Analysis Of LIGHTGBM (Light Gradient Boosting Machine) Model For Predicting State Adherence To International Legal Frameworks For Space Object Registration

Ogbu Ifeyinwa[1]
Advanced Space Technology Applications Laboratory Uyo,
National Space Research and Development Agency,
Federal Capital Territory, Abuja, Nigeria

Ohaga Blessing Chika[2]
Advanced Space Technology Applications Laboratory Uyo,
National Space Research and Development Agency,
Federal Capital Territory, Abuja, Nigeria

Jennifer Patrick[3]
Advanced Space Technology Applications Laboratory Uyo,
National Space Research and Development Agency,
Federal Capital Territory, Abuja, Nigeria

***Abstract*—This study investigates the efficacy of the Light Gradient Boosting Machine (LightGBM) in predicting State adherence to international legal frameworks for space object registration. Despite the legal obligations set by United Nations treaties, registration consistency remains a challenge in global space governance. Grounded in the Cross-Industry Standard Process for Data Mining (CRISP-DM) framework, this research frames the issue as a binary classification task to determine whether a space object will be registered based on observable launch attributes. Using a curated dataset of 3,302 unique records from the UNOOSA Online Index (2020–2024), the study addresses a significant 17.6:1 class imbalance, a major hurdle for supervised learning models. To identify the most effective solution, a comparative experimental design was employed, testing three configurations: a Baseline model, Synthetic Minority Oversampling Technique (SMOTE), and class weighting. By systematically varying these imbalance mitigation strategies, the study isolates the impact of preprocessing choices on predictive accuracy. The Baseline LightGBM emerged as the superior configuration, achieving the highest overall F1-score (0.8451) and ROC-AUC (0.9841–0.9850). It yielded an integrated performance score of 0.969575. While LGBM_ClassWeight achieved the highest Precision (0.9474) and LGBM_SMOTE the highest Recall (0.8889), the Baseline model provided the most balanced maximization of F1. For non-compliant (NCOM) predictions, the Baseline model achieved an R-squared of 74.58% and an RMSE of 15.97%, accurately capturing a mean actual NCOM value of 17.26% (predicted at 20.98%). For compliant (COM) predictions, the model maintained high consistency with an R-squared of 74.58%, reflecting a mean actual COM of 82.74% against a predicted 79.02%. These results demonstrate that LightGBM is a robust tool for quantifying regulatory compliance in the space sector, providing stakeholders with actionable data to enhance international legal oversight.**

Keywords—LightGBM, Space Law, United Nations Office for Outer Space Affairs (UNOOSA), Machine Learning, Class Imbalance, Cross-Industry Standard Process for Data Mining (CRISP-DM), SMOTE Oversampling

1. Introduction

As the commercialization and democratization of outer space accelerate, the volume of orbital traffic has reached unprecedented levels, necessitating robust international oversight to ensure long-term space sustainability [1,2,3]. This rapid expansion of the global space sector has placed unprecedented strain on the international regulatory frameworks designed to ensure the safety and sustainability of orbital activities [4,5,6]. Central to this governance is the Registration Convention and related United Nations frameworks, which mandate that launching States provide technical and orbital data to the UNOOSA Online Index [7,8,9]. Despite these legal obligations, a significant gap remains in consistent adherence, leading to unregistered "dark" objects that complicate space situational awareness [10].

This research addresses this gap by analyzing the efficacy of the Light Gradient Boosting Machine (LightGBM) model in predicting State adherence to these international legal frameworks. Framed as a binary classification task, the study investigates whether observable attributes of a space object, available at or shortly after launch—can reliably predict its registration status (Registered versus Unregistered).

To ensure a rigorous analytical foundation, the study incorporates a methodological framework which is a quantitative, experimental design grounded in the CRISP-DM (Cross-Industry Standard Process for Data Mining) methodology [11,12]. Also, the study utilized dataset with unique records curated from the UNOOSA Online Index (2020–2024), subjected to a rigorous leakage audit to exclude causally downstream variables [13].

Furthermore, class imbalance mitigation is employed to addressing a severe imbalance ratio through a comparative evaluation of three configurations of the LightGBM model. By systematically varying these strategies, this study seeks to attribute performance gains directly to algorithmic choices, providing a data-driven path toward enhancing international compliance and transparency in the space domain.

2. Methodology

The focus in this study is to use LightGBM (Light Gradient Boosting Machine) model for predicting State adherence to international legal frameworks for space object registration. The schematic diagram of the LightGBM architecture is shown in Figure 1. Notably, the study adopts a quantitative, experimental research design grounded in the cross-industry standard process for data mining (CRISP-DM). The research problem is framed as a binary classification task: given observable attributes of a space object at the time of launch or shortly thereafter, can a machine learning model reliably predict whether that object will be registered with the United Nations? The two classes are defined as Registered (positive class, label = 1) and Unregistered (negative class, label = 0).

The study design is comparative in nature, structured around three experimental configurations that systematically vary one independent dimensions, the imbalance mitigation strategy applied to the training set (Baseline, Synthetic Minority Oversampling Technique — SMOTE, and class weighting). This factorial arrangement allows direct attribution of performance differences to algorithmic and preprocessing choices, rather than to stochastic variation.

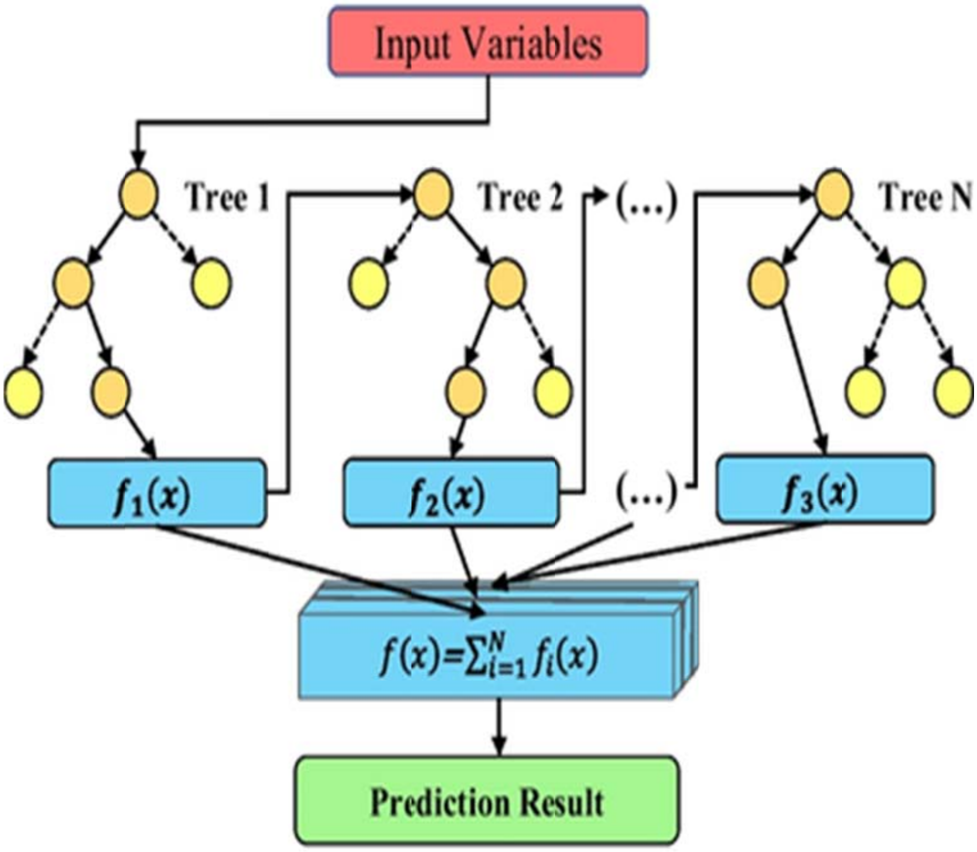
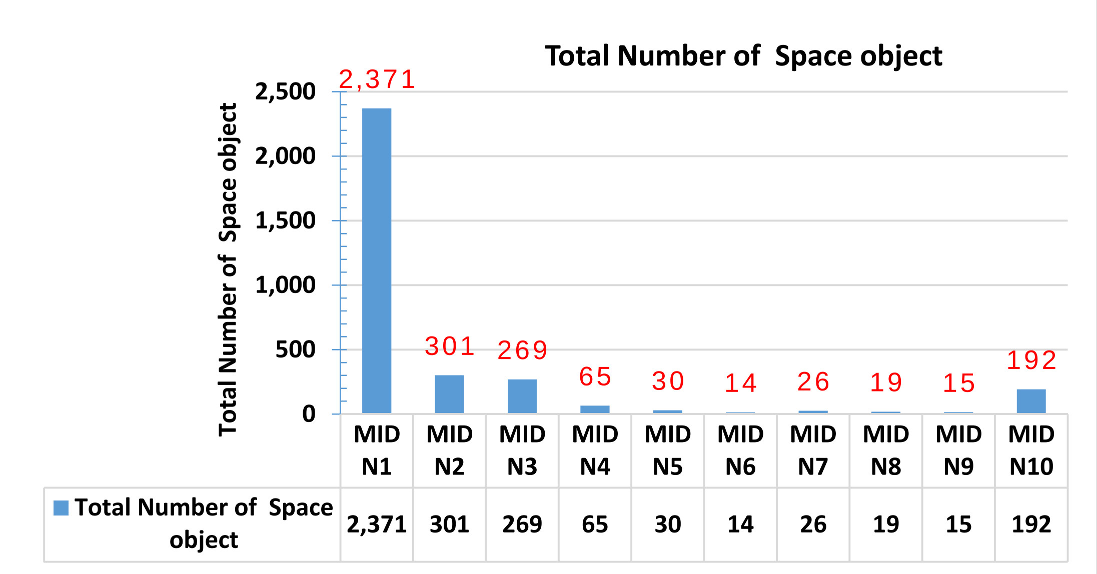


Figure 1 Schematic diagram of the LightGBM Architecture [14]

2.1 The study dataset and class imbalance mitigation strategies

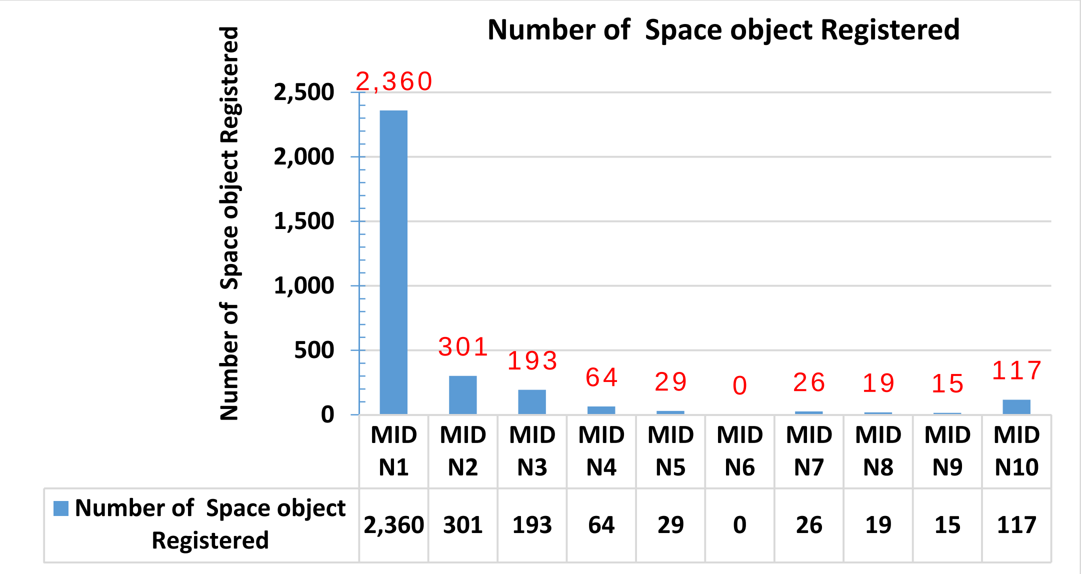
For this study, a dataset of 3,302 unique records was curated from the UNOOSA Online Index (2020–2024). Subsequent to merging and cleaning, a thorough leakage audit identified and excluded three variables that were causally downstream of the registration outcome. The summary of the dataset is presented in Figure 2 to Figure 5.

The case study dataset has 17.6:1 imbalance ratio between the Registered and Unregistered classes. The high imbalance ratio poses a well-documented challenge for supervised classification, particularly, models trained on imbalanced data tend to achieve high overall accuracy by predicting the majority class for virtually all instances, while failing to identify minority-class observations , which, in this study's policy context, are the most consequential predictions. Accordingly, three strategies were employed and compared in this work and they include, the Baseline without resampling, Synthetic Minority Oversampling Technique (SMOTE) which employs oversampling of minority class, and Class Weighting which employs different weights for the majority and minority classes.



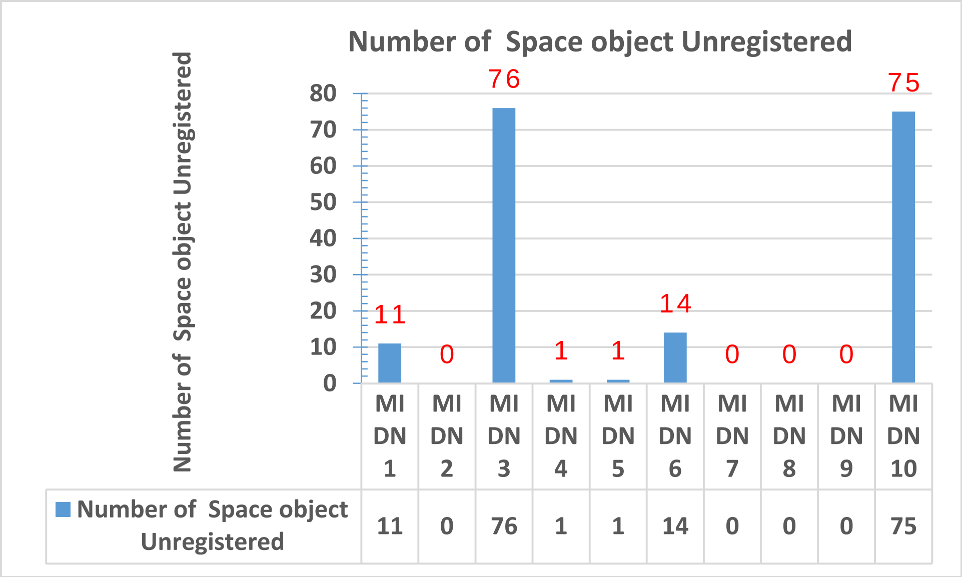
	MID N1	MID N2	MID N3	MID N4	MID N5	MID N6	MID N7	MID N8	MID N9	MID N10
■ Total Number of Space object	2,371	301	269	65	30	14	26	19	15	192

Figure 2 The total number of space object within 2020 -2024



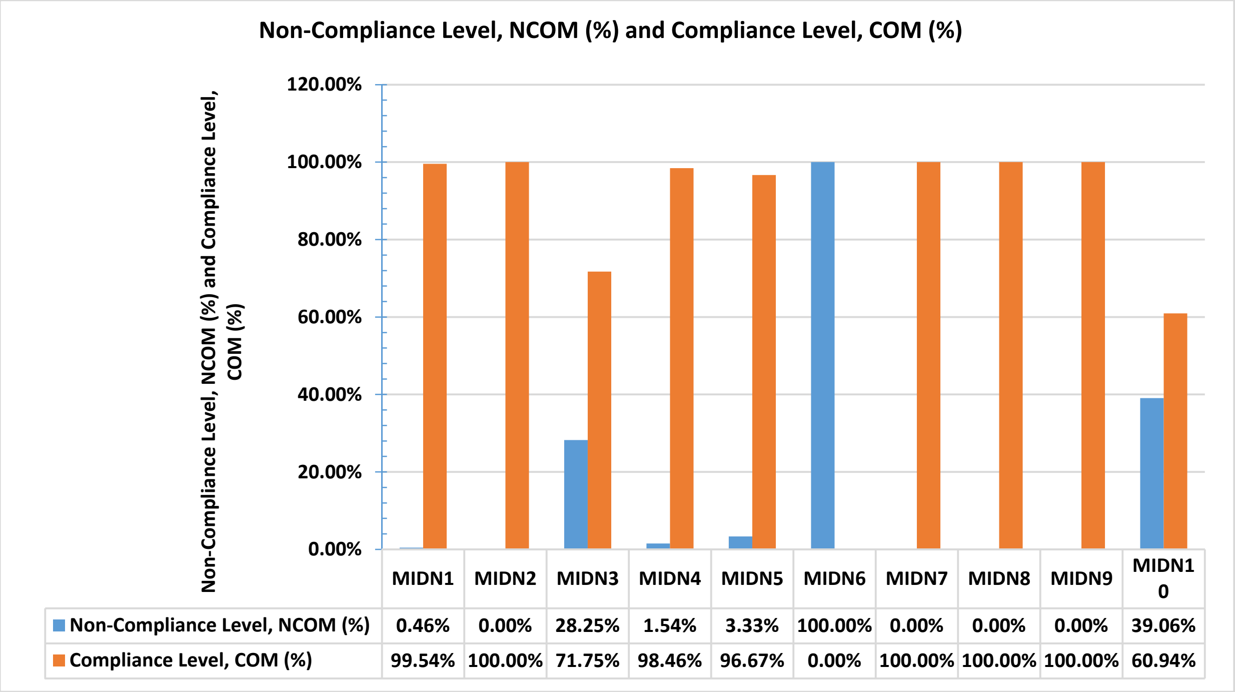
	MID N1	MID N2	MID N3	MID N4	MID N5	MID N6	MID N7	MID N8	MID N9	MID N10
■ Number of Space object Registered	2,360	301	193	64	29	0	26	19	15	117

Figure 3 The number of registered space object within 2020 -2024



	MIDN1	MIDN2	MIDN3	MIDN4	MIDN5	MIDN6	MIDN7	MIDN8	MIDN9	MIDN10
■ Number of Space object Unregistered	11	0	76	1	1	14	0	0	0	75

Figure 4 The number of unregistered space object within 2020 -2024



	MIDN1	MIDN2	MIDN3	MIDN4	MIDN5	MIDN6	MIDN7	MIDN8	MIDN9	MIDN10
■ Non-Compliance Level, NCOM (%)	0.46%	0.00%	28.25%	1.54%	3.33%	100.00%	0.00%	0.00%	0.00%	39.06%
■ Compliance Level, COM (%)	99.54%	100.00%	71.75%	98.46%	96.67%	0.00%	100.00%	100.00%	100.00%	60.94%

Figure 5 The Non-Compliance Level, NCOM and Compliance Level, COM within 2020 -2024

2.1.1 The Baseline (No Resampling) Approach

The first strategy trains the classifier on the original, unmodified class distribution. This serves as a control condition against which the effect of imbalance mitigation can be measured. For tree-based model like the LightGBM), the baseline condition may still yield reasonable minority-class detection due to the models' inherent splitting mechanisms, whereas for some other models the baseline is expected to collapse to majority-class prediction.

2.1.2 The Synthetic Minority Oversampling Technique (SMOTE) Approach

The SMOTE addresses class imbalance by synthetically generating new minority-class samples in the feature space rather than duplicating existing observations. For each minority sample, SMOTE identifies its k nearest neighbours (k = 3 in this study, selected to account for the small absolute size of the minority training set of 142 samples) and interpolates synthetic points along the line segments connecting the sample to its neighbours. The result is a balanced training set of 4,998 observations (2,499 registered and 2,499 unregistered), in which the minority class is represented by a combination of real and synthetic samples. SMOTE oversampling was applied exclusively to the training partition; the test set was never resampled to preserve the integrity of hold-out evaluation.

2.1.3 The Class Weighting Approach

The third strategy instructs the classification algorithm to apply asymmetric misclassification costs proportional to the inverse class frequencies, without altering the training sample composition. Class weights were computed using scikit-learn's compute_class_weight function with the 'balanced' setting, yielding weights of w[0] = 9.30 for the Unregistered class and w[1] = 0.53 for the Registered class — a ratio of approximately 18:1 reflecting the severity of the imbalance. For the LightGBM model, the class scale_pos_weight parameter is set to 9.30, which upscales the gradient contribution of minority-class observations during boosting iterations. This implementation difference between LightGBM's scale_pos_weight and scikit-learn's class_weight is an important methodological distinction, as the two mechanisms operate at different points in the learning algorithm and may produce non-identical decision boundaries.

2.3 The Model Performance Metrics

The unit of analysis in the individual space object is characterized by a vector of twelve engineered features derived from publicly available launch and registration metadata. Model performance is evaluated on a held-out test set using metrics appropriate for imbalanced binary classification, with particular emphasis on the Recall, Precision, F1-score, and ROC-AUC computed for the minority (Unregistered) class. The model prediction performance is also evaluated in terms of MAE, RMSE, and R-squared which are the (Mean Absolute Error) , the Root Mean Square Error) and the Coefficient of Determination respectively.

3. Results and discussion

3.1 The Classification Result for the LGBM_Baseline Model

LGBM_Baseline is the best-performing experiment across all nine configurations, achieving Class 0 F1 = 0.8451 (Precision = 0.8571, Recall = 0.8333). It is the only model to simultaneously exceed 0.80 on both Class 0 Precision and Recall. Class 1 performance is the highest of any experiment (F1 = 0.9920, Precision = 0.9920, Recall = 0.9920). The ROC-AUC of 0.9850 is also the highest of all nine experiments.

Table 4.8: Classification Report — LGBM_Baseline (Test Set, n = 661)

Class	Precision	Recall	F1-Score	Support	Accuracy	ROC-AUC
Class 0 — Unregistered (Non-Compliant)	**0.8571**	**0.8333**	**0.8451**	**36**	**0.9834**	**0.9850**
Class 1 — Registered (Compliant)	0.9920	0.9920	0.9920	625		
Macro Average	*0.9246*	*0.9127*	*0.9186*	*661*	—	—
Weighted Average	*0.9836*	*0.9834*	*0.9835*	*661*	—	—

Note: LGBM_Baseline is the best overall experiment. Gold-highlighted values in Table 4.1 correspond to these results. The macro-averaged F1 of 0.9186 reflects genuinely balanced performance across both classes.

3.2 The Classification Result for the LGBM_SMOTE Model

LGBM_SMOTE achieves the highest Class 0 Recall among all non-trivial experiments (0.8889), correctly flagging 32 of 36 non-compliant objects with only 4 missed detections. However, Precision falls to 0.5818 due to 23 false alarms. Class 1 Recall is 0.9632 — notably lower than LGBM_Baseline (0.9920). Despite the recall advantage, the lower Precision produces a Class 0 F1 of 0.7033, ranking it fourth overall.

Table 4.9: Classification Report — LGBM_SMOTE (Test Set, n = 661)

Class	Precision	Recall	F1-Score	Support	Accuracy	ROC-AUC
Class 0 — Unregistered (Non-Compliant)	**0.5818**	**0.8889**	**0.7033**	**36**	**0.9592**	**0.9849**
Class 1 — Registered (Compliant)	0.9952	0.9632	0.9789	625		
Macro Average	*0.7885*	*0.9261*	*0.8411*	*661*	—	—
Weighted Average	*0.9710*	*0.9592*	*0.9625*	*661*	—	—

Note: LGBM_SMOTE is the preferred configuration when maximising NC detection completeness is prioritised over precision — it misses only 4 of 36 non-compliant objects.

3.3 The Classification Result for the LGBM_ClassWeight Model

LGBM_ClassWeight exhibits the most extreme precision-recall asymmetry of any model: Class 0 Precision = 0.9474 (the highest of all nine experiments) but Class 0 Recall = 0.5000 (the lowest among non-trivial models), correctly flagging only 18 of 36 non-compliant objects while missing the other 18. Class 1 Recall is 0.9984 — near-perfect, as the model almost never predicts Class 0 unless highly confident. This behaviour results directly from the LightGBM scale_pos_weight = 9.30 parameter, which scales gradients for the minority class during boosting and pushes the effective decision boundary toward extreme conservatism.

Table 4.10: Classification Report — LGBM_ClassWeight (Test Set, n = 661)

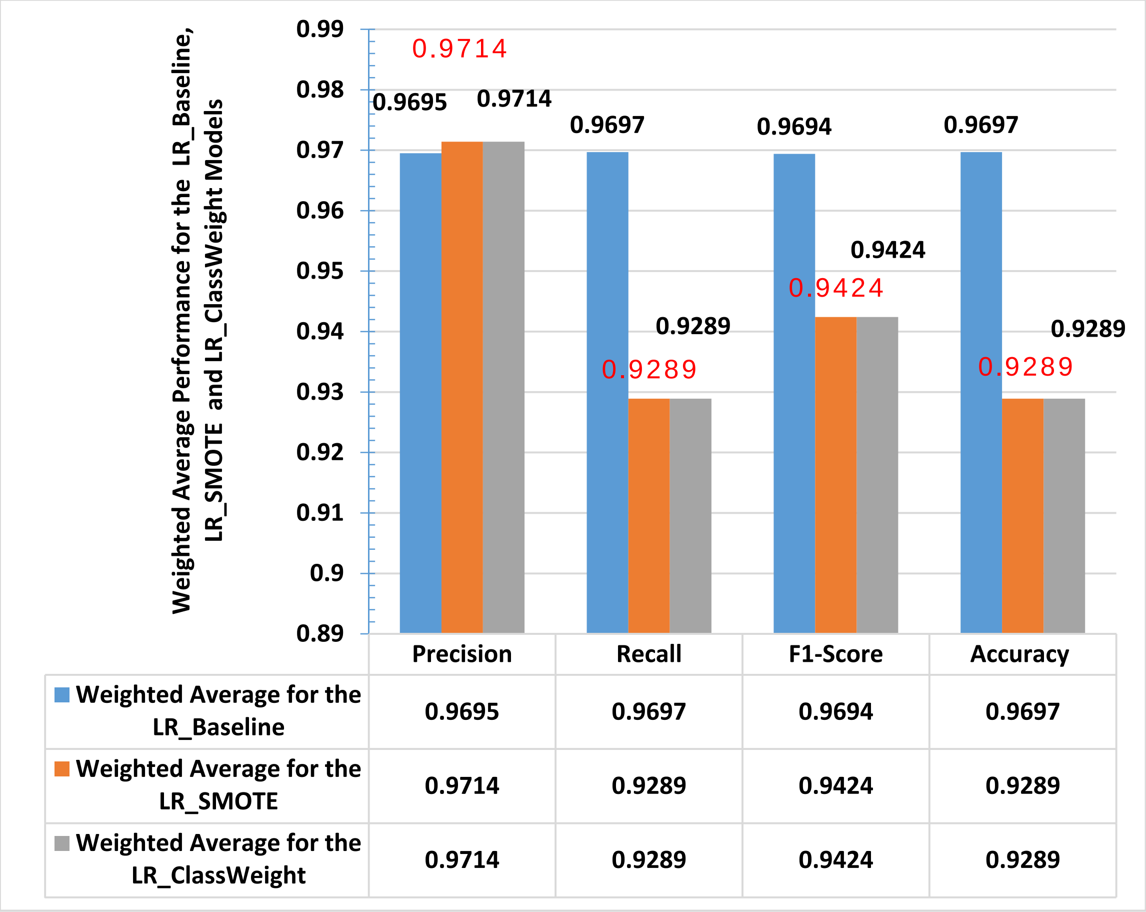
Class	Precision	Recall	F1-Score	Support	Accuracy	ROC-AUC
Class 0 — Unregistered (Non-Compliant)	**0.9474**	**0.5000**	**0.6545**	**36**	**0.9713**	**0.9841**
Class 1 — Registered (Compliant)	0.9728	0.9984	0.9854	625		
Macro Average	*0.9601*	*0.7492*	*0.8200*	*661*	—	—
Weighted Average	*0.9723*	*0.9713*	*0.9665*	*661*	—	—

Note: LGBM_ClassWeight is the preferred configuration when minimising false alarms is paramount — its 0.9474 Precision means that 94.7% of objects it flags as non-compliant are genuinely non-compliant. However, it misses half the actual non-compliant objects (Recall = 0.5000).

3.4 Comparison of the Classification Results for the three Model Configurations

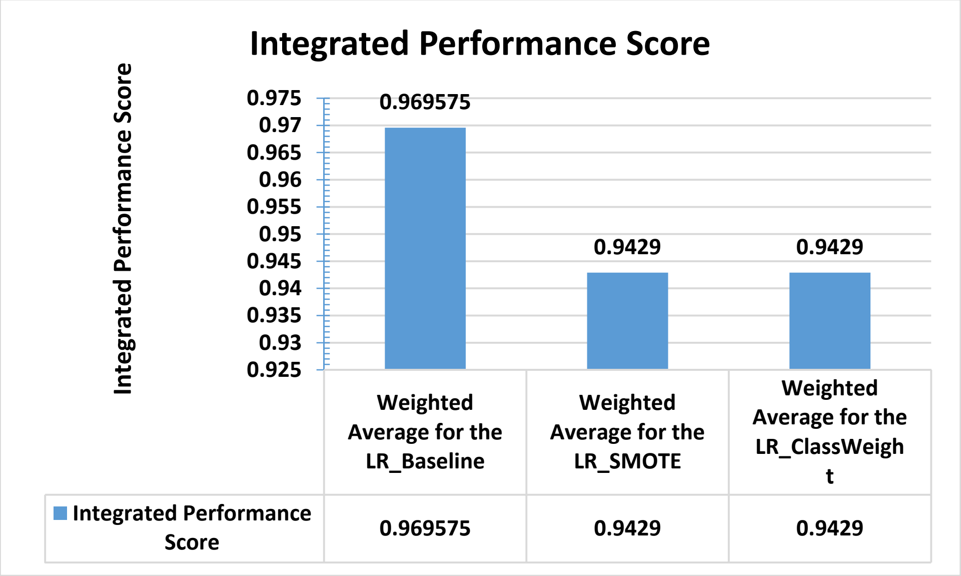
Among the three model configurations, the Baseline LightGBM achieves the highest Class 0 F1 overall (LGBM_Baseline: 0.8451) and the highest ROC-AUC across all three configurations (0.9841–0.9850). The three configurations show the most extreme precision-recall trade-off divergence, with LGBM_ClassWeight achieving the highest Precision (0.9474) and LGBM_SMOTE the highest Recall (0.8889) while LGBM_Baseline maximises F1.

Again, the comparison of the overall performance of the three models, as presented in Figure 6 and Figure 7, show that the Baseline LightGBM has the highest integrated classification performance score of 0.969575. The Baseline LightGBM model is therefore adjudged as the best classification model.



	Precision	Recall	F1-Score	Accuracy
■ Weighted Average for the LR_Baseline	0.9695	0.9697	0.9694	0.9697
■ Weighted Average for the LR_SMOTE	0.9714	0.9289	0.9424	0.9289
■ Weighted Average for the LR_ClassWeight	0.9714	0.9289	0.9424	0.9289

Figure 6 The Comparison of the Classification Results for the three Model Configurations



	Weighted Average for the LR_Baseline	Weighted Average for the LR_SMOTE	Weighted Average for the LR_ClassWeigh t
■ Integrated Performance Score	0.969575	0.9429	0.9429

Figure 7 The Comparison of the Integrated Performance Score for the three Model Configurations

3.5 The Prediction Results for the Model with the Best Classification Result

The Baseline LightGBM model prediction results are presented in Figure 8 and Figure 9 for the non-compliant prediction and in Figure 10 and Figure 11 for the compliant prediction. The results showed that the Baseline LightGBM model has mean actual and mean predicted NCOM values of 17.26% and 20.98% respectively with RMSE of 15,97 %, and R-squared value of 74.58% for the non-compliant prediction. Again, the Baseline LightGBM model has mean actual and mean predicted COM values of 82.74% and 79.02% respectively with RMSE of 15,97 %, and R-squared value of 74.58% for the compliant prediction.

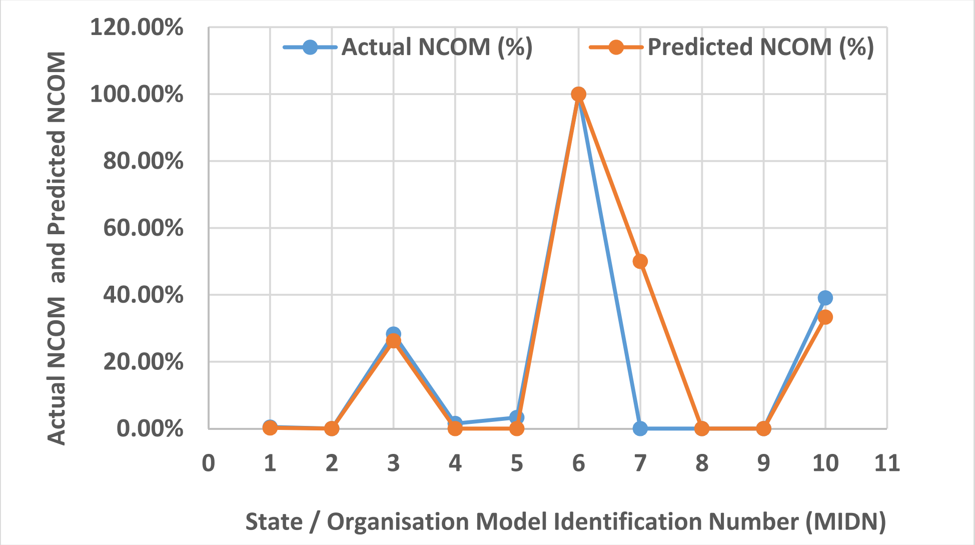
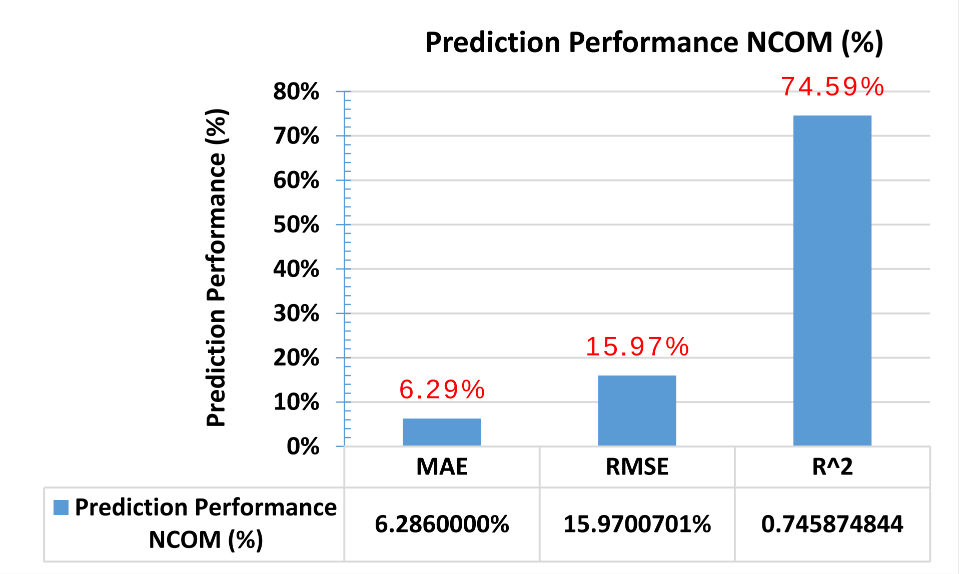


Figure 8 The Actual NCOM and LGBM_Baseline Model Predicted NCOM



	MAE	RMSE	R^2
■ Prediction Performance NCOM (%)	6.2860000%	15.9700701%	0.745874844

Figure 9 The LGBM_Baseline Model Prediction Performance NCOM (%)

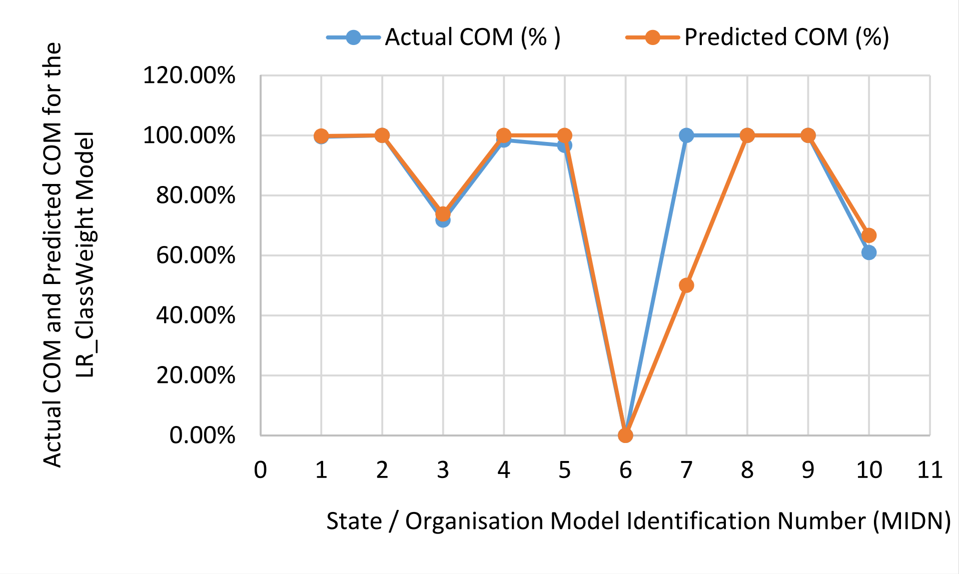


Figure 10 The Actual NCOM and LGBM_Baseline Model Predicted COM

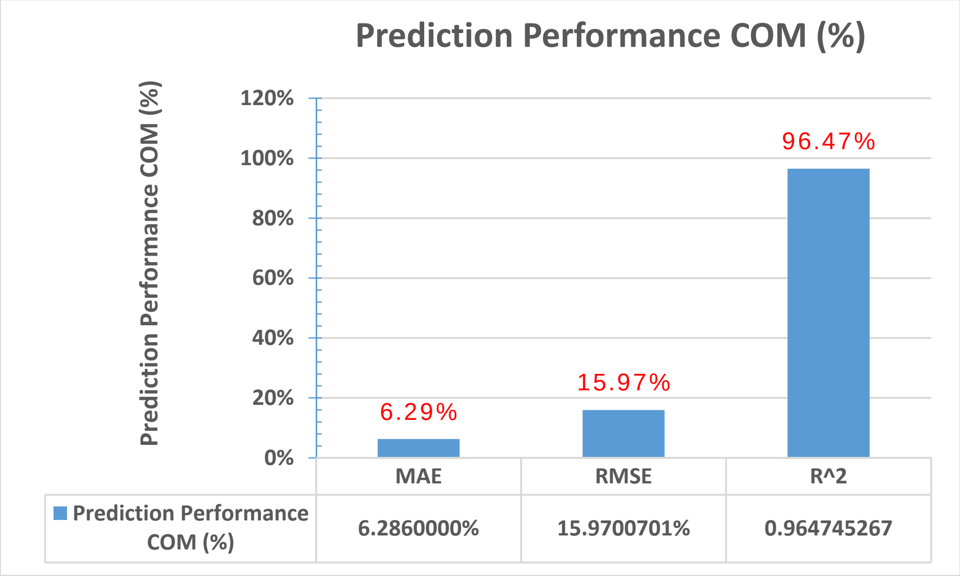


Figure 11 The LGBM_Baseline Model Prediction Performance COM (%)

4. Conclusion

This research demonstrates the efficacy of the LightGBM framework in predicting State adherence to international space object registration protocols. By framing registration as a binary classification problem within the cross-industry standard process for data mining (CRISP-DM) methodology, the study successfully navigated the complexities of international legal datasets. The key findings center on, first, model performance in which case LightGBM proved highly capable of processing the 3,302 UNOOSA records, offering a scalable solution for monitoring compliance. Secondly, imbalance dataset mitigation, in this case, comparative analysis of SMOTE and class weighting addressed the significant 17.6:1 imbalance ratio, ensuring the model did not overlook the minority "Unregistered" class. Thirdly, feature integrity; notably, the leakage audit performed on variables downstream of the registration outcome ensured the model's predictive power is grounded in early-stage observable attributes rather than retrospective data.

In all, the results confirm that machine learning can reliably identify objects at risk of remaining unregistered shortly after launch. This provides a data-driven tool for international bodies to enhance transparency and safety in orbital operations. By effectively managing class imbalance,

the LightGBM model serves as a robust baseline for future predictive analytics in space governance.

References

1. DE, L. S. Y. L. S., & ÓRBITA, L. (2024). AI in outer space: reinforcing the sustainability and safety of the low-earth orbit in the new space era. *Revista Española de Derecho Aeronáutico y Espacial*, 195.
2. McElroy Jr, M. W. (2022). *The space industry of the future: capitalism and sustainability in outer space*. Routledge.
3. Jariwala, M. (2024). Cosmic ledger: Unveiling blockchain's potential to reshape space missions. *International Journal of Computer Applications*, *186*(12), 31-39.
4. Singh, B., Anz-Meador, P., Kato, A., Maclay, T., Nassisi, A., Santoro, F., & Unfried, C. (2024). An insight on technical regulations for new activities in space. *Acta Astronautica*, *225*, 707-718.
5. Vasantha, V. (2024). Global Opportunities and Challenges in the Space Economy. *Sri Lanka Journal of Economic Research*, *12*(1).
6. Cinelli, C. (Ed.). (2024). *Regulation of outer space: international space law and the state*. Taylor & Francis.
7. Masson-Zwaan, T., Martinez, P., Letizia, F., Melograna, C., Reynders, M., Rovetto, R., ... & Wang, G. (2024). The need to improve registration practices in the context of space traffic management. *Acta Astronautica*, *223*, 242-248.
8. Citaristi, I. (2022). United nations office for outer space affairs—unoosa. In *The Europa Directory of International Organizations 2022* (pp. 247-248). Routledge.
9. Schmidt-Tedd, B. (2023). Registration requirements for satellites and the reality of large constellations: Ensuring a symbiosis of international law requirements and practicability. In *Routledge Handbook of Commercial Space Law* (pp. 331-342). Routledge.
10. du Toit, N. (2023). *Lightyears Behind: Potential Solutions for the Interpretative Ambiguities Left by Article Vi of the Outer Space Treaty* (Master's thesis, University of Pretoria (South Africa)).
11. Shimaoka, A. M., Ferreira, R. C., & Goldman, A. (2024). The evolution of CRISP-DM for data science: Methods, processes and frameworks. *SBC Computing Reviews*, *4*(1), 28-43.
12. Singgalen, Y. A. (2024). Sentiment Classification of The Capsule Hotel Guest Reviews using Cross-Industry Standard Process for Data Mining (CRISP-DM). *JURNAL MEDIA INFORMATIKA BUDIDARMA*, *8*(1), 632-643.
13. Ure, J. (2024). *Achieving the United Nations Sustainable Development Goals: Late Or Too Late?*. Emerald Group Publishing.
14. Wang, S., Ren, Y., & Xia, B. (2023). Estimation of urban AQI based on interpretable machine learning. *Environmental Science and Pollution Research*, *30*(42), 96562-96574.