Machine learning in healthcare

'Artificial intelligence is the new electricity. Just as 100 years ago electricity transformed industry after industry, AI will now do the same.' (Andrew Ng, Stanford AI professor)

Judith Houtepen, MBA, AAG
Sjors Altemühl, MSc, AAG

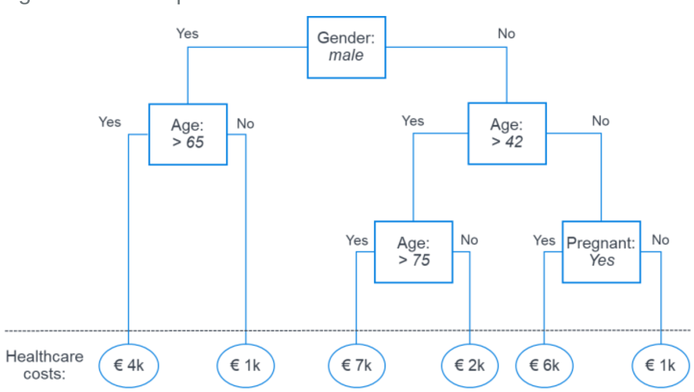
Artificial intelligence is a broad discipline. The subarea of **machine learning** is, however, closely related to techniques that actuaries have already been using for decades. Actuaries are thus in an excellent position to add machine learning to their existing skill set.

Just like traditional statistical models, machine learning algorithms make a prediction of the value of an uncertain variable (e.g., healthcare costs) on the basis of a series of characteristics. In formula form, this is $\boldsymbol{Y} = \boldsymbol{f}(\boldsymbol{X_i})$.

Most traditional techniques assume fixed **linear relationships** for which coefficients are estimated. This is also the case for the health insurance risk equalization model used in the Netherlands, where health insurers receive compensation from the government for insuring someone, based on an individual's risk profile. The exact amount is determined as the sum of the standard amounts for various health-related characteristics. In formula form, this is $\boldsymbol{Y} = \sum \boldsymbol{\beta_i X_i}$.

Due to their rigid structure, traditional techniques focus on 'average' relationships. Subtle relationships or cross-effects (e.g., differences in expenses between children and adults for specific diseases) are often ignored. 'Tree-based' machine learning algorithms are quite accessible to actuaries. These algorithms use **decision trees** to arrive at a prediction. Decision trees are nonlinear by design and capable of picking up on cross-effects. Furthermore, in contrast to traditional techniques, these algorithms determine relevant cross-effects in an automated way.

Figure 1: Example of a Basic Decision Tree



In order to also capture subtle relationships, machine learning repeats the estimation procedure a large number of times (typically 100 to 1,000 times). With **random forest**, this is done by creating new versions of the original data set using bootstrapping. Each estimation procedure can result in a different decision tree with different predictions. The final prediction value is the average of the predictions of all the individual decision trees.

With machine learning, the role of the **actuary** becomes more crucial than ever, as machine learning is not a replacement of traditional statistical or actuarial techniques. If the sole aim is to achieve the highest predictive power on an individual level, machine learning is often superior. But with machine learning, it can be difficult to trace back why a certain prediction has been made. Therefore, when explainability is important, it can be more suitable to use traditional techniques. Also, when the goal is only to achieve good predictive power on an aggregate level, the added value of machine learning can be limited. Actuaries are responsible for making correct modeling choices using their deep subject matter expertise.

One way machine learning can add value in the existing Dutch healthcare landscape is in analyses to **subpopulations** (e.g., regions or diseases). The Dutch risk equalization model estimates a standard amount for each relevant characteristic, but these standard amounts only represent average relationships valid at the population level. When these amounts are used to predict healthcare costs of specific subpopulations, there can be significant differences between the predicted and actual healthcare costs. The risk equalization system is also not designed for this purpose, but aims to compensate insurers on a total level for expected differences in healthcare costs as a result of the composition of their insurance populations. Also, the risk equalization system deliberately leaves out some characteristics.

Machine learning can increase the reliability of analyses to subpopulations. Machine learning models have flexible structures and arrive at predictions based on micro-relationships. As a result, there is less risk that predictions for a specific subgroup will be influenced by other parts of the insured population, thereby increasing the reliability of analyses to subpopulations. As such, machine learning is a useful tool for gaining a better understanding of the healthcare dynamics of a population.

Follow us on LinkedIn:
https://www.linkedin.com/company/milliman-benelux/