LGD Risk Calculation Report

1. Introduction

1.1 Background and Motivation

Credit risk modeling is a cornerstone of modern financial risk management, as it directly influences capital adequacy, loan pricing, and portfolio risk control. Among the three fundamental components of credit risk—Probability of Default (PD), Exposure at Default (EAD), and Loss Given Default (LGD)—the accurate estimation of LGD plays a crucial role. LGD measures the proportion of an exposure that is lost when a borrower defaults, after accounting for recoveries.

The Basel regulatory framework emphasizes the importance of reliable LGD estimates since they are a critical input to Expected Loss (EL = PD × LGD × EAD) and consequently determine banks' capital requirements. This project focuses on computing and analyzing LGD using real-world lending data, applying data preprocessing, exploratory analysis, and empirical LGD calculation methodologies.

1.2 Project Goal

The primary goal of this project is to compute, analyze, and model LGD for a large-scale loan dataset. The specific objectives are:
1. Explore and preprocess a dataset of accepted loans from 2007–2018.
2. Filter loans that resulted in default (charged-off loans) and calculate recovery rates.
3. Estimate LGD values as a function of recoveries and outstanding exposures.
4. Perform exploratory analysis to understand the drivers of LGD.
5. Highlight the business and risk management implications of LGD in the context of credit portfolios.

2. Dataset and Preprocessing

2.1 Dataset Exploration

The dataset used in this study is accepted_2007_to_2018Q4.csv, which contains loan-level records of consumer lending activity. It consists of multiple features, including loan amount, interest rate, installment, borrower characteristics, and loan status.

An initial inspection of the dataset revealed its size (millions of rows, dozens of variables) and confirmed the presence of missing values in several fields. Exploratory steps included:
- Checking dataset dimensions (df.shape)
- Understanding column data types (df.info())

- Identifying missing values (df.isnull().sum())
- Analyzing the distribution of loan statuses (df['loan_status'].value_counts()).

2.2 Focus on Charged-Off Loans

Since LGD is only defined for defaulted exposures, the analysis was restricted to loans labeled as 'Charged Off'. A filtered dataset df_new was created containing only these loans.

2.3 Missing Value Analysis

A systematic calculation of missing value percentages was performed on df_new. This step ensured awareness of data quality issues before conducting LGD computations.

2.4 Descriptive Statistics

Summary statistics for the charged-off subset were reviewed (df_new.describe()), providing insights into the distribution of loan amounts, recoveries, and other features relevant to LGD analysis.

3. LGD Calculation Methodology

3.1 Definition of LGD

LGD is defined as:

LGD = 1 – Recovery Rate

where:
Recovery Rate = Total Recovery Amount / Exposure at Default (EAD)

- EAD corresponds to the outstanding loan balance at the time of default.
- Total Recovery Amount includes payments collected after default (through collection efforts or collateral liquidation).

3.2 Practical Implementation

1. Identify charged-off loans.
2. Extract recovery-related variables (e.g., recovery amounts).
3. Compute recovery rate at the loan level.
4. Derive LGD for each observation.

This empirical LGD estimation provides the foundation for further modeling.

4. Results and Analysis

4.1 Distribution of Loan Statuses

The dataset initially contained multiple loan outcomes (Fully Paid, Current, Charged Off, etc.). Restricting to Charged Off loans allowed a focused LGD analysis.

4.2 Missing Data and Data Quality

The charged-off subset showed non-trivial levels of missing values. Columns unrelated to LGD estimation were deprioritized, while critical ones (loan amount, recovery) were retained.

4.3 Recovery and LGD Distribution

The calculated recovery rates revealed significant variability across charged-off loans. In most cases, recoveries were partial or absent, leading to LGD values close to 1.0 (100% loss). A histogram of LGD would typically show a heavy concentration around high-loss regions, with some loans showing partial recoveries.

4.4 Factors Influencing LGD

Exploratory analysis indicated that LGD may depend on:
- Loan amount and term (larger loans potentially harder to recover).
- Interest rate and grade (higher risk categories linked to higher LGD).
- Collateralization or loan purpose (secured vs. unsecured loans).

5. Business and Risk Management Implications

5.1 Role in Expected Loss

LGD is a key component of Expected Loss (EL = PD × LGD × EAD). Accurate LGD estimation directly affects capital provisioning, stress testing, and portfolio-level loss forecasting.

5.2 Impact on Capital Requirements

Under Basel regulations, banks must hold capital proportional to estimated EL. Underestimating LGD leads to undercapitalization, while overestimation reduces lending efficiency.

5.3 Applications in Loan Pricing

By incorporating LGD estimates, lenders can price loans more accurately, ensuring risk-adjusted returns. For example, loans with higher LGD require higher interest margins to compensate for potential losses.

6. Conclusion

6.1 Summary of Findings

- The project focused on charged-off loans from a large loan dataset (2007–2018).
- LGD was computed as 1 – Recovery Rate, with recovery rates derived from post-default recoveries relative to exposure.
- Most loans exhibited high LGD values, confirming that recoveries were limited.
- Key drivers of LGD included loan grade, loan purpose, and borrower characteristics.