Tecnológico de Monterrey

Instituto Tecnológico y de Estudios Superiores de Monterrey

Reporte completo de "El precio de los autos"

TC3004B.104 Inteligencia Artificial Avanzada para la Ciencia de Datos I

Profesores:

Ivan Mauricio Amaya Contreras

Blanca Rosa Ruiz Hernandez

Antonio Carlos Bento

Frumencio Olivas Alvarez

Hugo Terashima Marín

Julian Lawrence Gil Soares - A00832272

12 de Septiembre de 2023

Resumen:

Para nuestro reto del módulo de estadística se nos pidió que utilizaramos una base de datos conteniendo datos de algunos elementos de automóviles para generar predicciones acerca de sus precios. Para la solución utilice varias herramientas de estadística para poder brindar las mejores predicciones posibles a través de el uso de significancia y correlación para mi selección de variables utilizando, también de ciertas métricas de desempeño como el MSE y el valor de R^2 y un análisis de residuos para determinar si mi modelo cumple con los principios de homocedasticidad y media zero.

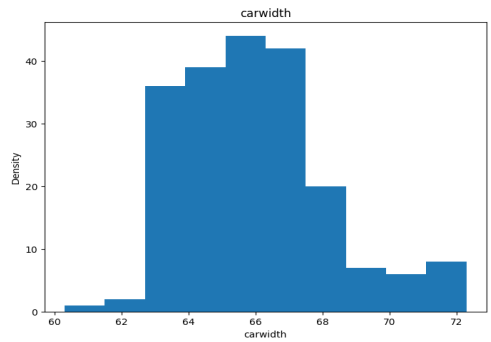
Introducción:

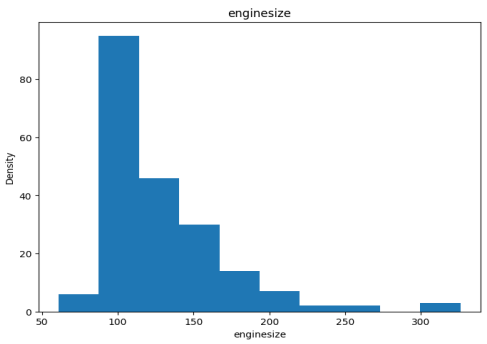
El problema a resolver es simple. ¿Utilizando los features de un automóvil se puede determinar su precio? ¿ Cuáles son las variables que se deben tomar en cuenta para esta predicción? El usar elementos de estadística e inteligencia artificial para realizar predicciones de todo tipo se está volviendo cada día más una comodidad que las empresas necesitan para ser competitivas en sus respectivos campos. Estas predicciones pueden ser de todo tipo, desde predicciones de factores ambientales hasta predicciones de desempeño y de precio. Dentro de la industria automotriz el uso de inteligencia artificial está impulsando a manufactureros y vendedores para tomar decisiones inteligentes en base a datos acumulados tras décadas. Específicamente el precio de los automóviles se está determinando usando predicciones basadas en información del mercado y de las maneras de ser de los consumidores(J.D. POWER, 2023).

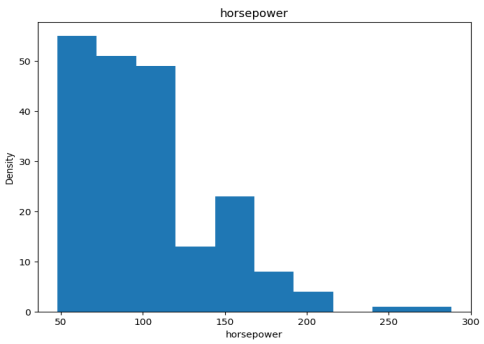
Análisis de los resultados:

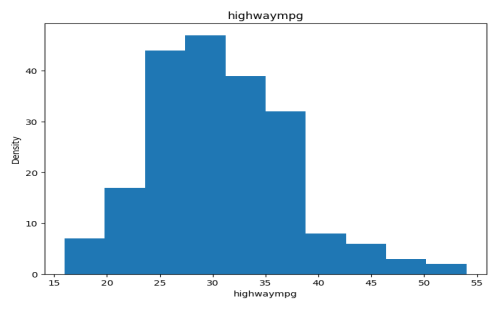
El primer paso en este proyecto fue determinar cuales son las variables que voy a utilizar para mi predicción. Se nos pidió que utilizaremos solamente 6 variables así que tuve que buscar la manera de determinar cuales son las mejores variables para mi problema. Mi primer paso fue observar a través de gráficas como boxplot e histogramas. Mi objetivo con esto fue determinar si los datos tienen buena simetría y si tienen una distribución normal.

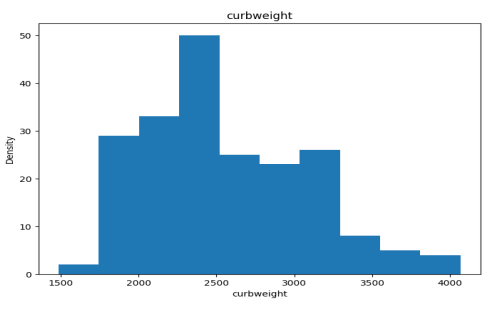
Histogramas.

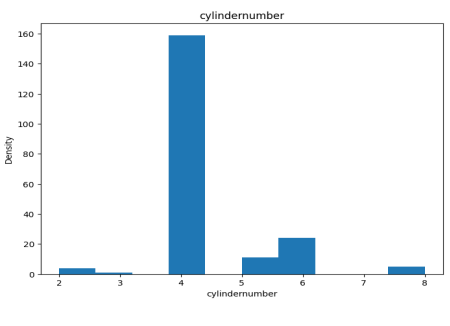




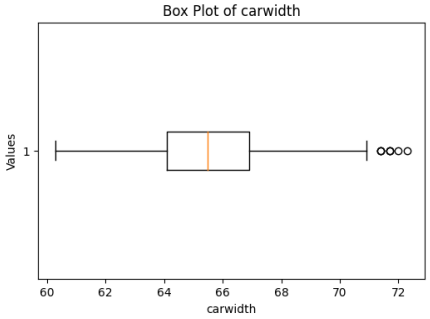


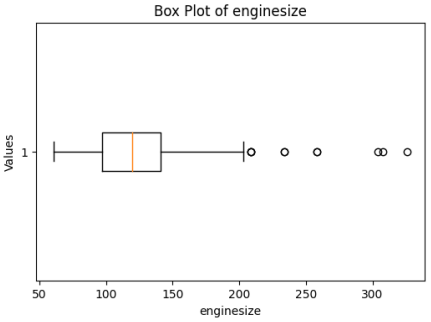


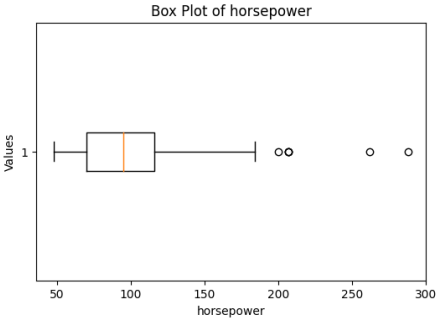


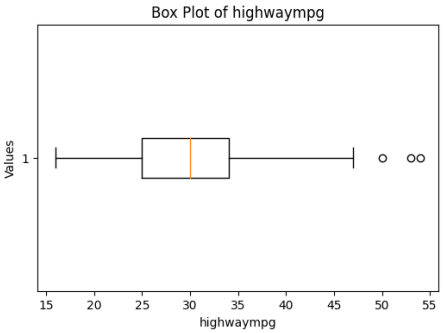


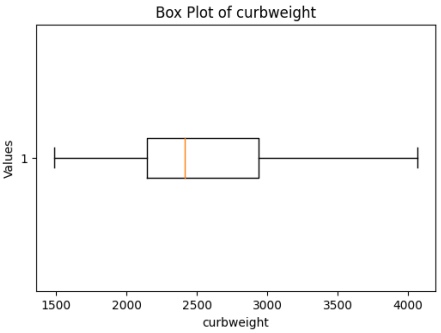
Boxplot





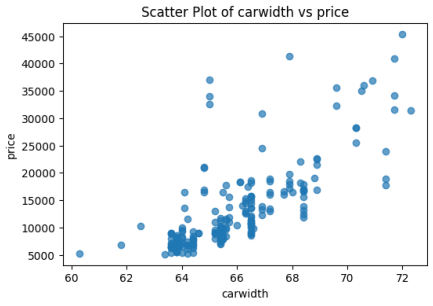


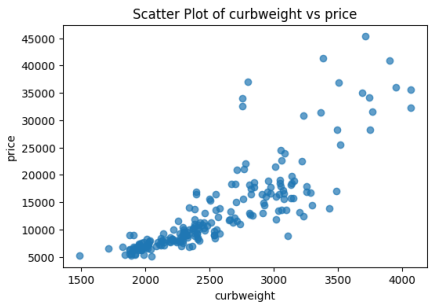


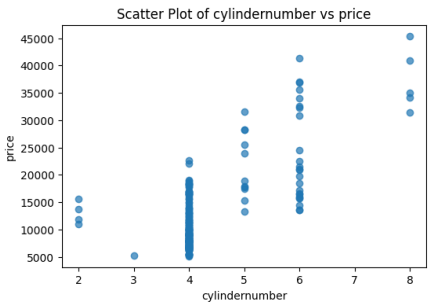


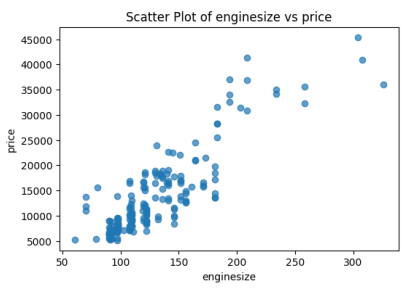
La importancia de este análisis es para ver si nuestros datos van a requerir alguna transformación por datos atípicos o si ya cuentan con una distribución normal. Como se puede observar algunas variables como: horsepower, enginesize, highwaympg y carwidth tienen datos que se desvían mucho de la media y estos podrían terminar influyendo negativamente en las predicciones. Después de observar esto busque la correlación entre las variables para poder determinar cuales son las que más influyen en el precio. Programe una función para solo mostrar las variables que tienen una correlación mayor a .7 y elegí las variables que tuvieran los mejores resultados.

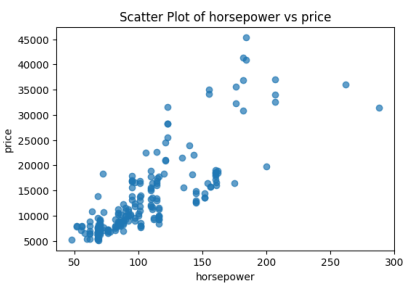
Pruebas de correlación.



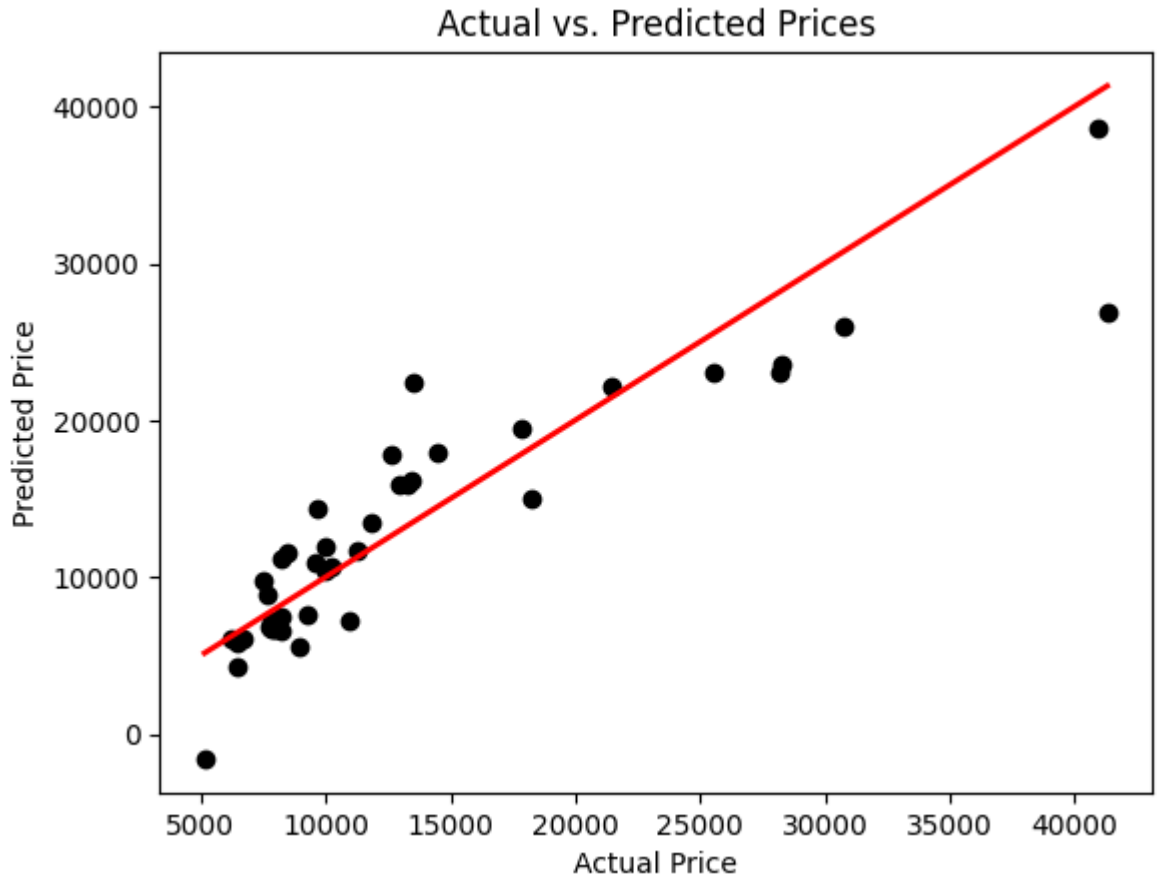








Una vez elegidas las variables empecé el proceso de generar la regresión lineal, decidí hacer una regresión lineal de múltiples variables para mi solución del reto y obtuve la siguiente gráfica.



Mis métricas de desempeño fueron los siguientes:

Mean Squared Error: 14316921.75872801

R^2 (: 0.8186446345245989

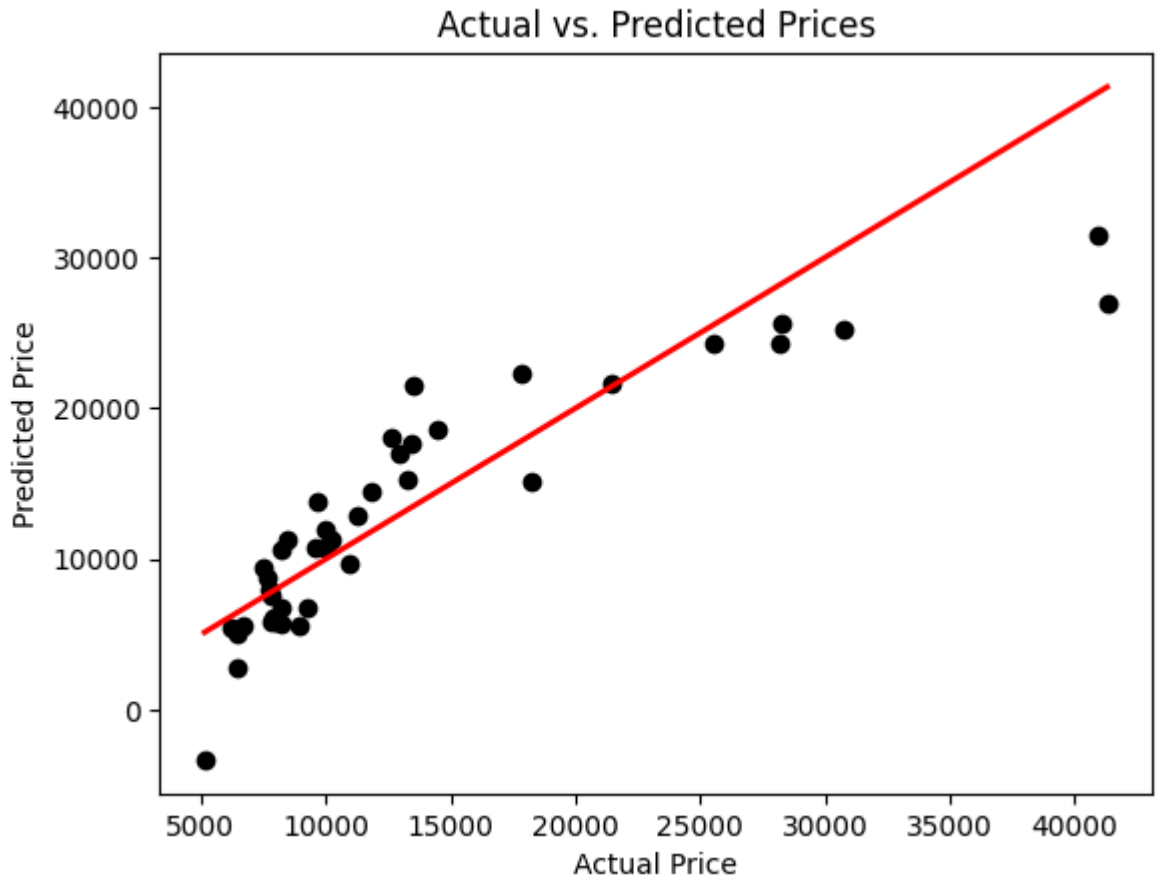
P-value: 1.1102230246251565e-16

Analizando nuestros valores podemos observar que nuestro modelo tuvo un desempeño malo, el valor del MSE salió demasiado alto así que hay un error promedio muy alto entre mis valores predecidos y los valores reales. Del lado positivo tuve un muy buen resultado del modelo puede “explicar” cerca del 82% de los datos generados en la predicción. Pero a pesar de este buen desempeño en el área de R^2 todavia me falta modificar mi modelo para poder tener un mejor margen de error en mis valores reales y predecidos. Otra métrica que tome para medir mi desempeño fue el valor P, este valor me muestra la significancia que mis datos tienen sobre la predicción, como se puede observar mi valor P fue muy cercano a 0 así que puedo inferir que mis datos tienen una muy alta significancia y son de alta relevancia cuando se intenta generar una predicción del precio.

Mi primer intento en mejorar estos valores fue a través de una transformación por valor de Z. El valor z encuentra cuanto un valor se desvía de la media de un conjunto de datos. Para mi primer intento hice una transformación sobre todos los puntos en mis valores independientes que tuvieran un valor de z mayor a 2 los cambie a la media de los datos, se cambiaron 59 datos y mis resultados fueron los siguientes.

Mean Squared Error: 64273612.3762614

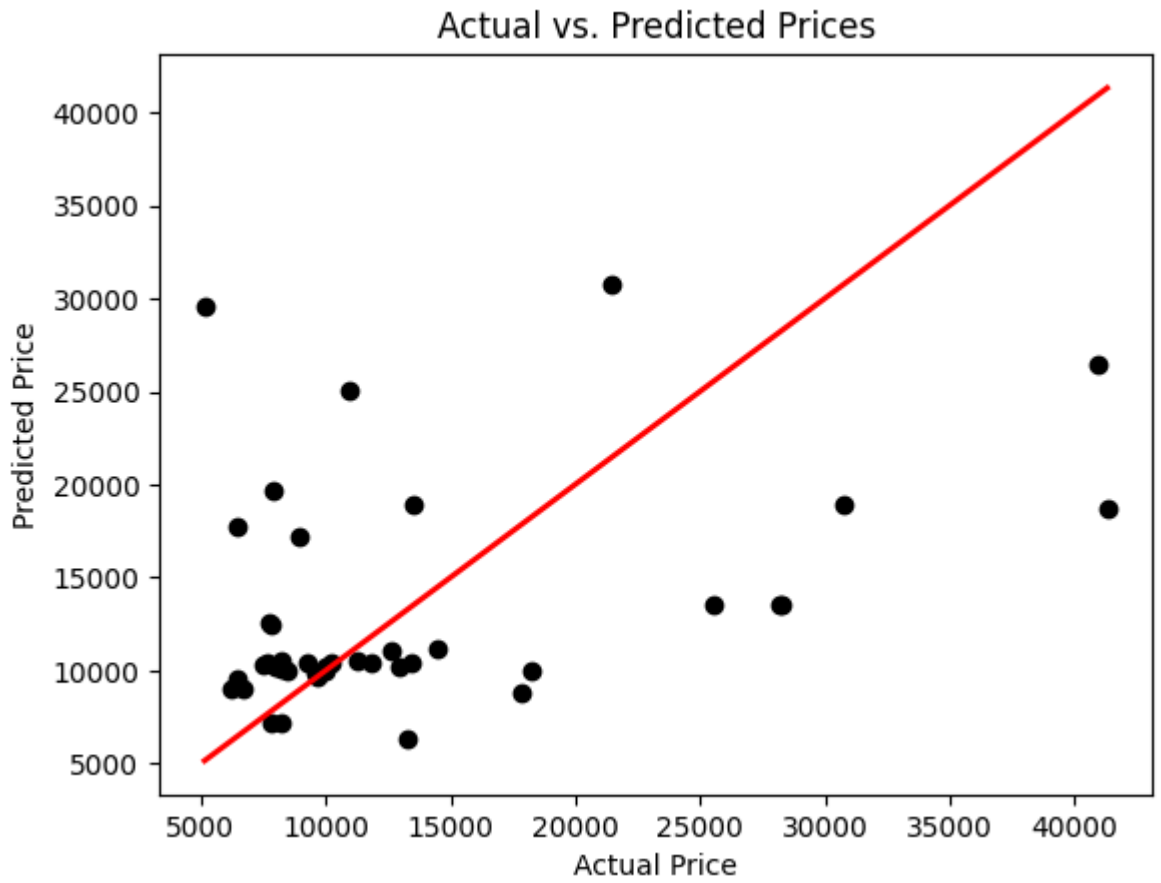
R-squared: 0.1858330541050076



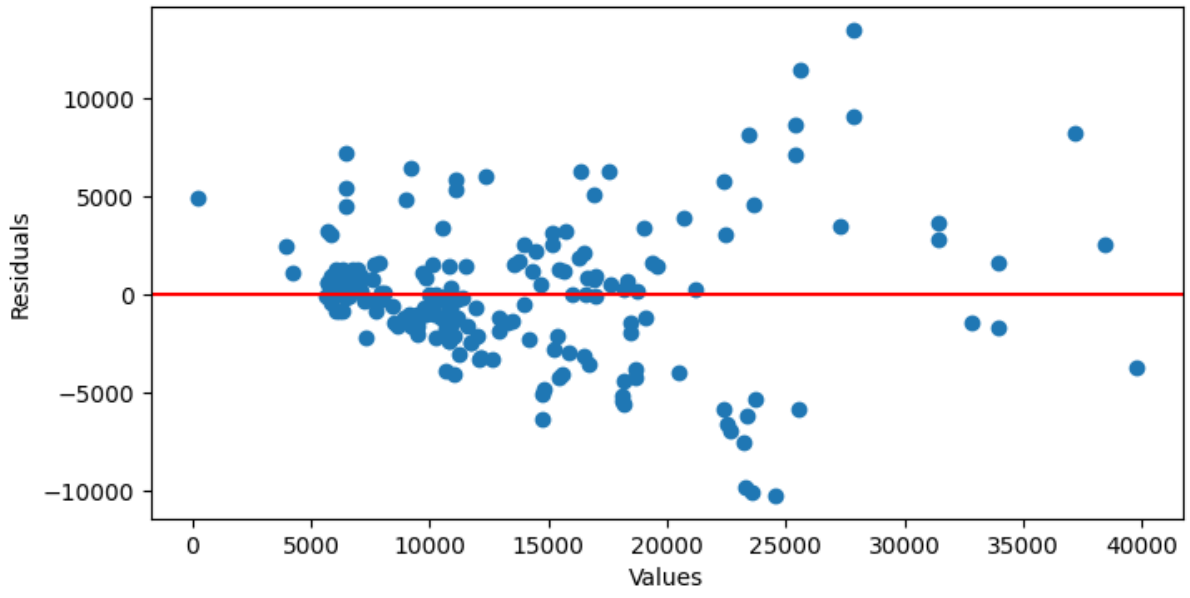
Como se puede observar mis métricas empeoraron mucho. Después hice otro intento con otra la misma transformación pero en vez de cambiarlos a la media hice una transformación logarítmica. Mis resultados fueron los siguientes.

Mean Squared Error: 73477786.58501509

R-squared: 0.06924190374056727



Después de ambas transformaciones puedo inferir que en realidad mis datos originales eran los que mejor sirvieron para realizar mis predicciones. Después realice una prueba para ver si se puede observar homocedasticidad y media de zero.



Mis residuos son la diferencia entre mis valores reales y mis valores predecidos. En la gráfica se puede observar que los residuos se ven distribuidos de forma semejante por encima y por debajo del 0 así que podemos inferir que mi modelo cumple con el principio de media de cero. El principio de homocedasticidad se observa cuando los puntos se ven semejantemente distribuidos a lo largo del eje x de la gráfica. Como podemos observar no hay una buena distribución en mi gráfica esto quiere decir que no hay una varianza equitativa y posiblemente existe bias en mis coeficientes.

Conclusión:

Después de generar mi modelo y analizar mis resultados puedo determinar que mis predicciones tienen valor pero pueden ser mejoradas. El uso de herramientas de aprendizaje máquina deben de ser complementados con fundamentos estadísticos de análisis de datos ya que estos son los que me permitieron determinar el desempeño de mi modelo así como donde son las áreas de mejora para el modelo. Específicamente creo que a la hora de determinar las variables que se tomarán en cuenta para el modelo es sumamente importante tener un conocimiento de las herramientas que se pueden utilizar para tener los mejores resultados posibles y una vez realizada la predicción es importante tener un conocimiento de las maneras en las que puedo medir que tan bien mi modelo está realizando su predicción.

Anexos:

https://colab.research.google.com/drive/1nk37oEleqVA-1xF6-RtFHfejgYJ4a8yR?usp=sharing

Referencias:

J.D. POWER. (2023, June 8). *Fueling Change:The Power of AI and Market Data in Transforming the Automotive Industry*. J.D. Power. Retrieved September 12, 2023, from https://www.jdpower.com/cars/shopping-guides/fueling-change-the-power-of-ai-and-market-data-in-transforming-the-automotive-industry