How Predictable is Household Financial Distress?

Elyas Saif

Abstract

I examine the ex-ante predictability of consumer financial distress (CFD) using nationally representative Finnish household-level data for the period 2008–2017. Unlike studies that focus on consumer default, this predictive study uses machine learning methods (ML) to analyse the predictability of CFD from the perspective of the consumer rather than the lenders. This study provides insights into the overall accuracy of ML in predicting CFD, revealing significant variability in predictability between households with and without prior experience of distress. Specifically, I find that predictability is nearly perfect for households with prior distress but significantly lower for those without. This demonstrates that predictions of CFD can be used to accurately target support measures to households most likely to experience persistent consumer financial distress. The findings of this study hence contribute to the literature on consumer financial protection — which focuses on causally identifying measures that mitigate suboptimal borrowing decisions and consumer financial distress — by considering a predictive dimension that complements this causal focus. With more accurate predictions of CFD, consumer protection measures, such as advice on how to cope with CFD, can be targeted to households that are at greatest risk.

1 INTRODUCTION

In standard economic models of intertemporal choice, rational and forward-looking consumers make consumption and saving choices to smooth their consumption over time (for an overview of these models see Attanasio and Weber, 2010). It follows from these models that in expectation, consumers should not borrow at a suboptimal level. However, deviations from optimal borrowing decisions that are linked to financial distress – also known as over-indebtedness or financial hardship – have been documented to happen in several different decision contexts (see, e.g., Campbell et al., 2011; Hyytinen and Putkuri, 2018; Parise and Peijnenburg, 2019). Taking this as the starting point, this paper uses unique data on consumers' financial distress (CFD) available for a large sample of Finnish households to study how predictable CFD is in advance (ex-ante or forward-looking) and whether the predictability of CFD differs for different groups of consumers.

In the current literature, deviations from optimal borrowing decisions are attributed to cognitive factors, such as biases associated with lack of financial literacy (Stango and Zinman, 2009; Campbell et al., 2011; Lusardi and Tufano, 2015) and optimistic errors in forecasting own future financial situation (Hyytinen and Putkuri, 2018); behavioural biases and limitations, such as matching own consumption to that of higher income peers (Georgarakos, Haliassos, and Pasini, 2014; Agarwal, Mikhed, and Scholnick, 2020), exerting low effort in examining the interest rates offered by various lenders when choosing a loan (Galenianos and Gavazza, 2022), and impulsivity (Ottaviani and Vandone, 2011; Gathergood, 2012). The behavioural and cognitive biases imply that consumers can ex-ante borrow sub-optimally and accumulate unsustainable levels of debt, thus increasing their likelihood to experience financial distress subsequently. In addition to the behavioural and cognitive biases, several studies show that non-cognitive personality traits, such as emotional instability, low conscientiousness (Xu et al., 2015; Parise and Peijnenburg, 2019), and low self-efficacy (Kuhnen and Melzer, 2018), are strongly linked to experiencing financial distress. Furthermore, consumers that exhibit certain traits, such as poor financial literacy, impulsivity, or lower non-cognitive ability, are more likely to experience ex-post financial shocks, including job loss, reduction in the level of income, and unexpected expenses (Gathergood, 2012; Parise and Peijnenburg, 2019). Therefore, the literature suggests that consumer financial distress could be related to both various ex-ante traits as well as to ex-post (financial) shocks.

The presence of biases and traits associated with financial distress provide a rationale for considering how consumers could be supported in avoiding suboptimal financial decisions and mitigate the consequences of consumer financial distress. One potential protective measure is to make financial education widely accessible to consumers. Using a natural experiment involving US army enlistees, Skimmyhorn (2016) shows that offering a personal financial management course reduces credit account balance and delinquencies. On the other hand, Campbell et al. (2011) proposes an agenda for consumer financial protection regulation, such as enforcing disclosure of borrowing costs in a manner that would make it easy for the average consumer to understand all the costs associated with

taking a loan. This proposition is supported by studies that have experimentally examined policy measures that could mitigate or debias the specific cognitive traits associated with sub-optimal borrowing and financial distress. For example, Bertrand and Morse (2009) use an experimental setting which incorporates the idea that payday borrowing is likely in-part a result of poor financial literacy (a cognitive limitation). They find that disclosing payday loan repayment fees in dollar terms instead of annual percentage rates (APR) is more effective in reducing payday loan borrowing. Similarly, Stango and Zinman (2011) showed that obscuring the APR was associated with higher borrowing costs among consumers with lower financial literacy. In addition, several studies have reported that external financial advice is likely to mitigate suboptimal borrowing decisions associated with consumer low financial literacy (Stango and Zinman, 2009; Disney, Gathergood, and Weber, 2015; Kondratjeva et al., 2021). Finally, Galenianos and Gavazza (2022) suggest using interest rate caps as a potential protective measure.

Despite the relatively large and still growing literature on consumer financial protection measures, there is limited research on the ex-ante predictability of CFD.[1] While several studies explored the predictability of consumer delinquency from the lenders' perspective (e.g., Albanesi and Vamossy, 2019; Fuster et al., 2022), to the best of my knowledge, only Salganik et al. (2019, 2020) have investigated the predictability of financial distress from consumers' perspective. They find that more sophisticated ML models with numerous predictors only moderately outperform a benchmark regression model that includes: i) only a lagged predictor of financial distress and ii) a lagged predictor of financial distress with a few additional predictors.

Hence, there is a clear need for further research along these lines, considering that accurate predictions of consumer financial distress may assist consumers in making important household economic decisions, such as borrowing and savings (Lusardi and Mitchell, 2014). For example, consumers can have imprecise probabilistic expectations of their individual likelihood of falling into financial distress. Hence, having access to a tool that would more accurately predict individuals' own likelihood of financial distress based on their observed covariates might allow consumers to avoid making sub-optimal borrowing decisions. The rationale for having such a tool parallels the need to predict individual risk of dementia, as discussed by Giustinelli, Manski, and Molinari (2022, p. 188). In this context too, individuals lack access to accurate probabilistic predictions of their dementia risk, and as such might rely on imprecise probabilistic expectations when making important economic decisions, such as retirement planning and purchase of long-term care insurance (Giustinelli, Manski, and Molinari, 2022, p. 211). In addition, accurate ex-ante predictions of CDF could, in theory at least, be used to target potentially scarce

[1]Previous studies on consumer financial protection made use of standard regression approaches, often utilising experimental settings to establish causality with the aim to suggest regulatory policy measures to reduce consumer suboptimal borrowing.

public financial counselling and debt restructuring advice to those most in need of it to help them cope with it.[2]

A similar rationale to predicting consumer financial distress can be seen in the literature on the predictability of consumer default, where default is defined as the probability of a 90+ days delinquency on a loan in the next 24 to 36 months (Albanesi and Vamossy, 2019; Netzer, Lemaire, and Herzenstein, 2019; Fuster et al., 2022). The primary purpose of these studies is to investigate how machine learning (ML) methods can improve the ability of lenders to discern borrowers' likelihood of default when making loan decisions. Improvements in the predictability of default is important from a societal point of view because it reduces the quantity of non-performing-loans issued by lenders (reduces financial losses) and minimises the negative economic consequences of consumer default (foreclosures) on borrowers and the wider economy. Nevertheless, improvements in consumer default prediction do not necessarily address the potential struggles that the financially distressed but non-defaulting consumers experience, such as the disruption to their established consumption pattern and standard of living that is brought upon by costly loan servicing and debt rollover fees (Betti et al., 2007; Athreya, Mustre-del-Río, and Sánchez, 2019). This is the novel viewpoint taken in this study.

When studying how predictable CFD is ex-ante, it is natural to use predictive ML methods (see Varian, 2014; Efron and Hastie, 2016; Athey and Imbens, 2019). The goal of such ML methods is to find a function that provides a good out-of-sample prediction of an outcome, such as the event of a consumer experiencing financial distress, as a function of predictors, such as the consumer's indebtedness, various socio-economic indicators and possibly his/her prior financial distress. Once an ML model is initially fitted to one part of the data, called the training set, its predictive accuracy is then typically evaluated on previously unseen data, called the test or hold-out set. The primary aim of the predictive ML methods is to predict unknown (and often future) outcomes as accurately as possible, without focusing on causal inferences (Hofman et al., 2021, pp. 183–184). However, some studies go beyond a mere statistical prediction and consider the economic implications of the accuracy of ML predictions in the context of specific decisions; these types of studies can be said to analyse *prediction policy problems* (see for example Kleinberg et al., 2015, 2018; Hyytinen et al., 2023). In such studies, a decision-theoretic framework with an uncertain outcome is considered and predictions are an input to the decision-making process. For example, in some of the studies, the aim is to minimise (maximise) the expected loss (utility) associated with a decision, while accounting for the costs and benefits associated with each decision-outcome combination (Lahiri and Yang, 2013, p. 1026; Elliot and Timmermann, 2016, p. 15). Many of these studies focus on evaluating the economic value of predictions for individual decision-making, rather than policy decisions (Granger and Pesaran, 2000; Granger and Machina, 2006; Lieli and White, 2010; Lieli and Springborn, 2013; Lahiri and Yang, 2013; Elliot and Timmermann, 2016). Finally, a more recent area of research examines the distributional impacts of predictions on dif-

[2] Public or non-profit debt and financial counselling services exist in many countries. For instance, there is the Guarantee Foundation in Finland or the National Foundation for Credit Counselling in the USA.

ferent population groups. For instance, Fuster et al. (2022) investigate the economic disparities between different racial groups arising from the use of ML models on credit markets.

Building on these strands of literature, I use ML methods to study how predictable CFD is ex-ante and explore how the accuracy of ML predictions differs between different population groups. More specifically, I address two main research questions in this study using several unique binary measures for CFD that are available from the Finnish Income Distribution Statistics (IDS) over 2008-2017. Taking each of them in turn:

For the first research question (RQ1), I examine how accurately various ML models can ex-ante predict CFD on average and whether the predictability of CFD vary depending on how it is defined. I also explore whether there is heterogeneity in the predictability of CFD, particularly between the sub-sample of households that have experienced prior CFD and those that have not. This comparison is motivated by the findings of Brown, Ghosh, and Taylor (2014), who find that CFD persists from one time-period to future periods. No prior study has examined this, i.e., the possible heterogeneity in predictability of CFD between the aforementioned subsamples. Yet, such a comparison has relevance for both households' borrowing decisions as well as possible policies designed to reduce or mitigate CFD.

For the second research question (RQ2), I investigate in more detail whether CFD can be predicted with higher accuracy for the subset of households that are at greatest risk of experiencing it; see Hyytinen et al. (2023) who consider this type of tail predictability in a different context. I also consider if there are notable differences in this tail predictability between the households with prior experience of CFD and those without? For the purposes of financial distress prediction, this approach could be relevant in the context of a (hypothetical) public financial counselling agency contemplating to organise a financial literacy/management course aimed at households that are at risk of experiencing future CFD. This is relevant given the aforementioned studies that seek to identify measures that can protect consumers from CFD. For instance, the natural experiment conducted by Skimmyhorn (2016) shows that offering a short financial management course reduces credit account balance and delinquencies. Therefore, the public agency would be interested in identifying, ex-ante, a small fraction of households at risk of CFD in order to be able to help them to cope with the distress. This is relevant since, due to budgetary constraints, the agency may not be able to provide the course to everyone identified as financially distressed within the population. As a result, the focus would be on correctly identifying a fraction of those that are most likely to experience future CFD rather than how predictable household CFD is on average. I will examine how setting the budget constraint influences the predictive accuracy for the relevant fraction of households in both the sub-samples. The relevance of this approach is that it can improve targeting efforts for potential scarce public financial counselling and debt restructuring advice to those at risk of experiencing future CFD.

A novelty of this study is that when answering RQ1 and RQ2, I consider a forward-looking and time-aware predictive scenario (can you add some ML/statistical refs. here?).

This means that the ML model is trained (fitted to) a training set in one time-interval but then its predictive performance is evaluated on the test set in a future time-interval – one period ahead (see e.g. Hyytinen et al., 2023). This forward-looking scenario is novel, as it mimics a situation where predictions are obtained using only data available at the time of when a policy decision is made (e.g., when deciding who is targeted when financial advice is provided). In contrast to the previous studies (e.g. Salganik et al., 2019, 2020), which are confined to concurrent (not forward-looking) predictive identification of households, the forward-looking scenario allows for an early implementation of support measures for households at risk to become financially distressed.

The rest of this study is structured as follows: In section 2, I describe the data, various binary outcome variables, and the predictors to be used. In Section 3, I address RQ1, i.e., the question of how accurately various ML models can predict CFD on average, and whether there is heterogeneity in the predictability of CFD between households that experienced financial distress in the prior period and those that did not. In Section 4, I turn to RQ2, and study the predictive accuracy of CFD among those who are at the greatest risk of experiencing it. Section 4.4 is dedicated to the discussion of Robustness measures and the limitations of this study. Finally, I discuss my findings in Section 5, and Section 6 concludes.

2 DATA AND VARIABLE DEFINITIONS

2.1 Data sources

The data used for this study comes from Statistics Finland's IDS dataset.[3] The IDS is a nationally representative household level dataset containing data on individual members of households that are permanently residing in Finland. The first IDS wave was collected in 1990. Initially it was collected annually using a rotating panel sampling method such that households in the dataset were surveyed for two consecutive years. Starting from 2009, households have been surveyed for four consecutive years, implying that about a fourth of households in each annual wave of IDS is new.

The IDS is composed of two parts. The first part is obtained from household phone-surveys conducted as a part of the Income and Living Conditions Survey. Each year, about 10,000 households are selected randomly from the Finnish population database to participate in the survey (Statistics Finland, 2022). The second part is obtained from administrative registers on income, pension, social benefits, and taxation. The survey component contains data on the outcome variable of interest, i.e., various measures of financial distress, as well as aggregate household variables, such as income, loans, household structure, and expectations about the development of household's future financial situation. The survey responses are obtained from the household reference person, who is above 16 years old and is usually the highest earning individual in the household. The household

[3]This data was previously used in a study of Finnish consumer over-indebtedness conducted by Hyytinen and Putkuri (2018) where it was described in detail.

reference person responds on behalf of all the members of the household. The administrative registers component contains data on individual household members, such as employment status, level of education, age, gender, marital status, number of children, and region of domicile.

2.2 Measuring financial distress

In this study, I use two main measures of financial distress as outcome variables. The first aims to be a more comprehensive measure of financial distress (henceforth referred to as *FD*). It is constructed as a binary variable that takes the value 1 if any of the following five indicators available in the IDS dataset are present: 1) rescheduled debt payment (*RDP*), 2) perceived over-indebtedness (*POI)*, 3) bill payment problems (*BPP*), 4) debt payment problems (*DPP)*, 5) expenditure difficulties (*EX*). Otherwise, it takes the value 0. All these measures are described in detail in Table 1. The second binary measure, *Debts*, is narrower and takes the value 1 if at least one indebtedness indicator is present, i.e., *RDP*, *POI*, and *DPP*. Otherwise, it takes the value 0.

Table 1: Indicators used in Constructing the Two Outcome Variables of Financial Distress (*FD* and *Debts*):

Indicator of Financial Distress Outcome Reported for the second period	**Survey Question Asked in the second period**	**Encoding**
1. Rescheduled Debt Payment (*RDP*)	Has the household agreed with a bank or other credit institution to change the debt payment schedule or interest rate during the accounting year due to financial survival?	= 1 if Yes; = 0 if No.
2. Perceived Over-indebtedness (*POI*)	Has the household been in a situation where the amount of debt has increased above the financial survival threshold during the statistical year?	= 1 if Yes; = 0 if No.
3. Bill Payment Problems (*BPP*)	Has the household been at least once during the statistical year in a situation where there has been no money to pay payments and invoices on the due date?	= 1 if Very often / Often / Every now and then / Once; = 0 if never.
4. Debt Payment Problems (*DPP*)	Has the household been in a situation where the money has not been sufficient to pay loan repayments or interest during the statistical year?	= 1 if Once or more often; = 0 if Never
5. Expenditure Difficulties (*EX)*	Taking into account all the household income, does the household cover its normal expenditure?	= 1 if With great difficulty, Difficulty, or Little difficulty; = 0 if Quite easily, Easily, or Very Easily

The reason for aggregating multiple binary measures to construct the *FD* and *Debts* outcome variables is that there is no consensus in the economic literature on how to define and hence measure financial distress. First, stand-alone indicators of financial distress might each individually reflect only one aspect or a single dimension of financial distress. The use of multiple measures aims at capturing a variety of dimensions of financial distress and, as a result, enables identifying a larger segment of households in financial distress. Moreover, several objective and subjective measures have been proposed that identify various facets of financial distress, but there is no clear conclusion which should be preferred.

Out of the five available measures (*RDP*, *POI*, *BPP*, *DPP* and *EX*), *RDP* can be considered objective. Even though it is self-reported, it reflects households' *actual* postponement of debt payments due to financial difficulty rather than merely reporting a sentiment of facing financial difficulties but still repaying their debts on time. This measure is also similar in nature to other objective measures of financial distress used in the economic literature, such as delinquency or being in arrears (Oxera, 2004; Collins, 2008; Duygan-Bump and Grant, 2009; Gathergood, 2012). A problem with these measures is that they might overlook households that meet all their debt payment commitments, but who nevertheless experience significant difficulties in meeting their normal expenditure needs (Betti et al., 2007; Eurofound, 2020). Other used objective measures of financial distress capture household debt relative to a pre-defined level of income, consumption, or assets, above which a household is considered financially distressed (Johnson and Li, 2010; D'Alessio and Lezzi, 2013). The main limitation of debt ratios is that there is no established cutoff at which a household can be considered financially distressed (Betti et al., 2007; Garz et al., 2020).

On the other hand, *POI*, *BPP*, *DPP*, and *EX* are based on households' self-assessment of the degree of their own financial difficulty and hence subjective. They are however similar to the subjective measures used in the earlier studies of financial distress (e.g., Betti et al., 2007; McCarthy, 2011; Lusardi, Schneider, and Tufano, 2011; Lusardi and Tufano, 2015; Hyytinen and Putkuri, 2018). The advantage of using subjective measures is that they i) identify households that might not be delinquent or fall below a debt ratio cutoff, but who nevertheless experience financial distress and ii) allow households to assess their own financial situation (Betti et al., 2007). The limitation of subjective measures is that they rely on households' own interpretations of what expenditure difficulties or debt burden might mean (Reijo, 2000, cited in Raijas, Lehtinen, and Leskinen, 2010, p. 218; Bertrand and Mullainathan, 2001). However, Betti et al. (2007) use EU data to document that there does not appear to be a substantial group of households that hide their financial difficulties from surveys.

Since both objective and subjective measures of financial distress have their limitations, I have opted for using as outcome variables both the broader *FD* measure and the narrower *Debts* measure, which focuses on difficulties associated with indebtedness. In particular, the *FD* measure aims at capturing a comprehensive range of financial struggles, whereas the *Debts* measure targets the various difficulties associated specifically with in-

debtedness. Indebtedness difficulties in particular were found to be negatively associated with both physical and psychological health outcomes (see, e.g., Drentea and Lavrakas, 2000; Sweet et al., 2013). Moreover, the global financial crisis in 2008 was a result of borrowers not able to repay their loans. This choice is in line with a number of recent studies that have used multiple objective and subjective measures to construct the outcome variable of financial distress (see for example Gathergood and Guttman-Kenney, 2016; EuroFound, 2020)). The aggregation of several measures is in spirit of the approach presented in Benjamin et al. (2020), where they suggest using multiple survey-based measures to capture different dimensions of wellbeing. In addition, this type of an aggregation has been used in Xu et al. (2015) that investigated the link between non-cognitive characteristics and experiencing financial distress.

2.3 Predictor variables

The predictor variables obtained from the IDS and used in the ML models of this study include: a) continuous variables: log of various debt-to-income ratios (debts such as mortgage, student loan, business, consumer and other loans); log of income; log of consumer debt servicing payments; age; number of children; monthly housing costs to-monthly-income ratio; Yearly interest payments on consumption & other debts- to-income ratio; Assistance received from another household for studies, bill payments, or other purposes, b) categorical variables: education level (3 categories); burden of housing costs (3 categories); lagged RDP, *POI*, *BPP*, *DPP*, and *EX* (2 categories each) gender (2 categories); marital status (4 categories); retirement status (2 categories); language (2 categories); regions of residence in Finland (20 categories); occupation (28 categories); household structure (19 categories); housing tenure (8 categories); development of a household's financial situation compared to previous year (5 categories) ;expectations on development of financial situation (6 categories); having savings (2 categories); having a car or a computer (2 categories each); reason for indebtedness (9 categories). In addition, the following macroeconomic variables are used: unemployment rate, real GDP growth, and inflation.

Similar set of predictors were used in past studies of consumer financial distress (see for example, Brunetti, Giarda, and Torricelli, 2016; Hyytinen and Putkuri, 2018; Kreiner, Leth-Petersen, and Willerslev-Olsen, 2020). Specifically, several of these predictors, such as various debt-to-income ratios (DeVaney and Lytton, 1995; Gathergood and Guttman-Kenney, 2016; Hyytinen and Putkuri, 2018) and lagged financial distress indicators (Burrows, 1997; McCarthy, 2011; Salganik et al., 2020) have been found to predict current financial distress and debt problems more generally, [4] Moreover, it is important to mention that ML models used in this study allow for data-driven identification of non-linear interactions between the different predictors (Fuster et al., 2022, p. 22).

[4] A few important predictors of financial distress used in the literature are not included in the IDS dataset, for instance measures such as financial literacy, stock of unsecured debt, and consumption behaviour. Exclusion of such predictors presents a potential limitation to the ability of ML to predict distress.

While Salganik et al. (2020) have investigated the predictability of financial distress from consumers' perspective, their study does not focus on ex-ante predictability (see also Salganik et al., 2019, 2020). In contrast, my study explicitly models a forward-looking scenario. To generate such a predictive scenario, the predictors for the household-year observations are lagged for one year relative to the outcome variable. This is made possible due to the rotating panel nature of the IDS dataset. It allows the linkage of predictors derived from the first year a household was surveyed (first period) with the outcome variable (*FD* and *Debts*) constructed using data from the subsequent year a household was surveyed (second period).

2.4 Total sample and training and test sets

The data sample used for this study includes years 2008-2017. The years prior to 2008 were excluded because the indicators used in constructing the outcome variables (and to a lesser degree some predictors) were missing in the earlier survey years. In addition, only observations that correspond to the household head are retained in the sample. This results in an annual sample of about 5000 households for the year 2008 and about 2500 households per year for years 2009-2017. After dropping household observations containing: 1) missing data and 2) non-meaningful non-response categories to some survey questions, such as "I do not want to say" and "I cannot say", the total number of observations amounts to 26,103 households.

To address RQ1 and RQ2, I predict the financial distress status of households a year into the future. To this end, I use supervised ML (see Gareth et al., 2013; Varian, 2014; Athey and Imbens, 2019 for general descriptions of ML methods). The objective of supervised ML is to predict an outcome variable (here: *FD* and *Debts*) based on one or more predictor variables. For this purpose, the data is divided into a training and test sets. Then an ML model is fitted to the training set by minimising a particular statistical loss function, such as the logistic loss. In addition to comparing different types of ML models, I also consider different (tuning) parameters for each type of ML model.

To evaluate an ML model's out-of-sample predictive performance and to follow standard ML pre-processing practices, I split the total sample into a training set and a separate test set. The training set contains a pooled set of households first surveyed in years 2008-2014, while the test set contains a pooled set of households first surveyed in 2015-2016. This corresponds to 81.2% (21,192) of data for training and 18.8% (4,911) for testing. The rotating panel characteristic of the IDS dataset enables me to link predictors from the first survey year (first period) with the realised outcome of *FD* and *Debts* from the survey conducted in the following year (second period). For instance, a household first surveyed in 2016 contains predictors from 2016 while the *FD* outcome is from 2017. The training dataset is used to fit an ML model and optimise its predictive performance, while the test set is used to evaluate whether the fitted ML model performs well on previously unseen data. It is worth noting at this point that separate pairs of training and test sets are constructed for *FD* and *Debts*.

Table 2 presents descriptive statistics for the outcome variables in both the training and test samples (sets) respectively. Descriptive statistics on predictors are reported in Appendix A.[5] As can be seen from Table 2, the proportion of the outcome *FD* = 1 and *Debts* = 1 are 30.37% and 7.65% for the training sample and 31.56% and 8.08% for the test sample, respectively. The share of *FD* compares to 19%-20% reported in a study of Finnish household indebtedness by Muttilainen and Reijo (2007). The higher share of households experiencing *FD* in my study can be attributed to the fact that the *FD* outcome variable is constructed using a larger number of financial distress indicators. On the other hand, the variable used by Muttilainen and Reijo (2007) utilises only household payment difficulties for debts and bills as a measure of financial distress. However, the *Debts* measure is directly comparable to Muttilainen and Reijo (2007). In particular, their reported proportions of the *RDP* and *POI* indicators match almost exactly the proportions in my study,

Overall, the values of the descriptive statistics of the predictor variables for the training sample and test sample are comparable (see Appendix A). Furthermore, they align with the descriptive statistics reported for similar measures (age, gender, language, education level, similar socio-economic categories, region of residence) in Hyytinen and Putkuri (2018) that use the IDS dataset.

Table 2: Weighted descriptive statistics for outcome variables in training and test sets

Panel A: Weighted Descriptive Statistics for Outcome Variable in Training Set					
Outcome variable (second period)					
Variable	**Mean**	**Median**	**St. dev.**	**Min**	**Max**
FD (1 indicates the occurrence of *FD*)	0.3037	0.0000	0.4599	0.0000	1.0000
Debts (1 indicates the occurrence of *Debts*)	0.0765	0.0000	0.2655	0.0000	1.0000
Panel B: Weighted Descriptive Statistics for Outcome Variable in Test Set					
Outcome variable (second period)					
Variable	**Mean**	**Median**	**St. dev.**	**Min**	**Max**
FD (1 indicates the occurrence of *FD*)	0.3156	0.0000	0.4647	0.0000	1.0000
Debts (1 indicates the occurrence of *Debts*)	0.0808	0.0000	0.2725	0.0000	1.0000

3 EVALUATING THE STATISTICAL PREDICTIVE PERFORMANCE OF ML MODELS (RQ1)

In this section, I evaluate the statistical predictive accuracy of several ML models. My approach is in line with applications of supervised ML in empirical research as well as in practice, which usually involve fitting multiple ML models and then comparing their performance (see Salganik et al., 2019, 2020; Fuster et al., 2022; Hyytinen et al., 2023).

[5] I use sampling weights, provided as a part of the IDS dataset, to calculate the descriptive statistics for both samples. This implies that the descriptive statistics shown in Table 2 and Appendix A are representative of the Finnish population.

3.1 Empirical framework and predictive ML models

To address RQ1, I examine the predictability of CFD, as measured by *FD* and *Debts*, one year ahead. To this end, I utilise and compare 1) Regularised Logistic Regression (RLR), 2) Support Vector Machines (SVM), 3) Decision Tree (DT), 4) Bagging, 5) Random Forest (RF), 6) Extra Trees (ET), 7) Gradient Boosting (GB), and 8) XGBoost. These ML models are described briefly in Table 3[6].

Table 3: Brief description of ML models used in selecting the best performing model

ML Model	Description
RLR	Models the relationship between the probability of an outcome and predictors using a logistic function. It utilises regularisation to shrink the coefficient estimates to reduce variance (James et al., 2023, pp. 240–247)
SVM	Uses a hyperplane to separate the observations by maximising the distance between the classes (0 or 1 in a binary context) of the outcome (James et al., 2023, pp. 367–384)
DT	A non-parametric hierarchical model composed of decision nodes where observations are recursively split according to IF-THEN rules based on predictors (Alpaydin, 2014, pp. 213–226)
Bagging	Bootstraps multiple DTs and averages their predictions (James et al., 2023, p. 343)
RF	Similar to Bagging, except that RF selects a random subset of predictors at each split-node (James et al., 2023, p. 346)
ET	Similar to RF but with randomly selected split-nodes and without bootstrapping (Geurts, 2006, p. 5)
GB	Grows DTs sequentially (without bootstrapping), successive DTs attempt to minimize the error of the prior DTs (James et al., 2023, pp. 347–348)
XGBoost	Similar to GB but with regularization

Each of these ML models has several tuning parameters that can be used to optimise the model's predictive performance. There are no specific rules governing the selection of values for these tuning parameters. Instead, a common approach is to choose a grid of tuning parameters with a range of values and to select the best performing model with the optimal combination of tuning parameters (for more details regarding parameter tuning, see Kuhn and Johnson, 2013, pp. 64–66) using an 8-Fold Expanding Window Cross-Validation (EWCV) technique (for more details regarding the implementation of parameter tuning using EWCV, see Appendix B). Finally, the test set is used to evaluate the selected model's predictive performance on previously unseen data.[7]

To evaluate the out-of-sample predictive performance for both model selection (validation) and testing, it is necessary to choose an evaluation metric that is used to select the best performing model and to test its predictive performance on previously unseen data. It is common that multiple evaluation metrics are used to provide a comprehensive assessment of ML models' predictive accuracy (Fuster et al., 2022). For this purpose, I

[6] I estimate these models using Python's scikit-learn package.

[7] Throughout this study, I utilize sampling weights, which are provided as a part of IDS dataset, in every ML specification that I implement.

use two evaluation metrics, called the Brier Score and the Average Precision Score. These measures allow me to evaluate how well financial distress can be predicted in advance, which in turn could be useful information for policy makers interested in e.g., designing tailored support measures that aid consumers to cope with CFD.

3.2 Evaluating predictive accuracy

Brier Score

The Brier Score (BS) is a widely used evaluation metric of probability predictions for binary outcomes (see Lahiri and Yang, 2013). In the binary setting, BS is equivalent to the mean squared error metric (MSE), i.e.,

$$BS = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{p}(x_i))^2 \tag{1}$$

where *N* denotes the number of households, *i* denotes the household, y_i is the occurrence of the binary outcome ($y_i = 1$ if household *i* experiences CFD and $y_i = 0$ otherwise), and $\hat{p}(x_i) \equiv \widehat{P}(y_i = 1|x_i)$ is the ML generated probability prediction of household *i* being in financial distress conditional on predictors x_i. The closer the value of BS is to 0, the better the predictive performance, whereas BS = 1 indicates the worst possible predictive performance.

Average Precision Score

Another suitable metric for evaluating the accuracy of binary predictions is the Average Precision Score (i.e., APS; see e.g., Fuster et al., 2022, p. 24). To describe it, let us consider a binary outcome y_i and the associated probability prediction $\hat{p}(x_i)$ for the event $y_i = 1$. We can convert $\hat{p}(x_i)$ into a binary point prediction $\hat{y}_i$ by considering threshold $c \in [0,\ 1]$ such that $\hat{p}(x_i) > c \Leftrightarrow \hat{y}_i = 1$. For instance, if *c* = 0.5, then $\hat{y}_i$ = 1 is predicted (CFD is predicted to occur) if and only if $\hat{p}(x_i) > 0.5$, and $\hat{y}_i$ = 0 (CFD is predicted not to occur) otherwise (Lahiri and Yang, 2013, p. 1059).

Applying a threshold value *c* to the probability predictions results in a joint probability distribution of point predictions $\hat{y}$ and occurrences y. This distribution is commonly known as the confusion matrix, like the one shown in Table 4. In the confusion matrix, $\pi_{\hat{y},y}$ denotes a joint probability, and $\pi_{\hat{y},.}$ and $\pi_{.,y}$ are the marginal probabilities of $\hat{y}$ and y, respectively (Lahiri and Yang, 2013, p. 1059).

Table 4: 2×2 Confusion Matrix

Prediction $\hat{y}$ Is financial distress predicted?	Outcome y Does financial distress occur? Yes y = 1	 No y = 0	 Row Total
Yes $\hat{y}$ = 1	$\pi_{1,1}$	$\pi_{1,0}$	$\pi_{1,.}$
No $\hat{y}$ = 0	$\pi_{0,1}$	$\pi_{0,0}$	$\pi_{0,.}$
Column Total	$\pi_{.,1}$	$\pi_{.,0}$	1

We can use the confusion matrix to define two further quantities, Precision and Recall. They are defined as follows:

$$\text{Precision} = P(y = 1 \mid \hat{y} = 1) = \pi_{1,1}/\pi_{1,.} \tag{2}$$

$$\text{Recall} = P(\hat{y} = 1 \mid y = 1) = \pi_{1,1}/\pi_{.,1} \tag{3}$$

In the context of this study, Precision asks what the probability of observing financial distress is, conditional on it being predicted, whereas Recall asks what the probability of predicting financial distress is, conditional on it being actually experienced. Note that both Precision and Recall share a common numerator, $\pi_{1,1}$, which is the probability of an individual being correctly identified to experience CFD. Such an occurrence is conventionally called a True Positive (henceforth referred to as TP).

The most common way to use these quantities is to consider Precision-Recall curves (PR-curves). To obtain a PR-curve, we can conceive a monotonically increasing sequence of thresholds, c_m , each of which generates a new confusion matrix. From each such matrix, we can calculate a value for Precision, and Recall. Using these paired values, we can plot Precision against Recall for different thresholds c_m . This PR-curve has the property that the closer it is to the top-right corner, the higher Precision and Recall. Specifically, if one uses the unconditional mean of the binary outcome as a naïve prediction, its PR-curve would be a constant line (Saito and Rehmsmeier, 2015).

The APS measure can be interpreted to be a summary measure of PR-curves.[8] It is defined as (Fuster et al., 2022, p. 24):

$$APS = \sum_{m}^{M} (R_m - R_{m-1}) P_m \tag{4}$$

where R_m and P_m refer to Recall and Precision associated with c_m (i.e., the m^{th} threshold). APS is highest when Precision and Recall equal to 1, which would be indicative of the

[8]The APS is a single value metric, whereas PR-curves show the trade-off between Precision and Recall across different thresholds.

case of perfect predictive ability. Therefore, values of APS would normally fall between 0 and 1.

APS provides additional information beyond what is provided by the BS. The BS measures the predictive accuracy of an ML model's probabilistic predictions. APS goes beyond this since it allows for the consideration of the average accuracy of predictions across various thresholds (c_m). In addition, one can expect APS to be particularly informative for the IDS dataset, since the event of a household being in financial distress is less common than its alternative (Fuster et al., 2022, p. 24). This is underscored by the fact that both Precision and Recall include $\pi_{1,1}$ in the numerator, which represents the probability of correct predictions of those experiencing CFD.

Baselines

Assessing the absolute performance of predictive ML on *Debts* and *FD* is not straightforward because there is no commonly accepted standard for what constitutes a good level of predictive accuracy for consumer financial distress. This issue is a common challenge in predictive studies, which therefore often utilize baselines to evaluate the comparative performance of the more sophisticated models. For instance, some studies use the unconditional mean of the sample as a benchmark (Lahiri and Yang, 2013) or compare human decisions to those of ML (Kleinberg et al., 2018). Hence, a key step in evaluating predictive accuracy is to choose an informative baseline for comparison (Hofman et al., 2021).

In order to assess the relative predictive accuracy, I use two baselines. The first is a logistic regression with only prior CFD predictor (B1). The second is a logistic regression with prior CFD and all the other available non-CFD predictors (B2), as described in Section 2.3. The motivation to use B1 as a baseline is that consumer financial distress is known to persist from one time-period to future periods (Brown, Ghosh, and Taylor, 2014), suggesting that prior CFD is a relatively good predictor of future CFD. Moreover, B1 aligns with a similar baseline approach used in a prior study by Salganik et al. (2019). The B2 baseline can be considered a more demanding baseline compared to B1, as it includes a broader set of predictors. As mentioned earlier, these predictors are selected based on prior economic literature on the topic of consumer financial distress (see for example, Brunetti, Giarda, and Torricelli, 2016; Hyytinen and Putkuri, 2018; Kreiner, Leth-Petersen, and Willerslev-Olsen, 2020). A more extensive baseline, with additional predictors, is also used in Salganik et al. (2019, 2020). The improvement in predictive performance of ML relative to baselines B1 and B2 is assessed using the percent improvement (%Δ) in BS and APS (in line with Fuster et al., 2022, p. 28).

3.3 Results: Average ex-ante predictability

ML's relative predictive performance evaluated using BS and APS scores

Comparing the predictive performance of the various ML models indicates that the best performing models are GB and RF for the *FD* and *Debts* measures, respectively (for details

of the model selection and validation procedure and how the GB and RF models perform against other ML models, see Appendix C). [9] Table 5 summarizes the comparative performance of these best performing GB and RF models in predicting consumer financial distress on previously unseen test data. Each ML model is evaluated against the two baselines B1 and B2, using the BS and APS evaluation metrics.

Table 5: Performance Comparison of Best ML Models Relative to B1 and B2 Baselines on the Test Set

FD		
Model	**BS**	**APS**
Baseline B1	0.1392	0.6076
Baseline B2	0.1311	0.7669
ML (GB Model)	0.1237	0.7902
Percent improvement (%Δ) GB over B1	+11.1%	+30.1%
Percent improvement (%Δ) GB over B2	+5.6%	+3.0%
Debts		
Model	**BS**	**APS**
Baseline B1	0.0573	0.3092
Baseline B2	0.0542	0.4852
ML (RF Model)	0.0528	0.5334
Percent improvement (%Δ) RF over B1	+7.9%	+72.5%
Percent improvement (%Δ) RF over B2	+2.6%	+9.9%

For *FD*, the predictive performance of ML is 0.1237 and 0.7902 as evaluated by the BS and APS metrics, respectively. In comparison, both B1 and B2 baselines obtain lower scores, with B1 scoring 0.1392 (BS) and 0.6076 (APS), and B2 scoring 0.1311 (BS) and 0.7669 (APS). The predictive results of ML demonstrate a moderate improvement of 11.1% over B1 and a smaller improvement of 5.6% over B2 in terms of BS. On the other hand, the improvement of the ML model in terms of the APS is 30.1% and 3.0% relative to B1 and B2, respectively. The improvement of ML over B1 is notably in the double digits based on both BS and APS. In particular, the improvement is more pronounced based on the APS metric, which is better suited for data scarce in those experiencing *FD*. On the other hand, only a slight improvement in ML performance is observed over the B2 baseline across both evaluation metrics.

Next, I consider the predictability of consumer financial distress when it is measured using *Debts*. The predictive performance of ML is 0.0528 and 0.5334 as evaluated by the BS and APS, respectively. The B1 baseline obtains a score of 0.0573 (BS) and 0.3092 (APS), while the B2 baseline scores 0.0542 (BS) and 0.4852 (APS). Overall, the comparative results for *Debts* follow a similar pattern as *FD*. The ML model shows an improvement of 7.9% over B1 and 2.6% over B2 in terms of BS. However, based on the APS metric, we

[9] It is important to note that the GB model, when selected based on the BS, differs in terms of tuning parameters to the GB model selected based on APS. A similar distinction applies to the RF model. For details on tuning parameters, see Section 3.1.

see that the relative improvement is much more significant over B1 at 72.5% and a modest improvement of 9.9% over B2.

The discrepancy between the scores across the two evaluation metrics (BS and APS) can be attributed to the differences in how each of these metrics emphasises the predictability of the rare observations of CFD. As mentioned earlier, the APS is more informative in these cases because it also focuses on the rare observations of experiencing CFD. In contrast, the BS focuses on the overall accuracy of probability predictions, where it treats each observation equally, without consideration for the rarity of the occurrence of CFD. Hence, given the differences in base rates between the sample of *FD* at 31.6% and *Debts* at 8.1%, the ML model's performance, as measured by BS, may better emphasise the predictability of the relatively more prevalent non-occurrence of *Debts* (compared to non-occurrence of *FD*). On the other hand, the better predictability for *FD* relative to *Debts*, as measured by the APS metric, could be due to either differences in base rate, a better ability of ML to predict *FD* relative to *Debts*, or a combination of both. That being said, a detailed analysis of the reasons for the disparity between the predictability of the *FD* and *Debts* measures is not the focus of this study and should be generally approached with caution, as these two outcomes measure different aspects of consumer financial distress.

Statistical significance

Following the procedure described in Fuster et al. (2022, p. 27), I verify whether the improvements over the baselines reported in Table 5 are statistically significant using bootstrapping. To accomplish this, I hold fixed the original training and test splits (for details on the data splitting process see Section 2.4) and resample observation with replacement to create 100 bootstrapped training datasets. I then fit each model (B1, B2, and the ML models GB and RF) on each bootstrapped training dataset and evaluate their predictive performance on the test set. For both *FD* and *Debts*, and across both BS and APS metrics, ML outperforms both the B1 and B2 baselines in 100% of these bootstrapped samples. Therefore, we can conclude that ML models can make more accurate forward-looking predictions of *Debts and FD* compared to the B1 and B2 baselines.

3.4 Results: Heterogeneity in predictability

I now take a closer look at how the predictability of CFD is influenced by the persistence of CFD. In particular, I analyse whether the predictability of CFD is influenced by the presence of prior period experience of CFD, as this could then result in heterogeneity in predictive performance across households depending on whether they experienced prior CFD or not. This heterogeneity can be investigated by comparing the Precision Recall Curves (i.e., PR-curves; see e.g., Fuster et al., 2022, p. 24) across the subsample of households with prior CFD and the sub-sample without prior CFD. In the analysis that follows, I will use the graphical tool of Precision-Recall curves to summarize the predictive accuracy of the estimated models for the different sub-samples discussed above as well as the full sample.

The PR-curves for the *FD* and *Debts* are presented in Figures 1 and 2, respectively. These figures plot the PR-curves of the best performing ML models (as determined by the APS and shown in Table 5) alongside the two baselines, RQ1-B1 and RQ1-B2[10]. In both Figure 1 (*FD*) and Figure 2 (*Debts*), there are thus three PR-curves. The Green curve represents the PR-curve of the ML model (GB for *FD* and RF for *Debts*), the Orange curve represents RQ1-B2, and the Blue dotted line represents RQ1-B1. Moreover, each subplot from left to right within Figures 1 and 2 represents the full sample (left), the subsample with no prior CFD (centre), and the subsample with prior CFD (right). As stated earlier in Section 3.2, each point on a PR-curve represents the P_m (Precision) and R_m (Recall) at the various thresholds c_m . Therefore, the PR-curves offer a comprehensive visualisation of the trade-offs between Precision and Recall at the various thresholds c_m . Higher thresholds are associated with higher levels of Precision. Conversely, lower thresholds tend to increase Recall at the cost of lowering Precision.

Baselines

To be more specific on how the baselines have been defined for this set of figures, the first baseline, RQ1-B1, is the base rate of CFD, computed separately for the full sample and subsamples, respectively. This baseline is constant across the various thresholds *m*. When I consider the full sample, this baseline reflects the unconditional mean of the sample, which can be considered a naïve baseline. The base rate is 31.6% for the full sample of *FD* and 8.1% for *Debts*. However, when I consider the subsamples with and without prior CFD, the base rates of these respective subsamples account for whether consumers experienced prior CFD or not, thereby factoring in persistence. The base rates for the subsample with no prior *FD* and *Debts* are 11.7% and 4.1%, respectively, while for those with prior *FD* and *Debts* the corresponding base rates are 70.2% and 50.6%. The second baseline, RQ1-B2, utilises the exact same baseline as B2, but is presented as a curve rather than as a single-value APS.

Findings based on PR-curves

Considering Figure 1 first, for the full sample (right), we observe that the ML model is relatively close to the top right corner and clearly lies above the corresponding RQ1-B1 baseline. This suggests that the predictive performance of the ML model for *FD* is relatively better compared to the RQ1-B1 baseline when considering the full sample. Furthermore, the ML model's PR-curve remains above the RQ1-B2 baseline for recall levels below 0.7, above which ML's performance appears to be the same as the RQ1-B2 baseline. This implies that the ML model is better suited in scenarios where a balance is sought between obtaining good levels of both Precision and Recall. However, if a higher Recall

[10]The PR-curves for *FD* and *Debts* are plotted for the best performing ML models as evaluated by the APS, which are shown to achieve statistically significant improvements over their corresponding baselines, as detailed in Section 3.3. This approach is consistent with Fuster et al. (2022, p. 27).

Figure 1: PR-curves demonstrating the predictability of the *FD* outcome

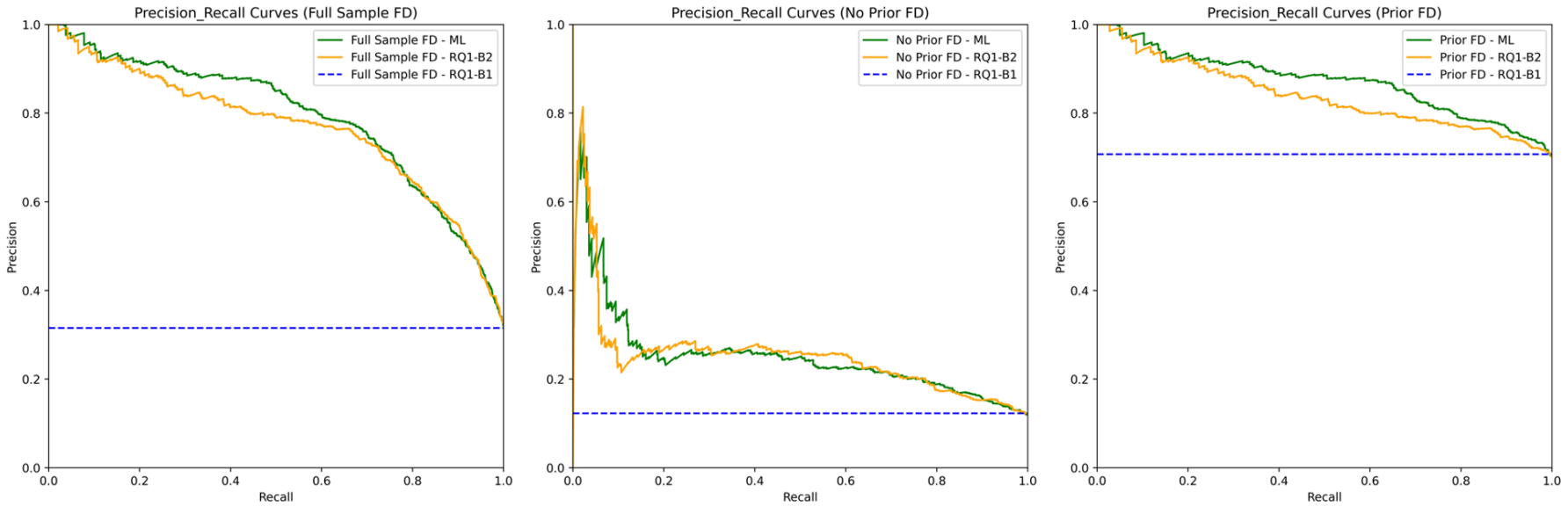
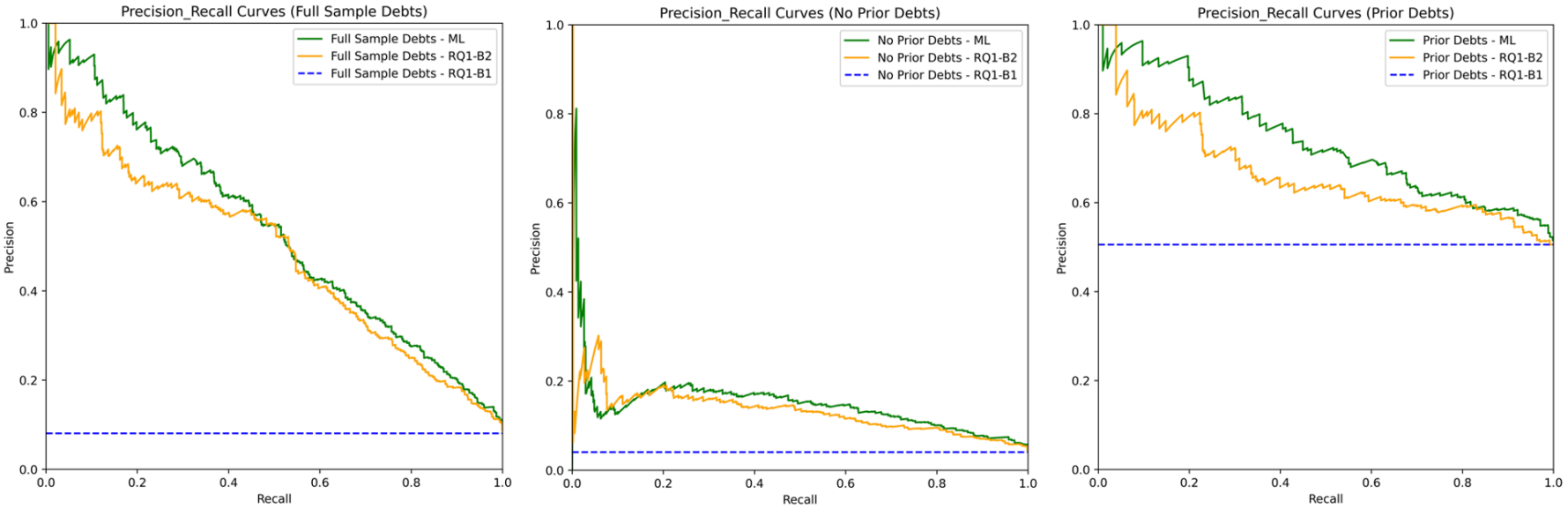


Figure 2: PR-curves demonstrating the predictability of the *Debts* outcome



level is preferred, at the unavoidable cost of lowering Precision[11], then the ML model's predictive performance for the full sample appears to be as good as the RQ1-B2 baseline. For the subsample with no prior *FD* (centre), we immediately observe a significant drop in the predictive performance of the ML model in this subsample compared to the that of the full sample, as evidence by the ML's comparatively lower PR-curve[12]. This result calls into question the practical suitability of using such *FD* predictions for households that have not experienced prior *FD*, due to their low predictability in absolute terms. Moreover, The ML model and RQ1-B2 exhibit comparable performance for the subsample with no prior *FD*. However, both appear to clearly outperform the RQ1-B1 baseline. On the other hand, we can observe improved performance by the ML model in predicting *FD* for the subsample with prior *FD* (right) compared to the full sample and especially to the subsample with no prior *FD*. In addition, for this subsample, ML clearly outperforms both RQ1-B1 and RQ1-B2.

Turning then to Figure 2, which presents the PR-curves for the *Debts* measure, the overall results closely resemble those discussed in connection with Figure 1. When considering the full sample (left), we observe that the PR-curves of the ML model clearly outperform both the RQ1-B1 and the RQ1-B2 baselines across the full range of precision-recall levels. In terms of the subsample with no prior *Debts* (centre), similar to *FD* as presented in Figure 1 (centre), there is a significant drop in the predictive performance of the ML model compared to the that of the full sample. On the other hand, for the subsample with prior *Debts* (right), the ML model demonstrates improved predictive performance over the full sample and the subsample with no prior *Debts*. The ML model's performance also clearly improves over the RQ1-B1 and RQ1-B2 baselines for this subsample.

When analysing the findings presented in Figures 1 and 2 in conjunction, we can make the following observations. First, we observe that across both CFD measures, the ML models' predictive performance is significantly lower in the subsample with no prior CFD. This suggests that the ML models used in this study are relatively accurate at predicting for the full sample, and particularly for the subsample with prior CFD. However, the same ML models are less accurate at identifying households with no prior CFD. These findings demonstrate that predictive accuracy for households that have experienced prior CFD can be improved using ML models, however, that does not appear to be the case for households that have not experienced prior CFD.

[11]This would be the case if the cost of incorrectly predicting *FD* = 1 ($\hat{y}$ = 0 instead of $\hat{y}$ = 1) is significantly high, thereby necessitating heavily favouring achieving higher Recall at the cost of lower Precision.

[12]Contrary to the usually expected PR-curve behaviour, which starts with high precision closest to the vertical-axis and then trends downwards, this curve begins with a precision of 0. This uncommon behaviour can be attributed to the low base rate of this subsample (11.7%), combined with the definition of precision according to equation 2 in Section 3.2. In this equation, the numerator, $\pi_{1,1}$, represents the probability of correct *FD*=1 predictions. However, at higher thresholds *m* (closer to the y-axis), if the predictive model fails to identify any TPs, due to their rare occurrence in this subsample, as well as, the resulting difficulty for the predictive model to accurately predict these cases, then $\pi_{1,1}$would be equal to 0. Hence, precision would also start at 0.

Second, it appears that the PR-curves for *FD* are consistently higher (better predictability) than their corresponding curves for *Debts*, as demonstrated by comparing Figure 1 and 2. To a certain extent, this disparity appears to be attributable to the differences in base rates between the full sample of *FD* at 31.6% and *Debts* at 8.1%, the subsample with prior *FD* at 70.2% and *Debts* at 50.6%, and the subsample with no prior *FD* at 11.7% and *Debts* at 4.1% (Berk, 2012, p. 5). This is also visible in Figures 1 and 2, where the point furthest to the right of each of the PR-curves (here referring to the actual PR-curves corresponding to the ML model and the RQ1-B2 baseline) is the base rate (the constant dotted blue line representing RQ1-B1) of its respective sample or subsample. An expected result given that a recall of 1 is usually obtained when all observations (households in this study) are predicted to be in CFD. This implies that the denominator of precision ($\pi_{\cdot,1}$), shown in equation 2, corresponds to the total number of observations, yielding a value for precision that is equivalent to the base rate. This, however, does not rule out that the ML model could just make more errors when predicting *Debts* compared to predicting *FD*.

4 PREDICTIVE ACCURACY UNDER A BUDGET CONSTRAINT (RQ2)

4.1 Decision Context

In this section, I analyse a context in which predictions are viewed as an input to decision-making, as opposed to merely evaluating the average predictability of CFD. This analysis shifts the focus from statistical accuracy to the practical usefulness of predictions, as viewed from the decision-maker's perspective (see, e.g., Granger and Machina, 2006; Lieli and White, 2010; Lieli and Springborn, 2013; Lahiri and Yang, 2013; Elliot and Timmermann, 2016; Hyytinen et al., 2023). The decision-context is analogous to that in the prediction policy problems considered in Kleinberg et al. (2015), where they utilize ML-predictions as inputs to forward-looking decisions.

Specifically, I consider the perspective of a public financial counselling agency operating under a budget constraint. The agency's objective is to help consumers who suffer or may potentially suffer from CFD. It therefore organises financial management and advice courses targeted at households at risk of experiencing CFD in future. Given the budget constraint, the agency needs to identify ex-ante a subset of households that are at the highest risk of experiencing CFD to offer them voluntary participation in the course. Skimmyhorn (2016) provides empirical evidence supporting the effectiveness of such short financial management courses in coping with CFD, i.e., in reducing credit account balance and delinquencies in the subsequent year. The idea here is to explore the extent to which ML algorithms can help the agency ex-ante identify the subset of financially distressed households while adhering to the budget constraint. The goal of the targeted advice is not to reduce the likelihood at which consumers experience CFD but rather to help mitigate its adverse consequences and help consumers to recover from it.

To consider the practical usefulness of ML predictions from decision-makers' perspective requires integrating two aspects into the ML approach. First, I use precision, also known as positive predictive value (PPV), as the accuracy metric of interest in the agency's decision-context, given that it measures the share of correctly predicted households in CFD among all households predicted to be in CFD. This metric is arguably of interest to the financial agency since it is a direct measure of how well it can identify consumers who are at risk of experiencing CFD and thus in need for advice. Second, to incorporate the budget constraint into the ML approach, I consider the predictability of CFD for a prespecified small fraction of households. For example, the budget of the agency might allow it to organize counselling for a certain fraction, say 2%, of its potential target population. The question of interest is, how could this fraction be chosen? To this end, I construct a budget-PPV curve that plots PPV on the y-axis against the ranked (ordered) predicted probabilities of CFD (screening budget) on the x-axis (see Hyytinen et al., 2023, p. 2471 for a similar approach). When evaluating how well the ML predictions work, the PPV is calculated recursively for different budget values using the realized values of CFD in the test sample, starting from the households with the highest predicted probabilities of experiencing CFD. The budget determines the percentage of highest ranked households that can be selected as candidates for the finance management course. For instance, if the budget can accommodate 2% of the relevant sample, then this methodology would select all the relevant households whose *predicted* probabilities are above the 98^{th} percentile of the distribution of predicted probabilities. The budget-PPV shows how well CFD can be predicted in the tail, i.e., for consumers who have the greatest risk of experiencing CFD. Note that such an approach does not require that the finance management course has previously been implemented in practice, as would be the case in treatment-control studies (Athey, Keleher and Spiess, 2023, p. 1).

The budget-PPV curves are plotted using the best performing ML model (as chosen by the APS, the summary measure of precision and recall; see Herlocker et al., 2004, p. 25; Aggarwal, 2016, p. 247) for both *FD* and *Debts* measures, with the accompanying predictors. Because there is heterogeneity in the predictability of CFD depending on whether individuals have experienced it in the prior period (see section 3.6), I calculate the PPV-budget curves separately for i) the full sample, ii) the subsample with no prior CFD, and iii) the subsample with prior CFD. The use of these different subsamples corresponds to different information sets that the financial counselling agency could be using when picking whom to advice (see below). I will be referring to this context-aware ML setup that accounts for the financial counselling agency's decision context as *ML* in the text and graphs that follow.

4.2 Baselines

To evaluate the comparative predictive performance of the context-aware *ML* when considering how ML predictions might help the financial counselling agency's decisions, I use two baselines. The first baseline, RQ2-B1, is the base rate of CFD, computed separately for each of the respective full sample and subsamples, as was done in Section 3.4. The use

of this baseline implies that the financial counselling agency just picks randomly a certain (budget-specified) fraction of consumers: Specifically, when I consider the full sample, this would mean that the agency ignores any information it might have on consumers. However, when I consider the subsamples with and without prior CFD, I am assuming the agency uses currently available information on consumers' experience of CFD (given that the agency is making forward-looking predictions at time t, "prior" CFD amounts to "current" CFD when making predictions for t+1). For example, when focusing on those with prior CFD, the financial counselling agency picks randomly a certain (budget-specified) fraction of consumers with prior CFD. Similarly, when focusing on those without prior CFD, the financial counselling agency picks randomly a certain (budget-specified) fraction of consumers without prior CDF. This baseline is thus informative, as it is motivated by the findings that prior-period experience of CFD predicts the next period's CFD relatively well (persistence).

The second baseline, RQ2-B2(ML), employs the same best performing ML model as above for *ML*, but with a constant threshold that does not account for the decision-context of the financial counselling agency. In the absence of any specific decision context, the commonly used *ad hoc* thresholds (c) are either 0.5 or the unconditional mean of the outcome variable (Greene, 2011; Lahiri and Yang, 2013, p. 1083; Elliot and Timmermann, 2016, p. 262). Given the imbalance in the outcome variables, with both *FD* and *Debts* equal to 1 being less represented in the sample (minority class) compared to the alternative (majority class), I set the threshold c to correspond to the unconditional mean of the event of *FD* and *Debts,* respectively, when constructing this baseline. The use of this baseline implies that when I focus on the full sample, the financial counselling agency picks randomly a certain (budget-specified) fraction of consumers for whom the predicted probability is above these thresholds. When I focus on the subsample of consumers with (without) prior CFD, the financial counselling agency picks randomly a certain (budget-specified) fraction of consumers for whom the predicted probability is above these thresholds in these subsamples.

RQ2-B2(ML) could be considered a more informative baseline compared to RQ2-B1, as it utilises the same underlying ML model with prior CFD (persistence) and all the additional predictors. However, unlike *ML*, RQ2-B2(ML) does not consider the budget-PPV decision-context of the counselling agency. By this I mean that the RQ2-B2(ML) baseline does not involve ranking households based on their predicted probabilities. Therefore, it can be argued that any observed improvement in predictive performance of the *ML* model over RQ2-B2(ML) are due to accounting for the agency's decision-context by means of the budget-PPV curve.

4.3 Results: Budget-PPV

The results in Figures 3 and 4 show the budget-PPV curves for the *FD* and *Debts* measures, respectively, across the full sample (left), subsample with no prior CFD (centre), and subsample with prior CFD (right). The budgets allow for the selection of 1% to 5%

of potential households into the agency's program, as reflected by the x-axis. The y-axis shows the precision at each budget-PPV percentile. In each figure, the ML model is represented by the green curve, the RQ2-B2 baseline by the orange curve, and the RQ2-B1 baseline by the blue dotted line. In addition, I plot the 95% confidence intervals for the ML model (green shaded area) and the RQ2-B1 baseline (orange shaded area). The construction of the budget-PPV curves and their corresponding confidence intervals for *ML* and the RQ2-B2(ML) baseline are based on Fuster et al.'s (2022, p. 27) bootstrapping strategy, which I also describe in Section 3.3.

When examining Figures 3 and 4, we can note a number of observations. First, for the full sample (left), *ML* clearly outperforms the RQ2-B1 and RQ2-B2(ML) baselines across both *FD* and *Debts* measures. The range of PPV values using *ML* spans from 99.1% at the 1st percentile down to 92.9% at the 5th percentile for *FD*, and from 86.7% at the 1st percentile to 62.2% at the 5th percentile for *Debts*. The comparative PPV for RQ2-B2(ML) is 69.6% - 69.6% and 20.7% - 20.8% for *FD* and *Debts*, respectively. On the other hand, the PPV values for the constant RQ2-B1 baseline are 31.6% and 8.1% for *FD* and *Debts*, respectively. In addition, *ML*'s improvement in performance over the RQ2-B2(ML) baseline is also statistically significant, as is suggested by the non-overlapping confidence intervals around its respective budget-PPV curve and that of the RQ2-B2(ML) baseline. This implies that *ML* can discern future CFD for the relevant budgets better compared to both the naïve RQ2-B1 baseline and the RQ2-B2(ML) baseline that does not account for the budget-PPV decision context.

Second, for the subsample with no prior CFD (centre), *ML* outperforms both the RQ2-B2(ML) and RQ2-B1 baselines on average for *FD* and *Debts*. Notably, for *FD* at the 1st percentile, *ML*'s PPV value is 50.3% compared to 30.6% and 11.7% for RQ2-B2(ML) and RQ2-B1 baselines, respectively. However, across both *FD* and *Debts*, *ML*'s improvement in PPV is only significant relative to the RQ2-B1 baseline, as the confidence intervals of *ML*'s budget-PPV curve overlaps with that of the RQ2-B2(ML) baseline for both *FD* and *Debts*. In addition, *ML*'s budget-PPV curves for this subsample are relatively lower compared to those of full sample, and as will be discussed later, they are also significantly lower compared to those of the subsample with prior CFD (right). Hence, the usefulness of *ML* predictions in this case appears to be limited, since the PPV remains low in absolute terms and *ML*'s improvement over the RQ2-B2(ML) baseline does not appear to be statistically significant.

Third, for the subsample with prior CFD (right), *ML* outperforms both the RQ2-B2(ML) and RQ2-B1 baselines on average for *FD* and *Debts*. *ML*'s improvement in PPV appears to be statistically significant to a large degree across the 1-5 percentile budgets for both *FD* and *Debts*. Notably, *ML*'s PPV is 99.8% at the 1st percentile and down to 98.4% at the 5th percentile for *FD*. Similarly for *Debts*, the comparative PPV values are 99.4% and 94.7%. These average PPV values are considerably higher compared to those obtained using the RQ2-B1 baseline, which yields 70.2% and 50.6% for *FD* and *Debts*, respectively. These results strongly suggest that *ML* can discern future CFD significantly more precisely compared to relying only on prior experience of CFD as an indicator of

Figure 3: Budget-PPV curves (*FD*)

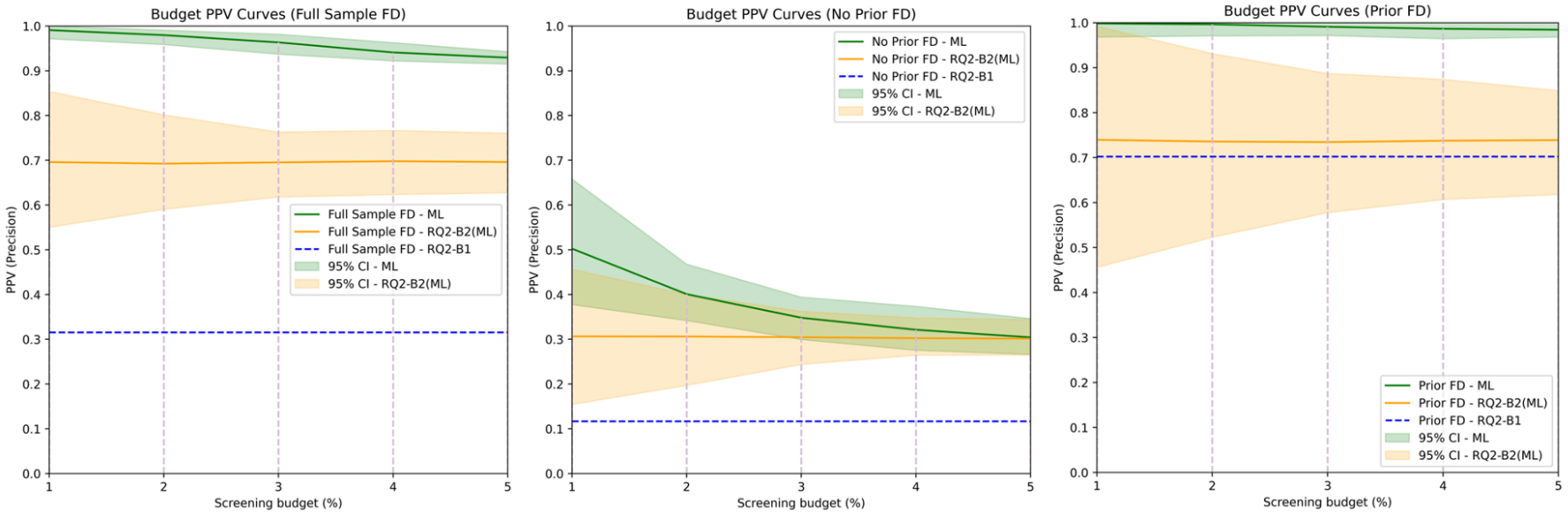
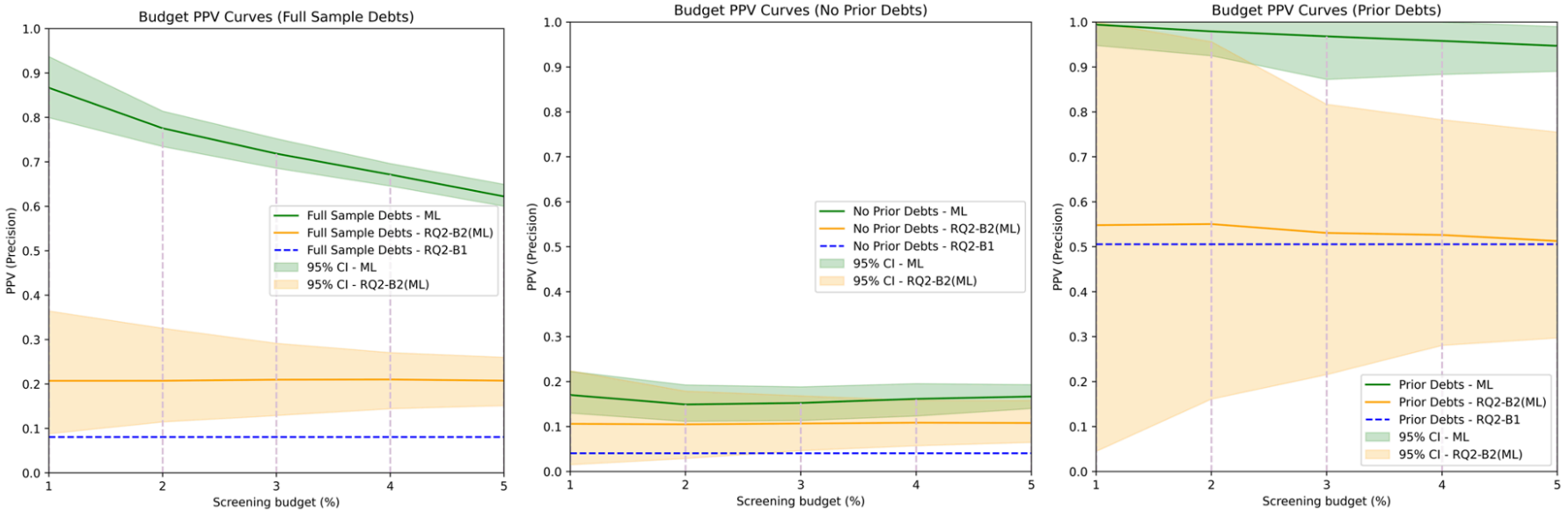


Figure 4: Budget-PPV curves (*Debts*)



future CFD (persistence). In terms of the RQ2-B2(ML) baseline, the comparative PPV values are 74.0% - 73.9% and 54.8% - 51.3% on average for *FD* and *Debts*, respectively. These PPV values are on average slightly better than those for the RQ2-B1 baseline, but considerably lower compared to those of *ML*. These results suggest that accounting for the budget-PPV decision context significantly improves precision for the relevant budgets compared to using the same ML model without consideration for this decision-context.

4.4 Robustness

In this sub-section, I discuss the robustness of my empirical findings. I consider various issues, each in turn.

Measurement of CFD

The robustness checks I use here are as follows: using two measures of CFD, namely, *FD* and *Debts*, to verify that performance of ML shows a consistent pattern across various measures of CFD.

In addition, I use these two different CFD measures to examine the heterogeneity in the predictability of CFD between households that experienced financial distress in the prior period and those that did not. The results demonstrate that the predictability of CFD for households with no prior CFD is significantly lower compared to the full sample and the subsample with prior CFD across two CFD measures. This consistent pattern of heterogeneity across the two measures of CFD provides evidence that the results are not specific to one CFD measure but are robust across both measures.

Similarly, I use these two CFD measures to ascertain that the results of the budget-PPV curves are robust and do not depend on a single measure of CFD.

Evaluation of predictive performance

Although the magnitude of ML's relative performance differs between the two CFD measures when assessed using the same evaluation metric, the pattern of improvement over the two baselines remains consistent across both measures, as shown in Table 5. This provides suggestive evidence that the relative performance of ML models may generalize across different measures of CFD

Choice of multiple baselines

I use two baselines to ensure that the relative improvement in ML models' PR-curves is not dependent on a single baseline. The ML models' PR-curves largely appear to outperform both baselines when considering the PR-curves for the full sample and the subsample with prior CFD across both measures of CFD, as shown in Figures 1 and 2.

Similarly, I use two baselines to ensure that the relative improvement in ML models' budget-PPV is not dependent on a single baseline. As an additional robustness check here,

I use a third baseline, namely, a logistic regression model with prior distress and the other predictors as outlined in Section 2.3. The threshold (c) in this case is the unconditional mean of the respective CFD outcome variable as in the case of RQ2-B2(ML). This baseline is referred to as RQ2-B2. The budget-PPV curves for this baseline can be found in Figures D1 and D2 of Appendix D for *FD* and *Debts*, respectively. The RQ2-B2 baseline's budget-PPV curves, as shown in Figures D1 and D2, concur with the budget-PPV curves shown in Figures 3 and 4, whereby ML generally outperforms the RQ2-B2 baseline as it outperforms RQ2-B2(ML).

Use of multiple subsets of the data and statistical significance

An additional robustness check was to test using EWCV (described in Appendix B) the performance of the different ML models on multiple subsets of the data. This allowed me to compute the average performance of each ML model (with its corresponding set of tuning parameters) across these various subsets. The results of EWCV can be viewed in Appendix C, Tables C1 and C2, under the column Val.Set for *FD* and *Debts*, respectively, across both BS and APS evaluation metrics. These results show that the top-performing ML models in these tables are not resulting from a particular draw of observations, instead, the top ML models' performance is reliably better compared to other ML models.

Furthermore, following the procedure described in Fuster et al. (2022, p. 27), I verify, using 100 bootstrapped samples, that the improvement of ML models over the two baselines B1 and B2 are statistically significant for *FD* and *Debts* across both BS and APS evaluation metrics. This provides evidence that the improvement in ML models' performance over the baselines is not just due to chance.

Moreover, I conduct bootstrapping to verify that the results are statistically significant and are not sensitive to a single sample draw of the dataset. In doing so, I plot confidence intervals, as shown in figures 3 and 4 as well as figures D1 and D2, to demonstrate that MLs' budget-PPV curves generally outperforms the baselines, showing statistically significant improvements for both the full sample and the subsample with prior CFD.

5 DISCUSSION

5.1 Discussion: Average ex-ante predictability

The findings shown in Table 5 and described in Section 3.3 contribute to the limited prior literature investigating the predictability of financial distress in two ways. First, unlike prior studies investigating the concurrent predictability of CFD, this study shows that consumer financial distress can be predicted ex-ante. Although this forward-looking setup is distinctive in the context of CFD predictions, it has been used in other decision contexts, such as the ex-ante prediction of: high-growth enterprises (Hyytinen et al., 2023); futile hip-replacement surgeries (Kleinberg et al., 2015); whether a defendant, if released, will

commit a crime (Kleinberg et al., 2018; Babii et al., 2020); consumer default (Fuster et al., 2022).

Second, when evaluating the relative predictive performance using the BS, the modest improvement of ML predictions relative to the B1 baseline (prior CFD as a predictor) is consistent with the findings by Salganik et al. (2019, 2020). This result, coupled with Brown, Ghosh, and Taylor's (2014) findings that show persistence of CFD over time, might suggest that using just prior CFD as predictor could yield predictive performance comparable to that of more complex ML models using multiple predictors. However, when using the APS metric, which is more suitable for evaluating relatively rare events such as experience of CFD, the ML models' results show a considerable improvement against the B1 baseline for both *FD* and *Debts*. This improvement in ML performance demonstrates that using an ML model with a selection of predictors, which are based on prior economic literature, can improve predictive performance considerably over the B1 baseline. This implies that ML predictions of CFD could potentially be used to better inform borrowing decisions or to pre-emptively target of financial counselling and debt restructuring advice to households at risk of future CFD. Therefore, these findings add to prior research by suggesting that ML can be utilised for the ex-ante prediction of CFD with relatively good accuracy (on average) in real-world forecasting scenarios and for decision-making purposes.

On the other hand, the improvement in ML models' performance against the B2 baseline, although statistically significant, is modest even when considering the APS evaluation metric. This suggests that using more complex ML models only modestly improves over a standard logistic regression model. This result calls on future research to utilise richer datasets with additional predictors documented in the literature to be associated with CFD. For instance, the following predictors could improve predictability of future CFD: financial literacy (Stango and Zinman, 2009; Campbell et al., 2011; Lusardi and Tufano, 2015); unsecured consumer debt-to-income ratio[13] (Del Rio and Young, 2008; Gathergood and Guttman-Kenney, 2016); expenditure behaviour, as consumers with impulsive expenditure behaviour (Ottaviani and Vandone, 2011; Gathergood, 2012) or gambling habits (Muggleton et al., 2021) are found to be more likely to experience CFD. The use of such predictors could increase the likelihood of ML models finding non-linear interactions between the different predictors that would further improve the accuracy of ML models' CFD predictions relative to those of the B2 baseline (Fuster et al., 2022). Nevertheless, these modest improvements in ML models' performance over the B2 baseline should not be dismissed, as they are statistically significant. Similar modest improvements of an ML model over a logistic regression baseline (corresponding to our B2 baseline) are also reported in the study predicting consumer default by Fuster et al. (2022). Moreover, the ML models' relative improvement in predictive performance over the two baselines across two different measures of CFD follows a consistent pattern. This allows me to cautiously conclude that the pattern of relative improvement in ML models can generalise to various measures of CFD.

[13]As the measure of consumer debt in the IDS also includes debt from holiday home purchase, which is likely to be collateralized (secured).

5.2 Discussion: Heterogeneity in predictability

The results outlined in Section 3.3 and discussed in Section 5.1 demonstrate that ML with multiple predictors significantly outperforms the B2 baseline in average predictive accuracy when evaluated using the APS metric. However, as discussed in Section 3.4 and demonstrated using the PR-curves in Figures 1 and 2, the ML models' predictability of CFD is significantly lower for the subsample with no prior period CFD compared to the full sample and the subsample with prior CFD. This demonstrates heterogeneity in the predictability of CFD and calls into question the efficacy of using ML predictions of CFD to improve borrowing-decisions or to target support measures to the subsample with no prior CFD. Therefore, this result is significant, as it reveals that analysing predictability of CFD solely based on the full sample, as was the case in prior research (see Salganik et al., 2019, 2020), can conceal the significantly lower predictive accuracy for the relevant subgroup, namely, the subsample with no prior CFD.

Furthermore, when evaluating the PR-curves relative to their respective baselines, additional insights ensue. First, the ML models' performance for the full sample shows clear improvement over the RQ1-B1 baseline for both *FD* and *Debts*, which demonstrates that ML models can considerably outperform the base rate baseline. This tells us that ML models can perform better than the absolute minimal baseline in making forward-looking predictions of CFD. Similarly, the predictive performance of ML models is generally better compared to the RQ1-B2 baseline, albeit less substantial. This is also consistent with the results reported in Table 5 of Section 3.3, where the improvement of the ML models over the B2 baseline[14], as measured by the APS metric[15], is minimal for *FD* (3.0%) and more considerable for *Debts* (9.9%). However, the PR-curves offer an additional insight not apparent from the APS alone. For *FD*, the APS suggests a marginal advantage of using an ML model over the B2 baseline. Yet, the PR-curves reveal that for recall levels above 0.7, the ML model, with the ability to capture non-linear interactions (Fuster et al., 2022) and utilising model selection and validation (see Section 3.1), offers no improvement over the RQ1-B2, which utilises a logistic regression model without model selection and validation. This insight is not evident from the APS metric alone. The implication of this additional insight from the PR-curves is that the ML model does not provide a better ability to discern those experiencing *FD* (*FD* =1) for higher recall levels. This would be relevant in scenarios where the cost of incorrectly predicting *FD* = 1 ($\hat{y}$ = 0 instead of $\hat{y}$ = 1) is considerably high. The particular specification of the costs would depend on the decision maker's decision-context, i.e., how the decision maker weights incorrectly predicting *FD* = 1 relative to incorrectly predicting *FD* = 0 (see for example Lahiri and Yang, 2013; Elliot and Timmermann, 2016). These predictive results are assessed based on previously unseen test set, which neither the ML model nor the RQ1-B2 baseline were trained on or utilised for model selection and validation, hence providing an unbiased evaluation of the predictive models' performance.

[14]As mentioned earlier, the RQ1-B2 utilises the same baseline as B2 but presents it as a curve across various thresholds rather than as a summary single-value APS.

[15]APS is a summary measure of PR-curves as described in Section 3.2.

Second, as previously mentioned, the absolute predictive performance of ML models for households with no prior CFD appears to be low for both *FD* and *Debts* when assessed on the test set. In addition, although there is a clear improvement of ML models over the RQ1-B1 baseline for both measures of CFD, there appears to be no improvement over the RQ2-B2 baseline. For these reasons, further discussion on ML models' performance relative to baselines for this subsample is not warranted, as it does not provide additional insights.

Third, using ML to predict CFD for the subsample with prior CFD shows improvement over both the RQ1-B1 and RQ2-B2 baselines for *FD* and *Debts* when assessed on the test set. The ML's model improvement over the RQ1-B1, is a notable finding, given that the RQ1-B1 baseline (prior CFD subsample base rate) is a relatively good predictor of CFD persistence, with 70.2% and 50.6% of households experience persistence of *FD* and *Debts*, respectively, from one period to the next. This result aligns with the findings of Brown, Ghosh, and Taylor (2014), who provide empirical evidence that CFD is persistent over time. Therefore, the improvement of ML models over the RQ1-B1 baseline demonstrates that ML models can substantially improve forward-looking predictions of CFD beyond what is inferred from prior CFD alone, which is a relatively strong predictor. This implies that ML models could be used to improve borrowing-decisions and more effectively target support measures to households experiencing CFD in a given period (as forward-looking predictions are made ex-ante at time t, "prior" CFD amounts to "current" CFD when making predictions for t+1). Similarly, ML models also outperform the RQ1-B2 baseline for this subsample for *FD* and *Debts*. This demonstrates that ML models have better predictive ability compared to a logistic regression that does not utilise model selection and validation, nor has the ability to capture non-linear interactions (Fuster et al., 2022).

The methodological approach of plotting PR-curves for the two subsamples offers a framework to evaluate the heterogeneity of CFD predictions in future studies on the predictability of CFD. This advances the understanding of ML predictions of CFD, by demonstrating the differences in predictive accuracy across households depending on their prior experience of CFD, which is not apparent when simply predicting CFD on average (as shown in Table 5 in Section 3.3). Moreover, the fact that this heterogeneity is present for both *FD* and *Debts*, suggests that it is a consistent feature across various definitions of CFD. In addition, this is relevant from a policy perspective, given this study's findings that ML generates significantly lower accuracy in predicting new (episodic) cases of CFD compared to predicting persistence of CFD. Hence, this approach offers a comprehensive and meaningful way to evaluate the accuracy of ex-ante predictions of CFD, by assessing whether a given ML model can improve predictability among households with no prior CFD. This contributes to the existing literature, which has not explored heterogeneity of CFD predictions across households depending on their prior experience of CFD (see Salganik et al., 2019, 2020) Nevertheless, in its examination of predictive heterogeneity, this study broadly aligns with other literature exploring heterogeneity in predictions in context such as criminal justice (Babii et al., 2020) and consumer loan defaults (Fuster et al., 2022).

5.3 Discussion: Budget-PPV

The findings in Section 4.3 suggest that using budget-PPV curves — in the context of ex-ante CFD predictions — can be a useful tool for budget-constrained financial counselling agency, which can accommodate only a subset of households when organising a finance management course. For the full sample of households, *ML* outperforms both baselines for *FD* and *Debts*. This is demonstrated by the comparatively higher levels of PPV at all relevant budgets for *ML* compared to the two baselines. The improvement over the RQ2-B1 baseline reflects a better performance compared to an absolute minimum baseline that represents the base rate for the full sample. On the other hand, the improvement of *MLs*' performance over the RQ2-B2(ML) signifies the benefits of applying the budget-PPV decision context. This decision-context involves ranking households based on their predicted probability of experiencing CFD. Thus, *ML* improves predictive performance relative to a setup that utilises the same underlying ML model (GB and RF models for *FD* and *Debts*, respectively), but does not consider the budget-PPV decision-context of the counselling agency. This implies that a financial counselling agency utilising *ML* can better target the course to households most at risk of experiencing future CFD relative to both the RQ2-B1 and RQ2-B2(ML) baselines for *FD* and *Debts*. This contributes to the economic literature on consumer financial protection, as the predictability of CFD for a subset of the relevant population has not been previously considered in this context. However, similar decision-contexts have been considered in prior studies, such as in the context of predicting futile hip replacement surgeries (Kleinberg et al., 2015) and predicting high growth enterprises (Hyytinen et al., 2023).

In discussing the budget-PPV curves for the households with no prior CFD, it can be deduced that on average *ML* outperforms both baselines for *FD* and *Debts*. However, the improvement in *MLs*' budget-PPV performance over that of the RQ2-B2(ML) baseline is not statistically significant, as evidenced by the overlapping confidence intervals. On the other hand, *ML* outperforms the RQ2-B1 baseline, as the confidence intervals of the of *MLs*' budget-PPV curves do not overlap with the RQ2-B1 baseline's curve. Hence, we can conclude that for both *FD* and *Debts*, *ML* improves predictions for the relevant budgets compared to the RQ2-B1 baseline. This implies that a financial counselling agency can better target the course to households at all the considered budget levels compared to randomly selecting a corresponding subset of households from those with no prior CFD. Another notable observation is that the *ML* budget-PPV curves for this subsample are considerably lower than the corresponding budget-PPV curves for the full sample and subsample with prior *FD* and *Debts*. These results are also consistent with the general trend observed in Section 3.4, where the PR-curves show a significant drop in predictive performance for the subsample with no prior CFD relative to both the full sample and the subsample with prior CFD. Arguably, these budget-PPV curves could still be used to pre-screen households for eligibility to the course. Then, the agency could apply their own expertise to additionally screen only those households that were predicted to experience CFD within the set budget limit, i.e., pre-screened. A similar double-screening scenario has been considered in the context of predicting high-growth enterprises for venture deal-

sourcing (Hyytinen et al., 2023), and more generally as a decision aid for judges in remand or release decisions based on a defendant's future crime prediction (Kleinberg et al., 2018; Babii et al., 2020). Overall, the budget-PPV curves for the subsample with no prior CFD demonstrate that making accurate forward-looking CFD predictions for new (episodic) cases of CFD is challenging, even when considering predictability for small subsets of households. This finding adds to the limited literature on CFD predictions by highlighting the heterogeneity in CFD predictions. In particular, it reveals the lower predictive accuracy of forward-looking CFD predictions for households with no prior CFD relative to the full sample and the subsample with prior CFD in the context of budget-PPV. Salganik et al. (2019, 2020) findings noted the relatively low accuracy of CFD predictions. However, these studies did not examine heterogeneity in predictability for the different subsamples. This study addresses this gap. It brings attention to the specific challenge of making accurate CFD predictions for the subsample with no prior CFD, which also reduces the overall predictive accuracy of CFD. Hence, this study presents an opportunity for future research to develop new models tailored to improving the accuracy of forward-looking CFD predictions by improving predictability for the subsample with no prior CFD.

Turning our attention to the subsample of households that experience prior CFD, it is worth noting that predicting CFD for these households involves forecasting whether CFD will persist into the next period (as opposed to being episodic, as in the case of no prior CFD). For this subsample, *MLs*' budget-PPV curves demonstrate a consistently high, almost perfect PPV across all the considered budgets for both *FD* and *Debts*. In addition, the improvement in *MLs*' budget-PPV curves compared to the RQ2-B1 and RQ2-B2(ML) baselines' is, overall, statistically significant. *MLs*' improvement over the RQ2-B1 baseline implies that *ML* can improve over a predictive model that makes forward-looking predictions of CFD based on prior CFD. As discussed earlier, CFD persists from one period to another (Brown, Ghosh, and Taylor, 2014). This implies that prior period CFD is indicative of a high likelihood of next period CFD. Hence, the RQ2-B1 baseline provides an informative benchmark against which to compare the predictive performance of *ML* (Hofman et al., 2021). In terms of the comparative performance of *ML* against RQ2-B2(ML), this improvement signifies the benefit of applying the budget-PPV decision-context. These results suggest that an effective targeting of the finance management course to households most at risk of future CFD can be achieved. Moreover, the ability to predict *Debts* accurately across the considered budgets also implies that programs that emphasise debt-management/ restructuring can be organised for those predicted to experience future (persistent) indebtedness problems. The suggested policy approach aligns with the approach advocated by Townley-Jones, Griffiths, and Bryant (2008), which calls for tailoring specific financial literacy programs to consumers identified as "chronic" debtors.

6 CONCLUSION

In this study, I examine forward-looking predictions of CFD using *FD* and *Debts* measures in three contexts. First, I find that ML can ex-ante predict CFD more accurately rel-

ative to a baseline model with only prior CFD as predictor, and a second model with prior CFD and additional predictors. Second, I demonstrate, using PR-curves, that predicting CFD is more challenging for households with no prior CFD than it is for households with prior CFD. This implies that the ML models I use are better at predicting persistence than episodical cases of CFD. The PR-curves graphical tool provides a framework for future research to evaluate performance of predictive models in predicting CFD, while considering the potential heterogeneity. Third, I consider the decision-context of a public financial counselling agency that aims to target a finance management course to households at the most risk of experiencing future CFD. The application of the budget-PPV approach here demonstrates that predicting persistence is almost perfectly precise (measured by PPV) for the tightest budget levels. On the other hand, ML predictions for those with no prior CFD appear to be less precise for the considered budgets.

The expectation is that these contributions would further inform the scholarly and practitioner discussion on consumer financial protection, by highlighting caveats, such as the heterogeneity in predictability between those with and without prior CFD, and introducing new approaches, such as the budget-PPV curves in the public agency's context. Therefore, these insights would hopefully contribute to the literature beyond the traditional focus on associative or causal relationships in addressing the challenge of consumer financial distress.

6.1 Limitations

Subjectivity bias in CFD outcome variables

Some of the indicators (outlined in Table 1) used to construct the *FD* and *Debts* measures are subjective in nature. As they are based on households' self-assessment of the degree of their own financial difficulty. Although Betti et al. (2007), using EU data, do not find a substantial group of households that hide their financial difficulties from surveys, the subjective indicators rely on households' own interpretations of what expenditure difficulties or debt burden might mean (Reijo, 2000, cited in Raijas, Lehtinen, and Leskinen, 2010, p.218; Bertrand and Mullainathan, 2001). This in particular is pertinent to the *POI* and *EX* indicators. Future research could try to ascertain the accuracy of households CFD reporting.

Data drift

The CFD predictions are generated using ML models trained on data obtained during the mostly relatively low interest rate period of 2008-2014. Hence, these ML predictions might not generalise well to new data from a period of high interest rates. However, this limitation can be mitigated by re-training the ML models on new data from the relevant period to ensure that the ML predictions maintain their accuracy.

6.2 Implications for future research

Using additional predictors to improve ML predictions of CFD

My findings show that ML predictions of CFD outperform baselines B1 and B2 on average, especially when evaluated using APS. However, the PR-curves show significantly lower predictive accuracy for households with no prior CFD than for those with prior CFD across both *FD* and *Debts*, with the same pattern in budget-PPV curves. This calls for consideration of additional predictors documented in the literature to be associated with CFD: financial literacy (Stango and Zinman, 2009; Campbell et al., 2011; Lusardi and Tufano, 2015; Hyytinen and Putkuri, 2018); unsecured consumer debt-to-income ratio (Del Rio and Young, 2008; Gathergood and Guttman-Kenney, 2016), as the IDS consumer debt measure also includes likely secured debt from holiday home purchases; and impulsive expenditure behaviour (Ottaviani and Vandone, 2011; Gathergood, 2012) or gambling habits (Muggleton et al., 2021). The proposed PR-curves graphical tool offers a framework to evaluate whether such additional predictors improve the predictive accuracy of future CFD across the respective subsamples.

Bridging causal and ML predictive studies of CFD

Prior studies examined factors contributing to CFD and measures to mitigate its risk and consequences. This study augments this literature by investigating CFD predictability from the consumer's perspective. In particular, the budget-PPV decision-context considers using ML predictions to target a financial management course to households at greatest risk of future CFD, similar to approaches predicting futile hip replacement surgeries (Kleinberg et al., 2015) and high growth enterprises (Hyytinen et al., 2023). Nevertheless, a pure predictive approach does not identify which consumers would benefit most from the course. This would require estimating heterogeneity in the effect of the course, which could reflect differences in participants' receptiveness to its learning outcomes (Athey, 2017, p. 483). Such estimation is not possible using the IDS dataset, as surveyed households were not assigned to a financial management course. For further details on targeting interventions based on predictive ML models and heterogeneous treatment effects see Athey, Keleher and Spiess (2023).

Consideration of CFD predictability in other decision-contexts

In this study, I considered the decision-context of a public financial counselling agency. However, another possible avenue for future research is to consider how consumer borrowing decision can be improved with ML predictions of future CFD. This would involve modelling consumers' individual trade-offs in making ex-ante inaccurate decisions of whether to borrow or not, when faced with a future prospect of experiencing CFD. Babii et al. (2020) integrates estimates of the costs and benefits of release and remand decisions of each defendant when applying ML methods to predict future crime. Therefore, ML predictions in this case are tailored to the individual risk assessment of each defendant.

A similar approach could be possible for CFD predictions, although it would require the modelling of individual costs and benefits of each decision conditional on the ex-ante unknown outcome of CFD. Note that in such studies, the decision is assumed to be informed by the prediction.

Considering other measures of CFD, other national datasets, and longer time horizons

As previously mentioned, there are multiple definitions of CFD in the economic and financial literature. In this study, I used *Debts* (measuring indebtedness) and *FD* (a comprehensive measure of CFD). Future studies could consider more objective measures unavailable in the IDS dataset, for instance a household being in arrears, which is a direct indicator of a household not being able to meet their financial commitments on time. Overall, the choice of the measure could be driven by the underlying policy intervention that would aim to reduce or mitigate the specific type of CFD. However, it should be noted that the specific intervention measures themselves are an active area of research. Furthermore, given that there is little prior research on CFD predictability, it would be of interest to utilise other national datasets to examine the predictability of CFD. In particular, assessing whether the accuracy of CFD predictions is relatively high on average or whether there is significant heterogeneity in CFD predictions. Finally, I consider the predictability of CFD one year in advance using one-year lagged predictors. Future studies could also look at longer-term predictability.

6.3 Implications for policy and practice

The budget-PPV curves facilitate the identification of the subset of households at the highest risk of future CFD, qualifying them for voluntary participation in the public agency's program. If such a program were to be implemented in practice (for similar implementation in the context of US army enlistees see Skimmyhorn, 2016). A further step could then be to estimate the effectiveness of the program on participants' experience of CFD. Furthermore, it is likely that different intervention programs are required depending on whether CFD is persistent (prior CFD) or episodic (no prior CFD) (Townley-Jones, Griffiths, and Bryant, 2008). Hence, assessing the effectiveness of a financial education program while accounting for the heterogeneity between the two groups of households might be beneficial.

REFERENCES

Agarwal, S., Mikhed, V. and Scholnick, B. (2020) 'Peers' income and financial distress: Evidence from lottery winners and neighboring bankruptcies', *The Review of Financial Studies*, 33(1), pp. 433–472.

Aggarwal, C.C. (2016) *Recommender Systems: The Textbook*. Springer.

Albanesi, S. and Vamossy, D.F. (2019) *Predicting Consumer Default: A Deep Learning Approach*. Technical report, National Bureau of Economic Research Working paper no. 26165.

Alpaydin, E. (2014) *Introduction to machine learning*. Third edition. Cambridge, Massachusetts; London, England: The MIT Press.

Athey, S. (2017) 'Beyond prediction: Using big data for policy problems', *Science*, 355(6324), pp. 483–485.

Athey, S. and Imbens, G.W. (2019) 'Machine learning methods that economists should know about', *Annual Review of Economics*, 11(1), pp. 685–725.

Athey, S., Keleher, N. and Spiess, J. (2023) *Machine learning who to nudge: causal vs predictive targeting in a field experiment on student financial aid renewal*. arXiv.org. doi:10.48550/arxiv.2310.08672.

Athreya, K., Mustre-del-Río, J. and Sánchez, J.M. (2019) 'The persistence of financial distress', *The Review of Financial Studies*, 32(10), pp. 3851–3883.

Attanasio, O.P. and Weber, G. (2010) 'Consumption and saving: models of intertemporal allocation and their implications for public policy', *Journal of Economic Literature*, 48(3), pp. 693–751.

Babii, A., Chen, X., Ghysels, E. and Kumar, R. (2020) *Binary Choice with Asymmetric Loss in a Data-Rich Environment: Theory and an Application to Racial Justice*. arXiv.org.

Benjamin, D., Cooper, K., Heffetz, O. and Kimball, M. (2020) 'Self-reported wellbeing indicators are a valuable complement to traditional economic indicators but are not yet ready to compete with them', *Behavioural Public Policy*, 4(2), pp. 198–209.

Berk, R. (2012) *Criminal Justice Forecasts of Risk: A Machine Learning Approach*. Springer.

Bertrand, M. and Morse, A. (2011) 'Information disclosure, cognitive biases, and payday borrowing', *The Journal of Finance*, 66(6), pp. 1865–1893.

Bertrand, M. and Mullainathan, S. (2001) 'Do people mean what they say? Implications for subjective survey data', *American Economic Review*, 91(2), pp. 67–72.

Betti, G., Dourmashkin, N., Rossi, M. and Yin, Y.P. (2007) 'Consumer over-indebtedness in the EU: measurement and characteristics', *Journal of Economic Studies*, 34(2), pp. 136–156.

Brown, S., Ghosh, P. and Taylor, K. (2014) 'The existence and persistence of household financial hardship: A Bayesian multivariate dynamic logit framework', *Journal of Banking & Finance*, 46(1), pp. 285–298.

Brunetti, M., Giarda, E. and Torricelli, C. (2016) 'Is financial fragility a matter of illiquidity? An appraisal for Italian households', *Review of Income and Wealth*, 62(4), pp. 628–649.

Burrows, R. and Ford, J. (1997) 'Who needs a safety-net? The social distribution of mortgage arrears in England', *Housing Finance*, 34, pp. 17–24.

Campbell, J.Y., Jackson, H.E., Madrian, B.C. and Tufano, P. (2011) 'Consumer financial protection', *Journal of Economic Perspectives*, 25(1), pp. 91–114.

Collins, D. (2008) 'Debt and household finance: evidence from the financial diaries', *Development Southern Africa*, 25(4), pp. 469–479.

D'Alessio, G. and Lezzi, S. (2013) *Household over indebtedness: definition and measurement with Italian data*. Bank of Italy, Occasional Papers (Questioni di economia e finanza), (149).

Del Rio, A. and Young, G. (2008) 'The impact of unsecured debt on financial pressure among British households', *Applied Financial Economics*, 18(15), pp. 1209–1220.

DeVaney, S.A. and Lytton, R.H. (1995) 'Household insolvency: A review of household debt repayment, delinquency, and bankruptcy', *Financial Services Review*, 4(2), pp. 137–156.

Disney, R., Gathergood, J. and Weber, J. (2015) 'Credit counseling: a substitute for consumer financial literacy?', *Journal of Pension Economics & Finance*, 14(4), pp. 466–491.

Drentea, P. and Lavrakas, P.J. (2000) 'Over the limit: the association among health, race and debt', *Social Science & Medicine*, 50(4), pp. 517–529.

Duygan-Bump, B. and Grant, C. (2009) 'Household debt repayment behaviour: what role do institutions play?', *Economic Policy*, 24(57), pp. 108–140.

Efron, B. and Hastie, T. (2016) *Computer age statistical inference: algorithms, evidence, and data science*. Cambridge University Press.

Elliot, G. and Timmermann, A. (2016) *Economic Forecasting*. Princeton University Press.

Eurofound (2020) *Addressing household over-indebtedness*. Publications Office of the European Union, Luxembourg.

Fuster, A., Goldsmith-Pinkham, P., Ramadorai, T. and Walther, A. (2022) 'Predictably Unequal? The Effects of Machine Learning on Credit Markets', *The Journal of Finance*, 77(1), pp. 5–47.

Galenianos, M. and Gavazza, A. (2022) 'Regulatory interventions in consumer financial markets: The case of credit cards', *Journal of the European Economic Association*, 20(5), pp. 1897–1932.

Gareth, J., Daniela, W., Trevor, H. and Robert, T. (2013) *An introduction to statistical learning: with applications in R*. Springer.

Garz, S., Giné, X., Karlan, D., Mazer, R., Sanford, C. and Zinman, J. (2020) *Consumer protection for financial inclusion in low and middle income countries: Bridging regulator and academic perspectives (No.* w28262). National Bureau of Economic Research.

Gathergood, J. (2012) 'Self-control, financial literacy and consumer over-indebtedness', *Journal of Economic Psychology*, 33(3), pp. 590–602.

Gathergood, J. and Guttman-Kenney, B. (2016) *Can we predict which consumer credit users will suffer financial distress?* FCA Occasional Paper No.20.

Georgarakos, D., Haliassos, M. and Pasini, G. (2014) 'Household debt and social interactions', *The Review of Financial Studies*, 27(5), pp. 1404–1433.

Geurts, P., Ernst, D. and Wehenkel, L. (2006) 'Extremely randomized trees', *Machine Learning*, 63(1), pp. 3–42.

Giustinelli, P., Manski, C.F. and Molinari, F. (2022) 'Precise or imprecise probabilities? Evidence from survey response related to late-onset dementia', *Journal of the European Economic Association*, 20(1), pp. 187–221.

Granger, C.W. and Machina, M.J. (2006) 'Forecasting and decision theory', in *Handbook of Economic Forecasting*, 1, pp. 81–98.

Granger, C.W. and Pesaran, M.H. (2000) 'Economic and statistical measures of forecast accuracy', *Journal of Forecasting*, 19(7), pp. 537–560.

Hastie, T., Tibshirani, R. and Friedman, J.H. (2009) *The elements of statistical learning: data mining, inference, and prediction*. Springer.

Herlocker, J.L., Konstan, J.A., Terveen, L.G. and Riedl, J.T. (2004) 'Evaluating collaborative filtering recommender systems', *ACM Transactions on Information Systems (TOIS)*, 22(1), pp. 5–53.

Hofman, J.M., Watts, D.J., Athey, S., Garip, F., Griffiths, T.L., Kleinberg, J., Margetts, H., Mullainathan, S., Salganik, M.J., Vazire, S. and Vespignani, A. (2021) 'Integrating explanation and prediction in computational social science', *Nature*, 595(7866), pp. 181–188.

Hyytinen, A., Rouvinen, P., Pajarinen, M. and Virtanen, J. (2023) 'Ex ante predictability of rapid growth: A design science approach', *Entrepreneurship Theory and Practice*, 47(6), pp. 2465–2493. doi:10.1177/10422587221128268.

Hyytinen, A. and Putkuri, H. (2018) 'Household optimism and overborrowing', *Journal of Money, Credit and Banking*, 50(1), pp. 55–76.

James, G., Witten, D., Hastie, T., Tibshirani, R. and Taylor, J. (2023) *An introduction to statistical learning: With applications in python*. Springer Text in Statistics.

Johnson, K.W. and Li, G. (2010) 'The debt-payment-to-income ratio as an indicator of borrowing constraints: Evidence from two household surveys', *Journal of Money, Credit and Banking*, 42(7), pp. 1373–1390.

Kleinberg, J., Lakkaraju, H., Leskovec, J., Ludwig, J. and Mullainathan, S. (2018) 'Human decisions and machine predictions', *The Quarterly Journal of Economics*, 133(1), pp. 237–293.

Kleinberg, J., Ludwig, J., Mullainathan, S. and Obermeyer, Z. (2015) 'Prediction policy problems', *American Economic Review*, 105(5), pp. 491–95.

Kondratjeva, O., Roll, S.P., Bufe, S. and Grinstein-Weiss, M. (2021) 'Using financial tips to guide debt repayment: Experimental evidence from low-and moderate-income tax filers', *Journal of Behavioral and Experimental Economics*, 90(1), pp. 101612.

Kreiner, C.T., Leth-Petersen, S. and Willerslev-Olsen, L.C. (2020) 'Financial trouble across generations: Evidence from the universe of personal loans in Denmark', *The Economic Journal*, 130(625), pp. 233–262.

Kuhn, M. and Johnson, K. (2013) *Applied predictive modeling*. Springer.

Kuhnen, C.M. and Melzer, B.T. (2018) 'Noncognitive abilities and financial delinquency: The role of self-efficacy in avoiding financial distress', *The Journal of Finance*, 73(6), pp. 2837–2869.

Lahiri, K. and Yang, L. (2013) 'Forecasting binary outcomes', in *Handbook of Economic Forecasting*, 1(2), pp. 1025–1106.

Lieli, R.P. and Springborn, M. (2013) 'Closing the gap between risk estimation and decision making: efficient management of trade-related invasive species risk', *Review of Economics and Statistics*, 95(2), pp. 632–645.

Lieli, R.P. and White, H. (2010) 'The construction of empirical credit scoring rules based on maximization principles', *Journal of Econometrics*, 157(1), pp. 110–119.

Lusardi, A. and Mitchell, O.S. (2014) 'The economic importance of financial literacy: Theory and evidence', *Journal of Economic Literature*, 52(1), pp. 5–44.

Lusardi, A., Schneider, D.J. and Tufano, P. (2011) *Financially fragile households: Evidence and implications*. National Bureau of Economic Research, Working Paper No.17072.

Lusardi, A. and Tufano, P. (2015) 'Debt literacy, financial experiences, and overindebtedness', *Journal of Pension Economics & Finance*, 14(4), pp. 332–368.

McCarthy, Y. (2011) *Behavioural Characteristics and Financial Distress*. European Central Bank, Working Paper No.1303.

Muggleton, N., Parpart, P., Newall, P., Leake, D., Gathergood, J. and Stewart, N. (2021) 'The association between gambling and financial, social and health outcomes in big financial data', *Nature Human Behaviour*, 5(3), pp. 319–326.

Muttilainen, V. and Reijo, M. (2007) *Kotitalouksien maksuvaikeudet 2002–2005 [Households' payment problems 2002–2005]*. National Research Institute of Legal Policy. Research Communications 81.

Netzer, O., Lemaire, A. and Herzenstein, M. (2019) 'When words sweat: Identifying signals for loan default in the text of loan applications', *Journal of Marketing Research*, 56(6), pp. 960–980.

Ottaviani, C. and Vandone, D. (2011) 'Impulsivity and household indebtedness: Evidence from real life', *Journal of Economic Psychology*, 32(5), pp. 754–761.

Oxera (2004) *Are UK households Over-indebted?* Report Commissioned by the Association for Payment Clearing Services (APACS), the British Bankers' Association (BBA), the Consumer Credit Association (CCA), and the Finance and Leasing Association (FLA). London: Oxera Consulting.

Parise, G. and Peijnenburg, K. (2019) 'Noncognitive abilities and financial distress: evidence from a representative household panel', *The Review of Financial Studies*, 32(10), pp. 3884–3919.

Raijas, A., Lehtinen, A.R. and Leskinen, J. (2010) 'Over-indebtedness in the Finnish consumer society', *Journal of Consumer Policy*, 33(3), pp. 209–223.

Saito, T. and Rehmsmeier, M. (2015) 'The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets', *PLOS ONE*, 10(3), pp. e0118432-e0118432.

Salganik, M.J., Lundberg, I., Kindel, A.T., Ahearn, C.E., Al-Ghoneim, K., Almaatouq, A., Altschul, D.M., Brand, J.E., Carnegie, N.B., Compton, R.J. and Datta, D. (2020) 'Measuring the predictability of life outcomes with a scientific mass collaboration', *Proceedings of the National Academy of Sciences*, 117(15), pp. 8398–8403.

Salganik, M.J., Lundberg, I., Kindel, A.T. and McLanahan, S. (2019) 'Introduction to the special collection on the fragile families challenge', *Socius*, (1)5, pp. 1–21.

Skimmyhorn, W. (2016) 'Assessing financial education: Evidence from boot camp', *American Economic Journal: Economic Policy*, 8(2), pp. 322–43.

Stango, V. and Zinman, J. (2009) 'Exponential growth bias and household finance', *The Journal of Finance*, 64(6), pp. 2807–2849.

Stango, V. and Zinman, J. (2011) 'Fuzzy math, disclosure regulation, and market outcomes: Evidence from truth-in-lending reform', *The Review of Financial Studies*, 24(2), pp. 506–534.

Statistics Finland (2022) *Survey on income and living conditions: General information about the data collection*. Available at: https://www.stat.fi/keruu/tuel/yleista_en.html (Accessed: 30 June 2022).

Sweet, E., Nandi, A., Adam, E.K. and McDade, T.W. (2013) 'The high price of debt: Household financial debt and its impact on mental and physical health', *Social Science & Medicine*, 91(1), pp. 94–100.

Townley-Jones, M., Griffiths, M. and Bryant, M. (2008) 'Chronic consumer debtors: the need for specific intervention', *International Journal of Consumer Studies*, 32(3), pp. 204–210.

Varian, H.R. (2014) 'Big data: New tricks for econometrics', *Journal of Economic Perspectives*, 28(2), pp. 3–28.

Xu, Y., Beller, A.H., Roberts, B.W. and Brown, J.R. (2015) 'Personality and young adult financial distress', *Journal of Economic Psychology*, 100(51), pp. 90–100.

APPENDIX

Appendix A: Descriptive Statistics for Predictors

Table A1: Descriptive statistics for continuous and binary predictors (training set)

Predictor variables (first period)					
Variable	**Mean**	**Median**	**St. dev.**	**Min**	**Max**
FD (1 indicates the occurrence of *FD*)	0.3091	0.0000	0.4621	0.0000	1.0000
Debts (1 indicates the occurrence of *Debts*)	0.0767	0.0000	0.2662	0.0000	1.0000
EX (1 indicates the occurrence of *EX*)	0.2335	0.0000	0.4231	0.0000	1.0000
RDP (1 indicates the occurrence of *RDP*)	0.0500	0.0000	0.2179	0.0000	1.0000
*DPP (*1 indicates the occurrence of *DPP)*	0.0381	0.0000	0.1915	0.0000	1.0000
BPP (1 indicates the occurrence of *BPP*)	0.1764	0.0000	0.3811	0.0000	1.0000
POI (1 indicates the occurrence of *POI*)	0.0270	0.0000	0.1622	0.0000	1.0000
Monthly housing costs to monthly income ratio	0.3867	0.1492	13.8109	0.0000	2604
Yearly income (EUR thousand)	37.5712	31.3790	30.8232	0.0000	2893.3880
Income generating debt-to-income ratio	0.0147	0.0000	0.3360	0.0000	99.4563
Mortgage debt-to-income ratio	1.4667	0.0000	56.3125	0.0000	4436.8333
Student debt-to-income ratio	0.0091	0.0000	0.6600	0.0000	275.0000
Consumption & other debts-to-income ratio	0.2863	0.0000	10.7064	0.0000	1499.1667
Yearly interest payments on consumption & other debts (non-mortgage & non-income generating debts) to income ratio	0.0112	0.0000	0.3447	0.0000	45.0323
Assistance received from another household for studies (>100 €)	20.9350	0.0000	211.2657	0.0000	16000.0000
Assistance received from another household for bill payments (€)	35.4751	0.0000	373.0164	0.0000	30000.0000
Assistance received from another household for other purposes (€)	65.3323	0.0000	417.1798	0.0000	15000.0000
Savings (1 = if savings left after expenses, otherwise = 0)	0.5548	1.0000	0.4970	0.0000	1.0000
Age	51.2205	51.0000	17.7072	16.0000	98.0000
Gender (1 = male)	0.5825	1.0000	0.4932	0.0000	1.0000
Language (1 = Finnish)	0.9508	1.0000	0.2162	0.0000	1.0000
Retired (1 = retired)	0.2737	0.0000	0.4458	0.0000	1.0000
Number of Children	0.3770	0.0000	0.8561	0.0000	10.0000

Table A2: Descriptive statistics for multi-category predictors (training set)

Proportion (%) of households in different categories of predictor variables (first period)

Education level	
No degree, or just primary education	25.15
Secondary education	40.30
Lower university level education	23.01
Higher university level degree (Master's or PhD)	11.54

Development of a household's financial situation compared to last year	
Changes significantly for the better	3.83
Changes somewhat for the better	17.24
Remains unchanged	56.74
Changes somewhat to the worse	16.65
Changes Significantly to the worse	5.54

Marital Status	
unmarried	33.33
married	41.71
divorced	16.21
widowed	8.75

Burden of housing costs	
Large burden	27.02
Moderate burden	53.29
No burden at all	19.69

Household structure	
single households, age under 25 years	3.68
single households, age 25-44 years	9.50
single households, age 45-64 years	12.97
single households, age 65+ years	13.67
childless couples, age of reference person under 25 years	1.90
childless couples, age of reference person 25-44 years	6.23
childless couples, age of reference person 45-64 years	12.91
childless couples, reference person age 65+ years	9.97
two-parent with children households, the youngest child under 3 years of age	6.01
two-parent households, children 3 to 6 years old	3.97
two-parent households, age 7-12 years	4.30
households with two parents, the youngest child aged 13-17	2.71
single-parent households with children, the youngest child under 3 years of age	0.39
single-parent households with children, age 3-6 years for the youngest child	0.64
single parent households with children, age of youngest child 7-12 years	1.01
single-parent households with children, age 13-17 years for the youngest child	0.92
married or unmarried couple and both minor and adult children or adult children only	5.45
an adult living alone and both minors and adult children or adult children only	1.50
Other households (e.g., three-generation households, households without a family))	2.28

Expectations on development of financial situation	
Changes significantly for the better	3.25
Changes somewhat for the better	16.12
Remains unchanged	62.11
Changes somewhat to the worse	12.71
Changes Significantly to the worse	2.76
Household cannot assess	3.05

Does the household have a car?	
Yes	74.92
No, because can't afford it	12.50
No for other reason	12.58

Table A2: Descriptive statistics for multi-category predictors (continued)

Does the household have a computer?	
Yes	80.55
No, because can't afford it	3.51
No for other reason	15.95

Socioeconomic status	
Agricultural entrepreneurs or employers	1.44
Other entrepreneur or employer	4.99
Freelancers	0.77
Senior white-collar employees	17.55
Junior managers	2.98
Junior white-collar employees	14.01
Blue-collar workers	17.17
Students and school pupils	3.86
Retired agricultural entrepreneurs	2.68
Other retired entrepreneurs	2.74
Retired senior civil servants	4.60
Retired junior civil servants	9.09
Retired employees	11.16
Other pensioners	1.08
Taking care of own household	0.51
Long term unemployed	4.62
Not elsewhere classified	0.75

Region of residence	
Uusimaa	27.81
Varsinais-Suomi	8.83
Satakunta	4.54
Kanta-Häme	3.35
Pirkanmaa	9.50
Päijät-Häme	3.76
Kymenlaakso	3.52
Etelä-Karjala	2.48
Etelä-Savo	3.00
Pohjois-Savo	4.57
Pohjois-Karjala	3.17
Keski-Suomi	5.28
Etelä-Pohjanmaa	3.32
Pohjanmaa	2.93
Keski-Pohjanmaa	1.13
Pohjois-Pohjanmaa	6.73
Kainuu	1.49
Lappi	3.24
Itä-Uusimaa	0.85
Ahvenanmaa	0.50

Housing situation	
Own detached house on own plot	34.03
own detached house on a rented plot	3.58
own apartment (ownership of shares in a housing company)	31.04
main tenant apartment	28.79
subtenant apartment	0.40
right-of-occupancy housing or co-ownership housing	1.50
service house for rent, support apartment or boarding school	0.61
something else, such as a caravan or boat, tent, or a mobile structure	0.04

Reason for indebtedness	
Not over-indebted	97.30
Becoming unemployed, laid off or on forced leave	0.71
Business difficulties	0.27
reduction of paid working time or income earned from paid work	0.18
Own or a family member's illness, accident, or death of a family member	0.27
The guarantee liability becoming payable	0.10
Dissolution of marriage or cohabitation	0.22
Unexpected expenses	0.47
Other individual reason	0.48

Table A3: Descriptive statistics for macroeconomic predictors (training set)

Macroeconomic Predictors (measured in % at the first period)					
Variable	**Mean**	**Median**	**St. dev.**	**Min**	**Max**
Inflation	2.5400	2.1800	0.9943	1.2100	3.9000
Real GDP Growth	-0.3962	-0.4000	3.3694	-8.1000	3.2000
Unemployment Rate	7.8016	8.2000	0.7844	6.4000	8.7000

Table A4: Descriptive statistics for continuous and binary predictors (test set)

Predictor variables (first period)					
Variable	**Mean**	**Median**	**St. dev.**	**Min**	**Max**
FD (1 indicates the occurrence of *FD*)	0.3395	0.0000	0.4735	0.0000	1.0000
Debts (1 indicates the occurrence of *Debts*)	0.0859	0.0000	0.2802	0.0000	1.0000
EX (1 indicates the occurrence of *EX*)	0.2509	0.0000	0.4336	0.0000	1.0000
RDP (1 indicates the occurrence of *RDP*)	0.0606	0.0000	0.2387	0.0000	1.0000
*DPP (*1 indicates the occurrence of *DPP)*	0.0380	0.0000	0.1912	0.0000	1.0000
BPP (1 indicates the occurrence of *BPP*)	0.2023	0.0000	0.4017	0.0000	1.0000
POI (1 indicates the occurrence of *POI*)	0.0294	0.0000	0.1689	0.0000	1.0000
Monthly housing costs to monthly income ratio	0.2316	0.1473	1.1283	0.0000	111.3333
Yearly income (EUR thousand)	39.7407	32.7910	31.9480	89.0000	748.7260
Income generating debt-to-income ratio	0.0096	0.0000	0.2944	0.0000	32.1861
Mortgage debt-to-income ratio	0.6652	0.0000	1.3667	0.0000	21.7615
Student debt-to-income ratio	0.0149	0.0000	0.0752	0.0000	0.7926
Consumption & other debts-to-income ratio	0.2355	0.0000	2.0420	0.0000	180.3596
Yearly interest payments on consumption & other debts (non-mortgage & non-income generating debts) to income ratio	0.0061	0.0000	0.0843	0.0000	8.5333
Assistance received from another household for studies	22.1795	0.0000	205.7732	0.0000	4800.0000
Assistance received from another household for bill payments	47.2744	0.0000	350.2193	0.0000	7000.0000
Assistance received from another household for other purposes	68.6083	0.0000	446.0115	0.0000	12000.0000
Savings (1 = if savings left after expenses, otherwise = 0)	0.5567	1.0000	0.4968	0.0000	1.0000
Age	51.8527	53.0000	18.0423	17.0000	95.0000
Gender (1 = male)	0.5645	1.0000	0.4953	0.0000	1.0000
Language (1 = Finnish)	0.9440	1.0000	0.2298	0.0000	1.0000
Retired (1 = retired)	0.2934	0.0000	0.4453	0.0000	1.0000
Number of Children	0.3528	0.0000	0.8784	0.0000	12.0000

Table A5: Descriptive statistics for multi-category predictors (test set)

Proportion (%) of households in different categories of predictor variables (first period)

Education level	
No degree, or just primary education	21.51
Secondary education	42.35
Lower university level education	22.96
Higher university level degree (Master's or PhD)	13.18

Development of a household's financial situation compared to last year	
Changes significantly for the better	3.77
Changes somewhat for the better	17.26
Remains unchanged	55.32
Changes somewhat to the worse	18.18
Changes Significantly to the worse	5.47

Marital status	
unmarried	33.94
married	40.91
divorced	16.03
widowed	9.11

Burden of housing costs	
Large burden	27.68
Moderate burden	54.04
No burden at all	18.28

Household structure	
single households, age under 25 years	3.52
single households, age 25-44 years	9.91
single households, age 45-64 years	12.98
single households, age 65+ years	14.73
childless couples, age of reference person under 25 years	2.38
childless couples, age of reference person 25-44 years	6.06
childless couples, age of reference person 45-64 years	11.86
childless couples, reference person age 65+ years	11.67
two-parent with children households, the youngest child under 3 years of age	5.26
two-parent households, children 3 to 6 years old	3.93
two-parent households, age 7-12 years	4.20
households with two parents, the youngest child aged 13-17	2.18
single-parent households with children, the youngest child under 3 years of age	0.45
single-parent households with children, age 3-6 years for the youngest child	0.46
single parent households with children, age of youngest child 7-12 years	0.89
single-parent households with children, age 13-17 years for the youngest child	0.88
married or unmarried couple and both minor and adult children or adult children only	5.53
an adult living alone and both minors and adult children or adult children only	1.29
Other households (e.g., three-generation households, households without a family))	1.82

Expectations on development of financial situation	
Changes significantly for the better	3.12
Changes somewhat for the better	17.39
Remains unchanged	62.58
Changes somewhat to the worse	11.76
Changes Significantly to the worse	2.73
Household can not assess	2.43

Does the household have a car?	
Yes	76.72
No, because can't afford it	12.86
No for other reason	10.42

Table A5: Descriptive statistics for multi-category predictors (continued)

Does the household have a computer?	
Yes	87.93
No, because can't afford it	1.79
No for other reason	10.28

Socioeconomic status	
Agricultural entrepreneurs or employers	1.20
Other entrepreneur or employer	4.37
Freelancers	0.96
Senior white-collar employees	17.08
Junior managers	1.88
Junior white-collar employees	14.97
Blue-collar workers	15.26
Students and school pupils	5.29
Retired agricultural entrepreneurs	1.90
Other retired entrepreneurs	2.36
Retired senior civil servants	5.18
Retired junior civil servants	10.51
Retired employees	10.73
Other pensioners	0.94
Taking care of own household	0.45
Long term unemployed	6.25
Not elsewhere classified	0.66

Region of residence	
Uusimaa	29.64
Varsinais-Suomi	8.54
Satakunta	4.22
Kanta-Häme	3.24
Pirkanmaa	9.69
Päijät-Häme	3.90
Kymenlaakso	3.31
Etelä-Karjala	2.71
Etelä-Savo	2.95
Pohjois-Savo	4.06
Pohjois-Karjala	3.26
Keski-Suomi	4.75
Etelä-Pohjanmaa	3.24
Pohjanmaa	3.10
Keski-Pohjanmaa	1.08
Pohjois-Pohjanmaa	7.04
Kainuu	1.45
Lappi	3.28
Itä-Uusimaa	0.00
Ahvenanmaa	0.54

Housing situation	
Own detached house on own plot	32.77
own detached house on a rented plot	4.20
own apartment (ownership of shares in a housing company)	28.97
main tenant apartment	30.98
subtenant apartment	0.78
right-of-occupancy housing or co-ownership housing	1.96
service house for rent, support apartment or boarding school	0.34
something else, such as a caravan or boat, tent, or a mobile structure	0.00

Reason for indebtedness	
Not over-indebted	97.06
Becoming unemployed, laid off or on forced leave	0.86
Business difficulties	0.29
reduction of paid working time or income earned from paid work	0.33
Own or a family member's illness, accident, or death of a family member	0.30
The guarantee liability becoming payable	0.07
Dissolution of marriage or cohabitation	0.10
Unexpected expenses	0.40
Other individual reason	0.59

Table A6: Descriptive statistics for macroeconomic predictors (test set)

Macroeconomic Predictors (measured in % at the first period)					
Variable	**Mean**	**Median**	**St. dev.**	**Min**	**Max**
Inflation	0.1133	-0.1500	0.2650	-0.1500	0.3800
Real GDP Growth	1.6427	0.5000	1.1400	0.5000	2.8000
Unemployment Rate	9.1019	9.4000	0.3000	8.8000	9.4000

Appendix B: Expanding Window Cross-Validation (EWCV)

EWCV partitions the temporally ordered training set into K evenly distributed train-validation subsamples or folds, which equates to K-1 iterations. In this study I use K = 8. Thus, an ML model is first fitted on the first fold and validated on the second. Then, for the next iteration, the first and second folds are now used for fitting the ML model, while a previously unused successive third fold is now used as a validation fold. This process of successively expanding the training set and moving the validation set forward is done a total of 7 times. This results in 7 evaluations of out-of-sample (on the validation folds) predictive performance for each of the ML models with a given combination of tuning parameters. These 7 evaluations are then averaged, the model with the best average predictive performance on the validation folds is selected. This procedure ensures that no future data is used to predict past data (for a graphical illustration of this procedure see Figure 3 in Cerqueira, Torgo, and Mozetič, 2020, p. 2000).

Appendix C: Comparative Performance of the best ML model relative to other ML models for FD and Debts

Table C1: Performance of Different ML Methods in Predicting Financial Distress (*FD*)

	Panel A: Brier Score (BS)		Panel B: Average Precision Score (APS)	
ML Method	**Val. Set**	**Test Set**	**Val. Set**	**Test Set**
GB	0.1240	0.1237	0.7768	0.7902
RLR	0.1256	-	0.7728	-
ET	0.1257	-	0.7756	-
RF	0.1266	-	0.7730	-
XGBoost	0.1303	-	0.7546	-
SVM	0.1346	-	0.7623	-
Bagging	0.1364	-	0.7595	-
DT	0.1638	-	0.6621	-

Table C2: Performance of Different ML Methods in Predicting Financial Distress (*Debts*)

	Panel A: Brier Score (BS)		Panel B: Average Precision Score (APS)	
ML Method	**Val. Set**	**Test Set**	**Val. Set**	**Test Set**
RF	0.0524	0.0528	0.4705	0.5334
ET	0.0527	-	0.4597	-
GB	0.0529	-	0.4604	-
RLR	0.0529	-	0.4649	-
XGBoost	0.0557	-	0.4107	-
Bagging	0.0573	-	0.4388	-
SVM	0.0595	-	0.4646	-
DT	0.0867	-	0.3362	-

Appendix D: RQ2-B2 budget-PPV curves for the different samples and their relative performance compared to RQ2-B1 and ML across both FD and Debts

Figure D1: Budget-PPV curves using the *FD* measure

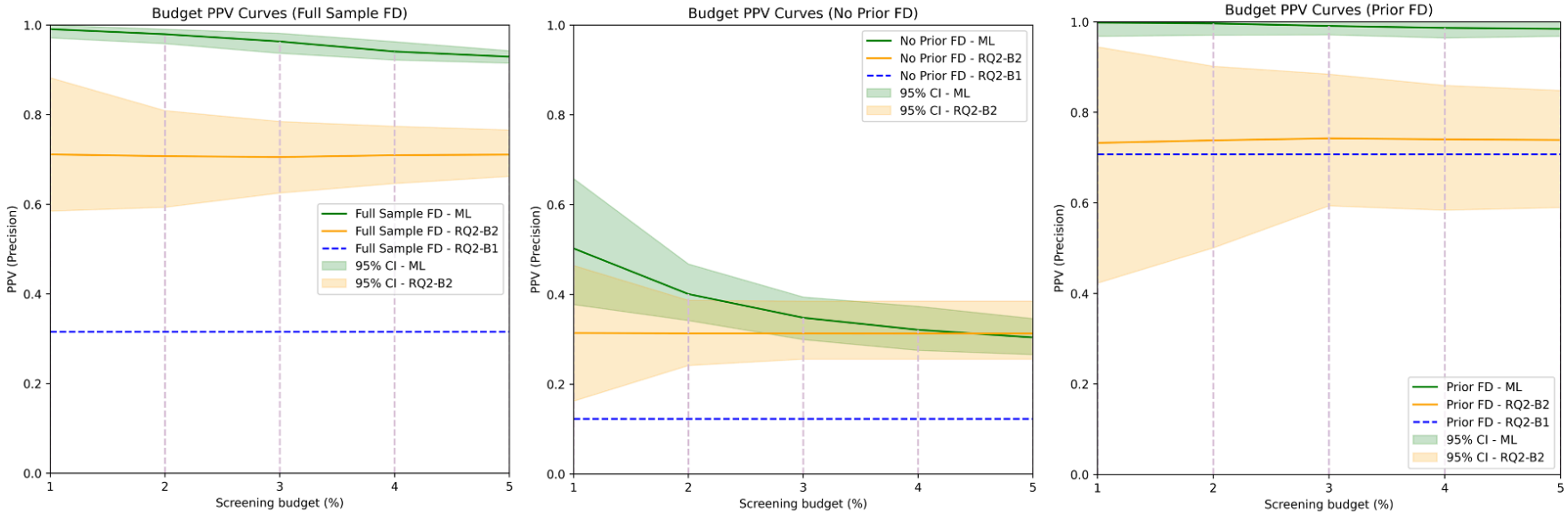
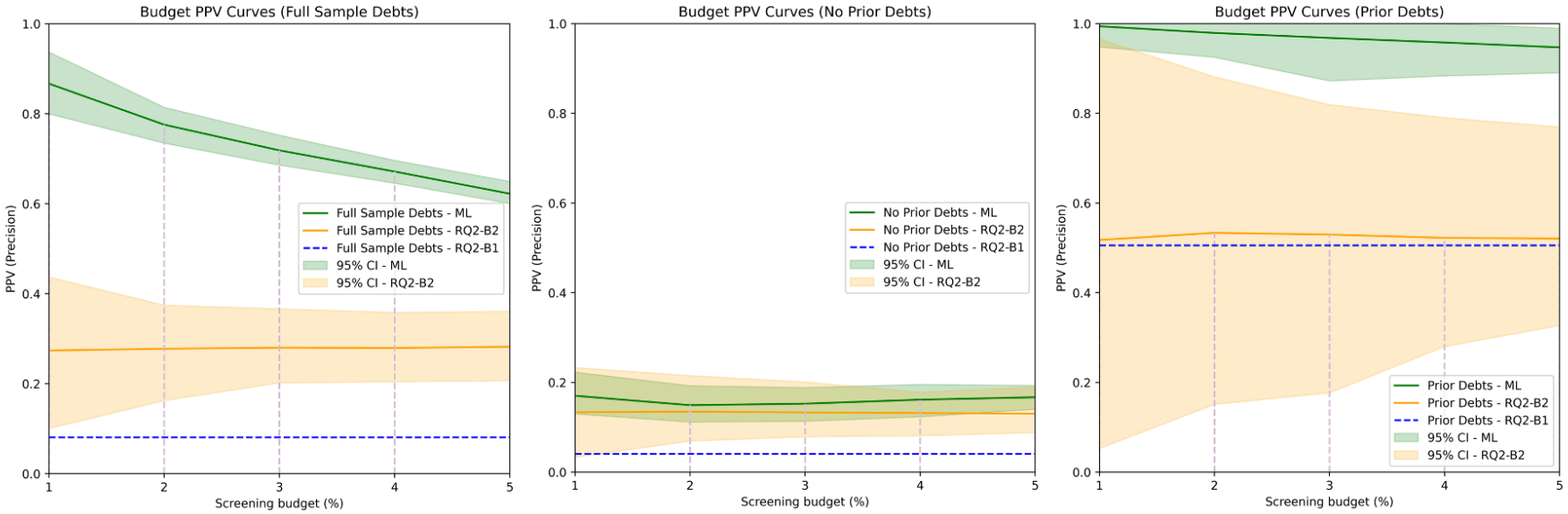


Figure D2: Budget-PPV curves using the *Debts* measure



REFERENCES FOR APPENDIX

Cerqueira, V., Torgo, L. and Mozetič, I. (2020) 'Evaluating time series forecasting models: An empirical study on performance estimation methods', *Machine Learning*, 109(11), pp. 1997–2028.