CUSTOMER SEGMENTATION FOR EFFECTIVE MARKETING DECISIONS

A report submitted as part of a project for CSCI-B 565: Data Mining

Keywords:
machine learning, EDA, data wrangling and cleaning, data preprocessing, bagging, boosting, svc, knn, decision tree, hyperparameter tuning

1. Abstract:

The aim of the customer segmentation project is to divide customers into different groups based on their common characteristics and behaviour. This allows businesses to better understand the needs and preferences of each group, and tailor their marketing efforts accordingly. By analyzing customer data, which often includes various demographic and behavioural characteristics such as age, profession, and family size, businesses can identify patterns and trends among different customer groups.

Data visualization techniques can be used to explore the relationships among different customer features, and help businesses identify key characteristics that can be used to segment customers. Machine learning algorithms can then be applied to the customer data to automatically classify customers into different segments based on their shared characteristics.

Different machine learning models can be used for customer segmentation, each with its strengths and weaknesses. For example, decision tree algorithms are easy to interpret and can handle both numerical and categorical data, but they are prone to overfitting. On the other hand, clustering algorithms are less interpretable but can handle large and complex datasets.

To evaluate the performance of different machine learning models, various metrics can be used, such as accuracy, precision, recall, and F1 score. These metrics provide different perspectives on the model's performance and can be used to compare and contrast the results of different

models. For example, a model with high accuracy may not be very useful if it has low precision or recall, as it may not be able to effectively identify the target customer segments.

2. Introduction:

Customer segmentation is the process of dividing a customer base into groups or segments based on shared characteristics or traits. This allows companies to tailor their marketing efforts and strategies to better target and engage with specific groups of customers. By understanding the needs and preferences of different customer segments, companies can create more effective marketing campaigns and offers that are more likely to resonate with each segment.

Effective customer segmentation can help improve the efficiency and effectiveness of a company's marketing efforts, leading to increased sales and revenue. It allows companies to better understand their customers and their purchasing habits, which can help inform marketing strategies and decisions. For example, a company selling outdoor gear may find that its customers can be segmented into two main groups: families with young children, and adventure seekers. By understanding the differences between these two segments, the company can create targeted marketing campaigns that appeal to the specific interests and needs of each group.

Overall, customer segmentation is an important tool for effective marketing decision-making. By segmenting their customer base, companies can create targeted marketing campaigns that are more likely to be successful, ultimately leading to increased sales and revenue.

3. Method:

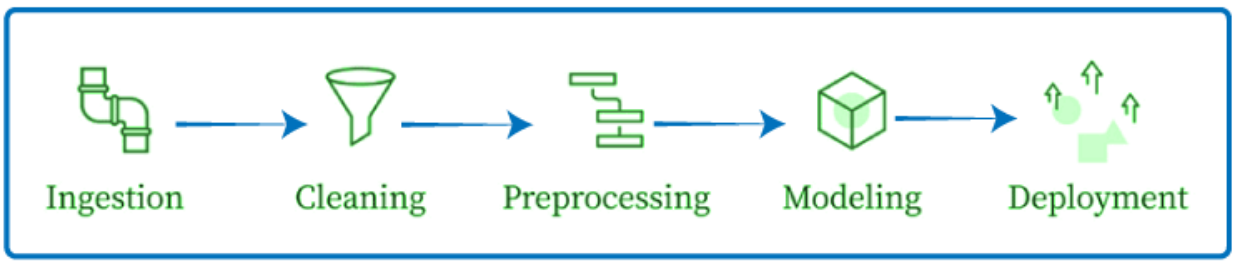


Fig 1: Process of Data Mining for this project

3.1. Data Understanding:

We obtained the dataset of market behaviour from Kaggle consisting of customers segmented into four classes. The data consists of 8068 objects and 11 features, out of which the ‘segment’ feature is dedicated to targeting prediction. Features like Gender, Ever_Married, and Graduated are categorical columns with binary values, and ‘Profession’, ‘Spending_Score’ and ‘Var_1’ are multivalued categorical columns where Var_1 is an anonymous category. The rest of the features, ‘ID’, ‘Age’, ‘Family_Size’ and ‘Work_Experience’, are numerical columns. Most of the columns contain NaN values, which were handled individually in data preprocessing. In order to analyze the values of these features in detail, we extracted the range of values of each of the features (fig1).

```
ID [458982, 467974]
Gender ['Female', 'Male']
Ever_Married ['Yes', nan, 'No']
Age [18, 89]
Graduated ['Yes', nan, 'No']
Profession ['Lawyer', 'Healthcare', 'Artist', nan, 'Homemaker', 'Doctor', 'Marketing', 'Engineer', 'Entertainment', 'Executive']
Work_Experience [0.0, 14.0, nan]
Spending_Score ['High', 'Low', 'Average']
Family_Size [1.0, 9.0, nan]
Var_1 [nan, 'Cat_6', 'Cat_5', 'Cat_7', 'Cat_2', 'Cat_4', 'Cat_1', 'Cat_3']
Segmentation ['A', 'C', 'D', 'B']
```

Fig2: Range of values of each feature

3.2. Exploratory Data Analysis:

Various Data mining techniques were kept in mind while doing the Exploratory Data Analysis part.

3.2.1: Exploring the balance of the dataset

For an effective data processing and machine learning model to learn it was essential to get to know if the dataset was balanced. That is if there were an equal number of data points in the segmentation. Looking at figure 3 we understand the data here is fairly balanced. So no additional steps like data augmentation or deleting the data to make it balanced would be needed.

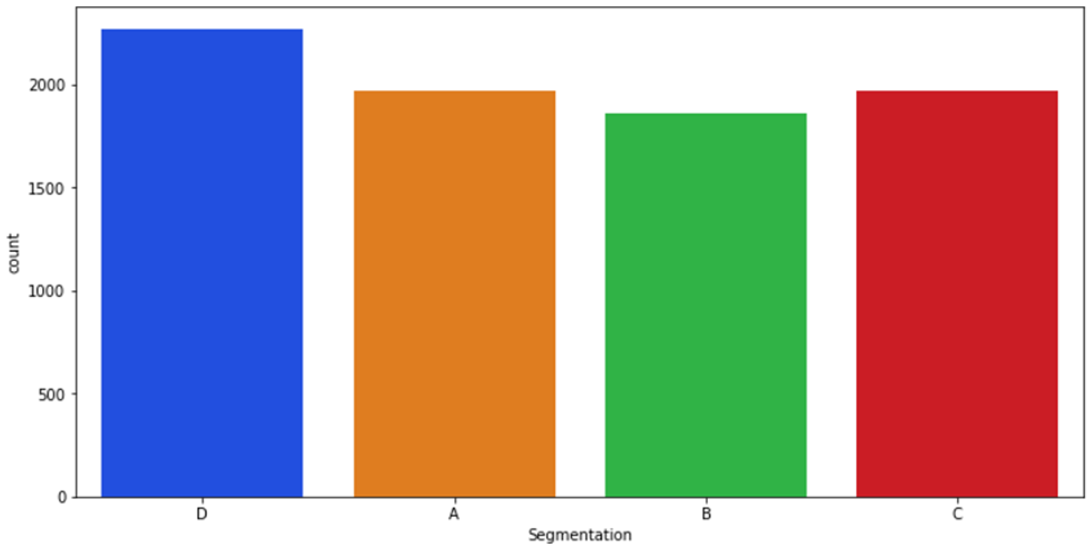


Fig 3: Distribution of Segmentation: Target feature

3.2.2: Exploring the features with respect to target feature:

3.2.2.1 Variance of Gender feature with the customer segmentation feature:

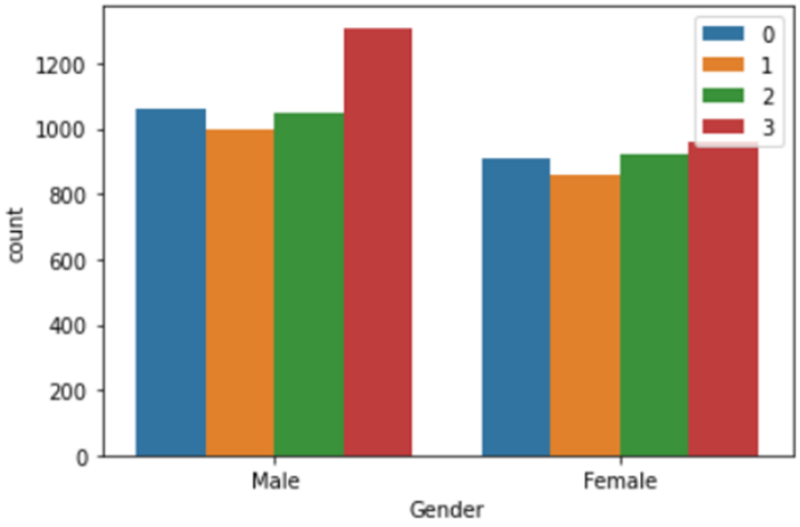


Fig 4: Distribution of Gender vs Segmentation classes

We understand from here that there are a total of 4417 males and 3651 females in the entire dataset. We can see a balanced behaviour of the data inside these two classes. There are over 1200 Males in segment 3 and the maximum number of Females can also be seen in Segment 3.

The conclusion from the graph:

- The company can develop a strategy to increase the number of females in each segment as compared to the males.
- One example strategy could be the company can make more Male oriented strategies for segment 3 and observe if the women count also increases as the difference between the two classes right now seems to be very less.

3.2.2.2 Variance of Marriage with Segmentation:

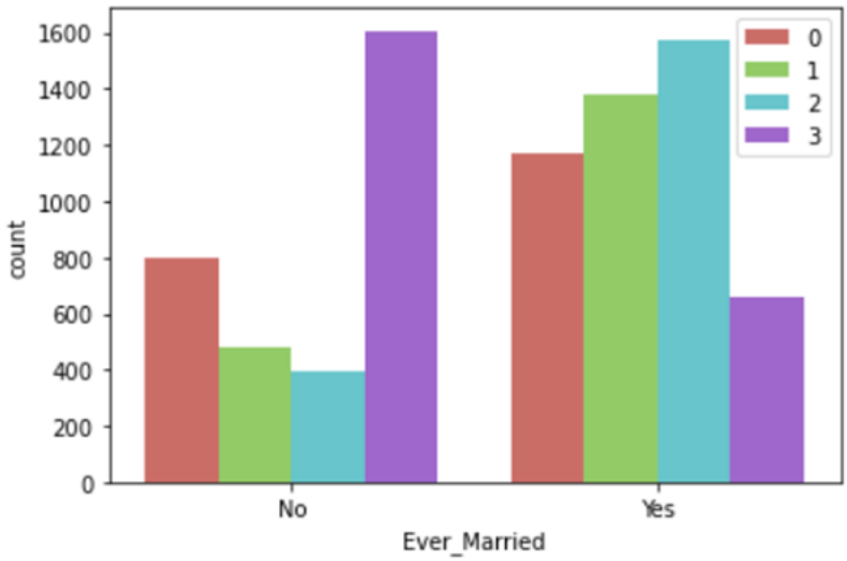


Fig 5: Distribution of Married Status vs Segmentation classes

The results here are surprisingly clear and interesting for targeting segment type 3 of the company they have to target the Non Married people. For Segment 2, people who are married are working well. Similar is the case from segment 2 for segment 0 more focus can be given to the married section.

The conclusion from the graph:

- There are some segments that have a heavy inclination towards one of the classes of marriage.
- Strategies can be developed in marketing or sales departments to give more importance to these individualities and inclinations.

3.2.2.3 Variance of Graduated people with Segmentation:

Referring to figure 6, we can say that there is a good statistic displayed when it comes to the variance of binary inclination of graduation of people and the segments they belong to. There are 4968 “Yes”, graduated labels and 3022 “No” graduated labels.

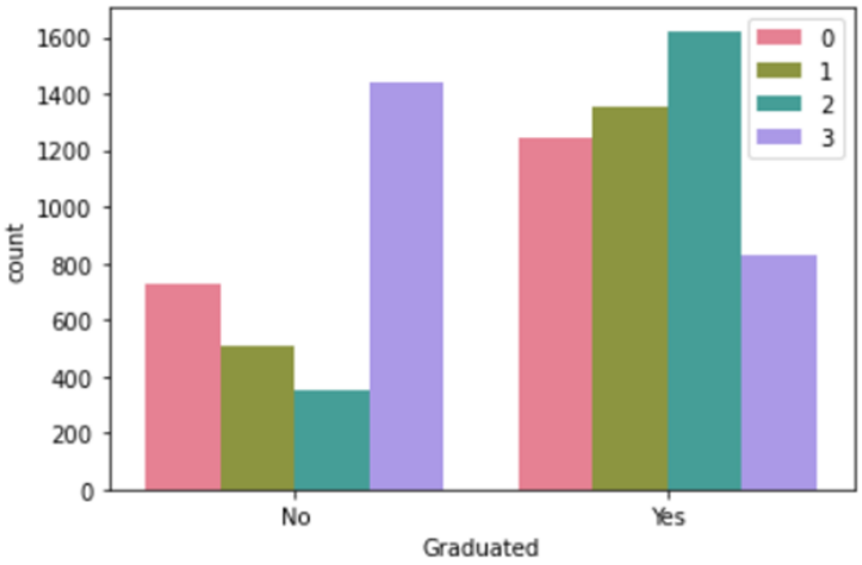


Fig 6: Distribution of Married Status vs Segmentation classes

Conclusion from the graph:

- We infer that most of the 3rd segmented category contains people who have not graduated.
- Special strategies could be made towards the non-graduated class for products of segment 3.
- Target campaigns can be made towards segments 2 and 1 for the graduated class as their inclination can be seen more in those particular segments.

3.2.2.4 Variance of Profession of People with Market Segmentation:

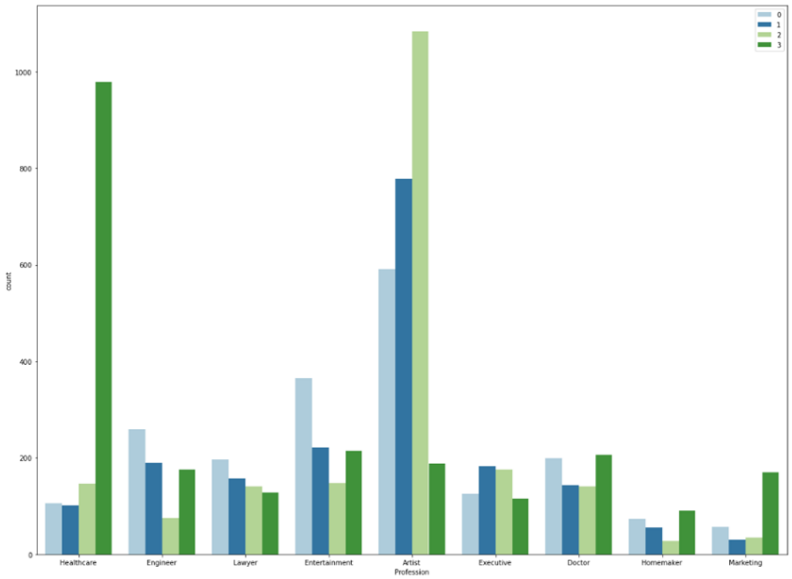


Fig 6: Distribution of Profession of people vs Segmentation classes

Referring to figure number 6 on the x-axis we see the number of different professions embedded in the professional class. There are 9 different types of professions in the feature profession ranging from Healthcare to Marketing. This graph provides a great insight into how different professions can be specifically targeted and how they affect the segmentation. The values also give us an understanding of the magnitude of the effect that it can have on the segmentation.

The conclusion from the graph:

- Specific target strategies can be made for specific classes like healthcare in segment 3 and artists in segment 2. The most important feature of those two professions can be seen very clearly.
- Understanding is developed like homemaker has a very less presence in the overall profession feature so significant strategies can be developed for increasing the number population in that class or in the homemaker sub-feature the company can target particularly on segment 3.
- Another interesting conclusion which can be drawn from the graph is that for professions like Artist the individually the in all the segments the number of this class exceeds all the other classes. So it says something about this company's products being more liked by artists in general. The company in that case could keep up with engaging more artists with all the segments as that could be that class is one of the biggest revenue-generating classes in the market.

3.2.2.5 Variance of Spending Score with Market Segmentation:

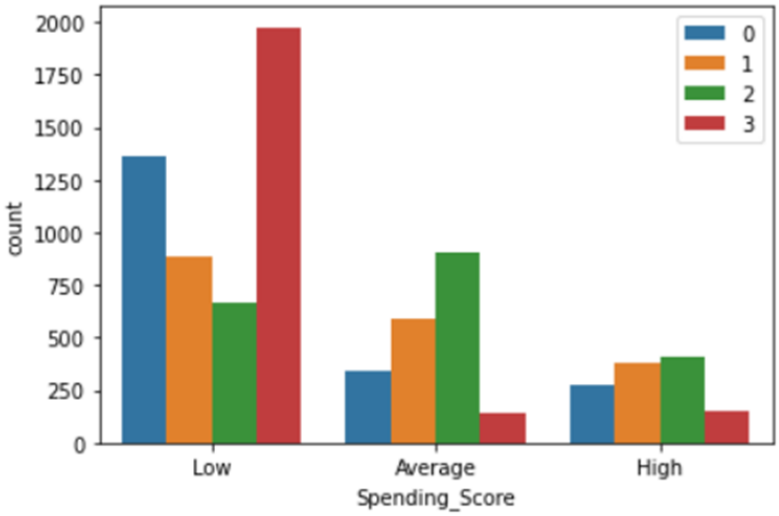


Fig 7: Distribution of Profession of people vs Segmentation classes

For any business the spending categorization of their customers is essential. This graph gives a detailed overview of how the strategies can be segmented based on the magnitude of people in each class.

Conclusions from the graph:

- Segment 3 contains the most number of low category of spending_score people
- Average spending customers can be targeted at most in segment 2.
- In the higher range the company can decide the policy of investing least in segment number 3.

3.2.2.6 Variance of Family Size with Market Segmentation:

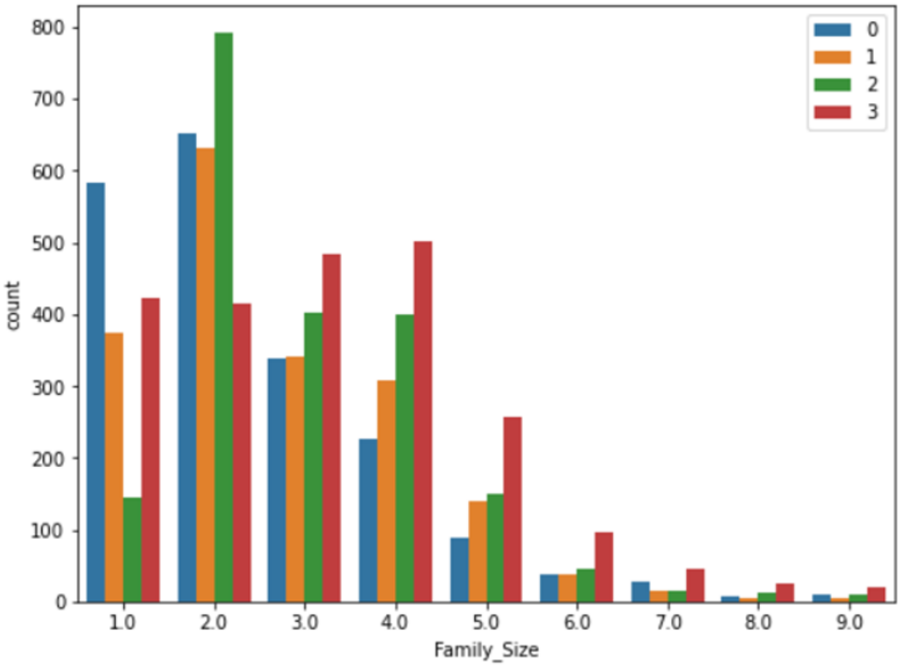


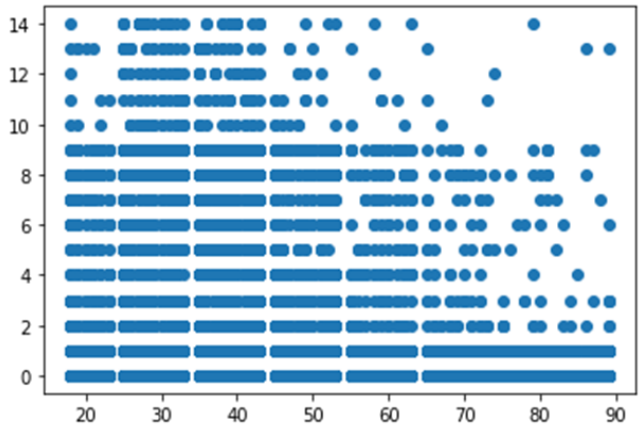
Fig 8: Distribution of Family Size vs Segmentation classes

This graph gives us an idea about targeting the optimum family sizes when it comes to segmented data targeting. It gives a general understanding that the families range from size 1 to 9 for the company's products. It can be clearly seen that as the family size increases its appearance in the dataset decreases. This plot gives us an understanding that for the strategic development of familial marketing or sales dimensions.

Conclusions from the graph:

- Strategy can be developed where most revenue can be targeted from families with size two.
- The company can make substantial efforts in improving more upon families with sizes 1,3, 4 and 5.
- The company may set up a research project for understanding the needs for their products with family sizes greater than 6. That is something the company could allocate resources to based on the magnitude of the histogram here.

3.2.2.7 Understanding some useful correlations between the important features:



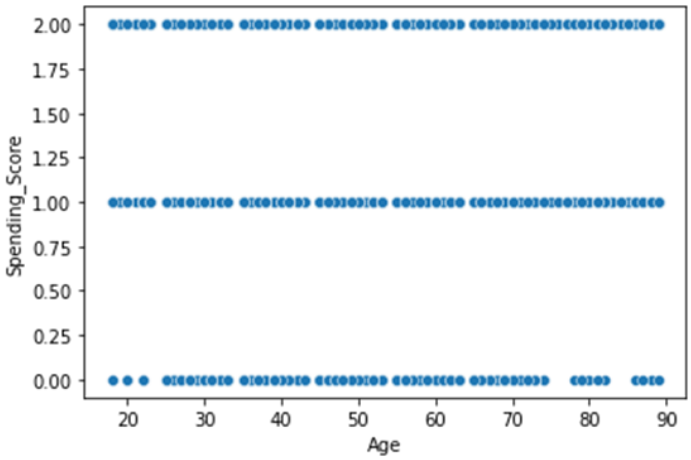


Fig 9: Age vs Work Experience Correlation

Fig 10: Age vs Spending Score Correlation

Referring to figure 9, We can see here as Age Increases the work experience increases but its density decreases. Referring to figure number 10, the Spending_Score population is really dense for the average and high types as age increases. However for Low ones the population is really sparse at an early age and at a later age.

3.2.2.8 Understanding the distribution of Categorical features:

Figure 11, signifies the distribution of the categorical features in the entire dataset.

- From the figure it can be inferred that around 55 % of the recorded data population is male and 45 % is female.
- Out of which 58% people are married and the rest are not. From the entire data, we have a good understanding that 62% of the people are graduated and around 38 % have not.
- A good segmentation has been deduced with the Spending score category where we can see out of the total population of the data around 60 % of the data is having a small spending score.
- Around 24 % of the population comes in the average spending category and around 15% of the people fall under the high spending category.
- These graphs are really helpful to make real-time strategic decisions to improve sales or in general the production in that particular segment.

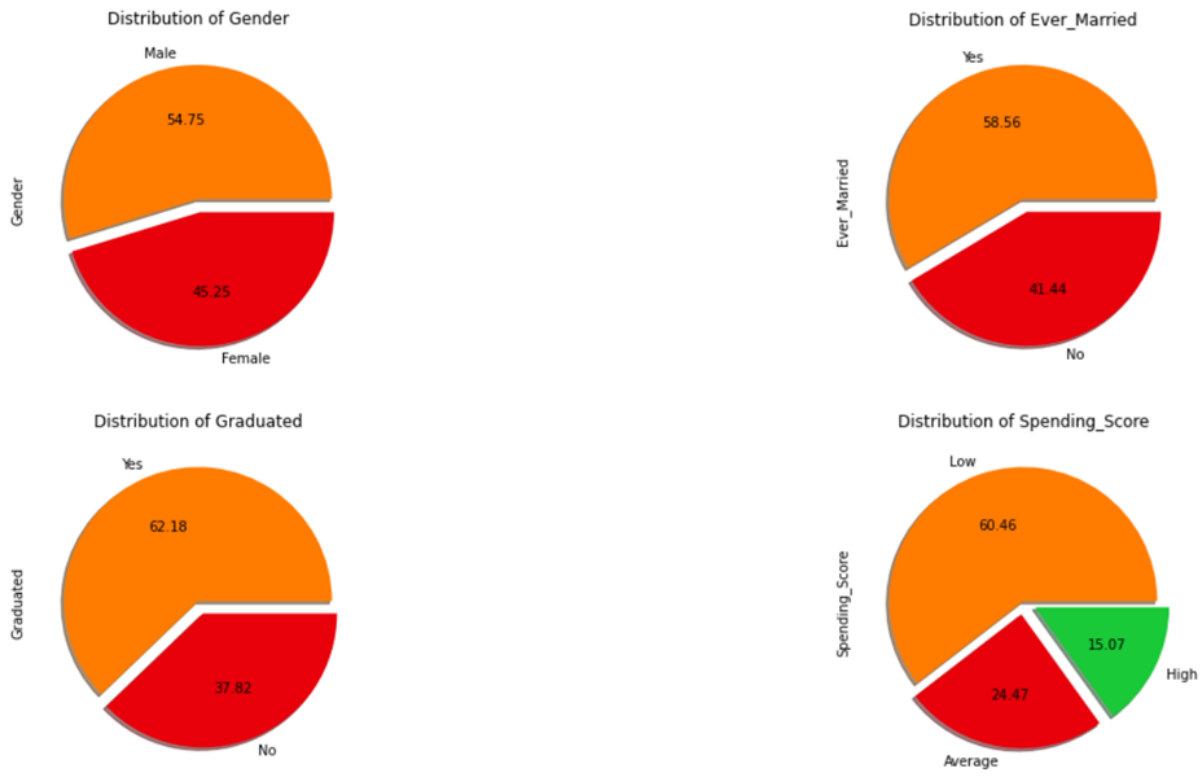


Fig 11: Distribution of categorical features

3.3. Data Preprocessing:

The main steps in preprocessing are handling values and duplicate objects. Since the dataset does not contain any duplicate rows, we focus mainly on the Nan values. We identified that 1371/8068 rows contain at least one Nan value. Since this accounts for 17% of the dataset, dropping these rows is not feasible. However, 2% of the dataset (162 rows) contains two or more Nan values and therefore was dropped before further preprocessing.

Since the feature ID has no contribution to the customer's behaviour, this feature was dropped entirely. Also, since Nan rows of Profession and Var_1 only contributed to 137 rows, these rows were lopped. The nan values of the 'Family_Size' feature were filled based on the mean value of the Family_Size of the corresponding Age since the family size is logically correlated with age. Similarly, 'Ever_Married' Nan values were filled based on the mode value of the corresponding Age and the Graduated feature was filled based on Profession.

For work experience, the KNN classification model was used to predict the Nan values. Before we proceed further, all the categorical features must be encoded to numerical values to convert them into machine-readable forms. This process was carried out using label_encoder and

get_dummies. The label encoder assigns numerical values from 0 to n-1 to each of the unique values of the feature, where n is the total number of unique values. The get_dummies, however, create a separate binary-valued (0, 1) column for each of the different values of the feature(fig2. and fig3). Both these methods were implemented and compared against each other with the model accuracies. This encoded data was then used to predict the work experience. Since the values of work experience range from 0-14, each of them can be treated as a separate class, and nan values can be predicted using classification models since regression models predict output-of-rangealues. At first, the data was split into train and test based on the Nan values of the work experience. The KNN model uses k nearest neighbors with similar features to train on work experience. Also, using gridSearchCV, hyperparameter tuning was performed on the nearest neighbor's parameter to identify the best k value(73). Finally, the model was trained with this best value and was used to predict the Nan values of work experience.

At last, the data was normalized using a standard scalar before training the model with SVC.

ID	Gender	Ever_Married	Age	Graduated	Profession	Work_Experience	Spending_Score	Family_Size	Var_1	Segmentation
462809	Male	No	22	No	Healthcare	1.00	Low	4.00	Cat_4	D
462643	Female	Yes	38	Yes	Engineer	NaN	Average	3.00	Cat_4	A
466315	Female	Yes	67	Yes	Engineer	1.00	Low	1.00	Cat_6	B
461735	Male	Yes	67	Yes	Lawyer	0.00	High	2.00	Cat_6	B
462669	Female	Yes	40	Yes	Entertainment	NaN	High	6.00	Cat_6	A

Fig 12: Before Label, Encoding using get_dummies

Gender	Ever_Married	Age	Graduated	Work_Experience	Family_Size	Segmentation	Spending_Score_High	Spending_Score_Low	Profession_Doctor	Pro
1	0	22	0	1.00	4.00	3	0	1	0	
0	1	67	1	1.00	1.00	1	0	1	0	
1	1	67	1	0.00	2.00	1	1	0	0	
1	1	56	0	0.00	2.00	2	0	0	0	
1	0	32	1	1.00	3.00	2	0	1	0	
0	0	33	1	1.00	3.00	3	0	1	0	
0	1	61	1	0.00	3.00	3	0	1	0	
0	1	55	1	1.00	4.00	2	0	0	0	
0	0	26	1	1.00	3.00	0	0	1	0	

Fig 13: After Label Encoding using get_dummies

3.3. Modeling:

We use various approaches such as

- classical machine learning model (SVC, decision tree)
- ensemble models (XGBClassifier)
- Bagging/Boosting for hyperparameter tuning.

Classical machine learning models, such as support vector machines (SVMs) and decision trees, are supervised learning algorithms that are commonly used in classification tasks. SVMs

are a type of algorithm that uses training data to find the best hyperplane that separates different classes of data points, while decision trees are algorithms that use a tree-like structure to make predictions based on the characteristics of the data.

Ensemble models, such as XGBoost (Extreme Gradient Boosting), are a type of machine learning model that combines the predictions of multiple individual models to make more accurate predictions. In the case of XGBoost, this is done by training multiple decision trees and combining their predictions using gradient boosting.

Bagging and boosting are two techniques that are often used in ensemble learning to improve the performance of individual models. Bagging, or bootstrap aggregating, trains multiple models on different subsets of the training data and combines their predictions to reduce overfitting and improve the generalization performance of the model. Boosting, on the other hand, trains multiple models in a sequential manner, with each model focusing on the mistakes made by the previous model in an effort to improve overall performance.

Hyperparameter tuning is the process of selecting the best set of hyperparameters for a machine-learning model in order to maximize its performance on a given task. This can be done using a variety of techniques, including grid search, random search, and Bayesian optimization. In the context of bagging and boosting, hyperparameter tuning is often used to optimize the parameters of the individual models that are combined in the ensemble.

4. Conclusion and future work:

The low accuracy of the model on this dataset may be due to a number of factors, including the presence of many missing or incomplete values (i.e. "Nan" values) and a relatively small training dataset.

To improve the accuracy of the model, a number of different approaches can be tried. Bagging with decision trees as the base model can help reduce overfitting and improve the generalization performance of the model by training multiple decision trees on different subsets of the training data and combining their predictions.

Using grid search with cross-validation (CV) for hyperparameter calculation can help identify the best combination of hyperparameters for the model, which can improve its accuracy and precision. Feature scaling with SVC using a StandardScalar model can also help improve the accuracy of the model by normalizing the features of the data so that they are on a similar scale.

In addition, further hyperparameter tuning can be performed to try and improve the accuracy and precision of the model. Using neural networks, which are a type of machine learning model that can learn complex relationships between the input data and the output predictions, can also potentially improve the accuracy and yield additional insights from the data.