Analítica para la toma de decisiones: Aprendizaje no supervisado

David Diaz Rodríguez

Nicolas Niño Valderrama

Valentina Jiménez Torres

Daniel Perea Mercado

Estudiantes de Ingeniería Industrial

Manuela Londoño Ocampo

Docente

Universidad de Antioquia

Facultad de Ingeniería

Ingeniería Industrial

2024

Introducción

El mercado de vehículos usados es complejo, con una amplia gama de factores que influyen en el precio y la demanda. Este proyecto se enfoca en aplicar técnicas de aprendizaje no supervisado para descubrir patrones ocultos y segmentar este mercado en grupos homogéneos, o clusters. El objetivo es identificar agrupaciones de vehículos usados con características similares, lo que puede proporcionar información valiosa para diferentes partes interesadas. El análisis se realizará utilizando un data set que contiene información sobre la venta de autos de segunda mano en el Reino Unido. Entre tanto, al aplicar un algoritmo de aprendizaje no supervisado, se espera identificar clusters de vehículos con perfiles similares y características importantes como; el modelo, el kilometraje, el cilindraje y el precio. Así mismo, se evaluarán diferentes algoritmos de aprendizaje no supervisado y se escogerá el que mejores métricas entregue en su desempeño esto, dependerá de las características del data set y de los objetivos específicos. Algunos de los algoritmos son; K-means, Hierarchical Clustering, Agrupación Jerárquica, DBSCAN o Gaussian Mixture Model estos, serán evaluados para determinar cuál proporciona la mejor representación de la estructura de los datos. La evaluación de la calidad de los clusters se realizará utilizando métricas apropiadas según el algoritmo para determinar la solidez y significado de los grupos encontrados finalmente, los resultados se visualizaran para una mejor interpretación.

Análisis exploratorio de datos y tratamiento de variables

La exploración de datos nos entrega resultados importantes en cuanto a cifras de la data set; esta base de datos cuenta con 50000 registros y 7 variables en donde se registraron la venta de autos de segunda mano en el Reino Unido, dejando en las variables datos de características y precio de los autos. Además, este data set no cuenta con datos nulos ni datos duplicados y contiene variables tipo objecto, flotantes y enteras.

Al analizar las distribuciones de las variables numéricas podemos encontrar que el cilindraje de los automóviles está entre 1000 y 5000 cc, el modelo de los vehículos está entre el año 2000 y el año 2022, las millas recorridas por estos vehículos oscilan entre 1001 y 199998 millas finalmente, los precios varían entre 1000 y 49999 Libras Esterlinas.

Así mismo, las cuatro variables analizadas (Engine Size (L), Year of Manufacture, Mileage, Price (£)) presentan distribuciones bastante uniformes, lo que indica una variedad de datos sin sesgos evidentes hacia valores específicos.

Las pruebas de normalidad de Shapiro, Kolmogorov y Anderson indican que estas variables no siguen una distribución normal. Aunque las variables no siguen una distribución normal, la falta de valores atípicos refuerza la calidad y la confiabilidad de los datos. Esto indica que los datos son consistentes y homogéneos dentro de los rangos esperados, lo cual es positivo para el análisis, ya que no se requieren medidas adicionales para manejar valores extremos.

Entre tanto, las variables categóricas entregan información no menos importante; el data set contiene registros de autos de segunda mano de 10 las marcas, Mercedes-Benz, Toyota, Audi, Nissan, Volkswagen, Chevrolet, BMW, Tesla, Honda, Ford. También, nos indica la referencia de los vehículos dentro de cada marca y el tipo de combustible de cada uno de ellos encontrando 3 tipos: eléctrico, gasolina y diesel.

Por otra parte, la matriz de correlación entre las variables numéricas indica que no hay correlaciones significativas entre las variables analizadas.

La variable Fuel type puede convertirse en una variable dummy que dé cuenta si el vehículo usa o no algún tipo de energía renovable, con el numero 1 para los que si (eléctrico e hibrido) y con el numero 0 para los que no (gasolina y diesel).

Se decidió trabajar sólo con las variables numéricas del dataframe ya que son las relevantes para el tipo de modelos que vamos a trabajar, así que se eliminan las variables 'Manufacturer', 'Model', 'Fuel Type'.

En el caso de Model y Manufacturer que son variables categóricas, podemos tratarlas con sus respectivas frecuencias para que puedan ser variables numéricas e incluirlas en el análisis de los clúster.

En el caso de la variable Year of Manufacture, para poder darle un mejor tratamiento como número, se ha convertido a Vehicle Age que hace referencia a la edad del vehículo, siendo esta la diferencia entre el año actual 2024 y el año de su manufactura.

Aplicación de Algoritmos

K-Means

En la aplicación de este algoritmo se tiene en cuenta sus características teniendo claro que es sensible a correlaciones altas entre variables, sesga sus resultados cuando encuentra outliers, sugiere dumizar variables categóricas y se optimiza con el método del codo y del coeficiente de silueta.

El análisis automatizado del punto de codo utilizando *KneeLocator* ha determinado que el número óptimo de clúster es 4.

Al aplicar este algoritmo y en la búsqueda de el numero óptimo de clúster, Gráficamente el método de la silueta nos sugiere un número de clúster de 8. Ahora, al aplicar el modelo K-Means con 4 clusters según el método *KneeLocator*  y evaluar métricas importantes nos entrega resultados como; Inertia: 35955.1810361024 este valor es relativamente alto, lo que sugiere que los puntos están dispersos y no están bien agrupados alrededor de sus centroides. Silhouette Score: 0.1908791012197668 este valor es muy bajo, indicando que la separación entre los clusters es pobre y que muchos puntos podrían estar mal asignados. Calinski Harabasz Score: 3324.8008283768954 este valor no es extremadamente bajo, tampoco es alto, lo que sugiere que la separación entre los clusters no es muy significativa. En conjunto, estas métricas indican que el clustering K-means realizado no ha sido exitoso en identificar clusters bien definidos y separados en el data set. Esto puede deberse a un número incorrecto K o que el algoritmo es inadecuado.

Entre tanto, se aplica este algoritmo con el número de 8 clusters según la métrica de silueta y nos entrega resultados así, Inertia: 23407.518529890833 esto significa que ha disminuido pero, sigue siendo relativamente alto lo que indica que los puntos están, en promedio, más cerca de los centroides de sus clusters. Silhouette Score: 0.2250938987768921 este valor indica que ha habido un ligero aumento en el Silhouette Score, pasando de 0.19 a 0.23. Si bien es una mejora, sigue siendo un valor bajo, indicando una separación entre clusters que no es óptima Calinski Harabasz Score: 3336.2170138383026 este valor ha aumentado ligeramente pero, nos indica que la separación entre clusters no es lo suficientemente significativa como para indicar un clustering exitoso.

También, se aplicó el algoritmo habiendo reducido a 4 componentes el data set, y que nos entrega resultados interesantes; Inertia: 8667.355489441105 este valor considerablemente menor que en los análisis previos indica cercanía de los centroides de sus clusters asignados, sugiriendo una mejor cohesión. Silhouette Score: 0.37041122994295606 este valor mayor que en los análisis anteriores aunque todavía no es excelente, esto indica que la mayoría de los puntos están relativamente bien asignados a sus clusters. Calinski harabasz score: 12435.202482312714 este valor es alto e indica una buena separación entre los clusters, Estas métricas indican que el clustering K-means ha mejorado sustancialmente. Finalmente, probamos el algoritmo con 6 clusters según sugerencia del score silueta y que nos dejó resultados; Inertia: 6173.831879695451 es un valor bajo, indicando que los puntos están, en promedio, cerca de los centroides de sus clusters. Esto sugiere una buena cohesión dentro de los clusters. Silhouette Score: 0.3508342502666487 este valor es moderadamente bueno, indicando una separación razonable entre los clusters. Calinski Harabasz Score: 11684.36380644751 Este valor es alto, indicando una buena separación entre los clusters. Las métricas sugieren un clustering de calidad aceptable.

Agrupación Jerárquica

Este algoritmo crea una jerarquía de clusters, representando la relación entre los clusters a diferentes niveles y se puede visualizar en un dendrograma. la agrupación jerárquica es sensible a la escala de las variables. La complejidad computacional de la agrupación es relativamente alta para conjuntos de datos grandes. Su principal ventaja es la construcción de la jerarquía, que proporciona información valiosa para comprender la estructura de los datos y elegir el número óptimo de clusters.

Al data set se le toma una muestra del 6% de los datos es así como, luego de comparar la distribución de una muestra del 6% de los datos originales con la distribución de estos mismos datos, concluimos que ambas son similares. Esto nos permite afirmar que la muestra del 6% será adecuada para entrenar el modelo, lo que nos ayudará a ahorrar capacidad computacional.

Es así como, se aplicó el algoritmo de agrupación Jerárquica implementando los métodos Single; muy sensible al ruido, el dendrograma se ve muy denso en las primeras fusiones, lo que sugiere que hay muchas fusiones pequeñas a distancias muy cortas y no genera agrupación de clústeres grandes.

Por el Método Complete; hay una distancia clara y significativa entre las últimas fusiones, lo que puede ser útil para decidir un corte en el dendrograma y los clústeres resultantes tienden a estar más equilibrados.

En el método Average; El dendrograma muestra un comportamiento intermedio entre los métodos "Single" y "Complete". Los clústeres no están tan fragmentados como en el método "Single", ni tan compactos como en "Complete". Finalmente, el método Ward; se observa que los clústeres se forman de manera muy jerárquica, con claras divisiones en niveles de distancia bien definidos. Hay pocas fusiones grandes al final, lo que sugiere que los clústeres son bien separados y pueden ser más fáciles de interpretar. Concluimos, que usaremos el método "Ward".

Entonces, empleamos el método "Ward" y cortamos el dendograma en el valor 1.0 del eje horizontal, obteniendo tres clústeres principales:

Un dendograma grande (verde), otro de tamaño intermedio (naranja), un tercero, en el medio, más pequeño (verde) esto, proporciona un equilibrio entre simplicidad y detalle, permitiendo analizar el comportamiento de los datos sin generar un número excesivo de grupos. Pero, realizaremos una comparación visual, entre el coeficiente de silueta y de Calinski con diferentes números de clústeres, para corroborar la hipótesis de que 3 clústeres son la mejor elección. Una vez aplicado, el número de clústeres que nos dan la mejor combinación entre el coeficiente de Silueta y de Calinski-Harabasz es 3. Posteriormente, se visualiza reduciendo a 3 dimensiones el PCA, la gráfica 3d nos muestra 3 clusters con agrupación alargada y definida. En cuanto al tamaño de los clúster, el clúster 1 es el mayor. El clúster 0 le sigue en tamaño, siendo un 4,07% más pequeño. Por último, el clúster 2 es el más pequeño de los tres, con un 18,67% menos de observaciones en comparación con el clúster 1. Además, al ordenar las 30 observaciones de forma descendente según su conteo, observamos que el clúster 2 aparece por primera vez en la fila 14 y ocupa las últimas 8 filas del conjunto de datos. La mayor diferencia de medias entre los clústeres se observa en la variable Mileage, donde el clúster 1 agrupa los vehículos con mayor millaje, mientras que los de menor millaje se encuentran en el clúster 2.

En cuanto a las demás variables, se puede notar que presentan medias similares en los tres clústeres. A pesar de esto, se pueden concluir algunas observaciones clave; los vehículos de mayor edad tienden a agruparse en mayor medida en el clúster 2, mientras que los de menor edad se localizan predominantemente en el clúster 0. El clúster 2 tiene la mayor presencia de vehículos que utilizan energía verde, en contraste con el clúster 0, que muestra la menor proporción de estos vehículos. Los vehículos más costosos se agrupan en el clúster 2, mientras que los más económicos se sitúan en el clúster 1. Por último, los motores de mayor tamaño se encuentran en el clúster 0, mientras que los de menor tamaño están en el clúster 1.

Gaussian Mixture Clustering

Este algoritmo supone una distribución normal en las características, es sensible a valores atípicos, agrupa los datos que pertenecen a cierta distribución, es menos sensible a las diferencias en escalas y usa las métricas AIC y BIC.

Se implementó este algoritmo separando las variables numéricas del data set original esto, debido a que las variables categóricas dumizadas no aportan mucho al modelo que aplicaremos debido a que su distribución va a variar entre dos valores y las distancias son las mismas, para hacer un ejercicio con cuidado solo se tuvieron en cuenta las variables numéricas en otro data set. Ahora, buscamos el numero óptimo de clusters, que arroja resultados en las métricas AIC y BIC para ambas métricas difieren un poco pero, estas muestran que el numero optimo está entre 4 y 5 clusters que podríamos analizar. Aplicamos Modelo GMM- entrenamiento del modelo con 4 componentes con una covarianza tipo full, estas métricas nos indican; Silhouette Score: 0.24384713 este valor es muy bajo e indica que la separación entre los clusters no es muy buena, y hay una considerable superposición entre ellos. Calinski harabasz score: 4996.3733 un valor moderado indica sugiriendo una separación entre los clusters que no es muy significativa. Ahora con 5 componentes, Modelo GMM- entrenamiento del modelo con 5 componentes con una covarianza tipo full obtenemos resultados; Silhouette Score: 0.2401499 este valor es bajo, indicando que la separación entre los clusters no es buena. Calinski harabasz score: 5023.585606 este valor es moderado, lo que sugiere una separación entre los clusters que no es muy pronunciada. ambos resultados apuntan a la necesidad de un ajuste significativo en el proceso de clustering para lograr una separación más clara entre los clusters. No hay una diferencia sustancial entre ambos conjuntos de resultados. Podemos concluir que entre ambos modelos con 4 y 5 componentes la diferencia no es significativa y ambos algoritmos no son viables.

Conclusiones

Se implementaron 3 algoritmos

Los resultados muestran que el kilometraje y la antigüedad son factores determinantes en la valoración de los vehículos, pero también destacan otros aspectos como el tamaño del motor y la calidad percibida. Mientras que los vehículos premium logran mantener su valor con el paso del tiempo (Clusters 1 y 2), los autos más accesibles (Clusters 0 y 3) brindan opciones adecuadas para aquellos que buscan equilibrio entre costo y funcionalidad. Esta segmentación proporciona una herramienta útil para adaptar las estrategias de mercado y optimizar la oferta de vehículos según las distintas necesidades y prioridades de los compradores, permitiendo atender tanto a quienes buscan lujo y durabilidad como a aquellos que valoran la asequibilidad.

Este estudio tiene como objetivo identificar las mejores características que influyen en la decisión de compra de un automóvil, específicamente en términos de la relación precio-estado del vehículo, que se refleja principalmente en el kilometraje y la antigüedad del mismo. Los resultados muestran que los vehículos con menor kilometraje y mayor antigüedad (Cluster 1) tienden a mantener un valor elevado, lo que los convierte en opciones premium para compradores que buscan calidad y durabilidad. Por otro lado, los autos más accesibles se encuentran en el Cluster 3, donde la antigüedad y el kilometraje moderado influyen en un precio más bajo, lo que podría atraer a aquellos con un presupuesto más limitado. La antigüedad y el kilometraje juegan un papel fundamental en la valoración de los vehículos, y la capacidad de un automóvil para mantener su precio a pesar de su edad (Cluster 1 y 2) sugiere que los compradores valoran la durabilidad y la calidad a largo plazo en su decisión de compra.