



UNIVERSITÀ
DI PAVIA

FACULTY OF UNIVERSITY OF PAVIA
Department of Economics and Business Sciences

MASTER'S DEGREE IN FINANCE

MASTER THESIS

**ANOMALY DETECTION ON POS
TRANSACTIONS**

Candidate: Alessandro Cammarata

Supervisor: Paolo Giudici

Academic Year 2025-2026

Contents

Abstract	7
1 Introduction	9
1.1 Problem Statement	10
1.2 Research Objectives	11
1.3 Significance of the Study	12
2 Literature review	13
2.1 Anomaly Detection	13
2.2 Fraud Detection with Machine Learning	14
2.3 Peer Group Analysis	15
2.4 Tax Evasion Detection Methods	15
2.5 Pseudo-Labeling	16
2.6 Conclusion	16
3 Methodology	18
3.1 Data Representation and Pre-processing	18
3.1.1 Standardization (z-score)	18
3.1.2 Target Encoding	19
3.1.3 Logarithmic Transformations	19
3.2 Clustering	20
3.2.1 K-Means	20
3.3 Peer Deviation for Anomaly Detection	21
3.4 Supervised Classification	22
3.4.1 Logistic Regression	23
3.4.2 Decision Trees	23
3.4.3 Gradient Boosting and LightGBM	24
3.4.4 Random Forest	24
3.5 Classifier Evaluation Metrics	25
3.5.1 Confusion Matrix	25
3.5.2 Precision, Recall and F1	25

3.5.3	ROC Curve and AUC	25
3.5.4	Cross-Validation	26
3.6	The Overall Pipeline	26
4	Experimental Setup	28
4.1	Dataset Description	28
4.1.1	Train/Test Split	31
4.2	Data Processing and feature Engineering	31
4.2.1	Merchant-level features	31
4.2.2	Encoding of categorical variables	32
4.2.3	Selection of the analysis population	32
4.3	Clustering Configuration	33
4.3.1	Clustering features	33
4.3.2	Transformations	34
4.3.3	Choice of the number of clusters	34
4.3.4	Behavioral families	36
4.4	Anomaly Score Definition	36
4.4.1	Flagging threshold	37
4.4.2	Application to the test set	37
4.5	Classifier Setup	37
4.5.1	Features	38
4.5.2	Models	38
4.5.3	Evaluation of the models	39
4.6	Internal Consistency Checks	39
5	Results and Experiments	40
5.1	Clustering Results	40
5.1.1	The three behavioral families	40
5.1.2	Geographic and sectoral composition	41
5.1.3	Cluster quality	43
5.2	Anomaly Detection Results	43
5.2.1	Distribution of anomalies across clusters	43
5.2.2	Profile of anomalous merchants	44
5.2.3	Overall profile of the flagged merchants	45
5.3	The Anomaly Scoring System	46
5.3.1	Separation between flagged and normal merchants	46
5.3.2	A graduated measure of risk	47
5.3.3	Extreme cases by family	48
5.4	Classifier Performance	48
5.4.1	Discriminative performance	48
5.4.2	Generalization over time	50

5.4.3	Feature importance	50
5.4.4	Classification at a fixed threshold	51
6	Methodology Validation	52
6.1	Stability and compactness of the clusters	52
6.2	Reproducibility of the flags	54
6.3	Economic Coherence of the flags	55
6.4	Absence of Geographic and Sectoral Bias	56
6.4.1	Geographic Bias	56
6.4.2	Sectoral Bias	56
6.5	Robustness of the threshold definition	57
6.6	Summary of Validation	58
7	Conclusion and Future Work	60
7.1	Summary of the work	60
7.2	Answers to the research questions	61
7.3	Limitations	62
7.4	Future Work	63

List of Figures

3.1	Overview of the methodology as a pipeline. The main stages describe the process from the raw data to the risk score produced by the classifier. The validation stage then assesses the internal consistency of the clustering, anomaly scores, and classifier.	27
4.1	Distribution of merchants by sector (ATECO). Merchants with unknown sector are excluded; the grey bar aggregates the remaining 15 sectors.	30
4.2	Distribution of merchants across the five geographic macro-areas.	30
4.3	Distribution of the cash ratio across merchants.	33
4.4	Effect of the logarithmic transformation on the POS volume. Merchants with zero POS volume are excluded from the plot. .	34
4.5	Silhouette coefficient and Davies-Bouldin index for K from 5 to 15. A higher silhouette and a lower Davies-Bouldin index indicate better-defined clusters. The dashed line marks the selected value $K = 8$	35
5.1	Distribution of the main variables across the eight clusters, grouped by behavioral family (cash-only, B2B, retail). Dashed lines separate the three families.	41
5.2	Distribution of the five macro-areas within each cluster.	42
5.3	Share of the eight most common sectors within each cluster. .	43
5.4	Share of flagged merchants in each cluster, on the training and test sets, grouped by behavioral family. Dashed lines separate the three families.	44

5.5	Distribution of the cash ratio, POS ratio, and cash amount for anomalous (A) and normal (N) merchants, by behavioral family, on the test set. The cash amount is shown on a logarithmic scale. In the retail and B2B families the flagged merchants are distinguished by the composition of their payments; in the cash-only family, where the composition is nearly identical, they are distinguished by the absolute level of cash.	46
5.6	Distribution of the predicted probability of anomaly for normal and flagged merchants on the test set, on a logarithmic count scale. The overlap between the two groups corresponds to the borderline cases.	47
5.7	ROC curves of the three classifiers on the test set. The dashed diagonal corresponds to a random classifier.	49
5.8	Confusion matrix of the LightGBM classifier on the test set, at a threshold flagging the top 5% by predicted probability. Rows are the true classes, columns the predicted ones.	51
6.1	Median silhouette coefficient of each cluster, grouped by behavioral family. All values are positive; the dotted line marks the overall silhouette of the partition. Dashed lines separate the three families.	53
6.2	Distribution of flagging frequencies across 200 bootstrap samples, restricted to the merchants flagged by the rule on the full training set.	54
6.3	Composition of the flagged set by behavioral family under the two threshold definitions.	59

List of Tables

4.1	Variables available in the raw dataset, grouped by category. .	29
5.1	Median profile of each cluster ($K = 8$) on the main variables. Monetary values are in euros.	40
5.2	Median profiles of anomalous and normal merchants, for one representative cluster per family (test set). Monetary values are in euros.	45
5.3	Merchants grouped by score band (test set), with the median cash amount (in euros) and median POS ratio in each band. .	47
5.4	One representative high-scoring merchant per behavioral family (test set). Cash amounts are in euros.	48
5.5	Discriminative performance of the three classifiers, in cross-validation on the training set and on the test set. The gap is the difference between the two.	49
5.6	Relative feature importance across the three models. For logistic regression the values are the normalized absolute standardized coefficients; for the tree-based models they are the built-in importance; the last column reports permutation importance.	51
6.1	Adjusted Rand Index (ARI) comparing the final partition with five partitions obtained using different random initializations. .	53
6.2	Median profile of flagged versus normal merchants on the training set. Cash amounts are in euros.	55
6.3	Median anomaly score and share of flagged merchants by macro-area in the test set.	57
6.4	Median anomaly score and share of flagged merchants for the most common sectors in the test set.	57
6.5	Composition of the flagged set by behavioral family, under the within-cluster threshold and under a single global threshold. .	58

Abstract

The increasing adoption of electronic payments provides a valuable source of information on how merchants receive and generate their revenues. In Italy, where cash is still widely used despite the mandatory adoption of POS terminals, a high reliance on cash compared to similar businesses may be a potential indicator of undeclared revenue.

This thesis, developed in collaboration with Cashberry, proposes a methodology that combines supervised and unsupervised machine learning to identify anomalous merchants in the absence of reliable anomaly labels.

The proposed methodology first uses K-Means clustering to group merchants with similar payment profiles. Anomaly scores are then computed within each cluster using z-scores, allowing each merchant to be evaluated relative to its peers. Merchants showing large deviations from the behavior of their peers are flagged as anomalous, and these flags are then used as pseudo-labels to train a supervised classifier that assigns a risk score to each merchant. Among the three models tested, LightGBM performs best with an AUC of 0.996 on the test set and a negligible gap from its cross-validated performance.

The reliability of the flags is then assessed through six internal validation checks. Bootstrap resampling shows that 96.2% of flagged merchants are consistently identified across 200 bootstrap samples, and neither the geographic nor the sectoral analysis shows any systematic concentration of flags in specific areas or business sectors.

The resulting risk score can then be used to rank merchants and prioritize those that may require further investigation. The same pipeline can also be extended to other anomaly detection settings where reliable labeled data are not available.

Sintesi

La crescente diffusione dei pagamenti elettronici fornisce una preziosa fonte di informazioni sulle modalità con cui gli esercenti ricevono e generano i propri ricavi. In Italia, dove il contante è ancora ampiamente utilizzato nonostante l'obbligo di adozione dei terminali POS, una forte dipendenza dal contante rispetto ad attività simili può rappresentare un potenziale indicatore di ricavi non dichiarati.

Questa tesi, sviluppata in collaborazione con Cashberry, propone una metodologia che combina tecniche di machine learning supervisionato e non supervisionato per identificare esercenti anomali in assenza di etichette affidabili.

La metodologia utilizza inizialmente il clustering K-Means per raggruppare gli esercenti con profili di pagamento simili. I punteggi di anomalia vengono calcolati all'interno di ciascun cluster mediante z-score, consentendo di valutare ogni esercente rispetto agli esercenti appartenenti allo stesso gruppo. Gli esercenti che mostrano deviazioni rilevanti rispetto al comportamento del proprio gruppo vengono identificati come anomali e infine tali segnalazioni vengono utilizzate come pseudo-etichette per addestrare un classificatore supervisionato, che assegna un punteggio di rischio a ciascun esercente. Tra i tre modelli testati, LightGBM ottiene le migliori prestazioni, con un AUC pari a 0.996 sul test set e una differenza trascurabile rispetto alle prestazioni ottenute in cross-validation.

L'affidabilità delle segnalazioni viene valutata attraverso sei controlli di validazione interna. Il bootstrap mostra che il 96.2% degli esercenti segnalati viene identificato in modo consistente nei 200 campioni bootstrap, mentre né l'analisi geografica né quella settoriale evidenziano una concentrazione sistematica di segnalazioni in specifiche aree o settori di attività.

Il punteggio di rischio ottenuto può essere utilizzato per ordinare gli esercenti per livello di rischio e dare priorità a quelli che potrebbero richiedere ulteriori approfondimenti. La stessa pipeline può inoltre essere estesa ad altri contesti di rilevazione di anomalie in cui non siano disponibili dati etichettati affidabili.

Chapter 1

Introduction

The digitalization of payments has transformed everyday economic activity into a continuous source of data. Every transaction that passes through a Point of Sale (POS) terminal leaves a digital record, providing a detailed picture of how merchant collect their revenue. Fintech companies increasingly use this information, together with machine learning techniques, to assess merchant creditworthiness, monitor business health, and to detect unusual patterns in payment behavior. In this context, merchants whose payment profiles differ from those of similar businesses can be detected by comparing them with their peers. Large differences in the use of payment channels may reveal unusual behavior, even when no labeled examples of anomalies are available.

This thesis proposed a methodology to identify merchants whose reliance on cash is unusually high and whose use of the POS channel is unusually low relative to their peers, since such profiles may be consistent with the under-reporting of revenue.

The phenomenon under study is far from marginal. According to [ISTAT \(2025\)](#), the value added concealed through the under-reporting of revenues and costs reached approximately 108 billion euros in 2023, equivalent to almost 5% of GDP. The informative value of this signal is further reinforced by the current regulatory framework. Following [Repubblica Italiana \(2022\)](#), accepting electronic payments through POS terminals is mandatory for all businesses and professionals selling goods or services to the public, so that a limited use of the POS channel can no longer be explained simply by the absence of a terminal. The combination of intensive cash use and limited POS activity therefore becomes a more informative signal of potentially anomalous behavior.

This thesis is developed in collaboration with Cashberry, a fintech company specialized in the analysis of merchant payment data. The proposed

methodology combines supervised and unsupervised machine learning techniques to analyze POS transaction data from a large population of merchants and identify payment profiles that differ significantly from those of similar businesses. Its purpose is to support the selection of cases that may deserve further investigation rather than provide definitive evidence of undeclared revenue.

1.1 Problem Statement

The challenge I try to address in this study is to identify merchants whose payment profiles differ significantly from that of similar businesses and may therefore indicate undeclared cash revenues. This task is challenging because of the characteristics of the problem, which make a simple approach insufficient and require a more structured methodology.

The first and most important challenge is the absence of labels. The available data does not contain information that tells us with certainty whether a given merchant is anomalous or not, so the problem is unsupervised by its nature. The only information available is the observed payment behavior of each merchant, from which the notion of anomaly is defined. Therefore, supervised classification techniques cannot be applied directly, since they require labeled observations for training.

The second challenge is the heterogeneity of the population. Evaluating all merchants against a single and uniform benchmark may wrongly classify as anomalous merchants whose payment behavior is consistent with that of their peers. Each merchant must be compared with a group of similar businesses that are as homogeneous as possible in terms of sector, size, and payment channel composition.

A challenge closely related to this problem is the risk that the identified signal does not reflect anomalous behavior but rather structural characteristics of the environment in which the merchant operates. Systematic differences across geographic areas and economic sectors may generate patterns that are mistakenly interpreted as anomalies. Therefore, the proposed methodology must be able to distinguish between anomalous behavior and structural differences, while avoiding biases that could systematically flag specific categories of merchants.

Finally, there is a fourth requirement, that is stability over time. Payment behaviors evolve over time due to both macroeconomic factors and regulatory or market changes. Therefore, the proposed methodology must be able to generalize well to data collected in different time periods while maintaining a good performance. This condition is necessary to have operational

value, useful for an institution that wants to use it as a tool for continuous monitoring.

These challenges show that the problem cannot be addressed using a simple approach. For example, applying the same threshold to all merchants does not account for differences across the population, as shown later in the analysis. This motivates the methodology developed in this thesis, which is designed to account for these differences.

1.2 Research Objectives

Considering the challenges declared above, the primary objective of this thesis is the development of a methodology for identifying, within a large population of merchants, anomalous payment profiles that may be consistent with the concealment of cash revenues. This objective can be divided into more specific ones.

The first objective is the construction of homogeneous peer groups. Since merchants must be compared with similar businesses, the population must be segmented into groups that are as similar as possible in terms of payment behavior, size, sector, and geographic area. With this goal, clustering techniques are employed to group merchants according to their characteristics without imposing predefined categories.

The second objective is to define an operational notion of anomaly. It is necessary, within each group, to assess how much a merchant's payment behavior deviates from that of its peers and to determine whether the deviation is anomalous. This leads to the development of an anomaly score and a corresponding flagging rule.

The third objective is the development of a predictive model based on anomaly flags generated in the previous stage. Once the anomalous profiles have been identified, a supervised classifier is trained to estimate the probability that a merchant is anomalous. This step transforms the notion of anomaly into a scoring tool that can be applied to new observations, while also allowing to assess whether the flagged profiles are characterized by learnable patterns.

The fourth and final objective is the evaluation of the overall methodology. In the absence of ground truth against which to assess the accuracy of the flags, validation must be based on a set of internal consistency checks, such as the stability of the peer groups, the robustness of the results, and the absence of geographic and sectorial bias. A part of this work is devoted to this aspect, since in a setting without ground truth the value of the methodology is not based on the number of flags it produces, but rather on how their

trustworthiness can be demonstrated.

These objectives lead to research questions that guide this thesis. I investigate whether, in the absence of labeled data, it is possible to define a notion of anomaly that is both economically interpretable and statistically robust, and whether the obtained flags can be identified by a classification model by exploiting the learnable characteristics of the anomalous merchants and translating the model into an operational scoring tool.

1.3 Significance of the Study

This study is relevant at three levels. The first is operational and concerns the financial institution involved in the analysis. The second is methodological and focuses on how the problem can be addressed in the absence of ground truth. The third has a broader scope and relates to the fight against the underground economy.

At the operational level, the proposed methodology represents a tool to complement the institution's existing monitoring activities. The resulting score reflects the likelihood that a merchant is anomalous, allowing cases that require closer examination to be identified and resources to be focused on the most relevant profiles. Unlike checks based on fixed threshold, this scoring system accounts for the heterogeneity of the population and provides a continuous measure of risk, giving analysts not only a set of flags but also an indication of their relative intensity.

At the methodological level, the main contribution of this thesis is the development of an end-to-end framework for anomaly detection in the absence of labeled data. Anomalous behavior must be defined from the data itself, and the reliability of the resulting flags must be assessed through several validation checks.

At the broader level, this study contributes to the fight against the underground economy, which is an important issue in Italy. Transactional data can be used to identify unusual behavior and may be useful not only to financial institutions but also to those working to combat tax evasion.

Chapter 2

Literature review

This work lies at the intersection of several research areas, including anomaly detection, financial fraud detection, peer group analysis, data-driven approaches to tax evasion, and methods that combine unsupervised anomaly detection with supervised learning. The literature review begins with the general methodological foundations and then moves to studies that address problems more closely related to the one considered in this thesis.

2.1 Anomaly Detection

Anomaly detection is a mature field of research, and one of its main references is the survey by [Chandola et al. \(2009\)](#). The authors define anomalies as "patterns in the data that do not conform to the definition of normal behavior" and provide a comprehensive review of anomaly detection methods. Their survey classifies existing approaches according to the underlying methodology and discusses applications across a wide range of domains, emphasizing that the concept of anomaly depends on the context in which it is studied.

The survey also distinguishes between point, contextual, and collective anomalies. The problem addressed in this thesis belongs to the second category, since the objective is not to identify isolated observations but merchants whose behavior is anomalous only when compared with that of similar merchants. Another important point highlighted by Chandola et al. concerns the availability of labeled data. Anomaly detection techniques based on classification require labels for both normal and anomalous instances, which are often not available in practice. This makes the validation of anomaly detection methods particularly challenging in settings where no ground-truth labels are available.

One of the most influential anomaly detection method is the Local Outlier Factor (LOF) proposed by [Breunig et al. \(2000\)](#). LOF assigns each observation an outlier score based on how much its local density differs from that of its neighbors, rather than comparing it with a global threshold. Although the method developed in this thesis does not estimate local density, it follows the same principle by comparing each merchant only with the members of its own peer group, identified through clustering.

Another widely used method is the Isolation Forest introduced by [Liu et al. \(2008\)](#). It does not model normal behavior or measure deviations from it, but instead identifies anomalies based on the fact that they are rare and different from the majority of data. The algorithm uses random partitions to separate observations, with anomalous ones typically being isolated in fewer splits than normal ones. Its computational efficiency has made it a standard method for anomaly detection in large datasets.

2.2 Fraud Detection with Machine Learning

Another methodological reference for this thesis is the financial fraud detection literature, for which the review by [Bolton and Hand \(2002\)](#) remains one of the main references. The authors survey the statistical machine learning methods used to detect fraud in a variety of applications, including money laundering, credit card fraud, telecommunications fraud, and computer intrusion.

The review distinguishes fraud prevention, which tries to stop fraudulent activity before it occurs, from fraud detection, which identifies it afterwards. It also separates supervised from unsupervised approaches, and highlights two challenges common to fraud detection. Fraudulent cases are rare and represent only a small fraction of all observations. More over, classification errors are asymmetric, since missing a fraud and flagging a legitimate case do not carry the same cost. Finally, the authors stress that fraud detection models must be updated over time because both legitimate and fraudulent behavior change.

The problem addressed in this thesis differs from the settings considered in most fraud detection studies. The objective is not to identify individual fraudulent transactions, but merchants whose payment behavior is unusual when compared with similar merchants. For this reason, the proposed methodology is more closely related to behavior anomaly detection than to traditional fraud detection based on individual transactions. It is therefore naturally connected to peer group analysis, introduced by the same authors in the next section.

2.3 Peer Group Analysis

The idea of comparing each subject with similar profiles, rather than against a fixed threshold, is central to the methodology developed in this thesis. This principle originates from the peer group analysis (PGA) proposed by [Bolton et al. \(2001\)](#).

Their work focuses on the detection of behavioral fraud using longitudinal data and introduces two complementary approaches. The first, called break point analysis, identifies abrupt changes in the behavior of an individual, by comparing current observations with his historical behavior. The second, peer group analysis, compares each subject with a group of similar subjects and detects cases whose behavior deviates from that of their peers over time. This second approach provides the main methodological inspiration for the present work.

[Weston et al. \(2008\)](#) further developed this idea by applying peer group analysis to real payment card transaction data provided by a UK bank. Their study demonstrates how the method can be applied on a large scale, with peer groups constructed for a large number of accounts, while only a limited number of investigations can realistically be conducted by an institution.

The methodology proposed in this thesis adopts the same principle of peer-based comparison but implements it differently. In the original formulation, peer group analysis is a dynamic method that monitors changes over time, while here, peer groups are obtained by clustering merchants according to their overall payment profiles, and each merchant is compared with the typical behavior of its own cluster. The common idea is that anomalies should be identified relative to comparable peers rather than by using a single global reference.

2.4 Tax Evasion Detection Methods

A related line of research focuses on the use of data-driven methods to detect tax evasion. Among these studies, [De Roux et al. \(2018\)](#) provide the closest methodological reference to the present work. They propose an unsupervised approach for identifying under-reported tax returns by first clustering tax declarations into homogeneous groups and then flagging observations in the tail of the within-cluster distribution using the quantiles.

The methodology used in this thesis is based on the same general idea, but there are some important differences. De Roux et al. use spectral clustering to define the peer groups, while this work is based on K-Means. Their anomaly threshold is obtained from a distribution estimated through kernel

density estimation, whereas here it is based on empirical percentile. Finally, the two studies use different types and amounts of data. De Roux et al. analyze just over one thousand tax returns, while this thesis considers a much larger population of merchants using aggregated POS transaction data rather than tax declarations.

[Savić et al. \(2022\)](#) propose a hybrid unsupervised outlier detection method that combines K-Means clustering with representation learning through an autoencoder for tax evasion risk management. Since ground-truth labels are not available, outliers identified by the two methods are cross-checked to obtain a set of internally validated anomalies. The present work shares the absence of ground-truth labels and uses clustering as part of the pipeline, but differs in how clustering is used. In their work, clustering serves directly as a detection mechanism, whereas here it defines peer groups within which anomalies are detected.

2.5 Pseudo-Labeling

A final line of research concerns methods that use the output of an unsupervised algorithm to train a supervised model, an approach commonly known as pseudo-labeling. Once the unsupervised method identifies the anomalies, the resulting flags are used as labels for a classifier. The classifier can then assign a continuous risk score to new observations and provide an additional check on the reliability of the original flags. If the classifier can learn these labels and apply them to new data, this suggests that the identified anomalies are not simply due to random noise.

The coarse-to-fine pseudo-labeling framework proposed by [Al-Lahham et al. \(2024\)](#) is particularly close to the approach adopted in this thesis. Although developed for anomaly detection in surveillance video, its overall structure is similar, as it combines clustering to generate coarse pseudo-labels with a statistical hypothesis test for identifying anomalous observations and a supervised classifier trained on the resulting labels to score new data.

2.6 Conclusion

Overall, these studies share a few common principles. Anomalies should be evaluated relative to comparable subjects rather than in isolation, clustering provides a natural way to define these reference groups, and the output of unsupervised methods can be used to train supervised predictive models. A common limitation of all these approaches is that, in the absence

of ground-truth labels, the accuracy of the detected anomalies cannot be assessed directly.

Chapter 3

Methodology

This chapter introduces and explains the techniques, models, and methods used in this research. The purpose is to describe how each of these techniques operates from a conceptual and formal perspective, while leaving the discussion of their implementation to the next chapter. The chapter presents the methods in the order in which they are used. It starts with data preparation and feature construction, then introduces the clustering procedure, the anomaly detection approach, and finally the supervised classification stage.

3.1 Data Representation and Pre-processing

The raw data must be preprocessed before any learning algorithm can be applied. This includes transforming the variables into a numerical representation, handling categorical features, making variables comparable, and reducing the effect of strongly skewed distributions.

3.1.1 Standardization (z-score)

Many machine learning algorithms, particularly distance-based methods such as K-Means, are sensitive to the scale of the variables. A variable with large numerical values will dominate the distance computation relative to a variable with smaller values, regardless of their importance. Z-score standardization ensures that each variable has zero mean and unit standard deviation, making the different dimensions comparable.

Given an observation x , with mean μ and standard deviation σ computed over the reference population, the standardized value is

$$z = \frac{x - \mu}{\sigma}.$$

The value z indicates how many standard deviations the observation lies from the mean. Positive values indicate observations above the mean, while negative values indicate observations below it.

In this work, standardization has two objectives. First, it makes the variable comparable before clustering, so that the distance computations will not be dominated by the variables with the largest numerical ranges. Second, it constitutes the definition of the anomaly measure, where the deviation of a merchant from its peer groups is expressed in standardized terms.

3.1.2 Target Encoding

Categorical variables, such as geographic area and the sector of activity, cannot be directly used by algorithms that are based on numerical data. Target encoding ([Micci-Barreca, 2001](#)) addresses this limitation by replacing each category with a numerical value derived from an associated target variable. Typically, this value corresponds to the mean of the target variable computed over all observations belonging to that category. This approach is particularly useful when categorical variables have many distinct levels, since alternative techniques such as one-hot encoding would generate many sparse binary columns, which can distort the distance between observations in methods such as K-Means.

3.1.3 Logarithmic Transformations

Many economic variables, such as transaction volumes or monetary amounts, are highly skewed. Most observations lie around relatively small values, while a small fraction takes on very large values, producing a long right tail. In this case, both distance and summary statistics are dominated by extreme observations. The logarithmic transformation compresses the right tail and stretches the lower part of the distribution, yielding a more symmetric shape. The transformation typically takes the form

$$x' = \log(1 + x),$$

where the addition of 1 allows zero values to be handled, for which the logarithm would otherwise be undefined. In this work the transformation is applied to the POS volume, whose distribution exhibits precisely the asymmetry described above.

3.2 Clustering

Clustering is an unsupervised learning technique that aims to partition data into groups (clusters) such that observations within the same cluster are more similar to one another than to those belonging to different clusters. Since no labels are available, these groups are not predefined but emerge from the underlying structure of the data. In this work, clustering is used to construct reference groups within which each subject is subsequently compared to its peers.

3.2.1 K-Means

K-Means ([Lloyd, 1982](#)) is one of the most widely used clustering algorithms, particularly for its simplicity and efficiency. The algorithm requires the desired number of clusters, denoted by K , to be specified in advance, and partitions the observations into K clusters by minimizing the within-cluster dispersion. Specifically, the algorithm minimizes the sum of the squared distances between each observation and the centroid of the cluster to which it is assigned, a quantity known as inertia or within-cluster sum of squares (WCSS):

$$\text{WCSS} = \sum_k \sum_{x \in C_k} \|x - c_k\|^2,$$

where c_k denotes the centroid of cluster k , defined as the mean of the observations assigned to it. After initializing the centroids, the algorithm proceeds iteratively by alternating between two steps until convergence.

In the first step, each observation is assigned to the nearest centroid, while in the second one, each centroid is updated as the mean of the observations assigned to it. These two steps repeated until the cluster assignments no longer change. The algorithm is run multiple times with different initializations using the K-Means++ seeding strategy developed by [Arthur et al. \(2007\)](#), which selects initial centroids to maximize their mutual distance and reduces the risk of convergence to suboptimal solutions, and at the end the solution with the lowest inertia is retained.

K-Means tends to identify clusters with approximately spherical shape and is sensitive to feature scaling, which is why it is typically preceded by standardization. This algorithm is particularly useful for segmenting large populations into compact, behaviorally homogeneous groups, and it scales well to datasets with many observations.

Selecting an appropriate value of K is an important decision, since its value must be specified in advance and there is no universally correct answer.

However, different evaluation metrics can assist the choice of K and the assessment of the quality of the segmentation.

The Silhouette coefficient ([Rousseeuw, 1987](#)) measures how well each observation matches its assigned cluster relative to the closest alternative cluster. It is defined as

$$s = \frac{b - a}{\max(a, b)},$$

where a is the mean distance between the observation and all other points in its own cluster, and b is the mean distance between the observation and the points in the nearest neighboring cluster. Its values range from -1 to $+1$, where values close to $+1$ indicate that the observation is well matched to its assigned cluster, values around 0 indicate that it lies near the boundary between two clusters, while negative values suggest that the observation may fit better in a different cluster. At the end, the average Silhouette coefficient across all observations provides an overall measure of clustering quality.

The Davies-Bouldin index ([Davies and Bouldin, 1979](#)) is based on a different mechanism, since it compares the compactness of each cluster with its separation from the others. For every pair of clusters, the ratio between the sum of their within-cluster dispersions and the distance between their centroids is computed. For each cluster, only the largest of these ratios is retained, as it represents the cluster that is most similar to it. The Davies-Bouldin index is then obtained by averaging these values across all clusters. Unlike the Silhouette coefficient, lower values of the Davies-Bouldin index indicate a better partition, corresponding to clusters that are both more compact and more clearly separated.

3.3 Peer Deviation for Anomaly Detection

Anomaly detection aims to identify observations that deviate significantly from expected behavior. In the absence of labels that indicate which observations are genuinely anomalous, the notion of anomaly cannot be learned from the data but must be defined in advance. The approach used in this thesis is based on a simple statistical idea: the more an observation departs from what is typical for a comparable subject, the more anomalous it is.

The main challenge associated with this approach is the heterogeneity of the population. Applying a single threshold uniformly to the whole population would flag as anomalous merchants that are simply unusual compared with the overall population, rather than relative to similar businesses. For this reason, the comparison must be relative rather than absolute, so

that each subject is evaluated against a set of comparable peers rather than against the population as a whole.

This principle combines two methodological elements introduced earlier. Peer groups are first defined through clustering, so that each subject is compared only with similar subjects rather than with the entire population. The departure of each subject from its peer group is then measured using the z-score, which expresses the distance from the cluster mean in units of standard deviations and is independent of the scale of the variable. Higher z-scores identify subjects whose behavior is unusual relative to their peers. A key advantage of this formulation is its interpretability, since the measure directly indicates how many standard deviations a subject departs from the typical behavior of comparable subjects. This is particularly important in a context where each flagged case must be explained and justified.

To turn the deviation measure into an actual flagging decision, a criterion that separates anomalous observations from the rest must be defined. Rather than fixing an arbitrary absolute cut-off in the score, the criterion can be expressed as the percentile of the score distribution, flagging as anomalous a predetermined fraction of the subjects with the most pronounced deviations. The resulting flags provide the basis for the two subsequent stages of the analysis. They define the anomalies from which a supervised classifier is later trained, and they represent the segmentation whose stability and coherence are subsequently examined.

3.4 Supervised Classification

Unlike clustering, supervised classification is applied when a set of labeled observations is available, allowing the model to learn the relationship between the input variables and the corresponding labels, and to predict the labels of new observations. In this thesis, the classification models are trained on the anomaly flags generated through the unsupervised procedure, transforming an analytically defined measure of anomaly into a predictive model.

Three classifiers of increasing complexity are considered: logistic regression as an interpretable linear baseline, and two tree-based ensemble methods, Random Forest and LightGBM, which can capture non-linear relationships. Using different models makes it possible to assess whether the anomaly pattern can be learned regardless of the specific model applied, and to compare a simple and transparent model like logistic regression against more flexible ones in terms of predictive performance.

3.4.1 Logistic Regression

Logistic Regression is one of the simplest and most interpretable classification models, and it often represents the baseline against which more complex models are compared. It models the probability that an observation belongs to the positive class as a function of its explanatory variables. To ensure that the predicted probabilities lie between 0 and 1, logistic regression applies the logistic (or sigmoid) function to a linear combination of the explanatory variables:

$$p = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \dots + \beta_n x_n)}},$$

where $x_1, \dots, x_n$ are the explanatory variables and $\beta_0, \dots, \beta_n$ are the coefficients estimated from the data. The linear combination in the exponent produces a continuous score for each observation, which the sigmoid function maps to a probability between 0 and 1. The observation is then assigned to one of the two classes according to whether the predicted probability exceeds a predefined threshold. The model parameters are estimated by maximizing the likelihood of the observed data, that is the probability of observing the data given a particular set of parameter values.

The main strength of logistic regression is its interpretability, since the sign and magnitude of each coefficient represent the direction and strength of the relationship between the explanatory variable and the probability of belonging to the positive class. Its main limitation is the linear nature of its decision boundary, which may be insufficient to capture complex or non-linear relationships among the variables. This limitation motivates the use of tree-based models presented in the following sections.

3.4.2 Decision Trees

The decision tree is the building block of the models considered in this section. It classifies an observation through a sequence of decision rules organized in a tree structure. At each internal node, the algorithm compares the value of a predictor with a threshold selected according to an impurity criterion and assigns the observation to the corresponding branch. This process continues until a terminal node (leaf) is reached, where the predicted class is assigned. Decision trees are easy to interpret because each prediction can be traced through the sequence of splits that leads to the final leaf. Their main limitation is that a single tree can adapt too closely to the training data and is sensitive to small changes in the observations, but ensemble methods address these limitations by combining the predictions of multiple trees.

3.4.3 Gradient Boosting and LightGBM

Gradient boosting ([Friedman, 2001](#)) is an ensemble learning technique that builds a predictive model by sequentially combining multiple decision trees. Instead of training trees independently and aggregating their predictions, each new tree is built to correct the errors made by the current ensemble. More precisely, at every iteration the new tree is fitted to the negative gradient of the loss function evaluated at the current predictions, thereby directing the learning process toward observations that are still difficult to classify. The contribution of each tree is scaled by a learning rate, a shrinkage parameter that controls how much each iteration influences the ensemble and helps reduce overfitting. Thanks to this iterative procedure, the model is able to capture complex, non-linear relationships among the variables while achieving high predictive accuracy.

LightGBM ([Ke et al., 2017](#)) is an efficient implementation of gradient boosting designed to reduce computational cost while maintaining predictive performance. It achieves this through several algorithmic refinements in the tree construction process, resulting in faster training and lower memory consumption than traditional gradient boosting implementations. Due to its computational efficiency and strong predictive performance, LightGBM is one of the most widely used machine learning algorithms for classification problems. Its structure is also useful in this application, as the relationship between payment features and the anomaly pattern may be not linear. A merchant may be atypical because of how several characteristics combine, not because of a single one of them, and boosting models are able to capture this type of relationship.

3.4.4 Random Forest

Random Forest ([Breiman, 2001](#)) is an alternative ensemble approach based on bagging rather than boosting. It constructs many decision trees independently, each trained on a bootstrap sample of the training data. At every split, only a randomly selected subset of the predictor variables is considered. The predictions of all trees are then aggregated, using majority voting for classification problems and averaging for regression. Because each tree is trained on a different bootstrap sample and considers different subsets of variables at each split, the trees are less correlated with one another. This reduces the variance of the ensemble and makes it more stable than an individual decision tree. Random Forest therefore provides a useful benchmark against which the performance of the gradient boosting model can be compared.

3.5 Classifier Evaluation Metrics

Class imbalance is an important issue when evaluating classification models. This is particularly relevant in anomaly detection, where anomalous cases usually represent only a small fraction of the observations. For this reason, model performance is evaluated using several metrics rather than relying on the proportion of correct predictions.

3.5.1 Confusion Matrix

The confusion matrix is the fundamental tool for evaluating the performance of a binary classifier. It represents the actual classes against the predicted ones, yielding four outcomes: true positives (anomalous observations correctly identified), true negatives (normal observations correctly identified), false positives (normal observations incorrectly flagged as anomalous), and false negatives (anomalous observations that remain undetected). All the evaluation metrics described below are derived from these four quantities.

3.5.2 Precision, Recall and F1

Precision measures the proportion of observations flagged as anomalous that are actually anomalous. Formally, it is defined as the ratio between the true positives and the sum of true positives and false positives.

Recall, by contrast, measures the proportion of truly anomalous observations that the model successfully identifies. It is defined as the ratio between the true positives and the sum of the true positives and false negatives. The two metrics involve a trade off, as increasing the sensitivity of the model typically improves recall at the expense of precision, and vice versa.

The F1-score combines precision and recall into a single measure by considering their harmonic mean:

$$F1 = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}}.$$

Unlike the arithmetic mean, the harmonic mean is high only when both precision and recall are high, penalizing models that perform well on one metric at the expense of the other.

3.5.3 ROC Curve and AUC

A classifier does not directly produce a binary decision, but rather a probability that an observation belongs to the positive class. Different thresholds

applied to this score lead to different trade-offs between correctly identifying positive cases and generating false alarms. The Receiver Operating Characteristic (ROC) curve (Fawcett, 2006) illustrates these trade-offs by relating the true positive rate (recall) to the false positive rate over the full range of possible thresholds. ROC curves closer to the upper-left corner indicate better classifier performance. The Area Under the ROC curve (AUC) summarizes the ROC curve in a single value ranging from 0 to 1. A value of 0.5 indicates performance similar to random guessing, while values closer to 1 indicate better discrimination between the two classes. The AUC does not depend on a specific decision threshold and is less affected by class imbalance than accuracy. For these reasons, it is useful for evaluating binary classification models when the classes are unevenly distributed.

3.5.4 Cross-Validation

A model should be evaluated on data that were not used during training, otherwise its performance is likely to be overestimated. In cross-validation, the dataset is divided into several folds, then the model is trained repeatedly, each time using a different fold for validation and the remaining folds for training, until every fold has served once as the validation set.

The most common variant is k -fold cross-validation, in which the data are divided into k folds and the procedure is repeated k times, each time using a different fold for validation and the remaining $k - 1$ folds for training. The performance obtained on the validation fold at each iteration is averaged to estimate how well the model is expected to perform on unseen data. Among the performance measures introduced in the previous section, the AUC is used as the main criterion for comparing the classifiers.

The models are evaluated through stratified cross-validation, which keeps the same proportion of anomalous cases in every fold. Their performance is then compared with the results obtained on data from a later period, to evaluate whether they generalize well to unseen future data.

3.6 The Overall Pipeline

Figure 3.1 shows the overall structure of the methodology. First, the raw transaction data are aggregated into a monthly panel and split into training and test sets based on a temporal cutoff. The training data are then used to construct merchant profiles and calculate the features described in the next chapter. The clustering algorithm divides merchants into 8 homogeneous peer groups belonging to three behavioral families. For each merchant, an

anomaly score is calculated within the assigned cluster, and a specific threshold for each cluster is used to identify the pseudo-labels. Finally, a supervised classifier is trained using these labels and applied to the test set to obtain a continuous risk score for each merchant. The dashed arrows represent the validation stage, which consists of a set of internal checks applied to the clustering results, anomaly scores, and classifier. Since no ground-truth labels are available, these checks are used to assess the reliability of the methodology.

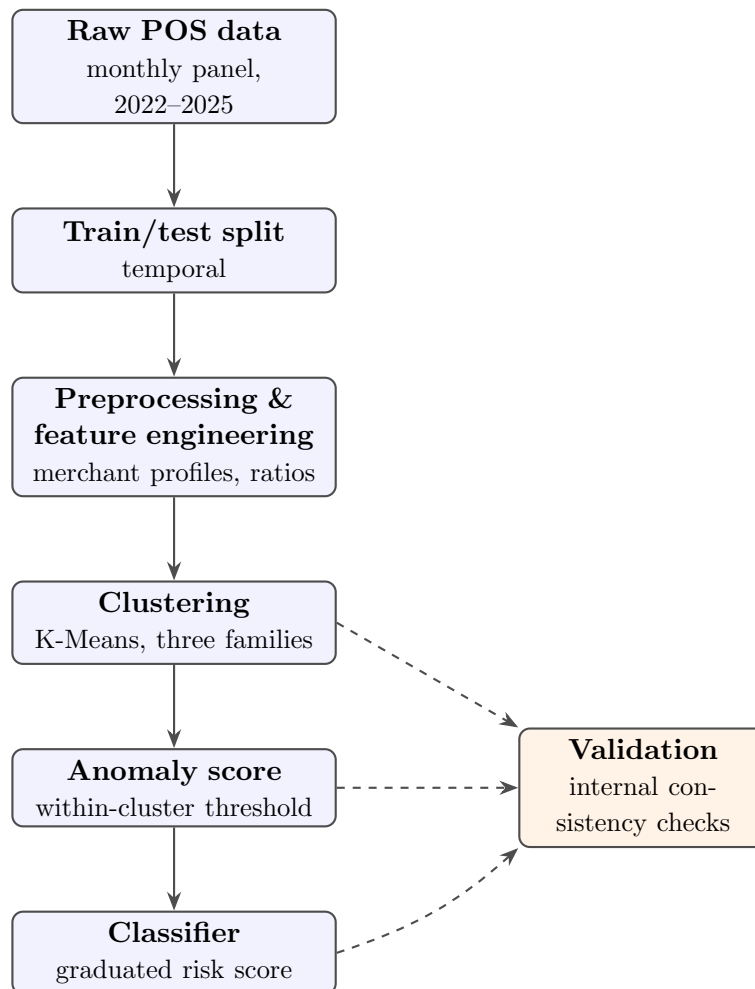


Figure 3.1: Overview of the methodology as a pipeline. The main stages describe the process from the raw data to the risk score produced by the classifier. The validation stage then assesses the internal consistency of the clustering, anomaly scores, and classifier.

Chapter 4

Experimental Setup

4.1 Dataset Description

The data used in this work are provided by Cashberry, the company with which this thesis was carried out, which in turn operates in collaboration with the banking group BPER. The dataset is composed of POS transaction records generated by a large number of merchants, and covers the period from January 2022 to August 2025.

The raw dataset contains approximately 26.6 million monthly observations referring to 243,933 distinct merchants. Each observation summarizes the activity of a single merchant for a given month across different dimensions. Table 4.1 lists the available variables, grouped into three broad categories.

The first category comprises the monetary variables describing the amounts collected through each payment channel, namely the total amount received in cash, the total amount received by bank transfer, and the amount collected through POS transactions. The second category contains variables that describe business activity, including the number of transactions and active terminals. The last category contains several categorical and administrative variables. These include the merchant’s ATECO code and sector description, the province of the registered office, the business opening date, the commercial service model, an anonymized merchant identifier, and the reference month.

From these records, a monthly panel is constructed, with each observation corresponding to a merchant in a given month, resulting in 6,622,611 observations. The merchant population is highly heterogeneous, both in terms of sector and geographic distribution. The dataset covers 28 sectors, identified by their ATECO code, with a predominance of activities characterized

Category	Variable	Description
Monetary	TOT_CONTANTI_AM	Total amount collected in cash
	TOT_BON_AM	Total amount received by bank transfer
	VOLUME_AM	Total amount collected through the POS terminal
Activity	TRANSAZIONI_QT	Number of transactions in the month
	TERMINALI_QT	Number of active POS terminals
Registry	ATECO_CD	Numeric code of the sector of activity
	categoria_ateco	Textual description of the ATECO sector
	PROVINCIA_SEDE_LEGALE_CD	Province of the registered office
	APERTURA_DT	Date the business was opened
	MODELLO_SERVIZIO_CD	Commercial service classification
	CLIENTE_CD_new	Anonymized merchant identifier
	MESE_RIFERIMENTO_CD	Month to which the observation refers

Table 4.1: Variables available in the raw dataset, grouped by category.

by frequent low-value retail transactions, such as food service, personal care services, bars, and various forms of retail trade. The composition of the merchants by sector is shown in Figure 4.1.

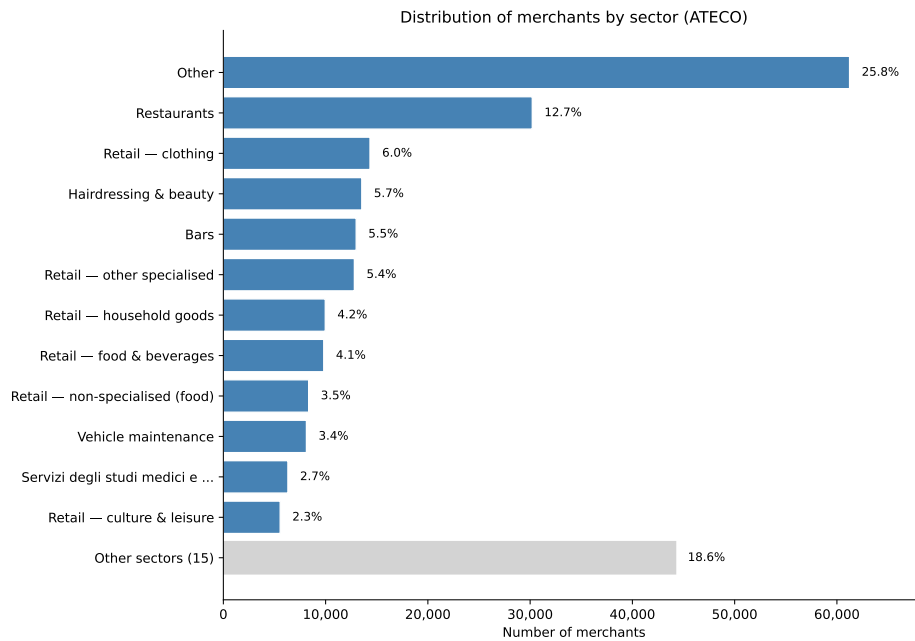


Figure 4.1: Distribution of merchants by sector (ATECO). Merchants with unknown sector are excluded; the grey bar aggregates the remaining 15 sectors.

Geographically, the merchants are distributed across 110 provinces, which are aggregated into five broader macro-areas for the purpose of the analysis: North-West, North-East, Centre, South, and Islands. The geographic distribution of the merchants is shown in Figure 4.2.

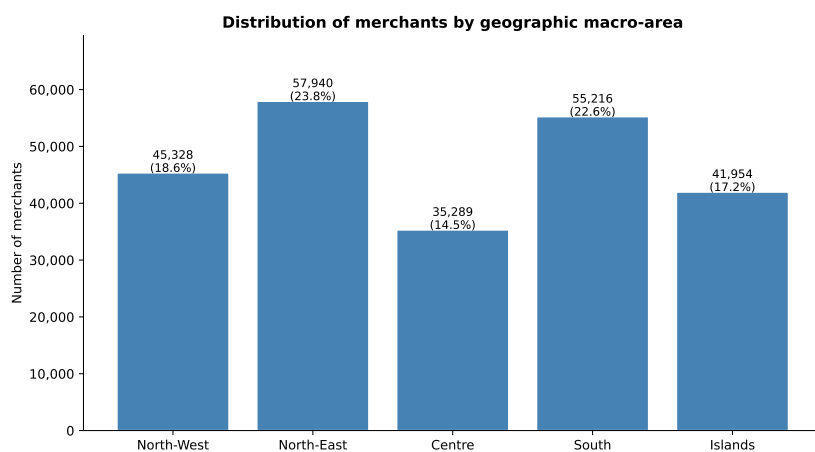


Figure 4.2: Distribution of merchants across the five geographic macro-areas.

The next sections describe the construction of merchant-level features and the filtering steps applied to obtain the final set of merchants used in the analysis.

4.1.1 Train/Test Split

Before any features are constructed, the dataset is divided into training and test sets. The partition is based on time, where the training set includes all merchant-month observations from January 2022 to December 2023, and the test set contains observations from January 2024 to August 2025. Before the feature engineering and filtering steps described in the following sections, the training period includes 233,796 distinct merchants, while the test period includes 139,007 merchants.

4.2 Data Processing and feature Engineering

The analysis uses a merchant-level representation, where each merchant is described by a single profile summarizing its payment behavior over the observed period. This section describes how these profiles are constructed and how the initial population is reduced to the set of merchants on which the analysis is performed.

4.2.1 Merchant-level features

Each merchant is represented by a set of features obtained by aggregating its monthly observations. The main features are based on monthly averages calculated over the months in which each merchant is observed. These include the average POS volume (*vol_mean*), the average cash collected (*cash_mean*), and the average amount received through bank transfers (*bon_mean*). Monthly averages are used instead of total amounts so that merchants with different observation periods can be compared directly. Three additional ratios are then calculated from these variables to describe the composition of each merchant's incoming flows. Let

$$total_flows = vol_mean + cash_mean + bon_mean$$

denote the sum of the three average monthly inflows. The ratios are then defined as

$$\begin{aligned} cash_ratio &= \frac{cash_mean}{total_flows}, & pos_ratio &= \frac{vol_mean}{total_flows}, \\ bon_ratio &= \frac{bon_mean}{total_flows}. \end{aligned}$$

and are set to zero whenever $total_flows = 0$. The three ratios sum to one by construction, and represent the share of each merchant’s payment inflows received through cash, POS transactions, and bank transfer. These ratios are central to the analysis because they describe how a merchant collects its revenue independently of the overall scale of its activity.

4.2.2 Encoding of categorical variables

The two categorical variables retained for the analysis, namely the geographic macro-area and the ATECO sector, are converted into numerical features using target encoding. Each category is replaced by the mean value of vol_mean computed across the merchants belonging to that category. As a result, both variables are represented by a single continuous feature reflecting the typical scale of activity of each category, denoted as $area_enc_mean$ and $ateco_enc_mean$, respectively. Compared with one-hot encoding, this approach keeps the number of features low and assigns each category a meaningful numerical value. This is particularly useful for the distance-based clustering method used later in the analysis.

4.2.3 Selection of the analysis population

Since the raw dataset includes merchants that are not relevant to the analysis, the initial population is progressively reduced through a series of filtering steps, each motivated by a specific criterion.

The first filter removes merchants that are inactive throughout the observed period, namely those whose average POS volume, cash collected, and bank transfers are all equal to zero. Since these merchants generate no payment activity, they provide no information on payment behavior.

A second filter excludes merchants with unknown geographic area. Since geographic information is used for target encoding and in the subsequent analysis, merchants with missing geographic information are excluded.

The final filtering step removes merchants with zero average cash inflows ($cash_mean = 0$), as the phenomenon under study does not apply to merchants that do not receive cash payments. Including them would not add relevant cases and would introduce merchants for whom the anomaly score cannot be meaningfully defined.

After applying these filters, the training population is reduced from 233,796 to 167,247 merchants, representing the sample used for the clustering and anomaly analyses. The same feature construction procedure is then applied to the test period, so that the merchants in the test set are described by the same set of features.

4.3 Clustering Configuration

The clustering step segments merchants into peer groups based on their payment behavior. This section describes the features used for clustering, the transformations applied to them, and the selection of the number of clusters. Figure 4.3 shows how the cash ratio varies across merchants and why merchants need to be compared within peer groups. A large share of merchants collects almost entirely in cash, while the others cover a much wider range, from limited cash usage to a more balanced mix of payment channels. Given these differences, a single threshold would not be suitable for all merchants, since a cash ratio that is common in one type of activity may be unusual in another.

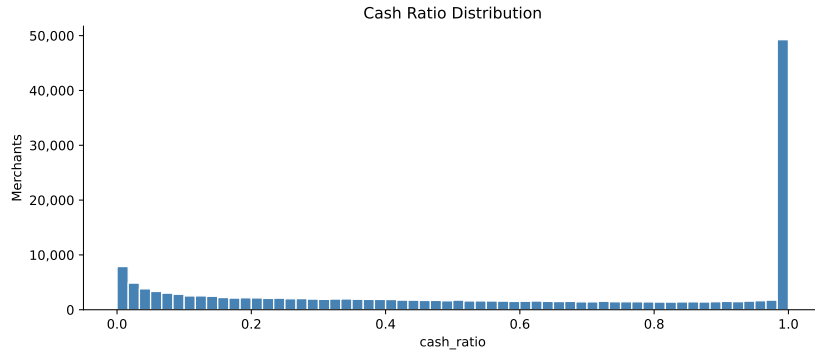


Figure 4.3: Distribution of the cash ratio across merchants.

4.3.1 Clustering features

Each merchant is represented by five features: the logarithm of the average POS volume ($\log(1 + vol_mean)$), the cash ratio, the bank-transfer ratio, and the two target-encoded variables representing the geographic area and the ATECO sector. These features are selected to group merchants according to the composition of their payment inflows and the scale of their activity, while also considering their geographic and sectoral context.

The monetary variables, *cash_mean* and *bon_mean*, are excluded from the clustering features because they would mainly group merchants by size, separating those with large payment volumes from smaller ones. The goal here is to group merchants that are homogeneous in payment profile, so that a merchant with an unusually large share of cash relative to its peers is more easily detected. For that reason the composition of incoming flows is captured through ratios, which don't depend on scale, while the overall scale

of activity is captured through the log transformation of POS volume. The two encoded variables add geographic and sectoral context while remaining scale-independent, as each category is represented by the typical volume of its members.

4.3.2 Transformations

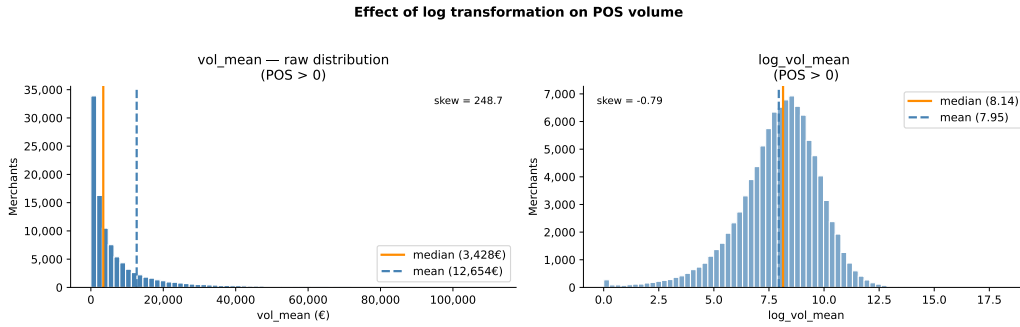


Figure 4.4: Effect of the logarithmic transformation on the POS volume. Merchants with zero POS volume are excluded from the plot.

The average POS volume has a strongly right-skewed distribution, as most merchants have relatively low volumes, while a small number have much higher values. Since these extreme observations can have a large influence on the distance calculation, the variable is log-transformed before clustering. The transformation compresses the higher values and reduces their influence, producing a more balanced distribution, as shown in Figure 4.4.

The features are then standardized to have zero mean and unit variance, so that all features are on the same scale and this prevents variables with larger numerical ranges from having a greater influence on clustering.

In addition, to further limit the influence of the most extreme observations on the position of the centroids, the features are clipped at their 99.5th percentile before standardization, while the ratios are, by construction, bounded to the $[0, 1]$ interval. This affects only a very small fraction of merchants at the far tail of the distribution, but prevents extreme outliers from distorting the resulting groups.

4.3.3 Choice of the number of clusters

Since K-Means requires the number of clusters K to be fixed in advance, an appropriate value must be selected. Two internal metrics are used to guide this choice, namely the Silhouette coefficient and the Davies-Bouldin index,

computed for values of K ranging from 5 to 15. Neither metric identifies a clear optimal value of K . As shown in Figure 4.5, the Silhouette coefficient increases up to $K = 7$, where it reaches 0.369, and then remains almost unchanged, fluctuating between 0.360 and 0.369 over the rest of the range. The Davies-Bouldin index leads to the same conclusion. Its lowest values are observed at $K = 6$ and $K = 7$, while the differences across the remaining values of K are small. Overall, both metrics indicate that the data do not identify a single geometrically optimal number of clusters, and that the choice cannot be made on the basis of these indices alone.

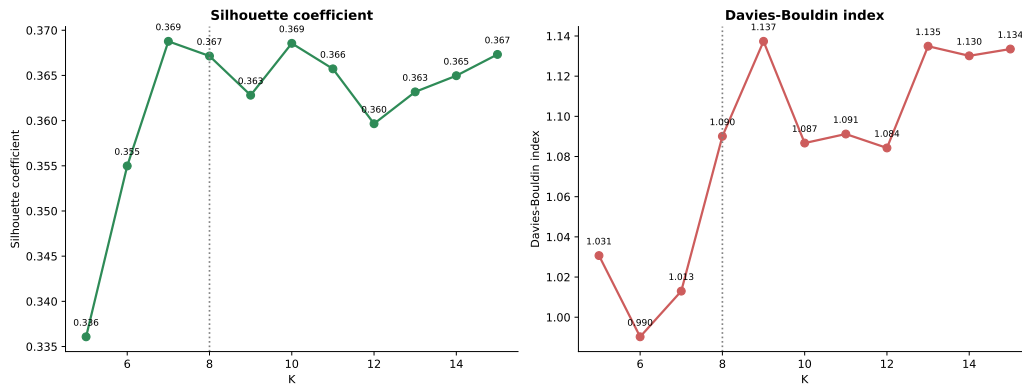


Figure 4.5: Silhouette coefficient and Davies-Bouldin index for K from 5 to 15. A higher silhouette and a lower Davies-Bouldin index indicate better-defined clusters. The dashed line marks the selected value $K = 8$.

$K = 8$ is not the geometric optimum according to the internal metrics, as $K = 7$ achieves a slightly higher score. However, the choice of eight clusters is driven by the purpose of the clustering. Since the anomaly score relies on a distinction between three behavioral families defined by the dominant payment channel, the clusters need to provide a clear separation between these families.

At $K = 7$ some clusters still mix profiles that belong to different families, so the segmentation does not separate the behaviors on which the score relies. At $K = 8$ each cluster belongs clearly to one of the three families. The small loss in the internal metrics is an acceptable price for a segmentation that fits better the structure used in the rest of the analysis. The value at $K = 8$ is 0.367, marginally below the maximum.

4.3.4 Behavioral families

The eight resulting clusters naturally group into three broad behavioral families defined by the dominant channel through which merchants collect their revenues.

The first family, indicated as *cash-only*, includes the clusters whose merchants collect almost entirely in cash, with a negligible POS volume. The second, denoted as *B2B*, comprises the clusters characterized by a high proportion of bank transfers. The third, referred to as the *retail* family, comprises clusters with a mixed payment profile, where cash and POS payments are both used. This grouping into three families represents the basis on which the anomaly score is defined, since the notion of anomalous behavior depends on the family a merchant belongs to, as described in the next section. The specific composition and profile of the eight clusters are presented in the results chapter.

4.4 Anomaly Score Definition

Once merchants have been grouped into peer clusters, an anomaly score is defined to quantify, for each merchant, how far its payment behavior departs from that of its peers. Following the peer-deviation principle introduced in the methodology, the score is built on within-cluster standardized deviations, but its exact form depends on the behavioral family to which the cluster belongs, since anomalous behavior does not mean the same thing across families.

For the retail and B2B families, the anomaly score is based on the composition of payment flows, and a merchant is considered anomalous when its payment composition is unusually skewed toward cash and away from the POS channel, relative to its peers. The anomaly score is defined as:

$$\text{score} = z_{\text{cash_ratio}} - z_{\text{pos_ratio}},$$

where $z_{\text{cash_ratio}}$ and $z_{\text{pos_ratio}}$ are the standardized deviations of the cash ratio and the POS ratio computed within each cluster. A high value of the score identifies merchants whose payment mix differs substantially from the typical profile of their peers, with a higher cash share and lower POS share than comparable merchants.

This approach cannot be applied in the same way to the cash-only family, as these merchants collect almost entirely in cash and therefore show little variation in their cash and POS ratios across observations. Standardized measures depend on the variability within each group, so a feature with

little variation cannot distinguish between merchants. In addition, when the standard deviation is close to zero, the resulting z-score becomes unstable.

What actually distinguishes an anomalous cash-only merchant is not the composition of its flows, which is fixed, but the absolute level of cash handled, since within a group of merchants that all operate almost in cash, those managing large amounts are flagged as anomalous. The score for these clusters is therefore defined as

$$\text{score} = z_{\text{cash_mean}},$$

the standardized deviation of the average cash from the cluster mean.

4.4.1 Flagging threshold

The score orders merchants within each cluster, but a criterion is needed to separate anomalous merchants from the normal ones. Rather than using a fixed absolute value, the threshold is calculated separately for each cluster as the 95th percentile of the corresponding score distribution. The 5% of merchants with the highest scores within each peer group are flagged as anomalous. This specific threshold ensures that merchants are evaluated relative to the typical behavior of comparable peers, in line with the peer comparison principle. In contrast, a single global threshold would be driven mainly by the clusters with the highest score values and could miss anomalies in the other groups. On the training set this procedure flags, by construction, exactly 5% of the merchants in each cluster, for a total of 8,366 flagged merchants.

4.4.2 Application to the test set

An important point is how the score is applied to the test period. The parameters needed to compute the score and the flagging thresholds are all estimated on the training set, and are then applied unchanged to the test set. A test merchant is standardized using the training statistics of its cluster, and flagged if its score exceeds the training threshold for that cluster. Keeping the thresholds fixed means that the proportion of flagged merchants in the test period is no longer constrained to 5% per cluster, but it may rise or fall depending on how the population has shifted between the two periods.

4.5 Classifier Setup

The final stage of the pipeline concerns the training of a supervised classifier on the anomaly flags produced by the unsupervised procedure. The goal is to

obtain a model that assigns each merchant a graduated probability of being anomalous and can be applied to new data while ensuring that the resulting flags correspond to a learnable pattern rather than noise.

4.5.1 Features

The classifier is trained on three merchant-level features derived directly from the raw monetary quantities:

$$cash_mean, \quad \frac{cash_mean}{bon_mean + 1}, \quad \frac{cash_mean}{vol_mean + 1}.$$

The first is the absolute average cash, while the other two express the predominance of cash over bank transfers and over the POS channel, respectively. The addition of one to the denominator avoids division by zero for merchants with no bank transfers or no POS volume. These features are not the within-cluster standardized deviations used to define the pseudo-label but quantities computed on the raw values, so that for the retail and B2B families the classifier must learn the pattern of anomaly from features different from those used for the definition of the labels.

For the cash-only family, however, this separation is not complete, since the pseudo-label of cash-only merchants is defined through their average cash, and *cash_mean* is at the same time one of the classifier's features. There is therefore a partial overlap for these clusters, between the quantity that defined the label and the quantity used to predict it. Cash-only merchants have, by definition, no meaningful POS or transfer activity, so the amount of cash they handle is the only informative signal available. Removing it from the classifier would mean discarding the only information available for judging these merchants. This limitation is confined to the cash-only family, whereas for the retail and B2B clusters the label and the classifier features remain distinct.

4.5.2 Models

Three classifiers of increasing flexibility are trained and compared. Logistic Regression is used as an interpretable baseline, while Random Forest and LightGBM are used to model more complex relationships between the features and the target.

Since the anomalous class represents only 5% of the merchants, the training data are strongly imbalanced. To reduce the effect of this imbalance, all three models are trained with balanced class weights, which assign more weight to the minority class during training. This prevents the models from

focusing mainly on normal merchants and gives greater importance to correctly identifying anomalies.

4.5.3 Evaluation of the models

The classifiers are evaluated using stratified five-fold cross-validation on the training set, with the area under the ROC curve (AUC) as the primary metric, while stratification preserves the 5% proportion of anomalies in each fold. The models are then refitted on the full training set and applied to the test set, where the same metrics are computed. Comparing the cross-validated performance on the training period with the performance on the test period indicates how well each model generalize over time.

The predicted probability that each model assigns to a merchant constitutes the final anomaly score, that is a continuous measure of risk that can be used to rank and prioritize anomalous merchants. The classifier is also evaluated at a specific operating point by setting the threshold at the 95th percentile of the predicted probabilities on the training set and applying the same threshold to the test set. This follows the flagging criterion used in the unsupervised stage. The resulting flags are then compared with the pseudo-labels generated in the previous stage over the same period.

4.6 Internal Consistency Checks

Since no ground truth is available, the flags cannot be directly compared with known anomalous cases. Their reliability must be assessed indirectly with several consistency checks performed alongside the pipeline. These checks evaluate the stability and compactness of the peer group, the reproducibility of the flags across different data partitions, the economic characteristics of flagged merchants, possible geographic and sectoral bias, and the robustness of the results to different threshold definitions. These checks and their outcomes are presented in Chapter 6.

Chapter 5

Results and Experiments

5.1 Clustering Results

The clustering step segments the merchants into eight groups, and rather than describing each single cluster, it is more informative to present them through the three behavioral families into which they fall. Table 5.1 reports the median profile of each cluster on the main variables, and Figure 5.1 shows the corresponding distributions.

	C0	C1	C2	C3	C4	C5	C6	C7
vol_mean	3754	0	2720	0	0	0	4170	9797
bon_mean	688	0	17682	0	0	5869	879	3573
cash_mean	4360	1163	2174	1125	1542	709	5017	11443
cash_ratio	0.5	1.0	0.1	1.0	1.0	0.1	0.5	0.4
pos_ratio	0.4	0.0	0.2	0.0	0.0	0.0	0.4	0.5
bon_ratio	0.1	0.0	0.7	0.0	0.0	0.9	0.1	0.1

Table 5.1: Median profile of each cluster ($K = 8$) on the main variables. Monetary values are in euros.

5.1.1 The three behavioral families

The *cash-only* family comprises clusters C1, C3 and C4. Their defining feature is a cash ratio equal to one and a POS volume equal to zero. The revenue of these merchants is collected almost entirely in cash, with no meaningful use of the POS channel or of bank transfers, and their median monthly cash is

modest, between 1,100 and 1,500 euros. Although their payment profiles are similar, the three clusters differ in geographic and sectoral terms, as shown later.

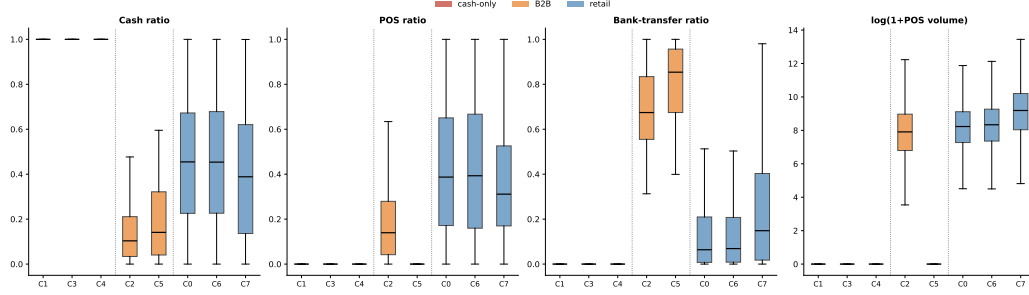


Figure 5.1: Distribution of the main variables across the eight clusters, grouped by behavioral family (cash-only, B2B, retail). Dashed lines separate the three families.

The *B2B* family comprises clusters C2 and C5, characterized by a high bank-transfer ratio and a low cash ratio, showing incoming flows dominated by bank transfers. The two clusters differ in the intensity and the scale of this pattern, more precisely C5 is more based on bank transfers with a ratio of 0.9 against 0.7 for C2, but it operates at a smaller scale with a median transfer amount of about 5,900 euros. C2 that is less based on bank transfers, handles larger amounts, with a median bank transfer exceeding 17,000 euros per month.

The *retail* family comprises clusters C0, C6 and C7, whose merchants have a mixed profile, with a cash ratio around 0.5 and most of the remaining incoming flows collected through the POS channel. These clusters differ in scale, more precisely C0 and C6 are similar to one another, operating at modest levels of POS volume and cash, while C7 is the largest cluster of the family in terms of both POS volume and cash handled.

5.1.2 Geographic and sectoral composition

The clustering features include geographic area and sector of activity, together with the payment ratios and the log-transformed POS volume. It is therefore useful to examine the geographic and sectoral composition of the clusters. Figure 5.2 shows the distribution of macro-areas within each cluster, and Figure 5.3 shows the five most frequent sectors in each cluster.

Within the *cash-only* family, the three clusters show different geographic and sectoral characteristics. C1 is strongly concentrated in the North-East

(around 75%), whereas C3 and C4 are mainly located in Southern Italy, with a substantial presence in the Islands. Their sectoral composition also differs, with C1 and C3 including mainly cash-intensive activities such as restaurants, bars, hairdressing, and clothing retail. In contrast, C4 is composed of non-specialized food retailers, hotels, fuel stations, and pharmacies.

The *B2B* family is the least characterized in both geographic and sectoral terms. Neither C2 nor C5 is concentrated in a specific macro-area, and both are largely composed of the residual sector category, which represents 38% of C2 and 54% of C5. This lack of clear geographic or sectoral profile is consistent with the nature of this family, whose merchants are grouped because they mainly collect revenue through bank transfers. Businesses with this payment pattern can belong to many different sectors and operate in different regions.

The *retail* family shows the greatest differences between its clusters. C0 and C6 have very similar payment and sectoral profiles, mainly including restaurants, bars, hairdressing, and other retail activities, but they differ geographically. C0 is concentrated in the South and the Islands, whereas C6 is located almost entirely in the North. C7 differs from both clusters, since it is more geographically balanced, and includes food retailers, hotels, pharmacies, and fuel stations.

The geographic and sectoral composition of the clusters is not directly used to define them. Merchants are grouped according to their payment behavior, while their sector and geographic characteristics emerge from the target-encoded features. As a result, belonging to a particular geographic area or sector does not directly influence the anomaly score, which is verified in the Chapter 6.

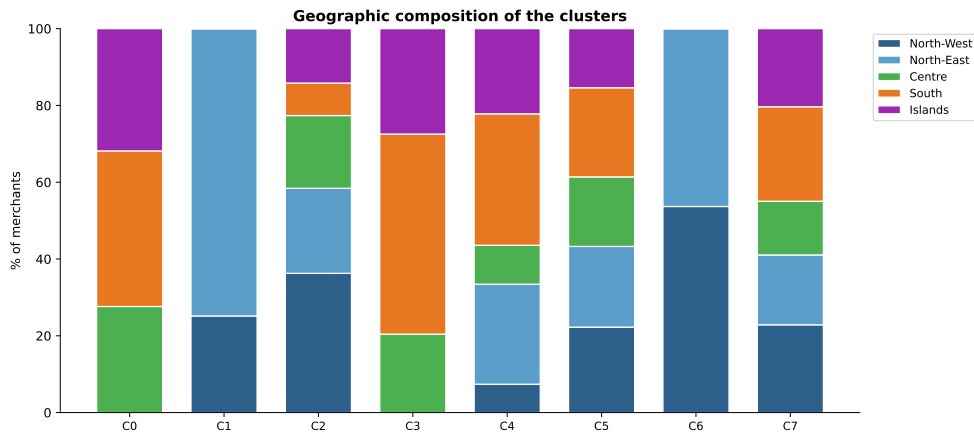


Figure 5.2: Distribution of the five macro-areas within each cluster.

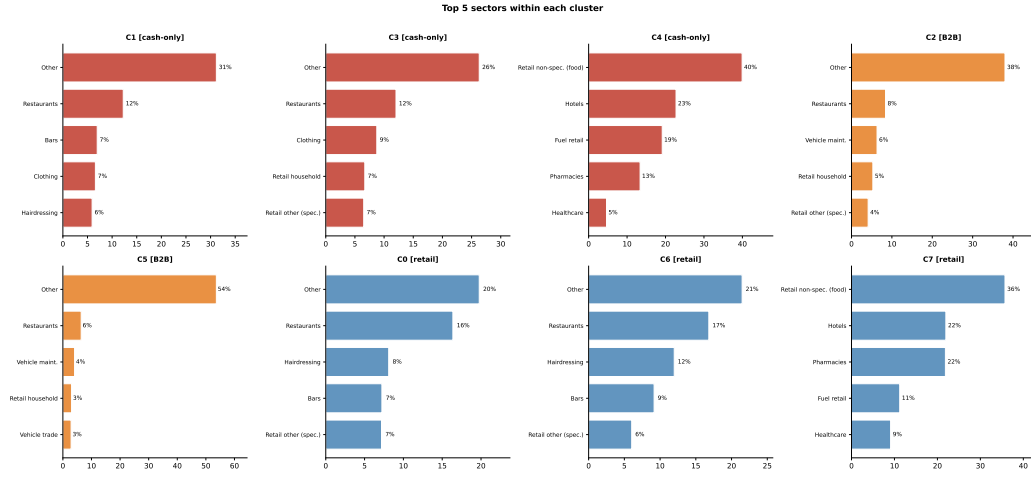


Figure 5.3: Share of the eight most common sectors within each cluster.

5.1.3 Cluster quality

The overall Silhouette coefficient of the partition is 0.370. At the cluster level, all Silhouette values are positive, ranging from 0.21 for the least compact cluster to 0.65 for the most compact one. Cluster stability and compactness are examined in greater detail in the validation chapter.

5.2 Anomaly Detection Results

The flagging rule described in the previous chapter identifies 8,366 anomalous merchants in the training set and 5,533 in the test set. This section examines how these flagged merchants are distributed across the clusters and the characteristics that distinguish them from the remaining merchants.

5.2.1 Distribution of anomalies across clusters

By construction, the threshold flags exactly 5% of the merchants in each cluster on the training set, while on the test set, where the thresholds estimated on the training period are held fixed, the proportion of flagged merchants is free to vary. Figure 5.4 compares the two periods.

The pattern is systematic, that is the three cash-only clusters (C1, C3, C4) rise well above the 5% baseline on the test set, reaching between 8% and 9%, while the B2B and retail clusters remain at or below it, with C5 falling to 3.2%.

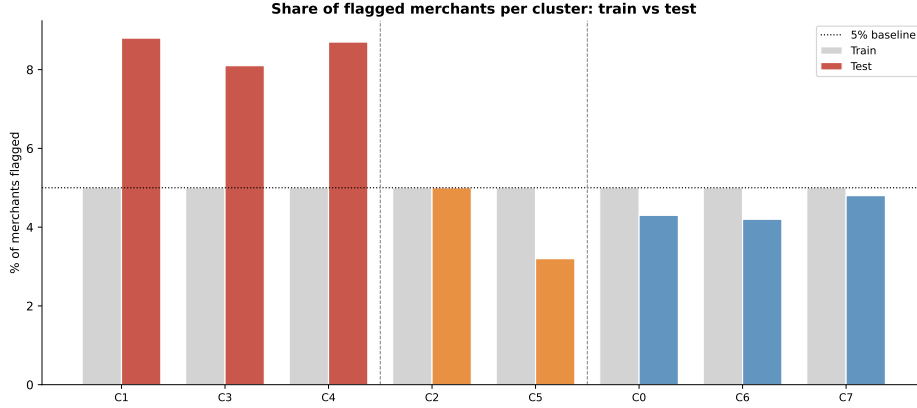


Figure 5.4: Share of flagged merchants in each cluster, on the training and test sets, grouped by behavioral family. Dashed lines separate the three families.

This difference is not caused by the method but reflects a change in the merchant population over time. In the more recent period, cash-only merchants handle larger amounts of cash on average, so more of them exceed the threshold calibrated on the training period.

5.2.2 Profile of anomalous merchants

Comparing flagged and non-flagged merchants within each cluster family shows that the anomaly is different across the three families, reflecting the way the score is defined. Table 5.2 reports the median profiles for one representative cluster per family on the test set.

In the *retail* family, the main difference is in the composition of payment flows. In cluster C0, for example, flagged merchants have a median cash ratio of 0.93, compared with 0.43 for the remaining merchants, while their median POS ratio is 0.03, compared with 0.40. Their payment inflows are almost entirely cash based, while the other merchants in the cluster make much greater use of the POS channel.

In the *B2B* family, the flagged merchants in cluster C2 have a median cash ratio of 0.40, compared with 0.08 among the remaining merchants. This indicates a greater reliance on cash in a group where bank transfers are the dominant payment channel.

In the *cash-only* family, the difference is based on the amount of cash handled, since flagged and non-flagged merchants have almost identical payment compositions, with a cash ratio close to 1 in both cases. In cluster C1, the median monthly cash collected by flagged merchants is approximately

Family	Feature	Anomalous	Normal	$\Delta\%$
Retail (C0)	cash_mean	82,256	4,646	+1671%
	cash_ratio	0.93	0.43	+119%
	pos_ratio	0.03	0.40	-93%
	vol_mean	1,948	4,634	-58%
B2B (C2)	cash_mean	19,150	2,305	+731%
	cash_ratio	0.40	0.08	+384%
	pos_ratio	0.02	0.10	-77%
	vol_mean	844	2,907	-71%
Cash-only (C1)	cash_mean	156,708	1,732	+8946%
	cash_ratio	0.98	1.00	-2%
	pos_ratio	0.00	0.00	0%

Table 5.2: Median profiles of anomalous and normal merchants, for one representative cluster per family (test set). Monetary values are in euros.

157,000, compared with about 1,700 for the other merchants, a difference of nearly two orders of magnitude. For these clusters, the anomaly is associated with the absolute amount of cash handled rather than with the composition of payment flows. Figure 5.5 shows these differences across the three families.

Another interesting result concerns the size of the flagged merchants in the retail and B2B families. Although one might expect merchants with unusually high cash payments to be among the largest businesses, the data suggest otherwise. In cluster C0, flagged merchants have a median POS volume of about 1,950, compared with 4,630 for the non-flagged merchants. A similar pattern is observed in cluster C2, where flagged merchants also have a lower POS volume than their peers.

These merchants are therefore not distinguished by the overall scale of their activity, but by the composition of their payment flows. Their reliance on cash is unusually high relative to comparable merchants, even if their business volume is not.

5.2.3 Overall profile of the flagged merchants

Considering all clusters together, flagged merchants show a consistent pattern. Their median monthly cash amount is approximately €87,000, compared with about €2,600 for the overall merchant population, while their

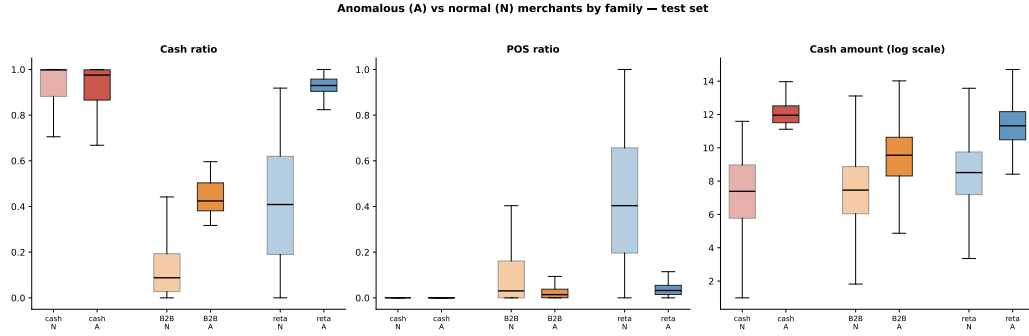


Figure 5.5: Distribution of the cash ratio, POS ratio, and cash amount for anomalous (A) and normal (N) merchants, by behavioral family, on the test set. The cash amount is shown on a logarithmic scale. In the retail and B2B families the flagged merchants are distinguished by the composition of their payments; in the cash-only family, where the composition is nearly identical, they are distinguished by the absolute level of cash.

median cash ratio is 0.94 against 0.60. This is consistent with the patterns observed within the individual cluster families, where the flagged merchants are not necessarily the largest businesses, but those that rely more heavily on cash than comparable merchants. Overall, the merchants identified by the method combine three characteristics, namely a high absolute amount of cash, a payment composition strongly dominated by cash, and a POS volume that is not higher than that of the average merchant.

5.3 The Anomaly Scoring System

The unsupervised procedure assigns each merchant a binary label, identifying it as either anomalous or non-anomalous. Training a classifier on these labels produces, for each merchant, a probability of being anomalous rather than a simply binary outcome. This section examines these predicted probabilities on the test set and evaluates whether they provide a meaningful ranking of merchants according to their anomaly risk.

5.3.1 Separation between flagged and normal merchants

The predicted probabilities show a clear separation between flagged and non-flagged merchants, specifically, the median predicted probability is 0.986 for flagged merchants and 0.000 for the remaining merchants. Even at the 90th percentile, non-flagged merchants reach a probability of only 0.019. Fig-

ure 5.6 shows the two distributions on a logarithmic count scale, with most of the mass concentrated at the two extremes. The normal merchants are concentrated near zero, whereas flagged merchants are concentrated near one. These results indicate that the classifier reproduces the distinction defined by the unsupervised procedure with a little overlap between the two groups.

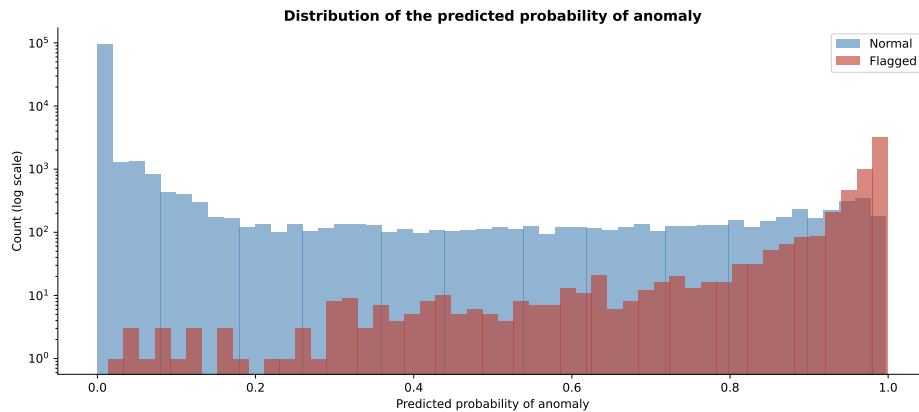


Figure 5.6: Distribution of the predicted probability of anomaly for normal and flagged merchants on the test set, on a logarithmic count scale. The overlap between the two groups corresponds to the borderline cases.

5.3.2 A graduated measure of risk

The predicted probability does more than separate flagged and non-flagged merchants, within the flagged groups it also provides a ranking according to the severity of the anomaly. Table 5.3 illustrates this pattern by grouping merchants into score bands.

Score band	Merchants	Median cash	Median POS ratio
Very high (> 0.9)	6,219	73,583	0.017
High (0.7–0.9)	1,804	30,196	0.041
Medium (0.5–0.7)	1,268	29,916	0.077
Low (0.3–0.5)	1,224	22,225	0.051

Table 5.3: Merchants grouped by score band (test set), with the median cash amount (in euros) and median POS ratio in each band.

As the predicted probability decreases, the median cash amount also decreases, while the POS ratio increases. Merchants in the highest score bands

handle the largest amounts of cash and make almost no use of the POS channel, whereas those in the lower bands show progressively less extreme payment profiles. Therefore, the predicted probability provides more than a binary classification, it also reflects the intensity of the anomaly, making it a useful criterion for prioritizing merchants for further investigation.

5.3.3 Extreme cases by family

To illustrate the types of merchants identified by the method, Table 5.4 reports the highest scoring merchant from each cluster family, where the three examples represent different forms of anomalous behavior.

The merchant from the cash-only family handles a very large amount of cash and does not use the POS channel. The retail example is characterized by an even larger cash amount, with about 95% of its payment inflows collected in cash despite the presence of an active POS terminal. The B2B example also handles a large amount of cash, but its cash ratio is around 0.55, an unusually high value for a group in which bank transfers are the dominant payment channel. Although the three merchants differ from one another, each one of them is anomalous relative to the merchants in its own cluster family.

Family	Cash amount	Cash ratio	POS ratio
Cash-only	271,378	1.00	0.00
Retail	376,206	0.95	0.05
B2B	199,460	0.54	0.00

Table 5.4: One representative high-scoring merchant per behavioral family (test set). Cash amounts are in euros.

5.4 Classifier Performance

This section reports the performance of the supervised classifier trained on the anomaly flags, how well it generalizes from the training to the test period, and shows which features drive its predictions.

5.4.1 Discriminative performance

Three classifiers, specifically logistic regression, Random Forest, and LightGBM are trained and compared. Table 5.5 reports their performance, mea-

sured by the area under the ROC curve, both in cross-validation on the training set and on the test set.

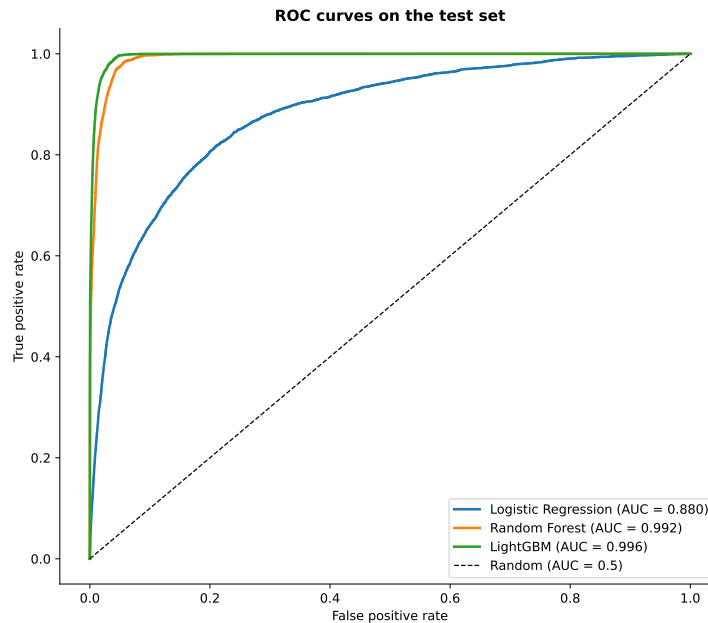


Figure 5.7: ROC curves of the three classifiers on the test set. The dashed diagonal corresponds to a random classifier.

All three models separate the two classes well, but with a clear ordering, specifically the linear baseline reaches an AUC of 0.88 on the test, while the two tree based ensembles perform better, with Random Forest at 0.992 and LightGBM at 0.996. The ROC curves are shown in Figure 5.7, where the ensemble models lie almost in the upper left corner, indicating a high detection rate combined with a low false positive rate.

Model	AUC (CV)	AUC (test)	Gap
Logistic Regression	0.905	0.880	-0.025
Random Forest	0.994	0.992	-0.002
LightGBM	0.997	0.996	-0.001

Table 5.5: Discriminative performance of the three classifiers, in cross-validation on the training set and on the test set. The gap is the difference between the two.

5.4.2 Generalization over time

An important question concerns the relationship between the classifier features and the anomaly labels. As discussed earlier, cash amount contributes to the definition of the label in the cash-only clusters, raising the possibility that the classifier simply reproduces the labels rather than learning a pattern that generalizes over time. Comparing cross-validation and test performance helps assess this issue, since if the model is relying mainly on memorization, a more substantial drop in performance would be expected when moving from the training period to the test one. Instead, the performance remains almost unchanged, as reported in Table 5.5, for LightGBM the AUC decreases only from 0.997 to 0.996, while for Random Forest it decreases from 0.994 to 0.992. Even logistic regression, despite its lower overall performance, loses only about two and a half percentage points. These results suggest that the models maintain the same discriminative ability when applied to data collected in a different period, indicating that the classifiers capture stable pattern rather than simply reproducing the training labels. Consequently, the overlap between the cash amount and the label definition in the cash-only clusters does not appear to produce artificially optimistic performance estimates.

5.4.3 Feature importance

The classification models are trained using three features, all derived from the cash handled by the merchant, namely the absolute cash amount and its ratios to bank transfers and to POS volume. Although all three features are related to cash, their importance differs across models, as shown in Table 5.6.

In logistic regression, the absolute cash amount has the largest standardized coefficient, indicating that it contributes more than the other two features. The tree-based models shows a more balanced distribution of feature importance, specifically in LightGBM the ratio of cash to bank transfers provides the largest contribution, whereas permutation importance, which measures the decrease in predictive performance after randomly shuffling a feature, identifies the ratio of cash to POS volume as the most influential variable.

Overall, the models rely on both the absolute amount of cash and its relationship with the other payment channels, indicating that anomalous merchants are identified not only by the amount of the cash they handle, but also by the share of cash relative to other payment channels.

Feature	LogReg	Random Forest	LightGBM	Permutation
cash_mean	0.747	0.428	0.203	0.279
cash_over_bon	0.123	0.225	0.434	0.146
cash_over_pos	0.130	0.346	0.363	0.575

Table 5.6: Relative feature importance across the three models. For logistic regression the values are the normalized absolute standardized coefficients; for the tree-based models they are the built-in importance; the last column reports permutation importance.

5.4.4 Classification at a fixed threshold

Although the AUC evaluates performances across all possible classification thresholds, the classifier can also be assessed at a specific threshold. Using the same 5% threshold adopted by the unsupervised procedure, LightGBM achieves a precision of 0.854, a recall of 0.85, and an F1-score of 0.852 on the test set. The corresponding confusion matrix is reported in Figure 5.8, showing that most flagged merchants are correctly classified, while the number of false positives and false negatives remain limited.

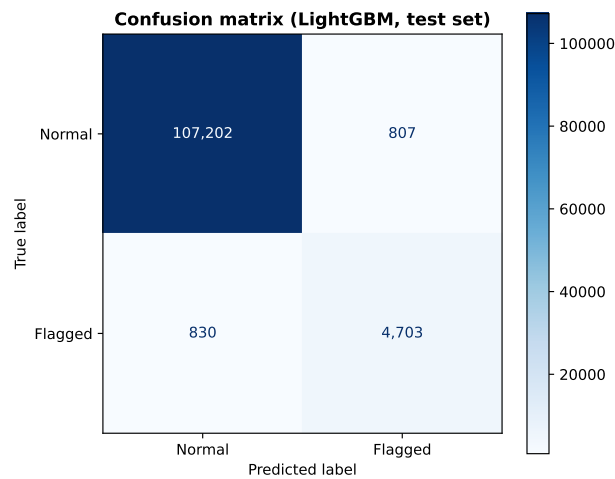


Figure 5.8: Confusion matrix of the LightGBM classifier on the test set, at a threshold flagging the top 5% by predicted probability. Rows are the true classes, columns the predicted ones.

Chapter 6

Methodology Validation

The method proposed in this thesis produces anomaly flags without using ground-truth labels. Therefore, there is no set of known anomalous merchants that can be used to directly evaluate the results. The method is evaluated through a series of validation analyses that examine different aspects of the methodology. These include the stability and compactness of the peer groups, the reproducibility of the anomaly flags, their economic coherence, possible geographic or sectoral biases, and the robustness of the results to a different definition of the flagging threshold.

None of these analyses can, on its own, show that the flagged merchants are truly anomalous, but together they provide evidence that the method produces stable and meaningful results.

6.1 Stability and compactness of the clusters

The clusters define the peer groups used to evaluate merchants, so their quality is important for the reliability of the overall pipeline. This section considers two aspects: the stability of the clustering across different random initializations and the compactness of the resulting clusters.

K-Means starts with randomly initialized centroids, so different initializations can produce different partitions. To assess its stability, the algorithm is run five times with different random seeds, and each partition is compared with the final solution using the Adjusted Rand Index (ARI), which measures the agreement between two partitions and takes a value of 1 when they are identical.

The results are reported in Table 6.1. The ARI is never below 0.997 and has an average value of 0.998, showing that the clustering remains almost unchanged across different initializations. The eight clusters therefore reflect

a stable structure in the data rather than a random partition.

Seed	Adjusted Rand Index
0	0.9968
1	0.9973
7	1.0000
12	0.9998
99	0.9971
Mean	0.9982
Minimum	0.9968

Table 6.1: Adjusted Rand Index (ARI) comparing the final partition with five partitions obtained using different random initializations.

Cluster compactness is evaluated using the Silhouette coefficient, computed separately for each cluster and shown in Figure 6.1. All eight clusters have a positive mean Silhouette value, ranging from 0.21, for the least compact cluster, to 0.65 for the most compact one, suggesting that none of the clusters is poorly defined. The cash-only clusters have the highest Silhouette values, reflecting the similarity of their payment profiles. The retail clusters are less compact, but their median values are still positive.

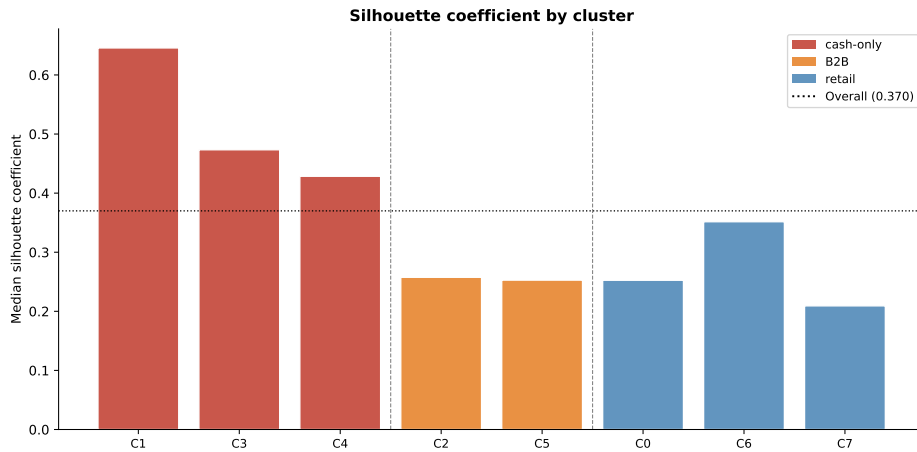


Figure 6.1: Median silhouette coefficient of each cluster, grouped by behavioral family. All values are positive; the dotted line marks the overall silhouette of the partition. Dashed lines separate the three families.

As a further check, each merchant is compared with the centroids of all

eight clusters, and in every case the closest centroid corresponds to the cluster to which the merchant is assigned. Thus, no merchant would be assigned to a different cluster based on the nearest-centroid rule. Overall, these results show that the clustering is stable and that the resulting clusters are well separated.

6.2 Reproducibility of the flags

The anomaly flags are based on cluster specific estimates of the mean, standard deviation, and flagging threshold. Their reproducibility is assessed through a bootstrap analysis to check whether the same merchants are consistently identified as anomalous across different samples from the training population. A total of 200 bootstrap samples are drawn from the training set, and for each sample the flagging rule is estimated within each cluster by calculating the mean, standard deviation, and threshold, while keeping the original cluster assignments fixed.

Among the merchants flagged in the original training set, 96.2% are flagged in at least 90% of the bootstrap samples. Figure 6.2 shows the distribution of flagging frequencies for the flagged merchants. Most merchants are concentrated near 1, with only a small tail towards lower values. This indicates that almost all flags are highly reproducible, while the few cases with lower frequencies correspond to the borderline merchants whose anomaly scores lie just above the flagging threshold.

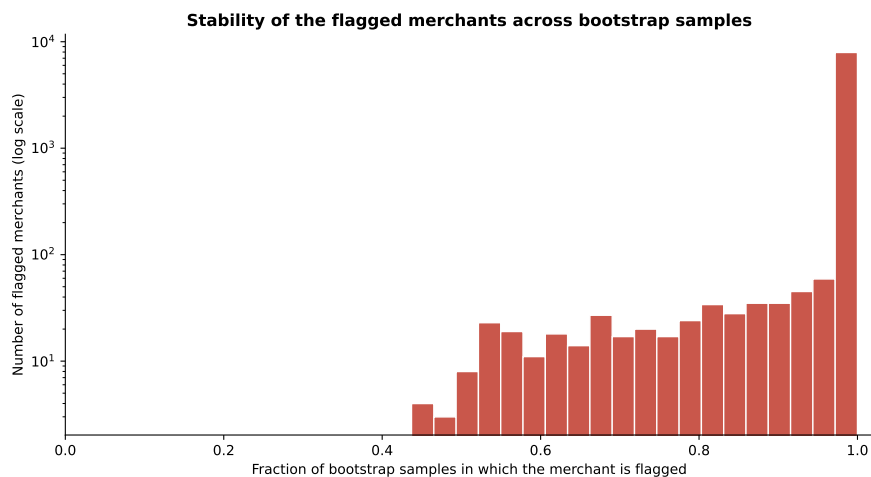


Figure 6.2: Distribution of flagging frequencies across 200 bootstrap samples, restricted to the merchants flagged by the rule on the full training set.

A similar pattern is observed for merchants that are not flagged, since 99.8% remain unflagged in at least 90% of the samples. Only 0.4% of merchants have a flagging frequency between 10% and 90%, indicating that their classification is less stable.

6.3 Economic Coherence of the flags

The flagged merchants should show characteristics consistent with the phenomenon investigated in this thesis. Since an anomaly is defined as an unusually high reliance on cash, flagged merchants are expected to have a high cash ratio and a low POS ratio. Merchants that collect most of their payments through POS would therefore not be consistent with this definition.

Three aspects are considered: the presence of contradictory flags, the differences between flagged and non-flagged merchants, and the payment profiles of the flagged merchants.

No contradictory cases are found since, among the 8,366 flagged merchants, none has a POS ratio above 0.5, and none of the merchants in the retail or B2B families has a cash ratio below 0.3. The rule therefore does not flag merchants whose payment profiles are not consistent with the type of anomaly being investigated.

The differences between flagged and non-flagged merchants also follow the expected pattern. As shown in Table 6.2, flagged merchants have higher cash ratios, lower POS ratios, and substantially larger cash amounts than non-flagged merchants. Overall, 78.3% of flagged merchants have a cash ratio above 0.7 and a POS ratio below 0.1, showing that most of these merchants have payment profiles strongly concentrated in cash.

Feature	Flagged	Normal
<code>cash_ratio</code>	0.937	0.572
<code>pos_ratio</code>	0.009	0.081
<code>cash_mean</code>	86,777	2,253

Table 6.2: Median profile of flagged versus normal merchants on the training set. Cash amounts are in euros.

6.4 Absence of Geographic and Sectoral Bias

The anomaly flags should not be concentrated in specific geographic areas or sectors, as this could indicate that the method is capturing structural differences between regions or industries rather than individual anomalies. This is assessed first on the test set, where the thresholds are fixed and the proportion of flagged merchants can vary across areas and sectors, while the results for the training set are then reported for comparison.

6.4.1 Geographic Bias

In the test set, the median anomaly score of flagged merchants is very similar across the five macro-areas, ranging from 2.94 to 3.01. The share of flagged merchants also varies only slightly, from 3.9% to 5.5%, with no area showing a consistently higher share of flags, as reported in Table 6.3.

Among the five macro-areas, the South deserves particular attention because it represents the case where a geographic bias would be most likely to emerge. Cash usage tends to be more prevalent in Southern Italy, since the sectoral composition differs from the rest of the country with a higher concentration of small cash-intensive activities. The results, however, show the opposite case. The South is not overrepresented, since its median anomaly score is similar to those of the other macro-areas, while its share of flagged merchants is the lowest. The training set shows the same pattern, since the South has both the lowest median anomaly score (1.99 compared with values between 2.86 and 3.01 in the other macro-areas) and the lowest share of flagged merchants.

These results suggest that the geographic differences in the data do not translate into a geographic bias in the anomaly score. This does not mean that anomalies are absent in the South, but only that the method does not systematically flag Southern merchants more often because of their geographic location and payment composition.

6.4.2 Sectoral Bias

The sectoral analysis leads to the same conclusion, as shown in Table 6.4.

In the test set, the median anomaly score is very similar across the most common sectors, ranging from 2.97 to 3.04. The share of flagged merchants varies more, from 2.4% to 7.8%. This difference is mainly due to the composition of the sector across the behavioral families. Sectors with more cash-intensive activities naturally contain a larger share of cash-dominant merchants. However, after standardizing the score within each cluster, no sector

Macro-area	Median score	% flagged
North-West	2.99	5.0
North-East	3.01	4.7
Centre	2.99	5.5
South	2.94	3.9
Islands	3.00	5.3

Table 6.3: Median anomaly score and share of flagged merchants by macro-area in the test set.

Sector	Median score	% flagged
Other	2.97	6.0
Restaurants	2.97	2.5
Hairdressing & beauty	3.02	2.4
Bars	3.00	6.0
Retail – food & beverages	2.97	7.8
Vehicle maintenance	3.03	6.0
Retail – other	3.04	3.7
Retail – household goods	3.04	6.1

Table 6.4: Median anomaly score and share of flagged merchants for the most common sectors in the test set.

shows systematically higher anomaly scores than the others.

6.5 Robustness of the threshold definition

The final validation compares the use of a separate threshold for each cluster with a single threshold applied to the entire population. As discussed in the methodology chapter, anomaly scores are not directly comparable across the different cluster families because they are based on different definitions. The two approaches are compared to assess how this choice affects the resulting flags.

The two approaches flag almost the same number of merchants, precisely 8,366 with the within-cluster threshold and 8,363 with the global threshold. However, the merchants selected by the two methods are different, and the

overlap between the two sets is limited, with a Jaccard index of 0.42. The main differences concern the distribution of flags across the three behavioral families. As shown in Table 6.5 and Figure 6.3, the global threshold almost eliminates the cash-only family, reducing its share of flagged merchants from 35.9% to 1.2%. At the same time, the share of the retail family increases from 44.7% to 82.2%. This is because the score of the three families are on different scales, since the cash-only score is based on the absolute cash amount, while the retail and B2B scores are based on payment ratios. Using a single threshold therefore favors families characterized by higher scores, while making it more difficult for observations from other families to be flagged.

Family	% within-cluster	% global
Retail	44.7	82.2
B2B	19.4	16.5
Cash-only	35.9	1.2

Table 6.5: Composition of the flagged set by behavioral family, under the within-cluster threshold and under a single global threshold.

The merchants flagged exclusively by the two approaches also differ in terms of cash amounts. Those flagged by the within-cluster threshold have median cash amount of about €117,000, compared with about €25,000 for those flagged only by the global threshold. The global threshold therefore misses merchants that handle unusually large amounts of cash relative to their peers, while flagging other merchants with much smaller amount. Using a separate threshold within each peer group avoids this problem and allows the different behavioral families to be evaluated according to their own distributions.

6.6 Summary of Validation

None of the validation analysis presented in this chapter can, on its own, show that the flagged merchants are truly anomalous. Without ground-truth labels this cannot be directly verified, but each analysis focuses on a different aspect of the methodology, providing several ways to assess its reliability. The clustering is stable across different random initializations and produces compact peer groups. The flagging procedure is reproducible, with most flagged merchants being identified consistently across different samples of the data. The flagged merchants are also consistent with the definition

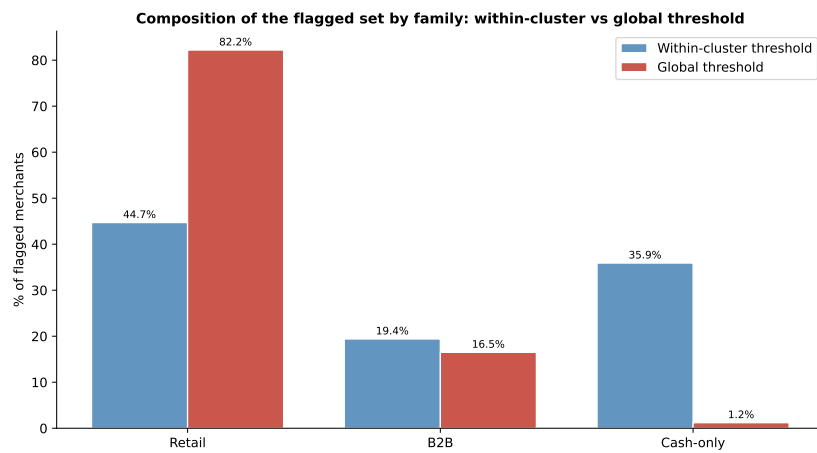


Figure 6.3: Composition of the flagged set by behavioral family under the two threshold definitions.

of a cash-based anomaly. No systematic geographic or sectoral bias is found when merchants are compared within their peer groups. Finally, using a separate threshold for each cluster avoids the distortions caused by a single global threshold.

Overall, these results show that the method produces stable, reproducible, and consistent results. However, confirming that the flagged merchants are actually involved in undeclared activity would require external evidence.

Chapter 7

Conclusion and Future Work

7.1 Summary of the work

This thesis addresses the problem of identifying merchants whose payment behavior differs from that of similar businesses, in a setting where no labeled data are available and the phenomenon of interest cannot be observed directly. Payment data provide a detailed and continuous record of how merchants receive their revenues, making it possible to identify signals that can be used to detect anomalous profiles. A higher reliance on cash than that observed among similar merchants may therefore indicate possible under-reporting of revenues. The methodology developed in this thesis aims to identify these cases using a transparent and interpretable approach. The methodology is divided into three sequential stages.

In the first stage, merchants are grouped into peer groups using K-Means clustering based on five features that describe the size and composition of each merchant's payment flows and its economic context. These features are the log-transformed POS volume, the cash ratio, the bank transfer ratio, and two target encoded variables representing the geographic area and the sector of activity. The eight resulting clusters can be further grouped into three behavioral families based on their payment profiles. The first family includes cash-only merchants, who receive almost all of their payments in cash. The second family consists of B2B merchants, whose payment flows are mainly based on bank transfers. The third family includes retail merchants characterized by a combination of cash and POS payments.

In the second stage, an anomaly score is calculated separately for each cluster. For the retail and B2B families, the score combines the standardized cash and POS ratios to capture deviations in the payment composition of a merchant relative to its cluster. For the cash-only family, where merchants

have very similar payment composition, the score is instead based on the absolute amount of cash handled. A merchant is flagged as anomalous when its score exceeds the 95th percentile of the corresponding cluster distribution, estimated using the training set and held fixed when applied to the test set.

Each merchant is evaluated relative to the other merchants in the same peer group, rather than against a single benchmark for the whole population. This is important because the merchants are heterogeneous and their payment behavior varies considerably across groups.

The flags obtained from the unsupervised rule are then used as pseudo-labels for the supervised stage. Logistic regression, Random Forest, and LightGBM are trained using the same payment features to learn the characteristics associated with the flags and assign each merchant a probability of being anomalous. This probability is then used as the final risk score to rank merchants for further examination. Among the three models, LightGBM performs best, with an AUC of 0.996 on the test set, and a difference of only 0.001 from the cross-validated results, showing that the model maintains similar performance on new observations.

The absence of ground-truth labels prevents the use of standard accuracy measures to assess whether the identified merchants are truly anomalous. The results are therefore examined through a separate set of validation checks. These include the stability of the peer groups across random initializations, the reproducibility of the flags across bootstrap samples, the economic characteristics of flagged merchants, possible geographic and sectoral biases, and the sensitivity of the results to an alternative flagging threshold.

7.2 Answers to the research questions

The thesis examines two research questions. The first is whether, in the absence of labeled data, it is possible to define a notion of anomaly that is both economically meaningful and statistically robust. The second is whether the resulting flags can be reproduced by a classification model and transformed into a scoring tool.

The within-cluster anomaly score identifies merchants with payment profiles that are consistent with the economic meaning of the anomaly. In particular, none of the flagged merchants has a POS-dominated profile that would be inconsistent with a cash-based anomaly, while 78% show a clearly cash-dominated payment profile. The results are also stable across different samples of the data. In the bootstrap analysis, 96.2% of the flagged merchants are flagged in at least 90% of the resampled datasets. The geographic and sectoral analyses provide further evidence that the method does not

systematically favor particular regions or industries. Overall, these results support the use of the proposed definition of anomaly as an economically meaningful and stable measure.

The results also support the second question. The classifier is able to reproduce the anomaly pattern and apply it to new observations. The small difference between cross-validated and test performance suggests that the model is not simply memorizing the training labels, but is learning a pattern that remains stable on data from a later period. The predicted probability also provides a continuous risk score, which can be used to rank merchants and prioritize cases for further investigation.

7.3 Limitations

Several limitations should be considered when interpreting the results of the methodology. The main limitation is the absence of ground-truth labels, which means that the validation checks cannot determine whether the flagged merchants are actually concealing revenue. Although the flags are stable, reproducible, and economically coherent, they should be interpreted as signals for further investigation rather than as evidence of revenue concealment.

Another limitation is specific to the cash-only clusters. For these merchants, the anomaly score is based on the absolute amount of cash handled rather than on payment composition. Absolute cash amounts are affected by nominal changes such as inflation and growth in business volume, which may explain the increase in the share of flagged cash-only merchants in the test set, from 5% to approximately 8%. As nominal cash amounts increase over time, more merchants may exceed a threshold calibrated on the earlier period, but this does not necessarily indicate an increase in concealment, since it may partly reflect the growth in nominal revenues. The ratio-based scores used for the retail and B2B families are not affected by this issue because they depend on the composition of payment flows rather than on their absolute level.

A third limitation concerns the choice of the number of clusters. Neither the Silhouette coefficient nor the Davies-Bouldin index identifies a clear optimum over the range considered. The choice of $K = 8$ is mainly based on interpretability and on the resulting alignment with the three behavioral families.

Finally, the partial overlap between the features used by the classifier and the definition of the anomaly score for cash-only merchants represents a structural form of data leakage. This overlap is difficult to avoid because there is no alternative signal that provides enough information for these mer-

chants. The small difference between cross-validated and test performance suggests that the leakage does not result in a large loss of generalization.

7.4 Future Work

There are several ways in which the methodology could be further developed. The most important improvement would be to extend the pipeline to a broader range of behavioral anomalies. This thesis focuses on a single type of anomaly, specifically an unusually high reliance on cash compared with similar merchants. The same approach of using unsupervised clustering to define peer groups, within-cluster z-scores to measure deviations, pseudo-labels to identify potentially anomalous cases, and a supervised classifier to obtain a continuous risk score, could also be applied to other types of anomalies. Three extensions are particularly relevant, each addressing a different business problem.

The first is *MultiPOS*, which concerns merchants that may use POS terminals from different providers within the same business, while only transactions processed through the BPER POS are observed. These merchants may have an unusually high share of bank transfers compared with similar businesses, for example when sales processed through another POS provider are transferred to the merchant's bank account. Another possible indicator is an unusually low level of POS revenue compared with total inflows, after accounting for the merchant's sector, province, and business size.

The second extension is *About-to-Leave*, which would identify merchants that may be gradually moving their payment activity from BPER POS to another provider. A possible indicator would be a negative within-cluster z-score for the trend in POS revenues or transaction volumes, together with stable or increasing total revenues or bank transfer volumes. This anomaly could be relevant because identifying a merchant before the migration allows the company to intervene and potentially prevent the loss of the customer.

The third extension is *Volatility*, which concerns merchants whose POS revenues vary substantially over time compared to similar businesses, since high variability is often associated with a persistent decline in revenues. This is also relevant from a credit risk perspective, as unstable or declining revenues can reduce the merchant's ability to repay outstanding credit facilities.

All three extensions could follow the same procedure used in this thesis, and the resulting scores could be combined to obtain a more complete risk profile for each merchant.

A final direction would be to consider how the anomaly score changes over time. In the current analysis, each merchant receives a single score

based on its average behavior over the observed period. A dynamic version of the method could track the score month by month and identify changes in payment behavior, providing an additional signal that would not be captured by a static version of the methodology.

Bibliography

- Al-Lahham, A., Tastan, N., Zaheer, M. Z., and Nandakumar, K. (2024). A coarse-to-fine pseudo-labeling (c2fpl) framework for unsupervised video anomaly detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 6793–6802.
- Arthur, D., Vassilvitskii, S., et al. (2007). k-means++: The advantages of careful seeding. In *Soda*, volume 7, pages 1027–1035.
- Bolton, R. J. and Hand, D. J. (2002). Statistical fraud detection: A review. *Statistical Science*, 17(3):235–255.
- Bolton, R. J., Hand, D. J., et al. (2001). Unsupervised profiling methods for fraud detection. *Credit scoring and credit control VII*, 22.
- Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5–32.
- Breunig, M. M., Kriegel, H.-P., Ng, R. T., and Sander, J. (2000). Lof: Identifying density-based local outliers. In *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data*, pages 93–104.
- Chandola, V., Banerjee, A., and Kumar, V. (2009). Anomaly detection: A survey. *ACM Computing Surveys*, 41(3):1–58.
- Davies, D. L. and Bouldin, D. W. (1979). A cluster separation measure. *IEEE transactions on pattern analysis and machine intelligence*, (2):224–227.
- De Roux, D., Perez, B., Moreno, A., Villamil, M. d. P., and Figueroa, C. (2018). Tax fraud detection for under-reporting declarations using an unsupervised machine learning approach. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 215–222.
- Fawcett, T. (2006). An introduction to roc analysis. *Pattern recognition letters*, 27(8):861–874.

- Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232.
- ISTAT (2025). L’economia non osservata nei conti nazionali. anni 2020–2023. Statistiche report, Istituto Nazionale di Statistica (ISTAT).
- Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., and Liu, T.-Y. (2017). Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems*, 30.
- Liu, F. T., Ting, K. M., and Zhou, Z.-H. (2008). Isolation forest. In *2008 Eighth IEEE International Conference on Data Mining*, pages 413–422. IEEE.
- Lloyd, S. (1982). Least squares quantization in pcm. *IEEE transactions on information theory*, 28(2):129–137.
- Micci-Barreca, D. (2001). A preprocessing scheme for high-cardinality categorical attributes in classification and prediction problems. *ACM SIGKDD explorations newsletter*, 3(1):27–32.
- Repubblica Italiana (2022). Decreto-legge 30 aprile 2022, n. 36: Ulteriori misure urgenti per l’attuazione del piano nazionale di ripresa e resilienza (pnrr). Gazzetta Ufficiale, Serie Generale n. 100 del 30 aprile 2022, convertito con modificazioni dalla Legge 29 giugno 2022, n. 79.
- Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65.
- Savić, M., Atanasijević, J., Jakovetić, D., and Krejić, N. (2022). Tax evasion risk management using a hybrid unsupervised outlier detection method. *Expert Systems with Applications*, 193:116409.
- Weston, D. J., Hand, D. J., Adams, N. M., Whitrow, C., and Juszczak, P. (2008). Plastic card fraud detection using peer group analysis. *Advances in Data Analysis and Classification*, 2(1):45–62.