**CAPSTONE PROJECT**

**ON**

**Debt Risk Analysis and Loan Default Prediction**

*Submitted in fulfillment of requirement for course under*

*DATA ANALYTICS FOR BUSINESS*

**SUBMITTED BY:**

Nerid Tamrakar (0852680)

Prabesh Ghimire (0858773)

Tasiu Oni (0853659)

Pramesh Babu Khadka (0860484)

UNDER GUIDANCE OF

Manjari Maheshwari

A green and white logo

Description automatically generated

**December 2024**

# **Summary/Abstract**

An essential procedure used by lenders, investors, and financial institutions to determine the probability that borrowers would miss payments on their debts is debt risk analysis. Several financial and non-financial elements that can have an impact on a borrower's capacity to repay loans are assessed as part of the examination. To forecast the likelihood of default, important factors like credit scores, debt-to-income ratios, work history, and economic conditions are looked at. To examine past data and spot trends in borrower behavior, sophisticated methods are employed, such as statistical models, machine learning algorithms, and credit scoring models. The findings of debt risk analysis assist organizations in making well-informed choices about interest rates, risk mitigation techniques, and loan approvals. Debt risk analysis has advanced in sophistication as the financial landscape changes, considering elements like shifts in industry laws and macroeconomic developments. Institutions can optimize portfolios, reduce losses, and guarantee financial stability by comprehending and controlling debt risks.

# **Introduction**

Effective debt risk management is essential to the stability and profitability of the banking industry. As the main providers of credit, Banks are always at danger from borrower defaults, which can cause large financial losses and cause instability in the economy. Because banks offer so many loans, it is critical that they determine each borrower's risk appropriately. Also, a financial institute plays a pivotal role in the economy of the country by facilitating flow of capital, enabling businesses and access of funds to the individuals. Hence, managing its core business, credits, and mortgages, is up most important in micro as well as macro level.

**Objectives:**

1. Find out how exposed the bank is to past due loans that fall into performing and non-performing categories.
2. Using machine learning models to determine the characteristics and factors causing loan default.
3. Find out which branches of banks are more likely to have bad loans so that the bank can concentrate on limiting the maximum amount of loans that can be approved in those branches.
4. Determine the exposure of loans under the retail, corporate and foreign sectors to conduct a segmentation analysis.
5. Make an alert system which reflects the red flag to the probable unhealthy borrowers.

# **Methodology (Data sources and how analysis is conducted and any research you have done so far)**

1. **Data Sources**

In this research, unpublished data from our internal analysis was used to examine the Debt Risk and Default Prediction of loan sanctioned by a bank.

The data is about the Loan Loss Provision to be prepared by bank for fiscal year 2022-23 (from 17th July 2022 to 16th July 2023) with respect to the rules and regulations of Central Bank of country.

1. **Data Preprocessing**

The data preprocessing phase includes critical activities that prepare the dataset for modelling and analysis. Outliers were discovered and resolved, and missing values were filled in or eliminated to ensure data integrity. To maintain consistency and analysis, categorical variables were encoded as numerical values.

1. **Exploratory Data Analysis**

Several visualizations and methods were used to obtain understanding of the dataset during the Exploratory Data Analysis stage. Numerical data were analyzed using bar plots and density plots, and categorical variables were examined using count plots. In order to identify and show outliers, box plots were used. For improved analysis, the datetime datatype was also changed into the proper format. A correlation matrix was created to determine the correlation between numerical variables. Negative values were replaced with zero because they turned out to be invalid or irrelevant for the analysis.

# **Data Analysis (present your main findings derived from the data. Add your visualizations here)**

1. **Findings through correlation matrix**

* Strong correlations (close to 1) are observed between PRINODDAYS and INTOVDDAYS (0.95), and between INTERESTOVERDUE, PENAL\_INT\_ON\_INT, and Total LLP (all above 0.8).
* Variables like INTRATE have weak correlations with most other variables.
* LCY\_OUTSTANDING and LIMIT are highly correlated (0.91), suggesting they might measure similar aspects.

1. **Findings from numerical variables analysis**

* The overall interest rate exposure was in between the range of 10 to 15 % in overall loan outstandings.
* On overall count of loan outstanding, loan outstanding under 1 million has the highest count with 26602 numbers but on overall count of limit of loan sanctioned loan sanctioned amount ranging from 1 million to 5 million has the highest count of 25870.
* Bank has exposure of 51025 number of Principle outstanding of amount ranging below 50K.

1. **Findings from categorical variables analysis**

* State NPST3 has highest loan exposure in numbers.
* More number of loans was sanctioned under Retail Segment.

1. **Findings during Visualization**

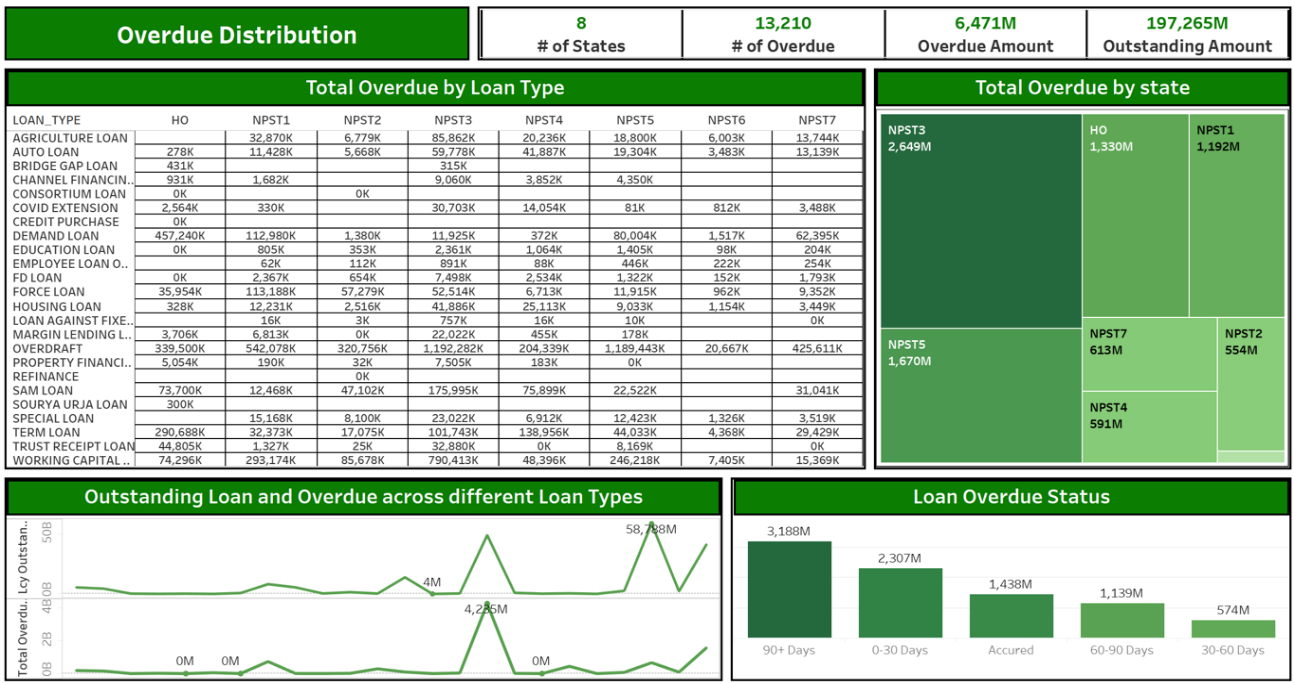
* **Dashboard 1: Portfolio as per Client Distribution**

A screenshot of a computer

Description automatically generated

Most clients are in the Pass category, indicating good loan performance while high volumes of non-performing loans indicate potential risk areas, with some good number of loans on the Watchlist. The bank offers core Overdraft and Term Loans, diversifying from Working Capital Loans to Overdrafts, with potential growth areas in Agriculture, Auto and Education Loans. NPST3 state has been exposed to most Loan outstanding reflecting majority of Bank’s portfolio with balanced client count and outstanding loans in NPST1. NPST2 state is under performing state. Corporate loan dominates Loan Outstanding compared to retail which reflect higher LLP for loan to corporate. The average percent of LLP with respect to Loan outstanding is 3.13% for corporate and 2.41% for retail.

* **Dashboard 2: Overdue Distribution**

****

There are a total of 13,210 overdue cases across 8 states, amounting to 6,471M. The largest overdue amounts are concentrated in NPST3 (2,649M) and NPST5 (1,670M), which represent high-risk zones. The majority of overdue loans, totaling 3,188M, fall into the 90+ days category, indicating significant delays in repayment. Overdrafts and term loans exhibit pronounced spikes in overdue amounts, while housing and employee salary loans show minimal overdue balances. The high-risk zones and loans, particularly in NPST3 and NPST5, require close monitoring and targeted interventions.

* **Dashboard 3: LLP distribution under various categories**

**A screenshot of a computer screen

Description automatically generated**

Overdrafts (28.62%) and Term Loans (27.23%) dominate LLP, while Housing, Auto and Education Loans have minimal contributions. Most LLP is concentrated in top classification (2,349 M) with smaller classifications contributing marginally. SM (1,719 M) and RE (942 M) lead segmentation, while smaller segments like AG contribute less. Corporate clients dominate LLP, with Retail clients and Segment PF plays smaller roles.

* **Dashboard 4: Account Opening and Loan Growth**

**A screenshot of a graph

Description automatically generated**

Increase in both the number of accounts opened and the limit sanctioned from 2015 onwards, indicating strong growth in business. Sanctioned loan limits shows consistent upward trajectory since 2015, reaching a significant high in 2023. Gap between total loan sanctioned and the outstanding loan balance has increased over the years. Increasing difference between loan limit sanctioned and the amount outstanding suggests an opportunity to better disburse the loans

# **Interpretations and Limitations (Discuss your findings and talk about limitations here)**

1. **Interpretations**

**Loan Performance:**

* Most clients fall under the Pass category, indicating good overall loan performance.
* High volumes of non-performing loans and a notable number of loans on the Watchlist highlight potential risk areas.

**Loan Portfolio and Diversification:**

* Core offerings include Overdrafts and Term Loans, with diversification into Working Capital Loans.
* Growth opportunities exist in Agriculture, Auto, and Education Loans.

**State-Wise Insights:**

* NPST3 state has the largest exposure to outstanding loans, representing a significant portion of the bank’s portfolio.
* NPST1 exhibits a balanced distribution of client count and loan amounts.
* NPST2 is an underperforming state.
* NPST3 (2,649M) and NPST5 (1,670M) are high-risk zones with the largest overdue amounts.

**Overdue Cases:**

* Total overdue cases: 13,210 across 8 states, amounting to 6,471M.
* 90+ days overdue loans dominate, contributing 3,188M.
* Overdrafts and Term Loans exhibit the highest spikes in overdue amounts, while Housing and Employee Salary Loans show minimal overdue balances.

**Loan Loss Provisions (LLP):**

* Overdrafts (28.62%) and Term Loans (27.23%) dominate LLP contributions.
* Minimal LLP contributions come from Housing, Auto, and Education Loans.
* Corporate loans dominate Loan Outstanding and LLP (average LLP: 3.13%) compared to retail loans (average LLP: 2.41%).
* Top classifications contribute significantly to LLP (2,349M), with smaller classifications being marginal contributors.
* Segments SM (1,719M) and RE (942M) lead in segmentation, while AG contributes the least.

**Growth Trends:**

* The number of accounts opened, and loan limits sanctioned has steadily increased since 2015.
* Loan limits sanctioned show a consistent upward trajectory, peaking significantly in 2023.
* The growing gap between loan limits sanctioned and outstanding loan balances indicates untapped opportunities for better loan disbursement.

1. **Limitation**

* **Missing and Incomplete Data:** The analysis may be hindered by missing or incomplete data, which can lead to gaps in understanding and inaccuracies in insights drawn from the dataset.
* **Inherent Uncertainty**: The dataset may not account for all variables or factors influencing outcomes, leading to inherent uncertainty in the findings and limiting the precision of predictions.
* **Privacy and Security Concerns**: Handling customer data comes with significant privacy and security concerns, requiring careful measures to ensure compliance with regulations and protection of sensitive information.
* **Lack of Historical Data for New Customers:** The absence of sufficient historical data for new customers limits the ability to assess their creditworthiness or predict potential risks effectively.

# **Recommendations and Future work**

1. **Recommendations on insights**

**Non-Performing Loans (NPL) Management:**

* Strengthen the data recovery team for non-performing loans.
* Periodically review watchlist categories and restructure bad and doubtful loans to improve borrowers' repayment capacity.
* Implement strategies to reduce debts overdue by more than 90 days.
* Address loans overdue by 30 to 60 days to prevent worsening situations.
* Modify terms for high-risk loans to reduce future overdue cases.

**Business Expansion:**

* Increase business exposure by focusing on Working Capital Loans, Overdrafts, and Term Loans.
* Promote low-performing Auto, Education, and Agriculture Loans in rural areas.
* Expand branch networks to improve reach in underperforming states like NPST2.
* Maintain growth in NPST3 by offering competitive products and catering to high-value corporate loans.

**Portfolio Diversification:**

* Diversify the portfolio, currently skewed toward corporate clients, by increasing retail client engagement.
* Conduct regular stress testing for corporate and retail clients to assess loan classifications.
* Leverage strong segments like SM and RE while fostering growth in smaller ones like AG.
* Strengthen high-value classifications and improve performance in underperforming ones.

**Customer Engagement:**

* Roll out targeted marketing campaigns to encourage customers to fully utilize their sanctioned loan limits.
* Promote specific loan products tailored to customer profiles for better loan utilization.
* Provide financial counseling and proactive payment reminders to borrowers.
* Educate customers on available loan products and offer incentives to utilize full sanctioned limits.

**Loan Recovery and Risk Management:**

* Prioritize high-risk zones such as NPST3 and NPST5 for focused interventions.
* Enhance loan recovery strategies to manage the increase in outstanding loans over the years.
* Improve collection techniques in key locations like HO, NPST1, and NPST2.

**Loan Growth and Product Focus:**

* Sustain growth in Overdrafts and Term Loans while increasing focus on Working Capital and Housing Loans.
* Promote corporate dominance while introducing innovative offerings for retail clients.
* Target growth opportunities in Agriculture, Auto, and Education Loans.

1. **Recommendations on Limitations**

* Strict data maintenance by bank
* Consider integrating expert judgment
* Strong data encryption
* Explore alternative data sources

1. **Recommendations on Limitations**

* Models Deployment:
  1. Logistic Regression Model
  2. Random Forest Model
  3. Decision Tree Model
* Web-based alerts on high-risk accounts

# **Appendix (Your main python file which shows your EDA as an Appendix. No need to submit a separate python file)**

Html file for the EDA is attached separately while submission.

# **References**

Stephanou, C., & Mendoza, J. C. (2005). Credit Risk Measurement under Basel II: An Overview and implementation issues for developing countries. In World Bank, *World Bank Policy Research Working Paper* (Vol. 3556) [Report]. <https://documents1.worldbank.org/curated/en/431871468330983467/pdf/wps35560corrected.pdf>

<https://www.nrb.org.np/contents/uploads/2024/01/ECL-Guidelines2024-2.pdf>

Economic Computation and Economic Cybernetics Studies and Research, Issue 2/2019; Vol. 53

<https://ecocyb.ase.ro/nr2019_2/9.%20Coser%20Al.%20Crisan%20Albu%20(T).pdf>

Credit Risk Analysis Using Machine and Deep Learning Models by Peter Martey Addo, Dominique Guegan and Bertrand Hassani

<https://www.mdpi.com/2227-9091/6/2/38>