Stability and Robustness of Model-Agnostic Interpretability Methods under Class Imbalance

Introduction

The rapid advancement of machine learning (ML) has revolutionized decision-making processes across various domains, including finance, healthcare, and agriculture. Understanding and interpreting their inner workings has become paramount as these models increasingly shape critical decisions. The lack of transparency in complex ML models, often called the "black box" problem, has raised significant concerns about accountability, fairness, and trust in the decisions they support. This issue has been particularly emphasized within the European Union's legal framework, where the General Data Protection Regulation (GDPR), effective as of May 2018, mandates companies to offer detailed explanations on automated decision-making processes involving customers and grants individuals the right to challenge these decisions (European Commission, 2018). In response to such legal and ethical imperatives, the field of Explainable AI (XAI) has emerged, focusing on developing methods that provide insights into model predictions without sacrificing performance, thereby fostering human-centric, sustainable, secure, inclusive, and trustworthy AI (Ribeiro et al., 2016; Gunning, 2017).

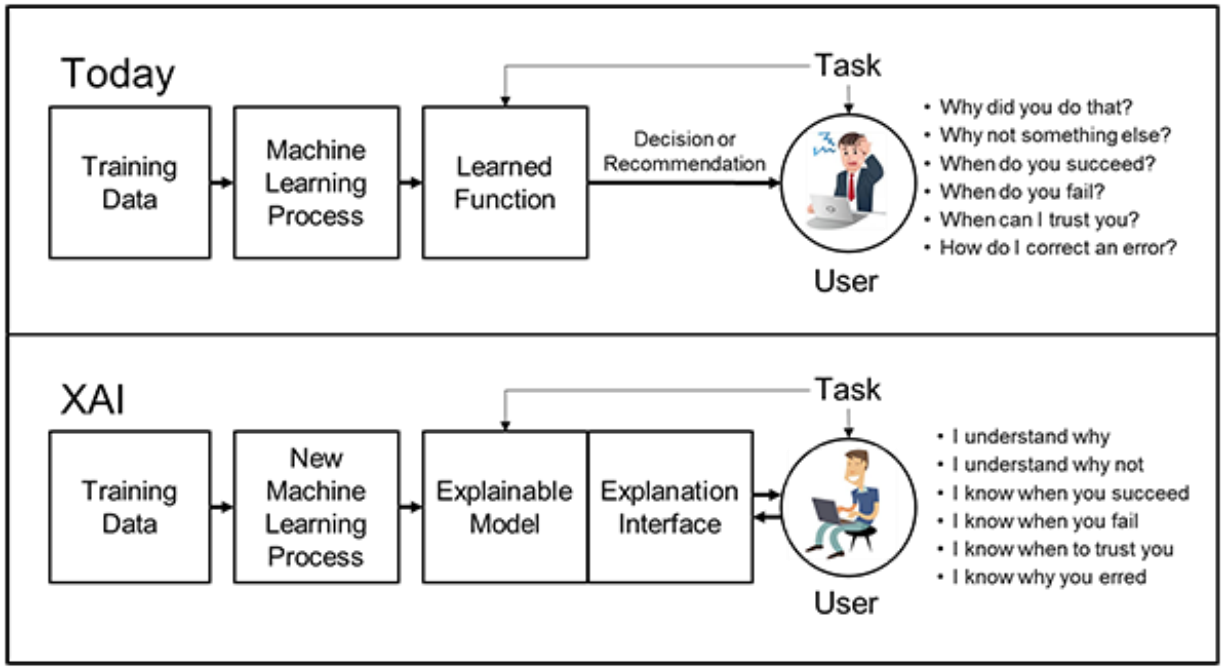

Figure 1: XAI Concept (Gunning, 2017)

Literature Review

Class Imbalance in Machine Learning

Class imbalance is a prevalent challenge in machine learning, occurring when the distribution of instances across different classes in a dataset is significantly disproportionate (He and Garcia, 2009). In imbalanced datasets, the majority class(es) have a substantially higher number of instances compared to the minority class(es), which can lead to biased and unreliable model performance (Haixiang et al., 2017). This issue is particularly concerning in critical applications such as healthcare, finance, and security, where the consequences of misclassification can be severe (Kaur et al., 2019).

Researchers primarily focus on enhancing the accuracy of predictive models to address class imbalance. They have introduced several strategies to tackle the challenges class imbalance poses to model efficacy. These include resampling techniques, such as augmenting the minority class or reducing the majority class, and algorithm modifications, like adopting cost-sensitive learning or leveraging ensemble approaches (Chawla et al., 2002; He and Garcia, 2009). However, these solutions often introduce their complexities and potential for overfitting, leading to concerns about the generalizability and reliability of the resulting models (Krawczyk, 2016).

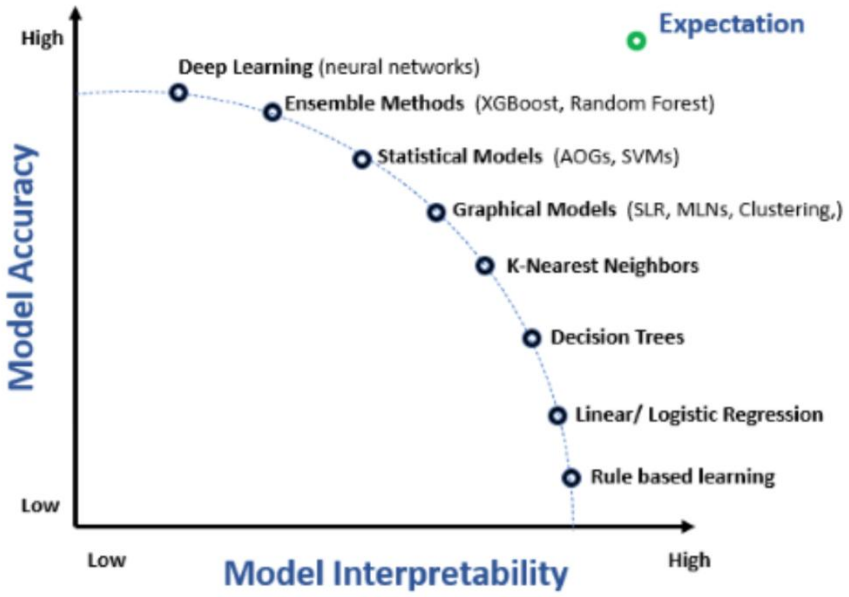

Figure 2: Model Explainability vs. Model Accuracy (Joshi, 2021)

Interpretability and Model-Agnostic Methods

Interpretability is a crucial aspect of machine learning, mainly when dealing with complex, "black box" models that are difficult to understand and explain (Doshi-Velez and Kim, 2017). In the context of class imbalance, interpretability becomes even more critical, as the biased performance of models on imbalanced datasets can lead to misleading or unreliable predictions (Loyola-González et al., 2016). Model-agnostic interpretation methods, such as Local Interpretable Model-Agnostic Explanations (LIME) (Ribeiro et al., 2016), SHapley Additive exPlanations (SHAP) (Lundberg and Lee, 2017), and TreeSHAP (Lundberg et al., 2020), have emerged as powerful tools for understanding the influence of predictor variables on model predictions, without requiring access to the model's internal workings.

Model-agnostic methods are instrumental in addressing class imbalance issues, enabling the identification of key features that drive model decisions, even when model performance is biased towards the majority class (Gao et al., 2022; Ochal et al., 2021). For example, LIME generates local explanations by approximating the behaviour of a black-box model around a specific instance, while SHAP and TreeSHAP provide global explanations by computing the average marginal contribution of each feature to the model's predictions (Lundberg and Lee, 2017; Lundberg et al., 2020). These methods can help practitioners understand how the model makes decisions and identify potential sources of bias or unreliability in the model's behaviour.

Research Gap

Despite the expanding research on XAI and its implementation across different fields, a considerable lack of insight persists regarding how class imbalance affects the consistency and dependability of model-agnostic interpretive approaches like TreeSHAP, LIME, and SHAP. Existing studies have primarily focused on improving the accuracy of predictive models in the presence of class imbalance (Haixiang et al., 2017; (Kaur et al., 2019) or evaluating the general performance of model-agnostic interpretation methods (Lundberg et al., 2020). Studies by Bussmann et al. (2020) hint at the complexity of class imbalances introduced to interpretation outputs, raising concerns over the potential for misinterpretation in practical applications. Furthermore, researchers have questioned the general robustness of interpretation methods against class imbalance and their performance with out-of-distribution samples. Researchers like Alvarez-Melis & Jaakkola (2018) have proposed Bayesian approaches as potential solutions to improve interpretation consistency and reliability in imbalanced contexts.

However, researchers have not thoroughly investigated how class imbalance affects the stability and consistency of feature importance explanations that TreeSHAP, LIME, and SHAP generate (Chen et al., 2024). A comprehensive evaluation comparing the effects of class imbalance on the stability and reliability of interpretations generated by these methods is conspicuously absent. This gap indicates a critical need for systematic research to evaluate and enhance the robustness and stability of interpretability methods in the presence of class imbalance, ensuring accurate and reliable model explanations that stakeholders can trust.

Objectives and Research Questions

By addressing these research questions, this study aims to fill the gap in understanding the impact of class imbalance on the stability and robustness of model-agnostic interpretation methods,

Research Question 1: How do different levels of class imbalance affect the stability and reliability of feature importance explanations generated by TreeSHAP, LIME, and SHAP across diverse benchmark datasets?

Research Question 2: To what extent can data-handling techniques, such as resampling and synthetic data generation, mitigate the impact of class imbalance on the stability and consistency of model-agnostic interpretation methods?

Research Question 3: How does the choice of a machine learning algorithm (e.g., XGBoost, CatBoost, LightGBM, and logistic regression) influence the robustness and interpretability of TreeSHAP, LIME, and SHAP explanations in class imbalance?

Research Question 4: What are the limitations of existing stability metrics, such as Sequential Rank Agreement (SRA) and Coefficient of Variation (CV), in assessing the reliability of feature importance explanations, and how can these metrics be improved or complemented to better capture the nuances of interpretation stability in imbalanced datasets?

Methodology

Benchmark Dataset

Each dataset provides valuable perspectives on using explainable artificial intelligence (XAI) to enhance model clarity and decision-making, primarily where class imbalances exist.

Banknote Authentication Dataset (Lohweg, 2019): This dataset is crucial for testing the ability of algorithms to identify genuine versus counterfeit banknotes using features obtained from wavelet-transformed image data.

Red Wine Quality Dataset (Cortez et al., 2009): This dataset includes detailed physicochemical properties essential for analyzing the performance of models that predict red wine quality.

Rice Dataset (Cinar and Koklu, 2019): Applied to classify rice varieties, this dataset tests models on morphological characteristics, underscoring significance in the agricultural sector.

Pima Indians Diabetes Dataset (Smith et al., 1988): Vital for healthcare predictive analytics, this dataset supports evaluations of model accuracy in predicting diabetes mellitus from diagnostic data.

Datasets	Features	Dataset size	Original Imbalance ratio
Banknote Authentication	4	1372	1: 1.22
Red Wine Quality	11	1599	1:8.63
Rice	7	18185	1:1.03
Pima Indians Diabetes	8	768	1:1.87

Table 1 Summary of Datasets

Research Methodology Structure

Data Preparation: Clean and analyze data.

Data Sampling: Randomly select 100 values for each class (major and minor)

Sampling Procedure: For the rest of the data, sample 10 dataset with same sample size but different imbalance class ratio from 1:2 to 1:1024 as training set.

Classifiers to Training and Predicting: Train Xgboost, Catboost, LightGBM and Logistic Regression (Class Weights or Threshold Adjustment) to training set then predict for each target obtained in data sampling.

Explainable AI (XAI) Interpretation: Generate TreeSHAP, SHAP, and LIME interpretation for prediction for each target obtained in data sampling.

Interpretation Stability Measurement: Measure ranking and stability on TreeSHAP, SHAP, and LIME.

Sampling Procedure

We follow a resampling procedure by avoiding synthetic sample generation to maintain data integrity. It involves:

1. **Initial Class Distribution Analysis**: Identifying minority and majority classes within each dataset.
   - **Wine Quality**: Let's consider high quality (quality > 6) as the minority class.
   - **Banknotes**: Assuming '0' (non-authentic) is a minority class.
   - **Rice**: Assuming: '1' is the minority class, the dataset is relatively balanced.
   - **Diabetes**: Assuming '1' (presence of diabetes) is the minority class.
2. **Resampling for Imbalance Creation**: Adjusting the dataset to create specified levels of imbalance, ranging from 1:2 to 1:1024, between minority and majority classes. This is achieved by under-sampling the majority class and over-sampling, including methods like SMOTE and Borderline SMOTE, the minority class, according to the target imbalance ratio.
3. **Ensuring Sample Size Consistency**: The total sample size is maintained across different imbalance levels to exclude the effect of sample size on model performance, mirroring approaches that fix sample size to achieve balanced dataset characteristics.

ML and XAI Models

Train tree-based models (XGBoost, CatBoost, LightGBM) and logistic regression are the primary predictive models due to their high performance in various machine-learning tasks, including those with imbalanced datasets. Then, the model optimizes hyperparameters and adjusts datasets for imbalance by using class weights or threshold adjustment.

- Evaluate model performance using cross-validation or separate test sets to ensure robustness and generalizability.
- Predict outcomes on tests obtained during initial sampling, focusing on accurately reflecting class distribution.
- Apply XAI techniques (TreeSHAP for tree-based models, SHAP and LIME for all models) to generate interpretation.

Analysis of Interpretation Stability

The core contribution of the study involves analyzing how the ranking and stability of interpretations provided by TreeSHAP, LIME and SHAP change with varying levels of class imbalance.

1. **Sequential Rank Agreement (SRA)** (Ekstrøm et al., 2019) is a methodological framework designed to quantify the concurrence among multiple features ranking lists, providing a robust measure of ranking stability.

The SRA computation commences by evaluating the ranking agreement for each feature across different lists. For a given feature (F_p), the ranking agreement value ($A(F_p)$) is determined using the formula:

$$A(F_p) = \frac{1}{L-1}\sum_{i=1}^{L}\left(R_i(F_p) - \overline{R(F_p)}\right)^2$$

Where $R_i(F_p)$: rank of the feature in the i^{th} List, and $\overline{R(F_p)}$: average rank of the feature across all L lists (as shown in Table 3).

Further, the SRA metric for each target X at every list depth d is calculated as the weighted expected agreement of the features within the unique set S_d, as denoted in Table 4. The expression formalizes this metric:

$$sra_d(X) = \frac{\sum_{p\in S_d}(L-1)A(X_p)}{(L-1)|S_d|}$$

Where $|S_d|$: cardinality of the set S_d. A lower SRA indicates greater consistency in feature rankings.

Ranking	L_1	L_2	L_3
1	A	A	B
2	B	C	C
3	C	B	A
4	D	D	E
5	E	E	D

Table 2

Feature	R_1	R_2	R_3
A	1	1	3
B	2	3	1
C	3	2	2
D	4	4	5
E	5	5	4

Table 3

Depth (d)	S_d
1	A, B
2	A, B, C
3	A, B, C
4	A, B, C, D, E
5	A, B, C, D, E

Table 4

2. **Coefficient of Variation (CV)** The CV measures variation relative to the mean; a higher CV indicates more fluctuation in feature importance across interpretations. The calculation is as follows:

For each feature, denoted by F_p, within our L different lists, we first determine the CV value $cv(F_p)$ using the following formula:

$$cv(F_p) = \frac{\sqrt{\frac{1}{L-1}\sum_{i=1}^{L}\left(V_i(F_p) - \bar{V}(F_p)\right)^2}}{\bar{V}(F_p)}, \bar{V}(F_p) = \frac{1}{L}\sum_{i=1}^{L} V_i(F_p).$$

Then, for each target outcome X, the composite CV value $cv(X)$ Over the entirety of L lists are calculated by computing the mean of the CV values for all features:

$$cv(X) = \frac{1}{P}\sum_{p=1}^{P} cv(F_p)$$

By integrating SRA and CV into our methodology, we aim to ensure that our predictive models' interpretability is robust and reliable, providing clear insights into the decision-making process.

Ethical Considerations

This research on the stability of model-agnostic interpretability methods in the presence of class imbalance must prioritize ethical considerations, including data privacy and security, bias and fairness, transparency and interpretability, responsible use and application, informed consent, and adherence to established ethical guidelines, to ensure the development of accountable and trustworthy AI systems.

References

1. Alvarez-Melis, D. and Jaakkola, Tommi S 2018. On the Robustness of Interpretability Methods. *arXiv.org*. [Online]. Available from: https://arxiv.org/abs/1806.08049.

2. Bussmann, N., Giudici, P., Marinelli, D. and Jochen Papenbrock 2020. Explainable Machine Learning in Credit Risk Management. *Computational Economics*. **57**(1), pp.203–216.

3. Chawla, N.V., Bowyer, K.W., Hall, L.O. and W. Philip Kegelmeyer 2002. SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*. **16**, pp.321–357.

4. Chen, Y., Calabrese, R. and Martin-Barragan, B. 2024. Interpretable machine learning for imbalanced credit scoring datasets. *European Journal of Operational Research*. **312**(1), pp.357–372.

5. Cinar and Koklu 2019. UCI Machine Learning Repository. *Uci.edu*. [Online]. Available from: https://archive.ics.uci.edu/dataset/545/rice+cammeo+and+osmancik.

6. Cortez et al. 2009. UCI Machine Learning Repository. *Uci.edu*. [Online]. [Accessed 10 March 2024]. Available from: https://archive.ics.uci.edu/dataset/186/wine+quality.

7. Doshi-Velez, F. and Kim, B. 2017. Towards A Rigorous Science of Interpretable Machine Learning. *arXiv.org*. [Online]. Available from: https://arxiv.org/abs/1702.08608.

8. Ekstrøm, Thomas Alexander Gerds and Andreas Kryger Jensen 2019. Sequential rank agreement methods for comparison of ranked lists. *Biostatistics*. **20**(4), pp.582–598.

9. European Commission 2018. GDPR and AI: Friends, foes or something in between? *Sas.com*. [Online]. [Accessed 25 March 2024]. Available from: https://www.sas.com/en_in/insights/articles/data-management/gdpr-and-ai--friends--foes-or-something-in-between-.html.

10. Gao, Y. 2022. Dealing with imbalanced data for interpretable defect prediction | Information and Software Technology. *Information and Software Technology*. [Online]. [Accessed 25 March 2024]. Available from: https://dl.acm.org/doi/abs/10.1016/j.infsof.2022.107016.

11. Gunning 2017. https://www.darpa.mil/ddm_gallery/xai-figure2.png. [Accessed 24 February 2024]. Available from: https://www.darpa.mil/ddm_gallery/xai-figure2.png.

12. Gunning 2024. https://www.darpa.mil/program/explainable-artificial-intelligence. *Darpa.mil*. [Online]. [Accessed 24 February 2024]. Available from: https://www.darpa.mil/program/explainable-artificial-intelligence.

13. Guo Haixiang, Yijing, L., Shang, J., Gu Mingyun, Huang Yuanyue and Bing, G. 2017. Learning from class-imbalanced data: Review of methods and applications. *Expert Systems with Applications*. **73**, pp.220–239.

14. He, H. and Garcia, E.A. 2009. Learning from Imbalanced Data. *IEEE Transactions on Knowledge and Data Engineering*. **21**(9), pp.1263–1284.

15. Joshi, K. 2021. Arya-XAI - A distinctive approach to explainable AI. *Arya AI Blog*. [Online]. [Accessed 15 March 2024]. Available from: https://blog.arya.ai/arya-xai-a-distinctive-approach-to-explainable-ai/.

16. Kaur, H., Pannu, H.S. and Malhi, A.K. 2019. A Systematic Review on Imbalanced Data Challenges in Machine Learning. *ACM Computing Surveys*. **52**(4), pp.1–36.

17. Krawczyk, B. 2016. Learning from imbalanced data: open challenges and future directions. *Progress in Artificial Intelligence*. **5**(4), pp.221–232.

18. Lohweg, V. 2019. UCI Machine Learning Repository. *Uci.edu*. [Online]. [Accessed 10 March 2024]. Available from: https://archive.ics.uci.edu/dataset/267/banknote+authentication.

19. Loyola-González, O., José Fco. Martínez-Trinidad, Jesús Ariel Carrasco-Ochoa and García-Borroto, M. 2016. Study the impact of resampling methods for contrast pattern-based classifiers in imbalanced databases. *Neurocomputing*. **175**, pp.935–947.

20. Lundberg, S.M., Erion, G., Chen, H., DeGrave, A., Prutkin, J.M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N. and Lee, S.-I. 2020. From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence*. **2**(1), pp.56–67.

21. Lundberg, S.M. and Lee, S.-I. 2017. A Unified Approach to Interpreting Model Predictions. *Advances in Neural Information Processing Systems*. **30**.

22. Ochal, M., Patacchiola, M., Storkey, A., Vazquez, J. and Wang, S. 2021. Few-Shot Learning with Class Imbalance. *arXiv.org*. [Online]. [Accessed 15 March 2024]. Available from: https://arxiv.org/abs/2101.02523.

23. Ribeiro, M.T., Singh, S. and Guestrin, C. 2016a. 'Why Should I Trust You?': Explaining the Predictions of Any Classifier. *arXiv.org*. [Online]. [Accessed 14 March 2024]. Available from: https://arxiv.org/abs/1602.04938.

24. Ribeiro, M.T., Singh, S. and Guestrin, C. 2016b. 'Why Should I Trust You?': Explaining the Predictions of Any Classifier. *arXiv.org*. [Online]. [Accessed 15 March 2024]. Available from: https://arxiv.org/abs/1602.04938.

25. Smith et al. 1988. Pima Indians Diabetes. *Data. World*. [Online]. [Accessed 10 March 2024]. Available from: https://data.world/uci/pima-indians-diabetes.