Inspira Crea Transforma

UNIVERSIDAD EAFIT

Estimación de ventas diarias para las categorías Hobbies y Foods en las tiendas de Walmart ubicadas en el estado de California

Proyecto Integrador para las asignaturas

Métodos estadísticos avanzada, Aprendizaje automático avanzado, Visualización de los datos

Alba M. Gómez, Angie K Barrera, Francisco J. Moya, José J. Muñoz, Juan Felipe Burbano

Maestría en Ciencia de los Datos y Analítica

Noviembre 2022

Contenido

1. Reto de analítica
2. Objetivos
3. Diseño Metodológico
4. Ciclo de trabajo CRISP-DM
5. Conclusiones
6. Trabajos Futuros

1.Reto de analítica

¿Es posible crear un modelo de Machine Learning o Deep Learning para predecir las ventas diarias que supere el modelo estadístico clásico (SARIMA)?

2.Objetivos

Objetivo general:

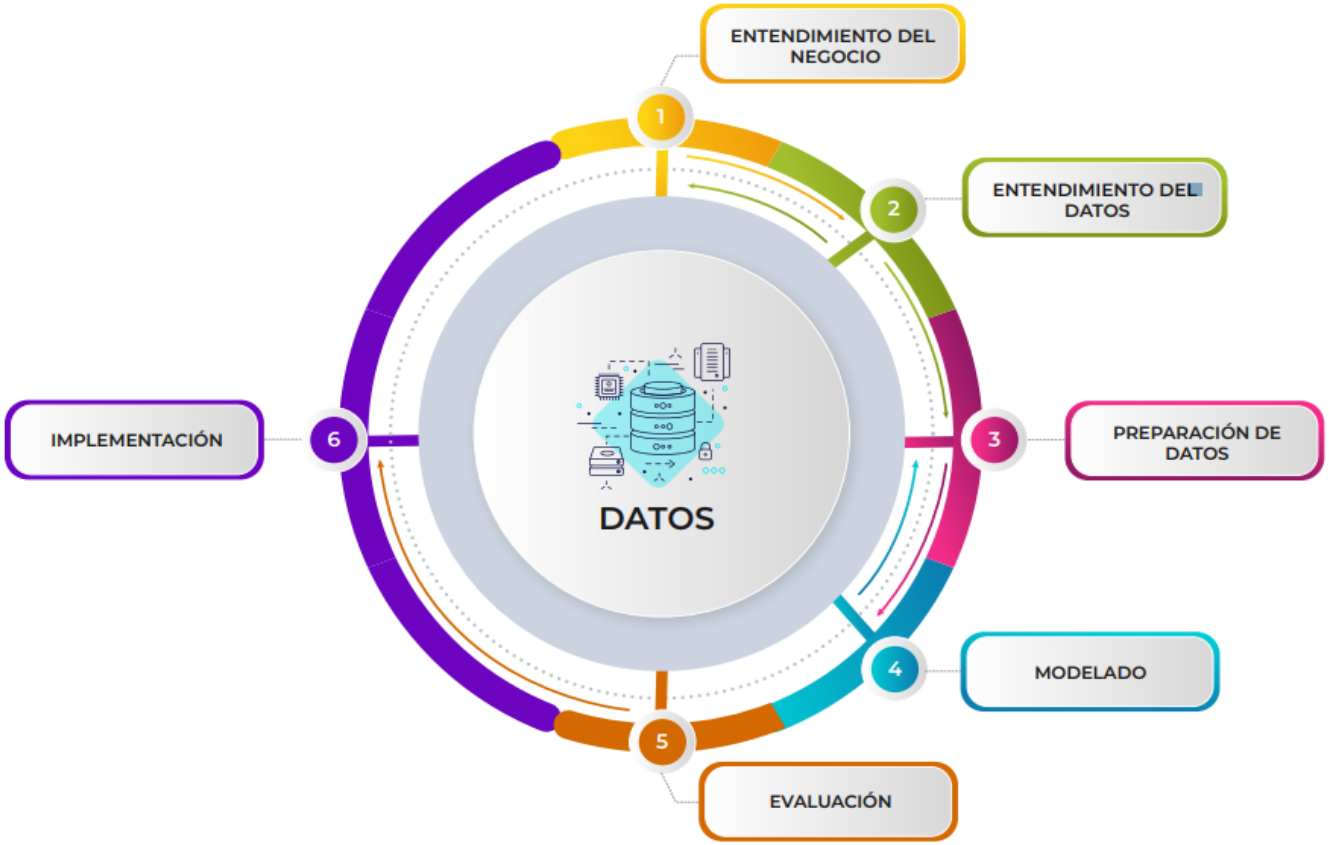
Comparar técnicas de predicción con series de tiempo para pronosticar las ventas en dólares de las tiendas Walmart en el Estado de California, Estados Unidos

Objetivos específicos:

- Explorar y analizar el conjunto de datos
- Establecer las técnicas de analítica disponibles para predicción de series de tiempo.
- Preprocesar y preparar el conjunto de datos
- Construir los modelos de serie de tiempo
- Evaluar los resultados de los modelos propuestos en función de sus métricas de calidad

3.Diseño Metodológico

Ilustración 2. Fases empleadas en la metodología CRISP-DM



Fuente: Adaptado de (IBM, 2022).

1.Entendimiento del negocio

Corporación multinacional Estadounidense de tiendas retail

Relevancia

- 11.000 tiendas a nivel mundial
- Presencia en 28 países
- E-commerce en 11 países
- 3 Estados de USA: California, Wisconsin, Texas

Necesidad

- Pronóstico de ventas

Caso de Estudio

Pronóstico de ventas para el estado de california en las 2 categorías : Foods y Hobbies

2.Entendimiento de los datos

Fuentes	Variables	Cantidad
Calendar	Fechas de eventos relevantes	30 eventos especiales
Calendar	Fechas de Ventas	29/01/2011 – 22/05/2016
Sales	Código de Fechas	Un identificador de días del año 365 por año
Sales	Unidades vendidas diariamente por Categoría, tienda, producto	3 Estados, 3 tiendas por estado. 23.672.436 registros
Sells Prices	Precio de Venta Unitario	Mas de 5000 referencias de productos

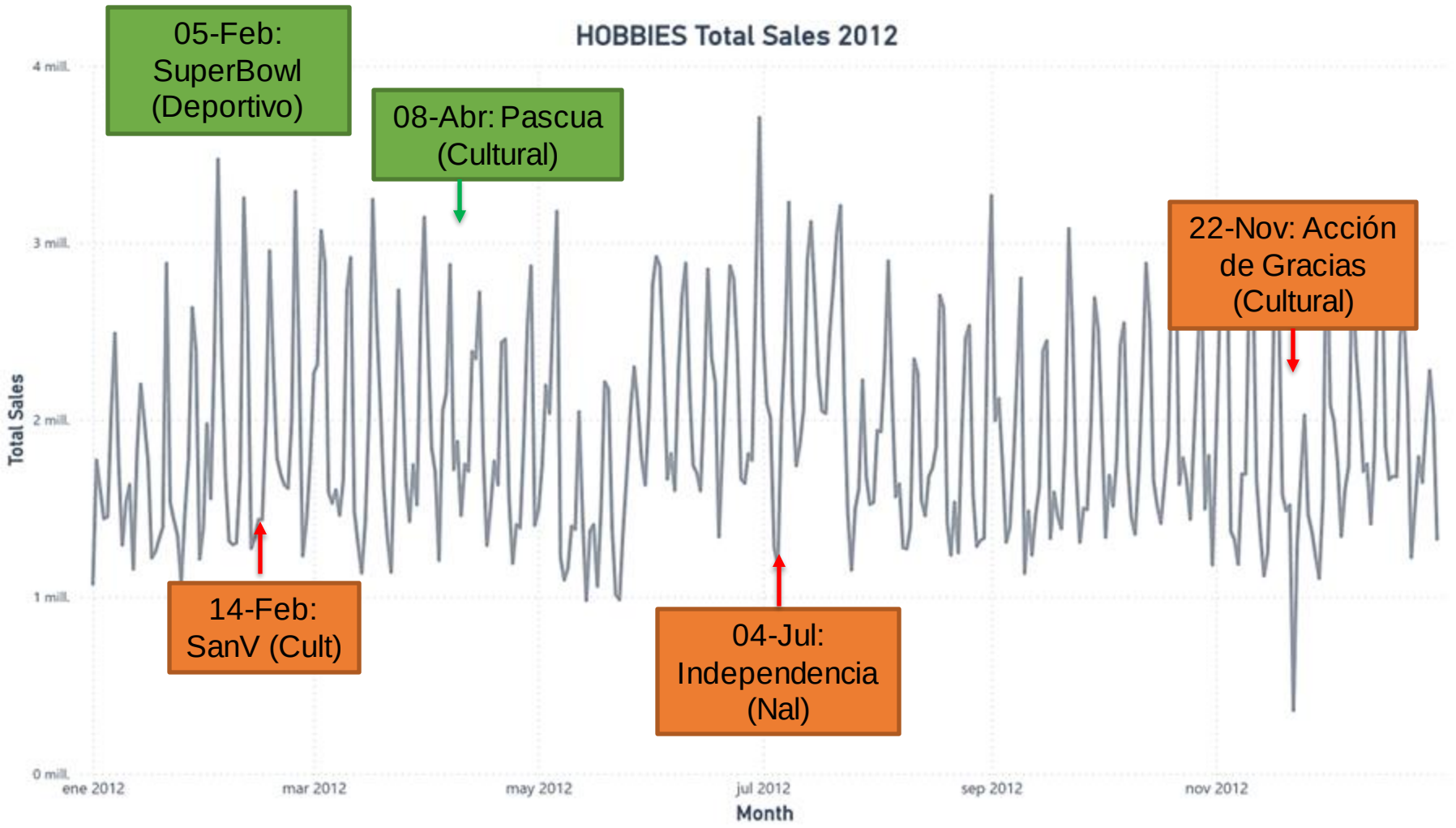
Kaggle: "Pronóstico M5 – Precisión: Estime las ventas unitarias de los productos minoristas de Walmart", que puede ser consultada en la siguiente url https://www.kaggle.com/competitions/m5-forecasting-accuracy/overview.

2.Entendimiento de los datos

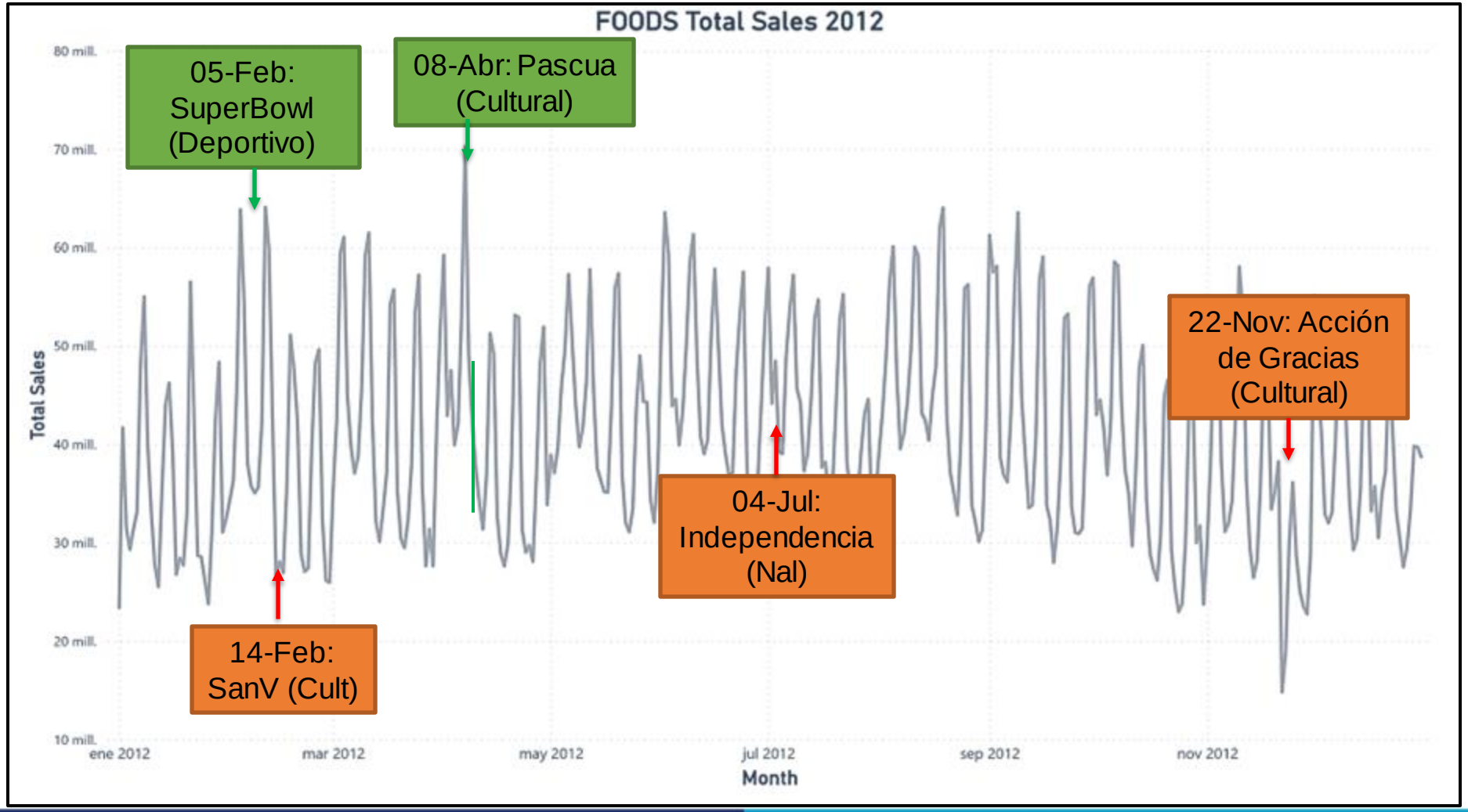
Fuentes	Variables	Cantidad
Ventas Categorías Foods, Hobbies en California	Date	29/01/2011-22/25/2016
	Total_sells	1941 registros por categoría

Kaggle: "Pronóstico M5 – Precisión: Estime las ventas unitarias de los productos minoristas de Walmart", que puede ser consultada en la siguiente url https://www.kaggle.com/competitions/m5-forecasting-accuracy/overview.

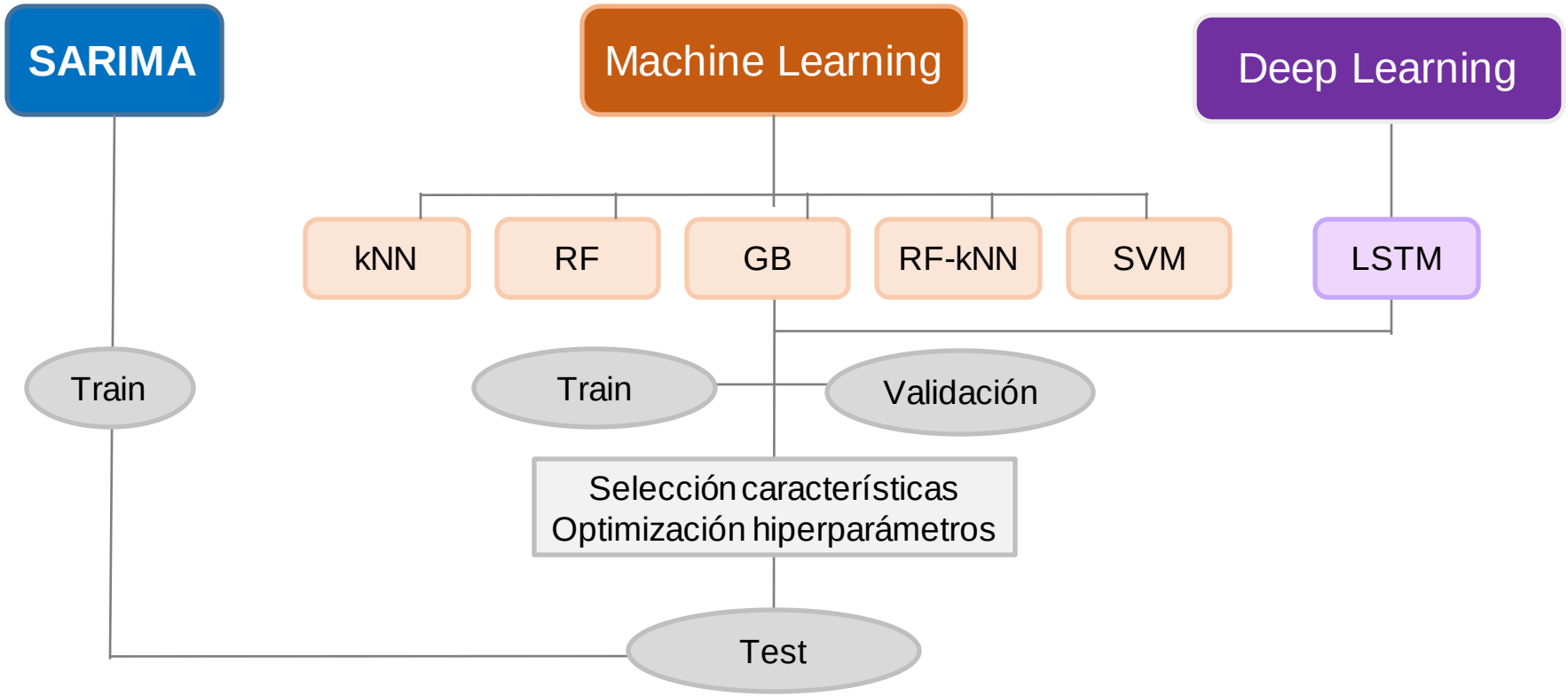
2.Entendimiento de los datos



2.Entendimiento de los datos



3. Preparación de los datos , 4.modelos, 5.evaluación



Modelos SARIMA

3.Preparación de los datos

Análisis de estacionariedad

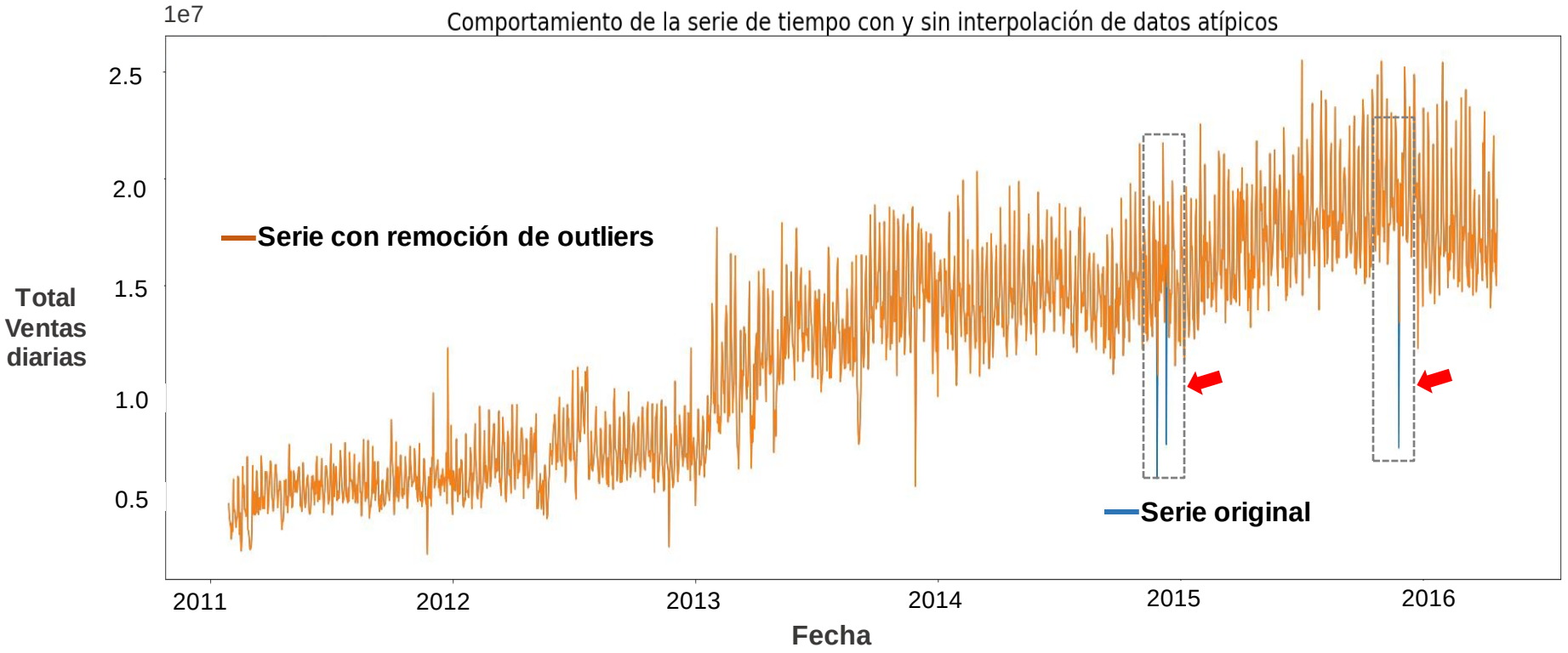
- Análisis gráfico
- Descomposición de la serie
- Tratamiento de Outliers
- Transformaciones a la serie
- Prueba de raíz unitaria

- Correlogramas

Orden de la serie

Modelos SARIMA

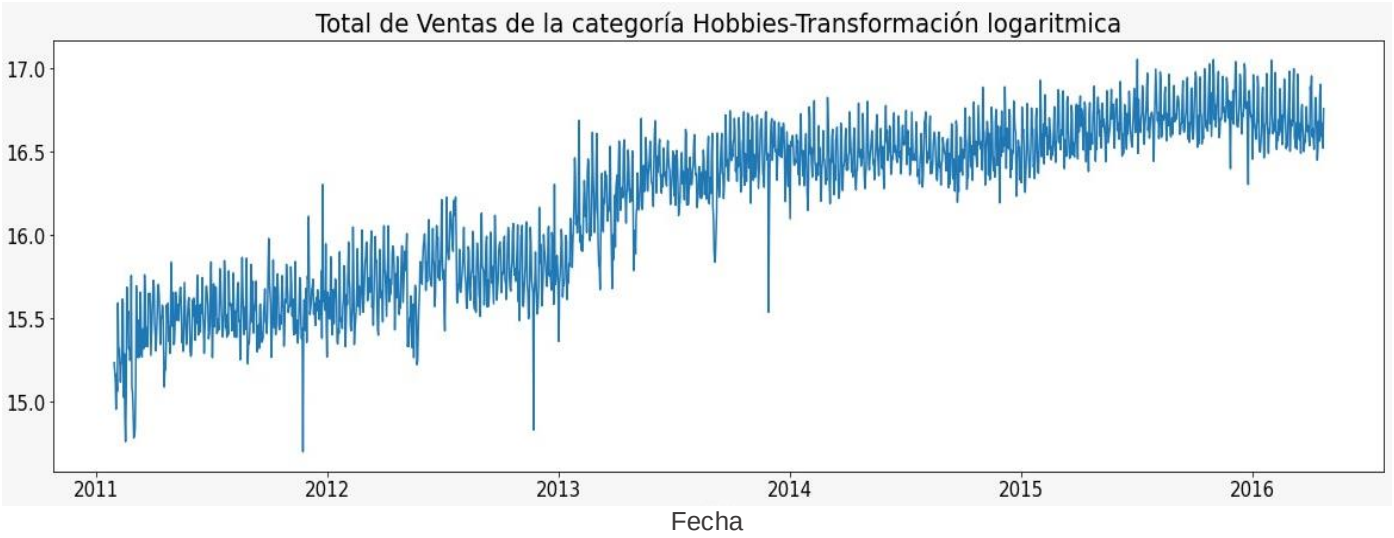
3.Preparación de los datos



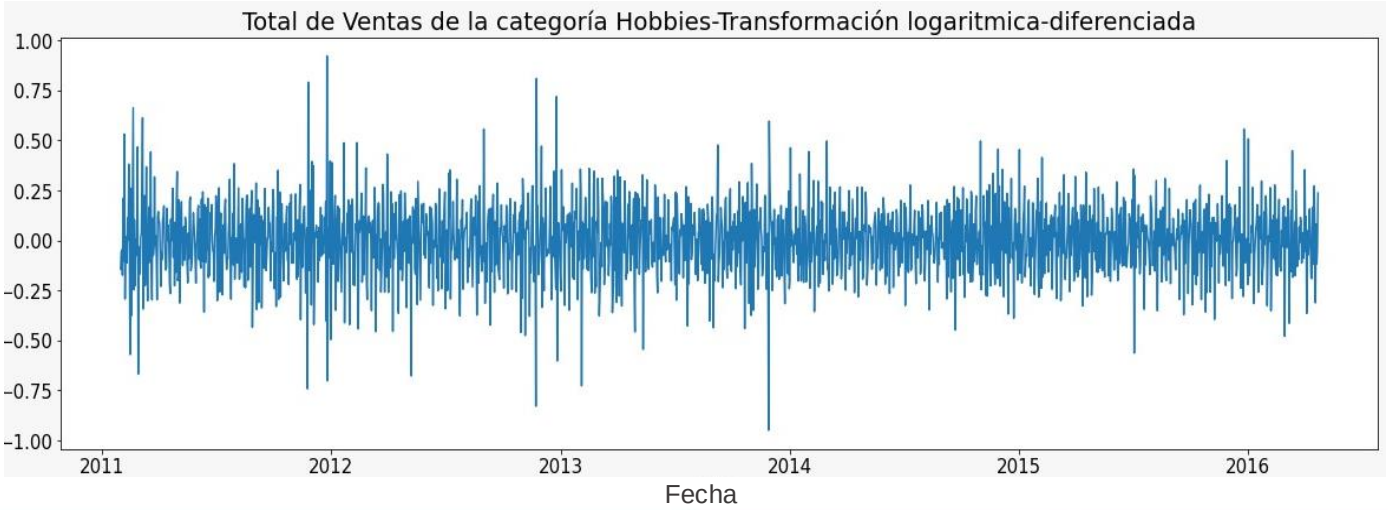
Modelos SARIMA

3.Preparación de los datos

Log (Total Ventas diarias)



Dif(Log (Total Ventas diarias))



Modelos SARIMA

3.Preparación de los datos

Test de Dickey – Fuller

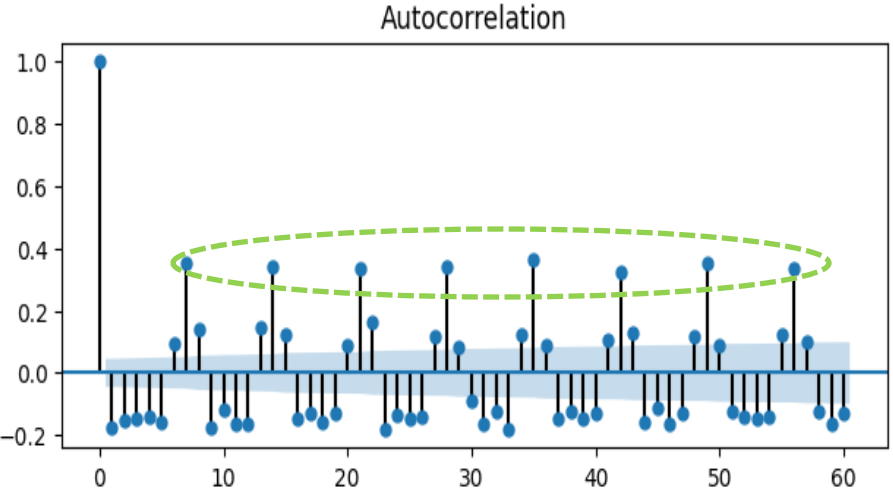
Hipótesis nula: La media de los datos es no estacionaria

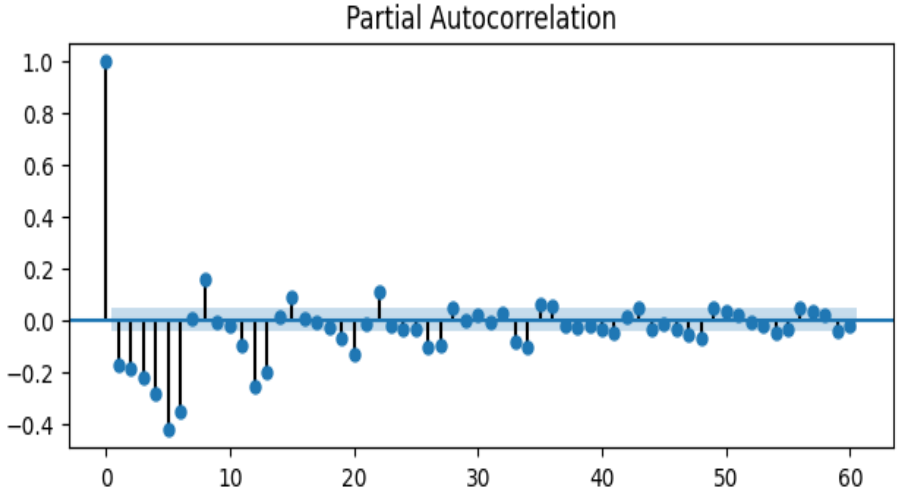
Hipótesis alternativa: La media de los datos es estacionaria

Valor p

4.3599e-26

Correlogramas





Modelos SARIMA

4.Modelado

Partición de los datos

1.911
Registros para entrenar

30
Registros para testeo

Modelos propuestos

Modelo 1.

Seleccionado analíticamente

SARIMA (6,1,0)(1,0,0,12)

Modelo 2.

Seleccionado del mejor modelo Autorima

Autoarima (0,1,3)(2,0,2,12)

Modelos SARIMA

4.Modelado

Resultados para la categoría Hobbies

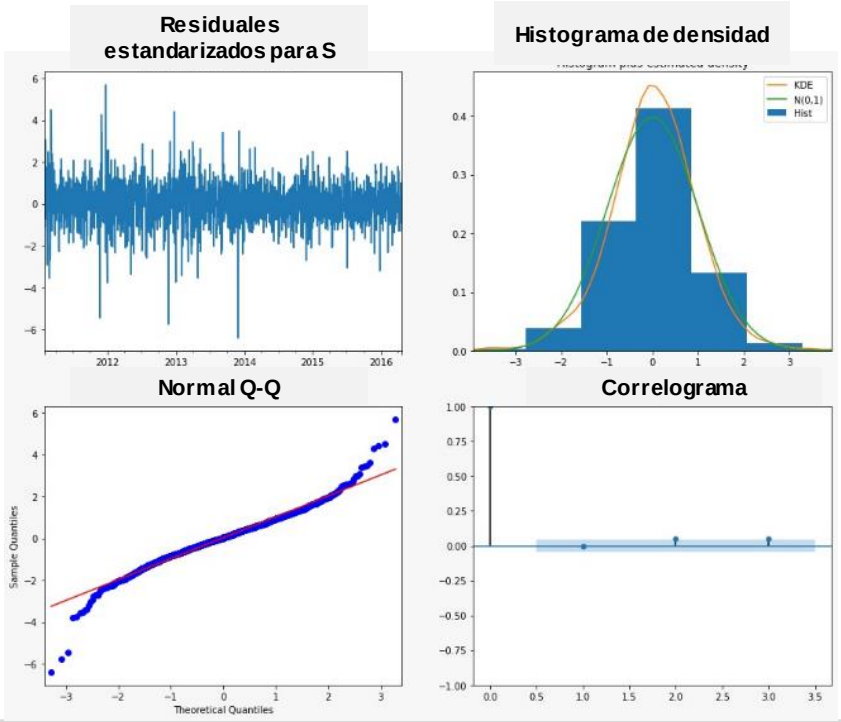
Modelo SARIMA (6,1,0)(1,0,0,12)

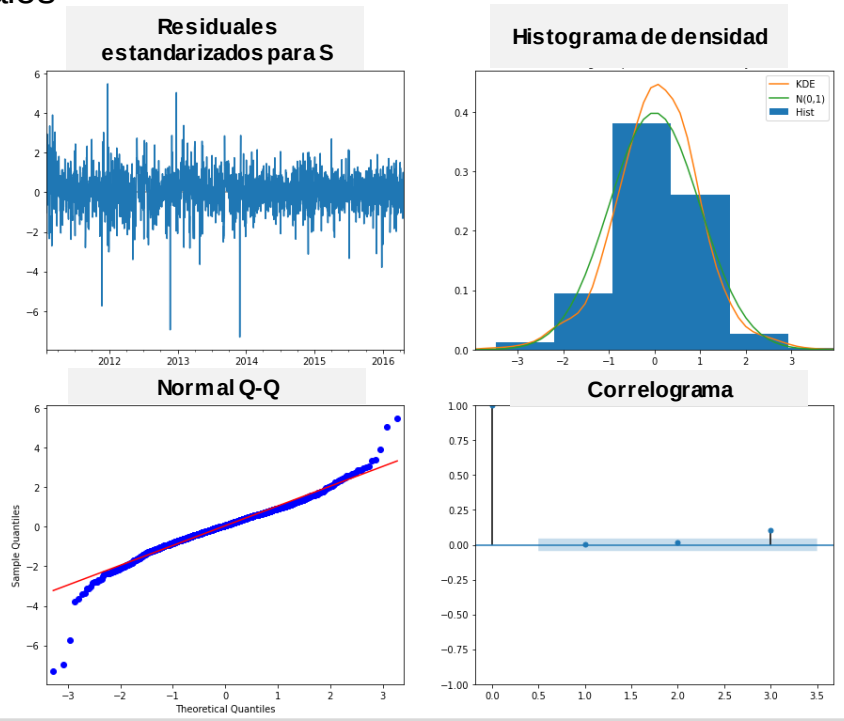
AIC: -2118.6

Modelo Autoarima (0,1,3)(2,0,2,12)

AIC: **-2315.9**

Residuales





Modelos ML

Modelos DL

3.Preparación de los datos

Estructura de datos para modelos ML y DL

Día 1	Día 2	…	Día 58	Día 59	Día 60	Día 61	Día 62	Día 63	…

1 x # Días

Día	D-14	D-9	D-7	D-2	Mediana 60 días	Std 60 días	Curtosis 14 días	Otras
61	Día 47	Día 52	Día 54	Día 59	Mediana(D1-60)	Std(D1-60)	Kurt(D47–60)	…
62	Día 48	Día 53	Día 55	Día 60	Mediana(D2-61)	Std(D2–61)	Kurt(D47–60)	…
63	Día 49	Día 54	Día 56	Día 61	Mediana(D3-62)	Std(D2–61)	Kurt(D47–60)	…
…	…	…	…	…	…	…	…	…

D x (# Días – Rezagos elegidos)

Lubba, C.H., Sethi, S.S., Knaute, P. et al. catch22: CAnonical Time-series CHaracteristics. Data Min Knowl Disc 33, 1821–1852 (2019).

Muestra de 22 características discriminatorias que se pueden probar para modelar la serie de tiempo como una regresión

Modelos ML

3.Preparación de los datos

- Mes y año (variables ordinales) a partir de la fecha

Variable a predecir: ventas en USD → Creación de variables a partir de la fecha → Creación de variables a partir de la serie de ventas

- Valor de las ventas para los rezagos 1 a 30 días
- Estadísticas de medida central de las ventas (de los rezagos 30, 60 y 180): media, máximo, mínimo, desviación estándar, curtosis

Eliminar nulos → Escalamiento de las variables → Selección de Características

Seleccionadas:
Rezagos 1, 14, 21, 24, 28
Media 1M, Std 1M, Std2M, Media 6M

Separar set de entrenamiento, validación y prueba

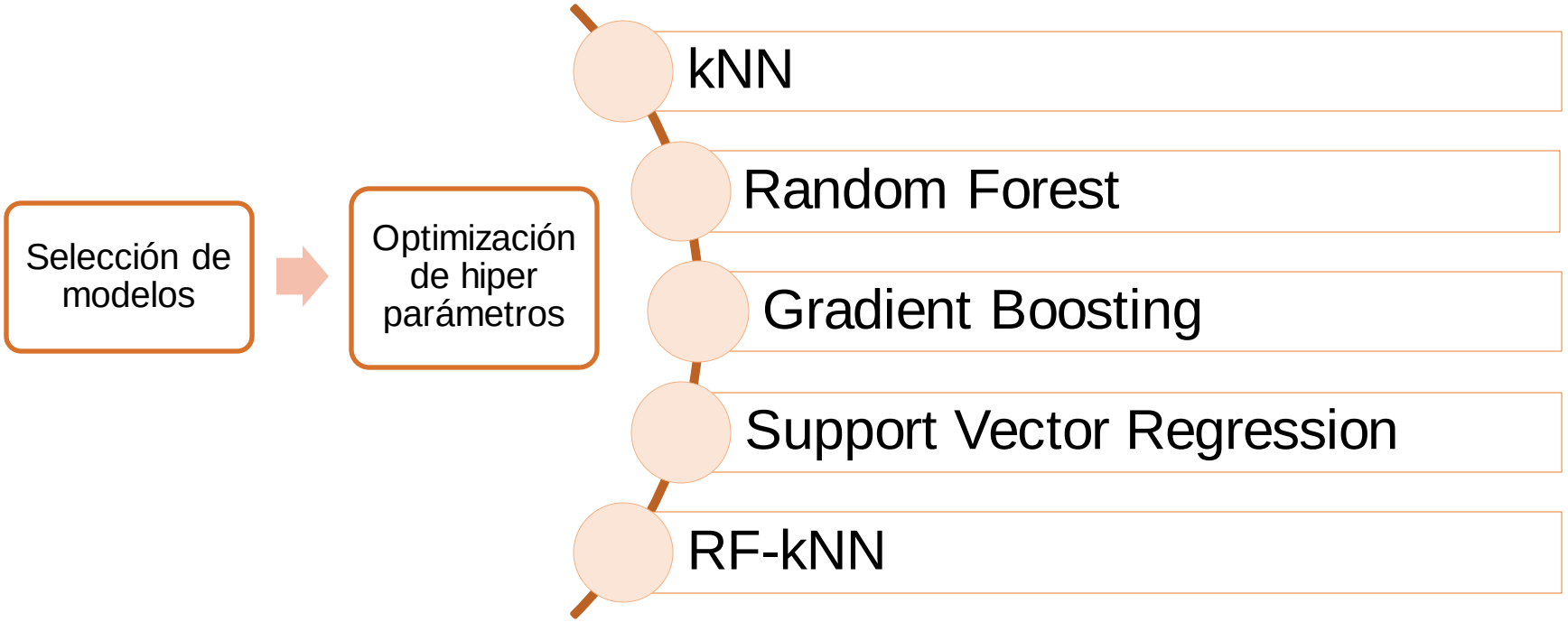
- Coef. corr de Pearson y Spearman
- Random Forest - Importancia
- Selección recursiva de variables
- Selección hacia atrás (Backward)

Modelos ML

4.Modelado

Partición de datos:

- Entrenamiento = 1891 registros consecutivos en días
- Validación = 15 registros consecutivos en días
- Prueba = 30 registros consecutivos en días



Modelos ML

4.Modelado

Resultado de entrenamiento para la categoría Hobbies

Modelos con todas las variables escaladas para categoría Hobbies

Modelo	MAE Train	MAE Validation
kNN	455,056.43	489,077.06
RF	423,522.97	527,205.10
Gradient Boosting	3427,712.86	5,007,366.38
SVR	3,427,703.46	5,007,347.17
RF-kNN	**456,893.1**	**479,575.37**

Modelos DL

4.Modelado

Transformación de datos
- Esquema supervisado
- Estandarización

Separación de datos
- Train/Validation/Test

Entrenamiento y optimización de modelos
- Train/Validation/Test
- Número de épocas

Selección del modelo

Modelo	MAE Train	MAE Validation
LSTM	7,073,574.64	9,484,482.28

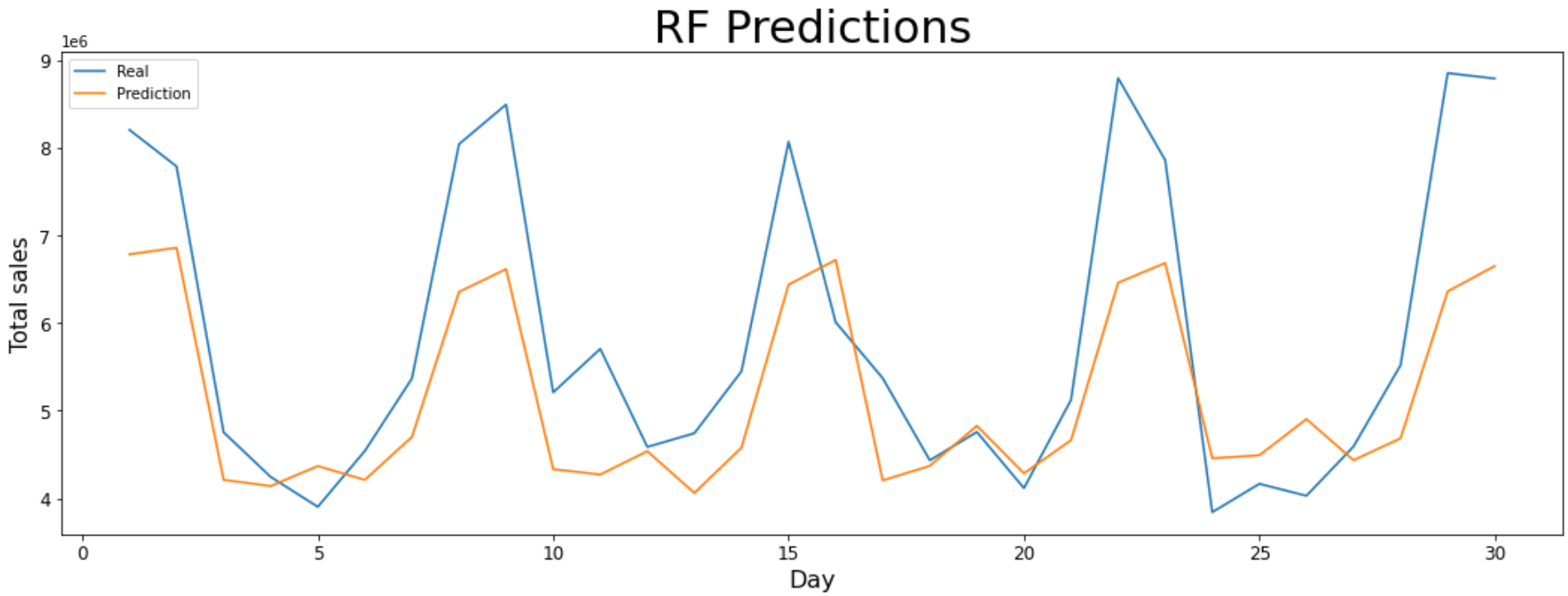
Evaluación

HOBBIES

CATEGORÍA	MODELO	MÉTRICA (MAE TEST) valores en USD
MODELOS ESTADÍSTICOS CLÁSICOS – SARIMA-	Modelo propuesto analíticamente (6,1,0)(1,0,0,12)	6,607,194.96
	Modelo hallado con Autoarima (5,1,2)(0,0,2,12)	7,252,768.01
MODELOS DE MACHINE LEARNING	kNN con variables escaladas	999,664.91
	RF con variables escaladas	**903,844.14**
	Gradient Boosting con variables escaladas	926,338.89
	SVR con variables escaladas	1,073,776.60
	RF-kNN con variables escaladas	971,490.96
MODELO DEEP LEARNING	Modelo LSTM	12,138,496.66

Evaluación

HOBBIES

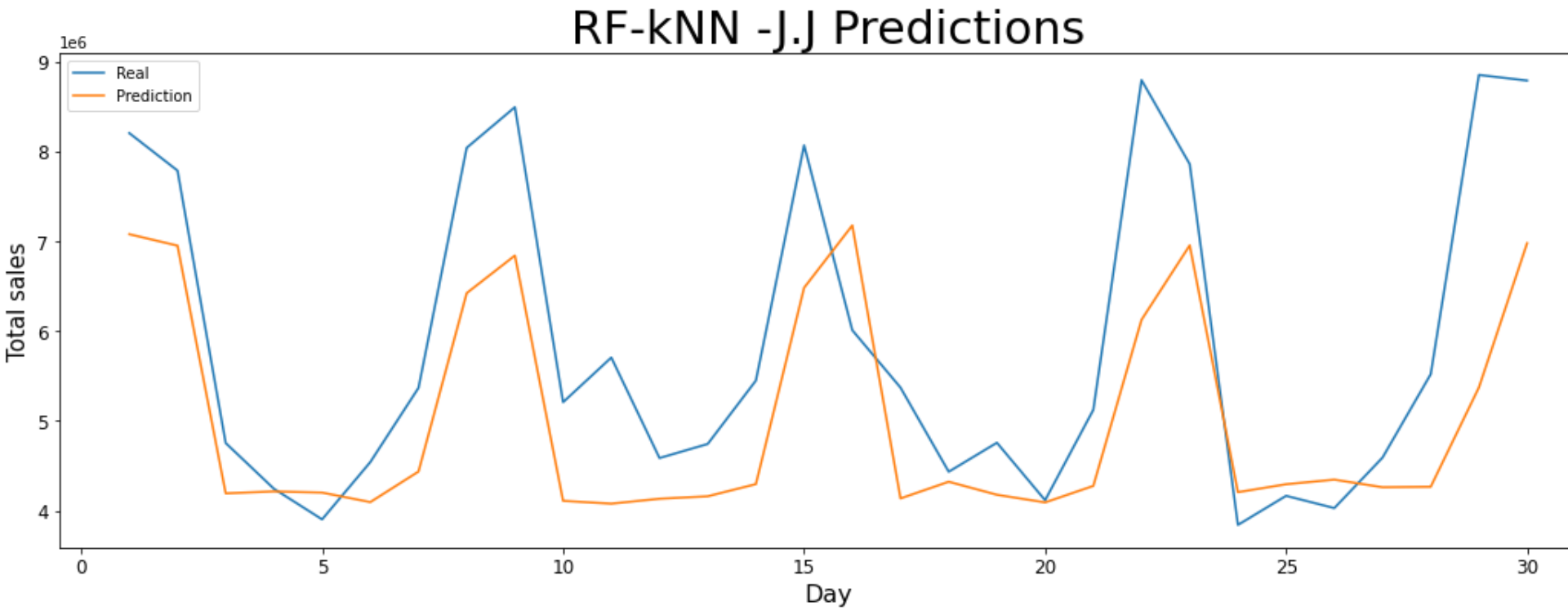


MAPE: 13,94%

n_estimators = 100,
max_features = 6,
max_leaf_nodes = 25

Evaluación

HOBBIES



MAPE: 15,01%

n_estimators = 35,
max_leaf_nodes = 2,
max_features = 9,
n_neighbors = 24

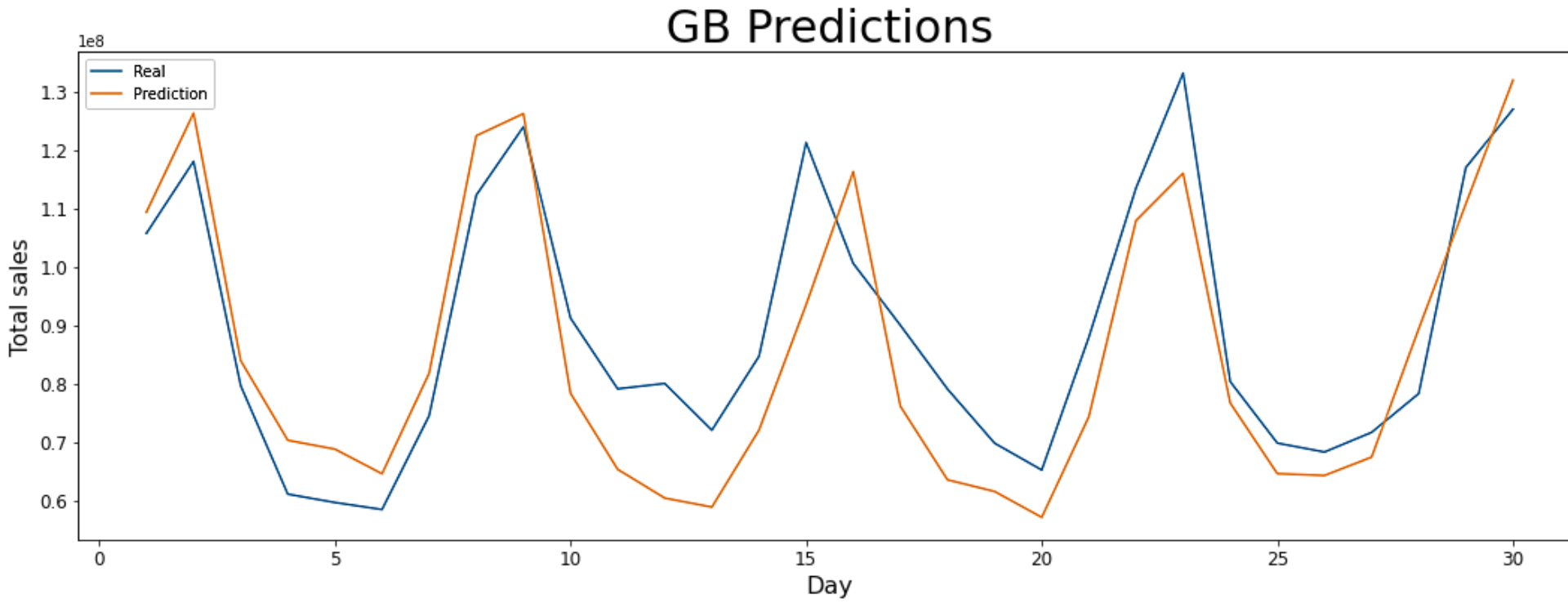
Evaluación

FOODS

CATEGORÍA	MODELO	MÉTRICA (MAE TEST) valores en USD
MODELOS ESTADÍSTICOS CLÁSICOS – SARIMA-	Modelo propuesto analíticamente (5,1,4)(1,0,0,12)	108,799,642.95
	Modelo Autoarima (4,1,5)(2,0,2,12)	122,990,454.14
MODELOS DE MACHINE LEARNING	kNN con variables escaladas	12,186,285.10
	RF con variables escaladas	11,727,245.21
	Gradient Boosting con variables escaladas	**9,913,601.50**
	SVR con variables escaladas	11,781,817.34
	RF-kNN con variables escaladas	11,822,498.31
MODELO DEEP LEARNING	Modelo LSTM	18,875,740.51

Evaluación

FOODS



MAPE: 11,50%

n_estimators = 150
max_features = 18,
max_leaf_nodes = 20

Conclusiones

- Específicamente para las ventas diarias en Walmart, la estructura univariada de los modelos SARIMA no logran captar la complejidad del modelo.

- De los 3 Frameworks usados para la predicción, los modelos de Machine Learning presentan un mejor desempeño, siendo de estos el Random Forest el mejor, aunque con un sobreajuste considerable. Si no se desea arriesgar con este sobreajuste se recomiendo usar el RF-kNN.

- La inclusión de transformaciones y variables exógenas puede influenciar positivamente en el desempeño de todos los modelos probados.

- Emplear un esquema de mayor complejidad para las redes neuronales puede mejorar el desempeño del modelo LSTM.

GRACIAS!

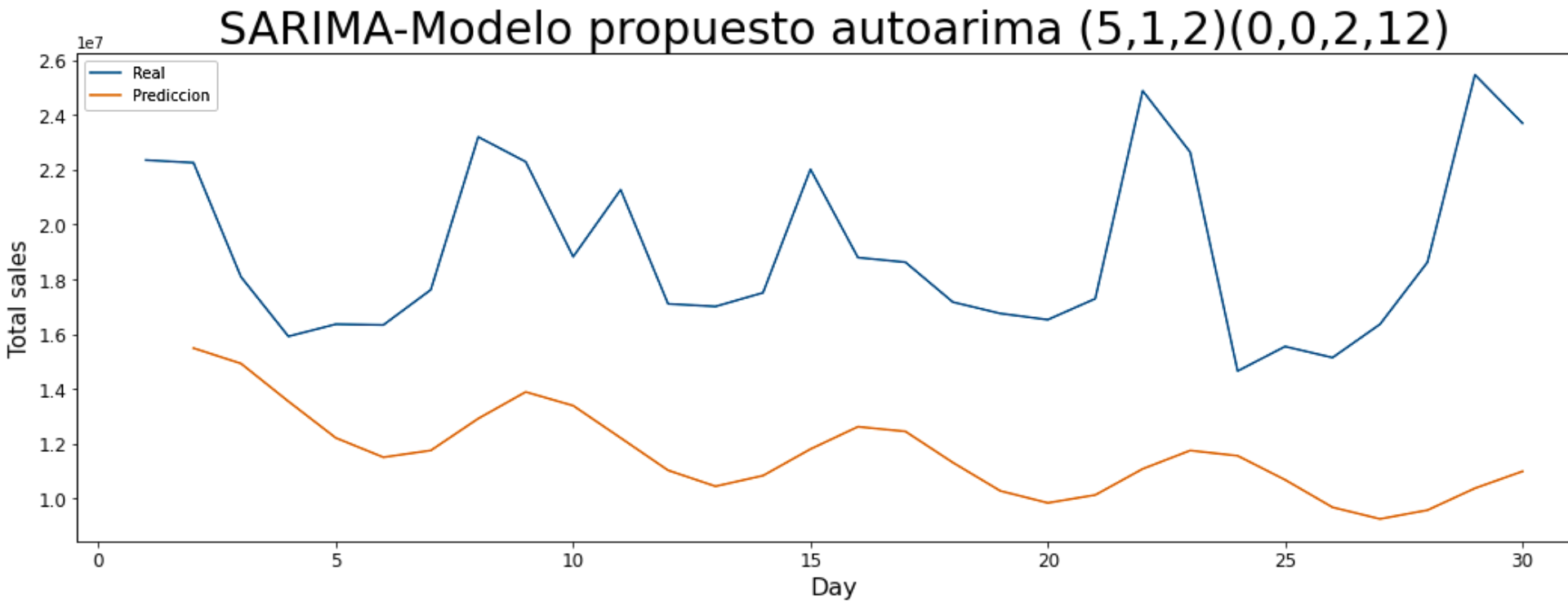
Espacio de Dudas, Sugerencias

UNIVERSIDAD EAFIT

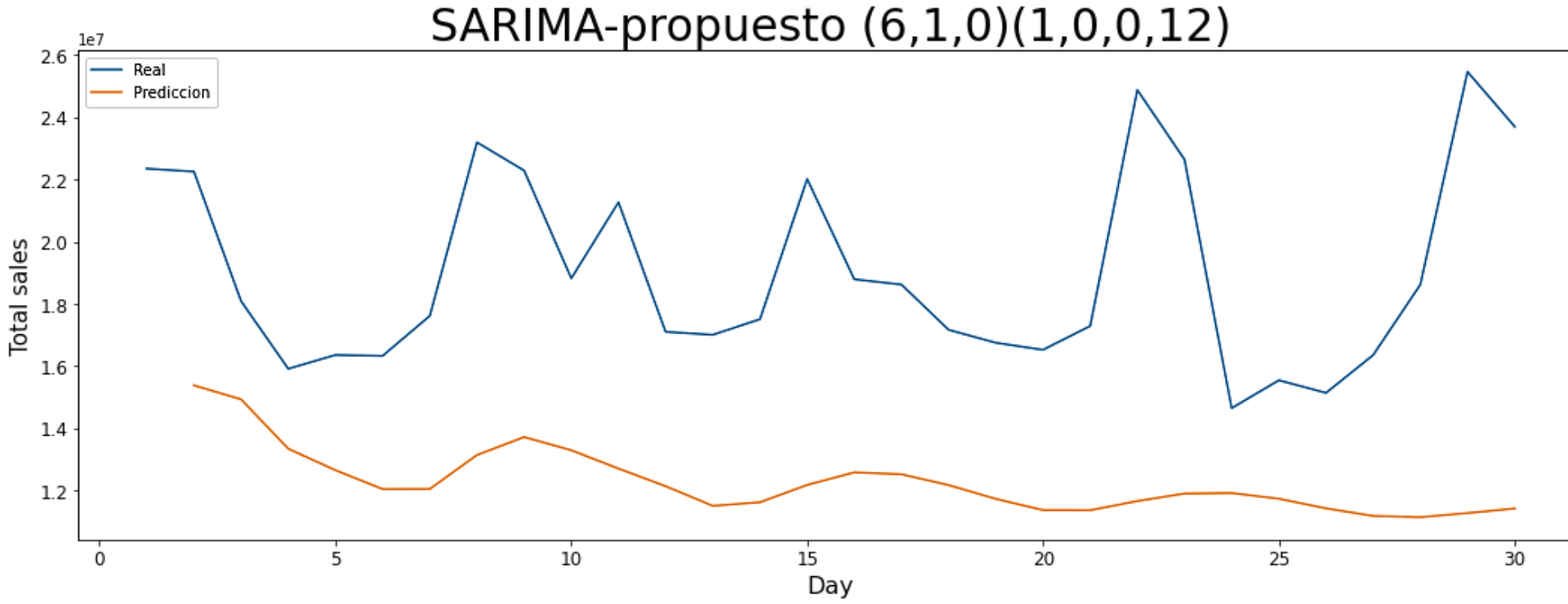
ANEXOS

UNIVERSIDAD EAFIT

HOBBIES

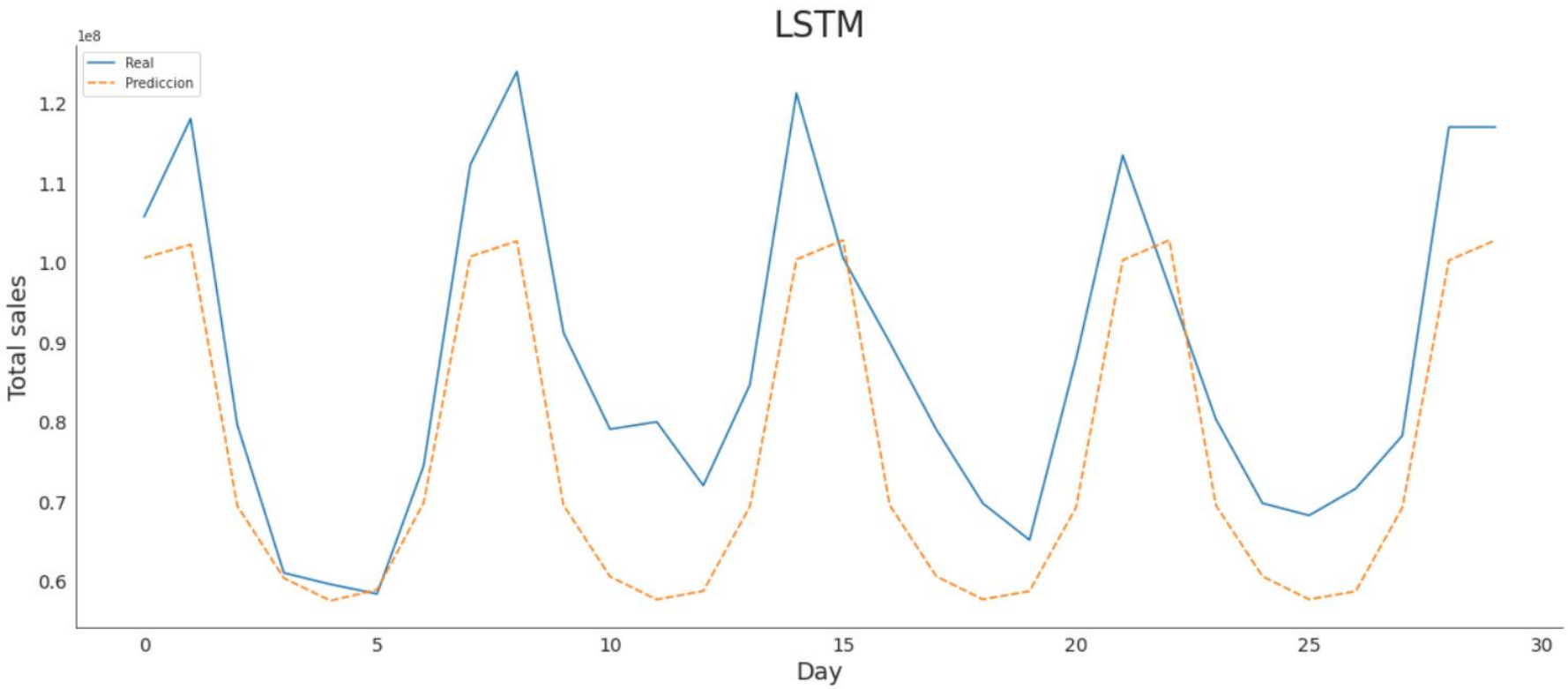

SARIMA-Modelo propuesto autoarima (5,1,2)(0,0,2,12)
1e7
Real
Prediccion
Total sales
Day
2.6
2.4
2.2
2.0
1.8
1.6
1.4
1.2
1.0
0
5
10
15
20
25
30

HOBBIES


SARIMA-propuesto (6,1,0)(1,0,0,12)
1e7
Real
Prediccion
Total sales
Day
2.6
2.4
2.2
2.0
1.8
1.6
1.4
1.2
0
5
10
15
20
25
30

Evaluación

HOBBIES



HOBBIES

Variables Seleccionadas (Backward):

Rezagos 1, 6-10, 12, 14, 15, 21-30
Kurt1M, Kurt2M, Media2M, Std6M, Mes