Predicting Bank Income Components Using Machine Learning Methods

Tan Schelling*

June 29, 2025

Abstract

This paper explores the performance of machine learning (ML) methods in predicting key components of bank income for the purpose of financial stress testing. Stress testing is an important tool for assessing financial stability and requires accurate forecasts of bank profitability under adverse macroeconomic scenarios. Using a comprehensive panel of U.S. commercial bank data from 2001 to 2025, I compare the predictive performance of a baseline linear regression model against several tree-based ML and an ensemble model. My analysis covers major income statement items, including interest and non-interest income and expenses, as well as net charge-offs. The results indicate that ML models, particularly tree-based ensemble methods like Random Forest and XGBoost, offer a significant improvement in predictive accuracy over the linear benchmark.

JEL Classification: C45, C53, G21, E47

Keywords: Machine Learning, Bank Profitability, Stress Testing, Financial Stability, Forecasting

*Senior Economist and Data Scientist. Affiliation: Swiss National Bank. Email: tan.schelling@gmail.com

1 Introduction

In the aftermath of the 2008 Global Financial Crisis (GFC), stress testing has emerged as a central toolkit for central banks and supervisory authorities to gauge the resilience of the financial system (Basel Committee on Banking Supervision, 2009). A crucial element of this exercise is the projection of bank profitability conditional on a set of severe but plausible macroeconomic scenarios. Bank profitability stands as an important line of defense against deteriorating economic conditions. Banks that can sustain profitability during a downturn are better positioned to absorb losses from credit defaults or asset price shocks through retained earnings, thereby preserving their capital base (Kashyap, Stein, and Hanson, 2010).

Stress scenarios represent a simulated adverse evolution of key economic indicators, such as Gross Domestic Product (GDP) growth, the unemployment rate, inflation, and asset prices. These variables serve as inputs for models designed to predict the trajectory of bank profitability. Consequently, the quality and informativeness of stress test results are dependent on the predictive power of these underlying models.

However, predicting bank profitability is a notoriously difficult task. First, there are data quality issues in bank accounting data, arising due to changes in accounting rules, differences in accounting methods and, sometimes, outright errors. Second, profitability may depend on discretionary management decisions, such as provisioning, wages or expenses. These are hard, if not impossible, to model with external data. Third, the banking sector is characterized by heterogeneous business models, with income streams that exhibit diverse sensitivities to economic conditions. Fourth, the relationship between macroeconomic drivers and bank performance may be subject to non-linearities and interaction effects, which are challenging to capture with conventional linear regression models. This study tackles this last difficulty by employing ML methods that have the inherent ability to discover such effects.

Non-linearities arise, for example, from the relationship between interest rates and net interest income; while moderate rate increases can widen lending margins, sharp increases may trigger a wave of defaults that erodes interest income, a dynamic particularly pronounced in low-rate environments (e.g., Borio, Gambacorta, and Hofmann, 2017). Similarly, the impact of an individual losing their job on loan losses is often non-linear, with unemployment shocks sharply accelerating the probability of default (see Gerardi, Herkenhoff, Ohanian, and Willen, 2018). Interaction effects can also be significant. For instance, the adverse impact of a GDP contraction on loan portfolios is often magnified when it occurs alongside a spike in corporate bond spreads, as credit market dislocations and weakening economic fundamentals simultaneously constrain the debt servicing capacity of corporate borrowers (Gilchrist and Zakrajšek, 2012).

Hunt, Myers, and Myers (2022) revisit the classic task of predicting the sign of future earnings changes. They find that while traditional logistic regression remains a solid baseline, non-parametric models like Random Forests can significantly improve out-of-sample forecast accuracy, leading to more profitable trading strategies based on these predictions. A comprehensive review of 70 studies from 2013 to 2023 found a clear predominance of ML techniques, with neural networks, logistic regression, and decision trees being the most frequently employed algorithms for predicting firm performance (Gogas, Papadimitriou, and Agathagelidis, 2024). The authors emphasize that a key advantage of ML is the ability to integrate a much broader range of data attributes beyond standard financial metrics, including operational, market, and even non-traditional data sources, to create more nuanced and accurate forecasts.

Recent work by supervisory bodies and academics highlights the potential for ML to enhance the accuracy and granularity of stress tests. Aldasoro, Hördahl, Schrimpf, and Zhu (2025) from the Bank for International Settlements demonstrate that tree-based models, specifically Random Forests, significantly outperform traditional time-series approaches in predicting the full distribution of future financial market stress, especially at longer horizons. Their use of Shapley values to interpret the model reveals that factors related to funding liquidity and investor overextension are key predictors of tail risk. Other research focuses on using ML for early-warning systems and bank failure prediction. Gogas, Papadimitriou, and Gounopoulos (2018) apply support vector machines (SVM) to build a bank failure forecasting model using a sample of U.S. banks during the Great Financial Crisis. They find that ML techniques can achieve high classification accuracy.

Furthermore, ML techniques are being developed to create more realistic and dynamic stress test simulations. Petropoulos, Siakoulis, Panousis, Papadoulas, and Chatzis (2022) propose "Deep-Stress," a framework that uses deep learning algorithms to simulate the entire bank balance sheet dynamically. This approach overcomes the static balance sheet assumption common in many regulatory models. Similarly, recent explorations into generative AI, such as Generative Adversarial Networks (GANs), aim to create synthetic data representing unprecedented but plausible market scenarios, moving beyond the limitations of purely historical data for stress testing (Metha, Lakhamraju, Miriyala, and Macha, 2025).

This study focuses on tree-based algorithms, as they have been shown in past studies to perform very well in financial and economic prediction tasks. They are able to capture such complex relationships without imposing restrictive a priori assumptions on the underlying functional form, and their strong performance has been well-documented in previous literature (e.g., Gu, Kelly, and Xiu, 2020; Mullainathan and Spiess, 2017; Gogas et al., 2024).

I focus on predicting the banks' sub-components of profitability. These are interest income and expenses, non-interest income and expenses as well as charge-offs on their loan portfolios. I systematically compare the out-of-sample performance of various models against a traditional linear regression benchmark. I find that tree-based machine learning algorithms offer a significant improvement over tradition linear models for predicting bank profitability components.

2 Data and Feature Engineering

My analysis relies on a comprehensive panel dataset constructed from the U.S. Federal Deposit Insurance Corporation (FDIC) Consolidated Reports of Condition and Income, commonly known as "call reports." The panel data spans the period from the first quarter of 2001 to the fourth quarter of 2024, 10,849 unique banks and 16,615 reporting items.

The data contains a wide range of balance sheet, off-balance sheet and income statement positions. Due to time and compute resource constraints, only a fraction of the data was used. From the whole sample, I selected 200 of the largest banks in terms of average total assets and focused on a few salient balance sheet positions.

The target variables were scaled by total assets to ensure comparability across institutions of varying sizes and to alleviate non-stationarity. The feature set includes a mix of bank-specific balance sheet characteristics and key macroeconomic variables, reflecting the standard approach in the stress testing literature. Table 1 and 2 describe the target and feature variables used in this study.

Table 1: Target Variables

Variable	**Description**
Interest Income to Total Assets	Income generated by interest-bearing assets like loans and securities.
Non-Interest Income to Total Assets	Income from fees, commissions, and other non-interest activities.
Interest Expense to Total Assets	Cost of borrowing, including interest paid on deposits and other debt.
Non-Interest Expense to Total Assets	Operating expenses, such as salaries, rent, utilities, and administrative costs.
Net Charge-offs to Total Assets	The difference between loans written off as losses and any subsequent recoveries.

Table 2: Feature Variables

Variable	Description
Bank-Specific Variables	
Log of Total Assets	Indicator for bank size, log-transformed to reduce skewness.
Deposits to Assets	Indicator for the funding structure and reliance on deposit base.
Loans to Assets	Indicator for the business model of the bank.
Macroeconomic Variables	
GDP Growth (QoQ)	Quarterly growth rate of real Gross Domestic Product.
CPI Growth (QoQ)	Quarterly growth rate of the Consumer Price Index.
Unemployment Rate	National unemployment rate.
Household Delinquency Rate	Measure of household loan defaults.
3-Month Treasury Bill	Short-term risk-free interest rate.
10-Year Treasury Yield	Long-term risk-free interest rate.
S&P 500 Growth (QoQ)	Quarterly return on the S&P 500 stock index.
Corporate Bond Spreads	Measures perceived credit-worthiness of companies.
VIX Growth (QoQ)	Measure of financial market volatility and fear.

2.1 Data Preprocessing

Preparing the data for modeling involved several preprocessing steps to handle common issues in large financial data panels and to engineer meaningful features.

Missing Values: A panel of bank accounting data spanning more than 20 years is naturally highly unbalanced. Many banks exit the sample due to mergers, acquisitions, or failures, while new banks enter. Of the nearly 11,000 banks in the full sample, only about 25% reported continuously throughout the entire period. This means approximately 75% of banks have at least one missing observation. Given the scale of missing values, standard imputation techniques would introduce significant bias and risk creating artificial patterns that ML algorithms could exploit, leading to inflated performance metrics. Instead, I adopted a more conservative approach. I removed banks with fewer than 12 consecutive quarters of data, as their short time series provide limited information for dynamic modeling. This procedure led to the removal of a subset of banks, but the panel was intentionally left unbalanced to reflect the true nature of the banking population. For banks where there was a reporting gap of 1 quarter in their data, I imputed the data by forward filling.

Feature Engineering: All bank-level balance sheet variables were scaled by total assets to control for heterogeneity in bank size and to alleviate non-stationarity concerns.

Total assets itself was log-transformed. For GDP and CPI, I used quarter-on-quarter growth rates. In the case of percentage indicators like the unemployment rate, interest rates, delinquency rates and default rates, I calculated the first difference quarter-on-quarter. To capture regular seasonal patterns, I included a set of dummy variables for each calendar quarter. To model time dynamics, I included four quarterly lags for all features.

Treatment of Structural Breaks: Bank balance sheets can experience sudden, sharp shifts due to M&A activity, major divestitures, or accounting changes. To prevent these breaks from introducing noise and estimation bias, I implemented a correction mechanism. For each bank, if its total assets changed by more than three standard deviations (calculated from historical time series of changes across all banks), I identified a structural break. I then calculated a "break factor" to scale all pre-break balance sheet data, harmonizing the series while preserving its growth dynamics. Additionally, I created a binary dummy variable that takes the value of one at the quarter of the break, allowing the model to explicitly account for the event.

Winsorization: To prevent extreme, one-off events from unduly influencing model training, I winsorized outliers in the target variables at the 2% level. Observations were capped at 2% and 98% of the distribution in the training set.

Restricting number of banks Due to time and compute resource constraints, I restricted the sample to the 200 largest banks by average total assets. This focus also mitigates potential data quality issues, as smaller institutions may have less sophisticated reporting and their data is often subject to less intensive regulatory inspection.

Train-Test Split: I used 80% of the data for training. To respect the timeseries structure, I split along time. The training set starts in 2001 and ends in 2019. The test set starts in 2020 and ends at the end of 2024.

The following tables show descriptive statistics of the resulting data set (train and test taken together).

Table 3: Descriptive Statistics for Bank-Specific Variables

	N	Mean	Std	Min	25%	50%	75%	Max
Interest Income / Assets	11812	0.01	0.00	-0.00	0.01	0.01	0.01	0.03
Interest Expense / Assets	11812	0.00	0.00	-0.00	0.00	0.00	0.00	0.01
Non-Int. Income / Assets	11812	0.00	0.01	-0.08	0.00	0.00	0.00	0.06
Non-Int. Expense / Assets	11812	0.01	0.00	-0.00	0.00	0.01	0.01	0.03
Net Charge-Offs / Loans	11729	0.00	0.00	-0.01	0.00	0.00	0.00	0.13
Deposit Ratio	11812	0.71	0.17	0.00	0.65	0.75	0.82	0.98
Loan-to-Asset Ratio	11812	0.61	0.18	0.00	0.54	0.66	0.73	1.02
Log(Total Assets)	11876	17.15	1.38	13.11	16.13	16.94	17.96	22.00
Structural Break Dummy	11876	0.01	0.11	0.00	0.00	0.00	0.00	1.00

Table 4: Descriptive Statistics for Macroeconomic Variables

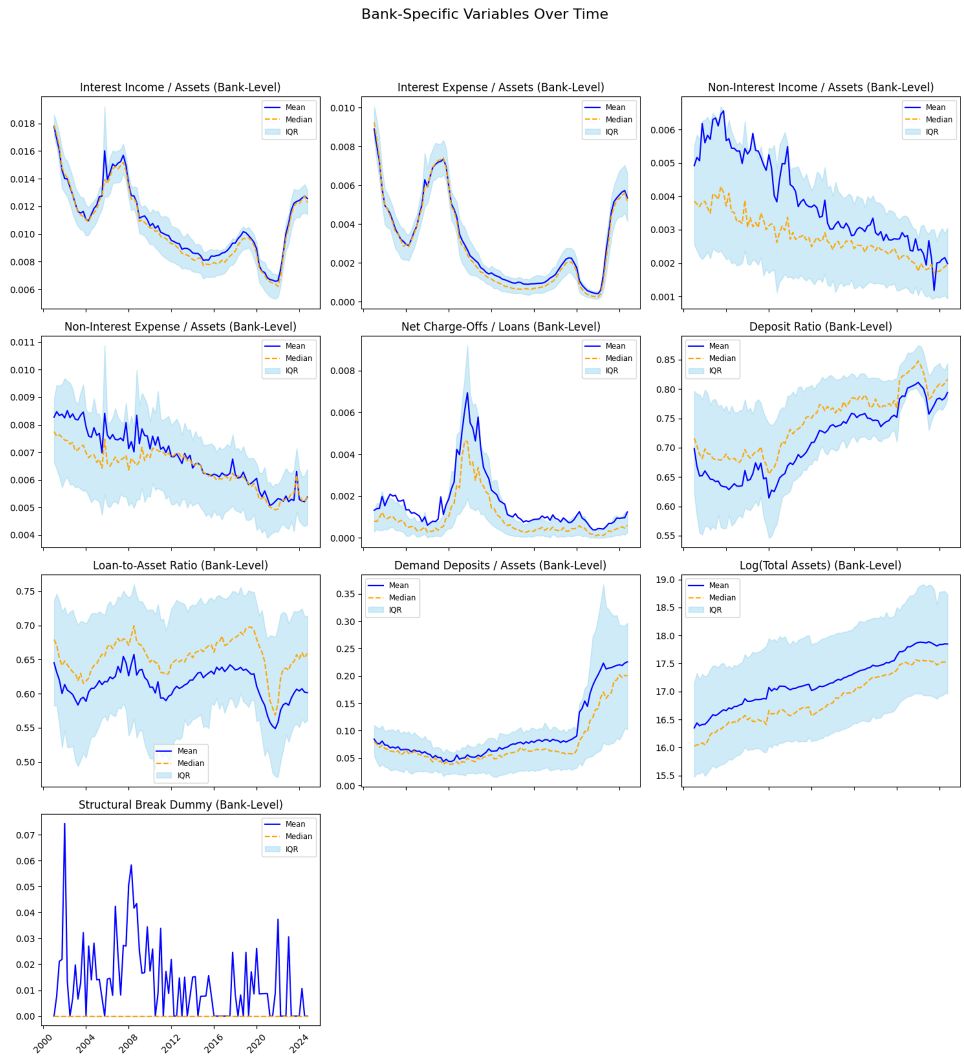
	N	Mean	Std	Min	25%	50%	75%	Max
Corp. Bond Spread (Diff)	96	-0.00	0.00	-0.02	-0.00	-0.00	0.00	0.02
CPI (QoQ)	96	0.01	0.01	-0.02	0.00	0.01	0.01	0.02
GDP (QoQ)	96	0.01	0.01	-0.08	0.01	0.01	0.01	0.09
Household Delinq. Rate (Diff)	96	-0.00	0.00	-0.01	-0.00	-0.00	0.00	0.01
S&P 500 (QoQ)	96	0.02	0.09	-0.24	-0.02	0.03	0.07	0.26
10Y T-Bill Rate (Diff)	96	-0.00	0.00	-0.01	-0.00	-0.00	0.00	0.01
3M T-Bill Rate (Diff)	96	-0.00	0.00	-0.01	-0.00	-0.00	0.00	0.02
Unemployment Rate (Diff)	96	-0.00	0.01	-0.04	-0.00	-0.00	0.00	0.09
VIX (QoQ)	96	0.02	0.27	-0.39	-0.12	-0.04	0.06	1.34

To get a sense of the evolution of the variables over time Figure 1 shows the mean (blue line), the median (orange dotted line) and the interquartile range (shaded area) of the bank-level variables. Figure 2 shows the evolution of the macro variables.

3 Methodology

I predict each of the five target variables separately. For each target, I estimate a baseline linear model via OLS and several more complex machine learning models.

Figure 1: Bank-level variables over time



3.1 Models

Ridge Regression A linear model that uses L2 regularization to address multicollinearity. It adds a penalty proportional to the square of the magnitude of coefficients, which shrinks them towards zero without eliminating them, thus stabilizing the model when features are correlated.

Decision Trees A non-linear model that creates a flowchart-like structure by recursively splitting the data based on feature values that best separate the target variable. While highly interpretable, a single decision tree is prone to overfitting and instability.

Random Forest An ensemble method that addresses the shortcomings of a single decision tree (Breiman, 2001). It constructs a large number of individual decision trees, with each tree trained on a different random bootstrap sample of the data. The final prediction is the average of the predictions from all trees in the forest. This

Figure 2: Macro variables over time

process of averaging across many decorrelated trees dramatically reduces variance and prevents overfitting.

XGBoost (Extreme Gradient Boosting) Another, more advanced ensemble method based on the principle of boosting (Chen, He, Benesty, Khotilovich, Tang, Cho, Chen, Mitchell, Cano, Zhou et al., 2015). Unlike Random Forest, which builds trees in parallel, XGBoost builds them sequentially in an adaptive manner to reduce bias. Each subsequent tree is trained specifically to correct the errors made by its predecessors. This methodical, error-focused learning process is why gradient boosting models often achieve state-of-the-art performance.

Voting Ensemble A meta-estimator that fits several base regressors and averages their individual predictions to form a final prediction. The rationale is that by combining conceptually different, well-performing models, the ensemble can achieve better performance and robustness, as the errors from one model may be offset by the strengths of another.

3.2 Hyperparameter Tuning and Feature Selection

To optimize model performance, hyperparameter tuning is applied. A randomized search cross-validation is used. To respect the time-series nature of the data, I employ a

`TimeSeriesSplit` cross-validator, ensuring that training data always precedes test data to prevent look-ahead bias. The primary scoring metric is the negative mean squared error.

For the Ridge regression, the key hyperparameter 'alpha' is tuned to find the optimal level of penalty on the model coefficients. For the Random Forest and XGBoost models, the tuning focuses on parameters that control model complexity and prevent overfitting, such as `n_estimators`, `max_depth`, and `learning_rate`.

Due to the high potential for multicollinearity, I also explored Recursive Feature Elimination (RFE). RFE iteratively fits a model, ranks features, and removes the weakest until an optimal subset is identified. For stability in the ranking step, I used a simple Linear Regression as the base estimator within the RFE process.

4 Results

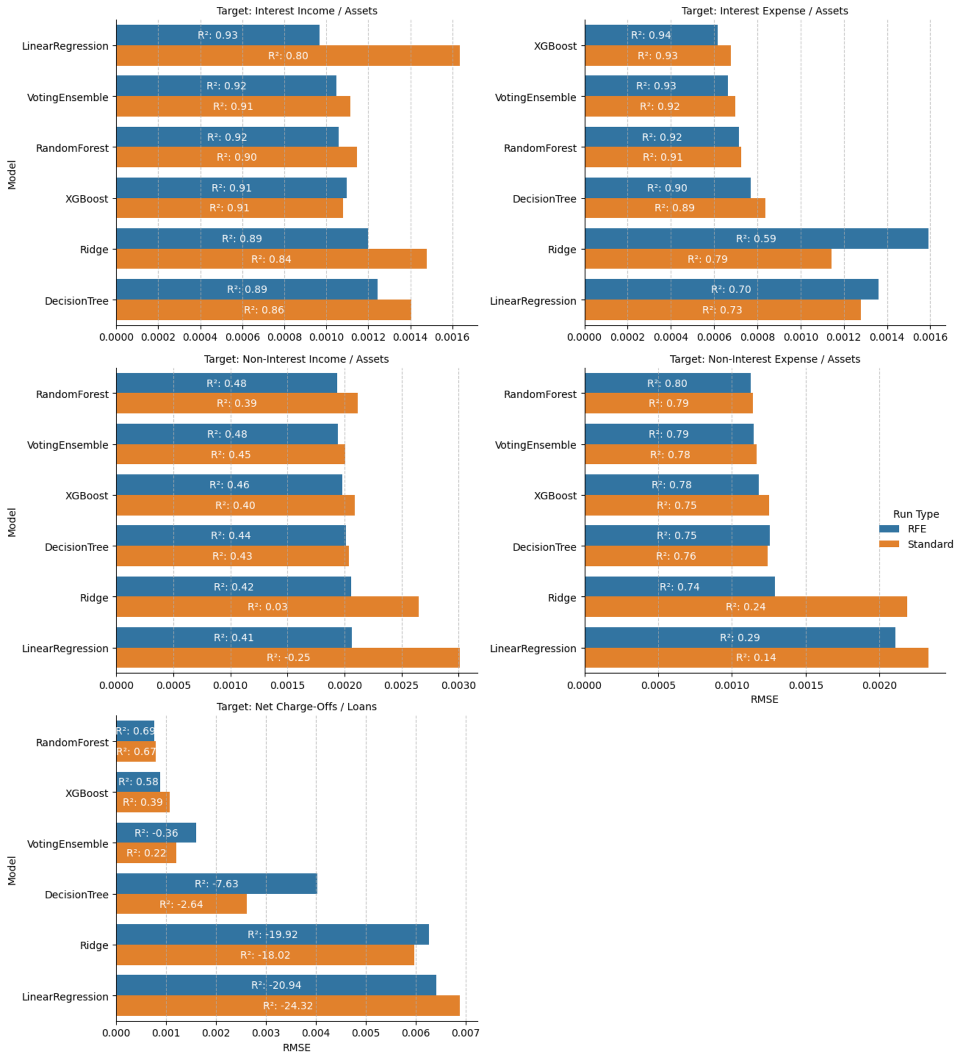
This section analyzes the out-of-sample predictive performance. The primary evaluation metric is the Root Mean Squared Error (RMSE), as its quadratic loss function heavily penalizes large prediction errors, which are of particular concern in stress testing. Figure 3 shows the predictive performance of the models for the various profitability items, ordered by RMSE.

The results reveal several key insights. First, there is a clear performance hierarchy, with the ensemble tree-based models—Random Forest and XGBoost—consistently outperforming the simpler linear and single-tree models across nearly all target variables. The one exception where Linear Regression performs best is for Interest Income / Assets, after applying RFE. Random Forest performs slightly better 3 out of 4 cases. However, XGBoost is much more efficient. Computational efficiency becomes more important when scaling the analysis to more banks and features.

Second, RFE leads to the largest performance gains for OLS and Ridge. As the number of multicollinear features increase, OLS and Ridge predictions become more unstable. Due to L2 regularization, Ridge can better deal with multicollinearity. The performance gains for tree-based ensemble models are small, even negative in some cases. This shows that these models are already well equipped to deal with a large amount of features.

Third, from a target variable perspective, net charge-offs and non-interest income prove to be the most challenging components to predict. In the case of net charge-offs, R2 is even negative for models other than Random Forest and XGBoost. Net charge-offs might hard to predict as there is an element of management discretion. Management may write off a loan or evergreen it to avoid recognizing a loss. Non-interest income comprises trading losses and gains, which can have a large idiosynchratic component due to market positioning. Also, fee and commission income is dependent on market movements and

Figure 3: Overall predictive performance



client behavior. Nevertheless, Random Forest and XGBoost still exhibit a reasonable performance.

Fourth, feature importance charts indicate that the models heavily rely on autoregressive terms. This is consistent with observations stated in Vukovic, Spitsina, Gribanova, Spitsin, and Lyzin (2023), who observe that previous profitability has a strong impact on performance.

5 Conclusion

This paper demonstrates that tree-based machine learning algorithms offer a significant improvement over traditional linear models for predicting bank income components. My findings show that tree-based ensemble methods like Random Forest and XGBoost consistently outperform Linear Regression. Banking income component data seems to follow complex data generation processes, with non-linearities and interactions.

I find that in the presence of a high number of features, recursive feature elimination may help improve predictive performance, especially for linear models. Among the top performing models, XGBoost presents a compelling case for practical application, offering predictive accuracy nearly on par with Random Forest but with significantly greater computational efficiency. This makes it a promising tool for banking authorities and risk managers who need to run complex models on large datasets.

However, my analysis also highlights some important areas of improvement. First, due to time and computational resource constraints, the data was restricted to the largest 200 banks. Second, as accounting data is inherently messy with a low signal-to-noise ratio, more extensive data cleaning avenues could be explored to improve prediction performance. Third, the addition of market-based bank data, like stock and Credit Default Swap data, could be a valuable addition. Finally, exploring other complex models, like neural networks, and more advanced dimensionality reduction techniques could yield further improvements and provide a deeper understanding of the drivers of bank profitability.

References

Aldasoro, I., P. Hördahl, A. Schrimpf, and S. Zhu (2025): "Predicting financial market stress with machine learning," BIS Working Papers 1250, Bank for International Settlements.

Basel Committee on Banking Supervision (2009): "Principles for sound stress testing practices and supervision," Tech. rep., Bank for International Settlements, Basel.

Borio, C., L. Gambacorta, and B. Hofmann (2017): "The influence of monetary policy on bank profitability," *International Finance*, 20, 48–63.

Breiman, L. (2001): "Random forests," *Machine learning*, 45, 5–32.

Chen, T., T. He, M. Benesty, V. Khotilovich, Y. Tang, H. Cho, K. Chen, R. Mitchell, I. Cano, T. Zhou, et al. (2015): "Xgboost: extreme gradient boosting," *R package version 0.4-2*, 1, 1–4.

Gerardi, K., K. F. Herkenhoff, L. E. Ohanian, and P. S. Willen (2018): "Can't Pay or Won't Pay? Unemployment, Negative Equity, and Strategic Default," *The Review of Financial Studies*, 31, 1096–1131.

Gilchrist, S. and E. Zakrajšek (2012): "Credit Spreads and Business Cycle Fluctuations," *American Economic Review*, 102, 1692–1720.

Gogas, P., T. Papadimitriou, and P. Agathagelidis (2024): "Machine learning in predicting firm performance: a systematic review," *Corporate Finance and Accounting Review.*

Gogas, P., T. Papadimitriou, and D. Gounopoulos (2018): "Forecasting bank failures and stress testing: A machine learning approach," *International Journal of Forecasting*, 34, 443–455.

Gu, S., B. Kelly, and D. Xiu (2020): "Empirical Asset Pricing via Machine Learning," *The Review of Financial Studies*, 33, 2223–2273.

Hunt, J. O., J. N. Myers, and L. A. Myers (2022): "Improving earnings predictions and abnormal returns with machine learning," *Accounting Horizons*, 36, 131–149.

Kashyap, A. K., J. C. Stein, and S. Hanson (2010): "An analysis of the impact of "subprime" mortgage lending on the financial system," Tech. Rep. w16091, National Bureau of Economic Research.

Metha, S., M. V. Lakhamraju, N. S. Miriyala, and K. Macha (2025): "Stress Testing Financial Systems– Simulating economic disruption using AI-driven risk models," *International Journal of Computational and Experimental Science and Engineering*, 11.

Mullainathan, S. and J. Spiess (2017): "Machine Learning: An Applied Econometric Approach," *Journal of Economic Perspectives*, 31, 87–106.

Petropoulos, A., V. Siakoulis, K. P. Panousis, L. Papadoulas, and S. Chatzis (2022): "A Deep Learning Approach for Dynamic Balance Sheet Stress Testing," in *Proceedings of the Third ACM International Conference on AI in Finance*, 53–61.

Vukovic, D. B., L. Spitsina, E. Gribanova, V. Spitsin, and I. Lyzin (2023): "Predicting the performance of retail market firms: regression and machine learning methods," *Mathematics*, 11, 1916.

A Number of features after RFE

Table 5: Number of Features Used by Model (Horizon 1)

BaseModel	Target / RunType	Interest Expense / Assets	Interest Income / Assets	Net Charge-Offs / Loans	Non-Interest Expense / Assets	Non-Interest Income / Assets
DecisionTree	RFE	28	17	55	39	37
	Standard	181	181	181	181	181
LinearRegression	RFE	56	36	53	57	15
	Standard	181	181	181	181	181
RandomForest	RFE	24	32	58	40	33
	Standard	181	181	181	181	181
Ridge	RFE	51	51	51	22	16
	Standard	181	181	181	181	181
VotingEnsemble	RFE	181	181	181	181	181
	Standard	181	181	181	181	181
XGBoost	RFE	28	42	28	28	28
	Standard	181	181	181	181	181

B Visual Prediction Comparison

Figure 4: Interest income to assets

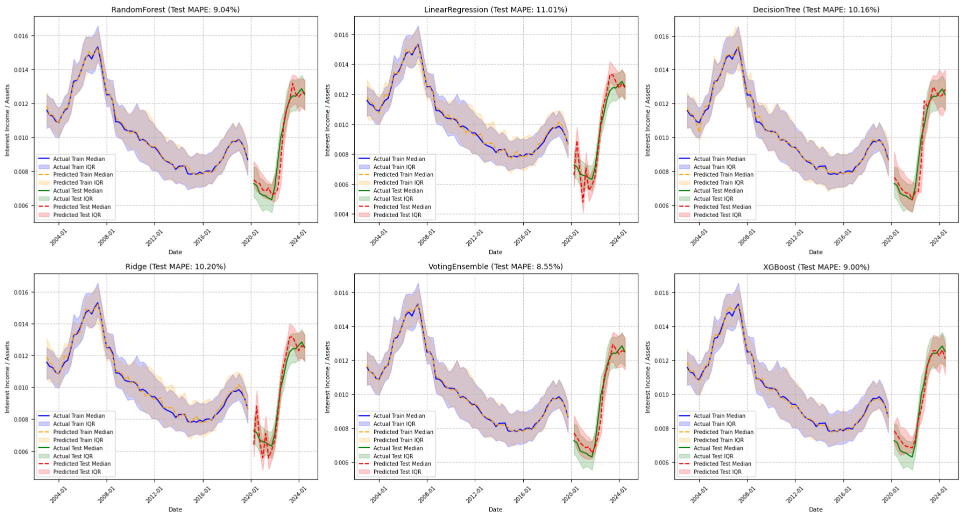


Figure 5: Interest expense to assets

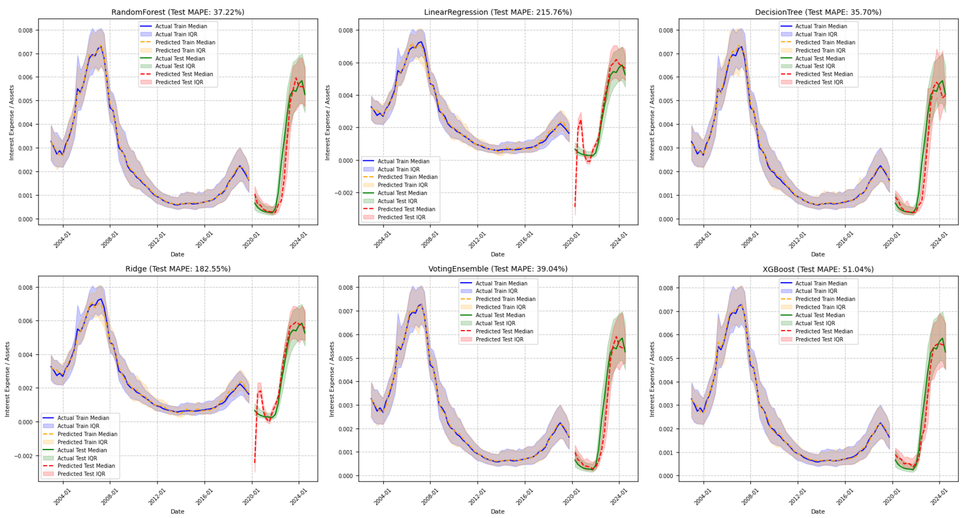


Figure 6: Non-interest income to assets

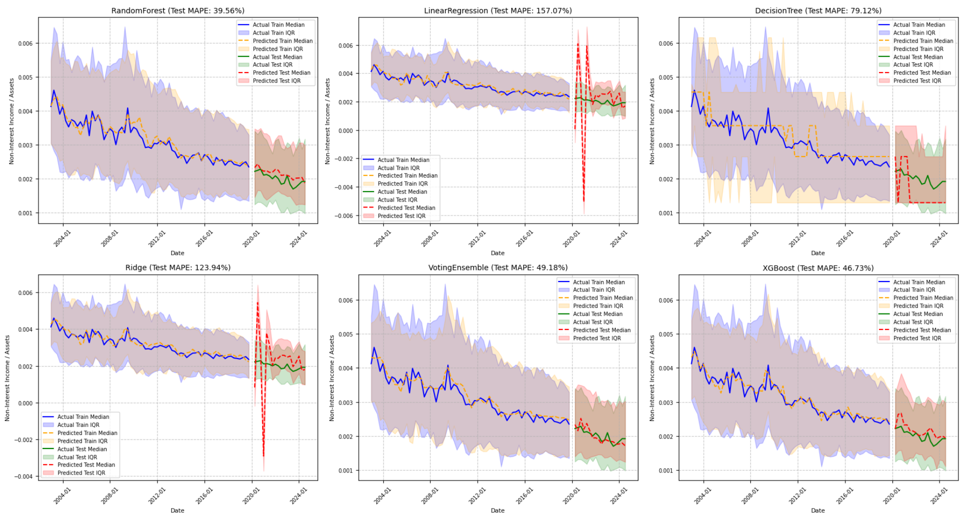


Figure 7: Non-interest expense to assets

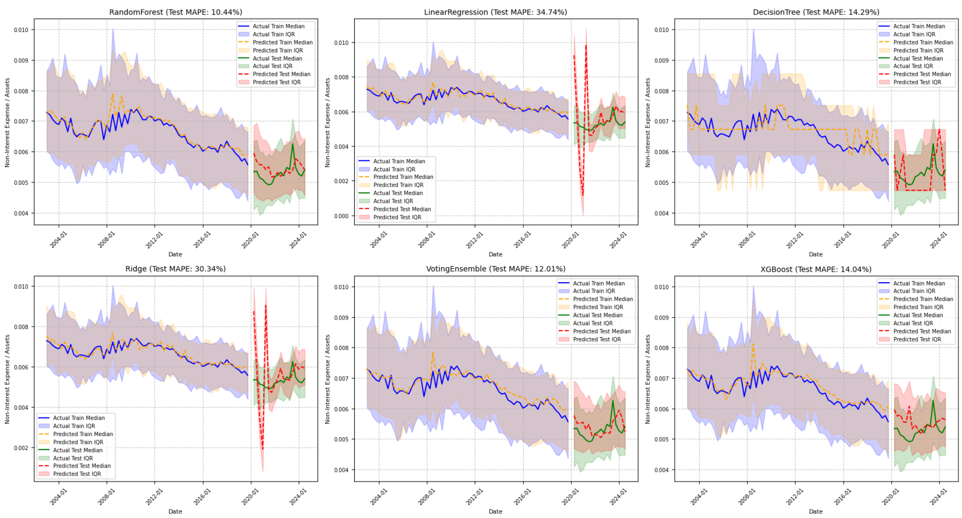
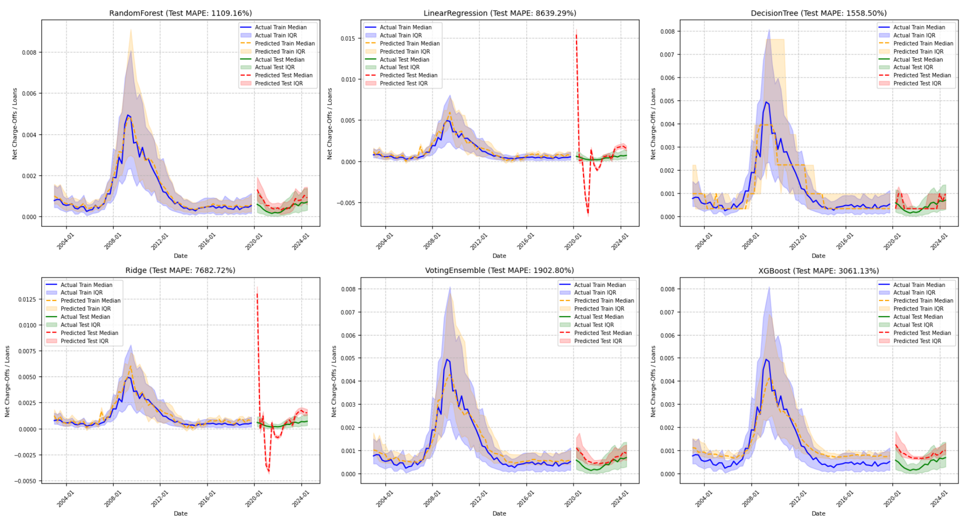


Figure 8: Net charge-offs on loans



C Feature Importances

Figure 9: Interest income to assets

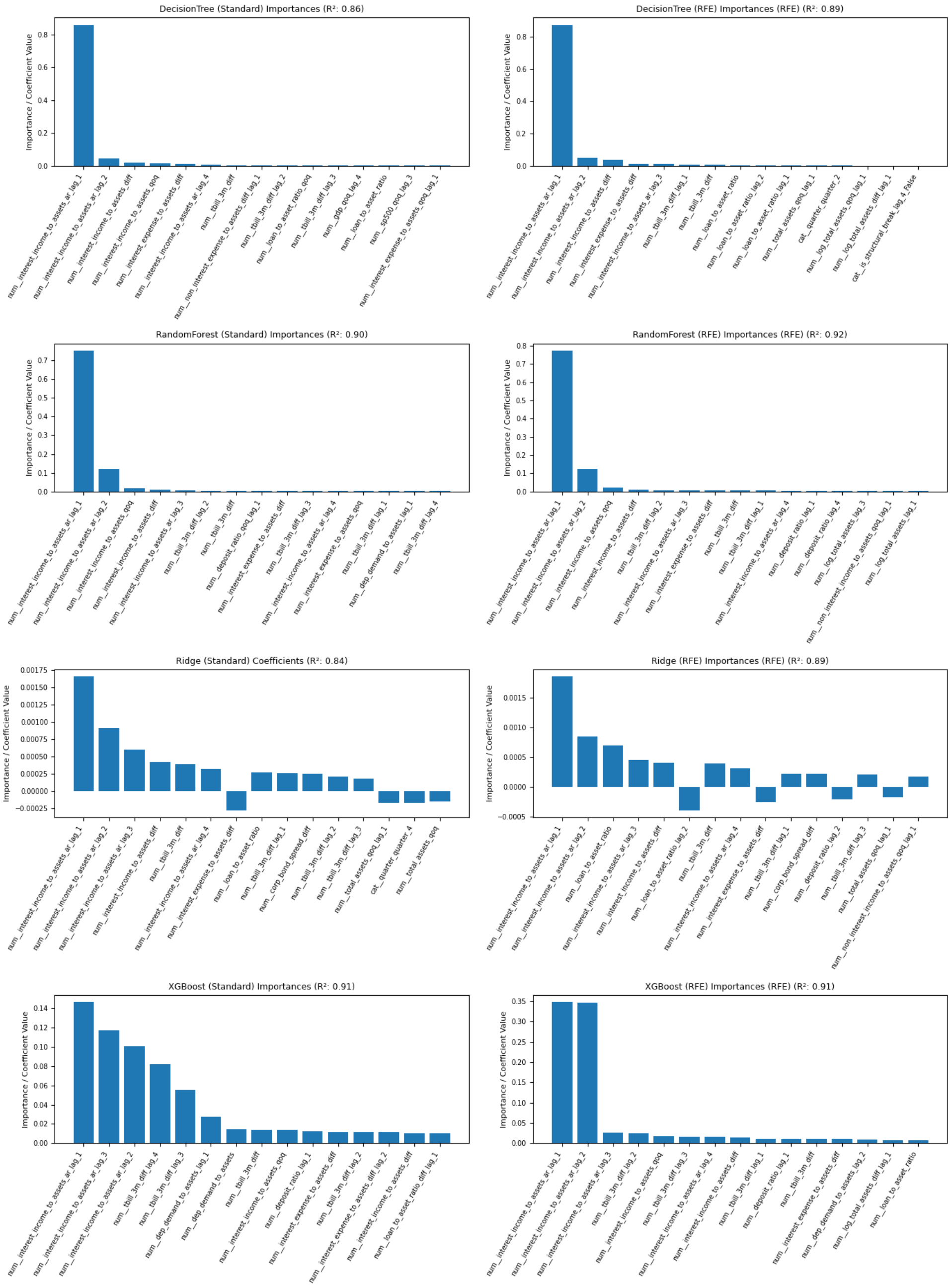


Figure 10: Interest expense to assets

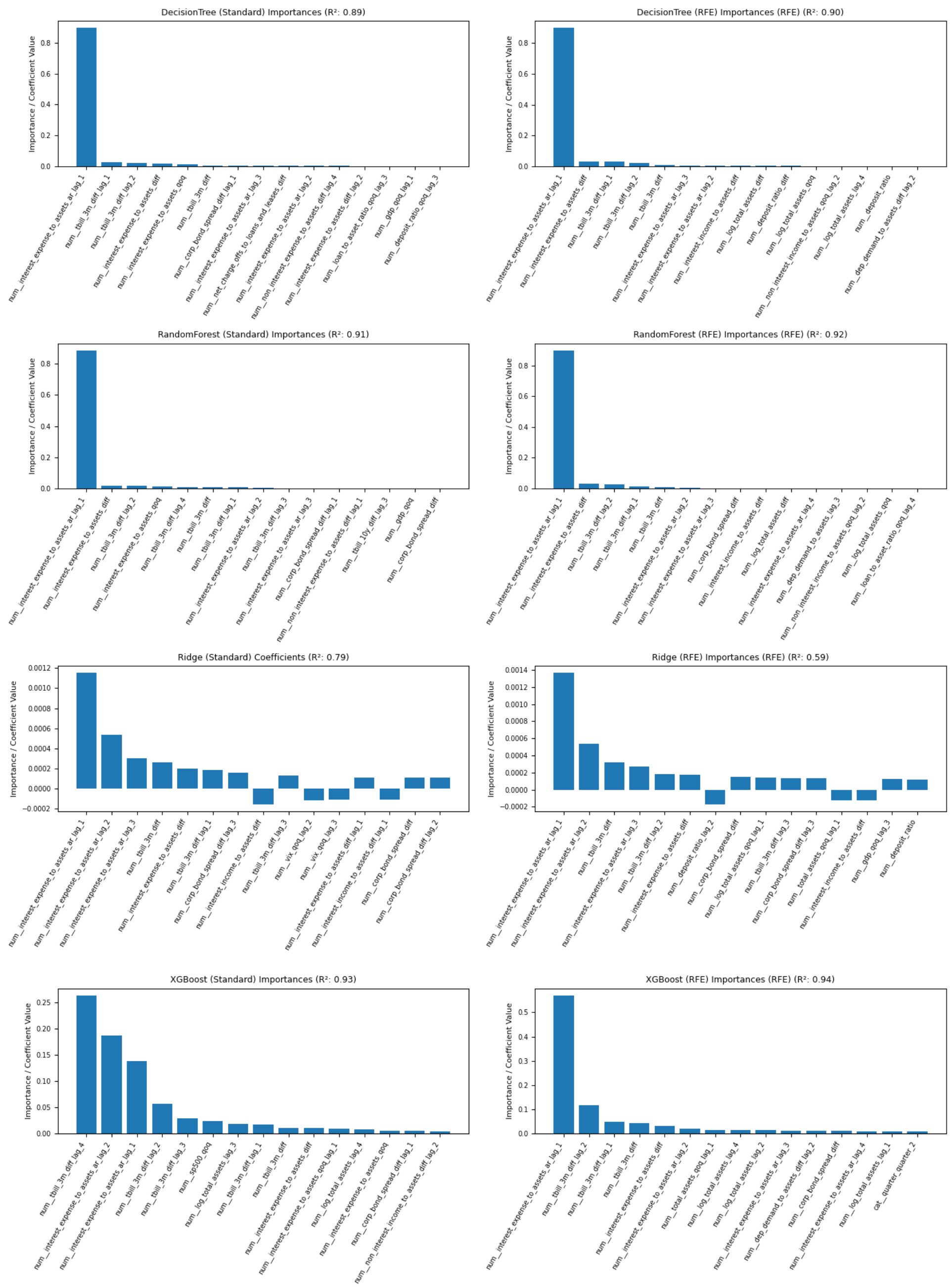


Figure 11: Non-interest income to assets

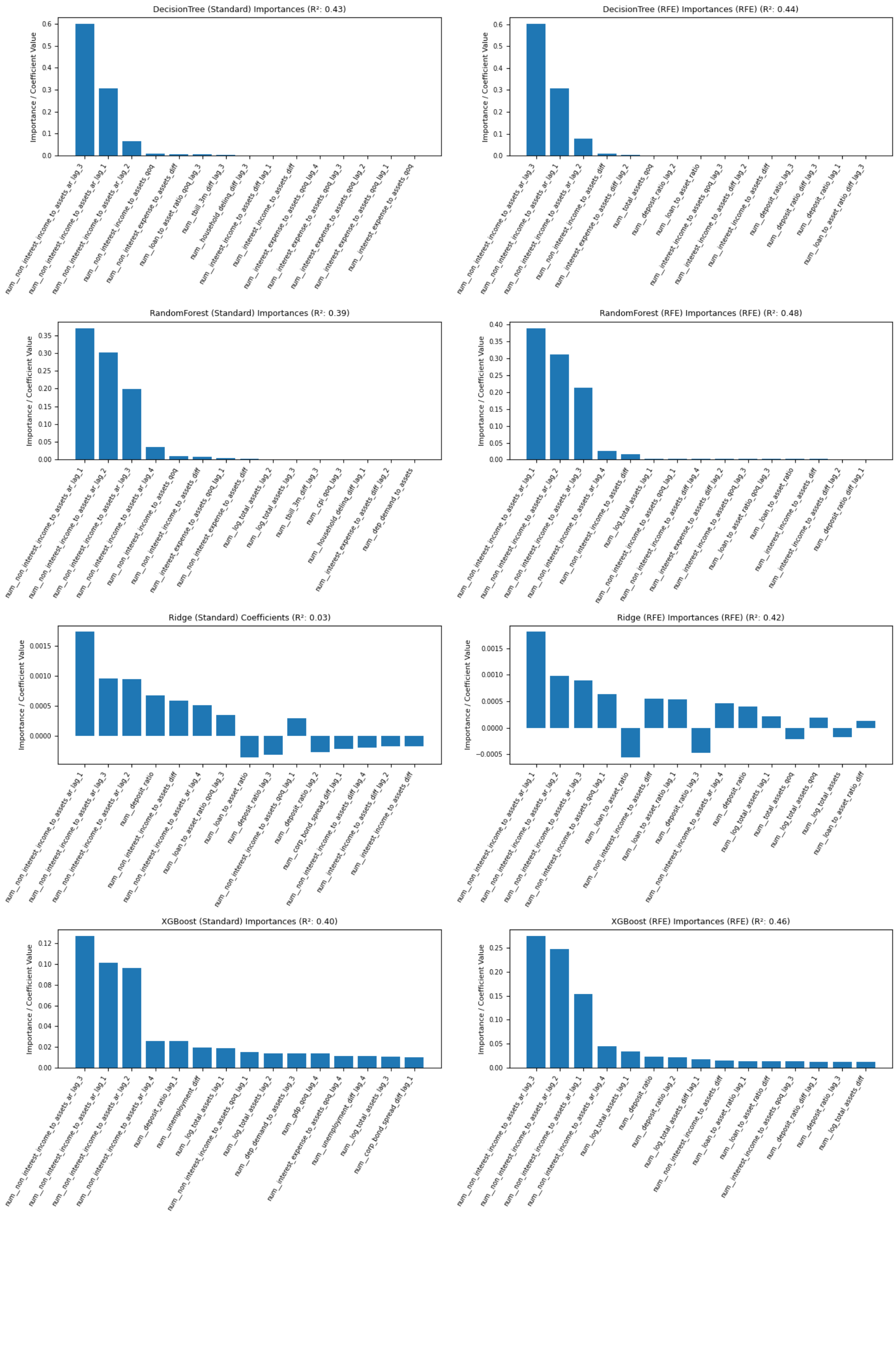


Figure 12: Non-interest expense to assets

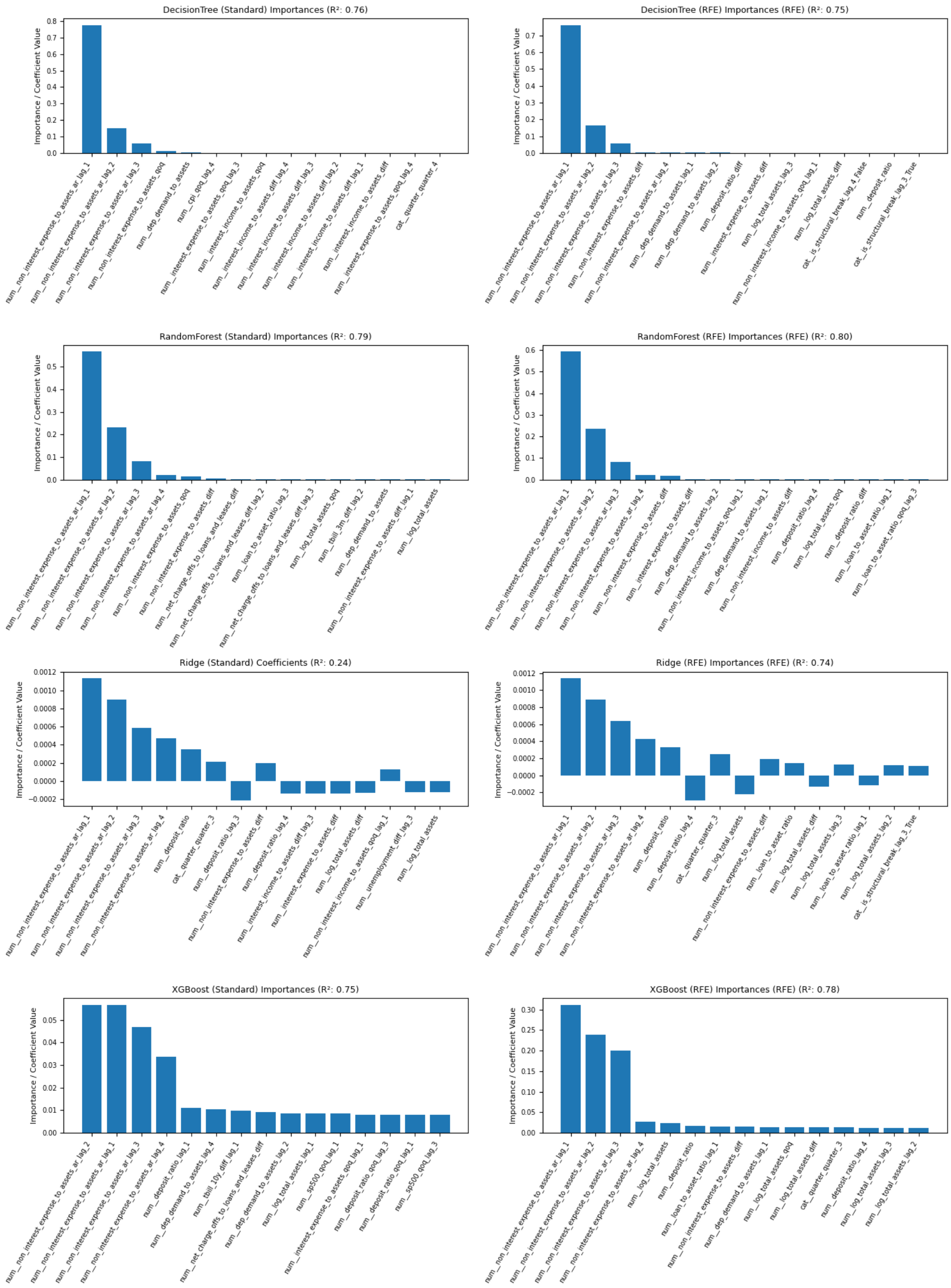


Figure 13: Net charge-offs on loans