APLICACIÓN DE ML PARA CONSTRUCCIÓN DE MODELOS PREDICTIVOS EN DISTRIBUIDORA IVANS

Sergio Geovany García Smith

Junio 2025

Resumen

Los modelos de inteligencia artificial (IA) se han convertido en herramientas clave para optimizar decisiones comerciales. Este trabajo implementa tres modelos predictivos para la empresa comercial Distribuidora Ivans:
inversión en mercadería, utilidad bruta y ventas por categoría, utilizando algoritmos de aprendizaje automático (XGBoost, LSTM, MLP) y técnicas clásicas (regresión polinomial, SVR). Los resultados destacan que:

- **XGBoost** superó a LSTM y MLP en predicción de inversión y ventas por categoría.
- **Regresión polinomial (grado 2)** fue óptima para utilidad bruta.
  El estudio revela que modelos sencillos pueden ser más efectivos que redes neuronales en datasets pequeños, mientras que el preprocesamiento riguroso (lags, filtrado de outliers) es crítico para mejorar la precisión.

Palabras Clave: series temporales, XGBoost, LSTM, preprocesamiento.

1. Introducción

Actualmente los modelos de IA están constituyendo una herramienta muy importante para el sector comercial, pues a través de estos se construyen sistemas que permitan predecir o aproximar a data de mucho interés que a su vez ayudan a clarificar la visión sobre las operaciones y con ello tomar decisiones más informadas en beneficio de la empresa.

Para empresas que se dedican a la venta y distribución de productos de consumo masivo (consumo diario) es muy importante contar con este tipo de sistemas ya que a través de ellos podrían agilizarse varios procesos como, por ejemplo: la cadena de suministro con sus proveedores, contar con un inventario actualizado, tener una idea más concreta del capital a invertir en ciertas temporadas, etc. La distribuidora comercial Ivans es una empresa de este tipo de negocios, que actualmente funciona en el occidente del país y hace más de un año tomó la decisión de computarizar sus procesos para operar de manera más eficiente y competitiva, por lo cual es una empresa ideal para la implementación de ese tipo de modelos.

Los modelos de predicción que se han implementado en el presente trabajo de acuerdo a la data histórica que hasta el momento se poseen son:

- Inversión en mercadería
- Utilidad bruta
- Ventas por categoría

2. Metodología

2.1 Modelo 1 – Inversión en Mercadería

Descripción:

Este modelo tiene como objetivo poder predecir la cantidad de dinero en Quetzales que se deberá invertir en la compra de mercadería para abastecimiento del inventario, el modelo recibirá como entrada los parámetros de fecha y se mostrara como salida la inversión en Quetzales.

Dataset:

Se utilizó el dataset llamado inversion mercaderia.csv, el cual esta conformado por las columnas id, fecha y total compra compuesto por 396 registros comprendidos entre el 01/04/2024 al 31/05/2025 fue extraído directamente de la base de datos de la empresa.

Preprocesamiento:

Al analizar la data se realizaron acciones como: agregar features adicionales: dia, mes, año, dia de la semana, dia el año, trimetstre, al ser un problema de series de tiempo se agregaron intervalos de tiempo(lags) y medias móviles (rolling_means) basado en ventanas de tiempo, es importante mencionar que para la división del set train/test se hizo de forma cronológica (a partir de una fecha), finalmente se realizó un escalado a la data.

Arquitectura y entrenamiento del modelo:

Los problemas de series de tiempo tienden a resolverse de mejor manera utilizando algoritmos como: XGBoost, Redes MLP y LSTM por lo cual se decidió realizar el entrenamiento con las siguientes arquitecturas:

-XGBoostRegressor()

-MLP con capa de entrada de 64 neuronas, capa oculta de 32 neuronas y capa de salida con 1 neurona, se entrenó a 100 épocas.

-LSTM con 50 neuronas en la capa de entrada y 1 neurona en la capa de salida, se entrenó a 50 épocas.

Evaluación y Selección del modelo

Luego del entrenamiento de cada modelo se obtuvieron las métricas:

Tabla 1: Métricas de los modelos entrenados – Modelo 1

Modelo	MAE	RMSE	R^2
XGBoost	9176.83	17185.51	0.14
MLP	8520.07	20171.76	-0.19
LSTM	10025.46	20871.14	-0.10

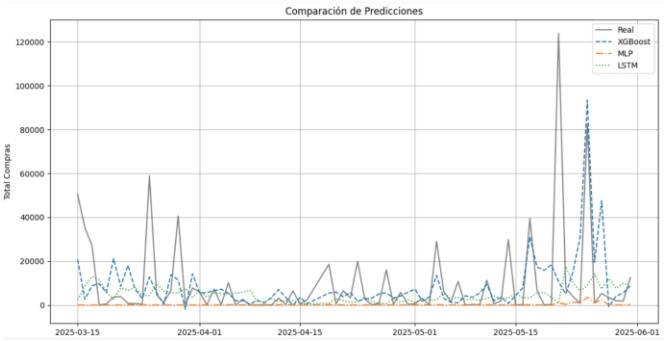

Figura 1. Comparación de predicciones – Modelo 1

De acuerdo a las métricas obtenidas y las comparaciones de las predicciones de cada modelo se puede observar que el que más se adaptó a los datos fue XGBoost aunque cabe mencionar que sus métricas y resultados siguen siendo bajas para considerarlo un modelo más efectivo en la solución al problema planteado.

2.2 Modelo 2 – Utilidad Bruta

Descripción:

Con éste modelo se busca predecir la utilidad bruta obtenida por la empresa de acuerdo al total monetario de las ventas que se realicen, el modelo recibiría como entrada el total de ventas en Quetzales y su salida seria el total de utilidad generada en Quetzales.

Dataset:

El dataset utilizado para este modelo se llama utilidad_bruta.csv, formado por las columnas vendido y utilidad con un total de 398 registros comprendidos entre el 01/04/2024 al 31/05/2025 y fue extraído directamente de la base de datos de la empresa.

Preprocesamiento:

Durante el análisis a los datos se observó que existían outliers por lo cual se procedió a filtrar y eliminarlas mediante el meto IQR obteniendo así una data con menos ruido y mejor distribución

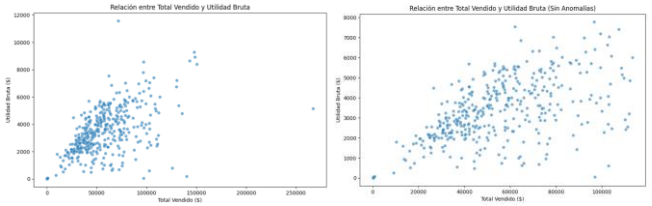

Figura 2. Filtración outliers – Modelo 2

Para facilitar el proceso de interpretación del modelo se realizó también un Escalado Estandar para que los valores quedaran comprendidos en una escala menor a la original.

Arquitectura y entrenamiento del modelo:

Este problema es de tipo regresión por lo que se utilizaron algoritmos clásicos de ML como:

- Regresión Lineal
- Regresión Polinomial (grado 2)
- SVR(Kernel RBF)
- Random Forest
- XGBoost

Evaluación y Selección del modelo

Al realizar el entrenamiento de los modelos se obtuvieron las métricas:

Tabla 2: Métricas de los modelos entrenados – Modelo 2

Modelo	RMSE	R^2
R. Lineal	1158,33	0.3327
R. Pol-g2	1090.86	0.4081
SVR	1161.69	0.3288
RF	1320.19	0.1332
XGBoost	1582.70	-0.24

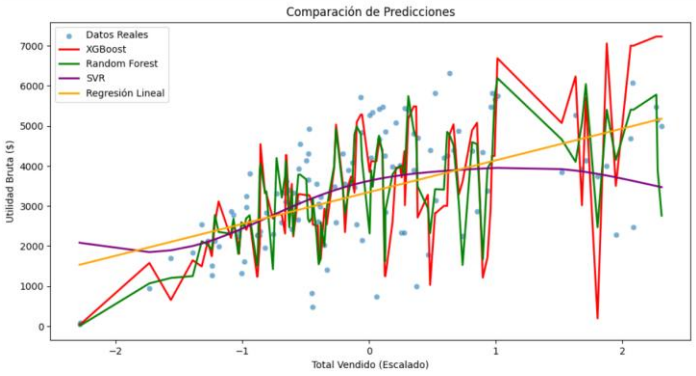

Figura 3. Comparación de predicciones – Modelo 2

Como puede notarse el modelo que mejor interpretación de los datos hizo fue La Regresión Polinomial de grado 2, aunque los modelos de Regresión Lineal y SVR también tuvieron una interpretación aceptable. Podemos considerar de forma general que este modelo presento un mejor rendimiento respecto al modelo 1.

2.3 Modelo 3 – Ventas por categoría

Descripción:

Con este modelo se busca poder predecir el total de ventas en Quetzales que se generaría por cierta categoría en una fecha específica, es decir que el modelo recibiría como entrada la fecha y la categoría y la salida seria la predicción del total monetario de ventas.

Dataset:

Se trabajó con el dataset llamado categorías_vendidas.csv que contiene las columnas fecha, categoría y total con más de 130mil registros comprendidos entre el 01/04/2024 al 31/05/2025 fue extraído directamente de la base de datos de la empresa.

Preprocesamiento:

En el análisis realizado a la data se observó que presentaba outliers por lo que se filtraron y eliminaron mediante el método IQR obteniendo con ello una data más equilibrada.

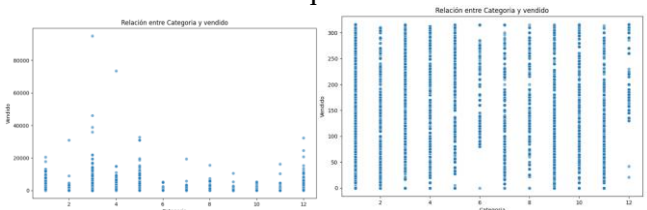

Figura 4. Filtración outliers – Modelo 3

Además se obtuvieron features derivados de la variables fecha como: dia de la semana, mes , se crearon lags y rollings_means ya que el problema trata sobre series de tiempo (similar al modelo 1) y para la categoría de realizo un one-hot encoding. El set train/test se obtuvo por medio de un Split cronológico considerando el 80/20 y para el modelo LSTM se aplicó Escalado MinMax a este set de datos.

Arquitectura y entrenamiento del modelo:

La naturaleza de este modelo es similar a la planteada en el modelo 1 ya que representan problemas de series de tiempo por lo que se utilizaron los algoritmos de XGBoost y una red LSTM, se entrenaron con las arquitecturas:

-XGBRegressor()

-LSTM con 50 neuronas en la capa de entrada y 1 neurona en la capa de salida, se entrenó a 50 épocas.

Evaluación y Selección del modelo

Las métricas obtenidas en el entrenamiento de los modelos fueron:

Tabla 3: Métricas de los modelos entrenados – Modelo 3

Modelo	MAE	RMSE	R^2
XGBoost	37.42	50.65	0.39
LSTM	46.34	60.08	0.14

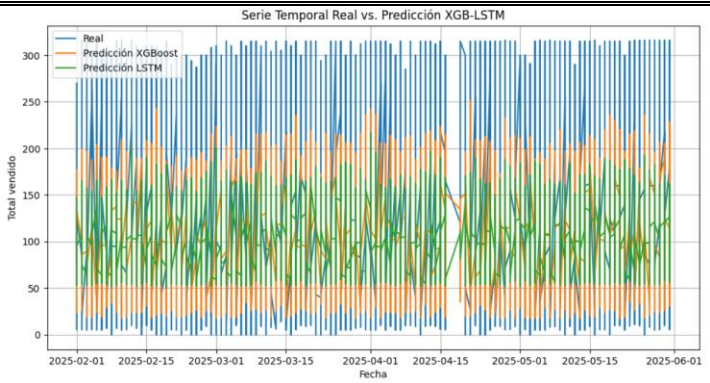

Figura 5. Comparación de predicciones – Modelo 3

Podemos observar que de acuerdo a sus métricas y predicciones el modelo que mejor interpreta los datos es XGBoost, su rendimiento es mejor comparado al Modelo 1.

3. Resultados

Los experimentos realizados evidenciaron lo siguiente:

1. **Modelo 1 (Inversión en Mercadería)**
   - **XGBoost** mostró el menor error (MAE: 9,176.83), aunque con un R^2 bajo (0.14), sugiriendo que factores externos no capturados en los datos podrían influir.
   - LSTM y MLP tuvieron peor rendimiento, posiblemente por el tamaño limitado del dataset (396 registros).
2. **Modelo 2 (Utilidad bruta)**:
   - **Regresión polinomial** superó a otros modelos (RMSE: 1,090.86 ; R^2: 0.41), indicando

una relación no lineal entre ventas y utilidad.
    - XGBoost tuvo el peor desempeño (R2*R*2: -0.24), probablemente por overfitting.
3. **Modelo 3 (Ventas por categoría)**:
    - **XGBoost** fue nuevamente el mejor (MAE: 37.42; R^2: 0.39), aprovechando features temporales (lags, medias móviles).
    - LSTM, pese a su capacidad para secuencias, no mejoró los resultados, posiblemente por requerir más datos.

Gráficos clave:

- Figuras 1, 3 y 5 muestran cómo las predicciones de XGBoost y regresión polinomial se alinean mejor con los valores reales.
- La eliminación de outliers (Figuras 2 y 4) redujo el ruido en los modelos 2 y 3.

4. Conclusiones

- XGBoost y regresión polinomial demostraron ser más robustos que LSTM en datasets pequeños, destacando la importancia de elegir algoritmos adecuados al volumen y naturaleza de los datos.

- Técnicas como lags, medias móviles y filtrado de outliers mejoraron significativamente la precisión, especialmente en series temporales

- La baja métrica R^2 en el Modelo 1 sugiere incluir variables externas (ej: eventos económicos, clima).

- Explorar ensembles (ej: combinar XGBoost con LSTM) o transformers para datasets más grandes.