**Group 10**

**Course**: Data Science and Analytics

**Instructor**: Noble Anumbe, PhD  
**Date of submission**: August 23, 2025

**Group Members:**

|  |  |  |
| --- | --- | --- |
| **Name** | **Osiri email** | **Phone No** |
| Adrian L. Diadula | [adiadula@osiriuniversity.org](mailto:adiadula@osiriuniversity.org) | +639472885413 |
| Ametelah Abubeker Aliyi | [amaliy@osiriuniversity.org](mailto:amaliy@osiriuniversity.org) | +251900296819 |
| Eboigbe Harrison | [eharri@osiriuniversity.org](mailto:eharri@osiriuniversity.org) | +2348167069292 |
| Metasebia Fekadu Wondimu | [mwondimu@osiriuniversity.org](mailto:mwondimu@osiriuniversity.org) | +251911985184 |
| Randy Boateng | [randyboateng@osiriuniversity.org](mailto:randyboateng@osiriuniversity.org) | +233552857293 |

# Final Project Summary: Credit Risk Analysis

# Introduction & Objective

Credit risk analysis helps financial institutions assess borrower default risk and minimize losses. This project uses anonymized loan records to identify high-risk loans, measure portfolio health, and detect breaches of risk limits. Key indicators include Non-Performing Loan (NPL) ratio, sector exposure, loan status, and compliance with assigned limits (CFI Team, 2025). Variables such as Risk Limit, Economic Sector, Loan Status, Interest Rate, Total Balance, and Principal enable calculations of NPL and sector vulnerabilities. The analysis supports better risk monitoring, loan approval strategies, and regulatory compliance, while addressing questions like: Which sectors pose *the highest risk? What share of loans are non-performing? Are risk limits being breached?*

## Dataset Description

The dataset “DSgroup10loan” consists of anonymized loan records from a private bank, with customer identifiers removed for confidentiality. Originally over 8,000 rows and 40 columns, it was refined to 7,577 rows and 13 key variables, including Loan Type, Risk Class, Interest Rate, Balance, Risk Limit, and Economic Sector. Data cleaning involved removing overdraft loans, 336 duplicates, invalid balances/interest rates, and handling missing values through median, mode, or predictive imputation. Date fields were standardized, and potentially erroneous records were excluded to maintain integrity.

## Methodology

The analysis began with Exploratory Data Analysis (EDA) to identify patterns, trends, and anomalies using summary statistics and visualizations (bar charts, pie charts, boxplots, and heatmaps) (Kevin, 2024) for sector exposure, loan status, and interest rates. Portfolio composition was further assessed by aggregating exposures by sector and type, while risk profiling classified loans into Pass, Special Mention, Substandard, Doubtful, and Loss.

For predictive modeling, a Logistic Regression classifier was applied to predict loan risk categories and evaluated using 5-fold cross-validation with classification reports and confusion matrices (Chumahenko, 2025). Python was the primary tool, leveraging pandas, numpy, matplotlib, seaborn, and scikit-learn for data manipulation, visualization, and modeling (Yana, 2024). These methods were chosen to combine descriptive insights with predictive power, ensuring both an understanding of loan portfolio patterns and the ability to anticipate credit risk outcomes.

## Key Findings & Results

* Sector Exposure & Risk Classes: The loan portfolio is heavily concentrated in a few economic sectors, creating vulnerability to sector shocks. Most loans are classified as “Pass,” with smaller shares in Special Mention and Non-Performing Loans.
* Interest Rates & Origination: Most loans carry single-digit interest rates, though some records lack valid rate data. Loan origination shows clear peaks, likely linked to seasonal or campaign-driven lending.
* Model Performance: Logistic Regression achieved ~97.3% mean accuracy under 5-fold cross-validation, showing strong robustness with low overfitting risk (Chumachenko, 2025).
* Dataset & Classification Challenges: The dataset is imbalanced, dominated by “Pass” loans. The model struggles with minority classes like Doubtful and Special Mention, and over-predicts “Bad Loss” due to high recall but low precision. Improvements require additional data, sampling techniques, or rebalancing.

## Insights & Recommendations

The results show that loans are concentrated in a few sectors, which makes the portfolio more vulnerable to sector-specific downturns. Most loans are performing well, but signs of migration toward higher-risk categories highlight the need for early monitoring. Interest rate gaps also point to data quality issues that may affect pricing decisions.

These findings imply that without proper controls, the bank may face rising credit risk and pricing inefficiencies. To address this, practical steps include setting sector exposure limits, conducting stress tests, improving data quality, adjusting pricing based on risk, and building dashboards for regular monitoring. Early warning systems and targeted remediation strategies should also be deployed to prevent performing loans from turning into non-performing ones.

## Limitations

The analysis was constrained by missing key borrower details such as income, history, and collateral value. Data quality issues, including outliers, duplicates, and missing or invalid interest rates, limited accuracy. Assumptions made during date conversions and reliance on provided sector/risk labels may have introduced minor errors. The dataset imbalance also restricted model performance, especially for minority loan classes.

**Video Link:**

**https://drive.google.com/file/d/1HnV3smsTdsg9LSgSSZID7RAoJBiYSjVk/view?usp=sharing**

## References & Acknowledgement

**References**

* Chen L. S.(2024), Introduction to the scientific method in research: A Guide for Student Scientists, University of Malasia Saba, Ch.4
* Chumachenko A. (2025), Methods of classification and clustering in Data Analysis”, International Journal of Scientific Engineering and Science, ISSN(online):2456-7361
* Hunter, J. D. (2007), Matplotlib: A 2D graphics environment, Computing in Science & Engineering, 9(3), 90–95.
* Kevin H. (2024), Step-by-step Explarotary Data Analysis (EDA) using python, Mediu.com
* McKinney, W. (2017), Python for data analysis: Data wrangling with pandas, NumPy, and IPython. 2nd Edition. O’Reilly Media.
* Pedregosa, F., Varoquaux, G., Gramfort, A., et al. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12, 2825–2830.

Works Cited:

* Yana Khare, (2024), <https://www.analyticsvidhya.com/blog/2024/11/python-libraries-for-data-analysis>, Accessed on August 14, 2025
* Jesus Paz, (2024),<https://blog.projecthelena.com/blog/top-10-python-tools-for-data-analysis>, Accessed on August 10, 2025
* CFI Team, (2025), <https://corporatefinanceinstitute.com/resources/commercial-lending/credit-risk-analysis-models/>, Accessed on July 25, 2025

### Acknowledgements

We sincerely thank our instructor, Dr. Noble Anumbe, PhD, for his guidance, and our group members—Adrian L. Diadula, Ametelah Abubeker Aliyi, Eboigbe Harrison, Metasebia Fekadu Wondimu, and Randy Boateng—for their contributions and collaboration throughout this project.