ECE225A Final Project Proposal
Overview of Tree-based Machine Learning Model

Enting Zhou *PID: A16159365*

November 2023

Introduction

Predictive modeling is a vital aspect of data analysis, machine learning, and artificial intelligence. Among the most popular techniques for predictive modeling are tree-based algorithms: decision trees, random forests, and gradient boosting trees. These models have demonstrated great performance in various scenario, outperforming other classification methods like Logistic Regression, Support Vector Machine, and Deep Neural Networks. This research project, in the review of major tree-based machine learning models, aims to understand and give insights into why tree-based models works well, as well as their limitations.

The project aims to review the following listed papers in the reference sessions, covering decision trees, random forest, and gradient boosted trees. This project will also conduct an empirical experiment on comparison on common decision trees implementations, as in those in scikit-learn[1] on a benchmark dataset credit score classification[2]. The goal of classification is to determine if a loan should be granted given the loaner's basic information and credit history. This benchmark dataset is suggested by Grinsztajn et al.[1], as the dataset have heterogeneous columns, I.I.D data, and non-deterministic data, attributes that similar to real-world applications.

Project Objective

This research project aims to achieve the following:

- To give overview on existing work on decision trees, and subsequent improvements on basic decision trees using ensemble learning methods.
- To assess the accuracy and predictive power of decision trees, random forests, and gradient boosting trees, through direct visualization of classification results on the selected benchmark dataset.

[1] https://scikit-learn.org/stable/index.html
[2] https://www.openml.org/search?type=data&status=active&id=45024

- To identify the strengths and weaknesses of each technique in terms of interpretability, handling of missing data, and resistance to overfitting.
- To explore the computational efficiency of these methods and their scalability with respect to dataset size.
- To provide recommendations and guidelines for choosing the most suitable technique for different predictive modeling tasks.

References

[1] L. Grinsztajn, E. Oyallon, and G. Varoquaux, "Why do tree-based models still outperform deep learning on typical tabular data?," in *Advances in Neural Information Processing Systems* (S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, eds.), vol. 35, pp. 507–520, Curran Associates, Inc., 2022.

[2] P. Geurts, D. Ernst, and L. Wehenkel, "Extremely randomized trees," *Mach. Learn.*, vol. 63, pp. 3–42, Apr. 2006.

[3] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, pp. 5–32, Oct. 2001.

[4] "Lightgbm: A highly efficient gradient boosting decision tree." https://proceedings.neurips.cc/paper/2017/hash/6449f44a102fde848669bdd9eb6b76fa-Abstract.html. Accessed: 2023-11-18.

[5] A. J. Myles, R. N. Feudale, Y. Liu, N. A. Woody, and S. D. Brown, "An introduction to decision tree modeling," *J. Chemom.*, vol. 18, pp. 275–285, June 2004.

[6] Y.-Y. Song and Y. Lu, "Decision tree methods: applications for classification and prediction," *Shanghai Arch Psychiatry*, vol. 27, pp. 130–135, Apr. 2015.

[7] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," *Ann. Stat.*, vol. 29, no. 5, pp. 1189–1232, 2001.

[8] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, (New York, NY, USA), pp. 785–794, Association for Computing Machinery, Aug. 2016.