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: \( y = \beta_0 + \beta_1 x_1 + \cdots + \beta_P x_P \)
- **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]
- **Add. + More Interactions**:
  \( y = \sum_i f_i(x_i) + \sum_{i,j} f_{ij}(x_i, x_j) + \sum_{i,j,k} f_{ijk}(x_i, x_j, x_k) + \cdots \)
- **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

<table>
<thead>
<tr>
<th></th>
<th>Pneumonia</th>
<th>Readmission</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Response</strong></td>
<td>mortality 9847:4352</td>
<td>30-day readmit 195901:100823</td>
</tr>
<tr>
<td><strong>train:test</strong></td>
<td>46 (bin., contin.)</td>
<td>3956 (bin., contin., counts)</td>
</tr>
<tr>
<td><strong>features</strong></td>
<td>46 (bin., contin.)</td>
<td>3956 (bin., contin., counts)</td>
</tr>
<tr>
<td><strong>Logistic Reg.</strong>*</td>
<td>0.8432</td>
<td>0.7523</td>
</tr>
<tr>
<td><strong>GAM</strong></td>
<td>0.8542</td>
<td>0.7795</td>
</tr>
<tr>
<td><strong>GA(^2)M</strong></td>
<td>0.8576</td>
<td>0.7833</td>
</tr>
<tr>
<td><strong>Random Forest</strong></td>
<td>0.8460</td>
<td>0.7671</td>
</tr>
</tbody>
</table>

* Used carefully handcrafted features instead of original raw data.
Intelligibility

- Intelligible September 3, 2015 5 / 1
Find out more:

Code: https://github.com/yinlou/mltk

References:

- **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.