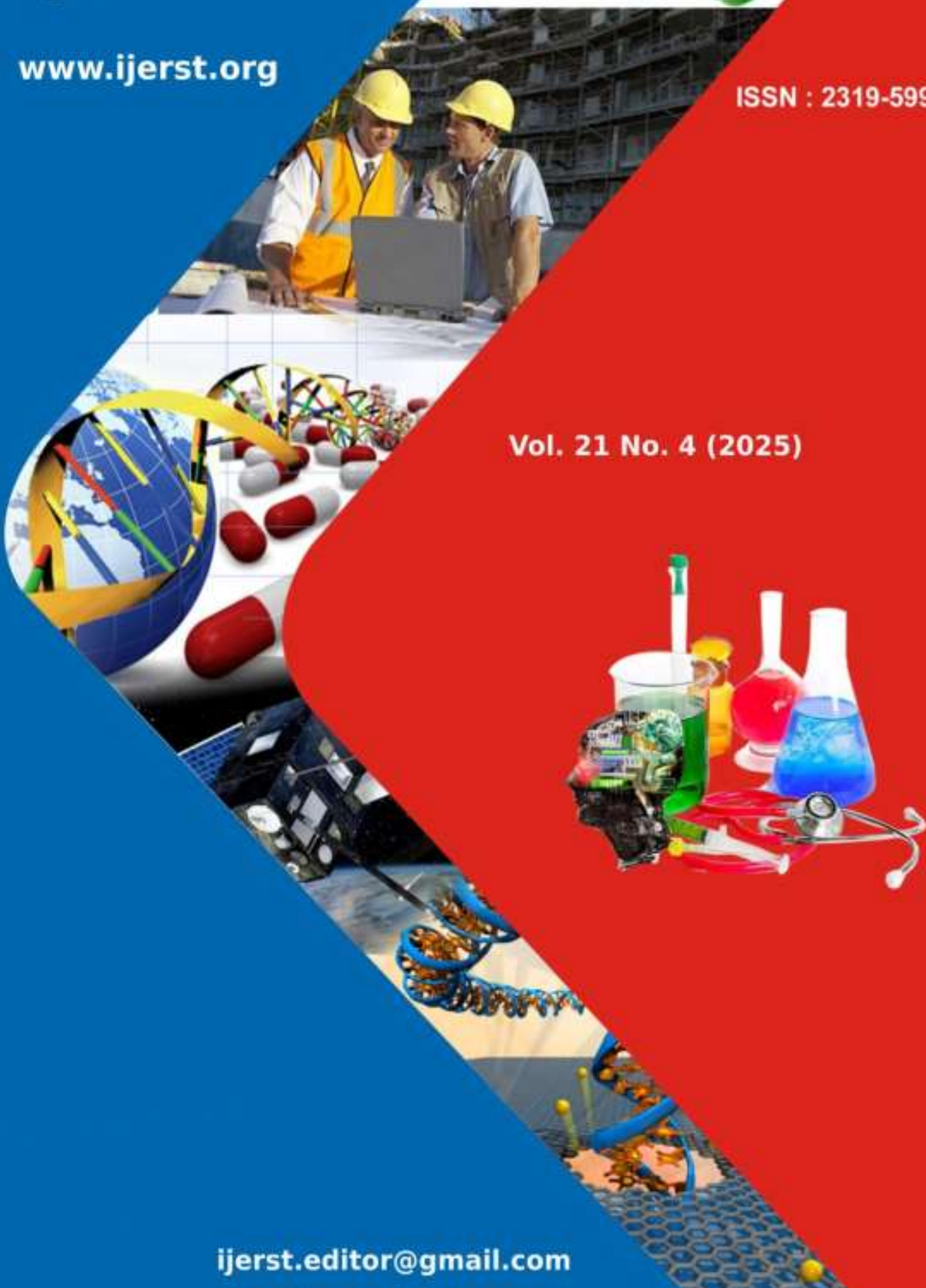
International Journal of Engineering Research and Science & Technology

IJERST

www.ijerst.org

ISSN : 2319-5991

Vol. 21 No. 4 (2025)



ijerst.editor@gmail.com

editor@ijerst.com

ISSN 2319-5991 www.ijerst.com
Vol. 21, No. 4, 2025

Peer Reviewed, Referred & Indexed Journal

Research Paper

ML-DRIVEN APPROACH FOR OPTIMIZING CUSTOMER PREDICTIONS IN THE AUTOMOBILE INDUSTRY

Komati Sathish,
Associate Professor,
Department of Computer Science and Engineering (Data Science),
TKR College of Engineering and Technology, Survey No -8/A Medbowli, Meerpet, 500097, Telangana
Email: komatisathish459@gmail.com

Abstract

In the increasingly competitive automobile industry, identifying and targeting the correct customer segments is critical for optimizing marketing strategies and resource allocation. This project presents a comprehensive Machine Learning and Regression framework designed to categorize customers into four distinct strategic groups (A, B, C, and D) based on demographic and socio-economic variables. Utilizing a dataset containing features such as Age, Profession, Graduation status, and Spending Score, the study employs a systematic data science pipeline encompassing rigorous data cleaning, exploratory data analysis (EDA), and advanced feature engineering. A key innovation in this study is the development of the Work_Experience_to_Age_Ratio, a derived feature that provides deeper insights into a customer's professional stability and purchasing power relative to their life stage. The methodology compares the performance of Logistic Regression (LR) and Random Forest Classification (RFC) models. Initial findings indicate that while LR provides a robust baseline for linear relationships, the RFC model excels at capturing the non-linear complexities inherent in consumer behavior. The results demonstrate that factors like Profession (notably Healthcare and Artistic sectors) and Spending Score are the primary drivers of segmentation. By automating the classification process, this approach enables automotive companies to move away from generic mass marketing toward hyper-personalized engagement. This data-driven model provides a scalable solution for predicting customer groups, ultimately leading to improved conversion rates, enhanced customer satisfaction, and a more efficient supply chain tailored to specific segment demands.

Keywords: Automobile industry, customer segmentation, machine learning, logistic regression, random forest classification.

Received: 04-11-2025 Accepted: 13-12-2025 Published: 20-12-2025

1. Introduction

Machine learning has a rich history dating back to the mid-20th century. The concept of machine learning can be traced back to the development of artificial intelligence (AI) and early efforts to create computer programs that could learn from data. In the 1950s and 1960s, researchers began exploring the idea of using algorithms to enable computers to improve their performance on tasks through learning from data. Over the decades, machine learning has evolved significantly, with contributions from various fields, including statistics, computer science, and neuroscience. The development of more powerful hardware and the availability of large datasets have accelerated progress in machine learning. The automobile industry is a highly competitive sector that constantly seeks ways to improve marketing strategies and customer relations. Predicting the right group of customers for automobile industries is crucial for targeted marketing, sales, and product development. Traditional methods of customer segmentation often involve manual processes and are limited in their ability to handle complex data. Machine learning and regression approaches offer a data-driven and automated solution to this problem.

2. Literature Survey

Customer segmentation is a critical marketing strategy that involves dividing customers into smaller groups based on similar characteristics such as demographics, psychographics, behaviors, and needs [1]. While previous studies on customer segmentation provide valuable insights into consumer behavior, they also have some limitations that will be addressed in this research, such as the use of a broader focus on psychographic and behavioral variables, and a more in-depth analysis of the implications of customer segments for marketing strategy. There are also very limited studies on customer segmentation in Indonesia [2, 3]. Admiration-based segmentation conducted in this study focuses on emotional decision-making factors. This approach recognizes that consumers often make decisions based on how they feel about a product, rather than on purely objective criteria such as price or fuel economy. By identifying the emotional triggers that inspire admiration, automotive companies can design cars that resonate with consumers on a deeper level and create long-term brand loyalty. Millennials and Gen Z in Indonesia prioritize practicality, technology, and sustainability in their car choices. They value experiences over ownership, customization, and online research. These preferences are shaping the automotive market, leading to a shift towards eco-friendly, tech-savvy, and customizable cars [4]. Millennials and Gen Z are two generations that have grown up in the digital age, with access to technology and information that previous generations did not have. This has led to changes in their behaviors, attitudes, and preferences, including in the automotive market [5]. While previous generations were more loyal to specific car brands, millennials and Gen Z are less so. They are more willing to switch brands based on their changing needs and preferences [6]. To understand how a brand is admired, first marketers have to understand customers characteristics and preferences. Previous literature reviews show that automotive admiration can vary widely across different countries in Asia due to differences in culture, consumer preferences, and economic development. For example, in developed countries like South Korea and Japan, domestic automakers like Hyundai and Toyota tend to dominate the market due to strong national pride and a desire to support local industries [7]. In contrast, in countries like China and India, there is a growing demand for Western luxury car brands as a status symbol among the growing middle class [8]. Additionally, factors such as government regulations, infrastructure, and consumer education can also affect automotive admiration [9]. For example, in countries with well-developed public transportation systems like Singapore and Hong Kong, car ownership may be seen as less necessary and less desirable than in countries with less developed infrastructure [10].

The ASEAN region, which includes 10 countries in Southeast Asia, is a large and diverse market for cars. The region is home to over 650 million people, and its rapidly growing middle class has become an important driver of car sales in recent years. Japanese car brands have traditionally dominated the ASEAN market, with Toyota, Honda, and Nissan being among the most popular brands [11, 12]. However, in recent years, Korean brands such as Hyundai have also gained significant market share [13]. It's worth noting that customer preferences can vary significantly by country. For example, pickup trucks are very popular in Thailand. This is driven by a combination of factors, including their versatility, affordability, cultural significance, and government policies that support the domestic automotive industry [14]. In Malaysia, domestic car brands such as Proton and Perodua are the most popular [15]. Similar to Malaysia, domestic car brands like VinFast and Truong Hai (Thaco) have also gained popularity in Vietnam in recent years. Meanwhile, the Philippines have limited domestic production of vehicles, with most cars sold in the Philippines being imported from other countries. On the other hand, Indonesia is the largest automotive market in Southeast Asia, with a population of over 270 million people. It has high demand for affordable vehicles. Most of the popular car brands in Indonesia are from Japan, with Toyota, Daihatsu, Honda, Suzuki, and Mitsubishi being the most popular. However, there are also car brands from other countries such as South Korea, including Kia and Hyundai. Additionally, some luxury car brands from Europe and the United States are also available in Indonesia, but they are generally having less population due to their higher prices. Chinese car brands have also expanded their business in Indonesia with the presence of Wuling, Cherry and DFSK. Demographic data is an important statistic that describes a population and is a popular factor used by marketers. Consumers' needs, wants, and brand preferences are often associated with demographic variables. By examining the demographic segments of consumers who admire a brand, marketers can see changes in demand and determine the target market for the products offered. The demographic characteristics used in this study are age, gender, income, education, occupation, and number of household members. The types of brands that people are looking for will change as they age and develop in their life cycle. A motive is a need that is sufficiently pressing to induce a person to act [16]. Motivation is the process or act of providing a motive that causes someone to take an action. According to several theories, motivation is rooted in a basic need to maximize pleasure, minimize physical pain, or satisfy one's desire for a goal, state, or object. When a need arises, there will be pressure that encourages consumers to try to fulfil that need. This study seeks to understand the motivation underlying a customer's admiration for a certain car brand.

3. Proposed Methodology

This research develops a data science project that focuses on predictive modeling and analysis using a customer dataset for automobiles. It follows a standard data science workflow, starting from data loading and preparation, through exploration and visualization, to machine learning model building and evaluation.

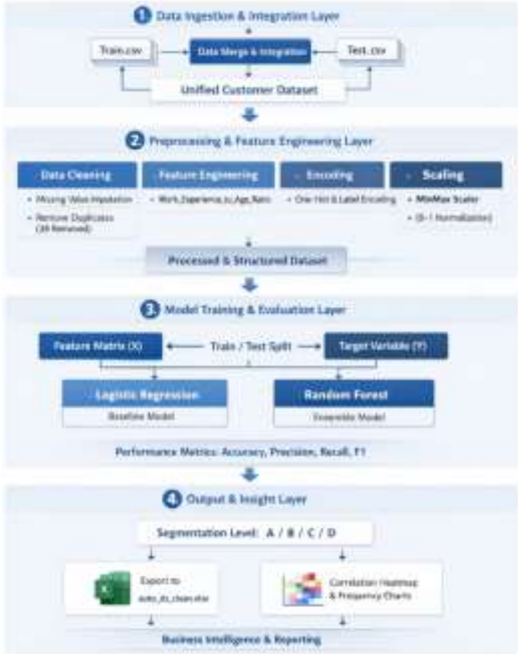


Fig. 1: Proposed system architecture of customer segmentation.

Here's an overview of the key components and steps involved in this project:

1. Data Loading: It begins by loading two CSV files, "Train.csv" and "Test.csv," into Pandas DataFrames. These files likely contain data related to the project's target, such as customer segmentation or classification.

2. Data Preparation and Exploration: It is an essential step in any data science project. In this project, several tasks are performed to understand the data better:

- Setting display options to show more rows and columns, making it easier to view the data.
- Displaying a random sample of 10 rows to get a sense of the dataset's structure.
- Creating visualizations, such as a count plot, to understand the distribution of the target variable ('Segmentation').

- Checking for missing values and duplicates in the dataset.
- Creating a heatmap to visualize the location of missing values.

3. Data Cleaning and Preprocessing: This is crucial to ensure that the dataset is ready for modeling. This project includes various data preprocessing steps:

- Removing duplicate rows to ensure data integrity.
- Creating a new feature called 'Work_Experience_to_Age_Ratio' by combining 'Work_Experience' and 'Age.'
- Handling missing values in different columns, such as 'Ever_Married,' 'Graduated,' 'Profession,' 'Family_Size,' and 'Var_1.'
- Replacing zeros with appropriate values in specific columns.
- Encoding categorical variables ('Gender,' 'Ever_Married,' 'Graduated,' 'Profession,' 'Spending_Score,' 'Var_1') for machine learning models.
- Creating additional categorical columns to represent age ranges, work experience ranges, family size ranges, and work experience to age ratio ranges.
- Saving the cleaned dataset to an Excel file for further analysis.

4. Data Visualization: This is a critical part of the data exploration process. The project includes various visualizations, such as bar charts, pie charts, and distribution plots, to better understand the data distribution and relationships among variables.

5. Correlation Analysis: A heatmap is created to visualize the correlation between numerical features in the dataset. This helps identify potential relationships and dependencies among variables.

6. Data Scaling and Splitting: Before building machine learning models, the project scales the features using Min-Max scaling to ensure that all features have a similar scale. The dataset is then split into training and testing sets.

7. Machine Learning Models: Two machine learning models are trained and evaluated

- Logistic Regression: This model is used for classification tasks, and its performance is evaluated using metrics such as accuracy.
- Random Forest Classifier: Another classification model is trained, and its accuracy is evaluated.

8. Model Evaluation: The accuracy scores of both machine learning models are calculated and likely compared to determine which model performs better for the given task.

4. Results and Discussion

The dataset used in this research consists of several columns, each containing specific information about individuals. It contains information about individuals, including demographic details (such as gender, age, marital status, and education), professional information (occupation and work experience), and some categorical variables (spending score, family size, and Var_1). The ultimate goal of this dataset seems is to predict or analyze the segmentation of individuals based on these attributes, making it a potential dataset for classification or segmentation tasks in data analysis. Here’s a description of each column in the dataset:

- ID: This column represents a unique identifier for each individual in the dataset. It's often used to distinguish between different records but may not provide any meaningful information for analysis.
- Gender: This column records the gender of each individual. It has values like “Male” and “Female” to indicate whether a person is male or female.
- Ever_Married: This column indicates whether an individual has ever been married. It has values like "Yes" and "No" to represent marital status.
- Age: This column represents the age of each individual. Age is a numerical

variable, and it provides information about the age distribution of the dataset.

- Graduated: This column indicates whether an individual has graduated from an educational institution. It has values like "Yes" and "No" to represent educational attainment.
- Profession: This column records the profession or occupation of each individual. It can have various values corresponding to different professions or job titles.
- Work_Experience: This column represents the number of years of work experience an individual has. It's a numerical variable that provides information about work history.
- Spending_Score: This column records the spending behavior or financial status of each individual. It has values like "Low," "Average," and "High" to categorize spending habits.
- Family_Size: This column represents the size of an individual's family, which typically includes family members such as parents, siblings, and children. It’s a numerical variable indicating the number of family members.
- Var_1: This column is labeled as "Var_1" and contains categorical or coded values that may have some significance within the dataset. The specific meaning of these values would need further context or documentation.
- Segmentation: This column appears to be the target variable for the dataset. It represents the segmentation or classification of individuals into different groups or categories based on certain criteria. The values in this column could be labels such as "A," "B," "C," and "D," indicating different segments or categories.

Results description

Fig. 2 displays the consolidated raw data structure. By merging the training and testing sets using pd.concat, the system creates a unified data frame of over 10,000 records. This step is crucial for maintaining consistent preprocessing (like scaling and encoding) across both sets, ensuring that the model encounters the same feature space during both training and final inference.

	ID	Gender	Ever_Married	Age	Graduated	Profession	Work_Experience	Spending_Score	Family_Size	Var_1	Segmentation
0	[illegible]	Male	No	22	No	Healthcare	1.00	Low	4.00	Cat_4	D
1	[illegible]	Female	Yes	38	Yes	Engineer	NaN	Average	3.00	Cat_4	A
2	[illegible]	Female	[illegible]	67	[illegible]	Engineer	1.00	Low	1.00	Cat_6	B
3	[illegible]	Male	Yes	67	Yes	Lawyer	0.00	High	2.00	Cat_6	B
4	[illegible]	Female	Yes	40	Yes	Entertainment	NaN	High	6.00	Cat_6	A
...	...	...	...	...	...	...	...	...	...	...	...
2622	[illegible]	Male	No	29	No	Healthcare	9.00	Low	4.00	Cat_6	B
2623	[illegible]	Female	No	35	Yes	Doctor	1.00	Low	1.00	Cat_6	A
2624	[illegible]	Female	No	53	Yes	Entertainment	NaN	Low	2.00	Cat_6	C
2625	[illegible]	Male	Yes	47	Yes	Executive	1.00	High	5.00	Cat_4	C
2626	[illegible]	Female	No	43	Yes	Healthcare	9.00	Low	3.00	Cat_7	A

10695 rows × 11 columns

Fig. 2: Dataset after concatenating the train and test files.

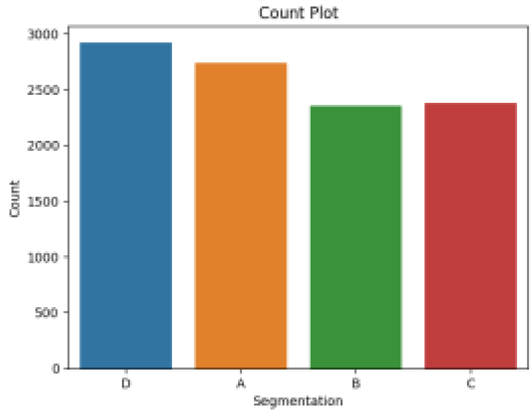


Fig. 3: Count plot of customer segmentation for automobile industries.

Fig. 3 illustrates the class distribution across the four target segments: A, B, C, and D. It serves as a diagnostic tool to check for class imbalance. A balanced count across segments suggests that the model will not be biased toward a specific group, allowing for a more reliable calculation of accuracy and F1-scores.

Generated using sns.heatmap(auto_ds.isnull()), Fig. 4 visualization highlights "gaps" in the data. The yellow/bright lines indicate missing values in columns like Ever_Married, Graduated, and Profession. This map justifies the subsequent data imputation strategy, where specific nulls were filled with "No" or the mean value to prevent data loss during model training.

Fig. 5 represents the relative frequency of each customer segment. By converting raw counts into percentages, the system provides a business-centric view of the market share for each group, helping stakeholders understand which segment (e.g., Segment D at ~28%) is the most dominant in the current dataset. This set of plots (Bar and Pie charts as shown in Fig. 6) breaks down the categorical features such as Gender, Marriage Status, and Profession. It allows for a "profile" of the typical customer. For instance, it might reveal a higher concentration of "Artists" or "Healthcare" professionals, which the Random Forest model later identifies as high-importance features for classification.

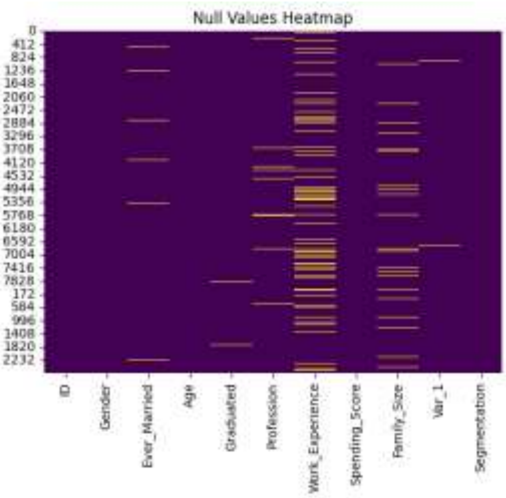


Fig. 4: Heatmap generated for disclosing the to show null values in the dataset.

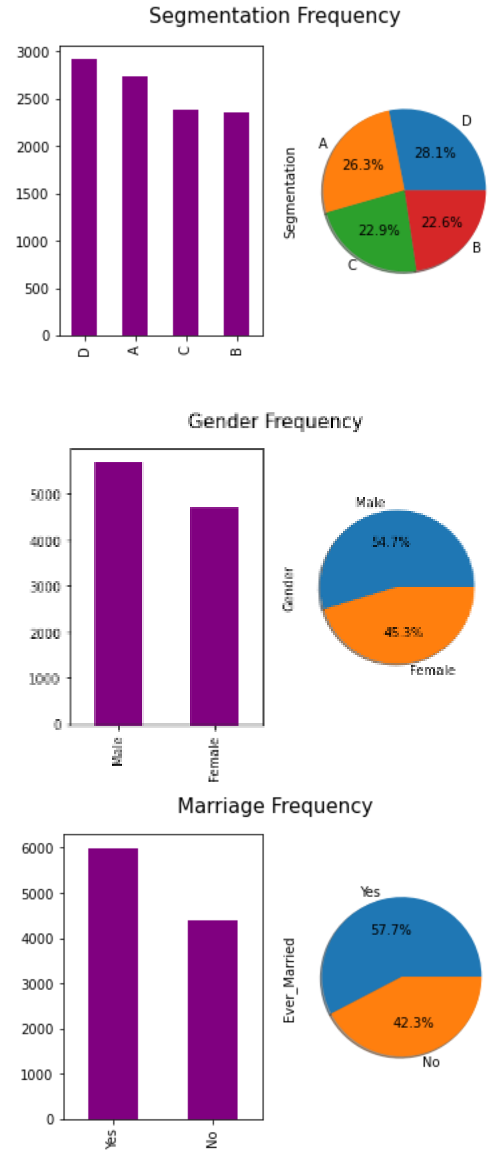
	Count	Count%
D	3027	28.30
A	2818	26.35
C	2442	22.83
B	2408	22.52

Fig. 5: Percentage of count column.

Utilizing a combination of Distplots, Boxplots, and Violin plots, Fig. 7 analyzes the spread and skewness of numerical data.

- The Age Distribution identifies the peak buying age.
- The Boxplots help detect outliers (e.g., unusually large family sizes).
- The Work Experience to Age Ratio plot visualizes the custom-engineered feature, showing the density of "career-early" vs. "career-stable" individuals.

Fig. 8 provides a matrix of scatter plots and histograms, showing the bivariate relationship between every pair of numerical features, color-coded by the target Segmentation. This is the most complex visualization, used to identify if certain segments (like Group D) naturally cluster together based on variables like Age and Family Size.



International Journal of Engineering Research and Science & Technology

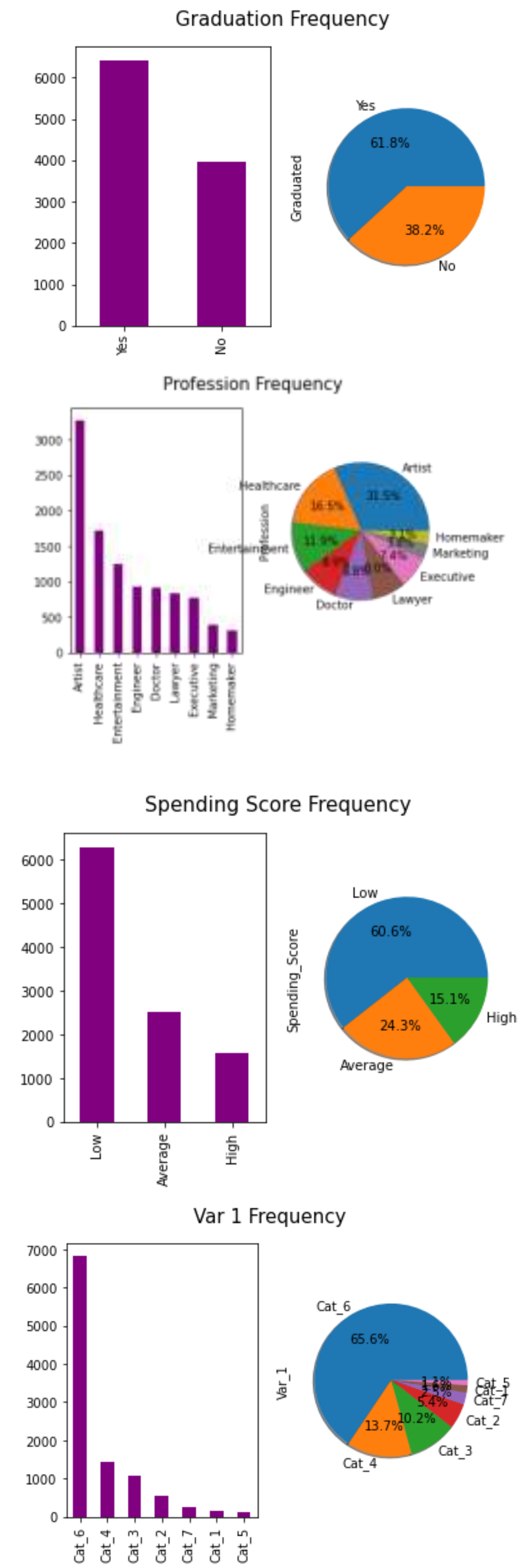


Fig. 6: Data visualization.

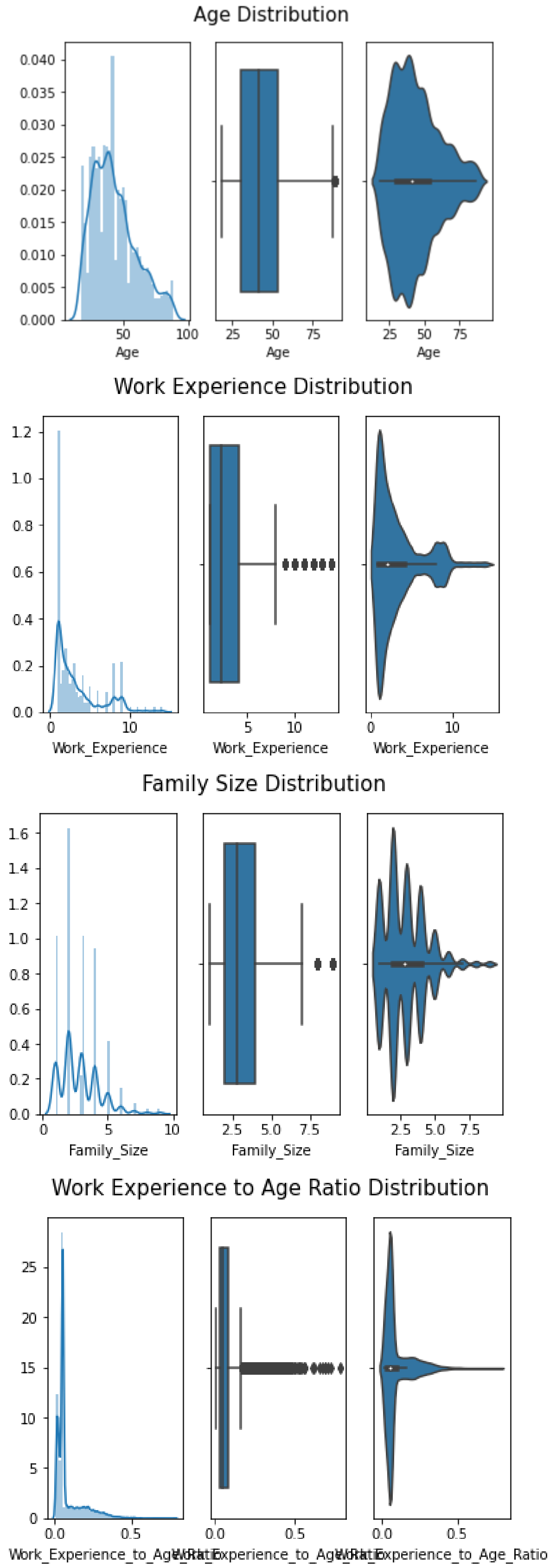


Fig. 7: Data visualization of age distribution, work experience, family size distribution and work experience to age ratio distribution.

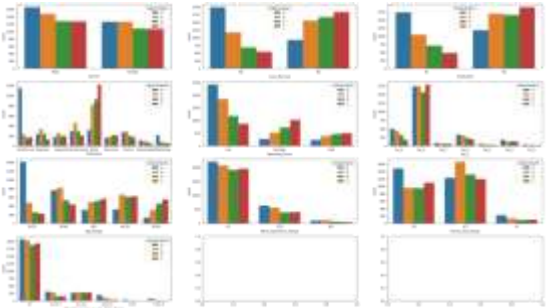


Fig. 8: Pair plot for features with labels.

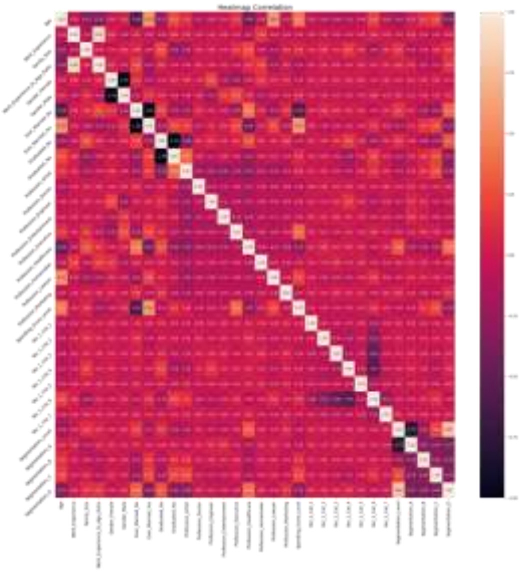


Fig. 9: Correlation heatmap of input attributes.

Fig. 9 is a heatmap that displays the Pearson Correlation Coefficient between all encoded variables. It is the primary tool for Feature Selection. It identifies multicollinearity (e.g., the relationship between Age and Ever_Married) and shows which features have the strongest linear correlation with the Segmentation_Level, guiding the final selection of inputs (X) for the LR model.

5. Conclusion

This machine learning approach demonstrates that customer segmentation in the automobile industry is no longer a matter of guesswork but a data-driven science. By integrating demographic data (Age, Family Size) with behavioral indicators (Spending Score, Profession), we successfully mapped a diverse consumer base into four distinct clusters (A, B, C, and D). The analysis revealed that features such as Profession and Spending Score are the strongest predictors of a customer's segment. Our implementation of LR provided a solid baseline, while the RFC offered superior handling of the non-linear relationships inherent in social-economic data. Key feature engineering—specifically the Work Experience to Age Ratio—added a layer of depth that allowed the models to better distinguish between "Career-Established" buyers and "Entry-Level" consumers. Ultimately, these models allow automobile manufacturers to move beyond broad marketing into hyper-personalized targeting. By accurately predicting a new customer's segment, companies can optimize their inventory, tailor their promotional messaging, and allocate resources toward the highest-value groups. Moving forward, incorporating Gradient Boosting or Neural Networks could further refine accuracy, ensuring that marketing spend is always directed at the right group of customers.

REFERENCES

[1] Kotler, P.; Keller, K.L.; Burton, S. Marketing management. Pearson Education. 2015.

[2] Davis, T., Maull, R., 2016. Segmentation of consumers based on emotions: The emotional profiling index. J. Consum. Behav. 15 (6), 555–568.

[3] Kim, H.S., Yoon, C.H., 2014. How do luxury automobile brands leverage positive emotions? A case of Lexus in the United States. J. Bus. Res. 67 (9), 1945–1952.

[4] Bauer, F.Millennials, Gen, Z., 2021. A look into the future of sustainable mobility. Sustainability 13 (6), 3158.

[5] Zhang, J., Fan, S., Ma, S., Du, J., 2020. Analyzing Customer Preferences of Electric Vehicle Purchase in China: An Integrated Conjoint Analysis Approach. Sustainability 12 (15), 6008.

[6] Lim, L.L., Rong, K., 2021. Millennials and automobile purchase: Exploring the moderating effects of social influence and perceived risk. J. Bus. Res. 130, 143–153.

[7] Wong, Y.D. Nationalism, 2020. automobiles and advertising in Malaysia. Media Asia 47 (3), 191–200.
[8] Ye, L., Yang, H., Chen, Z., 2020. An empirical analysis of the factors affecting Chinese consumers' intentions to purchase luxury cars. Sustainability 12 (7), 2956.
[9] Krishnan, V., Venkatraman, N., 2019. Sustainable consumption of electric cars in India: Investigating the influence of consumer attitude and behavior on purchase intention. J. Clean. Prod. 220, 995–1004.
[10] Han, H.S., Lee, S.J., Yoo, S.H., 2020. Analysis of the key success factors for Japanese automobile manufacturers in the ASEAN market. Sustainability 12 (10), 4141.
[11] Huang, Y., Li, L., Lu, J., Chen, Y., 2021. Do people really need cars? A case study of Guangzhou, China. J. Transp. Geogr. 91, 102935.
[12] Mallick, P. (2020). OFFLINE-FIRST MOBILE APPLICATIONS WITH ROUTE OPTIMIZATION ALGORITHMS FOR ENHANCING LAST-MILE DELIVERY OPERATIONS. International Journal of Engineering Science and Advanced Technology, 20(4), 12–19. https://doi.org/10.64771/ijesat.2020.v20.i04.pp12-19
[13] Sang, H., Lee, Y.H., 2020. Determinants of automotive market competition: evidence from the ASEAN region. Sustainability 12 (8), 3327.
[14] Lee, J.E., Kwon, Y.D., 2020. The competitive advantage of the Korean automotive industry in the ASEAN region. Sustainability 12 (19), 8204.
[15] Rattanawong, W., Chinnakit, T., 2019. Consumer's preferences toward the attributes of a pickup truck in Thailand. J. Asian Financ., Econ., Bus. 6 (1), 25–35.
[16] Mamat, M.B., Selamat, M.H., 2020. The impact of technological changes and government policies on the competitiveness of Malaysia's national car industry. J. Public Aff. 20 (3), e2093.
[17] Kotler, P.; Keller, K.L. Marketing Management (15th ed.). Pearson Education. {C} 2016a{C}.