

Inteligencia Artificial para las Ciencias e Ingenierías

Proyecto - Entrega 1

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

1. Definición del problema predictivo

La compra de vehículos usados por parte de concesionarios a través de subastas permite llegar a conclusiones totalmente diferentes a la hora de calificar la compra. Esto dado a que durante la misma es de vital importancia tener en cuenta aspectos característicos de un carro que podrían determinar si la compra se puede considerar beneficiosa (buena) o mala.

Desde aquí, se puede observar una problemática en cuanto la adquisición de carros que muchas veces pueden estar en condiciones cuestionables y que sin embargo, se tramita de igual manera la compra debido a la adulteración de condiciones en este, tales como su kilometraje, por ejemplo (La Vanguardia, 2023) o donde después, el concesionario puede acabar teniendo problemas no previstos con este; fallas disimuladas, antecedentes no tratados como multas vigentes, entre otros (Ortuya, 2022).

Por lo tanto, el problema predictivo está basado en determinar si la compra de un vehículo a través de una subasta es buena o mala. Para ello se debería tener en cuenta las diferentes características tanto del vehículo como del adjudicatario para poder realizar la predicción más acertada. Por ende, la variable objetivo o de interés es de tipo clasificatoria donde sus clases son: buena compra o mala compra.

Adicionalmente, para realizar el proceso de predicción se hace uso de herramientas predictivas enmarcadas dentro del área de Machine Learning (ML), y para ello se requiere de un conjunto de datos que permita entrenar y calibrar el modelo para obtener así predicciones lo más acertadas posible.

2. Dataset

El dataset utilizado se titula 'Don't get kicked' ([enlace](#)) y es de una competencia de Kaggle que tiene como desafío predecir si el auto comprado en la subasta es una buena o mala compra. En ella se proporcionan datos referentes a las condiciones y compras acompañados de variables sobre los vehículos, además, el dataset se encuentra dividido en un 60% como datos de entrenamiento y un 40% como datos de prueba. Cuenta con 34 columnas y un total de 72983 filas y tiene como objetivo la variable 'IsBadBuy' (que clasifica la compra como buena (no es mala) o mala). Las variables del dataset se definen a continuación:

- **RefID** - Número secuencial asignado al vehículo

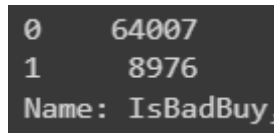
- **IsBadBuy** - Identifica si el vehículo era una compra evitable
- **PurchDate** - Fecha de compra del vehículo en la subasta
- **Auction** - Proveedor de subastas en el que se compró el vehículo
- **VehYear** - Año de fabricación de vehículo
- **VehicleAge** - Años transcurridos desde su año de fabricación
- **Make** - Fabricante del vehículo
- **Model** - Modelo del vehículo
- **Trim** - Nivel de equipamiento del vehículo
- **SubModel** - Submodelo del vehículo
- **Color** - Color del vehículo
- **Transmission** - Tipo de transmisión de vehículos (automática o manual)
- **WheelTypeID** - Tipo de identificación de la ruedas del vehículo
- **WheelType** - Descripción del tipo de rueda del vehículo (aleación o cubierta)
- **VehOdo** - Lectura del odómetro de los vehículos
- **Nationality** - País de fabricación
- **Size** - Categoría del tamaño del vehículo (Compacto, SUV, etc)
- **TopThreeAmericanName** - Identifica si el fabricante es uno de los 3 principales fabricantes estadounidenses
- **MMRAcquistionAuctionAveragePrice** - Precio de adquisición del vehículo en condiciones medias al momento de la compra.
- **MMRAcquistionAuctionCleanPrice** - Precio de adquisición del vehículo en condiciones medias anterior al momento de la compra
- **MMRAcquistionRetailAveragePrice** - Precio de adquisición del vehículo en el mercado minorista en condiciones medias al momento de la compra
- **MMRAcquistionRetailCleanePrice** - Precio de adquisición del vehículo en el mercado minorista en condiciones superiores a la media en el momento de la compra
- **MMRCurrentAuctionAveragePrice** - Precio de adquisición del vehículo en condiciones medias a partir del día actual
- **MMRCurrentAuctionCleanPrice** - Precio de adquisición del vehículo en condiciones anteriores a partir del día actual
- **MMRCurrentRetailCleanPrice** - Precio de adquisición del vehículo, en el mercado minorista en condiciones promedio a partir del día actual
- **PRIMEUNIT** - Identifica si el vehículo tendría una demanda más alta que una compra estándar
- **Acquisition Type** - Identifica cómo se adquirió el vehículo (compra de subastas, intercambio, etc.)
- **AUCGUART** - La garantía de nivel proporcionada por la subasta para el vehículo (Luz verde - Garantizada/ arbitrable, Luz Amarilla- Precaución/problema, Luz roja - se vende como está)
- **KickDate** - Fecha en la que el vehículo se devolvió a la subasta
- **BYRNO** - Número único asignado al comprador del vehículo
- **VNZIP** - Código postal donde se compró el vehículo
- **VNST** - Indica el Estado de Estados Unidos dónde se compró el vehículo

- **VehBCost** - Costo de adquisición pagado por el vehículo en el momento de la compra
- **IsOnlineSale** - Identifica si el vehículo se compró originalmente en línea
- **WarrantyCosts** - Precio de garantía (término = 36 meses y kilometraje = 36km)

3. Métricas de desempeño

3.1. Métrica de desempeño - Machine Learning (ML)

Para medir el desempeño del modelo de Machine Learning se hará uso de la métrica **F1-Score**. Se decide utilizar esta métrica ya que nos encontramos al frente de un conjunto de datos con resultados desbalanceados o desequilibrados (ver *Imagen 1*), que nos permite apreciar que, según los datos de entrenamiento más del 85% corresponde a buenas compras (predice el valor 0). Dicho esto, la métrica F1-score es la media armónica entre la precisión y la sensibilidad de un modelo clasificatorio, y evita las imprecisiones que genera utilizar la media aritmética tradicional (Bühl, 2019), lo que permite definir que F1 es una métrica favorable para medir el desempeño de conjuntos de datos que estén desequilibrados.



```
0      64007
1      8976
Name: IsBadBuy
```

Imagen 1. Predicciones de los datos de entrenamiento ('Don't get kicked')

Dicho esto, la media armónica utilizada genera una medida equilibrada entre precisión y sensibilidad, donde F1-Score se calcula (ver *Ecuación 1*) de la siguiente manera (Bühl, 2019):

$$F1 = 2 \left(\frac{\text{precisión} * \text{sensibilidad}}{\text{precisión} + \text{sensibilidad}} \right)$$

Ecuación 1. Ecuación de F1-Score

Donde la precisión (ver *Ecuación 2*) y la sensibilidad (ver *Ecuación 3*) se calculan así (Bühl, 2019):

$$\text{precisión} = \frac{VP}{VP+FP}$$

Ecuación 2. Ecuación de precisión

$$\text{sensibilidad} = \frac{VP}{VP+FN}$$

Ecuación 3. Ecuación de sensibilidad

Donde:

VP (verdaderos positivos): corresponde al caso donde el modelo predice correctamente que la compra fue mala (1)	VN (verdaderos negativos): corresponde al caso donde el modelo predice correctamente que la compra no fue mala (0)
FP (falsos positivos): corresponde al caso donde el modelo predice incorrectamente que la compra fue mala (1) pero realmente la compra no fue mala (0)	FN (falsos negativos): corresponde al caso donde el modelo predice incorrectamente que la compra no fue mala (0) pero realmente la compra fue mala (1)

Tabla 1. Análisis de precisión y sensibilidad del conjunto de datos ‘Don’t get kicked’

Además, F1-Score es un valor representado entre 0 y 1, donde 0 representa el resultado más deficiente y 1 representa el valor más excelente, y generalmente una puntuación alta de F1 representa un alto grado de precisión y sensibilidad simultáneamente por parte del modelo, y una puntuación baja de F1 representa dificultades para alcanzar ese equilibrio (Bühr, 2019).

3.2. Métrica de desempeño - Negocio

Como métrica de negocio se plantea un indicador de eficiencia asociado a la decisión de si invertir o no por parte de una empresa. Para ello, tener en cuenta el siguiente escenario:

Suponga que existe una empresa dedicada a la compra de vehículos para realizar reformas en él y venderlos una vez se modifiquen. Dicha empresa cuenta con un departamento dedicado a tomar la decisión de invertir o no en un vehículo, cuánto pagar por él, cuánto sería el costo a incurrir en términos de modificación y reformas del mismo, etc.

Lo ideal sería que éste departamento sea lo más eficiente y productivo posible, por lo que se podrían plantear los siguientes indicadores (KPI):

- **Indicador de ganancia económica** (ver Ecuación 4): Indicador que permita identificar el margen de ganancia económica obtenido al realizar una inversión en un vehículo.

$$KPI_{ganancia} = \text{dinero que ingresa venta del vehículo} - \text{dinero gastado en el vehículo}$$

Ecuación 4. KPI medidor de ganancia

Donde el dinero gastado en el vehículo representa el precio por el que adquirió el vehículo y la cantidad de dinero invertida para reformar el mismo. Y el dinero que ingresa por la venta del vehículo corresponde al precio por el cual se dio la venta del mismo.

- **Tasa de éxito en inversiones** (ver Ecuación 5): Indicador capaz de identificar o medir cuántas de las decisiones de inversión tomadas por el departamento representaron una ganancia para la empresa. Éste indicador se representa como un porcentaje y se esperaría que su valor sea lo más alto posible.

$$KPI_{\text{éxito}} = \frac{\# \text{ decisiones de inversión que representaron ganancia}}{\# \text{ de decisiones de inversión tomadas}} \times 100$$

Ecuación 5. KPI medidor de inversiones exitosas

4. Desempeño esperado

Partiendo del **indicador de ganancia económica**, se debe tener presente que una situación esperada es que, primero, el precio por el que se adquirió el vehículo sea el adecuado (sobre todo teniendo en cuenta su estado en el momento de la compra); segundo y por ende, que este mismo no presente deterioros o que la empresa deba realizarle demasiadas o costosas modificaciones inesperadas, ya que se puede incurrir a tener mayores gastos (*dinero gastado en el vehículo*) en comparación con lo que una persona puede estar abierta a pagar por un auto usado (*dinero que ingresa por venta del vehículo*), por ejemplo.

Así, lo que se esperaría es un comportamiento positivo y creciente de este indicador, en el que los gastos aplicados al vehículo sean siempre menores que la retribución al vender este. Entonces, en esta parte es donde la implicación favorable del modelo tiene un papel fundamental, debido a que si este es capaz de detectar de manera efectiva cuáles son buenas oportunidades de compra y cuáles no, se evitaría el comprar un carro en supuestas buenas condiciones, cuando en realidad es lo contrario y por ende, la inversión en arreglos en este sería al final, mayor, y posiblemente, la ganancia no sería la requerida o esperada para la empresa.

Por otro lado, en cuanto a la **tasa de éxito en inversiones**, esta se encuentra muy relacionada con lo anteriormente explicado, debido a que si el modelo (en caso ideal) tiene un nivel de precisión alto, esta daría como resultado un gran porcentaje de compras buenas de autos usados en cuanto calidad, condiciones y precio y por lo tanto, eso podría significar una menor inversión en reformas, por lo que el indicador de ganancia podría dar positivo, y finalmente, la tasa de éxito en inversiones aumentaría, ya que la mayor parte de las decisiones tomadas representan ganancias. Lo ideal sería que esta tasa sea lo más cercana posible al 100%, sin embargo, se podría definir que como mínimo esta tasa debe estar por encima del 70% ya que si este porcentaje es inferior empezaría ser problemático si se tiene en cuenta que entre más alejado esté del 100% pues representaría que se están tomando malas decisiones de inversión por parte del equipo.

5. Referencias Bibliográficas

Buhl, N. (18 de julio de 2023). *F1 Score in Machine Learning*. ENCORD.
<https://encord.com/blog/f1-score-in-machine-learning/>

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>