Estimación de las ventas de La Favorita de Ecuador en el periodo previo a la pandemia COVID-19

Gabriel Fernando Franco Calvo
Jessica Torres Franco
Asesor: Yohany Pemberthy Salas

Especialización en Analítica y Ciencia de Datos
Facultad de Ingeniería
Universidad de Antioquia

RESUMEN: En el mundo de las tiendas retail, los pronósticos son fundamentales para equilibrar la oferta y la demanda. Los propietarios de estas tiendas deben gestionar cuidadosamente la cantidad de inventario que compran, evitando tanto el exceso de productos perecederos como la falta de artículos populares que puedan generar pérdidas de ingresos.

Palabras claves: Retail; series de tiempo, Redes neuronales, Regresión logística

ABSTRACT: In the world of retail stores, forecasts are essential for balancing supply and demand. Owners of these stores must carefully manage the quantity of inventory they purchase, avoiding both an excess of perishable products and a shortage of popular items that could result in revenue losses.

Keywords: Retail; time series, neural networks, logistic regression

1. INTRODUCCIÓN

En el mundo de las tiendas retail, los pronósticos son una herramienta crucial para mantener el equilibrio entre la oferta y la demanda. Los dueños de estas tiendas tienen que gestionar cuidadosamente la cantidad de inventario que se debe comprar para evitar quedarse con productos perecederos con exceso de existencias o con una falta de los artículos populares que generan una pérdida de ingresos.

Ante este desafío, es importante contar con herramientas efectivas que permitan hacer pronósticos precisos. Una de las posibles soluciones es el uso de modelos de recomendación de productos y modelos de series de tiempo para predecir el consumo en los meses posteriores.

Los modelos de recomendación de productos son una forma efectiva de determinar qué productos se venden juntos con más frecuencia, lo que permite hacer pronósticos más precisos sobre la demanda de los mismos. Estos modelos se basan en el análisis de datos de ventas y la identificación de patrones que permiten sugerir productos complementarios que pueden ser atractivos para los clientes.

Por otro lado, los modelos de serie de tiempo son una herramienta clave para estimar la demanda futura de los productos. Estos modelos se basan en el análisis de datos históricos de ventas y en la identificación de tendencias y patrones estacionales que permiten hacer pronósticos precisos sobre el consumo en los meses posteriores. De esta forma, las empresas pueden ajustar su inventario y asegurar el suficiente stock para satisfacer la demanda en el momento adecuado.

En resumen, los pronósticos son cruciales para las tiendas retail tradicionales, y contar con herramientas efectivas como modelos de recomendación de productos y modelos de tiempo pueden ayudar a evitar la pérdida de ingresos y clientes molestos por falta o exceso de inventario. Los tenderos deben estar siempre al tanto de

las últimas tendencias y tecnologías en el ámbito de los pronósticos para asegurarse de estar un paso adelante en este competitivo mercado.

2. MARCO TEÓRICO

Una serie de tiempo es una serie de puntos de datos ordenados en el tiempo. Durante un evento en una serie de tiempo, las medidas son organizadas típicamente en tiempos sucesivos (6). Gracias a esto, existe la posibilidad de una correlación entre las observaciones. En gran medida, el análisis de las series de tiempo tiene como objetivo explicar esta correlación y las principales características de los datos, usando modelos estadísticos y métodos descriptivos apropiados (7). En una serie de tiempo, el tiempo es a menudo la variable independiente y el objetivo suele ser hacer un pronóstico para el futuro. Se pueden extraer diversas características de las series de tiempo, como las tendencias y variaciones estacionales que pueden ser modeladas de forma determinista con funciones matemáticas del tiempo (7).

Sea Y_t una serie temporal en la que t denota el momento en que se toma la observación, donde $t \in Z^+$. El objetivo es construir un modelo que describa la evolución de la serie a través del tiempo, para esto se asume que los datos se pueden expresar como una función de una componente de tendencia T_t, estacional S_t y un error E_t (8).

Ruido Blanco

Un proceso ϵ_t se denota ruido blanco de media 0 y varianza σ^2 si satisface

$$E(\epsilon_t) = 0,\ Var(\epsilon_t) = \sigma^2 < \infty,\ Cov(\epsilon_t, \epsilon_{t-k}) = 0$$

Para todo $k \neq 0$. En particular, una sucesión de variables aleatorias independientes e idénticamente distribuidas, con media 0 y varianza σ_ϵ^2 representa un caso especial de un proceso de ruido blanco, y que se denota por $\epsilon_t \sim N(0, \sigma 2_\epsilon)$. Si además ϵ_t se distribuye normalmente, la serie se denomina ruido blanco gaussiano (8).

Hay tres formas de comprobar si la serie temporal se asemeja al ruido blanco:

- Trazando la serie temporal
- Comparando la media y la desviación estándar a lo largo del tiempo
- Examinando los gráficos de autocorrelación

Modelos AR(p)

Los modelos de autorregresivos de orden finito p, son una representación de un proceso aleatorio, en el que la variable de interés depende de sus observaciones pasadas. En general, para denotar el modelo autorregresivo AR se usa $AR(p)$ (5). Así, un modelo (AR) de orden p se puede escribir como

$$Y_t = \phi_0 + \phi_1 y_{t-1} + \phi_2 y_{t-2} + ... + \phi_p y_{t-p} + \epsilon_t \quad (1)$$

para constantes $\phi_0, ..., \phi_p$ y $\epsilon_t \sim RB(0, \sigma^2)$.

Modelos MA(q)

Los modelos de medias móviles de orden finito q, son una aproximación común para las series de tiempo univariadas. El modelo de medias móviles especifica que la variable de salida depende linealmente del valor actual y varios de los anteriores. En general para denotar un modelo de medias móviles MA se usa $MA(q)$ (5). Así, un modelo MA de orden q se puede escribir como

$$Y_t = \theta_0 + \theta_1 \epsilon_{t-1} + ... + \theta_q \epsilon_{t-q} \quad (2)$$

para constantes $\theta_1, ..., \theta_q$ y $\epsilon_t \sim RM(0, \sigma^2)$.

Modelo ARMA

un proceso $ARMA(p, q)$ es un modelo que combina las propiedades de memoria larga de los $AR(p)$ con las propiedades de ruido débilmente autocorrelacionado en los $MA(q)$, y que tiene suficiente flexibilidad y parsimonia para representar una variedad grande de procesos estacionarios en covarianza.

$$Y_t = \phi_0 + \sum_{i=1}^{p} \phi_i y_{t-i} + \sum_{j=0}^{q} \theta_j \epsilon_{t-j} \quad (3)$$

donde $\epsilon_t \sim RB(0, \sigma^2)$ (5).

Modelo SARIMAX

Un proceso ARIMA estacional o SARIMA, utiliza la diferenciación con un retardo igual al número de estaciones (s) para eliminar los efectos estacionales aditivos. El modelo SARIMA tiene una componente ARIMA(P, D, Q) que modeliza la dependencia estacional, que está asociada a observaciones separadas por l periodos y contiene otra componente ARIMA(p, d, q) que modeliza la dependencia regular, que es la dependencia asociada a observaciones consecutivas y tambien tiene un proceso diferenciado $w_t = \nabla_l^D \nabla^d Y_t$ es un proceso estacionario que sigue el modelo ARMA estacional. Por tanto la ecuación general del modelo SARIMA es

$$\Phi_P(B^l)\phi_p(B)w_t = \Theta_Q(B^l)\theta_q(B)E_t \quad (4)$$

Con $E_{t t\in Z}$ un $RB \sim (0, \sigma^2)$ (5).

El modelo **SARIMAX** se define como

$$y_t = \beta_t x_t + u_t \tag{5}$$

$$\Phi_p(L)\tilde{\phi}_p(L^s)\Delta^d\Delta_s^D u_t = A(t) + \theta_q(L)\tilde{\theta}_Q(L^s)\zeta_t \tag{6}$$

donde β representa las variables externas, para el resto del modelos es similar al modelo SARIMA, donde

- p para el orden AR
- q para el orden MA
- I para el orden diferenciador
- P para el orden estacional AR
- Q para el orden estacional MA
- D para el diferenciador estacional
- s para los coeficientes estacionales

2.1. Métricas

La métrica utilizada para la evaluación del desempeño del modelo es el **RMSLE** (Root Mean Squared Logarithmic Error) propuesta por el negocio, sin embargo se utilizarán métricas adicionales.

2.1.1. RMSE

$$RMSLE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\log(1+\hat{y_i}) - \log(1-y_i))^2} \tag{7}$$

2.1.2. AIC y BIC

AIC (Crtiterio de Información Akaike) y BIC (Criterio de Información Bayesiano) son medidas estadísticas usadas para selección y comparación de modelos. Son usadas para la evaluación de la bondad de ajuste de distintos modelos y ayudar a seleccionar el modelo más apropiado de acuerdo a la información usada.

$$AIC = -2 * log - likehood + 2 * k \tag{8}$$

$$BIC = -2 * log - likelihood + log(n) * k \tag{9}$$

donde,

- log-likehood: Logaritmo natural de la función de probabilidad del modelo.
- k: Número de parámetros estimados en el modelo
- n: Número de observaciones en el dataset

2.1.3. MAE

MAE (Error Absoluto Medio) es una métrica usada para evaluar la precisión de un modelo de predicción. Mide la diferencia entre el valor predicho y el valor real

$$MAE = \frac{1}{n}\sum |y\_pred - y\_actual| \tag{10}$$

donde,

- **y_pred:** Valores predichos por el modelo
- **y_actual:** Valores reales
- **n:** Número de observaciones en el dataset

2.1.4. MSE

MSE (Error Cuadratico Medio) es una métrica usada para evaluar la precisión de un modelo de predicción. Mide la diferencia entre el valor predicho y el valor real

$$MSE = \frac{1}{n} * \sum (y\_pred - y\_actual)^2 \tag{11}$$

donde,

- **y_pred:** Valores predichos por el modelo
- **y_actual:** Valores reales
- **n:** Número de observaciones en el dataset

2.1.5. HQIC

HQIC (Criterio de Información de Hannan-Quinn) es un criterio de selección y comparasión en análisis de series de tiempo. Es una extensión de los criterios AIC y BIC.

El criterio incorpora tanto la bondad de ajuste como la complejidad del modelo penalizando el número de parámetros del modelo. EL objetivo es un balance entre los dos factores, sin embargo presenta una mayor penalización en los modelos que los criterios AIC y BIC.

$$HQIC = -2 * log - likelihood + 2 * k * log(log(n)) \tag{12}$$

- log-likehood: Logaritmo natural de la función de probabilidad del modelo.
- k: Número de parámetros estimados en el modelo
- n: Número de observaciones en la serie de tiempo

3. RESULTADOS

Para la evaluación de los modelos, se lleva a cabo la implementación de un modelo SARIMAX mediante una evaluación de una rejilla de hiperparámetros asociados al modelo. El objetivo es obtener resultados óptimos que mejoren la capacidad predictiva del modelo.

A través de la implementación de una rejilla de búsqueda exhaustiva, se evalúan todas las combinaciones posibles de hiperparámetros. Para cada combinación, se entrena el modelo SARIMAX y se evalúa su desempeño utilizando métricas relevantes, como el MSE, AIC, BIC, MAE, RMSLE. De esta manera, se busca encontrar la configuración de hiperparámetros que produzca los mejores resultados en términos de precisión y capacidad de predicción.

Tras el análisis exhaustivo, se obtienen los resultados correspondientes a cada combinación de hiperparámetros evaluados

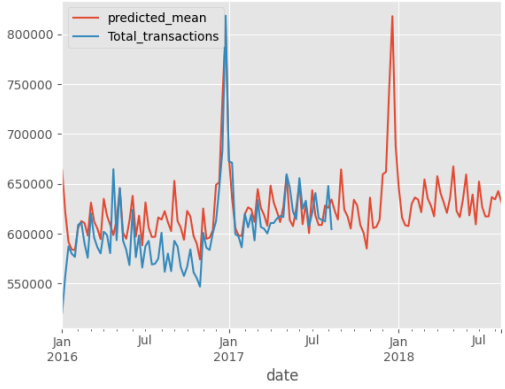
Parámetros	AIC	BIC	HQIC
(2, 0, 1, 0, 1, 1, 52)	2278.05	2291.27	2283.4
(2, 0, 1, 0, 1, 0, 52)	2280.02	2290.6	2284.31
(1, 0, 1, 1, 1, 1, 52)	2283.83	2297.05	2289.19
(2, 0, 2, 0, 1, 0, 52)	2284.86	2298.08	2290.22
(2, 0, 0, 1, 1, 1, 52)	2285.81	2299.04	2291.17
(2, 0, 1, 1, 1, 1, 52)	2285.94	2301.8	2292.37
(1, 0, 2, 0, 1, 0, 52)	2286.12	2296.7	2290.4
(2, 0, 2, 1, 1, 1, 52)	2287.46	2305.97	2294.96
(1, 0, 0, 1, 1, 1, 52)	2296.85	2307.43	2301.14
(0, 1, 1, 0, 1, 0, 52)	2330.82	2336.09	2332.96

Evaluación hiperparámetros SARIMAX

Luego de una evaluación de los hiperparámetros se escogen 3 modelos debido a su parsimonia y que presentan los menores valores, los modelos y sus hiperparámetros son SARIMAX(2,0,1)(0,1,1), SARIMAX(2,0,1)(0,1,0) y SARIMAX(1,0,2)(0,1,0)

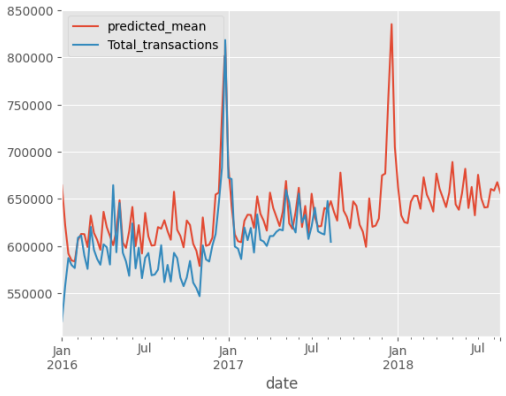
3.0.1. SARIMAX(2,0,1)(0,1,1)

El modelo presenta unas métricas de RMSE de 35001.47 y un RMSLE de 0.075



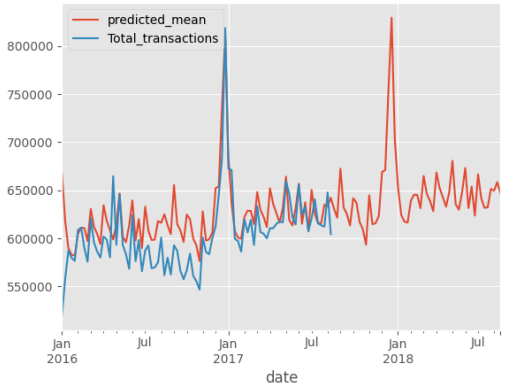
3.0.2. SARIMAX(2,0,1)(0,1,0)

El modelo presenta unas métricas de RMSE de 36620.05 y un RMSLE de 0.077



3.0.3. SARIMAX(1,0,2)(0,1,0)

El modelo presenta unas métricas de RMSE de 37765.82 y un RMSLE de 0.079



3.0.4. Comparación de modelos

Como se puede observar cada una de las gráficas anteriores, los modelos son capaces de poder obtener patrones y tendencias del problema planteado inicialmente que es predecir ventas futuras, la diferencia entre cada uno de los modelos son mínimas y no se logran observar diferencias signficativas entre los modelos.

Modelos SARIMAX	RMSE	RMSLE
(2, 0, 1, 0, 1, 1, 52)	35001.47	0.075
(2, 0, 1, 0, 1, 0, 52)	36620.05	0.077
(1, 0, 2, 0, 1, 0, 52)	37765.82	0.079

De acuerdo a las mediciones con la métrica del negocio (RMSLE) y una métrica adicional (RMSE), el mejor modelo que se ajusta a los datos es el modelo SARIMAX(2,0,1)(0,1,1) con un RMSE promedio de 35001.47 y un RMSLE promedio de 0.079. Estas métricas fueron obtenidas mediante el uso de validación cruzada.

4. CONCLUSIONES Y RECOMENDACIONES

En conclusión, la implementación del modelo SARIMAX a través de una evaluación de una rejilla de hiperparámetros permite mejorar la capacidad predictiva de este modelo en el análisis de series temporales. Los resultados obtenidos proporcionan una guía para seleccionar la configuración óptima de hiperparámetros y, así, optimizar el rendimiento del modelo en futuras predicciones y análisis de datos. Adicionalmente el modelo escogido es un modelo SARIMAX(2,0,1)(0,1,1) que posee las mejores métricas con respecto a los otros modelos.

En cuanto a los trabajos futuros, es recomendable realizar un estudio exhaustivo del comportamiento de las ventas tanto durante la pandemia como en el período postpandemia. Este análisis permitirá realizar una calibración adecuada del modelo existente o, en caso de que el modelo actual no sea apropiado, identificar y considerar los efectos posteriores a la pandemia en las proyecciones de ventas.

Además, sería beneficioso llevar a cabo una categorización de los distintos productos y agregarlos al modelo. Esto proporcionaría a La Favorita la capacidad de generar estimaciones más precisas y detalladas para cada tipo de producto. La incorporación de esta información adicional permitirá un análisis más completo y ayudará a la empresa a tomar decisiones estratégicas más informadas.

Referencias

[1] Adhikari, R. and R. Agrawal (2013). An Introductory Study on Time Series Modeling and Forecasting. Lap Lambert Academic Publishing GmbH KG.

[2] Scikit-learn: Machine Learning in Python, Pedregosa et al., JMLR 12, pp. 2825-2830, 2011.

[3] Hyndman, R. J., & Athanasopoulos, G. (2018). Forecasting: Principles and Practice. OTexts. Chapter 2: Time series graphics. Available at: https://otexts.com/fpp2/accuracy.html

[4] Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction (2nd ed.). Springer. Chapter 7: Model Assessment and Selection.

[5] Giraldo Gómez, N. D. (2006). Técnicas de pronósticos : aplicaciones con R

[6] Adhikari, R. and R. Agrawal (2013). An Introductory Study on Time Series Modeling and Forecasting. Lap Lambert Academic Publishing GmbH KG

[7] Paul S. P. Cowpertwait, A. M. (2009, June). Introductory Time Series with R. Springer-Verlag GmbH

[8] Jonathan D. Cryer, K.-S. C. (2009, October). Time Series Analysis with Applications in R. SpringerVerlag.

[9] Korstanje, J. (2021). The SARIMA model. Advanced Forecasting with Python, 115–122. https://doi.org/10.1007/978-1-4842-7150-6_7

[10] Pinedo Chapa, Joely Mireilli. Propuesta de un modelo de pronósticos de demanda y gestión de inventarios para la planeación de demanda en prendas de vestir juvenil. Edu.pe. Recuperado el 21 de abril de 2023, del Repositorio.