Customer Segmentation and Predictive Modeling for Enhanced Marketing Strategies in the Automobile Industry

Prepared By

Nathaniel Zhu, Youyi Fu, Xiaoya Wei, Xujuan Liang

Final Report - STAT 208

Professor Brandon Wales

University of California, Riverside

Section I: Introduction / Pre-analysis

Since the pandemic happened, the automobile industry has been suffering from a declining situation - car prices are higher, with market incentives added to the purchase, and fewer vehicles in stock since the shortage of semiconductors. But three years after the pandemic and nowadays, the automobile industry has been rapidly evolving and expanding - more cars in stock, car price has been fairly altered, and more customers are looking for cars. And now, it is crucial for automobile manufacturers and sellers to understand customer behavior and preferences to maintain a competitive advantage. As companies aim to expand into new markets, leveraging data-driven strategies to segment customers and predict their behavior can significantly enhance marketing effectiveness and operational efficiency. This report will bring understanding of a comprehensive data analysis project which focuses on customer segmentation and predictive modeling to support an automobile company in its venture to enter new markets with existing customers and products.

Why This Topic?

Customer segmentation and predictive analysis modeling are fundamental in creating targeted marketing strategies. By identifying distinct customer segments and understanding their spending behaviors, companies can match their marketing campaigns to address the specific needs and preferences of each segment. This targeted approach not only improves customer satisfaction but also optimizes marketing expenditures, leading to higher conversion rates and increased revenue.

In the context of the automobile industry, where customer preferences can vary widely based on demographic and behavioral characteristics, segmenting customers accurately and predicting their future value becomes even more important. This project addresses these challenges by utilizing advanced machine learning techniques to analyze customer data, segment them effectively, and predict key metrics such as spending scores.

Pre-analysis

Before we are going deep into descriptive analysis and model selections, a preliminary examination of the data was conducted to understand the structure and content. This dataset includes various features related to customer demographics, behavior, and interactions with the company. Initial data exploration revealed the presences of missing values and categorical variables, necessitating approach preprocessing steps:

- Aggregating and inspection of the dataset to identify key features and distributions
- Handling missing value using forward fill to maintain data integrity.
- Label encoding categorical variables to transform them into a format suitable for machine learning models.

This fundamental work set the stage for the subsequent detailed analyses and model building, ensuring a robust and reliable approach to achieving the project's objectives.

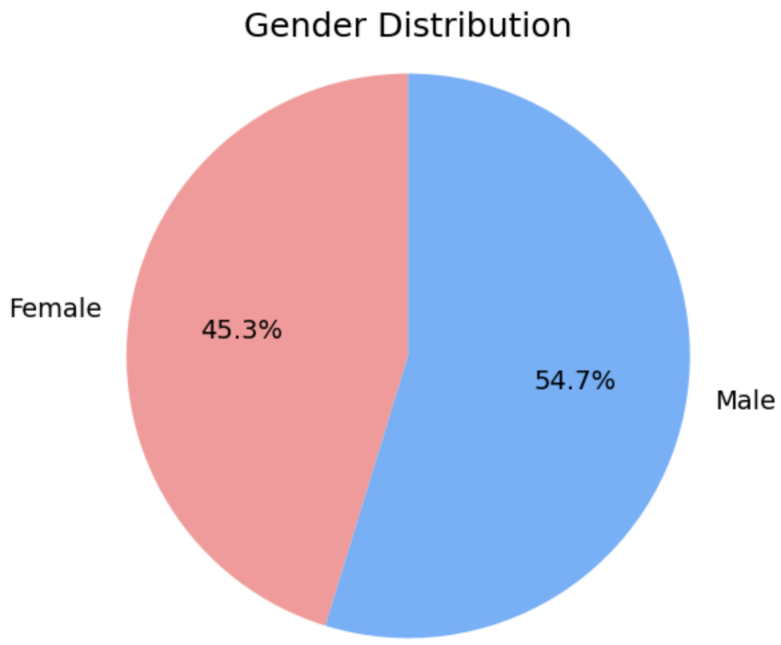
Now, we are going to pace into descriptive analysis.

Section II: Descriptive Analysis

Overview

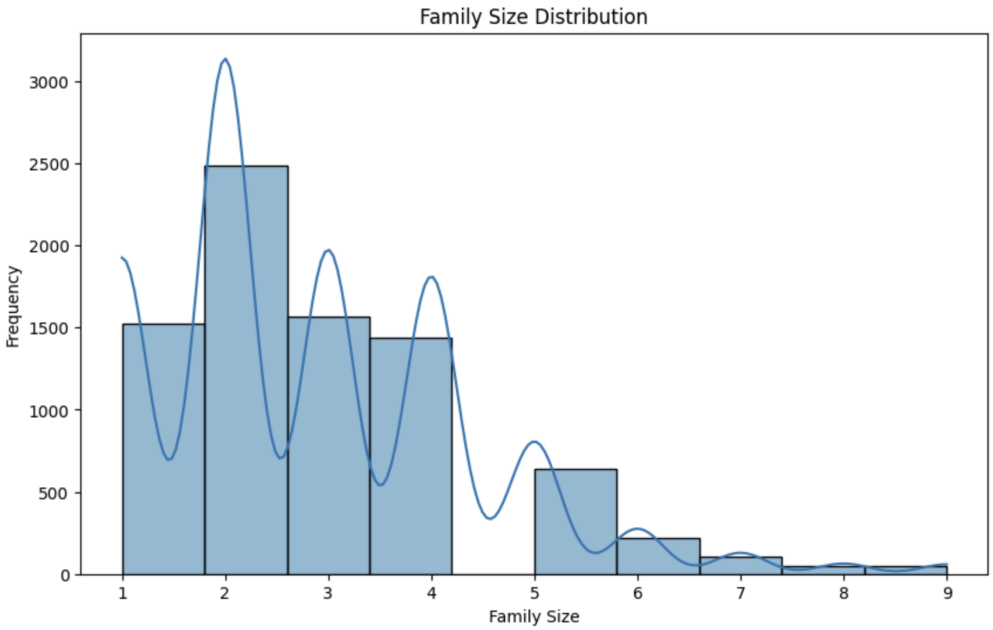
In Part II of our project analysis, we conducted a comprehensive descriptive analysis of the dataset with a primary focus on gender, family size, age, and spending score distributions. This part also includes visualizations to better understand the data's underlying patterns and distributions.

Gender Distribution



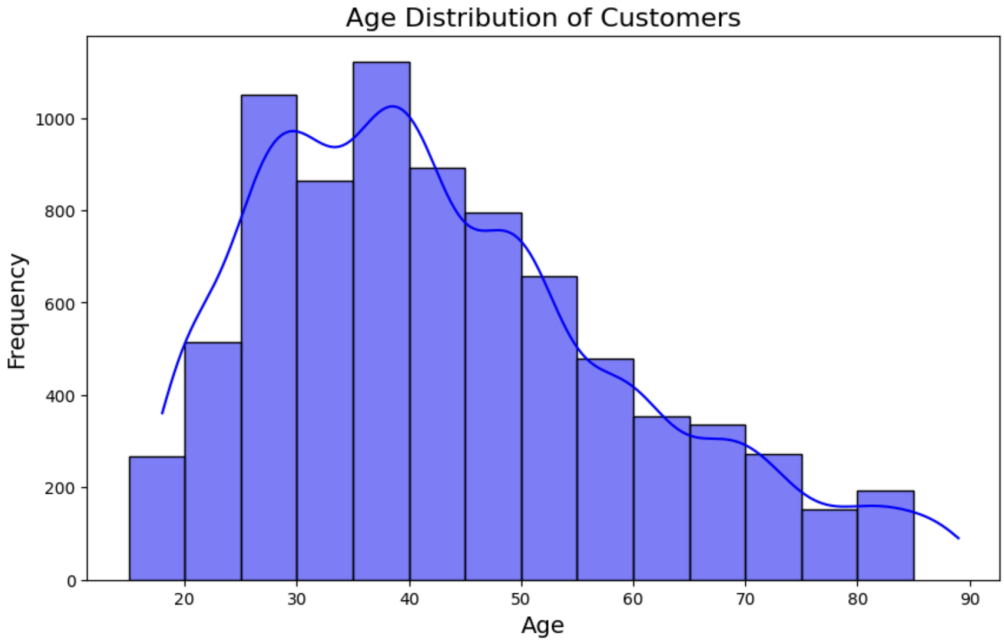
The dataset contains records from 8,068 respondents, consisting of 3,651 females and 4,417 males. The gender distribution was visualized using a pie chart, illustrating a slightly higher proportion of male respondents (54.7%) compared to female respondents (45.3%).

Family Size Distribution



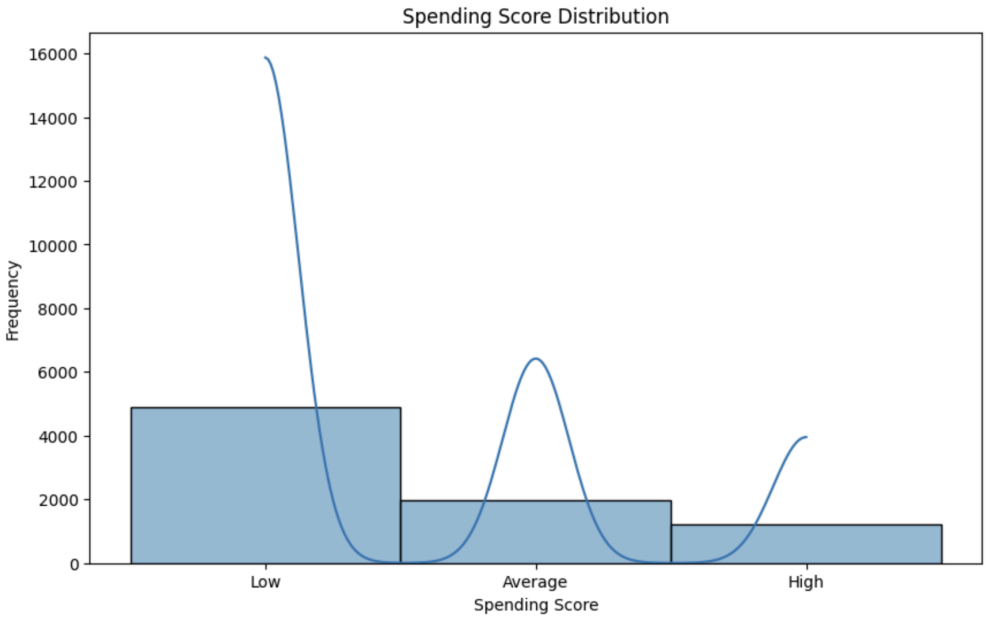
We analyzed the family size distribution which showed a continuous variable with a wide range of values. The histogram with a kernel density estimate (KDE) overlay provided insights into the commonality of different family sizes, highlighting a mode at smaller family sizes.

Age Distribution



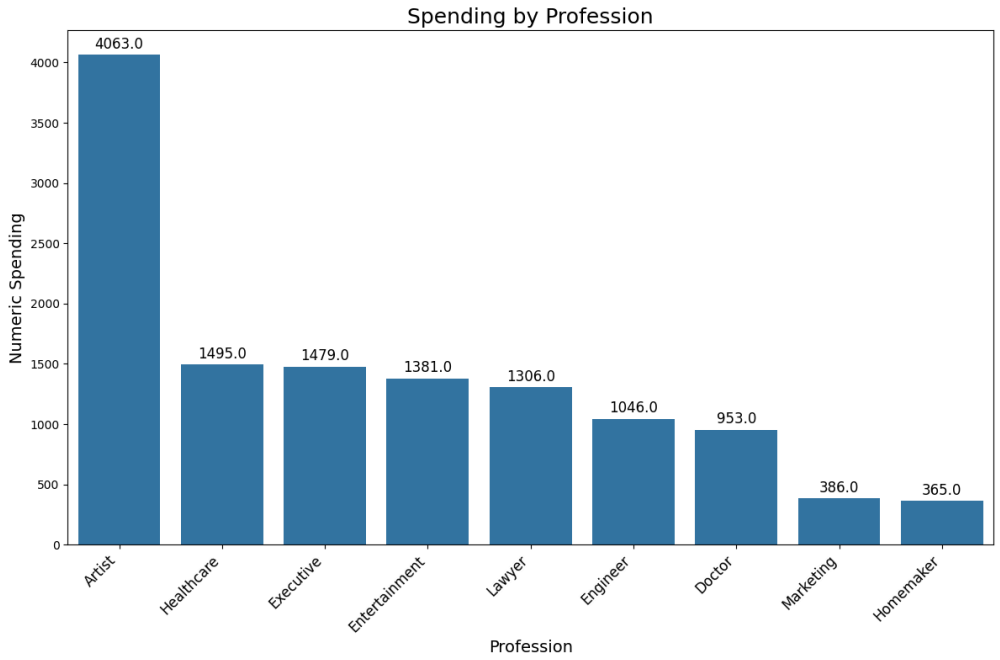
Age data was visualized using a histogram segmented into five-year intervals, which helped in identifying the age groups most prevalent in our dataset. The KDE showed a bimodal distribution, suggesting two major age groups within our respondents.

Spending Score Distribution



The distribution of spending scores was detailed through a histogram, indicating three distinct groups categorized as low, average, and high spenders. This visualization was integral in understanding the spending behaviors associated with different demographic groups.

Spending Power by Profession

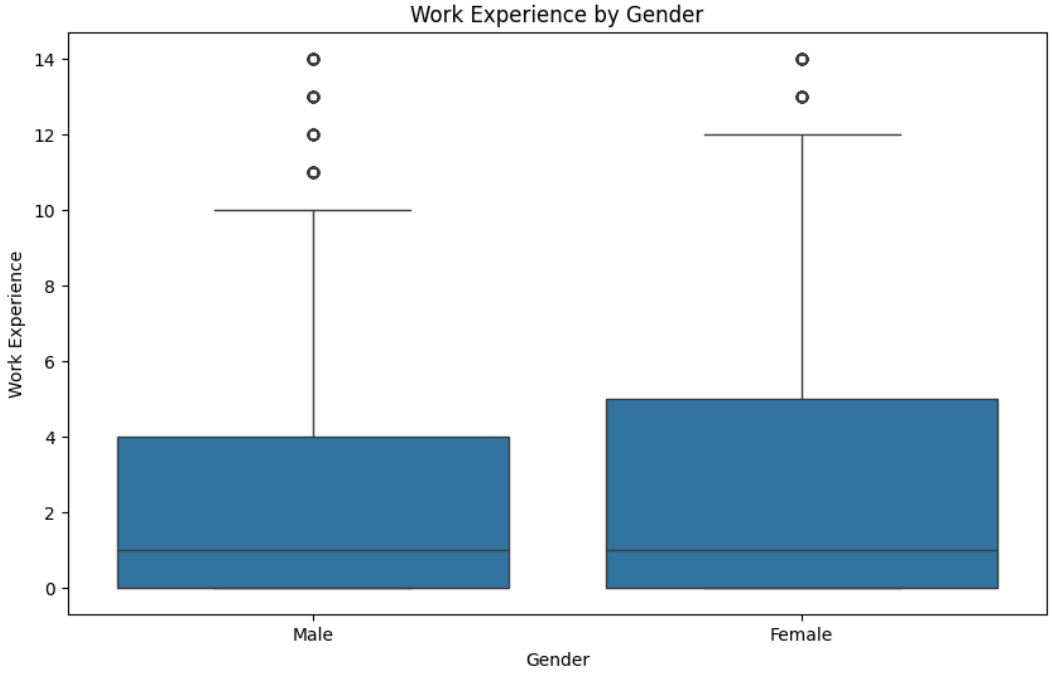


An analysis of spending power segmented by profession revealed significant variance among different occupational groups. Artists and healthcare professionals showed the highest and second-highest spending scores respectively, offering insights into economic behaviors by profession.

Spending by Gender

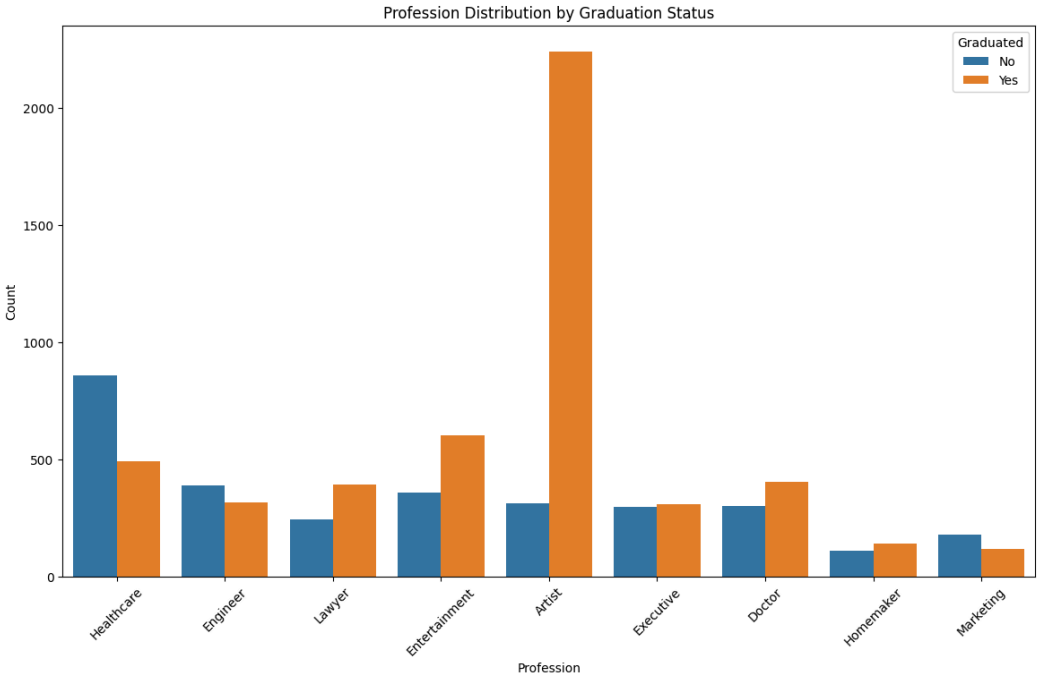
In this analysis of a Box and Whisker plot, we will understand the distribution of spending scores by gender, comparing male and female customers. Both genders exhibit a similar range in spending scores, with the interquartile range representing the middle 50% of the data being nearly identical for males and females. Since this is a categorical data, and from these two identical IQR range, this suggests that there is no significant difference in spending behavior between male and female customers.

Work Experience by Gender



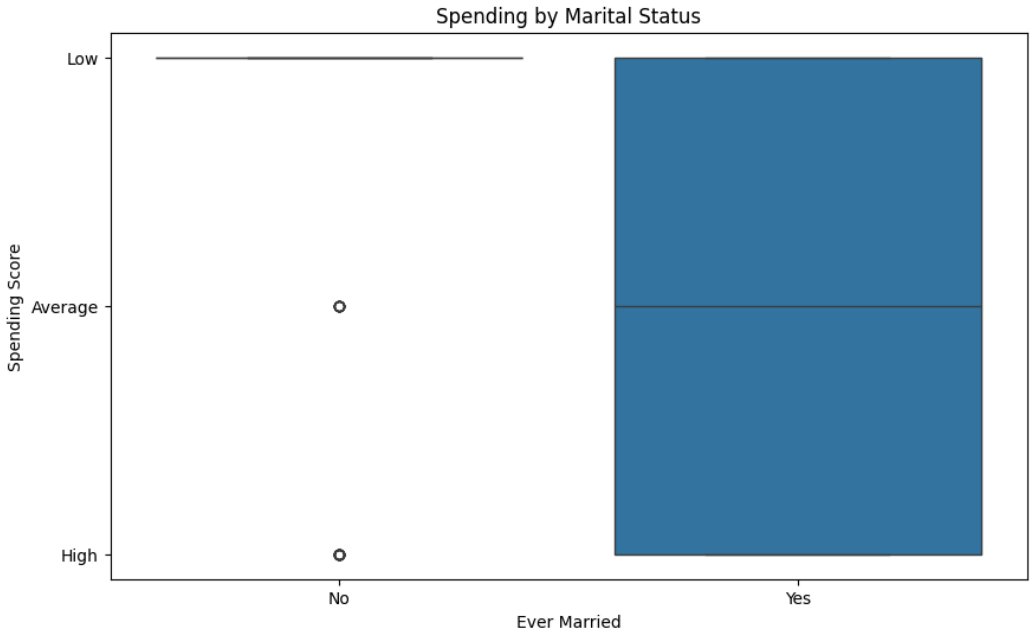
The work experience distribution by gender shows that males have a broader range of work experience compared to females. This indicates a wider variety of career lengths among male respondents.

Profession Distribution by Graduation Status



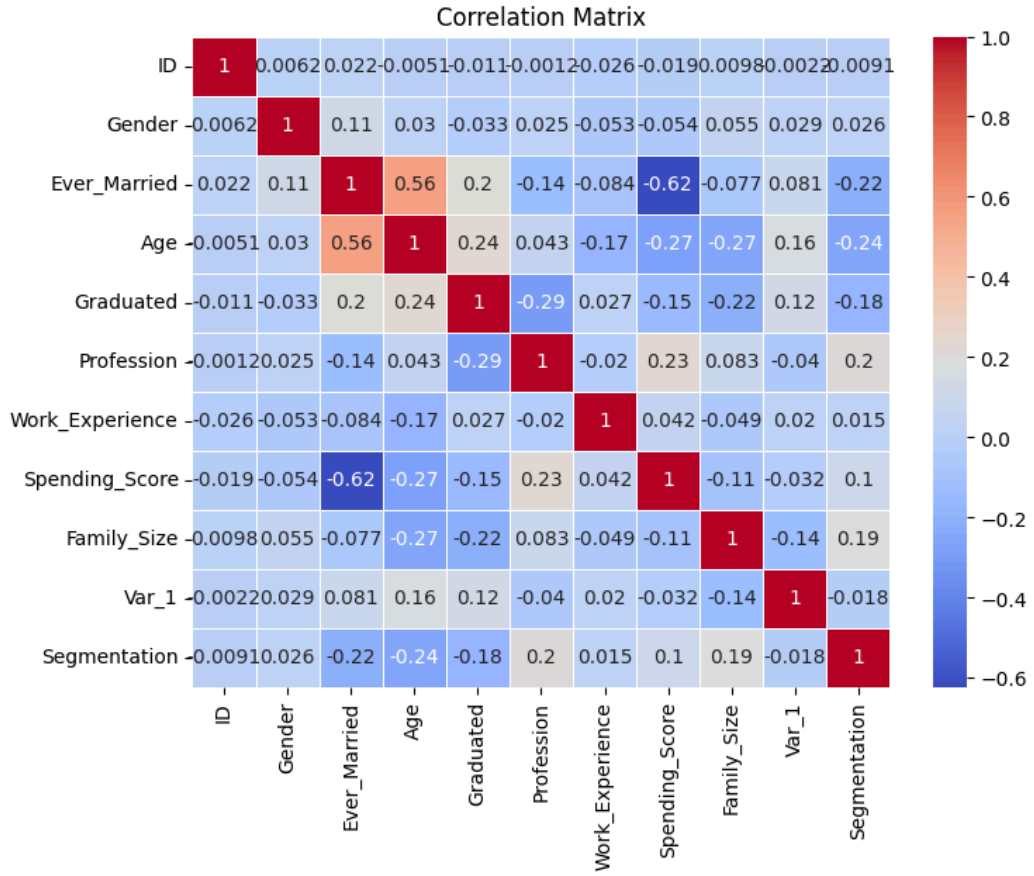
The profession distribution by graduation status reveals that a significant proportion of professionals across various fields are graduates. This underscores the importance of higher education in securing professional roles in the dataset.

Spending by Marital Status



An analysis of spending scores by marital status indicates that married individuals tend to have higher spending scores compared to those who have never married. This suggests that marital status may influence spending behavior, potentially due to dual-income households or family-related expenditures.

Correlation Analysis



A correlation matrix was created to examine the relationships between all numerical variables in the dataset after converting categorical variables via label encoding. This matrix is crucial for identifying potential predictors for more complex analyses and understanding the interdependencies between different variables.

Conclusion for Descriptive Analysis

The descriptive analysis in Section II provided foundational insights into the dataset, highlighting key demographic distributions and spending patterns. These analyses are critical for the subsequent predictive modeling and customer segmentation tasks outlined in Section III of the project.

Section III: Model Selection

Part 1: Predictive Modeling for Spending Score

In this part of the analysis, we focused on building predictive models to estimate the spending scores of customers. Five different regression models were evaluated: Linear Regression, Decision Tree Regressor, Random Forest Regressor, Gradient Boosting Regressor, and Support Vector Regressor. The performance of each model was measured using Mean Squared Error (MSE). Below are the outcome performances of these predictive models.

Initial Model Performance

The initial results indicated varying levels of accuracy among the models, with the Random Forest Regressor achieving the lowest MSE of 0.1973, demonstrating superior performance in predicting spending scores. The performance summary are as follows:

- Linear Regression MSE = 0.4134
- Decision Tree Regressor: MSE = 0.3689
- Random Forest Regressor: MSE = 0.1973
- Gradient Boosting Regressor: MSE = 0.3044
- Support Vector Regressor: MSE = 0.984 (initially)

Improvement of Support Vector Regressor

The initial MSE for the Support Vector Regressor was significantly higher than all of other models with a MSE of 0.9840, indicating potential issues with the choices of hyperparameters. To address this, a parameter grid search was performed using GridSearchCV to identify the optimal settings for the SVR. The parameter grid search included variations in kernel types (only chosen three since training model takes time), penalty parameter (‘C’ or ‘Cost’), and epsilon values. The parameter grid was as follow:

Kernel = Linear, Poly, RBF
C = 0.1, 1, 10
Epsilon = 0.01, 0.1, 1

We originally decided to pursue four penalty parameters and epsilon values, however, this model takes around 10-30 minutes to perform, which greatly lowers the reproducibility of this data analysis report. Such so, we pursue with only three penalty parameters and epsilon values.

Standardizing the data before implementing the SVR improvements also played a crucial role in enhancing the performance. By standardizing the features, the model become more sensitive to variations in the data, leading to more accurate predictions.

Final Model Performance

After performing the SVR improvement procedure, the MSE for the support vector regressor significantly improved to 0.3378, this marked a substantial reduction from the previous value, indicating that the model could make more accurate predictions. The final performance summary for the predictive models is as follows:

- Linear Regression MSE = 0.4134
- Decision Tree Regressor: MSE = 0.3689
- Random Forest Regressor: MSE = 0.1973
- Gradient Boosting Regressor: MSE = 0.3044
- Support Vector Regressor: MSE = 0.3378 (after tuning)

Conclusion for Section 3, Part 1

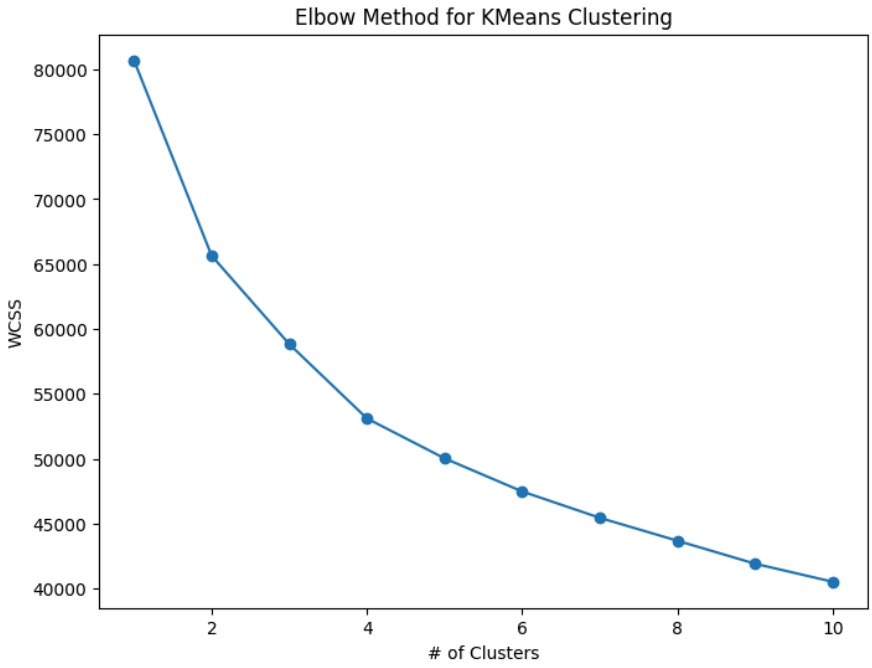
Among the evaluated models, the Random Forest Regressor performed the best for predicting spending scores with the lowest MSE. The significant improvement in the SVR's performance after tuning demonstrates the importance of hyperparameter optimization. These results provide a robust foundation for predicting customer spending behavior, which can be leveraged to create targeted marketing strategies and enhance customer engagement.

Part 2: Customer Segmentation Clustering - K-Means

The primary objective of this part of the analysis is to create new segments of customers based on clustering to better understand different customer groups. By identifying distinct clusters within the customer base, the company can tailor its marketing strategies to meet the specific needs and preferences of each segment.

Step 1: Determining the Optimal Number of Clusters

To identify the optimal number of clusters, the Elbow Method was employed. This method involves plotting the Within-Cluster Sum of Squares (WCSS) against the number of clusters and identifying the "elbow point" where the rate of decrease sharply slows. This point indicates the most appropriate number of clusters for the dataset. Below is the Elbow Chart



The Elbow Method plot suggests that the optimal number of clusters is 4, as indicated by the sharp bend at this point. This choice balances model complexity and the explanatory power of the clustering.

Step 2: Applying K-Means Clustering

With the optimal number of clusters determined, K-Means clustering was applied to the scaled training data. This method partitions the dataset into 4 clusters, each characterized by its centroid, which represents the average position of all the points within the cluster.

The table below displays the centroids of the 4 clusters, showcasing the average values for each feature within the clusters:

Cluster	ID	Gender	Ever_Married	Age	Graduated	Profession	Work_Experience	Spending_Score	Family_Size	Var_1	Segmentation
0	463462.096995	0.454463	0.414845	41.377049	0.828324	1.228142	3.112022	1.988160	1.913934	4.196721	0.801002
1	463554.754498	0.586565	0.997201	46.648541	0.757297	1.493003	2.572171	0.240304	3.207117	4.151939	1.410636
2	463415.966318	0.539372	0.086937	27.721438	0.284479	4.321802	2.865726	1.969504	3.739190	3.817478	2.609012
3	463468.670358	0.653322	0.959114	70.063884	0.581772	5.624361	1.333049	1.429302	2.153322	4.602215	1.348382

Step 3: Visualizing Customer Segments

To visualize the clustering results, a scatter plot was generated, displaying the relationship between age and work experience for each customer, color-coded by cluster membership. This visualization helps to understand the distribution and characteristics of each cluster.

Analysis and Insights

The clustering analysis revealed four distinct customer segments, each with unique characteristics:

1. Cluster 0: Represents a balanced mix of gender and marital status, with an average age around 41 and relatively high spending scores.
2. Cluster 1: Predominantly married individuals with higher ages and lower spending scores, suggesting a conservative spending behavior.
3. Cluster 2: Younger individuals with lower marriage rates and moderate spending scores, potentially indicating a less stable customer base.
4. Cluster 3: Older, predominantly married individuals with varying spending scores, indicating a diverse segment in terms of spending behavior.

Conclusion for Section III Part 2

The K-Means clustering approach effectively segmented the customer base into four distinct groups, each with specific demographic and behavioral characteristics. These insights can be leveraged to tailor marketing strategies, develop targeted campaigns, and improve customer engagement, ultimately driving higher customer conversion and retention rates and satisfaction.

By understanding the distinction of each cluster, the automobile company can optimize its market entry strategy and ensure that its marketing efforts are aligned with the needs and preferences of different customer segments.

Part 3: Customer Segmentation Prediction (Using Classification)

The objective of this analysis is to predict the customer segments for a new set of potential customers using three classification methods: Logistic Regression, Random Forest Classifier,

and Gradient Boosting Classifier. By comparing these models, we aim to identify the most reliable method for predicting customer segments, which are labeled as A, B, C, and D.

Step 1: Model Predictions

Three classification models were trained on the existing customer data to predict the segmentation of new potential customers. The predicted segments were then transformed back to categorical variables (A, B, C, D) and appended to the test dataset.

	ID	LogReg_Segmentation	RFC_Segmentation	GBC_Segmentation
0	458989	A	B	A
1	458994	C	C	B
2	458996	A	A	D
3	459000	A	C	B
4	459001	D	D	D

Figure: Predict Segments of the First Five Test_Data customers

Step 2: Distribution of Predicted Segments

	LogReg_Segmentation	RFC_Segmentation	GBC_Segmentation
A	713	629	718
B	244	568	472
C	783	638	609
D	887	792	828

The figure above shows the counts of each individual segment. They indicate variability in the segmentation results across the models, highlighting the need to evaluate their accuracy and reliability.

Step 3: Model Evaluation

The performance of each model was evaluated using cross-validation accuracy scores, classification reports, and confusion matrices.

Figure: Logistic Regression Classification Report and Confusion Matrix

```
Logistic Regression Classification Report:
              precision    recall  f1-score   support

           0       0.41      0.44      0.43      1972
           1       0.36      0.14      0.20      1858
           2       0.48      0.61      0.54      1970
           3       0.61      0.74      0.67      2268

    accuracy                           0.50      8068
   macro avg       0.47      0.48      0.46      8068
weighted avg       0.47      0.50      0.47      8068

Logistic Regression Confusion Matrix:
[[ 876  189  412  495]
 [ 586  260  749  263]
 [ 282  202 1193  293]
 [ 395   77  128 1668]]
```

Figure: Random Forest Classifier CR and CM

```
Random Forest Classifier Classification Report:
              precision    recall  f1-score   support

           0       0.41      0.41      0.41      1972
           1       0.35      0.32      0.34      1858
           2       0.51      0.50      0.51      1970
           3       0.62      0.67      0.64      2268

    accuracy                           0.49      8068
   macro avg       0.47      0.48      0.47      8068
weighted avg       0.48      0.49      0.48      8068

Random Forest Classifier Confusion Matrix:
[[ 807  460  262  443]
 [ 461  603  560  234]
 [ 259  456  988  267]
 [ 435  182  130 1521]]
```

Figure: Gradient Boosting Classifier CR and CM

```
Gradient Boosting Classifier Classification Report:
              precision    recall  f1-score   support

           0       0.45      0.45      0.45      1972
           1       0.41      0.33      0.37      1858
           2       0.58      0.58      0.58      1970
           3       0.63      0.73      0.68      2268

    accuracy                           0.53      8068
   macro avg       0.52      0.52      0.52      8068
weighted avg       0.52      0.53      0.53      8068

Gradient Boosting Classifier Confusion Matrix:
[[ 896  418  222  436]
 [ 458  621  543  236]
 [ 199  353 1138  280]
 [ 446  111   58 1653]]
```

Step 4: Cross-Validation Accuracy

The cross-validation accuracy scores for the three models indicate their overall performance:

	Logistic Regression CV:	Random Forest Classifier CV:	Gradient Boosting Classifier CV:
0	0.4954	0.4857	0.534

Analysis and Insights

The cross-validation accuracy scores suggest that the Gradient Boosting Classifier is the most reliable model for predicting customer segments, achieving the highest accuracy score of 0.534. Despite this, the accuracy scores for all three models are relatively low, indicating that while the Gradient Boosting Classifier is the best among the evaluated models, its predictive power is still limited. The inconsistency in the predicted segment counts further underscores the variability in model performance.

Conclusion for Section III Part 3

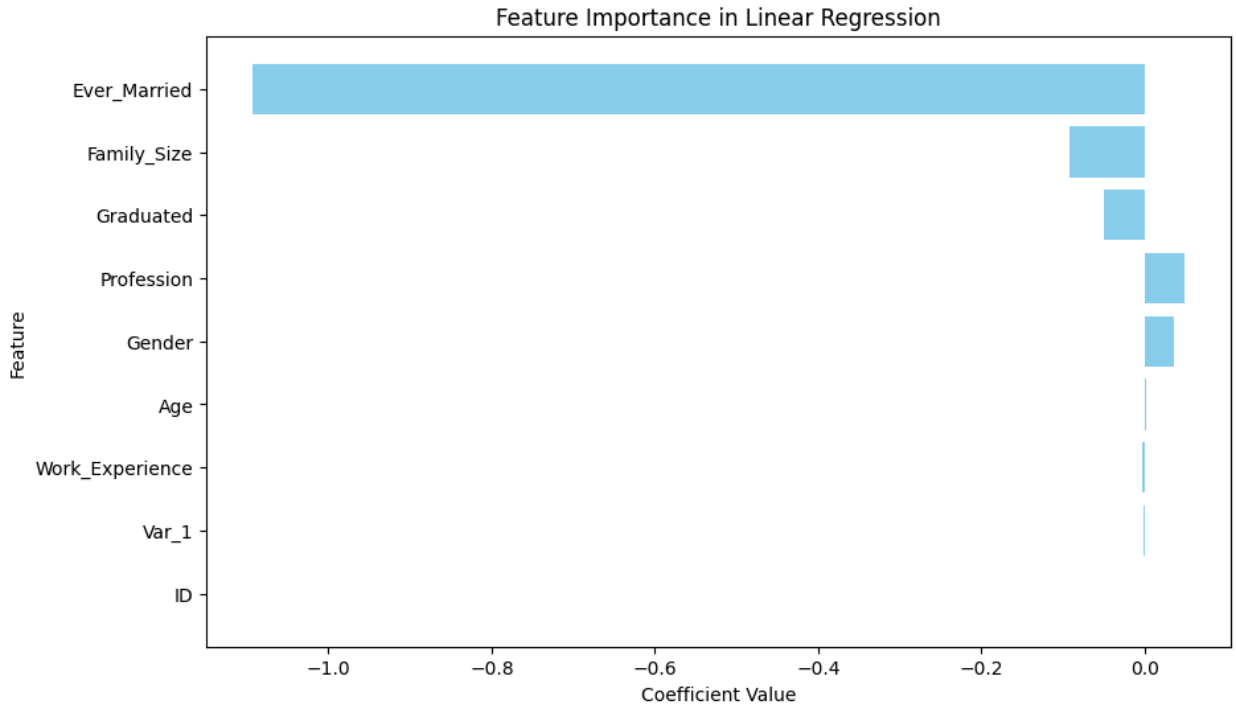
While the Gradient Boosting Classifier demonstrates the highest predictive power for customer segmentation, the overall accuracy scores highlight the need for further model refinement and potential exploration of additional features or more sophisticated techniques. The results suggest

that while the models can provide some insights into customer segmentation, they should be used cautiously and supplemented with additional analysis for more confident decision-making.

Part 4: Feature Importance Analysis

Feature importance analysis is a powerful tool in machine learning that helps to interpret models, reduce dimensionality, and gain insights into the data. Different algorithms provide different methods to calculate feature importance, each with its own advantages. Understanding and applying feature importance analysis can significantly enhance the effectiveness and interpretability of our machine learning models.

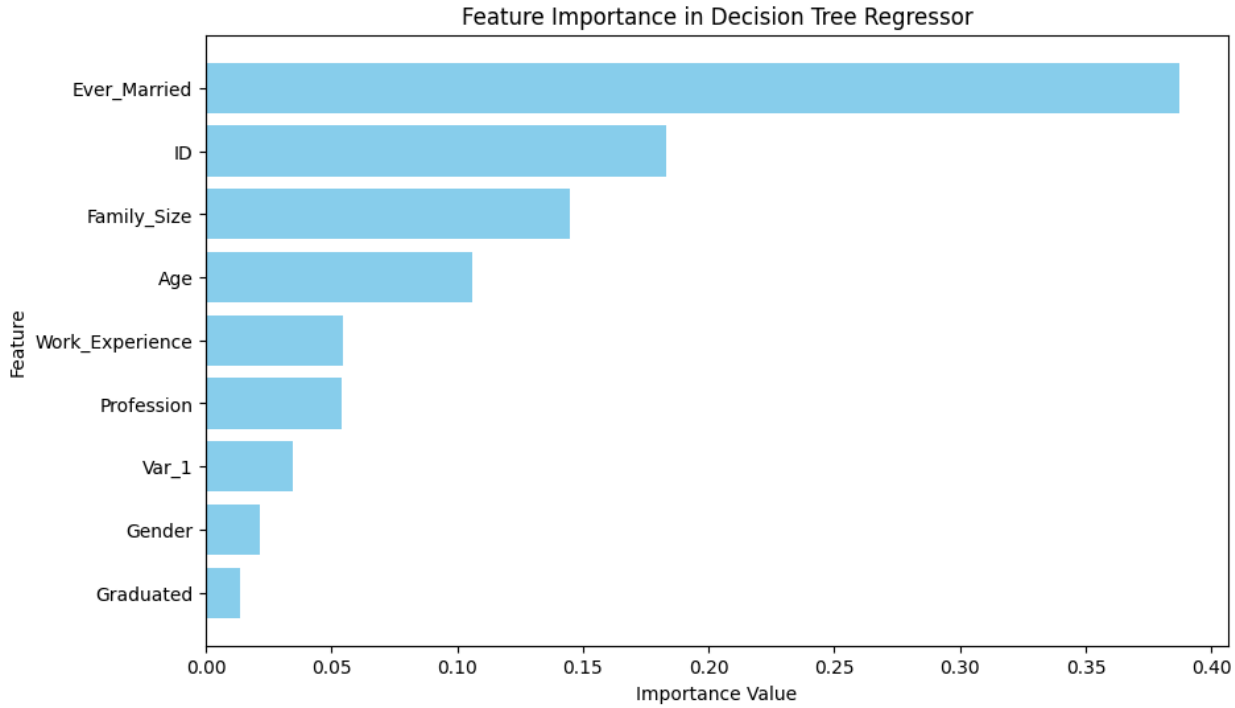
1.Linear Regression Model



The feature “Ever_Married” has a significantly larger coefficient in magnitude compared to all other features. This indicates that whether an individual is married has the strongest influence on the target variable among the features considered. The negative value suggests that being married is associated with a decrease in the target variable.

Features like “Family_Size,” “Graduated,” “Profession,” and “Gender” also influence the target variable but to a much lesser extent compared to “Ever_Married.” These features have smaller coefficients, indicating a weaker impact on the predictions.

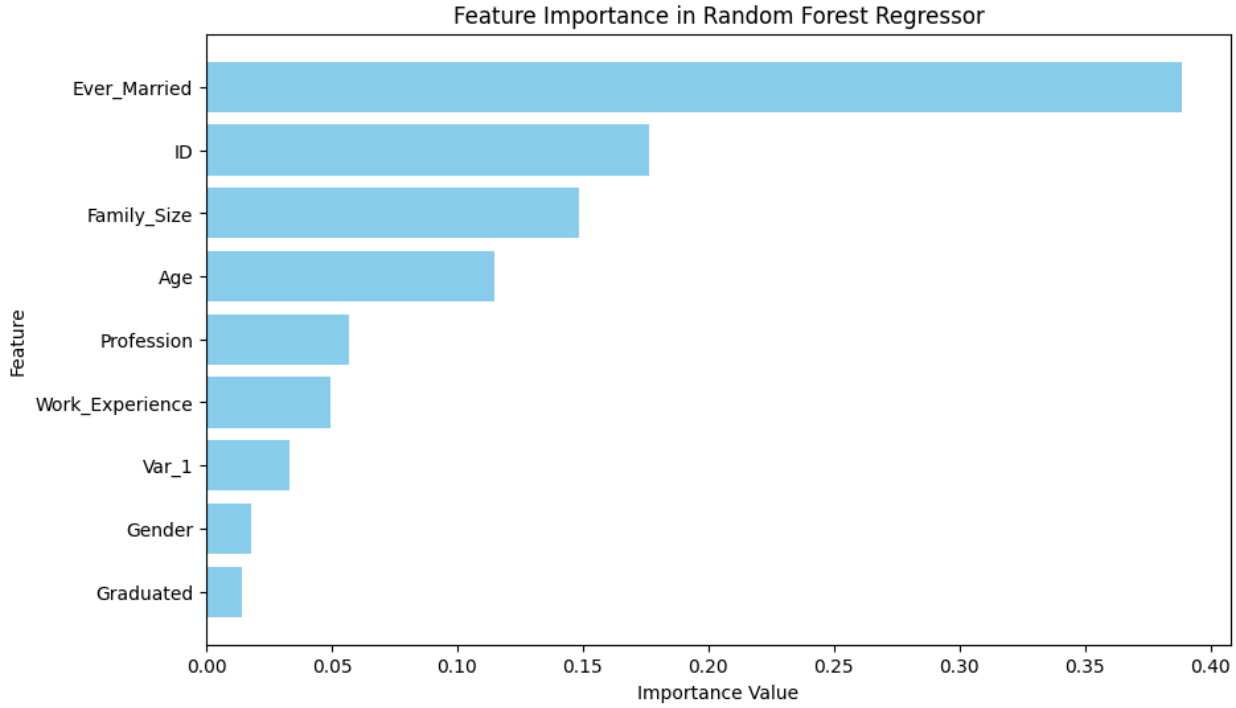
2.Decision Tree Model



The feature “Ever_Married” has the highest importance value, indicating it is the most influential feature in predicting the target variable. This suggests that whether an individual is married plays a significant role in the model’s predictions.The features “ID,” “Family_Size,” and “Age” also have relatively high importance values. This implies that these features significantly contribute to the model’s predictions, albeit to a lesser extent than “Ever_Married.”“Gender” and “Graduated” have the lowest importance values, indicating they have minimal influence on the model’s predictions. These features could potentially be excluded from the model without significantly affecting its performance.

The decision tree regressor relies heavily on a few key features, with “Ever_Married” being the most critical, while other features also contribute but to varying degrees. This analysis helps in understanding the relative contribution of each feature to the model’s performance and can guide feature selection and model refinement efforts.

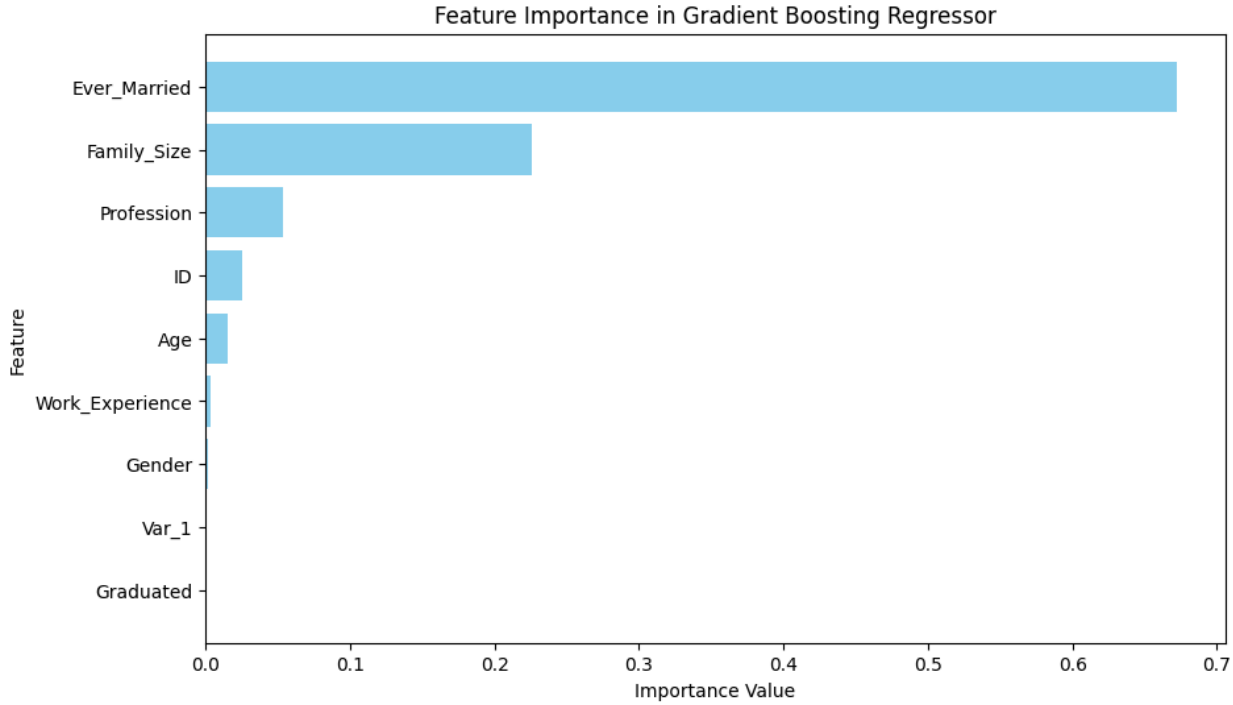
3.Random Forest Model



Similar to the previous models, “Ever_Married” emerges as the most important feature in the random forest model. It has the highest importance value, indicating that marital status is the most influential factor in predicting the target variable.The features “ID,” “Family_Size,” and “Age” also have relatively high importance values. These features significantly contribute to the model’s predictions and are crucial for the model’s performance.

The feature importance ranking is consistent with the decision tree regressor, with “Ever_Married” being the top feature in both models. However, the random forest model provides a more robust and averaged importance measure due to its ensemble nature.

4.Gradient Boosting Regressor



In the same situation as former models, the feature “Ever_Married” has the highest importance value by a significant margin. This indicates that marital status is the most influential feature in predicting the target variable. The feature “Family_Size” also has a high importance value, making it the second most influential feature. This suggests that the size of an individual’s family plays a critical role in the model’s predictions.

The analysis reveals that the gradient boosting regressor relies heavily on marital status and family size for making predictions, with these two features being far more important than the others. This information can be useful for feature selection, model optimization, and gaining insights into the factors that most influence the target variable in this dataset.

Section IV: Business Insights / Suggestions

This section synthesizes the findings from the predictive modeling and customer segmentation analyses to provide actionable business insights and strategic recommendations. The insights we

retrieved and analyzed from the data help to understand customer behaviors, preferences, and spending patterns, which are crucial for developing targeted marketing strategies. The suggestions offered here aim to enhance the effectiveness of these strategies, improve model accuracy, and ultimately support the company's efforts to penetrate new markets successfully. By leveraging these insights and recommendations, the automobile company can optimize its marketing efforts, tailor customer interactions, and drive sustainable business growth. Here is our insights, suggestions, and recommendations.

Business Insights from Data

Customer Segmentation and Behavior

- **Distinct Customer Clusters**: The K-Means clustering analysis revealed four distinct customer segments with unique characteristics. For example, Cluster 0 consists of a balanced gender mix with high spending scores, while Cluster 1 is predominantly married individuals with higher ages and lower spending scores. Understanding these clusters helps in tailoring marketing strategies to the specific needs and preferences of each segment.
- **Spending Score Predictors**: The Random Forest Regressor identified key predictors of spending scores, such as age, profession, and family size. These insights can guide targeted marketing efforts by focusing on the most influential factors driving customer spending.
- **Model Performance**: The Gradient Boosting Classifier emerged as the most reliable model for customer segmentation prediction, with a cross-validation accuracy of 0.534. Although its performance is better than other models, the accuracy is still moderate, indicating room for improvement.

Model Comparison and Reliability

- **Predictive Modeling for Spending Scores**: The Random Forest Regressor achieved the lowest Mean Squared Error (MSE) of 0.197269, making it the most accurate model for

predicting spending scores. This model's performance can be leveraged to forecast customer spending behavior accurately.

- **Classification Models for Segmentation**: Despite the Gradient Boosting Classifier being the top performer among the classification models, the overall accuracy scores indicate that none of the models provide highly confident predictions. This insight suggests the need for further refinement of models or exploration of alternative methodologies.

Business Suggestions

Targeted Marketing Strategies

- **Personalized Campaigns**: Utilize the distinct characteristics of the identified customer segments to design personalized marketing campaigns. For instance, focus on high-spending Cluster 0 with premium offers and targeted promotions, while offering value-based promotions to the more conservative Cluster 1.
- **Influential Factors**: Leverage key predictors of spending scores identified by the Random Forest Regressor. Marketing efforts can be optimized by targeting age groups, professions, and family sizes that significantly influence spending behavior.

Model Enhancement and Data Utilization

- **Improvement of Classification Models**: The moderate accuracy scores of the classification models suggest a need for enhancement. Consider incorporating additional features, such as detailed transaction histories or psychographic data, to improve model accuracy. Additionally, exploring advanced machine learning techniques like ensemble methods or deep learning could yield better predictive performance.
- **Continuous Model Training**: Implement a continuous learning approach where models are regularly updated with new data. This practice will help maintain the relevance and accuracy of the predictive models as customer behavior and market conditions evolve.
- **Data-Driven Decision Making**: Encourage a data-driven culture within the organization. Use the insights from predictive modeling and clustering analysis to inform strategic

decisions across marketing, sales, and customer service departments. Regularly review and adjust strategies based on the latest data insights to stay aligned with customer preferences and market trends.

Customer Engagement and Experience

- **Enhanced Customer Experience**: Use segmentation insights to improve customer experience by offering personalized interactions and tailored services. For instance, develop loyalty programs that cater to the specific needs of different segments, enhancing customer satisfaction and retention.
- **Feedback Mechanisms**: Establish robust feedback mechanisms to gather customer opinions and preferences continuously. Use this data to refine customer segments and adjust marketing strategies accordingly, ensuring they remain relevant and effective.

Section V: Conclusions

This report has provided a comprehensive analysis of customer segmentation and predictive modeling aimed at enhancing the marketing strategies for an automobile company entering new markets. Through a series of sophisticated data analyses, including predictive modeling for spending scores and customer segmentation clustering, we have identified key patterns and insights that can be leveraged to optimize marketing efforts and customer engagement. The use of models such as Linear Regression, Decision Tree Regressor, Random Forest Regressor, Gradient Boosting Regressor, and Support Vector Regressor provided different levels of accuracy in predicting spending scores, with the Random Forest Regressor being the most accurate model among all fives. Additionally, the clustering analysis using K-Means revealed four distinct customer segments, each with unique demographic and behavioral characteristics.

Despite these achievements, the classification models for predicting customer segments demonstrated moderate (~50%) accuracy, suggesting that there is still room for improvement. The Gradient Boosting Classifier, while the best performer, achieved a cross-validation accuracy of only 0.534, indicating that further improvement of the models is necessary, as discussed in Section IV. Future efforts could focus on incorporating additional data features, such as detailed transaction histories and psychographic data, to enhance model accuracy. Moreover, exploring advanced machine learning techniques, continuous model training, and implementing robust feedback mechanisms will help in maintaining the relevance and accuracy of the predictive models. These steps will ensure that the marketing strategies remain aligned with evolving customer preferences and market conditions.

In conclusion, the insights and recommendations derived from this analysis provide a solid foundation for the automobile company to develop targeted marketing strategies and improve customer engagement. By addressing the areas for improvement and continuously leveraging data-driven insights, the company can enhance its market entry strategy, drive higher conversion rates, and achieve sustainable business growth. The integration of advanced analytics into the decision-making process will be instrumental in maintaining a competitive edge in the dynamic automobile industry.

Section VI: Reference

Data Source:

https://www.kaggle.com/datasets/abisheksudarshan/customer-segmentation/data?select=test.csv