VIETNAM NATIONAL UNIVERSITY
HO CHI MINH CITY

UNIVERSITY OF SCIENCE

FACULTY OF MATHEMATICS AND COMPUTER SCIENCE

6th August 2025

Thesis Report

Data Science Major

Research topic: Machine Learning system for Credit Risk

Author:
Nguyen Minh Duy - 21280010

Supervisor:
Dr. Tran Anh Tuan

VIETNAM NATIONAL UNIVERSITY, HO CHI MINH CITY
UNIVERSITY OF SCIENCE

NGUYEN MINH DUY

MACHINE LEARNING SYSTEM FOR CREDIT RISK

BACHELOR'S THESIS IN DATA SCIENCE

SUPERVISOR
DR. TRAN ANH TUAN

Ho Chi Minh City - 2025

Acknowledgment

First and foremost, I would like to express my sincere gratitude to **The University of Science, Viet Nam National University Ho Chi Minh City** and **The Dean's Office of the Faculty of Mathematics and Computer Science** for providing me with favorable conditions to complete my academic program in general, and this graduation thesis in particular. The foundation of knowledge accumulated over four years of study, along with the experience gained during the thesis execution, has laid an excellent groundwork for my future scientific research endeavors.

Secondly, I extend my deepest and most heartfelt thanks to **Dr. Tran Anh Tuan**. He not only inspired me to pursue the field of Data Science but also enthusiastically guided me and provided the necessary knowledge and scientific materials to complete this thesis.

I am immensely grateful for the dedicated teaching and support from all the lecturers at the Faculty of Mathematics and Computer Science. Besides my supervisor, the lecturers have not only imparted knowledge and skills but also instilled in me a sense of responsibility and professional work ethic. On this occasion, I would like to express our third word of gratitude to all the esteemed lecturers, wishing them continued health and success in their teaching careers.

Finally, I wish to express my gratitude to my family, my friends and myself for always being by my side, supporting, and never stop believing throughout this journey.

Sincerely.

Ho Chi Minh City, 6th August 2025

Nguyen Minh Duy

Abstract

In recent years, credit risk modeling has increasingly shifted toward complex machine learning approaches focused narrowly on default prediction. However, this thesis argues for a more holistic and interpretable framework, demonstrating that traditional models when thoughtfully designed can deliver substantial insight and operational value in peer-to-peer (P2P) lending contexts.

The study develops a three-component credit risk framework encompassing Probability of Default (PD), Loss Given Default (LGD), and Exposure at Default (EAD). Each component is modeled independently using transparent linear techniques: logistic regression for PD, a two-stage approach for LGD, and linear regression for EAD. These models are then integrated into an Expected Loss formulation to quantify overall borrower risk.

A credit scorecard is derived from the PD model, scaling log-odds into an interpretable 300–850 credit score range. A rule-based credit policy is then implemented to automate lending decisions: low-risk applicants are approved, high-risk ones denied, and intermediate-risk cases evaluated based on estimated Return on Investment (ROI).

Test set results show meaningful improvements: default rate reduced from 6.71% to 5.65%, and expected loss lowered from 6.91% to 5.77%, while rejecting only 11% of applicants. Monitoring on a fully unseen out-of-time set reveals moderate population drift, supporting the need for long-term model maintenance.

The findings suggest that even with simple, interpretable models, it is possible to create a structured, risk-aware lending pipeline that aligns well with business goals and regulatory expectations in order to offering an effective alternative to black-box predictive systems.

Contents

Glossary **10**

1 Introduction **12**
1.1 Background and Context 12
1.2 Problem Statement 13
1.3 Objectives and Research Questions 14
1.4 Scope and Limitations 15

2 Literature Review **17**
2.1 Terminology Clarification: Risk vs. Scoring vs. Scorecard 17
2.2 Overview of Traditional Credit Scoring 18
2.3 Machine Learning Approaches in Credit Risk 18
2.3.1 Techniques in Use 18
2.3.2 Applications by Major Institutions 19
2.3.3 Opportunities and Challenges of Advanced Models 20
2.3.4 Regulatory Frameworks 20
2.3.5 Gaps and Motivation 21

3 Theoretical Framework **22**
3.1 CRISP-DM: Aligning Business Goals with Data Science 22
3.2 Regulatory Foundations: Basel III IRB & IFRS 9 ECL 23
3.3 Credit Risk and Expected Loss Decomposition 23
3.4 Weight of Evidence for Feature Transformation 23
3.5 Credit Scorecard Scaling and Industry Standards 24
3.6 Two-Stage LGD Modeling 24
3.7 Model Monitoring with Population Stability Index (PSI) 25
3.8 Credit Decisioning Policy Framework 25

4 Methodology **26**

4.1 Research Design 26
4.2 Data Collection and Description 27
4.2.1 Initial Data Cleaning and Feature Reduction 28
4.2.2 Exploratory Data Analysis (EDA) 29
4.3 Feature Engineering 33
4.4 Modeling Approach 35
4.4.1 Probability of Default (PD) Model 35
4.4.2 Loss Given Default (LGD) Model 37
4.4.3 Exposure At Default (EAD) Model 39
4.4.4 Expected Loss Integration 40
4.5 Evaluation 40
4.5.1 Validation Strategy 41
4.5.2 Model-Specific Evaluation Metrics 41
4.5.3 Business-Focused Evaluation 44
4.5.4 Calibration and Residuals 45
4.5.5 Limitations 46

5 Credit Scorecard Development **47**
5.1 Model Foundation and Score Construction 47
5.2 Reference Categories and Baseline Scoring 47
5.3 Intercept Calibration and Scaling Framework 48
5.4 Score Range and Component Scaling 48
5.5 Risk Classification Framework 49

6 Credit Policy Application **51**
6.1 Risk Classification Framework 51
6.2 Lending Decision Rules 52
6.3 Integration of Expected Loss and ROI 53
6.4 Policy Evaluation and Impact 54
6.5 Summary 54

7 Model Monitoring **56**
7.1 Overview and Purpose 56
7.2 Population Stability Index (PSI) Methodology 56
7.2.1 Definition and Formula 56
7.2.2 PSI Interpretation Guidelines 57
7.3 PSI Computation Process 57

7.4 PSI Results and Analysis 57
7.5 Implications for Model Maintenance 58
7.6 Conclusion 59

8 Discussion and Interpretation **60**
8.1 Summary of Findings 60
8.2 Answers to Research Questions 61
8.3 Practical Implications 62
8.3.1 Policy Scenario Analysis 62
8.4 Assumptions 63
8.5 Limitations 65
8.6 Future Work 66

9 Conclusion **68**

Preface

In recent years, machine learning has rapidly transformed credit risk assessment, with increasing reliance on complex models such as ensemble methods and deep learning. While these approaches often emphasize predictive accuracy, they tend to compromise interpretability known as an essential requirement in regulated lending environments.

This thesis is driven by a belief that **interpretable, well-structured models can still offer valuable insights and business utility without resorting to algorithmic complexity**. Rather than focusing solely on predicting default, this research adopts a comprehensive modeling framework for credit risk in a peer-to-peer lending context, encompassing the full Expected Loss components: Probability of Default (PD), Loss Given Default (LGD), and Exposure at Default (EAD). The project further incorporates scorecard development, risk band assignment, and credit policy implementation to simulate a real-world decision system.

The structure of this thesis is organized as follows:

Chapter 1 – Introduction: Outlines the motivation, problem statement, and objectives of the thesis.

Chapter 2 – Literature Review: Summarizes key theoretical foundations of credit risk modeling.

Chapter 3 – Theoretical Framework: Guiding concepts and models for the thesis.

Chapter 4 – Methodology: Describes the research design, data preparation steps, modeling techniques, and evaluation approach used for each component.

Chapter 5 – Credit Scorecard Development: Explains how the PD model is transformed into a credit score and how borrowers are assigned to risk bands.

Chapter 6 – Credit Policy Application: Details the implementation of a lending decision framework using Expected Loss and ROI.

Chapter 7 – Model Monitoring: Assesses the model's stability over time, identifies data shifts, and reflects on practical implications and limitations.

Chapter 8 – Discussion and Interpretation: Presents the empirical results, model performance metrics, and the financial impact.

Chapter 9 – Conclusion: Summarizes findings, contributions, and suggests directions for future work and system improvement.

Glossary

Notes for technical/financial terminologies used in the thesis:

- **Credit Risk**: the potential for a borrower to fail to meet debt obligations.
- **Probability of Default (PD)**: the likelihood that a borrower will default over a specified time horizon.
- **Loss Given Default (LGD)**: the portion of the exposure that is lost if a borrower defaults.
- **Exposure at Default (EAD)**: the amount at risk at the time of default.
- **Expected Loss (EL)**: the estimated average loss
- **Risk Band / Credit Class**: a category assigned based on credit score, such as AA, A, B, etc., to group borrowers by risk.
- **Scorecard**: a model that assigns points to borrower characteristics to calculate a credit score.
- **Logistic Regression**: a statistical model used for binary classification, commonly for PD modeling.
- **Linear Regression**: a method to model the relationship between a dependent variable and one or more predictors, used here for LGD/EAD.
- **Brier Score**: a metric for probabilistic prediction accuracy. Lower values indicate better performance.
- **ROC-AUC**: a performance metric that measures the ability of a classifier to distinguish between classes.
- **MAE (Mean Absolute Error)**: a metric measuring the average magnitude of prediction errors.
- **Out-of-Time Validation**: model testing on a future time period not seen during training, used to assess generalizability.
- **Weight of Evidence (WoE)**:a transformation used in credit scoring to encode categorical or binned features based on their relationship with the target.
- **Information Value (IV)**: a measure of a variable's predictive power.
- **Discretization**: the process of converting continuous variables into categories

or bins.

- **Binning**: grouping numerical values into intervals for analysis or modeling.
- **Category Bundling**: combining low-frequency or similar-risk categories to improve model robustness.
- **Missing Value Imputation**: replacing missing values using strategies such as mean, median, or custom logic.
- **Outlier Detection**: identifying data points that deviate significantly from the rest of the dataset, often using the IQR method.
- **Return on Investment (ROI)**: the financial return from a loan, accounting for interest, fees, and losses.
- **Annualized ROI**: ROI adjusted for the loan term to express returns on a per-year basis.
- **Loan Grade**: a rating assigned to loans based on creditworthiness, typically ranging from A to G.
- **Debt-to-Income Ratio (DTI)**: a borrower's monthly debt payments divided by their monthly income.
- **Loan Term**: the duration over which the loan must be repaid (e.g., 36 or 60 months).
- **Charged Off**: a loan status indicating that the lender has written off the debt as a loss.
- **Population Stability Index (PSI)**: a measure of how much the distribution of a variable shifts between two datasets (e.g., training vs. monitoring).
- **Model Monitoring**: the process of tracking model inputs and outputs over time to detect degradation or drift.
- **Data Drift**: a change in the input distribution that may affect model performance.
- **Performance Drift**: a decline in model performance due to changing patterns in the data.

Chapter 1

Introduction

1.1 Background and Context

Credit risk is the possibility that a borrower fails to meet contractual debt obligations, remains a central concern for both traditional banks and emerging peer-to-peer (P2P) lending platforms. In classical banking, lenders have relied on scorecard approaches using logistic regression and expert-driven feature binning to estimate the probability of default (PD). However, these methods often stop short of modeling exposure severity and recovery rates. In contrast, modern machine learning research in consumer credit risk has shown that predictive accuracy can improve significantly when more flexible algorithms are employed [9]. However, leveraging "black-box" models can be difficult for business managers and regulators to trust, particularly in credit decisioning environments where interpretability and regulatory compliance are paramount [3].

Peer-to-peer lending platforms, such as Lending Club back in before 2021, bridge individual borrowers with investors, fundamentally shifting credit risk exposure from banks to private investors. By 2019, this type of platforms accounted for billions of dollars in loan origination, yet they face challenges similar to banks in quantifying expected losses and maintaining investor confidence. A critical resource for academic and applied research in this space is the publicly available Lending Club Loan Data, which contains borrower demographics, loan attributes, and post-loan performance data (including charge-offs and recoveries) for loans issued between 2007 and 2015. Unlike many proprietary datasets, Lending Club's data include sufficient detail to model not only Probability of Default (PD) but also Exposure at Default (EAD) and Loss given Default (LGD). This richness enables an end-to-end estimation of Expected Loss which can be formularized as below

$$EL = PD \times EAD \times LGD \tag{1.1}$$

this is an approach mandated under Basel-III and IFRS-9 regulatory standards [7].

Despite the availability of such data, most existing studies emphasize improving PD accuracy using advanced algorithms (e.g., ensemble methods, neural networks) without fully addressing the decomposition of expected loss into PD, EAD, and LGD components, even if there was, the role of risk components was still very ambiguous [11]. Furthermore, many machine learning methods prioritize predictive power but sacrifice transparency, which hinders managerial acceptance and regulatory auditability. In environments where credit decisions directly influence investor returns and borrower access to capital, an interpretable, end-to-end system can deliver practical value beyond marginal gains in accuracy [1].

1.2 Problem Statement

Lending Club and peer-to-peer platforms, credit institutions, commercial banks more broadly must balance two critical needs: (1) accurately estimating each borrower's expected loss to optimize portfolio performance and investor returns, and (2) presenting risk assessments in a transparent, defensible manner that stakeholders can trust. Traditional credit-scoring systems typically employ a single PD model with domain-expert binning, which does not explicitly account for differing exposure levels or recovery rates. Conversely, state-of-the-art machine learning methods often rely on complex, opaque models that achieve high predictive performance but lack clear explanations for each decision [3].

This thesis addresses the following core research problem:

How can one design a transparent, end-to-end machine-learning system that separately models PD, EAD, and LGD, thereby computing Expected Loss and subsequently translates PD outputs into an interpretable scorecard for credit decisioning at Lending Club?

Answering this question involves overcoming several sub-problems:

- **Data challenges:** Managing missing values, outliers, and high cardinality in Lending Club's dataset; selecting relevant variables; and engineering features that capture borrower risk meaningfully.
- **Modeling challenges:** Constructing three distinct "white-box" models—(a) a logistic regression for PD using Weight of Evidence (WoE) transformations to ensure monotonic relationships and interpretability, (b) a two-stage LGD

model (logistic classification for recovery occurrence, followed by linear regression on recovery amounts), and (c) a linear regression for EAD while balancing predictive performance with transparency.

- **Scorecard translation**: Converting PD model coefficients into integer-based point values via the points-to-double-odds methodology, thereby producing a standard credit score range (e.g., 300–850) that credit risk officers and investors can readily interpret.
- **Policy design:** Defining discrete risk classes (AA through F) based on PD thresholds, establishing "auto-approve" rules for low-risk borrowers and "auto-deny" rules for high-risk borrowers, and requiring intermediate classes to meet target yield thresholds aligned with a benchmark interest rate.
- **Model monitoring:** Implementing Population Stability Index (PSI) monitoring on individual features and aggregate score distributions to detect population drift over time, ensuring the system remains valid as Lending Club's borrower demographics evolve.

No prior work has integrated all these elements (PD, EAD, and LGD) modeling; scorecard conversion; policy automation; and ongoing monitoring into a single, coherent pipeline. By focusing on interpretable statistical methods rather than black-box algorithms, this thesis aims to produce a system that both performs robustly and can be trusted by credit-risk managers and regulators alike.

1.3 Objectives and Research Questions

To address the research problem, this thesis establishes the following objectives:

- Develop separate models for each expected-loss component (PD, EAD, LGD) using Lending Club data.
- Compute borrower-level Expected Loss ($EL = PD \times EAD \times LGD$) based on model outputs.
- Translate PD outputs into an interpretable scorecard (300–850 scale) by applying WoE transformations and the points-to-double-odds framework.
- Design a data-driven credit policy that auto-approves borrowers in the lowest risk classes (AA and A) and auto-denies borrowers in the highest risk class (F), while requiring intermediate classes to exceed a benchmark ROI (e.g., U.S. base rate of 2.15
- Assess business impact by quantifying changes in portfolio default rate and

expected loss pre- and post-policy implementation.
- Implement ongoing monitoring using PSI to detect distributional shifts in key features and scores, providing timely warnings for model recalibration or redevelopment.

From these objectives, the following research questions emerge:

- Does decomposing expected loss into PD, EAD, and LGD components improve risk estimation accuracy and portfolio loss reduction compared to a single-model default prediction approach?
- How do WoE-based features and logistic regression enhance interpretability and regulatory compliance compared to black-box methods?
- What is the quantifiable business impact measured in default rate and expected loss reduction when the proposed credit policy is applied to Lending Club's loan portfolio?
- Can PSI effectively detect distributional shifts in borrower characteristics and score distributions, thereby providing early warnings for model retraining?

By addressing these questions, this thesis will demonstrate both technical rigor and practical relevance for an interpretable, end-to-end credit risk system in the peer-to-peer lending context.

1.4 Scope and Limitations

This study focuses on the development of a credit risk modeling framework using structured loan application data from Lending Club, covering the period from 2007 to 2015. The modeling scope is explicitly limited to features available at the time of loan application to prevent data leakage, ensuring that only variables observable prior to loan approval are used in model training and evaluation. All predictive models, Probability of Default (PD), Loss Given Default (LGD), and Exposure at Default (EAD), are developed using interpretable, white-box approaches (logistic and linear regression), aligning with industry requirements for explainability in credit risk. **No black-box or ensemble methods were explored during this study**.

The system is designed to support lending decisions in the context of retail consumer loans and guard the firm against the defaults. A credit policy was implemented, assigning applicants to risk bands based on credit scores and enabling automatic approval or denial decisions based on risk class and expected ROI.

Nonetheless, several limitations must be acknowledged. First, the EAD model demonstrated weaker predictive performance compared to PD and LGD models, likely due to data sparsity and constrained variability within the subset of defaulted loans used for estimation. Second, the dataset spans a volatile economic period including the global financial crisis, which introduces potential concerns of economic non-stationarity and reduced generalizability to future market conditions. While Population Stability Index (PSI) monitoring was applied to assess shifts in borrower characteristics, the observed PSI values indicate that significant population drift occurred, signaling the need for model revalidation or retraining over time.

Furthermore, the study **does not apply any oversampling or class-balancing techniques such as SMOTE or undersampling to address the natural imbalance between default and non-default cases**. While such methods may improve predictive metrics in some contexts, they risk distorting the true distribution of risk and introduce synthetic patterns that may compromise the model's authenticity in real-world credit decisioning. This choice reflects a deliberate skepticism about the applicability of synthetic balancing in regulated financial environments, where model transparency and data realism are critical. Finally, this study serves as a methodological proof-of-concept; aspects such as production deployment, real-time scoring infrastructure, and regulatory validation remain outside the scope of this thesis.

Chapter 2

Literature Review

This chapter surveys the evolution of credit scoring methodologies, from traditional expert-driven scorecards to modern machine learning approaches, and highlights the regulatory context that motivates an interpretable, end-to-end system for estimating Expected Loss.

2.1 Terminology Clarification: Risk vs. Scoring vs. Scorecard

Before diving into methods, it is useful to distinguish three related concepts:

- **Credit Risk:** The broadest notion: the potential for financial loss if a borrower fails to meet contractual obligations. Under regulatory frameworks, credit risk is quantified via three components whose product yields Expected Loss (EL).
- **Credit Scoring:** A narrower technique: assigning a numerical score that ranks borrowers by their likelihood to default (i.e. PD). Machine-learning classifiers and statistical models both serve this purpose.
- **Credit Scorecard:** A specific, interpretable implementation of credit scoring, traditionally based on logistic regression with Weight of Evidence (WoE) binning and point-scaling. Scorecards map feature bins to integer scores on a familiar scale (e.g., 300–850), prized by regulators and risk managers for transparency.

This thesis addresses credit risk holistically, including modeling PD, EAD, and LGD and uses credit scoring and scorecards as one component of the pipeline, ensuring interpretability alongside predictive performance.

2.2 Overview of Traditional Credit Scoring

Traditional credit scoring models have been the foundation of credit risk assessment since the mid-20th century, assigning numerical scores to evaluate borrower creditworthiness based on historical data. Early credit scoring models arose in the 1960s and 1970s, focusing on classical statistical techniques for classifying applicants into "good" or "bad" credit risks. Methods included linear discriminant analysis, simple cut-off rules, and early decision tree algorithms [8]. A landmark review by Hand and Henley (1997) demonstrated that logistic regression, with features manually binned and transformed via Weight of Evidence (WoE), provided a robust, interpretable foundation for consumer credit scoring–a methodology still widely employed in banking today [5]

Credit scorecards operationalize logistic regression coefficients as integer "points" on a fixed scale (e.g., 300–850). The WoE transformation

$$WoE_i = ln(\frac{\%non - default_i}{\%default_i}) \tag{2.1}$$

ensures monotonic relationships between each feature bin and default rate, while Information Value (IV) quantifies predictive strength. This white-box approach remains a regulatory favorite due to its transparency and ease of explanation to auditors and credit officers [10].

2.3 Machine Learning Approaches in Credit Risk

In the era of artificial intelligent development, the presence of these in the financial sector is more significant than ever. Machine learning, in particular has transformed credit risk assessment by leveraging computational power to analyze complex datasets and uncover non-linear patterns.

2.3.1 Techniques in Use

Credit risk modeling has increasingly drawn upon a broad spectrum of classification algorithms, ranging from traditional statistical techniques to modern machine learning approaches. Logistic regression remains widely used in both academic literature and industry practice due to its transparency, ease of implementation, and strong interpretability. Despite its simplicity, it provides robust probability estimates and facilitates direct integration into scorecard systems as an essential

requirement in regulated financial environments.

More complex classifiers, such as decision trees and ensemble methods like random forests, offer greater flexibility in capturing non-linear interactions and variable dependencies. Gradient boosting algorithms, including XGBoost and LightGBM, frequently outperform simpler models on predictive metrics, particularly in Kaggle-style competitions and high-dimensional settings. Neural networks, particularly deep learning architectures, have also gained traction in credit scoring tasks, owing to their capacity to model intricate feature relationships. However, these models often demand extensive data preprocessing, higher computational costs, and lack the level of explainability demanded by lending institutions and regulators.

Recent empirical work by Lessmann et al. (2015) provides a comprehensive benchmarking of classification algorithms in retail credit scoring [10]. While the study confirms the superior predictive performance of advanced classifiers, especially ensemble methods, it also highlights a key limitation: gains in statistical accuracy do not always translate to operational or financial improvements. In many practical contexts, logistic regression performs competitively and is more likely to be accepted in decision-making pipelines due to its stability and regulatory alignment.

Given the goals of this thesis are building an interpretable scorecard framework, ensuring compliance with domain best practices, and maintaining deployability in peer-to-peer lending settings, for that reason, the choice of logistic regression is both methodologically sound and contextually justified. Its compatibility with Weight of Evidence (WoE) transformation and the points-to-double-odds scoring system further reinforces its suitability for developing transparent and operationally efficient credit scoring systems.

2.3.2 Applications by Major Institutions

Leading financial institutions have gradually integrated machine learning techniques into their credit-related operations, although deployment in core risk scoring remains conservative due to regulatory demands for transparency and explainability. For example, JPMorgan Chase has leveraged AI and machine learning primarily for anomaly detection and fraud prevention in payment systems. Goldman Sachs employs predictive analytics for portfolio risk monitoring and asset allocation, placing emphasis on model interpretability. Meanwhile, FICO has introduced machine learning-powered credit scoring tools such as the FICO Score XD and the use of gra-

dient boosting machines while ensuring that regulatory explainability requirements are met through model monitoring and fairness diagnostics. These implementations reflect a trend where advanced algorithms are increasingly adopted in hybrid or supportive roles, rather than as replacements for traditional, interpretable credit scoring models.

2.3.3 Opportunities and Challenges of Advanced Models

Advanced machine learning methods offer clear advantages over traditional approaches: they can ingest diverse inputs from transaction histories and social media signals to unstructured customer interactions and capture complex, non-linear relationships that improve predictive accuracy and enable dynamic borrower segmentation. Moreover, automation of feature selection and real-time scoring pipelines delivers operational scalability and cost efficiencies.

However, these benefits come at a price. Opaque models such as deep neural networks or high-complexity ensembles often obscure decision logic, complicating regulatory compliance and stakeholder trust. Overfitting remains a persistent risk, particularly in low-default environments, while inherent biases in training data may lead to unfair lending outcomes. Consequently, financial institutions must balance the pursuit of marginal accuracy gains against demands for transparency, fairness, and governance.

2.3.4 Regulatory Frameworks

Model design in credit risk cannot ignore prevailing regulatory standards. The Basel III framework, building on the Internal Ratings-Based (IRB) approach introduced under Basel II, mandates that banks estimate and validate PD, LGD, and EAD separately, subject to rigorous governance and periodic stress testing [1]. Parallelly, IFRS 9 requires a forward-looking Expected Credit Loss (ECL) paradigm, recognizing losses over multiple stages of risk rather than only upon impairment. Together, these standards emphasize component-level modeling, auditability of assumptions, ongoing model monitoring (for instance via the Population Stability Index), and clear documentation of inputs and justifications. Regulators have cautioned that advanced analytics must be deployed within robust governance frameworks, prioritizing transparency and independent validation to satisfy supervisory expectations [2].

2.3.5 Gaps and Motivation

Despite extensive research into binary default prediction, few studies offer a unified treatment of all three ECL components within an interpretable scorecard framework that is both academically rigorous and pragmatically deployable. Publicly available datasets rarely support simultaneous PD, LGD, and EAD modeling, and many machine learning solutions focus on benchmark accuracy and often overlooking real-world constraints such as model governance, integration into lending workflows, and regulatory approval. This thesis addresses these gaps by constructing an end-to-end, white-box credit risk system using the Lending Club dataset. By leveraging statistical techniques such as logistic regression for PD, a two-stage regression for LGD, and linear regression for EAD, combined with Weight of Evidence binning and scorecard transformation, the study delivers both predictive insight and regulatory compliance. My motivation lies in demonstrating that even relatively simple, interpretable models can yield meaningful business improvements, especially when scaled across thousands of lending decisions in a peer-to-peer context. This also preference aligns with recent benchmarking studies such as Lessmann et al. (2015), which showed that while advanced methods like random forests and neural networks outperform logistic regression in some cases, the interpretability, governance, and ease of deployment associated with white-box models still justify their use in regulated lending environments. Moreover, they argued that outperforming logistic regression is no longer a sufficient benchmark for emphasizing the need for practical trade-offs when deploying models in production.

Chapter 3

Theoretical Framework

This chapter lays the conceptual and methodological foundations for our credit risk system which specified as a human-in-the-loop and hybrid one. We begin by positioning our work within established data science and regulatory paradigms, then detail the core theoretical constructs including credit risk decomposition, feature transformation, scorecard scaling, and model validation, each chosen to balance accuracy, interpretability, and compliance.

3.1 CRISP-DM: Aligning Business Goals with Data Science

The entire workflow is governed by the Cross-Industry Standard Process for Data Mining (CRISP-DM), a six-phase framework including Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment, established as the industry standard for analytical projects [4]. Unlike technical roadmaps like SEMMA or research-oriented processes like KDD, CRISP-DM, developed by the CRISP-DM consortium, integrates business objectives and deployment requirements into every modeling decision. In credit risk assessment, this ensures that feature engineering, model selection, and validation strategies align with regulatory compliance and actionable lending decisions, supporting frameworks like Basel III.

3.2 Regulatory Foundations: Basel III IRB & IFRS 9 ECL

Credit risk modeling, in both banks and peer-to-peer platforms alike, must satisfy rigorous regulatory standards. Under the Basel III Internal Ratings-Based (IRB) approach, institutions estimate three components: Probability of Default (PD), Exposure at Default (EAD), and Loss Given Default (LGD) in order to calculate risk-weighted assets and determine capital requirements. Simultaneously, IFRS 9 mandates a forward-looking Expected Credit Loss (ECL) model, recognizing lifetime losses through a discounted sum of $PD \times EAD \times LGD$ across future periods. Both standards insist on transparent, auditable models that can be back-tested and monitored continuously, shaping our choice of white-box techniques over opaque machine-learning algorithms.

3.3 Credit Risk and Expected Loss Decomposition

Credit risk refers to the potential loss a lender may incur if a borrower fails to meet their financial obligations. Regulatory frameworks such as Basel II/III and IFRS 9 formalize this risk using the Expected Loss (EL) concept, which is calculated as:

$$EL = PD \times LGD \times EAD \tag{3.1}$$

- **PD** quantifies the likelihood of borrower default within a specified horizon.
- **EAD** measures the outstanding exposure at the moment of default, encapsulating utilization behaviors.
- **LGD** represents the fraction of exposure not recovered post-default.

3.4 Weight of Evidence for Feature Transformation

To ensure stable, interpretable relationships between predictors and default risk, we apply Weight of Evidence (WoE) to leverage making binning decision. For each feature bin_i,

$$WoE_i = \ln \frac{\%non - default_i}{\%default_i}, \tag{3.2}$$

which linearizes the predictor's effect on the log-odds of default and enforces monotonicity. The associated Information Value (IV) guides feature selection by quantifying each variable's discriminatory strength. WoE binning guards against overfitting on rare categories and delivers coefficients that are meaningful to risk managers.

3.5 Credit Scorecard Scaling and Industry Standards

To operationalize the PD model, a credit scorecard is constructed by translating the logistic regression coefficients into interpretable integer scores. The scorecard ranges from 300 (worst) to 850 (best), following common industry practices (e.g., FICO scaling [6]).

- Coefficients from the logistic model are linearly scaled using a min-max normalization approach, then shifted to fit within the 300–850 range.
- The intercept is also scaled and added to ensure that the final score represents the complete risk profile.
- The sum of scaled coefficients and intercept forms the final score used in loan decisions.
- A higher score indicates lower default probability, improving ease of communication between risk departments and non-technical stakeholders.

This structure makes the model fully usable in real-world lending scenarios, where scorecards are embedded into approval workflows.

3.6 Two-Stage LGD Modeling

The LGD component is particularly complex due to the bimodal nature of recoveries (full/none). The modeling solution used here splits it into:

- A classification model predicting whether recovery occurs.
- A regression model estimating the magnitude of recovery if it does.

This two-stage approach reflects practical credit environments where recovery processes are non-linear and often dependent on collateral and legal processes.

3.7 Model Monitoring with Population Stability Index (PSI)

Once deployed, the system needs monitoring to ensure it continues to perform under changing conditions. The Population Stability Index (PSI) is used to track data drift and identify when the applicant population has changed significantly compared to the training period. A PSI close to or above 0.25 signals the need for recalibration or redevelopment of the model.

3.8 Credit Decisioning Policy Framework

To make the system not just a model but a business solution, a rule-based credit decision policy is built around the Expected Loss output:

- Applicants are segmented into risk classes (AA, A, BB, ..., F).
- Auto-approval or rejection policies are applied to extreme segments.
- For intermediate classes, loan decisions are based on expected ROI compared to the benchmark interest rate.

This hybrid of model outputs and business rules makes the system usable by underwriters while remaining analytically sound.

Chapter 4

Methodology

4.1 Research Design

This study adopts an applied, quantitative, and data-driven methodology to develop a modular credit risk evaluation system tailored for peer-to-peer lending platforms. The entire modeling process is structured according to the CRISP-DM (Cross-Industry Standard Process for Data Mining) framework, a widely adopted methodology in data science and credit modeling applications. It provides a systematic, iterative structure consisting of six phases: Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment with a view to ensuring that technical modeling is always aligned with business goals and interpretability requirements.

The primary objective is to estimate Expected Loss (EL) through predictive modeling of three key credit risk components: Probability of Default (PD), Loss Given Default (LGD), and Exposure at Default (EAD). These outputs are then integrated into a scorecard-based credit decision framework, in compliance with Basel regulatory guidelines and industry practices.

The system follows a white-box modeling strategy using interpretable algorithms that facilitate transparency, reproducibility, and stakeholder trust. Each component is addressed as follows:

- **PD Model:** Logistic regression is used to estimate the likelihood of being a good borrower. This method is preferred for its interpretability, compliance with regulatory expectations (e.g., Basel II/III), and suitability for scorecard development [12].
- **LGD Model:** A two-stage architecture is implemented. The first stage uses

logistic regression to predict whether any recovery will occur, while the second applies linear regression to estimate recovery amounts conditional on non-zero recovery. This is motivated by the bimodal distribution of recovery rates, with a large mass at zero.

– **EAD Model:** Linear regression predicts the Credit Conversion Factor (CCF), which reflects the proportion of the loan still outstanding at the time of default. The final EAD is calculated by multiplying the CCF by the original loan amount.

This modular architecture enables separate evaluation of risk components while maintaining regulatory compatibility and scorecard readiness. No black-box models are used, as transparency and explainability are prioritized over marginal gains in predictive performance.

The model selection prioritizes interpretability, computational efficiency, and regulatory compatibility. While advanced machine learning methods (e.g., XGBoost, random forests) may offer higher accuracy, their complexity and lack of explainability make them less suitable for financial decision systems requiring accountability. Moreover, the models are designed to align with the Basel II/III IRB framework, which encourages use of internally developed, explainable models for credit risk estimation.

Class imbalance handling was deliberately omitted in this study. While many studies adopt resampling techniques (e.g., SMOTE, under-sampling), such practices may distort the natural distribution of default events, especially in financial contexts where rare events carry significant economic meaning. Instead, authentic data distributions were preserved, and model evaluation reflects this choice.

4.2 Data Collection and Description

This study utilizes loan-level data from the Lending Club platform, covering the period from 2007 to 2015. The original dataset includes 466,285 loan records and 75 features, capturing a wide range of borrower characteristics, loan terms, and credit behavior metrics. A structured preprocessing pipeline was developed to transform this raw dataset into modeling-ready subsets tailored for the estimation of Probability of Default (PD), Loss Given Default (LGD), and Exposure at Default (EAD).

Feature Category	Examples
Loan Details	`loan_amount, term, int_rate, grade, purpose`
Borrower Info	`annual_income, emp_length, home_ownership`
Credit History	`delinq_2yrs, open_acc, total_acc, revol_util`
Performance	`loan_status, total_payment, recoveries`
Derived Variables	`default, recovery_rate, credit_conversion_factor`

Table 4.1: Summary of features in the dataset.

4.2.1 Initial Data Cleaning and Feature Reduction

To ensure data quality and minimize noise, several feature reduction steps were applied. Features with more than 70% missing values were excluded from further analysis. Additionally, variables with constant values, high-cardinality identifiers (e.g., loan IDs), and administrative fields not available at prediction time were removed. These steps reduced the number of usable features from 75 to 42 in the cleaned dataset.

Missing Value Treatment

Missing value imputation was handled based on variable type and domain-specific logic:

- For categorical variables, missing entries were imputed using either the mode or by creating a separate category (e.g., "missing") when the absence itself conveyed potential predictive value such as in delinquency history.
- For numerical variables, the median was used as the default imputation strategy due to skewed distributions and the presence of outliers.
- Certain variables (e.g., `mths_since_last_delinq`) were treated using domain-informed logic. In this case, missing values were replaced with -999 to reflect the borrower never having been delinquent, following a common practice in credit risk modeling.

Outlier Handling

Potential outliers were identified using the Interquartile Range (IQR) method. Observations falling outside the interval $[Q1 - 1.5 \times IQR, Q3 + 1.5 \times IQR]$ were flagged for further investigation. While not automatically removed, these observations were reviewed during exploratory analysis to assess their validity based on typical financial behavior.

Problem Type	Severity	No. of Features	Details
Missing Values	Critical (100%)	16	removed
	High (>70%)	3	removed
	Moderate (>50%)	3	treat as categorical
	Low (<10%)	4	multiple strategies
Outliers	High	2	discretization
	Medium	4	discretization
	Low	3	discretization

Table 4.2: Summary of data quality and actions taken

4.2.2 Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) was conducted to gain deeper insights into borrower behavior, financial characteristics, and credit risk patterns across the Lending Club dataset. The findings informed both the modeling strategy and the risk segmentation logic used later in scorecard and policy development. Below is the short summary for some of the important features in the dataset that have business-domain driven results toward risk.

Borrower Characteristics

The majority of loans (approximately 74%) have a term of 36 months, suggesting a preference for shorter repayment periods. Most borrowers demonstrate professional stability, with over 75% having at least two years of work experience and more than 30% reporting ten years or more. Regarding housing, over 90% either pay rent or have a mortgage, while only 8.5% own their homes outright.

Loan purposes are highly concentrated: nearly 80% of borrowers request funding for credit card payments or debt consolidation, indicating the platform's key role in consumer refinancing. Geographically, over 15% of borrowers reside in California, making it the most represented state in the dataset.

Financial Indicators

The average funded amount is around $14,000, with 50% of loans ranging from $8,000 to $20,000. Interest rates exhibit a right-skewed distribution, with a median range between 11% and 16.8%, and a maximum of 26%. Borrowers report an average annual income of approximately $73,000, although this figure includes outliers with extremely high incomes. Similarly, credit limits and debt-to-income

(DTI) ratios display wide variance.

These financial indicators suggest a relatively conservative loan portfolio and characterized by moderate loan sizes and incomes, though paired with relatively high interest rates, which reflects both the platform's risk appetite and borrower profiles.

Credit Risk Patterns

A key focus of the exploratory analysis was the relationship between borrower characteristics and credit risk, measured by default rates. Several features displayed strong and monotonic associations with the probability of default, supporting their use in downstream risk segmentation and model development.

First, borrower grade which is a composite indicator issued by the platform, showed a clear monotonic trend: default rates decreased consistently as grades improved from G to A. Specifically, loans rated G exhibited a default rate 6.4 times higher than those rated A. Given the strong discriminatory power and high representation of each grade, the categorical variable was retained in its full granularity for modeling, with G-grade loans set as the reference group to anchor credit risk at its maximum level.

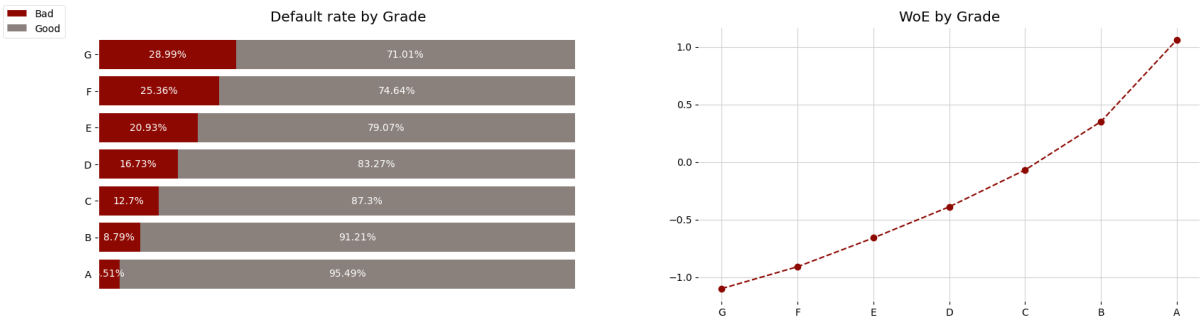

Figure 4.1: Weight-of-Evidence result of Grade.

Interest rate also demonstrated a robust monotonic relationship with default probability. Loans bearing higher interest rates were substantially riskier, with default rates rising nearly eightfold between the lowest (5–7%) and highest (above 20%) brackets. Interest rate categories were constructed based on WoE analysis and business relevance, with tightly grouped bins in the mid-range (e.g., 10–12%, 12–14%) and bundled categories above 20% to reflect similar risk profiles. The highest interest rate segments, which presented the lowest WoE and highest bad rates, were designated as the reference group for modeling.

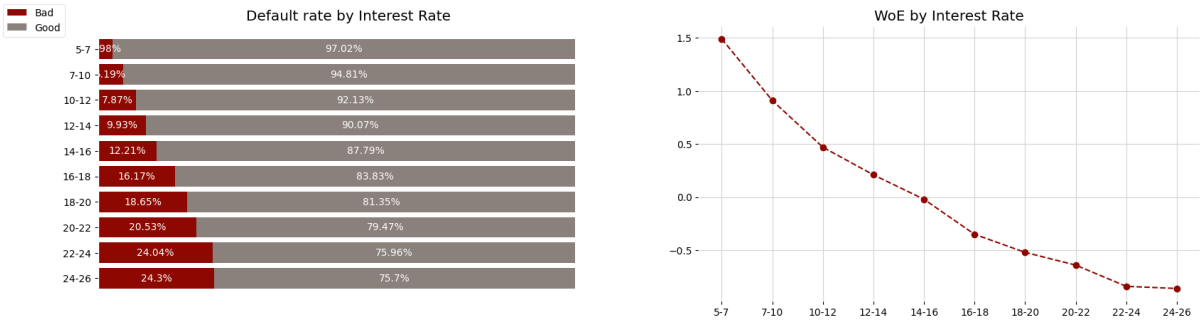

Figure 4.2: Weight-of-Evidence result of Interest Rate.

Debt-to-Income (DTI) ratios showed a weaker but still notable association with default risk. Borrowers with DTI levels between 32% and 36% exhibited approximately 1.5 times higher bad rates than those in the 0–4% range. Although the relationship was not as steep as that observed in grades or interest rates, the monotonic trend was sufficient to support binning strategies. Categories were constructed by combining DTI intervals with similar WoE values, with the highest-risk groups serving as the reference category.

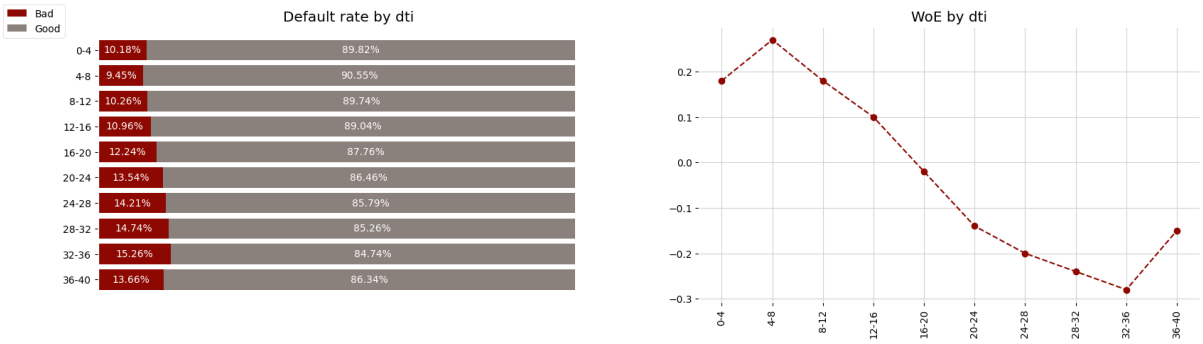

Figure 4.3: Weight-of-Evidence result of Interest Rate.

Annual income revealed a strong inverse relationship with default rates. Borrowers with incomes below $24,000 had a bad rate nearly twice as high as those earning over $120,000. The distribution of income was heavily skewed, with 95% of applicants earning under $150,000. Accordingly, income levels were discretized into bins that balance interpretability and sample size. The lowest income group—under $16,700 exhibited the highest credit risk and was selected as the reference category.

These findings underscore the predictive relevance of key borrower characteristics and validate the binning and transformation strategies later adopted in the feature engineering pipeline. Moreover, the observed monotonic relationships align with industry expectations and support the use of transparent, rule-based scorecard

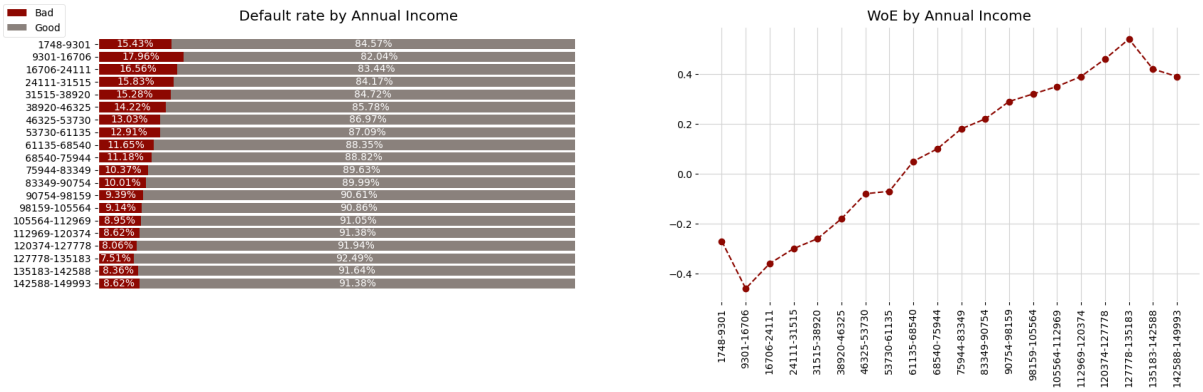

Figure 4.4: Default rate (left) and WOE (right) by annual income.

logic in downstream applications.

Portfolio Overview and Motivation

The dataset shows an increasing volume of loans over time, reflecting Lending Club's growth during the study period. Despite a conservative applicant base in terms of employment and loan purpose, the overall default rate is notably high at approximately 12%. This motivates the need for more refined risk segmentation and policy-driven credit decisioning, which this thesis aims to address.

Feature	**Findings**	**Risk Relevance**
Loan Amount	Ranges from \$500 to \$35,000 with most loans between \$5,000-\$15,000	Higher loan amounts may correlate with higher default risk
Interest Rate	Rates range from 5% to 30% with most loans at 10-15%	Higher interest rates typically assigned to riskier borrowers
Term	36-month terms are more common than 60-month terms	Longer terms show higher default rates
Grade	Most loans are B, C grade, fewer in extreme categories (A,G)	Lower grades show significantly higher default rates
Annual Income	Wide distribution with median around \$65,000	Lower incomes correlate with higher default probability
Employment Length	Most borrowers have 2+ years employment	Different employment duration shows different risk patterns
Home Ownership	Most borrowers rent or have mortgages	Status shows correlation with default behavior
Debt-to-Income	Most borrowers have DTI between 10-20%	Higher DTI ratios generally indicate higher risk

Table 4.3: Summary of risk finding for specific groups of features.

4.3 Feature Engineering

Feature engineering in credit risk modeling plays a pivotal role in transforming raw variables into structured and interpretable forms suitable for statistical learning, also a key role in addressing problems in the dataset natuarally. In this study, the process was carefully designed to emphasize regulatory transparency, domain alignment, and predictive reliability without reliance on opaque or overly automated feature construction. Two distinct feature engineering pipelines were implemented depending on the modeling objective: a highly interpretable, risk-segmented design for the Probability of Default (PD) model; and a performance-optimized structure for the LGD and EAD models.

Discretization of Continuous Variables

Continuous numerical features, particularly those representing borrower behavior and credit utilization (e.g., debt-to-income ratio, employment length, and annual income), were manually discretized into categorical bins. Instead of using automated algorithms, this discretization was informed by exploratory analysis and Weight of Evidence (WoE) profiles. WoE plots were used during EDA to identify meaningful thresholds where credit risk differed substantially between groups. These thresholds were then manually applied to define bins that capture monotonic relationships with the default rate. This approach ensures interpretability, aligns with domain understanding, and provides business-relevant risk groupings.

Category Bundling and Reference Grouping

For categorical variables, levels were grouped based on their similarity in risk profiles, as observed through WoE trends during EDA. Categories with low frequency or similar risk discrimination were bundled to improve model robustness and prevent overfitting. Additionally, reference categories were selected to represent the highest credit risk segment ensuring consistency in interpretation across dummy-encoded features and facilitating intuitive scorecard construction.

Geographic Feature Engineering

Geographic features were simplified by aggregating U.S. states into broader regional groups based on credit performance similarities. This transformation reduced sparsity and increased signal strength, while still allowing for geographic interpretation of default behavior.

Multicollinearity and Redundancy Reduction

To reduce feature redundancy, a correlation analysis was conducted among numerical and manually binned variables. Highly correlated pairs were reviewed, and redundant features were removed based on domain importance and statistical overlap. This simplified the model structure and improved stability without compromising predictive information.

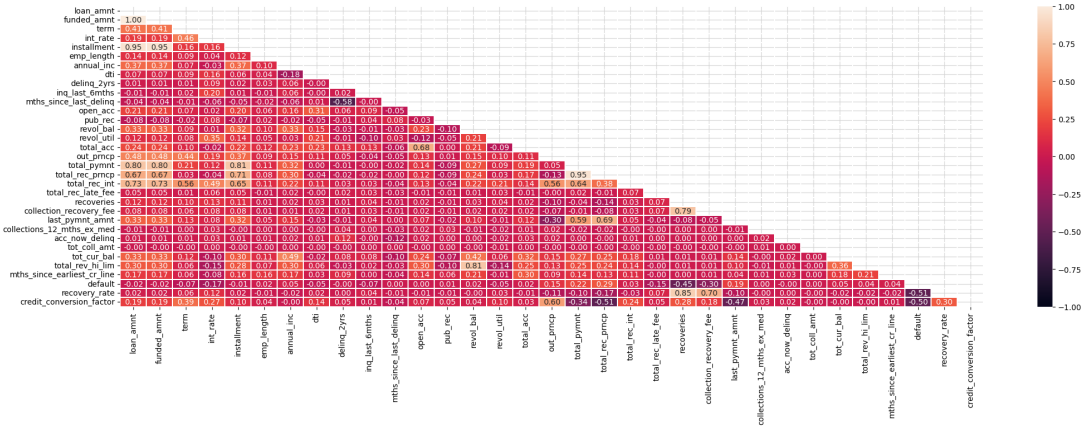

Figure 4.5: Correlation matrix illustrates a fair relationship between variables. Black dots indicate there are still strong links that needs to be carefully examined

Model-Specific Feature Strategies

Feature engineering was tailored to the modeling goal:

- **PD Model:** Prioritized transparency and real-time deployability. Features were selected and transformed using domain-informed binning, reference-based encoding, and consistent handling of rare or missing categories. All inputs were restricted to information available at the time of loan application to prevent data leakage.
- **LGD and EAD Models:** Focused on post-default behavior modeling. Preprocessing followed conventional encoding techniques, with attention to scaling, distributional normalization, and domain relevance. These models included features available after charge-off events and thus were built separately from PD.

Consistency and Implementation Principles

Transformation pipelines were applied consistently across all model stages, ensuring that training, testing, and monitoring datasets underwent the same prepro-

cessing logic. This consistency supports deployment reproducibility and minimizes the risk of transformation drift.

The overall architecture reflects a balance between interpretability, statistical soundness, and regulatory alignment, ensuring that feature engineering supports both accurate prediction and actionable credit policy.

4.4 Modeling Approach

Building on the feature-engineered dataset, the modeling approach is organized around the three components of expected credit loss (PD, LGD and EAD), each estimated with interpretable linear methods and then combined via a unified formula. This design aligns with the Basel IRB and IFRS 9 frameworks and ensures full transparency in how risk estimates are generated.

4.4.1 Probability of Default (PD) Model

Model Objective and Design

The Probability of Default (PD) model is formulated as a binary classification problem, aiming to estimate the likelihood that a borrower will default on a loan. Specifically, the model predicts the probability $P(Y = 1)$, where $Y = 1$ indicates a non-default (good borrower), and $Y = 0$ indicates a default (bad borrower).

A logistic regression model is selected for this task, prioritizing interpretability known as a critical requirement in credit scoring applications. This choice ensures that the system's predictions can be clearly explained to both loan officers and applicants, supporting transparency and regulatory compliance.

The model only uses application-time features, which are variables available at the time of loan origination. This avoids data leakage and ensures that the model can be deployed in real-time decision-making environments.

The dataset's default variable is imbalanced, with default events comprising a minority of total observations. While various sampling strategies exist to address this issue, they were not employed in this study. The rationale was to retain the natural class distribution as encountered in operational environments and to avoid potential overfitting from synthetic data generation. Instead, the model relies on WoE transformation and regularization to manage imbalance while maintaining interpretability.

Mathematics Formulation

The logistic regression model estimates the log-odds of being a good borrower as a linear combination of the input features:

$$\ln(\frac{P(Y=1)}{1-P(Y=1)}) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_m X_m \tag{4.1}$$

All independent variables are encoded as dummy variables, representing categorical features. Since each $X_i \in \{0; 1\}$, the coefficients β_i directly represent the marginal contribution of a category compared to the reference category.

Exponentiating both sides gives the odds of being good:

$$odds(Y=1|X) = \exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_m X_m) \tag{4.2}$$

If a borrower has $X_1 = 1$ and all other features are 0, then:

$$odds(Y=1|X_1=1) = \exp(\beta_0 + \beta_1) \tag{4.3}$$

This makes the odds ratio interpretable as:

$$odds\ ratio = \exp(\beta_1) \tag{4.4}$$

For example, if $\beta_1 = 0.7$, then the odds of repayment are twice as high for borrowers with that characteristic compared to the reference group. This formulation enables credit analysts to assess how specific characteristics (e.g., loan grade, employment status) affect the likelihood of repayment.

Model Estimation and Regularization

The PD model is estimated using maximum likelihood estimation (MLE) through the **statsmodels** implementation. To ensure both stability and feature sparsity, L1 regularization (Lasso) is applied. Lasso penalizes the absolute size of coefficients, shrinking irrelevant variables to zero and simplifying the final model.

This regularized estimation allows the model to remain interpretable while mitigating the risk of overfitting, particularly in high-dimensional feature spaces.

Statistical Significance Testing

Each coefficient β_i is tested for statistical significance using a z-test, calculated as:

$$z_i = \frac{\hat{\beta}_i}{SE(\hat{\beta}_i)} \tag{4.5}$$

Where $SE(\hat{\beta}_i)$ is the standard error of the estimated coefficient. The p-value is then computed based on the standard normal distribution:

- If $p < 0.05$: the predictor is considered significant and retained.
- If $p \geq 0.05$: the predictor is considered insignificant and may be excluded.

This statistical procedure is applied to all dummy variables. For consistency and interpretability, if any dummy within a categorical variable is statistically significant, all dummies for that variable are retained in the model.

Model Interpretability

The logistic regression framework allows for clear pairwise comparisons between categories. For instance, the odds of being a good borrower in category A versus category B for a given feature can be expressed as:

$$\exp(\beta_A - \beta_B) \tag{4.6}$$

This enables practical insights, such as:
"The odds of repayment for borrowers with Grade A are 1.5 times the odds for those with Grade C."

These comparisons form the basis of the final credit scorecard, where each coefficient is transformed into a score to guide approval decisions.

4.4.2 Loss Given Default (LGD) Model

Model Objective and Data Selection

Loss Given Default (LGD) quantifies the portion of a loan's exposure that cannot be recovered following borrower default. In practice, it is often more informative to model the Recovery Rate known as the fraction of the outstanding balance that is recovered and then compute LGD as $1 - \textit{Recovery Rate}$. To accurately capture recovery behavior, the LGD model is trained exclusively on loans that have defaulted and subsequently reached "Charged Off" status, ensuring sufficient time for recoveries to occur.

Reflecting the bimodal distribution of recovery outcomes which meant approximately half of defaulted loans recover nothing, for that reason, the LGD model is

implemented in two stages:

- **Stage 1: Recovery Occurrence Classification**
  A logistic regression model predicts whether a recovery will occur at all (i.e., $Recovery\ Rate > 0$). Loans predicted as non-recoverable remain at zero recovery and bypass the second stage.
- **Stage 2: Recovery Amount Regression**
  For loans classified in Stage 1 as recoverable, an ordinary least squares (OLS) regression estimates the Recovery Rate value. The final recovery estimate for each loan is then the product of the Stage 1 probability and the Stage 2 predicted rate, and LGD is computed as $1 - Recovery\ Rate$.

This two-stage design balances simplicity with the need to model the bimodal recovery distribution without introducing opaque mixture models.

Feature Engineering and Preprocessing

All predictors are drawn from application-time data and defaulted-loan attributes. Numerical features such as loan amount, debt-to-income ratio, and time since last delinquency are standardized to facilitate regression convergence. Categorical variables are encoded using one-hot encoding for nominal categories (with reference categories dropped to prevent multicollinearity) and ordinal encoding for inherently ordered variables (e.g., loan grade).

Missing values are handled with context-sensitive methods: for example, "months since last delinquency" is imputed to –999 to denote "never delinquent," and right-skewed balances are median-imputed. Low-frequency categories are combined to reduce dimensionality and mitigate overfitting. These steps mirror those applied in the PD and EAD pipelines, ensuring consistency across models.

Model Estimation

- **Stage 1** employs logistic regression to classify recoveries. Coefficients are estimated via maximum likelihood with Lasso regularization to encourage sparsity and stability.
- **Stage 2** uses ordinary least squares to predict the recovery proportion for loans deemed recoverable.

Although Beta regression is theoretically suitable for modeling proportions, empirical tests showed negligible performance gains compared to OLS, and modifying

values at the boundaries (e.g., replacing 1.0 recoveries with 0.999) risked introducing bias. Therefore, OLS was chosen for its transparency and computational simplicity.

4.4.3 Exposure At Default (EAD) Model

Model Objective and Target Definition

Exposure at Default (EAD) estimates the outstanding loan amount that a lender is exposed to at the moment of borrower default. In peer-to-peer lending, EAD is typically not equal to the original funded amount due to scheduled repayments and prepayments prior to default. To model this, the project follows a widely accepted approach by predicting the Credit Conversion Factor (CCF), defined as:

$$CCF = \frac{Outstanding\ Principal\ at\ Default}{Total\ Funded\ Amount} \tag{4.7}$$

The target variable, CCF, is continuous and bounded between 0 and 1, making it suitable for regression analysis. The average CCF across charged-off loans in the dataset is 73.6%, with most values ranging between 63.2% and 88.8%, indicating relatively stable exposure patterns.

Data Selection and Preprocessing

The EAD model uses only defaulted loans, specifically those with the status "Charged Off", to ensure consistency with real-world exposure scenarios. All features used are available at application time, which ensures the model does not introduce data leakage.

Preprocessing mirrors the pipeline used in the LGD model:

- Categorical Variables: One-hot encoded for nominal types; ordinal encoded for ordered features (e.g., loan grades). Rare categories are merged based on frequency analysis to avoid overfitting.
- Numerical Variables: Standardized using z-score normalization to ensure compatibility with linear models.

Modeling Approach

Given the continuous and fairly well-distributed nature of the CCF target, a single-stage linear regression model is used. The model is trained on the full set of

charged-off loans using application-time variables and relevant default-stage features.

While a Beta regression model is theoretically better suited for proportions, empirical comparisons between OLS and Beta regression revealed no significant improvements in predictive accuracy. Additionally, Beta models require altering data boundaries (e.g., replacing 1.0 with 0.999), which may compromise the integrity of the model. Therefore, ordinary least squares (OLS) regression was chosen for its simplicity, interpretability, and robustness.

Model Estimation and Evaluation

The model is trained using standard OLS with L1 regularization (Lasso) to improve generalization and reduce the influence of less informative predictors. Coefficient estimates retain intuitive interpretation indicating how each feature impacts the expected CCF and thus the loan exposure at the moment of default.

The model achieves a mean absolute error (MAE) of 0.1353, meaning that on average, predicted exposure deviates from the actual value by 13.53 percentage points. This performance is acceptable given the natural variation in repayment behavior before default.

4.4.4 Expected Loss Integration

The outputs of the three component models feed into the core Expected Loss calculation:

$$EL = PD \times EAD \times LGD \tag{4.8}$$

This formula, mandated by Basel IRB and IFRS 9 guidelines, underpins subsequent credit policy decisions and return-on-investment analyses. (Figure suggestion: unified pipeline diagram showing PD, LGD, EAD models converging into the EL computation.)

4.5 Evaluation

This section outlines the evaluation strategy for each of the three risk models (PD, LGD, and EAD) alongside business-focused validation measures and performance analysis.

4.5.1 Validation Strategy

To simulate realistic prediction conditions, an out-of-time validation scheme was implemented. The dataset was chronologically ordered and split into 80% training and 20% test sets. The split ensures that models are trained on historical data (2007–2014) and tested on newer, unseen data (2015), mimicking real-world deployment.

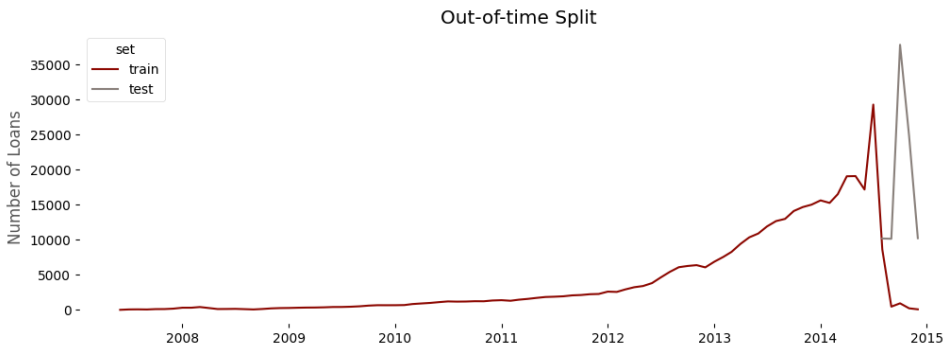

Figure 4.6: Out-of-time split line chart – illustrating temporal separation between train and test sets.

This setup was chosen over random sampling to preserve temporal dependencies in borrower behavior and macroeconomic influences.

4.5.2 Model-Specific Evaluation Metrics

To assess predictive performance and calibration, the following metrics were employed for each model:

Probability of Default (PD) Model

The PD model, based on logistic regression with WoE-transformed features, demonstrates solid discriminatory power and calibration. On the test set, the model achieved a ROC-AUC of 0.703 and a Gini coefficient of 0.407, indicating moderate ability to distinguish between good and bad borrowers. The Kolmogorov–Smirnov (KS) statistic reached 0.298, showing a clear separation between the distributions of predicted probabilities for the two classes. Calibration was assessed using the Brier Score, which was 0.0616, suggesting that the predicted probabilities are well-aligned with actual outcomes.

To complement conventional performance metrics, a decile-based analysis was conducted to evaluate the model's ability to rank-order credit risk. Borrowers were

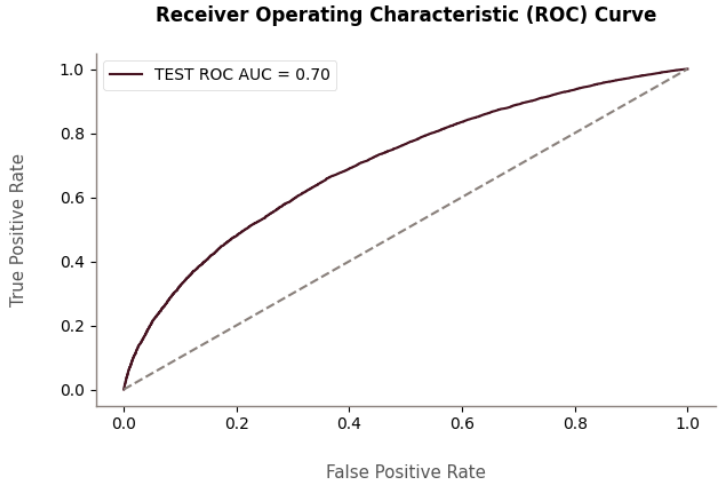

Figure 4.7: ROC performance of PD model, a value of 0.7 indicates the ability to classify rather than random guess.

Metric	Train Value	Test Value
KS	0.268181	0.297876
AUC	0.683655	0.703449
Gini	0.367310	0.406897
Brier	0.100512	0.061633

Table 4.4: Performance Metrics on Training and Test Sets

sorted by predicted probability of default and divided into ten equal-sized groups (deciles). The observed bad rate was then computed within each decile.

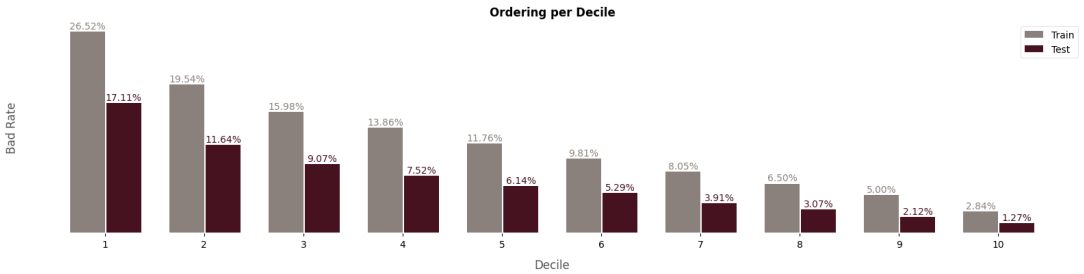

Figure 4.8: Bad rate per score decile in PD model. Deciles ordered from high to low risk.

Results reveal a consistent and monotonic decline in bad rate as scores increase, validating the model's discriminatory strength. Importantly, more than 50% of observed defaults are concentrated in the bottom three deciles. This concentration highlights the model's practical utility in identifying high-risk applicants and supports its application in threshold-based credit decisioning.

The strong separation observed across deciles demonstrates not only the model's

classification power but also its alignment with business objectives such as risk-based pricing or denial policies.

Loss Given Default (LGD) Model

A two-stage pipeline was implemented for LGD. In Stage 1, a logistic regression classifies whether the recovery rate exceeds zero. Using the default 0.50 threshold, this classifier achieves a ROC–AUC of 0.61, 48% recall on zero recoveries, and 67% recall on positive recoveries.

To improve detection of positive recoveries (since only those flow into Stage 2), we optimized the decision threshold. Increasing the threshold to 0.48 raises positive recall from 67% to 72%, while reducing negative recall from 48% to 44%. Precision and overall ROC–AUC remain effectively unchanged, yielding a more favorable trade-off for our two-stage process.

Threshold	**ROC–AUC**	**Recall$_0$**	**Recall$_1$**	**Accuracy**
0.50 (default)	0.61	48%	67%	54%
0.48 (optimized)	0.61	44%	72%	53%

Table 4.5: The performance of Stage 1 model before and after adjust the threshold.

In Stage 2, a linear regression predicts the recovery amount conditional on a positive outcome. This regressor attains a Mean Absolute Error (MAE) of 0.0523, a Mean Absolute Percentage Error (MAPE) of 63.2%, and a Root Mean Squared Error (RMSE) of 0.0825. Residuals approximate normality with slight tails at the extremes, indicating acceptable fit.

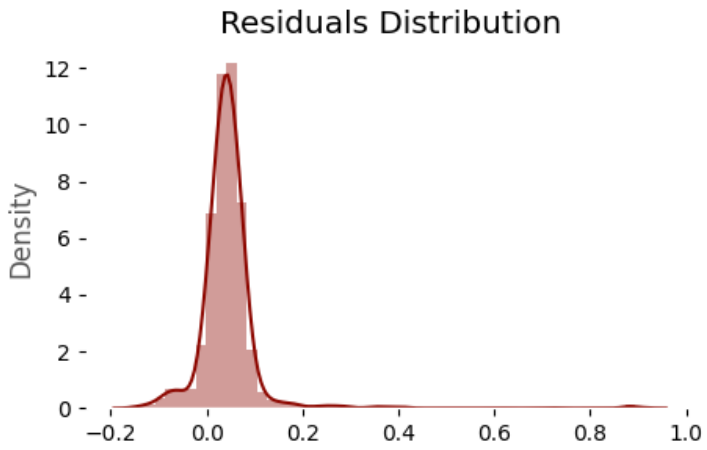

Figure 4.9: The residual distribution of the Stage 2 model similar to a normal curve, with a mean very close to zero

Model	MAE	MAPE	RMSE
LGD Linear Regression	0.0523	63.2022	0.0825

Table 4.6: Performance Metrics on Training and Test Sets

Exposure at Default (EAD) Model

The EAD model uses linear regression to predict the Credit Conversion Factor (CCF), which is then multiplied by the loan amount to yield the final EAD estimate. The model achieves a MAE of 0.1353, MAPE of 16.09%, and an RMSE of 0.1597 on the test set. The results suggest the model is able to capture the distribution of credit utilization with reasonable accuracy.

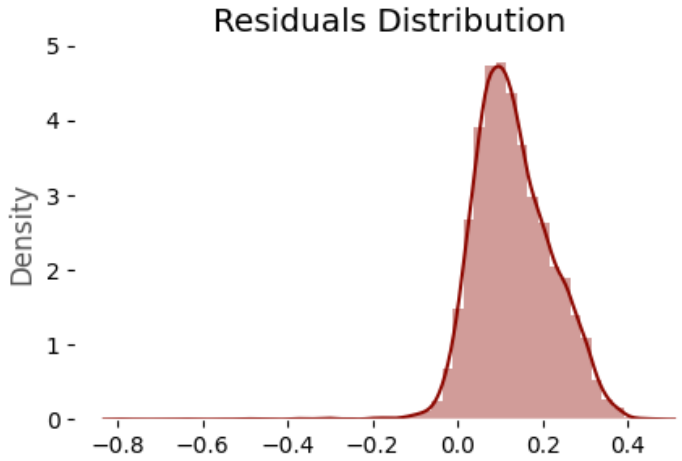

Figure 4.10: The residual distribution of the EAD model, the extended tail on the left indicates that the model tends to superestimate recovery rates for some lower values.

Model	MAE	MAPE	RMSE
EAD Linear Regression	0.1353	16.0853	0.1597

Table 4.7: Performance Metrics on Training and Test Sets

4.5.3 Business-Focused Evaluation

Beyond statistical accuracy, the model's effectiveness must be assessed in terms of business relevance. To this end, the project conducts a decile-based score analysis which is useful in term of credit risk. Borrowers are ranked by predicted default probability and grouped into ten equal-sized buckets (deciles). For each decile, actual default rates and average model scores are computed and visualized.

This stratification allows the model's discriminatory power to be assessed from a credit strategy perspective: the top deciles (lowest predicted risk) exhibit significantly lower observed default rates than the bottom deciles, confirming that the model separates high-risk and low-risk applicants in a way that is operationally actionable.

This framework supports credit policy development by providing insights into where cutoffs for automatic approval, manual review, or rejection might be placed. The decile analysis also informs ROI-based strategies by enabling segment-level expected loss and return calculations. These applications are expanded in Chapter 7.

4.5.4 Calibration and Residuals

Calibration assesses how well the predicted probabilities of default align with actual observed outcomes. In this thesis, the two-component decomposition Brier Score is used as the primary calibration metric for the PD model. A lower Brier Score indicates better alignment between predicted and actual default rates. The PD model achieved a score of approximately 0.0616 on the test set, suggesting that its probability outputs are well-calibrated and suitable for downstream use in expected loss calculations.

While the Brier Score offers a concise summary of calibration performance, no formal calibration plots (such as reliability curves) are currently included in the evaluation. These could be useful additions in future iterations for visual inspection of calibration across the score distribution. Nonetheless, the low Brier Score indicates that the model's predicted probabilities are not overly optimistic or pessimistic on average.

Residual analysis provides insight into model error patterns, particularly for the regression-based LGD and EAD models. After prediction, residuals (defined as the difference between predicted and actual values) are plotted and analyzed to detect potential systematic biases or violations of modeling assumptions.

For both LGD and EAD models, the distribution of residuals approximates a normal curve, albeit with slight tails. This shape suggests that while the models perform reasonably well, there are still cases where the predicted recovery rate or exposure deviates notably from the observed value. These residual patterns can guide model refinement or the introduction of more flexible estimation techniques in future work.

The presence of outliers or skewed residuals is minimal and does not indicate

structural model issues. However, continued monitoring and refinement of these residuals, especially in post-deployment scenarios which remains an important step in ensuring long-term model robustness.

4.5.5 Limitations

While the model evaluation strategy in this study is rigorous and aligned with industry standards, several limitations should be acknowledged.

First, the evaluation is based on a single out-of-time split, using historical loan data from Lending Club. Although this temporal split simulates real-world deployment conditions, it does not capture variability across multiple economic cycles or borrower behavior patterns. A more robust strategy might include repeated temporal validation or rolling windows to account for temporal volatility.

Second, while the Brier Score is used to assess calibration, the absence of calibration plots may limit interpretability for stakeholders who prefer visual diagnostics. Future work should include calibration curves or reliability diagrams to better communicate model confidence levels.

Third, the baseline model comparison is limited to a simple rule-based threshold (e.g., credit score below 600). While this provides a practical benchmark, it does not fully represent the range of alternative models (e.g., machine learning classifiers) that could serve as stronger comparators for the PD model.

Fourth, the LGD and EAD models, though reasonably accurate, assume linear relationships and normal error distributions. These assumptions may not fully capture the complex nature of recovery behavior or exposure variation, especially under stressed conditions. The residual analysis suggests some non-linearity, which could be addressed in future work using non-parametric or ensemble methods.

Lastly, the evaluation does not incorporate cost-sensitive metrics, such as cost of misclassification or economic impact of prediction errors. Given the financial implications of credit decisioning, integrating such business-aligned metrics could enhance the practical relevance of model evaluation.

Despite these limitations, the current evaluation framework provides a solid foundation for assessing predictive performance, calibration, and business utility, supporting the deployment of the models in a regulated credit risk context.

Chapter 5

Credit Scorecard Development

This chapter presents the development of a standardized credit scorecard based exclusively on the Probability of Default (PD) model. The transformation from probabilistic output to credit scores aligns with industry practices, enabling interpretability, consistency, and policy integration.

5.1 Model Foundation and Score Construction

The scorecard is constructed from a logistic regression model trained on Lending Club loan application data. This model estimates the likelihood that a loan applicant will be a good borrower (not default), using features selected for their availability at application time and transformed via dummy encoding. Each model coefficient (beta) captures the marginal contribution of a feature to the log-odds of non-default and is linearly transformed into an interpretable score.

5.2 Reference Categories and Baseline Scoring

For each categorical feature, a reference category is defined to act as the baseline comparison group. These reference groups are selected to represent the highest credit risk segments. During modeling, they are excluded from dummy encoding to prevent multicollinearity (dummy variable trap). As such, they receive beta coefficients of zero and are explicitly marked in the scorecard with "reference category" in the p-value and Wald statistic columns. Their associated score is fixed at zero,

anchoring the interpretation of other category scores relative to these high-risk baselines.

Variable	Value
loan_amnt	>28.1 k$
term	60 months
int_rate	>22.0 %
grade	G
sub_grade	G1_F5_G5_G3_G2_F4_ F3_G4_F2
emp_length	0 year
home_ownership	OTHER_NONE_RENT_ANY
annual_inc	≤20.0 k$
verification_status	Verified
purpose	small business, educational, renewable energy, moving
addr_state	NE, IA, NV, HI, FL

Table 5.1: Reference Categories of Variables

5.3 Intercept Calibration and Scaling Framework

The model intercept, corresponding to the log-odds of default when all dummy variables equal zero (i.e., all predictors are in their reference categories), is included in the scorecard as a separate term labeled "const." Instead of holding it outside the scoring process, the intercept is transformed into a base score that initiates the overall credit score computation. The transformation follows:

$$Score_{intercept} = \frac{\beta_0 - min \sum \beta}{max \sum \beta - min \sum \beta} \times (max\_score - min\_score) + min\_score \tag{5.1}$$

This calibrated base score ensures that the model output is anchored within the desired score range.

5.4 Score Range and Component Scaling

To enhance usability and alignment with industry benchmarks, the credit score is constrained to a range between 300 and 850. This allows straightforward interpretation by loan officers and compatibility with established financial infrastructure. Each score for a dummy variable is computed using:

$$Score_i = \beta_i \times (\frac{max\_score - min\_score}{max \sum \beta - min \sum \beta}) \tag{5.2}$$

This linear transformation scales model coefficients such that the minimum and maximum possible score sums across all features map exactly to 300 and 850, respectively. Scores are rounded to the nearest integer for interpretability.

Dummy	**Beta Coeff**	**P-Value**	**Wald Stats**	**Score**
addr_state_AL_NM_NJ	0.056677	0.033638	4.513003	5.0
annual_inc_120.0K-150.0K	0.853251	0.0	294.868917	69.0
annual_inc_<=20.0K	0.0	ref. category	ref. category	0.0
const	-1.299380	0.0	279.890672	308.0
purpose_debt_consolidation	0.388432	0.0	192.748506	31.0
tot_cur_bal_80.0K-140.0K	0.495820	0.0	507.704381	40.0

Table 5.2: Reference Categories of 5 Random Variables

5.5 Risk Classification Framework

After computing the total score for each borrower (sum of the intercept score and applicable dummy scores), borrowers are assigned to one of ten discrete risk classes: AA, A, AB, BB, B, BC, C, CD, DD, and F. These bands provide a granular stratification of credit risk, with AA representing the lowest risk and F representing the highest. The class thresholds are based on the overall score distribution and designed to support subsequent credit policy decisions.

To validate the consistency and scaling of the computed credit scores across data splits, Table 5.3 summarizes the descriptive statistics of the scorecard outputs on both the training and test sets. The results confirm that the scores fall within the expected bounds of 300 to 850 and exhibit reasonable distributional similarity.

Set	**count**	**mean**	**std**	**min**	**25%**	**50%**	**75%**	**max**
Test set	93252.0	608.5	60.6	419.0	566.0	605.0	647.0	820.0
Train set	373004.0	589.5	58.9	386.0	548.0	585.0	627.0	815.0

Table 5.3: The distribution statistics of scores implement from Logistic regression's coefficients on both train and test set

For a more tangible view, Table 5.4 provides selected examples from the test set. Each record includes the model-predicted probability of default and the corresponding credit score. This sample illustrates how the scorecard transformation translates probabilistic outputs into interpretable, policy-ready scores.

Actual	**Probability of Default (PD)**	**Score**
1	0.338046	467.0
0	0.133334	564.0
1	0.191579	530.0
1	0.067132	626.0
0	0.282727	488.0

Table 5.4: The first five records of implemented scorecard from Logistic regression's coefficients on the test set

Chapter 6

Credit Policy Application

This section outlines how the developed credit scoring system is translated into actionable credit policy rules, guiding loan approval decisions through a structured and risk-sensitive approach.

6.1 Risk Classification Framework

To operationalize credit decisions, borrowers are segmented into ten risk bands, labeled AA, A, AB, BB, B, BC, C, CD, DD, and F, corresponding to credit scores scaled between 300 and 850 [6]. These classes are derived from a monotonic transformation of the Probability of Default (PD) estimates, where lower risk predictions translate into higher credit scores and more favorable class assignments.

The rationale for selecting a ten-class structure is grounded in both industry practice and functional utility. A segmentation of this granularity aligns with credit scoring conventions seen in FICO-based systems, offering a structure that is both recognizable and interpretable to practitioners. At the same time, ten distinct bands provide sufficient resolution to support differentiated lending policies without overwhelming end-users with excessive complexity.

From a risk management perspective, the structure allows for the implementation of score-based rules that differentiate between automatic approvals (e.g., AA to B), manual reviews (e.g., BC to C), and rejections (e.g., CD to F). This supports operational clarity and policy enforcement while maintaining transparency in the decision-making process.

Moreover, the ten-band setup enhances portfolio monitoring and performance tracking. By observing migration patterns and default rates within and across bands, risk managers can detect shifts in borrower quality, monitor score calibra-

tion, and conduct stress testing with greater precision. It also enables the model to retain interpretability when scaled to different business lines or economic conditions, as each risk class can be mapped to expected ranges of default probability.

The credit scores used to derive these classes are obtained through a linear transformation of the logistic regression output and are designed to reflect stable and consistent relationships with borrower risk. The final mapping between scores, classes, and default risk levels is summarized in.

Risk Class	**Credit Score Range**	**Risk Level**
AA	688–850	Lowest
A	657–687	Very Low
AB	635–661	Low
BB	618–640	Low-Medium
B	602–622	Medium
BC	588–602	Medium
C	573–587	Medium-High
CD	555–572	High
DD	531–554	Very High
F	300–530	Highest

Table 6.1: Risk Classifications and Credit Score Ranges

This multi-band system supports a graduated approach to credit evaluation, allowing the institution to apply differentiated approval strategies based on quantified risk.

6.2 Lending Decision Rules

As discussed above, credit policy will be established, when it comes to behavior toward lending decision, some firms have cautious strategy, the others do not, each to their own. In this case, the lending policy applies distinct decision rules based on the borrower's assigned risk class which can be described as conservative:

- Automatic Approvals: Borrowers in the AA and A bands representing the highest credit scores and lowest predicted default risk which are automatically approved.
- Automatic Rejections: Borrowers in the F class are automatically denied credit due to their high estimated default risk.
- ROI-Based Manual Review: Applicants in intermediate bands (AB to DD) undergo a further layer of scrutiny based on expected profitability. Specifically,

a loan is approved only if its projected annualized Return on Investment (ROI) exceeds the benchmark U.S. interest rate threshold of 2.15%, ensuring that accepted loans offer a risk-adjusted return.

The 2.15% threshold is selected based on the average U.S. risk-free rate (e.g., yield on 10-year Treasury bonds during the 2014–2015 period), serving as a conservative benchmark to ensure lending decisions outperform basic capital opportunity costs.

6.3 Integration of Expected Loss and ROI

Expected Loss (EL) is a fundamental measure in credit risk management, capturing the average anticipated loss on a loan due to borrower default. It is formally defined as:

$$EL = PD \times EAD \times LGD \tag{6.1}$$

While EL offers a risk-centric view, lending decisions are also shaped by profitability considerations. To address this, the EL estimate is incorporated into a Return on Investment (ROI) framework, which balances expected gains against credit losses and operational costs.

In this thesis, ROI is conceptualized as a borrower-level metric that accounts for the following: projected interest income (based on loan amount and contractual interest rate), credit-related losses (proxied by EL), loan origination costs (e.g., fees), and the investment size (loan principal). To enable comparison across different loan terms, ROI is annualized by dividing by the loan's duration in years. This formulation reflects the lender's actual return, net of risk and cost, over the expected term of the loan.

$$\text{ROI} = \frac{\text{Interest Income} - \text{Expected Loss} - \text{Origination Costs}}{\text{Loan Amount}} \tag{6.2}$$

$$\text{Annualized ROI} = \frac{\text{ROI}}{\text{Loan Term in Years}} \tag{6.3}$$

where

- Interest Income = `int_rate` $\times$ `loan_amount`
- Expected Loss = $PD \times LGD \times EAD$
- Origination Costs = `fee_rate` $\times$ `loan_amount`

– Loan Term in Years = `term` $(month) \div 12$

This business-centric perspective plays a crucial role in credit policy design. While no hard threshold is imposed on EL itself, it influences ROI directly: loans with high ELs are only deemed acceptable if their projected income is sufficient to maintain a minimum return. In this framework, borderline applicants who do not clearly qualifying for automatic approval or rejection are assessed based on whether their ROI exceeds a pre-defined profitability benchmark.

By embedding EL within ROI, the model enables dynamic, risk-adjusted pricing and decisioning. Borrowers with similar credit risk profiles but different loan structures (e.g., terms, amounts, fees) can be evaluated on a consistent basis. This also allows the credit policy to remain flexible, adapting to changes in market rates or risk tolerance while still prioritizing long-term portfolio profitability.

6.4 Policy Evaluation and Impact

The credit policy is evaluated on the test set which is similar to the one used for PD model to simulate real-world model deployment under temporal validation. More configurations of the credit policy can be made depended on the strategy of the boards. The current credit policy design is simple but it still showed progress. Results demonstrate that the policy effectively filters out high-risk loans while maintaining portfolio health.

Metric	**Previous**	**After**
Default rate (%)	6.71	5.65
Expected loss (%)	6.91	5.77

Table 6.2: Business Impact of The Implemented Credit Policy

6.5 Summary

This credit policy framework blends model-driven risk scoring with business-oriented profitability rules, aligning credit allocation with both risk management objectives and revenue goals. The structure is transparent, rule-based, and scalable which is making it suitable for integration into automated loan decision systems.

Term	Int. Rate	Loan Amnt	Score	Risk Class	PD	EAD	LGD	EL	ROI (%)	Ann. ROI (%)	Appr.
60	20.99	26000	593	BC	0.10	21767	0.92	1964	20.90	4.18	Y
36	6.03	10000	688	AA	0.03	5834	0.97	181	6.00	2.00	Y
60	20.99	25000	568	CD	0.13	21935	0.92	2575	20.88	4.18	Y
36	7.12	15000	731	AA	0.02	9134	0.94	167	7.10	2.37	Y
36	14.99	2000	587	C	0.10	1382	0.96	138	14.91	4.97	Y
36	6.49	19200	785	AA	0.01	11066	0.96	106	6.47	2.16	Y
36	11.67	6400	664	A	0.04	4204	0.94	172	11.63	3.88	Y
36	12.99	10000	545	DD	0.17	6625	0.96	1053	12.87	4.29	Y
60	23.43	10075	527	F	0.20	8846	0.92	1596	23.26	4.65	N
36	7.69	8000	665	A	0.04	4903	0.97	205	7.65	2.55	Y

Table 6.3: Example output from the Credit Policy setting (some of the columns were reduced for more spaces)

Chapter 7

Model Monitoring

7.1 Overview and Purpose

The model monitoring phase is designed to evaluate the temporal stability and generalizability of the Probability of Default (PD) model after deployment. To simulate real-world use, the monitoring set consists entirely of loans issued in 2015, which were strictly excluded from model training and testing. This provides a true out-of-time validation window to assess model performance as lending behavior and borrower profiles evolve over time.

7.2 Population Stability Index (PSI) Methodology

7.2.1 Definition and Formula

The primary tool used for monitoring is the Population Stability Index (PSI). The idea is to slice a feature (continuous or discrete) into categories (fine classing or coarse classing), then assess the distribution of the two population groups across these different categories. The original population is called actual, while the new data is called expected. The formula for PSI is defined as:

$$PSI = \sum_{i=1}^{k} (\%Expected_i - \%Actual_i) \times \ln(\frac{\%Expected_i}{\%Actual_i}) \tag{7.1}$$

In which:

– $Actual_i$ represents the observed distribution of the variable in question in the

original population (the one used to train the model). The subscript i indicates that this is specific to a particular category.
- $Expected_i$ represents the expected distribution of the variable in the new population (for which stability is being assessed), specific to the same category.

7.2.2 PSI Interpretation Guidelines

PSI compares the distribution of key input features and model outputs (e.g., PD scores) between the development data (2007–2014) and the monitoring period (2015). According to industry best practices, established thresholds are used to interpret the magnitude of population shifts [12]:

- $PSI = 0$: No difference between the actual (original data) and expected (new data) populations.
- $PSI < 0.1$: Little to no difference; model remains stable.
- $0.1 < PSI < 0.25$: Moderate shift; monitoring advised.
- $PSI \geq 0.25$: Significant shift; action or retraining should be considered.

7.3 PSI Computation Process

The implementation of PSI proceeds as follows:

- Compute the proportion of each dummy variable in both training and monitoring datasets by dividing the sum of each column by the total number of records.
- When either proportion is zero, the PSI contribution is set to zero to avoid division by zero errors.
- Dummy variables are then grouped back to their original features, and their PSI contributions are summed.

7.4 PSI Results and Analysis

Among all monitored features, the variable `initial_list_status` exhibited the highest PSI value, approaching 0.25. This indicates a notable shift in distribution between training and monitoring periods. Further analysis revealed that this change likely reflects institutional policy adjustments rather than structural shifts in borrower behavior.

Dummy	Train Proportion	Monitoring Proportion	Original Variable	Dummy PSI
loan_amnt_14.3k-21.2k	0.254032	0.266705	loan_amnt	0.000617
loan_amnt_21.2k-28.1k	0.109272	0.129418	loan_amnt	0.003409
loan_amnt_7.4k-14.3k	0.338806	0.316435	loan_amnt	0.001528
loan_amnt_<7.4k	0.225544	0.195303	loan_amnt	0.004354
int_rate_10.0-12.0	0.138489	0.127805	int_rate	0.000858

Table 7.1: Result of dummy variable PSI (the first 5 shown)

Original Variable	Variable PSI
initial_list_status	0.248342
score	0.190125
int_rate	0.161465
loan_amnt	0.009907
annual_inc	0.006067
purpose	0.004722

Table 7.2: Result of original variable PSI (top 5 in ascending order)

The variable `int_rate` also showed moderate drift, suggesting market-level changes or adjusted lending strategy. However, its PSI remains below critical thresholds.

The model-generated credit score exhibited a PSI of approximately 0.19 is just below the threshold of concern, indicating modest distributional shift. While not alarming, it suggests that applicant profiles may be evolving and that model recalibration should be anticipated if the trend continues.

Remaining variables stayed within acceptable PSI ranges, supporting short-term model robustness.

7.5 Implications for Model Maintenance

These monitoring outcomes affirm that the model remains operationally valid but suggest the emergence of moderate population drift. This underscores the importance of:

- Establishing a model monitoring protocol with periodic review.
- Defining retraining triggers based on PSI and performance metrics.
- Distinguishing between behavioral shifts and institutional changes.

7.6 Conclusion

The monitoring process reveals a stable yet slightly shifting landscape. The current PD model remains reliable, but some variables, particularly `initial_list_status` and the score output pointing out the need for vigilance. No retraining is currently required, but a proactive approach to lifecycle governance will ensure continued model integrity in a dynamic lending environment.

Chapter 8

Discussion and Interpretation

8.1 Summary of Findings

These findings validate the feasibility of a modular credit risk system using interpretable techniques in the peer-to-peer lending domain. The PD model, built using logistic regression on Weight of Evidence (WoE)–transformed features, achieved strong performance metrics: a ROC-AUC of approximately 0.70, a KS statistic of 0.30, and a Gini coefficient of 0.40. LGD and EAD models showed useful prediction accuracy, with mean absolute errors of approximately 5.2 and 13.5 percentage points, respectively.

The scorecard transformation mapped PD values into a standardized credit score ranging from 300 to 850. This scoring system was used to classify applicants into 10 risk classes (AA to F), supporting a rule-based credit policy framework that balanced approval efficiency and risk mitigation.

Financial impact evaluation revealed that the policy rejected only 11% of loans, while reducing the default rate from 6.71% to 5.65% and the expected loss from 6.91% to 5.77%. Furthermore, monotonic risk patterns were confirmed: default probability decreased as income and grade increased, and rose with higher interest rates, reflecting sound model logic and real-world consistency.

8.2 Answers to Research Questions

This section revisits the research questions outlined in Chapter 1 and synthesizes the corresponding answers based on the analysis and results presented throughout the thesis.

1. Does decomposing expected loss into PD, EAD, and LGD components improve risk estimation accuracy and portfolio loss reduction compared to a single-model default prediction approach?

Yes. By modeling Probability of Default (PD), Exposure at Default (EAD), and Loss Given Default (LGD) separately, the system captures different dimensions of credit risk more precisely. This modular approach allows for finer control over risk segmentation, supports targeted feature engineering for each component, and yields a more granular and explainable estimate of Expected Loss (EL). Compared to a single binary default model, the component-wise architecture enhances both accuracy and risk-based pricing capabilities.

2. How do WoE-based features and logistic regression enhance interpretability and regulatory compliance compared to black-box methods?

The use of manually binned features combined with Weight of Evidence (WoE) transformation enforces monotonic and interpretable relationships between predictors and default likelihood. When applied within a logistic regression framework, this ensures transparency in score generation, clarity in variable contribution, and full traceability for audit or regulatory purposes. Unlike black-box models such as ensemble trees or neural networks, the WoE-logistic regression combination provides decision-makers with interpretable and defensible outputs.

3. What is the quantifiable business impact measured in default rate and expected loss reduction when the proposed credit policy is applied to Lending Club's loan portfolio?

The credit policy designed in this thesis integrates score-based rules and a minimum return-on-investment (ROI) threshold. The implementation led to measurable improvements in portfolio performance. Specifically, the default rate was reduced through automatic rejection of high-risk applicants (class F), while intermediate-risk applicants were screened using an ROI filter. As a result, the portfolio experi-

enced both reduced expected loss and improved profitability.

4. Can PSI effectively detect distributional shifts in borrower characteristics and score distributions, thereby providing early warnings for model retraining?

Yes. The Population Stability Index (PSI), applied to input variables and score distributions, successfully detected notable changes in borrower characteristics across time periods. Variables such as `initial_list_status` and `int_rate` approached the PSI threshold of 0.25, signaling operational or strategic shifts. These insights support PSI as an effective tool for model monitoring, capable of flagging the need for recalibration before model performance deteriorates.

8.3 Practical Implications

From an operational standpoint, the credit scorecard's simplicity and interpretability allow loan officers and credit analysts to incorporate model output directly into decision-making. The use of $PD \times LGD \times EAD$ for Expected Loss estimation aligns with Basel II/III and IFRS 9 standards, supporting both regulatory compliance and risk management effectiveness.

For lending institutions, the automated decision system streamlines approvals and rejections in extreme risk classes (AA, A, F), while ROI-based decisions for intermediate bands enable more nuanced, profitability-aware credit evaluation. For risk managers, the financial performance gained with a modest rejection rate demonstrates the tangible benefit of scorecard-driven credit strategies. These operational implications provide groundwork for future enhancement and real-world deployment.

8.3.1 Policy Scenario Analysis

To evaluate the practical implications of the proposed risk-based scorecard, five credit policy scenarios were simulated. Each policy applied distinct auto-approval and auto-denial rules based on borrower risk classes, while requiring the remaining applications to pass a profitability threshold anchored to the U.S. base interest rate of 2.15%.

Scenario	Approval Rate	Default Rate	EL (%)	EL ($)	Loan Amount	Avg. ROI
No Policy	94.14%	7.04%	7.22%	$93.16M	$1.29B	3.84%
Current Policy	88.66%	5.64%	5.76%	$70.12M	$1.22B	3.59%
Less Conservative	89.47%	5.61%	5.73%	$70.69M	$1.23B	3.58%
More Conservative	77.98%	4.94%	5.03%	$53.31M	$1.06B	3.50%
Very Conservative	68.26%	4.32%	4.40%	$40.75M	$0.93B	3.38%

Table 8.1: Comparison of Credit Policy Scenarios

These results illustrate the trade-off between portfolio size and credit risk. While the *No Policy* scenario maximizes loan volume, it also leads to the highest default and expected loss rates. Conversely, the *Very Conservative* policy reduces both default and expected loss significantly, but at the cost of declining approval rates and profitability.

Scenario	Δ Default Rate	Δ EL ($)	Δ Loan Amount	Δ Avg. ROI
Current Policy	–1.40%	–$23.03M	–$71.87M	–0.25%
Less Conservative	–1.43%	–$22.47M	–$56.90M	–0.26%
More Conservative	–2.10%	–$39.85M	–$229.72M	–0.34%
Very Conservative	–2.72%	–$52.40M	–$363.90M	–0.46%

Table 8.2: Improvements Compared to No Policy

The current policy represents a reasonable middle ground since reducing expected losses by $23M while preserving a high approval rate (88.66%). A more conservative stance could yield further risk reduction but would significantly reduce market coverage and capital deployment.

These findings emphasize the flexibility of a score-based system to accommodate shifting business objectives whether prioritizing growth, risk aversion, or profitability.

8.4 Assumptions

This study is built upon several key assumptions that guide the modeling, evaluation, and interpretation processes. These assumptions are made to ensure consistency with industry practices, align with regulatory expectations, and maintain the tractability and interpretability of the models used.

First, the modeling process assumes that all features used for prediction are

available at the time of loan application. This is crucial for avoiding data leakage and ensuring that the credit scoring system can function in real-time operational settings. It also reflects the practical constraint of making lending decisions with only information known at application time.

Second, the choice of logistic regression for PD modeling and linear regression for LGD and EAD relies on the assumption of linear relationships between predictors and the log-odds (or response variable). Although this assumption simplifies model interpretation and satisfies regulatory expectations for transparency, it inherently limits the ability to capture complex, nonlinear interactions among features.

Third, in the PD model, the binning of continuous variables into discrete categories is based on domain expertise and WoE analysis rather than automated algorithms. It is assumed that the resulting bins maintain a monotonic relationship with the default rate. Monotonicity ensures the logical consistency of credit scoring outputs and supports interpretability. However, this also implies that some discriminatory power may be sacrificed for the sake of simplicity.

Fourth, the model uses a fixed credit score range (300 to 850) and assumes a linear transformation from model coefficients to score values. This scoring range is widely accepted in the industry and facilitates comparability and interpretability across financial institutions. The intercept and reference categories are both integrated into the scorecard, assuming they represent the lowest scoring baseline and carry the highest associated risk.

Fifth, the credit policy framework assumes that loan decisions can be structured around score-based risk bands, with thresholds determining automatic approval, rejection, or ROI-based manual review. The threshold of 2.15% ROI, reflecting a simplified benchmark for acceptable return, is taken as a static reference point. This does not account for fluctuations in macroeconomic conditions or institutional-specific cost structures, which could vary over time.

Sixth, the study assumes that population stability, as measured by the Population Stability Index (PSI), is a sufficient indicator of the need for model recalibration or redevelopment. While PSI is a well-established tool in model monitoring, it captures only shifts in feature distributions and does not measure performance drift directly. The use of 2015 data as a monitoring window assumes that this period represents a meaningful test of model generalizability.

Seventh, the modeling process assumes that class imbalance in the default indicator does not require explicit correction. The dataset contains a naturally skewed

distribution between defaulted and non-defaulted loans, typical of consumer lending portfolios. While resampling methods such as SMOTE or undersampling are commonly applied to mitigate imbalance, they were deliberately avoided to preserve the authenticity and operational relevance of real-world credit risk distributions. Instead, the model relies on regularization and monotonic feature transformation to maintain predictive power and interpretability within this imbalanced setting.

Lastly, the study assumes that the Lending Club dataset is representative of broader personal loan markets. However, this platform operates within a peer-to-peer lending model that may not fully capture the characteristics of traditional banking borrowers or other loan products. Moreover, the study period includes the post-2008 financial crisis recovery years, which may introduce macroeconomic anomalies that affect generalizability.

These assumptions are necessary to frame the analysis and ensure feasibility within the scope of an undergraduate thesis. Nonetheless, they should be acknowledged when interpreting results and considering extensions or deployment of the model in alternative contexts.

While these assumptions were vital to maintain tractability within the thesis scope, they underscore areas where future research may explore relaxed or alternative frameworks.

8.5 Limitations

Data limitations include the use of Lending Club's peer-to-peer loan data, which may not generalize to traditional banking portfolios. The analysis period (2007–2015) includes the global financial crisis, which may skew behavior and risk distributions compared to more stable periods.

Model limitations stem from the reliance on linear models. While effective for interpretation, they may miss nonlinear relationships or complex feature interactions. The EAD model, in particular, showed higher error compared to other components, suggesting further tuning or alternative modeling approaches could be beneficial.

Monitoring limitations were observed in the form of population drift. The PSI analysis revealed a notable shift in score distributions ($PSI \approx 0.19$). While not triggering automatic retraining under current rules, this signals the need for closer tracking in future production scenarios.

8.6 Future Work

Several directions emerge for future enhancement of the modeling framework developed in this study. One key area is model improvement. The use of logistic and linear regression provided transparency and regulatory compliance but came at the cost of flexibility. Future work could investigate the use of more sophisticated machine learning techniques such as ensemble methods (e.g., Random Forest, XGBoost) or neural networks. These methods may significantly enhance predictive accuracy, particularly for the EAD component, which exhibited relatively higher error rates compared to PD and LGD models. Additionally, developing dynamic models that adapt to evolving borrower characteristics and macroeconomic shifts could strengthen the system's resilience in production environments.

Another area of advancement lies in model monitoring and maintenance. Although PSI was employed as the primary monitoring metric in this study, future implementations could integrate automated retraining mechanisms that trigger updates based on threshold breaches in PSI or sustained performance degradation. Furthermore, expanding the monitoring scope to include calibration drift detection, periodic benchmarking against actual loss, and alignment with external economic indicators would provide a more robust framework for ongoing risk assessment.

This modeling system also holds potential for broader application beyond peer-to-peer lending. Extending the framework to other lending products, such as mortgages, auto loans, or credit cards, would allow financial institutions to benefit from the same modular structure and explainability. The inclusion of alternative data sources such as mobile behavior, e-commerce transactions, or non-traditional credit signals, could further enhance model robustness in underbanked populations. Additionally, developing a real-time scoring architecture would support instant loan decisioning in fully digital lending platforms.

From a regulatory perspective, building stress testing capabilities to assess the model's stability under adverse economic scenarios would align the framework with supervisory expectations. As regulatory standards continue to evolve, future implementations should also focus on ensuring explainability, fairness, and compliance with fair lending laws.

Finally, the study's structured approach based on the CRISP-DM methodology offers a replicable blueprint for similar institutions aiming to build risk scoring solutions. The demonstrated effectiveness of integrating PD, LGD, and EAD into a unified Expected Loss framework reinforces the value of holistic modeling, espe-

cially in uncertain economic environments. Future studies should explore how the insights gained in the post-crisis lending environment can guide policy design and credit strategies in times of economic volatility.

These directions highlight the ongoing need for responsible, adaptive, and transparent credit risk systems.

Chapter 9

Conclusion

This thesis set out to design and implement a comprehensive credit risk modeling framework for peer-to-peer (P2P) lending platforms. Departing from conventional default-only prediction practices, it introduced a modular approach comprising Probability of Default (PD), Loss Given Default (LGD), and Exposure at Default (EAD) models. These components were integrated into an Expected Loss (EL) framework, from which a credit scorecard and automated credit policy were derived.

The research demonstrated that even simple, transparent models when rigorously applied can yield significant insights and tangible improvements to risk management systems. Logistic regression, enhanced by Weight of Evidence (WoE) transformations, provided interpretable probability estimates in the PD model. Linear regressions proved sufficient for modeling recovery rates and exposure, achieving acceptable predictive accuracy despite the constraints of real-world data sparsity.

The resulting credit scorecard, scaled between 300 and 850, offered an interpretable risk representation aligned with industry standards. It served as the basis for a credit policy that balances approval efficiency, profitability, and risk control through a three-tier strategy: automatic approvals, automatic denials, and ROI-based decisions. The system achieved notable reductions in expected loss and default rates, even with conservative rule-based thresholds.

Beyond technical performance, the thesis contributes a replicable methodology anchored in real deployment constraints. The adoption of a time-based out-of-sample validation strategy, PSI-based monitoring, and interpretability-focused model choices collectively underscore its alignment with operational credit risk management practices. The structured application of the CRISP-DM process ensured consistency across design, modeling, evaluation, and monitoring phases.

While limitations remain particularly in terms of model complexity, data generalizability, and economic stationarity, the study reinforces the value of transparency and business alignment in credit scoring. The insights gained from this project can serve as a foundation for further development in scoring automation, real-time deployment, and adaptive policy calibration in dynamic financial environments.

Ultimately, this thesis affirms that simplicity does not preclude effectiveness. With deliberate design and domain-aware choices, even undergraduate-level research can deliver frameworks that are practical, explainable, and scalable in the field of credit risk analytics.

Bibliography

[1] Basel Committee on Banking Supervision. International convergence of capital measurement and capital standards: A revised framework – comprehensive version. https://www.bis.org/publ/bcbs128.htm, 2006.

[2] Basel Committee on Banking Supervision. Sound practices: Implications of fintech developments for banks and bank supervisors. Technical report, Bank for International Settlements, February 2018.

[3] Diogo Carvalho, Eduardo M. Pereira, and Jaime S. Cardoso. Machine learning interpretability: A survey on methods and metrics. *Electronics*, 8(8):832, 2019.

[4] Pete Chapman, Julian Clinton, Randy Kerber, Tom Khabaza, Thomas Reinartz, Colin Shearer, and Rüdiger Wirth. CRISP-DM 1.0: Step-by-step Data Mining Guide. Technical report, CRISP-DM Consortium, 1999.

[5] Keqin Chen, Kun Zhu, Yixin Meng, Amit Yadav, and Asif Khan. Mixed credit scoring model of logistic regression and evidence weight in the background of big data. In Ajith Abraham, Aswani Kumar Cherukuri, Patricia Melin, and Niketa Gandhi, editors, *Intelligent Systems Design and Applications*, volume 940, pages 435–443, Cham, 2020. Springer International Publishing.

[6] FICO. Understanding fico scores. https://www.myfico.com/credit-education-static/doc/education/Understanding_FICO_Scores_5181BK.pdf, 2023. Accessed: July 12, 2025.

[7] Bank for International Settlements (BIS). Basel framework: Credit risk - standardised approach (chapter cre 32). https://www.bis.org/basel_framework/chapter/CRE/32.htm?inforce=20230101&published=20200327, 03 2020. Effective: 2023-01-01, Accessed: 2025-07-12.

[8] David J. Hand and William E. Henley. Statistical classification methods in consumer credit scoring: a review. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 160(3):523–541, 1997.

[9] Amir E. Khandani, Adlar J. Kim, and Andrew W. Lo. Consumer credit-risk models via machine-learning algorithms. *Journal of Banking & Finance*, 34(11):2767–2787, 2010.

[10] Stefan Lessmann, Bart Baesens, Hsin-Vonn Seow, and Lyn C. Thomas. Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. *European Journal of Operational Research*, 247(1):124–136, 2015.

[11] Anahita Namvar, Mohammad Siami, Fethi Rabhi, and Mohsen Naderpour. Credit risk prediction in an imbalanced social lending environment. https://arxiv.org/abs/1805.00801, 2018.

[12] Naeem Siddiqi. *Credit Risk Scorecards: Developing and Implementing Intelligent Credit Scoring*. John Wiley & Sons, 2006.