Market Customer Segmentation with SVM,Random Forest and K-Means

Team G:

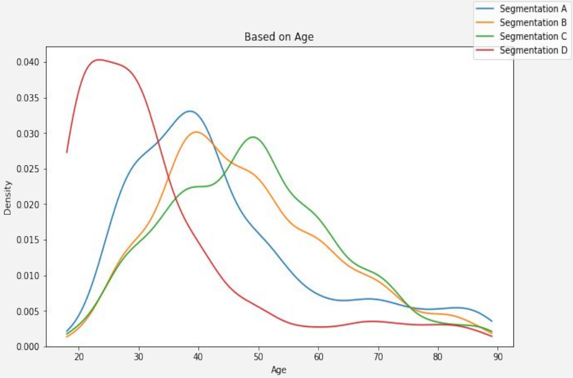
1. Alex Juffras
2. Amey Basangoudar
3. Riddhi Narayan

Customer segmentation is the practice of classifying consumers into groups based on shared traits so that businesses may effectively and appropriately sell to each group. Customers are divided into groups based on common behaviors and customs.

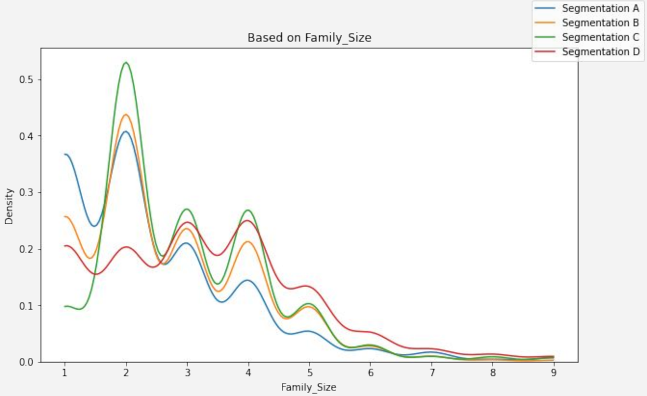
Dataset features

- Features
    - Gender (Male, Female)
    - Marital status (Yes, No)
    - Age (18-89)
    - Graduated (Yes, No)
    - Profession (Artist , Healthcare etc.)
    - Work Experience in Years (0-14)
    - Spending Score(Low, Med, High)
    - Family size(1-9)
    - Var1(Seven Categories)
    - Segmentation (A,B,C,D) -present in train and not test

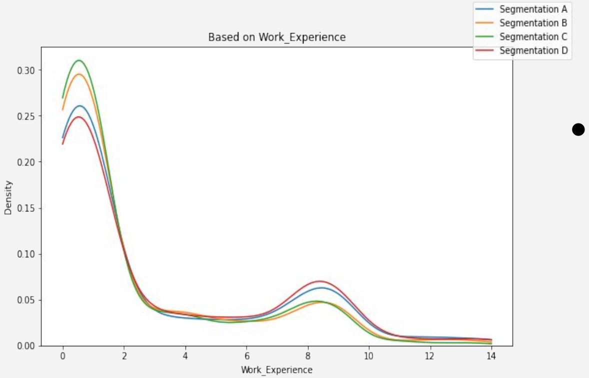
Effects of Numerical parameters on Segmentation



- People 30 years old or younger belong to segment D between 30 and 45, or older than 70 years old belong to segment A. People 45 to 70 years old belong to segment C

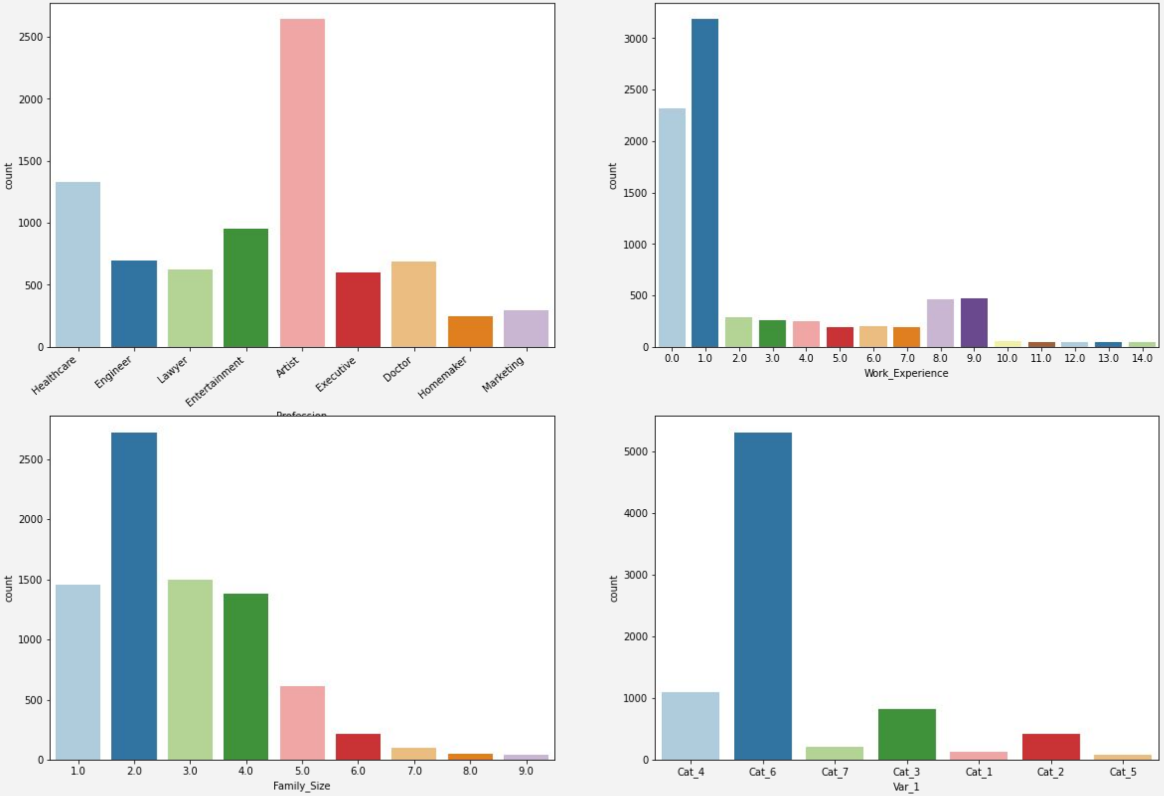


- Single people tend to be in segment A, while family sizes from 2-4 are more are in C. Those with families larger than four people were in segment D.



- Those with less than two years of work belong to segments C and B, while six to eleven years are in A and D.

Distributions

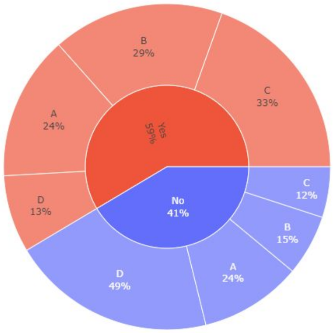


- Based on these graphs, our data is representative more of Artists, people with 0 to 1 year off work experience, and those with family sizes of 2.

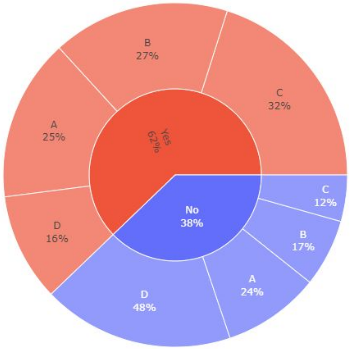
- The over-representation of work experience makes the segmentation based on age and work experience not line up.

Effects of Categorical parameters on Segmentation

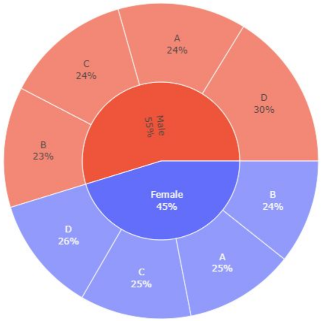
Affect of Ever_Married on Customer Segmentation



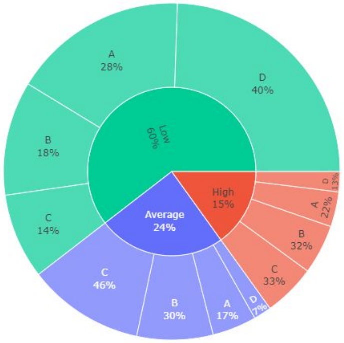
Affect of Graduated on Customer Segmentation

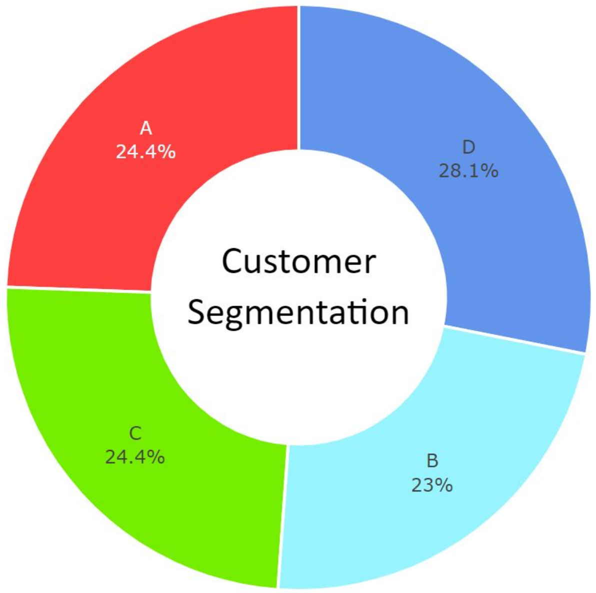


Affect of Gender on Customer Segmentation



Affect of Spending_Score on Customer Segmentation





MODELLING USING SUPERVISED TECHNIQUES

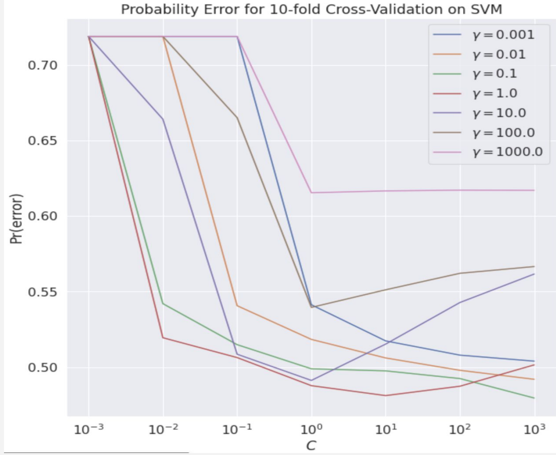
1. SUPPORT VECTOR MACHINES (SVMS):

Hyperparameters:

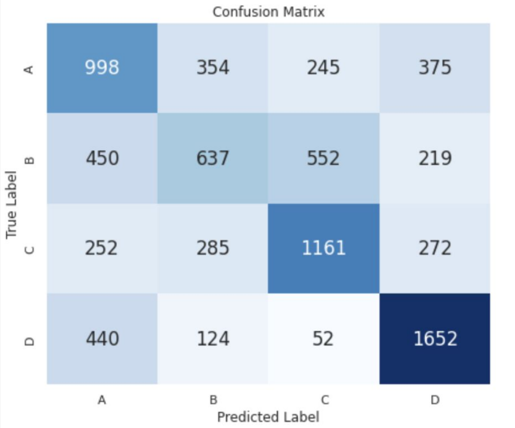
- Regularization parameter **C**
- Spread of the Kernel ***γ***

GridSearchCV Results:

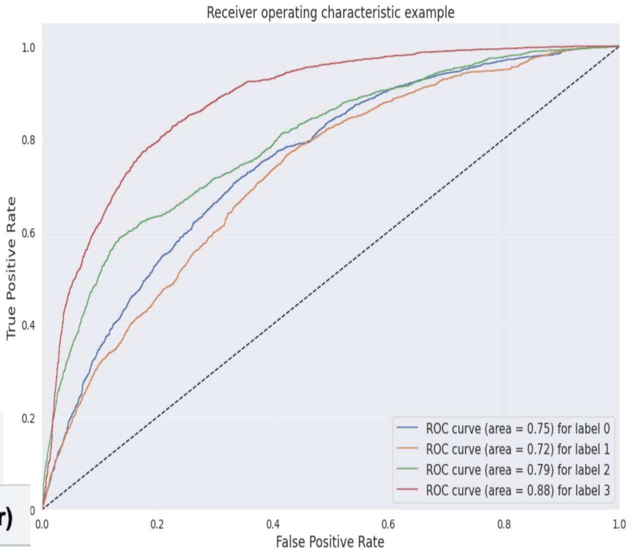
KERNEL	C	Gamma(γ)	Pr(Error)
RBF	1000	0.1	0.48



Classifier Results:



	A	B	C	D
A	998	354	245	375
B	450	637	552	219
C	252	285	1161	272
D	440	124	52	1652



PRECISION	RECALL	F1-SCORE	Pr(Error)
0.54	0.55	0.55	0.48

MODELLING USING SUPERVISED TECHNIQUES

2. RANDOM FOREST CLASSIFIER:

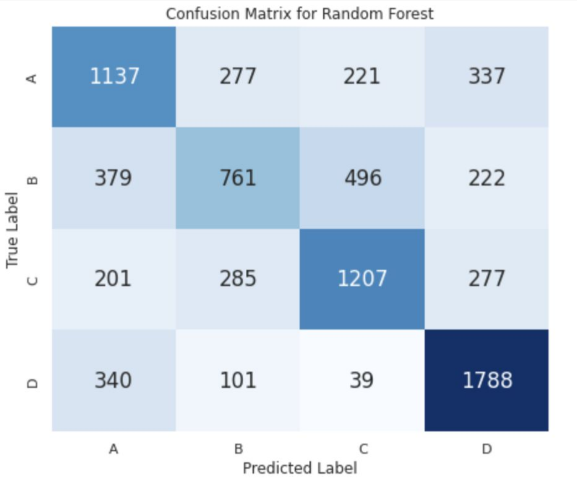
Hyperparameters:

- **n_estimators :** Number of Trees in the Forest
- **max_features :** Maximum number of features considered to split a node
- **Max_depth :** maximum levels in each decision tree
- **Criterion :** Loss Function (Gini Impurity or Entropy)

GridSearchCV Results:

n_estimators	max_features	max_depth	criterion
200	auto	8	Gini Impurity

Classifier Results:



PRECISION	RECALL	F1-SCORE	Pr(Error)
0.60	0.61	0.60	0.45

Data Preprocessing

Handling missing values:

-Rows that have 3 or more null Null Values have been dropped.

-Other remaining Nulls are imputed using KNN Imputation

Encoding:

-One hot encoding is done for the binary categorical variables such as Gender and ‘Ever_Married’

-Ordinal encoding is performed for features that have more than 2 values.

KNN Imputer: Offered via scikit-learn

KNN Imputer helps to impute missing values present in the observations by finding the nearest neighbors with the Euclidean distance matrix.

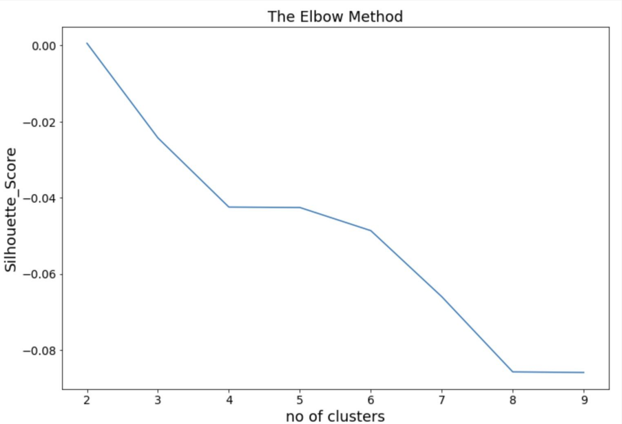
Min Max Scaler

Transform features by scaling each feature to a given range.

This estimator scales and translates each feature individually such that it is in the given range on the training set, e.g. between zero and one.

K-Means Clustering

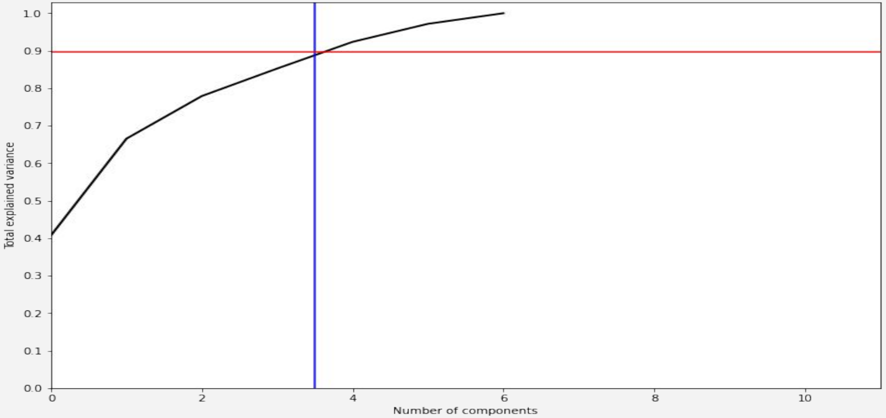
Choosing k: number of clusters:



Silhouette score:

The Silhouette Coefficient is calculated using the mean intra-cluster distance (a) and the mean nearest-cluster distance (b) for each sample

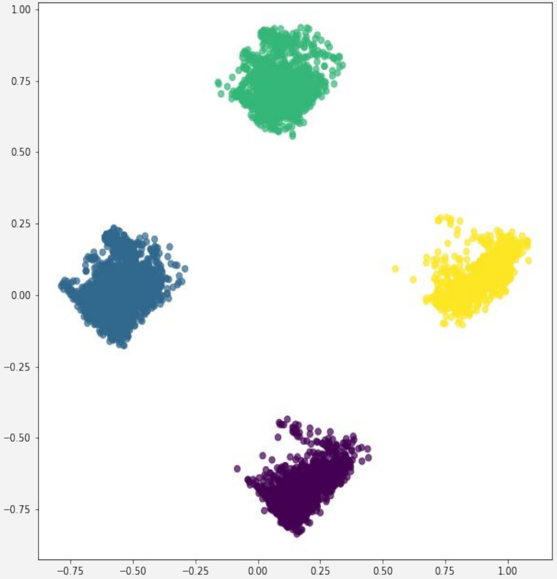
PCA Analysis



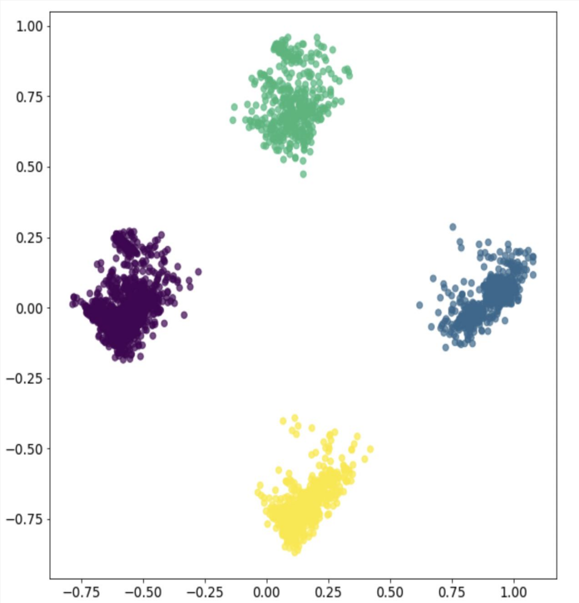
Dimensionality Reduction:

Eliminates noisy data dimensions and thus and improves accuracy in classification and clustering, in addition to reduced computational cost

K-Means Clustering and Conclusion



Clustering on train set



Clustering on test set

Solutions:

1.Feature Engineering

2. Increasing the size of our dataset

3. Using a dataset where the features have a substantial impact on the target variable

Thank you!

Any questions?

References:

>Customer Segmentation Using Machine Learning Techniques
https://www.sciencegate.app/document/10.1504/ijbidm.2022.10036753

> KNN Imputer
https://medium.com/@kyawsawhtoon/a-guide-to-knn-imputation-95e2dc496e

>Xiahou, Xiancheng, and Yoshio Harada. "B2C E-Commerce Customer Churn Prediction Based on K-Means and SVM." *Journal of Theoretical and Applied Electronic Commerce Research* 17.2 (2022): 458-475.