Machine Learning En Investigación Del Mercado Automotor

El mercado automotor está muy ligado a la cultura de cada país, según los gustos de cada uno, el mercado norteamericano, por ejemplo, valora mucho los motores y vehículos muy grandes, el mercado europeo prefiere el bajo consumo, el mercado latinoamericano, los precios finales bajos y así varía según región, país, nivel socioeconómico y cultura. Un mismo vehículo puede tener un valor muy distinto de un país al otro, y no solo por los impuestos o costos de producción, sino por cómo cotiza el modelo en el mercado.

Planteamiento de la problemática

Hemos sido contratados en el equipo de ciencias de datos en una consultora de renombre. Nos han asignado a un proyecto de estudio de mercado de una importante automotriz china. Nuestro cliente desea ingresar a nuestro mercado de automóviles, por lo que nos han encomendado analizar las características de los vehículos presentes en el mercado actual. Dado que tienen en su catálogo una amplia colección de modelos de todo tipo, cuyo catálogo está estratificado en gamas según el gusto de cada región, desean saber qué características presentan los vehículos de gama alta y los de gama baja en nuestro mercado, para poder abarcar todo los públicos objetivos ajustándose a toda la demanda y, en base a estos datos, poder cotizar correctamente los vehículos que ofrecerá.

Para ello, nuestro departamento de datos ha recopilado precios y características de varios de los modelos de vehículos disponibles en nuestro mercado, junto con sus

precios de venta al público y han armado un diccionario de datos donde se detalla los nombres de las variables y una breve descripción de cada una de ellas para comprender los datos que estaremos analizando.

Nuestro Data Lead nos ha recomendado que analicemos detalladamente los datos, los pre procesemos debidamente y que diseñemos dos modelos predictivos, uno para el precio y otro para distinguir vehículos de gama alta y de gama baja, utilizando la mediana de los precios como punto de corte. Desean obtener los archivos con las predicciones en formato de texto plano.

Además del análisis detallado de la exploración de los datos, estas son las dos predicciones posibles que les interesaría analizar:

1. Implementar un modelo de clasificación con aprendizaje supervisado que permita clasificar el precio de los vehículos en baratos y caros usando la mediana de los precios como punto de corte, utilizando los datos que se han puesto a su disposición.

2. Implementar un modelo de regresión con aprendizaje supervisado que permita predecir el precio final de los vehículos, utilizando los datos que se han puesto a su disposición.

Este proyecto constará de tres fases: Preparación de los datos, análisis exploratorio de datos y modelamiento / evaluación.

Preparación de los datos

Durante el proceso de limpieza de datos se aplicaron acciones significativas en el conjunto de datos 'original_dataset.csv' para garantizar la calidad y la preparación óptima de los datos. Estos resultados, junto con las métricas de memoria obtenidas, establecen una base sólida para análisis y modelos predictivos futuros, destacando la eficiencia del conjunto de datos en términos de tamaño de memoria. El conjunto de datos limpiado, ahora con 28 variables, se exportó como 'cleaned_dataset.csv', listo para análisis y modelado subsiguientes. Los resultados claves incluyen:

Tamaño del Conjunto de Datos

Número de variables: 26

Número de observaciones: 205

Características Generales

Ausencia de filas duplicadas y valores nulos en todas las columnas.

Columnas discretas: 8

Columnas continuas: 8

Categóricas: 10

Optimización de la Memoria

Tamaño total en memoria: 42.57 KiB

Tamaño promedio de registro en memoria: 0.21 B

Estos valores indican una eficiencia notable en el uso de memoria, crucial para análisis de datos eficientes y rendimiento.

Normalización y Transformación de Datos

Normalización de nombres de columnas.

Mapeo de valores a las columnas 'door_number', 'cylinder_number',' symboling'.

Agregadas las variables 'calificacion_riesgo' (Extremadamente Riesgoso, Riesgoso, Muy Riesgoso, Neutral, Poco Seguro, Seguro) y 'clasificacion_precio' (1= valor por debajo de la mediana, 0= si está por encima), permitiéndonos clasificar los vehículos como baratos o caros.

Redondeo del valor en columna 'price' y convertida a enteros.

Exploración de Outliers

Identificación visual de outliers, los cuales se retuvieron debido al tamaño relativamente pequeño del conjunto de datos y su posible relevancia en futuros análisis.

Análisis exploratorio de datos

El análisis exploratorio de datos proporciona una visión profunda de la estructura y distribución de nuestro conjunto de datos, extraído del archivo 'cleaned_dataset.csv '. Aquí presentamos las conclusiones clave derivadas de esta fase:

Distribución Estadística de Variables Numéricas

Las estadísticas descriptivas revelan la diversidad en la magnitud de las variables. Por ejemplo, la distancia entre ejes (wheel base) oscila entre 86.6 y 120.9, mientras que el peso en vacío (curb_weight) varía de 1488 a 4066. Este rango diverso destaca la importancia del escalado de datos en etapas posteriores del análisis.

Distribución de Variables Categóricas

La exploración de variables categóricas, como el tipo de combustible, sistema de transmisión y tipo de carrocería, revela patrones y tendencias en nuestro conjunto de datos. Marcas como Toyota, Nissan, Mazda y Honda son dominantes, según la nube de palabras, y las características más comunes incluyen vehículos a gasolina, con transmisión estándar, tipo sedan y tracción delantera.

Análisis Univariado de 'Price'

La variable 'price' es central para nuestro análisis. Observamos una amplia gama de precios, con una mediana de 10,295 y un promedio de 13,276. La presencia de valores atípicos, evidenciada por el máximo de 45,400, sugiere una distribución sesgada hacia la derecha. Se confirma esto al observar la skewness positiva y la kurtosis elevada.

El análisis por categorías muestra diferencias significativas en los precios según el tipo de combustible, tracción, carrocería y otras características. Por ejemplo, los

convertibles tienden a ser más costosos, y la tracción trasera (RWD) se asocia con precios más altos.

Relaciones entre Variables

El pairplot y el mapa de calor revelan relaciones entre variables numéricas. La correlación positiva entre 'price', 'curb_weight' y 'car_width' sugiere que, a menudo, precios más altos se asocian con mayor peso y anchura del automóvil.

Marcas y Precios

El análisis de marcas destaca que Toyota es la marca más común, mientras que BMW, Jaguar, Buick y Porsche tienden a tener precios más altos. Esto proporciona información valiosa para futuros análisis centrados en la marca.

En resumen, el análisis exploratorio de datos ha arrojado luz sobre la distribución, relaciones y patrones en nuestro conjunto de datos, proporcionando una base sólida para análisis más avanzados y modelado subsiguiente. Estas observaciones guiarán nuestras decisiones en etapas posteriores del proyecto.

Resultados: Comparación de Modelos de Regresión

En ésta fase se trabajó con el archivo 'cleaned_dataset.csv ', se realizaron transformación de variables categóricas y numéricas. Se entrenan y evalúan diversos modelos de regresión con el objetivo de estimar el precio de vehículos.

Para la evaluación se usará la validación cruzada, mientras que la identificación de características más relevantes se llevará a cabo con SequentialFeatureSelection de la biblioteca scikit-learn. El objetivo es comparar el rendimiento de distintos modelos y discernir cuál es el más apto para nuestro conjunto de datos.

A continuación, se presenta una tabla comparativa con los resultados obtenidos de los modelos de regresión utilizados en el proyecto. Esta tabla resume las métricas clave que evalúan el rendimiento de cada modelo en la tarea de predicción de precios de automóviles.

Métricas Utilizadas:

- **Error Absoluto Medio (MAE):**
  - Representa la media de las diferencias absolutas entre los valores reales y los valores predichos. Cuanto menor sea el MAE, mejor será el modelo.
- **Error Cuadrático Medio (MSE):**
  - Mide la media de los errores al cuadrado entre los valores reales y los predichos. Es más sensible a errores grandes. Un MSE más bajo indica un mejor rendimiento.
- **Raíz del Error Cuadrático Medio (RMSE):**
  - Es simplemente la raíz cuadrada del MSE. Proporciona una medida en la misma escala que la variable objetivo original. Un RMSE más bajo indica un mejor ajuste del modelo.

- **Coeficiente de Determinación (R2):**
  - Representa la proporción de la variabilidad de la variable dependiente que es explicada por el modelo. R2 varía de 0 a 1; un R2 más cercano a 1 indica un mejor ajuste del modelo.

	Modelo	Error Absoluto Medio	Error Cuadrático Medio	Raíz del Error Cuadrático Medio	Coeficiente de Determinación
0	LinearRegression	0.273362	0.155949	0.387908	0.815504
1	DecisionTreeRegressor	0.280416	0.159089	0.409177	0.827882
2	RandomForestRegressor	0.227652	0.101054	0.308399	0.897319
3	GradientBoostingRegressor	0.198691	0.086304	0.277776	0.921388
4	KNeighborsRegressor	0.264156	0.158912	0.379826	0.835219
5	SVR	0.258942	0.144982	0.365205	0.851188

En conclusión, El modelo que ha demostrado ser el más efectivo para este conjunto de datos es la **Potenciación del Gradiente**. Este modelo tiene un coeficiente de determinación más alto, alcanzando **0.921** en el conjunto de entrenamiento y **0.932** en el conjunto de prueba. Además, se destaca por presentar los valores mínimos en las métricas de error (MAE, MSE, RMSE). El modelo de **Potenciación del Gradiente** se considera como la mejor elección para la predicción de precios de automóviles en este proyecto.

Resultados: Modelos de Clasificación

En ésta fase se trabajó con el archivo 'cleaned_dataset.csv ', se realizaron transformación de variables categóricas y numéricas. Se entrenan y evalúan diversos de aprendizaje supervisado para la clasificación, con el propósito de determinar si un automóvil es económico o costoso. La evaluación de los datos de entrenamiento se realizará mediante validación cruzada, y la selección de características relevantes se llevará a cabo con la ayuda de SequentialFeatureSelector de la biblioteca scikit-learn. Posteriormente, compararemos el rendimiento de los modelos para identificar cuál es el más adecuado para estos datos.

A continuación, se presenta una tabla comparativa donde se resume las métricas clave que evalúan el rendimiento de cada modelo en la tarea de clasificación de automóviles en categorías de costos. Las métricas incluidas son:

Exactitud (Accuracy): Proporción total de predicciones correctas.

Precisión (Precision): Proporción de instancias positivas identificadas correctamente.

Sensibilidad (Recall): Proporción de instancias positivas identificadas correctamente respecto al total de instancias positivas reales.

Especificidad (Specificity): Proporción de instancias negativas identificadas correctamente respecto al total de instancias negativas reales.

Área bajo la Curva (AUC): Área bajo la curva ROC, proporcionando una métrica global del rendimiento del modelo en la clasificación binaria.

Tiempo de Ejecución: El tiempo requerido para entrenar y evaluar cada modelo.

	Modelo	Exactitud	Precisión	Sensibilidad	AUC	Tiempo de Ejecución (s)
0	DecisionTree	0.926705	0.937043	0.914118	0.975242	20.967000
1	RandomForest	0.920455	0.894474	0.951063	0.982288	60.736000
2	LogisticRegression	0.896402	0.918437	0.875960	0.980126	19.000000
3	SVM	0.920833	0.878747	0.984615	0.982498	20.519000
4	KNeighbors	0.920833	0.904645	0.939703	0.974136	14.440000
5	NaiveBayes	0.926894	0.929060	0.928592	0.971853	11.855000
6	GradientBoosting	0.926515	0.904225	0.952452	0.989821	185.359000

Considerando que nuestro dataset está balanceado, la exactitud es un buen indicador del desempeño del modelo. Notamos que, en el conjunto de entrenamiento, el modelo de **Naive Bayes** lidera con la mayor exactitud. No obstante, al evaluar los conjuntos de pruebas, el **Árbol de Decisión** logra la exactitud más alta. Aunque en el conjunto de entrenamiento ocupa el segundo lugar en la métrica de exactitud, para este conjunto de datos específico, el modelo más adecuado resulta ser el **Árbol de Decisión**.

Notas adicionales

Este proyecto, parte integral del Módulo 6 del Bootcamp, se presenta como una valiosa práctica anticipada para la fase de Labs. Sin embargo, es crucial destacar su **carácter opcional, exento de evaluación o calificación cuantitativa**.

La colaboración en equipo durante este desafío fue fruto del esfuerzo conjunto de **Gretel Sánchez** y **Johanna Rangel**. Aplicamos habilidades avanzadas en machine learning para explorar, comprender y modelar las distintas características de los vehículos en el actual mercado, lo cual no solo resultó fascinante sino también consolidó nuestros conocimientos.

Aprovechamos este espacio para expresar nuestro agradecimiento a **Mario Suaza**, nuestro instructor durante esta travesía educativa. Su paciencia y dedicación fueron fundamentales para nuestro aprendizaje.

Quedamos a disposición en nuestros respectivos perfiles de GitHub para aquellos interesados en explorar más a fondo nuestro trabajo. ¡Gracias por acompañarnos en este viaje de aprendizaje!.

Johanna Rangel

Gretel Sanchez