Pronóstico del número de ventas y evaluación de la satisfacción de los clientes de Walmart (US)

Natalia Castilla Reyes
ancastillar@unal.edu.co
Miguel Ángel Quintero
miaquinteroma@unal.edu.co

Universidad Nacional de Colombia
Departamento de Estadística

Agosto 10, 2021

Contenido

Introducción - Motivación

Diversos sectores han empezado a migrar y utilizar algoritmos de Machine Learning ya que son **propuestas de valor que permiten optimizar y predecir de acuerdo a los datos que se desean analizar**, como es el caso de la multinacional Walmart. Es por esto que estudiar y aprender a implementar estos algoritmos tanto desde el punto de vista estadístico como analítico es fundamental, ya que:

- Permite una correcta interpretación del algoritmo dependiendo de los objetivos del negocio.
- Un incorrecto uso de un algoritmo implican un riesgos que pueden conducir a pérdidas monetarias.
- Permiten utilizar al máximo las herramientas computacionales disponibles para este propósito.

Contenido

Planteamiento del Problema

Realizar correctamente el pronóstico del volumen de ventas para distintas tiendas de Walmart mediante la implementación de una red LSTM junto con un algoritmo de Machine Learning LightGBM para contrastar los resultados.

Adicionalmente, se realizará un análisis de sentimientos para evaluar la satisfacción de los clientes mediante Twitter. Para esto, los tweets de interés serán los relacionados a Walmart y seleccionados de acuerdo a la ubicación de las tiendas de estudio.

Objetivos

Para esto, se han planteado los siguientes objetivos para el desarrollo de este problema:

- Realizar un pronostico del volumen de ventas para las distintas tiendas de Walmart.
- Evaluar la satisfacción de los clientes de Walmart mediante análisis de sentimientos de tweets
- Encontrar relaciones entre la satisfacción de los clientes y el volumen de ventas en Walmart.
- Contrastar los resultados de las metodologías implementadas.

Contenido

Datos Usados en el Modelo de Pronóstico

Para el desarrollo de este proyecto, se utilizaron los datos suministrados en la siguiente competencia de Kaggle:
https://www.kaggle.com/c/m5-forecasting-accuracy/data

Los datos son recolectados en 3 diferentes estados de USA: California (CA), Texas (TX) y Wisconsin (WI) donde:

- La información registrada está desde enero de 2011 y junio de 2016.
- Comprende información de 3049 productos en 3 categorías.
- En total se tienen 10 tiendas: 4 en California, 3 en Texas y 3 en Wisconsin.
- Los datos se encuentran en 3 diferentes datasets, por lo cual es necesario su unión para obtener la información necesaria.

Carga y Preprocesamiento

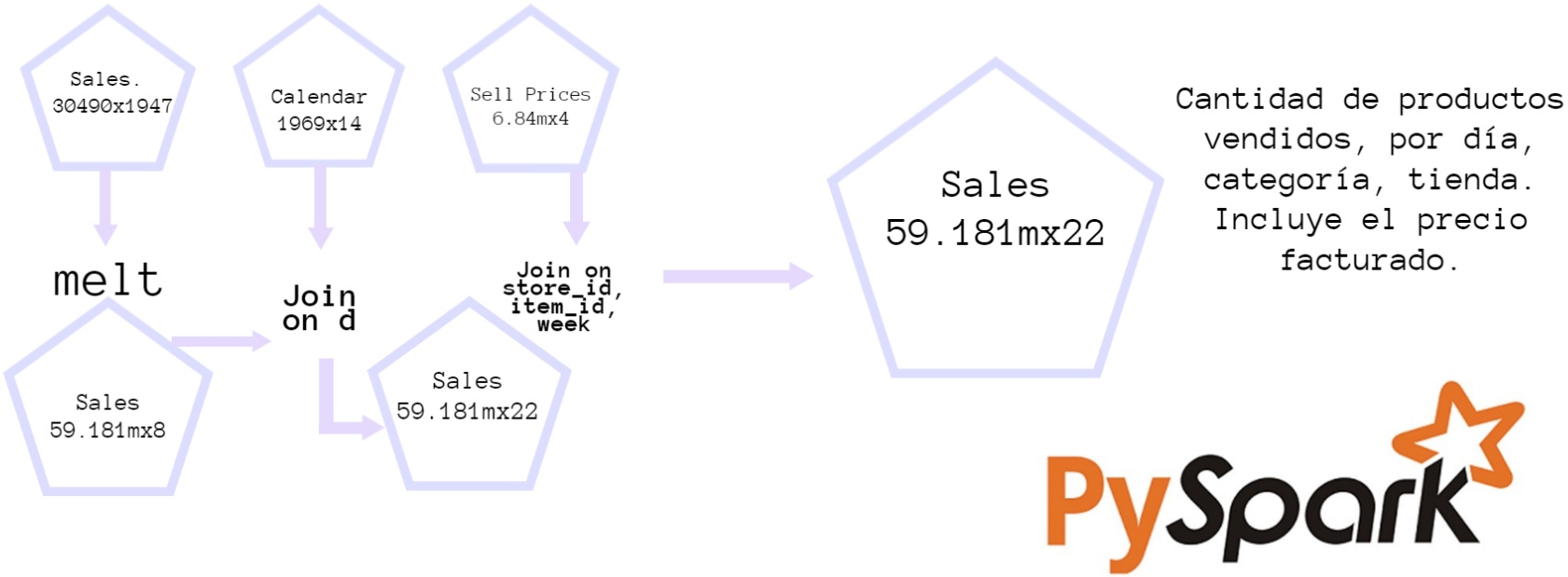

Figura 2: Esquema del procesamiento de datos realizado.

Clasificación de la información

Los datos se encuentran distribuidos en los 3 estados y se dividen en FOODS, HOUSEHOLD Y HOBBIES, los cuales a su vez se subdividen en distintos grupos.

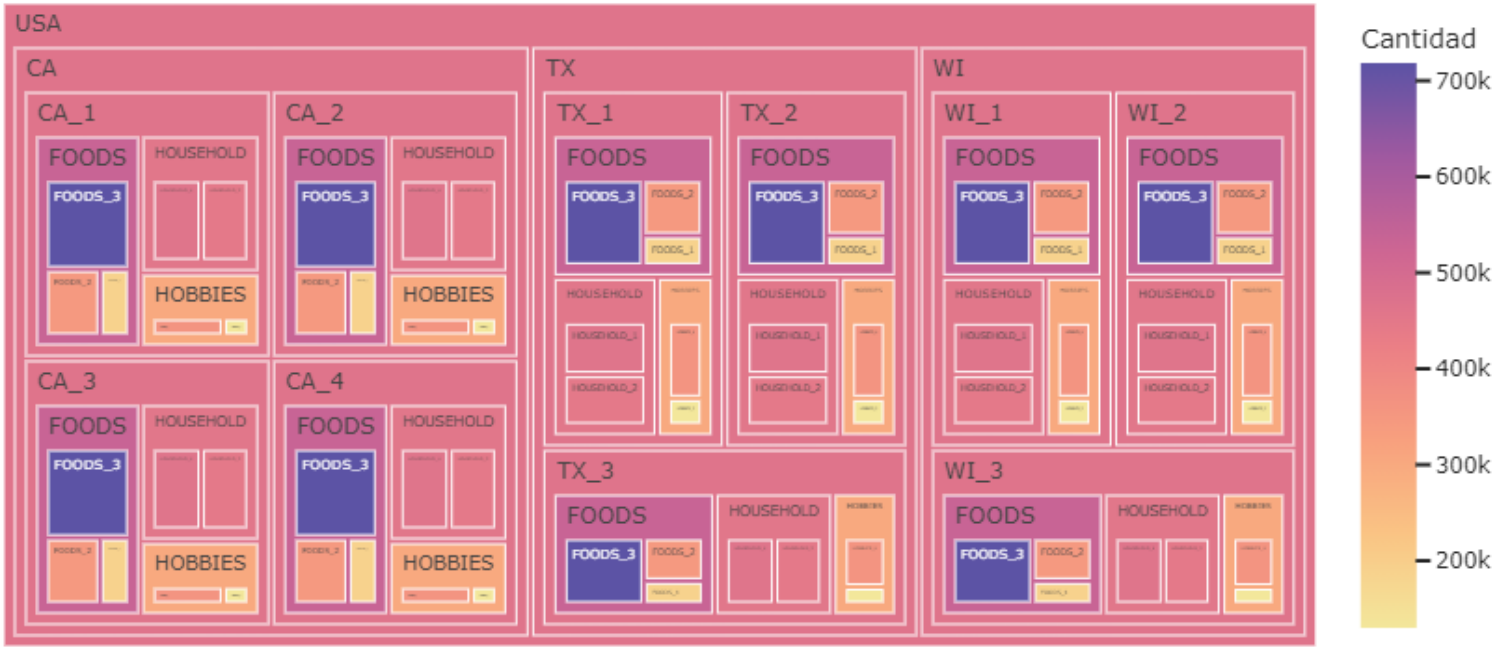

Figura 3: Clasificación de los datos

Distribución de los precios

Al observar el precio de venta de los productos, se observa que la categoría FOODS tienen menores precios que los otros items.

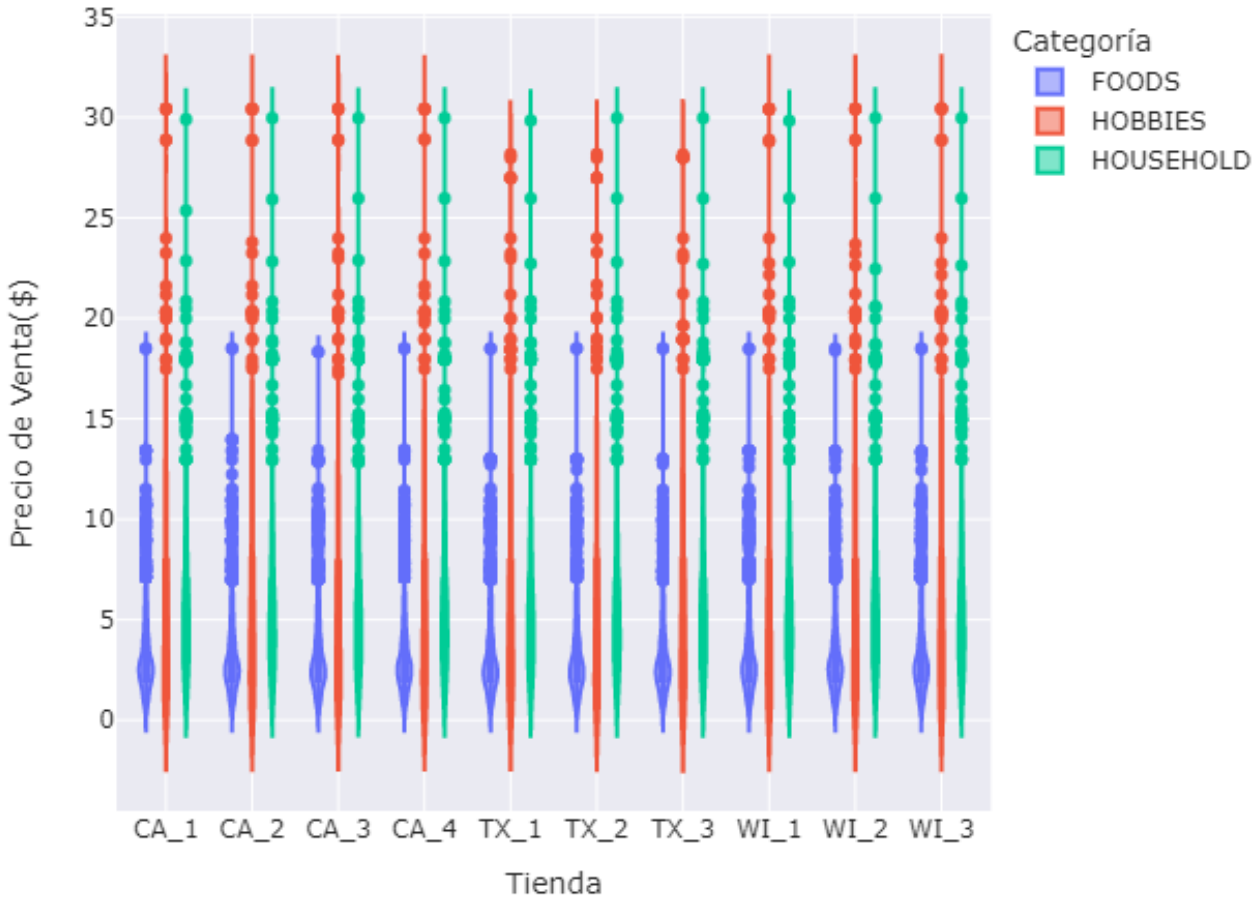

Figura 4: Distribución de los precios por categoría

Volumen de productos vendidos por tienda

Se observa que el volumen de ventas de cada tienda es bastante diferente a excepción de las tiendas en Texas. No obstante, se observan patrones similares en la caída y aumento de precios.

Figura 5: Productos vendidos en cada tienda respecto al tiempo de estudio

Volumen de productos vendidos por tienda

En este caso se analizan los datos para la tercera tienda de California.

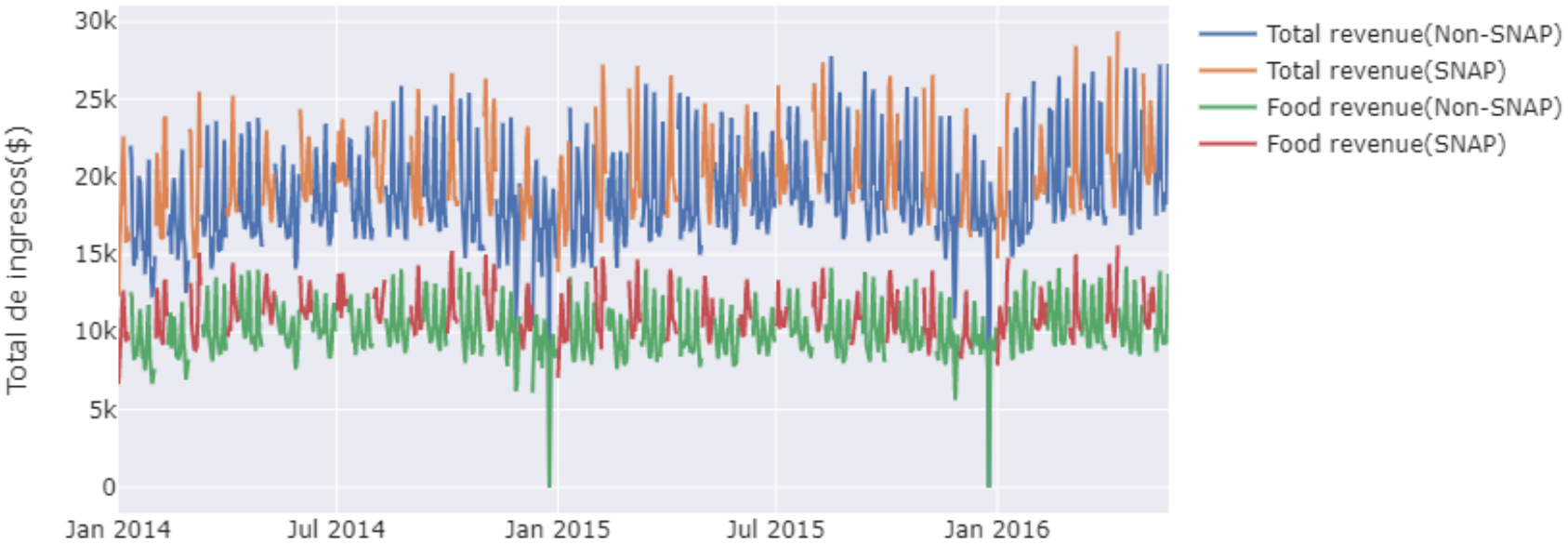

Figura 6: Productos vendidos para la tienda 3 en California (CA_3)

Contenido

Preprocesamiento de los datos

Antes de realizar el entrenamiento por medio de un modelo LSTM, es necesario realizar un preprocesamiento de los datos.

1. En primer lugar, es necesario realizar un reescalamiento de los datos ya que dentro del modelo LSTM solo se permiten valores dentro de un rango. Para esto, se utiliza la función *MinMaxScaler* de Keras, donde se escalan los datos entre 0 y 1.
2. Se particionan los datos para cada tienda en una base de entrenamiento y otra de prueba. En este caso el particionamiento será igual para todas las tiendas y será 80/20.
3. Es importante mencionar que dentro del modelo LSTM se utiliza un 10 % de los datos para validación.

Preprocesamiento de datos

En la siguiente figura se presentan los datos de entrenamiento y prueba para dos tiendas. Este mismo procedimiento se aplica para todas las 10 tiendas.

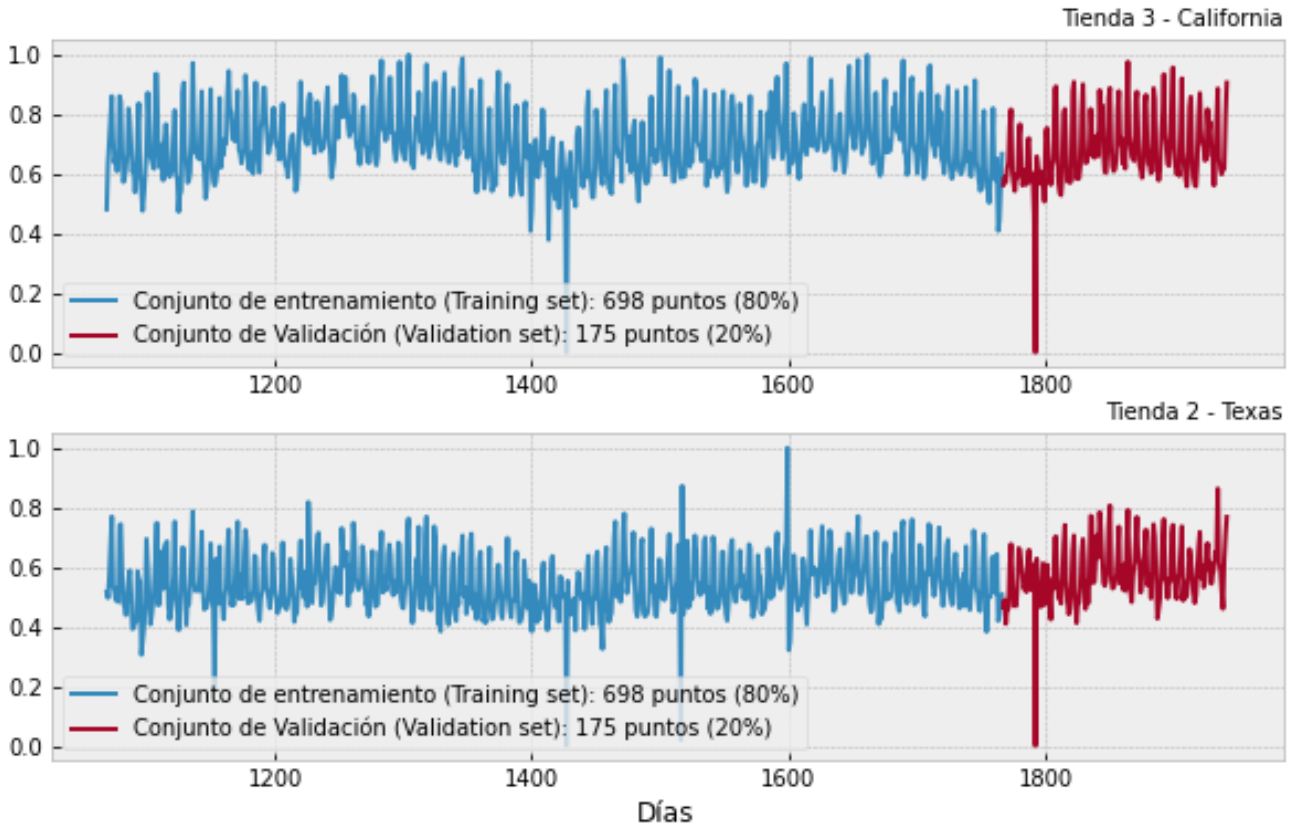

Figura 7: Datos escalados y divididos en entrenamiento y prueba

Estructura del Modelo LSTM

Para el modelo LSTM, se han utilizado los siguientes parámetros:

- Se han utilizado los últimos 30 retardos para el entrenamiento del modelo.
- Se tiene una primera capa tipo LSTM con un input de (30,1) y una salida de 60 neuronas.
- Se incluye una segunda capa densa con entrada de 60 neuronas y salida una neurona.
- Se utiliza como función de perdida el error cuadrático medio.
- El entrenamiento se realiza para 50 épocas.

Entrenamiento del Modelo

Se realiza el entrenamiento para cada tienda utilizando el modelo anteriormente propuesto. De esta manera se evalúa la función de pérdida para cada una.

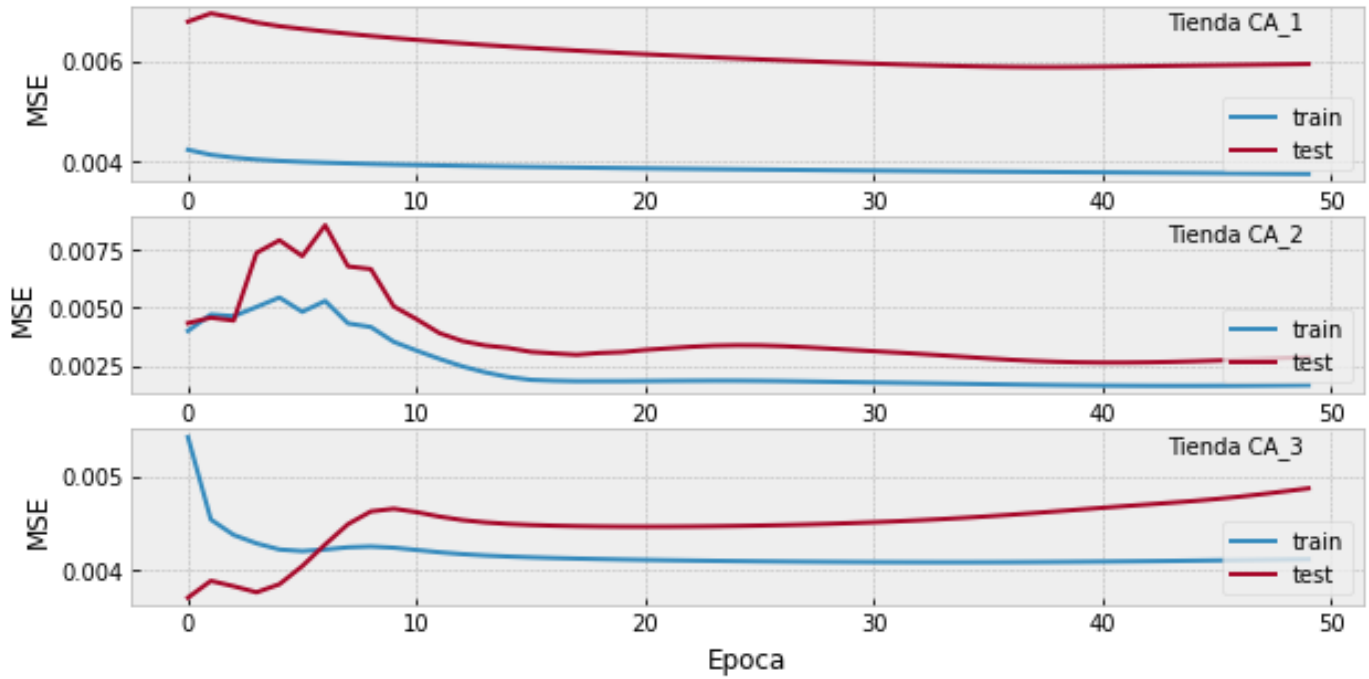

Figura 8: Función de pérdida del modelo para diferentes tiendas.

Resultados del Modelo LSTM

Finalmente se realiza la predicción de los datos de prueba, y se contrastan los resultados para los últimos 28 días en la siguiente gráfica.

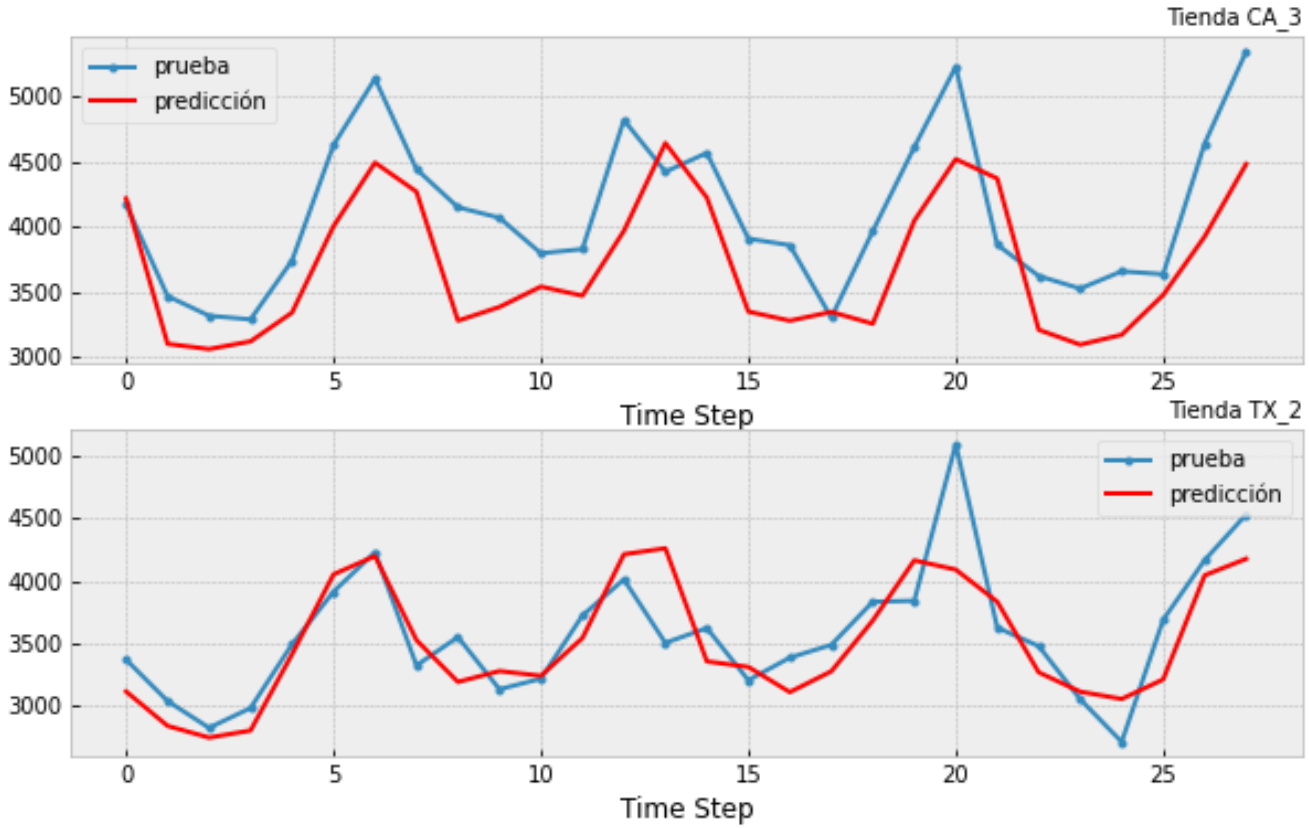

Figura 9: Resultados de la predicción en los últimos 28 días.

Contenido

Preprocesamiento de datos

Para realizar el pronóstico del volumen de ventas aplicando este modelo es necesario introducir algunas variables que le agreguen la temporalidad a la fuente de datos. Las variables creadas fueron:

- Creación de variables de retraso.
- Variables promedio de la cantidad de ventas, por tipo de producto, tienda y fecha.
- Variables tendenciales.
- Variables de ventana móvil.

Por otra parte, se tiene:

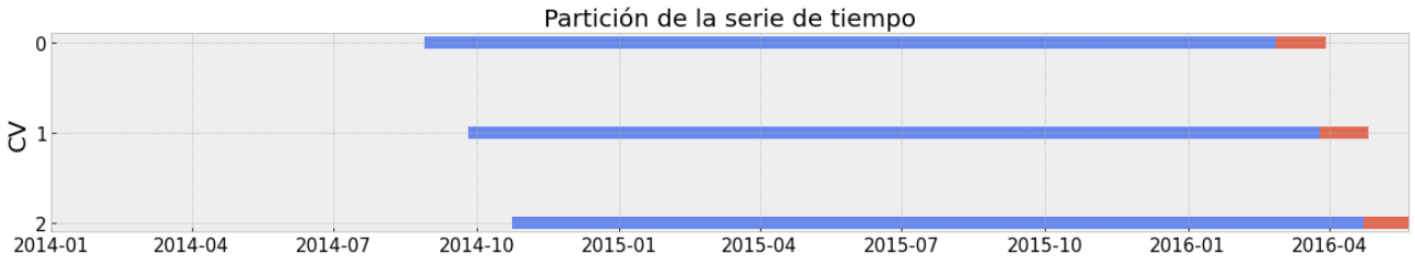

Figura 10: Partición de la series para realizar cross validation.

Resultados modelo LGBM

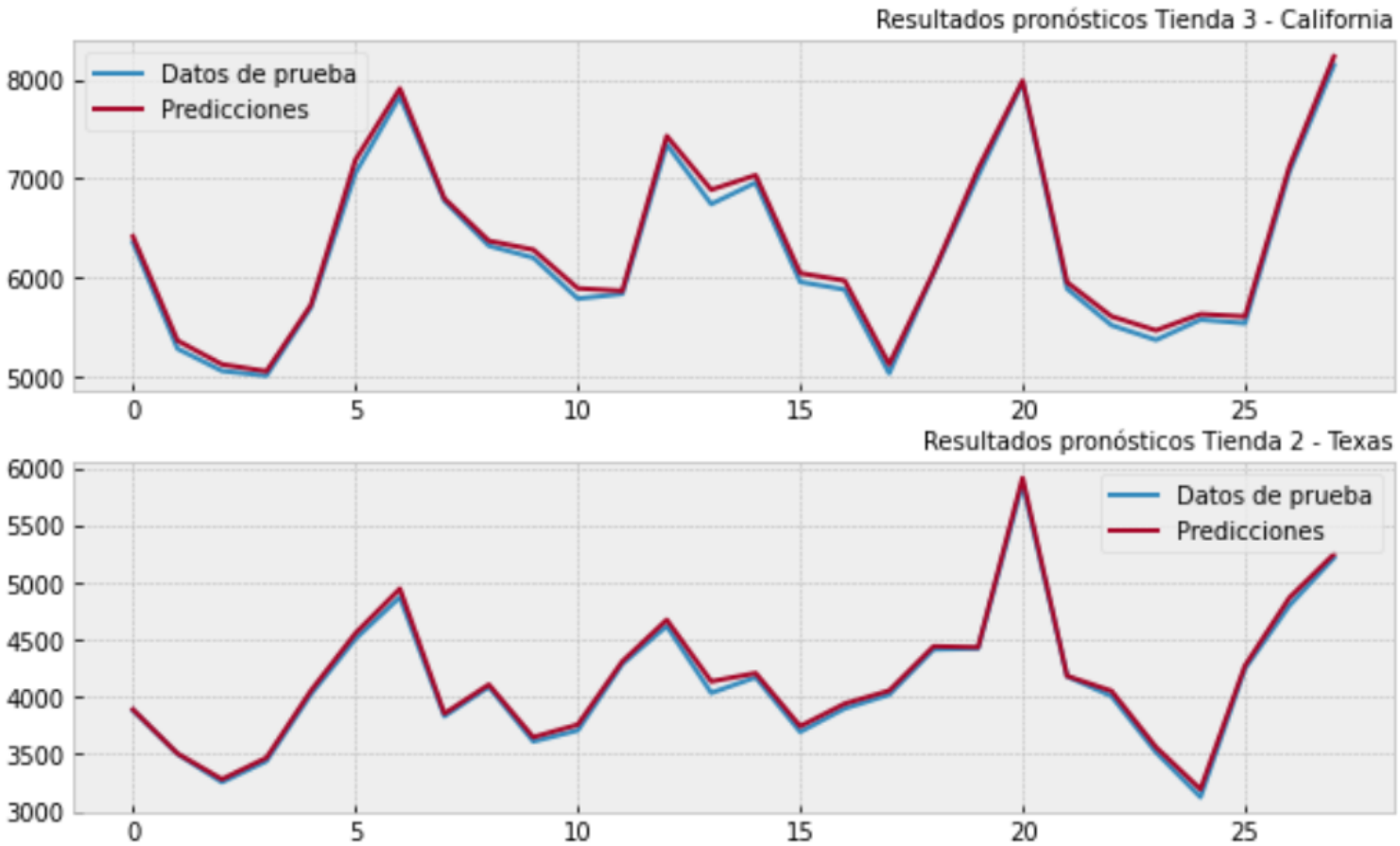

Figura 11: Resultados de la predicción de 28 días.

Variables importantes

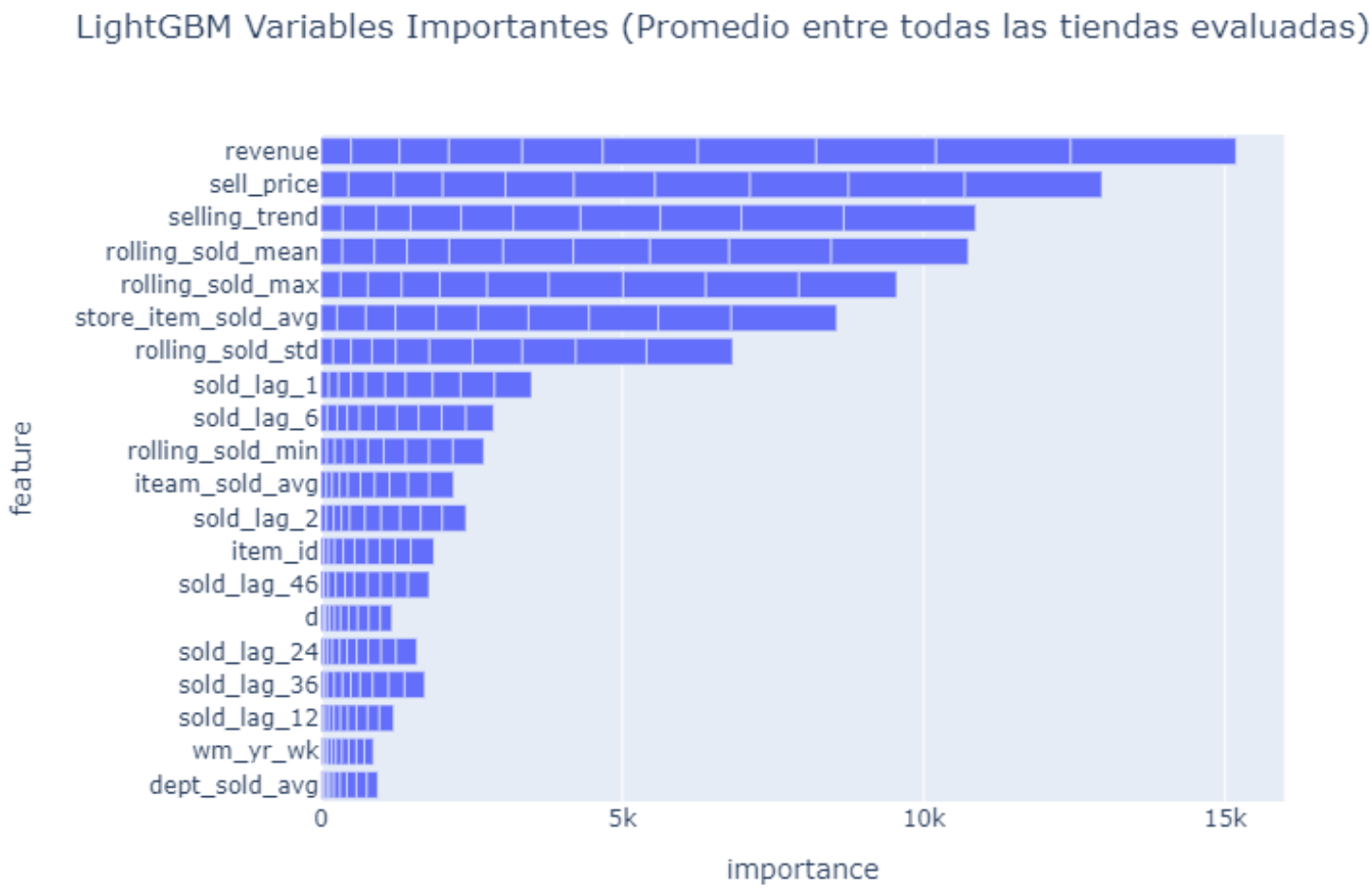

Figura 12: Variables importantes (promedio sobre los 10 modelos).

Función de pérdida de los modelos

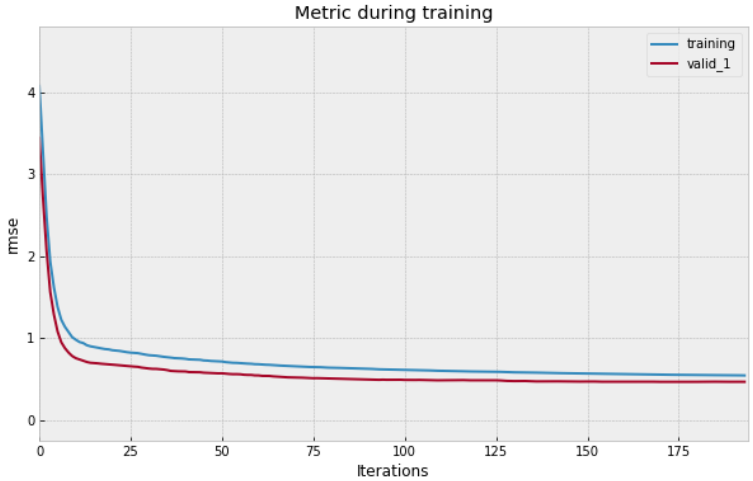

Figura 13: Función de pérdida para el modelo de la tienda 3: California

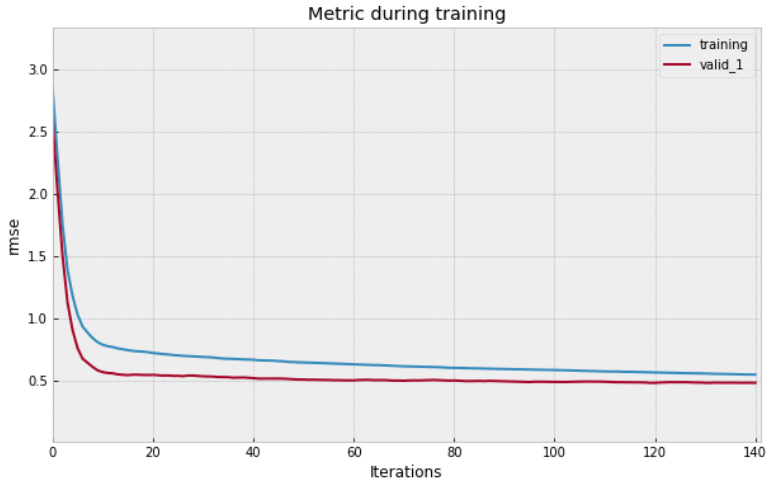

Figura 14: Función de pérdida para el modelo de la tienda 2: Texas

Contenido

Comparación de resultados

Finalmente, se realiza la comparación de ambos modelos utilizando la error cuadrático medio como métrica.

Tienda	LSTM	LGBM	Tienda	LSTM	LGBM
CA_1	546.21	44.87	**TX_2**	350.04	44.04
CA_2	600.72	47.22	**TX_3**	346.92	32.39
CA_3	588.94	76.88	**WI_1**	513.50	19.30
CA_4	587.07	22.99	**WI_2**	563.39	58.98
TX_1	427.37	33.42	**WI_3**	615.98	36.92

Cuadro 1: Comparación de la raíz del RMSE de cada tienda para cada modelo

Contenido

Medición de la satisfacción del cliente

Figura 15: Pipeline de la implementación del modelo de emociones

Desarrollo del modelo

Para desarrollar el modelo se emplearon 4 fuentes de datos:

- La base de entrenamiento tiene una dimensión de 16.000 registros y dos columnas. La primera variable corresponde a una frase aleatoria, la segunda es la etiqueta de la emoción asignada a esa frase.
- La base de prueba y validación tiene la misma estructura que la de entrenamiento cada una consta de 2.000 frases etiquetadas.
- fuente de datos para corregir palabras mal escritas.
- fuente de datos para corregir las contracciones de las palabras en inglés.

Los datos fueron obtenidos de Kaggle:
https://www.kaggle.com/ananthu017/emotion-detection-fer.

Datos de entrenamiento

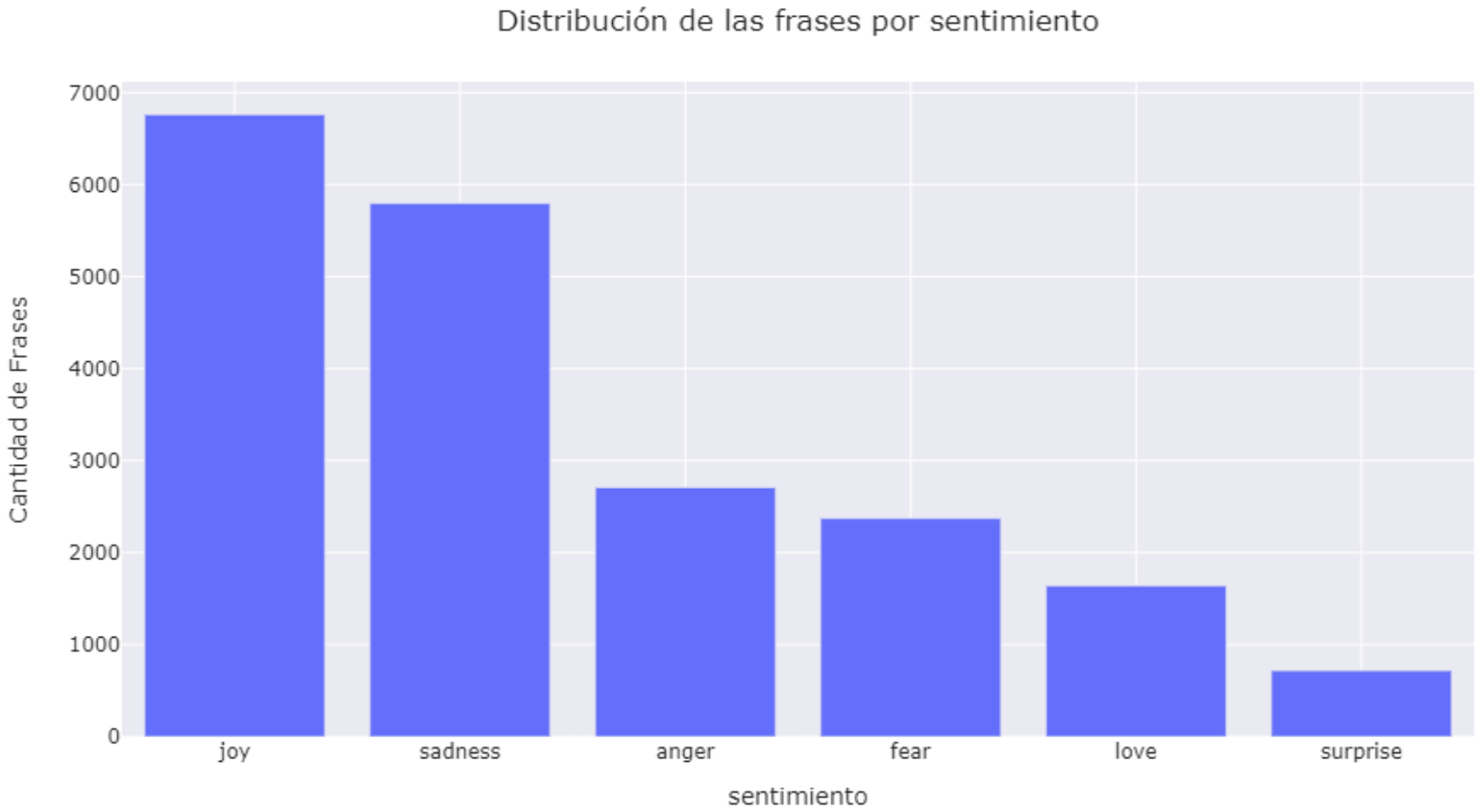

Figura 16: Entendimiento de la distribución de los datos de entrenamiento.

Procesamiento de datos

Figura 17: Resumen esquemático del procesamiento de datos hasta la aplicación del modelo.

Resultados conjunto de validación

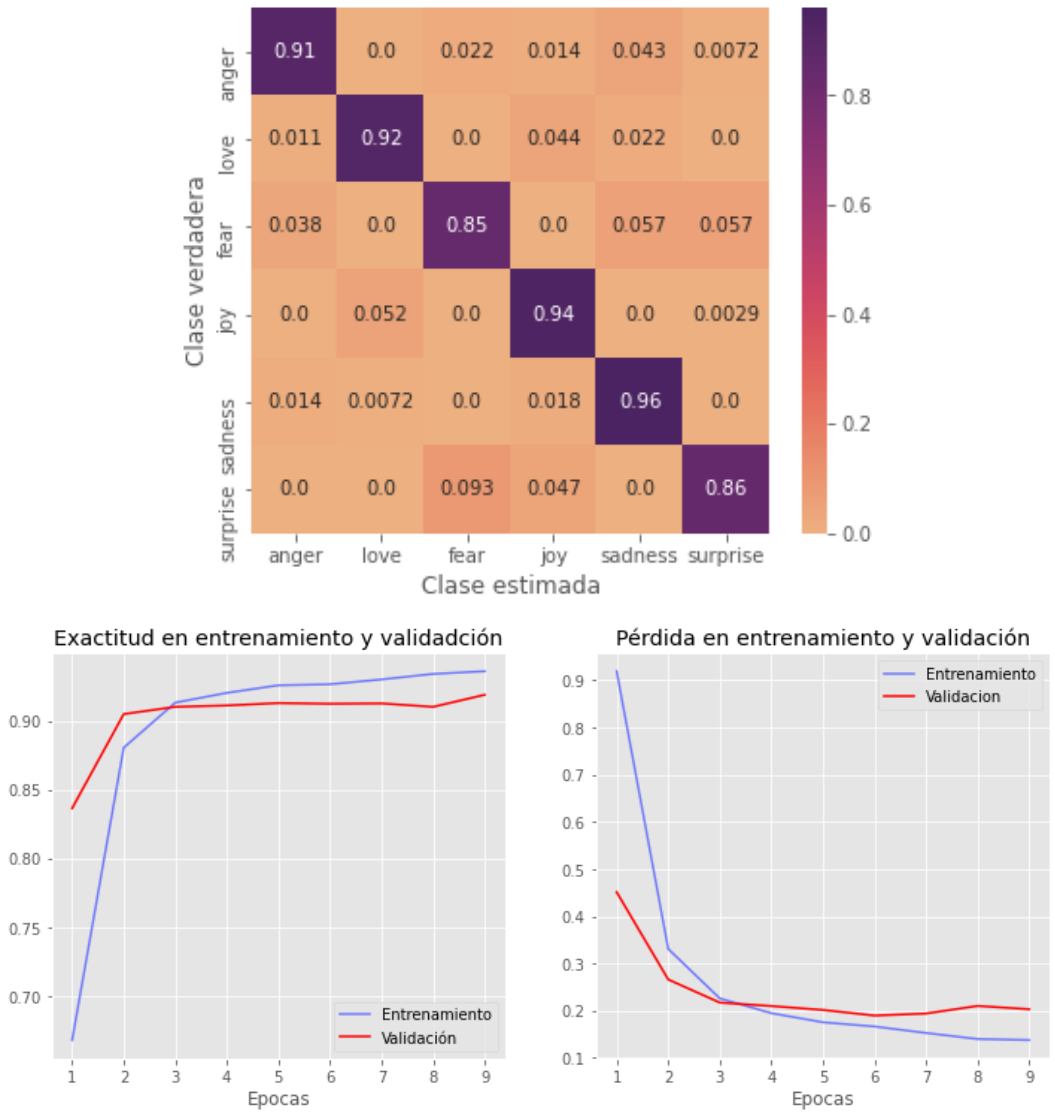

Figura 18: Resultados en el conjunto de validación para nuestro modelo de detección de emociones.

Resultados conjunto de prueba

Clase	Precisión	Exhaustividad	F1-score
0	0.93	0.91	0.92
1	0.81	0.92	0.86
2	0.93	0.85	0.89
3	0.96	0.94	0.95
4	0.95	0.96	0.96
5	0.82	0.86	0.84

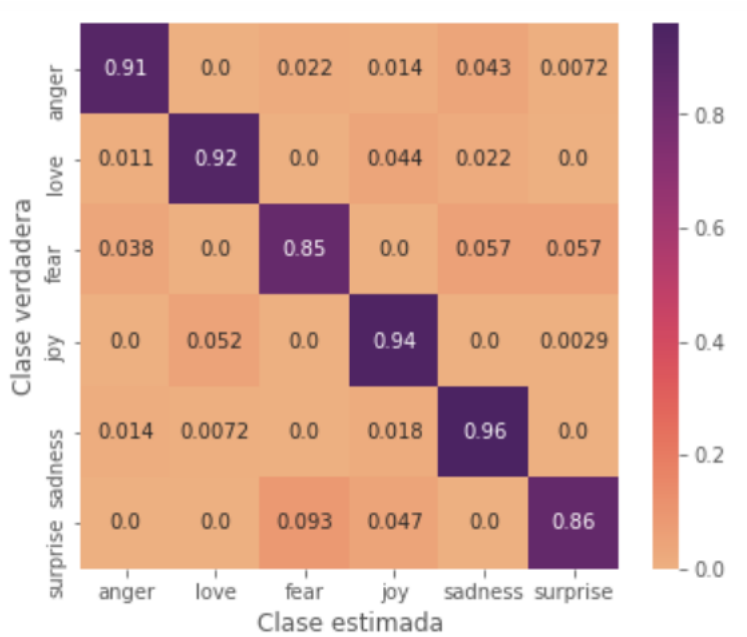

Figura 19: Resultados en el conjunto de de prueba.

Salida del modelo

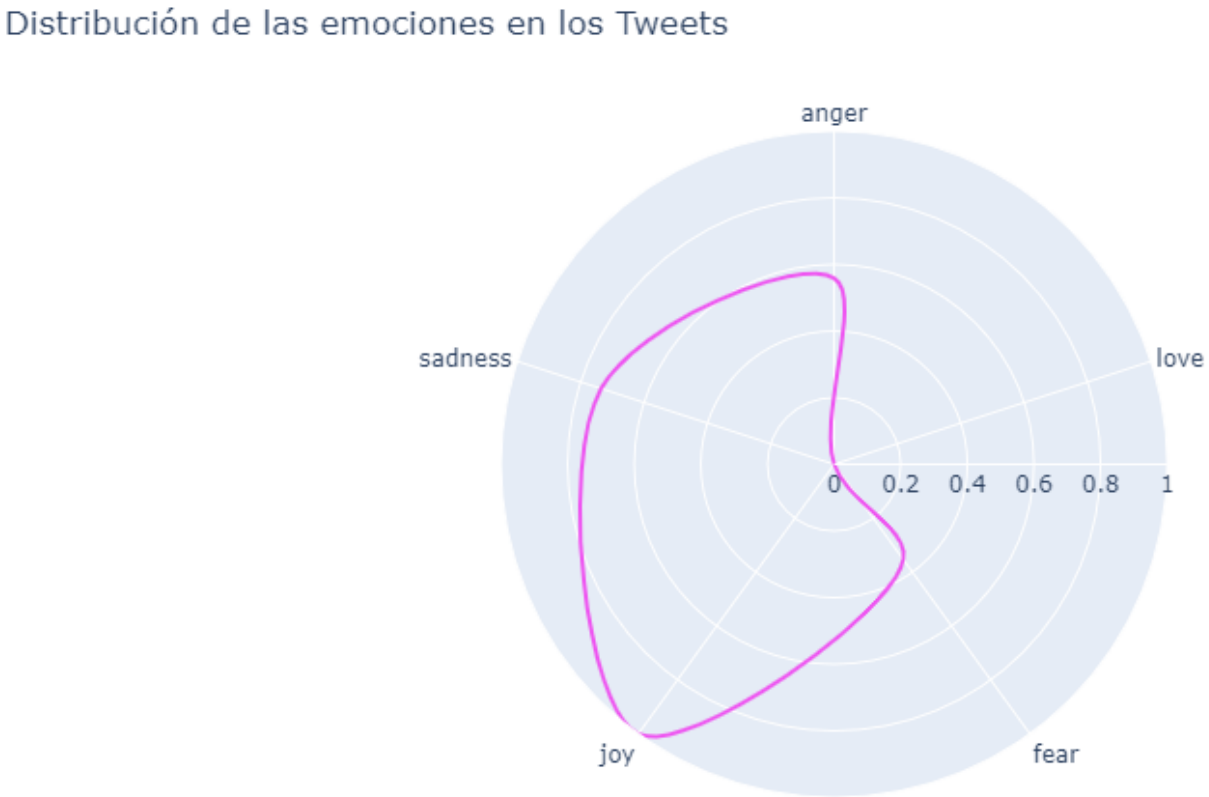

Figura 20: Salida del modelo de extracción de emociones de los Tweets.

Aplicación a Walmart California

- Podemos medir la satisfacción de los clientes respecto a un producto específico. De esta manera se pueden detectar aquellos productos que no están generando un impacto positivo lo cual es un insumo para la toma de decisiones.

Figura 21: Nube de palabras de los clientes de Walmart.

Aplicación a Walmart California

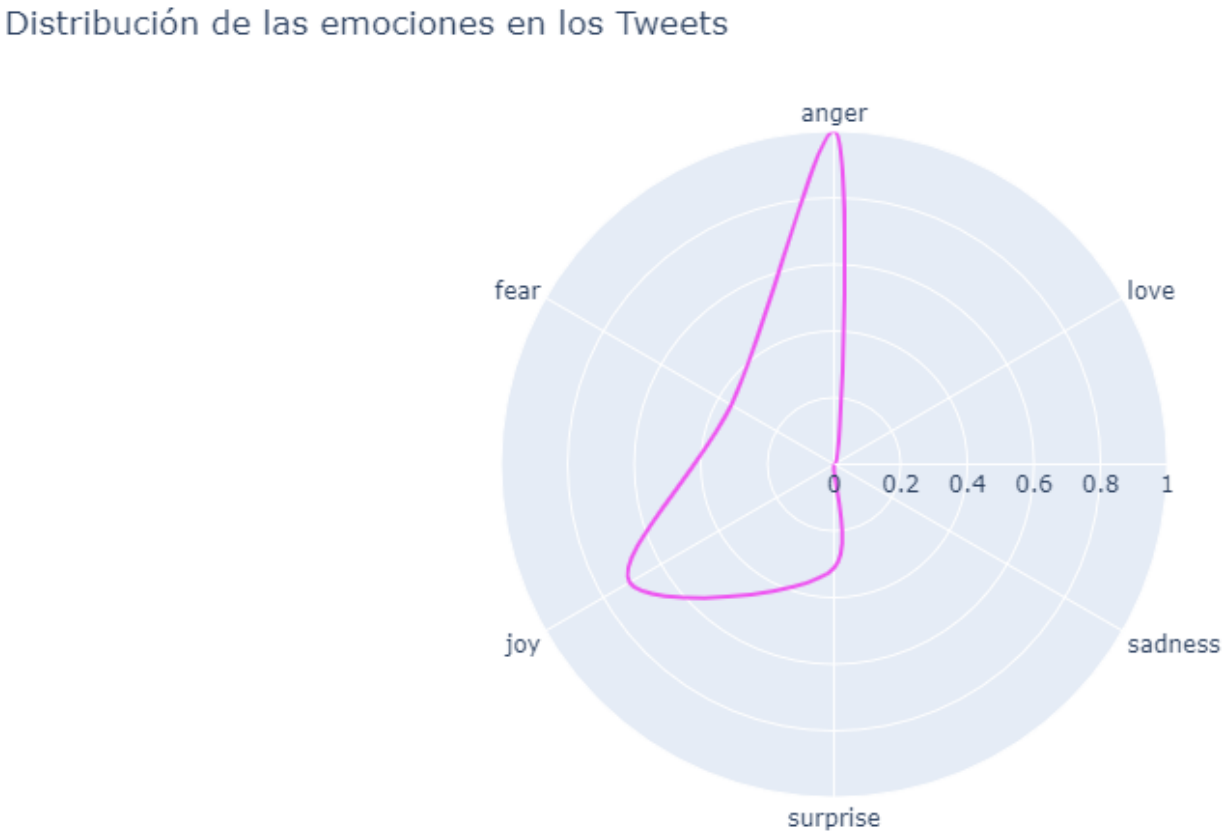

Figura 22: Satisfacción general de los clientes de Walmart.

Contenido

Conclusiones

- Se observa que el algoritmo basado en redes neuronales (LSTM) *no necesariamente ofrece un mejor rendimiento*, ya que en este caso el modelo LGBM presenta mejores resultados. No obstante, es importante tener en cuenta que el modelo LSTM se puede mejorar encontrando los mejores hiperparámetros y realizando un modelo diferente por tienda.
- A partir del análisis de los tweets de los clientes de Walmart California se puede concluir: El 40.4 % de los clientes expresan molestia, 20.9 % disfrute, 15.1 % miedo, 13.4 % sorpresa, 1.4 % amor y un 1.01 % tristeza.

Contenido

Trabajo Futuro

Al realizar el anterior proyecto se han podido determinar diferentes áreas en las que el trabajo se puede seguir desarrollando y mejorando:

- Realizar una búsqueda de los mejores parámetros de entrenamiento para el modelo LSTM, de tal manera que no sea el mismo para cada tienda.
- Realizar predicciones para productos específicos o áreas dentro de las tiendas.
- Recopilar suficiente información de Twitter para poder verificar si existe una relación entre la satisfacción de los clientes en esta plataforma y el volumen de ventas.

Trabajo Futuro

- Mediante minería de texto filtrar los Tweets por emociones. Posteriormente, se propone crear un modelo de tópicos (LDA),con los textos de cada emoción, con el objetivo de identificar cuales son los diversos temas que están generando esa emoción en específico en los clientes de Walmart.
- Extender el análisis para los demás estados. Podemos filtrar los Tweets por coordenadas más específicas para contrastar requerimientos entre grandes y pequeñas ciudades.
- Construir una aplicación que realice el pronóstico de ventas (para una base de datos similar) y la medición de la satisfacción del cliente mediante Twitter.