Ensemble Modeling

Toronto AI Summit | November 2017

Helen Ngo

Agenda

- Introductions
- Why Ensemble Models?
- Simple & Complex ensembles
- Thoughts: Post-real-life Experimentation
- Downsides of Ensembles
- A Model-agnostic Methodology for Interpretability

Hello!

About Me

- Data scientist
- Telecommunications industry
- Mathematics background
- Coffee enthusiast
- Women in STEM

www.linkedin.com/in/helen-ngo/

helen.ngo14@gmail.com

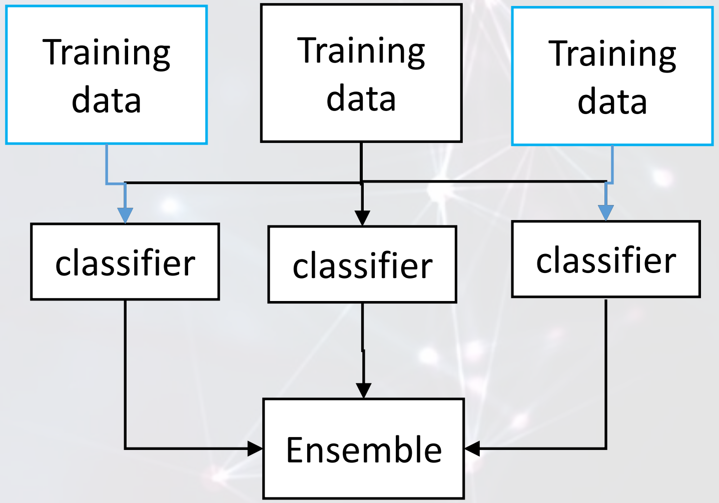
What’s more accurate than one great model?

- Sometimes, a lot of not-too-shabby models
- **Ensemble modeling**: learning algorithms to combine classifiers by weighting their predictions
- Lots of models; diverse algorithms; uncorrelated predictions
  - You’re more likely to be right most of the time
  - Ensembles only more accurate than individual models if the models disagree with each other

The simple ones

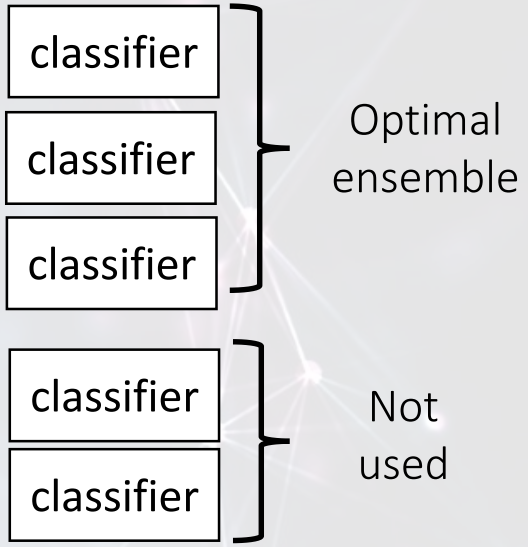
- Simple average, voting
- Bagging: sample with replacement
  - Random forests: random subset of features for many trees trained in parallel
- Boosting: iteratively reweight misclassified observations in training
  - XGBoost, AdaBoost

Anatomy of an ensemble model

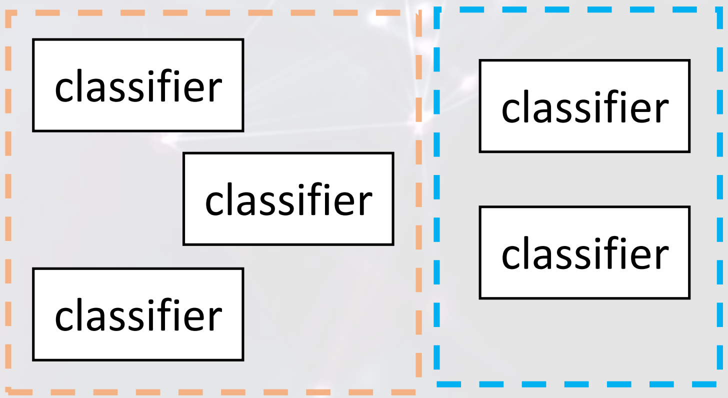


Slightly more sophisticated

- Top-T: for N models and t <= N,
  - Take best T models according to your accuracy measure of choice
  - Use validation data to select optimal value for T
  - More is not always better!

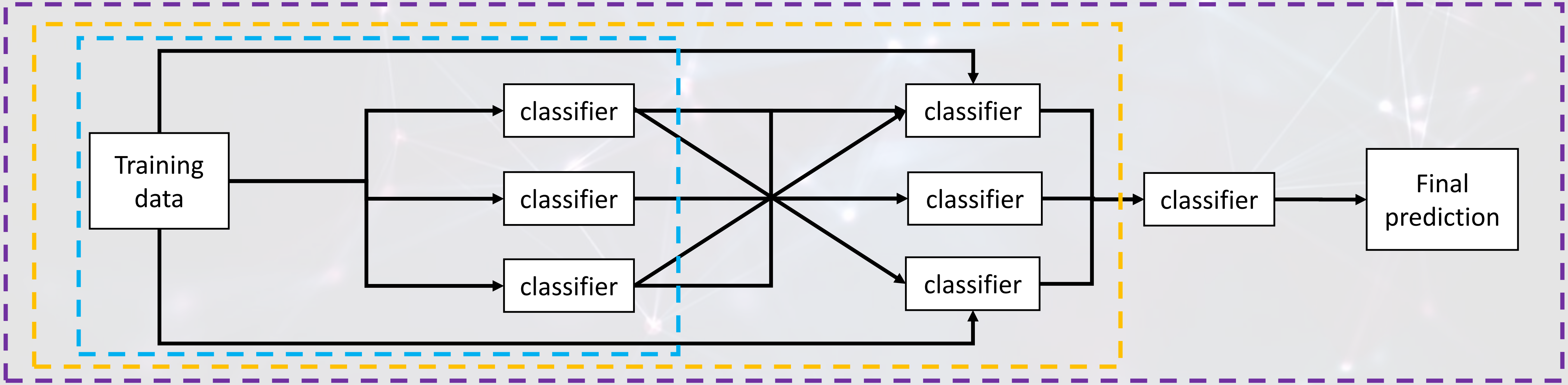


- Unsupervised cluster ensembling
  - Use PCA to assign probabilities from N models to clusters based on original features
  - Use the models in cluster with top-T method



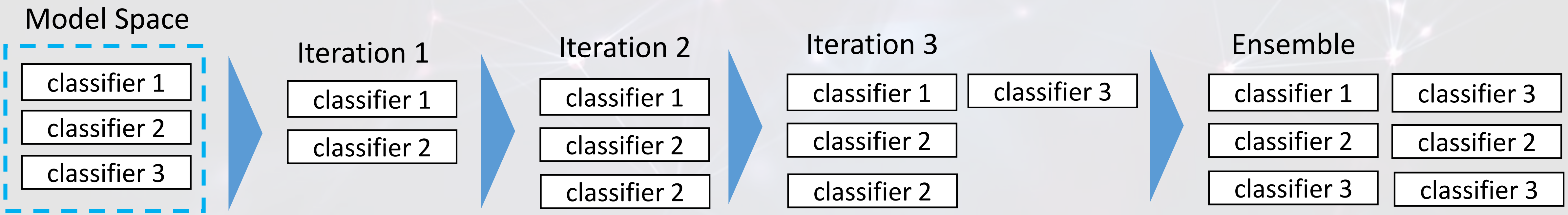
Popularized by our favourite data science competition

- Stacking & blending
  - Use posterior probabilities from trained models as numerical inputs to model original target variable
  - Can have several stacked layers
  - Stacking functions can be any supervised learning method



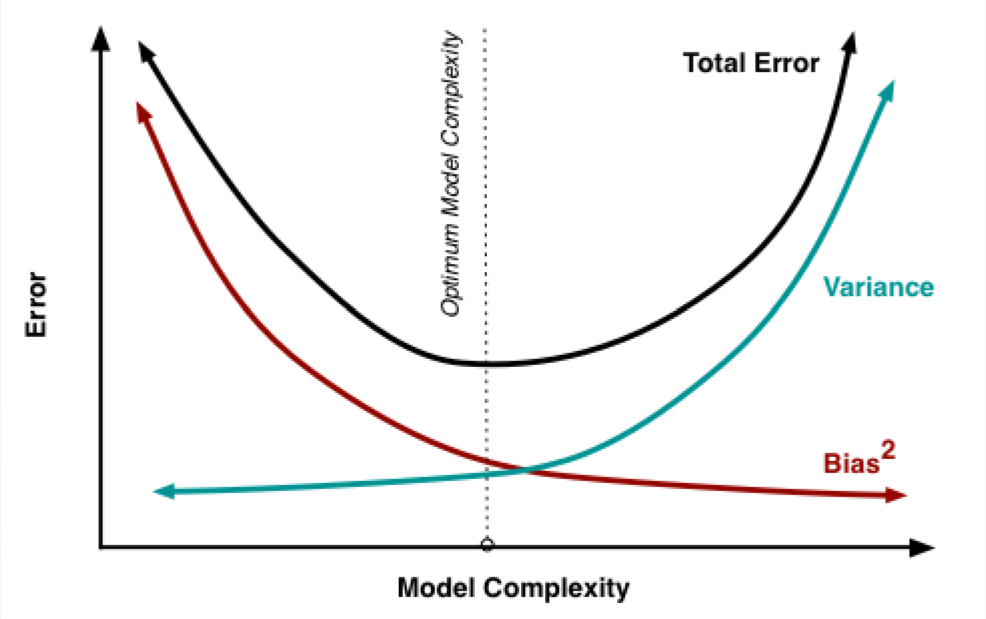
On the shoulders of giants

- Hill climbing
  - Rank models by some accuracy measure and calculate the incremental ensemble performance by adding one at a time
  - Greedy algorithm to choose next model to add
  - Final ensemble chosen based on overall accuracy metric
  - Models can be added multiple times and weighted differently; powerful models can be added many times



The reason why ensembles work

- Bias-variance error decomposition
  - Bias (underfitting) & variance (overfitting) traded off
  - More complexity → more overfitting
- What if we have a bunch of low-bias, high-variance models?
  - Ensemble them all
    → same bias, lower variance
  - Total error is lower!
  - Less correlation is better (higher reduction in variance)



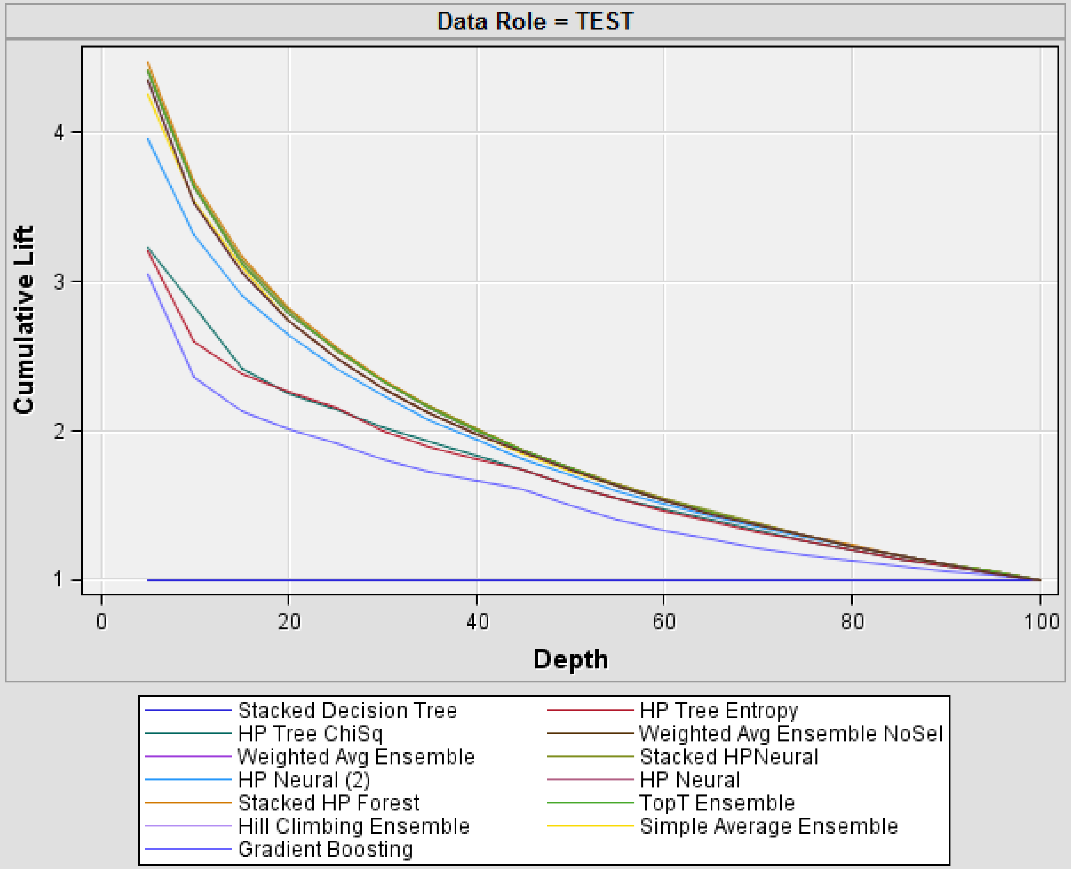
Source:
http://scott.fortmann-roe.com/docs/BiasVariance.html

Thoughts: post-real-life experimentation

- Success! Overcame “noisy” marketing data with ensembles

- Best ensembles created out of weak learners without many transformations

- Ensembles don’t necessarily win over simpler models—depends on your use case



The best ensemble models outperform the best non-ensemble significantly – 3.7x vs. 3.4x lift

Downsides of ensembles

- Search space for best ensemble is large
  - Time-consuming to train
- Feature impact on final score is unclear
- Lack of explanation for customer-facing decisions
  - Hard to sell to a business decision maker

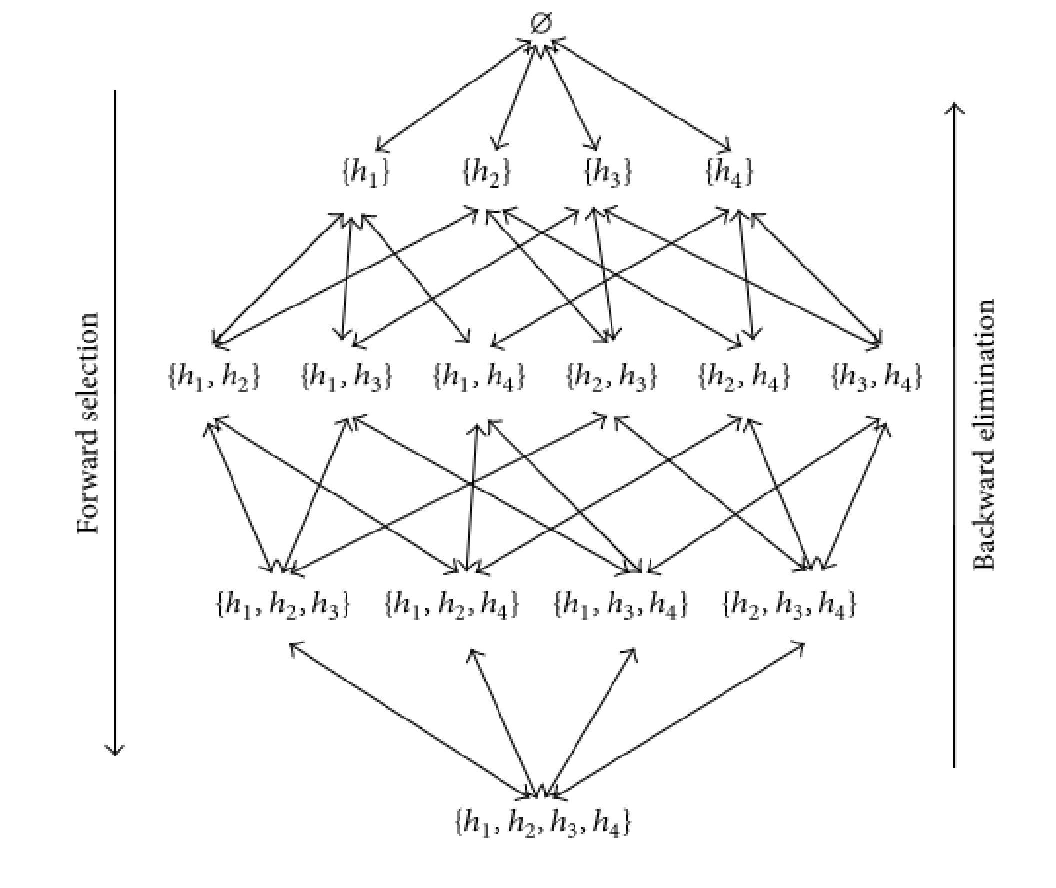

$\varnothing$
$\{h_1\}$ $\{h_2\}$ $\{h_3\}$ $\{h_4\}$
$\{h_1, h_2\}$ $\{h_1, h_3\}$ $\{h_1, h_4\}$ $\{h_2, h_3\}$ $\{h_2, h_4\}$ $\{h_3, h_4\}$
$\{h_1, h_2, h_3\}$ $\{h_1, h_2, h_4\}$ $\{h_1, h_3, h_4\}$ $\{h_2, h_3, h_4\}$
$\{h_1, h_2, h_3, h_4\}$
Forward selection
Backward elimination

Figure 1: The search space of DHCEP methods for an ensemble of 4 models.

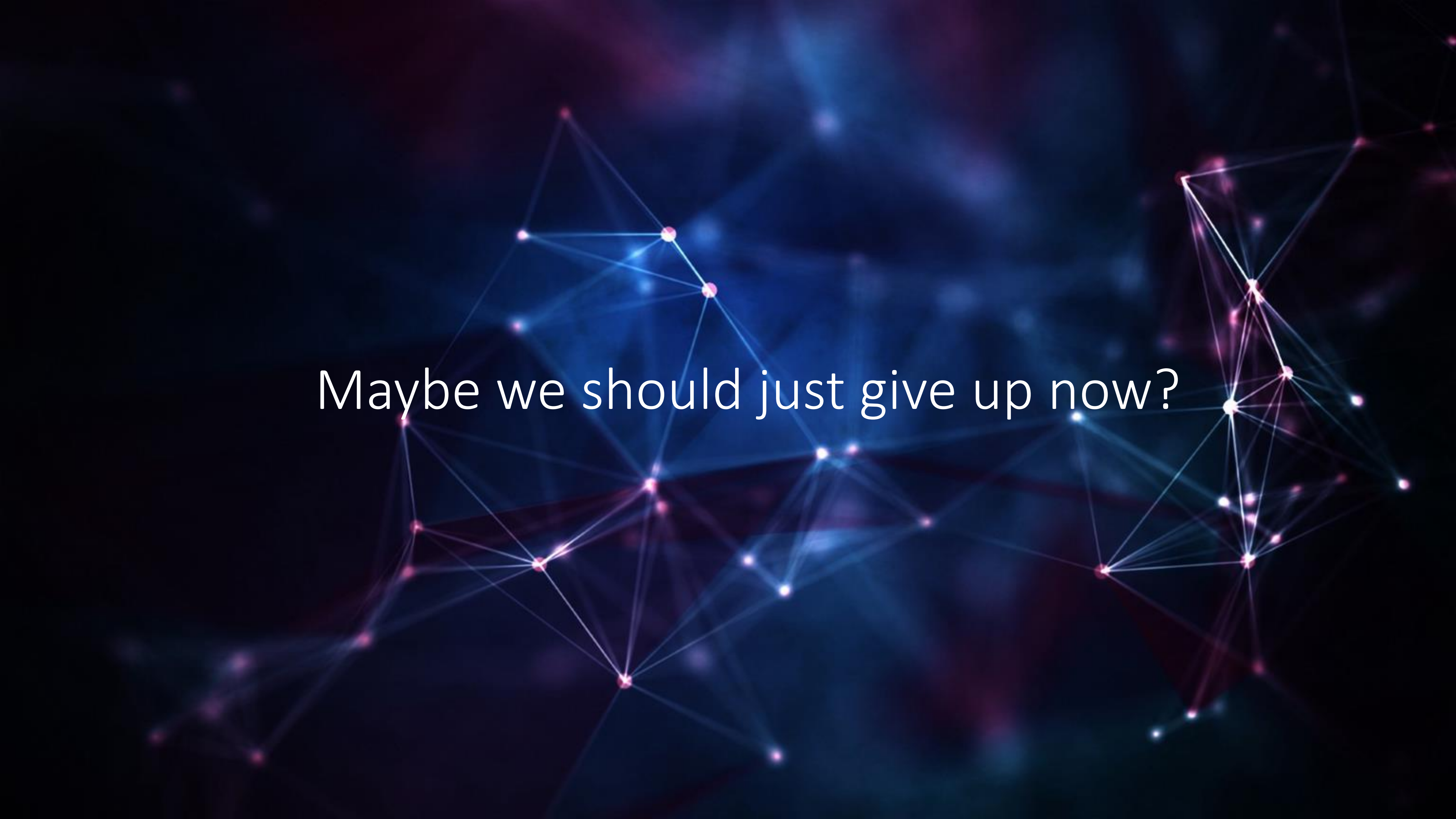
Source: https://www.hindawi.com/journals/mpe/2016/3845131/fig1/

Why does explanation matter?

- EU General Data Protection Regulation (GDPR) 2018
  - All EU citizens will have a “right to explanation” from vendors using their information, even if they are based outside the EU

- Cool ML products are redundant if they’re not adopted
  - Fear of the unknown outside of the ML/AI community
  - Exhibit A: self-driving cars

- Ethical machine intelligence
  - Weapons of Math Destruction: The Dark Side of Big Data – O’Neil (2016)

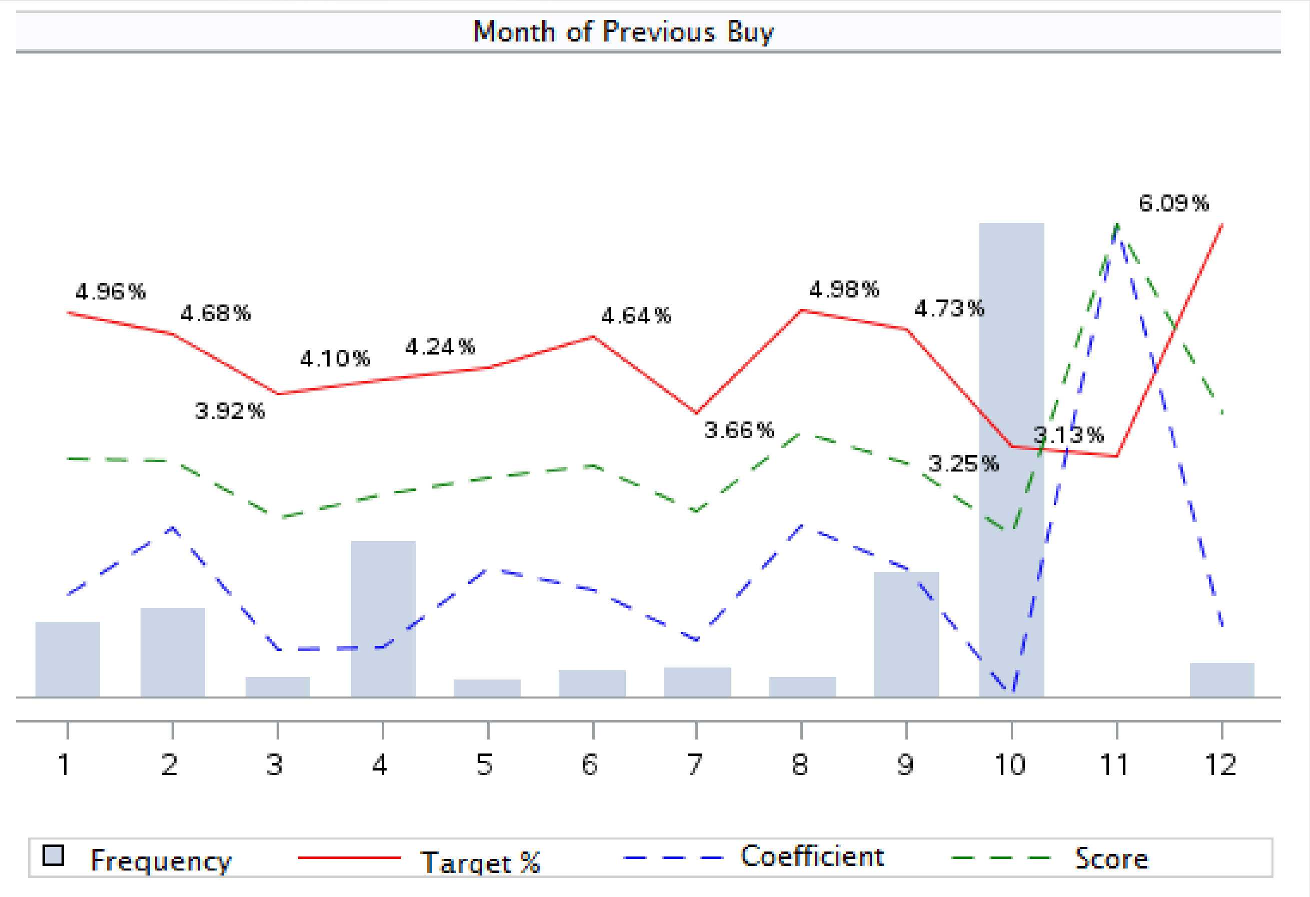


Maybe we should just give up now?

A model-agnostic methodology for interpretability

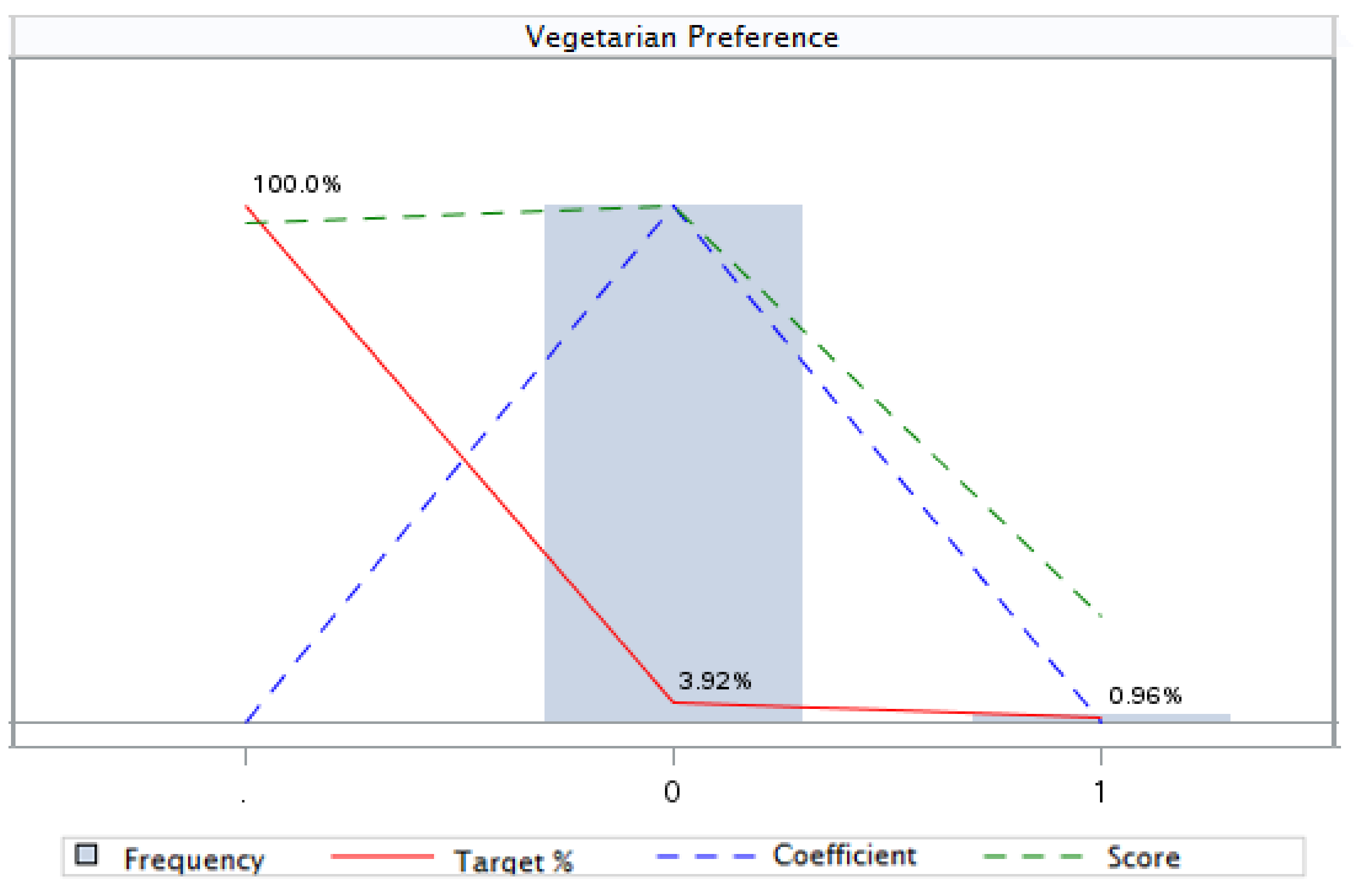
- Key insights
  - What is the impact of each feature on the score?
  - Plots to the rescue!
  - Bin numeric variables

- Variable importance
  - If we set each variable to be a constant, what is the negative impact on model performance?



Plotting also reveals poor model fit

- This feature is less great
- We should probably take it out of the model
  - Or transform it
- Domain knowledge
  - Skewed training set
  - Seasonality



Looking into the black box

- Surrogate modeling – “modeling the model”
  - Use original predictors to model the *scores* output by the ensemble
  - Decision trees recommended for capturing non-linear relationships
  - Statistical representation of model behaviour
  - Overfitting is okay!

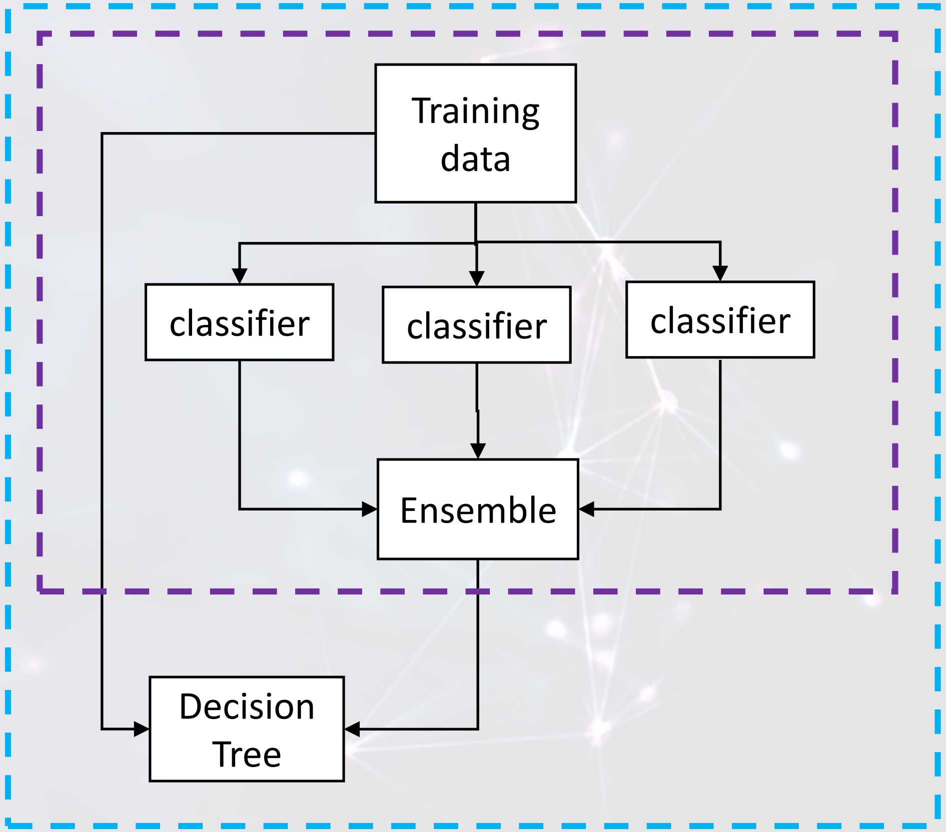
- Other neat methods
  - LIME (Locally Interpretable Model-Agnostic Explanations)
    “Why Should I Trust You?” (Ribeiro, Singh, Guestrin 2016)
  - Model Distillation (Tan, Caruana, Hooker, Lou 2017)

Questions?

Appendix: Step-by-step surrogate modeling

Post ensemble-creation

- Use the original features as inputs
- Instead of modeling the original target, model the scores generated
- Maximally overfit a wide decision tree to predict the scores
- Leverage the decision tree in model-agnostic explanation method


Training data
classifier
classifier
classifier
Ensemble
Decision Tree