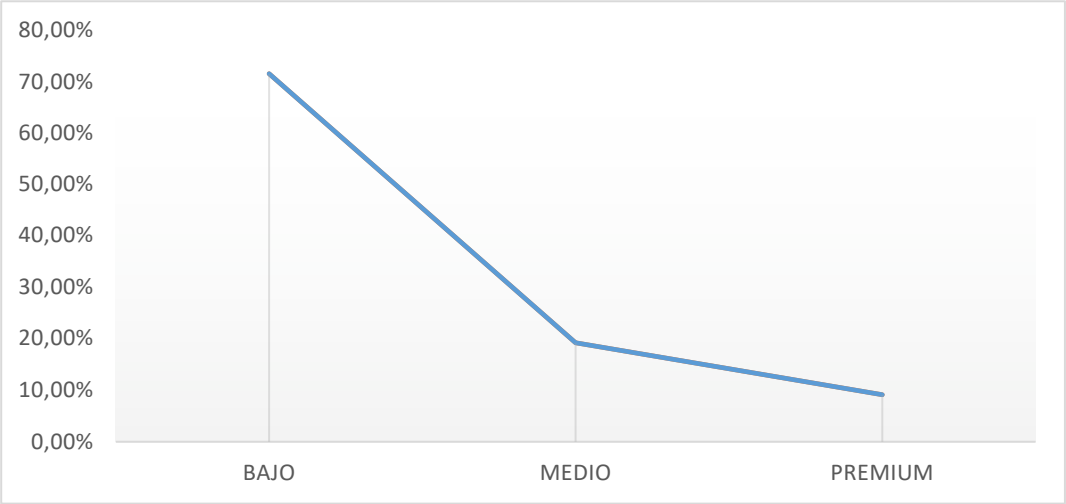


Ilustración 13. Porcentaje de ventas totales (según €) según rango de precio

En cuanto a las devoluciones, se comprobó que estas siguen el mismo comportamiento que las ventas. Los productos que pertenecen al rango de precios bajo son menos devueltos que los productos premium, tal y como se enseña en la Ilustración 14.

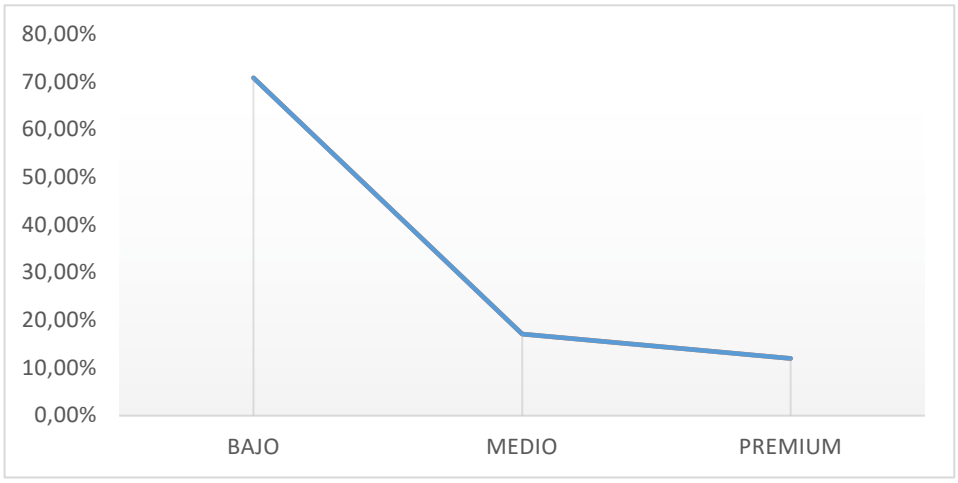


Ilustración 14. Porcentaje de devoluciones totales (según €) según rango de precio

- **Por proveedor**

Como se ilustra en la Tabla 1, los productos pueden proceder de 50 proveedores diferentes. En este apartado se estudió tanto cuales de ellos suponían el mayor porcentaje de ventas, como cuales de ellos suponían el mayor porcentaje de devoluciones. Para simplificar, en la Ilustración 15 y en la Ilustración 16 se muestra el top 10 de ambos.

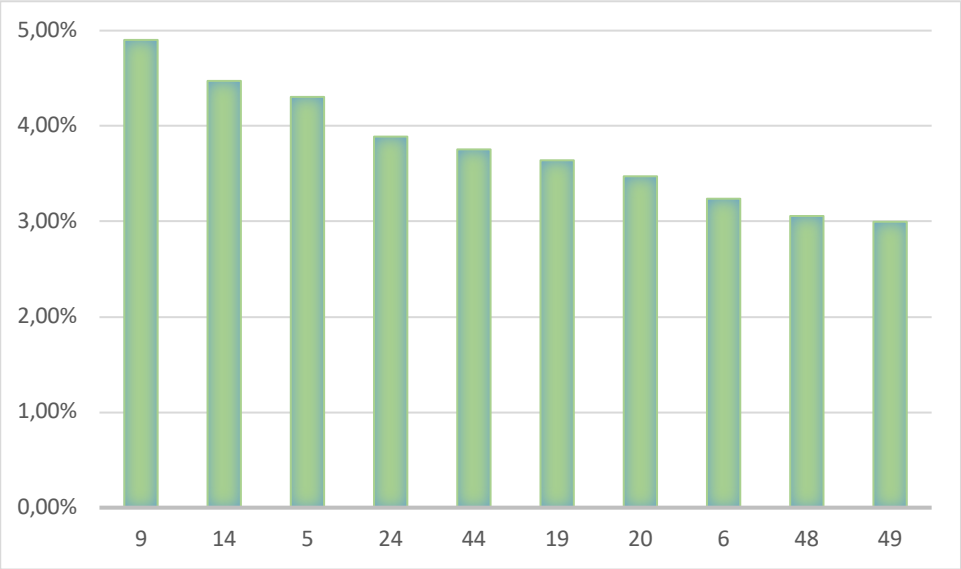


Ilustración 15. Top 10 proveedores con mayor porcentaje de ventas

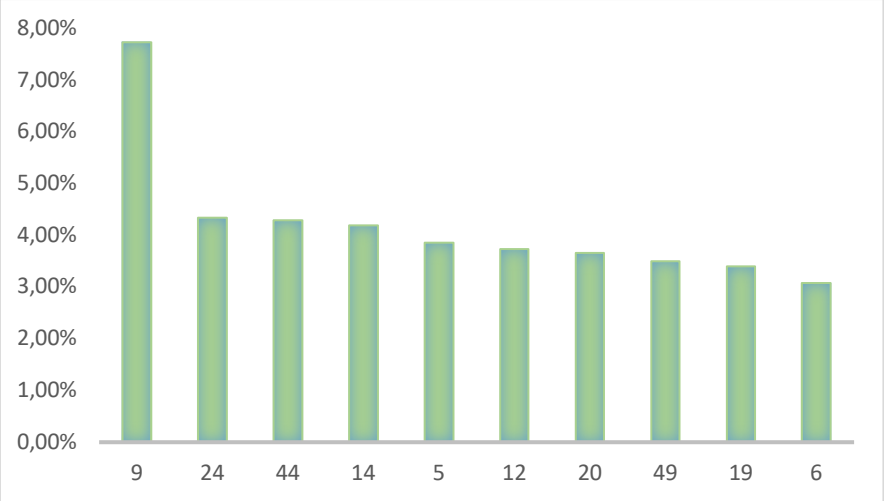


Ilustración 16. 10 proveedores con mayor porcentaje de devoluciones

Al analizar las gráficas, se puede decir que el hecho de que los proveedores pertenecientes al top 10 de ambas clasificaciones coincidan se debe a que al ser los más vendidos, tienen mayor oportunidad de ser devueltos.

Sin embargo, destaca el proveedor 9, pues sus productos se devuelven considerablemente más que el resto. La relación de este proveedor con otras variables como categoría, talla… se estudiarán en el capítulo 3.4.2.

- **Top 10 productos**

Por otra parte, se hizo un análisis de los 10 productos más vendidos. El número de unidades vendidas versus la descripción del producto se presentan en la Ilustración 17. Como se puede observar, los productos más vendidos se corresponden con prendas básicas como camisetas, jerséis y bufandas.

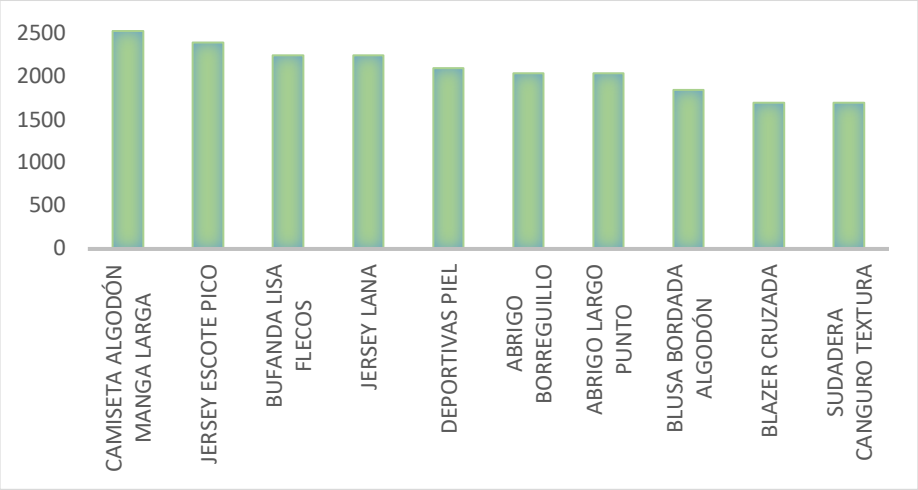


Ilustración 17. Top 10 productos más vendidos

3.4. MODELO PREDICTIVO

Para elaborar el modelo predictivo se siguieron diferentes pasos que se detallan a continuación:

3.4.1 Elección del módulo h2o y del modelo GBM

El módulo h2o de Python es una herramienta que permite acceder al software h2o de Java, y que admite la recolección de datos y la construcción de aplicaciones a usuarios que trabajan con Python [10].

Permite combinar Big Data con el aprendizaje estadístico y los algoritmos de machine learning. Además, una de sus principales ventajas es que es un módulo escalable, es decir, es útil tanto para volúmenes de datos reducidos como para grandes bases de datos. Asimismo, trata los datos de manera distribuida, es decir, en lugar de procesar los datos en un único servidor, divide la tarea en subtareas, que se llevan a cabo en múltiples nodos de manera paralela, reduciendo el

tiempo de cálculo. Por las características explicadas anteriormente, este es el módulo que se va a utilizar para realizar el modelo.

Es importante mencionar que a pesar de que los datos se encuentran en el cluster de h2o, la elaboración del modelo (iniciar el cluster, cargar los datos proporcionados por el marketplace, entrenar el modelo, predicciones de devoluciones…) se realizará desde Python. Sin embargo, es posible transferir datos de la memoria al cluster h2o y viceversa con las funciones necesarias. Esto será útil a la hora de estudiar las predicciones.

Por otra parte, de todos los modelos que proporciona el módulo h2o se ha escogido el GBM (Gradient Boosting Machine). Este combina de manera secuencial modelos más sencillos basados en árboles, por lo que se obtienen resultados de predicción cada vez mejores gracias a aproximaciones progresivamente refinadas.

Algunos de los hiperparámetros más importantes a la hora de diseñar el modelo son:

- Detención temprana: Métrica que evita un proceso de ajuste innecesario, una vez el modelo deja de mejorar. Se calcula con los datos de validación. En el caso de este proyecto se utilizará 'misclassification', que detiene el entrenamiento cuando el error durante dos mediciones consecutivas no se reduce más de un 1% [11].
- Comportamiento estocástico: Métrica utilizada para reducir el overfitting, es decir, que el modelo final se ajuste perfectamente a los datos de entrenamiento. Para conseguir este objetivo, los valores estarán entre el 0 y el 1.
- Complejidad de los árboles: Incluye el número de árboles, su profundidad…
- Velocidad de aprendizaje: Establece cuánto influye cada árbol en el conjunto final. Mientras más lenta, más árboles se necesitan, pero se reduce el riesgo de overfitting. En este modelo se utilizará un valor de 0,05.
- Balanceo de clases: En los datos proporcionados la cantidad de transacciones no devueltas es considerablemente mayor a las devueltas, es decir la clase mayoritaria es mucho más significativa que la minoritaria. Esto puede ocasionar dificultades a la hora de reconocerlas. Por este motivo es necesario llevar a cabo un balanceo de clases. En este caso se realizará mediante *oversampling,* es decir, sobre muestreando la clase minoritaria (devoluciones) para destacar su presencia.

3.4.2 Creación de variables

Este modelo analizará las transacciones en un espacio de tiempo, concretamente entre 2017 y 2018. Gracias a las variables que se han creado y que se explican a continuación, se estudiará el comportamiento de los clientes a lo largo de este período. Se analizará cada transacción, determinando sus condiciones (características del cliente y del producto, fecha…) con respecto a las anteriores. De esta manera, se examinarán nuevas transacciones para comprobar si estas siguen algún patrón de devolución, y se determinará si es probable que alguno de los productos comprados sea devuelto o no.

Con el fin de proporcionar una visión más clara de lo que estas variables analizan, se facilitan ejemplos de algunas de ellas a continuación, estudiando un número de clientes muy reducido.

- Variable 1: Naturaleza de la última transacción

Tras agrupar todas las transacciones por cliente y por fecha, se analiza cada una de ellas con respecto a las hechas anteriormente, con el objetivo de determinar si esta última fue una compra o una devolución.

Customer ID	Fecha	Devolución	UltTransac
26993	20/04/2018	0	NaN
26993	26/07/2018	1	0
26993	26/07/2018	0	1
26993	21/12/2018	0	0

Tabla 8. Ejemplo de variable 1: Naturaleza de la última transacción

La transacción correspondiente al 20 de abril no devuelve ningún número porque esta es la primera que ha realizado este cliente en concreto. En cambio, una de las realizadas en julio devuelve un 1, que representa que la transacción anterior fue una devolución y no una compra (representadas por un 0).

- Variable 2: Categoría del producto devuelto

Esta variable también segmenta las transacciones por cliente, y proporciona información sobre la categoría a la que pertenece el último producto devuelto.

Por otro lado, en el capítulo anterior se realizó un estudio para determinar qué categorías eran las más devueltas. Esto constituye un aspecto más que el modelo comparará para establecer si un producto tiene probabilidades de ser devuelto o no.

Customer ID	CategDev
26993	NaN
22165	NaN
22165	Camisetas

Tabla 9. Ejemplo de variable 2: Categoría del producto devuelto

El cliente 14992 es NaN porque este solamente ha realizado una o ninguna devolución, por lo que no se puede comparar con sus devoluciones anteriores. En cambio, la categoría del cliente 22165 de su segunda devolución (la primera no se puede comparar con nada, por lo tanto devuelve NaN), es 'Camisetas'. Esto significa que la devolución realizada justo antes de la transacción estudiada se trata de un producto perteneciente a la categoría 'Camisetas'.

- Variable 3: Color del producto devuelto

Esta variable funcionará de la misma manera que la anterior, pero devolverá el color del producto devuelto en lugar de su categoría.

Customer ID	ColorDev
26993	NaN
22165	NaN
22165	Marrón

Tabla 10. Ejemplo de variable 3: Color del producto devuelto

Al igual que en la variable 2, en el caso del cliente 14992 esta variable no devuelve nada porque, al ser la primera transacción de este cliente, no tiene nada con qué comparar. En el caso del

cliente 22165, se puede afirmar que el color del último producto que devolvió con respecto a su segunda transacción de naturaleza devolución era 'Marrón'.

- Variable 4: Edad del cliente que realizó la devolución

En este caso, la variable devolverá la edad del cliente que realizó la devolución anterior. Al ser el mismo cliente, esta variable permanecerá constante en cada cliente. Se analiza también porque este será uno de los aspectos que el modelo estudiará al analizar una nueva transacción. La variable 4 funciona de la misma manera que las anteriores.

Customer ID	EdadDev
26993	NaN
22165	NaN
22165	Adulto

Tabla 11. Ejemplo de variable 4: Edad del cliente que realizó la devolución

- Variable 5: Género del cliente que realizó la devolución

Esta variable funcionará de la misma manera que las anteriores, pero en este caso devolverá el género del cliente que realizó la devolución, que también será constante por cliente.

Customer ID	GeneroDev
26993	NaN
22165	NaN
22165	H

Tabla 12. Ejemplo de variable 5: Género del cliente que realizó la devolución

- Variable 6: Clasificación del producto devuelto: top 10

Esta variable analizará si el producto correspondiente a la última devolución realizada por un cliente coincide con la descripción de alguno que pertenezca al top 10 de productos devueltos que se determinaron en el capítulo anterior.

Customer ID	Devolución	Top10Dev
26993	1	NaN
22165	1	NaN
22165	1	Yes

Tabla 13. Ejemplo de variable 6: Clasificación del producto devuelto: top 10

Una vez más, en el caso del cliente 26993 la respuesta es NaN porque se trata de su primera transacción. La primera respuesta del cliente 22165 es NaN porque también se trata de su primera transacción, pero la segunda afirma que el producto devuelto en la transacción anterior si coincide con alguno de los top10.

- Variable 7: Clasificación del producto devuelto: rango de precios

En el capítulo anterior se explicó que el hecho de que un producto perteneciera al rango de precios bajo, es decir, si su precio está entre 7,95€ y 55,95€, tiene más posibilidades tanto de ser comprado como de ser devuelto. Por este motivo es importante al analizar para todas las transacciones si los productos pertenecen o no a este rango. Esta variable enseña a qué rango pertenece el último producto devuelto de cada cliente.

Customer ID	Devolución	RangoDev
26993	1	NaN
22165	1	NaN
22165	1	Bajo

Tabla 14. Ejemplo de variable 7: Clasificación del producto devuelto: rango de precios

- Variable 8: Agrupación de clientes de mismo género y edad. Análisis de la categoría devuelta.

Esta variable analizará si el producto correspondiente a la última devolución realizada por un cliente coincide con la descripción de alguno que pertenezca al top 10 de productos devueltos que se determinaron en el capítulo anterior.

Customer ID	Devolución	Top10Dev
26993	1	NaN
22165	1	NaN
22165	1	Yes

Tabla 15. Ejemplo de variable 8: Agrupación de clientes de mismo género y edad. Análisis de categoría devuelta

Una vez más, en el caso del cliente 26993 la respuesta es NaN porque se trata de su primera transacción. La primera respuesta del cliente 22165 es NaN porque también se trata de su primera transacción, pero la segunda afirma que el producto devuelto en la transacción anterior si coincide con alguno de los top10.

- Variable 9: Agrupación de clientes de mismo género y edad. Análisis del color devuelto.

Esta variable analizará si el producto correspondiente a la última devolución realizada por un cliente coincide con la descripción de alguno que pertenezca al top 10 de productos devueltos que se determinaron en el capítulo anterior.

Customer ID	Devolución	Top10Dev
26993	1	NaN
22165	1	NaN
22165	1	Yes

Tabla 16. Ejemplo de variable 9: Agrupación de clientes de mismo género y edad. Análisis del color devuelto

Una vez más, en el caso del cliente 26993 la respuesta es NaN porque se trata de su primera transacción. La primera respuesta del cliente 22165 es NaN porque también se trata de su primera transacción, pero la segunda afirma que el producto devuelto en la transacción anterior si coincide con alguno de los top10.

- Variable 10: Agrupación de productos de mismo color y categoría. Análisis de la edad del cliente.

Esta variable analizará si el producto correspondiente a la última devolución realizada por un cliente coincide con la descripción de alguno que pertenezca al top 10 de productos devueltos que se determinaron en el capítulo anterior.

Customer ID	Devolución	Top10Dev
26993	1	NaN
22165	1	NaN
22165	1	Yes

Tabla 17. Ejemplo de variable 10: Agrupación de productos de mismo color y categoría. Análisis de edad del cliente

Una vez más, en el caso del cliente 26993 la respuesta es NaN porque se trata de su primera transacción. La primera respuesta del cliente 22165 es NaN porque también se trata de su primera transacción, pero la segunda afirma que el producto devuelto en la transacción anterior si coincide con alguno de los top10.

- Variable 11: Agrupación de productos de mismo color y categoría. Análisis del género del cliente.

Esta variable analizará si el producto correspondiente a la última devolución realizada por un cliente coincide con la descripción de alguno que pertenezca al top 10 de productos devueltos que se determinaron en el capítulo anterior.

Customer ID	Devolución	Top10Dev
26993	1	NaN
22165	1	NaN
22165	1	Yes

Tabla 18. Ejemplo de variable 11: Agrupación de productos de mismo color y categoría. Análisis del género del cliente

Una vez más, en el caso del cliente 26993 la respuesta es NaN porque se trata de su primera transacción. La primera respuesta del cliente 22165 es NaN porque también se trata de su

primera transacción, pero la segunda afirma que el producto devuelto en la transacción anterior si coincide con alguno de los top10.

Además de las variables previamente ilustradas, de la misma manera se crearon otras que buscan relacionar otros aspectos. Todas ellas estudian comportamientos pasados, intentando identificar patrones de compra y devolución.

Como ya se ha explicado, el modelo utilizará todas esas variables (*'predictors'*) para determinar si un producto va a ser devuelto o no, es decir, la respuesta será la variable *'Devolucion'*. A continuación se presenta un resumen del nombre utilizado para cada una de estas variables, incluyendo también aquellas que no han sido explicadas detalladamente.

Variable 1	UltTransac
Variable 2	UltCateg
Variable 3	UltColor
Variable 4	UltEdad
Variable 5	UltGenero
Variable 6	DevTop10
Variable 7	RangoDevAnt
Variable 8	GenEdad-Categ
Variable 9	GenEdad-Color
Variable 10	ColorCateg-Edad
Variable 11	ColorCateg-Gen
Variable 12	TallaProv
Variable 13	ProvSubc-Talla

Variable 14	CustIDSubcat
Variable 15	DescripTalla
Variable 16	ProvFecha
Variable 17	ProvSubc-Fecha
Variable 18	ProvCust-Talla
Variable 19	ProvCust-Subcat
Variable 20	SubcProv

3.4.3 Entrenamiento, validación y test

Tras crear las variables sobre antiguas transacciones que el modelo comparará con las nuevas, se procederá a dividir los datos en tres conjuntos diferentes con el fin de cuantificar de manera precisa el error que comete el modelo creado. Estos conjuntos son:

- Entrenamiento (training)

Se utiliza una muestra de los datos disponibles para entrenar el modelo, es decir, el modelo observa y aprende de estos datos. En este caso, se ha decidido que el conjunto de entrenamiento constituya un 80% de los datos.

- Validación

Set de datos que se utiliza para evaluar imparcialmente el modelo y ajustar sus parámetros según los resultados obtenidos en caso de ser necesario. En esta fase el modelo no 'aprende' de los datos. Por lo tanto, se puede decir que se trata de un conjunto de datos que afecta al modelo indirectamente. Este conjunto supone un 10% de los datos.

- Testing

Consiste en utilizar una muestra de datos para comprobar la eficacia del modelo una vez superados las fases de entrenamiento y validación. Contiene datos que abarcan todas las

situaciones posibles a las que se podría enfrentar el modelo. Se trata de un último paso que solo se puede llevar a cabo cuando el modelo está completamente entrenado.

Este conjunto es necesario porque para poder cuantificar el error se necesita contar con observaciones de las que se sepa el resultado, pero que el modelo no conozca. Constituye el 10% restante.

3.4.4 Upsampling

Como ya se ha mencionado anteriormente, la respuesta de este modelo se puede distinguir en dos clases: devuelto (1) y no devuelto (0). Sin embargo, estas están muy desbalanceadas, es decir, el número de observaciones de una de ellas es significativamente mayor.

Este problema se presenta en modelos que buscan identificar enfermedades no comunes, a la hora de encontrar cuentas fraudulentas… y también a la hora de predecir devoluciones. Sucede porque los algoritmos de aprendizaje automático suelen estar diseñados para mejorar la precisión reduciendo el error. Por este motivo, no tienen en cuenta el equilibrio de las clases.

Una de las principales consecuencias de este inconveniente es que hace que una de las métricas más utilizadas para evaluar la validez de un modelo, *accuracy*, sea engañosa. Es una medida ampliamente utilizada porque mide el porcentaje de observaciones que el modelo ha acertado. En un caso como este en el que las clases están muy desbalanceadas, el resultado es usualmente muy elevado, lo que en un principio puede parecer muy positivo. Considerando que en los datos proporcionados sólo el 2% son devoluciones, si el modelo estimase siempre 0 la accuracy sería de un 98%. Esto es un ejemplo de como esta métrica puede resultar confusa, y el por qué no se va a tener en cuenta como medida de evaluación.

No obstante, existen diferentes iniciativas para solucionar este desequilibrio. La escogida en este proyecto se conoce como *upsampling*. Se trata de un proceso que consiste en duplicar aleatoriamente las observaciones de la clase minoritaria para reforzar su presencia. Los pasos seguidos son [12]:

1. En primer lugar, se separan las observaciones de cada clase en diferentes sets de datos.
2. Se reestablece el número de muestras minoritarias para que coincida con las de la clase mayoritaria.
3. Finalmente, se combinan ambos sets de datos.

Una vez concluido este proceso, el modelo es capaz de predecir ambas clases.

3.4.5 Grid Search

En un principio, los valores que tomarán los hiperparámetros se determinan según la tipología del modelo, su objetivo, los datos disponibles… El Grid Search es un proceso mediante el cual el programa proporciona los valores ideales que deben tomar para optimizar las métricas que evalúan el rendimiento del modelo.

En primer lugar, se establece el número de árboles, su profundidad, la tasa de aprendizaje… es decir, los hiperparámetros más influyentes. Una vez obtenidos, se hallan los valores de otros hiperparámetros mediante una búsqueda aleatoria.

3.4.6 Rendimiento del modelo

La validez del modelo desarrollado se estudió analizando la *matriz de confusión* de los tres sets en los que se dividieron los datos proporcionados: entrenamiento, validación y test. Estas matrices se utilizan para medir el rendimiento del modelo. Es una tabla que presenta cuatro combinaciones diferentes de valores predichos y reales:

	Estimado 0	Estimado 1
Realidad 0	Verdadero negativo (TN)	Falso Positivo (FP)
Realidad 1	Falso Negativo (FN)	Verdadero positivo (TP)

En este caso, 0 significa 'no devuelto' y 1 'devuelto'. Además, es importante mencionar que el modelo proporciona un 1 cuando la probabilidad de que un producto sea devuelto es mayor a 58,45%. Este valor se determinó estudiando las predicciones que devuelve automáticamente el modelo.

- Verdadero negativo (TN): El modelo predijo exitosamente que el producto no iba a ser devuelto.
- Falso positivo (FP): Casos en los que el modelo predijo que el producto iba a ser devuelto, pero no se produjo la devolución.
- Falso negativo (FN): El modelo predijo que el producto no iba a ser devuelto, pero sí se realizó la devolución.
- Verdadero positivo (TP): Casos en los que el modelo predijo que el producto iba a ser devuelto, y efectivamente se produjo la devolución.

Las matrices de confusión de los sets de entrenamiento, validación y test se presentan a continuación:

ENTRENAMIENTO	Estimado 0	Estimado 1
Realidad 0	318202	425
Realidad 1	42	318593
Total	318244	319018

Tabla 19. Matriz de confusión del set de entrenamiento

VALIDACIÓN	Estimado 0	Estimado 1
Realidad 0	39257	76
Realidad 1	7	39694
Total	39264	39770

Tabla 20. Matriz de confusión del set de validación

TEST	Estimado 0	Estimado 1
Realidad 0	39908	57
Realidad 1	18	39798
Total	39926	39855

Tabla 21. Matriz de confusión del set de test

A través de los valores que proporcionan las matrices de confusión se pueden obtener diferentes variables estadísticas que se utilizan para evaluar la validez del modelo. Sin embargo, dependiendo de lo que se estudia algunas son más relevantes que otras.

Como proclamar todas las observaciones negativas no es útil, resulta obvio que la evaluación se debe centrar en la identificación de los casos positivos. Por este motivo, se buscará maximizar, entre otras, la métrica *recall,* que mide la capacidad de un modelo para encontrar todos los casos positivos dentro de un conjunto de datos.

El significado de las métricas que se utilizaron para estimar la validez del modelo se especifica a continuación.

- Misclassification

Representa el error general que el modelo comete. Responde a la pregunta ¿qué tanto se equivoca? Se calcula con la fórmula siguiente:

$$Misclassification = \left(\frac{FP + FN}{Total}\right) \cdot 100$$

- Precision

Mide la calidad del modelo en cuanto a clasificación. En este caso representa qué porcentaje de productos serían realmente devueltos del total de devoluciones estimadas por el modelo (independientemente de la realidad, es decir TP y FP).

$$Precisión = \frac{TP}{TP + FP} \cdot 100$$

- Recall

Esta métrica esta orientada a evaluar la cantidad de devoluciones reales que el modelo es capaz de reconocer.

$$Recall = \frac{TP}{TP + FN} \cdot 100$$

CAPÍTULO 4. ANÁLISIS DE RESULTADOS

4.1. RESULTADOS DEL CASO BASE

4.1.1. Validez del modelo

Como se ha mencionado anteriormente, existen muchas métricas con las cuales se puede validar el rendimiento de un modelo. Sin embargo, dependiendo de que estudie este unas son mas adecuadas que otras. Por motivos explicados en el capítulo 3.4.5, las elegidas en este estudio son misclassification, precision y recall.

A continuación se presenta la Tabla 22 con el fin de proporcionar una visión global de los valores que toman estas métricas en los conjuntos de entrenamiento, validación y test del modelo propuesto en específico.

	Entrenamiento	Validación	Test
Misclassification	0,07%	0,11%	0,09%
Precision	99,99%	99,98%	99,95%
Recall	99,87%	99,81%	99,86%

Tabla 22. Resultado de métricas para la validación del modelo

Como se puede observar, los valores de misclassification en los tres casos son bastante reducidos, lo que quiere decir que el modelo es bastante exacto. Por otra parte, los porcentajes de precision y recall son muy altos, lo que indica no solo que el modelo es capaz de identificar los relativamente pocos casos positivos, si no también que lo hace de manera precisa.

Además, se puede decir que el modelo no está sobreentrenado (overfitted). *Overfitting* quiere decir que el modelo aprende de los datos proporcionados, pero es incapaz de reconocer nuevos casos, pues estos no tienen exactamente los mismos valores que las muestras de entrenamiento. Como el error del conjunto test es muy parecido al error del conjunto de validación, se puede decir que el modelo no presenta esta característica.

Por otro lado, es importante tener en cuenta cuales son las variables que, según el modelo, tienen más importancia a la hora de predecir si un producto va a ser devuelto o no. En la Ilustración 18 se presenta el top 5, siendo estas por orden de relevancia:

- CustIDSubcat: Siendo la más relevante de todas, estudia a qué subcategoría pertenece el último producto devuelto por un determinado cliente. Muestra ser útil porque da información individual sobre que subcategorías suele devolver cada cliente.
- UltColor: Analiza de qué color es el último producto devuelto por un determinado cliente. Esta variable también proporciona información individual sobre cada cliente.
- GenEdad-Color: Agrupando los clientes por género y edad, estudia de qué color era el último producto devuelto. Es útil porque explica qué colores son los más devueltos por ciertos colectivos.
- GenEdad-Categ: Agrupando los clientes por género y edad, estudia a qué categoría pertenecía el último producto devuelto. Al igual que la anterior, es útil porque explica qué categorías son las que más devuelven ciertos grupos de clientes.
- SubcProv: Estudiando proveedor por proveedor, analiza a que categoría de todas las que comercializa pertenece el último producto devuelto. Proporciona información sobre cuáles subcategorías son las más problemáticas por proveedor.

Ilustración 18. Top 5 de variables importantes según el modelo

4.2. CASO DE NEGOCIO

Como ya se ha mencionado anteriormente, el objetivo de este proyecto es el desarrollo de un modelo capaz de predecir devoluciones, con el fin de utilizar esta información para reducir los costes que la logística inversa tiene en una compañía.

El primer paso es analizar el coste total de la implantación del proyecto. Una vez obtenido, se plantearán y evaluarán diferentes estrategias, para luego escoger la más adecuada.

4.2.1 Costo del proyecto

Este proyecto tendrá una duración de 3 meses. Su coste total se ha dividido en los costes de desarrollo e implantación del proyecto, y el coste de mantenimiento durante los 9 meses restantes del año. Para cuantificar los primeros se formará un equipo compuesto por un data scientist, un consultor de negocio, y un manager que supervisará las actividades. Para los segundos, se incluye el coste de mantenimiento de la infraestructura, además del equipo de soporte (el valor de una semana del data scientist al mes). Esto se encuentra resumido en la Tabla 23.

Es importante mencionar que tanto estos valores, la duración y el coste de infraestructura se han estimado teniendo en cuenta experiencia previa en proyectos similares.

Coste de desarrollo (3 meses)		Coste de mantenimiento (9 meses)	
Data scientist	20.000,00 €	**Infraestructura**	45.000,00 €
Consultor de negocio	40.000,00 €	**Equipo de soporte**	15.000,00 €
Manager	10.000,00 €		
Infraestructura	20.000,00 €		
Total	90.000,00 €		60.000,00 €

Tabla 23. Desglose del coste mensual del proyecto

Como la duración del proyecto son 3 meses, el coste total del proyecto son **150.000€.**

4.2.2 Estimación del coste de la logística inversa

Con el fin de estimar los costes de logística inversa de esta empresa, se ha calculado el 1,5% de la facturación total.

Facturación total	28.314.490,69 €
% de logística inversa	1,5%
Coste total	450.200,40 €

Para estimar el costo que causan las devoluciones por producto, se tuvieron en cuenta los siguientes datos:

	Número de artículos	**Coste total**
Devoluciones reales	17.665	450.200,40 €
Devoluciones predichas	17.660	450.065,33 €

Por lo tanto:

$$\frac{450.200,40}{17.665} = 25,40€$$

Para obtener el costo asociado a logística inversa de las devoluciones predichas se multiplicó el número de artículos por el costo por unidad.

Además, es importante mencionar que el costo de la logística inversa se puede desglosar en diferentes secciones detalladas a continuación [13]. Su proporción se presenta en la Tabla 24.

- Costes de transporte: Se refieren a los costes relacionados con movilizar el producto desde el punto de recogida hasta el almacén, traslados si es necesario a centros para repararlo, y movimientos en general necesarios para reutilizar, reciclar… el embalaje del producto.

- Costes de devolución: Incluye la recepción del producto, comprobar si está en buenas condiciones, coordinar los departamentos implicados en la devolución, el coste del espacio asignado en el almacén…
- Costes de reparación: Incluye el coste de valorar si vale la pena que un producto sea reparado en casa de presentar daños, y si es así, el coste de este proceso. También se refiere al costo de aprovechar ciertas piezas, guardarlas…
- Costes de atención al cliente: Coste de acordar el sitio y la fecha de la devolución con el cliente, documentar los motivos de la devolución, confirmar que se ha llevado a cabo, reembolsos…
- Costes de reventa: En caso de que el producto no presente daños, es necesario tener en cuenta los costes de rempaquetado para que este esté listo para volver a salir a la venta.

		Por unidad	Dev Reales	Dev Estimadas
Transporte	30%	7,65 €	135.060,12 €	135.022,71 €
Devolución	20%	5,10 €	90.040,08 €	90.015,14 €
Recuperación	15%	3,82 €	67.530,06 €	67.511,35 €
Atención al cliente	15%	3,82 €	67.530,06 €	67.511,35 €
Reventa	20%	5,10 €	90.040,08 €	90.015,14 €
TOTAL	100%	25,40 €	450.200,40 €	450.075,70 €

Tabla 24. Distribución de costes de logística inversa

4.2.3 Estrategias propuestas para la monetización del modelo

En este apartado se evalúan diferentes iniciativas que la empresa podría implementar para reducir el impacto que tienen los costes de logística inversa en su margen. Se plantean estrategias a corto plazo y estrategias a largo plazo. Estas últimas tienen un carácter más innovador, por lo que para estudiar su factibilidad es necesaria otro tipo de información, con la que no se cuenta en este proyecto.

A. Estrategias a largo plazo

- Si el modelo predice que un cliente determinado que está registrado en el Marketplace es propenso a devolver productos de una determinada categoría, color… sería interesante estudiar el algoritmo de la página web, para que esta **no promocionase en un primer plano** los productos que cuentan con esas características.

- En el caso de desarrollar un modelo que no esté enfocado en analizar el comportamiento del cliente, si no que se base en estudiar el producto, es decir, que se base en datos sobre fallos de calidad, daños en el envío…, se puede proponer iniciativas que se enfoquen en **identificar en qué parte de la cadena se producen estos fallos, e intentar prevenirlos.**

- Si se es capaz de desarrollar a mayores una plataforma en la que el cliente justifique la causa por la cual no se queda el producto, y al predecir devoluciones se reconoce que ciertos productos son devueltos con cierta regularidad, se puede plantear una estrategia en la que los proveedores se **replanteen el diseño** de esos productos, los materiales, la calidad…

Sin embargo, el estudio de la viabilidad económica se centró en las estrategias que la empresa podría implementar a corto plazo, que son las que se indican a continuación. Se tiene en cuenta el coste del proyecto, diferentes fuentes de ingreso, la aceptación de los clientes…

B. Estrategias a corto plazo

- **<u>Estrategia 1:</u> Que el cliente pague una cierta cantidad de dinero por producto si el modelo predice que éste será devuelto.**

Como ya se ha mencionado, esta estrategia considera que el cliente pague una penalización por producto si el modelo estima que va a ser devuelto. Los resultados de analizar su factibilidad se presentan en la Tabla 25.

	3,00 €	5,00 €	6,00 €	7,00 €	9,20 €
Ingreso extra al año	52.980,32 €	88.300,53 €	105.960,64 €	123.620,75 €	162.499,77 €
Ingreso extra al mes	4.075,41 €	6.792,35 €	8.150,82 €	9.509,29 €	12.499,98 €
Payback (meses)	36,81	22,08	18,40	15,77	12,00
Payback (años)	3,1	1,8	1,5	1,3	1,0

Tabla 25. Análisis económico de la estrategia 1

El *payback* es un criterio para evaluar el período de retorno de una inversión. Se trata de determinar cuándo llega a esta al punto de equilibrio.

$$Payback = \frac{Inversión\ inicial}{Flujos\ de\ caja}$$

La viabilidad de esta estrategia no es óptima. A pesar de que la inversión se puede recuperar en un período de tiempo de un año o año y medio, el precio por producto es elevado.

- **<u>Estrategia 2:</u> Que el cliente pague un porcentaje de lo que cuesta el producto (según su categoría) si se estima que éste será devuelto.**

Esta estrategia consiste en cobrarle al cliente un porcentaje del precio del producto que el modelo ha predicho que va a ser devuelto. Para simplificar los cálculos, el porcentaje se ha calculado sobre el precio medio por producto por categoría.

Categoría	5%	10%	15%	20%
Abrigos	9.246,59 €	18.493,18 €	27.739,78 €	36.986,37 €
Accesorios	1.774,33 €	3.548,65 €	5.322,98 €	7.097,31 €
Bisutería	381,65 €	763,31 €	1.144,96 €	1.526,62 €
Bolsos	1.063,61 €	2.127,23 €	3.190,84 €	4.254,45 €
Botas	611,01 €	1.222,02 €	1.833,04 €	2.444,05 €
Camisas y tops	2.566,28 €	5.132,57 €	7.698,85 €	10.265,13 €
Camisetas	745,86 €	1.491,71 €	2.237,57 €	2.983,42 €
Cinturones	1.061,13 €	2.122,26 €	3.183,39 €	4.244,52 €
Faldas	2.758,43 €	5.516,86 €	8.275,29 €	11.033,71 €
Jerséis	4.907,99 €	9.815,97 €	14.723,96 €	19.631,94 €
Monos	579,34 €	1.158,67 €	1.738,01 €	2.317,34 €
Pantalones	2.490,65 €	4.981,30 €	7.471,95 €	9.962,61 €
Sastrería	661,29 €	1.322,58 €	1.983,87 €	2.645,16 €
Sudaderas	1.078,65 €	2.157,30 €	3.235,95 €	4.314,60 €
Trajes	904,69 €	1.809,37 €	2.714,06 €	3.618,74 €
Vaqueros	787,51 €	1.575,01 €	2.362,52 €	3.150,02 €
Vestidos	2.177,90 €	4.355,79 €	6.533,69 €	8.711,59 €
Zapatos	2.900,70 €	5.801,40 €	8.702,10 €	11.602,81 €
Ingreso extra al año	36.697,59 €	73.395,19 €	110.092,78 €	146.790,38 €
Ingreso extra al mes	2.822,89 €	5.645,78 €	8.468,68 €	11.291,57 €
Payback (meses)	53,1	26,6	17,7	13,3
Payback (años)	4,4	2,2	1,5	1,1

Tabla 26. Análisis económico de la estrategia 2

Con el fin de optimizar la estrategia, se calculó el porcentaje a pagar por el cliente para que la inversión se recupere en alrededor de un año. Como se puede observar en la Tabla 26, este porcentaje sería un 20%.

Según la categoría, los precios a pagar por el cliente que se corresponden a ese 20% son los que se indican en la Tabla 27. Se puede observar como algunos de ellos son elevados, por ejemplo, el de abrigos o sastrería. Esto indica que es posible que el cliente no esté dispuesto a pagarlo.

Categoría	Precio medio / categoría	20%
Abrigos	80,34 €	16,07 €
Accesorios	25,06 €	5,01 €
Bisutería	12,62 €	2,52 €
Bolsos	33,39 €	6,68 €
Botas	69,04 €	13,81 €
Camisas y tops	31,05 €	6,21 €
Camisetas	20,38 €	4,08 €
Cinturones	21,70 €	4,34 €
Faldas	40,51 €	8,10 €
Jerséis	48,45 €	9,69 €
Monos	53,15 €	10,63 €
Pantalones	35,38 €	7,08 €
Sastrería	70,35 €	14,07 €
Sudaderas	22,95 €	4,59 €
Trajes	56,37 €	11,27 €
Vaqueros	32,95 €	6,59 €
Vestidos	43,13 €	8,63 €
Zapatos	47,79 €	9,56 €

Tabla 27. Estrategia 5. Cantidad a pagar por producto por categoría

- **<u>Estrategia 3:</u> Que el cliente pague un porcentaje de lo que cuesta la devolución del producto (según su categoría) si se estima que éste será devuelto.**

En este caso se dividieron las devoluciones según la categoría a la que pertenezca el producto devuelto, pues no es lo mismo los costes que acarrea un abrigo a los de un accesorio. La diferencia con la estrategia anterior yace en que en este caso el porcentaje se calcula con respecto al coste de devolución, y no al precio del producto. Entonces, se plantean diferentes escenarios en los que varía el porcentaje del coste que el cliente pagaría. El ingreso extra que la empresa obtendría gracias a esta estrategia en cada uno de los casos, se indica en la Tabla 28.

Categoría	10%	20%	30%	40%
Abrigos	10.796,97 €	21.593,93 €	32.390,90 €	43.187,87 €
Accesorios	2.079,37 €	4.158,75 €	6.238,12 €	8.317,50 €
Bisutería	362,49 €	724,97 €	1.087,46 €	1.449,94 €
Bolsos	1.438,84 €	2.877,68 €	4.316,52 €	5.755,36 €
Botas	698,08 €	1.396,16 €	2.094,24 €	2.792,32 €
Camisas y tops	3.373,31 €	6.746,62 €	10.119,93 €	13.493,24 €
Camisetas	1.065,99 €	2.131,98 €	3.197,96 €	4.263,95 €
Cinturones	596,15 €	1.192,29 €	1.788,44 €	2.384,59 €
Faldas	2.987,63 €	5.975,26 €	8.962,89 €	11.950,52 €
Jerséis	3.692,22 €	7.384,44 €	11.076,66 €	14.768,89 €
Monos	1.605,14 €	3.210,28 €	4.815,42 €	6.420,55 €
Pantalones	2.707,33 €	5.414,66 €	8.121,99 €	10.829,32 €
Sastrería	648,81 €	1.297,62 €	1.946,43 €	2.595,24 €
Sudaderas	1.747,66 €	3.495,32 €	5.242,99 €	6.990,65 €
Trajes	1.342,74 €	2.685,48 €	4.028,22 €	5.370,96 €
Vaqueros	966,32 €	1.932,65 €	2.898,97 €	3.865,30 €
Vestidos	2.120,09 €	4.240,18 €	6.360,27 €	8.480,37 €
Zapatos	3.927,56 €	7.855,12 €	11.782,68 €	15.710,24 €
Ingreso extra al año	42.156,70 €	84.313,40 €	126.470,10 €	168.626,80 €
Ingreso extra al mes	3.242,82 €	6.485,65 €	9.728,47 €	12.971,29 €
Payback (meses)	46,3	23,128	15,4	11,6
Payback (años)	3,9	1,9	1,3	1,0

Tabla 28. Análisis económico de la estrategia 3

Se ha determinado que si el cliente pagase un 40% del costo de sus devoluciones, la inversión se recuperaría en un año. Se ha escogido este como el caso más viable con el fin de minimizar el payback, pero que la cantidad a pagar sea viable para el cliente.

CAPÍTULO 5. CONCLUSIONES

5.1. CONCLUSIONES SOBRE LOS RESULTADOS

Con el fin de proporcionar iniciativas eficientes a corto plazo que la empresa pueda llevar a cabo para reducir el impacto que tiene el coste de las devoluciones, se ha tenido en cuenta:

- El ingreso adicional anual que supondrían las diferentes estrategias.
- El coste de las devoluciones, calculado como un 3% de la facturación total.
- El tiempo de retorno de la inversión inicial, es decir, el payback.
- La posibilidad y disposición de los clientes a pagar dinero extra que contrarresten los costes.

Una vez analizados los resultados de la viabilidad económica de cada una de las estrategias, se determinó que la más rentable es **que el cliente pague un porcentaje de lo que cuesta la devolución del producto (según su categoría) si se estima que éste será devuelto.**

Es importante tener en cuenta la categoría a la que pertenece el producto, pues no es lo mismo retornar una prenda de invierno, es decir, pesada y que ocupa más espacio como un abrigo, a un accesorio.

Tras analizar diferentes proporciones a pagar por el cliente, se determinó que el porcentaje que este debe pagar es un 40%. De esta manera, la empresa obtendría 168.626,80 € de ingreso extra anuales, y recuperaría la inversión en un año. Además, para evaluar el rendimiento financiero se calculó el ROI (retorno de la inversión):

$$ROI = \frac{Ingresos - coste}{coste} \cdot 100 = 12{,}42\%$$

Los flujos de caja mensuales hipotéticos se calcularon teniendo en cuenta el coste de devolución por artículo según su categoría, y se presentan en la Tabla 29. Además, tal y como se enseña en la Ilustración 19, los valores más elevados coinciden en los meses donde se analizó previamente que se realizaban más devoluciones en el capítulo 3.3.

Diciembre 2017	14.304,64 €
Enero 2018	14.671,07 €
Febrero 2018	10.279,22 €
Marzo 2018	12.231,80 €
Abril 2018	10.187,10 €
Mayo 2018	5.454,28 €
Junio 2018	7.926,65 €
Julio 2018	24.987,84 €
Agosto 2018	11.302,37 €
Septiembre 2018	11.742,36 €
Octubre 2018	6.641,11 €
Noviembre 2018	11.459,02 €
Diciembre 2018	21.366,25 €
TOTAL	162.553,72 €

Tabla 29. Flujos de caja mensuales de la estrategia propuesta

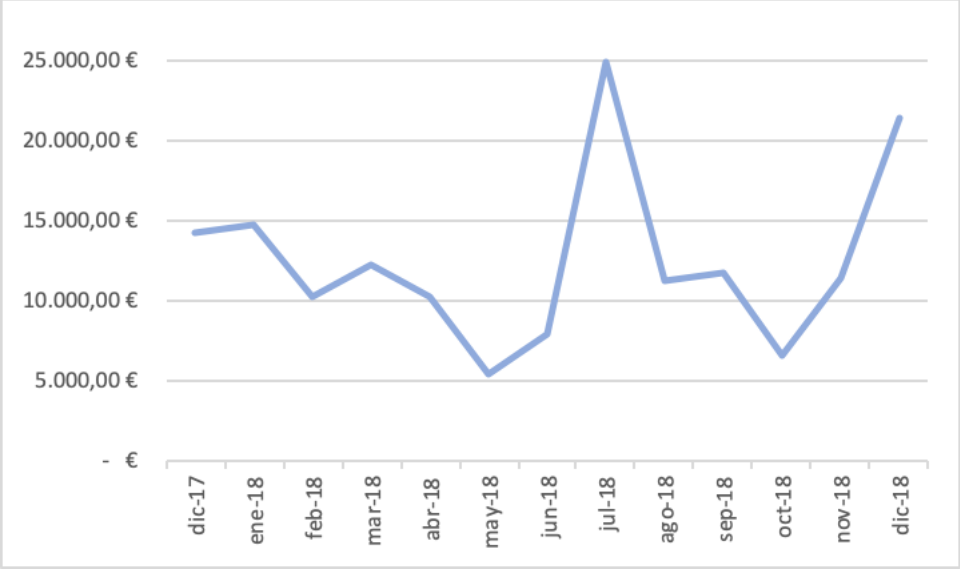


Ilustración 19. Gráfico de los flujos de caja mensuales de la estrategia propuesta

Con el fin de proporcionar una visión más clara de lo que supondría esta estrategia para los clientes, la penalización a la que estaría sujeto cada uno de ellos sería de **3,10 €** al año. Esto quiere decir que bastaría con añadir alrededor de 0,50 € por transacción como coste de envío, por ejemplo, si se supone que cada cliente realiza 6 compras al año.

5.2. RECOMENDACIONES PARA FUTUROS ESTUDIOS

Como ya se ha mencionado anteriormente, el objetivo de este proyecto es obtener un modelo que sea capaz de predecir devoluciones de la manera más precisa posible. Con el fin de optimizarlo, sería útil añadir otras variables que afecten a la decisión del cliente, como por ejemplo:

- Tiempo disponible para realizar una devolución.
- La calidad percibida de la prenda.
- La similitud entre el producto online y el real.
- El por qué de la devolución: el producto llegó más tarde de lo esperado, el producto es erróneo, fallo de tallas…

A pesar de que estos datos no son tan fácilmente conseguidos como la edad o el género que se obtienen cuando el usuario se registra en la web, sería interesante proponer un sistema que sea capaz de incluirlos. De esta manera, se proporcionaría un modelo que evalúa la transacción teniendo en cuenta estas otras variables que también son relevantes.
Más allá de esto, convendría ser capaz de ampliar el alcance del modelo, siendo capaz así de predecir otro tipo de información relevante como la fecha exacta de la devolución, el motivo…

Por otro lado, otra recomendación para futuros estudios sería incluir en la web iniciativas relacionadas con la realidad virtual, para que el cliente tuviese una visión más clara de lo que va a recibir. Así, se reducirían potencialmente las devoluciones por problemas de talla, calidad, color…

A pesar de que ya existen estudios sobre esto, es cierto que no hay muchas tiendas online que las ofrezcan. Además, sería una oportunidad para endurecer un poco las políticas de devolución de la empresa en cuestión. Como se ha explicado, hoy en día existe una ‘obligación’ de ofrecer devoluciones rápidas, cómodas y gratuitas para poder satisfacer al cliente. Es más probable que estos finalicen una compra si las devoluciones se llevan a cabo fácilmente. Por ello, la mayor parte de tiendas online cuentan con este tipo de pautas.

Por ejemplo, la marca de gafas de sol Ray-ban ofrece un sistema que permite que el usuario se las pruebe antes de comprarlas. Como consecuencia, su política de devolución no es gratuita, además de avisar de la posibilidad de que se realicen cargos adicionales si la mercancía devuelta no está en óptimas condiciones.

CAPÍTULO 6. BIBLIOGRAFÍA

[1] Catchoom, "eCommerce In The Fashion Industry. Industry Changes, Stats & Trends. Image Recognition, AR and Artificial Intelligence Solutions," 2018. [Online]. Available: https://catchoom.com/blog/how-is-ecommerce-changing-fashion-industry-stats-trends-predictions/?cn-reloaded=1.

[2] F. Ma, "The Study on Reverse Logistics for E-Commerce," IEEE Conference Publication, 2010. [Online]. Available: https://ieeexplore.ieee.org/document/5575577 .

[3] W. Kofler, "Artificial Intelligence in Retail – What to expect and how to act," PwC. [Online]. [Accessed 2019].

[4] S. Cullinane, M. Browne, E. Karlsson and E. Wang, "Retail Clothing Returns: A Review of Key Issues," 2019. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-030-14493-7_16.

[5] Y. Huang, S. Hu and S. Chih-Chieh Huang, "Exploration of Virtual Reality-Based Online Shopping Platform," 2019. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-030-23525-3_11 .

[6] P. Yildirim, D. Birant and T. Alpyldiz, "Data mining and machine learning in textile industry," 2017. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/widm.1228.

[7] X. Li and Y. Zhuang, "A trust-aware random walk model for return propensity estimation and consumer anomaly scoring in online shopping," 2019. [Online]. Available: http://scis.scichina.com/en/2019/052101.pdf.

[8] P. Cruz Bolívar, "Uso de Big Data para la toma de decisiones acordes a la estrategia empresarial en el sector retail," 2017. [Online]. Available: https://repository.unilibre.edu.co/bitstream/handle/10901/11204/Monograf%c3%ada%20Big%20Data%20-%20Paola%20Cruz%20Bol%c3%advar.pdf?sequence=1&isAllowed=y.

[9] T. Johnson, "The Future of Fashion: How Artificial Intelligence is Transforming the Apparel Industry," Tinuiti, 2019. [Online]. Available: https://tinuiti.com/blog/ecommerce/future-of-fashion/ .

[10] "The H2O Python Module," [Online]. Available: http://docs.h2o.ai/h2o/latest-stable/h2o-py/docs/intro.html.

[11] J. Amat Rodrigo, «Machine learning con H2O y Python,» 2020. [En línea]. Available: https://www.cienciadedatos.net/documentos/py04_machine_learning_con_h2o_y_python#Gradient-Boosting-Machine-(GBM).

[12] Elite Data Science, «How to handle imbalanced classes in machine learning,» [En línea]. Available: https://elitedatascience.com/imbalanced-classes.

[13] «Costes ocultos de devoluciones,» [En línea]. Available: https://www.mecalux.es/blog/gestion-de-devoluciones-costes.

[14] S. Davis, M. Hagerty and E. Gerstner, "Return policies and the optimal level of "hassle"." 1998. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0148619598000137.

[15] I. Bilgen and O. Sinan Saraç, "Prediction of return in online shopping," IEEE Conference Publication, 2015. [Online]. Available: https://ieeexplore.ieee.org/document/7130411.

[16] J. Hammond and K. Kohler, "Harvard Edu Project," 2010. [Online]. Available: http://projects.iq.harvard.edu/files/hctar/files/ec01.pdf .

[17] W. Jinfu and Z. Aixiang, "E-Commerce in the Textile and Apparel Supply Chain Management: Framework and Case Study," IEEE Conference Publication, 2009. [Online]. Available: https://ieeexplore.ieee.org/document/5209816.

[18] Forrester, «Crafting a Returns Policy that Creates a Competitive Advantage Online,» [En línea]. Available: https://www.ups.com/media/en/returns_forrester.pdf.

CAPÍTULO 7. APÉNDICES

A: OBJETIVOS DE DESARROLLO SOSTENIBLE (ODS) DE NACIONES UNIDAS

Introducción

Los objetivos de Desarrollo Sostenible, también conocidos como las Metas Globales, fueron adoptados por todos los estados miembros de las Naciones Unidas en 2015 con el fin de erradicar la pobreza, proteger el planeta y asegurar la paz y la prosperidad para todos los ciudadanos del mundo para 2030 [1]. Estos objetivos son:

Ilustración 20. Objetivos de Desarrollo Sostenible

Los ODS se aplican a todos los países, tanto a los desarrollados como a los que están en vías de desarrollo, y establecen las prioridades de sus gobiernos. Los cambios en los ámbitos demográfico y social, los cambios en el poder económico, urbanización, cambio climático, escasez de recursos, la desigualdad y los avances tecnológicos exigen una respuesta colectiva e internacional. Los ODS establecen la agenda de todas las naciones para hacer frente a estos desafíos con un énfasis en un crecimiento positivo que proporcione valor social y ambiental, así como un beneficio económico óptimo.

ODS en PwC

Existen numerosas razones para que una empresa o un proyecto se centren en prácticas comerciales sostenibles. Algunas de ellas son:

- Las prácticas comerciales sostenibles mejoran el *rendimiento*: las empresas con mejores ratings en asuntos medioambientales, sociales y de gobierno, tienen costos menores de deuda y equity, pues el mercado reconoce que son de menor riesgo y los recompensa en consecuencia. Además, generalmente superan el rendimiento del mercado a mediano y largo plazo.
- Gama de *beneficios comerciales* más amplia. Las empresas afirman que las estrategias de sostenibilidad aportan valor a través de la reducción de riesgos de operación, de reputación y de regulación. También disminuyen costos de operación y de la cadena de suministro, aumentan el valor del producto y el crecimiento a través de nuevos mercados o innovación de productos.

Como se puede observar en la Ilustración 21, PwC aborda todos los ODS mientras cumple sus dos propósitos de resolver problemas y de generar confianza en la sociedad [2]. Considera que el número 17 se encuentra implícito en todos los demás.

ODS	España	Impacto PwC	Oportunidad PwC	Factores Críticos Exito	Relevancia
1. No Poverty	●				BAJO
2. Zero hunger					BAJO
3. Good health and wellbeing		●			MEDIO
4. Quality education	●	●		●	ALTO
5. Gender equality		●	●	●	ALTO
6. Clean water and sanitation					BAJO
7. Affordable and clean energy			●		BAJO
8. Decent work and economic growth	●	●	●	●	ALTO
9. Industry, innovation and infrastructure				●	MEDIO
10. Reduced inequalities				●	MEDIO
11. Sustainable cities and communities					BAJO
12. Responsible consumption and production			●		MEDIO
13. Climate Action		●	●		MEDIO
14. Life below water	●				BAJO
15. Life on land	●				BAJO
16. Peace, justice and strong institutions				●	MEDIO
17. Partnerships for the goals					N/A

Ilustración 21. Priorización de los ODC en PwC

Los empleados de la firma a través de los diferentes equipos y en los proyectos que participan, contribuyen a proporcionar valor económico y social, además de ayudar a entender a otras empresas los ODS, mediante la promoción de temas relacionados con ellos.

En resumen, se podría decir que, en este momento, los ODS más relevantes para las actividades de la firma se presentan en la Ilustración 22 y en la Ilustración 23.

Ilustración 22. Foco principal: ODS de prioridad ALTA para PwC

Ilustración 23. Contribución directa: ODS de prioridad MEDIA para PwC

ODS en este proyecto

A pesar de que todos ellos están integrados, es decir, todos ellos persiguen el mismo fin, se puede considerar que los ODS que están más relacionados con los objetivos de este proyecto son el número 9 (‘Industria, innovación e infraestructura’) y el número 12 (‘Producción y consumo responsables’). Cabe destacar que ambos contribuyen directamente a la empresa dentro de la cual se realiza este proyecto.

Ilustración 24. ODS alineados con el proyecto

El *ODS 9* acentúa el hecho de que el progreso tecnológico es inevitable y solamente positivo si se lleva a cabo sin perjudicar el medioambiente [1]. Se trata de construir una infraestructura fuerte, promover la inclusión y la industrialización sostenible, además de fomentar la innovación. A pesar de que la creación de infraestructura se ha percibido tradicionalmente como responsabilidad de los gobiernos, factores actuales como el crecimiento de la población y los efectos del clima explican que cada vez hay más oportunidades para que las empresas apliquen sus recursos y experiencia en este campo.

El sector manufacturero es uno de los principales agentes de este ODS. Es por ello por lo que cada vez hay más empresas dispuestas a invertir en la optimización de sus procesos a través de

machine learning. Una inversión temprana e inteligente puede ahorrar costes de mantenimiento, ineficiencia y recursos naturales, además de asegurar un sistema mejor en el que poder prosperar.

Además de las iniciativas mencionadas en los capítulos previos, se puede decir que el desarrollo de un modelo capaz de predecir devoluciones es un claro ejemplo. El proceso de analizar datos de numerosas transacciones con el objetivo de obtener información sobre patrones de devolución es una propuesta de iniciativa innovadora que pueden tomar las empresas del sector retail para refrescar esta industria.

El *ODS 12* tiene como lema "garantizar modalidades de consumo y producción sostenibles" [1]. El hecho de que la población esté en aumento, la urbanización y el crecimiento económico están impulsando una demanda cada vez mayor de recursos naturales: energía, suelo, agua y minerales. Si las actuales tendencias de consumo continúan, los recursos naturales podrían agotarse rápidamente.

Este es un tema preocupante para todos. Numerosas iniciativas como la fabricación de prendas, mobiliario, envases… con materiales reciclados y movimientos para reducir el consumo masivo de plástico, son cada vez más populares. Mejorar la eficiencia de los recursos también significa menos desperdicio. Esto incluye un espectro muy amplio de iniciativas: materiales desechados, grandes cantidades de comida desaprovechada… La reducción de las devoluciones y la posibilidad de revender objetos devueltos son ejemplos de iniciativas que reducen el malgasto de recursos.

El objetivo final de este ODS es la creación de una economía circular, sin residuos ni contaminación, no sólo a través de la reutilización y el reciclaje, sino también a través de reparaciones, diseñándolas para que duren más y estableciendo modelos de negocio más sostenibles. Extender la vida del producto, reutilizarlo… reduce la necesidad de nuevos recursos y reduce los impactos de la eliminación de desechos.

Como ya se ha explicado, además de aumentar los márgenes de operación de las empresas del sector retail, la reducción de las devoluciones tiene como trasfondo la disminución de la

producción de unidades que más tarde van a ser desechadas, provocando un impacto negativo en el medio ambiente. Si una empresa es capaz de predecir el número y el tamaño de las órdenes que envían al departamento de fabricación, podrán reducir su inventario, y con ello el riesgo de que productos se vuelvan obsoletos y tengan que ser eliminados. Desechar prendas consume recursos y energía, además de contaminar el agua y el aire.

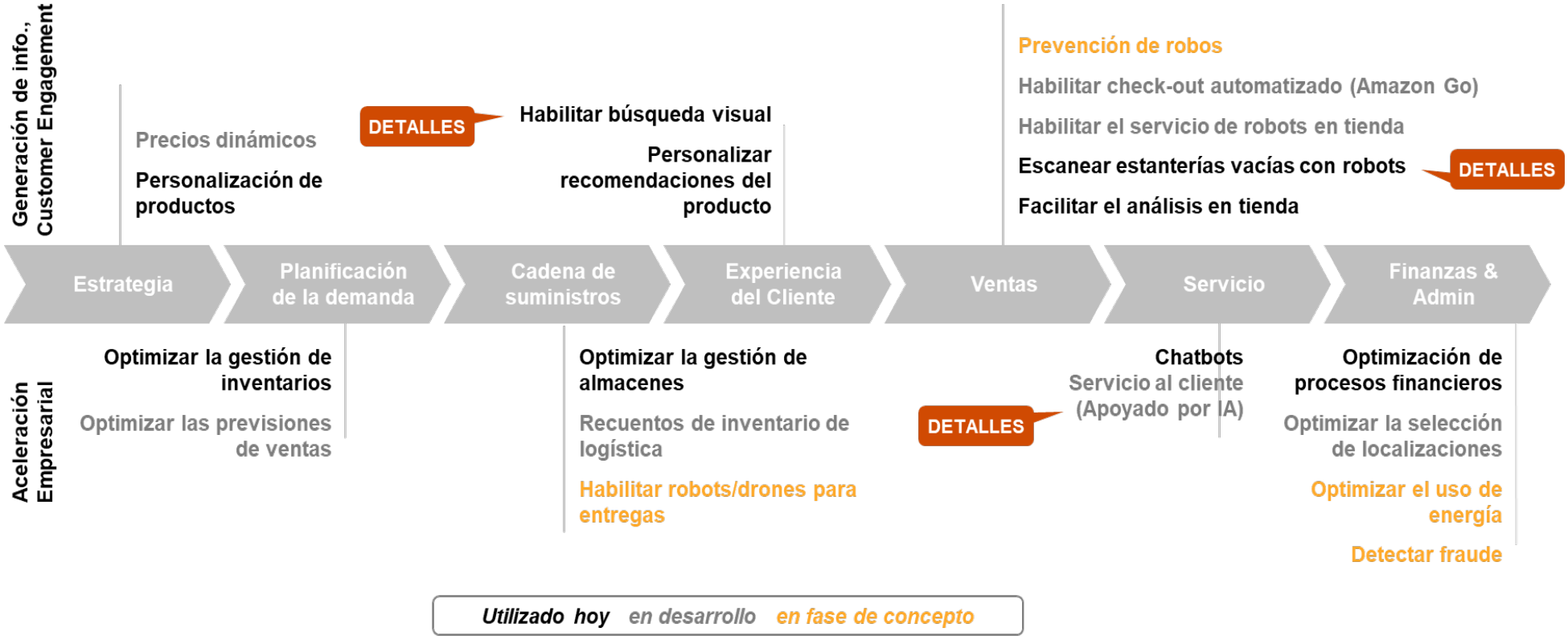
Tras desarrollar el modelo predictivo y analizar su impacto en la industria, se puede afirmar que los objetivos de este proyecto están alineados con los ODS propuestos por la Organización de las Naciones Unidas.

Referencias

[1] "Objetivos de Desarrollo Sostenible," Organización de las Naciones Unidas, 2015. [Online]. Available: https://www.un.org/sustainabledevelopment/es/sustainable-development-goals/.

[2] "Objetivos de Desarrollo Sostenible," PwC España, [Online]. Available: https://www.pwc.es/es/quienes-somos/rsc/objetivos-desarrollo-sostenible.html.

B: IMPACTO DE LA INTELIGENCIA ARTIFICIAL EN LA CADENA DE VALOR DE UN RETAILER



C: CÓDIGO EN PYTHON DEL MODELO PREDICTIVO

```
import h2o
import pandas as pd

from h2o.estimators.gbm import H2OGradientBoostingEstimator
from h2o.grid.grid_search import H2OGridSearch
from datetime import datetime


# Inicializar el modulo h2o

h2o.init()

# Tranformación de Datos

df=pd.read_csv(r'/Users/lauralegall/Desktop/TFMSpyder/Excel/DatosRaw
/datos.csv', encoding='utf-8', sep=';', decimal='.')

df['InvoiceDate'] = pd.to_datetime(df['InvoiceDate'], format =
'%d/%m/%y')

#                  CREACION DE VARIABLES

# ============================================================
#       VARIABLE 1: La última transacción fue una devolución?
# ============================================================

#Para estudiar lo que sucede justo en la transacción anterior
df['UltTransac'] = df.groupby(['CustomerID'])['Devolucion'].shift(1)

#Para ordenar por fecha

df['InvoiceDate'] = pd.to_datetime(df['InvoiceDate'])

df['UltTransac'] = (df.sort_values(by=['InvoiceDate'],
ascending=True).groupby(['CustomerID'])['Devolucion'].shift(1))


# ============================================================
#  VARIABLE 2: A qué CATEGORIA pertenece el producto devuelto?
# ============================================================

# Para quedarme solamente con las devoluciones (Devolucion=1)
```

```
dfDev=df[df.Devolucion.isin(df.groupby(['CustomerID'])['Devolucion']
.max())]

dfDev['UltCateg'] =
dfDev.groupby(['CustomerID'])['Categoria'].shift(1)
# ============================================================
#        VARIABLE 3: De qué COLOR es el producto devuelto?
# ============================================================

dfDev['UltColor'] = dfDev.groupby(['CustomerID'])['Color'].shift(1)


# ============================================================
#      VARIABLE 4: Qué EDAD tiene el cliente que realizó la
devolución?
# ============================================================

dfDev['UltEdad'] = dfDev.groupby(['CustomerID'])['Edad'].shift(1)

# ============================================================
#   VARIABLE 5: De qué GÉNERO es el cliente que realizó la
devolución?
# ============================================================

dfDev['UltGenero'] =
dfDev.groupby(['CustomerID'])['Genero'].shift(1)

# ============================================================
#         VARIABLE 6: El producto devuelto está en el TOP 10
vendidos?
# ============================================================

# Creo una lista con el top 10
top10 = ['CAMISETA ALGODON MANGA LARGA','JERSEY ESCOTE
PICO','BUFANDA LISA FLECOS','JERSEY LANA','DEPORTIVAS PIEL','ABRIGO
BORREGUILLO','ABRIGO LARGO PUNTO','BLUSA BORDADA ALGODON','BLAZER
CRUZADA','SUDADERA CANGURO TEXTURA']

dfDev ['PertTOP10'] = dfDev.apply(lambda row: 'Yes' if
row['Description'] in top10 else 'No', axis = 1)


dfDev['DevTop10'] =
dfDev.groupby(['CustomerID'])['PertTOP10'].shift(1)
```

```
# ============================================================
#        VARIABLE 7: En qué RANGO DE PRECIOS está el producto
devuelto?
# ============================================================

dfDev ['PertRango'] = dfDev.apply(lambda row: 'Bajo' if
(row['UnitPrice'] >= 7.95 and row['UnitPrice'] <= 55.95) else
('Medio' if (row['UnitPrice'] >= 59.95 and row['UnitPrice'] <=
105.95) else 'Premium'), axis = 1)

dfDev ['RangoDevAnt'] =
dfDev.groupby(['CustomerID'])['PertRango'].shift(1)


# ============================================================
# VARIABLE 8: Clientes mismo género, edad. Categoría anterior?
# ============================================================

dfDev['GenEdad-Categ'] = dfDev.groupby(['Genero',
'Edad'])['Categoria'].shift(1)


# ============================================================
#    VARIABLE 9: Clientes mismo género, edad. Color anterior?
# ============================================================

dfDev['GenEdad-Color'] = dfDev.groupby(['Genero',
'Edad'])['Color'].shift(1)


# ============================================================
#          VARIABLE 10: Productos mismo color, categoría. Edad
anterior?
# ============================================================

dfDev['ColorCateg-Edad'] = dfDev.groupby(['Color',
'Categoria'])['Edad'].shift(1)


# ============================================================
#         VARIABLE 11: Productos mismo color, categoría. Genero
anterior?
# ============================================================

dfDev['ColorCateg-Gen'] =
dfDev.groupby(['Color','Categoria'])['Genero'].shift(1)
```

```
# ============================================================
#      VARIABLE 12: Mismo proveedor. Talla última devolución?
# ============================================================

#Para estudiar si un determinado proveedor tiene problemas con x
tallas

dfDev['TallaProv'] = dfDev.groupby(['Proveedor'])['Talla'].shift(1)


# ============================================================
#   VARIABLE 13: Mismo proveedor y subcategoría. Talla última
devolución?
# ============================================================

# Por si un proveedor, fabrica mal las tallas de una subcategoría

dfDev['ProvSubc-Talla'] = dfDev.groupby(['Proveedor',
'Subcategoria'])['Talla'].shift(1)


# ============================================================
#        VARIABLE 14: Mismo cliente. Subcategoria última devolución?
# ============================================================

# Que subcategoría es más probable que devuelva un cliente
específico

dfDev['CustIDSubcat'] =
dfDev.groupby(['CustomerID'])['Subcategoria'].shift(1)


# ============================================================
#     VARIABLE 15: Misma descrip. Talla última devolución?
# ============================================================

# Analizar si un producto en concreto está mal tallado

dfDev['DescripTalla'] =
dfDev.groupby(['Description'])['Talla'].shift(1)


# ============================================================
#      VARIABLE 16: Mismo proveedor. Fecha última devolución?
# ============================================================
```

```
# Analizar si en ciertas fechas el proveedor fabrica con menos
calidad por la demanda

dfDev['ProvFecha'] =
dfDev.groupby(['Proveedor'])['InvoiceDate'].shift(1)

# ============================================================
#     VARIABLE 17: Mismo proveedor y subcat. Fecha última
devolución?
# ============================================================

dfDev['ProvSubc-Fecha'] = dfDev.groupby(['Proveedor',
'Subcategoria'])['InvoiceDate'].shift(1)


# ============================================================
#       VARIABLE 18: Mismo proveedor y cliente. Talla última
devolución?
# ============================================================

dfDev['ProvCust-Talla'] = dfDev.groupby(['Proveedor',
'CustomerID'])['Talla'].shift(1)


# ============================================================
#  VARIABLE 19: Mismo proveedor y cliente. Subcategoría última
devolución?
# ============================================================

dfDev['ProvCust-Subcat'] = dfDev.groupby(['Proveedor',
'CustomerID'])['Subcategoria'].shift(1)


# ============================================================
#                VARIABLE 20: Mismo proveedor. Subcateg última
devolución?
# ============================================================

dfDev['SubcProv'] =
dfDev.groupby(['Proveedor'])['Subcategoria'].shift(1)
```

```
# ============================================================
#                          Unir bases de datos
# ============================================================

dfResult = pd.merge(df, dfDev, how = 'outer')



# ============================================================
#                 1. UPSAMPLE MINORITY CLASS
# ============================================================


from sklearn.utils import resample

# Separar las clases mayoritaria y minoritaria
df_maj = dfResult[dfResult.Devolucion==0]
df_min = dfResult[dfResult.Devolucion==1]

# Upsample la clase minoritaria
df_minority_upsampled = resample(df_minority,
                                 replace=True,
                                 n_samples=397925,
                                 random_state=123)

# Combinar la clase mayoritaria con la clase minoritarian upsampled
df_upsampled = pd.concat([df_majority, df_minority_upsampled])

#Guardar como excel

df_upsampled.to_excel('/Users/lauralegall/Desktop/df_upSampled.xlsx'
, index=False)



# ============================================================
#                             MODELO
# ============================================================

from h2o.automl import H2OAutoML

# Pasar a h2o el Dataframe

dfResult =
h2o.import_file(path=r'/Users/lauralegall/Desktop/df_upSampled.csv')
```

```
# Asegurarse de que la variable de respuesta es de tipo factor
dfResult['Devolucion'] = dfResult['Devolucion'].asfactor()



# Variables
predictores = ['UltTransac', 'UltCateg', 'UltColor', 'UltEdad',
'UltGenero', 'DevTop10', 'GenEdad-Categ', 'GenEdad-Color',
'ColorCateg-Edad', 'ColorCateg-Gen', 'TallaProv','ProvSubc-
Talla','CustIDSubcat','DescripTalla', 'ProvFecha','ProvSubc-Fecha',
'ProvCust-Talla', 'ProvCust-Subcat', 'SubcProv']
var_respuesta = 'Devolucion'


# Establecer el criterio de separacion entre el training y la
validacion (simple)
# 80% train 10% valid 10% test

datos_train_h2o, datos_val_h2o , datos_test_h2o =
dfResult.split_frame(
                                                    ratios=[0.8,
0.1],

destination_frames= ["datos_train_H2O",

"datos_val_h2o",

"datos_test_H2O"],
                                                    seed = 123
                                                )


from h2o.estimators.gbm import H2OGradientBoostingEstimator
modelo_gbm = H2OGradientBoostingEstimator(
                        # Tipo de distribución (clasificación binaria)
                        distribution = "bernoulli",
                        # Número de árboles.
                        ntrees = 200,
                        # Complejidad de los árboles
                        max_depth = 20,
                        min_rows = 10,
                        # Aprendizaje
                        learn_rate = 0.01,
                        balance_classes = True,
                        # Detención temprana
```

```
                        sample_rate = 0.8,
                        col_sample_rate = 0.8,
                        stopping_rounds = 3,
                        stopping_metric = "misclassification",
                        stopping_tolerance = 0.001,
                        model_id = "modelo",
                        seed = 123
)

# Train

modelo.train(
    # Variable respuesta y predictores.
    y = var_respuesta,
    x = predictores,
    # Datos de entrenamiento.
    training_frame = datos_train_h2o,
    # Datos de validación para estimar el error.
    validation_frame = datos_val_h2o
)

#Para estudiar la métrica de los datos

perf = modelo.model_performance(datos_test_h2o)
perf

#Para evaluar la importancia de los predictores
modelo.varimp(use_pandas=True)
modelo.varimp_plot(num_of_features=13)
```