Advances in Reliability, Safety and Security, Part 1 – Kołowrocki, Dąbrowska (eds.)
 ISBN 978-83-68136-13-5 (printed), ISBN 978-83-68136-00-5 (electronic)

Advances in Reliability, Safety and Security

ESREL 2024
Monograph Book Series

Severity Prediction Of Maritime Accidents Based On Feature Selection And Data Balance Method

Tianyi Li[a], Xinjian Wang[a], Yinwei Feng[a], Huanxin Wang[a], Yuhao Cao[b], Zhengjiang Liu[a]

[a]*Navigation College, Dalian Maritime University, Dalian 116026, China;*
[b]*School of Engineering, Liverpool John Moores University, Liverpool L3 3AF, United Kingdom*

Abstract

With the rapid development of shipping industry, maritime accidents occur frequently, making research on maritime safety particularly crucial. In this study, a Feature Selection Balancing Framework (FSBF) is proposed, aiming to enhance the accuracy of predicting the severity of maritime accidents. Firstly, various oversampling methods are explored for their effectiveness in data balancing analysis, and the optimal oversampling method is identified. Secondly, a pre-training method of feature selection is employed to rank and select high-contributing features, thereby improving model performance and reducing computational costs. Thirdly, several machine learning methods are used to analyse their effectiveness in predicting accident severity, and establish a baseline model for maritime accident severity prediction. Fourthly, a series of ablation experiments are conducted to demonstrate the contribution of each module within the FSBF to the overall model performance. Finally, using the baseline model, accident severity prediction is conducted and the key factors that influencing accident severity are identified. The research results show that employing the KMeans SMOTE algorithm as an oversampling method, coupled with the use of Gradient Boosting Decision Tree, can make the prediction effect of model best. Furthermore, each module within the FSBF framework significantly improves the performance in predicting the severity of maritime accidents compared to predictions based on original data, offering an effective predictive tool to enhance maritime safety and mitigate the risk of maritime accidents.

Keywords: maritime safety, marine accidents, accident prediction, machine learning, oversampling, feature selection

1. Introduction

The maritime industry constitutes a pivotal component of the global trade network, where ensuring maritime safety is crucial for safeguarding lives, the environment, and cargo. Despite substantial efforts by concerned parties to enhance maritime safety, the occurrence frequency of maritime accidents has not reached the anticipated levels (Cao et al., 2023b). Chen et al. (2019) conducted a thorough survey of quantitative risk analysis methods for ship collision accidents. A classification system was proposed based on the technical features of these methods, concluding with proposed enhancements for several representative approaches. Fan et al. (2020) proposed a risk analysis method based on Bayesian networks, which systematically categorized factors of various types of maritime accidents, revealing their interrelationships. However, the study did not adequately address the issue of imbalanced data. Wang et al. (2021) explored the correlation between the severity of maritime accidents and various factors using an ordered logistic regression model, offering robust support to maritime authorities. However, the study focused solely on the analysis of objective factors, lacking a comprehensive examination of human and managerial factors. Wang et al. (2022) conducted an in-depth examination of factors influencing severity through a zero-inflated ordered probit model. However, the analysis was not sufficiently comprehensive, particularly in addressing managerial factors. Lan et al. (2023) developed a data-driven method that combined association rule mining (ARM), complex networks (CN), and random forests (RF). This innovative approach identified crucial risk factors for predicting the severity of ship collision accidents. However, the study refrained from exploring other machine learning predictive models. In general,

these above studies reveal the interconnectedness of various risk factors with the severity of maritime accidents, providing essential references for future predictions of maritime severity. However, those existing studies have not adequately addressed the issue of imbalanced data. Additionally, in the selection of models for predicting accident severity, there is insufficient comparison of the performance of different models, pointing towards a constrained approach in model selection.

This study makes a significant contribution by proposing a Feature Selection Balancing Framework (FSBF), providing a novel solution for predicting the severity of maritime accidents. The framework not only demonstrated notable performance improvements in experiments, but also offered valuable insights for addressing class imbalance and feature selection issues in related research areas. On the other hand, this study attempts to compare all combinations of five oversampling methods and four machine learning models, ultimately identifying the benchmark model combination that optimally enhances the effectiveness of the FSBF framework.

2. Materials and method

This study primarily focuses on predicting the severity of maritime accidents. However, the limited number of samples for severe accidents has resulted in an imbalance in classes. Firstly, in order to address the challenge of imbalanced data, this study explores diverse oversampling methods within the FSBF such as Synthetic Minority Over-Sampling Technique (SMOTE), Support Vector Machine- SMOTE (SVM-SMOTE), Borderline SMOTE (Borderline-SMOTE), KMeans SMOTE (KMeans-SMOTE), and Random Over Sampling (ROS), and find out the most effective oversampling strategy. Secondly, in order to find the features highly related to accident severity, a feature selection method based on model pre-training is developed. By ranking the features, the features with high contribution to the model performance are selected, which improves the model performance and reduces the calculation cost. Thirdly, through a comparative analysis of several machine learning models (Gradient Boosting Decision Tree, GBDT; eXtreme Gradient Boosting, XGBoost; Light Gradient Boosting Machine, LightGBM; Adaptive Boosting, AdaBoost) in predicting the severity of maritime accidents, the optimal model combination for severity prediction is determined. Finally, utilizing this optimal model, predictions of accident severity are made, accompanied by an in-depth analysis of the key risk factors influencing severity. The experimental framework is depicted in Figure 1.

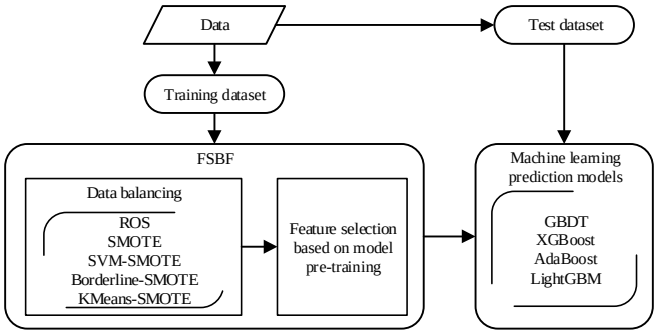

Fig. 1. The experimental framework of this study.

2.1. Data

The data utilized in this study are derived from the previous research (Cao et al., 2023a; Wang et al., 2021). The dataset encompasses marine accident investigation report spanning the years 2000 to 2019, sourced from seven maritime investigation agencies, including China Maritime Safety (China MSA), Federal Bureau of Maritime Casualty Investigation (BSU), National Transportation Safety Board (NTSB), Japan Transportation Safety Board (JTSB), Australian Transport Safety Board (ATSB), Canadian Transportation Safety Board (TSB), and Marine Accident Investigation Branch (MAIB). The data has been meticulously filtered and screened according to the principles of authenticity and completeness, resulting in a final set of 1294 marine traffic accident investigation reports.

2.2. Feature Selection Balancing Framework

Within the FSBF framework, two key components are feature selection and data balancing. Feature Selection refers to the selection of the most representative, relevant or important feature subset from the original feature set for model construction or analysis. It serves to reduce model complexity, enhance model generalization, mitigate overfitting, and accelerate model training speed, thereby conserving computational resources (Li et al., 2017). Common feature selection methods include Filter, Wrapper, and Embedded approaches. Given the necessity of comparing the effectiveness of various machine learning models during the prediction process and the emphasis on analysing the contribution of different features to the models, Filter methods, which do not involve machine learning model training (Lal et al., 2006), and Wrapper methods, which incur high computational costs (Chen et al., 2020b), are deemed less suitable for this study. Therefore, Embedded methods are chosen, involving the pre-training of machine learning models, ranking features based on their importance, and subsequently training the machine learning models in accordance with this feature order.

In the context of classification issues, achieving balance in the data is crucial for both model training and performance evaluation (Ghorbani and Ghousi, 2020). In practical datasets, there can be significant disparities in the quantity of samples across different classes, potentially leading to the model being overly trained on the majority class samples while performing inadequately on minority class samples. To address this imbalance in data categories, the proposed FSBF framework incorporates methods for balancing the dataset, with the aim of equalizing the sample counts across various classes. Commonly employed methods for balancing data include Oversampling, Undersampling, and Combining Sampling strategies. Given this study emphasis on addressing the issue of inadequate learning from minority class samples, the Oversampling strategy emerges as a direct and effective approach. In comparison, Undersampling has the potential to result in information loss, thereby impeding model learning (Agrawal et al., 2015). Additionally, the intricate nature of the Combined Sampling process introduces complexity, making it less straightforward (Chen et al., 2020a). In summary, this study employs the following Oversampling methods to achieve data balance.

Random Oversampling (ROS): Firstly, determine the multiplication factor for oversampling the minority class samples. Secondly, randomly select a subset of samples from the minority class and replicate these samples according to the multiplication factor for oversampling. Finally, merge the original minority class samples with the replicated ones to create a balanced dataset (Bora et al., 2022). The advantages of ROS lie in its simplicity and ease of implementation, without introducing complex computations or algorithms. However, in situations where noise is present in the minority class samples, randomly replicating samples may lead to overfitting.

Synthetic Minority Over-sampling Technique (SMOTE): Begin by selecting a minority class sample as the starting point and calculating its distances to all other minority class samples. Option for a few nearest neighbours from this pool of samples. Randomly choose a sample from the k-nearest neighbours of the initially chosen minority class sample. Calculate the distance between the selected nearest neighbour sample and the base sample. Generate the required number of synthetic samples by randomly determining a proportion multiplied by this distance (Ghorbani and Ghousi, 2020).

Support Vector Machine Synthetic Minority Over-sampling Technique (SVM-SMOTE): Start by employing the SMOTE algorithm to generate synthetic samples, effectively increasing the quantity of minority class samples. Form a classification model by training an SVM classifier with both the original and synthetic samples. Adjust the weights of the synthetic samples based on the error between the output of the SVM classifier and the actual labels. This adjustment aims to amplify the influence of synthetic samples as potential support vectors. Finally, retrain the SVM classifier using the re-adjusted weights of the synthetic samples and the original samples (Wang, 2008). SVM-SMOTE combines the advantages of SMOTE and SVM, uses SMOTE to synthesize samples to balance the data set, and adjusts the weights of synthesized samples by SVM classifier to improve the degree of attention to a few class samples.

Borderline Synthetic Minority Over-sampling Technique (Borderline-SMOTE): Start by employing the SMOTE algorithm to generate synthetic samples, effectively increasing the quantity of minority class samples. Form a classification model by training an SVM classifier with both the original and synthetic samples. Adjust the weights of the synthetic samples based on the error between the output of the SVM classifier and the actual labels. This adjustment aims to amplify the influence of synthetic samples as potential support vectors. Finally, retrain the SVM classifier using the re-adjusted weights of the synthetic samples and the original samples (Han et al., 2005). The fundamental concept of this method is to exclusively synthesize samples positioned around majority class samples and in close proximity to minority class samples. These samples, considered to be on the decision boundary, are the focus of Borderline-SMOTE. Therefore, Borderline-SMOTE concentrates on generating samples closer to the class boundary to better emulate the authentic data distribution and enhance the model's ability to delineate category boundaries.

KMeans Synthetic Minority Over-sampling Technique (KMeans-SMOTE): Employing the KMeans clustering

algorithm, the minority class samples are classified into K clusters, with each cluster's minority class samples serving as the foundational samples. The SMOTE algorithm is then implemented within each cluster to produce the required number of new samples. The final step involves amalgamating the original minority class samples with the synthesized samples to establish a well-balanced dataset (Qu et al., 2020). The primary rationale behind this methodology lies in the clustering of minority class samples using the KMeans algorithm, followed by the application of the SMOTE algorithm to generate new minority class samples within each cluster. This ensures that the synthesized samples are more representative, while the distinctive features of the original samples are preserved.

2.3. Advanced machine learning prediction models

To ascertain the efficacy of the FSBF framework across various models, this study experimentally employs four distinguished machine learning models—namely, GBDT, XGBoost, LightGBM, and AdaBoost—that excel in classification tasks. Then, the Unweighted Average Recall (UAR) is utilized as the benchmark metric to assess the model performance. The following content provides an introduction to the models and the evaluation metric:

Gradient Boosting Decision Trees (GBDT): GBDT falls under the category of Boosting models, combining decision tree models with ensemble learning techniques. By constructing multiple decision trees and progressively reducing prediction errors, GBDT employs gradient boosting during each training round. This involves calculating the gradient of the current model to determine the predicted values for the next tree (Wu et al., 2021). This iterative process allows the model to gradually minimize the loss function, enhancing predictive capabilities. Regularization techniques, such as constraining tree depth, minimum leaf node count, and learning rate, are applied to prevent overfitting.

eXtreme Gradient Boosting (XGBoost): XGBoost is a highly optimized gradient boosting tree algorithm that employs decision trees as its fundamental model. Building upon the foundation of the GBDT algorithm, XGBoost introduces optimizations and innovations to enhance model performance. By iteratively training multiple decision trees and utilizing gradient boosting techniques during each training round, XGBoost calculates the gradient of the current model to determine the predicted values for the next tree (Chen and Guestrin, 2016). This enables XGBoost to progressively minimize the loss function while improving model efficiency.

Light Gradient Boosting Machine (LightGBM): LightGBM is an efficient and powerful gradient boosting tree algorithm designed specifically to handle large-scale datasets and high-dimensional features. Similar to GBDT and XGBoost, LightGBM utilizes decision tree models. However, LightGBM introduces innovations such as the Gradient-based One-Side Sampling (GOSS) algorithm to reduce sample dimensions and the Exclusive Feature Bundling (EFB) algorithm to decrease feature dimensions (Ke et al., 2017). These advancements result in reduced memory usage and improved training speed.

Adaptive Boosting (AdaBoost): AdaBoost aims to enhance the effectiveness of weak classifiers by amalgamating multiple classifiers with incremental focus on previously misclassified samples. This strategic approach results in the creation of a formidable classifier, highlighting AdaBoost's prowess in tackling complex problem scenarios (Cai et al., 2022).

In the exploration of data balance analysis, this study investigates the learning impact of information from different categories of data. Given that Accuracy and area Under Curve (AUC) are not universally suitable for assessing the performance of imbalanced datasets, alternative evaluation metrics are sought for a more comprehensive understanding (Kim et al., 2015). Within imbalanced datasets, models may demonstrate a bias towards predicting the majority class with greater ease, thereby achieving high accuracy. Nevertheless, their performance in predicting minority classes may be inadequate. Thus, this study adopts UAR as the evaluation criterion (Chen et al., 2018). Designed for multi-class classification challenges, UAR functions as an evaluation metric by computing the recall for each class and subsequently determining the average of these recall values. The calculation formula is articulated as follows:

$$UAR = \frac{\sum_{i=1}^{Q_c} Recall_i}{Q_c}, \tag{1}$$

where Q_c is the quantity of category c, with $Recall$ denoting the recall rate.

In the context of imbalanced datasets, UAR stands out as a robust evaluation metric. It maintains equal consideration for each category, independent of their respective sizes. The variability in UAR scores reflects the influence of different oversampling techniques on model performance.

3. Results and analysis

This study experimentally evaluates the combined effects of various oversampling techniques, including SMOTE, SVM-SMOTE, Borderline-SMOTE, KMeans-SMOTE, and ROS, with multiple classifiers (GBDT, XGBoost, LightGBM, AdaBoost) to address the issue of imbalanced data categories. A comprehensive analysis of the experimental results reveals that, on the dataset, KMeans-SMOTE performs relatively well, especially when combined with GBDT, yielding optimal outcomes. This offers an effective oversampling choice for predicting the severity of maritime accidents. These findings hold significant reference value for practical applications in maritime accident prediction and safety management.

3.1. Performance analysis of FSBF

3.1.1. Balancing the data

Upon observing the data before and after balancing (Table 1), it is evident that SMOTE, Borderline-SMOTE, SVM-SMOTE, and ROS successfully balanced severe accidents with non-severe accidents, achieving a 1:1 ratio. However, KMeans-SMOTE concurrently augmented the count of both severe and non-severe accidents, achieving a ratio of 193:192. The primary distinction in the outcomes of KMeans-SMOTE compared to other methods arises from its unique ability to address both inter-class and intra-class imbalances. Given the imbalance within the majority class samples in the sample space, KMeans-SMOTE synthesizes majority class samples to achieve internal balance, thereby increasing the count of majority class samples. In contrast, other methods predominantly focus on generating fewer but more representative synthetic samples between classes to balance the class quantities.

Table 1.The changes in data before and after oversampling.

Oversampling methods	Non-serious accidents	Serious accident	Total	Minority proportions
SVM-SMOTE	768	768	1536	0.5000
Borderline-SMOTE	768	768	1536	0.5000
KMeans-SMOTE	772	768	1540	0.4987
SMOTE	768	768	1536	0.5000
RandomOverSampling	768	768	1536	0.5000
Original	768	267	1035	0.2580

Upon scrutinizing the amalgamation of oversampling methods with machine learning predictive models (Table 2), it is evident that SMOTE variants (SMOTE, SVM-SMOTE, Borderline-SMOTE, KMeans-SMOTE) exhibit diverse performances across different classifiers. Notably, SMOTE and Borderline-SMOTE yield relatively higher UAR scores when combined with LightGBM, positively influencing predictive outcomes. However, their effectiveness diminishes when coupled with AdaBoost. SVM-SMOTE and KMeans-SMOTE, when integrated with GBDT or LightGBM, yield relatively higher UAR scores, contributing to the enhancement of machine learning model performance. Conversely, their combination with AdaBoost results in inferior performance. ROS, when combined with XGBoost, attains higher UAR scores, indicating its beneficial impact on improving XGBoost's performance. These findings suggest that specific combinations of these methods can have a positive impact on model performance in particular scenarios.

Table 2. The average score and the highest score for each oversampling method.

Machine learning models	SMOTE	SVM-SMOTE	Borderline-SMOTE	KMeans-SMOTE	ROS	Average	Max
GBDT	0.6945	0.7221	0.6879	0.7341	0.7121	0.7102	0.7341
LightGBM	0.7143	0.7240	0.7318	0.7214	0.7065	0.7196	0.7318
XGBoost	0.7065	0.7016	0.6938	0.7166	0.7256	0.7088	0.7256
AdaBoost	0.6716	0.6872	0.6541	0.6966	0.7053	0.6830	0.7053
Average	0.6967	0.7088	0.6919	0.7172	0.7124		
max	0.7143	0.7240	0.7318	0.7341	0.7256		

Averaging the UAR scores of each oversampling method reveals that KMeans-SMOTE achieves the highest average UAR score, reaching 0.7172. When combined with GBDT, KMeans-SMOTE attains the highest score of 0.7341. By considering the clustering structure among samples, KMeans-SMOTE generates synthetic samples more reasonably, contributing to the diversity of the training set. This approach may positively impact the

model's generalization ability, resulting in excellent performance. In contrast, Borderline-SMOTE exhibits the lowest score, possibly indicating suboptimal performance when handling samples near category boundaries, leading to lower-quality synthetic samples. This might hinder the model's ability to capture crucial features between categories, thereby affecting performance. The relatively high score of ROS could be attributed to the dataset's features and distribution, allowing simple random oversampling methods to produce favorable effects on model performance.

3.1.2. Feature selection

In this study, pre-training of feature selection stands as a pivotal step in enhancing model performance and reducing computational costs. This method involves the ranking of features to pinpoint those making substantial contributions to model performance. Experiments encompassed different feature quantities, progressively increasing from one feature to encompassing all features. The augmentation of feature quantity might exhibit a pattern of initial enhancement followed by stabilization in performance metrics. The determination of the optimal feature quantity, where the model attains peak performance, is facilitated by monitoring diverse performance indicators, as depicted in Figure 2 (on the next page). Using the combination of KMeans-SMOTE and GBDT that showcasing the highest UAR score, this study illustrates the features optimizing model efficacy. These features are ranked based on their importance, as outlined in Table 3.

Table 3. Achieve the best feature ranking for model performance.

Ranking	Feature	Ranking	Feature
1	Engine power	10	Safety system
2	Accident type	11	Rectification of problems
3	Ship type	12	Time at sea
4	Time	13	Time in rank
5	Month	14	Gross tonnage
6	Depth draft ratio	15	Violation operation
7	Ship age	16	Width / length
8	Location	17	Company culture
9	Safety management		

Upon analysing Table 3, it becomes apparent that prominently influential features mostly pertain to the vessel's inherent characteristics, accident occurrence time and type, navigational environment, company management factors, and human-related factors. Primarily, the vessel's inherent characteristics encompass features such as Engine Power, Ship Type, Depth Draft Ratio, Ship Age, Gross Tonnage, among others. These features are directly related to the vessel's structure, capabilities, and performance, exerting a significant impact on the safety of maritime navigation. Additionally, the time and type of accidents emerge as crucial factors determining the severity of incidents. The accident severity is also impacted by factors such as the ratio of fairway width to ship length, depth draft ratio, and other navigational conditions. The impact on accident severity extends to company management factors, specifically Safety Management, Safety System, and the Rectification of Problems within the shipping operating company. Lastly, human factors are one of the crucial factors influencing the severity of accidents.

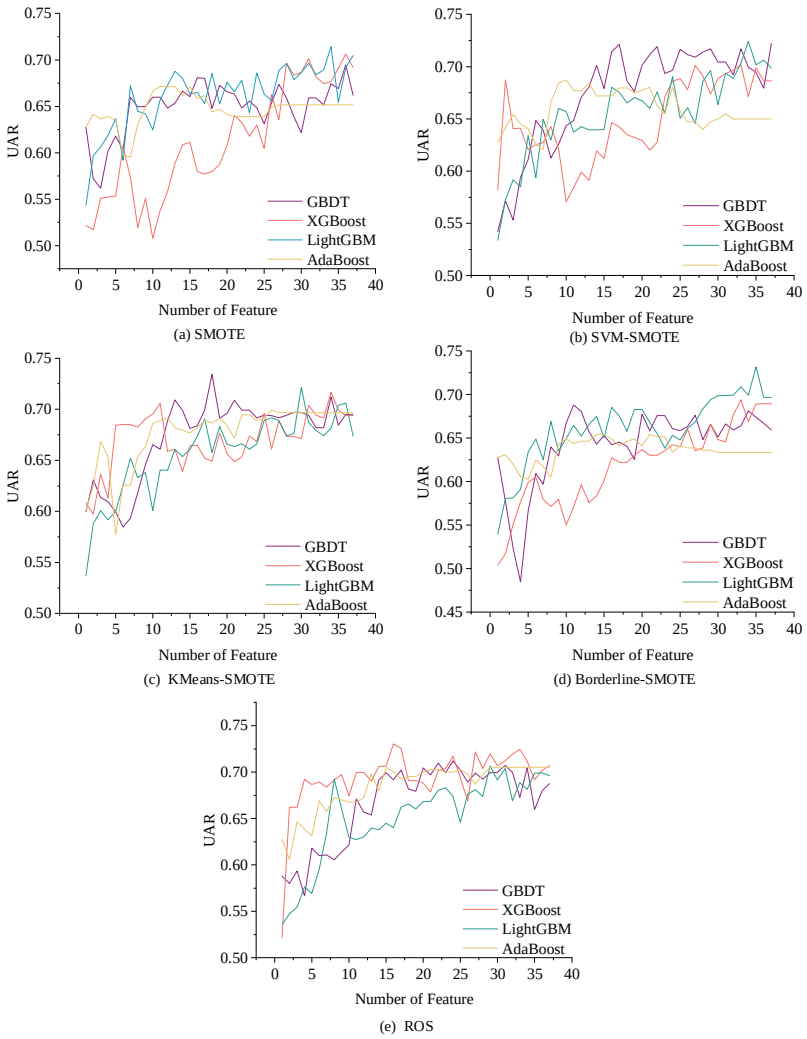

Fig. 2. Changes in the UAR scores after feature selection and oversampling.

3.2. Performance of machine learning model

A detailed examination of the mean and maximum UAR scores (as shown in Table 4) reveals a consistent pattern of excellence exhibited by GBDT, XGBoost, and LightGBM – three gradient boosting decision tree methods that rely on residual fitting. LightGBM, with a highest average UAR score of 0.7196, particularly shines when paired with Borderline-SMOTE, reaching its zenith at 0.7318. Innovative technologies such as Exclusive Feature Bundling are employed by LightGBM, reducing feature dimensions by amalgamating discrete values. This plays a crucial role in enhancing the model's generalization ability, which is a key factor in addressing class imbalance issues. GBDT, with a stellar highest average UAR score of 0.7102, closely followed by XGBoost, underscores the adaptability of these methods when synergized with oversampling techniques, showcasing remarkable information assimilation across diverse classes. This highlights the robust nature of gradient boosting decision tree methods in addressing imbalanced datasets. In contrast, AdaBoost, while securing a comparatively lower score, continues to display a commendable ability to adapt to different data processing methods. This propensity might be linked to AdaBoost's sensitivity to noise, making it responsive to the nuances

of imbalanced data. The ensemble learning strategy of AdaBoost could potentially contribute to overfitting on minority classes, influencing its performance on imbalanced datasets.

Table 4. UAR scores of machine learning methods.

	GBDT	LightGBM	XGBoost	AdaBoost
Average	0.7102	0.7196	0.7088	0.6830
Max	0.7341	0.7318	0.7256	0.7053

3.3. Ablation study

The ablation study serves as a crucial tool for explicating and evaluating the contributions of individual modules or features within machine learning models to the overall model performance. Through the systematic removal or alteration of specific modules in the model, the components influencing the final outcome can be discerned. Employing this method facilitates a more profound understanding of the internal mechanisms and critical components of the model. In this instance, the ablation study involves preserving feature selection and oversampling methods separately, using SVM-SMOTE oversampling as an illustrative example, and gauging the fluctuations in model performance assessed by the UAR score derived from the AdaBoost model. The ablation study results are shown in Figure 3.

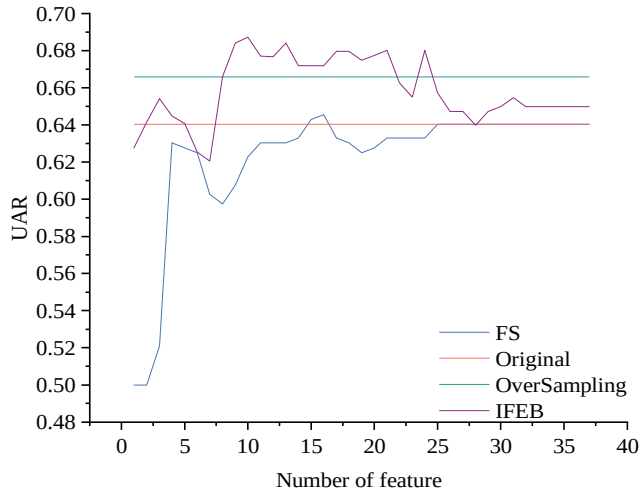

Fig. 3. UAR scores variation in ablation experiments.

Analysing the prediction results from AdaBoost with solely retained feature selection showcases a progressive enhancement in UAR values, culminating in the peak UAR score of 0.6456. Contrasting this with the original data's UAR score of 0.6404 demonstrates a positive impact of 0.81% attributed to feature selection. Under the influence of feature selection, the data is comprehensively understood by the model, leading to an enhanced model performance. The introduction of feature selection played a pivotal role in model performance. Through the utilization of pre-trained feature selection, key features associated with accident severity have been successfully identified., encompassing vessel-specific attributes, accident time and types, navigational contexts, company management variables, and human-related factors. These features hold substantial significance in maritime operations and contribute positively to the model's efficacy. Different oversampling methods exhibited diverse impacts on model performance. KMeans-SMOTE outperformes other oversampling methods, securing the highest average Unweighted Average Recall (UAR) scores across multiple machine learning methodologies, affirming its remarkable proficiency in maritime accident prediction. In contrast, Borderline-SMOTE and other oversampling approaches might prove less effective in specific scenarios. GBDT techniques, encompassing GBDT, XGBoost, and LightGBM, consistently demonstrated stellar performance across diverse oversampling methods, with LightGBM claiming the top spot in average UAR scores. Its incorporation of innovative techniques significantly enhances model generalization, showcasing robust adaptability to imbalanced datasets.

Through ablation experiments, the pivotal contributions of feature selection and oversampling methods to improved predictive performance are further underscored. Retaining feature selection or oversampling methods yielded appreciable performance enhancements, highlighting their effectiveness in strengthening predictive capabilities.

4. Conclusion

This study explored the predictive performance of various oversampling techniques (SMOTE, SVM-SMOTE, Borderline-SMOTE, KMeans-SMOTE, RandomOverSampling) combined with four machine learning methods (GBDT, XGBoost, LightGBM, AdaBoost) in forecasting the severity of maritime accidents.

Through conducting experiments and analysis, this study deeply investigates the roles of feature selection and oversampling methods in predicting the severity of maritime accidents. The strategic combination of models and oversampling methods significantly enhances model performance, offering reliable predictive and decision support for maritime safety. Nonetheless, the study has some limitations, such as not exploring a wider array of machine learning or deep learning models. In future research, additional combination strategies can be explored to identify the optimal pairing of feature selection and oversampling. Furthermore, this study does not introduce advanced technologies from other domains or explore more sophisticated oversampling methods, which could further elevate predictive performance. Subsequent research efforts may benefit from a more nuanced and comprehensive approach to analyse accident data, propelling advancements in the field of maritime accident risk prediction.

Acknowledgements

The authors gratefully acknowledge support from the National Natural Science Foundation of China (Grant No. 52101399), and the Bolian Research Funds of Dalian Maritime University (Grant No. 3132023617). This research is also funded by a European Research Council project under the European Union's Horizon 2020 research and innovation programme (TRUST CoG 2019 864724).

References

Agrawal, A., Viktor, H.L., Paquet, E., 2015. SCUT: Multi-class imbalanced data classification using SMOTE and cluster-based undersampling, 2015 7th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K), 226-234.

Bora, D.J., Navlani, A., Mohari, A., 2022. A Comparative Study and Impact Analysis of Different Oversampling Techniques for CIP. ECS Transactions 107 (1), 4261.

Cai, Q., Abdel-Aty, M., Zheng, O., Wu, Y., 2022. Applying machine learning and google street view to explore effects of drivers' visual environment on traffic safety. Transportation Research Part C: Emerging Technologies 135, 103541.

Cao, Y., Wang, X., Wang, Y., Fan, S., Wang, H., Yang, Z., Liu, Z., Wang, J., Shi, R., 2023a. Analysis of factors affecting the severity of marine accidents using a data-driven Bayesian network. Ocean Engineering 269, 113563.

Cao, Y., Wang, X., Yang, Z., Wang, J., Wang, H., Liu, Z., 2023b. Research in marine accidents: A bibliometric analysis, systematic review and future directions. Ocean Engineering 284, 115048.

Chen, C.W., Tsai, Y.H., Chang, F.R., Lin, W.C., 2020a. Ensemble feature selection in medical datasets: Combining filter, wrapper, and embedded feature selection results. Expert Systems 37 (5).

Chen, M., He, X., Yang, J., Zhang, H., 2018. 3-D Convolutional Recurrent Neural Networks With Attention Model for Speech Emotion Recognition. IEEE Signal Processing Letters 25 (10), 1440-1444.

Chen, P., Huang, Y., Mou, J., van Gelder, P.H.A.J.M., 2019. Probabilistic risk analysis for ship-ship collision: State-of-the-art. Safety Science 117, 108-122.

Chen, T., Guestrin, C., 2016. XGBoost: A Scalable Tree Boosting System, Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Association for Computing Machinery, San Francisco, California, USA, 785–794.

Chen, T., Shi, X., Wong, Y.D., Yu, X., 2020b. Predicting lane-changing risk level based on vehicles' space-series features: A pre-emptive learning approach. Transportation Research Part C: Emerging Technologies 116, 102646.

Fan, S., Yang, Z., Blanco-Davis, E., Zhang, J., Yan, X., 2020. Analysis of maritime transport accidents using Bayesian networks. Proceedings of the Institution of Mechanical Engineers, Part O: Journal of Risk and Reliability 234 (3), 439-454.

Ghorbani, R., Ghousi, R., 2020. Comparing Different Resampling Methods in Predicting Students' Performance Using Machine Learning Techniques. IEEE Access 8, 67899-67911.

Han, H., Wang, W.-Y., Mao, B.-H., 2005. Borderline-SMOTE: A New Over-Sampling Method in Imbalanced Data Sets Learning. Springer Berlin Heidelberg, Berlin, Heidelberg, 878-887.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., Liu, T.-Y., 2017. LightGBM: a highly efficient gradient boosting decision tree, Proceedings of the 31st International Conference on Neural Information Processing Systems. Curran Associates Inc., Long Beach, California, USA, 3149–3157.

Kim, J., Kumar, N., Tsiartas, A., Li, M., Narayanan, S.S., 2015. Automatic intelligibility classification of sentence-level pathological speech. Computer Speech & Language 29 (1), 132-144.
Lal, T.N., Chapelle, O., Weston, J., Elisseeff, A., 2006. Embedded Methods, in: Guyon, I., Nikravesh, M., Gunn, S., Zadeh, L.A. (Eds.), Feature Extraction: Foundations and Applications. Springer Berlin Heidelberg, Berlin, Heidelberg, 137-165.
Lan, H., Ma, X., Qiao, W., Deng, W., 2023. Determining the critical risk factors for predicting the severity of ship collision accidents using a data-driven approach. Reliability Engineering & System Safety 230, 108934.
Li, J., Cheng, K., Wang, S., Morstatter, F., Trevino, R.P., Tang, J., Liu, H., 2017. Feature Selection: A Data Perspective. ACM Comput. Surv. 50(6), Article 94.
Qu, Z., Li, H., Wang, Y., Zhang, J., Abu-Siada, A., Yao, Y., 2020. Detection of Electricity Theft Behavior Based on Improved Synthetic Minority Oversampling Technique and Random Forest Classifier, Energies.
Wang, H., Liu, Z., Wang, X., Graham, T., Wang, J., 2021. An analysis of factors affecting the severity of marine accidents. *Reliability* Engineering & System Safety 210, 107513.
Wang, H., Liu, Z., Wang, X., Huang, D., Cao, L., Wang, J., 2022. Analysis of the injury-severity outcomes of maritime accidents using a zero-inflated ordered probit model. Ocean Engineering 258, 111796.
Wang, H.Y., 2008. Combination approach of SMOTE and biased-SVM for imbalanced datasets, 2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence), 228-231.
Wu, W., Wang, J., Huang, Y., Zhao, H., Wang, X., 2021. A novel way to determine transient heat flux based on GBDT machine learning algorithm. International Journal of Heat and Mass Transfer 179, 121746.