PREDICTING HOUSEHOLD FINANCIAL FRAGILITY WITH TABNET

COMPARATIVE PERFORMANCE, INTERPRETABILITY, AND IMBALANCE ANALYSIS

ANRI RRUMBULLAKU

THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF SCIENCE IN DATA SCIENCE & SOCIETY
AT THE SCHOOL OF HUMANITIES AND DIGITAL SCIENCES
OF TILBURG UNIVERSITY

STUDENT NUMBER

2132222

COMMITTEE

dr. Görkem Saygili
dr. Travis Wiltshire

LOCATION

Tilburg University
School of Humanities and Digital Sciences
Department of Cognitive Science &
Artificial Intelligence
Tilburg, The Netherlands

DATE

December 1st, 2025

WORD COUNT

7871

ACKNOWLEDGMENTS

I would like to take this chance to thank my supervisor, dr. Görkem Saygili, for his guidance during the process of writing this thesis. His kindness and support was greatly appreciated.

PREDICTING HOUSEHOLD FINANCIAL FRAGILITY WITH TABNET

COMPARATIVE PERFORMANCE, INTERPRETABILITY, AND IMBALANCE ANALYSIS

ANRI RRUMBULLAKU

Abstract

Financial fragility, defined in this thesis as a household's inability to cover three months of essential expenses using liquid assets, remains a widespread challenge in household finance. While traditional models offer useful insights, their linear structure may limit their ability to capture the non-linear patterns underlying financial vulnerability. This thesis evaluates whether TabNet, an interpretable deep learning architecture for tabular data, can improve the prediction of liquidity-based financial fragility compared with established models such as Logistic Regression, Random Forests, and XGBoost.

Using the 2022 Survey of Consumer Finances (SCF), a binary fragility indicator is constructed, and the data are preprocessed through implicate pooling, categorical encoding, leakage-free feature selection, and standardisation. All models are trained and assessed using nested cross-validation and evaluated using ROC-AUC, precision, recall, and F1-score, with separate analyses for unbalanced and class-weighted training pipelines due to the minority share of fragile households.

Results show that XGBoost consistently achieves the highest predictive performance, with Random Forest and Logistic Regression performing strongly as well. TabNet underperforms in the unbalanced setting due to low recall but improves considerably when class weights are applied, reaching competitive yet still lower F1-scores. TabNet's feature masks indicate that housing wealth, wage income, retirement liquidity, emergency savings behaviour, and selected demographic variables are the most influential predictors.

Overall, the thesis provides a transparent and empirically grounded comparison of modern modelling approaches for financial fragility. The findings highlight both the potential and the limitations of deep learning in this context and offer evidence-based guidance for identifying households at risk of financial vulnerability.

1 DATA SOURCE, ETHICS, CODE, AND TECHNOLOGY STATEMENT

This thesis uses data from the Survey of Consumer Finances (SCF) 2022, provided by the Board of Governors of the Federal Reserve System (U.S.) and accessed through the Federal Reserve's official portal(Board of Governors of the Federal Reserve System, 2023). The dataset is publicly available, fully anonymized, and does not contain any personally identifiable information. No new data were collected from human participants or animals. The Federal Reserve Board, as the data owner, grants public research use of the SCF for academic purposes, and therefore no additional consent was required.

All figures, tables, and visualizations presented in this thesis are created by the author using Python (Matplotlib, Seaborn). No copyrighted or third-party images were reused.

All code used in this project was written by the author in Python 3.11. OpenAI (2025)'s ChatGPT (GPT-5) has been used as a de-bugging tool to resolve any coding errors. The analysis relies on standard open-source libraries, including pandas, numpy, scikit-learn, xgboost, pytorch-tabnet, matplotlib, and seaborn). Code fragments adapted from official documentation or open-access examples are properly cited, and the full implementation is stored in the following public GitHub repository for reproducibility https://github.com/AnriRr/Financial-Fragility.

For language and writing refinement, OpenAI (2025)'s ChatGPT (GPT-5) was used to improve phrasing and structure as well as grammar and spelling checks while maintaining full control over the scientific content. The thesis report was typed using Overleaf following the Tilburg University Data Science and Society Master Thesis Template.

2 INTRODUCTION

2.1 *Context: Scientific and Societal Relevance*

Financial fragility refers to the limited capacity of households to absorb short-term economic shocks, such as unexpected expenses, sudden income loss, or increases in the cost of living. Prior work by Lusardi et al. (2011) shows that a substantial share of U.S. households would struggle to raise an emergency amount within 30 days, highlighting the prevalence of liquidity constraints across income groups. Subsequent studies reinforce this view, linking fragility to insufficient liquid savings. (Babiarz & Robb, 2014; Hasler et al., 2017). Recent macroeconomic developments, including rising housing costs, increased price volatility, and increased income uncertainty, have further intensified concerns about the ability of households to with-

stand short-term financial stress. Evidence from the Survey of Consumer Finances (SCF) indicates that many households continue to hold limited liquid assets despite larger increases in wealth (Aladangady et al., 2023).

For policymakers, consumer protection agencies, and central banks, accurately identifying financially fragile households is essential for designing targeted interventions. The Federal Reserve's reports on household economic well-being regularly emphasise emergency savings and liquid asset buffers as indicators of financial resilience (Board of Governors of the Federal Reserve System, 2022). From this societal perspective, improving the prediction of financial fragility has immediate practical value for resource allocation, financial education, and stability-oriented planning.

Scientifically, research on financial fragility has traditionally relied on classical methods such as logistic regression, which offer transparency and a close link to economic theory. However, these models assume linear and additive relationships among predictors, which may fail to capture the complex, non-linear interactions in household financial behaviour. Modern machine learning approaches address this limitation by modelling higher-order interactions flexibly. Empirical work in credit risk and consumer distress prediction demonstrates that machine learning models, including ensemble methods and deep neural networks, can improve predictive performance over linear models (Addo et al., 2018; Albanesi & Vamossy, 2019; Liu et al., 2022). Yet despite their promise, the application of advanced machine learning techniques to household financial fragility has been limited.

A further scientific motivation arises from recent progress in deep learning for tabular data. While traditional neural networks often underperform on structured datasets, architectures such as TabNet (Arik & Pfister, 2021) and other specialised models (Borisov et al., 2021) have sometimes demonstrated competitive performance alongside improved interpretability. These models leverage attention mechanisms and feature sparsity to identify relevant predictors in a transparent manner. Nonetheless, TabNet's potential in the context of household finance and financial fragility in particular, has not yet been systematically examined. This gap provides a compelling rationale for the present study.

2.2 *Research Strategy*

To address the questions raised in the scientific and societal context, the present study evaluates whether the *TabNet* architecture can provide measurable improvements in predicting financial fragility relative to widely used baselines such as Logistic Regression, Random Forests, and XGBoost. The empirical analysis is conducted using microdata from the 2022 Sur-

vey of Consumer Finances (SCF), following a fully leakage-free modelling pipeline incorporating rigorous nested cross-validation, feature preprocessing, and class imbalance handling.

The analysis is guided by the following research questions:

Main Research Question

To what extent can TabNet improve the prediction of financial fragility compared to traditional classification models?

In addition, two sub-questions support a systematic examination of model behaviour and help uncover the mechanisms driving predictive performance:

Sub-Questions

1. Which household characteristics are most influential in predicting financial fragility according to TabNet's feature masks?
2. How does class imbalance impact the performance of the models used to predict financial fragility?

The overall objective of the study is therefore: (i)to determine whether a specialised deep learning architecture for tabular data offers improvements in predictive performance over established approaches, (ii)to generate insights into the household-level drivers of financial fragility that can support more effective and equitable policy interventions and (iii)to study the impact of class imbalance for each of these models.

3 RELATED WORK

This literature review combines prior research relevant to household financial fragility, liquidity adequacy, financial behaviour, and the use of machine learning to predict financial outcomes. Three strands of literature are central to this thesis: (i) conceptual and empirical research on financial fragility and emergency savings, (ii) evidence from the Survey of Consumer Finances (SCF) documenting household financial positions, and (iii) machine learning approaches for predicting financial distress using tabular data. These strands collectively motivate the construction of a liquidity-based definition of financial fragility and the application of modern machine learning techniques to predict it.

3.1 *Financial Fragility and Household Liquidity*

Financial fragility is typically understood as the inability of a household to absorb an economic shock without experiencing significant hardship. A fundamental contribution to this literature is provided by Lusardi et al. (2011), who define financial fragility as the inability to raise US$2,000 within 30 days. Their findings, based on nationally representative surveys, show that financial fragility is widespread in the United States and not limited to low-income groups. Even many middle-income households report difficulties in meeting modest emergency expenses, suggesting that traditional measures of economic well-being such as income or net worth may not fully capture a household's vulnerability to short-term shocks.

Subsequent research has expanded the framework used to study fragility. Hasler et al. (2017) examine both objective and perceived financial fragility and identify low levels of liquid savings and limited financial literacy as key contributors. Their analysis demonstrates that while income provides a partial buffer against shocks, the presence, or absence of liquid assets plays a decisive role in shaping a household's resilience. They also show that financially literate individuals tend to accumulate larger emergency savings and are less likely to experience fragility, reinforcing the importance of financial knowledge in shaping economic outcomes.

A related stream of research focuses on emergency savings and liquid asset adequacy. Babiarz and Robb (2014) analyse the determinants of emergency saving behaviour and find that many households maintain liquid balances well below recommended levels. Their work emphasises the importance of liquidity for coping with irregular expenses, unexpected events, and income volatility. They also show that emergency savings are strongly correlated with financial satisfaction and future financial security. This literature provides empirical support for defining fragility through observable liquidity shortages rather than self-reported expectations alone.

Institutional analyses strengthen this perspective by identifying liquidity sufficiency, particularly multi-month expense coverage as a key indicator of financial security. The Federal Reserve Board's report evaluates household preparedness for financial shocks by assessing the share of adults who could cover three months of expenses in an emergency (Board of Governors of the Federal Reserve System, 2022). This three-month benchmark is widely used in policy and financial education contexts and offers an objective, measurable threshold for identifying vulnerable households. Because these benchmarks are grounded in practical financial guidance and empirical measurement, they provide a strong justification for constructing a liquidity-based definition of financial fragility in this thesis.

3.2 *Household Finance and the Survey of Consumer Finances*

The Survey of Consumer Finances (SCF) is the primary source of data for analysing the financial circumstances of U.S. households. The SCF provides detailed information on income, assets, debts, and demographic characteristics, enabling rich analyses of household balance sheets. The SCF's sampling design includes oversampling of wealthy households, combined with multiple imputation for missing data, allowing researchers to accurately model the highly skewed distributions of wealth and income. The 2022 SCF documented in Board of Governors of the Federal Reserve System (2023) continues this approach and provides an up-to-date picture of household financial conditions.

Research based on the SCF highlights substantial variation in household liquidity and financial preparedness. Aladangady et al. (2023) examine changes in household finances between 2019 and 2022 and show that, while aggregate wealth increased, many households continue to hold minimal liquid assets relative to their expenses. The concentration of liquid assets among higher-wealth households implies that a large share of the population remains vulnerable to short-term disruptions. These findings align with the fragility literature, painting a consistent picture of widespread liquidity constraints.

Historical SCF analyses further demonstrate that liquid assets have not kept pace with overall wealth growth. For example, Aladangady et al. (2023) document changes in U.S. family finances from 2013 to 2016 and show that many households have very limited precautionary savings despite improvements in net worth. The combination of rising income volatility and persistently low liquidity buffers, suggests that fragility may be structurally embedded in the financial lives of many households. These insights provide a strong empirical rationale for modelling financial fragility using SCF data.

3.3 *Financial Behaviours and Determinants of Fragility*

Financial fragility does not arise solely from income and liquid asset levels. Financial behaviours, literacy, and debt management also play an important role. Allgood and Walstad (2021) study the role of perceived and actual financial literacy and find that both influence planning, saving, and borrowing decisions. Households with greater financial knowledge are more likely to maintain emergency savings and engage in budgeting behaviours that reduce the likelihood of financial strain.

Other machine learning-based analyses of financial distress emphasise the importance of behavioural and demographic predictors. de Waal et al.

(2023) develop interpretable ML models to predict consumer financial distress and show that spending patterns, liquidity, and debt-to-income ratios are strong indicators of vulnerability. Their results highlight the value of flexible, non-linear models that can capture interactions between financial and demographic variables.

Similarly, S. Chen et al. (2020) analyse financial distress across diverse household settings and show that ML models can provide accurate predictions using a mix of economic, behavioural, and demographic features. These studies broadly support the notion that fragility is shaped by multiple dimensions of household financial behaviour, making it suited for the heterogeneous modelling capacity of machine learning.

3.4 *Machine Learning for Financial Distress and Credit Risk*

Machine learning has been applied extensively to credit risk prediction and has shown great performance compared to traditional statistical models. Albanesi and Vamossy (2019) demonstrate that deep learning models improve the prediction of credit default, primarily by capturing non-linear relationships in transactional and balance sheet data. Their work provides strong evidence that ML methods can reveal patterns that standard logistic regression may miss.

A comprehensive comparison of ML models for credit risk by Liu et al. (2022) shows that ensemble methods such as Random Forests and XGBoost achieve high predictive performance across a variety of datasets. These models are effective at modelling complex interactions across financial features, making them appropriate baselines for predicting household financial fragility.

Addo et al. (2018) compare deep learning and traditional machine learning models under different macroeconomic conditions and find that tree-based and neural network models often outperform classical methods. Their work supports the use of advanced ML techniques in financial prediction tasks, including household fragility.

Further evidence comes from Qiu et al. (2024), who evaluate LightGBM, XGBoost, and TabNet for predicting consumer credit risk. Their findings suggest that modern hybrid models, including TabNet, can perform competitively with traditional gradient-boosted models. Because TabNet is designed specifically for tabular data, this provides preliminary evidence supporting its use in modelling household-level financial outcomes.

3.5 *Deep Learning for Tabular Data and TabNet*

Deep learning has historically struggled on tabular datasets compared to gradient-boosted trees. However, specialised architectures have emerged to address this gap. Shwartz-Ziv and Armon (2022) conduct a thorough comparison of deep learning and tree-based methods and conclude that deep models often underperform unless specifically designed for tabular data.

A detailed survey by Borisov et al. (2021) identifies several architectures, including TabNet, that offer competitive performance on tabular tasks. TabNet, introduced by Arik and Pfister (2021), uses sequential attention to generate sparse feature masks that highlight the most relevant predictors at each decision step. This design allows TabNet to combine the representational flexibility of deep learning with interpretable feature selection mechanisms. Its built-in interpretability makes TabNet particularly suitable for financial applications where transparency is essential.

Although attention-based interpretability is native to TabNet, broader ML literature demonstrates the utility of model-agnostic interpretability approaches. Lundberg and Lee (2017) introduce SHAP values as a unified framework for explaining ML model predictions. SHAP is widely used for tree-based and neural network models and serves as a useful interpretability metric for feature importance.

Class Imbalance in Financial Prediction

A well-documented challenge in modelling financial fragility and related credit-risk outcomes is the presence of substantial class imbalance, where instances of distress or default represent only a small fraction of the population. The machine learning literature emphasises that such imbalance can bias models toward majority-class predictions, reduce sensitivity to rare but economically important events, and distort evaluation metrics (He & Garcia, 2009). Krawczyk (2016) further note that imbalance affects not only predictive accuracy but also the stability of learned decision boundaries, making careful imbalance handling essential for reliable minority-class detection.

In financial applications specifically, several studies demonstrate that classifier performance declines sharply when minority events such as default, arrears, or financial hardship are under-represented in the training data. Brown and Mues (2012) provide one of the earliest systematic comparisons of algorithms on imbalanced credit-scoring datasets, showing that linear models, tree ensembles, and neural networks all exhibit substantial drops in recall when trained without cost-sensitive adjustments. More

recent work confirms these patterns, with boosting-based methods and class-weighting strategies improving sensitivity to distress cases (Addo et al., 2018; Liu et al., 2022).

The findings of Kanász et al. (2024) also underscore the importance of explicitly addressing class imbalance. Their study shows that TabNet is particularly sensitive to skewed class distributions and requires oversampling or cost-sensitive adjustments to achieve competitive sensitivity. This mirrors well-known challenges in imbalanced classification and supports the methodological choice in this thesis to evaluate both unbalanced and class-weighted training pipelines(He & Garcia, 2009; Krawczyk, 2016).

These findings directly motivate the imbalance-handling procedures adopted in this thesis, including class-weight adjustments for Logistic Regression and Random Forests, the tuning of `scale_pos_weight` in XGBoost, and the use of `pos_weight` in TabNet. Given the low prevalence of financially fragile households in the SCF, appropriate treatment of class imbalance is necessary to ensure meaningful comparative evaluation across models and to avoid systematically under-identifying fragile households.

3.6 *Research Gap*

Although prior research has identified several determinants of financial fragility and has demonstrated the usefulness of machine learning methods in predicting financial distress, two important gaps remain. First, existing studies largely rely on linear or tree-based models and do not evaluate modern deep learning architectures specifically designed for tabular data. As a result, it is unclear whether architectures such as TabNet can capture non-linear interactions in household balance sheets more effectively than traditional approaches. Second, while many contributions examine financial fragility or credit risk, few apply rigorous modelling pipelines to the SCF, particularly in combination with nested cross-validation, class-imbalance handling, and interpretable model outputs. Consequently, the predictive value and interpretability of deep tabular models in the context of liquidity-based financial fragility remain untested. This thesis addresses these gaps by providing a systematic comparison of TabNet and established models within a transparent and methodologically robust framework.

4 METHODOLOGY

This chapter describes the empirical strategy used to investigate the extent to which TabNet can improve the prediction of financial fragility compared to more traditional classification models, using data from the 2022 Survey of Consumer Finances (SCF). The methodology is designed to be transpar-

ent, reproducible, and aligned with proper practices in applied machine learning, with particular attention to class imbalance, and interpretability.

The chapter proceeds as follows. First, the overall research design is outlined with a flowchart. Second, the set of predictors and data preprocessing steps are explained. Third, Exploratory Data Analysis (EDA) on the dataset is explained. Fourth, the construction of the liquidity-based financial fragility target is described. Fifth, four classification models are discussed: Logistic Regression, Random Forest, XGBoost, and TabNet. Sixth, hyperparameter tuning procedure is explained and the hyperparameters search space is defined. Seventh, the treatment of class imbalance is motivated. Finally, evaluation metrics, and implementation choices are presented.

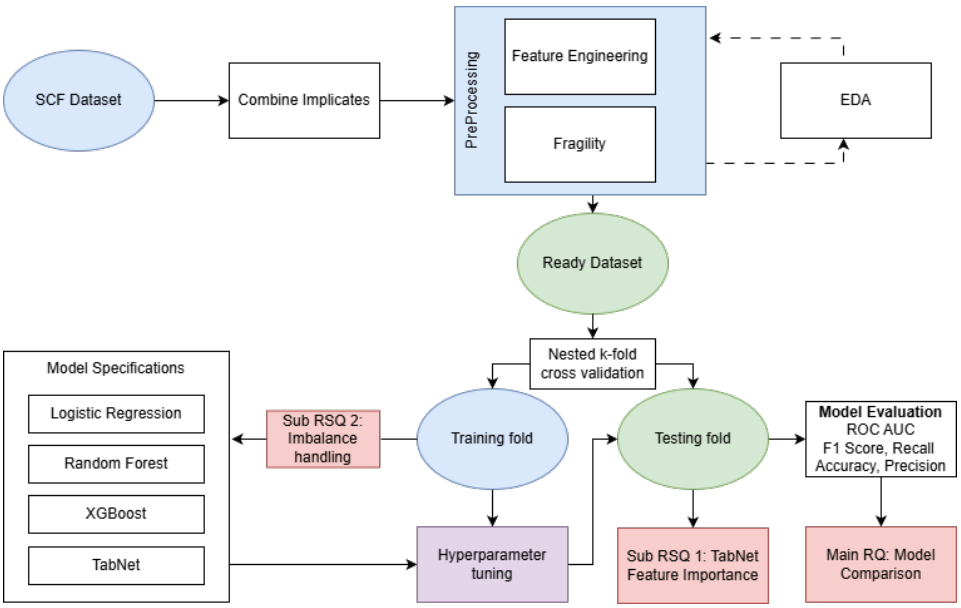


Figure 1: End-to-end modelling workflow for predicting financial fragility.

4.1 *Predictors and Preprocessing*

The predictor set is designed to capture a broad range of household characteristics that may influence financial fragility. These include demographic variables (such as age, marital status, household composition), indicators of education and financial literacy, labour market status, income and its components, housing tenure, debt positions, and selected variables capturing financial attitudes or recent financial difficulties. The choice of variables is informed by prior research on fragility, emergency savings, and financial behaviour (Allgood & Walstad, 2021; Babiarz & Robb, 2014; Hasler et al.,

2017; Lusardi et al., 2011), as well as by the practical constraints of working with SCF data.

The SCF contains a mixture of continuous variables (e.g. income, expenditure, balances) and categorical variables. Categorical predictors such as education level, family structure, or occupation are transformed into one-hot encoded dummy variables. When categories have very few observations, rare levels are grouped into an `OTHER` category. This reduces sparsity and helps models, especially deep networks, avoid overfitting to poorly supported categories (Borisov et al., 2021).

These steps yield a refined set of predictors that is both economically meaningful and methodologically sound. Variables that could cause leakage or were too sparse were removed or consolidated, while the remaining demographic, financial, and behavioural characteristics were encoded in a consistent and model-friendly format.

4.2 *Exploratory Data Analysis*

Before model development, an exploratory data analysis (EDA) was conducted to examine the structure of the pooled SCF dataset, identify general patterns in the predictors, and detect potential issues related to multicollinearity, class imbalance, and variable distributions. The EDA stage provides a descriptive overview of the data and informs subsequent preprocessing decisions.

A correlation heat map was computed for all continuous variables included in the initial feature set. As shown in Figure 2, the majority of pairwise correlations fall within a low to moderate range, indicating limited linear dependence between predictors. Strong correlations were observed only within logically related asset categories (e.g., liquid financial assets, retirement liquidity, or mutual fund holdings), which guided the removal of redundant asset subcomponents during feature selection.

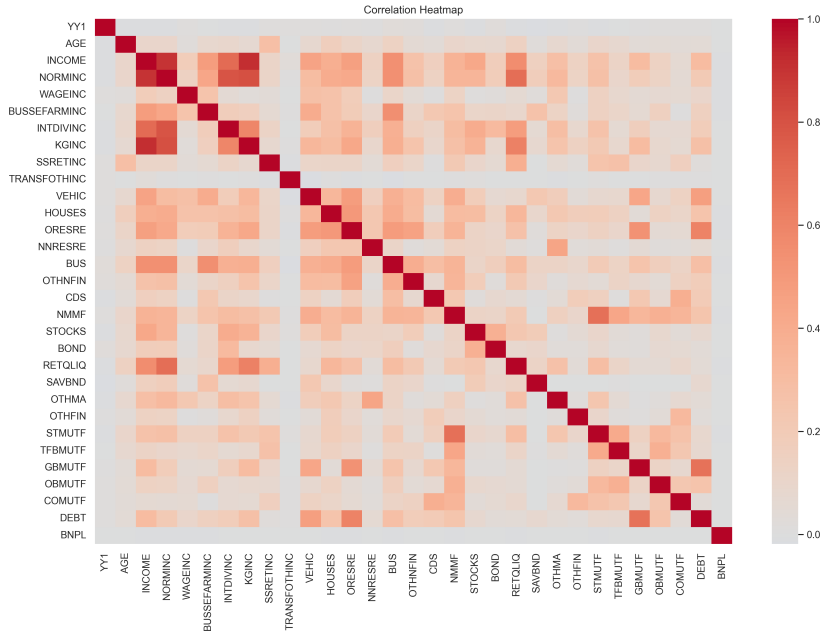


Figure 2: Correlation heatmap for continuous predictors in the SCF dataset.

The distribution of the binary target variable `fragile` was examined to assess the extent of class imbalance. As shown in Figure 3, the majority of households are classified as non-fragile, with approximately two-thirds of observations belonging to the negative class. This imbalance motivated the implementation of class-weighted learning for Logistic Regression and Random Forest, the use and tuning of `scale_pos_weight` in XGBoost, and the use and tuning of `pos_weight` for TabNet in the balanced training pipeline. The EDA thus directly informs the imbalance-handling strategy adopted in the modelling stage.

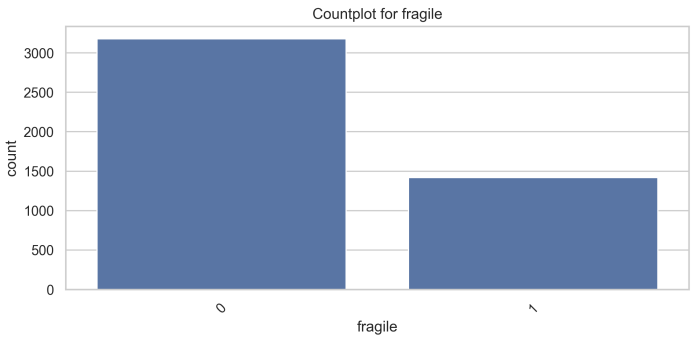


Figure 3: Class distribution of the target variable `fragile`.

Additional descriptive checks were performed to ensure the suitability of the dataset for supervised learning. Summary statistics confirmed substantial cross-sectional heterogeneity in household income, asset holdings, age, and expenditure levels, consistent with the design of the SCF. Visual inspection of distributions revealed a high degree of right skewness in wealth and income variables, which motivated the use of standardization for continuous predictors prior to model training.

Overall, the EDA stage provides a high-level understanding of the dataset and supports key methodological choices, including the removal of redundant features, the use of standardized continuous variables, and the implementation of class-imbalance mitigation techniques in the modelling pipeline.

4.3 *Constructing the Financial Fragility Target*

The definition of financial fragility adopted in this thesis builds on two strands of literature. On the one hand, Lusardi et al. (2011) and Hasler et al. (2017) conceptualise fragility as an inability to cope with relatively modest financial shocks, such as raising US$2 000 within a month. On the other hand, studies of emergency savings and liquid assets emphasise the importance of having several months of expenses in liquid form (Babiarz & Robb, 2014). Policy institutions, including the Federal Reserve, similarly evaluate financial resilience based on whether households could cover several months of expenses in an emergency (Board of Governors of the Federal Reserve System, 2022).

This thesis integrates these perspectives by defining financial fragility in objective, liquidity-based terms. Rather than asking whether households believe they could raise a certain amount, the SCF dataset allows a liquidity-based threshold to be implemented directly. The starting point is the idea that essential monthly expenses consist of basic consumption items and housing costs. In the SCF, food expenditure is reported in several components (e.g. food at home, food away from home, delivery), which are aggregated into an annual food spending variable `FOODSPEND`. Housing costs for renters are captured by the `RENT` variable. Monthly essential expenses are then defined as:

$$\text{monthly\_expenses} = \frac{\text{FOODSPEND}}{12} + \text{RENT}.$$

Liquid assets are represented by the standard SCF variable for immediately accessible financial resources. For each household, a three-month

liquidity threshold is constructed as $3 \times$ monthly_expenses. A binary indicator of financial fragility is then defined as:

$$\text{fragile} = \begin{cases} 1, & \text{if LIQ} < 3 \times \text{monthly\_expenses}, \\ 0, & \text{otherwise.} \end{cases}$$

This construct can be interpreted as a measure of *liquidity-based financial fragility*. Households that do not hold enough liquid assets to sustain three months of essential expenses are considered fragile, as they lack a buffer commonly recommended for emergency preparedness (Babiarz & Robb, 2014; Board of Governors of the Federal Reserve System, 2022). In contrast, households above this threshold are classified as non-fragile for the purposes of the model.

An important methodological consideration is the avoidance of target leakage. Because the fragility label is constructed from specific components (liquid assets, food expenditure, and rent), these variables are removed from the predictor set after the target has been created. This ensures that models cannot trivially "learn" the definition by reusing the exact inputs used to derive it, which would lead to artificially inflated performance (Kapoor & Narayanan, 2023).

4.4 *Algorithms*

This thesis evaluates four supervised classification models that differ in their flexibility, inductive biases, and suitability for tabular financial data: Logistic Regression, Random Forest, XGBoost, and TabNet. Together, these models provide a broad spectrum of representational capacity, ranging from linear decision boundaries to attention-based deep learning architectures.

Logistic Regression

Logistic Regression serves as a classical benchmark model widely used in empirical household finance. It models the log-odds of financial fragility as a linear function of the predictor variables.Although limited in its ability to capture non-linear interactions, Logistic Regression offers a transparent baseline whose coefficient structure remains widely interpretable for policy audiences.

Random Forest

Random Forests (Breiman, 2001) are ensemble methods that aggregate many decision trees trained on bootstrapped subsets of the data. By

averaging across trees, Random Forests reduce variance and improve robustness to noisy predictors, making them well suited to heterogeneous and high-dimensional survey data such as the SCF. hyperparameters such as tree depth, number of estimators, and split constraints are tuned via nested cross-validation.

XGBoost

XGBoost (T. Chen & Guestrin, 2016) is a scalable gradient boosting algorithm that constructs decision trees sequentially, with each new tree correcting errors made by previous ones. Its regularised objective, shrinkage, and handling of sparse features make it a state-of-the-art method for tabular prediction tasks.In this thesis, key hyperparameters such as learning rate, maximum depth, subsampling rates, and the class-imbalance weight (`scale_pos_weight`) are optimised through the nested cross-validation framework.

TabNet

TabNet (Arik & Pfister, 2021) is an attentive deep learning architecture explicitly designed for tabular data. The model uses sequential decision steps with feature-masking mechanisms to learn sparse, interpretable feature selection patterns. Unlike tree-based models or linear classifiers, TabNet describes non-linear interactions through attention-driven transformations, potentially capturing complex relationships between household characteristics. hyperparameters controlling the attention dimensions, number of decision steps, sparsity regularisation, and class imbalance weighting (`pos_weight`) are tuned via Optuna.

4.5 *Hyperparameter Tuning and Nested Cross-Validation*

To obtain robust and unbiased estimates of model performance, the thesis uses nested cross-validation. The outer loop consists of three stratified folds, ensuring that each fold contains a similar proportion of fragile households. For each outer split, the training data are further partitioned into three inner folds used for hyperparameter optimisation. The inner loop is implemented with Optuna, an efficient Bayesian optimisation library.

Within the modelling pipeline, continuous predictors are standardised to ensure that differences in scale do not affect optimisation or model convergence. Standardisation is performed inside each fold of the nested cross-validation procedure to prevent leakage: the mean and standard deviation are computed on the training portion of each fold and applied

only to its corresponding validation or test split. This z-score standardisation is required for models such as Logistic Regression and TabNet, which are sensitive to the relative scaling of input features. Tree-based models (Random Forest and XGBoost) do not rely on feature scaling but receive the same transformed inputs for consistency across models. Binary and dummy variables remain in their original 0/1 form.

The strict separation of training and test transformations is a central design choice. As Kapoor and Narayanan (2023) emphasise, even seemingly innocuous preprocessing steps can introduce data leakage if information from the test set is used during training. By computing all scaling parameters and model weights on training data exclusively, the pipeline avoids this problem and yields more reliable estimates of out-of-sample performance.

All models are optimised using Optuna, which performs a structured search over model-specific hyperparameter spaces within the inner loop of the nested cross-validation procedure. The search spaces are defined to balance flexibility, computational efficiency, and relevance to the model architectures under consideration. For each algorithm, Optuna samples hyperparameter configurations using a combination of Tree-structured Parzen Estimators (TPE) and pruning strategies based on early stopping.

The optimisation objective is the mean ROC-AUC score computed across the validation folds of the inner loop, ensuring that the selected hyperparameters maximise discriminatory performance while avoiding overfitting. Separate search spaces are defined for the unbalanced and class-weighted configurations of each model to reflect the different training dynamics introduced by imbalance handling.

Tables 1–4 present the full hyperparameter search spaces explored for Logistic Regression, Random Forest, XGBoost, and TabNet in both unbalanced and balanced pipelines.

Table 1: Hyperparameter search space for Logistic Regression.

Hyperparameter	Type	Search space
C (inverse regularisation strength)	log-uniform	10^{-4} – 10^{4}
class_weight	fixed	None (unbalanced) / balanced

4.6 *Class Imbalance Treatment*

Financial fragility, as defined in this thesis, is a minority condition. Meaning most households in the SCF are not classified as fragile. This imbalance poses a challenge for standard classification algorithms, which may simply learn to predict the majority class and thereby achieve deceptively high

Table 2: Hyperparameter search space for Random Forest.

Hyperparameter	Type	Search space
$n_{\text{estimators}}$	integer	200–1000
max_depth	categorical	{None, 5, 10, 15, 20}
min_samples_split	integer	2–10
min_samples_leaf	integer	1–4
max_features	categorical	{sqrt, log2, 0.3, 0.5, 0.8}
class_weight	fixed	None (unbalanced) / balanced

Table 3: Hyperparameter search space for XGBoost.

Hyperparameter	Type	Search space
learning_rate	log-uniform	0.05–0.15
max_depth	integer	3–10
min_child_weight	integer	1–10
subsample	continuous	0.5–1.0
colsample_bytree	continuous	0.5–1.0
γ(gamma)	continuous	0.0–0.5
$n_{\text{estimators}}$	integer	200–1000
scale_pos_weight(balanced)	continuous	$[1.0\, w_{\text{base}},\ 2.0\, w_{\text{base}}]$, $w_{\text{base}} = \frac{N_{neg}}{N_{pos}}$
scale_pos_weight(unbalanced)	fixed	1

Table 4: Hyperparameter search space for TabNet.

Pipeline	Hyperparameter	Type	Search space
Unbalanced & balanced	batch_size	categorical	{1024, 2048}
	mask_type	categorical	{entmax, sparsemax}
	n_d	integer	8–52 (step 2)
	n_a	integer	8–52 (step 2)
	n_{steps}	integer	2–8
	γ(gamma)	continuous	1.0–2.0
	$\lambda(lambda)_{\text{sparse}}$	log-uniform	10^{-6}–10^{-4}
	learning rate	log-uniform	10^{-3}–10^{-2}
Balanced only	pos_weight	continuous	$[1.0\, w_{base},\ 2.0\, w_{base}]$, $w_{base} = \frac{N_{neg}}{N_{pos}}$

accuracy. The broader literature on imbalanced learning shows that naive models underperform on the minority class and that specialised techniques are often required (He & Garcia, 2009; Krawczyk, 2016).

Two broad families of strategies exist: data-level methods, which modify the training distribution (e.g. oversampling, SMOTE, undersampling), and algorithm-level methods, which adjust the loss function to penalise minor-

ity class errors more heavily. In this thesis, the focus is on algorithm-level treatments for two reasons. First, they preserve the empirical distribution of the SCF, which is desirable when studying real-world prevalence of fragility. Second, they integrate naturally with the chosen models.

For Logistic Regression and Random Forest, class weights are used such that misclassifying a fragile household is penalised more than misclassifying a non-fragile household. XGBoost relies on the `scale_pos_weight` parameter, which is proportional to the ratio of non-fragile to fragile households in the training data but is refined through tuning rather than fixed. TabNet's loss function similarly incorporates a positive-class weight. Comparing the performance of models with and without these imbalance-aware configurations is a key way in which the second research sub-question, concerning the effect of class imbalance on learning dynamics and error patterns is addressed.

4.7 *Evaluation Metrics and Error Analysis*

Given the class imbalance and the substantive focus on identifying fragile households, evaluation relies on a set of complementary metrics. Overall accuracy is reported but not used as the sole criterion, because high accuracy can be obtained trivially by predicting the majority class. Instead, the thesis places greater emphasis on recall and F_1 score for the fragile class, as well as the ROC-AUC.

Recall for the fragile class measures the proportion of truly fragile households that the model correctly identifies. This is particularly important if the goal is to flag vulnerable households for further attention. Precision captures the proportion of predicted fragile households that are actually fragile, which matters for avoiding false alarms. The F_1-score balances these two quantities. ROC AUC summarises the model's ability to discriminate between fragile and non-fragile households across all possible probability thresholds.

In addition to scalar metrics, confusion matrices aggregated over outer folds are examined to understand the types of errors models make. For instance, a model may achieve high ROC-AUC but still miss a large fraction of fragile households at the chosen threshold, which would be problematic from a policy perspective.

4.8 *Interpretability Strategy*

A further methodological objective is to understand which household characteristics are most influential for predicting fragility. TabNet provides an intrinsic mechanism for this via its attention-based feature masks, which

can be aggregated to provide global importance scores for each predictor (Arik & Pfister, 2021). These scores are examined and related back to economic theory and prior empirical findings on fragility and emergency savings.

4.9 *Implementation and Reproducibility*

All analyses are conducted in Python. Data handling relies on `pandas` and `numpy`, models are implemented using `scikit-learn`, `XGBoost`, and `pytorch-tabnet`, hyperparameter tuning uses `Optuna`, and visualisations are created with `matplotlib` and `seaborn`. Random seeds are fixed where possible to enhance reproducibility. The modelling scripts are structured so that each step of the pipeline, from loading the SCF data to producing final metrics and feature importance summaries, can be rerun end-to-end, consistent with recommendations for transparent ML-based research (Kapoor & Narayanan, 2023).

Overall, the methodological framework integrates a carefully constructed target variable, a rich set of predictors, advanced machine learning models, and rigorous evaluation procedures. This provides a solid empirical basis for assessing the contribution of TabNet to predicting financial fragility and for understanding the role of class imbalance and household characteristics in shaping model performance.

5 RESULTS

This section presents the empirical findings from the predictive modelling analysis using the 2022 Survey of Consumer Finances (SCF). Four classification models, Logistic Regression, Random Forest, XGBoost, and TabNet, were evaluated under two training pipelines: (i) an unbalanced specification using the raw class distribution, and (ii) a balanced specification incorporating class-weight adjustments or tuned imbalance parameters. All results are obtained through three-fold nested cross-validation, ensuring a robust separation between hyperparameter optimisation and model evaluation. Performance is assessed using Accuracy, Precision, Recall, F1-score, and ROC-AUC, with particular emphasis on the fragile class.

5.1 *Hyperparameter Behaviour*

Hyperparameter tuning provides insight into how each model adapts to the structure of the SCF predictors. Tables 5–8 present the tuned hy-

perparameters selected by Optuna for each model with class imbalance handling.

Table 5: Tuned hyperparameters for Logistic Regression (balanced).

Hyperparameter	Fold 1	Fold 2	Fold 3
C	58.95	1099.33	227.58

Table 6: Tuned hyperparameters for Random Forest (balanced).

Hyperparameter	Fold 1	Fold 2	Fold 3
n_estimators	448	766	914
max_depth	8	5	6
min_samples_split	3	3	4
min_samples_leaf	3	3	3
max_features	log2	log2	sqrt

Table 7: Tuned hyperparameters for XGBoost (balanced).

Hyperparameter	Fold 1	Fold 2	Fold 3
n_estimators	441	216	212
max_depth	3	3	7
learning_rate	0.0744	0.0608	0.0609
min_child_weight	1.0	10.0	10.0
subsample	0.8495	0.7769	0.5014
colsample_bytree	0.5039	0.6477	0.4168
gamma	0.3197	0.2682	0.4168
scale_pos_weight	2.2413	2.2413	2.2389

These hyperparameters show that ensemble methods prefer moderately deep and well-regularised trees, while TabNet benefits from larger decision/attention dimensions and carefully chosen sparsity penalties. The tuned `pos_weight` values in the balanced TabNet model align closely with the true class imbalance ratio, reinforcing the importance of cost-sensitive loss functions for deep learning on tabular data. Additionally, the hyperparameters for the algorithms without class imbalance handling are presented in Appendix B (page 34).

Table 8: Tuned hyperparameters for TabNet (balanced).

Hyperparameter	Fold 1	Fold 2	Fold 3
batch_size	1024	2048	1024
mask_type	entmax	entmax	entmax
n_d	36	44	44
n_a	30	40	14
n_steps	6	6	7
gamma	1.263	1.201	1.093
lambda_sparse	3.91e-05	1.19e-06	6.29e-05
learning_rate	0.00986	0.00399	0.00404
pos_weight	3.79	3.38	4.19

5.2 *Overall Predictive Performance*

Tables 9 and 10 summarise the mean outer-fold performance of all models. Because financial fragility is a minority class, threshold-based metrics such as Recall and F1-score play a critical role in evaluating each model's suitability for identifying vulnerable households.

Table 9: Model performance without imbalance handling (nested CV means).

Model	Accuracy	Precision	Recall	F1	ROC-AUC
Logistic Regression	0.8474	0.7531	0.7525	0.7528	0.9163
Random Forest	0.8607	0.7856	0.7546	0.7698	0.9285
XGBoost	**0.8672**	**0.7887**	**0.7786**	**0.7836**	**0.9305**
TabNet	0.7952	0.7613	0.4894	0.5958	0.8727

Without imbalance handling, XGBoost achieves the highest performance across all metrics. Logistic Regression and Random Forest exhibit comparable Recall and Precision, reflecting a balanced decision boundary. In contrast, TabNet shows low Recall (0.489), indicating that it systematically fails to identify fragile households under the raw class distribution. Its ROC-AUC (0.873), however, suggests that the model still captures useful risk-ordering information, even if the threshold classification is poor.

When imbalance handling is applied, Recall improves substantially across all models. In TabNet's case, Recall increases by nearly 26 percentage points, from 0.489 to 0.749, illustrating that the model is highly sensitive to class imbalance. XGBoost remains the strongest performer overall, achieving the highest combined F1-score and ROC-AUC in both unbalanced and balanced settings. Additionally, a detailed breakdown

Table 10: Model performance with imbalance handling (nested CV means).

Model	Accuracy	Precision	Recall	F1	ROC-AUC
Logistic Regression	0.8337	0.6779	**0.8787**	0.7654	0.9152
Random Forest	0.8529	**0.7280**	0.8357	0.7781	0.9287
XGBoost	**0.8546**	0.7269	0.8477	**0.7825**	**0.9292**
TabNet	0.8039	0.6646	0.7490	0.6994	0.8739

of misclassification patterns for each model is available in Appendix A (page 33)., where confusion matrices for both balanced and unbalanced pipelines are presented.

5.3 *Impact of Class Imbalance on Learning Dynamics*

The impact of class imbalance is particularly pronounced in this classification task because fragile households form a clear minority. Prior research shows that imbalanced datasets can distort learning dynamics, leading to majority-class bias and low sensitivity to rare events (He & Garcia, 2009; Krawczyk, 2016). Our findings follow this pattern.

First, all unbalanced models exhibit lower Recall than their balanced counterparts. Logistic Regression's Recall improves from 0.752 to 0.879 after weighting, Random Forest from 0.755 to 0.836, and XGBoost from 0.779 to 0.848. These improvements demonstrate that class weights allow the models to better detect minority-class examples.

Second, Precision decreases in the balanced setting for all models. This reflects the well-known trade-off between detecting more minority instances and falsely flagging non-fragile households as fragile.

Third, ROC-AUC remains relatively stable across balanced and unbalanced training. This suggests that the models' underlying ability to rank households by fragility risk is robust to class imbalance; the main effect lies in the threshold choice.

The strongest imbalance effects occur in TabNet, reflecting findings from the deep learning literature: neural networks often underperform on tabular data without explicit imbalance controls (Shwartz-Ziv & Armon, 2022). Balancing ensures that TabNet receives stronger training signals from fragile households, allowing its attention modules to form more informative feature masks.

5.4 *Model-by-Model Interpretation*

Logistic Regression

Logistic Regression performs surprisingly well given its linear functional form. In the unbalanced setting, the model achieves balanced Precision and Recall (≈ 0.75). After accounting for class imbalance, Recall increases dramatically to 0.879, the highest among all models. This indicates that the linear separation between fragile and non-fragile households becomes more apparent when the cost of minority misclassification is explicitly increased. Similar behaviour is observed in credit-risk contexts, where logistic regression improves significantly when class costs are adjusted (Liu et al., 2022).

Random Forest

Random Forest achieves improved performance relative to Logistic Regression due to its ability to model non-linear interactions between income, demographics, and financial characteristics. Its Recall improves from 0.755 to 0.836 after balancing, while F1 increases from 0.770 to 0.778. Tree-based ensembles are known for their robustness to noisy or heterogeneous predictors (Addo et al., 2018), which likely contributes to the model's consistent performance across folds.

XGBoost

XGBoost consistently achieves the best performance across both pipelines, especially in terms of F1-score and ROC-AUC. Its strong performance aligns with the broad consensus that gradient boosting is the state-of-the-art for tabular financial data (Addo et al., 2018; Albanesi & Vamossy, 2019). After applying tuned `scale_pos_weight` values (≈ 2.24), Recall improves further, demonstrating that cost-sensitive boosting effectively handles class imbalance.

TabNet

TabNet's unbalanced performance is the weakest among the models, with Recall of 0.489. This result aligns with prior work showing that deep learning architectures can perform poorly on tabular datasets without tailored regularisation and balanced training signals (Shwartz-Ziv & Armon, 2022). After tuning `pos_weight`, TabNet's Recall improves dramatically to 0.749, and its F1-score increases by more than 10 percentage points. While TabNet does not surpass XGBoost, it provides valuable interpretability through its

attention-based masks, consistent with findings in Arik and Pfister (2021) and Borisov et al. (2021).

5.5 *TabNet Feature Importance*

Figure 4 reports the mean TabNet feature importance scores, averaged across the three outer folds of the nested cross-validation procedure. The results show a clear and consistent ordering of the top ten predictor contributions.

Across folds, HOUSES exhibits the highest mean importance value, substantially exceeding the contributions of all other variables. The second most important feature is WAGEINC, followed by RETQLIQ, which forms the third-largest contribution in the aggregate importance ranking.

Several additional variables show moderate but non-negligible importance scores. These include RACE_2.0, EMERGSAV, and VEHIC, all of which appear within the upper half of the ranked features. Lower but still measurable contributions are observed for EMERGBORR, SPENDMOR_3.0, MARRIED, and KNOWL_8.0. Together, these ten variables constitute the highest ranked set in the overall importance distribution produced by TabNet.

The remaining predictors exhibit substantially smaller importance values and do not contribute meaningfully to the aggregated ranking. Their relative magnitudes remain close to zero, indicating limited involvement in the model's attention mechanism across folds.

Overall, the mean importance distribution displays a steep decline from the top-ranked variables to the remainder of the feature set, reflecting a highly concentrated allocation of feature weight within TabNet's learned masks. The full ranking and importance values are visualised in Figure 4.

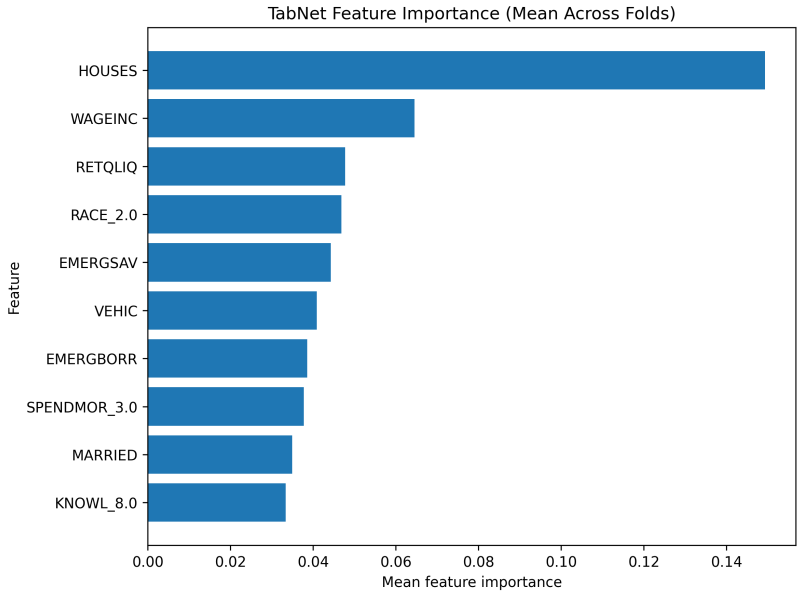


Figure 4: Mean TabNet feature importance across outer folds.

These findings directly address the research questions by showing (i) how TabNet compares to traditional models, (ii) what characteristics are most influential in predicting financial fragility, and (ii) how class imbalance shapes model dynamics and errors.

6 DISCUSSION

This chapter interprets the empirical findings presented in Section 5 and situates them within the broader literature on financial fragility, credit-risk modelling, and machine learning for tabular data. The discussion is organised around the three research questions examining: (i) whether TabNet improves predictive performance relative to traditional models, (ii) which household characteristics are most influential for predicting fragility, and (iii) how class imbalance shapes learning dynamics. The chapter concludes with implications for research and policy.

6.1 *MRQ: Can TabNet Improve the Prediction of Financial Fragility?*

The main research question addressed whether TabNet provides predictive benefits beyond traditional models. The results show that TabNet does not surpass the strongest baseline, XGBoost, in either the unbalanced or balanced setting. XGBoost consistently achieves the highest ROC-AUC, Recall, and F1-scores, confirming prior evidence that boosted tree ensembles

remain state-of-the-art for structured tabular data, particularly in financial applications (Addo et al., 2018; Albanesi & Vamossy, 2019; Liu et al., 2022).

These findings are consistent with insights from Grinsztajn et al. (2022), who argue that deep learning architectures often underperform on tabular data because such data lack the smooth hierarchical structure that neural networks are designed to exploit. Instead, tree-based models benefit from inductive biases that align more closely with the heterogeneous, noisy, and weakly structured nature of socio-economic variables such as those in the SCF.

TabNet performs particularly poorly in the unbalanced setting, detecting fewer than half of fragile households. This behaviour is expected: deep neural networks are known to struggle when minority-class gradients are insufficiently represented during training (Shwartz-Ziv & Armon, 2022). After applying a tuned `pos_weight`, TabNet's Recall improves substantially, and its performance becomes broadly comparable to Logistic Regression and Random Forest. Still, even with these adjustments, TabNet does not surpass XGBoost and exhibits greater variability across folds.

Therefore, it cannot be said that TabNet improves predictive accuracy over traditional models in this application.

6.2 *SRQ1: Which Household Characteristics Are Most Influential According to TabNet?*

The first sub research question examined which features TabNet identifies as most relevant through its attention-based feature masks. Despite differences in performance between the balanced and unbalanced settings, the importance rankings are remarkably stable across folds.

Housing wealth emerges consistently as the single most influential predictor. This is fully consistent with household finance research showing that home ownership and real estate values are strong indicators of financial resilience and long-term wealth accumulation. Wage income also plays a central role, reflecting its importance as a stable source of liquidity for meeting short-term expenses. Retirement liquidity, captured by balances in pension and retirement accounts, becomes particularly influential for older households, whose resilience increasingly depends on these assets rather than labour income.

Emergency savings and borrowing behaviours, such as the ability to handle unexpected expenses or the need to rely on borrowing in emergencies, also feature prominently. These variables align directly with the liquidity-based definition of fragility and indicate whether a household has the buffer needed to withstand short-term shocks. Education categories, particularly higher levels of attainment, consistently appear as influential

predictors as well, in line with research linking education to financial literacy, planning ability, and overall financial resilience (Allgood & Walstad, 2021; Hasler et al., 2017). Finally, indicators of financial strain, such as repeated late payments signal existing stress and are naturally correlated with fragility.

Taken together, these results show that TabNet identifies a coherent and economically intuitive set of characteristics. Despite its lower predictive accuracy, TabNet offers reliable interpretability outputs that reinforce established findings from the fragility and financial literacy literature (Babiarz & Robb, 2014; Lusardi et al., 2011). In response to SRQ1, the most influential predictors of fragility are housing wealth, wage income, emergency savings capacity, retirement liquidity, educational attainment, and signs of financial hardship.

6.3 *SRQ2: How Does Class Imbalance Affect Learning Dynamics and Errors?*

Class imbalance plays a central role in shaping model behaviour in this study. Without imbalance correction, all models exhibit substantial reductions in Recall, systematically under-identifying fragile households. This mirrors well-established findings in the imbalance literature: unless explicitly penalised, classifiers tend to prioritise the majority class, leading to poor detection of rare but important outcomes (He & Garcia, 2009; Krawczyk, 2016).

Tree-based models and logistic regression all show marked improvements in Recall when class weighting is applied. These improvements demonstrate that the underlying structure of fragility is sufficiently learnable, provided that minority cases receive adequate weight in the loss function. TabNet is the most sensitive to imbalance: without `pos_weight`, the model fails to detect more than half of fragile households; with it, its Recall increases by over 25 percentage points. This confirms the view that deep learning models require explicit mechanisms to overcome skewed gradient signals (Shwartz-Ziv & Armon, 2022).

Interestingly, ROC-AUC remains largely stable across balanced and unbalanced variants, indicating that class imbalance affects threshold metrics (Precision, Recall, F1) more than ranking ability. This phenomenon is well documented in boosting literature, where class weighting improves decision thresholds but does not significantly alter ranking performance (Addo et al., 2018).

In response to SRQ2, class imbalance substantially affects all models by suppressing minority-class sensitivity, with TabNet being disproportionately affected. Proper imbalance handling is therefore essential for meaningful prediction of financial fragility.

6.4 *Implications for Research and Policy*

The findings carry several implications. Methodologically, the results confirm that boosted tree ensembles, particularly XGBoost, remain the most effective models for structured financial data. While TabNet can challenge their performance when appropriately re-weighted, it does not surpass them. Its primary contribution lies in generating interpretable feature importance patterns, suggesting that hybrid pipelines combining boosting for prediction and TabNet for interpretability may be a productive direction for future research.

From a policy perspective, the findings highlight the need to prioritise Recall in fragility prediction. False negatives, fragile households incorrectly classified as resilient, carry greater social and economic costs than false positives. Class imbalance handling is thus not only a technical choice but also a normative one, ensuring that vulnerable households are not systematically overlooked. The features that consistently emerge as influential, emergency savings, income stability, and housing assets, offer further guidance on where policy interventions may have the greatest impact.

Overall, the results suggest that while deep learning contributes interpretability benefits, tree-based ensemble methods remain the most reliable tools for predicting liquidity-based financial fragility.

6.5 *Limitations and Future Research*

Although this study provides a systematic assessment of machine learning approaches for predicting liquidity-based financial fragility, several limitations should be acknowledged. First, the analysis relies on a single cross-sectional wave of the Survey of Consumer Finances. Without longitudinal information, the models can identify correlates of fragility but cannot capture transitions into or out of vulnerable states.

Second, the fragility definition employed here focuses on a three-month liquidity buffer. While widely used in the emergency savings literature, it does not incorporate other dimensions of vulnerability such as debt service burdens, income instability, or access to credit. Alternative definitions could yield different predictive patterns and may offer a more holistic view of household financial resilience. Exploring multiple fragility definitions in a unified modelling framework represents a promising extension.

Third, to ensure a leakage-free design, a substantial number of SCF variables were excluded because they mechanically overlapped with components of the fragility definition. This necessary restriction reduces the feature space and may limit the ability of complex models, especially deep learning architectures, to detect nuanced patterns. Future work could

pursue more advanced feature engineering, non-linear transformations, or latent representations to enrich the input space without introducing leakage.

Fourth, TabNet's underperformance may reflect not only its architectural limitations but also the relatively modest sample size of the SCF and the sparsity introduced by one-hot encoding. Deep learning models often require larger datasets or alternative regularisation strategies to realise their advantages. Evaluating more recent neural architectures for tabular data, such as FT-Transformer, SAINT, or NODE, could provide further insight into whether deep learning can match or exceed boosting models in household finance applications.

Finally, the present evaluation focuses on aggregate model performance and does not examine fairness or subgroup differences. Understanding whether prediction errors vary systematically across demographic groups is crucial for policy applications, particularly when identifying financially vulnerable households. Future studies should incorporate fairness metrics, distributional evaluation, and robustness checks under different sampling schemes and economic conditions.

Overall, these limitations highlight opportunities for methodological refinement and motivate further research aimed at improving the robustness, fairness, and interpretability of machine learning models for financial fragility prediction.

7 CONCLUSION

This thesis set out to evaluate whether modern deep learning architectures, specifically TabNet, can improve the prediction of liquidity-based financial fragility relative to established machine learning models. Using data from the 2022 Survey of Consumer Finances, the study implemented a rigorous leakage-free modelling pipeline with nested cross-validation, systematic preprocessing, and explicit treatment of class imbalance. Beyond predictive performance, the analysis also focused on interpretability through TabNet's attention-based feature masks and on understanding how class imbalance influences learning dynamics.

Several conclusions emerge from the findings. First, TabNet does not outperform the traditional models examined. XGBoost consistently achieves the strongest predictive performance in both unbalanced and class-weighted settings, followed by Random Forest and Logistic Regression. These results align with existing evidence that boosted tree ensembles remain the most effective methods for tabular financial data. While TabNet improves considerably when class imbalance is addressed, its performance remains below that of XGBoost and exhibits greater variability across folds.

Thus, the primary research question is answered in the negative: TabNet does not enhance predictive accuracy in this context.

Second, the interpretability analysis shows that TabNet identifies a coherent and economically intuitive set of predictors. Housing wealth, wage income, emergency savings capacity, retirement liquidity, and educational attainment consistently emerge as the most important characteristics associated with financial fragility. These results closely match established findings in the household finance literature, underscoring the relevance and stability of the underlying mechanisms captured by the model.

Third, the study demonstrates that class imbalance plays a crucial role in model performance. Without class weighting, all models, especially TabNet, systematically under identify fragile households. Applying class weights or imbalance-aware parameters substantially improves Recall and yields a more equitable distribution of errors. This highlights the importance of careful imbalance handling for any predictive system designed to support policy interventions targeted at financially vulnerable populations.

Overall, this thesis contributes to the growing literature on machine learning for household finance by providing a transparent and methodologically rigorous comparison of modern and traditional models. Although deep learning does not surpass tree-based methods in predictive accuracy, its interpretability features offer valuable insights into the drivers of financial fragility. The findings emphasise the importance of model transparency, robust validation, and careful treatment of class imbalance when developing tools for identifying vulnerable households. Together with the limitations and future research directions outlined earlier, this work provides a foundation for continued methodological innovation and improved understanding of financial fragility within the U.S. population.

REFERENCES

Addo, P. M., Guegan, D., & Hassani, B. (2018). Credit risk analysis using machine and deep learning models. *Risks*, *6*(2), 38. https://doi.org/10.3390/risks6020038

Aladangady, A., Bricker, J., Chang, A., et al. (2023). *Changes in u.s. family finances from 2019 to 2022: Evidence from the scf* (tech. rep.). Federal Reserve Board. https://doi.org/10.17016/8799

Albanesi, S., & Vamossy, J. (2019). *Predicting consumer default: A deep learning approach* (tech. rep. No. 26165). NBER. https://doi.org/10.3386/w26165

Allgood, S., & Walstad, W. B. (2021). The effects of perceived and actual financial literacy on financial behaviors. *Economic Inquiry*, *54*(1), 675–697. https://doi.org/10.1111/ecin.12255

Arik, S. O., & Pfister, T. (2021). Tabnet: Attentive interpretable tabular learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, *33*(1), 6679–6687. https://doi.org/10.1609/aaai.v35i8.16826

Babiarz, P., & Robb, C. A. (2014). Financial literacy and emergency savings. *Journal of Family and Economic Issues*, *35*(1), 40–50. https://doi.org/10.1007/s10834-013-9369-9

Board of Governors of the Federal Reserve System. (2022). *Economic well-being of u.s. households in 2021: Dealing with unexpected expenses* (tech. rep.). Federal Reserve Board. https://www.federalreserve.gov/publications/2022-economic-well-being-of-us-households-in-2021-dealing-with-unexpected-expenses.htm

Board of Governors of the Federal Reserve System. (2023). *Survey of consumer finances (scf), 2022*. Federal Reserve Board. https://www.federalreserve.gov/econres/scfindex.htm

Borisov, V., Leemann, T., Seßler, K., Haug, J., Pawelczyk, M., & Kasneci, G. (2021). Deep neural networks and tabular data: A survey. *arXiv preprint*. https://doi.org/10.48550/arXiv.2110.01889

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Brown, I., & Mues, C. (2012). An experimental comparison of classification algorithms for imbalanced credit scoring data sets. *Expert Systems with Applications*, *39*(3), 3446–3453. https://doi.org/10.1016/j.eswa.2011.09.033

Chen, S., et al. (2020). Financial distress prediction using machine learning techniques. *Asian Journal of Economics, Business and Accounting*.

Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794. https://doi.org/10.1145/2939672.2939785

de Waal, H., et al. (2023). Consumers' financial distress: Prediction and prescription using interpretable ml. *Journal of Consumer Affairs*. https://doi.org/10.1007/s10796-024-10501-1

Grinsztajn, L., Oyallon, E., & Varoquaux, G. (2022). Why do tree-based models still outperform deep learning on tabular data? *arXiv preprint arXiv:2207.08815*. https://doi.org/10.48550/arXiv.2207.08815

Hasler, A., Lusardi, A., & Oggero, N. (2017). Financial fragility in the us: Evidence and implications. *Global Financial Literacy Excellence Center Working Paper*.

He, H., & Garcia, E. (2009). Learning from imbalanced data. *IEEE TKDE*, *21*(9), 1263–1284. https://doi.org/10.1109/TKDE.2008.239

Kanász, R., Drotár, P., Gnip, P., & Zoričák, M. (2024). Clash of titans on imbalanced data: Tabnet vs xgboost. *2024 IEEE Conference on*

Artificial Intelligence (CAI), 320–325. https://doi.org/10.1109/CAI59869.2024.00068

Kapoor, S., & Narayanan, A. (2023). Leakage and the reproducibility crisis in ml-based science. *Patterns*, *4*(10), 100804. https://doi.org/10.48550/arXiv.2207.07048

Krawczyk, B. (2016). Learning from imbalanced data: Open challenges and future directions. *Progress in Artificial Intelligence*, *5*(4), 221–232. https://doi.org/10.1007/s13748-016-0094-0

Liu, Y., Yang, M., Wang, Y., Li, Y., Xiong, T., & Li, A. (2022). Applying machine learning algorithms to predict default risk. *Journal of International Financial Markets, Institutions and Money*, *78*, 101526. https://doi.org/10.1016/j.irfa.2021.101971

Lundberg, S., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems (NeurIPS)*.

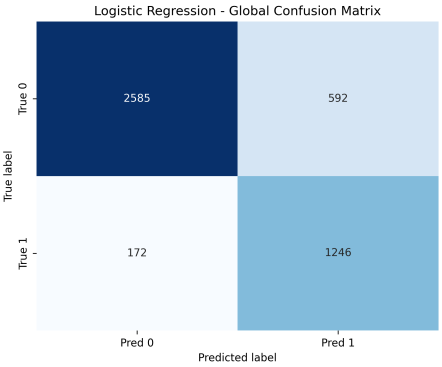
Lusardi, A., Schneider, D., & Tufano, P. (2011). Financially fragile households: Evidence and implications. *Brookings Papers on Economic Activity*, *2011*(1), 83–134. https://doi.org/10.1353/eca.2011.0002

OpenAI. (2025). Chatgpt (version 5.1) [Language model used for debugging and editing text].

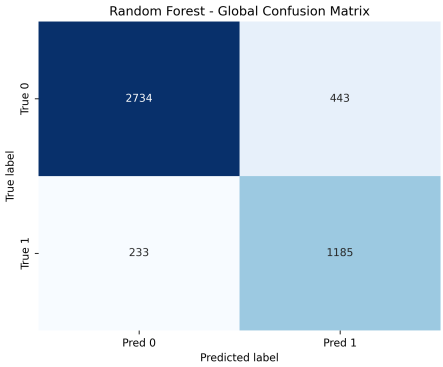
Qiu, J., Jin, Y., Meng, S., et al. (2024). Advanced user credit risk prediction model using lightgbm, xgboost and tabnet with smoteenn. *2024 IEEE 4th International Conference on Power, Intelligent Computing and Systems (ICPICS)*, 876–883. https://doi.org/10.1109/ICPICS62053.2024.10796247

Shwartz-Ziv, R., & Armon, A. (2022). Tabular data: Deep learning is not all you need. *Information Fusion*, *81*, 84–90. https://doi.org/10.1016/j.inffus.2021.11.011

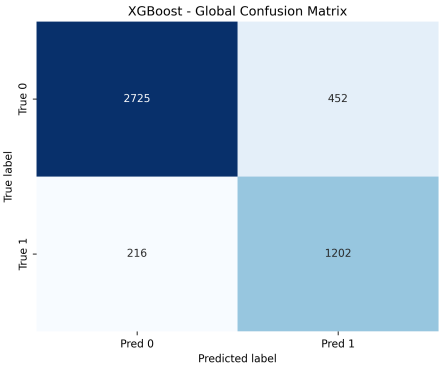
APPENDIX A: CONFUSION MATRICES



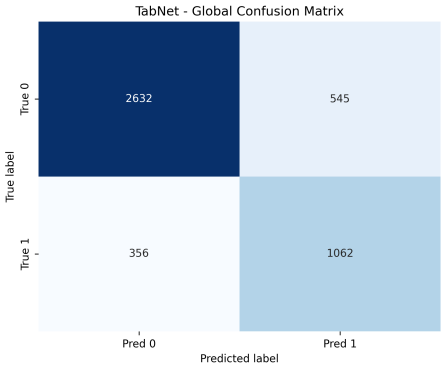
(a) Logistic Regression Confusion Matrix



(b) Random Forest Confusion Matrix

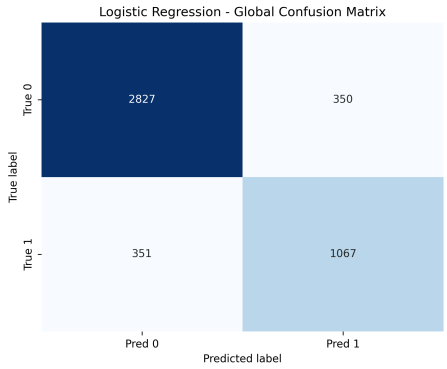


(c) XGBoost Confusion Matrix

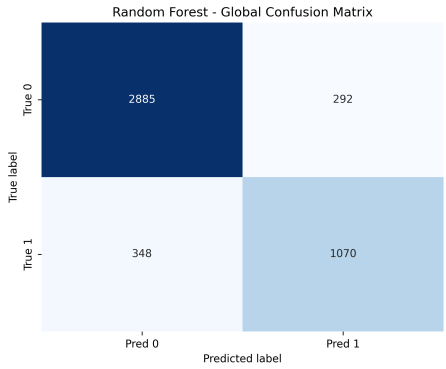


(d) TabNet Confusion Matrix

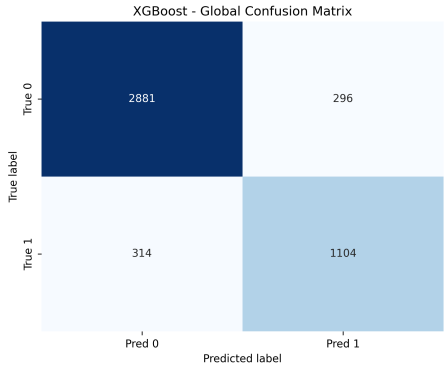
Figure 5: Confusion Matrices with Class Imbalance handling



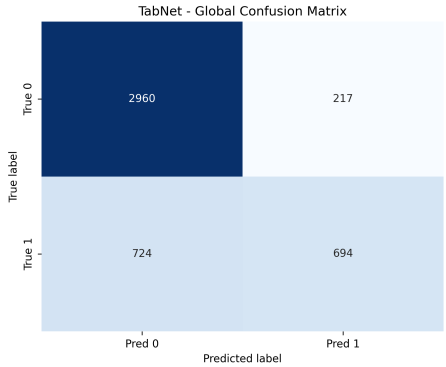
(a) Logistic Regression Confusion Matrix



(b) Random Forest Confusion Matrix



(c) XGBoost Confusion Matrix



(d) TabNet Confusion Matrix

Figure 6: Confusion Matrices without Class Imbalance handling

APPENDIX B: HYPERPARAMETERS WITHOUT CLASS IMBALANCE HANDLING

Table 11: Tuned hyperparameters for Logistic Regression (unbalanced).

Hyperparameter	Fold 1	Fold 2	Fold 3
C	52.03	1335.27	116.16

Table 12: Tuned hyperparameters for Random Forest (unbalanced).

Hyperparameter	Fold 1	Fold 2	Fold 3
n_estimators	353	404	338
max_depth	10	20	15
min_samples_split	7	2	8
min_samples_leaf	1	1	4
max_features	log2	log2	log2

Table 13: Tuned hyperparameters for XGBoost (unbalanced).

Hyperparameter	Fold 1	Fold 2	Fold 3
n_estimators	256	203	204
max_depth	3	4	8
learning_rate	0.0542	0.0569	0.0721
min_child_weight	4	1	5
subsample	0.9309	0.7872	0.8787
colsample_bytree	0.5431	0.8746	0.5534
gamma	0.2088	0.2496	0.1033

Table 14: Tuned hyperparameters for TabNet (unbalanced).

Hyperparameter	Fold 1	Fold 2	Fold 3
batch_size	1024	1024	1024
mask_type	sparsemax	entmax	entmax
n_d	26	16	32
n_a	52	24	34
n_steps	5	6	7
gamma	1.029	1.018	1.004
lambda_sparse	7.58e-06	2.25e-06	8.82e-06
learning_rate	0.00691	0.00447	0.00298