Technical Annex – Tracking Consumer Journeys

10/04/2026

Table of Contents

Annex Authors: Isabela Barra, Daniel Bogiatzis-Gibbons, Lawrence Charles, Wenjin Li

We would like to thank Dr. Raphael Sonabend-Friend (NICE) for input into an earlier survival analysis project, as well as Maria Jomy for undertaking that project. We would also like to thank the colleagues that reviewed this work and provided comments, as well as those that quality assured the work.

Data

Monthly Credit Reference Agency (CRA) data from February 2017 to February 2024, covering credit products, such as credit cards, loans and mortgages, and arrears records for over 400,000 consumers, is used to develop a refined and interpretable consumer segmentation. Survival analysis is then applied to better understand the risks of financial distress.

The data is grouped by consumer identifier in 6-month periods (we aggregate monthly observations into rolling 6-month windows; balances averaged; arrears and default taken as any occurrence). Two separate datasets were used here, one containing a 50% sample of consumers present in all data pulls (over 400,000 consumers), and the second with a 10% sample of all data pulls (over 6 million consumers), which includes consumers that are not present in later data pulls. The latter was only used to verify pre-exit behaviours and to check that leavers were not disproportionately in distress immediately before leaving the sample.

Rule-Based Consumer Segmentation

Motivation for Rule-Based Approach

A K-means clustering model was initially implemented to understand patterns and identify discernible sub-groups within the entire population (Steinley, 2006). To ensure model robustness over time, Gaussian Mixture Models (GMM) and Hierarchical Density-Based Spatial Clustering (HDBSCAN) were tested, with time stability and interpretability prioritised for supervisory use (Fraley, 2002; Campello et al., 2015). Clustering, however, did not reveal practically useful new groupings, and financial distress was difficult to discern with certain algorithms. When comparing these with a rule-based approach, the latter provided greater interpretability and time stability for segmentation in this case.

Cluster membership is less directly explainable than explicit rules. Even though it is possible to approximate explanations, for example, by using Shapley Additive Explanations (SHAP) values (Lundberg et al., 2020). In practice, post-hoc SHAP explains the surrogate classifier rather than the clustering objective. Clusters produced using unsupervised learning methods are not stable across different data pulls, as shown by low Jaccard overlap after re-fitting, resulting in differences in composition and overall proportions in the data. This is not useful when analysing cluster transitions over time.

Overall, the rule-based approach is more interpretable and supports casework and monitoring by helping supervisors understand consumer behaviour, track consumer journeys, and assess how different policies and firm actions affect consumers over time.

Segment Definitions

Table 1: Consumer Segment Definitions

Segment	Key Characteristics	Inclusion Criteria
Distress	Severe credit issues	CCJ, Bankruptcy, 3 or more months in arrears, default, debt buyer accounts
At Risk	Early warning signs	Recent missed payment, high utilisation, account instability, new unsecured debt
Secured Credit Users	Property owners	At least one active mortgage
Unsecured Credit Users	Active credit users	Two or more active credit accounts
Low Credit Engagement	No credit usage	Less than two active credit accounts

Precedence order: Distress overrides At Risk; At Risk overrides Secured Credit Users; Secured overrides Unsecured; Unsecured overrides Low Credit Engagement.

The consumer segments include recent credit behaviour. At any point in time, each consumer falls into one of the categories based on how they use credit and how well they manage repayments. Each segment is assigned hierarchically, with consumers classified into the first applicable category in Table 1, ensuring mutual exclusivity. If a consumer is in distress, they will not be in any of the other segments.

Segment Profiling

The following tables provide an overview of how consumers are distributed across segments and their typical credit behaviours. The largest cluster is low credit engagement, followed by secured credit users, unsecured credit users, distress and at-risk consumers. The latest Financial Lives Survey reported that, in May 2024, 8% of adults were in financial difficulty, a percentage slightly higher than the one presented here (6.4%) (FCA, 2025).

Table 2: Overall Segments' Proportions

Segment	Percentage
Low Credit Engagement	37.9%
Secured Credit Users	32.9%
Unsecured Credit Users	18.3%
Distress	6.4%
At Risk	4.4%

Credit Behaviour

Most consumers in distress (70.3%) have at least one account with three or more months in arrears, 49.2% have at least one account in default, 35.5% have a debt buyer account, 8.8% have a CCJ and 1.5% have an active bankruptcy.

Most consumers at risk (87.0%) have an account with 1 or 2 months in arrears, 74.0% have more than 2 missed repayments, 9.5% have an account with over 100% increase in arrear balance, and 4.5% of consumers had account instability, with 3 or more credit accounts opened in the last 3 months.

Most secured credit users have only one mortgage (57.8%), and 62.8% of consumers with low credit engagement do not have any credit accounts (current accounts and household bills were not considered).

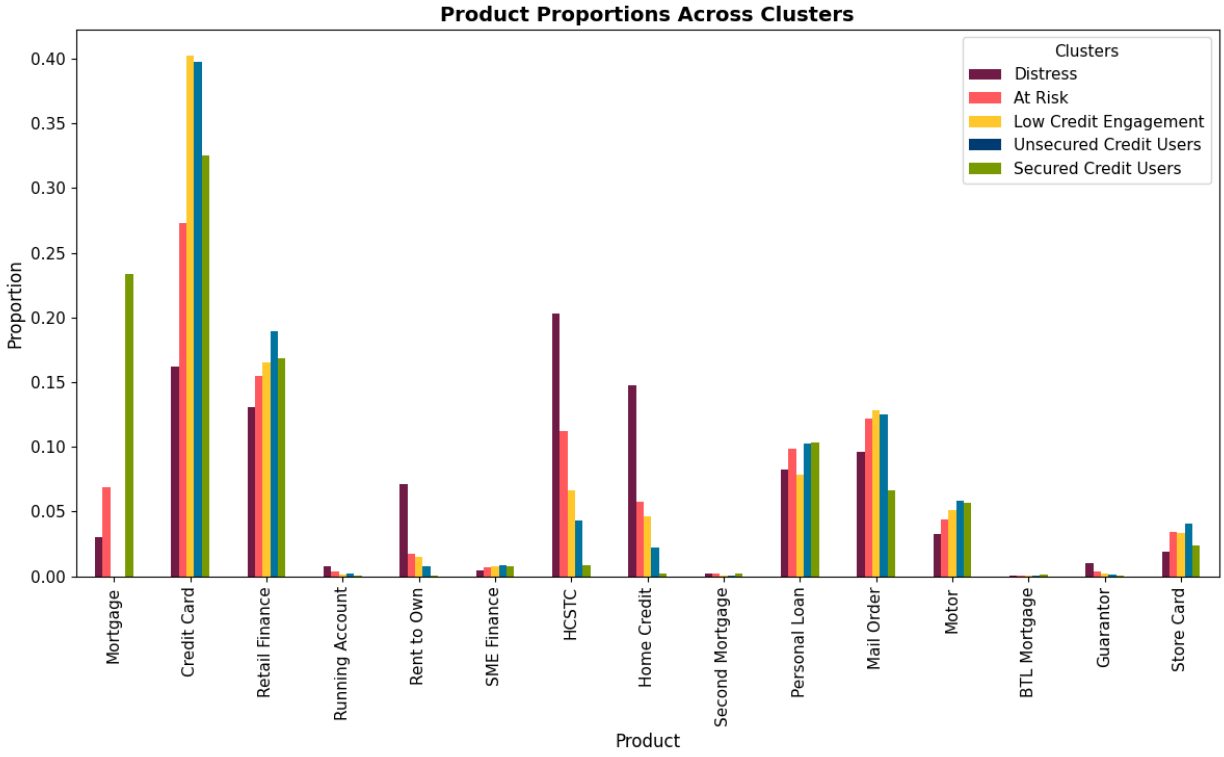
The median for different credit behaviour statistics is presented in Table 3. When comparing credit products, secured credit users have a significantly higher average balance, due to a higher proportion of mortgages.

Table 3: Credit Behaviour per Segment

Segment	Number of Accounts	Mortgage Count	Credit Card Count	Average Balance (£)
Distress	2	0	1	477
At Risk	3	0	2	793
Low Credit Engagement	0	0	0	0
Unsecured Credit Users	3	0	2	372
Secured Credit Users	3	1	2	13132
Overall	2	0	1	281

Consumers in financial distress have the highest proportions of high-cost short-term credit (HCSTC), home credit and rent to own, with the lowest percentage of credit cards and mortgages (Figure 1). Consumers at risk follow a similar pattern, with a high proportion of HCSTC and home credit, and a low percentage of credit cards and mortgages in comparison to other segments.

Figure 1: Credit Product Proportions Across Segments

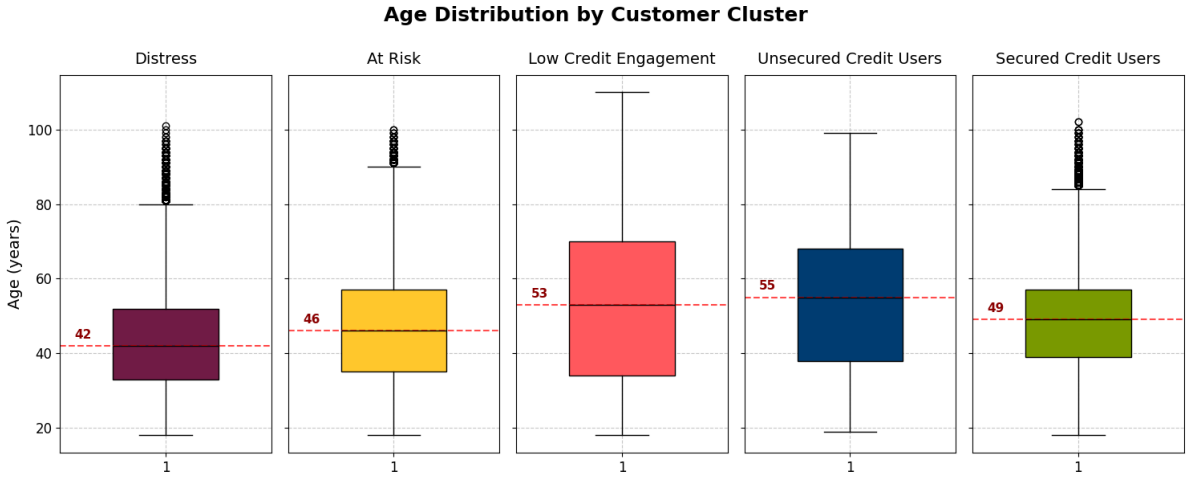


Aggregate product data from 2017 to 2024 for around 400,000 consumers.

Demographic Statistics

The average age across the dataset is 50 years. Figure 2 shows the age spread by consumer segment. Distress has the lowest median age (42), followed by consumers at risk (46), while the unsecured credit users cluster has the highest (54). The low credit engagement cluster has the largest spread of ages.

Figure 2: Age by Segment



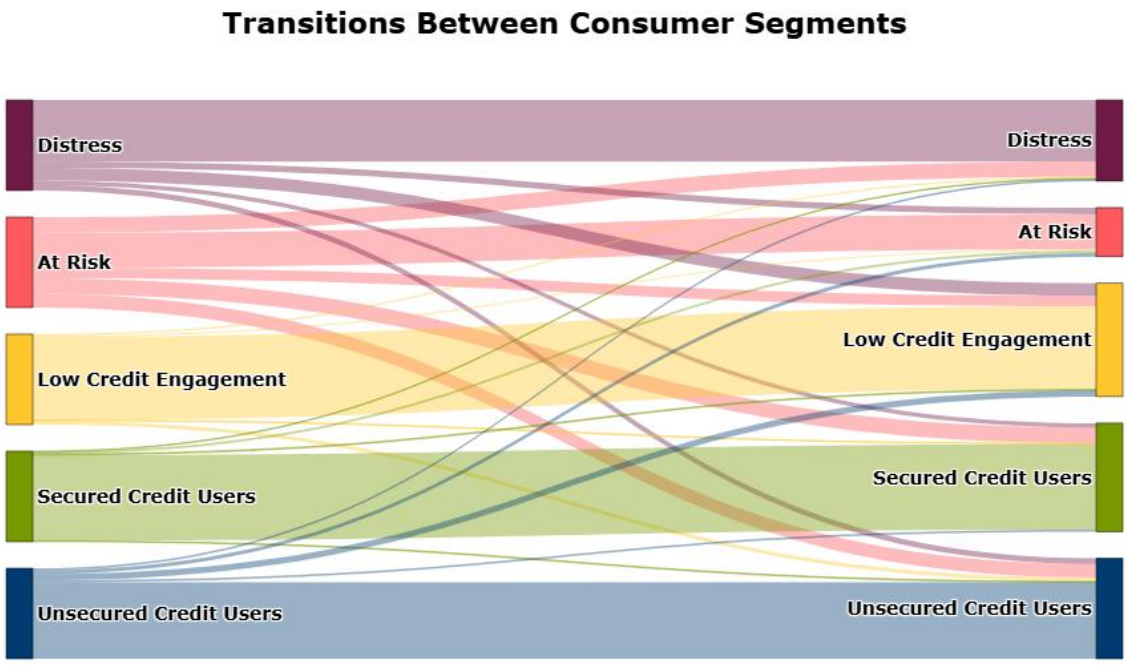
Aggregate demographic data from 2017 to 2024 for around 400,000 consumers. A consumer may appear in more than one segment based on their credit behaviour in the past 6-month period.

Transition Tracking Methodology

To understand the consumer journey, transitions between segments were measured on a 6-monthly basis from the start of 2017 to the end of 2024. For this analysis, any missing data was removed from the sample. Each individual transition was counted and the percentages of moves from each cluster calculated. Transitions include both movement between segments and persistence within a segment across two periods.

Over time, consumers often move between segments in response to changing financial behaviours. If payments are late or missed, or if there is high account utilisation with multiple new unsecured accounts, consumers are considered at risk of financial distress. If the situation worsens, with severe arrears, defaults or bankruptcy, for example, the consumer moves into financial distress.

Figure 3: Overall Segments' Transitions



Sankey diagram showing transitions between segments from t to t+1 (6-month windows) from 2017 to 2024 for around 400,000 consumers.

Table 4: Segments Transitions

Original Cluster	New Cluster	Percentage
At Risk	Distress	17.08%
At Risk	Low Credit Engagement	11.72%
At Risk	Secured Credit Users	17.20%

Original Cluster	New Cluster	Percentage
At Risk	Unsecured Credit Users	15.15%
Distress	At Risk	7.38%
Distress	Low Credit Engagement	13.61%
Distress	Secured Credit Users	4.90%
Distress	Unsecured Credit Users	6.25%
Low Credit Engagement	At Risk	1.60%
Low Credit Engagement	Distress	1.54%
Low Credit Engagement	Secured Credit Users	1.00%
Low Credit Engagement	Unsecured Credit Users	4.09%
Secured Credit Users	At Risk	2.40%
Secured Credit Users	Distress	0.67%
Secured Credit Users	Low Credit Engagement	1.12%
Secured Credit Users	Unsecured Credit Users	1.06%
Unsecured Credit Users	At Risk	4.25%
Unsecured Credit Users	Distress	2.27%
Unsecured Credit Users	Low Credit Engagement	6.86%
Unsecured Credit Users	Secured Credit Users	2.06%

From this aggregate data on transitions, the average retention of consumers per segment was calculated, and is as follows:

- At Risk: 3.8 months
- Distress: 12.8 months
- Unsecured Credit Users: 26.3 months
- Low Credit Engagement: 41.4 months
- Secured Credit Users: 51.5 months

Whilst consumers in most segments tend to stay in that category, consumers at risk are more likely to move to a different segment than remain at risk.

Consumers entering and leaving the sample were compared, using a sample with around a million consumers per data pull, with the overall goal to find whether the cluster consumers were in before leaving the sample differed from the overall. This was not the case, as their proportions stayed nearly identical.

Survival Modelling

Objective

Survival analysis models time-to-event data, and it has been traditionally used to model time-to-death in biostatistics. Here, it was implemented to model time-to-distress as defined above and assess the impact of different features in consumers falling into financial distress — especially to understand which features are associated with earlier or later signs of financial difficulty, helping anticipate emerging risks.

Models Used

We first use Kaplan–Meier estimators to describe survival by initial segment, then estimate adjusted risks using Cox Proportional Hazards (PH) and Random Survival Forests (Kaplan & Meier, 1958; Cox, 1972; Ishwaran, 2008).

The Kaplan-Meier estimator is a non-parametric statistic used to estimate the survival function. It provides a baseline estimate against which to compare model performance. Because it does not consider the effect of covariates in the survival function estimation, it is useful for identifying the effect of different features, such as different initial clusters, in the survival function.

The Cox Proportional Hazards Model is a semi-parametric model that assumes the hazard for a consumer is proportionally related to its other features, independent survival times between different consumers and constant hazard ratios over time. We fit a Cox proportional hazards model with time-fixed baseline covariates taken from the initial 6-month window; a sensitivity analysis uses time-varying covariates in a counting-process (start–stop) format. Elastic net regularisation, which combines Ridge and Lasso penalties, improved the model's predictability. The standard Cox model often fails when many features are included as it tries to invert a matrix that becomes non-singular due to correlations among features. Elastic net penalty uses a weighted combination of Ridge and Lasso, here only a small weight (0.9), to improve the stability of Lasso, reducing the number of features to a small subset of features that are most predictive.

Unlike Kaplan-Meier and Cox Proportional Hazards models, both of which are classical statistical models, Random Survival Forest is a machine learning model, which is more complex with lower explainability, but is less reliant on assumptions. It was used here as a non-linear benchmark, as it often achieves better results against evaluation metrics with high-dimensional, complex data in comparison to classical models. It is based on Random Survival Forests, creating and aggregating survival trees. The model parameters were optimised to balance performance and training time.

Feature Set

We construct features from rolling 6-month windows of monthly CRA data. Unless otherwise specified, features are calculated per consumer per window and are used either as baseline covariates (from the initial window) or, in a sensitivity analysis, as

time-varying covariates. The survival outcome is time to first entry into financial distress, measured in months from the first fully observed 6-month baseline window. Consumers without an event are right censored at their last observed window, and any observations after entry into financial distress are excluded from the risk set.

Table 5: Covariates in the analysis

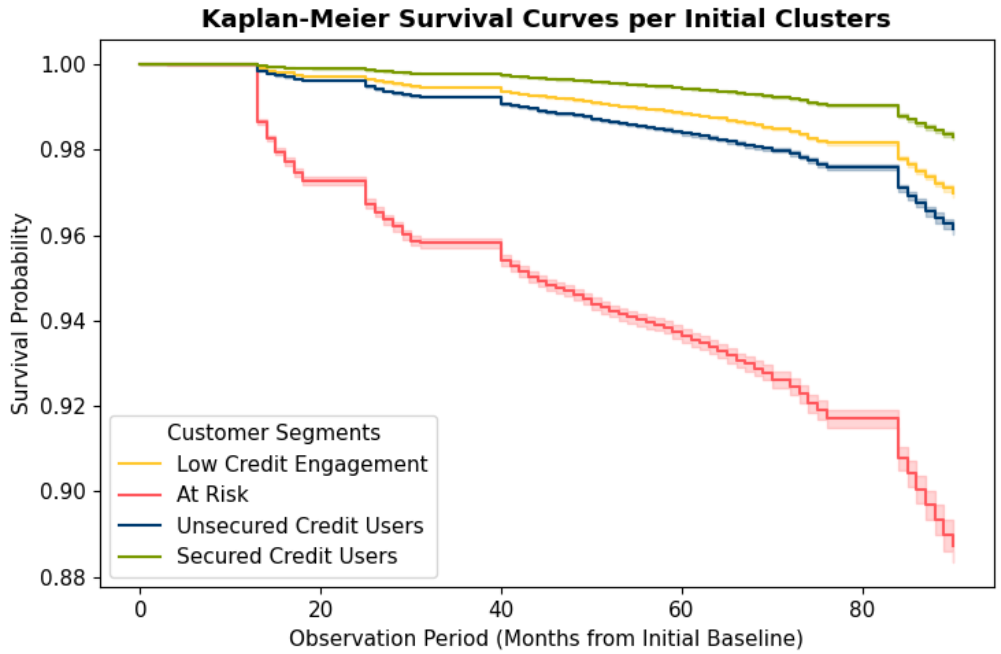
Covariate	Description
Active Accounts	Counts by secured and unsecured categories; total active accounts exclude household bills, telco and current accounts.
Product Counts	Individual columns per product type (excluding household bills, telco and current accounts).
Outstanding Balances	Average balance across all credit accounts.
Credit Limits	Average limit across all credit accounts.
Balance and Limit Change	Percentage change compared to the previous 6-month window.
Instability Flag	Three or more new credit accounts opened within the last 3 months.
Arrears Balance Growth Flag	Percentage increase over 100% in arrears balance.
Missed/Late Payments	Counts of accounts with at least one missed payment.
Age	Age at the start of the baseline window. We include age to control for lifecycle differences; no individual-level decisions are made using this variable.

The covariates were scaled for the Cox Proportional Hazards model using a standard scaler, which standardises features by removing the mean and scaling to unit variance.

Key Results

Consumers initially at risk have a lower survival probability (higher likelihood of falling into distress) than those in other segments. The survival probability at the end of the observation period, however, is still quite high at around 0.88 for consumers initially at risk, and it is not possible to estimate the median survival time per segment without large extrapolation (Figure 4).

Figure 4: Survival Curves per Initial Segment

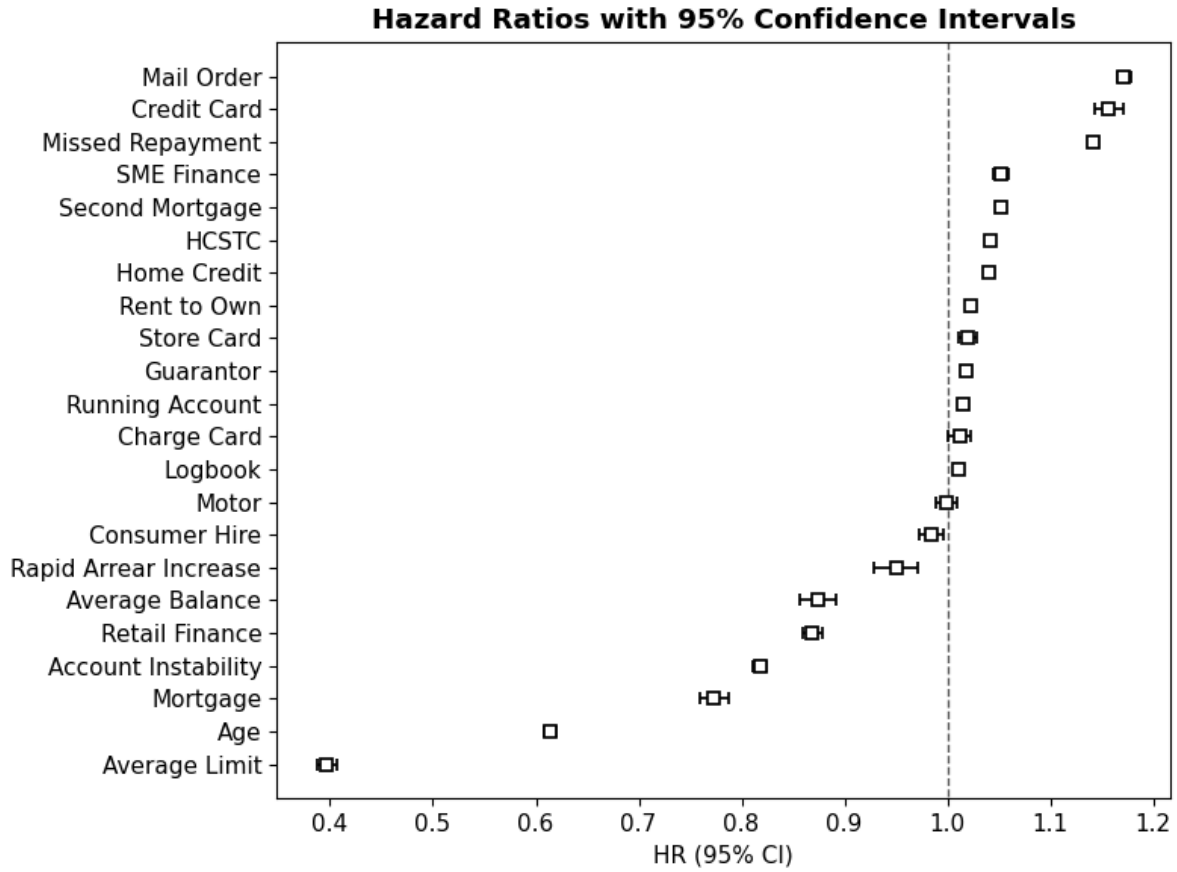


Survival curves for around 400,000 consumers. Time 0 is the start of the observation period (2017).

With the Cox PH model, hazard ratios (HRs) and statistical significance for each covariate in the move to distress can be analysed (Figure 5). The most statistically significant features are:

- Rent to own
- Home credit
- Mail order
- Account instability
- Missed repayment
- Average limit
- Age

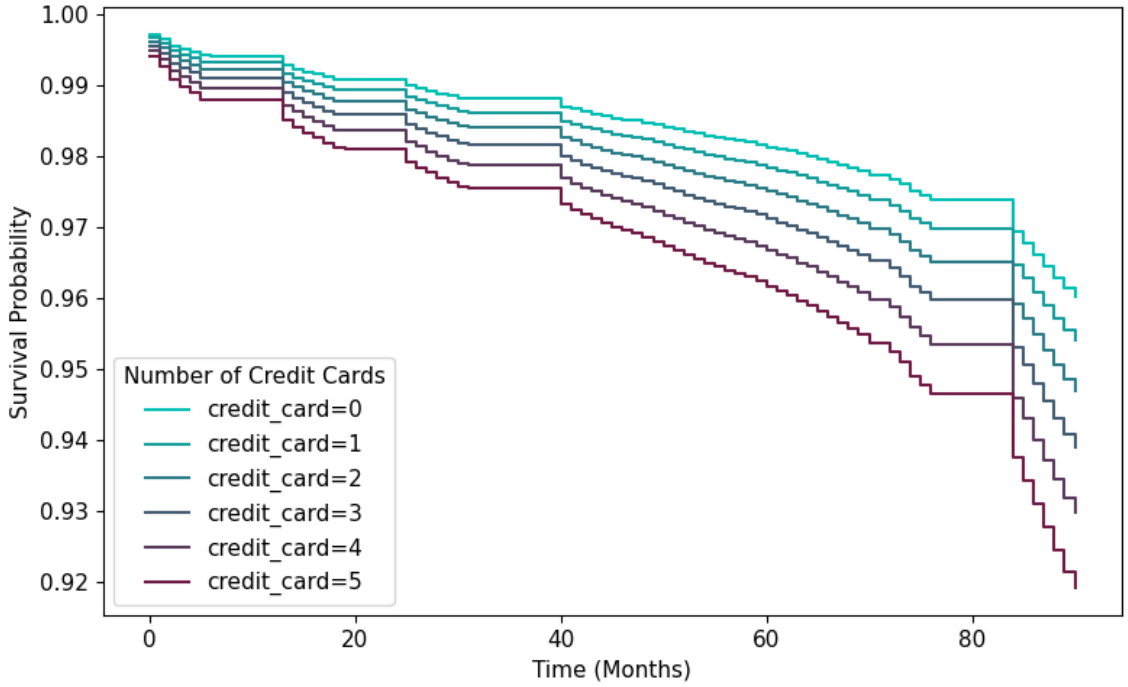
Figure 5: Hazard Ratios for Statistically Significant Features



Hazard ratios estimate the effect of covariates on event risk. Features with a hazard ratio greater than 1 increase the risk of falling into distress, while features less than 1 decrease this risk. This allows us to identify which credit behaviours are most predictive of early financial stress. Hazard ratios, however, are prone to selection bias and may mask significant time variation. It is important to analyse them in conjunction with survival curves and other evaluation metrics (Hernán, 2010).

The individual effects of features on the survival probability can be analysed with the Cox PH model, as it creates a baseline and only changes the specific feature to obtain survival functions for each individual change.

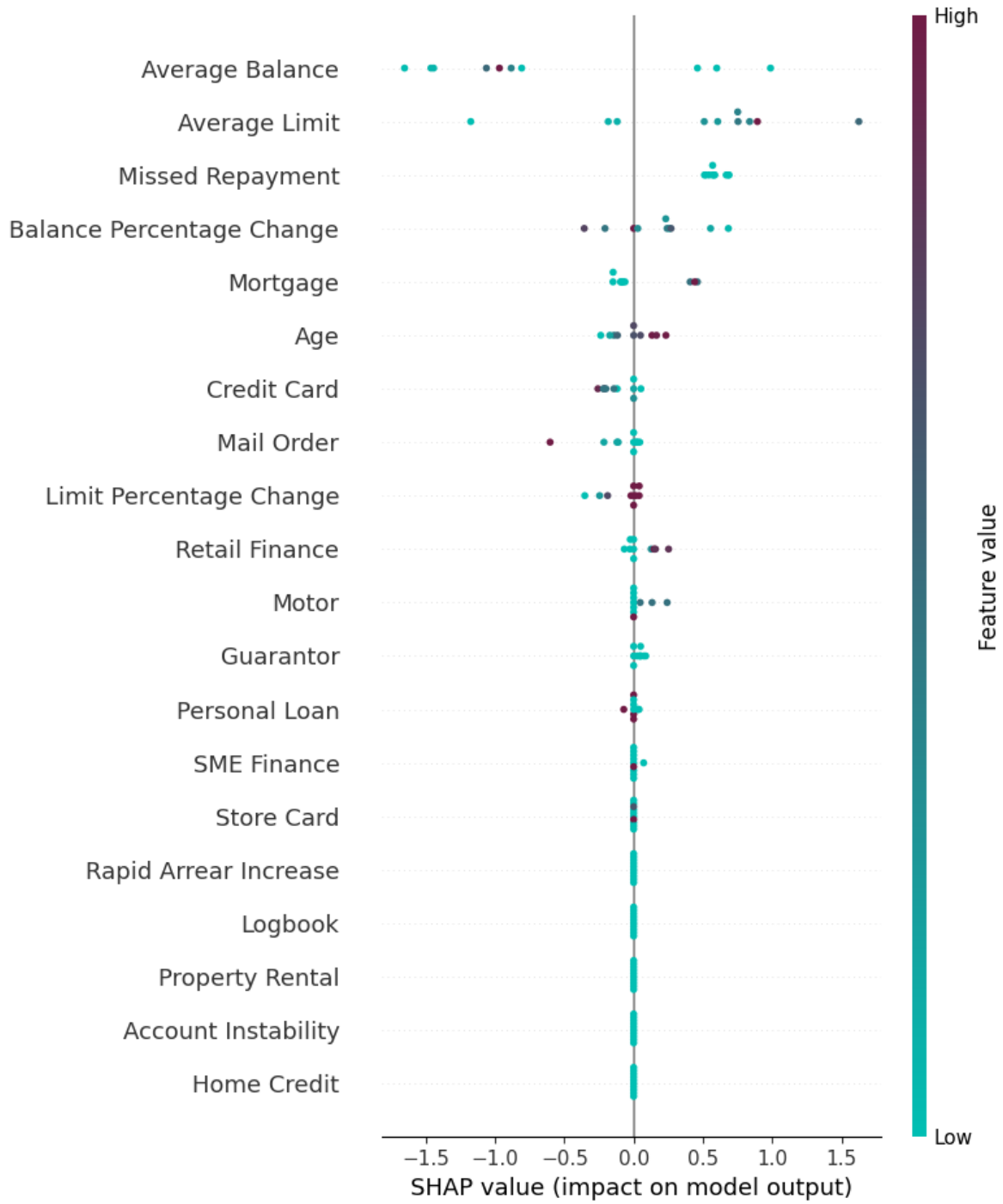
Figure 6: Effect of Credit Cards



Random survival forests achieved the highest results against the evaluation metrics. SHAP values, used to understand the impact of features in the move to distress, revealed a consistent feature ranking with the Cox PH model, enhancing confidence in the robustness of the findings. Features with positive SHAP values increase the risk of falling into financial distress, while features with negative values decrease this risk.

Figure 7 is ordered by most significant effects on financial distress risk (higher feature value). The top five features that increase the likelihood of distress are missed repayments, rent to own, mail orders, credit cards, and HCSTC, while the top five that decrease the likelihood of distress are average balance, average limit, percentage change in balance, mortgage and age.

Figure 7: SHAP Values from Random Survival Forest Model



Implementation Notes

The analysis was conducted in Python with the following libraries:

- NumPy and Pandas for data manipulation
- Matplotlib, Seaborn, Plotly and SHAP for data visualisation
- Lifelines and scikit-survival, which is built on scikit-learn, for survival analysis

Table 5: Training times per model

Model	Training Time (s)
Kaplan-Meier	0.1
Cox Proportional Hazards	20.2
Random Survival Forest	47852.0

Model Parameters

Cox Proportional Hazards Model:

- LASSO-to-ridge ratio for elastic net penalty: 0.9
- Step size: 0.005 (a small step size was required for convergence)

Random Survival Forest Model:

- Number of trees: 20
- Maximum tree depth: 15
- Minimum number of samples required for internal node split: 10
- Minimum number of samples in a leaf node: 20
- Maximum number of features considered for split: log2 of the number of input features

All other parameters for both models were kept as the default for the imported libraries, lifelines and scikit-survival, respectively.

Model Evaluation and Validation

Cross validation across models ensured consistent evaluation metrics, concordance index and brier score, and feature importance highlighted similar risk factors identified for both Cox PH and Random Survival Forest. The same train-test split (80:20) was also used for training and evaluation for consistency.

Concordance index (C-index) was the main evaluation metric. It assesses how well the model's predicted risk scores align with the actual observed order of events, thus evaluating the performance of a prediction model. Both the Cox PH and Random Survival

Forest model achieved high scores for this metric, 0.86 and 0.95, respectively. This indicates that both models have strong discriminative ability to correctly assess the risk of a consumer falling into financial distress, with the Random Survival Forest model showing superior predictive performance.

The Brier score is another metric that can be used to evaluate the accuracy of probabilistic predictions. It calculates the average squared difference between the predicted probability of survival and the actual observed outcome. This value was low for all three models at around 0.010, indicating the models' estimated survival probabilities closely matched the observed outcome with high overall prediction accuracy.

Statistical validation for Cox PH, with a log likelihood ratio test and a null hypothesis that none of the covariates had any effect on the risk of falling into financial distress, indicated the overall p-value for the model was under 0.005. Individual p-values for each feature were considered to evaluate statistical significance per covariate, with a null hypothesis that the feature has no effect on hazard, and only statistically significant features were considered.

Individual survival curves using the Random Survival Forest model for random samples were used to assess model behaviour at the individual level. The predicted curves closely matched observed event patterns, and consistency across different samples indicated stable performance. This also demonstrated that the model was able to capture relatively short- and long-term risk dynamics, supporting its suitability for this application.

Limitations

Sample composition varies over time due to consumers entering and exiting the dataset with new samples. A new sample may affect segment tracking. If across samples from different years, consumers are no longer present in the data, it is not possible to track the consumer journey over time. It would still be possible to analyse cluster proportions and behaviour over time, but not the effect of things such as new policies or financial crisis on consumers specifically.

We infer outcomes from credit behaviour, which may not always reflect underlying financial hardship. Certain consumers may be flagged as at risk or in distress, but not present financial difficulties. Similarly, secured credit users may find themselves in financial difficulties, but this would not be flagged in the data. These limitations highlight the importance of combining quantitative insights with other data sources and qualitative understanding of consumer experiences.

References

Campello, R. J. G. B., Moulavi, D., Zimek, A., & Sander, J. (2015). Hierarchical density estimates for data clustering, visualization, and outlier detection. *ACM Transactions on Knowledge Discovery from Data (TKDD),* 10(1), Article 5.

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society: Series B*, 34(2), 187–220.

FCA. (2025). *Financial Lives 2024: Key findings from the FCA's Financial Lives May 2024 survey*. Financial Conduct Authority. https://www.fca.org.uk/publication/financial-lives/financial-lives-survey-2024-key-findings.pdf

Fraley, C., & Raftery, A. E. (2002). Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association*, 97(458), 611–631.

Hernán, M. A. (2010). The hazards of hazard ratios. *Epidemiology*, 21(1), 13–15.

Ishwaran, H., Kogalur, U. B., Blackstone, E. H., & Lauer, M. S. (2008). Random survival forests. *Annals of Applied Statistics*, 2(3), 841–860.

Kaplan, E. L., & Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, 53(282), 457–481.

Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., & Lee, S.-I. (2020). From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence*, 2, 56–67.

Steinley, D. (2006). K-means clustering: A half-century synthesis. *Psychological Methods*, 11(1), 91–111.