SEGUNDA ENTREGA PROYECTO FINAL

CURSO

INTELIGENCIA ARTIFICIAL PARA CIENCIAS E INGENIERÍAS

PROFESOR

RAÚL RAMOS POLLÁN

PRESENTADO POR

MATEO TORO MOLINA

JOSÉ ALEJANDRO URREGO PABÓN

UNIVERSIDAD DE ANTIOQUIA

FACULTAD DE INGENIERÍA

DEPARTAMENTO DE INGENIERÍA MECÁNICA

OCTUBRE 22 DE 2023

MEDELLÍN

TABLA DE CONTENIDO

1. DESCRIPCIÓN DEL PROBLEMA

Los concesionarios que compran autos en gran cantidad durante las subastas son propensos a comprar autos en mal estado o que necesitan gran cantidad de reparaciones, debido precisamente a que esos autos se camuflan con los que están en buen estado, los concesionarios y la comunidad automovilista llama a estas desafortunadas compras "Kicks" o patadas.

Nosotros, como estudiantes de ingeniería mecánica, hemos detectado que con un pronóstico más preciso usando técnicas de Machine Learning se le podría dar herramientas a los concesionarios sobre posibles autos averiados y, de esta forma, poder ahorrar mucho dinero en compras "patada".

Los autos "pateados" o comprados con alguna avería son aquellos que cuentan con tacómetros manipulados, problemas en la carrocería que son imperceptibles pero que ponen en riesgo la vida del conductor, problemas con el traspaso de la propiedad del vehículo, entre otros. Estos percances hacen que los concesionarios pierdan mucho dinero en tratar de reparar las averías para revender el vehículo, porque inclusive muchas veces la propia avería es más costosa que el precio del auto.

1. ANÁLISIS DE DATOS

El dataset implementado se tomó de una competencia realizada en Kaggle, donde reportan 121756 autos recopilados entre los años 2009 y 2010; dicho dataset cuenta con 33 columnas de variables que permiten la correcta identificación y clasificación del auto.

El 60% del dataset fue destinado para el proceso de entrenamiento y el otro 40% para el proceso de testeo, además, cerca del 40% de las variables son categóricas y el otro 60% son variables numéricas, además, se tienen muchos valores faltantes o nulos en por lo menos 5 columnas del dataset.

El dataset cuenta con 4 archivos, el primero donde se explican cada una de las variables que conforman dichos archivos (Carvana_Data_Dictionary.txt), un archivo destinado al entrenamiento del modelo (Training.csv), otro archivo destinado al testeo de este (test.csv) y un archivo final con las entradas para el testeo (Example_entry.csv) [1]

2. ANÁLISIS DE DATOS

Se realiza un análisis de todos los datos, tanto numéricos como categóricos, para filtrar la cantidad de datos faltantes en cada una de las variables, obteniendo un resultado del 6,19% de datos faltantes en todo el dataset. Luego se procede a analizar las variables categóricas y numéricas individualmente.

2.1. Variables categóricas

Se filtran los datos para obtener la cantidad de datos faltantes entre las variables categóricas, obteniendo un total de 144662 datos faltantes. En base a lo anterior, se procede a elegir 6 variables categóricas que se consideraron las más importantes debido a que son las que, en nuestra opinión, afectan en mayor medida si se realizó una buena o mala compra. Las 6 variables elegidas son: auction, make, color, nationality, size y VNST. Las variables elegidas se contrastan con respecto a la variable IsBadBuy para saber cuantos datos se tienen, obteniendo los siguientes resultados:

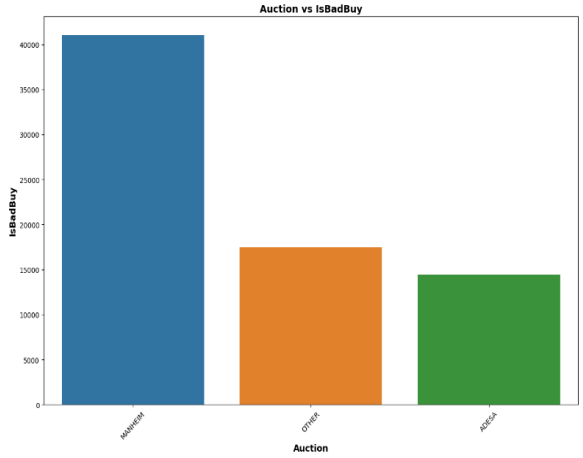

Ilustración 1. Auction vs IsBadBuy

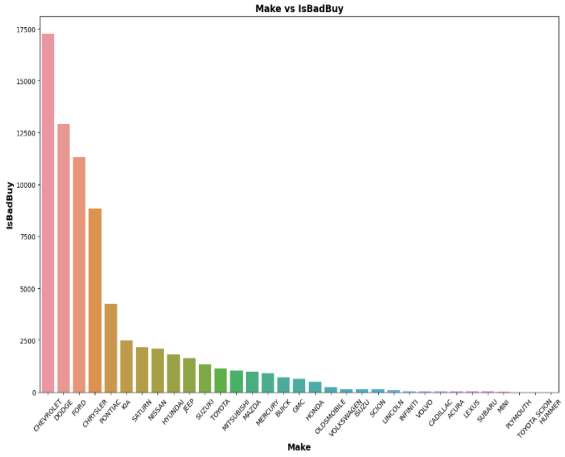

Ilustración 2. Make vs IsBadBuy

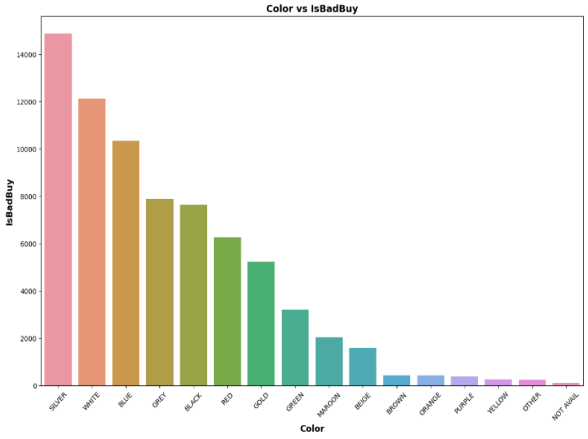

Ilustración 3. Color vs IsBadBuy

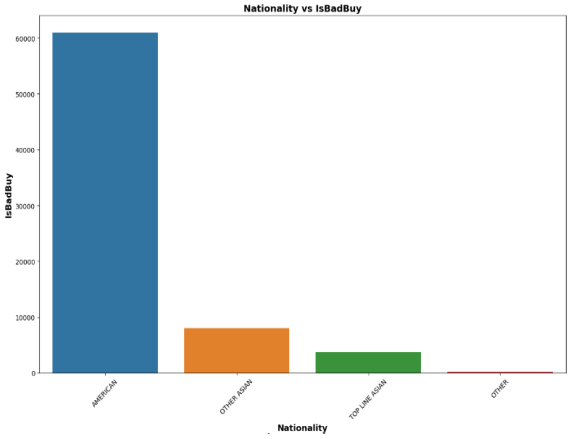

Ilustración 4. Nationality vs IsBadBuy

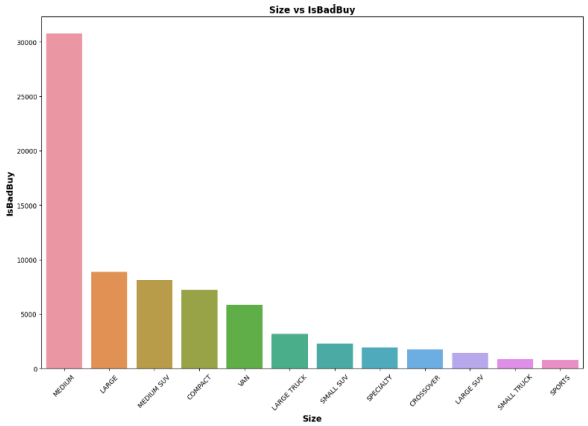

Ilustración 5. Size vs IsBadBuy

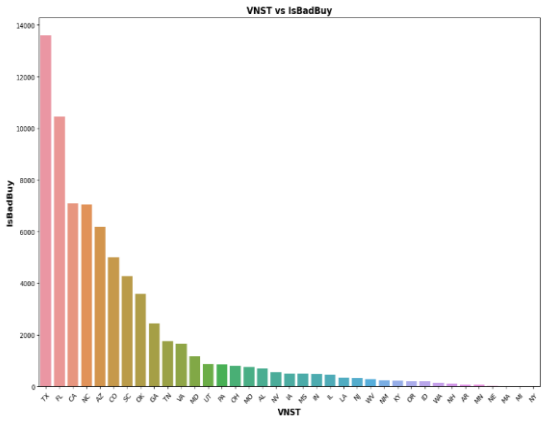

Ilustración 6. Size vs IsBadBuy

2.2. Variables numéricas

Se filtran los datos para obtener la cantidad de datos faltantes entre las variables numéricas, obteniendo un total de 1332 datos faltantes. Luego se procede a calcular las características de cada variable para lograr una comprensión mayor de los datos que se tienen. A continuación, se muestra la relación que tiene cada variable con respecto a las demás:

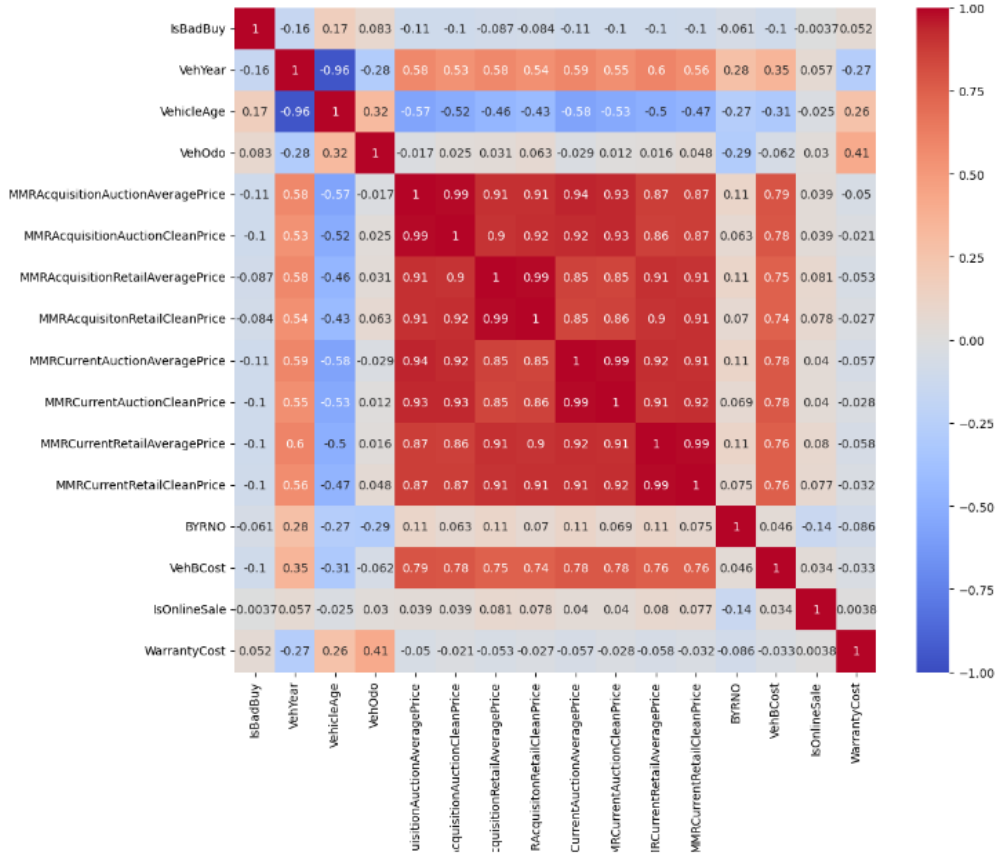

Ilustración 7. Correlación entre los datos

De la ilustración podemos saber la proporcionalidad entre las variables numéricas que se tienen; entre mayor sea el número, mayor será la relación entre las variables. Un número positivo significa que la relación entre las variables es directamente proporcional, un número negativo significa que la relación es inversamente proporcional y una relación de 1 significa que se está relacionando una variable consigo misma.

3. PROBLEMAS PRESENTADOS

Se presentaron varios problemas durante la elaboración del proyecto. Un problema que se tuvo fue la filtración de los datos, ya que se tenían varias variables que no servían debido a una gran cantidad de elementos faltantes, por lo tanto, se tuvo que realizar una filtración y limpieza de datos para obtener variables con datos mas representativos. Otro problema que se presentó fue la elección de las variables a utilizar, debido a que, como compañeros, tuvimos diferencias con respecto a lo que considerábamos variables importantes dentro de los datos que se tenían. Por último, se tuvo una gran dificultad para lograr entender la relación entre las variables que se tienen, pero, gracias al mapa de calor presentado en la ilustración 7, se logró una comprensión adecuada de cada una de las variables y como están relacionadas entre sí.

4. BIBLIOGRAFÍA

- https://www.kaggle.com/competitions/DontGetKicked/overview