





# Visualizing Prediction Provenance in Regression Random Forests

N. Médoc<sup>1</sup> , V. Ciorna<sup>2</sup> , F. Petry<sup>2</sup> , and M. Ghoniem<sup>1</sup> 

<sup>1</sup>Luxembourg Institute of Science and Technology, Luxembourg

<sup>2</sup>Goodyear Innovation Center Luxembourg, Luxembourg

## Abstract

Random forest models are widely used in many application domains due to their performance and the fact that their constituent decision trees carry clear decision rules. Yet, the provenance of the predictions made by an entire forest is complex to grasp, which motivates application domain experts to adopt black-box testing strategies. We propose in this paper a coordinated multiple view system allowing to shed more light on prediction provenance, uncertainty and error in terms of bias and variance at the global model scale or at the local scale of decision paths and individual instances.

## CCS Concepts

• **Human-centered computing** → *Visualization*; • **Computing methodologies** → *Classification and regression trees*;

## 1. Introduction

Decision tree models owe their popularity to the fact that any prediction can be explained as a relatable cascade of rules. A random forest (RF) is an ensemble method using a collection of decision trees as weak learners. Unlike single decision tree models, random forests do not overfit [Bre01], hence their wide use in many application areas. The provenance of their predictions is still complex to apprehend. To build trust in the behavior of a predictive model, including random forests, an application domain expert will often adopt a black-box testing strategy [Ost02] by trying different inputs and comparing the model outputs to her expectations. Regardless of the model evaluation work done by the model builder, the end user will still need a way to build trust in the model. We focus on the following goals related to the trust levels (TL) of Chatzimparmpas et al. [CMJ\*20]: ( $G_1$ ) what rules/criteria the model uses to reach a given prediction (TL3, understanding and explanation)? ( $G_2$ ) what data was used to learn the model or to form a decision path (TL3, diagnosis)? ( $G_3$ ) what is the degree of uncertainty for a prediction (TL4, performance, model bias and variance)?

Previous work using visual analytics to explain decision tree ensembles raises several known challenges, including: visualizing many decision trees [MNP21, NP21], summarizing the learned decision rules [ZWLC19], supporting hyper-parameter tuning and comparing model accuracy for multi-class prediction models [LXL\*18]. Most visual analytics solutions for machine learning focus more on classification and less on regression [CMJ\*20]. Also, most solutions provide global model quality scores, and rarely local instance-level provenance information, e.g., Neto and Paulovich [NP21] or Sawada and Toyoda [ST19] for classification.

Focusing on regression random forests, this paper presents a visual approach to assess prediction errors in terms of bias and vari-

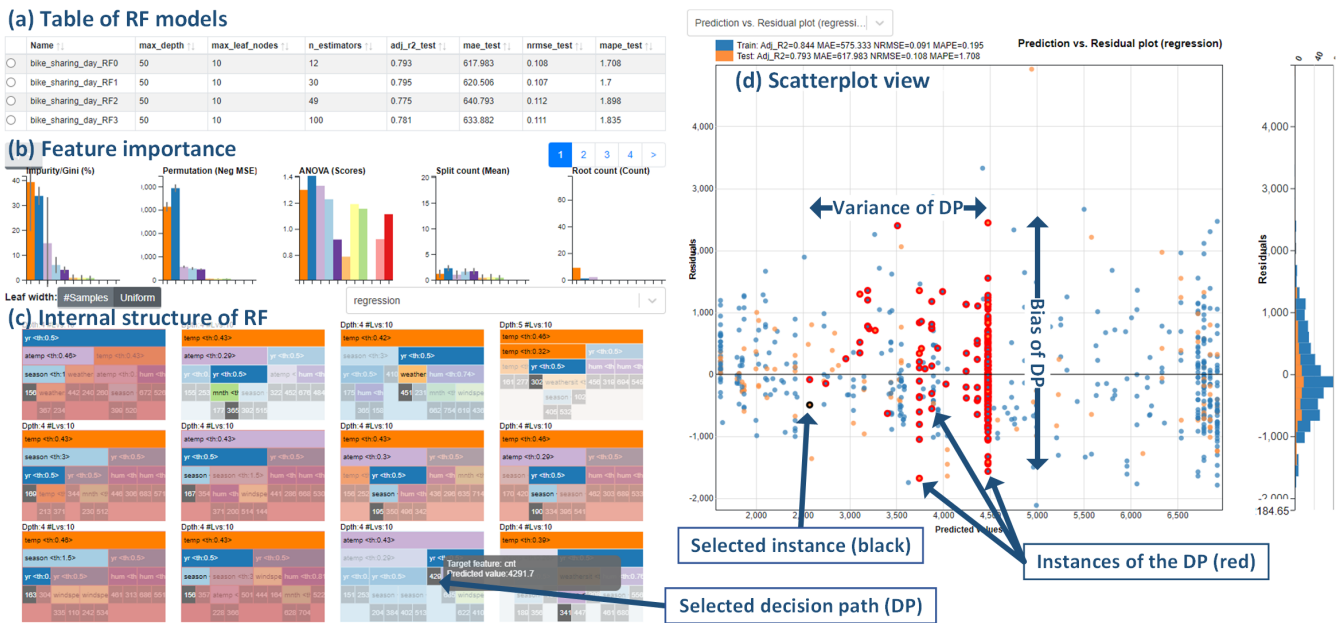
ance. By coordinating existing visualizations, we show how user interaction can support the analysis of prediction provenance at instance level or for all instances falling in a selected decision path.

## 2. System overview

Figure 1 shows the user interface of the SYLVIA system (from the latin silva, i.e. forest, and visual analytics) comprising four coordinated views. A table view (Figure 1 a) lists all models trained on a given data set, e.g. with different hyper-parameters. A set of bar charts (Figure 1 b) shows relative feature importance according to five different metrics, along with variance bars across all models. Features are distinguished by their color on a categorical palette.

A grid of icicle plots (Figure 1 c) gives an overview of the forest and the topology of each decision tree. Each internal node represents a decision rule based on a threshold of a feature in the data. Nodes are colored consistently with the bars of the bar charts. A path in the tree is a cascade of rules that lead to a decision reached at the level of its leaf node. Since leaf nodes do not correspond to any specific feature, they are colored in gray. Node size encodes the relative number of instances funneled through the node during model training. This shows, at the scale of a tree, which decision path (DP) was more frequently used during the learning phase. Some decision paths built with few instances may be too skinny. Switching to unit weights on all leaf nodes helps to see the tree topology.

A residual error vs. predicted scatterplot (Figure 1 d) supports the assessment of model quality, combined with a bar chart showing the distribution of residual errors. Usual model quality scores are displayed above the scatter plot. Instances colored in blue and orange belong to the training set and the test set respectively. The user gets a direct impression of global model bias from the his-



**Figure 1:** Overview of SYLVIA: (a) the list of models trained on a given data set; (b) features sorted by importance according to different metrics; (c) the internal structure of the random forest (RF); (d) residual vs. predicted scatterplot.

toграм of residuals. For instance, the multi-modal distribution of test data (orange) in Figure 1 shows an important bias of the model. From the scatterplot, the domain expert can also judge for each prediction the magnitude of residual error. For a given predicted value (abscissa), the spread of the residual errors along the Y-axis reflects prediction bias ( $G_3$ ), noticeable in Figure 1-d by the vertical alignment of instances. The scatterplot can also be used in a *predicted vs. actual* mode, which in passing leads to more accurate model quality judgment than the reverse *actual vs. predicted* mode [PPGP08].

Coordinating the icicle plot grid and the scatter plot helps to pinpoint local patterns that might explain model performance.

*Selecting an instance in the scatterplot* highlights the DP where the instance falls, i.e. the applicable decision rules ( $G_1$ ). The background color of each icicle plot encodes the difference between the selected instance and the prediction yielded by the DP on a diverging red-white-blue color palette (red for negative and blue for positive), as in Figure 1 c. This helps the analyst understand the prediction bias for a given instance at the level of each tree and identify precisely which trees and DPs provide good or poor predictions ( $G_3$ ). The distribution of background color gives a sense of the agreement between the trees, which may be seen as a visual indication of prediction uncertainty ( $G_3$ ). In Figure 1 c, a split node on “season” appears in all DPs with a red background only. This rule seems to explain the prediction disagreement between trees.

*Details on demand* are obtained as tooltips when hovering over internal tree nodes including feature name and the split threshold, or the predicted value for leaf nodes (Figure 1 c).

*Selecting a decision path* in the RF highlights in the scatterplot all instances falling in that path ( $G_2$ ). By inspecting the spread of the selected instances along the X-axis or the Y-axis, the analyst can as-

sess the prediction variance, respectively the prediction bias, within the selected DP ( $G_3$ ). When the highlighted instances are shifted to the right of the initial (black) instance the residual error at the DP level is positive, as in Figure 1 d.

*Persisting the selected instances after switching models* allows to compare variance and bias across models for these instances ( $G_3$ ), e.g. to analyze the impact of model hyper-parameters, e.g., number of trees, or tree morphology, on prediction quality. The proposed interactions aim to give an intuitive understanding of model bias and variance and possible trade-offs between model complexity and model performance, as shown in the supplementary material.

### 3. Discussion and future work

This work stems from an industrial collaboration, where domain experts develop products with critical safety consequences. Predictive models promise to speed up current development processes, but cannot be deployed based on faith in data science only [SMK\*18]. In SYLVIA, the grid of icicle plots exposes the internal structure of an RF and the scatter plot shows local patterns of bias and variance. An analyst can identify and analyze local patterns surrounding a prediction at the level of each decision tree and get a sense of certainty and a more precise context for the prediction. Yet, a forest with many intricate trees may be daunting. Prior work shows that the performance of an RF will plateau around 100 trees [OPB12]. To scale up this approach, sorting and clustering the decision trees may help to understand the variety of trees in the forest. One may also run an experiment to identify the pros and cons of this approach compared to other approaches, e.g. matrix of decision paths [NP21]. Finally we are extending SYLVIA to handle tree boosting approaches in which decision trees are interdependent.

## References

- [Bre01] BREIMAN L.: Random forests. *Machine learning* 45, 1 (2001), 5–32. [1](#)
- [CMJ\*20] CHATZIMPAMPAS A., MARTINS R. M., JUSUFI I., KUCHER K., ROSSI F., KERREN A.: The State of the Art in Enhancing Trust in Machine Learning Models with the Use of Visualizations. *Computer Graphics Forum* 39, 3 (2020), 713–756. [\\_eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/cgf.14034](#). [doi:10.1111/cgf.14034](#). [1](#)
- [LXL\*18] LIU S., XIAO J., LIU J., WANG X., WU J., ZHU J.: Visual Diagnosis of Tree Boosting Methods. *IEEE Transactions on Visualization and Computer Graphics* 24, 1 (Jan. 2018), 163–173. [doi:10.1109/TVCG.2017.2744378](#). [1](#)
- [MNP21] MAZUMDAR D., NETO M. P., PAULOVIH F. V.: Random Forest Similarity Maps: A Scalable Visual Representation for Global and Local Interpretation. *Electronics* 10, 22 (Jan. 2021), 2862. Number: 22 Publisher: Multidisciplinary Digital Publishing Institute. [doi:10.3390/electronics10222862](#). [1](#)
- [NP21] NETO M. P., PAULOVIH F. V.: Explainable Matrix - Visualization for Global and Local Interpretability of Random Forest Classification Ensembles. *IEEE Transactions on Visualization and Computer Graphics* 27, 2 (Feb. 2021), 1427–1437. Conference Name: IEEE Transactions on Visualization and Computer Graphics. [doi:10.1109/TVCG.2020.3030354](#). [1, 2](#)
- [OPB12] OSHIRO T. M., PEREZ P. S., BARANAUSKAS J. A.: How many trees in a random forest? In *Machine Learning and Data Mining in Pattern Recognition* (Berlin, Heidelberg, 2012), Perner P., (Ed.), Springer Berlin Heidelberg, pp. 154–168. [2](#)
- [Ost02] OSTRAND T.: Black-box testing. *Encyclopedia of Software Engineering* (2002). [1](#)
- [PPGP08] PIÑEIRO G., PERELMAN S., GUERSCHMAN J. P., PARUELO J. M.: How to evaluate models: Observed vs. predicted or predicted vs. observed? *Ecological Modelling* 216, 3–4 (2008). [doi:10.1016/j.ecolmodel.2008.05.006](#). [2](#)
- [SMK\*18] SUCIU O., MARGINEAN R., KAYA Y., DAUME III H., DUMITRAS T.: When does machine learning {FAIL}? generalized transferability for evasion and poisoning attacks. In *27th USENIX Security Symposium (USENIX Security 18)* (2018), pp. 1299–1316. [2](#)
- [ST19] SAWADA S., TOYODA M.: *Model-Agnostic Visual Explanation of Machine Learning Models Based on Heat Map*. The Eurographics Association, 2019. Accepted: 2019-06-02T18:21:14Z. [doi:10.2312/eurp.20191140](#). [1](#)
- [ZWLC19] ZHAO X., WU Y., LEE D. L., CUI W.: iForest: Interpreting Random Forests via Visual Analytics. *IEEE Transactions on Visualization and Computer Graphics* 25, 1 (Jan. 2019), 407–416. Conference Name: IEEE Transactions on Visualization and Computer Graphics. [doi:10.1109/TVCG.2018.2864475](#). [1](#)