# PHSX 499: Selected Paper Outline

William Jardee

March 3, 2022

This is an analysis of the work published in 2020 by Thibaut Vidal and Maximilian Schiffer, **Born-Again Tree Ensembles**. I go through section by section and try to give a one sentence synopsis of each paragraph. The paper can be found at citation [1].

## Abstract

¶ 1: Decision Trees and their ensemble counterparts, Decision Forests, are very important in modern applications. Thus, the authors study "born-again tree ensembles": "the process of constructing a single decision tree of minimum size that reproduces the exact same behavior as a given tree ensemble in its entire feature space", by a dynamic-program that uses pruning and bounding rules to generate a minimum decision tree from the ensemble [1].

## 1 Introduction

¶ 1: Types of Decision Tree are important to get right and casting into explainable models because they are used in a wide area of subjects.
¶ 2: Currently there is a choice between interpretability and performance, their model addresses this issue
¶ 3: Here they define their notation and define problem 1:
"**Problem 1 (Born-again tree ensemble)**: Given a tree ensemble $\mathcal{T}$, we search for a decision tree T of **minimal size** that is **faithful** to $\mathcal{T}$, i.e., such that $F_T(x) = F_{\mathcal{T}}(x)$ for all $x \in \mathbb{R}^p$."
¶ 4: They restate their problem as looking for a "new representation of the same classifier".
¶ 5: They propose that this problem is NP-hard (when optimizing depth or number of leaves) by suggesting a proof of reduction to the 3-SAT problem.[1]

### 1.1 State of the Art

¶ 1: They point towards other works that should be studied for a more complete understanding.

#### Thinning tree ensembles

¶ 2: Introduce the two perspectives in ensemble thinning.
¶ 3: The approach of ensemble thinning has been very successful, both in the case of doing static and dynamic algorithms.
¶ 4: The other "born-again" models of simplifying an ensemble to one tree are brought up and they end with the flaw that the models "do not guarantee faithfulness" but focus on trees that "remain

[1] These are all concepts that anyone who has studied computer science theory should understand, so they don't, and don't need to, go into elaboration about any of these concepts.

interpretable."
¶ 5: The authors then explain the current state of neural network studies into "model compression"

**decision trees**

¶ 6: The authors bring up the research of optimal decision tree generation, not from an ensemble but of their our right.

**Summary**

¶ 7: Recent work has been focused on creating an optimal tree from an ensemble rather than an explainable model, that is going to be the purpose of this paper

### 1.2 Contributions

¶ 1: Their aim of this work is: 1) formally define the problem (and prove it is NP-hard), 2) highlight the important characteristics of the problem that can be exploited, 3) design pruning strategies, 4) do numerical studies to analyze their algorithm.
¶ 2: We have succeeded in solving the problem.

## 2 Fundamentals

¶ 1: The authors define more essential terms.
¶ 2: Defining what a Tree Ensemble is.
¶ 3: Defining what a Cell is.
¶ 4: Defining what a Region is.
¶ 5: Stating the mathematically rigorous definitions.
¶ 6: Discussing the depth restriction defined on an arbitrary hyperplane cut of the Tree Ensemble.

## 3 Methodology

¶ 1: The authors presents the primary algorithm that allows the problem to be solved dynamically.
¶ 2: Two primary weaknesses of the algorithm are presented
¶ 3: A solution to the first weakness is addressed and a theorem of dimensionality is presented.
¶ 4: A solution to the second weakness is addressed.
¶ 5: The authors prepare the reader for the presentation of the algorithm structure. The remaining paragraphs of the section are on that topic
¶ 6: Check whether a leaf/branch should be added to the DP memory.
¶ 7: The search area is reduced and then the best covering of the space is computed (according to theorems and equations already brought up).
¶ 8: A point about simplifying memory consumption is made.
¶ 9: The time complexity of the algorithm is addressed.
¶ 10: This is more of a figure, but they present the pseudo-code of the algorithm.

## 4 computational Experiments

¶ 1: The computational experiments are laid out: 1) Evaluate the performance of the algorithm on a variety of criteria, 2) Study the complexity, 3) Measure the impact of a simple pruning strategy, 4) evaluate a fast heuristic algorithm. The language, computer specs, and github are then given.

### 4.1 Data Preparation

¶ 1: The datasets used are outlined and justified.

### 4.2 Computational Effort

¶ 1: "In a first analysis, [they] evaluate the computational time of Algorithm 1 (their algorithm) for different data sets and size metrics."
¶ 2: "In [their] second analysis, [they] focus on the FICO case and randomly extract subsets of sample and features to produce smaller data sets."
¶ 3: They explain the results (they are in line with what was expected).

### 4.3 Complexity of the Born-Again Trees

¶ 1: "We now analyze the depth and number of leaves of the born-again trees for different objective function and datasets in Table 2." That's the whole paragraph...
¶ 2: The results of each of their heuristics are explained.

### 4.4 Post-Pruned Born-Again Trees

¶ 1: The need for pruning is motivated.
¶ 2: "Unmotivated" leaves are pruned.
¶ 3: They claim that this pruning worked to improve simplicity and interpretability.
¶ 4: The impact of pruning of effectiveness is analyzed.
¶ 5: Pruning had little impact on performance.

### 4.5 Heuristic bone-Again Trees

¶ 1: The authors motivate a need for a heuristic of deciding tree size.
¶ 2: The heuristic for when to split into more branches is explained.
¶ 3: They claim that their heuristic worked well at decreasing computation time.

## 5 Conclusions

¶ 1: A summary of the paper is provided.
¶ 2: The authors present the future work on the project.

## References

[1] Thibaut Vidal and Maximilian Schiffer. Born-again tree ensembles. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 9743–9753. PMLR, 13–18 Jul 2020.