8995 CAPSTONE PROJECT PRESENTATION

Shop Customer Dataset

Jacinda Zou / U3145190

TUTORIAL GROUP – WEEK 13, 2 May 2023/16:30

Table of Contents

1. Introduction / Problem Statement

- Problem:
- To determine whether a customer is a potential value customer (spending score more than 50) based on the data of the Shop.
- The Objective:
- To help the shop owner understand their customers and make business decisions.

2. Dataset Details

- An in-depth investigation of the customers in a shop [1].
- It helps the business learn more about its customers.
- First 5 rows of Dataset

	CustomerID	Gender	Age	Annual Income ($)	Spending Score (1-100)	Profession	Work Experience	Family Size
0	1	Male	19	15000	39	Healthcare	1	4
1	2	Male	21	35000	81	Engineer	3	3
2	3	Female	20	86000	6	Engineer	1	1
3	4	Female	23	59000	77	Lawyer	0	2
4	5	Female	31	38000	40	Entertainment	2	6

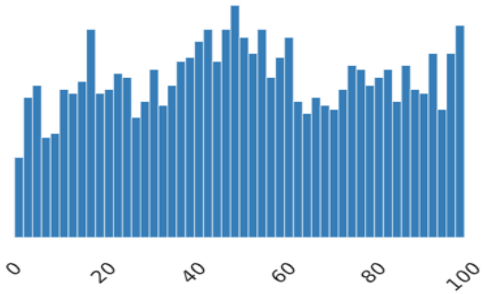
3. EDA (Exploratory Data Analysis) Outcomes

- EDA Report – Spending Score

Spending Score (1-100)

Real number (ℝ)

Distinct	101	**Minimum**	0
Distinct (%)	5.1%	**Maximum**	100
Missing	0	**Zeros**	2
Missing (%)	0.0%	**Zeros (%)**	0.1%
Infinite	0	**Negative**	0
Infinite (%)	0.0%	**Negative (%)**	0.0%
Mean	50.9625	**Memory size**	15.8 KiB



Statistics | Histogram | Common values | Extreme values

Quantile statistics

Minimum	0
5-th percentile	6
Q1	28
median	50
Q3	75
95-th percentile	95
Maximum	100
Range	100
Interquartile range (IQR)	47

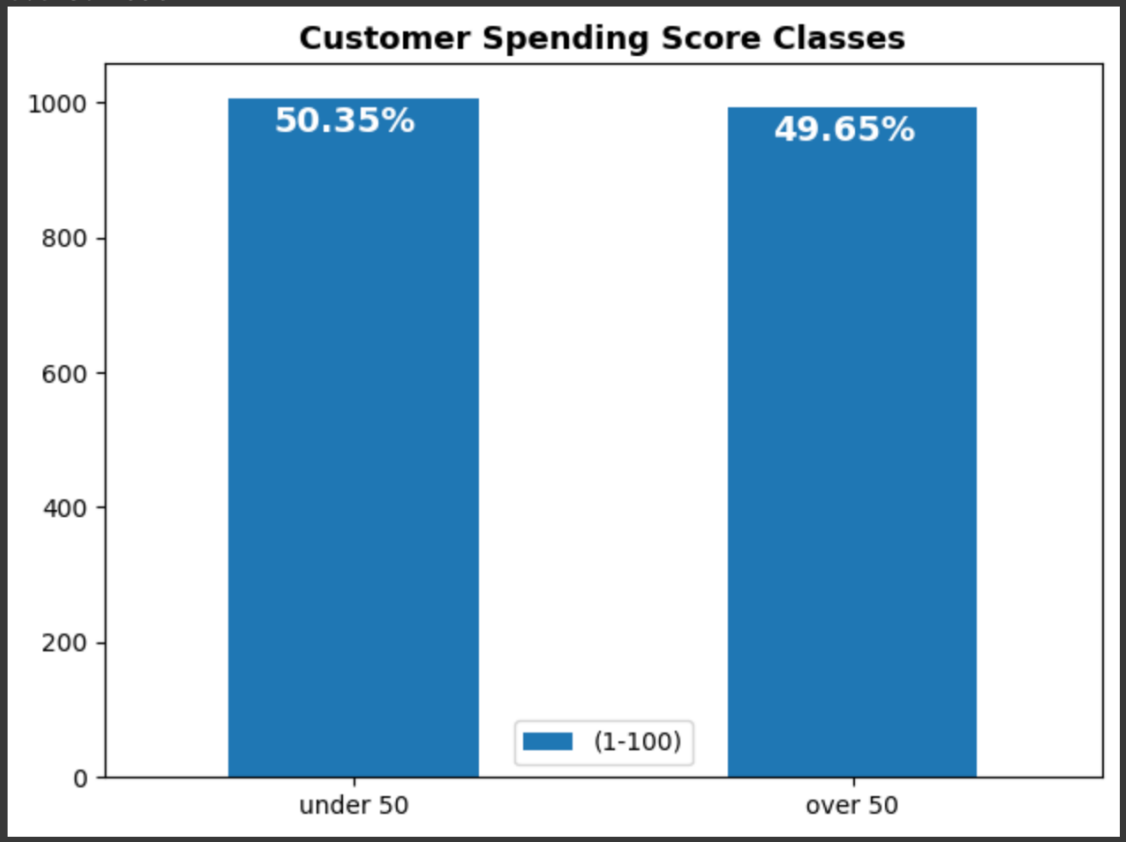
Descriptive statistics

Standard deviation	27.934661
Coefficient of variation (CV)	0.54814149
Kurtosis	-1.1007316
Mean	50.9625
Median Absolute Deviation (MAD)	24
Skewness	0.0045552482
Sum	101925
Variance	780.34527
Monotonicity	Not monotonic

3. EDA (Exploratory Data Analysis) Outcomes

Customer Spending Score Classes

```
under50  1007
over50  993
```



4. PDA (Predictive Data Analysis) Outcomes

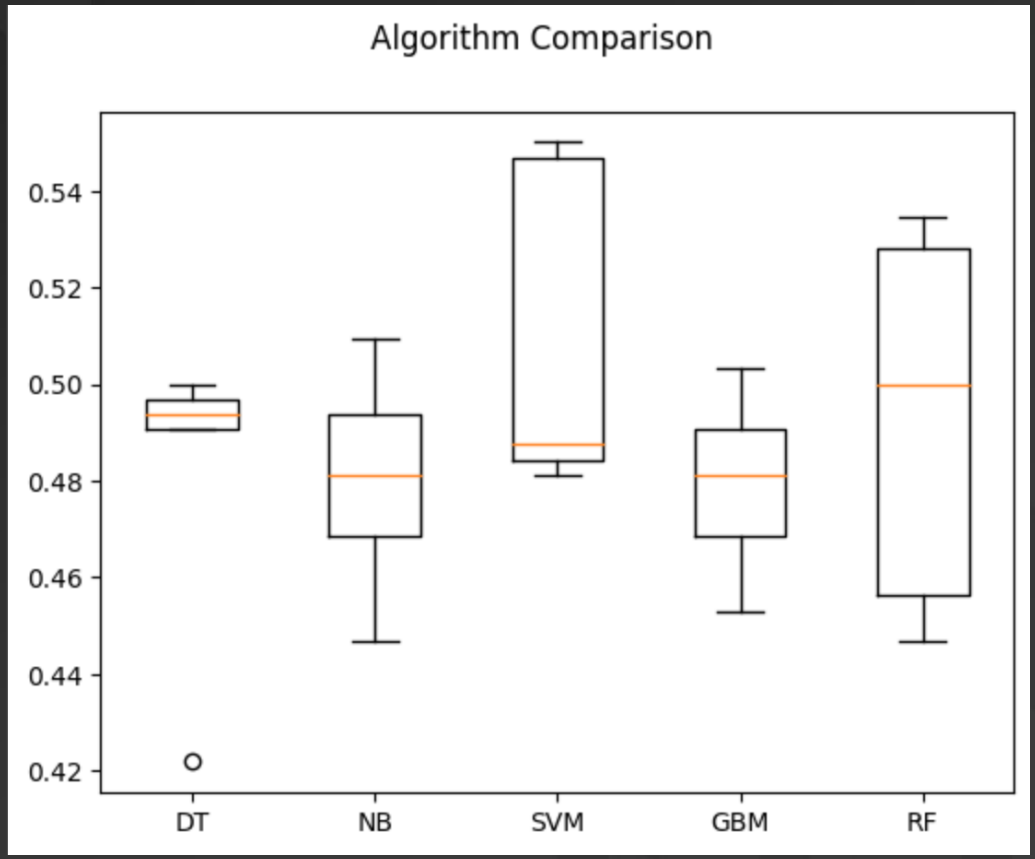
- Added a new column named ‘value customer’ based on spending score
- 1 = Spending score > 50
- 0 = Spending Score <50

```
class_label = df['Spending Score (1-100)']
df = df.drop(['Spending Score (1-100)'], axis =1)
df = (df-df.min())/(df.max()-df.min())
df['Spending Score (1-100)']= class_label
df['Value Customer'] = (df['Spending Score (1-100)'] > 50).astype(int)
df
```

4. PDA (Predictive Data Analysis) Outcomes

Algorithm Comparison

- Choose RF (Random Forest Classifier) for the model performance evaluation



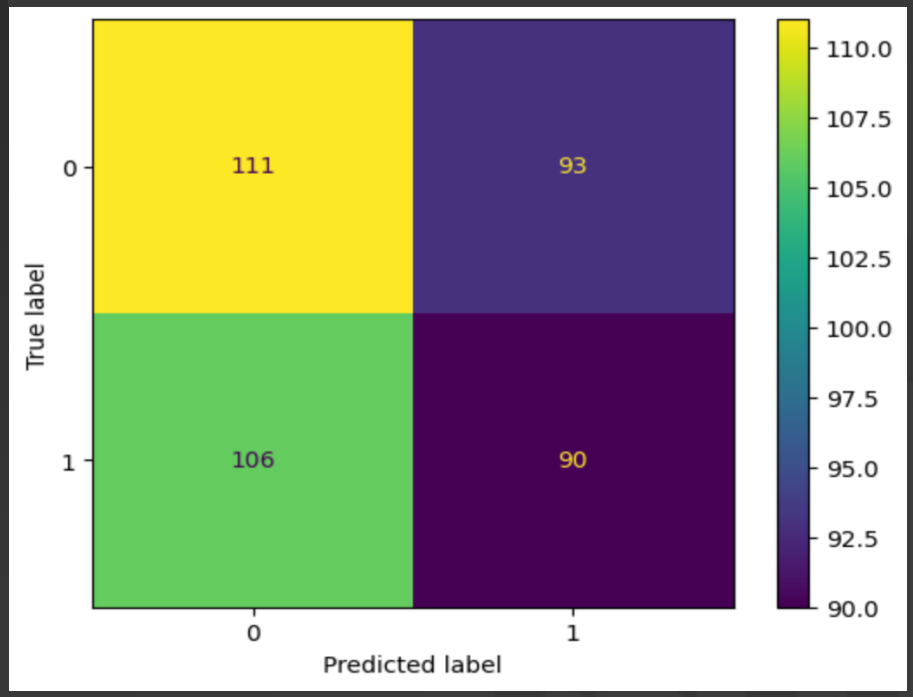
4. PDA (Predictive Data Analysis) Outcomes

Model Performance Evaluation

- Classification report

	precision	recall	f1-score	support
0	0.51	0.54	0.53	204
1	0.49	0.46	0.47	196
accuracy			0.50	400
macro avg	0.50	0.50	0.50	400
weighted avg	0.50	0.50	0.50	400

- Confusion matrix



5. Implementation and Deployment (TkInter) Plan and Status Update

- Implementation: TkInter
- Deployment plan: GitHub
- Current Status: Write a Gui program by using TkInter

References /Bibilography

Open Database, *Shop Customer Data*, [Online]. Available: https://www.kaggle.com/datasets/datascientistanna/customers-dataset (accessed: May 2, 2023)