ENTREGA 2 - PROYECTO INTRODUCCIÓN A LA INTELIGENCIA ARTIFICIAL

ISABELA PAREJA GIL

JULIÁN DAVID OLAYA RESTREPO

ESTEBAN CARO PELÁEZ

DOCENTE INTRODUCCIÓN A LA INTELIGENCIA ARTIFICIAL

Raúl Ramos Pollán

UNIVERSIDAD DE ANTIOQUIA

FACULTAD DE INGENIERÍA

DEPARTAMENTO DE INGENIERÍA INDUSTRIAL

MEDELLÍN -ANTIOQUIA

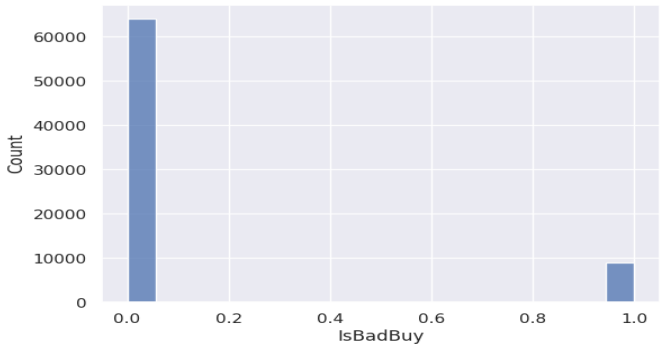
1. Exploración de los datos

El dataset tiene 72983 filas y 34 columnas, y el 6% son datos faltantes (de 2481558 datos, 149185 son faltantes)
Las siguientes son las columnas que tienen datos faltantes, y a su derecha la cantidad.

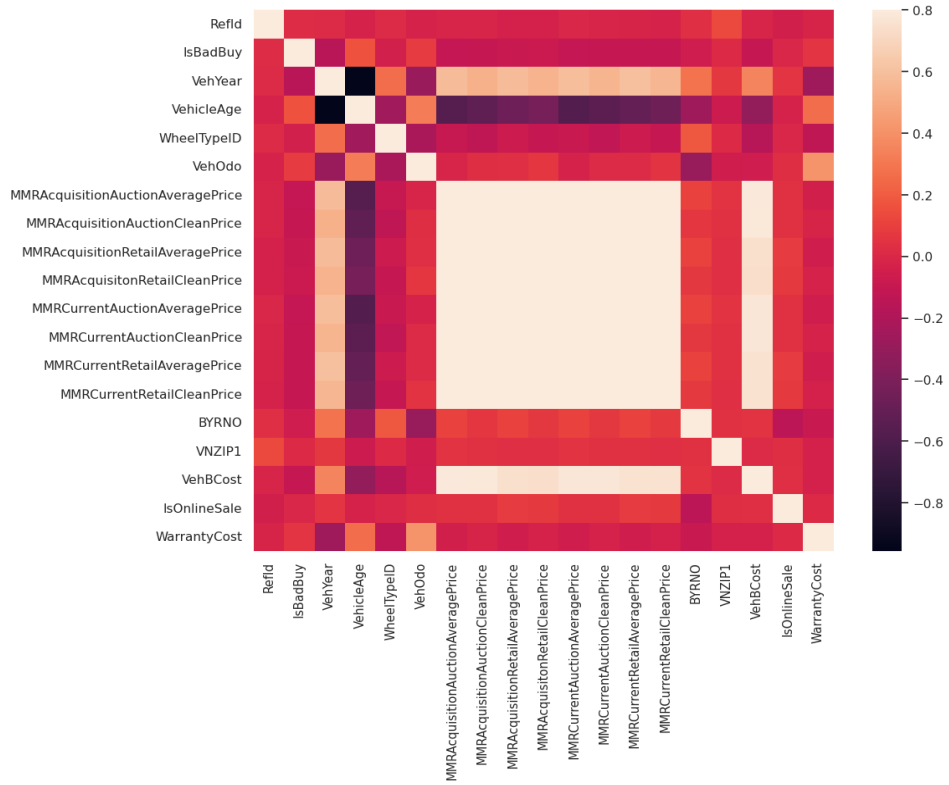
Trim	2360
SubModel	8
Color	8
Transmission	9
WheelTypeID	3169
WheelType	3174
Nationality	5
Size	5
TopThreeAmericanName	5
MMRAcquisitionAuctionAveragePrice	18
MMRAcquisitionAuctionCleanPrice	18
MMRAcquisitionRetailAveragePrice	18
MMRAcquisitonRetailCleanPrice	18
MMRCurrentAuctionAveragePrice	315
MMRCurrentAuctionCleanPrice	315
MMRCurrentRetailAveragePrice	315
MMRCurrentRetailCleanPrice	315
PRIMEUNIT	69564
AUCGUART	69564

El 95% de los datos de PRIMEUNIT y AUCGUART son faltantes, por lo cual se toma la decisión de eliminar estas columnas, ya que los datos existentes no representan mayor impacto en la variable objetivo.

Se realizó un histograma para inspeccionar la variable objetivo, el cual permite identificar con facilidad al ser esta una variable binaria. Se observa que aproximadamente el 12% de las compras fueron una mala decisión (1) y la mayoría fueron buena decisión (0)

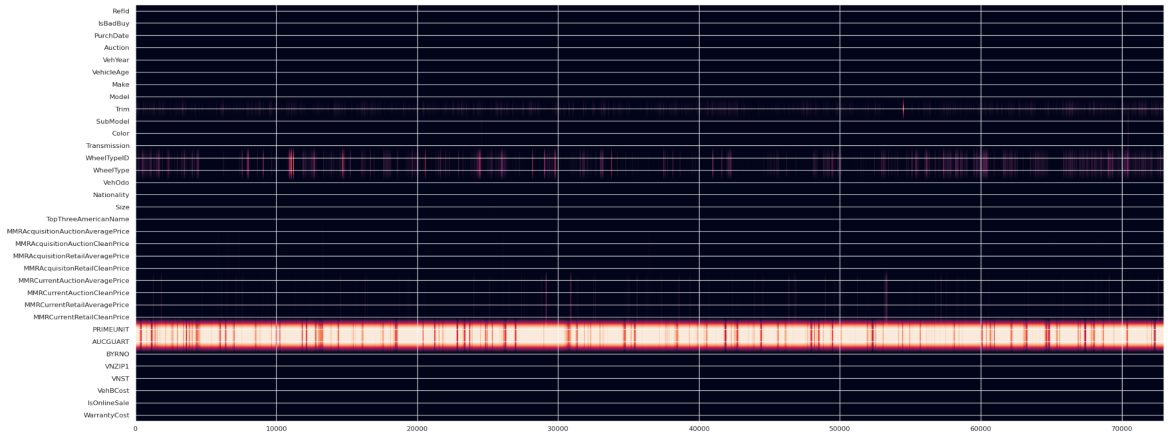


Al analizar la correlación entre las variables, es evidente que la mayoría, no presentan ningún tipo de correlación, por ejemplo nuestra variable objetivo que es IsBadBuy, ésta no presenta ninguna relación significativa. Las variables que mayor correlación presentan son: la de VehYear y VehicleAge, ya que mientras una va aumentando en año, la otra va disminuyendo en edad. Las variables MMRA tienen correlación positiva entre ellas mismas.



Visión de los datos faltantes

La base de datos no cuenta con tantos datos faltantes, como se había mencionado, estos solo corresponden al 6% de todos los datos, de los cuales, la mayor parte son de las variables PRIMEUNIT con 69564 y AUCGUART con 69564 datos faltantes.



2. Limpieza de datos

Variables numéricas

Para los datos faltantes en las variables numéricas, en la mayoría se decidió hallar la media de cada variable, para así permitir que los datos sigan su patrón de comportamiento. Para las variables como el VNZIP que es del código postal, al ser un dato tan específico, se llenaron los faltantes con 0, esto debido a la falta de relevancia. Para la variable IsOnlineSale, al ser binaria, se usó la función np.random. Las variables **PRIMEUNIT,AUCGUART** se eliminan dada la cantidad de datos faltantes que ellas tienen como se observó el gráfico de nulos. La variable **WheelType** nos da la misma información de **WheelTypeID,** razón por la cual decidimos eliminarla. Por otro lado, la variable **BYRNO,** que es un número asignado al comprador del vehículo, se elimina puesto que no es importante en nuestro estudio saber quién compró el auto. Las variables **VehYear** y **VehiceAge** nos dan la misma información, dado que una nos menciona el año de fabricación y otra la edad del auto, por lo cual decidimos eliminar **VehYear.**

Variables categóricas

Se eliminaron las columnas **WheelType** (porque ya había una variable numérica con su información), **PRIMEUNIT** y **AUCGUART** (por tener tantos datos faltantes). Finalmente**,** por el principio de parsimonia, eliminamos **Trim**, **Model** y **SubModel**, puesto que haciendo uso de variables dummies para poder hacer uso de dichas variables, se generan aproximadamente 800 columnas.

3. Modelos

El modelo que primero vamos a implementar es el de regresión logística, ya que es fácil de implementar, interpretar, muy eficiente para el entrenamiento y su resultado, y nuestro problema es esencialmente de clasificación. Se trató de implementar para la entrega, pero no funcionó muy bien por lo cual no sé dejó en el Colab.
Vamos a tratar de implementar otros algoritmos de clasificación como lo es Árboles de decisión, y el super vector machine.

4. Dificultades en el desarrollo del proyecto

A la hora de realizar la limpieza en los datos categóricos, nos encontramos con la incertidumbre de cómo realizarla, ya que por el método Dummies, se crearía un Data Frame muy grande, más de 800 columnas se generan con la variable Trim, por lo cual no lo hicimos, ya que complicaría el desarrollo del modelo.

Por otro lado, en la imputación de datos, estábamos utilizando la función .loc de Python para reemplazar los datos por sus medias o por cero según fuese el caso, y no estábamos utilizando bien dicha función, por lo cual toda la fila que contenía un valor nulo se terminaba reemplazando por dichas medias.