Predicting future healthcare expenses with machine learning

A new approach to population management

Neil Schneider, FSA, MAAA
Arthur L. Wilmes, FSA, MAAA

Milliman

INTRODUCTION

The payment landscape of healthcare is changing from fee-for-service to fee-for-performance. These payment arrangements will shift some or all of the risk to the healthcare provider. As such, providers will need to turn to either existing or new products to manage a patient population and the best process for prioritizing patients within those populations for care management.

Most analytic products in the market focus on ordering patients for care management by highest cost or highest level of morbidity: The sicker they are, the greater the need to manage them. Such approaches may ignore some critical business questions:

- Which patients represent the "greatest risk" under a risk-sharing arrangement?
- Which risks can be mitigated through care management?
- Can advanced metrics be developed to measure care management performance?
- How can the care management process be optimized?

These questions will require analytic approaches more advanced than prioritizing patients with the highest average expense.

MACHINE LEARNING VERSUS TRADITIONAL STATISTICS

With new contracting models, healthcare providers will want to analyze each member in the attributed population to assess what, if any, opportunities for care management may exist. How will this analysis be done? There are a number of analysis techniques in the market that rely upon traditional statistical methods. Traditional statistical methods place importance on data interpretability when defining relationships in the data. These methods result in statistical models where model fit is assessed through hypothesis testing. The result is that traditional analytic tools and regression techniques focus on the average outcome and do not necessarily provide insight into risk that is specific to an individual member. In other words, precision at the level of the individual member is compromised for the sake of improving precision for the average member.

In contrast, solutions based upon machine learning techniques (MLTs) focus on developing models that not only describe the data well but also perform well when making individual outcome predictions. Machine learning algorithms may also be better suited than traditional statistics for healthcare expense predictions. This is due to the number of potential dimensions (variables) involved in data available for analysis.

AVERAGE MEMBER VERSUS INDIVIDUAL OUTLIER

Conventional prospective risk adjusters predict the relative future cost of an average member with given medical attributes. These scores are designed to provide insight into a population's health and expenses but may perform poorly on predicting an individual's expense, which is due to some of the inherent weaknesses of traditional statistics. Such methods also provide no information related to the variability of potential healthcare expense outcomes around the average score or insight into where the best opportunities for improvement in prior performance may reside.

Milliman's approach to predictive analytics focuses on the concept of potentially avoidable healthcare expenses and patients having prospective characteristics with the largest risk for potentially avoidable healthcare expenses.[1] MLTs can be constructed to predict the distribution of an individual's total and potentially avoidable healthcare expenses.[2] Prospective models for the distribution provide insight into a provider's risk per member by not only providing information about the expected healthcare expenses but also attaching values to expenses under adverse outcome scenarios. Understanding an individual's distribution of both total healthcare expenses and potentially avoidable healthcare expenses allow decisions to be made based on the possibility of reducing expenses through changes in the current ambulatory care management. This focus on individual member risk, rather than on overall average population outcomes, gives the healthcare provider valuable information in selecting who and what to manage—in other words, answers to the business questions posed above.

1 Maslow, Katie & Ouslander, MD, Joseph G. (February 2012). Measurement of Potentially Preventable Hospitalizations: Prepared for the Long-Term Quality Alliance, pg. 1.
2 Stranges, E. & Stocks, C. (November 2010). Potentially Preventable Hospitalizations for Acute and Chronic Conditions, 2008. HCUP Statistical Brief #99. Agency for Healthcare Research and Quality.

BOOSTING THE INDIVIDUAL SIGNAL - DROWN OUT THE NOISE

Milliman has developed proprietary methods for predicting healthcare expense outcomes utilizing MLT. Variations of Friedman's gradient boosting machine were developed for these predictions.[3] A gradient boosting machine is a framework for creating boosted decision trees. Each decision tree is a simple, interpretable model. A series of true/false decision points lead to predictions for each terminal leaf in the tree.

A single decision tree is not competitive with other models on prediction accuracy; decision trees generally need to be blended to improve the level of accuracy. Boosting is a method of sequentially building small decision trees to slowly build a more robust and accurate prediction. Figure 1 illustrates how predictions are grown from decision trees.

While boosting can often result in dramatic improvements in accuracy, it is at the loss of interpretability and an increased risk of over-fitting.[4] A version of the random subspace method was custom-written for these analytics to reduce the models' over-fitting tendencies and improve computational efficiency.[5]

Random subspacing forces the model to learn from a broader amount of the information, creating trees with more statistical independence. This results in more robust predictions.

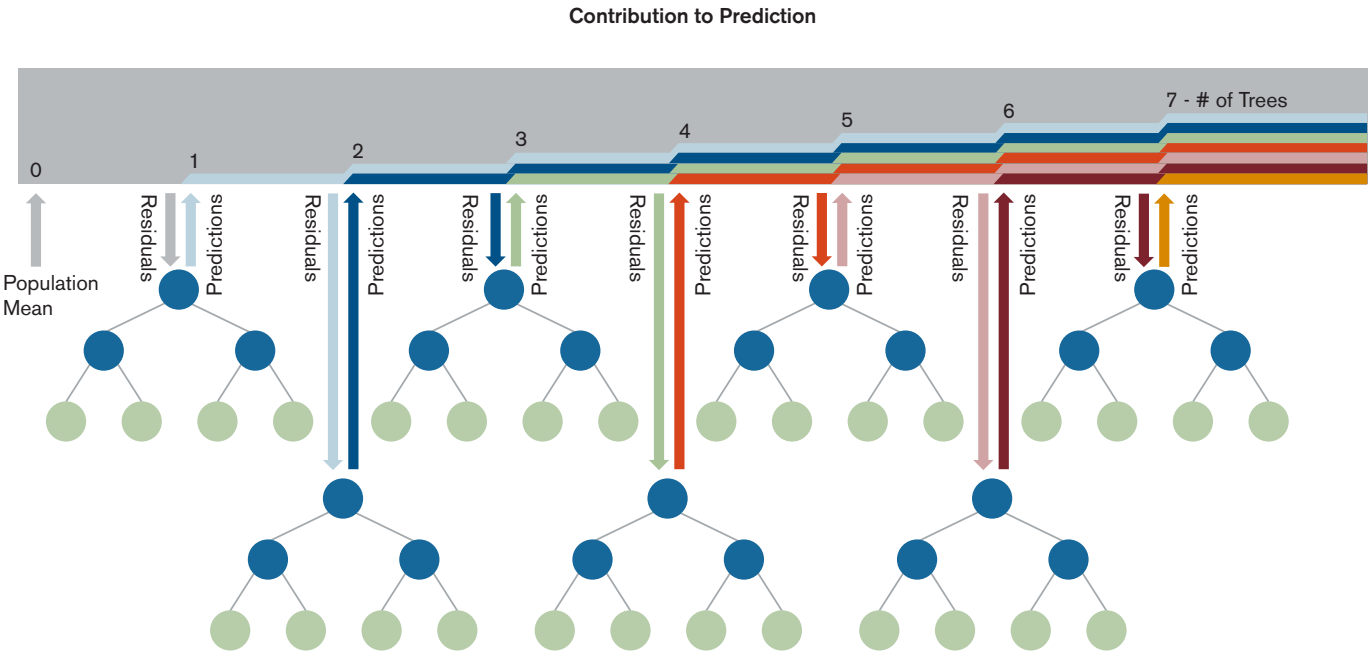
In contrast to other prediction algorithms and methods, gradient boosted decision trees:

- Are more robust against the influence of outlier data
- Reduce the influence of weak or highly collinear variables
- Handle missing values without complex imputations or deletions
- Are designed to model nonlinear relationships and interactions
- Employ multiple regularization techniques to reduce over-fitting

The gradient boosting framework allows for different types of predictions depending on the choice of loss functions to optimize (e.g., probabilities, quantiles, or averages). The appropriate loss functions are chosen to make predictions for the following outcomes:

- *Probability* of an inpatient visit
- *Probability* of an emergency room (ER) visit
- *Potential outlier size* of total healthcare expenses
- *Potential outlier size* of potentially avoidable healthcare expenses

FIGURE 1: BOOSTED DECISION TREES



3 Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics* 29(5):1189-1232.
4 James, G., et al. An Introduction to Statistical Learning: With Applications in R. Springer Texts in Statistics 103, p. 303 DOI 10.1007/978-1-4614-7138-7.
5 Ho, Tin (1998). The Random Subspace Method for Constructing Decision Forests. IEEE Transactions on Pattern Analysis and Machine Intelligence 20 (8): 832–844. doi: 10.1109/34.709601.

The ability of boosted decision trees to produce quality predictions relies on selecting an optimal number of trees. Optimizing the number of trees is done by *k*-fold cross-validation (*k*-fold CV), which involves randomly dividing the set of observations into *k* groups, or *folds*, of approximately equal size. The first fold is treated as a validation set, and a model is built on the remaining $k - 1$ folds. The appropriate error metric is then optimized on the validation set. This procedure is repeated *k* times; each time, a different group of observations is treated as a validation set. This process results in *k* estimates of the test error. The *k*-fold CV error is computed by averaging the testing error from each fold.[6] The optimal number of trees is then selected based on the lowest *k*-fold CV error. To improve consistency of the results, the entire *k*-fold cross-validation process is repeated multiple times and the final predictions are averaged. While it is computationally expensive to run all the repeated *k*-folds individually, they are independent models that can be calculated simultaneously with a large computer cluster.

GARBAGE IN, GARBAGE OUT

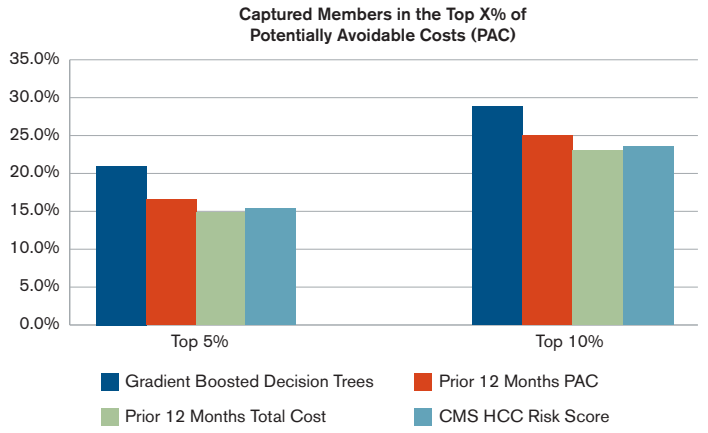
It is important to understand how to best "work your data." The "garbage in, garbage out" adage in this case refers to the need for quality variables to make quality predictions. For example, claim line detail or very specific clinical data is often too granular to make a good model. However, deriving variables from the data will condense the information into useful groups. Summarizing an individual's utilization and costs by month or quarter generally produces better predictive results. With regard to administrative claims and clinical data, the data is generally broken down and summarized into the following categories:

- Demographic information about an individual includes:
  - Gender
  - Age
  - Eligibility status
- Medical conditions: Centers for Medicare and Medicaid Services (CMS) Hierarchical Condition Categories (HCC)
- Risk scores: Varies by line of business
- Historical cost and utilization by:
  - Potentially avoidable services
  - In-network versus out-of-network services
  - Type of service (e.g., inpatient, ER, skilled nursing facility)
  - Date of service
- Electronic medical record data, for example:
  - Vitals signs (e.g., height, weight, and blood pressure)
  - Lab results
  - Smoking status

PERFORMANCE

Predictions of the potentially avoidable healthcare expenses provide healthcare entities with valuable information when selecting individuals for ambulatory management. Historical data is used to make predictions for a known time period. In the Milliman models, we generally predict in six-month prospective increments to isolate on near-term opportunities for additional management. The members within the prediction period are ranked based on their actual potentially avoidable healthcare expenses and then compared to the predicted values. The predictions from the gradient boosting machine are compared with rankings produced from the CMS-HCC community risk scores (for Medicare beneficiaries, for example) and rankings based on the prior 12 months of total and potentially avoidable costs. Figure 2 illustrates the ranking accuracy with the percentage of members in the actual top 5% or 10% who were also identified in the top 5% or 10% of the prediction algorithm.

FIGURE 2: RANKING ACCURACY



CONCLUSION

The advantages of gradient boosted decision trees over traditional predictive modeling techniques help produce more accurate projections of future healthcare expenses. The accuracy can be seen in the improved ranking of members with high potentially avoidable costs. This allows the care coordination teams to strategically target members to manage, saving time and resources–and, hopefully, reaching members before they incur potentially avoidable costs. In aggregate, the algorithm's predictions could ultimately form a basis for benchmarking care management outcomes.

Neil Schneider, FSA, MAAA, is an associate actuary with the Indianapolis office of Milliman. Contact him at neil.schneider@milliman.com.

Arthur L. Wilmes, FSA, MAAA, is a principal and consultant with the Indianapolis office of Milliman. Contact him at art.wilmes@milliman.com.

6 James, G., ibid., p. 181.