NavaJyoti, International Journal of Multi-Disciplinary Research

Volume 9, Issue 1, June 2025

Large Language Models and Interpretable Models in Data Science

Priyanka R[1], Maria John[2]

[1]Student, Department of Computer Science(PG), Jyoti Nivas College Autonomous, Bangalore, India
[2] Assistant Professor, Department of Computer Science(PG), Jyoti Nivas College Autonomous, Bangalore, India

Abstract

The rapid advancement of Large Language Models (LLMs) has opened new avenues in enhancing the interpretability of machine learning models. This study explores the synergy between LLMs and Generalized Additive Models (GAMs), demonstrating how LLMs can interpret, describe, and debug GAMs, thereby expanding the scope of explainable AI (XAI). By leveraging GAMs' ability to generate transparent, human-readable model representations, LLMs provide insightful explanations, anomaly detection, and data summarization. Experimental results show that GPT-4 significantly outperforms GPT-3.5 in interpreting model graphs, identifying statistical anomalies, and summarizing entire models based on their graphical outputs. However, the study also highlights the potential risk of hallucinations in LLM explanations and discusses methods to validate their outputs. Integrating LLMs with interpretable models enhances data scientists' ability to extract actionable insights while fostering trust and transparency in AI-driven decision-making. This research proposes an open-source LLM-GAM interface that enables domain experts to interact with machine learning models using natural language, improving accessibility and interpretability.

Keywords:

Large Language Models (LLMs), Generalized Additive Models (GAMs), Explainable AI (XAI), Model Interpretability, Data Science, Anomaly Detection, Human-AI Collaboration

1. Introduction

The use of machine learning in criminal justice has raised concerns about fairness, transparency, and bias in decision-making. The COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) dataset is widely used to predict recidivism risk but has been criticised for racial and gender bias. This study compares GAMs, Random Forest, and XGBoost to determine which model balances accuracy and interpretability for ethical AI use in law enforcement.

This paper responds to the following questions:

1. What model yields the best prediction?
2. How interpretable are the models?
3. What are the most influential features regarding recidivism risk?

2. Literature Review

Singh et al. (2022), in "Augmenting Interpretable Models with LLMs during Training," introduce Augmented GAMs (Aug-GAMs), which use LLMs during training to generate explanations and improve feature selection while keeping the model transparent. They report

gains in interpretability and efficiency over standard GAMs, but note that dependence on LLM outputs can introduce variability.

Singh et al. (2024), in "Rethinking Interpretability in the Era of Large Language Models," argue that natural language and interactive explanations offer a more intuitive alternative to static methods such as SHAP values and feature importance scores. They also caution that LLMs are opaque and computationally costly, making it hard to verify that their explanations are faithful.

Luo and Specia (2024), in "From Induction to Unification: A Survey on Explainability for Large Language Models," categorise explainability methods into local and global techniques and discuss mechanistic interpretability and model editing. They identify high computational cost and the difficulty of validating LLM-generated explanations as key barriers.

Lengerich et al. (2023), in "LLMs Understand Glass-Box Models, Discover Surprises, and Suggest Repairs," show that LLMs can read GAM outputs, detect anomalies and biases, and explain them in human-readable form. They also warn of over-reliance on LLM interpretations and the need to validate suggested repairs.

Baayen and Linke (2022), in "An Introduction to the Generalized Additive Model," explain how GAMs extend generalized linear models with smooth functions, capturing non-linear relationships while remaining interpretable and avoiding restrictive parametric assumptions.

Zhuang et al. (2021), in "Interpretable Ranking with Generalized Additive Models," apply GAMs to ranking tasks and show competitive performance against gradient-boosted trees and deep ranking models, with the added benefit of transparent feature effects.

Overall, the literature shows that GAMs offer a strong balance between accuracy and interpretability, and that LLMs can make such models easier to explain and debug, although the reliability and cost of LLM-generated explanations remain open challenges.

3. Methodology

3.1 Dataset Description

The COMPAS dataset has features such as age, prior convictions, charge type, race, and sex. The target variable two_year_recid signifies whether someone recidivates within two years.

3.2 Preprocessing Steps

Management of categorical variables: One-hot encoding for columns such as sex and charge degree.

a. Feature selection: Selecting appropriate variables like age, previous convictions, type of offense, and demographic.

b. Standardization: Scaling numerical features to ensure model training consistency.

c. Train-test split: 80% for training, 20% for testing to check model performance.

3.3 Models Compared

Generalized Additive Model (GAM)

GAMs are explainable machine learning models that capture non-linear relationships among features and outcomes.

Random Forest (RF)

RF is an ensemble learning technique that constructs multiple decision trees

to enhance accuracy and reliability.
XGBoost (XGB)
XGBoost is an efficient high-accuracy gradient boosting
algorithm commonly used for predictive modeling.

3.4 Model Evaluation Metrics

• Accuracy → Overall model correctness.
• AUC-ROC Score → Quantifies the ability of the model to separate recidivists from non-recidivists.
• Feature Importance (SHAP Values) → Determines features having the largest impact on predictions.

4. Results and Discussion

4.1 Model Performance Comparison

Model	**Accuracy (%)**	AUC-ROC Score	Interpretability
GAM	73.5%	0.72	High (Transparent)
Random Forest	79.2%	0.78	Low (Black-box)
XGBoost	83.1%	0.82	Low (Black-box)

1. XGBoost has the best AUC-ROC value, i.e., it is more accurate in predicting recidivism.
2. GAM is the most interpretable model, with direct human comprehension of relationships.

4.2 Feature Importance Analysis (SHAP Values)

SHAP analysis was performed to interpret feature influence in XGBoost.
Top Features Responsible for Recidivism Predictions:
1. Priors Count (Most impactful)
2. Age (Younger people have greater risk)
3. Charge Degree (Felony or Misdemeanor)
4. Race (Potential bias identified)
SHAP Summary Plot indicates that priors count has the largest effect on predictions.

4.3 Ethical Aspects and Explainability

Although XGBoost works best, it is not transparent.
GAMs, however, yield detailed insights into the influence of features on predictions. This is imperative when making judgments in court where explainability matters as much as accuracy.

5. Conclusion

This work contrasted GAM, Random Forest, and XGBoost on the COMPAS dataset to predict the risk of recidivism. XGBoost was most accurate in prediction, but GAMs were more interpretable.This work emphasizes the value of trading accuracy for fair ness in applying AI to the criminal justice system. We used GAM for transparency & fairness,

XGBoost for accurate predictions, and SHAP to explain sophisticated models. Future work would involve investigating bias-reducing methods and ensemble models that marry the interpretability of GAM with the high performance of XGBoost.

References

[1] Singh, Chandan, Armin Askari, Rich Caruana, and Jianfeng Gao. "Augmenting interpretable models with large language models during training." *Nature Communications* 14, no. 1 (2023): 7913.

[2] Singh, Chandan, Jeevana Priya Inala, Michel Galley, Rich Caruana, and Jianfeng Gao. "Rethinking interpretability in the era of large language models." *arXiv preprint arXiv:2402.01761* (2024).

[3] Luo, H., & Specia, L. (2024). From understanding to utilization: A survey on explainability for large language models. *arXiv preprint arXiv:2401.12874*.

[4] Lengerich, Benjamin J., Sebastian Bordt, Harsha Nori, Mark E. Nunnally, Yin Aphinyanaphongs, Manolis Kellis, and Rich Caruana. "LLMs understand glass-box models, discover surprises, and suggest repairs." *arXiv preprint arXiv:2308.01157* (2023).

[5] Baayen, R. Harald, and Maja Linke. "An introduction to the generalized additive model." *A practical handbook of corpus linguistics* 563 (2020): 591.

[6] Zhuang, Honglei, Xuanhui Wang, Michael Bendersky, Alexander Grushetsky, Yonghui Wu, Petr Mitrichev, Ethan Sterling, Nathan Bell, Walker Ravina, and Hai Qian. "Interpretable ranking with generalized additive models." In *Proceedings of the 14th ACM international conference on web search and data mining*, pp. 499-507. 2021.

[7] Caruana, R., et al. "Intelligible Models for Healthcare: Predicting Pneumonia Risk and Hospital Readmission." KDD, 2015.

[8] Lundberg, S. M., & Lee, S.-I. "A Unified Approach to Interpreting Model Predictions." NeurIPS, 2017.

[9] Rudin, C. "Stop Explaining Black Box ML Models and Use Interpretable Models Instead." Nature Machine Intelligence, 2019.