Carnegie Mellon University
Tepper School of Business

46-886: Machine Learning Fundamentals

Amr Farahat

Interpretable Machine Learning

Note:

- No office hours this coming Saturday, April 8th, (due to travel) Please email me for alternative times.
- Deliverable 3 posted tomorrow by 5:00 p.m. (work with same teams)

Some “Interpretable” Models

- **Linear / Logistic Regression**
  - Linear regression:
    - The notion of a weighted linear combination of features as a predictor is common and intuitive
    - The impact of each feature is separable and quantifiable using its coefficient
    - Even log-log or semi-log models have simple interpretations in terms of % change
  - Logistic regression:
    - Probability thresholds translate to “score” thresholds. The scores are weighted linear combination of features
    - Interpreting the coefficients is a bit tricky…

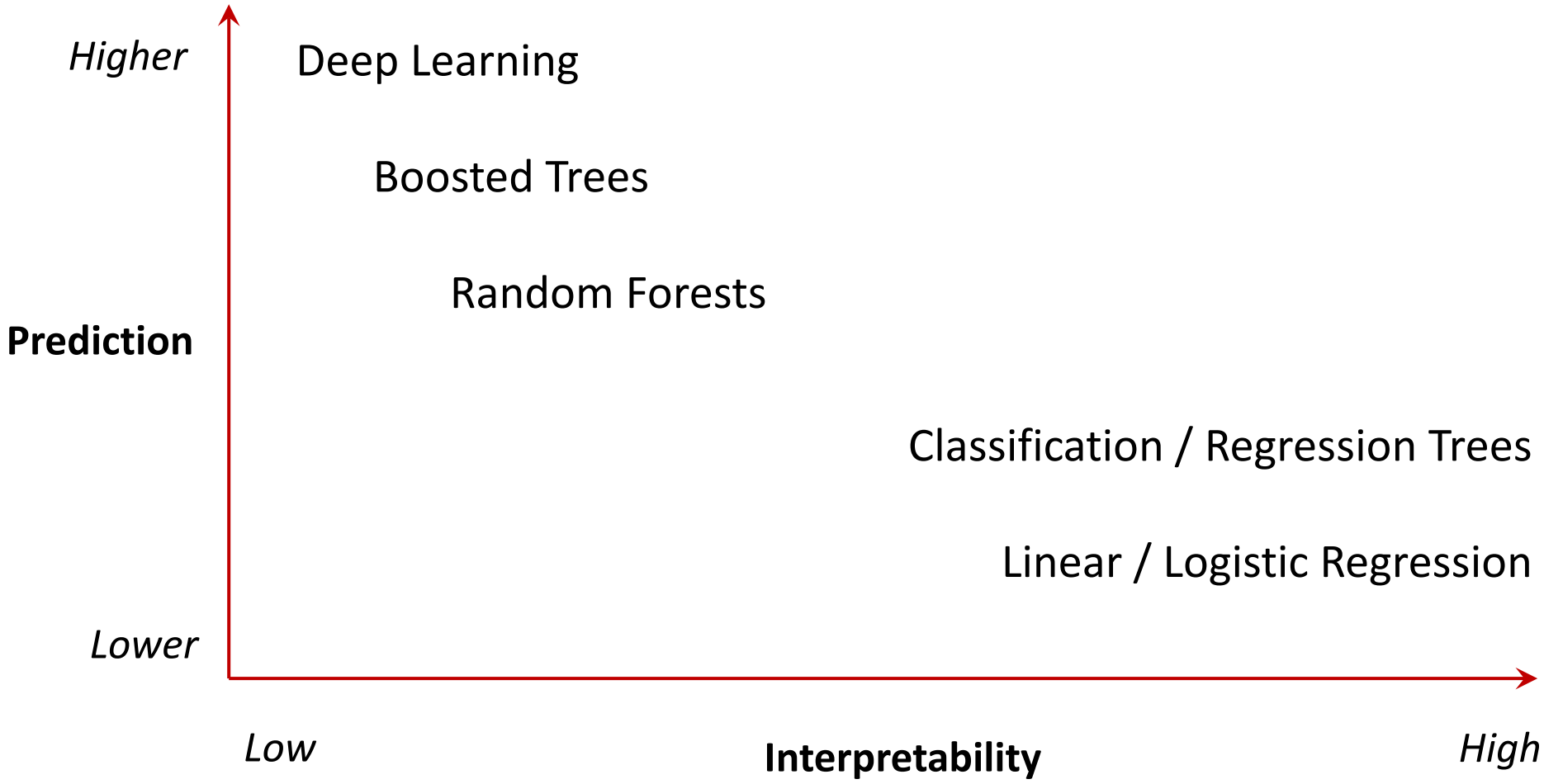
Some “Interpretable” Models

- **Classification / Regression Trees**
  - If – then statements/rules, coupled with a graphical representation, are intuitive and capture natural dependencies
  - “On interpretability, trees rate an A+” – Leo Breiman
  - However:
    - We’ve seen how complicated trees can be once they exceed 3 or so levels
    - The exact threshold values of branch node can be perplexing – best to interpret these nodes as high vs. low
    - Handling of categorical variables in Python through dummy variables is not ideal

Less Interpretable Models

- Random Forests
- Boosted Trees
- Neural Networks

Tradeoffs between Prediction and Interpretability



Discussion

- What exactly is “interpretability”?
- Why does it matter?
- Do you have examples where interpretability was an obstacle to leveraging analytics in practice?

Some thoughts

"Interpretability"

=

How well do we understand / explain the inner workings of the model

+

How well do we understand / explain the outputs of the model

Some thoughts

1. Understanding / Explaining the inner workings of a model
   - "What I cannot create, I do not understand" – Richard Feynman
   - Good news: Good communication can go a long way…
     - "If you can't explain it simply, you don't understand it well" – Albert Einstein (apocryphal)
   - Understanding / Transparency builds trust in the model and in the analyst!
   - Always attempt to do this:
     - Explain the data used to train the model and its (many) limitations
     - Explain the logic of the modeling framework chosen and how it's data-driven
     - Emphasize your metrics test the model on unseen data
     - Compare your chosen model's performance with other more interpretable alternatives

Some thoughts

2. Understanding / Explaining the outputs of a model

- Often, we have a responsibility to provide some reasonable justification of the factors underlying a given decision / prediction / classification
- Explaining the output, especially, surprising or edge cases can be helpful in doing sanity checks on the data and the model
- In some cases “The purpose of models is insight, not numbers” - Richard Hamming (with modification)

Some Tools that may Help Explain Model Outputs

- Surrogate models
- Partial dependence plots
- Feature importance metrics
- SHAP / Shapley values & plots

Reference

If you wish to read more:

- “Interpretable Machine Learning” by Christoph Molnar available online here: https://christophm.github.io/interpretable-ml-book/
- “The Mythos of Model Interpretability: In machine learning, the concept of interpretability is both important and slippery” by Zachary Lipton (2018). https://dl.acm.org/doi/10.1145/3236386.3241340