Bank Customer Churn Segmentation

Xiyi, Ye Joon, Stella, and Cailey

Data

	CLIENTNUM	Attrition_Flag	Customer_Age	Gender	Dependent_count	Education_Level	Marital_Status	Income_Category	Card_Category	Months_on_book
0	768805383	Existing Customer	45	M	3	High School	Married	$60K-80K	Blue	39
1	818770008	Existing Customer	49	F	5	Graduate	Single	Less than $40K	Blue	44
2	713982108	Existing Customer	51	M	3	Graduate	Married	$80K-120K	Blue	36
3	769911858	Existing Customer	40	F	4	High School	Unknown	Less than $40K	Blue	34
4	709106358	Existing Customer	40	M	3	Uneducated	Married	$60K-80K	Blue	21

- Size: 10127 rows x 23 columns
  - No null values
- Given labels: Existing or Attrited Customer
- Sufficient for Data Mining Techniques
  - Customer Attributes
  - Can predict label from the labels
  - Recency and Frequency metrics

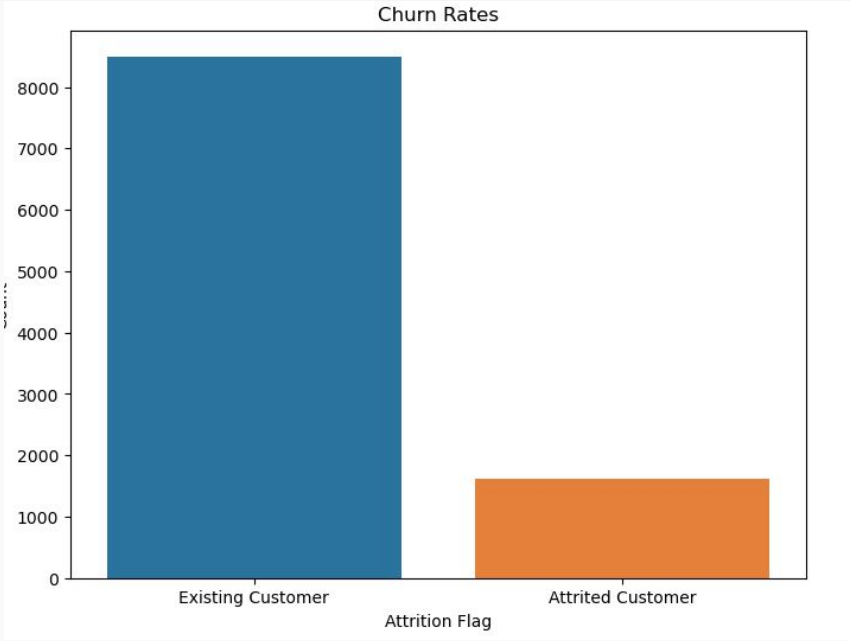
Source:
https://www.kaggle.com/datasets/thedevastator/predicting-credit-card-customer-attrition-with-m/data

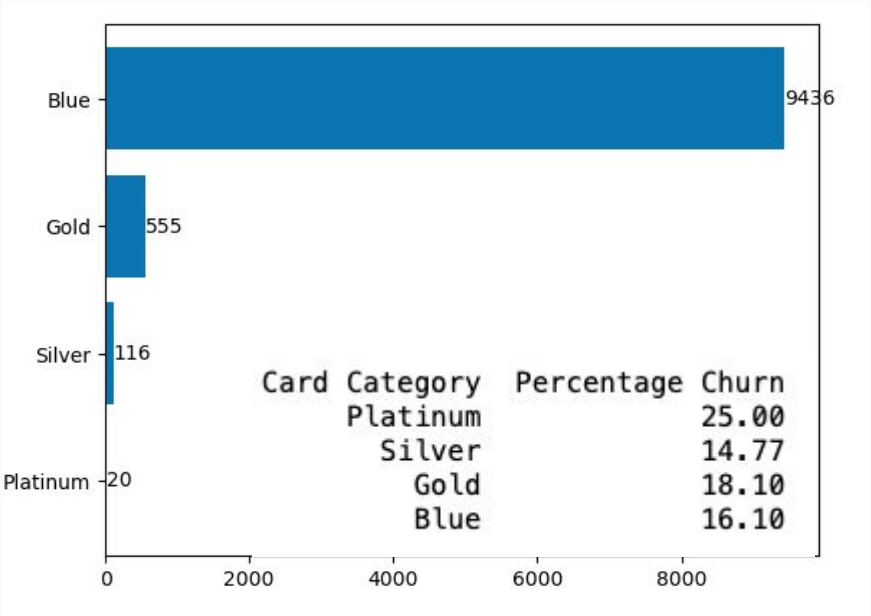
Can we optimize customer retention strategies?

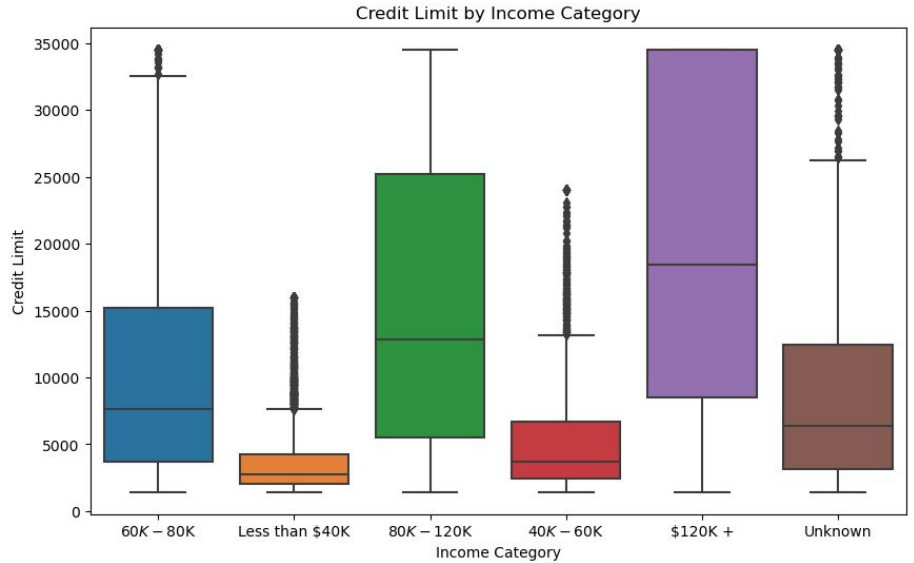
Objectives:

- Can we create clusters organically using unsupervised methods and then predict the labels from the label?
- Predict whether a customer will churn or not?
- Understand customer past behaviour

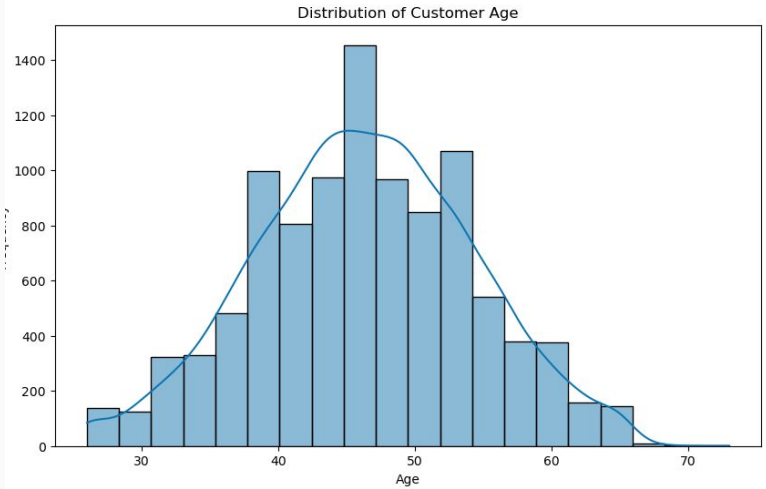
EDA - What do the customers look like?

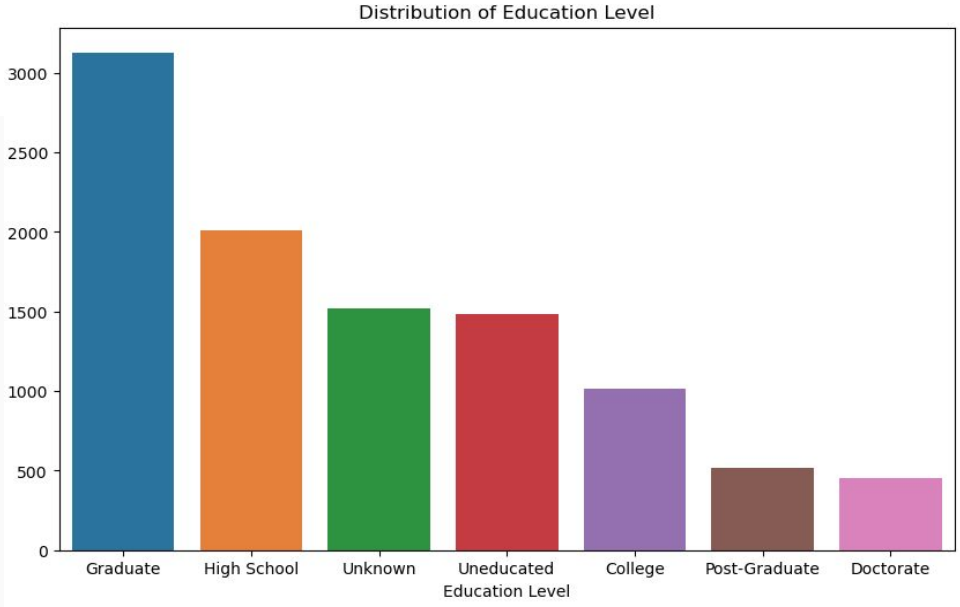






Customer Characteristics





Recency, Frequency, and Monetary Value

- **Months_Inactive_12_mon (a proxy for Recency)**: indicates how many months have passed since the customer's last active engagement.
  a. A lower number of 'Inactive_Months' would imply more recent activity (higher score)
- **Total_Trans_Ct (Frequency)**: a good indicator of transaction frequency over a given period, including all types of transactions such as deposits, withdrawals, payments, etc.
  a. A higher number of 'Total_Trans_Ct' would imply more frequent activity (higher score)
- **Total_Trans_Amt (Monetary)**: how much money is being transacted.

Formula: RFM Score = 0.7*Recency + 0.3*Frequency

RFM

Why not consider *Monetary*?

- **Churn Focus**: the focus might be more on identifying patterns or behaviors that precede account closure or inactivity, rather than on maximizing revenue from each customer. While transaction amounts can indicate customer value, they might not provide insights into why customers are leaving, which is a primary concern in churn analysis.
- **Data Interpretation Challenges**: Large transaction amounts don't necessarily equate to high customer value in a banking context. A few large transactions could be less valuable than many smaller, regular transactions (depends on rewards), which indicate a higher level of customer engagement and loyalty.

Humanization

```
RFM_Segment: Low, Attrited: 926, Existed: 1862, Proportion: 33.21%
RFM_Segment: Medium, Attrited: 648, Existed: 3841, Proportion: 14.44%
RFM_Segment: High, Attrited: 45, Existed: 1428, Proportion: 3.05%
RFM_Segment: Very High, Attrited: 8, Existed: 1369, Proportion: 0.58%
```

Under each segment designed, what's the proportion of attrited vs. existed (based on true labels)?

The lower the segment level,

- The higher the proportion of churn (or the larger the number of attrited customers)
- The higher risk of churn
- Potential Reason: Attrition due to low engagement, infrequent transactions, or recent inactivity

Great portion of the customers falls into the Medium level (largest base of existing customers)

- still room for improvement to move these customers to higher segments and reduce attrition further

High/Very High segment shows extremely strong customer loyalty or satisfaction within groups

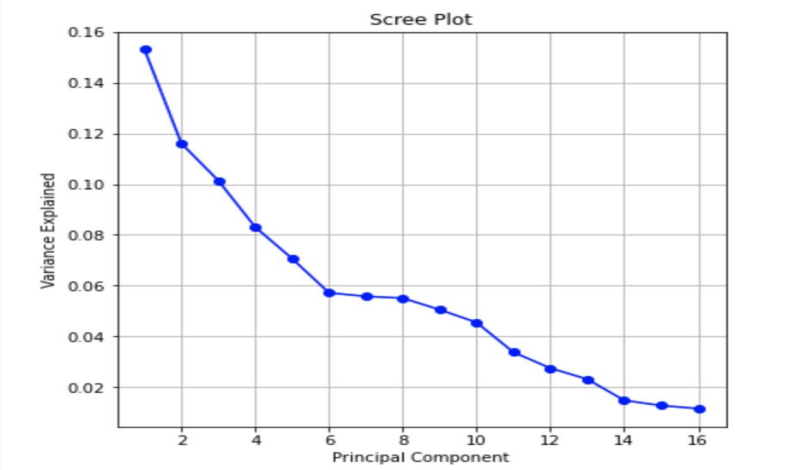
RFM Clustering (K=5) & Strategy

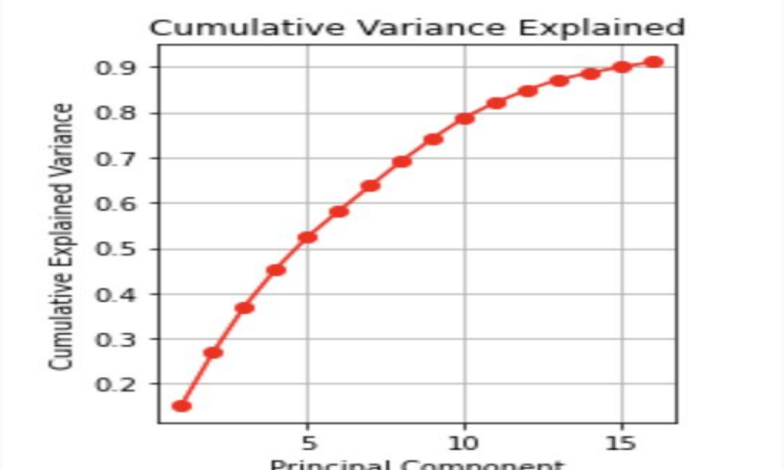
	Cluster	Recency			Frequency			Credit_Limit	Avg_Open_To_Buy	Total_Revolving_Bal	Avg_Utilization_Ratio	RFM_Segment	Attrition_Flag
		mean	min	max	mean	min	max	mean	mean	mean	mean	<lambda>	<lambda>
0	0	3.232194	3	6	45.901194	10	67	8340.441046	7253.245080	1087.195965	0.256716	Low	915
1	1	1.569354	0	2	95.104575	82	138	10137.774001	8881.357153	1256.416848	0.265763	Very High	8
2	2	1.601707	0	2	46.125835	10	67	8770.724016	7601.014662	1169.709354	0.263795	Medium	567
3	3	3.278087	3	6	83.692201	68	139	8539.255896	7362.487558	1176.768338	0.289722	Medium	92
4	4	1.575696	0	2	74.565513	68	81	7586.735438	6419.744942	1166.990496	0.312018	High	45

- **Risk of Churn**: Clusters 0, 2 show a high number of attrited customers, indicating they are at a higher risk of churn.
- -> Target Clusters 0 and 2 with loyalty programs and personalized outreach to reduce churn.
- **Engagement and Value (Credit_Limit)**: Clusters 1, 4 show very high engagement, with Cluster 1 having the most valuable customers based on RFM Segment classification and credit limit.
- -> Leverage the high engagement of Cluster 1 to introduce premium services and rewards.
- **Credit Utilization (Avg_Open_To_Buy)**: Clusters 3,4 have lower credit utilization ratios, which could be an indicator of different financial behaviors that might be of interest for cross-selling or upselling financial products.
- -> Offer credit management or savings products to Cluster 3, which has high revolving balances. And customize

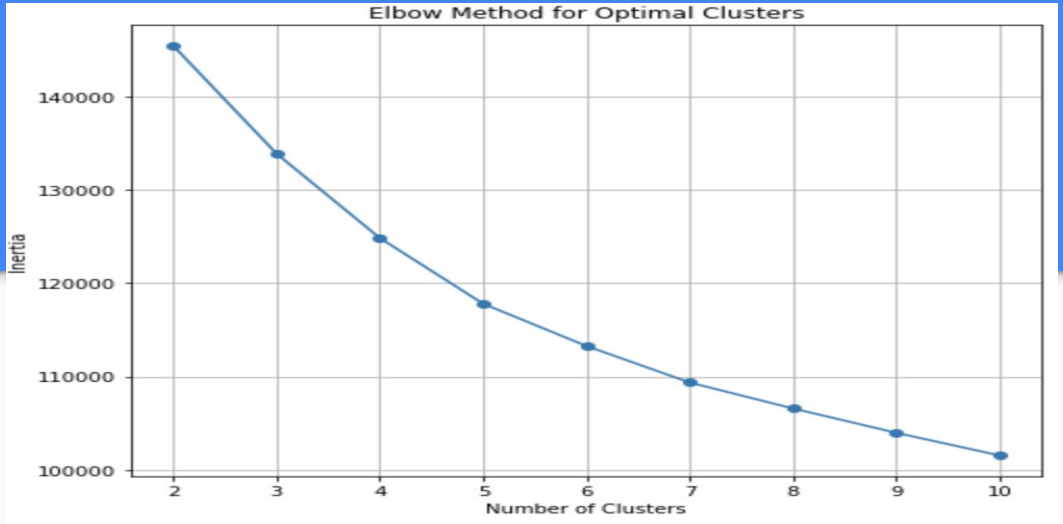
Principal Component Analysis

- The 16th component brings the cumulative variance explained to 90%, affirming the adequacy of our PCA model in preserving data integrity.
- This graphical representation confirms that the chosen components retain the majority of data variability, supporting subsequent analytics or model building.



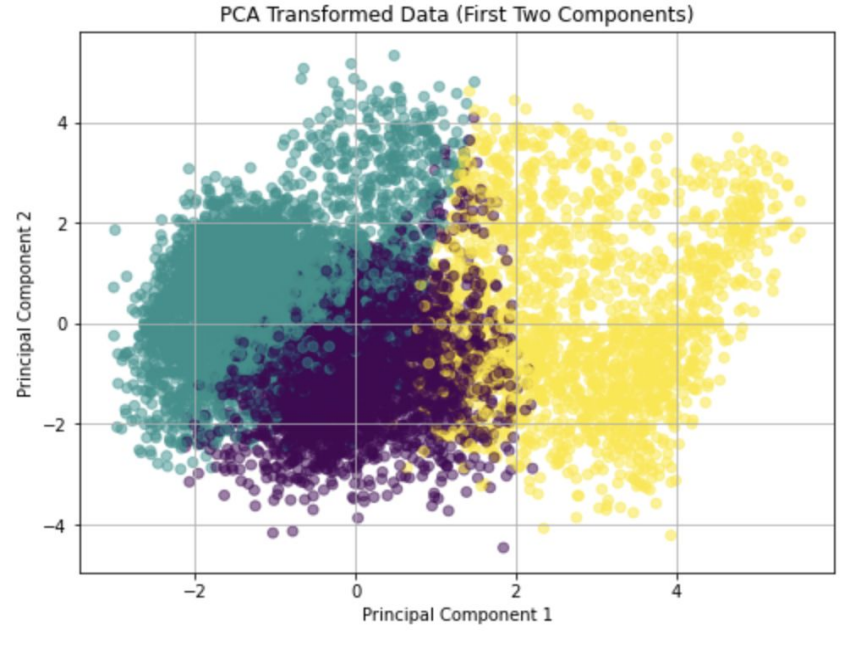


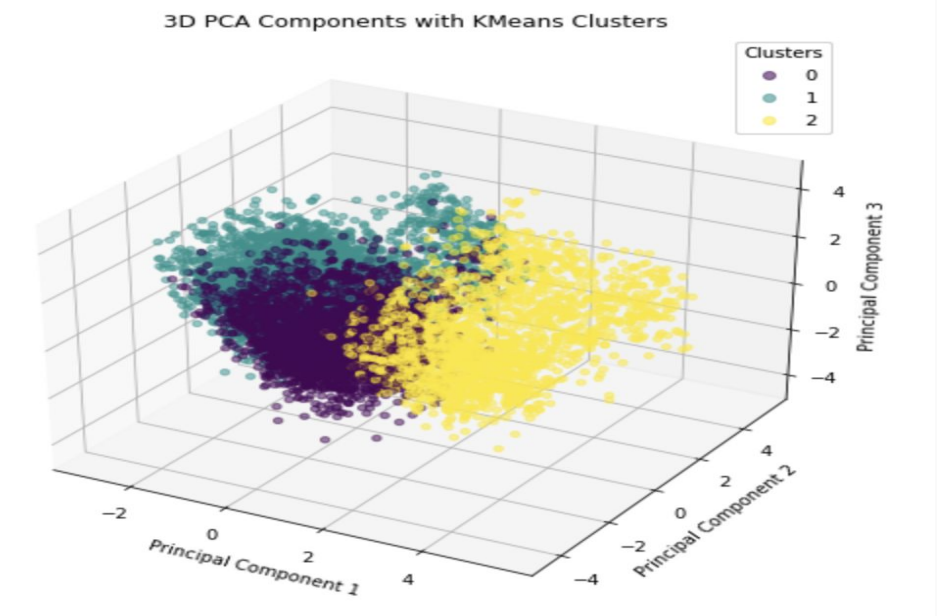
K-means



- Elbow Method for Optimal Cluster Determination: To determine the optimal number of clusters, we analyzed inertia for 2 to 10 clusters to identify the most appropriate cluster count, highlighting the balance between cluster coherence and total number.

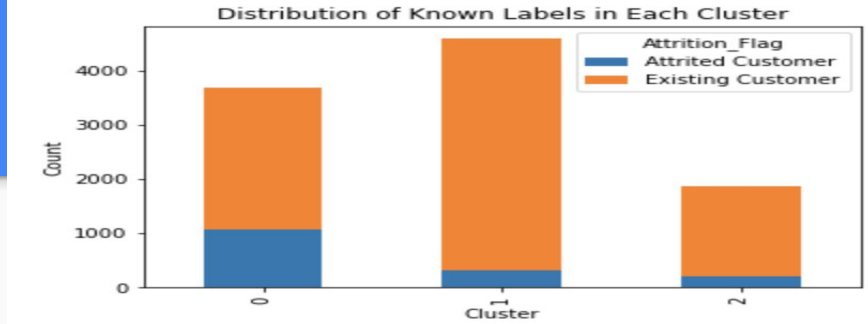
K-means


PCA Transformed Data (First Two Components)
Principal Component 2
Principal Component 1
4
2
0
−2
−4
−2
0
2
4


3D PCA Components with KMeans Clusters
Clusters
0
1
2
Principal Component 1
Principal Component 2
Principal Component 3

K-means

Attrition_Flag	Attrited Customer	Existing Customer
cluster		
0	1080	2600
1	330	4250
2	217	1650



Cluster Characteristics:

- **Cluster 0:** Mix of customers (2600 existing vs. 1080 attrited), which could indicate a segment with high loyalty or satisfaction. We analyze common features within this group to understand how to promote retention.
- **Cluster 1:** Contains a large number of existing customers (4250) and attrited customers (330). This cluster may represent engaged customers who are sensitive to certain triggers leading to attrition.
- **Cluster 2:** A balanced mix of attrited (217) and existing customers (1650). This group might contain newer customers or those with a neutral stance towards the services provided.

Clustering & Supervised Learning

Objective: Predicting attrition rates for various customer segments.

Importance: Insightful for devising targeted retention strategies and understanding customer churn patterns to enhance retention strategies and service offerings.

Clustering & Supervised Learning

How It Works:

- Feature Engineering: Implemented SMOTE to counteract class imbalance issue, integrated K-Means clustering results (cluster labels) with selected features including:
  - Demographic info such as customer age, gender, education levels, income category, etc.
  - Customer spending behavior such as credit limit, num of inactive months, total transaction count.
- Target Definition: Transformation of 'Attrition_Flag' into a binary outcome (0 for retained, 1 for churned customers).

Clustering & Supervised Learning

Model Performance

Model Choice	AUC Score	Precision	Recall	F1
Logistic Regression	0.77	0.46	0.70	0.55
Random Forest	0.93	0.86	0.88	0.87
Gradient Boosting	0.91	0.63	0.92	0.75

Why Random Forest?

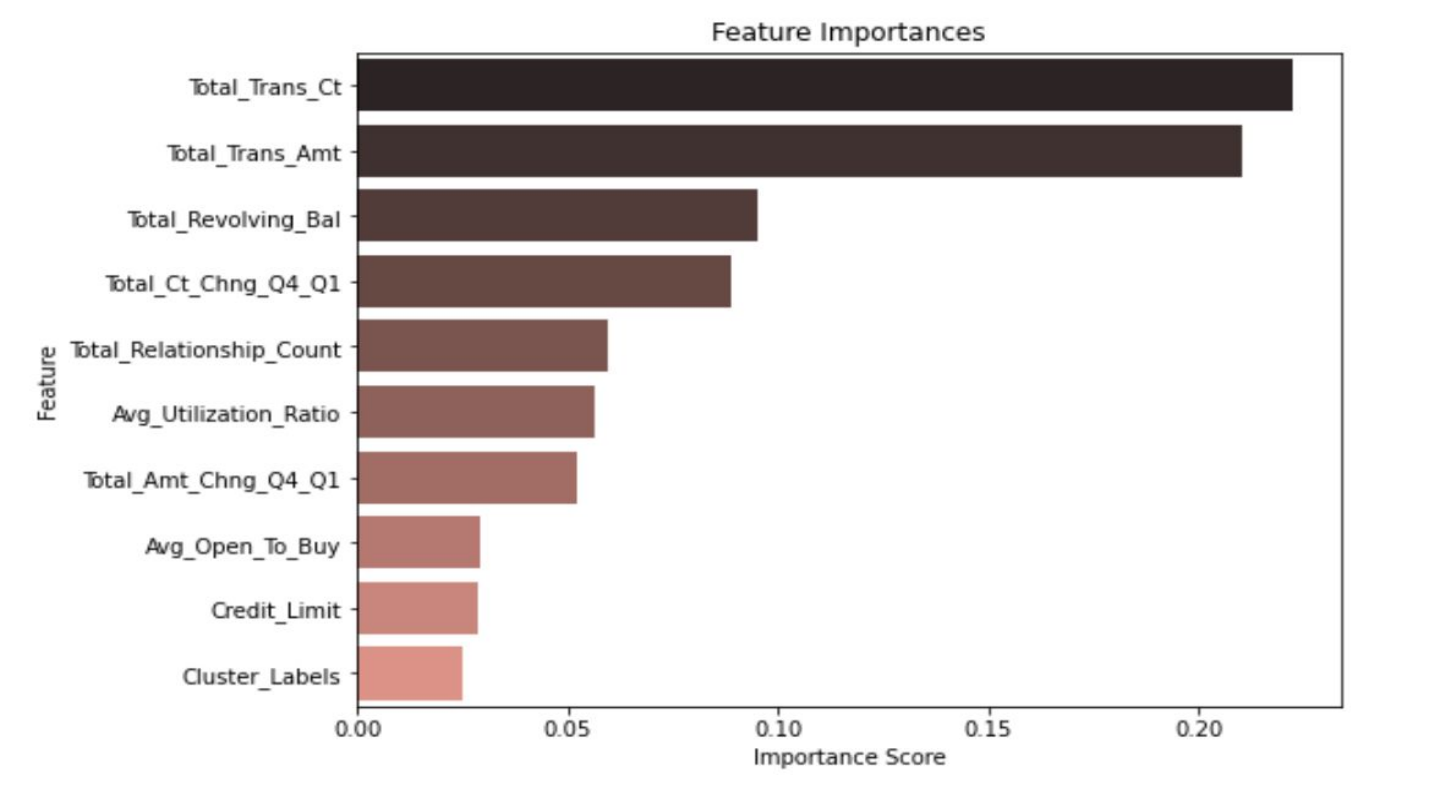
- The Random Forest model with 0.93 **AUC** is best at differentiating between customers who will and will not churn.
- For predicting attrition, a higher **precision** (Random Forest with 0.86) means that when the model predicts a customer will churn, it is correct most of the time.
- A high **recall** (Gradient Boosting with 0.92) is crucial if the cost of missing a customer who might churn is high.
- The Random Forest model shows the highest **F1 score** (0.87), suggesting a good balance between precision and recall.

Clustering & Supervised Learning

Model Evaluation & Comparison

Factor	Random Forest	Logistic Regression	Gradient Boosting	Why Random Forest?
Overfitting Resistance	High	Low	Moderate	Ensemble method provides robustness to overfitting.
Imbalance Data	Handles well	Handles moderately	Handles well	Effective in datasets with unequal class distribution.
Feature Importance	Provides insight	Limited insight	Provides insight	Useful for understanding and improving retention.

Clustering & Supervised Learning



Conclusion

Recommendations to reduce customer churn:

- Increase customer transaction frequency and amount through loyalty programs or personalized marketing.
- Improve credit options and benefits encourage customers to maintain a revolving balance.
- Provide tools for better credit management to help customers avoid over utilization of credit.
- After identifying the clusters with a higher churn risk, the company can optimize resource allocation by focusing on the most vulnerable customer segments.

Findings:

- Activity in the past 12 months is more important than frequency and monetary
- Total transaction amount and count is most important to predict churn
- Cluster 1 is at risk customers, Cluster 0 are loyal, Cluster 2 are neutral

Thank you and Q&A