PROYECTO
BI SCIENSE

“Estimación de Precio

para venta de autos usados”

Profesor: Norberto Leonel Gonzales

Tutor: Rogelio Tobar De La Cruz

Fecha de Entrega: 22/03/2023

Arturo Ledezma y Alex Arce

CONTENIDO

PROYECTO DATA SCIENCE: ESTIMAR PRECIO

DESCRIPCIÓN DEL CASO DE NEGOCIO

Con la escasez de chips para la fabricación de nuevos modelos y en consecuencia abastecer la demanda actual de autos, hemos visto un incremento en la demanda o decisión de compra sobre automóviles usados para la región de Estados Unidos, objeto del presente estudio.

Mediante el proyecto se pretende poder establecer una base para predecir el precio futuro de vehículos usados en función de variables que mejor describan su influencia en el precio.

El caso va dirigido hacia todas aquellas personas o compañías que se dedican a la venta-compra de vehículos y que pueda ser ayudarles mejor a la toma de decisiones, buscando incrementar la rentabilidad del negocio.

TABLA DE VERSIONADO

Fecha	Entregas	Referencia de cambios
31/11/2022	1.0	Se realizó la primera entrega donde realizamos DS con su respectivo Análisis de los datos (Data Wrangling, Data Adquisicion y EDA).
21/02/2023	2.0	Se realizo la segunda entrega donde ya se incluye PCA, RandomForest, Validación Simple y CV

OBJETIVOS DEL MODELO

La finalidad del proyecto es realizar un análisis exploratorio del dataset para encontrar patrones en la información que nos ayuden a estudiar las correlaciones más importantes que ayuden a elaborar un modelo de predicción (regresión) para las ventas de automóviles usados de los siguientes periodos para lo que se emplearan distintas librerías de ciencia de datos en Python (pandas, numpy, Matplotlib, Seaborn, etc.).

Esperamos además poder tener hallazgos que permitan optimizar estrategias de venta y con ello capturar mayor ingreso por unidad vendida.

DESCRIPCIÓN DE LOS DATOS

Los datos refieren a una compañía de Compra y venta de autos usados la cual quiere determinar qué factores son determinantes para el precio final del auto. EL Data Set contiene 24 columnas y 19514 filas.

Variables:

- Date = Fecha de venta del vehículo
- Dealer Name = Nombre del vendedor del vehículo
- Company Model = Marca del vehículo
- Year = Año del vehículo
- Body Style = Tipo de vehículo (sedan, SUV,minivan, pickup, etc)
- Transmission = Transmisión manual o automática)
- Color = Color del vehículo
- Price = Precio en dólares al que fue vendido el vehículo
- Millage KM = cantidad de KM acumulados del vehículo
- Fuel = Tipo de combustible empleado en el vehículo
- Engine capacity = capacidad cc del vehiculo
- Drivetrain = tren motriz (tracción delantera, trasera, o completa)
- Dealer add = dirección del vendedor (particular)
- Customer = nombre del cliente

- Customer address = dirección del cliente
- Counsilarea = condado de dirección particular vendedor
- Gender Anual = genero o sexo del comprador
- Annual Income = Ingerso anual del comprador
- Dealer Location = Dirección del local
- Dealer no = teléfono del vendedor
- Dealer región = Región del vendedor

Acciones realizadas al DS:

- Identificamos 24 columnas y 19514 filas y los valores nulos fueron eliminados DS, esto para evitar futuros inconvenientes
- Fueron Agregadas Columnas para completar el data set por falta de variantes

HALLAZGOS ENCONTRADOS EN EL EDA

GRÁFICO 1:

Identificamos que los valores obtenidos en el año 2021 registran un PIT en el mes de Julio.

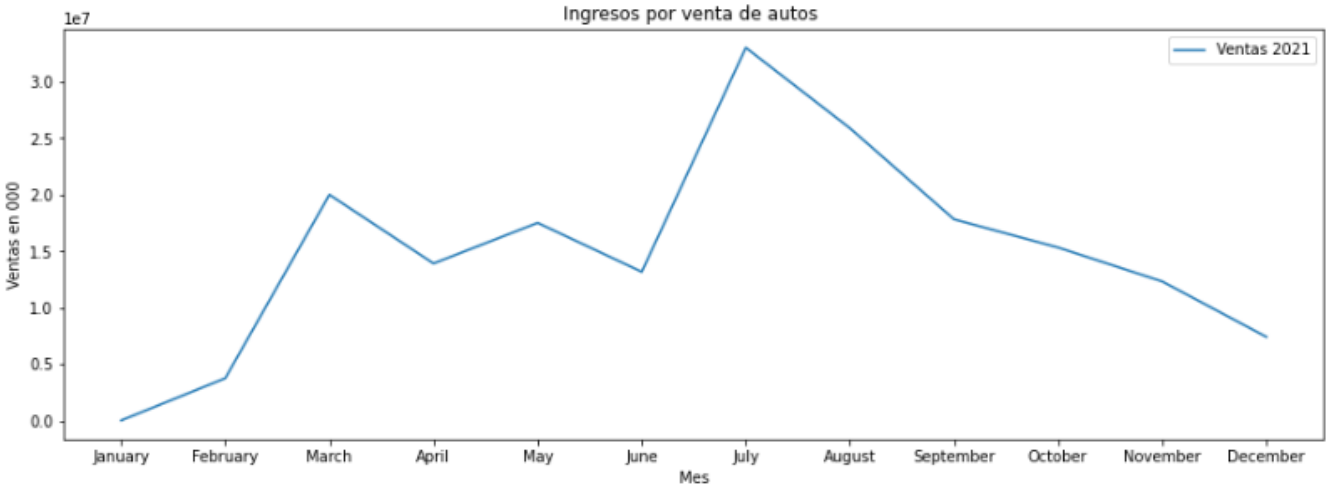


GRÁFICO 2:

El género que más compra autos es el de los hombres. Los meses con mayor volumen de transacciones son el marzo, julio y agosto. Los meses con volúmenes muy bajos de ventas son Enero, Febrero y Diciembre.

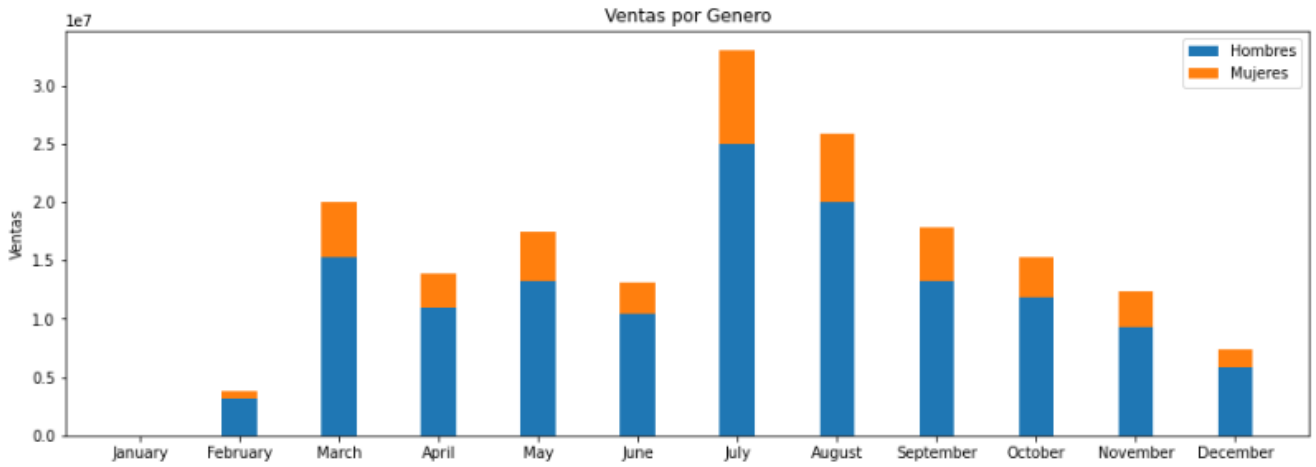


GRÁFICO 3:

Gráficos por Vetas de modelos de Autos, observamos que Los modelos con mayor venta son los que tienen una antigüedad de 6 a 15 años

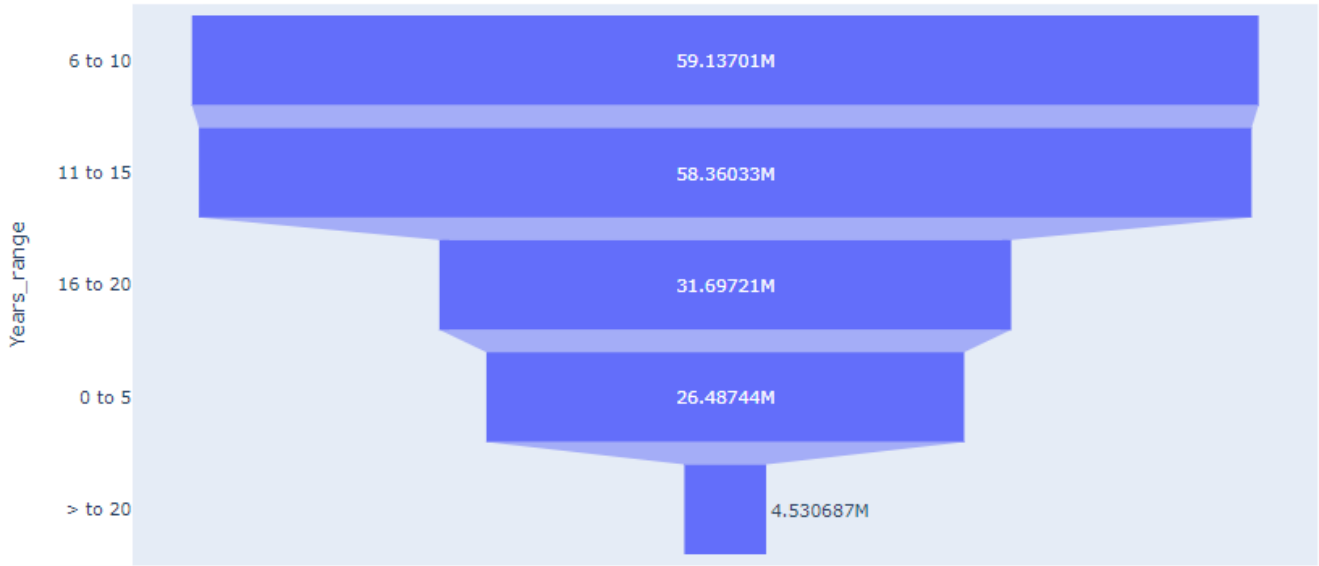
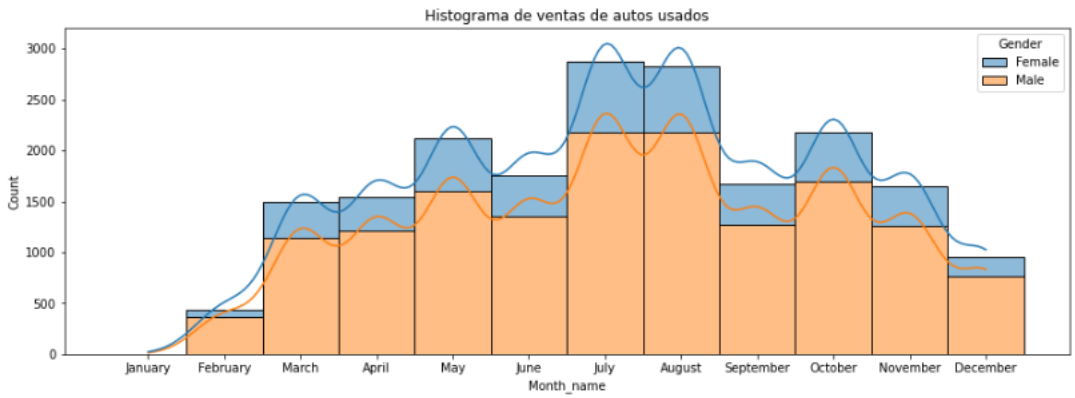


GRÁFICO 4:

Chequeamos como se distribuye las ganancias respecto en el transcurso de los años que se realizó el DS, Se registra un ingreso mayor en ventas realizadas durante los meses de Julio y Agosto, siendo el género masculino quien más compra autos usados en el transcurso del año.



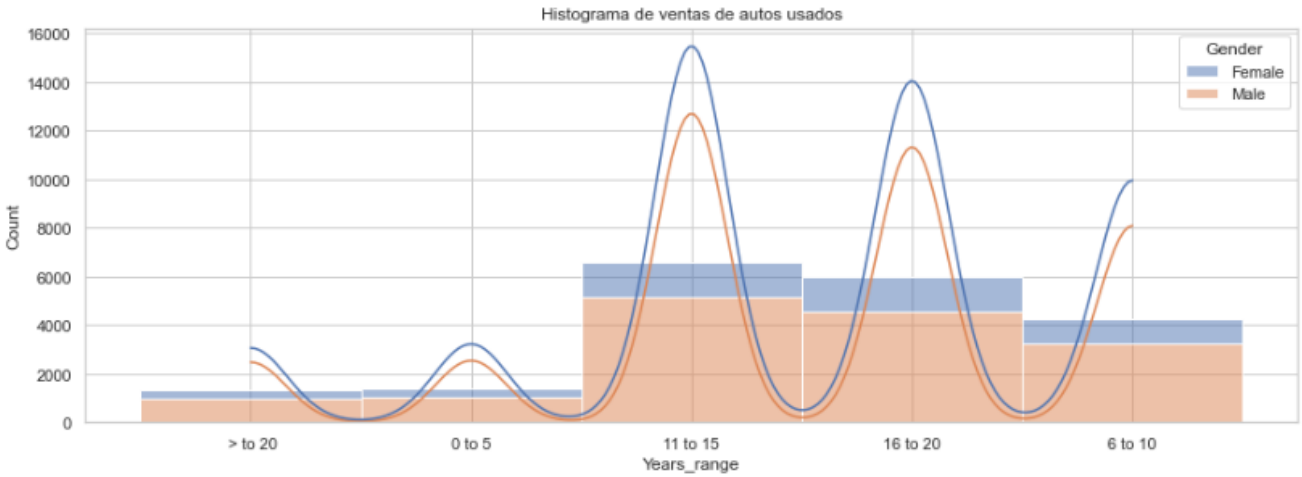


GRÁFICO 5:

El precio de los autos oscila entre los 150 a 50,000 USD, el promedio es de 9,235 USD (mediana de 7,500 USD) lo que habla de un sesgo positivo por ser la mediana<media; el precio que más se repite (moda) es de 9,235 USD lo que contrasta con una acumulación de los datos en el segundo percentil y con datos dispersos a revisar por arriba de los 22,000 USD.

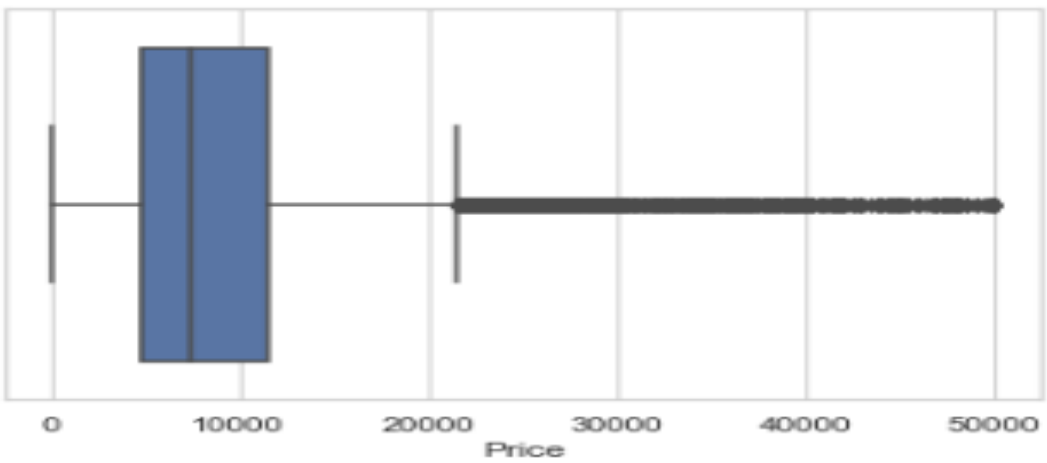


GRÁFICO 6:

Aunque se observa una correlación fuerte entre el precio y la antigüedad del auto, además del kilometraje acumulado, debemos modificar el DataFrame para codificar columnas como género, color, etc. asignándoles un valor numérico por categoría y analizar más a fondo la correlación con otras características del vehículo.

	Year	Price	Mileage_km	engine_capacity	Annual Income	Months	Years_old
Year	1.000000	0.671517	-0.673459	-0.187110	-0.005662	-0.178171	-1.000000
Price	0.671517	1.000000	-0.433368	0.274543	0.001774	-0.126966	-0.671517
Mileage_km	-0.673459	-0.433368	1.000000	0.175454	0.011438	0.134995	0.673459
engine_capacity	-0.187110	0.274543	0.175454	1.000000	0.007762	0.036772	0.187110
Annual Income	-0.005662	0.001774	0.011438	0.007762	1.000000	0.004334	0.005662
Months	-0.178171	-0.126966	0.134995	0.036772	0.004334	1.000000	0.178171
Years_old	-1.000000	-0.671517	0.673459	0.187110	0.005662	0.178171	1.000000

GRÁFICO 7:

Observamos que la mayor correlación con el precio del vehículo es el Modelo y el año del vehículo, es decir son los que en mayor medida influyen para determinar el precio y existen otras correlaciones como el tipo de vehículo, tren motriz, transmisión y kilometraje

	Year	Price	Mileage_km	engine_capacity	Annual Income	Months	Company Category	Model Category	Year Old Category	Old Range Category	Body Style Category	Transmission Category	Color Category	Fuel Category	Drivetrain Category	Dealer Add Category	CouncilArea Category	Gender Category	Dealer_Location Category	Dealer_Region Category
Year	1.000000	0.671517	-0.673459	-0.187110	-0.005662	-0.178171	0.082604	-0.021896	-1.000000	0.207718	0.067727	-0.196863	-0.005668	0.195708	-0.145114	0.004546	-0.003021	-0.003205	0.003022	0.009456
Price	0.671517	1.000000	-0.433368	0.274543	0.001774	-0.126966	-0.018746	0.055695	-0.671517	0.121709	0.201691	-0.372982	-0.068362	-0.004457	-0.273213	0.001094	-0.000934	-0.008251	0.001212	-0.003552
Mileage_km	-0.673459	-0.433368	1.000000	0.175454	0.011438	0.134995	-0.032673	-0.047169	0.673459	-0.101899	0.057222	0.198674	0.027918	-0.401835	0.181553	-0.002810	0.003019	0.004388	-0.005060	-0.008640
engine_capacity	-0.187110	0.274543	0.175454	1.000000	0.007762	0.036772	-0.206996	0.166980	0.187110	0.003804	0.273475	-0.420572	-0.117903	-0.037775	-0.227061	-0.004728	0.000756	0.007832	0.002934	-0.012863
Annual Income	-0.005662	0.001774	0.011438	0.007762	1.000000	0.004334	0.004768	0.000209	0.005662	-0.007983	-0.002923	0.000759	-0.002484	-0.000690	-0.005983	-0.012091	-0.074759	0.064917	0.003079	-0.002008
Months	-0.178171	-0.126966	0.134995	0.036772	0.004334	1.000000	-0.027369	0.058199	0.178171	0.149201	-0.011152	0.041950	0.000928	-0.057480	0.016336	-0.016369	0.014793	0.000504	0.008548	0.006061
Company Category	0.082604	-0.018746	-0.032673	-0.206996	0.004768	-0.027369	1.000000	-0.098514	-0.082604	-0.027039	-0.042325	0.163964	0.034950	-0.043517	-0.046528	0.000219	-0.008022	0.000332	0.007059	-0.004015
Model Category	-0.021896	0.055695	-0.047169	0.166980	0.000209	0.058199	-0.098514	1.000000	0.021896	0.009739	-0.100761	-0.069847	0.012464	0.074601	0.024590	0.000496	-0.000503	0.005402	0.007676	0.004343
Year Old Category	-1.000000	-0.671517	0.673459	0.187110	0.005662	0.178171	-0.082604	0.021896	1.000000	-0.207718	-0.067727	0.196863	0.005668	-0.195708	0.145114	-0.004546	0.003021	0.003205	-0.003022	-0.009456
Old Range Category	0.207718	0.121709	-0.101899	0.003804	-0.007983	0.149201	-0.027039	0.009739	-0.207718	1.000000	0.022187	-0.070508	0.012075	0.005917	-0.056092	-0.006045	-0.005251	0.002338	-0.004081	0.022333
Body Style Category	0.067727	0.201691	0.057222	0.273475	-0.002923	-0.011152	-0.042325	-0.100761	0.067727	0.022187	1.000000	-0.168361	-0.062626	-0.132766	-0.180750	-0.003609	-0.000372	-0.000506	0.003101	-0.004824
Transmission Category	-0.196863	-0.372982	0.198674	-0.420572	0.000759	0.041950	0.163964	-0.069847	0.196863	-0.070508	-0.168361	1.000000	0.106274	-0.243315	0.217852	0.003562	-0.007756	-0.004811	0.004985	0.001215
Color Category	-0.005668	-0.068362	0.027918	-0.117903	-0.002484	0.000928	0.034950	0.012464	0.005668	0.012075	-0.062626	0.106274	1.000000	-0.021516	0.065987	-0.000474	0.011011	0.003992	0.006824	-0.001401
Fuel Category	0.195708	-0.004457	-0.401835	-0.037775	-0.000690	-0.057480	-0.043517	0.074601	-0.195708	0.005917	-0.132766	-0.243315	-0.021516	1.000000	-0.075354	-0.000073	0.006205	-0.001058	0.000846	0.007725
Drivetrain Category	-0.145114	-0.273213	0.181553	-0.227061	-0.005983	0.016336	-0.046528	0.024590	0.145114	-0.056092	-0.180750	0.217852	0.065987	-0.075354	1.000000	0.000078	0.005733	-0.005614	-0.014486	-0.005213
Dealer Add Category	0.004546	0.001094	-0.002810	-0.004728	-0.012091	-0.016369	0.000219	0.000496	-0.004546	-0.006045	-0.003609	0.003562	-0.000474	-0.000073	0.000076	1.000000	0.002271	-0.011713	0.008144	-0.003545
CouncilArea Category	-0.003021	-0.000934	0.003019	0.000756	-0.074759	0.014793	-0.008022	-0.000503	0.003021	-0.005251	-0.000372	-0.007756	0.011011	0.006205	0.005733	0.002271	1.000000	0.016344	0.006408	0.003976
Gender Category	-0.003205	-0.008251	0.004388	0.007832	0.064917	0.000504	0.000332	0.005402	0.003205	0.002338	-0.000506	-0.004811	0.003992	-0.001058	-0.005614	-0.011713	0.016344	1.000000	0.003810	0.002285
Dealer_Location Category	0.003022	0.001212	-0.005060	0.002934	0.003079	0.008548	0.007059	0.007676	0.003022	-0.004081	0.003101	0.004985	0.006824	0.000846	-0.014486	0.008144	0.006408	0.003810	1.000000	0.069025
Dealer_Region Category	0.009456	-0.003552	-0.008640	-0.012863	-0.002008	0.006061	-0.004015	0.004343	-0.009456	0.022333	-0.004824	0.001215	-0.001401	0.007725	-0.005213	-0.003545	0.003976	0.002285	0.069025	1.000000

VARIABLES Y TARGET

Para poder realizar una regresión en el precio las variables a utilizar de acuerdo a su correlación:

- Mileage_km: Kilometraje -0.433368
- Engine_capacity: capacidad del motor 0.274543
- Years Old: Antigûedad del vehiculo 0.671517
- Transmission: Tipo de transmisión -0.372982
- Drivetrain: Tren Motriz -0.273213

Como se describe en el objetivo, la finalidad es identificar o predecir el precio óptimo (Target Precio) para los modelos actualmente en venta para los siguientes años, además de identificar la tendencia en modelos o tipos de auto más convenientes para mejorar el ingreso y dejar de lado el inventario que significa poco ingreso (compra-venta)

- Variables X= [Mileage,Engine capacity,Years Old,Transmission, Drivetrain]
- Target Y= [Price]

ALGORITMOS ELEGIDOS

Regresión Lineal Múltiple y Random Forest:

De acuerdo con el análisis de correlaciones y las variables seleccionadas, poder generar una proyección de lo que será el precio del vehículo para el siguiente año de venta y también el poder predecir la base de precios para nuevo inventario de vehículos a vender

Métricas de desempeño del modelo

Regresión lineal Múltiple vs Random Forest:

Como se describe en el objetivo, la finalidad es identificar o predecir el precio óptimo (Target Precio) para los modelos actualmente en venta para los siguientes años, además de identificar la tendencia en modelos o tipos de auto más convenientes para mejorar el ingreso y dejar de lado el inventario que significa poco ingreso (compra-venta)

- Error cuadrático medio (MSE).
- Error cuadrático medio (RMSE).
- Error absoluto medio (MAE)
- r2 COEFICIENTE DE DETERMINACIÓN

De la regresión lineal múltiple:

Mean Absolute Error (MAE): 3,165.95

- Mean Squared Error (MSE): 21,563,437.38
- Root Mean Squared Error (RMSE): 4,643.64
- r2 COEFICIENTE DE DETERMINACIÓN: 52.18 % de las veces se acierta en la predicción

Del modelo random forest:

- Mean Absolute Error (MAE): 2,538.30
- Mean Squared Error (MSE): 14,438,520.67
- Root Mean Squared Error (RMSE): 3,799.80
- r2 COEFICIENTE DE DETERMINACIÓN: 67.98 % de las veces se acierta en la predicción

El valor absoluto de los errores (MAE) es mayor en la regresión lineal múltiple (RLM) vs Random Forest (RF),siendo que para la RLM el máximo valor absoluto de error es de 31,822.82 y en para RF es de 27,051.94; en ambos modelos un valor muy alto para el precio de un vehículo, gráficamente observamos que existen muchos puntos (registros-vehículos) por arriba de los $40,000 en precio que pueden estar originando que sea tan alto el valor del error absoluto.

Para el MSE la predicción en ambos modelos muestra grandes errores pero al hacer la comparativa con RMSE considerando que ambos modelos utilizan la misma escala o unidad de medida ($)vemos que Random Forest nos da un menor margen de error promedio con +- $3,799.80 además de evaluar el Coeficiente de determinación que con RF podemos acertar en el 67.98% de las veces.

Se realizo una validación de modelo utilizando el K-fold en donde el MSE mejoro ligeramente siendo de 12,738,609.25 vs la validación simple en donde es es 14,438,520.67

Considerar además que los vehículos que son de marca o segmento premium con un valor mayor a los $40,000 hace que los modelos aumenten su MSE o error por lo que para futuros análisis valdría la pena separar el frame y modelos por segmentos de precio o target de clientes.

CONCLUCION

Hasta aquí finalizamos con el estudio y el análisis de la aplicación de los modelos de Machine learning sobre nuestra prediccion estimacion de precio para venta de autos usados. Pudimos ver, a modo de conclusión, cuáles son los modelos que mejor trabajan sobre las variables de nuestro set de datos, de acuerdo a su composición.

Finalizamos con la conclusión final de que el Modelo de predicción de Clasificación Random Forest es el acorde para nuestro proyecto debido a la problemática, los features, el set de datos en sí y los objetivos planteados.