Tecnológico de Monterrey

Actividad:

Módulo 2: Análisis y Reporte sobre el desempeño del modelo

Grupo:

TC3006C.101

Equipo:

Franco Quintanilla Fuentes - A00826953

Profesor:

Ivan Mauricio Amaya Contreras

Fecha:

12 de septiembre de 2022

Introducción

En este entregable se hace la implementación de una técnica de aprendizaje máquina (ML) sin el uso de alguna librería de aprendizaje máquina. En este caso se optó por usar la Regresión Lineal en base a las funciones lambda, tanto para las pruebas de hipótesis como las funciones de costo.

Desarrollo

Lo primero que hicimos, fue importar y visualizar los datos, los cuales se ven de la siguiente manera.

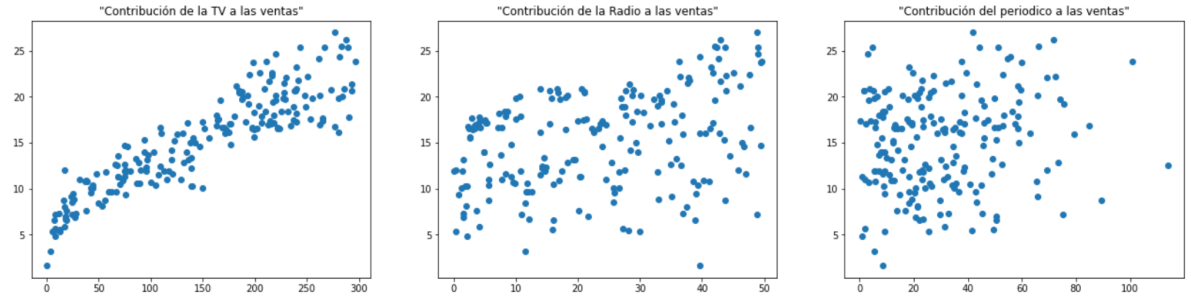

Figura 1: Tendencia de los datos.

Para este caso, después de visualizar el comportamiento de los datos, utilizamos la regresión lineal simple en base a la Contribución de la TV a las ventas, ya que las demás contribuciones no siguen una línea de tendencia tan significativa.

Para poder hacer el análisis de nuestra técnica, dividimos nuestro dataset en 80% datos para entrenamiento y 20% datos para la evaluación. En donde verificamos la tendencia de los datos seleccionados de manera aleatoria.

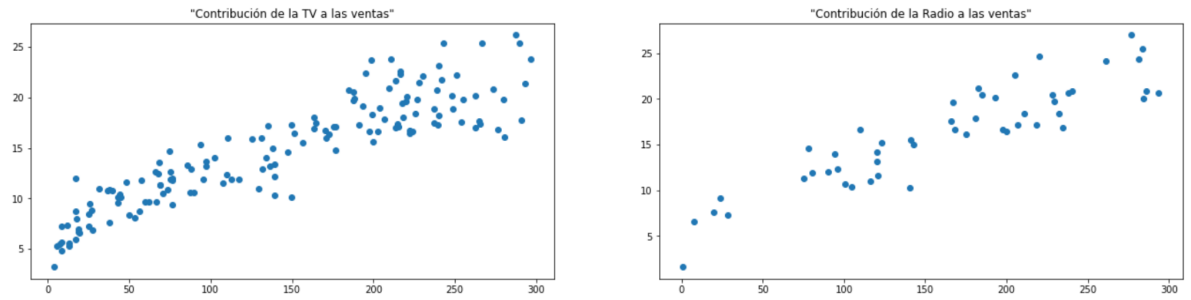

Figura 2: Comportamiento de la contribución de la TV a las ventas en 80% train, 20% test.

Para los casos iniciales, después de un poco de investigación, y de estar a prueba y error, los parámetros iniciales fueron los siguientes para nuestro modelo:

- alpha = 0.001
- theta = [7, 0.05]

En donde nuestro modelo trata de buscar el mejor "fit" para nuestros datos del modelo. Consecuentemente, graficamos la línea de tendencia que sigue nuestro modelo de Regresión Lineal, junto a los datos de nuestro dataset para que la visualización de la tendencia fuera más clara, junto con la regresión lineal del paquete de seaborn, para ver si tienen similitudes, las cuales si las tienen.

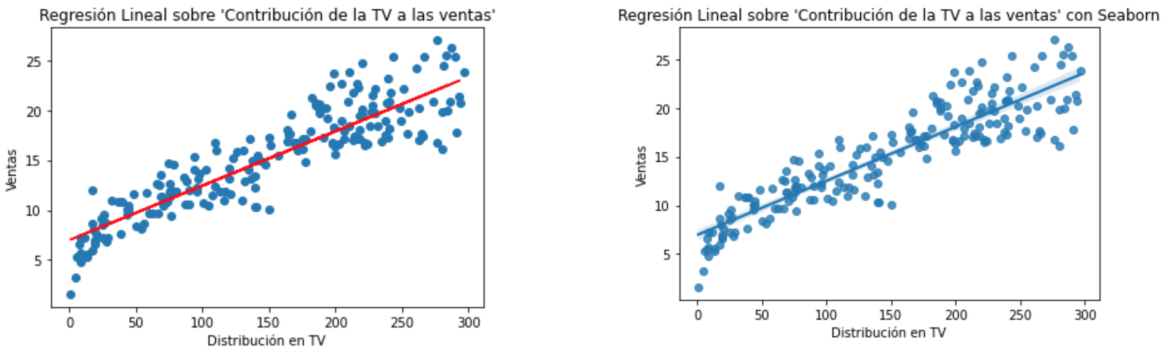

Figura 3: Comparación de nuestro modelo, con el modelo del paquete de Seabron.

Por último, hicimos algunas predicciones con los valores de la ecuación de nuestro modelo con el 20% de los datos de la evaluación, para después poder sacar los valores de:

- Coeficiente de determinación: 0.811
- Error cuadrático medio: 5.762

Análisis de datos

Después de hacer el análisis de los datos, hicimos el análisis de las siguientes características:

- Sesgo
- Varianza
- Underfitting o Overfitting

En donde obtuvimos los siguientes resultados.

Sesgo

Para el sesgo, graficamos los datos del dataset, en donde el comportamiento es de la siguiente manera.

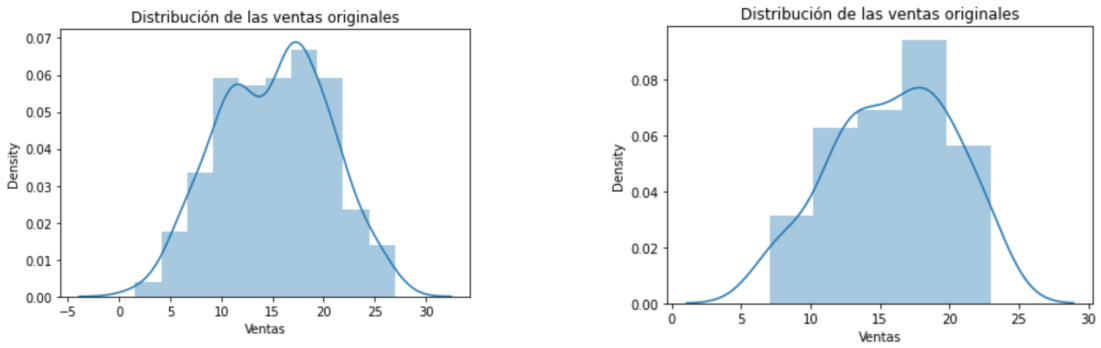

Figura 4: Comparación del sesgo del data frame con nuestra predicción de las ventas.

Varianza

Para la varianza, obtenemos la varianza de los datos para el test, como la varianza de nuestros resultados de la predicción, y obtuvimos los siguientes resultados.

- La varianza de y_test es 27.919
- La varianza de y_pred es 18.776

En donde podemos ver que nuestra predicción tiene menos varianza, por lo que es un buen modelo de regresión lineal, y los datos se ajustan de manera correcta.

Underfitting o Overfitting

Para saber si nuestro modelo es overfitted o underfitted, graficamos el **learning curve**, que en nuestro caso, nos queda de la siguiente manera.

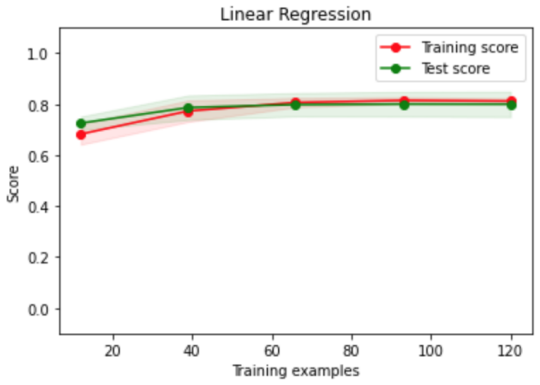

Figura 5: Curva de aprendizaje de nuestro modelo de regresión lineal simple.

Conclusión

A lo largo del análisis y la implementación de este modelo, podemos ver que nuestro modelo se comporta de buena manera al modelo de regresión lineal simple tomando en cuenta el aumento de las ventas a la hora de la contribución de los anuncios/publicidad hecha en la TV, vemos que nuestro modelo es bueno para la regresión y aprende con su entrenamiento y se comporta bien a la hora del test, es decir que se ajusta bien y que no cae en el underfitting ni en el overfitting, si no, que es óptimo y nos determina el comportamiento del modelo con un coeficiente de determinación del 0.811, lo cual es excelente.

Anexo

Dataset:

https://www.kaggle.com/code/ashydv/sales-prediction-simple-linear-regression/notebook