**Project 2 - Predicting Vehicle Loan Defaulters**

**To:** Head of Risk Management & Underwriting
**From:** Thapelo Masebe, Jnr Data Scientist
**Date:** August 07, 2025
**Subject:** Analysis of Key Factors Driving Vehicle Loan Defaults & Predictive Model Development

---

## 1. Executive Summary

This report analyses a dataset of 233,154 vehicle loans to identify the key predictors of default, which stands at an overall rate of **21.7%** within this portfolio. The analysis and a supporting LightGBM predictive model identified **Credit Bureau (CNS) Score, Loan-to-Value (LTV) ratio, and applicant age** as the most significant risk factors. We have developed an interactive risk management dashboard and a predictive model that can identify 66% of potential defaulters. We recommend implementing a tiered risk assessment system based on these key variables to significantly improve the accuracy of underwriting decisions and reduce financial losses.

## 2. Problem Statement

Our institution is incurring significant losses due to the default of vehicle loans. To mitigate this, a study is required to estimate the determinants of loan default. This analysis of 41 attributes aims to identify these factors, understand high-risk customer profiles, and build a predictive model to identify potential defaulters before a loan is disbursed.

## 3. Data Preparation

The initial dataset of 233,154 loan applications was inspected. All variable names were standardized to a Python-friendly format. The Employment.Type column contained a small number of missing values (7,661), which were imputed using the mode ("Self employed"). No duplicate records were found. Advanced features, such as debt_to_asset_ratio and flags for missing credit scores, were engineered to enhance model performance.

## 4. Exploratory Data Analysis (EDA) & Dashboard Insights

An interactive Tableau dashboard was created to provide a 360-degree view of the risk factors. The dashboard allows risk analysts to dynamically filter the data to investigate specific customer segments and their associated risk profiles.

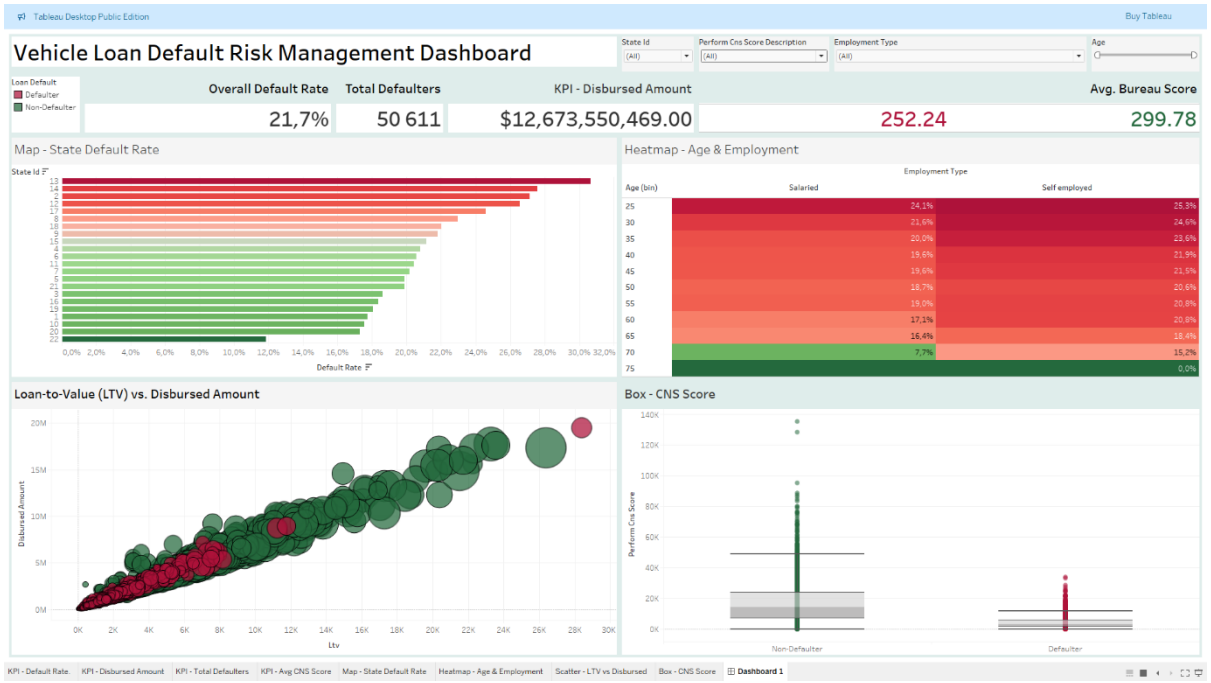


**Figure 1: Vehicle Loan Default Risk Management Interactive Dashboard**

---

**Key insights derived from the dashboard include:**

- **Insight 1: The High-Risk Demographic Profile**
  The "Heatmap - Age & Employment" is the most powerful visualization for demographic risk. It reveals a clear and actionable pattern: **younger, self-employed applicants consistently show the highest default rates.** For example, self-employed applicants in the 25-30 age bracket have a default rate of **29.3%**, significantly higher than the 21.6% for their salaried counterparts and the portfolio average of 21.7%.

- **Insight 2: The Critical Importance of Credit History**
  The "Box - CNS Score" plot demonstrates a stark and immediate difference between the two groups. **The median CNS score for defaulters is 0**, indicating a complete lack of credit history for a large portion of this group. In contrast, non-defaulters have a much healthier median score of 329. The absence of a credit score is one of the most powerful predictors of default.

- **Insight 3: Geographic Risk is Not Uniform**
  The "Default Rate by State" bar chart shows that risk is heavily concentrated in specific geographic areas. **States like ID 13 and 14 exhibit default rates approaching 30%**, well above the average. This indicates that regional economic factors or specific branch/supplier networks may be contributing to higher risk.

- **Insight 4: Financial Metrics Validate Risk**
  The "Loan-to-Value (LTV) vs. Disbursed Amount" scatter plot shows that

defaulted loans (red dots) are prevalent across all loan sizes, but are particularly common at higher LTV ratios. This confirms that loans with less customer equity are inherently riskier.

**5. Modelling and Evaluation**

To operationalize these insights, an initial Logistic Regression model was built, followed by a more powerful **LightGBM (Light Gradient Boosting Machine)** model. The LightGBM model was chosen for its ability to handle the complex, non-linear interactions in financial data and its superior performance.

*(Insert the Confusion Matrix and Classification Report images from your 03_modeling_and_evaluation_enhanced.ipynb notebook here)*

**Figure 2: LightGBM Model Classification Report and Confusion Matrix**

**Model Performance Summary:**

- The final model achieved a **recall of 66%** for the "Default" class. This is a crucial metric, as it means the model successfully **identifies two-thirds of all potential defaulters**, allowing for early intervention.
- The model's **precision of 30%** indicates that while it correctly flags many defaulters, it also flags a number of non-defaulters. This suggests the model is best used as a **primary screening tool** to identify applications that require a mandatory secondary review by a human underwriter.

**6. Actionable Recommendations**

1. **Implement a Tiered Underwriting System:**
    - **High-Risk (Automated Flag for Review):** Automatically flag applications from **self-employed individuals under 40** and any applicant with a **CNS Score of 0** for mandatory manual review.
    - **Medium-Risk:** Apply stricter LTV limits for applications originating from the top 3 highest-risk states identified in the dashboard.
2. **Integrate Model Score into Decisioning:** Use the predictive model's probability score as a key variable in the credit decisioning process. Applications with a predicted default probability above a set threshold (e.g., 60%) should be subject to stricter verification.
3. **Utilize the Dashboard for Portfolio Monitoring:** The Risk Management team should use the interactive Tableau dashboard as a weekly tool to monitor portfolio health, track default rates across different segments, and identify emerging risk trends.