Ensemble Learning and the Heritage Health Prize

Jonathan Stroud, Igii Enverga, Tiffany Silverstein, Brian Song, and Taylor Rogers

iCAMP 2012
University of California, Irvine
Advisors: Max Welling, Alexander Ihler, Sungjin Ahn, and Qiang Liu

August 14, 2012
The Heritage Health Prize

Goal: Identify patients who will be admitted to a hospital within the next year, using historical claims data.[1]

1,250 teams

Improve Healthcare, Win $3,000,000.

» Goal: Identify patients who will be admitted to a hospital within the next year, using historical claims data.[1]

» 1,250 teams
Purpose

- Reduce cost of unnecessary hospital admissions per year
- Identify at-risk patients earlier
Public Leaderboard  Round 2 Milestone Leaderboard

This leaderboard is calculated on approximately 30% of the test data. The final results will be based on the other 70%, so the final standings may be different.

Reminder: It’s against the rules to make submissions through multiple accounts. Contact us if you notice any ‘sock puppets’.

<table>
<thead>
<tr>
<th>#</th>
<th>Δ1w</th>
<th>Team Name</th>
<th>RMSLE</th>
<th>Entries</th>
<th>Last Submission UTC (Best Submission - Last)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>↑1</td>
<td>Opera Solutions *</td>
<td>0.450799</td>
<td>221</td>
<td>Sun, 12 Aug 2012 23:31:23 (-2.1d)</td>
</tr>
<tr>
<td>2</td>
<td>↓1</td>
<td>EXL Analytics *</td>
<td>0.450816</td>
<td>315</td>
<td>Fri, 10 Aug 2012 19:25:56 (-78.8d)</td>
</tr>
<tr>
<td>3</td>
<td>-</td>
<td>jack3</td>
<td>0.452671</td>
<td>298</td>
<td>Mon, 13 Aug 2012 01:34:27 (-5d)</td>
</tr>
<tr>
<td>4</td>
<td>-</td>
<td>Dolphin</td>
<td>0.453266</td>
<td>324</td>
<td>Mon, 13 Aug 2012 09:18:44 (-3.4d)</td>
</tr>
<tr>
<td>5</td>
<td>-</td>
<td>Market Makers</td>
<td>0.453613</td>
<td>225</td>
<td>Sat, 04 Aug 2012 00:06:19</td>
</tr>
<tr>
<td>6</td>
<td>-</td>
<td>Aroté Associates</td>
<td>0.454935</td>
<td>96</td>
<td>Mon, 13 Aug 2012 15:14:39 (-7.6d)</td>
</tr>
</tbody>
</table>

- Public online competitions
- Gives feedback on prediction models
Data

- Provided through Kaggle
- Three years of patient data
- Two years include days spent in hospital (training set)

<table>
<thead>
<tr>
<th>MemberID</th>
<th>ProviderID</th>
<th>Vendor</th>
<th>PCP</th>
<th>Year</th>
<th>Specialty</th>
<th>PlaceSvc</th>
<th>PayDelay</th>
<th>LengthOfStay</th>
<th>DSFS</th>
<th>PrimaryConditionGroup</th>
<th>CharlsonIndex</th>
</tr>
</thead>
<tbody>
<tr>
<td>42286078</td>
<td>8013252</td>
<td>172153</td>
<td>37976</td>
<td>Y1</td>
<td>Surgery</td>
<td>Office</td>
<td>28</td>
<td>8-9 months</td>
<td>NEUMENT</td>
<td>1-2</td>
<td></td>
</tr>
<tr>
<td>97903248</td>
<td>331666</td>
<td>726296</td>
<td>5300</td>
<td>Y3</td>
<td>Internal</td>
<td>Office</td>
<td>50</td>
<td>7-8 months</td>
<td>NEUMENT</td>
<td>1-2</td>
<td></td>
</tr>
<tr>
<td>2759427</td>
<td>2979752</td>
<td>140433</td>
<td>91972</td>
<td>Y3</td>
<td>Internal</td>
<td>Office</td>
<td>14</td>
<td>0-1 month</td>
<td>METAB3</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>73570559</td>
<td>7053364</td>
<td>240043</td>
<td>70119</td>
<td>Y3</td>
<td>Laboratory</td>
<td>Independent Lab</td>
<td>24</td>
<td>5-6 months</td>
<td>METAB3</td>
<td>1-2</td>
<td></td>
</tr>
<tr>
<td>11837065</td>
<td>7557067</td>
<td>496247</td>
<td>68696</td>
<td>Y2</td>
<td>Surgery</td>
<td>Outpatient Hospital</td>
<td>27</td>
<td>4-5 months</td>
<td>FXDISLC</td>
<td>1-2</td>
<td></td>
</tr>
<tr>
<td>48544661</td>
<td>1963488</td>
<td>4042</td>
<td>68283</td>
<td>Y3</td>
<td>Pediatrics</td>
<td>Office</td>
<td>25</td>
<td>3-4 months</td>
<td>NEUMENT</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>99829076</td>
<td>6721023</td>
<td>26273</td>
<td>91972</td>
<td>Y1</td>
<td>Rehabilitation</td>
<td>Office</td>
<td>162+</td>
<td>0-1 month</td>
<td>TRAUMA</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>54466321</td>
<td>9532074</td>
<td>35565</td>
<td>27294</td>
<td>Y1</td>
<td>Diagnostic Imaging</td>
<td>Office</td>
<td>29</td>
<td>0-1 month</td>
<td>RESPR4</td>
<td>1-2</td>
<td></td>
</tr>
<tr>
<td>60497718</td>
<td>363858</td>
<td>293107</td>
<td>64913</td>
<td>Y2</td>
<td>Rehabilitation</td>
<td>Office</td>
<td>42</td>
<td>1-2 months</td>
<td>INFECC4</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>72200595</td>
<td>6251529</td>
<td>97127</td>
<td>43465</td>
<td>Y3</td>
<td>Internal</td>
<td>Office</td>
<td>56</td>
<td>7-8 months</td>
<td>MISCHRT</td>
<td>1-2</td>
<td></td>
</tr>
<tr>
<td>12333431</td>
<td>606446</td>
<td>120531</td>
<td>20893</td>
<td>Y3</td>
<td>Anesthesiology</td>
<td>Outpatient Hospital</td>
<td>37</td>
<td>0-1 month</td>
<td>ARTHSPIN</td>
<td>1-2</td>
<td></td>
</tr>
<tr>
<td>66951188</td>
<td>469470</td>
<td>173509</td>
<td>75876</td>
<td>Y1</td>
<td>Diagnostic Imaging</td>
<td>Inpatient Hospital</td>
<td>51</td>
<td>0-1 month</td>
<td>ROAM0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>78859064</td>
<td>9121540</td>
<td>523791</td>
<td>21579</td>
<td>Y1</td>
<td>Emergency</td>
<td>Urgent Care</td>
<td>221 day</td>
<td>6-7 months</td>
<td>HEART2</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>35643369</td>
<td>716167</td>
<td>86849</td>
<td>782</td>
<td>Y3</td>
<td>Surgery</td>
<td>Office</td>
<td>21</td>
<td>4-5 months</td>
<td>MISC2a3</td>
<td>1-2</td>
<td></td>
</tr>
<tr>
<td>26477481</td>
<td>836265</td>
<td>68253</td>
<td>20990</td>
<td>Y1</td>
<td>Emergency</td>
<td>Urgent Care</td>
<td>143</td>
<td>0-1 month</td>
<td>ROAM0</td>
<td>1-2</td>
<td></td>
</tr>
<tr>
<td>83695244</td>
<td>524478</td>
<td>791272</td>
<td>322724</td>
<td>Y1</td>
<td>Internal</td>
<td>Office</td>
<td>162+</td>
<td>0-1 month</td>
<td>MISC2a3</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>63053005</td>
<td>2664249</td>
<td>167772</td>
<td>3582</td>
<td>Y1</td>
<td>Diagnostic Imaging</td>
<td>Urgent Care</td>
<td>62</td>
<td>10-11 months</td>
<td>PNEUM0</td>
<td>1-2</td>
<td></td>
</tr>
<tr>
<td>46891766</td>
<td>5094920</td>
<td>659281</td>
<td>71134</td>
<td>Y2</td>
<td>General Practice</td>
<td>Office</td>
<td>21</td>
<td>6-9 months</td>
<td>MISCL5</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>34087064</td>
<td>790212</td>
<td>91972</td>
<td>Y1</td>
<td>General Practice</td>
<td>Office</td>
<td>46</td>
<td>4-5 months</td>
<td>GIOBSEN0</td>
<td>1-2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>31334908</td>
<td>566224</td>
<td>521876</td>
<td>85074</td>
<td>Y2</td>
<td>General Practice</td>
<td>Office</td>
<td>31</td>
<td>0-1 month</td>
<td>RESPR4</td>
<td>1-2</td>
<td></td>
</tr>
<tr>
<td>777964</td>
<td>2540808</td>
<td>62417</td>
<td>76334</td>
<td>Y3</td>
<td>Internal</td>
<td>Office</td>
<td>34</td>
<td>0-1 month</td>
<td>MISC2a5</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>33383188</td>
<td>7053364</td>
<td>240043</td>
<td>68368</td>
<td>Y1</td>
<td>Laboratory</td>
<td>Independent Lab</td>
<td>24</td>
<td>6-7 months</td>
<td>HEART2</td>
<td>1-2</td>
<td></td>
</tr>
<tr>
<td>86882871</td>
<td>741817</td>
<td>943792</td>
<td>20909</td>
<td>Y1</td>
<td>Diagnostic Imaging</td>
<td>Independent Lab</td>
<td>162+</td>
<td>0-1 month</td>
<td>ARTHSPIN</td>
<td>1-2</td>
<td></td>
</tr>
<tr>
<td>78920475</td>
<td>7053364</td>
<td>240043</td>
<td>77779</td>
<td>Y3</td>
<td>Laboratory</td>
<td>Independent Lab</td>
<td>18</td>
<td>0-1 month</td>
<td>METAB3</td>
<td>1-2</td>
<td></td>
</tr>
<tr>
<td>831279</td>
<td>7053364</td>
<td>240043</td>
<td>38672</td>
<td>Y2</td>
<td>Laboratory</td>
<td>Independent Lab</td>
<td>23</td>
<td>7-8 months</td>
<td>RENAL3</td>
<td>1-2</td>
<td></td>
</tr>
<tr>
<td>65398881</td>
<td>5193362</td>
<td>576500</td>
<td>62284</td>
<td>Y2</td>
<td>Emergency</td>
<td>Office</td>
<td>31</td>
<td>3-4 months</td>
<td>TRAUMA</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>41932830</td>
<td>7053364</td>
<td>240043</td>
<td>85661</td>
<td>Y2</td>
<td>Laboratory</td>
<td>Independent Lab</td>
<td>162+</td>
<td>0-1 month</td>
<td>METAB3</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>72140833</td>
<td>661224</td>
<td>391178</td>
<td>72351</td>
<td>Y3</td>
<td>Internal</td>
<td>Office</td>
<td>42</td>
<td>6-7 months</td>
<td>GYNEC10</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>59324278</td>
<td>6303610</td>
<td>154478</td>
<td>44801</td>
<td>Y3</td>
<td>Internal</td>
<td>Urgent Care</td>
<td>481 day</td>
<td>0-1 month</td>
<td>AMI0</td>
<td>1-2</td>
<td></td>
</tr>
<tr>
<td>276553</td>
<td>7053364</td>
<td>240043</td>
<td>58880</td>
<td>Y2</td>
<td>Laboratory</td>
<td>Independent Lab</td>
<td>29</td>
<td>0-1 month</td>
<td>MISC2a3</td>
<td>1-2</td>
<td></td>
</tr>
</tbody>
</table>
Root Mean Squared Logarithmic Error (RMSLE)

\[ \varepsilon = \sqrt{\frac{1}{n} \sum_{i}^{n} [\log(p_i + 1) - \log(a_i + 1)]^2} \]

Threshold: \( \varepsilon \leq .4 \)
The Netflix Prize

- $1 Million prize
- Leading teams combined predictors to pass threshold
Blending

Blend several predictors to create a more accurate predictor
Prediction Models

- Optimized Constant Value
- K-Nearest Neighbors
- Logistic Regression
- Support Vector Regression
- Random Forests
- Gradient Boosting Machines
- Neural Networks
Feature Selection

- Used Market Makers method [2]
- Reduced each patient to vector of 139 features
Optimized Constant Value

- Predicts same number of days for each patient
- Best constant prediction is $p = 0.209179$

RMSLE: 0.486459
(800th place)
K-Nearest Neighbors

- Weighted average of closest neighbors
- Very slow
Eigenvalue Decomposition

Reduces number of features for each patient

\[ X_k = \lambda_k^{-1/2} U_k^T X_c \]
K-Nearest Neighbors Results

Neighbors: $k = 1000$
RMSLE: 0.475197
(600th place)
Logistic Regression

RMSLE: 0.466726
(375th place)
Support Vector Regression

\[ \varepsilon = 0.02 \]

RMSLE: 0.467152
(400th place)
Decision Trees
Random Forests

RMSLE: 0.464918
(315th place)
Gradient Boosting Machines

Trees = 8000
Shrinkage = 0.002
Depth = 7
Minimum Observations = 100
RMSLE: 0.462998
(200th place)
Artificial Neural Networks
Back Propagation in Neural Networking
Neural Networking Results

Number of hidden neurons = 7
Number of cycles = 3000
RMSLE: 0.465705
(340th place)
## Individual Predictors (Summary)

<table>
<thead>
<tr>
<th>Method</th>
<th>Score</th>
<th>Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>Optimized Constant Value</td>
<td>0.486459</td>
<td>800th</td>
</tr>
<tr>
<td>K-Nearest Neighbors</td>
<td>0.475197</td>
<td>600th</td>
</tr>
<tr>
<td>Logistic Regression</td>
<td>0.466726</td>
<td>375th</td>
</tr>
<tr>
<td>Support Vector Regression</td>
<td>0.467152</td>
<td>400th</td>
</tr>
<tr>
<td>Random Forests</td>
<td>0.464918</td>
<td>315th</td>
</tr>
<tr>
<td>Gradient Boosting Machines</td>
<td>0.462998</td>
<td>200th</td>
</tr>
<tr>
<td>Neural Networks</td>
<td>0.465705</td>
<td>340th</td>
</tr>
</tbody>
</table>
Individual Predictors (Summary)
Deriving the Blending Algorithm

Error (RMSE)

\[ \varepsilon = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (X_i - Y_i)^2} \]

\[ n\varepsilon_c^2 = \sum_{i=1}^{n} (X_i - Y_i)^2 \]

\[ n\varepsilon_0^2 = \sum_{i=1}^{n} Y_i^2 \]
Deriving the Blending Algorithm (Continued)

$X$ as a combination of predictors

\[ \tilde{X} = X_w \]

or

\[ \tilde{X}_i = \sum_c w_c X_{ic} \]
Minimizing the cost function

\[ C = \frac{1}{n} \sum_{i=1}^{N} (Y_i - \tilde{X}_i)^2 \]

\[ \frac{\partial C}{\partial w} = \sum_i (Y_i - \sum_c w_c X_{ic})(-X_{ic}) = 0 \]
Minimizing the cost function (continued)

\[
\sum_i Y_i X_{ic} = \sum_i \sum_c w_c X_{ic} X_{ic}
\]

\[
Y^T X = w_c^T X_c^T X
\]
Optimizing predictors’ weights

\[ w_c = (Y^T X)(X^T X)^{-1} \]

\[
\sum_i Y_i X_{ic} = \sum_i X_{ic}^2 + \sum_i Y_{ic}^2 - \sum_i (Y_i - X_{ic})^2
\]

\[
\sum_i Y_i X_{ic} = \sum_i X_{ic}^2 + n\epsilon_0^2 - n\epsilon_c^2
\]
Error (RMSE)

\[ \varepsilon = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (X_i - Y_i)^2} \]

\[ n\varepsilon_c^2 = \sum_{i=1}^{n} (X_i - Y_i)^2 \]

\[ n\varepsilon_0^2 = \sum_{i=1}^{n} Y_i^2 \]
Optimizing predictors’ weights

\[ w_c = (Y^T X)(X^T X)^{-1} \]

\[
\sum_i Y_i X_{ic} = \sum_i X_{ic}^2 + \sum_i Y_{ic}^2 - \sum_i (Y_i - X_{ic})^2
\]

\[
\sum_i Y_i X_{ic} = \sum_i X_{ic}^2 + n\varepsilon_0^2 - n\varepsilon_c^2
\]
X as a combination of predictors

\[ \hat{\mathbf{X}} = \mathbf{X}_w \]

or

\[ \hat{X}_i = \sum_c w_c X_{ic} \]
1. Submit and record all predictions $X$ and errors $\varepsilon$

2. Calculate $M = (X^T X)^{-1}$ and
   
   $$v_c = (X^T Y)_c = \frac{1}{2} \sum_i (X_{iC}^2 + n\varepsilon_0^2 - n\varepsilon_c^2)$$

3. Because $w_c = (Y^T X)(X^T X)^{-1}$, calculate weights $w = Mv$

4. Final blended prediction is $\tilde{X}_i = Xw$
Blending Results

RMSLE: 0.461432
(98th place)
Future Work

- Optimizing Blending Equation with Regularization Constant
  \[ w_c = (Y^T X)(X^T X + \lambda I)^{-1} \]
- Improved feature selection
- More predictors
Questions
Heritage provider network health prize, 2012.  

David Vogel Phil Brierley and Randy Axelrod.  
Market makers - milestone 1 description.  
September 2011.