# Intelligible High-Accuracy Models in HealthCare: Predicting Pneumonia Risk and Hospital 30-day Readmission

## Chosen by Dave Kale

Rich Caruana, et al.

Microsoft Research

September 3, 2015

# "Opaque" models are potentially dangerous

**Inspiration:** model risk of death in patients with pneumonia.

- Data set from *hospitalized* patients; no info about treatments.
- Best model: neural net, AUC=0.86 (vs. Log. Reg., AUC=0.77)
- Final model? **logistic regression**!
- Models learned that *asthma lowers risk* (treatment effects!)
- "Wrong" inferences more obvious in, e.g., linear models... but we want more powerful models!

**Solutions?**

- Getting "better" (i.e., clinical trial) data not always an option.
- Removing asthma patients could introduce other biases.
- Removing asthma feature may disperse spurious correlation.
- Changing label for asthma patients confounds treatment, outcome.
- **Better solution:** *transparent* (or *intelligible*) nonlinear models.

# Generalized additive models: balance complexity, intelligibility

- **Intelligible:** e.g., linear model: $y = \beta_0 + \beta_1 x_1 + \cdots + \beta_P x_P$

  $\vdots$

- **Additive:** $y = f_1(x_1) + \cdots + f_P(x_P)$ [HT 1990] [LC 2012]
- **Add. + 2x Interactions:** $y = \sum_i f_i(x_i) + \sum_{i,j} f_{ij}(x_i, x_j)$ [LC 2013]

  $\vdots$

- **Add. + More Interactions:**
  $y = \sum_i f_i(x_i) + \sum_{i,j} f_{ij}(x_i, x_j) + \sum_{i,j,k} f_{ij}(x_i, x_j, x_k) + \ldots$
- **High complexity:** $y = f(x_1, \ldots, x_P)$, e.g., neural net, random forest

$f$'s can be polynomials, splines, etc. [HT 1990]; Gaussian processes; etc.

**Here:** boosted regression trees with greedy selection of interactions [LC 2012] [LC 2013]

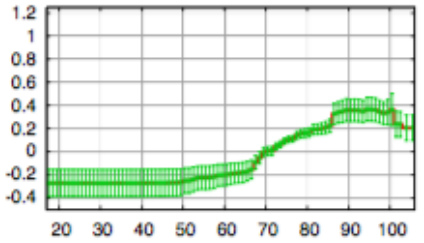
- $f_i$ outputs a 1-D risk curve for $x_i$,
- $f_{ij}$ outputs a 2-D risk "heat map" for $(x_i, x_j)$
- predictive performance comparable to high complexity models

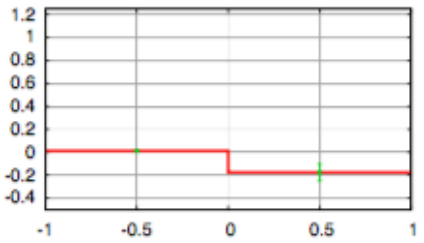
# Predictive Performance

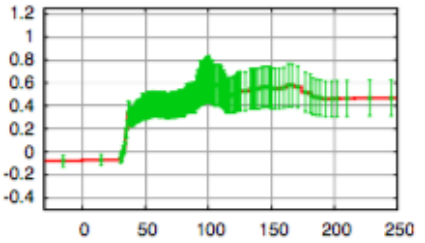
| | Pneumonia | Readmission |
|---|---|---|
| Response | mortality | 30-day readmit |
| train:test | 9847:4352 | 195901:100823 |
| features | 46 (bin., contin.) | 3956 (bin., contin., counts) |
| Logistic Reg.* | 0.8432 | 0.7523 |
| GAM | 0.8542 | 0.7795 |
| $GA^2M$ | 0.8576 | 0.7833 |
| Random Forest | 0.8460 | 0.7671 |

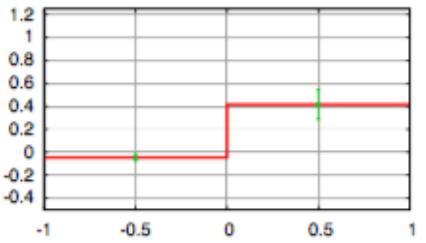
* Used carefully handcrafted features instead of original raw data.

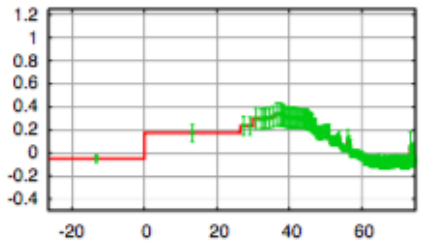
# Intelligibility

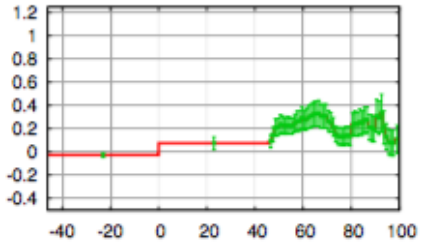


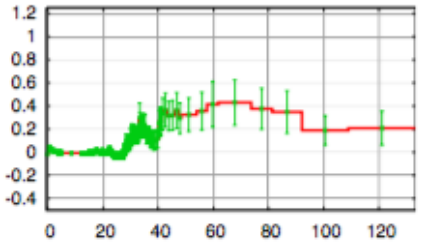



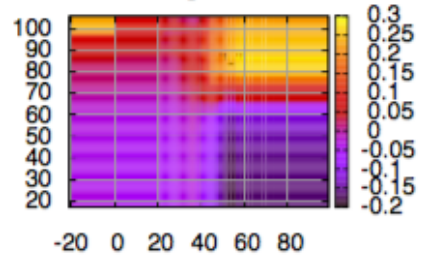



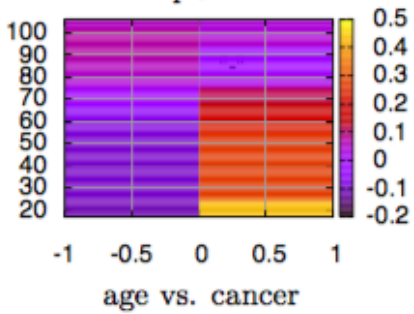



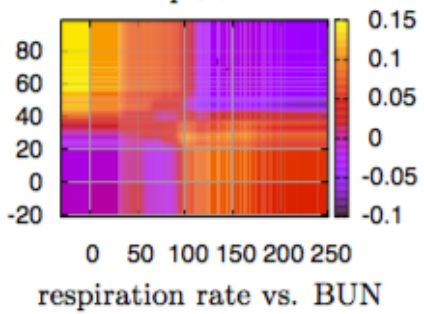



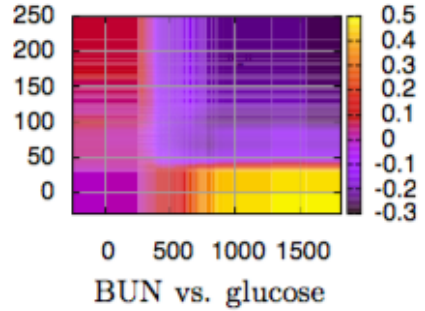



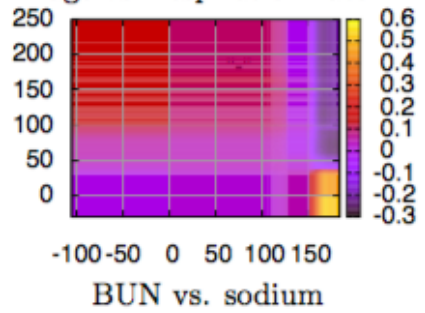



# Find out more:

Live demo: `https://dl.dropboxusercontent.com/u/1497184/medis/iModelTop10.html`
Code: `https://github.com/yinlou/mltk`

References:


- [HT 1990] T. Hastie and R. Tibshirani. *Generalized additive models.* Chapman & Hall/CRC, 1990.
- [LC 2012] Y. Lou, R. Caruana, and J. Gehrke. Intelligible models for classification and regression. KDD 2012.
- [LC 2013] Y. Lou, R. Caruana, J. Gehrke, and G. Hooker. Accurate intelligible models with pairwise interactions. KDD 2013.
- **This:** R. Caruana, Y. Lou, J. Gerke, P. Koch, and M. Sturm. Intelligible High-Accuracy Models in HealthCare: Predicting Pneumonia Risk and Hospital 30-day Readmission. SIGKDD 2015.