Inteligencia Artificial para las Ciencias e Ingenierías
Proyecto - Entrega 2

Karen M. Pérez, Santiago Restrepo, Yilian V. Melgarejo
Estudiantes de Ingeniería Industrial
Universidad de Antioquia, Colombia

1. Introducción

La realización de compras de vehículos usados por concesionarios a través de subastas muestra un panorama de múltiples factores que al final genera que se llegue, por lo tanto, a ciertas conclusiones en el momento de categorizar la adquisición. La razón de esto es que durante su realización es de crucial importancia analizar aspectos particulares de un carro, los cuales permiten establecer si la compra ha sido buena o mala.

A partir de ello, se evidencia una situación referente a la compra de vehículos usados en cuanto a que esta se puede llegar a tramitar a pesar de las condiciones reales del objeto en cuestión, lo anterior gracias a la adulteración de su estado o características como el kilometraje, por ejemplo (La Vanguardia, 2023), o dándose inconvenientes posteriores donde el concesionario puede acabar teniendo problemas no anticipados, tales como fallas disimuladas, asuntos previos pendientes, multas vigentes sin tramitar, entre otros (Ortuya, 2022).

De esta forma, el problema predictivo a tratar consiste en determinar si la compra de un vehículo a través de una subasta es satisfactoria o insatisfactoria y para ello se debería tener en consideración principalmente las propiedades intrínsecas del automóvil y las condiciones externas que lo rodean (proveedor de subastas en el que se compró el vehículo, por ejemplo). Así, se determina que la variable objetivo o de interés para es de tipo clasificatoria donde sus clases son: buena compra (0) o mala compra (1).

Finalmente, para el proceso de predicción se lleva a cabo la utilización de técnicas de aprendizaje supervisado basadas en Machine Learning, necesitando para esto un conjunto de datos que permita tanto entrenar como calibrar el modelo para así conseguir que las predicciones resultantes sean lo más acertadas posible y que las métricas de desempeño aplicadas den óptimamente.

2. Exploración descriptiva del dataset

Inicialmente el dataframe de entrenamiento (bautizado en el Notebook de Colab como df_kick) cuenta con 72983 observaciones (filas) y 34 variables (columnas). Este dataframe tiene como variable objetivo la columna ‘IsBadBuy’, la cual es una variable binaria donde 0 corresponde a que el carro comprado en la subasta no representa una mala compra y 1 corresponde a que sí corresponde a una mala compra. Dicha variable objetivo está

representada por aproximadamente 88% de valores 0 (64007 observaciones) y el 12% restante equivale a valores 1 (8976 observaciones) (ver *Ilustración 1*).

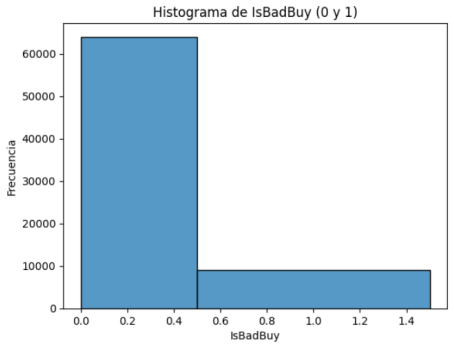

Ilustración 1. Histograma de la variable objetivo 'IsBadBuy'

Además, al realizar una descripción de los datos se encuentra que 19 de las 32 columnas son numéricas y las restantes son categóricas o fechas. Es por eso que al ser numéricas es posible obtener estadísticas importantes (ver *Ilustración 2*).

	count	mean	std	min	25%	50%	75%	max
RefId	72983.0	36511.428497	21077.241302	1.0	18257.5	36514.0	54764.5	73014.0
IsBadBuy	72983.0	0.122988	0.328425	0.0	0.0	0.0	0.0	1.0
VehYear	72983.0	2005.343052	1.731252	2001.0	2004.0	2005.0	2007.0	2010.0
VehicleAge	72983.0	4.176644	1.712210	0.0	3.0	4.0	5.0	9.0
WheelTypeID	69814.0	1.494299	0.521290	0.0	1.0	1.0	2.0	3.0
VehOdo	72983.0	71499.995917	14578.913128	4825.0	61837.0	73361.0	82436.0	115717.0
MMRAcquisitionAuctionAveragePrice	72965.0	6128.909217	2461.992768	0.0	4273.0	6097.0	7765.0	35722.0
MMRAcquisitionAuctionCleanPrice	72965.0	7373.636031	2722.491986	0.0	5406.0	7303.0	9021.0	36859.0
MMRAcquisitionRetailAveragePrice	72965.0	8497.034332	3156.285284	0.0	6280.0	8444.0	10651.0	39080.0
MMRAcquisitonRetailCleanPrice	72965.0	9850.928240	3385.789541	0.0	7493.0	9789.0	12088.0	41482.0
MMRCurrentAuctionAveragePrice	72668.0	6132.081287	2434.567723	0.0	4275.0	6062.0	7736.0	35722.0
MMRCurrentAuctionCleanPrice	72668.0	7390.681827	2686.248852	0.0	5414.0	7313.0	9013.0	36859.0
MMRCurrentRetailAveragePrice	72668.0	8775.723331	3090.702941	0.0	6536.0	8729.0	10911.0	39080.0
MMRCurrentRetailCleanPrice	72668.0	10145.385314	3310.254351	0.0	7784.0	10103.0	12309.0	41062.0
BYRNO	72983.0	26345.842155	25717.351219	835.0	17212.0	19662.0	22808.0	99761.0
VNZIP1	72983.0	58043.059945	26151.640415	2764.0	32124.0	73108.0	80022.0	99224.0
VehBCost	72983.0	6730.934326	1767.846435	1.0	5435.0	6700.0	7900.0	45469.0
IsOnlineSale	72983.0	0.025280	0.156975	0.0	0.0	0.0	0.0	1.0
WarrantyCost	72983.0	1276.580985	598.846788	462.0	837.0	1155.0	1623.0	7498.0

Ilustración 2. Descripción de datos numéricos del dataframe

Gracias a esto, es posible realizar un análisis de correlación con el gráfico de la matriz de correlaciones (ver *Ilustración 3*).

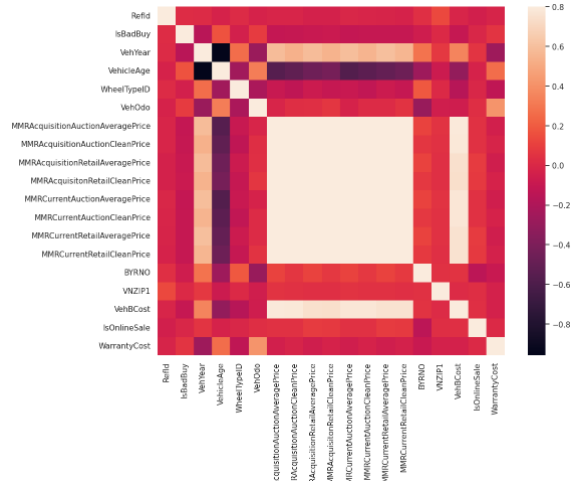

Ilustración 3. Matriz de correlación datos numéricos 'don't get kicked'

De la cual se puede concluir que las variables definidas como ‘MMR’, ‘Vehyear’ y ‘VehBCost’ representan una alta correlación con la variable objetivo.

3. Iteraciones de desarrollo. Para cada iteración, incluye los elementos que consideres de los siguientes:

- **Preprocesado de datos**

- **Eliminación de variables (columnas):** Gracias a la exploración de los datos se determinó que existen algunas variables o columnas que contienen datos que realmente no llegan a ser relevantes para un modelo predictivo. Dichas variables son ‘PRIMEUNIT’, ‘AUCGUART’, ‘RefId’, ‘PurchDate’, ‘BYRNO’, ‘VNZIP1’, ‘Trim’ y ‘WheelTypeID’, las cuales son eliminadas del dataframe.

- **Datos duplicados:** El dataframe no posee datos duplicados.

- **Datos faltantes**: El dataframe posee datos faltantes en varias de sus columnas (ver *Ilustración 4*).

  En el caso de ‘Color’, ‘Transmission’ y ‘Size’ se realiza la imputación de sus datos faltantes con la moda dado que son pocos datos y no generaría problemas reemplazarlos con estos valores.
  La variable ‘WheelType’ también es imputada con su moda dado que el tipo de llanta de un carro no debería ser problemático al momento de comprar un vehículo en subasta.
  Por otro lado, las 8 variables comenzadas por ‘MMR’ son imputadas con su media dado que no se cuenta con información suficiente para hacer otro tipo de imputación. Además, parecen ser variables relevantes dentro del modelo y por ende, no pueden ser eliminadas.

```
IsBadBuy                              0
Auction                               0
VehYear                               0
VehicleAge                            0
Make                                  0
Model                                 0
SubModel                              8
Color                                 8
Transmission                          9
WheelType                          3174
VehOdo                                0
Nationality                           5
Size                                  5
TopThreeAmericanName                  5
MMRAcquisitionAuctionAveragePrice    18
MMRAcquisitionAuctionCleanPrice      18
MMRAcquisitionRetailAveragePrice     18
MMRAcquisitonRetailCleanPrice        18
MMRCurrentAuctionAveragePrice       315
MMRCurrentAuctionCleanPrice         315
MMRCurrentRetailAveragePrice        315
MMRCurrentRetailCleanPrice          315
VNST                                  0
VehBCost                              0
IsOnlineSale                          0
WarrantyCost                          0
dtype: int64)
```

Ilustración 4. Datos faltantes

  Para el caso de ‘Nationality’ y ‘TopThreeAmericanName’ se realiza la imputación con valores un valor específico de su columna (‘OTHER’) dado que no se obtiene información para imputarlos de otra forma, y al clasificar sus datos faltantes en ‘OTHER’ no generaría problema, aparte son muy pocos datos faltantes.
  Finalmente, se eliminan las observaciones con datos faltantes en la columna ‘SubModel’ ya que son muy pocos datos e imputarlos con datos incorrectos podría generar problemas.

- **Modelos supervisados**

  Antes de ejecutar el modelo, es importante mencionar que el dataframe resultante del preprocesado de datos se llevó a formato dummy a través de .get_dummies(),

obteniendo así un dataframe ‘df_kick_dummies’ con 72975 observaciones y 2056 columnas. Además, se realiza el escalado de los datos a través de StandardScaler() y se dividen los datos en 40% de entrenamiento y 60% de prueba.

- **Regresión logística:** Para aplicar un modelo de regresión logística (clasificador lineal) se hace uso de LogisticRegression() entrenado con los datos de training y el .predict() para poder ejecutar la predicción sobre los datos de prueba.
- **Random forest:** En el caso de los bosques aleatorios se hace uso de RandomForestClassifier() que implementa por defecto 100 árboles de decisión para su ejecución y se someten a entrenamiento y predicción los mismos datos de la regresión logística.

- **Resultados**
- **F1 Score:** En el caso de la regresión logística se obtiene un valor de aproximadamente 0.08, y en el caso del bosque aleatorio se obtiene un valor aproximado de 0.03, lo que permite concluir que los datos aún requieren de ajuste y manipulación; esto, teniendo en cuenta que los valores de F1 Score cercanos a 0 determinan un resultado deficiente por parte del modelo.
- **Accuracy:** A pesar de que nuestra métrica de desempeño no es el accuracy, también se encontró este valor para los modelos aplicados. En el caso de la regresión logística se alcanzó un accuracy del 87% y para el bosque aleatorio se obtuvo un accuracy de 88%.

4. Conclusiones

- Los resultados obtenidos permiten concluir que el dataframe aún requiere de ajuste y manipulación dado que no arroja predicciones tan coherentes si se analiza el F1 Score, sin embargo, al observar los valores obtenidos de accuracy se evidencia que son sólidos. Esto se puede deber a que es una base de datos desequilibrada, ya que como veíamos al inicio, aproximadamente el 88% de la variable objetivo posee un valor de 0 (no es una mala compra), lo que genera que el modelo obtenga casi siempre una predicción de este estilo.
- Se debería analizar de manera más precisa la correlación existente entre variables y los comportamientos de cada una de ellas para determinar cuáles son las variables realmente influyentes dentro de un proceso de predicción.

5. Referencias Bibliográficas

La Vanguardia. (3 de enero de 2023). *El timo del kilometraje en los coches de segunda mano: estos son los modelos más trucados.* https://www.lavanguardia.com/motor/actualidad/20230103/8661617/timo-kilometraje-trucado-coches-segunda-mano-son-modelos-mas-manipulados.html

Ortuya, N. (1 de junio de 2022). *Riesgos de comprar un carro usado y cómo evitarlos*. Autofact. https://www.autofact.com.co/blog/comprar-carro/consejos/riesgo-carro-usado