

INTELIGENCIA ARTIFICIAL AVANZADA PARA LA CIENCIA DE DATOS

*Carolina Arratia Camacho
Frida Lizett Zavala Pérez
Fabián González Vera
Jazzareth Bernal Martínez*



Introducción

Este reto fue obtenido de una competencia en Kaggle cuyo objetivo es predecir las ventas de las miles de familias de productos que se encuentran en las tiendas Favorita, ubicadas en Ecuador.

PLANEACIÓN DE ANÁLISIS

Objetivo

Se busca crear un modelo que logre predecir de manera precisa las ventas unitarias de los productos en las diferentes tiendas de la marca Favorita, con el fin de que los minoristas tengan inventario de los productos al momento adecuado, se evite el desperdicio y se garantice el abasto.

PREGUNTAS

PRINCIPALES

- ¿Existe algún patrón respecto a las fechas para las ventas de cada familia de productos?

SECUNDARIAS

- ¿Las ventas se vieron influidas por el terremoto ocurrido en el país?
- ¿Cómo influyen los precios del petróleo (oil) en las ventas?
- ¿Existe alguna relación entre los días festivos y las ventas?

CONJUNTO DE DATOS QUE SE TIENEN

Se utilizaran los archivos obtenidos de la competencia "Store sale- Time Series Forecasting" de Kaggle considerando predicciones de Agosto 2017:

- **Train.csv**: Los datos de capacitación, incluyen fechas, información de la tienda y del producto, si ese artículo se estaba promocionando, así como las cifras de ventas.
- **Test.csv**: Incluye los mismos datos que el archivo de entrenamiento pero sin el valor de las ventas que es el que se busca predecir
- **Stores.csv**: Datos de cada tienda, incluidos ciudad, estado, tipo y clúster.
- **oil.csv**: Precio diario del petróleo.
- **holidays_events.csv**: Datos de las vacaciones y eventos celebrados en el país.

CRITERIOS DE INCLUSIÓN/EXCLUSIÓN

Se conservan datos que tengan relación con el aumento y decrecimiento de ventas

Se conservan aquellos que no sean repetitivos, y que no puedan ser inferidos con alguna otra variable

Se busca que los datos sean completos y abarquen un periodo en donde las tiendas analizadas ya hubieran sido inauguradas



VARIABLES QUE SE UTILIZARÁN EN EL ANÁLISIS

Para la realización del modelo se considera utilizar los datos contenido en el archivo de train.csv

date

Fecha en la que se vendieron dichos productos

store_nbr

Nos dice el número de la tienda, el cual funciona como un indicador

family

Se refiere al tipo de cada producto vendido. ¿A qué familia pertenece

sales

Da el total de ventas por familia de productos, en cada tienda.

onpromotion

Señala el número de productos en una familia que estuvieron en promoción

VARIABLES QUE SE UTILIZARAN EN EL ANÁLISIS

Otras variables a tomar en cuenta que nos pueden decir el por qué del comportamiento de nuestra predicción se encuentran en los demás archivos

holidays_events.csv

date

Fecha de dicho evento o suceso

description

Menciona de qué se trata dicho día festivo

oil.csv

date

Nos dice el número de la tienda, el cual funciona como un indicador

dcoilwtico

Se refiere al precio del crudo en un día.

MÉTODOS DE SOFTWARE ESTADÍSTICOS QUE SE EMPLEARÁN

python

librería de Python especializada en la manipulación y el análisis de datos.

Pandas

Numpy

Biblioteca para el lenguaje de programación Python que da soporte para crear vectores y matrices grandes multidimensionales

Kaggle

Plataforma de competencia de ciencia de datos

Librerías de análisis estadísticos

Librerías de python para análisis estadísticos y procesamiento de los datos como: sklearn entre otras.

TABLAS SIN DATOS TRAIN.CSV

date
tipo:
datetime

store_nbr
tipo:
Int

family
tipo:
String

onpromotion
tipo:
Int

Sales
tipo:
Int

Valores de:
31-12-2012
a
14-08-2017

Valores de:
1 a 54

Valores
Unicos

Valores de:
0 a 741

Valores de:
0 a 125k

TABLAS SIN DATOS (HOLIDAYS)

date

tipo:

datetime

description

tipo:

String

Valores de:

31-12-2012

a

14-08-2017

**Valores
Unicos**

TABLAS SIN DATOS (OIL)

date

tipo:

datetime

dcoilwtico

tipo:

Float

Valores de:

31-12-2012

a

14-08-2017

Valores de:

26.2 a 2.11

ESTIMACIÓN DE TIEMPO Y RECURSOS



4



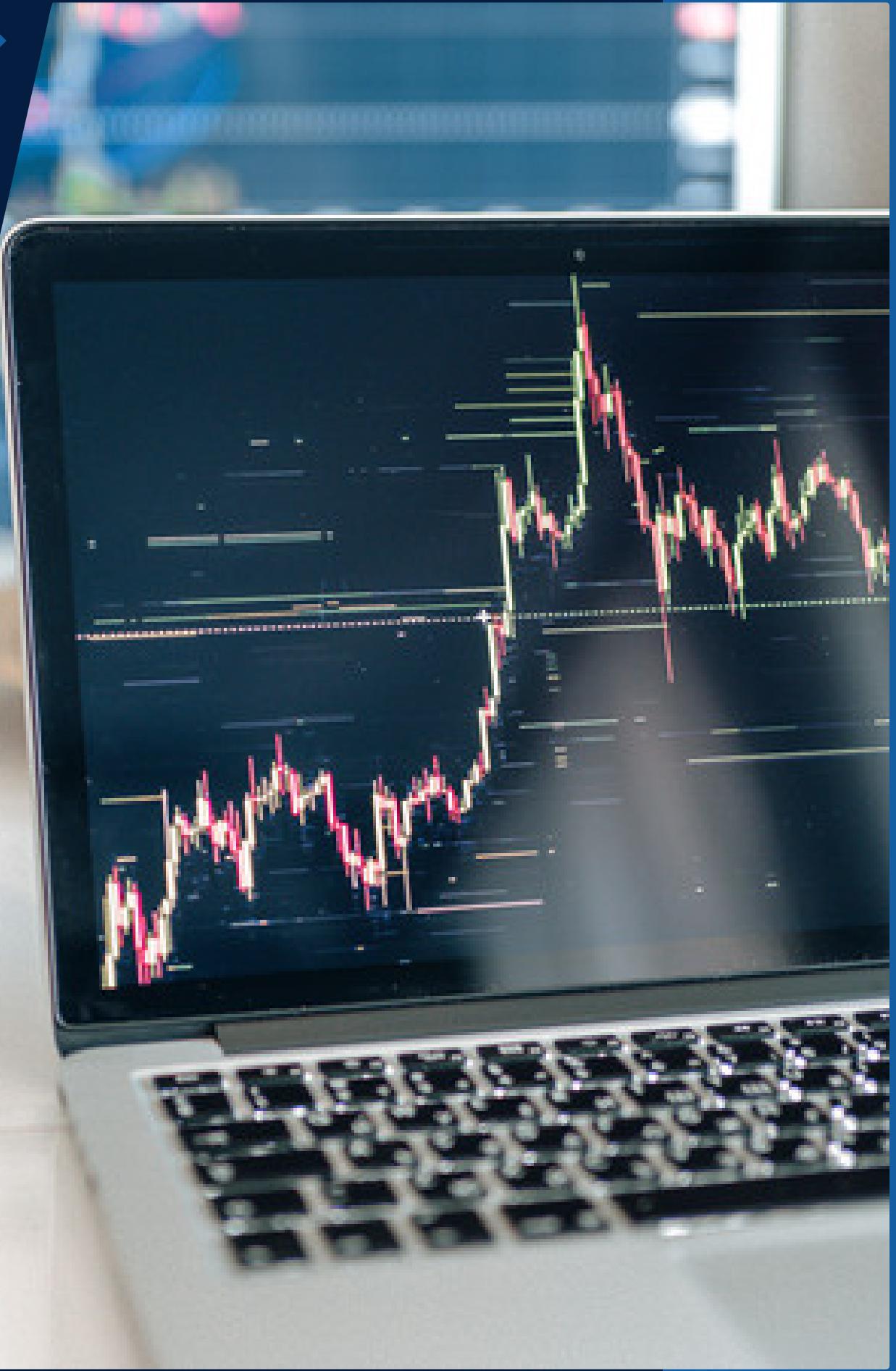
Analistas

Se estima que el tiempo necesario para el análisis es de 5 semanas.

- Parte 1. Limpieza y normalización de los datos
- Parte 2. Aplicación y prueba del modelo
- Parte 3. Análisis de resultados

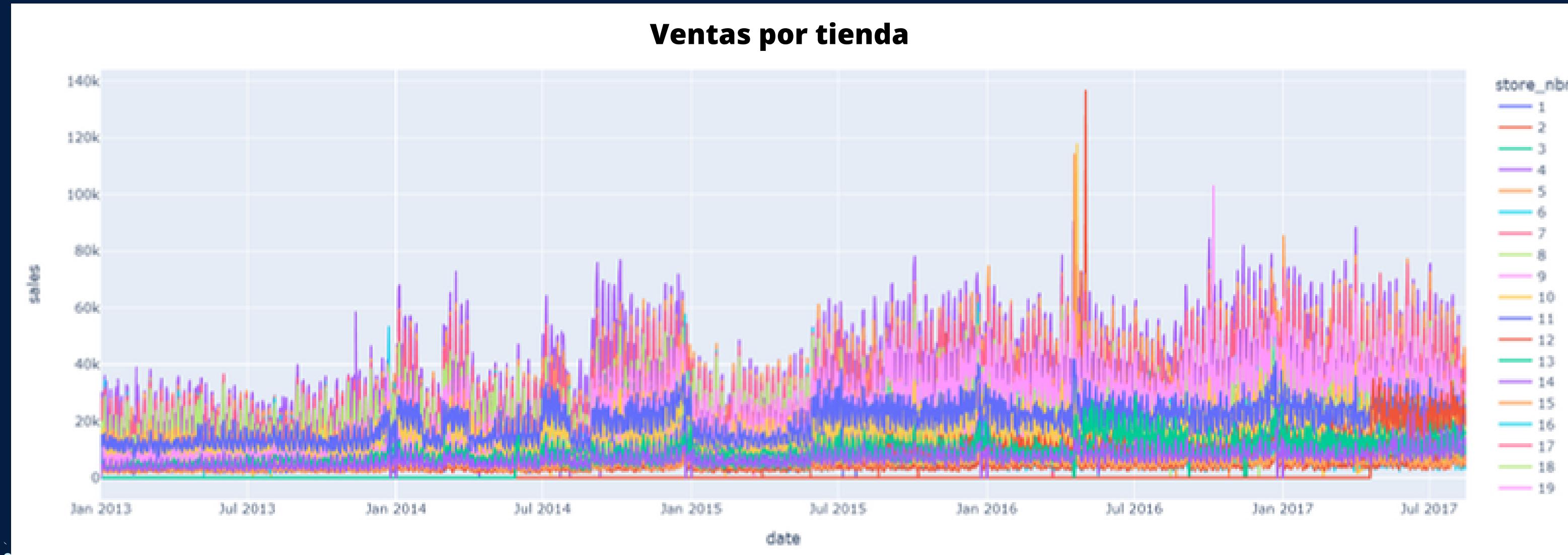


LIMPIEZA Y ANALISIS DEL CONJUNTO DE DATOS



VENTAS POR TIENDA

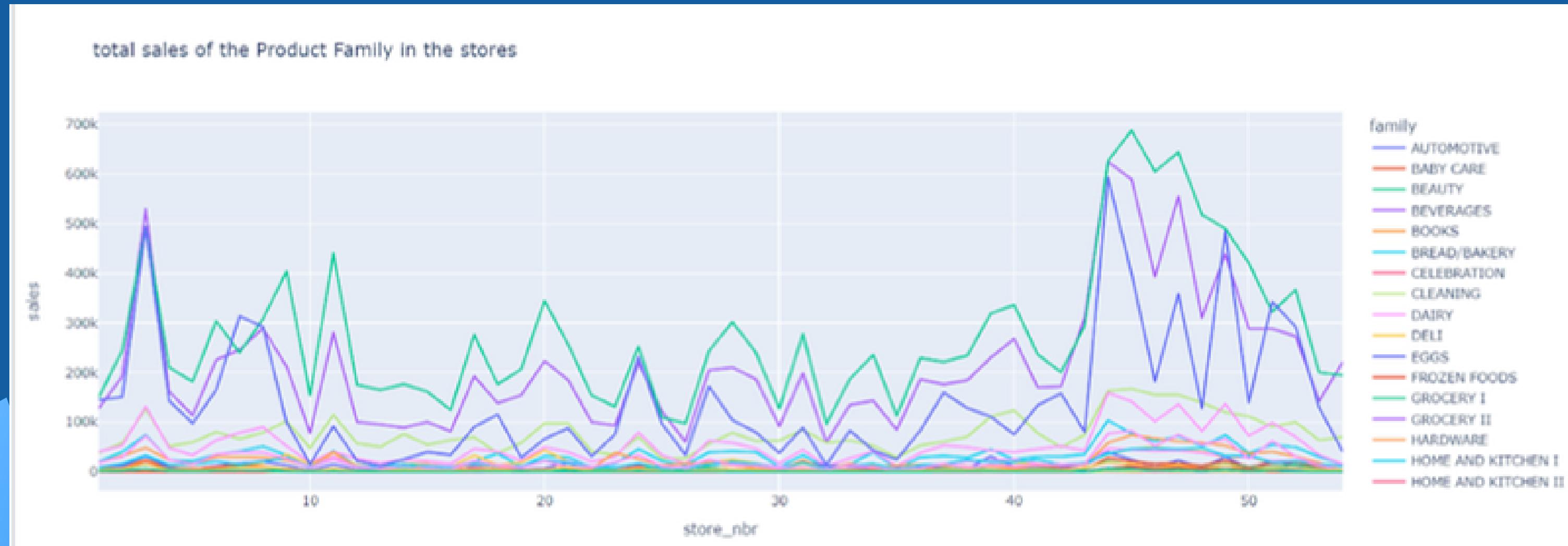
“Predecir las ventas unitarias de los productos vendidos en las diferentes tiendas de la marca Favorita.”



En la parte inferior tenemos tiendas cuyas ventas son de 0 hasta cierto periodo de tiempo, esto se debe a que algunas tiendas abrieron mucho después.

VENTAS POR FAMILIA

Por otra parte es importante revisar que no tengamos ninguna anomalía con respecto a nuestras ventas por familia de productos en las tiendas



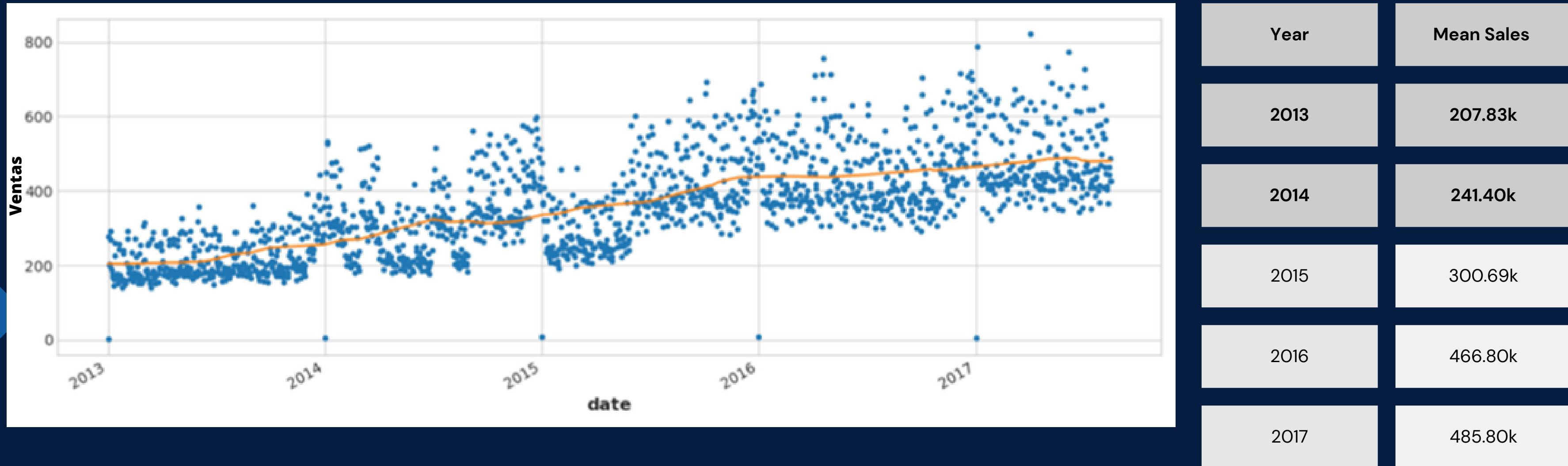
Como podemos observar en la parte inferior de la gráfica, existen productos cuya sumatoria total de venta de cierta familia de productos es igual a 0, esto es debido a que varias tiendas no ofrecen todos los productos.



¿EL TERREMOTO AFECTÓ DE MANERA SIGNIFICATIVA LAS VENTAS DE LAS TIENDAS?

“Un terremoto de magnitud 7,8 sacudió Ecuador el 16 de abril de 2016. La gente se unió a los esfuerzos de ayuda donando agua y otros productos de primera necesidad, lo que afectó en gran medida las ventas de los supermercados durante varias semanas después del terremoto.”

VENTAS ABRIL POR AÑO



Producto	Marzo	Abril-Mayo	Junio
Grocery	4,800.774	4,841.85	4,702.43
Beverage	3,733.14	3,592.93	3,580.13
Produce	2,442.18	2,404.91	2,403.45

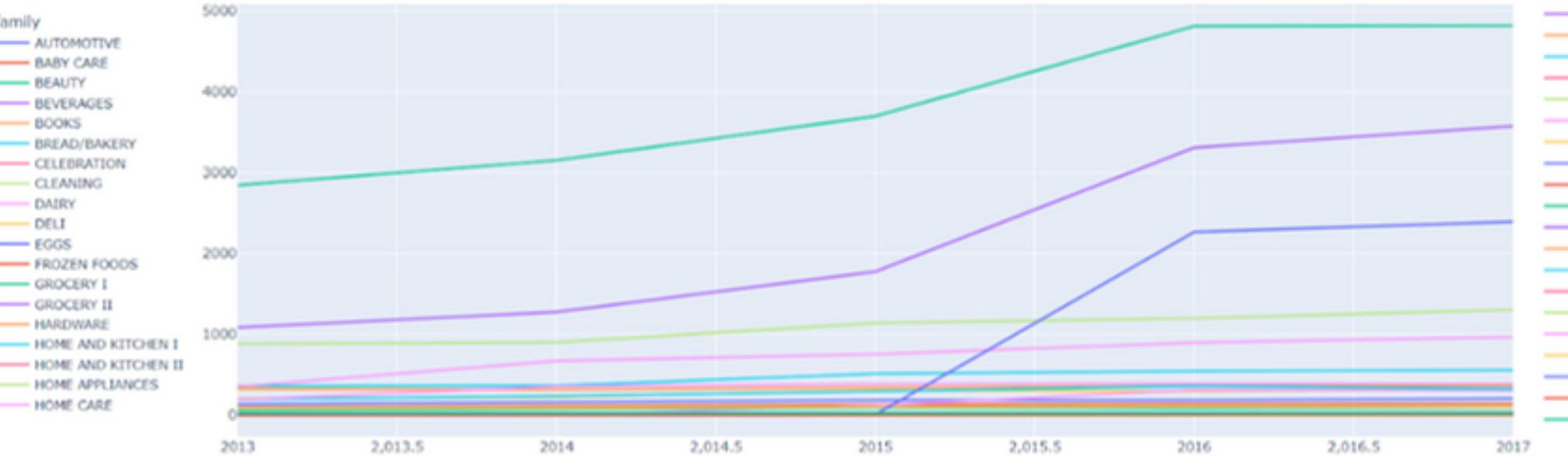
Podemos descartar que se tuvo un impacto relevante para las ventas generales de ese mes comparado a los anteriores años.

VENTAS POR TIENDA CADA MES EN 2016

Marzo 2016

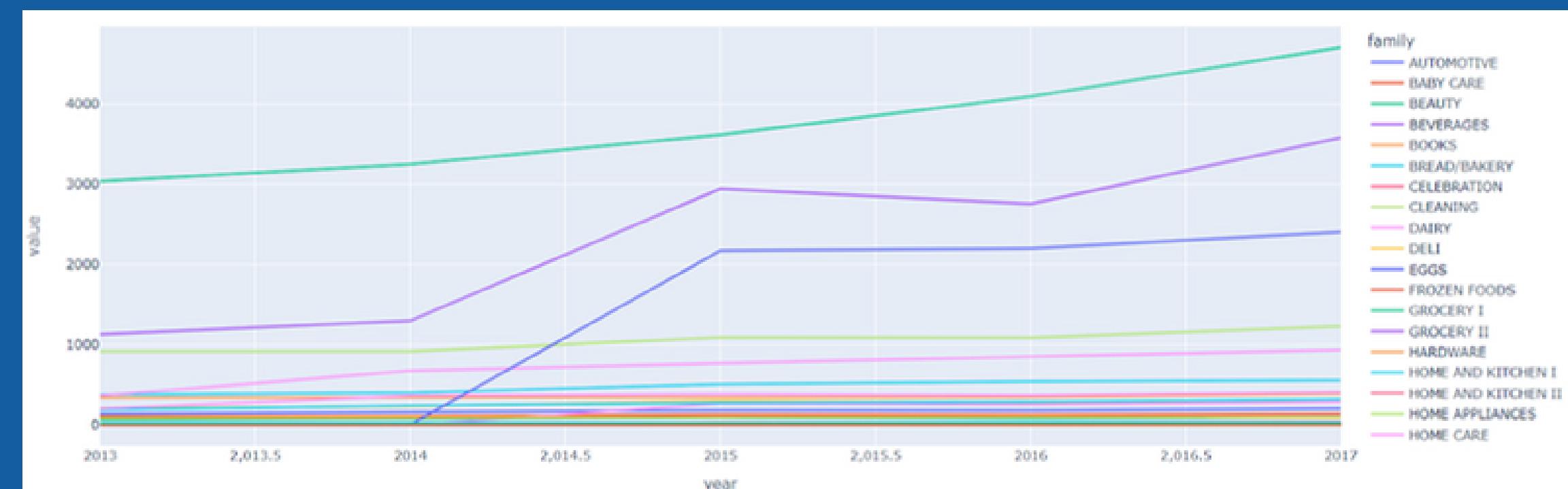


Abril-Mayo 2016



Productos más vendidos:

- Grocery
- Beverage
- Produce



¿EXISTE ALGUNA RELACIÓN ENTRE LOS DÍAS FESTIVOS Y LAS VENTAS?

Dentro del Dataset, se incluyen archivos adicionales que pueden contener información suplementaria para el modelo, dentro de estos se encuentra el archivo `holidays_events.csv`, el cual contiene información de eventos festivos y de diversos eventos como lo sería los partidos del mundial de fútbol.
Se buscó analizar si dichos eventos tienen algún efecto en las ventas,

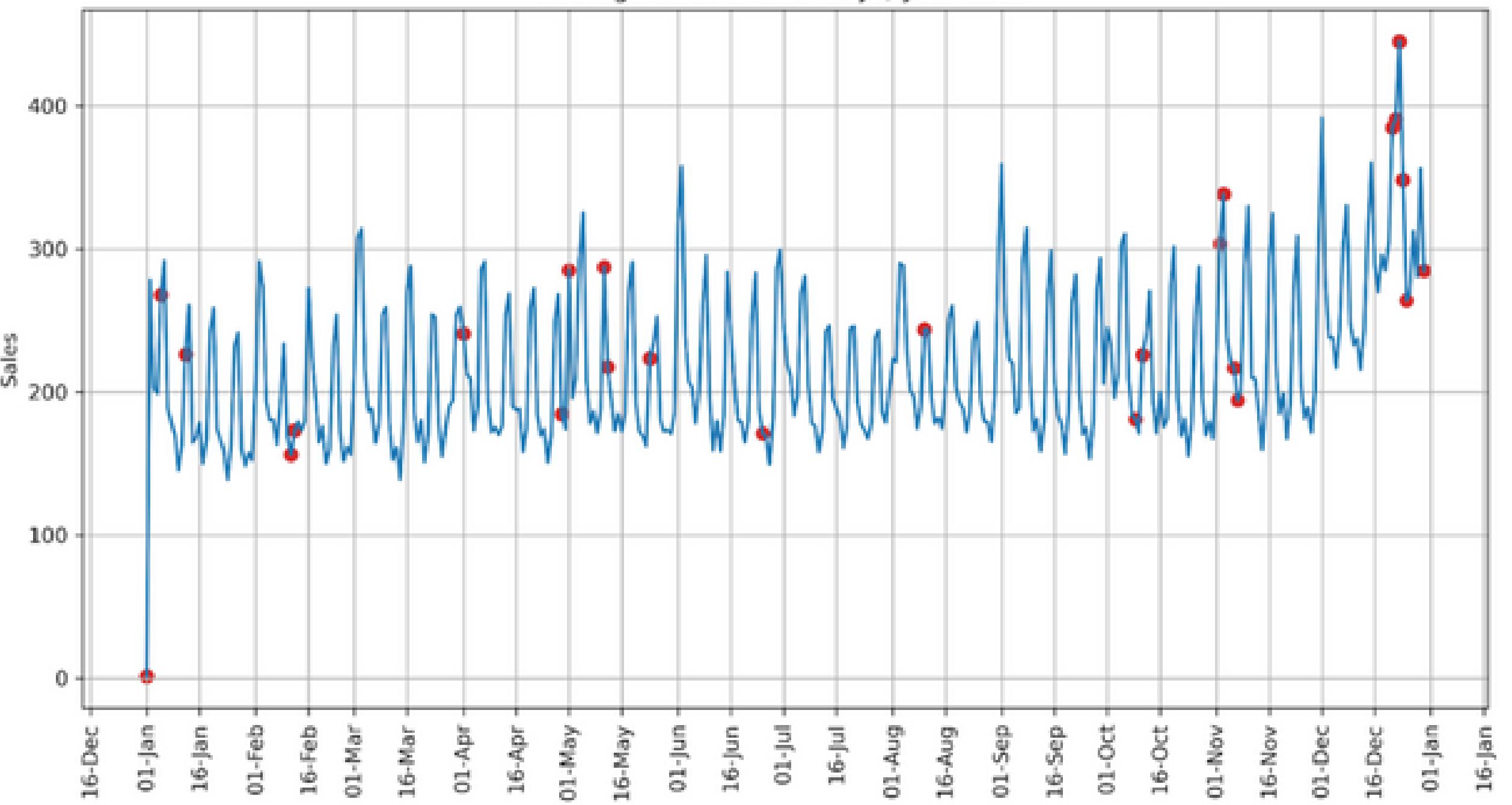


description

date

2013-01-01	Primer dia del año
2013-01-05	Recuperacion puente Navidad
2013-01-12	Recuperacion puente primer dia del año
2013-02-11	Carnaval
2013-02-12	Carnaval
2013-04-01	Provincializacion de Cotopaxi
2013-04-29	Tiernes Santo
2013-05-01	Dia del Trabajo
2013-05-11	Dia de la Madre - I
2013-05-12	Dia de la Madre
2013-05-24	Batalla de Pichincha
2013-06-25	Provincializacion de Imbabura
2013-08-10	Primer Grito de Independencia
2013-10-09	Independencia de Guayaquil
2013-10-11	Segunda Independencia de Guayaquil
2013-11-02	Dia de Difuntos
2013-11-03	Independencia de Cuenca
2013-11-06	Provincializacion de Santo Domingo
2013-11-07	Provincializacion Santa Elena
2013-12-21	Navidad-4
2013-12-22	Navidad-5

Average sales and holidays, year 2013



¿ES EL PRECIO DEL PETRÓLEO UN PARÁMETRO A TENER EN CUENTA?

Ecuador es un país dependiente del petróleo y puede ser vulnerable a los cambios del precio de este. Es por esto que se quiere averiguar si de verdad el precio del petróleo representa un parámetro considerable que podría afectar las ventas.

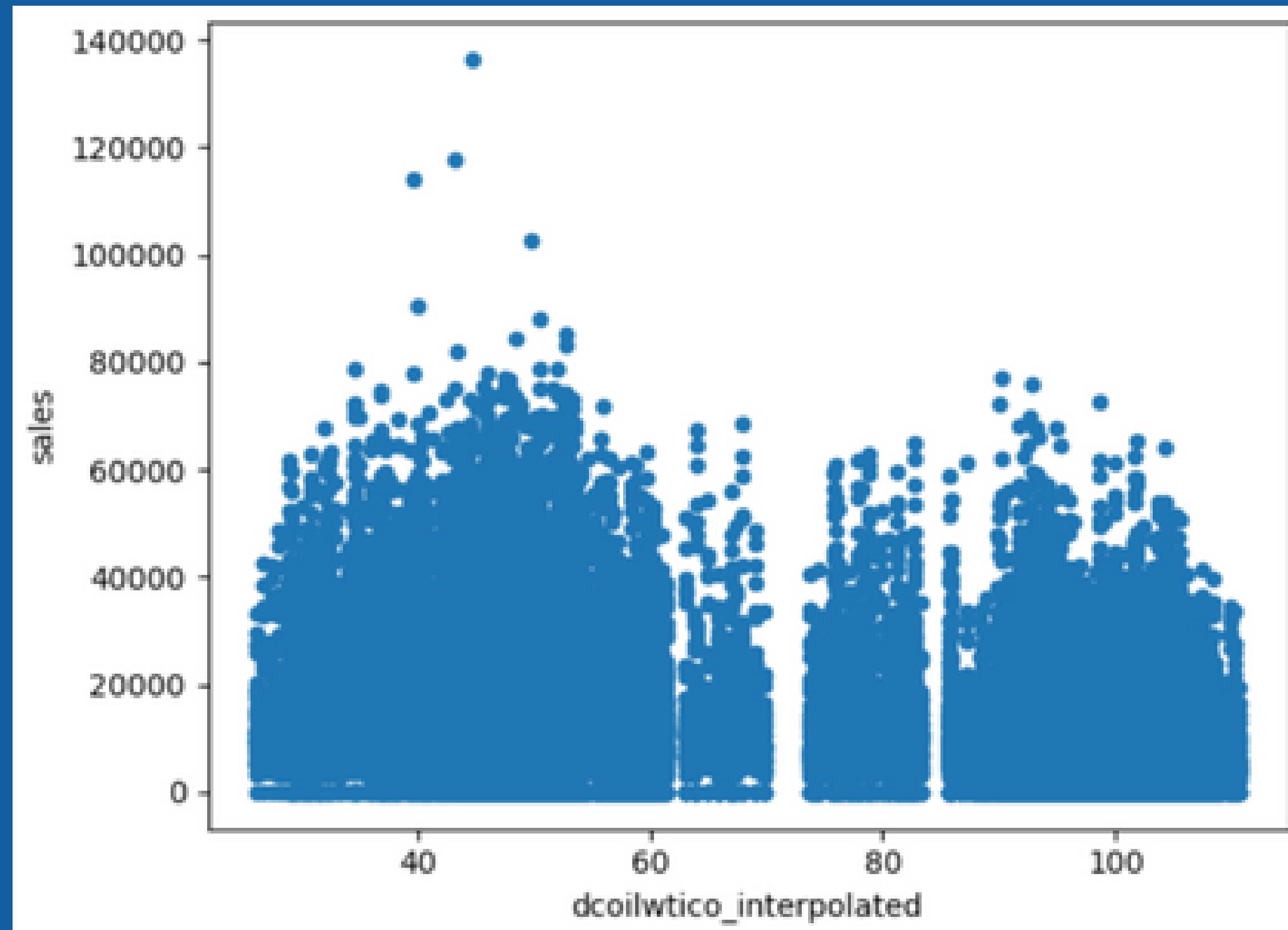


Precio diario del petróleo

- Los datos con los que se contaban para el precio del petróleo tenían algunos días faltantes, sin embargo estos se podían inferir.



No tienen una relación tan estrecha como para considerar al precio del petróleo como algo que repercute de manera importante en las ventas, tampoco existe una aparente correlación, por lo que no será tomado en cuenta.



Manejo de los datos

Transformar los datos de días festivos
con onehot

Transformar los datos de las familias
con onehot



MODELO DE PREDICCIÓN

CONFIGURACIÓN Y ENTRENAMIENTO



REGRESIÓN LINEAL

Utilizar todos los datos

En este modelo se adaptaron todas las columnas que se consideraron relevantes para el modelo para correr una regresión simple



[AvancesReg_reto - Version 7](#)

Complete · 10d ago

2.27321

El modelo era muy simple para la complejidad de los datos

XGBOOST

- Implementacion del algoritmo de árboles aumentados de gradientes.
- GBLinear.



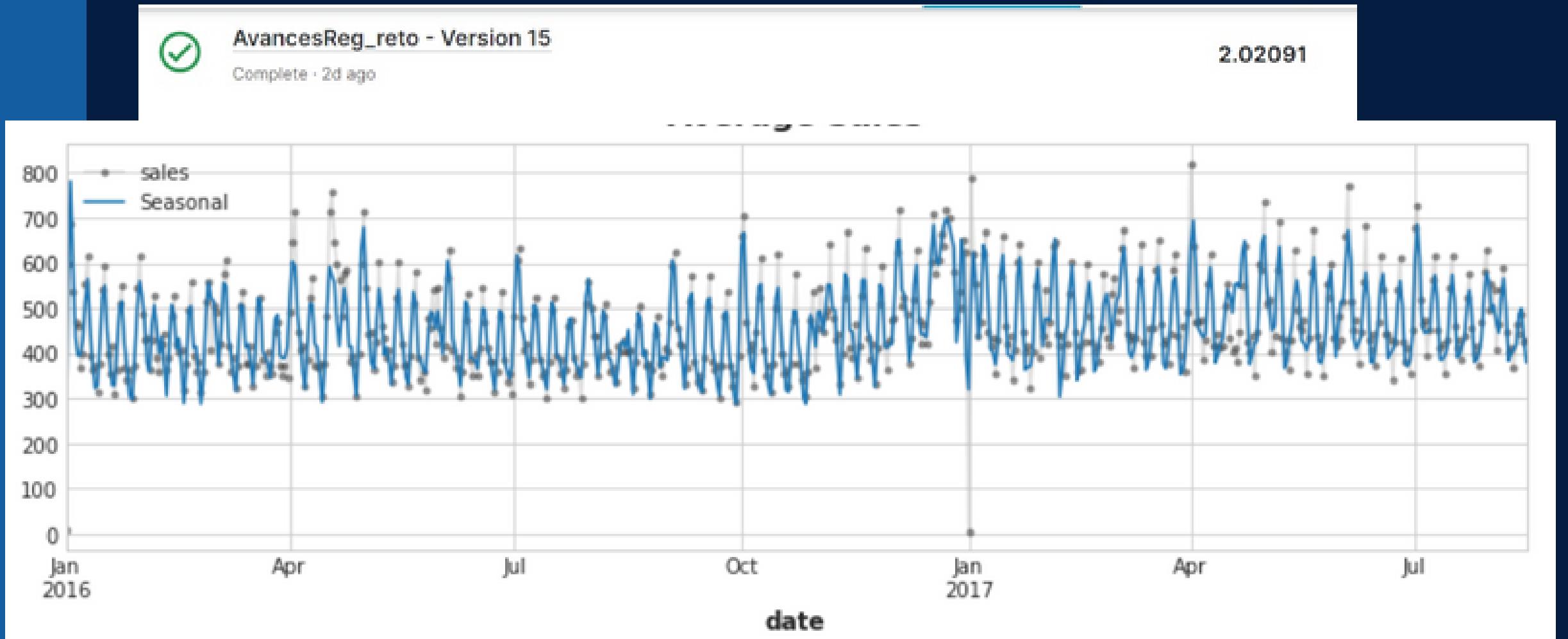
2.2564



SEASONALITY

Transformación de Fourier: La transformada de Fourier permite transformar una función de tiempo y señal en una función de frecuencia y potencia. Esto le indica qué frecuencias componen su señal y qué tan fuertes son.

En nuestro caso la señal son las ventas y podríamos esperar algún tipo de frecuencia semanal o diaria.



- Seasonality Features
- holidays 1hot Features
- Regresión lineal

RANDOM FOREST

- Implementacion del algoritmo de Random Forest
- Winsorización



AvancesReg_reto - Version 16

Complete · 2d ago

0.60695





MÉTRICAS PARA EVALUAR EL DESEMPEÑO DEL MODELO

RMSLE

La métrica de evaluación para este concurso es el error logarítmico cuadrático medio. El RMSLE se calcula como:

$$\text{RMSLE} = \sqrt{\frac{1}{n} \sum_{i=1}^n (\log(y_l + 1) - \widehat{\log(y+1)})^2}$$

Es una extensión del error cuadrático medio (MSE) que se utiliza principalmente cuando las predicciones tienen grandes desviaciones, como es el caso de esta competencia de predicción de energía. Los valores van desde 0 hasta millones y no queremos castigar las desviaciones en la predicción tanto como con MSE.



REFINAMIENTO DEL MODELO

- AJUSTE DE HIPERPARÁMETROS
- TÉCNICAS DE REGULARIZACIÓN



IMPLEMENTACIÓN DE INTERFAZ



- Django
- Interfaz web

Gracias

