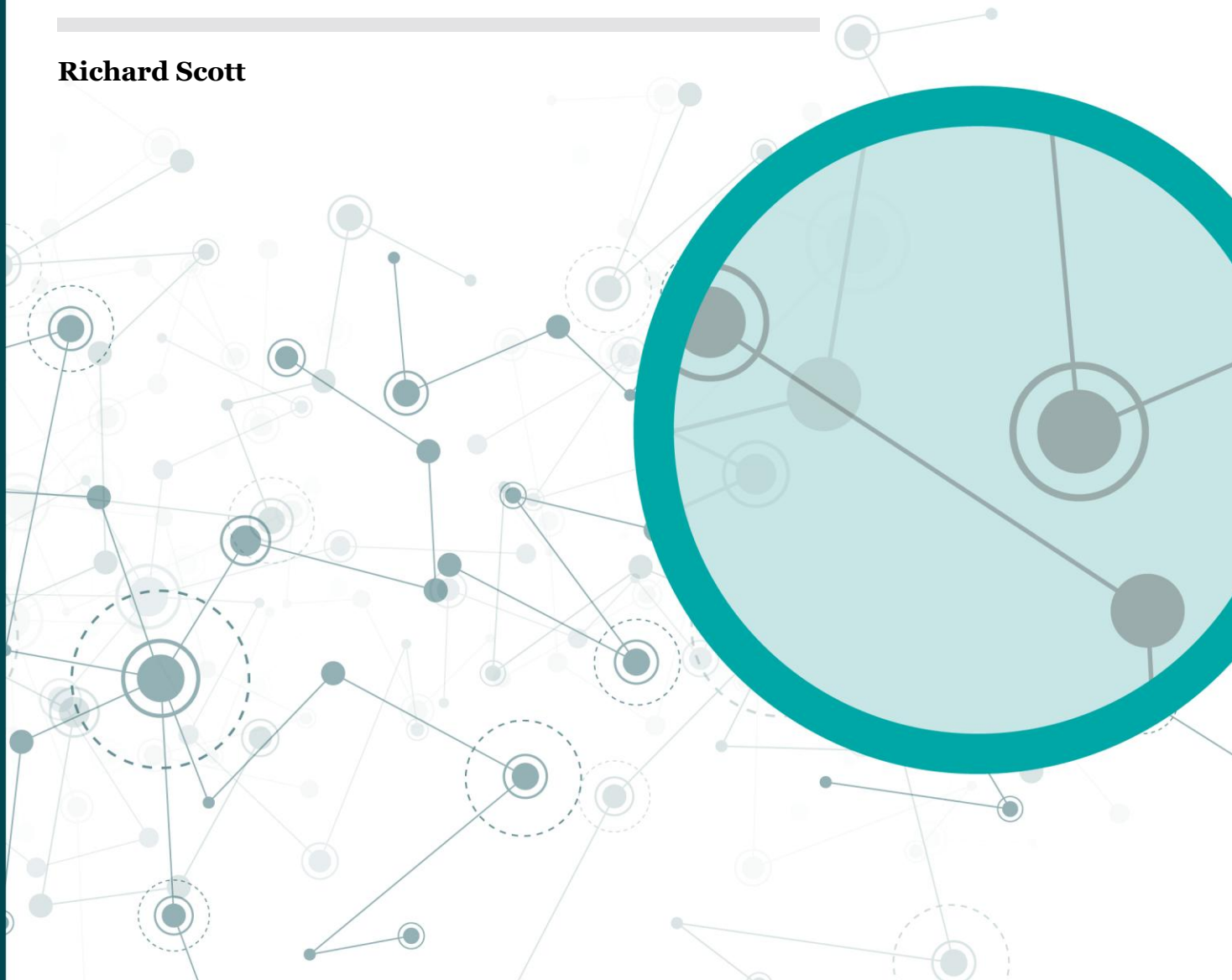
Financial Conduct Authority

Research Note

09/06/2026

Mortgage arrears risk among UK first-time buyers: A survival analysis approach

Richard Scott



FCA research notes in financial regulation

The FCA research notes

The FCA is committed to encouraging debate on all aspects of financial regulation and to creating rigorous evidence to support its decision-making. To facilitate this, we publish a series of Research Notes, extending across economics and other disciplines.

Disclaimer

Research notes contribute to the work of the FCA by providing rigorous research results and stimulating debate. While they may not necessarily represent the position of the FCA, they are one source of evidence that the FCA may use while discharging its functions and to inform its views. The FCA endeavours to ensure that research outputs are correct, through checks including independent referee reports, but the nature of such research and choice of research methods is a matter for the authors using their expert judgement. This note is provided for general information only. The FCA does not guarantee the accuracy, completeness, or reliability of this note. The FCA accepts no responsibility for any errors or omissions in this note, any loss or damage arising from reliance on this note, or for any action taken based on the information provided.

Authors

Richard Scott, Technical Specialist, FCA Economics Directorate.

Acknowledgements

We would like to thank Prof. John Gathergood at the University of Nottingham for academic review of the paper. Thank you to Guido Bodrato (former FCA), Daniel Wylie and Charlotte Woodacre (both FCA), for support on wider work linked to this paper, and to colleagues at the FCA for review comments.

All our publications are available to download from www.fca.org.uk. If you would like to receive this paper in an alternative format, please call 020 7066 9644 or email publications_graphics@fca.org.uk or write to Editorial and Digital Department, Financial Conduct Authority, 12 Endeavour Square, London E20 1JN.

Contents

Summary

This Research Note examines the drivers of mortgage arrears among UK first-time buyers using detailed loan-level Product Sales Data (PSD) between 2015 and 2025. It focuses on the probability (i.e. risk) that a borrower enters arrears for the first time – defined as a shortfall equivalent to two or more regular payments – and how this risk varies with borrower characteristics, loan features and economic conditions over the life of the mortgage observed within the dataset. The results will inform our work under the Mortgage Rule Review and beyond.

Arrears are an important early signal of financial strain in the mortgage market. Arrears do not, by themselves, demonstrate poor outcomes or inappropriate lending. For example, some arrears are unforeseeable, and some are resolved quickly. But arrears negatively affect household wellbeing and future credit access, and can lead to significant wider consequences, particularly for certain types of borrowers. Understanding the factors that influence the likelihood of falling into arrears helps us identify where repayment pressures are potentially most concentrated.

Our analysis uses a survival model which estimates how likely an event (in this case falling into arrears for first-time buyers) is to happen over time. Our model also allows us to exploit the semi-annual structure of PSD and to account for 'censoring', or the flow of people into and out of the mortgage market. Our data allow us to track borrowers across mortgage deals and through changing interest rate environments, from a prolonged period of low rates into the higher interest rate environment beginning in 2022. We complement our survival model with a machine-learning approach that informs our main model.

Findings

The results show that several factors are associated with lower arrears risk among first-time buyers. Higher income and, on average, incentivised interest rate offers are both strongly associated with lower risk of arrears. (Though incentivised rates appear to increase risk at later periods.) The findings also suggest the importance of consumer engagement in the mortgage market, as switching mortgage product may reduce arrears, risk even when controlling for gross interest rates and other factors. However, we note that switching could be correlated with unobserved consumer characteristics.

By contrast, higher first-time borrower leverage, particularly above 80% and 90% loan-to-value (LTV) rates (adjusted for local house price inflation), is one of the strongest risk indicators of entering arrears. This is consistent with economic intuition and the wider literature. More highly leveraged households have less capacity to absorb financial shocks. In addition, our focus on first-time buyers means our sample is a relatively young demographic—the median borrower age is 30—so borrowers may not have had time to build savings or equity buffers to protect them from the impact of adverse shocks.

A number of other mortgage features are also associated with higher risk of entering into arrears. Tracker mortgages, where the interest rate tracks Base Rates or another benchmark plus a markup, are associated higher arrears in our data. This may reflect the fact that our sample period includes a sharp increase in Base Rates in 2022 and 2023 especially which affected a large number of first-time buyers on tracker rates in our sample. Having previous credit impairment, being self-employed, buying a home under a government scheme like Help to Buy and having dependent children are also all linked with higher risk of arrears.

Arrears risk is also not concentrated at the time of origination. The probability of entering arrears rises over the first few years of the loan before flattening out, suggesting that borrowers' financial circumstances change over time, or that financial stress varies over time as borrowers face changing economic conditions. This highlights the importance of monitoring how risks evolve after origination, rather than focusing only on initial affordability assessments.

We also benchmark our survival model against a machine-learning based gradient-boosted model. The results imply the gradient-boosted model is better able to rank households by risk.

Limitations

Our data has some limitations. For example, we do not observe changes in first-time borrower circumstances after origination, such as job loss or ill-health, that are known from previous studies to cause arrears. In addition, while we observer borrowers over time, we cannot observe borrowers who sell or move home.

These data limitations mean that we cannot know how borrowers self-select into different mortgage products. Our focus on first-time buyers avoids some issues from selective survival of loans but also means that our findings our not necessarily applicable to understanding the arrears patterns of the wider stock of mortgages.

Our findings are therefore best read as indicating which characteristics are associated with higher first-arrears risk among first-time buyers in this period, and their relative importance, rather than as precise or causal estimates of any single factor's effect. Nonetheless, our results are consistent with prior peer-reviewed studies and are robust to a range of diagnostic checks.

Discussion

Taken together, the results provide a consistent picture of arrears risk among first-time buyers. The main drivers – leverage, income and household financial structure – are intuitive and estimates are stable across specifications, whilst being broadly aligned with prior research. This paper contributes to the evidence base by contributing up-to-date, UK-specific information on the relative risk factors facing first-time buyers. Our sample period covers a period of rapidly increasing interest rates and cost of living pressures. Moreover, our findings allow us to understand the contribution of less common mortgage characteristics such as employment status or incentivised rates.

Our models also provide a practical framework for assessing how changes in lending conditions could affect arrears risk. While extrapolating from historical data involves uncertainty and strong assumptions, our gradient-boosted model in particular could be

used internally to illustrate how different compositions of new lending could translate into different aggregate arrears outcomes, and be used to stress-test future policy options.

As such, the findings support the FCA's ongoing work under the Mortgage Rule Review and the wider theme of risk rebalancing. Expanding access to mortgage credit may increase exposure to arrears risk for some borrower groups, whereas other product features may ease short-term repayment pressures. While quantitative assessment of risk rebalancing policy changes is very difficult, the evidence in this paper helps us assess the potential where these risks are likely to sit and their potential magnitude.

1 Introduction

Policy context and motivation

Understanding which borrowers are at greatest risk of falling into arrears is an important question in mortgage policy. Falling behind on a mortgage can have large negative implications for household wellbeing. For lenders and regulators, arrears can signal underlying risks that require policy attention. While entering into arrears does not in itself imply poor consumer outcomes or inappropriate lending, it does provide a measurable indicator of repayment stress and consumer resilience over time.

This paper informs future work under the FCA Mortgage Rule Review (MRR). As originally set out in Discussion Paper DP25/2, the MRR aims to simplify and modernise aspects of our mortgage rules to support wider appropriate access to mortgage products. Feedback Statement FS25/6 confirmed broad support for this strategic direction and set out a roadmap of targeted reforms to expand access for creditworthy borrowers while maintaining core protections. We are publishing this Research Note alongside CP26/18, which sets out proposals including changes to our rules around affordability requirements, including for interest-only and part interest-only mortgages, and products tailored to borrowers with variable or non-standard income profiles.

This Research Note improves our understanding of risk rebalancing for first-time buyers in the mortgage market. Under the Mortgage Rule Review, this means balancing the prevention of poor consumer outcomes with maintaining access to credit, particularly for people currently outside of the property market and current underserved groups. We therefore assess how regulatory changes and lending conditions may affect arrears risk as part of testing 'tolerable harm'. The work also contributes to the FCA Strategy priority of helping consumers navigate their financial lives.

Research objective and contribution

This Research Note analyses how borrower characteristics, loan features, and economic conditions relate to the risk of entering mortgage arrears. We focus on first-time buyers and examine how factors observed at origination and over the life of the loan are associated with the likelihood of entering arrears for the first time. Our objective is to provide policy-relevant evidence on which factors are most strongly linked to arrears risk.

The analysis contributes to the existing evidence base in two ways. First, it provides up-to-date evidence – our January 2015 to June 2025 sample range covers both a prolonged period of low interest rates and the subsequent tightening in UK monetary conditions starting in 2022. Second, it allows us to examine the role of specific mortgage features that are not well captured in much of the existing literature.

2 Research design and data

This section sets out a high-level overview of the empirical approach used to estimate the risk of entering mortgage arrears. We also summarise prior literature on mortgage arrears. We provide further technical detail on the modelling approach, including robustness checks, in the annexes.

Research design

We model the determinants of mortgage arrears using a survival model. Survival models are a statistical form of regression analysis where the outcome is the time until an event, and where many observations never experience the event at all during the sample period. The model addresses the question: 'given that a fist-time mortgage has reached a particular point in its life without entering arrears, how does the probability of it entering arrears in the next period vary with borrower characteristics, current loan features and wider economic conditions?'

We estimate a discrete-time proportional hazard model, a survival model that is specifically suited to the repeated snapshot structure of our data. We model the probability of a mortgage entering arrears in each six-month period, conditional on having survived to the start of the period. Our approach tracks the evolving financial conditions of loans by following borrowers across successive mortgage deals, conditional on remaining in the same home.

This approach is well suited to the mortgage arrears setting for several reasons. Firstly, arrears are rare – most households never experience arrears in our sample (see Section 3 for descriptive statistics). A more conventional statistical approach that focused on the purely binary question of whether households experienced arrears would not account for how long each loan was observed, reducing the information content of the model. The survival framework uses the information from the time each loan spends at risk, even if a loan does not fall into arrears. Second, loans enter and leave the dataset at different points, and many are closed for reasons unrelated to arrears such as redemption, switching lender or the borrower moving house. This entry and exit is known as censoring, and survival models explicitly adjust for it.

The outcome of interest is first-time entry into mortgage arrears for first-time buyers, defined as the first shortfall equivalent to two or more regular payments. Once a household enters arrears, we remove all subsequent observations so that the estimated hazard captures the transition into arrears rather than repeat or prolonged default. This design ensures that the estimated relationships reflect conditions at the point repayment difficulties first arise. While repeat arrears dynamics are important, modelling time to first arrears isolates the underlying driver of distress without conflating it with the path-dependent consequences of having already defaulted.

Machine-learning model

Alongside the discrete-time hazard model, we estimate a machine-learning model of mortgage arrears. We use a gradient-boosted decision tree model using the LightGBM algorithm (Ke et al., 2017). This is a modern machine learning method widely used in industry, including for credit modelling and stress testing. Unlike the hazard model, which estimates the effect of each variable through a specified functional form, the gradient boosted model learns flexible patterns directly from the data, automatically detecting non-linearities and interactions between borrower characteristics. In exchange for this flexibility, gradient-boosted models are less interpretable — it does not produce coefficients like our main model — and can be prone to overfitting, meaning they can learn patterns that are specific to the data used to train them rather than patterns that hold more generally.

We use the gradient-boosted model for two purposes. First, in preliminary work to refine variable selection in the discrete time hazard model. (The modifications were marginal, so do not present risk that we are undertaking circular benchmarking.) Second, to benchmark and compare the predictive performance of the survival model, reflecting that our hazard model is better suited as an explanatory ex-post model but a machine-learning approach is likely to be stronger at prediction. The hazard model remains our primary tool because its interpretable coefficients provide clearer interpretation.

Data

The analysis uses UK Mortgage Product Sales Data (PSD) for all first-time buyer mortgages sold between January 2015 and December 2024, and subsequent 6-monthly updates on their status up to June 2025. We use an internal matched sample of PSD001 (sales data) and PSD007 (6-monthly performance data). Approximately 90% of mortgage sales in PSD001 appear in the matched sample.

PSD contain detailed information in two broad categories:

- Loan characteristics: E.g. loan-to-value ratio (LTV), loan-to-income ratio (LTI), payment-to-income measures, loan term, repayment type, and fixed versus variable rates at origination. PSD records interest rates, terms, payment types and arrears and possessions information on an ongoing basis in the 6-monthly snapshots.
- Borrower characteristics: E.g. borrower age, employment status, number of borrowers.

In addition, following prior literature we merge PSD observations with local output area-level unemployment rates from ONS, and an index of local house prices matched on both local area and dwelling type. The former allows us to proxy local economic conditions facing borrowers, while matching with house prices allows us to estimate the current LTV based on the current loan balance and uprated property value estimate.

Restricting attention to first-time buyers avoids statistical issues from the fact that the stock of mortgages any point is a selected subset that has already avoided arrears or default (known as left-truncation).

We define the start of the observation window after the introduction of the Mortgage Market Review (MMR), ensuring that the sample reflects a consistent regulatory regime. We note that our coronavirus guidance for lenders in 2020 that introduced repayment deferrals and, to a lesser extent, the Mortgage Charter from June 2023, may have

altered the path of some distressed borrowers into arrears. The definition of arrears itself has remained constant over our sample period.

A key strength of the data is the ability, subject to conditions, to follow households over time and model the probability of arrears at each time point. Each mortgage contributes observations at origination and at six-monthly intervals until it redeems, refinances, or enters arrears. The unit of observation is the household holding the mortgage—proxied by the combination of property postcode and borrower date of birth—rather than the individual mortgage. This structure allows us to follow borrowers across product switches and remortgages, though not across property moves, as a change in postcode breaks the data matching.

Data limitations

We note a number of limitations with our data. We note how these affect limitations of our analysis in Section 4:

- We observe households' circumstances at mortgage origination but not subsequently. We do not observe household-specific changes in income, employment status, family structure or health, which are known to be important factors in arrears.
- We do not observe borrowers who sell their property or move home – households remain in the data over time conditional on remaining in the same property.
- Third, our observation window can be short relative to the length of mortgages. The maximum period during which we observe a household is fewer than ten years.

Prior literature

We reviewed the relevant academic and regulatory literature on arrears in our annex to DP25/2, but provide a short synthesis here.

The literature in economics and finance on mortgage arrears and default has evolved from viewing mortgage default by borrowers as primarily arising a rational response to negative equity to a greater focus on involuntary repayment difficulties driven by adverse borrower shocks. In the US, recent evidence suggests purely strategic defaults – where mortgage default is a strategic response to the loan exceeding the value of the property – represent a small minority of the total volume of mortgage loans and far below earlier estimates (Ganong and Noel, 2023). Previous literature links mortgage default to a range of adverse shocks including unemployment (Gerardi et al., 2018), interest rate shocks (Campbell and Cocco, 2015), and other cash-flow pressures (Elul et al., 2010).

Previous studies also link mortgage arrears to borrower and loan characteristics that capture household financial strain. Various findings suggest that borrower leverage, e.g. loan-to-value levels (Stanga et al., 2020) or gearing, e.g. debt-service-to-income ratios (Kelly and McCann, 2015), the number of borrowers on the mortgage (Bergmann, 2020), employment status of borrower (Reserve Bank of Australia, 2019) and borrower age (Aarland & Santiago, 2023) are all associated with arrears risk. In addition, some cross-country studies have shown that institutional factors, including underwriting and recourse law, can explain substantial variation in how income and equity shocks translate into defaults (Linn & Lyons, 2020).

A handful of studies have estimated mortgage arrears using survival models or, increasingly, machine learning models. Slaymaker et al. (2019) use a discrete time survival model to estimate the association of arrears to a measure of current debt service ratio in Ireland, finding that a 100 basis-point increase in central bank interest rates would lead to a 0.5 percentage point increase in new defaults. Bolliger et al. (2024) estimate another common survival model (a Cox proportional hazards model) on Irish data, but conclude that a machine-learning random forest model outperforms the survival framework and is better able to handle complex features of the dataset. Azimi & Khaledian (2025) use data on US mortgage arrears to demonstrate the superiority of machine-learning approaches, including LightGBM – the algorithm we use in this paper, over traditional logit regression for prediction purposes, while Barbaglia et al. (2023) come to a similar conclusion, finding interest rates and local economic characteristics best explain loan default in their European data. One downside of machine learning models noted in these papers, however, is the results tend to be less interpretable for policy purposes.

3 Descriptive statistics

Descriptive analysis of our sample

We run our final model on a stratified sample of the data where we take all households that experienced arrears at any point, and 20% random sample of all other households that appear at any point in the full sample. This is common practice in rare-event analysis and does not bias our estimates (though we need to correct the estimated intercept before any use of the model's predicted values).

Table 1 summarises the estimation sample. It shows how many households we observe, how long we follow them on average, and how often first arrears occur over that window.

Table 1: Sample overview

Metric	Value (full sample)	Value (stratified sub-sample)
Unique households	3,109,238	682,064
Total household-period observations	27,308,949	5,896,583
First-arrears events	75,271	75,271
Event rate (%)	2.42%	11.04%
Number of periods observed - median	8	8
Number of periods observed - mean	8.8	8.6

Source: FCA analysis of PSD

Note: Periods are 6-months, meaning on average we follow households for 4-5 years

Table 2 summarises selected numerical variables at origination, the point at which we first observe first-time buyers in our stratified sample. Figures are nominal, not corrected for inflation.

Our focus on first-time buyers shapes the demographic profile of the sample. Borrowers are relatively young, with a median age of 30, and the median number of dependent children is zero. Reflecting this earlier lifecycle stage, leverage is relatively high: the median loan-to-value ratio is 85% and the median loan-to-income ratio is 3.6. The analysis on first-time buyers should therefore be interpreted as evidence on arrears risk

among a younger and more highly leveraged segment of mortgage borrowers, rather than as representative of the whole mortgage market.

Table 2: Summary of numeric variables at origination in analysis sample

Variable	Mean	Median	P25	P75
Age of first borrower	32	30	26	36
Initial interest rate (%)	3.05	2.69	2.08	3.89
Total household gross income (£)	53,664	44,960	32,193	62,039
Loan value (£)	180,501	152,000	106,000	224,000
Stress test differential (stress test rate minus initial interest rate)	4.1	4.3	3.5	4.9
Number of dependent children	0.4	0.0	0.0	0.0
Number of dependent adults	0.1	0.0	0.0	0.0
Loan-to-value ratio	77.4	84.8	72.8	90.0
Loan-to-income ratio	3.5	3.6	2.9	4.2
Expected monthly payment at origination (£)	789	653	461	964
Mortgage term in months	357	360	300	420
Previous credit impairment	0.01	0.00	0.00	0.00
Advised sale	0.98	1.00	1.00	1.00
Mortgage advanced under a gov. supported initiative	0.13	0.00	0.00	0.00

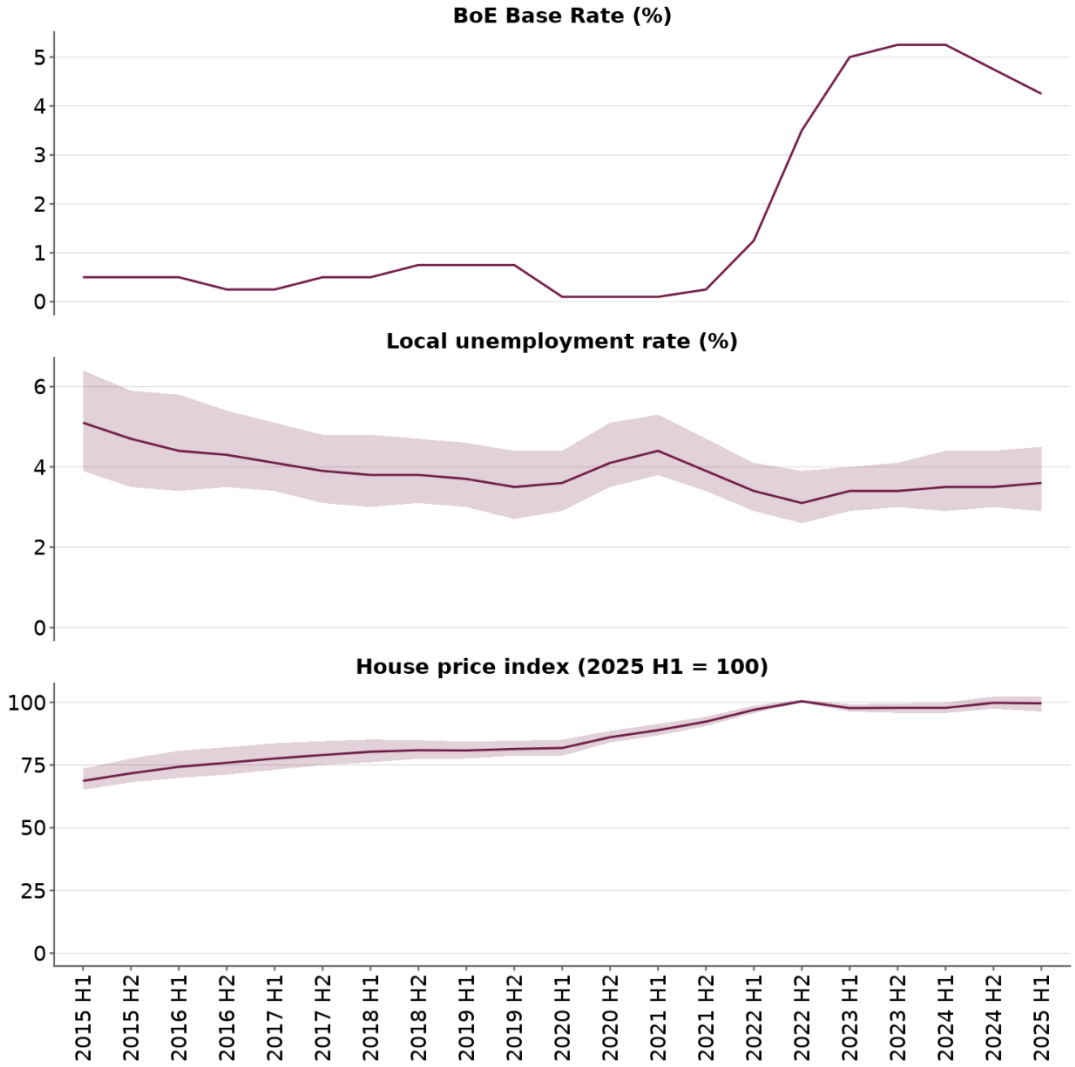
Source: FCA analysis of PSD001 and PSD007

Note: N = 682,075. Values for households at first period of loan only. Previous credit impairment (whether the borrower had previous arrears, bankruptcy, IVA or CCJ), Advised sale and government supported initiative (whether the loan was advanced under government schemes like Help to Buy) are dummy variables taking the value 1 for yes and 0 for no.

Macroeconomic and local economic conditions

Error! Reference source not found. shows how time trend for the macroeconomic variables in our analysis: Base Rates, local unemployment rates, and local house price index. For the latter two series, we plot the time trend of the median, 25th and 75th percentile local authorities in our data.

Figure 1: Trend in macroeconomic variables



Source: FCA analysis of Bank of England, Nomis and UK House Price Index data
Note: For the two lower plots, the line denotes the median region and the shaded zone denotes regions at the 25th and 75th percentiles.

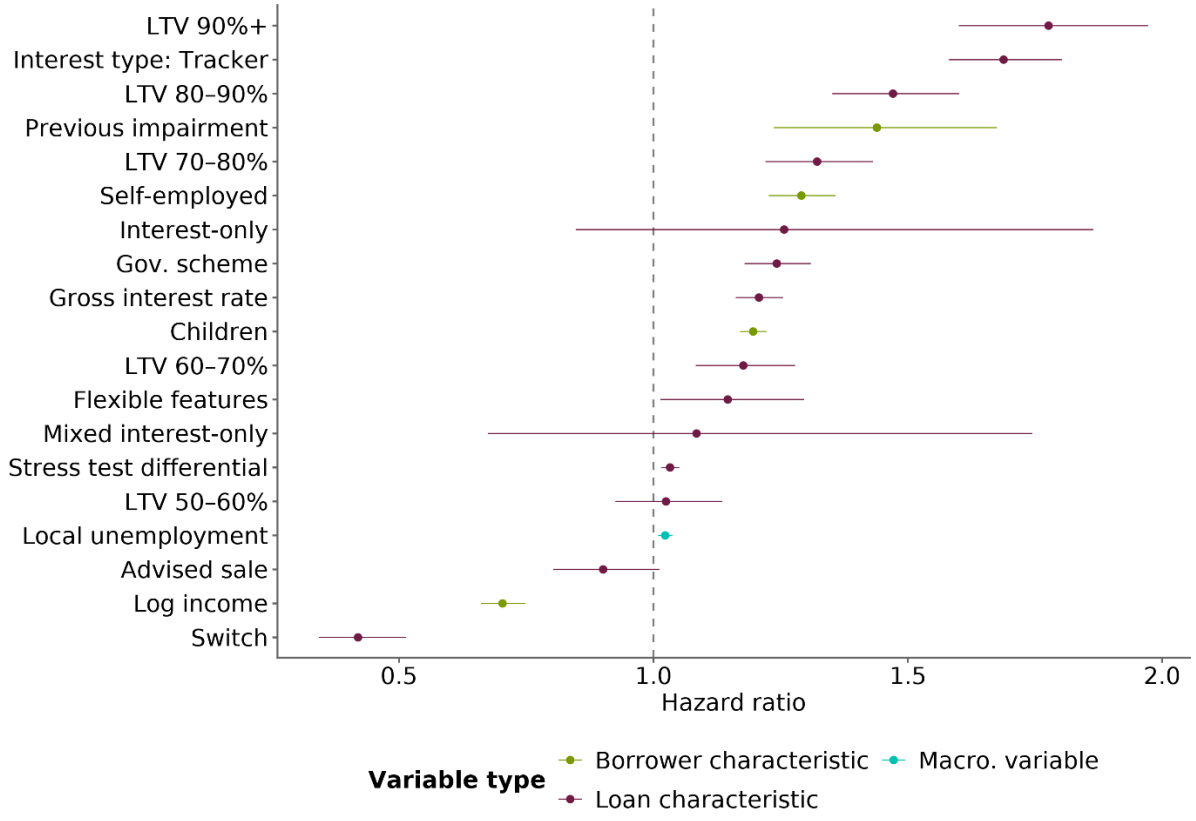
4 Survival model results

This section summarises the main results from the survival analysis, focusing on the borrower and loan characteristics most strongly associated with the risk of entering mortgage arrears.

Coefficient estimates represent the effect of each variable on the hazard of entering arrears, holding other factors constant. We report the results as hazard ratios — for example, a hazard ratio of 1.2 for a characteristic means that borrowers with that characteristic face roughly a 20% higher arrears risk in any given interval than otherwise similar borrowers, holding other factors constant.

Figure 2 sets out headline results, with the full results provided in the Annex. The lines represent 99% confidence intervals for the arrears hazard, holding all other factors constant. The results are independent rather than cumulative, i.e. they represent independent hazard ratios holding all other explanatory variables constant.

Figure 2: Estimated hazards of mortgage arrears, selected coefficients and 99% confidence intervals



Source: FCA analysis of PSD001 and PSD007

Notes: Estimates are hazard ratios (exponent of coefficients) expressed relative to baseline and, for categorical variables, relative to the excluded category. Point estimate and 95% confidence intervals. Time interactions not shown (see Annexes).

Borrower characteristics

The results show that several borrower characteristics are associated with arrears risk.

Higher income is associated with lower arrears hazard, even after controlling for loan characteristics. Because income enters the model in logs, the estimate implies that a 1% increase in income is associated with around a 0.3% reduction in arrears hazard.

Household composition is associated with arrears risk, but the pattern is mixed. Mortgages with dependent children have higher estimated arrears hazard, consistent with additional household expenditure pressures. The estimate for dependent adults is positive but less precise. The point estimate for two-borrower mortgages is slightly lower than for single-borrower mortgages, suggesting having multiple incomes may protect the household from adverse shocks. However, mortgages with three or more assessed borrowers—potentially reflecting more complex or stretched affordability assessments—have higher estimated hazard. This suggests that larger borrower groups may capture more complex household or affordability circumstances rather than simply additional income insurance.

Employment status also matters. Relative to employed borrowers, self-employed borrowers have an estimated 24% higher arrears hazard, while borrowers with unknown employment status have an estimated 25% higher hazard. This may reflect greater income volatility, unobserved borrower characteristics, or informational frictions at underwriting.

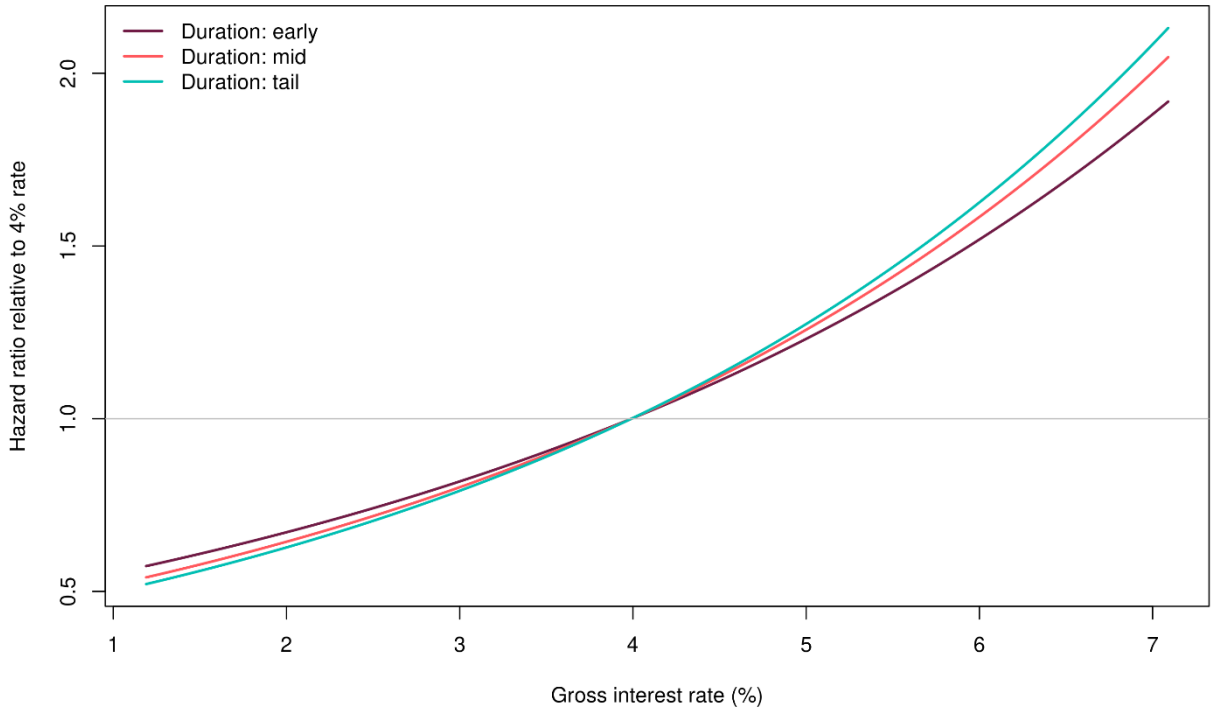
Borrowers with a recent credit impairment at origination are more likely to experience mortgage arrears, though this group is rare among first-time buyers. The impairment indicator is associated with around 32% higher arrears hazard than no recorded impairment. Preliminary testing implied that arrears on previous secured and unsecured loans drives this result rather than IVA, bankruptcy or debt relief orders, which are rarer in our data. The definition of impairment in PSD excludes arrears on revolving credit such as overdrafts or credit cards. Importantly, however, less than 1% of our sample of first-time borrowers have previous credit impairment and our results do not imply that all impaired borrowers are riskier.

Loan characteristics

Interest rates matter for arrears risk, but the model suggests the strength of that relationship changes only modestly over a loan's life. Figure 3 shows the estimated relative effect of interest rates on arrears hazard, separately by loan age. We group observations into duration bands (0–3, 4–7, and 8+ years since origination, though the household could have switched mortgage in that time). Each curve is normalised to 1 at a 4% interest rate within its own duration band. The chart therefore shows how arrears risk changes as rates move away from 4%, rather than comparing the overall level of risk across borrower types or duration bands. We hold other borrower and loan characteristics fixed, so the figure isolates the modelled interest-rate relationship.

Higher rates are associated with higher arrears hazard in all duration bands. The relationship is slightly steeper for households with older loans aged 8+ years and slightly flatter for younger loans, though the differences are modest and we place more weight on the overall positive rate relationship than on the precise shape of the curve.

Figure 3: Estimated effect of interest rate on arrears hazard, relative to a 4% rate



Source: FCA analysis of PSD

Consistent with prior literature, borrower leverage is among the strongest factors associated with of arrears, even after conditioning on borrower income and other loan features. To avoid multicollinearity (high correlation between explanatory variables), we include only one measure of leverage, current estimated loan-to-value ratios (LTV), though including loan-to-income or debt service ratio in preliminary work appeared to make relatively little difference to our conclusions. (We are able to include all leverage measures in our gradient-boosted model.) Higher current LTV bands are associated with progressively higher arrears risk, with materially larger hazards for loans above 80% and particularly above 90% LTV. Relative to loans below 50% LTV, loans above 90% LTV have an estimated 74% higher arrears hazard.

Over the period we study, repayment type has a less clear relationship with arrears risk in this specification. Interest-only mortgages from origination have a positive but imprecise estimate relative to standard repayment mortgages, while mixed repayment mortgages have a smaller positive but imprecise estimate. On one hand reduced monthly repayments may reduce the risks of non-payment compared with standard repayment mortgages in the early life of a loan, and interest-only mortgages also tend to be used by borrowers able to raise substantial deposits. On the other hand, the choice of interest-only repayment type at origination could be correlated with other unobserved characteristics of borrowers that lead to a higher arrears risk association. As noted above, we cannot comment on the longer-term risk of interest-only or mixed mortgages as they approach term, when arrears or non-repayment risk would theoretically increase.

Loans associated with government-supported initiatives such as Help to Buy and those secured on new-build properties exhibit higher arrears risk, even after conditioning on leverage and borrower characteristics. Longer mortgage terms are associated with slightly lower arrears risk at a point in time, reflecting lower required monthly payments, though the time period of the analysis does not capture risks that may emerge later in the mortgage term.

Some estimates require caution. The mortgage switch variable is associated with substantially lower arrears hazard, but switching behaviour is likely to reflect borrower selection and may be related to unobserved financial resilience.

Overall, the results are consistent with expectations and prior work that leverage, household structure and income stability as the most important correlates of arrears risk, with product features playing a secondary but still meaningful role.

Limitations of our analysis

While our results provide new insights into the drivers of mortgage arrears, we note some limitations. We explore these issues through robustness checks in Annex 3.

The model identifies associations rather than causal effects, although our results appear robust to persistent unobserved account-level factors. Many mortgage characteristics—such as loan-to-value ratios, product types and term lengths—are jointly determined by borrowers and lenders and may depend partly on unobserved factors, including risk preferences, expectations of income stability and lenders' private information. If these factors underpinning mortgage choice also affect arrears risk (e.g. borrowers may select mortgage products that leave them more exposed to repayment difficulties or lenders may tolerate differential levels of expected arrears across product types), unobserved heterogeneity could affect the baseline survival model. However, Annex 3 presents 'frailty' model results, which suggest that the main findings are not highly sensitive to persistent borrower-level unobserved factors.

Second, households may exit the sample in a non-random way, although our competing-risks checks suggest this is unlikely to drive the main results. Because we do not observe borrowers who sell their property or move home, the model may miss cases where households respond to early repayment strain by adjusting their housing situation before entering arrears. This could bias estimated associations if households that leave the sample differ systematically in unobserved arrears risk from those that remain. To assess this, we estimate a competing-risks specification that models arrears and sample exit as alternative outcomes. We do not find a persistent relationship between the coefficient estimates in the arrears and exit models, suggesting that the main results are not materially affected by selective exit from the sample.

Third, we only partially capture longer-term dynamics in our estimates. Because our sample is skewed towards the relatively early periods in the life of a mortgage, our findings predominantly apply to shorter-term arrears risk. Some mortgage features—most notably interest-only repayment—may in theory be associated with higher arrears risk later in the mortgage term, as the risk of being unable to repay the loan balance increases over time. However, the fact that we have matched borrowers where possible over a 10-year period means our data are longer-term in nature than most studies in prior literature.

While we control for local economic conditions and other observable factors, local economic indicators are imperfect proxies. Their omission could lead to two concerns: attenuation (downward) bias in the estimated impact of local unemployment on arrears (Gyourko & Tracy, 2014), and potential omitted variable bias that can affect other estimates (depending on if and how they correlate with these unobserved shocks). Our use of region and half-year fixed effects mitigates some of these concerns, as does that fact that most of our variables set at origination are unlikely to be closely related to future risk of adverse shocks. The coefficient on local unemployment, in particular, should be interpreted as the combined impact of local unemployment and the individual job losses that tend to come with it, bundled together.

Overall, we consider the results useful for informing policy development, but they require careful interpretation. While our estimates deliver precise parameter values, these may be specific to the period of analysis and have limited generalisability to other periods. A conservative approach to interpreting the analysis is to focus on the direction and relative magnitude of the estimates, rather than on the precise size of any single coefficient.

5 Gradient-boosted model results

Comparison

Our gradient-boosted model outperforms our survival model on discrimination (ranking households by risk), though the two are roughly equally accurate on average. Using a training period of observations prior to 2022 and test data 2022 and after, the two models achieve similar Brier scores (Table 3), which measure measures how close each model's predicted arrears probabilities are to what actually happened. However, the gradient-boosted model attains a materially higher AUC (0.751 vs 0.669), which measures how well a model separates first-time buyer households that go on to enter arrears from those that do not. Overall, both models are equally accurate on average, but the gradient-boosted model is better at telling higher-risk households apart from lower-risk ones.

The comparison points to possible non-linearities or interactions not captured by the simpler survival model. The results also suggest the gradient-boosted model would predict arrears risk more effectively than our survival model where the goal is ranking households by risk, though the two perform similarly at estimating the level of risk, and this advantage may not carry over to out-of-sample projection.

Table 3: Comparison of gradient-boosted and survival model

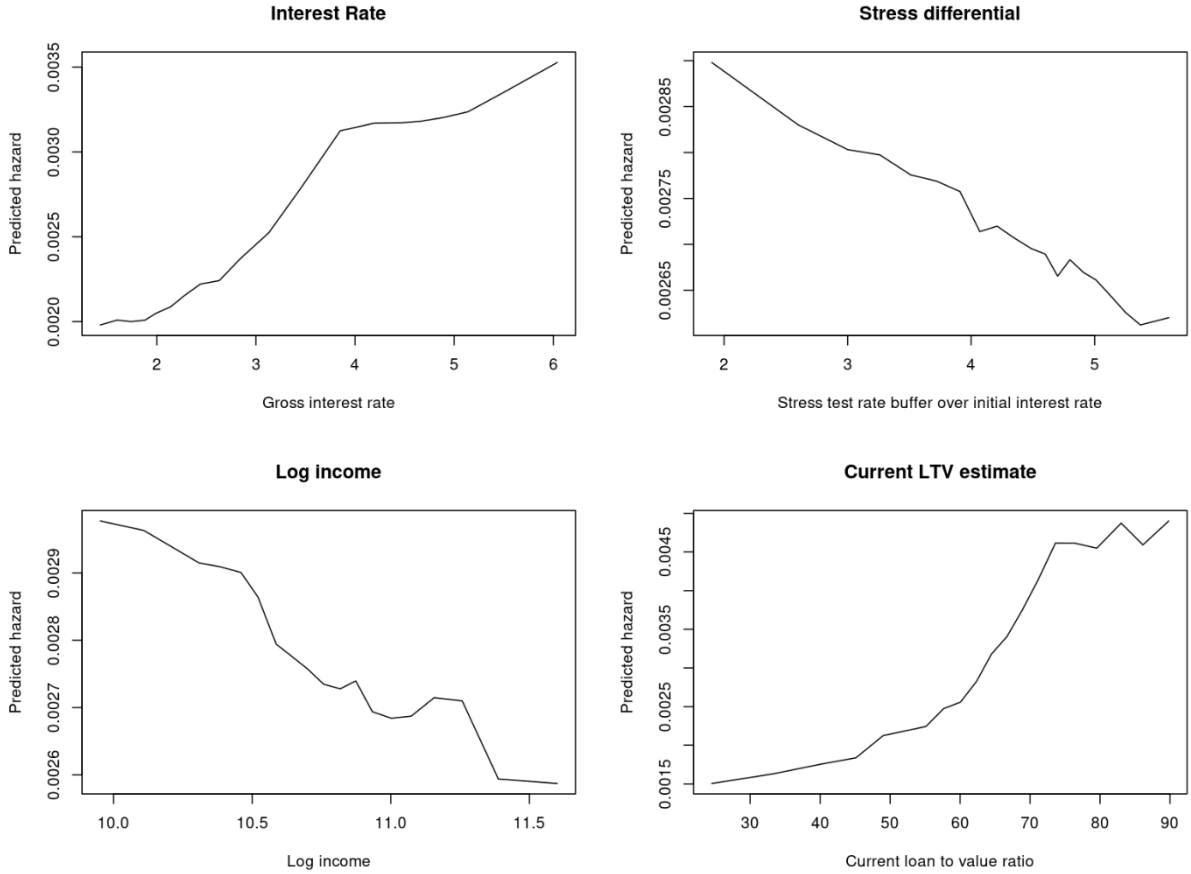
Model	Brier score	AUC score
Gradient-boosted (machine learning) model	0.0137	0.7497
Discrete time hazard (survival) model	0.0143	0.6750

Source: FCA analysis of PSD
Note: Results reflect out-of-sample comparison on the 2022- test set, with both models scored on identical rows. Scores are on the stratified sample, given a relative comparison rather than population-calibrated values. Brier scores are mean-squared squared prediction errors (lower is better); AUC is area under the ROC curve (higher is better, 0.5 = random)

Non-linearities

Figure 4 shows a selection of partial dependence plots, which show the shape of a variable’s association with arrears risk in the gradient-boosted model. The curves suggest the model is able to detect some non-linearities, though the curves are best interpreted qualitatively, rather than as an effect size in interpretable units.

Figure 4: Selected partial dependence plots based on our GBM model



Source: FCA analysis of PSD

Note: Plots show the model-implied average probability of entering arrears across the range of one variable, with other covariates held at their observed value.

6 Conclusion

This empirical paper examines the correlates of first-time buyer mortgage arrears using administrative data and a multivariate survival modelling framework. The results are consistent with the established literature, showing evidence of the role of borrower characteristics, mortgage product features and economic conditions in determining modelled arrears risk. The findings provide reassurance that the modelling approach captures well-understood patterns present in the mortgage market.

The paper provides new evidence on some specific loan features, such as flexible features, impairment and borrowers aided by government schemes. And we provide up-to-date evidence on the determinants of arrears, with mortgage originations up to 2024 and performance data up to June 2025.

Our results could be used in the future to understand consumer resilience in the mortgage market, for instance in response to macroeconomic shocks or future policy proposals. But there are many difficulties and assumptions involved in extrapolating beyond the estimation sample, for example to quantitively assess risk rebalancing, when firm and consumer behaviour reactions to future policies are highly uncertain.

One way we could extend the analysis in this paper would be further matching of PSD either with sources that contain non-mortgage arrears or household circumstances. Introducing variables for defaults or arrears on unsecured debt, for example, may improve short-term prediction of mortgage arrears and act as a type of leading indicator. Linking PSD to sources of household adverse events, particularly job loss, would deepen our short-run predictive power further, though real-time sources for this are less obvious.

Annex 1: Data preparation

Product Sales Data (PSD)

Our analysis uses mortgages product sales data – PSD001 captures all regulated mortgage sales and PSD007 records information on the performance of the stock of active mortgages (ongoing loan conditions, and arrears and possession information). Regulated mortgage lenders complete both returns every 6 months.

Borrower identification and matching

Given there is no common identifier across PSD001 and PSD007, we use an in-house sample that matches mortgages in these two data returns using a combination of first borrower's date of birth and postcode. We remove duplicate observations prior to analysis – these are either associated with multiple mortgages per household erroneous matches

We use the same date-of-birth–postcode combination for longitudinal matching. We track borrowers over time conditional on the postcode and the main borrower remaining unchanged. We stop observing borrowers following a property sale, change of address, or change in the main borrower on the mortgage.

Merge

Longitudinal tracking of mortgages requires several data adjustments. In some cases, an existing mortgage overlaps with a new mortgage within PSD, as both may be active within the same six-monthly reporting period. Where this occurs, we drop the older mortgage observation in the overlapping time period to maintain a unique borrower–time-period structure.

Where a first-time borrower drops out of our merged data, we cannot observe whether it was because of a property sale, house move or other reason. We infer non-arrears mortgage exits where the final observed outstanding balance or remaining term is recorded as zero. In the handful of instances where we record arrears and exit in the same reporting period, we treat the observation as an arrears rather than an exit.

Sampling

We use a stratified sub-sampling strategy to improve computational efficiency. We retain all IDs (borrower date of birth-postcode combinations) that experience arrears at any point, and a random 20% of all other ID. There are no time constraints in choosing the random 20% – a household that exists in every period of the sample is equally likely to be selected as a household appearing in one period. This preserves in the sample the full trajectory of each loan. There around 3.1 million unique households in the full data, and around 680,000 in the selected sample.

As arrears events are rare and the sampling is independent of the covariates, the sampling approach does not bias our estimates of slope coefficients or hazard ratios (King & Zeng, 2001). To correct the intercept term for projections, we add a constant to

the linear intercept that brings the mean predicted hazard into line with the actual arrears rate observed in the full data. We solve for this constant numerically, with the correction being very similar in magnitude to the rule of thumb correction of -log(0.2) (Scott & Wild, 2001).

Sample exclusions

We exclude mortgage products that differ from standard residential owner-occupier mortgages, where we would expect the determinants of arrears risk to be structurally different. The excluded categories are business lending, bridging loans, second-charge mortgages, lifetime mortgages, retirement interest-only mortgages, buy-to-let mortgages, high-net-worth lending, self-build mortgages, and secured overdraft products.

Sample restrictions and variable exclusions

We exclude some variables that may influence arrears due to limited variation over the estimation period. Reported income verification is almost always 'yes' for first-time buyers, with most mortgages showing evidenced income at origination.

PSD also includes an indicator for low-start mortgages. Despite their potential relevance to the Mortgage Rule Review proposals, these loans represent a negligible share of originations, with around 3,600 observations recorded in PSD001 since 2005. This sample size does not support reliable estimation, so we exclude the variable from the model.

Mortgages with loan-to-value ratios above 100% are similarly rare. Since 2016, around 6,900 mortgages in PSD001 have reported LTVs above 100%. We include these observations within the 90%+ category rather than modelling them separately.

Data cleaning

We remove a handful of observations that contained implausible values. The criteria for removing observations are:

- Stress test rate > 30%
- Number of dependents > 25
- Loan value < 0
- Loan-to-income ratio > 10
- Loan-to-value ratio > 125%
- Interest rate > 30%
- Term > 600 months

Variable construction and coding

We apply some variable coding decisions designed to preserve information while avoiding sparse or unstable categories. Where appropriate, we group variables into bands, retain meaningful "Unknown" categories, and pool categories with very few arrears events.

- Banded and continuous variables: We group LTV, mortgage term and number of bedrooms into bands to allow for non-linear relationships without imposing a specific functional form. The results tables report the relevant reference categories. We keep other continuous variables in their original form, except total gross income, which we use in logs because of its skewed distribution.

- Unknown categories: where categorical variables contain a large number of "Unknown" observations (principally borrower employment status and interest rate type), we retain this as a separate category rather than dropping these records.
- Impaired credit history: the impaired indicator equals 1 if any named borrower has a recent history of arrears, CCJ, bankruptcy or IVA, and 0 otherwise. (See Glossary definition.) We do not include specific impairment types separately because some groups are small. Most borrowers flagged as impaired have a history of arrears rather than a CCJ, IVA or bankruptcy.
- Mortgage switching: we code this variable to indicate whether the provider changes or the household remortgages with the same lender in the 6-monthly period, conditional on the borrower remaining at the same postcode.
- Lender categories: PSD contains around 150 lenders. A small number of lenders account for most arrears events, while many lenders record few or none. We therefore pool lenders with fewer than 100 arrears events over the full sample period into a residual category. This reduces instability from sparse cells in our model while also allowing us to control for lender fixed effects.

External data and matching

We match our data with local ONS Annual Population Survey model-based estimates of unemployment at the local authority level using property postcode. These are designed by ONS to overcome small sample issues at the local level. Since our mortgage data are half-yearly, we use unemployment in the 12 months to the end of each half year period (i.e. covering the current and preceding half-year period).

We match each property in our dataset with the UK House Price Index by local authority, year and dwelling type. Dwelling types in the HPI and PSD do not perfectly align – we code 'bungalows' in PSD within 'detached'. HPI does not provide a breakdown by dwelling type for Northern Ireland – we use the aggregate index for Northern Ireland postcodes. After matching, we estimate current estimated LTVs by uprating origination property value using the appropriate HPI and cumulative house price growth using the HPI over the life of each mortgage.

Finally, we match Base Rate using the account start date or, for subsequent observations, the date of balance.

Annex 2: Methodology details

Model specification

We estimate a discrete-time hazard model using a binary regression framework. The dependent variable equals 1 if a borrower enters arrears in a given period and 0 otherwise. Arrears are defined in PSD as a shortfall equivalent to two or more regular payments and we focus on the first observed arrears event.

We use a complementary log-log transformation, which models the log cumulative hazard rather than the probability of arrears directly. This gives the discrete-time model a proportional hazards interpretation, similar to a Cox proportional hazards model in continuous time. Coefficients indicate how each covariate shifts the hazard of entering arrears, holding other factors constant. We estimate the model by maximum likelihood on the person-period dataset in R.

The estimating equation is:

$$\text{cloglog}(h_{ij}) = \alpha_j + \mathbf{X}_i'\boldsymbol{\beta} + \mathbf{Z}_{ij}'\boldsymbol{\gamma} + \delta_r + \delta_c$$

where:

- $h_{ij}$is the probability that loan $i$enters arrears in duration interval j, conditional on not having entered arrears before the start of that interval;
- $\alpha_j$captures duration-interval effects, allowing baseline arrears risk to vary over the life of the mortgage;
- $X_i$contains borrower and loan characteristics measured at origination;
- Z_{ij}contains time-varying covariates, including local unemployment and current loan-to-value (see Annex 1);
- $\delta_r$captures region effects;
- $\delta_c$captures half-year reporting period.

The duration-interval effects allow the baseline arrears hazard to vary flexibly over the life of the mortgage. This avoids imposing a specific shape on how arrears risk changes with time since origination. We include region and half-year effects to control for broad geographic and age-of-loan differences in arrears risk.

We also include a flexible spline function of calendar time to absorb common time-varying factors, such as the interest-rate environment, macroeconomic conditions and policy changes. This avoids imposing a linear or otherwise restrictive form on aggregate time effects.

We include local unemployment and current loan-to-value ratio as time-varying covariates. This allows arrears risk to evolve with changing local economic conditions and property values, even though we do not observe all changes in borrower circumstances after origination.

We cluster standard errors at household level to account for repeated observations of the same borrower over time.

The baseline specification imposes proportional hazards. This means that each covariate has a constant multiplicative effect on the hazard across duration. We test this assumption and, where it fails for economically important variables, allow effects to vary with duration. In the reported model, we interact duration with current LTV band, gross interest rate, the number of dependent children and whether the current interest rate is incentivised.

We tested alternative specifications for interest-rate dynamics, including the change in interest rate since the previous period and the cumulative change since origination. These variables produced unstable coefficients that were sensitive to specification. In some cases, the coefficient on interest-rate change was negative. This likely reflects the fact that, conditional on the current rate, borrowers with larger increases often started from lower rates, rather than indicating that interest-rate increases reduce arrears risk. We therefore use the current interest rate level in the preferred specification.

We also tested two approaches to capturing changes in property values: using current loan-to-value in each period, or including LTV at origination alongside cumulative house price growth. The results were similar. We use current LTV in the preferred specification because it captures both the original underwriting position and subsequent changes in property values.

We exclude loan-to-income (LTI) and debt-service ratio (DSR) from the preferred specification following Variance Inflation Factor (VIF) diagnostics, which indicate problematic collinearity with the retained affordability measures. LTV, income, LTI and DSR capture closely related dimensions of borrower leverage and affordability. Including all of them produces unstable coefficients without adding clear independent information. Results for other covariates remain robust across alternative leverage specifications, but we interpret relative hazard estimates cautiously because there is no single definitive leverage measure.

Annex 3: Additional results and descriptive statistics

Descriptives

Table 3 shows descriptive statistics for categorical variables in our sample.

Table 4: Descriptives of categorical variables at origination in our sample (unique borrower observations)

Variable	Level	N	Share
First borrower employment status	Employed	626,399	91.8%
	Self-employed	50,341	7.4%
	Unknown	4,585	0.7%
	Retired	739	0.1%
Repayment type	Capital and Interest	679,913	99.7%
	Interest only	1,251	0.2%
	Mixed	900	0.1%
Interest rate type	Fixed	665,415	97.6%
	Tracker	13,911	2.0%
	Unknown	2,738	0.4%
Bedrooms	Unknown	69,121	10.1%
	2	210,049	30.8%
	3	282,960	41.5%
	1	55,818	8.2%
	4	56,083	8.2%
	5+	8,033	1.2%
Flexible feature mortgage	0	441,199	64.7%
	1	240,865	35.3%
Government supported initiative	0	594,388	87.1%
	1	87,386	12.8%
Advised sale	0	290	0.0%
	1	667,552	97.9%
New dwelling	0	14,512	2.1%
	1	561,197	82.3%
Borrower impairment history	0	120,595	17.7%
	1	272	0.0%

Source: FCA analysis of PSD
Notes: Variables evaluated at first observation of each household.

Main model results

Table 5: Survival model estimated hazards

Variable name	Estimated hazard	Standard error	Significance
(Intercept)	0.024	0.293	***
Age of first borrower	0.994	0.001	***
Self-employed borrower	1.291	0.020	***
Retired borrower	1.054	0.178	
Employment status unknown	1.296	0.059	***
Two assessed borrowers	1.019	0.045	
Three or more assessed borrowers	1.328	0.148	
Previous credit impairment	1.439	0.059	***
Dependent children	1.196	0.009	***
Dependent adults	1.037	0.041	
Gross income, log	0.703	0.024	***
Current LTV: 50–60%	1.025	0.040	
Current LTV: 60–70%	1.177	0.032	***
Current LTV: 70–80%	1.322	0.031	***
Current LTV: 80–90%	1.471	0.033	***
Current LTV: above 90%	1.777	0.041	***
Interest rate stress differential	1.033	0.007	***
Advised sale	0.901	0.045	
Government-supported initiative	1.243	0.020	***
Flexible mortgage feature	1.146	0.048	**
Direct sale channel	0.846	0.018	***
Incentivised current rate	0.616	0.035	***
Gross interest rate	1.207	0.015	***
Gross interest rate squared	1.002	0.002	
Mortgage switch	0.419	0.079	***
Remaining term: 200–300 months	0.780	0.037	***
Remaining term: 300–400 months	0.769	0.038	***
Remaining term: above 400 months	0.857	0.040	***
Interest-only repayment	1.257	0.153	
Part interest-only repayment	1.085	0.185	
Tracker rate	1.688	0.025	***
Interest rate type unknown	1.294	0.103	
New-build property	1.218	0.020	***
Two bedrooms	1.059	0.038	

Three bedrooms	1.107	0.038	**
Four bedrooms	1.312	0.042	***
Five or more bedrooms	1.724	0.057	***
Number of bedrooms unknown	4.061	0.064	***
Local unemployment rate	1.023	0.005	***
Log income : Middle loan age	1.038	0.031	
log income : Late loan age	1.179	0.103	
Current LTV: 50–60% : Middle loan age	1.378	0.047	***
Current LTV: 60–70% : Middle loan age	1.507	0.041	***
Current LTV: 70–80% : Middle loan age	1.583	0.043	***
Current LTV: 80–90% : Middle loan age	1.520	0.060	***
Current LTV: above 90% : Middle loan age	1.652	0.114	***
Current LTV: 50–60% : Late loan age	1.685	0.109	***
Current LTV: 60–70% : Late loan age	1.477	0.147	**
Current LTV: 70–80% : Late loan age	1.572	0.214	
Current LTV: 80–90% : Late loan age	1.246	0.336	
Current LTV: above 90% : Late loan age	0.000	0.074	***
Gross interest : Middle loan age	1.021	0.009	
Gross interest : Late loan age	1.035	0.031	
Dependent children – Middle loan age	1.009	0.010	
Dependent children – Late loan age	1.001	0.037	
Incentivised current rate – Middle loan age	1.134	0.044	**
Incentivised current rate – Late loan age	1.507	0.172	

Source: FCA analysis of PSD
Note: Significance levels are 0.001 (***), 0.001 (**). Excluded categories are: Employed borrowers; current LTV <50%; Remaining term <200 months; 1 bedroom; and early loan age for interaction terms

Additional results

Concordance

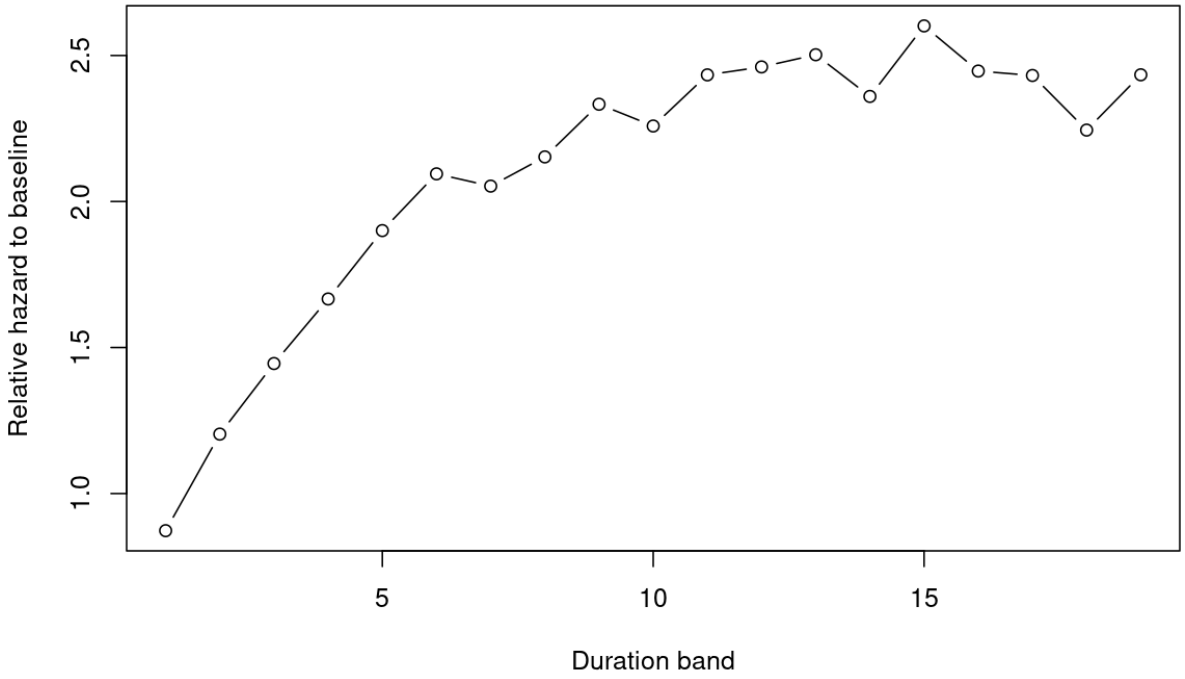
The model achieves a pooled person-period concordance of 0.73, meaning that when comparing a randomly chosen observation that went into arrears with one that did not, the model assigns a higher predicted hazard to the arrears case around three quarters of the time. This is in the range typically reported for arrears and default models built on origination data, reflecting the fact that by post-origination shocks that we do not observe (such as income and employment changes) drive a substantial share of arrears risk.

Duration dependence in arrears hazard

Figure 5 shows that the estimated arrears hazard rises with loan age, reaching around twice the baseline after roughly three years, before flattening into a higher but uneven plateau. The chart reflects both the underlying time dependence of arrears risk and the changing composition of who is still in the data. Each period is a six-month observation

band, so the horizontal axis represents a span of around 10 years (the maximum we could observe a household in our sample). The pattern suggests arrears risk is not concentrated immediately after origination but builds over time, as borrower circumstances change and financial buffers may erode. The pattern supports using flexible duration controls in the model, rather than assuming a constant baseline arrears risk over the life of the mortgage. Estimates at longer durations should be treated more cautiously because they rely on a smaller, more selected risk set.

Figure 5: How arrears risk evolves over the life of a mortgage



Source: FCA analysis of PSD.

Notes: chart shows duration dependence in the estimated baseline hazard by 6-month band.

Annex 4: Robustness checks

We report several robustness checks to assess how far key study limitations may affect the estimates.

Relaxing the proportional hazards assumption

The discrete-time hazard model assumes that covariate effects are constant across mortgage duration. This is the discrete-time equivalent of the proportional hazards assumption in a Cox model. To diagnose potential violations of this assumption, we estimated the Cox counterpart of our model and inspected Schoenfeld residual plots, which show residuals for each covariate against event time. A small number of variables showed slope or curvature, suggesting that their effects may vary with duration.

On this basis, we interact log income, current LTV band, number of borrowers, gross interest rate, number of dependent children and a current incentivised rate dummy with a coarsened duration term. The duration bands distinguish the early, middle and later stages of a loan's life and broadly match the turning points in Figure 5 above. We retained interactions where the coefficient pattern was economically interpretable and the terms were jointly significant.

We also tested duration interactions for interest rate type and number of dependent adults but did not retain them. The estimates appeared to reflect sampling noise rather than meaningful time variation.

For variables interacted with duration, the main coefficient gives the effect in the reference period: the first three years of the loan. The interaction terms show how that effect changes in later duration bands. Where this materially changes interpretation, we report effects by duration.

Endogenous time-varying conditions

In the model, the main time-varying conditions of interest are local unemployment and house prices. These vary at local level and are plausibly exogenous to an individual borrower's arrears status.

Time-varying mortgage characteristics from PSD007 are more complicated. Some variables, such as remaining balance and remaining term, largely reflect the original mortgage contract. Others, such as repayment type, may reflect borrower choices or firm actions that respond to unobserved financial stress.

This matters most for interest-only status. In the data, some mortgages in pre-arrears shortfall switch to interest only as a form of temporary forbearance. The risk is reverse causality: the model could associate interest-only status with higher arrears risk when the borrower's emerging financial difficulty caused the switch.

Joint modelling approaches can address this issue, but we take a simpler and more transparent approach. We limit the use of potentially endogenous time-varying covariates and favour origination characteristics where this risk is material. The main

remaining concern is the mortgage switch variable, since switching behaviour may itself be related to arrears risk.

Unobserved heterogeneity

Another identification concern is that borrowers who remain at risk at longer durations are not a random sample – they may look less risky – and this can lead to bias in covariate estimates that correlate with unobserved risk (e.g. factors like financial literacy, savings buffers, relationship stability). A related concern is selective censoring – lower risk borrowers may be more likely to move or exit our sample.

One way to assess this is to include a frailty term: a borrower-level random effect that captures underlying unobserved risk. Comparing specifications with and without the frailty term provides a diagnostic for unobserved heterogeneity, although it is less informative about selective censoring. A frailty variance significantly different from zero would suggest that unobserved heterogeneity is present.

Frailty models are computationally demanding on samples of this size. We therefore estimate the frailty specification on a random subsample of 20,000 borrowers, retaining all observation periods for those borrowers. This subsample is large enough to identify meaningful shifts in fixed-effect coefficients with reasonable precision.

The results show that borrower heterogeneity does exist (standard deviation is around 0.3 on cloglog, hazard ratio ≈ 1.35 per SD). However, Table 6 suggests that most estimated effects not are materially affected, though a small number (5 out of 39 estimates change by more than 10%). Overall, we conclude that unobserved heterogeneity does not appear to be causing major distortions.

Table 6: Comparison of frailty and non-frailty coeffcient estimates

Variable	No frailty	Frailty	Percentage change
(Intercept)	-0.94	-0.88	5.7
Age of first borrower	0.00	0.00	8.4
Self-employed borrower	0.21	0.21	1.2
Retired borrower	0.27	0.27	0.2
Employment status unknown	-0.06	-0.05	10.5
Two assessed borrowers	0.01	0.02	75.6
Three or more assessed borrowers	0.30	0.31	2.6
Previous credit impairment	0.71	0.71	0.1
Dependent children	0.13	0.13	0.0
Dependent adults	-0.04	-0.04	-9.4
Gross income, log	-0.45	-0.46	-1.2
Current LTV: 50–60%	-0.34	-0.34	-0.6
Current LTV: 60–70%	-0.04	-0.04	-4.2
Current LTV: 70–80%	0.04	0.04	2.6
Current LTV: 80–90%	0.07	0.07	6.4
Current LTV: above 90%	0.44	0.45	1.2

Interest rate stress differential	-0.01	-0.01	-13.8
Advised sale	-0.09	-0.09	0.2
Government-supported initiative	0.33	0.33	1.3
Flexible mortgage feature	0.11	0.11	1.2
Direct sale channel	-0.09	-0.09	0.9
Incentivised current rate	-0.29	-0.29	-0.6
Gross interest rate	0.19	0.19	-2.4
Gross interest rate squared	0.00	0.00	40.3
Mortgage switch	-0.80	-0.80	-0.2
Remaining term: 200–300 months	-0.50	-0.51	-1.0
Remaining term: 300–400 months	-0.55	-0.55	-0.9
Remaining term: above 400 months	-0.32	-0.33	-1.1
Interest-only repayment	-0.05	-0.06	-16.3
Part interest-only repayment	0.30	0.29	-2.5
Tracker rate	0.52	0.52	0.2
Interest rate type unknown	0.54	0.54	0.4
New-build property	0.24	0.24	2.0
Two bedrooms	0.13	0.13	1.1
Three bedrooms	0.13	0.13	1.2
Four bedrooms	0.37	0.38	1.7
Five or more bedrooms	0.79	0.79	0.8
Number of bedrooms unknown	1.41	1.40	-0.6
Local unemployment rate	0.04	0.04	2.5

Source: FCA analysis of PSD

Notes: N = 20,000

Out-of-sample validation

Out-of-sample validation checks whether the estimated relationships generalise to later cohorts, as the macroeconomic and mortgage market environment changes. This matters because the model is intended to inform forward-looking analysis.

We estimate the main model on loans originated between 2015 and 2024. As a robustness check, we re-estimate the model on originations up to and including 2019, then predict arrears hazards over a 36-month horizon using the training-sample coefficients. We estimate the model on the sampled data but apply the predicted hazards to the full dataset by adjusting the intercept.

The training specification predicts 94% of total arrears in the test period. The shape of arrears risk generally fits well over the first three years of loan life, although the model underpredicts more clearly in the second duration band (Table 7). We interpret this as evidence that the model is informative about the relative arrears risk of different borrower and loan types, and about the duration profile of risk. But the absolute hazards may be confounding noise in the model with structural differences in the sample before and after 2019 (e.g. because of the pandemic and the later rapid rise in base rates).

Table 7: Out of sample validation: actual arrears in 2020-2024 vs. those predicted from model on 2015-2019 data

Duration band	Observations (rows)	Arrears events	Predicted arrears	Ratio
0	1,547,228	2,592	2,003	0.77
1	1,539,326	2,764	1,800	0.65
2	1,359,044	2,849	2,637	0.93
3	1,204,246	2,894	2,890	1.00
4	1,028,336	3,042	3,220	1.06
5	903,932	2,740	3,327	1.21

Source: FCA analysis of PSD

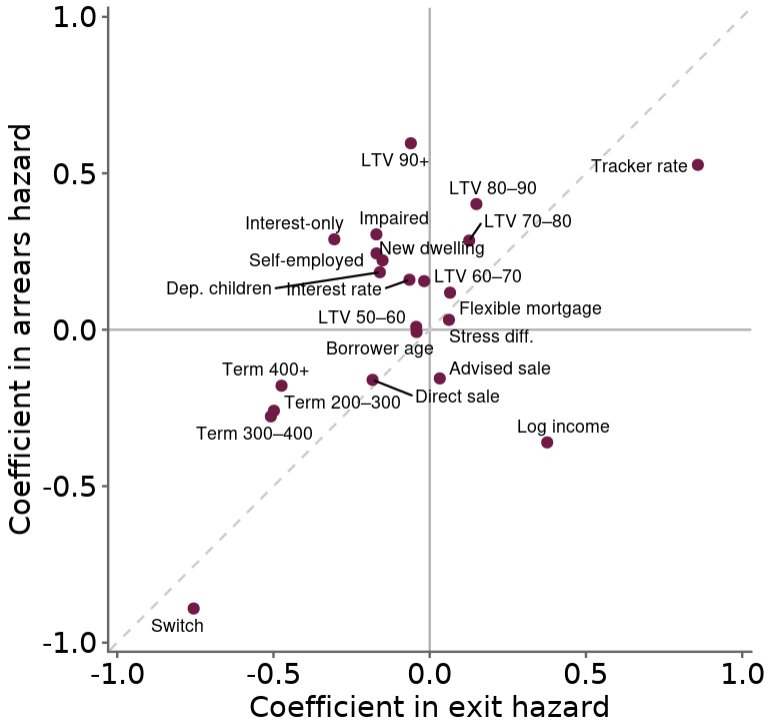
Competing risks

Borrowers may leave the observed risk set for reasons other than arrears, principally through voluntary redemption, remortgaging to another lender, or property sale. In a standard survival model, we treat these non-arrears exits as right censoring. This is valid if, conditional on covariates, borrowers who exit would have faced the same future arrears hazard as otherwise similar borrowers who remain in the risk set. Even where competing risks are present, relative risk estimates remain interpretable as cause-specific hazards.

To assess the potential for bias in the cause-specific hazard estimates, we estimate a companion model for non-arrears exit using the same specification. Following Deng et al. (2003), we also re-estimate the arrears hazard as a logistic model and treat non-arrears closures as a separate event. This allows us to examine whether the same covariates predict both arrears and exit, which would suggest that censoring may be informative.

While this is a diagnostic rather than a formal test, the results do not raise immediate concerns that selective exit is driving the main estimates (Figure 6). Bias depends on unobserved factors that are correlated across the arrears and exit processes, which we cannot observe directly. However, the pattern of coefficients provides a useful indication of whether selection on unobservables is likely to be material.

Figure 6: Comparison of arrears and non-arrears exit model coefficients



Source: FCA analysis of PSD

Annex 5: References

Aarland, K. & A. Santiago (2023), 'Staying Afloat or Going Under: Mortgage Arrears in Norway's Starter Mortgage Program', *Tidsskrift for boligforskning* 6(1)

Azimi A. & N. Khaledian (2025), 'Multi-stage mortgage default prediction using ensemble machine learning: a comparative framework', *Digital Finance*

Barbaglia, L. et al. (2023), 'Forecasting Loan Default in Europe with Machine Learning', *Journal of Financial Econometrics* 21(2)

Bergmann, M. (2020), 'The Determinants of Mortgage Defaults in Australia – Evidence for the Double-trigger Hypothesis', *Reserve Bank of Australia Research Discussion Paper* 2020-03

Bolliger, E. et al. (2024), 'Distressed Mortgages: A Machine Learning Assessment', Working Paper

Campbell, J. & J. Cocco (2015), 'A Model of Mortgage Default', *The Journal of Finance* 70(4)

Deng. Y et al. (2003), 'Mortgage Terminations, Heterogeneity and the Exercise of Mortgage Options', *Econometrica* 68(2)

Elul, R. et al. (2010), 'What "Triggers" Mortgage Default?', *American Economic Review* 100(2)

Ganong, P. & P. Noel (2023), 'Why do Borrowers Default on Mortgages?', *The Quarterly Journal of Economics* 138(2)

Gerardi, K. & K. Herkenhoff (2018), 'Can't Pay or Won't Pay? Unemployment, Negative Equity, and Strategic Default', *The Review of Financial Studies* 31(3)

Ke, G. et al (2017), 'LightGBM: A Highly Efficient Gradient Boosting Decision Tree'

Kelly, R. & F. McCann (2015), 'Some defaults are deeper than others: Understanding long-term mortgage arrears', *Central Bank of Ireland Research Technical Paper*

King, G. & L. Zeng (2001), 'Logistic Regression in Rare Events Data', *Political Analysis*

Linn, A, & R. Lyons (2020), 'Three Triggers? Negative Equity, Income Shocks and Institutions as Determinants of Mortgage Default', *The Journal of Real Estate Finance and Economics* 61

Reserve Bank of Australia (2019), 'Financial Stability Review – October 2019'

Scott A. & C. Wild (2001), 'Case-Control Studies with Complex Sampling', *Journal of the Royal Statistical Society* 50(3)

Slaymaker, R. et al. (2019), 'Monetary policy normalisation and mortgage arrears in a recovering economy: The case of the Irish residential market', *ESRI Working Pape*r 613

Stanga, I. et al. (2020), 'Mortgage arrears, regulation and institutions: Cross-country evidence', Journal of Banking & Finance 118

12 Endeavour Square, London E20 1JN
Telephone: +44 (0)20 7066 1000 Website: www.fca.org.uk