Running Head: Estimating Mortgage Prepayment Timing with Machine Learning

Understanding Mortgage Prepayment Timing and Its Drivers: A Comparative Machine Learning Approach

Sirui Zhang[1]

[1] The School of Operations Research and Information Engineering, Cornell University, Ithaca, NY, USA

Abstract

This study investigates the timing of mortgage prepayment using a comprehensive dataset of U.S. residential mortgage loans spanning 192 months from origination. Unlike most existing research that focuses on the likelihood of prepayment, we focus exclusively on loans that were actually prepaid, and model how many months after origination the prepayment occurred. In other words, we treat the prepayment timing as a continuous variable and aim to estimate its duration based on borrower and loan characteristics. We apply and compare three regression models—linear regression, random forest regression, and multilayer perceptron (MLP)—to examine the extent to which borrower, loan, and property characteristics explain variation in prepayment timing. Our findings show that the random forest model significantly outperforms linear regression and MLP, achieving the lowest prediction error and highest explanatory power. Mortgage interest rate emerges as the dominant predictor, confirming its role as a central driver of prepayment behavior. Linear regression underperforms due to violations of its statistical assumptions, while MLP performs moderately and presents opportunities for improvement through fine-tuning. This study contributes to the mortgage analytics literature by introducing a model comparison framework for analyzing prepayment timing and identifying its key influencing factors.

Introduction

Mortgage prepayment—the act of a borrower repaying their mortgage ahead of schedule—is a fundamental factor in mortgage risk analysis and plays a crucial role in evaluating the risk exposure of mortgage insurers. For institutions such as mortgage insurance companies, understanding prepayment timing is essential for assessing how long a policy remains in force and for estimating the number of loans that require coverage over time. Accurate predictions of prepayment behavior impact the valuation of mortgage portfolios, interest rate risk modeling, and the design of lending strategies. Since the 1980s, a substantial body of empirical research has investigated the determinants of mortgage prepayment, focusing primarily on whether or not prepayment occurs within a given timeframe. Commonly studied factors include borrower credit scores, loan-to-value (LTV) ratios, interest rates, loan terms, property types, and macroeconomic indicators such as

housing prices and unemployment rates (Campbell & Dietrich, 1983; DENG, 1997; Lin et al., 2010; Schwartz & Torous, 1993).

However, an important gap remains: while much attention has been given to prepayment likelihood, relatively few studies have addressed the exact timing of prepayment—how many months after loan origination a borrower tends to prepay. Understanding the timing, rather than the binary occurrence, of prepayment provides deeper insights into cash flow dynamics and improves forecasting precision for investors and financial institutions. Furthermore, past modeling approaches have often relied on restrictive assumptions such as linear relationships and additive effects, which may fail to capture the complex, nonlinear interactions among risk factors and borrower behavior.

Recent advancements in machine learning, especially tree-based models and neural networks, offer new opportunities to overcome these limitations (Hung & Lo, 2024; Krasovytskyi & Stavytskyy, 2024; Pillai et al., 2020; Sadhwani et al., 2021). These methods can model highly nonlinear relationships and capture complex feature interactions without imposing rigid parametric forms. For example, Giesecke et al., (2016) employed deep learning to explore borrower transitions across multiple mortgage states and demonstrated the importance of modeling nonlinearities and interactions for accurate mortgage risk prediction. However, even in that advanced modeling framework, the analysis primarily emphasized transitions such as delinquency or prepayment occurrence, without centering on prepayment timing as a continuous outcome variable.

This study aims to address this research gap by analyzing the specific timing of mortgage prepayment using a publicly available dataset of U.S. mortgage loans originated in 1991 and tracked monthly over a period of 212 months. Unlike most existing studies, we treat the number of months until prepayment as a continuous target variable and apply three regression-based predictive models—linear regression, random forest regression, and multilayer perceptron (MLP)—to examine which borrower, loan, and economic characteristics most strongly influence prepayment timing.

By comparing model performance and interpreting variable importance, this paper contributes to the literature in two key ways: (1) It introduces a regression-based framework for estimating prepayment timing rather than mere probability, and (2) it identifies nonlinear drivers of early repayment using interpretable machine learning techniques. Our findings offer practical implications for lenders seeking to optimize loan structures and for investors aiming to assess prepayment risks with greater granularity.

Methods

Data preparation

This study utilizes a publicly available mortgage dataset retrieved from the Kaggle repository Loan Prepayment Dataset, which tracks loan-level monthly performance records of U.S. residential mortgage loans originated in 1991 covering 192 months of mortgage repayment histories. A comprehensive set of variables was extracted, including borrower-level financial characteristics (e.g., credit score, debt-to-income ratio [DTI]), loan characteristics (e.g., original unpaid balance [OrigUPB], original interest rate, mortgage rate),

and property-level information (e.g., occupancy status, property state, loan purpose). To ensure the reliability of the model and minimize noise from outliers or data artifacts, we excluded observations with full-term repayments (i.e., loans repaid exactly on schedule with no early payoff) and extremely early prepayments (under 3 months), as these may reflect data errors, refinancing noise, or special programmatic circumstances rather than meaningful behavioral decisions.

We also created a new feature called *RateSpread*, defined as the difference between the original interest rate and the prevailing mortgage rate at a given time. This variable serves as a proxy for refinancing incentive—widely acknowledged in the mortgage literature as a primary driver of prepayment behavior (e.g., Schwartz and Torous, 1989; Agarwal et al., 2017). Higher RateSpread values indicate more favorable conditions for prepayment due to potential interest savings. The prevailing mortgage rate used for this calculation was sourced from the Federal Reserve Economic Data (FRED), which provides weekly averages of the 30-year fixed mortgage rate in the United States. Categorical variables such as loan purpose and occupancy status were label-encoded into numeric representations for modeling. For numerical features, we applied standard normalization (zero mean, unit variance) for models that are sensitive to scale. All models—including linear regression, random forest, and MLP—were trained on the same normalized dataset to ensure consistency across comparisons.

Comparison of linear and non-linear models

To explore the effectiveness of different predictive strategies, we applied and compared three regression models: linear regression, random forest regression, and multilayer perceptron (MLP). The goal was twofold: (1) to evaluate the extent to which prepayment timing can be predicted using borrower and loan features, and (2) to identify the most influential predictors. Model performance was evaluated using two commonly used indicators: Mean Absolute Error (MAE), which measures the average magnitude of prediction error in months, and R-squared (R^2), which represents the proportion of variance in the dependent variable explained by the model. A lower MAE and higher R^2 indicate better predictive performance. To determine feature importance, we relied on the mean decrease in impurity (MDI) metric derived from the random forest model. This indicator reflects how much each feature contributes to reducing variance in the model's decision trees. Features with higher MDI scores are considered more influential in predicting prepayment timing.

Results

Model performance

We evaluated the predictive accuracy of three models: linear regression, random forest regression, and multilayer perceptron (MLP), using Mean Absolute Error (MAE) and R-squared (R^2) as key indicators. The results showed that Linear Regression produced a moderate performance with an MAE of 14.08 months and R^2 of 0.655. While the model captured some linear trends between the input features and prepayment timing, residual

diagnostics revealed clear evidence of heteroscedasticity (non-constant variance) and non-normality. These findings suggest that the linear model may not fully account for nonlinear interactions or feature interdependencies. Furthermore, it struggled to predict very early or very late prepayment behaviors, likely due to its rigid assumptions.

However, for the non-linear models, Random Forest Regression significantly outperformed the other models, with an MAE of 3.01 months and an R^2 of 0.953. This model demonstrated excellent fit and predictive power, effectively capturing complex, nonlinear relationships between borrower characteristics, loan features, and prepayment timing. Its robustness to outliers and multicollinearity contributed to its superior performance. The relatively low MAE indicates that the model's predictions of the number of months until prepayment were close to the actual values across a broad range of observations. Multilayer Perceptron (MLP) achieved intermediate performance, with an MAE of 8.57 months and an R^2 of 0.835. While it captured some nonlinear patterns, it did not perform as well as the random forest model. This may be due to the neural network's sensitivity to data scale, architecture choice, and limited data volume. Its higher variance and longer training time, compared to random forests, could have limited its predictive capacity in this context.

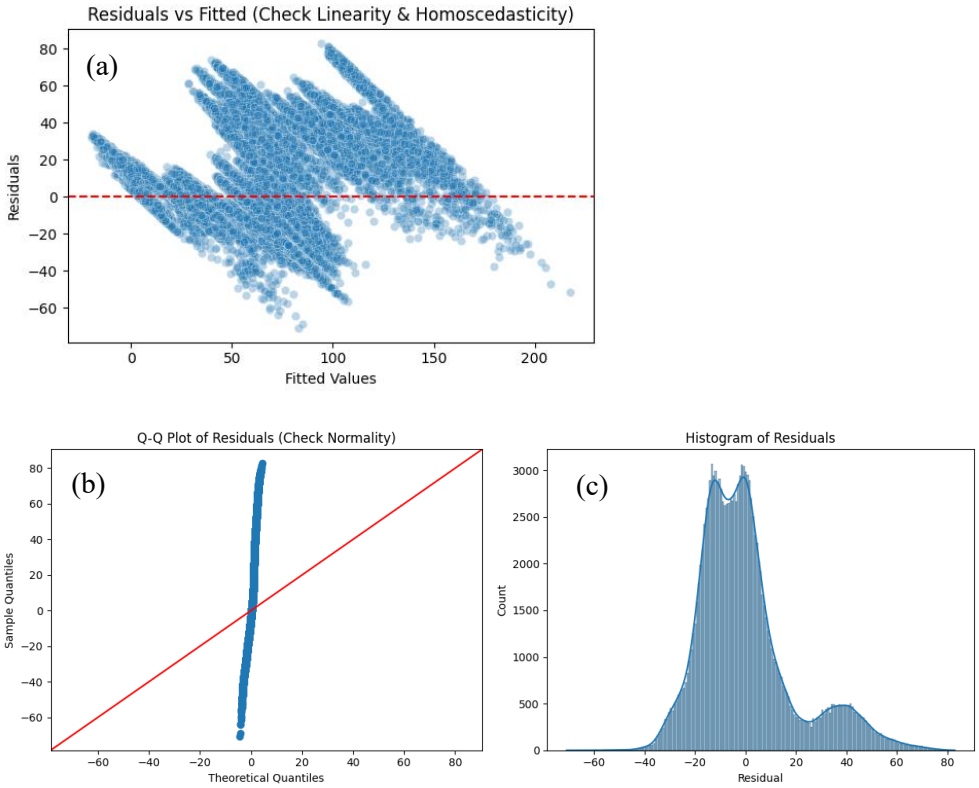

Figure 1.

(a) Residuals versus fitted values from the linear regression model. The fan-shaped pattern indicates violations of the homoscedasticity assumption. (b) Q–Q plot of residuals from the linear regression model. The strong deviation from the diagonal suggests non-normality of residuals. (c) Histogram of residuals from the linear regression model. The skewed and multimodal distribution further confirms the violation of normality.

The most influential factors of prepayment timing

To understand which variables contributed most to predicting prepayment timing, we analyzed feature importance using the Mean Decrease in Impurity (MDI) metric from the random forest model. We found that mortgage interest rate emerged as the most influential factor. This is consistent with economic theory and prior empirical work showing that lower interest rates create strong incentives for refinancing or early repayment. Original unpaid principal balance (OrigUPB) also played a major role, reflecting the relationship between loan size and borrower payoff strategies. Months delinquent indicated borrower repayment behavior and financial stress, influencing their ability or urgency to prepay. Credit score was moderately important, as it reflects borrower creditworthiness and access to alternative financing options. Other features such as DTI, LTV, occupancy status, and loan purpose had smaller but still detectable contributions.

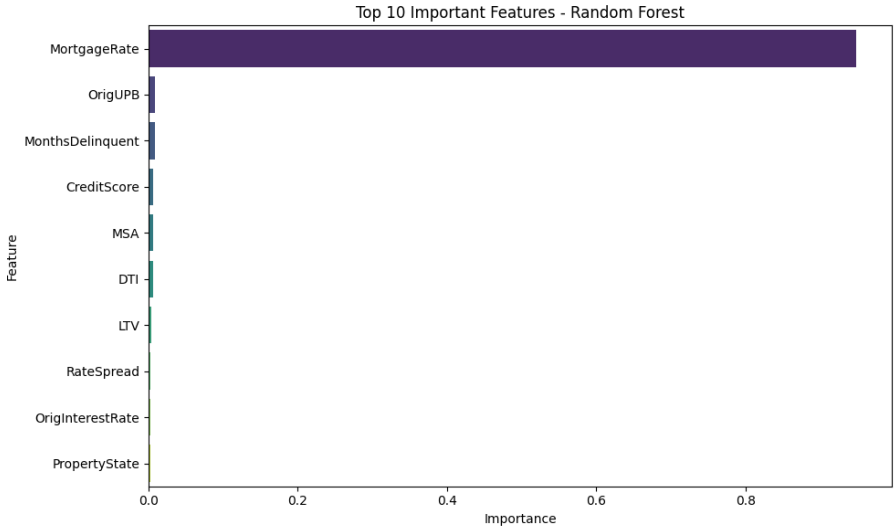

Figure 2.

Top ten influencing factors of mortgage prepayment timing, derived from the random forest model using the mean decrease in impurity (MDI).

Discussion

The results of this study highlight key differences in model performance and offer insights into the drivers of mortgage prepayment timing. Among the three models tested, linear regression performed the weakest. This can largely be attributed to its underlying assumptions—linearity, homoscedasticity, and normality of residuals—which were clearly violated in our dataset. As a result, the linear model failed to capture the complex, nonlinear relationships between borrower behavior and loan features (Sadhwani et al., 2021). The random forest model achieved the highest predictive accuracy, with a remarkably low MAE and high R^2, indicating that it was well-suited to the problem. MLP offered a moderate improvement over linear regression but was unable to match the performance of random

forests. Although it captured some nonlinear effects, its performance may have been constrained by architectural limitations or lack of parameter tuning. In future work, fine-tuning the MLP's hyperparameters and experimenting with deeper or more complex network structures could enhance its effectiveness.

Notably, the predictive performance of the random forest model was driven almost entirely by a single factor: the mortgage interest rate. This finding is consistent with established economic theory that identifies interest rate incentives as a central motivation for mortgage refinancing and early repayment (Agarwal et al., 2016). The strength of this one variable suggests that fluctuations in market rates relative to the original loan rate are critical to prepayment behavior and should be closely monitored by lenders and investors. Beyond performance metrics, the models collectively underscore that mortgage prepayment timing is a highly nonlinear process, dominated by financial incentives rather than borrower demographics alone. While random forest models are not as interpretable as linear regression, their ability to extract meaningful variable importance rankings makes them valuable for applied forecasting in mortgage analytics. These findings have direct implications for financial institutions. Lenders can use such models to forecast prepayment behavior and manage interest rate risk more effectively, while policymakers might consider incorporating such behavioral predictors into macroprudential monitoring frameworks.

Conclusion

This study demonstrates the applicability of machine learning models for analyzing mortgage prepayment timing, emphasizing the explanatory power of loan-level and borrower-level features. Rather than predicting whether prepayment occurs, we focused on estimating when it happens, framing prepayment timing as a continuous outcome. The random forest model achieved the best performance, revealing that mortgage interest rate overwhelmingly determines early repayment behavior. This suggests that interest rate dynamics play a central role in borrower decision-making and should remain a primary consideration in mortgage portfolio risk assessments. In contrast, linear regression exhibited limited effectiveness due to violations of key assumptions such as linearity and homoscedasticity. The MLP model performed moderately well, suggesting potential for further improvement with architectural tuning and larger-scale data. By reframing the problem from prepayment likelihood to prepayment timing, this study contributes to the literature on mortgage behavior and offers practical tools for lenders and analysts seeking granular forecasts of loan performance. Future research may build on this work by incorporating temporal macroeconomic indicators, conducting model fine-tuning, or integrating survival analysis to better capture time-to-event dynamics.

Appendix

The complete Python code for this project, including data preprocessing, model training, evaluation, and visualization, is available at the following link:
https://colab.research.google.com/drive/15nq_BJEDpe5yJYA2S6NyMDsdNS039S4k

References

Agarwal, S., Rosen, R. J., & Yao, V. (2016). Why Do Borrowers Make Mortgage Refinancing Mistakes? *Management Science*, *62*(12), 3494–3509. https://doi.org/10.1287/mnsc.2015.2272

Campbell, T. S., & Dietrich, J. K. (1983). The Determinants of Default on Insured Conventional Residential Mortgage Loans. *The Journal of Finance*, *38*(5), 1569–1581. https://doi.org/10.1111/j.1540-6261.1983.tb03841.x

DENG, Y. (1997). Mortgage Termination: An Empirical Hazard Model with a Stochastic Term Structure. *The Journal of Real Estate Finance and Economics*, *14*(3), 309–331. https://doi.org/10.1023/A:1007758412993

Giesecke, K., Sirignano, J., & Sadhwani, A. (2016). *Deep learning for mortgage risk*. Working paper, Stanford University. https://pdfs.semanticscholar.org/13b7/d1a49896182ea3fc38e656d26c3640d0d6f1.pdf

Hung, M.-T., & Lo, H.-C. (2024). Risk Analysis of Mortgage Loan Default for Bank Customers and AI Machine Learning. *Journal of Applied Finance & Banking*, *14*(6), 33–50.

Krasovytskyi, D., & Stavytskyy, A. (2024). Predicting mortgage loan defaults using machine learning techniques. *Ekonomika*, *103*(2), 140–160.

Lin, T. T., Lee ,Chia-Chi, & and Hu, M.-S. (2010). The determinants of mortgage defaults and prepayments. *Journal of Information and Optimization Sciences*, *31*(5), 993–1009. https://doi.org/10.1080/02522667.2010.10700007

Pillai, S. G., Woodbury, J., Dikshit, N., Leider, A., & Tappert, C. C. (2020). Machine Learning Analysis of Mortgage Credit Risk. In K. Arai, R. Bhatia, & S. Kapoor (Eds.), *Proceedings of the Future Technologies Conference (FTC) 2019* (Vol. 1069, pp. 107–123). Springer International Publishing. https://doi.org/10.1007/978-3-030-32520-6_10

Sadhwani, A., Giesecke, K., & Sirignano, J. (2021). Deep Learning for Mortgage Risk*. *Journal of Financial Econometrics*, *19*(2), 313–368. https://doi.org/10.1093/jjfinec/nbaa025

Schwartz, E. S., & Torous, W. N. (1993). Mortgage Prepayment and Default Decisions: A Poisson Regression Approach. *Real Estate Economics*, *21*(4), 431–449. https://doi.org/10.1111/1540-6229.00619