****

Universidad de los Andes

MIAD – Machine Learning y Procesamiento de Lenguaje Natural

30/04/2023

**Integrantes**

Daniel Rozo Isaza

Gabriel Gómez

Yolanda Franco

**Parte 1. Análisis descriptivo.**

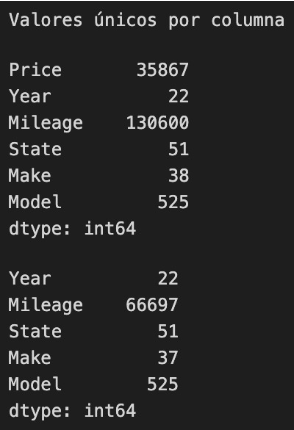
Las bases de datos a utilizar en este ejercicio ya habían sido segmentadas previamente en datos de entrenamiento y de prueba. La base de entrenamiento cuenta con un total de 400.000 observaciones, mientras que los datos de prueba constan de 100.000 observaciones. Adicionalmente, se cuentan con 6 columnas que corresponden al precio del vehículo, el año de fabricación, el kilometraje, el estado donde se vende el vehículo, el fabricante y el modelo. Las primeras tres variables son numéricas, dos de tipo continuo (precio y kilometraje) y una de tipo discreto (año de fabricación). Las tres variables restantes son de tipo categórico. La variable a predecir es el precio. Con los descriptivos realizados se tiene que el precio promedio es de 21,146.92 dólares, con una desviación estándar de 10,753.66 dólares, lo que indica una gran variabilidad en los precios. El precio mínimo registrado es de 5,001 dólares, mientras que el precio máximo es de 79,999 dólares. El 25% de los carros tienen un precio inferior a 13,499 dólares, el 50% tienen un precio inferior a 18,450 dólares y el 75% tienen un precio inferior a 26,999 dólares. Estos descriptivos indican que la distribución de precios es sesgada hacia la derecha, ya que la mediana (18,450 dólares) es menor que la media (21,146.92 dólares). El kilometraje promedio de los coches es de 55,072.96 millas, con una desviación estándar de 40,881.02 millas, lo que indica una gran variabilidad en los datos. El kilometraje mínimo registrado es de 5 millas, lo que podría ser atribuible a coches nuevos o a errores de entrada de datos. La mediana, que es el valor que separa la mitad inferior y la mitad superior de los datos, es de 42,955 millas, lo que sugiere que la mayoría de los coches han recorrido un número moderado de millas. El 25% de los coches tienen un kilometraje de menos de 25,841 millas, lo que indica que un cuarto de los coches son relativamente nuevos. El 75% de los coches tienen un kilometraje de menos de 77,433 millas, lo que indica que la mayoría de los coches han sido usados. El kilometraje máximo registrado es de 2,457,832 millas, lo que es una cifra bastante inusual y probablemente sea un error de entrada de datos o una medida exagerada. La siguiente tabla resume lo mencionado.

**Tabla 1. Estadísticas descriptivas de las variables numéricas.**



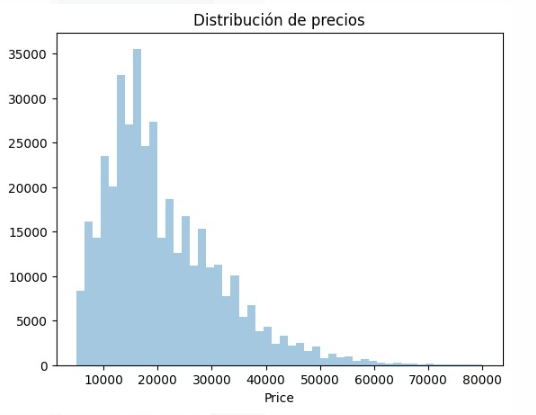
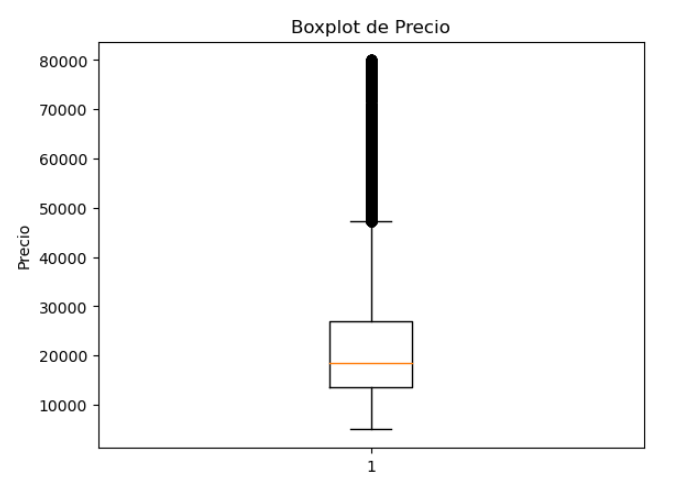
En general, no se encontraron valores nulos en ninguna de las columnas. En cambio, se cuenta con un alto número de valores únicos, incluso para las variables de tipo categóricas. La siguiente tabla muestra que, se tiene información de 22 años, 51 estados, 38 fabricantes y 525 modelos. El alto número de categorías que se presenta en una variable como modelos, eventualmente puede dificultad su inclusión en un modelo predictivo ya que, al transformar cada categoría en una variable dicotómica, el número de variables sube de manera exagerada. Por esta razón, más adelante se observará qué tipo de tratamiento recibirá esta variable.

**Tabla 2. Número de valores únicos por columna para las bases de prueba y entrenamiento.**



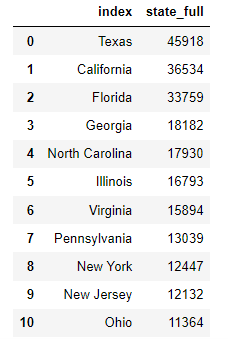
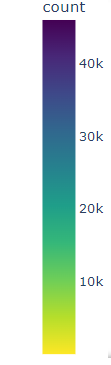
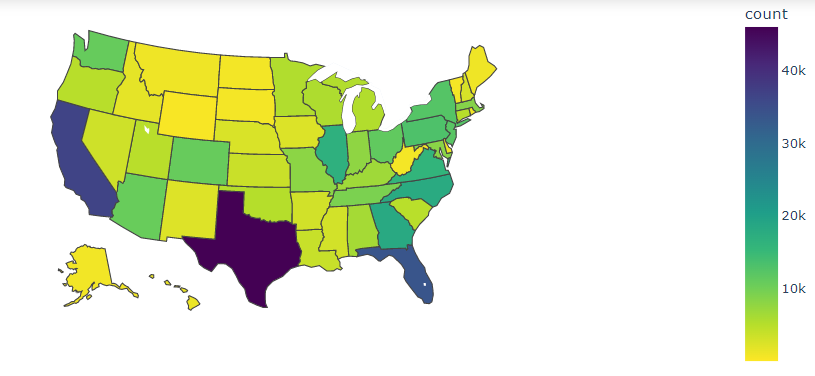
Tanto el histograma como el boxplot corroboran la asimetría de la variable precio hacia la derecha. Además, el boxplot muestra la presencia de datos atípicos altos. Para verificar si estos datos altos tienen sentido, se ubicó el fabricante y la marca del precio máximo encontrado. En esta base de datos el precio máximo es de 79,999 dólares y corresponde a un Land Rover del 2018. La búsqueda indica: "The Manufacturer's Suggested Retail Price (MSRP) for a base-model 2018 Land Rover Range Rover starts at about 88,500, which isn't bad, while long-wheelbase versions start at about 110,000", siendo el de la base de datos más bajo que el valor encontrado en la búsqueda, pero esto puede deberse a que no es un auto de primera mano, sino que cuenta con 649 millas.

**Gráfico 1. Distribución de la variable Precio.**



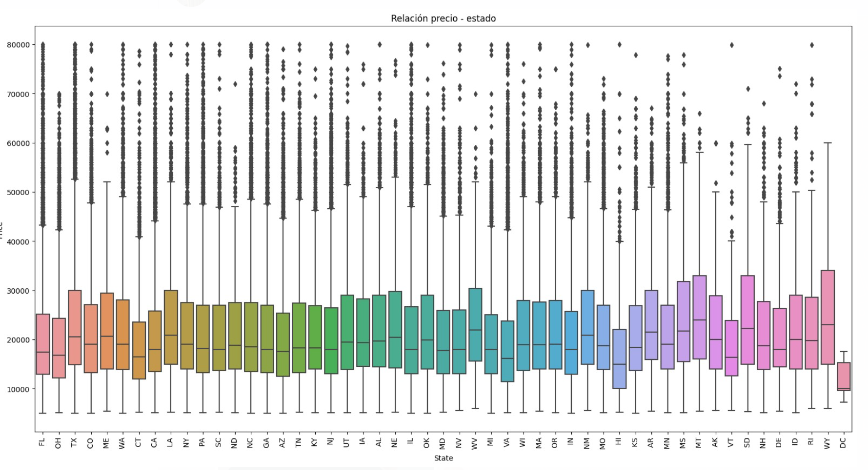
Sobre los estados en los cuales se listan los automóviles, se construyó un mapa para comprender su dinámica, además de una tabla con los primeros 10 estados con el mayor número de observaciones. Se encuentra que los vehículos de la muestra tienden a concentrarse en el sur, sobre todo en los estados de Texas, California y Florida.

**Gráfico 2. Distribución de la muestra por estados.**



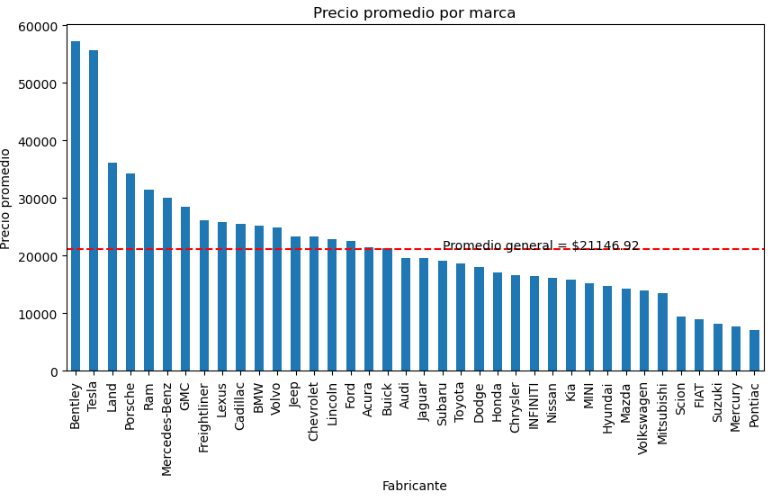
La distribución de los precios por estado muestra que no hay diferencias muy marcadas de precio entre un estado y otro. En realidad, la única excepción parece ser Washington DC, donde los precios se tienden a concentrar en valores inferiores a los que se encuentran en los demás estados.

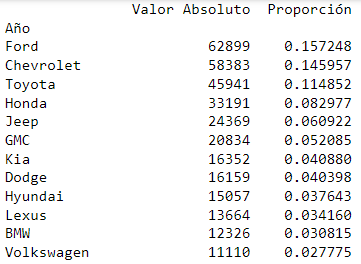
**Gráfico 3. Precios de los vehículos por estado.**



El fabricante que más aparece en la muestra es Ford con el 15.6% de las observaciones, seguido por Chevrolet con el 14.6% y Toyota con el 11.5%. Por otra parte, los fabricantes con los automóviles de mayor precio promedio son Bentley (57256) y Tesla (55646), mientras que el precio menor se encuentra para Pontiac (7167.8). Estos resultados se presentan en el gráfico 4.

**Gráfico 4. Número y proporción de automóviles por fabricante y precio promedio.**



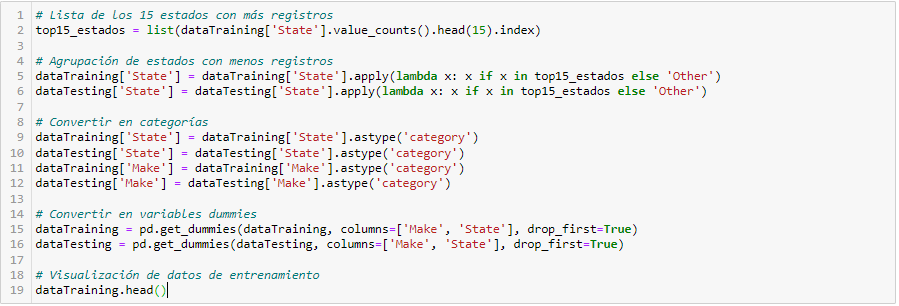


Para la parte de los modelos, debido al alto número de categorías resulta difícil realizar una descripción más detallada. Sin embargo, en el notebook que acompaña el desarrollo de este proyecto se añadió un gráfico interactivo que permite observar la frecuencia de modelos segmentando por fabricantes. En general, el modelo más habitual es el Silverado que corresponde al 4% de los datos de entrenamiento y el Grand que corresponde al 3%.

**Parte 2. Preprocesamiento.**

En el preprocesamiento no se presentó la necesidad de realizar imputación de datos faltantes ni valores atípicos. Dado que las bases de datos ya habían sido separadas en entrenamiento y prueba con anterioridad, tampoco existió la necesidad de incorporar este paso. En general, se hicieron algunas transformaciones. Como se mencionó anteriormente, el número de categorías en algunas variables es alto. Si bien, los modelos a utilizar podrían, en teoría trabajar con este número elevado de variables, ya que el tamaño de las observaciones también lo permite, fue necesario tener en cuenta el tiempo de ejecución de los modelos y, por lo tanto, buscar una estrategia que permitiera disminuir estas categorías. Para el caso de los estados, se crearon variables dicotómicas para los primeros 15 estados con mayor número de observaciones. El resto de los estados se agruparon dentro de la categoría “Other”. Para la variable de fabricante sí se generó una variable dicotómica para cada una de las marcas. Debido al número elevado de categorías, la variable de modelo fue excluida.

**Imagen 1. Transformación de las variables.**

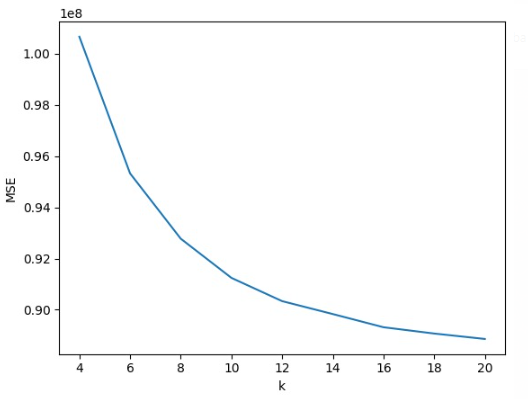


**Parte 3. Estimación y calibración de los modelos.**

En todos los casos que se presentan a continuación, los modelos fueron estimados utilizando la base de prueba y calibrados con indicadores de desempeño calculados a través de la base de entrenamiento.

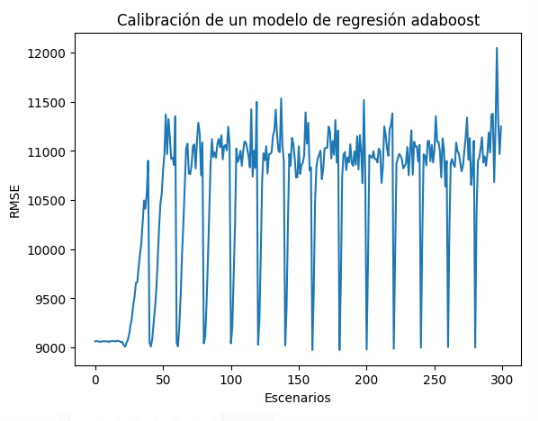
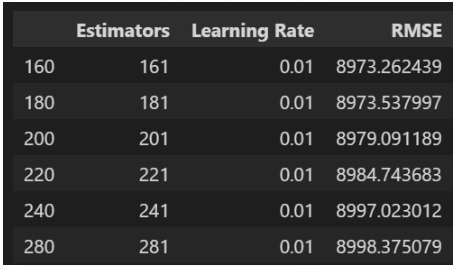
*Modelo de k vecinos:* El primer modelo que se evaluó fue el modelo de k-vecinos. El código correspondiente se encuentra en el notebook que acompaña esta entrega, pero como estrategia de calibración se evaluó el modelo con diferentes números de vecinos. Debido a que el tiempo de procesamiento es elevado por el tamaño del problema, no fue posible evaluar un número mayor de vecinos, pero el mejor resultado se presentó con 20 vecinos, dando un RMSE de 9426,56 y un MAE de 7313 en la base de prueba. El siguiente gráfico muestra cómo disminuye el MSE conforme se aumenta el número de vecinos.

**Gráfico 5. Calibración del modelo de k vecinos cercanos**



*Regresión con AdaBoost:* El segundo modelo evaluado es un modelo de regresión con adaptative boosting. Para este modelo, la estrategia de calibración consistió en probar diferentes escenarios en los cuales se modifica el número de estimadores y la tasa de aprendizaje. De nuevo, el código se adjunta en el notebook, pero en general, la siguiente gráfica da cuenta de las métricas de desempeño obtenidas en los diferentes escenarios. Se observa que el menor RMSE se obtuvo con 161 estimadores y una tasa de aprendizaje de 0.01, llavando a un RMSE de 8973,26, valor inferior al encontrado en el modelo previo, de manera que se mejora el poder predictivo al disminuir las distancias entre los valores reales y predichos.

**Gráfico 6. Calibración del modelo de regresión con AdaBoost.**



**Parte 4. Entrenamiento del modelo**

**Parte 5. Disponibilización del modelo**