Session 57PD, Predicting High Claimants

Presenters:
Zoe Gibbs
Brian M. Hartman, ASA
Using Asymmetric Cost Matrices to Optimize Wellness Intervention

ZOE GIBBS
Session 57, Predicting High Claimants
June 26, 2018
Active participation in the Society of Actuaries is an important aspect of membership. While the positive contributions of professional societies and associations are well-recognized and encouraged, association activities are vulnerable to close antitrust scrutiny. By their very nature, associations bring together industry competitors and other market participants.

The United States antitrust laws aim to protect consumers by preserving the free economy and prohibiting anti-competitive business practices; they promote competition. There are both state and federal antitrust laws, although state antitrust laws closely follow federal law. The Sherman Act, is the primary U.S. antitrust law pertaining to association activities. The Sherman Act prohibits every contract, combination or conspiracy that places an unreasonable restraint on trade. There are, however, some activities that are illegal under all circumstances, such as price fixing, market allocation and collusive bidding.

There is no safe harbor under the antitrust law for professional association activities. Therefore, association meeting participants should refrain from discussing any activity that could potentially be construed as having an anti-competitive effect. Discussions relating to product or service pricing, market allocations, membership restrictions, product standardization or other conditions on trade could arguably be perceived as a restraint on trade and may expose the SOA and its members to antitrust enforcement procedures.

While participating in all SOA in person meetings, webinars, teleconferences or side discussions, you should avoid discussing competitively sensitive information with competitors and follow these guidelines:

- **Do not** discuss prices for services or products or anything else that might affect prices
- **Do not** discuss what you or other entities plan to do in a particular geographic or product markets or with particular customers.
- **Do not** speak on behalf of the SOA or any of its committees unless specifically authorized to do so.
- **Do** leave a meeting where any anticompetitive pricing or market allocation discussion occurs.
- **Do** alert SOA staff and/or legal counsel to any concerning discussions
- **Do** consult with legal counsel before raising any matter or making a statement that may involve competitively sensitive information.

Adherence to these guidelines involves not only avoidance of antitrust violations, but avoidance of behavior which might be so construed. These guidelines only provide an overview of prohibited activities. SOA legal counsel reviews meeting agenda and materials as deemed appropriate and any discussion that departs from the formal agenda should be scrutinized carefully. Antitrust compliance is everyone’s responsibility; however, please seek legal counsel if you have any questions or concerns.
Presentation Disclaimer

Presentations are intended for educational purposes only and do not replace independent professional judgment. Statements of fact and opinions expressed are those of the participants individually and, unless expressly stated to the contrary, are not the opinion or position of the Society of Actuaries, its cosponsors or its committees. The Society of Actuaries does not endorse or approve, and assumes no responsibility for, the content, accuracy or completeness of the information presented. Attendees should note that the sessions are audio-recorded and may be published in various media, including print, audio and video formats without further notice.
The Data

967,031 members from a large insurance company

Predictors:
- Funding Arrangement
- Market Segment
- Total Cost 2012
- Diagnosis Count
- Age
- Prescribed Drug Types
- Health Risk Scores
- Gender
- 320 ETGs

Response:
- Binary Classification:
  - Low Claimant: <$100,000
  - High Claimant: >$100,000
2012 vs. 2013 Cost Comparison

<table>
<thead>
<tr>
<th></th>
<th>2012</th>
<th>2013</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean Cost</td>
<td>$4,912</td>
<td>$3,049</td>
</tr>
<tr>
<td>Median Cost</td>
<td>$1,102</td>
<td>$662</td>
</tr>
<tr>
<td>Maximum Cost</td>
<td>$2,631,172</td>
<td>$1,800,314</td>
</tr>
<tr>
<td>Number of High Claimants</td>
<td>4,351</td>
<td>1,783</td>
</tr>
</tbody>
</table>
The Data

• Mean ETG Count: 1.8145
• Mean Diagnosis Count: 4.876
• Mean Age: 34.63
• Classifying all members as low cost results in an accuracy of 0.998
Cost Matrices

\[
\begin{pmatrix}
\text{True Negative} & \text{False Positive} \\
\text{False Negative} & \text{True Positive}
\end{pmatrix}
\]
Cost Matrices

• Negative predictions imply zero interventions and zero additional savings

\[
\begin{pmatrix}
0 & \text{False Positive} \\
0 & \text{True Positive}
\end{pmatrix}
\]
Cost Matrices

• A false positive indicates a loss of the cost of intervention. For simplicity, we assume interventions for false positives yield no savings.

\[
\begin{pmatrix}
0 & -I \\
0 & \text{True Positive}
\end{pmatrix}
\]
Cost Matrices

• A true positive prediction yields total savings less the cost of intervention.

\[
\begin{pmatrix}
0 & -I \\
0 & S - I
\end{pmatrix}
\]
Cost Matrix

• By dividing by $I$ the matrix can be reduced to:

$$
\begin{pmatrix}
0 & -1 \\
0 & \frac{S}{I} - 1
\end{pmatrix}
$$

Where $\frac{S}{I}$ is the savings per dollar of intervention.
Threshold Tuning

• We used XGBoost to predict high claimants.
• Predictive probabilities were calculated using 3-fold cross-validation over the training set.
Threshold Tuning

• The threshold is optimized over the predicted values by maximizing the following equation:

\[
\left( \frac{S}{I} - 1 \right) \times (\# \text{ of True Positives}) - (\# \text{ of True Negatives})
\]
Tuning Threshold

Optimal Threshold by Savings Per Dollar of Intervention

Threshold

Savings Per Dollar of Intervention
Tuning Threshold
# Prediction Probability Values

<table>
<thead>
<tr>
<th>Quantile</th>
<th>Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.0000512</td>
</tr>
<tr>
<td>0.2</td>
<td>0.0000761</td>
</tr>
<tr>
<td>0.4</td>
<td>0.0000845</td>
</tr>
<tr>
<td>0.6</td>
<td>0.0001332</td>
</tr>
<tr>
<td>0.8</td>
<td>0.0004146</td>
</tr>
<tr>
<td>1</td>
<td>0.9597558</td>
</tr>
</tbody>
</table>
Comparing Thresholds

Predicted Probability of Positive

True Value

S/l=2
S/l=10
S/l=100
Extensions

• Expected savings per dollar of intervention can be adjusted according to ETGs
• Cost matrices may be adjusted to include some intervention-induced savings for false positive predictions
SOCIETY OF ACTUARIES
Antitrust Compliance Guidelines

Active participation in the Society of Actuaries is an important aspect of membership. While the positive contributions of professional societies and associations are well-recognized and encouraged, association activities are vulnerable to close antitrust scrutiny. By their very nature, associations bring together industry competitors and other market participants.

The United States antitrust laws aim to protect consumers by preserving the free economy and prohibiting anti-competitive business practices; they promote competition. There are both state and federal antitrust laws, although state antitrust laws closely follow federal law. The Sherman Act, is the primary U.S. antitrust law pertaining to association activities. The Sherman Act prohibits every contract, combination or conspiracy that places an unreasonable restraint on trade. There are, however, some activities that are illegal under all circumstances, such as price fixing, market allocation and collusive bidding.

There is no safe harbor under the antitrust law for professional association activities. Therefore, association meeting participants should refrain from discussing any activity that could potentially be construed as having an anti-competitive effect. Discussions relating to product or service pricing, market allocations, membership restrictions, product standardization or other conditions on trade could arguably be perceived as a restraint on trade and may expose the SOA and its members to antitrust enforcement procedures.

While participating in all SOA in person meetings, webinars, teleconferences or side discussions, you should avoid discussing competitively sensitive information with competitors and follow these guidelines:

- **Do not** discuss prices for services or products or anything else that might affect prices
- **Do not** discuss what you or other entities plan to do in a particular geographic or product markets or with particular customers.
- **Do not** speak on behalf of the SOA or any of its committees unless specifically authorized to do so.
- **Do** leave a meeting where any anticompetitive pricing or market allocation discussion occurs.
- **Do** alert SOA staff and/or legal counsel to any concerning discussions
- **Do** consult with legal counsel before raising any matter or making a statement that may involve competitively sensitive information.

Adherence to these guidelines involves not only avoidance of antitrust violations, but avoidance of behavior which might be so construed. These guidelines only provide an overview of prohibited activities. SOA legal counsel reviews meeting agenda and materials as deemed appropriate and any discussion that departs from the formal agenda should be scrutinized carefully. Antitrust compliance is everyone’s responsibility; however, please seek legal counsel if you have any questions or concerns.
Presentation Disclaimer

Presentations are intended for educational purposes only and do not replace independent professional judgment. Statements of fact and opinions expressed are those of the participants individually and, unless expressly stated to the contrary, are not the opinion or position of the Society of Actuaries, its cosponsors or its committees. The Society of Actuaries does not endorse or approve, and assumes no responsibility for, the content, accuracy or completeness of the information presented. Attendees should note that the sessions are audio-recorded and may be published in various media, including print, audio and video formats without further notice.
Acknowledgments

• Coauthors: Rebecca Owen and Zoe Gibbs
• The Health Care Cost Institute (HCCI) and its data contributors, Aetna, Humana, and UnitedHealthcare, for providing the claims data analyzed in this study.
• The SOA (led by Dale Hall) for funding this work.
• Brad Barney for insightful comments and suggestions
Why is this important?
Importance

• Insurers and policymakers are very interested in predicting which members will be high-cost next year for:
  • Assigning interventions (nurse, etc)
  • High-risk pools
  • General solvency
  • Group rate renewals
Data
## Size of the HCCI Datasets

<table>
<thead>
<tr>
<th>Year</th>
<th>Number of Members</th>
</tr>
</thead>
<tbody>
<tr>
<td>2009</td>
<td>48,511,544</td>
</tr>
<tr>
<td>2010</td>
<td>47,539,751</td>
</tr>
<tr>
<td>2011</td>
<td>46,193,435</td>
</tr>
<tr>
<td>2012</td>
<td>46,544,359</td>
</tr>
<tr>
<td>2013</td>
<td>47,351,996</td>
</tr>
<tr>
<td>2014</td>
<td>48,087,209</td>
</tr>
<tr>
<td>2015</td>
<td>47,782,320</td>
</tr>
</tbody>
</table>
## Explanatory Variables

<table>
<thead>
<tr>
<th>Variable Name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Z_PATID</td>
<td>Member ID number</td>
</tr>
<tr>
<td>RX_CVG_IND</td>
<td>Prescription drug coverage indicator (1 if the member has coverage). If 1, the pharmacy costs for the year are included in the total allowed costs below.</td>
</tr>
<tr>
<td>FEMALE</td>
<td>Gender (0 for male, 1 for female)</td>
</tr>
<tr>
<td>AGE</td>
<td>Age in years</td>
</tr>
<tr>
<td>MKT_SGMNT_CD</td>
<td>Market segment code (I-Individual market, G-Individual group conversion, L-Large, S-Small, O-Other)</td>
</tr>
<tr>
<td>CAT</td>
<td>Total allowed, adjudicated cost for the year, divided into five groups (&lt;100K, 100K-250K, 250K-500K, 500K-1M, &gt;1M)</td>
</tr>
<tr>
<td>CATLESS_1</td>
<td>CAT from one year prior</td>
</tr>
<tr>
<td>CATLESS_2</td>
<td>CAT from two years prior</td>
</tr>
</tbody>
</table>
## Number of High-cost Members

<table>
<thead>
<tr>
<th>Year</th>
<th>100K-250K</th>
<th>250K-500K</th>
<th>500K-1M</th>
<th>&gt;1M</th>
</tr>
</thead>
<tbody>
<tr>
<td>2009</td>
<td>96,554</td>
<td>17,738</td>
<td>4,162</td>
<td>661</td>
</tr>
<tr>
<td>2010</td>
<td>100,812</td>
<td>18,162</td>
<td>4,393</td>
<td>706</td>
</tr>
<tr>
<td>2011</td>
<td>108,965</td>
<td>20,375</td>
<td>4,773</td>
<td>841</td>
</tr>
<tr>
<td>2012</td>
<td>117,325</td>
<td>22,393</td>
<td>5,250</td>
<td>941</td>
</tr>
<tr>
<td>2013</td>
<td>126,099</td>
<td>24,275</td>
<td>5,458</td>
<td>998</td>
</tr>
<tr>
<td>2014</td>
<td>135,050</td>
<td>26,018</td>
<td>5,749</td>
<td>1,030</td>
</tr>
<tr>
<td>2015</td>
<td>147,220</td>
<td>28,425</td>
<td>6,517</td>
<td>1,200</td>
</tr>
</tbody>
</table>
### Prediction Datasets

<table>
<thead>
<tr>
<th>Prediction Year</th>
<th>Sample Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>2011</td>
<td>25,954,734</td>
</tr>
<tr>
<td>2012</td>
<td>26,539,732</td>
</tr>
<tr>
<td>2013</td>
<td>27,061,494</td>
</tr>
<tr>
<td>2014</td>
<td>26,425,810</td>
</tr>
<tr>
<td>2015</td>
<td>25,199,632</td>
</tr>
</tbody>
</table>
Inference
Inference

To help us understand what variables are really driving high-cost members, we fit logistic regressions to each:

- Year
- High cost cutoff (cut)

- We then compared the coefficient estimates (and confidence intervals) to look for trends.
• All effects relatively constant between years
• In the correct order (a priori more likely to be above 100K than above 250K)
RX_CVG_IND

- Positive effect for 100
- Smaller positive effect for 250
- Not much of an effect for 500 or 1000
FEMALE

- Slight negative effect for 100
- Larger negative effect for 250, 500, and 1000
INDV_FLAG

- Slight negative effect 100 or 250
- No significant effect for 500 or 1000
CATLESS1_100

- Large positive effect for all cuts.
- 100, 250, and 500 are in order from smallest to largest effect.
- Much larger uncertainty in 1000.
CATLESS1_250

- Larger effects than CATLESS1_100
- 100, 250, and 500 are in order from smallest to largest effect.
- Much larger uncertainty in 1000, though definitely a larger effect than for 100.
CATLESS1_500

- Continued increased separation.
- Stronger effects across years and cuts.
• Largest separation
• Largest effects
• Increased standard error
CATLESS2_100

- Biggest impact on 100
- Next largest impact on 250
- All impacts significantly smaller than those for CATLESS1_100
CATLESS2_250

- Impact on 100 relatively similar to CATLESS_100.
- Impact on all other cuts larger than CATLESS2_100.
• No significant difference between the various cuts, but all are significantly positive.
CATLESS2_1000

- Similar to CATLESS2_500, no significant difference between the cuts, but all are significantly positive.
AGE (2011)

- All years are very similar in their pattern.
- We have a linear term and several groups
  - 0-2
  - 3-18
  - 19-49
  - 50+
- Each group increases with age, except 1000
• All years are very similar in their pattern.
• We have a linear term and several groups
  • 0-2
  • 3-18
  • 19-49
  • 50+
• Each group increases with age, except 1000
AGE (2013)

- All years are very similar in their pattern.
- We have a linear term and several groups
  - 0-2
  - 3-18
  - 19-49
  - 50+
- Each group increases with age, except 1000
• All years are very similar in their pattern.
• We have a linear term and several groups
  • 0-2
  • 3-18
  • 19-49
  • 50+
• Each group increases with age, except 1000
• All years are very similar in their pattern.
• We have a linear term and several groups
  • 0-2
  • 3-18
  • 19-49
  • 50+
• Each group increases with age, except 1000
Prediction
Prediction

• The other main goal of this work is to explore a few possible models for predicting which members are likely to be high-cost next year.

• In all cases, we fit the models from training data in one year and use it to predict the following year.
Because predicting an extreme minority class can be very difficult, we compare predictive models based on three different training sets.
Methods

To predict which members will be high-cost, we will fit the following models:

• Logistic regression
• Extreme gradient boosted tree (xgboost) using default parameters
• 3 other xgboost models with optimized parameters
Hyperparameters

• Maximum tree depth, range = (3, 10) - maximum number of branch levels in any tree. A higher number here make it more likely that an individual tree is overfit.

• Minimum child weight, (1, 10) - This parameter tells the tree-building process when to stop. If splitting a node would make a child have less weight than this parameter, then the process stops. The larger this value, the simpler the trees will be.
Hyperparameters (continued)

• Subsample, (0.5, 1) - Proportion of the total training set used to build each tree. A smaller value will help to prevent overfitting.

• Column Sample by Tree, (0.5, 1) - Proportion of all the possible covariates used to build each tree.

• Eta, (0,1) - The learning rate. A higher eta will speed up convergence, while a lower eta may make the convergence more precise.
## Learners

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Untrained</th>
<th>Trained1</th>
<th>Trained2</th>
<th>Trained3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Maximum Tree Depth</td>
<td>6</td>
<td>3</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>Minimum Child Weight</td>
<td>1</td>
<td>9.77</td>
<td>2.98</td>
<td>9.26</td>
</tr>
<tr>
<td>Subsample by Tree</td>
<td>1</td>
<td>0.66</td>
<td>0.79</td>
<td>0.97</td>
</tr>
<tr>
<td>Column Sample by Tree</td>
<td>1</td>
<td>0.76</td>
<td>0.6</td>
<td>0.69</td>
</tr>
<tr>
<td>Eta</td>
<td>0.3</td>
<td>0.54</td>
<td>0.52</td>
<td>0.63</td>
</tr>
</tbody>
</table>
Area Under the Curve

- To compare the predictions, we calculate the area under the ROC curve.

At a threshold of 0.6
True Positive Rate = 0.1587  
False Positive Rate = 0.0013
2012

**Cut = 100**

- AUC
- Logistic: Undersampled, Standard, Oversampled
- Untrained, Trained1, Trained2, Trained3

**Cut = 250**

- AUC
- Logistic: Undersampled, Standard, Oversampled
- Untrained, Trained1, Trained2, Trained3

**Cut = 500**

- AUC
- Logistic: Undersampled, Standard, Oversampled
- Untrained, Trained1, Trained2, Trained3

**Cut = 1000**

- AUC
- Logistic: Undersampled, Standard, Oversampled
- Untrained, Trained1, Trained2, Trained3
2013
2014
2015
Prediction Notes

• Sampling method doesn’t matter too much for cuts 100, and 250 (plenty of positive cases).
• For cuts 500 and 1000, oversampling is best.
• Undersampled trained1 does almost as well, but trained2 and trained3 do much worse.
Conclusion

• While good for inference and understanding the drivers of high-cost members, logistic regression is not the best for prediction.

• Oversampling seems to be the best when you have an extreme minority class.

• Draft paper available (hartman@stat.byu.edu)

• This work is threshold-independent, Zoe’s work builds on this to incorporate costs.