Predicting High-cost Health Insurance Members through Boosted Trees and Oversampling: An Application Using the HCCI Database

Brian Hartman\textsuperscript{a,}\textsuperscript{*}, Rebecca Owen\textsuperscript{b}, Zoe Gibbs\textsuperscript{a}

\textsuperscript{a}Department of Statistics, Brigham Young University, Provo, UT, USA
\textsuperscript{b}HealthCare Analytical Solutions, Inc., Bend, OR, USA

Abstract

Using the Health Care Cost Institute data (approximately 47M members over 7 years), we examine how to best predict which members will be high-cost next year. We find that cost history, age, gender, and prescription drug coverage all predict high-costs, with cost history being the most predictive. We also compare the predictive accuracy of logistic regression to extreme gradient boosting and find that the added flexibility of the extreme gradient boosting improved the predictive power. Finally, we show that with our extremely unbalanced classes because high-cost members are so rare, oversampling the minority class provides a better predictive model than undersampling the majority class or using the training data as is.

1. Introduction

A small proportion of members are responsible for a large majority of the total healthcare costs. While most people use very few services, mainly preventive care or minor acute care, and others are regular consumers, but at a moderate cost, nearly 75% of all healthcare expenditures are made by only 17% of users (McWilliams and Schwartz, 2017). This high-cost care can be attributed to four main groups:

1. Completely unexpected - burns or serious car accidents or the transition of a normally mild disease into a crisis due to an unexpected and unavoidable situation, such as the development of encephalitis from a case of West Nile Virus.
2. Lack of due care and caution, as well as some terrible luck, for example septicemia.
3. Expected but not necessarily predictable, like cancer care.
4. Chronic disease that has worsened in severity, so that she is fighting for her life after years of debility.

Due to the contribution to costs of this small segment of the population, there is considerable interest to understand what portion of it can be predicted. The portion that is entirely unfortuitous, and rare, may be estimated by a distribution based on large population studies. It is the portion that could be estimated using a predictive model based on the characteristics of the population that is of great interest because it would allow for some predictions in future costs of a specific population, as well as identifying people for interventions and additional care. While risk adjustment models are good at predicting average costs of care for a category of people, they are still not effective at identifying particular people who may be at risk for very high claim costs in the near future.

This research explores the types of models that will identify individuals who are most likely to exceed a high cost threshold based on a number of characteristics available in the most common information source: administrative claims data. While it is true that the addition of other information from chart review and clinical recommendations are essential for really identifying a person at risk, this data is often difficult to incorporate into the actuarial studies for trend and pricing work. This work seeks to add another tool to the

\textsuperscript{*}Corresponding author

Email addresses: hartman@stat.byu.edu [Brian Hartman], rebecca@hcasolutions.com [Rebecca Owen], zgibbs@gmail.com [Zoe Gibbs]

Preprint May 2, 2019
risk quantification process for members whose costs form a large part of the overall costs of care as well as a significant contributor to the force of trend.

Many authors have examined the issue of high-cost claimants from different directions. A first group explored the common characteristics of high-cost members. Zook and Moore (1980) looked at 2238 patients (of which 13% were high-cost) and found that smoking and drinking were much more prevalent in the high-cost group than the low-cost one. Schroeder et al. (1979) found that very few (17%) of the high-cost claimants suffered from an actual medical catastrophe; most had chronic conditions. Joynt et al. (2013) found that only a small portion of the total spending for high-cost members was due to preventable acute care. Zulman et al. (2015) showed that multimorbidity is common among high-cost members of the U.S. Veterans Affairs Health Care System. They suggest that interventions are needed to help those members better manage multiple conditions. In order to effectively assign these interventions, we need to predict who will likely be high-cost.

Another group looked at which covariates are most likely to predict high-cost members. Garfinkel et al. (1988) used the National Medical Care Utilization and Expenditure Survey to look at predictors of high-cost patients. They found that health status, followed by economic factors best predict high-cost members. Meenan et al. (2003) compared many risk-adjustment models available at the time to determine which are the best at predicting high-cost patients. Fleishman and Cohen (2010) analyzed the Medical Expenditure Panel Survey (MEPS) and compared a risk score (diagnostic cost group) with a count of chronic conditions on their ability to predict which members would be in the highest cost decile the following year. They also checked whether self-rated health status and functional limitations improved predictions. They found that the risk score was the best predictor. After controlling for the risk score, the number of chronic conditions, self-reported health status, and functional limitations were significantly associated with future high-costs.

A final group, whose work most closely aligns with our paper, focus on developing optimal methodology for predicting high-cost members given available covariates. There are two main ways to predict who will be high-cost. Predicting the actual costs for the member will give an entire predictive distribution. Then calculating the probability of any cost, or of exceeding any threshold is trivial. Bayesian hierarchical models (Fellingham et al., 2005), Bayesian nonparametric regression (Fellingham et al., 2015; Hong and Martin, 2017; Richardson and Hartman, 2018), two-part models (Rosenberg and Farrell, 2008; Frees et al., 2011, 2013), and machine learning models (Duncan et al., 2016; Robinson, 2008; Moturu et al., 2007, 2009) can be used to solve this problem. Accurate prediction of probabilities in the (notably heavy) tails can be difficult due to the lack of extreme data and largely dependent upon model assumptions. For those reasons, we focus on predicting the probabilities that members will exceed certain thresholds. While the predictions are not as detailed as those obtained from the total cost models mentioned earlier, they are not as dependent on model assumptions and focus on a simpler question which the sparse extreme data are better able to answer. Exceedance probabilities also naturally answer the question of how likely certain members are to benefit from intervention, both in cost and member outcome.

2. Data

Our data was gathered by the Health Care Cost Institute. It consists of member information from three of the largest health insurers in the United States. When we performed our analysis, they had data for each year 2009-2015. The number of members in each year are listed in Table 1.

The variables we are interested in for our analysis are described in Table 2.

We divide the members into five groups based on their allowed, adjudicated costs for the year (<100K, 100K-250K, 250K-500K, 500K-1M, >1M). The vast majority of the members had less than $100,000 in total claims each year. To understand how rare the group we are exploring is, Table 3 shows the number of members in each high-cost group.

As is readily apparent, there are not many members in the extremely high-cost group. This is another reason to focus on the probability of exceeding a certain threshold, rather than attempt to estimate a predictive distribution of costs for each member. We are essentially taking the role of an intervention manager and trying to find those members which are most likely to be high-cost. The proportion of all members in each of the high-cost groups have also increased every year.
<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>

Table 1: Number of members in each dataset

<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>GDR</td>
<td>Gender (1 for male, 2 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) For inference, we focus only on the individual market (INDV_FLAG), but in prediction we use all segments.</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>CATLESS1</td>
<td>Total allowed, adjudicated cost for the member one year ago, divided into five groups (&lt;100K, 100K-250K, 250K-500K, 500K-1M, &gt;1M)</td>
</tr>
<tr>
<td>CATLESS2</td>
<td>Total allowed, adjudicated cost for the member two years ago, divided into five groups (&lt;100K, 100K-250K, 250K-500K, 500K-1M, &gt;1M)</td>
</tr>
</tbody>
</table>

Table 2: Variable names and descriptions

When predicting whether a member will be high-cost in a certain year, we only use data available at that time (similar to how the analysis will be done in practice). We will use data from the previous two years to predict if the member will be high-cost in the following year. For example, to predict whether the member will be high-cost in 2012, we will use data from 2010 and 2011. Because we have data from 2009-2015, we predict each member in 2011-2015. For each prediction year, we only use those members for which we have data for the year in question and the previous two. That reduces our sample sizes to those shown in Table 4.

Reducing our dataset to only those who were members for at least part of each of three sequential years impacts our data (and therefore our inference) in several ways. First, there are no members under the age of 2 in our prediction datasets. Partially because of those missing infants, the median age of the members in our dataset is about nine years older than that of those not in our set (39 vs. 30). Further, much of the lifetime medical spending occurs in the final year of life, so those who were expensive and then passed away in previous years will not be included in our dataset. About 25% of the high cost members (>100K) from any year are not in the dataset in the following year. Of those in our dataset, around 50% have prescription drug coverage, whereas of those not in our dataset about 60% have coverage. Additionally, there are about twice as many members in the individual market among those not in our dataset (about 8% to about 4%), potentially due to people moving to the Affordable Care Act’s exchanges and our data being unable to connect that person in the two different providers. Most importantly, the proportion of high-cost members is about the same between the two groups, except for a few more people (one or two per 100K members) above 1M in the group not in our dataset.
### Table 3: Number of members in each high-cost group

<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>96,354</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>135,099</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>

### Table 4: Sample size of each three-year dataset

<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>

### 3. Methods

For inference, we fit separate logistic regression models to four different measures of high cost, greater than 100K, 250K, 500K, and 1M in claims for each of the years. We are interested in the parameter estimates and how they changed over time. Looking at the parameter estimates over time will allow us to see both how consistent the estimates are and to notice any temporal changes or patterns (possibly due to the Affordable Care Act).

For prediction, we will compare two separate models, logistic regression and extreme gradient-boosted classification trees (Chen and Guestrin, 2016), also known as xgboost. Classification trees attempt to model which observations are likely to have a response variable of 1 by repeatedly splitting the dataset on different explanatory variables, trying to make the resulting subsets of observations as similar as possible (containing mainly positive or negative cases) while preventing overfitting. Xgboost refines standard classification trees by fitting new trees to the residuals resulting from earlier trees. The effectiveness of an xgboost model largely depends on the hyperparameter settings which we discuss and optimize later in this section.

When we compare the predictive accuracy of the two models, we fit the model to one year (say 2012), and use that model to fit the following year (2013). This will show which model is superior in a realistic situation. This is better than dividing each year into a training and test set and comparing model accuracy that way.

Classification can be difficult when the positive class is extremely rare, as it is in our case. To help mitigate that issue, we will train the 2012 models on three different datasets.

- Standard: The original 2011 data (say 1,000 high-cost members and 1,000,000 low-cost members for illustration)

- Under: A dataset with the 2011 low-cost members undersampled, making an equal number of high- and low-cost members. We randomly select (without replacement) 1,000 of the 1,000,000 low-cost members to be in the training set. This means that we have a training sample of 2,000 members.

- Over: A dataset with the high-cost members oversampled, again making an equal number of high- and low-cost members. We randomly select (with replacement) a sample of 1,000,000 from the 1,000 high-cost members. In this case, our training sample will include 2,000,000 members.
To tune the xgboost models, we adjust the following five hyperparameters:

- Maximum tree depth, ranging between (3, 10) - maximum number of branch levels in any tree. A higher number here makes 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.
- 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.

We created four different xgboost models. The first (untrained) uses commonly-held default values for each of the above hyperparameters. The next model (trained1) starts with the hyperparameters from the untrained model and then compares it to ten different possible settings, randomly drawn from the set of possible hyperparameters (in parentheses in the list above). The settings are compared through 3-fold cross-validation and by choosing the set of hyperparameters which minimizes the AUC in the cross-validation. This is done only with the data available through 2011, making sure that the optimization does not include any of the data we are trying to predict. The following model (trained2) starts with the chosen hyperparameters in trained1, and then compares that to ten additional randomly drawn possible sets. The final model (trained3), follows the same pattern. The chosen hyperparameters are in table 5.

<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</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>0.60</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>

Table 5: Hyperparameter settings for the four models

Trained1 and Trained3 have hyperparameter settings which are rather similar, though the subsample rates are much higher for trained3. That could make overfitting more likely in trained3 than in trained1. Trained2 has a much smaller minimum child weight, which can also lead to more complicated trees and potential overfitting.

4. Results

We divided our results section into two parts. First, we examine the inference results gathered from the logistic regressions on each year 2011-2015, using the two previous years to help predict current year cost. Afterward, we will discuss the prediction results of the various models.

4.1. Inference

To help visualize the impact of the various covariates on the high-cost predictions, we plot each of the coefficients across all the years and for each cut defining high-cost. We plot the regression coefficients across years to give the reader an understanding of the consistency (or lack thereof) of the estimates and also to potentially see an impact of the Affordable Care Act. The plots are available in Figures 1-4. The solid lines are the fitted regression coefficients, while the dashed lines are the 95% confidence interval bounds.
The intercepts seem relatively constant over time. They are also in the expected order with the 1M cutoff lowest and the 100K cutoff highest. Holding all other variables constant, the probability of exceeding 1M must be no greater than the probability of exceeding 100K.

Prescription coverage also has a relatively constant effect across years. The effect is largest on the 100K cutoff, and is relatively small for all the other cutoffs. There are few prescriptions which will push a member over the 250K, 500K, or 1M thresholds, but more that will push them over the 100K threshold.

The effect of gender is relatively small at the 100K cutoff with female members being slightly less likely to exceed the cutoff. For the other three cutoffs, the effect is significantly larger. In all cases, female members are less likely to be high-cost. This effect is also rather consistent over time.

Whether the member is in the individual market does not have a significant impact at the 500K or 1M cutoffs, but makes the member a little less likely to exceed the 100K or 250K cutoffs. It is also a little surprising that there is no noticeable change in this plot when the ACA began. This is potentially due to the fact that we are only using members with data in 3 consecutive years. Those which moved to the market from employer health insurance may not have those sets of records connected in our data. Similarly, those who were uninsured before the ACA have no data in our set before they purchased a policy.

Cost history has the biggest impact on the probability of being high-cost. When looking at the cost history from the previous year, if the member was high-cost last year, he is more likely to be high-cost this year. The impact increases as the costs last year increase. Additionally, the impact increases as the cutoff increases. In all cases, the uncertainty around the 1M cutoff estimates are the greatest. The cost history from two years ago had a similar, though slightly muted, impact.

The age pattern is relatively consistent across years. For the 100K, 250K, and 500K cutoffs, members are more likely to be high-cost as they age. While the linear term for each additional year is relatively small, there are significant jumps between the categorical groups (at ages 3, 18, and 50). The 1M cutoff has no significant age effect, due to the large amount of uncertainty in these estimates.

### 4.2. Prediction

We compare the predictions from a logistic regression to those from gradient boosted trees, both with default hyperparameter settings and settings optimized through cross-validation. We compare the accuracy using the area under the ROC curve (AUC). The ROC curve plots the true positive rate against the false positive rate as the threshold changes. Therefore, the larger the AUC, the better the model discerns between high- and low-cost members. Plots of the AUC values are available in Figures 5-8.

The resulting patterns are rather consistent across years. For cuts 100K and 250K, all of the xgboost models significantly outperformed logistic regression. Also, the sampling method doesn’t seem to have much of an impact. As the number of positive cases decreases (for cutoffs 500K and 1M), oversampling outperforms the other two sampling methods for the xgboost models. This is less true for trained1, where undersampling works essentially as well as oversampling. The hyperparameter settings for trained1 constrained the model to be relatively simple, possibly muting the benefit of the oversampling. For the other three xgboost models, undersampling is by far the worst method. Logistic regression performs equally well, regardless of sampling methodology.

### 5. Conclusion

In this paper, we use the Health Care Cost Institute data (approximately 47M members over 7 years) to examine how to best predict and describe high-cost members. Using a logistic regression, we find that cost history, age, gender, and prescription drug coverage all predict high-costs, with cost history being the most predictive. In addition to the logistic regression model, we compare the predictive accuracy of extreme gradient boosting and find that the added flexibility of the extreme gradient boosting greatly improves the predictive power. Finally, we show that with our extremely unbalanced classes, oversampling the minority class provides better predictions than undersampling the majority class or using the training data as is.
There are many potential avenues of future work to explore. With the HCCI data, it would be very interesting to further explore the many relationships among the members (spatially, temporally, and hierarchically). It would also be interesting to try and quantify the impact of wellness programs, or attempt to further explore the impact of the Affordable Care Act.

6. Acknowledgments

The authors acknowledge the assistance of the Health Care Cost Institute (HCCI) and its data contributors, Aetna, Humana, and UnitedHealthcare, in providing the claims data analyzed in this study. They are also grateful for the support of the Society of Actuaries which funded this work. Finally, they are grateful to Brad Barney and the anonymous referees for their suggestions and advice.

References


Figure 1: Coefficient Plots part 1
Figure 2: Coefficient Plots part 2
Figure 3: Coefficient Plots part 3
Figure 4: Coefficient Plots part 4
Area under the ROC Curve for Prediction Year 2012

Figure 5: AUC Plots 2012
Area under the ROC Curve for Prediction Year 2013

Figure 6: AUC Plots 2013
Area under the ROC Curve for Prediction Year 2014

Figure 7: AUC Plots 2014
Figure 8: Area under the ROC Curve for Prediction Year 2015