Pronóstico de Ventas para Walmart

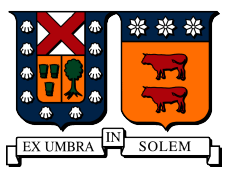
Proyectos estadísticos

Prof. Francisco Alfaro

Helena Bahamondes
Carlos Tapia
Mateo Hidalgo
Brian Valenzuela

Universidad Técnica Federico Santa María

Martes 2 de Diciembre 2025



Problema

- Trabajaremos con datos provenientes de Walmart.
- Nuestro objetivo es predecir ventas futuras (28 días).

Datos

- sales_train_validation.csv:

id	item_id	dept_id	cat_id	store_id	state_id	d_1	d_2	d_3	d_4	...	d_1904	d_1905	d_1906	d_1907	d_1908	d_1909	d_1910	d_1911	d_1912	d_1913
HOBBIES_1_001_CA_1_validation	HOBBIES_1_001	HOBBIES_1	HOBBIES	CA_1	CA	0	0	0	0	...	1	3	0	1	1	1	3	0	1	1
HOBBIES_1_001_CA_2_validation	HOBBIES_1_001	HOBBIES_1	HOBBIES	CA_2	CA	0	0	0	0	...	0	0	1	2	0	4	0	0	2	2
HOBBIES_1_001_CA_3_validation	HOBBIES_1_001	HOBBIES_1	HOBBIES	CA_3	CA	0	0	0	0	...	0	2	4	0	1	1	1	0	3	3
HOBBIES_1_001_CA_4_validation	HOBBIES_1_001	HOBBIES_1	HOBBIES	CA_4	CA	0	0	0	0	...	1	0	0	1	0	1	0	1	2	1
HOBBIES_1_001_TX_1_validation	HOBBIES_1_001	HOBBIES_1	HOBBIES	TX_1	TX	0	0	0	0	...	0	0	0	0	1	0	1	0	0	2
HOBBIES_1_001_TX_2_validation	HOBBIES_1_001	HOBBIES_1	HOBBIES	TX_2	TX	0	0	0	0	...	0	0	0	0	0	1	0	0	0	0
HOBBIES_1_001_TX_3_validation	HOBBIES_1_001	HOBBIES_1	HOBBIES	TX_3	TX	0	0	0	0	...	0	3	0	0	0	0	0	0	0	0
HOBBIES_1_001_WI_1_validation	HOBBIES_1_001	HOBBIES_1	HOBBIES	WI_1	WI	0	0	0	0	...	0	0	0	0	0	0	1	2	1	1
HOBBIES_1_001_WI_2_validation	HOBBIES_1_001	HOBBIES_1	HOBBIES	WI_2	WI	0	0	0	0	...	1	1	0	0	0	0	0	0	0	1
HOBBIES_1_001_WI_3_validation	HOBBIES_1_001	HOBBIES_1	HOBBIES	WI_3	WI	0	0	0	0	...	0	1	0	0	0	0	0	0	0	0

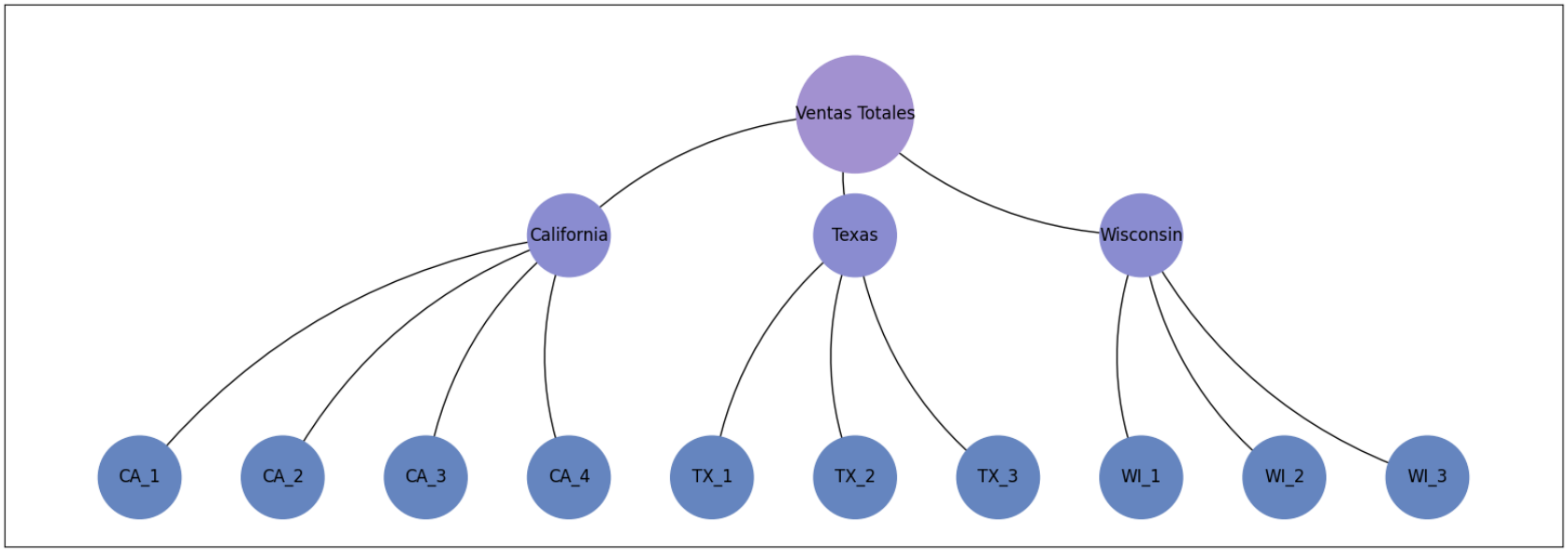
- calendar.csv:

date	wm_yr_wk	weekday	wday	month	year	d	event_name_1	event_type_1	event_name_2	event_type_2	snap_CA	snap_TX	snap_WI
2011-01-29	11101	Saturday	1	1	2011	d_1	NaN	NaN	NaN	NaN	0	0	0
2011-01-30	11101	Sunday	2	1	2011	d_2	NaN	NaN	NaN	NaN	0	0	0
2011-01-31	11101	Monday	3	1	2011	d_3	NaN	NaN	NaN	NaN	0	0	0
2011-02-01	11101	Tuesday	4	2	2011	d_4	NaN	NaN	NaN	NaN	1	1	0
2011-02-02	11101	Wednesday	5	2	2011	d_5	NaN	NaN	NaN	NaN	1	0	1

- sell_prices.csv:

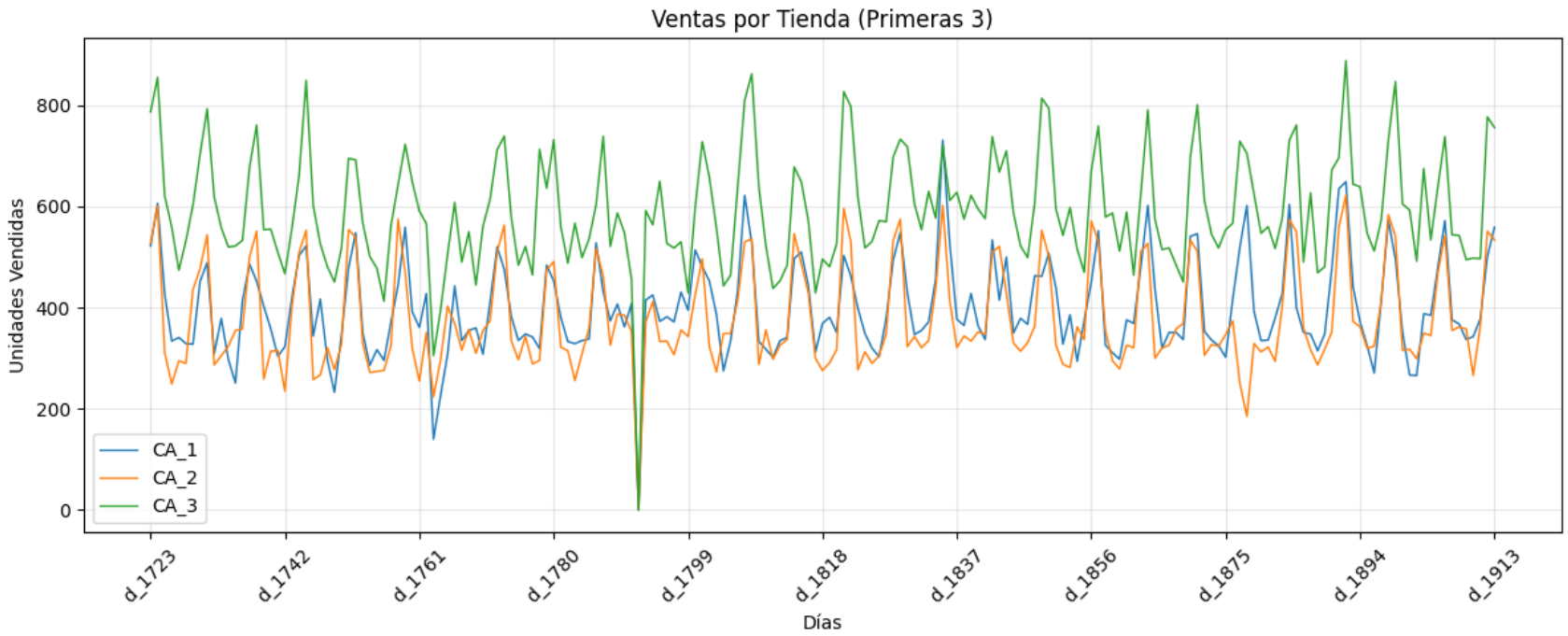
store_id	item_id	wm_yr_wk	sell_price
CA_1	HOBBIES_1_001	11325	9.58
CA_1	HOBBIES_1_001	11326	9.58
CA_1	HOBBIES_1_001	11327	8.26
CA_1	HOBBIES_1_001	11328	8.26
CA_1	HOBBIES_1_001	11329	8.26

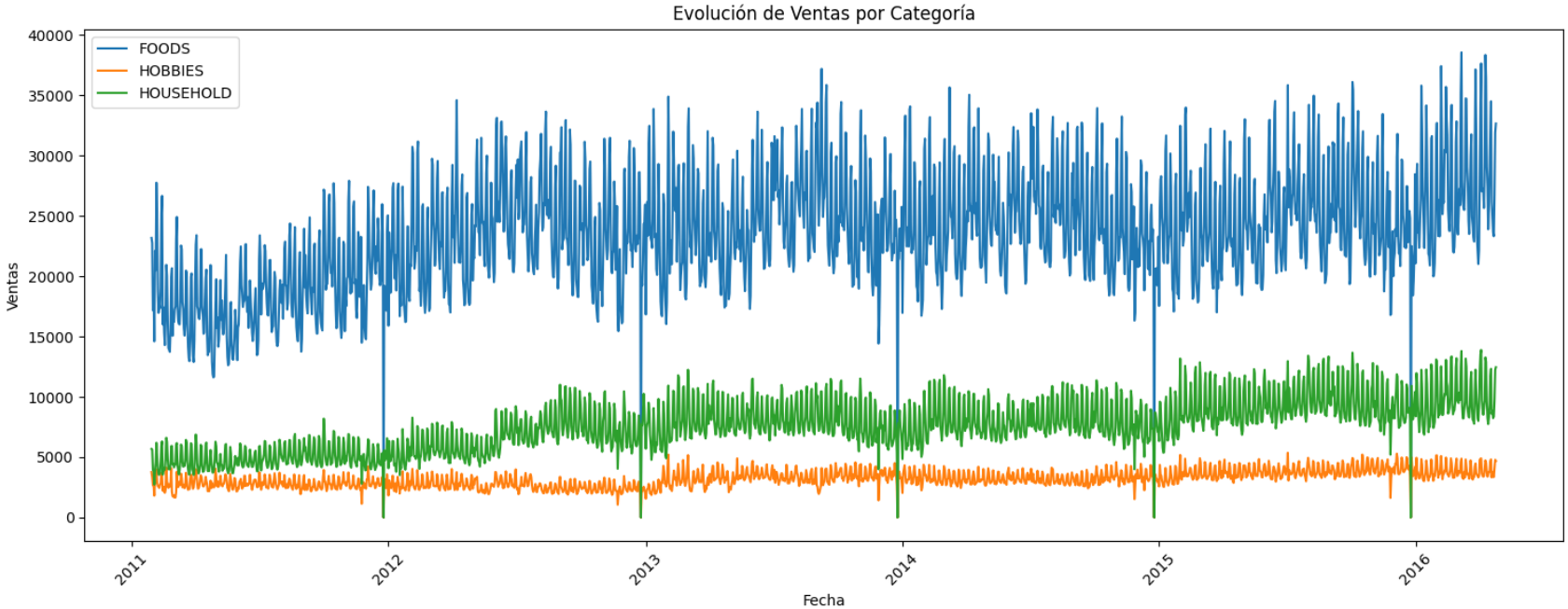
EDA

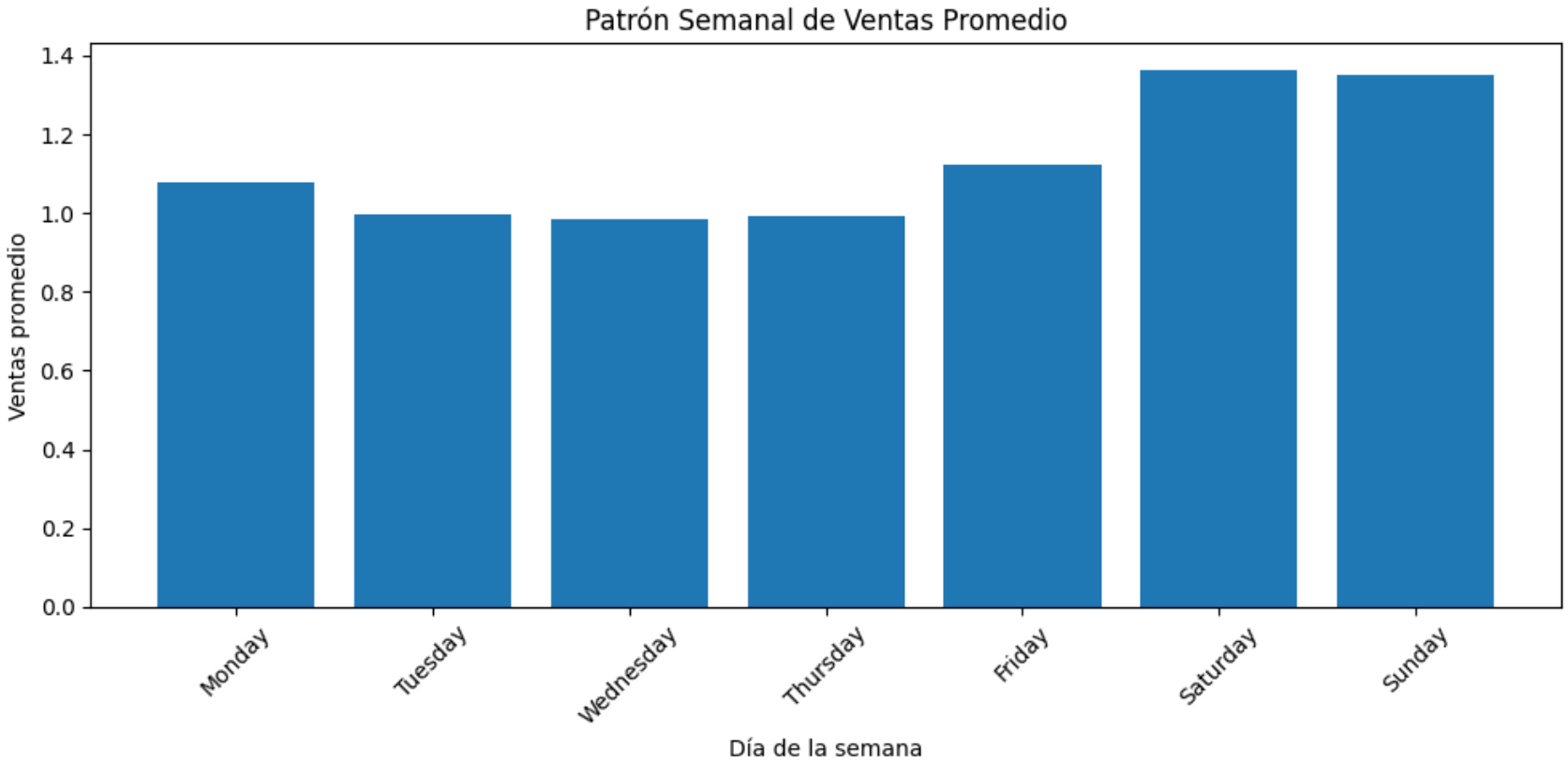


- 3049 productos distribuidos en todas las tiendas.
- 1969 días de registros de venta.

- Sólo tenemos datos faltantes en el archivo 'calendar.csv'.
- En un 68.20 % de los días no se presentan ventas.
- Hay 162 eventos registrados.


Ventas por Tienda (Primeras 3)
Unidades Vendidas
0
200
400
600
800
CA_1
CA_2
CA_3
d_1723
d_1742
d_1761
d_1780
d_1799
d_1818
d_1837
d_1856
d_1875
d_1894
d_1913
Días

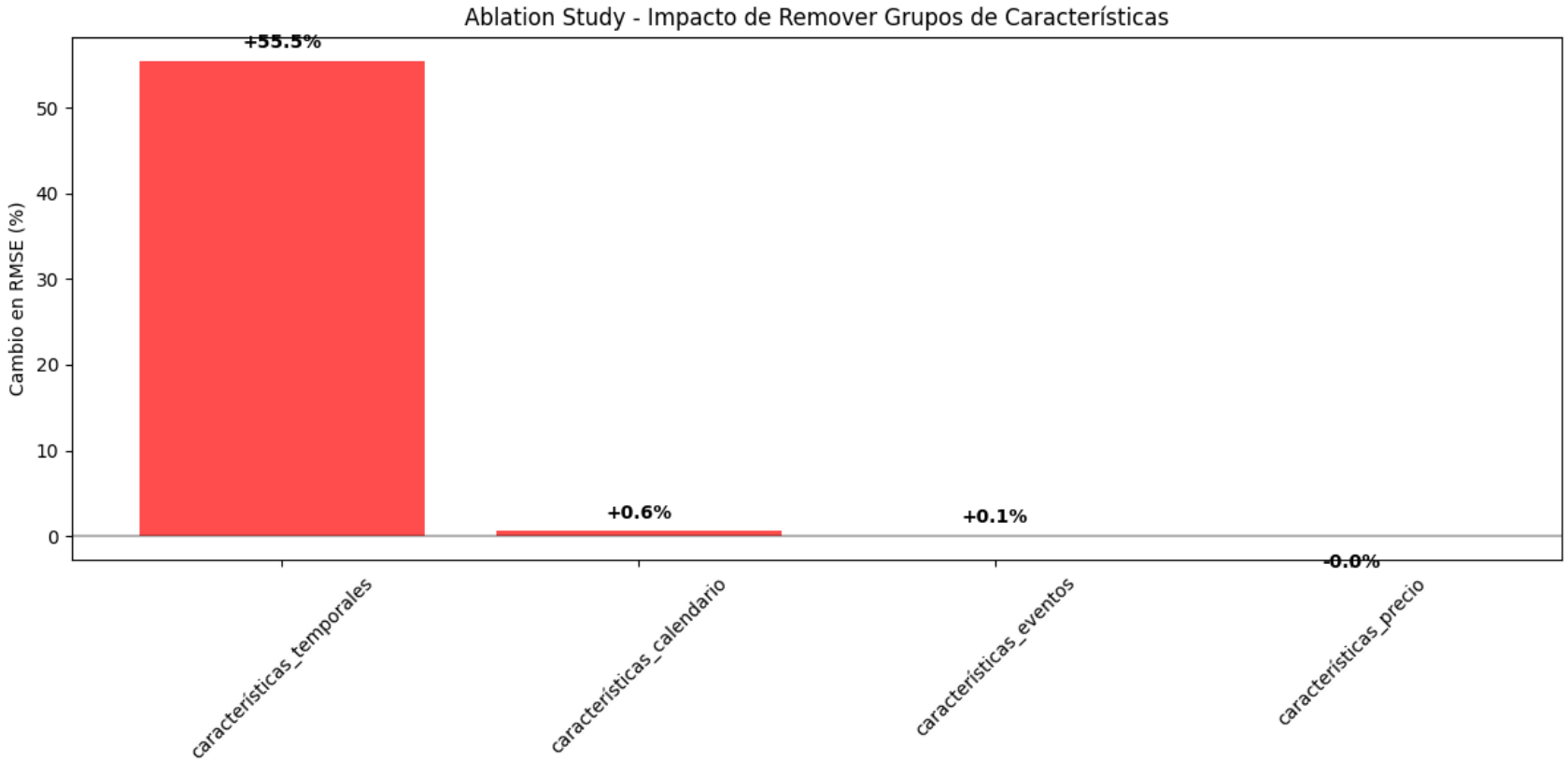

Evolución de Ventas por Categoría
FOODS
HOBBIES
HOUSEHOLD
Ventas
Fecha
40000
35000
30000
25000
20000
15000
10000
5000
0
2011
2012
2013
2014
2015
2016


Patrón Semanal de Ventas Promedio
Ventas promedio
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
Monday
Tuesday
Wednesday
Thursday
Friday
Saturday
Sunday
Día de la semana

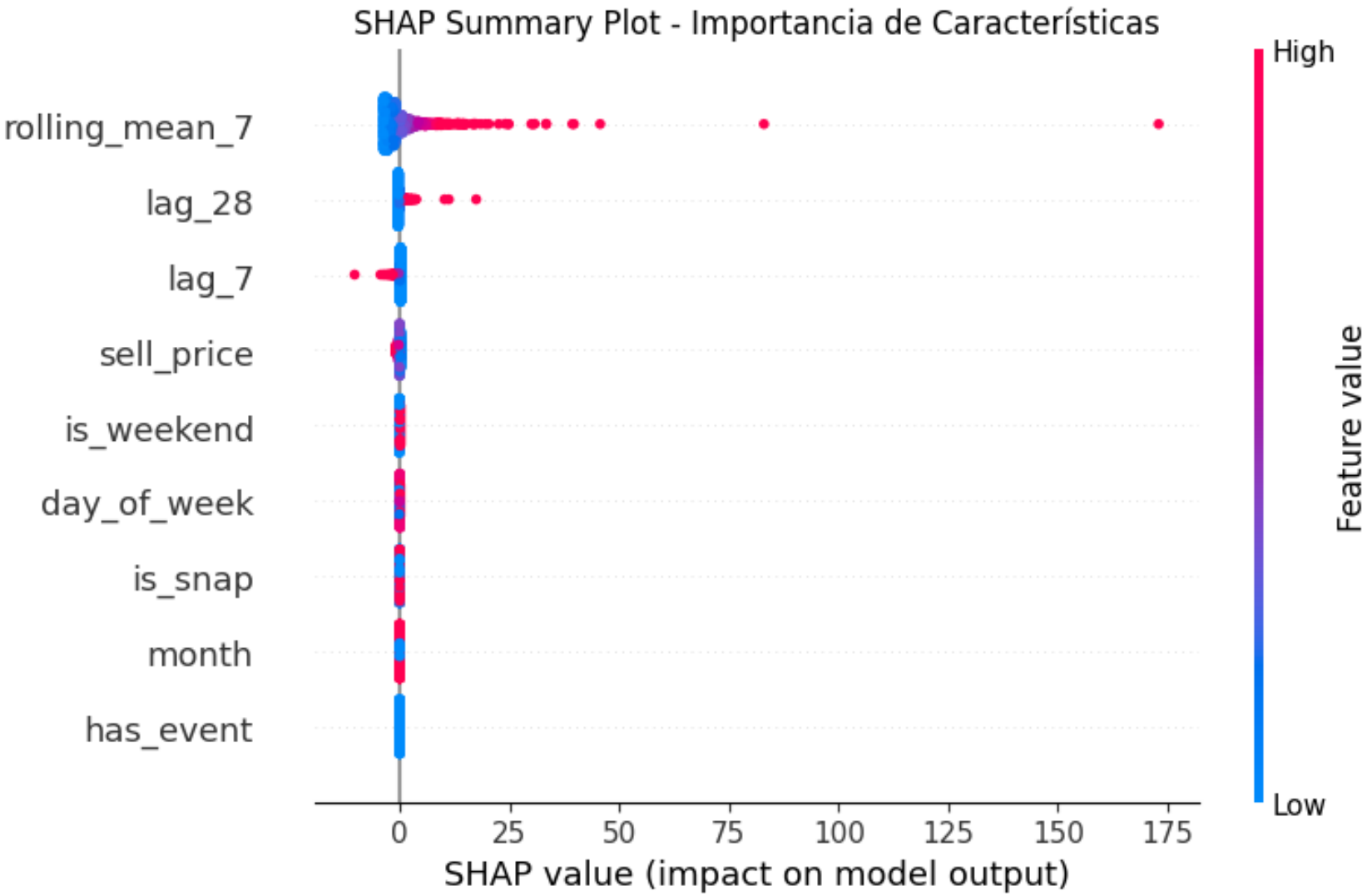
Modelos

1) Naive Estacional
2) Random Forest
3) LightGBM
4) Ridge
5) ARIMA
6) SARIMA
7) Suavizamiento Exponencial (ETS)

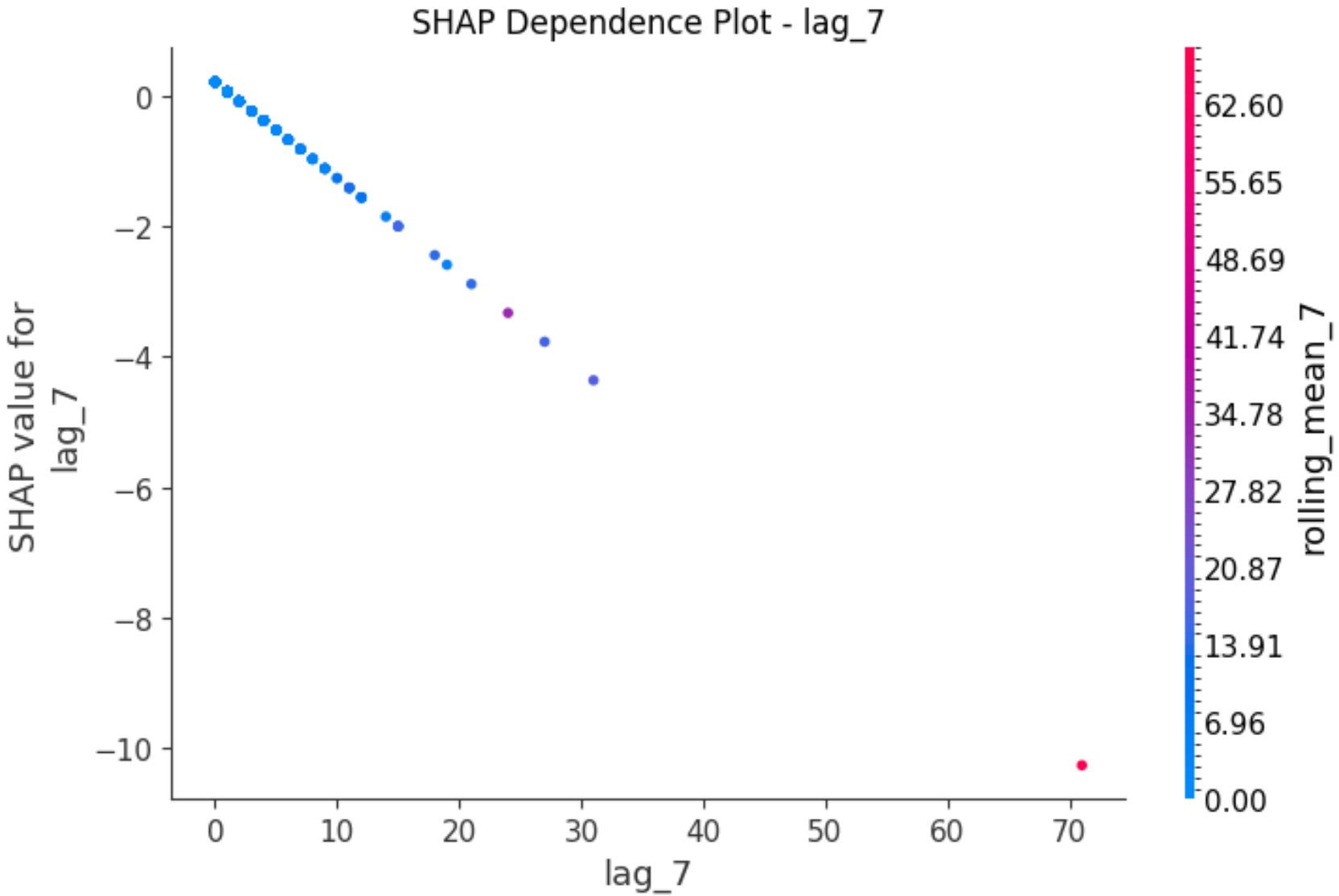
Impacto de características



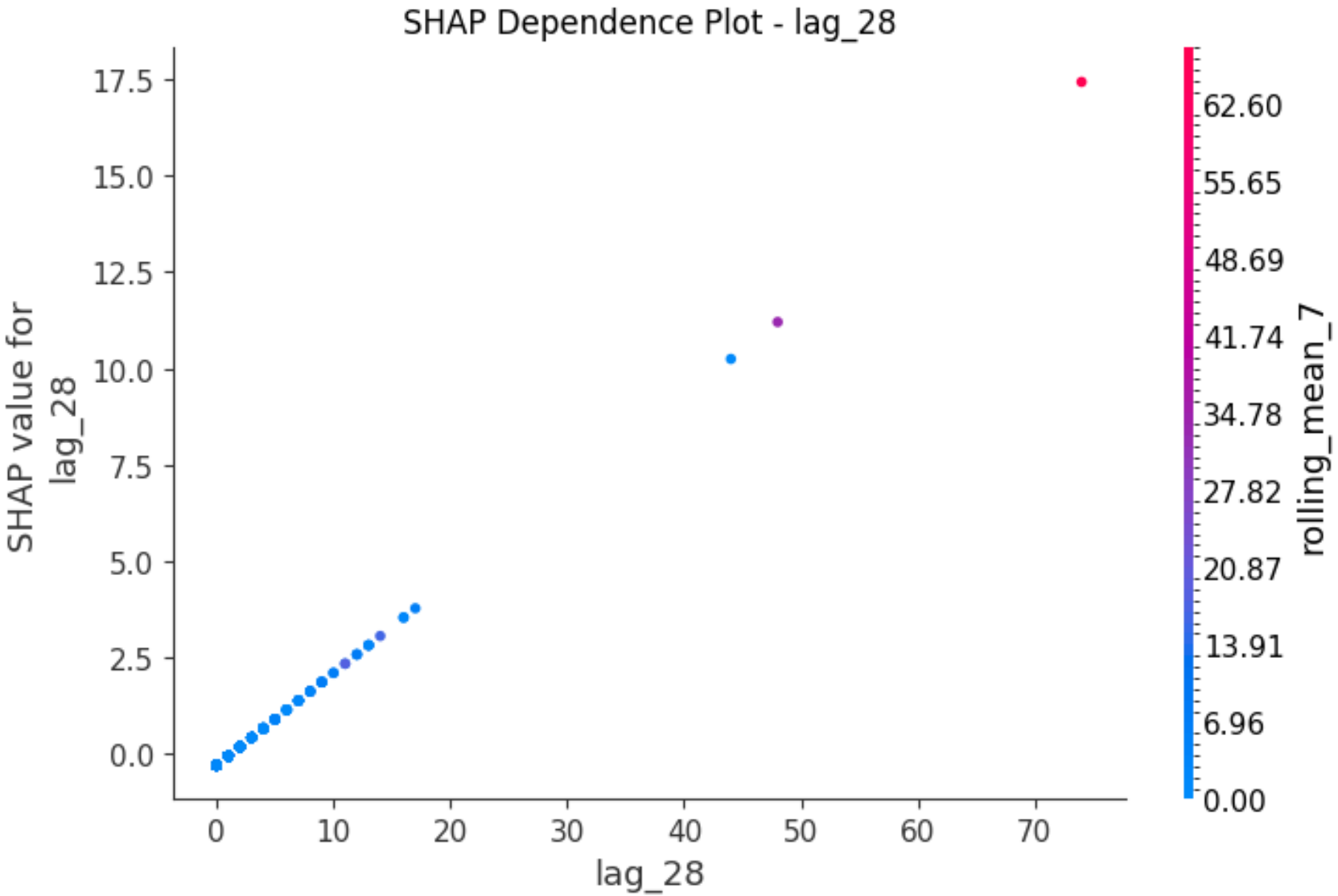
Grados de características



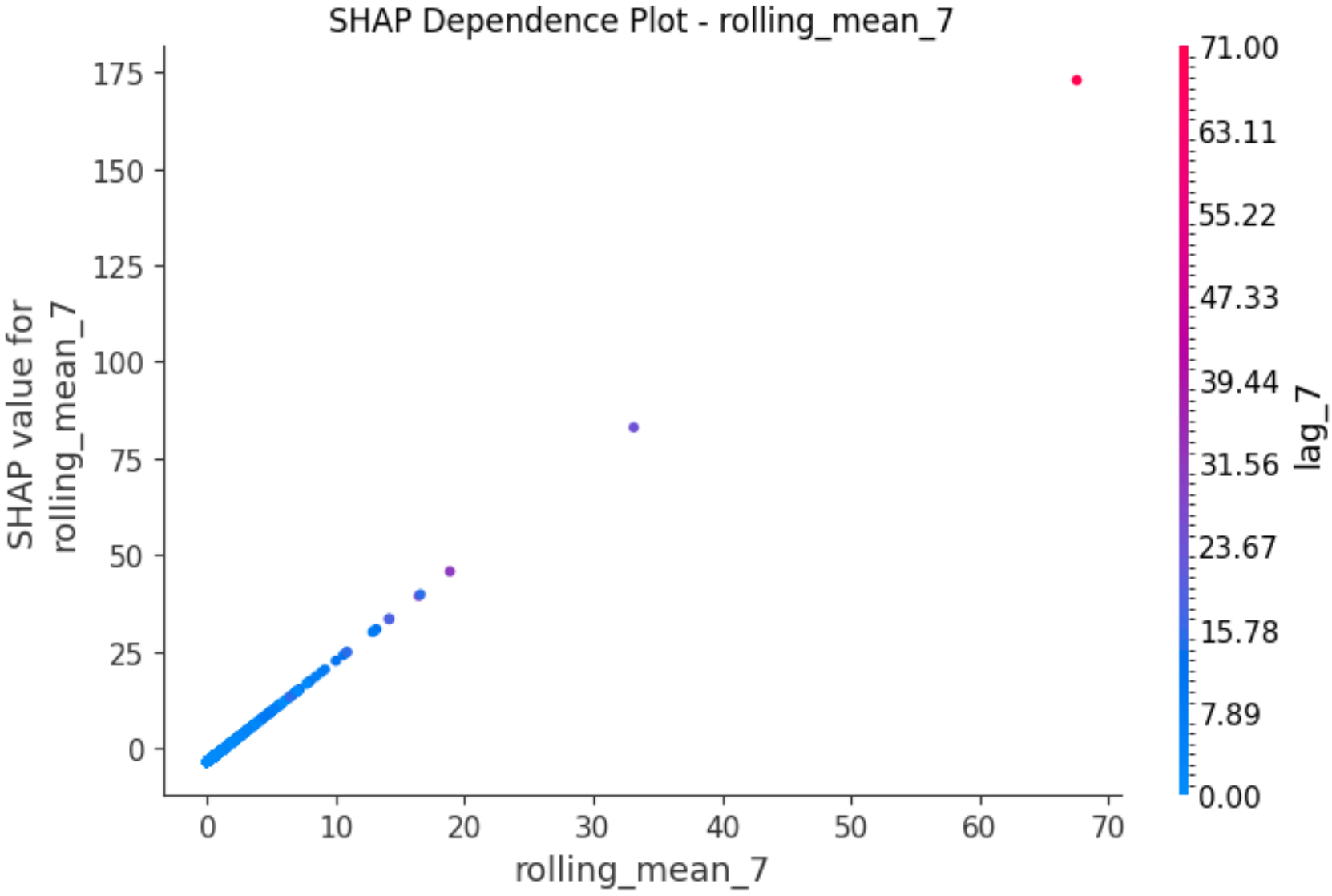
Impacto ventas a 7 días



Impacto ventas a 28 días

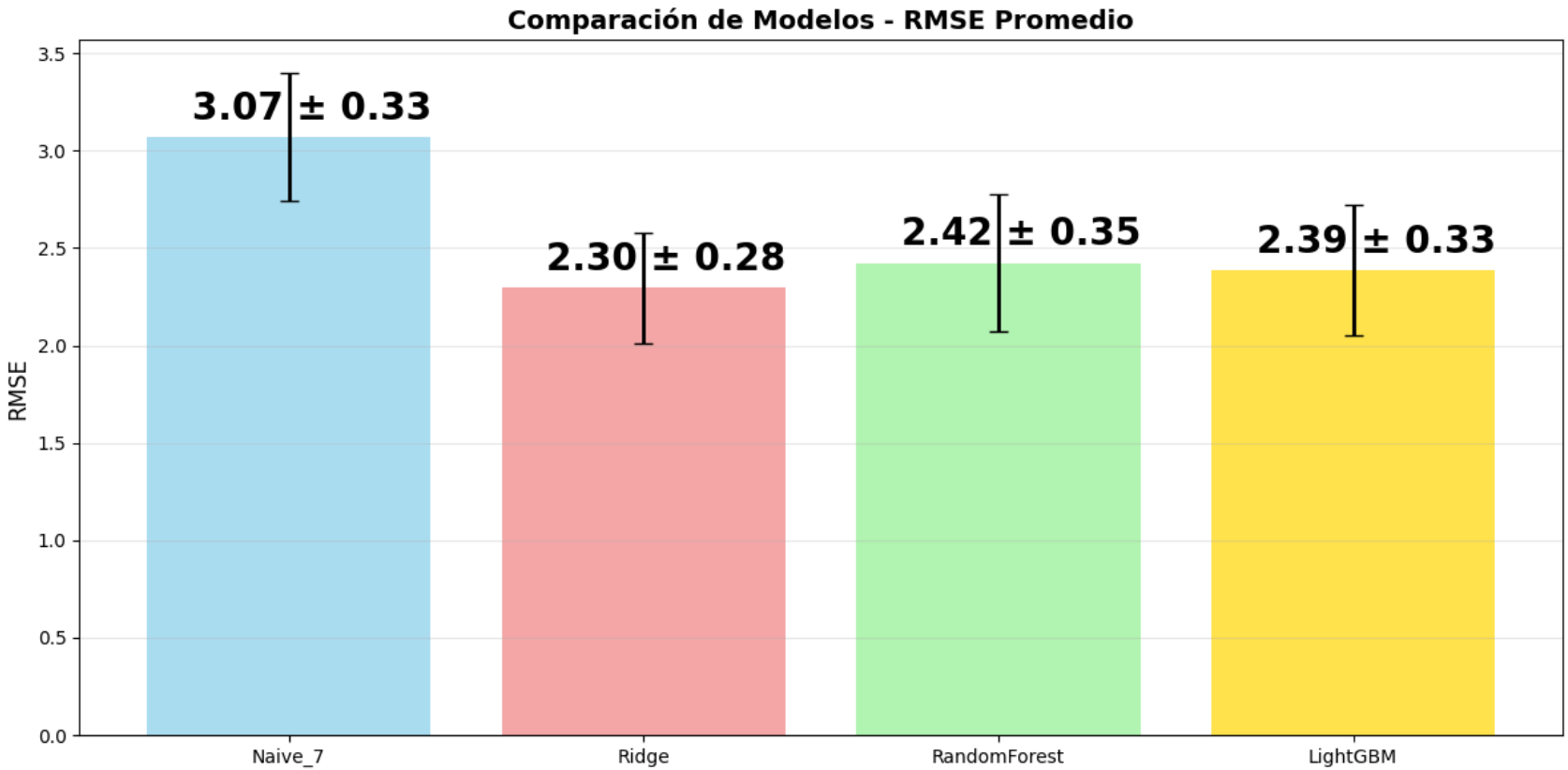


Impacto media móvil a 7 días

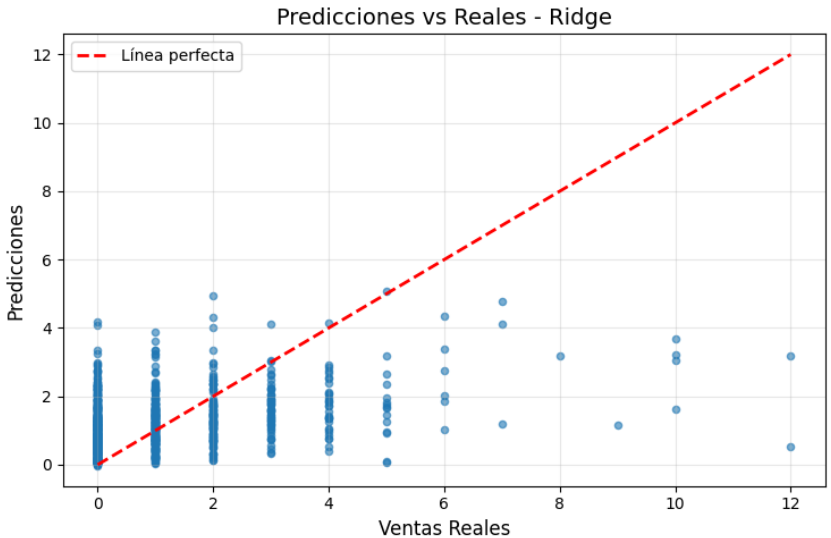


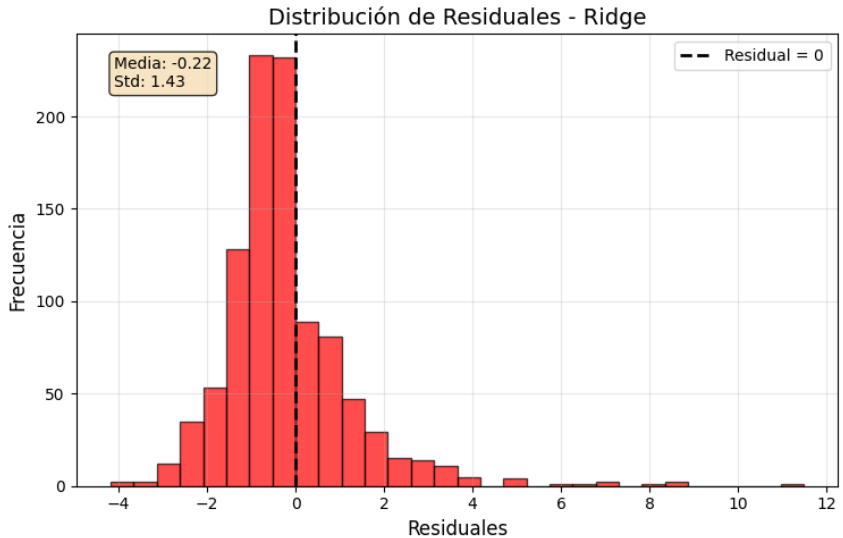
Modelos ML

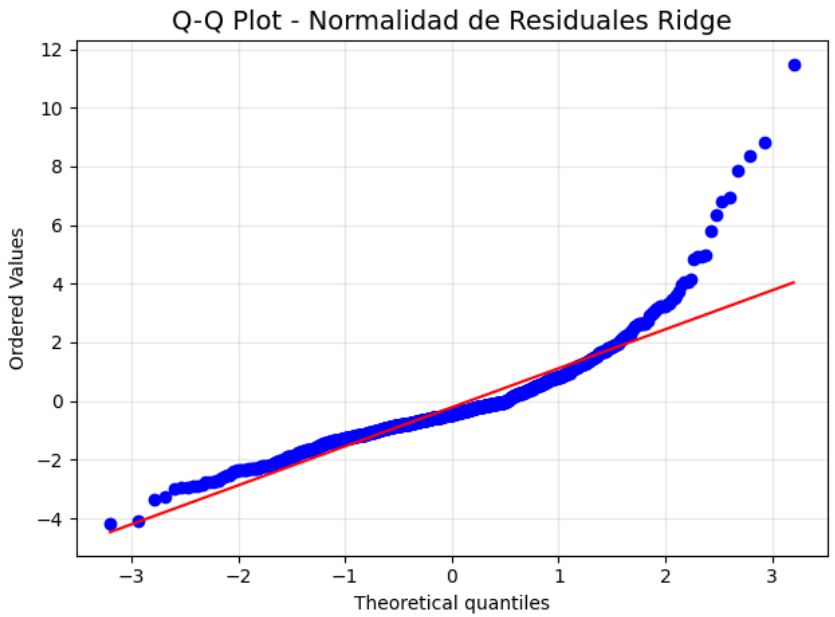
- Alrededor de 91.000 observaciones (40 % de los datos).
- 57 % son ceros.



Análisis Ridge

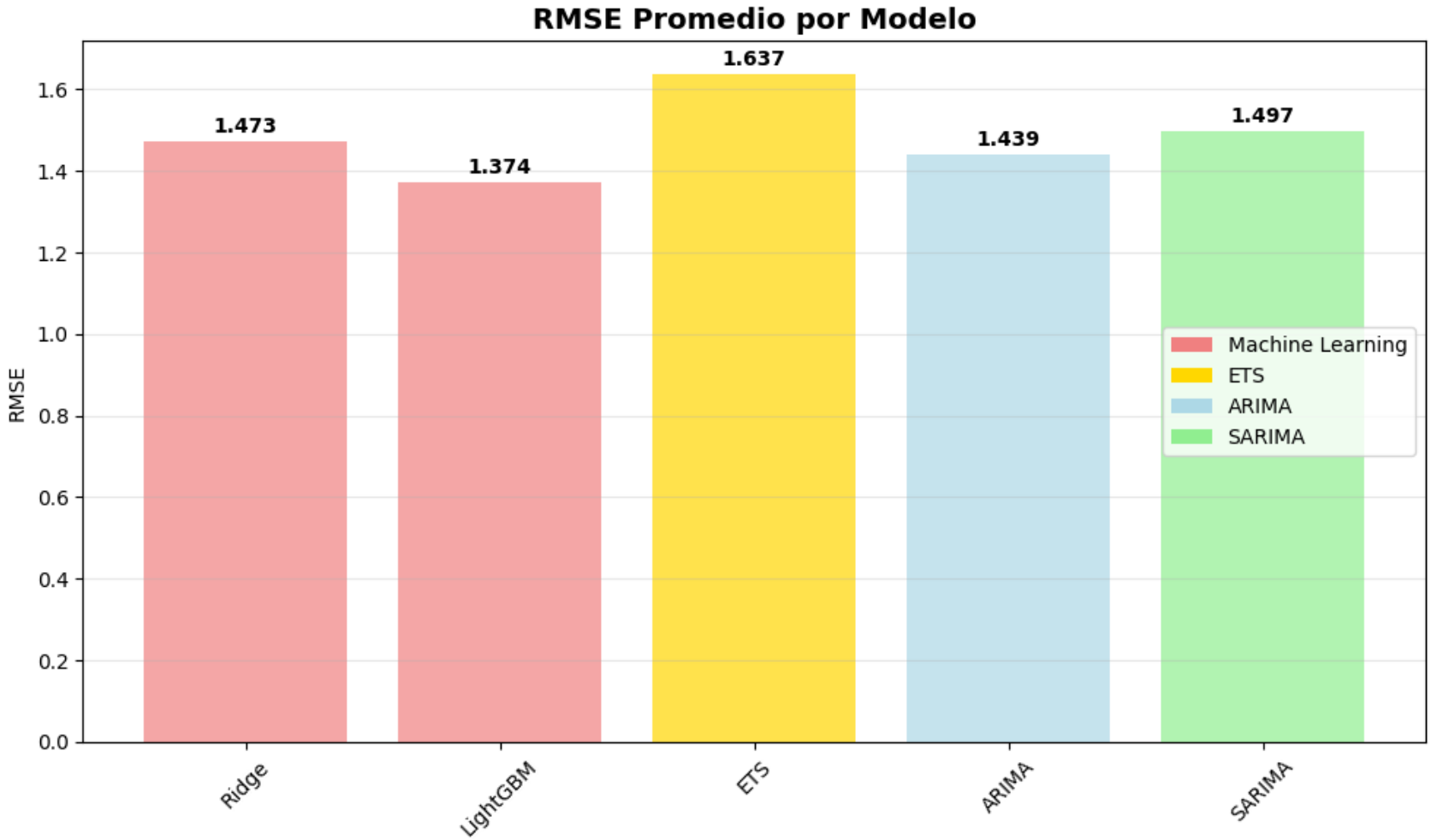




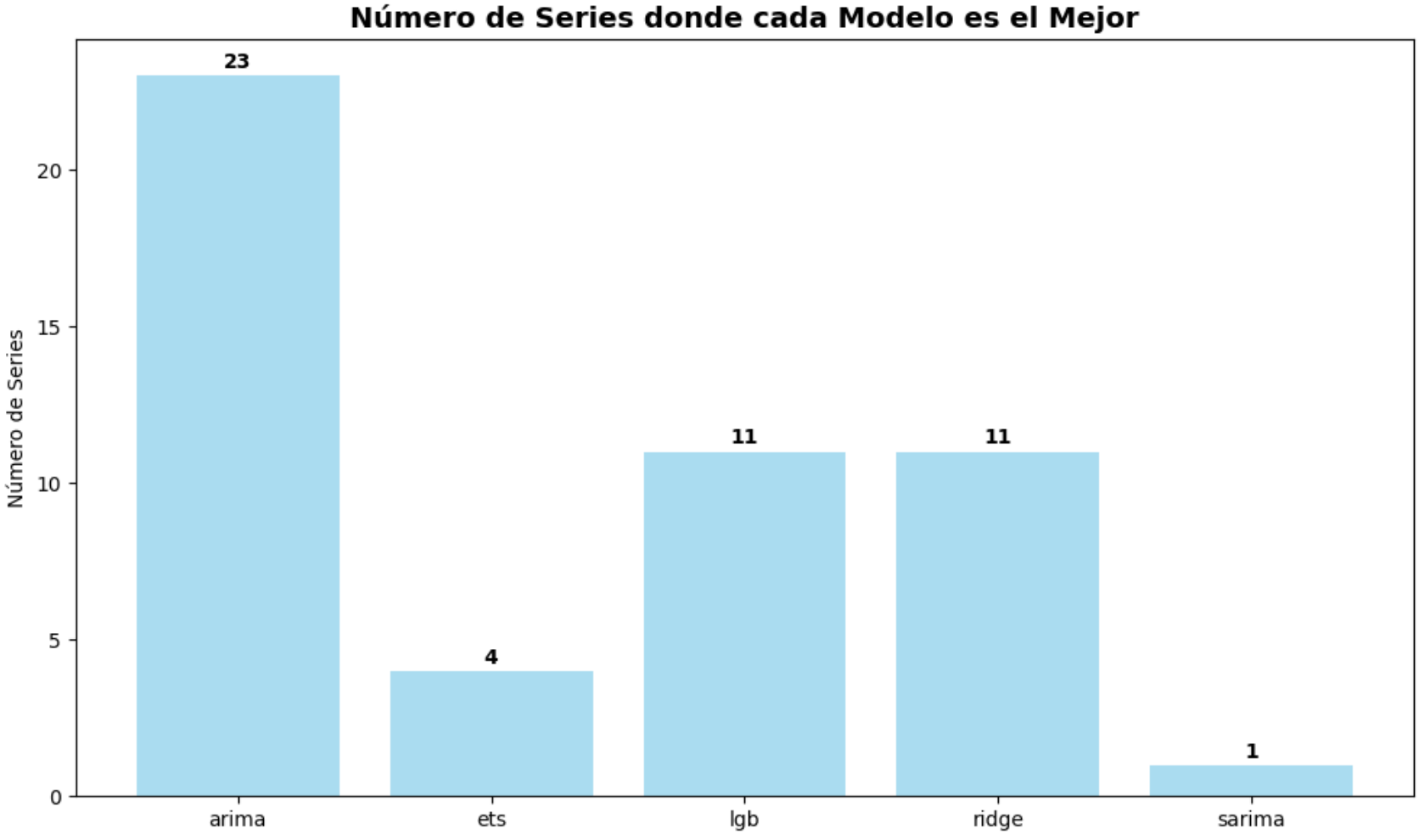


Comparación con Series de Tiempo

- Uso de 20 y 50 series muestreadas.
- Intermitentes (muchos ceros), estables y mixtas ($\approx$ 50 % ceros)



Comparación final



- 11 Intermitentes, 9 Mixtas y 3 Estables.

Parámetros ARIMA

(p,q,d)	Número de victorias	Porcentaje
(0, 0, 0)	14	(60.9 %)
(0, 0, 1)	2	(8.7 %)
(1, 0, 1)	2	(8.7 %)
(0, 1, 1)	1	(4.3 %)
(3, 1, 2)	1	(4.3 %)
(1, 1, 2)	1	(4.3 %)
(4, 0, 1)	1	(4.3 %)
(3, 0, 2)	1	(4.3 %)

Cuadro: Parámetros obtenidos para modelo ARIMA

Limitaciones

- Muchos datos y poca capacidad computacional.
- Serie muy estacional, lo que no se refleja con modelos ARIMA.
- Datos faltantes en calendario.

Trabajo a futuro

- Definir eventos como temporadas en lugar de días.
- Incluir promociones, niveles de inventario, información de competidores.

Conclusión

- El mejor desempeño correspondió a modelos ARIMA, donde la configuración (0,0,0) obtuvo la mayoría de las victorias en las series muestreadas.
- El problema es altamente intermitente y heterogéneo ($\approx$ 57–68 % ceros por serie), por lo que los modelos ML (p. ej. Ridge/LightGBM) siguen siendo útiles cuando hay patrones marcados y variables externas.

Muchas Gracias