Venta de televisores: análisis de series de tiempo

Inteligencia Artificial avanzada para la ciencia de datos - TC3007C

Grupo 501

Portafolio de Análisis - Módulo 5

Cristofer Becerra Sánchez - A01638659

***Resumen*—El presente reporte analiza las ventas de televisores de un histórico de 4 años, descomponiendo los datos en las diferentes componentes que constituyen una serie de tiempo. Se calculó el vector de valores irregulares de la serie para encontrar los índices estacionales que toman valores de 0.9322, 0.8378, 1.0933, 1.1433 para cada trimestre del año respectivamente. Utilizando estos índices se calculó la serie desestacionalizada, es decir, la tendencia de la serie. Se encontró un modelo lineal ajustado por mínimos cuadrados de la forma $y = 5.0996 + 0.1471x$. Se calcularon métricas de error para el modelo lineal encontrado. Finalmente, a partir de estos hallazgos se realizó un pronóstico de ventas del año siguiente (quinto): se prevé que el año siguiente se vendan 7,086 televisores en el primer trimestre, 6,491 en el segundo, 8,632 en el tercero, y 9,195 en el cuarto.**

***Index Terms*—Series de Tiempo, , Ventas, Televisores, TV, Television, Vender**

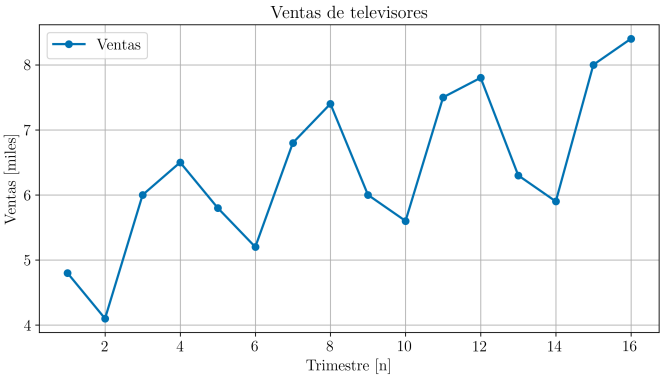

Figura 1. Gráfico de las ventas de televisores en miles. El eje x representa el número total de trimestres registrados, por eso es un número ascendente y no cíclico. El eje y representa la venta de televisores en miles de unidades.

I. Introducción

Es de suma importancia remarcar que partir de un histórico de ventas es posible extraer mucha información valiosa para un negocio. En particular, se pueden entender a profundidad los comportamientos de las ventas y extraer relaciones con otras variables como costos de producción, marketing, o externalidades. Además, y quizá aún más importante, el análisis de estos datos pasados permite realizar pronósticos de ventas a futuro. Toda esta información resulta ser de valor para cualquier negocio pues puede acercar a un negocio a una mejor toma de decisiones. El presente reporte analiza las ventas de televisores de un histórico de 4 años, descomponiendo los datos en las diferentes componentes que constituyen una serie de tiempo; a partir de estos hallazgos se realiza un pronóstico de ventas del año siguiente (quinto).

II. Resultados y análisis

II-A. Visualización de la serie

Se comienza el análisis visualizando el conjunto de datos en cuestión, es decir, el vector de datos con las ventas de televisores a lo largo de los cuatro años registrados. En la figura 1 se ilustra claramente que los datos representan una serie de tiempo.

Cristofer Becerra Sánchez, pertence al Tec de Monterrey Campus Monterrey, Monterrey, Nuevo León, Mexico, C.P. 64849

A primera instancia es evidente la naturaleza cíclica de los datos; además, se aprecia un patrón de crecimiento. Es decir que los datos no sólo tienen un período fijo, sino que los valores dentro de estos ciclos aumentan con respecto del tiempo. Con estas observaciones puede decirse que se trata de una serie de tiempo no estacionaria o serie de tiempo con tendencia.

II-B. Análisis de tendencia y estacionalidad

El análisis comienza descomponiendo la serie en sus irregularidades. Para hacerlo, primero se debe calcular la media móvil de la serie; en este caso se escogió una media móvil con ventana de cuatro (4) trimestres. Una vez calculado este nuevo vector de datos se procede a suavizar la serie aún más aplicando otra media móvil; en esta ocasión, sin embargo, la media móvil de la media móvil de la serie se calcula con una ventana de dos trimestres. A este vector que representa un segundo suavizado de la serie se le llama media móvil centrada de la serie. La figura 2 representa los dos vectores calculados: la media móvil y la media móvil centrada.

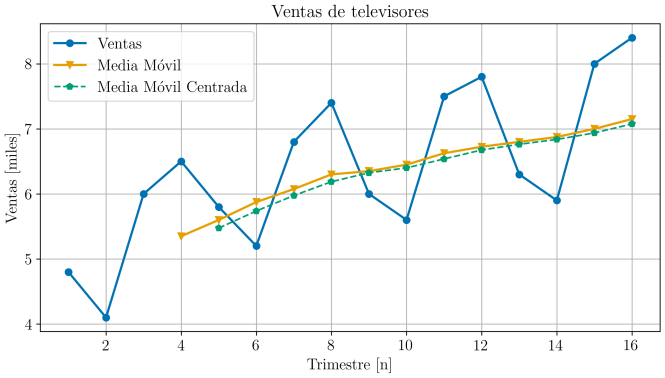

Figura 2. Gráfico de las ventas de televisores en miles y las dos curvas de la serie suavizada; en específico, se incluye la media móvil y la media móvil centrada de la serie.

Es notable el efecto de la ventana sobre la serie, y se resalta el hecho de que el tamaño de la ventana es igual al período de la serie (cuatro trimestres por año). Además, también se aprecia que, a pesar del minúsculo efecto del segundo suavizado, la curva se centra más en la serie y desciende un poco con respecto al primer suavizado; esto tiene sentido ya que los valores van aumentando con el tiempo y la ventana del suavizado jala los nuevos puntos hacia abajo.

Ahora, si se divide el vector de ventas y la media móvil centrada se obtendrán los valores estacionales irregulares de la serie. A partir de estos valores irregulares se extrae el índice estacionario al identificar el trimestre al que pertenece cada valor irregular. Se obtienen los promedios de valores irregulares por trimestre, es decir, el valor del componente estacional irregular; estos promedios representan los índices estacionales.

Habiendo calculado los índices estacionales, sólo falta dividir el valor de cada venta por su respectivo índice estacional (dependiendo del trimestre de la venta) para obtener la tendencia de la serie de tiempo en cuestión. Se visualiza la tendencia calculada en la figura 3

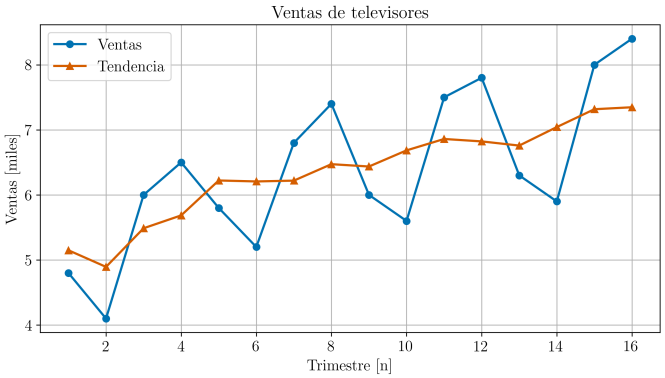

Figura 3. Gráfico de las ventas de televisores en miles y la tendencia de la serie calculada mediante la extracción de los valores irregulares y los índices estacionales consecuentes.

Es evidente que el vector resultante que representa la tendencia tiene un comportamiento mucho menos suave que las medias móviles calculadas anteriormente; no obstante, se debe hacer una regresión lineal para obtener la función lineal que constituye la tendencia.

II-C. Regresión lineal

Se prosigue el análisis realizando una regresión lineal con respecto al vector de datos calculados para la tendencia de la serie. El procedimiento por mínimos cuadrados entrega el siguiente modelo lineal:

$$y = 5.0996 + 0.1471x; \tag{1}$$

es decir que $\beta_0 = 5.0996$ y $\beta_1 = 0.1471$. Ambos coeficientes son significativos con un p-value de 0 redondeado a 3 decimales. La representación visual de este modelo puede apreciarse en la figura 4.

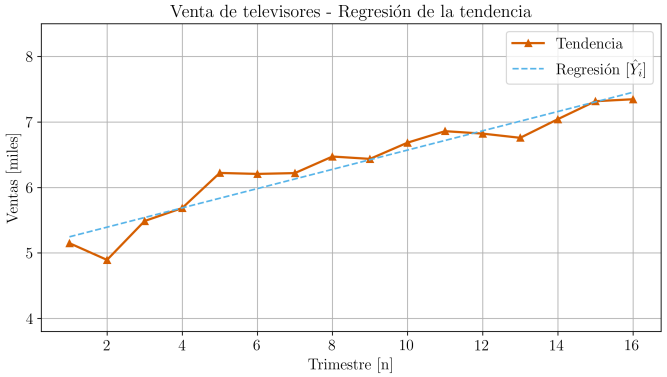

Figura 4. Serie de tiempo desestacionalizada con el modelo de regresión lineal obtenido.

Una vez obtenido un modelo lineal que representa la tendencia de la serie, se debe verificar la validez de la herramienta estadística; en particular, se deben corroborar los supuestos del modelo.

II-D. Verficación del modelo

Se empieza por corroborar la significancia del coeficiente β_1 que otorga la tendencia lineal. Se calcula el estimador del coeficiente,

$$\hat{\beta}_1 = \frac{\text{cov}(x, y)}{s_x^2}$$

y se procede a calcular su respectivo estadístico de prueba t^*,

$$t^* = \frac{\hat{\beta}_1 - \beta_1}{s_{\hat{\beta}_1}} = 12.3522$$

que arroja un p-value de $p = 6.447 \times 10^{-9}$; ya el p-value asociado es sumamente pequeño, es posible desechar la hipótesis nula de que el coeficiente $\beta_1 = 0$. Así, se concluye que el coeficiente β_1 de la regresión es significativo. Con este supuesto demostrado, se continúa con el análisis de los residuos, empezando por su distribución.

Se realiza una prueba de normalidad de Shapiro-Wilk sobre los residuos para demostrar que siguen una distribución normal, tal como indica el supuesto de normalidad. La muestra

de los residuos indica un estadístico de prueba $W = 0.9638$ con un respectivo p-value de $p = 0.7307$. Se obtiene un valor bastante mayor a un nivel de significancia de $\alpha = 0.05$, por lo tanto, no se rechaza la hipótesis nula que establece que los residuos siguen una distribución normal. Dicha conclusión puede visualizarse en la figura 5 que ilustra tanto el diagrama cuantil-cuantil de los residuos como su distribución.

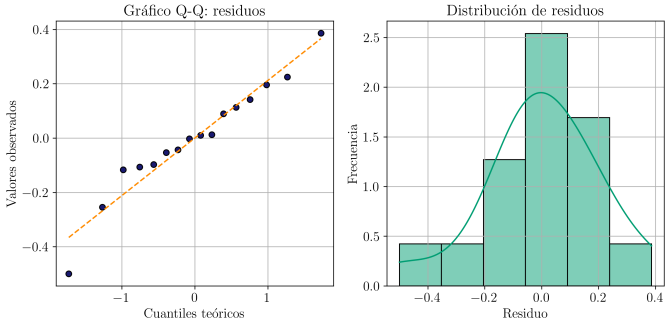

Figura 5. (a) Gráfico cuantil-cuantil de los residuos de la regresión lineal de la tendencia. (b) Histograma de la distribución de los residuos.

Se observa una curva de densidad muy similar a la de una distribución normal, con una cola izquierda ligeramente anormal; también el gráfico Q-Q ilustra esta cuestión de la cola izquierda, pero se aprecia mejor el buen ajuste a una normal al observar la poca diferencia entre los puntos y la línea teórica. En seguida se debe comprobar que la media de los residuos no es significativamente diferente de cero mediante otra prueba de hipótesis. Se emplea una prueba t de Student de una sola muestra con la condición de $\mu = 0$; la media muestral calculada numéricamente tiene un valor de $\bar{X} = -1.8873 \times 10^{-15}$. La prueba arroja un p-value de $p = 0.99999$ con un estadístico de prueba $t = -3.6744 \times 10^{-14}$; con un p-value tan cercano a 1, no es posible rechazar la hipótesis nula que indica que la media de los residuos es igual a cero. Se compureba este resultado con la condición de homocedasticidad en la figura 6 ya que se ilustra la media cero de los residuos y su independencia de manera simultánea.

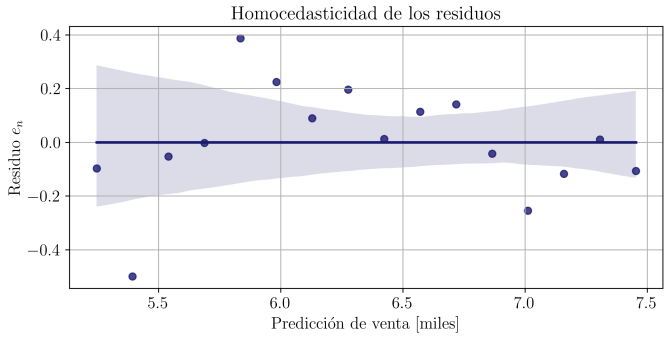

Figura 6. Ilustración de la varianza de los residuos. El eje x representa la predicción del modelo y el eje y representa el error obtenido a partir de esa predicción. Se observa claramente una media cero, una aparente simetría de los residuos y una aparente independencia.

Finalmente se revisa el coeficiente de determinación R^2 que indica la variabilidad explicada por el modelo. Se obtiene un coeficiente de determinación $R^2 = 0.9208$ y un coeficiente de determinación ajustado de $R^2_{\text{adj}} = 0.9151$. Ambos valores indican una alta variabilidad explicada por la regresión, por lo tanto, se concluye la verificación del modelo de manera exitosa concluyendo que el modelo es válido.

II-E. Errores de la predicción

A partir del modelo lineal calculado, se realizó una predicción de la serie de tiempo empleando los índices estacionales correspondientes. Se despliega la comparación de la serie real con las predicciones del modelo en la figura 7.

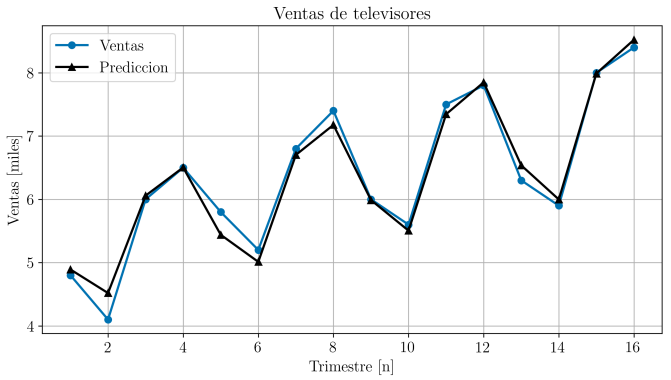

Figura 7. Gráfico de las ventas de televisores en miles y las predicciones calculadas a partir del modelo de regresión lineal interpolando los respectivos índices estacionales de los trimestres.

A partir de estas predicciones, se calcularon dos métricas de evaluación del modelo, a saber, el cuadrado medio del error y el promedio de los errores porcentuales. Se obtuvo un $MSE = 0.033$ y un $MAPE = 0.0244$. El promedio de los errores porcentuales indican que, en promedio, el error porcentual del modelo es el 2.44 % lo cual es un valor excelente considerando la simplicidad del modelo.

II-F. Pronóstico siguiente año

De forma similar a lo anterior, se realiza un pronóstico del año siguiente a partir del modelo calculado. De este modelo se prevé que el próximo año se vendan 7,086 televisores en el primer trimestre, 6,491 en el segundo, 8,632 en el tercero, y 9,195 en el cuarto. Este pronóstico de ventas se plasma en la figura 8.

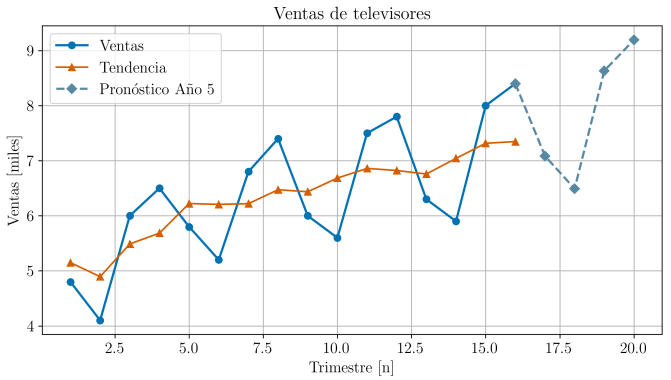

Figura 8. Gráfico de las ventas de televisores en miles, la tendencia de la serie y el pronóstico del siguiente año calculado utilizando el modelo de regresión lineal con los indices estacionales.

III. Conclusión

En resumen, se analizó un histórico de ventas de televisores de una empresa desconocida mediante un enfoque de series de tiempo. Se desglosaron los datos en sus constituyentes: irregularidades, estacionalidad y tendencia. A partir de estos elementos se calculó un modelo con un $MSE = 0.033$ y un $MAPE = 0.0244$ bastante buenos. En fin, se vuelve a resaltar la importancia de los datos y herramientas estadísticas y de análisis para mejorar la toma de decisiones de los negocios e incluso otras entidades o instituciones que requieran de un análisis similar.

Anexo

Enlace al repositorio de GitHub, *VentasTVs*, con todos los archivos del proyecto (Jupyter Notebook, Notebook en formato .py, la base de datos utilizada, y el presente documento PDF): https://github.com/crisb-7/VentasTVs.