LET'S STOP USING BLACK-BOX MODELS TO PREDICT STUDENT SUCCESS

A comparison between black box models and explainable models to Predict Student Success (PreSS) on introductory programming courses (CS1)

AUTHORS
Jan Glazenborg

AFFILIATIONS
Hogeschool Utrecht, TU Dublin

ABSTRACT

Predicting student success in introductory programming (CS1) could aid with early intervention for at-risk students. The PreSS model, using Naïve Bayes, has achieved ~70% accuracy, but recent attempts with black-box models like deep learning and LLMs have struggled with interpretability. Transparency is essential in educational AI, as teachers need to understand the factors influencing student outcomes. Explainable Boosting Machines (EBMs) have shown potential to match black-box accuracy while remaining interpretable. This study evaluates EBMs and other interpretable models using the PreSS dataset (692 students) to compare accuracy, feature importance, and explainability. This research aims to balance performance with trustworthiness. The findings could inform future AI-driven student success prediction models that are both effective and understandable.

INTRODUCTION

Predicting student success has been extensively researched, with AI models offering potential for early intervention [8]. The PreSS model, developed at TU Dublin, uses Naïve Bayes to predict CS1 performance with ~70% accuracy [2]. Recent attempts to improve accuracy with black-box models like LLMs and deep learning have resulted in lower transparency, making it difficult for educators to interpret predictions. Interpretability is crucial in education, where understanding contributing factors aids decision-making. Rudin argues for using inherently explainable models over black-box approaches [4]. Models like Explainable Boosting Machines (EBMs) have matched black-box accuracy while remaining transparent [6]. This study implements EBMs and other interpretable models on the PreSS dataset to assess their accuracy and explainability compared to black-box models.

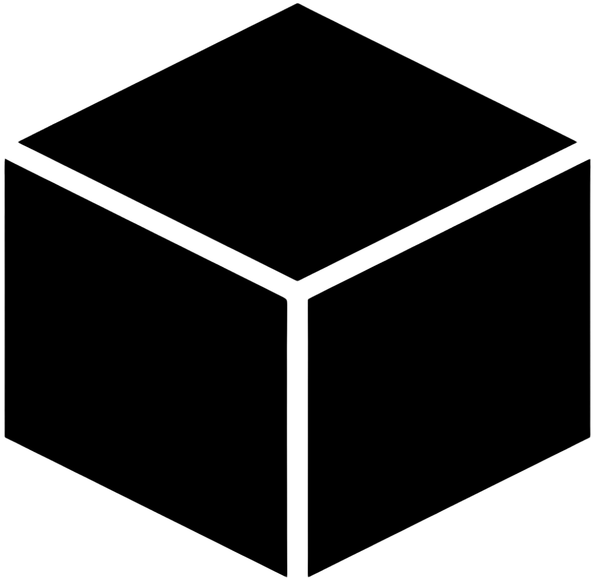
LITERATURE REVIEW

Predicting student success in CS1 has been widely studied, with early models like PreSS relying on Naïve Bayes to classify at-risk students [1] , [2]. More recent work has explored deep learning and LLMs, but these approaches have struggled with interpretability and have not significantly outperformed traditional models [3]. Interpretability is essential in educational AI, as black- box models provide little insight into the factors influencing student outcomes. Rudin argues that inherently explainable models should replace post-hoc explanations of black-box predictions [4]. In other domains, Explainable Boosting Machines (EBMs) have shown promise, performing comparably to deep learning models while maintaining transparency [6]. Efforts to apply XAI in CS1 have begun, but existing studies primarily focus on theoretical frameworks rather than practical implementation [5].

METHODOLOGY

This study evaluates the performance of traditional explainable ML models, following a structured approach to compare accuracy, interpretability, and feature importance against existing PreSS implementations. The study utilizes the PreSS dataset, which contains data from 692 students across 11 institutions in Ireland and Denmark [2]. The dataset includes student background information, academic performance, and psychological factors collected early in the CS1 course. Standard preprocessing techniques, such as handling missing values and normalizing numerical features, will be applied.

To address the research question on effective XAI for educators, a user study or survey may be conducted to assess whether teachers find interpretable models more trustworthy than black-box predictions. This study aims to determine whether EBMs can provide both high accuracy and transparency, ensuring AI-driven student success prediction is both effective and understandable.



PRELIMINARY CONSIDERATIONS

This study examines whether explainable models can match black-box accuracy while improving transparency in CaS1 prediction. Since educational AI must be interpretable, models like EBMs may offer a better alternative [4]. Ethical considerations include data privacy and potential bias in predictions [7]. Ensuring that models support, rather than mislead, educators is key. The findings could guide future AI tools for student success prediction, balancing accuracy with trustworthiness.

CONCLUSION

This study explores the balance between predictive accuracy and interpretability in student success modeling. While black-box models like deep learning have been investigated, their lack of transparency limits their usefulness in education. Explainable models offer a promising alternative, potentially maintaining high accuracy while providing insights into key success factors. By evaluating these models on the PreSS dataset, this research aims to determine whether interpretable approaches can replace or complement existing methods. The study also considers ethical implications, ensuring that predictions remain fair and trustworthy. If successful, this work could contribute to the development of AI-driven tools that help educators intervene early with at-risk students, providing a transparent and reliable approach to student success prediction.

ACKNOWLEDGEMENTS

The HCAIM (the Human-Centred AI Master's Programme) Project is Co-Financed by the Connecting Europe Facility of the European Union Under Grant NoCEF-TC-2020-1 Digital Skills 2020-EU-IA-0068. This poster was created as part of the Blended Intensive Programme organised under the Erasmus + Programme of the European Union.

REFERENCES

[1] S. Bergin, "A computational model to predict programming performance," Ph.D. dissertation, Maynooth University, 2006. [2] K. Quille and S. Bergin, "CS1: how will they do? How can we help? A decade of research and practice," Computer Science Education, vol. 29, no. 2–3, pp. 254–282, 2019 [3] P. Riello, K. Quille, R. Jaiswal, and C. Sansone, "Reimagining Student Success Prediction: Applying LLMs in Educational AI with XAI," [4] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," Nature Machine Intelligence, 2019. [5] K. Quille, L. Vidal-Meliá, K. Nolan, and A. Mooney, "Evolving Towards a Trustworthy AIEd Model to Predict at Risk Students in Introductory Programming Courses," Association for Computing Machinery, 2023 [6] V. Dsilva, J. Schleiss, and S. Stober, "Trustworthy Academic Risk Prediction with Explainable Boosting Machines," in Artificial Intelligence in Education, 2023 [7] K. Quille and S. Bergin, "Programming: predicting student success early in CS1. A re-validation and replication study," 2018, [8] R. Alamri and B. Alharbi, "Explainable Student Performance Prediction Models: A Systematic Review," 2021, doi: 10.1109/ACCESS.2021.3061368.