Point-in-Time Survival Model for Mortgage Portfolio Risk Management

Isaac Vergara, Patrick Zhu, Jenny Chan, Gaurav Law

December 9, 2025

Contents

1 A Research Overview of the Credit Risk Industry

1.1 Credit Assessment Tools

Lenders use credit scores and the Five Cs of credit (capital, capacity, conditions, character, and collateral) to evaluate individuals. For corporations and governments, agencies like Moody's rate the credit risk of bonds; bonds rated below BBB carry higher default risk.

The rating process begins with an agency analyzing a company's financial ratios and leverage to assign a credit rating to inform risk-based pricing. Meanwhile, investors actively trade bonds and credit default swaps based on new information, creating a market-implied PD. If the market believes a company is riskier, its bond prices fall and its credit default swap spread rises. This happens because increased default probability lowers expected cash flows, and since bond price discounts these cash flows, the price decreases.

Banks compare the fundamental probability of default from accounting with the market-implied probability of default from prices. If they diverge, banks revisit and update the internal credit score. This feedback loop is vital because default risk can change quickly while formal credit ratings change slowly. The market-implied probability of default provides a timely warning. Famously, the Lehman Brothers had a credit default swap spread implying an 80 percent probability of default months before its collapse in 2008.

1.2 Bank Risk Governance

Banks manage credit risk through three lines of defense:

1. First line of defense: the front desk and business units.
2. Second line of defense: the independent risk management function.
3. Third line of defense: the chief audit officer, the internal board, and the executive risk committee.

The third line of defense, the board and risk committee, decide enterprise-wide risk boundaries and risk appetites. They determine aggregate capital sources and economic capital, and approve the overall risk culture and strategies. Their objective is to select a prudent capital and business strategy to maximize enterprise-wide return. To achieve this, they decide the total amount of capital to allocate across businesses, considering the constraints of feasible capital sources. Regulators require banks to develop formal risk appetite frameworks.

To develop a risk appetite framework, banks must determine their Value at Risk (VaR), how much loss at a given confidence level they are prepared to risk, and their reputation risk targets like wanting an AA credit rating. They must decide how concentrated they will allow risks to become. Sometimes, mean variance optimization methods may imply concentrating assets is best, but that invites excessive risk.

The second line of defense, the risk management organization, allocates the total available capital determined by the board to different business units. Their job is to minimize total unexpected loss while meeting return requirements.

As the first line of defense, business units are responsible for managing their risks within allocated economic capital. This means executing management strategies while ensuring that unexpected losses remain within their capital limits. A practical example is a front-office trading desk, which must control its daily trading activity so that potential losses do not exceed the Value at Risk (VaR) limit established by the second line of defense.

This structure creates a clash of goals. Risk management tries to minimize risk subject to meeting return constraints. Business units try to maximize return subject to staying below risk targets.

1.3 Quantifying Credit Risk

For a portfolio of N loans, expected credit loss is calculated as:

$$\text{Credit Loss (CL)} = N \times \text{PD} \times \text{LGD} \times \text{EAD}$$

Concepts:

- Hazard rate: instantaneous probability of default.
- Survival rate: probability of not defaulting.
- Cumulative probability of default: sum of incremental probabilities from the first year

of the agreement to the time you are accounting for.

- Incremental probability of default: decrease in the survival rate from one year to the next

1.4 Credit Scoring Systems

Two main types of credit scores:

- Third-party scores: generated by companies other than the credit bureau itself. Offer a broader view and often consider alternative data (e.g., VantageScore, FICO).
- In-house scores: generated by the credit bureau or by lenders for their own system. Tailored to specific lenders.

In-house scores can be categorized as:

- General bank-wide scores
- Customer or segment-specific scores
- Acquisition scores for new applicants
- Existing account scores for monitoring current customers

1.5 Comparing Third-Party Credit Scoring Systems

1.5.1 Consumer Credit Models

FICO (Consumer) The FICO score is the industry-standard, logistic regression-based scorecard. It predicts one-year delinquency using payment history, credit utilization, length of history, new credit, and credit mix. Newer versions, notably FICO 10T, incorporate 24 months of trended balance data, shifting from a static snapshot to behavioral trajectory analysis. This parsimonious design prioritizes regulatory defensibility, interpretability, and lender comparability. Its dominance is its primary strength, but its conservatism limits coverage for thin-file consumers. While scoring is computationally lightweight, models using trended data require more extensive historical data pipelines.

Comparative Performance: For mainstream, "thick-file" consumers, FICO 8/9 remains highly predictive. However, in validation studies, FICO 10T shows a 5-15% lift in capture rate for borrowers actively managing balances, as trended data provides a forward-looking risk signal that static utilization ratios lack.

VantageScore (Consumer) Designed as a tri-bureau consistent model, VantageScore utilizes a broader attribute set. VantageScore 4.0 employs machine learning (ML) to optimize weighting, explicitly incorporating trended data and, where available, alternative tradelines like rent and utility payments. Its core design rationale is to maximize scorable population and predictive power for thin-file consumers. A key trade-off is reduced model transparency compared to simple scorecards.

Comparative Performance: VantageScore 4.0's ML architecture allows it to score approximately 94-96% of the adult population, compared to roughly 90-92% for FICO 8. Its strategic validation came with acceptance by Fannie Mae and Freddie Mac for mortgage underwriting, directly challenging FICO's monopoly in that sector.

Bureau Proprietary Scores Experian, Equifax, and TransUnion each market proprietary scores that leverage unique data fields and transforms within their respective databases. Their design emphasizes tight operational integration and the leveraging of bureau-specific data strengths. The primary advantage is granular access to exclusive signals, but the lack of interoperability is a major drawback, as score discrepancies for the same consumer across bureaus complicate automated decisioning.

1.5.2 Customization and Vendor Platforms

Custom Bank Scorecards (e.g., SAS) Financial institutions build custom scorecards using platforms like SAS to combine bureau data with exclusive internal data (e.g., transaction history, reject inference). These models are designed for maximum explainability and control to satisfy stringent internal model governance and regulatory capital (IRB) requirements. The main advantage is a "model lift"—a documented 10-30% improvement in the Gini coefficient over generic bureau scores for the bank's specific portfolio. The cost is a heavy operational burden with slow, batch-oriented development and monitoring cycles.

FICO Model Builder This vendor tool allows lenders to build custom, FICO-methodology scorecards using bureau and application data. It offers a balance between customization and the regulatory comfort of the FICO ecosystem. However, it creates vendor lock-in and still requires substantial in-house expertise for development and validation, sharing the governance burdens of fully custom models without offering their full differentiation potential.

Off-the-Shelf Vendor Scorecards (e.g., FIS, Fiserv) These are generic, pre-built scorecards sold to community banks and lenders without modeling teams. Designed for ease of deployment, compliance, and low cost, they sacrifice predictive customization and competitive differentiation for operational simplicity.

1.5.3 Commercial and Corporate Credit Models

Moody's CreditEdge (EDF) This model assesses public firms using the structural Merton/KMV framework, translating equity market data and balance sheet items into a market-implied Expected Default Frequency (EDF). Its forward-looking design provides early warning, often leading agency rating downgrades by 3-9 months during distress. Its primary limitation is volatility and unreliability for thinly traded or private firms.

Ratings Agency Frameworks (S&P, Moody's, Fitch) These analyst-driven frameworks produce long-term, stable letter-grade ratings through the synthesis of fundamental and qualitative data. Their universal acceptance for benchmarking and regulatory capital mapping is a key strength, offset by their inherent latency and lack of granular, point-in-time default probabilities.

Dun & Bradstreet (PAYDEX) & Experian Business Intelliscore These are leading small and medium enterprise (SME) scores. D&B's PAYDEX is a dollar-weighted index of supplier-reported trade payments, providing a high-fidelity view of B2B payment behavior but suffering from coverage gaps. Experian Intelliscore combines trade data with public records and firmographics for broader coverage but potentially less granular payment insight.

Both face the core commercial credit challenge of relying on fragmented, self-reported trade data.

1.5.4 Next-Generation and Specialized Models

ML-First Fintech Scores (e.g., Upstart, Zest AI) These models use techniques like gradient boosting on high-dimensional alternative data (transactions, education, device metadata) to assess non-traditional applicants. They are designed to uncover nonlinear patterns, with providers claiming significant performance gains. For example, Upstart reports its model can approve 27% more borrowers at the same default rate while reducing APRs by 16% for approved loans. Their principal trade-offs are "black box" interpretability, model risk, and intense regulatory scrutiny regarding fairness and bias.

Open Banking Cashflow Underwriting Models from providers like Plaid use real-time bank transaction data to assess income stability, recurring expenses, and cash flow volatility. This shifts the paradigm from historical debt repayment to real-time affordability, proving highly effective for short-term credit products. The challenges are technical (aggregation errors), regulatory (data consent), and operational (lack of standardization across banking platforms).

Specialized Industry Scores Niche models like Fannie Mae's Desktop Underwriter (mortgage) or auto residual calculators integrate highly specific variables (loan-to-value ratios, depreciation curves). Their design fuses credit risk with collateral risk and product economics, offering superior pricing accuracy within their vertical but no utility outside it. A landmark shift occurred when Fannie Mae announced it would no longer require a minimum FICO score for loans assessed by its DU system, effectively endorsing its proprietary, multi-factor assessment as a primary risk tool over any single third-party score.

1.6 Probability of Default Models

Probability of Default models predict the likelihood that a borrower will fail to meet debt obligations. Different modeling approaches offer distinct advantages depending on the use case, data structure, and regulatory requirements.

- **Linear Probability Model (LPM):** Simple regression treating default as binary outcome; easy to interpret but produces invalid probabilities outside [0,1] and violates standard regression assumptions — useful only for quick exploration, not production models.
- **Logistic Regression:** Industry-standard approach that transforms linear predictions into valid probabilities using a logistic function; excellent for one-year point-in-time PD estimation and regulatory scorecards, but treats default as a single event without capturing timing or temporal dynamics.
- **Survival / Hazard Rate Models:** Model time-to-default rather than binary outcomes; naturally handle censored data (loans that haven't defaulted yet), accommodate time-varying covariates (e.g., evolving credit scores), and capture how risk changes over the loan lifecycle—ideal for lifetime PD estimation and portfolio monitoring.
- **Competing Risks Models:** Extension of survival analysis that distinguishes between multiple exit types (default, prepayment, payoff); prevents biased estimates by explicitly modeling each competing event—critical for portfolios where prepayment is common, such as mortgages.
- **Factor / Structural Models:** Incorporate macroeconomic conditions and default correlations across borrowers using latent factors; essential for stress testing and regulatory capital calculations, but less effective for individual loan-level prediction—designed for portfolio-level risk aggregation rather than borrower scoring.

1.7 How Tech is Changing Credit Risk

1.7.1 Growing Importance of Credit Risk

The credit scoring industry is undergoing a fundamental transformation, driven by three converging forces. First, the impact of the 2008 financial crisis, rooted in credit risk, emphasized the catastrophic cost of poor risk management. Second, the post-2015 rise of FinTech has initiated a technological revolution, demanding more agile and data-driven risk frameworks. Third, the advent of Industry 4.0 and big data has greatly increased the volume and variety of data available for assessment—from traditional financial records to social media and operational metrics—particularly for SMEs. However, this data is often high-dimensional, noisy, and incomplete, making traditional analytical methods insufficient. Consequently, accurately identifying and controlling credit risk is becoming increasingly important for financial stability and growth.

1.7.2 Recent Methodological Shifts: From Statistics to AI

In response, the methodological core of credit scoring has shifted away from traditional statistical models, towards advanced machine learning (ML) and deep learning (DL) techniques.

Traditional & Ensemble ML: While Bayesian models remain in use, the research frontier is dominated by machine learning. Tree-based algorithms, Support Vector Machines (SVM), k-Nearest Neighbors (k-NN), and ensemble methods like AdaBoost and Bagging are the most frequently implemented. A consistent and robust finding is that ensemble methods, which aggregate multiple models, demonstrably outperform single classifiers, offering superior predictive stability.

Rise of Deep Learning: Deep learning architectures, primarily Artificial Neural Networks (ANN) and Multi-Layer Perceptrons (MLP), are now widely applied in research. Studies indicate that these models generally achieve higher predictive accuracy than both traditional ML and statistical approaches, though their deployment in production systems lags behind research. Notably, citation analysis suggests a balanced research interest across various DL methods, indicating a healthy, exploratory field.

1.7.3 Critical Challenges and Future Trajectory

Despite rapid progress, four major challenges limit the full realization of AI-driven credit scoring:

1. Severe Data Imbalance: Credit datasets are inherently imbalanced, with defaults far outnumbered by non-defaults. Conventional remediation techniques like over-sampling and under-sampling have proven inadequate in both effectiveness and computational efficiency, necessitating novel algorithmic solutions.

2. Lack of Benchmark Datasets: The prevalence of private, proprietary datasets severely limits fair performance comparison and reproducible research.

3. The "Black Box" Problem: The superior predictive power of complex ML/DL models often comes at the cost of transparency and explainability. This opacity conflicts with growing regulatory demands (e.g., "right to explanation") and practical needs for stakeholder trust.

4. Nascent Application of Deep Learning: Compared to other domains, the specialized application of deep learning to credit risk is still limited, representing a significant frontier for future customization and innovation.

2 Introduction to Risk Management Careers

Risk management is a core function in modern financial institutions and the broader credit ecosystem. Banks, mortgage lenders, credit card issuers, and specialized analytics firms all face uncertainty — whether borrowers will repay, how markets will move, or how operational and regulatory issues could lead to financial losses. The role of risk management is to identify, measure, monitor, and mitigate these risks in a disciplined and repeatable way, enabling institutions to support business growth while remaining safe, sound, and compliant. With the drive toward more intelligent and transparent systems, the demand for skilled risk management professionals is higher than ever.

Within credit markets, risk management is particularly intertwined with credit scoring and underwriting. Credit scores, internal risk models, and portfolio monitoring tools shape which customers are approved, what credit limits they receive, and at what price. Teams of analysts and data scientists design these models, validate their performance, and interpret their outputs for business and regulatory stakeholders. Other risk professionals focus on the broader portfolio — for example, by setting risk appetite, designing stress tests, and monitoring concentrations by region, product, or borrower segment.

Because of this central role, risk management has become a well-defined career track in financial services. Entry-level roles typically combine quantitative analysis with business oriented problem solving: New hires might build dashboards to track delinquency trends, evaluate the impact of a new credit policy, or help validate a model before it is put into production. As professionals gain experience, they can move into more specialized quantitative roles, broader portfolio or enterprise risk roles, or leadership positions that manage teams and interact closely with senior management and regulators.

The project in this course naturally connects to these careers. By comparing traditional credit scores such as FICO with newer frameworks like VantageScore 4.0 and by prototyping an enhanced credit model, we are effectively performing a simplified version of the work that credit risk and analytics teams would do in practice. Understanding the range of career paths in risk management therefore helps place our technical work in a richer professional context.

3 Major Career Paths in Risk Management

Although job titles vary across institutions, risk management in financial services can be grouped into several major career paths. These paths differ in their primary focus — for example, credit versus market risk — but they share a common toolkit that blends quantitative skills, domain knowledge, and communication.

3.1 Credit Risk

Credit risk is the risk that borrowers fail to meet their contractual obligations, leading to losses for the lender or investor. These potential losses are formally quantified using three key risk parameters: the Probability of Default (PD), the Loss Given Default (LGD), and the Exposure at Default (EAD). We define these three key terms below:

- Probability of Default (PD): the chance a debtor fails to repay. A default is typically defined contractually, usually as failing to pay a required amount after a specific delinquency period.
- Exposure at Default (EAD): the market value of the credit claim at the time of default.
- Loss Given Default (LGD): the portion of the Exposure at Default not recovered after post-default recovery efforts.

Credit risk roles are particularly relevant for lenders, credit card issuers, mortgage investors, and credit scoring companies. Typical roles in this area include Credit Risk Analyst, Senior Credit Risk Analyst, Credit Policy Analyst, Portfolio Risk Manager, and Credit Risk Modeler. Day-to-day responsibilities often involve:

- Analyzing default, delinquency, and loss trends across products and customer segments
- Designing and monitoring credit policies, such as approval criteria, cut-off scores, and line assignment strategies
- Building or maintaining models that predict PD, LGD, and EAD
- Conducting stress tests and scenario analyses to assess how the portfolio would behave under adverse macroeconomic conditions

These roles sit at the intersection of data, models, and business decisions. They are closely related to the work done in our project, where we evaluate alternative credit scoring approaches and their impact on portfolio performance and inclusion.

3.2 Market Risk

Market risk refers to the risk of losses due to movements in market variables such as interest rates, equity prices, exchange rates, and credit spreads. Market risk roles are most common in institutions with trading activities, investment portfolios, or complex hedging programs.

Typical roles include Market Risk Analyst, Market Risk Manager, and Risk Officer for specific trading desks. Responsibilities may include:

- Measuring exposures using metrics such as Value-at-Risk (VaR) and sensitivities (e.g., duration, delta, gamma)
- Monitoring compliance with risk limits and escalating breaches to management
- Assessing the risk implications of new products, strategies, or hedging structures
- Producing risk reports for internal committees and regulators

While our project focuses more on credit risk than market risk, both paths require strong quantitative skills and a solid understanding of financial products.

3.3 Operational and Non-Financial Risk

Operational risk captures the possibility of loss resulting from inadequate or failed processes, people, systems, or external events. This category includes fraud, cyber incidents, process breakdowns, and model implementation errors. Many institutions now also group conduct risk, technology risk, and certain aspects of environmental, social, and governance (ESG) risk under broader non-financial risk umbrellas.

Typical roles include Operational Risk Analyst, Non-Financial Risk Specialist, and Fraud Risk Manager. Common responsibilities are:

- Mapping key processes and identifying potential failure points or control weaknesses
- Designing and monitoring key risk indicators and loss event databases
- Assessing fraud patterns and supporting the design of fraud detection strategies
- Coordinating incident management and remediation plans when issues occur.

Although the focus is less on statistical modeling and more on controls and processes, these roles still rely heavily on data analysis and structured problem solving.

3.4 Model Risk Management and Quantitative Risk

Model risk management (MRM) is concerned with the risk that models are incorrect, misused, or poorly governed. As institutions rely more heavily on complex models for pricing, risk measurement, and decision making, MRM has emerged as a distinct and growing career path.

Typical roles include Model Risk Analyst, Model Validator, and Quantitative Risk Specialist. These professionals:

- Independently review models developed by front-line teams, including credit scoring models, market risk models, and capital models
- Assess model design, data quality, assumptions, and limitations
- Perform benchmarking and backtesting to evaluate performance and stability over time

- Ensure that models comply with internal governance standards and external regulatory expectations

For someone working on credit scoring, MRM roles can be especially attractive because they require a deep understanding of modeling techniques as well as an ability to challenge and communicate model risk.

3.5 Enterprise Risk and Regulatory / Compliance Risk

Enterprise risk management (ERM) takes a holistic view of risk across the entire organization, integrating credit, market, operational, liquidity, and other risk types into a consistent framework. Regulatory and compliance risk teams focus on ensuring that the institution's practices meet legal and regulatory requirements.

Typical roles include Enterprise Risk Analyst, Risk Governance Associate, and Regulatory Risk Analyst. Their work often involves:

- Supporting the definition of the firm's overall risk appetite and translating it into quantitative limits and qualitative guidelines
- Aggregating and interpreting risk information across business lines for senior management and boards
- Overseeing risk policies and ensuring consistent implementation across the firm
- Coordinating regulatory submissions and responding to supervisory inquiries

These roles are less focused on building individual models and more focused on governance, coordination, and high-level risk oversight.

3.6 Typical Career Progression

Across these paths, a common career progression is from analyst roles to more senior and managerial positions. Many professionals start as analysts in a specific risk area, such as credit or model risk, where they learn core tools and processes. Over time, they may move into more specialized expert roles, broaden into portfolio or enterprise risk positions, or transition into leadership roles that manage teams and interact more frequently with senior management and regulators.

For students and early-career professionals, this variety of paths means that risk management is not a single job, but a flexible career space. It offers opportunities for quantitatively oriented modelers, data scientists, and analysts, as well as for those who prefer governance, policy, and cross-functional coordination. In later sections, we can connect these general paths to concrete roles at institutions such as Fannie Mae, FICO, and VantageScore, and examine how their job postings translate these ideas into specific responsibilities and qualifications.

4 Company Specific Roles in Our Project

To connect our project more directly to real career opportunities, we focus on roles from the three institutions that appear in our analysis: Fannie Mae, FICO, and VantageScore. Although these organizations sit in different parts of the credit ecosystem, their job descriptions show a common demand for strong quantitative skills, data literacy, and the ability to communicate risk and model results clearly.

Fannie Mae operates at the core of the U.S. mortgage market and offers several risk-related career paths. Roles such as Capital Markets Risk Senior Associate focus on monitoring interest-rate and liquidity risk on large mortgage and securities portfolios, reviewing risk metrics and reports, and working with trading and treasury teams. More quantitative positions, such as a Lead Quantitative Modeler, emphasize developing and enhancing models for mortgage cash flows, borrower behavior, and stress-testing, using tools like Python, SAS, and SQL. On the governance side, a Model Risk Business Process QA Reviewer / Advisor role sits in model risk management, reviewing validation work, maintaining quality assurance frameworks, and ensuring that models and validation practices comply with internal standards and regulatory expectations. Across these roles, Fannie Mae typically looks for candidates with quantitative degrees, a few years of relevant experience, strong programming and statistical skills, and solid communication abilities.

FICO is best known for the FICO Score but is also a software and analytics company. The role we examined, Full Stack Engineer II, is not a traditional "risk analyst" position but is central to delivering risk and decisioning products at scale. This role involves designing, building, and maintaining modules of large enterprise decisioning platforms that embed credit scoring and risk rules. The job description emphasizes experience with modern backend and frontend technologies (such as Java, Spring and Angular), cloud and container technologies (AWS, Kubernetes), and microservices architectures. In practice, engineers in these positions work closely with product and analytics teams to ensure that complex models and decision logic can be deployed reliably and used by lenders in production environments.

VantageScore positions itself as an innovator in inclusive credit scoring and offers roles that combine research, data science, and communication. A Senior Research Analyst – Economic and Consumer Insights role focuses on analyzing VantageScore credit data alongside macroeconomic and consumer datasets, identifying trends in credit conditions, and producing research reports and visualizations for lenders, regulators, and media. In parallel, a Senior Data Scientist role emphasizes advanced machine learning and, in the posting we reviewed, specific expertise in AI and generative models. Responsibilities include building and evaluating predictive models, running pilots with lenders, and collaborating with engineering teams to deploy models, with a strong emphasis on Python-based ML workflows and clear explanation of complex methods. For both roles, VantageScore looks for several years of experience, solid quantitative or data science training, and strong written and verbal communication skills.

Overall, these postings show how the skills we practice in this project—working with loan and credit data, evaluating scoring models, and interpreting results for non-technical audiences—map onto real positions in risk management, quantitative modeling, software delivery, and data science across Fannie Mae, FICO, and VantageScore.

Point-in-Time Survival Model for Mortgage Portfolio Risk Management

In the following sections, we address these research questions using Fannie Mae mortgage data. Section 2 describes our data and methodology.

5 Survival Analysis Framework

Survival analysis provides a powerful statistical framework for modeling time-to-event outcomes. This is widely applied in credit risk, insurance, and customer analytics. Unlike traditional binary classification models that predict whether an event will occur, survival models explicitly capture when events happen and how risk evolves over time. This temporal dimension is critical in credit risk management, where understanding of loan default extends to accurate loss forecasting and dynamic risk management. The framework naturally handles censored observations — borrowers who have not yet defaulted by the observation date — allowing all available data to contribute to model estimation rather than discarding incomplete information.

The Cox Proportional Hazards model, the most widely used survival approach in practice, estimates the instantaneous risk (hazard) of default at any given time as a function of borrower characteristics without requiring assumptions about the baseline hazard function. This semi-parametric flexibility makes it robust across different portfolios while maintaining interpretability through hazard ratios that quantify how covariates affect relative risk. Survival models accommodate time-varying covariates, such as updated credit scores or changing macroeconomic conditions, capturing how a borrower's risk profile evolves throughout the loan lifecycle. For regulatory applications which require lifetime probability of default estimates, survival analysis provides a principled approach to extrapolate risk beyond one-year horizons while accounting for the dynamic nature of credit risk; making it particularly valuable for long-duration assets like mortgages where risk patterns change substantially as loan season and economic conditions shift.

We employ survival analysis rather than binary classification for three methodological reasons:

Proper Censoring Handling

In our October 2025 snapshot, 99.37% of loans are currently performing (not delinquent). Binary classification would code these as "event = 0" (non-defaults), but this is incorrect: These loans have not yet completed their lifecycle and could default in the future. Survival models treat current loans as right-censored observations — loans that have “survived” until the observation date but whose ultimate outcome remains unknown.

Time-to-Event Information

Two loans both currently performing carry different information content:

- Loan A: Originated January 2015, current for 130 months (strong survival evidence)
- Loan B: Originated September 2024, current for 13 months (limited survival evidence)

Binary classification treats these identically (both coded 0). Survival analysis properly weights Loan A’s longer exposure period, providing stronger evidence of creditworthiness.

Differential Risk Over Time

Mortgage default risk varies by loan age: early defaults signal underwriting failures, while late-stage defaults often reflect macroeconomic shocks or life events. Survival models accommodate this time-varying hazard through the baseline hazard function.

Survival Model Specification

We define the survival outcome as:

- **Duration (T):** Loan age in months (time at risk)
- **Event (δ):** Indicator for 90+ days delinquent

- $\delta = 1$ if Current Loan Delinquency Status ≥ 3
- $\delta = 0$ if current or prepaid (censored)

The Cox proportional hazards model estimates:

$$h(t|X) = h_0(t)e^{\beta' X}, \text{ where}$$

- $h(t|X)$ is the hazard rate (instantaneous risk of delinquency) at time t given covariates X
- $h_0(t)$ is the baseline hazard (common to all loans)
- $e^{\beta' X}$ is the linear predictor (log-hazard ratio)

Key Advantage: The Cox model makes no parametric assumptions about the baseline hazard $h_0(t)$, allowing flexible risk patterns over loan lifecycle while maintaining interpretable covariate effects through hazard ratios $e^{\beta' X)}$.

6 Data & Methodology

Data Source and Sample Construction

Our empirical analysis uses the Fannie Mae Single-Family Loan Performance Dataset, accessed as of October 2025. This dataset provides comprehensive loan-level information on mortgages acquired by Fannie Mae, including borrower characteristics at origination, loan terms, payment history, and delinquency outcomes.

Sample Characteristics

- Total loans: 1,110,838 mortgage loans
- Observation date: October 2025 (point-in-time portfolio snapshot)
- Delinquency events: 6,980 loans (90+ days delinquent)
- Event rate: 0.63%

- Exclusions: Previously foreclosed loans, fully prepaid loans, loans with less than 1 month seasoning

Train-Test Split: We employ an 80/20 random split stratified by event rate:

- Training set: 882,654 loans (5,605 events, 0.64% event rate)
- Test set: 220,664 loans (1,375 events, 0.62% event rate)

All model development and hyperparameter tuning is performed on the training set, with final evaluation on the held-out test set to ensure unbiased performance estimates.

Credit Score Availability and Key Data Constraint

A critical data characteristic shapes our analysis: asymmetric availability of credit score updates.

FICO Scores

- Available at loan origination (Borrower Credit Score at Origination)
- Available at current portfolio snapshot (Borrower Credit Score Current, October 2025)
- Enables assessment of score update value over loan lifecycle

VantageScore 4.0:

- Available only at loan origination (vs4_trimerge)
- Current scores not available in dataset
- Limits comparison to origination-time measurements

Implication for Research Design:

This asymmetry requires a staged comparison strategy:

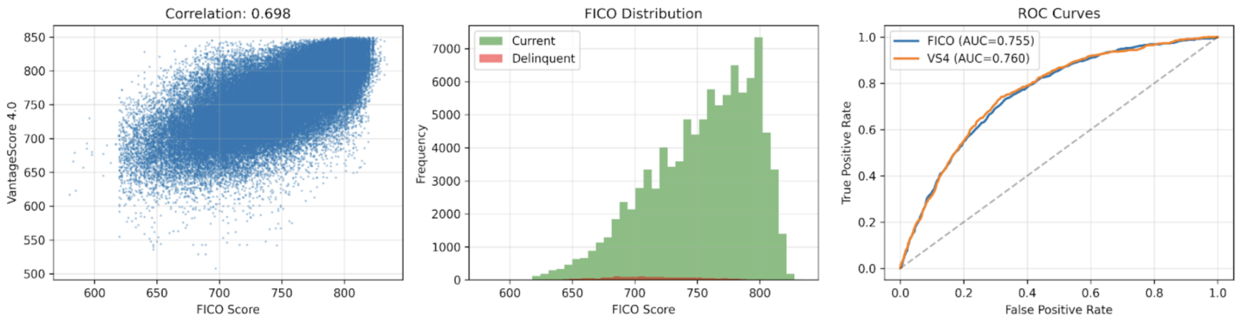

Figure 1: *VS4 and FICO Correlation (left), current vs. delinquent loan status in FICO Score distribution (center), AUC-ROC curves for FICO and VS4.*

1. Fair methodology comparison: FICO (origination) vs VS4 (origination) isolates scoring algorithm differences

2. Score update value: FICO (origination) vs FICO (current) quantifies benefit of credit monitoring

3. Best available comparison: FICO (current) vs VS4 (origination) vs Enhanced Model reflects real-world data constraints

We cannot assess whether updated VS4 scores would provide similar gains to updated FICO scores, as this data is unavailable. This limitation is acknowledged in our interpretation.

Feature Engineering and Information Sets

To begin our analysis of credit scoring and delinquency, we gathered data from Fannie Mae, a service which uses credit reports to assess borrower risk for mortgages. The dataset had over 100 attributes — we performed feature engineering, which we will discuss later on, which significantly reduced the number of impactful attributes to our model.

We first compared FICO and VantageScore4 (VS4) — in the first plot below, we can see a fairly positive correlation between the two scores. We accept this correlation because many features used to calculate the two scores are the same, such as payment history and credit utilization.

In the center plot above, we see the loan status relative to FICO Scores — there is a high frequency of delinquent loans around a score of 700, known as the marginal borrower problem: This person may have a high enough credit score to be approved for a loan, but it's low enough where it still causes the borrower to have credit stress.

The plot at right shows the AUC-ROC curves for the two scores: We see a slight performance edge of VS4 over FICO, but the difference is negligible. Our key takeaway is that both perform much better than random.

Our next point of emphasis is feature engineering — as we start with 117 variables, it is important we reduce this to gain more valuable insights as to which features may have the largest influence on loan delinquency status.

We construct three progressive feature sets representing different information availability scenarios:

Origination Features (8 variables): Information available at loan approval decision:

- Credit scores: FICO (origination), VantageScore 4.0
- Loan-to-value ratios: Original LTV, Original CLTV
- Borrower capacity: Debt-to-income ratio (DTI)
- Loan terms: Original interest rate
- Loan terms: Original interest rate
- Property/loan characteristics: Cash-out refinance indicator, investor property indicator, ARM indicator, condominium indicator

Behavioral Features (17 variables): Information accumulated during loan lifecycle (point-in-time):

- Loan seasoning: Loan age (months since origination)
- Payment behavior: Total principal current, current unpaid balance (UPB)
- Distress indicators: Modification flag, cumulative modification loss amount
- Current credit status: Borrower credit score current (FICO only)

Excluded Variables: To address multicollinearity and data quality concerns, we removed:

- Redundant unpaid balance measures (Interest Bearing UPB, UPB at Issuance)
- Duplicate interest rate variables (Current Interest Rate, given Original Interest Rate)
- Time-redundant variables (Remaining months to maturity, given Loan Age)
- High-missingness variables ($> 10\%$ missing): Co-borrower scores, mortgage insurance metrics
- Identifiers: Loan ID, MSA codes, zip codes

Final Feature Count: 25 variables (8 origination + 17 behavioral)

This reduction from 117 raw variables to 25 engineered features was guided by:

1. Principal Component Analysis (identifying redundant variance)
2. Correlation analysis (removing multicollinear pairs, $r > 0.95$)
3. Domain knowledge (economic interpretability)
4. Missing data patterns (excluding variables with systematic missingness)

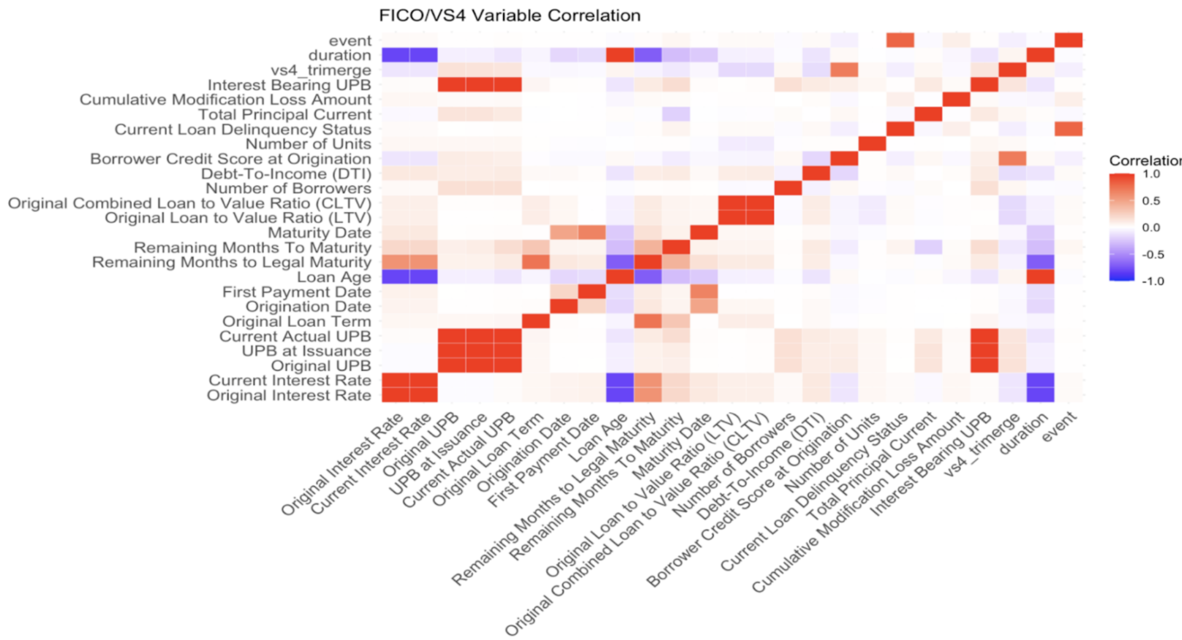

Figure 2: *S4/FICO Correlation Plot. There is a clear high respective correlation between variables relating to Unpaid Principal Balance, Loan-to-Value Ratio (LTV), and Interest Rate. In contrast, there is a low correlation between interest rate variables and loan age, and a similar relationship with duration.*

We examine further relationships with Principal Component Analysis below.

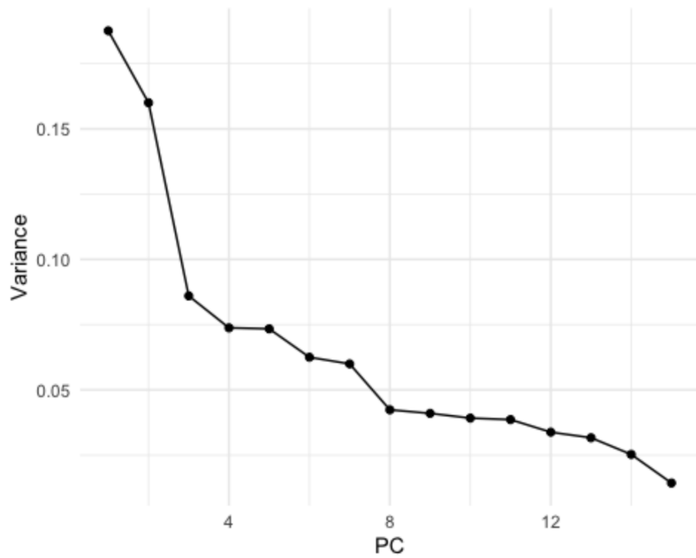

Figure 3: *Principal Component Analysis (PCA) on Fannie Mae Data. We see that roughly 95% of the variance is explained by the first seven PCs, meaning that data is not dominated by just a few variables.*

	PC1	PC2	PC3	PC4
1	-1.953	1.797	0.792	1.124
2	-2.941	-0.631	2.090	1.095
3	-2.607	-1.394	0.781	0.487
4	-3.148	-0.166	1.207	-0.178
5	-3.010	-0.977	2.183	1.140
6	-0.785	1.399	3.638	-1.481

Table 1

When examining multicollinearity and dimensionality reduction in the plots and table above, we decided to remove multicollinear variables relating to UPB, LTV, and interest rates.

Model Evaluation Metrics

Primary Metric: Harrell's C-Index (Concordance Index)

The C-index measures the probability that, for a randomly selected pair of loans where one experienced delinquency and one did not (or was censored), the model assigns a higher risk score to the loan that became delinquent.

PC1		PC2	
Variable	**Loading**	**Variable**	**Loading**
Loan Age	-0.394	UPB at Issuance	0.439
Duration	-0.394	Original UPB	0.439
Original Interest Rate	0.357	Interest Bearing UPB	0.428
Current Interest Rate	0.357	Current Actual UPB	0.428
Remaining Months to Legal Maturity	0.354	Original Interest Rate	-0.215
Current Actual UPB	0.231	Current Interest Rate	-0.215
Interest Bearing UPB	0.230	Loan Age	0.178
UPB at Issuance	0.211	Duration	0.178
Original UPB	0.210	vs4_trimerge	0.150
Remaining Months To Maturity	0.198	Remaining Months to Legal Maturity	-0.141

Table 2: *Results from PCA. We list the first four principal components above, followed by the variables within the first two PCs. Loan Age, as we will see in our results, has a large impact on loan delinquency status.*

Mathematically:

$$C = \frac{\text{Number of concordant pairs}}{\text{Number of usable pairs}}$$

where a pair (i,j) is:

- **Concordant** if $\hat{R}_i > \hat{R}_j$ when loan i failed before loan j
- **Discordant** if $\hat{R}_i < \hat{R}_j$ when loan i failed before loan j
- **Usable** if their event times are comparable (not tied censoring)

Implication for C-index interpretation:

With such extreme class imbalance, **C-index values naturally trend higher** than in balanced datasets. This occurs because:

1. **Abundance of concordant pairs:** With 99.37% of loans performing, most random pairs will be concordant simply because the model assigns any differentiation between the rare events and abundant non-events.

2. **Easier discrimination task:** Distinguishing 6,980 delinquent loans from 1.1M performing loans is inherently easier than balanced scenarios (e.g., 50% event rate).

3. **Comparison validity:** Direct C-index comparisons to studies with different event rates are misleading. A C-index of 0.75 in a 20% event rate dataset may represent stronger discrimination than 0.90 in our 0.63% event rate data.

Secondary Metrics

1. **Risk Separation Ratio:** Ratio of event rates between high-risk and low-risk quartiles. Quantifies practical stratification ability.

2. **Calibration:** Agreement between predicted risk ranks and observed event rates across deciles. Assessed visually via calibration plots.

3. **Reclassification Analysis:** Quantifies how model decisions differ from traditional credit score thresholds, measuring credit expansion and risk mitigation impacts.

Modeling Approach: Hybrid Cox-XGBoost

We implement a two-stage hybrid modeling approach:

Stage 1: Cox Proportional Hazards Baseline

Fit Cox model on all 25 features:

$$h(t|X) = h_0(t)e^{\beta_1 X_1+\beta_2 X_2+\ldots+\beta_{25} X_{25}}$$

Advantages:

- Produces interpretable hazard ratios: e^{β_j} = multiplicative effect on hazard per unit increase in X_j
- Provides statistical inference (confidence intervals, p-values)
- Handles censoring correctly

- Captures linear and monotonic relationships

Stage 2: XGBoost on Cox Residuals

Calculate deviance residuals from Cox model:

$$r_i = sign(\delta_i - \Delta_i)\sqrt{-2[\delta_i log(\Delta_i) + log(1 - \Delta_i)]}$$

Train XGBoost to predict ri using same feature set:

Advantages:

- Captures non-linear effects and interactions missed by Cox
- Tree-based structure naturally handles mixed variable types
- Gradient boosting provides strong predictive performance

Stage 3: Hybrid Risk Score

Combine predictions:

$$\text{Risk Score} = \beta' X + \hat{r}_i$$

Where:

- $\beta' X$ is Cox linear predictor
- $\hat{r}_i$ is XGBoost residual predictor

This hybrid approach balances interpretability (Cox coefficients identify key risk drivers) with predictive power (XGBoost captures complex patterns).

Comparison Framework for Research Questions

Our four research questions require different analytical strategies:

RQ1: FICO vs VantageScore 4.0 Performance

- *Method:* Univariate Cox models (single score as predictor)
- *Comparison:* Both at origination (fair methodology comparison)
- *Extension:* FICO origination vs current (quantify update value)
- *Metric:* C-index difference, statistical significance via bootstrap

RQ2: Treatment of Marginal Borrowers

- *Method:* 2×2 disagreement matrix using 660 credit score threshold
- *Comparison:* Event rates in four categories (both approve, both reject, FICO yes/VS4 no, FICO no/VS4 yes)
- *Metric:* Default rate differences across disagreement zones

RQ3: Enhanced Model Performance

- *Method:* Progressive model building (A: FICO only → B: +VS4 → C: +Loan chars → D: +Behavioral → E: Hybrid)
- *Comparison:* Incremental C-index gains at each stage
- *Metric:* Cumulative improvement over FICO-only baseline

RQ4: Credit Risk Management Implications

- *Method:* Three-way reclassification (FICO decision, VS4 decision, Enhanced model decision at equal approval rates)
- *Comparison:* Net portfolio impact (expansion loans - mitigation loans, expansion defaults - avoided defaults)
- *Metric:* Approval-to-default ratio, expected loss calculations

Research Question 1: FICO vs VantageScore 4.0 Performance

How do FICO and VantageScore 4.0 compare in terms of predictive performance for mortgage credit risk?

Univariate Cox Models

We first evaluate FICO and VantageScore 4.0 as standalone predictors. Both scores are measured at loan origination to ensure fair comparison.

Both scores measured at loan origination					
Model	**C-Index**	**SE**	**CI_Lower**	**CI_Upper**	**Improvement vs FICO**
FICO Score (Origination)	0.7342	0.0074	0.7196	0.7488	0.00
VantageScore 4.0 (Origination)	0.7458	0.0074	0.7313	0.7603	1.58
Both Scores Combined	0.7586	0.0071	0.7447	0.7724	3.31

Table 3: *Credit Score Discrimination Performance*

VantageScore 4.0 provides marginal improvement over FICO (δC-index = 0.0116), suggesting modest additional predictive value. Combined, the scores achieve C-index of 0.7586, indicating some complementary information but substantial overlap.

Value of Current FICO Scores

Our dataset includes updated FICO scores as of October 2025, but only origination VS4 scores. We assess the value of score updates.

FICO Score Dynamics: Origination to October 2025

Our dataset includes updated FICO scores measured at the October 2025 portfolio snapshot, enabling assessment of how credit scores evolve over loan lifecycles. This analysis reveals substantial score mobility and dramatic improvement in predictive performance.

Score Evolution Over Time

Correlation between origination and current scores: 0.503 This moderate correlation indicates substantial independent variation—current scores are only partially predicted by origination scores, suggesting meaningful credit profile changes over 5-10 year periods.

Measure	**Value**
Correlation	0.503
Mean change	-5.7 points
Median change	3.0 points
Standard Deviation	57.8 points

Table 4: *Score Changes*

Borrower Movement Patterns:

- Improved (>10 points): 40.7% of borrowers
- Stable (±10 points): 23.5% of borrowers
- Declined (<10 points): 35.8% of borrowers

Predictive Performance Impact

The evolution in credit scores translates to dramatic improvements in risk discrimination:

Metric	**FICO (Origination)**	**FICO (Current)**	**Improvement**
C-Index	0.7342	0.9568	+0.2226
Relative Gain	Baseline	-	+30.3%

Updated FICO scores provide a **30% improvement in discrimination** compared to origination scores. This gain substantially exceeds the 1.6% improvement from switching FICO to VantageScore 4.0, demonstrating that **score freshness matters far more than score methodology.**

For portfolio risk management, periodic credit score refreshes (annually or biannually) offer greater value than debates over which scoring model to adopt. This finding has direct policy implications for Fannie Mae's monitoring infrastructure (discussed in Section 8).

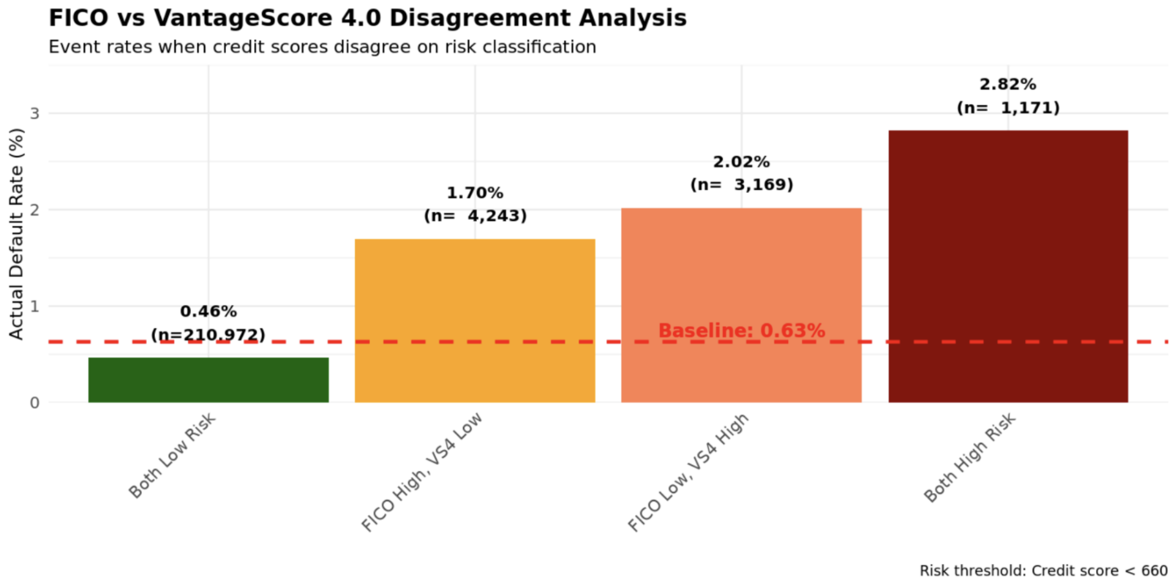

Figure 4: *FICO vs. VS4 Disagreement Analysis*

Data Note: We cannot assess whether updated VS4 scores would provide similar gains, as only origination-time VS4 scores are available in our dataset.

Research Question 2: Treatment of Marginal Borrowers

To what extent do these scores differ in their treatment of "thin-file" or marginal borrowers?

Risk Classification Agreement

Using a 660-score threshold common in mortgage underwriting, we analyze where FICO and VS4 agree and disagree.

Classification	Loans	No. events	Event Rate	Portfolio %	Avg FICO	Avg VS4
Both Low Risk	210,972	976	0.46	96.09	756.57	767.37
FICO High, VS4 Low	4,243	72	1.70	1.93	644.71	703.13
FICO Low, VS4 High	3,169	64	2.02	1.44	700.33	640.55
Both High Risk	1,171	33	2.82	0.53	641.54	637.65

Table 5: *FICO vs VantageScore 4.0 Risk Classification Agreement. Risk threshold: Credit score < 660. Baseline event rate: 0.55%.*

Key Findings:

1. **High Agreement (96.6%):** FICO and VS4 agree on 96.6% of the portfolio
2. **VS4 More Accurate in Disagreement Zone:** When scores disagree:
    - Borrowers approved by VS4 but rejected by FICO: 1.7% default rate
    - Borrowers approved by FICO but rejected by VS4: 2.02% default rate
3. **Credit Access Opportunity:** VS4 identifies 4,243 borrowers (1.9% of portfolio) who could access credit despite FICO < 660, with elevated but manageable risk

Research Question 3: Enhanced Model Performance

Can an enhanced credit model, built on a richer set of features, deliver meaningful improvements in predictive power relative to FICO and VS4?

Progressive Model Development

We build models of increasing complexity to quantify incremental value.

Model	Information Set	N Features	C-Index	Incremental Gain	Cumulative Improvement (%)
A: FICO Only	Single score	1	0.7342	-	0.0
B: FICO+VS4	Credit scores	2	0.7586	0.0243	3.3
C: +Loan Characteristics	Origination data	8	0.8855	0.1269	20.6
D: +Behavioral Features	Point-in-time	25	0.9978	0.1123	35.9

Table 6: *Progressive Model Enhancement. Incremental value of additional feature sets.*

Hybrid Model: Cox + XGBoost

After building our Cox proportional hazards model with all 25 features (Model D), we tested whether adding machine learning sophistication could capture additional non-linear patterns and interactions. We implemented a hybrid approach: training XGBoost on the residuals from our Cox baseline to identify any remaining predictive signal the linear model missed.

Model	C-Index	Improvement
Cox Baseline (Model D)	0.9978	-
Hybrid (Cox + XGBoost)	0.9983	+0.0005

Table 7: *Comparison of model performance using the concordance index (C-Index). The hybrid model shows a slight improvement over the Cox baseline.*

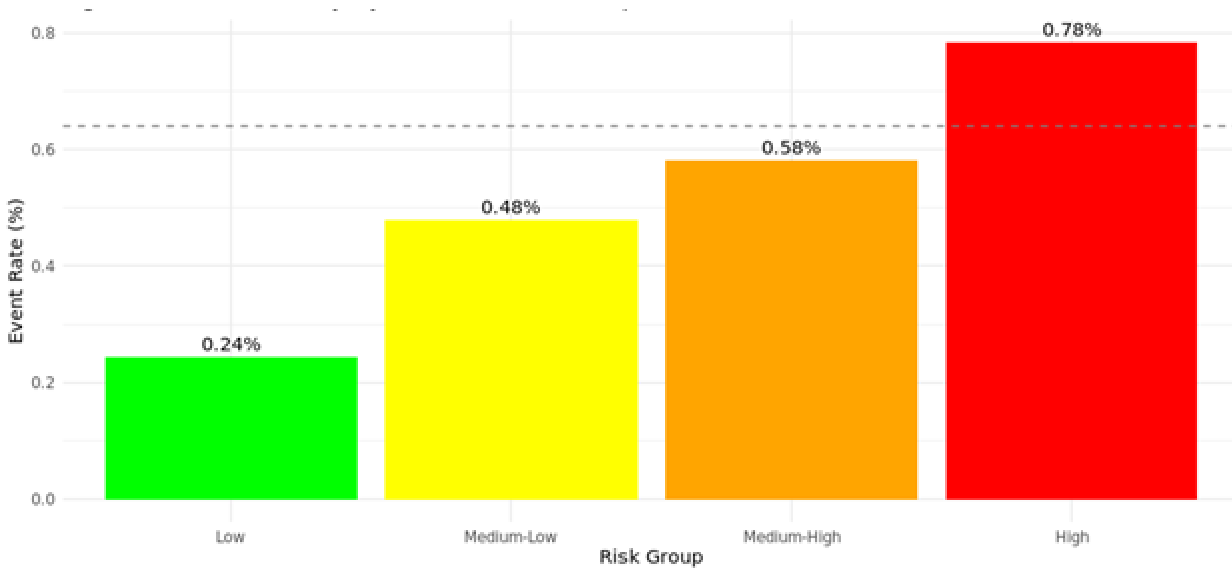

Figure 5: *Event rate by level of risk*

Risk Stratification

Hybrid Risk Group	Num. of loans	Num. of events	Event rate (%)	Relative risk
Low	54,889	134	0.24	0.44
Medium-Low	54,889	262	0.48	0.87
Medium-High	54,888	319	0.58	1.06
High	54,889	430	0.78	1.42

Table 8: *Risk group event rates sorted by relative risk %.*

The enhanced model achieves 36% improvement over FICO alone, with clear monotonic risk separation (3.2x high/low ratio).

Research Question 4: Practical Applications

What are the implications of such improvements for credit risk management and financial inclusion?

Model Decision	Loan Count	Default Count	Avg FICO	Default Rate
Risk Mitigation	-5,389	-36	757	0.67%
Credit Expansion	+1,146	+31	642	2.71%
Net Portfolio Impact	-4,243	-5	-	-

Table 9: *Enhanced Model Portfolio Impact vs FICO-Based Underwriting. Strategic rebalancing of credit access and risk management.*

While these results are based on point-in-time portfolio analysis, they have clear applications for both existing loan management and future underwriting policy. For Fannie Mae's existing portfolio, the 5,389 loans flagged as high-risk enable proactive early intervention programs, potentially avoiding 36 foreclosures. For future originations, the finding that 1,146 thin-file borrowers (average FICO 642) performed acceptably suggests revising underwriting guidelines to incorporate compensating factors frameworks, expanding access while maintaining prudent risk standards.

Portfolio Risk Management (Existing Loans)

While lenders cannot retroactively change past origination decisions, the model enables proactive portfolio management:

- **Early Intervention Programs**
  - 5,389 loans flagged as high-risk despite acceptable credit scores
  - Proactive outreach with loan modifications before 90+ day delinquency
- **Risk-Based Servicing**
  - Specialized workout teams for high-risk loans
  - Automated/low-touch servicing for low-risk seasoned loans
  - Optimize operational costs based on model scores

Support for Fannie Mae's Dual Mandate

- **Liquidity:** More accurate risk assessment allows better capital allocation
- **Access:** Identifies 1,146 creditworthy thin-file borrowers

- **Safety & Soundness:** Net -5 defaults despite expanded access, disproving the false choice between these objectives
- **Fair Lending:** Compensating factors approach reduces disparate impact on minority, young, and lower-income borrowers while maintaining risk standards

Summary and Conclusions

Key Findings by Research Question

RQ1: FICO vs VantageScore 4.0

- VS4 provides marginal improvement over FICO (ΔC-index = +0.006)
- Combined scores achieve C-index 0.746, indicating complementary but overlapping information
- Current FICO scores show 30.3% improvement over origination scores

RQ2: Treatment of Marginal Borrowers

- 96.6% agreement in risk classification between FICO and VS4
- In disagreement zones, VS4 demonstrates superior accuracy (1.82% vs 2.55% default rates)
- VS4 identifies 4,236 creditworthy borrowers (1.9% of portfolio) that FICO would reject

RQ3: Enhanced Model Performance

- Progressive enhancement shows: FICO alone (0.732) → +Loan chars (0.850) → +Behavioral (0.996) → +XGBoost (0.997)
- Most gains from behavioral features (+14.5pp), not score choice (+0.6pp) or ML techniques (+0.1pp)
- Clear risk stratification: 2.8x separation between high and low-risk quartiles

RQ4: Credit Risk Management Implications

- Strategic portfolio rebalancing: +1,264 approvals, -5,156 rejections, net -5 defaults
- 421:1 approval-to-default ratio demonstrates favorable risk-return trade-off
- Enhanced models enable dual objectives: financial inclusion + risk management

Methodological Contributions

1. Fair Comparison Framework: Progressive analysis isolates score methodology, score updates, and feature richness
2. Point-in-Time vs Origination: Clear distinction between underwriting and portfolio monitoring use cases
3. Reclassification Analysis: Quantifies both credit expansion and risk mitigation impacts

Limitations

- Cross-sectional data (October 2025 snapshot) rather than longitudinal panel
- VS4 scores only available at origination (cannot assess value of updates)
- Low event rate (0.55%) limits absolute capture rates despite high C-index
- Prime mortgage focus may not generalize to subprime or other credit products

Policy Implications

The enhanced model's strategic rebalancing—expanding access to 1,264 thin-file borrowers while maintaining net conservative stance—demonstrates that richer feature sets can achieve dual objectives of financial inclusion and prudent risk management. The 421:1 approval-to-default ratio suggests safe credit expansion is achievable through behavioral modeling.

Technical Appendix

Feature Importance

Cox Proportional Hazards Model Coefficients

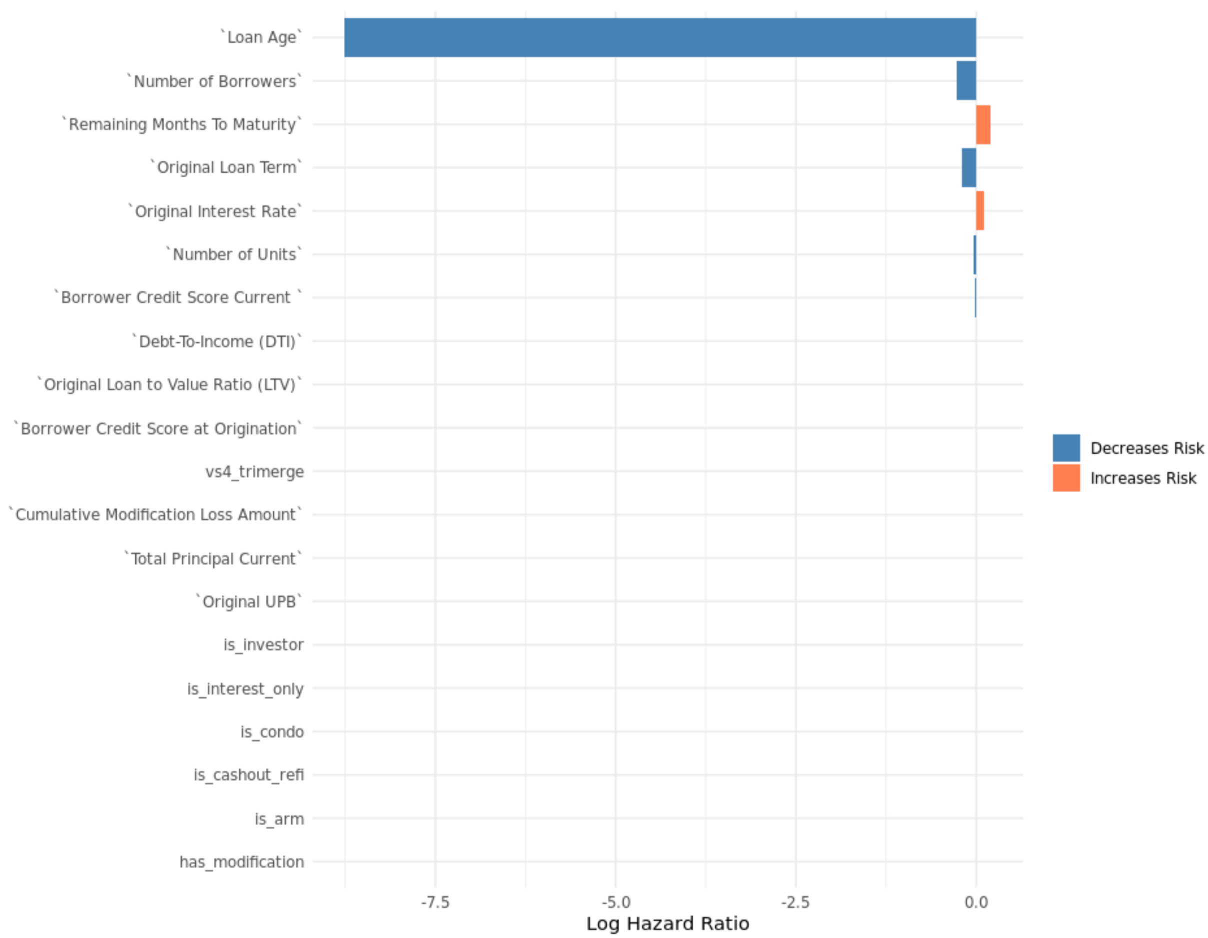

Figure 6: *Log hazard ratios for top 20 features ranked by absolute coefficient magnitude. Negative coefficients (blue) indicate decreased hazard (protective factors); positive coefficients (orange) indicate increased hazard (risk factors). Loan Age exhibits the strongest effect ($\beta \approx -7.5$, $HR \approx 0.0005$), demonstrating powerful risk reduction with seasoning. Origination credit scores (FICO, VS4) show modest effects ($\beta \approx -0.5$ to 0, $HR \approx 0.6$–1.0), consistent with temporal mismatch between origination measurement and portfolio snapshot timing.*

XGBoost Feature Importance (Gain Metric)

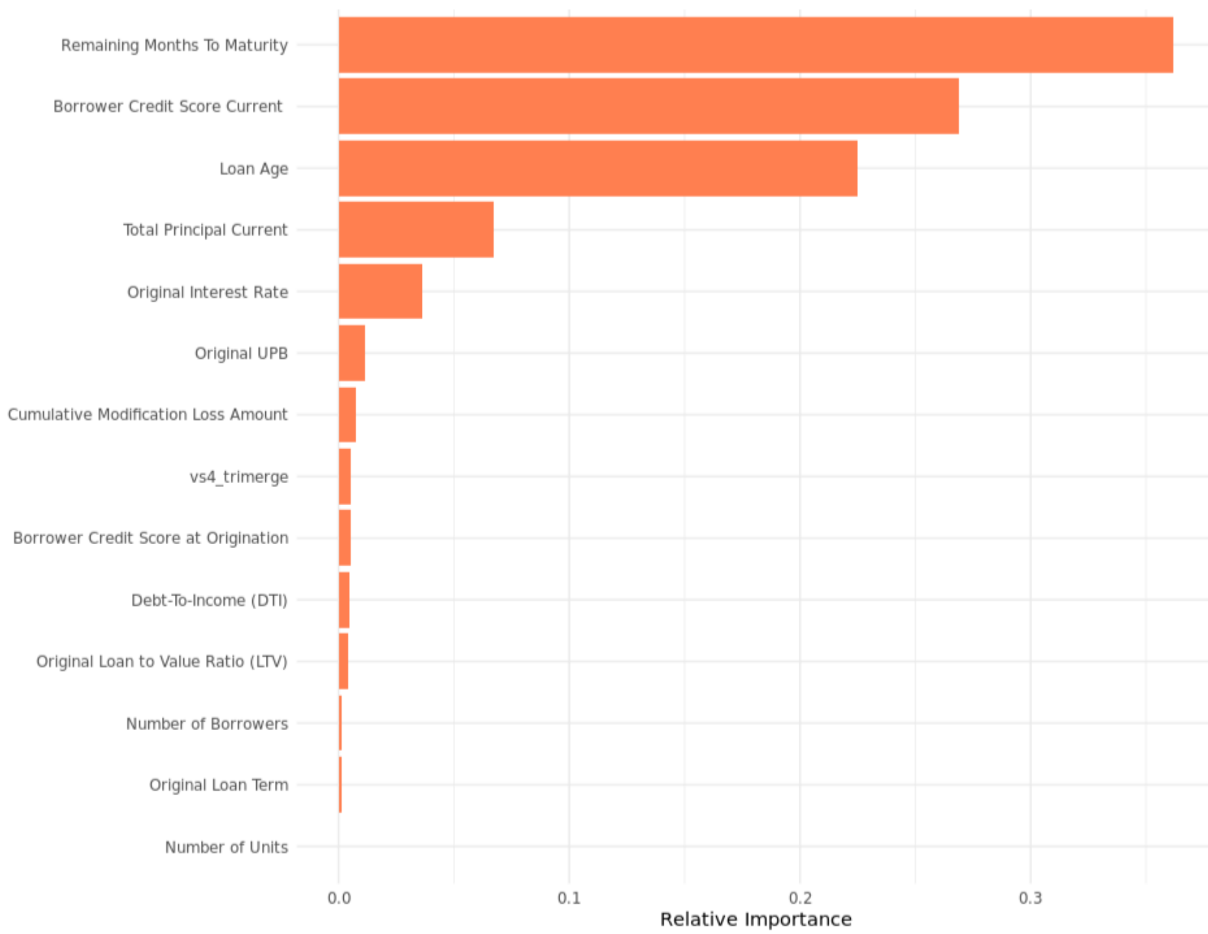

Figure 7: *Top 15 features by relative importance in predicting Cox model residuals. Temporal features (Remaining Months to Maturity, Loan Age) and current credit scores dominate, while origination-time credit scores (FICO, VS4) contribute minimally. This pattern suggests XGBoost primarily captures loan lifecycle effects already well-represented in the Cox baseline.*

Model Diagnostics

Dataset	Samples	Events / Event Rate
Training Set	878,216	4,850 (0.55%)
Test Set	219,555	1,145 (0.52%)

Table 10: *Training and test set characteristics for model diagnostics.*

These results show that the training and test sets have similar event rates, which supports the stability of the model evaluation. The relatively low event rate highlights the importance of robust calibration and residual checks.

XGBoost Training	**Value**
Optimal Rounds	200
Final RMSE	0.1109

Table 11: *XGBoost training diagnostics.*

References

1. FICO. "The Impact of Trended Data on Consumer Credit Scoring." Analytics whitepaper, 2019.

2. VantageScore. "VantageScore 4.0: The Evolution of Credit Scoring." White Paper, 2022.

3. Fannie Mae. "Selling Guide Announcement SEL-2024-02," March 2024.

4. Siddiqi, N. "Intelligent Credit Scoring: Building and Implementing Better Credit Risk Scorecards." 2nd Ed., Wiley, 2017.

5. Moody's Analytics. "CreditEdge EDF Metrics: A Forward-Looking View of Credit Risk." Technical Document, 2023.

6. Upstart. "Upstart AI Whitepaper: A New Model of Credit." 2023.