Customer Segments Classification

Classification task for automobile customers

- Maryam Omar
- Abdullah Bushnag

Description

An automobile company has plans to enter new markets with their existing products (P1, P2, P3, P4 and P5).

the sales team has classified all customers into 4 segments (A, B, C, D), and performed outreach on segments.

They plan to use the same strategy on new markets and have identified 2627 new potential customers

GOAL

Classify new customers into the right segment.

Data Description

8067 Rows, 12 Col

col	Description	Type
ID	Customer ID	string
Gender	Customer gender (male/female)	string
Ever_Married	Marital status	string
Age	customer age	int
Graduated	yes/no, gradtuation from college	string
Profession	customer's profession	string
Work_Experience	customer work experience in years	float
Spending_Score	a score of customer spending habbits (Low, Average, High)	string
Family_Size	number of family members	float
Var_1	anonymization feauter	string
Segmentation (target)	segment to which the customer belongs	string

STEPS

01

EDA

- Checking data balance
- Analyzing Segments

02

Data Prep

- Dealing with nulls
- Encoding
- Feature engineering
- Dropping Unnecessary Columns

03

Experiments

- All features w/ feature engineering/ 4 classes.
- merged segments
- balanced data

EDA: Segments Analysis

- all segments has almost balanced gender distribution.

- customers in segment A have married before, most are between the age of 26-45, most graduated from college, most are artists, work in the entertainment industry , or engineers, they have low spending score, most have small families

EDA: Segments Analysis

- segments B and C are very similar, they are artists, have small to average families, most are graduated and middle aged, the only difference is spending score.

- customers in segment D are mostly healthcare providers, they are young, and never been married unlike the rest of the segments.

EDA: Cleaning & Formatting

Dealing with nulls

- Fill na with mode if str
- Fill na with median if float

Encoding

- Label encoding and Ordinal Encoding
- Encoded Columns: Ever Married, spending score, graduated, gender

EDA: Cleaning & Formatting

Feature engineering

- AgeGroup
- Family Size

Dropping Unnecessary Columns

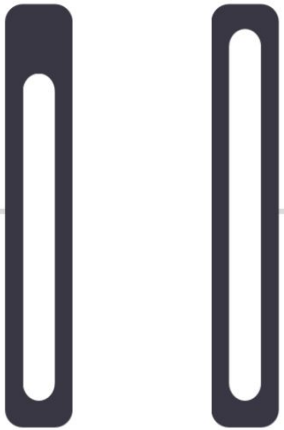
- Dropped ID column
- Dropped the Age column

Baseline Model

Logistic Regression

Numerical

Training : 0.466
Validation : 0.463



All Features

Training : 0.518
Validation : 0.518

Experiments

- All features w/ feature engineering/ 4 classes.
- merged segments
- balanced data

Results

Expr 1

Experiment	LR	KNN	DT	SVM	RF	Voting	Stacking
1 : All Features, Engineering, 4 classes	0.519 / 0.516	0.52 / 0.52	0.54 / 0.53	0.55 / 0.54	0.58 / 0.56	0.55 / 0.55	0.56 / 0.55

Merging Similar Segments

Expr 2

B

- 0.463 females, 0.537 males
- 0.738 have married before
- most are between the age of 36-55
- 0.724 graduated from college
- most are artists, work in the entertainment industry , or engineers
- low spending score
- small families
- anonymization category 4 and 6.

C

- 0.468 females, 0.532 males
- 0.796 have married before
- most are between the age of 36-55, 0.822
- graduated from college
- most are artists
- average spending score
- average sized families,
- anonymization category 4 and 6.

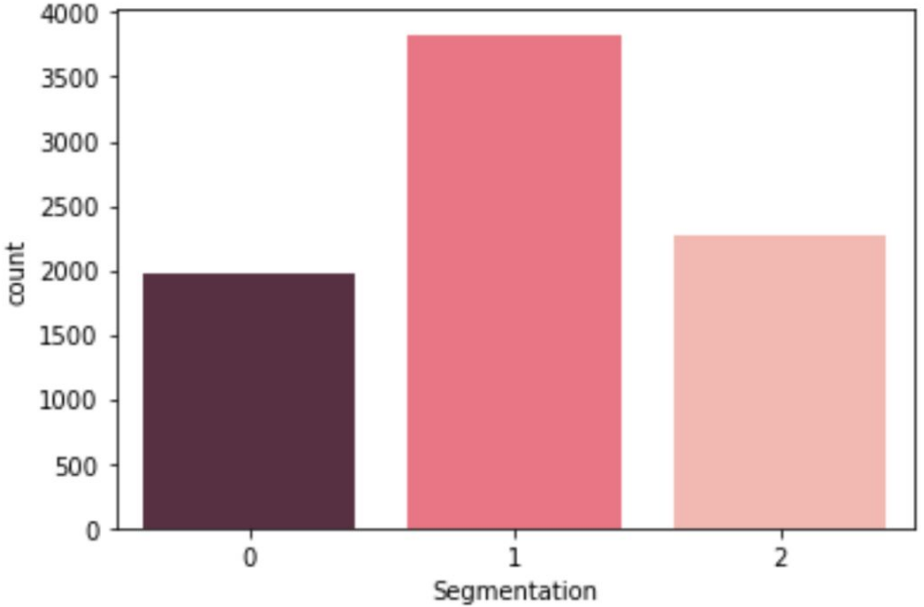
Results

Expr 2

Experiment	LR	KNN	DT	SVM	RF	Voting	Stacking
2 : with feature engineering	0.64 / 0.646	0.637 / 0.634	0.64 / 0.65	0.64 / 0.646	0.69 / 0.657	0.65 / 0.646	0.68 / 0.65

Fixing Data Imbalance

Expr 3



Fixing Data Imbalance

Expr 3

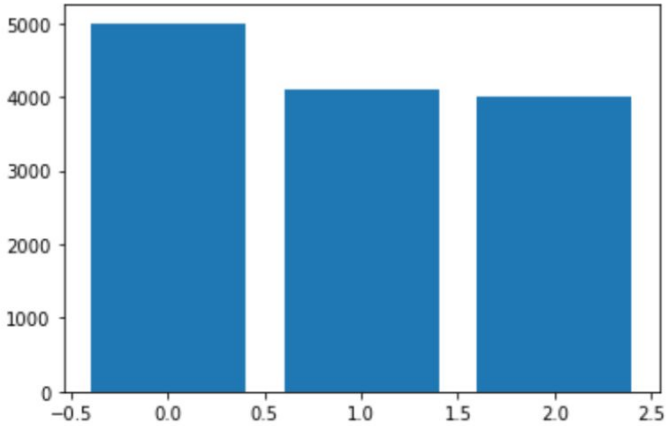
	precision	recall	f1-score	support
0	0.41	0.58	0.48	385
1	0.74	0.63	0.68	775
2	0.71	0.64	0.68	454
accuracy			0.62	1614
macro avg	0.62	0.62	0.61	1614
weighted avg	0.65	0.62	0.63	1614

Fixing Data Imbalance

Expr 3

SMOTE - OVERSAMPLING

Class=2, n=4000 (30.534%)
Class=0, n=5000 (38.168%)
Class=1, n=4100 (31.298%)



Results

Expr 3

Experiment	LR	RF
3 : Balanced Data set	0.65 /0.645	0.685 / 0.67

Conclusion

After running the mentioned models it showed that the model performed the best on validation is the RandomForestClassifier: 0.685 on train | 0.671 on val

we would suggest merging segments B & C for the sales team.