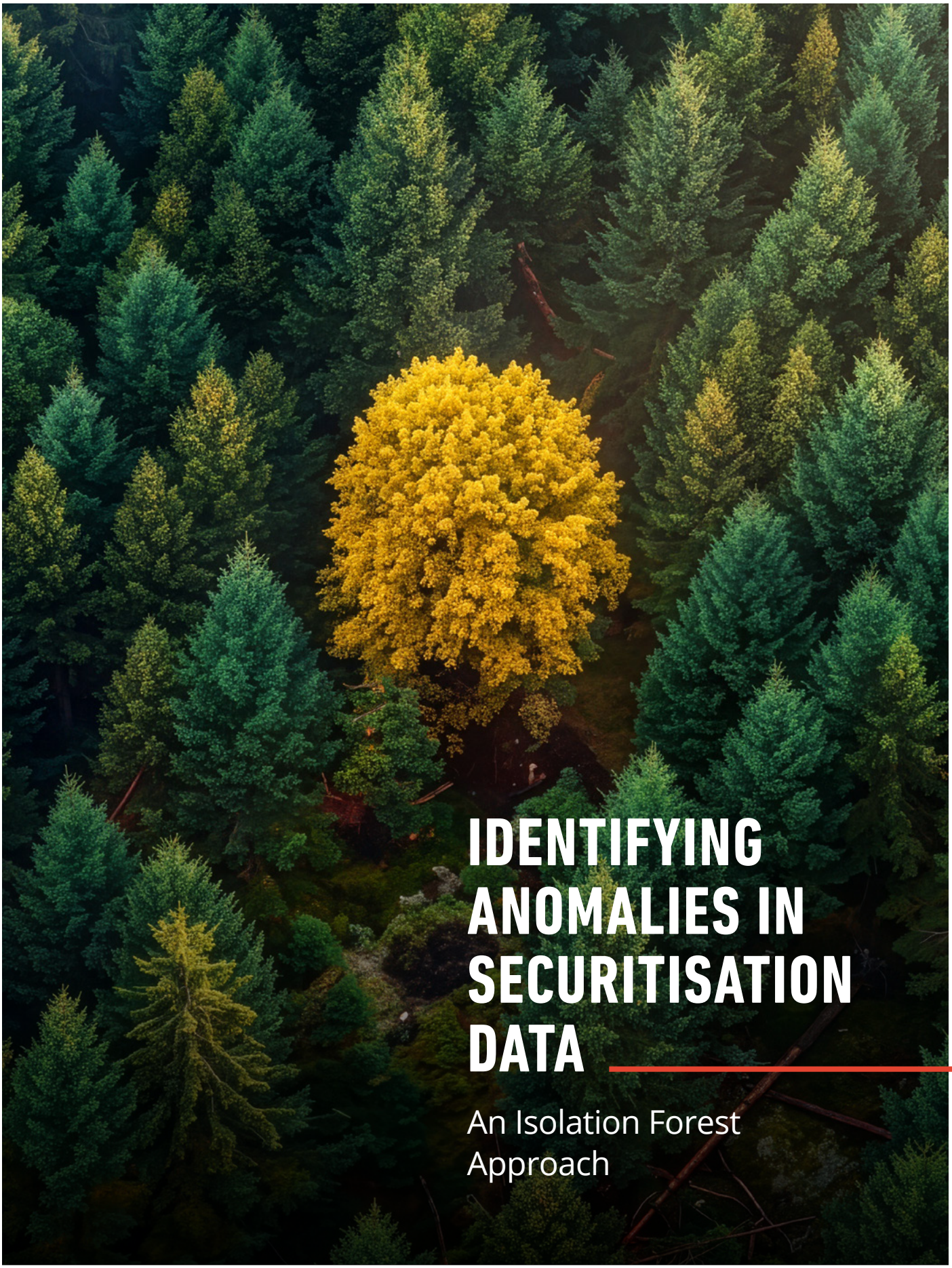
EUROPEAN DATAWAREHOUSE

EUROPE'S FIRST SECURITISATION REPOSITORY



IDENTIFYING ANOMALIES IN SECURITISATION DATA

An Isolation Forest Approach

WHAT'S IT ABOUT?

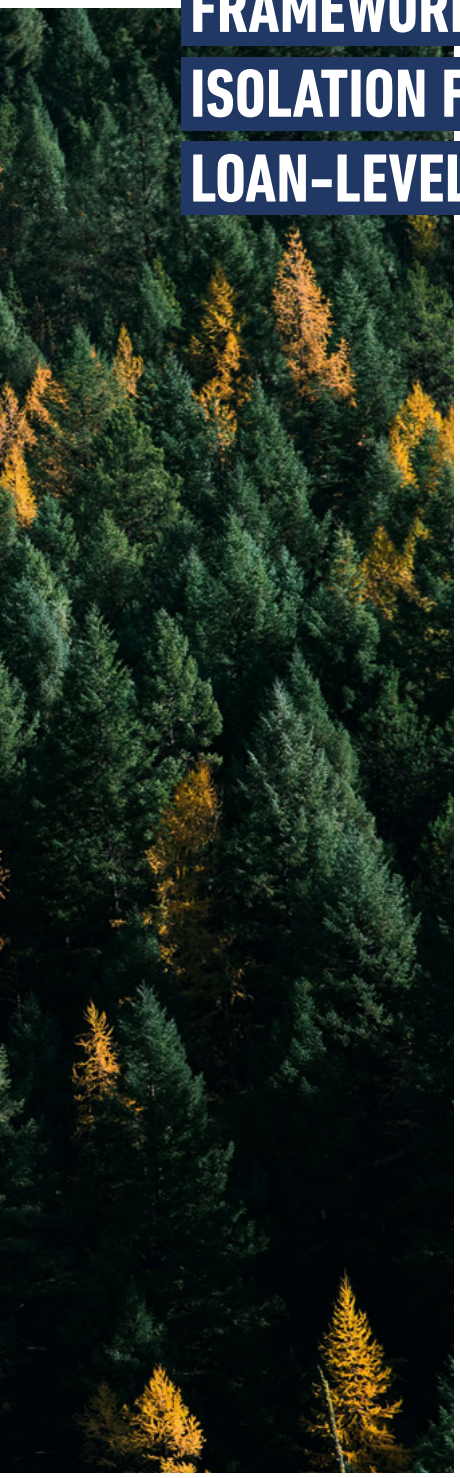
Data integrity issues within capital markets can long remain undetected before surfacing through inconsistencies in the dataset.

European DataWarehouse (EDW) thus runs a number of data quality checks on the data uploaded to its database, to check for possible errors and inconsistencies.

This paper explores the possibility of using an Isolation Forest based technique to complete our set of verifications.



ANOMALY DETECTION FRAMEWORK APPLICATION USING ISOLATION FOREST ON UK RMBS LOAN-LEVEL DATA.



EDW is uniquely positioned to support such efforts given its extensive repository of standardised loan-level data across a large universe of public European securitisation transactions.

While this data has traditionally been used for transparency, reporting and portfolio analysis purposes, its breadth and consistency also create an opportunity to apply advanced analytical techniques aimed at identifying observations that appear unusual relative to the wider market. By leveraging data already available within the EDW ecosystem, additional risk monitoring capabilities can be implemented without additional reporting requirements or data collection processes.

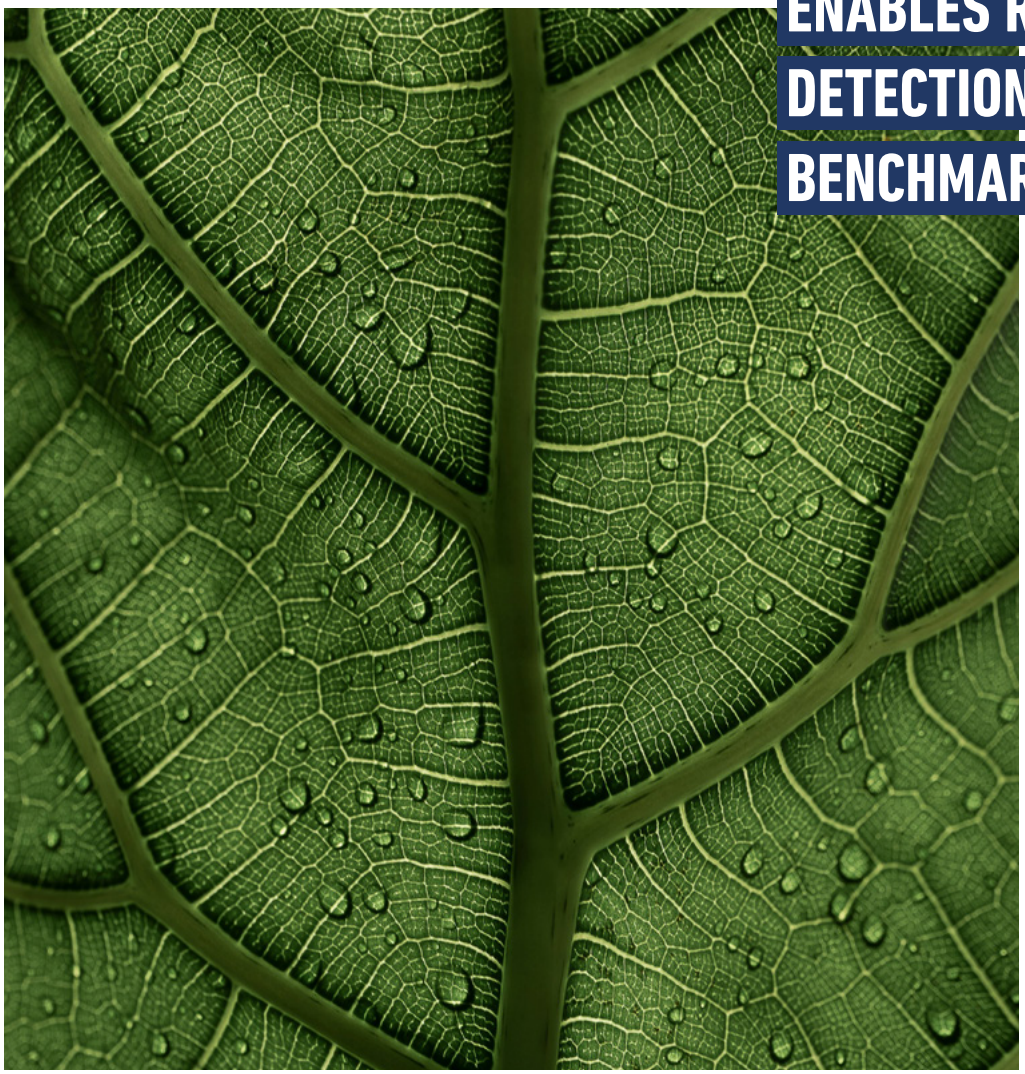
In this paper the Research team applies an unsupervised anomaly detection framework using Isolation Forest on UK RMBS loan-level data. The analysis seeks to determine whether statistical outliers can be identified within the broader market population and whether these results can provide a useful starting point for targeted reviews, control procedures and future enhancement of error detection and data quality monitoring activities.

WHY TRY ISOLATION FOREST FOR ERROR DETECTION?

A key strength of the Isolation Forest approach sits in the quality and consistency of the underlying EDW data it is to be applied to. EDW Data is historically consistent, standardised and validated across a large population of securitisations and provides a robust foundation for anomaly detection techniques which rely heavily on comparability across observations.

The data reported to EDW is embedded in standardised reporting templates across asset classes with each upload being validated individually; which significantly enhances the reliability of both the anomaly identification process and the resulting market-wide benchmarking analysis.

HIGH-QUALITY DATA ENABLES RELIABLE ANOMALY DETECTION AND MARKET-WIDE BENCHMARKING



METHODOLOGY

To explore the potential for enhanced checks in EDW with Isolation Forests, a set of anomaly detection rules and an algorithm were developed. Anomaly detection rule sets are designed to identify observations that differ materially from the expected behaviour of the population, providing scope for further investigations.

> Isolation Forest is an unsupervised algorithm that identifies unusual observations by progressively isolating records through a series of random partitions.

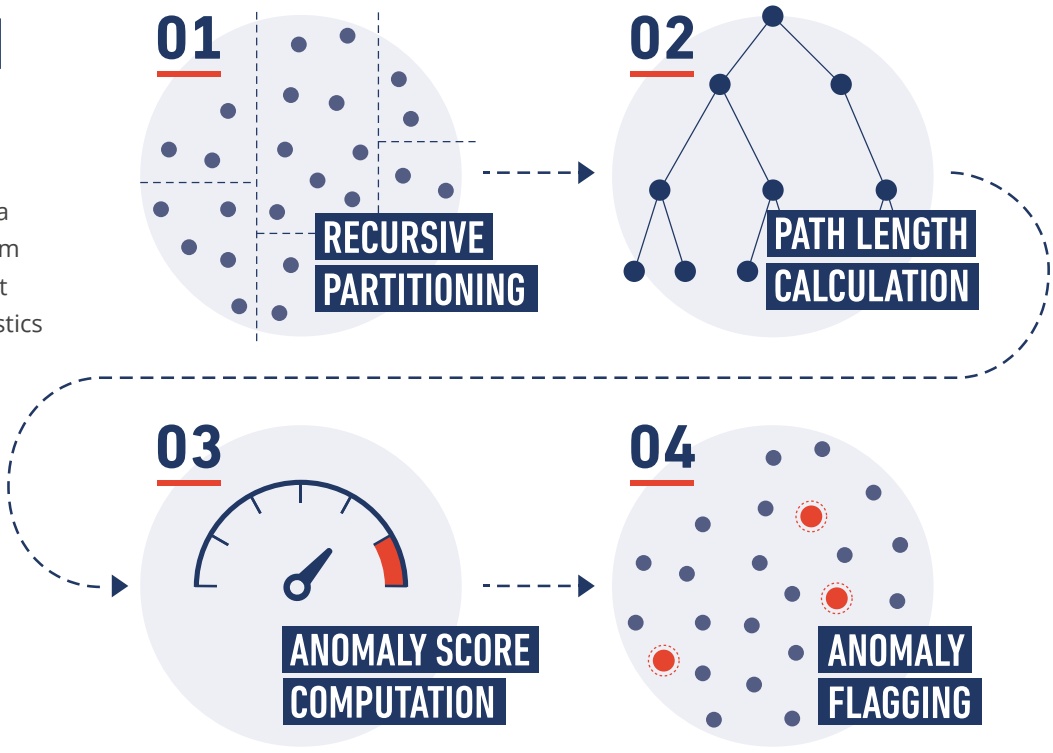
Records that require fewer partitions to isolate are considered more anomalous, while those embedded within dense clusters of similar observations are considered normal.

The primary output is an anomaly score between -1 and 1 assigned to each observation based on a prediction function, with a score of -1 corresponding to higher anomaly levels and a 1 to something very expected. Observations are then ranked to enable in-depth analysis of the most significant outliers. All anomalous observations from a selected transaction are then plotted against the entire UK RMBS active portfolio.

The underlying assumption is that anomalous records are both rare and distinct and can be isolated faster than typical observations. Unlike traditional supervised approaches, Isolation Forest does not require pre-labelled instances, making it suitable for large and diverse data sets where there might be unusual patterns. Isolation Forest scores observations are based on how easily they can be isolated through random data partitions.

HOW ISOLATION FOREST WORKS

Rather than determining whether a record is very strange, the algorithm highlights observations that exhibit atypical combination of characteristics in comparison to peers which may warrant further review by subject matter experts.



DATA PREPARATION

The research team have considered and tested the majority of the numerical format fields, as the algorithm required, from the EDW database.

For the asset class residential real estate, data fields used include but are not limited to RREL16 (Primary Income); RREL29 (Original Principal Balance); RREL30 (Current Principal Balance); RREL43 (Current Interest Rate); RREC12 (Current Loan-to-Value); RREC13 (Current Valuation Amount); RREC16 (Original Loan-to-Value), corresponding fields have been used for the rest of the asset classes.

The distribution of each variable has been tested and deemed to sit within the normal distribution. Missing values or NDs have been removed from this analysis. For this exercise, only the eligible UK transactions from the RMB asset class have been considered.

FOUR STEPS TOWARDS DATA QUALITY

01



NUMERICAL FIELDS

Majority of numerical format fields tested

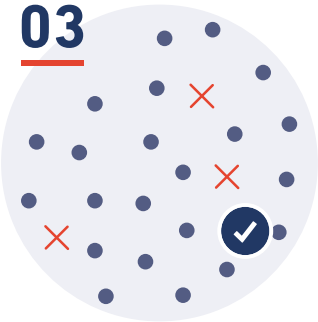
02



DATA DISTRIBUTION

Each variable tested for distribution

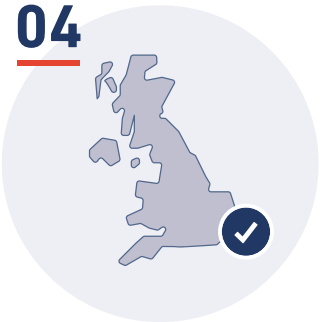
03



DATA CLEANING

Missing values
NDs removed

04



SCOPE

Eligible UK RMB transactions only

RESULTS

To support better feature selection, a Jaccard overlap matrix was constructed across multiple model iterations using different combinations of candidate variables. The objective was to identify variable sets that produces the highest degree of discrimination from a model iteration to the other. By measuring the degree of overlap between the tested populations, a set of stable discriminatory features was constructed that consistently contributed to the identification of unusual characteristics.

Securitisation ID	Loan ID	Primary Income	Original Principal Balance	Interest Rate	Original Valuation Amount	Anomaly Score
SEC_1	LOAN_1	€1.73m	€3.82m	1.17%	€8.91m	-0.1043
SEC_1	LOAN_2	€2.94m	€2.16m	2.43%	€5.37m	-0.1036
SEC_2	LOAN_3	€0.88m	€4.71m	0.64%	€6.28m	-0.1030
SEC_1	LOAN_4	€3.67m	€1.42m	3.11%	€12.54m	-0.1009
SEC_1	LOAN_5	€1.21m	€5.83m	1.76%	€3.49m	-0.1007
SEC_1	LOAN_6	€2.48m	€3.07m	2.88%	€9.73m	-0.1000
SEC_1	LOAN_7	€1.95m	€0.97m	1.32%	€4.16m	-0.0998
SEC_1	LOAN_8	€0.76m	€2.64m	2.57%	€7.85m	-0.0998
SEC_1	LOAN_9	€3.12m	€4.29m	1.94%	€2.71m	-0.0987
SEC_3	LOAN_10	€2.31m	€1.58m	3.47%	€10.22m	-0.0985

Table 1 Illustrative Extract of Highest-Ranked Anomalous Loans

As per the extract shown in Table 1 the highest ranked anomalies are characterised by combinations of borrower income, loan balance and collateral valuation that are uncommon within the broader mortgage population.

> While each individual characteristic may be plausible on its own, the algorithm identifies these loans because the overall combination of attributes occurs infrequently relative to the wider dataset.

For instance, a case in which the loan amount is very small compared with the income.

On one hand, results suggest that the algorithm primarily detects observations located at the extreme upper end of the portfolio distribution where borrower characteristics in terms of affordability metrics, property valuations and exposure sizes differ materially from the majority. The negative and low anomaly score indicates that these observations are amongst the most isolated ones from the multidimensional space analysed by the algorithm.

On the other hand, anomalous observations do not appear to be driven by one single variable such as interest rate, but multidimensional relationships between income, collateral value and loan size are considered showing that the anomaly detection is assessing the overall profile of each loan rather than applying simple threshold-based rules to individual fields.

While the anomaly table provides visibility over individual observations identified by the algorithm, the real analytical value lies in the visualisations which place anomalous observations in the context of the entire UK RMBS population.

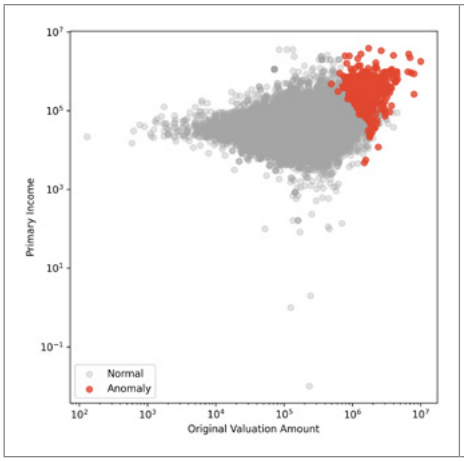


Figure 1: Primary Income vs Original Valuation Amount

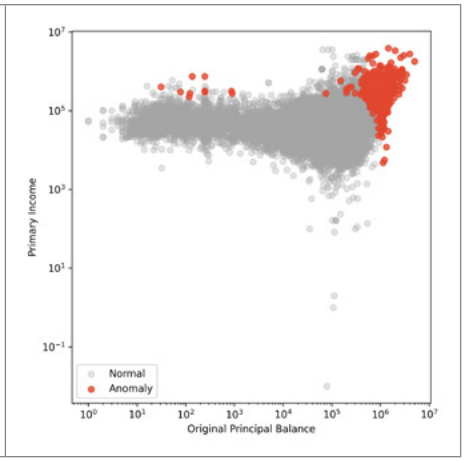


Figure 2: Primary Income vs Original Principal Balance

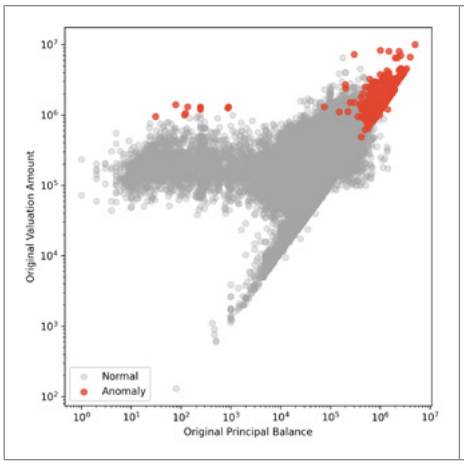


Figure 3: Original Valuation Amount vs Original Principal Balance

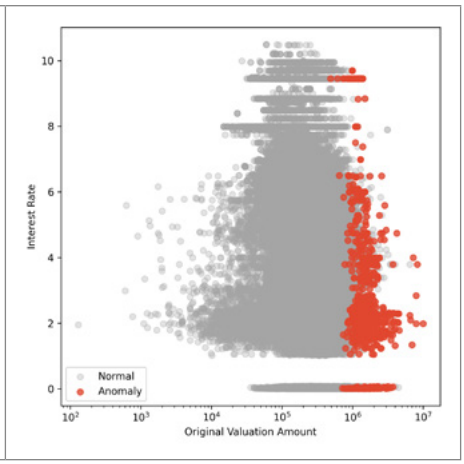


Figure 4: Interest Rate vs Original Valuation Amount

Overall, the visualisations demonstrate that the anomalous population occupies clearly defined regions of the market-wide distribution, suggesting systematic differences from the broader UK RMBS dataset.

A consistent pattern can be seen across all four charts; a majority cluster forms around central areas showing that observations are homogeneous to some degree.

Anomalous observations tend to appear in marginal regions of the feature space

and are particularly visible at the upper end of the distribution for Original Valuation Amounts and Primary Income (Figure 1 and Figure 2), Original Principal Balance and Original Valuation Amount (Figure 3). Rather than appearing randomly throughout the feature space, the anomalous loans form distinct clusters that separate from the main population, indicating borrower and collateral profiles that are uncommon relative to the wider market. Figure 4 further supports this observation, showing that anomalies are concentrated among loans with higher valuation amounts across a range of interest rates, while the bulk of the market remains densely grouped at lower valuation levels. Overall, the visualisations demonstrate that the anomalous population occupies clearly defined regions of the market-wide distribution, suggesting systematic differences from the broader UK RMBS dataset.

In the same graphs the EDW team then highlights the observations of one particular transaction in blue, providing additional context for data owners and data providers. The primary focus of these visualisations does not sit in the number of anomalies identified in the cloud but in the degree of overlap between the transaction population and the wider market.

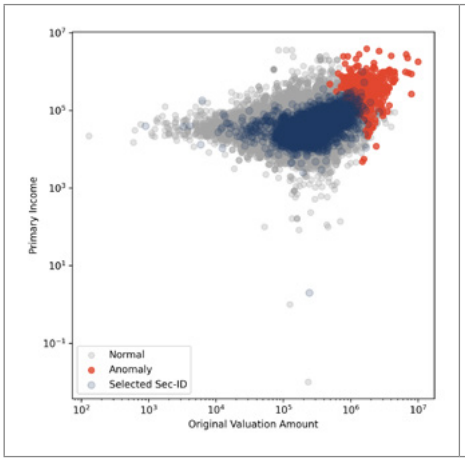


Figure 5: Primary Income vs Original Valuation Amount

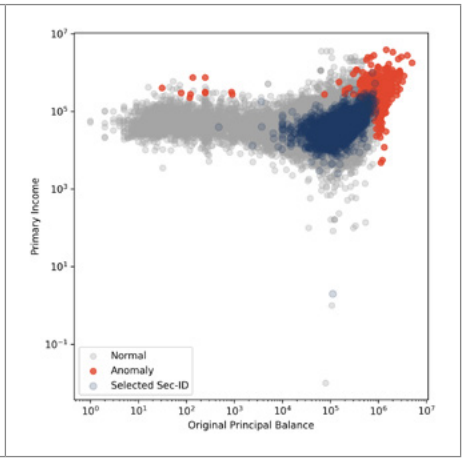


Figure 6: Primary Income vs Original Principal Balance

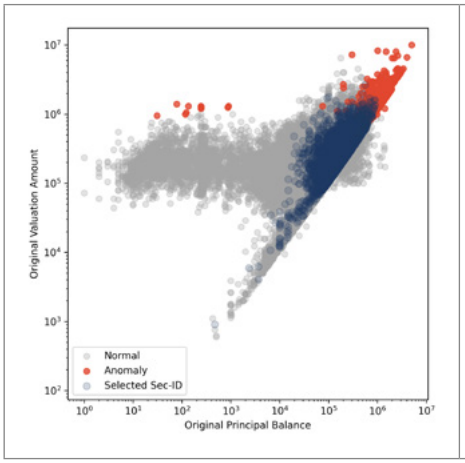


Figure 7: Original valuation Amount vs Original Principal Balance

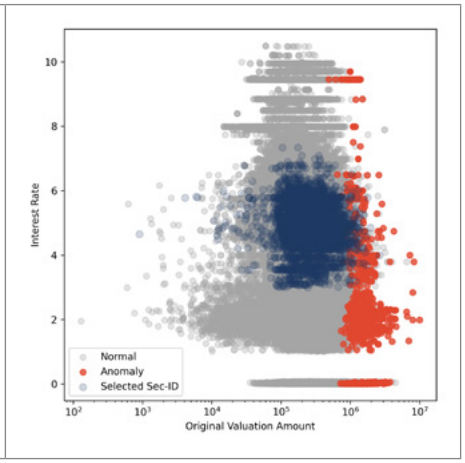


Figure 8: Interest Rate vs Original Valuation Amount

The charts ultimately serve as a benchmarking tool, enabling characteristics of a specific transaction to be assessed not only against the wider market but also against the population if loans identified as statistically unusual by the algorithm highlighting how closely the characteristics of loans align with those of the market.

Across the observed variable combinations, the selected transaction is concentrated within the upper range of the core market population and shows partial overlap with clusters of anomalous observations.

This is particularly evident in the relationships between Primary Income and Original Valuation Amount (Figure 5), Primary Income and Original Principal Balance (Figure 6), and Original Valuation Amount and Original Principal Balance (Figure 7), where the transaction extends towards regions associated with anomalous loans while remaining largely aligned with the broader market distribution. The relationship between Interest Rate and Original Valuation Amount (Figure 8) shows a similar pattern, with the transaction concentrated in a distinct but not isolated segment of the market.

The visualisations therefore provide an immediate indication of where the transaction sits relative to both typical and atypical market behaviour. A greater overlap between the transaction population and the anomaly clusters suggest stronger exposure to loan characteristics that are less frequently observed across the wider UK RMBS universe, whereas stronger concentration outside of the anomaly could indicate a closer alignment with broader market observations.

LIMITATIONS

Naturally, the results are influenced by several algorithm design choices, including the selection of the input parameters such as the contamination rate. While these parameters have been set according to practice, a degree of subjectivity remains in the model specification.

The identified observations provide a meaningful starting point for Agreed-Upon-Procedures review within the pool selection process of structured finance transaction. The anomalous classifications alone should not be interpreted as evidence of fraud, misreporting or credit weakness.

CONCLUSIONS & NEXT STEPS

01

PROVEN EFFECTIVENESS

The results demonstrate the effectiveness of applying the Isolation Forest algorithm on EDW's consistent and standardised loan-by-loan data. By leveraging a harmonised dataset across a large population of UK RMBS transactions, the approach identified statistically unusual loans and compare them to the market. The resulting anomaly clouds provide a clear and intuitive representation of market behaviour, allowing for in-depth transaction level assessments that would be difficult to achieve using traditional sampling or rule-based techniques alone.

02

FROM PILOT TO SCALE

This pilot implementation has successfully demonstrated the practical applicability of this methodology and can be fully automated within the EDW analytical environment. As a result, the framework is now ready to be extended across other asset classes enabling repeatable anomaly screening across a significantly larger transaction universe with minimal additional operational effort.

03

LOOKING AHEAD

Looking ahead, future enhancements will focus on expanding the use of advanced data analytics and machine learning techniques across asset classes. While approached as anomaly detection remain relatively new and experimental within the ABS market, they represent an opportunity to complement existing review procedures, improve risk identification and unlock additional value from the extensive loan-level information available on EDW. By continuously researching and refining methodologies, EDW is well positioned to take a leading role in the application of data-driven analytical techniques within the securitisation market.

ANOMALY DETECTION PROVIDES A NEW WAY TO ASSESS LOAN-LEVEL DATA AGAINST THE WIDER MARKET.

REFERENCES

1. Liu, F.T., Ting, K.M. and Zhou, Z.-H. (2008) 'Isolation Forest', Proceedings of the 2008 Eighth IEEE International Conference on Data Mining, pp. 413-422.

2. Scikit-learn Developers. (2025) IsolationForest. Available at: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.IsolationForest.html

3. European DataWarehouse GmbH. (2025) European DataWarehouse loan-level repository. Frankfurt am Main: European DataWarehouse. Available at: https://eurodw.eu/

STAY AHEAD OF THE MARKET.

Subscribe for regulatory updates, research insights and industry news from European DataWarehouse.

Copyright© 2026 by European DataWarehouse GmbH, Walther-von-Cronberg-Platz 2 60594 Frankfurt am Main. Telephone: +49 (0) 69 8088 4300. All rights reserved.

The information contained in this document (the "White Paper") has been produced by European DataWarehouse GmbH (EDW) in the context of its research activities. The information displayed in the White Paper results from an analysis of public securitisation data hosted at EDW. EDW assumes no responsibility and shall not be held liable for the results of the analysis or any conclusions drawn from it. It is published "as is" without any representation or warranty of any kind. EDW does not guarantee or warrant its accuracy or completeness. EDW does not warrant that it will meet the user's or any third party's individual requirements.

The reproduction and distribution of the information contained in the White Paper is prohibited without EDW's prior written consent. All reproductions and copies of the information shall include appropriate acknowledgement of EDW as the source of the data.

Certain information contained herein may include assumptions, opinions, analysis, estimates and views of EDW. The assumptions, opinions, analysis, estimates and views contained herein are solely opinions which are uncertain and subject to risks. The information contained herein is not intended to constitute investment, legal, regulatory, nor any other sort of advice. EDW assumes no obligation to update the content in any form or format. The White Paper includes a table with the highest-ranked anomalous loans. Such extract is illustrative and does not depict real loan-level data from any securitisation transactions hosted at EDW.

In no event shall EDW be liable for any inaccuracy, error or omission regardless of cause, or for any damages resulting therefrom. Under no circumstances shall EDW have any liability to any party for any direct, indirect, incidental, exemplary, compensatory, punitive, special or consequential damages, costs, expenses, legal fees, or losses in connection with any use of the information contained in the White Paper.

EUROPEAN UNION

European DataWarehouse GmbH
Walther-von-Cronberg-Platz 2
60594 Frankfurt am Main

www.eurodw.eu
enquiries@eurodw.eu
+49 (0) 69 50986 9017

UNITED KINGDOM

European DataWarehouse Ltd.
One Canada Square, 37th Floor
London E14 5AA

www.eurodw.co.uk
enquiries@eurodw.co.uk
+44 (0) 20 3997 6861

092026