Bank Debit Collection Analysis and Prediction

Naufal Sulthan Dila

1. Data Loading and Preparation

A. Import Dataset

To begin the analysis, I imported the dataset into Tableau. The process involves the following steps:

1) Open Tableau and connect to the data source. In this case, I used a .xlsx file which contains data on loan products, recovery rates, and outstanding balances.
2) After selecting the file, Tableau automatically detects the structure of the dataset, showing a preview of the rows and columns.
3) I verified the data types (e.g., numeric, string) and adjusted them as necessary, ensuring that fields such as loan amounts are recognized as numbers, and date columns are properly categorized.
4) Finally, I clicked on the "Sheet" tab to proceed to the data visualization workspace where the dataset was ready for analysis.

B. Data Cleaning

After importing the dataset into Tableau, I performed several data cleaning steps to ensure the dataset was ready for analysis:

1) Handling Missing Data:

- I reviewed the dataset to identify any null or missing values, particularly in important fields such as recovery rates, loan products, and outstanding balances.
- Missing values were either excluded from the analysis or replaced with appropriate placeholders depending on the context. For instance, if a recovery rate was missing, I analyzed other metrics to determine if it could be reasonably inferred.

2) Correcting Data Types:

- Tableau sometimes incorrectly categorizes data, especially date fields or numerical values. I ensured that all date fields were properly formatted as date types (e.g., Month-Year), and that numerical fields like Outstanding Balance and Payment_for_Fee were correctly classified as numbers.

3) Removing Duplicates:

- I scanned the dataset for any duplicate records, which can lead to inaccurate calculations, and removed any that were found to maintain data integrity.

4) Filtering Irrelevant Data:

- Some fields or records, such as test data or incomplete loan statuses, were filtered out as they were not relevant to the scope of the analysis.

2. Data Transformation

A. Convert Data

In the data transformation phase, converting data into the proper format is crucial to facilitate accurate analysis and visualization. Here are the key steps I took:

1) Converting Dates:

- The dataset included separate Month and Year fields. I combined these into a single Month-Year field to create a continuous timeline for trend analysis.
- I ensured that this new Month-Year field was treated as a date in Tableau to enable chronological sorting and accurate time-series visualizations in the dashboard.

2) Converting Categorical Data:

- Some fields, such as Product_final (loan product categories) and Range_DPD (Days Past Due ranges), were initially in text form. These were converted into dimensions to be used for segmentation and comparison in various visualizations.
- The Range_DPD field was also sorted logically (from 0-180 to 1080+ and Deadloan) to create meaningful visual breakdowns based on overdue periods.

3) Normalizing Financial Data:

- I standardized the numerical financial data such as Outstanding Balance and Payment_for_Fee by ensuring consistent units (e.g., converting to millions or billions where appropriate) so that they could be meaningfully compared in visualizations without scale distortions.

B. Create Calculated Columns

To calculate the average Days Past Due (DPD) from the Range_DPD field, we need to convert these text-based ranges into numerical values. Here's how to approach this:

1) Create a Calculated Field to Convert Range_DPD to Numeric Values:

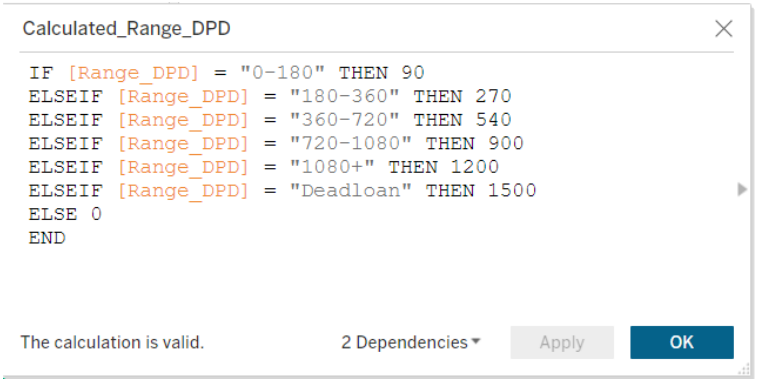
- We can create a calculated field in Tableau to assign a numerical value to each Range_DPD category. The simplest approach is to use the midpoint of each range as the numerical value. Here's the logic for assigning numeric values:

  0-180: Take the midpoint of this range, which is 90.
  180-360: The midpoint is 270.
  360-720: The midpoint is 540.
  720-1080: The midpoint is 900.
  1080+: You can use an assumed value of 1200 for more than 1080 days.
  Deadloan: You can assign a high value, such as 1500, or treat it as unrecoverable.

(blank): Blank values can be treated as 0, or ignored.

2) Formula for the Calculated Field in Tableau:

- To implement this conversion, create a calculated field in Tableau using the following formula:



- This field will now convert the categorical DPD ranges into numeric values, allowing us to perform calculations such as finding the average DPD for different loan products or partners.

3. Data Modeling

A. Create Measures

In this section, we focus on creating measures (KPIs) that help analyze and track key metrics for the bank's debt collection performance. Since the dashboard was built using Tableau, the following are the key calculated fields (measures) that were created:

1) Total Loan Recovery:

Formula: SUM([Payment_for_Fee])

Purpose: This measure tracks the total amount recovered during the reporting period across all loan products and partner companies. It helps visualize the success of recovery efforts.

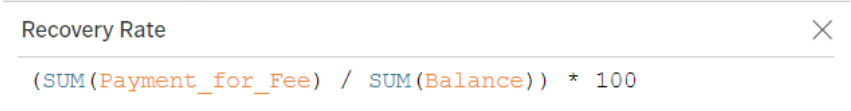
2) Total Outstanding Balance:

Formula: SUM([Balance])

Purpose: This measure shows the total outstanding loan amount that remains unpaid after recovery efforts. It provides an overall view of how much money is still owed.

3) Recovery Rate:

Formula: (SUM([Payment_for_Fee]) / SUM([Balance])) * 100



Purpose: The recovery rate provides a percentage indicating how much of the total outstanding balance has been successfully recovered. It gives insight into the effectiveness of the debt collection process.

4) Average Days Past Due (DPD):

Formula: AVERAGE([Range_DPD])

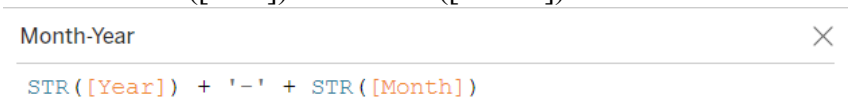
Purpose: This measure calculates the average overdue period for loans. Understanding the average Days Past Due helps the bank identify

whether overdue loans are a growing problem and target recovery efforts more effectively.

5) **Outstanding Balance by Partner:**
Formula: SUM([Balance]) (grouped by Os_Company)
Purpose: This KPI compares the total outstanding balance across different partner companies. It identifies which partners have the highest unpaid loan amounts, highlighting potential risks and areas for improvement.

6) **Recovery by Loan Product:**
Formula: SUM([Payment_for_Fee]) (grouped by Product_final)
Purpose: This KPI evaluates which loan products have the highest recovery amounts. It helps assess the performance of each loan type in terms of recovery efforts and identify products that are more challenging to recover.

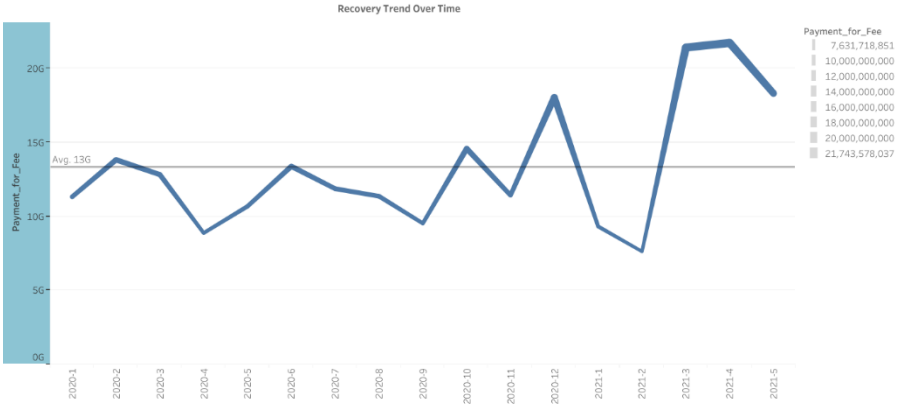
7) **Month-Year:**
Formula: STR([Year]) + '-' + STR([Month])



Purpose: A calculated field combining the Year and Month fields to generate a time series. This allows for plotting recovery trends over time, which is essential for monitoring the performance across different periods.

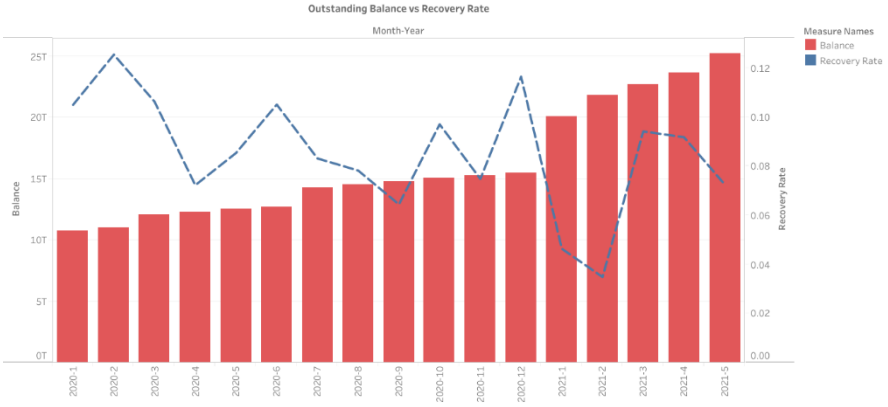
4. Dashboard Design

A. Recovery Trend Over Time



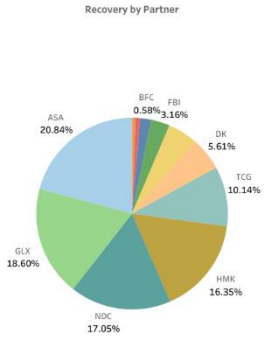
X-axis: Month-Year (combine Year and Month fields).
Y-axis: SUM(Payment_for_Fee) (Total Loan Recovery).
Reason: This chart will help visualize the recovery trend across months and years, providing insights into seasonality or recovery performance over time. It shows if recovery is improving or declining in recent periods.
Filters: Add slicers for Os_Company and Product_final to compare partner or product-specific trends.

B. Outstanding Balance vs Recovery Rate



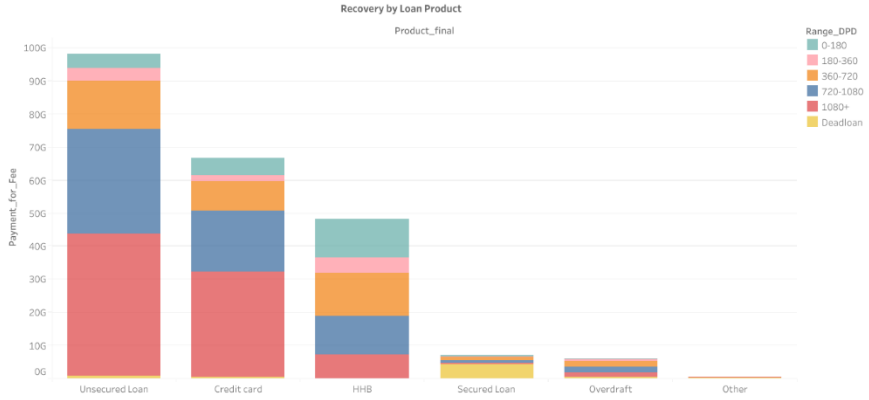
X-axis: Month-Year.
Bar (Y1): SUM(Balance) (Total Outstanding Balance).
Line (Y2): Recovery Rate.
Reason: This chart compares the outstanding balance against the recovery rate to understand the effectiveness of debt collection efforts. It reveals if higher balances result in better recovery or if recovery is lagging behind large outstanding amounts.

C. Recovery by Partner



Categories: Os_Company.
Values: SUM(Payment_for_Fee) (Total Loan Recovery).
Reason: The pie chart breaks down recovery amounts by partner company, providing insights into the contribution of each firm to the overall recovery. It helps identify the top-performing partners and those lagging behind in collections.

D. Recovery by Loan Product



X-axis: Product_final.
Y-axis: SUM(Payment_for_Fee) (Total Recovery).
Stack: Range_DPD to categorize by overdue periods.
Reason: This stacked bar chart compares the loan recovery by product type, showing which products are performing well in recovery efforts. The DPD stack adds another layer of analysis to see which products have the most overdue loans.

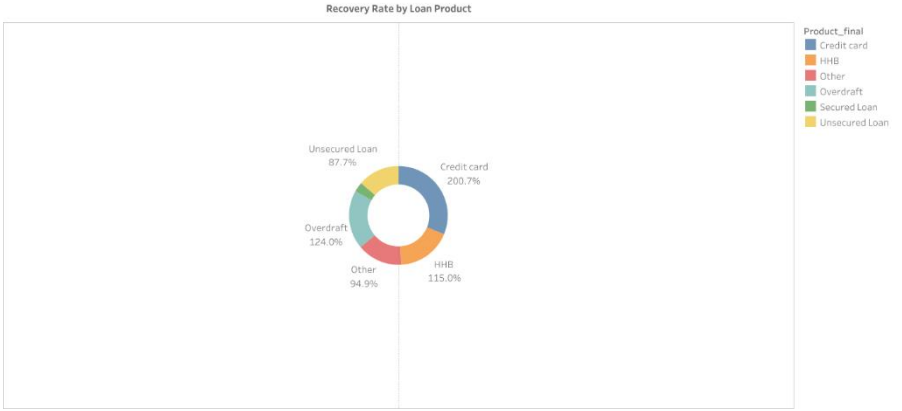
E. Average Days Past Due by Loan Product

X-axis: Product_final.
Y-axis: Range_DPD.
Color Scale: Represent the average Range_DPD.
Reason: The heat map will visually represent which loan products have longer overdue periods, allowing the bank to prioritize recovery strategies for those products with higher risk.

F. Partner-wise Outstanding Balance vs Recovery Amount

X-axis: Os_Company.
Bars (Y1): SUM(Balance) (Total Outstanding Balance).
Bars (Y2): SUM(Payment_for_Fee) (Total Recovery Amount).
Reason: This chart directly compares the total outstanding balance and the amount recovered by each partner. It highlights discrepancies between partners with high balances and low recovery, signaling areas for improvement.

G. Recovery Rate by Loan Product



Value: Recovery Rate by Product_final.
Reason: A gauge chart can be used to quickly assess how well each loan product is performing in terms of recovery rate. This helps in identifying the most successful products and those requiring attention.

5. Final Touches

In the final stage of building the dashboard, we focus on improving the presentation, usability, and functionality of the visualizations to ensure they are clear, interactive, and easy to interpret.

A. Formatting

1) Consistent Design:

I applied consistent colors, fonts, and styles throughout the dashboard to ensure a professional look and clear communication of information. For example, similar types of visuals share the same color schemes, making it easier to compare data across different charts.

2) **Titles, Axis Labels, and Legends:**
Each chart is given clear titles that describe the data being presented. Axis labels are used to indicate units and legends are added to clarify what different colors represent in multi-dimensional visuals like stacked bar charts and heat maps.

B. Tooltips and Interactivity

1) Tooltips:

I added tooltips to every chart, which provide detailed data points when users hover over elements. For example, in the recovery trend chart, hovering over a specific month shows the exact recovery amount and outstanding balance for that period.

2) Slicers and Filters:

Filters such as Os_Company, Product_final, and Range_DPD are implemented to allow dynamic interaction. The slicers ensure that all visuals update accordingly when selections are made, enabling users to explore data from multiple perspectives. This makes the dashboard more interactive and insightful for decision-makers.

C. Testing

1) Testing with Filters:

The dashboard was tested extensively using different combinations of slicers and filters. This ensures that the data updates correctly and consistently across all visualizations. For example, selecting specific partners or products results in accurate updates for both the outstanding balance and recovery rate charts.

2) Cross-visual Interaction:

I verified that interactions between different visualizations, such as clicking on a section of a pie chart to filter other charts, work seamlessly. This enhances the user's ability to drill down into specific details.

Conclusion for Bank Debit Collection Analysis and Prediction

1. Recovery Trend Over Time: Recovery payments have increased over time, but fluctuations in recent months are worth noting. At its peak in 2021, payments were over 20T.
2. Recovery by Partner: ASA contributed around 20.84% of total recovery, followed by GLX and NDC. Partners with low recovery, such as HNA, may require more attention.
3. Recovery by Loan Product: The largest recovery occurred in Unsecured Loans and Credit Cards with a significant contribution from delays above 720 days, while products such as HHB showed faster recovery although with lower total recovery.
4. Average Days Past Due by Loan Product: Products such as unsecured loans and credit cards have higher Days Past Due (DPD), especially in the category of more than 1080 days. This indicates a greater risk in these products.

5. Partner-wise Outstanding Balance vs Recovery Amount: Partners like GLX and ASA have high outstanding balances but relatively low recoveries. While some other partners, like DK, show significant recoveries despite lower balances.
6. Recovery Rate by Loan Product: Credit cards have the highest recovery rate, over 200%, while secured loans have a lower recovery rate (20.3%). This could be an indication of a strategy that is effective in some products but needs improvement in others.