Article Type: Research Paper

Ensemble learning with imbalanced data handling in the early detection of capital markets

Putri Auliana Rifqi Mukhlashin*, Anwar Fitrianto1, Agus M. Soleh1, and Wan Zuki Azman Wan Muhamad2

Abstract
Research aims: This study aims to create an early detection model to predict events in the Indonesian capital market.

Design/Methodology/Approach: A quantitative study comparing ensemble learning models with imbalanced data handling detected early capital market events. This study used five ensemble learning models—Random Forest, ExtraTrees, CatBoost, XGBoost, and LightGBM—to detect early events in the Indonesian capital market by handling imbalanced data, such as under sampling (RUS), oversampling (SMOTE, SMOTE-Broder, ADASYN), and over-under sampling (SMOTE-Tomek, SMOTE-ENN), and weighted (class weight). Global and regional stock markets, commodities, exchange rates, technical indicators, sectoral indices, JCI leaders, MSCI, net buys of foreign stocks, national securities, and national share ownership all predicted the lowest return of Crisis Management Protocol (CMP) binary responses.

Research findings: Hyperparameters and thresholds were tuned to produce the optimum model. The best model had the highest G-mean. ExtraTrees with SMOTE-ENN predicted the highest number of one-day events, with a G-Mean of 96.88%. LightGBM with SMOTE handling best predicted five-day events with an 89.21% G-Mean. With a G-Mean of 89.49%, CatBoost with SMOTE-Border handling was the best for a 15-day event. In addition, LightGBM with SMOTE-Tomek handling and 68.02% G-Mean was best for 30-day events. Further, performance evaluation scores decreased with increased prediction time.

Theoretical contribution/Originality: This work relates more imbalance handling methods and ensemble learning to capital market early detection cases.

Practitioner/Policy implication: Capital markets can indicate economic stability. Maintaining capital market efficacy and economic value requires a system to detect pressure.

Research limitation/Implication: This study used ensemble learning models to predict capital market events 1, 5, 15, and 30 days ahead, assuming Indonesian working days. The model’s forecast results are expected to be utilized to monitor the capital market and take precautions.

Keywords: Capital Market; Early Detection; Ensemble Learning; Imbalance Class; Risk Event

Introduction

Imbalanced data can be handled in various approaches, including under sampling and oversampling. While under sampling is done by reducing the
majority class randomly (Wang & Liu, 2021), oversampling is accomplished by randomly duplicating the minority class (Putri & Dhini, 2019).

The oversampling approach includes SMOTE, SMOTE-Borderline, and ADASYN. The Synthetic Minority Over-Sampling Technique (SMOTE) creates new samples from the minority class by taking the nearest minority class samples. SMOTE-Borderline generates new data along the line between the minority class and its nearest neighbors. Besides, Adaptive Synthetic Sampling (ADASYN) uses a distribution weight for instances in the minority class based on the model's difficulty level of learning data. Faris et al. (2020) showed that SMOTE improves the prediction strength in the geometric mean (G-Mean) and type II errors. However, both under sampling and oversampling techniques are limited (Rahardja et al., 2023). Under sampling can remove important parts of the majority class, while oversampling can cause overfitting in the model. Sir and Soepranoto (2022) found that SMOTE-ENN (Edited Nearest Neighbor), included in the over-under sampling approach, is the best resampling technique. Indrawati (2021) also uncovered that SMOTE-ENN improved the accuracy performance of SVM by 2%–23%. Another handling approach is class weight, which gives more weight to the minority class. This method efficiently handles class imbalance (Asundi et al., n.d.).

Ensemble models combine multiple machine learning models to improve prediction performance and accuracy (Mishraz et al., 2021). They are typically used in classification problems and are better than single models, such as logistic regression, SVM, and neural networks, because the prediction errors of a single model are not always made by another model (Lutfiani et al., 2023). Examples of ensemble models include bagging (bootstrap aggregating) and boosting methods. These methods combine predictions from multiple single models through voting or weighted voting (Mishraz et al., 2021).

Several models within the bagging method yield good evaluation results, such as Extremely Randomized Trees (ExtraTrees) and Random Forest (RF). Bagging and boosting models outperform the others in predicting bank failure, with the RF model having the highest area under the ROC curve (AUC) value at 93% (Liu et al., 2021). Thakkar and Chaudhari (2021) found that RF produced the best results, with an accuracy of 90.4%. According to research by Islam et al. (2019), Extremely Randomized Trees (ExtraTrees) generated comparatively better results with a ROC-AUC value of 94.2%. ExtraTrees also exhibited promising precision performance—more than one-period prediction (Aini et al., 2023). Meanwhile, models that fall under the boosting method include Categorical Boosting (CatBoost), Extreme Gradient Boosting (XGBoost), and Light Gradient Boosting (LightGBM). Research by Aly et al. (2022), using a four-period class imbalanced dataset, showed that CatBoost with oversampling SMOTE provides a high accuracy result in predicting bankruptcy in Poland, with an AUC value in each period of more than 90% (Santoso et al., 2022). Carmona et al. (2019) revealed that XGBoost provided the best result with an accuracy of 95%. With its ability to prevent overfitting and create predictions that can be applied generally, XGBoost capability increases accuracy in bank failure prediction. In the study, XGBoost outperforms both the conventional method (Logistic Regression) and the modern machine learning approach (Random Forest). A study by Wang et al. (2022) also demonstrated that LightGBM was better than Decision
Tree (DT), K-Nearest Neighbor, and RF in predicting bankruptcy with an F1-Score of 87.63%.

On the other side, the capital market is a platform that brings together parties that need capital and investors. The capital market can be an indicator of a country’s economic stability and needs to be monitored to provide optimal benefits. In this regard, early warning systems can help reduce negative impacts on the capital market. Pressure on the capital market can also be caused by various factors, such as technical indicators, macroeconomics, portfolio management, regional or global integration, monetary policy, the global financial crisis, and behavioral economics factors such as public mood, news, risk aversion, and consumer confidence (Hermawan et al., 2023). Specifically, in machine learning modeling, pressure events are rare occurrences called class imbalance problems. Imbalanced class classification is where the majority class distribution has a larger proportion than the minority class. The class ratio is divided into three categories: mild when the minority class proportion is 20% to 40%, moderate when it is 1% to 20%, and extreme when it is less than 1% of the total data available (Google Developers, 2021).

Regarding the Chinese capital market early warning system, the RF algorithm has the highest accuracy for identifying bond market crises using under sampling and double-sampling methods, with 96.08 and 90.2%, respectively (Zhang & Chen, 2022). In developing an early warning system for predicting stock market crises in China based on market indicators and mixed frequency investor sentiments, the Artificial Neural Network (ANN)-based model demonstrates more stable performance, with an accuracy range of 98% to 99% (Pramono et al., 2022). For other models, such as Support Vector Machine (SVM), Decision Tree (DT), Random Forest (RF), Gradient Boosting (GBDT), K-Nearest Neighbor (KNN), and Logistic Regression (LR), the prediction accuracy is greatly influenced by different stock markets (Lu et al., 2021). On early warning, Wang and Liu (2021) showed that a long short-term memory (LSTM) network produces satisfying performance with a test-set accuracy of 96.4% and an average of 2.8 days of forewarning. Cross-validation, back-testing, and a reality check demonstrate the model’s reliability and practical value in real-time decision-making.

Using machine learning models, research on class imbalances for the Early Warning System (EWS) or early detection in the capital markets is still rare, especially in Indonesia. This study, therefore, aims to use ensemble learning models for the early detection of events in the capital market in Indonesia with the handling of imbalanced data for predictions 1 day, 5 days, 15 days, and 30 days in advance, assuming working days in Indonesia are Monday through Friday. This research is anticipated to contribute knowledge about handling imbalances from multiple data periods with ensemble learning in situations involving the early detection of capital markets in Indonesia or other similar mattresses. In practice, this research can be used to monitor and detect events in the Indonesian capital market to assist with decision-making (Putri et al., 2023).
Literature Review and Hypotheses Development

The Early Warning System (EWS) in the field of the economic crisis has been developed for a long time. Many works of literature use statistical models as the EWS. However, with the advancement of technology, new modeling types known as ensemble learning have emerged (Bintoro et al., 2023). Many studies have been conducted to investigate whether these ensemble learning models can accurately predict crises (Candra et al., 2023).

Ensemble learning can avoid overfitting and bias in the model because it creates multiple machine learning algorithms to determine the optimal value by majority voting. In the case of the early warning system for bank failures, ensemble learning has the greatest accuracy compared to all other SMOTE-based methods for converting imbalanced to balanced data (Shrivastava et al., 2020). Gnip and Drotár (2019) have compared several ensemble machine learning methods applied to a recently acquired dataset of small and medium-sized enterprises in the Slovak Republic. In certain instances, the highest obtained prediction accuracy of the proposed classification models, measured by geometric mean, is nearly 100% (Hariguna et al., 2022).

A study (Bluwstein et al., 2021) using macro-financial data from 17 countries from 1870 to 2016 approached the machine learning model in the stock market crisis study, where it applied one of the ensemble models, Extremely Randomized Trees (Extra Trees) and produced the most accurate results. Tölo (2020), in predicting systemic financial crises one to five years ahead, which includes the crisis dates and annual macroeconomic series of 17 countries over the period 1870-2016, found that machine learning models can be significantly improved by using Long-Short Term Memory (RNN-LSTM) and Gated Recurrent Unit (RNN-GRU) neural nets (Marlina et al., 2023). In another study, Coffinet and Kien (2019) focused on detecting rare events in banking crisis cases, using data collected from 32 European and non-European countries from 2010 to 2017 (Pratama & Wijaya, 2023). The random forest method was found to be the best approach for computing an indicator for the probability of a banking crisis (Zanubiy et al., 2023). In the case of bank insolvencies with non-failed, failed, and assisted entity data from the Federal Deposit Insurance Corporation (FDIC) database from 2008 to 2014, the Random Forest model provided the best results by considering the flow of data and had more stable and consistent performance in all test samples (Petropoulos et al., 2020). The Random Forest, a modern classification tree ensemble technique, provides the best results with an accuracy of 90.4%, and includes global credit and real estate variables as predictors (Thakkar & Chaudhari, 2021).

Furthermore, the objective of Carmona et al. (2019) was to anticipate bank failure in the United States financial system (Mahardika & Irawan, 2022). Expanded Gradient Boosting was used for empirical analysis (Sipahutar et al., 2020). This method evolved from previous boosting methods such as AdaBoost and boosted classification trees (Widiastuti et al., 2023). The XGBoost algorithm’s applicability and ability are to increase the accuracy of bank failure prediction (Kosasi et al., 2022). In addition, an early warning system predicts banking crises to identify excessive credit growth and aggregate leverage. Junyu (2020) showed that the accuracy of XGBoost has reached approximately 80%, providing a
novel method for predicting financial crises. Lu et al. (2021) have also proven that, among single classifiers, the combination of XGBoost and random under sampling outperforms the random forest in predicting the probability of healthy enterprises. The probability of financial crisis enterprises in the t-period is maintained at 92.86%, the misjudgment rate of normal enterprises is reduced to approximately 12%, and the overall prediction accuracy is enhanced through a straightforward integration of the Random Forest and XGBoost (Tussa’diah & Kartika, 2023).

Research Method

Data

This study used data from the Financial Services Authority. The predictor variables used were all numeric variables that fell into several categories, including Global Regional Stock Market, Commodities and Exchange Rates, Technical Indicators, Sector Indices, IHSG Leaders, Morgan Stanley Capital Indonesia (MSCI), Foreign Net Buy/Sell Stocks, and Government Securities (SBN) and SBN ownership, with a total of 2424 variables. The response variable employed was the event of the lowest Crisis Management Protocol (CMP) return. If the value is less than minus 5%, it is "Pressure," while if the value is greater than or equal to minus 5%, it is "Normal" (Soesilo and Tinggi, 2021). The time range of the data was from January 2010 to November 2020. A list of variables used can be found at bit.ly/thesisvariables. Afterward, this study compared ensemble learning models with imbalanced data handling to further predict the Indonesian capital market. In this regard, ensemble learning creates a more accurate and complete strong model by combining weak classifier models. Bagging and boosting are two ways to combine poor classifiers with strong ones. The bagging method generates base classifiers in parallel, while the boosting method successively generates them and influences later classifiers.

Analysis Process

This study took the following steps: data preprocessing, feature selection, data division, handling class imbalance, ensemble learning modeling, model improvement, and evaluation. Each stage had subprocesses involved. The detailed flow is shown in Figure 1.

Data Preprocessing

This process includes data cleaning, Exploratory Data Analysis (EDA), and feature engineering. Data were cleaned to remove or modify irrelevant, duplicate, and unformatted data. In EDA and Feature Engineering, the addition of the return variable from each predictor variable, which is a daily price index, and the conversion of the daily price index predictor variable into stock return value was carried out. \( P_t - P_{t-1} \) is the stock price at period \( t \), and \( P_{t-1} \) is the stock price at the previous period \( t-1 \).
In addition, other variables were added by creating moving average (MA) values. For example, the previously used ihsg_daily variable, the IHSG daily price index, was replaced with the return value ihsg_daily_return, and the moving average value was added, for example, ihsg_daily_return_movavg_five, or the five-day moving average of its daily return data. The MA value in this research was added to MA(k), with k = 5, 10, 15, and 30. Subsequently, data exploration was performed to identify the data patterns before and after each predictor variable’s “Pressure” event. This pattern served as the basis for determining the change of the original variable into the time lag variable, such as in the example ihsg_daily_return_lag_i with i = 1, 2, 3, 4, and 5 to predict 1 day ahead. The original variable was replaced with the time lag variable (see Table 1) to avoid information leakage during the machine learning modeling.
Mukhlashin, Fitrianto, Soleh, & Muhamad
Ensemble learning with imbalanced data handling ...

Table 1 Creating a time lag variable

<table>
<thead>
<tr>
<th>Period</th>
<th>X</th>
<th>x lag 1</th>
<th>x lag 2</th>
<th>...</th>
<th>x lag i</th>
<th>flag</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-2.86</td>
<td>NA</td>
<td>NA</td>
<td>...</td>
<td>NA</td>
<td>Normal</td>
</tr>
<tr>
<td>2</td>
<td>-1.72</td>
<td>-2.86</td>
<td>NA</td>
<td>...</td>
<td>NA</td>
<td>Pressure</td>
</tr>
<tr>
<td>3</td>
<td>0.56</td>
<td>-1.72</td>
<td>-2.86</td>
<td>...</td>
<td>NA</td>
<td>Normal</td>
</tr>
</tbody>
</table>

X column is not utilized in the modeling process.

Feature Selection

Many initial variables were the basis for the feature selection process. Feature selection was performed using the Random Forest algorithm to remove weakly influencing variables and improve accuracy and classification performance (Chen et al., 2020). A feature selection was performed for each scenario, a prediction for 1 day, 5 days, 15 days, and 30 days ahead.

Data Splitting

The data division was carried out using the expanding window time series method (see Figure 2). This approach is called forward-chaining cross-validation (Vien et al., 2021). The data division was done annually by dividing the data into eight parts. However, 2014 and 2017 were not used as testing data due to the absence of "Pressure" events in those years.

Handling Class Imbalance Issue

Based on Table 2, the proportion of "Pressure" events is 0.017, or 44 days out of 2620. It can be categorized as an extreme to moderate imbalance (Google Developers, 2021). Handling imbalanced class techniques are divided into four categories: under sampling (RUS), oversampling (ROS, SMOTE, SMOTE-Borderline, ADASYN), over-undersampling (SMOTE-Tomek, SMOTE-ENN), and weighting (class weight).

Ensemble Learning Modeling

The modeling was carried out on the training data using five models: ExtraTrees (Alfian et al., 2022), RF (Speiser et al., 2019), CatBoost (Jabeur et al., 2021), XGBoost (Qiu et al., 2021), and LightGBM (Sun et al., 2020). Each model was approached with and without the handling of class imbalance. It was also done for each prediction scenario.
Table 2 The proportion of pressure event labels based on the year

<table>
<thead>
<tr>
<th>Year</th>
<th>Pressure Event</th>
</tr>
</thead>
<tbody>
<tr>
<td>2010</td>
<td>1.63%</td>
</tr>
<tr>
<td>2011</td>
<td>4.45%</td>
</tr>
<tr>
<td>2012</td>
<td>0.41%</td>
</tr>
<tr>
<td>2013</td>
<td>3.20%</td>
</tr>
<tr>
<td>2014</td>
<td>0.00%</td>
</tr>
<tr>
<td>2015</td>
<td>0.81%</td>
</tr>
<tr>
<td>2016</td>
<td>0.40%</td>
</tr>
<tr>
<td>2017</td>
<td>0.00%</td>
</tr>
<tr>
<td>2018</td>
<td>1.24%</td>
</tr>
<tr>
<td>2019</td>
<td>0.41%</td>
</tr>
<tr>
<td>2020</td>
<td>6.16%</td>
</tr>
</tbody>
</table>

Hyperparameter Optimization and Threshold

The hyperparameter optimization technique was performed using Optuna. Optuna is effectively used on the XGBoost model (Srinivas & Katarya, 2022). A total of 30 trials were performed to determine the best hyperparameter value of the model with imbalance class handling, such as the number of estimators, which is the number of decision trees, and max depth, which is the maximum depth of the decision tree. Threshold tuning was also performed using Optuna to obtain the optimal value based on the G-Mean.

Model Evaluation

This research aimed to predict the "Pressure" event for prevention and "Normal" for market capitalization monitoring. The evaluation was performed on the test data using the G-Mean value, maximizing both true positive and true negative while keeping both relatively balanced. The highest G-Mean value on the modeling algorithm is the basis for choosing the best model.

Results and Discussion

Data Exploration

The general pattern of the predictor variables with the pressure event was seen in this process. From the following Figure 3, to see the occurrence of pressure on the next day, on average, the predictor variables decreased in the previous 4 to 6 days. Then, the predictor variables decreased to see if there was pressure in the next five days, as seen in the previous 10 to 15 days. Meanwhile, to see pressure events in the next 15 days, the predictor variables decreased over the past 22 to 25 days. In addition, to predict the next 30 days, no pattern differed significantly from the predictor variable. Data exploration depicted that the predictor variables could detect pressure events before they happened.
Feature Selection

Feature selection was (see Figure 4) made to obtain variables that strongly influenced predicting pressure events in the capital market. In addition, it is expected to improve the performance of the classification algorithm in terms of computation. Fifty predictor variables were selected from the feature importance Random Forest model.

In the prediction of pressure events for the next 1 day, the variables with a high impact on the occurrence of pressure on the capital market were the IHSG, sectoral indices (financial, manufacturing), leading stocks (LQ45, Kompas) and the MSCI index (MXID) which were the highest 50 variables or 2.1% of all variables, based on RF prediction 1 day ahead feature importance covering 23.23% importance of the model. Then, the mining sector, indices (PCOMP, OMX), and exchange rates (YEN, YUAN) for the next 5 days covered 12.70%. Furthermore, for the next 15 days, foreign investors’ SBN holdings, the exchange rate (YEN), the mining sector, cons good, global/regional stock index returns (SZCOMP, IBEX, TOPIX), and commodity prices (BRENT) covered 10.77%. Then, for the next 30 days, foreign investors’ SBN holdings, exchange rates (YEN, USD), leading stocks (Kompas), the IHSG in the agribusiness sector, finance, and the MSCI index (MIXD) covered 16.03%.

Ensemble Model Results

The results of prediction modeling for the next 1-day (see Table 3) showed that the RF model with SMOTE-ENN treatment produced a G-Mean value of 0.9628 with a max depth of 2 and several decision trees of 42 at a threshold of 0.5393, in which this handling was better than other RF models. ExtraTrees with SMOTE-ENN handling produced the best performance value with a G-Mean of 0.9668, max depth of 4 and number of decision trees of 82 at a threshold of 0.5141. The CatBoost model with SMOTE-Border handling produced the highest G-Mean value of 0.9533 at max depth 4, and the number of decision trees was 710, the threshold of 0.4656. Likewise, with CatBoost, the XGBoost model handled SMOTE-Border with a max depth value of 4 and several decision trees of 86 and, at an optimal threshold of 0.5852, produced the highest G-Mean value of 0.9324. Whereas in LightGBM, SMOTE-ENN produced a good performance as seen from G-Mean 0.9486 with optimal threshold and hyperparameters of 0.5525 with a max depth of 1 and a total of 22 decision trees. In general, the ExtraTrees model with SMOTE-ENN handling was the best algorithm in predicting events in the capital market 1 day ahead with a higher
Mukhlashin, Fitrianto, Soleh, & Muhamad

Ensemble learning with imbalanced data handling...

G-Mean value of around 2% - 5% compared to other models at the same treatment and 1% - 3% compared to other best algorithms. Some models produced low G-Mean (<50%), i.e., without handling, RUS on LightGBM models, ROS on CatBoost and XGBoost models, class weight on CatBoost, XGboost, and LightGBM models.

Furthermore, for predicting events for the next 5 days (see Table 3), the RF model with SMOTE-Border handling produced the best value compared to other RF models with a G-Mean value of 0.8529 at max depth 4, and the number of decision trees was 562 at a threshold of 0.6384. The ExtraTrees-SMOTE-Border algorithm produced the highest G-Mean of 0.8614 with a max depth of 4 and 82 decision trees at a threshold of 0.5141. The CatBoost model with SMOTE handling produced the highest G-Mean value of 0.8543 at max depth 4, and the number of decision trees was 541 at a threshold of 0.6955. The XGBoost model with SMOTE-Tomek handling produced the highest G-Mean value of 0.8812 in the optimal hyperparameter combination max depth of 4 and the number of decision trees of 881 at a threshold of 0.6312. The LightGBM model with SMOTE handling...
with a max depth of 4 and several decision trees of 991 and at a threshold of 0.7590 produced the highest G-Mean value of 0.8921. Viewed as a whole, for predictions in the next 5 days, LightGBM with SMOTE treatment produced the highest G-Mean value. Compared to other models with SMOTE handling, the LightGBM-SMOTE algorithm showed a higher performance value, with a higher G-Mean value of about 1% - 14% higher than the same treatment and 1% - 4% compared to the other best algorithms. Some treatments produced a low G-Mean (0 - 50%), i.e., without overall handling producing a bad G-Mean value of less than 30%, RUS and class weight less than 40%, and ROS less than 50%.

The RF-SMOTE-ENN algorithm, with an optimal threshold of 0.5453, max depth of 4, and several decision trees of 897, was the best model for predicting events for the next 15 days, with the highest G-Mean of 0.7889 (see Table 3). The ExtraTrees model with SMOTE handling produced the highest G-Mean of 0.7573 with a max depth of 4, the number of decision trees was 60, and the threshold was 0.5285. The CatBoost model with SMOTE-Border handling produced the highest G-Mean value of 0.8949, optimal at a max depth of 4, number of decision trees of 798, and threshold of 0.3413. The XGBoost-SMOTE-Border algorithm revealed the best results with a G-Mean value of 0.8336 obtained after tuning with a threshold of 0.4006, a max depth of 2, and several decision trees of 490. In the LightGBM model, handling SMOTE-ENN produced the best G-Mean value of 0.8245 at the optimal max depth of 4, the optimal number of trees of 895, and the threshold of 0.7426. In general, the CatBoost-SMOTE-Border model was the best for predicting capital market events for the next 15 days. Compared with other SMOTE-Border models, it produced higher G-Mean values of 6% - 28% and 1% - 4% compared to the other best algorithms. Models that did not handle the imbalance class or use other handling methods (weight class, RUS, ROS) produced low G-Mean values (<30%).

Predicting the pressure events in the capital market earlier can help take more effective preventive measures. However, the predicted results for the next 30 days (see Table 3) were not good enough compared to predictions for other shorter periods. The RF model with SMOTE-Tomek handling produced the highest G-Mean value of 0.5968 at the threshold of 0.6007, max depth of 4, and the number of trees 32. The ExtraTrees-ADASYN algorithm produced the highest G-Mean value of 0.4162 at the threshold of 0.5255, max depth of 4, and the number of decision trees of 45. In the LightGBM model, the SMOTE-ENN treatment produced the highest G-Mean value of 0.6468 at a max depth of 4, a number of decision trees of 463, and a threshold of 0.4795. The XGBoost model with SMOTE handling produced the highest G-Mean value of 0.6632 at a max depth of 4, and the number of decision trees was 646. The LightGBM model with SMOTE-Tomek handling at a threshold of 0.6318 with a max depth of 3 and several decision trees of 224 produced the highest G-Mean value of 0.6802. Generally, compared to the SMOTE-Tomek handling model, LightGBM uncovered a higher G-Mean value of about 5% - 48% and 2% - 26% compared to the other best algorithms. Meanwhile, models with imbalanced class weight, RUS, and ROS handling produced G-Mean values below 50%.
The best model for every scenario yielded a different handling (see Figure 5). Not only did it have a high geometric mean value, but it also yielded high values in other metrics, such as the F1 score, indicating that it was correctly identifying most of the pressure events in the dataset and making accurate predictions, recall, or the proportion of actual pressure events correctly predicted.

**Table 3 Evaluation results of prediction model (in percent)**

<table>
<thead>
<tr>
<th>Scenario</th>
<th>Model</th>
<th>G-Mean value from Imbalanced Class Handling Results</th>
<th>Class Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>None</td>
<td>RUS</td>
</tr>
<tr>
<td>1 day ahead</td>
<td>RF</td>
<td>0.6628</td>
<td>0.9051</td>
</tr>
<tr>
<td></td>
<td>ExtraTrees</td>
<td>0.7672</td>
<td>0.8518</td>
</tr>
<tr>
<td></td>
<td>CatBoost</td>
<td>0.7147</td>
<td>0.9132</td>
</tr>
<tr>
<td></td>
<td>XGBoost</td>
<td>0.4252</td>
<td>0.7537</td>
</tr>
<tr>
<td></td>
<td>LightGBM</td>
<td>0.1925</td>
<td>0.4514</td>
</tr>
<tr>
<td>5 days ahead</td>
<td>RF</td>
<td>0.2108</td>
<td>0.3824</td>
</tr>
<tr>
<td></td>
<td>ExtraTrees</td>
<td>0.2324</td>
<td>0.1939</td>
</tr>
<tr>
<td></td>
<td>CatBoost</td>
<td>0.1096</td>
<td>0.389</td>
</tr>
<tr>
<td></td>
<td>XGBoost</td>
<td>0.2507</td>
<td>0.2213</td>
</tr>
<tr>
<td></td>
<td>LightGBM</td>
<td>0.0625</td>
<td>0.1537</td>
</tr>
<tr>
<td>15 days ahead</td>
<td>RF</td>
<td>0.0375</td>
<td>0.2042</td>
</tr>
<tr>
<td></td>
<td>ExtraTrees</td>
<td>0</td>
<td>0.2861</td>
</tr>
<tr>
<td></td>
<td>CatBoost</td>
<td>0.0437</td>
<td>0.1803</td>
</tr>
<tr>
<td></td>
<td>XGBoost</td>
<td>0.104</td>
<td>0.2627</td>
</tr>
<tr>
<td></td>
<td>LightGBM</td>
<td>0</td>
<td>0.2100</td>
</tr>
<tr>
<td>30 days ahead</td>
<td>RF</td>
<td>0.1462</td>
<td>0.1783</td>
</tr>
<tr>
<td></td>
<td>ExtraTrees</td>
<td>0</td>
<td>0.205</td>
</tr>
<tr>
<td></td>
<td>CatBoost</td>
<td>0.0422</td>
<td>0.4616</td>
</tr>
<tr>
<td></td>
<td>XGBoost</td>
<td>0.0707</td>
<td>0.3512</td>
</tr>
<tr>
<td></td>
<td>LightGBM</td>
<td>0.0371</td>
<td>0.2747</td>
</tr>
</tbody>
</table>

*The bold value indicates the highest evaluation value among all models.
*The highest evaluation value for the Random Forest (RF) model
*The highest evaluation value for the ExtraTrees model
*The highest evaluation value for the CatBoost model
*The highest evaluation value for the XGBoost model
*The highest evaluation value for the LightGBM model
model with SMOTE-Tomek handling was the best, with all evaluation metrics greater than 60%. As can be seen, the evaluation's effectiveness degrades as the prediction grows even further. Generally, this study aligns with the results of Indrawati (2021) and (Shrivastava et al., 2020), which found that SMOTE-Based oversampling is an effective method for handling class imbalance.

![Figure 5 Summary of the Best Model](image)

**Feature Importance**

Based on the best algorithm ExtraTrees-SMOTE-ENN, it can be seen that the variables at a time lag 1 had the most impact, which can be said that predicting the pressure event 1 day ahead was influenced by the event 1 day before (see Figure 6). The IHSG, industrial sector, trade, manufacturing, and MSCI Indonesia Index impact predicting the 1 day ahead event. The LightGBM-SMOTE algorithm showed that the most influential variables for predicting market events 5 days ahead were in the range of 5-14 days, namely indices and foreign stock exchanges such as the return of China's Yuan 8 days, the Philippine stock exchange 6 days, the Australian stock exchange 6 days, the return of Thailand's index 6 days ago, SBN, and the consumer goods industry sector. Based on the best algorithm CatBoost-SMOTEBorder, which predicts 15 days ahead, the most influential factors were in the range of 16-25 days ago, which include SBN on days 16 and 20, the USD index on day 22, Japan TOPIX on day 25, the XAUD sector which is the gold/silver sector, Brent oil on days 19 and 25, trade, industry, utility, and transportation sectors with significant impact. Based on the best algorithm LightGBM-SMOTE-Tomek, it can be seen that the moving average of foreign net buy/sell shares, the USD index, Shenzhen, Yuan, the moving average MSCI, and some sectors such as basic industry, agriculture, and manufacturing, Brent oil, SBN, and the CSI300 and S&P 500 indices 30 days before are important factors in predicting events in the market 30 days ahead.
Muklashin, Fitrianto, Soleh, & Muhamad

Ensemble learning with imbalanced data handling ...

Figure 6. Feature Importance Prediction Model (a) 1 Day Ahead; (b) 5 Days Ahead; (c) 15 Days Ahead; and (d) 30 Days Ahead

Conclusion

This study used ensemble learning modeling with and without handling the imbalance class to detect events in the Indonesian capital market. There are differences in the best algorithm for each scenario. For a 1-day prediction, the ExtraTrees model with SMOTEENN handling had the highest G-Mean value of 0.9668. Then, for a 5-day prediction, the LightGBM algorithm with SMOTE handling had a G-Mean value of 0.8921. For a 15-day prediction, the CatBoost algorithm with SMOTE-Border handling had a G-Mean value of 0.8949. Furthermore, for a 30-day prediction, the LightGBM algorithm with SMOTE-Tomek handling had a G-Mean value of 0.6802. In conclusion, the further the prediction, the weaker the model’s performance. In this study, effective methods for handling the class imbalance problem in machine learning models were oversampling techniques (SMOTE, SMOTEBorder) and over-under sampling (SMOTE-ENN, SMOTE-Border). On the other hand, random under sampling (RUS), random oversampling (ROS), and class weight
Mukhlashin, Fitrianto, Soleh, & Muhamad

Ensemble learning with imbalanced data handling...

were less effective methods for handling the class imbalance problem in machine learning models in this study.

Based on the study's findings, it is hoped it will be able to contribute knowledge about how to handle imbalances from multiple data points with ensemble learning in scenarios involving early detection of the Indonesian capital market or other similar situations in the application of machine learning, which has previously rarely been studied and can be further explored. It is also expected that this research can be utilized to monitor and detect events in the Indonesian capital market to assist in decision-making because the metrics evaluation findings show excellent performance in detecting events in the capital market. In this study, the ensemble learning model was only utilized for bagging and boosting purposes; there was no comprehensive description of the ensemble learning model's role in predicting how events would affect the capital market. This study also has limitations regarding the newest data. Therefore, for future research, the best existing models can be combined with other ensemble learning methods, such as stacking, and the feature importance of the best model can be explained with more advanced interpretations, such as LiME (Local Interpretable Model Agnostic Explanation).

References


Asundi, R. V., Prakash, R., & Kumar, K. (n.d.). Class Weight technique for Handling Class Imbalance.


Journal of Accounting and Investment. 2023 | 614


