Final Project Report

Vehicle-Loan Credit Assessment

Exploring Loan Default Patterns Through Advanced Visual Analytics and Data Driven Insights

Vishnuvardhan Reddy Kollu
viskollu@iu.edu

Santoshi Borra
vborra@iu.edu

Jaswanth Mandava
jmandava@iu.edu

ABSTRACT

The increasing rate of vehicle loan defaults, particularly in the first Equated Monthly Installment, has led to significant financial losses for lending institutions. In response, the need for an advanced credit risk scoring model has become risky and tough. A model that visualizes the data required to analyze the risks and cases can be developed. It helps to accurately predict the chances of loan default based on borrower features, ranging financial data, and credit history. With the large range of data available everywhere, we have more chances to assess the risks, but it is also important to choose the right features to avoid confusion. To see the borrower's credit risk, it helps to understand their past loan performance. The dataset used for this project includes a variety of important variables such as borrower demographics, financial details, and historical credit information. Demographic variables include indicators such as the borrower's age, state names, branch ID's, employment status, and even the type of identification documents provided. These social variables provide information about the borrower's security and trustworthiness, impacting their ability to repay loans. The geographical analysis helps to detect the hotness of the frauds occurring at certain locations, which are prone to scams.

INTRODUCTION

Motivation

Vehicle loans are an important part of the financial system because they help people buy cars by taking out loans from banks or credit companies. These organizations provide borrowers with loans under specific conditions, usually requiring the borrower to repay the money gradually or monthly, known as Equated Monthly Instalments (EMI). When borrowers fail to pay their EMI, it is a growing concern for financial institutions, resulting in major financial losses. A salaried employee might be considered more stable and reliable due to consistent income, while a self-employed individual might be viewed as riskier. The PAN and Aadhar cards (identity proofs) allow the parties to verify the borrower's identity and cross-check it with government databases. Lenders also look at the borrower's credit history, which includes their credit score, the number of active loans they currently have, whether they have defaulted on any loans in the past, and how much debt they are already carrying. Also borrowers in certain regions may experience unique economic pressures or employment trends that influence their repayment behavior. The rising default rate has tightened the situation of basic standards, resulting in higher loan rejection rates. The tough thing is to find a balance between financial security and ensuring that clients who have worthwhile credit are approved. So, what we have decided is to provide a detailed visualization dashboard so that individuals can make more informed, data-driven decisions and have a better understanding of the factors that have a large impact.

Existing Visualizations

There is a significant amount of research on uncertainty analysis and visualization with a variety of uncertainty metrics defined and reframed, particularly when it comes to understanding the variability and risk involved in financial decisions, like vehicle loan defaults. Here, uncertainty metrics such as credit scores, loan-to-value ratios (LTV), and risk classifications are crucial. Each of these metrics carries some degree of uncertainty, whether it's the potential fluctuation in a borrower's credit score over time, or the variability in how closely a borrower's risk category matches their actual repayment behavior. But incorporating interactive features could help decision-makers explore specific borrower profiles or further analyze how risk metrics change over time, providing a clearer picture of uncertainty in loan default predictions.

```
In [291]: plt.figure(figsize=(15,5))
          sns.countplot(df['State_ID'])
          plt.title("State_ID")
          plt.show()
```

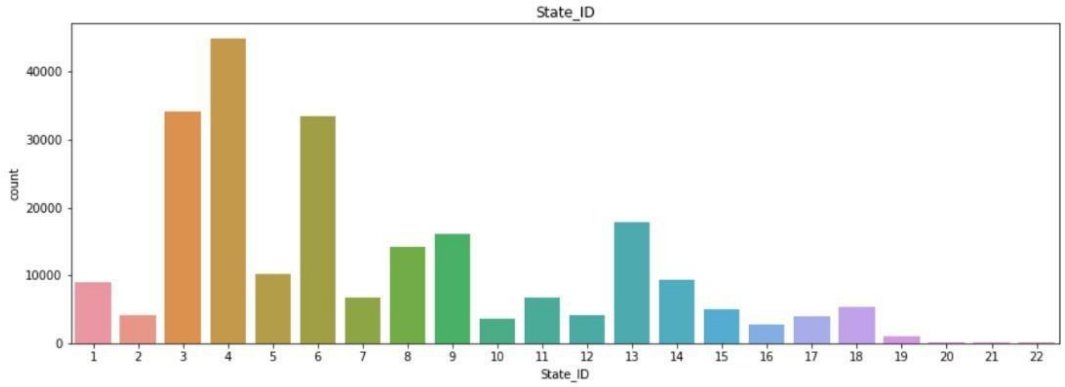


Fig 1: Count plot: Loan Distribution by State

This chart shows the distribution of total loans disbursed by state, without focusing on default rates. It helps in understanding which regions have more loan activity. We can merge this plot with loan default rates, overlaying or color-coding the default percentages. This would make it easier to identify states with higher loan activities and simultaneously high default rates.

```
In [340…
plt.figure(figsize=(10,5))
sns.scatterplot(df["disbursed_amount"],df["ltv"],hue=df['loan_default'],markers="x")
plt.show()
```

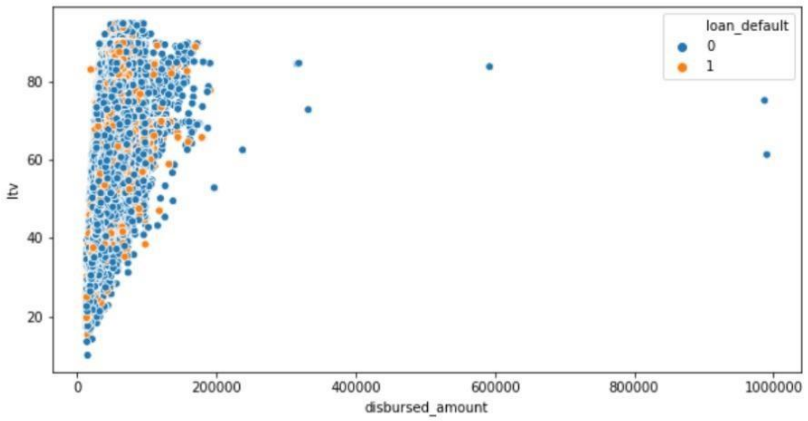


Fig 2: Scatter Plot: Loan Amount(vs)Loan-to-Value (LTV) Ratio with Default Indicator

This scatter plot examines the relationship between the loan disbursed amounts and the loan-to-value ratios while showing whether a loan defaulted or not. While the scatter plot is effective for identifying outliers, we could further improve it by adding trend lines or clustering similar data points. We could also segment the data by employment type or credit score to see if these factors shift the relationships.

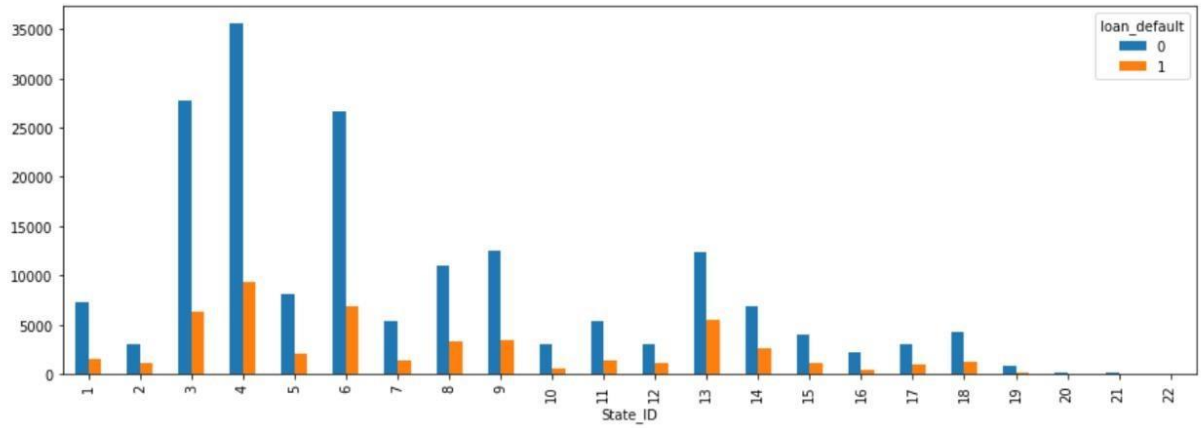


Fig 3: The bar graph: Loan Default by State

This chart visualizes the distribution of loan defaults across different states, helping financial institutions understand regional differences in default rates. This bar chart gives a clear picture but can be enhanced by normalizing the data. Instead of showing absolute counts, showing percentages or default rates per state would make comparisons more meaningful. We could also add another layer of analysis, such as filtering by borrower characteristics.

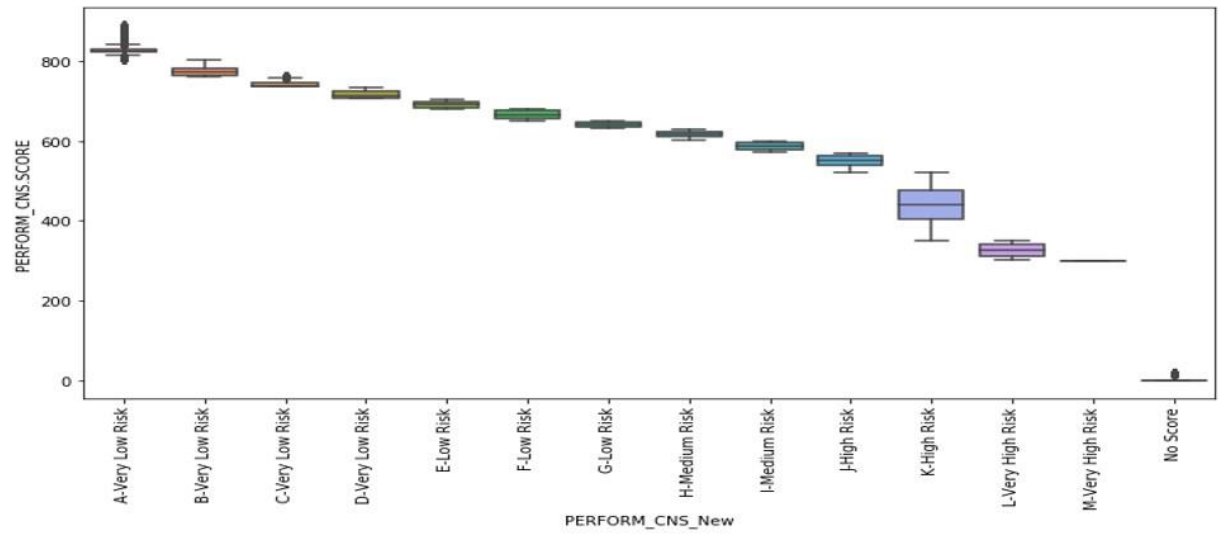


Fig 4: The box plot: Distribution of Credit Scores by Risk Category

The box plot helps to categorize borrowers into risk groups based on their Credit Bureau Score, allowing analysts to visually assess the distribution of scores across risk levels. Each box represents the interquartile range (IQR) of scores within a specific risk category, showing the spread of scores. While this static box plot provides useful insights, we can improve it by offering interactive elements where users can hover over data points to see more details about specific risk groups.

Contribution

Idea 1: Regional Analysis of Loan Defaults and Disbursement Patterns

This idea focuses on understanding how loan default rates, total disbursed amount, and asset costs vary across geographical areas. By identifying high-risk regions and areas with substantial loan activity, the analysis aims to uncover geographic trends that influence loan performance and credit risk. These insights can help prioritize regions for targeted interventions and policy development.

Idea 2: Impact of Employment Type on Default Risk

Examining the relationship between borrower employment types and default rates, this idea seeks to determine which employment categories are more prone to loan defaults. Also it explores the average loan amounts disbursed and repaid by each employment group, to show a clearer picture of borrower reliability and repayment patterns.

Idea 3: Temporal Trends in Borrower Performance

This idea studies the changes in disbursed amounts, asset costs, and loan default rates over time, highlighting trends and patterns in lending practices and borrower behavior. The goal is to identify shifts in credit risk or financial behaviors that align with economic or institutional changes.

Idea 4: Correlation Between Credit Metrics and Borrower Risk

This idea delves into the relationships between critical metrics such as loan-to-value ratio (LTV), credit history length, and delinquencies, seeking to understand how these factors collectively influence the likelihood of defaults. The analysis aims to identify the most significant predictors of credit risk, enabling better decision-making for lending institutions.

What do they answer?

1. What are the regional differences in loan metrics across states?
2. Which employment types are most prone to defaults, and how does this vary across states?
3. How have loan disbursements, asset costs, and defaults changed over time?
4. What are the key correlations between metrics like LTV, credit history length, and loan defaults?
5. Are there specific geographic and demographic groups more at risk of loan defaults?
6. How do borrower repayment behaviors correlate with loan performance and defaults?
7. What patterns can be observed in loan performance based on borrower demographics like age and employment type?
8. How do borrower behaviors, such as inquiries and delinquencies, affect their likelihood of default?

PROTOTYPES

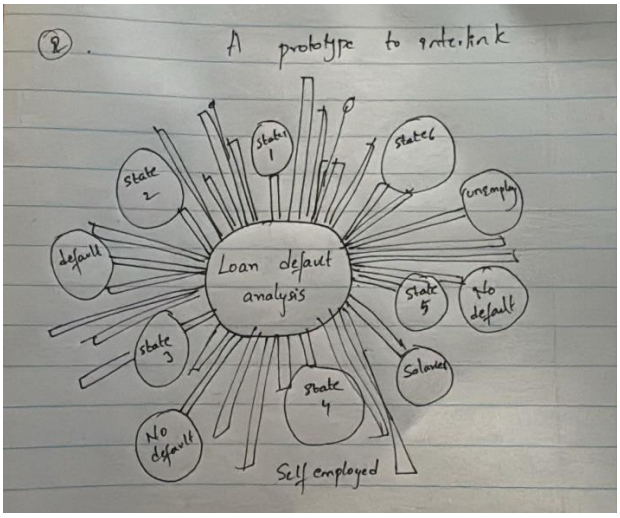


Fig 5: Links between state names, employment types and loan defaults

Imagine starting with all loan accounts as a single big circle (the root). This circle is then divided into smaller sections, representing states, employment types, and finally, whether the loans are in default or not. This is the initial prototype.

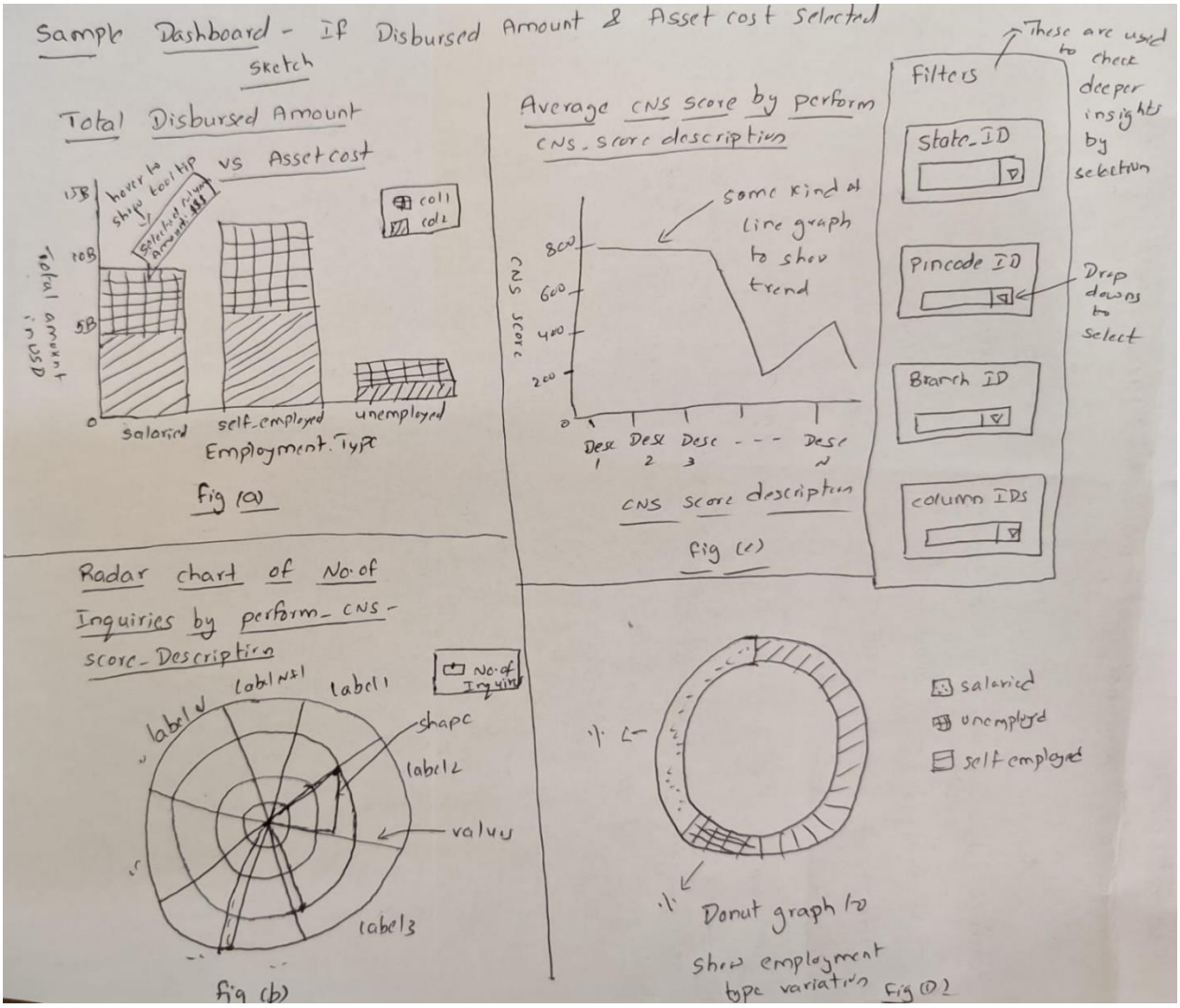


Fig 6: Prototype of Interactive dynamic dashboard

Imagine if we want to create static graphs to visualize geographical trend across country, it will be too clutter and highly difficult to understand trend easily. So, we came up to reduce manual efforts and create an interactive dashboard. The above prototype is our view to establish so that everything will be interactive and dynamically change

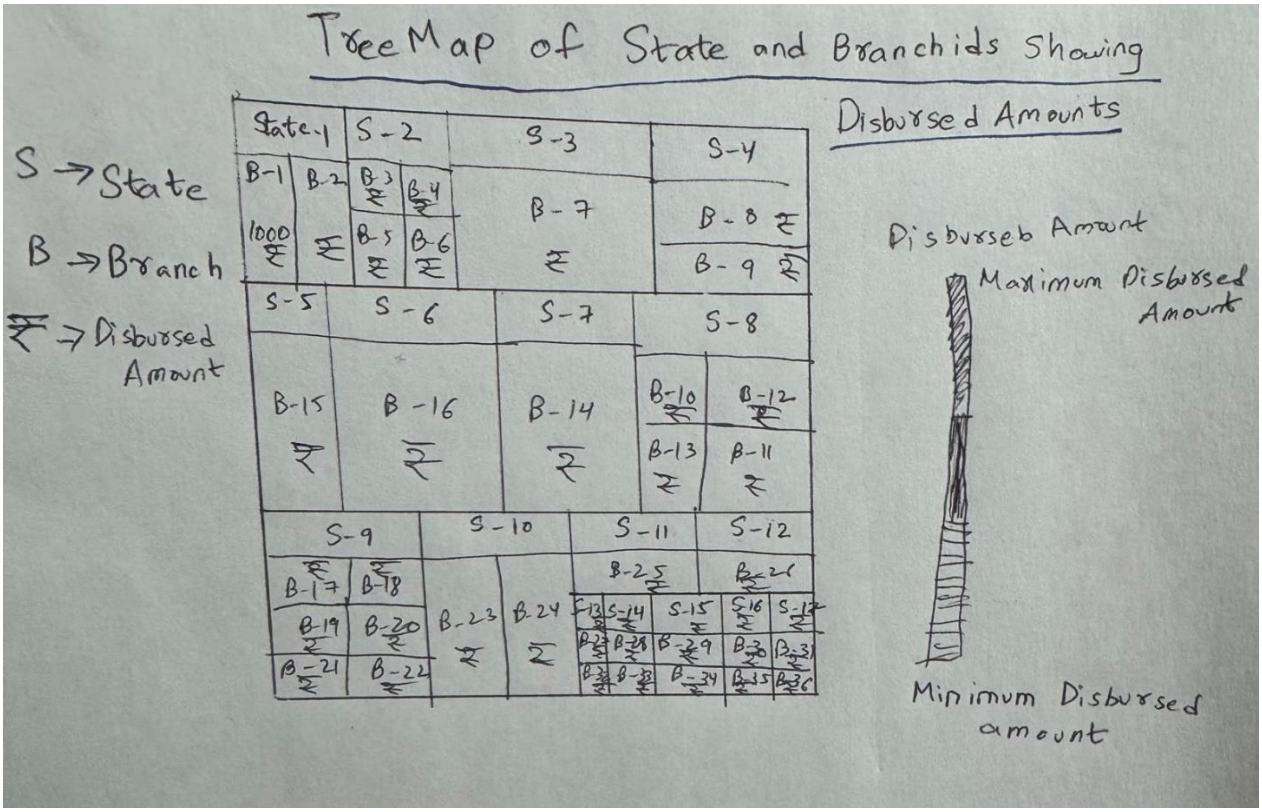


Fig 7: TreeMap of Disbursement Amounts

The sketch is a rough sketch for a tree map visualization, illustrating the distribution of disbursed amounts across various states and branches. Each large rectangle represents a state (labeled S-1 to S-12), subdivided into smaller rectangles, each representing a branch within the state (labeled B-1 to B-36 and beyond). The size of each branch rectangle correlates with the amount disbursed, enabling a quick visual assessment of financial distribution across geographic locations. The tree map employs a color gradient to represent disbursed amounts, with darker shades possibly indicating higher disbursals and lighter shades for lower disbursals, though specific color details are not included in the sketch. Adjacent to the tree map, a bar indicating the range of disbursed amounts from minimum to maximum, providing a scale for understanding the depth of disbursal values represented within the tree map.

DATASET AND METHODS

The dataset we have chosen for this analysis comes from a Kaggle competition focused on vehicle loan default prediction. It includes detailed records for 233,154 loan applications, structured in a tabular format with 41 diverse attributes. The dataset has null values for only one column, employment type (7661 values). Apart from that, all the columns are not having null values. This varied dataset not only allows for a multi-dimensional analysis of potential risk factors but also supports a thorough exploration of demographic impacts on loan performance. These attributes range across several data types:

Numeric: Financial amounts like disbursed_amount and asset_cost, and ratios such as ltv, loan to value etc. belong to Numeric data type.

Categorical: Variables, including employment.type and perform_cns.score.description.

Binary: Indicators such as whether identification documents were provided: Aadhar flag, PAN flag, Passport flag, Voter id flag, Driving flag.

Dates: Important temporal data like date of birth and disbursal date.

Variable Name	Description
UniqueID	Identifier for customers
loan_default	Payment default in the first EMI on due date
disbursed_amount	Amount of Loan disbursed
asset_cost	Cost of the Asset
ltv	Loan to Value of the asset
branch_id	Branch where the loan was disbursed
supplier_id	Vehicle Dealer where the loan was disbursed
manufacturer_id	Vehicle manufacturer(Hero, Honda, TVS etc.)
Current_pincode	Current pincode of the customer
Date.of.Birth	Date of birth of the customer
Employment.Type	Employment Type of the customer (Salaried/Self Employed)
DisbursalDate	Date of disbursement
State_ID	State of disbursement
Employee_code_ID	Employee of the organization who logged the disbursement
MobileNo_Avl_Flag	if Mobile no. was shared by the customer then flagged as 1
Aadhar_flag	if aadhar was shared by the customer then flagged as 1
PAN_flag	if pan was shared by the customer then flagged as 1
VoterID_flag	if voter was shared by the customer then flagged as 1
Driving_flag	if DL was shared by the customer then flagged as 1
Passport_flag	if passport was shared by the customer then flagged as 1
PERFORM_CNS.SCORE	Bureau Score
PERFORM_CNS.SCORE.DESCRIPTION	Bureau score description
PRI.NO.OF.ACCTS	count of total loans taken by the customer at the time of disbursement
PRI.ACTIVE.ACCTS	count of active loans taken by the customer at the time of disbursement
PRI.OVERDUE.ACCTS	count of default accounts at the time of disbursement
PRI.CURRENT.BALANCE	total Principal outstanding amount of the active loans at the time of disbursement
PRI.SANCTIONED.AMOUNT	total amount that was sanctioned for all the loans at the time of disbursement
PRI.DISBURSED.AMOUNT	total amount that was disbursed for all the loans at the time of disbursement
SEC.NO.OF.ACCTS	count of total loans taken by the customer at the time of disbursement
SEC.ACTIVE.ACCTS	count of active loans taken by the customer at the time of disbursement
SEC.OVERDUE.ACCTS	count of default accounts at the time of disbursement
SEC.CURRENT.BALANCE	total Principal outstanding amount of the active loans at the time of disbursement
SEC.SANCTIONED.AMOUNT	total amount that was sanctioned for all the loans at the time of disbursement
SEC.DISBURSED.AMOUNT	total amount that was disbursed for all the loans at the time of disbursement
PRIMARY.INSTAL.AMT	EMI Amount of the primary loan
SEC.INSTAL.AMT	EMI Amount of the secondary loan
NEW.ACCTS.IN.LAST.SIX.MONTHS	New loans taken by the customer in last 6 months before the disbursment
DELINQUENT.ACCTS.IN.LAST.SIX.MONTHS	Loans defaulted in the last 6 months
AVERAGE.ACCT.AGE	Average loan tenure
CREDIT.HISTORY.LENGTH	Time since first loan
NO.OF_INQUIRIES	Enquries done by the customer for loans

```
In [10]: df.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 233154 entries, 0 to 233153
Data columns (total 41 columns):
 #   Column                               Non-Null Count   Dtype
---  ------                               --------------   -----
 0   UniqueID                             233154 non-null  int64
 1   disbursed_amount                     233154 non-null  int64
 2   asset_cost                           233154 non-null  int64
 3   ltv                                  233154 non-null  float64
 4   branch_id                            233154 non-null  int64
 5   supplier_id                          233154 non-null  int64
 6   manufacturer_id                      233154 non-null  int64
 7   Current_pincode_ID                   233154 non-null  int64
 8   Date.of.Birth                        233154 non-null  object
 9   Employment.Type                      225493 non-null  object
 10  DisbursalDate                        233154 non-null  object
 11  State_ID                             233154 non-null  int64
 12  Employee_code_ID                     233154 non-null  int64
 13  MobileNo_Avl_Flag                    233154 non-null  int64
 14  Aadhar_flag                          233154 non-null  int64
 15  PAN_flag                             233154 non-null  int64
 16  VoterID_flag                         233154 non-null  int64
 17  Driving_flag                         233154 non-null  int64
 18  Passport_flag                        233154 non-null  int64
 19  PERFORM_CNS.SCORE                    233154 non-null  int64
 20  PERFORM_CNS.SCORE.DESCRIPTION        233154 non-null  object
 21  PRI.NO.OF.ACCTS                      233154 non-null  int64
 22  PRI.ACTIVE.ACCTS                     233154 non-null  int64
 23  PRI.OVERDUE.ACCTS                    233154 non-null  int64
 24  PRI.CURRENT.BALANCE                  233154 non-null  int64
 25  PRI.SANCTIONED.AMOUNT                233154 non-null  int64
 26  PRI.DISBURSED.AMOUNT                 233154 non-null  int64
 27  SEC.NO.OF.ACCTS                      233154 non-null  int64
 28  SEC.ACTIVE.ACCTS                     233154 non-null  int64
 29  SEC.OVERDUE.ACCTS                    233154 non-null  int64
 30  SEC.CURRENT.BALANCE                  233154 non-null  int64
 31  SEC.SANCTIONED.AMOUNT                233154 non-null  int64
 32  SEC.DISBURSED.AMOUNT                 233154 non-null  int64
 33  PRIMARY.INSTAL.AMT                   233154 non-null  int64
 34  SEC.INSTAL.AMT                       233154 non-null  int64
 35  NEW.ACCTS.IN.LAST.SIX.MONTHS         233154 non-null  int64
 36  DELINQUENT.ACCTS.IN.LAST.SIX.MONTHS  233154 non-null  int64
 37  AVERAGE.ACCT.AGE                     233154 non-null  object
 38  CREDIT.HISTORY.LENGTH                233154 non-null  object
 39  NO.OF_INQUIRIES                      233154 non-null  int64
 40  loan_default                         233154 non-null  int64
dtypes: float64(1), int64(34), object(6)
memory usage: 72.9+ MB
```

Fig 8: The dataset table with descriptions of the attributes

DATA PROCESSING

1. Data Loading and Initial Setup: Loaded the dataset from a CSV file. Replaced periods in column names with underscores to ensure compatibility with Python syntax.
2. Data Cleaning: Dropped rows with missing values in the 'Employment_Type' column to maintain data integrity. Addressed missing values across the dataset by applying appropriate imputation and removal methods depending on the nature and impact of the missing data.
3. Feature Engineering:

Transformed 'AVERAGE_ACCT_AGE' and 'CREDIT_HISTORY_LENGTH' from string representations to a numerical format that represents the total number of months. This was done to make easier calculations and comparisons.

Encoded the 'PERFORM_CNS_SCORE_DESCRIPTION' column into a numerical format using a predefined score mapping. This transformation was necessary to convert qualitative score descriptions into quantifiable metrics.

Combined data from primary and secondary accounts into single representative metrics, reducing the complexity and improving the interpretability of the data. New metrics included total accounts, active accounts, overdue accounts, current balance, sanctioned amount, disbursed amount, and installment amount.

Added new calculated features such as total number of accounts and installment amounts by summing corresponding primary and secondary account values to provide a more detailed view of the customer's finances.

4. Data Transformation:

Implemented custom functions to convert date strings into numerical age values, facilitating age-based analyses which are crucial in understanding demographic impacts on loan repayment behaviors.

Label Encoding: Converted categorical strings in 'Employment_Type' to numerical labels using Label Encoder, simplifying machine learning processes which require numerical input.

5. Visualization and Analysis:

Utilized various plotting techniques to visualize distributions and relationships in the data, including the distribution of the CNS performance scores post-transformation. These visualizations were key in identifying patterns and outliers in the data.

6. Data Summary and Export:

After processing, the cleaned and transformed data was summarized and exported for using in further analysis.

Interactive Graphs we implemented:

Analysing Risk:

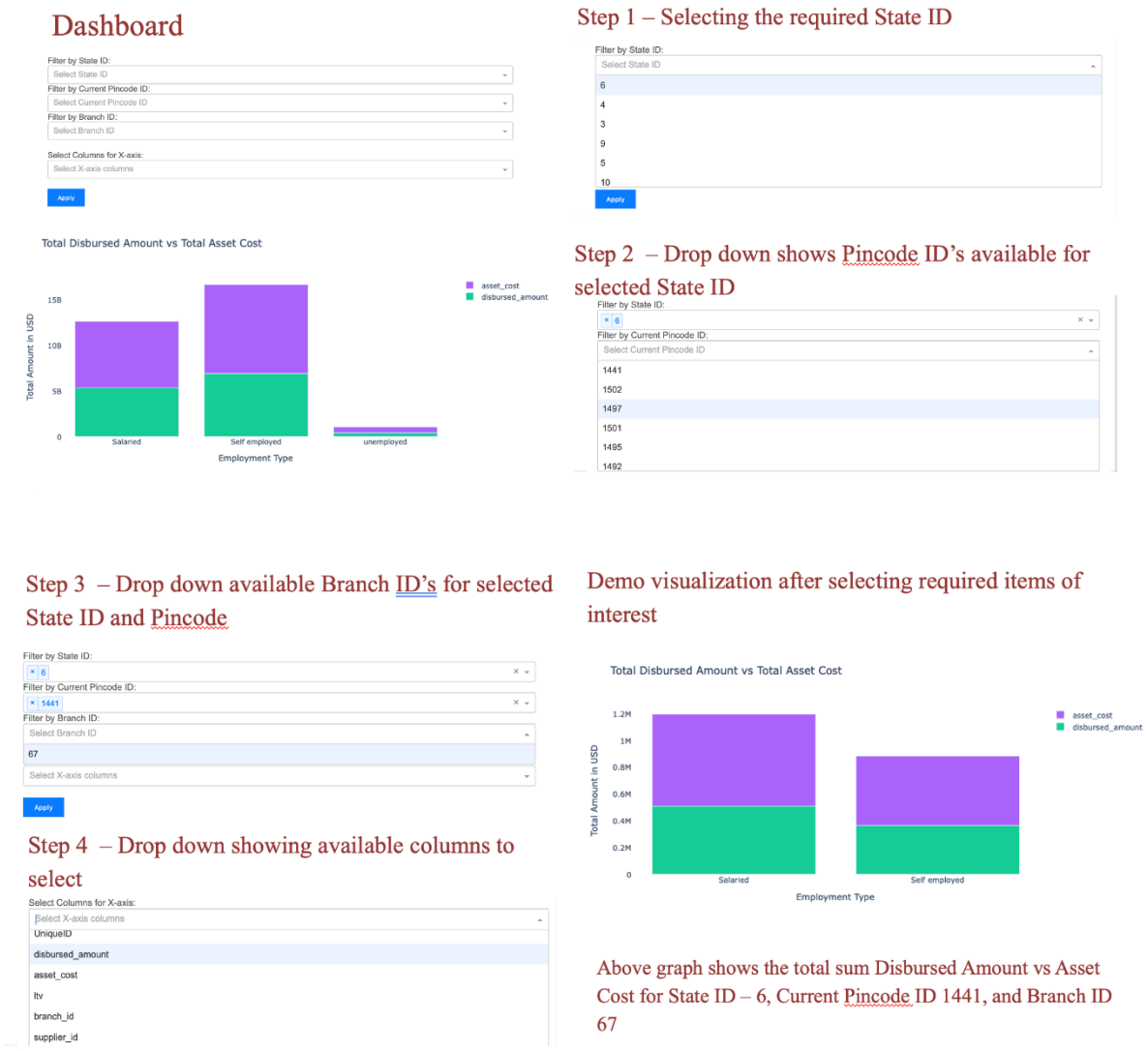


Fig 9: Dashboard with Stacked plot to analyze Disbursed Amount and Asset Cost for each Employment Type

The dashboard we provided offers a comprehensive and interactive way to analyze risk metrics in loan disbursement, unlike the static count plot of loan distribution by state (Fig. 1), which only provides a general overview of loan activity across regions. By incorporating filters for State ID, specific Pincode IDs based on the State ID, and related Branch IDs based on the selected State and Pincode IDs, the dashboard enables users to focus on specific geographic or operational segments. This targeted filtering makes it easier to identify patterns and trends relevant to particular areas or branches.

A key strength of our visualization is the comparison of multiple metrics, such as Total Disbursed Amount and Total Asset Cost, across categories like employment type. Through the use of a stacked graph, the visualization highlights how salaried, self-employed, and unemployed borrowers differ in the loans they access and their associated asset costs. This visual simplicity allows decision-makers to quickly understand loan distributions and borrower behavior, which is critical for assessing risk or optimizing loan offerings. Also, the dashboard includes robust features such as interactive tooltips. When users hover over the plots, it displays precise details, including exact amounts with fractions, providing granular insights into the data.

Moreover, the dashboard's interactivity enhances its utility for real-time data exploration. The dropdown options allow for the selection of multiple columns and customization of the X-axis, enabling Bankers to dynamically adjust the visualizations to fit their specific analytical needs. This flexibility makes the dashboard a powerful tool for various use cases, from pinpointing high-risk regions to evaluating the performance of individual branches. The ability to drill down into the data provides stakeholders and auditors with a deeper understanding of loan performance and associated risks, something that the static Fig. 1 lacks, also the dashboard includes robust features such as interactive tooltips. When users hover over the plots, it displays precise details, including exact amounts with fractions.

Fig 10: Radar Chart of Number of Inquiries for each CNS Score Description based on Selection

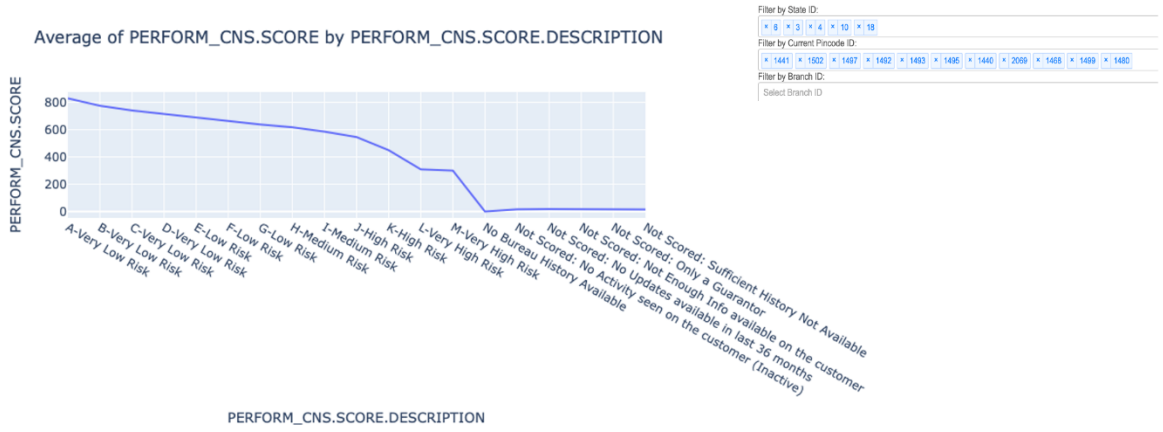


Fig 11: Average CNS score for each CNS score description based on Selection

Interactive Filters: We have included filtering options (by State ID, Current Pincode ID, and Branch ID). This level of interactivity enhances user engagement and allowing dynamic analysis for people to use conveniently.

Line Chart: The line chart provides a clear and effective visualization of the trend in average CNS scores across risk descriptions. Unlike box plots, which focus on the distribution of data, the line chart highlights changes and patterns over ordered categories, making it more suitable for identifying overall trends.

Radar Chart: The Radar chart enables comparative visualization across multiple CNS score descriptions. Its circular layout intuitively shows extremes and anomalies in the number of inquiries for each risk description. This format emphasizes the spread and concentration of data points, making it more engaging and informative for categorical comparisons than a traditional box plot.

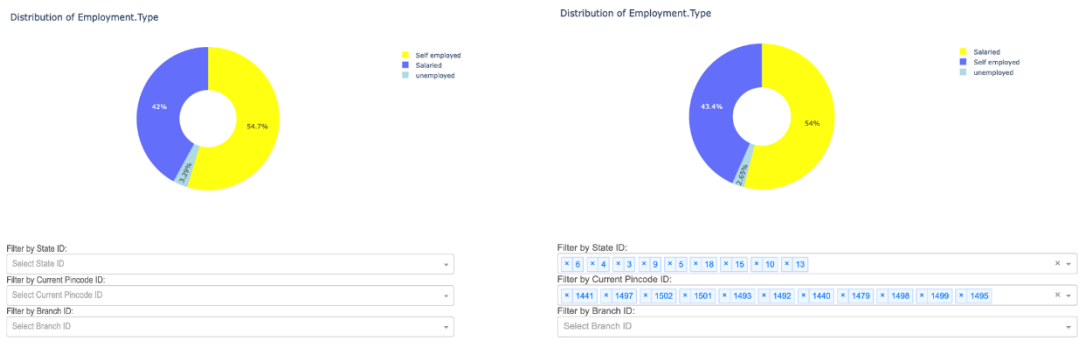


Fig 12: Donut Chart showing Employment Type distribution

The donut graph we provided offers a more interactive and focused visualization compared to traditional pie charts, which often lack clarity and flexibility. Its dynamic filtering options for parameters like State ID, Pin code ID, and Branch ID allow users to analyze employment-type variations (e.g., salaried, self-employed, unemployed) with precision. Clear percentage labels ensure quick readability at a glance.

A standout feature is its interactivity, enabling targeted insights by refining filters for specific regions or units. This clean design highlights trends without clutter, while the central "hole" display summary metrics, adding functionality beyond static pie charts.

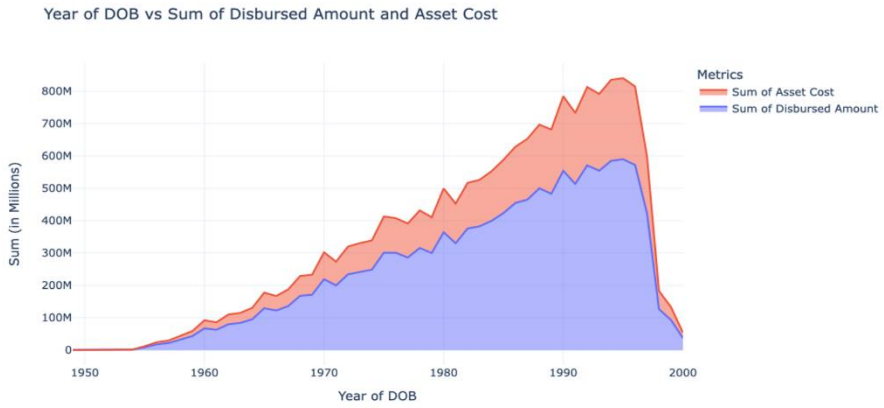


Fig 13: Area Plot of Amount Disbursed and Asset Cost based on Year of Date of Birth

The area plot we designed provides a powerful visualization of the relationship between the disbursed loan amounts and asset costs over the years, offering a clear and continuous view of trends. The graph highlights a proportional increase in the disbursed amount relative to asset costs, reflecting a consistent lending strategy by financial institutions. Notably, individuals born between 1990 and 1998 exhibit higher disbursed amounts and asset costs, suggesting that people between 25 to 35 years of age are purchasing more vehicles by taking loans.

As it effectively conveys the relationship between variables, we feel this visualization is far superior than traditional static plots. Also, the layered format of the area plot makes it easy to compare the total values while identifying overlaps and variations in trends. Additionally, the continuous design ensures clarity when analyzing patterns over a long-time span, avoiding the clutter or segmenting that can occur in traditional plots.

The interactivity of the plot further enhances its utility, enabling users to explore specific data points and trends dynamically. This makes it a valuable tool for financial analysts and decision-makers aiming to understand historical lending patterns, assessing risks, and optimize lending strategies.

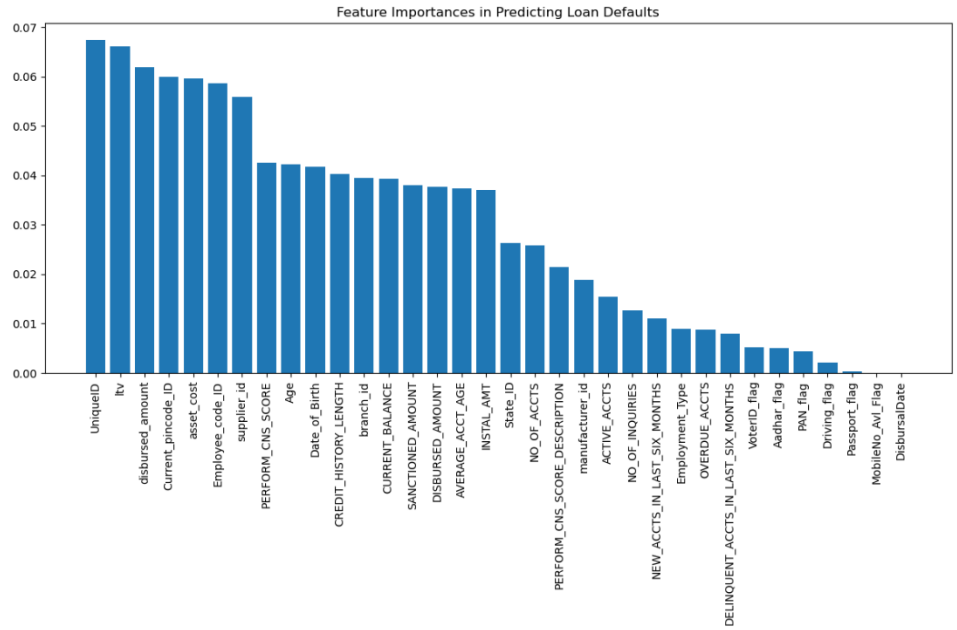


Fig 14: Feature Importances in Predicting Loan defaults

This graph explained the feature importance analysis conducted using a Random Forest Classifier. This analysis helps identify the most significant predictors of loan defaults in the dataset. The dataset was divided into training and testing sets with an 80-20 split, using a random state of 42 to ensure reproducibility. A Random Forest Classifier was initialized with 100 estimators and trained on the training data. The bar chart below illustrates the relative importance of each feature in predicting loan defaults.

UniqueID - This feature, surprisingly, shows the highest importance, suggesting a potential data leakage or overfitting issue that should be further investigated.

ltv (Loan to Value ratio) - As expected, the loan to value ratio is a significant predictor, indicating higher default risk with higher LTV ratios.

disbursed_amount - The amount of loan disbursed is closely related to default probabilities, with higher amounts likely reflecting higher financial risk.

Current_pincode_ID - Pincode plays a critical role, possibly reflecting regional economic conditions affecting the default rates.

asset_cost - The cost of the asset for which the loan is taken influences default likelihood, with more expensive assets possibly entailing higher financial stakes and risks.

Other significant features include the borrower's age, the length of credit history, and the amount sanctioned for the loan. Interestingly, features like employment type and various flags related to the borrower’s profile like 'aadhar_flag', 'pan_flag' show lesser importance in this model. It also underscores the need for careful feature selection and further investigation into features like UniqueID that exhibit unusually high importance. This insight can aid in refining predictive models and enhancing decision making processes in loan approvals.

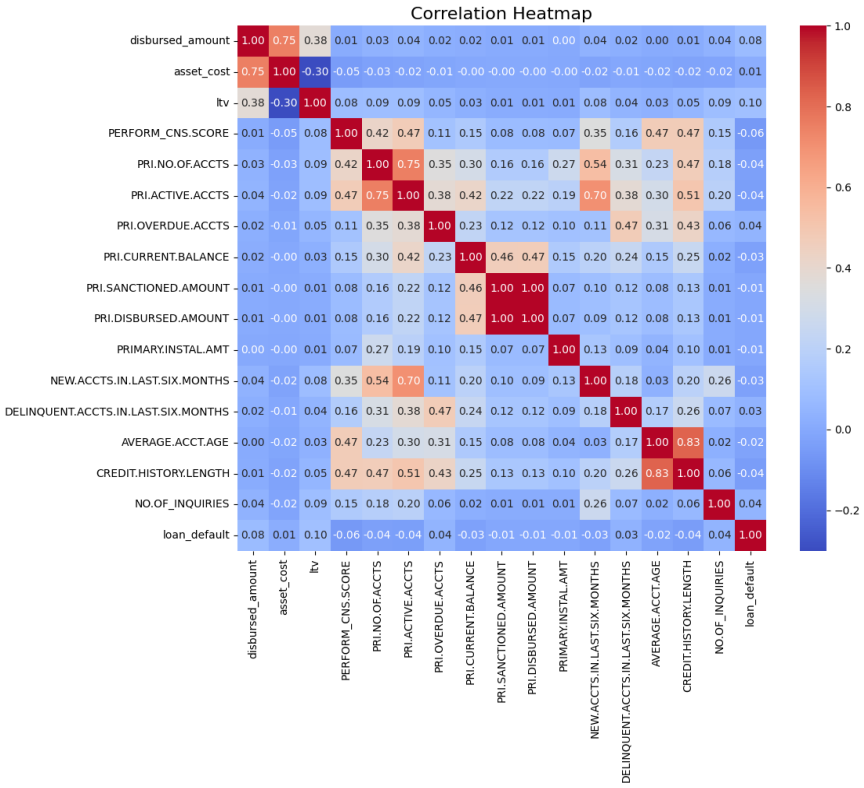


Fig 15: Correlation matrix using Pearsons's coefficient

The correlation heatmap provided illustrates the relationship between various financial and non-financial parameters in the loan dataset. This visualization is a key tool in identifying factors that might contribute to loan default, thereby assisting in enhancing risk assessment models. An analysis is made based on the heatmap observed.

High Correlation between Loan Amounts, the heatmap reveals a strong positive correlation (0.75) between 'disbursed_amount' and 'asset_cost', indicating that as the cost of the asset increases, so does the amount disbursed for the loan. This relationship is intuitive, as more expensive assets typically require larger loan amounts.

Influence of Account History: Variables related to the borrower's credit account history, like 'PRI.NO.OF.ACCTS' (total number of loan accounts) and 'PRI.ACTIVE.ACCTS' (number of active loan accounts), show moderate positive correlations (0.42 and 0.47 respectively) with the credit score ('PERFORM_CNS_SCORE'). This suggests that borrowers with more accounts, particularly active ones, tend to have better credit scores.

Impact on Default Rates: Interestingly, 'loan_default' does not show strong correlation with most of the financial parameters but has a slight negative correlation with 'CREDIT.HISTORY.LENGTH' and 'PERFORM_CNS_SCORE' (both -0.04). This implies that longer credit history and higher credit scores are marginally associated with lower default rates, though the strength of these relationships is weak.

Debt and Delinquency Impacts: The heatmap also shows that 'PRI.OVERDUE.ACCTS' (number of overdue accounts) has a low positive correlation with 'loan_default' (0.03). While this correlation is not strong, it highlights that delinquencies in repayment could have a slight impact on increasing default risks.

From the correlation heatmap, it is clear that while some parameters are strongly interlinked, the direct correlation of individual parameters with loan default is not pronounced. This suggests that loan default is a multifactorial issue, influenced by a combination of various factors rather than any single financial metric.

Fig 16: Combined plot of Bar Graph and Scatter plot showing Average Disbursed Amount and Transaction Count by Age

The dataset has been visualized using a dual-axis graph, one showing the average disbursed amount in blue and the other showing the count of people who took loans, categorized by age in red. This representation allows us to simultaneously evaluate how both the average loan amount and the frequency of loans vary with borrower age.

Peak Borrowing Age: The transaction count shows that the highest frequency of loans occurs around the age of 28, with a gradual increase from early adulthood and a sharp decrease after the mid-40s. This peak could reflect a period in life characterized by increased financial needs, possibly related to family expansion, home purchases, or mid-career investments.

Average Disbursed Amounts: The average disbursed amount remains relatively high and stable from ages 21 to 59, peaking slightly at age 45. This indicates that while younger and middle-aged adults frequently seek loans, the amounts they require do not vary dramatically with age until a significant drop is noted as borrowers approach retirement, post-50 years.

Risk Assessment: The stabilization of loan amounts in the earlier ages suggests that lenders might face relatively consistent risk levels when dealing with borrowers under 50. However, the decreasing number of loans and amounts in older age groups may indicate a shift in lending risk, possibly due to retirement and reduced income.

The data indicates that the most active borrowers are between the ages of 30 and 50. The average loan amount disbursed to individuals does not fluctuate extensively during these years, suggesting a stable borrowing behavior pattern in this age range.

Additionally, the decrease in both the number and amount of loans in older age groups suggests that companies need to adjust their risk assessments and lending strategies to accommodate the changing financial landscapes as individuals age.

This analysis not only helps in understanding borrower behavior across different life stages but also aids financial institutions in refining their approaches to customer engagement and risk management.

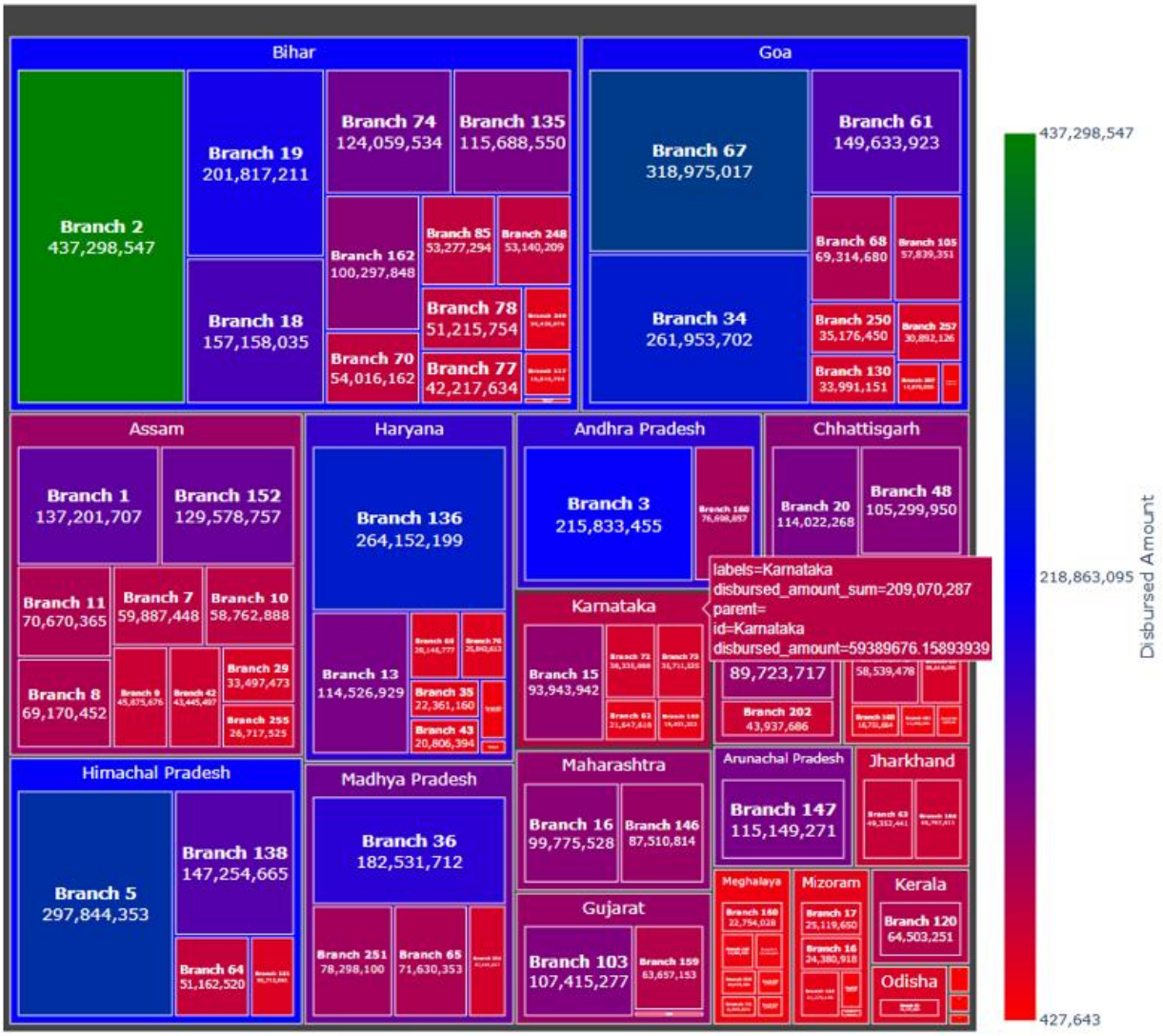


Fig 17: Treemap of the disbursed loan amounts by state and branch across india

In our analysis of disbursed loan amounts across various branches and states, we utilized a treemap to visually break down the distribution and magnitude of disbursements. This method allows us to identify patterns and outliers effectively, offering insights into regional lending behaviors.

The treemap categorizes disbursement amounts by state and individual branches within those states. Each rectangle in the treemap represents a branch, with the size indicating the total disbursed amount and the color signifying the relative magnitude compared to other branches. There is a significant variation in the disbursed amounts across different states. For instance, branches in Bihar display prominently due to their larger disbursed amounts, suggesting a high demand like a larger customer base in this region.

Branch Performance, the visualization highlights branches that are outperforming or underperforming. For example, Branch 2 in Bihar shows a particularly high level of disbursement, which could be due to several factors such as economic activity, the effectiveness of the branch's management, or marketing strategies.

The color gradient from red to green not only enhances visual appeal but also provides quick insights into the scale of disbursement. Red indicates lower disbursements, and green signifies higher amounts, offering an intuitive understanding of financial distribution.

By observing the disbursement patterns, financial institutions can make informed decisions about where to focus their marketing efforts, where to expand their presence. Understanding the geographic distribution of loan disbursements helps in assessing potential risks regionally associated with large disbursements.

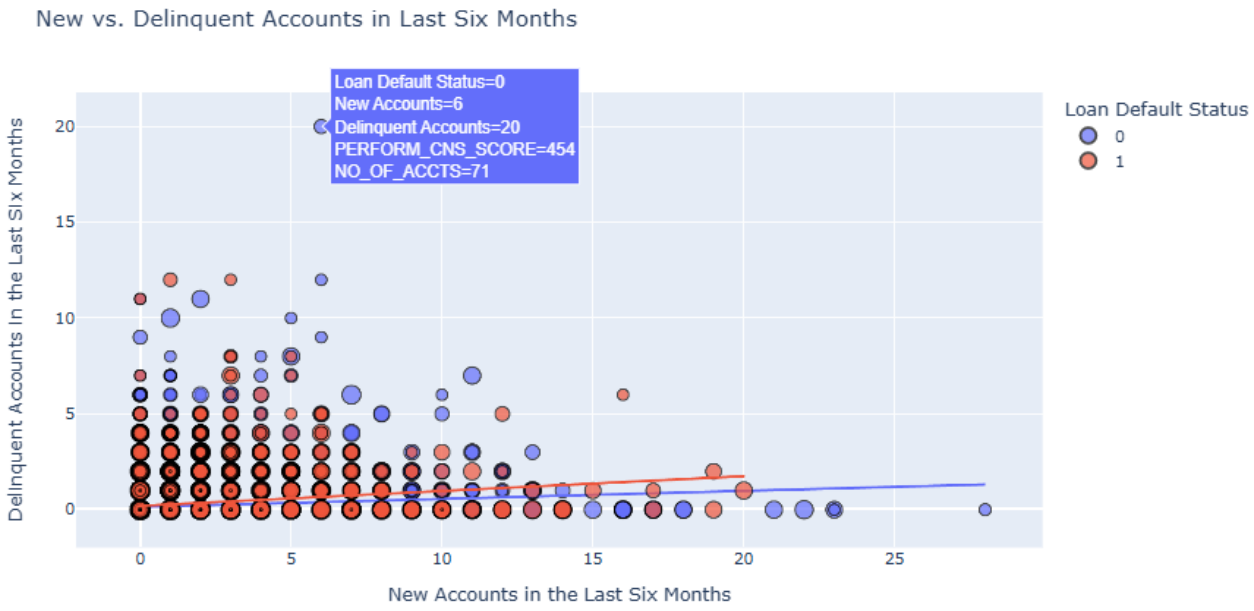


Fig 18: Scatter plot comparing the relationship between new and delinquent accounts

In this analysis, we can examine the relationship between new and delinquent accounts over the last six months and their impact on loan default status. The addition of a trend line using ordinary least squares, OLS regression is crucial in understanding the general relationship between these variables across the dataset. With a positive correlation, As borrowers open more new accounts, there is a tendency for the number of delinquent accounts to increase as well. The data visualized in the scatter plot provides insight into borrower behavior and credit risk, facilitating a deeper understanding of factors contributing to loan defaults.

The scatter plot contrasts the number of new accounts opened by borrowers against the number of their delinquent accounts in the same timeframe, with each point color-coded to indicate whether the loan defaulted (red) or not (blue). Additional variables such as the borrower's credit score ('PERFORM_CNS_SCORE') and total number of accounts ('NO_OF_ACCTS') are incorporated as hoverable data points to provide richer context.

Density of Data Points, The majority of data points cluster at lower numbers of both new and delinquent accounts, indicating that most borrowers neither open many new accounts nor fall into delinquency frequently within a six-month period.

Relationship Between New and Delinquent Accounts, There appears to be a general trend where borrowers with higher numbers of new accounts also tend to have more delinquent accounts. This suggests that individuals who are actively engaging in obtaining new credit may also be those who struggle to maintain current accounts, potentially increasing their risk of default.

Impact of Credit Behavior on Loan Default: Interestingly, the color differentiation shows that a significant number of borrowers with high delinquency do not necessarily default on loans. This could imply that while delinquency is an indicator of financial stress, it is not a definitive predictor of default within the observed period.

Influence of Credit Score: The size of the markers, representing the credit score, indicates that individuals with higher credit scores tend to have fewer delinquent accounts, aligning with conventional credit risk assessments.

Credit Risk Assessment: Financial institutions might consider these patterns when evaluating borrower risk. For instance, a high number of new accounts, especially coupled with delinquencies, could flag potential risks.

Loan Approval Criteria: This analysis could be used to refine criteria for loan approval, potentially incorporating metrics like the ratio of new to delinquent accounts to better see the borrower stability.

The scatter plot analysis highlights the complex interplay between new account openings, account delinquency, and loan defaults. By understanding these relationships, lenders can improve their credit risk models, leading to more informed decision-making and potentially lower default rates. This insight is invaluable for both mitigating risk.

Fig 19: Scatter plot showing the relationship between CNS score and length of credit history

In our analysis of the relationship between CNS, Credit News Score and credit history length, we explore how these factors correlate with perceived risk levels in financial behavior. This study utilizes a scatter plot visualization, which segments data points into risk categories based on CNS scores, illustrating trends and providing deeper insights into creditworthiness. The scatter plot displays CNS scores on the x-axis against the length of credit history in months on the y-axis. Each data point represents an individual, with colors indicating the risk level: red for high risk, orange for moderate risk, green for low risk, and blue for very low risk. The size of each point reflects the number of inquiries, adding another layer of detail regarding credit-seeking behavior.

High risk individuals (red points) tend to have lower CNS scores, often below 500, indicating a correlation between lower scores and higher credit risk.

Moderate risk individuals (orange points) typically cluster in the mid-range CNS scores, between 500 and 600.

Low and very low risk individuals (green and blue points) mostly appear with CNS scores above 600, highlighting that higher scores are associated with lower perceived risk.

Credit History Length, There is a visible trend where individuals with longer credit histories generally have higher CNS scores. This trend supports the notion that a longer history of managing credit responsibly can positively impact CNS scores. The distribution shows that individuals with very low risk often have both high CNS scores and relatively lengthy credit histories.

Impact of Inquiries, The varying sizes of the data points indicate the number of credit inquiries, with larger sizes denoting more inquiries. Interestingly, a higher number of inquiries does not necessarily correlate with a lower CNS score can influence the impact on the CNS score.

This analysis tells the importance of maintaining a healthy credit score and a stable, lengthy credit history in reducing perceived credit risk. It also highlights that while inquiries are an integral part of credit behavior. The insights garnered from this visualization not only enrich our understanding of credit behavior but also facilitate better credit risk management and financial planning strategies.

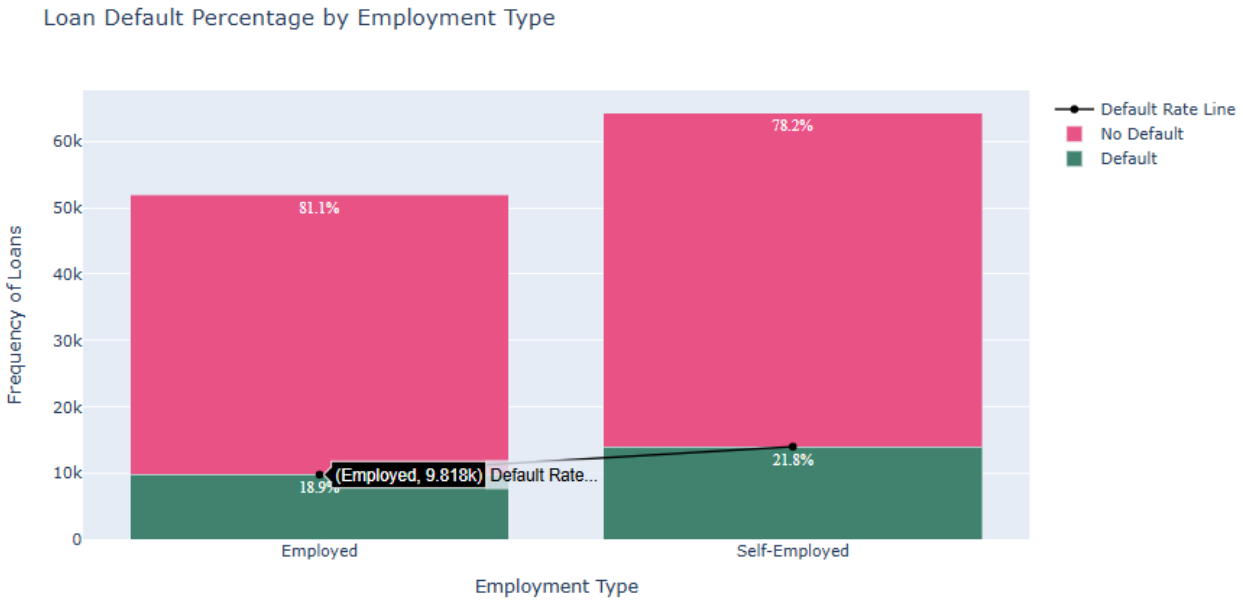


Fig 20: stacked bar chart showing the loan default percentages by employment type

In the analysis depicted in the bar chart, we examine the correlation between employment type and loan default rates. This visualization provides an insightful view into how different employment statuses affect the likelihood of loan defaults.

The chart contrasts two categories of employment: Employed and Self-Employed. Each category is represented by two bars, colored to distinguish between loans that have defaulted (green) and those that have not (pink). Additionally, a black line overlays the bars to indicate the difference in default rate percentage for each employment type.

Default Rates, the "Employed" category shows a default rate of approximately 18.9%, with a significant majority of loans not resulting in default. The "Self-Employed" category exhibits a higher default rate of 21.8%. This suggests that self-employed individuals might face more financial instability, which could contribute to a higher risk of default.

Loan Volume, the frequency of loans is higher among Self- employed individuals compared to employed ones. This could be attributed to the perceived stability of employed individuals by lenders, making them more likely to approve loans for this group.

Despite the higher number of loans, the lower default rate percentage in the Self-employed category highlights a stronger repayment criteria. The analysis clearly illustrates that employment type is a significant factor in loan default rates. Understanding these dynamics helps financial to address the distinct needs of employed and self-employed individuals while lending loan.

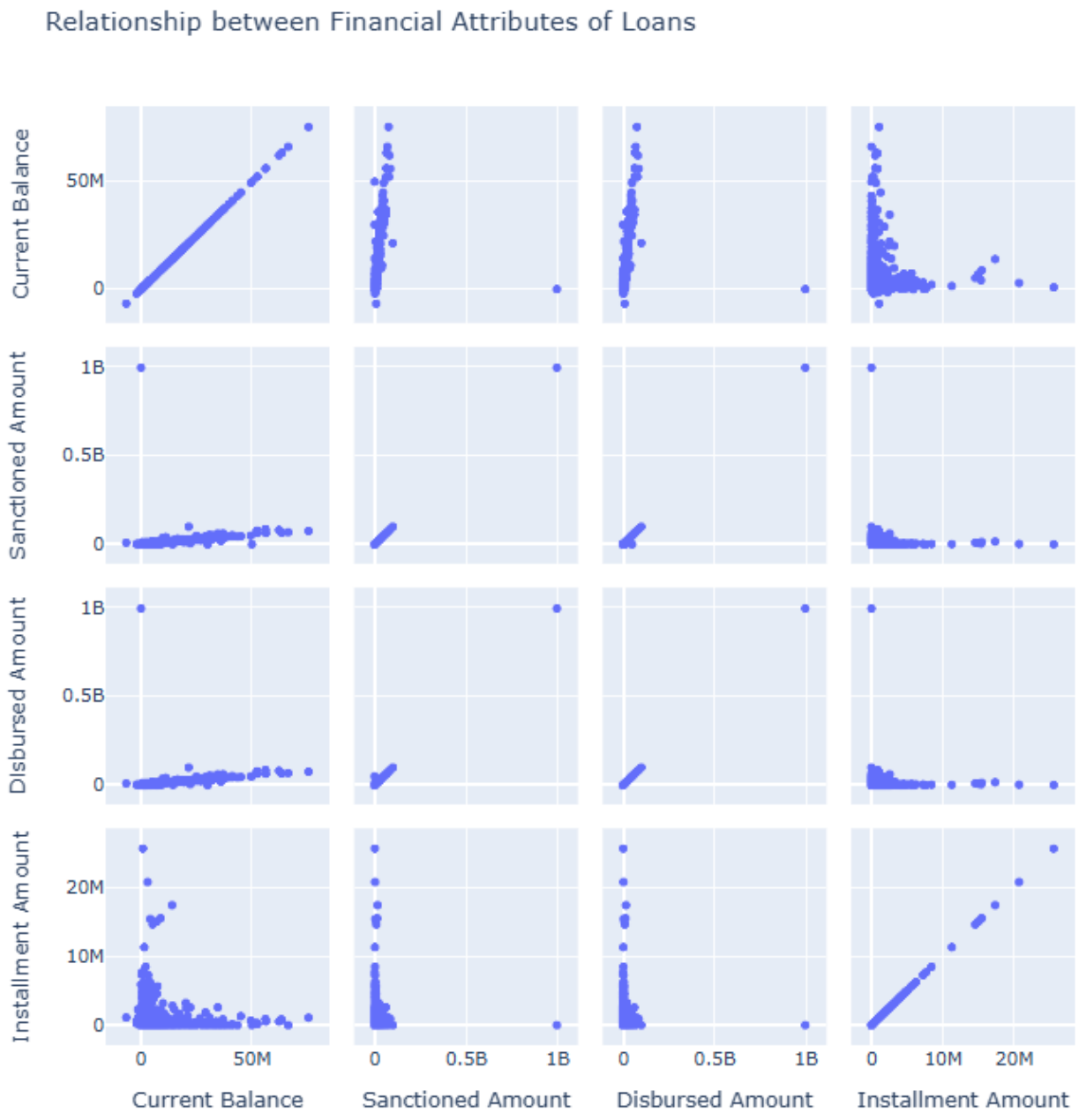


Fig 21: Scatterplot matrix showing relationship between various financial attributes

In this analysis, we explore the relationships between various financial attributes of loans, specifically focusing on current balance, sanctioned amount, disbursed amount, and installment amount. Using a scatter matrix plot, this visual representation allows us to observe the correlation and distribution patterns between these variables. This tool is highly effective for identifying potential relationships and outliers within the data set. The scatter matrix includes several pairwise scatter plots, each showing the relationship between two of the four financial attributes. Each plot provides a visual correlation between the variables, which is essential for understanding how different loan attributes interact with each other.

Current Balance and Sanctioned Amount:

There is a visible correlation between current balance and sanctioned amount, indicating that higher sanctioned amounts generally correspond to higher current balances. This pattern is expected as loans with higher sanctions might not be fully paid off yet, reflecting in larger current balances.

Disbursed Amount and Sanctioned Amount:

The disbursed amount closely follows the sanctioned amount, as shown by a dense line of points along the identity line where the disbursed amount equals the sanctioned amount. This suggests that for most loans, the disbursed amount is typically equal to the sanctioned amount.

Installment Amount and Disbursed Amount:

There is a positive correlation between the installment amount and the disbursed amount. Higher loan disbursements lead to larger installment amounts, which is intuitive since larger loans require larger repayments.

Outliers, several plots show clusters of outliers, such as extremely high current balances with relatively lower disbursed or sanctioned amounts. These could indicate special cases such as restructured loans or extended credit lines.

The scatter matrix plot is a powerful tool for visualizing and analyzing the relationships between different financial attributes of loans. By examining these relationships, both lenders and borrowers can gain valuable insights into loan behaviors, which can guide better. This analysis not only helps in pinpointing unusual patterns but also aids in predicting future trends based on historical data.

Choropleth Map : Geographical distribution of amounts

This visualization uses a choropleth map to display loan-related data across Indian states, showcasing metrics such as loan default count. Each state is represented with a color intensity that corresponds to the value of the selected metric, with darker shades indicating higher values. The map is created using Plotly's interactive tools, where data is layered into three separate maps average disbursed amount, average asset cost, and loan default count. These maps are combined into a single interactive dashboard with buttons allowing users to toggle between the metrics. A Mercator projection ensures accurate geographic representation, and interactivity is enhanced with tooltips that display specific details, such as the loan default count for a hovered state.

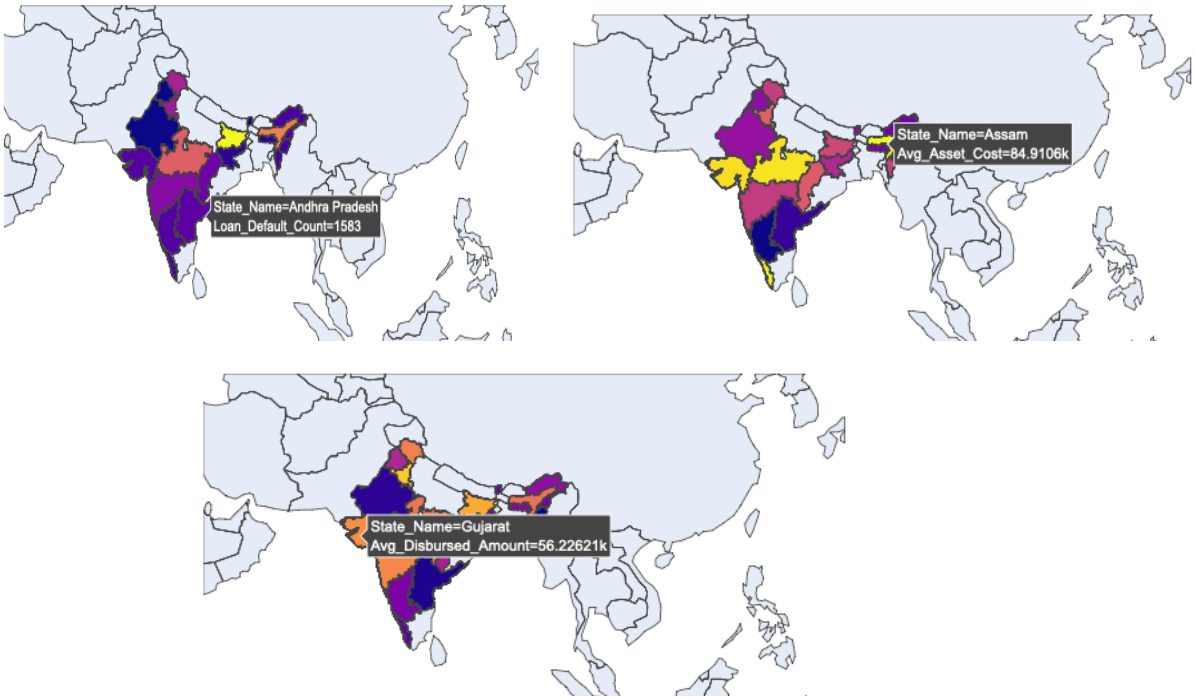


Fig 22: The resulting figure reveals state-wise differences in loan metrics, with color coding visually distinguishing areas of higher or lower values.

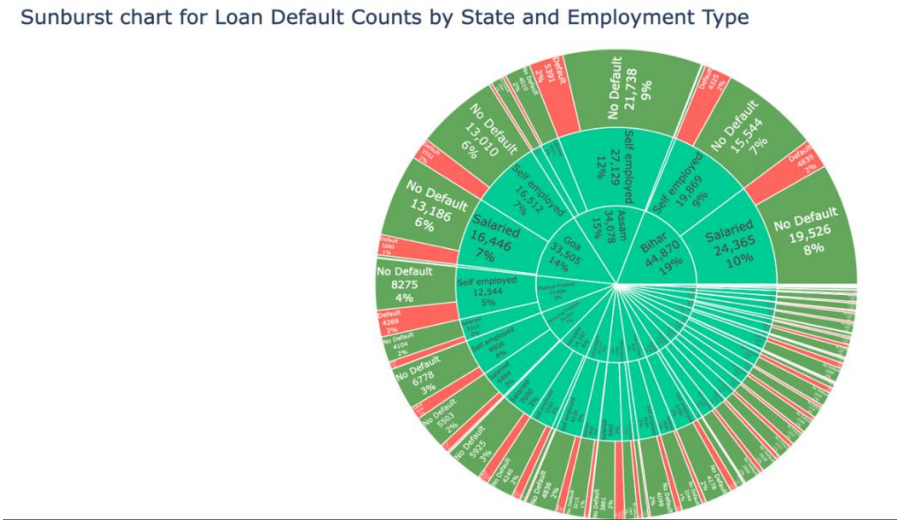
a) For instance, Andhra Pradesh is highlighted with a tooltip showing a loan default count of 1,583, and other states are shaded according to their respective metric values. This visualization makes it easy to spot geographic patterns, such as regions with higher loan default risks, by using contrasting colors. The interactivity of the map provides precise data points and facilitates comparisons between states.

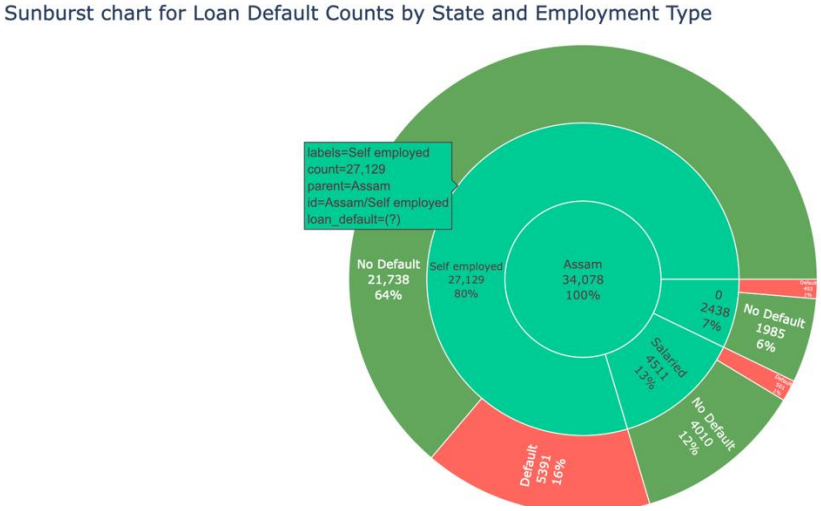
b) The second map depicts the average cost of assets financed through loans in each state, using a blue color scale to represent variations in asset costs. As shown in the tooltip, Assam has an average asset cost of approximately 84.91k.

c) The third map focuses on the average loan amount disbursed per state, represented with an orange color scale. For instance, Gujarat is highlighted with an average disbursed amount of approximately 56.22k.

Sunburst Map: Hierarchical Visualization

We use a visualization model that employs a sunburst chart to illustrate loan default counts segmented by state and employment type. The data is aggregated by grouping on state, employment type, and loan default status, with counts computed for each combination. The loan default status is mapped to distinct colors, with green representing "No Default" and red for "Default." The hierarchical structure is defined through a path comprising state names, employment types, and loan default statuses, enabling the creation of a multi-layered sunburst chart. Tooltips provide interactive details for each segment, showing label names, counts, and the percentage share of entries, ensuring a comprehensive view of the data hierarchy.





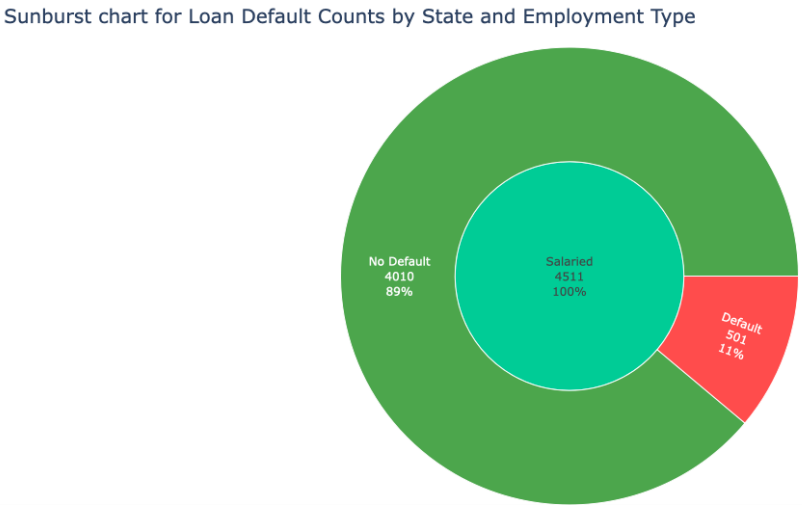


Fig 23: a) This sunburst chart extends the analysis nationwide, presenting loan default counts across all states and employment types. It reveals diverse patterns of loan defaults and employment type distributions across states. States are arranged in a hierarchical way that show larger counts of loan accounts, with a significant proportion marked as "No Default."

b) This chart focuses on a single state, Assam, and its loan default distribution by employment type. The chart highlights that 80% of loan accounts belong to self-employed individuals, followed by salaried individuals (13%) and others (7%). Among these, most accounts exhibit "No Default," with only a small proportion marked as "Default."

c) The third chart narrows the focus to a specific employment type: salaried individuals. The sunburst chart depicts that 89% of loan accounts for salaried employees have "No Default," while only 11% are classified as "Default."

FAILED EXPERIMENTS

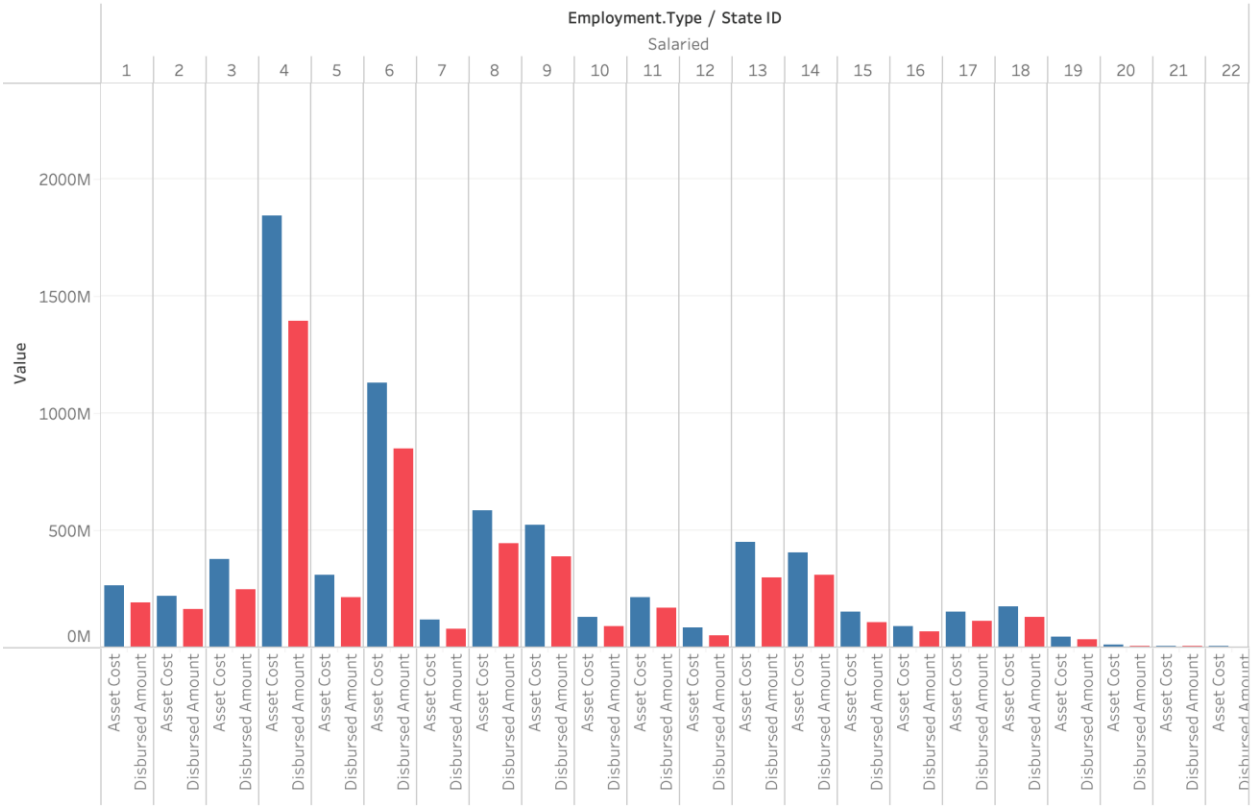


Fig 24: Bar chart displaying total Disbursed amount and Asset cost for each State ID

Initially, we planned to use static graphs, such as bar charts and pie charts, to represent relationships between loan default rates and borrower demographics for risk assessment. However, the high volume of graphs required to capture the complexity of our data led us to adopt more nuanced and interactive visualizations. For instance, when investigating the relationship between disbursed amount, asset cost, and states, the static graphs became cluttered and difficult to interpret. Imagine the added complexity of analyzing subparts of states like pincodes and branches. To address this, we developed an interactive and dynamic dashboard.

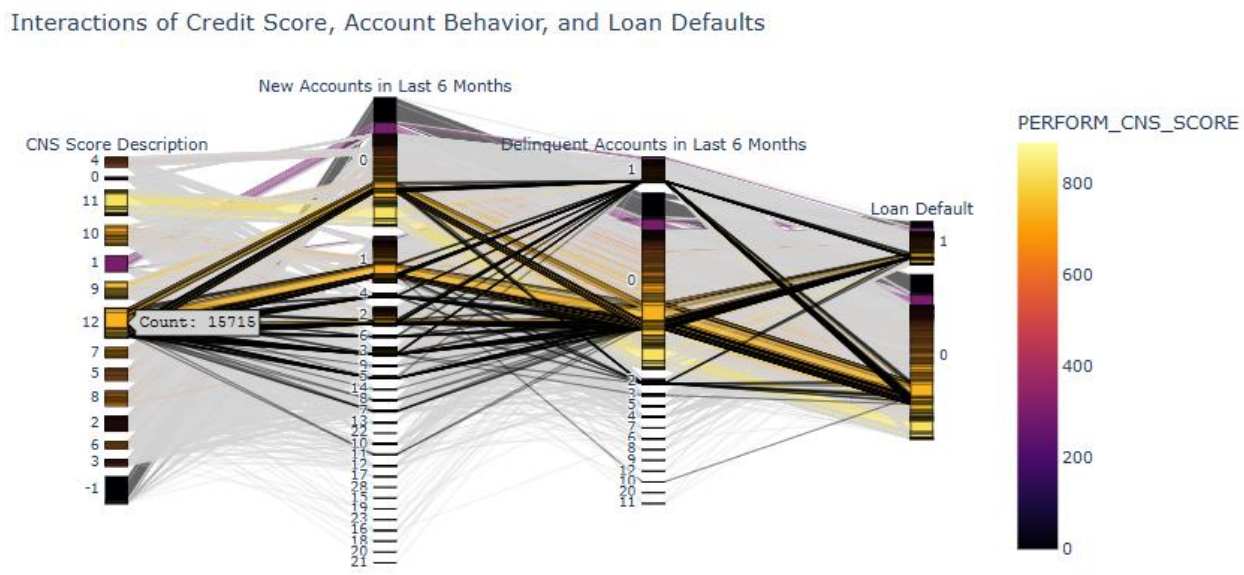


Fig 25: Parallel Chart for comparing interactions of credit score, account behavior, loan defaults

We attempted to analyze the interactions of credit scores, account behaviors, and loan defaults using a parallel categories diagram. The objective was to intuitively represent complex relationships and dependencies among these variables to better understand their influence on loan repayment outcomes. However, the experiment did not yield the clear insights we anticipated. Here are several reasons why the visualization did not meet our expectations:

Overcomplexity, the parallel categories diagram became overly complex with the addition of multiple dimensions. This complexity made it difficult to discern clear patterns or derive actionable insights from the visualization. The intertwining lines, although visually engaging, cluttered the view, leading to confusion rather than clarity.

Color Scale Issues: We used the 'PERFORM_CNS_SCORE' as a color scale to indicate the influence of credit scores across categories. However, the gradient was not distinct enough to effectively differentiate between the various score ranges. This made it challenging to visually assess how different credit scores impacted loan default risks.

Misinterpretation Risk: The visualization required a high level of interpretation skill, which might not be feasible for all stakeholders. The complexity of the diagram could lead to misinterpretations of the data, potentially resulting in misguided decisions based on the visualized information.

This attempt has been a valuable learning experience in understanding the limitations and challenges of using complex multivariate visualizations in data analysis. It underscores the importance of matching the visualization technique to the complexity and type of data to ensure clarity and utility.

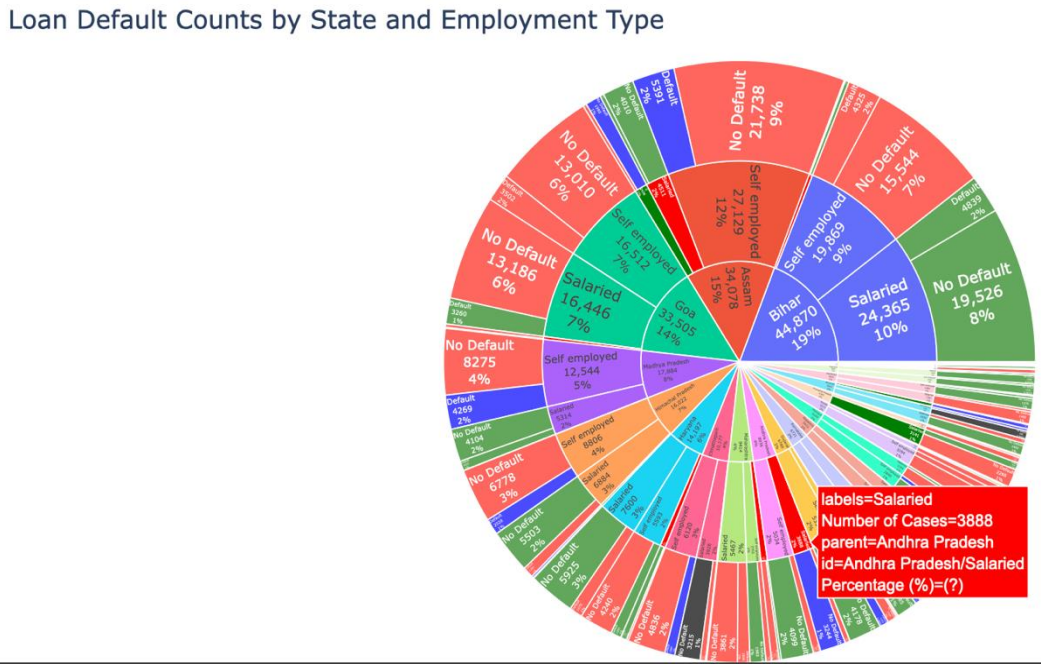


Fig 26: Sunburst failed model with overwhelming colors

The chart uses too many colors, with similar hues representing different states or categories. This makes it difficult to distinguish between segments, especially when multiple hierarchical levels are involved. Combined with the excessive color variation, this creates visual clutter and overwhelms the viewer, making it hard to see the relevance between the states.

CONCLUSION

In our research on vehicle loan default prediction, we sought to address the critical challenge of understanding and visualizing loan default risks through advanced data analytics. Our study aimed to move beyond traditional static visualization methods by developing an interactive dashboard that could provide deeper insights into the complex factors influencing loan defaults.

Through our analysis, we uncovered significant patterns in loan default risks across various dimensions. We found that employment type plays a crucial role in default probabilities, with self-employed individuals showing slightly higher default rates compared to salaried employees. Our age-related analysis revealed interesting borrowing patterns, particularly highlighting that individuals between 25 and 35 years demonstrate the most active loan-taking behaviors, with peak borrowing occurring around age 28.

WHAT HAVE WE LEARNED

The most impactful outcome of our research was the development of an interactive visualization tool. By creating a dashboard with dynamic filtering capabilities, we demonstrated a more sophisticated approach to credit risk assessment. Our visualization techniques allowed for granular exploration of data, enabling financial institutions to gain more nuanced insights into potential default risks across different demographic, geographical, and financial parameters.

FUTURE SCOPE

Visualization techniques represent a particularly exciting area for future research. While our interactive dashboard marked a significant improvement over static visualizations, we believe there is substantial room for developing even more intuitive and comprehensive visual representations of complex financial data. Future researchers could explore more innovative ways of depicting uncertainty and risk in loan default predictions.

We also see great potential in expanding the contextual analysis of loan defaults. By integrating broader economic indicators and more comprehensive socio-economic factors, researchers could develop more nuanced regional risk assessment models. Our study demonstrated the importance of looking beyond traditional financial metrics to understand the full context of borrower characteristics and economic conditions.

We could incorporate machine learning models to predict loan defaults more accurately based on real time data and emerging trends present it in a dashboard with much interactive features.

Geospatial analysis can be visualized more accurately when we can include area-wise codes, incorporating demographic segmentation. Time series visualizations can be used for temporal analysis.

Ultimately, our project underscores the complexity of loan default prediction as a multidimensional challenge. We hope our work contributes to the ongoing efforts to develop more precise and insightful tools for credit risk management. While we have made significant strides, we acknowledge that this is an evolving field that requires continuous research, innovation, and interdisciplinary approaches.

REFERENCES

1. https://www.kaggle.com/datasets/mamtadhaker/lt-vehicle-loan-default-prediction/data
2. https://medium.com/@zh2772/prediction-of-car-loan-default-results-based-on-multi
3. https://ieeexplore.ieee.org/document/9995969
4. https://ieeexplore.ieee.org/document/10056590
5. https://www.atlantis-press.com/proceedings/icifde-23/125993222
6. https://rpubs.com/A_Rodionoff/VehicleLoanDefaultPrediction
7. https://medium.com/@zh2772/prediction-of-car-loan-default-results-based-on-multi-model-fusion-9fbb84e402ab
8. https://ytian22.github.io/Lending-Club-Visualization/
9. https://www.matellio.com/blog/credit-analytics-in-banking/
10. https://www.lendingtree.com/auto/debt-statistics/