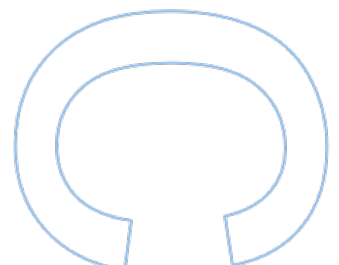
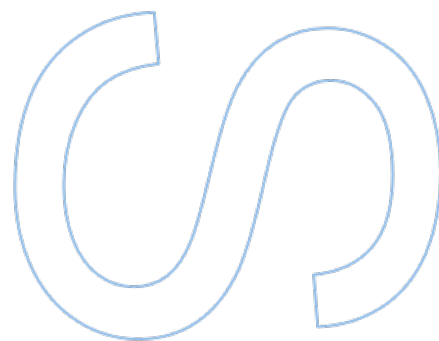
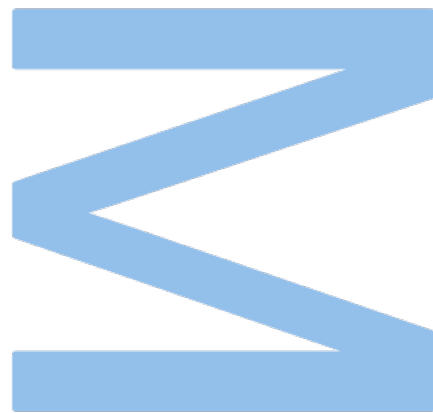


Statistical and Machine Learning Methods in Process Mining



Manuel Galhanas Germano da Silva
Condeço

Master in Computational Statistics and Data Analysis
Department of Mathematics
Faculty of Sciences of the University of Porto
2026

Statistical and Machine Learning Methods in Process Mining

Manuel Galhanas Germano da Silva
Condeço

Internship Report carried out as part of the Master in
Computational Statistics and Data Analysis
Department of Mathematics
2026

Supervisor

João Pedro Pedroso Ramos dos Santos, Associate Professor,
Faculty of Sciences of the University of Porto

External Host Supervisor

Henrique Daniel de Oliveira Duarte, Head of R&D and Data
Science, Breakawai



Acknowledgements

I would like to begin by expressing my sincere gratitude to Breakawai for providing me with the opportunity to undertake this internship. I am especially thankful for the warm welcome I received and for making me feel a valued member of the team from the very beginning.

To my host supervisor, Henrique Duarte, I am deeply grateful for his close and attentive guidance throughout this work, for the knowledge he so generously shared, and for the companionship that accompanied it from start to finish.

To my academic supervisor, Professor João Pedro Pedroso, I extend my sincere thanks for his constant availability, guidance, and knowledge, which made this process both easier and far more enriching.

To my parents, for their unconditional support, as always, and for being a fundamental pillar in my life.

To Beatriz, for always believing in me and for making every step of this journey lighter.

Resumo

Este relatório documenta três atividades desenvolvidas durante um estágio curricular de sete meses na área de process mining, disciplina que extrai conhecimento sobre a execução real de processos a partir de registos de eventos gerados por sistemas de informação. Os objetivos foram segmentar threads de correio eletrónico em perfis comportamentais interpretáveis, desenvolver uma ferramenta interativa de análise de causa-raiz para bases de dados de process mining, e estudar a consolidação de paletes num cliente do setor logístico. À primeira atividade foram aplicadas análise de componentes principais e métodos de agrupamento; na segunda foi desenvolvida uma aplicação multi-página em *Streamlit*, generalizando o desvio médio ponderado ao caso multivariável e integrando testes não paramétricos e uma árvore de decisão; a terceira combinou uma análise descritiva com um Modelo Linear Generalizado Misto e modelos de *ensemble* baseados em árvores interpretados via valores SHAP. Foram identificados quatro perfis comportamentais estáveis, com Índice de Rand Ajustado de 0,749 entre *k*-means e solução hierárquica. A análise de consolidação estimou redução anual de 7,8% nas paletes expedidas, menos 334 camiões e cerca de 188 toneladas de CO₂ equivalente. O GLMM, Random Forest e XGBoost atingiram AUCs de 0,793, 0,867 e 0,883, identificando a largura da paleta, a última entrega de uma ordem e a rota de envio entre os fatores mais determinantes. O trabalho ilustra como adequar os métodos à estrutura de cada problema apoia a tomada de decisão operacional.

Palavras-chave: process mining; análise de componentes principais; agrupamento; análise de causa-raiz; modelo linear generalizado misto; modelos de ensemble de árvores; otimização logística.

Abstract

This report documents three activities developed during a seven-month curricular internship in the process mining area, the discipline that extracts knowledge about how processes are actually executed from the event logs recorded by information systems. The objectives were, respectively, to segment e-mail communication threads into interpretable behavioural profiles, to develop an interactive root cause analysis tool for process-mining databases, and to study the consolidation and utilization of pallets at a logistics client, from the perspectives of optimisation potential and operational drivers. Principal component analysis and clustering were applied to the first problem; in the second activity was implemented a multi-page Streamlit application generalising the weighted mean deviation to the multivariate case, integrating non-parametric tests and a decision tree; the third combined a descriptive consolidation analysis with a Generalized Linear Mixed Model and ensemble tree-based models interpreted via SHAP values. Four stable behavioural profiles were identified, with an Adjusted Rand Index of 0.749 between the k -means and the hierarchical solutions. The pallet-consolidation analysis estimated an annual reduction of 7.8 per cent in pallets shipped, 334 fewer trucks and approximately 188 tonnes of CO₂ equivalent. The GLMM, Random Forest and XGBoost attained test AUCs of 0.793, 0.867 and 0.883 respectively, and identified pallet width, last delivery of an order and shipping route among the strongest drivers. The work illustrates how matching statistical and machine learning methods to the structure of each problem can support operational decision-making.

Keywords: process mining; principal component analysis; clustering; root cause analysis; generalized linear mixed model; ensemble tree-based models; logistics optimisation.

Table of Contents

List of Tables	vi
List of Figures	vii
List of Abbreviations	x
1. Introduction.....	1
1.1. Process Mining: a brief theoretical framing.....	1
1.2. Objectives and structure of the report.....	3
2. Behavioural Segmentation of E-mail Threads through Principal Component Analysis and Clustering	4
2.1. Problem, motivation and objective	4
2.2. Data and variables	4
2.3. Exploratory analysis and variable transformations	6
2.4. Principal Component Analysis	8
2.5. Clustering	10
2.6. Results and discussion	12
3. An Interactive Tool for Root Cause Analysis in Process Mining	15
3.1. Problem, motivation and objective	15
3.2. Implementation choices: the Streamlit framework	16
3.3. Description of the tool	16
3.4. Data loading and case-table construction.....	17
3.5. Target specification.....	18
3.6. The Weighted Mean Deviation and its multivariate generalisation	20
3.7. Detection of redundancy among variables.....	22
3.8. Non-parametric hypothesis tests	23
3.9. The interpretable decision tree	24
4. A Logistics Analysis of Pallet Consolidation and Utilization for an External Client	26
4.1. Phase 1: Quantifying the Logistics Optimization Potential of Pallet Consolidation	27

4.1.1.	Problem, motivation and objective	27
4.1.2.	Methodological decisions	28
4.1.3.	Estimation of the avoided CO2 emissions	29
4.1.4.	Results and discussion	30
4.2.	Phase 2: Identifying the Operational Drivers of Pallet Utilization	32
4.2.1.	Problem, motivation and objective	32
4.2.2.	Feature engineering and aggregation	33
4.2.3.	Exploratory analysis and preparation for modelling	34
4.2.4.	The Generalized Linear Mixed Model	36
4.2.5.	Assumption checking and diagnostics	37
4.2.6.	Odds, odds ratios and their interpretation	40
4.2.7.	Model interpretation	41
4.2.8.	Model validation	44
4.2.9.	Ensemble tree-based models	45
4.2.10.	Comparison of the modelling paradigms.....	48
5.	Conclusion and Future Work	50
5.1.	Conclusion	50
5.2.	Future Work	51
6.	References	53
7.	Annex	55
	Annex A - Supplementary figures and tables for Chapter 2.....	55
	Annex B - Supplementary figures for Chapter 4, Phase 1	59
	Annex C - Supplementary figures for Chapter 4, Phase 2.....	60

List of Tables

Table 1 - Thread-level behavioural indicators used in the segmentation, with a short description of what each one captures.	5
Table 2 - Interpretation of the first three principal components and the variables that dominate each one.	10
Table 3 - The four behavioural profiles identified by the clustering, with their location on the principal components.	13
Table 4 - Estimated optimisation potential from pallet consolidation, by route and in total, for the year 2024.	30
Table 5 - Variables defined at or aggregated to the pallet level.	33
Table 6 - Complete fixed-effect odds ratios of the GLMM, with 95% confidence intervals and p-values, relative to the respective reference categories.	41
Table 7 - Predictive performance of the three models on the held-out test set.	46
Table 8 - Principal-component loadings: contribution of each variable to each principal component.	58

List of Figures

Fig. 1 - Spearman rank correlation matrix of the thread-level indicators, reported on their original scale.....	8
Fig. 2 - Scree plot of the cumulative explained variance as a function of the number of principal components retained.....	9
Fig. 3 - Elbow curve of the within-cluster inertia (left) and silhouette coefficient (right) of the k-means solution as a function of the number of clusters. The silhouette attains a local maximum at four clusters.	11
Fig. 4 - Projection of the threads onto the first two principal components, coloured by the four-cluster k-means solution.	12
Fig. 5 - Spatial agreement between the k-means and the hierarchical solutions on the first two principal components. Disagreements concentrate at the boundaries between clusters.	14
Fig. 6 - Page flow of the multi-page Root Cause Analysis tool developed in Streamlit. Later pages remain locked until the earlier ones are completed.	17
Fig. 7 - Data loading and column-mapping page. The event log is uploaded as a CSV file and each column is assigned to a normalised role.....	17
Fig. 8 - Processing-preview page, showing the aggregated case table and the variable-type assignment, which the user can review before the analysis proceeds.	18
Fig. 9 - Target-specification page. A binary deviance label is defined from a chosen label-generating variable, a rule and an SLA-anchored threshold, and the resulting class balance is reported.	19
Fig. 10 - Single-feature Weighted Mean Deviation page. For each value of the selected feature, the subgroup mean, count, weight and signed influence are reported.....	21
Fig. 11 - Multi-condition Weighted Mean Deviation page. The influence is computed on the subset of cases that simultaneously satisfy all the specified conditions.....	22
Fig. 12 - Decision-tree page. The hyperparameters that govern the depth and the regularisation of the tree are exposed to the user, alongside the seeds that control reproducibility.	25
Fig. 13 - Number of pallets shipped per period on a representative route, distinguishing the small non-mixed pallets carrying one to three items (lower band) from the remaining pallets (upper band).....	29
Fig. 14 - Mean number of non-mixed, mixed and total pallets by day-group for the two warehouses, showing that the incidence of small non-mixed pallets varies across the week.	29

Fig. 15 - Number of mixed pallets per period before and after the proposed consolidation, for one warehouse. The dashed line marks the operational ceiling on new mixed pallets per period.	31
Fig. 16 - Distribution of the binary response variable in the training set, showing a moderate imbalance of approximately 73% good pallets to 27% bad pallets.	35
Fig. 17 - Bias-corrected Cramer's V matrix of the categorical predictors retained in the final model.	36
Fig. 18 - Normal quantile-quantile plot and histogram of the random-effect predictions-Best Linear Unbiased Predictors (BLUPs) of the GLMM, used for the visual assessment of normality.	39
Fig. 19 - Simulation-based residual diagnostics of the GLMM. The left panel shows a quantile-quantile plot of the simulated scaled residuals against the uniform distribution; the right panel shows the residuals against the model predictions.	40
Fig. 20 - Receiver operating characteristic curves of the Random Forest and the XGBoost models on the held-out test set.	46
Fig. 21 - SHAP summary plot for the XGBoost model on the held-out test set. Each point is an observation, the colour encodes the value of the predictor and the horizontal position its contribution to the predicted log-odds of a good pallet.	47
Fig. 22 - Principal-component biplot of the threads on the first two components, with the loading vectors of the original variables.	55
Fig. 23 - Dendrogram of the agglomerative hierarchical clustering with Ward linkage.	56
Fig. 24 - Three-dimensional projection of the threads onto the first three principal components, coloured by the four-cluster k-means solution.	57
Fig. 25 - Distribution of the mean number of small non-mixed pallets per period for one warehouse.	59
Fig. 26 - Proportion of good and bad pallets within the categories of selected categorical predictors, in the training set.	60
Fig. 27 - Spearman rank correlation matrix of the numerical predictors.	61
Fig. 28 - Empirical logit of the response against the three retained continuous predictors under their original, logarithmic and square-root scales, used for the visual assessment of the linearity of the logit.	62
Fig. 29 - Selected marginal effects from the GLMM, expressed as the predicted probability of a good pallet with 95% confidence intervals, translating the odds ratios onto the probability scale.	63
Fig. 30 - Receiver operating characteristic curve of the GLMM on the held-out test set.	64

Fig. 31 - Aggregated SHAP feature-importance rankings for the Random Forest and the XGBoost models.....	64
Fig. 32 - SHAP summary plot for the Random Forest model on the held-out test set..	65

List of Abbreviations

ARI	ADJUSTED RAND INDEX
AUC-ROC	AREA UNDER THE RECEIVER OPERATING CHARACTERISTIC CURVE
BLUPs	BEST LINEAR UNBIASED PREDICTORS
EPV	EVENTS PER VARIABLE
ETL	EXTRACTION, TRANSFORMATION AND LOADING
FCUP	FACULTY OF SCIENCES OF THE UNIVERSITY OF PORTO
GLEC	GLOBAL LOGISTICS EMISSIONS COUNCIL
GLMM	GENERALIZED LINEAR MIXED MODEL
GVIF	GENERALIZED VARIANCE INFLATION FACTOR
ICC	INTRACCLASS CORRELATION COEFFICIENT
OR	ODDS RATIO
PCA	PRINCIPAL COMPONENT ANALYSIS
RCA	ROOT CAUSE ANALYSIS
SHAP	SHAPLEY ADDITIVE EXPLANATIONS
SLA	SERVICE LEVEL AGREEMENT
WMD	WEIGHTED MEAN DEVIATION

1. Introduction

This document was prepared in the context of the curricular internship unit of the final year of the MSc in Computational Statistics and Data Analysis of the Faculty of Sciences of the University of Porto (FCUP), within the Department of Mathematics. Its purpose is to describe and reflect on the activities developed during the curricular internship, which was hosted by the company Breakawai. The internship took place at the company premises, in Porto, over a period of seven months, between the 6th of October 2025 and the 30th of April 2026. Over this period three distinct activities were carried out, each addressing a different problem and each requiring a different combination of statistical and machine learning techniques.

Breakawai is a technology and consultancy company specialising in data science, process mining, task mining and process intelligence solutions oriented towards the digital transformation of organisations. Founded in Denmark, the company operates across several sectors, including health, banking, insurance, retail, industry and aviation, supporting organisations in process optimisation and in the extraction of knowledge from complex data. The mission of the company is to help organisations improve operational efficiency and decision-making through advanced data analysis, artificial intelligence and continuous process improvement, combining technical competence with domain knowledge in a multidisciplinary approach. During the internship the company provided a collaborative environment oriented towards innovation, allowing the practical application of knowledge acquired throughout the academic path, particularly in the areas of data analysis, statistical modelling and the development of analytical solutions for decision support.

1.1. Process Mining: a brief theoretical framing

Process mining is a discipline positioned at the intersection of data science and business process management, and it aims to extract knowledge about how processes are actually executed from the data recorded by information systems [1]. Rather than relying on how a process is assumed or designed to run, process mining starts from the digital traces that the execution of the process leaves behind, and uses them to reconstruct, measure and reason about the real behaviour of the process. This evidence-based perspective is particularly valuable in complex organisations, where the gap between the idealised process and its actual execution can be substantial.

The starting point of any process-mining initiative is the event log, a collection of records in which each event is associated, at a minimum, with a case identifier, an activity and a timestamp. The case identifier denotes the entity whose process is being followed, for instance an order, a patient or, as in the third activity of this internship, a pallet. The activity denotes the step that was performed, and the timestamp denotes the moment at which it occurred. Additional attributes, both at the level of the individual event and at the level of the case, are frequently available and enrich the analysis [1], [2]. By grouping the events of a log by their case identifier and ordering them by their timestamp, one obtains, for each case, a trace, that is, the sequence of activities through which that case passed; the collection of all traces constitutes the behaviour captured by the log.

Three broad types of process-mining task are commonly distinguished [1], [2]. Process discovery aims to construct a process model from an event log without using any a priori information, producing for instance a graph of activities and their observed orderings. Conformance checking compares an existing model with an event log of the same process in order to detect, locate and explain deviations between the prescribed and the observed behaviour. Process enhancement seeks to extend or improve an existing model using information about the actual process recorded in the log, for instance by enriching it with performance information, such as bottlenecks and waiting times, or with root-cause information that explains why certain cases behave differently from others. The activities developed during the internship draw on data of this nature, namely event logs and the case-level tables derived from them, and they are oriented towards the analytical, diagnostic and predictive dimension of process mining rather than towards the automatic discovery of process models.

A recurring step in process mining, and one that is central to two of the three activities described in this report, is the construction of a case table from an event log. Whereas the event log has one row per event, the case table has one row per case and summarises, for each case, a set of features derived from its trace and its attributes, such as the total duration, the number of events, the number of distinct activities, and any case-level attributes of interest. This aggregation transforms a sequential, multi-row representation into a tabular, one-row-per-case representation that is well-suited to classical statistical and machine learning methods. The choice of the unit of observation, that is, of what constitutes a case, is therefore a decisive modelling decision, especially when, as in the logistics analysis of the third activity, the data simultaneously embed several plausible notions of case.

1.2. Objectives and structure of the report

The three activities described in this report are not directly related to one another and address distinct problems and data sets. They are nevertheless unified by their belonging to the field of process mining and by the type of data on which they operate, namely event logs and the case-level tables derived from them. The first activity was an exploratory, proof-of-concept analysis that segmented e-mail communication threads, represented as a process log, into behavioural profiles by means of dimensionality reduction and clustering. The second activity produced a general-purpose interactive tool to support Root Cause Analysis on process-mining databases. The third and most extensive activity was an analysis carried out for an external client, in which a logistics process was studied first from the perspective of optimisation potential and subsequently from the perspective of identifying the operational drivers of process quality.

The objective of this report is to document the three activities, describing for each one the underlying problem, the motivation and the objective, the methodological decisions taken, and, where applicable, the analysis and interpretation of the results. Particular emphasis is placed on the description and justification of the main statistical and machine learning techniques applied in each context, since one of the central purposes is precisely to document and justify the correct use of the methods and techniques employed. The activities are presented in chronological order, which also corresponds to a progressive increase in the time allocated to each one. After this introduction, the chapter 2 addresses the e-mail thread segmentation activity, the chapter 3 describes the Root Cause Analysis tool, and the chapter 4, divided into two phases, presents the logistics analysis carried out for an external client, which was the most extensive of the three. The report closes with a final chapter that summarises the work, reflects on the limitations encountered, and outlines directions for future work.

2. Behavioural Segmentation of E-mail Threads through Principal Component Analysis and Clustering

2.1. Problem, motivation and objective

The first activity developed during the internship was, to a large extent, an introductory exercise that served as an integration task within the company. It constituted a proof of concept aimed at exploring whether a particular analytical approach would be viable. The data set was not constructed during the internship, it was received already finalised, having been produced by an in-house extraction, transformation and loading (ETL) pipeline. The activity therefore began at the modelling stage, with a clean case-level table as its input.

The underlying idea was to treat e-mail communication as a process. When the exchange of messages within an organisation is viewed through the lens of process mining, each conversation, or thread, can be seen as a case, and each message as an event, with an activity attribute indicating whether the message started the thread, replied to it or forwarded it. Approached in this way, a mailbox becomes an event log, and the behavioural regularities of the communication, such as how widely conversations are distributed, how deeply they are nested, and how responsibility for them shifts among participants, become measurable. The objective of the analysis was to segment the threads into a small number of interpretable behavioural profiles, characterising the way in which conversations evolve in terms of audience size, depth, alternation of responsibility and influence of the initiator, to assess whether such a segmentation could yield meaningful and actionable categories.

2.2. Data and variables

The data originate from the mailbox of a single collaborator of the company. The raw collection was extracted from two archive files, namely the inbox and the sent items, and parsed into a process-mining-friendly format, in which each row represents an individual e-mail hop, that is, each successive message within a thread, together with a set of cumulative columns that store the state of every thread-level indicator at each hop. The unit of analysis, however, is the thread rather than the individual message. To collapse

the multi-hop structure into one row per thread, only the last hop of each thread was retained, since it carries the final cumulative state of every indicator. A categorical attribute identifying whether the thread had been initiated by the mailbox owner was also added.

A set of thread-level behavioural indicators is available, pre-computed in the ETL stage. The descriptions of these indicators are summarised in Table 1. The Responsibility Dilution Index measures the number of recipients relative to the number of distinct senders, so that high values indicate many recipients but few active contributors. The Passive Participation Rate captures the share of participants who were only carbon-copied. The Thread Expansion Rate measures how quickly a thread grows in number of recipients. The Responsibility Transfer Count records how many times the sender changes within a thread, and the Initiator Influence Score reflects the share of replies attributable to the thread initiator. The Action-to-Recipient Ratio relates the number of replies and forwards to the total number of recipients. The Avoidance Count records how often the previous sender is excluded from the next message, and the Cascade Depth corresponds to the total number of hops. The Cumulative Carbon-Copy Count and the categorical Thread Activity complete the set.

Table 1 - Thread-level behavioural indicators used in the segmentation, with a short description of what each one captures.

Indicator	What it captures
Responsibility Dilution Index	Recipients per distinct sender, high values indicate diluted ownership
Passive Participation Rate	Share of carbon-copy-only participants
Thread Expansion Rate	Speed at which the recipient set grows
Responsibility Transfer Count	Number of changes of sender within the thread
Initiator Influence Score	Share of replies attributable to the initiator
Action-to-Recipient Ratio	Replies and forwards per total recipient
Avoidance Count	Times the previous sender is excluded from the next message
Cascade Depth	Total number of hops in the thread
Cumulative CC Count	Cumulative number of carbon-copy recipients
Thread Activity	Whether the thread was initiated by the mailbox owner

Two filtering steps were applied before the analysis. Single-message threads were removed, since a thread with a single e-mail has no conversational dynamics to characterise. In addition, four threads were identified by visual inspection as non-legitimate communications, namely automated or no-reply messages whose indicators were anomalous, and were also removed: three were detected through an implausibly high Action-to-Recipient Ratio, typical of automated senders, and one through an inflated Thread Expansion Rate caused by an automatic reply sent within seconds. These threads do not reflect genuine human behaviour and would otherwise distort the downstream analysis.

2.3. Exploratory analysis and variable transformations

The analysis was conducted in Python. The exploratory analysis of the variables revealed three distinct kinds of distribution. Several indicators were strongly right-skewed, such as the Responsibility Dilution Index and the Action-to-Recipient Ratio. Others were bounded ratios confined to the unit interval, such as the Passive Participation Rate and the Initiator Influence Score. A third group consisted of discrete counts, such as the Cascade Depth, the Responsibility Transfer Count and the Avoidance Count. Because both Principal Component Analysis (PCA) and distance-based clustering benefit from variables with approximately symmetric distributions, a set of transformations was applied, guided by the shape of each distribution.

Strictly positive and right-skewed variables, such as the Responsibility Dilution Index and the Action-to-Recipient Ratio, were log-transformed, which compresses the long right tail and brings the distribution closer to symmetry. Count variables that could take the value zero were transformed with the $\log_{1p}$ function, which applies the logarithm to one plus the value and therefore handles zeros. Variables that were essentially binary in their information content were binarized, the Avoidance Count took the value zero in about eighty per cent of the threads and one in most of the remainder, so it was reduced to a binary indicator, and the Thread Activity, having only two categories, was likewise encoded as a binary variable. Introducing binary variables into PCA is a methodological compromise, since the method strictly assumes continuous inputs. The alternative of retaining the original integer counts would, however, have introduced spurious variability driven by a very small number of observations with values other than zero or one, without conveying additional information beyond the zero versus non-zero contrast. Finally, after these distribution-specific transformations, all the variables were standardised to zero

mean and unit variance, since both PCA and the distance-based clustering that followed are sensitive to differences in scale among the variables.

The dependence structure among the variables was examined using the Spearman rank correlation coefficient, shown in Fig. 1. This coefficient was preferred to the Pearson coefficient because, even after the log-transformations, the pairwise relationships are not necessarily linear. The Spearman coefficient is rank-based and therefore captures monotonic rather than only linear associations, is robust to skewness and outliers, and is meaningful for the discrete integer counts present in the data set [3]. For two variables, the Spearman coefficient is defined as the Pearson correlation between the ranks of the observations. In the absence of ties this reduces to,

$$\rho = 1 - \frac{6 \sum_i d_i^2}{n(n^2 - 1)} ,$$

where, d_i is the difference between the ranks of the i -th observation on the two variables and n is the number of observations. In the presence of tied ranks, as occurs for the discrete counts, the coefficient is computed directly as the Pearson correlation of the average ranks. Since the coefficient operates on ranks, it is invariant to any strictly monotonic transformation, such as the logarithm, which allowed the associations to be reported on the original, more interpretable scale. The correlation analysis revealed the structural relationships that PCA was subsequently expected to compress. Two examples illustrate the kind of structure present in the data: a strong positive association between the Cumulative Carbon-Copy Count and the Passive Participation Rate, of approximately 0.88, indicating that broader threads have proportionally more passive participants, and a strong negative association between the Responsibility Transfer Count and the Initiator Influence Score, of approximately -0.76, indicating that when ownership shifts often, the initiator drives the conversation less. These are only two of the many substantial pairwise associations visible in the correlation matrix, the presence of which justifies the use of Principal Component Analysis to summarise the redundant information into a smaller number of uncorrelated components.

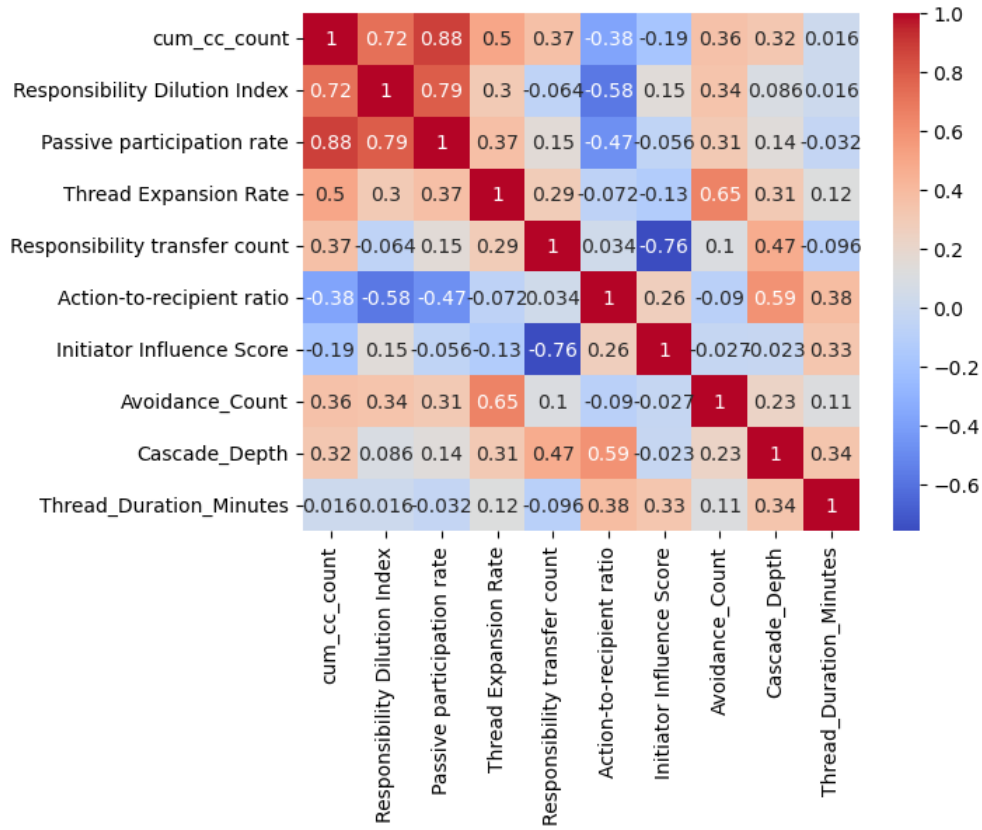


Fig. 1 - Spearman rank correlation matrix of the thread-level indicators, reported on their original scale.

2.4. Principal Component Analysis

Principal Component Analysis is a dimensionality-reduction technique that constructs a set of new, mutually orthogonal variables, the principal components, as linear combinations of the original variables, ordered so that each successive component captures the largest possible share of the remaining variance [4]. Given a data matrix whose columns have been standardised to zero mean and unit variance, the components correspond to the eigenvectors of the correlation matrix, and the variance explained by each component is given by the corresponding eigenvalue. Formally, the first principal component is the unit-norm linear combination w that maximises the variance of the projected data,

$$\max_w \text{Var}(Xw) \quad \text{subject to} \quad \|w\| = 1,$$

and each subsequent component maximises the same quantity subject to being orthogonal to all the previous ones. Working on the correlation matrix, that is, on

standardised variables, ensures that the decomposition reflects the dependence structure of the data rather than being dominated by the variables with the largest raw scales [4]. PCA was applied here both to summarise a moderately large set of correlated indicators into a few interpretable axes and to enable a low-dimensional visualisation of the segmentation structure.

The number of components to retain was decided using two complementary criteria. The cumulative-variance criterion, with a conventional threshold of seventy per cent, was satisfied at three components, while the Kaiser criterion, which retains components with an eigenvalue greater than one, suggested four, since the fourth component had an eigenvalue marginally above one. The scree plot of the cumulative explained variance is shown in Fig. 2. Three components were retained because the first three jointly captured approximately seventy-one per cent of the total variance, with the first three components explaining respectively 34.2, 22.3 and 14.3 per cent, and because retaining three components permitted a direct three-dimensional visualisation of the clustering structure, which was one of the goals of the analysis.

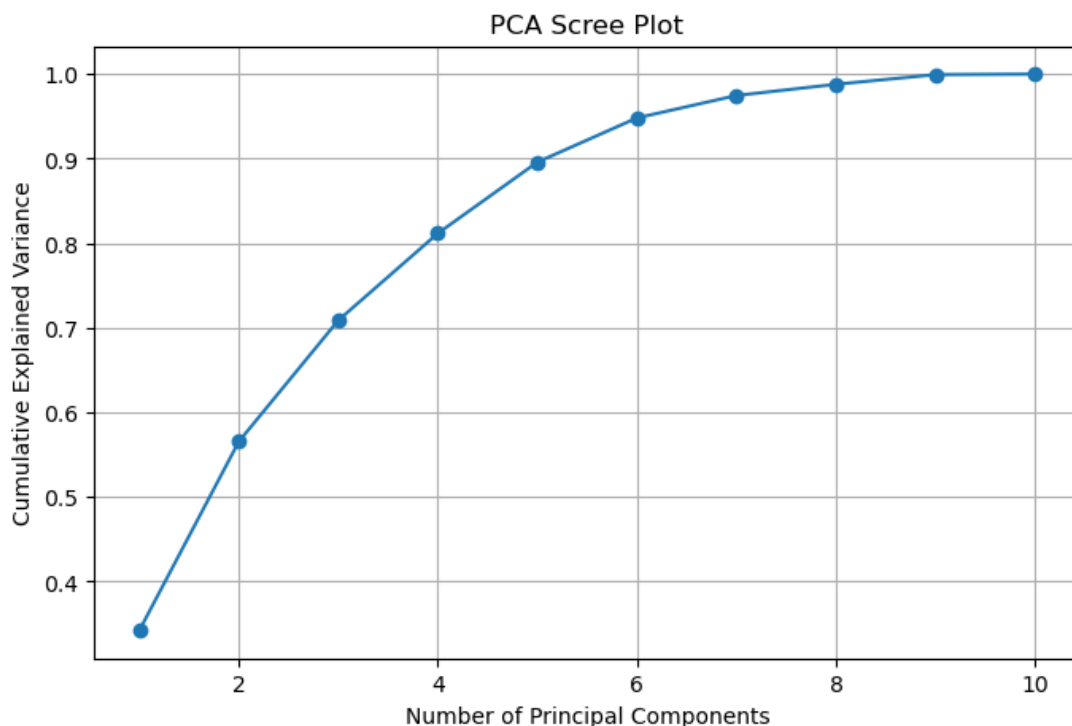


Fig. 2 - Scree plot of the cumulative explained variance as a function of the number of principal components retained.

The first three principal components admitted a clear interpretation, summarised in Table 2. The first component, accounting for approximately thirty-four per cent of the variance, contrasts threads with a broad audience and diffuse engagement, being dominated by

the cumulative carbon-copy count, the Passive Participation Rate, the Responsibility Dilution Index and the Thread Expansion Rate; it can be read as an axis of broad reach with passive diffusion. The second component, accounting for approximately twenty-two per cent of the variance, captures the depth and alternation of responsibility, being dominated by the Cascade Depth, the Responsibility Transfer Count and the Action-to-Recipient Ratio; high values mark long, sequential conversations in which ownership rotates among participants. The third component, accounting for approximately fourteen per cent of the variance, reflects the influence of the initiator, being dominated by the Initiator Influence Score; high values mark threads in which the initiator drives the conversation and retains control.

Table 2 - Interpretation of the first three principal components and the variables that dominate each one.

Component	Variance	Interpretation and dominant variables
PC1	34.2%	Broad reach with passive diffusion: CC count, passive participation, dilution, expansion
PC2	22.3%	Depth and alternation of responsibility: cascade depth, transfer count, action ratio
PC3	14.3%	Initiator influence: initiator influence score, action ratio

2.5. Clustering

The clustering was performed on the standardised feature space rather than on the PCA projection, to preserve the full information of the original variables, the principal-component coordinates were used purely for visualisation. This conservative approach is consistent with standard practice in unsupervised learning [3], since projecting onto a small number of components before clustering would discard the variance carried by the discarded components. Two complementary algorithms were used, chosen so that the agreement between them could serve as a check on the stability of the resulting partition.

The k -means algorithm partitions the observations into k groups by iteratively assigning each observation to the nearest centroid and recomputing the centroids to minimise the within-cluster sum of squared distances,

$$\min \sum_k \sum_{x \in \mathcal{C}_k} \|x - \mu_k\|^2,$$

where μ_k denotes the centroid of cluster C_k . Since minimising this objective exactly is NP-hard, the algorithm approximates it iteratively but may reach only a local minimum; it requires the number of clusters in advance and depends on random initialization of the centroids. Agglomerative hierarchical clustering with Ward linkage was applied as a complementary method: starting from each observation as its own cluster, it repeatedly merges the pair of clusters whose merger produces the smallest increase in the total within-cluster variance, until a single cluster remains [5]. The sequence of merges is represented by a dendrogram, whose horizontal cuts yield partitions at different levels of granularity. Both algorithms target compact, roughly spherical clusters, but they rest on different geometric mechanisms, k -means being centroid-based and Ward being variance-minimising and agglomerative, so their agreement is informative.

The number of clusters was selected by jointly reading the elbow curve of the within-cluster inertia and the silhouette coefficient calculated for the k -means solution, shown side by side in Fig. 3. The silhouette coefficient measures, for each observation, how similar it is to its own cluster relative to the nearest competing cluster, on a scale from minus one to one, with higher values indicating better-defined clusters [6]. The elbow curve did not exhibit a sharp bend, so it was compatible with several values, but the silhouette exhibited a local maximum at four clusters, followed by a dip at five and only then a monotonic rise that reflects progressive fragmentation rather than genuinely better-defined clusters. The dendrogram of the hierarchical solution, provided in the Annex, displayed a clear gap separating exactly four cohesive sub-trees. These three pieces of evidence converged on a four-cluster solution.

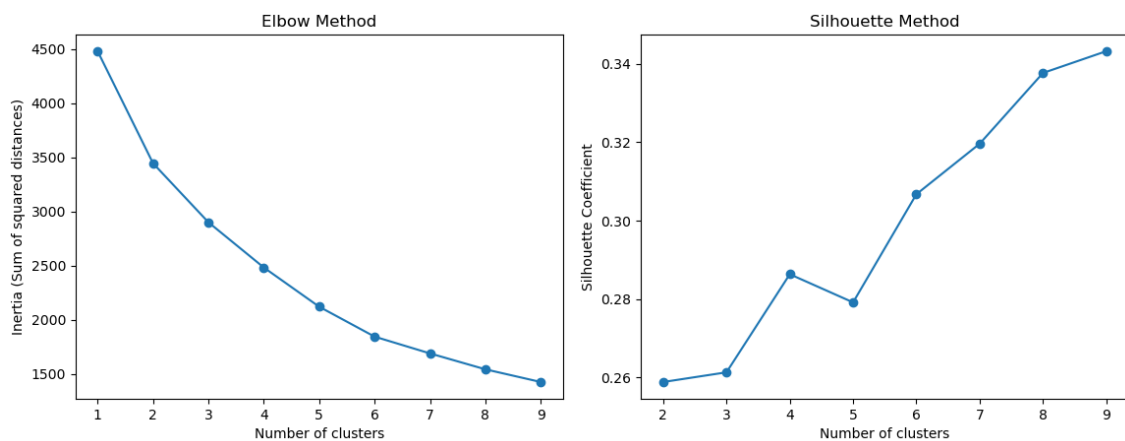


Fig. 3 - Elbow curve of the within-cluster inertia (left) and silhouette coefficient (right) of the k -means solution as a function of the number of clusters. The silhouette attains a local maximum at four clusters.

The agreement between the two clustering solutions was quantified using the Adjusted Rand Index (ARI), an external index that compares partitions of the same set of

observations and is invariant to permutations of the cluster labels, which is essential in cases where the labels of one partition do not correspond to those of the other. This can occur because the numeric labels produced by the two algorithms are arbitrary, so a naive label-by-label comparison would be meaningless.

The Rand Index counts the proportion of pairs of observations that are treated consistently by the two partitions, either co-clustered in both or separated in both, divided by the total number of pairs. The Adjusted Rand Index corrects this raw proportion for the value that would be expected by chance under a random labelling, by subtracting the expected agreement and dividing by the maximum possible agreement [7]. The resulting index ranges from minus one to one, takes the value zero for random partitions and the value one for identical partitions, and can be interpreted as the degree of agreement between the two partitions beyond what chance alone would produce.

2.6. Results and discussion

The projection of the threads onto the first two principal components, coloured by the k -means assignment, is shown in Fig. 4. Reading the cluster profiles together with their location on the principal axes, four distinct behavioural profiles emerged, summarised in Table 3.

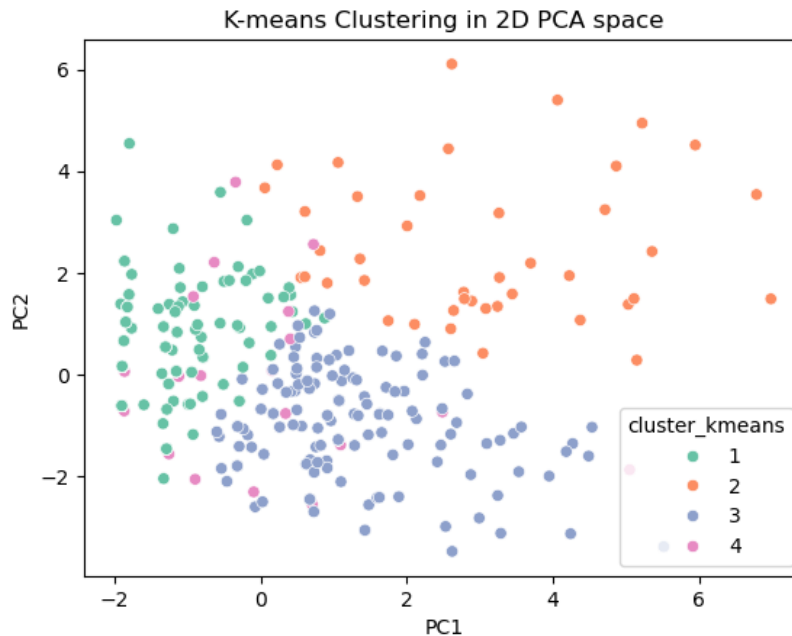


Fig. 4 - Projection of the threads onto the first two principal components, coloured by the four-cluster k-means solution.

The first profile is characterised by a low value on the first component, which corresponds to a narrow audience with little passive participation and limited diffusion of the

conversation, so that these are threads with a small recipient set and active engagement from the few participants involved. The second profile stands out on all three components, with high values on the first, the second and the third, and therefore combines a wide and largely passive audience, a deep and sequential structure with frequent rotation of ownership, and a dominant initiator whose contribution carries most of the conversation. The third profile is characterised by a high value on the first component together with a low value on the second, corresponding to threads with a broad audience and a high proportion of passive participants but limited depth and few transfers of ownership. The fourth profile is distinguished by a uniformly low value on the third component, while the first two components take moderate values without standing out in either direction, so that the defining feature of this group is the marked absence of initiator-driven dynamics, regardless of the audience size or the depth of the thread.

Table 3 - The four behavioural profiles identified by the clustering, with their location on the principal components.

Profile	Location	Behavioural description
Narrow and engaged	Low PC1	Small recipient set, little passive participation, active engagement
Broad, deep and initiator-driven	High PC1, PC2 and PC3	Wide passive audience, deep sequential structure, dominant initiator
Broad and shallow	High PC1, low PC2	Broad audience, many passive participants, limited depth
Low initiator influence	Low PC3	Moderate reach and depth, marked absence of initiator influence

The stability of the four-cluster solution was strongly supported. The k -means and the hierarchical solutions reached an Adjusted Rand Index of 0.749, above the conventional threshold of 0.7 that is usually interpreted as strong agreement, and the two algorithms assigned the same label to approximately ninety-two per cent of the threads. The spatial distribution of the agreement is shown in Fig. 5, where the threads on which the two algorithms differ are highlighted within the first-two-component projection. The disagreements are concentrated near the boundaries between clusters, as expected for threads whose profile sits between two behavioural groups, and which are therefore intrinsically harder to assign regardless of the algorithm used. The convergence of two

algorithms with different geometric assumptions indicates that the four clusters reflect a genuine pattern in the data rather than an artefact of a particular method.

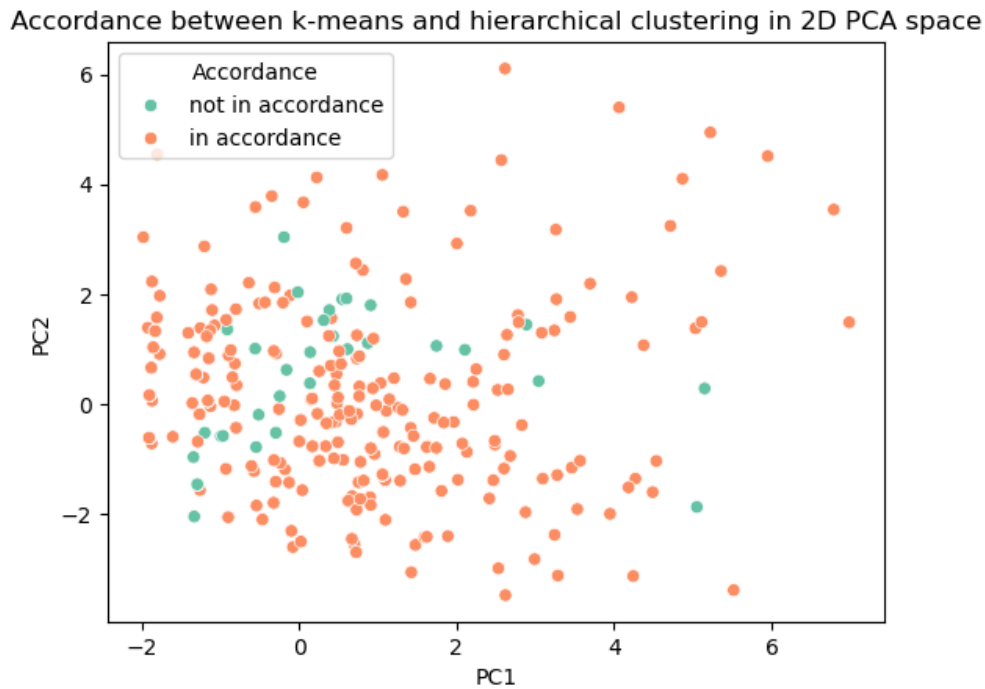


Fig. 5 - Spatial agreement between the k-means and the hierarchical solutions on the first two principal components. Disagreements concentrate at the boundaries between clusters.

As this activity was a proof of concept, the resulting segmentation was delivered as an exploratory characterisation. The analysis nevertheless demonstrated that the chosen approach, namely the combination of behavioural indicators, dimensionality reduction and clustering, can extract interpretable behavioural profiles from a communication log, which was the question that the proof of concept set out to answer. Two methodological reservations should be acknowledged. The first concerns the inclusion of binary variables within an analysis designed for continuous data: PCA strictly assumes continuous variables and Euclidean geometry, whereas the input retained two binary variables alongside the continuous and log-transformed ones, and a more rigorous alternative for mixed data would have been Factor Analysis of Mixed Data, which treats categorical and continuous variables on an equal footing. The second concerns the modest sample size, with approximately four hundred and sixty threads available for ten behavioural indicators, the stability of the resulting profiles may be limited, particularly when interpreting the segmentation at a more detailed level. Additional diagnostic figures, including the principal-component biplot, the dendrogram and the three-dimensional *k*-means projection, are provided in the Annex.

3. An Interactive Tool for Root Cause Analysis in Process Mining

3.1. Problem, motivation and objective

The second activity consisted of the development of a tool for internal use, created to address an existing need in the analysis of process-mining databases. For this type of analysis the company uses the *Microsoft Power Automate* software, in which one of the available metrics is the *Case Duration Influence*, the designation attributed by *Microsoft* to the Weighted Mean Deviation (WMD). For each subgroup of cases defined by a value of a given variable, this indicator measures the deviation that the subgroup introduces in the mean case duration relative to the global mean, weighted by the relative size of the subgroup. It therefore allows the identification of the values of a variable that are associated with slower or faster cases, which is a natural starting point for diagnosing the causes of long case durations.

The initial request was to extend this indicator so that, instead of evaluating the effect of a single variable at a time, it would allow several variables to be combined simultaneously, thereby quantifying the deviation of the cases that satisfy multiple conditions relative to the global mean. The limitation of the single-variable indicator is that it cannot reveal effects that only appear in conjunction, for instance a combination of a particular activity and a particular resource that is jointly associated with slow cases even though neither is markedly slow on its own. Starting from this motivation, and recognising the potential of a more comprehensive analysis, the decision was taken to develop an autonomous and more complete tool rather than a punctual extension of the metric.

Beyond the generalisation of the WMD to the multivariate case, two further capabilities were added. A set of non-parametric statistical tests was integrated, in order to complement and statistically substantiate the analysis by assessing whether the associations suggested by the descriptive indicators are statistically significant. An interpretable machine learning model, in the form of a decision tree, was also incorporated, with the aim of automatically identifying the combinations of attributes that best distinguish deviant cases from normal ones. The entire tool rests on the premise of supporting Root Cause Analysis (RCA), that is, the identification of the underlying causes of the deviations observed in a process, by progressively narrowing the analysis from

broad descriptive summaries to specific attribute combinations associated with deviant behaviour. The tool was tested and validated end-to-end using synthetic event-log data. As this activity produced a software tool rather than an analytical study, no discussion of substantive results is presented, the description that follows focuses on the design of the tool and on the justification of the methods it implements.

3.2. Implementation choices: the Streamlit framework

The tool was implemented in Python using the Streamlit framework. Streamlit is an open-source Python library for the rapid development of interactive web applications oriented towards data science, in which the entire interface, including interface widgets, is declared directly in a Python script and rendered as a web page that runs in the browser. Each user interaction triggers a re-execution of the script from top to bottom, with intermediate computations cached to avoid unnecessary recomputation. This reactive model removes the need to handle these aspects explicitly.

Streamlit was preferred to the alternative of a notebook-based prototype or a full web framework for three reasons. First, it allows the rapid translation of an analytical workflow into an interactive application without leaving Python, keeping the tool aligned with its analytical libraries. Second, the reactive execution model fits naturally with the analytical sequence that the tool is intended to enforce, in which each page depends on the choices made on the previous ones. Third, the resulting application can be deployed internally as a self-contained service, requiring no installation effort on the part of the analysts who use it, which is appropriate for an internal company tool that must be accessible to non-developer users.

3.3. Description of the tool

The tool is organised as a multi-page application in which a sidebar lists the seven analytical stages, and later pages remain locked until the earlier ones have been completed. This locking enforces a coherent analytical sequence and prevents, for instance, the application of a decision tree before the response variable has been defined. The internal logic was organised into a set of modules with clear responsibilities, namely the construction of the case table, the detection of variable types, the creation of the target, the production of plots, the computation of the Weighted Mean Deviation, the statistical tests, and the decision-tree pipeline. The overall workflow is summarised in Fig. 6.

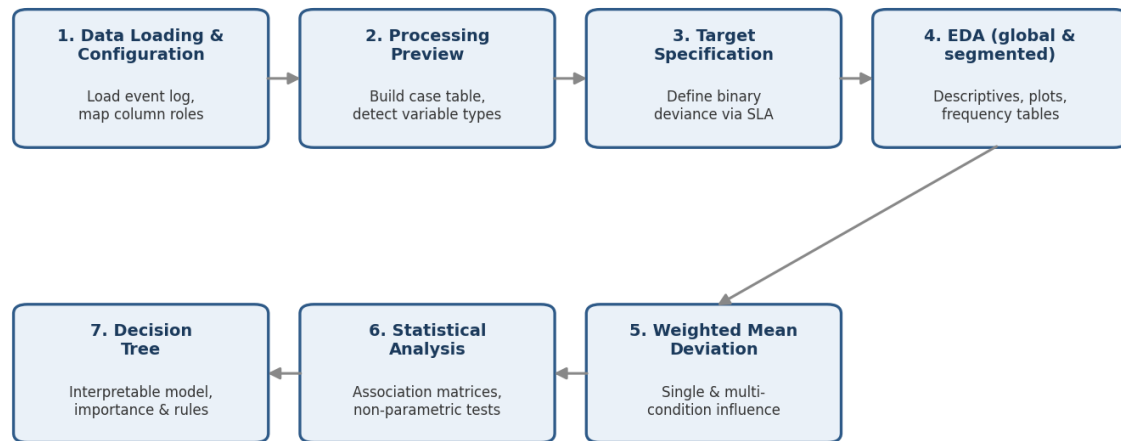


Fig. 6 - Page flow of the multi-page Root Cause Analysis tool developed in Streamlit. Later pages remain locked until the earlier ones are completed.

3.4. Data loading and case-table construction

The analysis begins with the ingestion of the event log and the mapping of its columns onto a set of normalised roles, namely the case identifier, the activity, the timestamps, and the event-level and case-level attributes, as illustrated in Fig. 7. This mapping decouples the tool from any particular column naming convention and allows it to operate on logs of different provenance, including logs whose columns were originally named in another language or with internal codes.

1. Load event log

CSV file

Drag and drop file here

Limit 200MB per file • CSV

Browse files

Procure-to-Pay.csv 1.1MB

X

Loaded 9119 rows × 9 columns.

	Case ID	Start Timestamp	Complete Timestamp	Activity	Resource	Role	Invoice amount	Discount	Country
0	339	16/02/2020 14:31:00	16/02/2020 15:23:00	Create Purchase Requisition	Nico Ojenbeer	Requester	135024.39	29576.82	JPN
1	339	17/02/2020 09:34:00	17/02/2020 09:40:00	Analyze Purchase Requisition	Maris Freeman	Requester Manager	135024.39	29576.82	JPN
2	339	17/02/2020 21:29:00	17/02/2020 21:52:00	Amend Purchase Requisition	Elvira Lores	Requester	135024.39	29576.82	JPN
3	339	18/02/2020 17:24:00	18/02/2020 17:30:00	Analyze Purchase Requisition	Heinz Gutschmidt	Requester Manager	135024.39	29576.82	JPN
4	339	18/02/2020 17:36:00	18/02/2020 17:38:00	Create Request for Quotation Requester Manager	Francis Odell	Requester Manager	135024.39	29576.82	JPN

2. Column mapping

Case ID

Start Timestamp

Case ID

Start Timestamp

Complete Timestamp

Activity

Complete Timestamp

Activity

Resource

Role

Event Level Attribute

Event Level Attribute

Fig. 7 - Data loading and column-mapping page. The event log is uploaded as a CSV file and each column is assigned to a normalised role.

From this mapping, the tool constructs the aggregated case table, in which the events are aggregated at the case level. For each case, a set of derived metrics is computed automatically, including the total case duration, the number of events, the number of unique activities, the number of repeated activities, the most frequent activity within the case, and a count of how often it occurred. The case-level attributes provided in the event log are collected and joined to this table. The type of each variable is detected automatically, as categorical, discrete numerical, continuous numerical or set-valued, the last category accommodating attributes that take a set of values per case, such as the set of resources that participated in a case. The automatic classification is exposed to the user for review and correction, as shown in Fig. 8, since an incorrect type assignment would lead to inappropriate plots and tests further down the workflow.

Aggregated Case Table										
	case_id	event_count	unique_activities_count	repeated_activities_count	most_frequent_activity	most_frequent_activity_count	case_duration	Invoice amount	Discount	Country
14	15	17	17	0	None	1	11353	253238.04	34253.26	NZL
15	16	24	20		2 Analyze Request for Quotation	4	9663	967995.03	33501.87	SGP
16	17	9	4	2	2 Analyze Request for Quotation	4	1624	316792.71	69919.91	NZL
17	18	18	18	0	None	1	10738	806678.48	25692.32	JPN
18	19	18	18	0	None	1	9086	812471.52	131148.54	AUS
19	20	29	21	2	2 Analyze Request for Quotation	6	17163	311028.11	8163.18	SGP
20	21	28	19	2	2 Analyze Request for Quotation	6	20209	869544.11	183320.81	NZL
21	22	22	20	2	2 Analyze Purchase Requisition ; Ana	2	14780	279084.77	8787.27	NZL
22	23	21	18	2	2 Analyze Request for Quotation	3	8982	727298.66	125474.62	SGP
23	24	19	18	1	1 Analyze Request for Quotation	2	17584	537988.66	55626.94	NZL

Case table shape: 608 rows × 10 columns.

Variable types ⓘ

event_count	unique_activities_count	repeated_activities_count
discrete	discrete	discrete
most_frequent_activity	most_frequent_activity_count	case_duration
categorical	discrete	continuous
Invoice amount	Discount	Country
continuous	continuous	categorical

Confirm variable types and continue

Fig. 8 - Processing-preview page, showing the aggregated case table and the variable-type assignment, which the user can review before the analysis proceeds.

3.5. Target specification

On the case table a binary response variable is defined that classifies each case as deviant or non-deviant, as illustrated in Fig. 9. The classification is built from a label-generating variable, typically the case duration, and a rule, such as greater than a given threshold, the threshold itself being derived from a Service Level Agreement (SLA), rather than from the empirical distribution of the durations. Anchoring the threshold in an externally agreed service level avoids the circularity of letting the duration distribution

define its own deviations and gives the label a clear business meaning. A case is deviant when it breaches a commitment set independently of the data, so the share of deviant cases is itself an informative outcome rather than a value fixed by construction. Once the target is confirmed, the tool reports the resulting class balance, so that the user can assess the degree of imbalance before proceeding.

Target Specification

Target ⓘ

Target name
case_deviance

Label-generating variable
case_duration

Rule
greater_than

Threshold (in the same units as the label-generating variable) ⓘ
14400.00

Confirm target

Target confirmed.

Class balance:

case_deviance	count	percentage
Deviant	392	64.47
Normal	216	35.53

Fig. 9 - Target-specification page. A binary deviance label is defined from a chosen label-generating variable, a rule and an SLA-anchored threshold, and the resulting class balance is reported.

With the target defined, the tool then offers a broad set of analyses. Descriptive statistics, histograms and bar charts, frequency tables and pairplots are produced, both globally and segmented by the response variable, so that the analyst can visually compare the distribution of each feature between deviant and non-deviant cases.

The Weighted Mean Deviation is computed, including the multi-condition version that combines up to three variables. Association matrices between variables are produced for the detection of redundancy, and hypothesis tests are provided to evaluate the association of each variable with the response.

Finally, a decision tree is fitted, whose hyperparameters can be adjusted and whose feature importance can be inspected, with an evaluation of the model. The remainder of this chapter describes and justifies the principal methods that the tool implements.

3.6. The Weighted Mean Deviation and its multivariate generalisation

The central indicator of the tool is the Weighted Mean Deviation. For a feature taking a particular value, let the cases sharing that value form a subgroup, the influence of the subgroup is defined as the deviation of the subgroup mean of the response from the global mean, weighted by the relative size of the subgroup and expressed relative to the global mean,

$$\text{influence} = \frac{(\bar{x}_{\text{subgroup}} - \bar{x}_{\text{global}}) \times (n_{\text{subgroup}}/n_{\text{total}})}{\bar{x}_{\text{global}}},$$

The weighting by the relative size of the subgroup is what distinguishes this indicator from a simple difference of means. A subgroup that is markedly slower than average but contains very few cases contributes little to the overall behaviour of the process and therefore receives a small influence, whereas a subgroup that is only moderately slower but very large can carry a substantial influence. The indicator is signed, so that positive values identify subgroups associated with slower-than-average cases and negative values identify subgroups associated with faster-than-average cases, and it is expressed as a percentage of the global mean, which makes the values comparable across features.

In its single-feature form, the indicator is computed once per pair of feature and value, and the resulting table can be filtered by feature or sorted by the absolute value of the influence, so that the values most strongly associated with deviant behaviour appear at the top, as illustrated in Fig. 10. A minimum subgroup-size threshold is exposed to the user so that subgroups too small to be operationally relevant can be excluded, an option that is particularly useful when the data set contains a long tail of rare categorical values.

For numerical variables, including continuous ones, the indicator is computed at the level of each unique observed value, in line with the implementation in Microsoft Power Automate; the weighting by the relative size of the subgroup ensures that values that appear in only one or two cases contribute negligibly to the ranking, so that no preliminary binning is required.

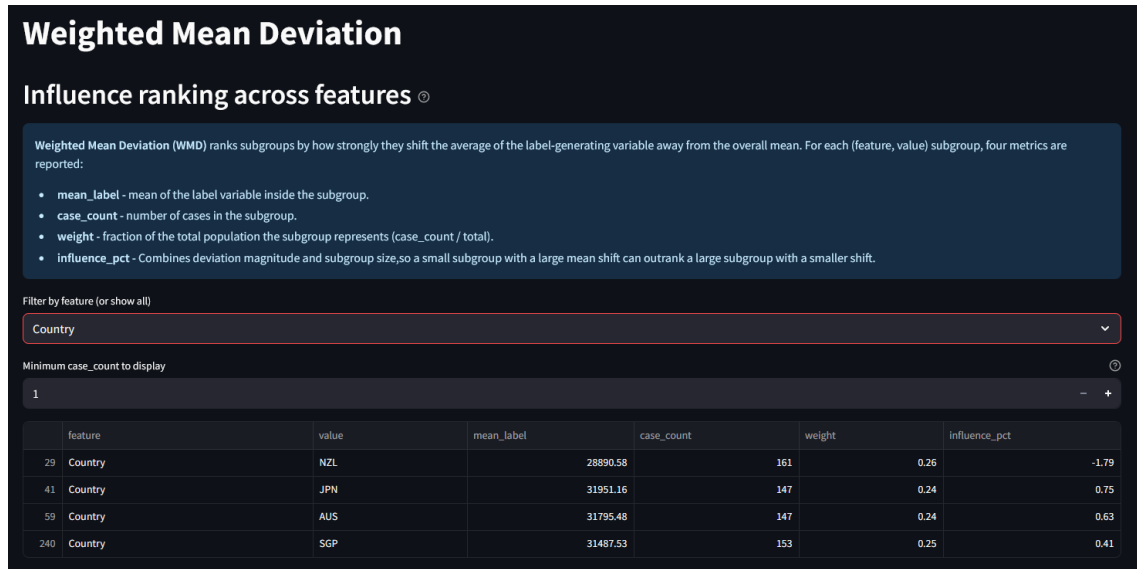


Fig. 10 - Single-feature Weighted Mean Deviation page. For each value of the selected feature, the subgroup mean, count, weight and signed influence are reported.

The generalisation implemented in the tool replaces the single-value subgroup by the set of cases that simultaneously satisfy a list of conditions on up to three variables, so that the influence of a conjunction of attribute values can be quantified in the same units, as illustrated in Fig. 11. Each condition is expressed as a triple of feature, comparison operator and value, and the conditions are combined by logical conjunction. This multi-condition formulation is the principal extension requested by the company and allows the analyst to isolate the joint effect of attribute combinations that single-variable analysis cannot capture, addressing directly the limitation that motivated the activity.

The interface guides the user step by step, with appropriate operators offered for each variable type, namely equality and inequality for categorical variables and the usual comparison operators for numerical ones, so that invalid combinations cannot be specified.

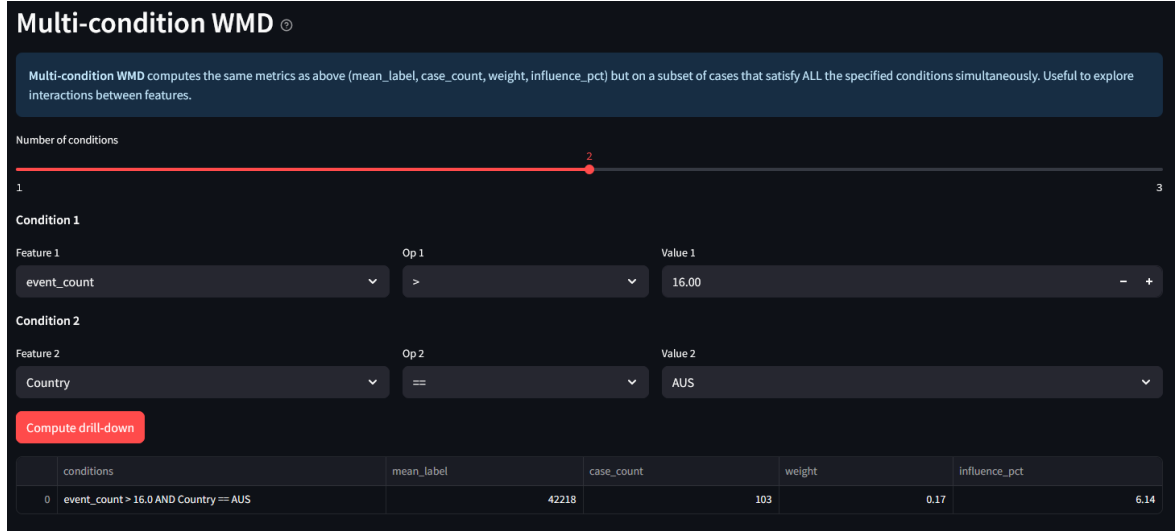


Fig. 11 - Multi-condition Weighted Mean Deviation page. The influence is computed on the subset of cases that simultaneously satisfy all the specified conditions.

3.7. Detection of redundancy among variables

Before the associations between the candidate explanatory variables and the response are tested, it is useful to detect redundancy among the explanatory variables themselves, since two strongly associated variables would convey overlapping information and could lead to misleading conclusions if treated as independent causes. Two association matrices are provided for this purpose, according to the type of the variables involved.

For pairs of categorical variables, the bias-corrected Cramer's V is used. Cramer's V is a measure of association derived from the chi-square statistic of the contingency table, that is, the cross-tabulation in which each row corresponds to a category of the first variable, each column corresponds to a category of the second variable, and each cell records the number of observations that take the combination of the two categories. For a table with n observations, r rows and k columns, Cramer's V is defined as

$$V = \sqrt{\frac{\chi^2}{n \times \min(r - 1, k - 1)}} ,$$

and ranging from zero, for no association, to one, for perfect association [8], [9]. The raw statistic is known to be biased upwards in small or sparse tables, where it can suggest association even between independent variables. The bias correction proposed by Bergsma adjusts the statistic and the table dimensions to remove this upward bias. This corrected version is the one that the tool computes and reports to the user in the

association matrix. For pairs of numerical variables, the Spearman rank correlation coefficient is used, for the same robustness reasons given in the previous chapter, namely its insensitivity to non-linearity, skewness and outliers [3]. The two matrices together allow the analyst to identify and, if appropriate, set aside redundant variables before interpreting the results.

3.8. Non-parametric hypothesis tests

The objective of this stage is to assess, in a statistically substantiated way, whether each candidate variable is significantly associated with the binary response. To this end, a set of non-parametric hypothesis tests is provided, chosen to avoid the distributional assumptions that process-mining variables rarely satisfy.

For categorical variables, the chi-square test of independence compares the observed cell counts of the contingency table that cross-tabulates the variable against the response with the counts expected under the null hypothesis of independence,

$$\chi^2 = \sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}},$$

where O_{ij} and E_{ij} denote, respectively, the observed and the expected counts in cell i, j . A small p -value indicates that the distribution of the variable differs between the deviant and the non-deviant cases.

For numerical variables, the Mann-Whitney U test is used to compare the distribution of the variable between the deviant and the non-deviant cases. This test is non-parametric and assesses whether one of the two groups tends to produce larger values than the other, by ranking all the observations jointly and comparing the rank sums of the two groups, without assuming normality of the underlying distributions [10]. This makes it appropriate for the skewed and heavy-tailed variables that are common in process data, where the assumptions of the classical t -test would be violated.

As a complementary procedure, a bootstrap test of the difference in medians is also provided. The bootstrap approximates the sampling distribution of a statistic by repeatedly resampling the observed data with replacement and recomputing the statistic on each resample [11]. In this case the difference in the median of the variable between the deviant and the non-deviant cases is recomputed on a large number of resamples, yielding an empirical distribution from which a confidence interval and a p -value are

obtained without any distributional assumption. Providing both a rank-based test and a resampling-based test gives the analyst two complementary perspectives on the same question, which is valuable when the conclusions are to be acted upon.

3.9. The interpretable decision tree

Finally, an interpretable machine learning model is incorporated in the form of a classification decision tree. A decision tree is a hierarchical model in which each internal node corresponds to a test on a single feature that divides the cases reaching it into two subsets, and each terminal node, or leaf, corresponds to a region of the feature space in which a single class is predicted. The tree is built by recursively partitioning the feature space into regions that are as homogeneous as possible with respect to the response, choosing at each node the split that most reduces the impurity of the resulting subsets. The Gini impurity of a node is defined as

$$\text{Gini} = 1 - \sum_c p_c^2 ,$$

where p_c is the proportion of cases of class c in the node, so that the impurity is zero when the node is pure and maximal when the classes are balanced [3]. The decision tree was chosen, rather than a more accurate but opaque model such as a gradient-boosting ensemble, precisely because of its interpretability, which is the property that matters most in the context of Root Cause Analysis. The path from the root of the tree to a leaf can be read directly as a conjunction of conditions on the features, and therefore yields human-readable rules that describe the combinations of attributes most associated with deviant cases, exactly the kind of output required to support a root-cause investigation.

Because the deviant cases are typically a minority, the tree is fitted with class weights inversely proportional to the class frequencies, so that the minority class is not neglected during training and the model does not collapse into always predicting the majority class. The hyperparameters that govern the trade-off between fidelity and interpretability are exposed to the user, as shown in Fig. 12, namely the maximum depth, the minimum number of cases per leaf and the proportion of cases held out for evaluation, together with the seeds that control the train-test split and the tie-breaking decisions of the tree, so that runs can be made reproducible. The maximum depth in particular controls the complexity of the rules extracted, since a shallow tree yields few, simple rules that are

easy to communicate but may miss structure, while a deeper tree captures more structure at the cost of producing rules that are harder to interpret.

Decision Tree

Decision tree classifier for root cause analysis. The tree recursively splits the case table on attribute thresholds to maximise separation between deviant and normal cases. Output is fully interpretable: each leaf is a subgroup of cases defined by an explicit conjunction of conditions, and we report the deviant rate within each subgroup.

Hyperparameters

Adjust the parameters that control tree complexity and training:

- **max_depth** - maximum number of splits from root to leaf. Higher values let the tree capture more patterns but risk overfitting; lower values keep it simpler and more interpretable.
- **min_samples_leaf** - minimum number of cases in any leaf. Avoids rules supported by very few cases.
- **test_size** - fraction of cases held out to evaluate generalisation.
- **class_weight = 'balanced'** - re-weights classes inversely proportional to their frequency, so the tree pays similar attention to both classes even when one is rare.
- **Random states** - fix the train/test split and the tree's tie-breaking for reproducible runs.

max_depth: 3 (range 2 to 100)
min_samples_leaf: 10 (range 1 to 100)
test_size: 0.30 (range 0.10 to 0.50)
class_weight: ☒ 'balanced'
Random state (split): 333
Random state (tree): 333
Fit decision tree

Fig. 12 - Decision-tree page. The hyperparameters that govern the depth and the regularisation of the tree are exposed to the user, alongside the seeds that control reproducibility.

Once the tree is fitted, the tool reports the test-set accuracy, the confusion matrix and the F1 score, the latter being particularly informative under class imbalance, as well as the feature importances aggregated across the splits and a textual representation of the tree itself, so that the extracted rules can be inspected. The tool therefore provides, in a single guided workflow, the descriptive, the inferential and the predictive perspectives on the same root-cause question, each of them substantively informed by the methodological choices justified above, namely the anchoring of the deviance threshold in an externally agreed SLA, the choice of non-parametric tests so that no normality of the underlying variables is required, and the deliberate choice of an interpretable decision tree over a more accurate but opaque model, since interpretability is the property that matters most in the context of RCA.

4.A Logistics Analysis of Pallet Consolidation and Utilization for an External Client

The third and most extensive activity of the internship was an analysis developed for an external client, here referred to simply as the client, operating in the manufacturing and distribution of items, namely windows and doors. As in the first activity, the data set was not constructed during the internship, it was built beforehand by the data engineer of the company, from data extracted from the SAP information system of the client referring to the year 2024. The resulting data set contains 695,891 rows and 106 columns and describes the logistics process surrounding the storage of items, their consolidation into pallets and the shipping of those pallets between the warehouses of the client in Central Europe.

The structure of the database is inherently complex because it incorporates several case-identifier dimensions simultaneously, each representing a different perspective of the logistics process, namely the sales delivery, the pallet, the freight-agent tracking unit and the shipment. This multiplicity of dimensions means that a single row of the database may combine information originating from distinct levels, which makes the analysis more demanding and requires a clear definition of the unit of observation before any modelling decision. The unit of observation adopted in this project is the pallet, identified by the field denoted here as the pallet delivery identifier, which corresponds to an individual shipped pallet. This choice was made because the pallet is the natural object of both questions that the client wished to answer, namely how pallets are loaded and how efficiently they are used.

Although the database contains more than one hundred columns, the great majority are not usable for the analysis, for two main reasons. First, many of the variables are unrelated to the problem under study, referring to operational or administrative aspects with no bearing on the use of the pallets. Second, a substantial share of the variables is defined at a hierarchical level different from the unit of observation, namely at the level of the sales delivery, the shipment or the freight-agent tracking unit. The analysis therefore retained only the variables that are both relevant to pallet usage and either defined at the pallet level or meaningfully aggregated to it. The set of variables that was ultimately used is described in the second phase of the project, in which it became the basis of the statistical models.

Two notions related to the composition of a pallet are central to the project. A pallet is considered mixed when it consolidates items originating from two or more distinct sales orders, and single-order, or non-mixed, when it carries items from a single sales order. Based on this distinction, the project was organised into two parts with distinct objectives. The first part analyses the use of non-mixed pallets with few items in the shipment of orders between warehouses and estimates the optimisation potential arising from a better consolidation, quantifying the reduction in the number of pallets and trucks and the avoided CO2 emissions. The second part identifies the operational factors with the greatest influence on pallet utilization, determining which characteristics of the logistics process are associated with the classification of a pallet as well or poorly used.

The first phase was conducted in Python. In the second phase, the feature engineering, the exploratory analysis and the ensemble tree-based models were also developed in Python, while the Generalized Linear Mixed Model (GLMM) was fitted in R, which offers a more mature ecosystem for the fitting and diagnostics of mixed models. A presentation of the results was delivered to the client at the end of each phase.

4.1. Phase 1: Quantifying the Logistics Optimization Potential of Pallet Consolidation

4.1.1. Problem, motivation and objective

A pallet that carries only one, two or three items makes inefficient use of the space available for transport, since the same physical pallet position and, ultimately, the same truck capacity are consumed regardless of how few items are carried. When several such small non-mixed pallets share an origin, a destination and a shipping time window, they could in principle be combined into a smaller number of mixed pallets, reducing the total number of pallets shipped and therefore the number of trucks required. The objective of this first phase was to quantify, by route and by period, the reduction in the number of pallets and trucks that would result from consolidating non-mixed pallets carrying between one and three items, together with the carbon-dioxide emissions that would be avoided through the reduction in the number of journeys.

This phase is descriptive and quantitative in nature, it does not fit a statistical model but rather constructs, from the event log, a to-be scenario in which the small non-mixed pallets are consolidated subject to the operational constraints described above, namely the requirement of a shared route and shipping time window and the ceiling on the

number of mixed pallets that each warehouse can produce in a period, and compares this scenario with the as-is scenario observed in the data. The value of the exercise lies in translating an operational intuition, namely that many pallets are shipped underfilled, into a defensible estimate of the magnitude of the resulting waste and of the benefit that consolidation would bring.

4.1.2. Methodological decisions

The consolidation candidates were defined as the non-mixed pallets carrying one, two or three items at the moment of dispatch. Pallets carrying four or more items are excluded as they already make adequate use of the available space and are therefore not candidates for consolidation. To respect the operational reality that pallets can only be combined when they are shipped together, the analysis was organised by route and by period, where each week was divided into three day-groups, corresponding to the early week, the mid week and the end of the week. This temporal granularity reflects the actual dispatch rhythm of the warehouses and ensures that only pallets that could realistically have travelled together are considered for consolidation.

Within each warehouse and period, the small non-mixed pallets were paired into new mixed pallets, subject to an operational ceiling on the number of mixed pallets that a warehouse can produce in a period. This ceiling accounts for the additional handling effort that the creation of mixed pallets entails and for the mixed pallets that the warehouse already creates in normal operation within the same period, so that the to-be scenario does not assume an unbounded capacity to build mixed pallets. The number of trucks was then derived from the number of pallets using the standard loading capacity of a full truck, since the cargo is volume-constrained and the number of pallet positions, rather than the weight, determines how many trucks are needed. Two outlier routes with negligible and irregular activity were removed from the analysis.

The share of small non-mixed pallets was found to vary substantially across routes and over time, as illustrated for one representative route in Fig. 13, where the small non-mixed pallets are shown as a distinct band within the total number of pallets shipped in each period. The same pattern, examined at the level of the warehouse and of the day-group, showed that the incidence of small non-mixed pallets is not uniform across the week, which justifies the per-period treatment adopted in the consolidation logic. The corresponding day-group breakdown for the two warehouses is shown in Fig. 14.

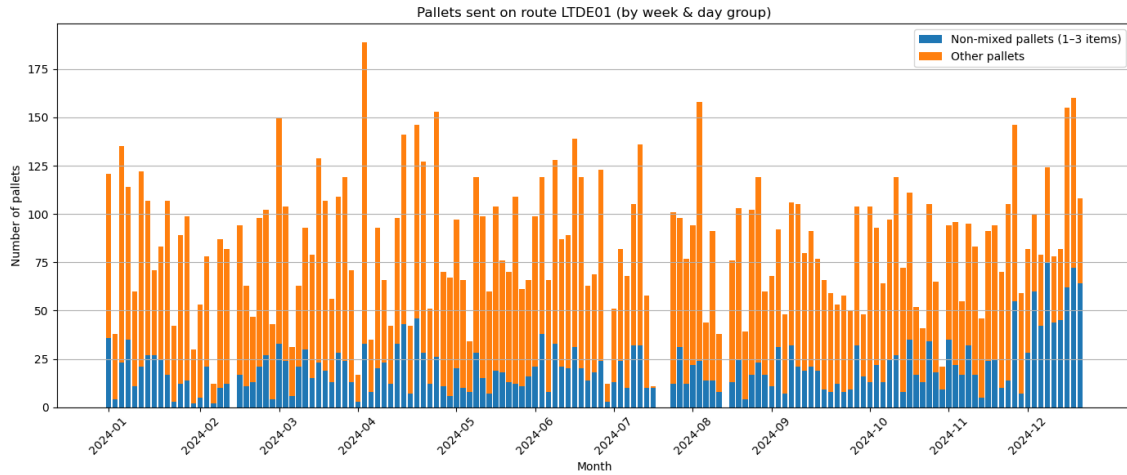


Fig. 13 - Number of pallets shipped per period on a representative route, distinguishing the small non-mixed pallets carrying one to three items (lower band) from the remaining pallets (upper band).

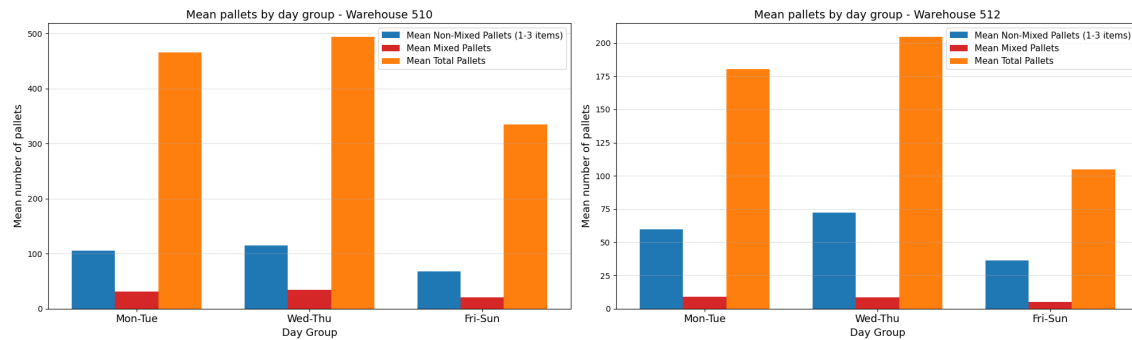


Fig. 14 - Mean number of non-mixed, mixed and total pallets by day-group for the two warehouses, showing that the incidence of small non-mixed pallets varies across the week.

4.1.3. Estimation of the avoided CO2 emissions

The avoided emissions were estimated following the Global Logistics Emissions Council (GLEC) Framework, version three [12], in alignment with the international standard for the quantification and reporting of greenhouse-gas emissions from transport-chain operations [13]. A well-to-wheel accounting basis was adopted, so that the emissions arising from the upstream production of the fuel are included in addition to those produced during combustion, as required for complete reporting. Because the cargo is volume-constrained rather than weight-constrained, the pallets occupy the floor of the trailer well before the maximum payload of the vehicle is reached; hence, a vehicle-kilometre method was used in preference to a tonne-kilometre method. The latter would underestimate fuel consumption in this setting, since aerodynamic drag and the tare weight of the vehicle, rather than the payload, dominate fuel use. The baseline emission factor was interpolated according to the actual load factor of the vehicle, yielding a corrected emission factor per vehicle-kilometre that was then multiplied by the route

distance and by the number of journeys avoided. The physical parameters used in the estimation, namely the average number of items per pallet, the average item weight, the pallet weight, the number of pallets in a full load and the maximum payload, were agreed with the client so that the estimate would rest on assumptions that the operational team recognised as realistic.

4.1.4. Results and discussion

Aggregated over the whole of 2024 and across all routes, the consolidation of small non-mixed pallets was estimated to reduce the total number of pallets shipped from 84,791 to 78,165, a reduction of 7.8 per cent corresponding to 6,626 fewer pallets. The effect was much stronger when restricted to the targeted small non-mixed pallets, whose number fell from 21,889 to 15,263, a reduction of 30.3 per cent. The reduction in pallets translated into a reduction in the number of trucks from 4,569 to 4,235, a saving of 334 trucks, and into an estimated saving of approximately 188 tonnes of CO₂ equivalent. The route-level results are summarised in Table 4.

Table 4 - Estimated optimisation potential from pallet consolidation, by route and in total, for the year 2024.

Route	Pallets before	Pallets after	Pallet reduction (%)	Truck savings	CO ₂ e savings (t)
LTPL01	11,541	9,048	21.6	127	42.7
PLDE01	47,351	45,447	4.0	93	59.1
LTDE01	12,261	10,844	11.6	72	72.6
PLPL01	9,059	8,393	7.4	34	0.5
PLIE01	4,579	4,433	3.2	8	13.5
Total	84,791	78,165	7.8	334	188.0

The results revealed a marked heterogeneity across routes. The route denoted LTPL01 presented the largest relative reduction in the total number of pallets, at 21.6 per cent, reflecting both a high incidence of small non-mixed pallets, almost half of which could be eliminated, and a comparatively short route over which trucks are used intensively. Routes with a high absolute volume, such as PLDE01, exhibited a smaller relative reduction in total pallets, because the small non-mixed pallets represent a smaller share of an already large flow, yet still contributed a substantial absolute number of avoided trucks. The relationship between truck savings and avoided emissions is mediated by the route distance, which explains why a route with a modest number of avoided trucks

but a long distance, such as LTDE01, accounts for the largest share of the avoided emissions, while a very short route, such as PLPL01, contributes negligibly to the emissions saving despite a non-trivial truck saving. This dissociation between truck savings and emission savings is an important nuance for the client, since it implies that the routes on which consolidation yields the greatest operational benefit are not necessarily those on which it yields the greatest environmental benefit.

The effect of consolidation on the composition of the shipped pallets is illustrated in Fig. 15, which contrasts the number of mixed pallets before and after consolidation for one warehouse. The consolidation systematically increases the number of mixed pallets, since the procedure converts pairs of small non-mixed pallets into mixed ones, up to the operational ceiling, which is visible in the figure as the level at which the post-consolidation bars saturate.

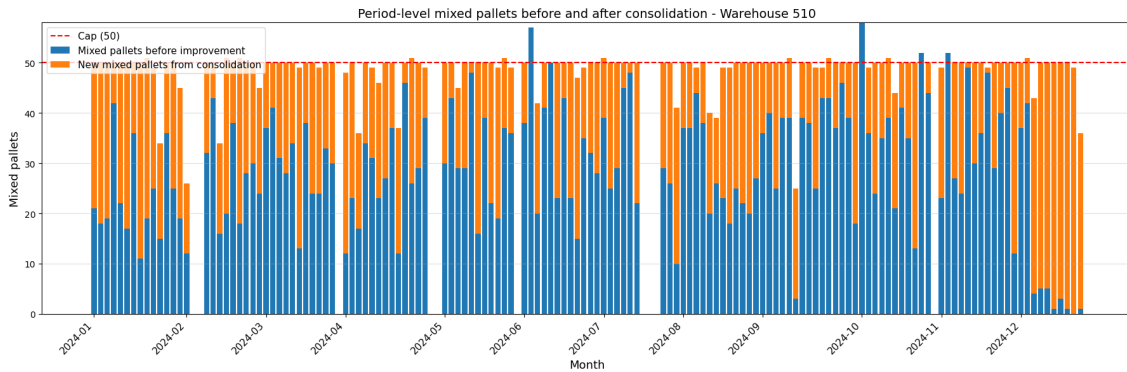


Fig. 15 - Number of mixed pallets per period before and after the proposed consolidation, for one warehouse. The dashed line marks the operational ceiling on new mixed pallets per period.

These figures should be read as an estimate of the optimisation potential under the stated assumptions rather than as a guaranteed saving. In particular, the ceiling imposed on the number of mixed pallets per warehouse and period, the assumed truck loading capacity and the physical parameters of the emission model all influence the magnitude of the estimated savings. The analysis nevertheless provides a defensible order of magnitude for the benefit of consolidation and identifies the routes on which such an intervention would be most worthwhile. Additional figures, including the distribution of small non-mixed pallets per period, are provided in the Annex.

4.2. Phase 2: Identifying the Operational Drivers of Pallet Utilization

4.2.1. Problem, motivation and objective

Whereas the first phase quantified the potential benefit of consolidation, the second phase sought to understand why pallets are well or poorly used in the first place, by identifying the operational factors associated with the quality of a pallet. The response variable of the study is the pallet utilization, a binary variable that classifies each shipped pallet as a good or a bad pallet according to how efficiently its space is used. The definition of what constitutes a good and a bad pallet was provided by the client, on the basis of operational criteria established internally, and was not derived from the data. It combines two criteria, namely the number of items carried and whether the pallet is mixed: a non-mixed pallet is considered good when it carries at least four items, while a mixed pallet is considered good when it carries at least six items. Any pallet not meeting the corresponding threshold is classified as bad. The rationale is that a well-used pallet makes adequate use of the space and of the transport, whereas a pallet with few items represents an inefficient use. The more demanding threshold applied to mixed pallets follows from the additional operational effort required to create them, which must be compensated by a larger number of items for the pallet to be considered well used. The objective of this phase was to determine which characteristics of the logistics process are associated with the classification of a pallet as good or bad, offering a substantiated basis for discussion with the operational team.

Two distinct modelling strategies were pursued in parallel and subsequently compared. A GLMM was fitted as the core inferential model, prioritising the interpretability of the effects and the quantification of their uncertainty, since the principal goal of the phase is explanatory. Two ensemble tree-based models, a Random Forest and a gradient-boosting model, were fitted as complementary approaches because they rely on far fewer assumptions than the GLMM. Consequently, they can accommodate a broader set of features, including variables that the GLMM could not retain because of multicollinearity or violations of the linearity of the logit assumption, while automatically capturing non-linear relationships and interactions among predictors. Because the explanatory objective of the phase requires interpretability, which these models lack by themselves, their fit was complemented by the computation of Shapley Additive exPlanations (SHAP) values, which recover variable-level explanations from the otherwise opaque ensemble predictions. An isolated decision tree was not chosen as an alternative, even though it is intrinsically interpretable, because it is typically less

accurate than ensemble methods, and because the combination of an ensemble model with SHAP values delivers both higher predictive accuracy and a post-hoc interpretability that, although different in nature from that of a single tree, is sufficient for identifying the variables most associated with the response. The use of both paradigms therefore allows the conclusions of the inferential model to be cross-validated against a more flexible class of models, and allows variables that the inferential model could not accommodate to be examined.

4.2.2. Feature engineering and aggregation

The starting point was the raw event-level log, in which each row records a single activity performed on a pallet. From this log a table with one row per pallet was constructed. For each pallet, the reference row was taken to be its dispatch activity, since it is the last row in the trace of the pallet and therefore carries the most up-to-date values of the event-level attributes for that case. This was complemented by the earliest recorded start timestamp of the pallet. A set of derived variables was engineered at the pallet level, including the processing duration, indicators of delivery completeness, seasonality descriptors such as the quarter and the day of the week of dispatch, and a number of shipment-level measures summarised per pallet. The original SAP extract was reduced to the subset of variables that are relevant to pallet utilization and either defined at the pallet level or meaningfully aggregated to it, keeping the data set aligned with the chosen unit of observation. The engineered table was exported for the subsequent modelling stages. The variables that were ultimately retained, together with a short description of each, are listed in Table 5.

Table 5 - Variables defined at or aggregated to the pallet level.

Variable	Description
Pallet duration	Hours between the first and the last recorded activity of the pallet
Quarter of dispatch	Quarter of the year in which the pallet was shipped
Day of the week of dispatch	Day of the week on which the pallet was shipped
Is the last delivery	Whether the pallet is the final shipment for at least one of its sales deliveries
Shipment duration	Hours spanned by the shipment to which the pallet belongs
Shipping route	Route along which the pallet is shipped

Storage unit type	Type of storage location where the pallet is held until shipped
Identity	Product model carried by the pallet
Is delivery with broken item	Whether a damaged item was detected before dispatch
Items moved between pallets	Whether items passed through an intermediate pallet
Pallet type	Pallet width, 80 cm or 120 cm
Pallet area	Surface area of the pallet, in square metres
Total items per shipment	Total number of items in the shipment
Number of pallets per shipment	Total number of pallets in the shipment
Occupied pallet area per shipment	Total pallet area occupied within the shipment
Rack allocation in shipment	Whether racks are present in the shipment
Is mixed pallet	Whether the pallet contains items from different sales orders
Items in delivery sent	Number of items on the pallet at the moment of dispatch
Pallet utilization	Binary response: good or bad pallet according to item count and mixing

4.2.3. Exploratory analysis and preparation for modelling

The data were split into a training set and a test set before any exploratory analysis or modelling decision, so that the test set would remain entirely unseen throughout the analytical process and no information could leak from it into the modelling choices. The split was performed at the shipment level, so that every shipment appears entirely in either the training or the test set. This is the methodologically correct approach given that pallets belonging to the same shipment are not independent, and it ensures that the test set contains shipments completely unseen during training. A fixed random seed was used to guarantee the reproducibility of the partition. Three variables were removed prior to the split because they define the target, directly or indirectly: the number of items on the pallet and the mixed-pallet indicator define the response by construction, and the number of sales orders defines the mixed-pallet indicator in turn. Retaining any of these would create perfect prediction and invalidate the evaluation.

All subsequent exploratory analysis, transformations and assumption checks were performed exclusively on the training set. The response variable exhibited a moderate imbalance, with approximately seventy-three per cent of good pallets against twenty-seven per cent of bad pallets, as shown in Fig. 16. This imbalance rules out accuracy as a reliable evaluation metric, since a naive classifier that always predicted the majority class would attain seventy-three per cent accuracy without learning anything, the area under the receiver operating characteristic curve (AUC-ROC), and the Brier score were therefore adopted as the principal evaluation metrics, complemented by the sensitivity and the specificity at the default classification threshold. Given the moderate degree of imbalance, resampling techniques were not required.

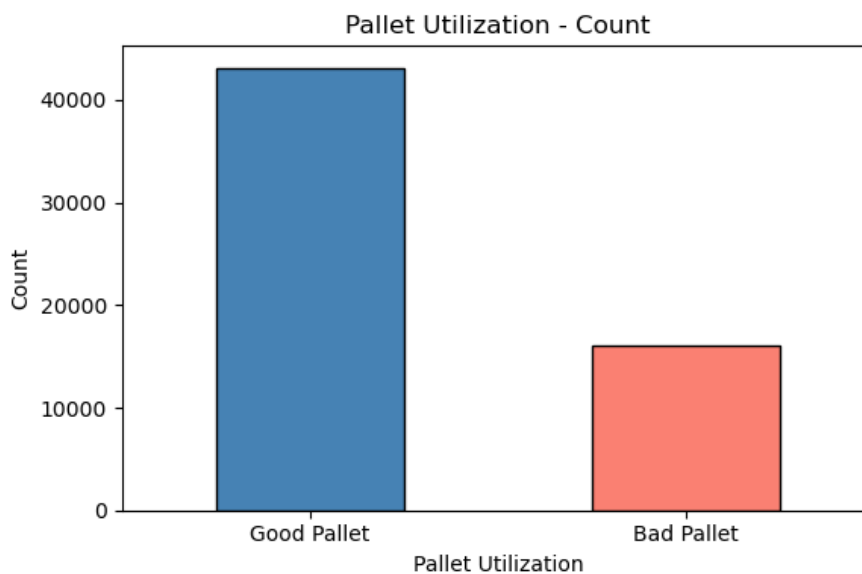


Fig. 16 - Distribution of the binary response variable in the training set, showing a moderate imbalance of approximately 73% good pallets to 27% bad pallets.

The relationship between each categorical predictor and the response was inspected by means of stacked bar charts segmented by the response, a selection of which is provided in the Annex. These plots reveal at a glance which categories are associated with a higher or lower proportion of good pallets, and were used to anticipate the directions later confirmed by the model. Categories with very few observations, namely fewer than one thousand in the training set, were grouped into an Other category to ensure sufficient representation for stable parameter estimation, and the same groupings were subsequently applied to the test set.

The dependence structure among the predictors was examined in two ways: the Spearman rank correlation was used for the numerical variables and the bias-corrected Cramer's V for the categorical variables. Several strong associations were identified that

informed the modelling, most notably a perfect association between the warehouse and the shipping route, since each warehouse exclusively serves a specific set of routes, a near-perfect association between the warehouse and the product model, reflecting that each warehouse stores specific models, and a near-perfect association between the storage-unit type and the pallet type, since the pallet width determines the storage unit assigned. These redundancies guided the exclusion of variables in the model that is sensitive to multicollinearity. The Cramer's V matrix of the categorical predictors retained in the final model is shown in Fig. 17, confirming that, after the exclusions, no pair of predictors exhibits a substantial association beyond a moderate link between the shipping route and the product model, which reflects a genuine operational pattern rather than redundancy.

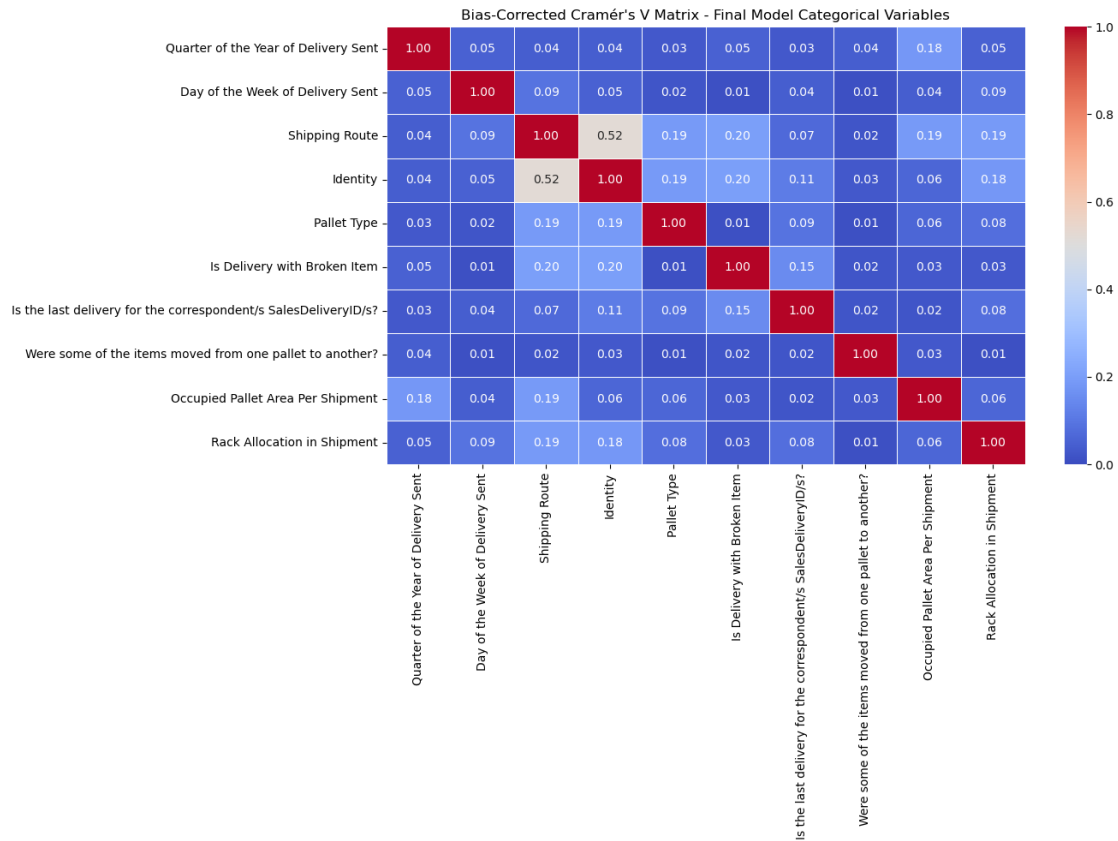


Fig. 17 - Bias-corrected Cramer's V matrix of the categorical predictors retained in the final model.

4.2.4. The Generalized Linear Mixed Model

The core inferential model of the analysis is a GLMM with a binomial response and a logit link, explaining the probability that a pallet is classified as good. A mixed model extends the generalized linear model by including, in addition to the fixed effects common

to all observations, random effects that vary across groups and capture the correlation between observations belonging to the same group [14], [15]. Because pallets belonging to the same shipment are not independent, a random intercept for the shipment was included. For a pallet i belonging to shipment j , the model is

$$\text{logit}\left(P(y_{ij} = 1)\right) = \beta_0 + \sum_k \beta_k x_{kij} + u_j, \quad u_j \sim \mathcal{N}(0, \sigma_u^2)$$

where the x_{kij} are the fixed-effect predictors, the β_k are the corresponding coefficients and u_j is the shipment-specific random intercept, assumed normally distributed with variance σ_u^2 [15]. The need for the random intercept was confirmed by the Intraclass Correlation Coefficient (ICC), which measures the proportion of the total variance attributable to differences between groups. For a logistic model with a latent-variable formulation the ICC is given by

$$\text{ICC} = \frac{\sigma_u^2}{\sigma_u^2 + \pi^2/3},$$

where $\pi^2/3$ is the variance of the standard logistic distribution, which is the residual variance assumed for the latent-variable formulation of the logistic model [15]. The ICC was estimated at 0.268, indicating that approximately twenty-seven per cent of the total variance in pallet utilization is attributable to differences between shipments. This violates the independence assumption of standard logistic regression and confirms that a mixed model is the appropriate framework, since fitting an ordinary logistic regression would understate the uncertainty of the estimates [15].

4.2.5. Assumption checking and diagnostics

Several assumptions were verified before and after fitting. The sample size was assessed in terms of the number of events per variable (EPV), which requires a minimum number of minority-class observations per predictor for stable estimation, and in terms of the number of groups, which must be sufficiently large for the random-effect variance to be reliably estimated, both were comfortably adequate given the very large data set and the more than three thousand shipments [16]. A formal separation-detection procedure confirmed that no predictor perfectly separated the two classes, so that finite maximum-likelihood estimates exist for all coefficients [17]. Multicollinearity among the fixed effects was reassessed using variance inflation factors (VIF). For the multi-level

categorical predictors, the ordinary VIF is mechanically inflated by the number of categories, so the appropriate measure is the adjusted form, namely the generalized variance inflation factor (GVIF). All values were well below the respective cut-offs, confirming that collinearity does not substantially inflate the variance of the estimates.

The linearity of the logit was examined by inspecting the empirical logit of the response against each continuous predictor under the original scale and under candidate transformations, namely the logarithm and the square root, as shown in the Annex. A formal test was deliberately avoided, since the very large sample size would render it uninformative by flagging trivial departures from linearity as statistically significant. As none of the continuous predictors displayed a satisfactorily linear relationship under any of the transformations considered, the two duration variables were removed in the interest of parsimony, given their limited explanatory role. The occupied pallet area per shipment was retained and discretised at a threshold of thirty-three square metres, corresponding to the standard loading area of a tautliner truck, which removes the linearity requirement for that predictor and provides an operationally meaningful cut-off that is independent of the data.

The fitted model was subjected to a range of diagnostics. The distribution of the random-effect predictions was assessed by visual inspection of a normal quantile-quantile plot and a histogram, shown in Fig. 18, which indicated approximate normality with some deviations in the lower tail. A formal normality test was deliberately not applied to the random-effect predictions, because with more than three thousand groups such a test possesses enough statistical power to flag minor and practically irrelevant departures from normality as significant. The diagnosis was therefore based on visual inspection, which is the appropriate approach in this large-sample setting.

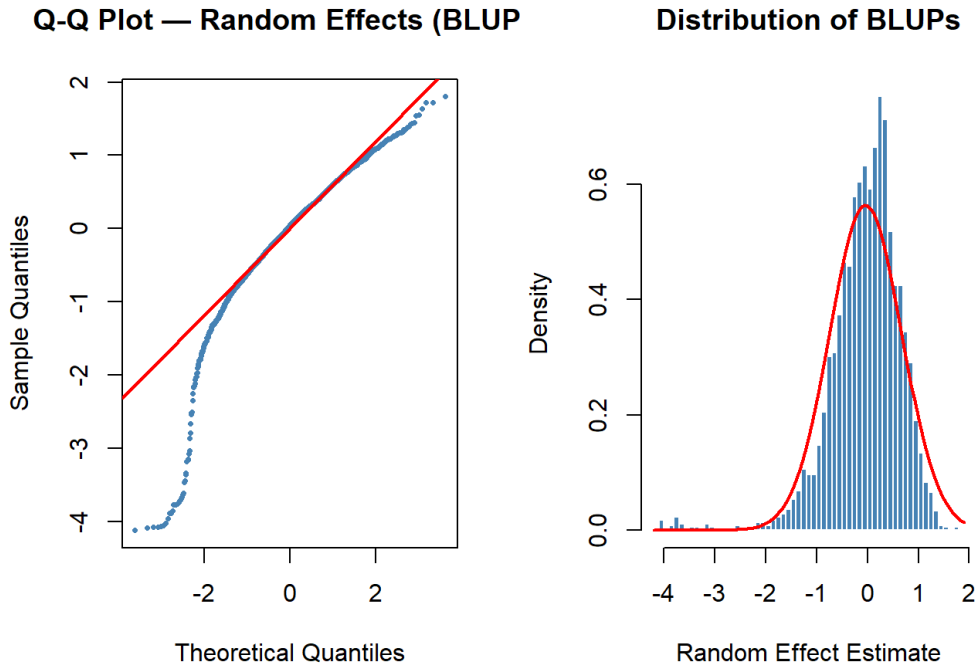


Fig. 18 - Normal quantile-quantile plot and histogram of the random-effect predictions- Best Linear Unbiased Predictors (BLUPs) of the GLMM, used for the visual assessment of normality.

Residual diagnostics based on simulated scaled residuals, shown in Fig. 19, indicated no evidence of over- or under-dispersion and no excess of outliers, supporting the adequacy of the binomial variance assumption [18]. While the Kolmogorov-Smirnov test for uniformity returned a statistically significant deviation, the graphical diagnostics showed only minor departures, consistent with the high sensitivity of formal goodness-of-fit tests when applied to large sample sizes.

A likelihood-ratio test, comparing the full model against the null model containing only the random intercept, confirmed that the fixed effects collectively contributed significantly to the fit ($\chi^2 = 11643.3$, $df = 27$, $p < 0.001$), with both the Akaike and the Bayesian information criteria favouring the full model (AIC dropping from 62520.7 to 50931.4 and BIC from 62538.7 to 51192.1).

The marginal coefficient of determination, which measures the variance explained by the fixed effects alone, was 0.346, while the conditional coefficient of determination, which additionally accounts for the shipment-level random effect, was 0.481 [19]. The difference between the two reflects the additional variance accounted for by the shipment-level random effect, in line with the substantial between-shipment heterogeneity already indicated by the ICC.

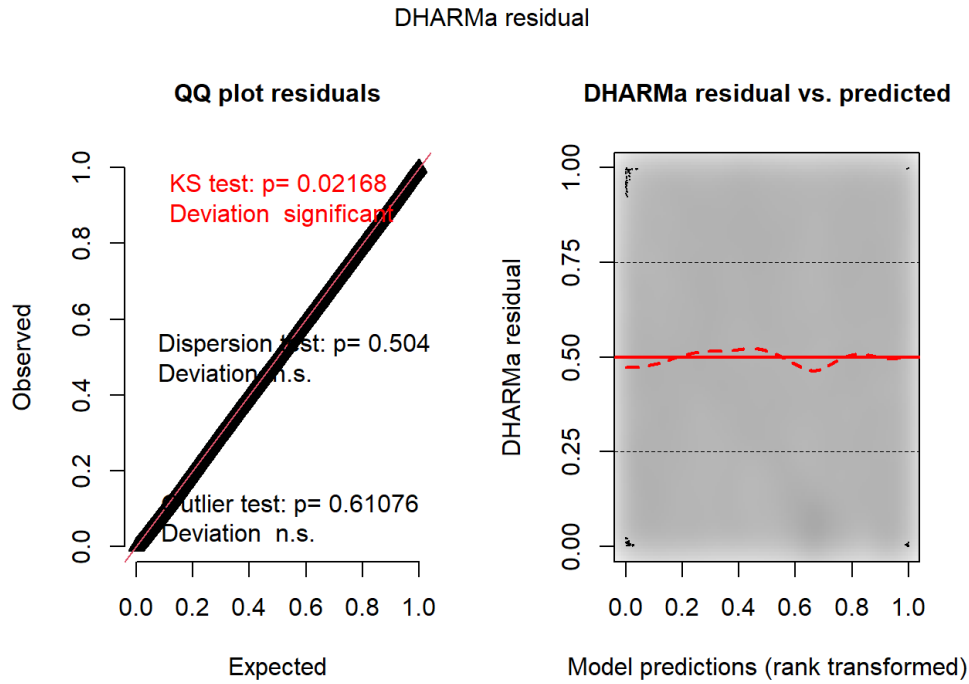


Fig. 19 - Simulation-based residual diagnostics of the GLMM. The left panel shows a quantile-quantile plot of the simulated scaled residuals against the uniform distribution; the right panel shows the residuals against the model predictions.

4.2.6. Odds, odds ratios and their interpretation

Before the results of the model are presented, it is useful to clarify the quantities through which they are expressed. The odds of an event are defined as the ratio of the probability that the event occurs to the probability that it does not,

$$\text{odds} = \frac{p}{1 - p} ,$$

so that a probability of one half corresponds to odds of one, a probability of two thirds to odds of two, and so on. A logistic model is linear in the logarithm of the odds, which is why its coefficients are most naturally expressed on the odds scale. The odds ratio (OR) associated with a predictor is the multiplicative factor by which the odds of the outcome change when the predictor increases by one unit, or, for a categorical predictor, when one moves from the reference category to the category in question, holding all other predictors constant. The odds ratio is obtained by exponentiating the corresponding coefficient, $\text{OR} = \exp(\beta)$, so that an odds ratio greater than one indicates higher odds of

a good pallet relative to the reference, and an odds ratio less than one indicates lower odds.

A crucial point, which is essential for the correct reading of the results that follow, is that a high odds ratio does not mean a proportional increase in probability. The relationship between the odds and the probability is non-linear, and the change in probability that corresponds to a given odds ratio depends on the baseline probability: the same odds ratio produces a large change in probability when the baseline probability is near one half, but only a small change when the baseline probability is already close to zero or to one. The odds ratio is therefore a scale-invariant measure of association that is constant across observations, but it is not directly a statement about probabilities. Precisely because of this distinction, marginal-effects plots were produced to translate the odds ratios onto the probability scale, with the aim of communicating the results in a more intuitive way to a less technical audience, since predicted probabilities are easier to relate to than odds ratios. Each effect is expressed as a predicted probability of a good pallet averaged over the observed distribution of all other predictors, a representative selection of these plots is provided in the Annex.

4.2.7. Model interpretation

The complete set of fixed-effect odds ratios, with their ninety-five per cent confidence intervals and p -values, is reported in Table 6, relative to the reference categories AURAPLUS2015 for the product model, the first quarter, Monday, the 510-PLDE01 route, the eighty-centimetre pallet width, and an occupied area at or below thirty-three square metres.

Table 6 - Complete fixed-effect odds ratios of the GLMM, with 95% confidence intervals and p -values, relative to the respective reference categories.

Predictor (vs reference)	OR	95% CI	p-value
Quarter Q2 (vs Q1)	1.29	[1.16, 1.45]	<0.001
Quarter Q3 (vs Q1)	1.23	[1.10, 1.38]	<0.001
Quarter Q4 (vs Q1)	1.10	[0.98, 1.23]	0.104
Friday (vs Monday)	1.23	[1.09, 1.40]	<0.001
Saturday (vs Monday)	1.08	[0.85, 1.37]	0.520
Thursday (vs Monday)	1.04	[0.92, 1.17]	0.580
Tuesday (vs Monday)	1.11	[0.98, 1.25]	0.112

Wednesday (vs Monday)	0.98	[0.86, 1.10]	0.703
Route 510-PLIE01	0.79	[0.66, 0.94]	0.007
Route 510-PLPL01	0.28	[0.25, 0.32]	<0.001
Route 512-LTDE01	0.37	[0.28, 0.50]	<0.001
Route 512-LTPL01	0.13	[0.10, 0.18]	<0.001
More than one product model	3.13	[2.48, 3.96]	<0.001
AURA2015	0.85	[0.74, 0.97]	0.015
CLASSIC_ALU_2	0.97	[0.81, 1.15]	0.710
FORMA2015	0.71	[0.61, 0.83]	<0.001
FORMAPLUS2015	1.12	[1.01, 1.24]	0.036
Other model	0.59	[0.47, 0.75]	<0.001
RIBO_ALU_2	0.50	[0.47, 0.54]	<0.001
V200	0.72	[0.53, 0.96]	0.028
V200E	0.82	[0.62, 1.08]	0.157
Pallet width 120 cm (vs 80 cm)	6.80	[6.45, 7.16]	<0.001
Broken item detected	1.68	[1.53, 1.83]	<0.001
Last delivery of order	0.24	[0.23, 0.25]	<0.001
Items reallocated	0.55	[0.47, 0.65]	<0.001
Occupied area > 33 m2	0.74	[0.65, 0.84]	<0.001
Rack allocation	0.37	[0.32, 0.43]	<0.001

The strongest positive effect was that of the pallet width, pallets of one hundred and twenty centimetres had almost seven times higher odds of being good than the narrower pallets. This effect is not as straightforward as it may seem at first, because the relationship between pallet width and the number of items carried is not direct: wider pallets are not used arbitrarily but are typically reserved for the transport of larger items, while the narrower eighty-centimetre pallets are typically used for smaller items, so part of the association reflects the type of product placed on each pallet rather than the dimension of the pallet alone. Despite this nuance, the finding is highly relevant from an operational standpoint, since it confirms an observation that the company was already considering, namely that it is in general easier to obtain a well-utilised pallet when one uses one-hundred-and-twenty-centimetre pallets rather than the eighty-centimetre ones, which supports the option, already under discussion, of transitioning to the exclusive use of pallets one hundred and twenty centimetres wide.

The last delivery of an order had markedly lower odds of being a good pallet, with an OR of 0.24, which is operationally expected, as the final pallet of an order typically carries only the residual items that remain after the bulk of the order has already been shipped, and these are fewer in number and less likely to meet the minimum thresholds.

The five shipping routes correspond to pairs of origin warehouse and destination country in Central Europe. Two of them terminate in Poland, namely the routes labelled 510-PLPL01 and 512-LTPL01, while the reference route 510-PLDE01 terminates in Germany. Pallets shipped on the routes to Poland faced the strongest penalties, with odds ratios of 0.28 and 0.13 for the two warehouses, suggesting that orders destined for that market are systematically more fragmented regardless of the warehouse of origin.

The product model carried by the pallet, encoded in the variable Identity, had a substantial and varied effect, all relative to the reference category AURAPLUS2015. The categories of this variable are internal product codes of the client, denoted by labels such as AURAPLUS2015, FORMA2015 or RIBO_ALU_2, each corresponding to a specific window or door model with its own physical characteristics. The strongest positive contrast was observed when more than one product model was placed on the same pallet, with an OR of 3.13, reflecting the role of consolidation in filling pallets and pointing in the same direction as the other consolidation-related findings. Among the single-model categories, FORMAPLUS2015 had slightly higher odds of being good, with an OR of 1.12, while the remaining single-model categories had odds ratios below one, ranging from CLASSIC_ALU_2, which did not differ significantly from the reference at 0.97, down to RIBO_ALU_2 at 0.50 and the residual Other category at 0.59, both with very small p -values. The variability of these effects indicates that the product model is genuinely informative about pallet utilization, plausibly because different product models have different physical characteristics, such as size and weight, that constrain how many of them fit on a single pallet and how readily they can be combined with other items.

Two further effects with substantial magnitude and clear operational meaning are the reallocation of items between pallets and the occupied pallet area within the shipment. Pallets whose items had been moved from another pallet at some point during the process had lower odds of being good, with an OR of 0.55. This is consistent with the operational picture: items reallocated to a pallet typically correspond to a residual portion of an order that did not fit elsewhere or that had to be split, so the resulting pallet ends up carrying few items and is less likely to meet the minimum thresholds. The occupied pallet area within the shipment, discretised at the threshold of thirty-three square metres that corresponds to the standard tautliner loading area, had an OR of 0.74 for the higher

level, meaning that pallets travelling in shipments that fill the truck above this threshold are less likely to be good pallets than those travelling in less heavily loaded shipments. This effect, while seemingly counter-intuitive, is consistent with shipments that approach the full capacity of the truck having to accommodate additional small pallets to use up the remaining space, rather than being entirely composed of well-utilised pallets.

Two further effects deserve a cautious interpretation. The detection of a damaged item before dispatch was, counter-intuitively, associated with higher odds of a good pallet, with an OR of 1.68, which may reflect that more heavily used pallets are handled more often and are therefore more exposed to damage, or that the replacement process introduces additional time during which more items can be incorporated into the rectified pallet. The presence of racks in the shipment was associated with lower odds of a good pallet, with an OR of 0.37, the precise operational mechanism is not fully clear, and this association should be reported with the caveat that a complete interpretation would require additional domain information, since racks alter the physical organisation of pallets within the truck and may be associated with specific product types or pallet sizes that are inherently harder to fill efficiently. The temporal effects were comparatively mild: the second and third quarters showed modestly higher odds than the first, and Friday showed higher odds than Monday, while the remaining days did not differ significantly.

4.2.8. Model validation

Although the objective of the model was explanatory rather than predictive, its out-of-sample behaviour was assessed in order to confirm that the fixed-effect structure generalises beyond the training data and that the model is not severely overfit. A five-fold cross-validation at the shipment level, in which complete shipments were allocated to each fold, yielded an area under the receiver operating characteristic curve of 0.796 with a small standard deviation across folds, and the held-out test set yielded an area under the curve of 0.793, with a sensitivity of 0.933 and a specificity of 0.406 at the default classification threshold. The lower specificity reflects the class imbalance and the fact that predictions for unseen shipments rely on the fixed effects only, since the shipment-specific random effects are not estimable for new shipments, this is consistent with the gap between the conditional and the marginal coefficients of determination. The close agreement between the cross-validation and the test estimates confirms that the reported performance is stable and not an artefact of a particular partition. The receiver operating characteristic curve of the model on the test set is provided in the Annex.

The principal limitation of the model is that the data have a four-level hierarchical structure, with pallets nested within shipments, which are in turn nested within shipping routes and warehouses, whereas only the shipment level was modelled through a random intercept. The route and warehouse levels have too few groups, namely five routes and two warehouses, for a reliable estimation of a random-effect variance, and were instead represented through fixed effects [20]. This is methodologically defensible, but it implies that any residual variation attributable to systematic differences between routes or warehouses that is not captured by the fixed effects remains unaccounted for. Furthermore, the reported coefficients quantify associations under the assumptions of the model and do not establish causal relationships, since unmeasured confounders at the route or warehouse level may influence both the predictors and the outcome.

4.2.9. Ensemble tree-based models

To complement the inferential model, two ensemble tree-based models were fitted to the same binary response, namely a Random Forest [21] and a gradient-boosting model in its XGBoost implementation [22]. These models rely on far fewer assumptions than the GLMM. They capture non-linear relationships and interactions automatically and admit a broader feature set, including variables that had to be excluded from the GLMM due to multicollinearity or to violations of the linearity of the logit. A Random Forest builds a large collection of decision trees on bootstrap samples of the data, each tree using a random subset of the features at every split, and aggregates their predictions by averaging, thereby reducing the variance of a single tree without increasing its bias. Gradient boosting instead builds the trees sequentially, each new tree being fitted to the residual errors of the current ensemble, which tends to reduce bias and often yields higher predictive accuracy at the cost of a greater risk of overfitting if not properly regularised.

The evaluation strategy mirrored that of the GLMM, with a five-fold cross-validation at the shipment level and a final evaluation on the held-out test set, so that the three models are directly comparable. Their performance on the test set is summarised in Table 7. Both ensemble models outperformed the GLMM both in discrimination (AUC-ROC) and in the overall quality of the probabilistic predictions (Brier score), which is expected given that they are not constrained by the assumptions of the GLMM, can capture interactions without explicit specification, and were trained on a richer feature set. The two ensemble models occupy different operating points along the sensitivity-specificity trade-off: XGBoost achieved the highest area under the curve and a markedly higher specificity,

making it better suited to contexts where the identification of bad pallets is prioritised, while the Random Forest achieved the highest sensitivity, at the cost of a lower specificity. The corresponding receiver operating characteristic curves are shown in Fig. 20.

Table 7 - Predictive performance of the three models on the held-out test set.

Model	AUC-ROC	Brier score	Sensitivity	Specificity
GLMM	0.793	0.150	0.933	0.406
Random Forest	0.867	0.133	0.977	0.399
XGBoost	0.883	0.113	0.944	0.577

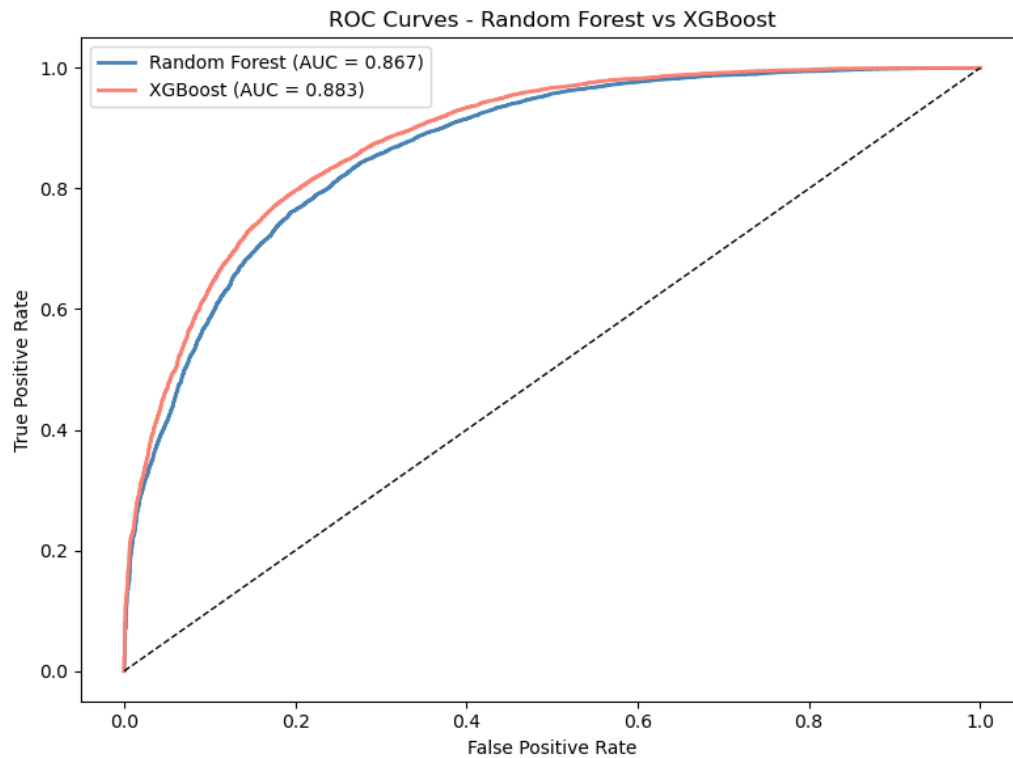


Fig. 20 - Receiver operating characteristic curves of the Random Forest and the XGBoost models on the held-out test set.

The models were interpreted using SHapley Additive exPlanations, a unified, game-theoretic approach that attributes to each predictor its contribution to each individual prediction, in such a way that the contributions of all predictors sum to the difference between the prediction and a baseline [23]. A positive value indicates that the predictor increased the predicted probability of a good pallet for that observation, and a negative value the opposite. It should be noted that the tree explainer returns values on the

probability scale for the Random Forest and on the log-odds scale for XGBoost, so that the magnitudes are not directly comparable between the two models and only the rankings and directions of the effects should be interpreted across models. The most important predictors were consistent across the two models, the strong alignment of their importance rankings being confirmed by a Spearman rank correlation of 0.882 between the aggregated rankings. The pallet type and the last-delivery indicator, which were among the strongest fixed effects in the GLMM, also appeared consistently among the most important predictors, providing a reassuring cross-method confirmation. The SHAP summary for the XGBoost model is shown in Fig. 21, and the aggregated importance rankings of both models are provided in the Annex.

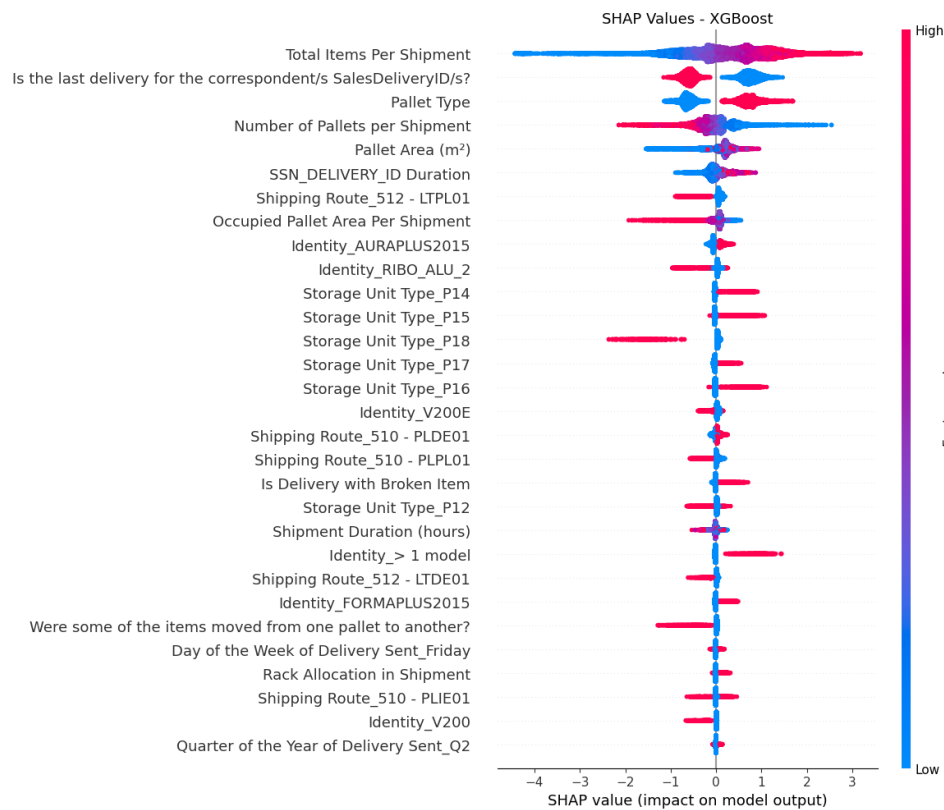


Fig. 21 - SHAP summary plot for the XGBoost model on the held-out test set. Each point is an observation, the colour encodes the value of the predictor and the horizontal position its contribution to the predicted log-odds of a good pallet.

Beyond confirming the findings of the GLMM, the ensemble models provided additional insight that the inferential model could not, because they could accommodate variables that had been excluded from the GLMM due to severe multicollinearity. The total number of items per shipment and the number of pallets per shipment, in particular, emerged as highly important and with opposite directions: a larger number of items per shipment was associated with a higher probability of a good pallet, whereas a larger number of pallets per shipment was associated with a lower probability. Taken together, these two

variables capture the item-density structure of a shipment, a dimension that is operationally coherent, since a shipment that spreads many items across few pallets produces well-used pallets, whereas one that spreads few items across many pallets produces poorly used ones.

The storage unit types also emerged as important predictors in the ensemble models, which was particularly informative because this variable had been excluded from the GLMM owing to its near-perfect association with the pallet type and could therefore not be studied within the inferential framework. The SHAP values revealed that the contribution of the storage unit type is highly heterogeneous across its levels, several storage unit types showed a clear positive association with the probability of a good pallet, while others showed an equally clear negative association, indicating that the physical storage location at which a pallet is held until dispatch carries genuine information about its eventual utilization. This finding complements the inferential analysis by surfacing structure that the assumptions of the GLMM had forced out of its scope, and it is one of the clearest illustrations of the methodological complementarity between the two paradigms.

4.2.10. Comparison of the modelling paradigms

The two modelling paradigms are best understood as complementary rather than competing. The GLMM provides interpretable, inferentially grounded coefficient estimates with associated uncertainty, directly supporting the explanatory objective of identifying which operational variables contribute to pallet utilization and in which direction, and quantifying the confidence with which each association is established. The ensemble models sacrifice these inferential properties in exchange for superior predictive performance and the ability to capture complex, non-linear structure, with SHAP values offering a post-hoc approximation of variable importance that complements, but does not replace, the formal inference of the GLMM. The GLMM characterises the direction and statistical significance of each predictor under explicit modelling assumptions, while the ensemble models confirm these findings in a non-parametric framework and additionally reveal variables and patterns that the GLMM could not include.

Several limitations specific to the ensemble models should be acknowledged. They do not account for the hierarchical structure of the data, treating pallets within a shipment as independent, which may lead to an underestimation of the prediction uncertainty. The SHAP values, while locally accurate and globally consistent, do not carry the inferential

guarantees of regression coefficients and provide neither standard errors nor p -values, so they should not be read as measures of statistical significance. Finally, the models were trained with conservative hyperparameters without systematic tuning, so that optimised hyperparameters might yield somewhat improved predictive performance, although the explanatory conclusions are unlikely to change substantially. Taken together, the two approaches converge on a consistent and operationally coherent picture of the drivers of pallet utilization, which was the goal of this phase.

5. Conclusion and Future Work

5.1. Conclusion

This report described the three activities developed during the curricular internship hosted by Breakawai, all of which are situated within the field of process mining and operate on data derived from event logs, while addressing distinct problems and serving distinct purposes.

The first activity, an introductory proof of concept, applied PCA and clustering to the segmentation of e-mail threads, producing four interpretable behavioural profiles whose stability was confirmed by the agreement of two clustering algorithms with different geometric assumptions. The second activity delivered a general-purpose interactive tool to support RCA on process-mining databases, generalising the WMD indicator to the multivariate case and combining it with non-parametric statistical tests and an interpretable decision tree. The third and most extensive activity, carried out for an external client, studied a logistics process first by quantifying the optimisation potential of pallet consolidation, with an estimated annual reduction of approximately three hundred and thirty-four trucks and one hundred and eighty-eight tonnes of CO₂ equivalent, and subsequently by identifying the operational drivers of pallet utilization through a GLMM complemented by ensemble tree-based models.

Across the three activities, a consistent methodological posture was maintained. The choice of each technique was guided by the structure of the data and the nature of the question, and explicitly justified. The preference for rank-based and non-parametric procedures in the presence of skewed process variables, the careful separation of training and test data at the appropriate hierarchical level to prevent leakage, the modelling of dependence structures via random effects where appropriate, and the deliberate pairing of an interpretable inferential model with more flexible models complemented by SHAP-based interpretation, all illustrate this posture.

Several limitations were encountered over the course of the internship and should be acknowledged. The first activity was exploratory and was not deployed, so that the practical value of the resulting segmentation was not validated in a production setting, and its inclusion of binary variables within an analysis designed for continuous data is a known simplification. The sample size of approximately four hundred and sixty threads is also modest in relation to the ten behavioural indicators considered, which constrains the robustness of any inference about the resulting profiles and limits the granularity at

which the segmentation can be examined. The Root Cause Analysis tool, although functional, was developed under time constraints and would benefit from more extensive validation on a wider range of event logs. In the logistics analysis, the estimates of the first phase depend on operational assumptions, most notably the ceiling on the number of mixed pallets, while the conclusions of the second phase are associational rather than causal and rest on a single year of data, which limits their generalisability. The hierarchical structure of the logistics data was only partially modelled, since the route and warehouse levels had too few groups to support random effects. Despite these limitations, the internship provided an opportunity to apply a broad range of statistical and machine learning techniques to genuine business problems, and to translate analytical results into a form that supports operational decision-making.

5.2. Future Work

Several directions for future work follow directly from the activities described in this report and from the limitations identified above.

For the first activity, the methodological compromise of treating binary variables within a Principal Component Analysis would be more rigorously addressed by Factor Analysis of Mixed Data, which handles categorical and continuous variables on an equal footing. The behavioural segmentation could also be extended to additional mailboxes, in order to enlarge the sample size and to allow an external validation of the four profiles identified.

For the second activity, the most natural next step is to validate the Root Cause Analysis tool with real client data in a production setting, complementing the validation already performed with synthetic event-log data.

For the first phase of the third activity, the consolidation analysis would benefit from acquiring data on the physical dimensions of the items, which would make it possible to explore alternative consolidation scenarios that take into account the spatial feasibility of combining specific items on the same pallet. The integration of the analysis with business intelligence systems could also be explored, so that the consolidation diagnostics can be regularly refreshed and consulted by the operational team.

For the second phase of the third activity, the explanatory analysis would benefit from a systematic exploration of interactions between the most important predictors, either by extending the GLMM with explicit interaction terms or by inspecting SHAP interaction

values computed from the ensemble models, which would allow joint effects to be identified beyond the main effects reported here.

6. References

- [1] W. M. P. van der Aalst, *Process Mining: Data Science in Action*, 2nd ed. Berlin, Germany: Springer, 2016.
- [2] W. M. P. van der Aalst et al., "Process mining manifesto," in *Business Process Management Workshops, Lecture Notes in Business Information Processing*, vol. 99, Berlin, Germany: Springer, 2012, pp. 169–194.
- [3] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*, 2nd ed. New York, NY, USA: Springer, 2009.
- [4] I. T. Jolliffe, *Principal Component Analysis*, 2nd ed. New York, NY, USA: Springer, 2002.
- [5] J. H. Ward, "Hierarchical grouping to optimize an objective function," *Journal of the American Statistical Association*, vol. 58, no. 301, pp. 236–244, 1963.
- [6] P. J. Rousseeuw, "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis," *Journal of Computational and Applied Mathematics*, vol. 20, pp. 53–65, 1987.
- [7] L. Hubert and P. Arabie, "Comparing partitions," *Journal of Classification*, vol. 2, no. 1, pp. 193–218, 1985.
- [8] H. Cramér, *Mathematical Methods of Statistics*. Princeton, NJ, USA: Princeton University Press, 1946.
- [9] W. Bergsma, "A bias correction for Cramér's V and Tschuprow's T," *Journal of the Korean Statistical Society*, vol. 42, no. 3, pp. 323–328, 2013.
- [10] H. B. Mann and D. R. Whitney, "On a test of whether one of two random variables is stochastically larger than the other," *The Annals of Mathematical Statistics*, vol. 18, no. 1, pp. 50–60, 1947.
- [11] B. Efron and R. J. Tibshirani, *An Introduction to the Bootstrap*. New York, NY, USA: Chapman and Hall, 1993.
- [12] Smart Freight Centre, *Global Logistics Emissions Council (GLEC) Framework*, version 3.0. Amsterdam, The Netherlands: Smart Freight Centre, 2023.
- [13] International Organization for Standardization, *ISO 14083:2023 Greenhouse Gases — Quantification and Reporting of Greenhouse Gas Emissions Arising from Transport Chain Operations*. Geneva, Switzerland: ISO, 2023.
- [14] D. Bates, M. Mächler, B. Bolker, and S. Walker, "Fitting linear mixed-effects models using lme4," *Journal of Statistical Software*, vol. 67, no. 1, pp. 1–48, 2015.
- [15] A. Gelman and J. Hill, *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge, U.K.: Cambridge University Press, 2007.

- [16] P. Peduzzi, J. Concato, E. Kemper, T. R. Holford, and A. R. Feinstein, "A simulation study of the number of events per variable in logistic regression analysis," *Journal of Clinical Epidemiology*, vol. 49, no. 12, pp. 1373–1379, 1996.
- [17] D. W. Hosmer, S. Lemeshow, and R. X. Sturdivant, *Applied Logistic Regression*, 3rd ed. Hoboken, NJ, USA: Wiley, 2013.
- [18] F. Hartig, DHARMA: Residual Diagnostics for Hierarchical (Multi-Level / Mixed) Regression Models, R package version 0.4.6, 2022. [Online]. Available: <https://cran.r-project.org/package=DHARMA>
- [19] S. Nakagawa and H. Schielzeth, "A general and simple method for obtaining R^2 from generalized linear mixed-effects models," *Methods in Ecology and Evolution*, vol. 4, no. 2, pp. 133–142, 2013.
- [20] C. J. M. Maas and J. J. Hox, "Sufficient sample sizes for multilevel modeling," *Methodology*, vol. 1, no. 3, pp. 86–92, 2005.
- [21] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.
- [22] T. Chen and C. Guestrin, "XGBoost: a scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*, 2016, pp. 785–794.
- [23] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems*, vol. 30, 2017, pp. 4765–4774.

7. Annex

Annex A - Supplementary figures and tables for Chapter 2

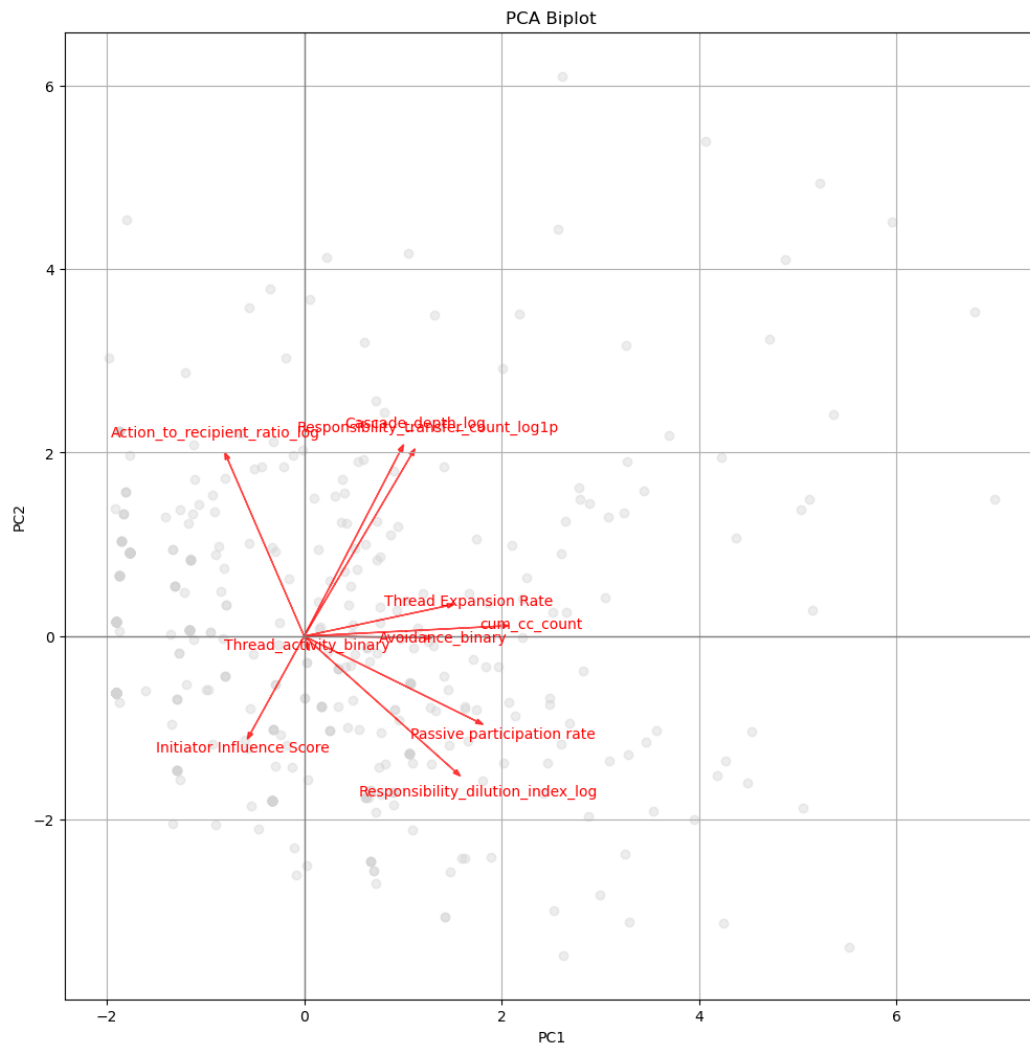


Fig. 22 - Principal-component biplot of the threads on the first two components, with the loading vectors of the original variables.

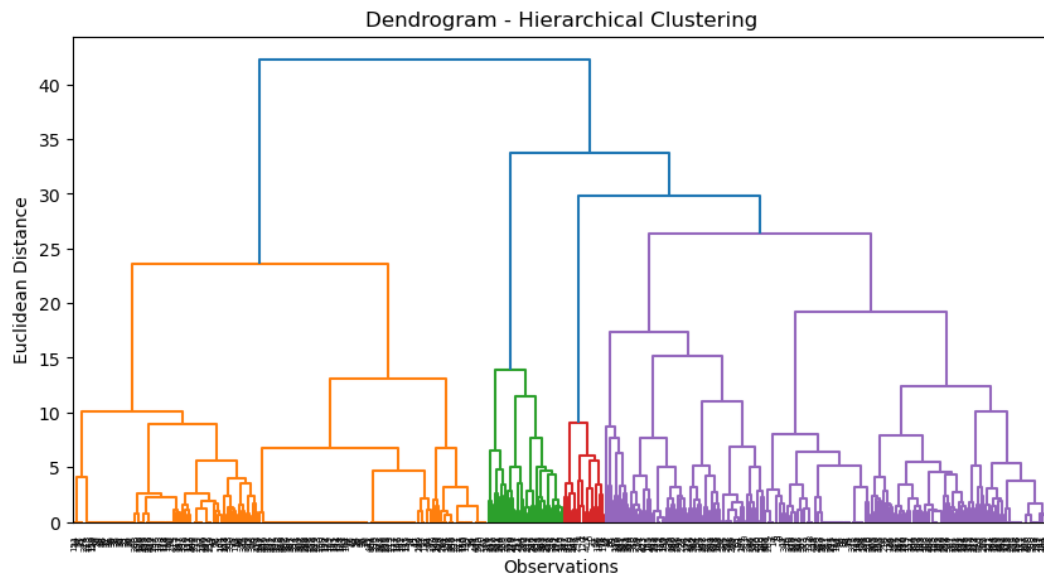


Fig. 23 - Dendrogram of the agglomerative hierarchical clustering with Ward linkage.

K-means Clustering in 3D PCA Space

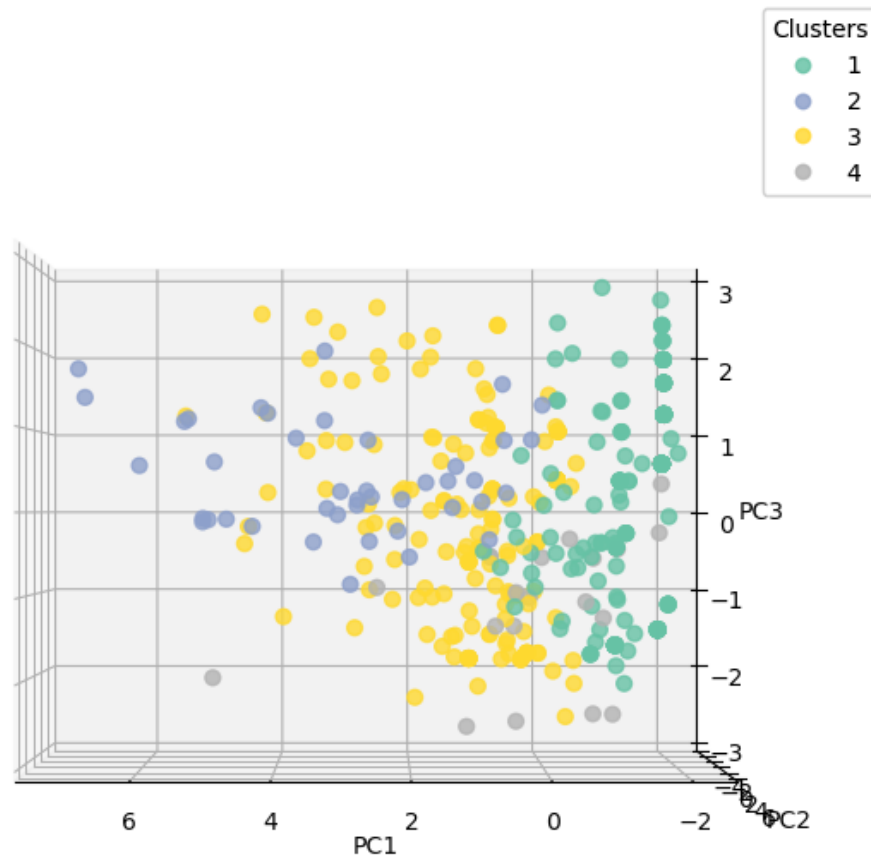


Fig. 24 - Three-dimensional projection of the threads onto the first three principal components, coloured by the four-cluster k-means solution.

Table 8 - Principal-component loadings: contribution of each variable to each principal component.

Variable	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10
Cumulative Count	0.50	0.03	0.01	0.08	-0.21	0.15	0.18	0.63	-0.50	-0.07
Thread Activity (binary)	0.01	-0.02	-0.18	0.90	0.34	-0.18	0.07	0.01	-0.01	0.01
Responsibility Dilution Index (log)	0.38	-0.37	0.16	0.11	-0.19	-0.14	-0.58	-0.31	-0.11	-0.42
Passive Participation Rate	0.44	-0.23	0.00	0.06	-0.32	-0.07	0.64	-0.35	0.33	0.06
Thread Expansion Rate	0.36	0.08	0.13	-0.02	0.52	0.72	-0.04	-0.21	0.08	0.00
Responsibility Transfer Count (log1p)	0.27	0.49	-0.29	0.00	-0.11	-0.06	-0.22	0.26	0.62	-0.28
Action-to-Recipient Ratio (log)	-0.19	0.48	0.46	0.11	-0.07	-0.01	0.30	-0.20	-0.17	-0.59
Initiator Influence Score	-0.14	-0.26	0.69	0.19	-0.08	0.10	-0.07	0.41	0.44	0.10
Avoidance (binary)	0.31	0.00	0.24	-0.31	0.60	-0.61	0.11	0.11	0.01	-0.01
Cascade Depth (log)	0.24	0.51	0.30	0.14	-0.21	-0.14	-0.24	-0.23	-0.13	0.62

Annex B - Supplementary figures for Chapter 4, Phase 1

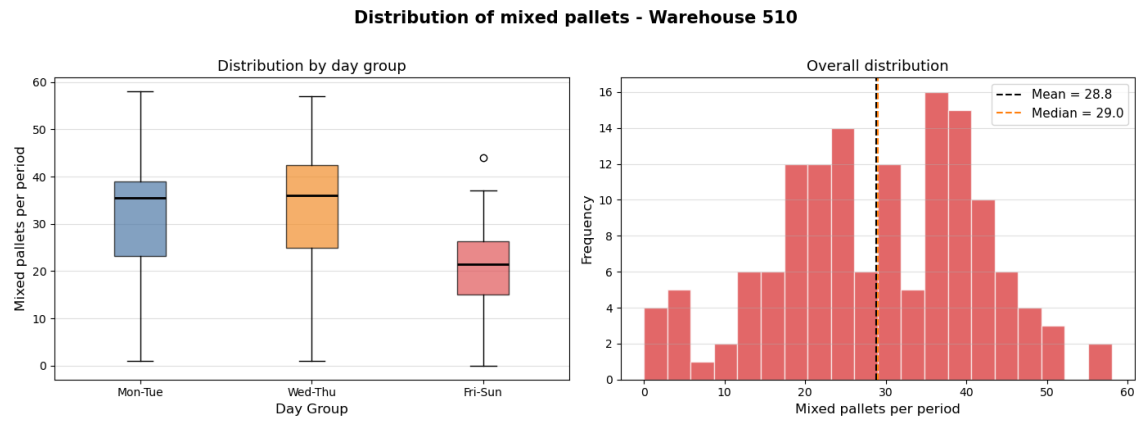


Fig. 25 - Distribution of the mean number of small non-mixed pallets per period for one warehouse.

Annex C - Supplementary figures for Chapter 4, Phase 2

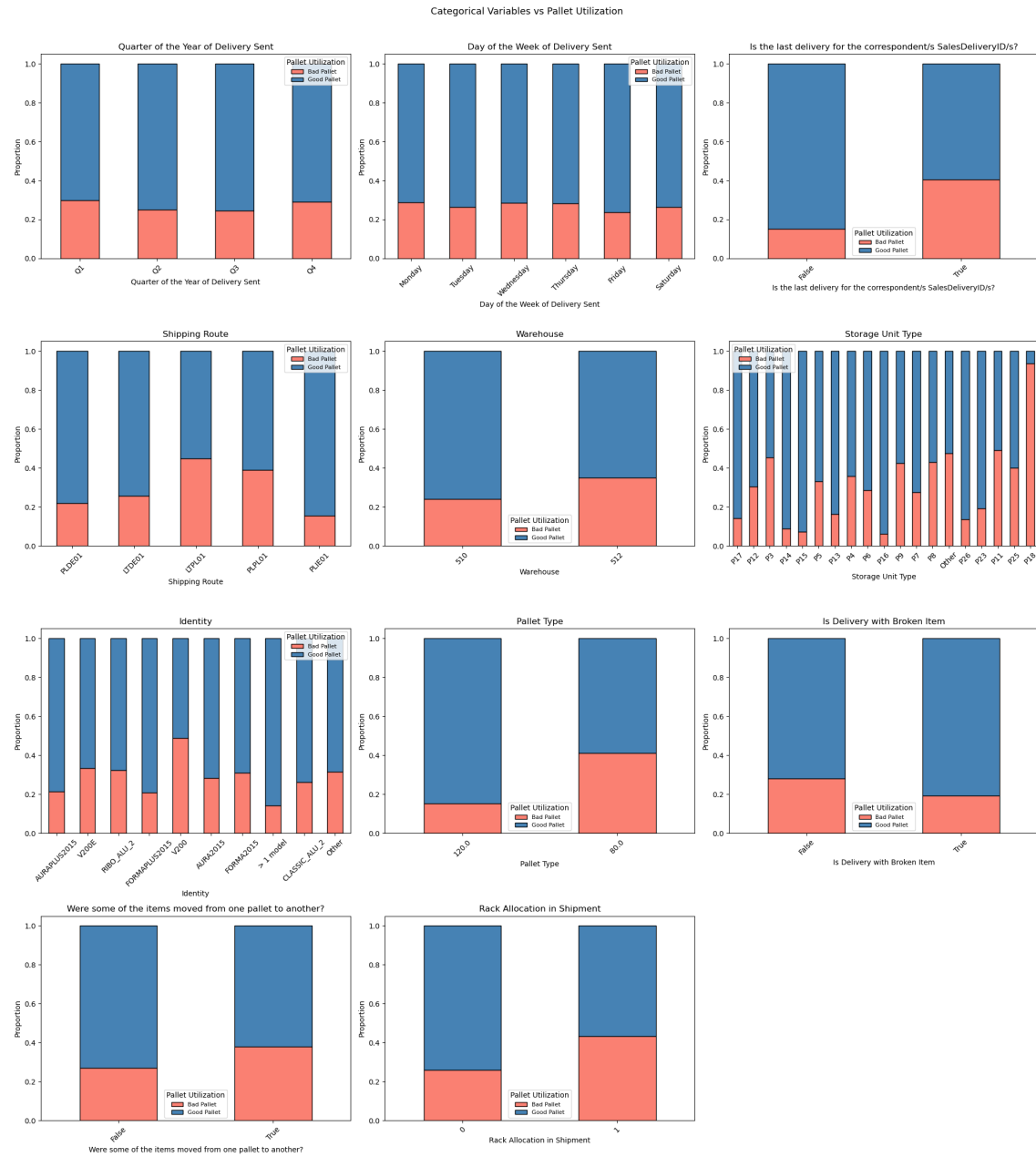


Fig. 26 - Proportion of good and bad pallets within the categories of selected categorical predictors, in the training set.

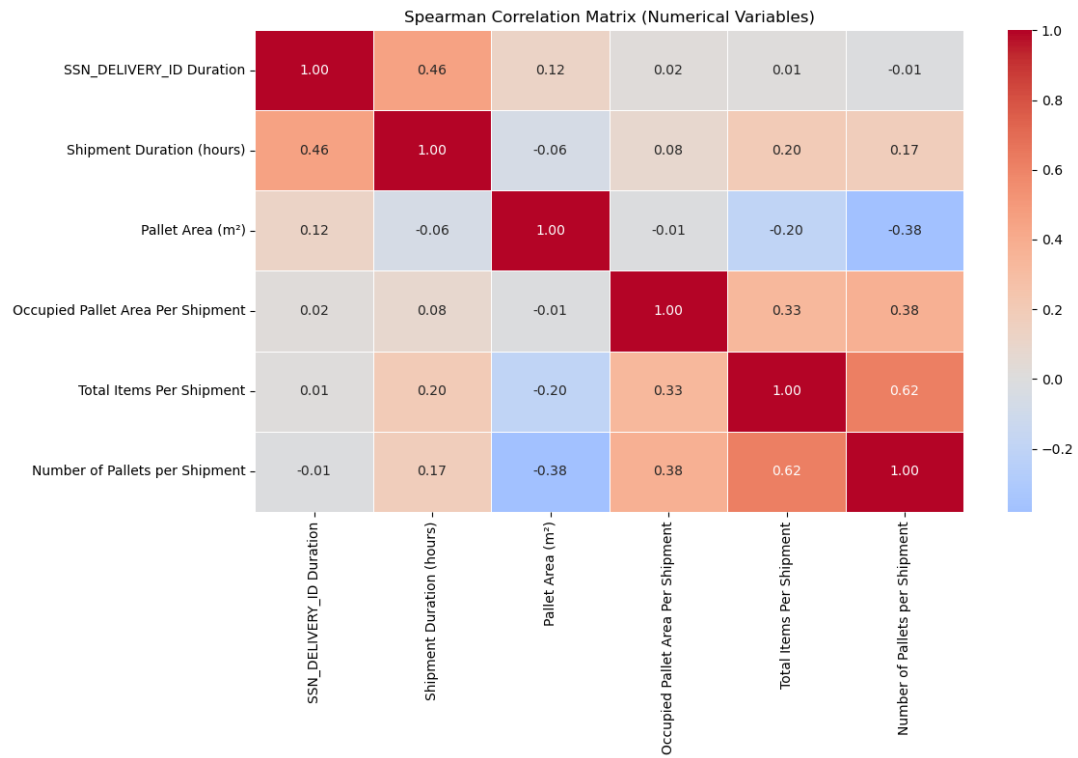


Fig. 27 - Spearman rank correlation matrix of the numerical predictors.

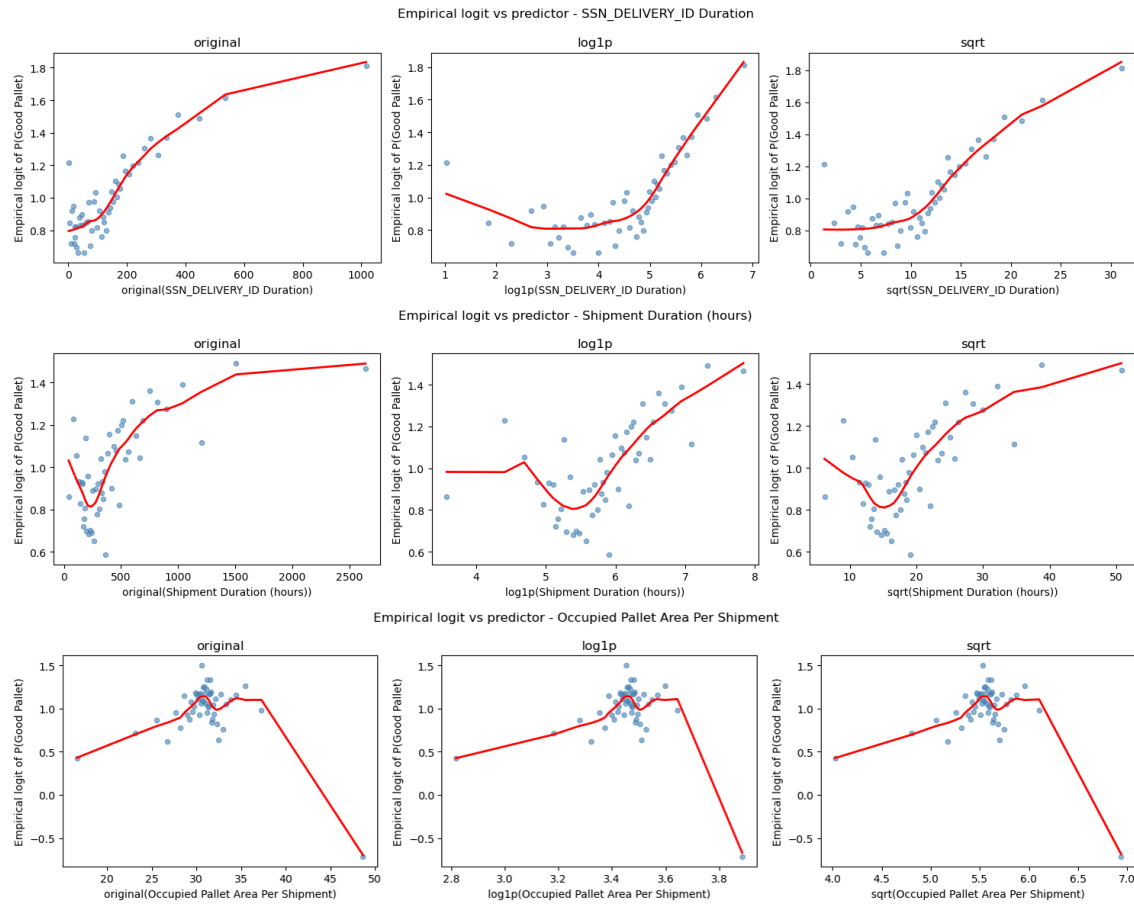


Fig. 28 - Empirical logit of the response against the three retained continuous predictors under their original, logarithmic and square-root scales, used for the visual assessment of the linearity of the logit.



Fig. 29 - Selected marginal effects from the GLMM, expressed as the predicted probability of a good pallet with 95% confidence intervals, translating the odds ratios onto the probability scale.

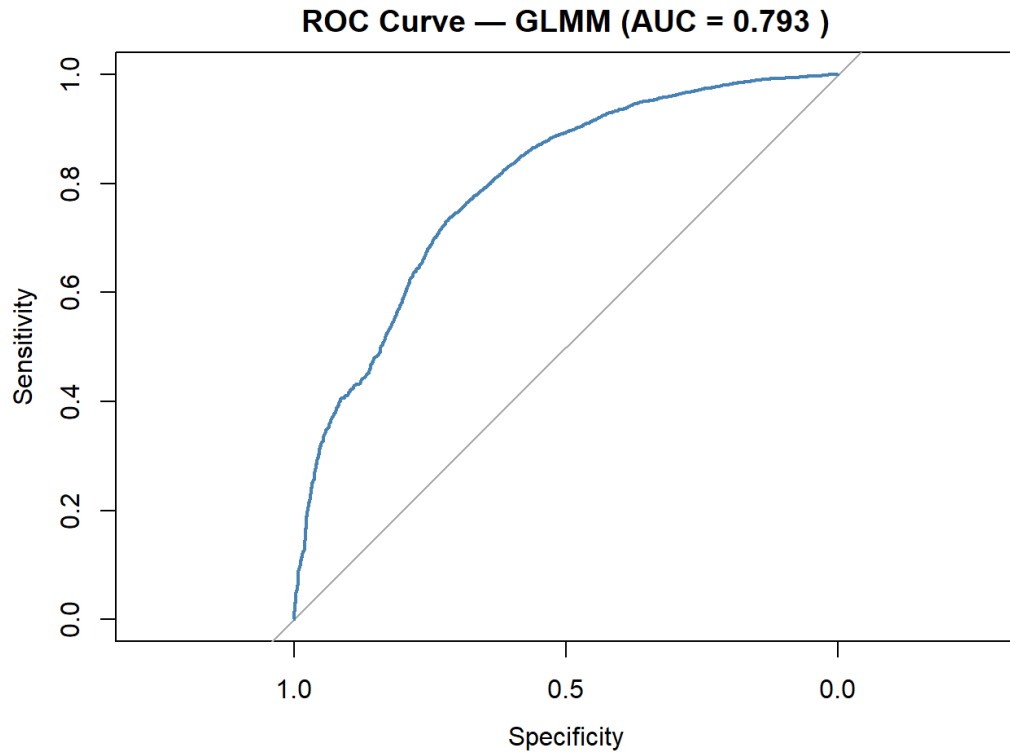


Fig. 30 - Receiver operating characteristic curve of the GLMM on the held-out test set.

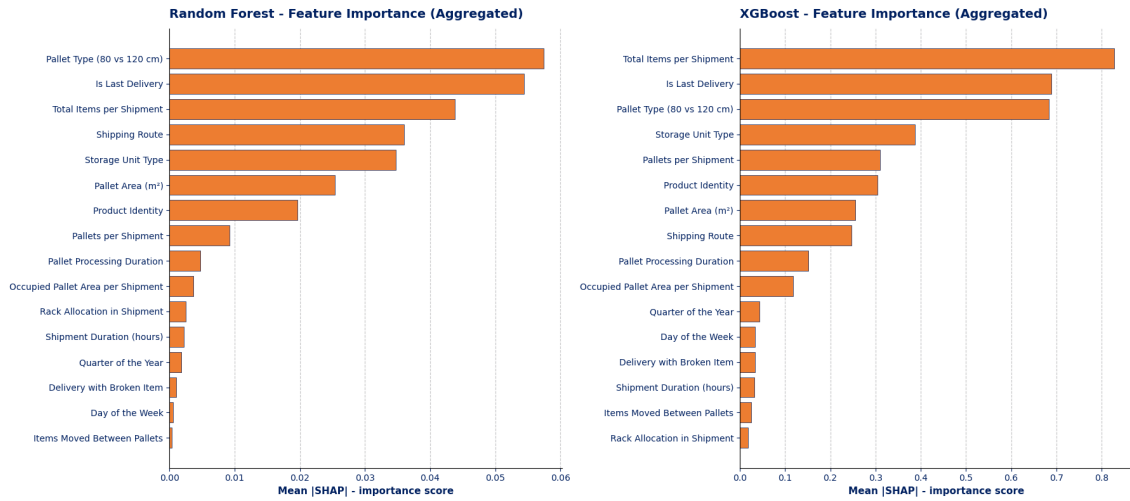


Fig. 31 - Aggregated SHAP feature-importance rankings for the Random Forest and the XGBoost models.

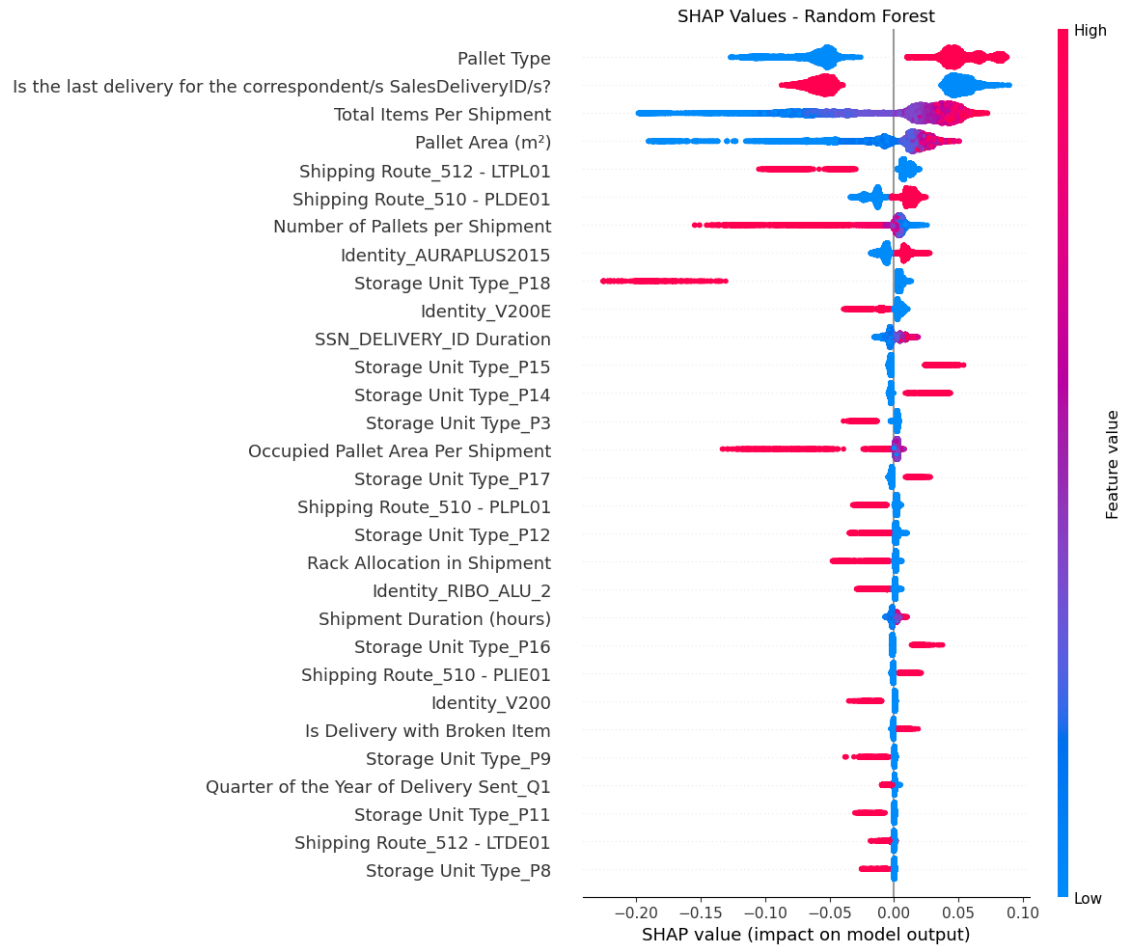


Fig. 32 - SHAP summary plot for the Random Forest model on the held-out test set.