“Inteligencia de Negocios: Laboratorio 2”

Felix S. Rojas Casadiego, Juan S. Alegría Z., Daniel Reales
Universidad de los Andes, Bogotá, Colombia
{fs.rojas, j.alegria, da.reales}@uniandes.edu.co
Fecha de presentación: marzo 5 de 2023

Tabla de contenido

Modelo analítico: Felix Rojas
Pipeline: Daniel Reales
Visualización: Juan Alegría

1. Entendimiento de los datos

En esta etapa, se realizará un análisis de la calidad de los datos para comprender el conjunto de datos de MotorAlpes. El conjunto de datos contiene 11 columnas principales que incluyen información sobre los vehículos usados, como el año en que fue comprado, el número de kilómetros recorridos, el número de propietarios, el tipo de vendedor, el número de asientos, el combustible, la transmisión, el rendimiento, el tamaño del motor, la potencia máxima y el precio de venta. Primero, se examinará el contenido y la estructura de los datos. El conjunto de datos contiene 7115 filas y 11 columnas. La mayoría de las variables son numéricas, aunque algunas son categóricas.

En cuanto a los datos numéricos, se puede observar que en algunos casos como el selling price se supera el límite de valores posibles indicados en el diccionario del cliente, también engine y max_power tienen un valor mínimo incorrecto.

	year	km_driven	seats	mileage	engine	max_power	selling_price
count	6876.000000	6.917000e+03	7115.000000	6917.000000	6835.000000	6847.000000	6714.000000
mean	2013.980948	6.911118e+04	5.411103	19.523473	1835.489539	141.981595	11261.208041
std	3.852565	5.796521e+04	0.953555	4.241574	2363.919253	274.956684	40765.694516
min	1994.000000	1.000000e+00	2.000000	0.000000	4.000000	1.070000	1.910000
25%	2012.000000	3.400000e+04	5.000000	16.800000	1197.000000	68.050000	3210.560000
50%	2015.000000	6.000000e+04	5.000000	19.330000	1248.000000	83.100000	5451.900000
75%	2017.000000	9.400000e+04	5.000000	22.320000	1597.000000	104.680000	8480.740000
max	2020.000000	2.360457e+06	14.000000	46.816000	19972.000000	1995.640000	598983.440000

En cuanto a los datos categóricos, se puede observar que en la columna “Dueño” hay un 3% de datos inválidos, y la mayoría de los vehículos usados son de primer dueño.

```
First Owner             0.635278
Second Owner            0.249192
Third Owner             0.061982
NaN                     0.033591
Fourth & Above Owner    0.019396
Test Drive Car          0.000562
Name: owner, dtype: float64
```

Aproximadamente el 83% de los vehículos tienen a un tipo de vendedor que es individual.

```
Individual          0.827969
Dealer              0.141673
Trustmark Dealer    0.030358
Name: seller_type, dtype: float64
```

En cuanto al tipo de combustible, más de la mitad de los vehículos usados en venta son de combustible Diesel y la otra mitad petróleo, mientras que un 1% de los vehículos utilizan CNG o LGP.

```
Diesel    0.540126
Petrol    0.448630
CNG       0.006746
LPG       0.004498
Name: fuel, dtype: float64
```

Por otro lado, la mayoría de los vehículos en venta son de transmisión manual, sólo 13% son automáticos.

```
Manual       0.86676
Automatic    0.13324
Name: transmission, dtype: float64
```

Ahora bien, hay varios datos faltantes, pero no representa un problema mayor ya que la columna con mayor número de datos faltantes es el precio de venta con aproximadamente un 6%.

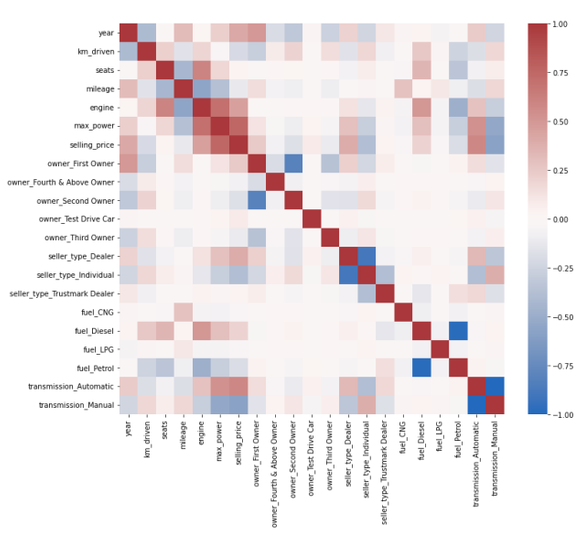
```
# Ver el porcentaje de atributos vacios
df_porcentajes = (100*df_cars.isna().sum()/len(df_cars)).to_frame()
df_porcentajes.sort_values(0, ascending = False)
```

✓ 0.1s

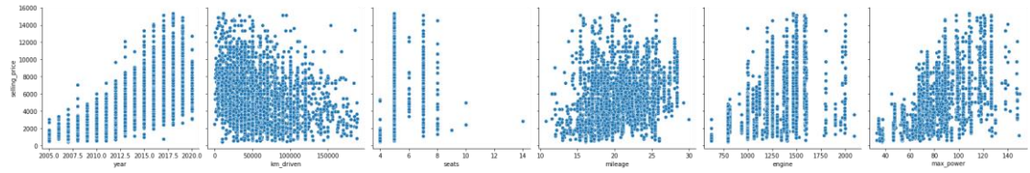
	0
selling_price	5.635980
engine	3.935348
max_power	3.766690
year	3.359100
owner	3.359100
km_driven	2.782853
mileage	2.782853
seller_type	0.000000
seats	0.000000
fuel	0.000000
transmission	0.000000

2. Identificación de variables a utilizar

Después de hacer one hot encoding para las variables categóricas, se decidió eliminar aquellas variables con
una correlación mayor a ~|0.8|. Estas variables son 'owner_Second Owner', 'seller_type_Individual', 'fuel_Petrol', y 'transmission_Manual'



Luego, para identificar las variables más relevantes, se hizo un scatterplot de cada una de las variables en relación con el precio de venta de los vehículos usados, llegando a la conclusión de que las variables 'year', 'km_driven', 'mileage', 'engine' y 'max_power' son las que tienen una mayor relación con la variable objetivo.



3. Preparación de datos

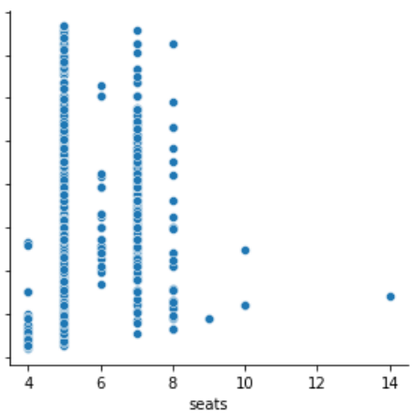
Para esto, se eliminaron los valores inválidos de acuerdo con el diccionario de datos proporcionado por el cliente.

```
# Eliminar valores inválidos de acuerdo al diccionario de datos proporcionado
df_cars = df_cars[(df_cars['year'] >= 1994) & (df_cars['year'] <= 2020) &
                  (df_cars['km_driven'] >= 1) & (df_cars['km_driven'] <= 2360457) &
                  (df_cars['seats'] >= 2) & (df_cars['seats'] <= 14) &
                  (df_cars['mileage'] >= 0) & (df_cars['mileage'] <= 46.816) &
                  (df_cars['engine'] >= 624) & (df_cars['engine'] <= 3604) &
                  (df_cars['max_power'] >= 32.8) & (df_cars['max_power'] <= 400) &
                  (df_cars['selling_price'] >= 363.45) & (df_cars['selling_price'] <= 121153.38)]
```

Después de este proceso, el porcentaje de atributos vacíos para cada variable disminuyó a 0 y la cantidad de filas disminuyeron de 7115 a 5458.

	0
year	0.0
km_driven	0.0
owner	0.0
seller_type	0.0
seats	0.0
fuel	0.0
transmission	0.0
mileage	0.0
engine	0.0
max_power	0.0
selling_price	0.0

Se decidió hacer uso del preprocesamiento recomendado para regresiones que implica manejar los datos atípicos. Para esto, se eliminaron 1271 filas utilizando la técnica de los cuartiles que elimina aquellos datos que no se encuentren dentro del rango intercuartílico. Se tuvo en cuenta únicamente en las variables numéricas que tengan datos suficientemente distribuidos, este no es el caso para la variable seats, por ejemplo.




```
q1 = df_cars.quantile(0.25)
q3 = df_cars.quantile(0.75)
IQR = q3 - q1
df_cars = df_cars[~((df_cars[['year', 'km_driven', 'mileage', 'engine', 'max_power', 'selling_price']] < (q1 - 1.5 * IQR)) |
                   (df_cars[['year', 'km_driven', 'mileage', 'engine', 'max_power', 'selling_price']] > (q3 + 1.5 * IQR))).any(axis=1)]
```

Por otro lado, al hacer regresión lineal es necesario que todas las variables sean númericas, así que se aplicó one hot encoding a las variables categóricas.

```
# Convertir variables categoricas a numericas
df_cars = pd.get_dummies(df_cars, columns=['owner', 'seller_type', 'fuel', 'transmission'])
```

4. Modelamiento

Primeramente, se tuvo en cuenta las variables identificadas en el punto 2 del documento para el modelo. También, se dividió el conjunto de datos en train y test para poder evaluar el rendimiento del modelo.

```
df = df_cars.copy()
df = df.reset_index(drop = True)
X_train, X_test, y_train, y_test = train_test_split(df[['year', 'km_driven', 'mileage', 'engine', 'max_power']], df['selling_price'], test_size = 0.3, random_state = 1)
```

El score de entrenamiento y test es el siguiente:

```
# Ver score de entrenamiento y test
print('Train:', regression.score(X_train, y_train))
print('Test:', regression.score(X_test, y_test))
✓ 0.0s

Train: 0.7257661352472929
Test: 0.7193060753361276
```

Luego, se decidió utilizar todas las variables para el modelo.

```
df = df_cars.copy()
df = df.reset_index(drop = True)
X_train, X_test, y_train, y_test = train_test_split(df.drop(['selling_price'], axis = 1), df['selling_price'], test_size = 0.3, random_state = 1)
```

Esto da un mejor resultado:

```
# Ver score de entrenamiento y test
print('Train:', regression.score(X_train, y_train))
print('Test:', regression.score(X_test, y_test))
✓ 0.1s

Train: 0.7537636056125889
Test: 0.7460481658476747
```

5. Evaluación cuantitativa

5.1 ¿Su equipo recomienda instalar el modelo de estimación en producción o es mejor continuar usando expertos para la tarea?

R// Para responder a esta pregunta, es necesario evaluar el desempeño del modelo de regresión en términos de precisión y eficiencia en comparación con los expertos humanos. Si el modelo proporciona una estimación más precisa y eficiente, es recomendable instalarlo en producción. Sin embargo, si el modelo no cumple con las expectativas o si el costo de implementación es prohibitivo, puede ser mejor continuar utilizando expertos para la tarea.

5.2 En caso de no recomendar el uso de un modelo de regresión ¿Qué otras posibilidades tiene la empresa? ¿Hacia dónde debe seguir con esta tarea?

R// Si el modelo de regresión no es recomendado, la empresa puede considerar otras técnicas de modelado, como modelos no lineales o de series de tiempo, o bien puede optar por mantener el uso de expertos humanos. También es importante evaluar los costos y beneficios de cada opción y considerar los posibles riesgos y limitaciones.

6. Evaluación cualitativa

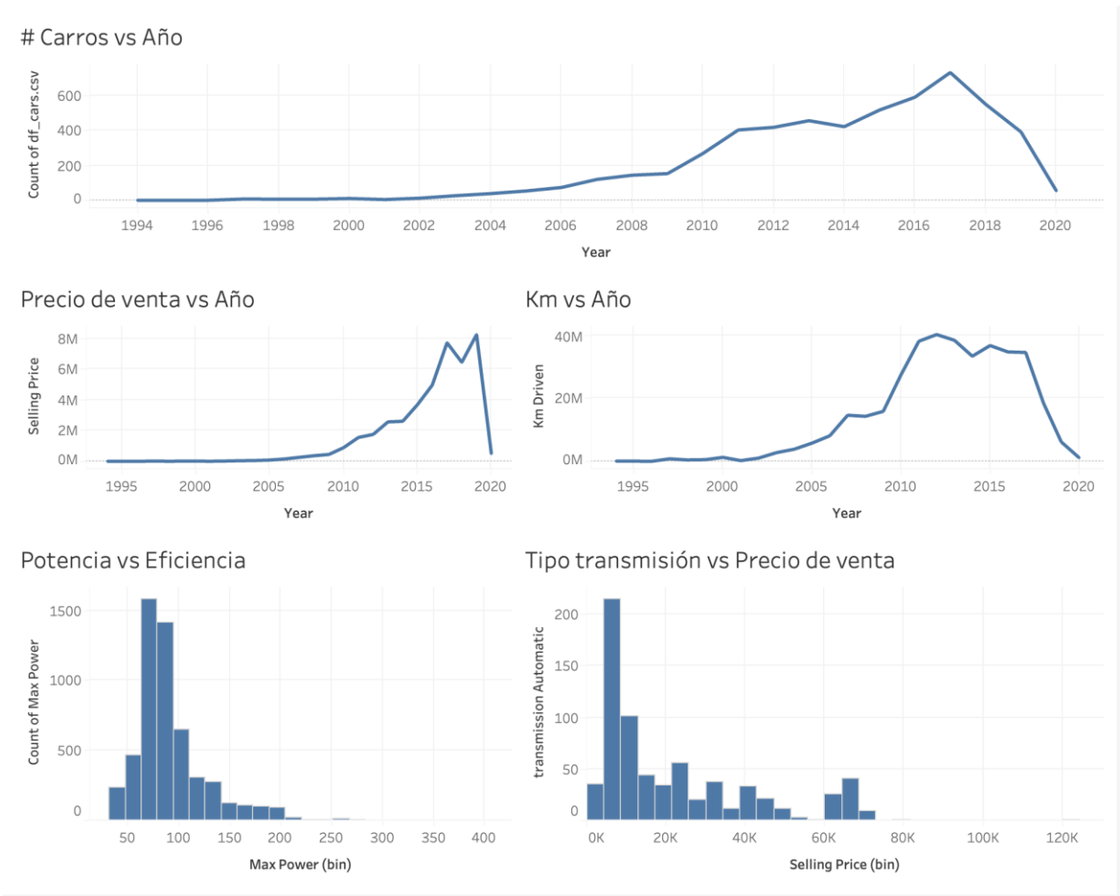
6.1 Validación de supuestos

La validación de supuestos en el modelo de regresión es fundamental para asegurarse de que los supuestos del modelo se cumplen. Estos supuestos incluyen la linealidad, normalidad, homocedasticidad, independencia y ausencia de multicolinealidad. La validación de estos supuestos puede realizarse mediante pruebas estadísticas, como la prueba de Shapiro-Wilk para normalidad y la prueba de Breusch-Pagan para homocedasticidad.

6.2 Interpretación de los coeficientes

La interpretación de los coeficientes en el modelo de regresión es importante para comprender la relación entre las variables. Los coeficientes representan la magnitud del efecto que cada variable tiene sobre la variable dependiente. Es importante tener en cuenta que la interpretación de los coeficientes puede verse afectada por la presencia de variables predictoras correlacionadas y la posible presencia de variables omitidas o variables de interacción que pueden afectar la interpretación de los resultados.

7. Visualizar el resultado del modelo



El dashboard creado muestra cinco gráficas de visualización de datos que se enfocan en diferentes aspectos importantes del conjunto de datos de carros del cliente.

La primera gráfica, "Carros vs Año", muestra el número de carros en el conjunto de datos para cada año de fabricación. Esta hoja puede ser útil para identificar cualquier tendencia en la cantidad de carros fabricados a lo largo de los años. Por ejemplo, si hay un aumento en la cantidad de carros fabricados en los últimos años, esto puede ser una indicación de la creciente demanda de carros.

La segunda gráfica, "Precio de venta vs Año", muestra la relación entre el precio de venta de los carros y su año de fabricación. Esta hoja puede ayudar a identificar cualquier tendencia en los precios de venta a lo largo del tiempo. Si se observa un aumento en el precio de venta de los carros a lo largo del tiempo, esto puede ser una indicación de que los carros se están volviendo más costosos para fabricar o de que los compradores están dispuestos a pagar más por los carros más nuevos.

La tercera gráfica, "Kms vs Año", muestra la relación entre los kilómetros recorridos por los carros y su año de fabricación. Esta hoja puede ser útil para identificar cualquier tendencia en la cantidad de kilómetros recorridos por los carros a lo largo del tiempo. Si se observa un aumento en la cantidad de kilómetros recorridos por los carros más nuevos, esto puede ser una indicación de que los compradores están utilizando más sus carros en la actualidad.

La cuarta gráfica, "Potencia vs Eficiencia", muestra la relación entre la potencia del motor y la eficiencia de combustible del carro. Esta hoja puede ser útil para identificar cualquier tendencia en la relación entre la potencia del motor y la eficiencia de combustible. Si se observa una tendencia en la que los carros más potentes son menos eficientes en términos de combustible, esto puede ser una indicación de que los compradores están dispuestos a sacrificar la eficiencia de combustible por una mayor potencia.

Finalmente, la quinta gráfica, "Tipo de transmisión vs precio de venta", muestra la relación entre el tipo de transmisión de un carro y su precio de venta. Esta hoja puede ser útil para identificar cualquier tendencia en la relación entre el tipo de transmisión y el precio de venta. Si se observa una tendencia en la que los carros con transmisión automática son más costosos que los carros con transmisión manual, esto puede ser una indicación de que los compradores están dispuestos a pagar más por la comodidad y la facilidad de uso que ofrecen las transmisiones automáticas.

8. Exportar el modelo

Se exportó el modelo con la librería joblib.