Actuarial & Finance

The Language Debate & Intro to Tree Based Algorithms

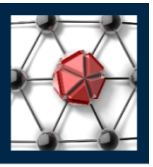
Joe Long
Assistant Actuary and Data Scientist
Milliman, Minneapolis

March 21, 2018

ILTCI

18th Annual Intercompany Long Term Care Insurance Conference

Limitations

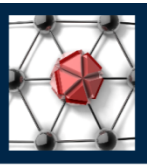


This presentation is intended for informational purposes only. It reflects the opinions of the presenter, and does not represent any formal views held by Milliman.

Milliman makes no representations or warranties regarding the contents of this presentation.

Milliman does not intend to benefit or create a legal duty to any recipient of this presentation.

Overview

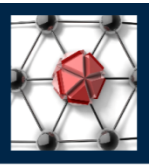


- The language debate: R vs SAS vs Python
  - Which tools should I learn?
- Building blocks of decision tree algorithms
  - Single decision trees, bagging, and boosting
- Opening the machine learning black box
  - What is driving the predictions?
- How to keep up as stuff evolves?
  - Review of educational resources

The language debate

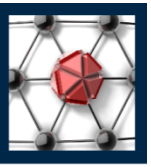
Source: https://www.analyticsvidhya.com/blog/2017/09/sas-vs-vs-python-tool-learn/

The language debate



- Things to consider
  - Statistical methods and techniques
  - Ease of learning
  - Support
  - Visualization
  - Costs
  - Scalability
  - Familiarity among coworkers

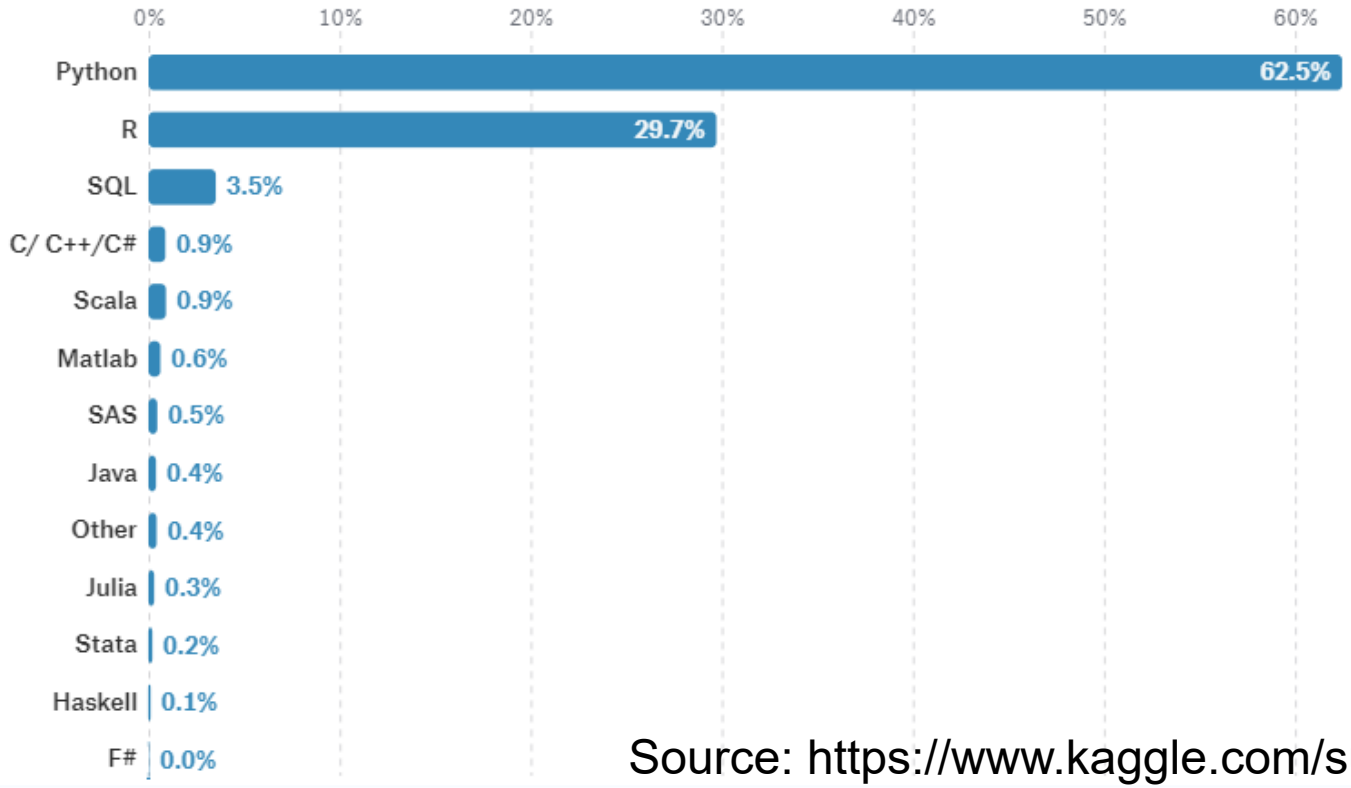
The language debate



What language would you recommend new data scientists learn first?

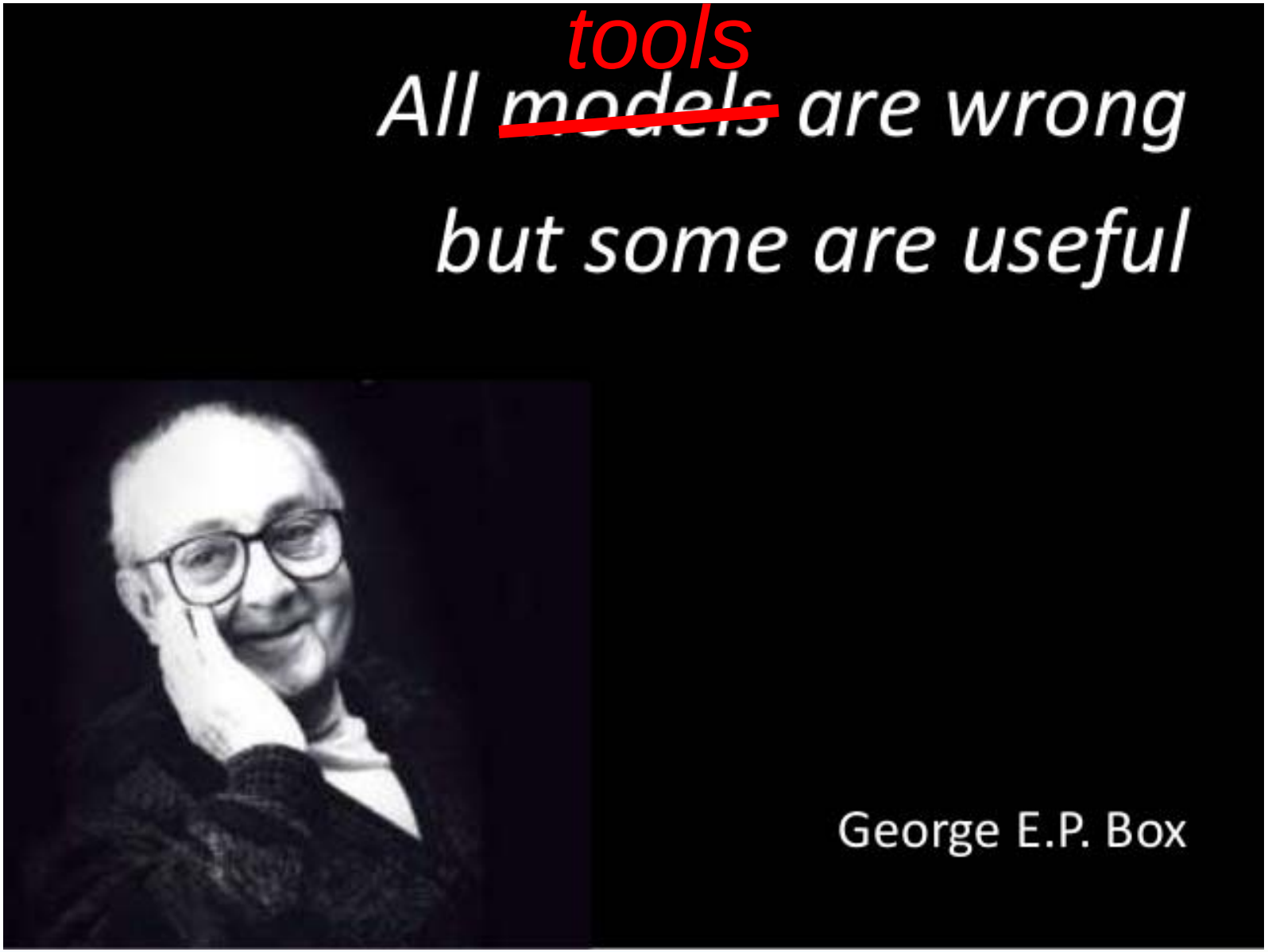
Everyone data scientist has an opinions on what language you should learn first. As it turns out, people who solely use Python or R feel like they made the right choice. But if you ask people that use both R and Python, they are twice as likely to recommend Python.

FILTER FROM USERS OF All Both Python R

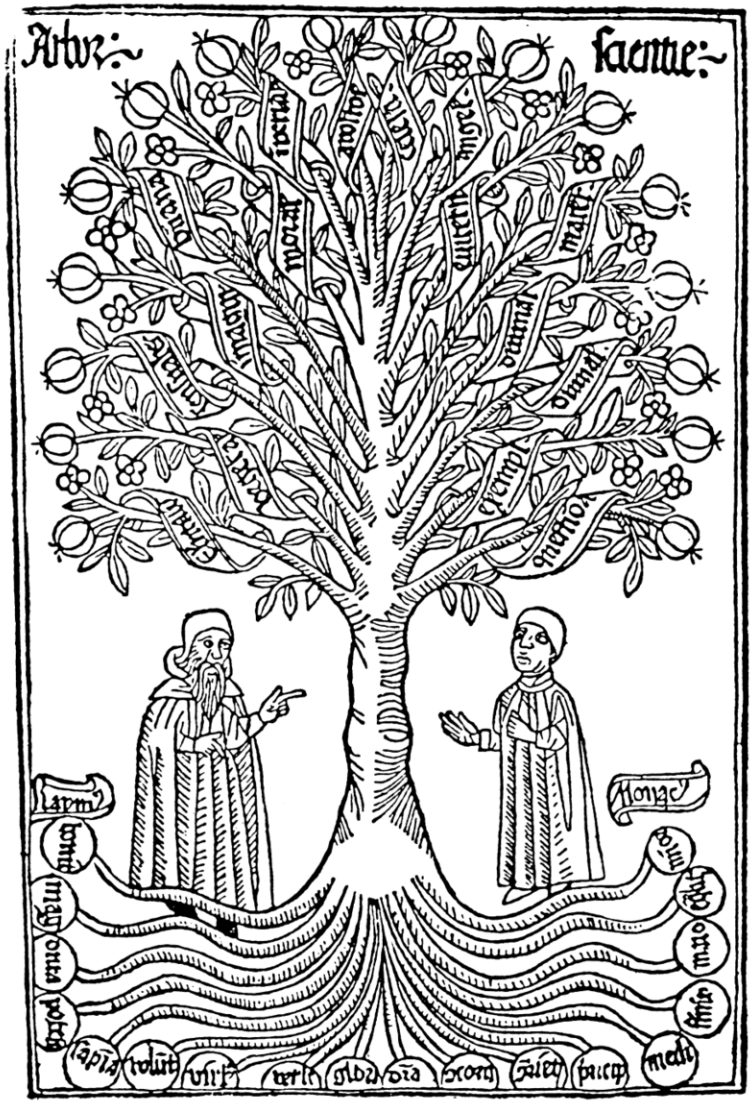


Source: https://www.kaggle.com/surveys/2017

The language debate

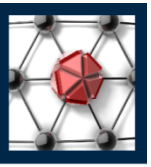


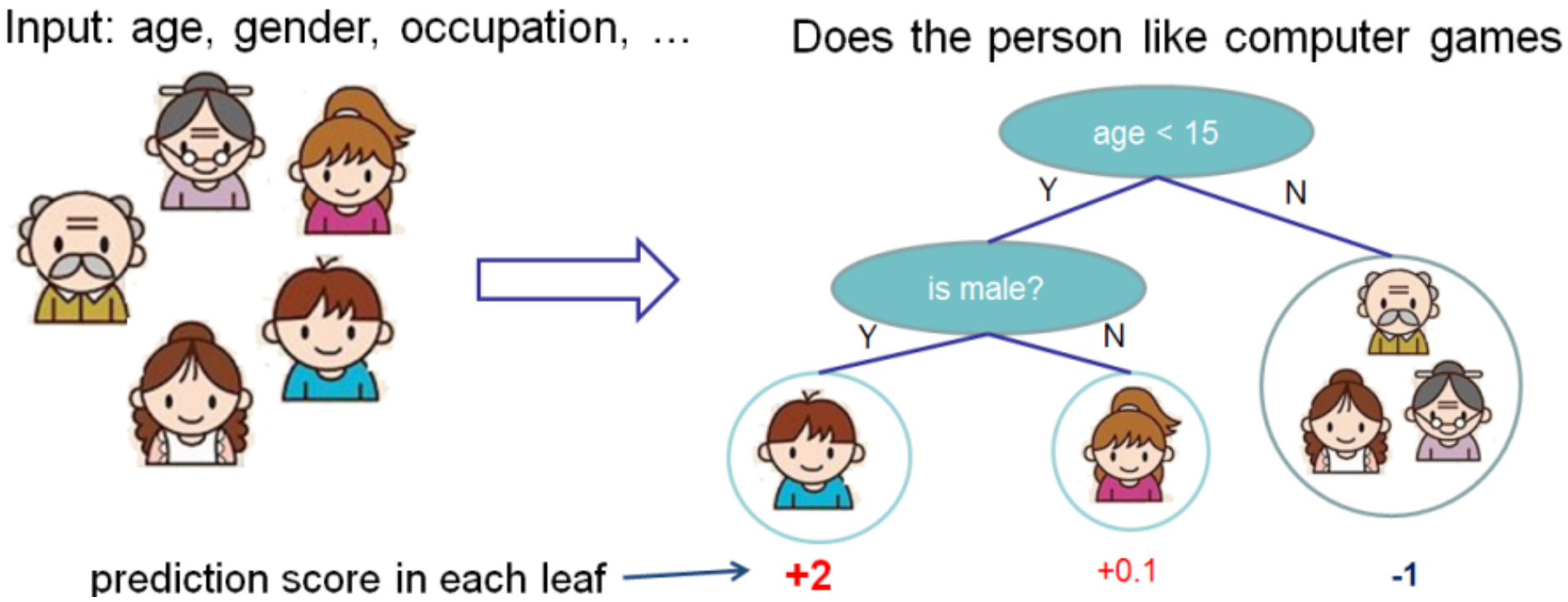
Intro to Tree Based Algorithms



Tree of Science (Arbre de la ciència, Arbor Scientiae) - Ramon Llull ~ 1295

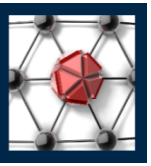
Decision Tree



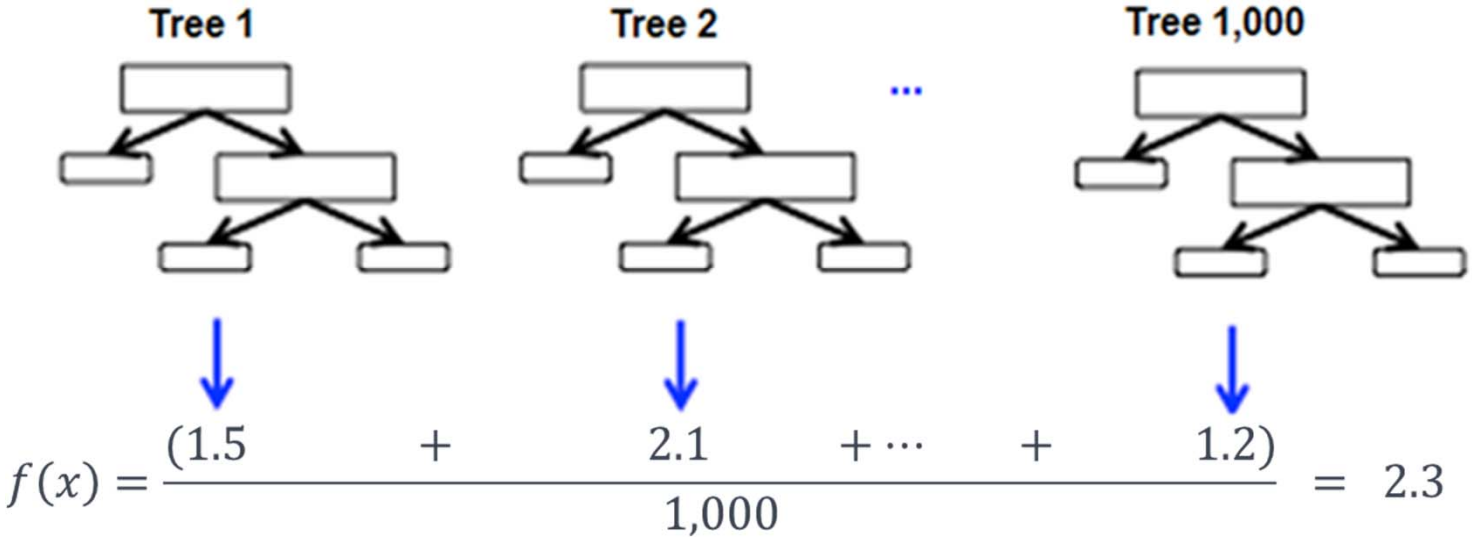


Source: http://xgboost.readthedocs.io/en/latest//model.html

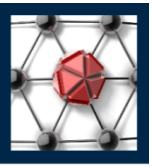
RandomForest™



- Bagging
  - Build 100's of trees independently
    - Random sampling of observations and features
  - Final prediction is the average across all the trees
    - Typically the more trees, the better the prediction will be

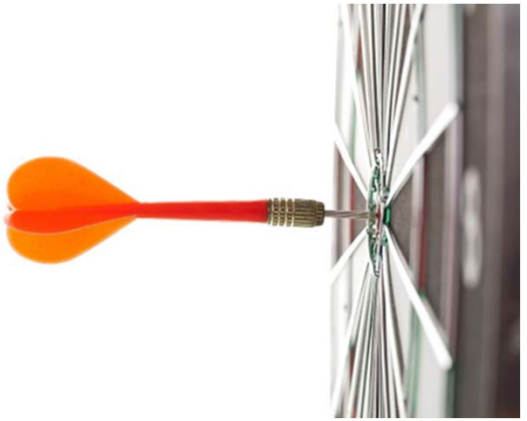


RandomForest™ Implementations



- RandomForest™, Breiman and Cutler
  - https://www.stat.berkeley.edu/~breiman/RandomForests/
  - R implementation – randomForest
- H2o - Distributed Random Forest
  - Interface through - H2o's API, Java, Scala, Python, R
- Python – sklearn
- SAS – PROC HPFOREST

RandomForest™



Pros

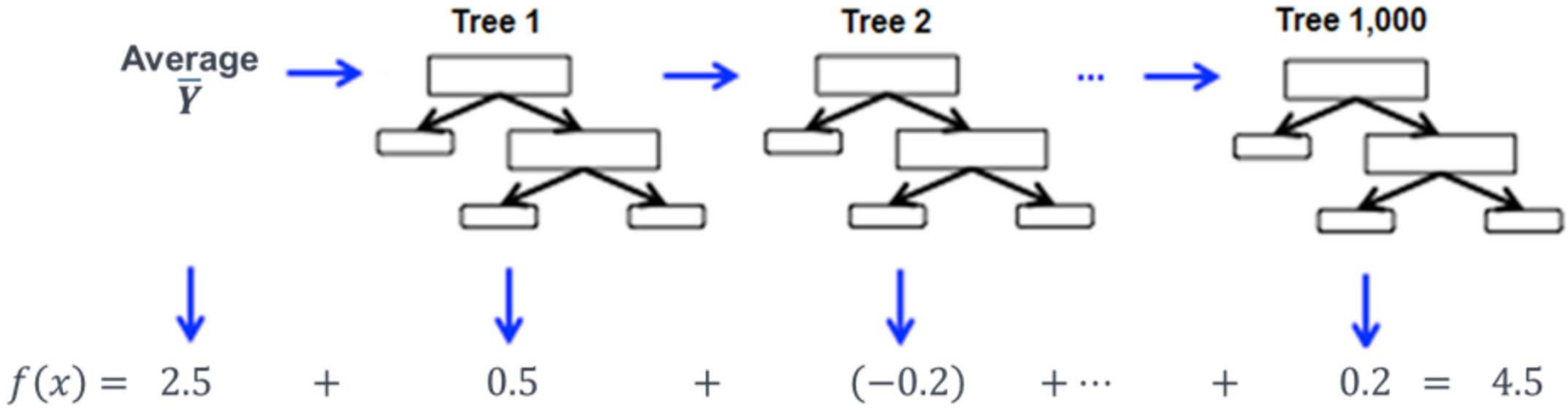
- Automatically creates interactions
- Better predictor than single decision tree

Cons

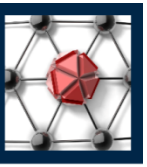
- Computationally expensive
- Black box harder to interpret
  - But there are clever ways to do so

Gradient Boosting Machine (GBM)

- Boosting
  - Build 100's of trees sequentially
  - Final prediction is the sum across all the trees
    - Too many trees can result in overfitting
    - Control for overfitting by tuning many hyperparameters



GBM Implementations



dmlc XGBoost

- Interface through - C++, Java, Julia, Scala, Python, R,

H_2O.ai

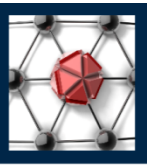
- Interface through - H2o's API, Java, Scala, Python, R

LightGBM

Microsoft

- Interface through – Python, R

Gradient Boosting Machine (GBM)



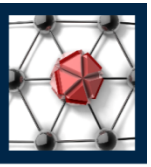
Pros

- Tends to have better performance than RandomForest™
  - Boosting may uncover relationships that RandomForest™ may miss out on

Cons

- Slower runtime because trees are built sequentially
- Many hyperparameters to tune to avoid overfitting
- Black box harder to interpret

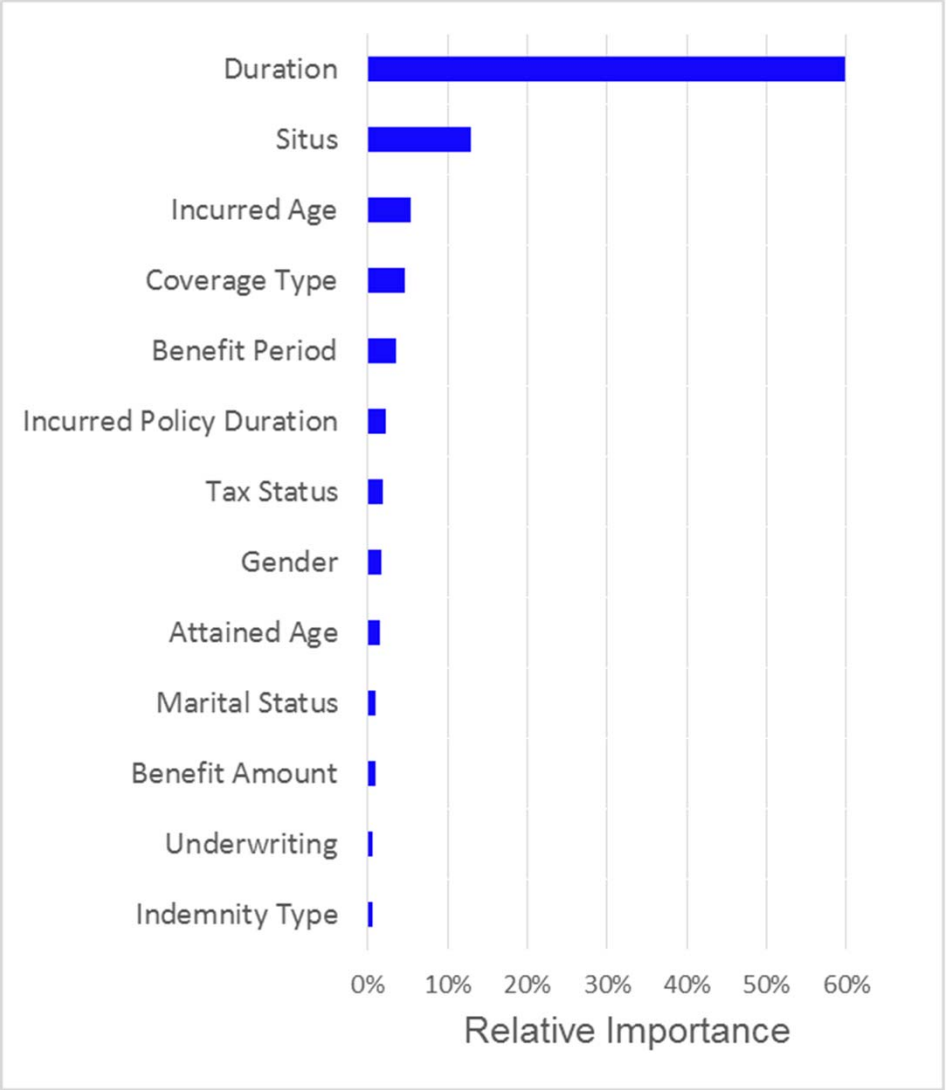
Opening up the black box



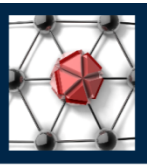
- Can we explain how our model is making predictions?
- Can we use tree based models to help us build GLMs?

Relative Importance

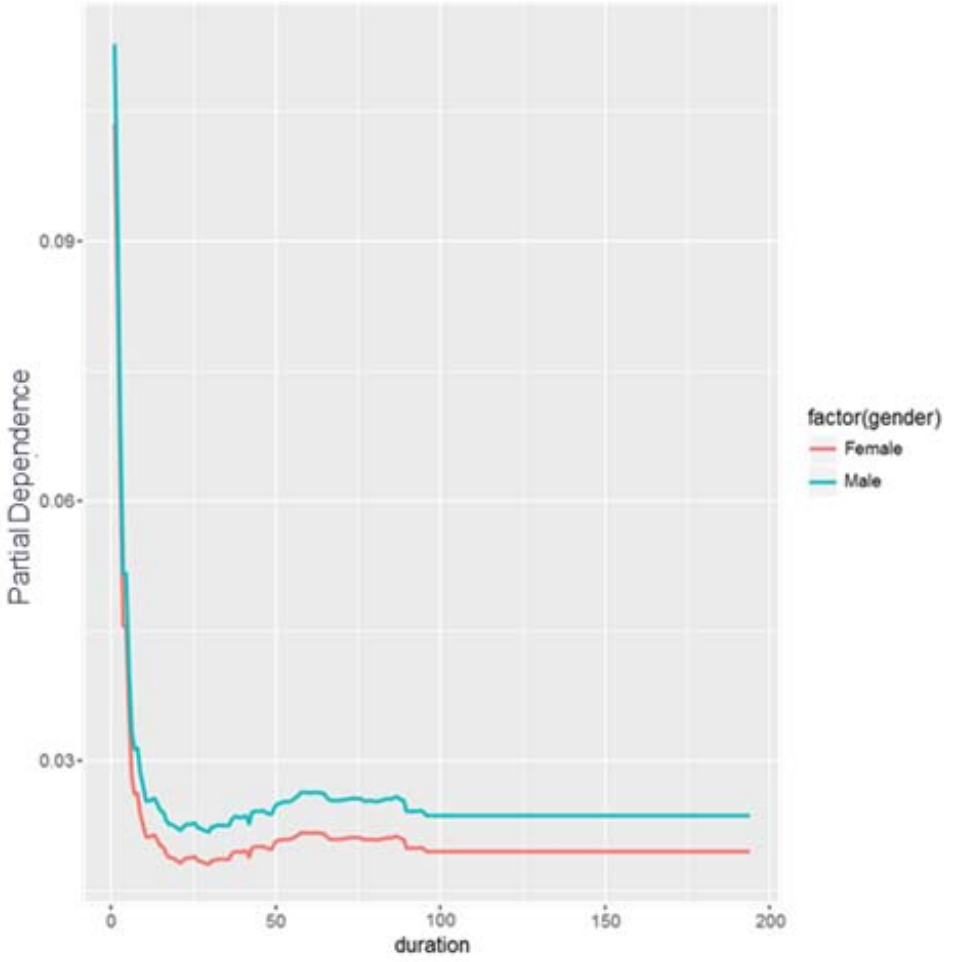
- What variables are most influential in driving the prediction?



Partial Dependence Plots

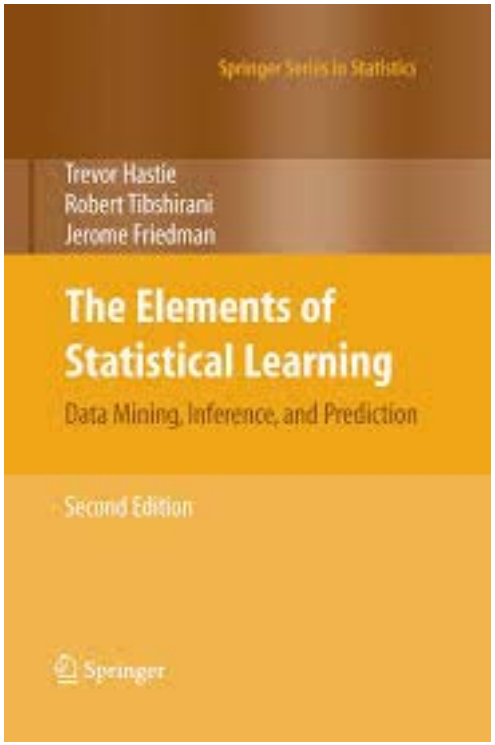


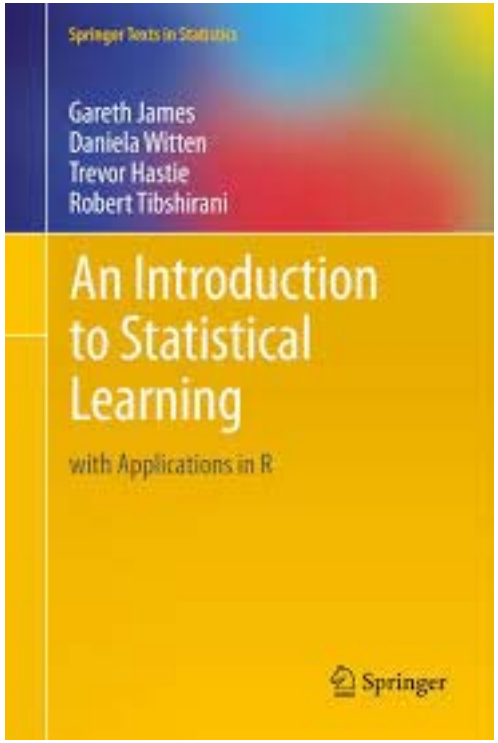
- What is the relationship between the response and independent variables?



How to keep up as stuff evolves?

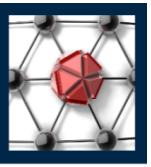
Hit The Books





https://statweb.stanford.edu/~tibs/ElemStatLearn/
http://www-bcf.usc.edu/~gareth/ISL/

Online Education



https://www.datacamp.com/

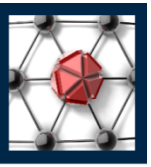
https://www.udacity.com

https://www.coursera.org

http://machinelearningmastery.com

https://www.kaggle.com/

Blogs\Local User Groups



Data Elixir

https://dataelixir.com/

R-bloggers

https://www.r-bloggers.com/

StatsBlogs

http://www.statsblogs.com/

Win-Vector LLC

http://www.win-vector.com/blog/

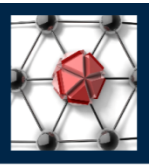
http://blog.revolutionanalytics.com/local-r-groups.html

pymntos
Twin Cities Python User Group

https://wiki.python.org/moin/LocalUserGroups

meetup

https://www.meetup.com/



Questions?