Convolve 3.0:
A Pan IIT AI/ML Hackathon

B.Tech 2nd Year

Department Of Artificial Intelligence

Sardar Vallabhbhai National Institute of Technology

Team Name: deepdblm

Members:

Deep Das
Nilang Bhuva
Archit Savaliya

Problem Statement: Development of a behaviour scoring model for existing credit card customers

Objective: Predict probability of default for risk management

Dataset size: 96,806 records (development) and 41,792 records (validation)

- **Results:** We have made a neural network of 4 hidden layers, 6 convolutional layer for trying to predict the probability of given credit card holding customer defaulting.

```
Neural Network Model accuracy: 0.96
Neural Network F1-score: 0.96
Neural Network Classification Report:
              precision    recall  f1-score   support

           0       0.95      0.98      0.97     19160
           1       0.96      0.90      0.93      9455

    accuracy                           0.96     28615
   macro avg       0.96      0.94      0.95     28615
weighted avg       0.96      0.96      0.96     28615

Neural Network Model, Scaler, and Imputer saved.
1/1 [==============================] - 0s 103ms/step
Neural Network Prediction: [[1]]
```

- Model performance metrics were the ratio of defaulters to ratio of non-defaulters as we didn't have any reference value for the probability which we took out to calculate other evaluation metrices
- **Model Evaluation Metrices for predicting the bad-flags:**
  Model Accuracy: 0.96
  F1 score: 0.96
  Recall: 0.9

Table of Contents

Data Insights

Development dataset: 96,806 records with bad_flag and Account Number

Other Variable Categories

1. **On-us Attributes**
    - Credit limit related variables
    - There are 48 Attributes
2. **Transaction Attributes**
    - Merchant-specific transaction data
    - Transaction counts and values
    - There are 664 Attributes
3. **Bureau Tradeline Attributes**
    - Product holdings
    - Historical delinquencies
    - There are 452 Attributes
4. **Bureau Enquiry Attributes**
    - Recent loan enquiries
    - There are 50 Attributes

1)Anomaly Count

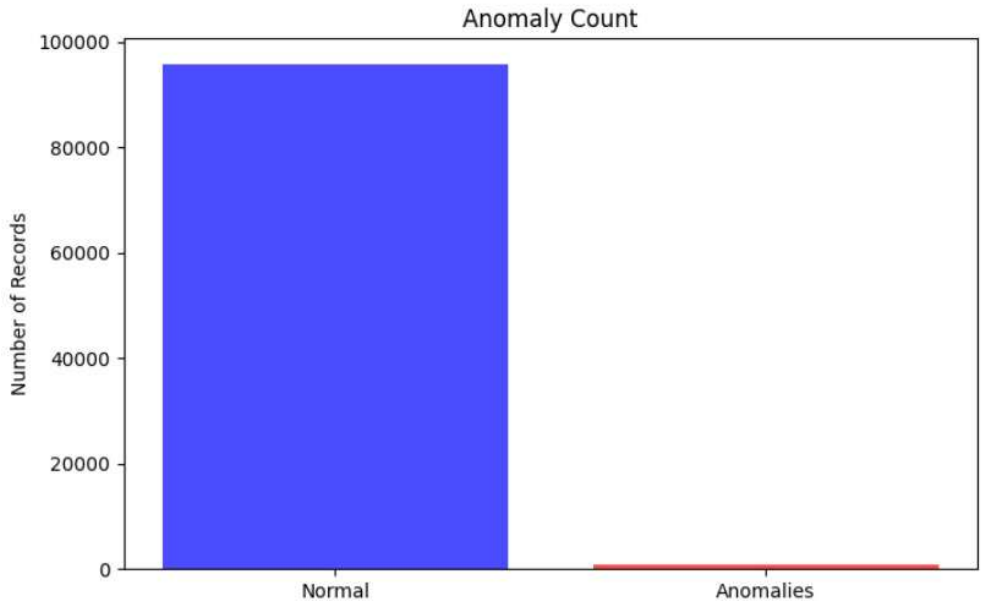


Fig. 1: Anomaly Count

The chart depicts the distribution of Transaction Homogeneity Scores, showing how frequently certain scores occur within the dataset

Observation

The bulk of the points are centered at low Homogeneity Scores, peaking sharply near the bottom of the x-axis.
The frequency drops very rapidly as the Homogeneity Score rises, with almost no higher scores occurring.

Insights

High concentration of low scores means that most transactions follow a similar or uniform pattern, hence high consistency in transaction behavior among most accounts.
The scattered distribution of more significant values can relate to anomalies or outlier accounts of highly diverse or distinctive transaction flows.

Ratio of Good behaviour score to Bad behaviour score: 69.56

So Out of Every 70 Account there is one Account with Bad behaviour score

2) Correlation Matrix

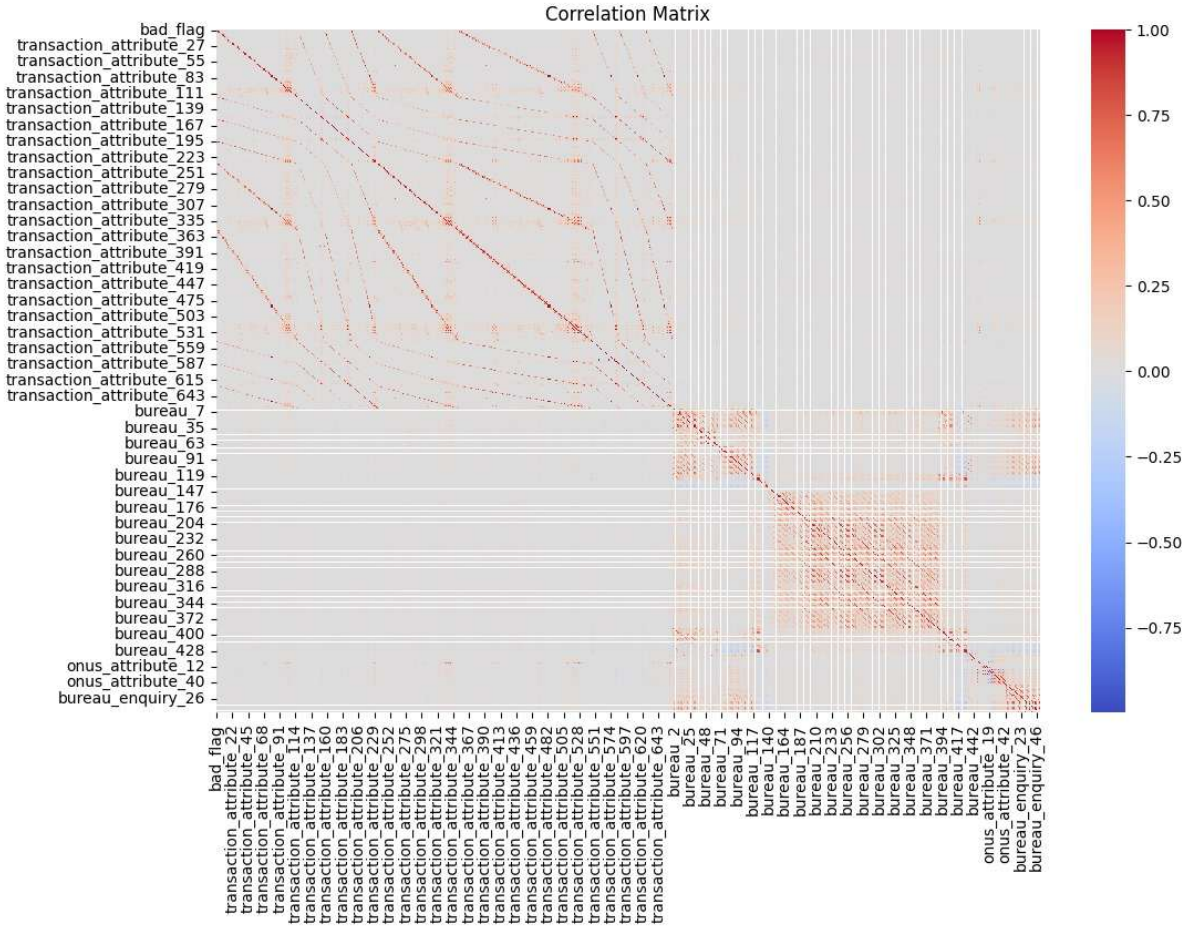


Fig. 2: Correlation Matrix

The Correlation Matrix, a heatmap where each cell represents the correlation coefficient between two variables in the dataset. The colour intensity indicates the strength and direction of the correlation:

- Red (closer to 1): Positive correlation (both variables increase together).
- Blue (closer to -1): Negative correlation (one variable increases as the other decreases).
- Gray/Neutral (around 0): No significant correlation.

Key Insights and Derivations

Highly Correlated Groups (Red Clusters):

- Red blocks show highly related variables those that appear to measure the same thing.
- Actionable Insight: we can use it to apply dimensionality reduction or feature selection to remove redundancy.

Negative Correlation (Blue Regions):

- Sparse blue regions indicate variables that inversely impact one another often representing trade-offs.
- Actionable Insight: we can use negatively correlated features to identify risk indicators (e.g., bureau inquiries vs. spending habits).

Sparse or Neutral Relationships (Gray Areas):

- Variables with no significant correlation often contribute independently to the target variable.
- Actionable Insight: indicates non-linear relation or the factors which have maybe the least impact on the factor under consideration.

Correlation with bad_flag

- Attributes that are positively correlated with bad_flag are potential risk indicators. Negative correlations may indicate stability indicators.
- Actionable Insight: Include correlated features in predictive models; weak correlations might need further feature engineering.

3) Important Correlations

This bar chart shows the top correlations between all attributes and bad_flag

- The correlations are relatively weak overall, with all values around 0.10 (or 10%), suggesting no single attribute is a strong predictor of default on its own.
- "onus_attribute_2" has the highest correlation at approximately 0.11, followed closely by "onus_attribute_17" at about 0.105.
- There's a consistent pattern where all five attributes show similar levels of correlation (between 0.095 and 0.11), with only small differences between them.

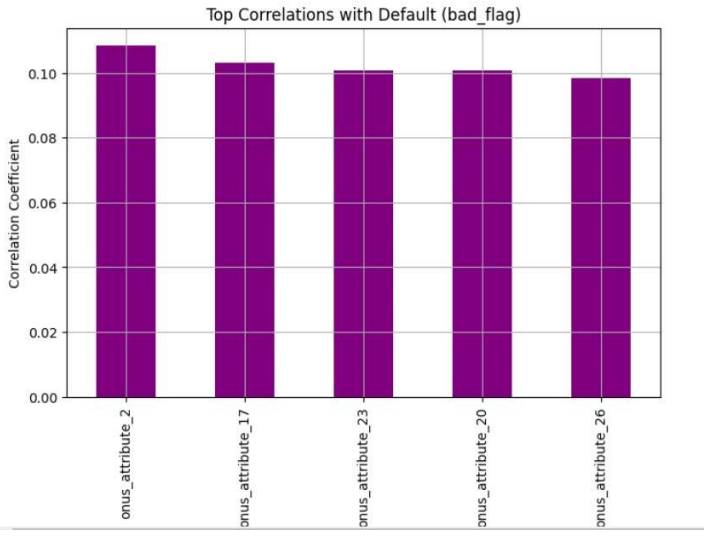


Fig.3 Top Correlation

- Despite being the "top" correlations, these relatively low values suggest that:

- A multivariate approach might be more effective than using single variables
- Other types of relationships (non-linear) might be worth exploring
- Additional features or feature engineering might be needed for better default prediction

4) Inquiry Load Distribution

```
Insight 6: Inquiry Load
<ipython-input-3-a2c16588f8e3>:67: PerformanceWarning: DataFrame is highly fragmented.  This is usually the result of calling
  df['inquiry_load'] = df[bureau_enquiry_cols].sum(axis=1)
```

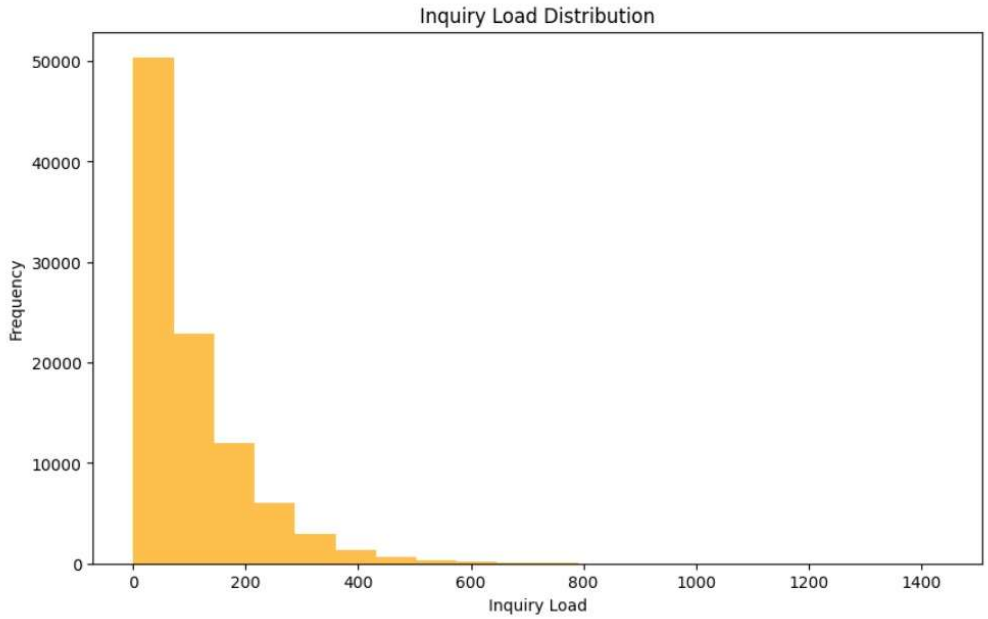


Fig.4 Inquiry Load Distribution

The histogram visualizes the distribution of inquiry load across individuals.

Key Observations

Skewed Distribution: The histogram is highly skewed to the right, showing that most people have a low inquiry load, while a few have very high values.

Long Tail: There is a small segment of people with very high inquiry loads (more than 400 inquiries), which would be considered outliers or unusual behaviour.

Derivations

Low Inquiry Load (Dominant Segment):

The majority of the population has a low inquiry load, indicating a stable financial profile or less dependence on credit inquiries.

High Inquiry Load (Sparse Segment):

Those with a very high inquiry load could be people who have a high credit-seeking behaviour and could be an indication of financial distress or risky profiles.

Frequency Decline:

The way frequency drops sharply with increasing inquiry load suggests concentration in stable behaviour in credit usage in the population.

Insights and Actionable steps

Risk Identification:

- A high inquiry load indicates increased reliance on credit or financial instability.
- These individuals are the ones that are needed to be looked at more closely for potential risks, including loan defaults or overleveraging.

Stable Profile Identification:

- Low inquiry load identifies a segment that most likely has better creditworthiness and hence is fit for low-risk financial products or services.

Outliers:

- Extreme values in the long tail (e.g., >400 inquiries) needs further investigation to verify data accuracy or identify fraud or unusual patterns, based on the exact names and characteristic it represent in the banking sector.

Actionable Steps:

1.Risk Monitoring:

Credit card company should focus on individuals with higher inquiry loads for risk mitigation strategies or tailored financial products.

2.Segment Analysis:

We can create customer segments based on inquiry load to design differentiated strategies for high-risk and low-risk groups.

5) Financial Stress

```
Insight 2: Financial Stress Prediction
<ipython-input-9-237823ea298e>:66: PerformanceWarning: DataFrame is highly fragmented.  This is usually the result
  df['financial_stress_score'] = (
```

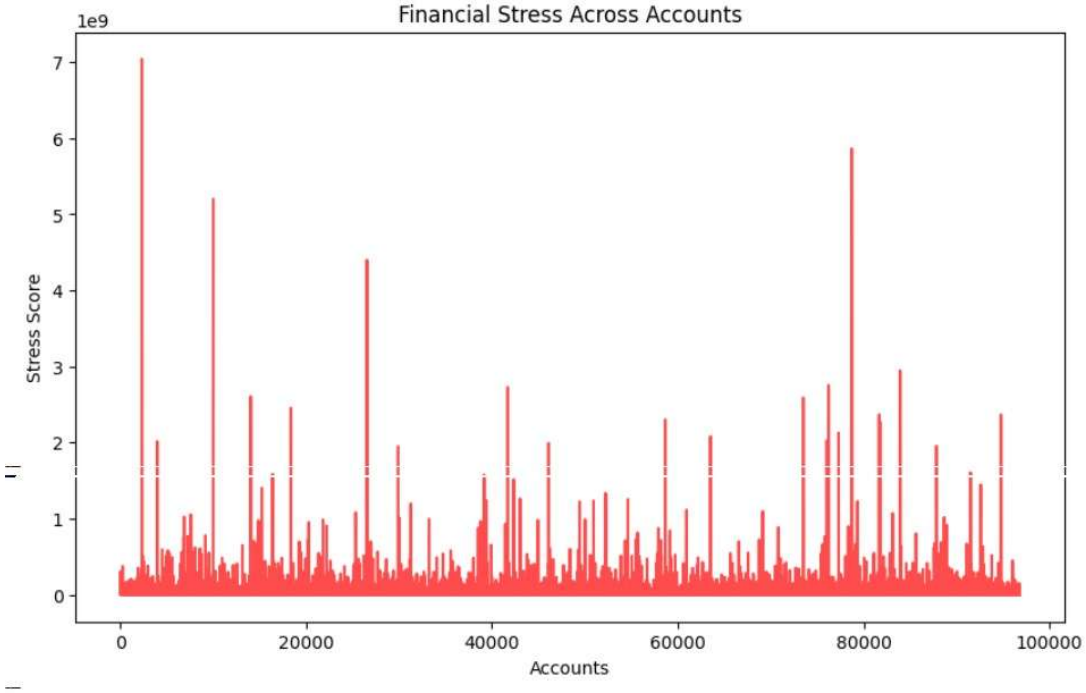


Fig. 5: Financial Stress Across Accounts

The chart shows the financial stress scores calculated for individual accounts, displaying their distribution in a population. Financial stress scores reflect the level of financial risk or the degree of difficulty associated with each account.

Key Observations

- Stress Scores Variability: Financial stress scores have very high variability, with scores ranging from almost zero to extremely high values (>7e9).
- Frequent Spikes: There are many spikes in the stress scores, which indicate that some accounts are highly stressed relative to the others.
- Predominantly Low Stress: Most accounts have relatively low financial stress scores, mostly spread near the bottom of the axis.

Derivations

Outliers:

- Accounts with extremely high stress scores indicates serious financial distress, fraud, or data errors.
- These peaks are sparse but they have a substantial effect on risk modeling in general.

Majority Stability:

- Many accounts have low stress scores, indicating that most people are financially stable or that the majority are not exposed to much risk.

Insights

Risk Segmentation:

- Spikes in stress scores indicates high-risk accounts that deserve closer examination or focused intervention.
- Credit card company can provide counselling on finances, loan rescheduling, or credit monitoring to these type of account holders.

Population Segregation:

- Population distribution calls for segregation in low-stress and high-stress groups to focus on them

Actionable Items

- Outlier Analysis: Detailed investigation into individual accounts with stress scores above threshold (to be set by the company).
- Feature Normalization: Transformation or normalization of the log stress score can be used in predictive models to dampen extreme values.
- Focused Interventions: Devise strategies targeted at high-stress accounts in order to prevent financial risk or to provide individualized solutions.
- Stress Threshold: Define thresholds that can be implemented in terms of "low," "medium," and "high" levels for easy understanding during operations.

6) Spend Distribution Skewness

The graph visualizes the skewness in spend distributions across transactions, providing insights into the asymmetry of the data. Skewness measures the extent to which the distribution deviates from symmetry, with values close to 0 indicating balanced (symmetric)

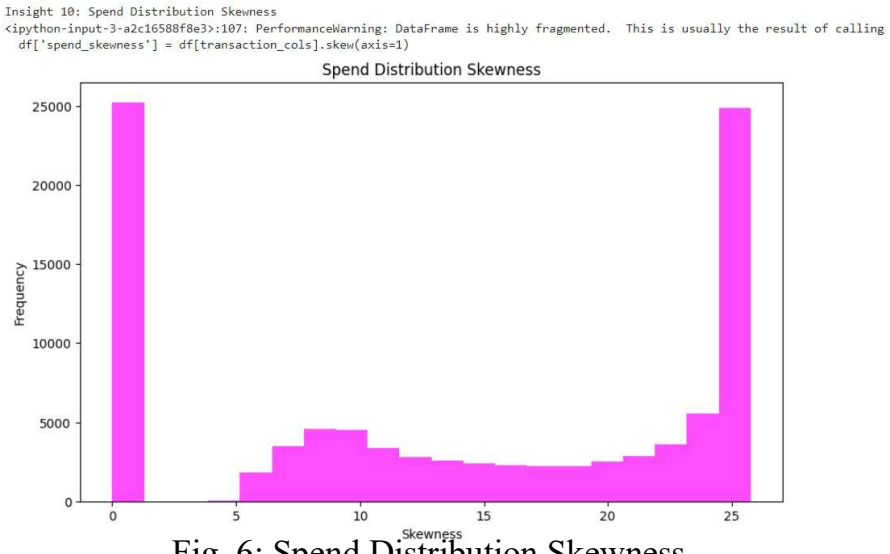


Fig. 6: Spend Distribution Skewness

distributions, while higher positive skewness suggests distributions with long tails on the right. This histogram illustrates the frequency of skewness values, allowing us to identify patterns in transaction behaviours.

Observations

High Frequency at Zero Skewness:

- A large proportion of the data is concentrated around skewness values close to 0, meaning that most transaction distributions are nearly symmetric and do not have significant outliers.

Positive Skewness Cluster:

- There is a noticeable range of skewness values from 5 to 25, indicating distributions with long tails on the right. This implies the existence of high-value transactions that are less frequent but highly influential in the overall distribution.

Insights

- Symmetric Spending Patterns: Most of the data represents regular transaction behaviour with evenly distributed values, which is perfect for standard statistical modelling.
- Outliers and Irregular Behaviour: The spread of higher skewness values indicates that the transactions may contain extreme, infrequent events, which may include unique customer behaviour, seasonal peaks, or very high-value transactions.

Derivations

- Segmentation Opportunities: Based on skewness values, it is possible to group transactions for segments that consistently spend in the same pattern and those with more irregular spending.
- Risk Implications: High skewness can suggest financial risks or opportunities with respect to outliers that need further review of the origin and effect.

Recommendations

- Targeted Analysis: Credit card company can identify the transactions that are causing skewness to identify patterns such as high-value customer groups or seasonal trends.
- Customer Segmentation: Credit card company can use clustering algorithms to divide the data into segments for tailored strategies based on spending patterns.

7) Average Spend by Cluster

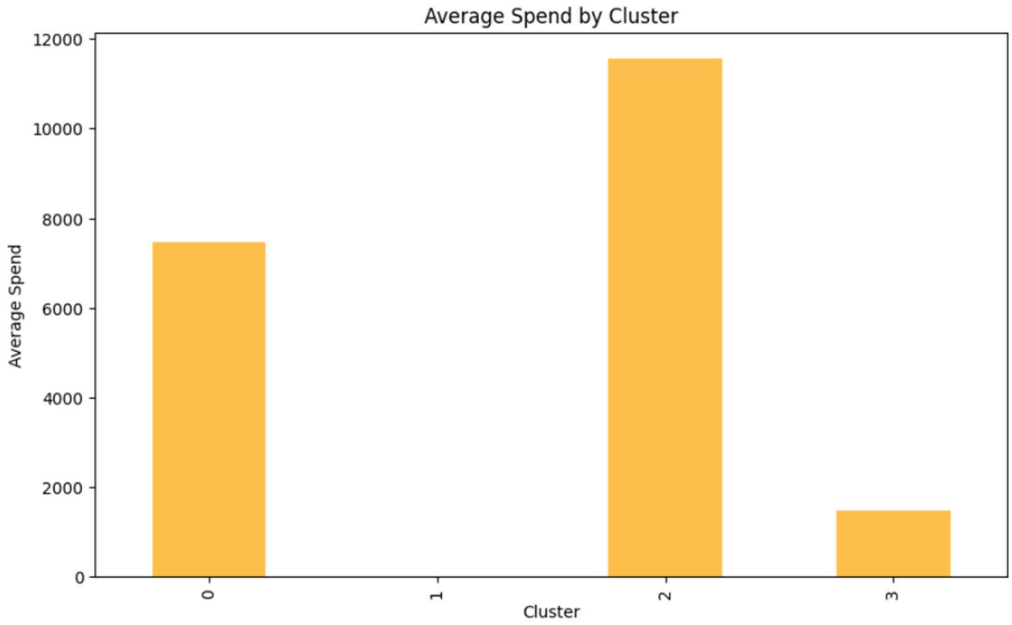


Fig. 7: Average Spend by Cluster

The bar chart illustrates the average spend across different clusters, providing insights into customer behavior and spending patterns. Each cluster represents a distinct grouping of customers based on shared characteristics or behaviours, such as spending habits, borrowing tendencies, or credit history.

Key Findings

- Cluster 2 has the highest average spend: The average spend for customers in Cluster 2 is significantly higher than any other cluster, meaning that this group comprises high-value customers. The customers probably belong to a segment with strong financial capacity or are highly engaged in spending.
- Cluster 0 shows moderate spending behaviour: Customers in Cluster 0 have an average spend that is relatively high but significantly lower than Cluster 2. This cluster may represent middle-tier customers who are active but not as profitable as Cluster 2.
- Cluster 3 has the lowest average spend: The average spend in Cluster 3 is low compared to other clusters. This cluster may be a low-value customer or an individual with a very low financial involvement or restrictions.
- Cluster 1 is either not present or weakly represented

Insights

- Marketing to High-Paying Customers, Cluster 2: Marketing and premium service efforts should focus on Cluster 2, as it is likely that this is where the business gets most of its revenue. Customized offers or loyalty programs can be used as a retention effort for this cluster.

- Opportunity to Engage Mid-Tier Customers (Cluster 0): Cluster 0 can be an area of growth. Upsell or cross-selling products and services may encourage such customers to increase their spending towards Cluster 2 behaviour.
- Improve Interaction with Low-Value Customers (Cluster 3): For Cluster 3, it would be possible to find out what barriers exist in terms of spending. Providing simple or introductory products combined with educational tools could encourage higher interaction.

Example of insights by selecting random 10 accounts from the validation dataset.

```
Sampled Account Insights:
       spending_elasticity  financial_stress_score  overleveraged  \
60498            21.858597            2.400344e+07           True
56514             0.000000            4.165104e+06           True
88379             3.223360            2.418326e+07           True
47016             6.397961            4.993423e+06           True
83012             4.714727            4.154530e+07           True
81438             8.464373            6.431539e+06           True
49612             0.000000            2.314204e+07           True
75524             4.228647            2.007428e+06           True
80762             4.787577            1.531433e+06           True
61008             4.613278            2.461708e+06           True

       inquiry_load  delinquency_score  transaction_concentration  \
60498          46.0       521813.389362                   0.091493
56514          12.0       347090.992285                        NaN
88379          32.0       755726.114056                   0.917893
47016          83.0        60160.796740                   0.156235
83012          90.0       461613.504440                   0.565638
81438          94.0        68419.713233                   0.338195
49612          97.0       238576.771396                        NaN
75524          28.0        71693.010428                   0.671251
80762          99.0        15468.066331                   0.400981
61008         256.0         9615.064805                   0.625158

       credit_limit_utilization  spend_skewness
60498                 21.858597        6.368508
56514                  0.000000        0.000000
88379                  3.223360       25.751398
47016                  6.397961        8.528458
83012                  4.714727       23.875813
81438                  8.464373       18.493973
49612                  0.000000        0.000000
75524                  4.228647       25.090581
80762                  4.787577       21.895932
```

8) Spending Behaviour Volatility

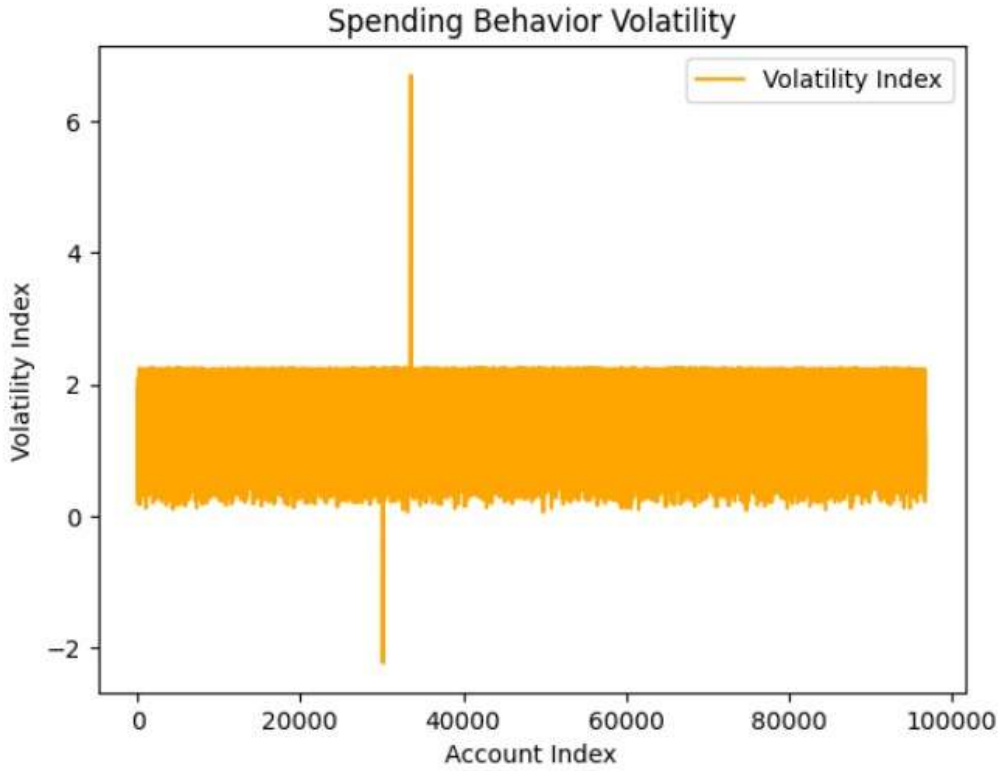


Fig.8: Spending Behaviour Volatility

The graph illustrates the "Spending Behaviour Volatility" of Credit Card customers across the development dataset. Each account is represented along the X-axis (Account Index), while the Y-axis indicates the corresponding **Volatility Index**. The Volatility Index measures fluctuations in spending behaviour, where higher values imply greater variability and potential risk.

Observations and Derivations

Total Distribution:

- Mostly, the clients' Volatility Index is centralized between 1.5 to 2.0. This would indicate a significant degree of expenditure stability for the majority of the accounts.
- Some few accounts would be observed as extreme on volatility at both high and low values as against the mainstream.

Outlier Anomalies:

- There is a prominent spike in volatility for a particular account or group of accounts, indicated by the tall peak at an index around 40,000. This might reflect unusual spending behaviour, potentially tied to fraudulent activity, erratic usage patterns, or atypical financial behaviour.

- Zero or Negative Volatility: Accounts that have negative or close to zero volatility are present, but such cases are very few. They may be representing dormant accounts or accounts with smooth spending patterns, that is, no high spikes.

- Most of the Accounts Uniform Band: The flat uniform band of data reflects that most customers' spending pattern is predictable with no high spikes of volatility. Such customers can be less risky for default.

Insights

- Focus on outliers: Accounts that may have high volatility or extreme hike should be flageed for further underlines.
- Low-Risk Category: The stable Volatility Index in most accounts can be classified as low-risk customers. Targeting them with respect to high value-added products or expanding credit lines could be an optimal choice.
- Behavioural Clusters: The data suggests that customers can be clustered based on their Volatility Index. The risk management framework can be improved by segmenting customers into low-risk, moderate-risk, and high-risk categories.

- Actionable Outliers: The large spike at ~40,000 and similar anomalies should be cross-referenced with other variables (e.g., delinquency history, credit utilization) to determine whether these are systematic issues or genuine high-risk cases.

9) Credit to Inquiry Ratio

The graph represents the distribution of the **Credit-to-Inquiry Ratio** for customers in the dataset. The X-axis denotes the ratio value, while the Y-axis represents the density or frequency of these ratios. This metric evaluates the relationship between the credit limit assigned to a customer and the number of credit inquiries made by that customer. Higher values suggest higher credit assigned relative to inquiries, while outliers indicate potential anomalies or unusual behaviours.

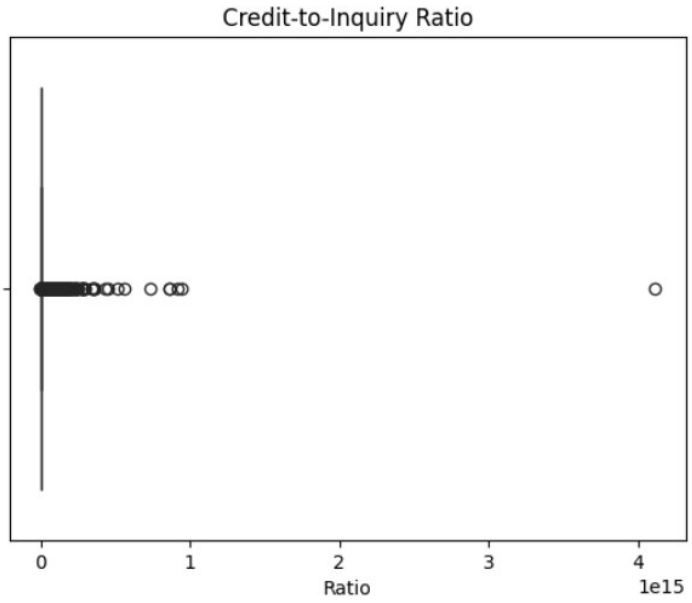


Fig.9: Credit to inquiry ratio

Observations and Inferences

Centralization Towards Low Ratios:
Mostly data points were distributed around low Credit-to-Inquiry Ratio - around zero indicates the majority have the credit to inquire ratio not being very lopsided - at least relative to their high frequency of inquiring.

Prevalence of Anomalies: There is a significant outlier at an extremely high Credit-to-Inquiry Ratio ($\sim 4 \times 10^{15}$). This is anomalous and could be due to an erroneous account or data. This ratio is far beyond the observed range of typical customer behavior.

Spread of Data Points: Apart from the concentration near zero, there are a few scattered points with moderate ratios, which indicates a small group of customers with high credit limits relative to their inquiry activity. These accounts may be of high-value customers or exceptional cases.

Skewness in Distribution: The distribution seems to be heavily skewed to the right, mainly because of the presence of extreme outliers. This calls for normalization or further investigation to understand the nature of these anomalies.

Insights for Risk Management:

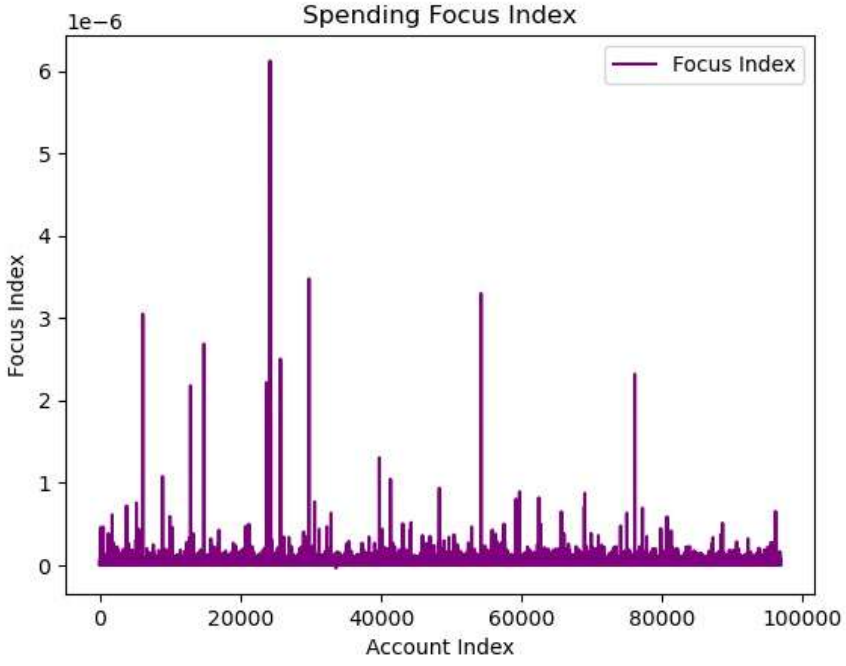
Anomaly Investigation: Extreme outlier must be investigated for data entry errors or an unusual customer scenario (for example, high-value customers with minimal inquiries). Such cases may distort the model's predictions and should be treated separately.

Customer Segmentation: Customers with moderate Credit-to-Inquiry Ratios could represent a low-risk segment, as their credit utilization aligns proportionally with their inquiries. These accounts might be suitable candidates for credit expansion or premium services.

High-Risk Indicators: Extremely low ratios might indicate excessive inquiries relative to credit limits, suggesting potential credit-hungry behavior or financial instability. These customers could be flagged for closer monitoring.

10) Spending Focus Index

The graph visualizes the Spending Focus Index for a range of accounts. The X-axis represents the Account Index, while the Y-axis denotes the Focus Index values. The Spending Focus Index is likely a metric that quantifies how concentrated or focused an account's spending behavior is, with higher values indicating more pronounced or erratic spikes in spending across specific categories or timeframes.



What the Data Tells Us:

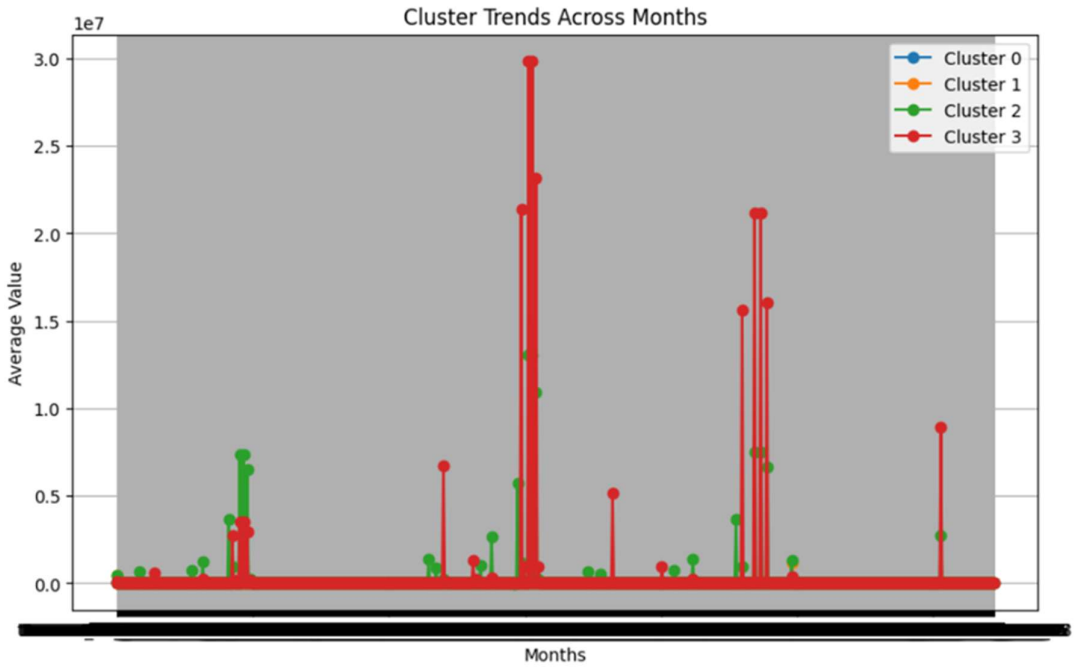
1. **Most Accounts Show Stable Behavior:**
    a. For the majority of accounts, the Spending Focus Index remains at very low levels. This suggests that their spending patterns are generally consistent and not skewed towards specific areas or timeframes.
2. **Unusual Spikes in Specific Accounts:**
    a. There are some accounts that stand out with sharp spikes in their Focus Index, particularly around the **20,000th Account Index**. These spikes highlight accounts where spending behavior is noticeably concentrated or irregular.
3. **A Few Outliers Among a Stable Baseline:**
    a. While the data mostly appears stable, the sharp jumps in a handful of accounts clearly indicate outliers. These outliers could represent unique spending behaviors, rare events, or even potential anomalies in the data.
4. **Small but Significant Values:**
    a. The Y-axis is scaled to very small numbers (*10−610^{-6}*10−6), which means even the biggest spikes in the Focus Index reflect minor deviations. However, these small variations could still have a meaningful impact, especially for identifying rare behaviors.

What This Means:

1. **Spotting Anomalous Accounts:**
    a. The accounts that show significant spikes are worth a closer look. These might represent high-value customers with unique spending habits or could even signal unusual activity like fraud or targeted spending campaigns.
2. **Clustering Accounts by Behavior:**
    a. From this data, we can broadly group accounts into two categories:
        i. **Stable accounts**: Consistent spending patterns with low Focus Index values.
        ii. **Anomalous accounts**: Those with sharp spikes, indicating concentrated spending behavior.
3. **Investigating the Outliers:**
    a. The spike around the **20,000th Account Index** is particularly striking and should be investigated further. It could point to a rare spending event or, alternatively, an issue with the data that needs validation.
4. **Using This for Future Predictions:**
    a. The Spending Focus Index could serve as a valuable input for predictive models in areas like fraud detection, customer segmentation, or even credit risk analysis. Properly analyzing and normalizing these outliers would make such models more accurate.
5. **Customer Behavior Insights:**
    a. Accounts with higher Focus Index values might represent:
        i. High-value customers with niche preferences.
        ii. Individuals engaging in seasonal or festival-related spending surges.

iii. Potential cases of irregular or risky financial behavior.

11) Cluster Trends Across Months



Insights

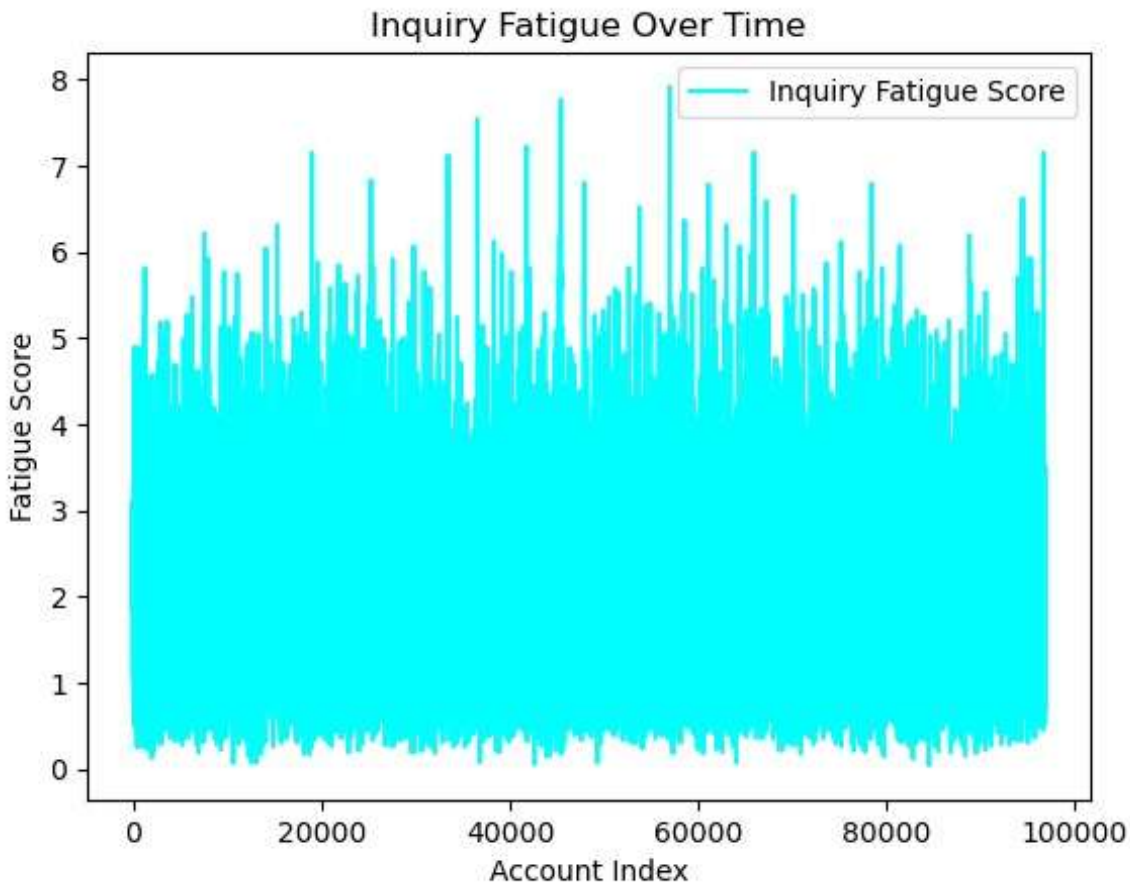
- **Cluster 0 (Blue)**: Low and stable usage across most months.
- **Cluster 1 (Orange)**: Moderate spikes in a few months but relatively stable otherwise.
- **Cluster 2 (Green)**: Sharp spikes in select months (sporadic behavior).
- **Cluster 3 (Red)**: Extremely high spikes, indicating anomalies or special events.

Ivferneceσ

- **Cluster 0**: Represents low-usage customers who may not be a priority for targeted campaigns or interventions.
- **Cluster 1**: Users with consistent medium usage; potential candidates for upselling or promotional offers.
- **Cluster 2**: Sporadic high-utilization users; could be responding to specific events or emergencies.
- **Cluster 3**: Extremely high usage in specific months. These may be:
    - Outliers or anomalies in the data.
    - High-priority users with significant financial contributions or risks.

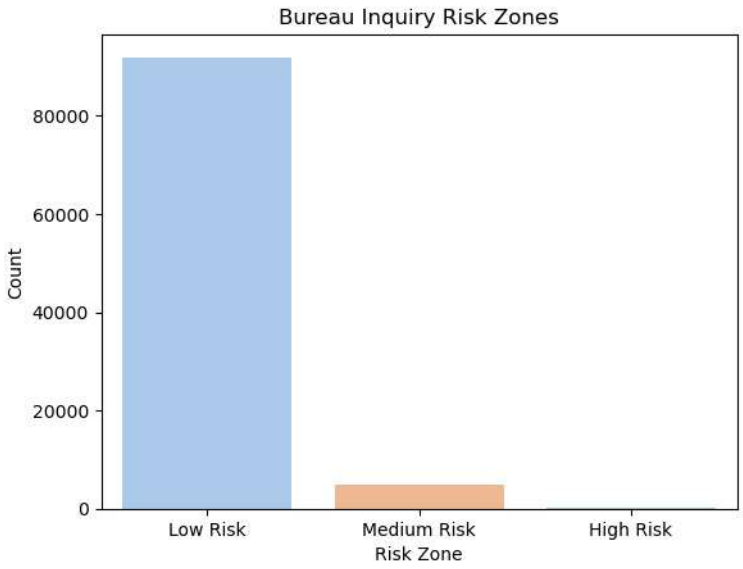
12) Inquiry Fatigue

The chart titled "Inquiry Fatigue Over Time" displays the variation in fatigue scores across different account indices. The X-axis represents the account index, while the Y-axis represents the fatigue score, ranging from 0 to 7. The analysis reveals that fatigue scores are relatively consistent, with no significant upward or downward trend over the account indices. However, occasional spikes in fatigue scores are observed, indicating accounts with unusually high levels of inquiry fatigue. These outliers could be indicative of specific scenarios such as excessive inquiries or unique stress points in the system. The general variability suggests that inquiry fatigue is likely influenced by external factors, such as inquiry volume or account-specific characteristics, rather than by account indices alone. To address this, it is recommended to investigate accounts with higher fatigue scores to identify root causes, explore correlations with other variables like inquiry frequency or account activity, and implement strategies to mitigate fatigue in high-risk accounts.



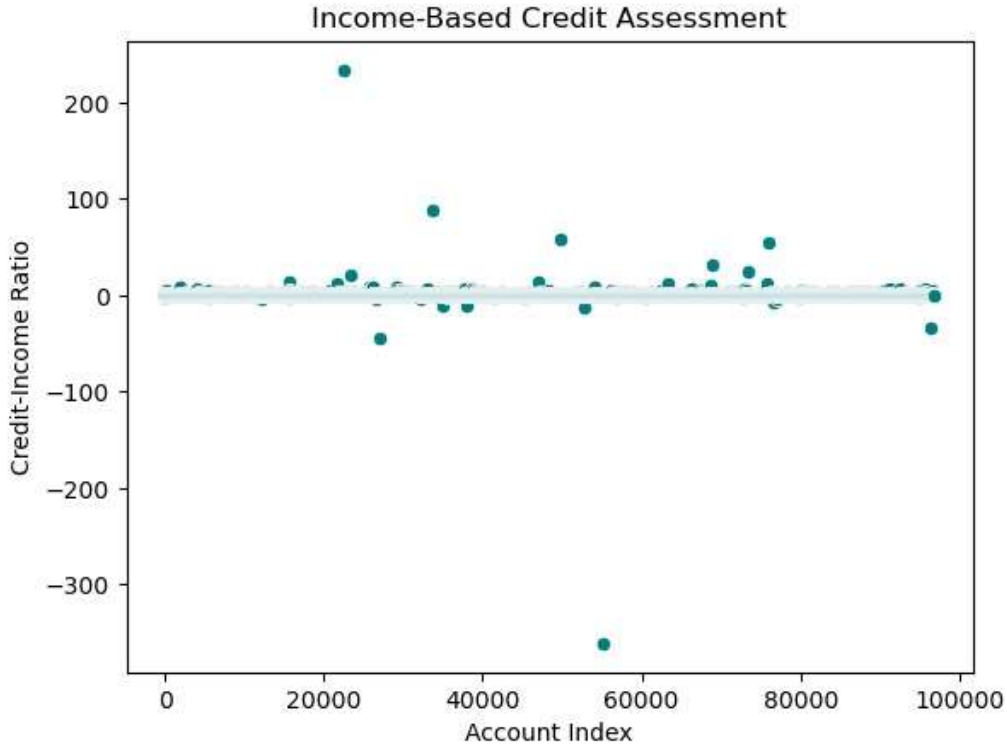
13) Bureau Inquiry Risk Zone

The chart shows the distribution of individuals categorized into three different Bureau Inquiry Risk Zones: **Low Risk**, **Medium Risk**, and **High Risk.**



- **Observation**:
  The majority of individuals fall into the **Low Risk** category, with a count significantly higher than the other risk zones.
  The **Medium Risk** category accounts for a smaller but noticeable proportion of individuals.
  The **High Risk** category is almost negligible, with minimal representation.
- **Insights**:
  The data suggests that most individuals assessed by the bureau have a relatively safe risk profile, which indicates a generally stable credit or behavioral profile.
  The low representation in the **High Risk** zone implies that a very small percentage of individuals are considered high-risk. This could indicate effective screening or a generally low-risk population.
- **Derivations**:
  Financial or operational strategies can focus on maintaining and catering to the **Low Risk** population, as they form the largest segment.
  Monitoring systems and policies should be enhanced for the **Medium Risk** category to prevent migration to the **High Risk** zone.
  The minimal size of the **High Risk** group suggests limited exposure to potential risks; however, this group should be closely monitored to mitigate any adverse outcomes.

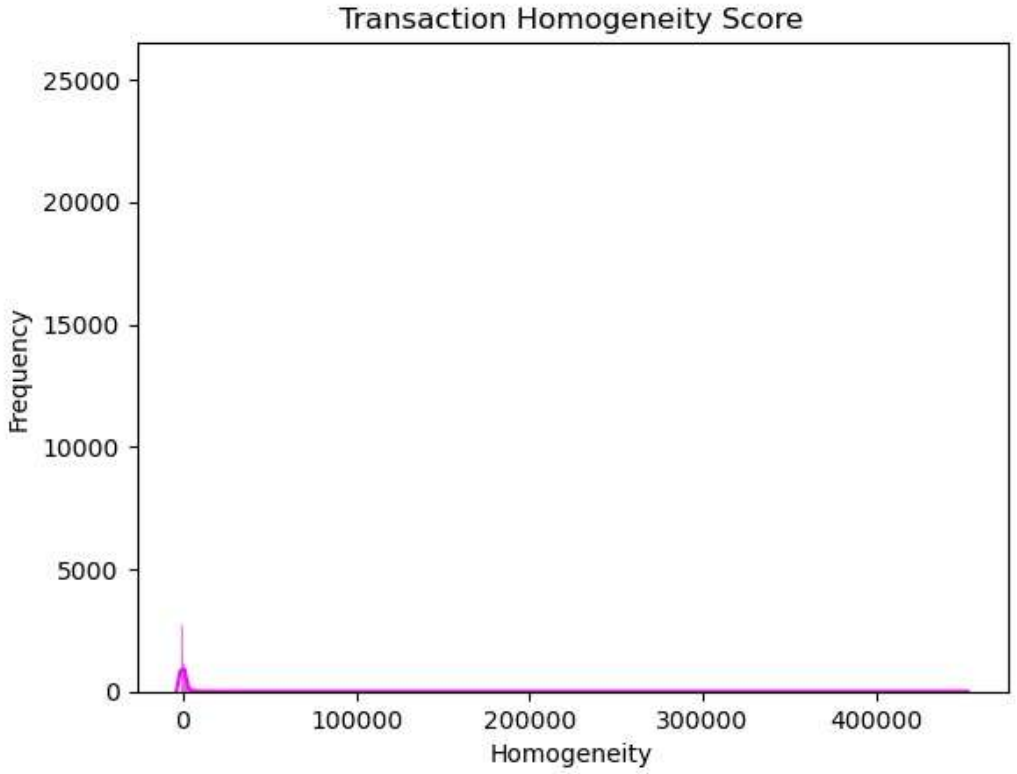
14) Income Based Credit Assesment



The chart shows a scatter plot of **Credit-Income Ratio** against an **Account Index**

- **Observation**:
  Most data points are concentrated around a **Credit-Income Ratio** of 0, indicating a balanced relationship between credit and income for the majority of accounts.
  A few outliers exhibit significantly higher or lower credit-income ratios, with values extending beyond 100 and below -300.
- **Insights**:
  The clustering around a ratio of 0 suggests that most accounts maintain a proportionate credit utilization relative to income, indicating financial stability for this group.
  The presence of outliers with extremely high or low ratios could point to anomalies, such as individuals with either excessive borrowing or disproportionately high income compared to credit usage.
  Negative credit-income ratios may indicate accounts with either excess liabilities or misclassified data.
- **Derivations**:
  Accounts with extreme positive ratios may represent potential credit risks due to over-leveraging and should be monitored or investigated further.
  Negative ratios might indicate financial distress or data recording issues, and these accounts may require intervention or correction.
  Given the concentration of points near zero, credit policies and strategies could target maintaining this balanced segment as the majority.

15) Transition Homogeneity Score



- **Observation**:
  The majority of data points are concentrated near low **Homogeneity Scores**, creating a sharp peak around the lower end of the x-axis.
  The frequency drops off rapidly as the **Homogeneity Score** increases, with very few occurrences of higher scores.
- **Insights**:
  The high concentration of low scores suggests that most transactions exhibit a similar or uniform pattern, indicating a high level of consistency in transaction behavior for the majority of accounts.
  The sparse distribution of higher scores may correspond to anomalies or outlier accounts with highly varied or unique transaction patterns.
- **Derivations**:
  The high uniformity reflected in the data indicates a stable and predictable transaction environment for most accounts.
- The outliers with significantly higher scores should be further analyzed as they may represent accounts with irregular activity, potential fraud, or special circumstances.
- Policies or systems could leverage the homogeneity of the majority by designing standardized solutions while maintaining a mechanism to flag and review outliers.

Data Preprocessing

1. **GPU Availability Check**

- Ensured that GPU resources are available for model training.

2. **Load Dataset**

- Loaded the development dataset.

3. **Drop Unnecessary Columns**

- Dropped the account_number column if present.

4. **Handle Missing Values**

- Dropped columns with all missing values.
- Dropped columns with more than 80% missing values

5. **Define Target and Features**

- Defined bad_flag as the target variable and the rest as features.

6. **Impute Missing Values**

- Used SimpleImputer to impute missing values with the mean.

7. **Handle Class Imbalance**

- Used SMOTETomek to handle class imbalance by oversampling the minority class and undersampling the majority class.

8. **Split the Data**

- Split the balanced dataset into training and testing sets.

9. **Standardize the Features**

- Standardized the features using StandardScaler.

10. **Custom F1 Score Metric**

- Defined a custom F1 score metric for model evaluation.

Neural Network Model Architecture

Overview

The neural network model in code is designed to perform binary classification using a combination of convolutional layers and fully connected layers. The architecture is built to handle and learn from complex patterns in the data.

Detailed Architecture

Input Layer:

The input layer accepts data with the shape (input_dim, 1), where input_dim is the number of features in the dataset. The input data is reshaped to fit this requirement.

Convolutional Layers:

First Conv1D Layer:

Filters: 64

Kernel Size: 3

Activation Function: ReLU (Rectified Linear Unit)

This layer applies 64 filters to the input data with a kernel size of 3, extracting local features from the data.

Batch Normalization:

This layer normalizes the output of the preceding Conv1D layer, which helps in accelerating the training process and improving the model's performance.

Dropout:

Rate: 0.5

This layer randomly drops 50% of the neurons during training to prevent overfitting.

Second Conv1D Layer:

Filters: 64, Kernel Size: 3, Activation Function: ReLU

Another convolutional layer with the same configuration to further extract features.

Batch Normalization: Normalizes the output of the second Conv1D layer.

Dropout:

Rate: 0.5, Another dropout layer to prevent overfitting.

Flatten Layer: This layer flattens the output of the last Conv1D layer into a one-dimensional array, preparing it for the fully connected layers.

Fully Connected (Dense) Layers

First Dense Layer: Units: 64

Activation Function: ReLU

This layer has 64 neurons and applies the ReLU activation function. It learns complex patterns from the flattened input.

Dropout:

Rate: 0.4, Drops 40% of the neurons during training to prevent overfitting.

Second Dense Layer:

Units: 32

Activation Function: ReLU

Reduces the number of neurons to 32, continuing to learn from the previous layer's output**.**

Dropout:

Rate: 0.4, Another dropout layer with a 40% dropout rate.

Third Dense Layer:

Units: 16

Activation Function: ReLU

Further reduces the number of neurons to 16, refining the learned patterns.

Output Layer

Units: 1

Activation Function: Sigmoid

This layer has a single neuron with a sigmoid activation function, which outputs a probability value between 0 and 1, indicating the class of the input data.

Note: The threshold for a credit card holder defaulting is considered to be **0.4** instead of 0.5 because the credit card issuing company will not like to take chances so being on the safer side we can arguably give higher accuracy in preventing and identifying the defaulter.

Compilation and Optimization

Optimizer: Adam, Learning Rate: 0.001

Adam optimizer is used for training, which adjusts the learning rate adaptively based on the training process.

Loss Function: Binary Crossentropy.**This is suitable for binary classification problems.**

Metrics: Accuracy and Custom F1 Score

The model is evaluated based on accuracy and a custom F1 score metric, which balances precision and recall.