Customer segmentation and churn prediction in the maritime shipping industry using machine learning techniques: A Sri Lankan case study

Ushara Prabash Melder
Dept. of Software Engineering
University of Kelaniya
Kelaniya, Sri Lanka
usharaprabash@gmail.com

Shanilka Haturusinghe
Dept. of Software Engineering
University of Kelaniya
Kelaniya, Sri Lanka
shatur251@kln.ac.lk

S P Kasthuri Arachchi
Dept. of Software Engineering
University of Kelaniya
Kelaniya, Sri Lanka
sandelik@kln.ac.lk

***Abstract*—Customer churn is a critical issue in industries with high acquisition costs, such as maritime shipping and logistics, where losing a single client can result in substantial financial losses. While churn prediction is well-researched in domains such as telecom and SaaS, limited work has been done in the maritime shipping sector, particularly in Sri Lanka. This research aims to design and deploy a machine learning–based churn prediction and customer segmentation framework tailored for maritime shipping, which uses very common features in the maritime shipping industry. Two primary datasets from the ERP system were combined: the Detailed Income Report, containing financial data, and the operational data report, capturing operational records spanning nine years. Three models were implemented: the logistic regression model, the random forest model, and the XGBoost. As the real-world data had class imbalance present, it was handled using SMOTE for logistic regression and built-in methods for random forest and XGBoost. Results showed that XGBoost achieved the best performance, outperforming the other models. Feature importance analysis revealed the strongest common predictors for the framework. The front-end for the framework was developed using Streamlit and integrated with Power BI for real-time churn monitoring and customer segmentation. This research contributes to filling the gap in maritime churn prediction by demonstrating how machine learning and business intelligence can provide actionable insights for customer retention strategies.**

***Keywords*— Customer Churn, Customer Segmentation, Machine Learning, Maritime Shipping, XGBoost**

I. Introduction

Maritime shipping accounts for nearly 70–80% of global trade, making it a critical backbone of the world economy [1]. In this industry, customer relationships are long-term and high-value, which means that customer churn has a giant financial impact. For Sri Lanka, an island nation with a strategically important geographic position, the potential of the maritime sector has not been fully realized. Despite its importance, the maritime industry in Sri Lanka has been slow to adopt modern data-driven decision-making practices and remains underexplored in academic research.

Customer churn prediction has been widely studied in other industries such as telecom, retail, and SaaS, where predictive models have successfully been used to reduce revenue loss and improve retention. Ensemble methods such as Random Forest, XGBoost, and LightGBM have consistently been reported as effective in handling imbalanced datasets and complex relationships. However, in maritime shipping, most studies have focused only on customer segmentation or qualitative insights, leaving a clear research gap in churn prediction models tailored to this sector. Moreover, a practical, generalizable framework for maritime companies remains absent from the literature.

The aim of this research is to design and deploy a machine learning–based churn prediction and customer segmentation framework for the maritime shipping industry. Specifically, the study integrates exploratory data analysis, predictive modeling, and deployment into business intelligence tools to provide actionable insights for decision-making.

The work aims to address the following research questions:

RQ1: Are machine learning based methods effective in predicting customer churn probabilities in the maritime industry based on rule-based churn labeling methods ?

RQ2: Can machine learning based methods be employed in the segmentation of customer interactions through the transactional and operational data obtained through ERP data, without labels.

This work was conducted as a Sri Lanka-specific case study within the maritime industry, using real-world data from a leading Sri Lankan maritime company.

The contributions of this work include:

- Constructing a domain-specific dataset from ERP systems
- Applying machine learning models to predict churn with a focus on recall
- Deploying the final model with Streamlit and Power BI to bridge technical outputs with business operations

By addressing the gap in maritime churn prediction, this research demonstrates how AI-driven approaches can enhance competitiveness and sustainability in the industry.

979-8-3315-5723-2/26/$31.00 ©2026 IEEE

II. Related Work

Customer churn prediction has been widely studied across multiple industries, yet its applications in the maritime and container shipping sector remain limited. Early work in maritime customer analytics mainly focused on segmentation rather than churn. [2] and [3] analyzed the segmentation of freight forwarders and the competitive position among ocean carriers, providing foundational insights into customer behavior but not predictive churn modeling.

In adjacent logistics and service sectors, churn prediction using machine learning has been extensively explored. [4] highlighted the potential of traditional classifiers for churn prediction in logistics, while research in telecom and subscription domains has driven methodological advances. Random Forest, XGBoost, and ensemble models have repeatedly shown strong performance in churn predicting and classification tasks [4]–[8] with recent studies demonstrating that tree-based models often outperform deep learning for tabular business data [9]. Several works also stress the importance of handling class imbalance, recommending techniques like SMOTE and hybrid resampling [10]–[12] which helps to mitigate the biases of these models.

Recent studies broaden the context by demonstrating churn prediction success in SaaS products [13], appliance rental services [14], credit card usage [15], and customer segmentation research [16]. These collectively highlight the importance of feature engineering, behavioural data, and profit-driven modelling for improving churn detection accuracy [17].

Despite this progress, studies specifically targeting churn prediction in the maritime shipping industry remain scarce. Most existing maritime research focuses on segmentation rather than predictive analytics. Therefore, the current study addresses this gap by exploring machine learning-based churn prediction tailored to maritime shipping customers, leveraging proven modeling techniques from related industries.

III. Methodology

The research followed a structured six-phase methodology. Data were collected from the company's ERP system, consisting of Detailed Income Reports and Job Count Reports spanning nine years. This was followed by data cleaning and preprocessing, which included removal of null values, definition of churn as 24 consecutive months of inactivity, and application of logarithmic and Yeo-Johnson transformations to reduce feature skewness. Exploratory Data Analysis (EDA) identified arrival ratios, billing amounts, and running balance as the strongest and most consistent predictors of churn. Three classification models were developed and evaluated: Logistic Regression, Random Forest, and XGBoost. Class imbalance was handled using SMOTE [17] for Logistic Regression and the `scale_pos_weight` parameter for XGBoost. Finally, the solution was deployed via a Streamlit interface for interactive predictions and a Power BI dashboard for comprehensive visualization and business intelligence reporting. An overview of this methodology is presented in Figure 1.

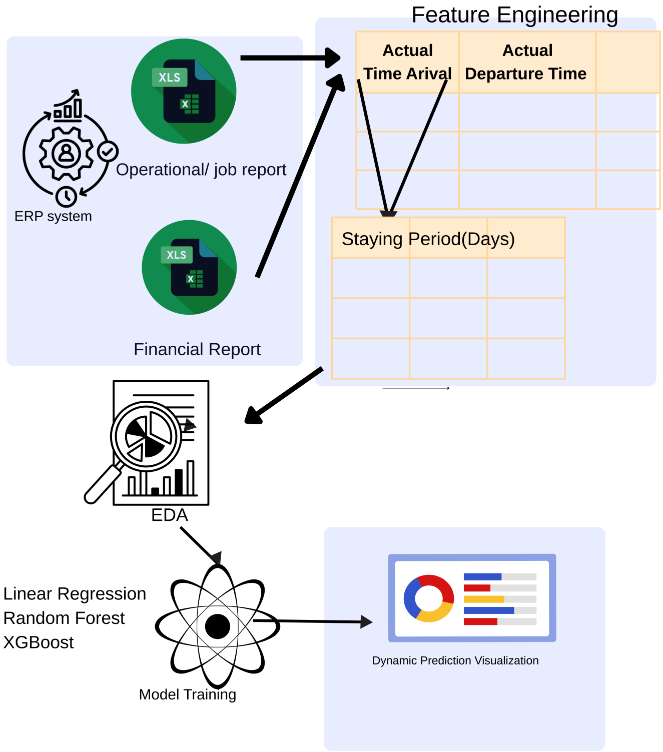


Fig. 1. Overview of the proposed churn prediction framework.

A. Dataset Acquisition

The dataset used in this research was obtained from a leading Sri Lankan maritime shipping company's ERP system, which captures both financial and operational aspects of shipping service activities. Two primary reports were extracted and integrated to construct a customer-level dataset spanning approximately nine years(2016–2024).

a) Detailed Income Report: This report included financial attributes for each job (a single shipping service transaction): **Billed Amount** – total amount invoiced to the customer; **Cost Amount** – total expenses incurred by the company; **Running Balance** – outstanding balance remaining after the transaction.

b) Job Count Report: This report contained operational attributes capturing customer interaction patterns: **Arrival Ratio** – yearly frequency of customer vessel arrivals at Sri Lankan ports; **Staying Period** – average duration (in days) vessels spent in port; **Office** – company office handling the transaction; **Department** – responsible internal department; **Port** – specific port of arrival.

To prepare the dataset for churn prediction, records were aggregated from job-level to customer-level using the principal account holder as the customer identifier. Feature engineering was performed to derive variables such as annual arrival frequency, average billed amount, total cost, cumulative running balance, and number of active years. These engineered features captured both behavioral and financial dimensions of customer engagement. Following initial feature selection and correlation analysis, the most predictive features were retained for modeling, as detailed in Section III-C.

B. Data Pre-processing

Building on the initial two reports, this framework employed a data pre-processing pipeline which excluded attributes with more than 70% missing values, while essential categorical fields such as `Office` were imputed using mode replacement. Duplicate records were removed, and inconsistent identifiers were standardized to maintain referential integrity across merged datasets.

A churn labeling process was implemented where a customer was classified as churned if no recorded transactions or operational activities were observed for 24 consecutive months in the ERP system. This time-window labeling approach is widely adopted in churn modeling [4], particularly where customer attrition is implicit rather than explicitly recorded.

Feature transformation was conducted to prepare the data for modeling. Outliers in continuous features were detected using the IQR method [15] and extreme values were capped or removed. Several financial variables exhibited strong positive skewness, particularly cost amount, which was corrected using Logarithmic and Yeo-Johnson transformations to stabilize variance and improve feature distribution.

Additionally, the dataset exhibited class imbalance, as the proportion of churners was considerably smaller than retained customers. To address this, SMOTE [12] was applied for Logistic Regression to achieve balanced class distribution; crucially, this oversampling was conducted strictly on the training set to prevent data leakage and ensure unbiased evaluation on the test set, while Random Forest and XGBoost used their inbuilt mechanisms, specifically class weighting and the `scale_pos_weight` parameter, to ensure balanced learning without artificially generating data.

After these preprocessing steps, the final dataset was clean, normalized, and balanced, providing a robust and high-quality foundation for exploratory data analysis (EDA) and subsequent predictive modeling.

C. Exploratory Data Analysis

Exploratory Data Analysis (EDA) was conducted to uncover hidden patterns, assess relationships between variables, and understand feature behavior before model training. The primary objective of EDA was to identify the key financial and operational factors influencing customer churn. Various statistical and visualization techniques were applied using Python libraries such as Pandas, Matplotlib, and Seaborn to summarize data characteristics and detect anomalies.

The analysis revealed that arrival-related variables were strong behavioral indicators of churn. Customers with a consistently high arrival ratio mean and a positive arrival ratio trend showed a significantly lower probability of churning, implying that frequent engagement reduces attrition risk. Conversely, declining arrival activity was found to be an early warning sign of potential churn.

TABLE I
HYPERPARAMETER SETTINGS FOR EVALUATED MODELS

Model	**Hyperparameter**	**Value**
Logistic Regression	Penalty	L2
Logistic Regression	C	1.0
Random Forest	n_estimators	100
Random Forest	max_depth	10
Random Forest	class_weight	'balanced'
XGBoost	n_estimators	200
XGBoost	learning_rate	0.1
XGBoost	scale_pos_weight	4.56
XGBoost	max depth	6

In terms of financial variables, billed amount mean and cost amount sum demonstrated a clear relationship with churn behavior. Customers with higher average billing and cost engagement were less likely to churn, suggesting that high-value customers tend to maintain long-term business relationships. However, the cost amount variable exhibited substantial right skewness, which was later corrected using Logarithmic and Yeo–Johnson transformations in during preprocessing. The running balance variable displayed weak linear correlation with churn but proved to be an influential predictor in tree-based models due to its nonlinear relationship with retention behavior.

Correlation analysis using a Pearson heatmap revealed that arrival ratios, billed amount mean, and cost amount sum were positively correlated with retention, while running balance contributed in a nonlinear fashion. Categorical features such as department and office exhibited minimal direct correlation but were retained for their potential domain relevance. Outlier analysis identified a few extreme financial cases, which were handled appropriately during preprocessing.

The EDA phase thus provided valuable insight into the data structure and confirmed that both financial intensity (billed and cost values) and engagement consistency (arrival ratios) play critical roles in customer retention. These findings guided the subsequent model development process, particularly in feature selection and hyperparameter tuning. Similar exploratory studies emphasizing behavioral and financial features for churn prediction have also been reported in previous works [2], [13], [14], confirming the robustness of this analytical approach.

D. Model Development and Evaluation

The modeling stage aimed to identify the most effective machine learning algorithm for predicting customer churn in the maritime shipping domain. Three supervised classification models were implemented: Logistic Regression, Random Forest, and XGBoost, selected based on their established success in prior churn prediction research [5], [13], [14], [16]. The hyperparameters utilized in model development are presented in Table I.

a) Logistic Regression: served as a baseline model due to its simplicity and interpretability. It provided a reference for evaluating improvements achieved through ensemble methods. However, Logistic Regression assumes linear relationships between features and the outcome, which limited its performance

TABLE II
MODEL PERFORMANCE COMPARISON

Model	Acc.	Churn Recall	F1	AUC
Logistic Regression	0.700	0.85	0.79	0.69
Random Forest	0.851	0.93	0.89	0.88
XGBoost	0.854	0.94	0.89	0.89

in this dataset characterized by complex nonlinear feature interactions.

b) Random Forest: an ensemble method based on bagging and decision trees, was implemented to capture nonlinear relationships and reduce overfitting. The model was trained using multiple randomized trees with parameter tuning for n_estimators (number of trees) and max_features. The class_weight = balanced parameter was applied to manage class imbalance. Random Forest achieved significant improvements over Logistic Regression, consistent with findings from prior studies [14], [15].

c) Extreme Gradient Boosting (XGBoost): The final and most effective model. It is a gradient-boosted tree algorithm recognized for its efficiency and superior performance in tabular data tasks [6], [11]. XGBoost was chosen for its ability to handle missing values, manage skewed data distributions, and incorporate built-in class imbalance handling through the scale_pos_weight parameter. Key hyperparameters such as max_depth, learning_rate, n_estimators, and scale_pos_weight were optimized using cross-validation with grid and randomized search methods. The objective function was set to binary:logistic to model churn probability as a binary classification task..

Evaluation metrics included Accuracy, Precision, Recall, F1-score, and AUC–ROC, as recommended by earlier research [12]. Accuracy measured overall correctness, while Recall (Sensitivity) for churners was prioritized because identifying potential churners is more critical than avoiding false alarms in a business context. The F1-score balanced Precision and Recall, providing a robust performance indicator under class imbalance conditions. The AUC–ROC metric was used to assess model discrimination ability.

IV. Results and Discussion

Regarding RQ1, the results in Table II indicate that while machine-learning methods are applicable to churn prediction in the maritime sector, linear models such as logistic regression show limited effectiveness when the churn labels are generated using rule-based heuristics. The confusion matrix (Fig. III-D0c(a)) highlights this limitation: the model correctly identifies only 51 of the 124 non-churn customers, misclassifying 73 of them as churners. This produces a low recall of 0.41 and an F1-score of 0.48 for the non-churn class, demonstrating that the linear decision boundary cannot accurately capture the behavioral and operational patterns of stable customers. Even though the model performs substantially better for the churn class (recall = 0.85, F1-score = 0.79), this imbalance results in a moderate overall ROC–AUC score of 0.696. These findings confirm that the maritime churn problem exhibits non-linear, heterogeneous patterns, likely influenced by operational anomalies, seasonality, and variation in customer logistics profiles, which rule-based labels alone cannot linearize.

This contrasts with industries such as telecommunications, SaaS platforms, rental subscription services, and certain logistics environments, where customer behavior tends to follow more monotonic and predictable feature–churn relationships. In those domains, linear and semi-linear models like logistic regression have historically demonstrated competitive performance because churn is strongly associated with structured patterns such as usage frequency, billing regularity, contract length, and service quality [10], [13], [14].This reaffirms that more advanced ML models like Random Forest, XGBoost are required to effectively predict churn probabilities in this domain, validating the motivation behind RQ1.

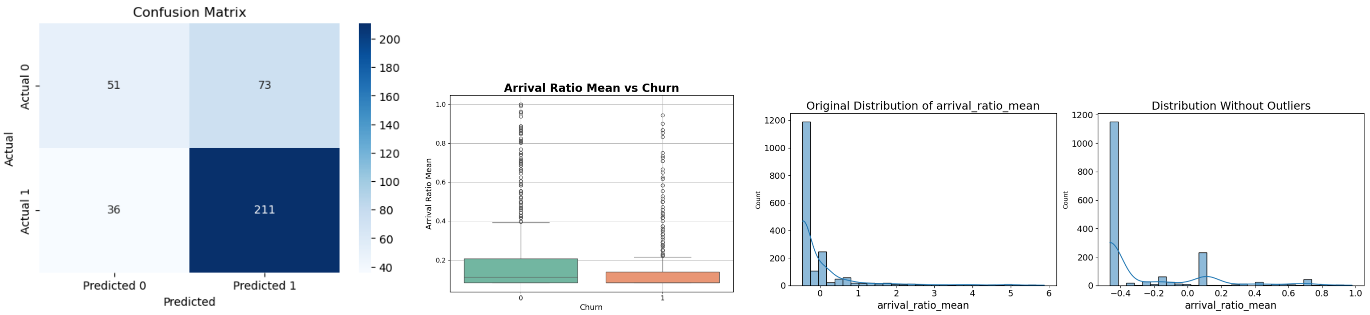
Traditional linear models like Logistic Regression struggle to capture complex, non-linear relationships in churn datasets. In our results, Logistic Regression achieved 0.70 accuracy and 0.83 churn-recall, indicating limited effectiveness in modeling variable interactions. Tree-based models inherently outperform linear models by learning hierarchical decision boundaries, automatically capturing feature interactions, and adapting to non-linear patterns without manual feature engineering [9]. This is particularly important in churn prediction, where user behavior and service interactions create complex non-linear relationships. Both Random Forest and XGBoost significantly surpass the linear model, confirming that tree ensembles produce more stable and generalizable results.

Random Forest shows significant improvement, achieving 0.85 accuracy, 0.93 churn recall, and 0.88 AUC-ROC, validating ensemble bagging methods in noisy, imbalanced datasets. Random Forest reduces variance by aggregating decorrelated decision trees, making it robust to outliers and irregular customer behavior [14], [17]. However, it cannot iteratively correct errors because each tree is built independently. This limitation has been documented in telecom and SaaS churn studies [15], [16]. Therefore, a boosting-based approach like XGBoost becomes necessary to push predictive accuracy further.

According to Table II, XGBoost achieved the highest overall accuracy but produced a non-churn recall of only 0.68, which is problematic from an industry perspective. Misclassifying a large portion of retained customers introduces substantial operational risk.

To address this issue, hyperparameter tuning was applied to improve class-wise recall and overall generalisation. As shown in Table I, XGBoost configuration with a maximum depth of 6, learning rate of 0.1, 200 estimators, and class imbalance handling through the scale _pos _weight method resulted in improved performance, delivering 0.854 precision, a 0.76 recall for non-churners, and a 0.93 recall for churners, significantly enhancing the practical viability of the model.

After evaluating all candidate models and applying hyperparameter tuning, XGBoost produced the strongest overall performance, and therefore it was selected as the most suitable



(a) (b) (c)

Fig. 2. (a) Confusion matrix of the Logistic Regression model. (b) Distribution of mean arrival ratio across customers, stratified by churn status (churned = 1, retained = 0). (c) Effect of outlier removal (using IQR method) on the distribution of the mean arrival ratio.

TABLE III
Final Customer Segmentation Profiles ($k = 3$) after Outlier Removal

Cluster	Segment Name	Proportion	Key Behavioral Characteristic (Centroid Trend)	Business Implication
0	**Standard Portfolio**	89.6%	**Baseline Behavior.** Metrics hover near population mean ($Z \approx 0$). Low debt, moderate frequency.	Represents the “Long Tail” of stable, low-maintenance customers suitable for automated engagement.
1	**Frequent Flyers**	6.6%	**High Operational Intensity.** Highest *Arrival Ratio Mean* ($Z = 2.75$) with moderate financial value.	High service cost relative to revenue. Primary risk vector is operational friction or delays.
2	**The Whales (VIPs)**	3.8%	**High Risk / High Reward.** Highest *Billed Amount* ($Z = 3.95$) and deepest *Running Balance* ($Z = -1.64$).	Critical revenue drivers leveraging significant credit. Primary risk vector is involuntary churn (Credit Default).

TABLE IV
Cluster Validity Analysis using Silhouette Coefficient

Clusters (k)	Silhouette Score	Observation
2	0.674	High separation, over-generalized
3	**0.676**	**Optimal Configuration**
4	0.671	Comparable to $k = 3$, no gain
5	0.572	Significant drop in cohesion
6	0.308	Structural failure (overlapping)
7–10	< 0.350	Poor model fit

model for this dataset. However, several observations from the feature–behavior analysis led to the following discussion. According to Fig. III-D0c(b), the Arrival Ratio Mean does not provide a clearly separated distribution between churners and non-churners. The box–whisker plot shows that the median value for retained customers is approximately 0.12, while the median for churners is only slightly lower at around 0.08. Although a difference exists, it is marginal and insufficient to create a strong discriminative boundary. Furthermore, the presence of substantial outliers in both classes suggests inconsistent behavioral patterns, which may dilute the predictive strength of this feature. This reinforces the need for robust non-linear models such as XGBoost, which can capture subtle interactions that linear methods or weakly separable features fail to represent effectively.

Fig. III-D0c(c), illustrates the distributions of the Arrival Ratio Mean before and after outlier removal, and despite the filtering process, the feature still exhibits noticeable skewness. A similar pattern is observed in the Billed Amount, Cost Amount, and Running Balance Amount distributions, where even after removing extreme values, the data remains highly skewed and non-symmetric. These variables do not follow typical well-behaved distributions, such as normal or near-normal patterns, which are generally more suitable for linear models. In reality, the behavior reflected in these plots strongly supports the conclusion that linear modelling approaches are not appropriate for this dataset, as they assume proportional and structured relationships between features and outcomes. Instead, the irregular, skewed, and non-linear nature of the maritime customer behavioral data aligns much more effectively with tree-based models, which are capable of capturing complex interactions, handling skewed distributions, and managing feature heterogeneity without requiring strict statistical assumptions. This further justifies the superiority of models such as Random Forest and XGBoost for the maritime churn prediction task.

To investigate data abnormalities necessitating non-linear modeling, an unsupervised clustering analysis was conducted. This revealed distinct data topologies that differentiate maritime ERP data from conventional churn benchmarks. Unlike standard subscription datasets, the maritime data exhibited extreme skewness and heavy tails in financial variables. The initial K-Means clustering iteration served as an unsupervised anomaly detection mechanism, revealing the population’s standard liability profile. This underscores that maritime customer data is not merely transaction frequency but a complex ledger of credit risk and capital flow.

The optimal cluster topology was determined using a two-stage validation process in Table IV. The Silhouette Coefficient provided definitive validation for tripartite segmentation ($k = 3$), achieving a maximum score of 0.676. Increasing granularity beyond this resulted in sharp degradation of cluster cohesion (dropping to 0.572 at $k = 5$).

The resulting behavioral profiles in Table III reveal a "Risk-Reward Paradox" unique to this industry. The impact of a False Negative is highly segment-dependent. In the 'Whales' segment, an FN represents a critical financial failure due to high billed amounts, whereas in the 'Frequent Flyers' segment, an FN leads to sudden operational voids and under-utilized capacity. Thus, the framework's high recall is essential for mitigating these varying but equally significant business risks. Cluster 1 (Frequent Flyers) demonstrated highest operational activity (Arrival Ratio Mean $Z \approx 2.75$) yet moderate financial exposure. Cluster 2 (Whales) exhibited highest revenue generation (Billed Amount $Z \approx 3.95$) but deepest debt liability (Running Balance $Z \approx -1.64$). This challenges the assumption that high interaction frequency correlates linearly with high value.

V. Conclusion and Future Work

This research investigated machine learning techniques for predicting customer churn and understanding engagement dynamics in the maritime shipping industry using ERP data, addressing two key research questions: Are machine learning methods effective in predicting customer churn probabilities based on rule-based labeling, and can they segment customer interactions through transactional and operational data without labels.

The results provide affirmative answers to both questions, with critical nuances regarding data distribution. For RQ1, ERP data contains valuable behavioral and financial Indicators, but these indicators are not linear. The separation between retained and churned customers was subtle in key features like `Arrival Ratio Mean`, where retained customers showed a median of approximately 0.12 compared to churners at 0.08. As Fig. III-D0c(c) shows, non-churn customers had more outliers with high arrival ratios, suggesting outliers cannot be ignored for modeling. Key features exhibited heavy right-skewness and long tails even after outlier removal, leading to the conclusion that traditional linear metrics are insufficient. The best accuracy was achieved with tree models trained with outliers but excluding extreme outliers. These results established that domain observations should be prioritized over conventions.

For RQ2, the unsupervised segmentation revealed that maritime customer behavior is fundamentally non-monolithic. Three distinct clusters emerged: Standard Portfolio (most ships' behavior, at risk of churn), Frequent Flyers (high frequency interaction but moderate revenue, not at churn risk), and Whales (medium churn risk while generating high revenue). These findings debunk the assumption that high interaction frequency directly indicates financial stability.

This study presents a novel domain-specific framework for customer segmentation and churn prediction framework in the maritime industry, demonstrating that the skewed and irregular nature of maritime ERP data requires sophisticated non-linear modeling.

Future research should integrate external datasets such as macroeconomic indicators and employ hybrid models combining ERP data with unstructured communications to address current limitations. Additionally, applying explainable AI techniques such as SHAP [11] would enhance transparency, enabling stakeholders to interpret why customers with specific profiles are flagged as high risk.

References

[1] International Chamber of Shipping, "Shipping and world trade – world seaborne trade," https://www.ics-shipping.org/shipping-fact/shipping-and-world-trade-world-seaborne-trade/, accessed: 2024.

[2] C.-H. Wen and W.-W. Lin, "Customer segmentation of freight forwarders and impacts on the competitive positioning of ocean carriers," *Maritime Policy & Management*, vol. 43, no. 4, pp. 420–435, 2016.

[3] G. Balci and I. B. Cetin, "Benefit segmentation of the container shipping market in turkey," *Maritime Policy & Management*, vol. 47, no. 6, pp. 797–814, 2020.

[4] P. Pradeep, K. P. Kumar, and S. S. Rani, "Analysis of customer churn prediction in logistic industry using machine learning," *International Journal of Scientific and Research Publications*, vol. 7, no. 11, pp. 91–96, 2017.

[5] H. Suh, "Ml-based customer churn prediction in home appliance rental business," *Sustainability*, vol. 15, no. 4, pp. 1873–1889, 2023.

[6] A. Imani, "Evaluating an ensemble of rf and xgboost with gnus for churn prediction," *International Journal of Data Science and Analytics*, vol. 9, no. 2, pp. 103–115, 2024.

[7] Y. Chen, "Investigation of machine learning approaches for customer segmentation: A review," *Journal of Intelligent & Fuzzy Systems*, vol. 38, no. 4, pp. 3491–4491, 2025.

[8] A. Lemmens and C. Croux, "Bagging and boosting classification trees to predict churn," *Journal of Marketing Research*, vol. 43, no. 2, pp. 276–286, 2006.

[9] L. Grinsztajn, E. Oyallon, and G. Varoquaux, "Why do tree-based models still outperform deep learning on tabular data?" arXiv preprint arXiv:2207.08815, 2022.

[10] L. Zhang, H. Wang, and J. Yang, "Churn prediction in telecom using machine learning in big data platform," *Cluster Computing*, vol. 23, pp. 1393–1405, 2020.

[11] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 785–794.

[12] J. Burez and D. Van den Poel, "Handling class imbalance in customer churn prediction," *Expert Systems with Applications*, vol. 36, no. 3, pp. 4626–4636, 2009.

[13] P. Jain, "Machine learning models developed for telecom churn prediction," *International Journal of Computer Applications*, vol. 178, no. 26, pp. 15–19, 2021.

[14] I. Ullah, M. Islam, M. Habib, and N. A. Rahman, "A churn prediction model using random forest: Analysis of machine learning techniques for churn prediction in telecom," *International Journal of Advanced Computer Science and Applications*, vol. 10, no. 5, pp. 220–225, 2019.

[15] A. Idris, A. Khan, and Y. S. Lee, "Intelligent churn prediction in telecom: Employing mrmr feature selection and rofboost-based ensemble classification," *Applied Intelligence*, vol. 51, no. 1, pp. 104–121, 2021.

[16] R. Khare and A. Arora, "Predicting customer churn in saas products using machine learning," *International Journal of Information Management Data Insights*, vol. 4, p. 100119, 2024.

[17] A. Amin, S. Anwar, A. Adnan, M. Nawaz, N. Howard, J. Qadir, and A. Hawalah, "Comparing oversampling techniques to handle the class imbalance problem: A customer churn prediction case study," *IEEE Access*, vol. 7, pp. 68 160–68 173, 2019.