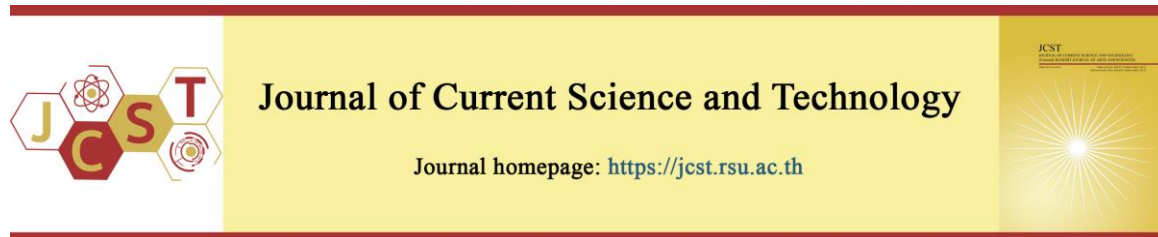


Cite this article: Asfi, M., Warsito, B., & Wibowo, A. (2026). COSHAP: An explainable AI framework integrating conformal prediction and SHAP for Uncertainty-aware consumer segmentation. *Journal of Current Science and Technology*, 16(4), Article 213. <https://doi.org/10.59796/jcst.V16N4.2026.213>



COSHAP: An Explainable AI Framework Integrating Conformal Prediction and SHAP for Uncertainty-Aware Consumer Segmentation

Marsani Asfi^{1,2,*},
Budi Warsito³,
Adi Wibowo⁴,

¹Doctoral Program of Information System, School of Postgraduate Program Studies, Diponegoro University, Semarang Indonesia

²Department of Information Systems, Catur Insan Cendekia University, Cirebon, Indonesia

³Department of Statistics, Faculty of Science and Mathematics, Diponegoro University, Semarang Indonesia

⁴Department of Informatics, Faculty of Science and Mathematics, Diponegoro University, Semarang Indonesia

*Corresponding author; E-mail: marsani.asfi@cic.ac.id

Received 16 March 2026; Revised 23 April 2026; Accepted 22 May 2026; Published 15 September 2026

Abstract

Data-driven decision-making in digital marketing is often constrained by the lack of transparency and uncertainty estimation in conventional machine learning models. This study proposes COSHAP (*Conformal Segmentation with Hierarchical Attribution Paths*), a unified framework that integrates Conformal Prediction (CP) and SHAP to model uncertainty while enhancing interpretability in customer segmentation. Using the *Online Retail II* dataset, segmentation was performed based on Recency, Frequency, Monetary (RFM), Tenure, and Average Basket Size features using K-Means, which was subsequently approximated using supervised models (Random Forest, XGBoost, Naïve Bayes, and SVM). CP was applied to generate prediction sets with guaranteed statistical coverage, while SHAP analysis was extended by conditioning feature attributions on prediction set membership through *Hierarchical Attribution Paths*. Experimental results demonstrated that the XGBoost model achieved the highest approximation accuracy of 97.36%. At a 95% confidence level, the framework achieved empirical coverage consistent with the statistical target for the ensemble models, with an average prediction set size close to 1.0, indicating high certainty in the majority of cases. However, a subset of customers at decision boundaries yielded multi-label predictions, accurately reflecting behavioral ambiguity. SHAP analysis identified *Recency* and *Frequency* as the primary drivers of segmentation, whereas COSHAP successfully uncovered the push-pull feature dynamics that trigger prediction uncertainty. The COSHAP framework effectively transforms point-based segmentation into a reliable, calibrated, and transparent decision-support system.

Keywords: conformal prediction, consumer segmentation, explainable artificial intelligence, machine learning, SHAP, RFM analysis, uncertainty quantification

1. Introduction

Data-driven decision-making has become a central pillar of modern retail and digital marketing, where organizations increasingly leverage machine learning to analyze large-scale transactional and behavioral data to identify meaningful customer

segments. Customer segmentation enables firms to group individuals based on similarities in purchasing behavior, engagement patterns, and preferences, thereby supporting targeted marketing strategies and personalized services (Alves Gomes & Meisen, 2023). Consequently, predictive models are widely adopted to

classify customers into behavioral categories such as loyal customers, discount seekers, or those at risk of churn. Despite these advances, many machine learning models remain inherently opaque, limiting transparency and reducing trust in automated decision-making systems (Kim et al., 2018; Molnar, 2019). Prior studies emphasize that while artificial intelligence significantly enhances segmentation, targeting, and personalization, its limited interpretability continues to hinder broader adoption in real-world marketing applications (Efendioglu, 2023; Ljepava, 2022).

Traditional segmentation and classification models typically produce deterministic outputs in the form of a single predicted label or a probability distribution. Most machine learning approaches assign customers to a single segment without explicitly quantifying the uncertainty associated with the prediction (Resheff et al., 2014; Vovk et al., 2005). Although such models may achieve high predictive accuracy, they often fail to communicate the confidence or reliability of their outputs. In practice, customer behavior is inherently dynamic and influenced by multiple contextual factors, making uncertainty unavoidable. The absence of explicit uncertainty quantification may lead decision-makers to over-rely on model outputs, potentially resulting in suboptimal or misleading marketing strategies (Kendall & Gal, 2017). This limitation highlights the need for predictive systems that not only provide accurate classifications but also explicitly represent uncertainty in a reliable and interpretable manner.

Conformal prediction (CP) has recently emerged as a statistically grounded framework to address this limitation (Portela et al., 2025; Vovk et al., 2005). Under relatively mild assumptions, this approach provides mathematical guarantees (Farag et al., 2025) and distribution-free guarantees that the true label is contained within the prediction set with a predefined probability (Romano et al., 2020). Instead of a single label, CP generates a set of all plausible segments that cannot be rejected at a given significance level. In consumer segmentation, larger sets indicate higher model uncertainty regarding a specific customer's profile (Correia et al., 2024). While this significantly improves the reliability and validity of predictive models, it introduces a new challenge in interpretability. Prediction sets may contain multiple plausible segments, reflecting ambiguity in the model's decision boundary. For practitioners, such set-valued outputs are often difficult to interpret, particularly when seeking to understand why a customer simultaneously belongs to multiple potential segments.

To address interpretability challenges, the field of Explainable Artificial Intelligence (XAI) has gained substantial attention. XAI methods aim to provide transparent insights into model behavior, enabling users to understand how predictions are generated and which features drive the outcomes (Kim et al., 2018; Molnar, 2019). In marketing analytics, explainability is crucial for building trust, ensuring accountability, and supporting actionable decision-making. Among various XAI techniques, SHapley Additive exPlanations (SHAP) has emerged as a widely adopted method due to its solid theoretical foundation in cooperative game theory and its ability to provide consistent and locally accurate feature attribution (Lundberg & Lee, 2017). SHAP enables both local explanations for individual predictions and global analysis through the aggregation of feature attributions, making it particularly suitable for customer-level and segment-level interpretation (Molnar, 2019). In addition to SHAP, other XAI techniques such as LIME and Partial Dependence Plots (PDP) have been widely used to interpret machine learning models (Molnar, 2019; Ribeiro et al., 2016). Despite these advances, most XAI approaches are designed for deterministic, single-label predictions and do not explicitly account for predictive uncertainty. Furthermore, XAI approaches have also been successfully integrated with clustering and classification models to enhance interpretability in segmentation tasks (Hu et al., 2023; Takaew & Romsaiyud, 2026). However, despite these advancements, existing approaches predominantly focus on point predictions and post-hoc explanations, often overlooking the inherent uncertainty associated with model outputs. This limitation becomes critical in customer segmentation, where behavioral patterns often overlap, leading to ambiguous classification boundaries. In such cases, interpreting only a single predicted label may provide an incomplete or even misleading understanding of customer behavior. On the other hand, conformal prediction offers a principled framework for uncertainty quantification by producing set-valued predictions with statistical coverage guarantees. However, while conformal prediction improves reliability, it lacks the ability to explain why multiple labels are assigned to a given instance, particularly at the feature level. Therefore, a fundamental gap remains between interpretability and uncertainty quantification, where existing methods either explain deterministic predictions without accounting for uncertainty, or quantify uncertainty without providing interpretable insights. To the best of our knowledge, there is still no unified framework

that systematically integrates conformal prediction with feature attribution methods to explain set-valued predictions in customer segmentation.

To address this limitation, this study proposes COSHAP (Conformal Segmentation with Hierarchical Attribution Paths), a unified framework that integrates conformal prediction and SHAP to jointly model predictive uncertainty and interpretability in a single analytical pipeline. Unlike conventional approaches, COSHAP treats conformal prediction outputs as structured, interpretable objects and extends SHAP analysis to operate on these uncertainty-aware predictions. In particular, the framework introduces a novel Hierarchical Attribution Path, which organizes feature contributions based on prediction confidence, enabling a deeper understanding of both what drives predictions and how reliable those drivers are. By transforming point-based segmentation into a set-based, interpretable decision process, COSHAP provides a principled approach for bridging the gap between predictive performance, statistical validity, and interpretability (Msaddi & Kumbasar, 2023). Although conformal prediction provides statistically valid coverage guarantees and improves uncertainty quantification in classification, it may introduce additional computational overhead, particularly for large-scale datasets or complex predictive models. This trade-off between predictive reliability and computational efficiency remains an important practical consideration and motivates efficient uncertainty-aware, explainable frameworks such as the proposed COSHAP approach.

This study makes several key contributions. First, it proposes COSHAP, a unified framework that integrates conformal prediction and SHAP for uncertainty-aware customer segmentation. Second, the framework introduces uncertainty-aware segmentation by leveraging statistically valid prediction sets. Third, it develops a hierarchical attribution mechanism to interpret feature contributions under ambiguity. Finally, the proposed approach is validated using the Online Retail II dataset across four machine learning models (Random Forest, XGBoost, Naïve Bayes, and SVM), demonstrating its effectiveness in providing both reliable and interpretable segmentation outcomes.

2. Objectives

This study aims to develop an uncertainty-aware and interpretable framework for customer segmentation that integrates predictive modeling, uncertainty quantification, and explainability. Specifically, the objectives of this study are:

- To develop the COSHAP framework, which integrates conformal prediction and SHAP to enable joint modeling of predictive uncertainty and feature attribution in customer segmentation.

- To quantify predictive uncertainty by constructing conformal prediction sets with statistical coverage guarantees for segment assignments.

- To enhance interpretability by analyzing feature contributions at both global and local levels using SHAP-based explanations.

- To evaluate the proposed framework on the Online Retail II dataset using multiple classifiers, focusing on consistency between clustering-based segmentation and supervised approximation, as well as uncertainty reliability.

- To investigate the relationship between uncertainty and feature attribution, particularly in cases involving ambiguous or overlapping customer segments.

3. Materials and Methods

3.1 Overview of the COSHAP Framework

This study proposes COSHAP (Conformal Segmentation with Hierarchical Attribution Paths) as a unified analytical framework that integrates prediction, uncertainty quantification, and explainability into a single pipeline. Unlike conventional approaches that treat clustering, prediction, and explanation as separate processes, COSHAP jointly models these components to enable uncertainty-aware and interpretable customer segmentation. The objective of this study is not to improve predictive performance against true ground-truth labels, but to enhance the interpretability and reliability of segmentation outcomes derived from data-driven methods.

3.2. Data Source and Preprocessing

The dataset used in this study is the *Online Retail II* dataset from the UCI Machine Learning Repository, which contains transactional records of a UK-based online retail company from 2009 to 2011, including attributes such as invoice number, stock code, quantity, unit price, customer ID, and country (Chen, 2012). These data capture diverse customer purchasing behaviors and serve as a foundation for evaluating the COSHAP framework, particularly for analyzing uncertainty in customer segmentation. From the raw transactions, customer-level features are constructed, including Recency, Frequency, and Monetary (RFM), along with additional attributes such as Tenure, Quantity, Unit Price, and Country, as

summarized in Table 1. The preprocessing stage involves data cleaning by removing invalid entries (e.g., negative Quantity or UnitPrice) and records without Customer ID, followed by feature engineering to derive behavioral indicators, where Recency measures time since last purchase, Frequency captures transaction count, and Monetary represents cumulative spending. Additional attributes such as Tenure and average purchase quantity are included to enhance interpretability. To ensure robustness and compatibility with machine learning models, outlier handling, standardization, and logarithmic transformation are applied to skewed variables. Finally, customers are segmented into four behavioral groups: Champions, Loyal Customers, At-Risk, and Potential Loyalists using RFM-based scoring and clustering, where these segment labels are treated as proxy targets for model approximation and subsequently extended into uncertainty-aware prediction sets within the COSHAP framework.

To maintain data integrity, the preprocessing phase focused on excluding records with missing or invalid values, particularly missing Customer IDs, to avoid bias in the RFM (Recency, Frequency, Monetary) segmentation. By prioritizing record exclusion over imputation, the initial dataset of 1,067,371 rows was refined to 805,549 clean transactions, removing 261,822 entries in the process. Ultimately, the preprocessing stage produced a final modeling dataset consisting of $N = 5,878$ unique customers, ensuring that all analyses were conducted

on complete and valid customer records without missing values.

3.3 Segment Approximation Model

To enable instance-level explanation and uncertainty quantification, a supervised classifier f was trained to approximate the segmentation structure generated during clustering. K-Means was selected because it is well suited for continuous behavioral variables such as Recency, Frequency, Monetary, Tenure, and Average Basket Size, while efficiently minimizing intra-cluster variance and producing behaviorally coherent customer groups (Kumar, 2025). The method is widely adopted in RFM-based segmentation due to its computational efficiency, scalability, centroid-based interpretability, and strong performance on well-separated behavioral patterns (Battur & Totad, 2026; Pradeepa & Jeyakumar, 2022). The number of clusters was determined using the Elbow method and validated using the Silhouette coefficient to ensure coherent cluster separation (Kumar, 2025; Thakur & Mittal, 2024). Prior studies also report that K-Means provides competitive clustering accuracy with relatively low computational complexity and stable performance across practical applications (Al-Kerboly et al., 2023; Pakpahan et al., 2025), while optimization strategies such as improved centroid initialization further enhance convergence and reduce sensitivity to initial assignments (Lv, 2022).

Table 1 Description of Variables in the Online Retail II Dataset

Variables	Description
Recency (R)	The number of days between a customer's most recent transaction and the final observation date recorded in the dataset.
Frequency (F)	The total number of transactions or unique orders made by a customer during the study period.
Monetary (M)	The total monetary value representing the cumulative spending generated from all transactions made by a customer.
Tenure	The time span (in days) from the customer's first recorded transaction in the system to the final observation date in the dataset.
Quantity	The average number of product units purchased by a customer in each transaction.
UnitPrice	The average unit price of products purchased by a customer, serving as an indicator of purchasing power.
Country	A categorical variable representing the geographical location or country from which the customer conducts transactions.
Customer Segment	The target label (segment) consisting of Champions, Loyal Customers, At-Risk, and Potential Loyalists.

Importantly, the supervised models are not intended to validate or improve the clustering results, but rather to approximate the underlying segmentation boundaries in a form compatible with the COSHAP framework. This design separates segmentation discovery from interpretability and uncertainty modeling, allowing the framework to analyze segmentation outcomes in a structured and uncertainty-aware manner. It should be noted that these cluster-derived labels represent an approximation setting for evaluating uncertainty-aware explanations, rather than externally validated behavioral classes, and this constitutes a methodological limitation of the study.

3.4 COSHAP Framework Definition

To establish COSHAP, shown in Figure 1, as a unified and systematic framework for uncertainty-aware explainability in customer segmentation rather than a mere sequential workflow, it is defined as a structured mapping that unifies prediction, uncertainty quantification, and explainability within a single analytical process. This formulation ensures that the framework is reproducible, interpretable, and mathematically well-defined.

Formally, COSHAP is defined as:

$$\text{COSHAP}: (X, f, \alpha) \rightarrow (\hat{y}, \Gamma(x), \phi, H) \quad (1)$$

where:

- $X \subseteq \mathbb{R}^d$ denotes the input feature space (e.g., RFM-based features),
- $f: X \rightarrow Y$ is a trained predictive model,
- $\alpha \in (0, 1)$ is the significance level used in conformal prediction.

The outputs of the framework are defined as:

- $\hat{y} = f(x) \in Y$: the predicted class label,
- $\Gamma(x) \subseteq Y$: the conformal prediction set capturing uncertainty,
- $\phi = (\phi_1, \phi_2, \dots, \phi_d) \in \mathbb{R}^d$: the SHAP feature attribution vector,

$H \in \mathcal{H}$: the hierarchical attribution structure that integrates uncertainty into feature-level explanations.

This formulation explicitly defines the input–output relationship of COSHAP and distinguishes it from conventional pipelines by jointly modeling prediction, uncertainty, and interpretability within a single functional mapping. To ensure mathematical rigor, each component of COSHAP is formally defined as follows.

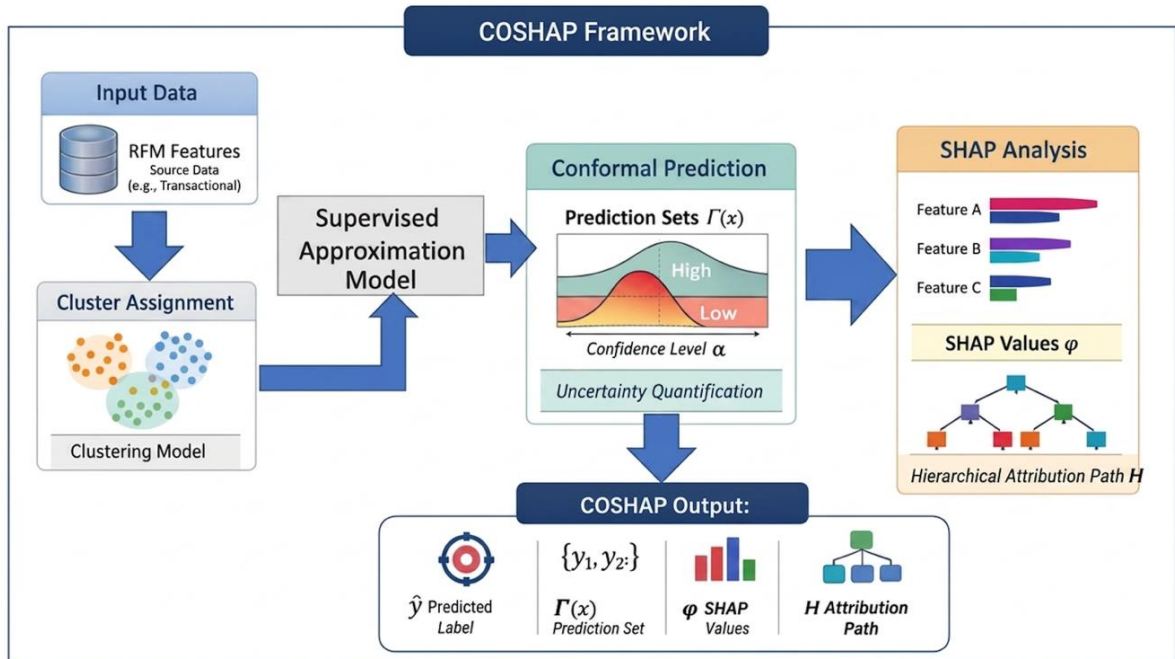


Figure 1 COSHAP framework for uncertainty-aware customer segmentation and explainable set-valued predictions

3.4.1 Conformal Prediction Set

The conformal prediction set is defined as:

$$\Gamma(x) = \{y \in Y : s(x, y) \leq q_{1-\alpha}\} \quad (2)$$

where the non-conformity score is given by:

$$s(x, y) = 1 - \hat{P}(y|x) \quad (3)$$

and $q_{1-\alpha}$ denotes the empirical $(1-\alpha)$ -quantile of non-conformity scores computed on a calibration dataset.

This formulation guarantees statistical coverage:

$$P(y \in \Gamma(x)) \geq 1 - \alpha \quad (4)$$

where $\Gamma(x)$ denotes the conformal prediction set, ensuring that the true label is included in the prediction set with high probability.

3.4.2 SHAP-Based Feature Attribution

Feature contributions are computed using SHAP values, defined as:

$$\phi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!(|F|-|S|-1)!}{|F|!} [f(S \cup \{i\}) - f(S)] \quad (5)$$

where:

- F is the set of all features,
- S is a subset of features excluding feature i .

This formulation ensures that feature attributions are consistent and additive, enabling both local and global interpretability.

3.4.3 Hierarchical Attribution Structure

To incorporate uncertainty into explanation, COSHAP introduces a hierarchical attribution structure defined as:

$$H = H(\phi, \Gamma(x)) \quad (6)$$

where H is a mapping that organizes feature attributions based on prediction set membership and associated confidence levels.

Specifically, this structure enables:

- differentiation between features contributing to confident predictions and ambiguous predictions,
- grouping of features according to their influence across multiple candidate labels,
- interpretation of uncertainty at the feature level.

3.4.4 Algorithmic Implementation

The operational procedure of COSHAP is summarized in Algorithm 1.

Algorithm 1: COSHAP Framework

Input: Dataset $D = \{(x_i, y_i)\}_{i=1}^n$, predictive model f , significance level α

Output: $\hat{y}$, $\Gamma(x)$, ϕ , H

1. Train predictive model f using dataset D
2. Split data into training and calibration sets
3. Compute non-conformity scores on calibration data
4. Estimate threshold $q_{1-\alpha}$
5. For each instance x :
 - a. Compute prediction $\hat{y} = f(x)$
 - b. Construct prediction set $\Gamma(x)$ using Eq. (2)
 - c. Compute SHAP values ϕ
 - d. Construct hierarchical attribution $H = H(\phi, \Gamma(x))$
6. Return $(\hat{y}, \Gamma(x), \phi, H)$

3.4.5 Hierarchical Attribution Path

The proposed Hierarchical Attribution Path (HPA), denoted HH and shown in Figure 2, extends conventional SHAP-based explanations by explicitly incorporating uncertainty information derived from conformal prediction. While standard SHAP methods provide a flat representation of feature importance with respect to a single predicted label, the proposed structure organizes feature contributions into multiple levels conditioned on prediction set membership and confidence. In COSHAP, feature attribution is not interpreted solely based on predicted probability, but is explicitly conditioned on whether a label is included in the conformal prediction set $\Gamma(x)$. This formulation enables a shift from deterministic explanation to uncertainty-aware interpretation, where feature contributions are evaluated relative to multiple candidate labels rather than a single outcome.

Specifically, this approach allows the framework to distinguish between: (i) stable features, which consistently drive predictions across all candidate labels within $\Gamma(x)$, and (ii) ambiguity-inducing features, which contribute to competing evidence and lead to the inclusion of multiple labels in the prediction set. As a result, H provides a structured explanation of prediction behavior under uncertainty, capturing both the magnitude and the reliability of feature contributions. This capability fundamentally differentiates COSHAP from standard SHAP approaches, which are limited to explaining deterministic outputs and cannot explicitly characterize the sources of prediction ambiguity.

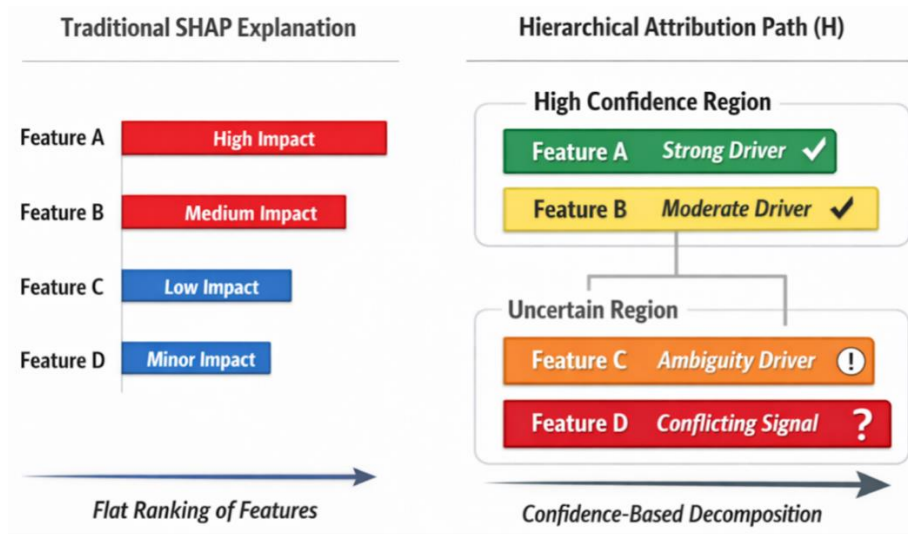


Figure 2 Traditional SHAP vs. COSHAP hierarchical attribution for explaining prediction confidence and ambiguity

Table 2 A comparison between SHAP, conformal prediction, and COSHAP

Aspect	SHAP	Conformal prediction	COSHAP
Feature Attribution	✓	✗	✓
Uncertainty	✗	✓	✓
Integrated Explanation	✗	✗	✓
Hierarchical Insight	✗	✗	✓

Note: ✓ = supported; ✗ = not supported.

3.4.6 Comparative Perspective

To further highlight its distinct contribution, Table 2 presents a comparison between SHAP, conformal prediction, and COSHAP.

3.5. Model Training

3.5.1 Data Partitioning Strategy

The processed dataset was partitioned into three primary subsets to ensure statistical validity and reliable predictive evaluation. The training set (60%) was utilized to develop the base model and optimize its internal parameters. A separate calibration set (20%) played a critical role in the Conformal Prediction framework; rather than being used for model training, this subset was employed to compute non-conformity scores, which serve as the basis for determining the probability threshold used in constructing prediction sets. Finally, the test set (20%) was reserved for assessing the consistency of model approximation and the reliability of uncertainty-aware predictions. To maintain a balanced representation of customer segments across all subsets, the dataset partitioning was performed using a stratified split, ensuring that the distribution of each segment

remained proportionally consistent in the training, calibration, and testing sets.

3.5.2 Base Algorithm Selection and Optimization

The experiments utilized four machine learning algorithms to evaluate the robustness and consistency of the COSHAP method across different modeling approaches. Random Forest (RF) and XGBoost represented ensemble tree-based models capable of capturing complex non-linear relationships, while Support Vector Machine (SVM) was implemented with probability estimation to generate probabilistic outputs required for conformal score computation. Naïve Bayes (NB) was included as a simple probabilistic baseline for comparison. Hyperparameter tuning was conducted using RandomizedSearchCV with 3-fold cross-validation to identify optimal parameter configurations while reducing overfitting and improving model generalization. The search space included parameters such as `n_estimators`, `max_depth`, `min_samples_split`, `learning_rate`, `C`, and various kernel functions, whereas Naïve Bayes required no hyperparameter tuning due to its simpler structure.

Table 3 Hyperparameter configurations and optimal parameter selections for machine learning models

Model	Hyperparameter	Parameter list	Best Parameter
Random Forest	n_estimators	[100, 200, 300]	200
Random Forest	max_depth	[none, 10, 20, 30]	10
Random Forest	min_samples_split	[2, 5, 10]	5
XGBoost	n_estimators	[100, 200]	200
XGBoost	Learning_rate	[0.01, 0.1, 0.2]	0.01
XGBoost	Max_depth	[2, 3, 4]	4
Naïve Bayes	N/A	[]	N/A
SVM	C	[0.1, 1, 10]	10
SVM	kernel	['linear', 'rbf', 'poly']	rbf

As summarized in Table 3, the tuning process explored parameters such as *n_estimators*, *max_depth*, and *min_samples_split* for Random Forest; *n_estimators*, *learning_rate*, and *max_depth* for XGBoost; and *C* along with several kernel functions (linear, RBF, and polynomial) for SVM, while Naïve Bayes did not require hyperparameter tuning due to its simpler probabilistic structure.

3.5.3 Conformal Prediction (CP) Implementation

After the base models were trained, the Conformal Prediction method was applied through several systematic steps. First, non-conformity scores were computed using the calibration dataset to measure how unusual or inconsistent each sample was relative to its true label, based on the predicted probability outputs produced by the base model (Cohen et al., 2023). Second, a threshold value $q_{1-\alpha}$ was determined by calculating the quantile of the calibration scores according to a predefined confidence level, such as 95%. A significance level of $\alpha = 0.05$ (95% confidence) was adopted as a practical and widely accepted standard in machine learning, predictive analytics, and uncertainty quantification studies to ensure statistical robustness while reducing the probability of Type I errors (Dracup, 2007). In conformal prediction, this threshold provides a balanced trade-off between statistical reliability, coverage guarantees, and prediction specificity by ensuring that the true label is included in the prediction set with probability at least $1-\alpha$, while avoiding excessively large prediction sets. The use of a 95% confidence level is also consistent with established practices in predictive analytics and explainable artificial intelligence research (Levenstien et al., 2003). However, the suitability of this threshold may vary depending on the application domain, where high-risk settings often require stricter significance levels, whereas exploratory analyses may adopt more flexible thresholds to support broader pattern discovery (Chaudhuri et al., 2025).

$$P(y \in \Gamma(x)) \geq 1 - \alpha \quad (7)$$

Where $\Gamma(x)$ denotes the conformal prediction set. Increasing confidence levels (e.g., from 95% to 99%) generally enlarges prediction sets, improving coverage but reducing prediction specificity. This trade-off between reliability and informativeness is an important practical consideration in conformal prediction. Because conformal reliability depends on the calibration and predictive quality of the underlying model, multiple base classifiers were evaluated and selected based on predictive performance and calibration behavior prior to conformal implementation. Finally, a prediction set was constructed for each new customer in the test dataset. Instead of producing a single predicted label, the model generated a set of possible labels that guarantees the inclusion of the correct label with the specified confidence level, thereby providing a statistically reliable representation of predictive uncertainty.

3.5.4 Explainability Analysis (SHAP)

To enhance the transparency of the model's decision-making process, SHAP was incorporated into the training pipeline (Salih et al., 2025). The implementation of TreeExplainer for tree-based models such as Random Forest and XGBoost (Sharma et al., 2020), along with KernelExplainer for models like SVM (Wang, 2022) and Naïve Bayes (Karahan et al., 2025), enabled the identification of the most influential RFM features at both global and local levels. This approach ensures that the resulting segmentation is not only statistically accurate but also interpretable from a business perspective, as the contribution of each feature can be traced through a hierarchical attribution pathway.

3.6 Uncertainty-aware Explanation using SHAP and Conformal Prediction

Unlike conventional SHAP analysis that explains a single deterministic prediction, the proposed COSHAP framework extends feature attribution to set-valued predictions generated by Conformal Prediction (CP). In this setting, each instance is associated with a prediction set $\Gamma(x)$, which may contain multiple plausible customer segments under a specified confidence level. SHAP is applied to decompose the predicted probabilities $\hat{P}(L|X)$ for each candidate label, enabling both global and local interpretation of uncertainty-aware predictions. At the global level, SHAP identifies how key behavioral features such as Recency, Frequency, and Monetary define segmentation boundaries across the dataset. At the local level, instance-specific explanations (e.g., waterfall plots) reveal how individual features contribute differently to each candidate label within the prediction set, thereby providing a structured interpretation of ambiguity in model decisions.

The integration between SHAP and CP is formally established through the non-conformity score $s_x = 1 - \hat{P}(L|X)$, where SHAP values are used to decompose the underlying probability $\hat{P}(L|X)$ into feature-level contributions. By comparing these contributions with the calibration threshold $q_{\hat{f}}$, COSHAP explains why multiple labels may be included in the prediction set under a given confidence level. This mechanism enables the construction of a Hierarchical Attribution Path (H),

which organizes feature contributions according to prediction confidence. Specifically, features that consistently support high-probability labels are distinguished from those contributing to ambiguity across competing labels. As a result, COSHAP not only explains which features drive predictions but also contextualizes their reliability under uncertainty, addressing the interpretability gap in conformal classification.

From a practical perspective, this uncertainty-aware explanation provides actionable insights for decision-making. In cases where prediction sets contain multiple segments (e.g., {Loyal Customers, Potential Loyalists}), SHAP reveals the competing influence of features through a “push-pull” dynamic. For example, a high Monetary value may support classification into a high-value segment, while unfavorable Recency may simultaneously indicate potential disengagement. Such insights allow practitioners to identify transitional customer states and design more adaptive marketing strategies, moving beyond rigid single-label classification toward confidence-aware decision support.

3.7 Experimental Setup

The experimental setup follows the COSHAP pipeline summarized in Figure 1 and detailed as an end-to-end workflow in Figure 3, which integrates data partitioning, model configuration, uncertainty quantification, and explainability within a unified process.

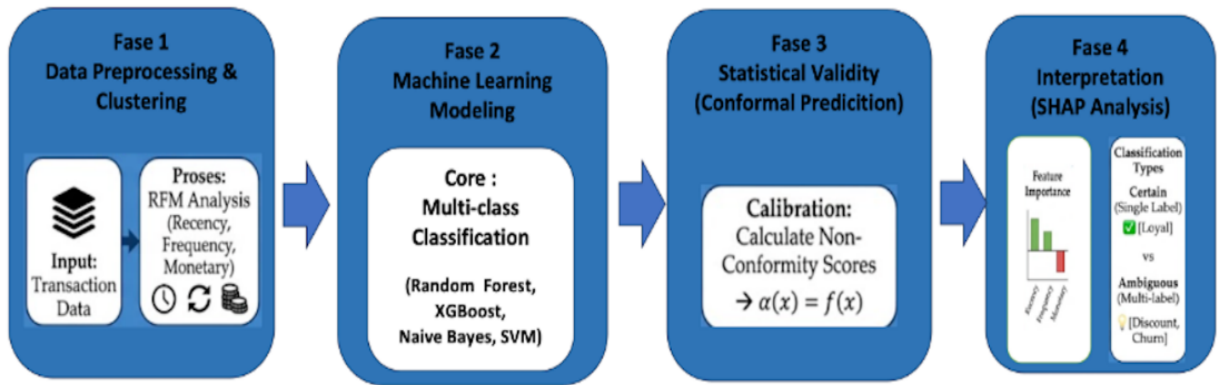


Figure 3 Flowchart of the end-to-end workflow of the COSHAP framework for uncertainty-aware customer segmentation and explainability

The dataset was divided into three subsets according to the split conformal prediction protocol: training (60%), calibration (20%), and testing (20%). The training set was used to fit the predictive models, the calibration set is used to compute non-conformity scores and determine the confidence threshold, and the testing set is used for final evaluation. A stratified sampling strategy is applied to preserve the distribution of customer segments across all subsets. To ensure reproducibility, all stochastic processes were executed using a fixed random seed ($seed = 42$), including dataset partitioning, K-Means centroid initialization, hyperparameter optimization, and model training procedures.

To approximate the segmentation structure, four supervised learning models are employed, including Random Forest, XGBoost, Support Vector Machine (SVM), and Naïve Bayes. Hyperparameter tuning is conducted using randomized search with cross-validation to ensure robust performance. The optimized models produce probabilistic outputs, which are subsequently used in the conformal prediction stage to generate prediction sets $\Gamma(x)$ with statistical coverage guarantees. SHAP is then applied to decompose the model outputs and provide feature-level explanations for both individual predictions and candidate labels within the prediction sets.

All experiments are implemented in a Python-based environment using standard libraries, including scikit-learn, XGBoost, and SHAP. The conformal prediction procedure follows a split conformal framework to ensure valid uncertainty estimation. This setup ensures that the COSHAP pipeline is reproducible and scalable for uncertainty-aware and interpretable customer segmentation.

4. Results

4.1 Data Characteristics and Predictive Performance

The Online Retail II dataset reflects diverse customer purchasing behaviors captured through Recency, Frequency, and Monetary (RFM) features. Across all evaluated models (Table 4), XGBoost achieves the strongest approximation performance, with an accuracy of 97.36% in reproducing the cluster assignments generated by the K-Means segmentation model. These results should not be interpreted as real-world predictive validity against externally verified customer labels, but rather as the model's ability to approximate the underlying clustering structure. The conformal prediction framework further ensures statistical reliability, achieving coverage aligned with the predefined confidence level while maintaining an average prediction set size close to 1.00. This indicates that most customers are assigned a single label with high confidence, while a subset of customers exhibits ambiguity reflected through multi-label prediction sets.

When the confidence level increases to 99% (Table 5), all models demonstrate higher coverage at the expense of larger prediction sets. This behavior highlights the expected trade-off between confidence and specificity, where models become more conservative in uncertain regions of the data space. Importantly, although predictive performance remains consistently high across models, these results alone do not capture the model's ability to explain uncertainty. Therefore, beyond standard evaluation metrics, further analysis is required to understand how ambiguity arises and how it can be interpreted within customer segmentation.

TTable 4 Performance metrics (Confidence: 95%)

Model Variant	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	Coverage (%)	Avg. Set Size	Threshold	Time (s)
Random Forest	97.11	95.77	96.49	96.03	95.75	0.98	0.6356	1.36
XGBoost	97.36	96.07	96.98	96.40	97.36	0.98	0.3782	0.58
Naïve Bayes	86.48	88.39	84.50	84.97	96.00	1.30	0.0663	0
SVM	94.81	93.40	94.02	93.58	94.81	1.00	0.4944	0.70

Table 5 Performance metrics (Confidence: 99%)

Model Variant	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	Coverage (%)	Avg. Set Size	Threshold	Time (s)
Random Forest	97.11	95.77	96.49	96.03	98.47	1.05	0.2450	2.54
XGBoost	97.36	96.07	96.98	96.40	98.72	1.06	0.2983	0.58
Naïve Bayes	86.48	88.39	84.50	84.97	99.08	1.76	0.0033	0
SVM	94.81	93.40	94.02	93.58	99.57	1.15	0.1572	0.71

Figure 4 illustrates the receiver operating characteristic (ROC) curves for the four evaluated models, Random Forest, XGBoost, Naïve Bayes, and Support Vector Machine demonstrating their ability to discriminate between customer segments. Overall, all models exhibit strong classification capability, as indicated by ROC curves that approach the upper-left corner and lie well above the diagonal baseline representing random prediction. The highest area under the curve (AUC) value of 0.9978, followed by XGBoost (0.9966), SVM (0.9955), and Naïve Bayes (0.9796). SVM was the only model whose empirical coverage (94.81%) fell marginally below the nominal 95% level, which is within the expected finite-sample fluctuation for a calibration set of this size. These results indicate that the models are highly effective in distinguishing customer behavioral patterns derived from the RFM-based features of the Online Retail II dataset.

The ROC analysis is consistent with the performance metrics reported in Table 4 and Table 5, where XGBoost and Random Forest demonstrate the strongest predictive performance in terms of accuracy, precision, recall, and F1-score. For example, Table 4 shows that XGBoost achieves the highest accuracy of 97.36%, closely followed by Random Forest with 97.11%, while Table 5 confirms the robustness of these models under a higher confidence level within the conformal prediction

framework. This consistency between the ROC results and the metrics in Table 4 and Table 5 indicates that ensemble-based models provide more reliable classification performance, forming a strong foundation for subsequent uncertainty and explainability analyses.

The confusion matrix in Figure 5 illustrates the classification performance of the Random Forest model in predicting customer segments. The results show that most instances are correctly classified along the diagonal, indicating strong predictive accuracy. The At-Risk segment demonstrates perfect classification with 470 correctly predicted instances and no misclassifications. Similarly, the Champions segment is largely predicted correctly, with 290 accurate predictions and only 6 instances misclassified as Loyal Customers.

For the Loyal Customers segment, the model correctly predicts 220 instances, while a small number are misclassified as Champions (3) and Potential Loyalists (22). The Potential Loyalist segment also shows strong performance, with 162 correct predictions and only 3 instances misclassified as Loyal Customers. Overall, the confusion matrix indicates that the model effectively distinguishes between most customer segments, although slight overlaps occur between Loyal Customers and Potential Loyalists, reflecting similarities in their purchasing behavior.

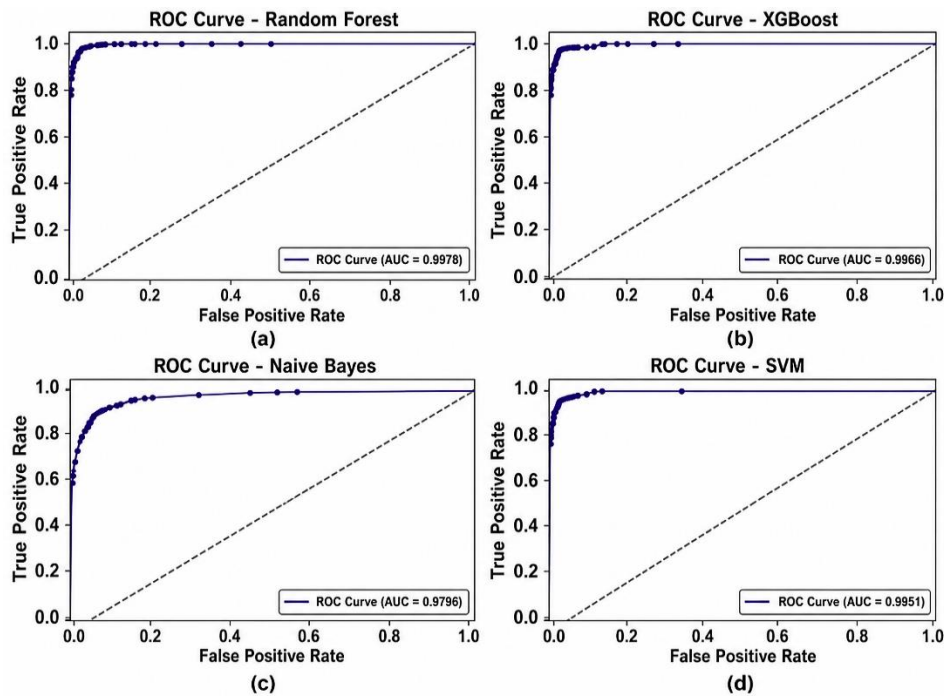


Figure 4 The ROC curves for the four evaluated models (a) Random Forest, (b) XGBoost, (c) Naïve Bayes, and (d) SVM

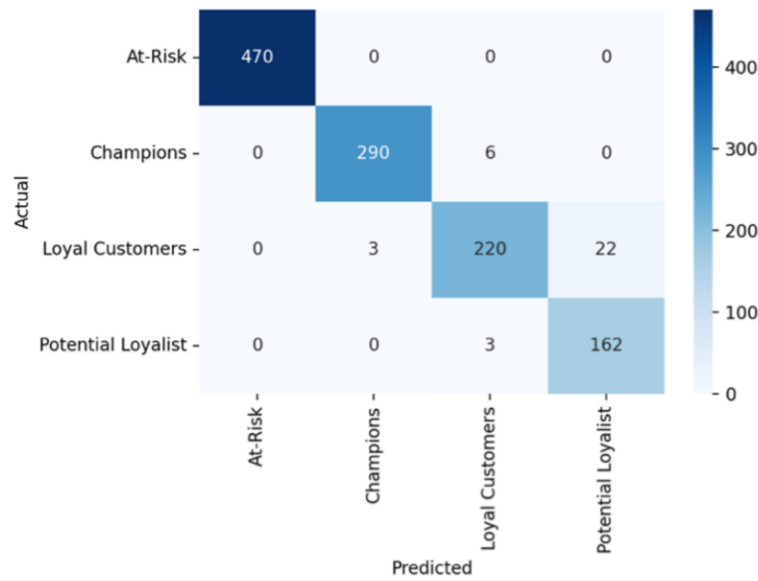


Figure 5 Confusion matrix for customer segment classification (Random Forest)

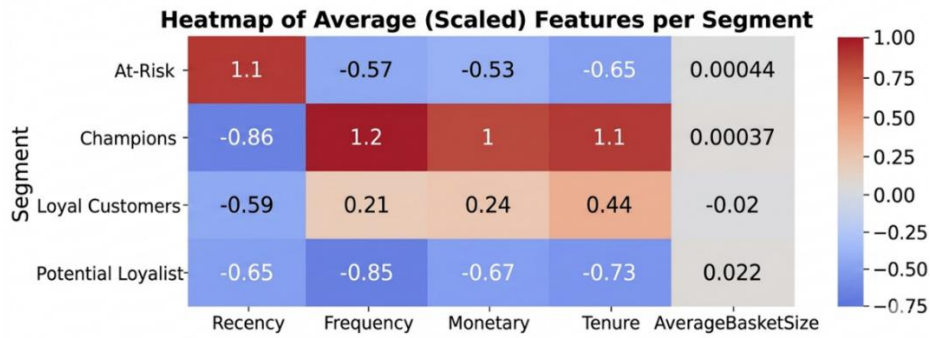


Figure 6 Heatmap illustrating feature overlap across customer segments

To further analyze uncertainty, we examine the distribution of prediction set sizes produced by the conformal framework. The results show that the majority of customers are associated with prediction sets of size 1, indicating confident classification. However, a non-negligible proportion of customers exhibit prediction sets of size greater than 1 ($|\Gamma(x)| > 1$), representing ambiguous cases where multiple segments are statistically plausible. These ambiguous instances typically occur near decision boundaries, particularly between closely related segments such as Loyal Customers and Champions.

To further illustrate segment overlap and the source of prediction ambiguity, Figure 6 presents a heatmap of average standardized feature values across

segments based on Recency, Frequency, Monetary, Tenure, and Average Basket Size. Clear contrasts are observed for extreme segments, with At-Risk characterized by high Recency and low activity, and Champions showing the opposite pattern with high Frequency and Monetary. However, partial overlap appears among intermediate segments. In particular, Loyal Customers exhibit similar Frequency and Monetary patterns to Champions while maintaining moderate Recency. This overlap indicates that some customers share characteristics across multiple segments, making them difficult to distinguish using deterministic models. As a result, these instances are assigned multi-label prediction sets under the conformal framework.

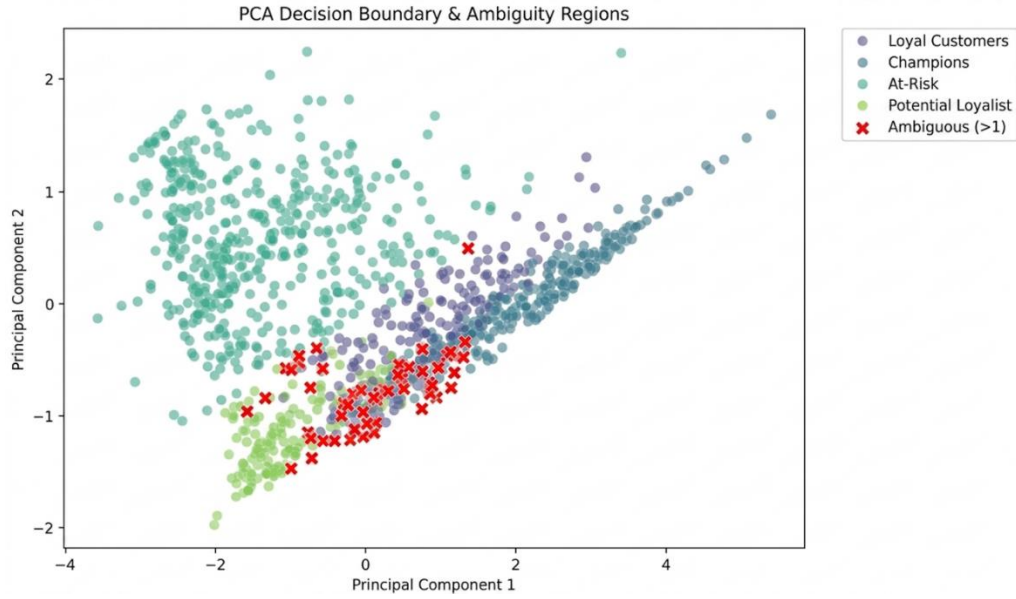


Figure 7 PCA decision boundary showing segment overlap and ambiguity

Furthermore, Figure 7 presents a decision boundary projection in a reduced feature space using Principal Component Analysis (PCA). The visualization shows that, although the main customer segments form relatively distinct clusters, there are clear boundary regions where segment distributions overlap. Notably, ambiguous instances identified by the conformal prediction framework ($|\Gamma(x)| > 1$), highlighted in red, are concentrated in these overlapping regions. This indicates that prediction uncertainty is not randomly distributed, but systematically arises near class boundaries where customer characteristics are not clearly separable. In particular, the overlap between Loyal Customers and Champions is evident in the central region of the projection, where customers exhibit similar behavioral profiles. This explains the occurrence of multi-label predictions, as multiple segment assignments remain statistically plausible for these instances.

While standard SHAP produces similar feature rankings for deterministic predictions, COSHAP provides additional explanatory power by identifying the features responsible for prediction ambiguity in multi-label prediction sets. For example, in ambiguous cases, Recency often contributes to uncertainty by pushing predictions toward the At-Risk or Loyal segments, whereas Monetary and Frequency may simultaneously support assignment to the Champions segment. This competing influence

among features explains why certain customers cannot be confidently assigned to a single segment, highlighting the importance of uncertainty-aware interpretability.

4.2 Customer Segment Discovery

The unsupervised clustering process applied to the customer transaction data successfully identified several meaningful behavioral segments based on the Recency, Frequency, and Monetary (RFM) features. As summarized in Table 6, four main customer profiles were discovered: At-Risk, Champions, Loyal Customers, and Potential Loyalist. The Champions segment represents the most valuable customers, characterized by very low recency (20.37), high purchase frequency (15.66), and the highest monetary value (8,294.97), with a total of 1,482 customers. These customers can be considered loyal or VIP customers who purchase frequently and generate the highest revenue. In contrast, the At-Risk segment shows the highest recency value (427.62) and very low purchasing frequency (2.55), indicating customers who have not made transactions for a long time and may be at risk of churn. The Loyal Customers segment represents customers with relatively recent transactions (78.57) and moderate purchasing behavior, while the Potential Loyalist group consists of customers who recently started purchasing but still exhibit relatively low transaction frequency and monetary value.

Table 6 RFM Segment Profiles

Segment	Recency	Frequency	Monetary	Customer Count	Description
At-Risk	427.62	2.55	980.44	2,347	At-Risk Customers (Have not purchased for a long time, high churn risk)
Champions	20.37	15.66	8,294.97	1,482	Champion Customers (Purchase very frequently, high spending, very recent transactions)
Loyal Customers	78.57	5.42	2,087.58	1,221	Loyal Customers (Frequent buyers, high loyalty, strong repeat purchase behavior)
Potential Loyalist	64.84	1.41	724.95	828	Potential Customers (Recently purchased, moderate transaction level, can be nurtured into loyal customers)

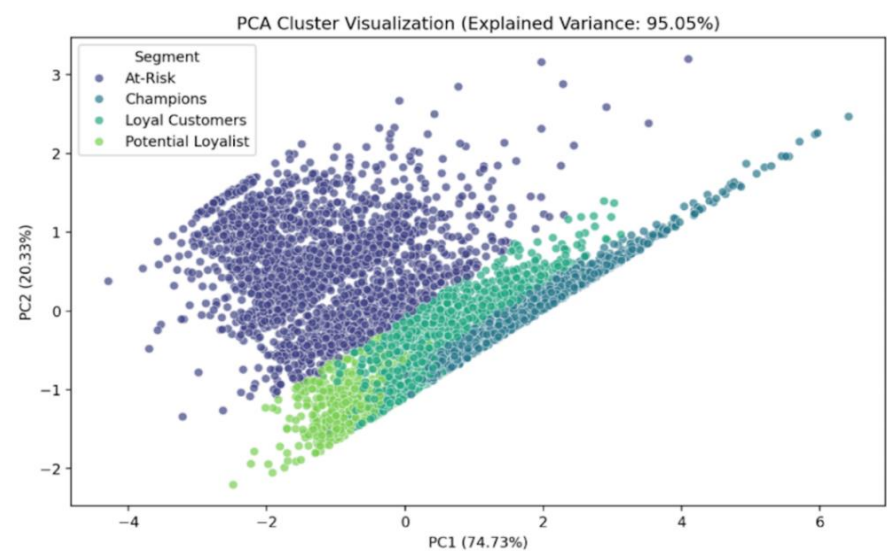


Figure 8 PCA Cluster Scatter Plot

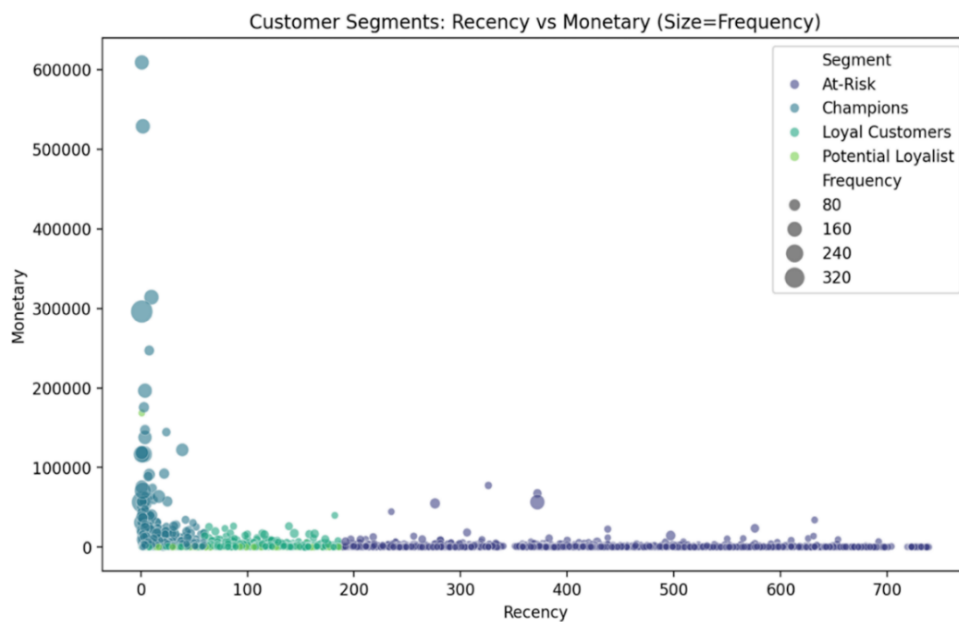


Figure 9 RFM Scatter Plot

The clustering structure is further illustrated in Figures 8 and 9. The PCA cluster visualization (Figure 8) shows that the first two principal components explain 95.05% of the variance, indicating that the dimensionality reduction effectively preserves most of the original information while clearly separating the customer groups. Meanwhile, the Recency–Monetary scatter plot with frequency represented by point size (Figure 9) provides additional insight into purchasing behavior across segments. Customers in the Champions cluster appear with very high monetary values and larger point sizes, reflecting high purchasing frequency, whereas the At-Risk customers are located in the region with very high recency and relatively low monetary values. The Loyal Customers and Potential Loyalist segments occupy intermediate regions, indicating moderate purchasing activity. These segmentation results provide an important behavioral foundation for the subsequent COSHAP explainability analysis, as they enable the model explanations to be tailored according to distinct customer transaction profiles.

4.3 Global and Segment-wise Explainability

Figure 10 presents the SHAP interaction values, illustrating the overall contribution of each feature to the model predictions. As reported in Table 7, Recency emerges as the most influential variable (mean $|SHAP| = 0.255$), followed by Frequency (0.113) and Monetary (0.040), while Tenure (0.018) and Average Basket Size (0.004) exhibit relatively minor contributions. This ranking indicates that the recency of customer activity is the primary driver of segmentation decisions, with purchase frequency

providing additional discriminatory power. The SHAP color gradients further reveal that feature effects are not uniform; instead, high and low feature values contribute in different directions depending on the segment, reflecting the non-linear decision boundaries learned by the model.

Table 7 Global feature importance ranking based on mean absolute SHAP values

Rank	Feature	Mean $ SHAP $
1	Recency	0.255
2	Frequency	0.113
3	Monetary	0.040
4	Tenure	0.018
5	Average Basket Size	0.004

At the segment level (Figure 11), distinct behavioral patterns become evident. For the At-Risk segment, high Recency and low Frequency consistently push predictions toward churn-prone behavior, confirming that inactivity and infrequent transactions are the strongest indicators of disengagement. In contrast, the Champions segment is characterized by the opposite pattern, where low Recency (recent activity), high Frequency, and strong Monetary contributions collectively reinforce high-value customer status. The Loyal Customers segment is primarily supported by sustained Frequency and Tenure, indicating stable engagement over time, while the Potential Loyalist segment reflects transitional behavior, where relatively recent activity combined with moderate Frequency signals early-stage customer development.

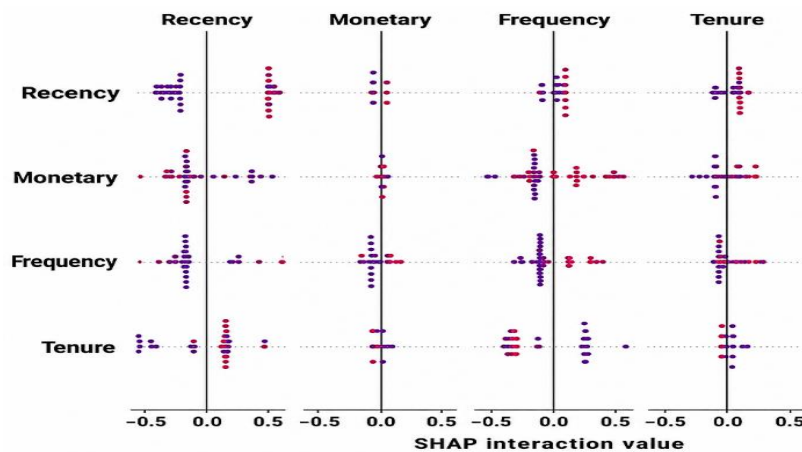


Figure 10 SHAP Interaction Value Beeswarm Plot

Figures 11a–b show the cluster-level SHAP summaries for At-Risk and Champions customers. In the At-Risk cluster, as shown in Figure 11a, Recency has a strong negative contribution, meaning that customers who have not made a recent purchase are more likely to be classified into this group. Low Frequency also pushes the prediction toward the At-Risk cluster, indicating that infrequent transactions are another key warning sign. Monetary and Tenure have smaller effects, while Average Basket Size contributes very little. In contrast, for the Champions cluster, as shown in Figure 11b, the pattern is reversed. Low recency values, meaning very recent purchases, strongly contribute positively to the model output. High Frequency also has a clear positive impact, showing that customers who purchase often are much more likely to be identified as Champions. Monetary and Tenure provide additional positive support, suggesting that high spending and longer customer relationships reinforce this premium segment.

Figures 11c–d show the Loyal Customers and Potential Loyalist clusters. In the Loyal Customer cluster, as shown in Figure 11c, Frequency and Tenure appear as important supporting variables after Recency, indicating that customers with repeated transactions and longer relationships with the company are more likely to be categorized as loyal. Monetary contributes moderately, whereas Average Basket Size remains negligible. Meanwhile, in the Potential Customer cluster, as shown in Figure 11d, Recency and Frequency are again the most decisive features. Customers with relatively recent transactions and increasing purchase frequency tend to show positive SHAP values, suggesting that they are in the early stage of becoming more valuable customers. Tenure also plays a role, implying that newer customers who have started to engage actively still hold growth potential. Overall, the SHAP results consistently confirm that Recency is the dominant feature across all clusters, followed by Frequency, while Monetary, Tenure, and especially Average Basket Size have more limited influence on customer segmentation.

While these results confirm that SHAP effectively captures global and segment-specific

feature importance, they remain inherently deterministic, focusing on single-label predictions. As a result, they do not fully explain cases where customer behavior lies near segment boundaries. In contrast, the COSHAP framework extends this analysis by conditioning feature attribution on prediction set membership, enabling the identification of features that contribute not only to prediction outcomes but also to prediction ambiguity. This distinction is critical, as it allows the model to reveal how competing feature influences, such as Recency indicating disengagement while Monetary suggests high value, lead to multi-label prediction sets. Consequently, COSHAP provides a deeper, uncertainty-aware interpretation of customer behavior that goes beyond traditional feature importance analysis. This demonstrates that feature importance alone is insufficient to fully explain customer segmentation, particularly in regions where behavioral patterns overlap, reinforcing the need for uncertainty-aware explainability as provided by COSHAP.

4.4 Individual-Level Explanation under Prediction Uncertainty

To demonstrate the practical advantage of the proposed framework, we analyze individual-level predictions under uncertainty using the COSHAP approach. Table 8 summarizes the feature profile of a selected customer (ID: 16928), characterized by moderate Recency (80 days), high Frequency (29 transactions), substantial Monetary value (\$4,113.23), and long Tenure (594 days). These attributes indicate a customer with strong engagement and purchasing activity, yet not clearly belonging to a single behavioral segment.

Table 8 Feature Values of the Selected Customer Instance (ID: 16928)

Feature	Value	Description
Recency	80 days	Time since last purchase
Frequency	29	Number of transactions
Monetary	\$4,113.23	Total spending
Tenure	594 days	Duration as a customer
Average Basket Size	95.0	Average number of items per purchase

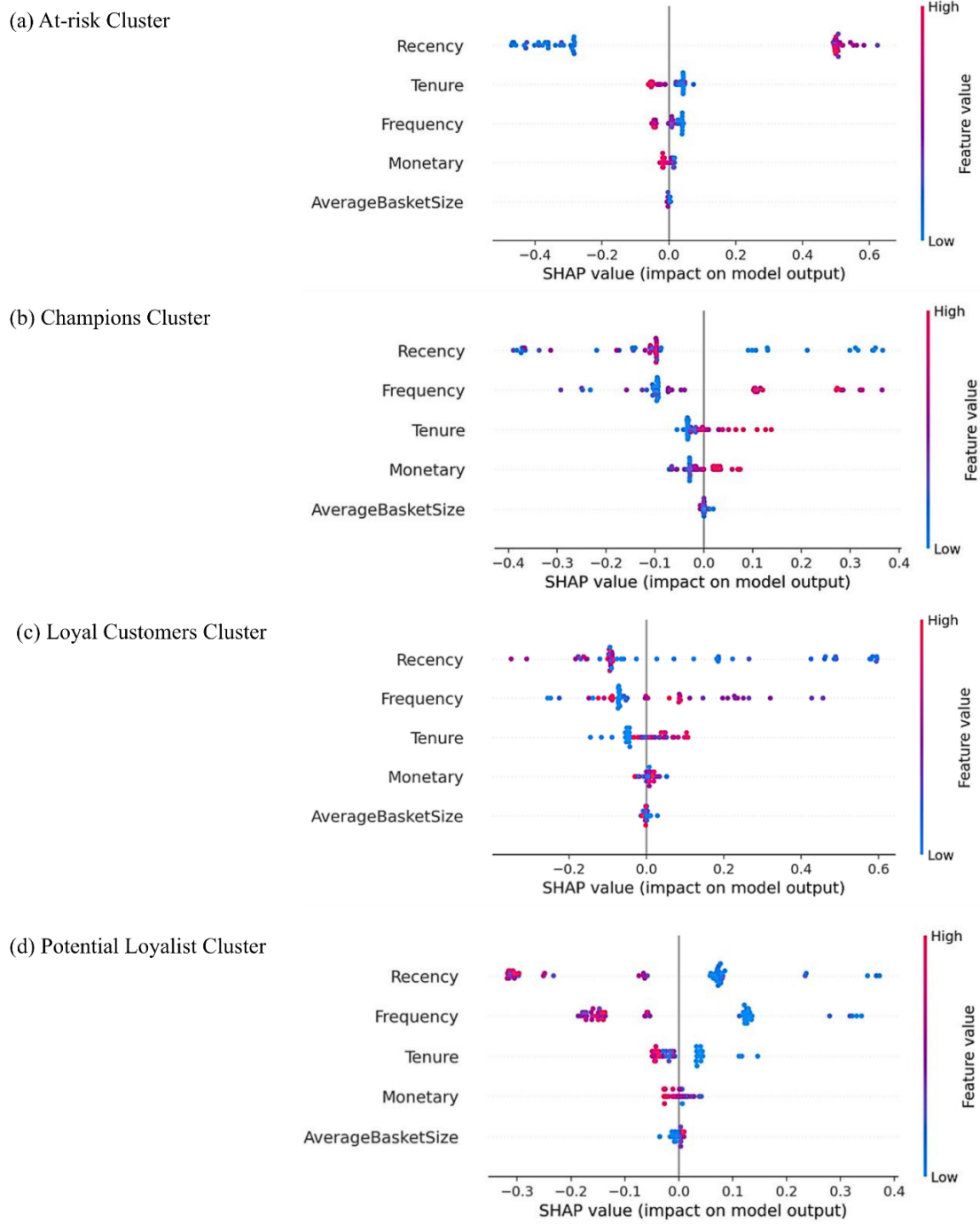


Figure 11 Segment-wise SHAP feature importance showing the dominant drivers for each customer segment

The conformal prediction results in Table 9 show that this customer is assigned a multi-label prediction set, consisting of *Champions* (52.97%) and *Loyal Customers* (47.03%). This reflects inherent ambiguity in the customer's behavioral profile, where both segments are statistically plausible. Unlike

traditional classification models that force a single-label decision, the conformal framework preserves this uncertainty by allowing multiple candidate labels. However, while conformal prediction quantifies *what* is uncertain, it does not explain *why* the ambiguity occurs.

To understand how the model behaves at the individual level, local interpretability was examined using SHAP waterfall plots for one selected customer. This analysis was conducted to evaluate how each feature contributed to the lower-bound probability of the conformal prediction for several customer classes. The SHAP waterfall plots visually decompose the contribution of each customer attribute, allowing the prediction mechanism to be interpreted more transparently. In this case, four class-specific plots were examined, namely At-Risk, Champions, Loyal Customers, and Potential Loyalist.

Conformal prediction, as shown in Table 9, measures predictive uncertainty without strong distributional assumptions and produces prediction sets that contain the true segment with a chosen confidence level. In customer segmentation, these sets help identify cases where the model is uncertain about segment assignment. This allows human review or alternative strategies to reduce mis-segmentation risk.

To address this limitation, SHAP-based explanations are applied at the class level to decompose the prediction for each candidate label. Figures 12(a–d) present the SHAP waterfall plots, which reveal how individual features contribute differently across classes. For the *Champions* class, Frequency, Monetary, and Tenure provide strong positive contributions, reflecting high purchasing intensity and long-term engagement. In contrast, for the *Loyal Customers* class, Recency becomes the dominant positive driver, indicating relatively recent activity. At the same time, these features exhibit competing effects across classes: for instance, Frequency pushes the prediction toward *Champions*,

while Recency supports *Loyal Customers*. This “push-pull” dynamic across feature contributions explains why the model is unable to confidently assign a single segment.

Crucially, this analysis highlights the key novelty of the COSHAP framework. While standard SHAP explains why a single label is predicted, COSHAP explains why multiple labels are simultaneously plausible. By conditioning feature attribution on prediction set membership, COSHAP identifies not only the dominant drivers of prediction, but also the features responsible for prediction ambiguity. In this case, the coexistence of high Frequency and Monetary (favoring *Champions*) with moderate Recency (favoring *Loyal Customers*) results in a boundary condition between segments. This indicates that the customer is in a transitional behavioral state, rather than being misclassified.

From a practical perspective, this ambiguity-aware explanation provides actionable insights for decision-making. Instead of treating such cases as model uncertainty to be ignored, COSHAP enables analysts to interpret them as meaningful signals of customer transition. For example, the customer may be evolving from a loyal profile toward a high-value segment, suggesting targeted strategies such as personalized promotions or retention incentives. Therefore, ambiguity is not merely a limitation of the model, but a valuable source of insight when properly interpreted through the COSHAP framework. This demonstrates that ambiguity is not an artifact of model imperfection, but a reflection of overlapping behavioral patterns that can only be revealed through uncertainty-aware explainability.

Table 9 Prediction sets for selected customers using conformal prediction

Customer ID	Recency	Frequency	Monetary	Prediction Set (Probabilities)	True Label
16928	−0.5796	2.2877	1.0706	Champions (52.97%), Loyal Customers (47.03%)	Loyal Customer
14698	−0.9522	0.2992	0.2131	Champions (6.91%), Loyal Customers (92.89%)	Champions
17371	−0.8806	0.4896	−0.244	Champions (67.48%), Loyal Customers (32.52%)	Champions
17702	−0.8185	0.4896	0.8769	Champions (61.34%), Loyal Customers (38.66%)	Champions
15306	−0.65608	0.9303	0.9510	Champions (18.72%), Loyal Customers (81.28%)	Loyal Customers
14484	−0.72296	0.0739	0.1013	Champions (10.18%), Loyal Customers (89.82%)	Loyal Customers

Note: Recency, Frequency and Monetary are standardized values; raw values for customer 16928 are given in Table 8

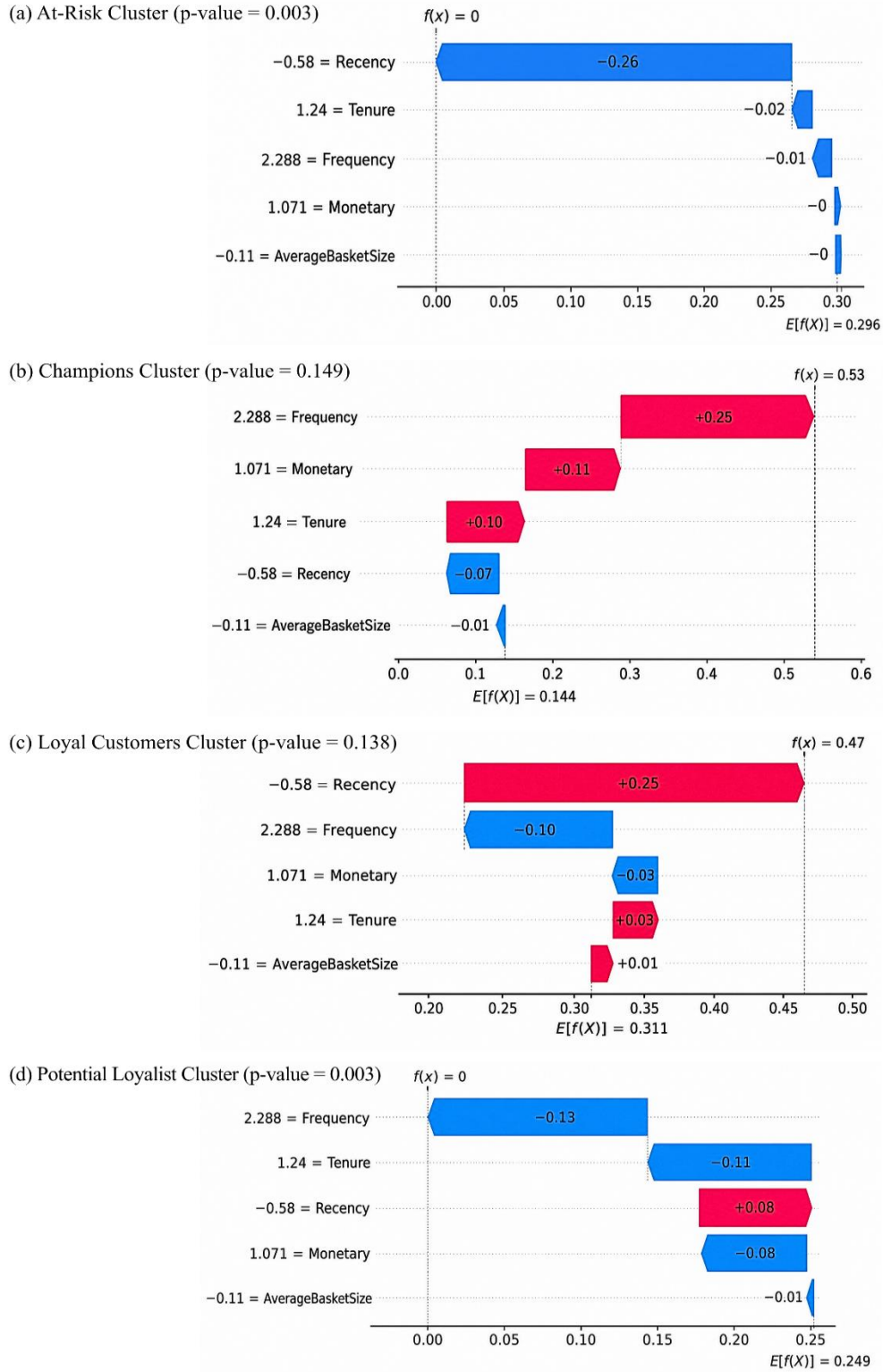


Figure 12 SHAP waterfall plots for an individual customer's prediction (ID: 16928): (a) At-Risk, (b) Champions, (c) Loyal Customers, and (d) Potential Loyalist

Table 10 Stability index metric

Segment	COSHAP Mean Corr (Intra Cluster)	COSHAP Std Dev	Standard Mean Corr (Global)	Standard Std Dev.
At-Risk	0.986	0.023	-0.008	0.947
Champions	0.513	0.480	0.261	0.764
Loyal Customers	0.692	0.282	0.017	0.686
Potential Loyalist	0.963	0.028	0.051	0.751

Table 11 Stability of feature attributions across models

Feature	Random Forest	XGBoost	Naïve Bayes	SVM
Recency	0.93	0.92	0.58	0.57
Frequency	0.92	0.97	0.87	0.93
Monetary	0.44	0.00	0.76	0.73
Tenure	0.86	0.42	0.92	0.75
Average Basket Size	0.70	0.00	0.78	0.24

4.5 Explanation Stability

The stability of model explanations was assessed by comparing the correlation of feature attributions generated by the COSHAP approach and the standard SHAP method. In the standard SHAP setting, explanations were analyzed globally across all customers without considering segmentation, resulting in a very low average correlation of 0.075 across all customers (segment-wise values are reported in Table 10). This occurs because customers exhibit diverse behavioral characteristics and purchasing patterns, leading to highly varied feature attributions across instances. In contrast, COSHAP evaluates explanation consistency within each customer segment, where individuals share similar behavioral profiles. As shown in Table 10, this approach produces significantly higher intra-cluster correlation values, particularly for the *At-Risk* (0.986) and *Potential Loyalist* (0.963) segments, indicating highly consistent explanations within these groups. The *Loyal Customers* segment shows moderate stability (0.692), while the *Champions* segment exhibits greater variability due to more diverse purchasing behaviors.

In addition, the stability of feature attributions was examined across multiple machine learning models, including Random Forest, XGBoost, Naïve Bayes, and SVM. As summarized in Table 11, several behavioral features demonstrate consistent importance across models. In particular, Recency and Frequency exhibit high stability in tree-based models such as Random Forest and XGBoost, with correlation values exceeding 0.90, indicating that these features consistently influence customer segmentation predictions. Meanwhile, Naïve Bayes and SVM show relatively stable contributions for the Frequency and

Monetary features. The values of 0.00 in the XGBoost column indicate that the model effectively eliminates redundant features (Monetary and Average Basket Size) and focuses primarily on Recency and Frequency without significantly reducing accuracy. This behavior arises because XGBoost employs a highly greedy and efficient approach when constructing its decision trees. Specifically, once XGBoost identifies features that provide strong data separation (such as Frequency and Recency), it tends to completely ignore other features whose information overlaps or is redundant. Consequently, features with overlapping explanatory information may receive near-zero attribution stability, not because they are irrelevant in an absolute sense, but because their predictive contribution is largely absorbed by stronger correlated features during tree construction. Overall, these findings suggest that incorporating segmentation through COSHAP improves the consistency of model explanations while also highlighting key behavioral attributes that remain influential across different predictive models.

4.6 Hierarchical Attribution Path (HPA)

To further examine prediction ambiguity, we analyze a representative case where the model produces a set-valued output, as illustrated in Figure 13. In this case, customer ID 16928 is assigned to the prediction set {Champions, Loyal Customers}, indicating that the model cannot confidently distinguish between the two segments under the specified confidence level. This ambiguity reflects overlapping behavioral patterns in the RFM space. Using standard SHAP analysis (Table 12), feature attributions are presented as separate, flat rankings for

each class. For instance, Frequency (+0.25) and Monetary (+0.11) support the Champions class, while Recency (+0.25 for Loyal and -0.07 for Champions) shifts the prediction toward Loyal Customers. However, this representation remains fragmented, as it does not explicitly capture the interaction or conflict between features across competing labels, making the source of ambiguity difficult to interpret.

In contrast, the proposed Hierarchical Attribution Path (HPA), visualized in Figure 13, restructures these contributions into layered roles that explicitly reflect prediction uncertainty. Frequency and Monetary emerge as dominant drivers that consistently support the Champions segment, while Recency acts as a competing driver that introduces

ambiguity by pulling the prediction toward Loyal Customers. Meanwhile, Average Basket Size (-0.01) functions as a shared baseline feature with minimal influence on class separation. This layered decomposition reveals the underlying “push-pull” dynamics between features, where dominant signals reinforce a primary class while competing signals generate uncertainty across alternative labels. As summarized in Table 12, this structured interpretation enables COSHAP to identify features responsible for prediction ambiguity rather than only their contribution magnitudes, providing a clearer and more actionable explanation compared to conventional SHAP.

Table 12 Comparison of SHAP and HPA Explanations for customer ID 16928

Feature	SHAP (Champions)	HPA Role	Interpretation	SHAP (Loyal Customers)
Recency	-0.07	Competing	Drives loyal customers	$+0.25$
Frequency	$+0.25$	Dominant	Supports champions	-0.10
Monetary	$+0.11$	Dominant	Supports champions	-0.03
Tenure	$+0.10$	Dominant	Supports champions	$+0.03$
Average Basket Size	-0.01	Shared	Baseline factor	$+0.01$

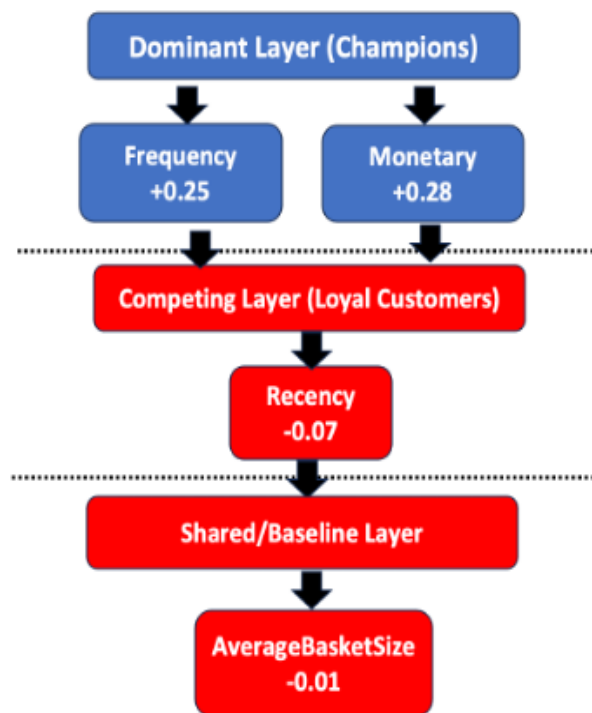


Figure 13 Hierarchical Attribution Path (HPA) for ambiguity explanation

4.7 XAI Comparison: COSHAP vs PDP vs LIME

To further evaluate the effectiveness of the proposed framework, COSHAP is compared with widely used explainability techniques, namely Partial Dependence Plots (PDP) and Local Interpretable Model-agnostic Explanations (LIME), representing global (Szepannek & Lübke, 2022) and local interpretability approaches (Salih et al., 2025), respectively.

Figure 14 (PDP) illustrates the global effect of Recency, showing a clear decreasing trend where higher recency values are associated with lower predicted likelihood of belonging to high-value

segments. While this confirms the overall importance of Recency at the population level, PDP averages across all instances and fails to capture local variations or overlapping segment behavior.

Figure 15 (LIME) presents a local explanation for an individual customer, where feature contributions such as Recency, Frequency, and Monetary influence the prediction. Although LIME provides an intuitive breakdown, it explains the prediction toward a single class and does not account for alternative plausible segments, thereby ignoring prediction uncertainty.

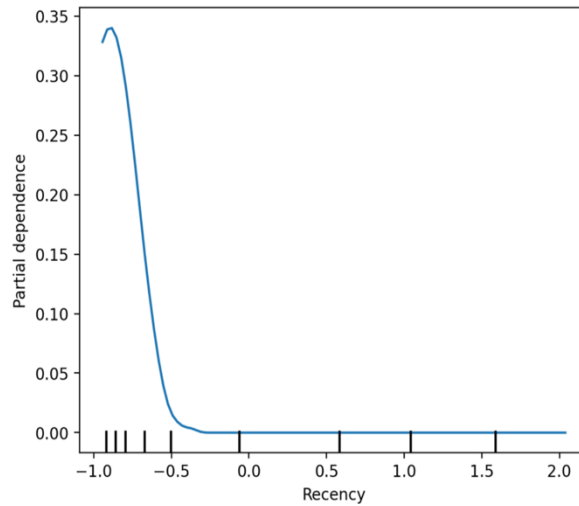


Figure 14 PDP of Recency showing its global effect on model predictions for ID: 15306

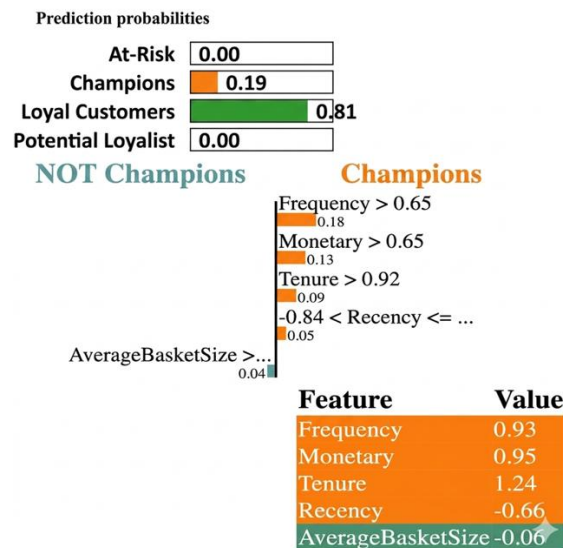


Figure 15 LIME-based local explanation for an individual customer's prediction for ID: 15306

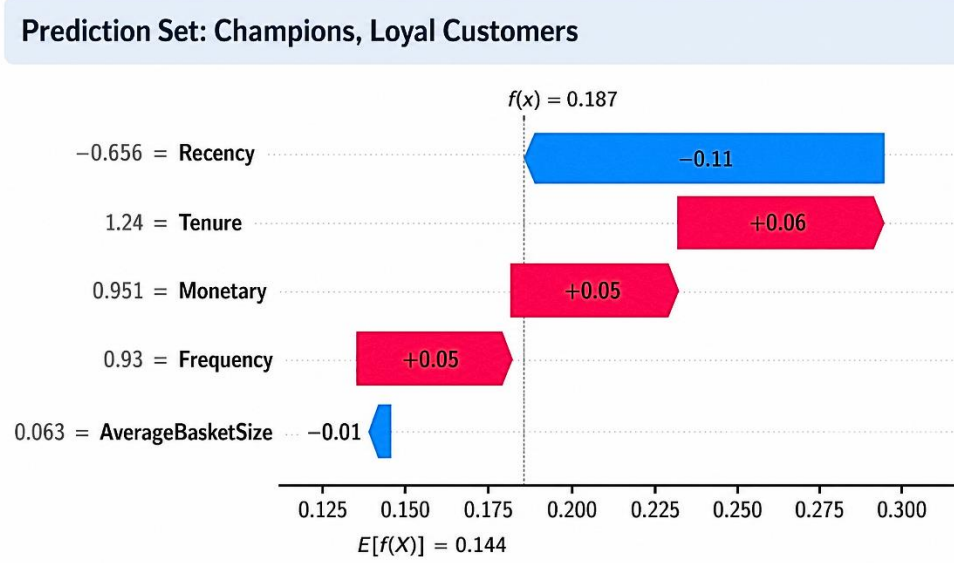


Figure 16 COSHAP waterfall plot explaining feature attribution for customer loyalty prediction

In contrast, COSHAP, as shown in Figure 16, identifies that the same customer belongs to a prediction set {Champions, Loyal Customers}, explicitly capturing ambiguity. Through its attribution mechanism, COSHAP reveals the competing influence of features, where Recency pushes the prediction away from high-value segments while Frequency and Monetary support them.

Overall, while PDP provides global trends and LIME offers local interpretability, both fail to capture ambiguity in overlapping regions. COSHAP addresses this limitation by jointly explaining predictions and uncertainty, providing a more comprehensive and reliable interpretation of customer segmentation.

5. Discussion

The results show that high predictive performance does not necessarily imply reliable decision-making in the presence of uncertainty. While ensemble models such as XGBoost and Random Forest achieve strong classification results, they predominantly operate in a low-uncertainty regime, with an average prediction set size of approximately 0.98. Values slightly below one indicate that a small number of highly atypical instances received empty prediction sets, a known behaviour of split conformal prediction at the chosen threshold. However, the presence of non-singleton sets confirms that ambiguity persists in a subset of instances,

particularly in less confident models such as Naïve Bayes, which generate larger prediction sets (≈ 1.30).

Importantly, the reported classification performance should be interpreted within the context of cluster approximation rather than supervised prediction against externally validated ground-truth labels. Since the target labels are derived from K-Means clustering, the reported accuracy values primarily reflect the consistency with which the supervised models reproduce the segmentation structure generated during the unsupervised stage. Therefore, these metrics indicate approximation fidelity rather than real-world predictive validity and should be interpreted accordingly.

This finding is consistent with prior work in conformal prediction, which focuses on quantifying uncertainty through calibrated prediction sets, but does not address its interpretability (Portela et al., 2025; Vovk et al., 2005). Conversely, SHAP-based approaches provide feature-level explanations but are typically limited to single-label predictions (Lundberg & Lee, 2017). The proposed COSHAP framework bridges this gap by enabling conditional feature attribution over prediction sets, allowing both confident and ambiguous predictions to be interpreted within a unified framework.

From a practical perspective, prediction set size can be interpreted as a decision-support signal. Near-singleton sets (≈ 1.00) indicate high-confidence predictions suitable for direct action, while larger sets (e.g., > 1.20) signal increasing ambiguity and require

more cautious or flexible strategies. This transforms conformal prediction from a statistical guarantee into an operational tool for managing decision risk.

Importantly, COSHAP reveals that ambiguity does not merely reflect a lack of confidence but arises from competing feature contributions. While global SHAP identifies dominant drivers such as Recency and Frequency, COSHAP captures a “push–pull” dynamic, where different features simultaneously support different segment memberships. This provides a concrete explanation of why ambiguity occurs. However, this multi-label interpretability introduces a potential circular-label issue, where overlapping feature attributions support multiple labels without clear differentiation. Although limited in low-uncertainty settings, this issue becomes more relevant as prediction sets grow and requires careful interpretation.

Finally, differences in intra-cluster stability highlight segment heterogeneity. The high stability observed in the At-Risk segment suggests consistent behavioral patterns, whereas lower stability in the Champions segment indicates more diverse feature combinations. This difference can be explained by the behavioral heterogeneity of high-value customers. While At-Risk customers exhibit relatively uniform characteristics, such as high Recency and low purchasing activity, the Champions segment contains customers with diverse combinations of Frequency, Monetary value, and Tenure. Some Champions customers are highly frequent buyers, while others achieve high value through large spending despite lower transaction frequency. This variability produces less consistent feature attribution patterns and consequently lower explanation stability. This suggests that stability metrics can provide additional insight into segment complexity beyond conventional clustering. Overall, COSHAP functions as an adaptive interpretability framework that integrates prediction, uncertainty quantification, and explanation. Its value is most evident in ambiguous cases, while still remaining consistent with standard explanations in high-confidence scenarios. These findings demonstrate that prediction ambiguity is model- and confidence-dependent.

6. Conclusion

This study introduces COSHAP (Conformal Segmentation with Hierarchical Attribution Paths) as a unified framework that integrates predictive modeling, uncertainty quantification, and explainable artificial intelligence for customer segmentation.

Using the Online Retail II dataset, multiple machine learning models including Random Forest, XGBoost, Support Vector Machine (SVM), and Naïve Bayes were evaluated using RFM features. The results show that ensemble-based methods, particularly Random Forest and XGBoost, achieve superior predictive performance, while SVM and Naïve Bayes serve as comparative baselines, highlighting trade-offs between model complexity and performance. However, the findings emphasize that predictive accuracy alone is insufficient for reliable and interpretable decision-making, especially in the presence of overlapping customer behaviors.

By incorporating conformal prediction, COSHAP generates prediction sets with formal statistical coverage guarantees, enabling a principled representation of predictive uncertainty. While most customers are assigned to a single dominant segment, a non-negligible proportion exhibits multi-label prediction sets, reflecting intrinsic ambiguity in behavioral patterns. This challenges the conventional reliance on deterministic segmentation and underscores the importance of uncertainty-aware analytical frameworks in practical marketing applications.

Furthermore, COSHAP enables systematic interpretation of uncertainty through conditional feature attribution. By integrating SHAP within the conformal prediction framework, the approach provides fine-grained insights into feature contributions for both confident and ambiguous predictions. This dual-level interpretability facilitates a deeper understanding of customer behavior, particularly in regions where segment boundaries overlap.

Despite these contributions, several limitations should be acknowledged. The use of clustering-based labels instead of ground-truth annotations may introduce bias in evaluation, and the framework’s effectiveness depends on the calibration of underlying models. Additionally, conformal prediction may impose computational overhead in large-scale settings. Future work should explore applications on fully labeled real-world datasets, extensions to streaming environments, and alternative uncertainty quantification methods, and should address the challenge of overlapping feature attributions in multi-label explanations.

7. Data and Code Availability

The dataset used in this study, Online Retail II, is publicly available and can be accessed through the

UCI Machine Learning Repository or its original data provider. All data preprocessing steps and feature engineering procedures are fully described in the manuscript to ensure reproducibility. The code will be made publicly available upon publication.

8. Conflicts of Interest

The authors declare no conflicts of interest.

9. Acknowledgements

The authors express their sincere gratitude to the Institute for Research and Community Services (LPPM), Catur Insan Cendekia University, for their financial support through a research grant. This research is also part of a doctoral dissertation program at Diponegoro University.

10. Abbreviations

Abbreviation	Full Term
AI	Artificial Intelligence
CP	Conformal Prediction
COSHAP	Conformal Segmentation with Hierarchical Attribution Paths
SHAP	SHapley Additive exPlanations
LIME	Local Interpretable Model-agnostic Explanations
PDP	Partial Dependence Plots
RFM	Recency, Frequency, Monetary
SVM	Support Vector Machine
XAI	Explainable Artificial Intelligence

11. CRediT Statement

Marsani Asfi: Conceptualization, Methodology, Software, Validation, Formal Analysis, Investigation, Resources, Data Curation, Writing – Original Draft, Visualization, Project Administration, Writing – Review & Editing.

Budi Warsito: Conceptualization, Methodology, Validation, Formal Analysis, Writing – Review & Editing, Supervision.

Adi Wibowo: Conceptualization, Methodology, Validation, Formal Analysis, Writing – Review & Editing, Supervision.

11. References

Al-Kerboly, D. M. A., Hamad, M. M., & Dawood, O. A. (2023). *Clustering algorithms comparison for University of Anbar researchers' Google Scholar profiles* [Conference presentation]. In 2023 Al-Sadiq International Conference on Communication and Information Technology (AICCIT). IEEE.

Al-Muthana, Iraq.

<https://doi.org/10.1109/AICCIT57614.2023.10217951>

Alves Gomes, M., & Meisen, T. (2023). A review on customer segmentation methods for personalized customer targeting in e-commerce use cases. *Information Systems and e-Business Management*, 21, 527–570.

<https://doi.org/10.1007/s10257-023-00640-4>

Battur, S., & Totad, S. G. (2026). Data clustering algorithms for incremental datasets: A comparative study. In *International Conference on Data Science and Applications* (pp. 421–433). Springer Nature Switzerland. https://doi.org/10.1007/978-3-032-10940-8_33

Chaudhuri, S. E., Ben Chaouch, Z., Hauber, B., Mange, B., Zhou, M., Christopher, S., et al. (2025). Use of Bayesian decision analysis to maximize value in patient-centered randomized clinical trials in Parkinson's disease. *Journal of Biopharmaceutical Statistics*, 35(5), 981–1000. <https://doi.org/10.1080/10543406.2023.2170400>

Chen, D. (2012). *Online Retail II [Data set]*. UCI Machine Learning Repository. <https://doi.org/10.24432/C5CG6D>

Cohen, K. M., Park, S., Simeone, O., & Shitz, S. S. (2023). Calibrating AI models for wireless communications via conformal prediction. *IEEE Transactions on Machine Learning in Communications and Networking*, 1, 296–312. <https://doi.org/10.1109/TMLCN.2023.3319282>

Correia, A. H., Massoli, F. V., Louizos, C., & Behboodi, A. (2024). An information theoretic perspective on conformal prediction. *Advances in Neural Information Processing Systems*, 37, 101000–101041. <https://doi.org/10.52202/079017-3203>

Dracup, C. (2007). *Confidence intervals*. Encyclopedia of Statistics in Quality and Reliability. Wiley. <https://doi.org/10.1002/9780470061572.eqr218>

Efendioğlu, İ. (2023). *The change of digital marketing with artificial intelligence* [Conference presentation]. The 7th International Conference on Applied Research in Management, Economics and Accounting. Dublin Republic of Ireland. <https://doi.org/10.33422/7th.iarme.2023.07.101>

- Farag, M., Emam, A., Leonhardt, J., & Roscher, R. (2025). Enhancing decision support in crop production: Analyzing conformal prediction for uncertainty quantification. *Computers and Electronics in Agriculture*, 237, Article 110559. <https://doi.org/10.1016/j.compag.2025.110559>
- Hu, X., Liu, A., Li, X., Dai, Y., & Nakao, M. (2023). Explainable AI for customer segmentation in product development. *CIRP Annals - Manufacturing Technology*, 72(1), 89-92. <https://doi.org/10.1016/j.cirp.2023.03.004>
- Karahan, S. N., Güllü, M., Karhan, D., Cimen, S., Osmanca, M. S., & Barışçi, N. (2025). Realistic performance assessment of machine learning algorithms for 6G network slicing: A dual-methodology approach with explainable AI integration. *Electronics*, 14(19), Article 3841. <https://doi.org/10.3390/electronics14193841>
- Kendall, A., & Gal, Y. (2017). *What uncertainties do we need in bayesian deep learning for computer vision?* [Conference presentation]. 31st Conference on Neural Information Processing Systems (NIPS 2017). Long Beach, CA USA.
- Kim, B., Wattenberg, M., Gilmer, J., Cai, C., Wexler, J., & Viegas, F. (2018). *Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV)* [Conference presentation]. International Conference on Machine Learning. PMLR. Stockholm, Sweden.
- Kumar, N. (2025). Intelligent customer segmentation: Unveiling consumer patterns with machine learning. *Journal of Umm Al-Qura University for Engineering and Architecture*, 16(3), 774-783. <https://doi.org/10.1007/s43995-025-00180-7>
- Levenstien, M. A., Yang, Y., & Ott, J. (2003). Statistical significance for hierarchical clustering in genetic association and microarray expression studies. *BMC Bioinformatics*, 4(1), Article 62. <https://doi.org/10.1186/1471-2105-4-62>
- Ljepava, N. (2022). AI-enabled marketing solutions in marketing decision making: AI application in different stages of marketing process. *TEM Journal*, 11(3), 1308-1315. <https://doi.org/10.18421/TEM113-40>
- Lundberg, S. M., & Lee, S. I. (2017). *A unified approach to interpreting model predictions* [Conference presentation]. 31st Conference on Neural Information Processing Systems (NIPS 2017). Long Beach, CA, USA.
- Lv, M. (2022). *Application of an K-means improved clustering analysis algorithm in the design of resource management information system* [Conference presentation]. 2022 World Automation Congress (WAC). IEEE. San Antonio, TX, USA. <https://doi.org/10.23919/WAC55640.2022.9934387>
- Molnar, C. (2019). *Interpretable machine learning, a guide for making black box models explainable*. Retrieved from <https://christophm.github.io/interpretable-ml-book/shap.html>
- Msaddi, S., & Kumbasar, T. (2023). Generating high-quality prediction intervals for regression tasks via fuzzy C-means clustering-based conformal prediction. In *International Conference on Intelligent and Fuzzy Systems* (pp. 728–739). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-39777-6_63
- Pakpahan, S. I., Mukid, M. A., Saputra, B. A., & Rochayani, M. Y. (2025). *Simulation study and application to compare the performance of the K-means and K-medoids algorithms* [Conference presentation]. 2025 International Conference on Computer Sciences, Engineering, and Technology Innovation (ICoCSETI). IEEE. Jakarta, Indonesia. <https://doi.org/10.1109/ICoCSETI63724.2025.11019227>
- Portela, A., Banga, J. R., & Matabuena, M. (2025). Conformal prediction for uncertainty quantification in dynamic biological systems. *PLOS Computational Biology*, 21(5), Article e1013098. <https://doi.org/10.1371/journal.pcbi.1013098>
- Pradeepa, P., & Jeyakumar, M. K. (2022). Data redundancy removal using KMAD-based self-tuning spectral clustering and CKD prediction using ML techniques. *Journal of Current Science and Technology*, 12(3), 517–537. <https://doi.org/10.14456/jcst.2022.40>
- Resheff, Y. S., Rotics, S., Harel, R., Spiegel, O., & Nathan, R. (2014). AcceleRater: A web application for supervised learning of behavioral modes from acceleration measurements. *Movement Ecology*, 2(1), Article 27. <https://doi.org/10.1186/s40462-014-0027-0>

- Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should i trust you?" *Explaining the predictions of any classifier* [Conference presentation]. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, US.
<https://doi.org/10.1145/2939672.2939778>
- Romano, Y., Sesia, M., & Candes, E. (2020). Classification with valid and adaptive coverage. *Advances in Neural Information Processing Systems*, 33, 3581-3591.
- Salih, A. M., Raisi-Estabragh, Z., Galazzo, I. B., Radeva, P., Petersen, S. E., Lekadir, K., & Menegaz, G. (2025). A perspective on explainable artificial intelligence methods: SHAP and LIME. *Advanced Intelligent Systems*, 7(1), Article 2400304.
<https://doi.org/10.1002/aisy.202400304>
- Sharma, P., Mirzan, S. R., Bhandari, A., Pimpley, A., Eswaran, A., Srinivasan, S., & Shao, L. (2020). *Evaluating tree explanation methods for anomaly reasoning: A case study of SHAP TreeExplainer and TreeInterpreter* [Conference presentation]. International Conference on Conceptual Modeling. Vienna Austria. https://doi.org/10.1007/978-3-030-65847-2_4
- Szepannek, G., & Lübke, K. (2022). Explaining artificial intelligence with care: Analyzing the explainability of black box multiclass machine learning models in forensics. *KI-Künstliche Intelligenz*, 36(2), 125-134.
<https://doi.org/10.1007/s13218-022-00764-8>
- Takaew, S., & Romsaiyud, W. (2025). An explainable approach to sentiment analysis of Thai hotel reviews using a fine-tuned language model and SHAP. *Journal of Current Science and Technology*, 15(4), Article 149.
<https://doi.org/10.59796/jcst.v15n4.2025.149>
- Thakur, Y., & Mittal, N. (2024). *Customer segmentation using K-mean clustering* [Conference presentation]. Biennial International Conference on Future Learning Aspects of Mechanical Engineering. Singapore: Springer Nature Singapore.
https://doi.org/10.1007/978-981-96-7214-1_32
- Vovk, V., Gammernan, A., & Shafer, G. (2005). *Algorithmic learning in a random world*. Boston, MA: Springer US.
<https://doi.org/10.1007/b106715>
- Wang, Q. (2022). *Support vector machine algorithm in machine learning* [Conference presentation]. 2022 IEEE International Conference on Artificial Intelligence and Computer Applications (ICAICA). IEEE. Dalian, China.
<https://doi.org/10.1109/ICAICA54878.2022.9844516>