Unsupervised Learning Capstone

Finding Customer Segments for a Credit Card Company

Megan Dibble

Introduction

Motivation & Explanation of Data

Motivation

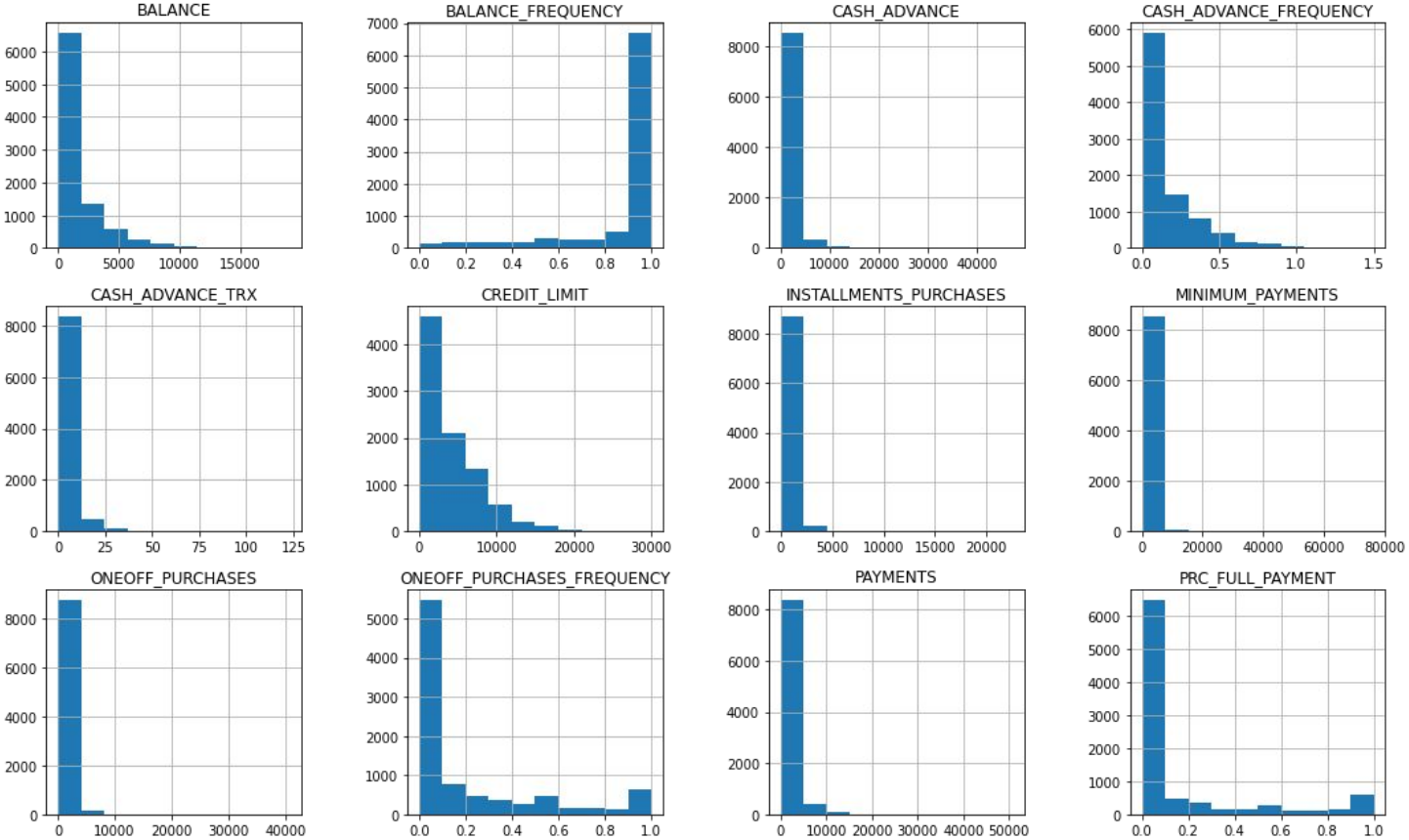
- Credit Card Company has collected data on customers
  - Need marketing strategy
- No prior indication of customer segments
  - Application for clustering models
- Customizing marketing to a “customer type” (or segment)
  - Could increase profit
  - Customers more likely to sign up for promotional offers
- Goal: use segmentation to increase profit & customer satisfaction

Explanation & Structure of Data

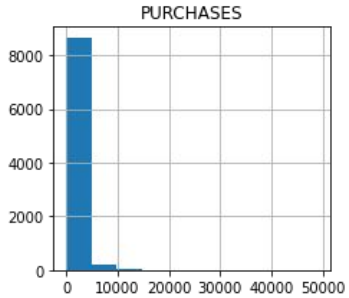
- Dataset summarizes the usage behavior of about 9000 active credit card holders
  - During the last 6 months
- Source: Kaggle.com
- 18 columns, all continuous variables except unique Customer ID
- 8950 rows

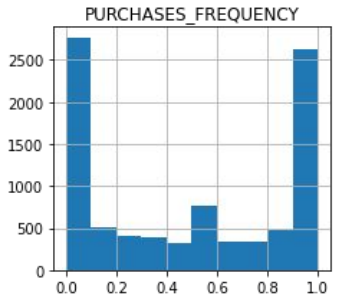
Data Exploration & Cleaning

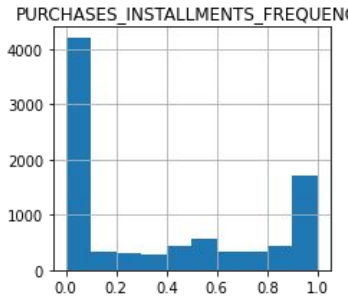
Distributions

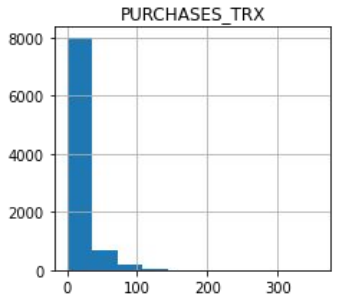


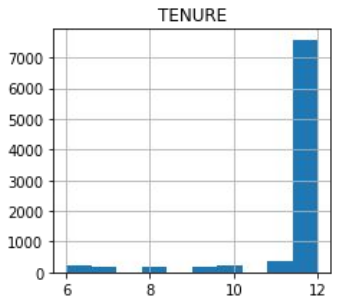
Distributions











Takeaways

- Lots of distributions are skewed right,
  - Outliers that may need to be dealt with later
  - Left them for now
  - Did not appear to be errors after deep dive
- PURCHASES_FREQUENCY --not a lot of customers are in the middle
  - Could be a potential cluster split
- 75% of customers have less than 1,110 purchases during the last 6 months
  - Less than 185 purchases/month on average
- Minimal null values (3% or less for 2 variables)
  - Filled with median or mode, depending on the variable

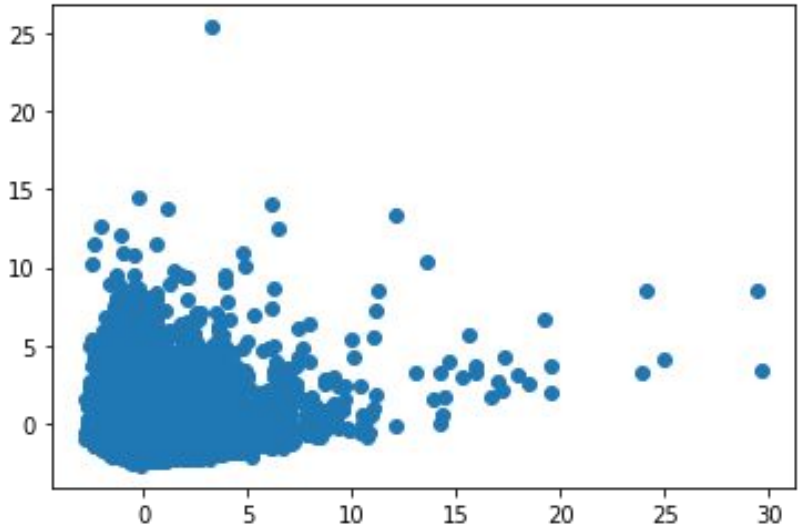
Modeling Process & Visualizations

Model Inputs, Visualizations, & Results

Model Inputs

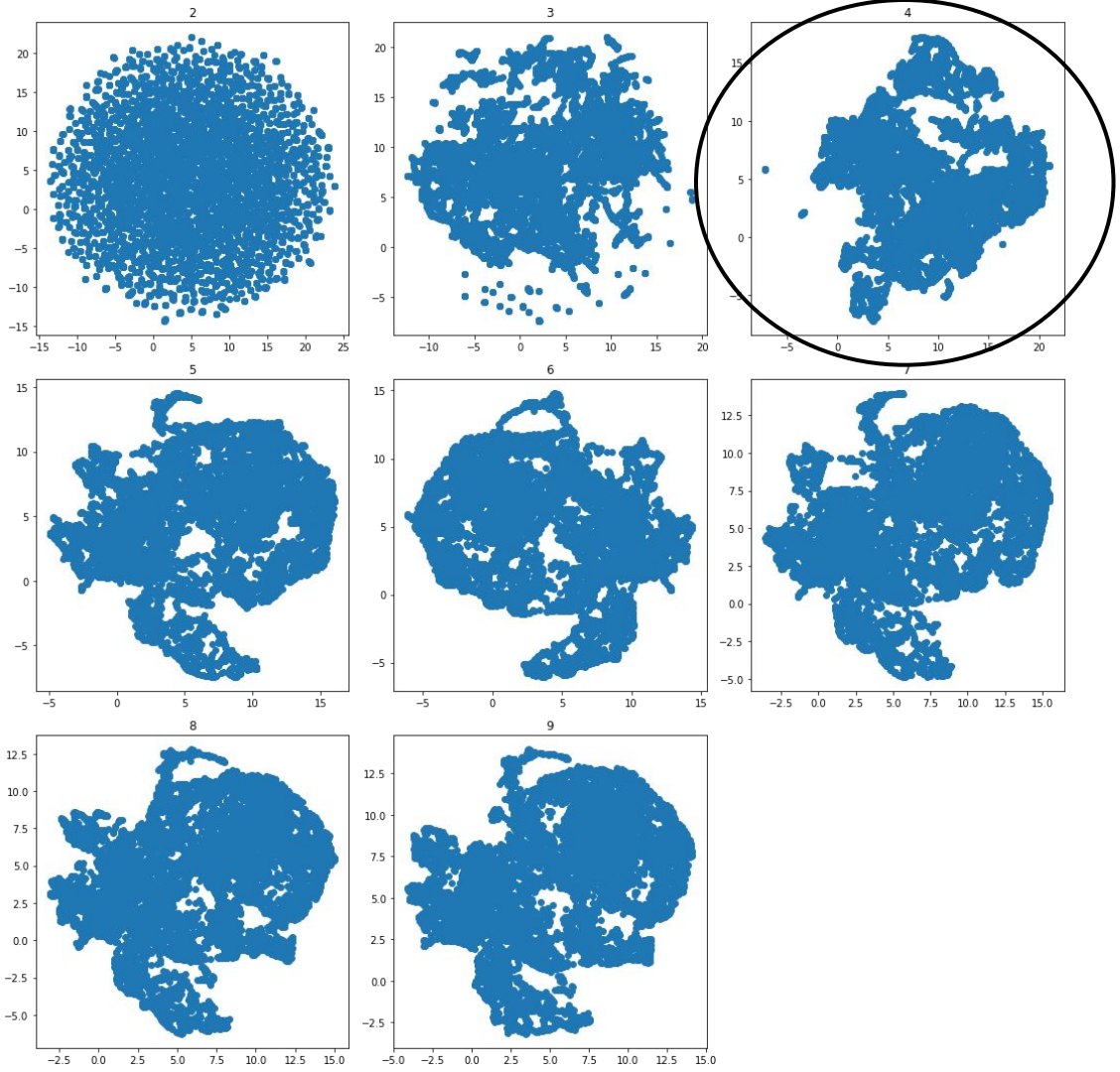
- Feature engineering: standardized all variables
- Data shape: 8949 observations and 18 features

Visualizing the data in 2D with PCA:



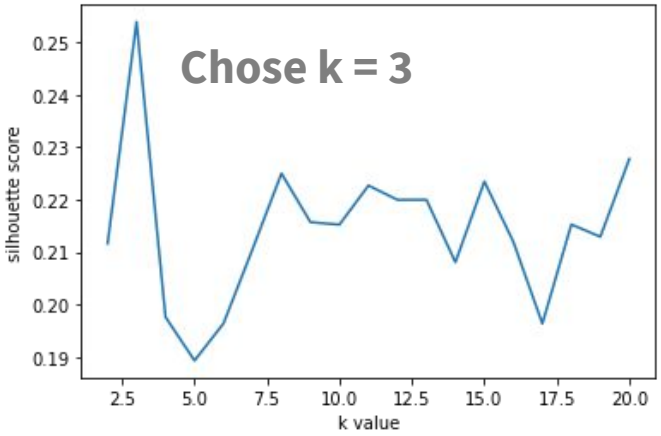
Visualizing the Data with UMAP

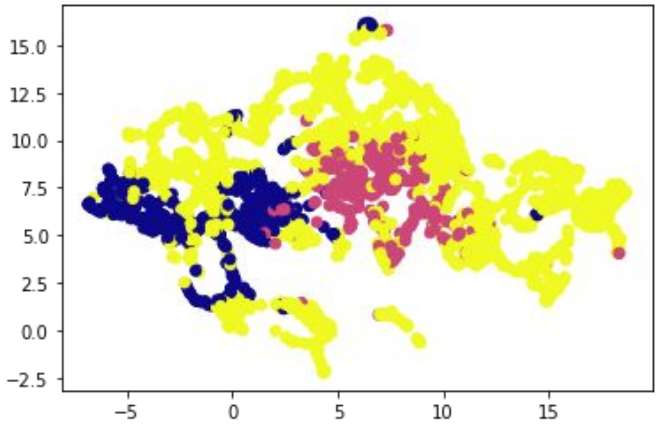
- Small numbers at top of graphs are tuning parameter n_neighbors
- Looking for most distinct clusters
- chose n_neighbors = 4



K Means

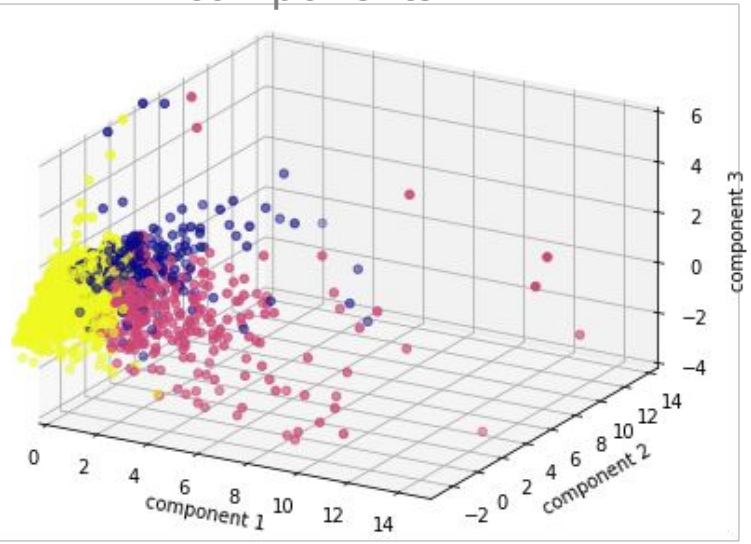
Silhouette Score = .254

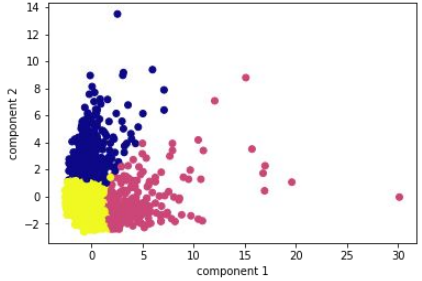




UMAP

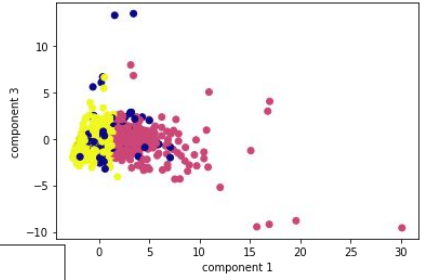
PCA with 3 components

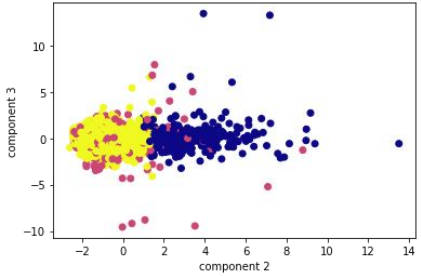




X-Y axis

Y-Z axis





X-Z axis

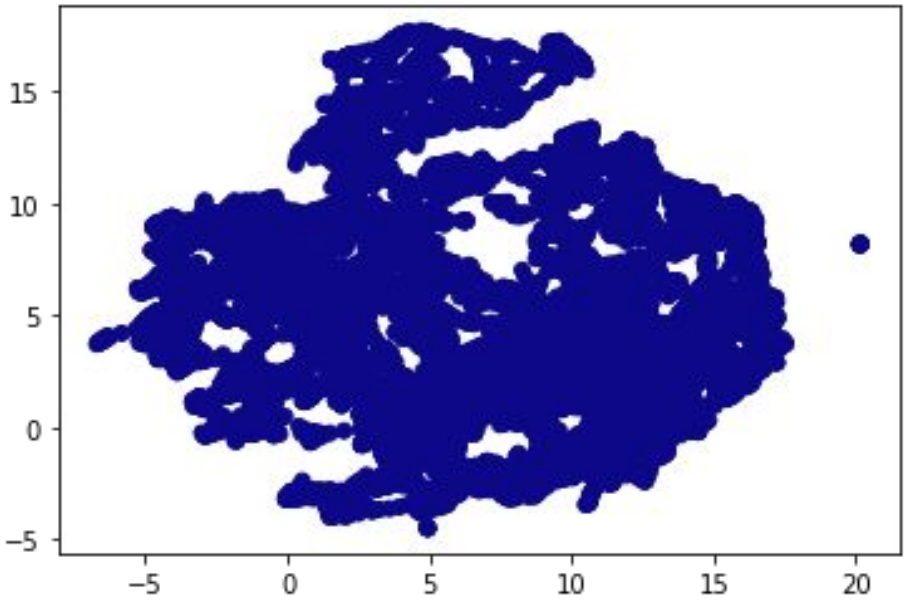
Hierarchical Clustering

Silhouette Scores:

- Complete: 0.787
- Ward: 0.181
- Average: 0.841

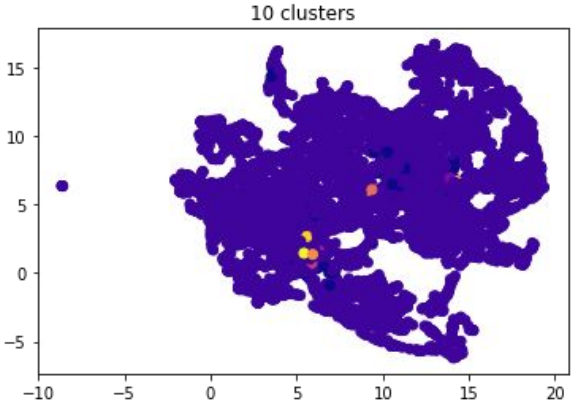
Average method found 2 clusters

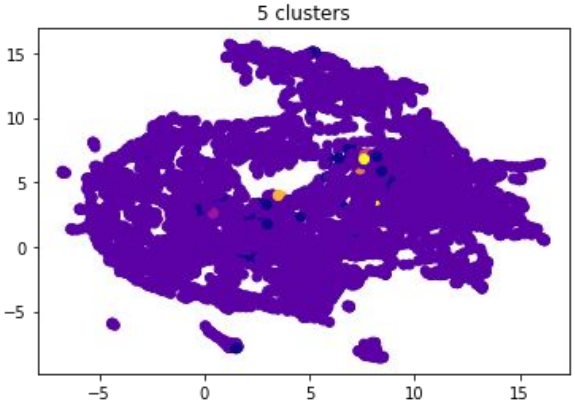
- Good silhouette score, but model is not informative
- **Data is not hierarchical in nature**

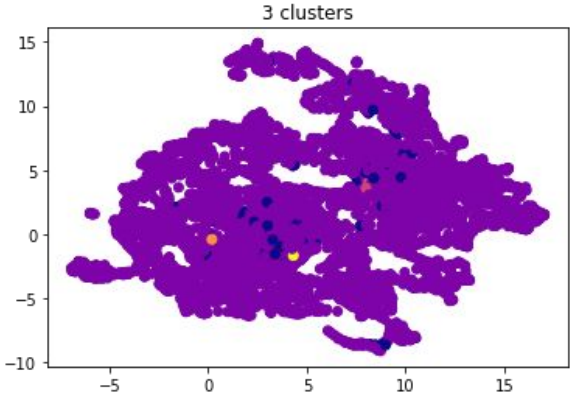


UMAP (2 clusters)

DBSCAN





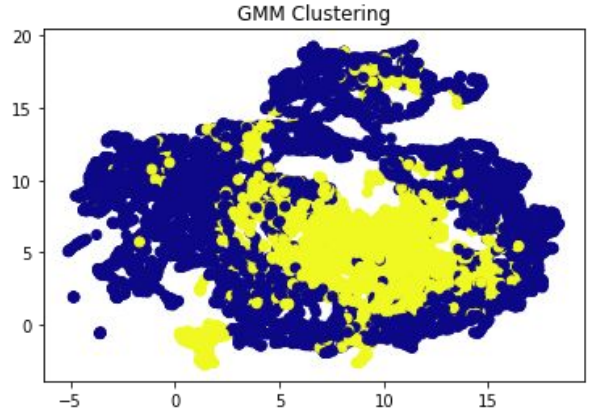


Originally DBSCAN found 156 clusters

- Tuned min_samples hyperparameter
- Graphs (left to right) are min_samples = 2, 3, 4
- **Not a good model choice for the data**

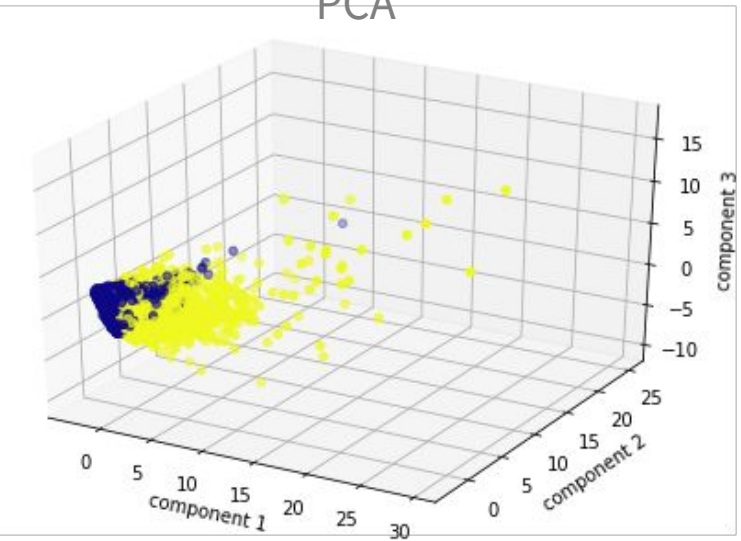
GMM Clustering

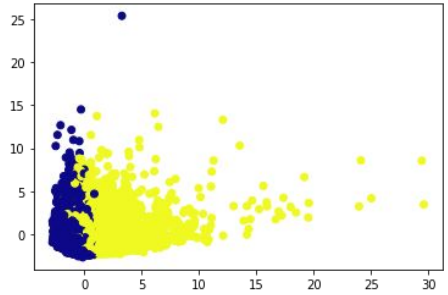
After tuning, n_components = 2 gave the highest silhouette score (**.185**)



UMAP

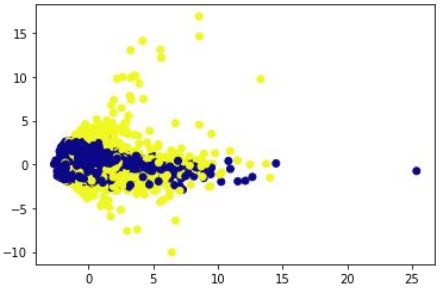
PCA

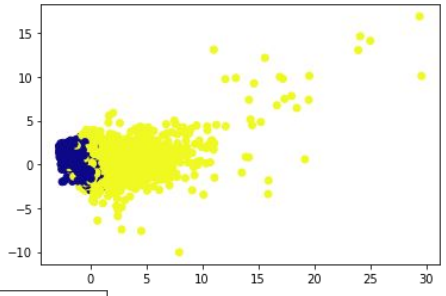




X-Y axis

Y-Z axis



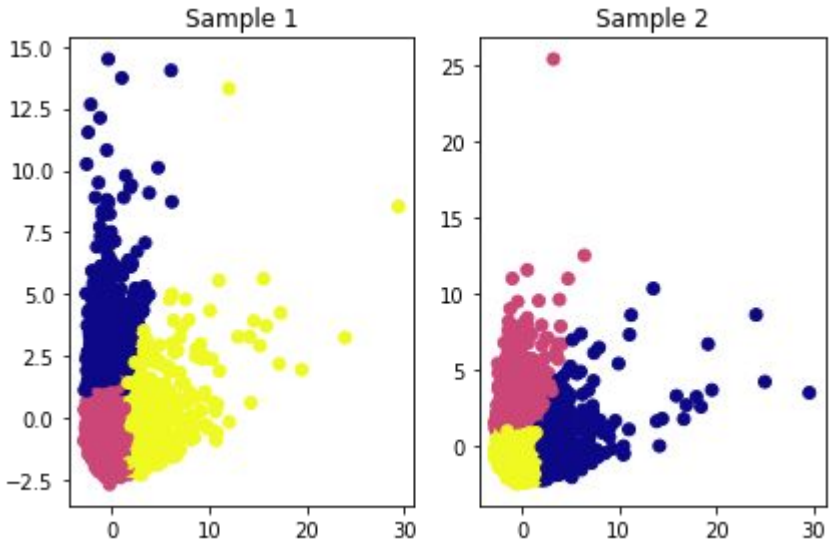


X-Z axis

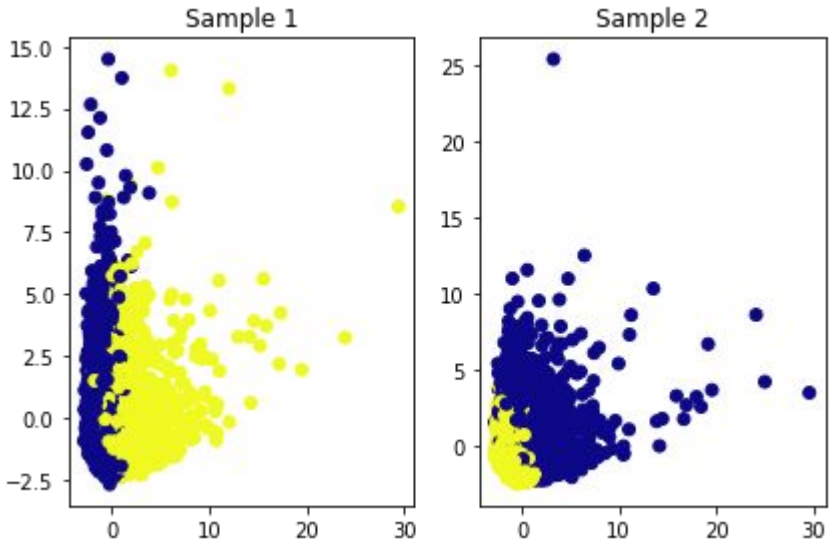
Evaluating Consistency

- Comparing K-Means and GMM, K-Means was more consistent
- Split data evenly and then used PCA to visualize clusters

K Means



GMM



Conclusion

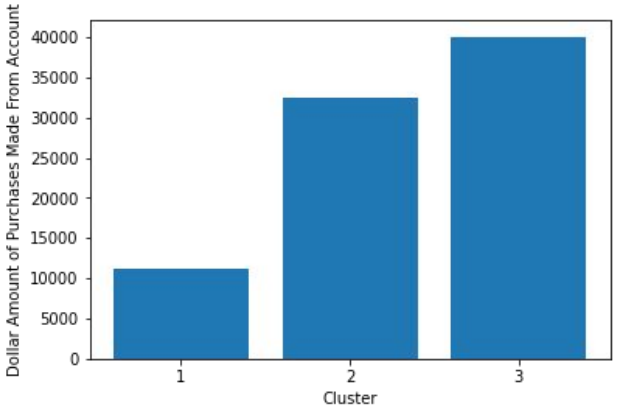
Business Impact & Recommendations

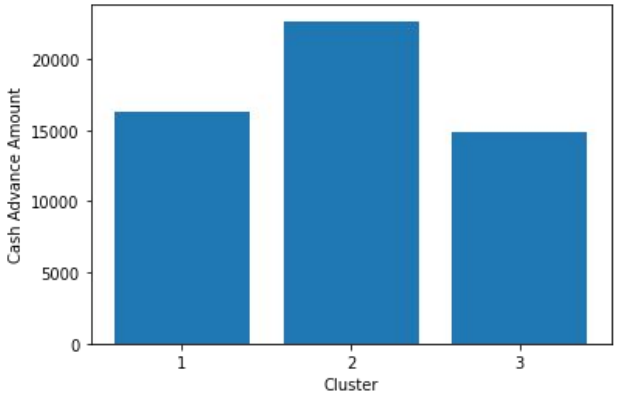
Business Implications of Results

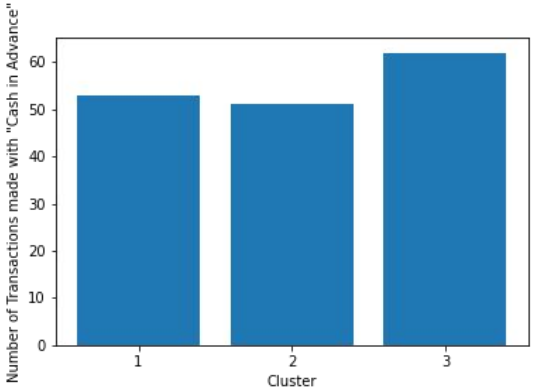
Fed data labeled with clusters into a decision tree

Variable	Importance
CASH_ADVANCE	0.143418
PURCHASES	0.133598
CASH_ADVANCE_TRX	0.114313

- Cluster 1: "small spenders"
- Cluster 2: "medium spenders with large cash advances"
- Cluster 3: "big spenders"







Recommendations

- Personalize marketing to each cluster with different credit card offerings
- Collect data on effectiveness of marketing strategy
    - Evaluate change
- If I had more time
    - Go back and spend time to remove or set threshold on outliers
    - Work on more tuning to increase silhouette score
    - Evaluate sensitivity of random forest feature importance

Thank You
For Listening!

Any Questions?