Customer Segmentation Classification

1. Introduction:

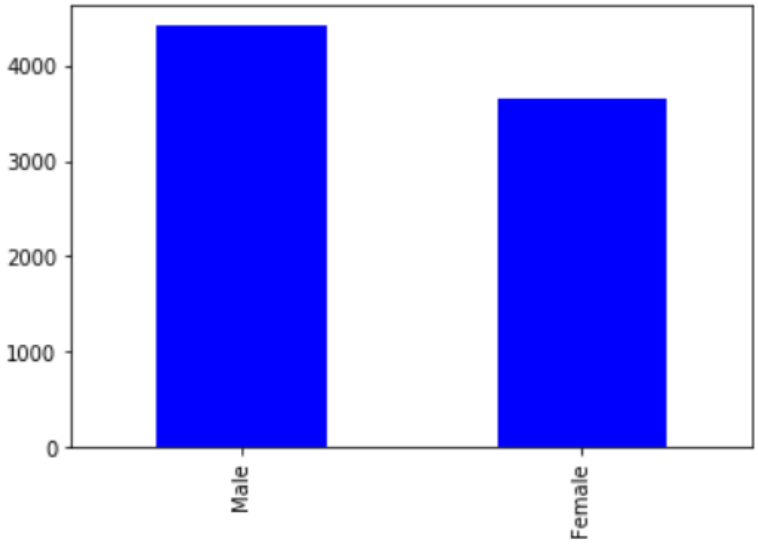
The sales team in their current market has split all customers into four categories (A, B, C, D). After that, they did segmented outreach and networking with a particular consumer segment. This technique has proven to be extremely effective for them. They plan to proceed with the same approach in the emerging markets, and they have established 2627 new potential customers. This is a multiclassification use case where We need to estimate the appropriate community of new customers in this analysis.

2. Analysis:

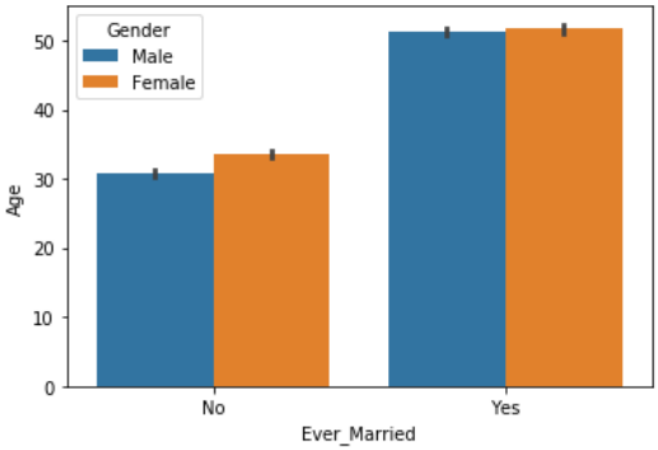
2.1 Data Exploration

While performing the exploratory data analysis on the dataset, we could see the below insights:

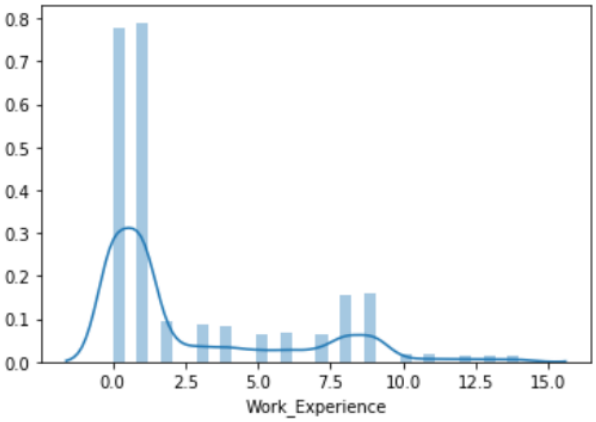
The below plots tells us that there are around 4500 males customer and around 3500 female customers.



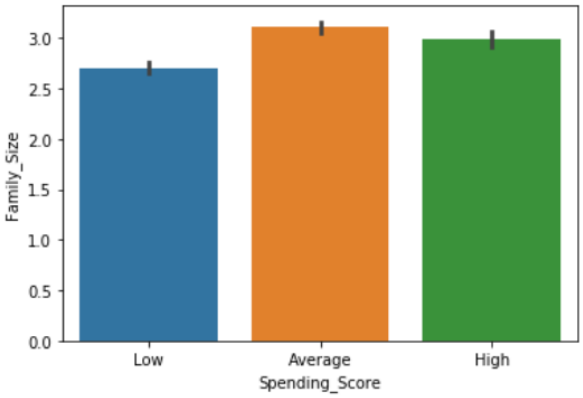
Based on the gender, we could see from the below plot that the customers of age group of 50 are married than the age group of 30 who are not married.



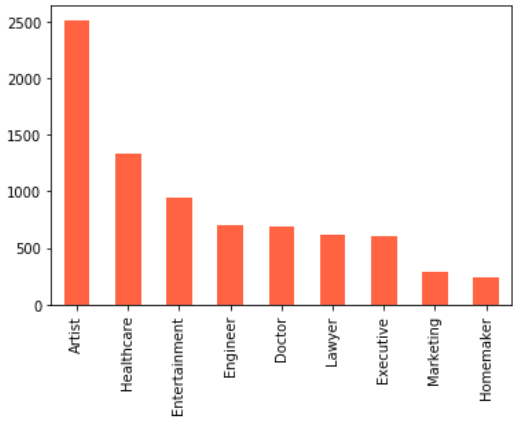
The below plot says most of the customers have around 0 to 2.5 years of work experience.



The below plot tells us that the lesser the family size, the lesser the spending score. The higher the family size, the spending score is average.

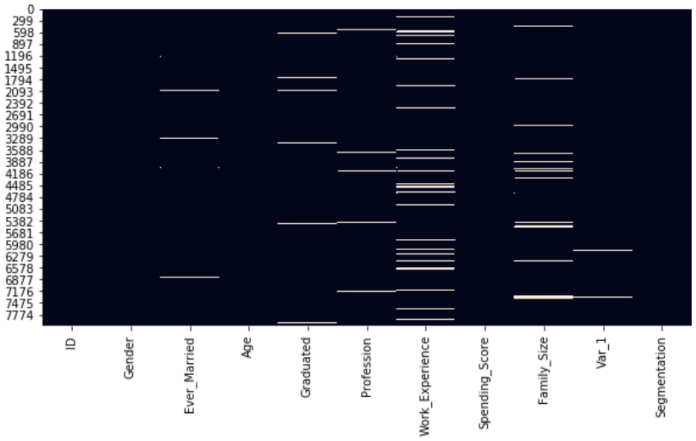


By looking at the below plot, we can say that there are more number of customers whose profession is “Artist” than any other.



2.2 Data Cleaning

Before building the model, first we need to check with the null values in our dataset and remove or replace the null values according to the features.

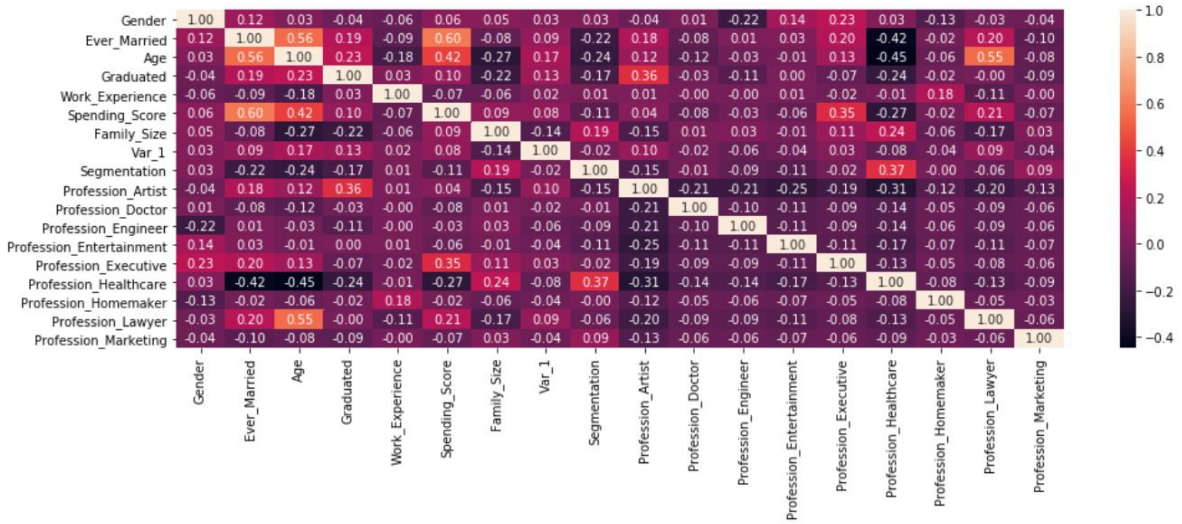


The white lines for each column indicate the null values. So according to the features, we need to replace null values ore remove them. After replacing the null values with the mean for “Ever_Married”, “Graduated”, “Family_Size”, and “Work_Experience” variables, we drop the null values for “Profession” and “Var_1”. The next step would be feature selection where we select the important features required for our model building. Here, we will be dropping “ID” variable which is of no use for our model.

Before building the model, we convert all the categorical variable into numerical ones by 0’s and 1’s or using one hot encoding.

2.3 Data correlation and features selection

When we look at the co-relation between each variables, we can see that “Ever_Married” and “Family_size” of around 0.6.



Test dataset:

We need to apply the same changes that we did for the training dataset including replacing the null values and feature selection.

2.4 Model Building:

We will be assigning the values for X_train, y_train from training dataset and X_test, y_test from testing dataset.

2.5 Machine Learning Algorithms

We will be using two machine learning algorithms – Random Forest Classifiers and Support Vector machine.

2.5.1 Random Forest Classifier

Random forest is a learning algorithm that is supervised. It produces a "forest" out of an ensemble of decision trees, which are normally trained using the "bagging" process. The bagging method's basic premise is that combining different learning models improves the overall outcome.
Random forest is also a valuable algorithm because the default hyperparameters it utilises often make accurate predictions. Understanding the hyperparameters is easy, and there aren't many of them to begin with.

By using random forest classifier, the test dataset could achieve an accuracy of around **32%** which is lower compared to other models.

2.5.2 Support Vector Classifier

Libsvm is used in this implementation. The fit time scales at least quadratically with the number of samples, and above tens of thousands of samples, it may be impractical. Consider using LinearSVC or SGDClassifier instead, probably after a Nystroem transformer, for large datasets. A one-to-one scheme is used to manage multiclass support.
By using support vector classifier, the test dataset could achieve an accuracy of around **33%** which is lower compared to other models which is almost the same as random forest classifier.

3 Conclusion

By analysing the dataset, we could see that customers who have less than 2.5 years of experience have been targeted more. The company should target the customers who have more than 2.5 years of experience as well to increase the company profits.

Currently, the company is targeting the customers whose spending score is low and people who have less spending score tend to spend less compared to people with more scores. So, the company needs to target people with higher spending score so that people spend more money which would eventually increase companies profits.

When we went through the entire phase in this analysis, starting with identifying the business goal, collecting data, exploring features and distributions, treating data, recognising correlations, choosing appropriate features, data modelling, and providing two separate algorithms with metrics to choose the best to predict Customer Segmentation. What will assist the organisation in implementing the best marketing strategy for each of them in order to improve market share and sales. Support vector classifier was chosen because it is not the most accurate model, but it does have drawbacks and does not have a high accuracy. With more data about customers, we might create a more robust model; this is something to study further with the business team and the data engineer to see if more applicable features are usable.