TEAM 6 FCS5
Tan Jun Hern
Hudzaifah
Adrian

SC1015

ONLINE SHOPPING!

MINI PROJECT

Agenda

01 Introduction

02 Exploratory Data Analysis

03 Machine Learning

04 Optimization

05 Insights & outcome

1

Introduciton

a Problem Motivation

b Problem Statement

Practical Motivation

Photo: (StraitTimes, 2023)

Practical Motivation

Compound Annual Growth Rate of 9.47% from 2023 - 2027

(YahooFinance, 2024)

Photo: (wallpapercave)

Practical Motivation

E-commerce gross merchandise value in Singapore is expected to surge from **US$14.7 billion in 2022 to US$23.5 billion by 2027**

(yahoofinance, 2024)

Photo: (canva)

Dataset Used

Search

ZEE SOLVER · UPDATED 6 MONTHS AGO

165 | New Notebook | Download (150 kB)

Consumer Behavior and Shopping Habits Dataset:

E-Commerce Transaction Trends: A Comprehensive Dataset:

Data Card | Code (27) | Discussion (3) | Suggestions (0)

About Dataset

Context:

The Consumer Behavior and Shopping Habits Dataset provides comprehensive insights into consumers' preferences, tendencies, and patterns during their shopping experiences. This dataset encompasses a diverse range of variables, including demographic information, purchase history, product preferences, shopping frequency, and online/offline shopping behavior. With this rich collection of data, analysts and researchers can delve into the intricacies of consumer decision-making processes, aiding businesses in crafting targeted marketing strategies, optimizing product offerings, and enhancing overall customer satisfaction.

Content:

The Consumer Behavior and Shopping Habits Dataset provides a detailed overview of consumer preferences and purchasing behaviors. It includes demographic information, purchase history, product preferences, and preferred shopping channels (online or offline). This dataset is essential for businesses aiming to tailor their strategies to meet customer needs and enhance their shopping experience, ultimately driving sales and loyalty.

Dataset Glossary (Column-wise)

Customer ID:A unique identifier assigned to each individual customer, facilitating tracking and analysis of their shopping behavior over time.

Age: The age of the customer, providing demographic information for segmentation and targeted marketing strategies.

Gender: The gender identification of the customer, a key demographic variable influencing product preferences and purchasing patterns.

Usability
10.00

License
CC0: Public Domain

Expected update frequency
Annually

Tags

Business | Tabular | Retail and Shopping | Data Visualization | Exploratory Data Analysis | E-Commerce Services

https://www.kaggle.com/datasets/zeesolver/consumer-behavior-and-shopping-habits-dataset/data

Problem Statement

Is there a relationship between gender and purchasing behavior evident in the Consumer Behavior and Shopping Habits Dataset?

To find out if females the differences in shopping habits between males and females

2

Exploratory Data Analysis

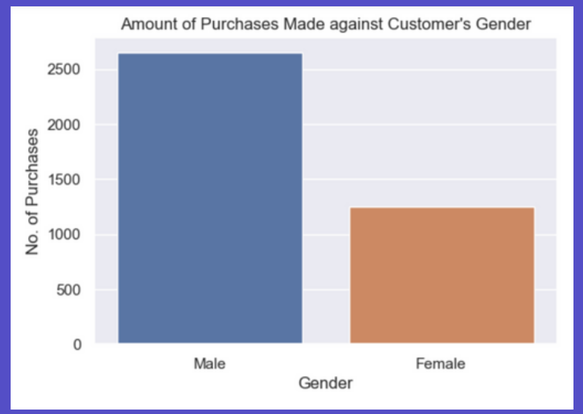
01 Data Extraction

02 Curation

03 Prep & Cleaning

Exploratory Analysis:

There are twice more males than females



Boxplot has no outliers

Numeric data are uniformly distributed.

	Age	Purchase Amount (USD)	Review Rating	Previous Purchases
count	3900.00	3900.00	3900.00	3900.00
mean	44.07	59.76	3.75	25.35
std	15.21	23.69	0.72	14.45
min	18.00	20.00	2.50	1.00
25%	31.00	39.00	3.10	13.00
50%	44.00	60.00	3.70	25.00
75%	57.00	81.00	4.40	38.00
max	70.00	100.00	5.00	50.00

No duplicates and no null values

```
df.isnull().sum()
```

```
Customer ID               0
Age                       0
Gender                    0
Item Purchased            0
Category                  0
Purchase Amount (USD)     0
Location                  0
Size                      0
Color                     0
Season                    0
Review Rating             0
Subscription Status       0
Shipping Type             0
Discount Applied          0
Promo Code Used           0
Previous Purchases        0
Payment Method            0
Frequency of Purchases    0
dtype: int64
```

```
As seen above, the dataset does not contain null values.
Next, we check for duplicates.
```

```
value=len(df[df.duplicated()])
print(value)
```

```
0
```

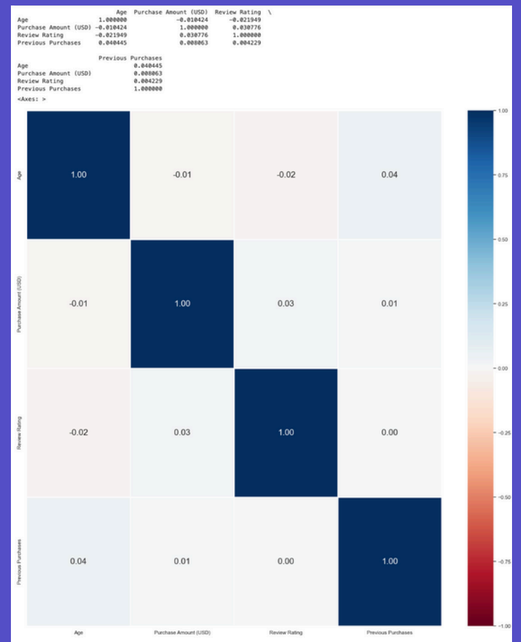
Exploratory Analysis:

M-size shirts are popular choice

No numeric variable have significant correlation with each other.

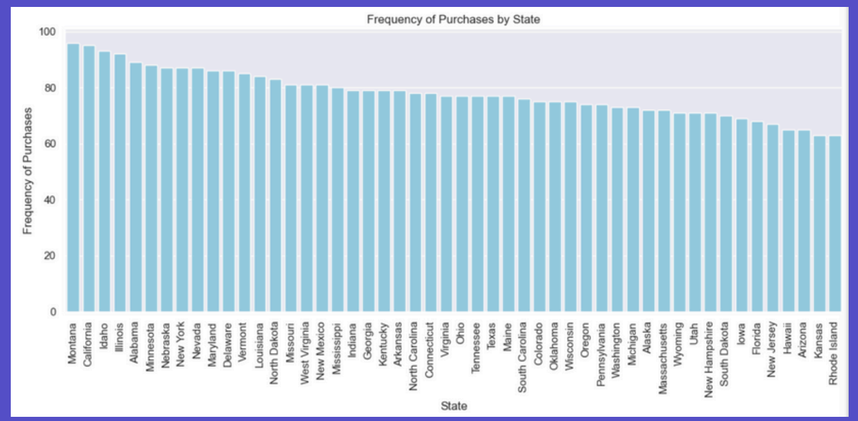
```
                        Age  Purchase Amount (USD)  Review Rating  \
Age                1.000000              -0.010424      -0.021949
Purchase Amount (USD) -0.010424            1.000000       0.030776
Review Rating     -0.021949               0.030776       1.000000
Previous Purchases 0.040045               0.000063       0.004229

                   Previous Purchases
Age                          0.040045
Purchase Amount (USD)        0.000063
Review Rating                0.004229
Previous Purchases           1.000000
<Axes: >
```



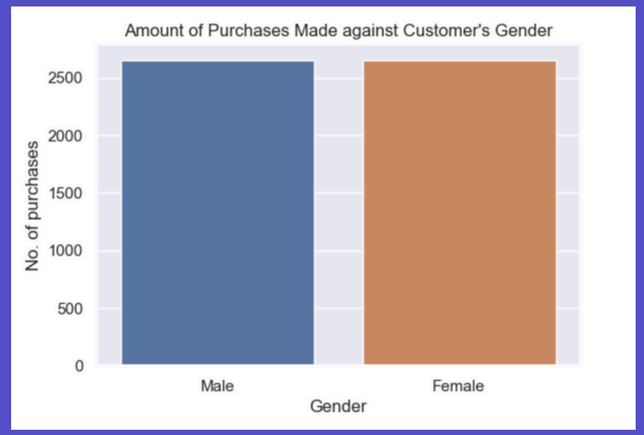
It is evident that no numeric variables have significant correlation with another.

This shows us that the region matters; Montana has the highest frequency of purchases.

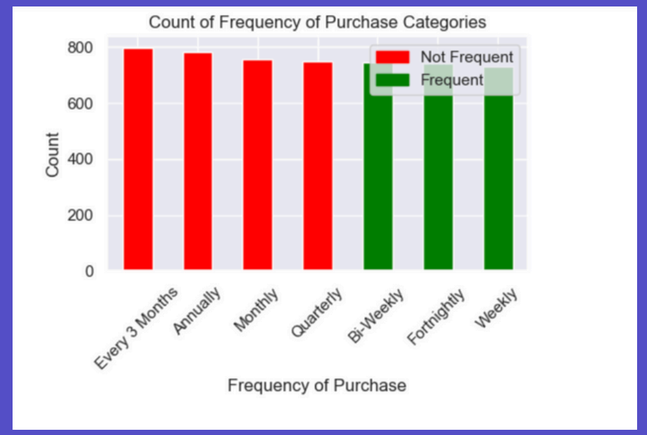


Data preparation:

Upsampling for imbalance



Feature engineering



One hot encoding

	Purchase Amount (USD)	Previous Purchases	Age	High Frequency	Location_Alabama	Location_Alaska
0	53	14	55	True	False	False
1	64	2	19	True	False	False
2	73	23	50	True	False	False
3	90	49	21	True	False	False
4	49	31	45	False	False	False
...	...	...	...	...	...	...
5299	30	42	61	True	False	False
5300	52	46	58	False	False	False
5301	60	31	23	True	False	False
5302	100	48	28	False	False	False
5303	63	7	58	True	False	False

5304 rows × 132 columns

3

Machine Learning

01 Importance features

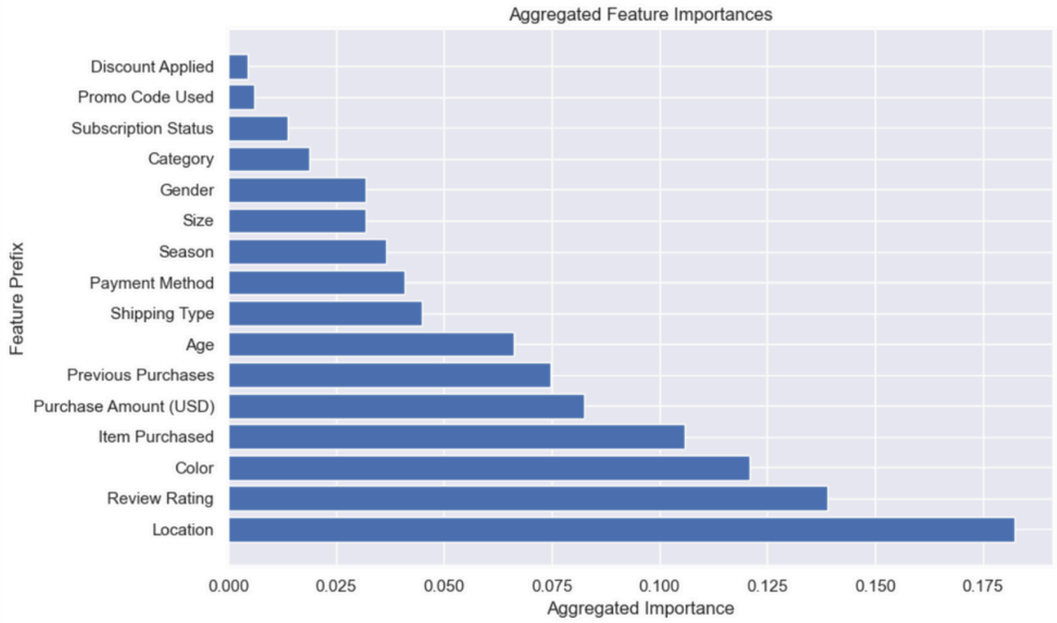
02 Decision Tree Classification

03 Random Forest Classification

What are the importance features?

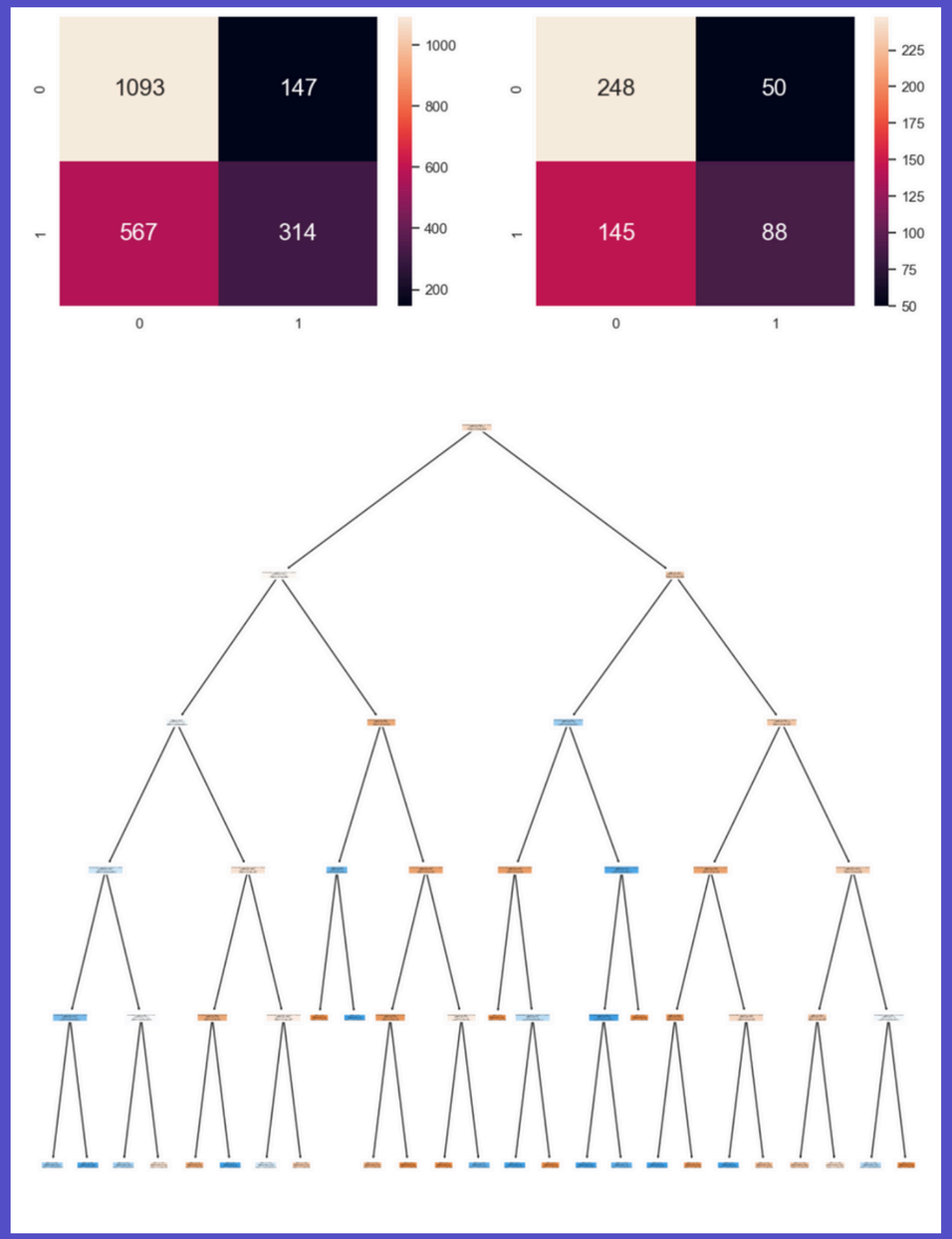
- Aggregated Feature Importance based on prefix
- Location - most important
- Gender - considerably important
- Discount doesn't affect the high-frequency purchasing behaviour.

```
Aggregated Feature Importance:
1. Location: 0.18225019538589574
2. Review Rating: 0.13895217926511141
3. Color: 0.12079647687164931
4. Item Purchased: 0.10580019980076319
5. Purchase Amount (USD): 0.08261095779632148
6. Previous Purchases: 0.07463545171486036
7. Age: 0.06622477197642464
8. Shipping Type: 0.04476823684572869
9. Payment Method: 0.040830761479895464
10. Season: 0.03670495289437293
11. Size: 0.03179149602734477
12. Gender: 0.03170352805657481
13. Category: 0.018681681541286466
14. Subscription Status: 0.013871679935173614
15. Promo Code Used: 0.0059037247718159895
16. Discount Applied: 0.004473705636781098
```



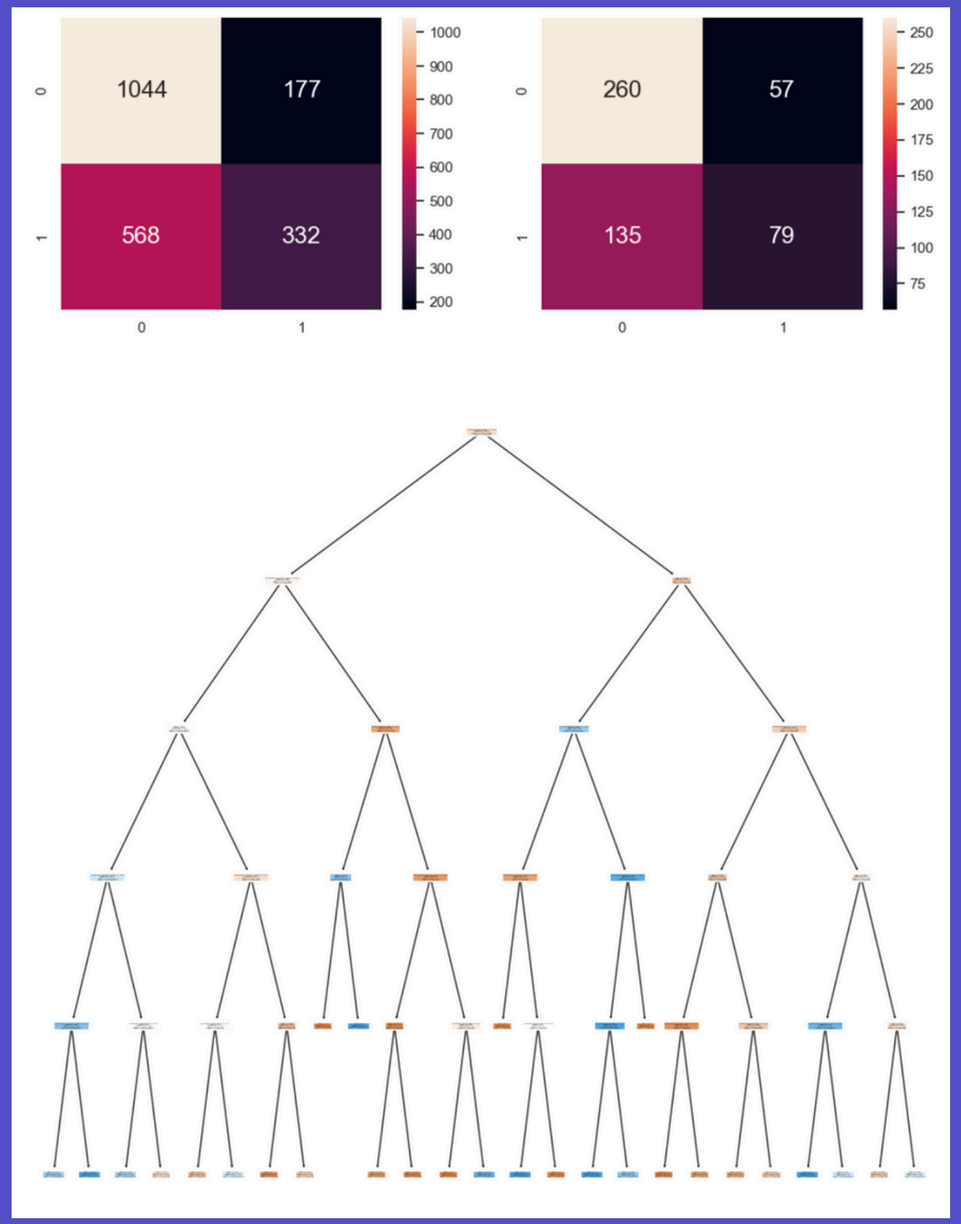
Decision Tree Classification (M)

- Train Dataset Accuracy: 66.34%
- Test Dataset Accuracy: 63.28%
- There's a slight drop in accuracy from the training to the test dataset.
- The model performs reasonably well but may benefit from further tuning to improve generalization. We apply hyper tuning later on.



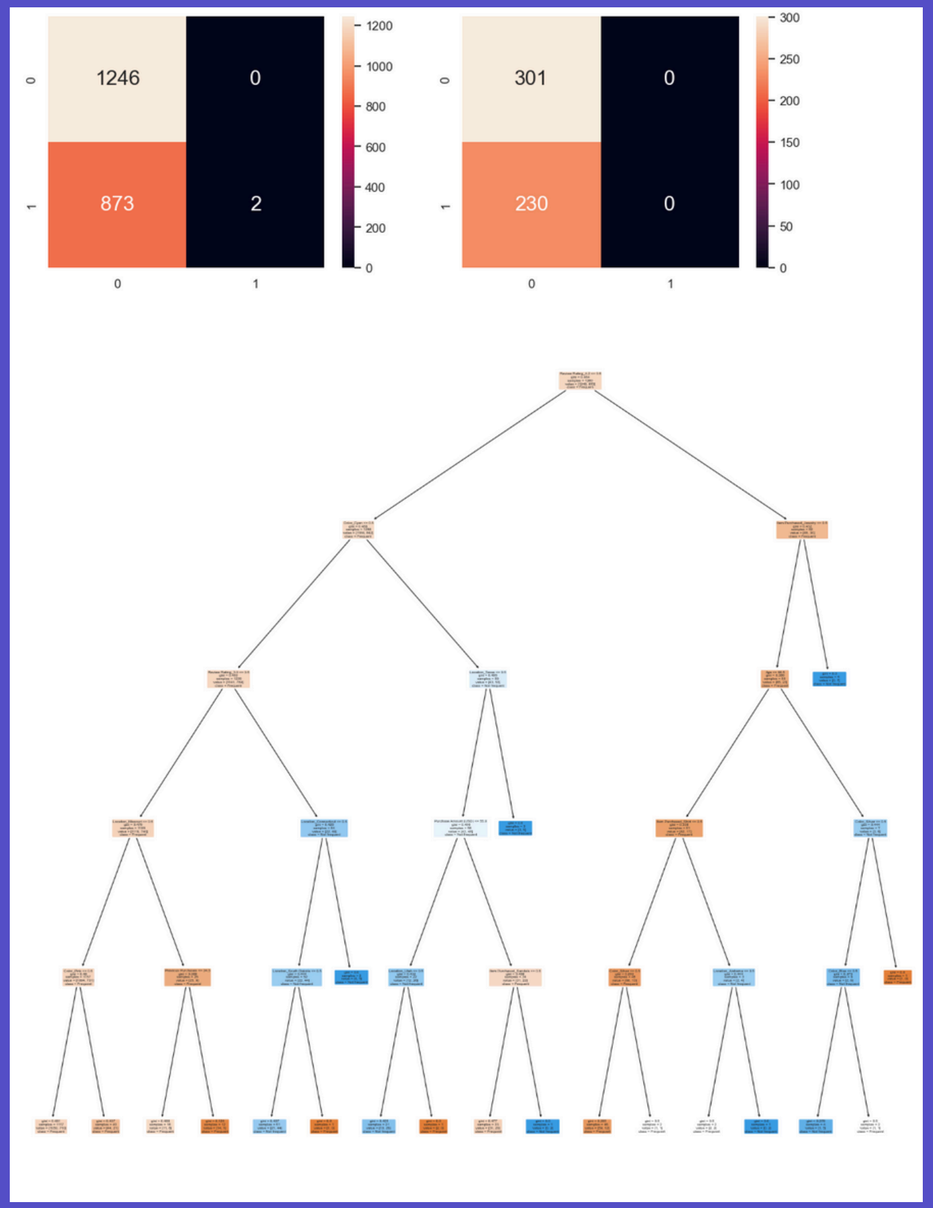
Decision Tree Classification (F)

- Train Dataset Accuracy: 64.88%
- Test Dataset Accuracy: 63.84%
- The model demonstrates similar performance between training and test datasets.
- Accuracy is slightly lower compared to the male dataset, suggesting potential differences in predictive factors. Possibly overfitting.



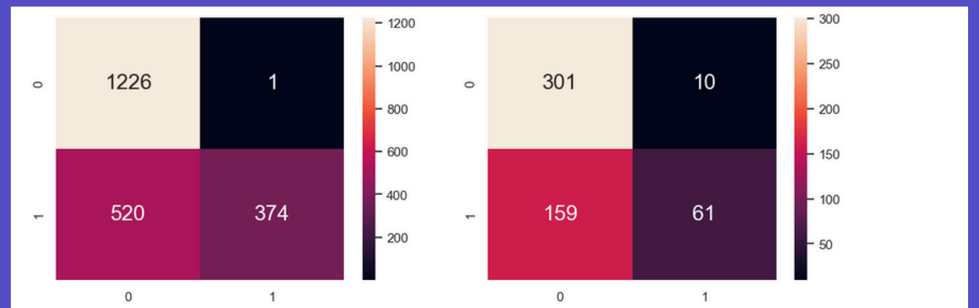
Random Forest Classification (M)

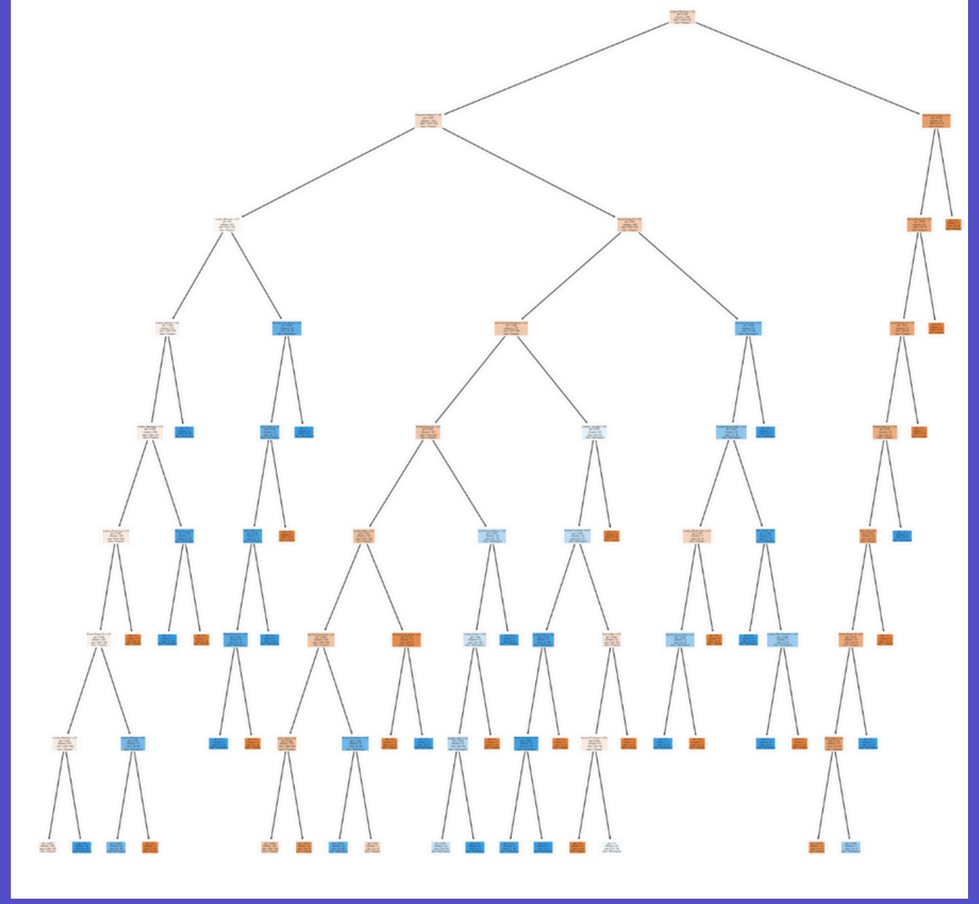
- Train Dataset Accuracy: 58.84%
- Test Dataset Accuracy: 56.69%
- Both train and test accuracies are lower compared to the female dataset.
- Similar to the decision tree, we may need further tuning or feature selection to improve performance.



Random Forest Classification (F)

- Train Dataset Accuracy: 75.44%
- Test Dataset Accuracy: 68.17%
- Higher accuracy compared to the male dataset, indicating better predictive power.
- Note that test accuracy is slightly lower than the training accuracy, suggesting potential overfitting.





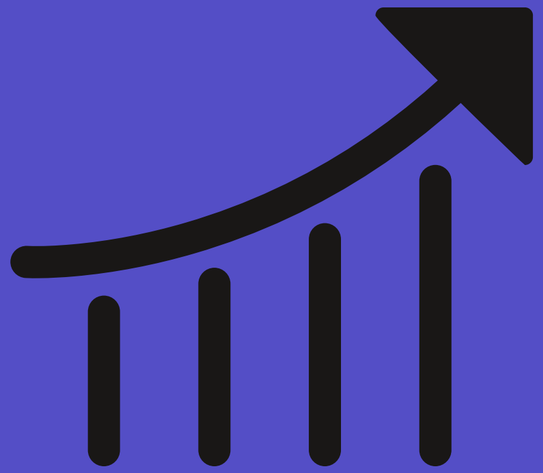
4

Optimization

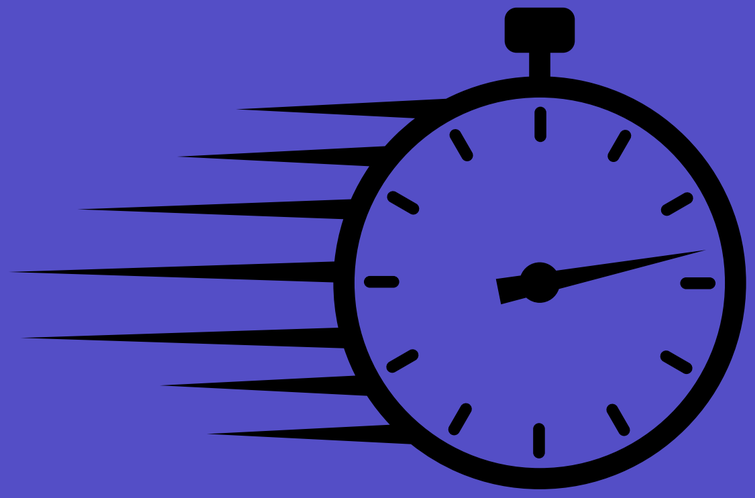
a Hypertuning

Hypertuning

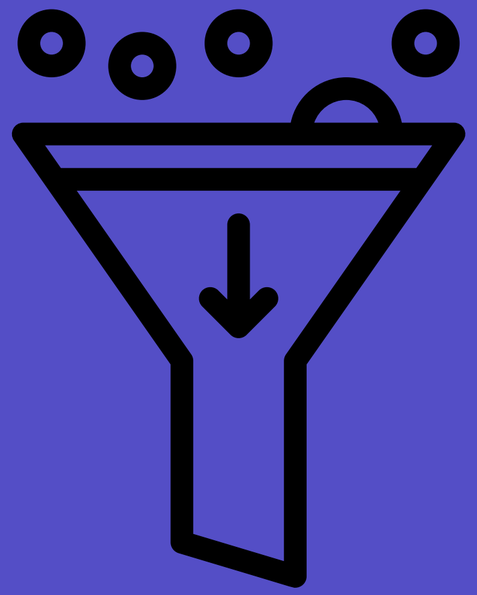
Improves model Performance



Time & resource

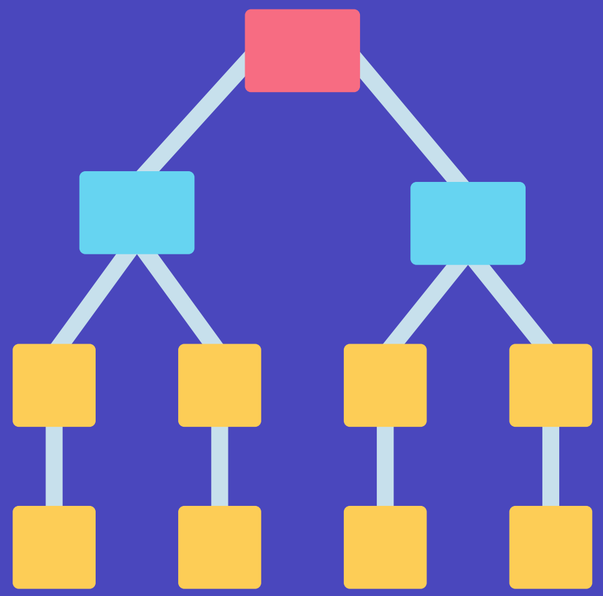


Enhanced Generalization



Application of Hypertuning

1. n_estimators (Random Forest): Sets the number of decision trees built to aggregate the model. Spots intricate patterns at higher computational costs
2. max_depth: Restricts depth. More interpretable model & enhance generalization.
3. test-train split: Dataset Partitioning. Giving the model a good balance of training data and performance evaluation



5

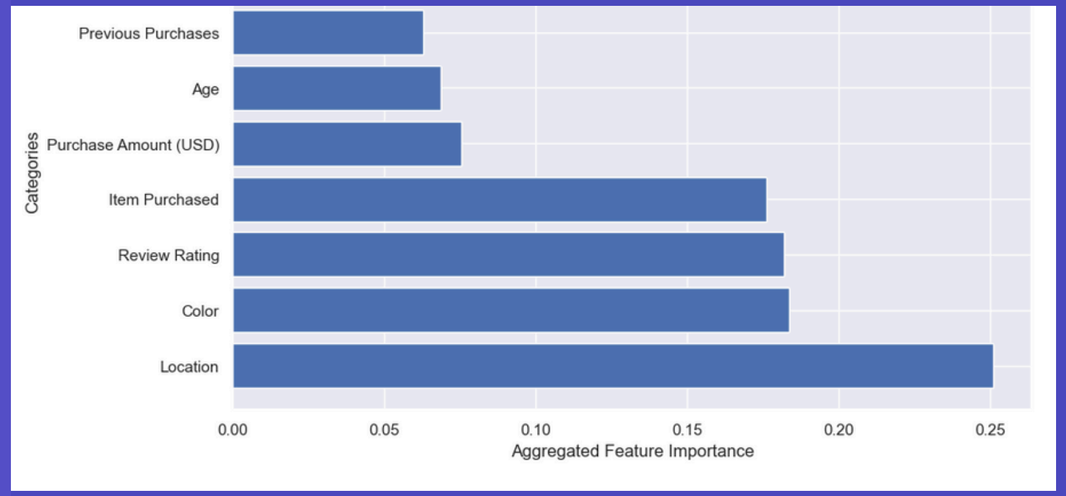
Insights & outcome

a Project outcomes

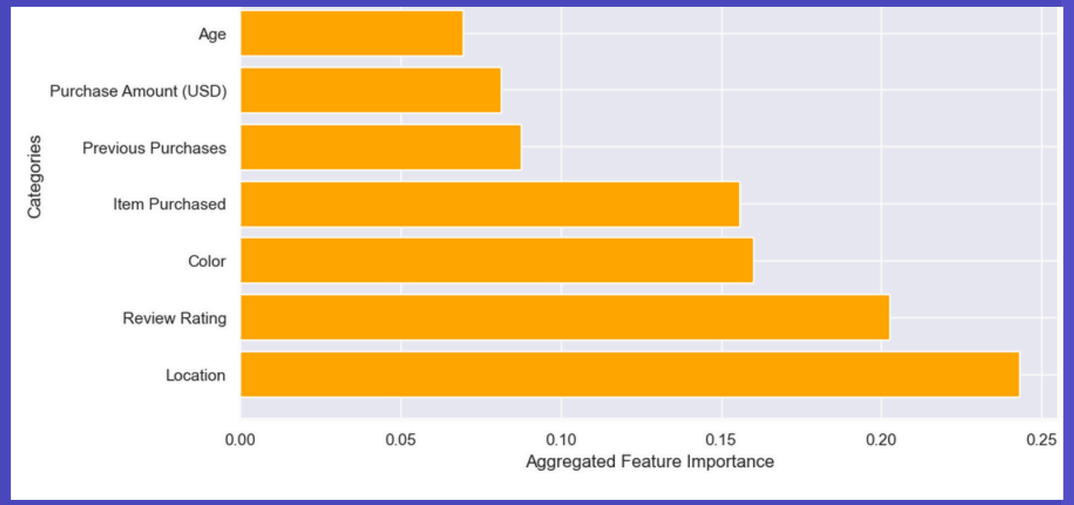
b Interesting insights

c Recommendations

Findings after Hypertuning random forest classification



Important features for Males



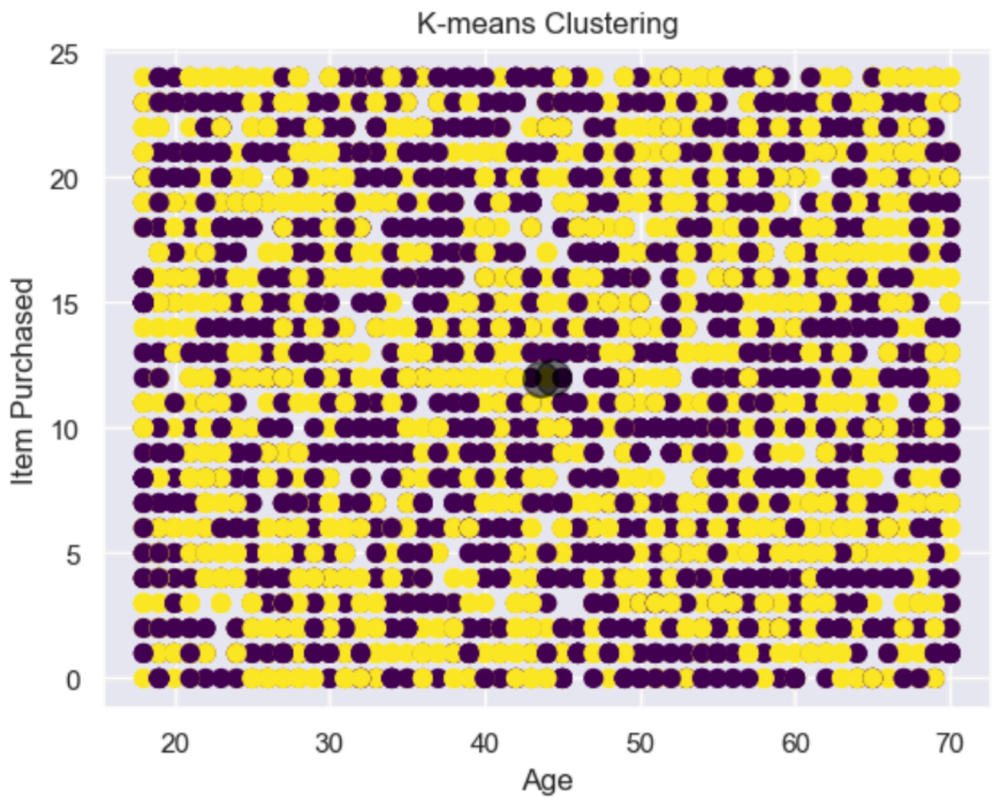
Important features for Females

Interesting insights

It is worth noting that we experimented with K means clustering - a nonsupervised machine learning technique.

The algorithm automatically finds out any relations of gender to other categories by clusters.

- Age and Item purchased are the clusters the algorithm picked up.
- Found that the dominant gender in both clusters is females.




```
In Cluster 0, the dominant gender is Female
In Cluster 1, the dominant gender is Female
```

Recomendations

- Implement random forest classification for predictive performance and optimize marketing efforts.
- Female customers are overlooked in the dataset.

Recommendations

Utilize classifiers to improve marketing strategies to understand customer behavior and segmentation that maximizes sales.

Photo: (StraitTimes, 2022)

Thank you

References:

1. ZeeSolver. (n.d.). Consumer Behavior and Shopping Habits Dataset. Kaggle. https://www.kaggle.com/datasets/zeesolver/consumer-behavior-and-shopping-habits-dataset
2. Analytics Vidhya. (2021, June). Understanding Random Forest. https://www.analyticsvidhya.com/blog/2021/06/understanding-random-forest/
3. Microsoft. (n.d.). https://www.microsoft.com/en-sg/
4. Project Jupyter. (n.d.). https://jupyter.org

Special Thanks to Ma'am Kethmi & Dr. Smitha for your guidance.