PROYECTO MACHINE LEARNING

MEMORIAS

El presente trabajo tiene como objetivo analizar el mercado de los coches usados en Estado Unidos, esto con el fin de lograr predecir el precio usando técnicas de Machine Learning y determinar cuál es el mejor modelo para dicho propósito.

Para lo anterior fue necesario usar un dataset que se descargó de Kaggel, sin embargo fue necesario realizar una limpieza de los datos utilizando las herramientas aprendidas en el curso. Realizando un primer vistazo de los datos nos damos cuenta que se necesitan eliminar las columnas ('Unnamed: 0','vin', 'lot') ya que no nos aportan al análisis. Adicionalmente se transforma la columna ‘mileage’ a kilómetros para una mayor comprensión y se convierte la columna ‘year’ a la cantidad de años del vehículo para lograr una mayor comprensión del dataset; por último, teniendo en cuenta que no contamos con una cantidad de datos muy amplia maximizamos la capacidad de análisis transformando las variables categóricas a numéricas.
En este punto comenzamos a graficar las variables, para este fin usaremos un boxplot para determinar si las variables precio y brand cuentan con alguna relación. Es posible observar la tendencia del mercado de vehículos en Estados Unidos; vemos que marcas como Ford, Chevrolet, BMW, Cadillac y Lexus son vistas como marcas de lujo y que mayoritariamente los vehículos son de grandes dimensiones.

En las siguientes gráficas observamos la relación entre precio y año así como entre precio y kilometraje, lo cual nos muestra que existe una estrecha relación entre estas variables, es decir tanto el kilometraje como los años afectan el precio de los vehículos. en otras palabras la correlación es negativa moderada entre las variables ‘price’ y ‘year’, lo cual significa que generalmente a medida que aumenta el año del vehículo el precio tiende a disminuir, sin embargo también podemos observar que los vehículos más nuevos tienden a depreciarse más rápido aunque esto puede deberse al tamaño del dataset éstos se encuentran menos representados.

Coeficiente de Pearson price / year:

- Coeficiente de Pearson: -0.3199722034249613
- Valor p: 9.30762902833853e-57

Coeficiente de Pearson precio / Km:

- Coeficiente de Pearson: -0.361493250493227

- Valor p: 4.816421511041697e-73

Adicionalmente se observa que el p_value es bajo, lo que indica una evidencia estadística fuerte contra la hipótesis nula, lo cual sugiere que la correlación negativa entre price / year y price / km es significativa.

Ya teniendo nuestro 'target' claro iniciamos con el análisis de encontrar cual es el mejor modelo predictivo para nuestro caso en concreto. iniciamos evaluando una regresión lineal con L1 y L2. Resultados :

Regresión Lineal

- MSE: 110980519.44682021, R2: 0.15673439918134902 , MAPE:0.7904462559562402.

Error Cuadrático Medio (MSE):
El MSE es bastante alto, lo que indica que hay una diferencia considerable entre los valores predichos y los valores reales. Esto sugiere que el modelo tiene un error de predicción significativo.

Coeficiente de Determinación (R2):
El R2 es muy bajo. Este valor indica que el modelo sólo explica aproximadamente el 15.67% de la variabilidad en los datos. En otras palabras, el 84.33% de la variación en el precio no se explica por las variables independientes (año y kilometraje) en este modelo lineal.

Error Porcentual Absoluto Medio (MAPE):
El MAPE de 0.7904 (79.04%) es extremadamente alto, lo que significa que, en promedio, las predicciones del modelo se desvían un 79.04% de los valores reales. Un MAPE tan alto indica que el modelo tiene un rendimiento muy pobre en términos de precisión de predicción.

Dado el rendimiento tan deficiente del modelo intentaremos poner a prueba distintos modelos para conseguir el mejor resultado. A continuación se presentan los diferentes modelos con sus resultados.

con un modelo de regresión lineal, en consecuencia los resultados son los siguientes:

Regresión Polinomial

- MSE: 109590395.668717, R2: 0.1672970057432711, MAPE:0.7724752971182547

Error Cuadrático Medio (MSE):
El MSE es ligeramente menor que el de la regresión lineal (110,980,519.45). Esto indica una pequeña mejora en la precisión de las predicciones, aunque sigue siendo un valor alto.

Coeficiente de Determinación (R2):
El R2 es marginalmente mejor que el de la regresión lineal (0.1567). El modelo polinomial explica aproximadamente el 16.73% de la variabilidad en los datos, lo que sigue siendo un porcentaje bajo.

Error Porcentual Absoluto Medio (MAPE):
El MAPE (77.25%) es ligeramente mejor que el de la regresión lineal (79.04%), pero sigue siendo muy alto. Esto indica que, en promedio, las predicciones del modelo polinomial se desvían un 77.25% de los valores reales.

Random Forest

Error cuadrático medio: 44753135.51773846, R2 Score: 0.6599513148884841, MAPE:0.4219039936124184

Rendimiento moderado: El R2 Score de 0.6599 indica que el modelo explica aproximadamente el 66% de la variabilidad en los datos1. Esto sugiere un rendimiento moderado, pero hay margen para mejoras.

Error considerable: El Error Cuadrático Medio (ECM) de 44,753,135.52 es relativamente alto. Este valor está en las unidades al cuadrado de la variable objetivo (probablemente el precio del vehículo), lo que indica que las predicciones pueden desviarse significativamente de los valores reales.

Sesgo en las predicciones: El MAPE (Error Porcentual Absoluto Medio) de 0.4219 sugiere que, en promedio, las predicciones del modelo tienen un error del 42.19%. Esto indica un sesgo considerable en las estimaciones.

Al graficar la importancia de las características en este modelo vemos que le dio una gran importancia a la variable Km muy por encima de la variable ‘year’. lo que nos da a entender que no necesariamente entre mayor la cantidad de años mayor kilometraje, es así que se observa que el mercado de automóviles es complejo y puede haber factores adicionales que influyen en el precio, como la marca, el modelo y las condiciones del mercado.

Para profundizar en el análisis usaremos un Grid Search para cuatro modelos diferentes, con el fin de que encuentre la mejor configuración de hiperparametros para cada modelo y nos permita ser las exactos en poder predecir el precio de los automóviles.

Es así que los resultados obtenidos bajo esta nueva perspectiva son los siguientes.

Mejor rendimiento general: El modelo de Gradient Boosting muestra el mejor rendimiento en general, con el MAE más bajo (4162.74), el RMSE más bajo (6386.08), el R2 más alto (0.6901) y el MAPE más bajo (0.3956).

1. Comparación de modelos:
    - Gradient Boosting > Linear Regression > Random Forest > Decision Tree.
2. Precisión de las predicciones:
    - El Gradient Boosting tiene un error promedio de alrededor de $4,163 (MAE) y un error porcentual promedio del 39.56% (MAPE).
    - El modelo lineal tiene un rendimiento sorprendentemente bueno, superando a Random Forest y Decision Tree.
3. Variabilidad explicada:
    - El Gradient Boosting explica aproximadamente el 69% de la variabilidad en los datos (R2 = 0.6901).
    - El modelo lineal explica cerca del 66% de la variabilidad.
4. Complejidad vs. rendimiento:
    - El Decision Tree tiene el peor rendimiento, lo que sugiere que puede estar subajustando los datos.
    - Random Forest mejora sobre el Decision Tree, pero no supera al modelo lineal, lo que indica que la relación puede ser más lineal de lo esperado.
5. Margen de mejora:
    - Incluso el mejor modelo (Gradient Boosting) tiene un MAPE de 39.56%, lo que sugiere que hay margen para mejorar las predicciones.

En resumen, el Gradient Boosting ofrece las mejores predicciones para este conjunto de datos, pero todos los modelos muestran errores significativos. Esto sugiere que podría ser beneficioso explorar técnicas adicionales de ingeniería de características o considerar otros algoritmos más avanzados para mejorar aún más las predicciones.
Para esto y teniendo en cuenta los resultados anteriores del mejor modelo aplicaremos un PCA para obtener una mejora considerable, esto nos da como resultado que el modelo se

ve afectado negativamente ya que se necesiten 204 componentes principales para explicar el 95% de la varianza indica que el conjunto de datos es altamente dimensional y complejo.
Por su parte el rendimiento del modelo (MAPE) del modelo Gradient Boosting con PCA es de 0.5129, lo que significa que, en promedio, las predicciones tienen un error del 51.29%.
En comparación con el modelo original: El MAPE del modelo Gradient Boosting original era de 0.3956 (39.56%). Al aplicar PCA, el MAPE aumentó a 0.5129 (51.29%), lo que indica un empeoramiento significativo en el rendimiento del modelo. Teniendo esto en cuenta, la aplicación de PCA, aunque reduce la dimensionalidad, parece haber eliminado información relevante para las predicciones, resultando en un modelo menos preciso.

En conclusión podemos observar que el predecir el precio de un vehículo es una tarea más difícil de lo que aparenta ya que pueden existir muchos factores diferentes que pueden afectar el precio, adicionalmente se recomienda evaluar a futuro las condiciones del mercado, teniendo lo anterior en mente vemos que el mejor modelo para predecir el precio es el gradient boosting, sin embargo, si bien es una gran ayuda para el propósito que en este trabajo nos atañe, tiene una gran oportunidad de mejora si se tienen en cuenta las variables que puedan enriquecer el análisis.