Academic Journal of Science and Technology
ISSN: 2771-3032 | Vol. 8, No. 3, 2023

Construction of Objective Segmentation Model of Automobile Users Based on Unsupervised Learning

Ziyi Chen, Decheng Liu and Tiantian Wan

China Auto Information Technology (Tianjin) Co.,Ltd, China

Abstract: With the rapid development of new energy vehicles and intelligent connected vehicles, today's automotive consumption has shifted from product centric to user centric. Grouping automobile users and conducting in-depth research on their characteristics and needs can provide certain support and guidance for manufacturers in planning and developing vehicle models, market validation, etc. However, current research on car user segmentation models is relatively lagging behind, and existing models use subjective factors such as values/automotive views/consumption views as modeling indicators. In the actual business development process of car companies, there are limitations of inconsistent user subjective cognition in various business processes, resulting in unclear vehicle development direction and inaccurate user positioning. In order to solve the above problems, this research innovatively constructs a set of automobile user segmentation model which is easy to understand and use by using unsupervised clustering algorithm based on objective indicators, and provides certain references for the development of new models and user positioning of automobile enterprises.

Keywords: Passenger car market, Unsupervised learning, Objective indicators, User segmentation models.

1. Introduction

At present, China's automobile market has changed from an incremental market to a stock market, the automobile industry intelligent, network, electric, sharing of the four changes have changed the original connotation of the automobile, the whole industry in the new energy, intelligent technology and other industries under the impact of the industry chain is disintegrating and restructuring, to the automobile companies have brought product definition, research and development and marketing and other challenges.[1]Under such an industry background and trend, car companies have adjusted their competitive strategies, user awareness, from product as king to user-centered, pay more attention to user experience, the importance of user research in the automotive industry has become increasingly prominent, and the accurate classification of users and the positioning of target users have become an important fulcrum of car companies.

At present, the population segmentation system based on subjective indicators such as values is used more in the process of model planning and development, but in the actual model development process, it is found that this system is easy to cause the problem of different cognition of users in each business link, which brings the risk of development deviation.At the same time, due to the lack of a relatively unified and objective subdivision system, opportunities and trend research is also easy to fragment, lack of systematic comparability research, it is not convenient to track the population change trend.In addition, there is no unified segmentation system from the user, and it is easy to over-rely on product attributes in the process of planning and innovation, and it is difficult to find a breakthrough point. Therefore, building a set of population segmentation model that is easy to understand and use and suitable for the needs of automobile enterprises can improve the accuracy, system and comprehensiveness of the relevant work of automobile enterprises planning.

2. Research Progress of Automobile User Segmentation Model

In recent years, in the domestic new force car brand, Weilai, ideal, Xiaopeng and other enterprises took the lead in introducing the user-centered thinking, breaking through the traditional car enterprises to the standard oriented car model, and becoming the user-oriented car pioneer, at the same time, the major car companies in the industry have turned to the user as the center.Since American scholar Wendell-R-Smith put forward the concept of user segmentation in 1956, domestic and foreign university scholars and market research institutions have never stopped the research on the theory and model of user segmentation.[2,3]Throughout the past research on user segmentation model, most of them focus on e-commerce, online games, finance, retail and other industries with high user consumption frequency, while the automobile industry is different from other consumer categories due to the influence of multiple factors such as price, industry, policy and users, so the research on automobile user segmentation model is relatively lagging behind. At present, there are many researches on automobile user segmentation model by market research institutions, and the Profiler population segmentation model of Roland Berger Strategy Consulting Company is highly recognized in the industry.The model uses values and consumption views to classify users.Based on the psychological needs of the crowd, through statistical cluster analysis, the typical consumer groups in the market are obtained, and each group of people share a similar set of core values.Finally, the car crowd is divided into: decent and smart type, warm and harmonious type, personality enterprising type, traditional rational type, conservative and peaceful type, trendy self-type six categories of people.Another group classification is the top ten group classification of the National Information Center, which adopts the three dimensions of social class, values and car view as the key variables. Among them, social class is divided into five levels according to economic resources, power resources and

cultural resources.At the same time, the car concept is divided into four levels of needs (basic needs, experience needs, personality, identity).On this basis, 25 subgroups are obtained by crossing the one-dimensional classification of values and social classes, and ten subgroups are finally obtained by clustering according to the sub-groups of car view.

Although the population segmentation model mentioned above has been recognized and applied by the market, it still has major limitations, and most subjective factors such as values are used as modeling indicators. This study is committed to building a set of population segmentation model system based on objective indicators, so that personnel of all business links can reach a unified cognition, and help improve the accuracy of vehicle planning and development and user positioning.

Table 1. Data of sample database

tag	Tag value	Take up a proportion of
Fuel type	Fuel truck (including HEV)	55.0%
	New energy -EV	30.0%
	New energy -PHEV	10.0%
	New Energy -REEV	5.0%
Body form	sedan	48.5%
	SUV	45.0%
	MPV	6.5%
region	northeast	5.0%
	North China	16.0%
	East China	22.0%
	South China	21.0%
	Central China	18.0%
	northwest	4.0%
	southwest	14.0%
Line level	1	17.0%
	2	43.0%
	3	18.0%
	4	14.0%
	5	8.0%
Age range	18-25 years old	9.0%
	26-30 years old	22.0%
	31-35 years old	34.0%
	36-40 years old	22.0%
	41-45 years old	10.0%
	46-50 years old	2.0%
	Over 51 years old	1.0%
Annual household income	Less than 50,000	1.0%
	50,000 to 100,000	5.0%
	100,000 to 150,000	14.0%
	150,000 to 200,000	23.0%
	200,000 to 300,000	31.0%
	300,000 to 500,000	21.0%
	Half a million to a million	5.0%
First increase and replacement purchase	First purchase	65.0%
	Buy more	16.5%
	trade-in	18.5%
Marital status	Be married	82.9%
	spinster	17.1%
sex	male	80.5%
	female	19.5%
Educational background	High school and below	4.7%
	College degree or equivalent	38.4%
	Undergraduate course	54.9%
	Master degree or above	2.1%
Living condition	native	63.7%
	Move to live locally for school/work	19.6%
	outlander	6.6%

3. Route Method for Establishing Passenger Car Market Population Classification Model Based on Objective Indicators

3.1. Overview

The consumer database accumulated by the brand consulting Department of China Automotive Information Technology in the past two years is used as the data source for this modeling. The first step is to clean the database samples and adjust the modeling sample structure based on the market situation to reflect the real situation of the market. The second step: determine the subdivision index system and the priority of each modeling index; The third step: modeling method determination. Mainly through unsupervised machine learning of a variety of clustering algorithms (K-Means, DBSCAN, etc.) and other algorithms to try, and timely adjustment of the process, the establishment of passenger car market population segmentation model.

3.2. Database Establishment and Cleaning

The market research project data (22,899 items) accumulated by the brand consulting Department of China Automotive Information Technology in the past two years is used as the database of this modeling, which includes 16 objective indicators such as line level, age, whether you have children, education background, family annual income,

region, gender, marriage, age range, and generation. According to the established cleaning standards, the database is cleaned, including the processing of missing sample outliers, the unification of label values, and the supplement of derivative indicators. In order to make the sample structure more truly reflect the market situation, the sample structure is adjusted, and the priority of sample structure adjustment and adjustment principle are confirmed according to the passenger car market situation and business needs. The final sample base structure is shown in the table 1:

3.3. Confirming Indicators and Priorities

After the database is built, the priorities of the modeling indicators are determined according to the data, actual business and past research experience: the first priority is the urban regional line level, education background, marital status, family annual income, gender, whether there are children, and the intergenerational age-age range; the second priority is the nature of the unit, the first increase and purchase.

3.4. Modeling method

Supervised algorithm and unsupervised algorithm are two common types of algorithms in machine learning, and they have different characteristics in data processing and problem solving.[4,5]

A supervised algorithm is an algorithm that uses labeled training data to train a model and make predictions. In supervised learning, we have input data and corresponding outputs (or labels), and the goal of the algorithm is to learn a function (model) from the input data that maps the input to the correct output data. Common supervised learning algorithms include decision tree, support vector machine, logistic regression and neural network.

Unsupervised algorithm is an algorithm that uses unlabeled training data for pattern recognition and data clustering. In unsupervised learning, we only have input data, no labels or category information, and the goal of the algorithm is to discover intrinsic structures, patterns, or regularities in the data. It is mainly used in clustering, reduction and anomaly detection. Clustering is the task of dividing data into similar groups (clusters); Dimensionality reduction is the reduction of data dimensions to remove redundant features. Anomaly detection is the identification of outliers that differ from other samples. Application scenarios include data mining, market analysis, and recommendation systems.

There are obvious differences between supervised algorithm and unsupervised algorithm in training process and application scenario. Supervised algorithms need labeled training data to make predictions by learning the relationship between inputs and corresponding outputs. Unsupervised algorithms, on the other hand, model and analyze data based on its own statistical characteristics, and discover patterns and structures in it.

Based on the above analysis and business needs, unsupervised algorithm is selected as the modeling method of this model.

4. Model Establishment and Evaluation

4.1. Modeling algorithm

4.1.1. Modeling algorithm selection

According to the objective database situation and modeling purpose, for data sets containing different types of variables, this modeling uses clustering algorithms based on distance or similarity, such as K-Means, K-Medoids, DBSCAN, etc.[6-8]

K-means algorithm is a distance-based iterative clustering algorithm, whose goal is to divide the data set into K non-overlapping clusters, so that the samples within each cluster are as close to each other as possible, and the samples between different clusters are as far away as possible. K-means uses distance as the similarity index to find K classes in a given data set, and the cluster center of each class is obtained according to the mean value of all values in the class, and the center of each class is described by the cluster center.

The L-Medoids algorithm is an improved version of the K-Means algorithm, mainly for processing data in non-Euclidean Spaces. Unlike K-Means, the cluster center of the K-Medoids algorithm is taken from the samples that actually exist in the data set, rather than simply calculating the mean of the samples within the cluster. Its execution steps are as follows:Select K initial center points at random → Calculate the distance between each sample and each center point and divide it into the cluster represented by the nearest center point → For each cluster, select other samples in the cluster as the new center point so as to minimize the total distance between all samples in the cluster and the new center point → Repeat the above two steps.Until the position of the center point no longer changes or reaches a predetermined number of iterations.

DBSCAN (Density-Based Spatial Clustering of Applications with Noise) is a density-based clustering algorithm that can efficiently find clusters with arbitrary shapes. Compared with K-Means and K-Medoids, DBSCAN does not need to specify the number of clusters in advance. The main idea is to determine the boundary of the cluster based on the density between the samples. Perform the following steps to perform DBSCAN: Select an unvisited sample at random → If the neighborhood density of the sample is greater than the set threshold, divide the sample and its density reachable samples into a new cluster → Continue to perform density reachable judgment on each sample in the new cluster until no new density reachable samples can be found → determine noise points or smaller clusters in all unvisited samples.

Considering the situation of sample database, the K-Means algorithm is mainly selected for this modeling.

4.1.2. Evaluation index of modeling algorithm

A common algorithm evaluation index is the Silhouette Coefficient, which takes into account in-cluster tightness and inter-cluster separation. The closer the value is to 1, the better the clustering effect.

The contour coefficient takes into account the tightness of samples within clusters and the separation of samples between clusters. For each sample, calculate its average distance (a) from other samples in the same cluster and its average =-098 distance (b) from samples in the nearest cluster, and then calculate the profile coefficient as (b-a)/max(a,b). The value of the contour coefficient ranges from -1 to 1, where closer to 1 means that the sample is more clustered in the correct cluster, and closer to -1 means that the sample is more likely to be assigned to the wrong cluster.

Another common evaluation criterion for modeling algorithms is the Elbow method, which helps us find a suitable K value by plotting the sum of squares of error (SSE) of clustering results with different K values over K. Up to a certain point, subsequent increases in the value of K do not significantly reduce the sum of squares of error. An inflection

point is formed where the value of K is such that a balance is found between the model complexity (i.e., the value of K) and the clustering effect (SSE), which can be considered the optimal number of clusters K.

4.2. Data Preparation

Commonly used clustering algorithms such as K-Means algorithm, DBSCAN algorithm, etc. are suitable for numerical variables because distance measures are needed to determine the similarity between data points. This means that the input data should be composed of real numbers or numerical features, i.e. the algorithm cannot be directly applied to the category variable and needs to be encoded to convert the category variable into a numerical variable.

4.2.1. Nominal variables

Nominal variables are encoded by One-Hot coding, which is a data coding technique used to convert discrete class data into numerical vectors. It represents categorical variables as binary vectors, each representing a category with a value of 1 in the corresponding column for that category and 0 in all other columns.

The method of unique thermal Encoding first determines the different values in the class variable, and then assigns a unique integer number to each value, either using Label Encoding or a custom numbering method. Next, each integer number is converted into a corresponding binary vector, usually into a matrix of 01, where each row represents a non-repeating class and each column represents a data sample. If a data sample belongs to a class, the column of that class has a value of 1, and the other columns have a value of 0.

The nominal variables in this database include: region, gender, marriage, and whether you have children.

Table 2. Database nominal variables

Variable name	Number of variable value categories	Coded result
region	7	Region _ East China, Region _ North China, Region _ Southwest, Region _ Northeast, Region _ Central China, Region _ South China, Region _ Northwest
sex	2	Gender _ Male, Gender _ Female
matrimony	2	Marriage _ Unmarried, marriage _ Married
Do you have children?	2	Do you have children _ No, do you have children _ Yes

4.2.2. Ordered variables

Ordered variables are encoded with labels. Label Encoding is a method of converting categorical variables in a data set into corresponding numeric labels. In label coding, we assign a unique integer number to each different class that has a size meaning, such as from 1 to n, where n is the number of different values in the class variable. With label encoding, ordered variables can be converted to numeric variables, and the order relationship between categories is preserved.

The ordered variables of this database include: generation, age range, educational background, and annual household income.

Table 3. Database ordered variables

Variable name	Number of variable value categories	Coded result
Between generations	8	0: 'after the' 50 s, 1: 'after 55 s, 2:' after the '60 s, 3: after the' 65 s', 4: 'after 70 s', 5:' after 75 s', 6: 'after 80 s', 7:' after 85 s', 8: 'after the' 90 s, 9: 'after the' 95 s, 10: '00 s after'
Age range	7	0:18 to 25, 1: '26 to 30 years old, 2:' 31 and 35, 3: '36 to 40 years old, 4:' 41-45 years old, 5: '46-50 years old, 6:' 51 years of age or older
Annual household income	7	0: '1 million', 1:50-1 million, 2:30-500000, 3:20-300000, 4:15-200000, 5:10-150000, 6:5-100000, 7: '50000'
Educational background	4	0:' High school and below ',1:' College and equivalent ',2:' Bachelor's Degree ',3:' Master's degree and above '

4.3. Modeling parameter optimization

Through the K-Means algorithm, according to the contour coefficient method, the number of possible clusters is [2,3,4,5,6,7,8].

When cluster number is 2, the mean profile coefficient is 0.4861621464423962.

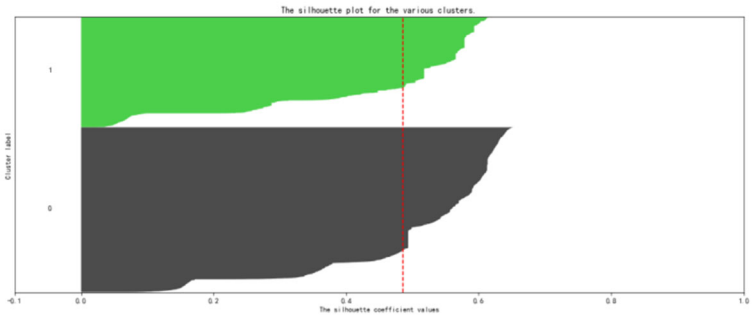

Figure 1. K-means algorithm K=2 contour coefficient

When cluster number is 3, the average contour coefficient is 0.4114124688456909.

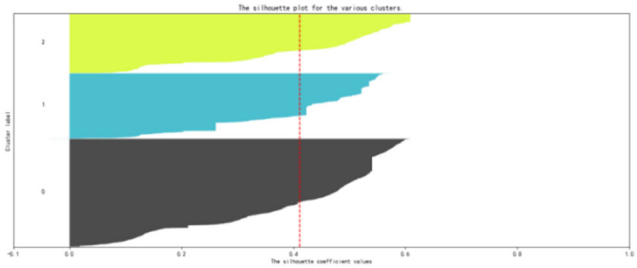

Figure 2. K-means algorithm K=3 contour coefficient

When cluster number is 4, the mean profile coefficient is 0.3936975719552483.

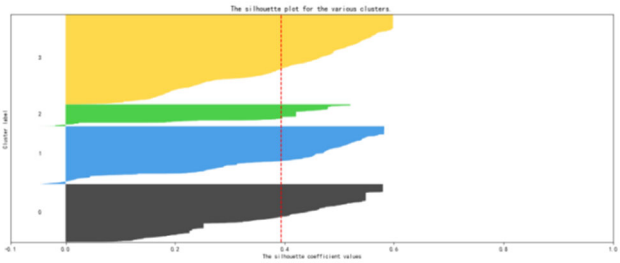

Figure 3. K-means algorithm K=4 contour coefficient

When cluster number is 5, the mean profile coefficient is 0.32937614485010025.

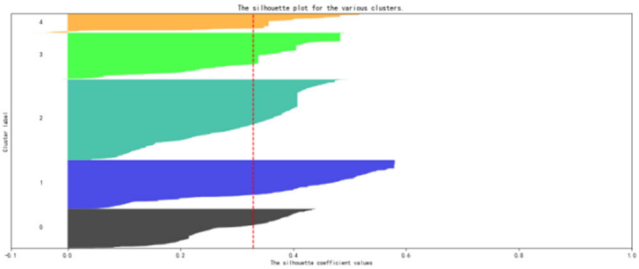

Figure 4. K-means algorithm K=5 contour coefficient

When the cluster number is 6, the mean profile coefficient is 0.29903198575008144.

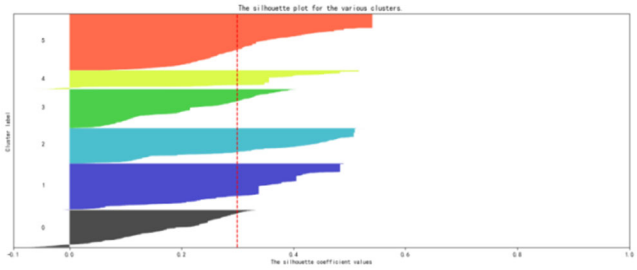

Figure 5. K-means algorithm K=6 contour coefficient

When cluster number is 7, the mean profile coefficient is 0.2954374503507193.

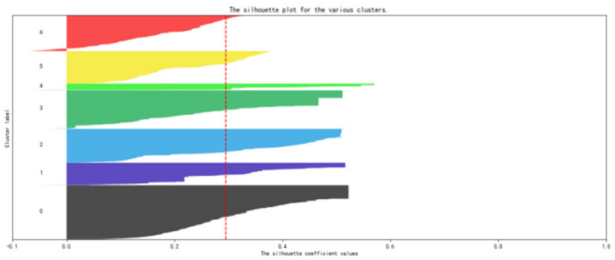

Figure 6. K-means algorithm K=7 contour coefficient

When cluster number is 8, the mean profile coefficient is 0.29937152132335443.

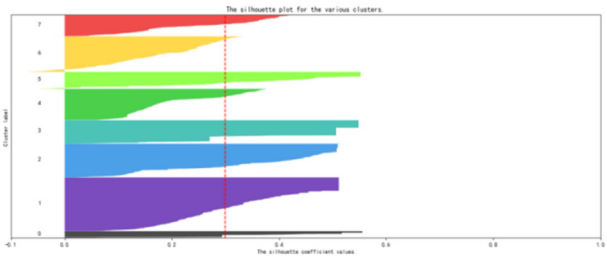

Figure 7. K-means algorithm K=8 contour coefficient

According to the bending diagram method, the inflection point and the decline rate of SSE(K) in the figure are mainly considered as the number of clusters [2,3,4].

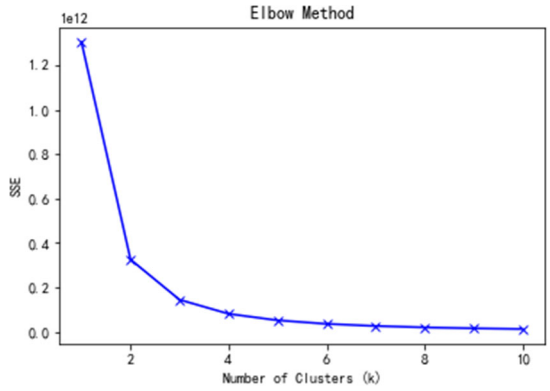

Figure 8. Evaluation results of K-Means bending graph method

4.4. Modeling Results

The modeling results are mainly divided into 2 clusters, 3 clusters and 4 clusters.

4.4.1. Number of clusters is 2

When K=2 and the number of clusters is 2, the original scatter plot is as follows:

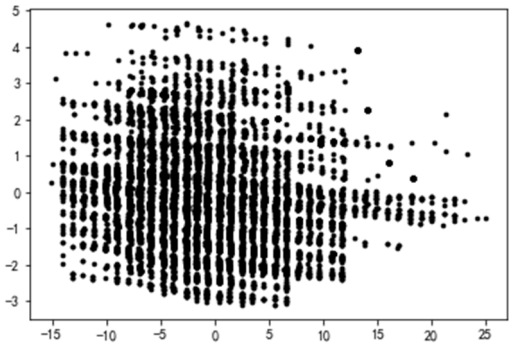

Figure 9. K=2 Original scatter plot

The visual scatter diagram is as follows:

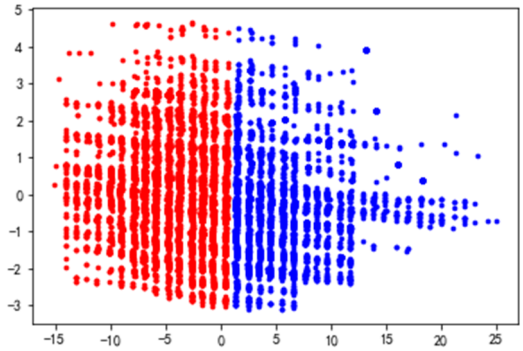

Figure 10. K=2 Visual scatter plot

The portrait features of the two clusters are as follows (using the mean rounded to represent the cluster) :

Table 4. Portrait features at K=2

index	Cluster 0	Cluster 1
Source data quantity	14999	10011
region	East China	South China
Line level	2	3
Educational background	Undergraduat e course	College degree or equivalent
matrimony	Be married	Be married
Annual household income	200,000 to 300,000	200,000 to 300,000
sex	male	male
Do you have children?	is	is
age	29, 59	39, 58
Age range	26-30 years old	36-40 years old
Between generations	After 90s	After 80s

4.4.2. Number of Clusters is 3

When K=3 and the number of clusters is 3, the original scatter plot is as follows:

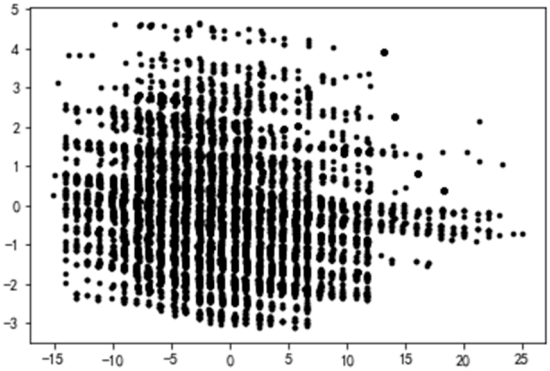

Figure 11. K=3 Original scatter plot

The visual scatter diagram is as follows:

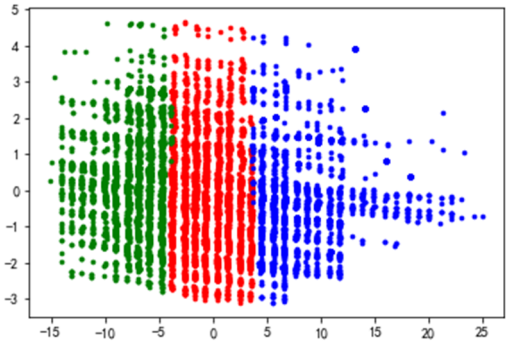

Figure 12. K=3 Visual scatter plot

The portrait features of the three clusters are as follows (using the mean rounded to represent the cluster) :

Table 5. Portrait features at K=3

index	Cluster 0	Cluster 1	Cluster 2
Source data quantity	11587	7026	6397
region	East China	South China	North China
Line level	2	3	2
Educational background	Undergrad uate course	College degree or equivalent	Undergraduat e course
matrimony	Be married	Be married	spinster
Annual household income	200,000 to 300,000	150,000 to 200,000	150,000 to 200,000
sex	male	male	male
Do you have children?	is	is	no
age	32 years old	41.2 years old	26 years old
Age range	31-35 years old	41-45 years old	26-30 years old
Between generations	After 90s	After 80s	95s later

4.4.3. Cluster Number is 4

When K=4 and the number of clusters is 4, the original scatter plot is as follows:

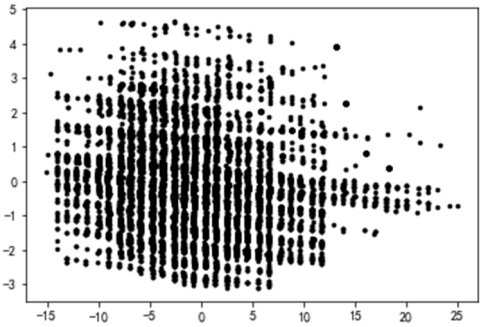

Figure 13. K=4 Original scatter plot

The visual scatter diagram is as follows:

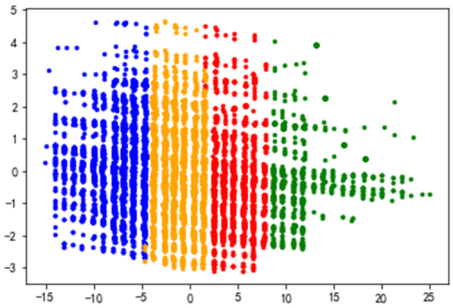

Figure 14. K=4 Visual scatter plot

The portrait features of the four clusters are as follows (using the mean rounded to represent the cluster) :

Table 6. Portrait features at K=4

index	Cluster 0	Cluster 1	Cluster 2	Cluster 3
Source data quantity	6369	6360	2401	9880
region	South China	North China	North China	East China
Line level	3	2	3	2
Educational background	College degree or equivalent	Undergraduate course	Undergraduate course	Undergraduate course
matrimony	Be married	spinster	Be married	Be married
Annual household income	150,000 to 200,000	150,000 to 200,000	200,000 to 300,000	200,000 to 300,000
sex	male	male	male	male
Do you have children?	is	no	Be married	is
age	38 years old	26 years old	45.2 years old	32 years old
Age range	36-40 years old	26-30 years old	41-45 years old	31-35 years old
Between generations	After 85s	95s later	After 75s	After 90s

5. Conclusion and Prospect

Through unsupervised machine learning of multiple clustering algorithms (K-Means, DBSCAN, etc.) and related algorithm effect evaluation attempts, based on 16 objective indicators such as line level, age, children, education, family annual income, region, gender, marriage, age range, generation, etc., database construction, sample cleaning and database sample structure adjustment.The classification model of the passenger car market population is obtained, and the innovative breakthrough achieved on the basis of the previous subjective and objective labels has formed a preliminary exploration of the overall passenger car market population classification based on pure objective indicators.

On the basis of this preliminary modeling attempt, the algorithm needs to be optimized in the future, and a more perfect modeling method can be preliminarily explored based on this model, such as hierarchical clustering for each layer. At the same time, more algorithms and data processing methods can be tried to form a complete set of objective population classification model for passenger car market users, so as to unify the cognition of users in all business links.Thus avoid the risk of development deviation, and realize the continuous tracking of population change trends.

References

[1] Wang Zhiyuan. What opportunities and challenges will the "New Four Modernizations of Automobile" bring [N]. China Youth Daily,2023-07-27(011).

[2] Chang H. C., Tsai H. P.. Group RFM analysis as a novel framework to discover better customer consumption behavior[J]. Expert Systems with Applications, 2011, 38(12): 14499-14513.

[3] Han S. H., Lu S. X., Leung S.. Segmentation of telecom customers based on customer value by decision tree model[J]. Expert Systems with Application, 2012, 39(4): 3964-3973.

[4] Chen Zhengyu. Research on Vehicle Rerecognition Algorithm Based on unsupervised Learning [D]. Dalian Maritime University, 2020.

[5] Zhou Feng, Yu Yi, Lin Xin et al.Device Fault diagnosis Technology and Algorithm fusion based on supervised and unsupervised learning algorithms [J].Industrial Technology Innovation, 2002,9(04):30-38.

[6] Huang Yaping, Li Yuanjiang.Research on E-commerce User Segmentation based on K-means Algorithm [J]. Electronic Design Engineering, 2017(2): 63-66.

[7] Mehar A. M., Matawie K., Maeder A.. Determining an optimal value of K in K-means clustering[C]. IEEE International Conference on Bioinformatics & Biomedicine. IEEE, 2014.

[8] Xu F. Research on cigarette retail customer classification based on DBSCAN clustering algorithm [J].China Market, 2023 (23): 121-124.