Capstone Project

Customer Behaviour Segmentation

Sakina Sakdun

What is customer segmentation?

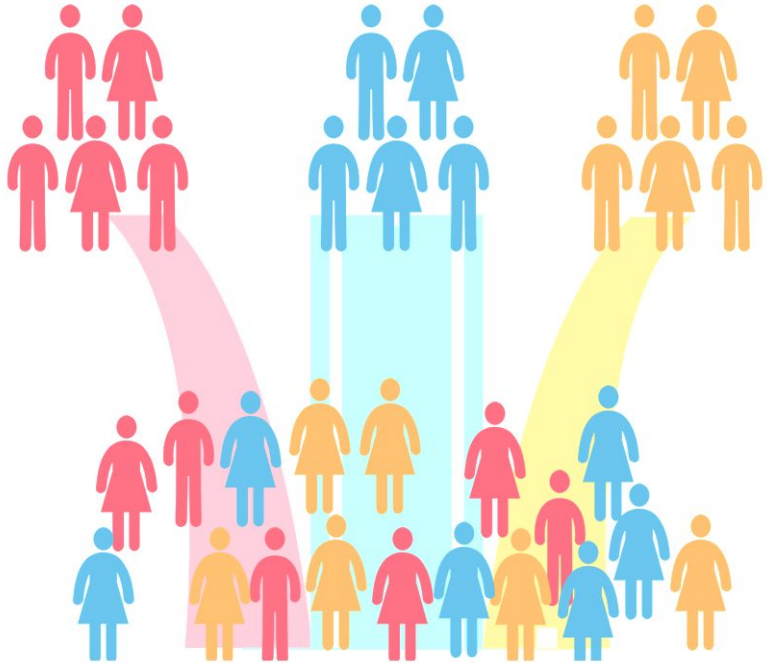
Homogeneous groups

Helps marketers better understand similar traits within each group

Customise marketing campaigns on each target group

⇩

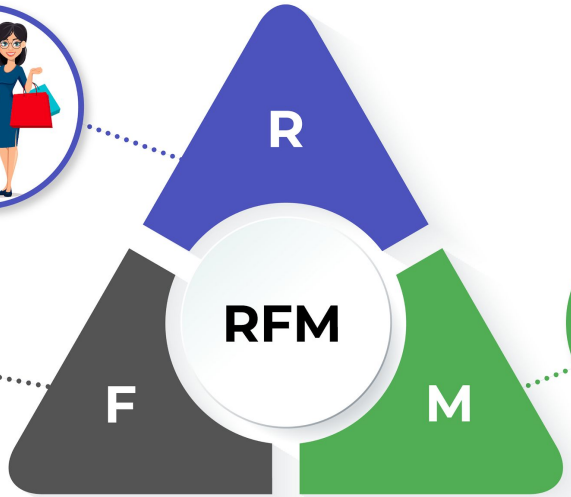
Business Value
Efficient resource allocation
thus **maximising** profits



RFM Framework

The freshness of customer's activity
(i.e. days from last purchase)

Recency



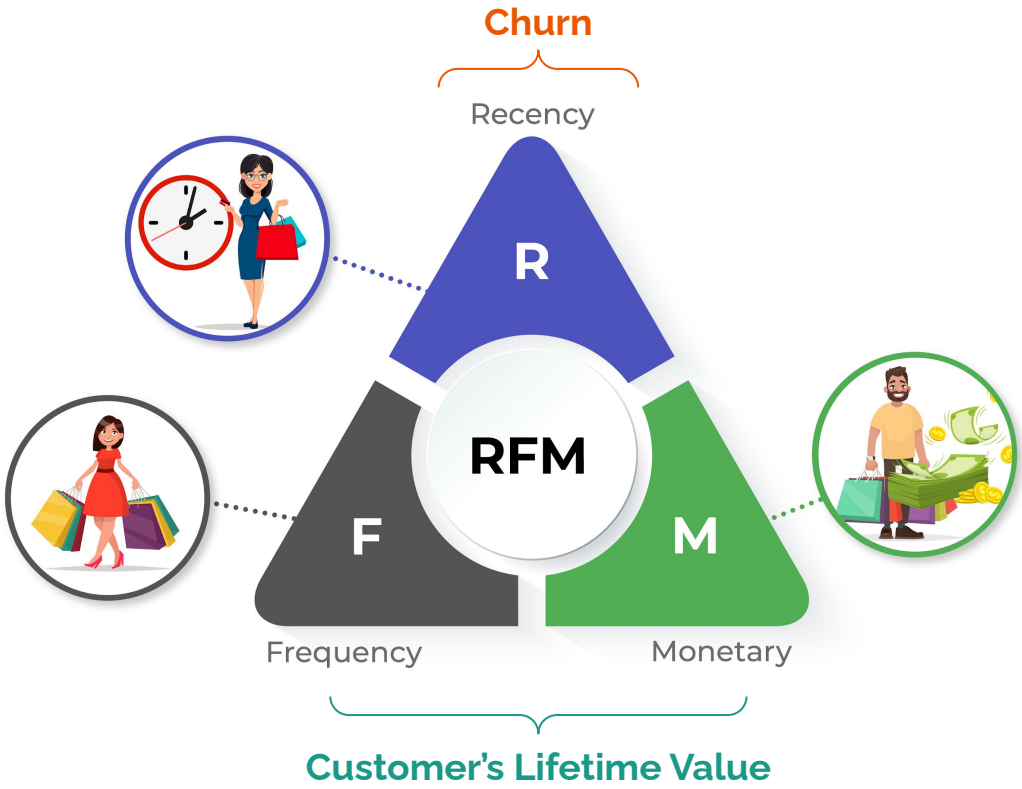
Frequency

Monetary

The frequency of customer transactions
(i.e. total purchases)

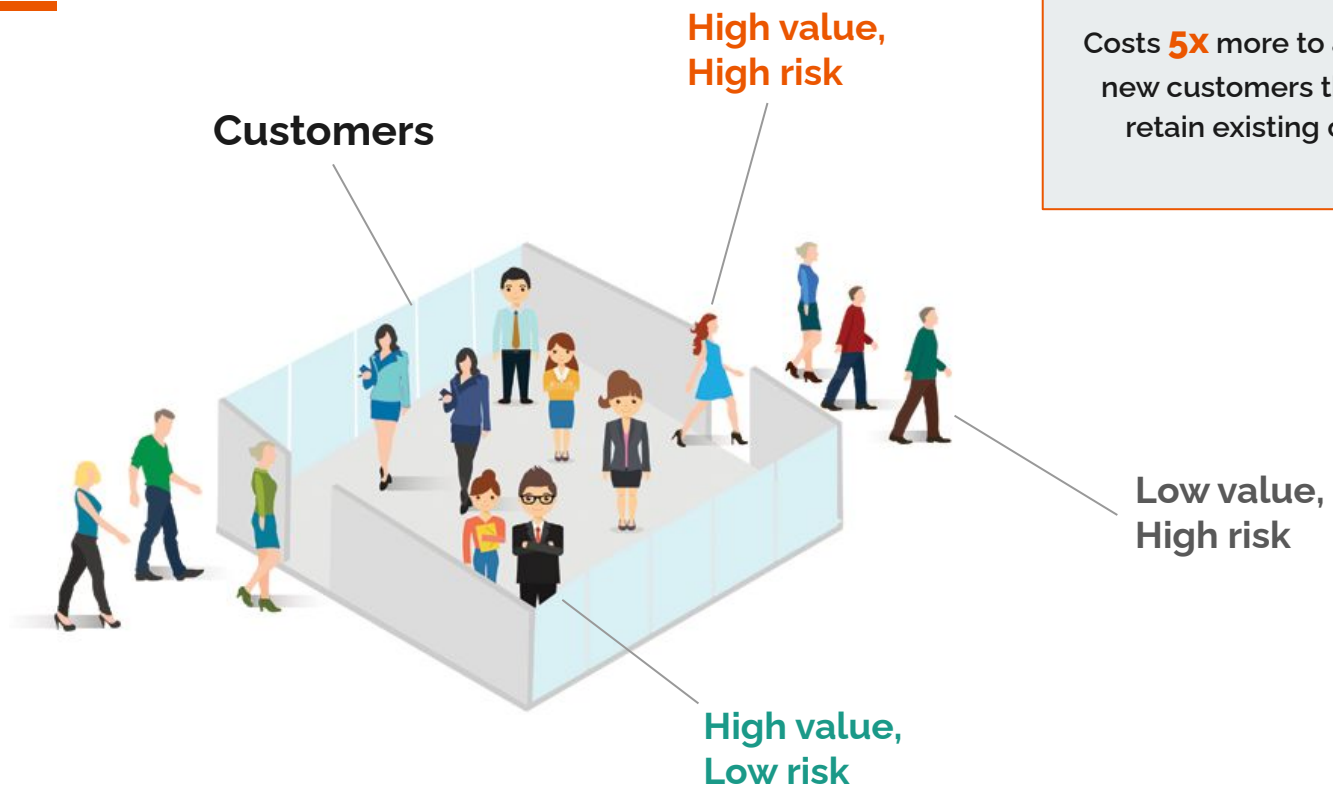
Purchasing power of customer
(i.e. total spending $$)

Why is RFM important?



Why segment customers using RFM?

Costs **5x** more to acquire new customers than to retain existing ones



Business Problem

Which of our customers that are worth retaining?

- *Head of Marketing Team of an E-commerce site (stakeholder)*

Data Question

Can we use unsupervised clustering algorithm to identify high value & high risk customers, better than a manual RFM calculation?

Process Workflow

E-commerce User Behaviour Dataset

from multi-category store [Kaggle]

Raw Data

167 million rows, 8 columns
Event time: Jan-Mar 2020

Raw Data Columns
event_time
event_type (i.e. view, cart, purchase)
product_id
category_id
category_code
brand
price
user_id
user_session

+

RFM Framework

Recency
'days_from_last_purchase'

Frequency
'purchase_total'

Monetary
'purchase_amount_total'

Feature Engineering on **User-Id** Level

Scope:
Repeat Customers
(more than 1 purchasing day)

Processed Data

338,000+ rows, 4 columns

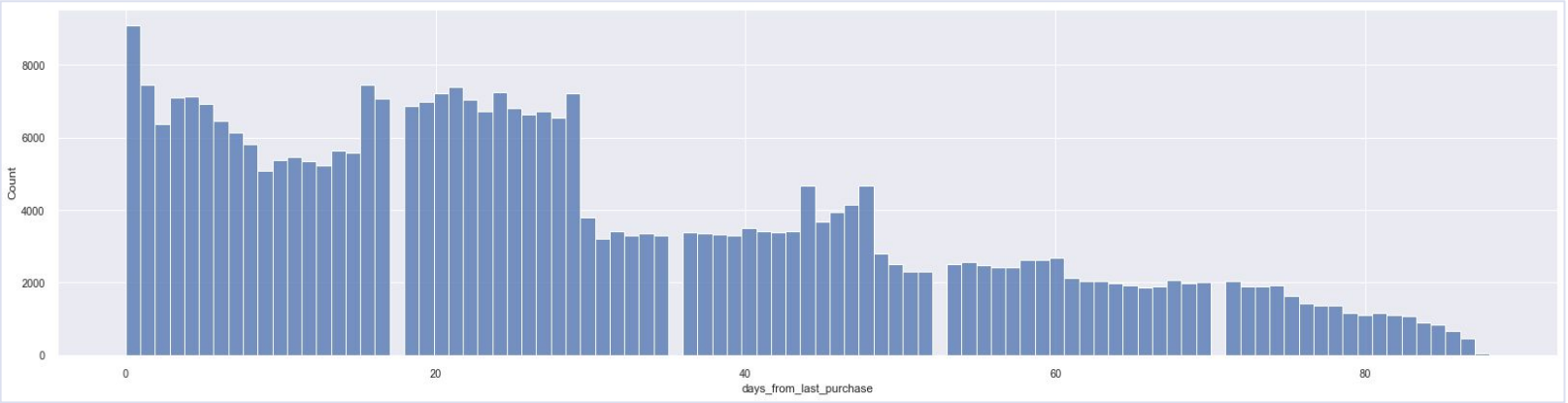
DataFrame Columns
user_id
recency
frequency
monetary

RFM Manual Calculation

RFM with Unsupervised Clustering

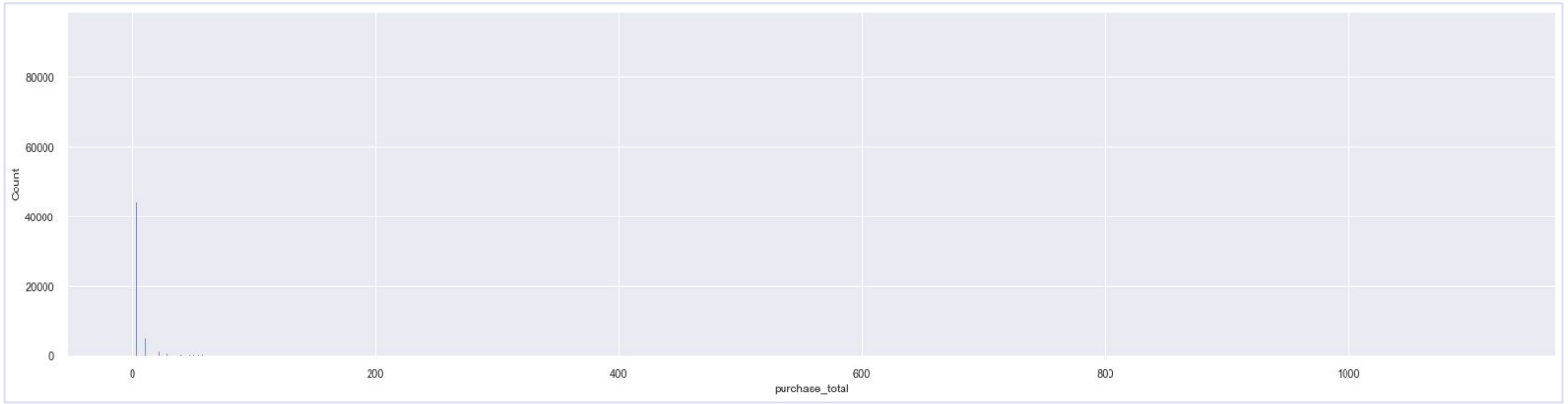
Exploratory Data Analysis (EDA)

‘recency’



Exploratory Data Analysis (EDA)

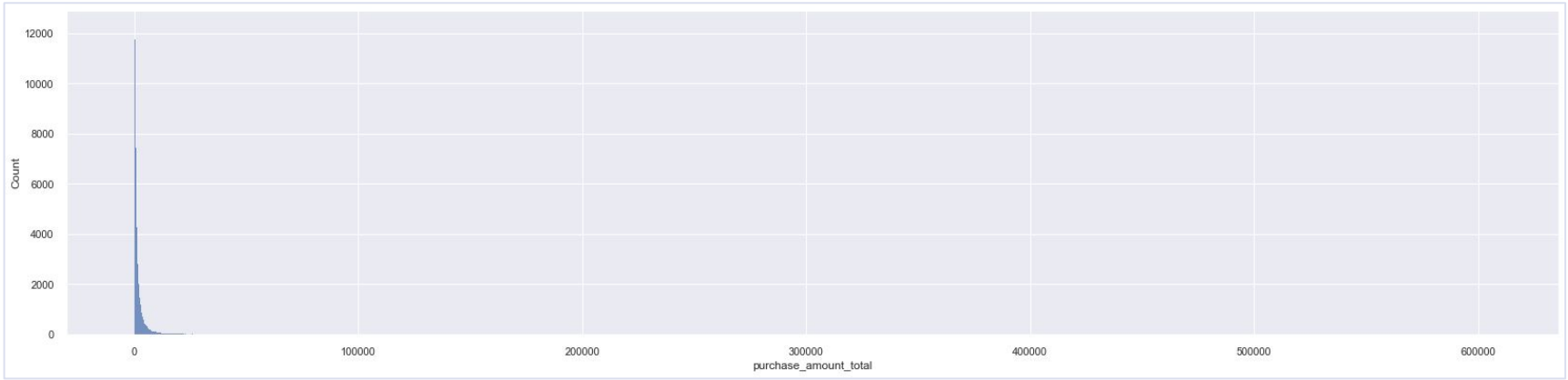
‘frequency’



Extremely right-skewed

Exploratory Data Analysis (EDA)

‘monetary’



Extremely right-skewed

Manual RFM Calculation

Using quartiles,

- Assign a score from 1 to 4 to R, F, M
  - 4 - most ideal
  - 1 - least ideal
- Final RFM score is calculated simply by combining individual R, F, M value numbers

  (best customer = '444')

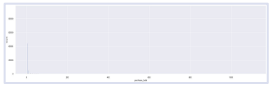
user	recency	frequency	monetary	R	F	M	RFM_score	
A	34.0	2.0	117.84	2	1	1	211	Low value, Mid risk
B	11.0	11.0	6429.41	4	4	4	444	High value, Low risk
C	55.0	11.0	12028.97	1	4	4	144	High value, High risk

RFM + Mini Batch KMeans

Processed Data

R

F

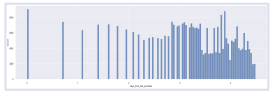


M

Log Transform

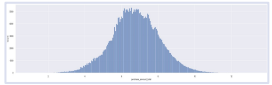
Model Input Features

R



F

M

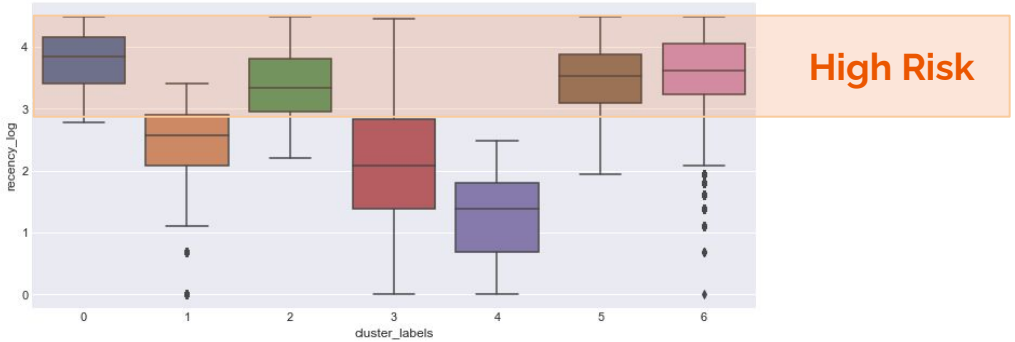


Unsupervised Clustering
Mini Batch KMeans

Elbow Method
Optimal Cluster Size = **7**

RFM + Mini Batch KMeans - Clusters

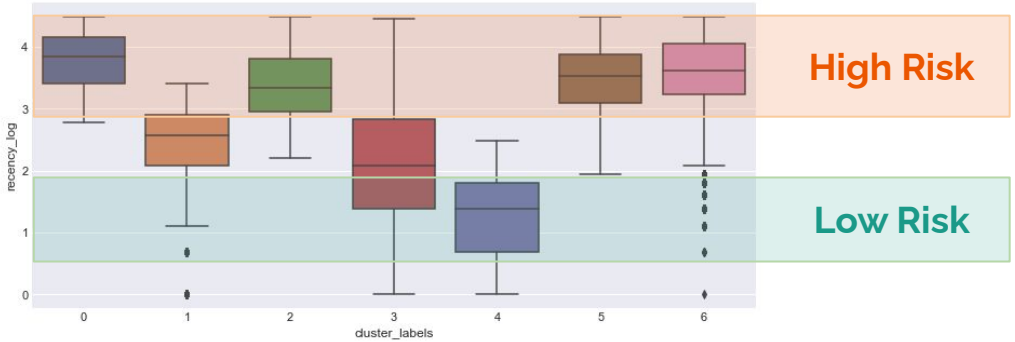
R



Cluster Labels	Risk Level	
0	High	
1		
2	High	
3		
4		
5	High	
6	High	

RFM + Mini Batch KMeans - Clusters

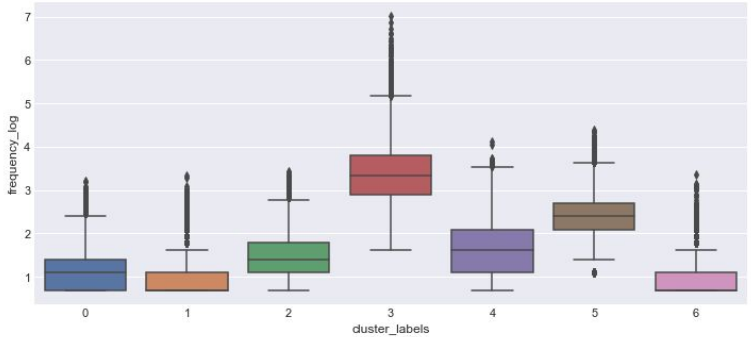
R



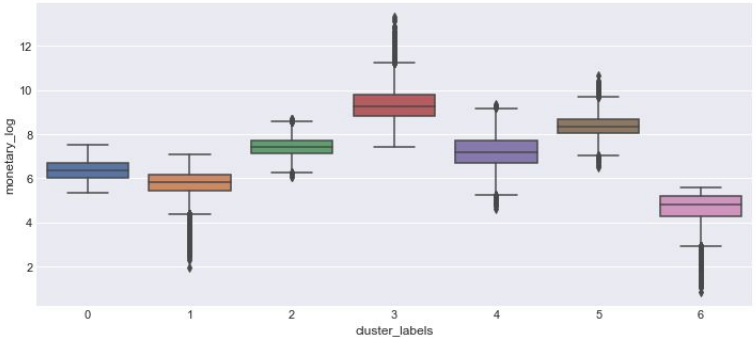
Cluster Labels	Risk Level	
0	High	
1	Mid	
2	High	
3	Mid	
4	Low	
5	High	
6	High	

RFM + Mini Batch KMeans - Clusters

F



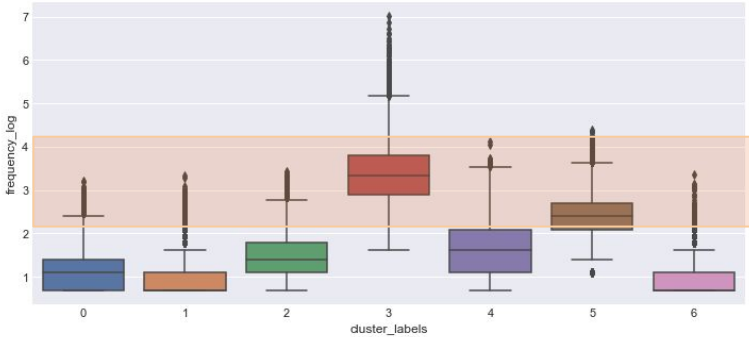
M



Cluster Labels	Risk Level	
0	High	
1	Mid	
2	High	
3	Mid	
4	Low	
5	High	
6	High	

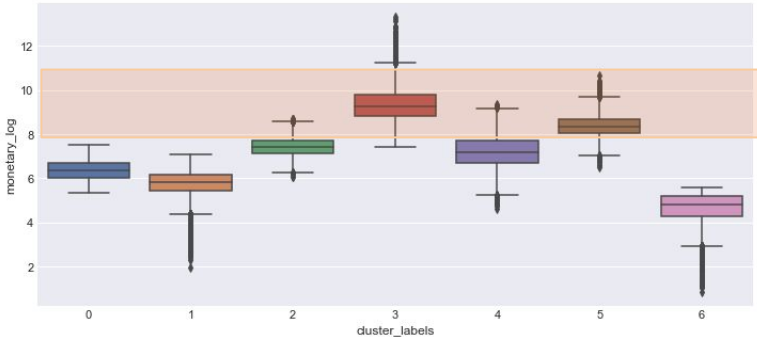
RFM + Mini Batch KMeans - Clusters

F



High Value

M

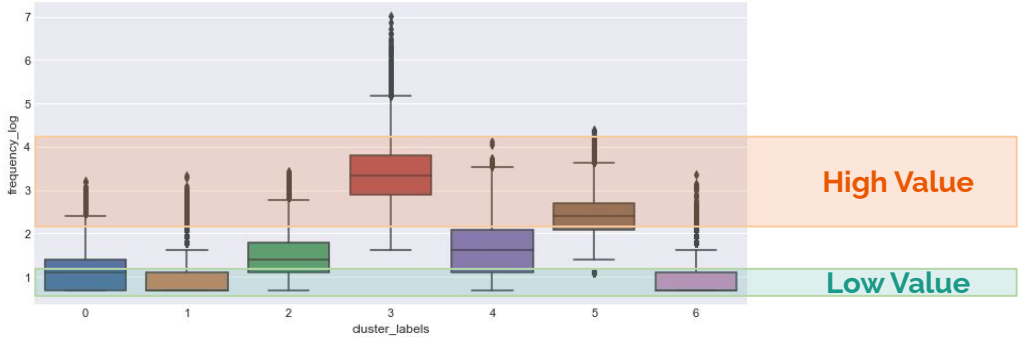


High Value

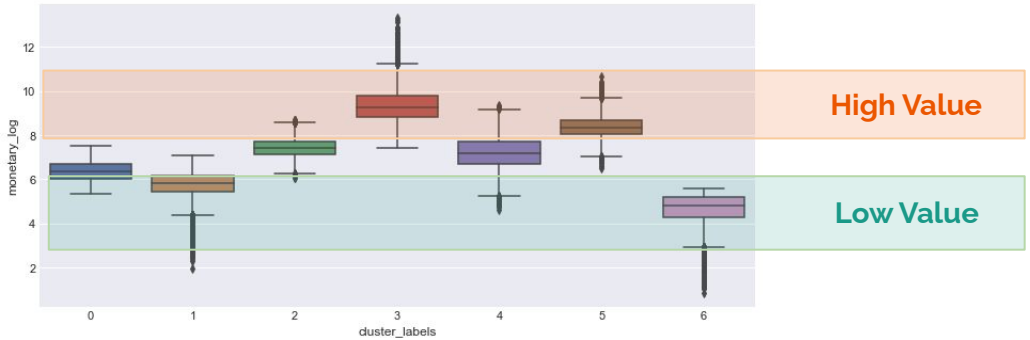
Cluster Labels	Risk Level	Value Level
0	High	
1	High	
2	Mid	
3	Low	High
4	Low	
5	High	High
6	High	

RFM + Mini Batch KMeans - Clusters

F



M



Cluster Labels	Risk Level	Value Level
0	High	Mid
1	High	Low
2	Mid	Mid
3	Low	High
4	Low	Mid
5	High	High
6	High	Low

RFM Manual Calculations

True: High Value, High Risk

(low recency, high frequency, high monetary)

R	F	M
‘1	4	4’
‘1	3	4’
‘1	4	3’

Accuracy Score: **0.67**

RFM with Mini Batch KMeans

Predicted: High Value, High Risk

(low recency, high frequency, high monetary)

Cluster Label: ‘5’

RFM Manual Calculations

‘1 4 4’
‘1 3 4’
‘1 4 3’

14400 customers
(4% of total)

‘2 4 4’ ‘1 3 3’
‘2 4 3’ ‘2 4 2’
‘1 2 3’ ‘2 3 3’
...
..
.

‘Undetected’
high value, high risk
due to human
subjectivity

64 groups total
Difficult to optimally split or combine the groups

RFM with Mini Batch KMeans

‘High Value, High Risk’

37200 customers
(11% of total)

If this churns

Possible Loss of Revenue
$ 136 million

(14% of 3-month revenue)

(assuming users spend the same amount if they are targeted for retention)

Conclusion

Can we use unsupervised clustering algorithm to identify high value & high risk customers, better than a manual RFM calculation?

Yes, better, in terms less subjectivity in splitting and the revenue amount saved

Next Steps

Recency Frequency Engagement (RFE)

To measure users' engagement level such as visit duration, pages per visit
(not related to purchases)

Product Display

To predict which items that customers are likely to buy, based on their search history

A/B Testing on UX Improvements

To analyse if there is improvement in engagement after new releases