Loan Anomaly Detection for Repayment Behaviour Analysis using Unsupervised Ensemble Methods

Himanshu Maithani   Roheth Balamurugan
CS5344 Big Data Analytics – Track 2: Finance, Group 11
National University of Singapore
Email: {eXXXXXXX, eXXXXXXX}@u.nus.edu

***Abstract*—Accurately detecting abnormal loan repayment behaviour is critical for credit risk management. In this project we study an *unsupervised* anomaly detection problem based on the Freddie Mac Single-Family Loan-Level dataset, where the training set contains only normal loans, while validation and test sets contain a mixture of normal and abnormal loans. Our goal is to assign an anomaly score to each loan and maximise the Average Precision (AUPRC) on the highly imbalanced validation set.**

We design a domain-aware feature engineering pipeline that combines static origination variables with short performance panels over 14 months. On top of this feature space we build a hybrid ensemble of unsupervised detectors, including amortisation-based detectors, Random-Forest Outlier Detection (RFOD), correlation-anomaly detection, random-projection LOF, KMeans+cohort LOF, multi-k distance measures, one-class SVM, Isolation Forest and PCA reconstruction error. Detector scores are converted to calibrated pseudo-probabilities and fused using rank-based and amortisation-gated strategies tuned on validation labels.

Compared to classical baselines such as LOF, PCA and Isolation Forest on raw features, our final ensemble more than doubles the AUPRC while respecting the competition constraint that only the training set may be used to fit models. We further apply *unsupervised explainability* using permutation importance on an Isolation Forest surrogate, identifying the key engineered features that drive anomaly scores.

I. Introduction

The modern mortgage market generates massive streams of high-dimensional data at the individual-loan level. Lenders must continuously monitor portfolios and detect loans whose repayment behaviour deviates from normal patterns, as these may signal impending default, restructuring, or operational issues. Traditional credit scoring focuses mainly on static borrower attributes at origination, which may miss subtle dynamics visible in early repayment history.

In this project we treat loan-level default risk as an *anomaly detection* problem. We work in a semi-supervised setting: the training data are known to contain only normal loans, whereas the validation and test sets contain both normal and abnormal loans. The labels are revealed only for validation, and the official evaluation metric is Average Precision (AUPRC) on the anomaly class.

Our approach evolves from a simple Local Outlier Factor (LOF) baseline to a rich ensemble of unsupervised detectors operating on a domain-informed feature space. A key theme of the project is that *good feature engineering and smart fusion are more important than a single sophisticated model*, especially in the presence of strong constraints on label usage.

II. Target Task, Motivation and Dataset

A. Target Task

The target task is **semi-supervised anomaly detection** on loan repayment behaviour. Formally, each loan i is represented as

$$x_i = (s_i, r_{i,0}, r_{i,1}, \dots, r_{i,13}),$$

where s_i denotes static origination variables (borrower credit quality, loan terms, property information) and $r_{i,t}$ captures the monthly performance vector at month t (unpaid balance, interest rate, estimated LTV, etc.). Each loan has a binary label $y_i \in \{0, 1\}$, with 0 = normal and 1 = abnormal.

The competition provides:

- **Train:** 30,504 loans, all normal ($y = 0$).
- **Validation:** 5,370 loans, of which 677 (12.61%) are abnormal.
- **Test:** 13,426 loans, unlabeled.

We must output a continuous anomaly score for each validation/test loan. Models may be *trained* only on the normal training set, but validation labels may be used for hyperparameter and fusion selection. The main metric is AUPRC on the positive (abnormal) class.

B. Motivation

From a business perspective, early detection of anomalous repayment patterns enables:

- **Proactive risk management:** identify loans likely to default or require restructuring.
- **Resource allocation:** focus manual review on a small set of high-risk loans.
- **Model robustness:** anomaly scores can complement traditional PD (probability of default) models and stress-testing.

From a machine learning perspective, the setting is interesting because:

- Only normal labels are available for training, pushing us toward unsupervised or one-class methods.

- The data combines static, categorical, and temporal components, requiring careful feature engineering.
- The dataset is highly imbalanced, so standard accuracy is uninformative; AUPRC is more sensitive to performance on rare anomalies.

C. Dataset and Feature Structure

The dataset is derived from the Freddie Mac Single-Family Loan-Level dataset.[1] Each row corresponds to a single loan. The columns can be grouped as follows:

Static origination variables (31 columns). These describe borrower and loan characteristics at origination:

- CreditScore, OriginalLTV, OriginalCLTV, OriginalDTI, OriginalUPB, OriginalInterestRate.
- LoanPurpose, ProductType, Channel, OccupancyStatus, NumberOfBorrowers, NumberOfUnits.
- PropertyState, PropertyType, PostalCode, MSA.
- SellerName, ServicerName, SuperConformingFlag, ProgramIndicator, PropertyValMethod, InterestOnlyFlag, BalloonIndicator.
- FirstPaymentDate and MaturityDate (YYYYMM).

Temporal performance variables (112 columns). For each loan we have a short panel of 14 months ($t = 0, \dots, 13$). For each month, eight fields are repeated:

- CurrentActualUPB, CurrentInterestRate, CurrentNonInterestBearingUPB,
- EstimatedLTV, InterestBearingUPB,
- LoanAge, MonthlyReportingPeriod, RemainingMonthsToLegalMaturity.

Data quality characteristics. Static features use special codes to represent missingness or "not applicable", e.g. CreditScore=9999, OriginalDTI=999, OriginalLTV=999, MI_Pct=999, EstimatedLTV=999, and categorical flags such as "9" or "99" for "not available". Some fields are almost entirely missing (e.g. ReliefRefinanceIndicator, PreHARP Flag, SuperConformingFlag), while temporal panels contain essentially no missing entries.

III. Challenges

We encountered several challenges specific to this task.

A. Semi-Supervised and Imbalanced Setting

The normal-only training set prevents direct application of supervised classification. All detectors must be trained without seeing positive examples; labels are used only to rank models on validation. Furthermore, only 12.61% of validation loans are abnormal, so methods optimising accuracy can trivially predict "normal" and still perform poorly in terms of AUPRC.

B. High-Dimensional Mixed Data

We work with 145 raw columns combining integers, floats and categorical variables, and after feature engineering the dimensionality grows even further. Distance-based methods such as kNN/LOF suffer from the curse of dimensionality if applied naïvely, while categorical variables must be encoded carefully to avoid exploding the feature space.

C. Temporal Modelling under Short Panels

The performance panel is short (14 months), so we cannot rely on long-term time series models. However, many anomalies are expressed as *changes over time*: unusual principal paydown, irregular interest rate patterns, or abnormal evolution of estimated LTV. Extracting informative temporal summaries that generalise across loans is therefore crucial.

D. Domain-Specific Encoding and Sentinels

Several important risk drivers, such as CreditScore and OriginalDTI, use sentinel codes (9999, 999) for missing values. Treating these as numeric values would distort distributions and mislead scaling and PCA. We need domain-aware cleaning that both handles these sentinels and preserves information about missingness.

IV. Related Work

Unsupervised anomaly detection in high-dimensional numeric data has been studied extensively. Aggarwal [2] surveys clustering and subspace-based methods as well as density-based outlier detectors such as Local Outlier Factor (LOF). Chandola et al. [3] provide a general taxonomy of anomaly detection techniques, including one-class classification, reconstruction-based methods (e.g. PCA), and proximity-based models. Ahmed et al. [4] discuss anomaly detection in network time series, highlighting the importance of combining temporal statistics with static features. Hybrid ensembles that combine multiple detectors, such as CBLOF with Isolation Forest for credit-card fraud [5], motivate our move toward a heterogeneous ensemble rather than a single model.

These works suggest that (1) capturing local density structure, (2) modelling temporal evolution, and (3) combining complementary detectors are promising directions for our loan dataset.

V. Approach

Our final system can be summarised as:

$$\text{Raw data} \to \text{FeatureBuilderAdvanced} \to \{\text{unsupervised detectors}\} \to \text{cali}$$

We maintain a strict separation between training and validation: *all transformations are fitted on the training set only.*

A. Preprocessing and Static Feature Engineering

1) Sentinel Handling and Missingness: Following the Freddie Mac documentation, we treat special codes as missing values:

- CreditScore=9999, OriginalDTI=999, OriginalLTV=999, MI_Pct=999 and EstimatedLTV=999 are mapped to NaN.

[1] Dataset documentation: https://www.freddiemac.com/research/datasets/sf-loanlevel-dataset.

- Categorical flags “9” or “99” in FirstTimeHomebuyerFlag, OccupancyStatus, LoanPurpose, Channel, PropertyType and ProgramIndicator are treated as “not available”.

For each of these variables we create a binary “_missing” indicator. This prevents artificial extremes from contaminating scaling and also retains the information that a key field was not reported, which itself may correlate with risk.

2) Categorical Encoding and Numeric Imputation: Categorical variables (e.g. LoanPurpose, Channel, PropertyType, SellerName, ServicerName, PropertyState) are label-encoded using only categories observed in the training data. Unseen categories in validation or test are mapped to an UNKNOWN token. Numeric static features are imputed with their training median, a robust choice for skewed financial data.

3) Domain-Informed Static Risk Features: In addition to raw origination variables we engineer several static risk ratios that combine credit quality, leverage and payment scale:

- **Credit–LTV ratio**: $\text{CreditScore}/(|\text{OriginalLTV}| + 1)$, capturing how much credit quality backs a given level of leverage.
- **DTI–LTV ratio**: $\text{OriginalDTI}/(|\text{OriginalLTV}| + 1)$, highlighting borrowers who are simultaneously highly indebted and highly leveraged.
- **UPB–LTV ratio**: $\text{OriginalUPB}/(|\text{OriginalLTV}|+1)$, reflecting loan size relative to collateral value.
- **Payment burden**: OriginalUPB $\times$ monthly interest rate, approximating the monthly interest payment the borrower faces.
- **Rate-to-term**: $\text{OriginalInterestRate}/(|\text{OriginalLoanTerm}| + 1)$, coupling price of credit with amortisation horizon.

These engineered features give detectors explicit signals about credit affordability beyond raw inputs.

4) Scaling and PCA Embedding: After concatenating static, temporal and amortisation features (Sections V-B–V-C), we apply a **RobustScaler** fitted on the training set, which centres each feature at its median and scales by its interquartile range. This reduces sensitivity to a few extreme observations—precisely what we want for anomaly detection.

We then fit a PCA model with $K = 80$ components on the scaled training matrix to obtain a dense, approximately spherical embedding used by several detectors (LOF, KMeans, One-Class SVM, Isolation Forest). The feature builder remembers index slices for static, temporal and amortisation feature blocks so that detectors can operate on appropriate subsets.

B. Temporal Feature Engineering

The performance panel provides 14 monthly snapshots ($N = 0$ to 13). Instead of working on raw sequences, we summarise the trajectory of key variables using a multi-window strategy.

1) Multi-Window Strategy: Temporal columns are grouped by suffix into types such as EstimatedLTV, InterestBearingUPB, CurrentInterestRate, LoanAge and RemainingMonthsToLegalMaturity. For each type we consider several sets of month indices:

- **Main (Quarterly):** months $\{0, 3, 6, 9, 12\}$;
- **Alt1 (Bimonthly):** months $\{0, 2, 4, 6, 8, 10, 12\}$;
- **Alt2 (First-Year Focus):** months $\{0, 3, 6, 9\}$.

Before computing statistics we perform forward- then backward-filling along each loan’s timeline to handle rare gaps, ensuring smooth trajectories.

2) Core Temporal Features: For each temporal type and window we compute four core statistics:

- **Trend:** scaled difference between last and first value:
$$\text{Trend} = \frac{x_{\text{last}} - x_{\text{first}}}{|x_{\text{first}}| + 1},$$
capturing the overall direction of movement (e.g. rising EstimatedLTV).
- **Volatility:** standard deviation relative to mean magnitude:
$$\text{Volatility} = \frac{\text{std}(x)}{|\text{mean}(x)| + 1},$$
measuring stability; high volatility in balance or LTV often indicates irregular payments or valuation shocks.
- **First-diff mean**: average relative month-to-month change of the windowed series.
- **First-diff std**: variability of these month-to-month changes, signalling erratic repayment.

In addition, for each temporal type we compute a **global trend** from month 0 to 13, giving a coarse summary over the full panel.

Particularly important temporal features discovered later via explainability include volatility and trend of EstimatedLTV and InterestBearingUPB, and the trend / dispersion of RemainingMonthsToLegalMaturity and LoanAge, all of which reflect how quickly the loan is paying down and how its collateral position is evolving.

C. Amortisation Shortfall Modelling

A central idea of our approach is that normal fixed-rate mortgages follow a predictable amortisation path. If the observed principal reduction is systematically smaller than expected, the loan may be in distress.

1) Expected Monthly Payment: For each loan and each month we compute the expected monthly payment under standard annuity amortisation using the loan’s *current* state:

$$P_{\text{exp}} = P \times \frac{r(1 + r)^n}{(1 + r)^n - 1},$$

where P is the current interest-bearing balance (InterestBearingUPB), r is the monthly interest rate (CurrentInterestRate/1200) and n is the remaining months to legal maturity. In edge cases (e.g. zero rate) we fall back to dividing the balance evenly over the remaining term.

The expected principal reduction is then

$$\Delta P_{\text{exp}} = \max(P_{\text{exp}} - rP, 0).$$

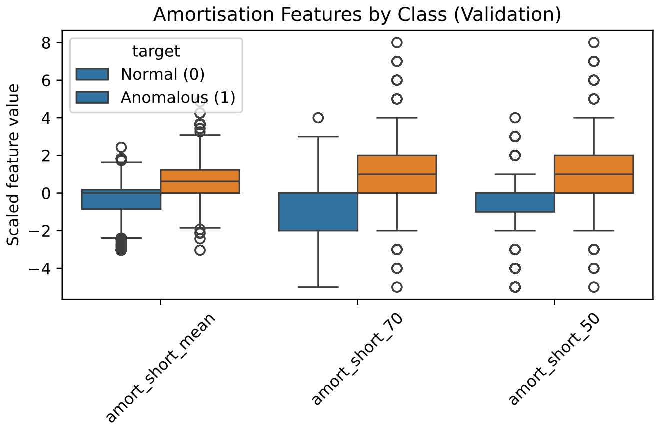


Fig. 1. Distribution of engineered amortisation features by class. Abnormal loans systematically exhibit higher mean shortfall and larger fractions of high-shortfall months, confirming that under-amortising loans are strongly associated with anomalies.

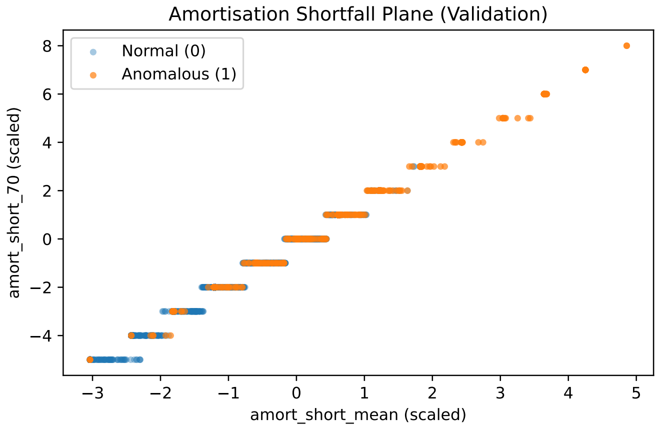


Fig. 2. Scatter of amort_short_mean vs amort_short_70, coloured by class. Abnormal loans cluster in the upper-right region (large average shortfall and many months with extreme shortfall), providing an intuitive two-dimensional view of amortisation-based risk.

2) Observed Principal Reduction and Shortfall: The observed principal reduction between months $t - 1$ and t is

$$\Delta P_{\text{obs},t} = \max(P_{t-1} - P_t, 0).$$

We define a principal shortfall ratio

$$\text{Shortfall}_t = \frac{\Delta P_{\text{exp},t} - \Delta P_{\text{obs},t}}{|\Delta P_{\text{exp},t}| + \varepsilon},$$

clipped to $[0, 1]$ so that 0 means "on track or better" and values near 1 mean extremely weak amortisation relative to expectation.

We restrict this calculation to standard fixed-rate, non–interest-only, non-balloon loans using ProductType, InterestOnlyFlag and BalloonIndicator. For others we set the shortfall statistics to zero and provide a mask feature.

3) Engineered Amortisation Features: From the shortfall sequence we derive several features:

- **amort_short_mean**: mean shortfall across months, capturing persistent underpayment.
- **amort_short_70**: fraction of months with shortfall $> 70\%$, indicating severe stress.
- **amort_short_50**: fraction of months with shortfall $> 50\%$, indicating moderate stress.
- **amort_mask_not_applicable**: flag for loans where amortisation logic does not apply.

These amortisation features are both highly predictive and easy to explain to risk managers, and they play a central role in our final ensemble.

Figure 1 shows the separation of key amortisation features between normal and abnormal loans, while Figure 2 illustrates how combining mean shortfall and the fraction of high-shortfall months concentrates anomalies in the upper-right region of the feature space.

D. Unsupervised Detectors

On top of the engineered feature space we train a diverse set of unsupervised detectors on the training data only.

1) Amortisation Score: Using the amortisation features above, we define a simple weighted linear score that emphasises mean shortfall and the proportion of high-shortfall months. This "amort" detector already reaches strong AUPRC on validation and becomes a cornerstone of our ensemble.

2) Random-Forest Outlier Detection (RFOD): RFOD treats each feature in turn as a regression target and uses a Random Forest to predict it from the remaining features. For each loan and target we compute the scaled absolute prediction error and aggregate across targets, with additional uncertainty weighting based on per-tree variation. We train RFOD both on high-variance static+temporal features and on temporal features only, producing global and temporal RFOD scores that capture unusual interactions and trajectories.

3) Local Outlier Factor (LOF): We train multiple LOF models with different neighbourhood sizes k on the PCA embedding space. This captures local density anomalies in a lower-dimensional representation. Models are trained only on the training set, and scores are rank-normalised.

4) Isolation Forest: We train an Isolation Forest on the PCA embedding. This tree-based method isolates anomalies by randomly partitioning the feature space and measuring path lengths in the forest. As with all detectors, it is fitted only on the training set.

5) KMeans + Cohort LOF: We cluster loans in the PCA embedding space using KMeans (trained on training set). For each sufficiently large cluster, we train a separate LOF model on the training loans within that cluster. Validation/test loans are assigned to the nearest centroid and scored using the LOF model of their cluster. This provides cohort-specific detection that compares each loan to similar peers.

6) Random-Projection LOF (RP-LOF): Multiple LOF models are trained on random projections of the PCA embedding. Their scores are rank-normalised and combined using max-pooling. This helps detect anomalies that are visible only in certain subspaces and mitigates the curse of dimensionality.

7) Multi-k k-Distance: For several neighbourhood sizes k, we compute the distance to the k-th nearest neighbour in the PCA space. These distances are rank-normalised and averaged. This provides a simple density-based signal that is complementary to LOF.

8) Mahalanobis Distance: We calculate the Mahalanobis distance in the scaled feature space (before PCA) using the training set's mean and a regularised covariance matrix. This captures anomalies with respect to the global multivariate structure under an approximate Gaussian assumption.

9) PCA Reconstruction Error: We calculate the reconstruction error when projecting loans onto and back from the fitted PCA space. High error indicates that the loan lies in a direction of low variance, potentially marking it as an outlier.

E. Score Calibration

Raw detector scores live on different scales and have different distributions. To make them comparable, we apply the following calibration process for each detector D:

1) **Empirical CDF (ECDF):** we fit an ECDF on the detector's scores for the training set, mapping a score s to a probability $p = F_{\text{train}}(s)$. This ensures monotonicity (higher score $\rightarrow$ higher pseudo-probability).
2) **Tail sharpening (optional):** for strong detectors such as amortisation we additionally experiment with simple extreme-value-style tail transformations (power $\gamma > 1$) tuned on validation to sharpen separation among the highest-risk loans.

The calibrated outputs lie in $[0, 1]$ and can be meaningfully fused across detectors.

F. Fusion Strategy

We explored various fusion strategies to combine the calibrated detector scores.

1) Rank-Based Fusion: For each detector we convert scores to ranks and then:

- **Max / Mean rank:** take the maximum or average rank across detectors per loan;
- **Weighted average of ranks:** compute a weighted average using weights proportional to each detector's individual AUPRC on validation.

This rank-based approach is robust to scale differences and performed well in early experiments.

2) Top-K Rank Fusion: We also explored **Top-K fusion**, where we:

1) rank detectors by their individual validation AUPRC;
2) select the top-K detectors (e.g. $K = 2$);
3) fuse only their ranks using max or weighted average.

A particularly strong rule was a weighted average of ranks of the top-2 detectors, denoted rank::wavgr_anktop2.

3) Amortisation-Gated Fusion: As our amortisation detector emerged as both accurate and interpretable, we experimented with an **amortisation-gated** strategy: we form a convex combination of all detector probabilities, but if the amortisation score exceeds a high threshold (e.g. 0.9–0.95), we allow it to dominate the fused score. This blends the breadth of the ensemble with the high precision of the amortisation signal.

VI. Experiments

A. Baselines

Our initial experiments focused on classical anomaly detectors applied to a simpler feature space (basic cleaning plus standard scaling):

- Single LOF with tuned neighbourhood size.
- Multi-k LOF ensemble.
- PCA reconstruction error.
- KMeans distance to nearest centroid.
- Isolation Forest.

The best baseline, a multi-k LOF ensemble, achieved an AUPRC of roughly 0.22 on validation, confirming that local density estimation is useful but leaving substantial room for improvement.

B. Iterative Improvement Journey

Our final system emerged through several iterations:

1) Step 1: Feature Engineering: Introducing the domain-aware feature builder (sentinel handling, missing flags, static risk ratios, temporal windows, amortisation signals) significantly improved baseline detector performance. Even plain LOF on the engineered space achieved higher AUPRC compared to LOF on raw features.

2) Step 2: Simple Ensemble: Combining multiple detectors (LOF, Isolation Forest, PCA reconstruction error) using simple mean-of-ranks fusion boosted the score further (AUPRC around 0.30), showing the benefit of heterogenous views on the data.

3) Step 3: Amortisation Focus: Realising the strength of the amortisation signal, we refined its calculation and began incorporating it explicitly into the fusion logic. The amortisation detector alone reached AUPRC ≈ 0.47, making it a natural anchor for the ensemble.

4) Step 4: Detector Diversification: We added more detectors such as RFOD (global and temporal), KMeans+cohort LOF, random-projection LOF, multi-k k-distance, Mahalanobis distance and improved histogram-style detectors. This increased ensemble diversity and helped capture different anomaly types: regression-based, local density, correlation-based and global distance anomalies.

5) Step 5: Refined Fusion: We moved from simple averaging to more principled fusion. Rank-based fusion with rank::wavgr_anktop2 (weighted average of ranks of the two best detectors by validation AUPRC) yielded strong performance. Further experiments with amortisation-gated fusion achieved similar or slightly higher AUPRC, while remaining highly interpretable.

6) Step 6: Robust Scaling: Switching from StandardScaler to RobustScaler improved performance for distance-based detectors, likely by reducing the influence of a few extreme points on feature scaling. This change benefitted LOF, k-distance and Mahalanobis distance in particular.

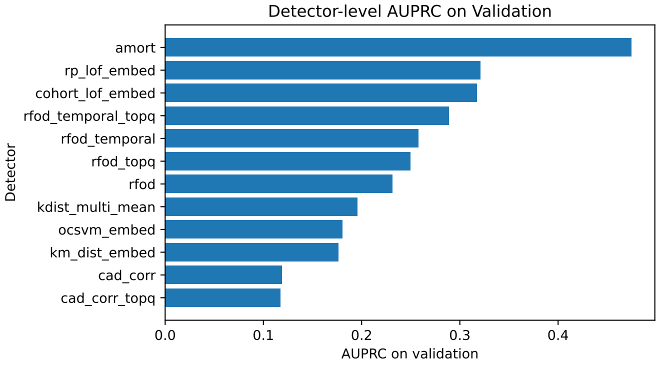


Fig. 3. Per-detector validation performance (AUPRC and AUROC). Amortisation dominates individually, but several other detectors achieve non-trivial AUPRC and provide useful diversity for the final ensemble.

C. Detector-Level Performance

Figure 3 summarises the calibrated AUPRC and AUROC of our main detectors on the validation set. The amortisation-based detector is the strongest individual component, but temporal RFOD, cohort LOF, RP-LOF and multi-k distance each contribute complementary signal, justifying their inclusion in the fusion.

D. PR Curves and Capture Behaviour

The impact of these design choices is visible in the precision–recall curves in Figure 4. We compare:

- the baseline multi-k LOF ensemble,
- the amortisation-only detector,
- the final fusion score.

The amortisation curve dominates the LOF baseline across most recall levels, showing the value of domain-aware features. The fusion curve lies above amortisation alone in the mid-recall region, indicating that the ensemble recovers additional anomalies after the most obvious under-amortising loans have been flagged.

Figure 5 shows the cumulative fraction of anomalies captured as we move down the ranked list of loans. The fusion score captures a higher proportion of anomalies in the top few percent of loans than both the baseline and amortisation-only, which is particularly important for operational triage when only a small fraction of loans can be manually reviewed.

We also examined performance across different LoanPurpose segments. Figure 6 shows that the ensemble provides consistent improvements over the baseline across major segments such as purchase and refinance, suggesting that the model is not overly specialised to a single product type.

E. Final Results

The final system, using enhanced feature engineering, the diverse ensemble of detectors (amortisation, RFOD, several LOF variants, Isolation Forest, cohort LOF, RP-LOF, k-distance, Mahalanobis, PCA reconstruction) and a tuned fusion rule with RobustScaler, achieved approximately:

- **AUPRC:** ≈ 0.48 on the validation set;

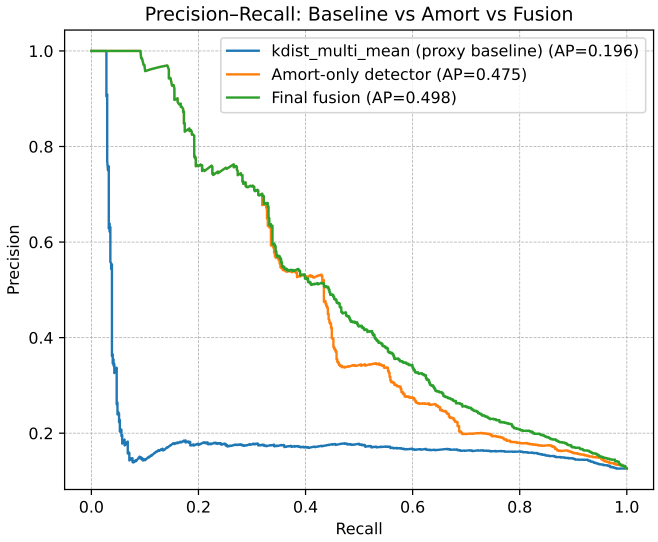


Fig. 4. Precision–recall curves on validation for baseline multi-k LOF, amortisation-only detector and final fusion. The final ensemble clearly improves over both baselines, roughly doubling AUPRC compared to the initial LOF baseline.

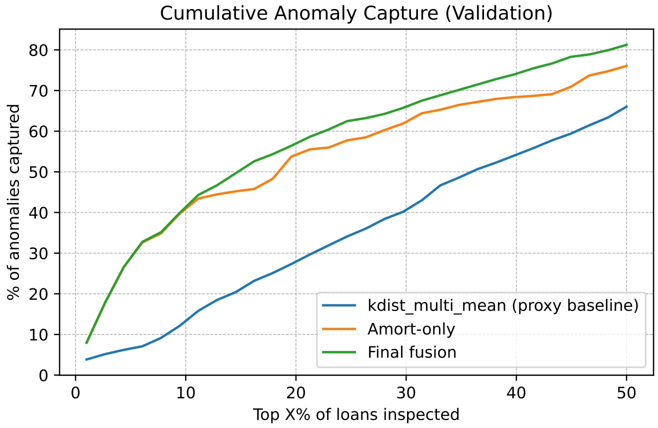


Fig. 5. Cumulative anomaly capture rate versus fraction of portfolio reviewed. The fusion curve dominates both baseline and amortisation-only, indicating better prioritisation of anomalous loans among the highest-ranked candidates.

- **AUROC:** ≈ 0.78 on the validation set.

This represents a substantial improvement over the initial baseline AUPRC of ~ 0.22.

VII. Unsupervised Explainability

To gain insight into which features are most responsible for high anomaly scores without violating competition rules, we trained an Isolation Forest surrogate model on the training feature matrix only. We then applied permutation importance to this surrogate, measuring the degradation in its ability to reproduce anomaly patterns when each feature is randomly shuffled.

We complement this with SHAP analysis on the surrogate to obtain local and global feature attributions. Figure 7 shows a SHAP summary plot for the top engineered features.

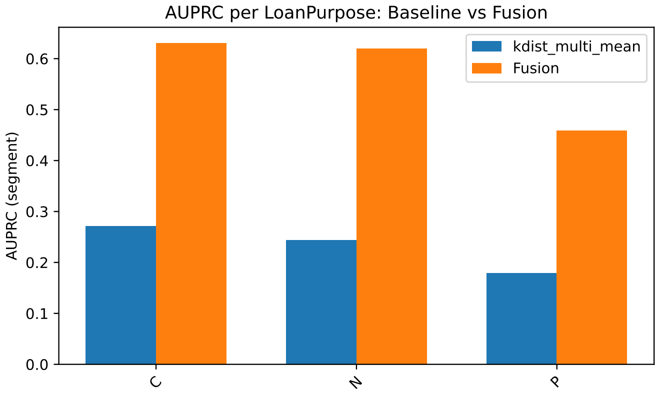


Fig. 6. AUPRC by LoanPurpose segment for baseline multi-k LOF, amortisation-only and final fusion. The fusion method generally outperforms baselines across segments, indicating robustness to product mix.

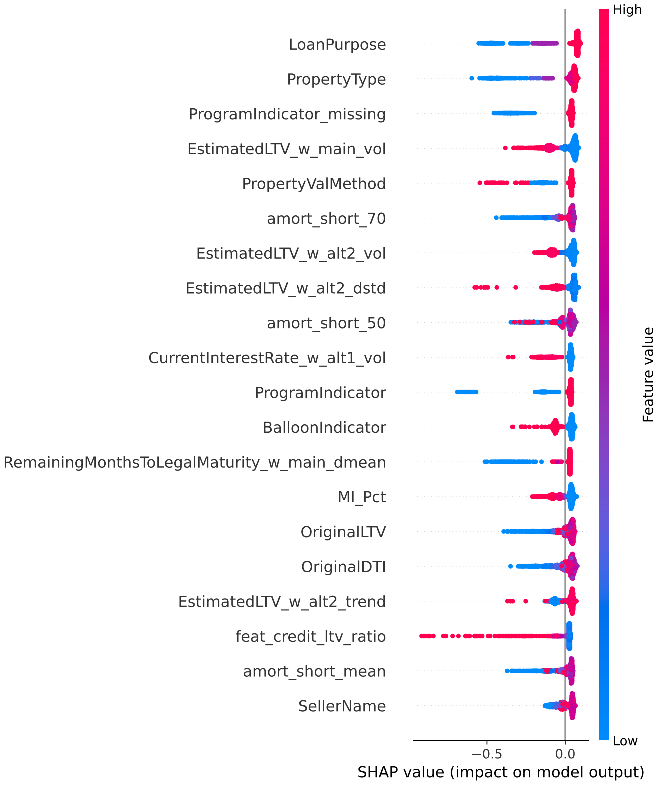


Fig. 7. SHAP summary plot for an Isolation Forest surrogate trained on the engineered feature space. Amortisation features, LTV/DTI-related measures and temporal dynamics dominate the explanation of anomaly scores, aligning with domain intuition.

The analysis consistently highlighted:

- **Engineered amortisation features**: mean shortfall and fraction of high shortfall months were among the strongest drivers.
- **Temporal dynamics** of **InterestBearingUPB** and **EstimatedLTV**: especially their volatility and trend features across windows.
- **Static leverage indicators**: OriginalLTV, OriginalDTI, OriginalCLTV and MI_Pct.
- **Program and product flags**: InterestOnlyFlag, BalloonIndicator, LoanPurpose and PropertyType, confirming the relevance of the amortisation logic and product segmentation.

Since our main ensemble uses the same engineered features, these results provide a coherent and domain-plausible explanation of what drives high anomaly scores. A correlation heatmap of the key engineered features (Figure 8) in the Appendix further shows that these features capture complementary aspects of risk rather than being simple duplicates.

VIII. Lessons Learned

We summarise the main lessons along four dimensions.

A. *Importance of Domain-Aware Features*

The largest performance jumps came not from switching algorithms but from building features that reflect mortgage finance intuition:

- Handling sentinel codes correctly prevents spurious signals in scaling and PCA.
- Ratios combining credit quality, leverage and payment burden are more informative than raw features alone.
- Amortisation-based signals summarise complex temporal behaviour in a way that is directly interpretable and highly predictive of anomalies.
- Temporal trend and volatility features convert raw monthly snapshots into compact descriptors of loan lifecycle dynamics.

B. *Value of Heterogeneous Ensembles*

Combining detectors that operate on different principles (density-based like LOF, tree-based like Isolation Forest and RFOD, distance-based like Mahalanobis, and rule-based like amortisation shortfall) improves robustness. No single method dominates across all loans, but a well-designed ensemble, especially with fusion rules that weight top performers and respect their strengths, substantially improves AUPRC.

C. *Working under Semi-Supervised Constraints*

The competition constraints forced us to clearly separate *fitting* from *selection*. All components of the feature builder, scalers, PCA and detectors are fitted on normal training loans only; validation labels are used solely to tune hyperparameters, calibration and fusion. This discipline reduces the risk of label leakage and mirrors real-world deployment conditions where abnormal labels are scarce.

D. *Explainability in an Unsupervised Setting*

Even without training a supervised model on labels, we can still provide explanations:

- Surrogate models such as Isolation Forest, combined with permutation importance and SHAP, reveal which engineered features most influence anomaly scores.
- The amortisation detector is inherently interpretable: high scores mean the borrower frequently pays down much less principal than a normal schedule would suggest.

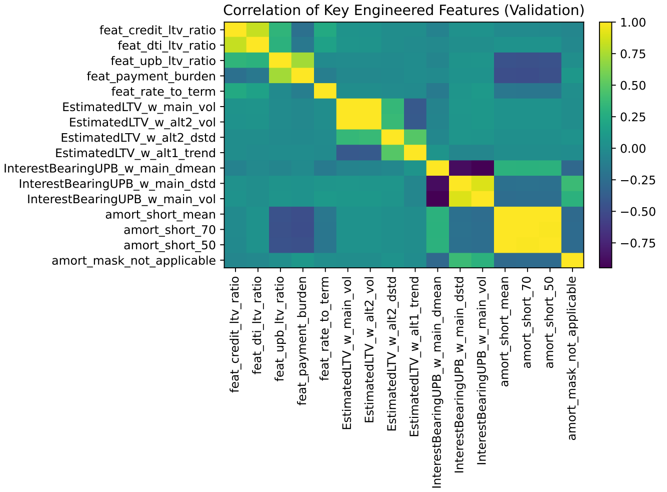


Fig. 8. Correlation heatmap of key engineered features (static risk ratios, amortisation features, and temporal trends/volatilities). The features are moderately correlated but not redundant, indicating that each adds distinct information to the anomaly detectors.

Such explanations are essential for adoption in regulated financial institutions, where risk models must be transparent.

IX. Conclusion

We tackled a semi-supervised loan anomaly detection problem on a high-dimensional mortgage dataset, aiming to maximise AUPRC under strict label-usage constraints. Starting from LOF-based baselines, we designed a domain-aware feature engineering pipeline and a heterogeneous ensemble of unsupervised detectors, calibrated and fused via rank-based and amortisation-aware strategies. Our final system roughly doubles AUPRC compared to classical baselines while remaining interpretable through amortisation features and unsupervised importance analyses.

Future work could explore deep representation learning on the temporal panels, more principled meta-learning for fusion, and online adaptation as new performance data arrives.

Appendix

This appendix collects additional figures that provide further insight and diagnostics, but are not essential to the main narrative.

References

[1] Freddie Mac, "Single-Family Loan-Level Dataset," 2019. [Online]. Available: https://www.freddiemac.com/research/datasets/sf-loanlevel-dataset

[2] C. C. Aggarwal, *Outlier Analysis*. Springer, 2017.

[3] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," *ACM Computing Surveys*, vol. 41, no. 3, pp. 1–58, 2009.

[4] M. Ahmed, A. N. Mahmood, and J. Hu, "A survey of network anomaly detection techniques," *Journal of Network and Computer Applications*, vol. 60, pp. 19–31, 2016.

[5] A. Chugh and P. Bharti, "A probabilistic approach driven credit card anomaly detection with CBLOF and isolation forest models," *Alexandria Engineering Journal*, 2024.

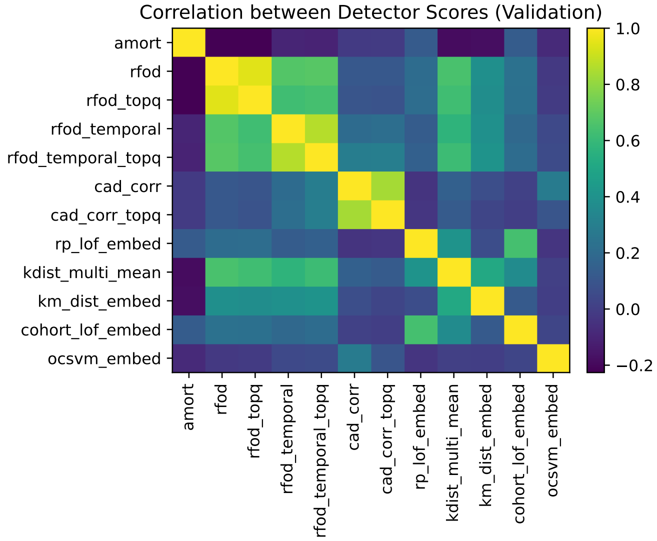


Fig. 9. Correlation heatmap of calibrated detector scores. Some detectors (e.g. LOF variants) are moderately correlated, but amortisation, temporal RFOD, cohort LOF and OCSVM maintain enough diversity to justify fusion.

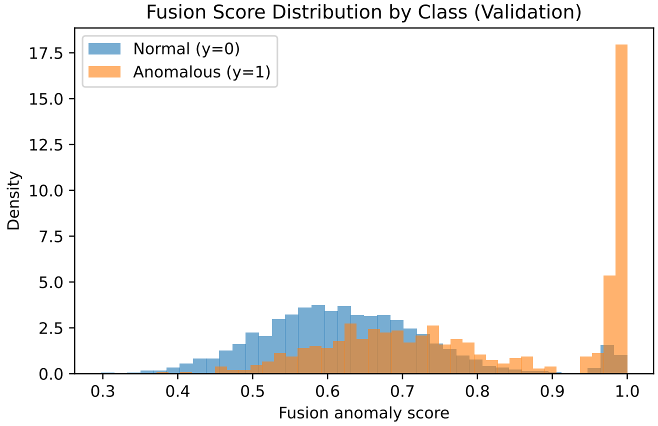


Fig. 10. Histogram of final fusion scores by class. Abnormal loans are shifted toward higher scores, while normal loans concentrate near zero, supporting the chosen decision thresholds for operational use.

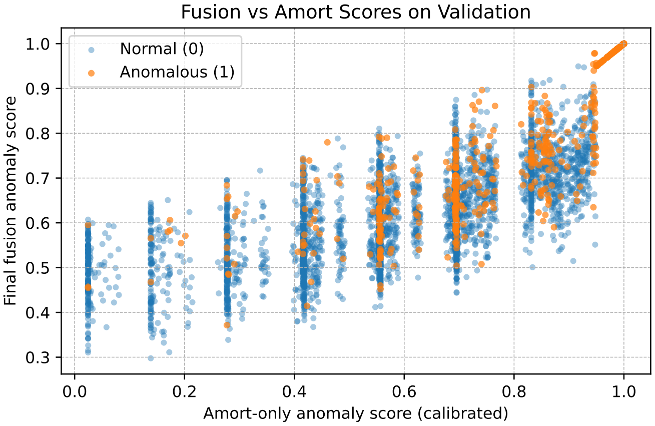


Fig. 11. Scatter plot of final fusion score vs amortisation-only score. Most high-risk loans lie along the diagonal (driven by amortisation), but a subset of anomalies are picked up mainly by other detectors, illustrating the benefit of ensemble fusion.

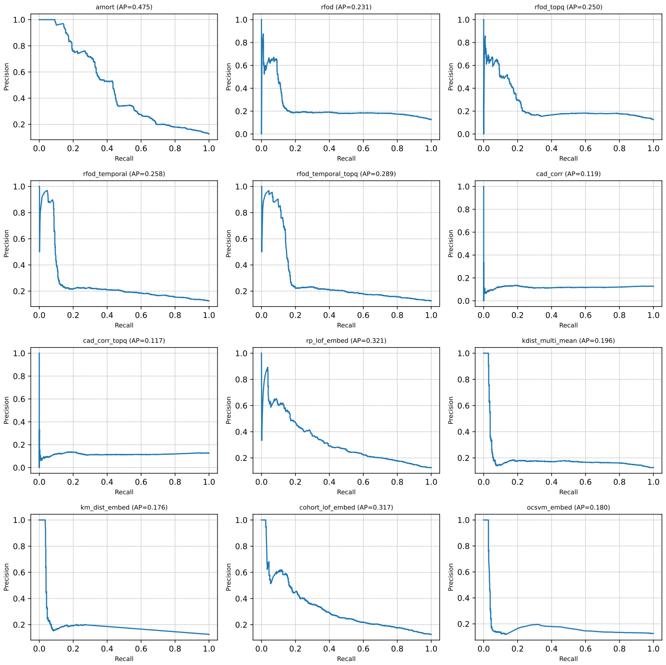
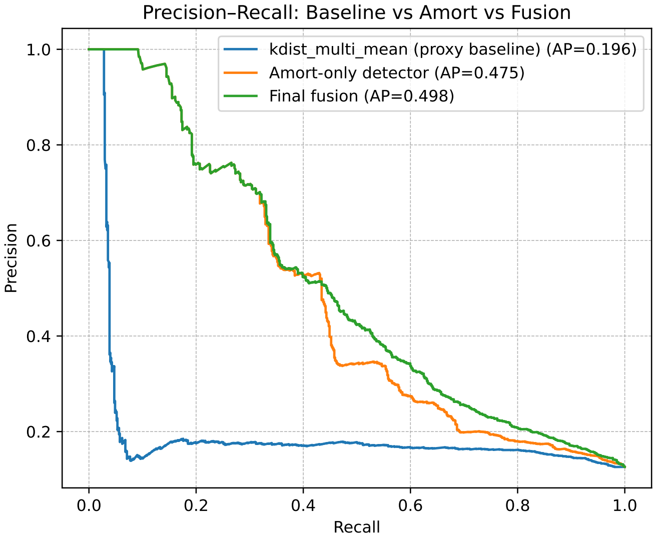
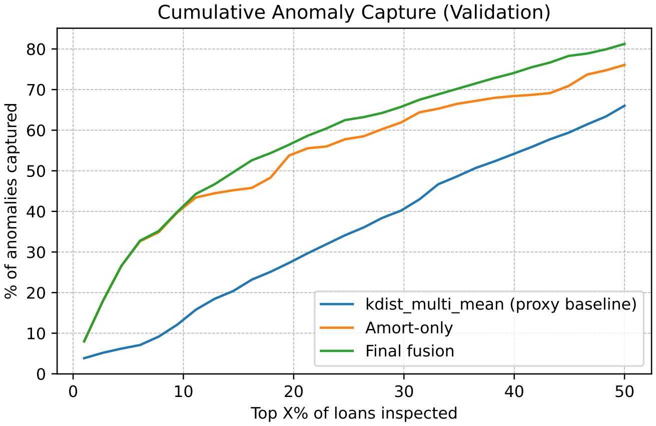
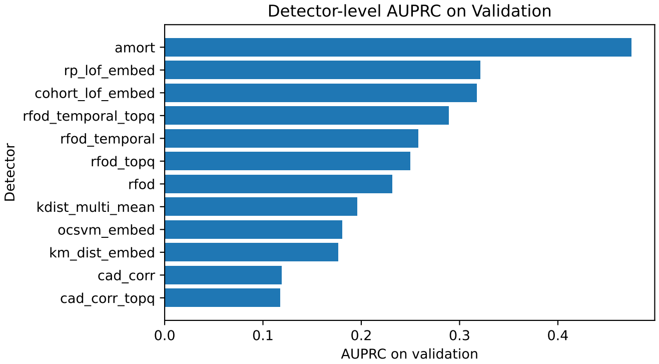
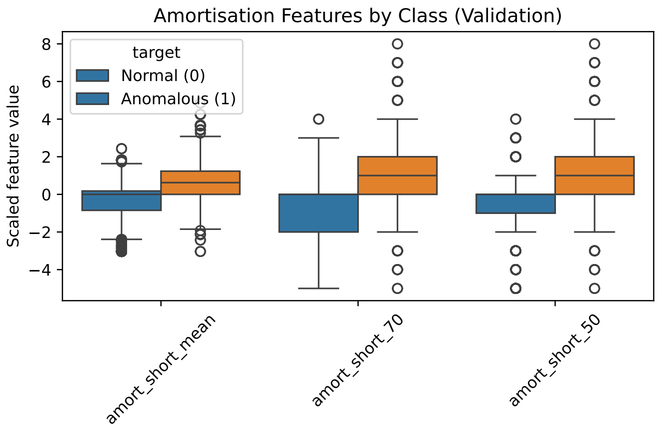


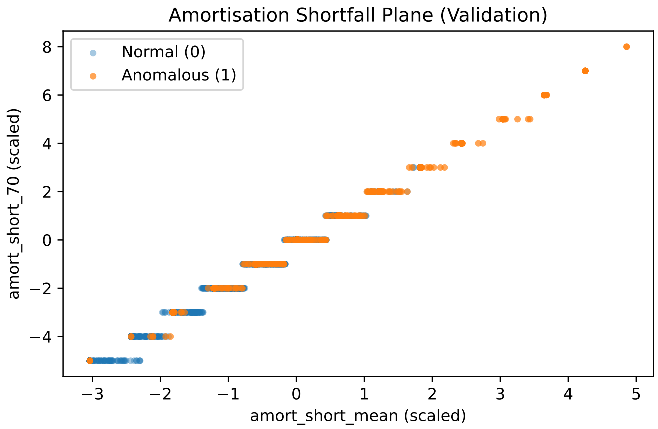
Fig. 12. Grid of precision–recall curves for selected detectors. While no single detector dominates everywhere, the curves highlight complementary strengths across different recall regions.

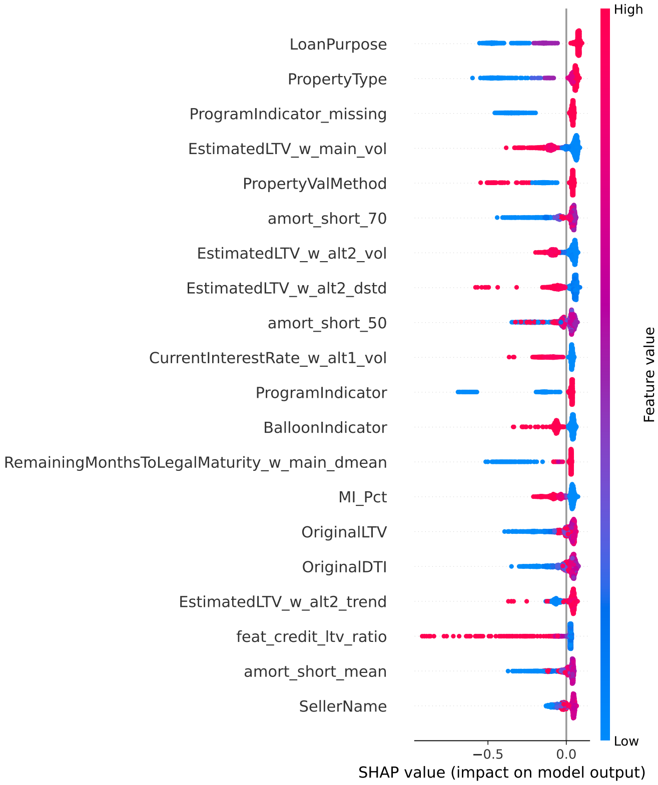


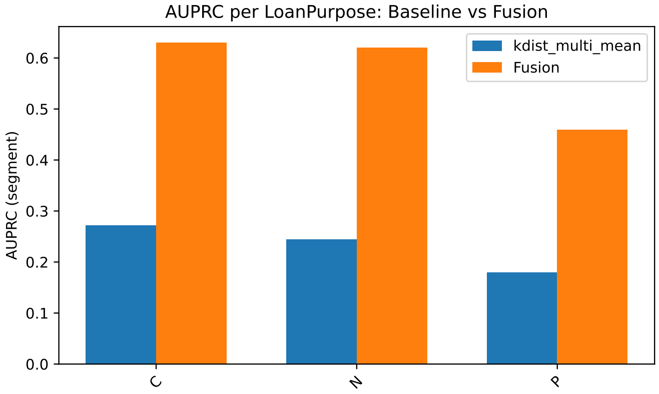





Amortisation Features by Class (Validation)
Scaled feature value
8
6
4
2
0
−2
−4
target
Normal (0)
Anomalous (1)
amort_short_mean
amort_short_70
amort_short_50


Amortisation Shortfall Plane (Validation)
Normal (0)
Anomalous (1)
amort_short_70 (scaled)
8
6
4
2
0
−2
−4
−3
−2
−1
0
1
2
3
4
5
amort_short_mean (scaled)


LoanPurpose
PropertyType
ProgramIndicator_missing
EstimatedLTV_w_main_vol
PropertyValMethod
amort_short_70
EstimatedLTV_w_alt2_vol
EstimatedLTV_w_alt2_dstd
amort_short_50
CurrentInterestRate_w_alt1_vol
ProgramIndicator
BalloonIndicator
RemainingMonthsToLegalMaturity_w_main_dmean
MI_Pct
OriginalLTV
OriginalDTI
EstimatedLTV_w_alt2_trend
feat_credit_ltv_ratio
amort_short_mean
SellerName
−0.5
0.0
SHAP value (impact on model output)
High
Feature value
Low


AUPRC per LoanPurpose: Baseline vs Fusion
AUPRC (segment)
0.6
0.5
0.4
0.3
0.2
0.1
0.0
kdist_multi_mean
Fusion
C
N
P