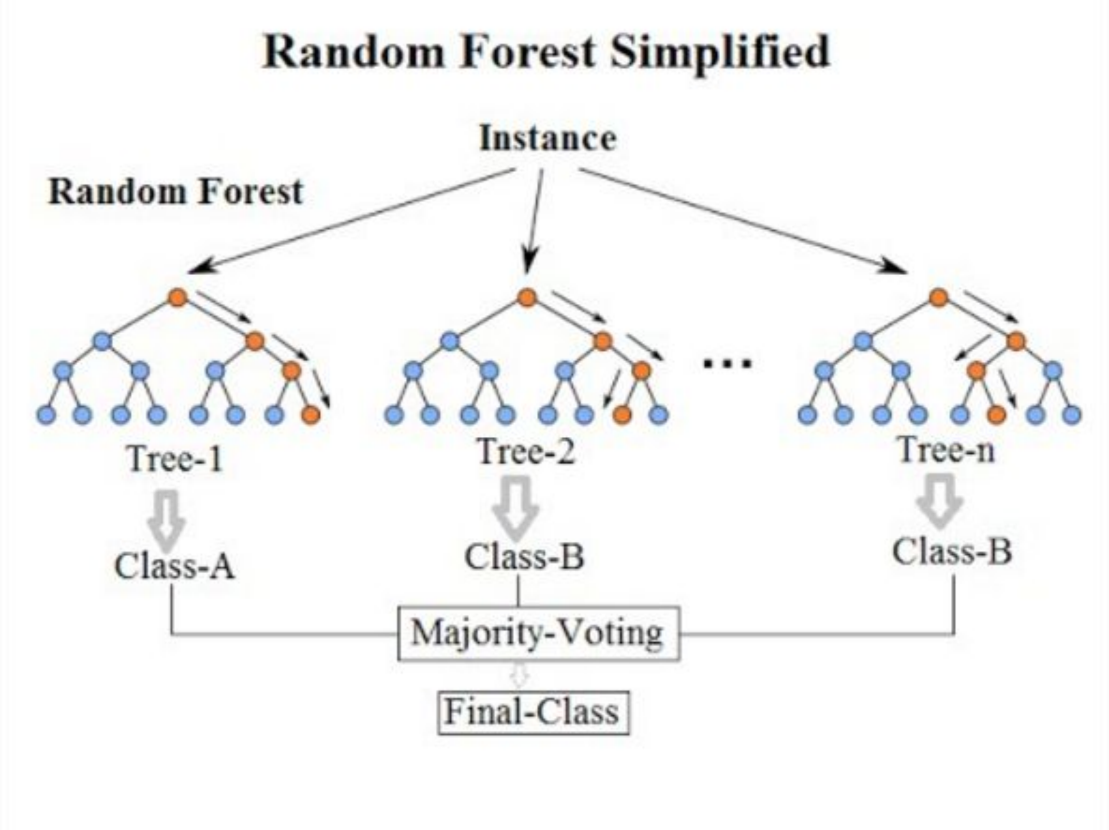
Dynamic Semi-random Forest

Nikhil Alladi, Jacob Dipasupil, Petr Kisselev

Random Forest Algorithm

- Ensemble Learning Algorithm
- Bootstrapping
- Random attribute subset selection
- Aggregating



Limitations of Random Forest

01

Stochastic Nature

Run-to-run variance and inconsistent results

02

Small Dataset Performance

Although RF improves with higher sample sizes, this leads to unnecessary computational bloating for little improvement

03

Uniform random attribute selection

Assumes all features are equally correlated with the class

Related Works

Wang et al.

Recursive Feature Selection

Lifandali et al.

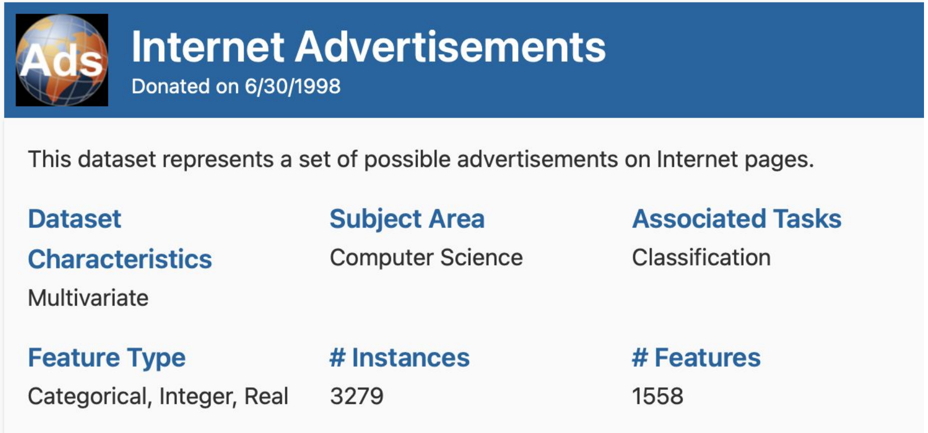
Two Phase Model

Abdellatif et al.

Weighted Random Feature Selection

Datasets

Internet Advertisements + Jobs Dataset



Internet Advertisements

Donated on 6/30/1998

This dataset represents a set of possible advertisements on Internet pages.

Dataset Characteristics	**Subject Area**	**Associated Tasks**
Multivariate	Computer Science	Classification
Feature Type	**# Instances**	**# Features**
Categorical, Integer, Real	3279	1558

- Represents possible internet advertisement images
- Task is to determine if the image is an internet advertisement based on the geometry of the image (aspect ratio), phrases in the URL, alt text, etc.
- Chose for high dimensionality

- Smaller dataset of synthetic job postings
- 1.62 million instances, 23 features
- Task is to predict salary range
- Obviously uncorrelated features (Job Portal, Contact)

Our Implementation

Dynamic Semi-Random Forest

Two Phase Model

Phase one: feature selection phase
Phase two: forest generation phase

Weighted Feature Selection

Weigh each feature based on the performance of decision trees containing those features

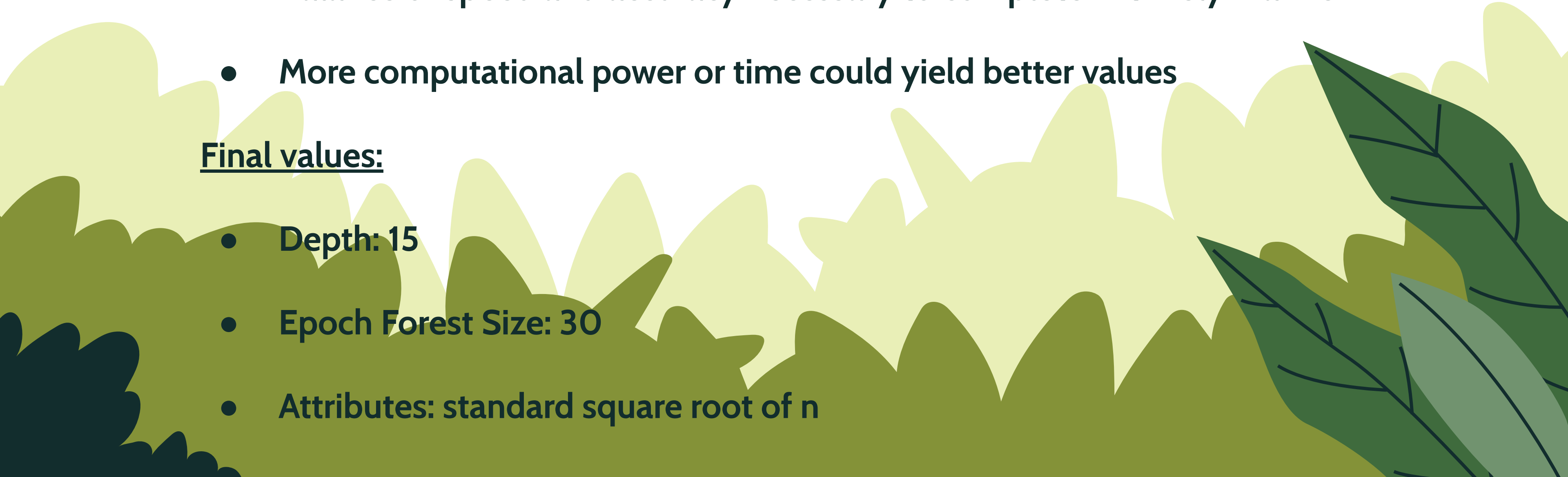
Testing Methodology

Main focus – Hyperparameter tuning

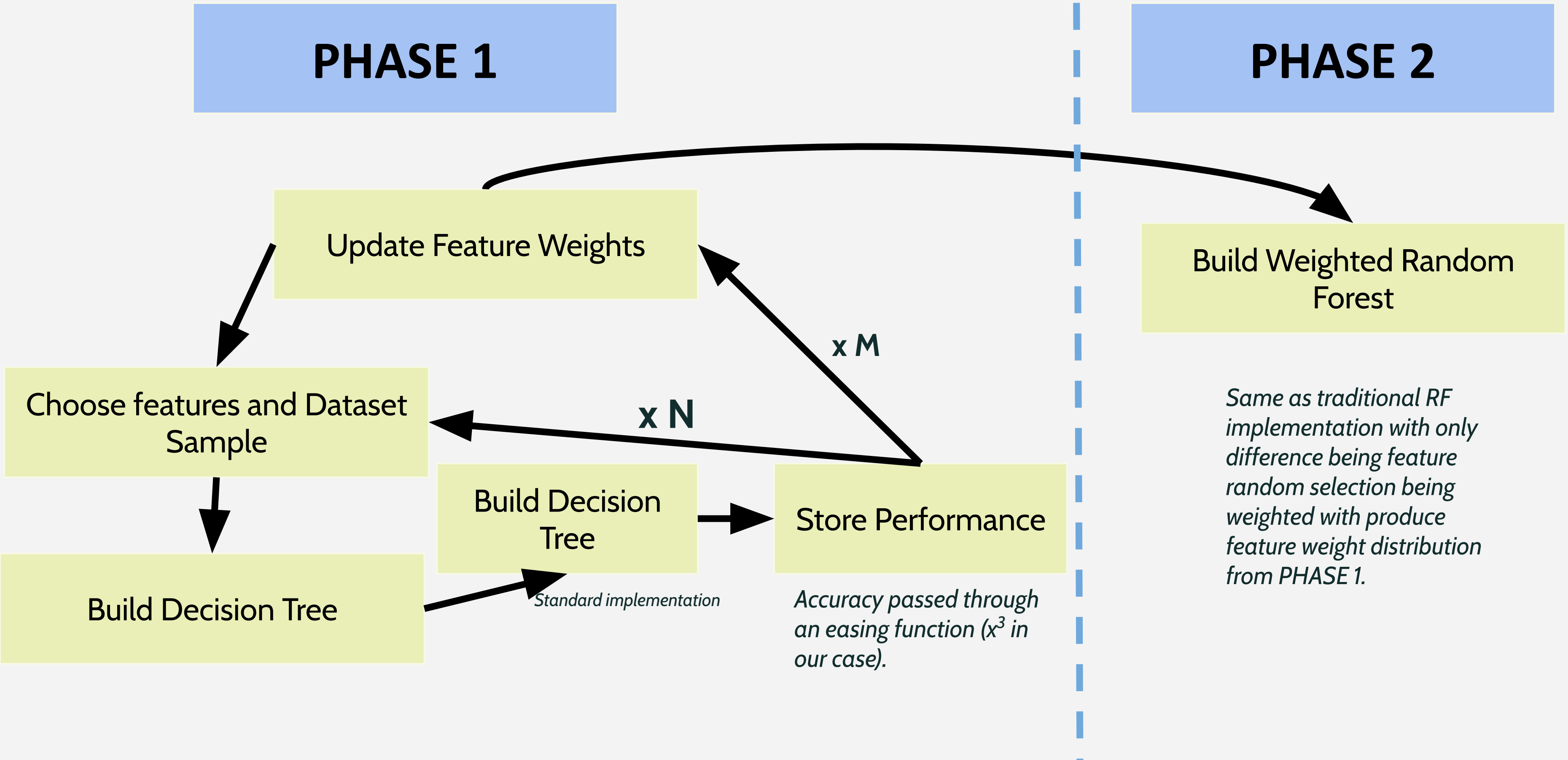
- Difficult due to high running time
- Balance of speed and accuracy necessary to complete in timely manner
- More computational power or time could yield better values

<u>Final values:</u>

- Depth: 15
- Epoch Forest Size: 30
- Attributes: standard square root of n



DSRF MODEL ARCHITECTURE



PHASE 1

PHASE 2

Update Feature Weights

Choose features and Dataset Sample

Build Decision Tree

Build Decision Tree

Store Performance

Build Weighted Random Forest

x M

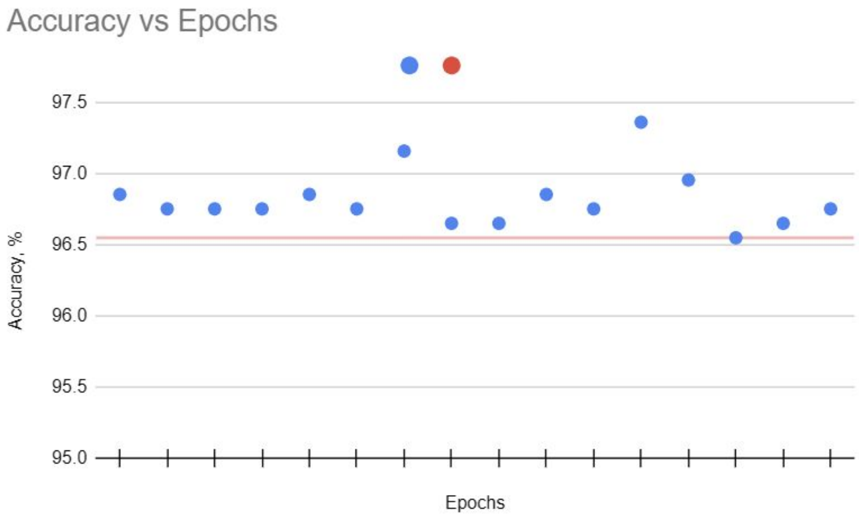
x N

Standard implementation

Accuracy passed through an easing function (x^3 in our case).

Same as traditional RF implementation with only difference being feature random selection being weighted with produce feature weight distribution from PHASE 1.

RESULTS

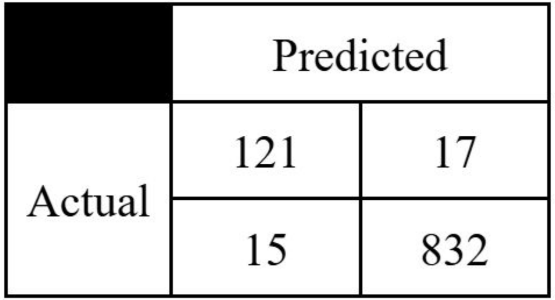


Empirically found that 11 epochs had the highest performance.

Performance not well correlated with number of epochs trained – suggests fundamental algorithmic issue

Our peak accuracy was that of about **97%**, a strong performance. Must be compared to **96%** accuracy of standard Random Forest algorithm.

4% -> **3%** missed is actually a **25%** improvement!



	Predicted	
Actual	121	17
	15	832

Precision: 0.877
Recall: 0.89

Discussion

What limited performance?

- Small room for improvement
- Getting stuck in local minimums - once attribute better than others it will receive an abundance of attention
- Grouped attributes receiving same score (no matter actual impact within Decision Tree)

Future Work

- Explore getting relative performances from Decision Tree level to leverage full scoring system
- Incorporate more advanced metrics for feature scoring
- More challenging dataset choice to have more room for improvement
- Improve efficiency to speed up runtime

Conclusion

Good

- Conceptually strong - needs deeper analysis
- Noticeable, albeit small, relative improvement

Bad

- Small *absolute* improvement
- Run time
- Local min issue

References

[1] Data source website: https://archive.ics.uci.edu/dataset/51/internet+advertisements

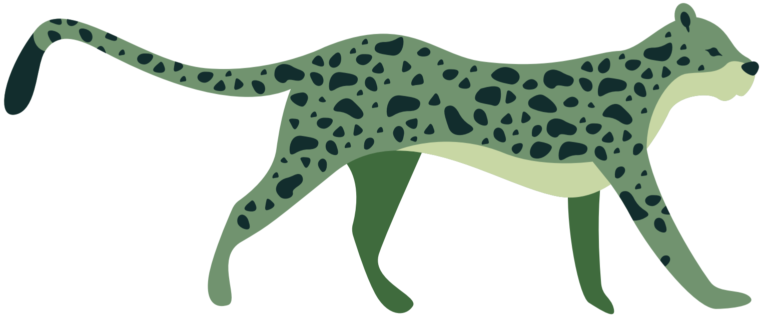
[2] https://www.sciencedirect.com/science/article/abs/pii/S0167865512001274

[3] https://www.sciencedirect.com/science/article/pii/S0957417423020511

[4] https://www.sciencedirect.com/science/article/pii/S1877050923006415

[5] https://onlinelibrary.wiley.com/doi/full/10.1155/2021/5529389

[6] https://ieeexplore.ieee.org/abstract/document/9802107



Any questions?