

Data Science en el ámbito industrial - Previsión de la demanda promocional

**CRISTIAN SUBIRANA FIDONE**

Máster en Data Science

Área 5

**Lorena Polo Navarro**

**Jordi Conesa Caralt**

25/06/2021

  
Esta obra está sujeta a una licencia de Reconocimiento-NoComercial-SinObraDerivada [3.0 España de Creative Commons](http://creativecommons.org/licenses/by-nc-nd/3.0/es/)

**FICHA DEL TRABAJO FINAL**

|  |  |
| --- | --- |
| **Título del trabajo:** | *Data Science en el ámbito industrial - Previsión de la demanda promocional* |
| **Nombre del autor:** | ***Cristian Subirana Fidone*** |
| **Nombre del consultor/a:** | ***Lorena Polo Navarro*** |
| **Nombre del PRA:** | ***Jordi Conesa Caralt*** |
| **Fecha de entrega (mm/aaaa):** | 06/2021 |
| **Titulación:** | *Máster en Data Science* |
| **Área del Trabajo Final:** | *5* |
| **Idioma del trabajo:** | Castellano |
| **Palabras clave** | *Machine Learning, Minería de datos, estadística* |
| **Resumen del Trabajo (máximo 250 palabras):** *Con la finalidad, contexto de aplicación, metodología, resultados y conclusiones del trabajo.* | |
| Este proyecto nace de la necesidad analítica de una empresa en concreto, en que se requiere de encontrar un modelo capaz de prever la demanda de productos a fin de ajustar su stock, generando una reducción tanto de coste como de stock.  El objetivo del estudio es principalmente generar un modelo capaz de generar mejores resultados del actual modelo que ya dispone la empresa, a fin de saber si los resultados son mejores nos basaremos en la precisión que nos ofrecerá el modelo estudiado.  Para conseguir la meta propuesta inicialmente se ha de estudiar las variables disponibles en el dataset proporcionado por la empresa y realizar los ajustes necesarios a fin de poder iniciar un análisis de datos óptimo. Acto seguido, al disponer de los datos preparados para iniciar el análisis se realizará un estudio de múltiples modelos que puedan encajar con el ámbito de los datos de estudio, a fin de acotar las múltiples posibilidades de estudio de modelos disponibles. De entre estos modelos, se estudiará para cada uno de ellos cuál ofrece mejores resultados para conseguir la meta propuesta.  Finalmente, el modelo seleccionado nos ha de ofrecer un mejor ajuste de la demanda de productos, ha de presentar mejor precisión que el actual modelo de la empresa y principalmente, ha de mostrar una reducción de coste gracias a un mejor ajuste de la demanda. | |
| **Abstract (in English, 250 words or less):** | |
| This project arises from the analytical need of a specific company, in which it is required to find a model capable of forecasting the demand for products to adjust its stock, generating a reduction in both cost and stock.  The objective of the study is mainly to generate a model capable of generating better results than the current model that the company already has, to know if the results are better, we will base ourselves on the precision that the studied model will offer us.  To achieve the proposed goal, we must initially study the variables available in the dataset provided by the company and make the necessary adjustments to be able to start an optimal data analysis. Then, once the data are ready to start the analysis, a study of multiple models that may fit the scope of the data under study will be carried out, to narrow down the multiple possibilities of studying the models available. Among these models, we will study for each one of them which one offers the best results to achieve the proposed goal.  Finally, the selected model must offer us a better adjustment of the product demand, it should present better precision than the current model of the company and, mainly, it must show a cost reduction thanks to a better adjustment of the demand. | |

**Índice**

[1. Introducción 1](#_Toc72768191)

[1.1 Contexto y justificación del Trabajo 1](#_Toc72768192)

[1.2 Motivación personal 2](#_Toc72768193)

[1.3 Objetivos del Trabajo 3](#_Toc72768194)

[1.3.1 Objetivo Principal 3](#_Toc72768195)

[1.3.2 Objetivos Secundarios 3](#_Toc72768196)

[1.4 Enfoque y método seguido 4](#_Toc72768197)

[1.5 Planificación del Trabajo 6](#_Toc72768198)

[2.Marco teórico y revisión bibliográfica 8](#_Toc72768199)

[2.1 Marco teórico 8](#_Toc72768200)

[2.2 Revisión bibliográfica 12](#_Toc72768201)

[2.3 Conclusiones 14](#_Toc72768202)

[3. Diseño e implementación 16](#_Toc72768203)

[3.1 Recogida inicial de datos y almacenamiento 16](#_Toc72768204)

[3.2 Preparación, limpieza y preprocesado de los datos 19](#_Toc72768205)

[3.3 Análisis de los datos 28](#_Toc72768206)

[3.3.1 Fases de entrenamiento 28](#_Toc72768207)

[3.3.2 Correlación entre variables 31](#_Toc72768208)

[3.3.3 Estudio de la estacionalidad de los datos 33](#_Toc72768209)

[3.3.4 Conjunto de datos de entrenamiento y de test 36](#_Toc72768210)

[3.3.5 Tratamiento de Outliers 37](#_Toc72768211)

[3.3.6 Método de evaluación de los modelos 40](#_Toc72768212)

[3.4 Algoritmos de Machine Learning 41](#_Toc72768213)

[3.4.1 XGBoost 41](#_Toc72768214)

[3.4.2 LightGBM 46](#_Toc72768215)

[3.4.3 SARIMA 51](#_Toc72768216)

[3.4.4 LTSM 52](#_Toc72768217)

[3.4.5 Análisis comparativo de los resultados 55](#_Toc72768218)

[4. Conclusiones 58](#_Toc72768219)

[4.1 Logro objetivo 58](#_Toc72768220)

[4.2 Lecciones aprendidas 61](#_Toc72768221)

[4.3 Líneas de trabajo futuro 61](#_Toc72768222)

[5.Bibliografía 62](#_Toc72768223)

[Anexos 63](#_Toc72768224)

**Lista de Figuras**

[Figure 1 CRISP-DM metodología 4](#_Toc72767370)

[Figure 2 Diagrama de Gantt 7](https://d.docs.live.net/e00db0ff5955c623/Escritorio/TFM_CristianSubirana_es.docx#_Toc72767371)

[Figure 3 Dataset Ventas 16](#_Toc72767372)

[Figure 4 Dataset Promociones 17](#_Toc72767373)

[Figure 5 Dataset Stock 17](#_Toc72767374)

[Figure 6 Dataset Previsiones empresa 18](#_Toc72767375)

[Figure 7 Descripción datos Ventas 21](#_Toc72767376)

[Figure 8 Rango de fechas de los datos 21](#_Toc72767377)

[Figure 9 Descripción datos Stock 23](#_Toc72767378)

[Figure 10 Descripción datos Promociones 25](#_Toc72767379)

[Figure 11 Estructura dataset Promociones 26](#_Toc72767380)

[Figure 12 Rangos de fechas de los datasets 27](#_Toc72767381)

[Figure 13 Rango de fechas con Stock 28](#_Toc72767382)

[Figure 14 Muestra de datos Fase 2 29](#_Toc72767383)

[Figure 15 Muestra de código División datos test y training 29](#_Toc72767384)

[Figure 16 Variables creadas en Fase 3 30](#_Toc72767385)

[Figure 17 Matriz de correlación 31](#_Toc72767386)

[Figure 18 Matriz de correlación con variables extra 32](#_Toc72767387)

[Figure 19 Test de Dickey-Fullers 33](#_Toc72767388)

[Figure 20 Estudio estacionalidad - Año 34](#_Toc72767389)

[Figure 21 Estudio estacionalidad - Mes 35](#_Toc72767390)

[Figure 22 Estudio estacionalidad - Día 35](#_Toc72767391)

[Figure 23 Fragmento código Test y Training 36](#_Toc72767392)

[Figure 24 Outliers por día de la semana 37](#_Toc72767393)

[Figure 25 Fragmento de código tratamiento outliers 37](#_Toc72767394)

[Figure 26 Limpiez outliers Venta 38](#_Toc72767395)

[Figure 27 Limpieza outliers Precio 39](#_Toc72767396)

[Figure 28 Limpiez outliers Stock 39](#_Toc72767397)

[Figure 29 Fórmula RMSE 40](#_Toc72767398)

[Figure 30 Mejores parámetros fase 2 XGBoost 42](#_Toc72767399)

[Figure 31 Fragmento de código estudio parámetros XGBoost Fase 2 42](#_Toc72767400)

[Figure 32 Fragmento de código - definción XGBoost 42](#_Toc72767401)

[Figure 33 Evaluación RMSE XGBoost Fase 2 42](#_Toc72767402)

[Figure 34 Importancia de variables XGBoost Fase 2 43](#_Toc72767403)

[Figure 35 Mejores parámetros fase 3 XGBoost 44](#_Toc72767404)

[Figure 36 Evaluación RMSE XGBoost Fase 3 44](#_Toc72767405)

[Figure 37 Importancia de variables XGBoost Fase 3 45](#_Toc72767406)

[Figure 38 Funcionamiento LightGBM 46](#_Toc72767407)

[Figure 39 Mejores parámetros fase 2 LightGBM 47](#_Toc72767408)

[Figure 40 Fragmento de código estudio parámetros LightGBM Fase 2 47](#_Toc72767409)

[Figure 41 Fragmento de código - definción LightGBM 48](#_Toc72767410)

[Figure 42 Importancia de variables LightGBM Fase 2 49](#_Toc72767411)

[Figure 43 Mejores parámetros fase 3 LightGBM 49](#_Toc72767412)

[Figure 44 Importancia de variables LightGBM Fase 3 50](#_Toc72767413)

[Figure 45 Fragmento de código - definición SARIMA 51](#_Toc72767414)

[Figure 46 Mejores parámetros fase 2 LTSM 52](#_Toc72767415)

[Figure 47 Fragmento de código - definción LTSM 53](#_Toc72767416)

[Figure 48 Fragmento de código estudio parámetros LTSM Fase 3 53](#_Toc72767417)

[Figure 49 Mejores parámetros fase 3 LTSM 54](#_Toc72767418)

[Figure 50 Comparativa algoritmos 55](#_Toc72767419)

[Figure 51 Comparativa resultados - Original vs Predicho 55](#_Toc72767420)

[Figure 52 Comparativa resultados - Original vs Predicho empresa 57](#_Toc72767421)

[Figure 53 Ejemplo comparativa Venta Real vs Previsión Empresa vs Previsión 57](#_Toc72767422)

[Figure 54 Dimensión unidades empresa 59](#_Toc72767423)

[Figure 55 Dimensión unidades predicción 59](#_Toc72767424)

# 1. Introducción

## 1.1 Contexto y justificación del Trabajo

Hasta hace relativamente pocos años, las empresas tomaban sus decisiones basándose en la experiencia o instinto. Esto se debía a que las empresas eran un sector falto de información. Actualmente, han de adaptarse a un nuevo entorno mucho más analítico y mucho más rápido. Debido a esta velocidad se hace más difícil gestionar la información y tomar decisiones adecuadas, provocando que la mayoría acabe apostando por la experiencia y la intuición.

Para resolver este problema nace el concepto de Data driven, en que resumidamente se refiere a basarse en datos, el dato pasa a ser el foco principal.

Actualmente las empresas cuentan con más información que nunca, pudiendo registrar, clasificar y llegar a utilizarla para la toma de decisiones.

Encontramos muchas razones por las que las empresas que se basan en el enfoque “Data driven” puede ayudar a mejorar a las empresas.

Entre estos beneficios podemos encontrar:

* Mejor gestión del stock
* Reducción de costes
* Previsión de la demanda
* Mejor experiencia del cliente
* Mejor desarrollo de productos
* Claridad

Estos son algunas de las muchas ventajas que podemos encontrar, pero en este proyecto nos centraremos en la previsión de la demanda

## 1.2 Motivación personal

Los motivos por los que se ha elegido realizar este proyecto es debido a varios factores. Principalmente por mi actual empleo, ya que este tipo de proyectos están muy relacionados con el área en la que me encuentro y pueden aportar valor en la empresa.

Para llevar a cabo este proyecto no se utilizarán los datos de la empresa en la que trabajo, pero si los resultados obtenidos en este proyecto son buenos se reproducirá lo más similar posible en mi actual empresa.

Además, se espera que al final del proyecto ser capaz de gestionar y tener claro el desarrollo de este tipo de proyectos de analítica teniendo en cuenta las casuísticas diferentes que pueden presentar los datos.

Finalmente se espera poder demostrar que las decisiones basadas en datos pueden dar resultados mucho mejores que los basados en la experiencia o creencias.

## 1.3 Objetivos del Trabajo

### 1.3.1 Objetivo Principal

El objetivo de este trabajo es poder ofrecer un modelo estadístico capaz de prever la demanda de productos de una empresa en concreto proporcionando una precisión superior a la ya existente por parte de la empresa.

Durante el estudio se establecen los siguientes objetivos:

* Estudio del dataset. Se pretende realizar un estudio general de los datos presentados por la empresa a fin de conocer de qué tratan junto con ver su formato.
* Estudiar cada una de las variables. En este punto se estudiará cada una de las variables a fin de saber todas sus casuísticas y su formato.
* Preparar dataset. Una vez finalizado el estudio comentado será necesario realizar las modificaciones necesarias de los datos a fin de poder obtener unos resultados futuros óptimos.
* Estudiar posibles modelos candidatos según la tipología de datos.
* Seleccionar modelos más adecuados para iniciar el estudio a fin de saber cuál se adecua mejor a los datos.
* Validar resultados de precisión para cada modelo. En este apartado se pretende. Se espera obtener un modelo de entre el resto que proporcione la mayor precisión.
* Presentación de resultados en base a resultados obtenidos y explicación de los beneficios que se obtienen del modelo.

### 1.3.2 Objetivos Secundarios

Como objetivo secundario se espera que este proyecto gracias a ser revisado por expertos en el área sirva de ejemplo o guía para futuros proyectos parecidos en lo referente al ámbito de los datos, permitiendo realizar proyectos parecidos con una base revisada y estudiada.

Además, se espera que este proyecto pueda aportar una mejora real a la empresa en cuestión en lo referente a previsión de la demanda.

## 

## 1.4 Enfoque y método seguido

Para iniciar este proyecto es necesario disponer de dos partes. La primera el dataset o conjunto de datos a estudiar y la segunda y más importante entender la necesidad de negocio que requiere la empresa en cuestión a fin de aportar la solución deseada.

Para desarrollar el proyecto se seguirá la metodología de CRISP-DM, común en proyectos de minería de datos.

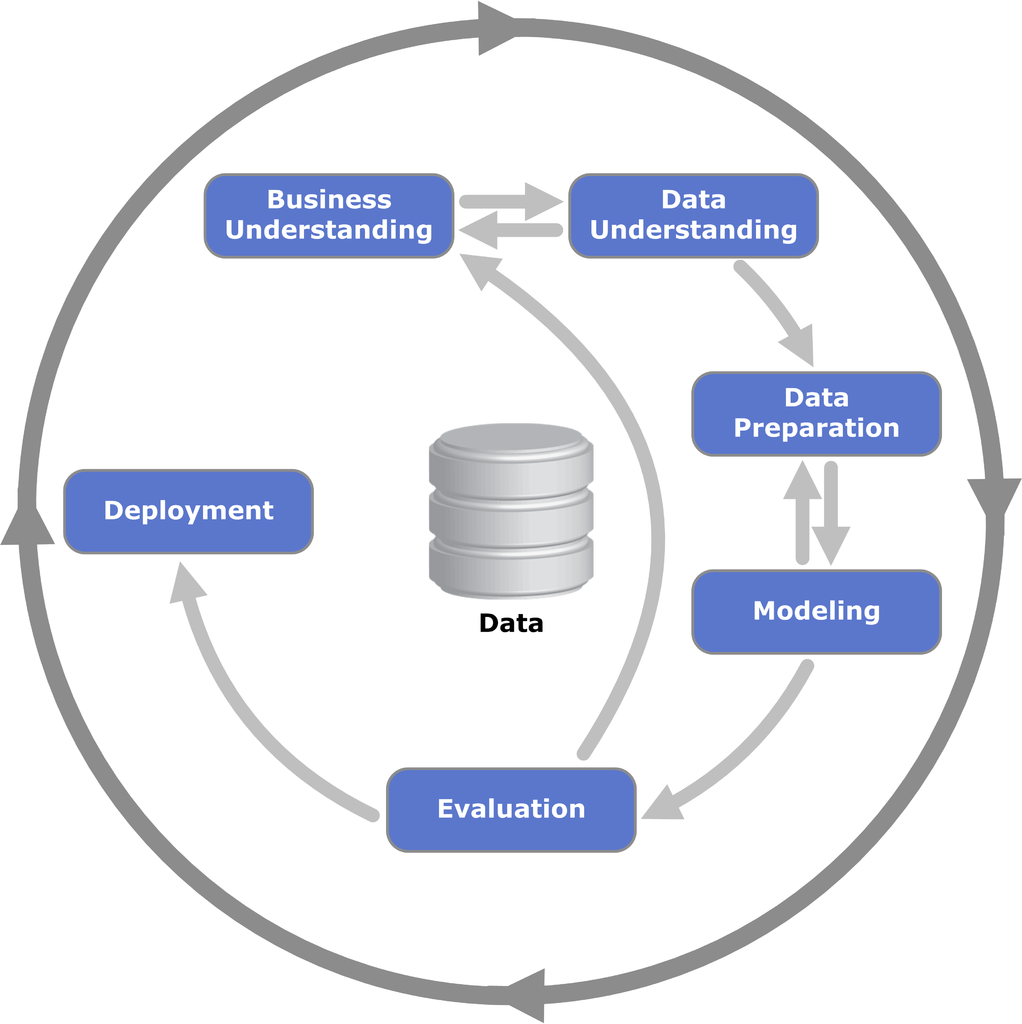
[](https://commons.wikimedia.org/wiki/File:CRISP-DM_Process_Diagram.png)

Figure 1 CRISP-DM metodología

Siguiendo las etapas de la metodología seleccionada:

**1.4.1 Entender el negocio**

En este apartado se trata de conocer a qué se dedica la empresa y cómo funciona, ya que por ejemplo es importante saber si su stock es propio del sector de la alimentación o propio de la industria.

**1.4.2 Conocer los datos**

En esta fase hemos de conocer los datos, los campos que los componen, su distribución y estructura.

**1.4.3 Preparación de los datos**

Una vez conozcamos los datos hemos de realizar un proceso de preparación de ellos a fin de poder realizar un análisis óptimo, aplicando tareas de selección, limpieza de datos, formatear, etc.

**1.4.4 Modelado**

Al disponer de los datos preparados en que no se requiere de ninguna modificación ni tratamiento se procede a modelar los datos preparados a fin de obtener un modelo capaz de ofrecer la mayor precisión posible. Se seleccionarán las técnicas más idóneas según los datos en cuestión junto con técnicas de estudio de los parámetros.

**1.4.5 Evaluación**

Se evaluarán los resultados de cada modelo basándose principalmente en la precisión de estos, teniendo en cuenta que se deberán seleccionar los parámetros óptimos para cada modelo.

**1.4.6 Entorno de producción**

En la última fase se debe disponer de un modelo capaz de proporcionar una precisión óptima junto con los parámetros correctos. El modelo ha de ser capaz predecir los futuros datos de la empresa si fuera seleccionado para ser utilizado habitualmente.

## 1.5 Planificación del Trabajo

Para realizar la planificación se ha tenido en cuenta las fechas definidas de entregas de las PAC’s.

Para definir la planificación correctamente su utilizará un diagrama de Gantt y una tabla en que simplemente se reflejarán las fechas de cada paso.

|  |  |  |  |
| --- | --- | --- | --- |
| Fase Principal | Fase | Inicio | Fin |
| PAC 1 | Definición objetivos, interés personal | 2021/02/17 | 2021/02/20 |
| PAC 1 | Redacción PAC 1 | 2021/02/21 | 2021/02/28 |
| PAC 2 | Estudio del dataset | 2021/03/01 | 2021/03/02 |
| PAC 2 | Estudio del negocio | 2021/03/03 | 2021/03/04 |
| PAC 2 | Estudio de trabajos parecidos | 2021/03/05 | 2021/03/08 |
| PAC 2 | Modelos más idóneos para el proyecto | 2021/03/09 | 2021/03/14 |
| PAC 2 | Redacción PAC 2 | 2021/03/15 | 2021/03/21 |
| PAC 3 | Estudio y preparaciones variables dataset | 2021/03/22 | 2021/03/30 |
| PAC 3 | Construir modelos | 2021/03/31 | 2021/04/09 |
| PAC 3 | Evaluación modelos | 2021/04/10 | 2021/05/02 |
| PAC 3 | Evaluación parámetros | 2021/05/03 | 2021/05/11 |
| PAC 3 | Modelo final y conclusiones | 2021/05/12 | 2021/05/15 |
| PAC 3 | Redacción PAC 3 | 2021/05/16 | 2021/05/23 |
| PAC 4 | Revisión final memoria | 2021/05/24 | 2021/05/31 |
| PAC 4 | Redacción PAC 4 | 2021/06/01 | 2021/06/06 |
| PAC 5 | Preparación presentación | 2021/06/07 | 2021/06/09 |
| PAC 5 | Preparación video para la defensa | 2021/06/10 | 2021/06/11 |
| PAC 5 | Redacción PAC 5 | 2021/06/12 | 2021/06/12 |

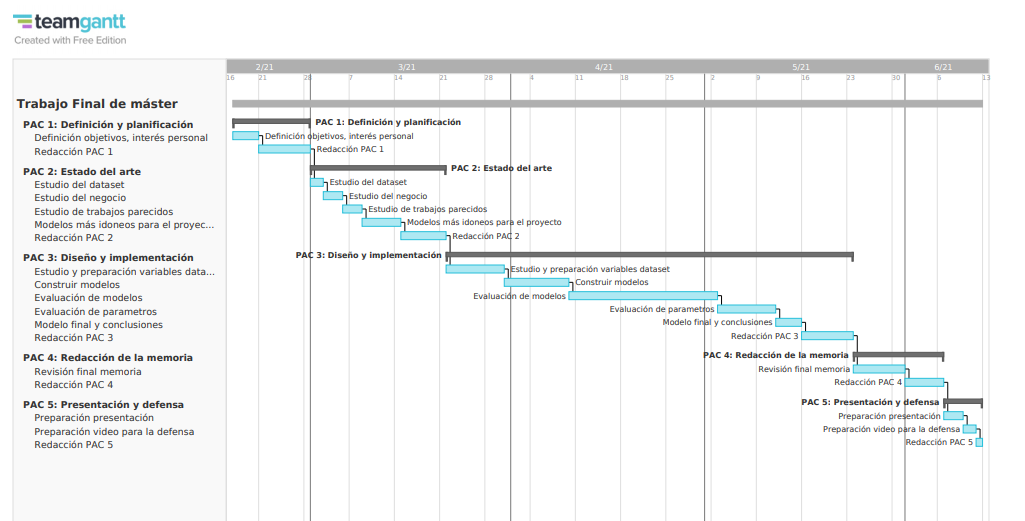


Figure 2 Diagrama de Gantt

# 2.Marco teórico y revisión bibliográfica

En este apartado se tratarán los antecedentes y artículos de investigación relacionados con la temática del proyecto. La función de este apartado es conocer el máximo posible la temática a fin de llegar a ser un experto o lo más cerca posible.

Este conocimiento nos permitirá tener un enfoque correcto de los pasos a seguir, ahorrándonos tiempo de investigación y permitiéndonos tener una idea mejor enfocada.

### 2.1 Marco teórico

La predicción de la demanda se refiere a desarrollar la mejor previsión de la demanda futura. En la práctica esto se refiere a analizar el impacto de las variables en la futura demanda, des de datos históricos, decisiones de negocio y factores externos.

En lo referente a la granularidad, podemos encontrar varios niveles, des de mensual, semanal, diario o a nivel de horas. Este factor es importante según la tipología de productos que estudiemos, por ejemplo, si los productos fueran alimentos frescos o no.

A lo largo de este apartado hablaremos de los múltiples factores para tener en cuenta y que pueden beneficiar obtener una mayor precisión en el análisis.

**Beneficios de la predicción de la demanda**

Des del punto de vista de los usuarios que se basan en una predicción de la demanda, encontramos múltiples beneficios, en el que destacaríamos principalmente, reducir la incerteza, permitiendo planificar mejor el stock, hacer pedidos a proveedores más precisos, optimizar el almacén, entre otros factores.

Resumiríamos estos beneficios en los siguientes puntos:

* Incrementan las ventas gracias a disponer de una mejor disponibilidad
* Se reduce la cantidad de productos que quedan estancados en el almacén por largos periodos de tiempo.
* Reducción de coste personal ya que se optimiza el trabajo indirectamente.
* Se mejora capacidad de utilización del almacén.
* Permite tratar mejor los posibles cuellos de botella.

**¿Como se está prediciendo actualmente la demanda?**

Des de empresas grandes a empresas pequeñas se basan en la predicción de la demanda a fin pedir la cantidad idónea de productos que esperan vender de la manera más precisa posible.

La diferencia recae en qué se basan, dónde podemos encontrar el formato más tradicional, en que se basan en hojas de cálculo y cálculos manuales, o los que utilizan software especializado que se basan en análisis estadísticos. Finalmente encontraríamos el formato más moderno basado en software automatizado basado en modelos propios de machine learning.

**Factores a tener en cuenta**

**Decisiones de negocio internas**

Las decisiones propias de negocio tales como promociones, cambio de precios, nuevos productos o cambios simplemente en como mostramos los productos, tienen un impacto muy grande en el volumen de ventas. Estas decisiones pueden producir una variación muy difícil de predecir, por lo que son factores importantes para tener en cuenta al realizar la predicción.

Las variables que se espera tratar son:

* Tipo de promociones, tales como reducción de precio o pack’s.
* Actividades de marketing
* Promoción visual de productos
* Impacto al cambiar de precio sobre el resto de los productos parecidos o de las familias parecidas. Es decir, si bajamos el precio de un producto puede generar bajar la demanda del resto de los productos de la categoría, este producto absorbería las ventas de sus parecidos.
* Introducción de nuevos productos. Este caso puede ser difícil de determinar su efecto ya que no contamos con datos históricos que nos sirvan de apoyo

**Factores externos**

En los referentes a los factores externos nos referimos a aquellos fuera de nuestro alcance, en los que no tenemos ningún control ni como desarrolladores ni como empresarios.

Encontraríamos como ejemplos, los festivos locales, la predicción del tiempo, las decisiones de negocio de nuestra competencia, todos estos factores pueden impactar en nuestra predicción.

Estos factores son muy difíciles de tratar con la mente humana, ya que por ejemplo si quisiéramos predecir el tiempo en base al histórico en todas las tiendas de una empresa nos resultaría muy difícil, pero gracias a algoritmos propios por ejemplo de Machine learning conseguiríamos aproximarnos.

Los factores para tener en cuenta en este apartado son:

* Meteorología local
* Eventos locales
* Precios de la competencia

### 2.2 Revisión bibliográfica

El artículo (Abolghasemi, Mahdi, Gerlach, Richard, Tarr, Garth, & Beh, Eric, 2009) nos presenta una visión de un estudio realizado utilizando múltiples métodos o algortimos para poder preveer la demanda, evaluando cada uno de ellos su rendimiento para poder obtener una idea sobre sun rendiemiento según el tipo de datos.

Este estudio primeramente nos da una visión de como podemos evaluar nuestro futuro modelo para saber si los resultados obtenidos son buenos, además nos proporciona un conjunto de modelos candidatos a ser estudiados para la necesidad de este proyecto.

Para evaluar los modelos estudiados se han basado en el error medio absoluto escalado (MASE) debido a que se encuentra variaciones en la demanda al escalar los datos. De este estudio podemos destacar que los modelos que mejor comportamiento tienen son ARIMA y ANN, en que ARIMA presenta un buen rendimiento y es robusto a la volatibilidad, en cambio ANN cuando la volatibilidad es alta su rendimiento baja notablemente. Finalmente podríamos concluir que estudiar el modelo de ARIMA y sus variantes (SARIMAX, ARIMAX entre otros) puede ser un buen punto de partida.

En el artículo (Sustrova, 2016) se nos presenta un estudio de un conjunto de metodos enfocados en el uso de redes neuronales con el fin de optimizar la previsión en la demanda. En este articulo se analizan múltiples tipos de redes neuronales evaluandolos en base a su coeficiente de determinación () y al error medio al cuadrado (RMSE). Finalmente en este estudio nos presentan las ANN como un buen candidato a fin de preveer la demanda, concretamente las redes neuronales “feed-forward back propagation learning” con una arquitectura 6-8-1 han resultado ser el mejor candidato debido a su alto rendimiento.

El tercer artículo tratado (He & Jiang, 2013) se centra en el uso de la red neuronal back-propagation para optimizar el control del inventario. En este artículo el dataset utilizado es relacionado con el mundo de la automoción por lo que presenta diferencias en logísticas respecto al dataset que se trabajará en nuestro proyecto, pero nos ofrece una visión de estudio interesante y candidata a revisar. El método de estudio demuestra ser capaz de ofrecer buenos resultados pero tambien nos presenta la problemática de saber determinar el numero optimo de capas a utilizar.

A partir de este punto, podríamos considerar tener una idea de los algoritmos o procesos más idóneos a fin de conseguir el objetivo del proyecto, pero llegado a este punto, es interesante conocer los algoritmos comentados anteriormente y que se difiere entre ellos, a fin de saber cual de ellos puede ser el más idóneo sin iniciar su estudio.

En lo referente al entorno académico, hemos encontrado múltiples procesos que serán candidatos de estudio en este proyecto, pero desconocemos cual de ellos son utilizados en entornos productivos, y cuales de ellos con preferidos por la comunidad.

Para conocer este apartado se ha revisado la comunidad de Kaggle, dónde se realizan concursos de análisis de datos, donde los concursantes han de ser capaces de conseguir el objetivo analítico propuesto por la propia plataforma, compitiendo a la vez con el resto de los concursantes en base a los resultados obtenidos por cada uno de ellos.

Dentro de esta comunidad, se ha filtrado por proyectos relacionados con predicción de la demanda y forecasts. Se ha encontrado que múltiples usuarios utilizan los algoritmos citados en artículos anteriores, pero destacar un proceso popular entre los usuarios, el XGBoost, este algoritmo es muy popular en la comunidad y es ampliamente utilizado en múltiples necesidades, des de procesos de predicción, clasificación, con un rendimiento en lo referente a procesamiento de máquina muy bueno. Este algoritmo también será tratado en el estudio como candidato.

El siguiente artículo (Mohit Gurnani, et al., 24-26 Feb. 2017) estudia todos los algoritmos hasta ahora comentados, incluyendo XGBoost, en que se comparan principalmente los algoritmos ARIMA, redes neuronales y XGBoost, junto con un conjunto de modelos híbridos que básicamente son resultado de mezclar los algoritmos de estudio entre ellos.

En este artículo nos ofrece un punto de vista muy interesante para nuestro proyecto ya que se comparan cada uno de los algoritmos basándose en RMSE, dónde podemos apreciar que para datos lineales ARIMA es un claro candidato por ofrecer mejores resultados, pero en el caso que los datos presenten cierta no linealidad, el resultado de la predicción varia bastante de la realidad.

Para el caso de algoritmos propios de ARNN podemos ver que tienen un muy buen resultado en caso a linealidad como de no linealidad, finalmente en el caso de XGBoost vemos que es capaz de ofrecer mejores resultados incluso que ARNN excepto en las primeras y últimas fases de los datos.

Las conclusiones del autor recalcan que obtenemos los mejores resultados en modelos híbridos, pero ARNN es capaz de alcanzar los resultados de los modelos híbridos.

El artículo Tianqi Chen, & Carlos Guestrin. (2016) no esta enfocada a la misma tematica que el proyecto que realizamos, pero el estudio que realizan los autores de los algoritmos XGBoost y LightGBM es muy interesante para tener en cuenta en nuestro estudio. En el artículo los autores muestran que parametros son los más importantes a optimizar junto con las pruebas que realizan. Este estudio de parametros nos sera muy útil en nuestro caso ya que cada uno de los algoritmos de estudio contienen multitud de parametros posibles a optimizar, permitiendonos gracias a este artículo conocer cuales de ellos realmente pueden ayudar a obtener mejores resultados.

### 2.3 Conclusiones

Para este proyecto podríamos considerar muchos algoritmos o modelos interesantes a tratar, ya que, en el caso de nuestro objetivo, predecir la demanda, podemos encontrar un gran número de candidatos. Sin embargo, gracias al estudio previo realizado en los artículos académicos podemos minimizar esta búsqueda o estudio de procesos.

Los autores de cada uno de los artículos ya han estudiado previamente cada uno de los algoritmos tratados comparando su rendimiento en múltiples datasets.

El estudio realizado nos ha permitido ganar gran cantidad de tiempo de estudio a fin de poder centrarnos en conseguir el objetivo del proyecto.

Del estudio proporcionado de todos los artículos tratados obtenemos un grupo reducido de algoritmos candidatos a tratar. Cada uno de estos algoritmos que serán tratados a lo largo del proyecto, estudiando su rendimiento y capacidad de predicción, además se tendrá en cuenta su capacidad de funcionar en entornos productivos, es decir, que el proceso sea capaz de proporcionar resultados en una cantidad de tiempo aceptable, a fin de evitar algoritmos que en lo referente a resultados quizás son óptimos, pero requieren de días para procesarse.

En este sentido, se estudiará encontrar el equilibrio entre resultados y rendimiento de máquina. En este caso se priorizará el resultado por encima de la velocidad de ofrecer resultados óptimos, pero en caso de obtener una velocidad muy baja se planteará la opción de escoger un modelo menos optimo pero que sea capaz de ofrecer resultados en un margen de tiempo aceptable.

Los algoritmos de estudio podemos encontrarlos dentro de las categorías, como arboles de decisión, redes neuronales y métodos de regresión, de los cuales específicamente se estudiarán XGBoost, debido a los buenos resultados ofrecidos principalmente en base a la velocidad de procesamiento, es muy popular dentro a la comunidad y además es capaz de ofrecer muy buenos resultados en lo referente a predicción.

El método de regresión ARIMA también será estudiado en el proyecto, ya que es el método más común para el tipo de proyecto de estudio y es capaz de adaptarse perfectamente el tipo de datos de nuestro proyecto siendo un claro candidato para proporcionar buenos resultados.

En lo referente a redes neuronales, se estudiarán dos de sus modelos, primeramente, la ANN con “feed-forward back propagation learning” en base a los buenos resultados obtenidos en el artículo (Sustrova, 2016) ya que este tipo de modelo presenta una gran adaptación al tipo de datos que se estudiaran en este proyecto.

El siguiente método dentro la categoría es ARNN, en base a los buenos resultados obtenidos en el artículo (Mohit Gurnani, et al., 24-26 Feb. 2017) ya que este modelo ha sido capaz de ofrecer buenos resultados respecto a modelos híbridos que eran mejores candidatos, además de proporcionar una gran adaptación al tipo de datos variables del autor.

Principalmente los algoritmos comentados serán los primeramente estudiados, pero no se descarta añadir más durante el proceso, por ejemplo, podría ser interesante añadir procesos propios de Deep Learning para predecir la demanda, pero este tipo de proceso se mantendrá en un segundo plano en caso de que los algoritmos estudiados nos sean capaces de ofrecernos el resultado esperado.

# 3. Diseño e implementación

## 3.1 Recogida inicial de datos y almacenamiento

Para la realización de este proyecto se dispone de básicamente 4 datasets. Todos los datasets contienen datos de una empresa real de productos de alimentación en que estos datos han sido anonimizados.

Los datasets disponibles son los siguientes:

* 01\_Ventas.xlsx
* 02\_Promociones.xlsx
* 03\_Stock.xlsx
* 04\_PrevisionesEmpresa.xlsx

**Dataset 01\_Ventas.xlsx**

En este dataset disponemos de datos referente a las ventas de la empresa dónde disponemos de las unidades de venta de cada artículo a nivel de día.

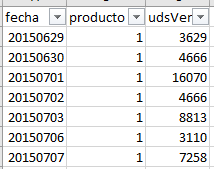


Figure 3 Dataset Ventas

**Dataset 02\_Promociones.xlsx**

El dataset referente a promociones disponemos para cada artículo su precio y el precio en cada que exista una promoción.

Además, disponemos de varios campos referentes a fecha que nos indican inicio y fin de las promociones junto con la semana en que se inició la promoción hasta su fin.

Finalmente disponemos de un campo que nos informa del tipo de promoción aplicada.

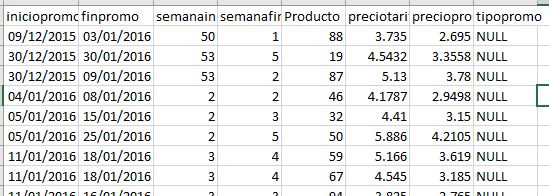


Figure 4 Dataset Promociones

**Dataset 03\_Stock.xlsx**

El dataset referente a stock, al igual que en el dataset de ventas, encontramos las unidades de stock por cada producto a nivel de día.

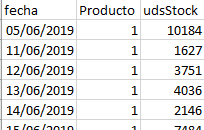


Figure 5 Dataset Stock

**Dataset 04\_PrevisionesEmpresa.xlsx**

El último dataset disponible podemos observar la previsión de unidades de venta que la empresa espera vender para cada artículo a nivel de día.

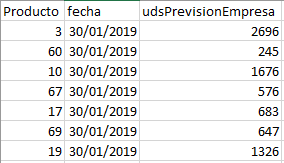


Figure 6 Dataset Previsiones empresa

## 3.2 Preparación, limpieza y preprocesado de los datos

En base a las indicaciones de la empresa hemos de considerar los siguientes factores ya que pueden impactar el análisis de datos que se realice.

Factores:

* Los días en que no hay venta para un producto, es que la venta es cero.
* En el fichero de promociones el precio de tarifa es el precio en un periodo no promocional.
* Los días en que no hay stock para un producto, probablemente se deba a que ese día el stock no se ha actualizado correctamente debido a un fallo en la sincronización de la información.
* Probablemente haya periodos de tiempo en que no estén disponibles las previsiones de la empresa.
* Las fechas correspondientes a festivos en el ámbito geográfico de la

Empresa son

* Año Nuevo Lunes 01/01/2018
* Reyes Sábado 06/01/2018
* Jueves Santo Jueves 29/03/2018 (varía de año en año)
* Viernes Santo Viernes 30/03/2018 (varía de año en año)
* San Jorge Lunes 23/04/2018
* Fiesta del trabajo Martes 01/05/2018
* Asunción de la Virgen Miércoles 15/08/2018
* Fiesta Nacional de España Viernes 12/10/2018
* Todos los Santos Jueves 01/11/2018
* Día de la Constitución Española Jueves 06/12/2018
* La Inmaculada Concepción Sábado 08/12/2018
* Navidad Martes 25/12/2018

**Observaciones sobre los datos de estudio**

En base a los factores a considerar y a disponer de los datasets, podemos empezar a estudiar la preparación del dataset final con el que iniciaríamos el análisis.

Inicialmente nos encontramos con 4 datasets, de los cuales, el dataset referente a las previsiones de la empresa solo lo utilizaremos al final únicamente para poder afirmar si nuestro resultado es mejor o no respecto al que prevé la empresa.

Debido a que cada uno de los dataset presenta una casuística diferente se comentaran las observaciones para cada uno de ellos.

**Ventas**

En el dataset de ventas observamos la siguiente descripción de los datos

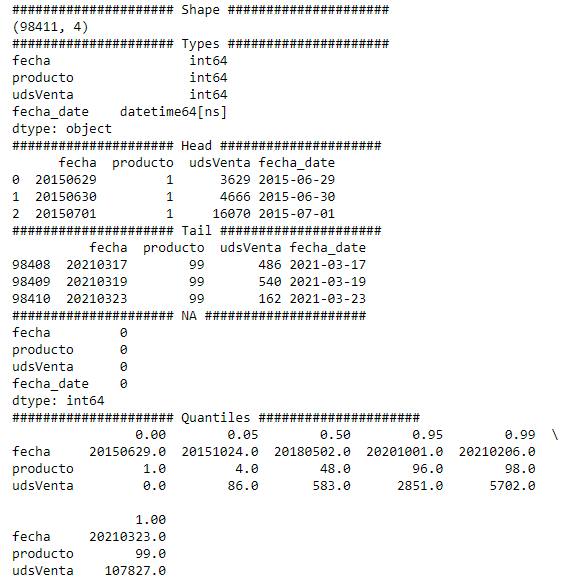


Figure 7 Descripción datos Ventas

Podemos apreciar que en los datos no encontramos nulos, de una dimensión de 98411 filas y 4 columnas.

Disponemos de datos des de 20150629 hasta 20210323



Figure 8 Rango de fechas de los datos

En el dataset hemos de tener en cuenta los factores comentados anteriormente, en que los días que no haya venta para un producto la venta es cero. Es decir, que hemos de modificar el dataset para que añada la venta de aquellos artículos que no aparezcan para cada día faltante.

Si del artículo 1 tenemos por ejemplo, venta el 2016-01-01, y no volvemos a observar venta hasta el 2016-01-04, hemos de añadir una línia por cada día faltante en el artículo, pero informando de una venta igual a 0.

El resultado del dataset ha de contener para cada día, des de 2015-06-29 hasta 2021-03-23, una linia referente a venta para cada artículo sin ningún día faltante.

**Stock**

En el caso de stock, hemos de considerar un nuevo factor, comentado anteriormente, en el caso que no aparezca stock para un producto en un día concreto, significa que ha habido un error y no se ha informado.

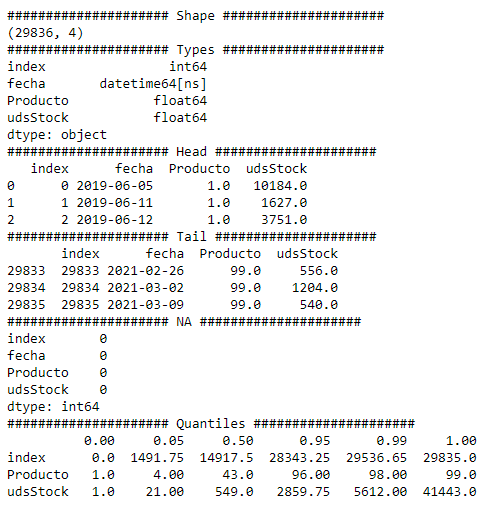


Figure 9 Descripción datos Stock

Observamos que en el dataset de stock no encontramos nulos tampoco, con una dimensión de 29836 filas y 4 columnas.

Si tenemos en cuenta la cantidad de filas de stock respecto a ventas, vemos que disponemos de muchas menos.

Esto es debido a que en el caso de stock disponemos de datos des de 2019-06-05 hasta 2021-03-24.

Nos encontramos con la problemática de no disponer de datos antes de 2019-06-05, y de haber días entre las fechas de datos, en que tampoco disponemos de datos para algunos artículos en días concretos.

En la primera casuística se decide no complementar los datos anteriores a la fecha comentada, ya que intentar completar estos datos tan antiguos con cualquier solución, por ejemplo la media de stock, podría presentar una desviación de la realidad demasiado importante.

En el caso de la segunda casuística, para los días que no tengamos información de un artículo para un día en concreto, se ha decidido completar la información, con la media de stock del artículo en cuestión.

**Promociones**

En el dataset de promociones observamos la siguiente descripción:

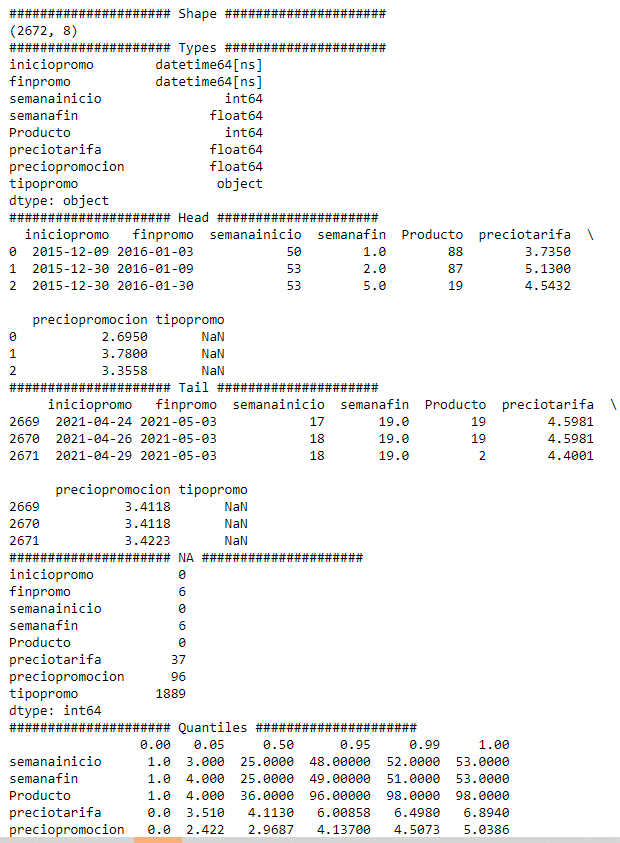


Figure 10 Descripción datos Promociones

Observamos que solo disponemos de 2672 filas y 8 columnas.

Además podemos apreciar que en este dataset encontramos nulos en la mayoría de campos.

En concreto en este dataset se ha decidido derivar del dataset original de promociones, en dos datasets, uno referente a las propias promociones y otro referente a precios.

En el dataset derivado de promociones dónde únicamente encontraríamos las promociones, podremos ver para cada día y por cada artículo si existe una promoción o no. Hay que tener en cuenta que originalmente teniamos el rango de fechas en que disponíamos de promoción para cada artículo, ahora a partir de la transformación desarrollada tendremos para todos los días si disponemos de promoción o no.

Referente a los nulls, se ha decidido eliminarlos debido al bajo volumen que suponen, en el caso de tipo promo, el volumen es muy grande, pero este campo no se utilizará para el estudio en ninguna fase.

El dataset derivado referente a precios, tenemos para cada día y para cada artículo su precio, en caso que haya una promoción el precio que aparece es el de la promoción, si no, aparace el precio normal. Es decir que si el artículo 1 esta en promoción des de 2016-01-01 hasta el 2016-01-03, durante este rango de fecha veremos el precio de promoción, y fuera de este rango el precio normal.

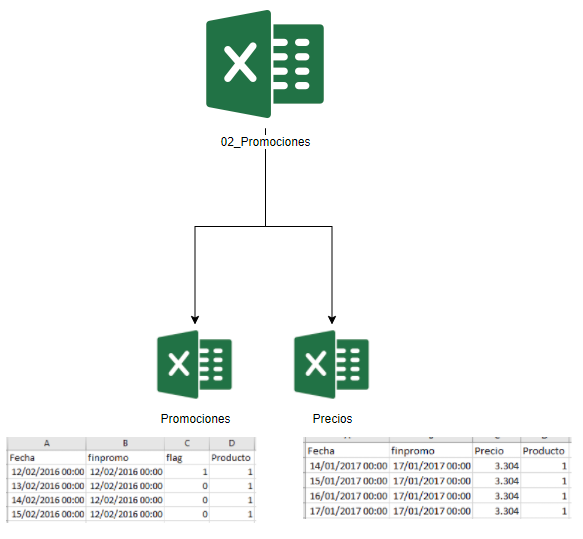


Figure 11 Estructura dataset Promociones

**Dataset final**

El dataset final es la combinación de los datasets tratados hasta ahora, ventas, stock, promociones y precios en un único dataset. De esta manera dispondremos para cada día y cada artículo todos los campos de interés para el análisis.

Finalmente, al dataset final también se ha añadido información referente a fechas, incluyendo los festivos marcados por la empresa.

A la hora de unir todos los datasets en uno nos encontramos con un problema referente a como conectar estos datasets entre sí.

Para poder conectar los dataset todos ellos disponen del campo fecha y producto. En el caso de unir dos dataset, uno de ellos no dispusiera de un producto o fecha que el otro dataset si contiene, este quedaría excluido del dataset final. Podríamos imaginar esta conexión como un INNER JOIN de SQL.

Pero además nos encontramos que el rango de fechas entre dataset es diferente:

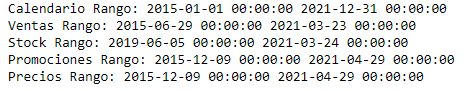


Figure 12 Rangos de fechas de los datasets

Si observamos en la imagen anterior la mayoría de dataset inician en 2015 y finalizan en 2021, excepto stock que inicia en 2019 y finaliza en 2021.

Debido a la casuística de stock se ha decidido dividir el dataset final en dos datasets finales idénticos, pero uno con stock y otro sin stock. El objetivo es poder observar si la variable stock es importante en el análisis, pero hemos de considerar que el volumen de datos sin stock es el mejor candidato en parte a fin de disponer de datos des de 2015 hasta 2021, en cambio con stock solo disponemos des de medianos de 2019 hasta 2021, apenas 1 año de datos.

Finalmente, una vez estudiado si es mejor disponer de los datos de stock o no, se creará un tercer dataset final idéntico al anterior sin stock, pero con el añadido de variables que se consideren que pueden mejorar el análisis.

Cada dataset creado tendrá su propia fase, en que siguiendo la metodología de CRISP-DM, cada fase se deberá poder presentar un entregable, entendiendo como entregable el análisis de los algoritmos ha estudiar junto con sus resultados.

## 3.3 Análisis de los datos

### 3.3.1 Fases de entrenamiento

**Fase 1 – Con Stock**

En esta fase la explicación es diferente del resto de fases. Inicialmente al disponer de stock, el histórico de datos se ve reducido considerablemente solo disponiendo de datos des de medianos de 2019 a medianos de 2021.

Además, la variable referente a las unidades de stock no la podemos utilizar sin realizar ninguna transformación, ya que es una variable que a futuro desconocemos, por este motivo a partir de la variable de stock se han generado 3 variables nuevas referentes a la media de stock de los últimos 90, 180 y 360 días, pero ha reducido el dataset aún más hasta tal punto de solo disponer datos de 2021.

El dataset se ha visto reducido debido a que al crear las nuevas tres variables estas, están realizando la media de fecha anteriores que no existen en las primeras filas del dataset, es decir, si el dato más reciente es del uno de febrero de 2021, al realizar el stock medio de los últimos 90 días para este registro no encuentra nada, generando nulos. Posteriormente de crear las variables eliminamos los nulos generando el dataset tan reducido.

Por este motivo comentado anteriormente se ha decidido eliminar la variable de stock a fin de disponer de un volumen de datos suficiente para realizar el análisis.

Muestra del dataset que obtendríamos final, mostrando el rango de fechas



Figure 13 Rango de fechas con Stock

**Fase 2 – Sin Stock**

En esta fase gracias a haber eliminado la variable referente a stock disponemos de un histórico de datos mucho más amplio des de 2015 hasta 2021.

En la siguiente imagen podemos ver una muestra de los datos del dataset



Figure 14 Muestra de datos Fase 2

Podemos apreciar que las variables que utilizaremos para el análisis son la Fecha, el producto, flag para saber si existe promoción o no, el precio, otro flag referente a los festivos y finalmente un último flag para saber si es fin de semana.

Una vez revisado el dataset realizamos la eliminación de outliers y escalamos los datos. Al finalizar la limpieza y estandarización de datos dividimos el dataset en un conjunto de entrenamiento y otro de test.

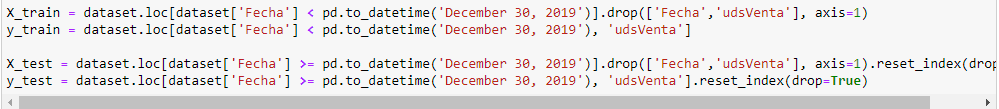


Figure 15 Muestra de código División datos test y training

**Fase 3 – Sin Stock más variables de interés**

En la última fase de estudio se han añadido un conjunto de variables en base a las unidades de venta a fin de que puedan ayudar a mejorar los resultados de la fase 2.

Debido a que la variable “udsVenta” no la podemos utilizar directamente para el estudio, ya que en parte es la variable que queremos predecir, pero sí que podemos utilizar sus datos pasados, es decir, podemos utilizar información resumida referente a las unidades de venta para poder predecir a futuro.

En esta fase exactamente las variables añadidas hacen referencia a las unidades de ventas en un período pasado. Encontramos dos tipos de variables añadidas, una en base a las ventas medias y las segundas, referentes al valor exacto en el período pasado.

Variables referentes a la media de ventas en días pasados:

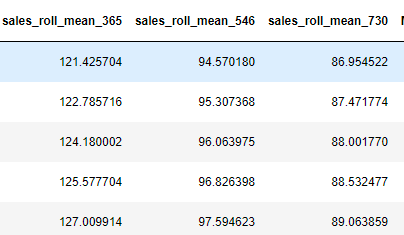


Figure 16 Variables creadas en Fase 3

En esta imagen vemos en la primera columna las ventas medias de los últimos 365 días.

En referencia a las variables en un período pasado se han definido los períodos 7,15,30,60,90,120,150,180,210,240, 270,300, 330,360 que referencian a días. Es decir, en el primer caso veremos las unidades de ventas exactamente de hace 7 días atrás.

### 3.3.2 Correlación entre variables

Teniendo en cuenta que en el proyecto se han utilizado dos datasets diferentes, uno sin stock y el siguiente con multitud de variables añadidas, hemos estudiado en cada uno de los datasets la matriz de correlación de cada uno de ellos.

🡪Matriz de correlación del dataset sin variables extra añadidas:

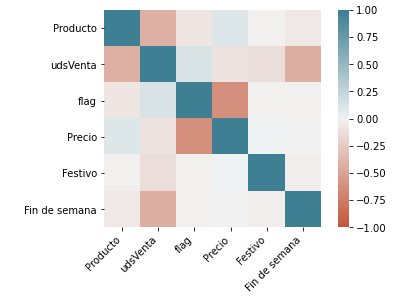


Figure 17 Matriz de correlación

Observamos que hay una correlación destacable entre flag y precio, el cual era de esperar ya que provienen del mismo dato.

Hay que destacar la correlación entre Fin de semana y udsVenta significando que la variable referente al fin de semana tiene un efecto importante en las ventas.

🡪Matriz de correlación del dataset con variables extra añadidas:

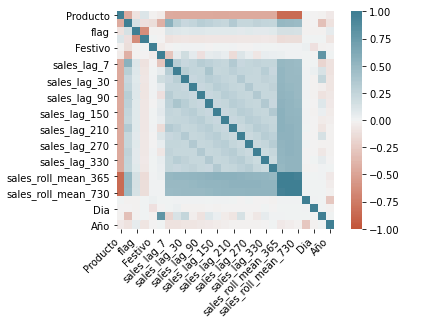


Figure 18 Matriz de correlación con variables extra

En la imagen anterior podemos apreciar que hay alta correlación entre algunas de las variables creadas por ejemplo en el caso de sales\_roll\_mean\_365 y sales\_roll\_mean\_730, pero esta correlación es causada en parte a que provienen de la misma variable origen.

### 3.3.3 Estudio de la estacionalidad de los datos

A fin de conocer mejor los datos de estudio se ha realizado un estudio de la estacionalidad de los datos a fin de ver si observamos una tendencia en base al tiempo, por ejemplo, en base a la semana o mes del año.

Además, necesitamos sabes si los datos son estacionarios para poder aplicar el modelo de ARIMA, ya que en función de si los datos presentan estacionalidad o no, deberemos aplicar un modelo u otro dentro de la familia de ARIMA.

Para estudiar la estacionalidad utilizaremos el método de Dickey-Fullers test

Dentro de este método podemos utilizar varios parámetros en función de método de regresión que decidamos, en nuestro caso se ha estudiado con los dos métodos disponibles



Figure 19 Test de Dickey-Fullers

Observamos que, en las tres pruebas diferentes realizadas, utilizando el método comentado, el p-value es inferior a 0.05 por los que podemos definir que los datos presentan estacionalidad.

El resultado además nos ayuda a decidir el método a seleccionar, dentro de los métodos disponibles dentro de ARIMA, el cual es SARIMA, en que es el idóneo en caso de tratar con datos que presentan estacionalidad.

En este apartado también se ha observado el comportamiento de las ventas en diferentes franjas de tiempo a fin de entender mejor los datos.

En las siguientes imágenes podemos observar las ventas en función del año, del mes y del día de la semana.

En base al año observamos que las unidades de venta claramente están bajando, pero hay que considerar que en el caso de 2021 no disponemos de los datos del año completos, pero podemos observar una clara tendencia de bajada.

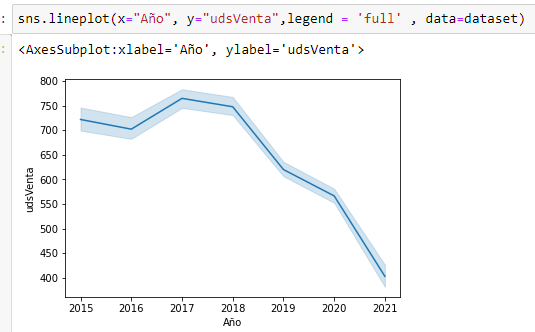


Figure 20 Estudio estacionalidad - Año

En base al mes:

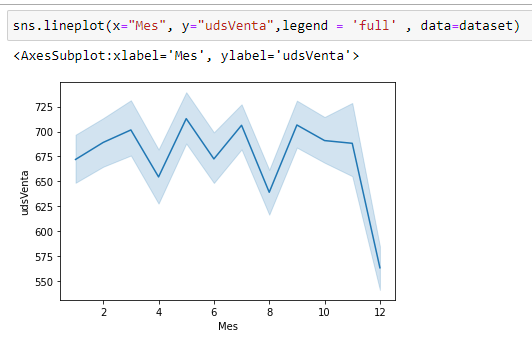


Figure 21 Estudio estacionalidad - Mes

Observamos que diciembre las ventas bajan de manera destacable.

Finalmente observamos a nivel de día de la semana

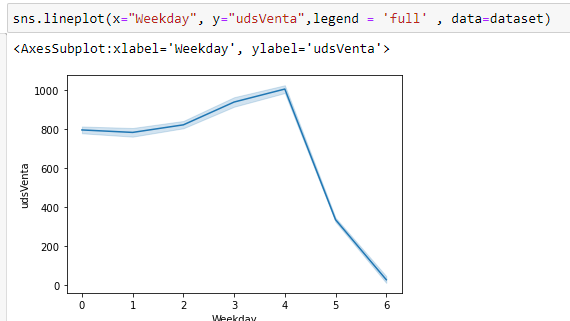


Figure 22 Estudio estacionalidad - Día

De la imagen podemos intuir que las ventas son más altas al final de la semana, en que el Domingo las ventas son nulas.

Si tenemos en cuenta la matriz de correlación observada anteriormente, dónde veíamos que la variable “fin de semana” y “udsVenta” estaban correlacionadas, en este grafico podemos entender que los fines de semana las ventas bajan respecto el resto de los días.

### 3.3.4 Conjunto de datos de entrenamiento y de test

En base a disponer del dataset final, necesitamos dividir este conjunto de datos en dos subconjuntos:

* Conjunto de entrenamiento: Subconjunto para poder entrenar el modelo
* Conjunto de prueba: Subconjunto para testear el modelo entrenado

Es habitual en los algoritmos de estudio que vamos a utilizar, dividir el dataset basándonos en el tan por cien del volumen de datos original a fin de distribuirlo entre el conjunto de entrenamiento y el de prueba. Exactamente el 80% se suele utilizar para entrenar y el 20% restante para testear. En nuestro case, el dataset original se basa en una línea temporal, es decir que des de una fecha inicial disponemos de los datos hasta una fecha fin.

Dividiremos el dataset original basándonos en la fecha, depositando en el conjunto de entrenamiento los datos anteriores a una fecha, y el resto serán utilizados para la prueba.

Inicialmente disponemos de datos des de 2015 hasta 2021, en que des de 2015 hasta 2019 se utilizaran para el entrenamiento y des de entonces hasta 2021 se utilizaran para test.

Además de este enfoque inicial se estudiará el efecto de utilizar diferentes franjas de tiempo para el entrenamiento y test, por ejemplo, para entrenamiento se observará el efecto de utilizar solo 2 años de datos, dónde al final se pueda observar que volumen de datos ofrece el mejor resultado.

En la siguiente figura podemos observar cómo se crean los subconjuntos de datos

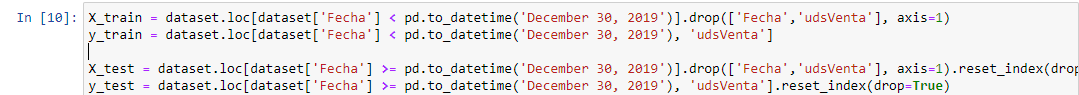


Figure 23 Fragmento código Test y Training

### 3.3.5 Tratamiento de Outliers

La mayoría de las modelos de estudio son sensibles a los outliers pudiendo generar resultados no esperados o con resultados muy alejados a los esperados.

Por este motivo se ha estudiado si el dataset contiene outliers a fin de tratarlos.

Primeramente, en base a los días de la semana observamos si hay outliers en el dataset

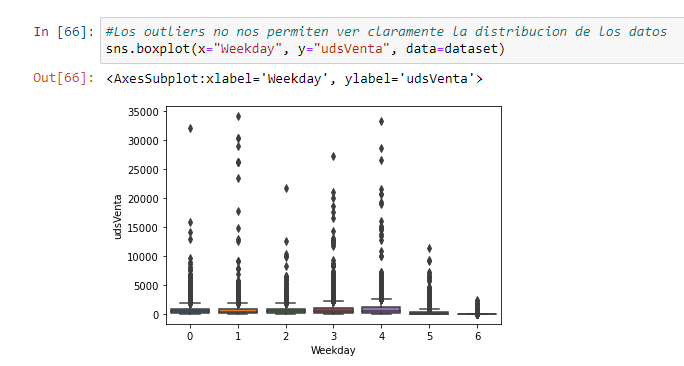


Figure 24 Outliers por día de la semana

En la imagen claramente observamos que el dataset contiene outliers.

Para tratar los outliers nos basaremos en el IQR score, basado en los rangos intercuartílicos para definir si el dato es un outlier o no.

Para tratar los outliers referentes a las unidades de venta utilizaríamos el siguiente fragmento de código:

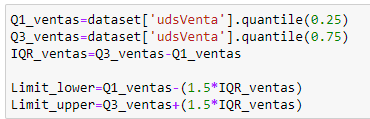


Figure 25 Fragmento de código tratamiento outliers

En la imagen podemos observar que definimos dos límites uno superior y otro inferior a fin de descartar si el dato es un outlier o no.

Finalmente, si el dato se encuentra por debajo del límite inferior o se sitúa por encima del límite superior eliminamos el registro.



Ahora al revisar visualmente el mismo dato dónde encontrábamos outliers observamos que después de eliminar los outliers podemos visualizar la distribución de las unidades de venta mucho mejor.

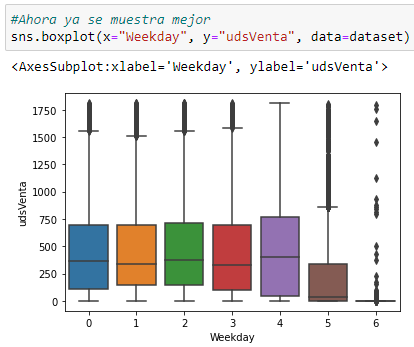


Figure 26 Limpiez outliers Venta

El proceso de eliminación de outliers se ha realizado para los campos referentes a unidades de venta, precio y stock.

Resultado después de limpieza de outliers en precio:

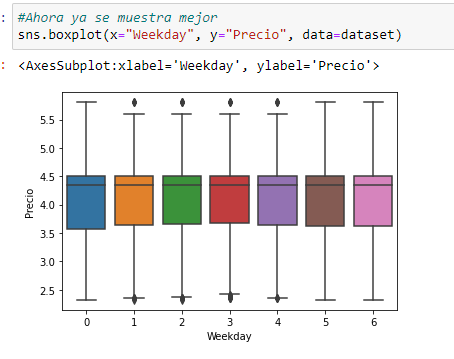


Figure 27 Limpieza outliers Precio

Resultado después de limpieza de outliers en stock:

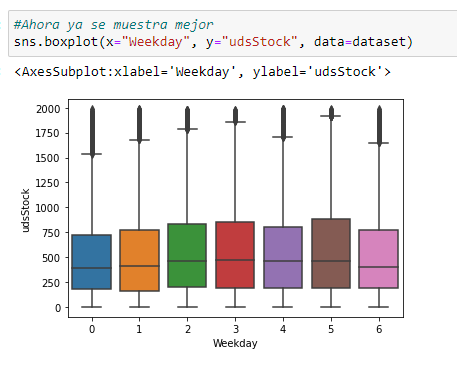


Figure 28 Limpiez outliers Stock

### 3.3.6 Método de evaluación de los modelos

Para evaluar y poder decidir que método es más apto, podemos encontrar multitud de métricas que nos ayudaran a conocer como de buenos son los modelos de estudio.

Un subconjunto de métricas que podemos utilizar para el estudio puede ser los siguientes:

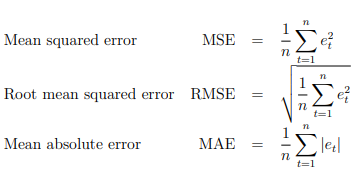


Figure 29 Fórmula RMSE

En este proyecto nos basaremos en el uso de RMSE ya que en los artículos académicos estudiados es la recomendada en la mayoría. El valor que nos muestre el RMSE, ha de ser lo más bajo posible, de esta manera el modelo que nos ofrezca el menor valor de RMSE será el mejor candidato.

Este valor nos permite calcular la diferencia entre los valores predichos y los reales.

## 3.4 Algoritmos de Machine Learning

### 3.4.1 XGBoost

El primer método que se estudiara en este proyecto es XGBoost el cual fue desarrollado por Tianqi Chen, & Carlos Guestrin. (2016), el cual es una técnica de aprendizajo automático dentro de la familia de metodos por gradiente utilizado para problemas de predicción. Además es un método que nos permite la computación distribuida, la paralelización y la optimización del caché.

Hay que considerar que este método pertenece a la familia de algoritmos de regresión basados en gradiente. En que podemos imaginar este método como un árbol de decisión que crece horizontalmente.

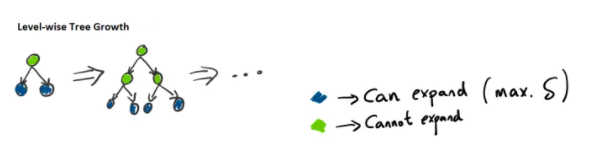


Figure 30 Funcionamiento XGBoost

Level-wise vs leaf-wise by [Felipe Sulser](https://felipesulser.github.io/ashrae-kaggle/)

Primeramente, a fin de utilizar el método de XGBoost de la forma más optima posible necesitamos estudiar cuales son los mejores parámetros que nos puedan ofrecer el mejor resultado posible. Para poder saber cual es el mejor resultado posible, tal y como se ha comentado anteriormente nos basaremos en el RMSE.

**Fase 2:**

En la fase en que no disponemos de stock y sin haber añadido las variables de interés, los mejores parámetros obtenidos han sido:



Figure 30 Mejores parámetros fase 2 XGBoost

El fragmento de código creado para obtener los resultados ha sido el siguiente:

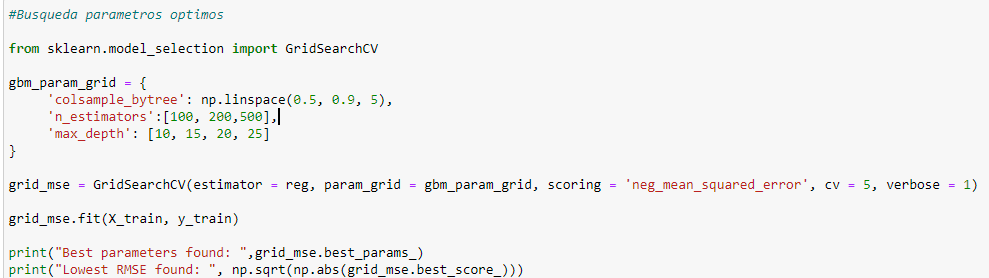


Figure 31 Fragmento de código estudio parámetros XGBoost Fase 2

Definimos el modelo con los parámetros óptimos:

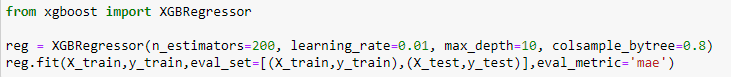


Figure 32 Fragmento de código - definción XGBoost

🡪 Evaluamos su RMSE

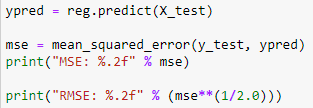


Figure 33 Evaluación RMSE XGBoost Fase 2



🡪 Estudiamos la importancia de variables:

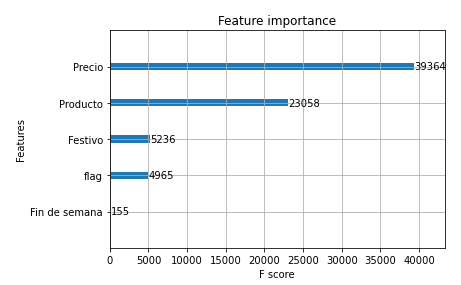


Figure 34 Importancia de variables XGBoost Fase 2

Observamos que en el caso de XGBoost la variable con más importancia es el precio seguido del propio producto, en cambio la variable fin de semana vemos que realmente no aporta valor al resultado.

**Fase 3:**

En esta última fase no disponemos de stock pero se han añadido un conjunto de variables de interés a fin de mejorar los resultados generados para cada uno de los algoritmos estudiados. Los mejores parámetros obtenidos han sido:

1. 

Figure 35 Mejores parámetros fase 3 XGBoost

A fin de saber cuáles han sido los mejores parámetros posibles junto con la definición del modelo se ha utilizado exactamente el mismo fragmento de código de la fase anterior.

🡪 Evaluamos su RMSE

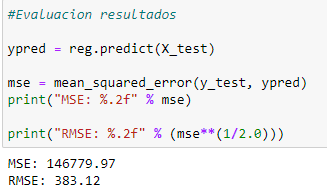


Figure 36 Evaluación RMSE XGBoost Fase 3

Observamos que en esta fase se ha conseguido mejorar los resultados gracias a haber añadido variables nuevas de interés.

🡪 Estudiamos la importancia de variables:

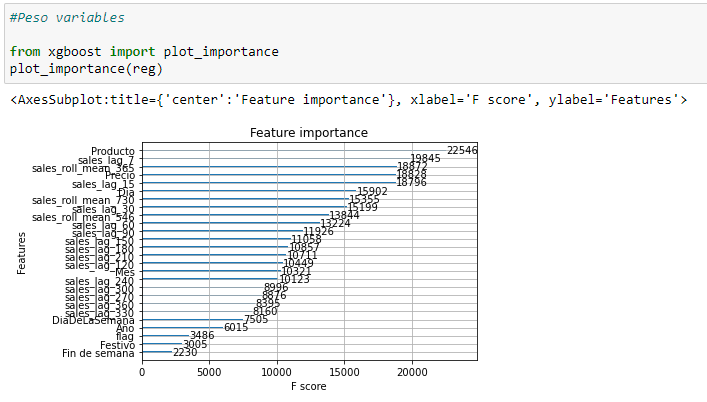


Figure 37 Importancia de variables XGBoost Fase 3

En esta fase al haber añadido múltiples variables podemos volver a observar su importancia. En este caso producto es la variable más importante en que anteriormente con menos variables era el precio. Además, vemos que las variables añadidas han contribuido positivamente en este modelo ya que la mayoría si sitúan en las más importantes.

### 3.4.2 LightGBM

El método LightGBM pertenece a la misma familia que XGBoost ya que ambos son métodos de regresión basados en gradiente, pero podemos encontrar diferencias entre ellos debido a que, en este algoritmo, el árbol de decisión en lugar de crecer horizontalmente como en el caso de XGBoost crece verticalmente.

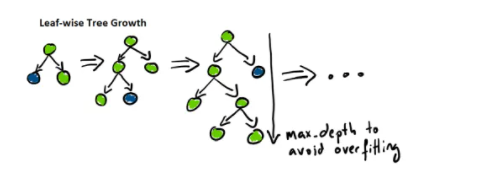


Figure 38 Funcionamiento LightGBM

Level-wise vs leaf-wise by [Felipe Sulser](https://felipesulser.github.io/ashrae-kaggle/)

Encontramos diferencias entre estos dos modelos comentados sobretodo en lo referente a la velocidad de procesamiento, en que debido a que LightGBM es un método enfocado para datasets grandes, su velocidad es superior a XGBoost, pero en cambio XGBoost nos ofrece resultados más robustos.

**Fase 2:**

En la fase en que no disponemos de stock y sin haber añadido las variables de interés los mejores parámetros obtenidos han sido:

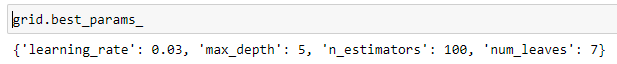


Figure 39 Mejores parámetros fase 2 LightGBM

El fragmento de código creado para obtener los resultados ha sido el siguiente:

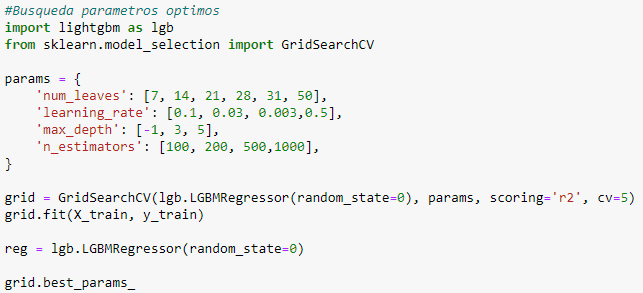


Figure 40 Fragmento de código estudio parámetros LightGBM Fase 2

Definimos el modelo con los parámetros óptimos:



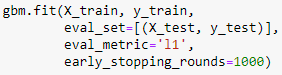


Figure 41 Fragmento de código - definción LightGBM

🡪 Evaluamos su RMSE



🡪 Estudiamos la importancia de variables:

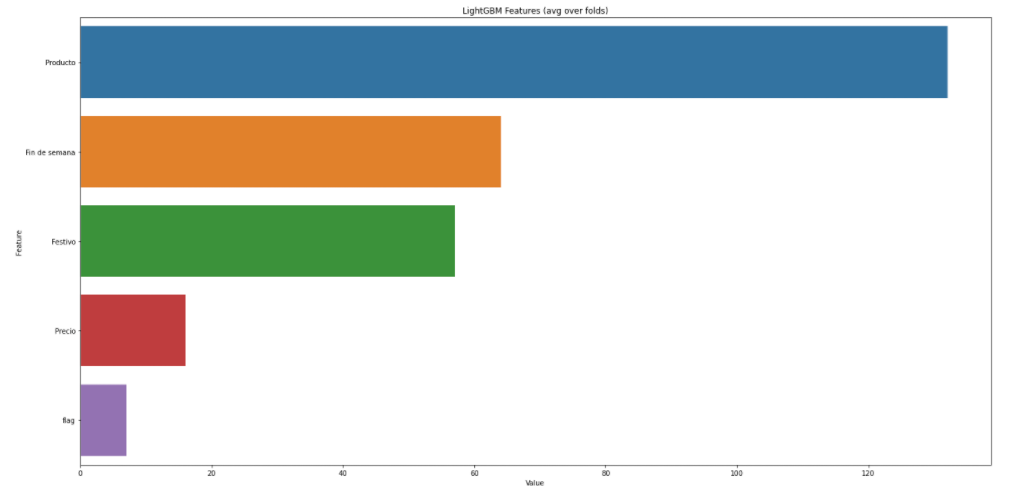


Figure 42 Importancia de variables LightGBM Fase 2

En el modelo de LightGBM vemos que la importancia de las variables es diferente de XGBoost en que la variable más importante es el “producto” seguido de “fin de semana” y “festivo”, dónde “fin de semana” no tenía importancia en XGBoost.

**Fase 3:**

En esta última fase no disponemos de stock pero se han añadido un conjunto de variables de interés a fin de mejorar los resultados generados para cada uno de los algoritmos estudiados. Los mejores parámetros obtenidos han sido:

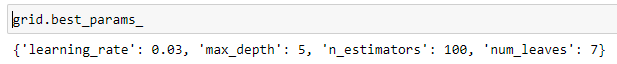


Figure 43 Mejores parámetros fase 3 LightGBM

🡪 Evaluamos su RMSE



🡪 Estudiamos la importancia de variables:

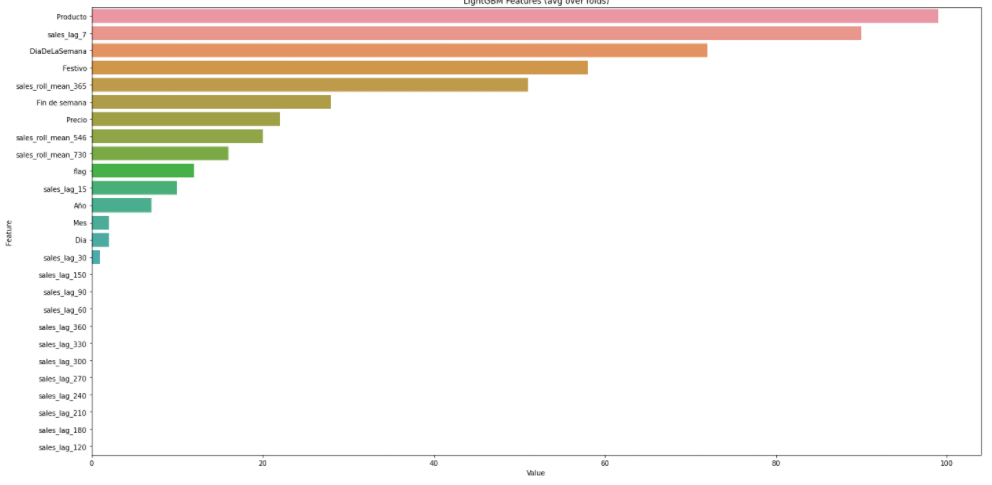


Figure 44 Importancia de variables LightGBM Fase 3

En este modelo vemos que la variable producto sigue siendo la variable más importante. Hay que destacar que en este caso las variables añadidas no aportan mucho valor al estudio ya que la mayoría se encuentran con un valor mínimo. Podemos concluir que este modelo no se ha visto beneficiado de añadir variables nuevas.

### 3.4.3 SARIMA

El método SARIMA es un método basado en series temporales en que poseen un componente estacional que se repite en las observaciones cada cierto tiempo.

Dentro de la familia de métodos que podemos encontrar dentro de ARIMA, podemos encontrar el propio método ARIMA y SARIMA, en que se diferencian por tratar la estacionalidad de las observaciones en el caso de SARIMA. Es decir, si al estudiar los datos observamos que presentan estacionalidad, debemos utilizar el método de SARIMA en lugar de ARIMA.

Hay que considerar que este método básicamente se basa en la variable a predecir y la fecha de la observación, por este motivo no se verá afectado su resultado si se añaden nuevas variables de interés.

1. Los mejores parámetros encontrados han sido:
2. SARIMA (1, 0, 1)x(1, 1, 1, 12)

Aplicamos los parámetros encontrados al modelo final.

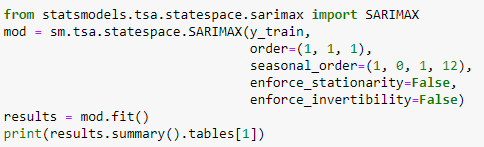


Figure 45 Fragmento de código - definición SARIMA

🡪 Evaluamos su RMSE



En este modelo no podemos estudiar ni la matriz de correlación ni la importancia de variables ya que este modelo solo utiliza la variable predictora y el campo fecha, además que no utiliza ninguna otra variable tampoco se añadirán más variables a fin de mejorar el resultado.

### 3.4.4 LTSM

El método LTSM es una extensión de las RNN (Redes neuronales recurrentes) que básicamente aplican su memoria para aprender de experiencias importantes que han ocurrido hace tiempo. Permiten recordar las entradas durante un período de tiempo largo.

Este tipo de red neuronal es muy útil en aquellos datos que pueden presentar ciclos recurrentes como estacionalidad ya que estos ciclos retroalimentan las memorias internas del método. Proporciona a la red información relevante del pasado a tiempo más reciente.

**Fase 2:**

En la fase en que no disponemos de stock y sin haber añadido las variables de interés los mejores parámetros obtenidos han sido:

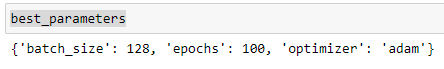


Figure 46 Mejores parámetros fase 2 LTSM

El fragmento de código creado para obtener los resultados ha sido el siguiente:



Figure 47 Fragmento de código - definción LTSM

Definimos el modelo con los parámetros óptimos

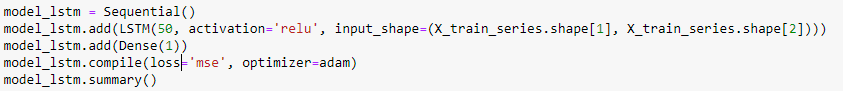
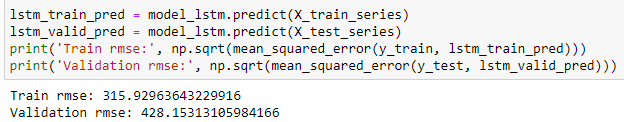




Figure 48 Fragmento de código estudio parámetros LTSM Fase 3

🡪 Evaluamos su RMSE



Debido a que el tipo de red neuronal elegida no contiene ningún parámetro para mostrar la importancia de las variables, no ha sido posible mostrarlo en este caso.

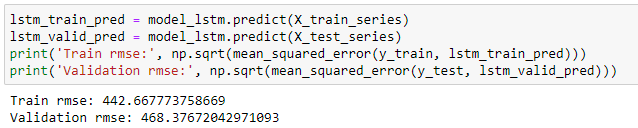
**Fase 3:**

En esta última fase no disponemos de stock pero se han añadido un conjunto de variables de interés a fin de mejorar los resultados generados para cada uno de los algoritmos estudiados. Los mejores parámetros obtenidos han sido:



Figure 49 Mejores parámetros fase 3 LTSM

🡪 Evaluamos su RMSE



Observamos en esta fase que haber añadido nuevas variables de interés no ha mejorado el resultado, sino que lo ha empeorado. Podemos concluir que este modelo nos ofrece mejores resultados en la fase 2.

### 3.4.5 Análisis comparativo de los resultados

En base a las fases estudiadas, y todos los componentes que han intervenido resumiríamos el estudio en la siguiente table basándonos en el RMSE:

|  |  |  |  |
| --- | --- | --- | --- |
|  | FASE 2 | FASE 3 | Mejora |
| **XGBoost** | 407 | 383 | +24 |
| **LightGBM** | 425 | 442 | -17 |
| **SARIMA** | 551 | 551 | 0 |
| **Red neuronal LTSM** | 428 | 468 | -40 |

Figure 50 Comparativa algoritmos

En la tabla de resultados finales podemos apreciar que el único modelo que se ha visto beneficiado de las nuevas variables ha sido XGBoost, dónde el resto de los modelos han empeorado sus resultados considerablemente.

Finalmente, el modelo que ha ofrecido el mejor resultado ha sido XGBoost con un RMSE de 383.

En la siguiente imagen podemos observar su resultado comparando el valor predicho y el real.

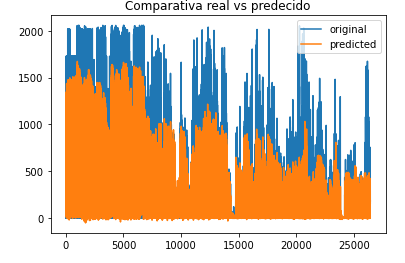


Figure 51 Comparativa resultados - Original vs Predicho

**Comparativa resultados obtenidos vs resultados de la empresa**

La propia empresa realizó su propia predicción de la demanda, siendo el resultado en el fichero 04\_PrevisionesEmpresa.xlsx.

Este fichero no contiene las unidades de venta reales, simplemente contiene fecha, producto y la previsión. Por este motivo se ha añadido a este dataset las unidades de venta a fin de poder saber como de acertada es su predicción.

Hay que considerar que este dataset hay días en que no hay datos para algunos productos, en este caso no se ha realizado ningún tratamiento, evitando alterar cualquier resultado. Se ha añadido para aquellos días existentes en el fichero las unidades de venta reales a partir del dataset generado en la fase 3, para aquellos días que coincidan con las previsiones de la empresa.

El RMSE de la previsión de la empresa es:



Si lo comparamos visualmente:

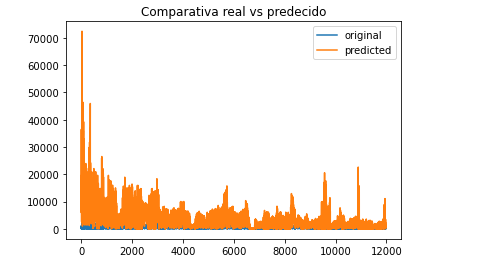


Figure 52 Comparativa resultados - Original vs Predicho empresa

Si nos fijamos el RMSE obtenido es mucho mayor a los obtenidos en el estudio.

Visualmente observamos que la predicción de la empresa esta claramente alejada de la realidad, llegando a encontrar para un producto en concreto en un día puntual, ventas reales iguales a 0, mientras la predicción de la empresa sobrepasa las 30mil unidades.

Ejemplo:

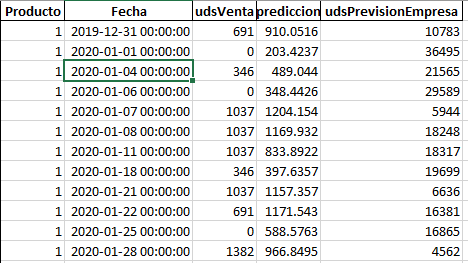


Figure 53 Ejemplo comparativa Venta Real vs Previsión Empresa vs Previsión

# 4. Conclusiones

## 4.1 Logro objetivo

Uno de los objetivos principales de este proyecto era poder cuantificar el beneficio que podía aportar este proyecto a la empresa en cuestión, si esta siguiera los pasos estudiados. En este apartado se mostrará el coste del stock que actualmente presenta la empresa y el del resultado del estudio realizado, con el objetivo de mostrar el beneficio y ahorro que puede aportar realizar este proyecto a la empresa.

En base a la previsión obtenida de la propia empresa y a los resultados obtenidos de nuestro estudio se ha podido comparar los dos métodos, valorar monetariamente el coste de cada uno a fin de considerar cuál de ellos sería idóneo para la empresa.

Primeramente, se ha calculado el stock de seguridad para los dos casos a fin de saber la cantidad de stock extra necesario para mitigar las variaciones de la demanda. Seguidamente se le ha añadido la cantidad de stock prevista para el caso de la empresa, y para el caso de nuestra previsión se ha añadido la cantidad prevista de venta. El resultado de este proceso nos muestra para cada uno de los dos casos la cantidad de stock disponible.

En base a conocer el precio de disponer de cada unidad de stock, podemos calcular el coste total de disponer de todos los artículos en el tiempo.

En el caso de la empresa se ha observado que la cantidad de stock esta sobredimensionada, ya que se supera en gran magnitud a las unidades vendidas. El resultado de estar sobredimensionado genera que no se produzca en ningún momento una rotura de stock, en que en ningún momento se produce una situación en que no se pueda vender un artículo por falta de stock.

En la siguiente imagen podemos observar las unidades vendidas y las unidades de stock de la empresa reales.

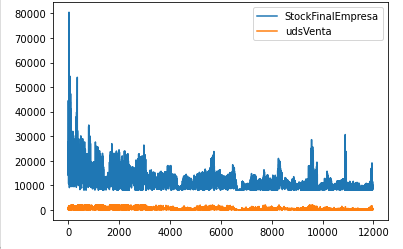


Figure 54 Dimensión unidades empresa

Para la previsión realizada en el estudio en cambio, vemos que la cantidad de stock se ajusta bastante bien a la venta real. En contrapartida en nuestro caso se producen dos escenarios a considerar. Hay días en que disponemos de más stock del necesario, y hay días en que no disponemos del stock necesario para vender generando una rotura de stock.

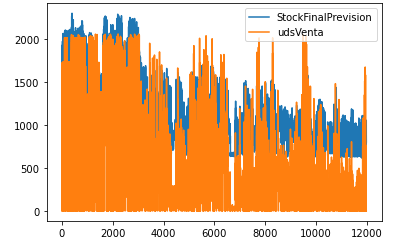


Figure 55 Dimensión unidades predicción

En el caso de la previsión estudiada el coste de las ventas perdidas por falta de stock es de 89.000€, en el caso de la empresa es de 0 debido a estar sobredimensionada la previsión de la empresa.



Teniendo en cuenta que en el caso de nuestra previsión tenemos perdidas por falta de stock, junto con la imagen que generaría de la empresa, hay que considerar el coste del stock de la empresa como de nuestro estudio a fin de poder decidir cual nos resultado nos puede aportar mayor ahorro.

En el caso de la empresa debido a estar sobredimensionada su previsión, el coste total de todo su stock es diez veces mayor que el del estudio realizado.



En base al resultado obtenido podemos afirmar que la empresa puede ahorrarse hasta 9’69 veces el coste total de stock en todo el histórico gracias al uso del modelo estudiado en este proyecto.

## 4.2 Lecciones aprendidas

Gracias a haber desarrollado el proyecto de previsión de la demanda, se ha conocido los algoritmos más destacables para la necesidad planteada junto con su implementación.

Se ha definido una estructura clara de los pasos a realizar a fin de iniciar un proyecto de analítica como el estudiado, permitiendo en un futuro aplicar los mismos pasos en nuevos proyectos personales y laborales.

En lo referente a proyectos de analítica con una necesidad diferente, se ha aprendido la importancia del estudio de proyectos similares realizados por académicos junto con el estudio de proyectos presentados en Kaggle. En base a este estudio el desarrollador ahorra muchas horas de estudio personal y además con base académica.

## 4.3 Líneas de trabajo futuro

En este proyecto encontraríamos varios puntos de mejora, especialmente en lo referente a algoritmos, debido a que en la fase de estudio el volumen de candidatos a estudiar era elevada, se decidió por utilizar aquellos más idóneos, pero dejando fuera algunos que podrían haber sido buenos candidatos, pero en menor medida. Considerar más algoritmos, aunque pudieran no ser los mejores hubiera aportado al estudio mayor diversidad de algoritmos junto con más opiniones.

Añadiríamos también el estudio de los parámetros de los algoritmos estudiados, en que se han estudiado los parámetros principales de cada uno de ellos, pero en realidad cada uno de ellos contiene una gran variedad, en que estos podrían haber mejorado los resultados de algunos de ellos.

Añadir más variables al estudio, puede aportar en un futuro mejoras en los resultados, mejorando la precisión de algunos algoritmos, aunque en algunos puede llegar a empeorar.

# 5.Bibliografía

[1] Abolghasemi, Mahdi, Gerlach, Richard, Tarr, Garth, & Beh, Eric. (2009). *Demand forecasting in supply chain: The impact of demand volatility in the presence of promotion.*

He, W., & Jiang, Z. (2013). An Inventory Controlled Supply Chain Model Based on Improved BP Neural Network. *Discrete Dynamics in Nature and Society*. doi:10.1155/2013/537675

Marsh, J. A., Pane, J. F., & Hamilton, L. S. (2006). *Making Sense of Data-Driven Decision Making in Education*. Retrieved 2 27, 2021, from https://rand.org/content/rand/pubs/occasional\_papers/op170

Mohit Gurnani, Y. Korke, P. Shah, S. Udmale, V. Sambhe, & S. Bhirud. (24-26 Feb. 2017). Forecasting of sales by using fusion of machine learning techniques. *2017 International Conference on Data Management, Analytics and Innovation (ICDMAI)* (pp. 93-101). Pune, India: IEEE.

Sustrova, T. (2016). A Suitable Artificial Intelligence Model for Inventory Level Optimization. *Trends Economics and Management*, 48. doi:10.13164/trends.2016.25.48

Tianqi Chen, & Carlos Guestrin. (2016). *XGBoost: A Scalable Tree Boosting System.* Washington: University of Washington. Retrieved from https://arxiv.org/pdf/1603.02754.pdf

(Liang, Weizhang, Luo, Suizhi, Zhao, Guoyan, & Wu, Hao, 2020)

# Anexos

Listado de apartados que son demasiado extensos para incluir dentro de la memoria y tienen un carácter autocontienido (por ejemplo, manuales de usuario, manuales de instalación, etc.)

Dependiente del tipo de trabajo, es posible que no haya que añadir ningún anexo.