Credit Card Customer Analysis

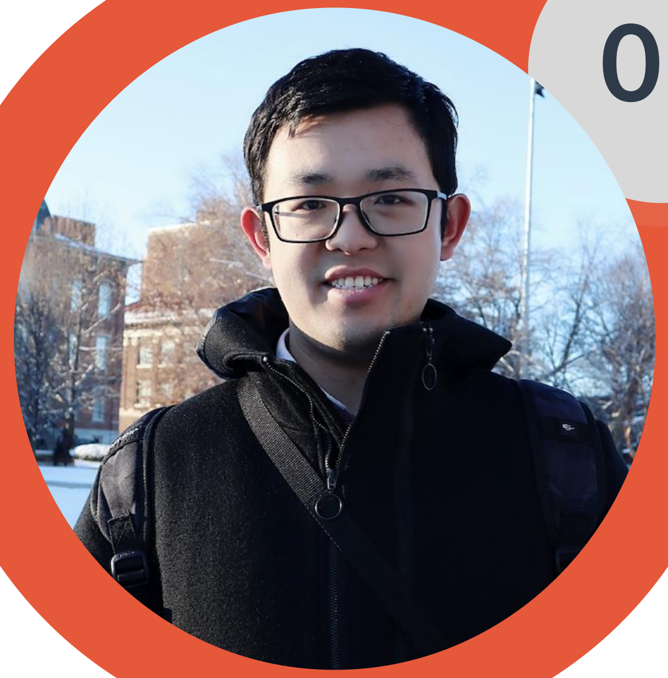
Presented by Team 2

Yulong Gong, Muyan Xie, Yangyang Zhou, Yichi Zhang

Team Member

Your great subtitle in this line

01

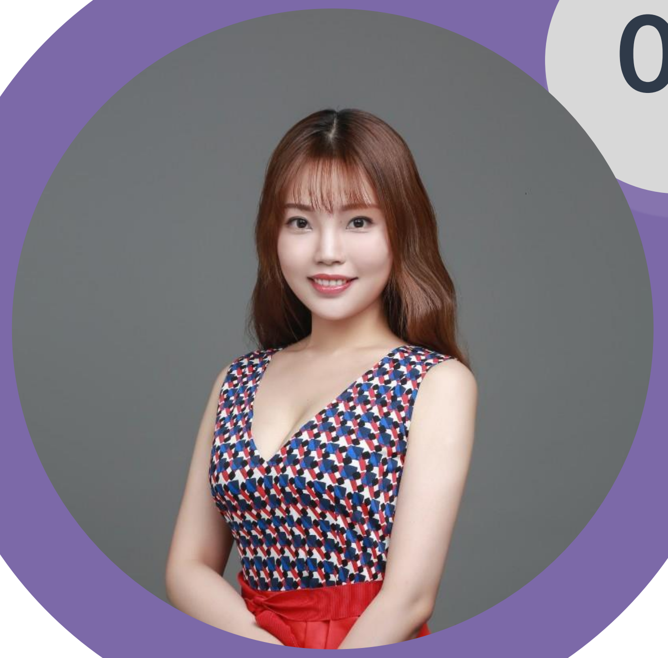


Yulong Gong

02

Muyan Xie

03



Yangyang Zhou

04

Yichi Zhang

Outline

1 Business Problem

2 Dataset

3 EDA

4 Card type

4 Churn/ no churn

5 Results & Recommendation

6 Q&A

Business Problem

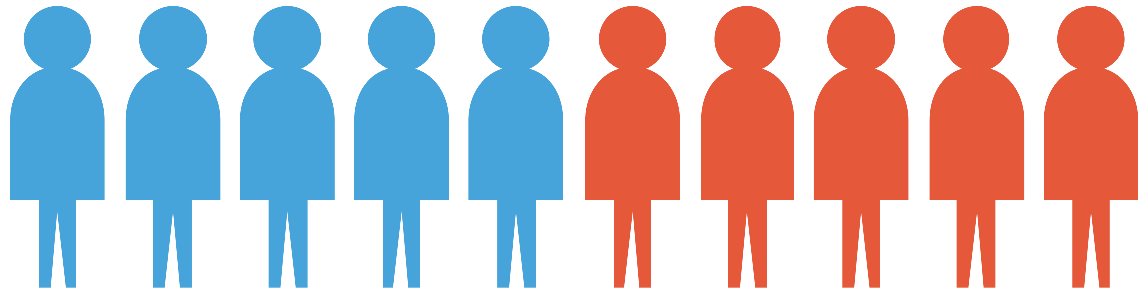
- Help business entities identify natural customer clusters to help them evaluate products design.
- Help business entities understand customer behavior patterns based on the artificial cluster label.

Dataset

- Kaggle: https://www.kaggle.com/sakshigoyal7/credit-card-customers
- 10,127 observations, 21 variables, 1 index, 6 categorical, 14 numerical
- Only 16.07% of customers who have churned
- There is no null value

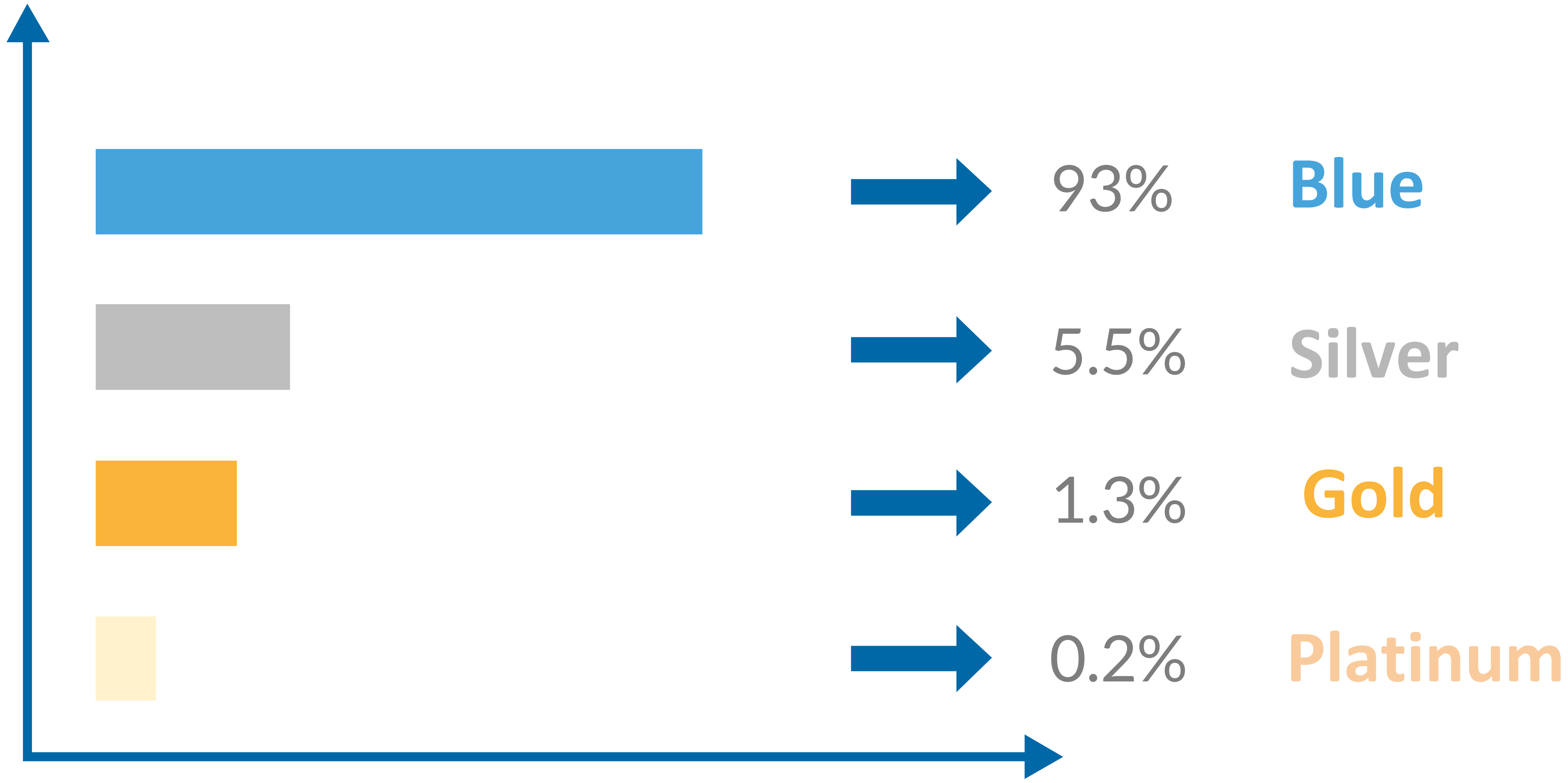
EDA

47 % Male/ 53%Female



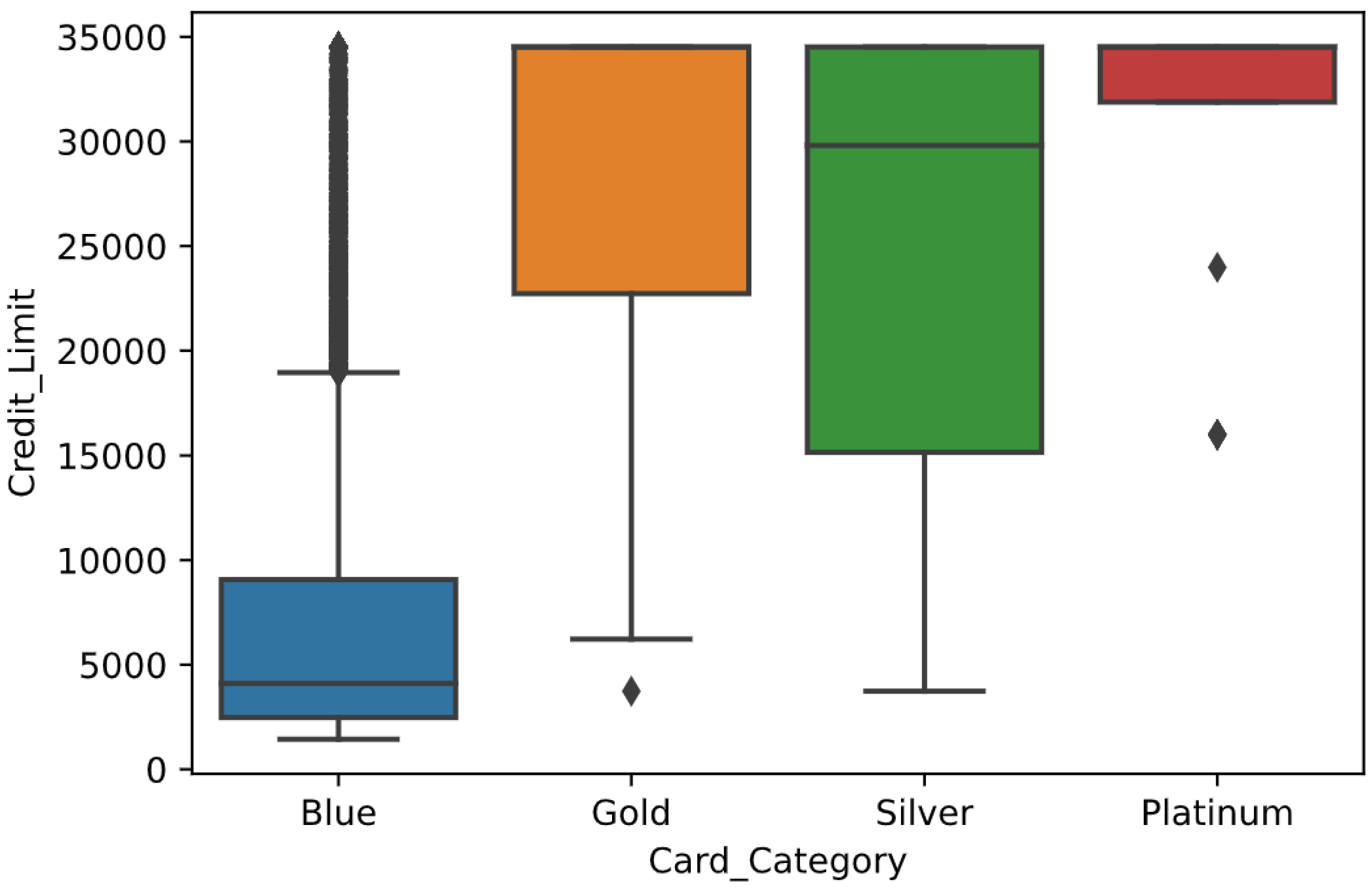
84% No churn/ 16% churn

Distribution of card types



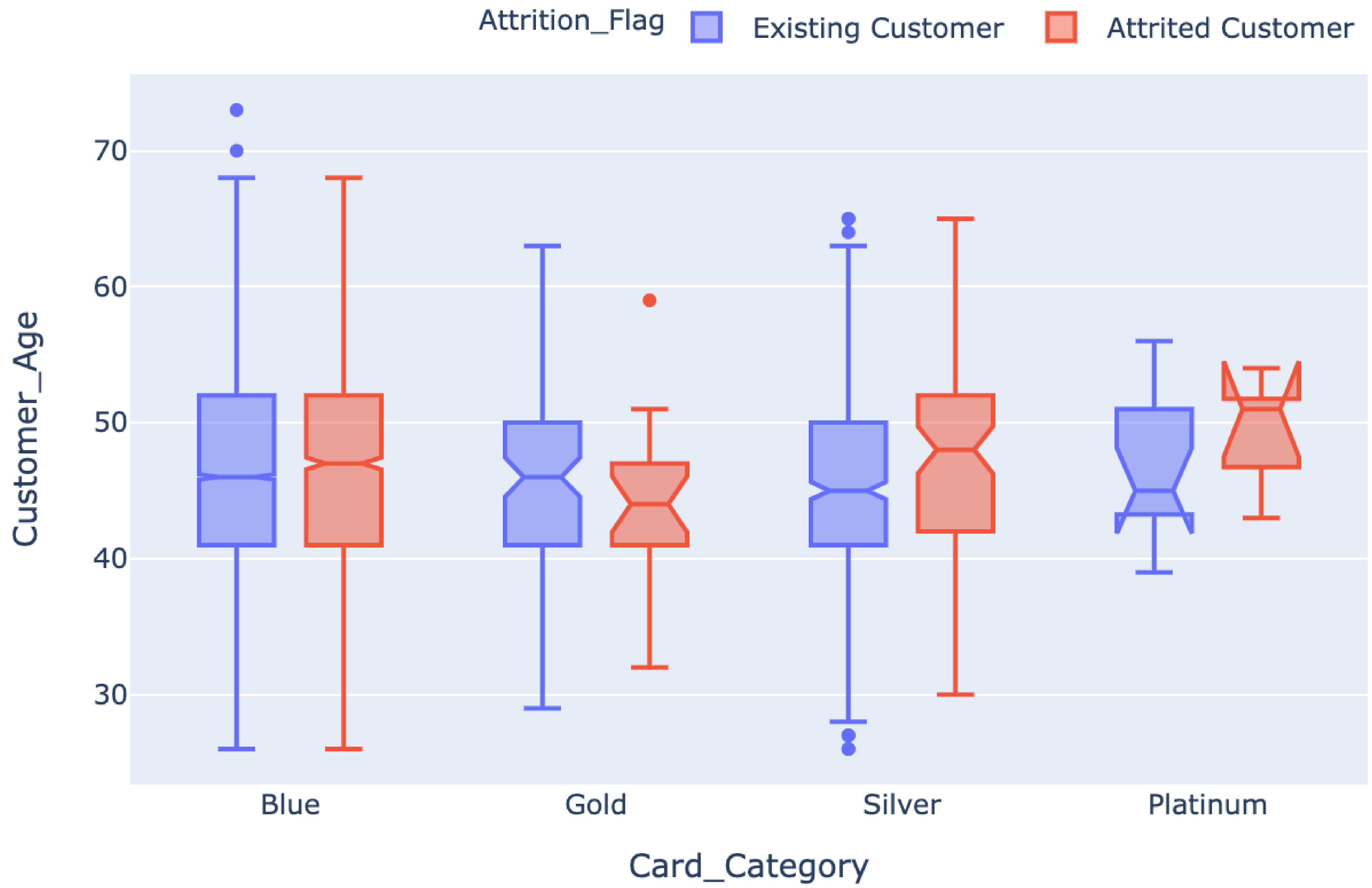
EDA

Credit Limit by Card Category



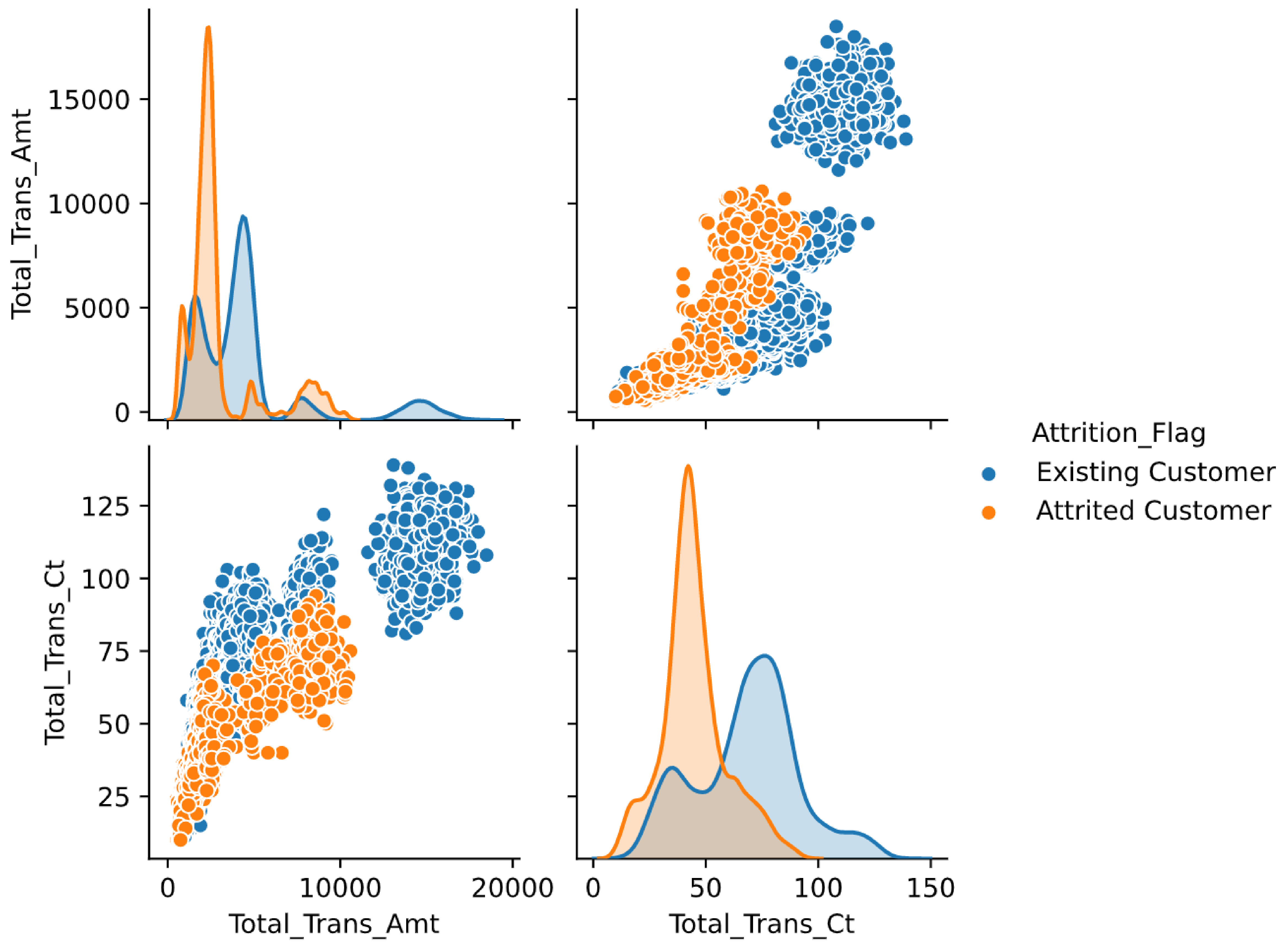
- Maximum credit limits for all cards is the same 34,516.

Customer Age by Card Category and gender



- The distribution of customer age is almost identical for all card types, around 45.

EDA



- Existing customer tends to have more transactions than attrited customers.
- Also, the existing customer tends to have higher total transaction amount.

Preprocessing

Numeric Columns

- Set client id as index
- Standardize

Categorical Columns

- Create dummy variables
- Target: Attrition Flag

Machine Learning - Card Type

Unsupervised Models

Blue Card

Gold Card

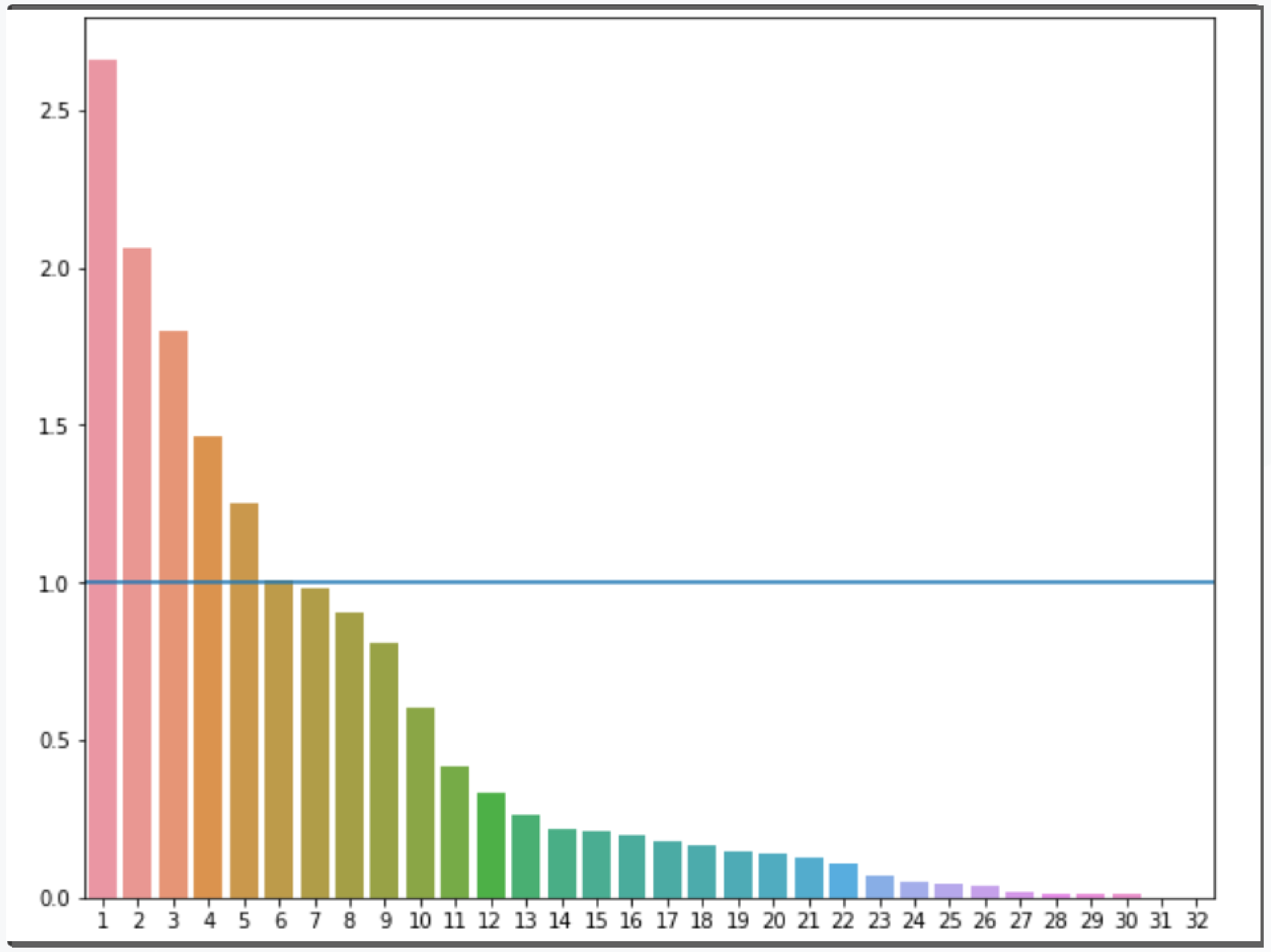
Silver Card

Platinum Card

PCA

Cumulative Variance

Reaching 90% Variance with 14 components

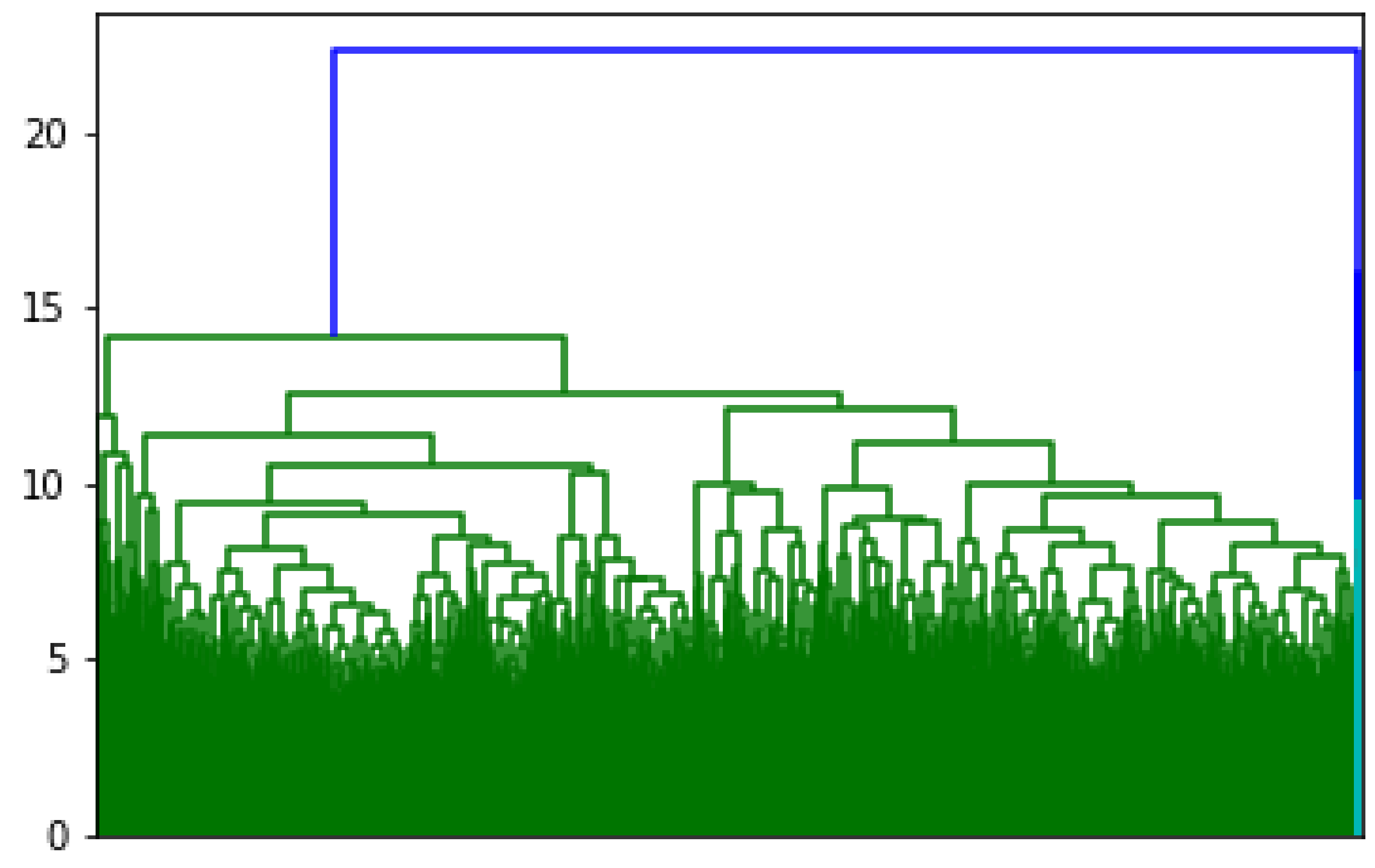


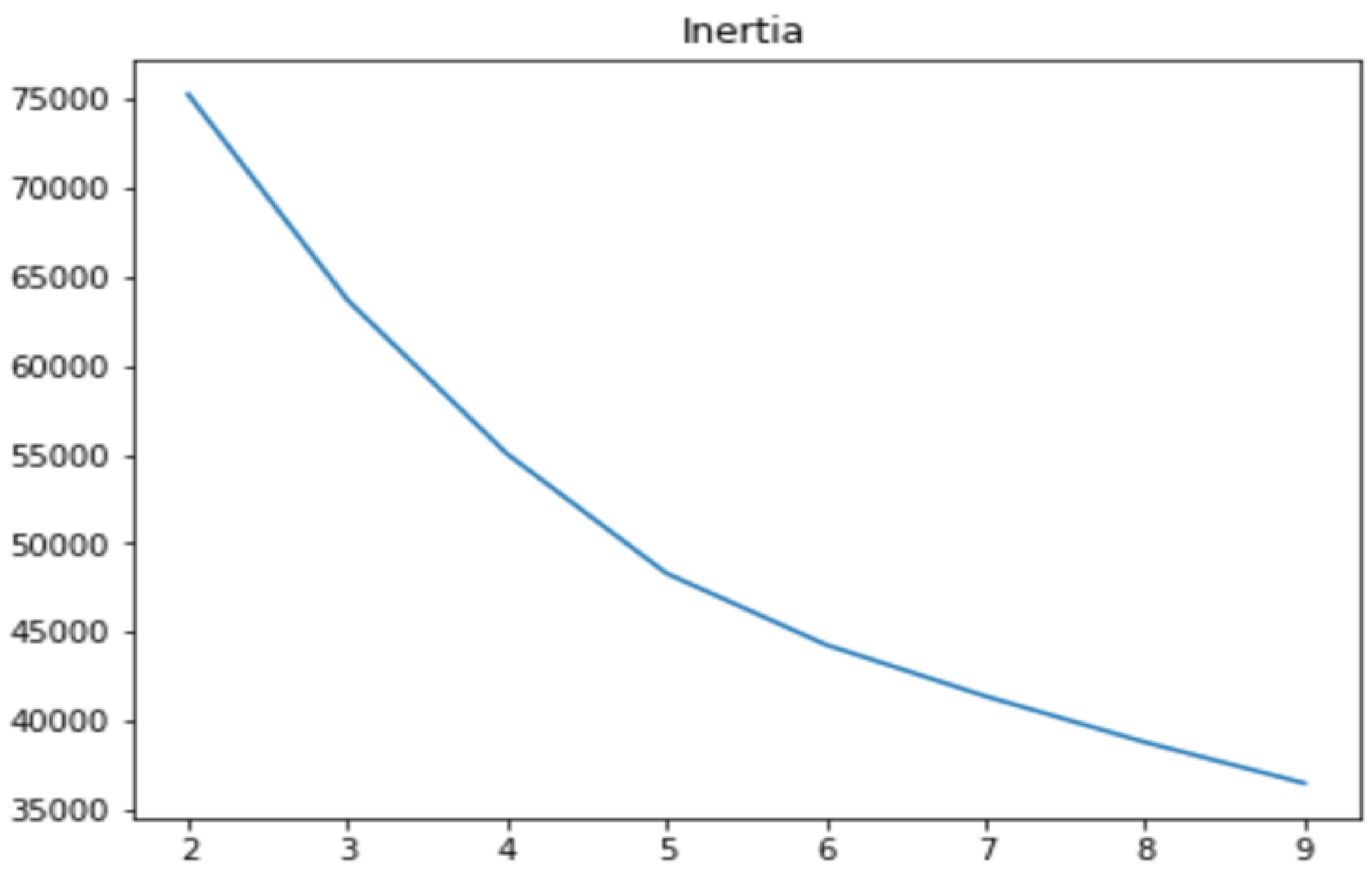
Components

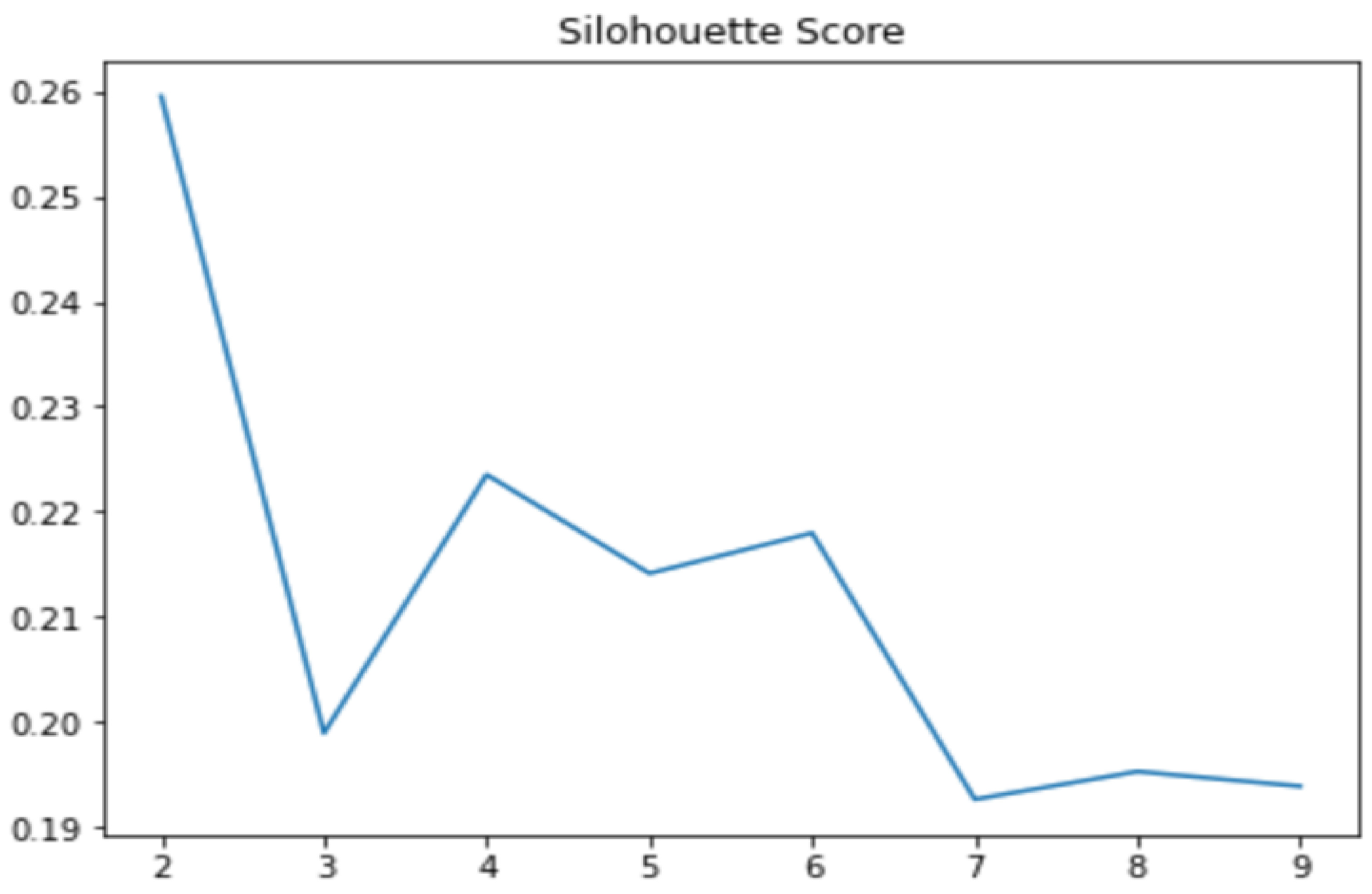
Keep 5 most significant components

PCA + KMeans

Unsupervised Models



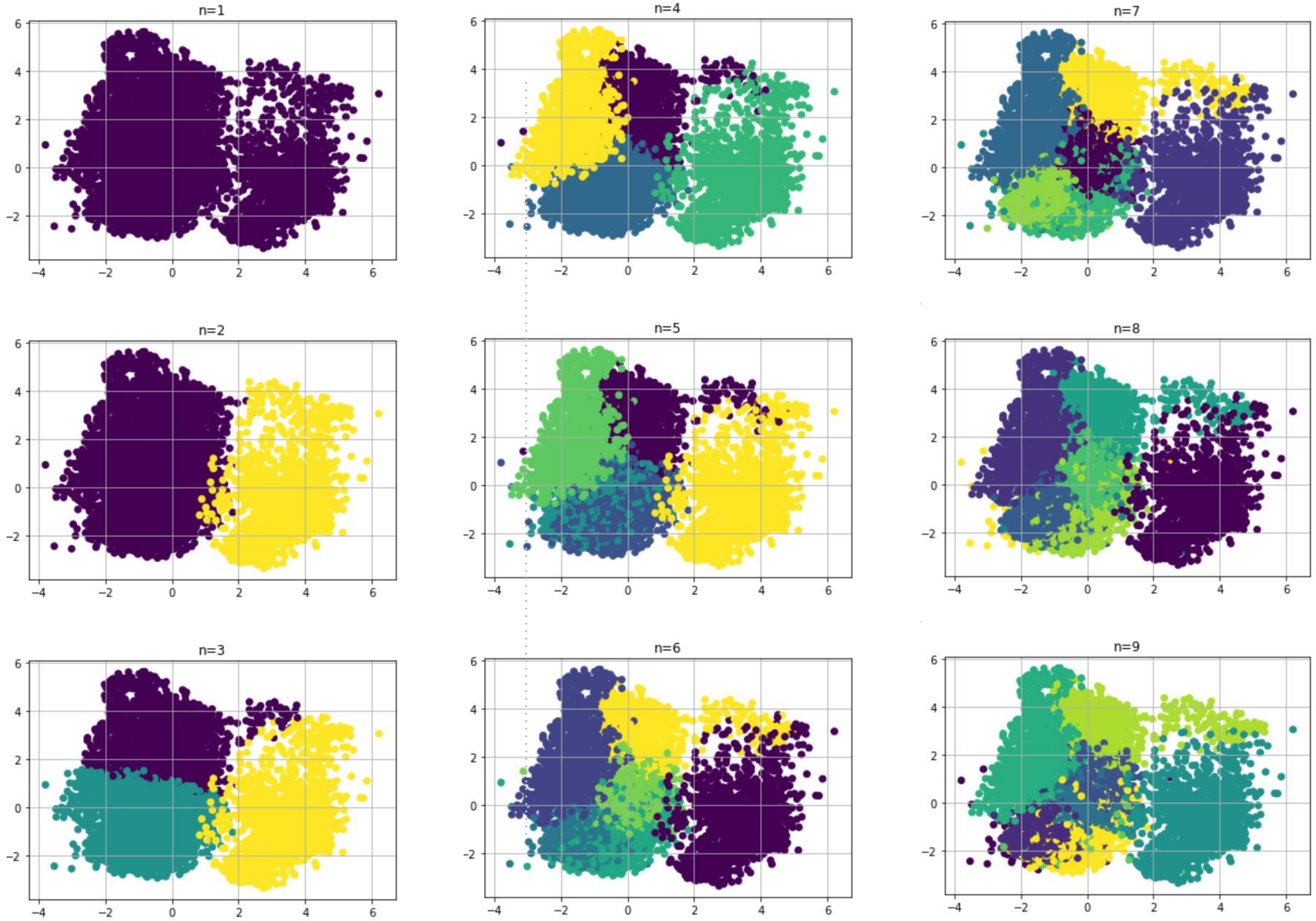




- k=4 is a good choice
- Low inertia + High silhouette score

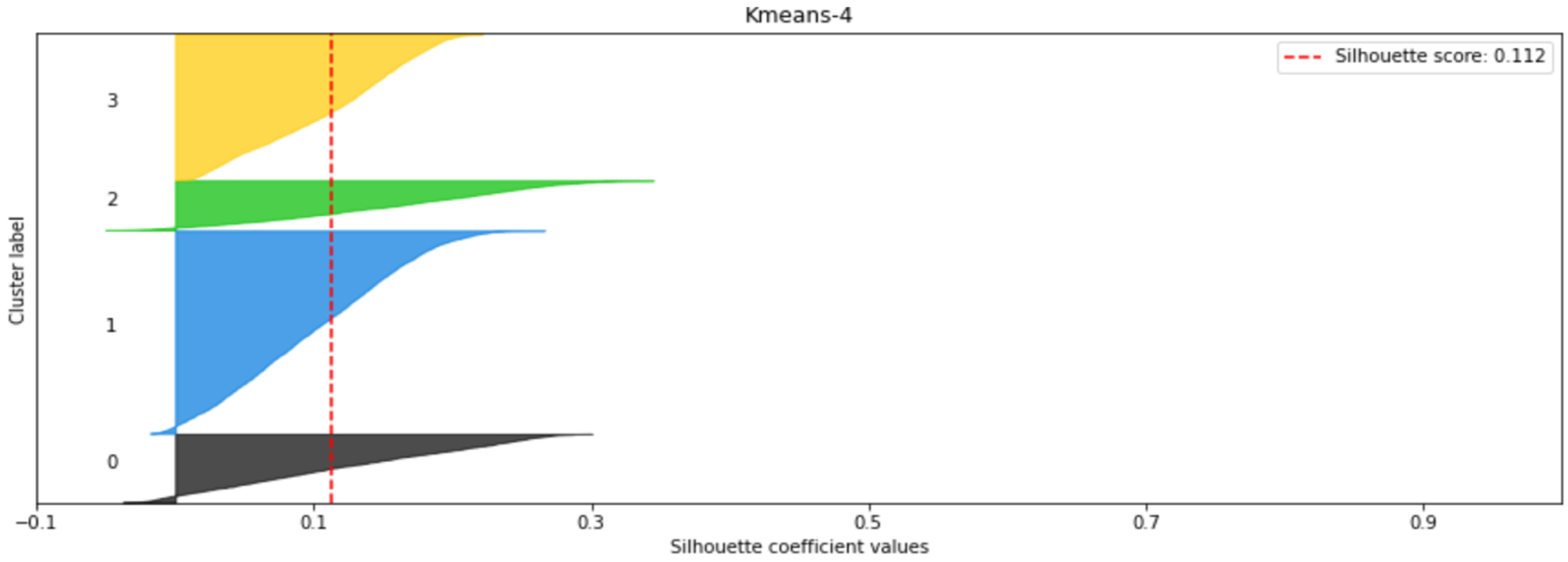
KMeans plot

Unsupervised Model



PCA + KMeans

Unsupervised Models



Blue	9436
Silver	555
Gold	116
Platinum	20

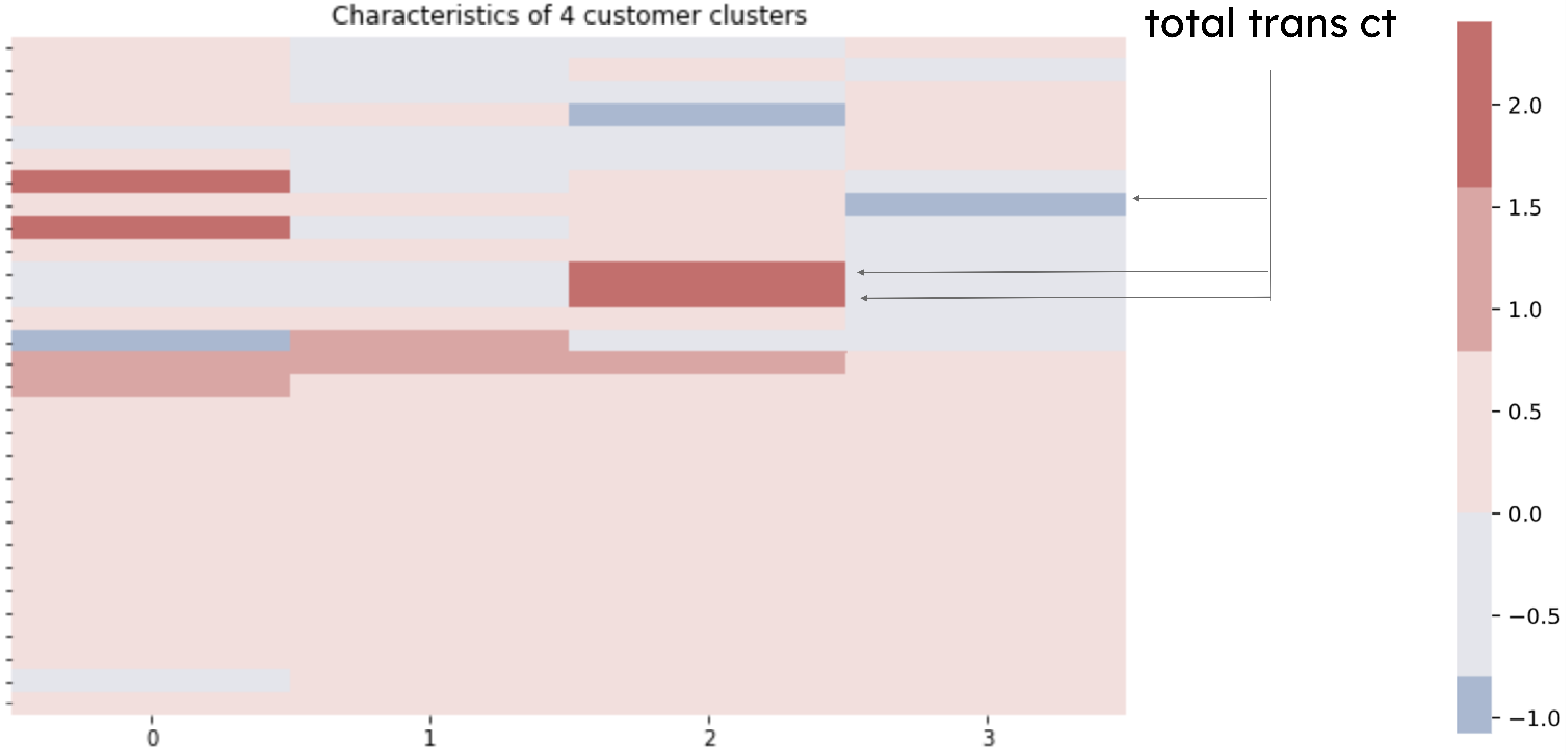
- Performed well overall

Results

Unsupervised Models

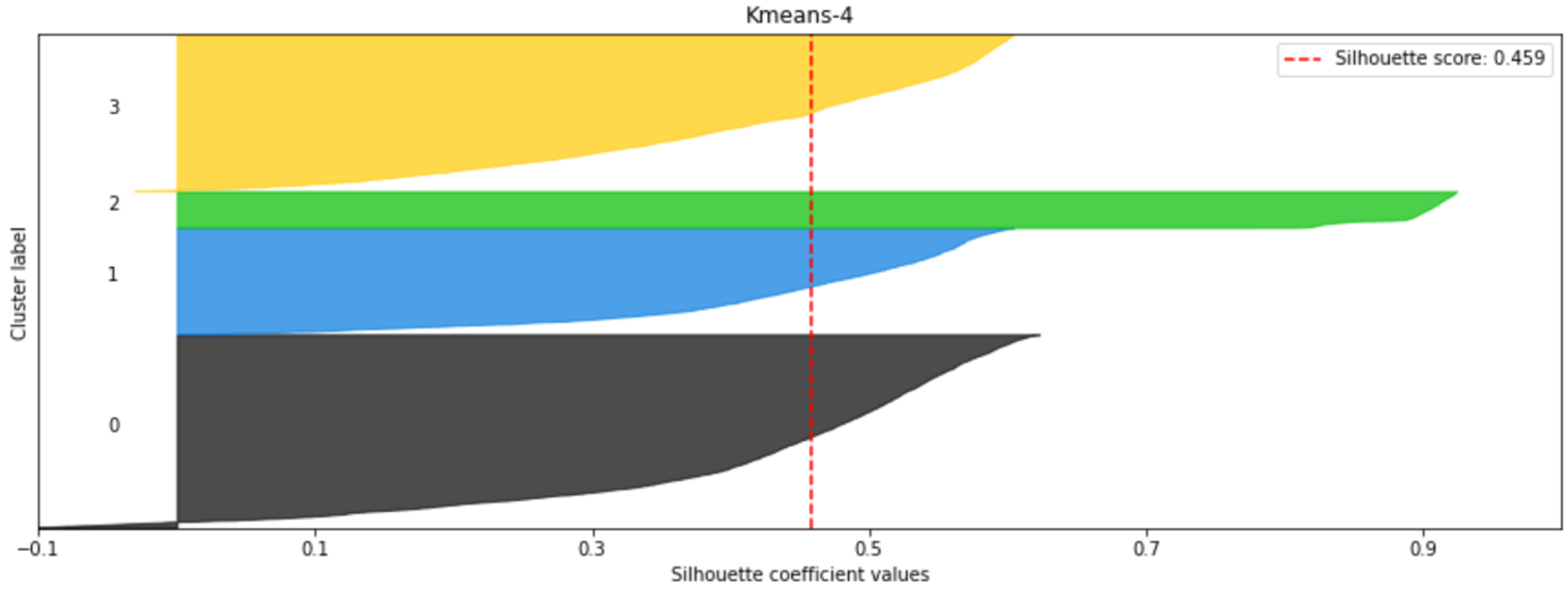
Credit limit
Avg open to buy
Avg utilization ratio

total revol bal
total trans amt
total trans ct



UMAP + KMeans

Unsupervised Models



- Fewer miscluster
- Higher silhouette score

Results

Unsupervised Models

total trans amt
total trans ct

Machine Learning - Churn or No Churn

Supervised Model

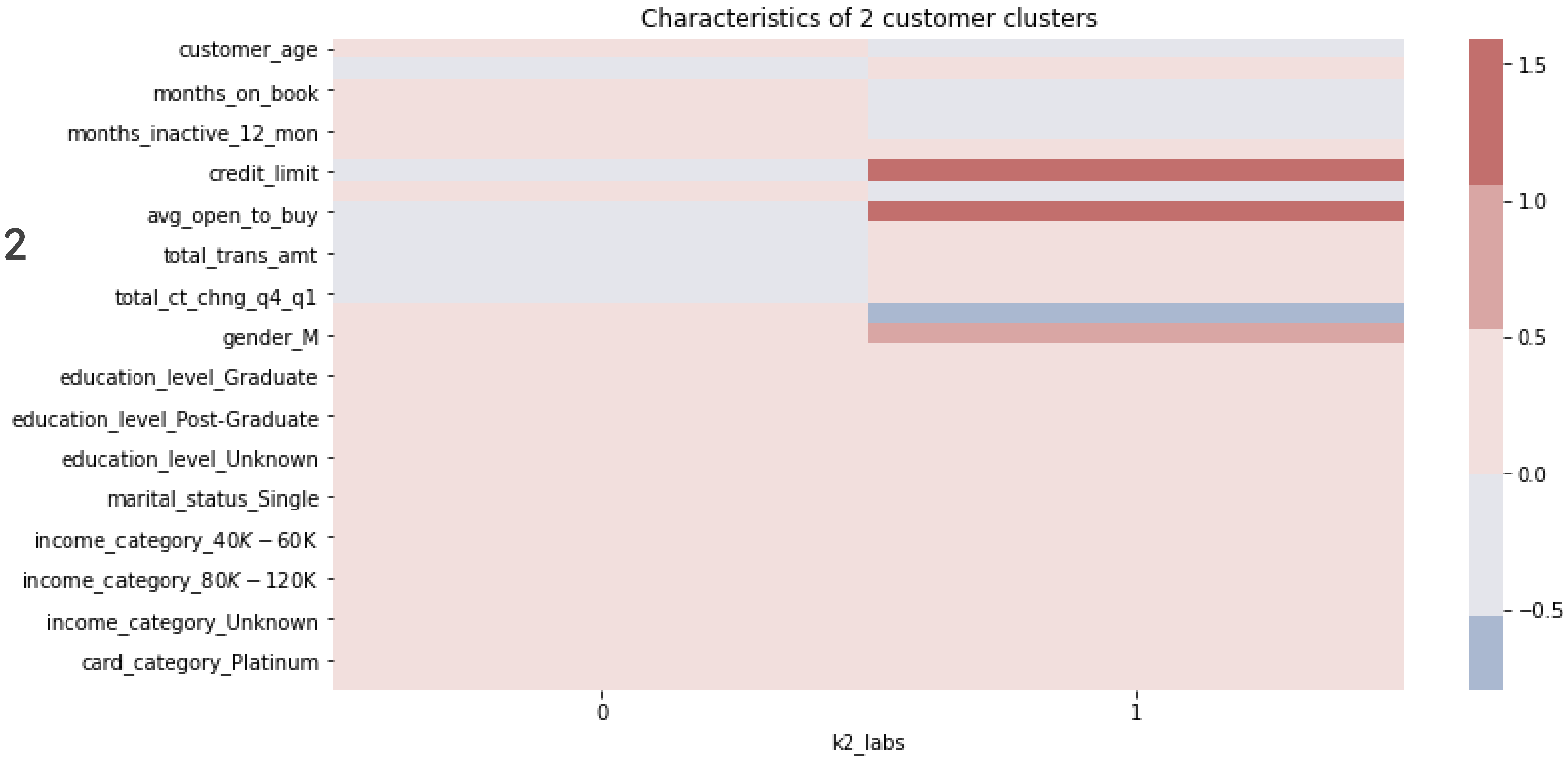
Attrited customers

Existing customers

Churn or No Churn

KMeans

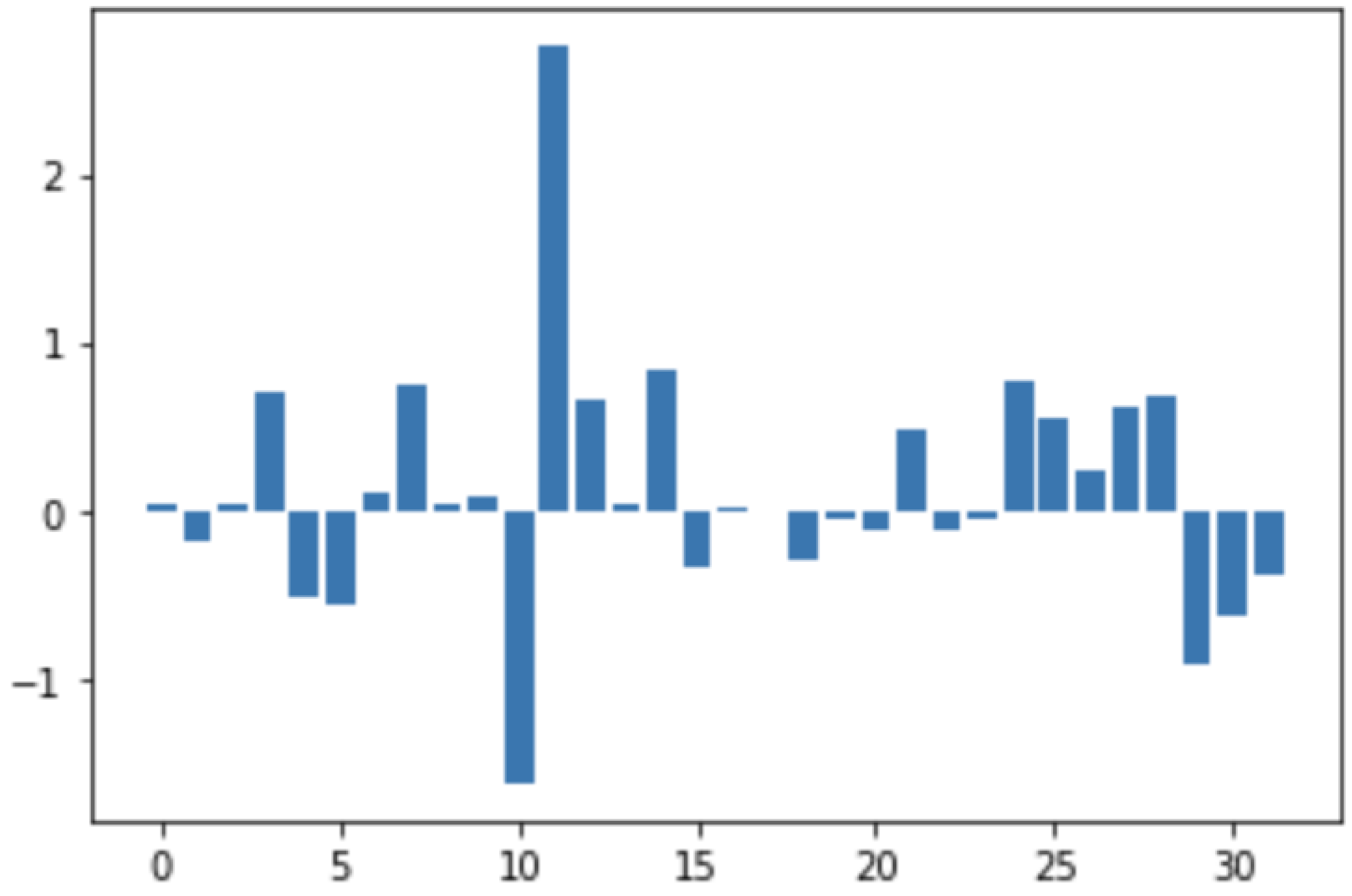
- Pool performance at k=2
- patterns not obvious



Supervised Models

Non PCA

F1 score
Logistic regression: 0.9
Random Forest: 0.88



total revolving balance
total trans amount
total trans count

Supervised Models

After PCA

F1 score

LogisticRegression: 0.93

RandomForest: 0.94

AdaBoost: 0.93

Gradient Boosting: 0.93

Conclusion

Non PCA

DBSCAN

Low accuracy according to unbalance variable

Logistic regression: 0.9

Random Forest: 0.88

Unsupervised methods | Supervised methods | Other optimization

After PCA ✓

K-means

No significant patterns for k = 2 and k = 4

LogisticRegression: 0.93

RandomForest: 0.94

AdaBoost: 0.93

Gradient Boosting: 0.93

TSNE

Pool perform ance

UMAP

Relatively high silhouette score: 0.45

Conclusion

Strengths

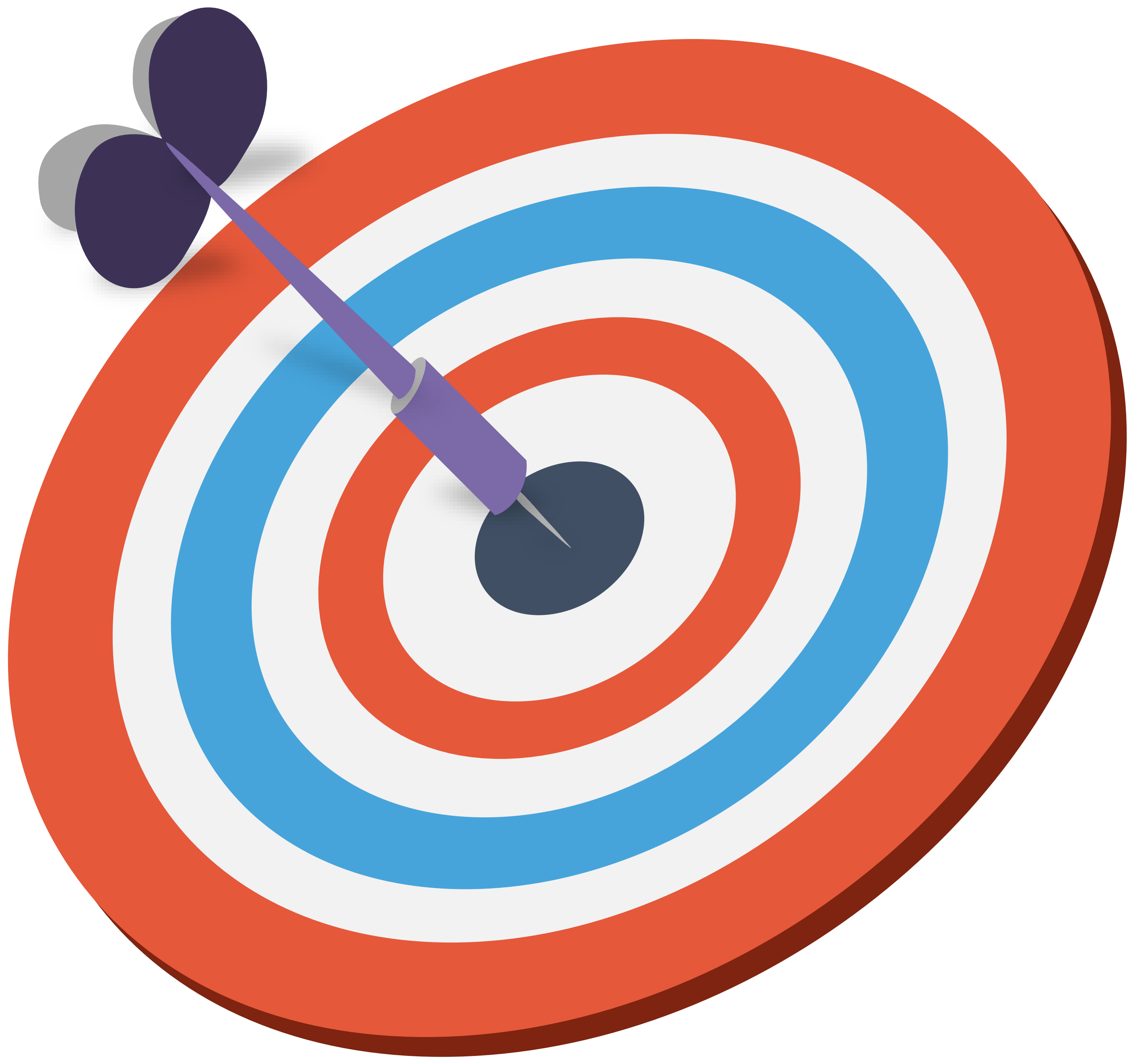
- Good silhouette score with Umap.

Weakness

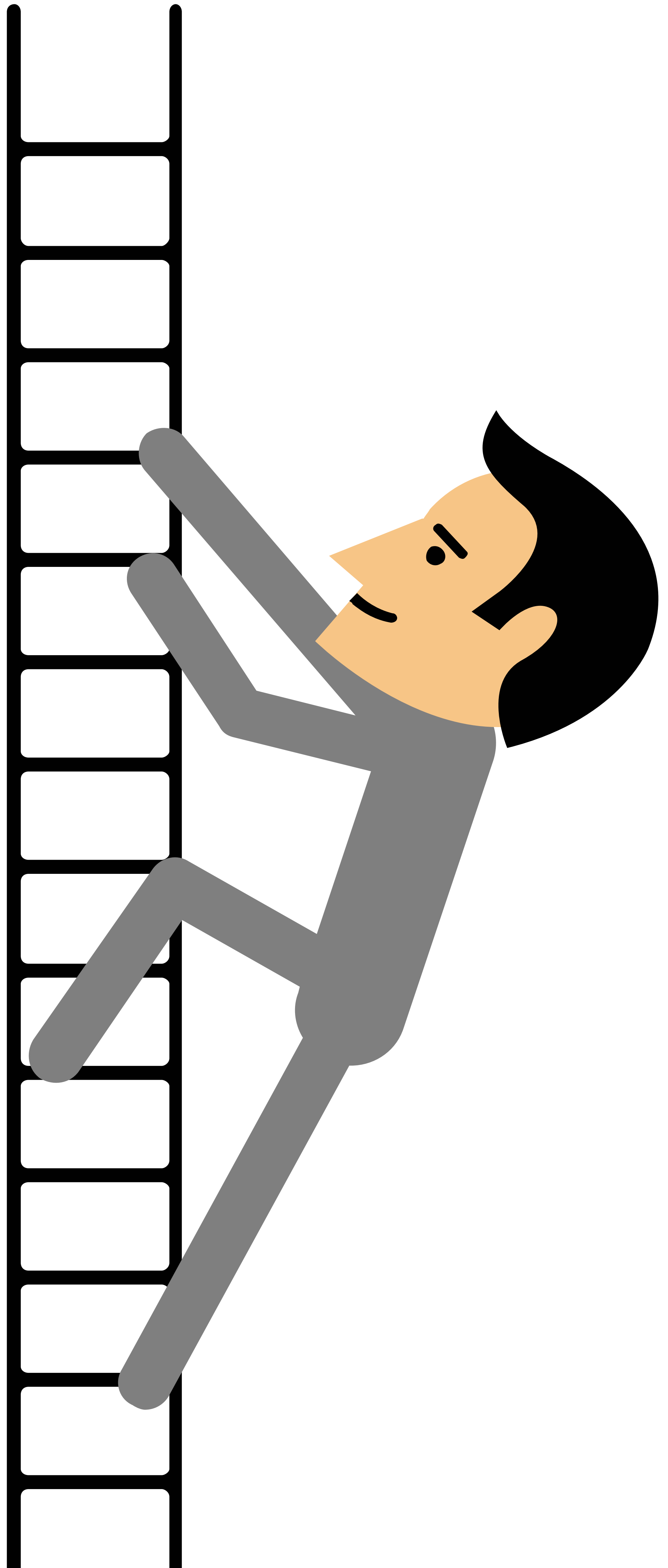
- Did not acquire ideal patterns for clustering.
- Result of Umap is not really interpretable even with a high silhouette score.

Choice

- Pick the result of K-means after PCA for card type clustering.
- Card type tend to be a more significant indicator.



Recommendations



01 Identifiable Card Type

No enough difference between different card types like Gold and Platinum.

02 Consider Gender

Gender difference account for the attrition condition.

03 Cooperation

Cooperate with business to launch promotion and policy aiming at different card users.

04 Customer relation

Customers with more relation with the bank tend to stay. Expand business with existing customers.

Q & A