Exploring and segmenting 3Rivers customers through machine learning

Ayomide Ajiboye [1] Anushka Bhat [1] Priscilla Chamenyi [1] Chimdia Kabuo [1] Pauline Korukundo [1] Grace Mahama [1]
Suchir Santosh Naik [1] Esther Okolie [1] Tanisha Raskar [1] Kofi Sarpong [1] James Soronnadi [1] Sai Teja Yarlagadda [1]
Supervisors: John Dever [2] Brent Hepner [2] Alessandro Maria Selvitella [1]

[1]Purdue University Fort Wayne [2]3Rivers Federal Credit Union

3Rivers Federal Credit Union

Since 1935, 3Rivers Federal Credit Union has been empowering our community to achieve financial wellness by offering personalized service, tools, and education. The member-owned, not-for-profit cooperative has $2.2 billion in assets, more than 110,000 members, 24 branches, and nearly 500 employees. 3Rivers offers a wide range of financial solutions, in addition to trustworthy, lifelong guidance and relationships. For more information, visit `3riversfcu.org`.

Project Description and Data Sources

Goal. Our goal is to segment 3Rivers customers into groups based on similar needs and behaviors. These clusters will be used to study 3Rivers customers for product recommendations. We will develop unsupervised learning algorithms (members are unlabeled) to create homogeneous and distinct subgroups of customers.

Data. De-identified datasets have been provided to PFW STAT 490 & MA 598 students by 3Rivers. These files include variables such as accounts held, balances, transaction information, and some demographic data. These files have been provided in RDS and HDF5 formats. Data dictionaries have been also provided with each file. Models have been built using either R or Python.

Study Methodology

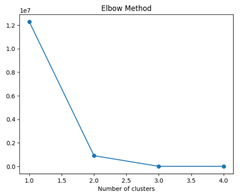
We used unsupervised learning algorithms to cluster 3Rivers customers. The data available is vast and heterogeneous and includes diverse data types: numerical and categorical variables, but also multi-variate time-series. The complicated dependency and high-dimensionality of the problem makes it hard to find meaningful clusters. *Pricinpal Component Analysis* (PCA) and K*-means* were developed to segment observations based on continuous variables, while K*-Prototype* was used when also categorical variables were considered. Methods that do not use pre-determined number of clusters, such as *Hierarchical Clustering*, were also implemented.

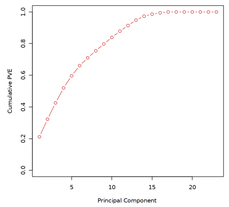
Research Objectives

- **Objective 1**: Describe, explore, understand the complex dependency of 3Rivers customers and the available variables.
- **Objective 2**: Segment 3Rivers clients for the purpose of product recommendation.

Elbow Method and PVE

The Elbow Method consists in determining the explained variation as a function of the number of clusters and picking the elbow of the curve as the optimal number of clusters to use. The Number of principal components is often selected using Proportion of Variance Explained (PVE).

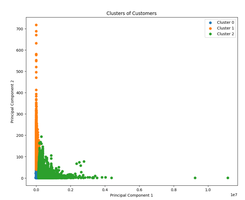


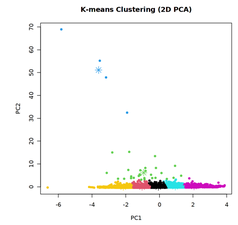


Results: KLIs, Transactions, & Monthly Balance

Key Lifestyle Indicators (KLIs) . Segmenting Savings Account customers into three distinct groups was achieved by categorizing KLIs' variable into 11-factor levels [Left Figure]. The top 10 most frequent KLIs in the dataset were identified, while all other indicators were grouped into a category labeled 'Other.' Utilizing this refined categorical variable, as well as the Monthly Balance data, we identified 3 customer segments. These customer segments are delineated by the frequency of transactions on an account, the membership period, and the monthly account balances maintained along with the KLI.

The following variables contribute to each of the 3 customer segments. *Cluster 1*: The variables that contribute to the first segment include: Primary account holders, Employed, Competitive Bill Pay, Home Improvement Store Patron, Deep Discount or Dollar Store, Ice Cream Lover, Delivery and Takeout Restaurant Customers, as well as the difference between their maximum and minimum account balance. *Cluster 2*: The customers that are not primary account holders, have Other KLI, the difference in their monthly account balance and their period of membership with the bank contribute to this cluster. *Cluster 3*: The Customers that are primary account holders, and their average balance, number of transactions, and those that have 'Other' as their KLI.





Transactions & Monthly Balance. Clustering customers based on their average monthly balance and debit frequency results in: *Cluster 0: Constant or ideal customers whose average running balance and debit transaction seems loyal*; *Cluster 1*: Does frequent POS transactions; and *Cluster 2*: Has higher average running balance and stores more money over time. [Middle Figure].

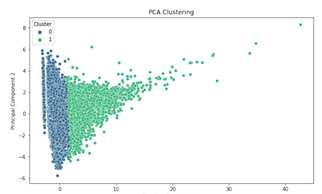
Using transactions and customer tables, we obtain 7 separated clusters [Right Figure], while Hierarchical clustering did not show well-separated clusters. This algorithm required filtering first for features with variance greater than 100 to avoid for algorithmic convergence purposes. The final features utilized in the K-means clustering are as follows: *average number of debit transactions per month*, *average monthly debit amount*, *average monthly credit amount*, *average number of credit transactions per month*, *age*, *age at joining*, *closest branch distance*, *current credit score*, and *years since joining*.

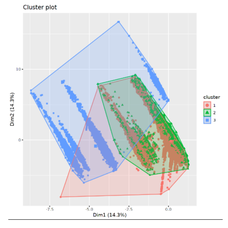
Acknoweldgements

This project is part of the Indiana Data Mine initiative at Purdue University Fort Wayne. The work has been performed by the students during Spring 2024 as part of their class work for STAT 490 and MA 598 Corporate Partners II, sponsored by 3Rivers Federal Credit Union and instructed by Prof. Selvitella. Prof. Selvitella acknowledges the support of Lilly Endowment Inc. through the "Indiana Data Mine" grant. We thank 3Rivers for the generous support and John Dever and Brent Hepner for the kind guidance and contributions.

Results: Lending, Census, Customer to Account, & Monthly Balance

Lending Table. *Cluster 1*: Cluster 1 consists of customers characterized by moderate interest rates of 7.6%, relatively lower monthly payments of around 325.43 dollars, and smaller loan amounts averaging $18,315.80. These individuals may represent a segment seeking manageable financial commitments over a moderate repayment period of approximately 64.45 months. With a positive FICO score change averaging around 16.76 points, they demonstrate an improvement in creditworthiness. *Cluster 2*. Cluster 2 comprises customers with lower interest rates 0f 3.8%, significantly higher monthly payments averaging 1,274.32 dollars, and larger loan amounts averaging $189,785.60. These individuals likely have substantial borrowing capacity and may be pursuing significant purchases or investments, reflected in the extended repayment period of approximately 304.04 months. Despite the positive FICO score change averaging around 16.30 points, they might prioritize long-term financial planning and larger-scale ventures.





Census, Customer to Account, & Monthly Balance. *Clusters 1, 2, and 3*: High account balances; inactive accounts with primary holders; checking accounts with high transaction volume; high-income customers; more likely to have children; more investment accounts with unknown status. *Cluster 4*: Medium/low balances; inactive/active accounts; primary/non-primary holders; Checking/Savings; high transaction volume; high and medium income; more likely to have children; more likely to be retired; investment/non-investment, known/unknown status. *Cluster 5*: Medium/low balances; inactive/active accounts; primary/non-primary holders; more Savings with low transactions; slightly above medium income; primary/non-primary holders; more > 65. *Cluster 6 and 7*: Very low balances; active accounts with non-primary holders; savings accounts with low transactions; low-income customers; predominantly non-primary holders; predominantly > 65; non-investment accounts with known status.

Conclusions

In this research, we explored and segmented 3Rivers customers data using unsupervised learning algorithms. The problem we attacked was complex and high-dimensional. The analysis faced computational challenges and required skilled insights and the development of fast algorithms suitable for big data. In many of the implementations, it was crucial to strike a balance between feature significance and computational feasibility. We were able to find hidden structures in the data and group 3Rivers customers based on low dimensional representations of the data. We favoured interpretable clustering methods and more flexible models, such as DBSCAN, Spectral Clustering, will be explored in future endeavours.

Lilly Endowment Inc.
A private foundation since 1937