Open Peer Review on Qeios

Chronic disease treatment default prediction, potential benefits with applied random sampling optimization techniques

Michael Owusu-Adjei[1], James Ben Hayfron-Acquah[1], Twum Frimpong[1], Gaddafi Abdul-Salaam[1]

1 Kwame Nkrumah University of Science and Technology

Funding: No specific funding was received for this work.

Potential competing interests: No potential competing interests to declare.

Abstract

A characteristic feature of real-world applications is the occurrence of dataset class imbalance in the output class distribution common in practical business applications such as spam filtering and fraud detection. Predictive modeling contributions from the minority or underrepresented class are overlooked by most learning algorithms. Addressing this challenge includes applying re-sampling techniques that eliminate class distribution imbalance for a more balanced output class distribution in the training examples. Random sampling techniques such as random over-sampling of the minority class duplicates the minority class examples to achieve a more balanced distribution or random under-sampling to delete training examples in the majority class for a balanced distribution to eliminate class imbalance in the dataset and finally Synthetic minority over-sampling technique which is designed to generate synthetic examples for the minority class to address class imbalance. The usefulness of these random sampling techniques has received attention in several research studies, particularly for binary classifications in two-class or multi-classification problems. This application to many is aimed at achieving equal class distribution meant to determine optimal model performance. This comparative assessment of random sampling optimization show significant difference in model performance on applied random sampling technique use and also confirms that even with poor performance in false alarms, high prediction accuracy score and roc_auc score can be achieved.

Michael Owusu-Adjei[1,*], **James Ben Hayfron-Acquah**[1], **Twum Frimpong**[1], **Gaddafi Abdul-Salaam**[1]

[1] *Kwame Nkrumah University of Science and Technology, Kumasi, Ghana*

[*]Corresponding author's e-mail: mowusuadjei@st.knust.edu.gh

Keywords: Sampling, benefits, prediction, accuracy, impact, disease, treatment, SMOTE.

1. Introduction

Common among real-world applications is the occurrence of unequal class distribution in target labels. In supervised learning, variables are labeled for identification. Output class may consist of two or more classes for predictive classification modeling. Predictive modeling techniques that are capable of correctly identifying predominantly higher number of target classes define its performance potency. However, the challenge of predictive modeling techniques making biased predictions with skewed class distribution persists and continues to attract research interest. This phenomenon leads to a situation in which minority class contribution, effect and impact are ignored by learning algorithms. Meanwhile, in many real-world applications such as healthcare systems, fraud detections etc., minority class contributions are particularly of paramount interest. For example, the detection of fraudulent transactions in a banking system, spam email detection system etc., may reveal an exceedingly higher number of normal transactions than fraudulent ones, but for which the detection of fraudulent transactions is of utmost priority. Similarly, the prediction of infectious diseases(Yang et al., 2020)(Santangelo et al., 2023)(Ray & Reich, 2018) among patients could also show that the majority of patients who may not have the disease but knowledge of patient minority with the disease is of paramount importance to contain any potential spread or large scale outbreak and any devastating impact on public health. Additionally, risk assessment of default by patients to disease treatment(Gichuhi et al., 2023)(Korneev et al., 2022) and other risk assessments for default may also encounter class distribution imbalance. Addressing dataset class imbalance to improve prediction accuracy scores continues to gain attention in many related research works. Solutions to dataset imbalance range from using sampling techniques that address class imbalance through over-sampling the minority class or under-sampling the majority class to other optimization techniques, such as applying class weight optimization to penalize the minority class for each prediction error made or using synthetic minority oversampling technique designed to generate synthetic samples for the minority class. A key concern for addressing class imbalance is why and for what purpose and to what benefit and for what impact? This legitimate concern arises because a characteristic feature of real-world applications is the phenomenon of dataset class imbalance. Assessing potential impact on model performance to determine benefits of applied techniques that includes pre-sampling performance evaluation creates an in-depth understanding of its use.

1.1. *Research Objectives*

The potential benefits of applied random sampling techniques for predictive modeling is to determine model performance benefits and to assess its impact through a rigorous evaluation process that includes model generalization concepts on such datasets with minority class involvement common in practical business applications. Research focus includes extensive impact assessment to determine potential benefits of other evaluation metrics such as model prediction accuracy score, balanced accuracy score, false alarms (true negative rates-TNR, false negative rates-FNR, true positive rates-TPR, false positive rates-FPR), model generalization that explains model behavior on acceptable number of training examples with five most commonly used predictive models. This key challenge has received little attention in related research works concerning practical business applications such as healthcare systems where the effect of incorrect prediction on disease treatment management process could result in significant damage to patient-healthcare personnel relationship including unfavorable treatment outcome. Impact of patient age on treatment is examined to estimate whether particular age group is associated with treatment default for target intervention.

1.2. *Research significance*

Predictive modeling offers great insight into patterns of change necessary for decision making. Its application in healthcare and other practical business applications offer greater opportunity for better insight that facilitates faster, safer, accurate and timely decision making in critical situations. Applied model learning curves in settings where repetitive task is performed as in healthcare has the potential to generate insight into how limited resources could be applied more efficiently to improve job performance. Comparative assessment of model performance using under-sampling, over-sampling and synthetic minority over-sampling techniques creates a better understanding of impact assessment to determine best model evaluation approach. By using five distinct classification algorithms, this study offers detailed, well rounded comprehensive assessment of model performance in this scenario providing insight into how feature evaluation of model performance in such practical business applications with dataset class imbalance could be enhanced. Additionally, the use of machine learning curves could also help in the selection of future model techniques for similar tasks.

1.3. *Related Works*

In this section, a preview of related research works showcasing random sampling technique use is examined for emphasis. Using transaction details of customers, the application of random sampling techniques showed high over-sampling technique performance over under-sampling (Mohammed et al., 2020). Applied over-sampling techniques on large datasets to address challenges associated with class imbalance drew conclusions about its efficiency against other sampling techniques (Rodríguez-Torres et al., 2022). Similarly, a comparative study of several re-sampling techniques to determine efficiency also concluded that over-sampling is an efficient sampling approach (Overview, 2014). Improving prediction accuracy in instance method selection (Hernandez et al., 2013) with over-sampling and under-sampling techniques showed accuracy improvements in minority class prediction. Understanding superior differences in performance of over and under-sampling techniques with an exploration of dataset inner structure (García et al., 2020) to explain the superior performance of over-sampling technique. A new approach (Lee & Kim, 2021) for determining the required over-sample size, which is less than the required sample for achieving class balance, has been evaluated to conclude that this approach improves classification performance when used with over-sampling. Comparative analysis of over-sampling and under-sampling influence on physical activities with ensemble machine learning techniques determines that under-sampling with refined features addresses class imbalance more effectively(Jeong et al., 2022). Another approach to effectively address class imbalance includes the use of synthetic minority over-sampling (Sowjanya & Mrudula, 2023).A combination of three sampling techniques – over-sampling, under-sampling and combi-sampling (Saul & Rostami, 2022) – on healthcare datasets with artificial neural networks showed varying performance for different re-sampling techniques, including varying performance on different datasets by ANN. The consequence of misclassification of abnormal example as normal is considered higher than the reverse as indicated by (Chawla et al., 2002) which concludes that the combination of over-sampling and under-sampling improves receiver operating characteristic curve score. The occurrence of dataset class imbalance is seen by machine learning models as deterrence to achieving satisfactory performance results. Effectiveness of applying these methods based on Deep neural and Convolutional neural networks as investigated (Joloudari et al., 2023) showed that using mixed Synthetic Minority Oversampling Technique (SMOTE) outperforms different methodologies to achieve 99.08% accuracy score. The application of SMOTE in sentiment analysis of hotel reservation service using classifiers such as Naïve Bayes (NB), Logistic Regression (LR), and Support Vector Machine (SVM) classification algorithms concluded that the use of SMOTE was effective in improving classification performance when dataset was imbalanced (Satriaji & Kusumaningrum, 2018). The performance of SMOTE-CD in compositional data operations using tested regressors such as Gradient Boosting tree, Neural Networks and Dirichlet regressor on real datasets assess predictive performance in areas such as prediction accuracy, cross-entropy, F1- score, R2 score and Root Mean Squared Error demonstrated result improvements for all metrics. However, the impact of applied over-sampling on model performance varied depending on model and the data used. Oversampling in some instances led to a decrease in performance for the majority class. This is in sharp contrast to best performance achieved when real data is used (Nguyen et al., 2023). Oversampling the minority class examples near the borderline with two SMOTE methods (borderline SMOTE1 and borderline SMOTE2) resulted in improvements in true positive values and F-values better than SMOTE and over-sampling (de Carvalho & Prati, 2020). Applied extreme gradient boosting technique with SMOTE on bond issuers to determine bond defaulters showed its effectiveness in applied imbalanced dataset situations (Zhang & Chen, 2021).

2. Methods

To address the challenge of minority class use in predictive modeling, we adopt two approaches. One is to perform predictive modeling using healthcare context-based datasets with extreme class imbalance without sampling techniques to evaluate model performance on prediction accuracy score, model generalization and model behavior regarding training examples with five (5) classification models. The second approach is to use the same dataset with class imbalance with two sampling techniques (over-sampling the minority class and under-sampling the majority class for comparative evaluation analysis to determine the impact of addressing dataset class imbalance. For this purpose, we adopt a real-world electronic healthcare dataset from a district hospital in Ghana noted for its long-standing involvement in managing chronic diseases. Identifiable features, such as patient names, were blocked from the records for privacy protection. Records obtained showed biological data, clinical notes, patient visits and performance metrics for prescriber evaluation. Records with no relevancy to this research were excluded in the collection process, and some of these were body temperature readings, clinical notes on eye treatments, dental notes, obstetric notes etc. Gold standards used for feature labeling include patient attendance records and clinical note descriptions without identifiable patient features to address patient privacy concerns. We applied five classification-based machine learning techniques: Gradient boosting, extreme gradient boosting, Logistic regression, Support vector machines and Random forest classifiers on three random sampling techniques for all predictive modeling.

Ethical approval and Consent: Real-world electronic healthcare record dataset obtained with approval notice referenced

DCS/S.1/VOL.1 on 30th march, 2022 from the management of Kwahu Government hospital in Ghana.

2.1. *Exploratory dataset analysis*

Dataset exploration for age, gender and target class patterns is displayed in Fig 1 which has three graphs. First graph is a display of age distribution from the sampled dataset and can be described as normal or with a Gaussian distribution. The second graph is the sampled gender distribution which shows the distribution of female and male patients. This graph shows a higher number of female patients than males in the sampled dataset. The last graph (third graph) shows the distribution of the target or output class (patients described as defaulters of treatment default and those who are described as non-defaulters).The distribution of female numbers in the sampled population was 4,312, constituting 80.86%, and 1,021 males made up 19.14% of the entire sampled population, bringing the total sample population to 5,333. Fig 1 third graph is the presentation of target class distribution. Output class distribution shown in Fig 1 (third graph from left) indicates unequal class distribution between defaulting and non-defaulting patients which is displayed in Fig 2 visualization. Probability of patient default by age is also presented in Fig 30.

2.2. *Data preprocessing stages*

Predictive data modeling involves data preprocessing steps as shown in Fig 3. Some of these steps include problem definition necessary to identify data collection types. Preprocessing involves other sub-processes such as removal of redundant and duplicate data, addressing dataset variable outliers, addressing null values (removal of null value cells), data scaling and formatting, addressing dataset dimensionality issues (dimensionality reduction or addition), label encoding etc. These are important preprocessing steps necessary to ensure smooth predictive modeling.

2.3. *Supervised learning types*

Fig 4 shows supervised model selection types for each problem definition with learning application choices for each problem domain. Supervised learning has two main branches of use, classification and regression. Each selection choice has predefined model selections applicable in each problem domain.

2.4. *Simplified model building process*

Simplified model process display in Fig 5 shows how data modeling occurs. It has an input process in the form of data given to a machine learning (supervised) algorithm with computational functions to train for predictions as an output label. One of the key advantages of Supervised modeling use is the prediction of an output based on prior experience for which knowledge of dataset class use is required.

3. Results

The presentation of results is in two parts. The first part deals with a display of results without random sampling techniques and covers performance metrics such as receiver operating characteristics and learning curves with cross-validation, true positive rates, true negative rates, false positives, false negatives, positive predicted values, negative predicted values and the second part deals with the display of results obtained from applied random sampling techniques.

3.1. *Evaluation metrics*

Results obtained from pre-sampling evaluation and those obtained after performing over-sampling, under-sampling and applied synthetic minority oversampling technique is presented in tables 3.0, 3.1, 3.2 and 3.3. Evaluated metrics for each model choice include base prediction accuracy score, percentage share of positive results truly positive (TPR), true negative rates (TNR), positive predicted values (PPV), negative predicted values (NPV), false alarm metrics (False positive rates- FPR and False negative rates- FNR), average prediction accuracy score (balanced accuracy) and area under the receiver operating characteristic curve score (auc_roc score). Graphical visualization of roc_auc score values obtained by each modeling technique using pre-sampling or base model is displayed in Fig 6, applied over-sampling results in Fig 7, under-sampling in Fig 8 and applied SMOTE over-sampling in Fig 24. Additionally, learning curves useful to determine model performance behavior against training examples over a period for each random technique applied is recorded as follows; over-sampling, under-sampling, pre-sampling and SMOTE oversampling (extreme gradientboosting classifier Fig 9, Fig 14, Fig 19 and Fig 25, logistic regression Fig 10, Fig 15, Fig 20 and Fig 28, random forest Fig 11, Fig 16, Fig 21 and Fig 27, gradientboosting classifier Fig 12, Fig 17, Fig 22 and Fig 29 and support vector machine Fig 13, Fig 18, Fig 23 and Fig 26) respectively. Each of these graphs show model performance prediction accuracy behavior on training example over a period. It also represents how well these predictive models would perform on unseen datasets and the extent to which these models can converge and be generalized as validation criteria to determine performance behavior over training examples.

T**able 3.** Sensitivity and Specificity evaluation results (pre-sampling)

Model	FNR(%)	TNR(%)	FPR(%)	PPV(%)	NPV(%)	TPR(%)	AUC SCORE(%)	PREDICTION ACCURACY(%)	BALANCED ACCURACY(%)
Gradient boosting	0.13	9.09	90.91	98.75	50	99.87	97.00	99.00	55.00
XGBC	0.51	13.64	86.36	98.80	27.27	99.49	99.00	98.00	57.00
Logistic regression	0.06	13.64	86.36	98.81	75	99.94	94.00	99.00	57.00
Random forest	0.57	9.09	90.91	98.74	20.16	99.25	97.00	98.00	54.00
SupportVector machine	0.00	0.00	100	98.62	66.67	100	89.00	99.00	57.00

Table 3.1. Sensitivity and Specificity results (Over-sampling)

Model	FNR(%)	TNR(%)	FPR(%)	PPV(%)	NPV(%)	TPR(%)	AUC SCORE(%)	PREDICTION ACCURACY(%)	BALANCED ACCURACY(%)
Gradient boosting	20.84	98.91	1.09	98.67	82.19	79.16	95.00	88.90	89.03
XGBC	2.38	100	0.00	100	97.61	97.62	100.00	98.79	98.81
Logistic regression	17.46	94.59	5.41	94.01	84.04	82.54	93.00	87.84	88.57
Random forest	1.81	100	0.00	100	98.17	98.19	100.00	99.08	99.09
SupportVector machine	20.53	91.63	8.37	90.71	81.28	79.47	91.00	85.47	85.55

Table 3.2. Sensitivity and Specificity evaluation results (Under-sampling)

Model	FNR(%)	TNR(%)	FPR(%)	PPV(%)	NPV(%)	TPR(%)	AUC SCORE(%)	PREDICTION ACCURACY(%)	BALANCED ACCURACY(%)
Gradient boosting	11.25	98.26	1.74	98.00	90.06	88.75	98.00	93.59	93.5
XGBC	2.13	100.00	0.00	100.00	97.99	97.87	100.00	98.97	98.93
Logistic regression	16.61	94.58	5.42	93.68	85.52	83.39	93.00	89.09	88.98
Random forest	1.29	100	0.00	100.00	98.77	98.71	100.00	99.37	99.35
SupportVector machine	18.55	91.15	8.85	89.87	83.6	81.45	92.00	86.39	86.3

Table 3.3. Sensitivity and Specificity evaluation results (SMOTE- Synthetic minority oversampling technique)

Model	FNR(%)	TNR(%)	FPR(%)	PPV(%)	NPV(%)	TPR(%)	AUC SCORE(%)	PREDICTION ACCURACY(%)	BALANCED ACCURACY(%)
Gradient boosting	11.61	97.69	2.31	97.51	89.16	88.39	97.00	93.37	93.04
XGBC	4.96	98.84	1.16	98.83	95.12	95.04	99.00	96.92	96.94
Logistic regression	14.68	92.43	7.57	92.02	86.02	85.32	94.00	89.85	88.87
Random forest	4.14	98.52	1.48	98.52	95.88	95.86	99.00	97.46	97.19
SupportVector machine	22.9	91.46	8.54	90.23	79.61	77.1	92.00	85.98	84.28

Discussions

In this section, research results obtained and its implications are presented. In Fig 1 three graphs represent the dataset distribution of patient age, gender and target class. These are represented as first graph- age distribution, the second graph- distribution by gender and the third graph is represented by target or output distribution. The mean age is approximated 64 years, the minimum age is 18 years, and the maximum age is 111. The distribution of age, as shown in Fig 1 (first graph) follows a normal Gaussian distribution. Effect of age on treatment default outcome as shown in Fig 30 indicate a high concentration of default patients between the ages of 20-40 years and above 80 years. The distribution of gender, as shown in the second graph, indicate an unequal distribution between female and male. Output class distribution, as indicated (third graph) show imbalance output class distribution which is confirmed in Fig 2 visualization plot. Evaluation metrics and predictive performance results presented in Table 3.0, 3.1, 3.2 and 3.3 show model performance on truly positive, truly negative and other misclassifications; positive predictions wrongly predicted as negative and negative predictions predicted as positive that raises false alarm (false positives, false negatives) under all optimization techniques. Determining the impact and potential benefits of applied optimization techniques on imbalance class distribution dataset in this scenario, a comprehensive assessment of performance using pre-sampling, over-sampling, under-sampling and SMOTE oversampling strategies was adopted. Table 3.0 showcases results obtained before applying random sampling techniques (pre-sampling) with poor performance on the classification of both true negative and false positive examples for all models used. Ironically, recorded high prediction accuracy score and roc_auc score values and average balanced accuracy scores are indicated. Table 3.1 on over-sampling show improved model performance on prediction accuracy score, roc_auc score and balanced accuracy. It also shows superior model performance exhibited by extreme gradient boosting classifier and random forest classifier on FNR, TNR and FPR. Table 3.2 on under-sampling also indicate better performance on prediction accuracy, roc_auc score, balanced accuracy, TNR, FNR and FPR for which extreme gradient boosting and random forest classifiers show a much improved FNR scores reducing from 2.38 in over-sampling for extreme gradient boosting to 2.13 in under-sampling and from 1.81 in over-sampling for random forest to 1.29 in under-sampling. Using SMOTE oversampling approach (Table 3.3) as compared to over-sampling results obtained in table 3.1, SMOTE oversampling showed similar scores much lower than what was obtained in under-sampling as shown in table 3.2

Evaluation of performance with area under the curve (auc_roc) for pre-sampling (Table 3.0), over-sampling (Table 3.1), under-sampling (Table 3.2) and SMOTE oversampling show relatively high performance in prediction accuracy, auc_roc score and balanced accuracy score for over-sampling, under-sampling and SMOTE oversampling. High auc_roc score and prediction accuracy score in pre-sampling show limited impact on overall model performance even with recorded poor performance in TNR, NPV and FPR. This therefore confirms that relying on prediction accuracy or roc_auc score as justification for model performance in real-world applications with skewed class distribution may not be beneficial and could lead to undesirable outcome for critical decision making in specialized areas such as healthcare. However, as recorded in Table 3.1, 3.2 and 3.3, better and improved performance for all models in areas such as FNR, TNR, FPR and improved balanced accuracy score from over-sampling, under-sampling and SMOTE oversampling show statistically significant impact of its use.

Evaluating model performance behavior with Learning curves indicates that for SMOTE oversampling, all models except random forest show improvements in performance accuracy score and convergence with increasing training examples. In over-sampling, gradient boosting classifier show consistent improvements in performance with additional training examples over time as indicated in Fig 12. Continuous performance improvement by gradient boosting classifier is replicated in under-sampling as shown in Fig 17, creating the best possible convergence with additional training examples.

Similar performance behavior by all models is observed in over-sampling, under-sampling and SMOTE oversampling. Significant difference in performance exits in applied random sampling and pre-sampling. Overall performance of gradient boosting classifier shows continuous improvements over time with more training examples in over-sampling, under-sampling and SMOTE oversampling. It generalizes well on unseen datasets on cross-validation in applied random sampling techniques. It is significant to indicate that random forest classifier show extreme divergence with additional training examples in SMOTE oversampling as indicated in Fig 27. High prediction accuracy and roc_auc scores recorded in pre-sampling and in (over-sampling, under-sampling and SMOTE oversampling) show minimal impact of random sampling on model performance. Further assessment of patient age on treatment default as shown in Fig 30 indicate that treatment default is predominant among the ages 20 to 40 years and from 80 years above. The most compliant age group is identified to be between 41-60 years.

Conclusions

This comprehensive evaluation of model performance with applied random sampling (over-sampling, under-sampling and SMOTE oversampling) techniques, show that under-sampling the majority class in an imbalance dataset leads to improvements in model performance far more than over-sampling and SMOTE for gradient boosting classifier including better performance in reducing false alarms (FNR and FPR). This research has demonstrated how certain age grouping is associated with treatment default together with the identification of the most compliant age grouping. Model evaluation with learning curves demonstrates model convergence, generalization and performance on unseen datasets, effect of performance with increasing training examples and how this can be applied in settings with repetitive tasks.

Research application and Recommendation

Performance evaluation metrics used can be replicated in similar settings with class imbalance. Practical business applications can benefit from the comprehensive assessment of model performance to determine model of choice in similar scenarios. Learning curves can be used to assess resource utilization and to justify the need for less or additional resource in a given task. Given the high prediction accuracy score and roc_auc score recorded in both pre-sampling and applied sampling scenarios, future model evaluation for performance can benefit from study to determine its appropriateness.

Figures

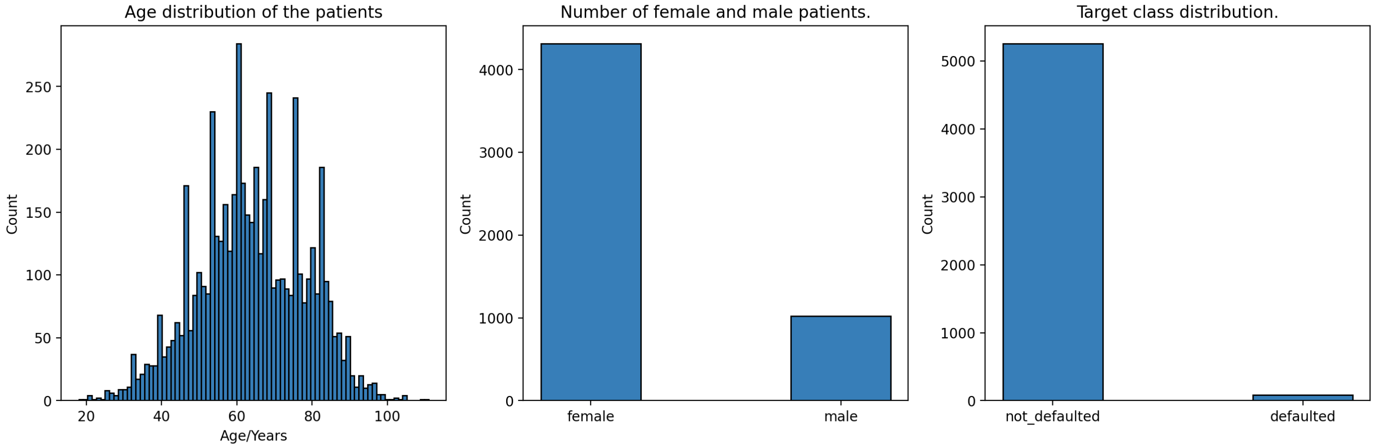

Fig 1. Class distributions. Caption: age, gender and target class distribution graphs. Age distribution among sampled gender together with target class distribution is illustrated by the individual graphs showed in Fig 1.

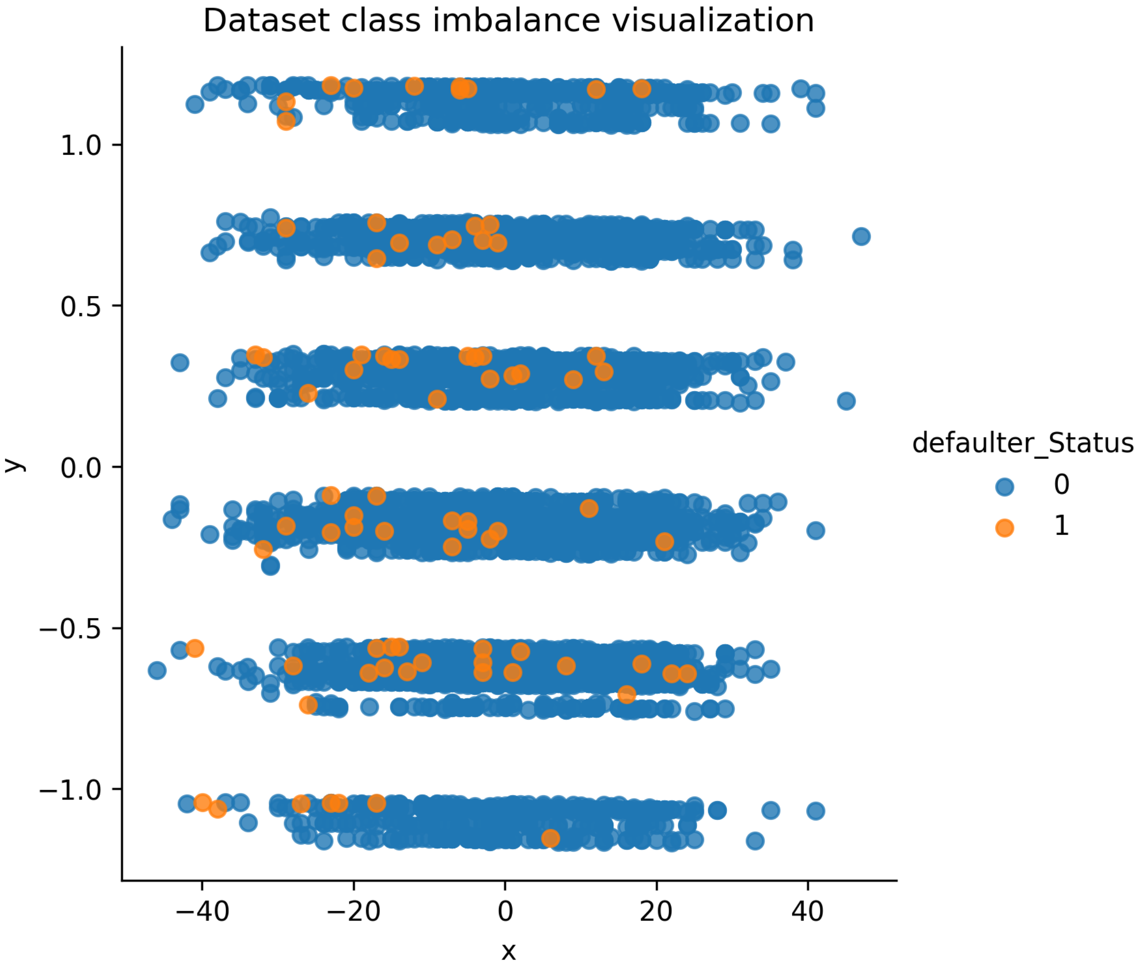

Fig 2. Dataset class imbalancevisualization. Caption: Target class distribution visualization. This graph captures target class distribution in a visualization that best explains minority and majority class examples.

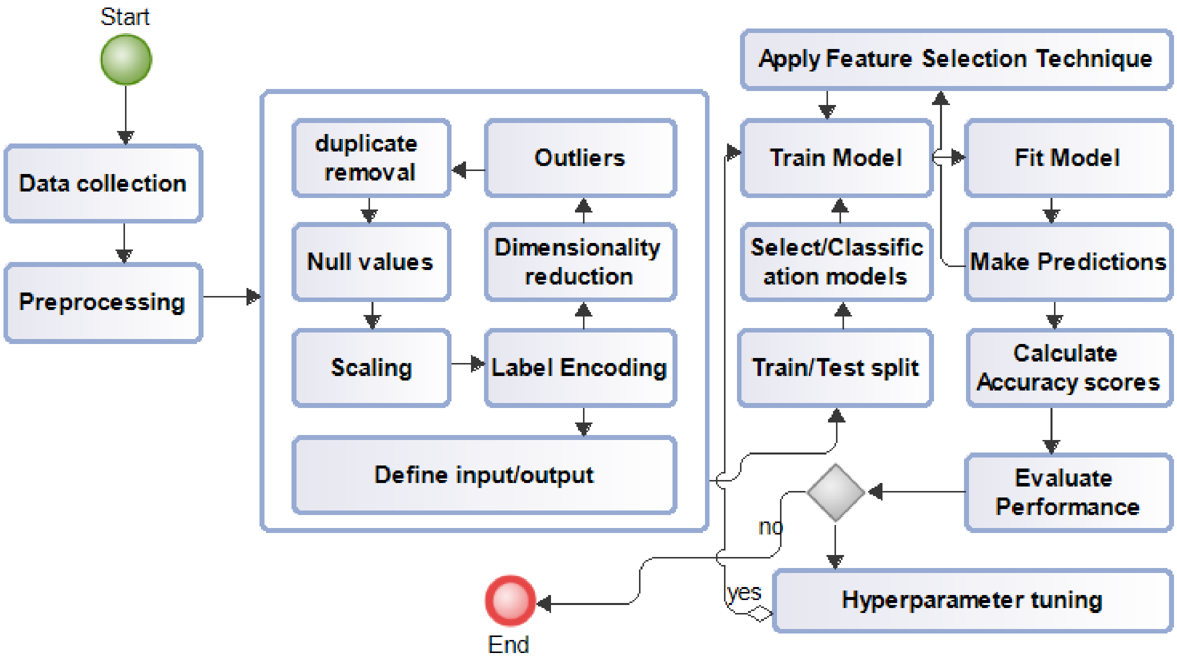

Fig 3. Data modelingflowchart. Caption: Data pre-processing and machine learning modeling stages to describe how data cleaning is achieved.

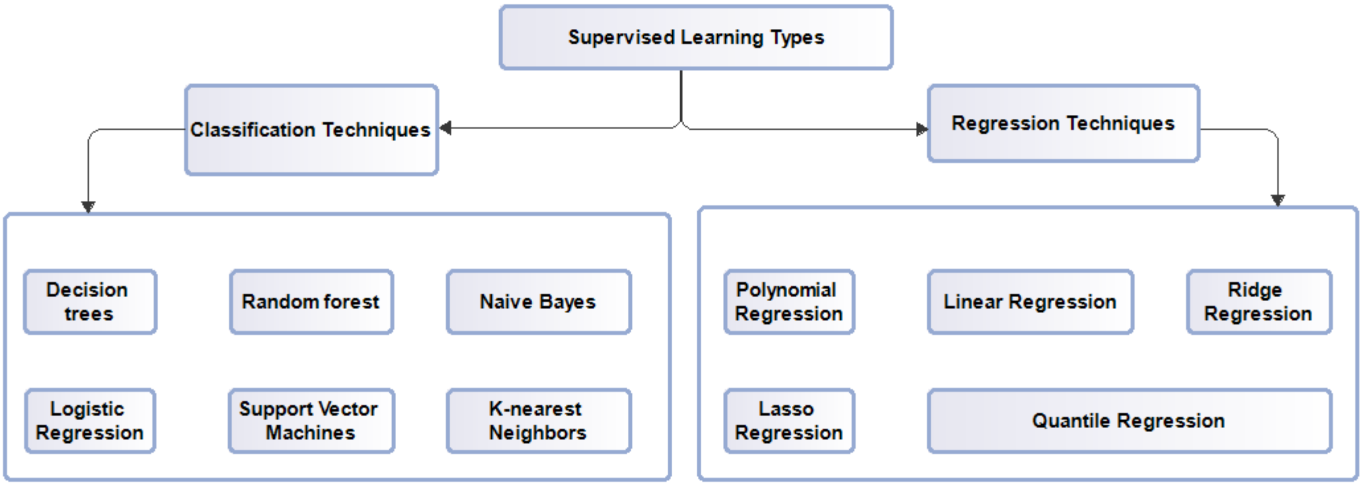

Fig 4. Supervised learning types : Caption. Supervised model selection categories listing classification and regression technique types. Each supervised category shows applied techniques listed under it.

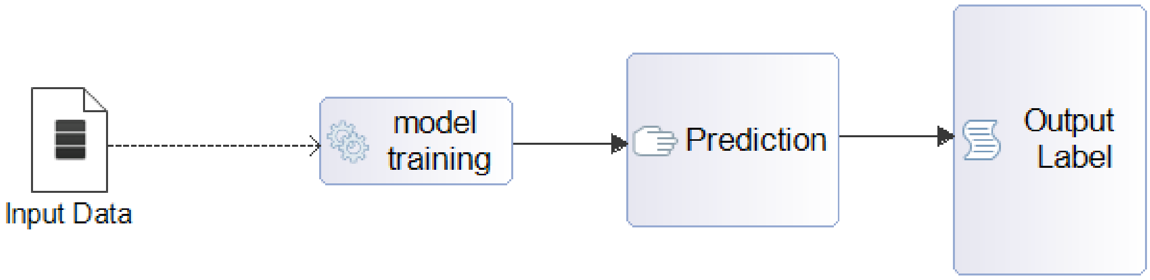

Fig 5. Modeling process : Caption. Simplified modeling process to indicate predictive modeling stages to show input,

processing and output formations is shown.

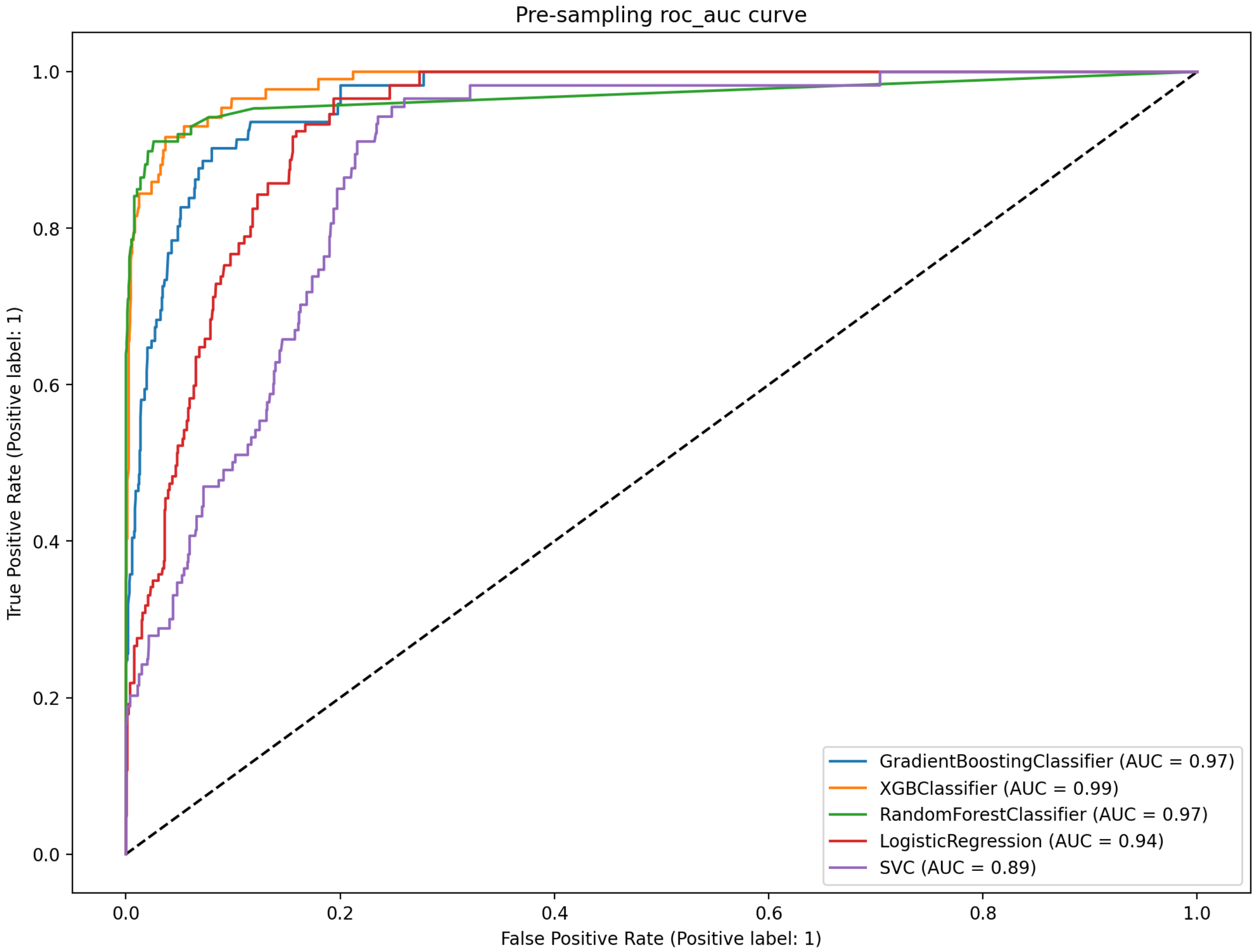

Fig 6. Pre-sampling receiver operating curve: Caption. Initial model performance before sampling was applied. Roc_auc scores for the individual models is indicated in this graph.

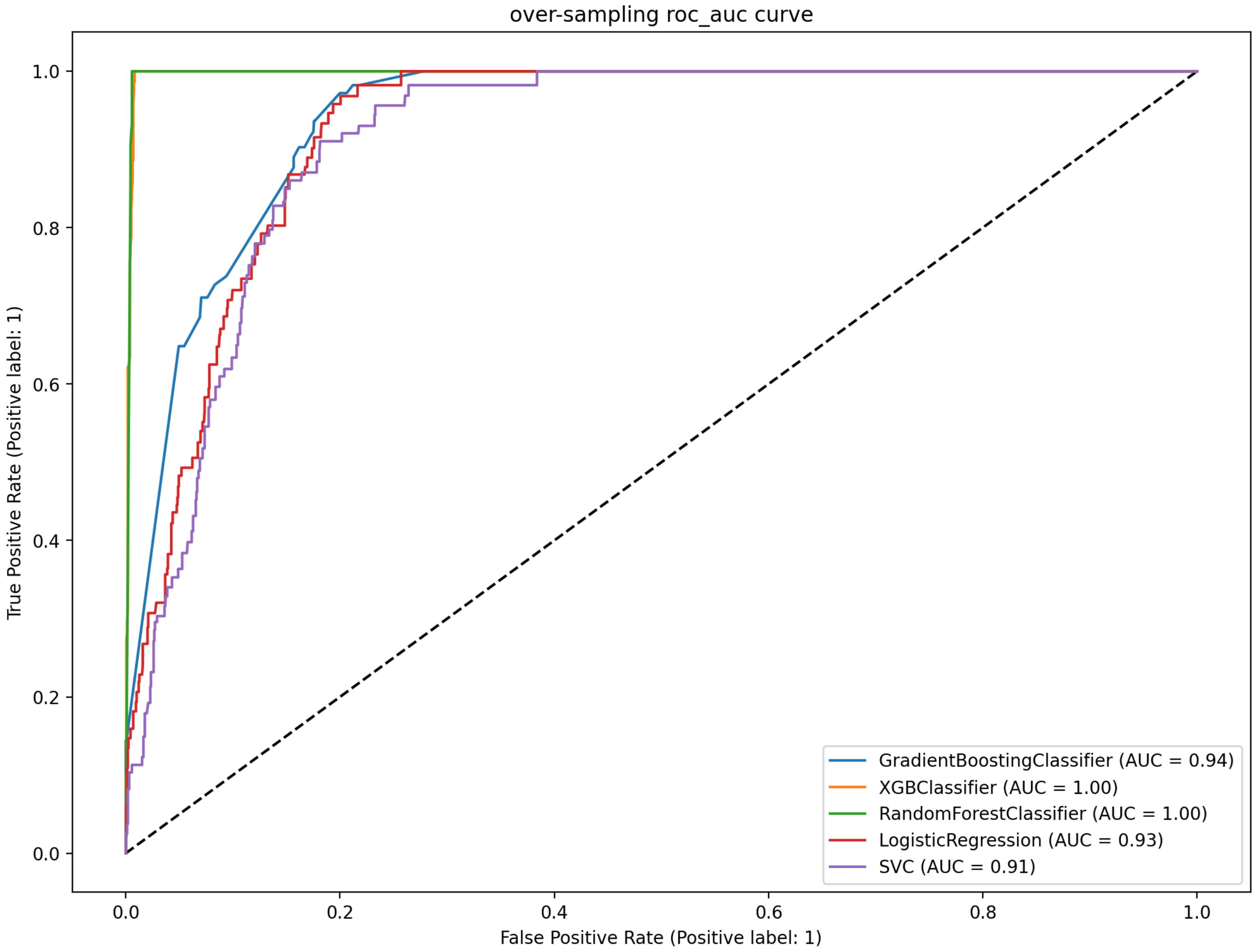

Fig 7. Over-sampling receiver operating curve : Caption. Roc_auc score obtained by each model in applied over-sampling.

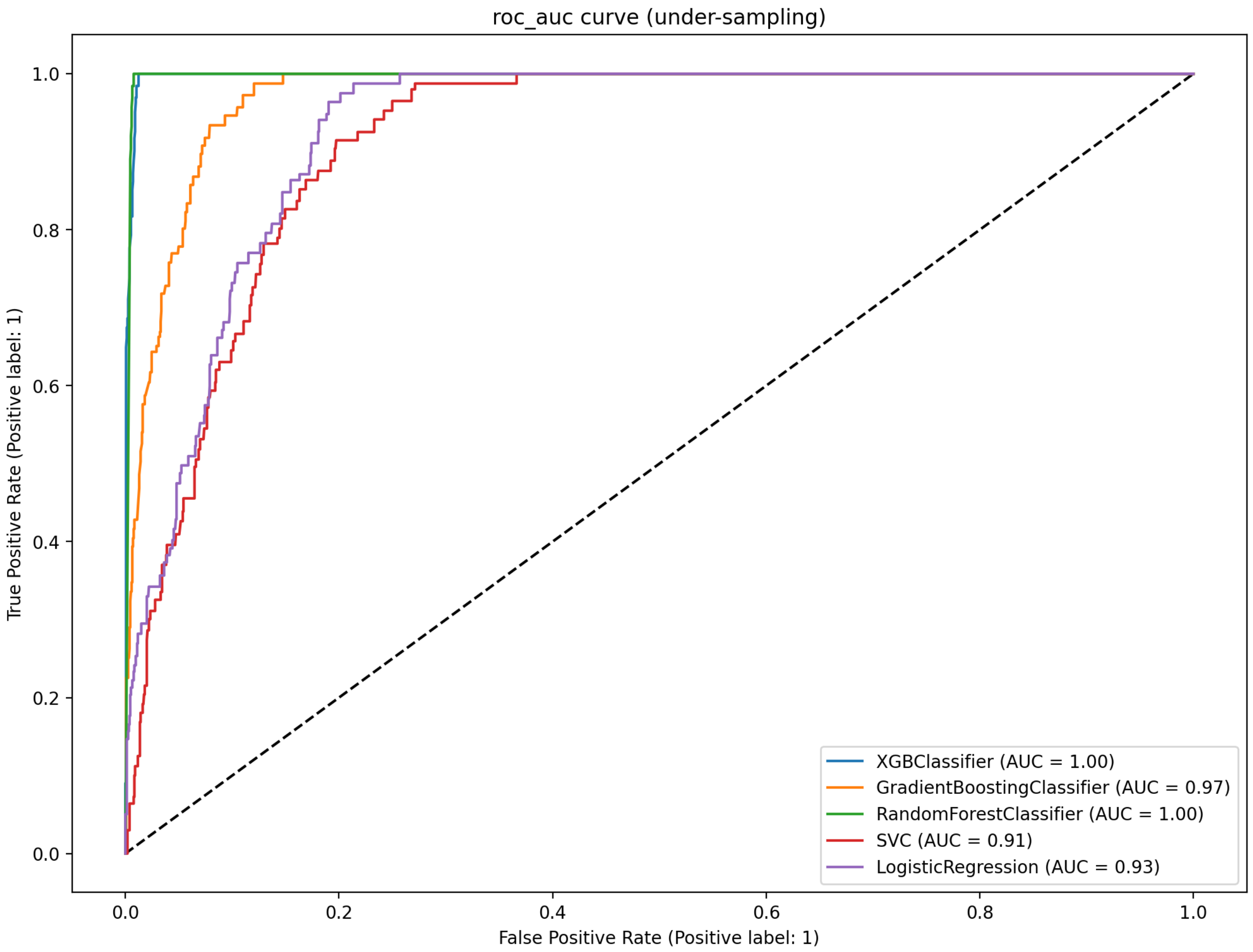

Fig 8. Under-sampling receiver operating characteristic curve : Caption. Applied under-sampling model performance curve.

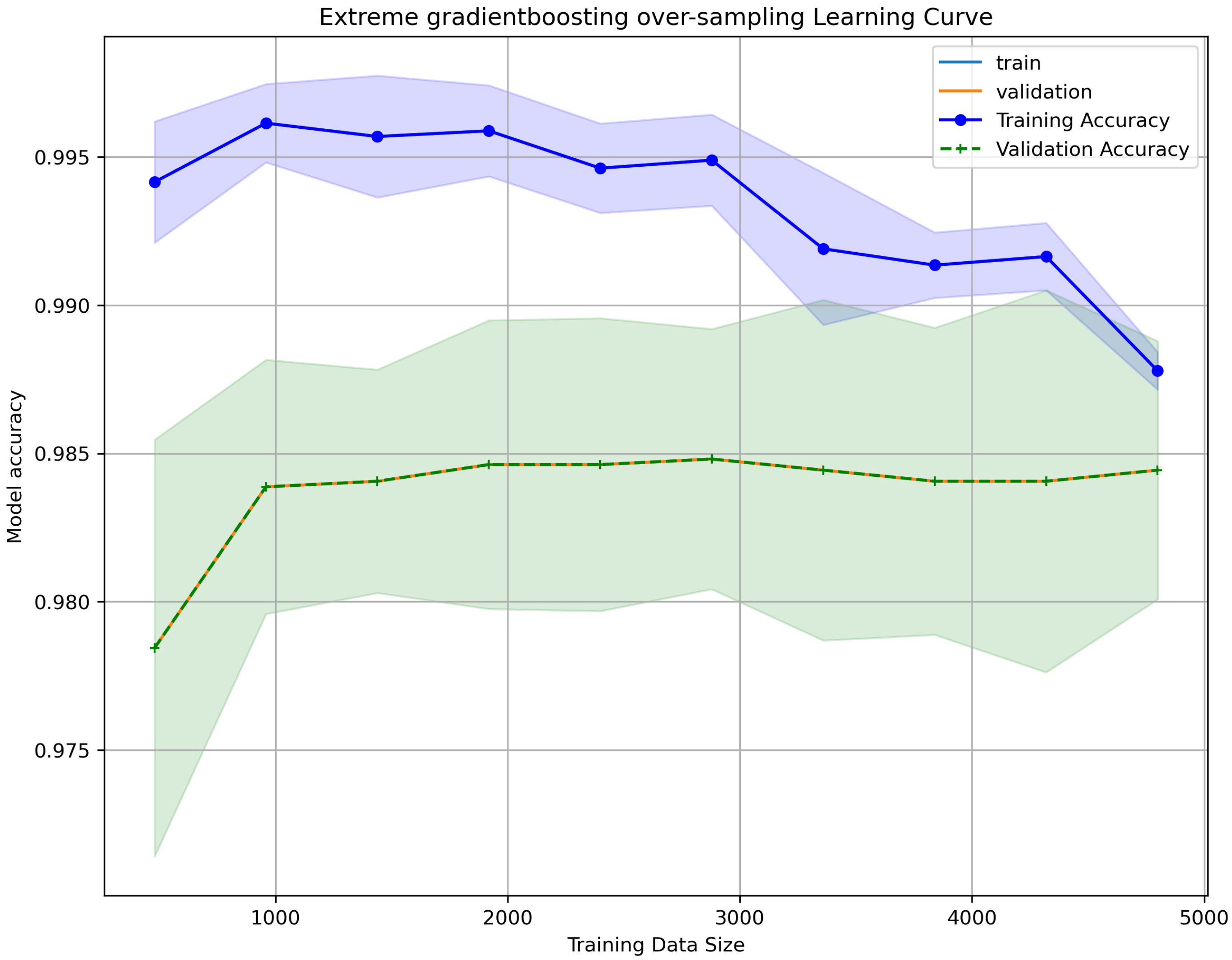

Fig 9. Learning curve graph : Caption. Extreme grading boosting performance in over-sampling to estimate model generalization and prediction accuracy score on unseen dataset.

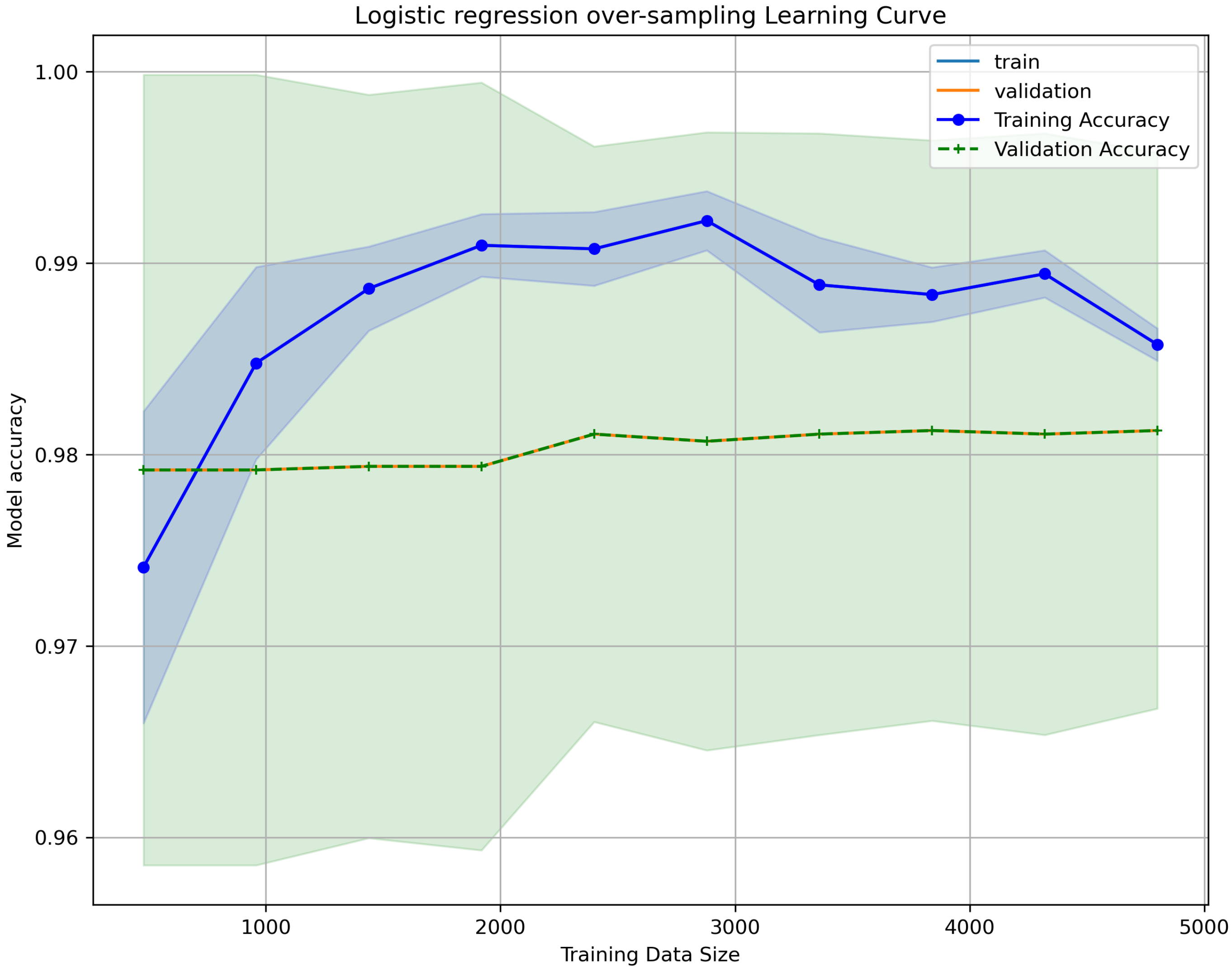

Fig 10. Learning curve graph: Caption. Logistic regression performance in over-sampling

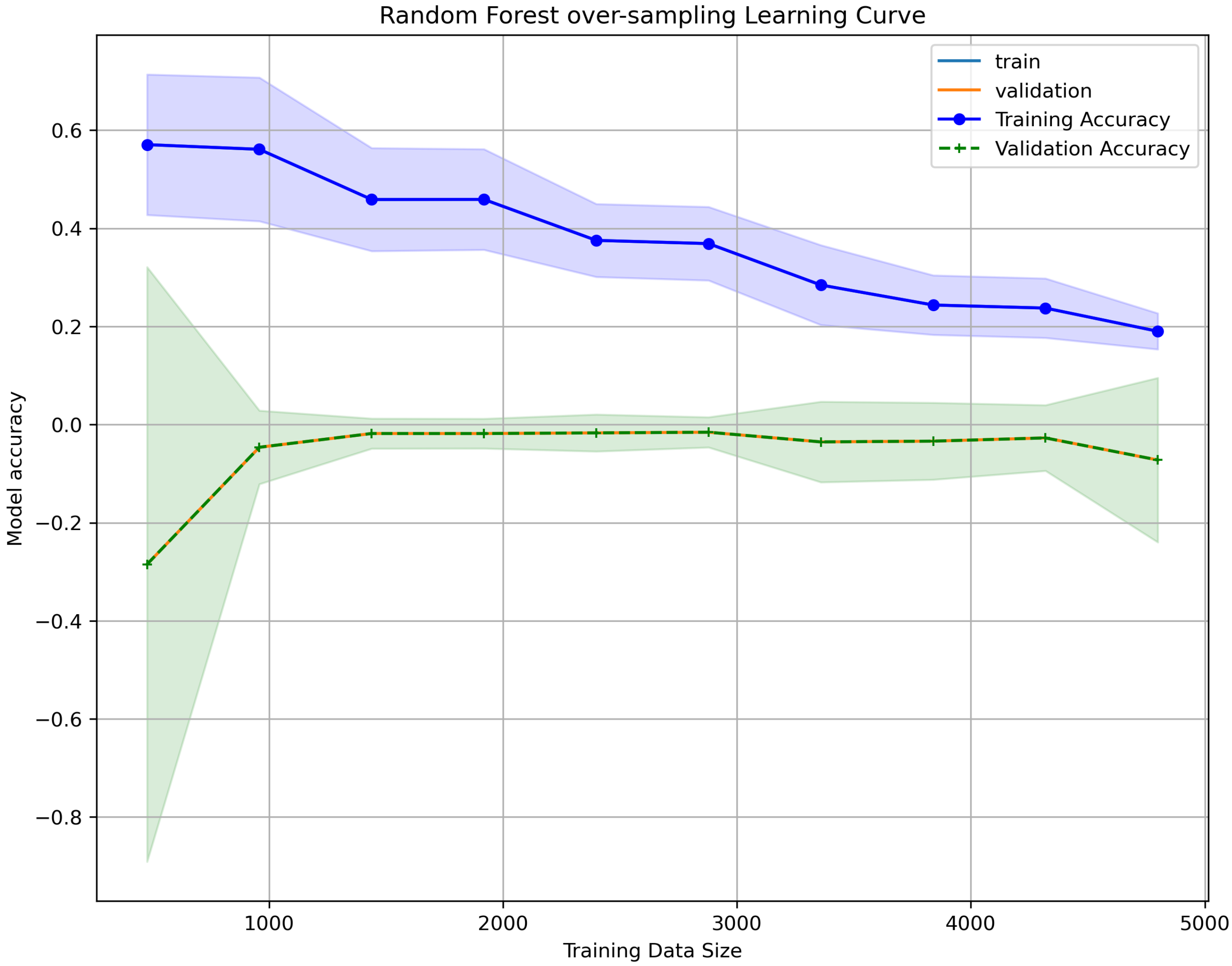

Fig 11. Learning curve graph: Caption. Random forest classifier performance in over-sampling

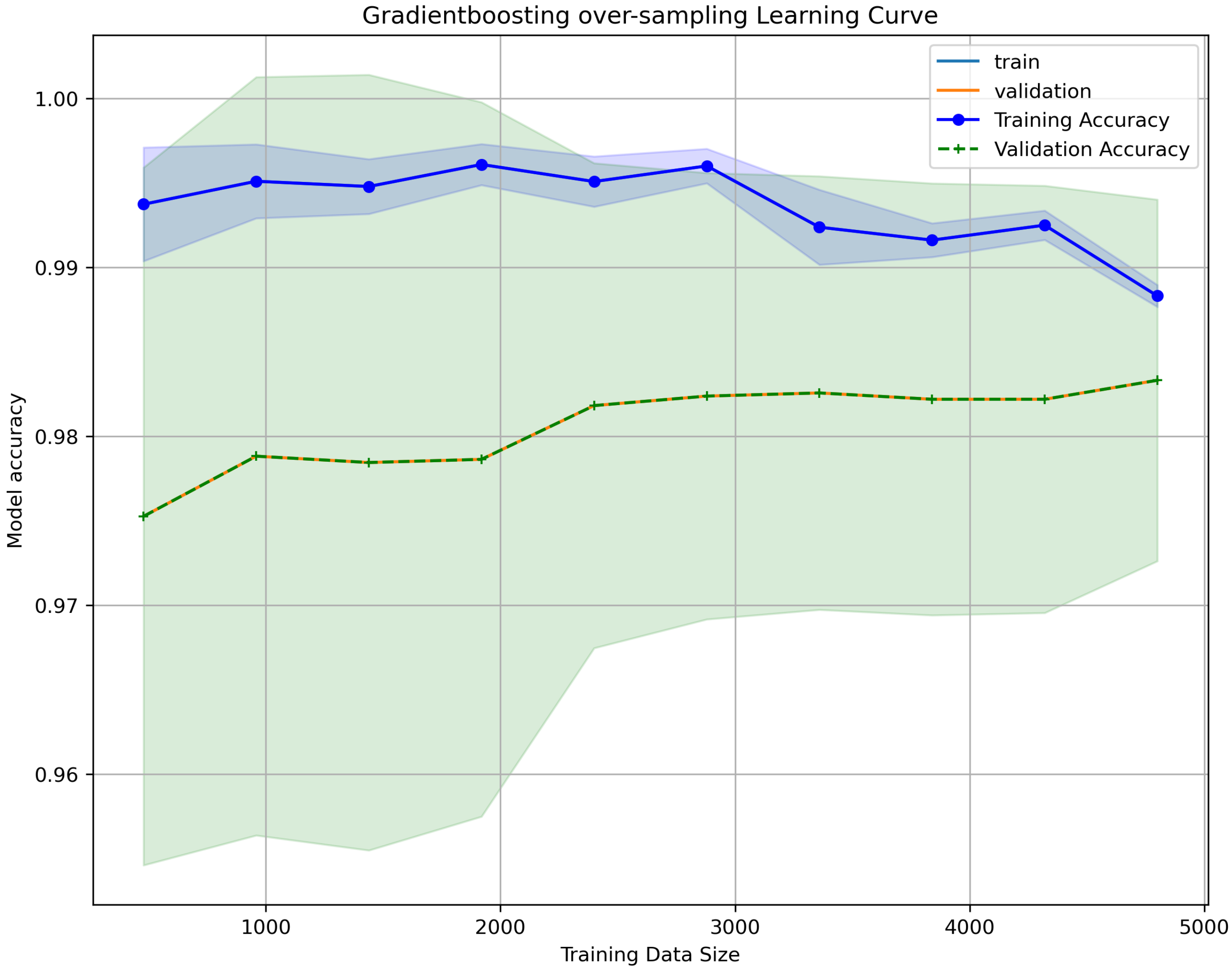

Fig 12. Learning curve graph. Caption. Gradientboosting classifier performance learning curve graph.

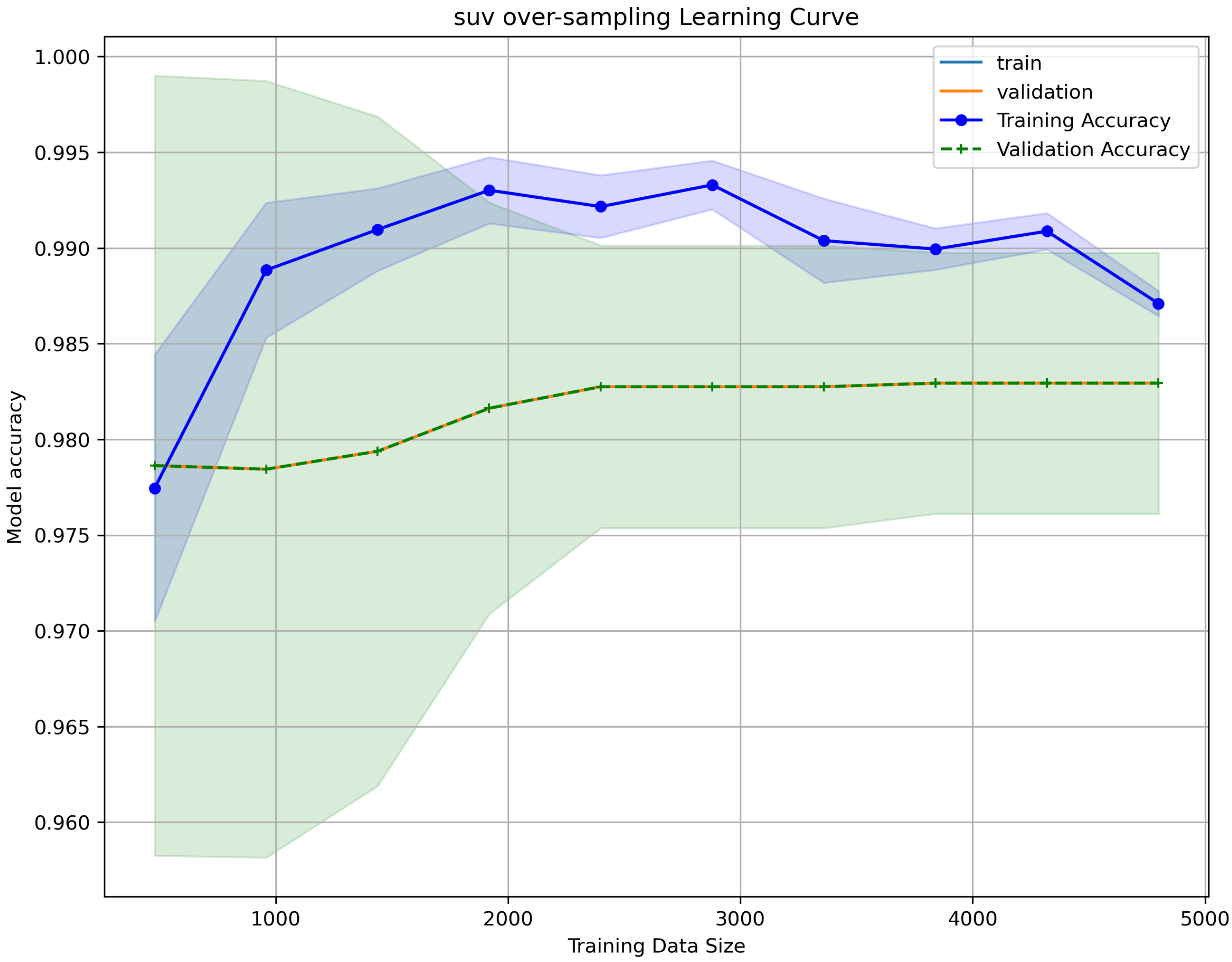

Fig 13. Learning curve graph. Caption. Support vector machine performance learning curve graph.

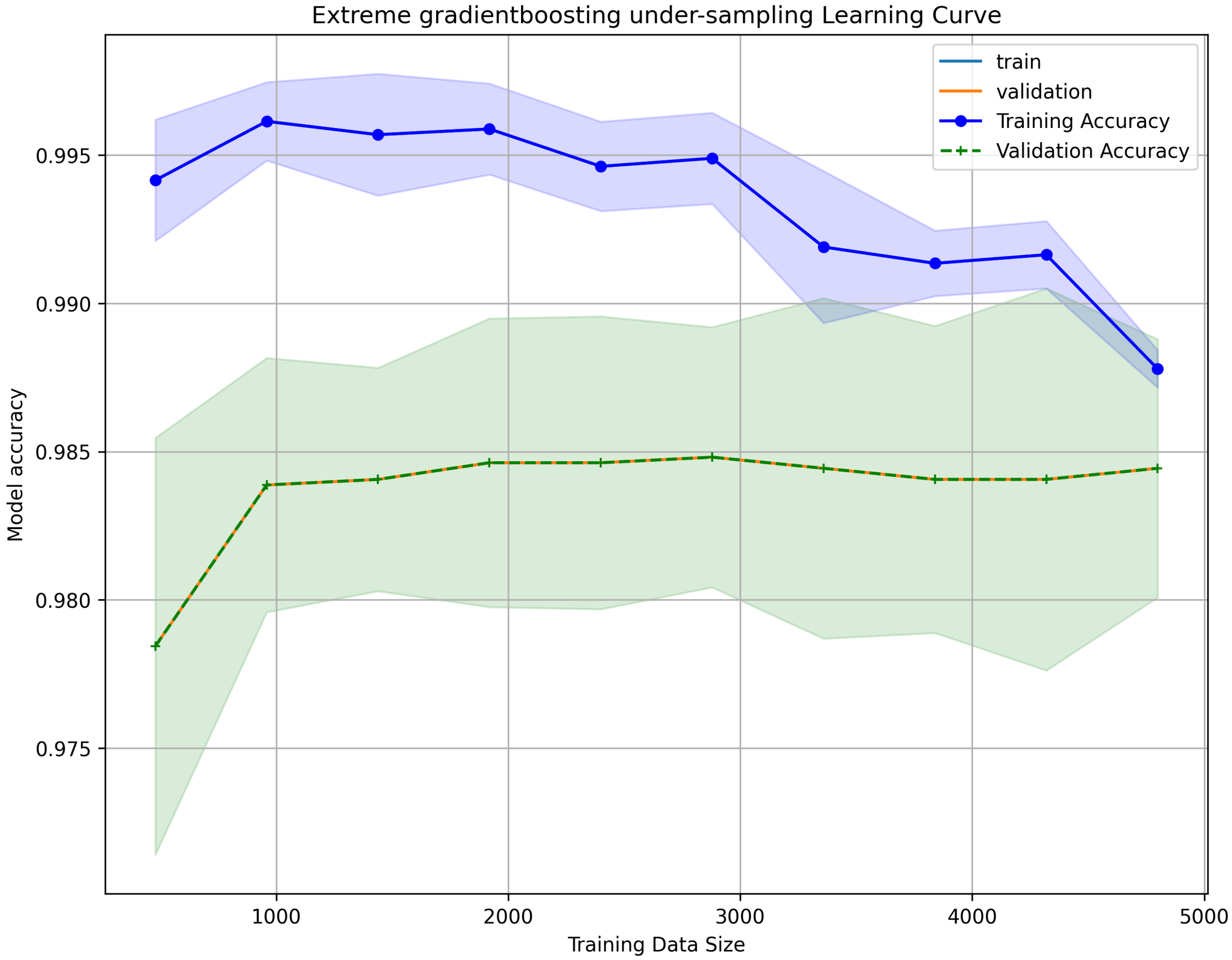

Fig 14. Learning curve graph: Caption. Extreme gradient boosting classifier under-sampling learning curve graph.

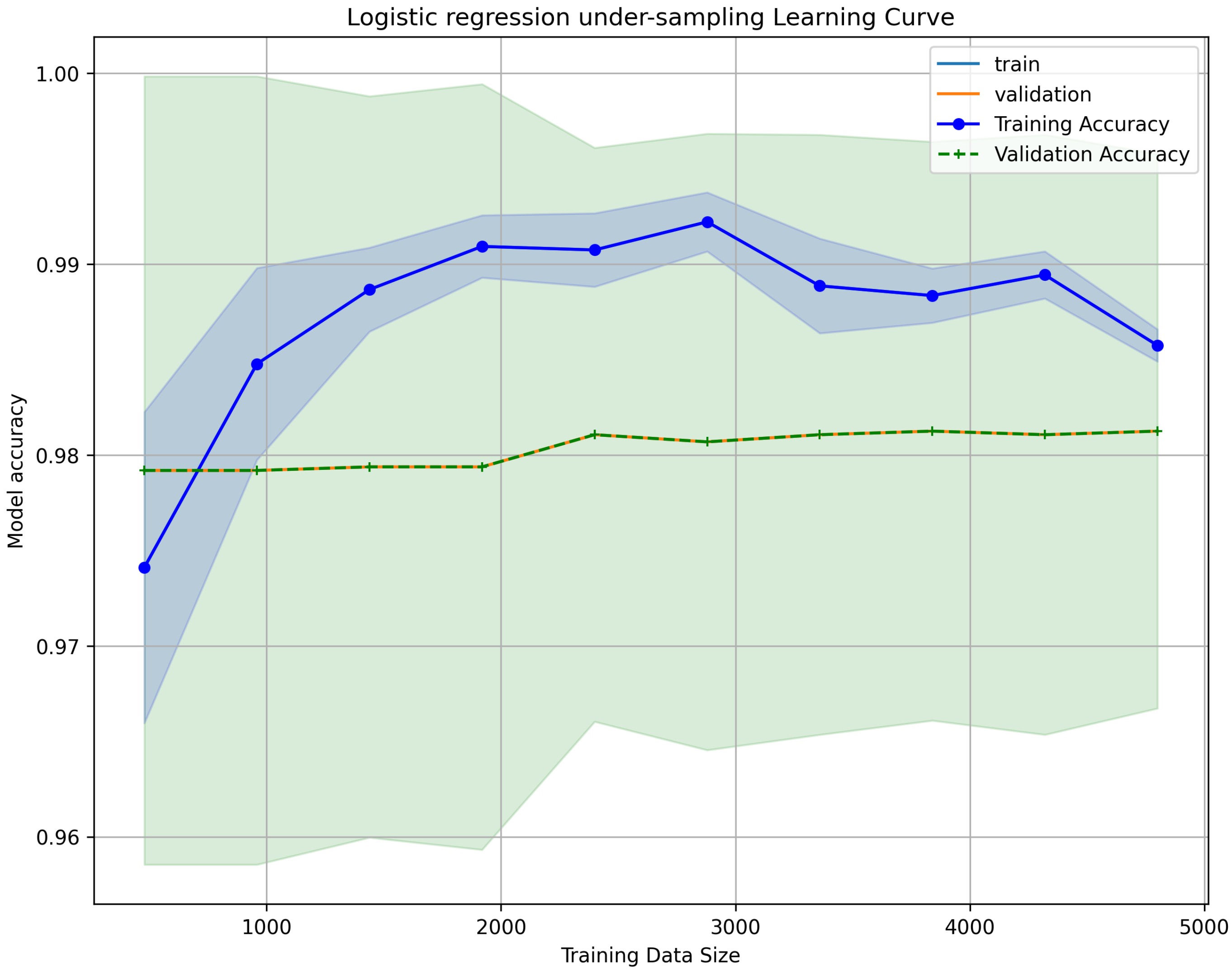

Fig 15. Learning curve graph: Caption. Logistic regression under-sampling learning curve graph.

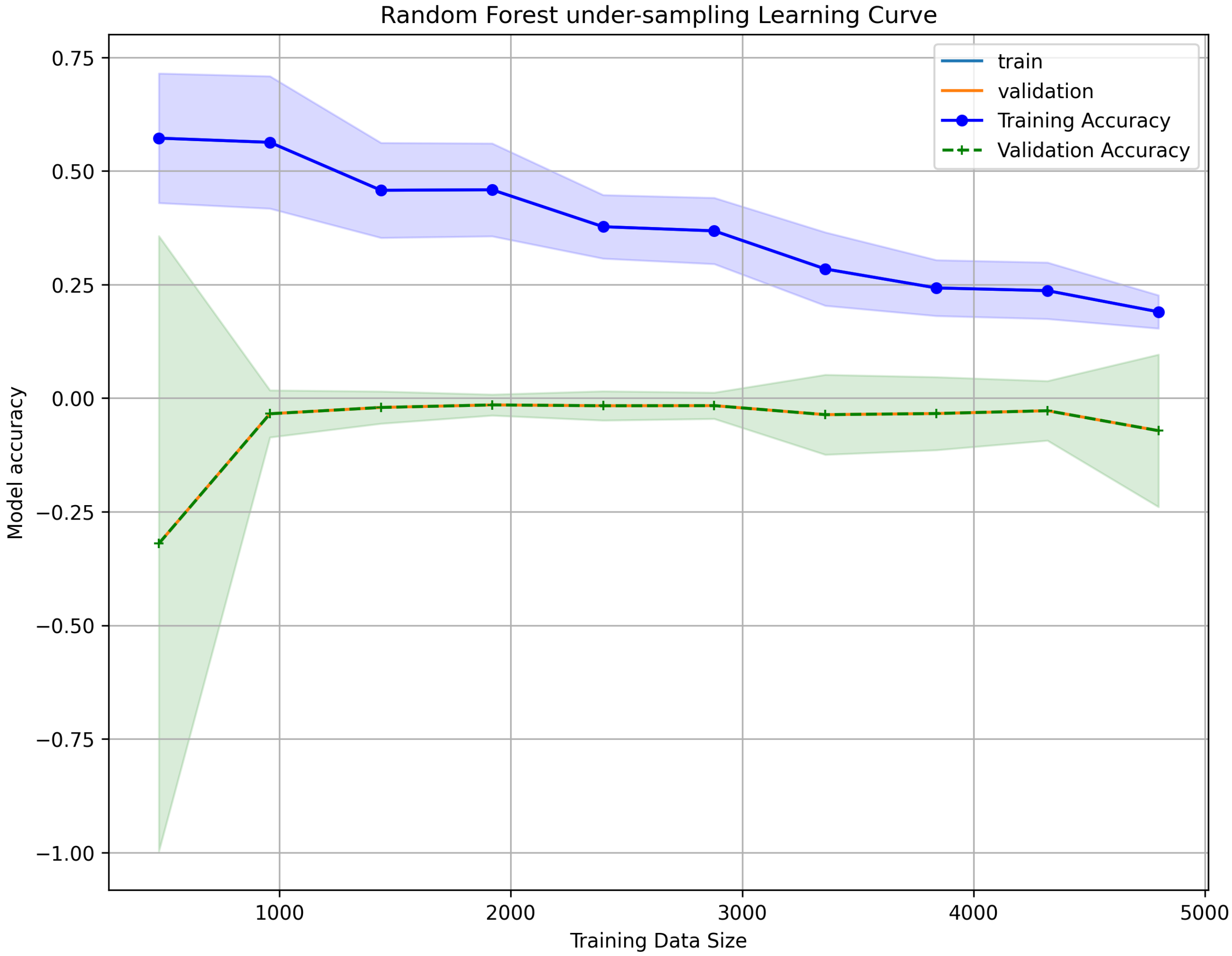

Fig 16. Learning curve graph: Caption. Random forest under-sampling learning curve graph.

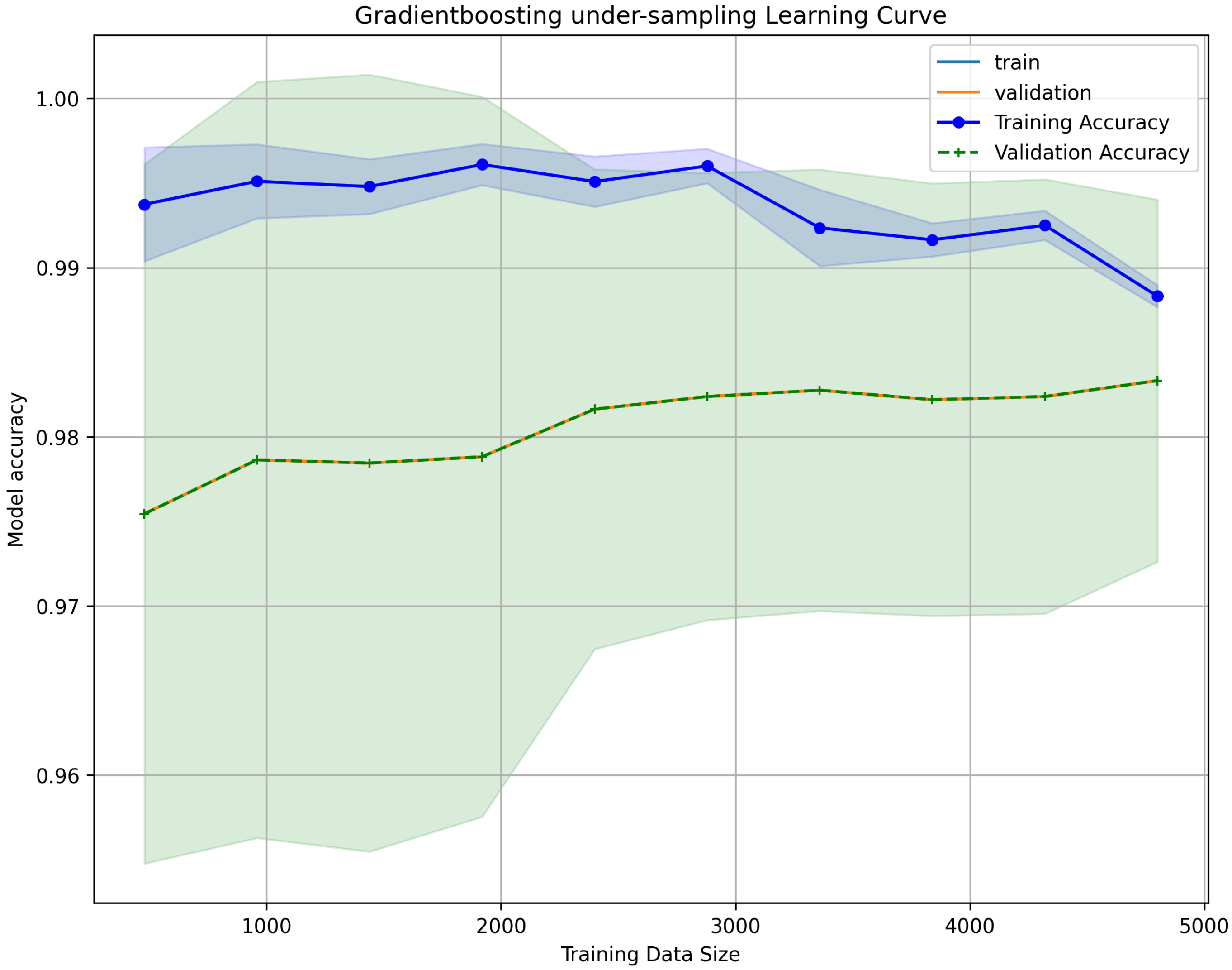

Fig 17. Learning curve graph: Caption. Gradient boosting classifier under-sampling learning curve graph.

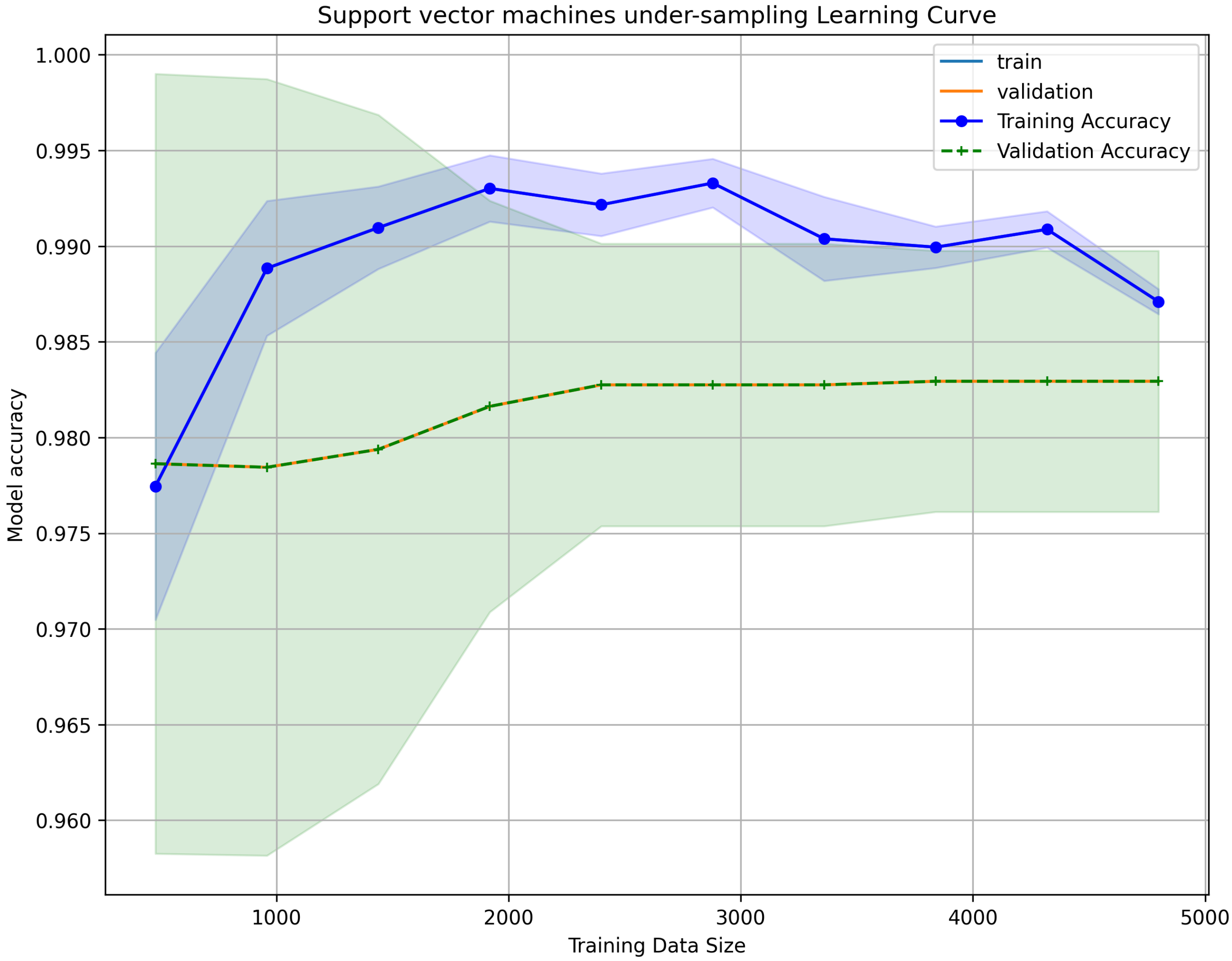

Fig 18. Learning curve graph: Caption. Support vector machine under-sampling learning curve

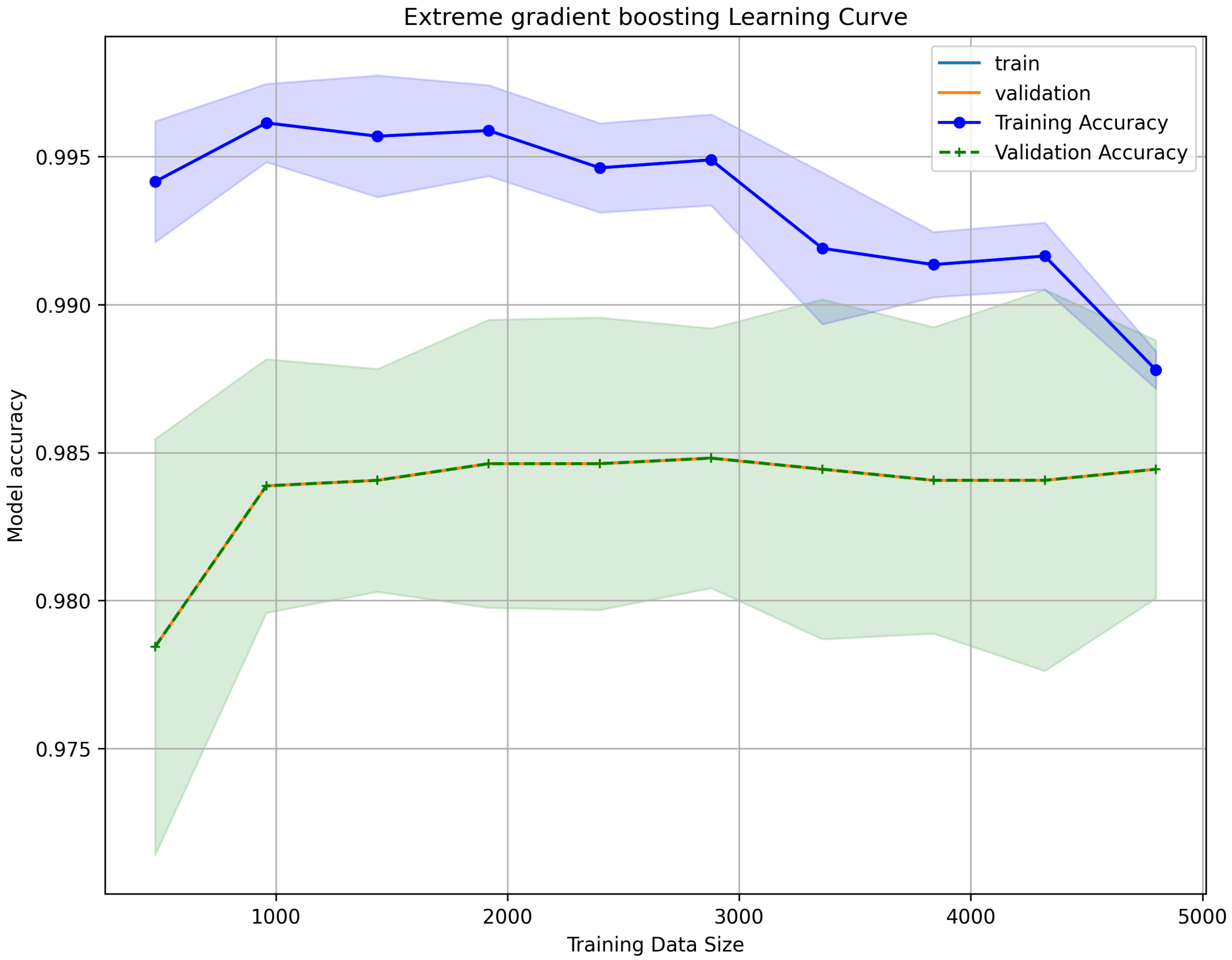

Fig 19. Learning curve graph. Caption. Extreme gradient boosting pre-sampling learning curve graph.

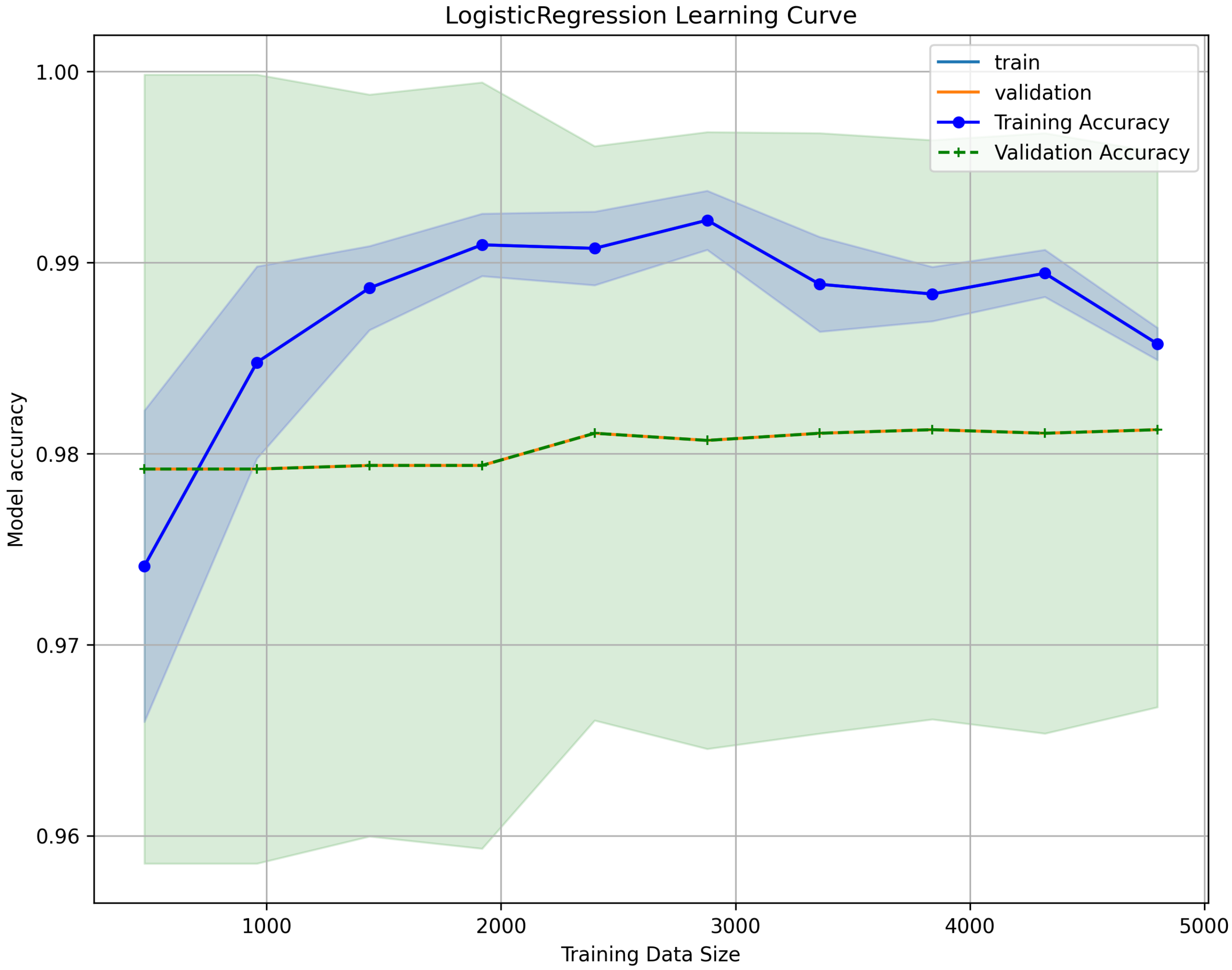

Fig 20. Learning curve graph: Caption. Logistic regression pre-sampling learning curve.

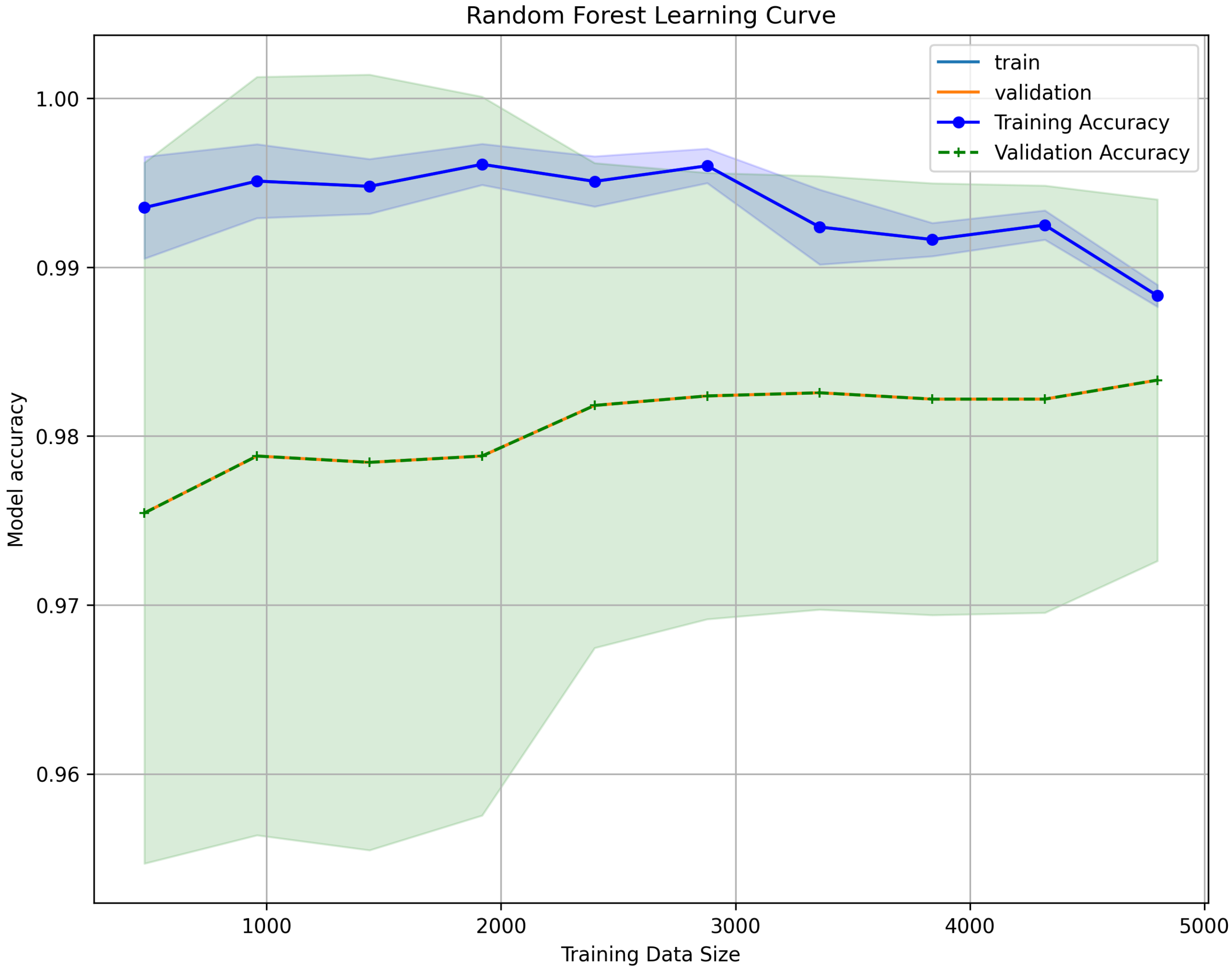

Fig 21. Learning curve graph: Caption. Random forest pre-sampling learning curve

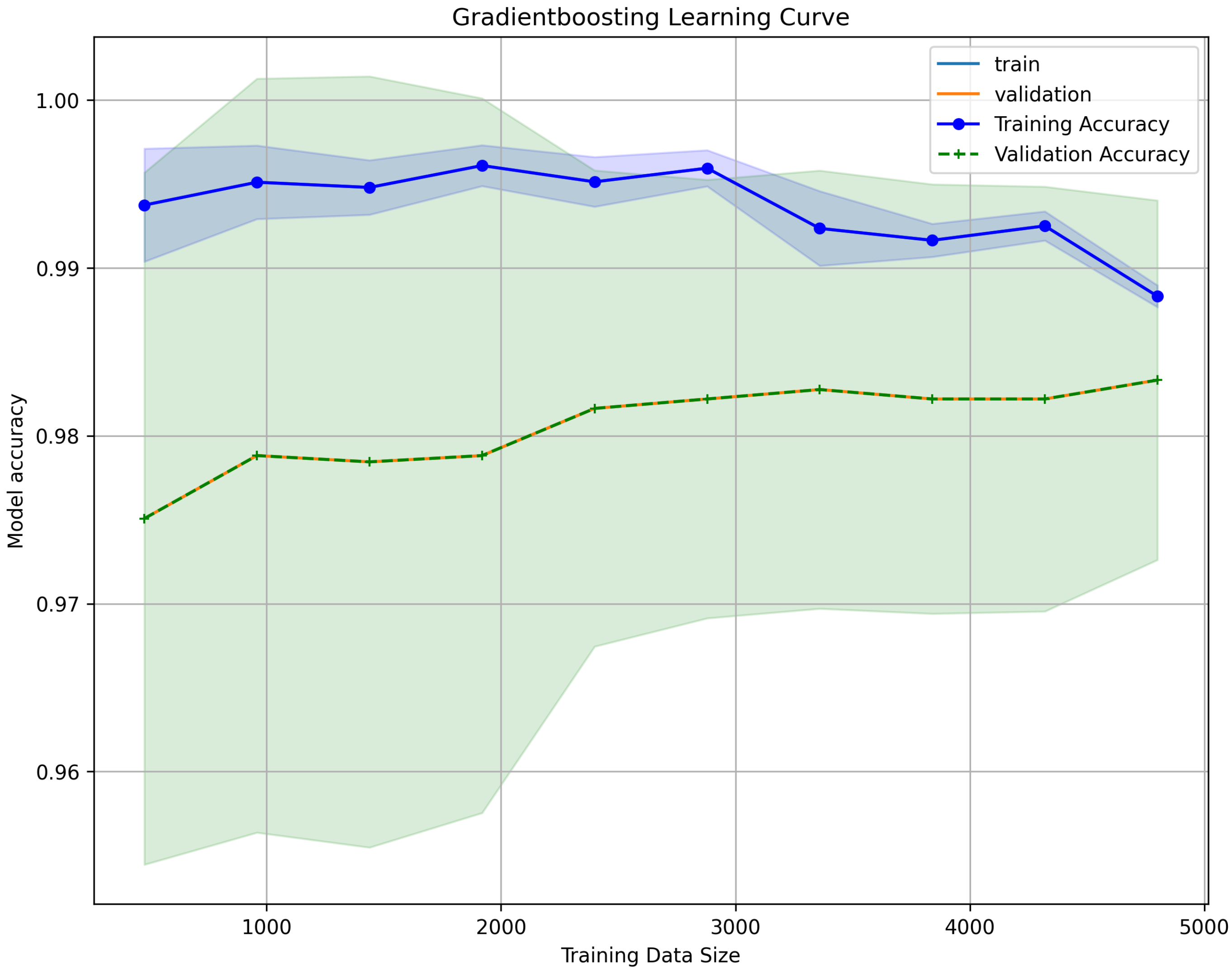

Fig 22. Learning curve graph: Caption. Gradient boosting classifier pre-sampling learning curve

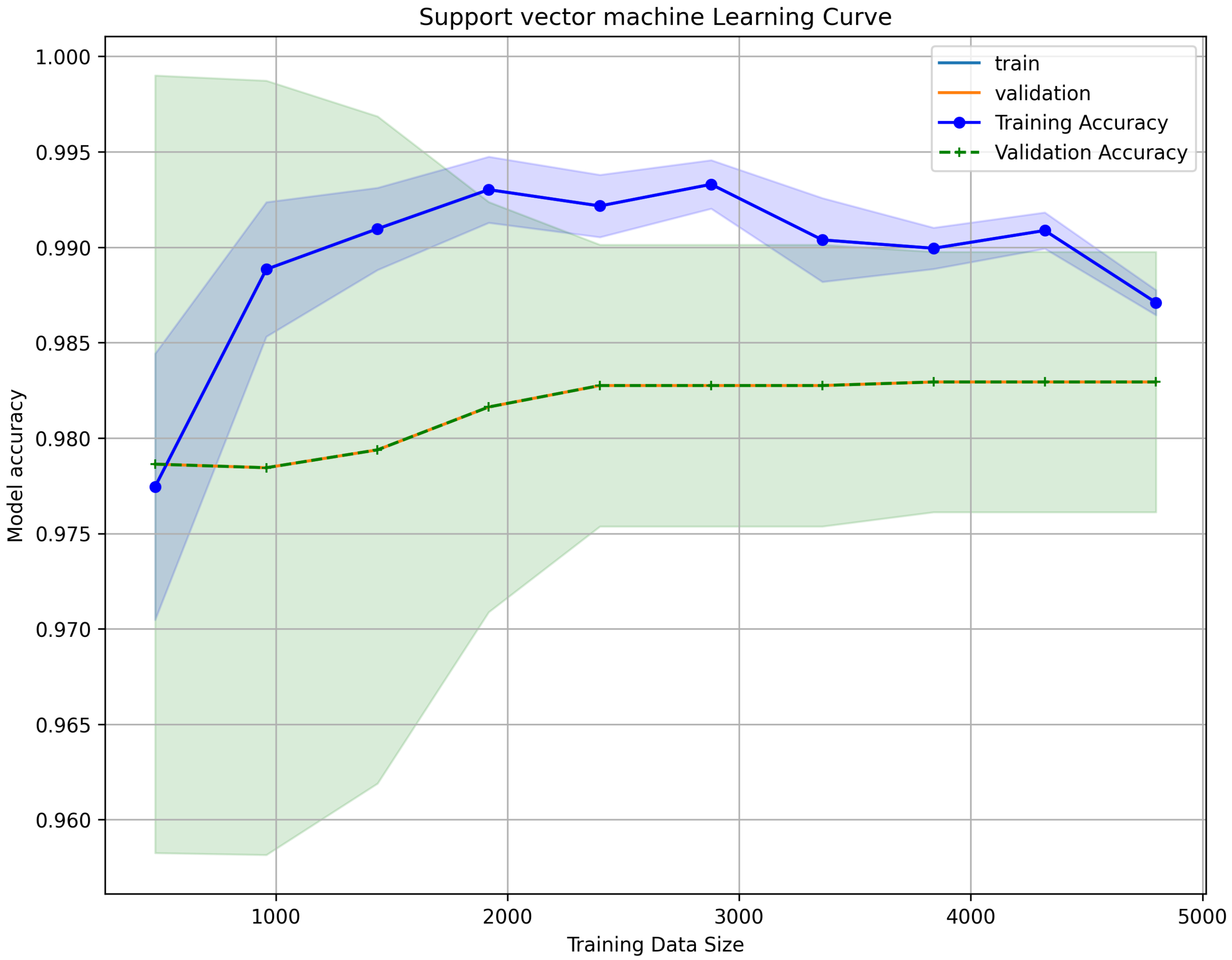

Fig 23. Learning curve graph: Caption. Support vector machine pre-sampling learning curve

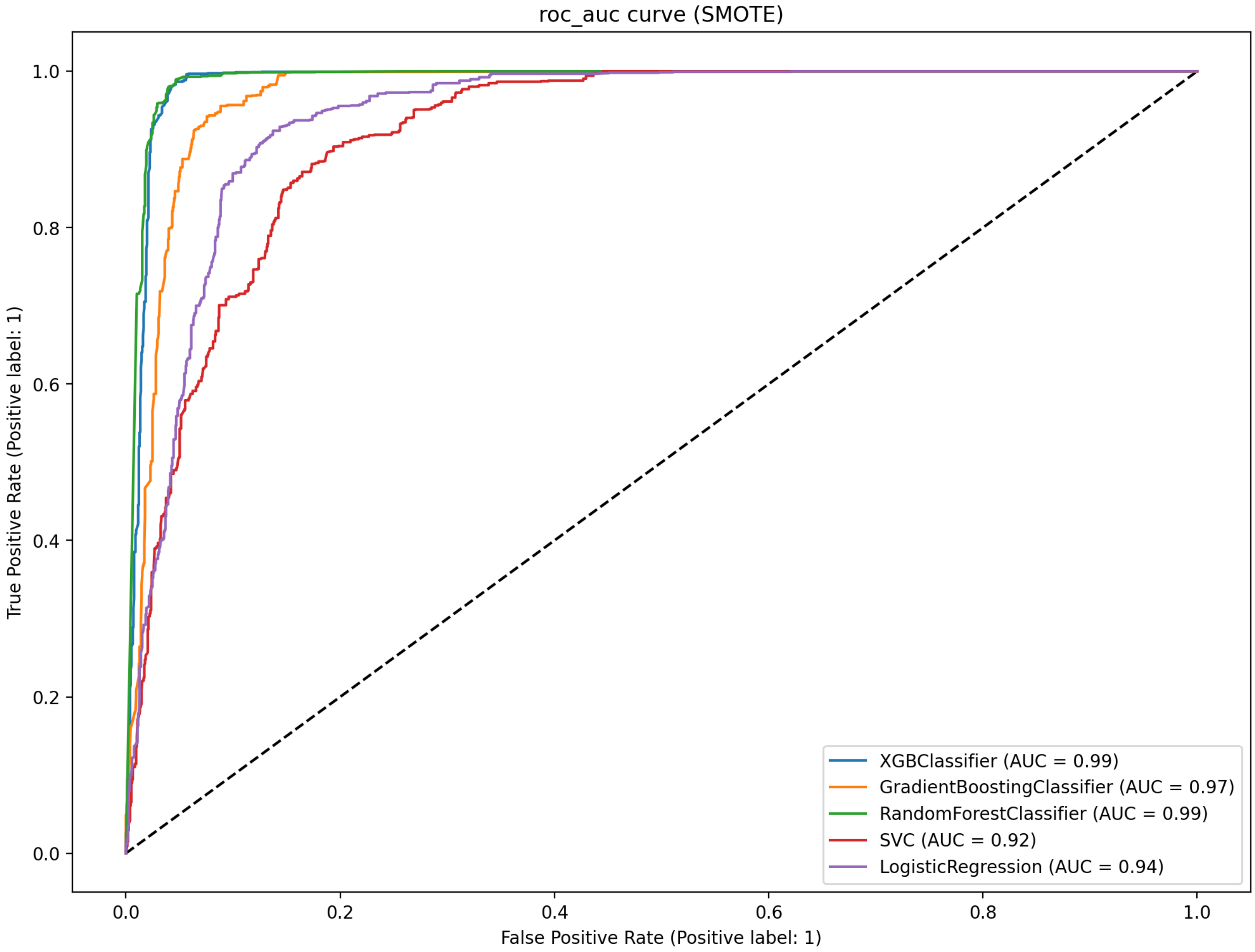

Fig 24. SMOTE over-sampling: Caption. Applied SMOTE roc_auc score curve for over-sampling.

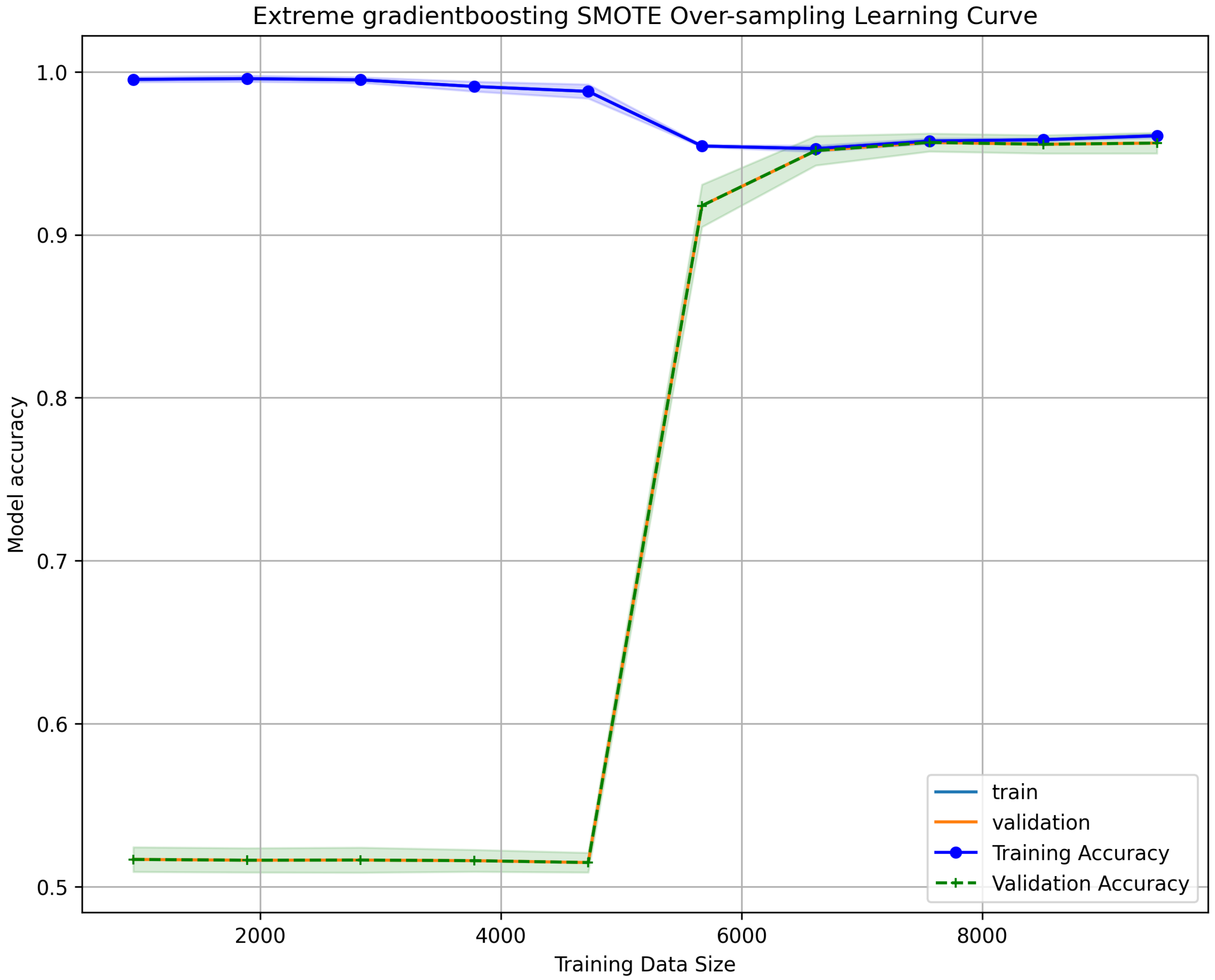

Fig 25. SMOTE over-sampling: Caption. Extreme gradient boosting classifier performance in applied SMOTE over-sampling.

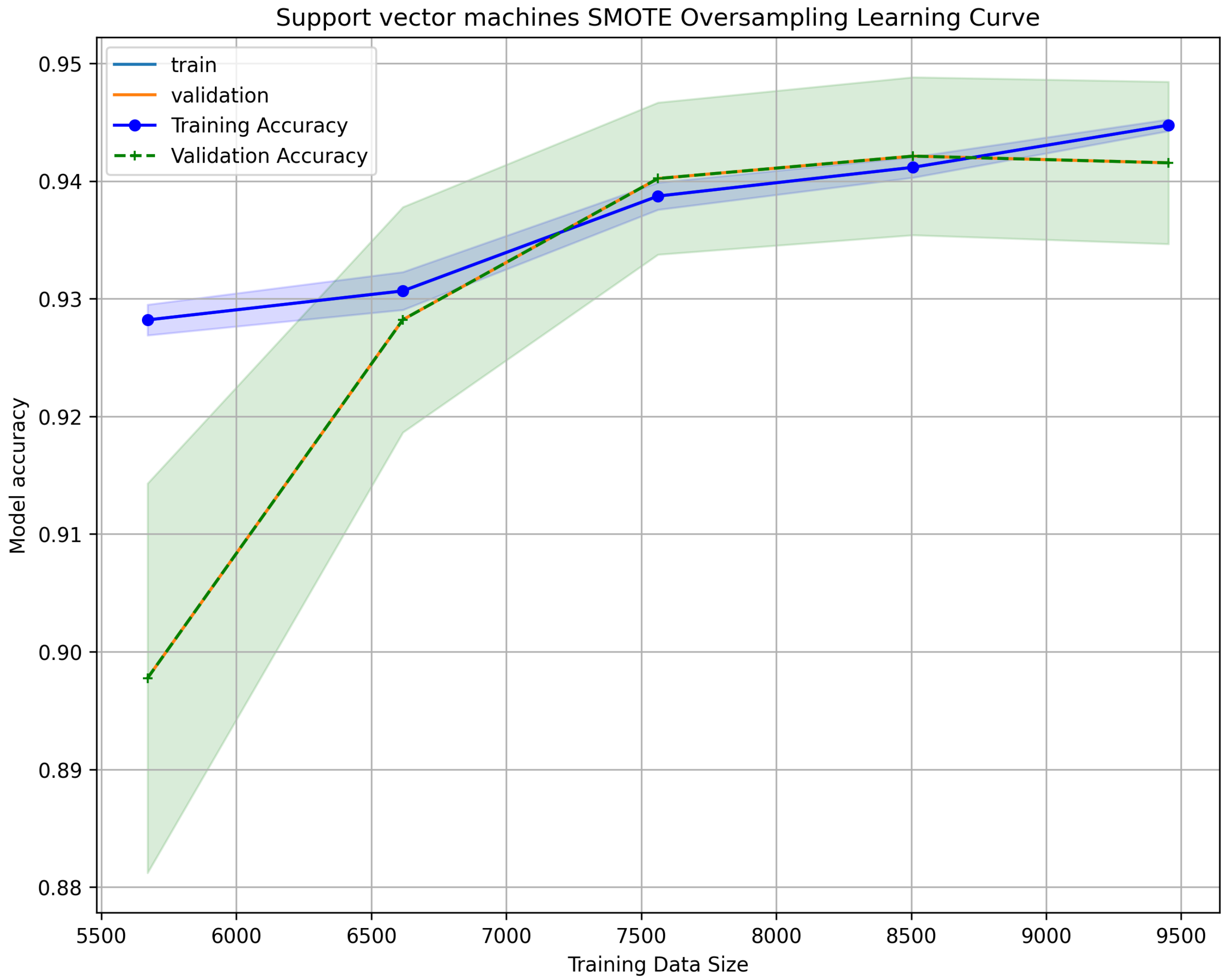

Fig 26. SMOTE over-sampling: Caption. Support vector machine classifier performance in applied SMOTE over-sampling

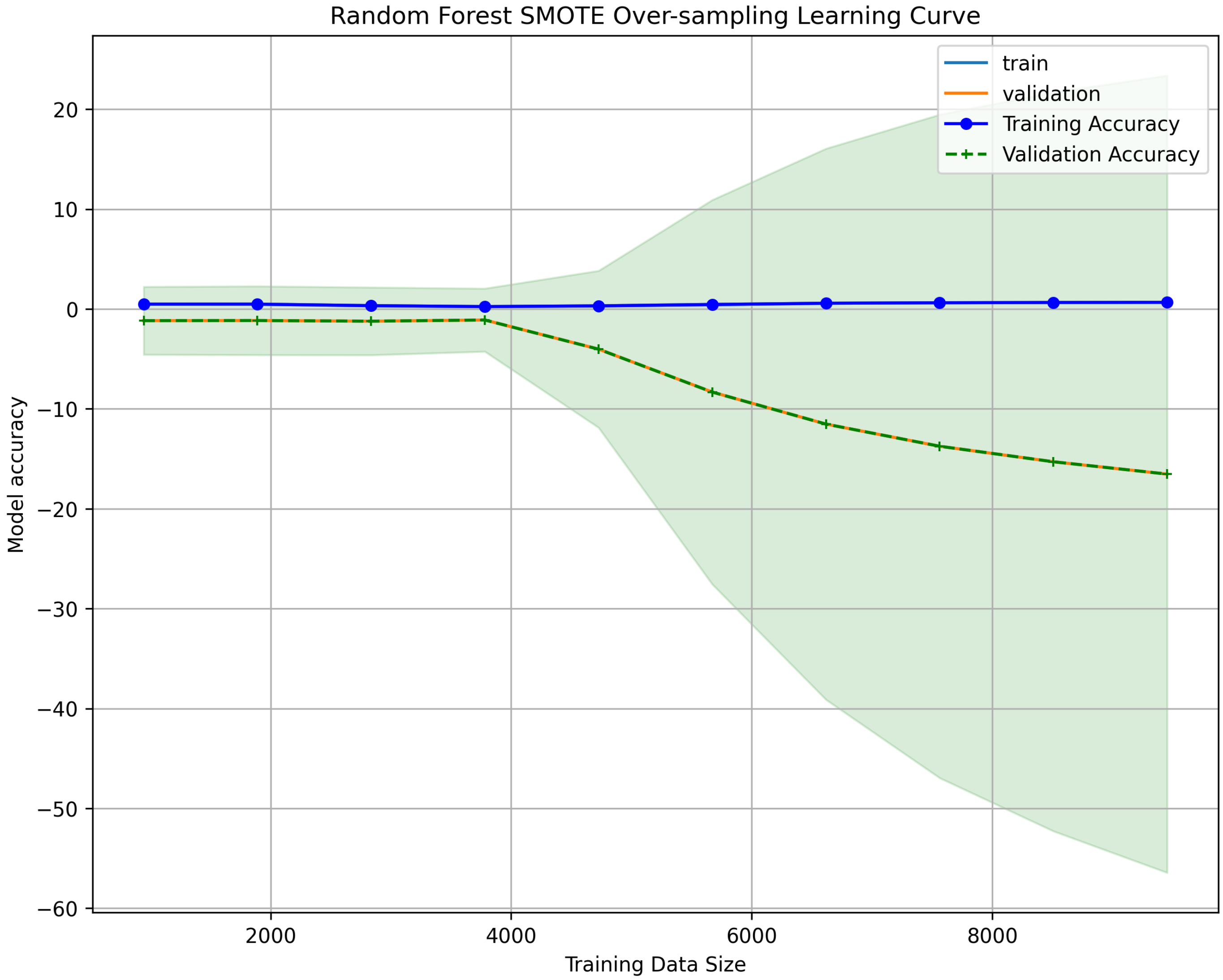

Fig 27. SMOTE over-sampling: Caption. Random forest classifier performance in applied SMOTE over-sampling

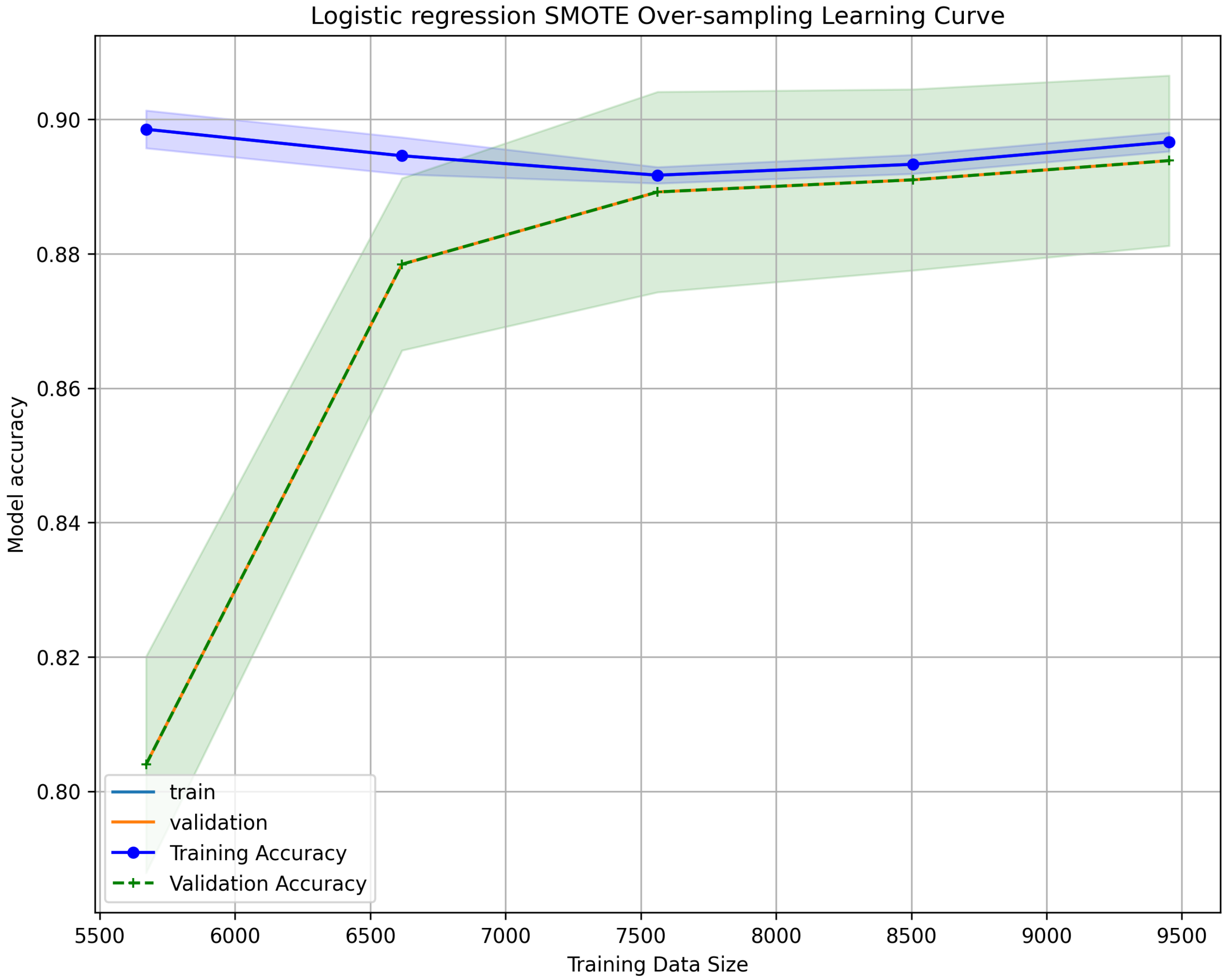

Fig 28. SMOTE over-sampling: Caption. Logistic regression performance in applied SMOTE over-sampling

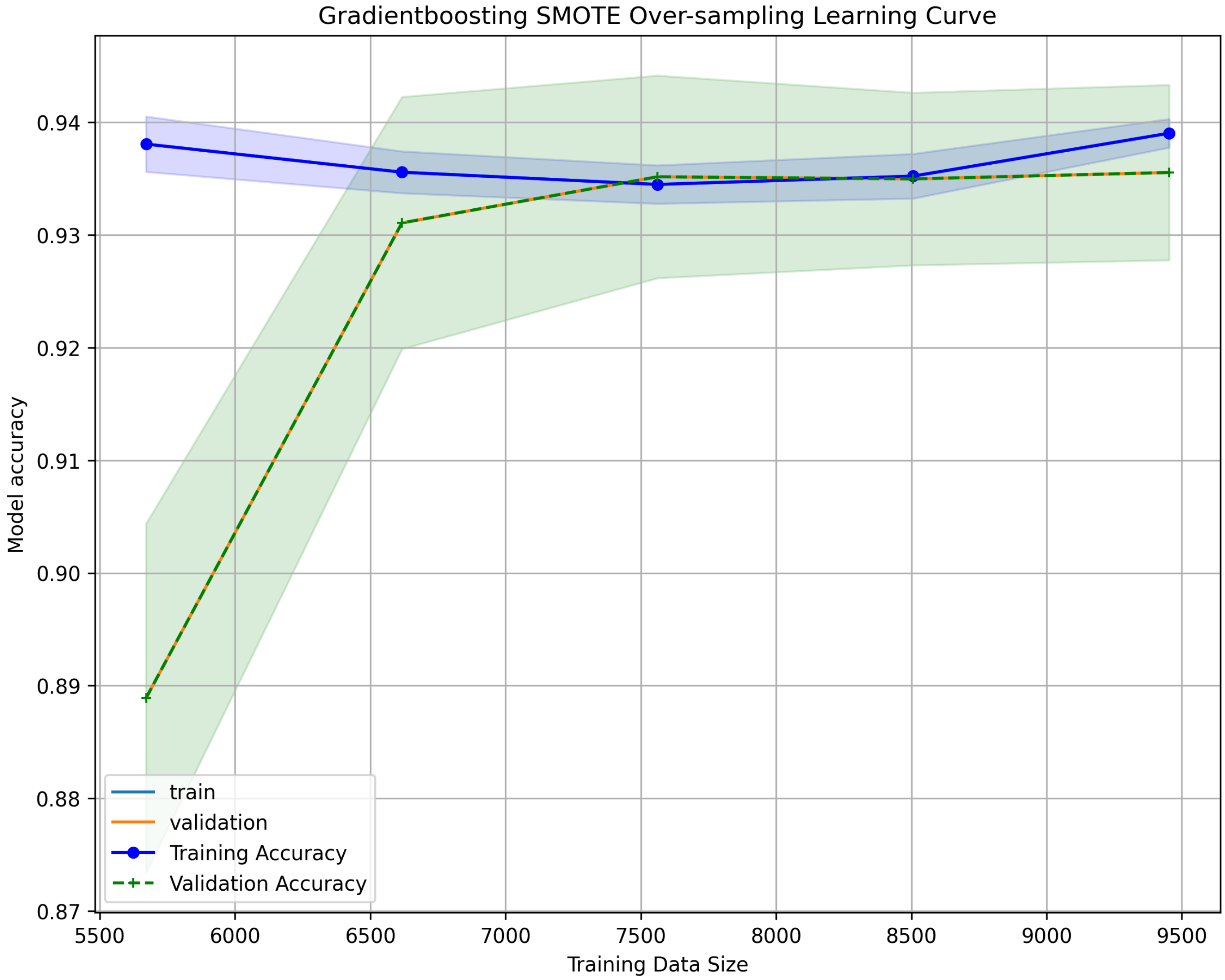

Fig 29. SMOTE over-sampling: Caption. Gradient boosting classifier performance in applied SMOTE over-sampling

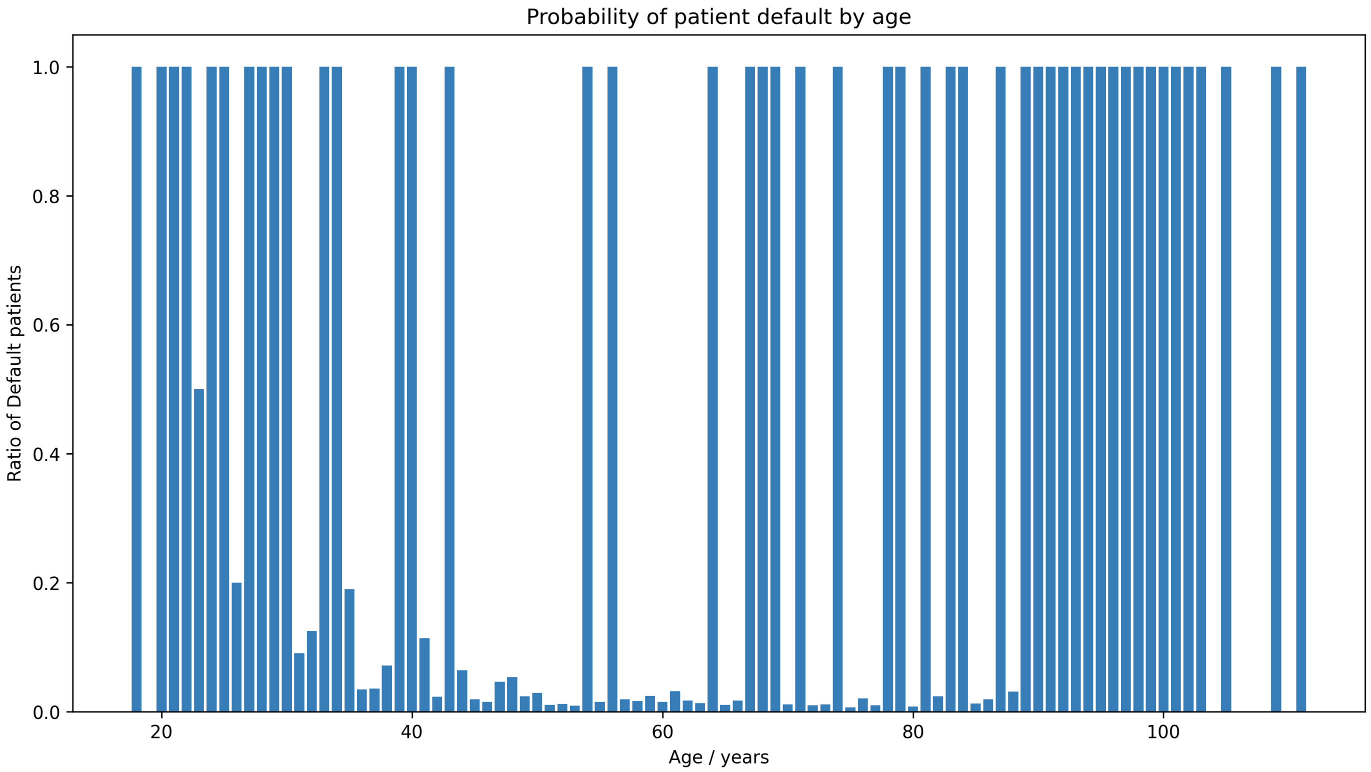

Fig 30. Patient default by age: Caption. Probability of patient default by age.

Statements and Declarations

Authors Contributions: Concept: Michael Owusu-Adjei; Methodology**:** Michael Owusu-Adjei, Twum Frimpong, Gaddafi Abdul-Salaam; Supervision: James Ben Hayfron-Acquah.

Funding: No funding.

Availability of data and Materials: The dataset used for analysis is available through upon request.

Conflicts of Interest: The author(s) declare that there is no conflict of interest regarding the publication of this paper.

Acknowledgments: We acknowledge the support and cooperation of the management and staff of Kwahu Government Hospital, Ghana.

References

- Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, *16*(Sept. 28), 321–357. https://arxiv.org/pdf/1106.1813.pdf%0A http://www.snopes.com/horrors/insects/telamonia.asp
- de Carvalho, A. M., & Prati, R. C. (2020). DTO-SMOTE: Delaunay tessellation oversampling for imbalanced data sets. *Information (Switzerland)*, *11*(12), 1–22. https://doi.org/10.3390/info11120557

- García, V., Sánchez, J. S., Marqués, A. I., Florencia, R., & Rivera, G. (2020). Understanding the apparent superiority of over-sampling through an analysis of local information for class-imbalanced data. *Expert Systems with Applications*, *158*, 113026. https://doi.org/10.1016/J.ESWA.2019.113026
- Gichuhi, H. W., Magumba, M., Kumar, M., & Mayega, R. W. (2023). A machine learning approach to explore individual risk factors for tuberculosis treatment non-adherence in Mukono district. *PLOS Global Public Health*, *3*(7), e0001466. https://doi.org/10.1371/journal.pgph.0001466
- Hernandez, J., Carrasco-Ochoa, J. A., & Martínez-Trinidad, J. F. (2013). An empirical study of oversampling and undersampling for instance selection methods on imbalance datasets. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *8258 LNCS*(PART 1), 262–269. https://doi.org/10.1007/978-3-642-41822-8_33/COVER
- Jeong, D. H., Kim, S. E., Choi, W. H., & Ahn, S. H. (2022). A Comparative Study on the Influence of Undersampling and Oversampling Techniques for the Classification of Physical Activities Using an Imbalanced Accelerometer Dataset. *Healthcare (Switzerland)*, *10*(7). https://doi.org/10.3390/healthcare10071255
- Joloudari, J. H., Marefat, A., Nematollahi, M. A., Oyelere, S. S., & Hussain, S. (2023). Effective Class-Imbalance Learning Based on SMOTE and Convolutional Neural Networks. *Applied Sciences (Switzerland)*, *13*(6). https://doi.org/10.3390/app13064006
- Korneev, N. V., Korneeva, J. V., Yurkevichyus, S. P., & Bakhturin, G. I. (2022). An Approach to Risk Assessment and Threat Prediction for Complex Object Security Based on a Predicative Self-Configuring Neural System. *Symmetry*, *14*(1). https://doi.org/10.3390/sym14010102
- Lee, D., & Kim, K. (2021). An efficient method to determine sample size in oversampling based on classification complexity for imbalanced data. *Expert Systems with Applications*, *184*, 115442. https://doi.org/10.1016/J.ESWA.2021.115442
- Mohammed, R., Rawashdeh, J., & Abdullah, M. (2020). Machine Learning with Oversampling and Undersampling Techniques: Overview Study and Experimental Results. *2020 11th International Conference on Information and Communication Systems, ICICS 2020*, 243–248. https://doi.org/10.1109/ICICS49469.2020.239556
- Nguyen, T., Mengersen, K., Sous, D., & Liquet, B. (2023). SMOTE-CD: SMOTE for compositional data. *PLoS ONE*, *18*(6 June), 1–19. https://doi.org/10.1371/journal.pone.0287705
- Overview, I. C. S. (2014). on the Performance of. *IEEE Vehicular Technology Conference*, *24*(1), 645–660. https://doi.org/10.1007/978-3-030-47436-2
- Ray, E. L., & Reich, N. G. (2018). Prediction of infectious disease epidemics via weighted density ensembles. *PLoS Computational Biology*, *14*(2), 1–23. https://doi.org/10.1371/journal.pcbi.1005910
- Rodríguez-Torres, F., Martínez-Trinidad, J. F., & Carrasco-Ochoa, J. A. (2022). An Oversampling Method for Class Imbalance Problems on Large Datasets. *Applied Sciences (Switzerland)*, *12*(7). https://doi.org/10.3390/app12073424
- Santangelo, O. E., Gentile, V., Pizzo, S., Giordano, D., & Cedrone, F. (2023). Machine Learning and Prediction of Infectious Diseases: A Systematic Review. *Machine Learning and Knowledge Extraction*, *5*(1), 175–198. https://doi.org/10.3390/make5010013
- Satriaji, W., & Kusumaningrum, R. (2018). Effect of Synthetic Minority Oversampling Technique (SMOTE), Feature

Representation, and Classification Algorithm on Imbalanced Sentiment Analysis. *2018 2nd International Conference on Informatics and Computational Sciences, ICICoS 2018*, 99–103. https://doi.org/10.1109/ICICOS.2018.8621648

- Saul, M., & Rostami, S. (2022). Assessing performance of artificial neural networks and re-sampling techniques for healthcare datasets. *Health Informatics Journal*, *28*(1). https://doi.org/10.1177/14604582221087109
- Sowjanya, A. M., & Mrudula, O. (2023). Effective treatment of imbalanced datasets in health care using modified SMOTE coupled with stacked deep learning algorithms. *Applied Nanoscience (Switzerland)*, *13*(3), 1829–1840. https://doi.org/10.1007/s13204-021-02063-4
- Yang, W., Zhang, J., & Ma, R. (2020). The prediction of infectious diseases: A bibliometric analysis.*International Journal of Environmental Research and Public Health*, *17*(17), 1–19. https://doi.org/10.3390/ijerph17176218
- Zhang, Y., & Chen, L. (2021). A Study on Forecasting the Default Risk of Bond Based on XGboost Algorithm and Over-Sampling Method. *Theoretical Economics Letters*, *11*(02), 258–267. https://doi.org/10.4236/tel.2021.112019