Customer Car Category

Sarah alrashidi
Lama saeed

Table of counts

introduction

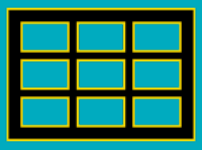
Customer Segmentation in an Automobile Company In this dataset we will try to classify.

they've deduced that the behavior of the new market is similar to their existing market. In their existing market, the sales team has classified all customers into 4 segments (A, B, C, D).

Goal

" To Know classify customer into the right segment "

Data



From Kaggle

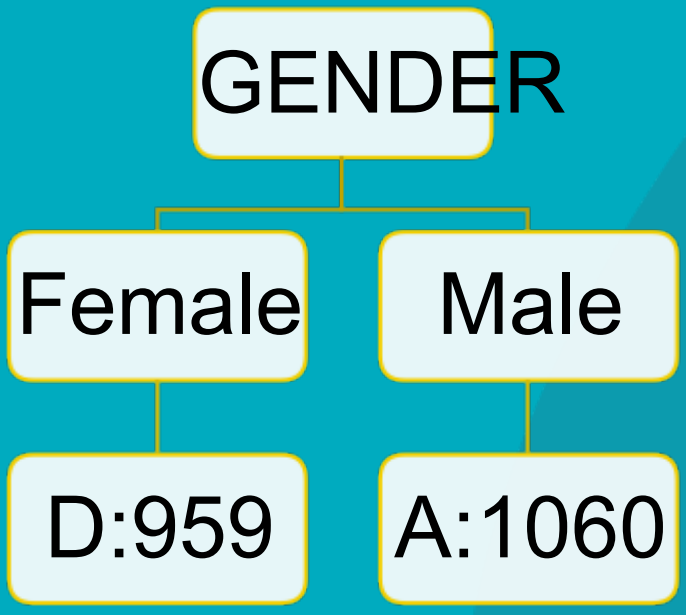
Shape data

10695

10 Features

Target :
Segmentation

Segmentation Analysis



CLEAING Data

01

Family size

- Change type for int to float
- Change null for median NUM

02

Gradate

- Change null for mode (Yes)

03

Married

- Change null for mode (Yes)

04

Work

Change null for mode (Yes)

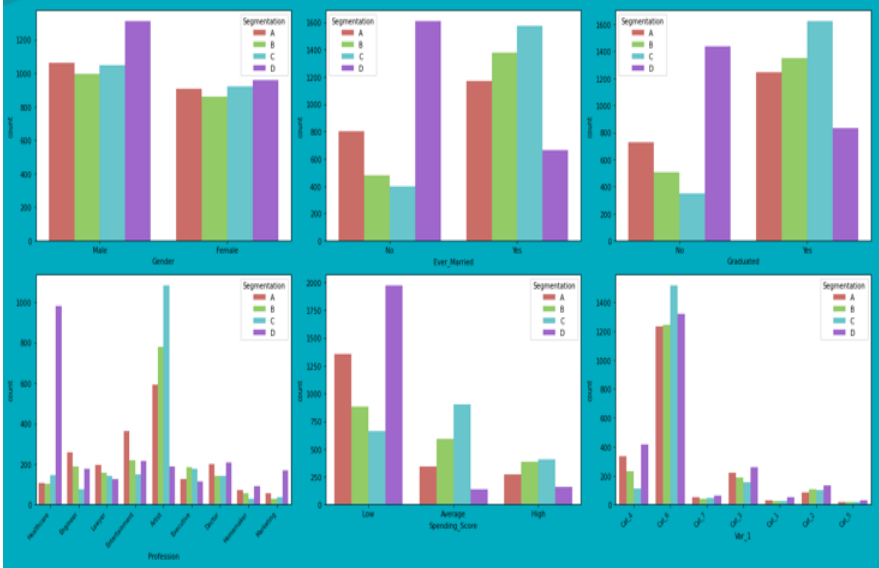
Feature Engineering

Data encoding we use it one-hot encoding .

The feature split to numerical and categorical.

	Gender	Ever_Married	Graduated	Profession	Spending_Score	Var_1	Segmentation
0	1	0	0	5	2	3	D
1	0	1	1	2	0	3	A
2	0	1	1	2	2	5	B
3	1	1	1	7	1	5	B
4	0	1	1	3	1	5	A

Multivariate Analysis



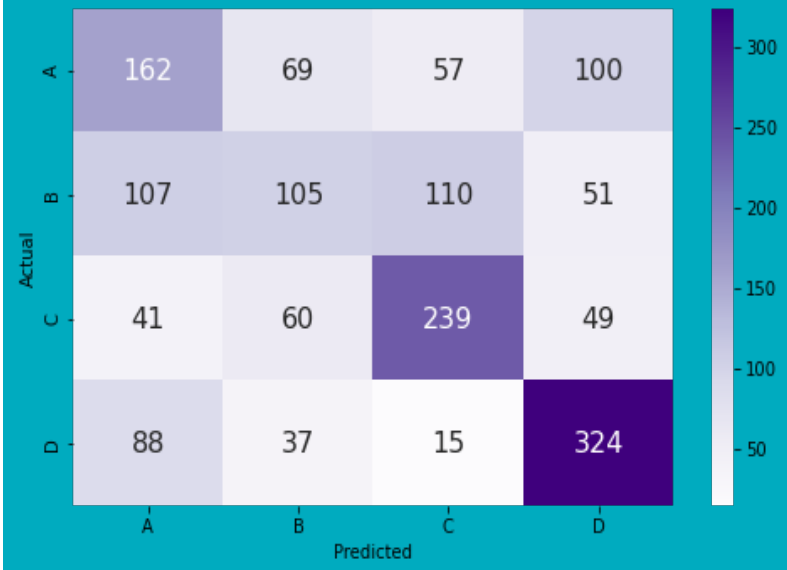
With target

nsights:
Most Gender Male in segment D, Female C & D
There are many married customers in segment C, many unmarried customers in segment D
Many customers who have graduated are in segment C, not many are in segment D
Most healthcare professions are in segment D, artists are in group C
Spending Score customers who are low mostly enter the D . segment
Customers with category 6 (cat6) are mostly included in segment C

Models

	Data	Mode	Train Score Accuracy	Test Score Accuracy	Precision	Recall	F1 Score
4	df	Gradient Boosting	0.61	0.54	0.53	0.53	0.53
0	df	Random Forest	1.0	0.51	0.49	0.5	0.49
3	df	Naive Bayes	0.48	0.48	0.45	0.47	0.45
2	df	Decision Trees	1.0	0.44	0.43	0.43	0.43
1	df	k-Nearest Neighbors	0.62	0.41	0.42	0.41	0.41
5	df	Logistic Regression	0.34	0.36	0.18	0.33	0.23
6	df	Support Vector Machine	0.28	0.29	0.07	0.25	0.11

Confusion Matrix



The best class is D in terms of Predaict and acul

Tools

1-Pandas
2-Numpy
3-Matplotlib
4-Seaborn
5-sklearn

Conclusion

The Best model on validation is the RndomForestClassifier:

- Train : 0.96
- Test : 0.47

Thank you