CREDIT RISK MODELING

MODEL RISK MANAGEMENT AND ENSEMBLE METHODS IN CREDIT RISK MODELING

By SEAN SEXTON, MSc, MA, BA

A Thesis Submitted to the School of Graduate Studies in Partial Fulfillment of the Requirements for the Degree Doctor of Philosophy

McMaster University DOCTOR OF PHILOSOPHY (2022) Hamilton, Ontario (Economics)

TITLE: Model Risk Management and Ensemble Methods in Credit Risk Modeling, AUTHOR: Sean Sexton MSc. (McMaster University), MA (McMaster University), BA (University of Waterloo), SUPERVISOR: Professor Jeffrey S. Racine NUMBER OF PAGES: 155

Abstract

The number of statistical and mathematical credit risk models that financial institutions use and manage due to international and domestic regulatory pressures in recent years has steadily increased. This thesis examines the evolution of model risk management and provides some guidance on how to effectively build and manage different bagging and boosting machine learning techniques for estimating expected credit losses. It examines the pros and cons of these machine learning models and benchmarks them against more conventional models used in practice. It also examines methods for improving their interpretability in order to gain comfort and acceptance from auditors and regulators. To the best of this author's knowledge, there are no academic publications which review, compare, and provide effective model risk management guidance on these machine learning techniques with the purpose of estimating expected credit losses. This thesis is intended for academics, practitioners, auditors, and regulators working in the model risk management and expected credit loss forecasting space.

Preface

This thesis provides an overview of the evolution of Model Risk Management (MRM) in financial institutions to date and also reviews different modeling techniques used to estimate a mortgage portfolio's credit losses. I have had the fortunate opportunity of completing this thesis while working full time at a bank and as a consultant in Toronto and New York City, acquiring applied experience on the aforementioned topics. In what follows, I have drawn both upon the most current academic and regulatory publications, as well as my industry experience to describe how these publications are interpreted and used in practice.

I began my PhD as a full time student with the intent of specializing in econometrics. This inspired me to complete my Masters in Statistics while simultaneously completing my economics comprehensive exams at McMaster University. Upon the completion of my Masters in Statistics at McMaster, I originally accepted a PhD offer at Queens University. But I instead decided to join TD Bank's quantitative modeling department in Toronto to gain applied econometrics/statistics experience.

While transitioning to the private sector, McMaster University agreed that if I stayed to do my PhD in Economics there, they would create a part-time PhD program to accommodate my situation. As such, I decided simultaneously to complete my PhD in Economics at McMaster University part-time under the supervision of Jeffrey Racine, who holds the McMaster Chair in Econometrics. Since I was hired to work in TD Bank's quantitative credit risk modeling department, it seemed practical to pursue a thesis topic that aligned with this subject.

To summarize quickly my approximately 6 years of private sector experience that occurred over the course of writing this thesis, I was promoted to Manager of MRM (bypassing the senior quantitative analyst role which is generally a precursor for the role of manager) after only 5 months as a quantitative analyst at TD. After 18 months at TD, I left the bank to join KPMG Canada's Financial Risk Management consulting team. After 18 months at KPMG Canada, I transferred to KPMG US's Modeling and Analytics Consulting team in New York, New York where I served for 2 years as a manager. In October 2021 I was promoted to my current role as Director of Modeling and Analytics Consulting in New York, New York, where I manage various work streams with over 20 quantitative analysts and consultants. My experience to date has largely focused on MRM, econometrics, and credit risk modeling.

This thesis consists of 3 chapters, where Chapter 1 provides an industry review of the evolution of MRM and evaluates how to effectively account for, monitor, and manage credit risk models in a financial institution. This process is illustrated via an empirical demonstration using Freddie Mac's loan performance data on a portion of its single-family mortgage loans. The content in this chapter draws upon academic literature, regulatory guidance & publications, and my experience consulting on these topics at top tier financial institutions across the US, Canada, Europe, and Asia in the private sector. Given the MRM space is still new terrain that many financial institutions are still learning to navigate, this type of industry review has not been performed and will benefit both practitioners and academics.

Chapters 2 and 3 dive deeper into the econometrics/statistics underpinning these credit risk models, and respectively provides comparative assessments of credit risk Probability of Default (PD) and Loss Given Default (LGD) modeling methodologies for use in financial institutions. A key innovation of this chapter is to demonstrate the utility of boosting and bagging machine learning techniques, both of which have been underutilized in the credit risk modeling space. Readers will be able to implement these new models effectively in their work and will be equipped with the knowledge to improve interpretability to a degree that satisfies auditors and regulators. They will also gain insights into the historical background of financial modeling and thereby understand more deeply the rationales behind them.

Holistically, these chapters offer new insights into the theory and practice of credit risk modeling departments; they trace the historical and institutional factors leading to their evolution; they consider the position they occupy in the institutional ecosystems in which they are embedded; and they provide cutting edge insights into how leaders of these departments can adopt unconventional strategies of machine learning quantitative credit risk modeling to leverage their full potential.

First and foremost, I would like to thank my advisor Jeffrey Racine, who has supported my unconventional journey since day one. I am also grateful for my committee members John Maheu and Seungjin Han for their insightful feedback, comments, and suggestions. I would also like to thank Denis O'Donoughue, who served as a mentor in the early stages of my career and continues to do so. He supported my goal to complete my PhD part-time when many others suggested otherwise. Finally, I want to thank my parents, Margot and Steve Sexton, who have always been a moral compass in my life. They are, of course, directly responsible for any errors or omissions herein.

Contents

List of Figures

List of Tables

Glossary

Table 1: Glossary

Term	Description
AIG	American International Group
AIRB	Advanced Internal Ratings-based Approach
AUC	Area Under the ROC curve
BCBS	Basel Committee on Banking Supervision
BHC	Bank Holding Companies
CCAR	Comprehensive Capital Analysis and Review
CLTV	Combined Loan to Value
CPR	Capital Plan Rule
DPD	Days past Due
DTI	Debt-to-Income
EAD	Exposure at Default
EC	Economic Capital
ECDF	Empirical Distribution Function
ECL	Expected Credit Losses
EL	Expected Loss
IFRS	International Financial Reporting Standard
FRTB	Fundamental Review of the Trading Book
FHFA	Federal Housing Finance Agency
FN	False Negative
FPR	False Positive Rate

Table 1: Glossary *(continued)*

Term	**Description**
GDP	Gross Domestic Product
GSE	Government Sponsored Enterprises
HPI	Housing Price Index
ICAAP	Institution's Capital Adequacy and Internal Review Process
IV	Information Value
KS	Kolmogorov–Smirnov
LASSO	Least Absolute Shrinkage and Selection Operator
LGD	Loss Given Default
LOB	Line of Business
LOD	Line of Defense
LTV	Loan-to-Value
MD	Model Development
MBS	Mortgage Backed Securities
MRM	Model Risk Management
MSA	Metropolitan Statistical Area
MV	Model Validation
OOS	Out-of-Sample
OOT	Out-of-Time
PD	Probability of Default
PRA	Prudential Regulation Authority
RCA	Regulatory Capital Arbitrage

Table 1: Glossary *(continued)*

Term	**Description**
ROC	Receiver Operating Characteristic
SICR	Significant Increase in Credit Risk
SREP	Supervisory Review and Evaluation Process
TP	True Positive
TRIM	Targeted Review of Internal Models
TPR	True Positive Rate
UL	Unexpected Losses
VIF	Variance Inflation Factor
VaR	Value at Risk
WOE	Weight of Evidence

1 Chapter 1: Model Risk Management and Credit Risk Modeling

1.1 Overview

Models are a ubiquitous part of life in the financial sector, and continue to grow in number and complexity. Advances are constantly being made in econometric and statistical theory while a rapidly expanding corpus of rules and regulations governing their use requires vigilance and flexibility from modeling professionals. The tendency of these rules to be vague demands that industry practitioners and institutions interpret them independently and collectively. This results in what is called an "industry standard," or a set of practices that are known to modeling experts but not necessarily to those from outside the industry. For these reasons, the modeling department may appear to be a 'black-box' to those who are not practitioners working in the industry.

In what follows, this chapter provides an industry review of the evolution of Model Risk Management (MRM), and outlines how effectively to account for[1], monitor, and manage a mortgage credit risk model[2] in a financial institution. The collection and curation of regulatory requirements governing the creation of an effective MRM function is still an emerging field that many financial institutions are trying to understand, navigate, and implement. Therefore, this type of detailed industry review and guidance is not available, and will benefit both practitioners and academics who want to understand the emerging MRM space and how it should be applied to a mortgage credit risk model.

[1] The action or process of keeping financial accounts.

[2] This chapter defines a model as a quantitative method, system, or approach that applies statistical, economic, financial, or mathematical theories, techniques, and assumptions to process input data into quantitative estimates (FED 2011).

For demonstrative purposes this chapter empirically examines a Probability of Default (PD) model commonly used in the financial industry using Freddie Mac's loan performance credit data on a portion of fully amortizing fixed-rate mortgages that Freddie Mac purchased or guaranteed from 1999 to 2019. Through this empirical examination, I provide guidance on how financial institutions should choose, develop, review, approve, monitor, and decommission a PD mortgage model. I also identify key risk functions that impact each of the model life cycle stages, such as Model Development (MD), Model Validation (MV), MRM, and Audit (internal and external). PD models are widely used in the financial sector and are a fundamental component in calculating credit portfolio's Expected Credit Losses (ECL), which, when managed properly, helps mitigate systemic risk.

Section 2 overviews key (global) stakeholders, regulatory guidelines, and historically significant events that are important for understanding the current state of MRM. Section 3 provides more granularity on MRM policies around credit risk models reviewing in particular detail the background of global regulatory practices and credit risk modeling methodology used to model ECL for the Advanced Internal Ratings-based Approach (AIRB) outlined by the Basel Committee on Banking Supervision. The AIRB ECL methodology implemented by financial institutions often resembles the methodology used when forecasting ECL for IFRS-9 and CECL (the application and purpose is, of course, quite different). While Section 2 and Section 3 historically contextualizes the current state of MRM, Section 4 more thoroughly defines MRM and discusses current state-of-the-art practices. Section 5 provides an empirical example of a popular mortgage PD model development procedure, while Section 6 assesses the model's performance using objective statistical measures. Section 7 and 8 respectively discuss procedures expected by regulators and auditors for model validation and performance monitoring. Section 9 concludes.

1.2 Historical Background

The banking industry is a fundamental component of the financial system and the economy. It is a popular research topic in areas such as macroeconomics, business cycles, monetary economics, and financial economics. The regulation, supervision and risk management of banks is so important to the global economy that 60 central banks, representing countries from around the world that together account for about 95% of the world's Gross Domestic Product (GDP) actively own and support the Bank for International Settlements (BIS). The BIS is "an international organisation that serves central banks and other financial authorities across the globe to build a greater collective understanding of the world economy, fosters international cooperation among them and supports them in the pursuit of global monetary and financial stability" (Bank for International Settlements 2019). The BIS supports central bank cooperation and provides an independent voice to sound policy making. It acts as a forum for discussion and a platform for cooperation among policymakers. Through this organization, participating central banks have agreed upon the Basel III international regulatory framework for banks, which is an internationally agreed upon set of measures developed in response to the financial crisis of 2007-2009. The Basel III standards are minimum requirements that apply to internationally active banks, where members are committed to implementing and applying standards in their jurisdiction within the time frame established by the BIS Committee.

This section provides an overview of how the global economy manages monetary and financial stability, with particular focus on credit risk models. It identifies key global stakeholders and their roles, accords and regulatory guidelines, and historically significant events that influenced the current state of MRM. These accords and regulations safeguard consumers but restrict how financial intermediaries are able to construct their quantitative analysis. Understanding these various accords and regulations is therefore necessary before proceeding to the notion of “model risk management”, in which mortgage loan default risk plays an important role.

A fundamental component of Basel III is the credit risk framework, which seeks to improve the credibility in the calculation of risk-weighted-assets (RWAs) and facilitate comparability of banks’ capital ratios (BIS 2018). In this framework, banks adopt the advanced internal ratings based (IRB) approach, which often involves the combination of a PD, loss given default (LGD), and an exposure at default (EAD) model to calculate financial instruments’ ECL. Commercial banks’ credit portfolios (e.g. mortgage, credit card, corporate, automobile, and sovereign) pose severe systemic risk, which can be defined as the risk of an event which adversely affects a number of systemically important intermediaries or markets including potentially related infrastructures (Hartmann et al. 2009). The cause of a systemic risk can be either an exogenous (outside the financial system) or endogenous (within the financial system or economy at large) shock. The severity of a systemic risk or event is often measured in terms of macroeconomic variables such as consumption, investment, and growth or economic welfare (Hartmann et al. 2009). This chapter focuses on a mortgage PD model because of the potential systemic impact the housing market has on economies, its relevance in the recent 2007-2009 financial crisis, and its influence on regulatory measures and MRM functions that unfolded afterward. This is discussed in more detail below.

A well known systemic event is the financial crisis of 2007-2009 (commonly referred to as the “Great Recession”), which has motivated significant changes in the banking industry in recent years (Gregory 2015). Because of the Great Recession, skepticism surrounding the ability of credit risk models (such as PD, LGD, and EAD models) to foresee potential risks inherent in credit portfolios has grown. Major banks were taking excessive risks to generate profits for shareholders and employees without holding a sufficient level of capital to survive such a crisis. In response, regulations surrounding large financial institutions have tightened in order both to enhance a firm's ability to survive under a broad range of internal or external stresses, as well as to reduce the impact on the financial system and broader economy in the event of a firm's failure or material weakness (Board of Governors of the Federal Reserve System 2012). Information is considered material if omitting, misstating or obscuring it could reasonably be expected to influence the decisions that the primary users of general purpose financial statements make on the basis of those financial statements, which provide financial information about a specific reporting entity (IASB 2018). This is an important definition, and is a primary consideration for financial auditors.

As financial institutions have become more sophisticated and complex, quantitative models and analytical tools used for decision making purposes have proliferated and become exponentially more critical. Current and upcoming global and domestic regulatory guidelines that banks have to comply with include Basel Advanced IRB (AIRB) (Bank for International Settlements 2019) discussed above, the Dodd-Frank Act stress test (DFAST) (Board of Governors of the Federal Reserve System 2019b) and Comprehensive Capital Analysis and Review (CCAR) (Board of Governors of the Federal Reserve System 2019a) in the US, the International Financial Reporting Standard (IFRS) 9 (IFRS 2018) published by the International Accounting Standards Board (IASB), the Fundamental Review of the Trading Book (FRTB) (Committee on Banking Supervision 2013), Current Expected Credit Losses (CECL, which is the US equivalent of IFRS-9), and IFRS 17 (IFRS 2017). Because these guidelines, collectively, rely heavily on models, a bank's model inventory must expand at a dramatically increased rate over time. Large financial institutions have upwards of thousands of actively used models which periodically require choosing, developing, reviewing, approving and monitoring. What is more, each of them is often a time consuming process requiring skilled, costly labor and pose a systemic economic risk at the aggregate. Quantitative resources (i.e. qualified individuals with statistics, mathematics, economics, computer science, and financial experience/knowledge) are scarce and in high demand, creating concerns for senior management from a cost sustainability perspective. Effectively managing these models is critical for managing systemic risk, hence the ever increasing regulatory focus.

An effective and efficient MRM function can considerably reduce costs, risks, and redundancies of labor efforts, while ensuring that each stage of a model's life cycle abides by governing regulation and internal/domestic/global policies. Globally, regulators have recognized the importance of MRM and have issued supervisory guidelines such as SR Letter 11-7 Supervisory Guidance on MRM (FED 2011) in the US and E-23 Enterprise-Wide MRM for Deposit-Taking Institutions (Office of the Superintendent of Financial Institutions Canada 2017) in Canada. SR 11-7 is often considered a canonical publication which has influenced similar government-published MRM guidelines across the globe. Government enforcement of these guidelines has increased significantly in recent years in response to problems. For example, McKinsey & Company reports one financial institution suffering a loss of several hundred million dollars due to a coding error in a defective risk model (Crespo et al. 2017). Another instance occurred when improper MV and governance for a Value at Risk (VaR) model resulted in losses that ran into the billions. Failing to abide by government model regulations can also result in costly yet avoidable fines (Federal Reserve Board 2011)[3] (Federal Reserve Board 2017b)[4].

Another major motivator for the increased scrutiny around MRM was the great recession credit crisis (2007-2009). Two of the largest contributors to the Great Recession, Fannie Mae and Freddie Mac, held $5.5 trillion in financial obligations, which made up nearly half of all residential mortgage debt outstanding as of June 30th, 2008 (Frame 2015). Both entities are government sponsored enterprises (GSEs) which were established in 1938 and 1970, respectively, to assist individuals in the purchase of a home. They currently provide housing finance for home buyers and renters in the US, while providing liquidity to the single-family market by purchasing and guaranteeing mortgage loans. Between 2005 and 2007, Fannie Mae and Freddie Mac acquired over $1.011 trillion in sub-prime and Alt-A loans (considered riskier than prime but less risky than sub-prime), becoming the largest purchasers of AAA tranches of these sub-prime pools (Peter and Charles 2009).

Some argue that simultaneously being a shareholder-owned company and a GSE resulted in moral hazard. On one hand, the GSEs were motivated by the government mission to support affordable housing through low mortgage interest rates. On the other, the shareholders incentive was to capitalize on their government subsidy and maximize profits (Peter and Charles 2009). With government support, accountability for risk was largely shifted to taxpayers, not to shareholders. This reduced incentives to appropriately assess and take risk. The decline in real estate values seen in 2007 resulted in many borrowers' mortgage values to begin exceeding the value of their

[3] A $3 Million fine levied against BNY Mellon for improperly assigning a lower risk weighting to a portfolio of assets, reducing the firm's risk-based capital ratios.

[4] Deutsche Bank AG was required to pay a combined $156.6 million for unsafe and unsound practices in the FX markets, as well as failure to maintain an adequate Volcker rule compliance program prior to March 30, 2016.

homes, which is considered as a necessary condition for mortgage default. Both agencies witnessed huge losses, effectively leading to bankruptcy and ultimately resulting in the Federal Housing Finance Agency (FHFA) placing both in conservatorship in 2008 (FHFA 2019).

During the Great Recession in the US between May 2007 and October 2009, over 7.5 million people lost their jobs, the unemployment rate rose from 4.4% to 10.1%, long-term unemployment increased sharply, US GDP contracted 3.4%, nearly $11 trillion in household wealth vanished, and 4 million families lost their homes due to foreclosure (Mamonov and Benbunan-Fich 2017) (Grusky, Western, and Wimer 2011). Along with Fannie Mae and Freddie Mac's conservatorship, other financial institutional giants began failing due to the credit crisis. Much of the crisis was due to complex mortgage packages such as mortgage backed securities (MBS), which were now owned by financial institutions that did not originate them. Historical events such as the first run on a bank (Northern Rock, a UK bank) in over a century occurred, which finally resulted in state ownership of Northern Rock in 2008. In March 2008, Bear Stearns was purchased by JP Morgan Chase for merely $2 a share (a historically low price per share), assisted by the Treasury Secretary (Henry Paulson), Fed Chairman (Ben Bernanke), and New York Fed President (Timothy Geithner). Lehman Brothers filed for bankruptcy, Bank of America provided a $50 billion rescue of Merrill Lynch, and American International Group (AIG), considered too big to fail, received $85 billion in exchange for four-fifths ownership from the US government. By 2009, the crisis had spread well beyond the US, resulting in what today is widely recognized as the largest recession since the 1930s Great Depression. In July 2010, the Dodd-Frank Act was signed into law by President Barack Obama, outlining rules and regulations designed to help prevent a repeat of the Great Recession. SR 11-7 was published shortly after in 2011 to provide MRM guidance for financial institutions in the US. Requirements to abide by global regulations have resulted in costly and massive restructuring in banks.

Ensuring that banks' risk models effectively forecast an amount of capital required to withstand different recessionary scenarios has become a huge focal point for financial regulators. Similarly, the importance of accurate financial and accounting estimates/forecasts has also significantly increased (see IFRS-9 and CECL accounting standards for estimating allowances for credit losses). For retail credit risk, this often requires a PD model in conjunction with a LGD model and EAD model. PD, LGD and EAD models are used to estimate ECL under portfolio-specific conditions (often macroeconomic) or long-run averages in order to determine allowances or provisions, regulatory and internal stress testing, risk weights and capital calculations under the AIRB framework, IFRS-9 and CECL accounting requirements. The AIRB approach allows institutions to use their own internal measures for key drivers of credit risk as primary inputs to the capital calculation, subject to meeting certain conditions and subject to explicit supervisory approval (Basel Committee on Banking Supervision 2005).

1.3 Background on Global Regulation and Credit Risk

This section provides more granularity on credit and MRM policies and regulatory requirements. These are canonical policies and requirements that heavily influence how financial institutions build and manage credit risk models, and thus play a large part in managing systemic risk. These globally recognized standards make financial institutions comparable and measurable across participating countries.

Reserve capital at large, complex Bank Holding Companies (BHCs) protects the institution against unexpected losses such as default and also instills stability and effective functioning of a country's financial system (Board of Governors of the Federal Reserve System 2013). The Federal Reserve's Capital Plan Rule (CPR) and CCAR show the increasing attention on large BHCs' abilities to survive in severe economic conditions in the US. The Federal Reserve's Capital Plan Rule requires all top-tier BHCs domiciled in the US with average total consolidated assets of $50 billion or more ($50 billion asset threshold) to develop and maintain a capital plan supported by a robust process for assessing their capital adequacy (Federal Reserve Board 2017a). MRM, a function designed to manage the enterprise-wide model risk, has transitioned from a largely US centric issue (SR 11-7/OCC 2011-12 (FED 2011)) to a key emerging theme across banks in the EU (e.g. CRD IV/CRR (EU-Kommission 2013) (parliament 2013) & Supervisory Review and Evaluation Process (SREP) (European Central Bank 2017), Targeted Review of Internal Models (TRIM) (European Central Bank 2019), and Prudential Regulation Authority (PRA) Stress Test Model Management (Bank of England Prudential Regulation Authority 2018)), Canada (E-23) (Office of the Superintendent of Financial Institutions Canada 2017), and others.

Regulatory capital is broken into two tiers and is comprised of several elements. Tier 1 (core) capital is comprised of common shares and retained earnings. Tier 2 (supplementary) capital is comprised of collective allowance and subordinated debt issuance[5]. Regulatory capital exists so that a bank can absorb sufficient losses through its shareholders' equity rather than through customer deposits or other funding sources. Unlike most companies, banks are in the business of issuing loans to individuals and businesses; as such, if borrowers default on their loans, the bank loses money. Hence, regulatory capital helps control the riskiness of banks and increases the stability of

[5] This is a very high level summary and the details vary by country.

the financial system and the economy as a whole.

To determine a bank's capital adequacy, the amount of regulatory capital a bank holds needs to be compared to its asset base. This is done through minimum capital ratios, which are enforced by a country's regulator, who actively owns, supports, and participates in the BIS. Inability to satisfy minimum ratios can restrict the institution from performing certain activities such as making acquisitions, paying dividends to shareholders, paying bonuses, and buying back shares. The core equity tier 1 ratio is shown below:

$$\text{Core Equity Tier 1 Ratio} = \frac{\text{Core Tier 1 Capital}}{\text{RWA}}, \tag{1}$$

where RWA stands for Risk Weighted Assets. The risk weights assigned to assets are outlined by the Basel Committee on Banking Supervision (BCBS), which is the primary global standard setter for the prudential regulation of banks. The Committee was first established in 1974 by the central bank Governors of the Group of Ten countries in the aftermath of serious disturbances in international currency and banking markets (BCBS 2018). The Committee was created to enhance financial stability by improving the quality of banking supervision worldwide and to serve as a forum for regular cooperation between its member countries on banking supervision. Capital adequacy quickly became the main focus of the Committee's activities. A brief timeline of the Basel Accord is provided in Table 2 below.

There are currently revisions to Basel III, which are referred to in the industry as Basel IV (although Basel IV technically does not exist). We will informally refer to these revisions to Basel III as Basel IV in what follows. The guidance provided in the Basel Accords are global standards; national regulators of financial institutions,

Table 2: A Brief History of The Basel Accord

Date	**Event**	**Description**
July 1988	Basel I	A capital measurement system referred to as *Basel I: The Basel Capital Accord* is established and released to banks. The framework was introduced to virtually all countries with active international banks.
January 1996	Basel I amendments	The Committee issued the amendment to the capital accord to incorporate market risks. This was designed to incorporate a capital requirement for market risks arising from banks' exposures to foreign exchange, traded debt securities, equities, commodities, and options. Another important amendment was that banks could now use internal models (value-at-risk models) to measure their market risk.
June 2004	Basel II	*Basel II: the new capital framework* is released, which is comprised of three pillars: 1. minimum capital requirements, which sought to develop and expand the standardized rules set out in the 1988 Accord, 2. supervisory review of an institution's capital adequacy and internal assessment process, and 3. effective use of disclosure as a lever to strengthen market discipline and encourage sound banking practices.
December 2010	Basel III	Following the 2007-2009 financial crisis, Basel III was agreed upon, which revises and strengthens the three pillars established in Basel II, and extends it to several areas which were to be phased in between 2013-2019. The Committee completed its post-crisis reforms in 2017.

however, are responsible for setting guidelines within their jurisdiction, setting a time frame for implementation, and assessing the compliance of its member institutions against the guidelines.

Issues surrounding the Basel I Accord led to the evolution of Basel II. Some of the criticisms leading to the update were that there was insufficient sensitivity to credit risk, did not account for loan maturity, and had limited granularity of risk weights. Limitations resulted in Regulatory Capital Arbitrage (RCA) through the use of securitization between assets with the same regulatory risk, but different economic risks. This means that banks could artificially inflate the measures of capital appearing in the numerator of regulatory capital ratios or artificially deflate the measures of total risk appearing in the denominator. For further discussion on this issue, see Jones (2000) (Jones 2000). The Basel II Accord Introduced the "3 Pillars" for capital adequacy:

1. Minimum capital requirements.
2. Supervisory review, and
3. Market discipline.

Pillar I put more emphasis on a bank's own internal methodologies - the Foundation or Advanced Internal Ratings-Based (IRB) approach - which led to an increased demand for quantitative analysts in the risk management space. A standardized approach is also an option, which is based on external credit ratings, which applies fixed risk weightings to assets. Pillar II requires banks to have an Institution's Capital Adequacy and Internal Review Process (ICAAP) in relation to their strategies, risk appetite, and actual risk profile. Regulators are expected to review these ICAAP assessments. Finally, Pillar III requires public disclosures by banks to assist users of the information to better understand the risk profile.

While Basel II was certainly an improvement over Basel I, the Great Recession predictably led to further amendments. Basel III was an extension of the existing Basel II framework and introduced new capital and liquidity standards. In particular, it aimed to increase bank liquidity and decrease bank leverage (through more capital and higher quality capital). It also attempted to eliminate/reduce RCA based on the use of credit risk mitigants and minority-owned, non-consolidated subsidiaries. Capital ratios are therefore a common topic during quarterly investor calls with big banks. They help to assess stability, future financial performance, ability to pay dividends, and the likelihood of the need to raise additional capital.

1.3.1 Background on Credit Modeling

This subsection provides an overview of fundamental credit risk modeling, particularly those that calculate expected and unexpected losses. In the retail credit risk modeling environment, calculating estimated Expected Loss (EL) helps determine allowances, regulatory and internal stress testing, and capital calculations to create a capital cushion for covering losses arising from defaulted loans. EL is also applied as part of the risk premium charged to the borrower. The EL of a portfolio is assumed to equal the proportion of obligors that might default within a given time frame (1 year in the Basel context), multiplied by the outstanding exposure at default and once more multiplied by the loss given default rate (i.e. the percentage of exposure that will not be recovered by sale of collateral etc.) (Basel Committee on Banking Supervision 2005). While EL will typically account for the majority of capital held, additional capital is often held to account for things like operational risks, possible emerging risks, reputational risks, future earnings visibility, or any other risks determined through a board, expert judgement, and oversight.

Separate models are built to estimate the PD, LGD and EAD parameters, and the expected loss can be defined as

$$EL = PD \cdot LGD \cdot EAD. \tag{2}$$

For a portfolio consisting of n loans with borrower i, this equation can also be written as

$$\sum_{i=1}^{n} EL_i = \sum_{i=1}^{n} PD_i \cdot LGD_i \cdot EAD_i. \tag{3}$$

Underlying the model is a probability space (Ω, V, P), where Ω is a sample space, V represents the measurable events within the space, and P is a probability measure. For a given account i, this formula essentially calculates the probability of defaulting (PD), multiplies it by the loss the bank would expect to lose (LGD), which is often a percentage between 0-100 (in some cases the losses may be less than 0 or exceed 100), and multiplies this by the exposure the bank could potentially lose (EAD). It should be noted that it is often assumed that the PD, LGD, and EAD in equation 2 and 3 are independent. In reality, this is highly unlikely. However, these are fundamental equations in credit risk modeling. For a study showing the relationship between PD and LGD see E.I. Altman et al. (Altman, Resti, and Sironi 2001). While modeling without assuming independence is possible, for simplicity and to remain consistent with what is commonly implemented in the industry, this chapter will assume independence. The three factors mentioned above correspond to the risk parameters upon which the Basel II IRB approach is built (Basel Committee on Banking Supervision 2005):

- PD per rating grade, which gives the average percentage of obligors that default in this rating grade in the course of one year (in the Basel context).

- EAD, which gives an estimate of the amount outstanding (drawn amounts plus likely future drawdowns of yet undrawn lines) in case the borrower defaults.
- LGD, which gives the percentage of exposure the bank might lose in case the borrower defaults. These losses are usually shown as a percentage of EAD, and depend, among other factors, on the type and amount of collateral as well as the type of borrower and the expected proceeds from the work-out of the assets.

A bank can expect to experience losses above EL which are usually referred to as Unexpected Losses (UL). The costs associated with UL cannot be transferred to the borrowers, because the market will not support prices sufficient to cover all the unexpected losses. As a result, additional capital is needed to cover the risks during peak losses. Banks have the incentive to minimize the capital held so that economic resources can be directed to profitable investments, while holding sufficient capital in order to meet its own debt obligations. The calculation of UL uses the exact same risk parameters as the ones used to calculate EL, i.e., PD, LGD and EAD.

The amount of capital a bank should hold can be determined in a number of ways. Capital is set to ensure that there is a very low, fixed probability of unexpected losses exceeding the probability of bank insolvency. The likelihood that losses will exceed the sum of EL and Unexpected Losses (UL), i.e., the likelihood a bank will not be able to meet its own credit obligations by its profits and capital (Basel Committee on Banking Supervision 2005), is set to maintain a supervisory fixed confidence level under Basel II.

When loans within a portfolio have a low correlation, there is a less likely chance that economic fluctuations will impact the entire portfolio. Banks' capital management will create internal policies to ensure consistency with regulatory requirements and recommendations under Basel III, including their own assessment of capital adequacy as per Pillar 2 of Basel II. The discussion above provides a preliminary overview of the governance, theory, and application of the fundamental PD, LGD and EAD parameter calculations.

The methods used to determine EL on retail portfolios range across products and banks. Since retail portfolios at large BHCs often have a large pool of accessible internal data, external data is infrequently used, and more complex methods can be utilized in comparison to (for example) wholesale portfolios. A common practice is segmentation, which helps capture different distributions that may be observed across accounts. Large BHCs often segment by lien position, risk characteristics such as credit score, loan-to-value (LTV) ratio, collateral, underlying collateral information (Board of Governors of the Federal Reserve System 2013), or delinquency status. Stronger practices are able to capture exposures that react differently to risk drivers under stressed conditions.

1.3.2 Qualitative Factors

Banks are becoming increasingly dependent on quantitative, model-based estimates. However, historical experiences may not fully reflect an entity's expectations about the future and, as such, it is expected that management will make adjustments to historical loss information to better reflect current economic conditions and forecasts qualitatively (FASB 2016). Qualitative adjustments, or "qualitative factors", should always be considered by an entity and should be clearly outlined in a documented qualitative framework/policy. This section is important to include to ensure the reader is aware of the importance of considering and including replicable and reproduceable qualitative frameworks that are used when there are limitations in the quantitative estimates.

The Financial Accounting Standards Board (FASB) in the US listed the following thirteen examples of qualitative adjustments which an entity may consider. They are commonly referred to as the thirteen Q-factors (also see ASC 326-20-55-4). Many FI's in the US use these Q-factors as a reference when building their qualitative frameworks:

1. The borrower's financial condition, credit rating, credit score, asset quality, or business prospects.
2. The borrower's ability to make scheduled interest or principal payments.
3. The remaining payment terms of the financial asset(s).
4. The remaining time to maturity and the timing and extent of prepayments on the financial asset(s).
5. The nature and volume of the entity's financial asset(s).
6. The volume and severity of past due financial asset(s) and the volume and severity of adversely classified or rated financial asset(s).
7. The value of underlying collateral on financial assets in which the collateral-dependent practical expedient has not been utilized.
8. The entity's lending policies and procedures, including changes in lending strategies, underwriting standards, collection, writeoff, and recovery practices, as well as knowledge of the borrower's operations or the borrower's standing in the community.
9. The quality of the entity's credit review system.
10. The experience, ability, and depth of the entity's management, lending staff, and

other relevant personnel.

11. The environmental factors of a borrower and the areas in which the entity's credit is concentrated, such as:

    i.) Regulatory, legal, or technological environment to which the entity has exposure.

    ii.) Changes and expected changes in the general market condition of either the geographical area or the industry to which the entity has exposure.

    iii.) Changes and expected changes in international, national, regional, and local economic and business conditions and developments in which the entity operates, including the condition and expected condition of various market segments.

The Federal Reserve similarly lists 9 "universal" qualitative factors (The Office of the Comptroller of and Currency 2020) (similarly see SR 06-17). In addition to considering the Q-factors above, an entity's qualitative framework should also include model output imprecision considerations, as well as a structured process to calibrate the quantitative model's output based on actual loss experience. These qualitative adjustments might be applied to two elements: qualitative adjustments based on recent and near future observables, and qualitative adjustments based on model imprecision or uncertainty. While these adjustments are qualitative in nature, it is often said that the qualitative adjustments should be quantitatively supported. What this means is that wherever possible, the qualitative adjustment should be defensible from thorough research, statistics, and/or trends from peer banks in comparable markets and/or from proxy datasets. While qualitative adjustments are a fundamental component of credit risk modeling in financial institutions, for the purpose of this chapter, they lie beyond our scope.

1.4 Model Risk Management (MRM)

This section will provide an overview of MRM, which is an emerging and growing field among financial institutions. As mentioned above, the current state of MRM is heavily influenced by the events described in the previous sections. Domestic and global regulators have recently published guidelines which require financial institutions to manage models at an enterprise level. This is discussed in further detail below. Large BHCs often have hundreds to thousands of actively used models, with the number of models and their complexity growing exponentially. The number of models is growing by 10 to 25 percent annually at large institutions according to an article by McKinsey Company (Crespo et al. 2017). This model growth is fueled by the benefits observed from utilizing quantitative methods made available through big data and advanced analytic evolutions, along with pressures from regulatory requirements. While the number of models continues to grow, these modeling tools range in complexity and materiality, from sophisticated statistical techniques to much simpler and less-material analytical methods.[6] Models certainly have the ability to improve business decisions; however, models have the risk of being improperly specified which may produce inaccurate, inconsistent, and/or biased outputs. Even if a model is perfectly specified, models entail the risk of potentially being used incorrectly or inappropriately.

An MRM function within a bank actively accounts for, manages, and assesses the inherent risk associated with all model usage within the institution. A guiding principle for managing model risk is the “effective challenge” of models, that is, critical analysis by objective, informed parties who can identify model limitations and assumptions and

[6] Recall that we consider information to be material if omitting, misstating or obscuring it could reasonably be expected to influence the decisions that the primary users of general purpose financial statements make on the basis of those financial statements, which provide financial information about a specific reporting entity.

produce appropriate changes (FED 2011). As business units within a bank innovate and grow, newly developed tools and models run the risk of going undetected (being used within the bank without appropriate management and risk oversight) for validation and monitoring purposes. As such, a proper level of oversight should ensure that emerging models have a clear owner and are identified, tracked and elevated to model status, or are appropriately decommissioned as characteristics change. Effectively identifying model risk requires an inventory of all models, which is a powerful tool for assessing the risk and required monitoring procedures. A centralized MRM function helps set consistent bank-wide standards and avoids the duplication of research efforts. New and emerging model risk also arises from changes within a Line of Business (LOB), new initiatives, new products, mergers or acquisitions, or regulatory changes. These models are not necessarily always built in-house and may be a vendor model, which is often a ‘black-box’ (i.e., the inner workings of the model/system may be masked by the vendor). Monitoring the activities discussed above is an evolving challenge for a MRM function.

Once models are identified, banks should have an effective system that ensures the levels of independent challenge and oversight correspond to the following key model risk assessment considerations:

i. Materiality.
ii. Model complexity.
iii. The institution’s reliance on the model.
iv. Financial statement, external reporting, and regulatory impact.
v. The amount of model uncertainty.
vi. The model’s operational impact.
vii. Frequency of use.

viii. The amount of manual intervention required.

Assessing the severity of these risk components can help determine the validation activities that should be performed by a bank's Line of Defense (LOD). What constitutes a low risk vs a high risk model can be defined by each business unit based on criteria that are relevant to their LOB and the types of tools used, while remaining in line with a bank's risk appetite. Admittedly, while the term "model complexity" is commonly used in practice, what constitutes a complex model is not strictly defined. Often analytical methods requiring statistical techniques such as regression, machine learning, or financial mathematics are considered complex and will likely fall within an MRM framework. For simplicity, we will refer to models as having low or high risk over a spectrum. In reality, banks may use a more granular risk rating and assessment system to determine the level of oversight and effective challenge required (such as 1-5, 1-10, 1-20, low medium or high). This should always be well documented. Models which have low risk should not be expected to follow the same rigorous validation efforts taken for high risk models. Models with low risk do not pose much of a threat to a bank's well-being during stable and recessionary economic periods, and allocating large amounts of costly resources toward them is often unreasonable and unnecessary. Determining these thresholds can be a difficult task and requires a combination of both quantitative and qualitative expert judgement. While independently a model may pose no threat, risk models collectively may signal risk if they show similar variable sensitivities, rely on common assumptions, data or methodologies, or any other factors that could affect several models adversely. Regulators expect a well documented, transparent and objective risk rating methodology.

The labor within a MV division is costly. Low risk models may pose a limited threat to a bank's financial stability and, hence, should require minimal to no oversight by

MV. Figure 1 shows a simplified visualization of a potential model's materiality and complexity assessment. Models falling in category **A** are very low risk and do not require the full validation procedures typically performed by a MV risk function. Models in category **B** are deemed risky enough that they require MV review and oversight. The threshold is ideally determined by the LOB and confirmed with MV/MRM, which will acknowledge that the level of model risk is low enough to require minimal involvement. Models falling in category **A** still require testing activities performed by the LOB to confirm the model is performing adequately and appropriately and is in line with internal expectations and external regulatory requirements.

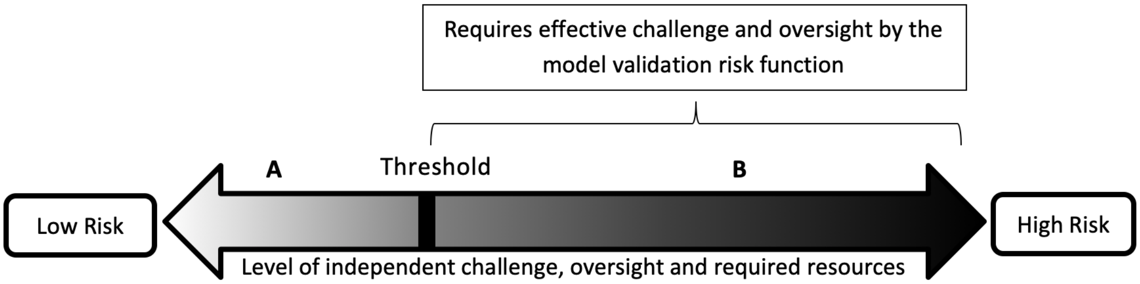
Requires effective challenge and oversight by the model validation risk function
A
Threshold
B
Low Risk
High Risk
Level of independent challenge, oversight and required resources

Figure 1: Level of Model Risk Management Oversight and Effective Challenge thresholds

The following subsection will discuss in further detail a financial institution's LOD, which is the most common benchmark for assigning control and risk management responsibilities to business functions in an organization (Arndorfer and Minto 2015).

1.4.1 Lines of Defence

Coordinating MRM responsibilities across financial departments and divisions is a challenging task that requires a clear definition of responsibilities, authority, independence, resources, and access to the board of directors. In most financial institutions, the "three lines of defense" model has been used to model the interaction between corporate governance and internal control systems (Arndorfer and Minto 2015). The three LOD defines three groups and their respective involvement in effective risk management (The Institute of Internal Auditors (IIA) 2013). A summary of the three LOD with a focus on modeling is outlined below.

1.4.1.1 First Line of Defence: Functions That Own and Manage Risk

The first LOD identifies, manages, and controls the ongoing operations of model risks. In theory, those who are responsible for the basic control and risk from using a model should be held accountable.

The three LOD model assumes individuals in the first line are likely most familiar with the intricacies, inherent risks, and overall modeling environment, and are therefore best suited to determine the design, theory, and logic underlying the model. They are best suited to ensure the model is conceptually sound and functions appropriately. As active users of the model, they are potentially best suited quickly to detect any inherent weaknesses in the model, and are able to escalate and remedy any issues. The Great Recession suggests that there was insufficient awareness of risks and control procedures by risk-taking units, so this approach has been encouraged in recent practice (Arndorfer and Minto 2015).

The first line should articulate a sound understanding of all limitations and conditions and actively monitor their models on an ongoing basis to demonstrate adequate model performance. This is often done through rigorous documentation, including statistical and mathematical tests, simulations, and expert judgement.

1.4.1.2 Second Line of Defence: Functions That Oversee Risks

The second LOD has seen considerable growth in response to tighter regulatory requirements and more complex products and methodologies. This line is responsible for setting standards, providing independent review, oversight, and effective challenge to the first line. These functions include MRM and MV. MV (discussed in further detail below) effectively challenges, oversees, and analyzes new and existing models objectively and critically. Successful approval of a model results in continued use by the line of business (first LOD). Thorough documentation is critical in demonstrating an adequate understanding and description of the processes undertaken. MRM is discussed above, and oversees the standards and guidelines for all three lines of defense.

1.4.1.3 Third Line of Defence: Functions That Provide Independent Assurance

The modeling internal audit team is the third LOD and provides a final independent oversight and challenge of the bank's model risk function as a whole. The third line validates the effectiveness of the first and second line. A visual summary outlined by the Institute of Internal Auditors is displayed in Figure 2. In practice, the audit function has to conduct at least once annually a risk assessment of the organisation and identify business units or processes that exhibit a high level of residual risk (Arndorfer and Minto 2015).

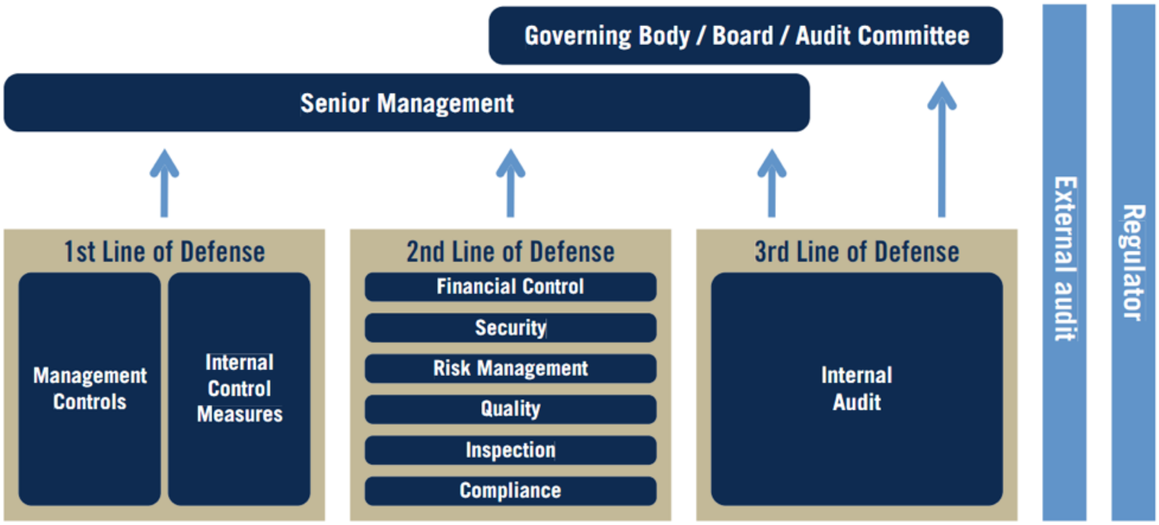

Figure 2: The Three Lines of Defense Model in Effective Risk Management and Control

A four LOD model/concept is also sometimes used (Arndorfer and Minto 2015), where external auditors and banking supervisors who provide a final level of assurance on the governance and internal financial processes are considered the fourth line.

1.4.1.4 The Model Life Cycle

This subsection summarizes the typical 'life' of a new model, where each material model follows a "model life cycle". The life cycle is "a perpetual activity that is continually refurbished and updated as the model evolves with the passage of time" (Office of the Superintendent of Financial Institutions Canada 2017), involving a process such as the following (Office of the Superintendent of Financial Institutions Canada 2017):

- The rational for modeling.
- Model development.
- Independent review (vetting, or sometimes referred to as an initial validation).
- Model approval.

- Ongoing monitoring and review (validation), and
- Modification/decommission of the model.

The frequency of ongoing monitoring and review is dependent on how ‘risky’ the model is, along the lines discussed above. While all models follow a model life cycle, models falling in the A category in Figure 1 above should follow a less rigorous process than those in category B. Models in category A potentially may not require as rigorous vetting, ongoing monitoring, and review requirements, with more reliance on the LOB and/or the risk management function within the LOB.

Quantitative methods in many cases allow for objective decision making, which previously may have relied on expert judgement, e.g., decision making based on work experience. Expert judgement is still extremely important due to the fact that modeling by nature is a simplification of reality. It helps account for insufficient and imperfect data, which may lead to improper representations of true population distributions and characteristics.

The following section presents an empirical example of the model development process.

1.5 Model Development

The sections above provides the reader with a foundational understanding of the history and evolution of MRM, with particular focus on its application to credit risk models. This section provides an empirical example of a mortgage PD credit risk model, which will give the reader a general idea of what procedures are taken in a credit risk model development team within a FI. To provide the reader with a comprehensive idea of the model life cycle as a whole, the other stages of the model life cycle are also discussed and examined. Even practitioners within these FIs’ risk departments often do not have a full understanding of all model life cycle stages, or for that matter, of these other departments, as risk departments are often siloed.

The level of risk inherent in the model development process is largely influenced by the developer's experience, education, training, work environment, and awareness of existing and emerging risks/trends. Further, financial modeling is often an activity that requires expert knowledge in multiple fields such as economics, finance, statistics, mathematics, and computer science. To help manage these risks, the purpose for the development of a model should be clear, and the suitability of the data, methodology, required programming, expected marginal improvements to the business, and level of additional oversight and monitoring should be taken into consideration. While these guiding principles are theoretically useful at a high level, it is often helpful to see an empirical example for demonstrative purposes. This section focuses on the development of a PD mortgage credit risk model using Freddie Mac's loan performance credit data on a portion of fully amortizing fixed-rate mortgages the company purchased or guaranteed from 1999 to 2019. The methodology is one commonly used by practitioners in credit risk modeling.

Retail portfolios are often modeled either at the loan level or the portfolio/segment level. A loan-level model has the benefit of providing information at the account level and is generally the preferred method in the industry when data is available. A portfolio-level model averages out the noise observed at the loan level; however, a loan-level model theoretically should outperform a portfolio-level model simply because more information is available. Smaller financial institutions with less systematic risk on a domestic or global economy will be under less scrupulous regulatory requirements, at which point portfolio-level modeling may be more common. The following subsections will discuss the following model building stages in detail:

1. Dataset background and relevance.
2. Model specification.

1.5.1 Dataset Background and Relevance

Freddie Mac is a GSE established in 1970 by Congress as a private company to help ensure an affordable supply of mortgage funds throughout the US. Today this is done by purchasing mortgage loans from lenders so that they are able to continue providing loans to qualified borrowers. Freddie Mac's publicly available Single-Family Fixed Rate Mortgage Loan Performance dataset ("the data") contains a subset of Fannie Mae's 30-year (or less), fully amortizing, full documentation, single-family, conventional fixed-rate mortgages. Excluded from the data are adjustable-rate mortgage loans, balloon mortgage loans, initial interest, step rates, government-insured mortgage loans, mortgages delivered under alternative agreements, mortgages associated with mortgage revenue bonds purchased by Freddie Mac, and mortgages delivered to Freddie Mac with credit enhancements other than primary mortgage insurance, with the exception of certain lender-negotiated credit enhancements (Freddie Mac 2020). This dataset is widely used and referenced among credit risk modeling practitioners in the industry, as it was made public by Freddie Mac for the purpose of helping practitioners build more accurate credit performance models in support of ongoing risk-sharing initiatives.

Two datasets are combined to build the full dataset used below: one containing acquisition data and another that monitors the performance data on a monthly basis. The full dataset contains mortgages originated between January 1st, 1999 and June 30th, 2019. The dataset is updated every quarter to include newly acquired mortgage loans and any updates observed in performance. For the purpose of this empirical demonstration, 6,000 loans from each origination quarter are sampled, and each loan's performance is tracked across time. Federal Housing Finance Agency (FHSA) Housing Price Index (HPI) data at the Three-Digit ZIP code level (Federal Housing Finance Agency 2020) are used, as well as unemployment data from the Bureau of Labor Statistics (Bureau of Labor Statistics 2020). The data is then split into an in-sample training dataset, an out-of-sample (OOS) test dataset, and an out-of-time (OOT) test dataset. OOS refers to a test dataset with the same date ranges as the training dataset, while OOT refers to a test dataset with date ranges outside of the training dataset. It is common to sample randomly 70% of the data for training the model and use 30% for validating the model on the OOS test dataset. A hold-out period of one year or greater is often used for the OOT test dataset and will be used for the empirical portion of this chapter.

1.5.2 Model Specification

By way of illustration, common PD modeling approaches used in the industry are linear regression analyses on default risk, logistic regression, optimization models of default, vintage loss models, cohort loss models, decision trees, state transition models and hazard models. For demonstrative purposes, this chapter looks at a variant of the most commonly used logistic regression method, built on survival data, where the dependent variable is a binary outcome; default (1) or non-default (0). To remain consistent with Freddie Mac's dataset, default is defined to have occurred when a mortgage's balance was reduced for the following reasons:

- Third party sale.
- Short sale or charge off.
- Repurchase prior to property disposition.
- Real Estate Owned (REO) disposition.

- Note sale / reperforming sale.

The variable of interest is whether or not an account defaults within a one-year (12 month) window or does not. This is a common AIRB prediction window. A one-year forward-looking default prediction is also a requirement for accounts that have not experienced a Significant Increase in Credit Risk (SICR) in the IFRS-9 accounting standard. For loans that have experienced a SICR since origination, under IFRS-9, expected credit losses must be estimated over the lifetime of the loan. Under US CECL guidance, credit losses over the lifetime of all loans must be estimated. For the logistic model, let Y_i represent the binary (Bernoulli) default dependent variable, where $Y_i = 1$ equals default (within 12 months), and $Y_i = 0$ equals no default. Using maximum likelihood, the parameters in the following equation can be estimated:

$$\ln\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + ... + \beta_m x_{m,i} + \epsilon, \tag{4}$$

where $x_i, x_{2,i}, ..., x_{m,i}$ represent m explanatory variables, and $p_i = E[Y_i|x_i, x_{2,i}, ..., x_{m,i}]$ is the expected value of Y_i given $x_i, x_{2,i}, ..., x_{m,i}$. The intercept β_0 and $\beta_1, \beta_2, ..., \beta_m$ are the respective logistic model coefficients.

For this particular use case, a Weight of Evidence (WOE) method is used, where each variable is first discretized up to a maximum of 15 bins, and then the transformed WOE values are used in place of the raw values for all cases in the bin range. This is a common transformation used in the credit industry and in scorecard models and is generally approved by validators and auditors (both internal and external). More complexity may be expected for more material portfolios. Both the dependent logit function and WOEs are in fact the log of probability odds. The WOE is calculated by taking the log of the proportion of 'goods' (non-default) in an attribute divided by the proportion that are bad (default) in an attribute. The WOE of the jth attribute of the ith characteristic is given by

$$w_{ij} = \ln\left(\frac{p_{ij}}{q_{ij}}\right), \tag{5}$$

where p_{ij} is the proportion of those classified good in attribute j of characteristic i and q_{ij} is the proportion of those classified bad in attribute j of characteristic i (Henley and Hand 1996).

An important step in the model development process is variable selection. Variable selection is the process of selecting the best set of predictors, which removes unnecessary noise, collinearity, overfitting, computation cost and improves interpretability. It is common to try to keep the number of independent variables to an interpretable amount (between 5-15 for example). A common variable selection technique leveraging the WOE is to use the Information Value (IV):

$$\text{IV} = \sum_{i,j}(p_{ij} - q_{ij}) \cdot w_{ij}, \tag{6}$$

where w_{ij} is defined above. The information value demonstrates a variable's ability to discriminate between defaults and non-defaults. There are no universally accepted thresholds; in practice, however, IV values below 0.02 are considered not useful for prediction, values between 0.3 - 0.5 are considered strong predictors, and values above 0.5 are considered very strong. A visual display of the binning for the original LTV variable is shown in Figure 3. The IV value for LTV is 0.27. A similar graph can be found for each variable in the final model in the appendix.

Other common variable selection techniques are forward and/or backward stepwise

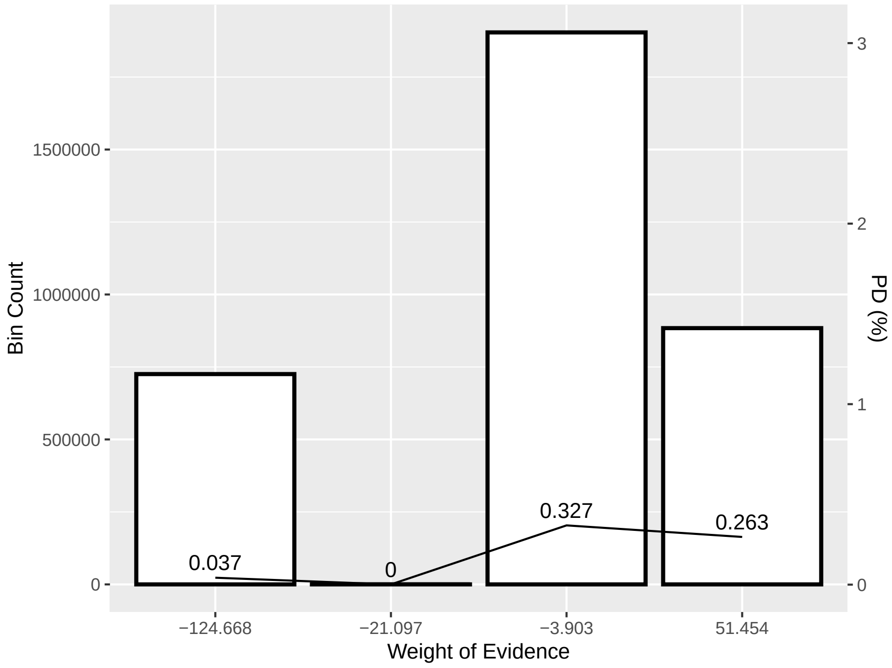

Figure 3: Original Combined Loan-To-Value Weight of Evidence Bin Count and Default Probabilities

regressions, decision trees, Least Absolute Shrinkage and Selection Operator (LASSO), Single Variable Analysis (SVA), and expert judgement. The variable selection process should be defendable and replicable. The dataset begins with 57 independent variables. Variables including future information are removed (e.g., variables indicating whether or not the loan will default in the future) or those that would not be available during production, and a few others which are intuitively not reasonable independent variables, which results in 44 variables. The two statistical variable selection techniques used are backward stepwise regression and IV. The backward variable selection technique begins with all the potential candidate variables. The algorithm selects a model based on the Akaike Information Criterion (AIC) by removing one variable at each step. Variables that are statistically most significant in both variable selection methods and are also in line with expert judgement and business sense are chosen for the final independent variable set.

The final variables selected are summarized in Table 3 below. Descriptions are taken directly from Freddie Mac's formal descriptions.

To check for collinearity amongst independent variables, the Variance Inflation Factor (VIF) is calculated for each variable, shown in Table 4. The VIF checks how much the variance of an estimated regression coefficient increases if your independent variables are correlated. While there is no universally accepted threshold for what constitutes a high VIF, values below 5 generally indicate that there appears to be no indication of multicollinearity. Sensitivity analysis, or 'stress tests', should also be performed, as these are useful methods for assessing a financial institutions health under different financial and macroeconomic conditions. As stated by the federal reserve, "They provide a systematic, disciplined framework for assessing whether firms have adequate capital to absorb losses and continue to fulfill their roles as financial intermediaries under various economic scenarios" (Federal Reserve Board 2020).

1.6 Model Performance

This section evaluates the model's performance using common statistical methods. First, the Population Stability Index (PSI) test is performed on the OOS and OOT dataset (or a "Characteristic Stability Index" when evaluating one variable). The PSI provides a statistical method for determining whether there is a significant difference in two variables' distributions. It is a common test performed by an internal MV team (second LOD), internal audit (third LOD), or an external auditor (fourth LOD) to assess the appropriateness of a training dataset's ability to capture underlying characteristics of an OOS dataset or its continued performance in an OOT dataset. If the PSI number is high, it may indicate that the training dataset is not representative of the OOS or OOT dataset and may require further investigation, since the data used

Table 3: Final Model Independent Variables

Variable	Description
Delinquency Indicator	A variable indicating whether the borrower is currently delinquent or not.
HPI Change	This is calculated as the origination HPI/current HPI. This is done for each loan's 3 digit zip code.
Occupancy Status	Denotes whether the mortgage type is owner occupied, second home, or an investment property.
Interest Rate	The original note rate as indicated on the mortgage note.
Original Combined Loan to Value (LTV) Ratio	The ratio is obtained by dividing the original mortgage loan amount on the note date plus any secondary mortgage loan amount disclosed by the Seller by the lesser of the mortgaged property's appraised value on the note date or its purchase price.
Number of Borrowers	The number of Borrower(s) who are obligated to repay the mortgage note secured by the mortgaged property.
Credit Score (FICO)	A number, prepared by third parties, summarizing the Borrower's creditworthiness, which may be indicative of the likelihood that the Borrower will timely repay future obligations. Generally, the credit score disclosed is the score known at the time of acquisition and is the score used to originate the mortgage.
Property Type	Denotes whether the property type secured by the mortgage is a condominium, leasehold, planned unit development (PUD), cooperative share, manufactured home, or Single Family home.
Loan Age	The number of months since the note origination month of the mortgage.
Debt to Income (DTI) Ratio	Disclosure of the debt to income ratio is based on (1) the sum of the Borrower's monthly debt payments, including monthly housing expenses that incorporate the mortgage payment the Borrower is making at the time of the delivery of the mortgage loan to Freddie Mac, divided by (2) the total monthly income used to underwrite the loan as of the date of the origination of the such loan.
Original Value	The value of the house, calculated by Original UPB / LTV / 100.
Unemployment Rate	The unemployment rate by state is used, lagged by 3 months.

Table 4: Variance Inflation Factor Test for Multicollinearity

	VIF
Delinquency Indicator	1.282
HPI Change	1.536
Occupancy Status	1.049
Interest Rate	1.138
Original Combined LTV	1.071
Number of Borrowers	1.053
Credit Score (FICO)	1.213
Property Type	1.035
Loan Age	1.202
DTI	1.051
Original Value	1.216
UR (3 month lag)	1.380

to develop the model may be different than the data to which the model is applied to. The PSI is defined as follows (Yurdakul 2018). Let N be the sample size for the training (or expected) dataset, and M be the sample size for the OOS and OOT (or actual) dataset. Then we can express the PSI as follows:

$$\begin{aligned} PSI &= \sum_{i=1}^{B} \left(\frac{n_i}{N} - \frac{m_i}{M} \right) \cdot \left(\ln\left[\frac{n_i}{N}\right] - \ln\left[\frac{m_i}{M}\right] \right) \\ &= \sum_{i=1}^{B} (\hat{p}_i - \hat{q}_i) \cdot (\ln(\hat{p}_i) - \ln(\hat{q}_i)), \end{aligned} \tag{7}$$

where n_i and m_i are counts in the i^{th} bin, $\sum n_i = N$, $\sum m_i = M$, $\hat{p}_i = n_i/N$, and $\hat{q}_i = m_i/M$. While there is no universally accepted threshold for what PSI value indicates a significant difference/shift in the distribution of two datasets, industry standard practice typically considers a PSI less than 0.10 to indicate that a small shift could have occurred and, hence, that the training sample is appropriately representative of the testing dataset. A PSI between 0.10 and 0.25 indicates a moderate shift, which will likely motivate further investigation for what may be causing the change in distribution. Finally, a PSI value greater than 0.25 means a significant shift and notion of overlap of distribution, which is often motivation for the model developer to acquire a more relevant/recent training sample if possible. Note that it is necessary to check and understand how the distributions between the training and test datasets differ, because not all changes in distribution necessarily mean the model will perform poorly on the test dataset (e.g., if the new data is now more heavily concentrated around the training dataset mean or if no extrapolation along the x-domain is occurring). For more discussion on this see (Taplin and Hunt 2019).

Given that most datasets have more than one variable (or “Key Risk Drivers”), it is natural to ask which variables’ change in distribution will contribute most to a potential deterioration in model performance. To be comprehensive, PSI is performed on each variable in the model. In Figure 4, density plots for the interest rate variable for the OOS and OOT datasets, respectively, are compared to the interest rate variable density plot in the training dataset. The PSI value is also included in the graph.

As shown, the PSI for the OOS dataset interest rate is 0.03. This implies that there has not been a significant difference/shift in the distribution. However, the OOT dataset PSI is 1.56, indicating that there has been a significant change in distribution. While there has been a change in distribution, the data is now concentrated more around lower rates, which the model was still trained on. Given there has not been a shift in distribution to values that the model was not trained on (for example, extrapolation to higher rates previously unseen), there is no reason to believe there is a significant deterioration in model performance in this instance. In this example, the model and interest rate variable would be monitored closely in subsequent quarters to determine

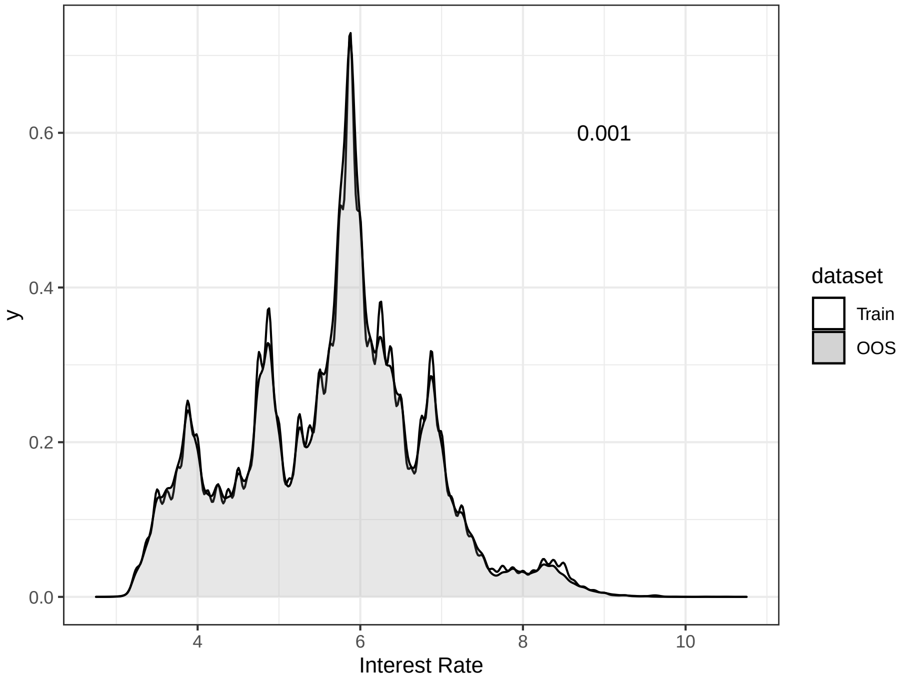

Figure 4: Interest Rate Training and Out-of-Sample Data Density Plots and Population Stability Index Test Result

whether any persistent model deterioration is occurring.

The model performance will be investigated next on the OOS and OOT datasets using the Receiver Operating Characteristic (ROC) curve, the Kolmogorov–Smirnov (KS) test, and the Gini coefficient (other common monitoring approaches might use internally determined performance threshold breaches). To properly describe these tests, some preliminary definitions are first provided. For the problem at hand, the model is built to predict either default (1) or non-default (0), where Y denotes the outcome, and $\hat{Y}$ denotes the prediction. Given the model estimate, there are four potential outcomes:

i. True Positive (TP): The model correctly predicted the loan would default ($Y = 1, \hat{Y} = 1$).

ii. False Positive (FP): The model predicted the loan would default, but the loan

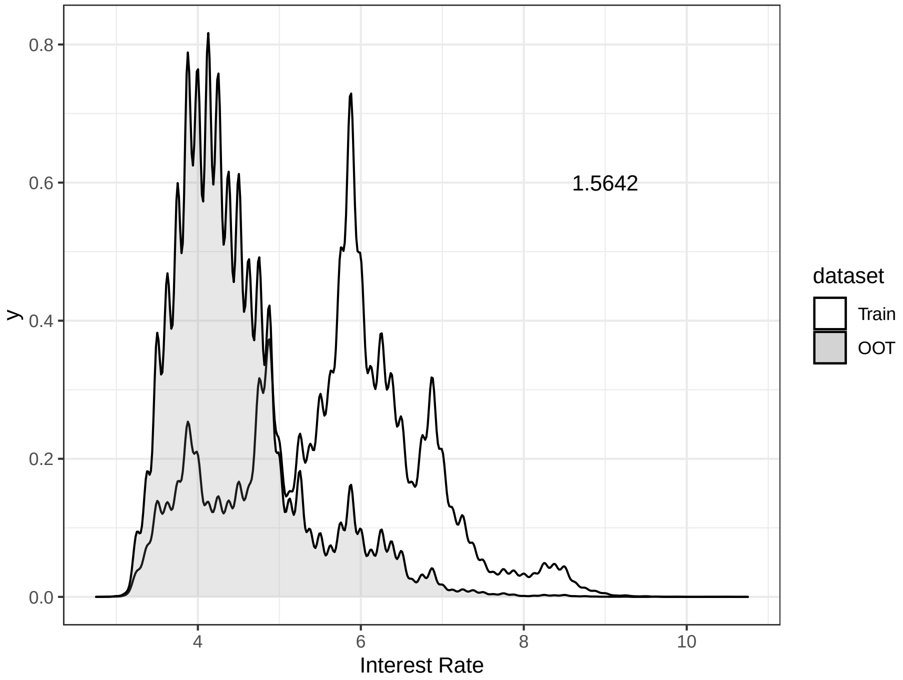

Figure 5: Interest Rate Training and Out-of-Time Density Data Plots and Population Stability Index Test Result

remained performing ($Y = 0, \hat{Y} = 1$).

iii. True Negative (TN): The model correctly predicted the loan would remain performing ($Y = 0, \hat{Y} = 0$).

iv. False Negative (FN): The model predicted the loan would remain performing, but the loan defaulted ($Y = 1, \hat{Y} = 0$).

These four outcomes are used to calculate both the True Positive Rate (TPR) and the False Negative Rate (FPR):

$$TPR = \frac{TP}{TP + FN} \tag{8}$$

$$FPR = \frac{FP}{FP + TN}. \tag{9}$$

The logistic regression model, however, does not immediately provide a binary prediction. For the predictor $\hat{\Pr}(Y = 1|X)$ and a classification threshold c, there is either $\hat{\Pr}(Y = 1|X) \leq c$ (i.e. non-default) or $\hat{\Pr}(Y = 1|X) > c$ (default).

Using the $TPR(c)$ and $FPR(c)$, the ROC curve can be calculated. The ROC curve is a monotone increasing function that helps determine the discrimination ability of a model by plotting the TPR against the FPR at various logistic regression classification thresholds c. By considering any possible cut-off c, the ROC curve can be written as (Calì and Longobardi 2015):

$$ROC(\cdot) = \{FPR(c),\ TPR(c),\ c \quad \in [0, 1]\}. \tag{10}$$

The TPR and (1-TRN) are commonly referred to as the sensitivity and specificity, respectively. The discrimination ability of a model is often determined based on the area under the ROC curve (AUC), which is a value ranging from 0 to 1. An AUC equal to 1 indicates the model perfectly predicts default and non-defaulted loans, depicted by line A in Figure 6. Line C in Figure 6 represents an AUC equal to 0.5, which means the model does no better than chance and is considered uninformative.

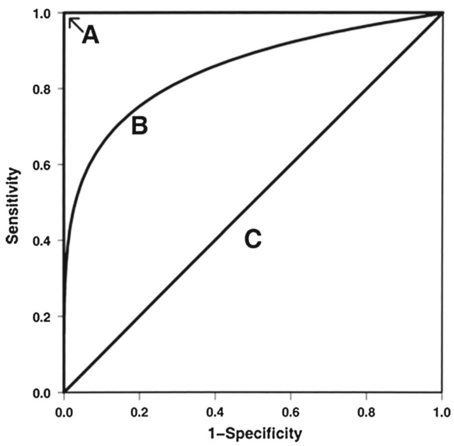

Figure 6: Three Hypothetical (perfect = A, uninformative = C) Receiver Operating Characteristic Curves

As test accuracy improves, the ROC curve will move toward line A, and AUC will move closer to 1. The AUC can be defined as

$$AUC = \int_0^1 ROC(x)dx. \tag{11}$$

The second model performance test we will be using is the Kolmogorov–Smirnov (KS) test. The KS test statistic ranges from 0% to 100% and is able to determine how different the TPR and FPR distributions are, where larger KS statistics imply the model has a stronger discrimination ability. Measuring the maximum vertical separation between two cumulative distributions TPR and FPR gives the KS statistic:

$$KS = \max(TPR - FPR). \tag{12}$$

The final performance statistic reported is the Gini test, also known as the Accuracy Ratio (AR). It is a single number that represents the area under the cumulative lift curve relative to the area under a uniform distribution. The lift curve measures the effectiveness of the models predictions by calculating the percentage of correctly estimated defaults, relative to a baseline ‘random’ diagonal line (the uniform distribution). The AR is computed by sorting scores from a distribution from low to high, determining the corresponding cumulative lift (Lorenz curve), or the Cumulative Accuracy Profile (CAP) curve, computing the area between the cumulative uniform distribution and the Lorenz curve on the interval [0, 1], and dividing the result by the area under the cumulative uniform distribution on the same interval (Greene and Milne 2010). An interesting relationship important to note is that the AR test is linearly related to the AUC through the following equation:

$$AR = 2 \times AUC - 1. \tag{13}$$

While reporting both the AUC and AR may be redundant, it is still reported separately for the benefit of the reader given they are both very common performance metrics. The ROC curve, KS test, and AR test for both the OOS and OOT dataset are shown below.

Figure 7 for both OOS and OOT show the empirical cumulative distribution function (ECDF) of the model estimates for the actual non-defaulted accounts ($Y = 0$) and the defaulted accounts ($Y = 1$), and the KS test is visualized by the black dotted vertical line, representing the farthest distance between the two ECDFs.

Figure 8 plots the sensitivity (true positive rate) and specificity (true negative rate) and calculates the area underneath this curve (the AUC).

Figure 9 plots the Cumulative Accuracy Profile (CAP) and its respective AR. Given the small OOT sample size, the KS and CAP curves are less smooth than the respective OOS tests. For this particular dataset and classification exercise, the model

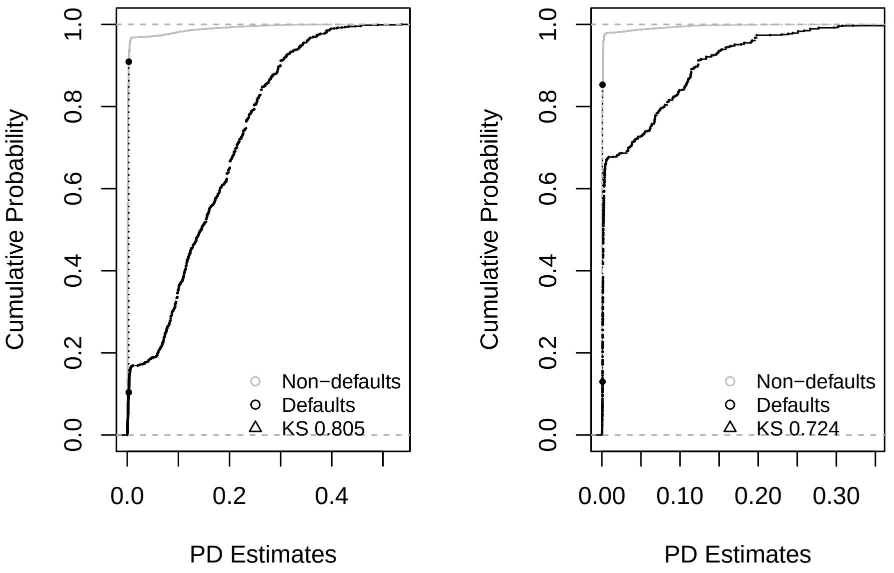

Figure 7: Kolmogorov–Smirnov Model Performance Test on Out-of-Sample (left) and Out-of-Time (right) Dataset

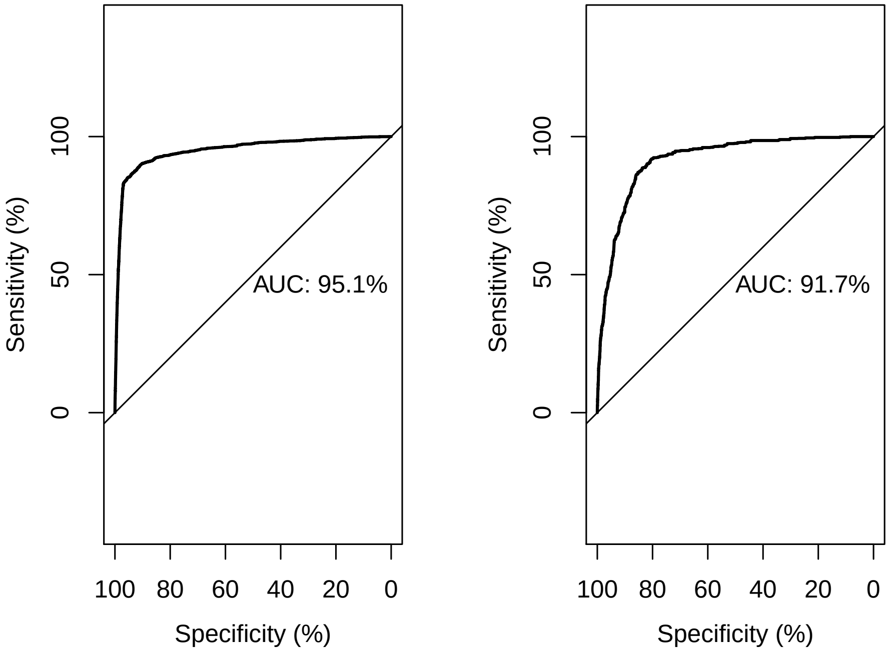

Figure 8: Receiver Operating Characteristic and Respective Area Under the Curve Model Performance Test on Out-of-Sample (left) and Out-of-Time (right) Dataset

is performing well both OOS and OOT.

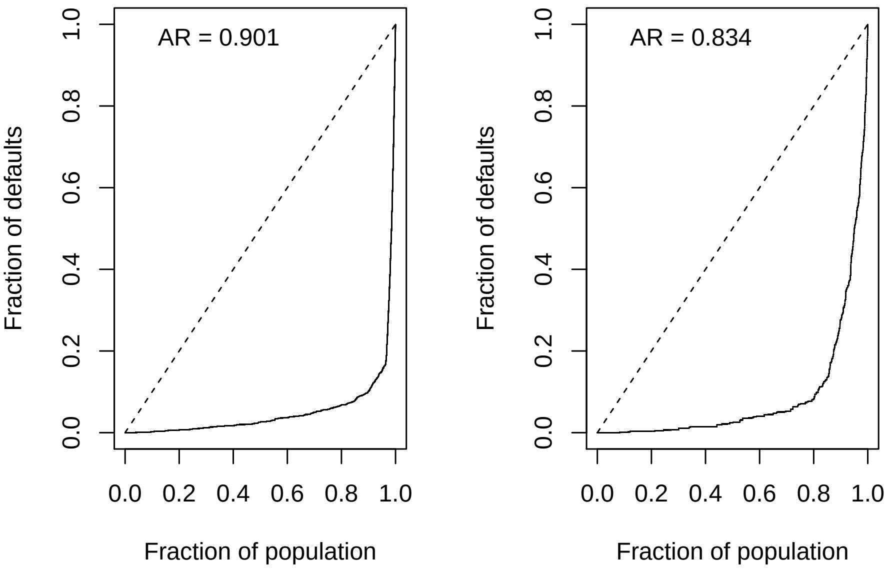

Figure 9: Cumulative Accuracy Profile Curve and Respective Accuracy Ratio Model Performance Test on Out-of-Sample (left) and Out-of-Time (right) Dataset

1.7 Model Validation

Once the model has been developed and the procedures undertaken are well documented, the next stage in the model life cycle is model validation. As such, this section provides a brief overview of what an internal/external auditor and/or regulator would expect to see from a financial institution's model validation risk function. Once model development is complete, an initial model validation should be conducted where the actions taken by the model validation team are thoroughly documented in a model validation report. This is an independent verification process, which should include back-testing, discriminatory power testing, and sensitivity testing. Model validation findings should include detailed recommendations from the validation team to the model owner/developer(s), and it should be clear what the severity level of each finding is. Ultimately, the model validation report should state the final conclusion as to whether the model is appropriate for its intended use. The model validation report should identify clearly the tests performed (or re-performed) by the model validation team and those performed by the model developer. Once a model has passed an initial model validation, there should be a periodic review (at least annually performed by individuals independent of the model development process) which evaluates the model's performance and key model assumptions to determine whether the model is still performing well, whether it remains appropriate for its intended use, or whether another full model validation is required prior to the next scheduled model validation date. These ongoing validation activities help identify any changes in markets, products, exposures, activities, clients, or business practices that might create model limitations.

There are sometimes situations, sometimes referred to as 'trigerring events', which result in a redevelopment or recalibration of a model. This could result either in another full model validation, or in some cases a targeted (limited-scope) validation outside the regularly scheduled validation periods. So as to help reduce the amount of labor required for a full model validation, a targeted validation will only look at certain aspects of the model. Changes in the model scope; data population; model inputs, theory, or code; or changes to the economic environment are all examples of situations which might warrant a re-validation (targeted or full).

As stated in SR Letter 11-7, "the rigor and sophistication of validation should be commensurate with the bank's overall use of models, the complexity and materiality of its models, and the size and complexity of the bank's operations". This statement alone is admittedly somewhat vague; however, a bank's peers/competitors, auditors, external consultants, and other relevant stakeholders help guide the appropriate level of rigor and sophistication. SR Letter 11-7 identifies three core elements of an effective validation framework:

- Evaluation of conceptual soundness, including developmental evidence, which involves assessing the quality of the model design and construction.

- Ongoing monitoring, including process verification and benchmarking. Once the model is approved, this becomes a joint responsibility of model users, owners, and validators.
- Outcomes analysis, including back-testing. This is a comparison of the model predicted values to actual values. A variety of quantitative and qualitative tests and analytical techniques can be used to measure performance. Clear model performance thresholds, performance breaches, and methods for identifying model deterioration should be documented. In the event that model outcomes persistently fall outside the bank's thresholds, model adjustments, recalibration, or redevelopment may be needed.

1.8 Performance Monitoring

Once the model has passed the model validation process, it can go into production. This section provides a brief overview of what an internal/external auditor and/or regulator can expect to see from a financial institution's model performance monitoring function. It is expected that the model owner/developer perform regularly scheduled performance monitoring, which generally evaluates whether the model is performing well, as intended, with a frequency appropriate with the nature of the model (i.e., how often is the model used, how often is new data available, what is the risk of the model, etc.). The performance monitoring tests and their respective thresholds for acceptable levels of error (through analysis of the distribution outcomes around expected or predicted values (FED 2011)) should be outlined in a model monitoring framework/policy; typically, it includes back-testing, methods for detecting distributional shifts in new data, forecast accuracy, and coefficient stability. There are periods in time when a model may be performing poorly due to some explainable exogenous or endogenous shock/force. As such, the models performance should be evaluated holistically, taking into consideration all tests, thresholds, and performance over multiple periods when assigning a model's performance monitoring rating. Model performance monitoring risk ratings vary across financial institutions, but traffic light systems (red, yellow, or green) or numeric rating systems (.e.g., 1-5) are commonly used.

1.9 Model Decommission

Over time, a model may be decommissioned due to deterioration in performance, obsolescence, or irrelevance. When a model is decommissioned, the model life cycle often continues as a new one may replace the old. The decommissioned model may often be used as a benchmark against the newly developed model for comparative purposes. There should be appropriate policies and procedures in place to ensure all relevant and impacted stakeholders are aware of an upcoming model decommission.

1.10 Conclusion

Globally there have been significant changes in financial institution's quantitative departments since the great recession of 2007-2008. This change was largely motivated by many newly issued regulatory requirements and policies such as Basel Advanced IRB (AIRB) (Bank for International Settlements 2019), the Dodd-Frank Act stress test (DFAST) (Board of Governors of the Federal Reserve System 2019b) and Comprehensive Capital Analysis and Review (CCAR) (Board of Governors of the Federal Reserve System 2019a) in the US, the International Financial Reporting Standard (IFRS) 9 (IFRS 2018) published by the International Accounting Standards Board (IASB), the

Fundamental Review of the Trading Book (FRTB) (Committee on Banking Supervision 2013), Current Expected Credit Losses (CECL, the US's IFRS-9 equivalent), and IFRS 17 (IFRS 2017), to name a few. The models developed in financial institutions' quantitative departments are heavily dependent on these guidelines and others, which makes developing theoretically sound econometric/mathematical models even more complex.

Effective Model Risk Management (MRM) is a complex set of tasks that requires a mastery of both econometric theory and the landscape of domestic and global regulations. This chapter provides both practitioners and academics with holistic and comprehensive guidance on navigating these challenges for a mortgage credit risk model. It sets out the mathematic and historical principles that underpin the development of a credit risk ECL mortgage model and builds upon it to offer practical guidance on the implementation, use, validation, and monitoring procedures for an effective credit risk modeling department. The reader is now familiar with the respective governance, controls, policies, and procedures present at all stages of the MRM lifecycle. As a result, they are better equipped to meet regulatory expectations while taking advantage of industry best practices.

2 Chapter 2: Ensemble Probability of Default Credit Risk Methodologies

2.1 Introduction

Binary choice classification models are fundamental and widely used in the field of study of economics. For example, in labor economics it helps analyze questions related to whether individuals are employed vs not employed, or what factors affect those who claim unemployment insurance vs those who do not. In health economics it helps analyze factors which impact whether individuals have a commorbidity or not, which factors lead to individuals dying or surviving, or whether individuals use a health service or not. In educational economics it can help determine what characteristics affect the likelihood of obtaining higher education. In financial economics, classification models are extremely important for determining the likelihood that a loan or investment will default, for detecting fraud, or detecting whether an amortizating loan will prepay on its contract or not, to name a few common applications.

This chapter provides a comparative assessment of applied bagging and boosting machine learning classification methods in Probability of Default (PD) credit risk modeling. I demonstrate the utility of these methods by way of an empirical comparison of their performance relative to a traditional benchmark model using Freddie Mac's loan performance data on a portion of its single-family mortgage loans. Bagging and boosting methods are considered "ensemble" learning techniques. Ensemble methods combine multiple classification estimates into a single weighted or unweighted estimate. Those familiar with "combined forecasts" and "model averaging" will immediately appreciate these approaches. While ensemble methods are among the most popular machine learning methods and often deliver the best results (Abellán and Castellano 2017), they are infrequently used in applied credit risk modeling. Methods for how these so-called “black box” methods can be used and interpreted by and for risk practitioners are also reviewed in this chapter. This is of particular interest for both academics and practitioners, as there is still significant uncertainty as to how these types of models can and should be integrated into the credit risk modeling industry. To the best of this authors knowledge, there exists neither a benchmark/review of these models, nor academic guidance on how effectively to implement a PD mortgage credit risk model using these techniques accepted by model validators and auditors in the financial industry.

Section 2 provides an overview of bagging and boosting ensemble methods with particular focus on classification. Section 3 reviews the ensemble literature and its application in credit risk. Section 4 reviews the dataset and develops a bagging model and boosting model that can be applied in a credit risk framework and assesses the practicality of each in turn. The WOE PD model in Chapter 1 is used as a benchmark model for comparative purposes. Section 5 concludes.

2.2 Overview of Ensemble Methods

This section reviews the historical and theoretical background of ensemble classification methodologies, with particular focus on bagging and boosting. Classical algorithms rely on one model (“model selection”) for analysis. Generally speaking, ensemble methods average different machine learning algorithms. Machine learning, as defined in this study, uses data and algorithms to make classifications or predictions and to uncover key insights within the data. We are aware of the plethora of other machine learning techniques that could be used in this study such as neural networks, support vector machines, k-nearest neighbor, naive bayes classifier, to name a few. However, given the increasing popularity of boosting and bagging methods due to their success in machine learning competitions as well as their reputation for being some of the most successful methods (Berk 2017), we direct our attention to boosting and bagging.

This chapter distinguishes between three classification algorithm methods (Lessmann, Baesens, and Seow 2015), namely individual classifiers, homogeneous ensemble classifiers, and heterogeneous ensemble classifiers:

- *Individual classifiers:* Individual classifiers pursue different objectives to develop a single classification model. Some examples include k-nearest neighbor, linear discriminant analysis, support vector machines, Classification and Regression Trees (CART), neural networks (multilayer perceptron, radial basis function), survival analysis, quadratic discriminant analysis, and logistic regression.

- *Homogeneous Ensemble Classifiers:* Homogeneous ensemble classifiers pool the predictions of multiple base models (or 'base selectors') using the same classification algorithm. This can be performed in either an independent (bagging) or dependent (boosting) fashion. Bagging ensemble methods, for example, create N homogeneous independent classifiers from N bootstrap samples of the original data (Breiman 1996). Boosting ensemble methods, on the other hand, actively try to force the added base model to change its hypothesis by changing the distribution over the training examples as a function of the errors made by a previously generated hypothesis (Freund, Schapire, and Hill 1996).

- *Heterogeneous Ensemble Classifiers:* Similar to the homogeneous classifier, heterogeneous ensemble classifiers combine multiple classification models but create these models using different classification algorithms. The idea is that different algorithms have different views about the same data and can complement each other. These are also referred to as a "stacking" approach. While stacking is briefly discussed in this chapter, given the regulatory scrutiny on black-box credit risk modeling techniques and time investment, stacking is not reviewed in the same level of detail as homogeneous ensemble classifiers (bagging and boosting).

Generally speaking, the challenge with individual classifiers is choosing a single (best) model from a set of models, the classical "model selection problem". All models are a subset of a general superset (overlapping) model which contains all submodels as special cases (Hansen 2020). Since there are theoretically an infinite number of possible submodels, the chance of locating the best, "correctly specified" parametric model is essentially zero. By reusing the same dataset in multiple ways (i.e., ensemble methods), we allow for multiple views of the same data, which as it turns out will always reduce variance and inches us closer toward the best ("correct") view. The challenge then becomes a question of how best to control bias. Averaging several independently trained regressions will never increase the expected error (Schapire and Freund 2012). Unfortunately, this is not transitive to classification, majority vote ensembling (this will be discussed in more detail below). The bias can be interpreted as the persistent error that would remain even if we had an infinite number of independently trained classifiers, while the variance measures the error due to fluctuations that are part of generating a single classifier (Schapire and Freund 2012). The following subsections respectively describe bagging, boosting, and stacking in further detail. We include stacking in the discussion below for informative purposes as it is a natural extension of bagging and boosting; however, this chapter directs attention to bagging and boosting in later empirical sections.

2.2.1 Bagging

This sub-section provides an overview of bootstrap aggregation, which is commonly referred to as 'bagging'. Bagging is a homogeneous ensemble method. Bagging is performed by bootstrapping independently and identically distributed (i.i.d) sample datasets with replacement B times of size n from a training dataset (refer to Efron 1979 for further discussion on bootstrapping theory (Efron 1979)). To generate a bagging estimate, let $m(x) = \mathrm{E}[y_i|x_i = x]$ be an unknown conditional mean, and let $\hat{m}(x)$ be the respective regression estimator. For each b bootstrapped dataset, recalculate the same estimator $\hat{m}_b(x)$ B times for each bootstrapped dataset. The final bagging estimator of $m(x)$ is

$$\hat{m}_{bag}(x) = \frac{1}{B}\sum_{b=1}^{B}\hat{m}_b(x). \tag{14}$$

For classification, instead of averaging the output, each classifier estimator $\hat{m}_b(x)$ output is collected and counted and the event with the highest number of counts is the victor. Put simply, classification uses a vote instead of an average. Ties are broken arbitrarily. This method works the same for both binary and multi-class classification. Each bagging estimater b is independent of other bagged estimates, so in practice parallel processing can be leveraged to reduce computational efforts. The bagging equation for a classification problem is illustrated below:

$$\hat{m}_{bag}(x) = \arg\max{}_{y \epsilon Y}\sum_{b=1}^{B}\mathrm{I}(\hat{m}_b(x) = y), \tag{15}$$

where I is an indicator. To explain in words, take a random sample of size N with replacement from the data and create B bootstrap samples. Now consider the example where each estimator $\hat{m}_b(x)$ is a binary classification estimating either 0 or 1. The total number of 0s and 1s estimated from each b bootstrap estimator is counted, and the algorithm will choose either 0 or 1 as its final classification based on whichever one has occurred more than 50% of the time. The ensemble will only result in an error if at least half its base classifiers make an error. The number of 0s and 1s estimated are divided by the total number of B bootstraps to arrive at a percentage, or proportion. It is important to note that the winning proportion is not an estimate of the probability that the imputation or forecast is correct (Berk 2017).

Bagging has the benefit of reducing variance, though it comes at the cost of accentuating bias. That is, if $\hat{m}(x)$ is biased, then bagging will increase the bias. Since bagging reduces variance but increases bias, it is considered a good approach when the estimator being bagged is known to have low bias and high variance. When this is the case, a lower Mean Squared Error (MSE) can be achieved, which potentially makes bagging a good approach for prediction (since it can be expressed as $Bias^2 + Variance$). While bagging does reduce variance, this reduction may be small since the final bagging is done on identically distributed but correlated bootstraps; each bootstrapped sample is likely quite similar to each other and may result in similar classification decisions.

A commonly used modification of bagging known as the “random forest” technique is designed further to reduce estimation variance. Random forests introduce additional randomness and decreases dependence among each bootstrapped estimate, which helps lower the variance. To create a random forest, bagging is applied to Classification and Regression Trees (CART). CARTs, as their name suggests, are typically displayed in a tree-like structure. They are machine learning algorithms that recursively partition the data space and fits a simple prediction model within each partition (Loh 2011). Classification trees are for dependent variables with finite unordered values, while regression trees are for continuous or ordered dependent variables.

Random forest makes three enhancements to the traditional bagging technique:

1. For each bootstrapped sample, a number of variables $m < k$ are selected at random for each CART estimator, where k represents all the variables in the candidate variable set. It is recommended to set $m = k/3$ (Hansen 2020), which is the default in the R RandomForest package for regression; $\sqrt{k}$ is the default for classification. Randomly selecting m variables from k candidate variables for each bootstrap estimator helps prevent the CARTs from repeatedly choosing the best predictors, which is what helps reduce correlation among the predictors and, as a result, reduces the variance.
2. The dependent variable in the random forest bagging exercise is sampled without replacement, whereas the independent variables are still resampled with replacement.
3. The bagging averages are performed on the Out-of-Bag (OOB) cases (these are described in further detail below).

Items 1, 2, and 3 above each decrease the dependence among the estimates, which is a key determinant for how effective an ensemble of trees will be. These items help resolve the issue of overfitting. Recall that in the random forest method, bagging is applied to CARTs. Choosing or ‘tuning’ the optimal parameters while avoiding model overfitting plays an important part in both the CART and random forest Out-of-Sample (OOS) and/or Out-of-Time (OOT) test dataset performance. Some important parameters which require tuning are listed below:

- The number of variables k.
- The subset of variables m (the maximum features).

- The number of trees (the number of bootstraps B).
- The depth of each tree.
- The minimum number of observations required to be in a leaf node (which also determines the minimum number of observations for a tree split).

In practice, parameters controlling the complexity and size of the tree should be controlled in order to reduce computational memory consumption (generally speaking, important ones to consider are m, B, and the minimum number of observations required in a leaf). The number of trees should be at least several hundred, and it is likely sufficient to have several thousand (Berk 2017). Breiman recommends starting with the square root of k to choose m, then trying more or less to assess performance. Unless the number of candidate variables k is large (say, greater than 100) it is likely sufficient from a performance perspective to choose the subset of variables m to be 2 or 3, since there will be sufficient opportunity for each variable to weigh in as the number of trees grows. There are other decisions that must be made (the splitting criterion, leaf weights, sample size, etc.) that can add to the overall complexity. There are, however, industry accepted defaults that can sometimes be leveraged and that often perform well in practice (of course, they must still be understood and considered). Tuning is both an art and a science, since knowing preemptively the best parameters for performance on new, never-before-seen data is impossible.

One advantage of the bootstrap is that for each bootstrapped dataset of size n (with replacement), each observation has the probability $1 - (1 - 1/n)^n$ of being selected at least once. As $n \to \infty$, it can be shown that this probability approaches $1 - 1/e = 0.632$, where e is Euler's number. This means on average 63.2% of the available observations will be in the bootstrapped dataset, leaving 36.8% omitted observations (Efron and Tibshirani 1994). These omitted observations are referred to as the OOB observations. These OOB observations can be used for cross validation on all the bagged CART models to determine the generalization error in the random forest. In some instances, only the OOB estimates can be used in the bagging averaging technique. For demonstrative purposes, an empirical example of a random forest will be presented later in this document.

2.2.2 Boosting

This sub-section provides an overview of boosting, also known as Adaptive Resampling and Combining (ARCing). Boosting is a homogeneous ensemble approach similar to bagging. However, one primary difference is that while the training stage for bagging can be done in parallel, boosting is done sequentially due to its dependency on the previous iteration(s). For each bootstrapped sample in bagging, each observation has the same probability of being randomly sampled (i.e., sampling is done with replacement). For each subsequent sample in boosting, observations which were poorly classified by the previous model(s) in the overall system are more likely to be sampled (or assigned a larger weight). This is done so that the subsequent models which well classify these previously misclassified observations can be identified, and then these can be added to the system to improve overall performance.

Boosting combines base, weak learning algorithms and their respective models (or classifiers) into a single learner whose overall predictions can become quite accurate. The assumption is that the weak classifier's MSEs are at least a little better than a random guess. This is referred to as the *weak learner assumption*, and it is a fundamental component of boosting. While the weak classifiers on their own may perform poorly, boosting will sample (or assign weights to) the training datasets in such a way that each new base classifer contributes something different than those preceding it. Boosting can be performed not only on weak learners, but also on well performing learning algorithms (for example, using the C4.5 algorithm (Schapire and Freund 2012)). In fact, when the base learners are trees used for classification problems, more complex base learners perform better (Berk 2017). However, the important takeaway is that boosting will improve performance as long as i) the weak learner assumption holds, ii) there are sufficient data, and iii) the base classifiers are not overly complex. To conceptualize how boosting works in a bit more detail, we will first direct attention to adaptive boosting, or 'AdaBoost', which is a heavily researched algorithm that won its discoverers the Godel Prize in 2003 (Freund and Schapire 1997).

AdaBoost takes as inputs a training dataset consisting of $(x_1, y_1), \ldots, (x_m, y_m)$, where x_i is observation i from the domain space X and y_i is the respective classification event taking values $\{-1, +1\}$. Each iteration uses the training dataset, which has a distribution denoted D_t for each iteration $t = 1, 2, ..., T$, and weights $D_t(i)$, where i refers to the respective weight for observation i. In the first iteration, all observations are assigned an equal weight of $1/n$, where n is the number of observations in the training dataset. As the algorithm progresses through each iteration t, the weights are increased on the previously incorrectly classified observation(s) to represent the importance of correctly classifying them on the current round. For each weak learner estimate $m_t : X \to \{-1, +1\}$ trained on dataset $D_t(i)$ for $i = 1, ..., m$, the aim is to select m_t (a decision tree stump is commonly used in AdaBoost) which minimizes the weighted error:

$$\begin{aligned} \epsilon_t &= Pr_{i \sim D_t}[m_t(x_i) \neq y_i] \\ &= \sum_{i: m_t(x_i) \neq y_i} D_t(i). \end{aligned} \tag{16}$$

The error ϵ_t represents the chance of m_t misclassifying a randomly sampled observation i from the distribution D_t. Recall that to satisfy the weak learner assumption, $\epsilon_t < 1/2$ must hold (i.e., ϵ_t must be slightly better than random guessing).

The first iteration of D_1 assigns equal weighting to each observation of $1/n$. Once the first classifier m_1 is determined, we can calculate its measure of importance denoted a_t using the following formula:

$$\alpha_t = \frac{1}{2}\ln\left(\frac{1-\epsilon_t}{\epsilon_t}\right). \tag{17}$$

From here, we can now calculate the new weights for $D_{t+1}(i)$ using the following formula:

$$\begin{aligned} D_{t+1}(i) &= \frac{D_t(i)}{Z_t}\begin{cases} e^{-\alpha_t} \text{ if } m_t(x_i) = y_i \\ e^{\alpha_t} \text{ if } m_t(x_i) \neq y_i \end{cases} && (18) \\ &= \frac{D_t(i)\text{exp}(-\alpha_t y_t m_t(x_i))}{Z_t}, && (19) \end{aligned}$$

where Z_t is a normalization factor chosen such that the probabilities sum to 1 (i.e., is a proper distribution) (Freund, Schapire, and Avenue 1999). Note that this equation holds when y_t takes values $\{-1, +1\}$. There are different approaches to using these weights at each iteration. One can randomly sample with replacement from D_t using the newly calculated weights $D_{t+1}(i)$ to create a new data sample D_{t+1}, which will (likely) over-sample the misclassified observations and under-sample the correctly classified ones from the previous estimator m_t. This is called “boosting by resampling”. A new estimator m_{t+1} is then calculated on the new distribution D_{t+1}, which is designed to focus on the difficult-to-classify observations. This process is continued T times to create T estimators, at which point they are all combined to create a single classifier $m_{boost}(x)$. Similar to bagging, this can be achieved by a simple weighted vote of all the classifiers using the formula below. Alternatively, the weights can be applied directly, which is called “boosting by reweighting”.

$$m_{boost}(x) = \text{sign}\left(\sum_{t=1}^{T} \alpha_t m_t(x)\right). \tag{20}$$

We can think of α_t as the coefficients for the linear combination used in this equation.

AdaBoost is fast, simple, and easy to implement in practice. It also only has one tuning parameter, i.e., the number of rounds T. It can be shown that the in-sample training error falls exponentially as a function of the number of weak classifiers. Boosting is also known to be resistant to overfitting (which is beneficial for classification performance on OOS/OOT data) (Freund, Schapire, and Avenue 1999); this is, however, not guaranteed (i.e., performance may deteriorate on OOS/OOT data as the number of base learners in the algorithm increase). During the development of any model, if attention is solely concentrated on minimizing the training error, then the model may overfit spurious patterns which appeared in the training data purely by chance and, as a result, have poor generalization error. The model’s success is dependent on how well the model fits the data, that it has sufficient data, and that it is simple (Schapire and Freund 2012).

While not designed with this purpose in mind, AdaBoost is actually performing coordinate descent while greedily minimizing an exponential loss function. At the time it was a new, statistical way of thinking about AdaBoost (Breiman 1998) (Breiman 1999) (Friedman 2001) (Rätsch, Onoda, and Müller 2001) (Duffy and Helmbold 1999)

(Mason et al. 1999). The exponential loss function does a good job of heavily penalizing misclassified observations, and not penalizing the correctly classified observations. Once it was recognized that AdaBoost is just coordinate descent optimizing a specific (exponential) loss function, AdaBoost was then generalized to any loss function. Formally, the algorithm minimizes a loss function $L(\lambda_1, \ldots, \lambda_N)$, where λ_j represents a specific weight for an estimator $\tilde{m}_j$ over the finite space $M = \tilde{m}_1, \ldots, \tilde{m}_N$. Note that we are using the notation $\tilde{m}_j$ instead of m_t as described above in the AdaBoost algorithm. Both $\tilde{m}_j$ and m_t are from the same finite space M, where for any $\tilde{m}_j$ there must be an equivalent m_t. To draw parallels to the AdaBoost algorithm, the exponential loss function we intend to minimize would be expressed as follows:

$$\begin{aligned} L(\lambda_1, \ldots, \lambda_N) &= \frac{1}{m} \sum_{i=1}^{m} \exp(-y_i F_\lambda(x_i)) & (21) \\ &= \frac{1}{m} \sum_{i=1}^{m} \exp\left(-y_i \sum_{j=1}^{N} \lambda_j \tilde{m}_j(x_i)\right). & (22) \end{aligned}$$

In AdaBoost, each round t is essentially adjusting one of the weights λ_j; i.e., it is minimizing the objective function L by iteratively descending along one coordinate at a time (Schapire and Freund 2012). This concept can be extended to Gradient Boosting (which uses gradient descent instead of coordinate descent), which adjusts all the weights $\lambda_1, \ldots, \lambda_N$ simultaneously during each iteration. The goal is then to minimize $L(F)$, which in its general form is called “AnyBoost” or “gradient boosting”. When each trained ensemble is performed on a subset of the training dataset, it is referred to as “stochastic gradient boosting”, which has the benefits of potentially improving the generalization of the model’s performance. Finally, if we consider also using the second-order derivatives of the loss function and incorporate L1 and L2 regularization, then we would be using the Extreme Gradient Boosting (XGBoost) method (Chen and Guestrin 2016), which has been popularized in recent years due to its success in machine learning competitions. An empirical example of XGBoost will be presented later in this document for demonstrative purposes.

2.2.3 Stacking

This sub-section provides an overview of stacking. As was previously mentioned, stacking is a heterogeneous ensemble classifier, which combines classification models which were built using different classification algorithms. Each model in the ensemble is trained on the same data. Various ensemble methods are used to create a suitable subset of all possible models using some selection criteria. Stacking is a relatively simple concept; namely, the combination of many models will likely do better than one single classifier. The simplest approach to building a stacked ensemble is to use the voting method. In a study by Lessman et al. (2015) which compares 41 different classifier methods and ranks their performance in credit risk probability of default forecasting, heterogeneous stacking ensembles ranked in the top 11 (Lessmann, Baesens, and Seow 2015). A shortfall of stacking is that the developer must build multiple models, which is quite a timely and, hence, costly approach. There are ways of implementing stacking procedures (refer to H_2O AutoML automatic machine learning, for example[7]). These approaches, however, tend to venture further into the black box territory, which may encounter push-back from a validation and/or internal/external audit team. There is also the argument that if a method is *too* easy to implement, then there is the risk that the developer might make unknown errors.

[7] https://h2o.ai/.

2.3 Literature Review

Estimating a financial instrument's PD is an important procedure in banking and is a fundamental component in forecasting Expected Capital Losses (ECL), which ultimately helps manage systemic risk. As financial institutions become more sophisticated and complex, the quantity and criticality of PD models used for decision making purposes has increased exponentially. Current and upcoming global and domestic (US) regulatory guidelines that banks have to comply with such as Basel Advanced IRB (AIRB) (Bank for International Settlements 2019), the Dodd-Frank Act stress test (DFAST) (Board of Governors of the Federal Reserve System 2019b), Comprehensive Capital Analysis and Review (CCAR) (Board of Governors of the Federal Reserve System 2019a), the International Financial Reporting Standard (IFRS) 9 (IFRS 2018) published by the International Accounting Standards Board (IASB), Current Expected Credit Loss (CECL, the US equivalent of IFRS-9), the Fundamental Review of the Trading Book (FRTB) (Committee on Banking Supervision 2013), and IFRS 17 (IFRS 2017) rely heavily on models, thus increasing dramatically the rate at which a bank's model inventory grows over time and highlighting the necessity for robust models. This section provides a review of the classification literature in credit risk modeling, and provides some insight into what situations bagging, boosting, or stacking might be best suited.

It is worth emphasizing that no single best "universal" method exists, since certain approaches may be more suitable given the situation and resources available (data, time, computational ability, experience, etc.). Benchmarking methods on a single dataset may result in conflicting conclusions, since it fails to examine the robustness of the models ability to generalize across different environments. It is also challenging to get an objective comparison when benchmarking methodologies because the authors may unintentionally tune one approach more than another (Lessmann, Baesens, and Seow 2015). This section attempts to provide an objective, theoretical review of where methodologies tend to excel or fail. That being said, the empirical component of this chapter is applied on a single, publicly available dataset, and is intended to objectively analyze and compare the ease of implementation, interpretability, and its overall suitability in applied credit risk modeling in industry, with the understanding that failing to analyze on multiple datasets may limit our ability to assess these methodologies ability to generalize across different environments.

2.3.1 Class Imbalance

We will first discuss the issue of “class imbalance”, which is prevalent in essentially all credit portfolios. There are significantly more loans which do not default versus those that do in credit portfolios, which is a well known phenomenon in industry, and publicly available default rates are commonly published by Moody’s for reference.[8] In machine learning techniques (as we have discussed thus far), methods may have a tendency to improve the true negatives (in our case, increase the correctly classified non-default loans) while also increasing the false negatives, rather than improve the true positives, to improve overall accuracy (Galar et al. 2012). This, of course, would give the misleading impression that the model is performing well. In practice, the cost of misclassifying an abnormal (defaulted) example as a normal (non-defaulted) example has a much higher cost than its converse.

To address the class imbalance issue one can under-sample the majority class (non-default), over-sample the minority class (defaulted), or some combination of both. One of the better known and utilized approaches is the Synthetic Minority Oversampling

[8] https://www.moodys.com/.

Technique (SMOTE) (Chawla V. et al. 2002), which is well researched and has demonstrated improvement in ensemble bagging and boosting classification methods (Galar et al. 2012). Instead of oversampling from the minority group with replacement, SMOTE over-samples by randomly selecting synthetic observations between all minority class examples using the k Nearest Neighbors (kNN) approach. This allows for a larger minority class decision boundary, which spreads into the majority class space. The SMOTE technique can be extended to both bagging and boosting, and is naturally called SMOTEBagging and SMOTEBoosting, respectively. These methods have been shown to improve classification performance (Galar et al. 2012) and as such, the SMOTE technique will be leveraged in the empirical exercise below.

2.3.2 When to Use Bagging and Boosting

An overview of how bagging, boosting, and stacking theoretically and mechanically works is provided above. This section discusses the situations where each of these approaches perform well or poorly. This is intended to provide practitioners with a more comprehensive understanding of when bagging, boosting, and stacking might be the most suitable method for a given application.

During a bagging procedure and under the right circumstances, if each single classifier has a high variance, averaging will smooth the estimator and hence smooth and reduce the variance (random forest is designed to reduce the variance even further). Essentially, bagging will tend to cancel out idiosyncrasies observed in the data. For this reason, bagging is largely viewed to be a variance reduction technique that reduces generalization error - an important feature when forecasting. Another benefit is that bagging allows for more complex classifiers in the ‘bag’, which normally have the risk of overfitting and increasing generalization error. Bagging will help stabilize the classifier estimates. This is why the random forest approach discussed above performs so well in practice. In addition, the random forest technique will not overfit as more trees are grown (an important feature) (Breiman 2001). There are cases where bagging can potentially reduce bias as well (see Section 4.3.7: Bagging and Bias in (Berk 2017)), however, this is not typically the primary reason for opting to use a bagging technique.

A notable pitfall of bagging is that since there is no single classifier nor an average classifier we can analyse and interpret, bagging is what is commonly referred to as a 'black box' statistical procedure. This has huge consequences in many fields in banking, where interpretability and justification for decision making and forecasts are important, e.g., lending decisions, economic forecasts, or determining future risks and losses. If a forecast or estimate cannot be rationalized or understood, there is a risk there may be bias, idiosyncrasy, racism, ageism, etc., unintentionally driving the estimate (as Berk notes, "statistical inference can certainly be useful, but are worse than useless when a credible rationale cannot be provided" (Berk 2017)). While not necessarily a limitation, bagging will differ from a single estimate only when the latter is a non-linear or adaptive function of the data (Hastie, Tibshirani, and Friedman 2009) (for example, bagging a linear regression would not result in any added benefits). Something for practitioners to be cognizant of, particularly in unbalanced data, is that the underlying classifier $\hat{m}(x)$ should perform reasonably on its own when bagging. If it performs poorly, bagging may result in worse estimates because the majority vote will lean toward the majority class (i.e., non-default vs defaulted). In a similar vein, if the base classifier persistently makes errors, bagging will simply continue to reproduce these errors. In the unlikelihood that an individual classifier is relatively stable, bagging may not be the best option as it has the potential to make things worse (see Section *4.4.2: Sometimes Bagging can Make the Bias Worse* in (Berk 2017)).

Alternatively, if we are working with unstable classifiers like CART methods (which are well known to have high variance), bagging is a very powerful and useful technique. Overall, the random forest has proven to be one of the best forecasting tools out there. As described by Berk, "if forecasting accuracy is one's main performance criterion, there are no other general purpose tools that have been shown to consistently perform any better" (Berk 2017).

Along with random forests, AdaBoost and Gradient Boost (and its derivatives like XGBoost) are considered some of the best off-the-shelf machine learning methods out there. Boosting can be shown to reduce both variance and bias (Schapire and Freund 2012). While bagging is appropriately considered a variance reduction tactic, boosting can be an effective tool when the model has large bias and low variance. In general, boosting will not perform well if there is i) insufficient data relative to the complexity of the base classifiers, and ii) the training errors of the base classifiers grow too quickly. In addition, boosting works particularly well with weak learners. If the pool of independent variables is significant and the process generating the data is well understood, a parametric regression may not benefit much from boosting and it will have the added benefit of ease of interpretability. Boosting also suffers from the same black box limitation as bagging, as does stacking.

2.4 An Empirical Comparison

This section provides an empirical comparison between bagging and boosting. In addition, the binomial logistic regression using the Weight-of-Evidence (WOE) technique from Chapter 1 is included as a benchmark comparison model. For further details on the WOE model development process, refer to Section "1.5 Model Development". Similarly, this chapter uses Freddie Mac's Single-Family Fixed Rate Mortgage Loan

Performance dataset (“the data”), where the final dataset and respective preparation and cleaning procedures taken are identical to the ones described in detail in Chapter 1, Section “1.5.1 Dataset Background and Relevance”. The same 70% of the data is randomly sampled for training the model, 30% is used for validating the model on the OOS test dataset, and a hold-out period of one year is used for the OOT test dataset. The variable of interest is whether an account defaults within a one-year (12 month) window or does not.

It is important to keep in mind that only one dataset is being used. Using multiple datasets would provide a more robust comparison of how well these methods perform in different environments (e.g. one method may perform better on one dataset, while another method may perform better on a different dataset). As such, in addition to model performance, this section is intended to analyze and assess objectively the ease of implementation, interpretability, and overall suitability of each methodology in applied credit risk modeling in the industry. Each model begins with the same set of independent variables.

Sub-sections 1 and 2 respectively presents the empirical model development of i) a random forest (bagging) model, and ii) an XGBoost (boosting) model. Sub-section 3 summarizes the results of each method, including the benchmark WOE logistic model from Chapter 1, and provides narrative on the pros and cons of each.

2.4.1 Random Forest Model Specification

This section describes the model development process undertaken for the random forest bagging technique. One benefit of the random forest is that tuning the model is rather simple when compared to, say, XGBoost.

The tuning parameters chosen are listed below:

- *The number of variables k:* the variables used are the same as the ones used in the benchmark model, i.e., $k = 31$. The intent is for each model to begin with the same information set.
- *The subset of variables m (the maximum features):* The default $m = \left\lfloor \sqrt{k} \right\rfloor = 5$ is used to choose the subset of variables. The tuning process is described in more detail below.
- *The number of trees (the number of bootstraps B):* The number of trees/bootstraps is 501. The tuning process is described in more detail below.
- *The depth of each tree:* Trees are grown to the maximum possible, with no limit on the node size.
- *The minimum number of observations required to be in a leaf node:* This also determines the minimum number of observations for a tree split. No minimum is set.

The subset of variables are tuned using $m = 2, 3, 4, 5, 6, 7, 9, 11, 13$. It is recommended to use $m = \left\lfloor \sqrt{k} \right\rfloor = 5$ for classification as a benchmark, so values around this are run to verify the optimal k. This tuning exercise is visualized in Figure 10 using the ROC on the OOB values. This means that the optimal tuning parameter is chosen based on whichever parameter has the highest ROC value when estimated on OOB values. As Figure 10 shows, the optimal subset of variables for ROC performance is $m = 4$, which is used in the final random forest model.

The second tuning parameter considered is the optimal number of trees to run. While tuning the number of trees is not recommended (Probst and Boulesteix 2018), it is still prudent to verify the stability of the probability estimations of all observations and ensure there are monotonic improvements. What this means is we should expect

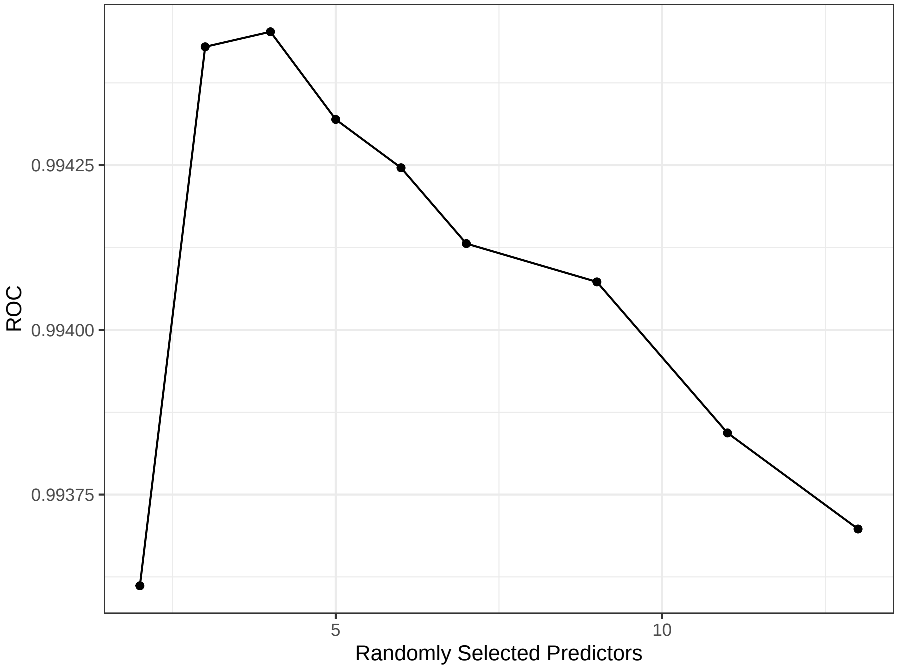

Figure 10: Optimal Subset of Random Forest Variables m Tuning

the model error to decline as we increase the number of trees. If it increases, then further investigation is required. It is observed that after 128 trees, there typically is no significant gains in ROC performance (Oshiro, Perez, and Baranauskas 2012) (Probst and Boulesteix 2018); this is observed here as well. The general consensus is that in most cases more trees are better. The improvement for OOB error rate is displayed in Figure 11.

A common critique of random forest models is that they are considered too opaque to be used for prudent risk management; this is commonly referred to as a black box. While analyzing each tree individually may indeed be unrealistic, very useful information can be extracted from a random forest algorithm which should alleviate the concerns credit modelers and respective management/stakeholders may have. The following suggestions should provide practitioners with some guidance on how to interpret their random forest model results, and how to be confident that their model

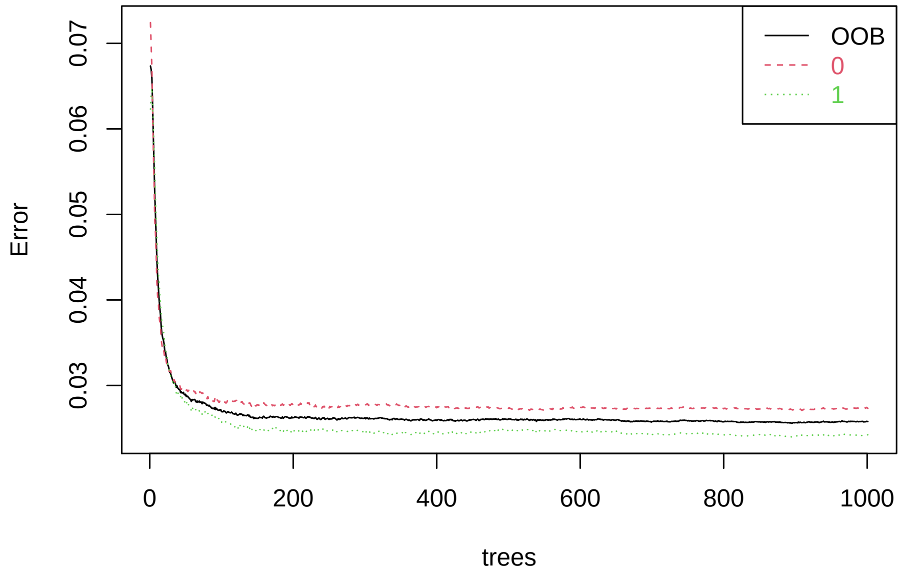

Figure 11: Random Forest Error Rates on Training data

is performing as intended and will continue to do so.

After running the random forest model, each variable's importance can be rank ordered. This is done by determining how much the OOB model accuracy decreases when each variable is excluded. Variables with a higher number are considered to be more important. Practitioners can use this to ensure that the variable's rank ordering is intuitive and in line with industry expectations. Each variable's importance, which is determined by how much the model's accuracy improves when including it in the model using the mean decrease in accuracy on OOB data, is ordered by importance as shown in Table 5. This method can also be used as an effective variable selection technique.

Another consideration that credit risk practitioners consider important is the effect each variable has on the dependent variable. In a regression, it is important that the coefficient has the correct sign. For example, unemployment rates should intuitively have a positive coefficient when modeling default rates, as we would expect default rates

Table 5: Random Forest Error Rates on Training data

	Variables	**Mean Decrease Accuracy**
1	HPI Change	128.62
2	Credit Score (FICO)	118.95
3	Original Value	101.53
4	Last UPB	98.86
5	Current UPB	95.65
6	Original UPB	92.90
7	Property Type	91.87
8	DTI	85.36
9	OCTV	82.33
10	Loan Age	80.81
11	Occupancy Status	79.96
12	Interest Rate	79.51
13	Modification Flag	73.91
14	LTV	68.78
15	FTHB flag	65.19
16	Delinquency Status	63.85
17	Mortgage Insurance	62.95
18	Current Interest Rate	61.20
19	# of Borrowers	56.69
20	Number of Units	53.26
21	Loan Purpose	52.93
22	UR (L4)	50.33
23	Delinquency Indicator	47.44
24	Channel	43.75
25	UR (L3)	42.53
26	UR (L1)	41.68
27	UR	41.37
28	UR (L2)	41.27
29	Super Conforming Flag	8.57

to be positively correlated with unemployment rates. With a random forest model, partial dependence plots can be easily obtained, which gives a graphical depiction of the marginal effect a variable has on the class probability (Friedman 2001). Partial dependence functions can be used to help interpret black box models (Friedman 2001) or put differently, any supervised learning model. Partial dependence plots on the four most influential variables are displayed in Figure 12.

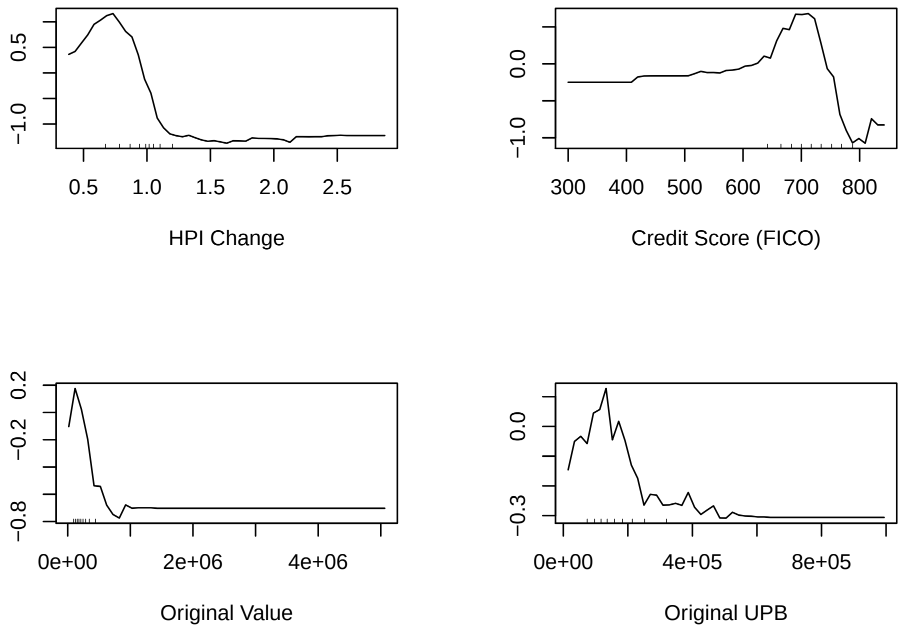

Figure 12: Random Forest Partial Dependence Plots

The partial dependence plots show that larger increases in HPI relative to when a house was first purchased result in a lower default probability (top left in Figure 12). As credit scores improve, there is a decrease in default probability, and default rates appear to be unaffected below 600 (top right in Figure 12). Default rates decrease as the value of a house increases (bottom left in Figure 12), which is in line with industry expectations since individuals who qualify for larger loans typically have higher credit scores, higher salaries, and pay higher down payments. Similarly, higher original UPBs have lower default rates (bottom right in Figure 12). This exercise can be performed on all variables in the random forest model, and helps provide risk management professionals with the ability to assess whether the independent variables have “correct” relationships with the dependent variable.

2.4.1.1 Random Forest Model Performance

This section evaluates the random forest model’s performance using the following statistical methods on both the OOS and OOT datasets:

- The Receiver Operating Characteristic (ROC) curve.
- The Kolmogorov–Smirnov (KS) test.
- The Gini Coefficient (AR Test).

The ROC curve, KS test, and AR test for both the OOS and OOT dataset are shown below:

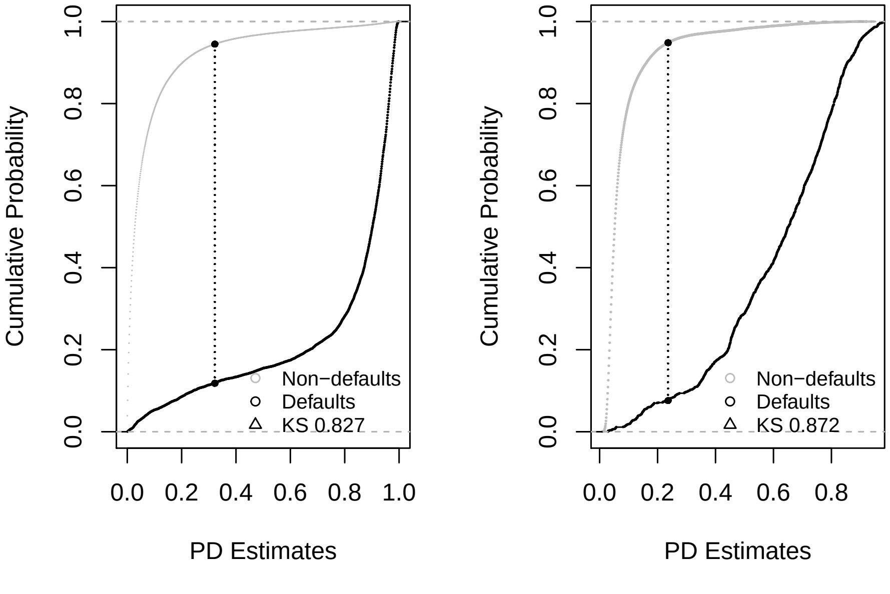

Figure 13: Kolmogorov–Smirnov Random Forest Model Performance Test on Out-of-Sample (left) and Out-of-Time (right) Dataset

As presented in the KS, AUC, and AR test figures above, the model is performing

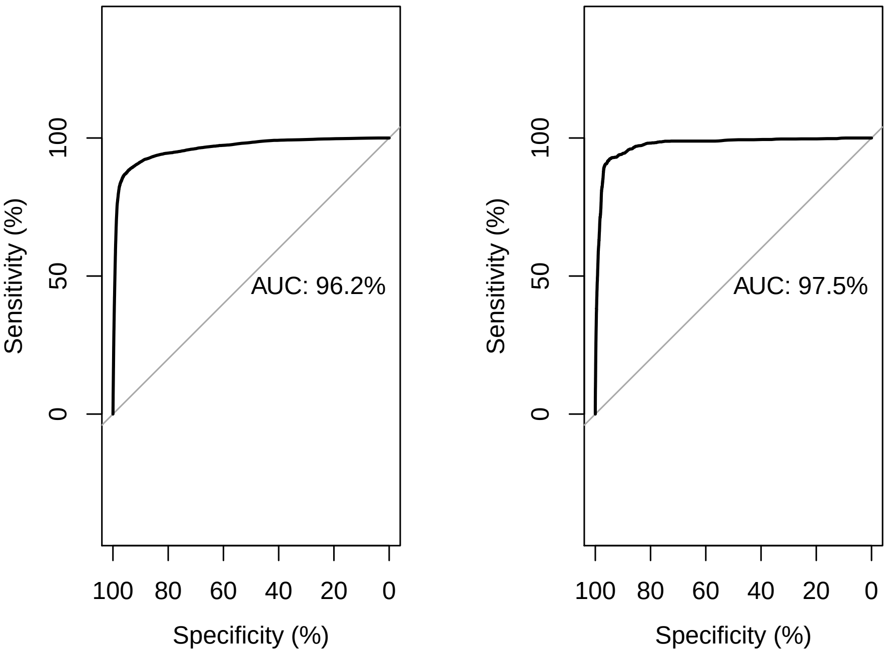

Figure 14: Receiver Operating Characteristic and Respective Area Under the Curve Random Forest Model Performance Test on Out-of-Sample (left) and Out-of-Time (right) Dataset

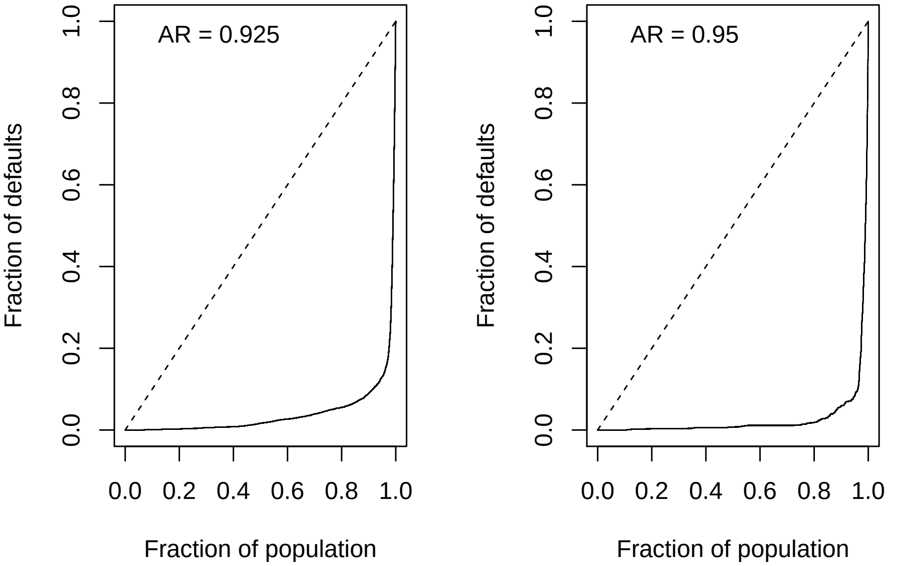

Figure 15: Cumulative Accuracy Profile Curve and Respective Accuracy Ratio Random Forest Model Performance Test on Out-of-Sample (left) and Out-of-Time (right) Dataset

well both OOS and OOT.

2.4.2 XGBoost Model Specification

This section describes the model development process undertaken for the XGBoost boosting technique. XGBoost tuning is more rigorous than random forest, and often takes multiple iterations. The parameters which typically have the biggest effect on model performance are the following:

- *Learning Rate:* the learning rate *eta* scales the contribution/improvement in error of each tree by a factor of $0 < r < 1$ when it is added to the current approximation. Smaller values of r may result in better predictions, though at the cost of more iterations to reach the optimum (i.e., more computation time). Values are often below 0.3.
- *Max Depth:* The maximum depth determines the depth of a tree at each iteration. The deeper the tree, the more complex the model becomes resulting in potential over-fitting in the OOS and OOT datasets.

- *Gamma: Gamma* sets a minimum loss reduction that is required for a further partition on a leaf node of the tree.
- *Minimum Sum of Weight (Hessian) needed to Create a Node:* Smaller values allow for nodes with fewer samples which allows for more complex trees, but are more likely to over-fit.
- *Number of Rounds:* The optimal number of rounds to perform.

For this model, the tuning parameters listed above are the ones tuned for this XGBoost model. In the first tuning iteration, $eta = 0.1, 0.05, 0.01$ and maximum depths of *depth* $= 2, 4, 6, 8$ are considered. Both of these are visualized in Figure 16.

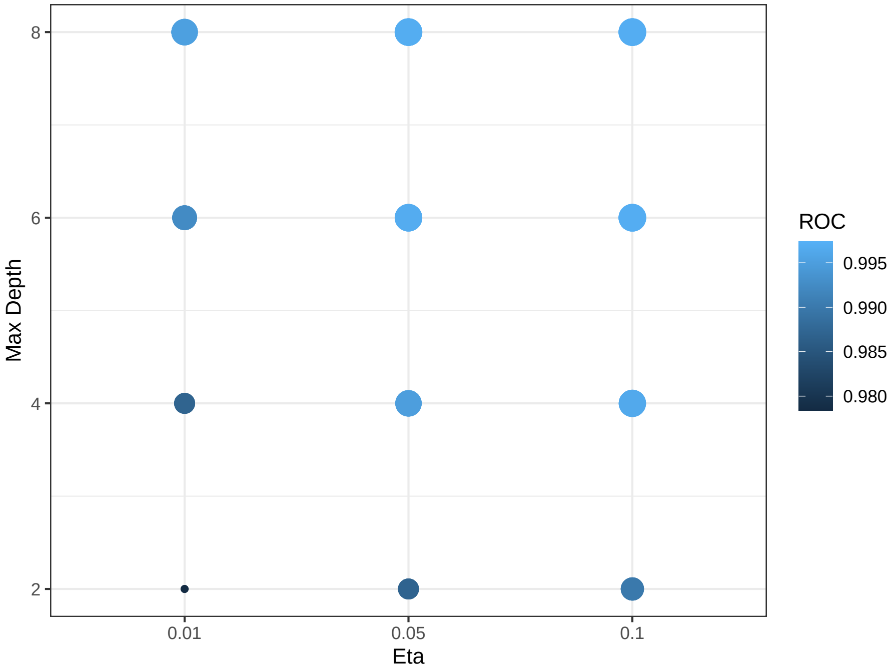

Figure 16: Optimal XGBoost Learning Rate and Maximum Depth Tuning

Similarly, *eta* and *gamma* $= 0, 0.25, 1$ are considered and visualized in Figure 17.

While the optimal maximum depth is 8, the improvements from 4 to 8 are marginal. Hence, to prevent over-fitting, an optimal depth of 4 is chosen. The optimal learning rate is 0.1. Instead of checking to see whether learning rates beyond 0.1 perform better, the number of iterations are increased instead (at the cost of increased computation). Optimal gamma is 0, hence this is used going forward. Minimum sum of weight (Hessian) needed to create a node values of 1, 3, and 5 are considered. Tuning shows 1 is optimal.

Finally, the optimal number of rounds are considered. If no improvement to the model using the Root Mean Squared Error (RMSE) has occurred after 20 model iterations, the algorithm terminates. The RMSE is calculated by comparing the models predicted probability estimates using $\hat{y}$ with the actual values in both the training and test

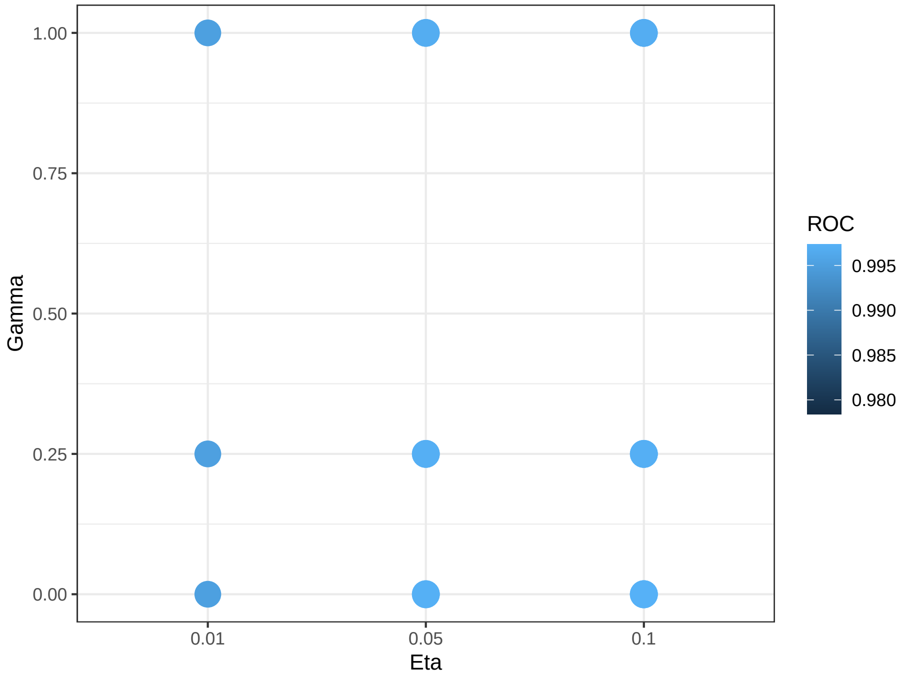

Figure 17: Optimal XGBoost Learning Rate and Gamma Tuning

dataset (predicted vs actual). This helps prevent model over-fitting. The improvement in RMSE model performance is presented in Figure 18, and the improvement in AUC model performance is presented in Figure 19. The improvement in the test data for both RMSE and AUC appears to approximately level off at 1500 rounds, which is used in the final model.

Similar to the random forest model, a common critique of XGBoost models is that they are considered to be a black box. However, also similar to the random forest model, variables can be rank ordered by importance, and partial dependence plots can also be obtained. Table 6 lists the variables in order by importance. The gain represents the contribution of each variable to the model; higher values indicate the variable is more important in the prediction. XGBoost can also be used as an effective variable selection technique using this approach.

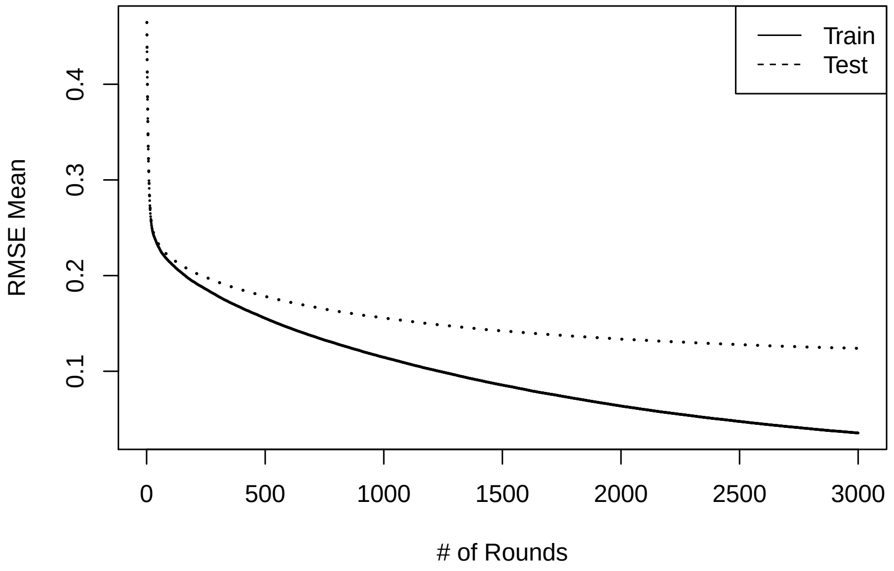

Figure 18: XGBoost Optimal Number of Rounds Tuning Using RMSE

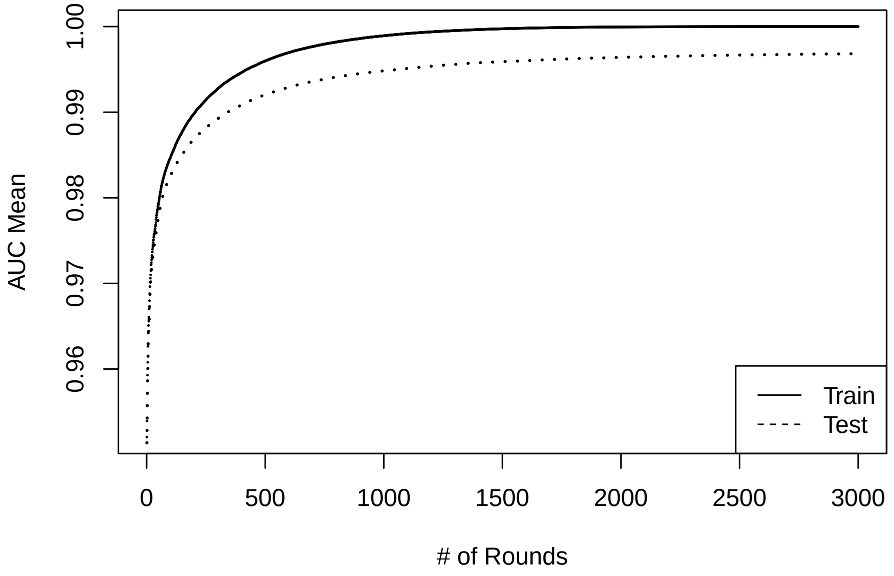

Figure 19: XGBoost Optimal Number of Rounds Tuning Using AUC

Table 6: XGBoost Importance

Variable	**Gain**
Delinquency Status	0.629

Table 6: XGBoost Importance *(continued)*

Variable	**Gain**
HPI Change	0.061
Interest Rate	0.037
Credit Score (FICO)	0.034
LTV	0.029
Current Interest Rate	0.022
DTI	0.020
Loan Age	0.018
Original Value	0.018
Mortgage Insurance	0.016
Current UPB	0.016
Occupancy Status	0.011
OCLTV	0.010
Original UPB	0.010
UR (L4)	0.009
Modification Flag	0.008
# of Borrowers	0.007
Property Type	0.007
UR	0.006
FTHB Flag	0.006
UR (L1)	0.005
UR (L2)	0.005

Table 6: XGBoost Importance *(continued)*

Variable	**Gain**
UR (L3)	0.005
Channel	0.004
Loan Purpose	0.003
Number of Units	0.003
Delinquency Indicator	0.000

The partial dependence plots for the top four most influential variables in the XGBoost model are displayed in Figure 20. These give a graphical depiction of the marginal effect a variable has on the class probability, and provides risk practitioners the ability to interpret the effect a variable has on the variable of interest. Figure 20 shows that default rates i) increase with delinquency status, ii) decrease as the change in HPI since origination increases, iii) increase as interest rates increase, and iv) decrease as credit scores increase. These are all in line with economic/business expectations.

2.4.2.1 XGBoost Model Performance

This section evaluates the XGBoost model's performance using the following statistical methods on both the OOS and OOT datasets:

- The Receiver Operating Characteristic (ROC) curve.
- The Kolmogorov–Smirnov (KS) test.
- The Gini Coefficient (AR Test).

The ROC curve, KS test, and AR test for both the OOS and OOT dataset are shown

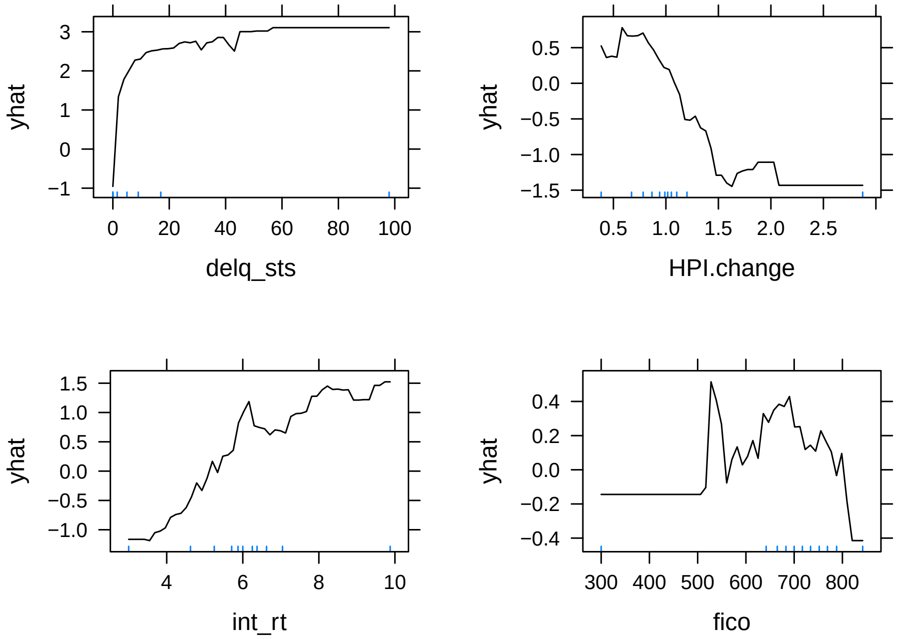

Figure 20: XGBoost Partial Dependence Plots

below.

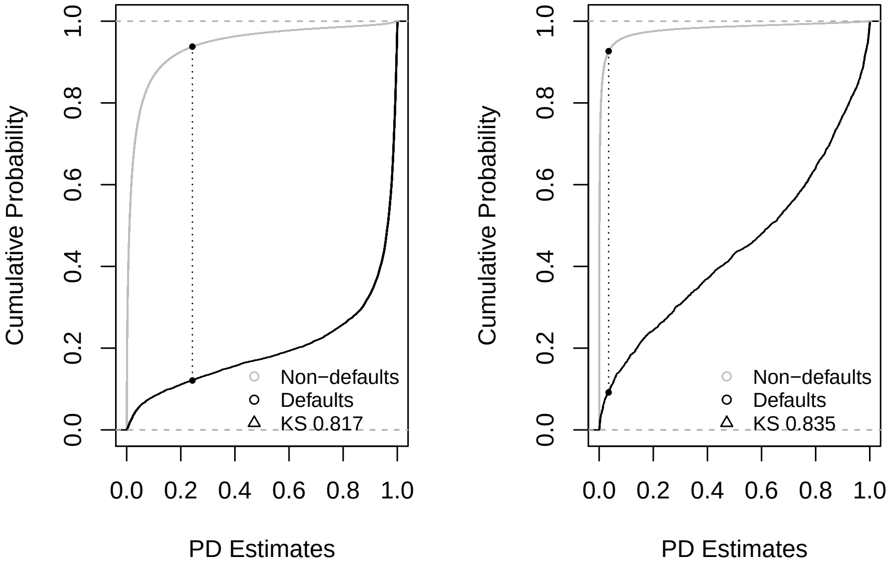

Figure 21: Kolmogorov–Smirnov XGBoost Model Performance Test on Out-of-Sample (left) and Out-of-Time (right) Dataset

As shown in the KS, AUC, and AR test figures above, the model is performing well

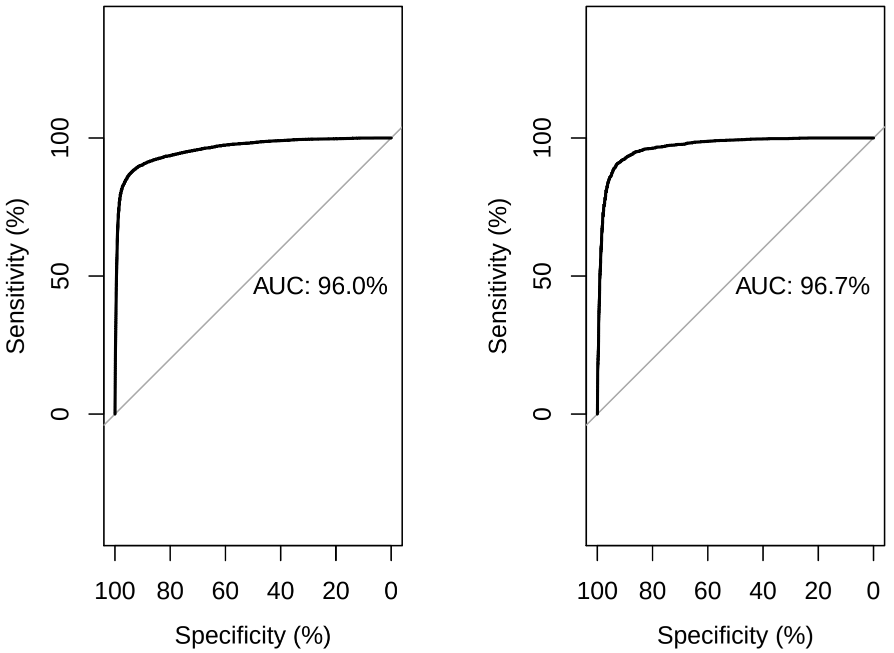

Figure 22: Receiver Operating Characteristic and Respective Area Under the Curve XGBoost Model Performance Test on Out-of-Sample (left) and Out-of-Time (right) Dataset

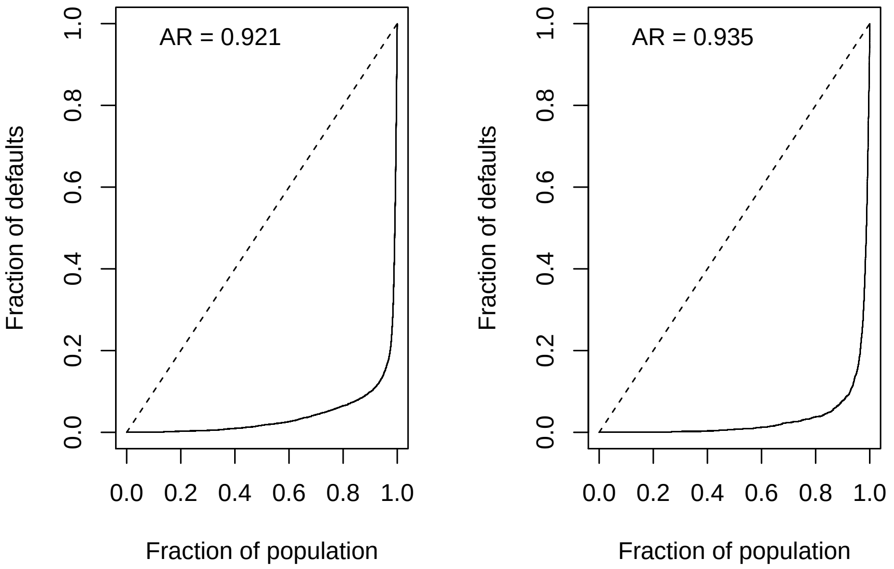

Figure 23: Cumulative Accuracy Profile Curve and Respective Accuracy Ratio XGBoost Model Performance Test on Out-of-Sample (left) and Out-of-Time (right) Dataset

Table 7: KS, AUC, and AR OOS Model Test Results Comparison

Test	**Logistic Regression**	**Random Forest**	**XGBoost**
KS (OOS)	0.805	0.827	0.817
AUC (OOS)	0.951	0.962	0.96
AR (OOS)	0.901	0.925	0.921

Table 8: KS, AUC, and AR OOT Model Test Results Comparison

Test	**Logistic Regression**	**Random Forest**	**XGBoost**
KS (OOT)	0.724	0.872	0.835
AUC (OOT)	0.917	0.975	0.967
AR (OOT)	0.834	0.95	0.935

both OOS and OOT.

2.4.3 Model Comparison

The previous sections provide an overview of the development process for a random forest model and an XGBoost model. The OOS and OOT KS, AUC, and AR test results for these models, including the benchmark WOE logistic model from Chapter 1 are summarized in Table 7 and Table 8, respectively.

The random forest and XGBoost models outperform the benchmark logistic regression model in both the OOS and OOT dataset. The random forest and XGBoost perform noticeably better in the OOT dataset, indicating that they may be better for generalization, and may have more reliable estimates for ongoing performance monitoring as new OOT data continues to be appended. The random forest methodology performs slightly better than XGBoost in OOS or OOT.

2.5 Conclusion

While logistic regression is a popular and widely accepted method in the industry, this chapter shows that the bagging and boosting methods clearly outperform a common and popular logistic regression specification in the empirical exercise. Common arguments against the usage of machine learning models is that they are “black boxes”. This, however, can be resolved with variable importance rankings and partial dependence plots. Partial dependence plots visualize the average partial relationship between the dependent and independent variable(s), which is an important component in the applicability and interpretability in credit risk modeling. PDPs do not capture the potential heterogeneity across independent variable ranges, which is an assumption a modeler may wish to analyse. Goldstein et al. (2015) proposed the Individual Conditional Expectation (ICE) method, which is designed to visualize the relationship between each observation and the dependent variable, which allows the reader to identify potential heterogeneity more easily (Goldstein et al. 2014). While ensemble methods are among the most popular machine learning methods and often have the best results (Abellán and Castellano 2017), they are infrequently used in applied credit risk modeling for the reasons above. The additional tests and graphs presented in this chapter are a constructive step in the right direction to resolve the common critique that bagging and boosting methods are considered too opaque to be used for prudent risk management.

3 Chapter 3: Benchmarking Ensemble and Traditional Loss Given Default Credit Risk Methodologies

3.1 Introduction

Fractional response variables commonly occur in many economic settings. Familiar examples include the fraction of total weekly hours spent working, the proportion of income spent on pension plans or charitable contributions (Papke and Wooldridge 1996), industry market shares, the proportion of students who pass standardized tests, and the fraction of land allocated to agriculture (Papke and Wooldridge 2008). In financial economics, a fundamental fractional response variable is the Loss Given Default (LGD), which is the observed percentage, often bounded between 0 and 1, of the actual exposure a financial institution can expect to lose in the event a borrower defaults on their loan.

LGD is one of three main parameters used in determining a bank's Expected Losses (EL) under Basel II international regulations (Basel Committee on Banking Supervision 2005), along with the Probability of Default (PD) and Exposure at Default (EAD). While PD models have been the focal point of credit scoring over the last 60 years, LGD modelling was not addressed adequately until the introduction of Basel regulations (Zhang and Thomas 2012). Basel II was implemented at the end of 2006 and requires banks to use a more risk-sensitive method for calculating credit risk capital requirements (Schuermann 2004). As a result, LGD has recently garnered research attention in both academia and industry. Generally speaking, it is more challenging to build accurate LGD consumer portfolio models than build PD models. Two reasons are i) the data are often right censored (debts are still being paid), and ii) borrowers typically have different repayment patterns (Zhang and Thomas 2012). As such, LGD industry models tend to have low predictive performance, particularly for consumer lending portfolios (Loterman et al. 2012).

This chapter provides a comparative assessment of credit risk Loss Given Default (LGD) modeling methodologies for use in financial institutions, with particular focus on a retail single-family mortgage portfolio. An empirical comparison of these model methodologies' performance using Freddie Mac's loan performance data on a portion of its single-family mortgage loans is used for demonstrative purposes. This chapter examines and compares the following four methods:

- Fractional response models (Papke and Wooldridge 1996) (Papke and Wooldridge 2008).
- A linear model.
- Random forest (Breiman 2001).
- XGBoost (Chen and Guestrin 2016).

For reasons this chapter provides, fractional response models are one of the most commonly used approaches at top tier banks. The linear model is a widely used method, despite the fact that its functional form is almost always misspecified when modeling LGD (it is typically linear and additive, which is limiting). We examine theoretically why that is, and empirically show an example of the linear model's performance shortcomings in practice, with the hopes of motivating practitioners to consider less restrictive methods. Finally, while random forest and XGBoost models have received ample attention in recent years, they are rarely used in practice for credit risk modeling. This chapter explains why that is, and discusses when and how these methods could be used effectively in credit risk modeling. This chapter is motivated to solve these problems, as well as to help fill the gap in limited LGD research and to explore potential improvements to predictive modeling. This chapter is also intended to provide practitioners and researchers with some guidance when choosing an appropriate methodology for modeling credit risk LGD models, particularly in mortgage portfolios.

Section 2 provides some helpful background knowledge on credit risk LGD in a banking environment. Section 3 reviews the methodological literature and industry best practices, guided by regulatory feedback. Section 4 describes the data. Section 5 discusses the model development component, compares the performance of these models, and summarizes the pros and cons of these methods. Section 6 concludes.

3.2 LGD Overview

Estimating financial instruments LGD (which is equal to 1 minus recoveries, which are defined in more detail below) is an important procedure in banking and is a fundamental component in forecasting Expected Losses (EL), which ultimately helps manage systemic risk. Along with the estimation of PD, LGD estimates and forecasts play a critical role in global and domestic (US) regulatory guidelines that banks must comply with. For more detail on these regulatory guidelines, refer to Sections 1.2 - 1.3 in Chapter 1. LGD is defined as the observed percentage of the actual exposure lost in the event of a borrower defaulting (Basel Committee on Banking Supervision 2005). It is typically measured as a percentage of the EAD and is used together with PD to estimate the EL (in currency amounts):

$$EL = PD \cdot LGD \cdot EAD \tag{23}$$

Larger and more mature banks will usually have separate quantitative models for each PD, LGD, and EAD parameter, and often build these models at the loan level, as opposed, for instance, to an aggregate estimate at the segment level, for example. Segmentation is the act of dividing large and diverse financial portfolios into smaller groups with similar characteristics. Losses technically only occur once a loan has formally defaulted, which generally happens when any of the following occur (Schuermann 2004):

- A loan is placed on non-accrual.
- A charge-off has already occurred.
- The obligor is more than 180 days past due (DPD)[9].
- The obligor has filed for bankruptcy.

Once an account defaults, both its cash inflows (recoveries) and outflows (charge-offs) are used to calculate the LGD. Charge-offs typically occur at the time of default and equal the amount of debt the creditor estimates it will be unable to collect from the borrower. Recoveries are observed in subsequent time periods, and are the sum of the amount the bank is able to recover. Recoveries need to be discounted appropriately back to the default date. There are three types of losses associated with LGD (Schuermann 2004):

- The loss of principal.
- The carrying costs of non-performing loans, e.g., interest income foregone.
- Workout expenses (collections, legalities, etc.).

Depending on the loan and/or product, it sometimes takes many years for all recoveries to be realized. A mortgage loan, for example, often takes 7 or more years to realize over

[9] Other retail exposures usually have a 120 DPD threshold and wholesale exposures have a 90 DPD threshold.

90% of all recoveries, as it must go through legal processes, the sale of the property, etc. This sometimes poses a challenge when building LGD models, because enough time must pass before a sufficient amount of recoveries has been observed to get accurate LGD values. Not waiting to observe a sufficient amount of recoveries would result in an LGD value that has upward bias (i.e., LGDs would be lower if more time was allocated to receive more recoveries). Choosing an appropriate 'hold out' period (i.e., how much time most pass before a sufficient amount of recoveries are realized) often requires portfolio-specific analysis by the model developers.

Across most credit portfolios, general key factors driving significant differences in LGDs are the capital structure, the presence and quality of the collateral, the industry, and the timing of the business cycle (Schuermann 2004). These, where possible, should be taken into consideration when modeling LGD percentages. Recoveries during recessions are systematically lower in recessions and, as a result, LGDs are often higher in such periods. Another important and commonly observed LGD characteristic is that distributions are often bimodal; i.e., LGD percentages (usually between 0-100%) are either relatively high or relatively low, and/or are skewed with modes close to the boundary values. This means that assuming a normal distribution or relying on 'average' LGD values might result in misleading estimates. Whether a loan is secured or unsecured also plays an important role in LGD. However, for the purposes of this chapter, the loans evaluated are all secured by the underlying mortgage property (condominium, leasehold, planned unit development (PUD), cooperative share, manufactured home, or single family home).

3.3 LGD Methodology Overview

This section reviews the LGD methodological literature and industry best practices, influenced by regulatory feedback. LGD models are often built using a single-step or a two-step approach. Single-step techniques are discussed first; two-step approaches are discussed second.

3.3.1 Single-Step LGD Methodology Review

Single-step methods typically model LGD as the dependent variable using a single linear or non-linear regression/approach. While linear regression is commonly used, conceptually it is not ideal because the dependent variable and the residuals will often not be normally distributed and the dependent variable values are usually bounded at 0 (no loss) and 1 (100% loss), with a large portion of values near or at the bounds. This results in unreliable confidence intervals and invalid hypothesis testing. Also, the effects from any particular independent variable x cannot be constant through the range of x (unless the range is limited), as this might result in estimates falling outside the 0 and 1 bounds. Said another way, the drawbacks of using a linear model for fractional data (where observations lie in [0, 1]) are analogous to the drawbacks of the linear probability model for binary data (Papke and Wooldridge 1996). Another approach which has received attention in recent years is the Frye Jacobs LGD modeling approach, which includes PD as a function of LGD (Frye 2013). This review does not assess methodologies that include PD as a function of LGD.

Non-linear techniques are often considered in an attempt to capture the non-linear characteristics commonly found in LGDs. Loterman et al. (2012) reviewed 24 techniques on 6 different loss datasets from major international banks and found that non-linear regression techniques like neural networks and Support Vector Machines (SVM) clearly outperform traditional linear models (Loterman et al. 2012). Other non-linear methods have been considered, such as regression trees (Bastos 2010), zero-adjusted gamma models (to accommodate for the excess number of zeroes and skewed nature of LGDs) (Tong, Mues, and Thomas 2013), SVMs, and Multivariate Adaptive Regression Splines (MARS). One of the more prominent methods deployed at top tier banks in the US, Canada, and Europe is “fractional logistic regression”, which is a special case of the “fractional response model” literature (discussed in more detail below) (Papke and Wooldridge 1996) (Papke and Wooldridge 2008).

Ensemble methods such as boosting and bagging are rarely used in banks’ credit risk models. Sun & Zin (2016) show that an ensemble approach using stochastic gradient boosting and random forests outperform a single decision tree when modeling LGD (Sun and Jin 2016). They also conclude that these are likely more appropriate methods for modeling LGD for credit risk portfolios than simple linear regression.

3.3.2 Two-Step LGD Methodology Review

When there is a large concentration of loans at the lower bound with 0% losses, or at the upper bound with 100% losses, a two-step approach might be more appropriate than a single stage approach. Two-step approaches use a combination of single-step models, and will first model a discrete component using a binary or multinomial model in order to separate/capture non-normal characteristics (e.g., the LGD concentration at the bounds and/or the data skewness) and then will model the continuous component second.

Choosing to model LGD as [0,1), (0,1], or [0,1] in one of the stages is up to the discretion of the developer (Ramalho, Ramalho, and Murteira 2011). To elaborate: in the first stage, one could first estimate the likelihood of a loan falling at the lower bound 0, the continuous (0,1) interval, and/or the upper bound 1. LGDs that have three occurrences of either 0, being between 0 and 1, or 1 are categories with a natural ordinal sequence and could be estimated using an ordinal or multinomial regression (note that an ordinal model has proportional odds that assume the independent variables have the same slopes for all categories, which is an assumption that needs to be tested). The second stage would estimate the continuous [0,1] interval. The expected value of the LGD for loan i is then calculated using the probabilities from stage 1 and the predicted LGD in stage 2 (Li et al. 2014):

$$\hat{\mathrm{E}}(LGD_i) = (0 \cdot \hat{P}_0^i) + \mathrm{L\hat{G}D}^i_{(\text{stage 2})} \cdot (1 - \hat{P}_0^i - \hat{P}_1^i) + (1 \cdot \hat{P}_1^i), \tag{24}$$

where $\hat{P}_0^i$ is the probability of loan i having LGD = 0, $(1 - \hat{P}_0^i - \hat{P}_1^i)$ is the probability of LGD falling between 0 and 1, $\hat{P}_1^i$ is the probability of LGD = 1, and $\mathrm{L\hat{G}D}^i_{(\text{stage 2})}$ is the predicted LGD in stage 2.

A common two-step approach is first to estimate whether the LGD is 0 (a full recovery) or greater than 0, then to use a second model to estimate the non-zero LGDs (Loterman et al. 2012). This approach is often used at top tier banks in the US, where the second stage is a fractional logistic regression (discussed in more detail below). The first stage is often a logistic regression. However, Tanoue et al. (2020) state that this could result in biased and inconsistent results due to the non-linear relationships commonly seen between the explanatory variables and the LGD dependent variable link function (Tanoue, Yamashita, and Nagahata 2020). They propose that using machine learning classification methods such as Support Vector Machines (SVM), neural networks, naive

Bayes, k-nearest neighbor, or random forest in the first step will better capture these non-linear relationships.

A two-step method proposed by Gurtler (2013) is first to model the probability of whether a loan results in a write-off or a workout (i.e. the lender and borrower have renegotiated the terms of the loan and the borrower is no longer in default), as the underlying characteristics of each type of default are arguably quite different (Gürtler and Hibbeln 2013). The second step would be to separately model the LGDs of recovered loans and loans which are written off separately. As mentioned above, another method first estimates whether the LGD is 0, 1, or between 0 and 1. Possible first-stage link functions are logit, probit, cauchit, log-log, and complementary log-log. Common second-stage methods are, similarly, logit, probit, cauchit, log-log, complementary log-log, or beta regression. Linear regression is used quite often as well (even though it is not perceived to be best practice) (Bellotti and Crook 2009).

Zhang and Thomas (2009) apply a classification tree algorithm first to segment unsecured personal loans that defaulted for different reasons. They then use linear regression and survival analysis to estimate LGD. This is beneficial because survival models are able to handle censored data, so loans which are still in the recovery process can be used, instead of having to wait until the recovery process is (nearly) complete (Zhang and Thomas 2012).

Many banks opt to use the single-step approach in mortgage portfolios when there does not seem to be a material improvement in performance when using the two-step model. Li et al. (2014) show that more complex parametric methods (two-step, inflated beta, Tobit, censored gamma, and two-tier gamma regressions) perform similarly to less complex methods such as standard linear regression and single stage Fractional Response Models (FRMs) (Li et al. 2014).

The four families of models examined and compared in this chapter are:

- Fractional response models (Papke and Wooldridge 1996) (Papke and Wooldridge 2008).
- Random Forest (Breiman 2001).
- XGBoost (Chen and Guestrin 2016).
- A linear model.

As mentioned above, variations of the Fractional Response Model (FRM) are one of the more commonly used methods at top tier banks in the US, Canada, and Europe, particularly the fractional logistic regression, which is a special case of FRM. While using random forests to model LGD in credit risk has been examined in academia, to the best of this author's knowledge and at the time of writing, it has yet to be compared with FRMs, nor has it been compared using a mortgage portfolio. Similarly, while stochastic gradient boosting has been reviewed for LGD credit risk modeling (Sun and Jin 2016), its variant Extreme Gradient Boosting (XGBoost) has yet to be examined for LGD in credit risk modeling. These three models are each discussed in greater detail in the following subsections.

3.3.3 Fractional Response Models

FRMs are used to model fractional (or proportional) response variables, which take on values in the standard unit interval [0, 1]. They were first introduced in Papke & Wooldridge's (1996) seminal paper, which includes an application to 401(k) participation rates. For loan i, the approach requires an assumption of the functional form of the dependent variable y, given a set of predictors x:

$$E(y_i|x_i) = G(x_i\beta), \tag{25}$$

where

- $G(\cdot)$ is a non-linear function which typically takes on any cumulative distribution function satisfying $0 \leq G(\cdot) \leq 1$,
- $0 \leq y_i \leq 1$ (which in our case is the LGD),
- x_i is a $1 \cdot k$ vector of the independent variables, and
- β is a vector of parameters to be estimated.

The logit model, where the Cumulative Distribution Function (CDF) is the logistic function, is a common choice. Other popular choices are the cauchy, probit (standard normal CDF), and the complementary loglog, which is asymmetric, and is used when the probability of an event is very small or very large (Ramalho, Ramalho, and Murteira 2011). Note that the logit, probit, and cauchy are (vertically) symmetric around the point 0.5 and approach 0 and 1 at the same rate. Papke & Wooldridge propose a quasi-likelihood method (QLM) based on the Bernoulli log-likelihood function which is given by the following equation:

$$LL_i(\beta) = y_i\text{log}[G(x_i\beta)] + (1 - y_i)\text{log}[1 - G(x_i\beta)]. \tag{26}$$

The Bernoulli distribution is a member of the linear exponential family (LEF) and, as such, β can be estimated by:

$$\hat{\theta} = \arg\max \sum_{i=1}^{N} LL_i(\theta), \tag{27}$$

which is consistent and asymptotically normal (Ramalho, Ramalho, and Murteira 2011). This form is often implemented as a weighted binary logistic regression. A historically popular alternative was to model the log-odds ratio as a linear function:

$$E\left(\log\frac{y}{(1-y)}|x\right) = x\beta. \tag{28}$$

However, this only works if y lies strictly between zero and one. Adjustments are possible, but not ideal (see Papke & Wooldridge 1996). This is particularly the case when there is a large concentration of observations at the bounds, which is certainly the case for LGD values.

This chapter focuses on Papke & Wooldridge's (1996) QLM method mentioned above due to its success and popularity in both academia and industry. The binary component of the two-part model is estimated using logistic regression maximum likelihood, which was chosen because it is by far the most commonly used approach in the industry and will be most familiar and relevant to practitioners.

3.3.4 Random Forests

For a detailed overview of the Random Forest technique, the reader is referred to Chapter 2. The primary difference with the Random Forest technique here, is that the model is estimating a continuous variable that is bounded between 0% and 100%, LGD, whereas the Random Forest model in Chapter 2 is estimating a binary outcome, default vs non-default. An important thing to consider when estimating LGD (or any continuous variable) is that by design Random Forests are unable to produce estimates of y which are outside the range of y in the training dataset. As such, it is important that if a Random Forest model is chosen, there should be a sufficient amount of y values across the full range of y the modeler would expect to see in the training dataset. In the case of modeling LGD, the training dataset should have values spanning the unit interval [0, 1].

3.3.5 XGBoost

This subsection provides a brief overview of Extreme Gradient Boost (XGBoost) (Chen and Guestrin 2016), which is a special case of boosting or "Adaptive Resampling and Combining" (ARCing). As described in greater detail in Chapter 2, XGBoost is a more recent variation of gradient boosting and has received considerable attention in recent years due to its success in machine learning competitions. Gradient boosting first originated with AdaBoost (Freund and Schapire 1997), which performs coordinate descent while greedily minimizing an exponential loss function (Breiman 1998), (Breiman 1999), (Friedman 2001), (Rätsch, Onoda, and Müller 2001), (Duffy and Helmbold 1999), (Mason et al. 1999). Friedman later generalized this exercise to the optimization of any loss function using gradient descent (Friedman 2001), called gradient boosting. When each trained ensemble is performed on a subset of the training dataset, it is referred to as stochastic gradient boosting (Friedman 2002), which has the benefit of potentially improving the generalization of the model's performance. Boosting has some similarities to bagging; however, one primary difference is that while the training stage for bagging is done in parallel, boosting is done sequentially.

Finally, gradient boosting optimizes $f(x)$ and makes use of its gradient $f'(x)$, whereas XGBoost optimizes $f(x)$ and makes use of both its gradient $f'(x)$ and its second derivative ("hessian" in a multivariate setting) $f''(x)$), and includes L1/L2 regularization. Similar to other tree-based algorithms, XGBoost also struggles with extrapolation outside the training dataset, so it is only good at making predictions on data on which it has previously been trained on.

3.3.6 Splines

This subsection reviews regression splines, which are commonly used in the industry to capture non-linear effects of continuous explanatory variables in statistical regression analysis. Splines are made up of two or more points, called ‘knots’, within a data range, which are then connected using polynomial functions of different order. What defines the type of splines is the type of polynomial, number of knots, the knot placement, and whether to include a penalty function. Splines including a penalty function are called smoothing splines and use the data points themselves as potential knots. Splines omitting a penalty function are called regression splines and place the knots at equidistant/equiquantile (varying widths but equally populated) points (Racine 2012). This chapter omits the penalty function, as some practitioners believe it adds a degree of additional complexity that might be unappealing in an applied setting (Perperoglou et al. 2019). A common knot placement is the ‘quantile’ knot sequence, which places the interior knots at the quantiles of the variables’ empirical distribution. This is done instead of spacing the knots at equal distances. For a fixed knot sequence and fixed polynomial degree d, the spline can be written in the following basis function format:

$$f(x) = \sum_{k=1}^{K+d+1} \beta_k B_k(x), \tag{29}$$

where B_k are a set of basis functions defining the vector space, K is the number of interior knots, and β_k is the respective spline coefficients. Three of the more popular spline basis are the truncated power series basis, the B-spline basis, and the natural spline basis. As an example, using equation (29), a cubic spline truncated power series basis with three knots τ_1, τ_2, τ_3 is presented in the following equation:

$$f(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \beta_4 (x - \tau_1)^3 + \beta_5 (x - \tau_2)^3 + \beta_6 (x - \tau_3)^3. \qquad (30)$$

While the truncated power series basis is conceptually straightforward, its form can lead to severe rounding errors with powers of large numbers (Hastie, Tibshirani, and Friedman 2009). B-splines (De Boor 2001) are a computationally efficient and equivalent (equivalent in the sense that the two bases span the same set of functions) representation of the truncated power basis with numerical stability and serve as a good alternative. B-splines are based on the following knot sequence:

$$\xi_1 \leq ... \leq \xi_d \leq \xi_{d+1} < \xi_{d+2} < ... < \xi_{d+K+1} < \xi_{d+K+2} \leq \xi_{d+K+3} \leq ... \leq \xi_{2d+K+2}, \quad (31)$$

where $\xi_{d+2} = \tau_1$, ..., $\xi_{d+K+1} = \tau_K$ are the inner knots, and ξ_{d+1} and ξ_{d+K+2} are the lower and upper boundary knots, respectively (Perperoglou et al. 2019). The additional arbitrary end knots are required for the Cox de Boor recursion formula, which is how B-splines are constructed. It is customary to set the arbitrary end knots to the boundary knots (Hastie, Tibshirani, and Friedman 2009). The B-spline basis $B_k^d(x)$ (degree d and knot(s) k) is defined by the following recursive formula:

$$B_k^0(x) = \begin{cases} 1 & \xi_k \leq x < \xi_{k+1} \\ 0 & \text{otherwise} \end{cases} \qquad (32)$$

for degree zero, and

$$B_k^d(x) = \frac{x - \xi_k}{\xi_{k+d} - \xi_k} B_k^{d-1}(x) - \frac{\xi_{k+d+1} - x}{\xi_{k+d+1} - \xi_{k+1}} B_{k+1}^{d-1}(x) \tag{33}$$

for $d > 0$, where $k = 1, 2, ..., K + d + 1$, and $0/0 = 0$. It is essentially taking a weighted average of the $B_k^{d-1}(x)$ and $B_{k+1}^{d-1}(x)$ functions. B-splines/polynomials have the limitation that the variability (standard errors) of the predictions can increase substantially at the boundaries of the data inputs. To address this issue, natural splines enforce $f''(x) = f'''(x) = 0$ at each boundary such that the tails beyond the boundary knots are linear.

When choosing the spline degree, cubic splines are the standard choice, since polynomials with order greater than 3 are often indistinguishable, and orders of 1 and 2 are often considered too “jagged” (Perperoglou et al. 2019). Hastie et al. state that there is seldom any good reason to go beyond a cubic spline unless one is interested in smooth derivatives (Hastie, Tibshirani, and Friedman 2009). However, there are situations where a higher degree might be more appropriate, depending on the Data Generating Process (DGP). Increasing the number of knots might help improve fit, though this runs the risk of overfitting the data and increasing the variance, and may suffer from poor generalization. On the other hand, having too few knots may result in a restrictive function with more bias.

3.4 The Data

The dataset chosen for the empirical component of this chapter is Freddie Mac’s Single-Family Fixed Rate Mortgage Loan Performance dataset, which is similarly used in Chapter 1 and 2. Two datasets are combined to build the full dataset; one containing acquisition data, and another which monitors on a monthly basis the performance data. The full dataset contains mortgages originated between January 1st, 1999 and June 30th, 2019. The dataset is updated every quarter to include newly acquired mortgage loans, as well as any updates observed in performance. Only loans which defaulted and have actual loss information are used in the final LGD dataset. In addition, to allow for sufficient time for Loss Components (expenses and proceeds) to be recorded, Freddie Mac implements a 90-day lag based on the Zero Balance Date (the Zero Balance Date is the date the loan's balance was reduced to zero. The denominator in the LGD calculation is the total amount of UPB remaining on the loan immediately prior to this) (Freddie Mac 2020). The resulting dataset has 19,864 observations. Federal Housing Finance Agency (FHSA) Housing Price Index (HPI) data at the Three-Digit ZIP code level (Federal Housing Finance Agency 2020) are included, as well as unemployment data from the Bureau of Labor Statistics (Bureau of Labor Statistics 2020).

The data are split into an in-sample training dataset, an OOS test dataset, and an OOT dataset. We hold out two years of OOT data, from May 2017 to May 2019. The data from January, 1999 to May 2017 is split 70% for the training dataset and 30% for the OOS testing dataset. The training, OOS, and OOT datasets have 12,227, 5,240, and 2,397 observations, respectively.

In practice, as more data become available, banks will continue to monitor their model's performance on the newly observed OOT datasets. The frequency of monitoring depends on things such as the model's materiality, complexity, and purpose, among other things. However, it is common to monitor on a quarterly or bi-annual basis for e.g., CECL, CCAR, Basel capital, and IFRS-9 models.

Actual Loss data, which are used to calculate LGD, are calculated using the following equation (Freddie Mac 2020):

$$\begin{aligned}\text{Actual Loss} = {} & (\text{Default UPB - Net Sale Proceeds}) + \\ & \text{Delinquent Accrued Interest - Expenses -} \\ & \text{MI Recoveries - Non MI Recoveries,}\end{aligned} \tag{34}$$

where

$$\begin{aligned}\text{Delinquent Accrued Interest} = {} & (\text{Default UPB - Non Interest bearing UPB})\times \\ & \text{min(Current Interest Rate - 0.35,} \\ & \text{Current Interest Rate - Servicing Fee})\times \\ & \text{(Months between last Principal \&} \\ & \text{Interest paid to date and zero balance date})\times \\ & 30/360/100.\end{aligned} \tag{35}$$

Note that the 35 bps is used as a proxy for the servicing fee when the servicing fee is not available. The 30/360 is an assumption that each month has 30 days and that the calendar year has 360 days. It is then further divided by 100 to convert it from a percentage to a decimal. Using Actual loss, LGD is calculated in the following way:

$$\text{LGD} = \frac{\text{Actual Loss}}{\text{Defaulted UPB}}, \tag{36}$$

where Defaulted UPB is the amount of total UPB remaining on the loan immediately prior to default, or the Exposure at Default (EAD).

For modeling purposes, LGD is often floored and capped at 0% and 100%. In some cases, LGD values above 100% can occur, which would mean the losses are beyond the initial investment from the FI. This could occur if the asset has depreciated in value (e.g., a loss of principal, property damages, etc.) or if workout expenses are excessive. Similarly, LGD values can also be negative if, for instance, the property has significantly appreciated in value since the FI's initial investment. If the percentage of LGD values below or above 0% and 100% are low (say, less than 10%), then they may be viewed as outliers. Some may argue that outliers should be removed from a dataset. However, practitioners are often interested in determining whether a loan will result in a 0% loss or a 100% loss, and values below and above 0% and 100%, respectively, contain valuable information that likely can help train the model to identify these loans. When the majority of defaulted loans LGD fall within the 0% to 100% range, practitioners will often prefer estimates to be restricted to this range as well, and developing models that predict values outside this range may result in estimates that are not in line with industry accounting, finance, and/or risk departments' expectations.

The distribution of LGD in this dataset is shown in Figure 24, where the dotted lines indicate 0% and 100%. There are 6.8% and 6.2% loans, which fall above and below the upper and lower bounds, respectively. Only 0.48% of LGD values are above 1.5, and 0.08% are above 2. The maximum LGD is 6.63, which is causing the long tail observed in Figure 24. After introducing the lower and upper bound of 0% and 100%, respectively, the first quantile, second quantile, and median of the dataset are unchanged. The mean has a minor change from 45.18% to 44.11%. This implies the characteristics of the distribution are largely preserved after the floor and cap are implemented. For the reasons discussed above, a lower and upper bound of 0% and

100% are used.

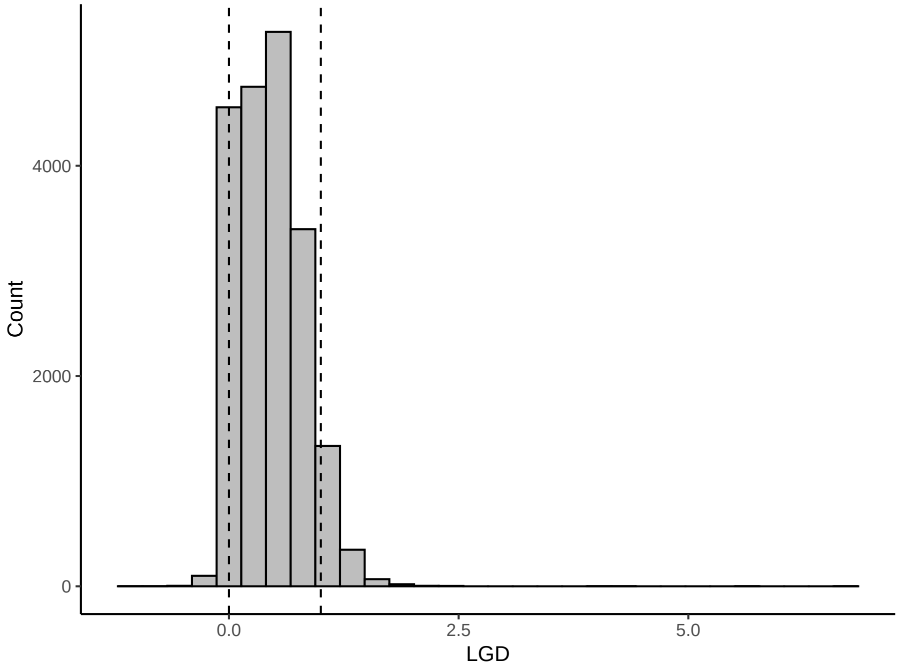

Figure 24: Loss Given Default Distribution

Figure 25 shows the distribution with the 0% and 100% floor and cap. The bimodal distribution commonly observed in LGD values is seen in this figure. For the remainder of this chapter, any references to LGD will pertain to the floored and capped LGD, unless explicitly stated otherwise.

Table 9 shows the number of LGD observations, and the LGD median, mean, and standard deviation by the default year. Since the dataset consists of mortgage loans originating in 1999, there are very few defaulted loans in the earlier years, which is expected in mortgage portfolios. We see a spike in defaults in the recessionary years after 2009 which is also in line with industry expectations and economic theory. For example, recessions typically have higher unemployment, which results in borrowers having less disposable income and therefore are more likely to miss mortgage amortization payments, leading ultimately to default. The increase in unemployment rates

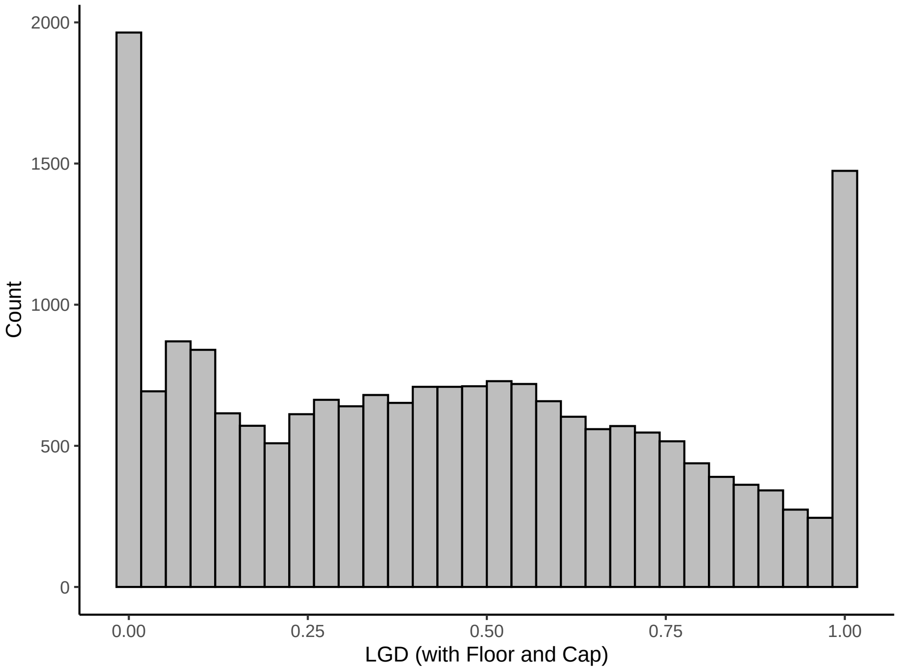

Figure 25: Loss Given Default Distribution with 0 percent Floor and 100 percent Cap

typically lag behind the observed increase in defaults, as borrowers typically have enough disposable income to continue to make amortization payments for a quarter or two. We also see an increase in the LGD mean and median during these years. The lower LGD rates in more current years are likely due to accounts resolving faster and receiving more recoveries due to strong housing price growth. It is common to remove some of the more recent years of LGD data in order to allow sufficient time for the defaulted loans to resolve.

Alternatively, Table 10 shows the number of LGD observations, and the LGD median, mean, and standard deviation by the loan origination year. As we can see, the number of defaults decreases significantly in more recent years, which is in line with industry expectations, given newly originated loans typically have a lower likelihood of defaulting in the earlier years of their origination. Note that in both Table 9 and 10, the standard deviation is relatively high, which indicates that there is a wide range of

Table 9: LGD Summary Statistics, by Loan Default Year

Default Year	**Count**	**Median**	**Mean**	**Std**
2000	5	0.24	15.58	21.46
2001	55	0.88	7.47	15.32
2002	160	3.16	14.07	22.80
2003	244	5.24	15.41	22.39
2004	334	7.59	18.51	24.55
2005	342	15.12	23.07	25.28
2006	318	13.71	24.69	28.29
2007	308	15.64	26.32	29.21
2008	520	30.62	36.18	31.16
2009	1087	40.19	41.27	27.70
2010	2095	48.36	48.35	27.79
2011	2630	53.53	53.04	26.36
2012	2853	51.69	51.77	27.60
2013	2029	49.03	50.49	29.21
2014	1803	54.29	55.05	30.17
2015	1335	56.24	56.81	29.84
2016	1031	51.67	52.60	32.39
2017	1010	10.76	27.14	31.57
2018	1366	7.85	17.47	25.54
2019	339	6.71	13.21	21.51

Table 10: LGD Summary Statistics, by Loan Origination Year

Origination Year	**Count**	**Median**	**Mean**	**Std**
1999	490	11.56	26.99	32.90
2000	623	12.91	27.71	32.44
2001	738	28.84	37.68	35.25
2002	866	37.78	42.00	34.85
2003	1144	36.03	40.04	32.10
2004	1744	37.58	41.12	31.74
2005	2518	43.23	44.11	29.47
2006	3848	50.82	49.53	29.05
2007	4459	47.63	48.15	29.69
2008	2567	43.76	45.79	31.78
2009	440	34.31	39.05	29.48
2010	177	33.77	36.77	28.56
2011	93	19.77	28.87	29.97
2012	30	33.74	33.65	29.79
2013	37	14.94	18.23	19.27
2014	49	22.42	30.14	30.18
2015	21	9.25	19.16	23.40
2016	11	2.01	9.17	14.88
2017	7	2.49	6.23	11.00
2018	2	4.28	4.28	0.12

LGD values across these time horizons.

There also appears to be a clear upward trend in LGD until 2017. Figure 26 shows the average LGD percentage and the weighted LGD across time, by quarter. Both appear to follow a similar trend. In addition, Figure 26 also shows the trend of both the numerator (actual losses) and the denominator (balance remaining at default) in the weighted LGD calculation. Both actually have downward trends after the recession around 2012, which is as expected. The reason we see the continuous upward trend in the actual average LGD percentage and weighted LGD is that the denominator in the LGD calculation is decreasing at a faster rate than the numerator. In 2017, actual losses continue to decrease, and the balance at default begins to increase again, resulting in the decreasing LGD visible after 2017.

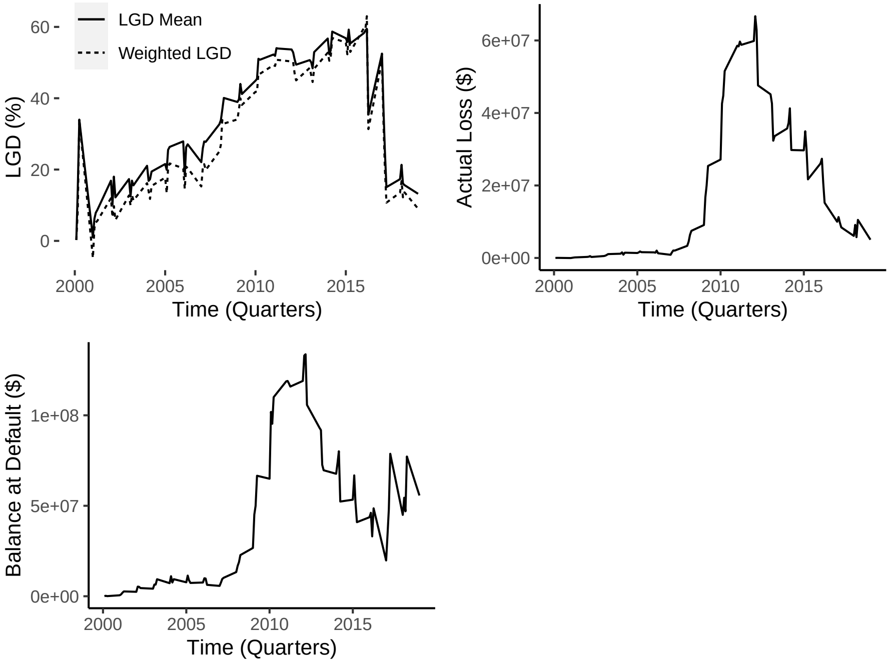

Figure 26: Balance Weighted Loss Given Default, Unweighted LGD, Actual Losses, and Balance at Default Trends

The box plots in Figure 27 and Figure 28 provide some insight into the relationships some of the categorical variables in the dataset have with LGD. The box plots in each figure are ordered by mean LGD from smallest to largest. It appears that Occupancy Status (Primary Residence = P, Investment Property = I, Secondary Home = S), First Time Home Buyer, Number of Units, Property Type (Condo = CO, Planned Unit Development = PU, Manufactured Household = MH, Single Family = SF, Co-op = CP), Loan Purpose (Purchase = P, Refinance Cash Out = C, Refinance No Cash Out = N)[10], and Number of Borrowers are all able to discriminate LGD characteristics. However, the boxplots all overlap, indicating that they do not perfectly discriminate.

[10] A Cash-out Refinance mortgage loan is a mortgage loan in which the use of the loan amount is not limited to specific purposes.

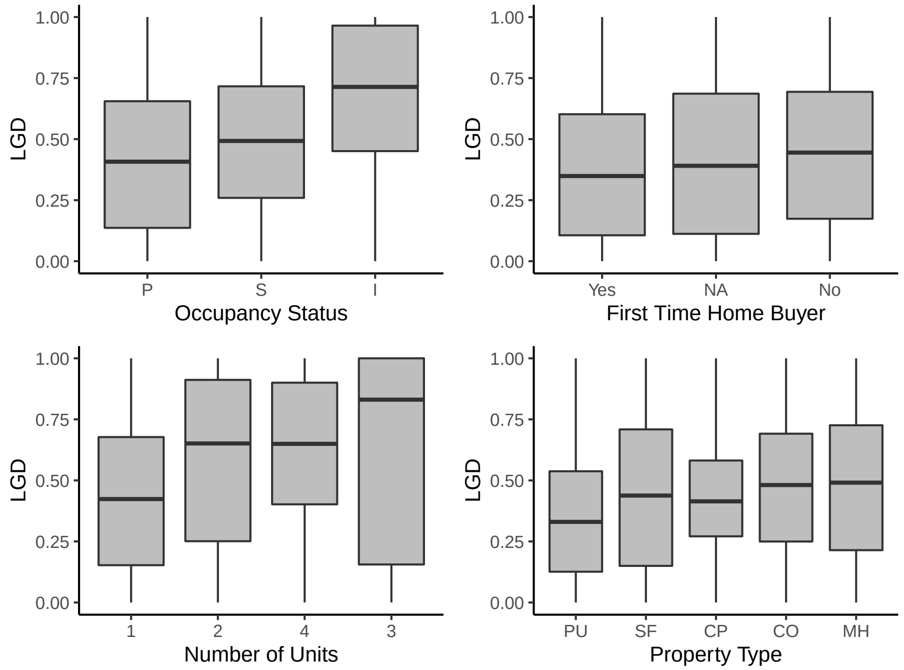

Figure 27: Boxplots of Important Variables and Loss Given Default

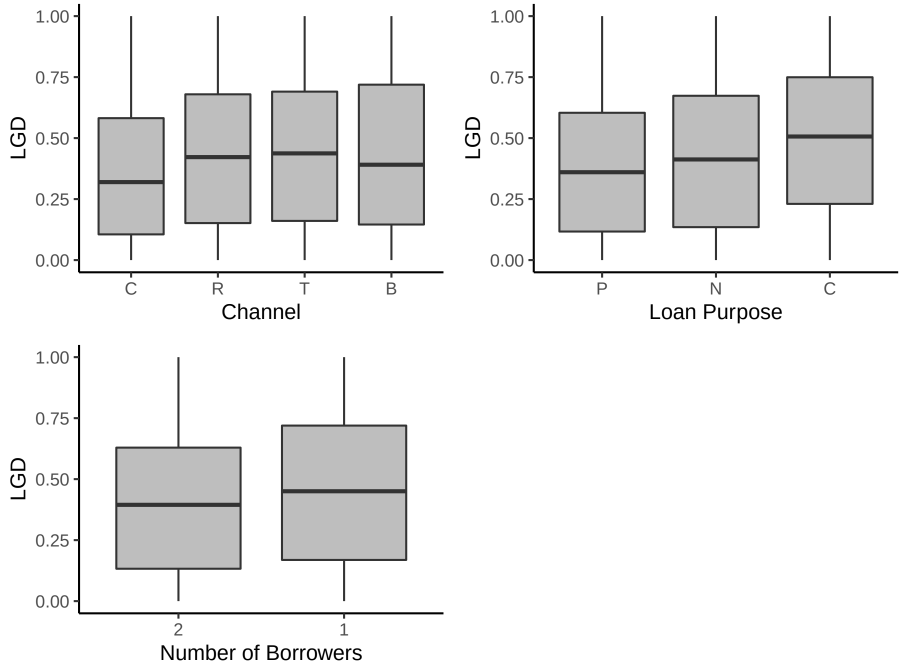

Figure 28: Boxplots of Important Variables and Loss Given Default (Continued)

In addition to the boxplot analysis on the categorical variables, the graphs in Figure 29 plot the average LGD across different numeric variables in the dataset. This helps visualize and interpret the relationship that these variables have with LGD as they increase. For example, we can see that as the change in HPI since origination (1-quarter lag) increases, LGD has a downward trend, which aligns with economic intuition. Similarly, as unemployment rate increases, LGD appears to increase. While the Pearson correlation only captures linear relationships, we have included it in each graph, despite there being non-linear relationships (such as loan age). The next section walks us through the model development process.

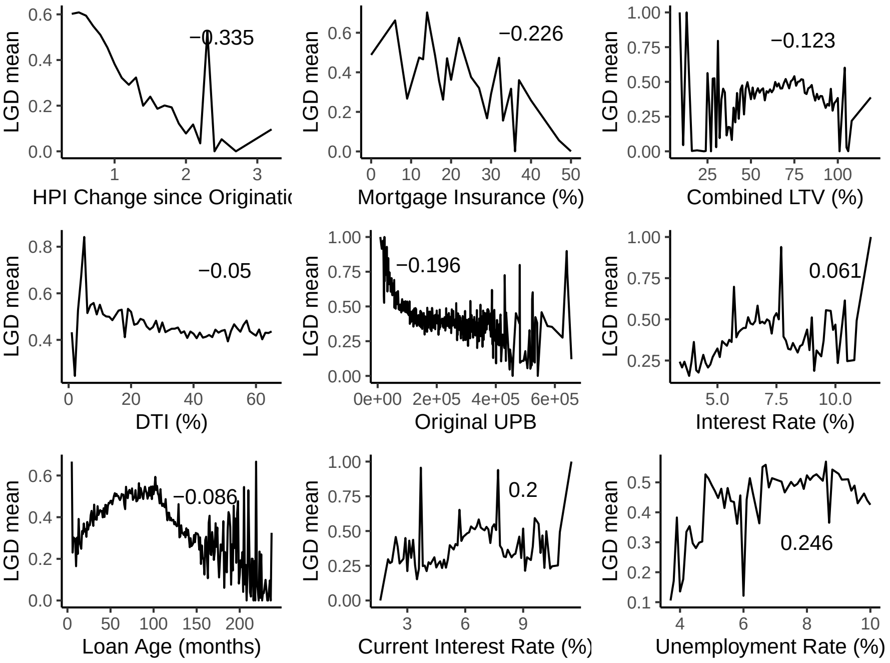

Figure 29: Numeric Variables Relationship with Loss Given Default

3.5 Model Development

This section provides an overview of the development process undertaken for each model, summarizes the results, and compares their performance. When comparing model performance, Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE) are used. Here, RMSE is calculated using the models predicted values and the test and training actual values. The lower the error rate, the better the model performance. In addition to model performance, this section attempts objectively to analyze and assess the ease of implementation, interpretability, and overall suitability of each methodology in applied credit risk LGD modeling.

3.5.1 Variable Selection

Each model begins with the same independent variable set, which is prescreened only to include variables in line with industry expectations and economic/finance theory. For the FRMs and the linear model, backward stepwise regression is used for variable selection. The random forest and XGBoost methods use the entire set of predictors in their models. Variable selection is the process of selecting the best set of predictors, which removes unnecessary noise, collinearity[11], overfitting, computation cost and improves interpretability. We expect the model to have intuitive and well-fit estimates, and we expect the variables to have coefficient signs that are in line with economic theory and statistical significance. It is common to keep the number of independent variables to an interpretable amount, such as 5-15.

B-splines are used for binning the continuous independent variables prior to variable selection for the FRMs. As discussed above, splines are a powerful way to capture non-linear relationships. For example, logistic regression assumes a linear relationship between the independent variable and the log odds. If the relationship is non-linear, then the regression may be mispecified. Splines are an effective way to address this issue.

[11] Multicollinearity is tested using the Variance Inflation Factor (VIF).

The dataset begins with 30 independent variables. Variables including future information are removed (e.g., recovery amounts, expenses, legal costs, taxes, etc.), and only variables that are in line with industry expectations and economic/finance theory are kept. Note that balance at default information is not included in the models, since this would not be known during production in practice. This results in 18 independent variables. These variables are listed in the appendix (the descriptions are taken directly from Freddie Macs data user guide) (Freddie Mac 2020).

3.5.2 Results

This section presents the results and compares the performance of the different LGD modeling methods discussed above. The variable selection process for each model begins with the same 18 independent variables. The random forest and XGBoost models use all 18 independent variables. The RMSE and MAE model performance on the training, OOS, and OOT datasets is presented in Tables 11, 12, and 13, respectively. Note that RMSE and MAE are calculated using the predicted and actual values from the training and test datasets. Each table is ordered by RMSE from lowest to highest. As we can see, the XGBoost model performs best in both the training and OOS test dataset, followed by the random forest model. The naive linear model is also included for reference, and ranks third in terms of MRSE and MAE performance in the training and OOS dataset. The four fractional response models' performance closely follows, and are all very similar to each other in the training and OOS test dataset performance. A fractional logistic model without splines is also included as a reference point.

Interestingly, the FRMs with B-splines all perform better than XGBoost, Random Forest, and the Linear Model in the OOT dataset, with the 2-step FRM performing best. Each model's OOS and OOT averaged predicted LGD values by quarter are plotted against actual values in Figure 30 and Figure 31 to better understand where the strengths and shortfalls may be occurring, and to determine whether the models are able to capture the trends. The predicted LGD values are the dashed lines and the actual LGD values are the solid lines. Everything to the right of the solid vertical black line (May 2017) is the OOT test dataset, and everything to the left is the OOS test dataset. In this particular case, it appears that the FRMs with B-splines are better than the other models at capturing the peaks and troughs in both the OOS and OOT test datasets, which is important when choosing a model in credit risk departments.

The OOT test dataset succeeds in visualizing the shortfalls of using a linear model. As we can see, the linear model predictions fall well below 0% in the OOT test dataset (and also in the first quarter of the OOS test dataset), which results in the high model RMSE and MAE. As discussed above, this is not in line with industry expectations for predicted or observed LGD percentages. While a floor and cap could be assigned to the model forecasts (a floor of 0% and cap of 100%, as an example), this does not resolve the underlying weaknesses associated with using a linear model in this situation. Hence, the linear model is not recommended.

While the XGBoost and Random Forest models perform well on the training and OOS test datasets, one limitation discussed above is that these methods are limited to data they were trained on. As such, if the training dataset LGD values do not sufficiently span a full 0-100% range, interpolating and extrapolating LGD forecasts in OOS or OOT test datasets will not be adequately predicted. This situation is actually quite likely to occur in banking stress tests, where different recessionary macroeconomic scenarios are provided by regulators and the macroeconomic variables are then fed through the models, and it is expected that the models will adequately predict the respective LGD values in these hypothetical scenarios. The inability to do so will

Table 11: Model Performance on Training Dataset, Ordered by RMSE

Models	**RMSE**	**MAE**
XGBoost	0.2110	0.1688
Random Forest	0.2312	0.1860
Linear Model	0.2409	0.1940
2-Step FRM (logistic/logistic)	0.2455	0.1984
Fractional Probit	0.2461	0.1992
Fractional Logit (No Splines)	0.2463	0.1994
Fractional Logit	0.2476	0.1996
Fractional c-loglog	0.2482	0.2003

Table 12: Model Performance on OOS Dataset, Ordered by RMSE

Models	**RMSE**	**MAE**
XGBoost	0.2245	0.1785
Random Forest	0.2276	0.1827
Linear Model	0.2393	0.1915
Fractional Probit	0.2424	0.1954
2-Step FRM (logistic/logistic)	0.2427	0.1955
Fractional Logit	0.2431	0.1955
Fractional c-loglog	0.2438	0.1965
Fractional Logit (No Splines)	0.2440	0.1965

likely be heavily scrutinized by regulators. This is one argument for why models like XGBoost and Random Forest are not used in banking credit risk models.

With respect to the FRMs, it appears that the two-step method does add some lift in performance, but only slightly. This marginal improvement in performance is sometimes deemed unnecessary in practice when compared to a slightly more interpretable and easier-to-implement method like the one-step FRMs.

Table 13: Model Performance on OOT Dataset, Ordered by RMSE

Models	**RMSE**	**MAE**
2-Step FRM (logistic/logistic)	0.2643	0.1652
Fractional c-loglog	0.2680	0.1646
Fractional Logit	0.2705	0.1669
Fractional Probit	0.2726	0.1708
XGBoost	0.2911	0.2543
Random Forest	0.3184	0.2915
Fractional Logit (No Splines)	0.3673	0.3198
Linear Model	0.3707	0.2911

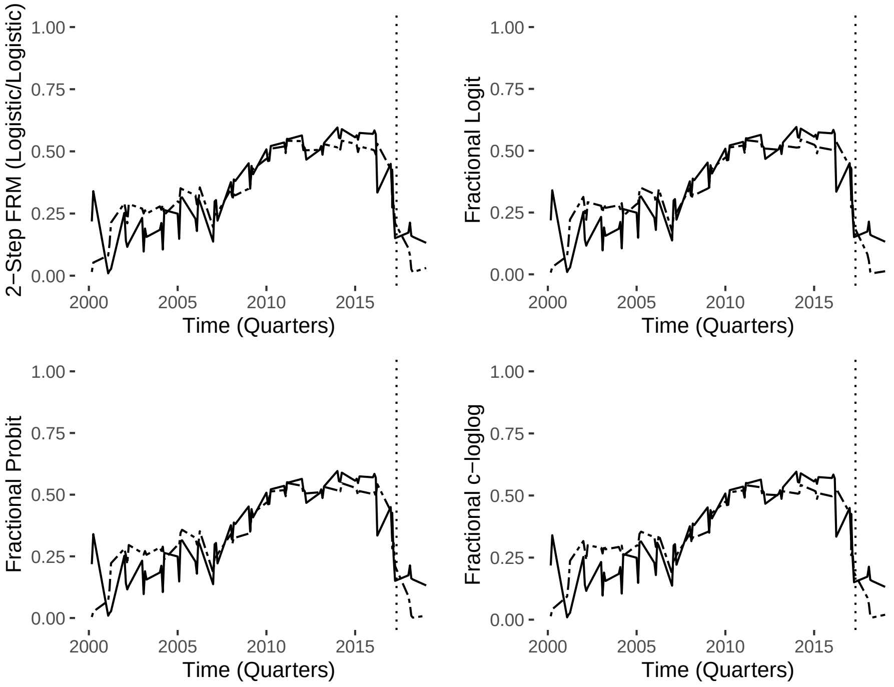

Figure 30: Model Predicted vs Actual Plots on OOS and OOT Data

3.6 Conclusion

This chapter compares various econometric models with the purpose of estimating and forecasting LGD percentages, which are bounded between 0% and 100%. In particular, we examine various one-step and two-step Fractional Response, XGBoost, Random

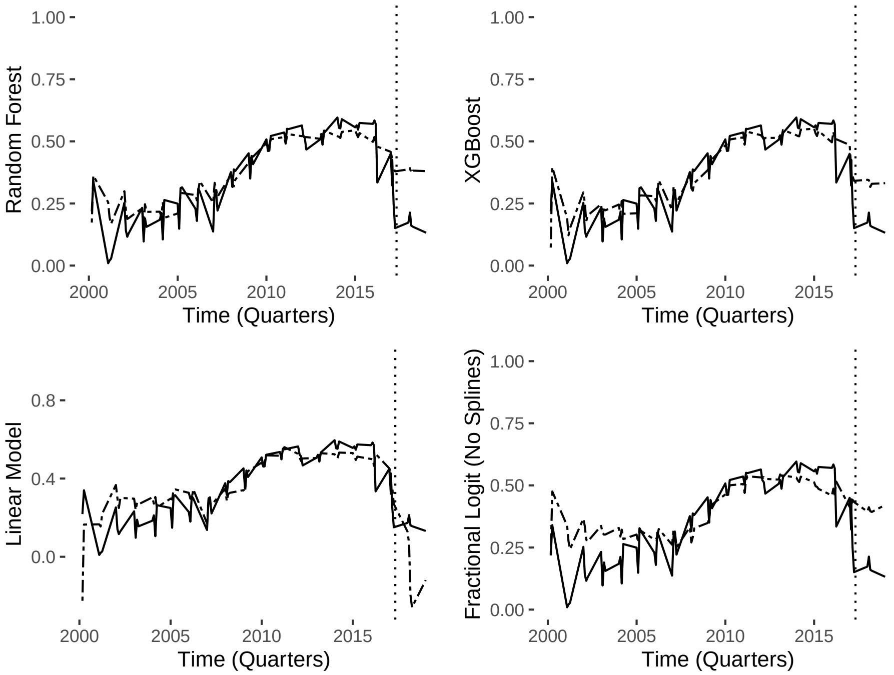

Figure 31: Model Predicted vs Actual Plots on OOS and OOT Data (Continued)

Forest, and linear models. Our findings indicate that the Fractional Response Models (FRMs) using splines have the lowest RMSE and MAE test statistics in the OOT test datasets, and appear to best capture peak and trough trends in the observed LGD data when compared to the other models. The two-step FRM performs slightly better than the one-step FRMs. However, the marginal improvement in performance might not be worth the added complexity for a practitioner, given how similar the one-step FRMs are in performance (in this particular dataset and example).

The XGBoost and Random Forest models have the lowest RMSE and MAE in the training and OOS test dataset, though their ability to generalize in the OOT test dataset seems to fall short compared to the FRMs. By design, XGBoost and Random Forests are limited in their ability to forecast outside the range of the dependent variable (LGD) in the training dataset, which means that if a practitioner intends to use one of these approaches, the training dataset dependent variable must sufficiently cover the range of values we expect and intend to see when forecasting, with sufficient data. In addition, tree-based methods might also not be appropriate for credit risk stress testing exercises, which are intended to feed recessionary-like macroeconomic scenarios (which might extrapolate in the independent variable domain) into the models to determine how overall expected losses and respective parameters like PD, LGD, and EAD will be impacted. The inability to do so may be heavily scrutinized by regulators.

The linear model is a widely used econometric method for estimating LGDs, despite the fact that its functional form is almost always misspecified. This particular dataset successfully highlights one major limitation: its inability to bound its estimates between 0% and 100%. This results in the linear model having the highest RMSE value in the OOT test dataset. While practitioners will often enforce a lower and upper bound of 0% and 100%, respectively, it is more suitable simply to choose a model designed to handle fractional values such as FRMs.

4 Appendix A

4.1 Weight of Evidence (WOE) Tests

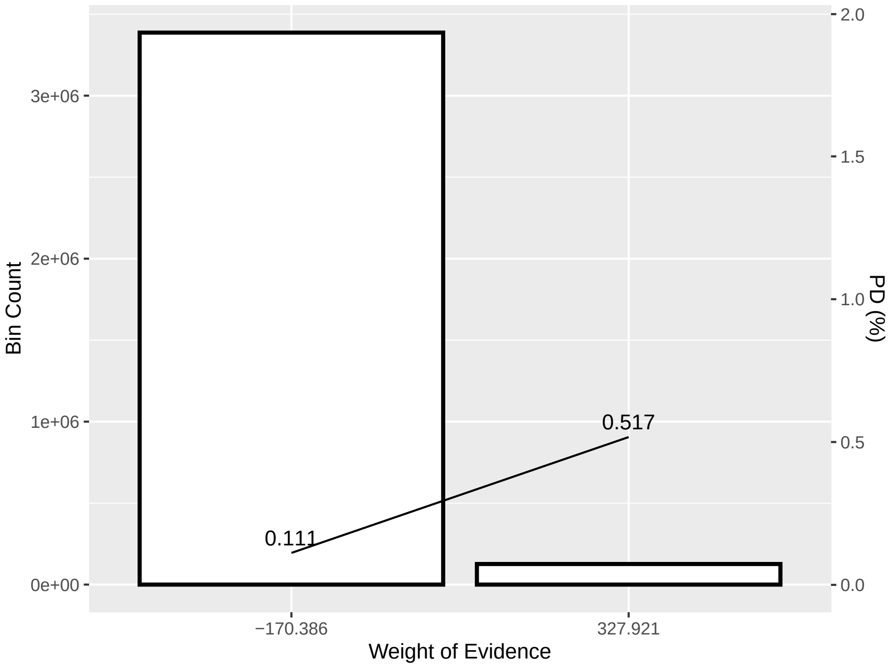

Figure 32: Delinquency Indicator Weight of Evidence Bin Count and Default Probabilities

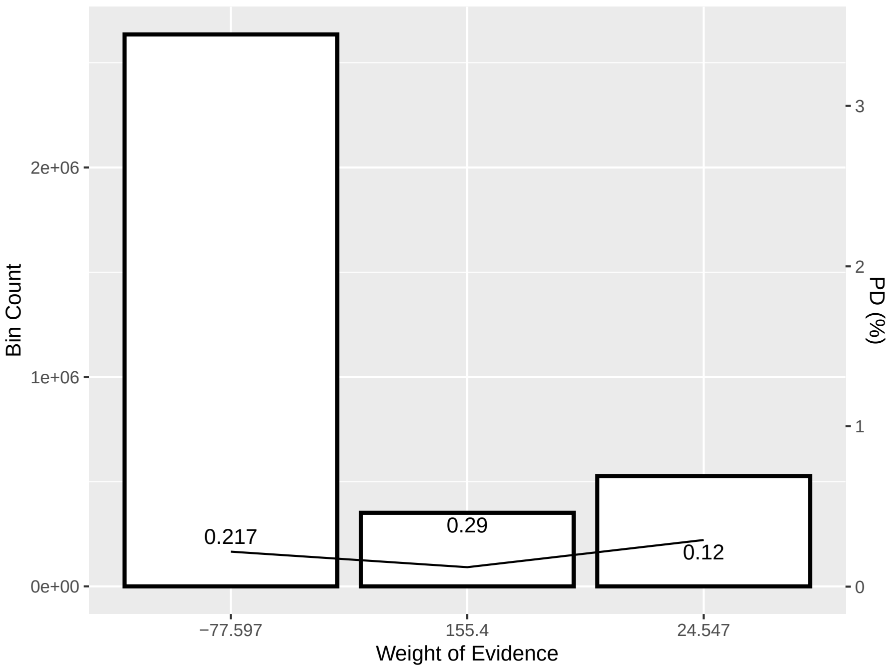

Figure 33: HPI Change Weight of Evidence Bin Count and Default Probabilities

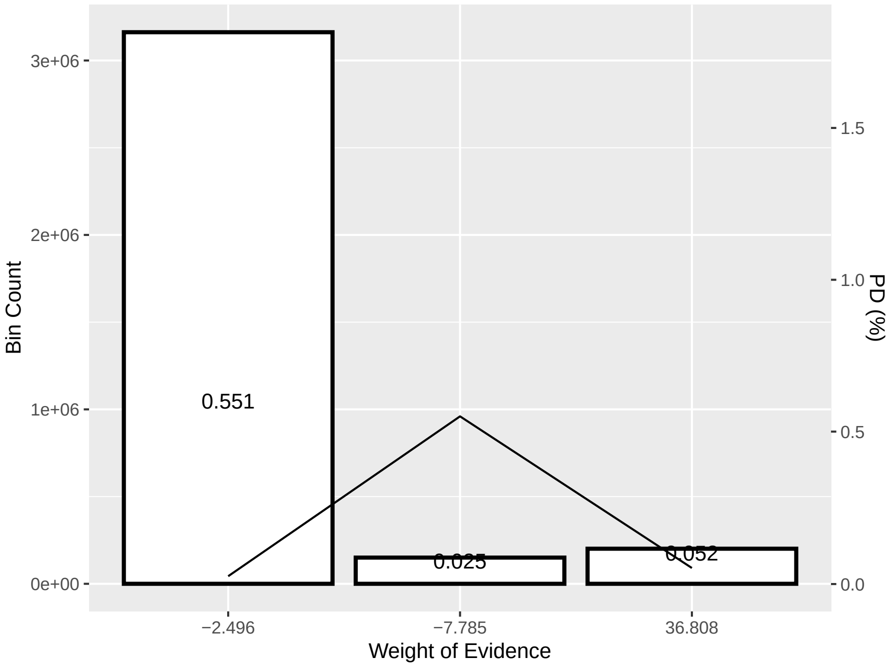

Figure 34: Occupancy Status Weight of Evidence Bin Count and Default Probabilities

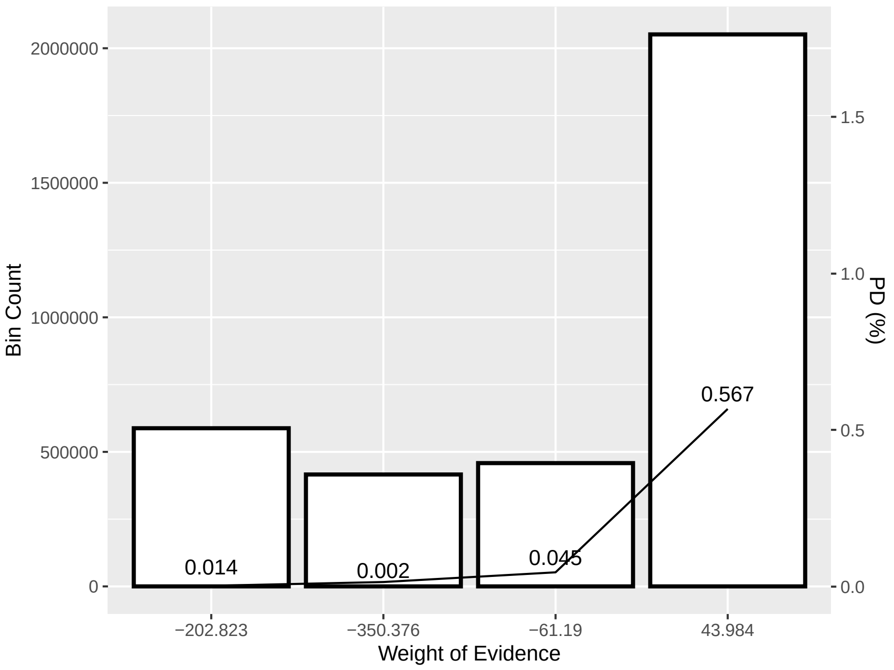

Figure 35: Interest Rate Weight of Evidence Bin Count and Default Probabilities

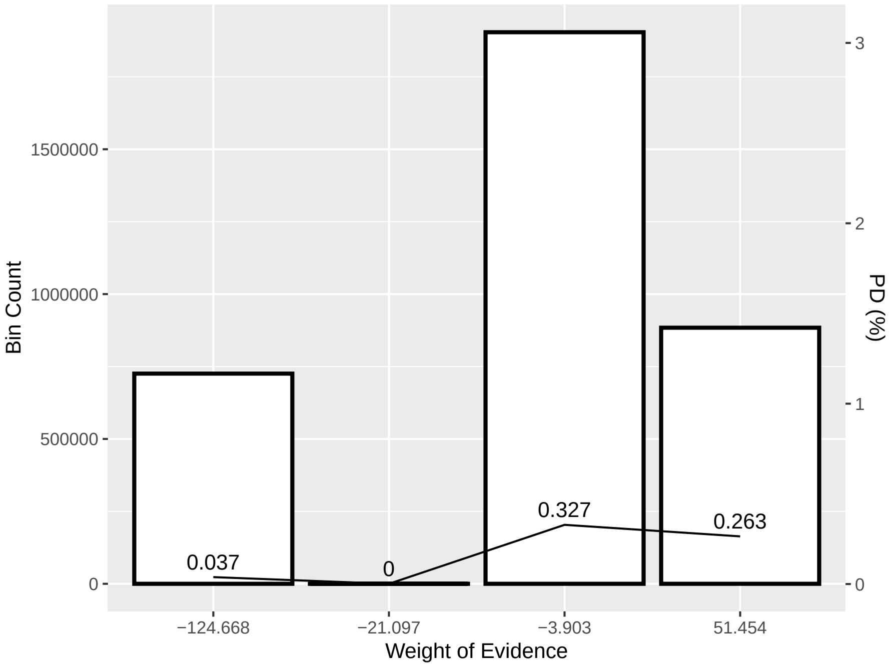

Figure 36: Original Combined LTV Weight of Evidence Bin Count and Default Probabilities

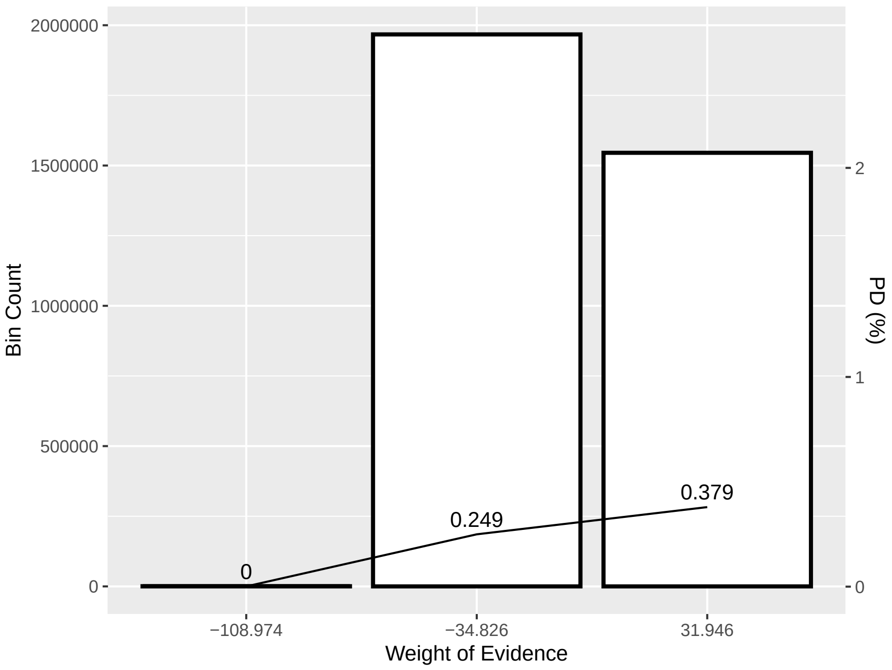

Figure 37: Number of Borrowers Weight of Evidence Bin Count and Default Probabilities

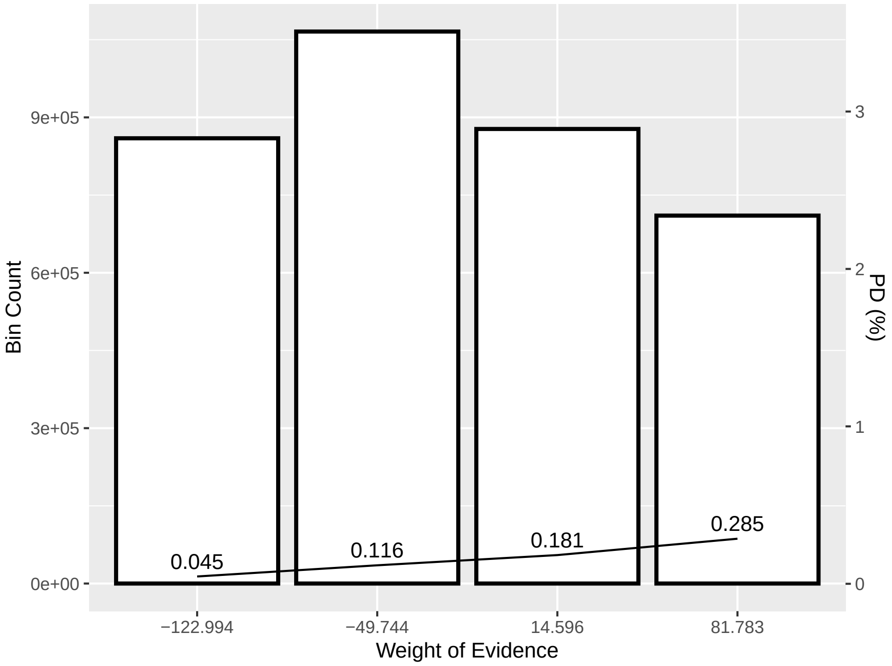

Figure 38: Credit Score (FICO) Weight of Evidence Bin Count and Default Probabilities

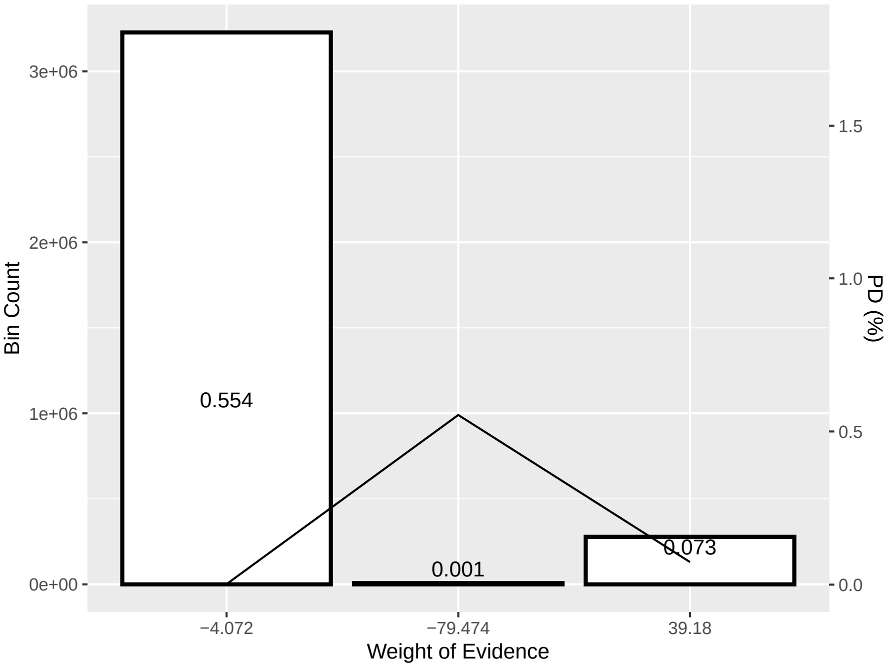

Figure 39: Property Type Weight of Evidence Bin Count and Default Probabilities

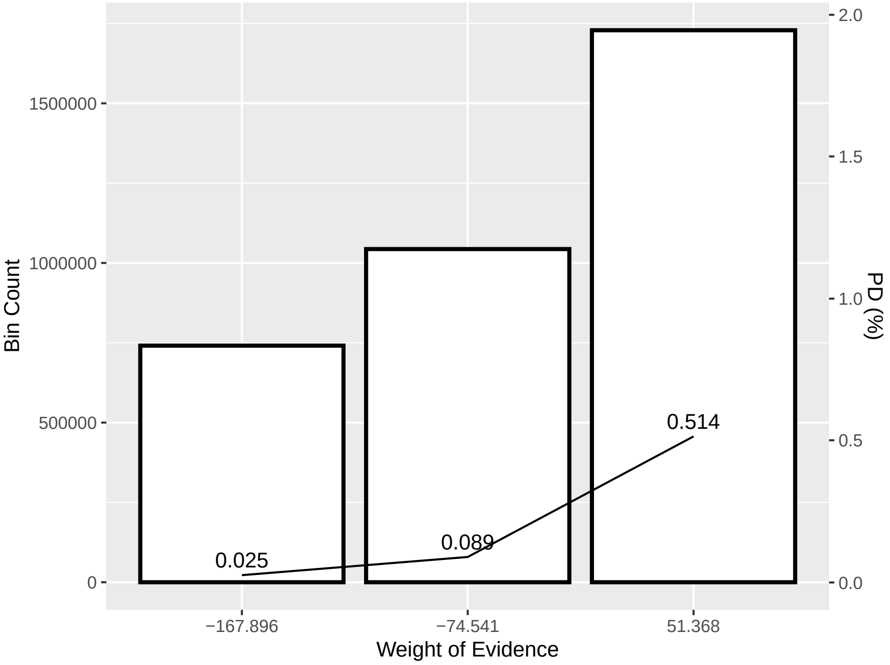

Figure 40: Loan Age Weight of Evidence Bin Count and Default Probabilities

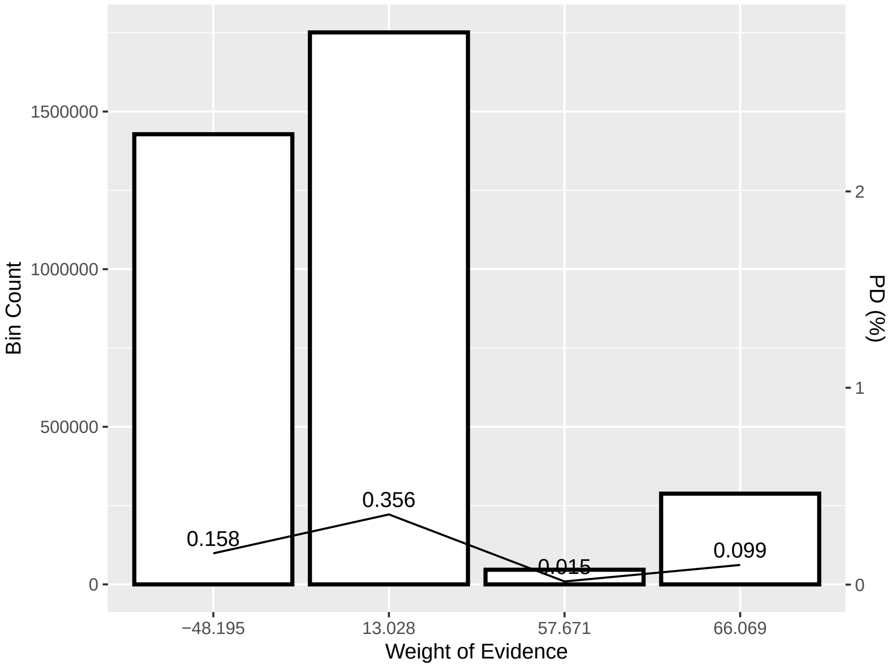

Figure 41: DTI Weight of Evidence Bin Count and Default Probabilities

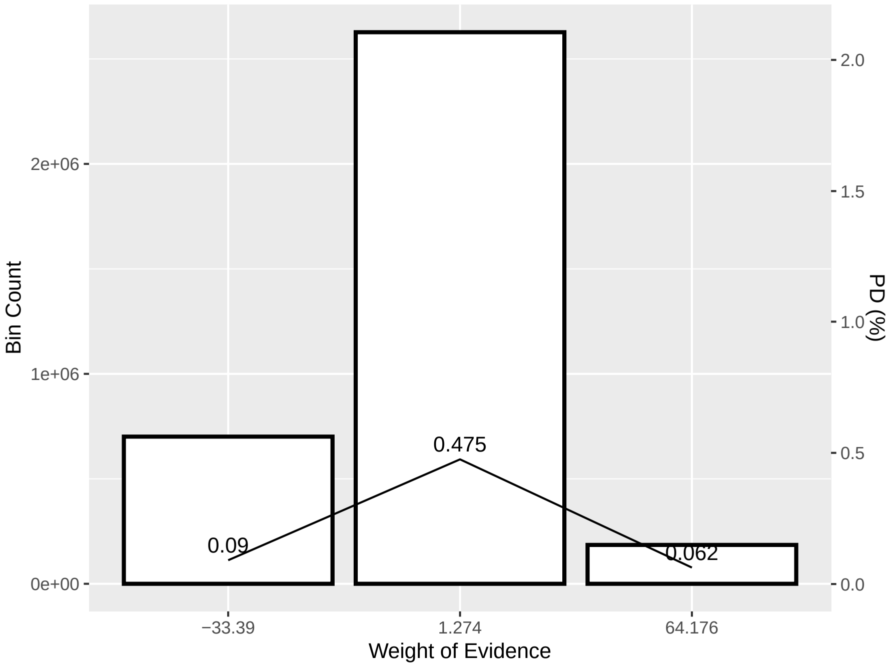

Figure 42: Original Value Weight of Evidence Bin Count and Default Probabilities

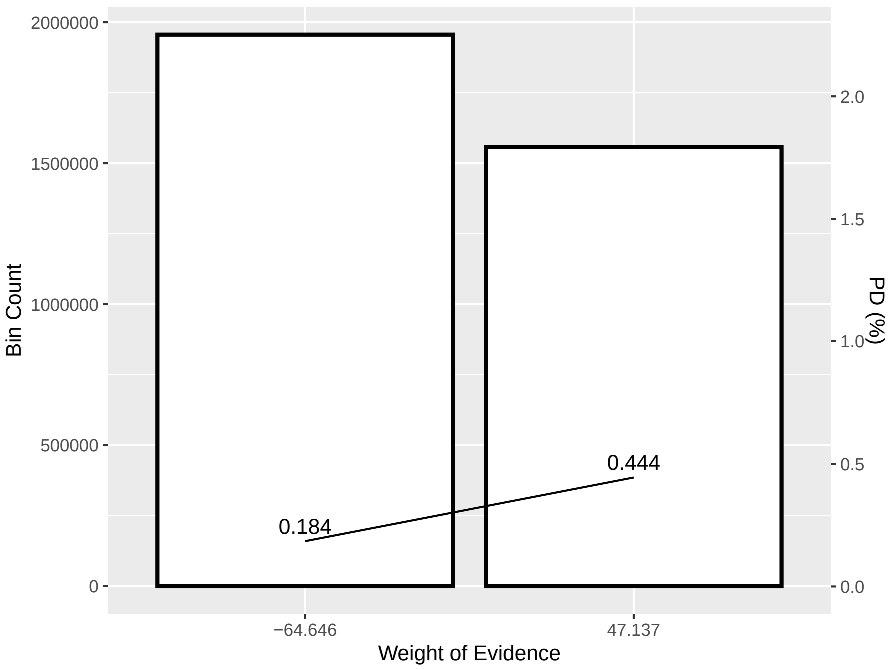

Figure 43: Unemployment Rate (3 month lag) Weight of Evidence Bin Count and Default Probabilities

4.2 Population Stability Index (PSI) tests

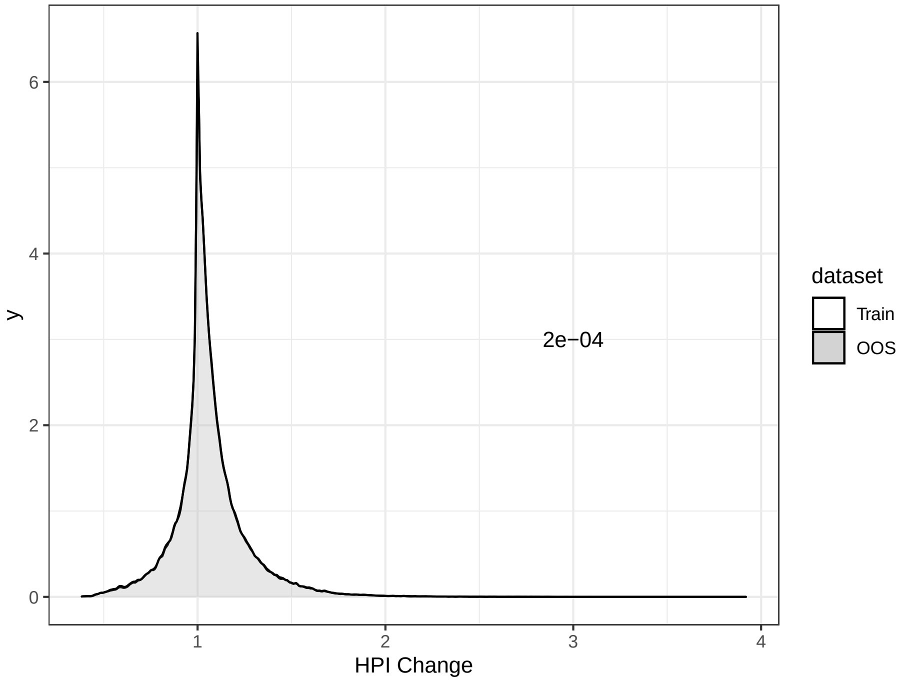

Figure 44: HPI Change Training and Out-of-Sample Dataset Bin Distribution, and Population Stability Index Test Result

5 Appendix B

Table 14: Candidate Variable List

Variable	**Description**
Credit Score	A number, prepared by third parties, summarizing the borrower's creditworthiness, which may be indicative of the likelihood that the borrower will timely repay future obligations. Generally, the credit score disclosed is the score known at the time of acquisition and is the score used to originate the mortgage.

Table 14: Candidate Variable List *(continued)*

Variable	**Description**
First Time Homebuyer Flag	Indicates whether the Borrower, or one of a group of Borrowers, is an individual who (1) is purchasing the mortgaged property, (2) will reside in the mortgaged property as a primary residence, and (3) had no ownership interest (sole or joint) in a residential property during the three-year period preceding the date of the purchase of the mortgaged property.
Mortgage Insurance Percentage	The percentage of loss coverage on the loan, at the time of Freddie Mac's purchase of the mortgage loan, that a mortgage insurer is providing to cover losses incurred as a result of a default on the loan
Number of Units	Denotes whether the mortgage is a one-, two-, three-, or four-unit property.
Occupancy Status	Denotes whether the mortgage type is owner occupied, a second home, or an investment property.

Table 14: Candidate Variable List *(continued)*

Variable	**Description**
Combined Loan to Value	In the case of a purchase mortgage loan, the ratio is obtained by dividing the original mortgage loan amount on the note date plus any secondary mortgage loan amount disclosed by the Seller by the lesser of the mortgaged property's appraised value on the note date or its purchase price. In the case of a refinanced mortgage loan, the ratio is obtained by dividing the original mortgage loan amount on the note date plus any secondary mortgage loan amount disclosed by the Seller by the mortgaged property's appraised value on the note date. If the secondary financing amount disclosed by the Seller includes a home equity line of credit, then the CLTV calculation reflects the disbursed amount at closing of the first lien mortgage loan, not the maximum loan amount available under the home equity line of credit. In the case of a seasoned mortgage loan, if the Seller cannot warrant that the value of the mortgaged property has not declined since the note date, Freddie Mac requires that the Seller must provide a new appraisal value, which is used in the CLTV calculation. In certain cases, where the Seller delivered a loan to Freddie Mac with a special code indicating additional secondary mortgage loan amounts, those amounts may have been included in the CLTV calculation.

Table 14: Candidate Variable List *(continued)*

Variable	**Description**
Debt to Income Ratio	Disclosure of the debt to income ratio is based on (1) the sum of the borrower's monthly debt payments, including monthly housing expenses that incorporate the mortgage payment the borrower is making at the time of the delivery of the mortgage loan to Freddie Mac, divided by (2) the total monthly income used to underwrite the loan as of the date of the origination of the such loan.
Original Unpaid Balance	The original Unpaid Balance on the loan.
Original Interest Rate	The original note rate as indicated on the mortgage note.
Channel	Disclosure indicates whether a Broker or Correspondent originated or was involved in the origination of the mortgage loan.
Property Type	Denotes whether the property type secured by the mortgage is a condominium, leasehold, planned unit development (PUD), cooperative share, manufactured home, or Single Family home.
Loan Purpose	Indicates whether the mortgage loan is a Cashout Refinance mortgage, No Cash-out Refinance mortgage, or a Purchase mortgage.
Number of Borrowers	The number of Borrower(s) who are obligated to repay the mortgage note secured by the mortgaged property.
Original Home Value	Value of the home at loan origination.

Table 14: Candidate Variable List *(continued)*

Variable	**Description**
Loan Age	The number of months since the note origination month of the mortgage.
Current Interest Rate	The current interest rate on the mortgage note, taking into account any loan modifications.
Unemployment Rate	The unemployment rate in the state of the loan.
Housing Price Index Change	The change in the HPI since the origination of the loan.

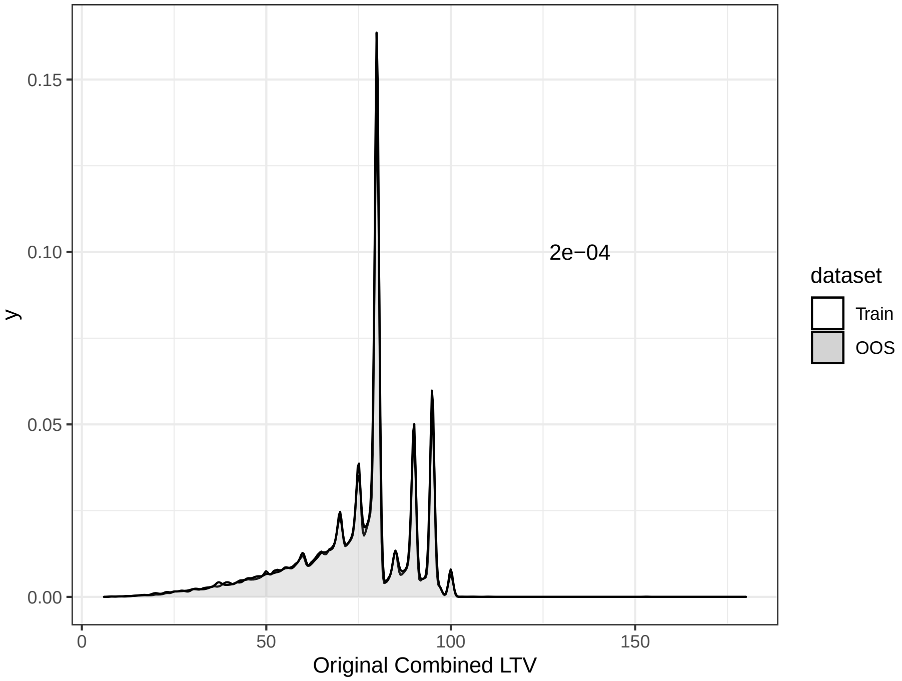

Figure 45: Original Combined LTV Training and Out-of-Sample Dataset Bin Distribution, and Population Stability Index Test Result

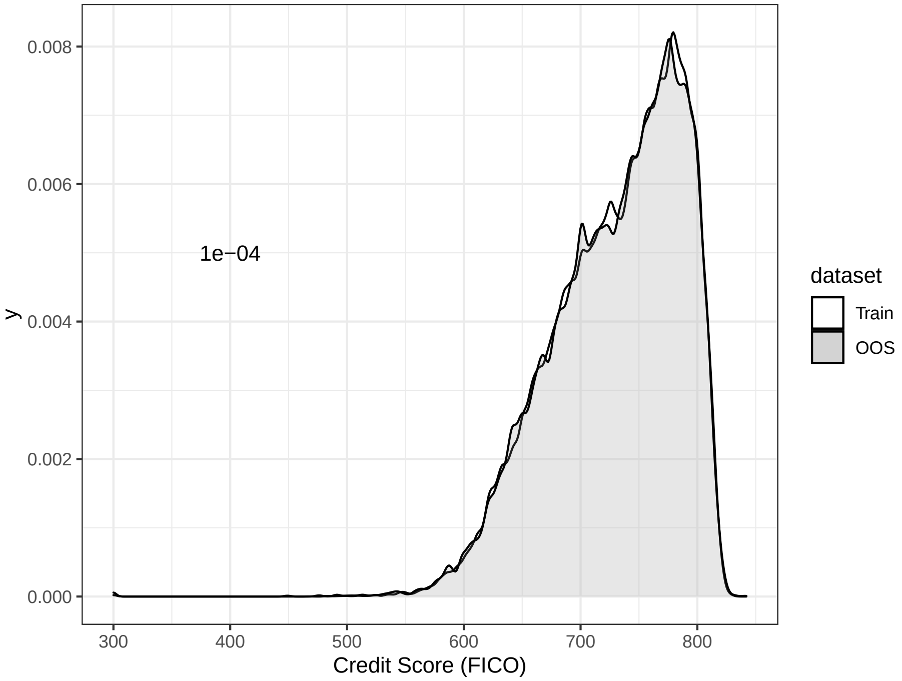

Figure 46: Credit Score (FICO) Training and Out-of-Sample Dataset Bin Distribution, and Population Stability Index Test Result

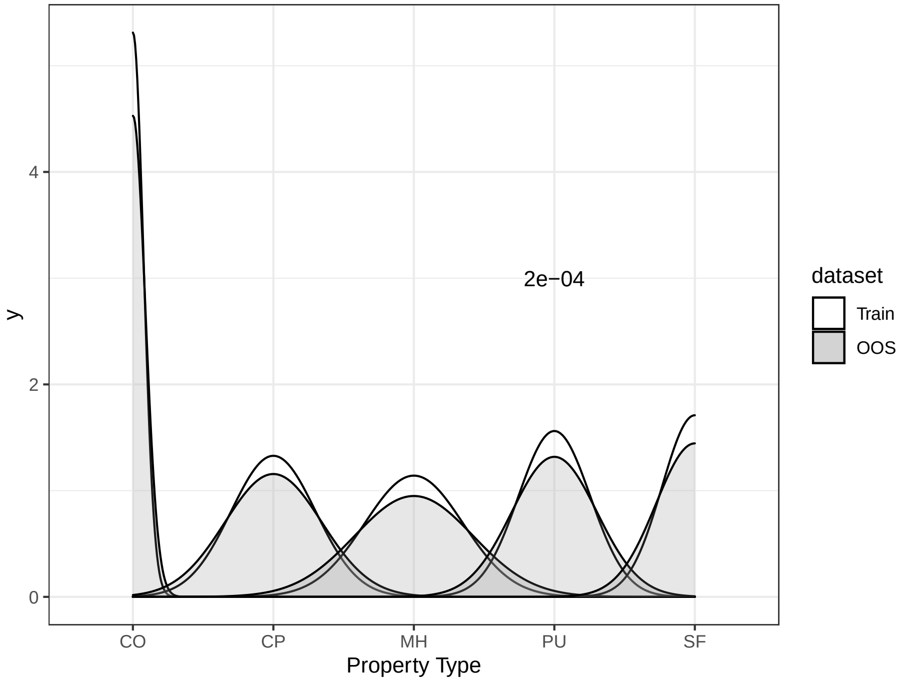

Figure 47: Property Type Training and Out-of-Sample Dataset Bin Distribution, and Population Stability Index Test Result

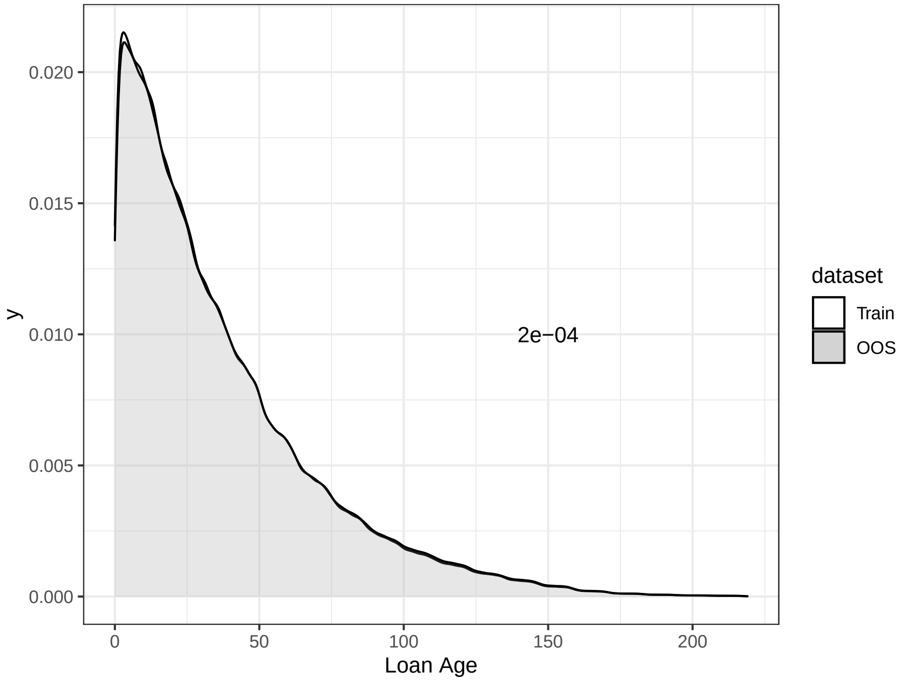

Figure 48: Loan Age Training and Out-of-Sample Dataset Bin Distribution, and Population Stability Index Test Result

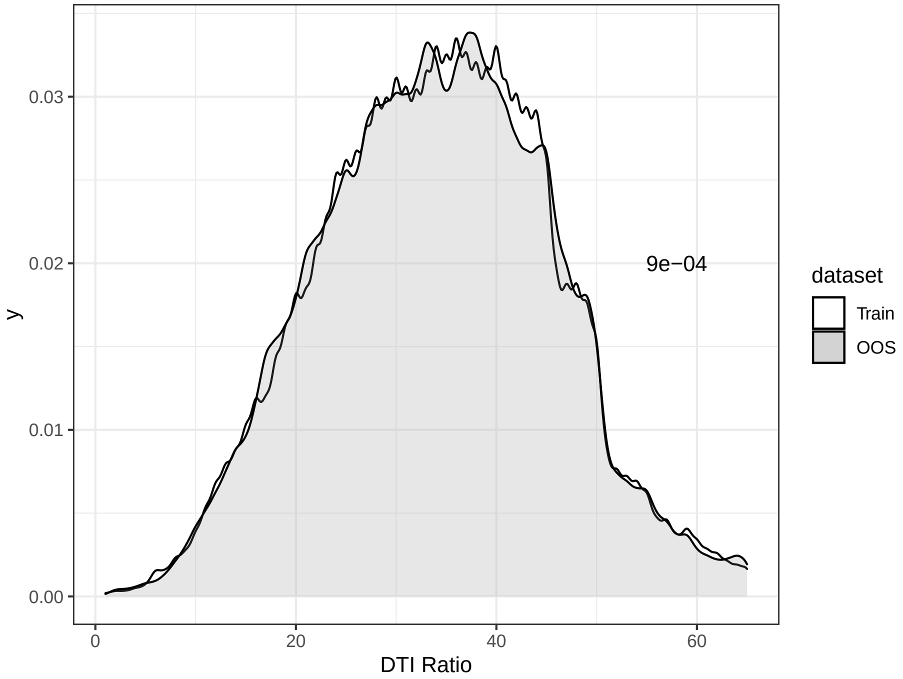

Figure 49: DTI Ratio Training and Out-of-Sample Dataset Bin Distribution, and Population Stability Index Test Result

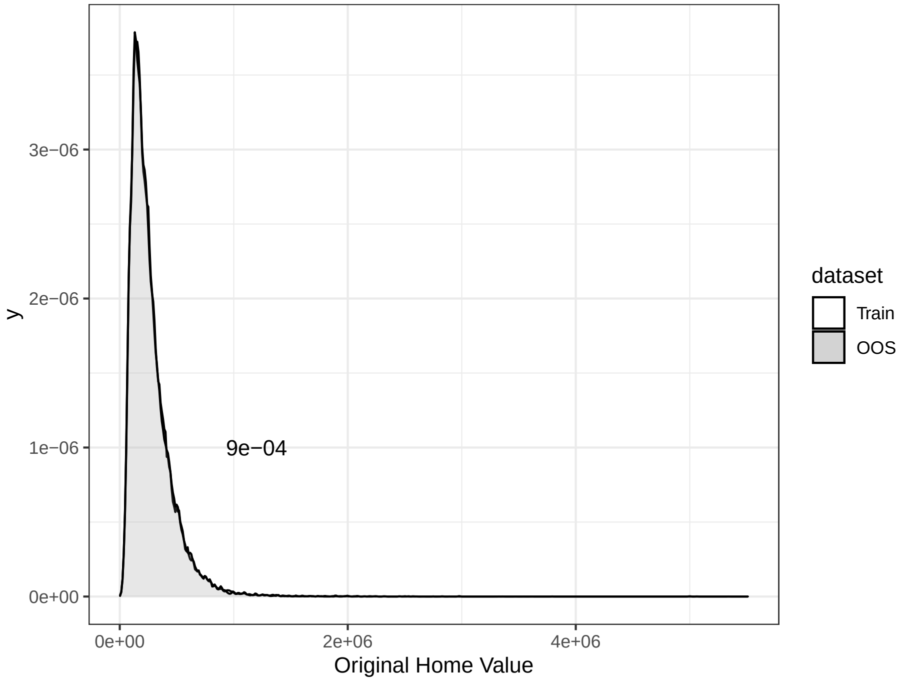

Figure 50: Original Home Value Training and Out-of-Sample Dataset Bin Distribution, and Population Stability Index Test Result

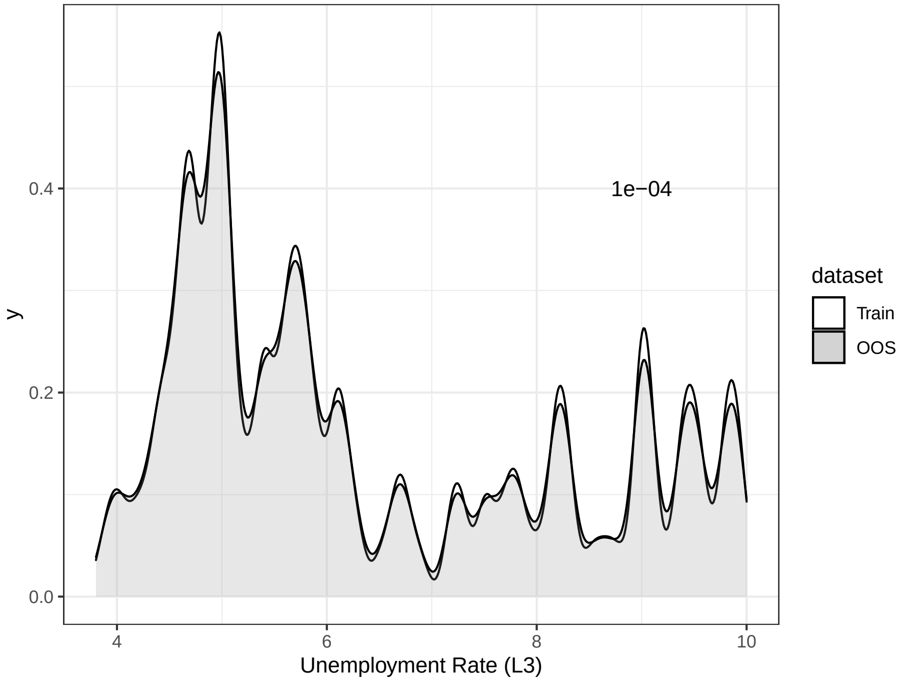

Figure 51: Unemployment Rate (3 month lag) Training and Out-of-Sample Dataset Bin Distribution, and Population Stability Index Test Result

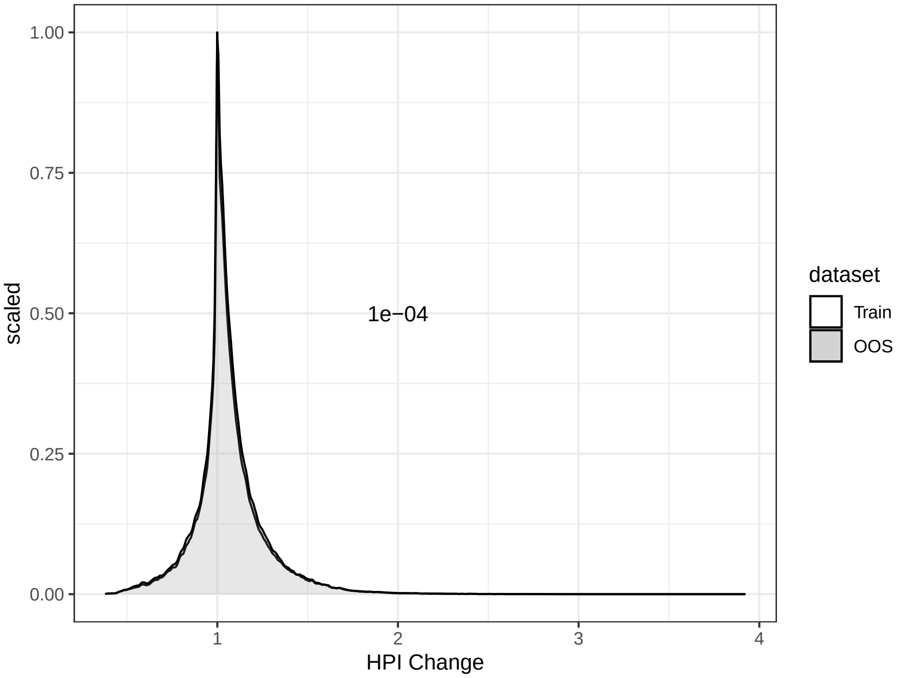

Figure 52: HPI Change Training and Out-of-Time Dataset Bin Distribution, and Population Stability Index Test Result

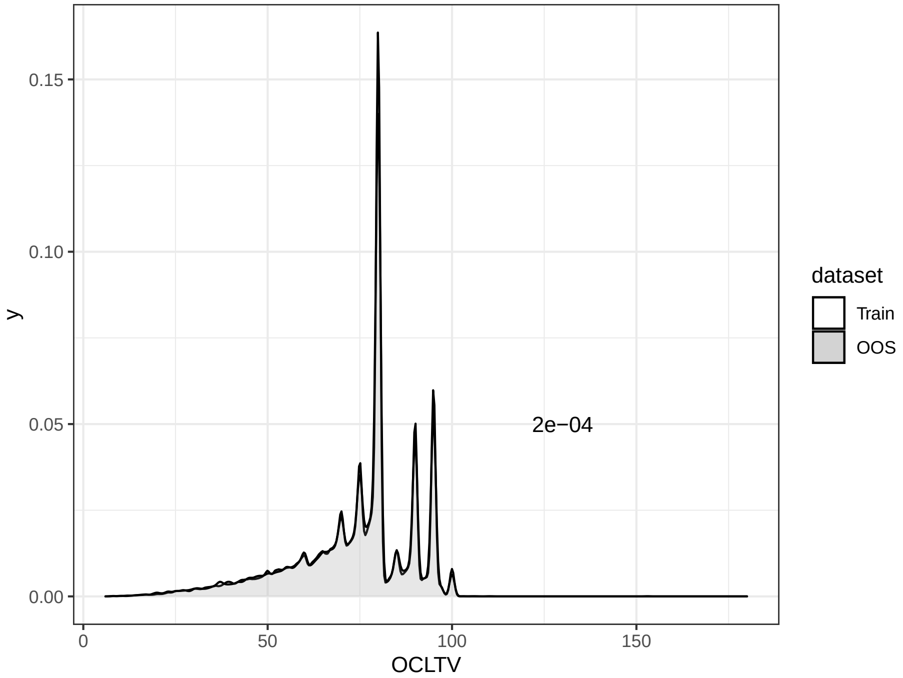

Figure 53: Original Combined LTV Training and Out-of-Time Dataset Bin Distribution, and Population Stability Index Test Result

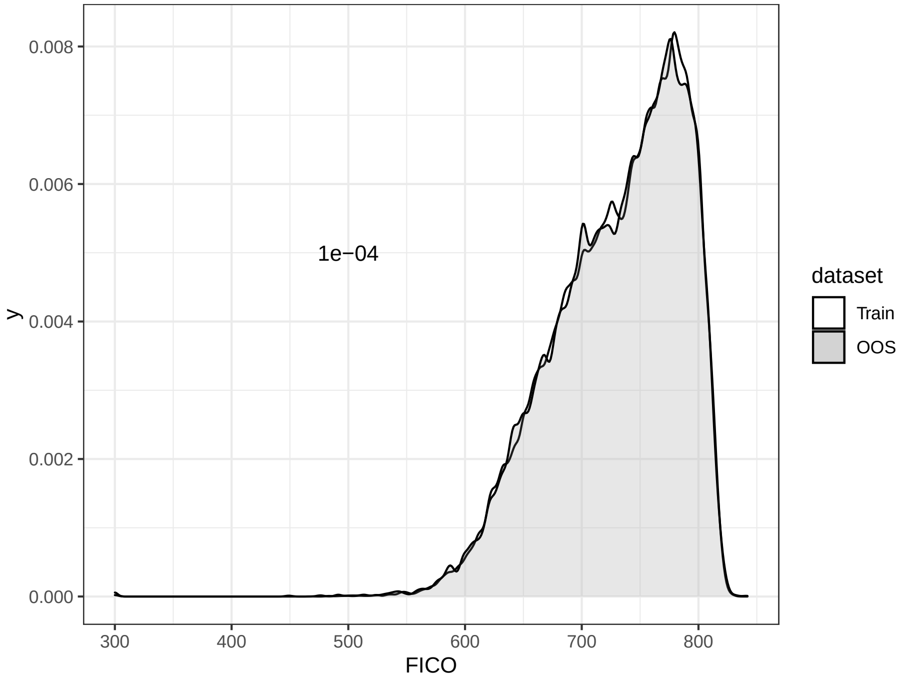

Figure 54: Credit Score (FICO) Training and Out-of-Time Dataset Bin Distribution, and Population Stability Index Test Result

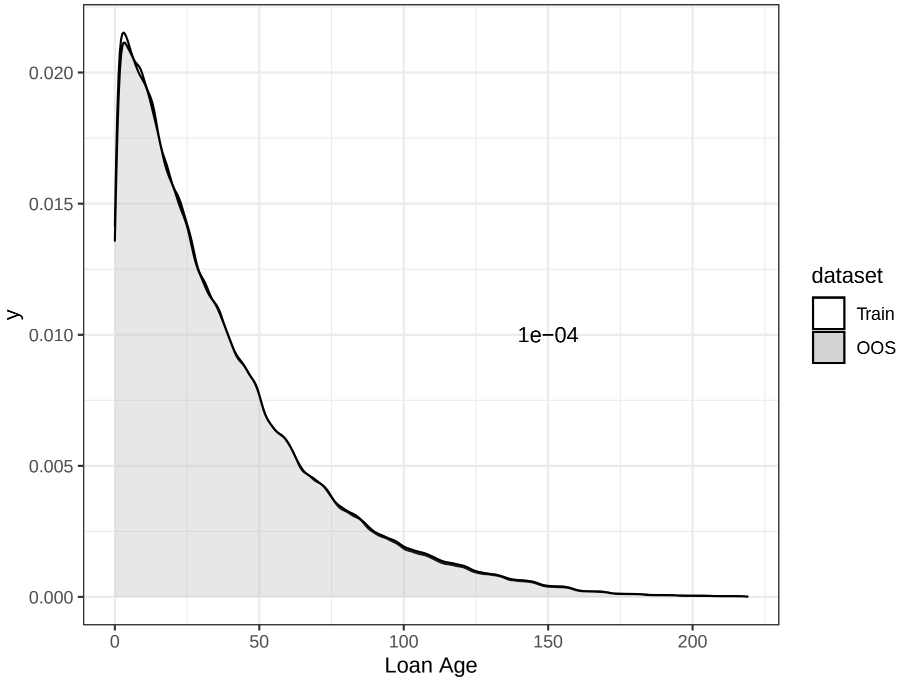

Figure 55: Loan Age Training and Out-of-Time Dataset Bin Distribution, and Population Stability Index Test Result

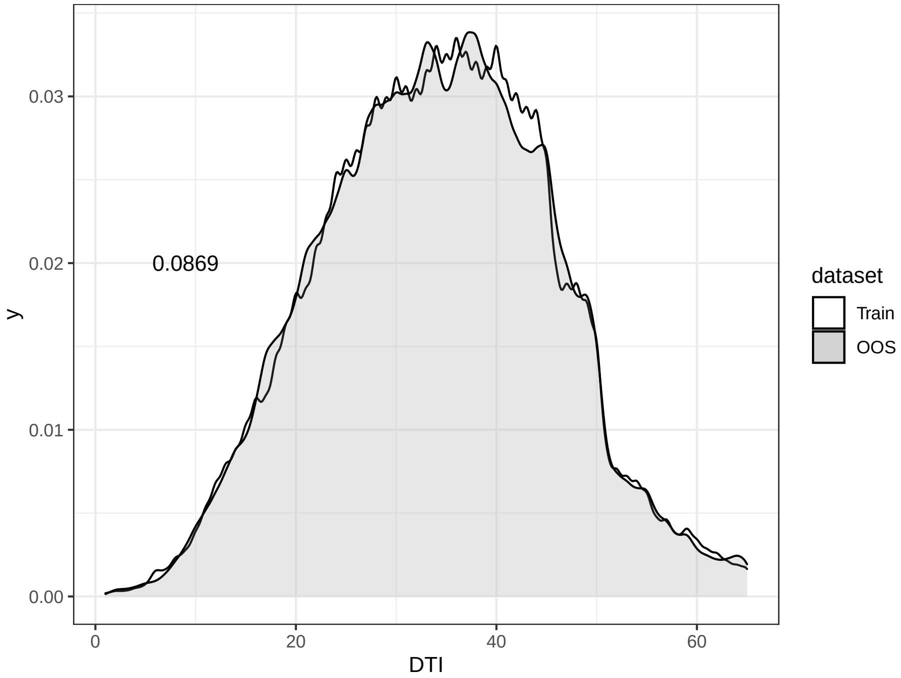

Figure 56: DTI Ratio Training and Out-of-Time Dataset Bin Distribution, and Population Stability Index Test Result

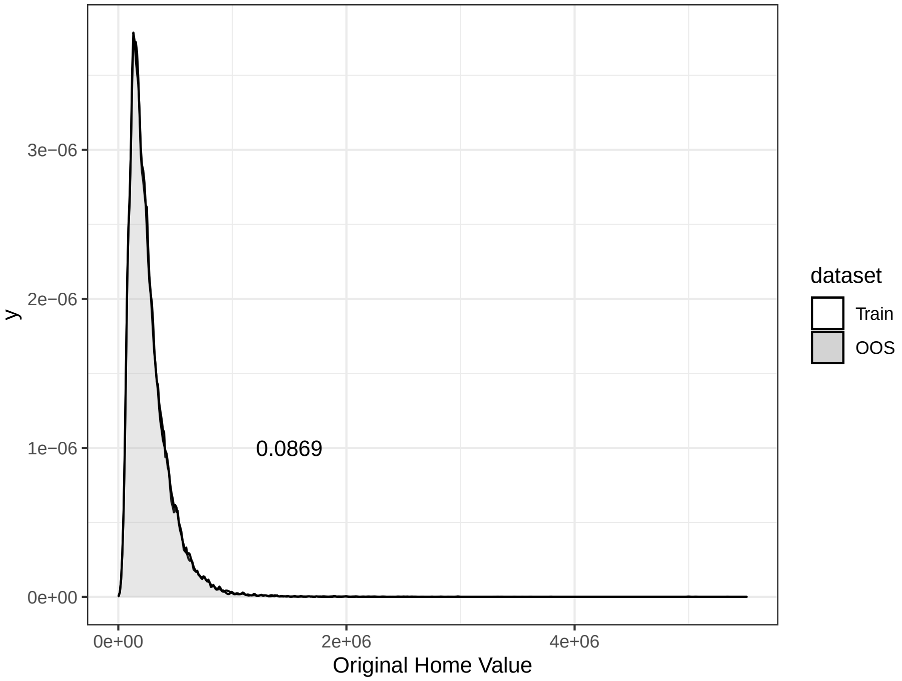

Figure 57: Original Home Value Training and Out-of-Time Dataset Bin Distribution, and Population Stability Index Test Result

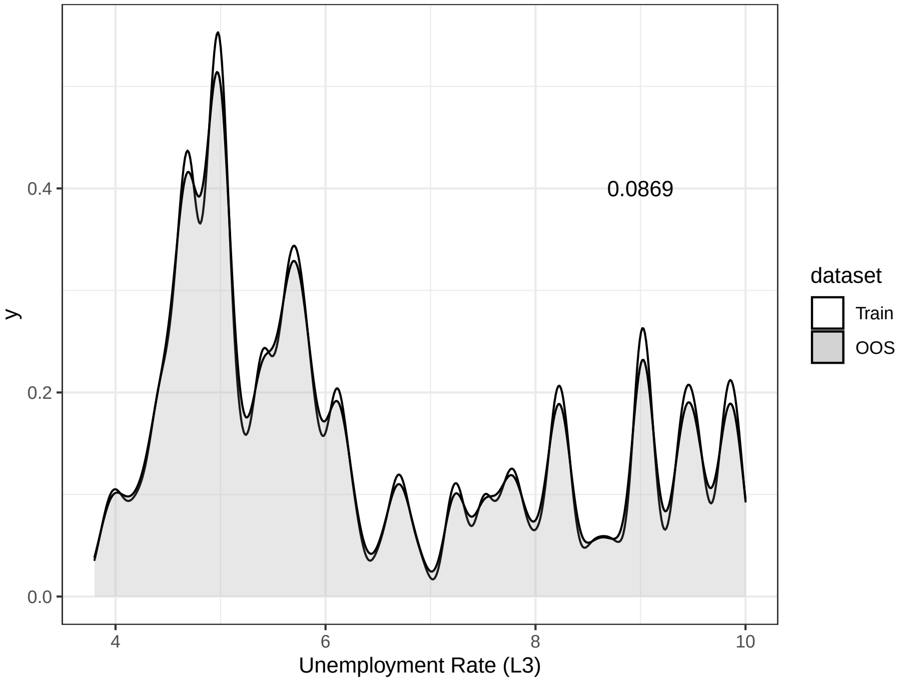

Figure 58: Unemployment Rate (3 month lag) Training and Out-of-Time Dataset Bin Distribution, and Population Stability Index Test Result

References

Abellán, Joaquín, and Javier G Castellano. 2017. "A comparative study on base classifiers in ensemble methods for credit scoring" 73: 1–10. doi:10.1016/j.eswa.2016.12.020.

Altman, EI, Andrea Resti, and Andrea Sironi. 2001. "Analyzing and explaining default recovery rates." *ISDA Report*, no. December. http://www.financerisks.com/filedati/WP/Credit risk/.

Arndorfer, Isabella, and Andrea Minto. 2015. "The 'four lines of defence model' for financial institutions. Taking the three-lines-of-defence model further to reflect specific governance features of regulated financial institutions," no. Occasional Paper 11: 1–29. http://www.bis.org/fsi/fsipapers11.pdf.

Bank for International Settlements. 2019. "The BIS - Promoting global monetary and financial stability through international cooperation." https://www.bis.org/about.

Bank of England Prudential Regulation Authority. 2018. "Model risk management principles for stress testing supervisory statement SS3/18," 9.

Basel Committee on Banking Supervision. 2005. "An Explanatory Note on the Basel II IRB Risk Weight Functions." *Bank for International Settlements*, 1–15. www.bis.orgb/cbsirbriskweight.pdf.

Bastos, João A. 2010. "Forecasting bank loans loss-given-default." *Journal of Banking and Finance* 34 (10): 2510–7. doi:10.1016/j.jbankfin.2010.04.011.

BCBS. 2018. "History of the Basel Committee." https://www.bis.org/bcbs/history.htm.

Bellotti, Tony, and Jonathan Crook. 2009. "Loss Given Default models for UK retail credit cards.” *Credit Research Centre University of*, no. January 2009: 1–28. http://www.crc.man.ed.ac.uk/publications/papers/workingpapers/workingpaper09-1.pdf.

Berk, A Richard. 2017. *Statistical Learning from a Regression Perspective*. Vol. 65. 4. doi:10.1111/j.1541-0420.2009.01343_5.x.

BIS. 2018. “Overview of the revised credit risk framework – Executive Summary,” 1–2.

Board of Governors of the Federal Reserve System. 2012. “12-17 / CA 12-14 Consolidated Supervision Framework for Large Financial Institutions.”

———. 2013. “Capital Planning at Large Bank Holding Companies: Supervisory Expectations and Range of Current Practice.” Washington, DC: FED.

———. 2019a. “Comprehensive Capital Analysis and Review 2019: Assessment Framework and Results.” *Federal Reserve Board*, no. June: 61. https://www.federalreserve.gov/publications/files/2017-ccar-assessment-framework-results-20170628.pdf.

———. 2019b. “Dodd-Frank Act Stress Test 2019: Supervisory Stress Test Methodology and Results,” no. March: 1–67. http://www.federalreserve.gov/newsevents/press/bcreg/.

Breiman, Leo. 1996. “Bagging Predictors” 140: 123–40.

———. 1998. “Arcing Classifiers.” *The Annals of Statistics* 26 (3): 801–49.

———. 1999. “Prediction Games and Arcing Algorihms.” *Neural Computation* 11 (7): 1493–1517.

———. 2001. “Random forests.” *Machine Learning* 45 (1): 5–32. doi:10.1201/9780429469275-8.

Bureau of Labor Statistics. 2020. “Local Area Unemployment Statistics.” Accessed

September 1. https://www.bls.gov/lau/.

Calì, Camilla, and Maria Longobardi. 2015. "Some mathematical properties of the ROC curve and their applications." *Ricerche Di Matematica* 64 (2). Springer Milan: 391–402. doi:10.1007/s11587-015-0246-8.

Chawla V., Nitesh, Kevin Bowyer W., Lawrence Hall O., and W. Philip Kegelmeyer. 2002. "SMOTE: Synthetic Minority Over-sampling Technique." *Journal of Artificial Intelligence Research* 16 (1): 321–57. doi:10.1613/jair.953.

Chen, Tianqi, and Carlos Guestrin. 2016. "XGBoost: A scalable tree boosting system." *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data*, 785–94. doi:10.1145/2939672.2939785.

Committee on Banking Supervision, Basel. 2013. *Basel Committee on Banking Supervision Consultative Document Fundamental review of the trading book: A revised market risk framework.* January. www.bis.org.

Crespo, Ignacio, Pankaj Kumar, Peter Noteboom, and Marc Taymans. 2017. "The evolution of model risk management." *McKinsey on Risk* 2.

De Boor, Carl. 2001. *A practical guide to splines.* Springer.

Duffy, Nigel, and David Helmbold. 1999. "A geometric approach to leveraging weak learners." *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 1572: 18–33.

Efron, Bradley. 1979. "Bootstrap Methods : Another Look at the Jackknife." *The Annals of Statistics* 7 (1): 1–26.

Efron, Bradley, and R Tibshirani. 1994. *An introduction to the bootsrap.* New York:

Chapman & Hall.

EU-Kommission. 2013. "Capital Requirements Directive (2013/36/EU; CRD IV)." *Official Journal of the European Union*, 338–436. https://www.europex.org/eu-legislation/crd-iv-and-crr/.

European Central Bank. 2017. "What is the SREP?" https://www.bankingsupervision.europa.eu/about/ssmexplained/html/srep.en.html.

———. 2019. "Interim update on the Targeted Review of Internal Models (TRIM); Second update on TRIM outcomes (as of March 2019)," no. April: 1–12.

FASB. 2016. "Financial Instruments — Credit Losses (Topic 326) Measurement of Credit Losses on Financial Instruments." *Financial Accounting Series* 437: 1–285.

FED. 2011. "SR 11-7: Supervisory Guidance on Model Risk Management," 1–21.

Federal Housing Finance Agency. 2020. "House Price Index Datasets." https://www.fhfa.gov/DataTools/Downloads/Pages/House-Price-Index-Datasets.aspx.

Federal Reserve Board. 2011. "Federal Reserve Board announces $3 million fine against Bank of New York Mellon Corporation (BNY Mellon) for unsafe and unsound practices."

———. 2017a. "12 CFR § 225.8 - Capital planning." FED.

———. 2017b. "Federal Reserve announces two enforcement actions against Deutsche Bank AG that will require bank to pay a combined $156.6 million in civil money penalties." https://www.federalreserve.gov/newsevents/pressreleases/enforcement20170420a.htm.

———. 2020. "Assessment of Bank Capital during the Recent Coronavirus Event, June 2020," no. June. https://www.federalreserve.gov/publications/files/2020-sensitivity-analysis-20200625.pdf.

FHFA. 2019. “History of Fannie Mae and Freddie Mac Conservatorship.” Accessed March 14.

Frame, W. Scott. 2015. “The 2008 Federal Intervention to Stabilize Fannie Mae and Freddie Mac.” *Journal of Applied Finance* 18 (2): 13.

Freddie Mac. 2020. “Freddie Mac Single Family Loan-Level Dataset.” http://www.freddiemac.com/research/datasets.

Freund, Yoav, and Robert E. Schapire. 1997. “A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting.” *Journal of Computer and System Sciences* 55 (1): 119–39. doi:10.1006/jcss.1997.1504.

Freund, Yoav, Robert E Schapire, and Park Avenue. 1999. “A Short Introduction to Boosting” 14 (5): 771–80.

Freund, Yoav, Robert E Schapire, and Murray Hill. 1996. “Experiments with a New Boosting Algorithm.” *Machine Learning: Proceedings of the Thirteenth International Conference*, 148–56.

Friedman, Jerome H. 2001. “Greedy function approximation: A gradient boosting machine.” *Annals of Statistics* 29 (5): 1189–1232. doi:10.2307/2699986.

———. 2002. “Stochastic gradient boosting.” *Computational Statistics and Data Analysis* 38 (4): 367–78. doi:10.1016/S0167-9473(01)00065-2.

Frye, Jon. 2013. “Loss given default as a function of the default rate.” *Federal Reserve Bank of Chichago*, no. September: 1–13.

Galar, Mikel, Alberto Fernandez, Edurne Barrenechea, Humberto Bustince, and Francisco Herrera. 2012. “A review on ensembles for the class imbalance problem: Bagging-, boosting-, and hybrid-based approaches.” IEEE. doi:10.1109/TSMCC.2011.2161285.

Goldstein, Alex, Adam Kapelner, Justin Bleich, and Emil Pitkin. 2014. “Peeking Inside the Black Box: Visualizing Statistical Learning With Plots of Individual Conditional Expectation.” *Journal of Computational and Graphical Statistics* 24 (1): 44–65. doi:10.1080/10618600.2014.907095.

Greene, Henry J., and George R. Milne. 2010. “Assessing model performance: The Gini statistic and its standard error.” *Journal of Database Marketing and Customer Strategy Management* 17 (1): 36–48. doi:10.1057/dbm.2010.2.

Gregory, Jon. 2015. *The xVA Challenge: Counterparty Credit Risk, Funding, Collateral and Capital*. 3rd ed. John Wiley & Sons, Ltd.

Grusky, David B., Bruce Western, and Christopher Wimer. 2011. *The Great Recession*. Rutgers University Press.

Gürtler, Marc, and Martin Hibbeln. 2013. “Improvements in loss given default forecasts for bank loans.” *Journal of Banking and Finance* 37 (7): 2354–66. doi:10.1016/j.jbankfin.2013.01.031.

Hansen, Bruce. 2020. *Econometrics*. University of Wisconsin, Department of Economics.

Hartmann, P, O De Bandt, P Molyneux, and J Wilson. 2009. “The Concept of Systemic Risk.” *European Central Bank Financial Stability Review*, no. December 2006: 134–42.

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*.

Henley, W.E., and D.J Hand. 1996. “A k-Nearest-Neighbour Classifier for Assessing

Consumer Credit Risk." *Journal of the Royal Statistical Society. Series D (the Statistician)* 45 (1): 77–95.

IASB. 2018. "IASB clarifies its definition of 'material'." https://www.ifrs.org/news-and-events/2018/10/iasb-clarifies-its-definition-of-material/.

IFRS. 2017. "IFRS 17 Insurance Contracts." https://www.ifrs.org/issued-standards/list-of-standards/ifrs-17-insurance-contracts/.

———. 2018. "IFRS 9 Financial Instruments." https://www.ifrs.org/issued-standards/list-of-standards/ifrs-9-financial-instruments/.

Jones, David. 2000. "Emerging problems with the Basel Capital Accord: Regulatory capital arbitrage and related issues." *Journal of Banking and Finance* 24 (1-2): 35–58. doi:10.1016/S0378-4266(99)00052-7.

Lessmann, Stefan, Bart Baesens, and Hsin-vonn Seow. 2015. "Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research." doi:10.1016/j.ejor.2015.05.030.

Li, Phillip, Min Qi, Xiaofei Zhang, and Xinlei Zhao. 2014. "Further investigation of parametric loss given default modeling." *Office of the Comptroller of the Currency* 2. doi:10.21314/JCR.2016.215.

Loh, Wei Yin. 2011. "Classification and regression trees." *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 1 (1): 14–23. doi:10.1002/widm.8.

Loterman, Gert, Iain Brown, David Martens, Christophe Mues, and Bart Baesens. 2012. "Benchmarking regression algorithms for loss given default modeling." *International Journal of Forecasting* 28 (1). Elsevier B.V.: 161–70. doi:10.1016/j.ijforecast.2011.01.006.

Mamonov, Stanislav, and Raquel Benbunan-Fich. 2017. “What can we learn from past mistakes? Lessons from data mining the Fanie Mae mortgage portfolio.” *Journal of Real Estate Research* 39 (2): 28.

Mason, Llew, Jonathan Baxter, Peter Bartlett, and Marcus Frean. 1999. “Boosting algorithms as gradient descent in Function space.” *Nips.* doi:10.1109/5.58323.

Office of the Superintendent of Financial Institutions Canada. 2017. “E-23: Enterprise-Wide Model Risk Management for Deposit-Taking Institutions Category: Sound Business and Financial Practices,” 1–15.

Oshiro, Thais Mayumi, Pedro Santoro Perez, and Jose Augusto Baranauskas. 2012. “How Many Trees in a Random Forest?” no. July: 154–68. doi:10.1007/978-3-642-31537-4.

Papke, Leslie E., and Jeffrey M. Wooldridge. 1996. “Econometric methods for fractional response variables with an application to 401 (k) plan participation rates.” *Journal of Applied Econometrics* 11 (6): 619–32. doi:10.1002/(SICI)1099-1255(199611)11:6<619::AID-JAE418>3.0.CO;2-1.

———. 2008. “Panel data methods for fractional response variables with an application to test pass rates.” *Journal of Econometrics* 145 (1-2): 121–33. doi:10.1016/j.jeconom.2008.05.009.

parliament, European. 2013. “Capital Requirements Regulation (575/2013; CRR).” *Official Journal of the European Union*, no. 27.6.2013: 1–337.

Perperoglou, Aris, Willi Sauerbrei, Michal Abrahamowicz, and Matthias Schmid. 2019. “A review of spline function procedures in R.” *BMC Medical Research Methodology* 19

(1). BMC Medical Research Methodology: 1–16. doi:10.1186/s12874-019-0666-3.

Peter, J Wallison, and W. Calomiris Charles. 2009. "The Last Trillion-Dollar Commitment: The Destruction of Fannie Mae and Freddie Mac." *The Journal of Structured Finance* 15 (1): 71–80.

Probst, Philipp, and Anne Laure Boulesteix. 2018. "To tune or not to tune the number of trees in random forest." *Journal of Machine Learning Research* 18: 1–8.

Racine, Jeffrey S. 2012. "A Primer on Regression Models," no. 1990: 82–146. doi:10.1002/9781118150528.ch3.

Ramalho, Esmeralda A., Joaquim J.S. Ramalho, and José M.R. Murteira. 2011. "Alternative estimating and testing empirical strategies for fractional regression models." *Journal of Economic Surveys* 25 (1): 19–68. doi:10.1111/j.1467-6419.2009.00602.x.

Rätsch, G., T. Onoda, and K. R. Müller. 2001. "Soft margins for AdaBoost." *Machine Learning* 42 (3): 287–320. doi:10.1023/A:1007618119488.

Schapire, Robert E., and Yoav Freund. 2012. *Boosting: Foundations and Algorithms.* MIT Press.

Schuermann, Til. 2004. "What do we know about Loss-Given-Default?" *Wharton Financial Institutions Center Working Paper* 04 (1): 30. doi:10.17816/ecogen844-9.

Sun, Han Sheng, and Zi Jin. 2016. "Estimating Credit Risk Parameters Using Ensemble Learning Methods: An Empirical Study on Loss Given Default." *Journal of Credit Risk*, no. Forthcoming: 27.

Tanoue, Yuta, Satoshi Yamashita, and Hideaki Nagahata. 2020. "Comparison study of two-step LGD estimation model with probability machines." *Risk Management* 22

(2).

Taplin, Ross, and Clive Hunt. 2019. “The population accuracy index: A new measure of population stability for model monitoring.” *Risks* 7 (2): 1–11. doi:10.3390/risks7020053.

The Institute of Internal Auditors (IIA). 2013. “The three Lines of Defense in effective Risk Management and Control.” *IIA Position Paper*, no. January. doi:10.1039/c1cc12161h.

The Office of the Comptroller of, and Currency. 2020. “Interagency Policy Statement on Allowances for Credit Losses” 85 (105): 15–28.

Tong, Edward N.C., Christophe Mues, and Lyn Thomas. 2013. “A zero-adjusted gamma model for mortgage loan loss given default.” *International Journal of Forecasting* 29 (4). Elsevier B.V.: 548–62. doi:10.1016/j.ijforecast.2013.03.003.

Yurdakul, Bilal. 2018. “Statistical Properties of Population Stability Index.” *Dissertations.*

Zhang, Jie, and Lyn C. Thomas. 2012. “Comparisons of linear regression and survival analysis using single and mixture distributions approaches in modelling LGD.” *International Journal of Forecasting* 28 (1). Elsevier B.V.: 204–15. doi:10.1016/j.ijforecast.2010.06.002.