CSCI 381 Final Proposal

Sebastian Mejia, Mauricio Monje, Aditya Dwivedi

June 2025

Research Question

For our final report, we want to ask and find the following: "How do SHAP and LIME compare in explaining individual predictions of different classifiers (logistic regression and random forest) on real-world tabular classification datasets, particularly in terms of feature importance rankings, consistency, interpretability, and runtime performance?"

The reason as to why we find the research for this field important is because we want to provide clear explanations as to why some ML models are providing predictions. Obviously, we understand the process and idea behind the ML, but we do lack the interpretability which does limit its trust and adoption into other fields where it could find it useful. SHAP and LIME tries to help bridge this gap by providing insight into why a model made a certain decision. Understanding how these tools perform allows others to choose the right method for their specific task, ensuring a much more transparent and responsible use for MLs.

As for the timely portion, we find that this is the best time to learn about it because of how there are increasing regulations as well as demand for MLs. Simply researching how common classifiers like logistic regression and random forests with different and new datasets helps update and deepen our understanding of how these classifiers came to their results.

Research Papers

- **"Evaluating the Correctness of Explainable AI Algorithms":** This paper introduces a synthetic dataset framework with ground-truth explanations to quantitatively evaluate the accuracy of XAI methods like SHAP and LIME, showing SHAP performs more reliably.
- **A Perspective on Explainable Artificial Intelligence Methods: SHAP and LIME:** This paper offers a comparative analysis of SHAP and LIME in tabular data contexts, highlighting their theoretical differences, practical limitations (e.g., collinearity issues), and suggesting improvements for more robust explanations.

- **Fooling LIME and SHAP: Adversarial Attacks on Post hoc Explanation Methods:** This study reveals how SHAP and LIME can be deceived by adversarial models, allowing biased classifiers to produce misleadingly fair explanations, especially in sensitive domains like credit scoring and criminal justice.
- **Provably Stable Rankings with SHAP and LIME:** Proposes statistical methods to stabilize the top feature rankings from SHAP and LIME, ensuring reliable and reproducible explanations through adaptive sampling and hypothesis testing.
- **ExplainBench: A Benchmark Framework for Local Model Explanations in Fairness-Critical Applications:** This work presents a comprehensive benchmarking toolkit to evaluate and compare local explanation methods (like SHAP and LIME) in fairness-critical settings, focusing on fidelity, sparsity, and stability of explanations.

Proposed Plan

1. Choose two real-world tabular classification datasets from a variety of sections (e.g., healthcare, education, finance) and making sure they have both numerical and categorical features for generalization.
2. Perform thorough EDA to understand feature distributions, correlations, handle missing or imbalanced data to demonstrate the type of data we're working with.
3. We want to perform preprocessing on the data through standardizing numerical features and one-hot encode categorical variables. The lasts part would be to split the data into training and test sets.
4. Model Training would occur with Logistic Regression and Random Forest with the preprocessed data.
5. Apply XAI Methods (SHAP and LIME) where SHAP gets the global and local feature importances and LIME to generate local explanations. After we get their results, we just create an evaluation and explanation on it.
6. Discussion and Conclusion regarding our results and all the processes taken.

Datasets

- **Heart Disease Dataset:** This heart disease dataset contains 2,877 patient records with features like age, sex, chest pain type, blood pressure, cholesterol levels, and exercise-related metrics. The target variable, HeartDisease, indicates the presence or absence of heart disease, making it suitable for binary classification tasks.

- **Student Alchohol Consumption:** This student alcohol consumption dataset contains detailed social, academic, and lifestyle information on secondary school students, including variables related to family background, study habits, and alcohol use. The main objective is to analyze patterns or predict students' final grades (G3) using factors like parental education, study time, social behaviors, and weekday/weekend alcohol consumption.
- **Titanic Passengers Dataset:** This Titanic dataset includes passengers with features such as age, sex, passenger class, fare, and family relationships, aimed at predicting survival outcomes (Survived) from the 1912 shipwreck. It provides both numerical and categorical data, making it a good dataset for binary classification and feature importance analysis in machine learning.

References

1 D. Slack, S. Hilgard, E. Jia, S. A. Singh, and H. Lakkaraju, "Evaluating the correctness of explainable ai algorithms," arXiv preprint arXiv:2105.09740, 2021. [Online]. Available: https://arxiv.org/abs/2105.09740

2 Y. Sun, S. Basak, and S. Roy, "A perspective on explainable artificial intelligence methods: Shap and lime," arXiv preprint arXiv:2305.02012, 2023. [Online]. Available: https://arxiv.org/abs/2305.02012

3 D. Slack, S. Hilgard, E. Jia, S. A. Singh, and H. Lakkaraju, "Fooling lime and shap: Adversarial attacks on post hoc explanation methods," arXiv preprint arXiv:1911.02508, 2020. [Online]. Available: https://arxiv.org/abs/1911.02508

4 W. Goldwasser and S. Hooker, "Provably stable rankings with shap and lime," arXiv preprint arXiv:2401.15800, 2024. [Online]. Available: https://arxiv.org/abs/2401.15800

5 B. McKinney, L. Cohen, A.-H. Karimi, A. Weller, and C. Jung, "Explainbench: A benchmark framework for local model explanations in fairness-critical applications," arXiv preprint arXiv:2506.06330, 2025. [Online]. Available: https://arxiv.org/abs/2506.06330