Archivo Fotográfico De La Universidad De Antioquia: Valoración Histórica De Las Fotografías, 1997 - 2003

Federico Gómez Betancur

Monografía presentada para optar al título de Especialista en Analítica y Ciencia de Datos

Facultad de Ingeniería, Universidad de Antioquia
Especialización en Analítica y Ciencia de Datos
Jose David Ruiz Alvarez
Medellín, Antioquia, Colombia
2021

Cita	(Muñoz Zapata & Martínez Naranjo, 2018)
Referencia **Estilo APA 7 (2020)**	Muñoz Zapata, L., & Martínez Naranjo, J. A. (2018). *Archivo fotográfico de la Universidad de Antioquia: valoración histórica de las fotografías, 1997 - 2003* [Trabajo de grado especialización]. Universidad de Antioquia, Seleccione ciudad UdeA (A-Z).

Especialización en Analítica y Ciencia de Datos, Cohorte II.

UNIVERSIDAD DE ANTIOQUIA
Vicerrectoría de Docencia

Sistema de Bibliotecas

Centro de Documentación Ingeniería (CENDOI)

Repositorio Institucional: http://bibliotecadigital.udea.edu.co

Universidad de Antioquia - www.udea.edu.co

Rector: John Jairo Arboleda Céspedes
Decano/Director: Jesús Francisco Vargas Bonilla
Jefe departamento: Diego Jose Luis Botia Valderrama

TABLA DE CONTENIDO

Link del repositorio donde se realizó el procesamiento de modelos y datos.

LISTA DE TABLAS

LISTA DE FIGURAS

Resumen Ejecutivo

Tienda Registrada es una empresa que captura información de ventas del canal de las tiendas de barrio por medio de un sistema POS; Gutiérrez y Hernández (2007) sostienen que la estructura de los datos y la construcción de unos maestros definidos a través del tiempo permiten realizar agrupaciones por múltiples variables que dan valor a los reportes.

El consumidor final de tienda registrada son los productores de las grandes marcas en el país, actualmente la empresa comercializa información para marcas como Postobón, Alpina, Familia, Colanta, Super, Ramo, Roa, Entre Otras.

La información se comercializa de manera personalizada en forma de reporte, es decir, con el cliente se llega a un entendimiento de la necesidad y ahí se propone el mejor esquema de datos con el cual se responderá su pregunta de negocio.

Actualmente la empresa cuenta con un panel de 500 tiendas distribuidas en los departamentos de Cundinamarca, Antioquia, Valle Del Cauca y Atlántico, y se contó con la información de dichas ventas desde 2017, está información a pesar de que se trae limpia, para el análisis que se planteó fue necesario realizar algunas transformaciones.

El objetivo del modelo es predecir las ventas de la semana siguiente a partir de la información del último año en todas las tiendas para una marca especifica.

Una vez se realizó la transformación de los datos necesaria se definió que el problema era de regresión, en donde buscamos predecir un valor numérico correspondiente a la suma de ventas de los 7 días posteriores dada la información agrupada del pasado año.

Durante el proceso de entrenamiento se probaron diferentes modelos de machine learning tales como regresión Lineal, regresión Lasso, random Forest, Gradient Boosting y RNN, se calculó el

error con la técnica de error absoluto medio y el modelo que menor error presento fue Gradient Boosting.

Palabras Clave: POS, modelo, machine learning, ventas, predicción.

Abstract:

Registered Store is a company that captures sales information from the neighborhood stores channel through a POS system; Gutiérrez and Hernández (2007) argue that the structure of the data and the construction of defined masters over time allow groupings by multiple variables that give value to the reports.

The final consumer of the registered store are the producers of the big brands in the country, currently the company markets information for brands such as Postobón, Alpina, Familia, Colanta, Super, Ramo, Roa, Among Others.

The information is marketed in a personalized way in the form of a report, that is, with the client an understanding of the need is reached and there the best data scheme is proposed with which their business question will be answered.

Currently the company has a panel of 500 stores distributed in the departments of Cundinamarca, Antioquia, Valle Del Cauca and Atlántico, and the information on these sales was available since 2017, this information despite being clean, for analysis that was raised it was necessary to carry out some transformations.

The objective of the model is to predict the sales of the following week from the information of the last year in all the stores for a specific brand.

Once the necessary data transformation was carried out, it was defined that the problem was regression, where we seek to predict a numerical value corresponding to the sum of sales of the 7 subsequent days given the grouped information from last year.

During the training process, different machine learning models were tested, such as Linear regression, Lasso regression, Random Forest, Gradient Boosting and RNN, the error was

calculated with the mean absolute error technique and the model with the lowest error was Gradient Boosting.

Keywords: POS, machine learning, pattern, sale, prediction.

1 Descripción Del Problema

La marca objetivo actualmente realiza la planeación de su suministro en base a los datos del pasado, esta planeación la realiza una persona que siempre se ha encargado de ello por lo cual puede estar sesgada a proporcionar un número en base a su experiencia y no en los datos, así mismo la marca únicamente analiza información de ventas y no tienen en cuenta otras variables del producto y de la tienda.

1.1. Problema De Negocio

Tienda registrada identificó la necesidad de la marca para poder realizar una oferta comercial en el momento que se tengan resultados tangibles, la idea es poder automatizar y comercializar con cierta periodicidad el reporte.

2 Datos

2.1. Origen De Los Datos

Los datos utilizados para crear la predicción de ventas fueron extraídos directamente de la base de datos de tienda registrada con previa firma de acuerdo de confidencialidad, los datos representan las ventas de 971 tiendas de 4 departamentos de Colombia entre el 2017 y el 2021 de la marca objetivo.

Actualmente la empresa cuenta con 500 tiendas activas, pero en el trascurso del tiempo algunas tiendas han entrado y salido del panel, es por esto que entre 2017 y 2021, 971 tiendas han reportado información de ventas.

2.2. Datos Originales

Cada registro corresponde a la venta de un producto en una tienda, los productos están divididos en categoría y las tiendas tienen algunas propiedades.

Tabla 1

Diccionario de datos del dataset

Campo	Descripción
IDHora	Hora de la transacción en formato numérico.
Fecha	Día de la transacción.
Unidades	Cantidad de Unidades del producto.
Precio	Precio total de la venta
idTienda	Id de la tienda donde se registró la venta
CodigoDeBarras	Código de barras del producto vendido
DescripcionProducto	Descripción del producto vendido
categoriaProducto	Categoría del producto vendido
VolumenTotal	Volumen total del producto vendido
longitud	Coordenadas longitud de la ubicación de la tienda
latitud	Coordenadas latitud de la ubicación de la tienda
Departamento	El departamento de la tienda donde se realizó la venta
Tamaño	Tamaño en metros cuadrados de la tienda.
idVenta	Identificador único de una venta, ayuda a agrupar los productos llevados en una venta

Nota: Elaboración propia (2021).

Dado que tenemos solo un set de datos tendremos que usar estos para entrenar y probar el modelo.

2.3. Transformación Datos.

Inicialmente el dataset contiene 396.303 filas y 14 columnas.

El primer cambio a los datos se hace en la columna fecha, la lectura del archivo no identifico el formato por lo cual se lo asignamos manualmente.

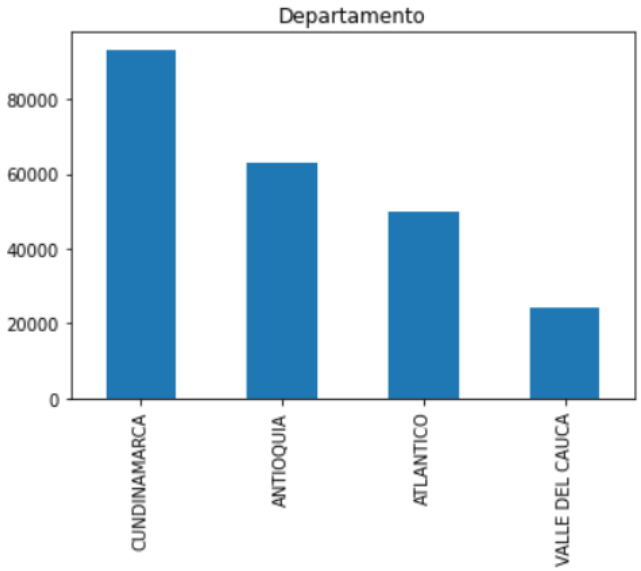
Para poder mostrar la información de la fecha separada por mes del año o semana del año de manera ordenada se realizó un método que concatena bien sea año – mes o año – semana, teniendo en cuenta que, si la semana o el mes contenían un carácter, este debía tener ser antecedido por un 0.

El dataset contiene información de 971 tiendas, muchas de ellas no contienen información de todos los años lo cual genera un ruido en los datos haciendo que para determinado año existan más tiendas activas que en otros, para esto únicamente se tomaron las ventas de las tiendas que tuvieran participación todos los años. Durante este proceso se descartaron 734 tiendas las representan el 41,3% de los datos totales, el dataset quedo con 237 tiendas que representan el 58.6% de los datos.

2.4.Análisis Descriptivo.

Figura 1.

Numero de ventas por departamento.

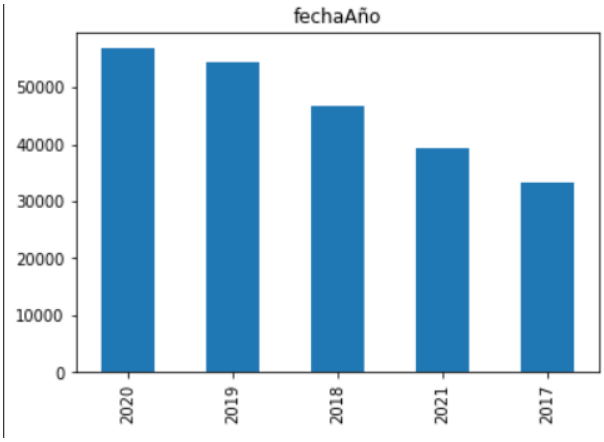


Nota. La figura muestra la cantidad de ventas por cada departamento. Elaboración propia (2021)

Se puede ver que el departamento que más datos de ventas tiene en el dataset en Cundinamarca, que es proporcional a la cantidad de tiendas que hacen parte del panel de tienda registrada para dicha zona.

Figura 2.

Numero de ventas por año

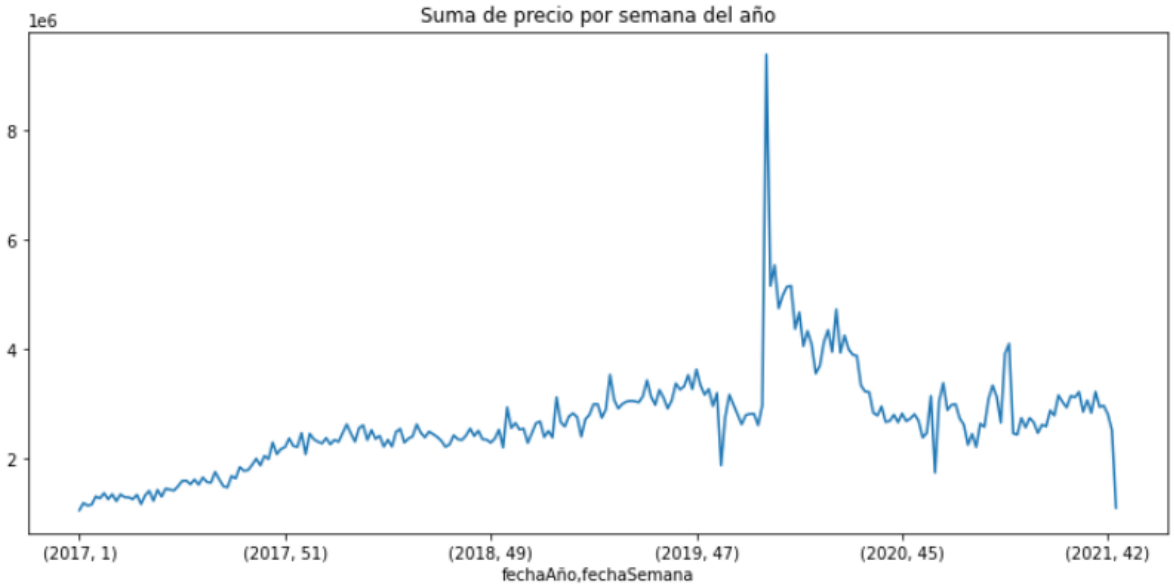


Nota. La figura muestra la cantidad de ventas por cada año. Elaboración propia (2021)

Se puede observar que el año 2020 es el año que más ventas tiene, esto se explica por el abastecimiento de productos de necesidad básica que realizo la población en general en marzo y abril que marcó el inicio de la pandemia.

Figura 3.

Suma de precio por semana del año

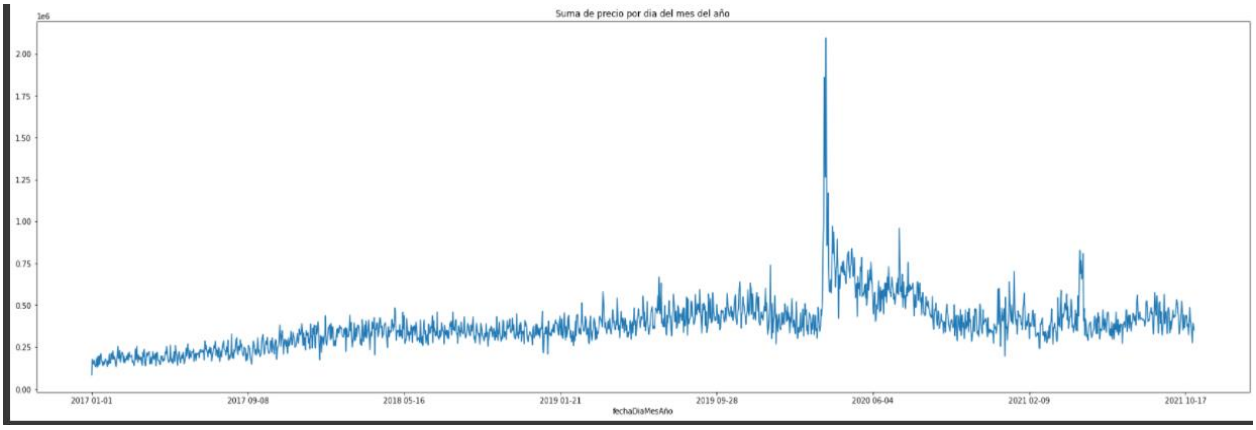


Nota. Elaboración propia (2021)

La suma de ventas por semana tenía un crecimiento estable hasta marzo de 2020 cuando se anunció el inicio de la cuarentena, durante esas semanas las personas realizaron compras masivas de productos de necesidad básica, una vez superado en inicio de la pandemia las ventas retomaron su tendencia normal.

Figura 4.

Suma de precio por dia del mes del año



Nota. Elaboración propia (2021)

La grafica de ventas por día permite apreciar mejor el pico de inicio de pandemia, evidenciando que su duración fue muy corta.

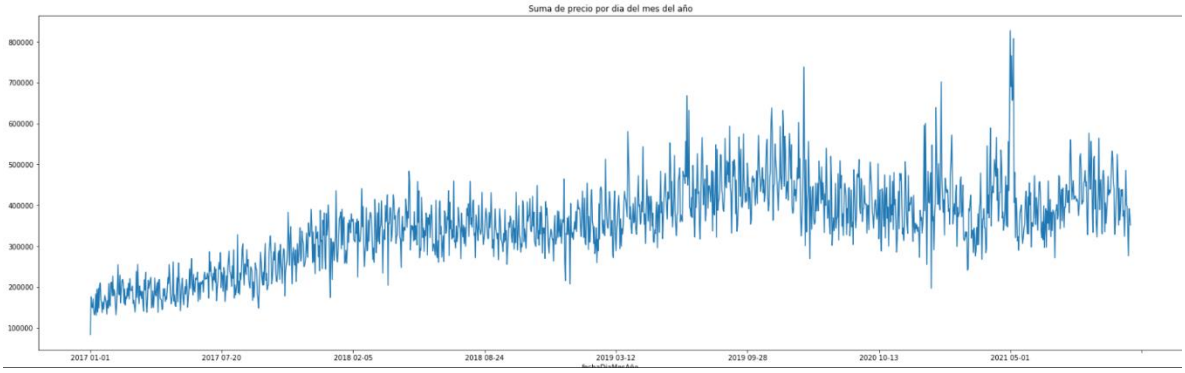
2.5. Eliminación De Atípicos.

El inicio de pandemia creo una necesidad de abastecimiento en la sociedad que hizo que todo el periodo de cuarentena absoluta fuese evidentemente atípico. la marca objetivo comercializa principalmente harinas lo cual es un producto de primera necesidad en muchos hogares del país. Dentro del análisis de ventas en el tiempo se puede evidenciar un gran pico comprendido entre los meses de abril y agosto de 2020, durante este periodo en el país hubo restricciones permanentes en todas las ciudades.

La decisión tomada para mitigar las ventas de este periodo fue eliminar estas ventas del dataset lo cual redujo significativamente al menos visualmente la presencia de datos atípicos.

Figura 5.

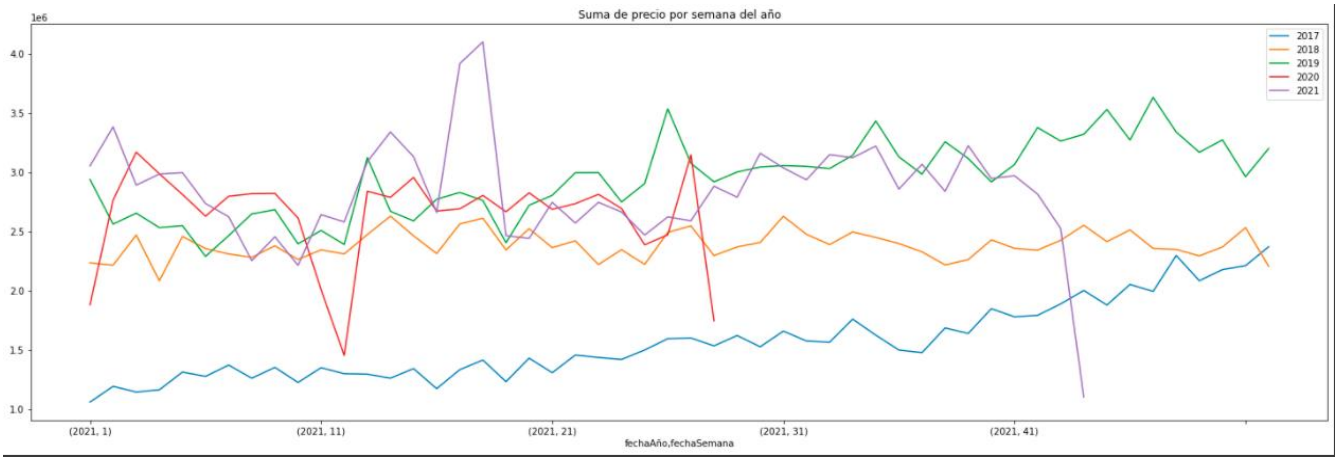
Suma de precio por día del mes del año



Nota. Elaboración propia (2021)

Figura 6.

Suma de precio por semana del año



Nota. Elaboración propia (2021)

La suma de precios por semana del año nos permite identificar comportamientos que se presentan en las ventas en ciertos periodos del año.

2.6. Preparación De Datos Para El Modelo.

El objetivo del modelo es predecir las ventas de todo el panel de tiendas para 7 días posteriores, para esto la data a nivel de transacción como viene inicialmente de base de datos se vuelve

tediosa de manipular y puede llegar a generar ruido con las ventas atípicas, es por esto que se tomó la decisión de presentar al modelo la data agrupada por día.

2.6.1. Estrategia De Preparación De Datos 1.

Se planteo una estrategia que toma cada 7 días la información del año anterior agrupado por año para predecir la suma de ventas de los 7 días posteriores a la predicción siendo

X= data agrupada por día de 365 días atrás.

Y= suma de ventas de la semana siguiente.

Se calcula una vez por semana.

Tabla 2.

Estrategia de preparación de datos 1

X	**Y**
Datos Agrupados por día del 2017-01-01 al 2018-01-01	Suma de ventas de 7 días posteriores al 2018-01-01
Datos Agrupados por día del 2017-01-07 al 2018-01-07	Suma de ventas de 7 días posteriores al 2018-01-07
Datos Agrupados por día del 2017-01-14 al 2018-01-14	Suma de ventas de 7 días posteriores al 2018-01-14

Nota. Elaboración Propia (2021).

Con esta estrategia obtenemos 200 datos para todo el dataset (174 después de la eliminación de inicio de pandemia).

Evidentemente 174 datos para entrenar un modelo es poco teniendo en cuenta que no poseemos una tendencia marcada a nivel de semana del año, es por esto que, con el fin de aumentar la cantidad de datos para el modelo sin perder la integridad de los mismos, se procedió a plantear una nueva estrategia sin perder la estructura de la entrada y la salida.

2.6.2. Estrategia De Preparación De Datos 2.

X: data agrupada por día de 365 días atrás.

Y: suma de ventas de la semana siguiente.

Se calcula una vez por día.

Tabla 3.

Estrategia de preparación de datos 2

X	Y
Datos Agrupados por día del 2017-01-01 al 2018-01-01	Suma de ventas de 7 días posteriores al 2018-01-01
Datos Agrupados por día del 2017-01-02 al 2018-01-02	Suma de ventas de 7 días posteriores al 2018-01-02
Datos Agrupados por día del 2017-01-03 al 2018-01-03	Suma de ventas de 7 días posteriores al 2018-01-03

Nota. Elaboración Propia (2021).

Con esta estrategia obtenemos 1201 datos.

Logramos aumentar la cantidad de filas del dataset un 690,22% lo cual permitirá. con un modelo y ajuste de parámetros correcto generalizar mejor.

3. Proceso De Analítica

3.1.Pipeline Principal

3.1.1. Conversión De Columna Fecha

A partir de los datos obtenidos de la fuente se convierte la columna de fecha al tipo datetime, la lectura del csv la toma como string.

3.1.2. Creación De Nuevas Columnas Para Graficar Y Segmentar.

Se crean nuevas columnas para facilitar la visualización y el particionamiento posterior de los datos, estas columnas son la concatenación de rangos de fecha ej. Mes-día, año-mes, año-semana. Se realiza una transformación de dichos valores para poder obtener graficas ordenadas, evitando por ejemplo que sea más reciente 2021-12 que 2021-2, para esto se agregó un cero al inicio de los campos de longitud uno haciendo que 2021-12 > 2021-02.

3.1.3. Eliminación De Tiendas Que No Vendieron En Todos Los Años Distintos Del Dataset.

Para evitar ruido en los datos y poder garantizar que las ventas sean constantes en el tiempo, se eliminaron las tiendas que no tuvieran ventas en todos los años distintos del dataset, garantizando así que todos los años las tiendas activas son las mismas.

3.1.4. Cambio De Tipos De Variable.

Las variables “idTienda", "idVenta", "CodigoDeBarras", "hora", tomaron por defecto el tipo de dato int32, estas variables son más identificadores que valores numéricos operables. Son usadas durante la exploración de datos para entender un poco que contienen.

3.1.5. Eliminación De Variables Que No Aportan Al Análisis.

Se eliminan las variables creadas en el punto 5.1.2 una vez se hayan usado para los análisis correspondientes, además de estas variables se eliminarán también las variables "idTienda”, “hora", "idVenta", “latitud” y “longitud, las cuales no presentan relevancia para el análisis y no tiene relación con la variable objetivo

3.1.6. Creación De Variables Dummies / One Hot Encoder.

Para poder convertir las variables categóricas a valores que entienda un modelo se aplicó la técnica de One Hot Encoder, que permite generar una columna por cada valor distinto que toma una variable categórica con el fin de evitar dar más peso a unas que a otras, esta técnica marca la

fila con un valor de 1 en la columna que corresponda a la aparición del feature que contiene, el resto de los valores de esa columna lo marca con 0 (2). el one hot encoder se les aplica a aquellas variables que no posean orden jerárquico, en este caso todas las categóricas.

3.1.7. Creación De La Serie De Tiempo.

Se preparo los datos, de tal manera que el resultado tenga la estructura y las reglas del grafico 4.2

3.1.8. División De Los Datos Para Entrenamiento Y Pruebas.

Se dividió el 25% de los datos para probar el rendimiento de los distintos modelos, para esto hacemos uso del paquete de sklearn especializado para dicho fin.

3.1.9. Entrenamiento Y Validación De Varios Modelos.

Se entreno con distintos modelos y distintos parámetros para analizar las métricas de desempeño, en especial el error que para este caso es el error cuadrático medio, error absoluto medio y r2, los algoritmos usados para crear los modelos fueron: Regresión Lineal, Regresión lasso, random forest, gradient boosting y una red neuronal recurrente.

4. Metodología.

4.1. Métricas De Los Distintos Modelos.

Para medir el rendimiento de los modelos se usaron 3 estimadores de error que permiten conocer cuantitativamente que tan bueno es el modelo, los 3 estimadores son: Error absoluto medio, Error cuadrático medio, R2.

4.1.1. Regresión Lineal

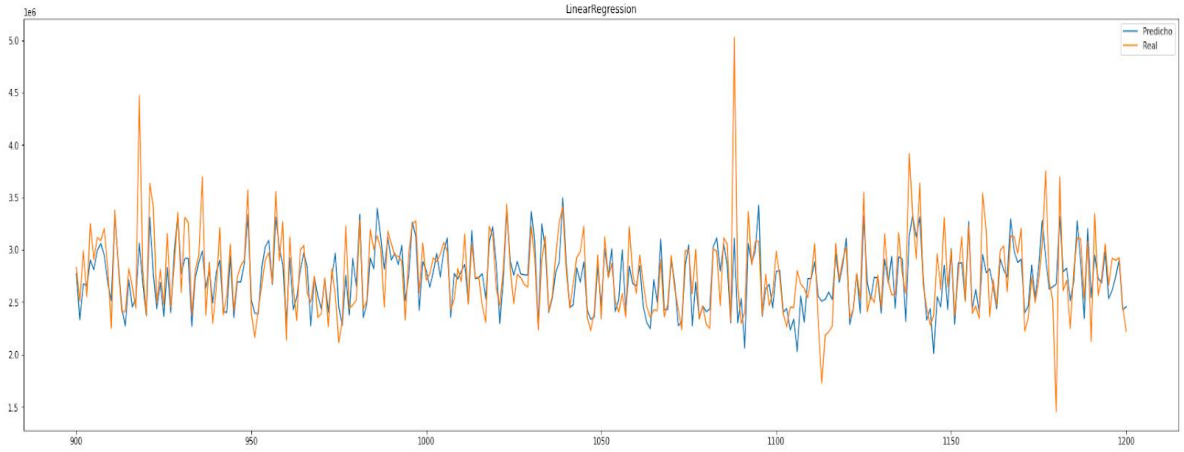
Error absoluto medio: 180137.7289668095

Error cuadrático medio 70707416030.97879

Error r2 0.5586482965306911

Figura 7.

Test con regresión lineal.



Nota. Elaboración propia (2021)

4.1.2. Regresión Lasso.

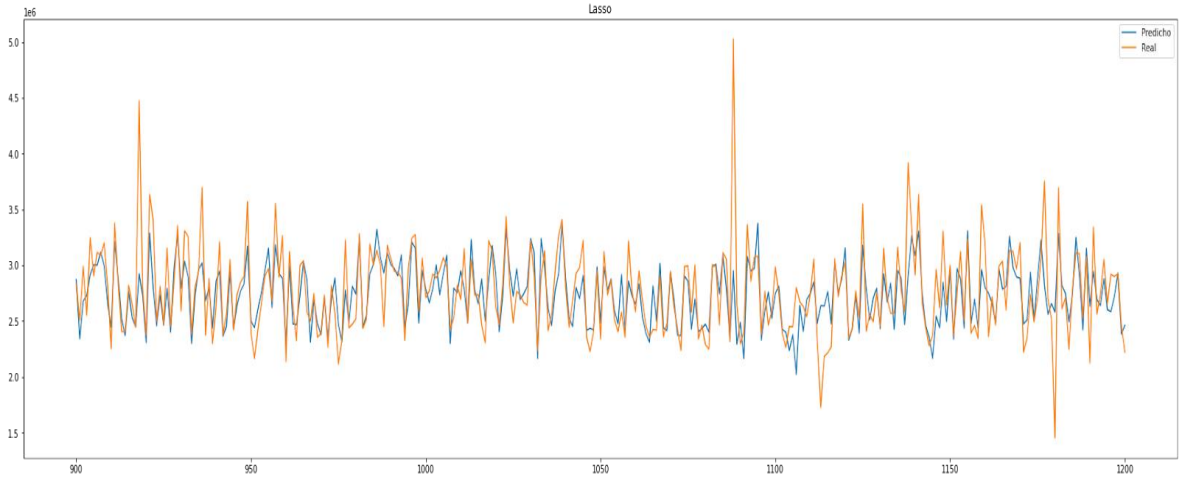
Error absoluto medio: 179975.3921097904

Error cuadrático medio 76527226685.87225

Error r2 0.5223213666188316

Figura 8.

Test con regresión lasso.



Nota. Elaboración propia (2021)

4.1.3. Random Forest.

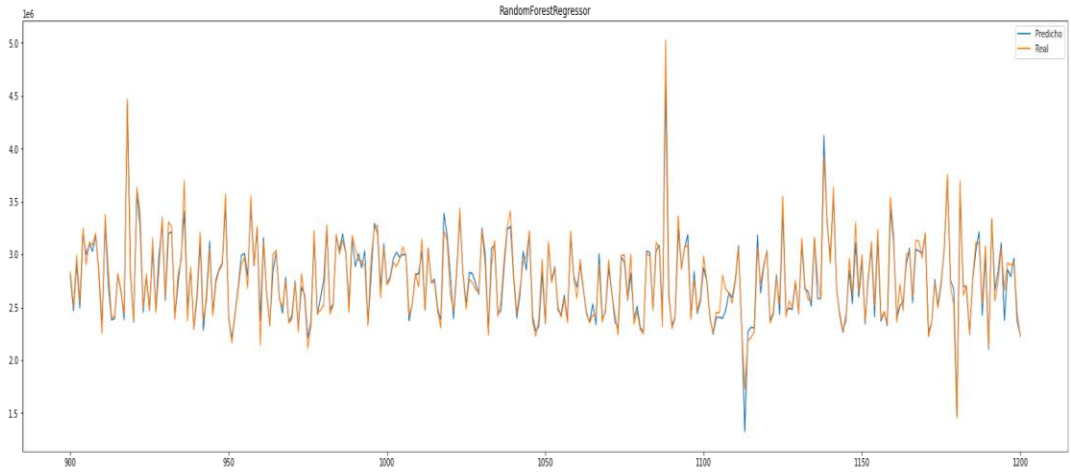
Error absoluto medio: 60820.17873078466

Error cuadrático medio 7353017063.874265

Error r2 0.954102882145234

Figura 9.

Test con random forest.



Nota. Elaboración propia (2021)

La gráfica representa el periodo de tiempo que se tomó para pruebas, es decir el modelo nunca conoció estos datos durante su entrenamiento, aun así, logro predecir de una manera muy acertada las ventas de dicho periodo

4.1.4. Gradient Boosting.

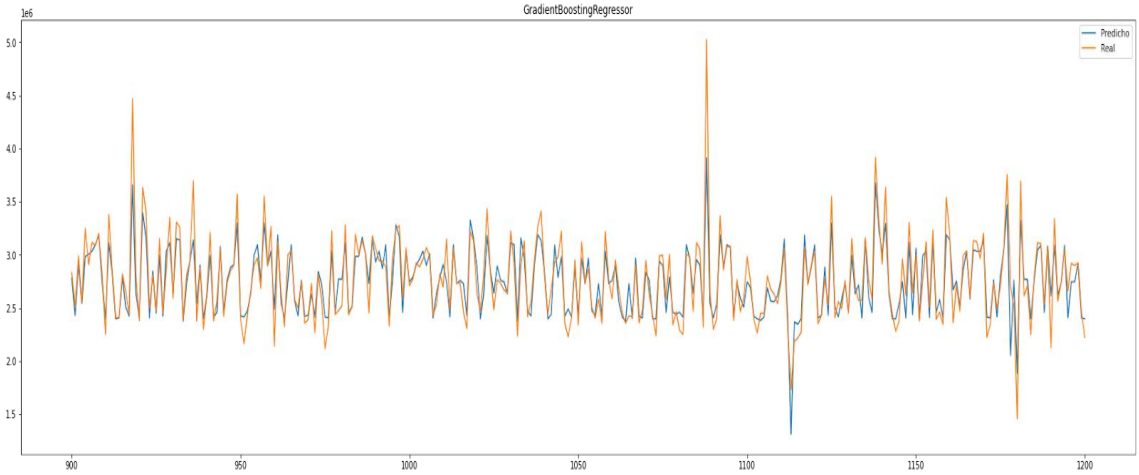
Error absoluto medio: 118403.54100337006

Error cuadrático medio 28741523154.286907

Error r2 0.8205970332887087

Figura 10.

Test con gradient boosting.



Nota. Elaboración propia (2021)

4.1.5. RNN.

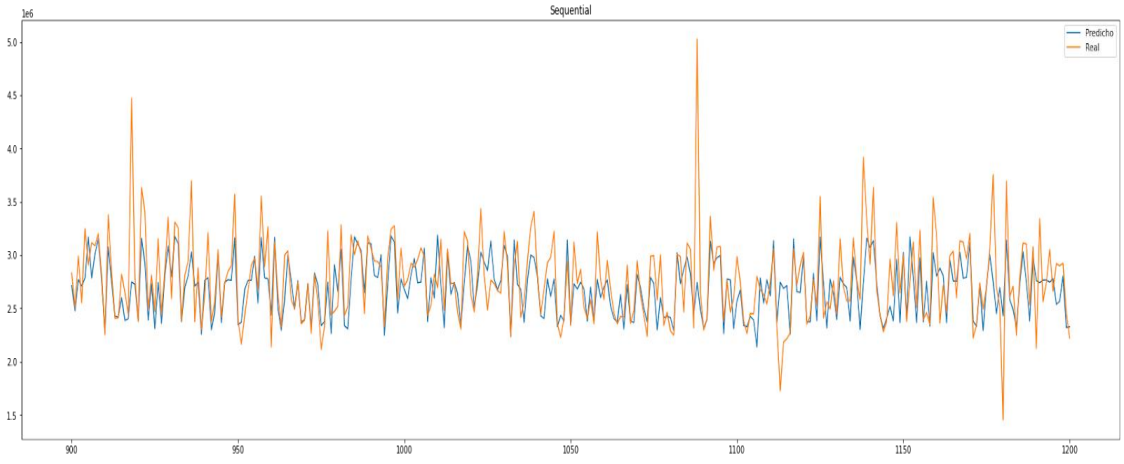
Error absoluto medio: 195638.37126245847

Error cuadrático medio 92037000373.60527

Error r2 0.425510233901079

Figura 11.

Test con RNN



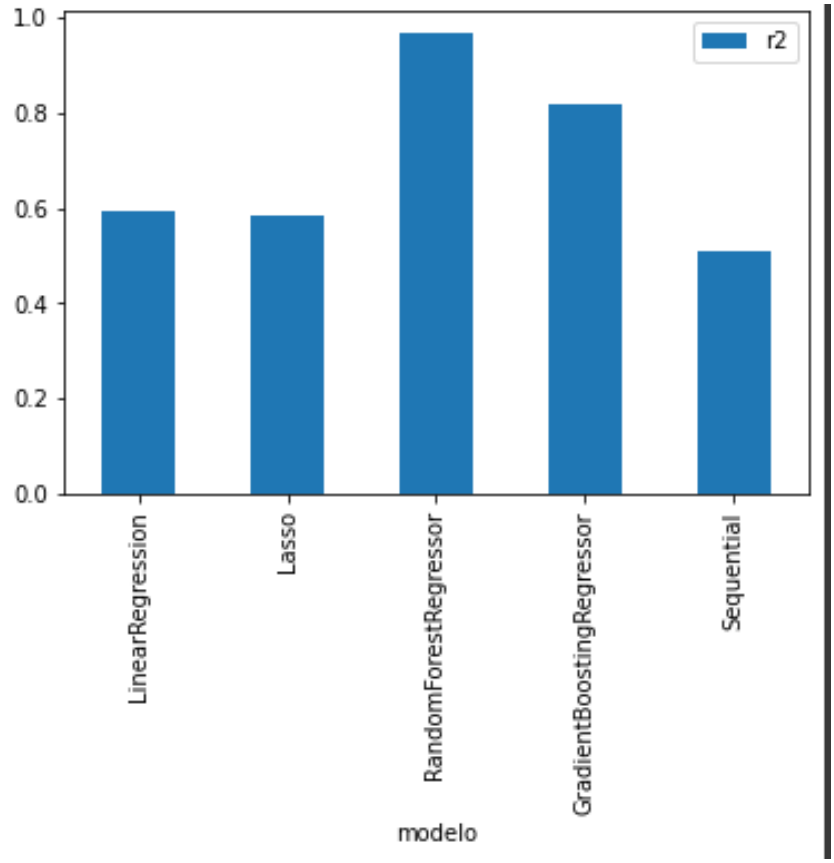
Nota. Elaboración propia (2021)

4.2. Métricas De Los Distintos Modelos.

4.2.1. R2

Figura 12.

Error R2 por modelo



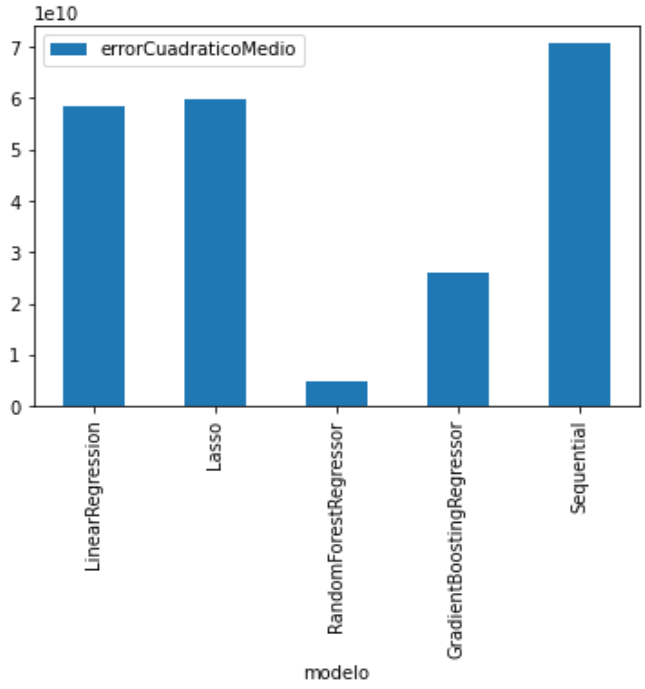
Nota. Elaboración propia (2021)

A mayor valor de esta métrica, mejor será el modelo.

4.2.2. Error Cuadrático Medio.

Figura 13.

Error Cuadrático medio por modelo

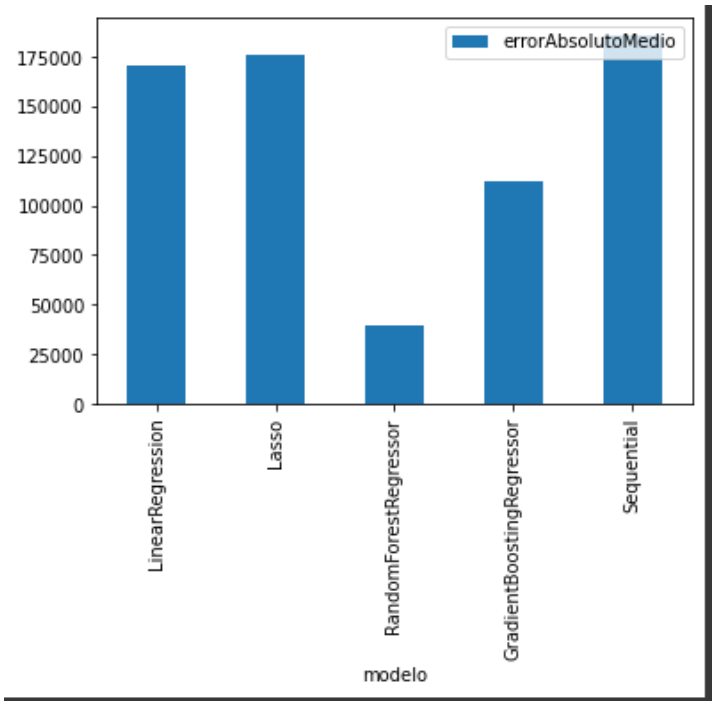


Nota. Elaboración propia (2021)

4.2.3. Error Absoluto Medio.

Figura 14.

Error absoluto medio por modelo



Nota. Elaboración propia (2021)

5. Bibliografía

Gutiérrez y Hernández (2007). *¿Qué es un POS?* Oracle.

https://www.oracle.com/co/industries/what-is-pos/

SK learn preprocessing. One Hot Encoder. (2021)

https://scikitlearn.org/stable/modules/generated/sklearn.preprocessing.OneHotEncoder.html