Interpretable Modeling in Machine Learning

Cynthia Rudin
Associate Professor
Computer Science Department
Electrical and Computer Engineering Department
Duke University
Cynthia’s Principles of Machine Learning

1) The Rashomon Effect: There is no "best" model for a finite dataset.

2) Thou shalt not make up a model using domain expertise alone.

3) People do not like to trust models that they don't understand.

4) Thou shalt not mistake “computationally fast” for “better.”
<table>
<thead>
<tr>
<th>Risk factors</th>
<th>Points</th>
</tr>
</thead>
<tbody>
<tr>
<td>CHF</td>
<td>1</td>
</tr>
<tr>
<td>HTN</td>
<td>1</td>
</tr>
<tr>
<td>Age ≥ 75</td>
<td>1</td>
</tr>
<tr>
<td>DM</td>
<td>1</td>
</tr>
<tr>
<td>Stroke/TIA/embolism</td>
<td>2</td>
</tr>
</tbody>
</table>

Max 6
Data

X: patient histories
Y: whether patient had stroke next year

Machine Learning Algorithm

Model

<table>
<thead>
<tr>
<th>CHADS\textsubscript{2}</th>
<th>Points</th>
</tr>
</thead>
<tbody>
<tr>
<td>Risk factors</td>
<td>1</td>
</tr>
<tr>
<td>CHF</td>
<td>1</td>
</tr>
<tr>
<td>HTN</td>
<td>1</td>
</tr>
<tr>
<td>Age≥75</td>
<td>1</td>
</tr>
<tr>
<td>DM</td>
<td>1</td>
</tr>
<tr>
<td>Stroke/TIA/embolism</td>
<td>2</td>
</tr>
<tr>
<td>Max</td>
<td>6</td>
</tr>
</tbody>
</table>
Data

X: patient histories
Y: whether patient had stroke next year

Model

<table>
<thead>
<tr>
<th>CHADS₂</th>
<th>Points</th>
</tr>
</thead>
<tbody>
<tr>
<td>Risk factors</td>
<td></td>
</tr>
<tr>
<td>CHF</td>
<td>1</td>
</tr>
<tr>
<td>HTN</td>
<td>1</td>
</tr>
<tr>
<td>Age ≥ 75</td>
<td>1</td>
</tr>
<tr>
<td>DM</td>
<td>1</td>
</tr>
<tr>
<td>Stroke/TIA/embolism</td>
<td>2</td>
</tr>
</tbody>
</table>

Max 6
<table>
<thead>
<tr>
<th>Historical</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Age 65-74</td>
<td>2 points</td>
</tr>
<tr>
<td>≥ 75</td>
<td>3 points</td>
</tr>
<tr>
<td>DM/HTN or angina</td>
<td>1 point</td>
</tr>
<tr>
<td>Exam</td>
<td></td>
</tr>
<tr>
<td>SBP &lt; 100</td>
<td>3 points</td>
</tr>
<tr>
<td>HR &gt; 100</td>
<td>2 points</td>
</tr>
<tr>
<td>Killip II-IV</td>
<td>2 points</td>
</tr>
<tr>
<td>Weight &lt; 67 kg</td>
<td>1 point</td>
</tr>
<tr>
<td>Presentation</td>
<td></td>
</tr>
<tr>
<td>Anterior STE or LBBB</td>
<td>1 point</td>
</tr>
<tr>
<td>Time to rx &gt; 4 hrs</td>
<td>1 point</td>
</tr>
</tbody>
</table>

Risk Score = Total (0 - 14)
<table>
<thead>
<tr>
<th>Question Number</th>
<th>Risk Factor</th>
<th>Codes</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Young</td>
<td>Aged 25 or older</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Aged 18 – 24.99</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>Ever Lived With</td>
<td>Ever lived with lover for</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>at least two years?</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Yes</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>No</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Index non-sexual violence - Any Convictions?</td>
<td>No</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Yes</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>Prior non-sexual violence - Any Convictions?</td>
<td>No</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Yes</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>Prior Sex Offences</td>
<td>Charges</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Convictions</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>None</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>3</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>4</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>4+</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>Prior sentencing dates (excluding index)</td>
<td>3 or less</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4 or more</td>
<td>1</td>
</tr>
<tr>
<td>7</td>
<td>Any convictions for non-contact sex offences</td>
<td>No</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Yes</td>
<td>1</td>
</tr>
<tr>
<td>8</td>
<td>Any Unrelated Victims</td>
<td>No</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Yes</td>
<td>1</td>
</tr>
<tr>
<td>9</td>
<td>Any Stranger Victims</td>
<td>No</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Yes</td>
<td>1</td>
</tr>
<tr>
<td>10</td>
<td>Any Male Victims</td>
<td>No</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Yes</td>
<td>1</td>
</tr>
<tr>
<td><strong>Total Score</strong></td>
<td><strong>Add up scores from individual risk factors</strong></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Points</th>
<th>Risk Category</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1</td>
<td>Low</td>
</tr>
<tr>
<td>2.3</td>
<td>Moderate-Low</td>
</tr>
<tr>
<td>4.5</td>
<td>Moderate-High</td>
</tr>
<tr>
<td>6+</td>
<td>High</td>
</tr>
</tbody>
</table>

Suggested Nominal Risk Categories
<table>
<thead>
<tr>
<th>Question Number</th>
<th>Risk Factor</th>
<th>Codes</th>
<th>Score</th>
</tr>
</thead>
</table>
| 1               | Young                            | Aged 25 or older  
Aged 18 – 24.99                                  | 0     |
| 2               | Ever Lived With                  | Ever lived with lover for at least two years?  
Yes  
No                                                | 0     |
| 3               | Index non-sexual violence - Any Convictions?  
No  
Yes                                           | 0     |
| 4               | Prior non-sexual violence - Any Convictions?  
No  
Yes                                           | 0     |
| 5               | Prior Sex Offences                | Charges  
Convictions  
None  
1-2  
3-5  
6 +  
None  
1  
2  
3                                           | 0     |
| 6               | Prior sentencing dates (excluding index)  
3 or less  
4 or more                                      | 0     |
| 7               | Any convictions for non-contact sex offences  
No  
Yes                                           | 0     |
| 8               | Any Unrelated Victims             | No  
Yes                                           | 0     |
| 9               | Any Stranger Victims              | No  
Yes                                           | 0     |
<p>| | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>Prior sentencing dates (excluding index)</td>
<td>1-2</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3-5</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>6+</td>
<td>3</td>
</tr>
<tr>
<td>7</td>
<td>Any convictions for non-contact sex offences</td>
<td>No</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Yes</td>
<td>1</td>
</tr>
<tr>
<td>8</td>
<td>Any Unrelated Victims</td>
<td>No</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Yes</td>
<td>1</td>
</tr>
<tr>
<td>9</td>
<td>Any Stranger Victims</td>
<td>No</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Yes</td>
<td>1</td>
</tr>
<tr>
<td>10</td>
<td>Any Male Victims</td>
<td>No</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Yes</td>
<td>1</td>
</tr>
<tr>
<td>Total Score</td>
<td>Add up scores from individual risk factors</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Suggested Nominal Risk Categories</th>
<th>POINTS</th>
<th>Risk Category</th>
</tr>
</thead>
<tbody>
<tr>
<td>0-1</td>
<td>Low</td>
<td></td>
</tr>
<tr>
<td>2-3</td>
<td>Moderate-Low</td>
<td></td>
</tr>
<tr>
<td>4-5</td>
<td>Moderate-High</td>
<td></td>
</tr>
<tr>
<td>6+</td>
<td>High</td>
<td></td>
</tr>
</tbody>
</table>
Location Model
(Decision Tree Diagram)

Current Cell Tower = 7838
- Time of Day ≤ 8:41 AM
  - In Transit: Bus/Train
- Time of Day > 8:41 AM
  - In Transit: Biking/Walking

Current Cell Tower = 0 or 21999 or 7832 or 21212
- Work/School: My Office/Desk

Current Cell Tower = 34608
- Work/School: My Office/Desk

Current Cell Tower = 7828
- Longitude ≤ 87° 39' 35"
  - Acceleration: Y Axis ≤ 34
    - My Spaces: My Home
  - Acceleration: Y Axis > 34
    - In Transit: Biking/Walking
- Longitude > 87° 39' 35"
  - Latitude ≤ 41° 55' 33"
    - In Transit: Bus/Train
  - Latitude > 41° 55' 33"
    - My Spaces: My Home
Location Model
(Decision Tree Diagram)

Current Cell Tower = 7838

Time of Day <= 8:41 AM
In Transit: Bus/Train

Time of Day > 8:41 AM
In Transit: Biking/Walking

Current Cell Tower = 0 or 21999 or 7832 or 21212
Work/School: My Office/Desk

Current Cell Tower = 34608
Work/School: My Office/Desk

Current Cell Tower = 7828

Longitude <= 87° 39' 35"
Acceleration: Y Axis <= 34
My Spaces: My Home
In Transit: Biking/Walking

Longitude > 87° 39' 35"
Acceleration: Y Axis > 34
In Transit: Bus/Train

Latitude <= 41° 55' 33"
Latitude > 41° 55' 33"
My Spaces: My Home
Decision Trees

• Why trees?

• CART (Breiman 1993) is arguably the most widely used predictive modeling method used in industry currently.
Decision Trees

• Example: Will the customer wait for a table at a restaurant?
  • OthOptions: Other options, True if there are restaurants nearby.
  • Weekend: This is true if it is Friday, Saturday or Sunday.
  • Area: Does it have a bar or other nice waiting area to wait in?
  • Plans: Does the customer have plans just after dinner?
  • Price: This is either $, $$, $$$, or $$$$ 
  • Precip: Is it raining or snowing?
  • Genre: French, Mexican, Thai, or Pizza
  • Wait: Wait time estimate: 0-5 min, 6-15 min, 16-30 min, or 30+
  • Crowded: Whether there are other customers (no, some, or full)

Credit: Adapted from Russell and Norvig
## Decision Trees

### Example: Will the customer wait for a table at a restaurant?

<table>
<thead>
<tr>
<th></th>
<th>OthOptions</th>
<th>Weekend</th>
<th>Area</th>
<th>Plans</th>
<th>Price</th>
<th>Precip</th>
<th>Genre</th>
<th>Wait</th>
<th>Crowded</th>
<th>Stay?</th>
</tr>
</thead>
<tbody>
<tr>
<td>x₁</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>$$$</td>
<td>No</td>
<td>French</td>
<td>0-5</td>
<td>some</td>
<td>Yes</td>
</tr>
<tr>
<td>x₂</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>$</td>
<td>No</td>
<td>Thai</td>
<td>16-30</td>
<td>full</td>
<td>No</td>
</tr>
<tr>
<td>x₃</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>$</td>
<td>No</td>
<td>Pizza</td>
<td>0-5</td>
<td>some</td>
<td>Yes</td>
</tr>
<tr>
<td>x₄</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>$</td>
<td>No</td>
<td>Thai</td>
<td>6-15</td>
<td>full</td>
<td>Yes</td>
</tr>
<tr>
<td>x₅</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>$$$</td>
<td>No</td>
<td>French</td>
<td>30+</td>
<td>full</td>
<td>No</td>
</tr>
<tr>
<td>x₆</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>$</td>
<td>Yes</td>
<td>Mexican</td>
<td>0-5</td>
<td>some</td>
<td>Yes</td>
</tr>
<tr>
<td>x₇</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>$</td>
<td>Yes</td>
<td>Pizza</td>
<td>0-5</td>
<td>none</td>
<td>No</td>
</tr>
<tr>
<td>x₈</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>$</td>
<td>Yes</td>
<td>Thai</td>
<td>0-5</td>
<td>some</td>
<td>Yes</td>
</tr>
<tr>
<td>x₉</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>$</td>
<td>Yes</td>
<td>Pizza</td>
<td>30+</td>
<td>full</td>
<td>No</td>
</tr>
<tr>
<td>x₁₀</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>$$$</td>
<td>No</td>
<td>Mexican</td>
<td>6-15</td>
<td>full</td>
<td>No</td>
</tr>
<tr>
<td>x₁₁</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>$</td>
<td>No</td>
<td>Thai</td>
<td>0-5</td>
<td>none</td>
<td>No</td>
</tr>
<tr>
<td>x₁₂</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>$</td>
<td>No</td>
<td>Pizza</td>
<td>16-30</td>
<td>full</td>
<td>Yes</td>
</tr>
</tbody>
</table>
Decision Trees

• Example: Will the customer wait for a table at a restaurant?

- Crowded?
  - None: No
  - Some: Yes
  - Full:
    - Plans?
      - Yes: No
      - No: Weekend?
        - Yes: No
        - No: Genre?
          - French: Yes
          - Mexican: Yes
          - Pizza: No

- None

- Some

- Full
Standard Way to Build a Decision Tree

• Start at the top of the tree.
• Grow it by “splitting” features one by one. To split, look at how “impure” the node is.
• Assign leaf nodes the majority vote in the leaf.
• Which of these two features should we split on?

Crowded?

None

- \(x_7, x_{11}\)

+ \(+ x_1, x_3, x_6, x_8, x_{12}\)
- \(- x_2, x_5, x_7, x_9, x_{10}, x_{11}\)

Some

+ \(+ x_4, x_{12}\)
- \(- x_2, x_5, x_9, x_{10}\)

Full

+ \(+ x_1\)
- \(- x_5\)

Genre?

French

+ \(+ x_1, x_3, x_4, x_6, x_8, x_{12}\)
- \(- x_2, x_5, x_7, x_9, x_{10}, x_{11}\)

Thai

+ \(+ x_6\)
- \(- x_{10}\)

Mexican

+ \(+ x_4, x_8\)
- \(- x_2, x_{11}\)

Pizza

+ \(+ x_3, x_{12}\)
- \(- x_7, x_9\)
• Which of these two features should we split on?

[.08,.99] is good

[.50,.50] is bad
• Which of these two features should we split on?

[.08,.99] is good
low entropy is good

[.50,.50] is bad
high entropy is bad
• Next we’ll split on Plans

```
Crowded?

None
- x_7, x_{11}

Some
+ x_{4,12}
- x_2, x_5, x_9, x_10

Full
+ x_4, x_{12}
- x_2, x_5, x_9, x_{10}

Plans?

Yes plans
+ x_4, x_{12}
- x_2, x_{10}

No plans
+ x_5, x_9

Yes

No
```


Standard Way to Build a Decision Tree

- Start at the top of the tree.
- Grow it by “splitting” features one by one. To split, look at how “impure” the node is.
- Assign leaf nodes the majority vote in the leaf.
- At the end, go back and prune leaves to reduce overfitting.

Why is this a good way to build a tree?
• Bottom line: Decision trees don’t optimize anything

• No wonder there’s a tradeoff between accuracy and interpretability in predictive modeling…
What we want
Scalable Bayesian Rule Lists

- accurate
- principled
- interpretable
- scalable
Step 1 of Scalable Bayesian Rule Lists: Mine Frequent Patterns

- If age < 50 and aspirin \rightarrow no stroke
- If past stroke and warfarin \rightarrow stroke
- If age < 40 \rightarrow no stroke
- If diabetes and hypertension \rightarrow stroke
Step 2 of Scalable Bayesian Rule Lists: Choose and Assemble Patterns into a Decision List

Example coming...
An Example of a Model for Predicting Customer Churn (IBM Watson Telco Customer Churn Data)

Input: data about each customer, and whether they churned

Output: predictive model on the next slide
An Example of a Model for Predicting Customer Churn (IBM Watson Telco Customer Churn Data)

if Contract= 1 Year & StreamingMovies=Yes )    -> P(churn) = 0.20
else if ( Contract= 1 Year )                                     -> P(churn) = 0.05
else if ( Tenure<1 year & InternetService=FiberOptic )    -> P (churn) = 0.70
else if ( Contract=2 year),                                     ->  P(churn) = 0.03
else if ( InternetService=FiberOptic & OnlineSecurity=No ) -> P(churn) = 0.48
else if ( OnlineBackup=No & TechSupport=No ),   -> P(churn) = 0.41
else                                                    ->  P(churn) = 0.22
Stroke Prediction Model (work in progress)

Input: data about each patient, and whether they had a stroke later
Output: predictive model

IF past history of stroke THEN P(stroke) = 40.5%
ELSE IF patient takes warfarin reliably THEN P(stroke) = 5.6%
ELSE IF age<70 THEN P(stroke) = 7.2%
ELSE IF Blood Pressure>110 THEN P(stroke) = 45%
ELSE IF age<75 THEN P(stroke) = 6.8%
ELSE IF high BMI THEN P(stroke) = 18.4%
ELSE P(stroke) = 7.2%
Step 2 of Scalable Bayesian Rule Lists: Choose and Assemble Patterns into a Decision List

Solves a special optimization problem over decision lists.

\[
\text{maximize}_{\text{models}} \quad \text{Posterior}(\text{model}): \\
\text{pile of rules} \quad \text{size of list}
\]

\[
\text{Posterior}([c_j]_j, m, \{N_{jl}\}_{jl}) = \left( \prod_{j=0}^{m} \frac{\prod_{l=1}^{L} \Gamma(N_{jl} + \alpha_l)}{\Gamma\left(\sum_{l=1}^{L} N_{jl} + \alpha_l\right)} \right) \left( \frac{\lambda_j^m}{m!} \right) \left( \prod_{j=1}^{m} \frac{\sum_{\lambda \lambda' \lambda''} \left( \frac{\lambda_j^m}{m!} \right) \prod_{j=1}^{m} \frac{\left( \frac{\eta_{c_j}^j}{c_j!} \right)}{\sum_{k \in R_{j-1}(c_{j-1})} \left( \frac{\eta_k^j}{k!} \right)} }{\sum_{\lambda \lambda' \lambda''}} \right),
\]
What do you want in a model anyway?

• **Accuracy**
  – Data should look like it could have been generated by the model

• **Gorgeousness**
  – Sparsity, Your own beliefs about the truth
\[ P(\text{data} \mid \text{model}) \]

Likelihood of data to come from model
\[ P(\text{model}) \times P(\text{data} \mid \text{model}) \]

- Prior of model
- Likelihood of data to come from model
\[ P(\text{model} \mid \text{data}) \propto P(\text{model}) \times P(\text{data} \mid \text{model}) \]

- **Posterior of model**
- **Prior of model**
- **Likelihood of data to come from model**
Step 2 of Scalable Bayesian Rule Lists: Choose and Assemble Patterns into a Decision List

Solves a special optimization problem over decision lists.

\[ P(\text{model} \mid \text{data}) \propto P(\text{model}) \times P(\text{data} \mid \text{model}) \]

- Posterior of model
- Prior of model
- Likelihood of data to come from model
Step 2 of Scalable Bayesian Rule Lists: Choose and Assemble Patterns into a Decision List

Solves a special optimization problem over decision lists.

\[
\text{maximize}_{\text{models}} \quad \text{Posterior}(\text{model}): \\
\text{pile of rules} \quad \text{size of list} \\
\]

\[
\text{Posterior}(\{c_j\}_j,m,\{N_{jl}\}_jl) = \left( \prod_{j=0}^{m} \frac{\prod_{l=1}^{L} \Gamma(N_{jl} + \alpha_l)}{\Gamma \left( \sum_{l=1}^{L} N_{jl} + \alpha_l \right)} \right) \left( \frac{\lambda^m / m!}{\sum_{j=0}^{\lambda} \left( \frac{\lambda_j / j!}{\sum_{k \in R_{j-1}(c_{<j})} \eta^k / k!} \right)} \right) \\
\text{counts for label l, rule j} \\
\]

Other SBRL ingredients

• Very fast bit-vector manipulation. Computational reuse. Pre-processing expensive computation.

• Theoretical bounds that are used as optimization cuts.
C4.5 (gray) - models are more accurate but not small

CART (blue) - models are small but not accurate
SBRL
An Example of a Decision List for Predicting Customer Churn (IBM Watson Telco Customer Churn Data)
Model from Fold 1

if Contract= 1 Year & StreamingMovies=Yes )    -> P(churn) = 0.20
else if ( Contract= 1 Year )                                     -> P(churn) = 0.05
else if ( Tenure<1 year & InternetService=FiberOptic )    -> P (churn) = 0.70
else if ( Contract=2 year),                                         ->  P(churn) = 0.03
else if ( InternetService=FiberOptic & OnlineSecurity=No ) -> P(churn) = 0.48
else if ( OnlineBackup=No & TechSupport=No ),    -> P(churn) = 0.41
else                                                    ->  P(churn) = 0.22
Model from Fold 2

\[
\begin{align*}
\text{if} \ ( \text{Contract=One\_year} \ & \text{& StreamingMovies=Yes} \ ) & \rightarrow P(\text{churn}) = 0.20 \\
\text{else if} \ ( \text{Tenure<1 year} \ & \text{& InternetService=Fiber\_optic} \ ) & \rightarrow P(\text{churn}) = 0.70 \\
\text{else if} \ ( \text{Tenure<1 year}&\text{&OnlineBackup=No} \ ) & \rightarrow P(\text{churn}) = 0.44 \\
\text{else if} \ ( \text{InternetService=Fiber\_optic}&\text{&Contract=Month-to-month} \ ) & \rightarrow P(\text{churn}) = 0.43 \\
\text{else if} \ ( \text{Contract=Month-to-month} \ ) & \rightarrow P(\text{churn}) = 0.22 \\
\text{else} & \rightarrow P(\text{churn}) = 0.03
\end{align*}
\]
Model from Fold 3

if ( Contract=One_year & StreamingMovies=Yes ) -> P(churn) = 0.25
else if ( Contract=One Year )               ->  P(churn) = 0.03
else if ( Contract=Two Year )                           ->  P(churn) = 0.05
else if (Tenure<1year & InternetService=Fiber_optic ) -> P(churn) = 0.69
else if (TechSupport=No&OnlineSecurity=No )            ->  P(churn) = 0.45
else                                                ->  P(churn) = 0.25
## Runtime in seconds

<table>
<thead>
<tr>
<th></th>
<th>BRL</th>
<th>LR</th>
<th>SVM</th>
<th>CART</th>
<th>C4.5</th>
<th>RF</th>
<th>ADA</th>
</tr>
</thead>
<tbody>
<tr>
<td>fold 1</td>
<td>0.81</td>
<td>0.30</td>
<td>2.45</td>
<td>0.12</td>
<td>0.42</td>
<td>2.70</td>
<td>6.20</td>
</tr>
<tr>
<td>fold 2</td>
<td>0.85</td>
<td>0.19</td>
<td>2.28</td>
<td>0.12</td>
<td>0.22</td>
<td>2.56</td>
<td>6.07</td>
</tr>
<tr>
<td>fold 3</td>
<td>0.87</td>
<td>0.18</td>
<td>2.34</td>
<td>0.11</td>
<td>0.23</td>
<td>2.60</td>
<td>6.05</td>
</tr>
</tbody>
</table>

Limits: 1 million data points, hundreds of rules: 2700 sec
50K data points, 50K rules: 9000 sec
1) The Rashomon Effect: There is no "best" model for a finite dataset.

2) Thou shalt not make up a model using domain expertise alone.

3) People do not like to trust models that they don't understand.

4) Thou shalt not mistake “computationally fast” for “better.”

5) If you want both accuracy and interpretability….

   optimize for them
Implications in

- Energy (equipment failure prediction)
- Healthcare (scoring systems, diagnose/predict)
- Criminology (who will commit a crime?)
- Marketing (understanding customer preferences)
*Code for SBRL is publicly available on CRAN and on my website. Creative Commons License.
Thanks

*code for SBRL is publicly available on CRAN and on my website*