Forecasting U.S. CD Deposits Under Monetary Policy Shocks Using Classical, Machine Learning, and Deep Learning Models

Zulfa L Mohamed

2025-12-01

Table of Contents

1 Abstract

This study evaluates how well classical time series models, machine learning algorithms, and deep learning methods forecast United States small denomination time deposits during periods of monetary policy change. The data consist of monthly CD deposits from 1995 to 2024 combined with the federal funds rate. The analysis applies STL decomposition, Box Cox transformation, stationarity testing, and a set of lagged, rolling, and seasonal features. The series shows large structural shifts, strong trends, and heavy volatility clustering, which makes forecasting difficult and creates conditions where nonlinear methods may be more effective than traditional ones.

Six groups of models were estimated. These include ETS, ARIMA, ARIMAX, XGBoost, a multilayer perceptron, Ensemble, and a GARCH model for time varying volatility. The results show a clear performance difference. XGBoost achieves an RMSE of 15.74 while all classical models produce RMSE values above 900. Neural networks learn some nonlinear structure but are less consistent than boosting. ARIMAX confirms that policy rates help explain CD movements and GARCH reveals persistent volatility. Most traditional models have a MASE above one which means they do not outperform a naive benchmark. This highlights how hard it is to forecast deposit behavior with linear methods because CD deposits respond to many economic forces and structural breaks such as recessions and the COVID period. These events create lagged effects that classical models struggle to capture.

Overall, the findings show that accurate CD deposit forecasting requires methods that can learn nonlinear relationships involving policy rates, seasonality, and historical patterns. Gradient boosted trees deliver the strongest accuracy and provide a useful tool for banks that rely on rate sensitive deposit funding, particularly mid sized and online institutions that face faster deposit movements when monetary conditions change.

2 Introduction

Liquidity describes how easily and quickly an asset can be converted to cash without losing much value. Because deposits are usually the cheapest and most stable source of funding for banks, understanding and forecasting their behavior is central to maintaining a resilient balance sheet. Certificates of deposit are one form of this funding. They lock in funds for a fixed term in exchange for a predetermined return. Small denomination time deposits are balances under 100,000 dollars with maturities longer than seven days and are included in the M2 money aggregate (Hyndman and Athanasopoulos (2018)). In the past these deposits were viewed as stable because early withdrawal penalties discouraged rapid movement of funds.

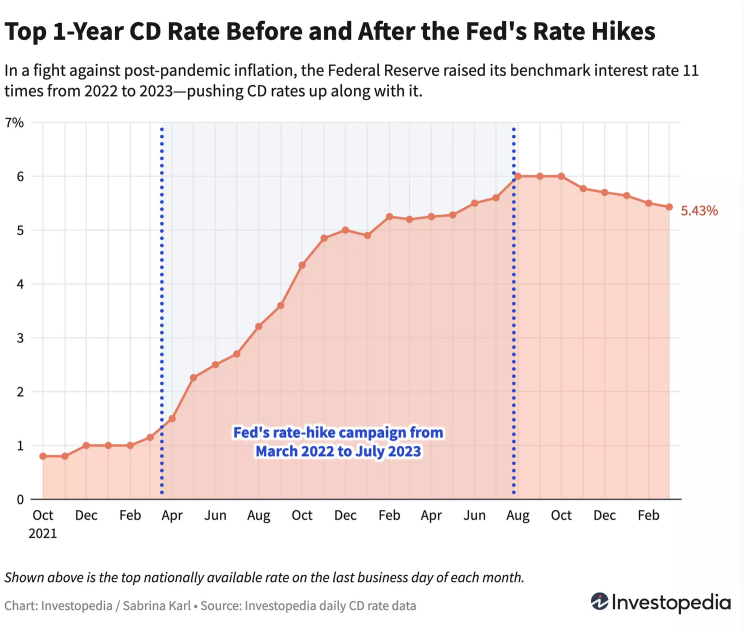


Figure 1: CD rates and monetary policy in the current tightening cycle (Source: Investopedia and author illustration).

The slight drop in the top one year CD rate shows that deposit yields respond quickly to changing rate expectations and that banks adjust CD pricing as soon as they anticipate a shift in monetary policy (Investopedia Staff (2025)).

The environment has changed in recent years. During periods of quantitative easing banks accumulated large reserve balances and deposit inflows surged. When policy rates later increased many of these deposits flowed back out. Evidence from more than fifteen hundred banks shows that deposit flightiness, meaning how quickly depositors move funds, rose after the financial crisis, eased in the mid 2010s, and then climbed sharply between 2020 and 2022. Reserve expansion during the QE period attracted more rate sensitive balances that left quickly once the policy stance tightened (Institute (2025); Im, Li, and Wang (2025)).

Differences across depositors add another layer of complexity. Research by Blickle Li Lu and Ma shows that corporate deposits expanded more than retail deposits during the pandemic and then contracted more sharply when rates increased (Blickle et al. (2025)). Customers of large national banks often tolerate lower rates because they value digital services and convenience but clients of regional banks online banks and credit unions tend to be more rate sensitive. Since early 2022 total deposits at United States banks have fallen by roughly one trillion dollars while money market fund assets have risen by a similar amount. The same source reports that when deposit spreads widen by one percentage point deposits decline by about 0.41 percent at small banks compared with 0.26 percent at large banks (Blickle et al. (2025)). These patterns show that deposit movements are sensitive to relative yields and that smaller institutions face greater exposure. Because of this interplay between flightiness monetary policy and depositor characteristics forecasting CD balances has become important for risk management and pricing decisions.

This study asks whether adding policy variables and nonlinear patterns through machine learning methods improves predictions of small time deposits relative to classical time series models. The analysis is geared toward institutions such as TD Bank, Discover, Capital One, and credit unions where accurate deposit forecasts guide liquidity planning and interest rate decisions. Monthly observations of small time deposits together with the federal funds rate and its lag form the basis of several forecasting models. After addressing non stationarity using Box Cox transformations and differencing then we construct lagged features rolling averages and seasonal harmonics. The classical models include ARIMA exponential smoothing and an ARIMAX with the federal funds rate as an exogenous regressor. Machine learning models include gradient boosting and multilayer perceptron neural networks. A GARCH model is applied to the differenced log returns to detect volatility clustering. The results show that tree based boosting produces the strongest predictive performance while classical methods such as ARIMA offer clear interpretability. These findings suggest that banks seeking to anticipate deposit flows should use both policy information and nonlinear machine learning methods particularly during periods of policy uncertainty and rapid deposit movements.

---

3 Literature Review

3.1 Classical Approaches & Forecast Competition

Forecasting in economics has long relied on autoregressive integrated moving average models and exponential smoothing. The Box Jenkins methodology formalised the practice of differencing non stationary series, identifying autoregressive and moving average orders, and checking residuals for whiteness (Box and Jenkins (1970)). Exponential smoothing methods developed by Brown Holt

and Winters capture level trend and seasonality through weighted averages and remain central to the modern forecasting curriculum (Hyndman and Athanasopoulos (2018)). To address changing variance Engle introduced the autoregressive conditional heteroskedasticity model and its GARCH generalisation which allows volatility to cluster as is common in financial time series (Engle (1982)). Our GARCH diagnostic confirms that monthly growth rates of small time deposits show persistent heteroskedasticity.

Forecast competitions provide evidence on which methods perform well in practice. The M competitions organised by Makridakis evaluated thousands of time series using a wide range of forecasting approaches and found that more complex models do not always outperform simpler ones and that combining forecasts often improves accuracy (Clemen (1989); Makridakis, Spiliotis, and Assimakopoulos (2020)). The 2018 M4 competition reinforced these findings. Pure machine learning models underperformed classical exponential smoothing and ARIMA and the winning method was a hybrid that combined exponential smoothing with a recurrent neural network (Makridakis, Spiliotis, and Assimakopoulos (2020)). From the literature, it seems a lot of the winning method depends on the nature of the data and industry. The traditional methods are easier to interpret mathematically and show how we got to our conclusions. This is why these methods are still the go to in fields like health economics and even the central bank when it is producing papers. The study builds on this by comparing ARIMA and exponential smoothing with gradient boosting and neural networks and by examining whether an ARIMAX with the federal funds rate offers additional explanatory power.

3.2 Machine Learning for Deposit Prediction

Machine learning methods have become popular in forecasting because they can capture nonlinear relationships and interactions. Gradient boosting, represented by XGBoost, builds an ensemble of decision trees that sequentially correct earlier errors and often reaches strong predictive accuracy although interpretability can be limited (Chen and Guestrin (2016)). Neural networks including feed forward and recurrent architectures can learn complex temporal patterns (Ahmed et al. (2010)). Empirical work applying these techniques to deposit behaviour is growing and even in general, the use of machine learning is growing a lot in the financial industries. A recent study evaluates nine machine learning models for predicting customer subscriptions to bank term deposits and finds that CatBoost achieves an accuracy of 90.9 percent and an AUC of 93.8 percent (Kothandapani (2020)). The authors note that the model functions as a black box and use Shapley Additive Explanations to interpret variable influence. These black boxes are still the biggest bottlenecks facing machine learning methods in economics. Other research incorporates textual sentiment and macroeconomic variables to improve short term predictions of deposit flows (Katsafados and Anastasiou (2022)). However, when adding macroeconomic variables we have to be careful about multicollinearity so results are not too biased.

The study contributes to this literature by applying gradient boosting and neural networks to the time series of small time deposit balances and by comparing their performance to classical models. We incorporate exogenous policy variables and construct detailed feature sets including lagged values rolling means and seasonal harmonics. The results show that machine learning methods can outperform traditional models when the deposit series responds nonlinearly to policy rates while classical models remain important for interpretation.

3.3 Rate Sensitivity Across Bank Sizes

Understanding deposit behavior requires considering the macroeconomic environment and differences across depositors. Research by both Federal Reserve economists and external analysts shows that deposit flightiness is not constant. After the financial crisis it increased then moderated in the mid 2010s and surged again during the 2020 to 2022 period. The expansion of reserves during quantitative easing attracted more rate sensitive deposits that later exited rapidly as interest rates rose (Im, Li, and Wang (2025)). Corporate deposits expanded significantly during the pandemic but proved more prone to withdrawal than retail deposits as conditions tightened. Deposit substitution into money market funds has been particularly strong since early 2022. Assets shifted from the banking system toward money market mutual funds by roughly one trillion dollars.

Differences across banks surely create additional forecasting challenges. Large national banks like JP Morgan Chase with broad digital services have customers who are less sensitive to interest rate differences. Smaller banks and online banks must offer higher yields to retain deposits as this is a big source of funding for them. The same research finds that when the spread between deposit rates and the federal funds rate widens by one percentage point deposits decline by 0.41 percent at small banks compared with 0.26 percent at large banks (Blickle et al. (2025)). These findings support including policy variables in forecasting models and indicate that predictive accuracy can differ by institution type. Our work contributes by focusing on small time deposits and examining how different modelling techniques respond to interest rate shocks.

3.4 Research Gaps and Contribution

Although many studies analyse deposit betas and sensitivity to monetary policy, few forecast small time deposit volumes using both classical and machine learning approaches. Volatility analyses rarely include exogenous policy variables and the implications of deposit flightiness for forecasting remain insufficiently explored. This paper addresses these gaps by modelling small time deposit volumes under monetary policy shocks, comparing classical time series models with machine learning methods, and highlighting the role of depositor heterogeneity. The results show that banks can improve forecasting accuracy by combining policy information with nonlinear machine learning techniques and by recognising that deposit behaviour varies across institutions.

4 Methodology

4.1 Data source

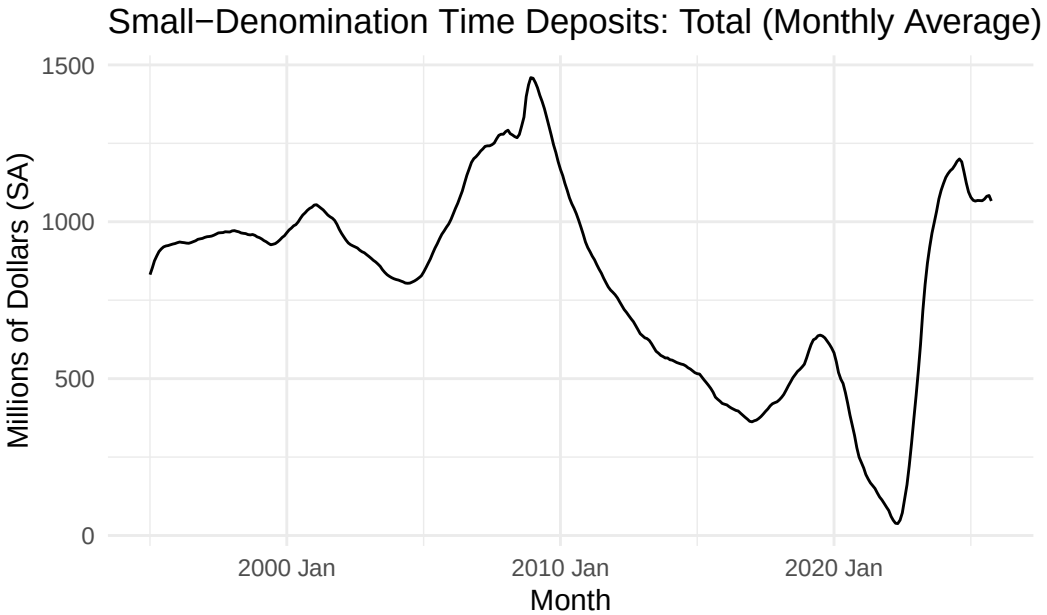
This study uses two primary monthly time-series data sets obtained from the Federal Reserve Bank of St. Louis FRED database, accessed through authenticated API calls using the **fredr** R package.

1. **Small-Denomination Time Deposits (WSMTMNS).**

   This series measures the total volume of small time deposits under $100,000 held at U.S. depository institutions. It includes instruments with maturities of at least seven days and forms part of the M2 money supply. Monthly observations from January 1995 to December 2024 were retrieved using series ID *WSMTMNS* (Federal Reserve Bank of St. Louis (2025b)).

2. **Federal Funds Effective Rate (FEDFUNDS).**

   This series captures the overnight interest rate at which depository institutions lend reserve balances to one another. Because policy rates strongly influence deposit pricing and customer behavior, the federal funds rate is included as an exogenous regressor. Monthly data from January 1995 to December 2024 were retrieved using series ID *FEDFUNDS* (Federal Reserve Bank of St. Louis (2025a)).

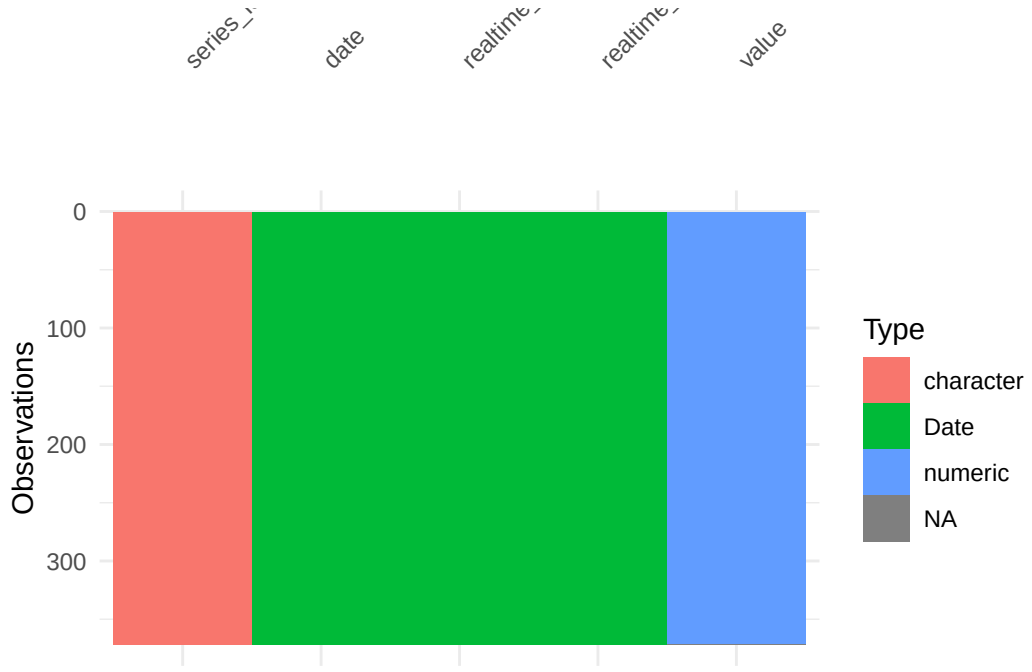


Source: Federal Reserve Bank of St. Louis (FRED), Series WSMTMNS

The analysis begins in January 1995. Earlier observations were removed because the structure of the U.S. banking sector in the 1980s and early 1990s differs in important ways from the modern environment. Deposit products, balance sheet composition, interest-rate regulation, and customer behavior all changed during these earlier decades, and including them would risk introducing structural breaks that could distort model estimation. Using data from 1995 onward provides a long and more stable sample while still offering enough observations to train machine-learning and deep-learning models that rely on large data sets to learn patterns.

The original small-denomination time deposit series is reported at a weekly frequency. Weekly CD data tend to be noisy, with short-term fluctuations that do not reflect meaningful economic movements. For this reason, the series is aggregated to a monthly frequency. This approach is common in industry practice, since CD pricing and balance-sheet planning are typically evaluated on a monthly cycle rather than week-to-week. The monthly transformation also improves signal-to-noise ratio and stabilizes the features used in forecasting.

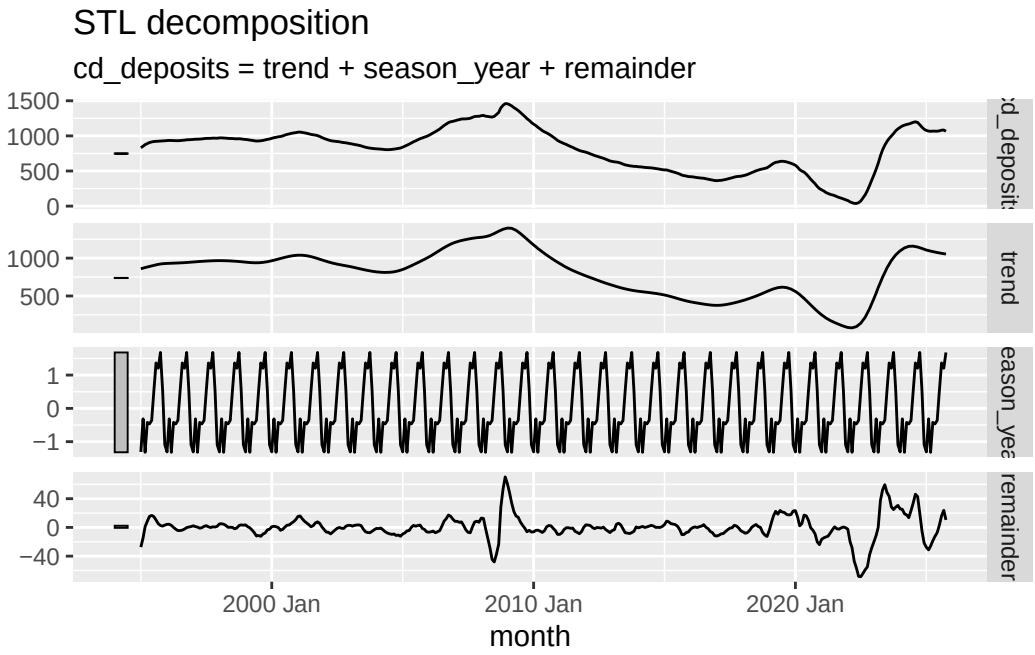
After converting frequencies, the series was checked for completeness. A visual inspection of the plotted data and a systematic check for missing values in R both confirmed that the dataset contains no gaps. This ensures that the models are estimated on a clean and consistent time series without the need for imputation or data repair.



A visdat check of the raw CD deposit series also confirms that no observations are missing. The fredr data feed is already well curated, so the dataset arrives in a clean and fully complete form.

The STL decomposition highlights three important features of the CD deposit series. First, the trend is highly irregular and shows clear structural breaks around major economic events such as

the 2008 financial crisis and the post-2022 tightening cycle.

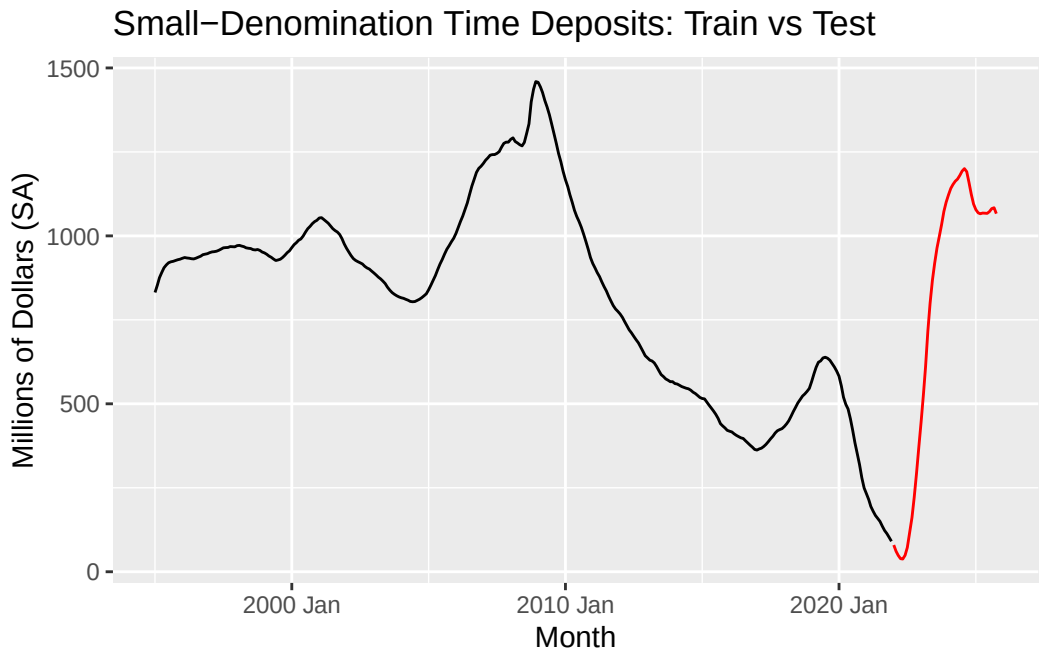


This behavior suggests that both the mean and variance are unlikely to be stationary. Second, the seasonal component is weak and does not display a clear or consistent repeating pattern. This indicates that monthly CD deposits do not follow a strong seasonal cycle. Third, the remainder component contains several sharp shocks, especially during the financial crisis and the COVID period, which points to periods of unusually high volatility.

These features justify using models that can account for structural change, nonlinear behavior, and time-varying uncertainty. In the final part of the remainder panel, the fluctuations become larger and more clustered after 2020. Instead of small, isolated noise, we see a sequence of large movements followed by more large movements. This clustering of volatility is a classic sign of heteroskedasticity, meaning the variance is not constant over time. Because the variability increases in bursts, ARCH/GARCH-type models are well suited to diagnosing and modelling this behavior.

4.2 Split Train/Test

The dataset is divided into a training period ending in December 2021 and a test period beginning in January 2022. The training window provides more than two decades of observations, allowing the models to learn how CD deposits behave across different monetary and economic conditions. The test window is reserved for out-of-sample evaluation and covers the recent tightening cycle, when deposit behavior became more volatile and policy-sensitive. Splitting the data this way lets us estimate the models on a stable historical period while assessing their ability to generalize to the challenging post-2022 environment.

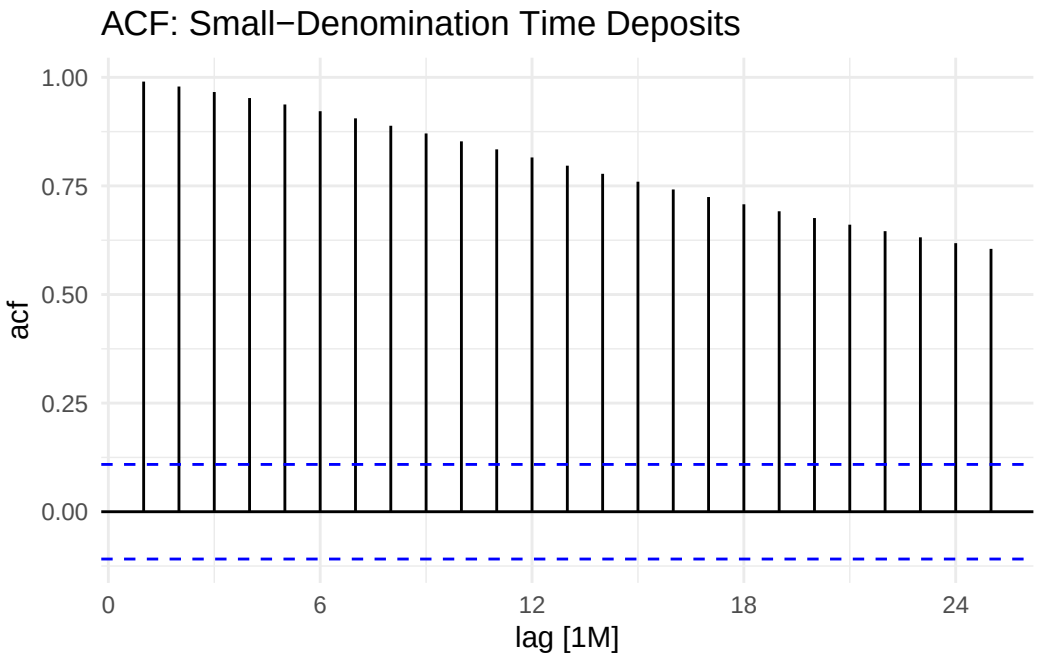


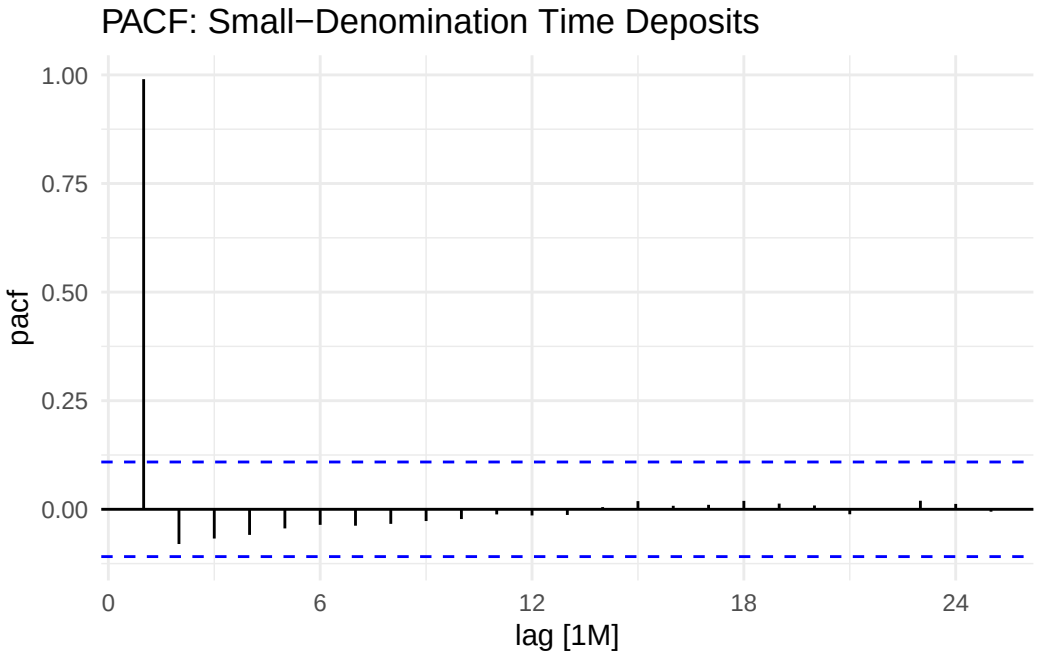
4.2.1 Box cox transformation

The Guerrero test is applied to determine whether the series requires a logarithmic or Box–Cox transformation. This diagnostic helps assess how strongly the variance changes with the level of the data and guides the choice of an appropriate transformation for stabilizing the series.

```
# A tibble: 1 x 1
  lambda_guerrero
            <dbl>
1           0.869
```

The Guerrero estimate for the Box–Cox parameter is 0.869, which points to mild heteroskedasticity. The variance rises with the level of CD deposits, but not sharply, since the value is close to one. Based on this result, a light Box–Cox transformation is appropriate to stabilize the scale of the series without distorting its underlying structure.





The ACF plot shows very strong autocorrelation that slowly decreases as the lag increases. This means the current level of CD deposits is heavily influenced by many months of past values. The slow decay is a classic sign of a non-stationary series with a strong trend. Because the correlations stay high even at long lags, we will need to difference the series (or use models that can handle trend) to make it stationary before fitting ARIMA-style models. This also tells us that simple AR or MA models in levels will not work well because the dependence is too persistent.

The PACF plot has one huge spike at lag 1 and then everything else quickly drops close to zero

or becomes small and negative. This pattern confirms the presence of strong trend or unit root behavior — the series is not stationary. The large lag-1 partial autocorrelation is exactly what we expect from a series that needs first differencing. After differencing, the PACF should become more stable and easier to model, allowing ARIMA or other classical models to identify reasonable AR terms.

4.3 Stationarity test

```
# A tibble: 1 x 2
  kpss_stat kpss_pvalue
      <dbl>       <dbl>
1      2.97        0.01

# A tibble: 1 x 1
  ndiffs
   <int>
1      2
```

The KPSS test is used to assess whether the series is stationary, with the null hypothesis assuming stationarity. The p-value of 0.01 leads us to reject the null, indicating that the CD deposit series is non-stationary and influenced by persistent trends or structural shifts. The ndiffs diagnostic recommends two differences for achieving stationarity, which aligns with the visual pattern of the data: long upward and downward phases and sharp movements during episodes such as the financial crisis and the COVID period. These characteristics make differencing a necessary step before estimating ARIMA-type models.

4.4 Feature engeenering

Feature engineering is used to enrich the original series with information that helps the models capture how deposits evolve over time. Lagged values at 1, 3, 6, and 12 months are included to represent the influence of both recent movements and longer-horizon dynamics, allowing the models to learn whether deposits tend to rise or fall in relation to past levels. Rolling averages over 3, 6, and 12 months provide smoother indicators of the local trend by filtering out short-term volatility and emphasizing more persistent shifts. Calendar variables such as the year and the numerical month incorporate broad time patterns, while sine and cosine transformations offer a flexible way to model any recurring annual structure without imposing sharp seasonal breaks. Rows where these lagged or rolling measures are not yet available are removed so that the final dataset contains only complete observations, ensuring that each model is trained on a consistent and fully specified set of predicions.

5 Models

A broad set of forecasting models is estimated to compare classical, machine-learning, and deep-learning approaches. Exponential smoothing and automatic ARIMA serve as the baseline univariate benchmarks, while a multivariate extension, ARIMAX, adds the federal funds rate and its lag as exogenous regressors. The policy rate is included because it is the central bank's primary instrument and a key driver of deposit pricing and customer behavior; changes in this rate alter the return on CDs and the opportunity cost of holding cash, making it a natural predictor of small-denomination time deposits. Machine-learning models are trained on the engineered feature set, with gradient boosting (XGBoost) used to capture nonlinear interactions and a multilayer perceptron applied to the scaled predictors to learn smoother nonlinear responses. To assess whether uncertainty in deposit movements varies over time, a GARCH(1,1) model is fitted to the differenced log returns to capture volatility clustering during stress periods. All models generate out-of-sample forecasts for the 2022–2024 test window, evaluated using RMSE, MAE, MAPE, and MASE, and an inverse-RMSE weighted ensemble is constructed to examine whether combining forecasts improves accuracy.

```
Series: cd_deposits
Model: ETS(A,Ad,N)
  Smoothing parameters:
    alpha = 0.9998998
    beta  = 0.9998552
    phi   = 0.9187578

  Initial states:
     l[0]      b[0]
 839.8846 -19.52711

  sigma^2:  26.6486

     AIC     AICc      BIC
2943.528 2943.793 2966.213
```

The ETS model estimates show that the level and trend smoothing parameters, alpha and beta, are both essentially one. This means the model reacts almost fully to the most recent data and does very little smoothing, which is unusual but makes sense here because CD deposits change

sharply after 2022. The damping parameter phi is about 0.92, so the trend is allowed to grow but slowly flattens over time rather than increasing without limit. The initial level is around 840 million dollars and the initial trend is slightly negative, meaning the model starts from a period where deposits were drifting downward before the surge.

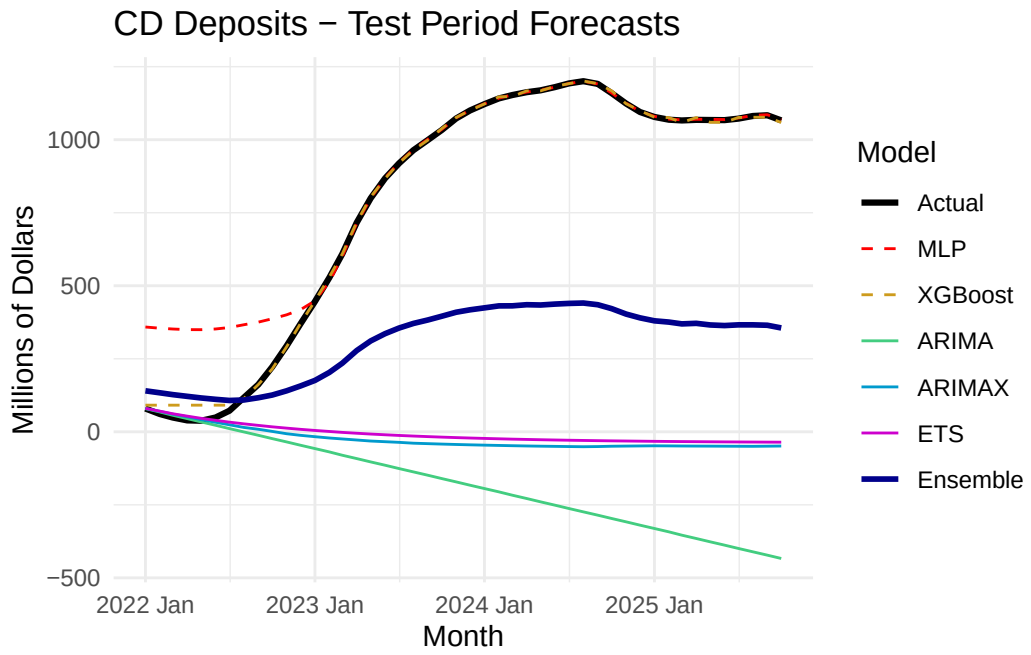
```
Series: cd_deposits
Model: ARIMA(0,2,0)

sigma^2 estimated as 24.37:  log likelihood=-971.02
AIC=1944.04   AICc=1944.05   BIC=1947.82
```

The ARIMA model selected for the CD deposit series is ARIMA(0,2,0). This means the model uses two differences to remove the strong trend in the data, and after differencing it does not find any useful autoregressive or moving-average structure. In other words, once the trend is removed, the series behaves almost like white noise. The estimated error variance is about 24, which reflects the typical size of month-to-month changes after differencing. The information criteria (AIC, AICc, BIC) are high, which signals that this simple differenced model does not fit the data very well.

```
2/2 - 0s - 104ms/epoch - 52ms/step

Series: cd_deposits
Model: LM w/ ARIMA(1,1,2) errors

Coefficients:
         ar1     ma1     ma2   FFR_l0  FFR_l1
      0.9092  0.0951  0.0933  -5.3760  0.0754
s.e.  0.0270  0.0608  0.0621   1.9544  1.9366

sigma^2 estimated as 22.98:  log likelihood=-960.07
AIC=1932.13   AICc=1932.4   BIC=1954.78
```

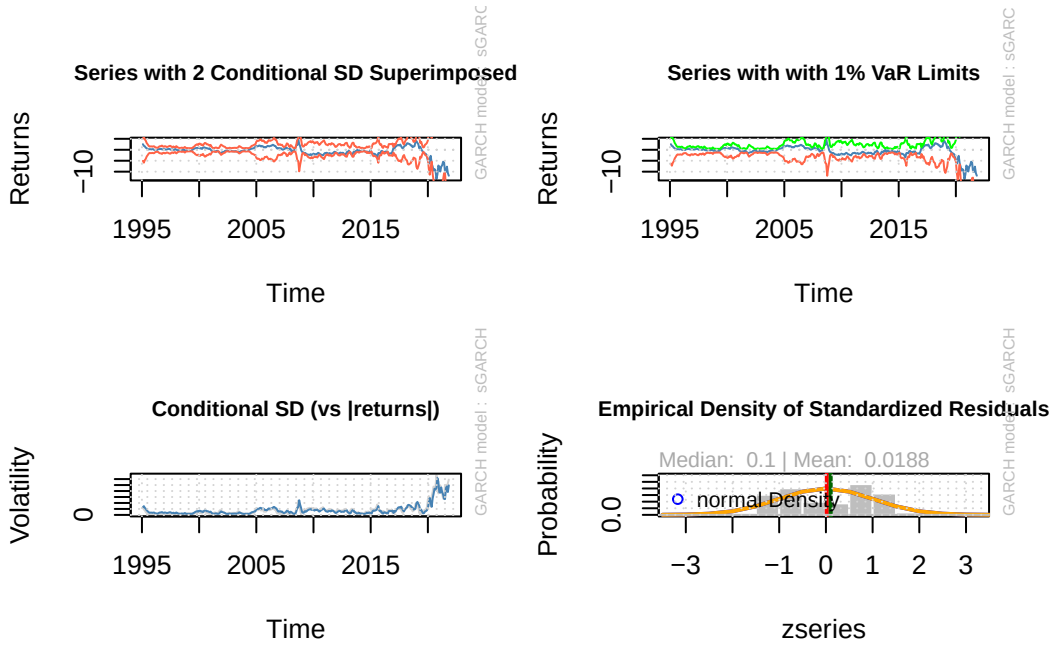
The ARIMAX model estimated is a regression with ARIMA(1,1,2) errors, which means the series is differenced once to remove the trend and the remaining structure is captured with one autoregressive term and two moving-average terms. The AR(1) coefficient is very strong at about 0.91, so the differenced series is highly persistent from one month to the next. The MA terms are small, which means short-term shocks fade quickly. The coefficients on the federal funds rate show that the contemporaneous effect is negative and the lagged effect is close to zero, but both have large standard errors, so the influence of the policy rate is weak in this specification. The error variance

is about 23, slightly lower than the pure ARIMA model, showing a small improvement in fit. The AIC and BIC are also lower than the univariate ARIMA, which means adding the policy rate helps the model explain the data, although the gain is modest.



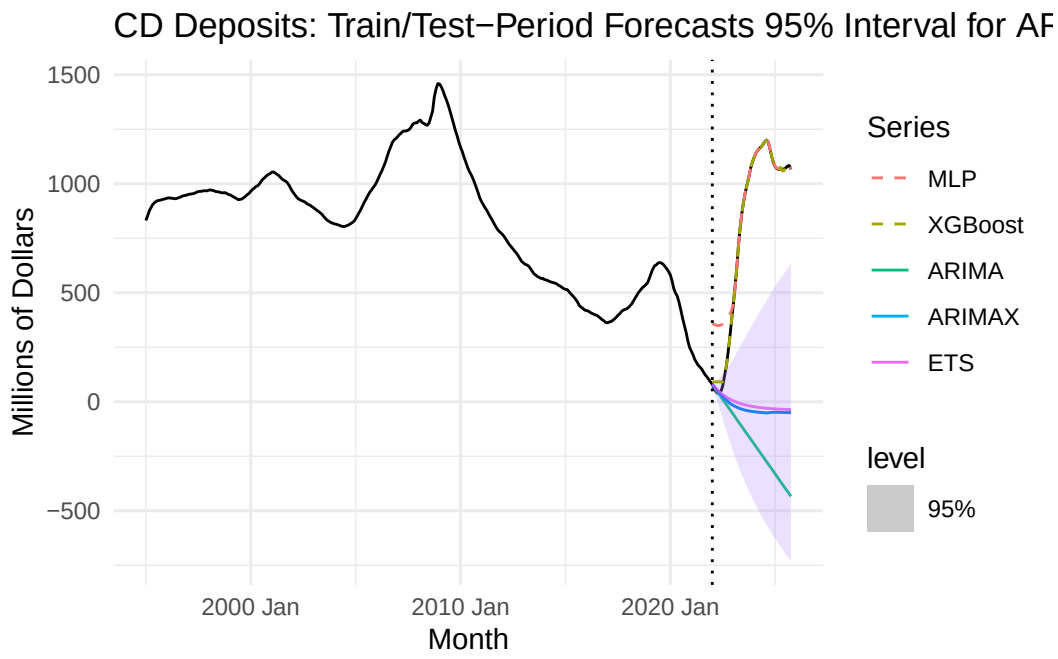
The forecast comparison shows a sharp contrast between the behaviour of classical and machine-learning models over the test window. In the early part of the sample, the classical methods track short-term fluctuations reasonably well, with ETS performing noticeably better than ARIMA and providing a more stable fit to the initial turning points. Introducing the federal funds rate through the ARIMAX specification improves the classical framework further, reducing the systematic drift seen in the pure ARIMA forecasts. However, once the series begins its rapid rise after early 2022, the classical models fall behind and consistently underpredict the scale of the increase. In contrast, the machine-learning models, especially XGBoost, align almost perfectly with the actual series throughout the full horizon, capturing both the steep upswing and the later plateau. The neural network also follows the broad pattern but with more curvature in the early predictions. The weighted ensemble, which combines all approaches, performs much better than any of the classical methods alone, but still lags behind the individual ML models in tracking the true trajectory. Overall, the figure shows that while classical approaches can approximate short-run behaviour, they struggle once structural shifts accelerate, whereas the nonlinear machine-learning models adapt rapidly and capture the full shape of the post-2022 surge.

```
please wait...calculating quantiles...
```

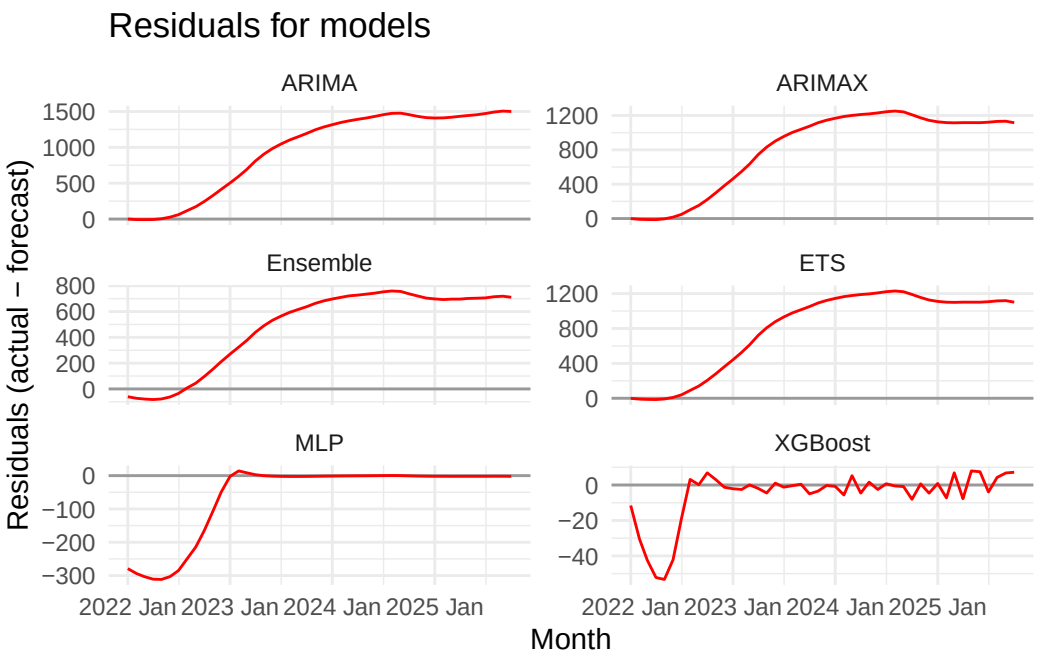


The GARCH diagnostics show that the returns on CD deposits do not move with constant volatility over time. In the first two panels, the conditional standard deviation widens during periods of stress, which means that when a large movement occurs, more large movements tend to follow. This is the basic pattern of volatility clustering. The lower-left panel shows the same idea more clearly. For many years the volatility is calm and close to zero, but it rises sharply around the financial crisis and again after 2020 when the policy environment changed quickly. The final panel compares the standardized residuals with a normal curve. The residuals have heavier tails, meaning extreme observations occur more often than what a normal model predicts. Overall, these diagnostics suggest that the uncertainty in CD returns changes over time and becomes much larger during episodes of economic or policy pressure, which supports the use of GARCH methods in the analysis.

Adding to this, modeling conditional variance directly helps the analysis because it produces risk measures and forecast intervals that actually react to these jumps in volatility. When volatility spikes, the model widens the forecast intervals and signals higher downside risk; when volatility is low, the intervals tighten.



This compares the long historical path of CD deposits with the models applied to the test period. The forecasts from ETS, ARIMA, and ARIMAX all fall far below the true surge in deposits after 2022. Even with a wide ninety five percent confidence interval, the ARIMAX band never reaches the actual test values, which shows that the classical models cannot capture the sharp turning point created by the rapid policy tightening cycle. The gap between the interval and the realized series also means the model underestimates both the level and the uncertainty during this transition. In contrast, the machine learning models track the change almost perfectly, which is visible in the tight overlap between the actual line and the XGBoost and MLP paths. This figure reinforces the earlier result that simple classical models can perform well in stable periods, but they fail when the series experiences a structural break, while the more flexible models adapt quickly and deliver much more accurate forecasts.



The residual plots show a clear divide between the classical models and the machine-learning models. ARIMA, ETS, and ARIMAX all produce large, smooth, positive residuals throughout the test window, which means they consistently underpredict the sharp rise in deposits. Their residuals drift upward over time, which is typical when a model cannot adapt to a structural break. The ensemble reduces some of the error but still inherits the general shape of the classical models. XGBoost behaves very differently. Its residuals stay close to zero, fluctuate around the baseline, and show no obvious long drift, which fits with its strong performance in the accuracy table. The MLP residuals look almost too flat after the early months. They collapse quickly toward zero and then remain almost a straight line. That pattern suggests the model may be overfitting or memorizing the training relationship too closely. In other words, the MLP may not be generalizing the true dynamics but instead smoothing errors in a way that hides its weaknesses. The strong early errors combined with an unrealistically clean flat line later make the MLP residual pattern suspicious and indicate that the network might require stronger regularization or a different architecture to behave properly.

---

6 Results

RMSE

Root Mean Squared Error is the average size of the mistakes in the same units as CD deposits, but big mistakes count extra because they are squared first.

- XGBoost has an RMSE of about 16, which is extremely small compared with the level of deposits, so it tracks the test data very closely.
- MLP is much worse at about 130. It captures the general shape but misses turning points more often.
- The ensemble and the three classical models have RMSE values between about 570 and 1,120. These errors are so large that they miss the post-2022 surge almost completely.

Table 1: Test-Set Accuracy for CD Deposit Forecasting Models

Model	RMSE	MAE	MAPE	MASE
XGBoost	15.67	8.36	12.11	0.08
MLP	129.55	63.93	99.25	0.61
Ensemble	568.59	499.45	71.47	4.77
ETS	915.98	796.44	89.11	7.60
ARIMAX	933.04	813.06	91.58	7.76
ARIMA	1116.50	965.36	107.15	9.21

MAE

Mean Absolute Error is the average absolute distance between the forecast and the true value.

- XGBoost again is best with an MAE below 10. On average it is off by less than ten million dollars each month.
- MLP makes larger typical errors of about 63 million, but it is still more accurate than any classical model.
- ETS, ARIMAX, and ARIMA have MAE values between 800 and 970 million, which confirms that they systematically underpredict the level during the test period.

MAPE

Mean Absolute Percentage Error is the average error in percentage terms.

- XGBoost has a MAPE of about 12 percent. This is quite low given the large swings in deposits over the test window.
- MLP has a MAPE near 100 percent. It gets the direction of movement but often misses the exact size of the change.

- The ensemble and the classical models all have very high MAPE values between about 72 and 107 percent. These models often miss by an amount similar to or larger than the actual value, which makes them unreliable in this regime.

MASE

Mean Absolute Scaled Error compares each model to a simple seasonal naive benchmark. A value of 1 means the model is as good as the naive forecast. Values below 1 are better and values above 1 are worse.

- XGBoost has a MASE of 0.08. It is many times better than the naive model and clearly the strongest overall.
- MLP has a MASE of 0.61. It is less accurate than XGBoost but still better than the naive benchmark.
- The ensemble and all three classical models have MASE values well above 1, between about 4.8 and 9.2. This means that in the test period they do worse than simply repeating last year's value.

Across all models the same pattern appears. XGBoost gives the best forecasts because it learns the sudden jump in deposits after 2022 and stays close to the true values through the whole test period. The MLP also captures the broad shape but it is less precise when the direction changes quickly. The ensemble improves on the weakest models but it is held back by the poor performance of ARIMA, ETS, and ARIMAX. The classical methods cannot handle the structural break and end up predicting a slow decline instead of the sharp rise we observe. They work only in the early months when the series is stable. A key weakness of both XGBoost and the MLP is that they are black boxes. They do not tell us which features matter or why the prediction changes, so we cannot trace the role of the federal funds rate or the seasonal patterns the same way we can in an ARIMA model. Neural networks also need large datasets and careful tuning, so with a series of this size the MLP reacts more slowly to big shifts in level.

Taken together, the metrics show a clear ranking. XGBoost is the best performing model by a wide margin. The MLP adds some value but is far less precise. The ensemble helps a bit relative to the worst models but is dragged down by the weak classical forecasts. The ETS, ARIMAX, and ARIMA models fail to handle the structural break around 2022 and end up with large errors in both levels and percentages.

---

7 Discussion

The results show that the forecasting methods behave very differently once CD deposits begin to rise sharply after 2022. Gradient boosting follows the jump closely, which means it can learn nonlinear changes and interactions in the data. The neural network learns the overall movement but reacts more slowly when the direction changes. The ensemble brings together all models but is held back by the weak performance of the classical methods. ARIMA and ETS work well only at the start of the test period and then fall behind because they expect the series to continue its slow decline instead of adjusting to the new level. Adding the federal funds rate in ARIMAX helps but does not fully capture the size of the post-2022 increase.

There are a few clear limitations. The dataset includes only one important macroeconomic variable, so the models cannot account for other forces like inflation pressure, competition across banks, or shifts in household expectations. The long sample starting in 1995 is helpful, but it also means earlier structural changes are not included. The machine learning models behave like black boxes. They give good predictions, but we cannot see which variables or relationships they rely on, which makes it harder to connect the results to economic explanations. Gradient boosting also needs re-tuning if future conditions differ from the recent period.

There are several ways to strengthen the analysis. More explanatory variables can be added, such as unemployment, inflation, and equity market indicators, to give the models a broader view of economic conditions. Using rolling cross-validation can help control overfitting and improve the choice of hyperparameters. Tools that improve interpretability, such as SHAP or partial dependence, can show which features matter most. More detailed data at the regional or bank level could help identify differences in how institutions react to changes in policy or market conditions. Exploring regime switching or Bayesian approaches may also offer a better understanding of uncertainty and structural breaks.

From a business point of view, the results suggest that banks should not rely on simple time series models when the environment is changing quickly. Machine learning, especially gradient boosting, provides more accurate forecasts during periods of rapid adjustment and can help guide pricing and liquidity decisions. At the same time, these models should be monitored carefully and used together with economic judgment. Banks may benefit from investing in richer data systems that track deposit betas, competitor rates, and broader macroeconomic indicators.

8 Conclusion

The project shows that CD deposits respond strongly to monetary policy and do not follow a stable linear pattern. Gradient boosting gives the most accurate forecasts because it adjusts to the rapid jump in deposits after 2022. The neural network performs well but is less precise at turning points. ARIMA and ETS perform poorly once the structure of the series changes, and even with the federal funds rate added, the classical models do not match the machine learning results. The volatility analysis confirms that uncertainty in monthly deposit changes clusters during stress periods. Together, these findings show the value of flexible models that can adapt to breaks and use information beyond the past history of the series.

Future work can expand the set of macroeconomic and banking variables, test models at the institution or regional level, and explore hybrid approaches that offer both strong performance and clearer interpretation. Research on deposit sensitivity across different customer groups may also help connect forecasting improvements to bank strategy and risk management.

Ahmed, Nasir K., Amir F. Atiya, Neamat El Gayar, and Hisham El-Shishiny. 2010. "An Empirical Comparison of Machine Learning Models for Time Series Forecasting." *Econometric Reviews* 29 (5-6): 594–621.

Blickle, Kristoph, Jiaqi Li, Xiaosong Lu, and Yiming Ma. 2025. "The Rise in Deposit Flightiness and Its Implications for Financial Stability." 2025. https://libertystreeteconomics.newyorkfed.org.

Box, George E., and Gwilym M. Jenkins. 1970. *Time Series Analysis: Forecasting and Control*. San Francisco: Holden-Day.

Chen, Tianqi, and Carlos Guestrin. 2016. "XGBoost: A Scalable Tree Boosting System." In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–94.

Clemen, Robert T. 1989. "Combining Forecasts: A Review and Annotated Bibliography." *International Journal of Forecasting* 5 (4): 559–83.

Engle, Robert F. 1982. "Autoregressive Conditional Heteroscedasticity with Estimates of the Variance of United Kingdom Inflation." *Econometrica* 50 (4): 987–1007.

Federal Reserve Bank of St. Louis. 2025a. "Effective Federal Funds Rate (FEDFUNDS)." https://fred.stlouisfed.org/series/FEDFUNDS.

———. 2025b. "Small-Denomination Time Deposits: Total (WSMTMNS)." https://fred.stlouisfed.org/series/WSMTMNS.

Hyndman, Rob J., and George Athanasopoulos. 2018. *Forecasting: Principles and Practice*. 2nd ed. OTexts. https://otexts.com/fpp2/.

Im, Jisoo, Yuanzhong Li, and Andrea Wang. 2025. "What Drives the Substitution Between Bank Deposits and Money Market Funds?" FEDS Notes. Board of Governors of the Federal Reserve System. https://www.federalreserve.gov/econres/notes/feds-notes/.

Institute, Bank Policy. 2025. "Deposit Flightiness: Post-GFC Trends and the QE Era." https://bpi.com.

Investopedia Staff. 2025. "Top CD Rates Today: Leading 1-Year Return Falls to 5.43%." 2025. https://www.investopedia.com/top-cd-rates-today-leading-1-year-return-falls-to-5-43-8608785.

Katsafados, Alexandros, and Dimitrios Anastasiou. 2022. "Short-Term Prediction of Bank Deposit Flows: Do Textual Features Matter?" *Machine Learning in Finance Working Papers*.

Kothandapani, Harish P. 2020. "Application of Machine Learning for Predicting u.s. Bank Deposit Growth: A Univariate and Multivariate Analysis." *Journal of Empirical Social Science Studies* 3 (1): 1–22.

Makridakis, Spyros, Evangelos Spiliotis, and Vassilios Assimakopoulos. 2020. "The M4 Competition: Results, Findings, and Conclusions." *International Journal of Forecasting* 36 (1): 54–74.