Want more?

https://github.com/matthewbrems/cif-ml-talk

Coding it Forward: Machine Learning & The Wisdom of The Crowds

Matt Brems

Introduction: Matt Brems (he/him)

Managing Partner & Principal Data Scientist, BetaVector

Distinguished Faculty, General Assembly

Marketing & Comms Director, Statistics Without Borders

Previously:

- Growth + Computer Vision @ Roboflow
- R&D Fairness/Bias in Data @ FINRA
- Data Science Education @ General Assembly
- Data Science @ 0ptimus Consulting
- Enterprise Analytics @ Smucker's
- M.S. Statistics @ The Ohio State University

Recommended Reads:

- Data-Driven Thinking: "Factfulness"
- Data Visualization: "Storytelling with Data"
- Data Science: "Introduction to Statistical Learning with Applications in R"

Agenda

1. **What is machine learning?**

2. Ensemble models.
   a. Bagged Decision Trees
   b. Random Forests
   c. Adaboost Models
   d. Gradient Tree Boosted Models

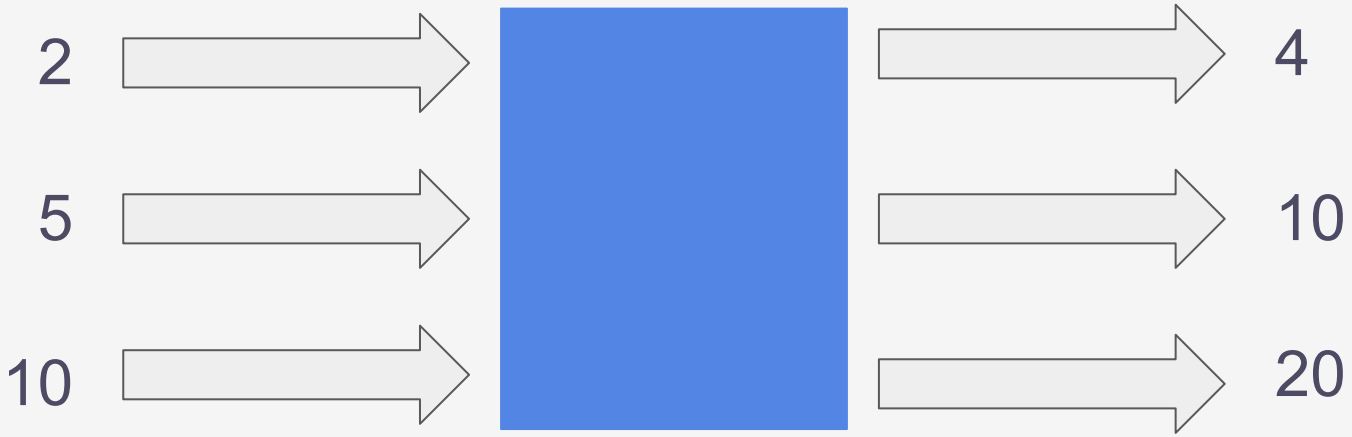
3. Model drift.

4. *Pending time, an AMA! (Other ML / data science questions, running a consultancy, my experiences being LGBTQ+ in tech, and more.)*

What is machine learning?

Machine learning (**ML**) is the study of computer algorithms that improve automatically through experience and by the use of data.[1] It is seen as a part of artificial intelligence. Machine learning algorithms build a model based on sample data, known as "training data", in order to make predictions or decisions without being explicitly programmed to do so.[2] Machine learning algorithms are used in a wide variety of applications, such as in medicine, email filtering, speech recognition, and computer vision, where it is difficult or unfeasible to develop conventional algorithms to perform the needed tasks.[3]

Let’s go back to fourth grade: inputs and outputs.



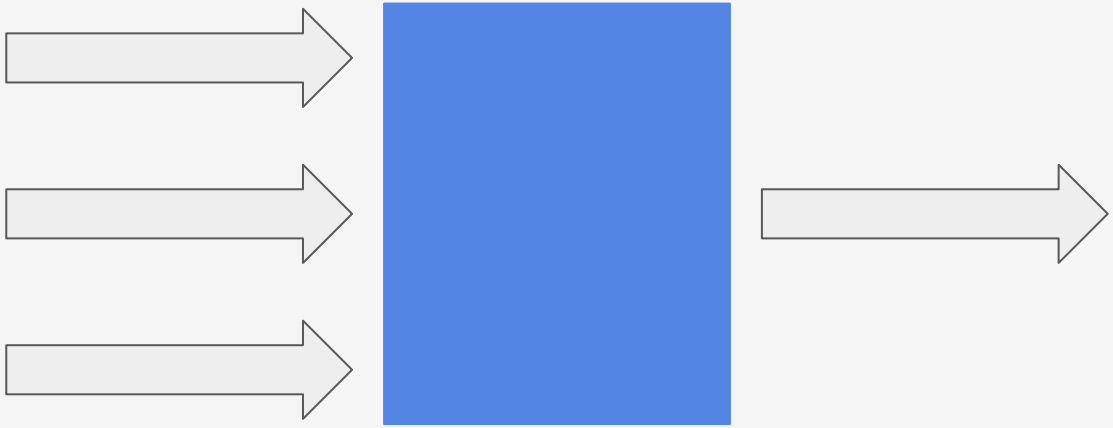
Let’s go back to fourth grade: spot the pattern.

You can I can understand what that blue box should be.

1. We have context.

2. The relationship between the inputs and outputs is ***exact***.

You know what isn't exact? The real world.



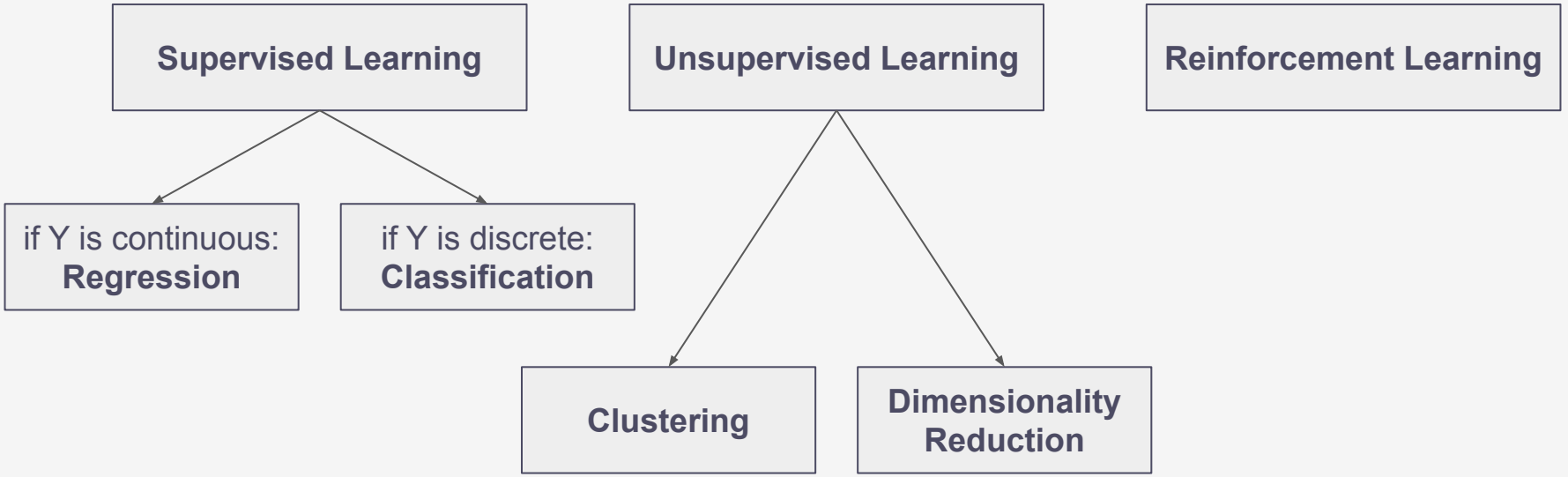
There are many, *many* types of models you can build with inputs and outputs.

- Linear regression
- Logistic regression
- Poisson regression
- Gamma regression
- Decision tree
- Bagged decision trees
- Random forest
- Adaboost
- Gradient boosted trees
- Support vector machines
- Naive Bayes
- Linear discriminant analysis

- Feedforward neural networks
- Convolutional neural networks
- Recurrent neural networks

- Generalized linear models
- Regularized models
- Time series models
- Spatiotemporal models
- Spatial models
- Time series models
- Hierarchical models
- Bayesian models

There are a lot of ways to categorize ML models… this is a common one.



Agenda

1. What is machine learning?

2. **Ensemble models.**
   a. **Bagged Decision Trees**
   b. **Random Forests**
   c. **Adaboost Models**
   d. **Gradient Tree Boosted Models**

3. Model drift.

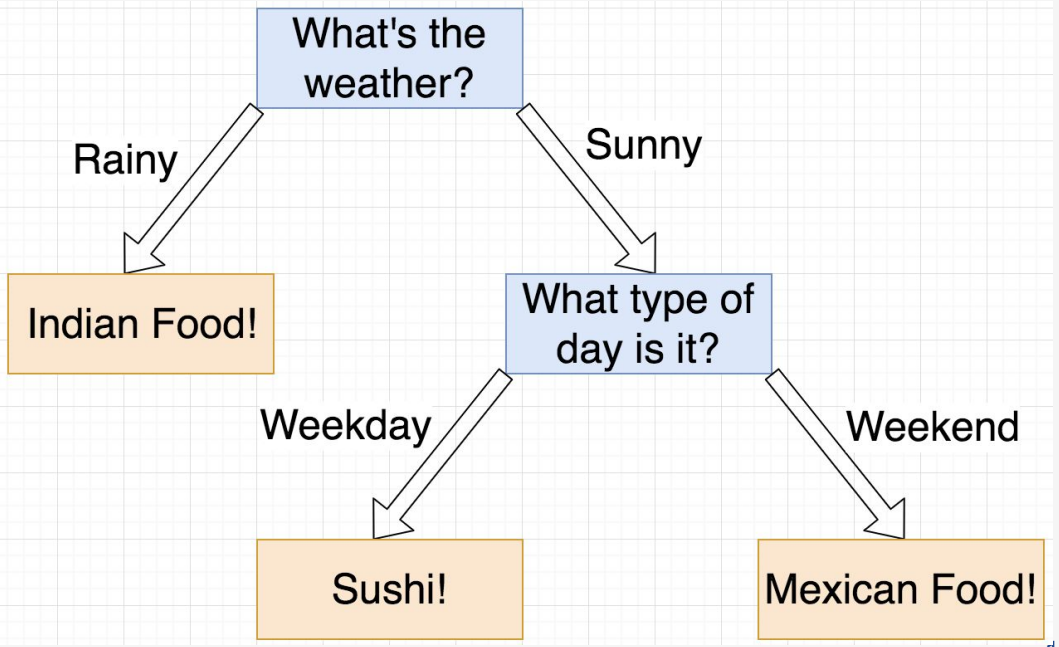
4. *Pending time, an AMA! (Other ML / data science questions, running a consultancy, my experiences being LGBTQ+ in tech, and more.)*

What will I eat tonight? Using a CART (decision tree).

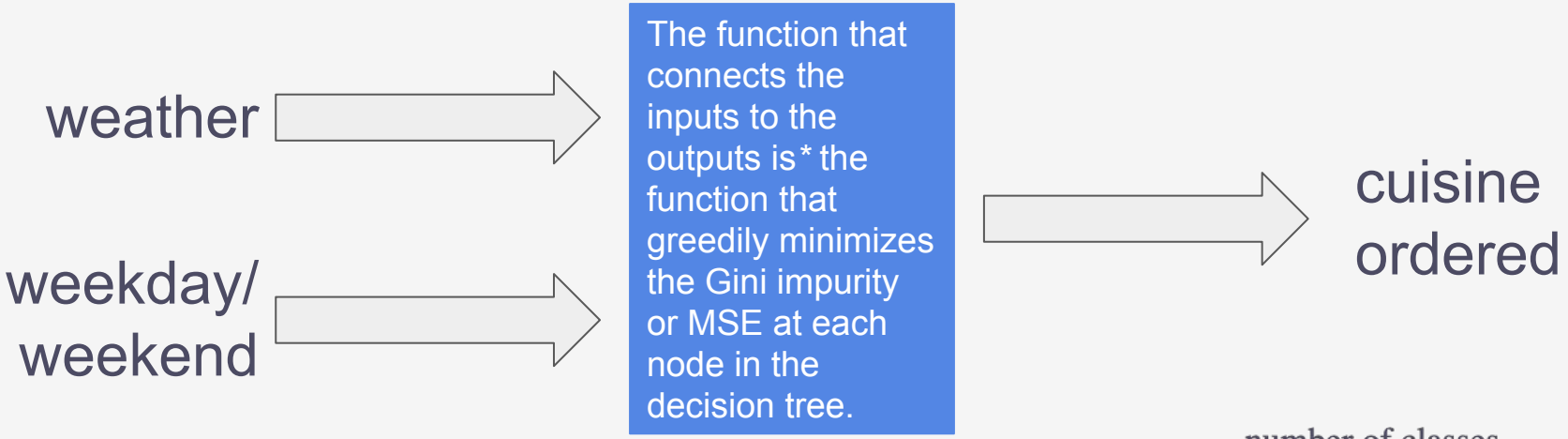
$Y =$ **Food**	$X_1 =$ **Weather**	$X_2 =$ **Day**
Indian	Rainy	Weekday
Sushi	Sunny	Weekday
Indian	Rainy	Weekend
Mexican	Sunny	Weekend
Indian	Rainy	Weekday
Mexican	Sunny	Weekend

What will I eat tonight? Using a CART (decision tree).

$Y =$ **Food**	$X_1 =$ **Weather**	$X_2 =$ **Day**
Indian	Rainy	Weekday
Sushi	Sunny	Weekday
Indian	Rainy	Weekend
Mexican	Sunny	Weekend
Indian	Rainy	Weekday
Mexican	Sunny	Weekend



You know what isn't exact? The real world.



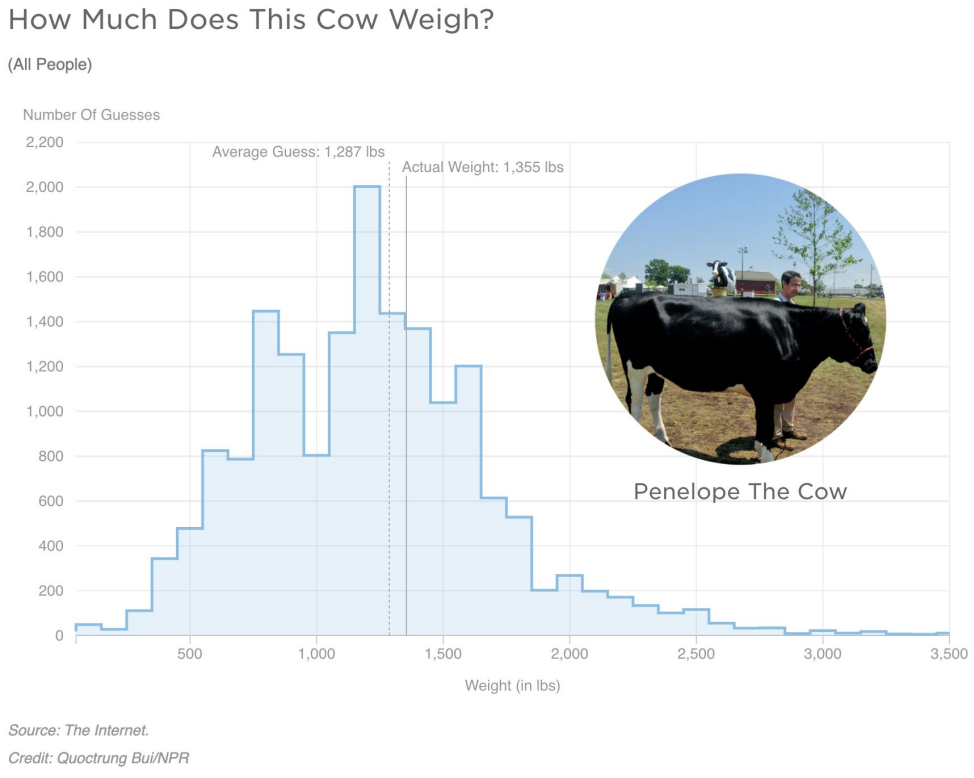
$$\text{Gini impurity} = 1 - \sum_{i=1}^{\text{number of classes}} p_i^2$$

$$\text{mean squared error} = \sum_{i=1}^{\text{n = sample size}} (\hat{y}_i - y_i)^2$$

What is the wisdom of the crowd?

How much does Penelope weigh? *(No cheating!)*

The crowd of people, on average, do better than the vast majority of people.



Average Guess: 1,287 pounds

Actual Weight: 1,355 pounds

Difference: 68 pounds (~5%)

https://www.npr.org/sections/money/2015/08/07/429720443/17-205-people-guessed-the-weight-of-a-cow-heres-how-they-did

This works for models, too!

Ensemble models are quite literally just an ensemble (group) of models.

Bagged Decision Trees, Random Forests, Adaboost Models, and Gradient Tree Boosted Models are types of ensemble models.

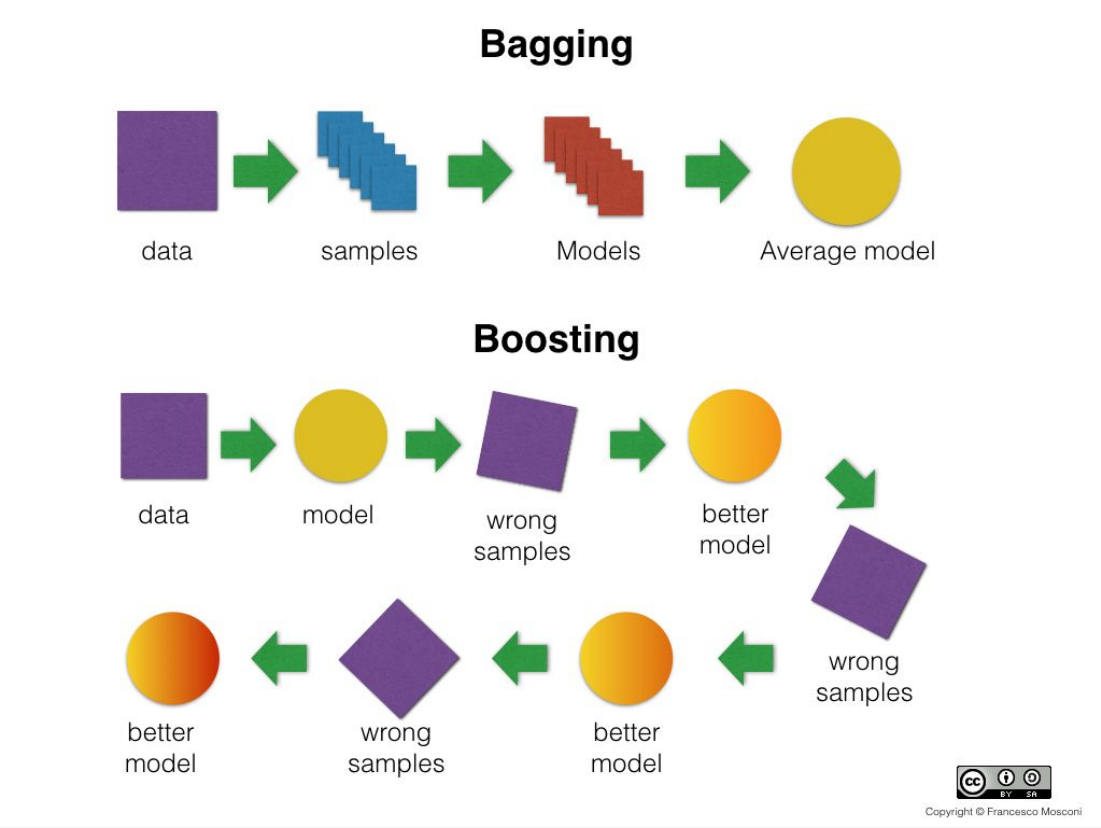
This is a non-exhaustive list above. Theoretically, you could build lots of *different types of models* and aggregate their predictions together.

For example, you could build a logistic regression model, a random forest, and a support vector machine, then combine their predictions together.

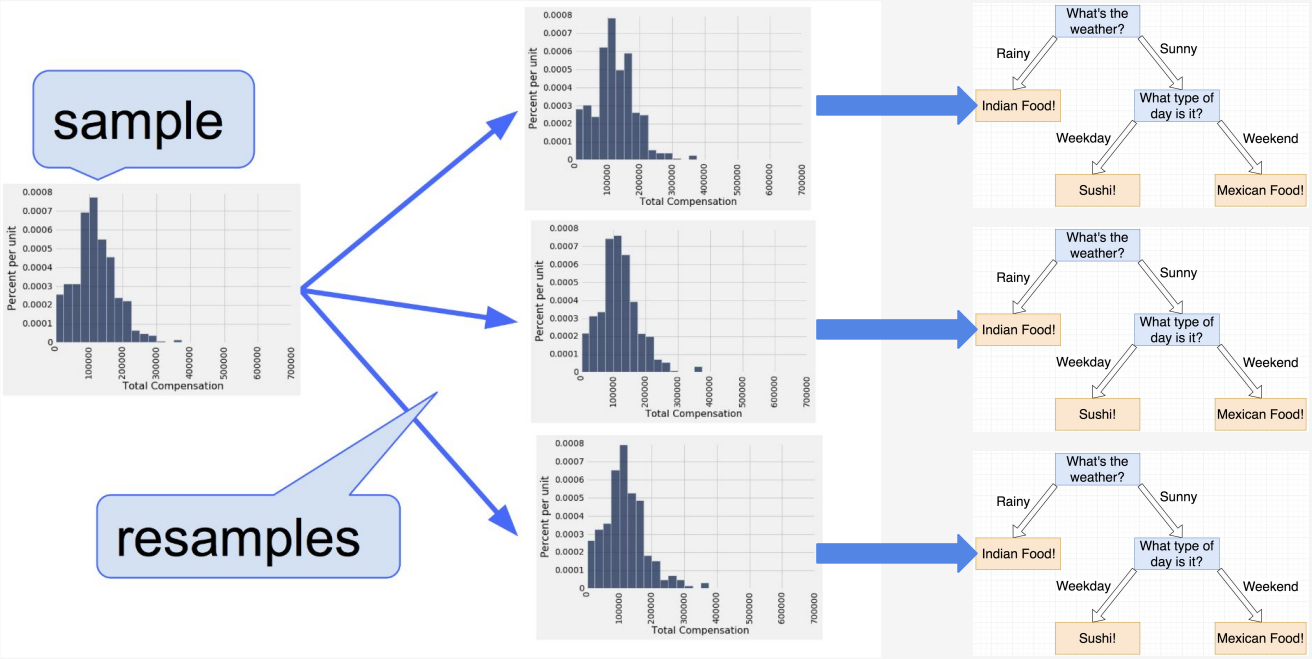
We often break ensemble models down into two common types: bagging models and boosting models

Bagging: build many good models in parallel, then aggregate models together.

Boosting: build simple models sequentially, adjusting for improper predictions.



Bagging: bootstrap aggregated (decision trees)



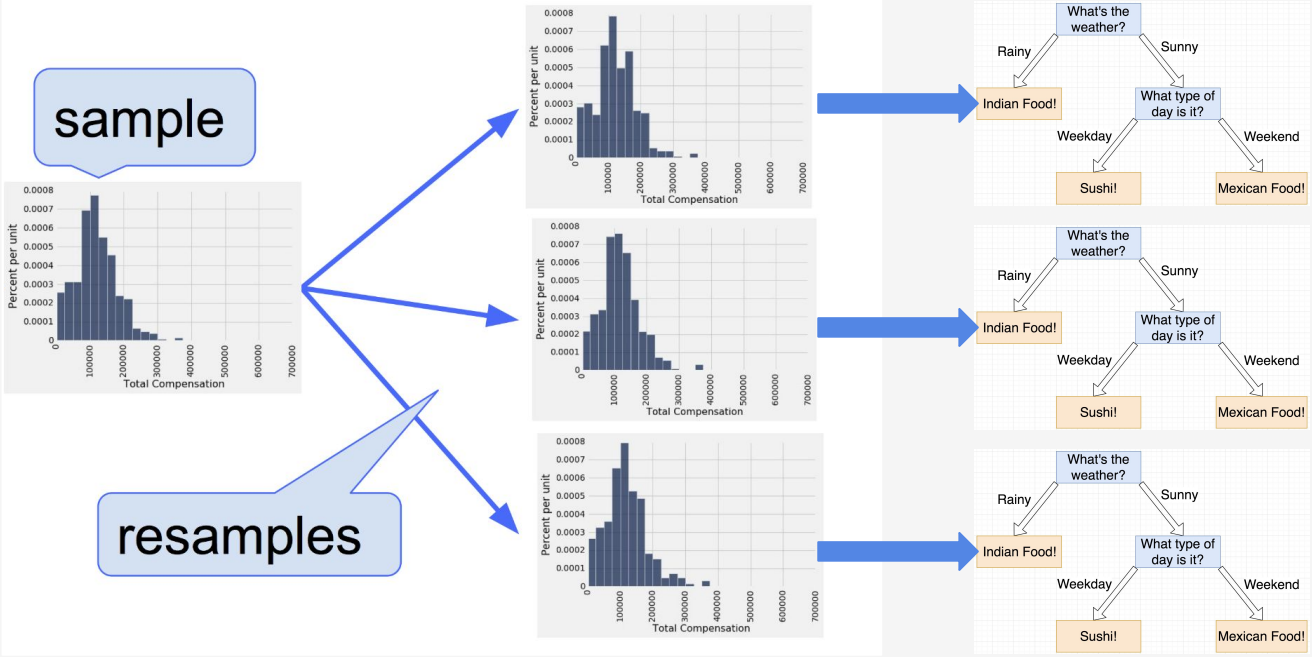
Every night, instead of turning to one decision tree to make my decision, I'll harness the "wisdom of the crowd" and ask all of my decision trees!

The majority (or plurality) vote wins!

Random forests: *bagging plus an extra source of randomness*

Bagged decision trees can suffer from high error (due to variance), because your trees are likely very similar.

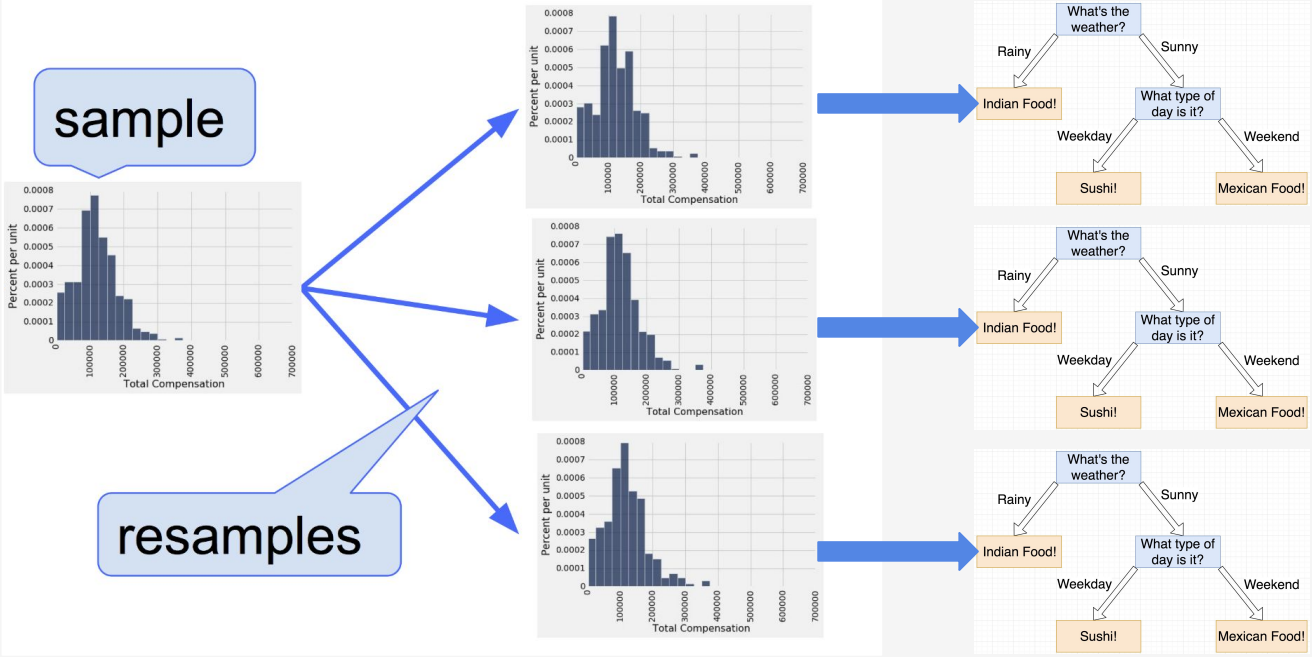
Random forests bootstrap samples **AND** randomly select a subset of features to split on at a given node.



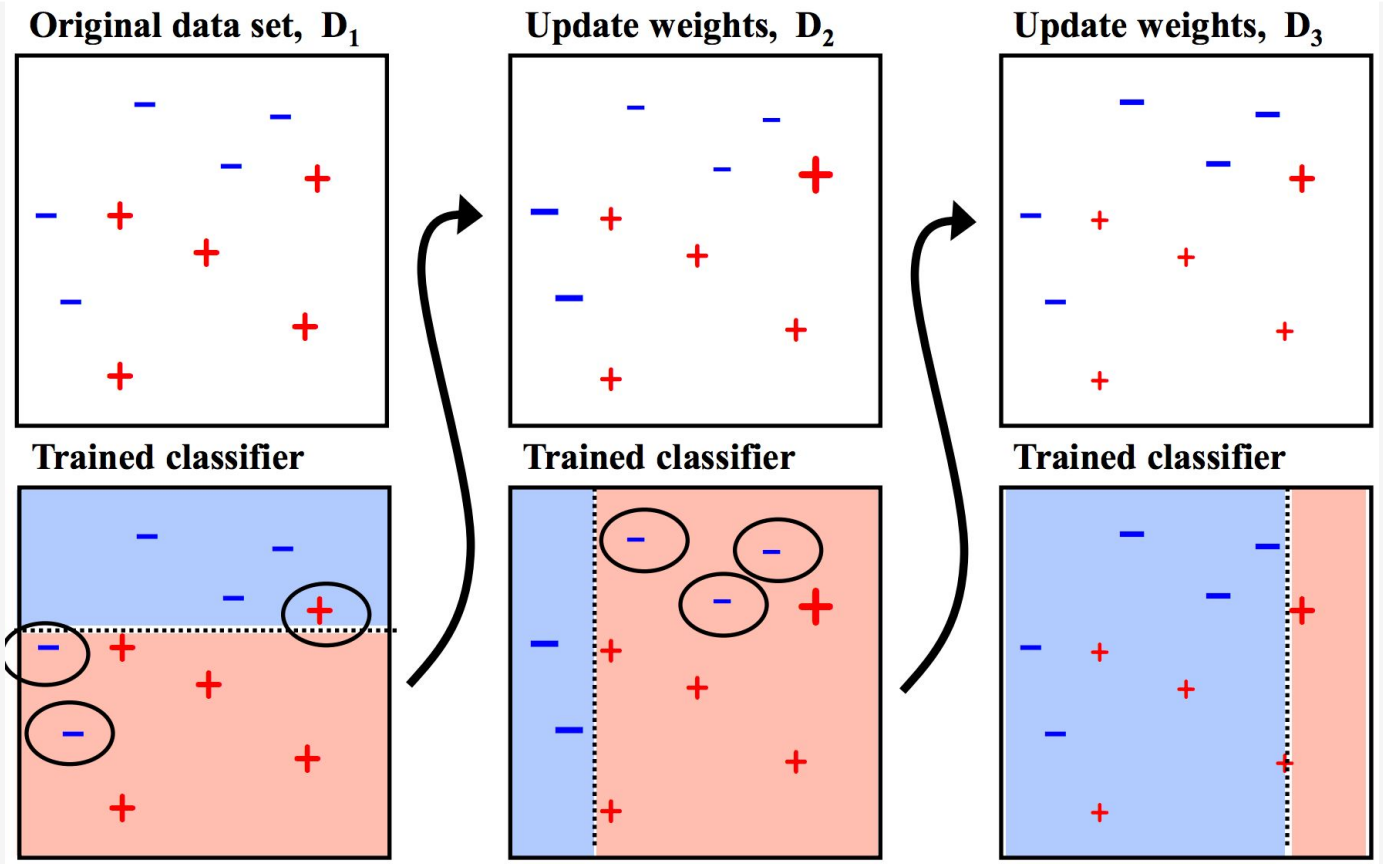
Extremely Randomized Trees: *random forests plus (you guessed it) another source of randomness!*

Random forests *may* suffer from high error due to variance. Extra randomness can help fix that.

ExtraTrees bootstrap samples **AND** randomly select a subset of features to split on at a given node **AND** randomly select splits in nodes.



Boosting: build simple models sequentially, adjusting for improper predictions.

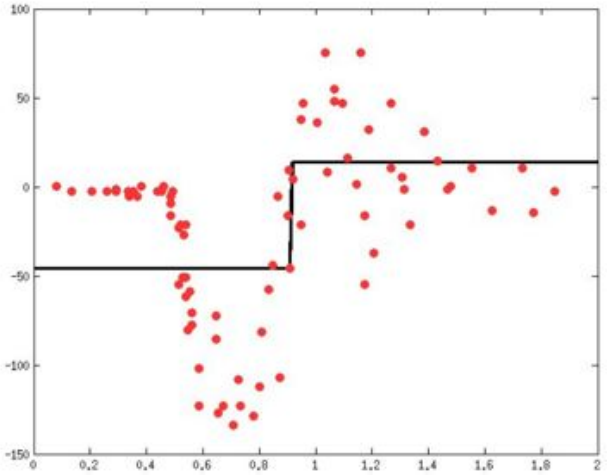


How do AdaBoost and Gradient Boosting differ? When fitting subsequent models...

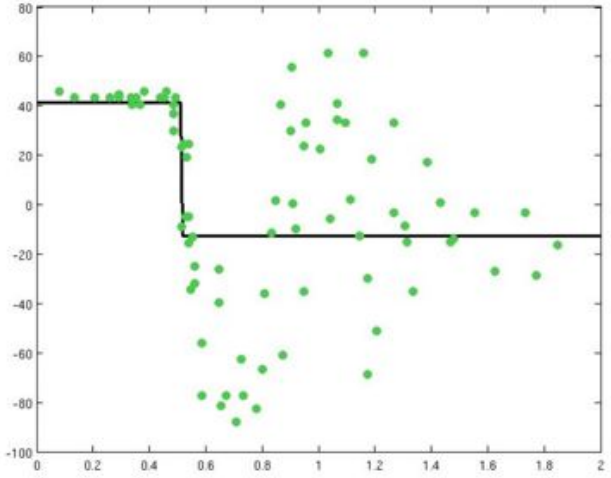
AdaBoost will reweight improperly predicted observations.

Gradient Boosting will fit a new model directly to the errors of the last model.

Learn a simple predictor…



Then try to correct its errors



Downsides of boosting

1. It can be very challenging to tune hyperparameters correctly.

2. Boosting models **must** be run sequentially. This takes significant time, because it means that calculations cannot be parallelized. *(Not great for larger datasets!)*

3. Hard to predict observations (often outliers) generally receive more and more influence in the final model fit.

To the notebook!

`00_ensemble_methods.ipynb`

Agenda

1. What is machine learning?

2. Ensemble models.
   a. Bagged Decision Trees
   b. Random Forests
   c. Adaboost Models
   d. Gradient Tree Boosted Models

3. **Model drift.**

4. *Pending time, an AMA! (Other ML / data science questions, running a consultancy, my experiences being LGBTQ+ in tech, and more.)*

Your model *(likely)* won't be great forever!

- A Fortune 500 company wants to estimate customer lifetime value. (How much will a customer spend in the next X months?) They use 2017-2018 data, evaluate their model on 2019 data, then deploy their model (e.g. put their model into production so that people can use it for decision-making purposes) in January 2020.

- A company specializing in tech bootcamps wants to predict time-to-employment for their graduates. They take all graduates from their first two cohorts of students, build a regression model, and use that to better serve students in future cohorts.

This is model drift.

Often used interchangeably: model drift, model shift, concept drift, population shift, data drift, model decay, model performance degradation.

This all boils down to the fact that even if you evaluate your model properly when training (e.g. use a training dataset, a validation dataset, a testing dataset, and prevent any leakage among these), **your model may perform worse upon deployment and over time**.

In fact… we kind of expect it to.

What can we do?

1. Accept that model drift is almost entirely unavoidable.

2. Consider holding out deployment until new data (*truly* never-before-seen data) is gathered.

3. Monitor your model's performance. Use the evaluation metric from your testing set as a starting point and track performance over time. If the model degrades more than [*insert acceptable threshold here*], then raise flags and/or pull the model out of production.

Good, quick, easy read: https://c3.ai/glossary/data-science/model-drift/

AMA! (pending time)

Feel free to ask me anything. A non-exhaustive but perhaps helpful list of topics:

- starting a data science consultancy/doing your own freelance work
- working for a startup
- technical interviews/negotiating (or even some technical content)
- my experiences as an LGBTQ+ community in tech
- my previous work experiences (includes consumer-packaged goods, politics, education, finance, computer vision)

Thank you!

LinkedIn: Matthew Brems

Twitter: @MatthewBrems

Github: MatthewBrems

Email: matt@betavector.com