Using Rashomon Sets for Robust Active Learning

Simon Dovan Nguyen
Department of Statistics
University of Washington
Seattle, WA 98101
simondn@uw.edu

Kentaro Hoffman
Department of Statistics
University of Washington
Seattle, WA 98101
khoffm3@uw.edu

Tyler H. McCormick
Department of Statistics
University of Washington
Seattle, WA 98101
tylermc@uw.edu

Abstract

Active learning is based on selecting informative data points to enhance model pre-
dictions, often using uncertainty as a selection criterion. However, when ensemble
models such as random forests are used, there is a risk of the ensemble containing
models with poor predictive accuracy or duplicates with the same interpretation.
To address this, we develop a novel approach to only ensemble only the set of
near-optimal models called the Rashomon set in order to guide the active learning
process. We demonstrate how taking a Rashomon approach can improve not only
the accuracy and rate of convergence of the active learning procedure, but can also
lead to improved interpretability compared to traditional approaches.

1 Introduction

Collecting labeled data to train data-hungry modern artificial intelligence (AI) and machine learn-
ing (ML) models can be expensive or time-consuming. This challenge arises in a wide range of
applications: sentence classification [17], image labelling [23] [10], and verbal autopsy [4] . In
such scenarios, strategically determining which observations merit labeling will greatly reduce data
redundancy and improve the learning of covariate-label relationships.

To address time and budget constraints, active learning allows researchers the freedom to adaptively
choose which observations to label. The key task in active learning is choosing the most informative
observations that will enhance the predictive quality of the model when labelled. Amongst the many
metrics of informativity [12] [5] [15], uncertainty is the most commonly employed [13].

Due to their ease in measuring uncertainty, ensemble techniques such as random forests are a popular
model used for active learning [14]. Since the weak learners of ensemble methods are independent
by design, the individual base learners naturally form a committee. When used in active learning, the
disagreement in "votes" between the ensemble committee members is often used as a measure of
uncertainty and informativity [11] [3].

While ensemble methods offer a natural way to quantify uncertainty through the random diversity
of their weak learners, this diversity comes with a potential drawback. Specifically, most ensemble
methods tend to aggregate over the space of all models, even if some of the models may have
relatively poor accuracy. While some approaches such as Bayesian model averaging account for this
by weighting the models by how likely they are given the data, having a large number of mediocre
models are known to make such weighting approaches difficult, especially in cases of limited, noisy,
or high-dimensional data [6]. Aggregating such poor and implausible models compromises the
query-selection criteria, potentially leading to a suboptimal query in the active learning process.

To address this limitation, we propose a novel approach to improving the quality of ensemble methods
used in active learning. Specifically, we propose an algorithm that enhances active learning with
random forests by restricting aggregation to a subset of well-performing, high-evidence models
known as the Rashomon set. The Rashomon set consists of near-optimal models that have strong

support from the observed data. By ensembling only models within the Rashomon set, our approach ensures that the active learning process is driven by models with high evidence, leading to better query-selection criteria and improved query-selection.

The main results illustrating the benefits of aggregating across the Rashomon set in ensemble learning can be seen in Figure 2, in which the TreeFarms approach (blue and orange lines, whose distinction will become clear later) consistently outperforms the traditional random forest approach. We also demonstrate that one can further restrict the Rashomon set to only select models with similar "explanations" while still preserving the performance of the active learning process. This Rashomon-based method ensures that the ensemble incorporates the interpretability of the weak learners in our ensemble while maintaining prediction accuracy.

2   Rashomon Sets

When constructing machine learning models, researchers face two district types of uncertainty. The first originates from the variability in the predicted outcomes generated by a given model, often referred to as a model's intrinsic risk. The second originates from selecting the right model from a vast and diverse hypothesis space, a phenomenon known as *model ambiguity*. This distinction, originally articulated by economist Kenneth Arrow in 1951, separates the uncertainty of prediction from given a model from the uncertainty of choosing among many plausible models [1]. In his 2013 Nobel lecture, Lars Hansen further highlights this idea by characterizing the distinction as "uncertainty outside and inside [economic] models" [7].

Current machine learning approaches have become exceedingly good at reducing the predictive uncertainty within a model, but often fail to fully account for model ambiguity. Methods such as LASSO search for a single optimal model while ensemble methods such as Bayesian Model Averaging [9] sample across the full hypothesis space. However, both approaches overlook model ambiguity and are ambivalent to how many models with similar predictive power exist for a given dataset [16]. This oversight is underscored by the Rashomon Effect, first noted by Leo Breiman in 2001 [2]. The Rashomon Effect highlights the existence of near-optimal models that have similarly high predictive performance, but explain the data in different, potentially conflicting ways. Rudin furthers this idea by noting that this phenomenon exposes a core issue in the current machine learning paradigm: a reliance on a single predictive model that is overly-sensitive [16] [18]. This reliance fails to notice the complexity of modeling heterogeneity, where different models can explain the data nearly equally well but offer substantively different insights.

To quantify the number of near-equivalent models exist for a given dataset, one can employ techniques from Rashomon Theory. Rashomon Theory is focused on the Rashomon sets — a collection of models that are all near-optimal in terms of predictive accuracy. By enumerating the Rashomon set, researchers can explore the full range of plausible explanations supported by the data. Traditional ensemble methods on the other hand such as random forests aggregate base learners based on the random sampling of features and data. This leads to diverse but potentially suboptimal ensembles due to the inclusion of implausible models with no way of removing such poor models. In contrast, Rashomon sets allow for the targeted aggregation of models that are only high-performing, reducing the risk of incorporating poor models in the query-selection process.

In the space of decision trees, Xin et al. is the first to provide an algorithm that completely enumerate the Rashomon set for sparse decision trees [22]. Their algorithm, `TreeFarms`, provides an exhaustive yet computationally feasible method to generate, store, and view the entire Rashomon set of decision trees. However, due to the inherent structure and geometry of decision trees, many trees in this set may offer redundant explanations of the data. This can be seen in Figure 1 and more deeply in Figure 3 in the appendix. As such, duplicate trees in the ensemble method have the potential to further skew our metric of uncertainty in the committee by artificially inflating agreement in votes.

To address this limitation, section 4 will propose a method to group trees based on their unique explanations of the data and select a representative from each group. We will then show how ensembling the Rashomon set to account for model ambiguity in the active learning process will improve our query-selection criteria.

2

3 Active Learning

3.1 Notation

Borrowing notation from Liu et. al (2022) [13], let observation i be composed of data $(\mathbf{x}_i, y_i)$ for vector $\mathbf{x}_i$ in covariate space $\mathcal{X}$ and label y_i in output space $\mathcal{Y}$. The data is sent through a supervised learning model $F(\cdot) : \mathcal{X} \to \mathcal{Y}$. When $F(\cdot)$ is an ensemble method consisting of base learners, denote the base learners at $\{f_m\}_{m=1}^M$. The model is learned from a training dataset $D_{tr} = \{(\mathbf{x}_i, y_i)\}_{i=1}^I$ and tested by an independent dataset $D_{ts} = \{(\mathbf{x}_j, y_j)\}_{j=1}^J$. The goal is to train $F(\cdot)$ to predict the labels of the out-of-sample test set with a budget-constrained number of labeled observations.

Active learning seeks to adaptively and strategically choose which unlabelled observations should be queried for oracle labeling and to then be used in the supervised learning model. Let the query iteration in the active learning framework be denoted by n. Denote the reservoir of unlabelled candidate observations as $D_{cdd}^{(n)} = \{(\mathbf{x}_k, y_k)\}_{k=1}^K$ with y_k initially unknown. A selector $S(\cdot)$ is the strategy used to select samples from $D_{cdd}^{(n)}$ to be oracle labelled. At each iteration, $S(\cdot)$ will sample a subset of observations, denoted $B^{(n)}$, from the candidate dataset $D_{cdd}^{(n)}$ without replacement to query for oracle labeling. $B^{(n)}$ is then added to the training set and removed from the candidate set: $D_{cdd}^{(n+1)} = D_{cdd}^{(n)} \cup B^{(n)}$ and $D_{cdd}^{(n+1)} = D_{cdd}^{(n)} \backslash B^{(n)}$. The model is then retrained on the new training set as $F^{(n+1)}\left(D_{tr}^{(n+1)}\right)$. As such, the $B^{(n)}$ is chosen so as to find the observations that are most informative to improving predictive performance.

The process is repeated, gradually expanding the training set with informative observations, until the labelling budget is reached or a desired classification metric threshold is met.

3.2 Query-By-Committee Metrics

Picking a selector metric is a key topic in the active learning literature. Common methods are uncertainty [12], Query-By-Committee metrics [5] [19], or expected error [15]. Due to the ensembling nature of our methods, we choose to measure informativity by Query-By-Committee (QBC) metrics, particularly Argamomn-Engelson and Dagan's vote entropy [3]:

$$\delta(y, \mathbf{x}, \mathcal{C}) = \max_{\mathbf{x}} - \sum_{y \in \mathcal{Y}} \frac{\text{vote}_\mathcal{C}(y, \mathbf{x})}{|\mathcal{C}|} \log \frac{\text{vote}_\mathcal{C}(y, \mathbf{x})}{|\mathcal{C}|} \tag{1}$$

where $\text{vote}_\mathcal{C}(y, \mathbf{x}) = \sum_{c \in \mathcal{C}} \mathbb{I}\{c(\mathbf{x}) = y\}$ is the number of "votes" that label $y \in \mathcal{Y}$ receives for $\mathbf{x}$ amongst the members c of committee $\mathcal{C}$. This selector metric is a committee-based generalization of uncertainty measures that considers the confidence of each committee member and is essentially a Bayesian adaptation of Shannon's 1948 uncertainty sampling entropy [20]. One can observe, from Equation 1 that ensembling duplicate models has the potential to overinflate the vote entropy [14] with trees from the best performing explanation group.

4 Algorithm/Methods

In our proposed work, the committee $\mathcal{C}$ in Equation 1 is the Rashomon set of decision trees $\mathcal{R}$. For predictor F, we construct a Rashomon set $\mathcal{R}$ of "near-equal" decision trees defined as the set of models whose objective function is within ϵ of the overall best model given the data. Since each near-equal model in the Rashomon set will describe the data differently, conflicting prediction labels/probabilities will arise amongst models in the Rashomon set.

To enumerate the Rashomon set of decision trees, we use Xin et al.'s TreeFarms approach [22]. TreeFarms exhaustively enumerates the Rashomon set of decision trees, allowing us to aggregate the best models in our ensemble method. However, unlike random forests, TreeFarms lacks the random sampling of features and data, making the models in TreeFarms correlated. This correlation in decision trees is a significant limitation, as correlation amongst committee members may both artificially inflate agreement in the vote and complicate interpretability [14]. To address this issue, we reduce the redundancy in TreeFarms by grouping trees based on their unique explanation of the data and selecting a single representative tree from each of these groups to ensemble. This ensures that
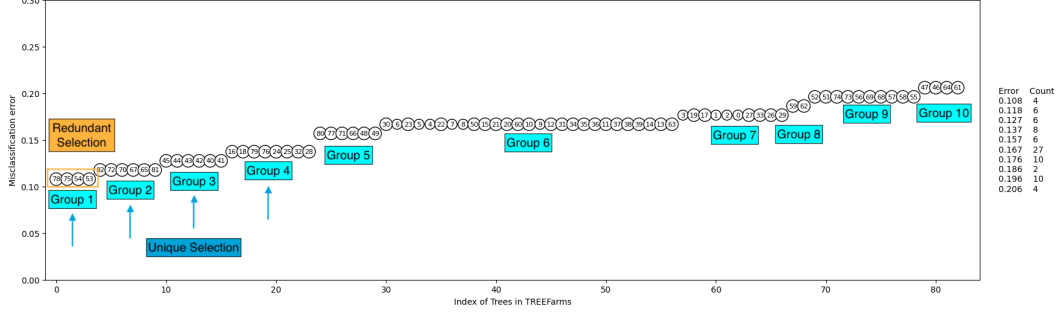
Figure 1: A depiction of how to ensemble Rashomon trees: redundantly and uniquely. This plot shows the classification error by the ordered indices of the tree. As shown, many trees have the *exact same* misclassification rate. For instance, the top 4 trees (whose geometry can be seen in Figure 3 of the appendix) share a misclassificaiton rate of 0.108. If this redundancy in the trees is not accounted for, committee agreement will be overinflated and dominated by the best performing group.

each chosen tree is meaningfully distinct while faithfully representing the Rashomon set, ultimately leading to a valid query-by-committee voting approach.

Our approach can be visualized in Figure 1. Suppose we want to define the committee of vote entropy by ensembling the top four decision trees of TreeFarms. If we ignore the the redundancy of explanations in TreeFarms, then trees 53, 54, 75, and 78 will be ensembled despite offering the same explanation and prediction. This, as noted in Equation 1, will artificially inflate agreement amongst our committee by own ensembling the trees in the top explanation groups. If we instead account for the redundancy of the trees, the unique selection method will instead choose one tree arbitrarily from Groups 1, 2, 3, and 4, diversifying our committee and more fully representing the Rashomon set. Our method is summarized in Algorithm 1.

Algorithm 1: Unique Tree Farms Active Learning

Input : $D_{tr}^{(0)}$; D_{ts}; $D_{cdd}^{(0)}$; ϵ;

1 **repeat**
2   Train $\mathbf{F}$ on $D_{tr}^{(n)}$;
3   Test $\mathbf{F}$ on $D_{ts}^{(n)}$;
4   Enumerate the Rashomon set $\mathcal{R}$ of predictor F with TreeFarms;
5   *(Optionally)* Reduce the Rashomon set $\mathcal{R}$ to the top k models in $\mathcal{R}$;
6   Predict labels $\hat{y}_{tr,m}^{(n)}$ and calculate the classification error for each tree f_m in $\mathcal{R}$;
7   Define the the smallest classification error from the $\mathcal{R}$ as the current iteration error;
8   Compute the vote-entropy metric $\delta^{(n)}(y, x, \mathcal{C})$ from equation 1 with $\mathcal{R}$ as the committee;
9   Resample $B^{(n)}$ from $D_{cdd}^{(n)}$ based on the observation with the highest vote entropy:
      $B^{(n)} := \arg\max_x \delta^{(n)}(y, x, \mathcal{C})$;
10   Query $B^{(n)}$ for oracle labeling;
11   Set $D_{tr}^{(n+1)} = D_{tr}^{(n)} \cup B^{(n)}$ and $D_{cdd}^{(n+1)} = D_{cdd}^{(n)} \setminus B^{(n)}$;
12 **until** *labelling budget is depleted or test error is sufficiently small*;

---

5 Experiments

One hundred active learning simulations were ran and averaged on the 1978 Boston Housing dataset of Harrison and Rubinfeld [8]. The goal was to classify whether the median value of a home was in the top 25% quantile based on five covariates: capita crime rate per town, nitric oxides concentration, average number of rooms per household, pupil-teacher ratio by town, and percent of lower status of the population. Due to the structure of TreeFarms, we discretized the five covariates into three categories (low, medium, and high) and one-hot encoded the variables. This resulted in the covariates being encoded in 15 binary variables. Open source code for the simulation is available on Github.

We compared the active learning process ensembling the top ten and 100 trees with unique explanations from the Rashomon set to its counterpart without considering the redundancy in tree explanations. At each iteration, TreeFarms was refitted with a regularization penalty on splits of 0.01 and a Rashomon ϵ of 0.05. Results can be seen in Figure 2. Figure 4 of the appendix shows the same simulation results without the baseline random forest.

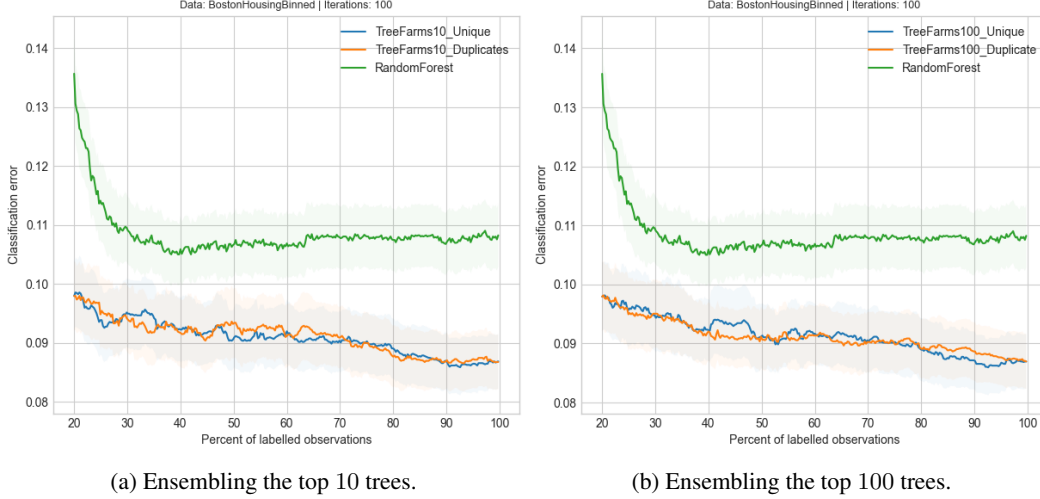(a) Ensembling the top 10 trees.
(b) Ensembling the top 100 trees.

Figure 2: This plot gives the classification error comparing three ensemble methods: TreeFarms selecting redundant trees, TreeFarms selecting unique trees, and random forests.

Our findings demonstrate the remarkable predictive power of the Rashomon set. Ensembles of decision trees from the Rashomon set (both unique and duplicate) significantly outperformed random forests, highlighting the robustness and prediction accuracy achievable with this approach. Furthermore, removing redundant explanations in the Rashomon set by only ensembling trees with unique explanations of the data maintained classification accuracy.

This result has profound implications for interpretability. While redundancy in explanations can hinder the interpretability of an ensemble, the Rashomon framework allows us to overcome this challenge by selecting a smaller, coherent set of unique trees while maintaining prediction accuracy. This approach combines the robustness of random forests with the interpretability of individual trees

6 Concluding thoughts and future work

Our work offers two key insights. Firstly, we demonstrate that ensembling over the Rashomon set of decision trees enhances the active learning process by a significant margin. Unlike traditional ensemble methods which aggregate over the entire space of models, potentially including models that are poor performing or implausible, the Rashomon set only contains models with high posterior probability. This allows active learning processes to form a commmitee with only the strong and plausible models whose disagreements will then provide a more robust measure of uncertainty for more efficient query selection.

Secondly, we address the issue of redundant and duplicate explanations when constructing a Rashomon set by only considering trees with unique explanations. Redundant explanations can inflate query-by-committee metrics and obscure interpretability. By only ensembling over the Rashomon's subset of trees with unique explanations, we ensure that the ensemble remains parsimonious and interpretable while maintaining prediction accuracy.

This work *plants the seeds* for future research into other methods that form the Rashomon set without an inherent geometric structure. In particular, Rashomon Partition Sets (RPS) [21] offer a promising framework for comprehensively enumerating the Rashomon set. Investigating the use of RPS in active learning may further deepen our understanding of the Rashomon's benefits in both prediction and interpretability.

References

[1] Kenneth Arrow. Alternative approaches to the theory of choice in risk taking situation. *Econometrica*, 19, 10 1951. doi: 10.2307/1907465.

[2] Leo Breiman. Statistical Modeling: The Two Cultures (with comments and a rejoinder by the author). *Statistical Science*, 16(3):199 – 231, 2001. doi: 10.1214/ss/1009213726. URL https://doi.org/10.1214/ss/1009213726.

[3] Ido Dagan and Sean P. Engelson. Committee-based sampling for training probabilistic classifiers. In *Proceedings of the Twelfth International Conference on International Conference on Machine Learning*, ICML'95, page 150–157, San Francisco, CA, USA, 1995. Morgan Kaufmann Publishers Inc. ISBN 1558603778.

[4] Shuxian Fan, Adam Visokay, Kentaro Hoffman, Stephen Salerno, Li Liu, Jeffrey T. Leek, and Tyler H. McCormick. From narratives to numbers: Valid inference using language model predictions from verbal autopsy narratives. *CoRR*, abs/2404.02438, 2024. doi: 10.48550/ARXIV.2404.02438. URL https://doi.org/10.48550/arXiv.2404.02438.

[5] Yoav Freund, H. Sebastian Seung, Eli Shamir, and Naftali Tishby. Selective sampling using the query by committee algorithm. *Mach. Learn.*, 28:133–168, September 1997. ISSN 0885-6125. doi: 10.1023/A:1007330508534. URL http://portal.acm.org/citation.cfm?id=263100.263123.

[6] Andreas Graefe, Helmut Küchenhoff, Veronika Stierle, and Bernhard Riedl. Limitations of ensemble bayesian model averaging for forecasting social science problems. *International Journal of Forecasting*, 31(3):943–951, 2015. ISSN 0169-2070. doi: https://doi.org/10.1016/j.ijforecast.2014.12.001. URL https://www.sciencedirect.com/science/article/pii/S0169207014001769.

[7] Lars Peter Hansen. Nobel lecture: Uncertainty outside and inside economic models. *Journal of Political Economy*, 122(5):945–987, 2014. ISSN 00223808, 1537534X. URL http://www.jstor.org/stable/10.1086/678456.

[8] David Harrison and Daniel Rubinfeld. Hedonic housing prices and the demand for clean air. *Journal of Environmental Economics and Management*, 5:81–102, 03 1978. doi: 10.1016/0095-0696(78)90006-2.

[9] Jennifer A. Hoeting, David Madigan, Adrian E. Raftery, and Chris T. Volinsky. Bayesian model averaging: a tutorial (with comments by M. Clyde, David Draper and E. I. George, and a rejoinder by the authors. *Statistical Science*, 14(4):382 – 417, 1999. doi: 10.1214/ss/1009212519. URL https://doi.org/10.1214/ss/1009212519.

[10] Miriam Huijser and Jan C. van Gemert. Active decision boundary annotation with deep generative models. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[11] Seho Kee, Enrique del Castillo, and George Runger. Query-by-committee improvement with diversity and density in batch active learning. *Information Sciences*, 454-455:401–418, 2018. ISSN 0020-0255. doi: https://doi.org/10.1016/j.ins.2018.05.014. URL https://www.sciencedirect.com/science/article/pii/S0020025518303700.

[12] David D. Lewis and William A. Gale. A sequential algorithm for training text classifiers. In Bruce W. Croft and C. J. van Rijsbergen, editors, *SIGIR '94*, pages 3–12, London, 1994. Springer London. ISBN 978-1-4471-2099-5.

[13] Peng Liu, Lizhe Wang, Rajiv Ranjan, Guojin He, and Lei Zhao. A survey on active deep learning: From model driven to data driven. *ACM Comput. Surv.*, 54(10s), sep 2022. ISSN 0360-0300. doi: 10.1145/3510414. URL https://doi.org/10.1145/3510414.

[14] Prem Melville and Raymond J. Mooney. Diverse ensembles for active learning. In *Proceedings of the Twenty-First International Conference on Machine Learning*, ICML '04, page 74, New York, NY, USA, 2004. Association for Computing Machinery. ISBN 1581138385. doi: 10.1145/1015330.1015385. URL https://doi.org/10.1145/1015330.1015385.

[15] Nicholas Roy and Andrew McCallum. Toward optimal active learning through sampling estimation of error reduction. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, page 441–448, San Francisco, CA, USA, 2001. Morgan Kaufmann Publishers Inc. ISBN 1558607781.

[16] Cynthia Rudin, Chudi Zhong, Lesia Semenova, Margo Seltzer, Ronald Parr, Jiachang Liu, Srikar Katta, Jon Donnelly, Harry Chen, and Zachery Boner. Amazing things come from having many good models. *ArXiV*, 07 2024.

[17] Raphael Schumann and Ines Rehbein. Active learning via membership query synthesis for semi-supervised sentence classification. In Mohit Bansal and Aline Villavicencio, editors, *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 472–481, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/K19-1044. URL https://aclanthology.org/K19-1044.

[18] Lesia Semenova, Cynthia Rudin, and Ronald Parr. On the existence of simpler machine learning models. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, page 1827–1858, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450393522. doi: 10.1145/3531146.3533232. URL https://doi.org/10.1145/3531146.3533232.

[19] Burr Settles. *Active Learning*. Morgan & Claypool Publishers, 2012. ISBN 1608457257.

[20] C. E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423, 1948. doi: 10.1002/j.1538-7305.1948.tb01338.x.

[21] Aparajithan Venkateswaran, Anirudh Sankar, Arun G. Chandrasekhar, and Tyler H. Mc-Cormick. Robustly estimating heterogeneity in factorial data using rashomon partitions. *ArXiv*, abs/2404.02141, 2024. URL https://api.semanticscholar.org/CorpusID:268857158.

[22] Rui Xin, Chudi Zhong, Zhi Chen, Takuya Takagi, Margo Seltzer, and Cynthia Rudin. Exploring the whole rashomon set of sparse decision trees. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA, 2024. Curran Associates Inc. ISBN 9781713871088.

[23] Jia-Jie Zhu and José Bento. Generative adversarial active learning. *ArXiv*, abs/1702.07956, 2017. URL https://api.semanticscholar.org/CorpusID:14631999.

7 Appendix

7.1 Geometry of trees in Group 1

The images in Figure 3 give insight into the decision rules of the top four decision trees. Note that feature 0 represents whether an observation is within an area with lowest levels of crime (bottom 33% quantile), and feature 9 represents an observation in an area with the lowest pup-teacher ratio (bottom 33% quantile). Features 6, 7, and 8 represent whether an observation is in one of the three categories of rooms per household respectively: low, medium, and high.

As seen, the trees exhibit very similar decision paths to each other, resulting in each one having the *exact* same misclassification error of 0.108. As described in the main corpus, ensembling these four trees as a committee and calculating the vote entropy metric off this committee will result in an inflated agreement and will recommend the observation that the best decision tree is most uncertain of rather than consider the uncertainty of the ensemble as a whole.

7.2 Experimental results plotted with random forests

Figure (4) gives the classification error comparing the two ensemble methods: TreeFarms selecting redundant trees and its counterpart with only unique trees It is the same plot as Figure 2 but removes the classification error line for random forests to better visualize the differences between the inclusion of redundant vs. duplicate trees in the ensemble method. The value given in each subfigure represents the p-value from a Wilcoxon ranked-sign test.
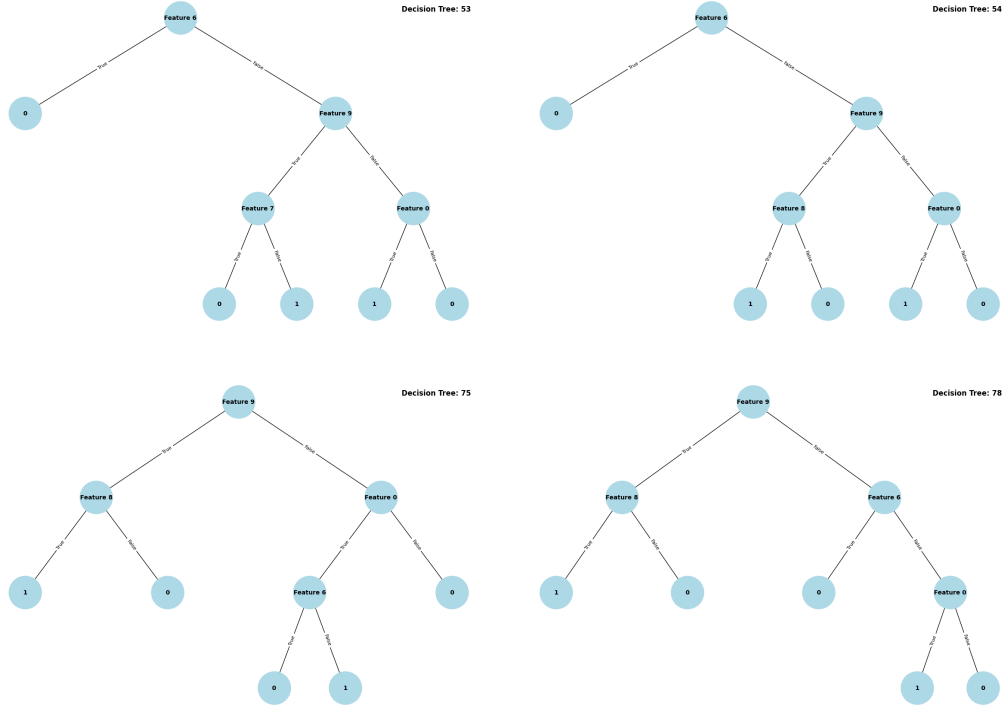
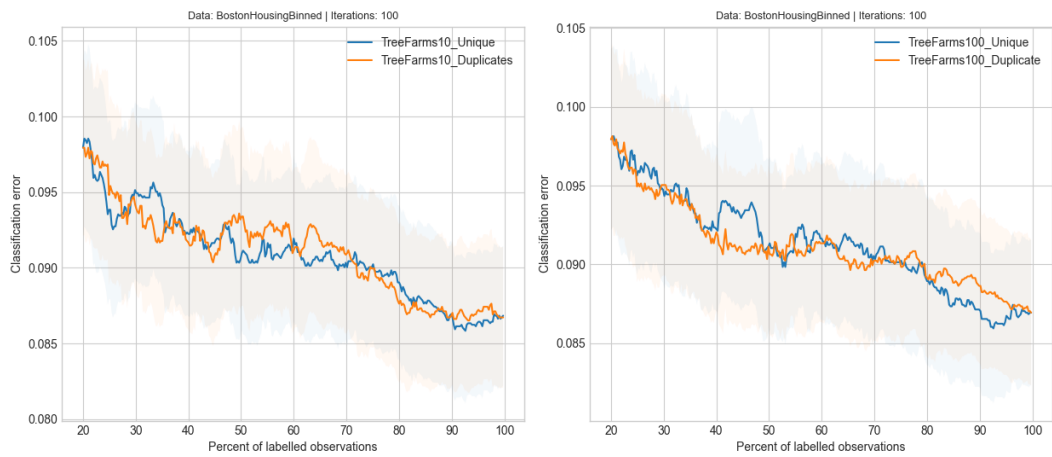Figure 3: Geometry of the best four decision trees in Group 1 of Figure 1.

(a) Ensembling the top 10 trees ($p = 0.00031$).

(b) Ensembling the top 100 trees ($p = 0.10174$).

Figure 4